Exploring Environmental Data: Supervised and Unsupervised Learning Insights through Multiple Linear Regression, Multiple Logistic Regression, and Clustering

1st Rizwan Karim Rafat
School of Computer, Data and Mathematical Sciences
Western Sydney University
Sydney, Australia
22061272@student.westernsydney.edu.au

***Abstract*—In this research paper we conduct a detailed analysis of a large environment dataset using a combination of supervised and unsupervised learning techniques. We use multiple linear regression and multiple logistic regression to establish relationship between environmental factors and observed results along with binary event likelihood and factors which cause them. Simultaneously, k-means clustering is applied on the dataset to unveil hidden patterns and structures. This research demonstrates the success of machine learning in extracting knowledge from environmental dataset and highlights the significance of multidimensional approach to understand the natural world to make responsible and informed environmental management practices.**

***Index Terms*—environmental, data, supervised, unsupervised, regression, clustering, modeling, analysis, groupings, factors, likelihood**

I. Introduction

Environmental data is complex and not easy to understand. Machine learning algorithms helps us to find patterns and relationships in this data. This analysis and report focuses on supervised and unsupervised approach which includes multiple linear regression, multiple logistic regression and k-means clustering.

II. Data Exploration

In order to extract knowledge from the dataset we need to explore the data from different angles. Initially, it is important to understand the number of observations and variables we have in the dataset. As shown in Appendix A, there are 108 observations and 21 variables in the environment dataset. Among the variables there are chr, int, and num types.

Tables (see Table I and Table II) containing the summary of the dataset has been provided below as illustrated in Appendix A. The summary includes the minimum, 1st quantile, median, mean, 3rd quantile and maximum value for each variable.

The distribution of the length_of_stay_minutes has been illustrated in Figure 1. From the figure we can observe that the histogram is right skewed, meaning most of the data points has a length of stay of around 150 to 200 minutes.

TABLE I
Summary of Environment Dataset

Variable	Description	Value
length_of_stay_minutes	Minimum	37.0
	1st Quantile	123.2
	Median	181.0
	Mean	234.0
	3rd Quantile	277.8
	Maximum	979.0
co	Minimum	0.000
	1st Quantile	0.100
	Median	0.200
	Mean	0.2241
	3rd Quantile	0.300
	Maximum	1.200
o3	Minimum	0.00
	1st Quantile	12.00
	Median	18.00
	Mean	18.87
	3rd Quantile	23.25
	Maximum	60.00
no2	Minimum	0.000
	1st Quantile	3.750
	Median	6.000
	Mean	8.343
	3rd Quantile	12.000
	Maximum	28.000
so2	Minimum	0.0000
	1st Quantile	0.0000
	Median	0.0000
	Mean	0.7407
	3rd Quantile	1.0000
	Maximum	7.0000
ppm10	Minimum	2.90
	1st Quantile	10.60
	Median	15.75
	Mean	19.26
	3rd Quantile	22.38
	Maximum	70.20
visibility_reduction	Minimum	0.2200
	1st Quantile	0.3400
	Median	0.4000
	Mean	0.5266
	3rd Quantile	0.6300
	Maximum	1.7300

TABLE II
SUMMARY OF ENVIRONMENT DATASET (CONTINUES)

Variable	Description	Value
aqi	Minimum	11.00
	1st Quantile	19.00
	Median	25.00
	Mean	29.89
	3rd Quantile	35.50
	Maximum	88.00
precipitation	Minimum	0.0000
	1st Quantile	0.0000
	Median	0.0000
	Mean	0.7037
	3rd Quantile	0.2000
	Maximum	29.0000
relativehumidity	Minimum	17.00
	1st Quantile	50.75
	Median	67.50
	Mean	66.04
	3rd Quantile	79.25
	Maximum	100.00
vapourpressure	Minimum	5.600
	1st Quantile	9.825
	Median	11.800
	Mean	12.604
	3rd Quantile	14.525
	Maximum	25.000
windspeed	Minimum	0.000
	1st Quantile	2.100
	Median	3.100
	Mean	3.863
	3rd Quantile	5.700
	Maximum	13.400
winddirection	Minimum	0.0
	1st Quantile	110.0
	Median	170.0
	Mean	169.4
	3rd Quantile	240.0
	Maximum	360.0
maxwindspeed	Minimum	0.00
	1st Quantile	3.10
	Median	5.10
	Mean	5.86
	3rd Quantile	8.20
	Maximum	18.00

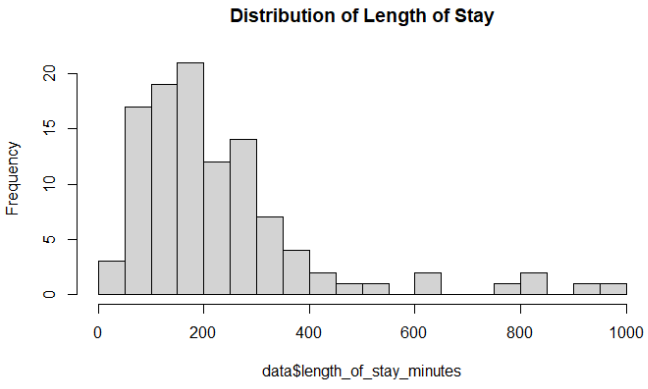


Fig. 1.

III. DATA PRE-PROCESSING

Data pre-processing is an important step before starting any analysis. It is the process of preparing the raw data so that it is ready for any different type of data processing procedure. It helps the data to transform that can be easily and effectively be used for machine learning [1].

To successfully carry out the analysis, we have checked for the following issues as listed below and made amend mends accordingly.

- Check for missing values (if missing values found, replace them with the mean)
- Check for appropriate data type
- Check for outliers
- Select the relevant features
- Normalize data if needed

For the purpose of this analysis and report, all the numeric and and integer columns have been separated into a new data frame named **data_new** for multiple linear regression. Similarly for the multiple logistic regression, the column **asthma** has been converted into factors.

IV. AIMS AND OBJECTIVES

The main goal of this study is to dive deep into the environment dataset which includes numerous factors that influence the environmental conditions. This analysis will help to clearly understand how these factors are interconnected.

Additionally, we look forward to building predictive models using multiple linear regression and multiple logistic regression. These models will help us to classify environmental outcomes based on the variables available. Simultaneously, we aim to create a k-means clustering algorithms to discover natural groupings in the dataset. These clusters will help reveal hidden patterns and categories.

To successfully achieve our aims, we have outlined a series of processes. We will start by exploring the data and then preprocess the dataset where necessary. We will apply multiple linear regression technique to model and quantify the relationship between environmental outcomes and specific outcomes of interest.

Simultaneously, we will use multiple logistic regression to predict binary outcomes such as if a person has asthma or not. This classification can be beneficial for early detection and response to environmental challenges.

V. SUPERVISED LEARNING: MULTIPLE LINEAR REGRESSION

Multiple linear regression is the statistical method used to model the relationship between multiple independent variables and a single dependent target variable. It is an extension of the simple linear regression method [2].

Multiple linear regression can be represented by the following equation:

$$E(Y) = \hat{a} + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n \qquad (1)$$

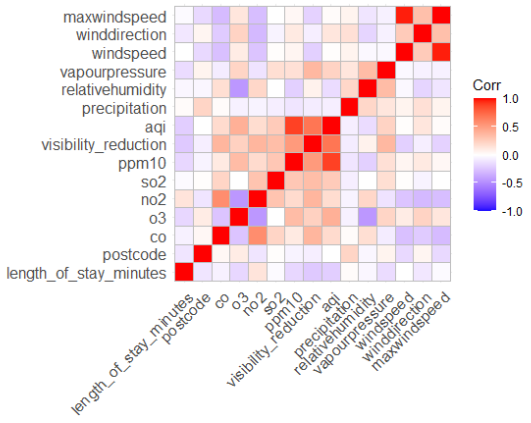


Fig. 2. Correlation between Environmental Variables

- Y represents the dependent variable
- $X_1, X_2 + \ldots + X_n$
- $\hat{a}$ is the intercept
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables

Multiple linear regression relies on different assumptions which includes linearity between variables, independence between errors, homoscedasticity and normailty of error.

Our research question for this analysis is as follows:

What is the combined influence of air quality parameters (CO, O3, NO2, SO2, PPM10, visibility reduction, AQI), weather conditions (precipitation, relative humidity, vapor pressure, windspeed, wind direction, max windspeed) on the length of stay for patients?

For the purpose of our analysis, we have created a new data frame named **data_new**. This data frame only included the numeric and integer variables.

The first thing we need to check here is the correlation between the variables. We have plotted a correlation plot (as shown in Figure 2) which helps to understand the which variables we can exclude before we start the multiple linear regression.

As shown in Figure 2 we can observe that ppm10 has strong correlation with aqi (1.00000000) and maxwindspeed has strong correlation with windspeed (1.00000000). The correlation values has also been calculated using the cor function as mentioned in Appendix B.

We have build our first model, **model1**, with postcode, co, o3, no2, so2, visibility_reduction, aqi, precipitation, relativehumidity, vapourpressure, windspeed and winddirection.

Estimate of Intercept = $\hat{a}$ = 664.14630

Estimate of the Slope corresponding to postcode = β_1 = -0.09744

Estimate of the Slope corresponding to co = β_2 = -110.97013

Estimate of the Slope corresponding to o3 = β_3 = 0.18389

Estimate of the Slope corresponding to no2 = β_4 = 9.99191

Estimate of the Slope corresponding to so2 = β_5 = 2.55109

Estimate of the Slope corresponding to visibility_reduction = β_6 = -167.05947

Estimate of the Slope corresponding to aqi = β_7 = -0.88205

Estimate of the Slope corresponding to precipitation = β_8 = 4.08728

Estimate of the Slope corresponding to relativehumidity = β_9 = -1.06282

Estimate of the Slope corresponding to vapourpressure = β_10 = 1.32790

Estimate of the Slope corresponding to windspeed = β_11 = -0.67520

Estimate of the Slope corresponding to winddirection = β_12 = -0.11318

Now using backward selection method i.e. by removing the variable with the highest p-value one after another, we have established 11 different models among which **model11** seems to give the best result.

The model thus obtained is:

$$length\_of\_stay\_minutes = 275.412 + (7.912) * no2 + (-203.912) * visibility\_reduction$$

This model gives an AIC of 1115.73 which is the lowest among all the other models.

R^2 = 0.1152

This implies that about 11.52% of the variation in the data set is explained by the model.

Residual Standard Error = 172.7

This implies that on average, the predicted values deviate from the true regression line by 172.7.

It can be seen clearly from the anova table that, F-statistics = 6.834 on 2 and 105 degrees of freedom which is greater than 3.082852. Hence, the model is adequate and no2 and visibility_reduction have significant linear relationship with length_of_stay_minutes.

The residuals vs fitted values graph (as shown in Figure 3) shows a fan-shaped pattern. This indicates that the variance of the residuals is not constant across all fitted values. It is possible that the regression model is not a good fit to the data, or that the assumptions of linear regression are not met.

The histogram of residuals (as shown in Figure 4) shows that it is a bell-shaped distribution and that the residuals are normally distributed with the center around zero.

Overall, the histogram of residuals suggests that the regression model is a good fit to the data and that the assumptions of linear regression are met.

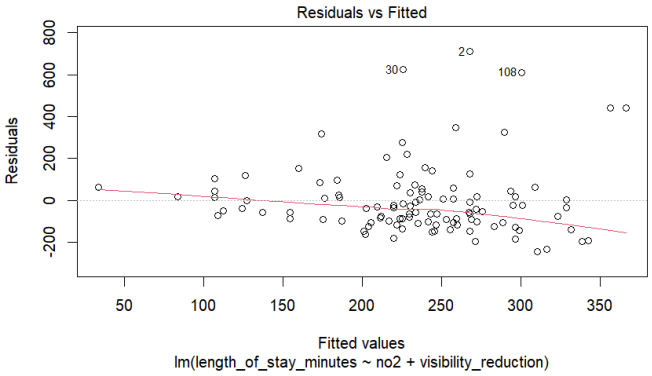


Fig. 3. Residuals vs Fitted Values

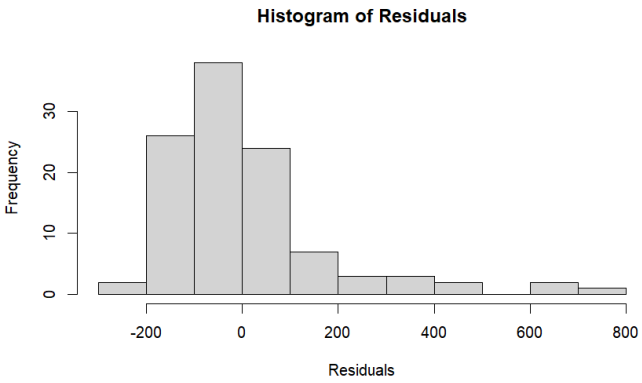


Fig. 4. Histogram of Residuals

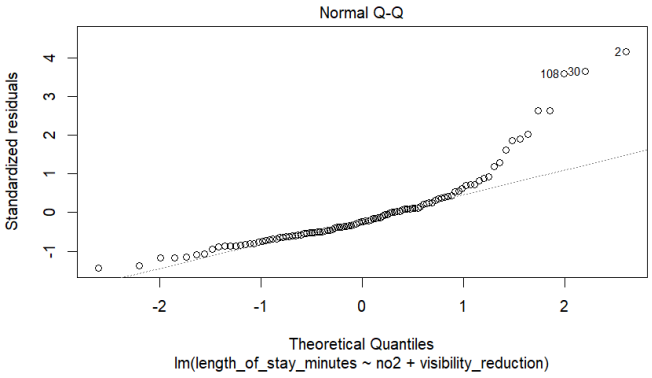


Fig. 5. Normal Q-Q Plot

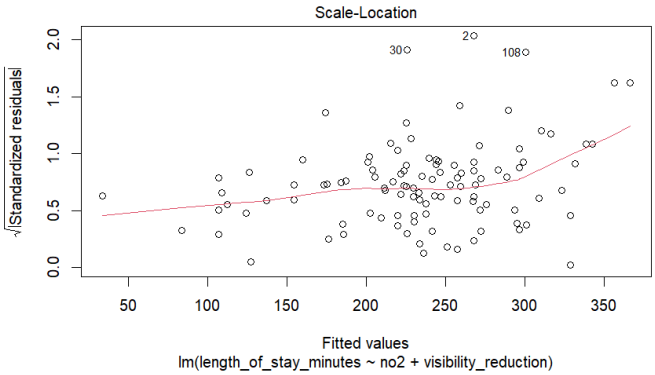


Fig. 6. Scale vs Location Plot

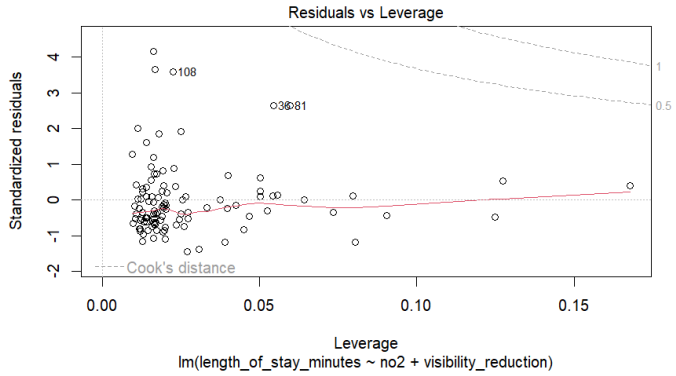


Fig. 7. Residuals vs Leverage Plot

The Q-Q plot (as shown in Figure 5) shows a slight curvature to the right. This suggests that the residuals are slightly skewed to the right, meaning that there are slightly more positive residuals than negative residuals. However, the curvature is not severe, so it is unlikely to have a significant impact on the results of the regression analysis.

The scale-location plot (as shown in Figure 6) shows a horizontal band. This indicates that the variance of the residuals is constant across all fitted values. This is a good sign, as it suggests that the homoscedasticity assumption of linear regression is met.

The residuals vs leverage plot (as shown in Figure 7) shows a few high-leverage points. High-leverage points are observations that have a strong influence on the regression model. These points may be outliers or simply observations that are very different from the rest of the data.

VI. Supervised Learning: Multiple Logistic Regression

Multiple logistic regression is the statistical method used for modeling the relationship between multiple predictor variables and a binary target variable [3].

The formula for logistic regression is:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}} \quad (2)$$

- $P(Y=1)$ is the probability of the outcome variable
- e is the base of the natural logarithm
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients estimated by the model
- $X_1, X_2, \ldots, X_p$ are the predictor variables

Logistic regression assumes that the log-odds of the binary outcome is a linear combination of the predictor variables and that the residuals are independent and follow a logistic distribution.

Our research question for this analysis is as follows:

Can we predict the likelihood of asthma attacks based on patient characteristics, environmental factors, and triage information using logistic regression?

For the purpose of our analysis, we have converted the asthma column into factors and then used the **glm** function to initiate the multiple logistic regression. Using backward selection method, 9 different models have been created. Please refer to Appendix C for more details.

A summary of the AIC and AUC values have been provided in the Table III.

TABLE III
AUC AND AIC FOR THE 9 MODELS

Model	AUC	AIC
1	0.7965368	139.0665
2	0.7987013	137.1284
3	0.7936508	135.2372
4	0.7940115	133.4179
5	0.7954545	131.5602
6	0.7907648	130.7563
7	0.7867965	129.5179
8	0.7676768	128.6508
9	0.754329	127.8868

From the table we can observe that model 2 has highest AUC (area under the curve) of 0.7987013 as shown in Figure 8 with the ROC curve. Although the AIC for model 9 is not the lowest, meaning that the model has more complexity preventing it from overfitting. Thus we can conclude that model 2 is the best model among all the models.

The equation of the model is:

$$log(odds\_of\_asthma) = -6.295517 + (0.001664 * length\_of\_stay\_minutes) + (0.001833 * postcode) + (1.704273 * co) - (0.064458 * o3) - (0.504912 * so2) + (0.896287 * visibility\_reduction) + (0.008011 * aqi) + (0.429285 * precipitation) + (0.009311 * relativehumidity) - (0.032544 * vapourpressure) + (0.093680 * windspeed) - (0.004286 * winddirection)$$

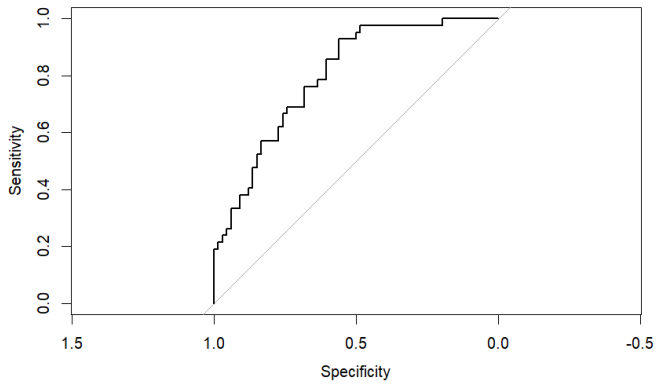


Fig. 8. ROC Curve

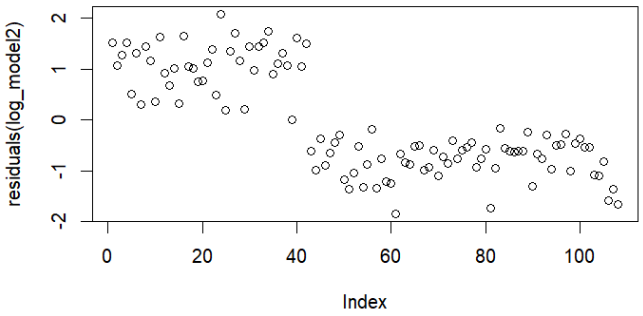


Fig. 9. Residuals vs Index Scatterplot

To obtain the estimated probability of asthma, we can apply the logistic function (sigmoid function) to the log-odds:

$$P(asthma) = 1/(1+e^{-log(odds\_of\_asthma)})$$

We have also generated a misclassification matrix that gives us the following:

- Misclassification Rate: 0.2685185
- False Positive Rate: 0.1666667
- False Negative Rate: 0.4285714

In Figure 9 we can observe that the residuals are randomly distributed around the index. This means that the model is not overfitting the data. It also shows that there is a slight downward trend in the residuals. This suggests that the model is slightly underpredicting the values for higher values of the index.

VII. UNSUPERVISED LEARNING: K-MEANS CLUSTERING

K-means clustering is a simple and unsupervised machine learning algorithm that is used to group similar data points together. It works by iteratively assigning data points to clusters based on their distance to the cluster centers [4].

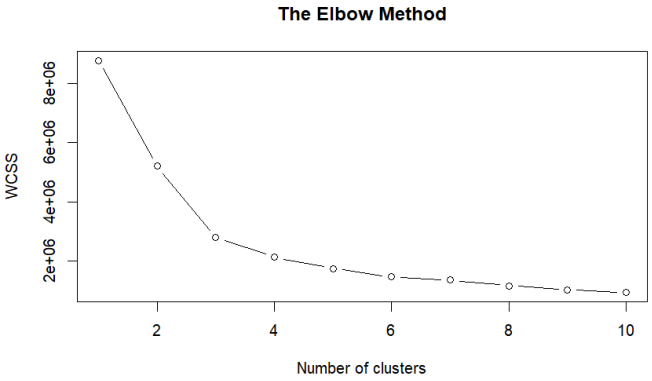


Fig. 10. The Elbow Method

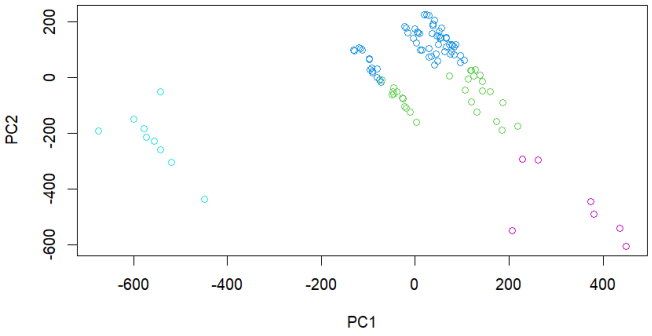


Fig. 11. PC1 vs PC2 Scatterplot

For our analysis we have first tried to understand how many clusters are ideal for this dataset. Figure 10 illustrates that 4 clusters are ideal for this dataset using the elbow method. Thus we have used the **kmeans** function as shown in Appendix D to initiate the clustering.

After successful clustering, we get between_SS / total_SS ratio of 75.6% which means that 75.6% of the total variation in the data is explained by the clustering. This is a relatively high ratio, which suggests that the clustering is meaningful.

Figure 11 clear shows the 4 different clusters. The data points that are close together in the PC1-PC2 space are more similar than the data points that are far apart.

VIII. Model Comparison

Multiple linear regression is a supervised learning algorithm that is used to predict a continuous target variable based on a set of predictor variables. It works by fitting a linear equation to the data, where the coefficients of the equation represent the relationship between the predictor variables and the target variable [6].

Multiple logistic regression is a supervised learning algorithm that is used to predict a binary target variable based on a set of predictor variables. It works by fitting a logistic curve to the data, where the coefficients of the curve represent the relationship between the predictor variables and the target variable [6].

K-Means clustering is an unsupervised learning algorithm that is used to group data points into clusters based on their similarity. K-Means clustering works by initializing a set of cluster centers and then iteratively assigning data points to the cluster centers that are closest to them. The cluster centers are then updated based on the data points that have been assigned to them [5].

IX. Results and Recommendation

Recommendations for improving multiple linear regression:

- We can regularize the model to prevent overfitting.
- We can use cross-validation to evaluate the model's performance on unseen data.

Recommendations for improving multiple logistic regression:

- We can regularize the model to prevent overfitting.
- We can use cross-validation to evaluate the model's performance on unseen data.

Recommendations for improving k-means clustering:

- We can use feature scaling to normalize the data.
- We can use a variety of initialization methods.

References

[1] https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing
[2] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.
[3] Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). Applied Logistic Regression. Wiley.
[4] Dhanachandra, N., Manglem, K. and Chanu, Y.J. (2015) 'Image segmentation using K -means clustering algorithm and subtractive clustering algorithm', Procedia Computer Science, 54, pp. 764–771. doi:10.1016/j.procs.2015.06.090.
[5] https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
[6] https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

Appendixes

Appendix A

Data Exploration

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.2.3
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.2.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
data <- read.csv("Envdata.csv")
attach(data)
head(data)
```

```
##   patient_id            triage length_of_stay_minutes postcode      age gender
## 1   PJY1ZY7M Triage 3 - Urgent                     166     3030 30 to 34 Female
## 2   PTX0ZI2D Triage 3 - Urgent                     979     3030 00 to 04 Male
## 3   PVX0GR6F Triage 3 - Urgent                      38     3030 20 to 24 Male
## 4   WOE7QE3M Triage 3 - Urgent                     184     3753 05 to 09 Male
## 5   WYU6CP0J Triage 3 - Urgent                      37     3036 05 to 09 Male
## 6   PYU6ZP0K Triage 3 - Urgent                     372     3085 05 to 09 Male
##           suburb  co o3 no2 so2 ppm10 visibility_reduction aqi precipitation
## 1       Werribee 0.4 20  13   0  12.9                0.40  20             0
## 2        Pt Cook 0.1 12   6   2   4.5                0.27  12             0
## 3 Werribee South 0.2 11   9   1  22.3                0.71  30             0
## 4      Beveridge 0.2 33   2   0   8.7                0.35  33             0
## 5         Keilor 0.6  4  15   0  17.7                1.40  60             0
## 6        Macleod 0.2 15  12   0  10.6                0.30  15             0
##   relativehumidity vapourpressure windspeed winddirection maxwindspeed asthma
## 1               76           11.2       1.5           160          2.1    Yes
## 2               77           14.4      10.8           140         13.9    Yes
## 3               74           13.8       7.2           110          9.8    Yes
## 4               17           11.6       5.1           170          7.2    Yes
## 5               86           11.9       1.5           100          2.1    Yes
## 6               73           11.6       7.7           250         13.4    Yes
```

```r
colnames(data)
```

```
##  [1] "patient_id"             "triage"                 "length_of_stay_minutes"
##  [4] "postcode"               "age"                    "gender"
##  [7] "suburb"                 "co"                     "o3"
## [10] "no2"                    "so2"                    "ppm10"
## [13] "visibility_reduction"   "aqi"                    "precipitation"
## [16] "relativehumidity"       "vapourpressure"         "windspeed"
## [19] "winddirection"          "maxwindspeed"           "asthma"
```

```r
str(data)
```

```
## 'data.frame':    108 obs. of  21 variables:
##  $ patient_id            : chr  "PJY1ZY7M" "PTX0ZI2D" "PVX0GR6F" "W0E7QE3M" ...
##  $ triage                : chr  "Triage 3 - Urgent" "Triage 3 - Urgent" "Triage 3 - Urgent" "Triage 3
##  $ length_of_stay_minutes: int  166 979 38 184 37 372 392 258 181 97 ...
##  $ postcode              : int  3030 3030 3030 3753 3036 3085 3095 3211 3094 3073 ...
##  $ age                   : chr  "30 to 34" "00 to 04" "20 to 24" "05 to 09" ...
##  $ gender                : chr  "Female" "Male  " "Male  " "Male  " ...
##  $ suburb                : chr  "Werribee" "Pt Cook" "Werribee South" "Beveridge" ...
##  $ co                    : num  0.4 0.1 0.2 0.2 0.6 0.2 1 0.1 0.2 0.2 ...
##  $ o3                    : int  20 12 11 33 4 15 1 12 4 12 ...
##  $ no2                   : int  13 6 9 2 15 12 14 6 7 6 ...
##  $ so2                   : int  0 2 1 0 0 0 0 2 0 0 ...
##  $ ppm10                 : num  12.9 4.5 22.3 8.7 17.7 10.6 11.8 4.5 15.1 11.9 ...
##  $ visibility_reduction  : num  0.4 0.27 0.71 0.35 1.4 0.3 0.58 0.27 0.41 0.38 ...
##  $ aqi                   : int  20 12 30 33 60 15 25 12 19 16 ...
##  $ precipitation         : num  0 0 0 0 0 0 0.2 0 0 8.4 ...
##  $ relativehumidity      : int  76 77 74 17 86 73 99 77 79 99 ...
##  $ vapourpressure        : num  11.2 14.4 13.8 11.6 11.9 11.6 11.1 14.4 13.3 11.2 ...
##  $ windspeed             : num  1.5 10.8 7.2 5.1 1.5 7.7 0 10.8 2.1 2.1 ...
##  $ winddirection         : int  160 140 110 170 100 250 0 140 170 280 ...
##  $ maxwindspeed          : num  2.1 13.9 9.8 7.2 2.1 13.4 0 13.9 3.6 3.1 ...
##  $ asthma                : chr  "Yes" "Yes" "Yes" "Yes" ...
```

```r
sapply(data, class)
```

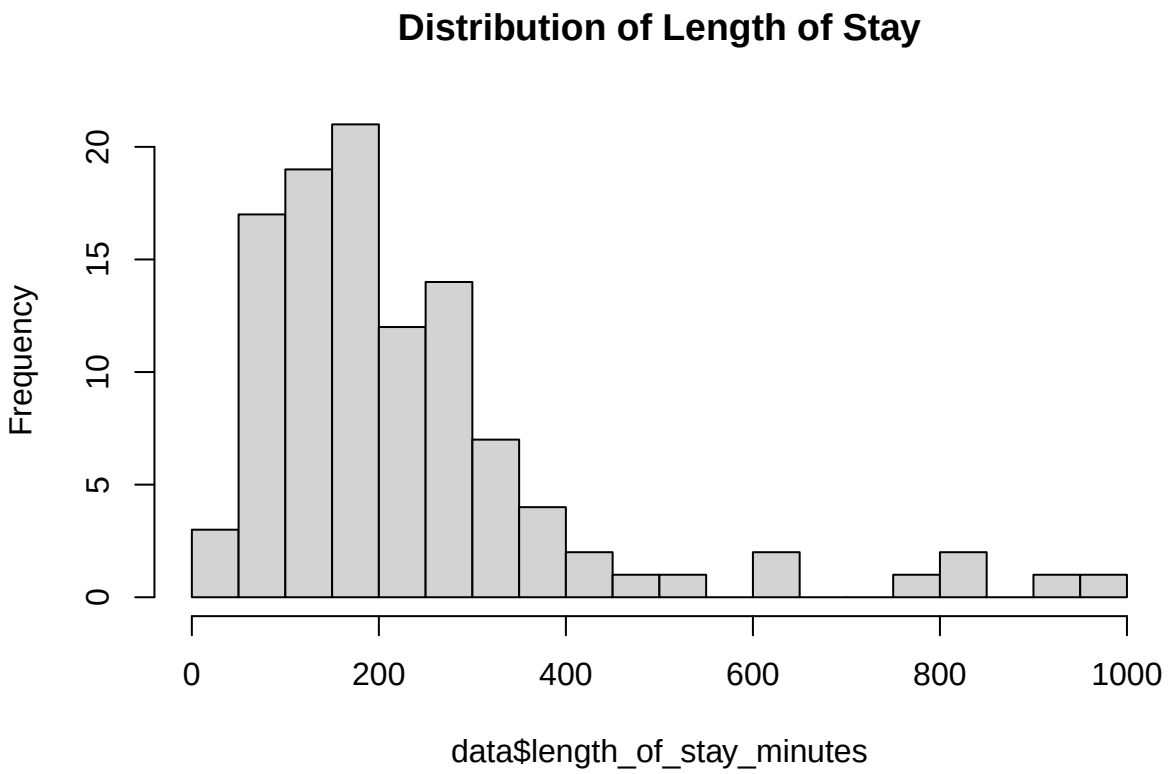
```
##             patient_id                 triage length_of_stay_minutes
##            "character"              "character"                "integer"
##               postcode                    age                   gender
##              "integer"              "character"              "character"
##                 suburb                     co                       o3
##            "character"              "numeric"                "integer"
##                    no2                    so2                    ppm10
##              "integer"              "integer"                "numeric"
##   visibility_reduction                    aqi           precipitation
##              "numeric"              "integer"                "numeric"
##       relativehumidity          vapourpressure                windspeed
##              "integer"              "numeric"                "numeric"
##          winddirection            maxwindspeed                   asthma
##              "integer"              "numeric"              "character"
```

```r
summary(data)
```

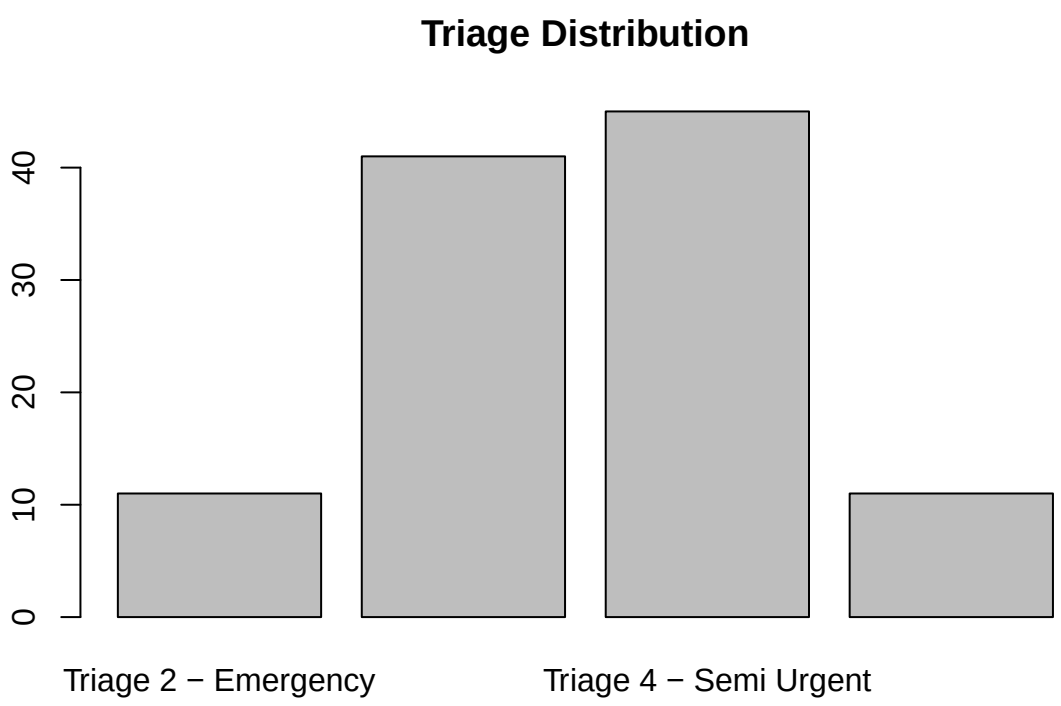
```
##   patient_id           triage          length_of_stay_minutes    postcode
##  Length:108         Length:108         Min.   : 37.0          Min.   :3013
```

```
##  Class :character   Class :character   1st Qu.:123.2          1st Qu.:3037
##  Mode  :character   Mode  :character   Median :181.0          Median :3079
##                                        Mean   :234.0          Mean   :3154
##                                        3rd Qu.:277.8          3rd Qu.:3214
##                                        Max.   :979.0          Max.   :3840
##      age                gender             suburb                co
##  Length:108         Length:108         Length:108         Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.1000
##  Mode  :character   Mode  :character   Mode  :character   Median :0.2000
##                                                           Mean   :0.2241
##                                                           3rd Qu.:0.3000
##                                                           Max.   :1.2000
##        o3              no2               so2             ppm10
##  Min.   : 0.00   Min.   : 0.000   Min.   :0.0000   Min.   : 2.90
##  1st Qu.:12.00   1st Qu.: 3.750   1st Qu.:0.0000   1st Qu.:10.60
##  Median :18.00   Median : 6.000   Median :0.0000   Median :15.75
##  Mean   :18.87   Mean   : 8.343   Mean   :0.7407   Mean   :19.26
##  3rd Qu.:23.25   3rd Qu.:12.000   3rd Qu.:1.0000   3rd Qu.:22.38
##  Max.   :60.00   Max.   :28.000   Max.   :7.0000   Max.   :70.20
##  visibility_reduction      aqi         precipitation     relativehumidity
##  Min.   :0.2200       Min.   :11.00   Min.   : 0.0000   Min.   : 17.00
##  1st Qu.:0.3400       1st Qu.:19.00   1st Qu.: 0.0000   1st Qu.: 50.75
##  Median :0.4000       Median :25.00   Median : 0.0000   Median : 67.50
##  Mean   :0.5266       Mean   :29.89   Mean   : 0.7037   Mean   : 66.04
##  3rd Qu.:0.6300       3rd Qu.:35.50   3rd Qu.: 0.2000   3rd Qu.: 79.25
##  Max.   :1.7300       Max.   :88.00   Max.   :29.0000   Max.   :100.00
##  vapourpressure     windspeed      winddirection    maxwindspeed
##  Min.   : 5.600   Min.   : 0.000   Min.   :  0.0   Min.   : 0.00
##  1st Qu.: 9.825   1st Qu.: 2.100   1st Qu.:110.0   1st Qu.: 3.10
##  Median :11.800   Median : 3.100   Median :170.0   Median : 5.10
##  Mean   :12.604   Mean   : 3.863   Mean   :169.4   Mean   : 5.86
##  3rd Qu.:14.525   3rd Qu.: 5.700   3rd Qu.:240.0   3rd Qu.: 8.20
##  Max.   :25.000   Max.   :13.400   Max.   :360.0   Max.   :18.00
##     asthma
##  Length:108
##  Class :character
##  Mode  :character
##
##
##
```

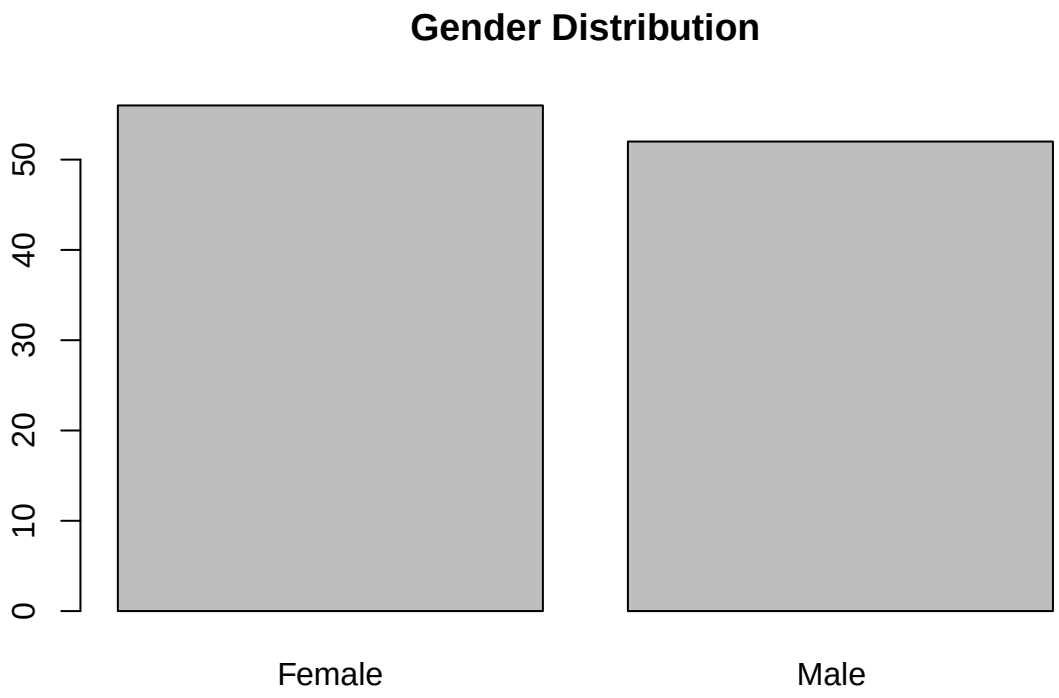
```r
# Histogram for 'length_of_stay_minutes'
hist(data$length_of_stay_minutes, breaks = 20, main = "Distribution of Length of Stay")
```




```r
# Bar chart for 'triage'
barplot(table(data$triage), main = "Triage Distribution")
```




```
# Bar chart for 'gender'
barplot(table(data$gender), main = "Gender Distribution")
```




```
# Cross-tabulation of 'gender' and 'asthma'
cross_table <- table(data$gender, data$asthma)
print(cross_table)
```

```
##
##          No Yes
##   Female 34  22
##   Male   32  20
```

Appendix B

Supervised Learning : Multiple Linear Regression

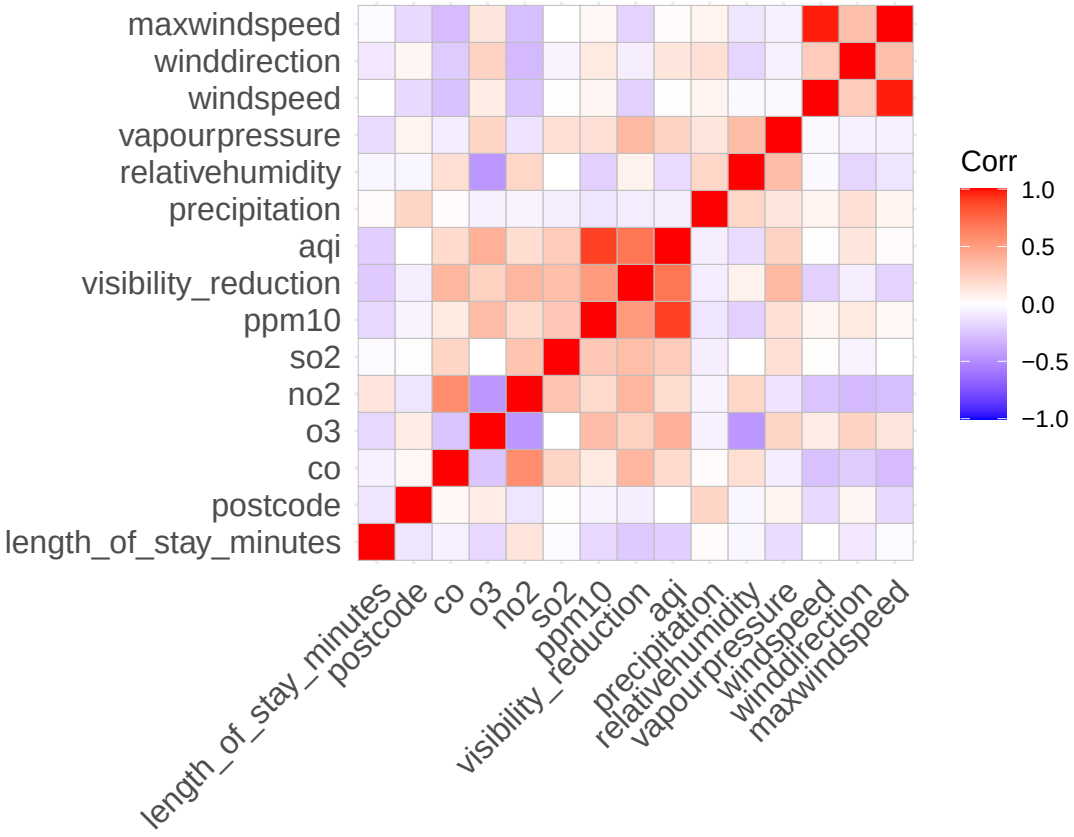
Research Question: What is the combined influence of air quality parameters (CO, O3, NO2, SO2, PPM10, visibility reduction, AQI), weather conditions (precipitation, relative humidity, vapor pressure, windspeed, wind direction, max windspeed) on the length of stay for patients?

```
data_new <- data[, sapply(data, is.numeric)]
str(data_new)
```

```
## 'data.frame':    108 obs. of  15 variables:
##  $ length_of_stay_minutes: int  166 979 38 184 37 372 392 258 181 97 ...
##  $ postcode              : int  3030 3030 3030 3753 3036 3085 3095 3211 3094 3073 ...
##  $ co                    : num  0.4 0.1 0.2 0.2 0.6 0.2 1 0.1 0.2 0.2 ...
##  $ o3                    : int  20 12 11 33 4 15 1 12 4 12 ...
##  $ no2                   : int  13 6 9 2 15 12 14 6 7 6 ...
```

```
##  $ so2                   : int  0 2 1 0 0 0 0 2 0 0 ...
##  $ ppm10                 : num  12.9 4.5 22.3 8.7 17.7 10.6 11.8 4.5 15.1 11.9 ...
##  $ visibility_reduction  : num  0.4 0.27 0.71 0.35 1.4 0.3 0.58 0.27 0.41 0.38 ...
##  $ aqi                   : int  20 12 30 33 60 15 25 12 19 16 ...
##  $ precipitation         : num  0 0 0 0 0 0 0.2 0 0 8.4 ...
##  $ relativehumidity      : int  76 77 74 17 86 73 99 77 79 99 ...
##  $ vapourpressure        : num  11.2 14.4 13.8 11.6 11.9 11.6 11.1 14.4 13.3 11.2 ...
##  $ windspeed             : num  1.5 10.8 7.2 5.1 1.5 7.7 0 10.8 2.1 2.1 ...
##  $ winddirection         : int  160 140 110 170 100 250 0 140 170 280 ...
##  $ maxwindspeed          : num  2.1 13.9 9.8 7.2 2.1 13.4 0 13.9 3.6 3.1 ...
```

```
cor_data_new <- cor(data_new)
ggcorrplot(cor_data_new)
```




```
cor_data_new
```

```
##                         length_of_stay_minutes     postcode          co
## length_of_stay_minutes             1.000000000 -0.106768257 -0.06495017
## postcode                          -0.106768257  1.000000000  0.03666563
## co                                -0.064950169  0.036665634  1.00000000
## o3                                -0.173150980  0.097963301 -0.24783822
## no2                                0.144415271 -0.105438799  0.58409547
## so2                               -0.021240363  0.006343169  0.22155629
## ppm10                             -0.168292838 -0.046020396  0.11051891
## visibility_reduction              -0.228880954 -0.066495813  0.38275962
## aqi                               -0.209319954  0.002352980  0.18854575
## precipitation                      0.022768498  0.215523118  0.01721102
```

```
## relativehumidity                -0.038884939 -0.042328221  0.16816804
## vapourpressure                  -0.145722544  0.056428189 -0.07744400
## windspeed                        0.001159248 -0.156200350 -0.26718361
## winddirection                   -0.098696351  0.046377123 -0.21634202
## maxwindspeed                    -0.022590026 -0.156867756 -0.28850380
##                                          o3          no2          so2       ppm10
## length_of_stay_minutes -0.173150980  0.14441527 -0.021240363 -0.16829284
## postcode                0.097963301 -0.10543880  0.006343169 -0.04602040
## co                     -0.247838218  0.58409547  0.221556292  0.11051891
## o3                      1.000000000 -0.44595830 -0.002552239  0.34864818
## no2                    -0.445958299  1.00000000  0.305434804  0.19459509
## so2                    -0.002552239  0.30543480  1.000000000  0.29398168
## ppm10                   0.348648180  0.19459509  0.293981676  1.00000000
## visibility_reduction    0.235205836  0.38116369  0.344134718  0.52427910
## aqi                     0.405070366  0.18095495  0.267316535  0.89113231
## precipitation          -0.055233274 -0.04914091 -0.069091631 -0.10750789
## relativehumidity       -0.447182708  0.20830090  0.004990065 -0.20138462
## vapourpressure          0.222427627 -0.12111837  0.174702769  0.15506876
## windspeed               0.102201364 -0.25105120  0.008340053  0.04517993
## winddirection           0.232471491 -0.30446415 -0.049739423  0.10612328
## maxwindspeed            0.126413436 -0.27686370  0.003399990  0.04364915
##                        visibility_reduction         aqi precipitation
## length_of_stay_minutes          -0.22888095 -0.20931995    0.02276850
## postcode                        -0.06649581  0.00235298    0.21552312
## co                               0.38275962  0.18854575    0.01721102
## o3                               0.23520584  0.40507037   -0.05523327
## no2                              0.38116369  0.18095495   -0.04914091
## so2                              0.34413472  0.26731654   -0.06909163
## ppm10                            0.52427910  0.89113231   -0.10750789
## visibility_reduction             1.00000000  0.69347167   -0.07773575
## aqi                              0.69347167  1.00000000   -0.06902767
## precipitation                   -0.07773575 -0.06902767    1.00000000
## relativehumidity                 0.06869811 -0.15259472    0.21009309
## vapourpressure                   0.36565242  0.23262894    0.12900055
## windspeed                       -0.19532582  0.01036262    0.06171287
## winddirection                   -0.07283250  0.12531838    0.16260790
## maxwindspeed                    -0.19005270  0.02119570    0.05850602
##                        relativehumidity vapourpressure    windspeed
## length_of_stay_minutes     -0.038884939    -0.14572254  0.001159248
## postcode                   -0.042328221     0.05642819 -0.156200350
## co                          0.168168044    -0.07744400 -0.267183608
## o3                         -0.447182708     0.22242763  0.102201364
## no2                         0.208300899    -0.12111837 -0.251051195
## so2                         0.004990065     0.17470277  0.008340053
## ppm10                      -0.201384619     0.15506876  0.045179926
## visibility_reduction        0.068698107     0.36565242 -0.195325824
## aqi                        -0.152594720     0.23262894  0.010362625
## precipitation               0.210093088     0.12900055  0.061712872
## relativehumidity            1.000000000     0.35112901 -0.025790829
## vapourpressure              0.351129005     1.00000000 -0.032403077
## windspeed                  -0.025790829    -0.03240308  1.000000000
## winddirection              -0.180438042    -0.06313943  0.268901636
## maxwindspeed               -0.107068803    -0.05751113  0.971359565
##                        winddirection maxwindspeed
```

```
## length_of_stay_minutes   -0.09869635  -0.02259003
## postcode                  0.04637712  -0.15686776
## co                       -0.21634202  -0.28850380
## o3                        0.23247149   0.12641344
## no2                      -0.30446415  -0.27686370
## so2                      -0.04973942   0.00339999
## ppm10                     0.10612328   0.04364915
## visibility_reduction     -0.07283250  -0.19005270
## aqi                       0.12531838   0.02119570
## precipitation             0.16260790   0.05850602
## relativehumidity         -0.18043804  -0.10706880
## vapourpressure           -0.06313943  -0.05751113
## windspeed                 0.26890164   0.97135956
## winddirection             1.00000000   0.32512831
## maxwindspeed              0.32512831   1.00000000
```

Drop ppm10 because strong correlation with aqi.

Drop maxwindspeed because strong correlation with windspeed.

```
model1 <- lm(length_of_stay_minutes ~ postcode + co + o3 + no2 + so2
             + visibility_reduction + aqi + precipitation + relativehumidity
             + vapourpressure + windspeed + winddirection, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + o3 + no2 +
##     so2 + visibility_reduction + aqi + precipitation + relativehumidity +
##     vapourpressure + windspeed + winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -257.03  -98.56  -42.54   42.22  695.46
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            664.14630  318.94280   2.082   0.0400 *
## postcode                -0.09744    0.09350  -1.042   0.3000
## co                    -110.97013   99.57023  -1.114   0.2679
## o3                       0.18389    2.42036   0.076   0.9396
## no2                      9.99191    4.32751   2.309   0.0231 *
## so2                      2.55109   15.74024   0.162   0.8716
## visibility_reduction -167.05947  101.23384  -1.650   0.1022
## aqi                    -0.88205    1.65785  -0.532   0.5959
## precipitation           4.08728    5.93697   0.688   0.4928
## relativehumidity       -1.06282    1.16854  -0.910   0.3654
## vapourpressure          1.32790    5.45105   0.244   0.8081
## windspeed              -0.67520    6.90474  -0.098   0.9223
## winddirection          -0.11318    0.20491  -0.552   0.5820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 177.6 on 95 degrees of freedom
## Multiple R-squared:  0.1534, Adjusted R-squared:  0.04649
## F-statistic: 1.435 on 12 and 95 DF,  p-value: 0.1639
```

```r
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: length_of_stay_minutes
##                       Df  Sum Sq Mean Sq F value   Pr(>F)
## postcode               1   40359   40359  1.2792 0.260893
## co                     1   13207   13207  0.4186 0.519193
## o3                     1  121076  121076  3.8376 0.053048 .
## no2                    1   82919   82919  2.6282 0.108296
## so2                    1    5340    5340  0.1693 0.681692
## visibility_reduction   1  226701  226701  7.1855 0.008665 **
## aqi                    1   10612   10612  0.3364 0.563310
## precipitation          1    5723    5723  0.1814 0.671150
## relativehumidity       1   22862   22862  0.7246 0.396772
## vapourpressure         1    3606    3606  0.1143 0.736039
## windspeed              1    1166    1166  0.0370 0.847964
## winddirection          1    9625    9625  0.3051 0.582014
## Residuals             95 2997252   31550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Drop o3

```r
model2 <- lm(length_of_stay_minutes ~ postcode + co + no2 + so2 + visibility_reduction
             + aqi + precipitation + relativehumidity + vapourpressure + windspeed
             + winddirection, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + so2 +
##     visibility_reduction + aqi + precipitation + relativehumidity +
##     vapourpressure + windspeed + winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -257.51  -97.76  -43.57   40.99  695.05
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           667.11366  314.89898   2.119   0.0367 *
## postcode               -0.09696    0.09280  -1.045   0.2987
## co                   -111.31239   98.95186  -1.125   0.2634
## no2                    9.85395    3.90770   2.522   0.0133 *
## so2                    2.49571   15.64172   0.160   0.8736
## visibility_reduction -165.45527   98.49335  -1.680   0.0962 .
## aqi                    -0.85278    1.60411  -0.532   0.5962
## precipitation           4.08911    5.90610   0.692   0.4904
## relativehumidity       -1.10137    1.04715  -1.052   0.2955
## vapourpressure          1.40564    5.32635   0.264   0.7924
## windspeed              -0.65526    6.86393  -0.095   0.9241
## winddirection          -0.11298    0.20383  -0.554   0.5807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 176.7 on 96 degrees of freedom
## Multiple R-squared:  0.1534, Adjusted R-squared:  0.05637
## F-statistic: 1.581 on 11 and 96 DF,  p-value: 0.1166
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: length_of_stay_minutes
##                       Df  Sum Sq Mean Sq F value   Pr(>F)
## postcode               1   40359   40359  1.2926 0.258399
## co                     1   13207   13207  0.4230 0.517003
## no2                    1  157502  157502  5.0444 0.026999 *
## so2                    1   11135   11135  0.3566 0.551788
## visibility_reduction   1  264467  264467  8.4702 0.004487 **
## aqi                    1    6286    6286  0.2013 0.654668
## precipitation          1    4826    4826  0.1546 0.695073
## relativehumidity       1   30518   30518  0.9774 0.325324
## vapourpressure         1    3988    3988  0.1277 0.721600
## windspeed              1    1134    1134  0.0363 0.849252
## winddirection          1    9592    9592  0.3072 0.580684
## Residuals             96 2997434   31223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Drop windspeed

```
model3 <- lm(length_of_stay_minutes ~ postcode + co + no2 + so2 + visibility_reduction
             + aqi + precipitation + relativehumidity + vapourpressure
             + winddirection, data = data)
summary(model3)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + so2 +
##     visibility_reduction + aqi + precipitation + relativehumidity +
##     vapourpressure + winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255.60  -99.27  -43.01   39.92  691.15
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           659.04534  301.79196   2.184   0.0314 *
## postcode               -0.09509    0.09025  -1.054   0.2947
## co                   -110.28410   97.86015  -1.127   0.2625
## no2                     9.89936    3.85878   2.565   0.0118 *
## so2                     2.28363   15.40387   0.148   0.8825
## visibility_reduction -163.71971   96.30526  -1.700   0.0923 .
## aqi                    -0.87957    1.57130  -0.560   0.5769
## precipitation           4.05362    5.86420   0.691   0.4911
## relativehumidity       -1.11180    1.03611  -1.073   0.2859
## vapourpressure          1.43314    5.29132   0.271   0.7871
## winddirection          -0.11622    0.19995  -0.581   0.5624
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 175.8 on 97 degrees of freedom
## Multiple R-squared:  0.1533, Adjusted R-squared:  0.066
## F-statistic: 1.756 on 10 and 97 DF,  p-value: 0.07912
```

```
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: length_of_stay_minutes
##                       Df  Sum Sq Mean Sq F value   Pr(>F)
## postcode               1   40359   40359  1.3059 0.255944
## co                     1   13207   13207  0.4274 0.514837
## no2                    1  157502  157502  5.0964 0.026217 *
## so2                    1   11135   11135  0.3603 0.549732
## visibility_reduction   1  264467  264467  8.5576 0.004285 **
## aqi                    1    6286    6286  0.2034 0.652998
## precipitation          1    4826    4826  0.1562 0.693575
## relativehumidity       1   30518   30518  0.9875 0.322827
## vapourpressure         1    3988    3988  0.1290 0.720220
## winddirection          1   10442   10442  0.3379 0.562407
## Residuals             97 2997718   30904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Drop so2

```
model4 <- lm(length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction + aqi
             + precipitation + relativehumidity + vapourpressure + winddirection, data = data)
summary(model4)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     aqi + precipitation + relativehumidity + vapourpressure +
##     winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -256.92  -99.01  -42.94   39.35  694.83
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           656.07997  299.62199   2.190  0.03092 *
## postcode               -0.09456    0.08973  -1.054  0.29457
## co                   -109.74535   97.30345  -1.128  0.26213
## no2                    10.03112    3.73625   2.685  0.00852 **
## visibility_reduction -162.65139   95.55487  -1.702  0.09189 .
## aqi                    -0.87073    1.56231  -0.557  0.57857
## precipitation           3.99748    5.82269   0.687  0.49400
## relativehumidity       -1.12763    1.02543  -1.100  0.27417
## vapourpressure          1.57167    5.18211   0.303  0.76231
## winddirection          -0.11468    0.19868  -0.577  0.56512
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 174.9 on 98 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.07533
## F-statistic: 1.968 on 9 and 98 DF,  p-value: 0.05106
```

Drop vapourpressure

```
model5 <- lm(length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction + aqi
             + precipitation + relativehumidity + winddirection, data = data)
summary(model5)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     aqi + precipitation + relativehumidity + winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -260.50  -97.95  -45.63   42.40  698.21
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           658.99880  298.09088   2.211   0.0294 *
## postcode               -0.09232    0.08901  -1.037   0.3022
## co                   -115.23535   95.16543  -1.211   0.2288
## no2                    9.69429    3.55097   2.730   0.0075 **
## visibility_reduction -151.75442   88.13569  -1.722   0.0882 .
## aqi                    -0.84258    1.55238  -0.543   0.5885
## precipitation           4.17168    5.76766   0.723   0.4712
## relativehumidity       -0.99772    0.92740  -1.076   0.2846
## winddirection          -0.12303    0.19585  -0.628   0.5313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 174.1 on 99 degrees of freedom
## Multiple R-squared:  0.1523, Adjusted R-squared:  0.08381
## F-statistic: 2.223 on 8 and 99 DF,  p-value: 0.03178
```

Drop aqi

```
model6 <- lm(length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction
             + precipitation + relativehumidity + winddirection, data = data)
summary(model6)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     precipitation + relativehumidity + winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -257.16  -96.24  -45.27   39.66  702.81
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           655.62680  296.97313   2.208  0.02955 *
## postcode               -0.09512    0.08855  -1.074  0.28531
## co                   -113.03574   94.74316  -1.193  0.23566
## no2                    9.67963    3.53832   2.736  0.00737 **
## visibility_reduction -185.03611   63.08449  -2.933  0.00416 **
## precipitation           4.20464    5.74696   0.732  0.46611
## relativehumidity       -0.88292    0.89977  -0.981  0.32883
## winddirection          -0.14321    0.19161  -0.747  0.45657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.5 on 100 degrees of freedom
## Multiple R-squared:  0.1498, Adjusted R-squared:  0.09027
## F-statistic: 2.517 on 7 and 100 DF,  p-value: 0.01999
```

Drop precipitation

```
model7 <- lm(length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction
             + relativehumidity + winddirection, data = data)
summary(model7)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     relativehumidity + winddirection, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -254.38 -100.61  -44.50   47.75  699.89
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           600.21749  286.49253   2.095  0.03867 *
## postcode               -0.08116    0.08627  -0.941  0.34905
## co                   -110.02782   94.43591  -1.165  0.24672
## no2                    9.63723    3.52970   2.730  0.00747 **
## visibility_reduction -188.56619   62.75483  -3.005  0.00335 **
## relativehumidity       -0.71324    0.86736  -0.822  0.41284
## winddirection          -0.11553    0.18741  -0.616  0.53898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173.1 on 101 degrees of freedom
## Multiple R-squared:  0.1452, Adjusted R-squared:  0.09446
## F-statistic:  2.86 on 6 and 101 DF,  p-value: 0.01293
```

Drop winddirection

```
model8 <- lm(length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction
             + relativehumidity, data = data)
summary(model8)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     relativehumidity, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -245.54  -99.95  -36.82   51.25  703.33
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           576.60187  283.05562   2.037  0.04423 *
## postcode               -0.08230    0.08599  -0.957  0.34077
## co                   -106.71929   93.99631  -1.135  0.25889
## no2                    10.09695    3.43952   2.936  0.00411 **
## visibility_reduction -190.87628   62.45218  -3.056  0.00286 **
## relativehumidity       -0.64846    0.85835  -0.755  0.45171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172.6 on 102 degrees of freedom
## Multiple R-squared:  0.142,  Adjusted R-squared:  0.09996
## F-statistic: 3.377 on 5 and 102 DF,  p-value: 0.007288
```

Drop relativehumidity

```
model9 <- lm(length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction, data = data)
summary(model9)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -258.01 -101.01  -36.50   51.78  695.38
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           530.97551  275.96066   1.924  0.05710 .
## postcode               -0.08037    0.08577  -0.937  0.35096
## co                   -111.39977   93.59626  -1.190  0.23670
## no2                    9.74807    3.40127   2.866  0.00504 **
## visibility_reduction -189.62154   62.29987  -3.044  0.00297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172.2 on 103 degrees of freedom
## Multiple R-squared:  0.1372, Adjusted R-squared:  0.1037
## F-statistic: 4.095 on 4 and 103 DF,  p-value: 0.004033
```

Drop postcode

```
model10 <- lm(length_of_stay_minutes ~  co + no2 + visibility_reduction, data = data)
summary(model10)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ co + no2 + visibility_reduction,
```

```
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.34  -95.59  -41.33   49.68  705.79
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           274.607     35.924   7.644 1.09e-11 ***
## co                   -122.879     92.736  -1.325  0.18806
## no2                    10.200      3.365   3.031  0.00307 **
## visibility_reduction -186.335     62.165  -2.997  0.00341 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172.1 on 104 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1048
## F-statistic: 5.174 on 3 and 104 DF,  p-value: 0.00226
```

```
anova(model10)
```

```
## Analysis of Variance Table
##
## Response: length_of_stay_minutes
##                       Df  Sum Sq Mean Sq F value   Pr(>F)
## co                     1   14935   14935  0.5042 0.479248
## no2                    1  178692  178692  6.0324 0.015702 *
## visibility_reduction   1  266145  266145  8.9847 0.003405 **
## Residuals            104 3080676   29622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Drop co

```
model11 <- lm(length_of_stay_minutes ~ no2 + visibility_reduction, data = data)
summary(model11)
```

```
##
## Call:
## lm(formula = length_of_stay_minutes ~ no2 + visibility_reduction,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -245.49 -103.79  -39.18   43.11  711.17
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           275.412     36.048   7.640 1.07e-11 ***
## no2                     7.912      2.899   2.730  0.00744 **
## visibility_reduction -203.912     60.951  -3.346  0.00114 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172.7 on 105 degrees of freedom
## Multiple R-squared:  0.1152, Adjusted R-squared:  0.09832
```

```
## F-statistic: 6.834 on 2 and 105 DF,  p-value: 0.001622
```

```
anova(model11)
```

```
## Analysis of Variance Table
##
## Response: length_of_stay_minutes
##                        Df  Sum Sq Mean Sq F value  Pr(>F)
## no2                     1   73839   73839  2.4749 0.11869
## visibility_reduction    1  333926  333926 11.1924 0.00114 **
## Residuals             105 3132684   29835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.95,2,105)
```

```
## [1] 3.082852
```

```
step(lm(length_of_stay_minutes ~ postcode + co + no2 + so2 + visibility_reduction + aqi
        + precipitation + relativehumidity + vapourpressure + windspeed + winddirection,
        data = data), direction = "backward")
```

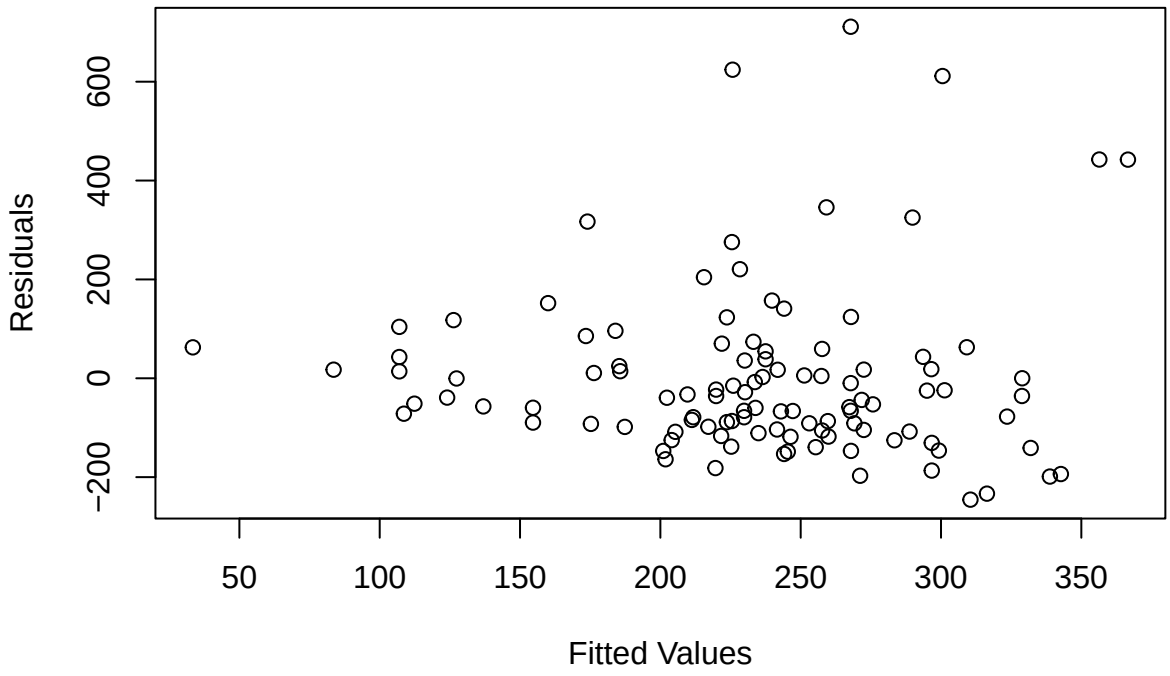
```
## Start:  AIC=1128.96
## length_of_stay_minutes ~ postcode + co + no2 + so2 + visibility_reduction +
##     aqi + precipitation + relativehumidity + vapourpressure +
##     windspeed + winddirection
##
##                         Df Sum of Sq     RSS    AIC
## - windspeed              1       285 2997718 1127.0
## - so2                    1       795 2998229 1127.0
## - vapourpressure         1      2175 2999608 1127.0
## - aqi                    1      8824 3006258 1127.3
## - winddirection          1      9592 3007026 1127.3
## - precipitation          1     14967 3012401 1127.5
## - postcode               1     34084 3031517 1128.2
## - relativehumidity       1     34540 3031974 1128.2
## - co                     1     39511 3036945 1128.4
## <none>                               2997434 1129.0
## - visibility_reduction   1     88110 3085544 1130.1
## - no2                    1    198544 3195977 1133.9
##
## Step:  AIC=1126.97
## length_of_stay_minutes ~ postcode + co + no2 + so2 + visibility_reduction +
##     aqi + precipitation + relativehumidity + vapourpressure +
##     winddirection
##
##                         Df Sum of Sq     RSS    AIC
## - so2                    1       679 2998398 1125.0
## - vapourpressure         1      2267 2999985 1125.0
## - aqi                    1      9684 3007402 1125.3
## - winddirection          1     10442 3008160 1125.3
## - precipitation          1     14767 3012485 1125.5
## - postcode               1     34307 3032026 1126.2
## - relativehumidity       1     35585 3033303 1126.2
## - co                     1     39249 3036968 1126.4
## <none>                               2997718 1127.0
```

```
## - visibility_reduction  1     89314 3087033 1128.1
## - no2                   1    203392 3201110 1132.1
##
## Step:  AIC=1125
## length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     aqi + precipitation + relativehumidity + vapourpressure +
##     winddirection
##
##                         Df Sum of Sq     RSS    AIC
## - vapourpressure         1      2814 3001212 1123.1
## - aqi                    1      9504 3007901 1123.3
## - winddirection          1     10194 3008591 1123.4
## - precipitation          1     14421 3012818 1123.5
## - postcode               1     33977 3032374 1124.2
## - relativehumidity       1     36998 3035396 1124.3
## - co                     1     38921 3037318 1124.4
## <none>                               2998398 1125.0
## - visibility_reduction   1     88649 3087046 1126.1
## - no2                    1    220541 3218939 1130.7
##
## Step:  AIC=1123.1
## length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     aqi + precipitation + relativehumidity + winddirection
##
##                         Df Sum of Sq     RSS    AIC
## - aqi                    1      8931 3010143 1121.4
## - winddirection          1     11963 3013175 1121.5
## - precipitation          1     15859 3017071 1121.7
## - postcode               1     32607 3033819 1122.3
## - relativehumidity       1     35087 3036299 1122.3
## - co                     1     44450 3045662 1122.7
## <none>                               3001212 1123.1
## - visibility_reduction   1     89875 3091087 1124.3
## - no2                    1    225943 3227155 1128.9
##
## Step:  AIC=1121.42
## length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     precipitation + relativehumidity + winddirection
##
##                         Df Sum of Sq     RSS    AIC
## - precipitation          1     16113 3026255 1120.0
## - winddirection          1     16815 3026958 1120.0
## - relativehumidity       1     28985 3039127 1120.5
## - postcode               1     34735 3044878 1120.7
## - co                     1     42847 3052990 1121.0
## <none>                               3010143 1121.4
## - no2                    1    225273 3235416 1127.2
## - visibility_reduction   1    258973 3269116 1128.3
##
## Step:  AIC=1120
## length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     relativehumidity + winddirection
##
##                         Df Sum of Sq     RSS    AIC
```

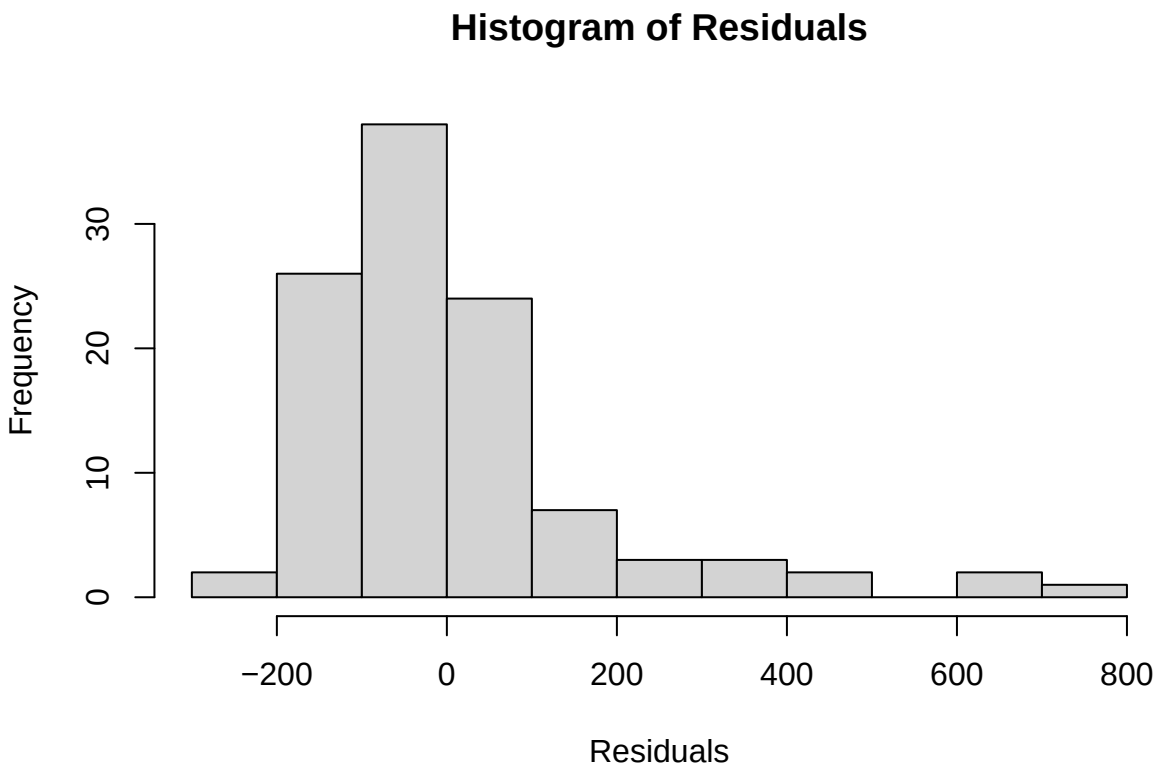
```
## - winddirection         1     11387 3037642 1118.4
## - relativehumidity      1     20261 3046516 1118.7
## - postcode              1     26521 3052776 1118.9
## - co                    1     40674 3066929 1119.4
## <none>                              3026255 1120.0
## - no2                   1    223364 3249620 1125.7
## - visibility_reduction  1    270531 3296787 1127.2
##
## Step:  AIC=1118.4
## length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction +
##     relativehumidity
##
##                         Df Sum of Sq     RSS    AIC
## - relativehumidity       1     16997 3054639 1117.0
## - postcode               1     27282 3064924 1117.4
## - co                     1     38388 3076031 1117.8
## <none>                               3037642 1118.4
## - no2                    1    256638 3294280 1125.2
## - visibility_reduction   1    278192 3315834 1125.9
##
## Step:  AIC=1117
## length_of_stay_minutes ~ postcode + co + no2 + visibility_reduction
##
##                         Df Sum of Sq     RSS    AIC
## - postcode               1     26037 3080676 1115.9
## - co                     1     42012 3096651 1116.5
## <none>                               3054639 1117.0
## - no2                    1    243601 3298240 1123.3
## - visibility_reduction   1    274741 3329380 1124.3
##
## Step:  AIC=1115.92
## length_of_stay_minutes ~ co + no2 + visibility_reduction
##
##                         Df Sum of Sq     RSS    AIC
## - co                     1     52008 3132684 1115.7
## <none>                               3080676 1115.9
## - visibility_reduction   1    266145 3346821 1122.9
## - no2                    1    272172 3352848 1123.1
##
## Step:  AIC=1115.73
## length_of_stay_minutes ~ no2 + visibility_reduction
##
##                         Df Sum of Sq     RSS    AIC
## <none>                               3132684 1115.7
## - no2                    1    222293 3354977 1121.1
## - visibility_reduction   1    333926 3466610 1124.7
##
## Call:
## lm(formula = length_of_stay_minutes ~ no2 + visibility_reduction,
##     data = data)
##
## Coefficients:
##          (Intercept)                   no2  visibility_reduction
```

```
##              275.412                 7.912              -203.912
```

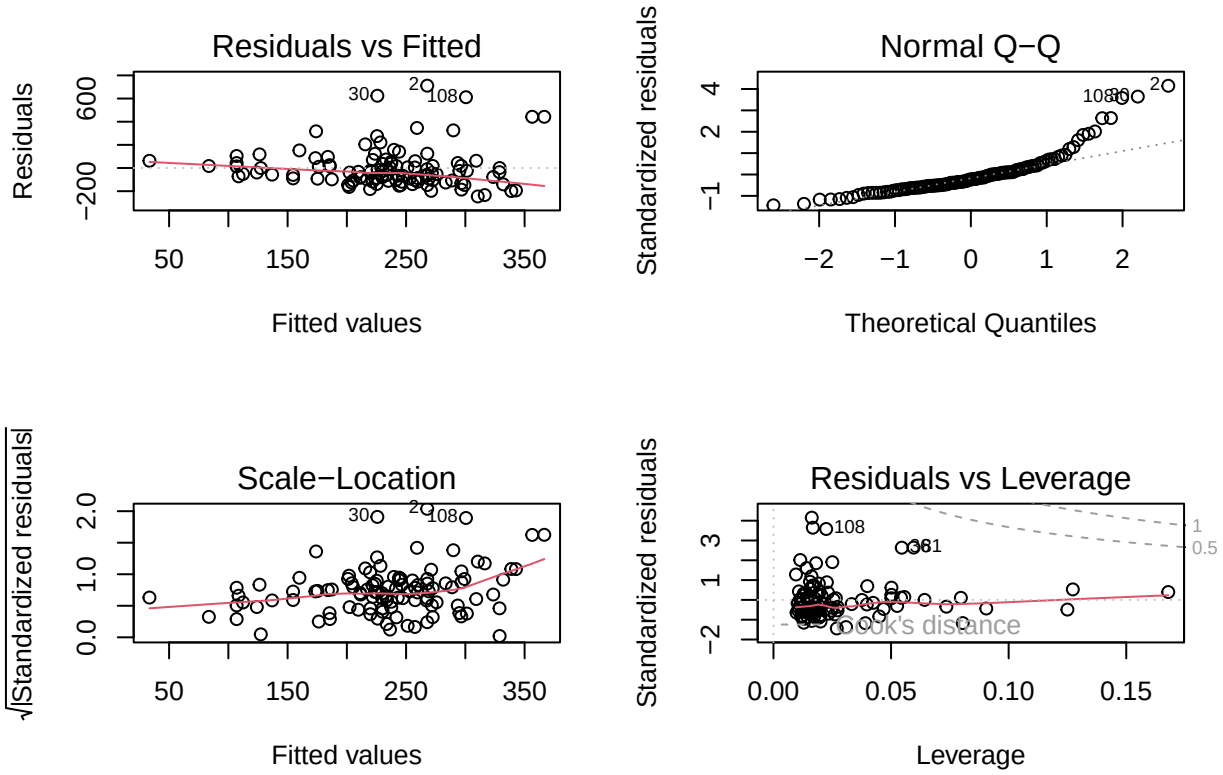
```
plot(predict(model11),resid(model11), xlab = "Fitted Values", ylab = "Residuals")
```




```
hist(resid(model11), main = paste("Histogram of Residuals"), xlab = "Residuals")
```




```
par(mfrow=c(2,2))
plot(model11)
```



Appendix C

Supervised Learning : Multiple Logistic Regression

Research Question: Can we predict the likelihood of asthma attacks based on patient characteristics, environmental factors, and triage information using logistic regression?

```
data$asthma <- as.factor(data$asthma)
str(data)
```

```
## 'data.frame':    108 obs. of  21 variables:
##  $ patient_id            : chr  "PJY1ZY7M" "PTX0ZI2D" "PVX0GR6F" "WOE7QE3M" ...
##  $ triage                : chr  "Triage 3 - Urgent" "Triage 3 - Urgent" "Triage 3 - Urgent" "Triage 3
##  $ length_of_stay_minutes: int  166 979 38 184 37 372 392 258 181 97 ...
##  $ postcode              : int  3030 3030 3030 3753 3036 3085 3095 3211 3094 3073 ...
##  $ age                   : chr  "30 to 34" "00 to 04" "20 to 24" "05 to 09" ...
##  $ gender                : chr  "Female" "Male  " "Male  " "Male  " ...
##  $ suburb                : chr  "Werribee" "Pt Cook" "Werribee South" "Beveridge" ...
##  $ co                    : num  0.4 0.1 0.2 0.2 0.6 0.2 1 0.1 0.2 0.2 ...
##  $ o3                    : int  20 12 11 33 4 15 1 12 4 12 ...
##  $ no2                   : int  13 6 9 2 15 12 14 6 7 6 ...
##  $ so2                   : int  0 2 1 0 0 0 0 2 0 0 ...
##  $ ppm10                 : num  12.9 4.5 22.3 8.7 17.7 10.6 11.8 4.5 15.1 11.9 ...
##  $ visibility_reduction  : num  0.4 0.27 0.71 0.35 1.4 0.3 0.58 0.27 0.41 0.38 ...
##  $ aqi                   : int  20 12 30 33 60 15 25 12 19 16 ...
##  $ precipitation         : num  0 0 0 0 0 0 0.2 0 0 8.4 ...
```

```
##  $ relativehumidity      : int  76 77 74 17 86 73 99 77 79 99 ...
##  $ vapourpressure        : num  11.2 14.4 13.8 11.6 11.9 11.6 11.1 14.4 13.3 11.2 ...
##  $ windspeed             : num  1.5 10.8 7.2 5.1 1.5 7.7 0 10.8 2.1 2.1 ...
##  $ winddirection         : int  160 140 110 170 100 250 0 140 170 280 ...
##  $ maxwindspeed          : num  2.1 13.9 9.8 7.2 2.1 13.4 0 13.9 3.6 3.1 ...
##  $ asthma                : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
```

```
log_model1 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + no2 + so2
                  + visibility_reduction + aqi + precipitation + relativehumidity
                  + vapourpressure + windspeed + winddirection, family = binomial, data = data)
summary(log_model1)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + no2 + so2 + visibility_reduction + aqi + precipitation +
##     relativehumidity + vapourpressure + windspeed + winddirection,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8567  -0.8680  -0.4778   1.0163   2.0598
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.079526   4.258534  -1.428   0.1534
## length_of_stay_minutes  0.001747   0.001315   1.329   0.1840
## postcode                0.001807   0.001236   1.462   0.1437
## co                      1.824463   1.465036   1.245   0.2130
## o3                     -0.068482   0.037044  -1.849   0.0645 .
## no2                    -0.014455   0.058090  -0.249   0.8035
## so2                    -0.496523   0.277888  -1.787   0.0740 .
## visibility_reduction    0.967451   1.533582   0.631   0.5281
## aqi                     0.009000   0.024287   0.371   0.7110
## precipitation           0.429356   0.274752   1.563   0.1181
## relativehumidity        0.009713   0.016447   0.591   0.5548
## vapourpressure         -0.037078   0.079540  -0.466   0.6411
## windspeed               0.091016   0.091949   0.990   0.3222
## winddirection          -0.004376   0.002775  -1.577   0.1148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 111.07  on  94  degrees of freedom
## AIC: 139.07
##
## Number of Fisher Scoring iterations: 6
```

```
roc_curve <- roc(data$asthma, predict(log_model1, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
auc(roc_curve)
```

```
## Area under the curve: 0.7965
```

```r
log_model2 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + so2
                  + visibility_reduction + aqi + precipitation + relativehumidity
                  + vapourpressure + windspeed + winddirection, family = binomial, data = data)
summary(log_model2)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + so2 + visibility_reduction + aqi + precipitation + relativehumidity +
##     vapourpressure + windspeed + winddirection, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8437  -0.8421  -0.4738   1.0077   2.0749
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.295517   4.170439  -1.510   0.1312
## length_of_stay_minutes  0.001664   0.001273   1.306   0.1914
## postcode                0.001833   0.001231   1.489   0.1366
## co                      1.704273   1.379339   1.236   0.2166
## o3                     -0.064458   0.033029  -1.952   0.0510 .
## so2                    -0.504912   0.275545  -1.832   0.0669 .
## visibility_reduction    0.896287   1.498011   0.598   0.5496
## aqi                     0.008011   0.023940   0.335   0.7379
## precipitation           0.429285   0.271849   1.579   0.1143
## relativehumidity        0.009311   0.016317   0.571   0.5683
## vapourpressure         -0.032544   0.077174  -0.422   0.6732
## windspeed               0.093680   0.091421   1.025   0.3055
## winddirection          -0.004286   0.002753  -1.557   0.1196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 111.13  on  95  degrees of freedom
## AIC: 137.13
##
## Number of Fisher Scoring iterations: 6
```

```r
roc_curve <- roc(data$asthma, predict(log_model2, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
auc(roc_curve)
```

```
## Area under the curve: 0.7987
```

```r
glm_prob<-predict(log_model2, type="response")
glm_pred<-rep("No",108)
glm_pred[glm_prob>0.5]="Yes"
table(glm_pred,asthma)
```

```
##         asthma
## glm_pred No Yes
##      No  55  18
##      Yes 11  24
```

```r
Misclassification_Rate <- (18+11)/(55+18+11+24)
Misclassification_Rate
```

```
## [1] 0.2685185
```

```r
False_Positive_Rate <- 11/(55+11)
False_Positive_Rate
```

```
## [1] 0.1666667
```

```r
False_Negative_Rate <- 18/(18+24)
False_Negative_Rate
```

```
## [1] 0.4285714
```

```r
log_model3 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + so2
                  + visibility_reduction + precipitation + relativehumidity
                  + vapourpressure + windspeed + winddirection, family = binomial, data = data)
summary(log_model3)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + so2 + visibility_reduction + precipitation + relativehumidity +
##     vapourpressure + windspeed + winddirection, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7027  -0.8544  -0.4805   1.0102   2.0812
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.322843   4.165055  -1.518   0.1290
## length_of_stay_minutes  0.001648   0.001271   1.297   0.1947
## postcode                0.001856   0.001229   1.510   0.1310
## co                      1.693894   1.375133   1.232   0.2180
## o3                     -0.062088   0.032015  -1.939   0.0525 .
## so2                    -0.496959   0.273070  -1.820   0.0688 .
## visibility_reduction    1.216873   1.153156   1.055   0.2913
## precipitation           0.441737   0.275403   1.604   0.1087
## relativehumidity        0.008518   0.016115   0.529   0.5971
## vapourpressure         -0.032801   0.077654  -0.422   0.6727
## windspeed               0.097476   0.090287   1.080   0.2803
## winddirection          -0.004188   0.002739  -1.529   0.1263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 111.24  on  96  degrees of freedom
## AIC: 135.24
##
## Number of Fisher Scoring iterations: 6
```

```
roc_curve <- roc(data$asthma, predict(log_model3, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
auc(roc_curve)
```

```
## Area under the curve: 0.7937
```

```
log_model4 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + so2
                  + visibility_reduction + precipitation + relativehumidity
                  + windspeed + winddirection, family = binomial, data = data)
summary(log_model4)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + so2 + visibility_reduction + precipitation + relativehumidity +
##     windspeed + winddirection, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7539  -0.8468  -0.4672   1.0337   2.1399
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -6.256852   4.141937  -1.511   0.1309
## length_of_stay_minutes  0.001645   0.001279   1.286   0.1985
## postcode                0.001808   0.001212   1.492   0.1357
## co                      1.832669   1.334519   1.373   0.1697
## o3                     -0.064701   0.031159  -2.076   0.0378 *
## so2                    -0.505856   0.271203  -1.865   0.0621 .
## visibility_reduction    1.050863   1.077538   0.975   0.3294
## precipitation           0.464006   0.280686   1.653   0.0983 .
## relativehumidity        0.005327   0.014137   0.377   0.7063
## windspeed               0.095162   0.089732   1.061   0.2889
## winddirection          -0.004177   0.002736  -1.526   0.1269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 111.42  on  97  degrees of freedom
## AIC: 133.42
##
## Number of Fisher Scoring iterations: 6
```

```r
roc_curve <- roc(data$asthma, predict(log_model4, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
auc(roc_curve)
```

```
## Area under the curve: 0.794
```

```r
log_model5 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + so2
                  + visibility_reduction + precipitation + windspeed
                  + winddirection, family = binomial, data = data)
summary(log_model5)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + so2 + visibility_reduction + precipitation + windspeed +
##     winddirection, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7625  -0.8314  -0.4793   1.0513   2.1527
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -5.875186   4.008291  -1.466   0.1427
## length_of_stay_minutes  0.001610   0.001276   1.261   0.2072
## postcode                0.001819   0.001210   1.503   0.1327
## co                      1.824742   1.334859   1.367   0.1716
## o3                     -0.069600   0.028468  -2.445   0.0145 *
## so2                    -0.515421   0.270926  -1.902   0.0571 .
## visibility_reduction    1.131613   1.057043   1.071   0.2844
## precipitation           0.488806   0.277020   1.765   0.0776 .
## windspeed               0.097772   0.089545   1.092   0.2749
## winddirection          -0.004268   0.002720  -1.569   0.1167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 111.56  on  98  degrees of freedom
## AIC: 131.56
##
## Number of Fisher Scoring iterations: 6
```

```r
roc_curve <- roc(data$asthma, predict(log_model5, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
auc(roc_curve)
```

```
## Area under the curve: 0.7955
```

```r
log_model6 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + so2
                  + visibility_reduction + precipitation + winddirection,
                  family = binomial, data = data)
summary(log_model6)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + so2 + visibility_reduction + precipitation + winddirection,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8620  -0.8219  -0.5118   1.0271   2.0930
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -4.574984   3.794391  -1.206   0.2279
## length_of_stay_minutes   0.001542   0.001268   1.216   0.2241
## postcode                 0.001526   0.001171   1.303   0.1924
## co                       1.552373   1.290026   1.203   0.2288
## o3                      -0.065635   0.027608  -2.377   0.0174 *
## so2                     -0.460756   0.258093  -1.785   0.0742 .
## visibility_reduction     0.879859   1.011892   0.870   0.3846
## precipitation            0.538422   0.295832   1.820   0.0688 .
## winddirection           -0.003619   0.002609  -1.387   0.1653
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 112.76  on  99  degrees of freedom
## AIC: 130.76
##
## Number of Fisher Scoring iterations: 6
```

```r
roc_curve <- roc(data$asthma, predict(log_model6, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
auc(roc_curve)
```

```
## Area under the curve: 0.7908
```

```r
log_model7 <- glm(asthma ~ length_of_stay_minutes + postcode + co + o3 + so2
                  + precipitation + winddirection, family = binomial, data = data)
summary(log_model7)
```

```
##
## Call:
## glm(formula = asthma ~ length_of_stay_minutes + postcode + co +
##     o3 + so2 + precipitation + winddirection, family = binomial,
##     data = data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8508  -0.8381  -0.5131   1.0318   2.0210
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.817491   3.656622  -1.044   0.2965
## length_of_stay_minutes  0.001315   0.001234   1.065   0.2867
## postcode                0.001400   0.001152   1.215   0.2243
## co                      1.887196   1.241910   1.520   0.1286
## o3                     -0.062302   0.028392  -2.194   0.0282 *
## so2                    -0.431845   0.256559  -1.683   0.0923 .
## precipitation           0.522539   0.290653   1.798   0.0722 .
## winddirection          -0.003698   0.002598  -1.423   0.1547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 113.52  on 100  degrees of freedom
## AIC: 129.52
##
## Number of Fisher Scoring iterations: 6
```

```
roc_curve <- roc(data$asthma, predict(log_model7, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
auc(roc_curve)
```

```
## Area under the curve: 0.7868
```

```
log_model8 <- glm(asthma ~ postcode + co + o3 + so2 + precipitation
                  + winddirection, family = binomial, data = data)
summary(log_model8)
```

```
##
## Call:
## glm(formula = asthma ~ postcode + co + o3 + so2 + precipitation +
##     winddirection, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5639  -0.8602  -0.5158   1.0193   2.0243
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.011502   3.562972  -0.845   0.3980
## postcode       0.001280   0.001142   1.120   0.2626
## co             1.659955   1.206918   1.375   0.1690
## o3            -0.064633   0.027996  -2.309   0.0210 *
## so2           -0.404455   0.250477  -1.615   0.1064
## precipitation  0.523457   0.290107   1.804   0.0712 .
```

```
## winddirection -0.003944   0.002581  -1.528   0.1265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 114.65  on 101  degrees of freedom
## AIC: 128.65
##
## Number of Fisher Scoring iterations: 6
```

```
roc_curve <- roc(data$asthma, predict(log_model8, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
auc(roc_curve)
```

```
## Area under the curve: 0.7677
```

```
log_model9 <- glm(asthma ~ co + o3 + so2 + precipitation + winddirection,
                  family = binomial, data = data)
summary(log_model9)
```

```
##
## Call:
## glm(formula = asthma ~ co + o3 + so2 + precipitation + winddirection,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5699  -0.8858  -0.5233   1.0031   1.9526
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.918031   0.712867   1.288   0.1978
## co             1.746150   1.196468   1.459   0.1444
## o3            -0.060243   0.027047  -2.227   0.0259 *
## so2           -0.395722   0.247147  -1.601   0.1093
## precipitation  0.512092   0.290830   1.761   0.0783 .
## winddirection -0.003872   0.002551  -1.518   0.1290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 144.34  on 107  degrees of freedom
## Residual deviance: 115.89  on 102  degrees of freedom
## AIC: 127.89
##
## Number of Fisher Scoring iterations: 6
```

```
roc_curve <- roc(data$asthma, predict(log_model9, type = "response"))
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```
auc(roc_curve)
```

```
## Area under the curve: 0.7543
```

```
list_of_models <- list(log_model1, log_model2, log_model3, log_model4,
                       log_model5, log_model6, log_model7, log_model8, log_model9)

aic_values <- lapply(list_of_models, function(model) {
  AIC(model)
})

for (i in 1:length(aic_values)) {
  cat("Model", i, "AIC =", aic_values[[i]], "\n")
}
```

```
## Model 1 AIC = 139.0665
## Model 2 AIC = 137.1284
## Model 3 AIC = 135.2372
## Model 4 AIC = 133.4179
## Model 5 AIC = 131.5602
## Model 6 AIC = 130.7563
## Model 7 AIC = 129.5179
## Model 8 AIC = 128.6508
## Model 9 AIC = 127.8868
```

```
auc_values <- lapply(list(log_model1, log_model2, log_model3, log_model4,
                          log_model5, log_model6, log_model7, log_model8,
                          log_model9), function(model) {
  roc_obj <- roc(response = data$asthma, predictor = predict(model, type = "response"))
  auc(roc_obj)
})
```

```
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```

```
for (i in 1:length(auc_values)) {
  cat("Model", i, "AUC =", auc_values[[i]], "\n")
}
```

```
## Model 1 AUC = 0.7965368
## Model 2 AUC = 0.7987013
## Model 3 AUC = 0.7936508
## Model 4 AUC = 0.7940115
## Model 5 AUC = 0.7954545
## Model 6 AUC = 0.7907648
## Model 7 AUC = 0.7867965
## Model 8 AUC = 0.7676768
## Model 9 AUC = 0.754329
```

```
auc_values <- lapply(list(log_model1, log_model2, log_model3, log_model4,
                          log_model5, log_model6, log_model7, log_model8,
                          log_model9), function(model) {
  roc_obj <- roc(response = data$asthma, predictor = predict(model, type = "response"))
  auc(roc_obj)
})
```

```
## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases

## Setting levels: control = No, case = Yes

## Setting direction: controls < cases
```

```
model_names <- c("Model 1", "Model 2", "Model 3", "Model 4", "Model 5", "Model 6",
                 "Model 7", "Model 8", "Model 9")
```

```r
best_model_index <- which.max(auc_values)

best_model_name <- model_names[best_model_index]

cat("The best model is:", best_model_name, "\n")
```

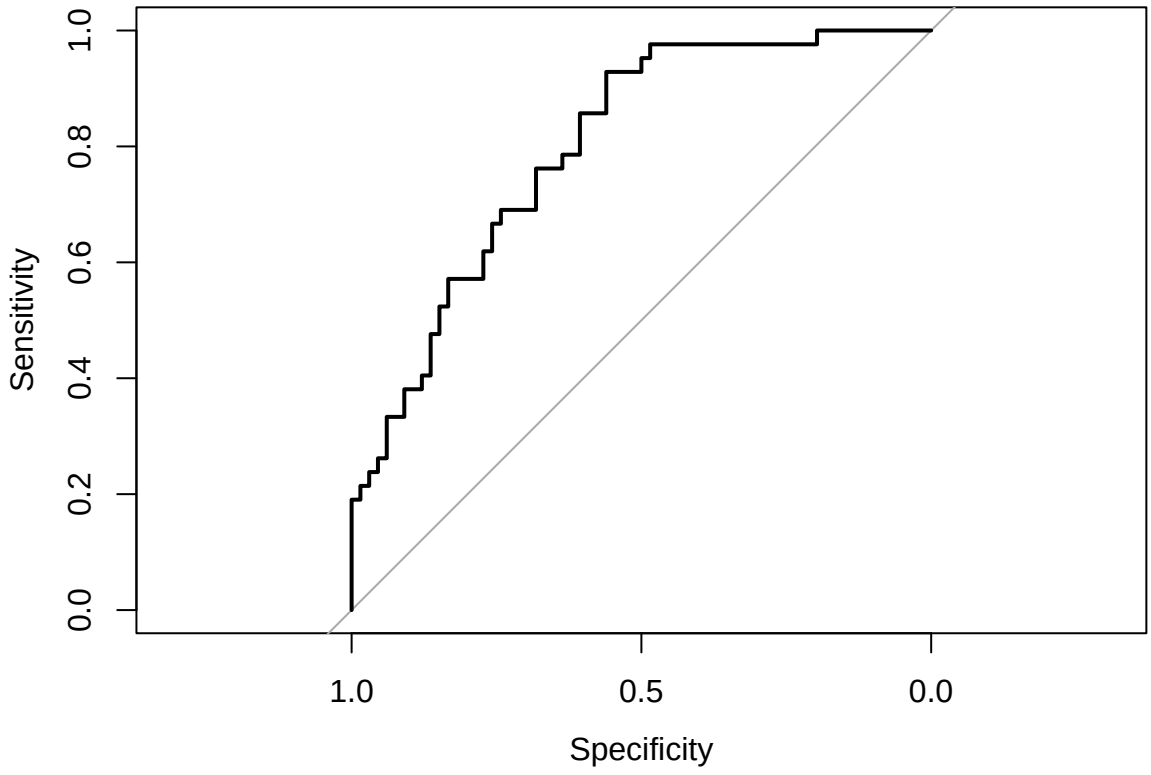
```
## The best model is: Model 2
```

```r
# Assuming you have fitted a logistic regression model (log_model)
library(pROC)
roc_obj <- roc(response = data$asthma, predictor = predict(log_model2, type = "response"))
```

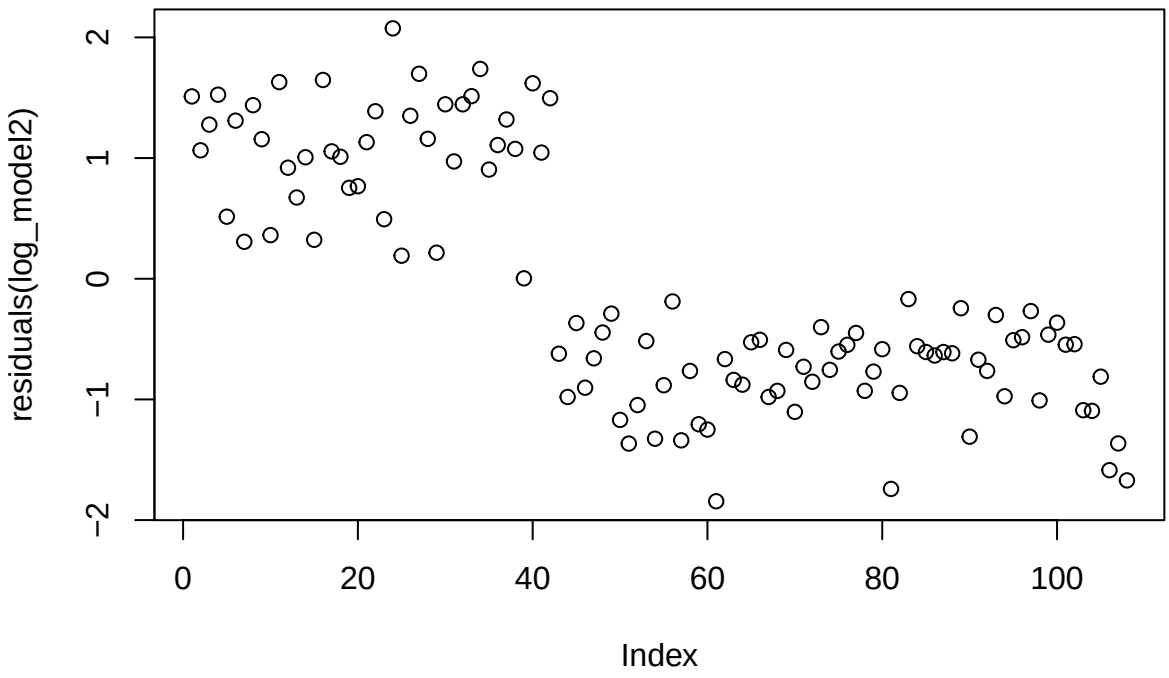
```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
plot(roc_obj)
```




```r
# Assuming you have fitted a logistic regression model (log_model)
plot(residuals(log_model2), type = "p")
```



Appendix D

Unsupervised Learning: K-means Clustering

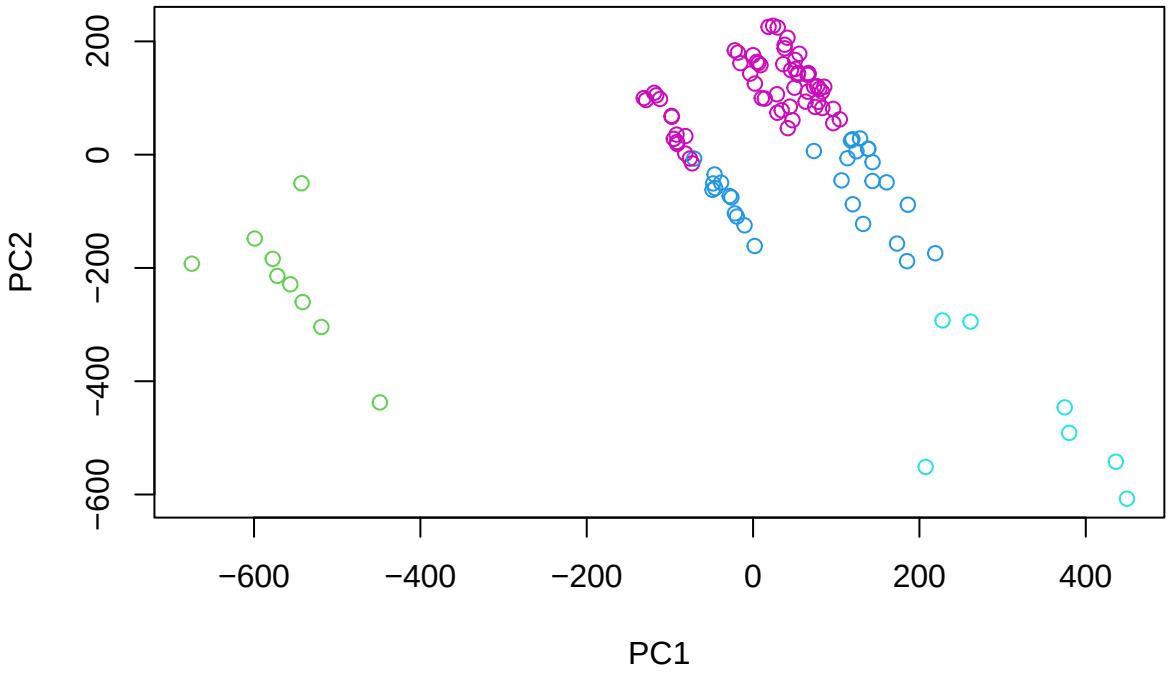
```
wcss = vector()
for (i in 1:10) wcss[i] = sum(kmeans(data_new, i)$withinss)
plot(1:10,
     wcss,
     type = 'b',
     main = paste('The Elbow Method'),
     xlab = 'Number of clusters',
     ylab = 'WCSS')
```


```
km <- kmeans(data_new, centers = 4, nstart = 20)
km
```

```
## K-means clustering with 4 clusters of sizes 9, 30, 7, 62
##
## Cluster means:
##   length_of_stay_minutes postcode        co       o3       no2       so2
## 1               180.3333 3753.222 0.2111111 22.44444  5.000000 0.3333333
## 2               316.6333 3118.267 0.2533333 17.46667  8.566667 0.7666667
## 3               795.5714 3066.286 0.1285714 12.14286 12.142857 0.5714286
## 4               138.4839 3093.516 0.2225806 19.79032  8.290323 0.8064516
##      ppm10 visibility_reduction      aqi precipitation relativehumidity
## 1 14.28889            0.4388889 28.11111     3.2666667         64.77778
## 2 20.27667            0.4803333 29.53333     0.4600000         66.63333
## 3 11.27143            0.3742857 19.14286     1.2285714         67.28571
## 4 20.39677            0.5788710 31.53226     0.3903226         65.79032
##   vapourpressure windspeed winddirection maxwindspeed
## 1       13.54444  3.244444      193.3333     4.911111
## 2       11.85667  4.050000      132.3333     6.170000
## 3       11.37143  3.757143      141.4286     5.071429
## 4       12.96774  3.874194      186.9355     5.937097
##
## Clustering vector:
##   [1] 4 3 4 1 4 2 2 2 4 4 2 4 1 2 4 4 4 4 4 2 2 4 2 4 3 2 4 4 2 3 4 4 2 4 2 3 2
##  [38] 4 1 1 1 1 4 4 2 4 4 4 4 4 4 2 4 4 4 4 4 4 2 4 2 4 2 4 4 4 4 4 2 4 2 2 2 4 4 4 2
##  [75] 2 4 4 2 4 2 3 2 4 4 4 4 4 4 2 4 4 4 1 4 3 4 2 4 2 4 4 4 4 4 4 4 4 1 1 3
##
```

```
## Within cluster sum of squares by cluster:
## [1]  191247.4  697414.8  198976.9 1048095.4
##  (between_SS / total_SS =  75.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
pp = prcomp(data_new)
plot(pp$x[,1:2], col=fitted(km, "classes")+2)
```



Variables	Description
patient_id	Patient ID
triage	Triage
postcode	Postcode of the admitted patients
age	Age at the time of admission
gender	Gender
suburb	Suburb of the admitted patients
co	Hourly records of atmospheric concentration of Carbon monoxide in parts per million (ppm)
o3	Hourly records of atmospheric concentration of ground level Ozone in parts per billion (ppb)
no2	Hourly records of atmospheric concentration of Nitrogen dioxide in parts per billion (ppb)
so2	Hourly records of atmospheric concentration of Sulphur dioxide in parts per billion (ppb)
ppm10	Hourly records of atmospheric concentration of particulate matter less than 10µm in diameter ($\mu g/m^3$)
visibility_reduction	Hourly records of visibility reduction i.e. minimum visible distance – 20km
aqi	Air Quality Index
precipitation	Half-hourly records of rainfall in millimetre (mm)
relativehumidity	Half-hourly records of Relative Humidity (%)
vapourpressure	Half-hourly records of Vapour Pressure
windspeed	Half-hourly records of Wind Speed (km/hr.)
winddirection	Half-hourly records of Wind Direction
maxwindspeed	Half-hourly records of Maximum Wind Speed (km/hr.)
asthma	Binary indicator of asthma attacks