MSc Applied AI and Data Science

A study of the impact of ambient air pollution on the severity of asthma patients in London

Lekshmi Vijayakumar Sudha Kumari

FACULTY OF BUSINESS, LAW AND DIGITAL TECHNOLOGIES

September 2022

SOLENT UNIVERSITY

FACULTY OF BUSINESS LAW AND DIGITAL TECHNOLOGIES

MSc Applied AI and Data Science
Academic Year 2021-2022

Lekshmi Vijayakumar Sudha Kumari

A study of the impact of ambient air pollution on the the severity of asthma patients in London

Supervisors: Dr. Hamidreza Soltani, Dr. Olufemi Isiaq
September 2022

This report is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc Applied AI and Data Science.

Abstract

The impact of air pollution has been proven detrimental to human life all over the world. Epidemiological research has revealed that short-term exposure to air pollution can lead to asthma exacerbation and increased respiratory and cardiovascular hospital admissions and mortality. This study conducts a correlation analysis to measure the relationship between air pollution and asthma casualties in London using multiple techniques.

For this analysis, only fourteen of the thirty-two boroughs in London were chosen based on the availability of pollutant data and a new dataset was compiled with the daily measured values of Nitrogen-dioxide (NO_2), Ozone (O_3), Particulate Matter-10 (PM_{10}), Particulate Matter-2.5 ($PM_{2.5}$), Wind speed, Wind direction and Temperature from the fourteen boroughs for the period 2013-2019. After correlation analysis, an autoencoder-based deep learning clustering model called Deep Embedded Clustering (DEC) was applied to the compiled dataset. The performance of the DEC model was compared against the traditional K-Means clustering using the cluster evaluation metrics like the Silhouette Coefficient score, Calinski-Harabasz index and Davies-Bouldin index.

In the correlation analysis, Pearson's correlation, Spearman's correlation and Kendall's correlation methods indicated a low correlation between air pollutant concentrations with asthma admissions. A moderate positive correlation of asthma with five variables like no, borough, o3, no2 and nox was detected using the Distance correlation metric. The Mutual Information method derived a strong positive correlation for the variables borough, no, nox and no2, a moderate positive correlation for the variables pm10, and o3, and a low positive correlation for the variables ws and wd. Besides the only negative correlation of asthma was with air_temp as the asthma exacerbations are triggered at low temperatures.

It was observed that the values of all three metrics for DEC improved as compared to that of the baseline K-Means clustering. Silhouette Coefficient and Calinski-Harabasz Index for DEC was more than double the value of baseline metric values while Davies-Bouldin Index reduced by 36%, which indicated better cluster outputs. However, the clustering time increased to 24 times the baseline value. The ANOVA test showed that there is no correlation between cluster labels and asthma admissions.

Contents

List of Figures

List of Tables

Abbreviations

ANN	Artificial Neural Network
API	Application Programming Interface
AQE	Air Quality England
AURN	Automatic Urban and Rural Network
CO	Carbon-monoxide
CO_2	Carbon-dioxide
COPD	Chronic Obstructive Pulmonary Disease
CRISP-DM	Cross-industry standard process for data mining
DEC	Deep Embedded Clustering
DNN	Denoising Neural Network
EDV	Emergency Department Visits
GAM	Generalised Additive Model
GDP	Gross Domestic Product
GLA	Greater London Area
ICD-10	International Classification of Diseases Tenth Revision
KCL	King's College, London
KL	Kullback–Leibler
KMO	Kaiser-Meyer-Olkin
LUR	Land Use Regression
MVR	Missing Value Ratio
NHS	National Health Service
NO_2	Nitrogen-dioxide
ONS	Office for National Statistics
PCA	Principal Component Analysis
PM_{10}	Particulate Matter 10 (particles < 10μm in size)
$PM_{2.5}$	Particulate Matter 2.5 (particles < 2.5μm in size)
PPMCC	Pearson Product-Moment Correlation Coefficient
ReLU	Rectified Linear Unit
SAE	Stacked Auto-Encoder
SGD	Stochastic Gradient Descent
SO_2	Sulphur-dioxide
WHO	World Health Organisation

Acknowledgement

I would like to express my sincere gratitude to my supervisor Dr Hamidreza Soltani for his support, guidance, and encouragement throughout this project which made this research a truly inspiring experience. I would also like to thank my module leader Dr Olufemi Isiaq for his timely feedback and thoughtful suggestions. Furthermore, I am extremely grateful to my tutors Dr Shakeel Ahmad, Dr Drishty Sobnath and Prins Butt for their dedicated and patient efforts to equip me with the knowledge and skills to complete this research. I extend my heartfelt gratitude to all staff in various departments, especially the Faculty of Business Law and Digital Technologies at Solent university who have been a part of my memorable learning experience at the university.

I am thankful to the Office for National Statistics and the NHS Digital team for providing me with the asthma dataset and necessary information regarding the same.

Finally, I am forever grateful to all my family and friends for their constant motivation and deep belief in my capabilities which propelled me to push my limits further to achieve my goals.

1. INTRODUCTION

This chapter gives a brief overview, and insight into the topic, aims and objectives of this research project.

1.1 Overview

Asthma is one of the most common and widely discussed respiratory diseases which has attracted the attention of many researchers around the globe. It is a condition in which the small airways in the lungs become narrow due to inflammation resulting in symptoms like wheezing, coughing, shortness of breath and chest tightness.

The high prevalence of asthma in industrialised countries can be related to the increased concentrations of air pollutants like particulate matter ($PM_{2.5}$, PM_{10}), gaseous pollutants (Ozone, Nitrogen dioxide, Sulphur dioxide) etc [1]. The main sources of air pollution include road transport, the burning of fuel, emissions from power generation, industries, and even natural sources like dust, volcanoes, pollen, sandstorms, and soil. Lai et. al (2009) concluded in their study that even though the more affluent countries have a higher prevalence of asthma, its severity is higher in less affluent countries due to the lack of timely diagnosis and treatment [2].

It is important to note that in the year 2019, asthma affected the lives of around 262 million people and caused 461,000 deaths as per the WHO reports[3]. Both outdoor and indoor air pollution can cause asthma exacerbations and flare-ups in individuals with existing lung conditions, particularly in children leading to life-threatening conditions [4]. A recent report by Asthma and Lung UK (Asthma UK and British Lung Foundation partnership) reveals that 53% of people with asthma and 47% of people with COPD have symptoms triggered by toxic air[5]. The highly contaminated air has also led to a spike in the number of hospital admissions and emergency department visits in various countries. For example, between the years 2017-2019, the poor air quality in London alone caused more than 1,700 hospital admissions for asthma and other serious lung conditions in the UK according to a new analysis by Imperial College, London. Exposure to air pollution during pregnancy can affect the baby's lung development in the womb. Such babies are more likely to experience asthma and other lung complications after birth.

Air pollution has affected even the daily lives of asthma patients since they are afraid to leave their homes and go out for leisure activities, daily exercise, visit their friends and families etc. Even though it was previously believed that air pollution only triggers asthma exacerbation, clinical research has proved that exposure to air pollution for a prolonged period might lead to the outset of asthma and delayed lung development in children.

Despite being the topic of interest for many researchers, there are only a few works of literature analysing the correlation between air pollution and asthma severity in the UK where 12% of the total population has been diagnosed with asthma. Therefore, this study using machine learning techniques is conducted as an attempt to make up for the deficit of research on this topic.

For administrative purposes, the Greater London Area (GLA) is split into thirty-two local authority districts also known as boroughs out of which twenty are outer London and the remaining twelve are inner London boroughs. The City of London is not considered a borough. This study is based on the data from the monitoring sites located in the fourteen boroughs of London selected based on the amount of data available for the analysis.

Inner London	Outer London
Camden	Barking and Dagenham
Greenwich	Barnet
Hackney	Bexley
Hammersmith and Fulham	Brent
Islington	Bromley
Kensington and Chelsea	Croydon
Lambeth	Ealing
Lewisham	Enfield
Southwark	Haringey
Tower Hamlets	Harrow
Wandsworth	Havering
Westminster	Hillingdon
	Hounslow
	Kingston upon Thames

	Merton Newham Redbridge Richmond upon Thames Sutton Waltham Forest

Table 1: Boroughs in London

1.2 Research Question

Is there a relationship between ambient air pollution and the severity of asthma patients in the different boroughs of London?

1.3 Aims and Objectives

The proposed research aims to explore the impact of the various air pollutant concentrations on the number of casualties caused by asthma in the different boroughs of London for the period 2013 to 2019 using machine learning and deep learning techniques.

The objectives of this study include

- To determine the correlation between the major outdoor air pollutant quantities and the number of casualties due to asthma in different boroughs of London.
- To apply a deep clustering algorithm namely Deep Embedded Clustering (DEC) to the collected data to group the data samples into multiple clusters.
- To explore the DEC output clusters and analyse the pollutant data and the asthma casualty count in the boroughs grouped under each cluster separately.

The traditional K-Means clustering is used as a baseline method to compare the results of the deep clustering performance in which clustering is applied to the learned feature representation vectors. The main purpose of using deep clustering is that the performance of the traditional clustering algorithms is usually superior when applied to the features in latent space.

The research background and literature review, methodology, implementation, project plan, results and discussion, as well as the conclusion of the research study are discussed in the following chapters.

2. BACKGROUND AND LITERATURE REVIEW

As mentioned before, this study is intended for a detailed analysis of the data from the fourteen selected boroughs of London for the years 2013-2019 only. For this research, London was selected since the air quality index for London is lower when compared to other areas of the UK. In the year 2020, after the outbreak of the COVID-19 pandemic around the globe, air pollution levels significantly dropped due to reduced traffic and industrial operations. However, the pandemic had a massive impact on asthma patients as it worsened their symptoms and even led to death in some cases. Therefore, data after 2019 are not considered for this study since the main reason behind the increased asthma admissions and mortality rates from 2020 onwards would be the pandemic rather than air pollution.

The related works which coalesce both the correlation analysis of air pollution with asthma and deep clustering based on air pollutant concentrations were not found. Hence the review of the studies which explore the correlation of asthma with air pollution and weather using different machine learning and statistical techniques are included. The review articles are selected from Google Scholar and IEEE Digital Library websites using a combination of the keywords like “asthma”, “air pollution”, “machine learning”, “time series” etc. using the year range 2010- 2022 and sorted by relevance.

In the research conducted by Aditya Narayan et al.,[6] to find the correlation between air pollution and asthma/Gross Domestic Product, the data contained the daily readings of the main pollutants namely $PM_{2.5}$, Carbon Monoxide, Sulphur Dioxide, PM_{10}, Ozone, and Nitrogen Dioxide were collected from 20 American states for the previous 20 years. Among the six pollutants, only 3 namely $PM_{2.5}$, Carbon Monoxide (CO) and Sulphur dioxide (SO_2) were selected for further analysis since they had a high value for Pearson’s coefficient which implied a strong correlation. Using PCA, the dimensionality was reduced and the target variables i.e., Asthma cases/GDP were divided into five categories. The data was then modelled using Random Forest (RFC), Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) for asthma prediction and the results were compared using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-Squared metrics among which SVM was found to be the most accurate with an average accuracy of 70%

for asthma prediction. It was concluded that a positive correlation exists between asthma rates and air pollution/GDP.

Another research on correlational analysis was conducted by Hajmohammadi, H et al.[7] aimed to find the relationship between asthma exacerbations and daily concentration of the oxides of Nitrogen by developing a time series regression model using the data available from eight different monitoring stations in East London. The study was centred on the data of 33,672 patients diagnosed with asthma who were prescribed oral steroids from primary health centres because of asthma exacerbations. The daily ambient temperature measurements from a station in East London and the daily measurements of NO_x from eight roadside and background monitoring stations were the independent features. A significant positive relationship was found between the daily NO_x concentration procured particularly from roadside stations and the count of oral steroids prescriptions in the following three weeks and a negative relationship with ambient temperature.

Along with air pollution, Lee, Eu Sun et al.[8] included weather also on the impact study on the emergency department visits of patients with respiratory diseases. The study which implemented RandomForestRegressor on 525,579 patients' data revealed that PM_{10}, temperature and steam pressure was the main cause for most of the patients visiting ED with acute upper respiratory infections, influenza and pneumonia. Also, the top three reasons for emergency admissions were mostly influenced by $PM_{10.}$

Air pollution has a major impact on pediatric asthma and both short-term and long-term effects of air pollution components caused by emissions from industries and traffic vehicles in Cleveland, OH, were evaluated by Khatri, Sumita B et al.[9]. The research used data from Cleveland Multiple Air Pollutant Study (CMAPS) and used the LUR (Land Use Regression) model, which is a method to describe concentrations of air pollutants at a smaller number of sites and develop a stochastic model using predictor variables. The LUR models determined the long-term exposures to NO_2 and PM_{10} whereas US EPA's Positive Matrix Factorization (EPA PMF 5.1) estimated the short-term exposures to $PM_{2.5}$ and PM_{10} components which were used to analyse the ED visits due to asthma and the seasonal-averaged time series for both short-term and long-term effects. The study found NO_2 and PM_{10} associated with motor traffic causing an increase in the ED visits of asthmatic

children. Further, the short-term exposures to PM_{10} and $PM_{2.5}$ associated with steel production also significantly affected the increased ED visits due to pediatric asthma.

Delamater P. L. et. al.[10] carried out another analysis related to the effect of air pollution and weather on asthma hospitalisation in Los Angeles, a highly polluted county in California. The CO, NO_2, O_3, PM_{10}, $PM_{2.5}$, maximum temperature, and relative humidity monthly measurements were collected from all over the county fofrom001 to 2008 and modelled using Bayesian Regression with temporal random effects. The model evaluation was performed using the goodness of fit criterion and the ability of the model to predict asthma hospitalisations. The variables CO, NO_2, and $PM_{2.5}$ showed a significant positive relationship and O_3 was found non-significant in all the models whereas PM_{10}, relative humidity, and maximum temperature exhibited mixed relationships.

A cross-over analysis of air pollutant concentrations and the EMS ambulance cases due to asthma attacks was conducted by L.H.[11]. This study investigated the relationship between O_3, $PM_{2.5}$, NO_2, SO_2 and CO pollution and 11754 EMS system calls in Houston, Texas over eight years from 2004 to 2011. The conditional Logistic Regression model was selected based on the low AIC criterion and the results revealed that O_3 and NO_2 were the main factors triggering asthma attacks. Besides, their risk was found to be increased at high concentrations when both were combined.

The correlation analysis of air pollutants and respiratory diseases investigated by Qiumin Zhai et al.[12] used Pearson Product Moment Correlation Coefficient (PPMCC) as the metric on the monthly air pollutant data and the record of inpatients with respiratory diseases from a hospital in the selected city. The air pollutant variables include PM_{10}, NO_2, SO_2 etc. and the respiratory diseases included asthma, bronchitis and pneumonia. The computed correlation coefficients showed that the patient admissions due to bronchitis showed a high correlation with NO_2 and PM_{10} quantities with coefficient values of 0.80 and 0.72 respectively. The variable PM_{10} had a positive linear relationship with asthma admissions with a coefficient value of 0.7 while pneumonia inpatient data had a correlation coefficient value of 0.55 with SO_2 concentration.

In the research conducted by Akinbami, L.J. et al[13] to find the correlation between the prevalence of childhood asthma and the air pollutants in metropolitan areas using an average of 12 months of data for each county in the US. The data was extracted from a

survey(2001-2004) which included 34,073 samples of children between the ages 3 and 7 from US metropolitan areas. The estimation of the relationship between asthma prevalence and each pollutant was modelled separately using Logistic regression both as continuous values and by splitting into equal parts. The counties with O_3 and PM pollution were observed to have a high number of cases currently suffering from asthma or had a recent attack whereas SO_2 and NO_2 concentrations showed no relationship with asthma prevalence.

Cox, L.A.[14] included the socio-economic factors along with air pollution in his study to analyse the correlation with the risk of adult-onset asthma, stroke and heart attack among US citizens. The study was based on the variables like self-reported asthma, heart attack, stroke experiences, gender, age, education, income and smoking status of adults over 50 years of age as well as the average yearly concentrations of O_3 and $PM_{2.5}$. The three health conditions had a positive correlation with each other. In the logistic regression model, if $PM_{2.5}$ is regressed against age, sex, and ever-smoking status only, then $PM_{2.5}$ showed a significant negative correlation with both stroke and heart attack risk. However, when income was also considered, $PM_{2.5}$ did not have any significant relation with health risks. In all three models Logistic Regression, Multiple Linear Regression, and Regression Tree, $PM_{2.5}$ had a significantly negative association with asthma risk. The author suggested that this could be due to the confounders and residual confounding.

The cause-effect relationship between air pollution and asthma also led to the development of an asthma attack prediction system, an android application with the aid of a pollutant monitoring device by M. N. Hoq et al[15]. After system installation, patient registration and the blue-tooth connection to the air pollution monitoring device which measures CO, NO, smoke and dust are performed. It then starts measuring air pollutant concentrations and the learning phase by asking a few questions to the user regarding the quality of the air and the asthmatic symptoms of the patient in regular short intervals. Based on the time series data of air pollutants measured and the input from the user on his/her asthmatic condition, the system learns to predict the asthma attack possibility using supervised learning. When the air pollutant's threshold level is exceeded, the system alerts the user using an alarm tone and suggests the user adopt precautionary measures.

3. METHODOLOGY

This chapter explains the methodology and the method followed in the implementation of this project.

3.1 Introduction

The primary objective of this research i.e., to determine the correlation between atmospheric pollutant concentrations and asthma severities is conducted using the quantitative correlational research methodology. This research uses this methodology to correlate two or more variables and identify the patterns, relationships, and trends between variables using statistical analysis methods using already available data from the website of different organisations. The CRISP-DM (Cross-industry standard process for data mining) has been used as a process model guide to carry out the different tasks involved in the realisation of this project.

3.2 CRISP-DM

CRISP-DM[16] is a complete blueprint for effective data mining which was proposed in 1996. It can be easily customised according to the needs of the topic of the study and can be applied to any industry. According to this generic process model, the entire life cycle of any data mining project can be divided into six phases namely Business understanding, Data understanding, Data preparation, Modelling, Evaluation and Deployment[17].

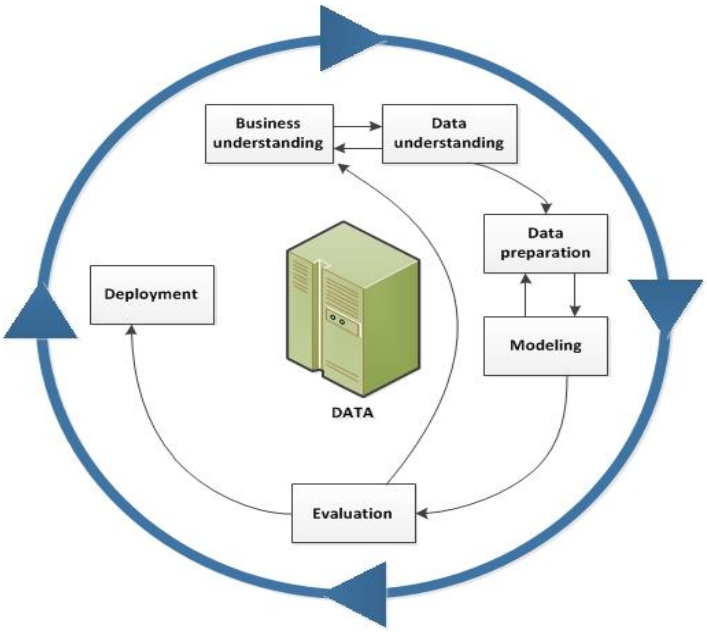


Figure 1: Phases of CRISP_DM (Source: IBM)

The sequence of the process flow indicated using the arrows is the most frequent and important one. However, the order is not exactly followed in most cases as the outcome of a particular phase decides the next phase or task to be performed. The cyclic pattern of the data mining is represented by the outer circle since the project cannot be considered complete even after the deployment of the solution. This is because the entire process of deriving the solution can unveil new research areas which are benefitted from past studies in a particular research area.

A brief overview of the tasks performed in each of the six phases of this methodology is discussed below based on the process adopted in this research study.

3.2.1 Business Understanding

This is the most important phase in a data mining project in which the objectives and success criteria of the project from a business perspective are decided first and then the requirements, risks and contingencies are analysed. Based on this, data mining goals are determined, and a project plan is outlined with an initial assessment of intended tools and techniques.

3.2.2 Data Understanding

This phase is closely related to the Business Understanding phase since the conversion of business objectives to data mining goals requires minimal knowledge of the data. This phase starts with the data collection followed by familiarisation of the data, data exploration and verification of the data quality.

3.2.3 Data Preparation

The Data Preparation phase focus on the conversion of the collected raw data to the final data which is used for modelling purposes and involves various tasks such as data integration from multiple sources, data formatting, data cleaning etc. This phase is often repeated multiple times to create an informative dataset to maximize the accuracy of the model outputs.

3.2.4 Modelling

This phase includes the selection of appropriate modelling tools and techniques, model implementation, test design, model evaluation and optimisation of the parameters. It may be required to return to the data preparation phase to ensure that the model yields the desired results.

3.2.5 Evaluation

In this phase, the selected model is first evaluated to determine how and then reviewed to ensure that none of the business objectives is overlooked and the model outputs align closely with the data mining goals. Based on the model evaluation and review results, the next phase is decided on whether to navigate back to the business understanding phase or to proceed with the model deployment phase.

3.2.6 Deployment

A data mining project may not enter this last phase if the results obtained during the previous phases are unsatisfactory which might require the reiteration of the entire process particularly when the requirements are too complex. However, the deployment phase can be concluded with a detailed report if the requirements are simple.

The tasks performed in the various phases of this data mining project are discussed in detail in their respective sections. The deployment phase is not carried out as it is not within the scope of this project.

4. IMPLEMENTATION

A detailed explanation of the various steps in the implementation is mentioned in this chapter.

4.1 Overview

The data employed in this research is downloaded using R and the data wrangling, analysis and deep clustering are performed using Python. The source code can be found here in the GitHub repository. The hardware and software details of the resources used in the implementation of this project are given in the below table.

System specifications	
OS	Windows 10 Pro
Processor	Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.20 GHz
RAM	8GB
Tools and Software	
Data download tools	RStudio 2022.07.1+554 R for Windows 4.2.1 Openair (R package)
IDE	PyCharm 2021.3.1
Interpreter	Python 3.8
Data preparation and analysis libraries	Jupyter 1.0.0 Pandas 1.3.5 NumPy 1.21.0 SciPy 1.8.0 Scikit-learn 1.0.2 Yellowbrick 1.4

	dcor 0.5.6 ennemi 1.2.0 data-science-utils 1.7.1
Deep Clustering libraries	Tensorflow 2.9.1
Visualisation libraries	Seaborn 0.11.2 Matplotlib 3.5.1

Table 2: Hardware and Software used in the implementation

4.2 Data Ingestion

The main challenge faced in the data collection step was to obtain the air pollution data and asthma data grouped by the borough for the period 2013-2019. The data for this study was gathered from government and non-government sources which provide comparatively reliable and accurate information. The annual mortality rate for London grouped by the boroughs was collected from the Office for National Statistics (ONS) team and the data regarding the annual hospital admissions for asthma in the different London boroughs were obtained from the NHS Digital team which includes the cases in which the root cause of death was mentioned as J45-J46 'Asthma', using the International Classification of Diseases Tenth Revision (ICD-10) codes.

The London boroughs operate and fund several monitoring sites at around a hundred distinct locations to monitor the air quality of London continuously. The real-time and historical data of these sites are available on the UK-Air, London Air and Air Quality England websites. These data can also be conveniently downloaded using RStudio by installing a package called openair [18] developed by David Carslaw for analysing data air quality data or atmospheric composition data. The openair-R package provides a family of functions and a wide range of parameters which provide extensive access to the air quality data from several air-quality networks in the UK namely Automatic Urban and Rural Network (AURN), Air Quality England (AQE), King's College, London (KCL) etc. which are updated every day.

Before importing the data, meta-data of each network such as the pollutants measured, measured units, site codes, the type of site (Urban or Rural), the local authority of the sites etc. can be obtained using the importMeta() function specifying the source as “AURN”, “AQE” or “KCL”. For example, the following line of code will return all the metadata for AURN as a tibble.

```
importMeta(source = "kcl", all = TRUE)
```

In this project, the importMeta() function has been used to identify the site codes corresponding to each of the London boroughs and any one of the functions mentioned in the below table can be used to download the corresponding network’s data.

Function	Description
importAURN()	Import air-quality data from the sites in the main UK network i.e., AURN.
importAQE()	Import air-quality data from sites under the AQE network.
importKCL()	Import air-quality data from sites operated by the KCL.

Table 3: Functions to import data from different network sources

The following line of code can be used to retrieve the hourly data of all the pollutants from the sites HR1 and HR2 from the year 2013 to 2019.

```
importKCL(site=c("HR1","HR2"), year=2013:2019, pollutant="all")
```

The imported data from each source for individual boroughs were written as comma-separated files into respective folders in the data directory. Also, the data from additional sites were downloaded from the London Air website in the form of CSV files and saved in a different folder. It is important to note that not all the networks provide the data for all the boroughs as they do not operate monitoring sites in all the London boroughs. In

addition, some sites were closed in between and at the same time some new sites were opened during the seven years from 2013 to 2019.

4.3 Data Wrangling

The data integration, outlier treatment, missing value replacement and resampling of the data were performed in this step.

4.3.1 Data Merging

As the first step in data integration, the data files from different source folders were simply merged to create a single file for each borough and saved in a different folder in the same path as the source folders. This folder had thirty-three files corresponding to the 32 boroughs and the City of London. The annual borough-wise asthma admissions were also compiled into a file named “asthma.csv”.

Secondly, the missing value ratio of all the features for the individual borough files was calculated after temporary downsampling from hourly data to daily data and saved to a file for selecting the boroughs and features. This downsampling based on the mean value was done to reduce the missing occurrence as the data was not available for all the hours for some days. By careful consideration of the missing value ratio (MVR), fourteen boroughs and 11 features including date and borough were selected for the final analysis.

4.3.2 Outlier and Missing value treatment

As the next step, treatment of outliers and missing values was performed to cleanse the data. For this first, a resampling was done because the data has multiple rows for the same hour due to the merging of data from multiple sources. This ensured that there was only a single record corresponding to a particular hour value for the individual boroughs. The strategy was to handle the outliers first and then manage the missing values since the outliers can affect adversely the mean, median, and mode values used in the missing value imputation. The data consisted of many negative and zero values which might have occurred because of the calibration errors in the sensors measuring the values or the processing errors while storing or retrieving from the database. The following boxplot shows the data distribution and outliers for the borough Barking ad Dagenham.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61344 entries, 0 to 61343
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   date      61344 non-null  object
 1   borough   61344 non-null  object
 2   nox       60168 non-null  float64
 3   no2       60168 non-null  float64
 4   no        60168 non-null  float64
 5   o3        59925 non-null  float64
 6   pm10      58657 non-null  float64
 7   pm2.5     58126 non-null  float64
 8   ws        59232 non-null  float64
 9   wd        59232 non-null  float64
 10  air_temp  59232 non-null  float64
dtypes: float64(9), object(2)
memory usage: 5.1+ MB
```

Figure 2: Data for Barking and Dagenham before outlier and missing value treatment

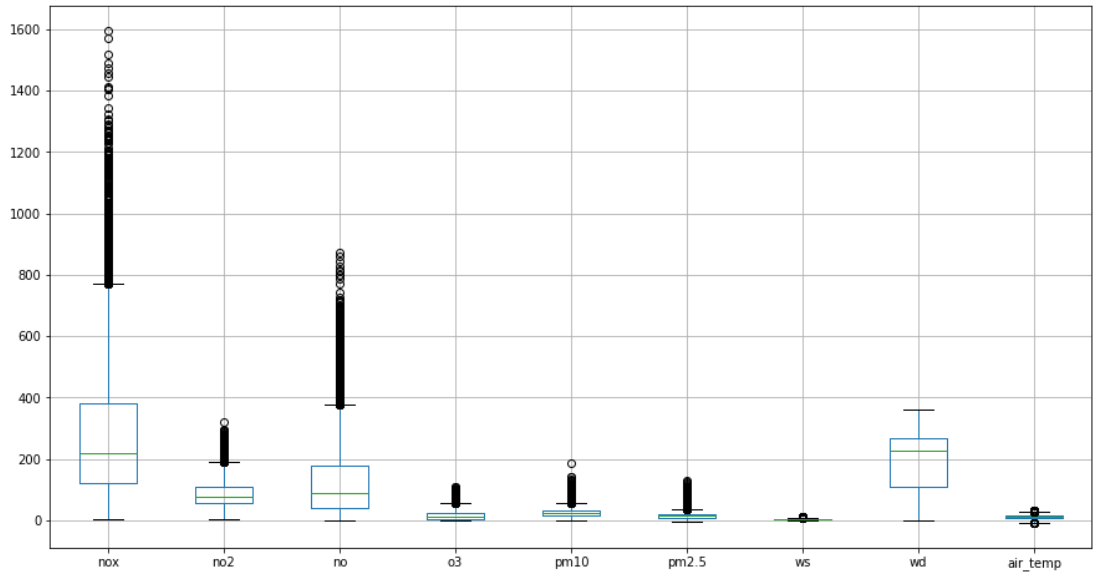
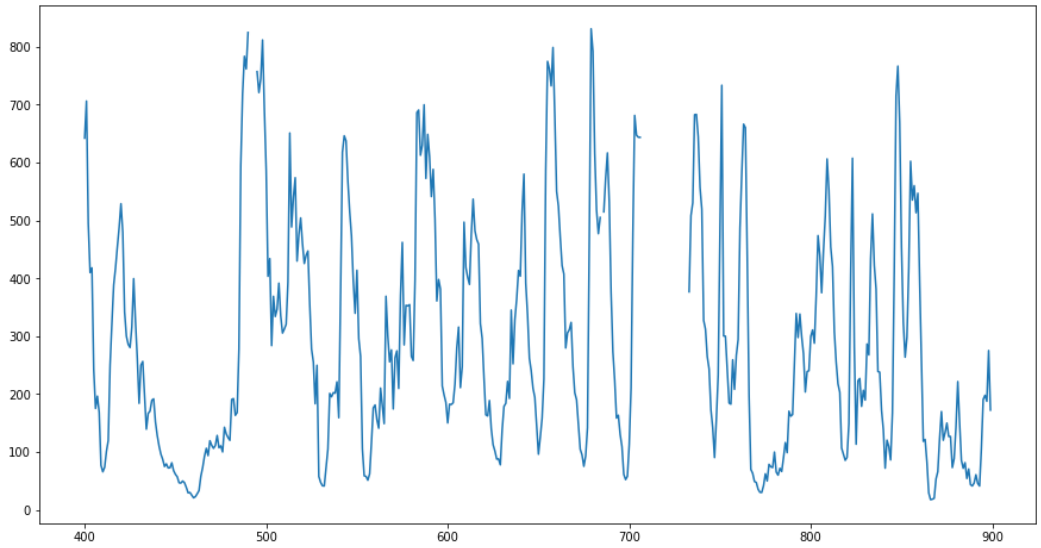


Figure 3: Boxplot for Barking and Dagenham before outlier and missing value treatment



was

Figure 4: "nox" plot for Barking and Dagenham before outlier & missing value treatment

As a part of the strategy to treat the outliers rather than deleting them, the zero and negative values except for the air_temp column were replaced with NaN values. For the air_temp column, negative values were valid as the temperature might fall below 0 during winter. To handle this the air_temp in degrees Celsius was converted to Fahrenheit by multiplying the value by 1.8 and then adding 32 to the product. In the next step, the Interquartile range (IQR) which is the difference between the 75th percentile (Q3) and the 25th percentile(Q1) values calculated for each column to determine the upper limit and lower limit values. All the values which do not lie in the range of the upper and lower limit values were considered outliers and replaced with NaN values during the missing value imputation process.

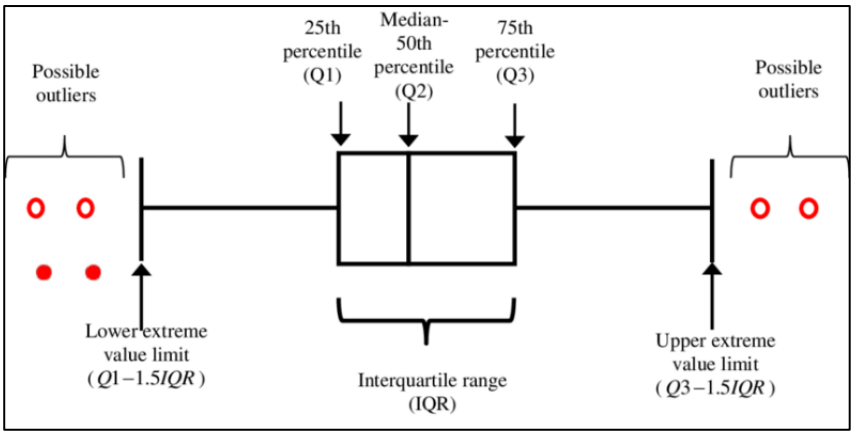


Figure 5: IQR in outlier detection (Source:ai-ml-analytics.com)

$$IQR = Q3 - Q1 \quad (1)$$

$$\text{Lower limit} = Q1 - 1.5 * IQR \quad (2)$$

$$\text{Upper limit} = Q3 + 1.5 * IQR \quad (3)$$

Following the treatment of outliers, as a trial, the missing values were imputed using the spline interpolation method which is one of the common methods for handling time series missing values[19]. But the imputation was performed as a smooth curve joining the two points between which the missing values exist. Hence instead of spline imputation, a multivariate imputing algorithm named sklearn.impute.KNNImputer was adopted which imputes missing values in each sample with the mean of "n" nearest neighbour's values rather than using the naïve approach of imputation by mean or median. The value of n is 5 by default and can be specified explicitly according to the data requirement.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61344 entries, 0 to 61343
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   nox        61344 non-null  float64
 1   no2        61344 non-null  float64
 2   no         61344 non-null  float64
 3   o3         61344 non-null  float64
 4   pm10       61344 non-null  float64
 5   pm2.5      61344 non-null  float64
 6   ws         61344 non-null  float64
 7   wd         61344 non-null  float64
 8   air_temp   61344 non-null  float64
 9   date       61344 non-null  object
 10  borough    61344 non-null  object
dtypes: float64(9), object(2)
memory usage: 5.1+ MB
```

Figure 6: Data for Barking and Dagenham after outlier and missing value treatment

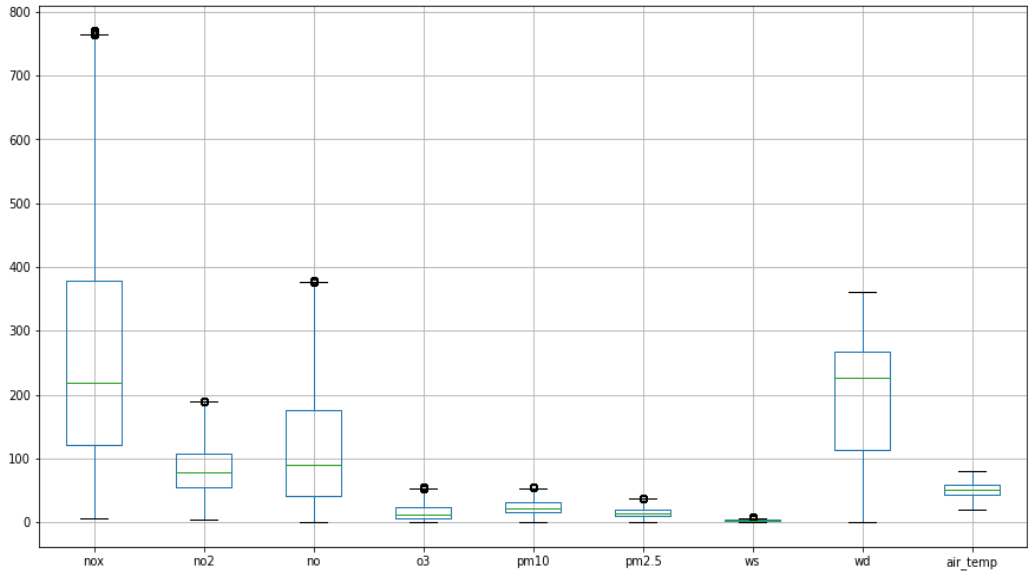


Figure 7: Boxplot for Barking and Dagenham after outlier and missing value treatment

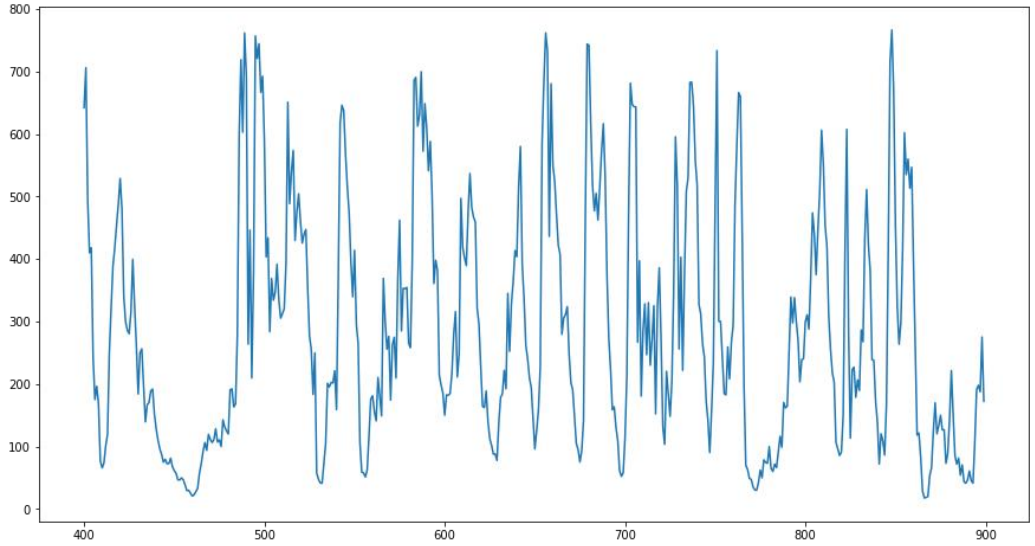


Figure 8: "nox" plot for Barking and Dagenham after outlier & missing value treatment

From Figure 4 and Figure 8, it can be observed that the missing values have been imputed whereas a few outliers very close to upper and lower limits are still showing in above Figure 7 because of the slight shift in IQR values after the imputation. However, most of them have been corrected by KNN imputation. The cleansed data is rounded to two decimal points and the data for each borough is then saved to a separate folder. Each data file will

contain 61344 records corresponding to the total no of hours in 7 years i.e., the sum of 7x365x24 hours and an additional 24 hours in the leap year 2016.

4.3.3 Data Resampling

Finally, the processed data from all the boroughs are first resampled to daily mean values and combined to create a single file for ease of analysis. The final file has 35784 rows and 11 columns namely date, borough, nox, no2, no, o3, pm10, pm2.5, ws, wd and air_temp.

4.4 Exploratory Data Analysis for correlation detection

The asthma data regarding hospital admissions were available only on an annual basis. Hence for conducting correlational analysis, the final pollutant data file is resampled with the annual mean value and combined with annual asthma data.

Variable	Description	Data type	Variable type
date	The date on which the reading was measured	object	Ordinal
borough	Name of the borough where the reading was measured	object	Nominal
nox	Nitrogen Oxide measurement.	float64	Ratio
no2	Nitrogen Dioxide measurement.	float64	Ratio
no	Nitric Oxide measurement.	float64	Ratio
o3	Ozone measurement.	float64	Ratio
pm10	Particulate matter 10 measurement	float64	Ratio
pm2.5	Particulate matter 2.5 measurement	float64	Ratio
ws	Wind Speed measurement	float64	Interval
wd	Wind Direction measurement	float64	Interval
air_temp	Air Temperature measurement	float64	Interval

admissions	Total number of hospital admissions where the primary cause was asthma	int	Ratio

Table 4: Feature details

4.4.1 Feature Encoding

The only relevant categorical feature used in the correlation analysis was the borough variable which was encoded to integer values starting from zero based on the alphabetical order of the values present in the variable. This step was performed using sklearn.preprocessing.LabelEncoder is essential as the machine can learn o on numeric values.

borough_names	borough_labels
Barking and Dagenham	0
Bexley	1
Camden	2
City of Westminster	3
Ealing	4
Greenwich	5
Haringey	6
Harrow	7
Hillingdon	8
Kensington and Chelsea	9
Lewisham	10
Richmond	11
Southwark	12
Tower Hamlets	13

Figure 9: Label Encoding of borough feature

4.4.2 Time series plot

The time series plots are used to analyse the trend of variables over time. The time series plots for the different pollutant variables are listed below.

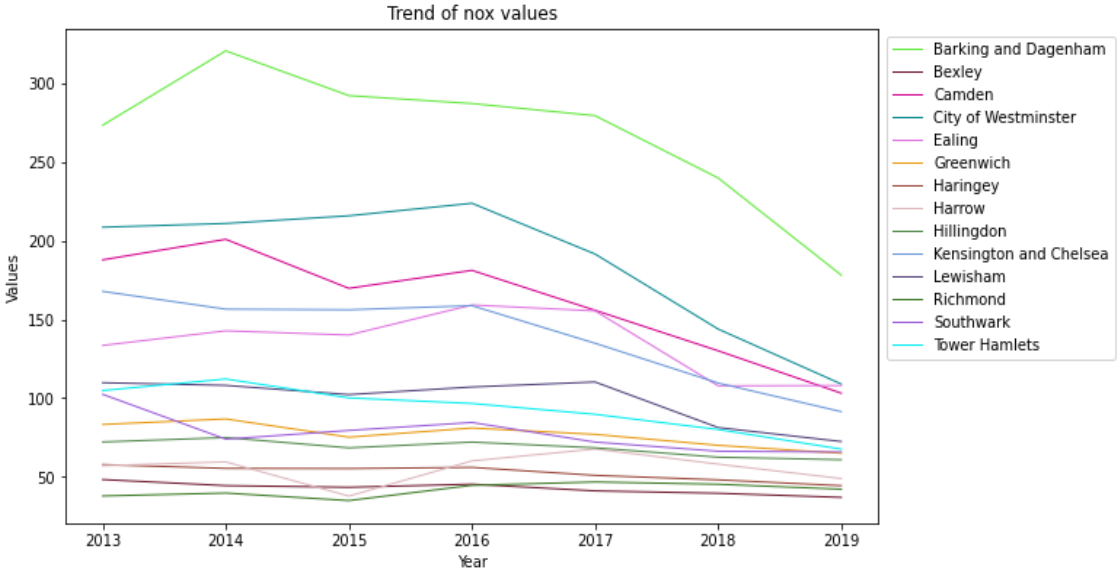


Figure 10: Trend of nox in different boroughs

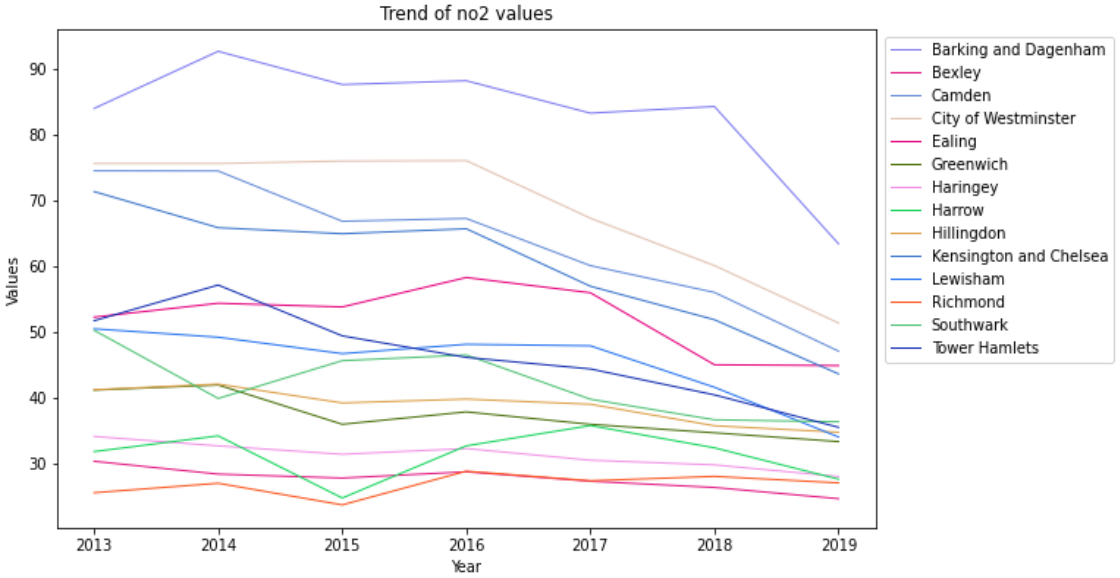


Figure 11: Trend of no2 in different boroughs

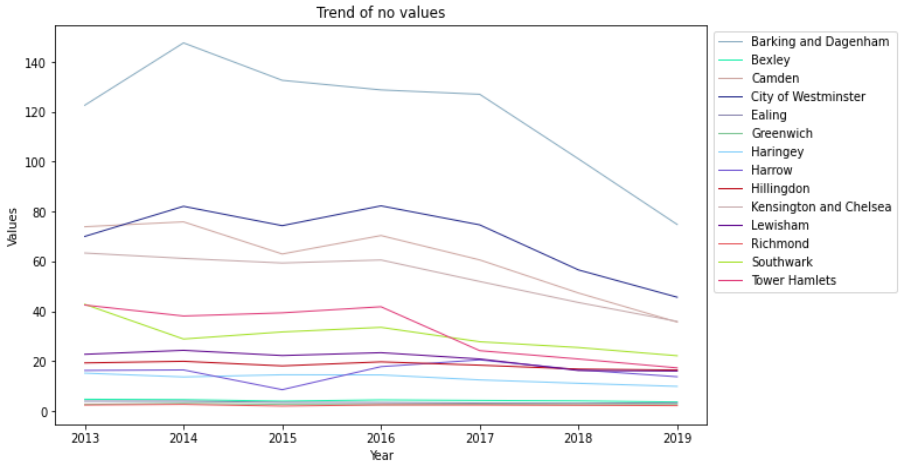


Figure 12: Trend of no in different boroughs

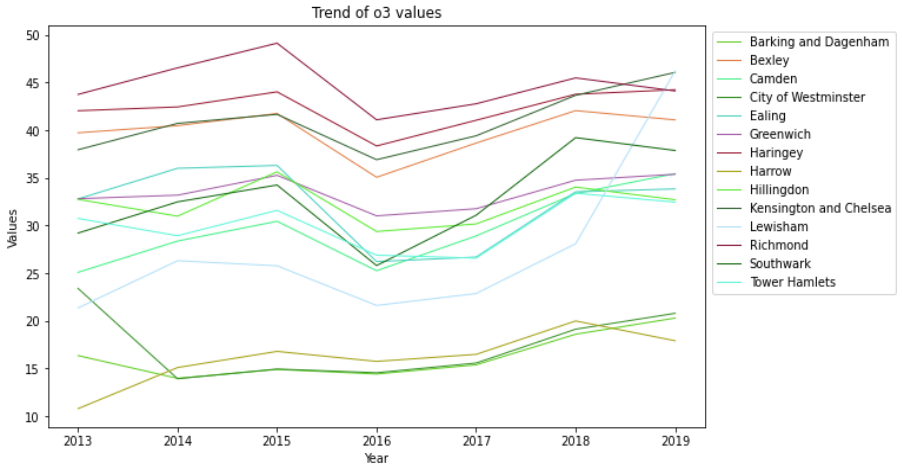


Figure 13: Trend of o3 in different boroughs

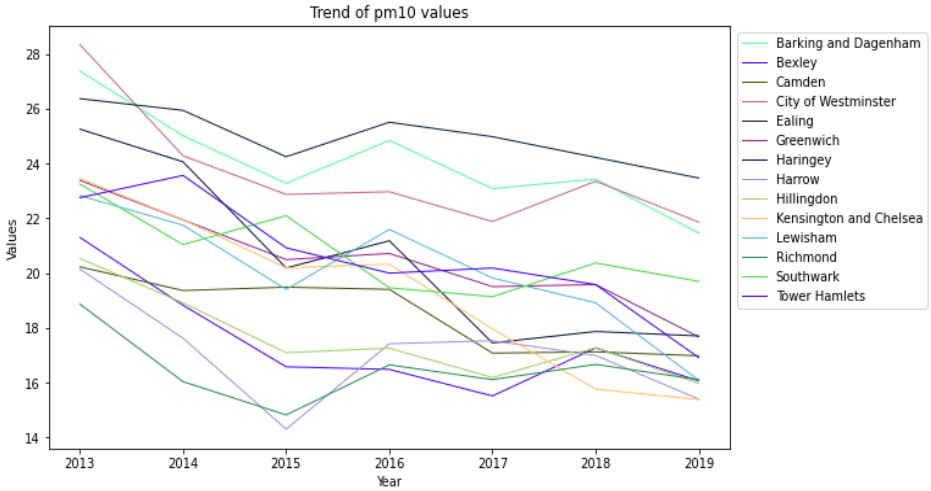


Figure 14: Trend of pm10 in different boroughs

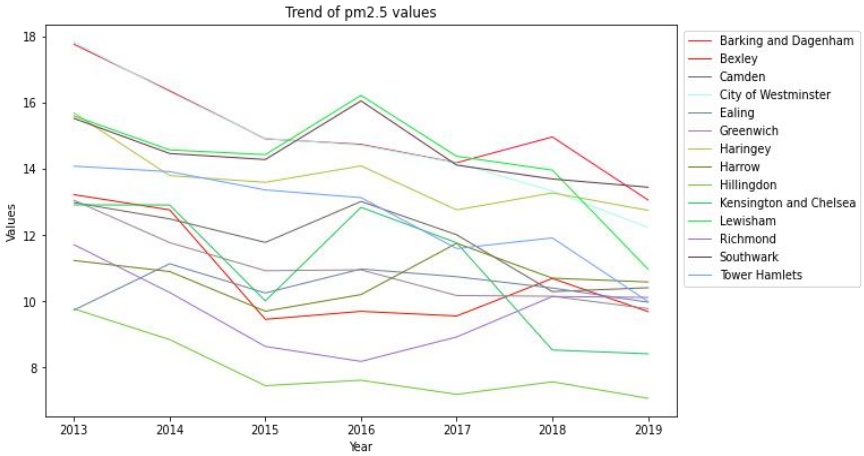


Figure 15: Trend of pm2.5 in different boroughs

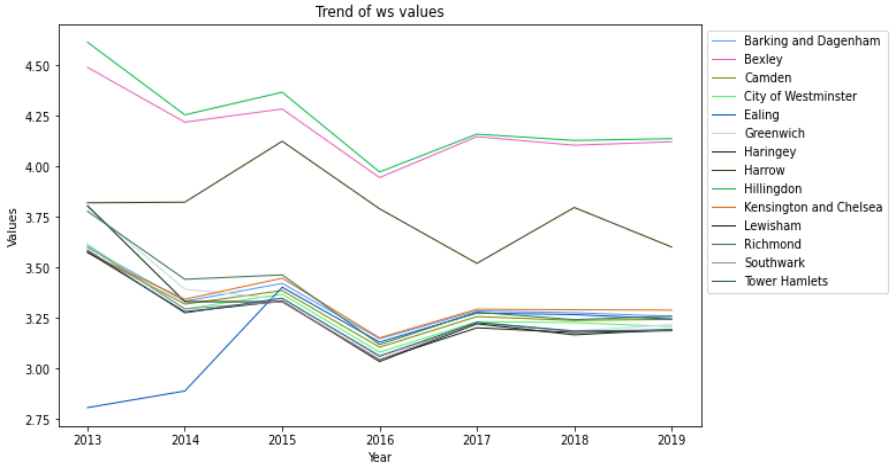


Figure 16: Trend of ws in different boroughs

Figure 17: Trend of wd in different boroughs

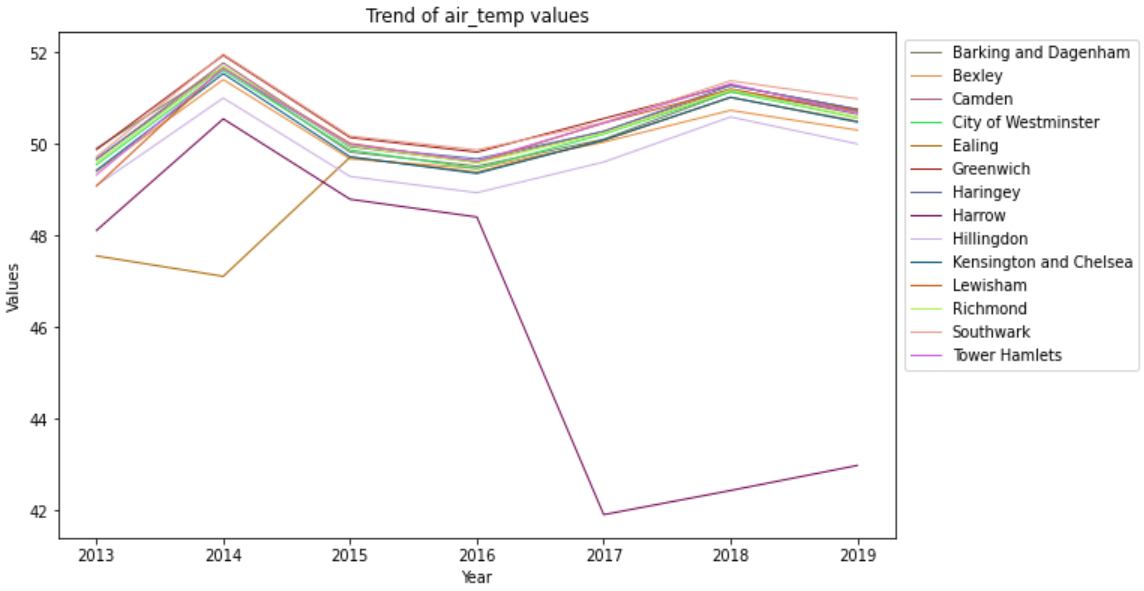


Figure 18: Trend of air_temp in different boroughs

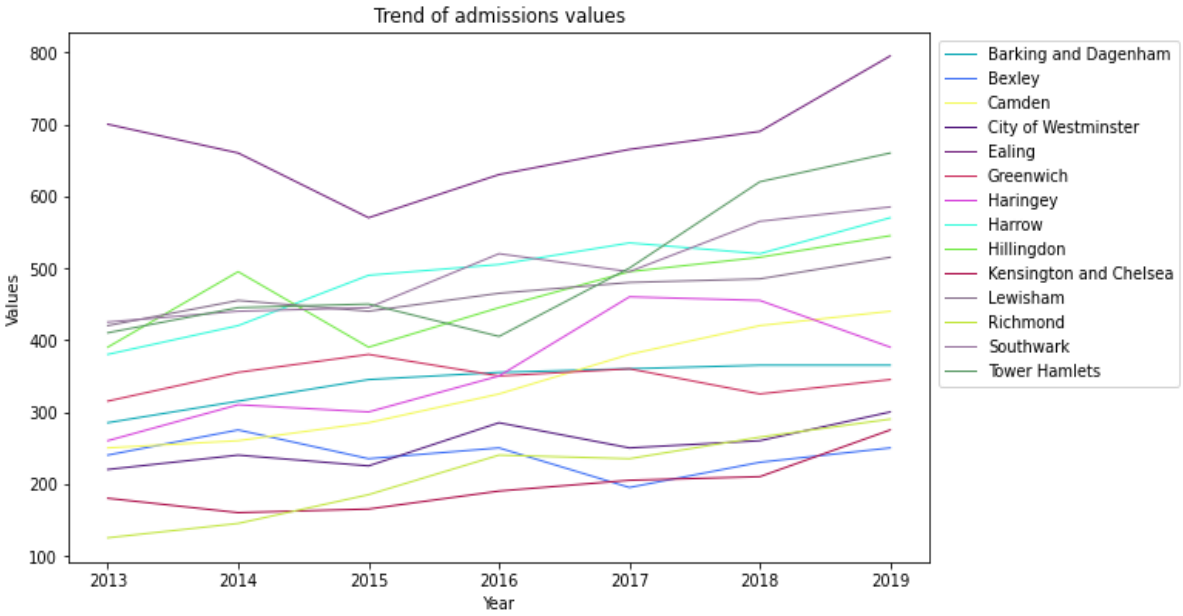


Figure 19: Trend of admissions in different boroughs

4.4.3 Histogram plot

The histogram plots are useful in analysing the data distribution, skewness and kurtosis of the data.

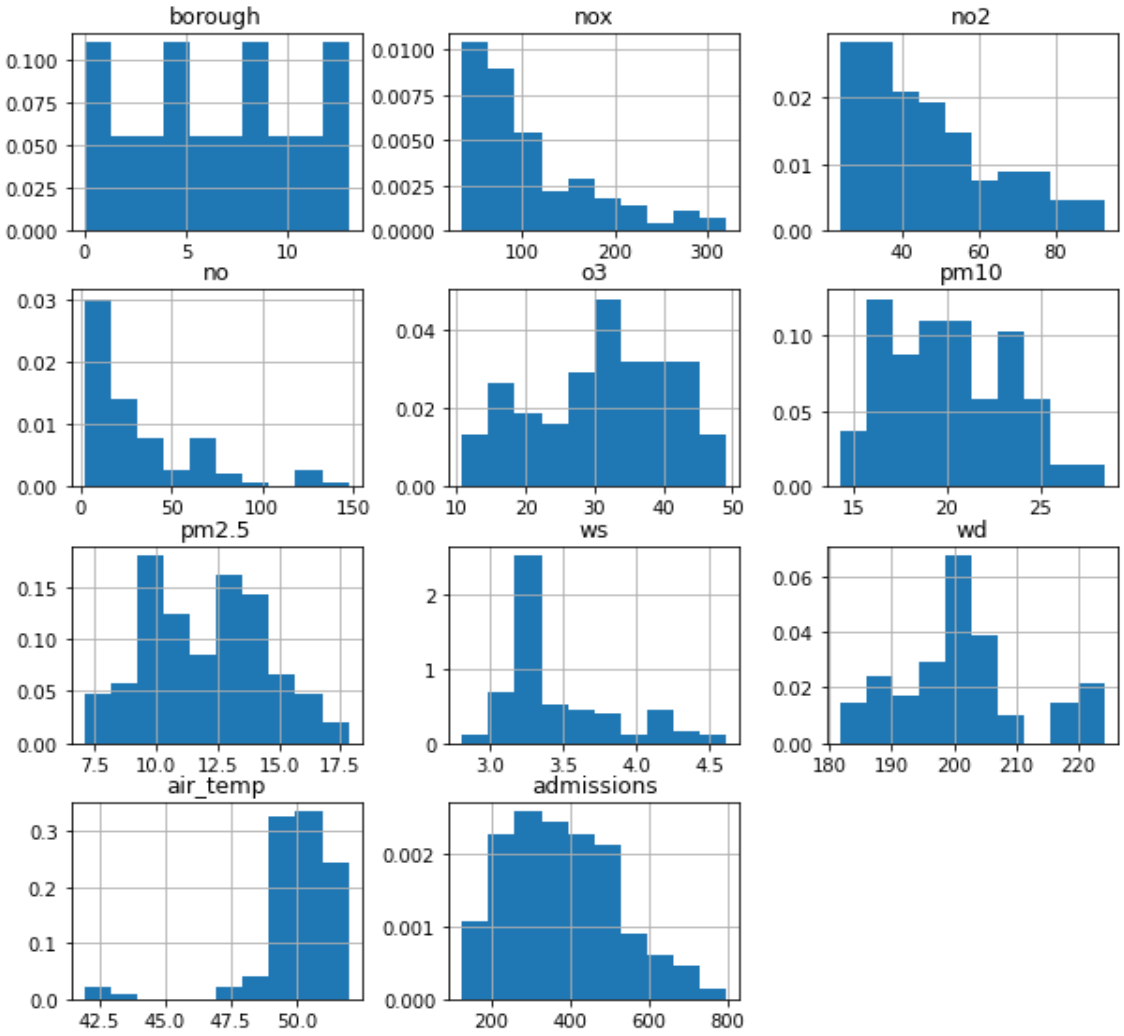


Figure 20: Histogram plot of the variables

The variables follow a skewed normal distribution with the variables nox, no2, no, pm10, ws, and admissions with the most skewness.

4.4.4 Box plot

The Box plot is used to identify the data distribution and the presence of outliers. As the outlier and missing value treatment have already been performed for individual borough files before resampling, the outliers shown in the figure below can be ignored.

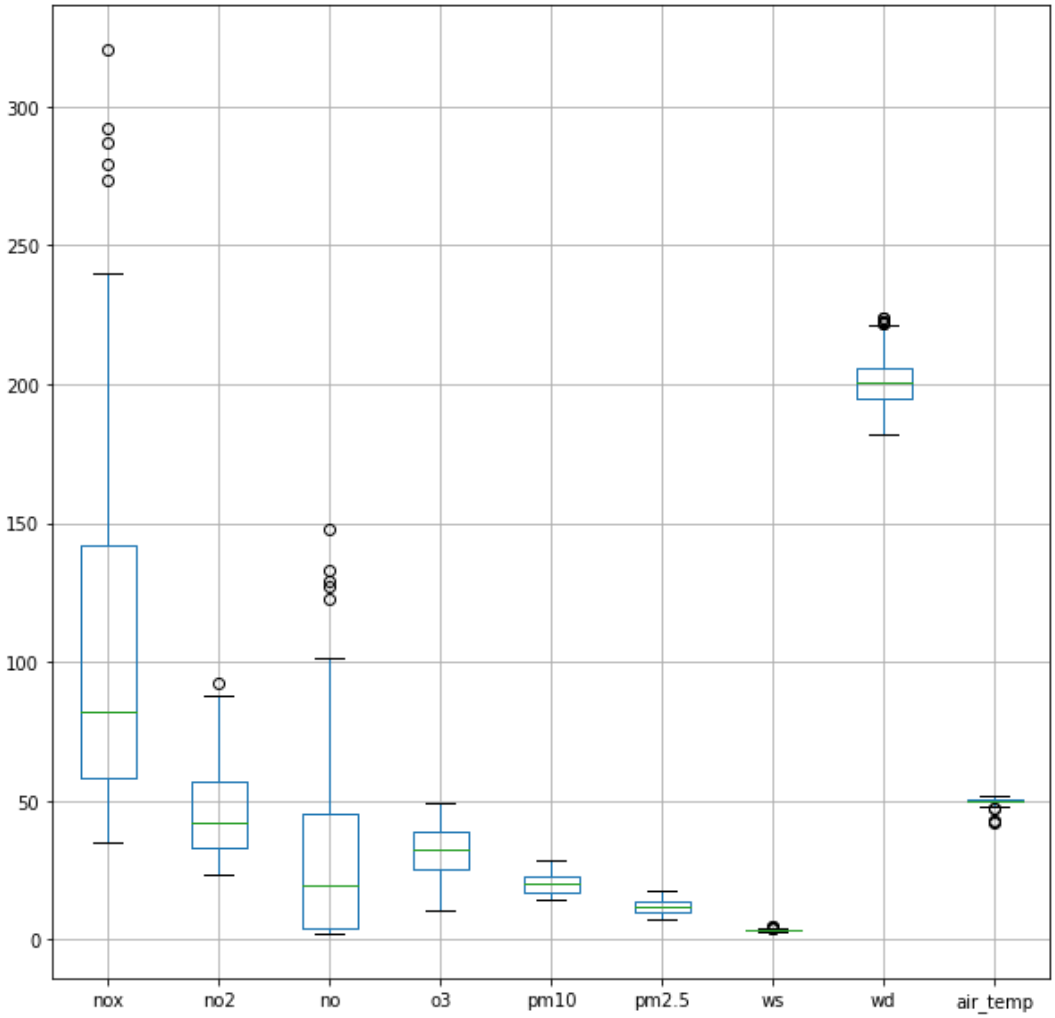


Figure 21: Boxplot for pollutant variables

4.5 Correlation Analysis

The initial aim of the project is to analyse the relationship between air pollutant levels and asthma admissions. Hence the air pollutant variables along with the borough variable are considered the independent variables while the asthma admissions variable is the dependent variable.

4.5.1 Pair-plot

The pair-plots can be used to visualise the distribution of single variables as well as for bivariate analysis to explore the relationship between two variables.

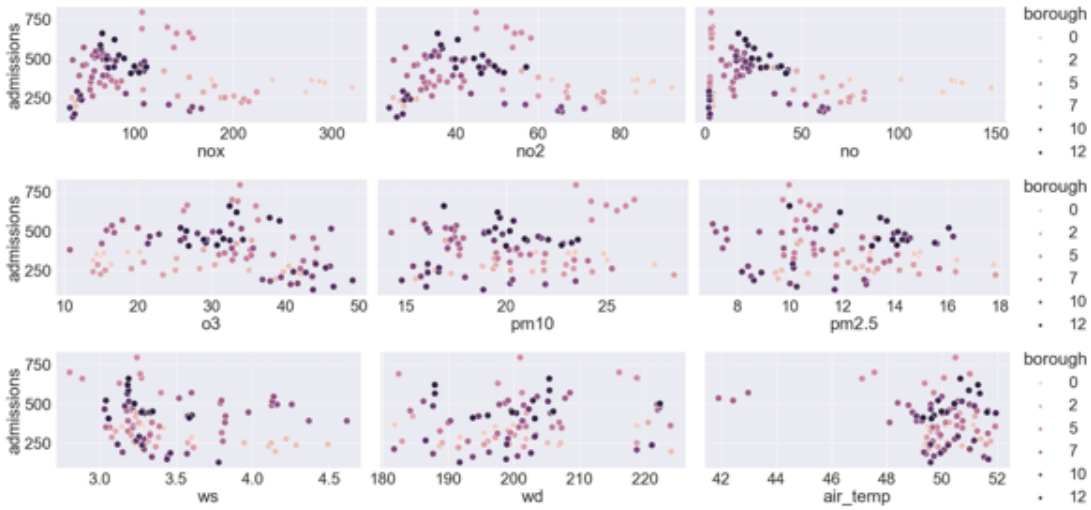


Figure 22: Pair plot for different variables with the admission variable

In Figure 22, the air pollutant variables are plotted against the target variable asthma admissions to analyse their pattern and trends. It can be assumed that the relationship is non-linear, and the nitrogen oxide features ("nox", "no2", "no") have a similar distribution.

4.5.2 Pearson's Correlation Coefficient (PPMCC)

Pearson's correlation coefficient[20] is a widely used method to determine the correlation between two variables. It is also known as Pearson Product-Moment Correlation Coefficient and the value ranges between -1 and +1. A value of 0 denotes no correlation whereas the values +1 and -1 denote strong positive correlation and strong negative correlation between the variables, respectively. However, it measures only the linear correlation and is not suitable for measuring non-linear relationships.

The Pearson Correlation Coefficient between variables X and Y is calculated as

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (4)$$

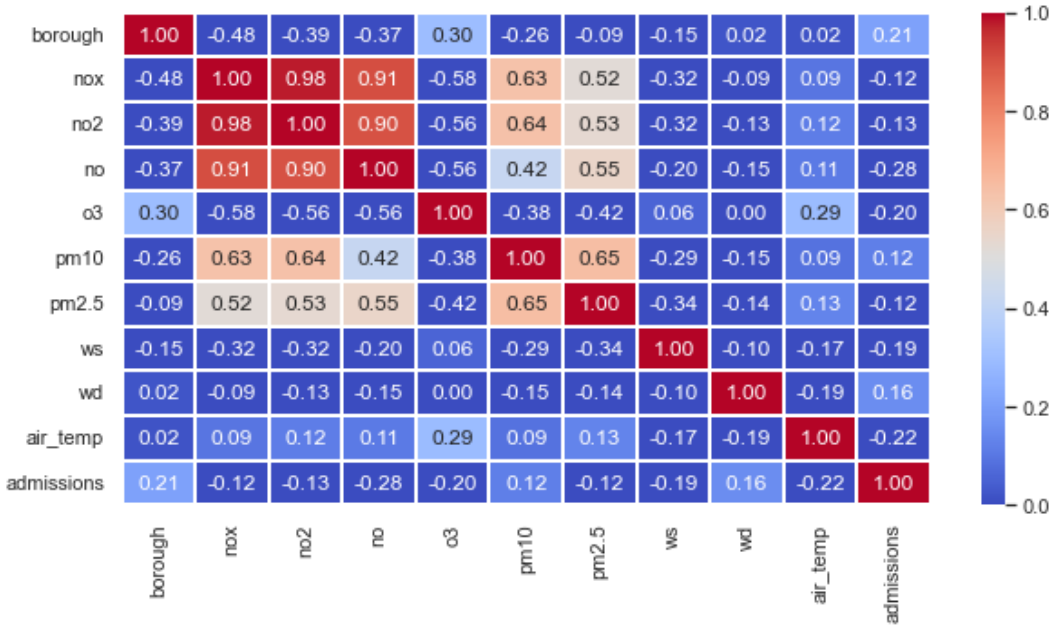


Figure 23: Pearson's correlation matrix

The above correlation plot shows a high correlation between nox, no2 and no which is obvious as the nox value is the sum of all nitrogen oxide emissions. It can also be seen that the pm10 and pm2.5 features are highly correlated with each other and with the nox and no2 features whereas the admissions feature has a low correlation with all other features.

4.5.3 Spearman's Correlation Coefficient

Spearman's correlation coefficient[21] is used to determine the strength and direction of the monotonic relationship between two variables based on rank orders. It is denoted by ρ or r_s. As in the case of Pearson's correlation, the value is bound between +1 and -1 where +1 indicates strong positive and -1 denotes strong negative relationships. When the ranks are distinct, ρ is calculated using the formula

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (5)$$

where d_i is the difference between the ranks of x and y of the same observation and n is the number of observations.

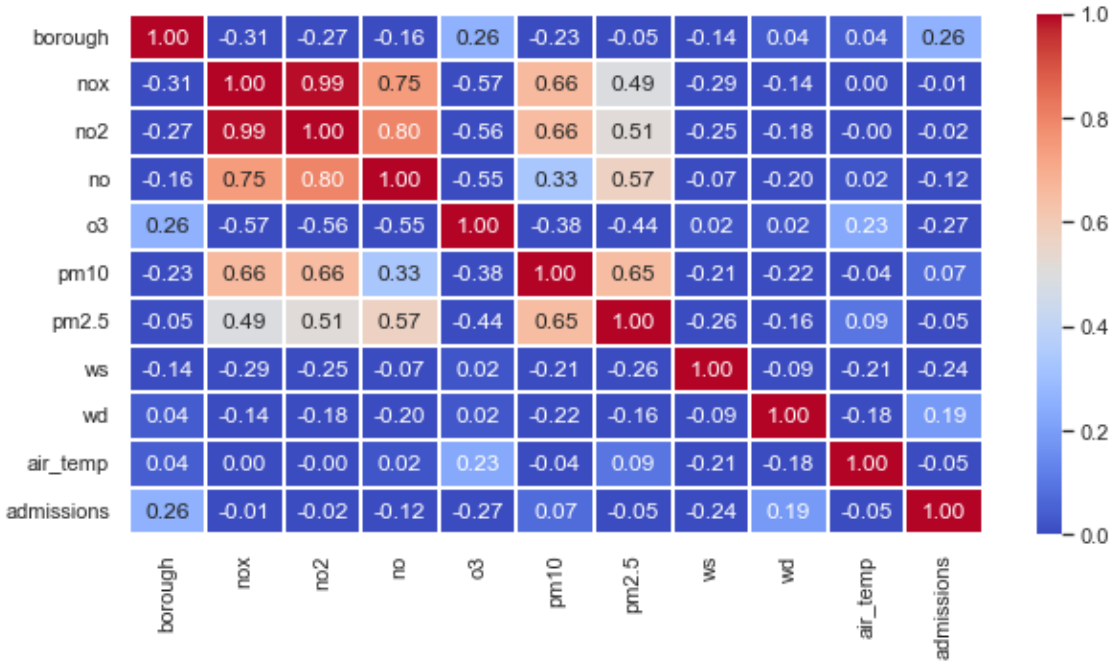


Figure 24: Spearman's correlation matrix

Spearman's correlation coefficient between admissions and other variables also shows a weak correlation.

4.5.4 Kendall's Correlation Coefficient

Kendall's correlation coefficient[22] is also a non-parametric measure for the relationship between two variables. The values of -1, +1 and 0 signify strong negative, strong positive and no relationships between the variables, respectively. It is denoted by τ and computed as

$$\tau = \frac{C - D}{C + D} \tag{6}$$

where C and D are the numbers of Concordant pairs and Discordant pairs, respectively.

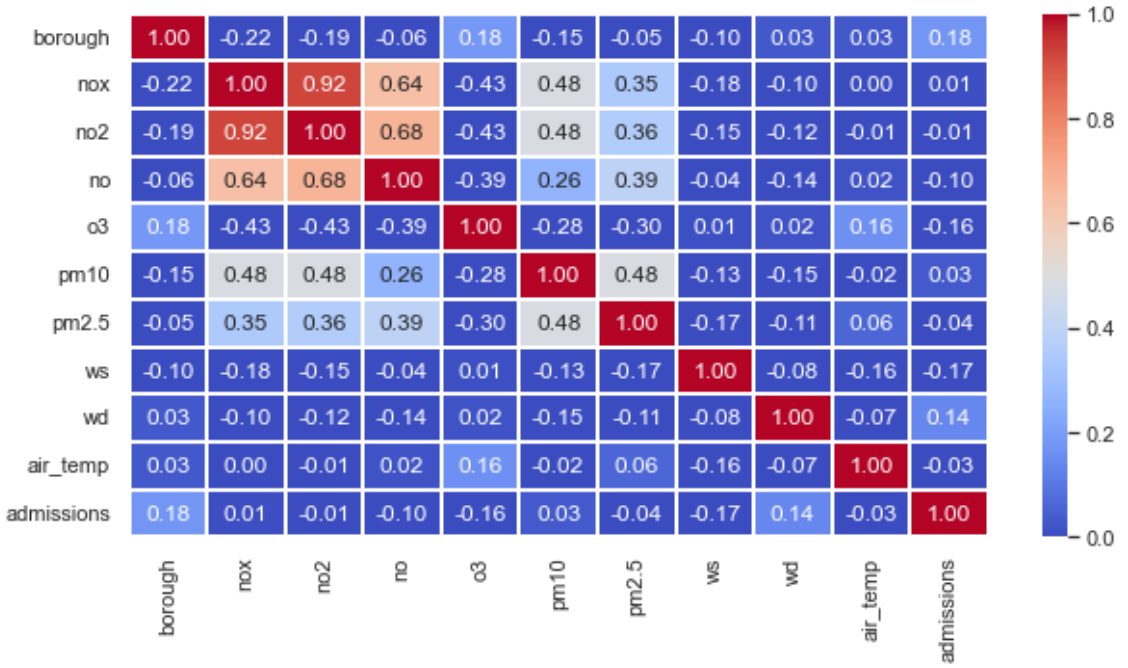


Figure 25: Kendall's correlation matrix

In this method also, no strong relationships are observed between asthma admissions and air pollutant levels.

4.5.5 Distance Correlation

Distance correlation can be used to detect both linear and non-linear correlation between random variables and is calculated as the ratio of their distance covariance and the product of their distance standard deviations. Unlike Pearson's coefficient, this value is always greater than zero or can be equal to zero only when X and Y are independent.

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X).dVar(Y)}} \quad (7)$$

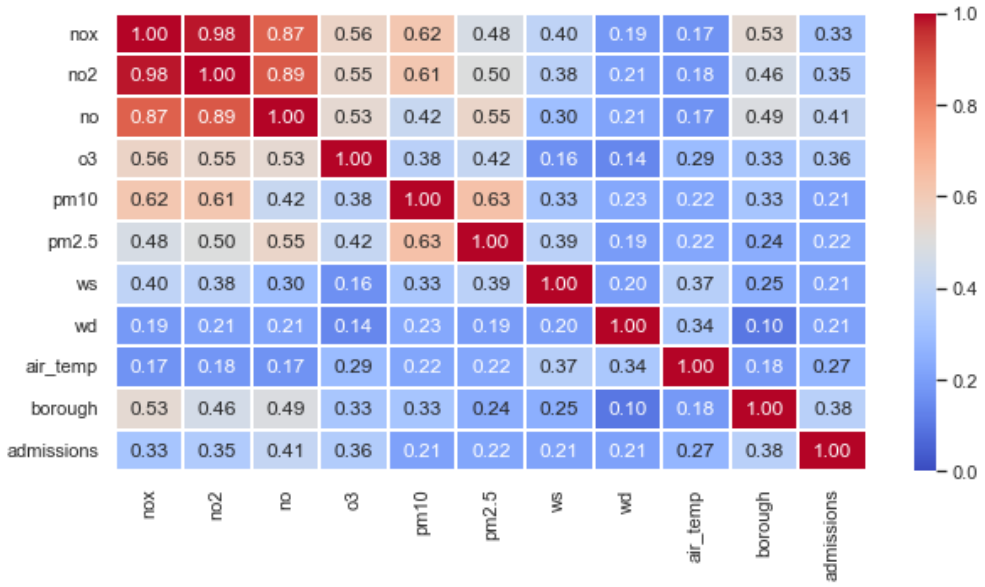


Figure 26: Distance correlation matrix for non-linear correlation

The Distance Correlation method showed improved correlation scores for the admissions variable with other variables.

4.5.6 Mutual Information Correlation

Mutual Information of two random variables can also be used to determine whether there exists a correlation between them that are non-linear in nature. It determines the quantity of information that can be obtained from another variable using the concept of entropy. The MI values are calculated using a python package namely ennemi[23], which was initially developed at the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki. This package was tested and validated using large atmospheric datasets and hence can be considered the most suitable for this study.

$$I_{(X,Y)} = \iint_{y\,x} p(x,y).\log\left(\frac{p(x,y)}{p(x)p(y)}\right) dx\,dy \tag{8}$$

It can range from to ∞ but can be bounded between 0 and 1 by normalisation as given by

$$\rho_{I_{(X,Y)}} = \sqrt{1 - \exp\left(-2I(X,Y)\right)} \tag{9}$$

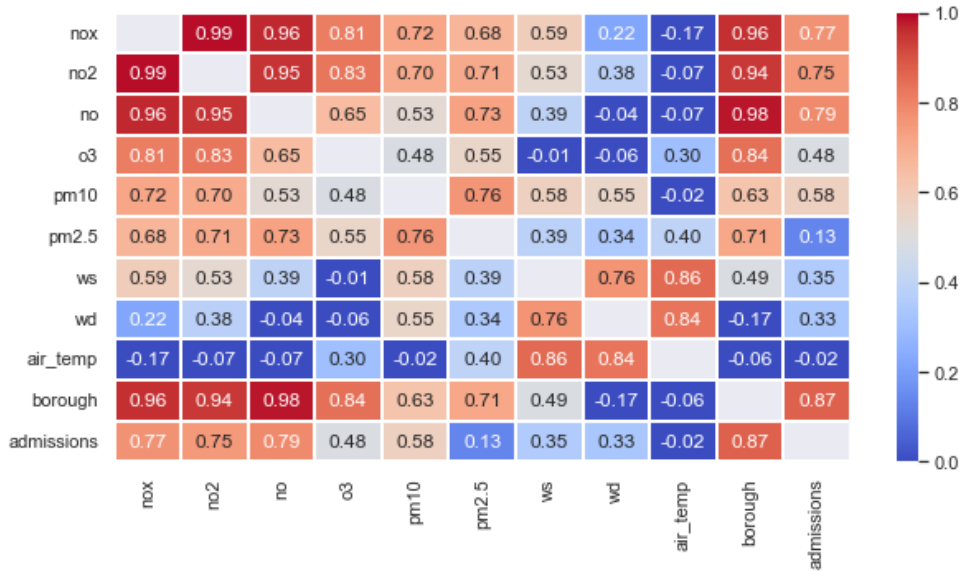


Figure 27: Correlation matrix based on pairwise mutual information

Using this method, the variables like “nox”, “no2”, and “no” show a strong correlation whereas “o3” and “pm10” show a moderate correlation with asthma admissions.

4.5.7 Data Transformation

The data is normalised using sklearn.preprocessing.StandardScaler which transforms the data to a common scale with mean zero and unit variance. It is performed to ensure that all the variables are given the same weights and treated equally by the model and to increase the model performance.

4.5.8 Feature Selection

The relative importance of different independent features in determining the dependent variable i.e., asthma admissions are analysed using various methods like RandomForestRegressor, F-statistic and p-values as well as Mutual Information.

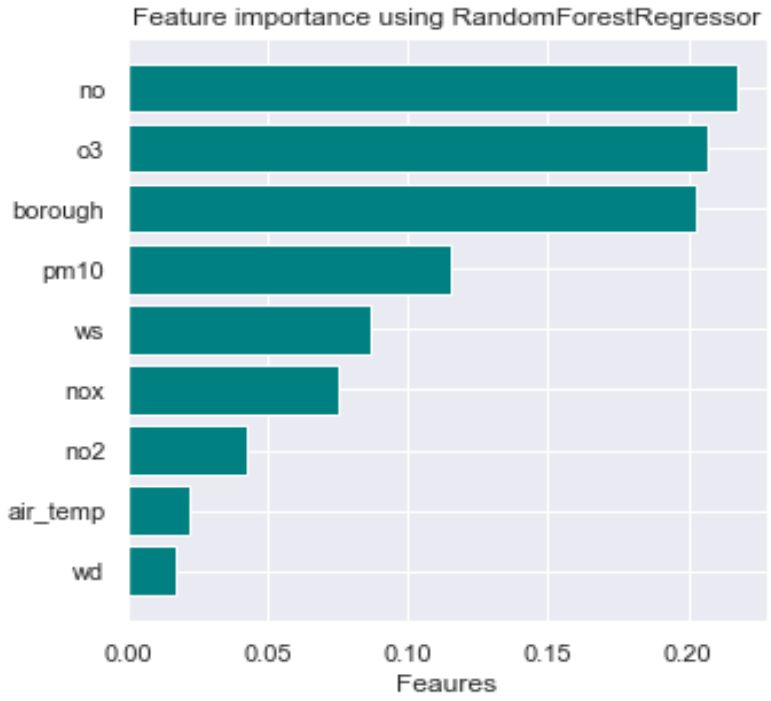


Figure 28: Feature importance using RandomForestRegressor

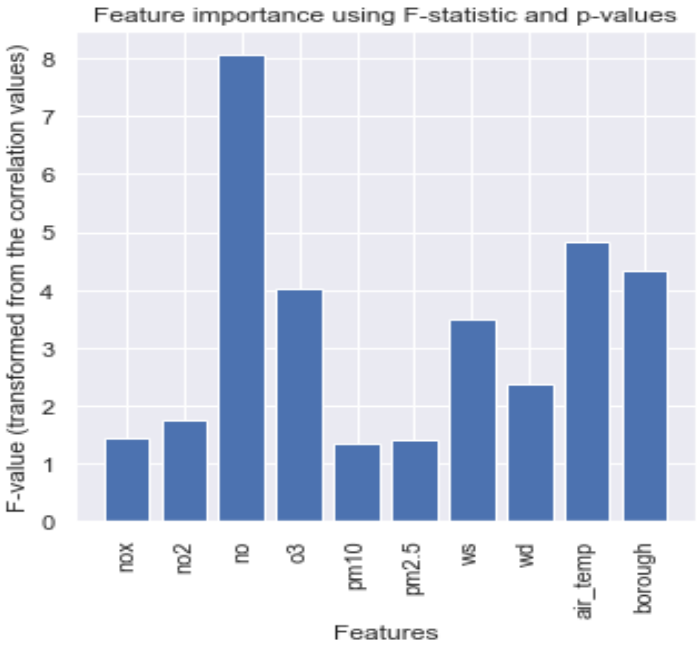


Figure 29: Feature importance using score F-statistic and p-values

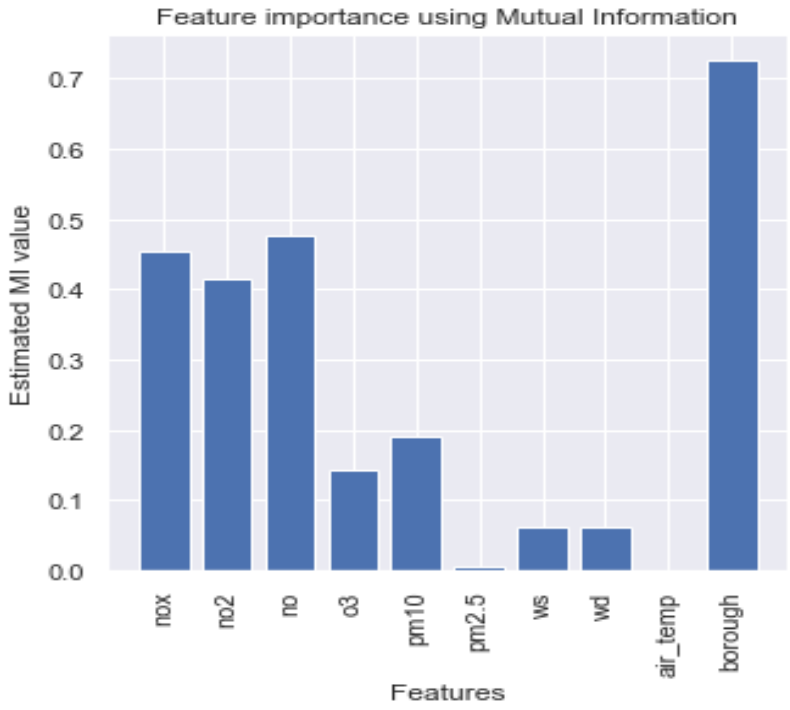


Figure 30: Feature importance using score Mutual Information

From Figure 28, Figure 29, and Figure 30 it is observed that the importance of the features estimated using different algorithms shows varied results.

4.6 Exploratory Data Analysis for unsupervised clustering

For unsupervised clustering, the hourly resampled data file is used and does not involve the asthma admissions feature as the data on an hourly basis were not available for the same. It is important to note that the data used for clustering purposes is different from the one used in correlation analysis. For correlation analysis the data for the target variable i.e., annual asthma admissions were available and hence yearly resampling was performed on the air pollutant data also.

4.6.1 Feature Encoding and Standardisation

Before starting the analysis, the categorical feature borough is converted into a numerical feature by converting each feature category into separate dummy or indicator variables. The number of dummy variables will be equal to the number of distinct categories in the variable. Using the sklearn.preprocessing.StandardScaler, the remaining features were

standardised to zero mean and unit variance. If u and s are the mean and standard deviation of the samples respectively, then the standard score of sample x is given as

$$z = \frac{(x - u)}{s} \tag{10}$$

After one hot encoding, an extra 14 columns will be added to the feature list eliminating the need for the original "borough" variable. However, it is not at all a concern for deep neural networks that are efficient in handling high-dimensional data.

4.6.2 Handling Skewness

The numerical features in the data with positive (right) skewness were converted to normal (Gaussian) distributions using sklearn.preprocessing.QuantileTransformer.

A comparison between Quantile transformation and Power transformation was done to determine the most appropriate one for the dataset. Quantile transformation applied to individual features tends to spread out the most frequent values and is robust to outliers since the transformation is based on quantile information.

Table 5: Comparison of Quantile transformation and Power transformation

From the above figure, it is evident that Quantile transformation is better than Power transformation for the data used in this study.

4.6.3 Correlation analysis

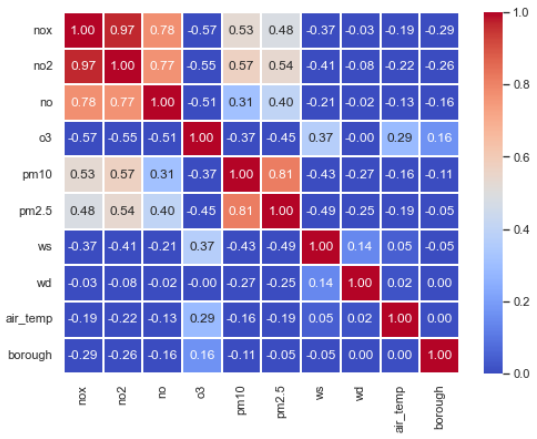


Figure 31: Pearson's correlation

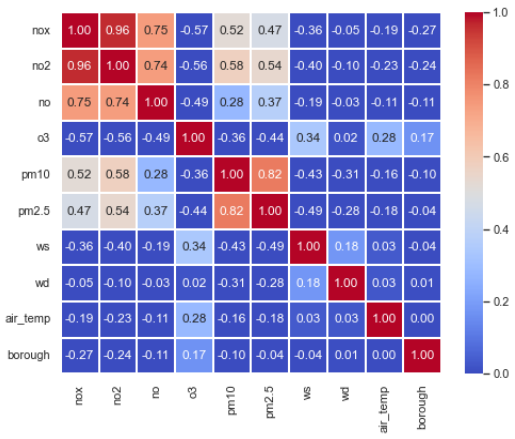


Figure 32:Spearman's correlation

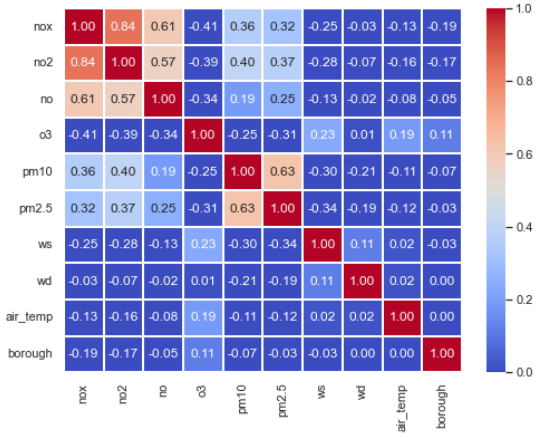


Figure 33: Kendall's correlation

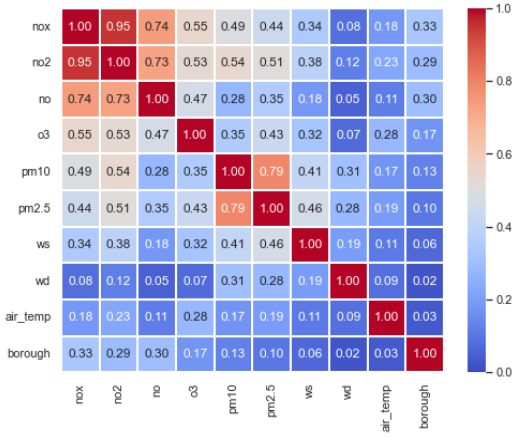


Figure 34: Distance correlation

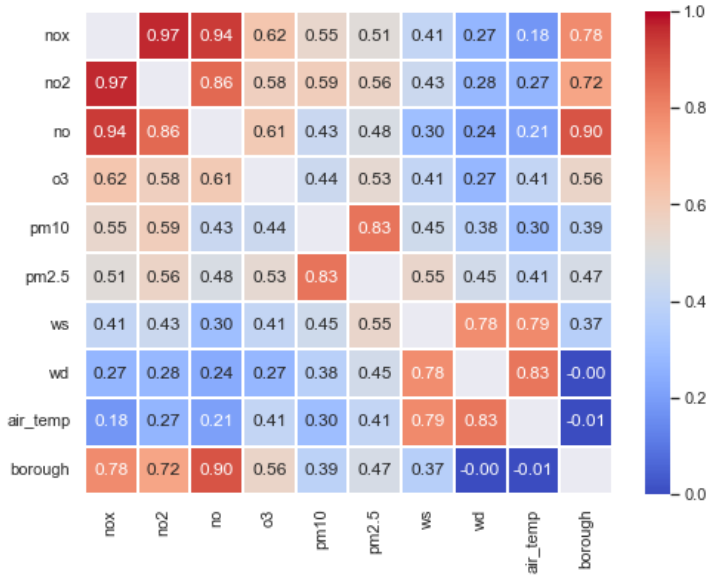


Figure 35: Mutual Information correlation matrix

Pearson's correlation, Spearman's correlation, Kendall's correlation and Distance correlation methods show the same information although the value slightly changes from one method to another. An interesting observation noticed was that the correlation between wind speed, wind direction and air temperature is evident in the mutual information correlation which could not be identified by the other methods.

4.6.4 Principal Component Analysis (PCA)

Principal Component Analysis is a technique of dimensionality reduction and is used when the features have a significant correlation between them. It can also be useful to represent the information in a high-dimensional data table using fewer dimensions called principal components which can be easily analysed, interpreted and visualised.

Using PCA, the 23 features are reduced to two principal components PC1 and PC2 which explained only 20.45% and 7.67% of the total variance in the data, respectively. In other words, PC1 and PC2 attributed to a total of 28.12% of the total variance in the data with PC1 containing most of the information. The component count needs to be increased to extract a higher percentage of the variance in data.

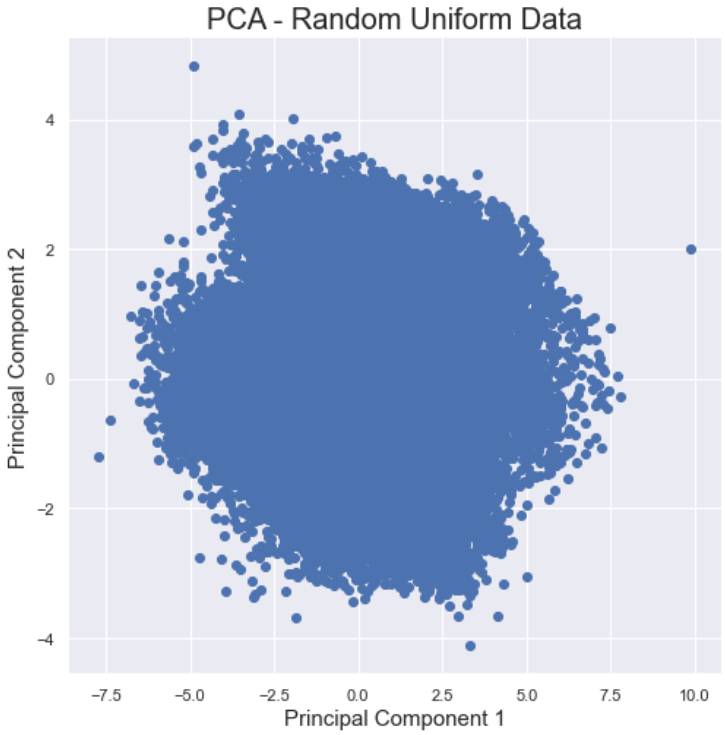


Figure 36: Principal components in 2-dimensional space

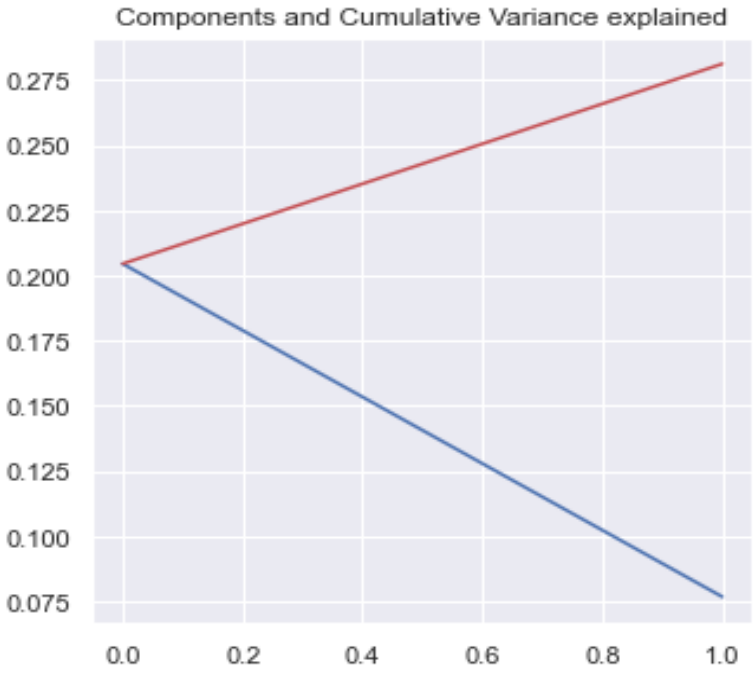


Figure 37: Principal components and their cumulative variance plot

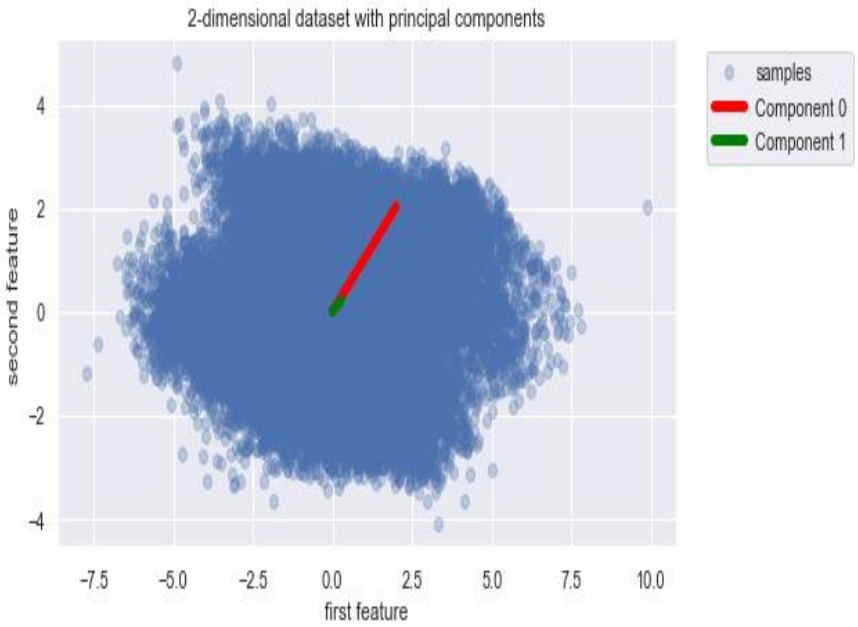


Figure 38: Principal component vectors(Eigenvectors)

4.6.5 Bartlett's Sphericity Test

Bartlett's test of sphericity[24] tests the null hypothesis that the correlation matrix is an identity matrix which implies that the variables do not correlate and are not suitable for factor analysis. If the p-value is less than an alpha value of 0.05, the null hypothesis can be rejected, and it can be concluded that the factor analysis can be performed on the data. It is implemented by the factor_analyzer.calculate_bartlett_sphericity function of the factor_analyzer package and the formula is given as

$$T = -log(det(R))(N - 1 - (2p + 5)/6) \tag{11}$$

The values of chi-square and p were shown as nan which might be a result of the encoded columns adding to the homogeneity of variances. To confirm this, when the encoded columns were removed from the data, the test gave a Chi-squared value of **230675.53** and a p-value of **0.0** for the data used in this study confirming the fact that the variables are correlated and ideal for factor analysis.

4.6.6 Kaiser-Meyer-Olkin(KMO) Test

KMO Test is another conducted to evaluate the partial correlation strength between the variables and how well each variable can explain other variables in the data. This test is implemented by factor_analyzer.factor_analyzer.calculate_kmo function of the factor_analyzer package. A value below 0.5 is unsatisfactory and it is considered that values above 0.8 are perfect for conducting factor analysis.

$$MO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u_{ij}} \tag{12}$$

where R=[r_{ij}] is the correlation matrix and U=[u_{ij}] is the partial covariance matrix.

The test gave an average KMO score of **0.4652** and the scores for individual variables were obtained as given below.

```
[0.58554758 0.73829435 0.46799778 0.71905968 0.71632547 0.68910745
 0.8077027  0.69112815 0.60408705 0.36541448 0.39192811 0.21489363
 0.35434459 0.0828073  0.08194629 0.30680664 0.28391848 0.23268326
 0.1836557  0.3016298  0.2522505  0.11560746 0.15694365]
```

All the encoded variables have a KMO value below 0.5 which reduced the final KMO score to drop below 0.5.

4.6.7 Factor Analysis

Factor analysis is also a method for dimensionality reduction which reduces multiple variables into fewer number of underlying factors without losing any information in the data. The first step in factor analysis is to find the ideal number of factors which can be performed using multiple methods.

Scree plot

The Scree plot is a graph in which the eigenvalues are plotted on the y-axis with a corresponding number of factors plotted on the x-axis.

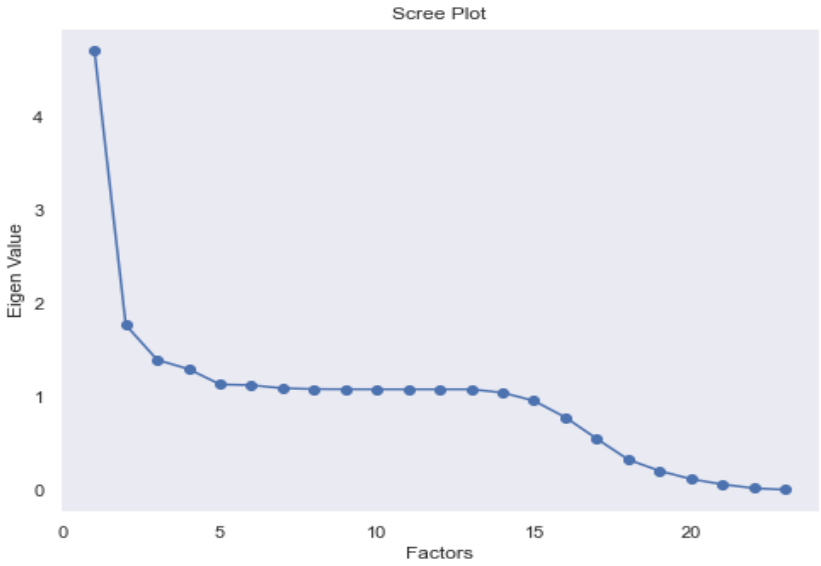


Figure 39: Scree plot

According to the general cut-off value of eigenvalue ≥1, the corresponding factor count is 5. The presence of encoded features has flattened the curve at y=1 making it difficult to decide the factor count.

Total Percent Variance Explained

The Total Percent Variance Explained is the total amount of variance of original variables attributed by the factor components in percentage. It is increased by adding more variables, but it is better to select the minimum number of factors explaining most of the variance. From the scree plot, it can be assumed that the first five factors contribute to the major portion of the total variability in the original variables.

Factor Loadings

Factor Loadings is a matrix which shows the correlation or relationship between the original variables in the data and the underlying factors that explains them.

	Factor1	Factor2	Factor3	Factor4	Factor5
nox	0.912822	0.297928	0.160439	-0.132982	-0.244925
no2	0.884181	0.368417	0.154279	-0.112706	-0.166773
no	0.952230	0.076698	-0.076589	0.005920	0.410459
o3	-0.506235	-0.130689	0.190728	0.833398	0.016358
pm10	0.268417	0.781976	0.120529	-0.202332	-0.060741
pm2.5	0.242556	0.887451	0.050061	-0.278462	0.299370
ws	-0.194475	-0.453557	-0.088857	0.201197	0.025086
wd	0.015789	-0.301114	0.026998	-0.036702	-0.031014
air_temp	-0.133313	-0.115806	0.099615	0.188931	0.053577
0	0.378906	0.027993	0.037948	-0.061200	0.042288
1	-0.240433	-0.086641	0.040026	-0.014999	0.072785
2	0.220140	-0.061146	0.028219	0.069151	0.012739
3	0.227443	0.040770	0.046802	-0.120681	0.045003
4	-0.060440	0.148491	0.099423	0.010623	-0.511755
5	-0.169560	0.015085	0.090491	-0.088704	-0.150432
6	-0.166234	0.112325	0.018457	0.098084	0.185705
7	-0.074933	0.007568	-0.992228	-0.113132	-0.097076
8	-0.003764	-0.218756	0.026813	-0.012429	0.003789
9	0.203559	0.001721	-0.006665	0.309243	-0.015377
10	-0.001264	0.037285	0.045712	-0.070402	0.067210
11	-0.296740	-0.042981	0.063056	0.028618	-0.040710
12	-0.050287	0.072783	0.026702	-0.002386	0.168537
13	0.034031	0.002950	0.032202	-0.010271	0.076065

Figure 40: Factor loadings

The observation from the above factor loadings can be summarised as

- Factor1 has high factor loadings for nox, no2, and no
- Factor2 has high factor loadings for pm10 and pm2.5
- Factor4 has high factor loadings for o3

	Factor1	Factor2	Factor3	Factor4	Factor5
Variance	3.465176	2.061349	1.144905	1.076364	0.732641
Proportional Var	0.150660	0.089624	0.049778	0.046798	0.031854
Cumulative Var	0.150660	0.240284	0.290062	0.336861	0.368715

Figure 41: Factor variances

The above image shows the variance, proportional variance and cumulative variance of each factor and it can be interpreted that around 36.9% of the cumulative variance is explained by the five factors.

4.6.8 Hopkins Statistic for Clustering Tendency

Hopkin's Statistic[25] is a measure of the clustering tendency of the data. A value close to one indicates highly clustered data and 0.5 denotes random distribution of data. The uniform distribution of data is indicated by a value close to zero.

The statistic value for the scaled dataset was obtained as **0.93** which confirms the presence of clusters in the data.

4.6.9 Determining the number of clusters

The number of clusters present in the data can be estimated using several methods based on the cluster evaluation metric scores.

Elbow Method - Distortion Score, Silhouette score and Calinski Harabasz score

The Elbow method estimation of optimal cluster count can be performed using the yellowbrick.cluster.KElbowVisualizer API of Yellowbrick package. The calculation of the k value is based on the distortion score by default but other metrics like the Silhouette score and Calinski Harabasz score can also be used with this method. It also automatically finds the inflexion point of the curve to estimate the optimal value of k and shows the training time for the clustering model for different cluster values(k). The distortion score is defined as the total sum of the squared distances of all the data points in a cluster to their cluster centre. The below plot shows the visualisation using KElbowVisualiser where the scoring parameter is "distortion" by default.

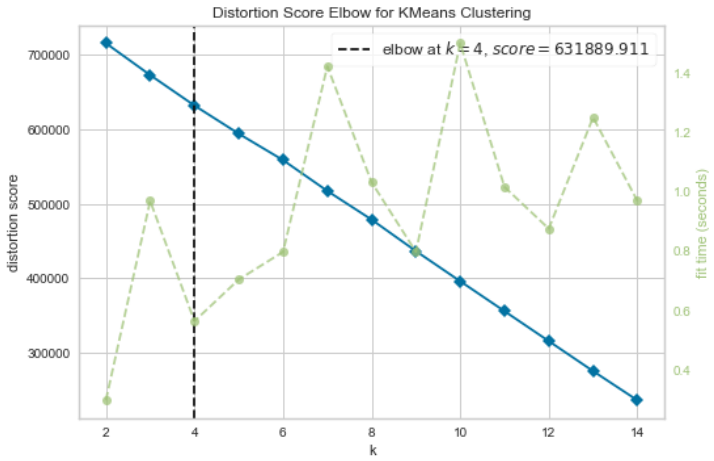


Figure 42: Distortion Score Elbow plot

If the scoring parameter is set to “silhouette”, then the elbow method calculation will be based on the mean value of the Silhouette scores of all the samples as shown in the figure below.

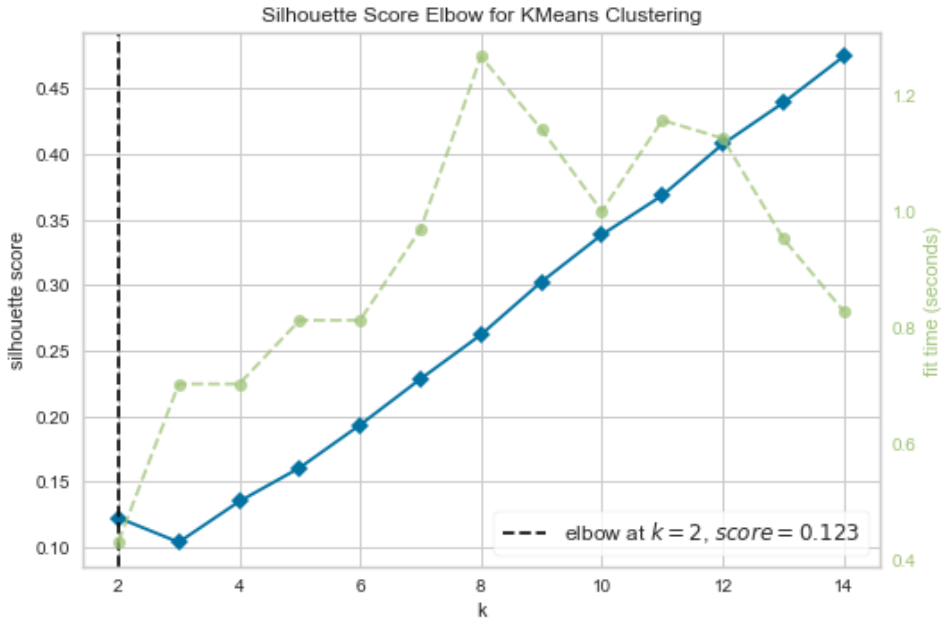


Figure 43: Silhouette Score Elbow plot

Similarly, the elbow method calculation will be based on the ratio of dispersion within and between clusters if the scoring parameter is set as “calinski_harabasz”. The visualisation based on the Calinski Harabasz score is shown in the figure below.

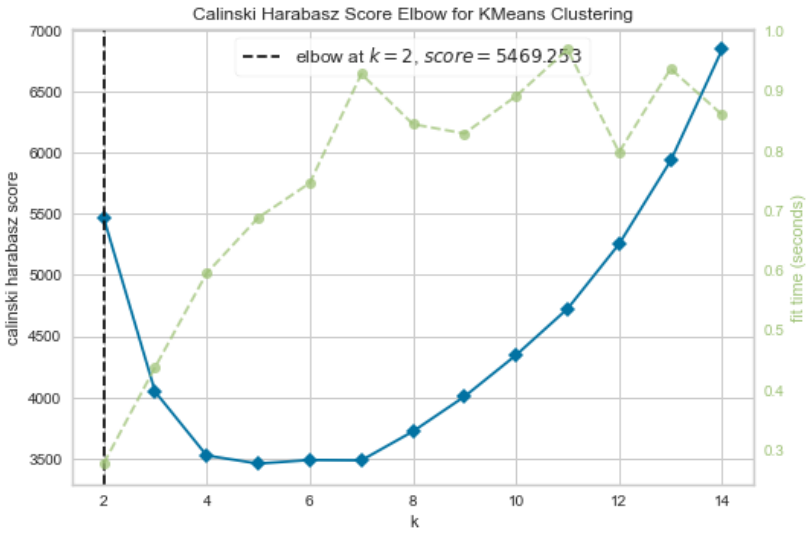
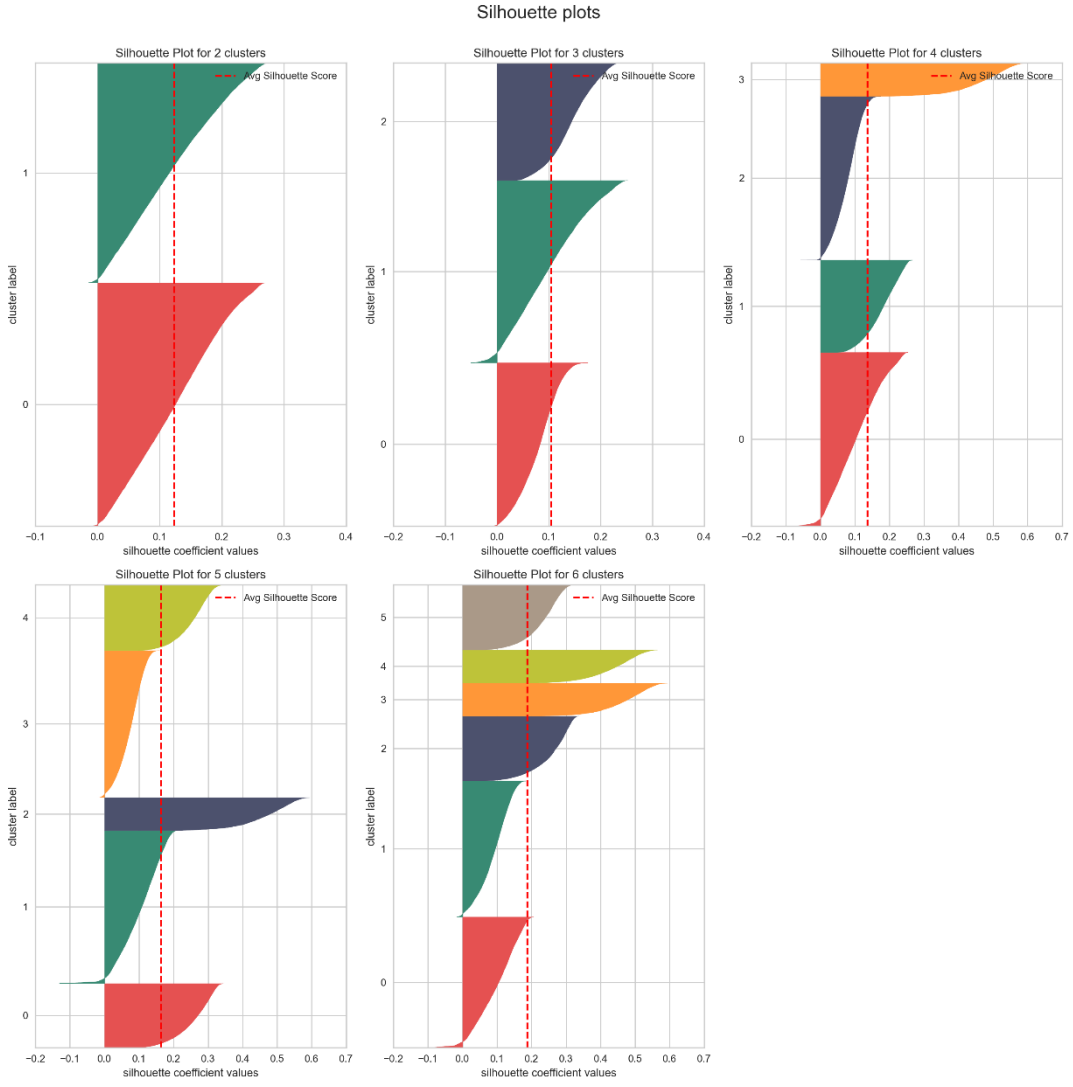


Figure 44: Calinski Harabasz Elbow plot

The optimal values of k estimated using the Distortion Score, Silhouette Score and Calinski Harabasz Score were 4, 2, and 2, respectively. Other methods were also considered as a common k value could not be suggested by the above three methods.

Silhouette Analysis

Silhouette analysis is used to determine the separation of the clusters and the size of the clusters. Based on this the appropriate value of k can be determined. The below figure shows the Silhouette plot for cluster values from 2 to 6.



All the above plots show below-average scores, the highest being close to 0.2 for 6 clusters. Based on the size of individual clusters and the Silhouette score, 2 can be considered the most appropriate one.

Elbow Method using K-Means Inertia

Inertia is defined as the distance of each data point from its centroid and indicates the accuracy of the K-Means algorithm. The below figure is a plot of the sum of squared distances of the sample points from the centre of its nearest cluster against different k values.

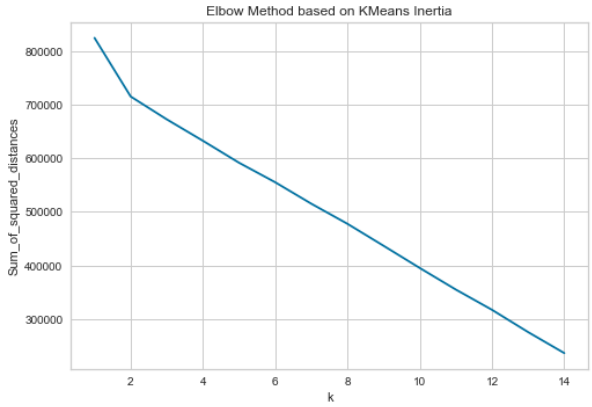


Figure 45: Elbow method based on K-Means inertia

The above graph points to two as the optimal k value since the decrease in the sum of squared distances tend to be slowed down at this point.

Davies Bouldin Plot

The graph plotting Davies Bouldin's score against different ka values is shown below. The lower value of this score, the more distinct and compact the clusters will be.

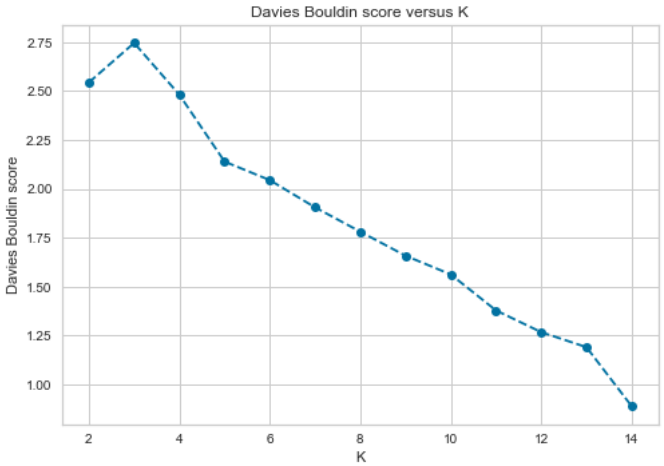


Figure 46: Davies Bouldin plot

From the above graph, the optimal k value cannot be interpreted but can be used along with other methods to choose the right value of k.

Gap Statistic

Gap Statistic[26] was proposed by Tibshirani et al. to estimate the optimum value of k based on the intra-cluster dispersion and is calculated using the formula

$$Gap_n(k) := E_n^* \log(W_k^*) - \log(W_k) \tag{13}$$

where W_k is the sum of the squared distances from the cluster means.

$$W_k := \sum_{r=1}^{k} \frac{1}{2n_r} D_r \tag{14}$$

where D_r is the sum of the pairwise distances for all points in cluster r.

However, as presented in a paper by Mojgan et. al.[27], the alternative equation without the logarithm function is superior due to better performance. Hence the following equation is used for calculating gap statistics.

$$Gap_n(k) := E_n^*(W_k^*) - W_k \tag{15}$$

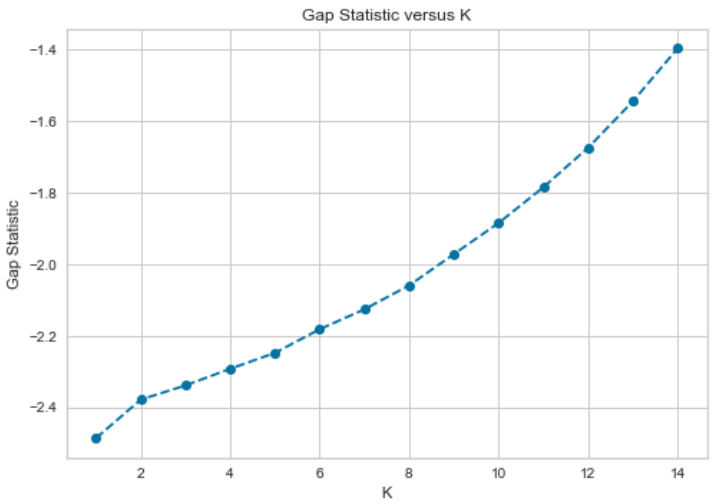


Figure 47: Gap Statistic plot with the log function

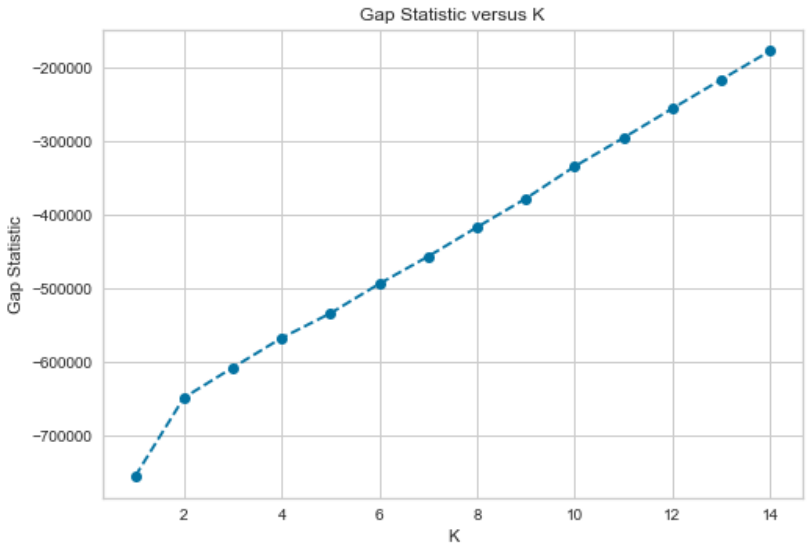


Figure 48: Gap Statistic plot without log function

The optimal value of k corresponds to the point at which the increase in gap statistic tends to slow down and from the above plot k can be assumed as two.

4.7 Clustering

As the initial step in clustering, the traditional K-Means clustering is applied using the "k-means++" initialiser to confirm whether the clustering has improved after DEC. Unsupervised Deep Clustering is the task of grouping the unlabelled data points into various clusters based on the similarities and dissimilarities between them using deep learning techniques.

4.7.1 Deep Embedded Clustering (DEC)

Deep Clustering is the task of grouping the data points into various clusters based on the similarities and dissimilarities between them using deep learning techniques. In this study, the deep clustering of the data is performed using the Deep Embedded Clustering (DEC) method proposed by Junyuan Xie, Ross Girshicka and Ali Farhadi [28] and is implemented using the Keras, an open-source software library in Python that provides numerous APIs for artificial neural networks (ANN). This method uses deep neural networks (DNN) to learn the feature representations and cluster assignments at the same time.

The figure below shows the architecture of the DEC network presented in the original paper.

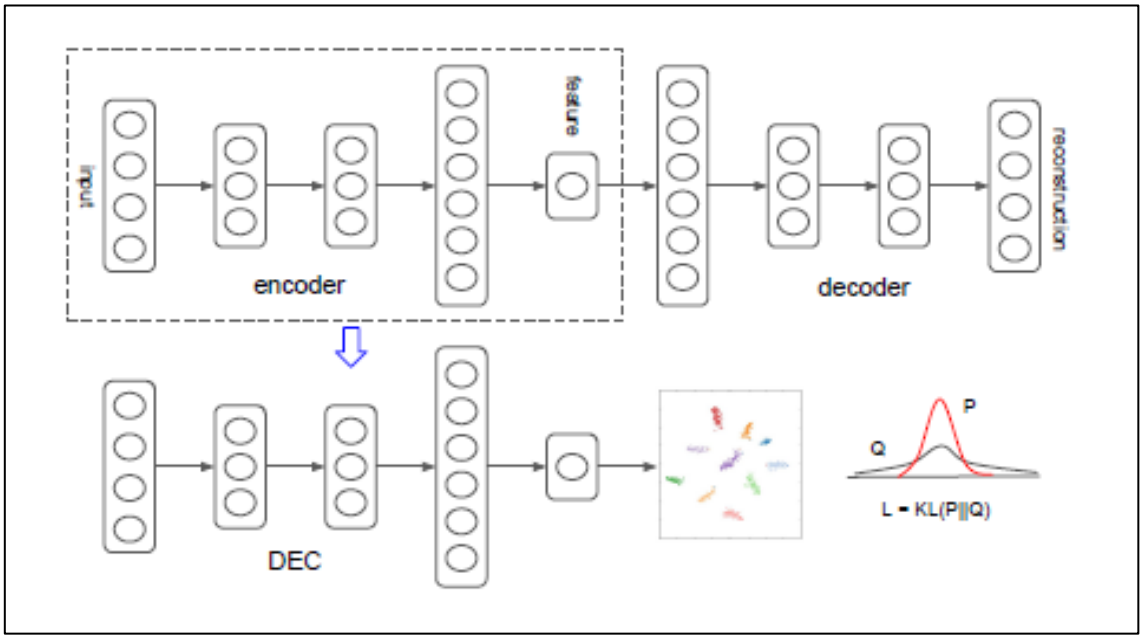
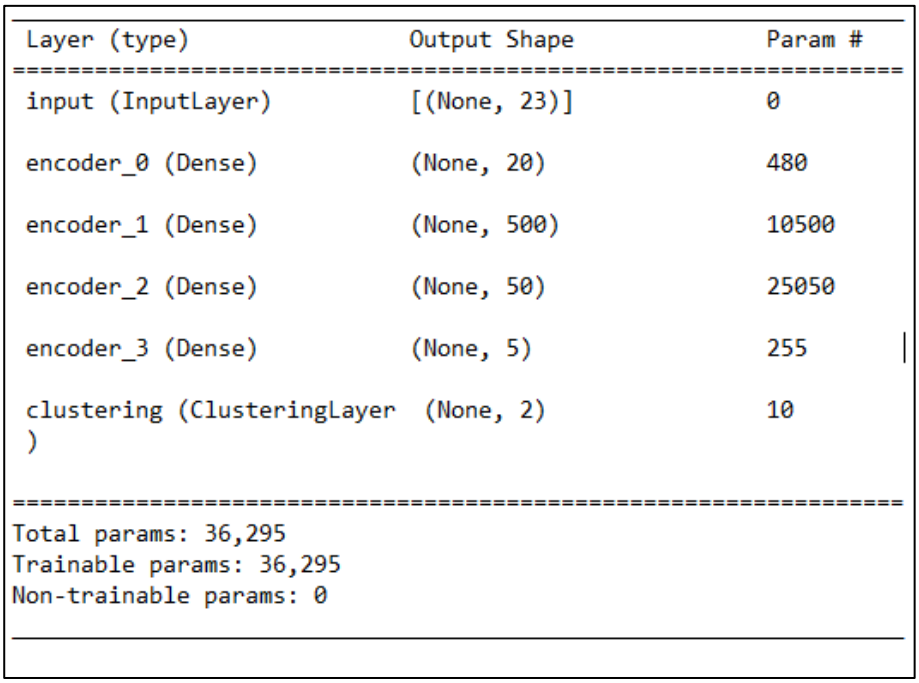


Figure 49: DEC network proposed by X. Junyuan et. al.

The dimensions of the DEC network used in this study are different from the one proposed in the paper and were modified based on the input feature count. The implemented DEC architecture consists of an autoencoder with an input layer, seven hidden layers and an output layer. The input and output have 23 nodes each, which is equal to the number of features in the data. The three internal layers in the encoder and decoder have the dimensions(20, 500, and 50) and the bottleneck layer's dimension is 5. The structure of the DEC model implemented is given below.



Layer (type)	Output Shape	Param #
input (InputLayer)	[(None, 23)]	0
encoder_0 (Dense)	(None, 20)	480
encoder_1 (Dense)	(None, 500)	10500
encoder_2 (Dense)	(None, 50)	25050
encoder_3 (Dense)	(None, 5)	255
clustering (ClusteringLayer)	(None, 2)	10

Total params: 36,295
Trainable params: 36,295
Non-trainable params: 0

Figure 50: DEC Model summary

The DEC algorithm consists of two phases. The first phase involves parameter initialization using a deep autoencoder and the second phase performs parameter optimization in which clustering with KL divergence is performed.

4.7.2 Parameter Initialization

First, training is performed to minimize the least-square errors and when one layer is trained, its output (h) is fed as the input for the next layer to train. The activation function used is ReLU except for the first pair and the last pair, respectively.

Once the training process is completed, the encoder layers and the decoder layers are joined to create a deep autoencoder after which fine-tuning is performed to minimize the reconstruction loss. Then the decoder layers are removed, and only the encoder layers are used for the mapping between the actual data and feature spaces.

The data is then fed to the initialized DNN to initialise the cluster centres. The k initial centroids are obtained by performing traditional K-Means clustering in the low-dimensional feature space.

4.7.3 Parameter Optimization (Clustering)

The parameter initialisation phase provides an initial estimation of the non-linear mapping fθ between data space S and latent space Z and the cluster centroids denoted by $\{\mu j\}_{j=1}^{k}$.

In this phase, as a first step, the Student's t-distribution is used as a measure to identify the similarity between centroid μ_j and embedded point z_i. The soft assignment or the probability that a sample i is assigned to cluster j is

$$q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j'}(1 + \|z_i - \mu_{j'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{16}$$

where the degree of freedom (α) is set to 1.

In the second step, the model training is performed by matching the soft assignment to an auxiliary target distribution for refining the clusters. The KL divergence loss is computed as

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{17}$$

where qi and pi are the soft assignment and the auxiliary distribution, respectively.

Then, the target distribution, P is computed using the soft cluster frequency fj as

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \tag{18}$$

The optimisation of the cluster centres {μj} and DNN parameters θ is performed and the gradients of KL divergence loss with respect to zi and μj are calculated using the below equations.

$$\frac{\partial L}{\partial z_i} = \frac{\alpha + 1}{\alpha} \sum_j (1 + \frac{\|z_i - \mu_j\|^2}{\alpha})^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j) \tag{19}$$

$$\frac{\partial L}{\partial \mu_j} = -\frac{\alpha + 1}{\alpha} \sum_i (1 + \frac{\|z_i - \mu_j\|^2}{\alpha})^{-1} \times (p_{ij} - q_{ij})(z_i - \mu_j) \tag{20}$$

The DNN receives the gradients ∂L/∂zi as input and calculates the parameter gradient of the DNN, ∂L/∂θ using the input gradients in the standard backpropagation. The process is stopped when the percentage of data points which change their clusters between two successive iterations is less than the threshold convergence value.

4.8 Evaluation metrics used for unsupervised clustering

The evaluation of unsupervised deep clustering quality where the ground truth labels are unknown was evaluated using intrinsic measures like Silhouette Coefficient, Calinski-Harabasz index and Davies-Bouldin Index in which the cluster qualities are determined using the inter-cluster and intra-cluster distances. The above scores for baseline K-Means clustering were also calculated and compared with that of the DEC output to evaluate the performance of the DEC algorithm.

4.8.1 Silhouette Coefficient

Silhouette coefficient[29] is a widely used metric used to calculate the clustering efficiency based on the average distance between a sample and all other data points within the same cluster (cluster cohesion) and the average distance between a sample and all other data points in the nearest cluster (cluster separation). If the cluster separation and the cluster cohesion values of the sample are represented by a and b respectively, then the Silhouette coefficient s defined as

$$S = \frac{(b-a)}{\max(a,b)} \tag{21}$$

The sklearn.metrics.silhouette_score implements the computation of this metric and values of the Silhouette coefficient range between -1 and +1. A value close to +1 indicates the presence of well-separated, dense, and distinct clusters whereas a value of 0 implies that the sample is near the decision boundary of two or more neighbouring clusters or the possibility of overlapping clusters. A value of -1 denotes incorrect cluster assignment for the sample.

4.8.2 Calinski-Harabasz Index

Calinski-Harabasz index[30] also called the Variance Ratio criterion is the ratio of the sum of inter-cluster(within-cluster) dispersion and the sum of intra-cluster(between-cluster) dispersion for all clusters. If k is the number of clusters obtained by performing clustering on a set of data points, E with n_E samples, then the Calinski-Harabasz index is computed using the formula

$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{22}$$

where $tr(B_k)$ is a trace of the between-group dispersion matrix and $tr(W_k)$ is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)\,(x - c_q)^T \tag{23}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)\,(c_q - c_E)^T \tag{24}$$

where C_q and n_q denote the sample points and the number of samples in the cluster q, respectively. c_q is the cluster centre of cluster q and c_E is the cluster centre of the data sample set E.

The sklearn.metrics.calinski_harabasz_score implements this metric computation which does not have an upper bound value. However, a higher score indicates better clustering since the individual sample points in a cluster are packed closely whereas the clusters are far from each other.

4.8.3 Davies-Bouldin Index

Davies-Bouldin index[31] denotes the average similarity between the clusters and its most similar one based on the comparison of the distance between the clusters and their cluster sizes.

If C_i represent the clusters and C_j represents the most similar cluster, then the similarity measure, R can be calculated using the formula

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{25}$$

where s_i, and s_j are the average distance between each point of cluster and the centroid of that cluster (cluster diameters) of the clusters i and j respectively and d_{ij} is the distance between cluster centroids i and j.

The Davies-Bouldin index is computed using the above similarity measure as

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{26}$$

The sklearn.metrics.davies_bouldin_score implements the calculation of the metric and better cluster outputs are identified by lower index values. Zero being the lowest possible value indicate better cluster partition and quality.

4.9 DEC analysis

The clustering output of the DEC is saved to a file for analysis which can be effectively done using distinct types of visualisations[32].The samples were grouped into two clusters and the cardinality or the number of samples in each cluster has an enormous difference as shown in the figure below. Cluster 0 has less than 5000 samples whereas cluster 2 has more than 30,000 samples.

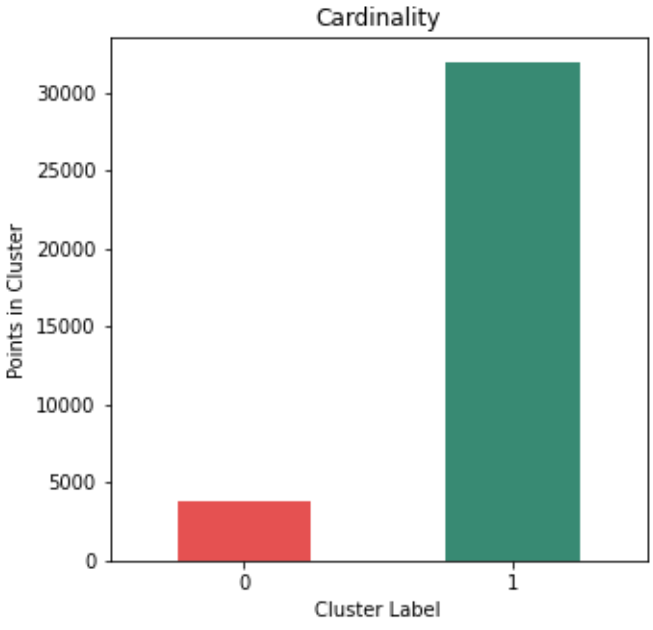


Figure 51: Cardinality of clusters

The figure below shows the feature distributions of each pollutant variable in the two clusters separately. Most of the variables in cluster1 have a substantial number of values beyond the upper limit (Q3 +1.5 * IQR) compared to that in cluster0. The values of no variable in cluster1 are extremely less when compared to that in cluster1 and the wd variable in cluster0 has many values below the lower limit (Q1 - 1.5 * IQR).

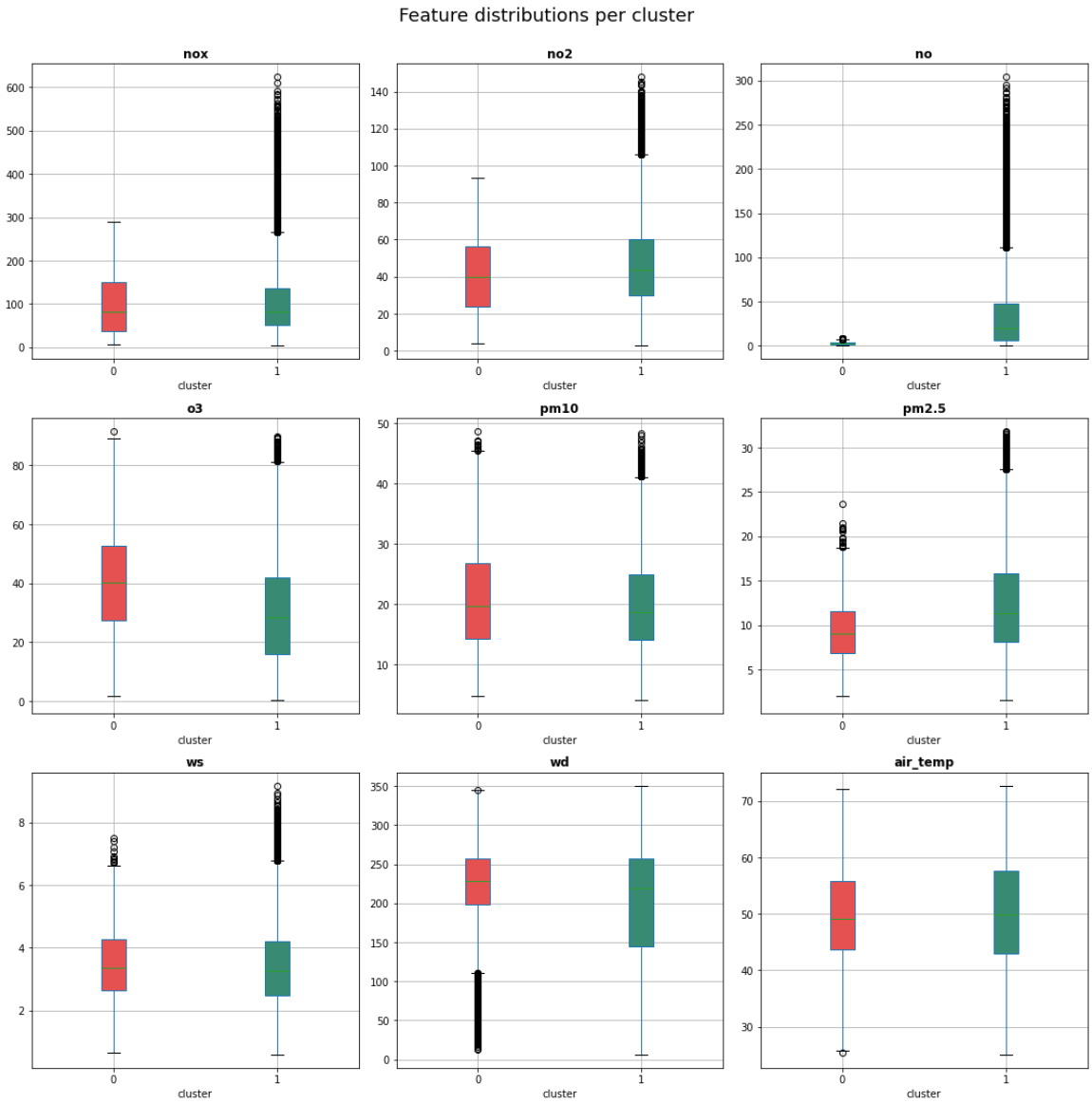


Figure 52: Feature distributions in clusters

The characteristics of both clusters can be consolidated using a radar chart using the mean values of the features for each cluster together in a single chart. The mean value of o3, wd and ws in cluster0 is higher than that of cluster1.

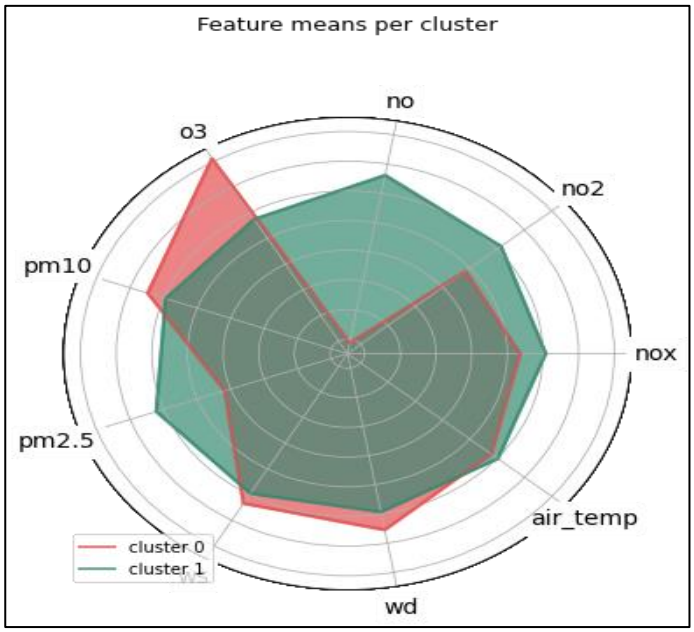


Figure 53: Plot of feature means per cluster

The following plot which displays the percentage deviation of the feature values in each cluster from the overall mean value can also be used to compare the features in the two clusters. The cluster0 feature values show high deviation when compared with the feature deviation in cluster1.

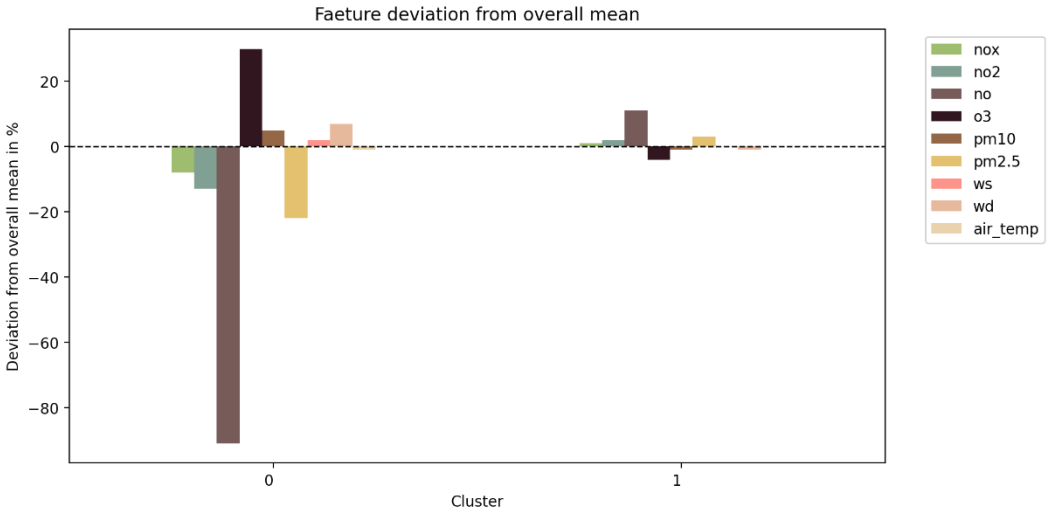


Figure 54: Feature deviation from the overall mean value

ANOVA Test

ANOVA test is the statistical test for estimating the correlation between a categorical variable and a numerical variable. In this study, this test is used to find the correlation of cluster labels with asthma admissions. The F-score and p-value were 0.002 and 0.963 indicating the absence of correlation among variables. Ideally, F-score should be as high as possible, and the p-value should be less than 0.05 for the correlation to be significant.

5. PROJECT PLAN

The following Gantt chart outlines all the tasks involved in this study in their respective order, starting date and finishing date of each task, and the submission deadlines planned against a timescale. The scheduled time for each task is denoted by a blue rectangle and a grey diamond is used to indicate the submission deadlines.

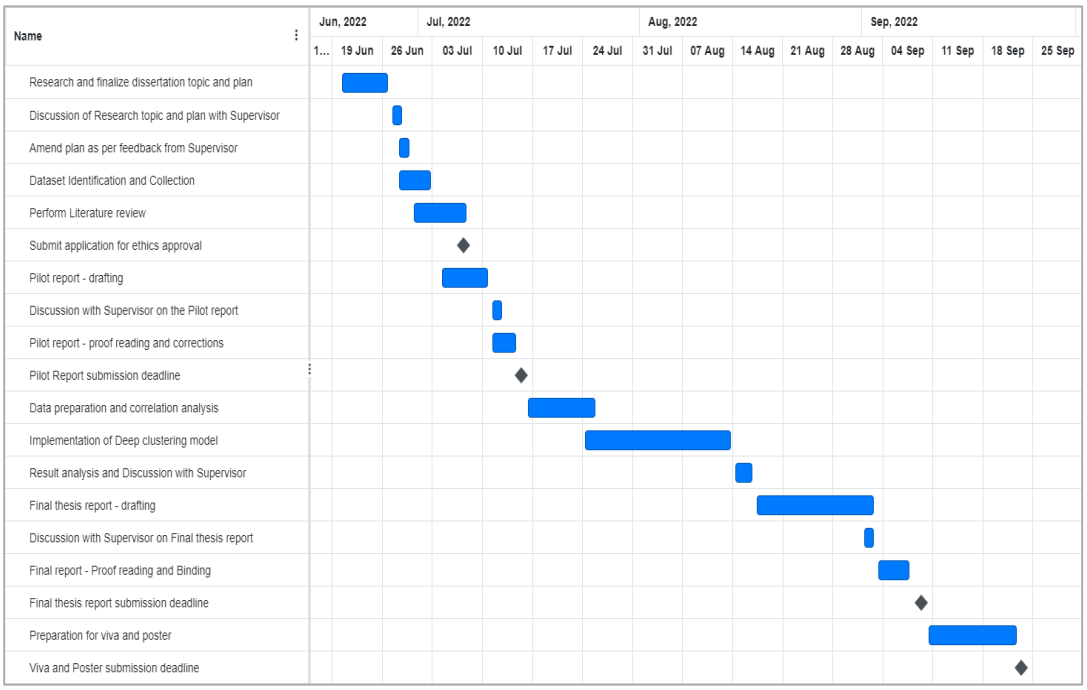


Figure 55: Gantt chart showing project timelines

The initial plan was to perform the study with air pollution datasets already available on the websites of different organisations. The deep clustering demanded more data for accurate results, hence the hourly raw data of air pollutants measured at multiple monitoring stations were downloaded using a package. Hence there has been a shift in these dates during the actual course of the project.

6. RESULTS

As mentioned before the aim of the study is to analyse the correlation between the air pollutant variables and asthma admissions and to analyse the clusters generated after performing DEC on the air pollution data. The main results are summarised in this chapter and the discussion of these results is presented in the following chapter.

The pair plots in Figure 22 which plotted individual air pollutants against asthma admissions show that the relationship between them could be considered **non-linear**. Hence different algorithms for linear and non-linear were chosen to calculate the correlation scores and a comparison was done.

Features	Pearson's correlation	Spearman's correlation	Kendall's correlation	Distance correlation	Mutual Info. Correlation
nox	-0.12	-0.01	0.01	0.33	0.77
no2	-0.13	-0.02	-0.01	0.35	0.75
no	-0.28	-0.12	-0.10	0.41	0.79
o3	-0.20	-0.27	-0.16	0.36	0.48
pm10	0.12	0.07	0.03	0.21	0.58
pm2.5	-0.12	-0.05	-0.04	0.22	0.13
ws	-0.19	-0.24	-0.17	0.21	0.35
wd	0.16	0.19	0.14	0.21	0.33
air_temp	-0.22	0.05	-0.03	0.27	-0.02
borough	0.21	0.26	0.18	0.38	0.87

Table 6: Correlation of air pollutant variables with asthma admissions

Among the multiple steps conducted for the correlation analysis, the two distinct methods were the calculation of correlation coefficients and the feature importance scores of independent variables with the target variable asthma. The above table summarises the correlation coefficient values between each of the dependent variables and the target variable using different concepts is summarised in Table 6.

Order of Importance	RandomForest Regressor	F-statistic and p-values	Mutual Information Scores
1	no	no	borough
2	o3	air_temp	no
3	borough	borough	nox
4	pm10	o3	no2
5	ws	ws	pm10
6	nox	wd	o3
7	no2	no2	wd
8	air_temp	nox	ws
9	wd	pm2.5	pm2.5
10	pm2.5	pm10	air_temp

Table 7: Feature importance in predicting asthma admissions

The importance of the independent features in predicting the asthma hospital admissions count in the decreasing order of their importance is listed in Table 7.

After correlation analysis, Hopkin's test for clustering tendency was performed on the samples and produced a value of **0.93**, which confirmed the presence of clusters. Regarding the optimal number of clusters, as most of the methods used in the estimation suggested a value of two, the "k" value was decided as two.

As the next step, the traditional K-Means clustering with initialiser as "k-means++" was applied to the samples to compare the performance of the DEC algorithm which produced the results as shown in the figure below.

```
Baseline Kmeans sh_score: 0.2293,  ch_score: 13960.3206,  db_score: 1.5550====>
clustering time:  27.29494333267212
```

Figure 56: Baseline K-Means metric scores

The reconstruction loss for each epoch in the pre-training phase of the DEC algorithm is depicted in the figure shown below.

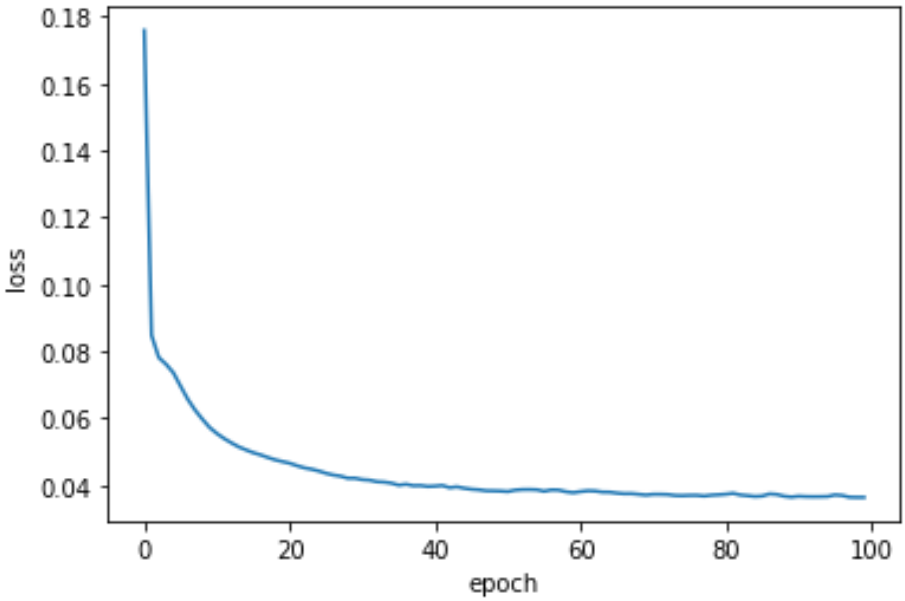


Figure 57: Epoch versus reconstruction loss for DEC pretraining phase

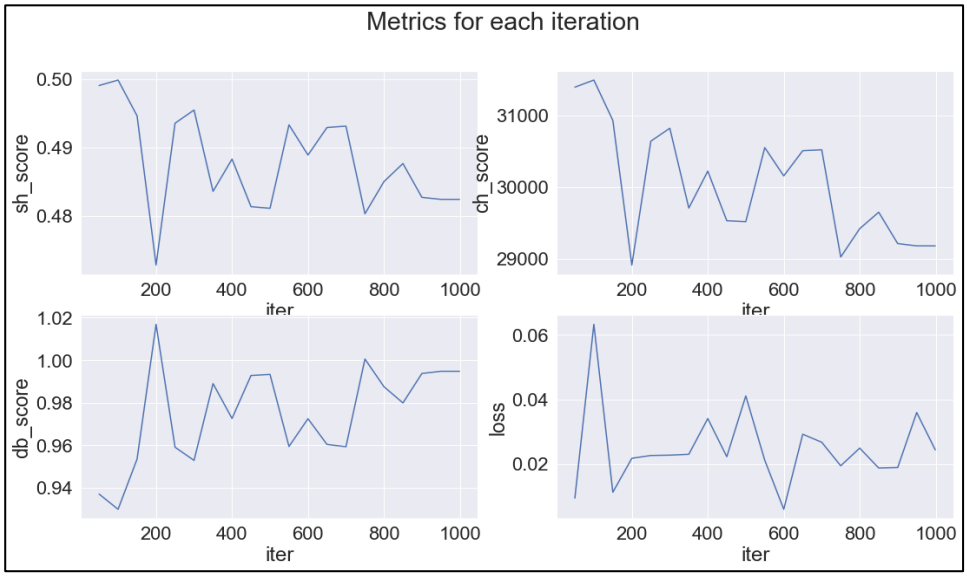


Figure 58: Metrics for DEC clustering phase

The autoencoder is compiled using “adam” as the initialiser and pre-trained for 100 epochs with a batch size of 128. The results mentioned in this report were produced by the DEC model trained for a maximum of 1000 iterations with weights updated at regular intervals of 50 or until the threshold convergence reached 0.00001.

The values of the metric scores for the baseline K-Means clustering and the Deep Embedded Clustering methods can be compared using the table below.

Phase	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index	Clustering Time
Baseline K-Means	0.2293	13960.3206	1.5550	27.29 seconds
DEC (1000 iterations)	0.48239	29176.87	0.99476	662.27 seconds

Table 8: Comparison of baseline K-Means and DEC metric scores

7. DISCUSSION

This chapter presents a detailed insight into the results presented in the previous chapter. The results extracted from this research can be grouped into two sections. The first section discusses the correlation between air pollutants and asthma admissions variables and the second section elaborates on the clustering results.

As shown in Table 6, there exists an extremely **low** correlation between different air pollutants factors and asthma admission counts according to Pearson's correlation, Spearman's correlation and Kendall's correlation methods. Among these, the score calculation for the former is based on feature values whereas the remaining ones implement a rank-based approach. The features with comparatively higher values are "no" (-0.28) for Pearson's correlation, "o3" (-0.27) for Spearman's correlation and "borough" (-0.18) for Kendall's correlation.

The Distance correlation and pairwise Mutual Information scores are distance based and exhibit improved correlation with asthma as shown in Table 6: Correlation of air pollutant variables with asthma admissions. The Distance correlation detected a **moderate positive** correlation of asthma with five variables like "no", "borough", "o3", "no2" and "nox" with values of 0.41, 0.38, 0.36, 0.35 and 0.33 respectively. As mentioned earlier, the MI score was computed using a package which was evaluated and validated using large atmospheric datasets like the raw data set used in this study. The MI approach was able to derive a **strong positive** correlation for the variables "borough", "no", "nox" and "no2" with scores (0.87, 0.79, 0.77, 0.75), a moderate positive correlation for the variables "pm10", and "o3" with values (0.58, 0.48), and a **low positive** correlation for the variables ws and wd with values (0.35, 0.33). Besides the only negative correlation of asthma is with air_temp as the asthma exacerbations are triggered at low temperatures.

In the feature importance method, if only the five most prominent features are considered for each method like RandomForest Regressor, F-statistic and p-values, and Mutual Information Scores, "no" and "borough" were in the top five for all the three methods. Also, two methods had "o3", "pm10" and "ws" whereas only one method had "air_temp", "nox" and "no2" in their top five list.

As shown in Figure 56, the Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index for the baseline K-Means clustering are 0.2293, 13960.3214, and 1.5551 respectively. The value of the Silhouette Coefficient should be close to one and that of the Davies-Bouldin Index should be close to 0 for perfect clustering. Although Calinski-Harabasz does not have a bound value, higher values imply better clustering. The results of the baseline method indicate poor clustering output and are improved by applying DEC.

Figure 57 shows the reconstruction loss in the pretraining phase which steadily decreased and converged to 0.024295637 after 100 iterations. Figure 58 shows the clustering metric score for 1000 iterations. It can be observed that the Silhouette Coefficient and Calinski-Harabasz Index follow a similar trend. However, the values are fluctuating throughout the entire process. Xu Wang and Yusheng Xu proposed an improved index, Peak Weight Index[33] (PWI) based on Silhouette Coefficient and Calinski-Harabasz Index to evaluate the clustering results. The authors claim that the sensitivity of both the index values which causes them to fluctuate and produce different results for the same clustering function can be resolved by combining the characteristics of both index values.

Table 8 displays the comparison of the metric scores obtained after baseline K-Means clustering and Deep Embedded Clustering methods. Silhouette Coefficient and Calinski-Harabasz Index produced after DEC is more than double the value of baseline metric values while Davies-Bouldin Index reduced by 36%, which indicates better cluster outputs. However, the clustering time increased to 24 times the baseline value.

The ANOVA test showed that no correlation exists between cluster labels and asthma admissions.

7.1 Limitations

The data for most of the air pollutants were missing for the remaining boroughs. The availability of these data would have given a coherent picture by enhancing the clustering output.

Also, the hourly air pollutant values of the selected fourteen boroughs had numerous missing values which were managed using interpolation techniques. In some scenarios, this might have led to the addition of incorrect data which would have affected the results adversely.

In addition, the pollutant values consisted of negative values which might have entered the database because of the calibration error in the sensors measuring them or a processing error. The in-depth knowledge in this domain would have been helpful to decide the most appropriate techniques to handle them and to increase the reliability of the study outcomes.

Further, the daily data on asthma admissions and mortality were not available for the study and the inclusion of this data into the analysis would have increased the robustness of the research findings.

Finally, the hyperparameters of the DEC algorithm like number of hidden layers, nodes in each hidden layer, pretraining epochs, maximum iterations, update interval, initializers etc., can be further fine-tuned to produce better clustering results.

8. CONCLUSIONS

This study was an attempt to investigate the relationship between air pollution and asthma severities in the London area. The distance-based metrics were able to detect strong correlations between air pollutant variables and asthma admissions, which the value-based and rank-based correlation metrics were not able to. The clustering outputs of traditional clustering algorithms can be improved and fine-tuned by applying them to learned low-dimensional (latent space) features. The cluster labels and admissions had no correlation between them.

There is a large scope for future work on this topic which links two different yet related fields i.e., environment and healthcare. It is proven that asthma exacerbations can also result from indoor air pollution arising from gas stoves, central heating and cooling systems, building materials, pressed wood products, moisture, bacteria, fungus, dust, fireplaces, pets etc., which are not contemplated in this study. Better outcomes can be obtained if other relevant factors triggering the asthma symptoms are also included as part of the study.

The deep learning techniques give comparatively better results with high dimensional data and hence the DEC performance can be improved by increasing the dimensionality of the data with the addition of more relevant feature variables.

The approach used in this study can be used to identify the areas with high atmospheric pollutants for the health care professionals to focus on. Also, an investigation can be performed for finding the reason behind the higher concentrations of air pollutants in particular areas and to adopt preventive measures to reduce them, thereby making the lives of asthma patients easier.

Deep clustering is highly beneficial in analysing complex data which are unstructured and high-dimensional like images, texts, sequences, expressions etc. The superior performance of deep clustering algorithms especially in handling high dimensional data compared to traditional clustering algorithms has been proven many years ago. Despite all these advantages, identifying the hidden patterns and trends in the data in the absence of labels is indeed an area demanding intense research.

9. REFERENCES AND BIBLIOGRAPHY

1. KOENIG, J.Q., 1999. Air pollution and asthma. Journal of Allergy and Clinical Immunology, 104(4), 717-722

2. Lai CK, Beasley R, Crane J, Foliaki S, Shah J, Weiland S; International Study of Asthma and Allergies in Childhood Phase Three Study Group. Global variation in the prevalence and severity of asthma symptoms: phase three of the International Study of Asthma and Allergies in Childhood (ISAAC). Thorax. 2009 Jun;64(6):476-83. DOI: 10.1136/thx.2008.106609. Epub 2009 Feb 22. PMID: 19237391.

3. WHO Asthma-Fact sheets b. [viewed Jul 6, 2022]. Available from: https://www.who.int/en/news-room/fact-sheets/detail/asthma

4. TRASANDE, L. and G.D. THURSTON, 2005. The role of air pollution in asthma and other pediatric morbidities. Journal of Allergy and Clinical Immunology, 115(4), 689-699

5. ANON., a. https://www.asthmaandlung.org.uk/1265-2/. In: Asthma & Lung UK. [viewed Jul 6, 2022].Available from: https://www.asthmaandlung.org.uk/1265-2/

6. A. N. S., A. Y. Nair, and V. S., "Determining the Effect of Correlation between Asthma/Gross Domestic Product and Air Pollution," 2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), 2022, pp. 44-48, DOI: 10.1109/WiSPNET54241.2022.9767145.

7. HAJMOHAMMADI, H. et al., 2022. Association between short-term NOx exposure and asthma exacerbations in East London: A time series regression model. Urban Climate, 44, 101173

8. Lee, Eu Sun et al., 2022. A Machine Learning-Based Study of the Effects of Air Pollution and Weather in Respiratory Disease Patients Visiting Emergency Departments. Emergency Medicine International

9. Khatri, Sumita B., 2021. Associations of Air Pollution and Pediatric Asthma in Cleveland, Ohio. The Scientific World Journal

10. DELAMATER, P.L., A.O. FINLEY and S. BANERJEE, 2012. An analysis of asthma hospitalizations, air pollution, and weather conditions in Los Angeles County, California. Science of The Total Environment, 425, 110-118

11. Raun, L.H., Ensor, K.B. & Persse, D. Using community level strategies to reduce asthma attacks triggered by outdoor air pollution: a case crossover analysis. Environ Health 13, 58 (2014). https://doi.org/10.1186/1476-069X-13-58

12. QIUMIN ZHAI et al., 2011. Correlation analysis of air pollution and respiratory disease in City A. - 2011 International Conference on Remote Sensing, Environment and Transportation Engineering. pp.197-200

13. AKINBAMI, L.J. et al., 2010. The association between childhood asthma prevalence and monitored air pollutants in metropolitan areas, United States, 2001–2004. Environmental Research, 110(3), 294-301

14. COX, L.A.(., 2017. Socioeconomic and air pollution correlates of adult asthma, heart attack, and stroke risks in the United States, 2010–2013. Environmental Research, 155, 92-107

15. M. N. Hoq, R. Alam, and A. Amin, "Prediction of a possible asthma attack from air pollutants: Towards a high-density air pollution map for smart cities to improve living," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-5DOI: 10.1109/ECACE.2019.8679335.

16. Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22.

17. 4. R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining", Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining Citeseer, pp. 29-39, 2000.

18. Carslaw, D. C., and K. Ropkins. 2012. “openair — An R package for air quality data analysis.” Environmental Modelling & Software 27–28 (0): 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008.

19. Lu, Y., Fang, J., Tian, L. and Jin, H., 2015. Advanced Medical Statistics. 2nd ed. World Scientific Publishing Company, p.1384.

20. C. J. Kowalski, “On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient” Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 21, No. 1 (1972), pp. 1-12.

21. Zwillinger, D. and Kokoska, S. (2000). CRC Standard Probability and Statistics Tables and Formulae. Chapman & Hall: New York. 2000. Section 14.7

22. Maurice G. Kendall, “A New Measure of Rank Correlation”, Biometrika Vol. 30, No. 1/2, pp. 81-93, 1938.

23. LAARNE, P., M.A. ZAIDAN and T. NIEMINEN, 2021. ennemi: Non-linear correlation detection with mutual information. SoftwareX, 14, 100686

24. Bartlett, M. S. (1951). The effect of standardization on a Chi-square approximation in factor analysis. Biometrika, 38, 337-344.

25. LAWSON, R.G. and P.C. JURS, 1990. A new index for clustering tendency and its application to chemical problems. Journal of Chemical Information and Computer Sciences, 30(1), 36-41

26. TIBSHIRANI, R., G. WALTHER and T. HASTIE, 2001. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. Journal of the Royal Statistical Society Series B, 63, 411-423ss

27. Mohajer, Mojgan & Englmeier, Karl-Hans & Schmid, Volker. (2011). A comparison of Gap statistic definitions with and without logarithm function. Computing Research Repository - CORR.

28. Xie, J., Girshick, R., and Farhadi, A., “Unsupervised Deep Embedding for Clustering Analysis”, arXiv e-prints, 2015.

29. ROUSSEEUW, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65

30. CALIŃSKI, T. and H. JA, 1974. A Dendrite Method for Cluster Analysis. Communications in Statistics - Theory and Methods, 3, 1-27

31. D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, April 1979, DOI: 10.1109/TPAMI.1979.4766909.

32. RINK, K., 2022. Best Practices for Visualizing Your Cluster Results [viewed Sep 8, 2022]. Available from: https://towardsdatascience.com/best-practices-for-visualizing-your-cluster-results-20a3baac7426

33. WANG, X. and Y. XU, 2019. An improved index for clustering validation based on the Silhouette index and Calinski-Harabasz index. IOP Conference Series: Materials Science and Engineering, 569(5), 052024

34. XifengGuo .https://github.com/XifengGuo/DEC-keras/blob/master/DEC.py

35. Deng Cai, Xiaofei He, and Jiawei Han. Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering (TKDE), 23(6):902–913, 2011.

36. Strehl, Alexander and Joydeep Ghosh. “Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions.” Journal of Machine Learning Research (JMLR), 3(Dec):583–617, 2002.

37. Harold W Kuhn. The Hungarian method for the assignment problem. Naval Research Logistics (NRL), 2(1-2):83–97, 1955.

38. J. Noh, J. Sohn, S. Cho, Y. Choi, C. Kim, and D. Shin, "Short-term effects of ambient air pollution on emergency department visits for asthma: An assessment of effect modification by prior allergic disease history", J. Preventive Med. Public Health, vol. 49, no. 5, pp. 329-341, 2016.

39. J. Halonen, T. Lanki, T. Yli-Tuomi, M. Kulmala, P. Tiittanen and J. Pekkanen, "Urban air pollution and asthma and COPD hospital emergency room visits", Thorax, vol. 63, no. 7, pp. 635-641, 2008.

40. L. Hubert and P. Arabie, Comparing Partitions, Journal of Classification 1985 https://link.springer.com/article/10.1007%2FBF01908075

41. JACQUEMIN BÉNÉDICTE et al., 2015. Ambient Air Pollution and Adult Asthma Incidence in Six European Cohorts (ESCAPE). Environmental health perspectives, 123(6), 613-621

42. Health matters: air pollution [viewed Aug 27, 2022]. Available from: https://www.gov.uk/government/publications/health-matters-air-pollution/health-matters-air-pollution

43. Gábor J. Székely. Maria L. Rizzo. Nail K. Bakirov. "Measuring and testing dependence by correlation of distances." Ann. Statist. 35 (6) 2769 - 2794, December 2007. https://doi.org/10.1214/009053607000000505

44. Calinski-Harabasz Index for K-Means Clustering Evaluation using Python 2022. [viewed Sep 5, 2022]. Available from: https://pyshark.com/calinski-harabasz-index-for-k-means-clustering-evaluation-using-python/

45. Encyclopedia of bioinformatics and computational biology 9780128114148 [viewed Sep 9, 2022]. Available from: https://dokumen.pub/encyclopedia-of-bioinformatics-and-computational-biology-9780128114148.html

46. ennemi on Pypi [viewed Sep 9, 2022]. Available from: https://libraries.io/pypi/ennemi

47. User Guide PDF | PDF | Thread (Computing) | Matrix (Mathematics) [viewed Sep 9, 2022]. Available from: https://www.scribd.com/document/396695783/User-Guide

Appendix 1:Ethical clearance for research and innovation projects

Project status

Status

Approved

Actions

Date	Who	Action	Comments
17:44:00 12 July 2022	Femi Isiaq	Supervisor approved	Ensure you are adhere to the use of publicly provided dataset as discussed.
19:25:00 07 July 2022	Lekshmi Vijayakumar Sudha Kumari	Principal investigator submitted	

Ethics release checklist (ERC)

Project details

Project name: A study of the impact of ambient air pollution on the severity of asthma patients in London

Principal investigator: Lekshmi Vijayakumar Sudha Kumar

Faculty: Faculty of Business, Law and Digital Technologies

Level: Postgraduate

Course: MSc Applied AI and Data Science

Unit code: COM726

Supervisor name: Femi Isiaq

Supervisor search:

Other investigators:

Checklist

Question	Yes	No
Q1. Will the project involve human participants other than the investigator(s)?	○	●
Q1a. Will the project involve vulnerable participants such as children, young people, disabled people, the elderly, people with declared mental health issues, prisoners, people in health or social care settings, addicts, or those with learning difficulties or cognitive impairment either contacted directly or via a gatekeeper (for example a professional who runs an organisation through which participants are accessed; a service provider; a care-giver; a relative or a guardian)?	○	●
Q1b.Will the project involve the use of control groups or the use of deception?	○	●
Q1c. Will the project involve any risk to the participants' health (e.g. intrusive intervention such as the administration of drugs or other substances, or vigorous physical exercise), or involve psychological stress, anxiety, humiliation, physical pain or discomfort to the investigator(s) and/or the participants?	○	●
Q1d. Will the project involve financial inducement offered to participants other than reasonable expenses and compensation for time?	○	●
Q1e. Will the project be carried out by individuals unconnected with the University but who wish to use staff and/or students of the University as participants?	○	●
Q2. Will the project involve sensitive materials or topics that might be considered offensive, distressing, politically or socially sensitive, deeply personal or in breach of the law (for example criminal activities, sexual behaviour, ethnic status, personal appearance, experience of violence, addiction, religion, or financial circumstances)?	○	●
Q3. Will the project have detrimental impact on the environment, habitat or species?	○	●
Q4. Will the project involve living animal subjects?	○	●
Q5. Will the project involve the development for export of 'controlled' goods regulated by the Export Control Organisation (ECO)? (This specifically means military goods, so called dual-use goods (which are civilian goods but with a potential military use or application), products used for torture and repression, radioactive sources.) Further information from the Export Control Organisation *	○	●
Q6. Does your research involve: the storage of records on a computer, electronic transmissions, or visits to websites, which are associated with terrorist or extreme groups or other security sensitive material? Further information from the Information Commissioners Office *	○	●

Declarations

I/we, the investigator(s), confirm that:

- [x] The information contained in this checklist is correct.

- [x] I/we have assessed the ethical considerations in relation to the project in line with the University Ethics Policy.

- [x] I/we understand that the ethical considerations of the project will need to be re-assessed if there are any changes to it.

- [x] I/we will endeavour to preserve the reputation of the University and protect the health and safety of all those involved when conducting this research/enterprise project.

- [x] If personal data is to be collected as part of my project, I confirm that my project and I, as Principal Investigator, will adhere to the General Data Protection Regulation (GDPR) and the Data Protection Act 2018. I also confirm that I will seek advice on the DPA, as necessary, by referring to the Information Commissioner's Office further guidance on DPA and/or by contacting information.rights@solent.ac.uk. By Personal data, I understand any data that I will collect as part of my project that can identify an individual, whether in personal or family life, business or profession.

- [x] I/we have read the prevent agenda.