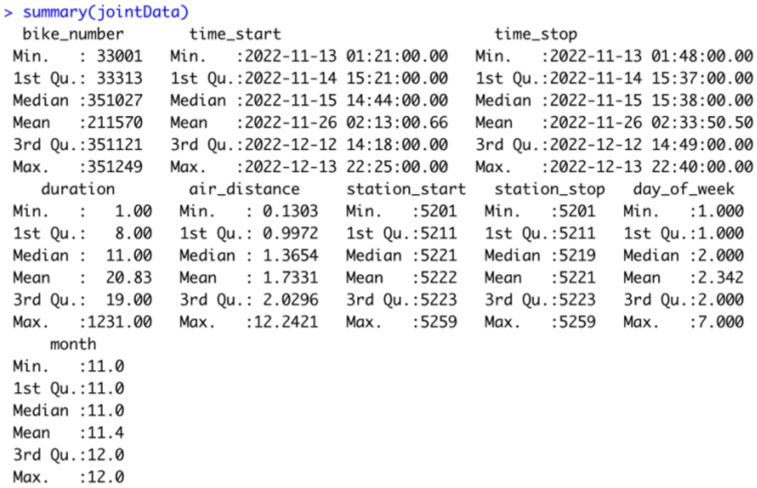
Case Study Master Data Science (Module 3)
Prof. Dr. Heike Trautmann, Prof. Dr. Pascal Kerschke
Submitted by: Rohan Francis Pinto
WWU Münster, Germany
rohanfrancispinto@gmail.com

Exploratory data analysis:

In the dataset, there are 1189 observations and 9 variables. The variables are bike number, time start, time stop, duration, air distance, station start, station stop, day of week, month. The day of the week and month were extracted from the time start variable.

The variables in the dataset consist of integers, numerical values, and time and date formats.

Key information using summary function:



```
> summary(jointData)
  bike_number        time_start                         time_stop                   
 Min.   : 33001   Min.   :2022-11-13 01:21:00.00   Min.   :2022-11-13 01:48:00.00  
 1st Qu.: 33313   1st Qu.:2022-11-14 15:21:00.00   1st Qu.:2022-11-14 15:37:00.00  
 Median :351027   Median :2022-11-15 14:44:00.00   Median :2022-11-15 15:38:00.00  
 Mean   :211570   Mean   :2022-11-26 02:13:00.66   Mean   :2022-11-26 02:33:50.50  
 3rd Qu.:351121   3rd Qu.:2022-12-12 14:18:00.00   3rd Qu.:2022-12-12 14:49:00.00  
 Max.   :351249   Max.   :2022-12-13 22:25:00.00   Max.   :2022-12-13 22:40:00.00  
    duration         air_distance      station_start   station_stop    day_of_week   
 Min.   :   1.00   Min.   : 0.1303   Min.   :5201   Min.   :5201   Min.   :1.000  
 1st Qu.:   8.00   1st Qu.: 0.9972   1st Qu.:5211   1st Qu.:5211   1st Qu.:1.000  
 Median :  11.00   Median : 1.3654   Median :5221   Median :5219   Median :2.000  
 Mean   :  20.83   Mean   : 1.7331   Mean   :5222   Mean   :5221   Mean   :2.342  
 3rd Qu.:  19.00   3rd Qu.: 2.0296   3rd Qu.:5223   3rd Qu.:5223   3rd Qu.:2.000  
 Max.   :1231.00   Max.   :12.2421   Max.   :5259   Max.   :5259   Max.   :7.000  
     month     
 Min.   :11.0  
 1st Qu.:11.0  
 Median :11.0  
 Mean   :11.4  
 3rd Qu.:12.0  
 Max.   :12.0  
```

From the above provided readings, following are the key findings:

Bike Number: The dataset includes bike numbers ranging from 33,001 to 351,249.

Time Start and Time Stop: The earliest recorded start and stop times are on November 13, 2022, at 01:21:00 and 01:48:00, respectively. The latest recorded start and stop times are on December 13, 2022, at 22:25:00 and 22:40:00, respectively.

Duration: The duration of bike rides ranges from 1 minute to a maximum of 1,231 minutes. The average duration of approximately 20.83 minutes and Median of 11 minutes. The duration has a standard deviation of 49.65, which shows the significant variations and wide range in the durations of bike trips.

Air Distance: The air distance traveled during the rides ranges from 0.1303 to 12.2421 miles. The average distance is 1.7331 and median is 1.36. The standard deviation is 1.225, which is slightly less than the mean.

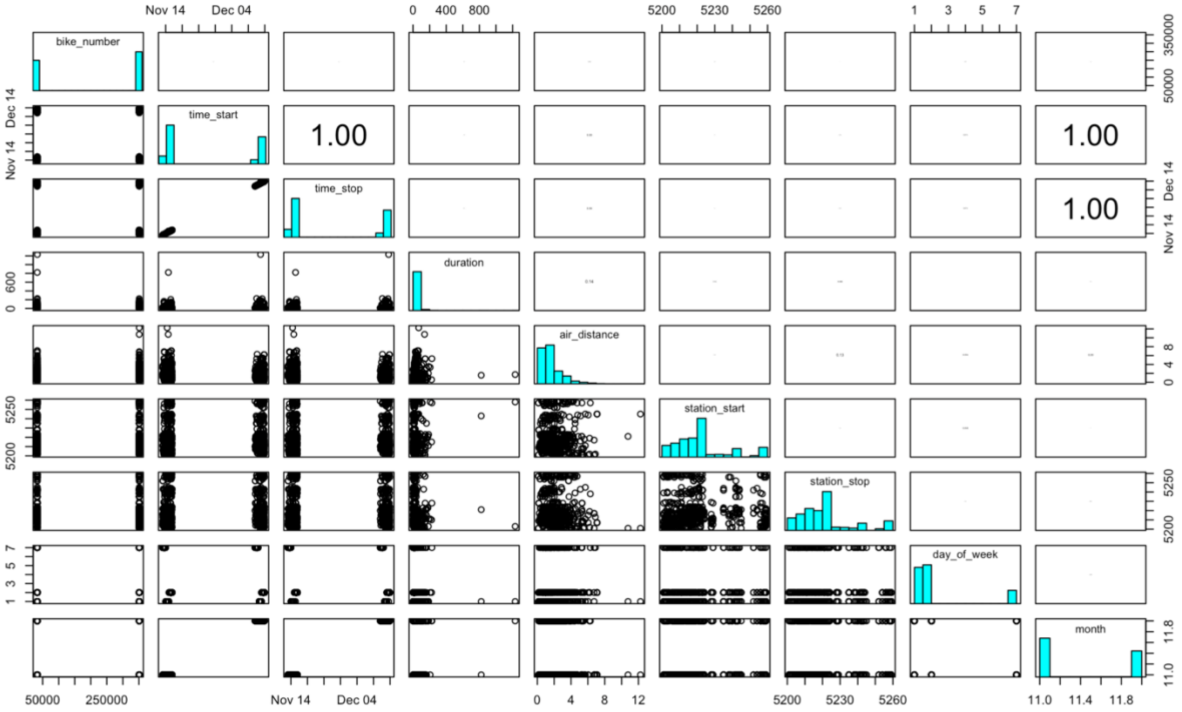
Station Start and Station Stop: This provides the information regarding the station numbers where the rides were started and stopped. The stations number starts from 5201 to 5259.

Day of Week: The min and max value gives and impression that the bikes were ridden throughout the week. However, the average value is approximately 2.34. Further information is gained from the plot.

Month: The rides were recorded in November and December. Mean value indicates that most of the observations are for the month of November.

Finding NA values:
There are no NA values found in the dataset.

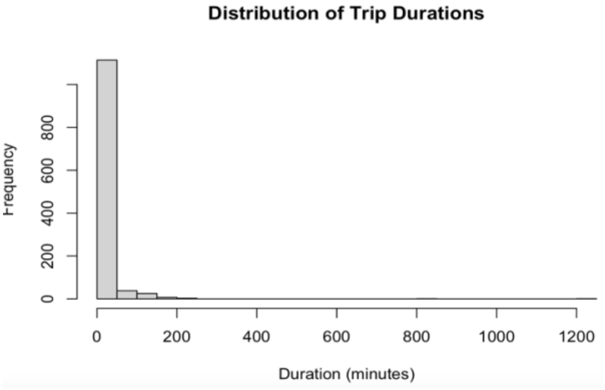
Visualizations and Insights:

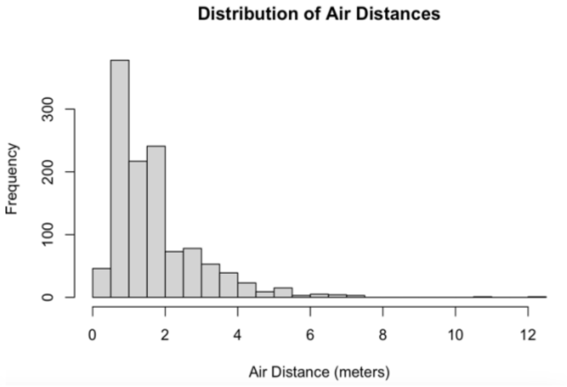


From the above plot we can see that there is perfect correlation between time start and time stop. This indicates that the trip start and stop has happened on the same day.

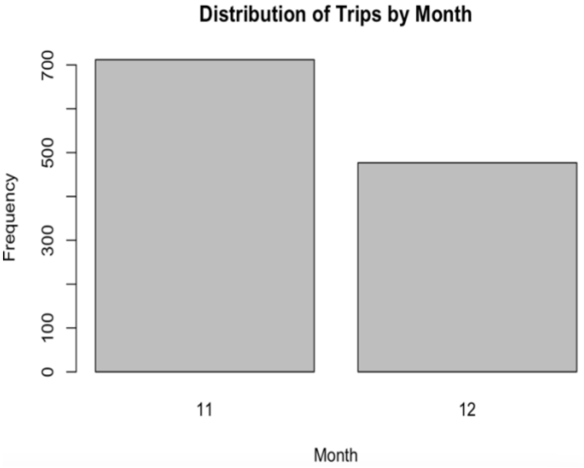
The data is not normally distributed. If we look at the prominent features such as duration and air distance, the data exhibits right skewed distribution. There is no linear relation between any of the other variables.

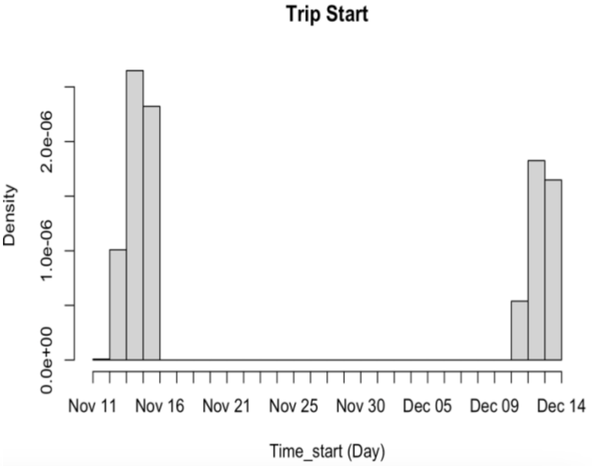
Histograms for Duration and Air distance:

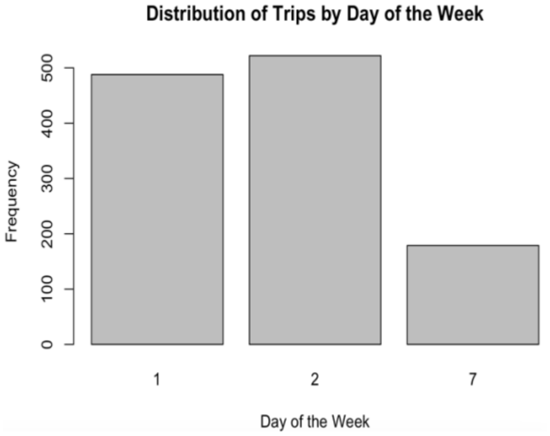




From the above two histograms, it is noticeable that the data is right skewed. In the duration, more than 90% of the data points has a duration below 50 minutes. However, when it comes to air distance the data points are spread across the wide range.





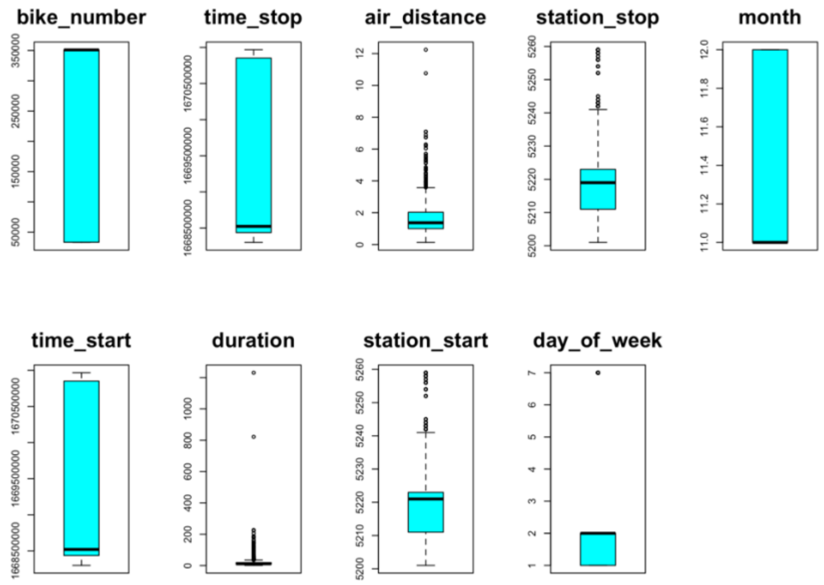


By analyzing the Month plot, it becomes apparent that the majority of bike rides occurred in the months of November and December. Specifically, a higher frequency of rides was observed in November compared to December. The observations are from November 11th to November 16th and from December 9th to December 14th.

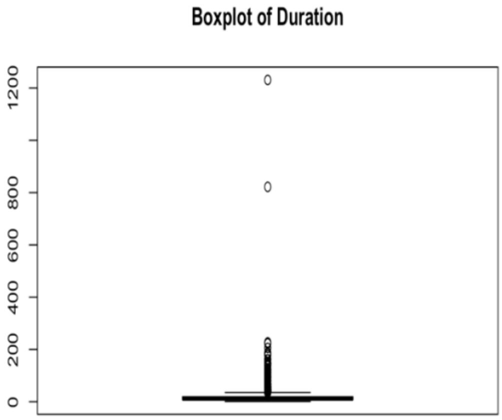
The bikes were ridden only on Mondays, Tuesdays, and Sundays. Most of the data points are for Monday and Tuesday. This indicates that the data is not widely spread across all the days of the week.

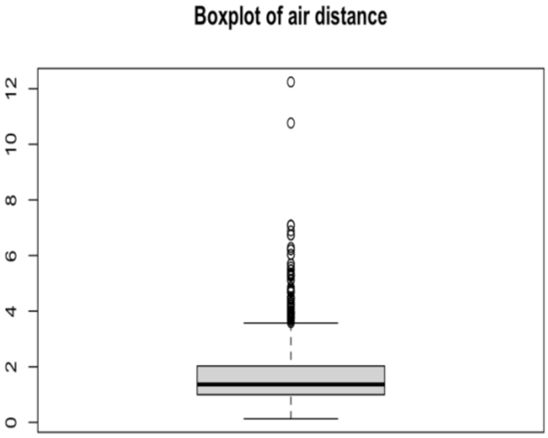
Outlier Detection:

Univariate Outliers:



Observing the above boxplots for all the variables, it is evident that there are several outliers present in the variables of air distance, duration, station start, and station stop. However, it is important to note that station start and station stop are irrelevant as they represent station IDs rather than meaningful numerical values. Therefore, our focus primarily lies on the variables of air distance and duration, which exhibit noticeable outliers.

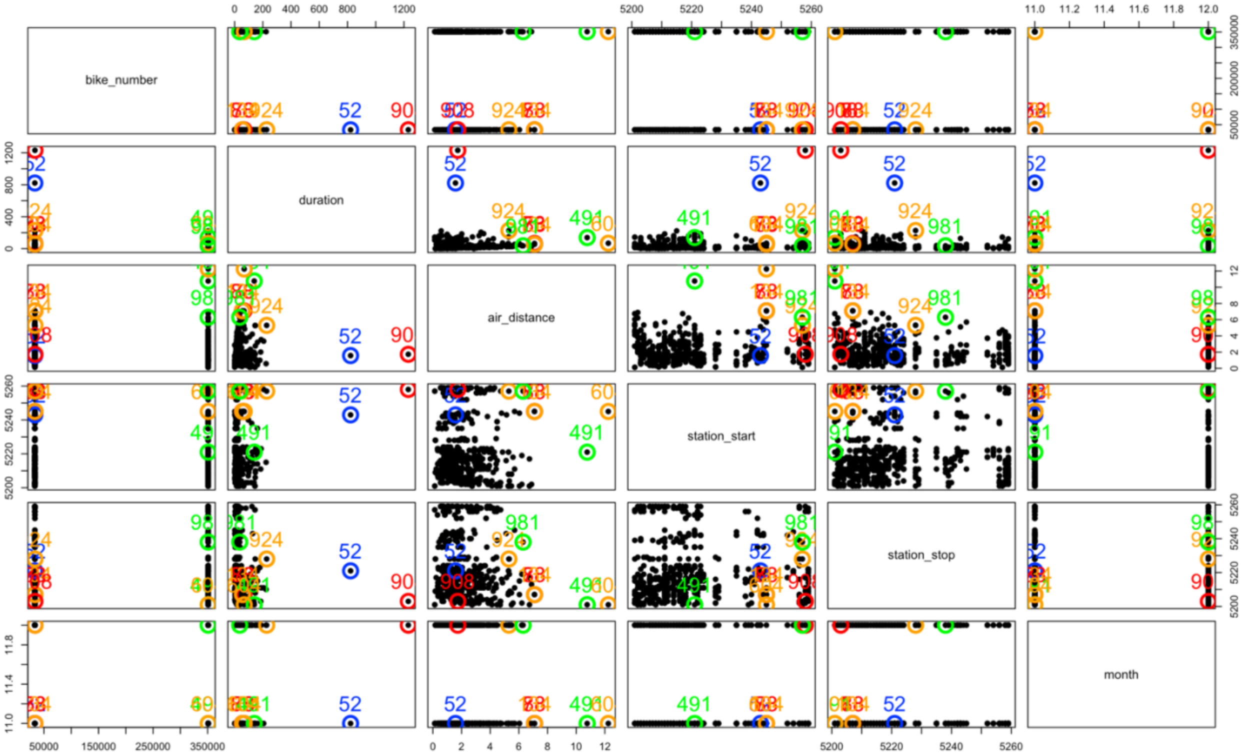




On further inspection we can see two extreme values in both the above boxplots. The values above 800 minutes are considered as extreme in duration and above 10 for air distance. When the values are filtered out, for the duration indices of the identified outliers are 52 and 908 and 491 and 604 respectively. This shows that the outliers are of different data points.

The outliers are not removed at this point, so that we can analyse these during multivariate outliers.

Multivariate Outliers:

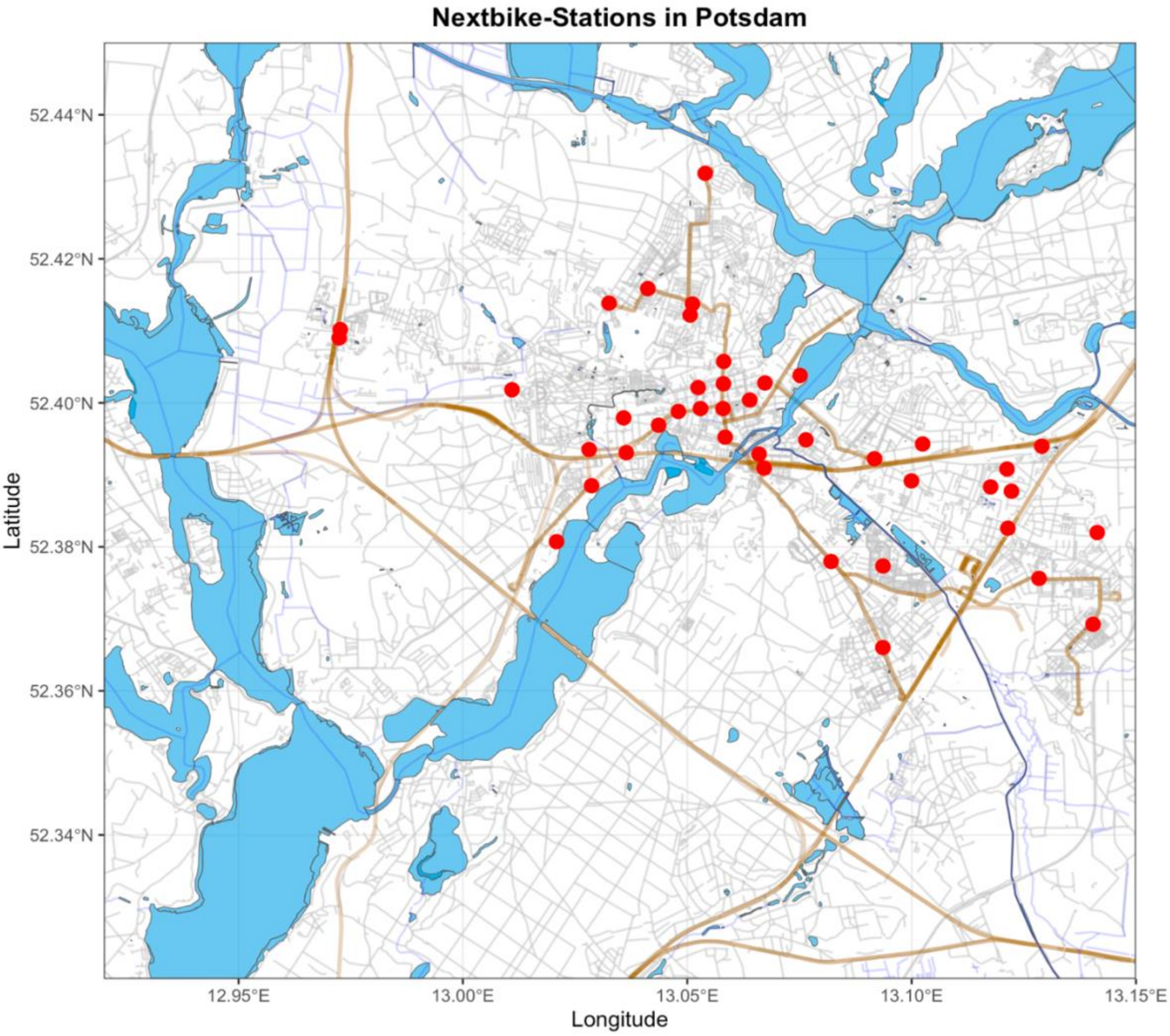


The above pairs plot shown is constructed with a threshold based on the 99th percentile. Any data points above that threshold are considered as outliers. The plot visualizes the relation between various variables along with highlighting all the data points which are considered as outliers. 52, 73, 88, 134, 491, 604, 908, 924, and 981 are indices identified as the outliers.

However, on further inspection we can see that 52 and 908 are the prominent ones in the most plots. They are completely segregated away from rest of the points. Also, these outliers are found in the duration boxplot during univariate outlier detection.

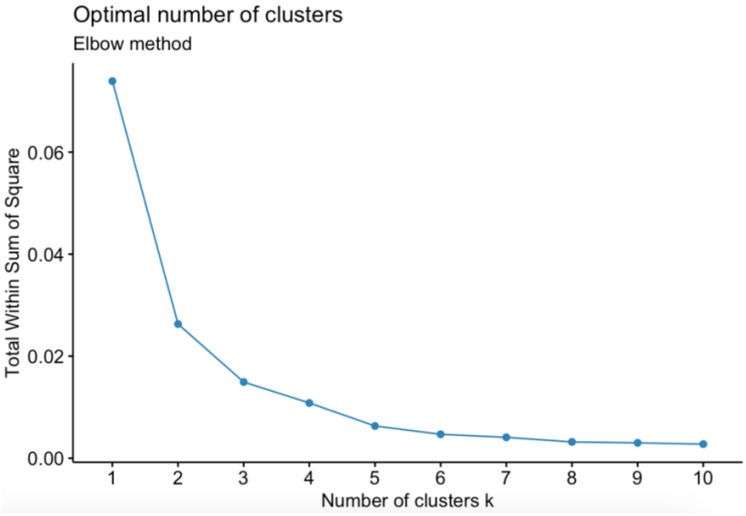
Hence, the outliers with the indices, 42, 908 and univariate outliers 491 and 604 are deleted from the dataset.

Map of Potsdam along with the respective bike stations:

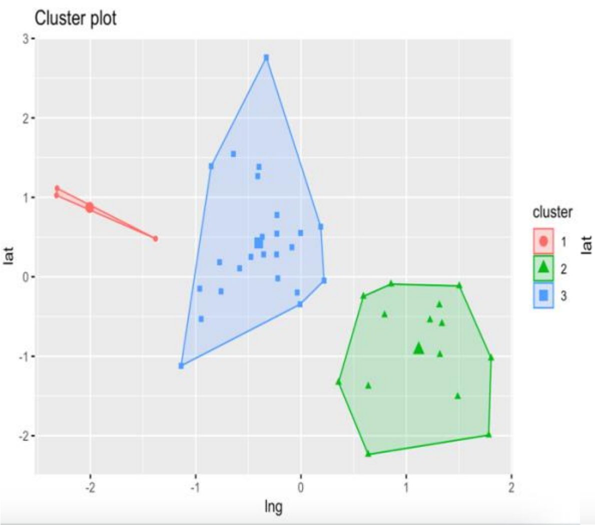


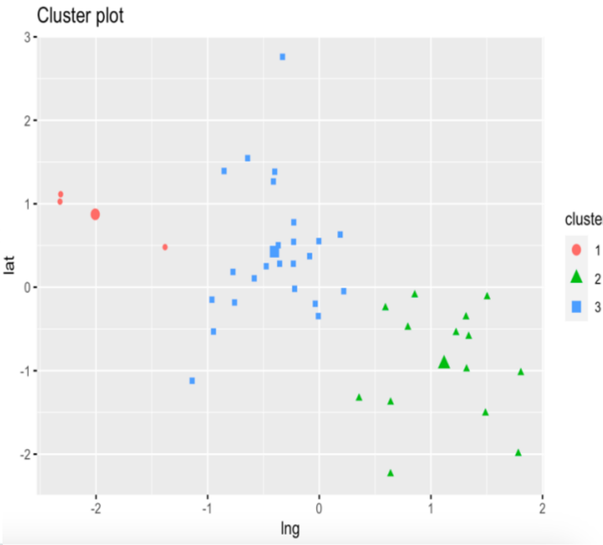
Clustering of the NextBike-stations:

K – means clustering:



From the analysis of Elbow curve, there are distinct deviations occurring after the third data point. There is significant change in the variation below the elbow. This suggests that there can be 3 clusters created for the stations. Hence for the k means, the value of k is taken as 3.

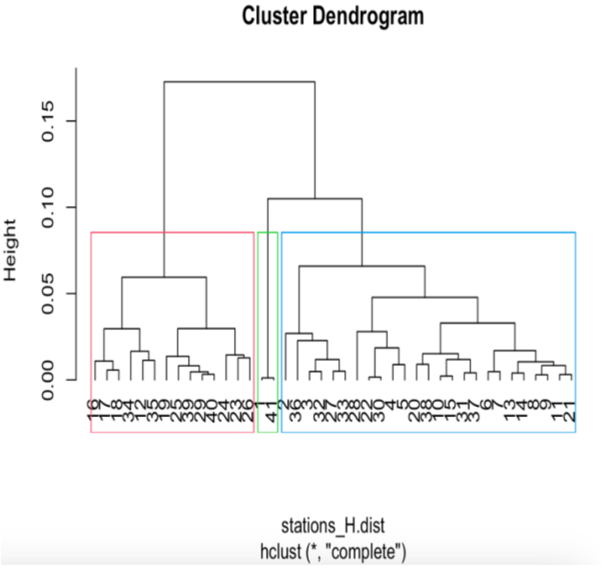


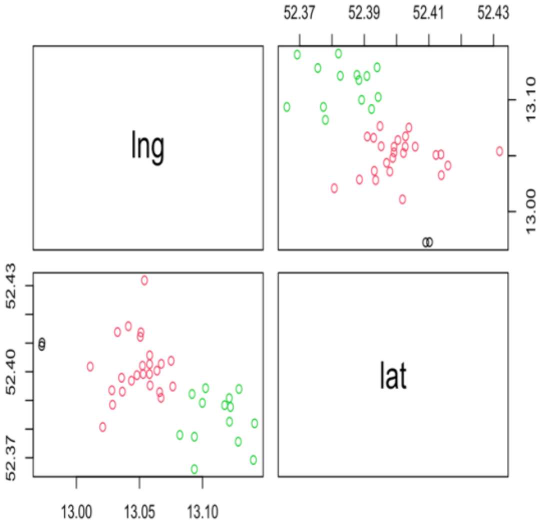


From the cluster plot, we can see that there are 3 cluster formations. There is no overlap between any of the clusters. The first cluster have a few data points compared to the other two clusters; however, the data points are at a considerable distance from the other clusters. So, we can consider it as a distinct group.

Overall, the cluster plot provides a visual representation of the clustering algorithm's results, highlighting the separation and composition of the different clusters.

Hierarchical clustering:

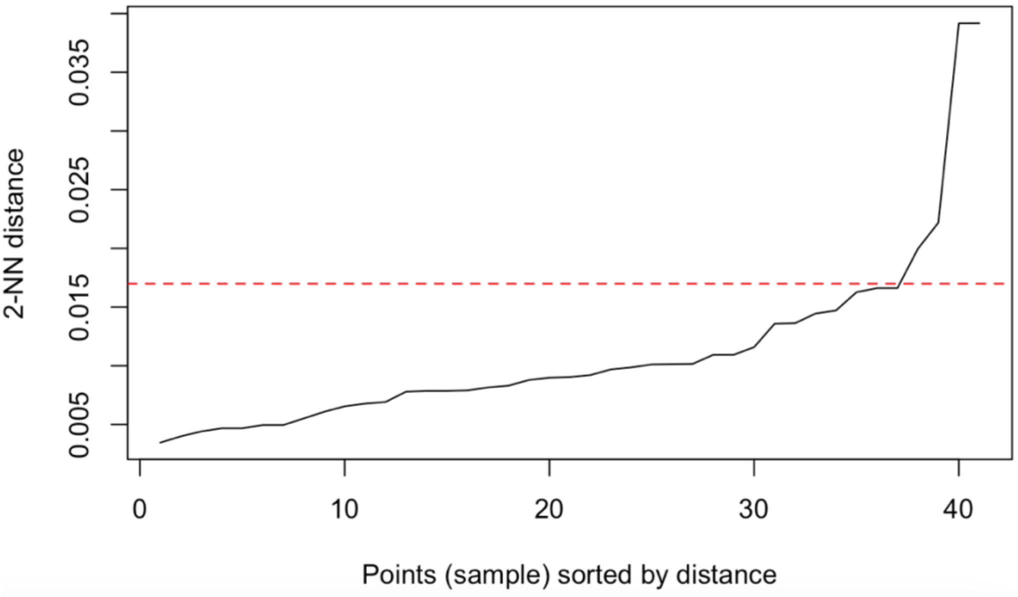




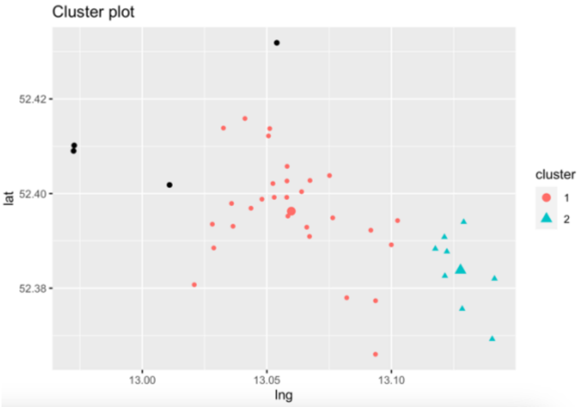
The second clustering technique used is Hierarchical clustering. From the dendrogram it is noticeable that there are three cluster formations. This can be observed based on the heights of the branches in the dendrogram, the gap between the first two branches is highest. So, the value of K i.e., number of clusters is taken as three. Here, the complete linkage method is used for the cluster formation.

From the second plot we can see the perfect formation of 3 clusters without any overlap between the data points or clusters.

DBSCAN:



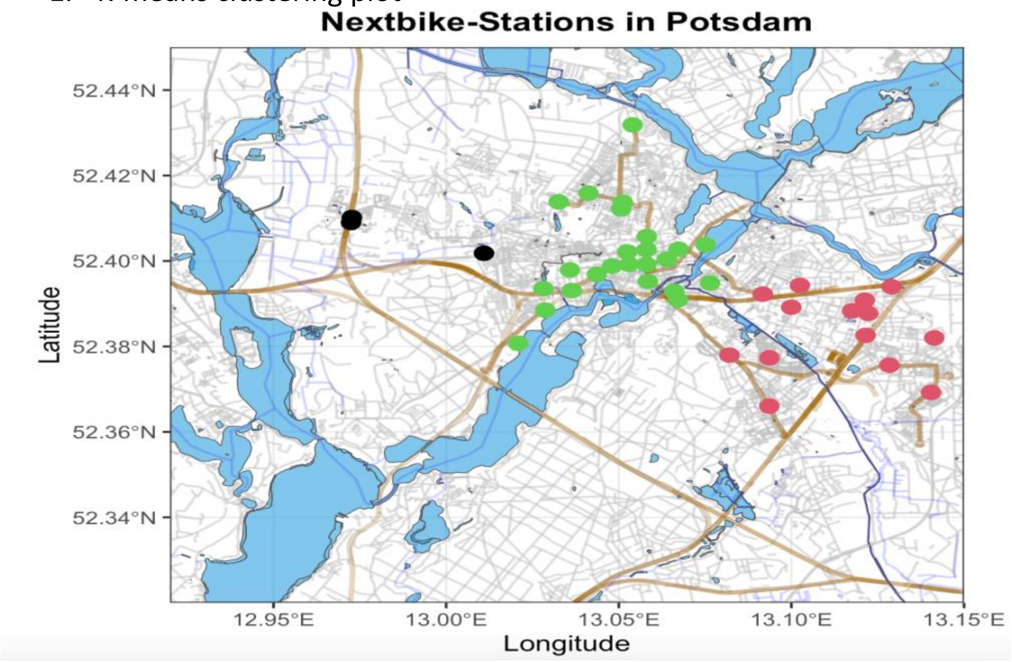
From the above KNNdistplot, we can see an elbow at 0.017. This suggests that values above it is a noise and the values below 0.017 are considered for the cluster. The value chosen for `eps` is 0.017 and minimum points for the neighbors is considered as 5.



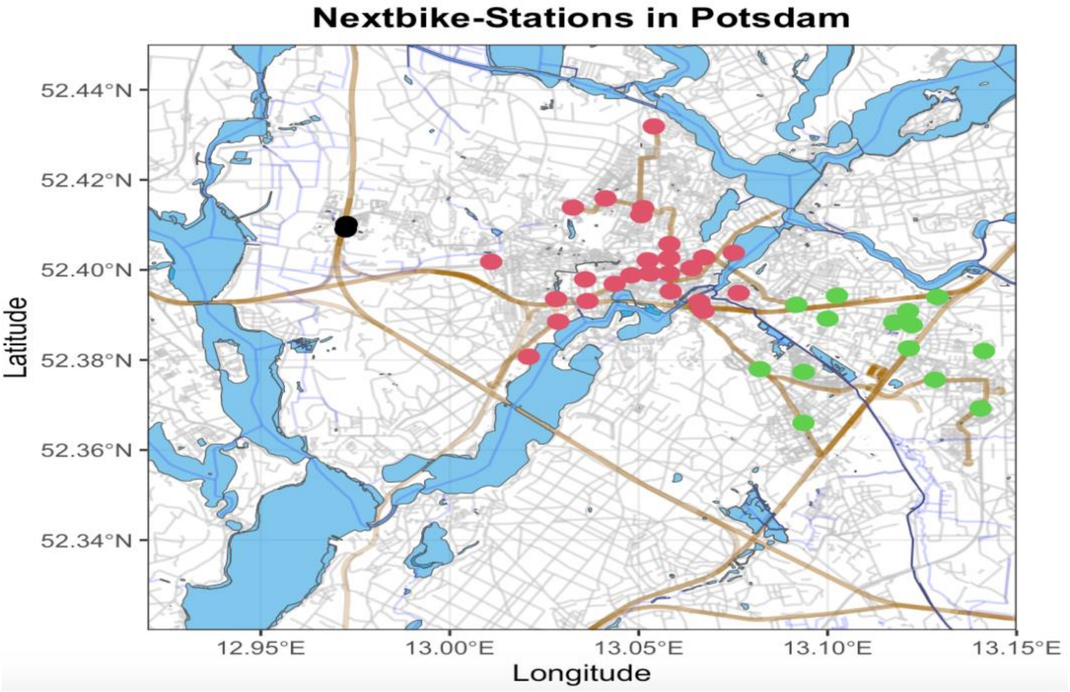
Upon observing the results, it is evident that the data exhibits the formation of two clusters; however, there are also a few missing stations. This clustering outcome does not represent an ideal scenario as the stations are not effectively grouped into the clusters.

Visualizing cluster results by color-coding:

1. K-means clustering plot



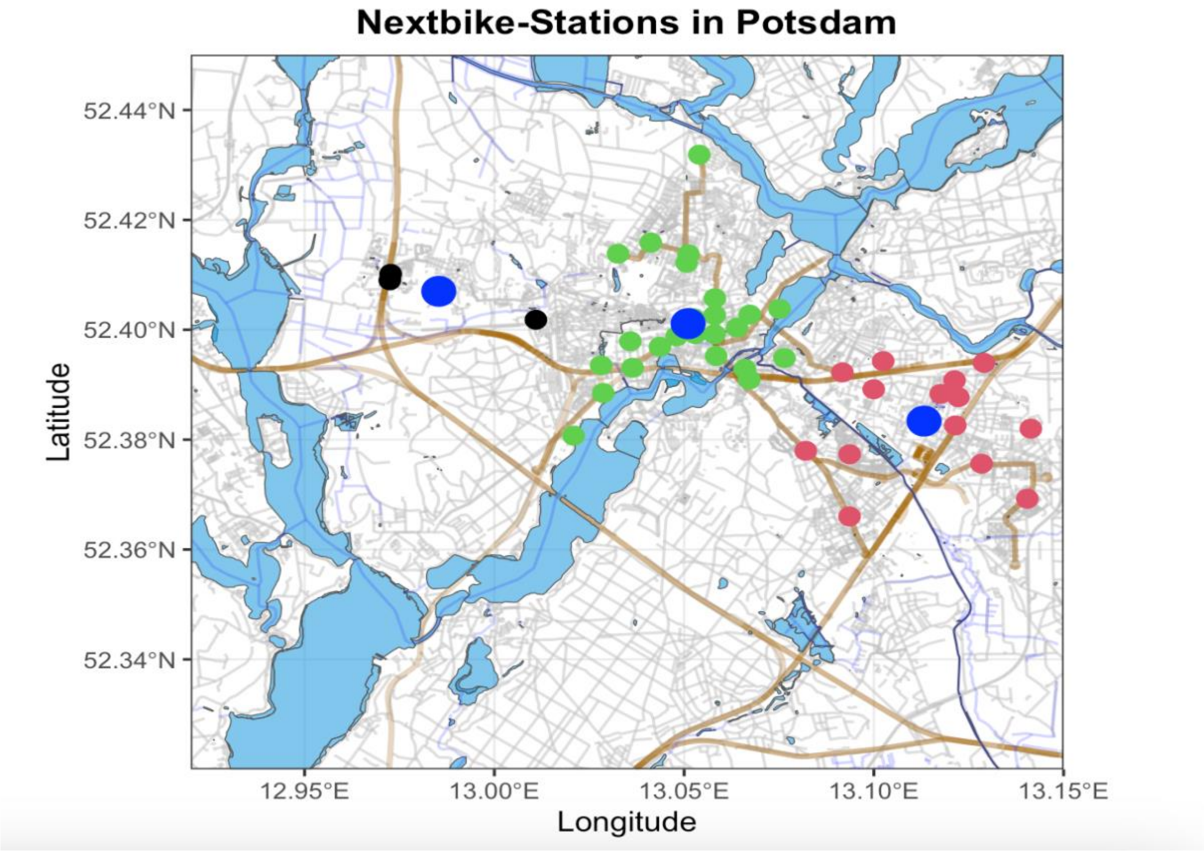
2. Hierarchical clustering plot



In order to create a clustering representation on the map, K-means and Hierarchical clustering algorithms are used. DBSCAN did not encompass all the data points within the cluster formations so it is excluded.

Based on the above two maps, the K-means clusters were chosen due to their clear and distinct categorization. The three clusters can be formed based on city, down town and suburbs.

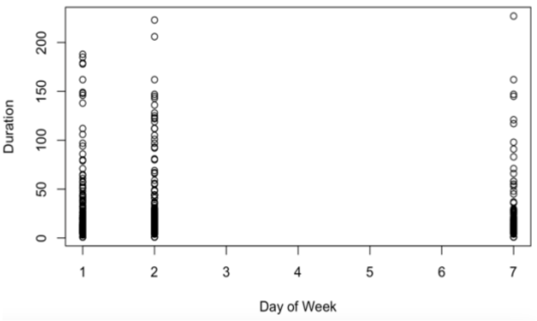
Prominent service station for each cluster:

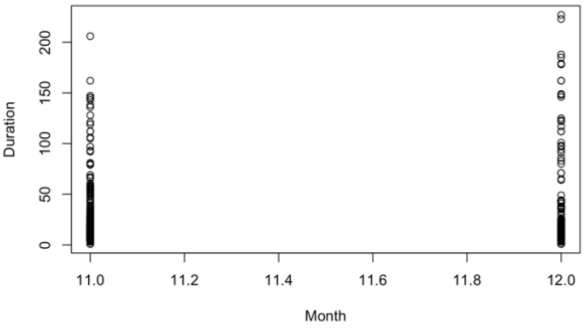


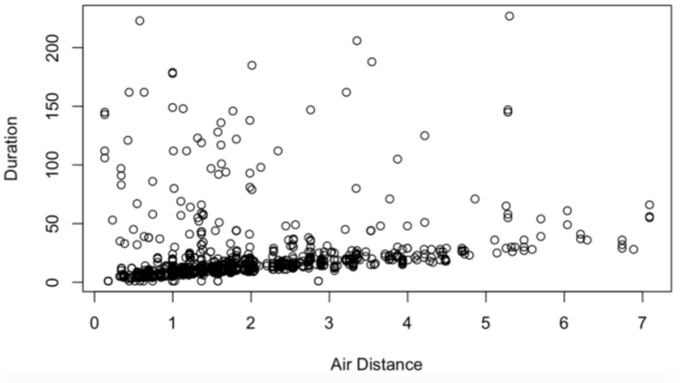
To ensure accessibility for riders from different stations within each cluster, it is crucial to establish service stations strategically. In this context, the approach employed involves determining the centroids for each cluster. By locating the centroids, which represent the geometric center of each cluster, an ideal station placement can be achieved that is positioned in the middle of the cluster. The blue highlighted points in the map are the ideal spots to establish the service centers, which are assessable and convenient from all other stations.

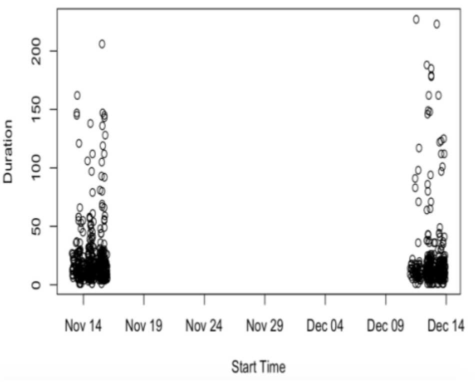
Understanding relationship between variable using regression techniques:

To identify the relationship between variables, the scatter plots are drawn considering the prominent features with respect to duration.

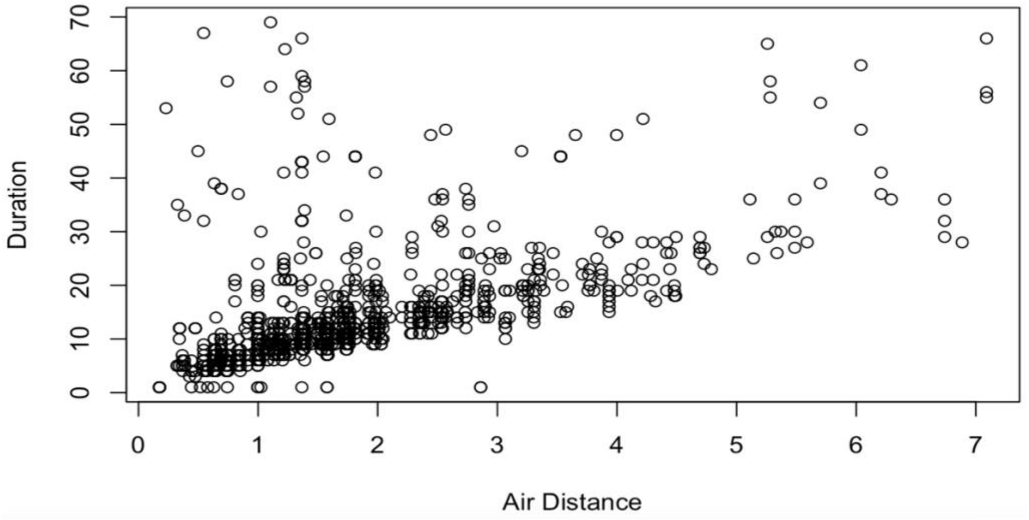






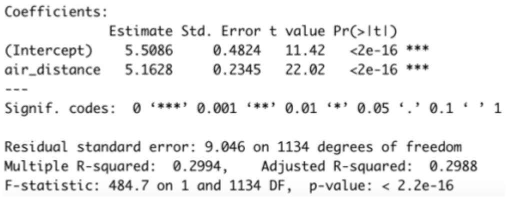


From the above plots, its noticeable that only air distance and duration have a certain amount of linear relationship compared to the other features. However, there are many dispersed points above the duration value of 70 and on finding the correlation, the value obtained was just 0.25, which is very low. As a result, data points with a duration value exceeding 70 minutes were excluded, leading to an observed increase in the correlation coefficient to 0.55.



Scatter plot after removing data points

Since duration has linear relation and higher correlation coefficient only with Air distance, the linear regression technique opted here.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5086     0.4824   11.42   <2e-16 ***
air_distance  5.1628     0.2345   22.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.046 on 1134 degrees of freedom
Multiple R-squared:  0.2994,    Adjusted R-squared:  0.2988
F-statistic: 484.7 on 1 and 1134 DF,  p-value: < 2.2e-16
```

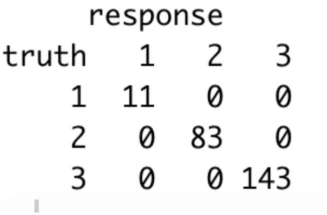
From the above table it is evident that there is relationship between the duration and air distance based on the provided information. The coefficient estimate for air distance is 5.1628, which suggests that, on average, for every unit increase in air distance, there is an increase in the duration 5.1628 units. Additionally, it has a very low p-value.

For further analysis multi linear regression was opted. However, apart from the air distance, rest of the p-values were way higher. In summary, the analysis indicates a relationship between the duration and air distance variables. The other variables can be considered redundant.

Different classifiers with the benchmark study:

The drop-off station ids were assigned the cluster numbers based on the K means clustering.
The data set is split into training and test data, with 80-20 split.
Following are the classifiers used for the study

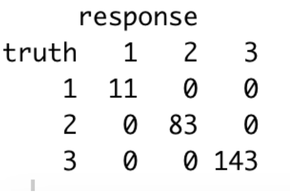
1. Decision Tree:



```
     response
truth   1   2   3
    1  11   0   0
    2   0  83   0
    3   0   0 143
```

From the confusion matrix it is evident that, the model hasn't misclassified any data between the 3 clusters. This indicates 100 percent accuracy.
The misclassification error is 0 and with an accuracy value of 1.

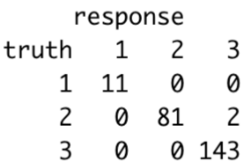
2. Random Forest:



```
     response
truth   1   2   3
    1  11   0   0
    2   0  83   0
    3   0   0 143
```

The results obtained from the random forest evaluation are comparable to those achieved by a decision tree model.

3. Support vector machines:



```
     response
truth   1   2   3
    1  11   0   0
    2   0  81   2
    3   0   0 143
```

The confusion matrix shows two misclassifications for second cluster. The misclassification error is 0.0084 and with an accuracy value of 99.1 percent. However, the other two classifiers performed better.

Following are the results from the benchmark study:

Decision Tree: Aggregated Result: mmce.test.mean =0.0000000, acc.test.mean=1.000

```
Random Forest:   Aggregated Result: mmce.test.mean=0.0010526,acc.test.mean=0.9989474
SVM:             Aggregated Result: mmce.test.mean=0.0095297,acc.test.mean=0.9904703
```

From the benchmark study we can conclude that, Decision tree classifier gives the highest accuracy.

Feature Selection:

In the feature selection process, the decision tree classifier was chosen to identify the most influential features. The analysis revealed that the "lng_stop" and "lat_stop" features, which represent the longitude and latitude coordinates of station stops, respectively, are the most prominent factors in the dataset. These features were found to have a significant impact on the classification outcome.