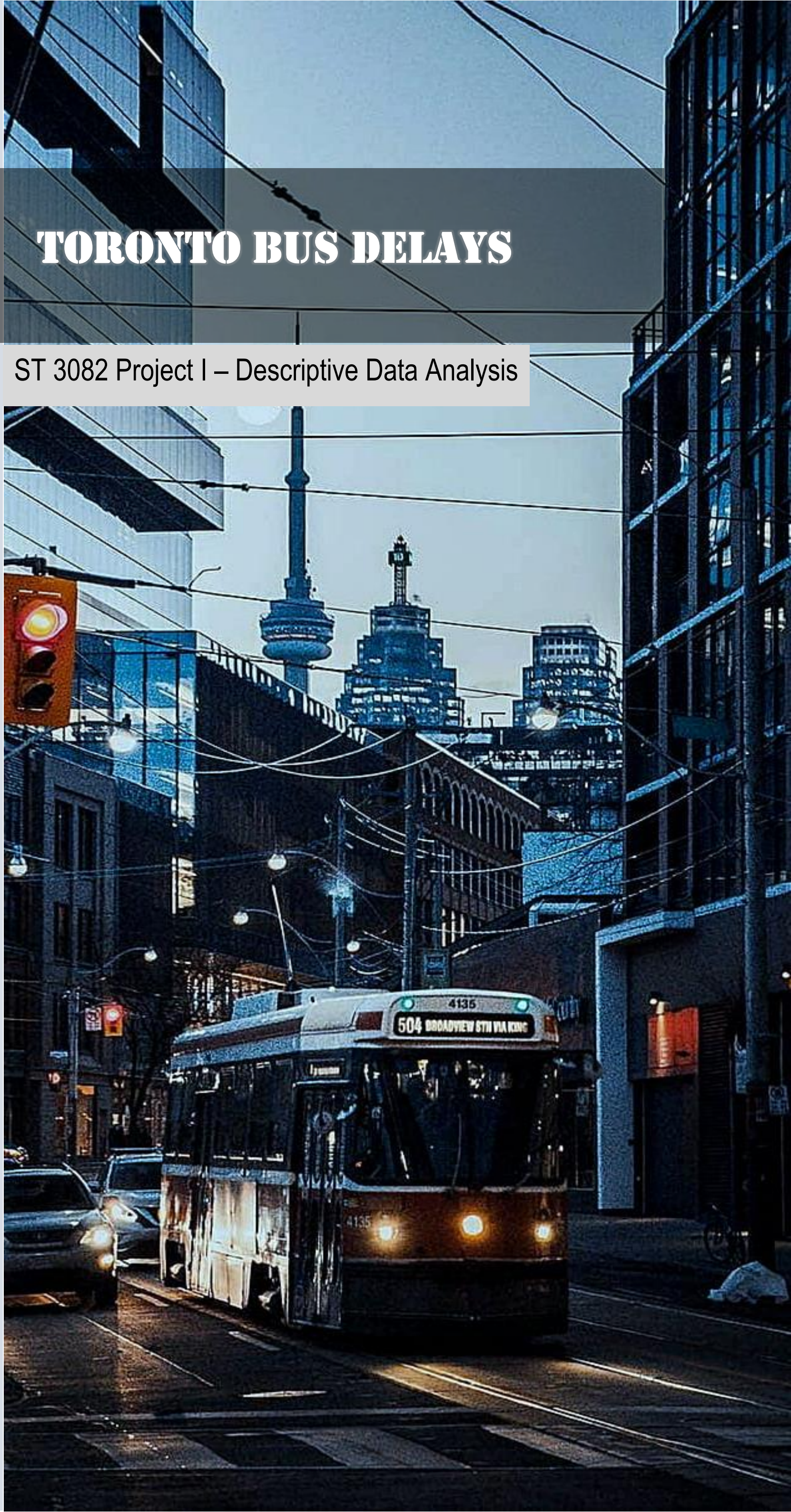
TORONTO BUS DELAYS

ST 3082 Project I – Descriptive Data Analysis



Group 4

s14853 – Pramudi Rajamanthri

s15030 – Vidura Chathuranga

s15091 – Chalani Wijamunige

Abstract

This report aims to outline the findings of the exploratory data analysis which was carried out based on the Toronto bus delays 2022 data set obtained from the open source data site, Kaggle. Furthermore, this report focuses on identifying the frontline reasons for Toronto bus delays along with their inter relationships and suggestions for fitting a suitable model to predict the delay time of buses with high accuracy.

Contents

List of Figures

List of Tables

1. Introduction

In Toronto, buses play a major role in public transportation. As per records in 2021, TTC bus system has 192 bus routes in operation including 28 night bus routes carrying over 264 million riders over 6686 kilometers of routes with buses travelling 143 million kilometers in the year. The system had a ridership of about 8,74,300 per weekday by the end of first 6 months of 2022. Bus routes extend throughout the city and are integrated with the subway system and the steer car system with free transfers among the 3 systems. In the recent past years, TTC is constantly being complained on the delays of the bus system despite of having a wide network. Numerous research studies have been done on the factors leading to these delays in order to fix this issue with the guidance of the professionals, since the primary objective of transportation efficiency is directly affected from these delays in the bus system. With the dataset in hand, an analysis can be performed to find the factors that have the most influence on bus delays in Toronto and a model can be built in order to predict the delay time in minutes accurately for the ease and the awareness of the riders. To build the necessary

foundation for this model, an exploratory data analysis will be conducted on the dataset which carries data on bus delays within the first 6 months of 2022 in Toronto, Canada.

2. Description of the Question

It is important and timely to address the issue in drop of efficiency in the Toronto bus system due to delays, in order to fix the problematic spots in the bus network, as it is directly connected with the day today life activities of millions of people. Hence, identifying and understanding the main factors resulting this issue is essential in setting the path to predict the delay time in minutes out of the visible causes such as route types, incidents causing delays, week of the day etc, to make the passenger lives easy by being aware of the possible amount of delay times.

3. Description of the data set

The Toronto bus delays 2022 dataset contains 27351 records and 10 variables, including Delay time as the response variable.

Variable Name	Variable Type	Description
Date	Qualititative	Date of the delay incident
Route	Qualititative	Route number which the incident happened
Time	Quanititative	Time of the day which the incident happened
Day	Qualititative	Day of the week which the incident happened
Location	Qualititative	Location in Toronto which the delay incident happened
Incident	Qualititative	The incident which caused the delay
Min.Delay	Quantitative	Amount of time which the bus got delayed
Min.Gap	Quantitative	Time gap in minutes with the next bus scheduled
Direction	Qualititative	Direction of the route which the particular bus travels
Vehicle	Qualititative	Vehicle number of the bus

Table 3.1

4. Pre-processing

Toronto bus delay dataset contains data under 10 variables where 7 are categorical and 3 are numerical. In the dataset, 154 duplicated entries were discovered and those were removed. Then it was found that there are 164 missing data records in the "Route" variable and 5529 missing records in the "Direction" variable. Afterwards, all the records having 0 as the "Min.Delay" value were removed from the dataset as only the buses which got delayed are being considered in the analysis. By doing so, the count of missing values of "Route" and "Direction" variables got reduced to 26 and 4987 respectively with a total of 25898 records.

4.1 Feature Engineering

A separate "Month" variable was created by extracting the months in the "Date" variable. Distinct values in the "Direction" column were identified and entries with error values discovered in that variable were filtered out. Considering the "Route" variable, routes labelled in text form were replaced by a random number for the ease of cleaning process.

1.Regular and Limited Service Routes	7 – 189 series	Weekdays – 6am-1am Weekends - 8am-1am
2.Blue Night Routes	300 -399 series	Weekdays – 1am-6am Weekends - 1am-8am
3.Community Routes	400 – 499 series	Only in rush hours of weekdays Rush hours (According to TTC statistics) :- Morning – 6am- 10am Evening - 3pm- 7pm
4.Express Routes	900 – 999 series	Any time with in the 24 hours (according to requirements)

Table 3.2

Based on the above details we got from the TTC official website, the new variable "Route_New" was created by categorizing all the distinct routes in the "Route" variable into 5 categories *(Blue Night Routes, Community Routes,*

Express Routes, Regular and limited service routes, Others) based on their route numbers. The "Hour" variable was introduced to the dataset by extracting the hours in the "Time" variable. The new variable "is_Weekday" was formed by categorizing the days in the "Day" column into 2 categories *(Weekday, Weekend)*. The "Incident_New" variable was created by taking the distinct categories in the "Incident" column and combining 3 categories with lowest number of cases *(Cleaning – Disinfection, Held By, Late Entering Service)* into one category *(Others)*, and renaming the category *"Road Blocked – NON–TTC Collision"* into *"Road Blocked"*. The "Vehicle" variable was removed from the dataset as it acts as an identifying variable and it causes no effect for the delay analysis.

Furthermore, by referring to the website of TTC bus system, the operating days and times of the bus routes were identified and the records which do not follow this day and time system under its corresponding bus route were filtered out as they are contradictory with the system data which left us with a total of 23853 records. Finally, the dataset was split into training and test data sets as 80% and 20% from the original data respectively for further analysis. Afterwards the variables with missing values of the training set were taken into consideration and we identified missing values in "Route","Route_New" and "Direction" variables as 24, 24, 3528 respectively. As "Route" variable is not considered for further analysis while "Direction" variable leads to misinterpretations with imputing using the mode as all its categories have approximately equal number of data currently, we only imputed the "Route_New" variable using the "*Hour"* and the "is_Weekday" observations relevant to those records with data in table 3.2.

5. Main Results of the Descriptive Analysis

The purpose of our model is to answer the question of how the delayed times of buses in Toronto are behaving with its associated factors. Therefore firstly, we consider the distribution of the delayed time in minutes. As shown in the figure 5.1, the distribution is clearly right skewed with a skewness of 10.69332 and a kurtosis of 148.4564. Also, it is dispersed within a wide range of 0 to 1000 mins due to its outliers.

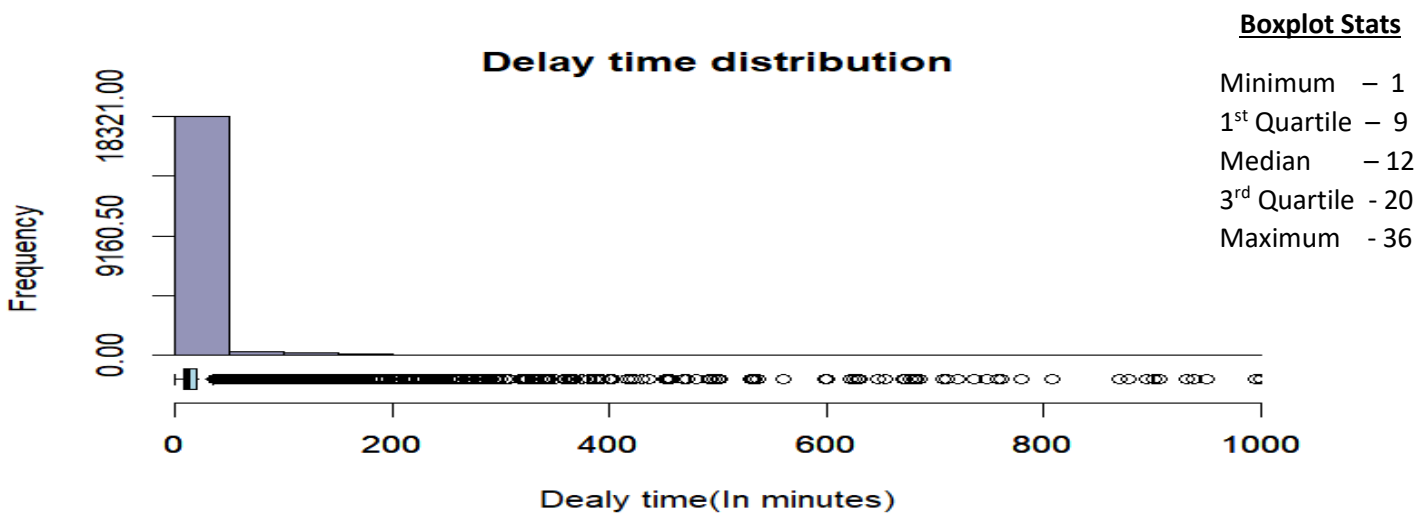


Figure 5.1

Most frequently recorded delay times falls in to the class of 0 – 50 mins, and as the time increases the frequency drastically drops. This is an acceptable scenario, but the problem is the presence of many extreme outliers. In this case it should be noted that the mean can be pulled towards these extreme values. Here the mean value of the delay times is 20.9 mins which suggests that it is not much affected by the outliers as majority of the observations are gathered in 0-50 range as shown above in figure 5.1. We identified 994 outliers in the response variable from the boxplot.

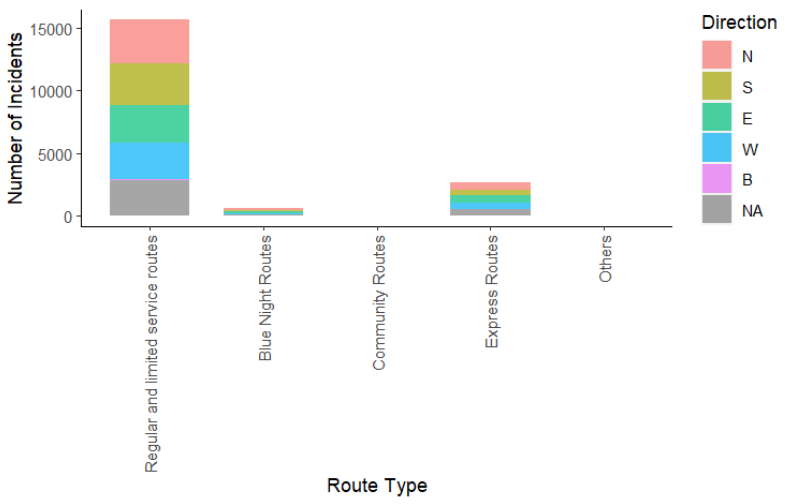


Figure 5.2

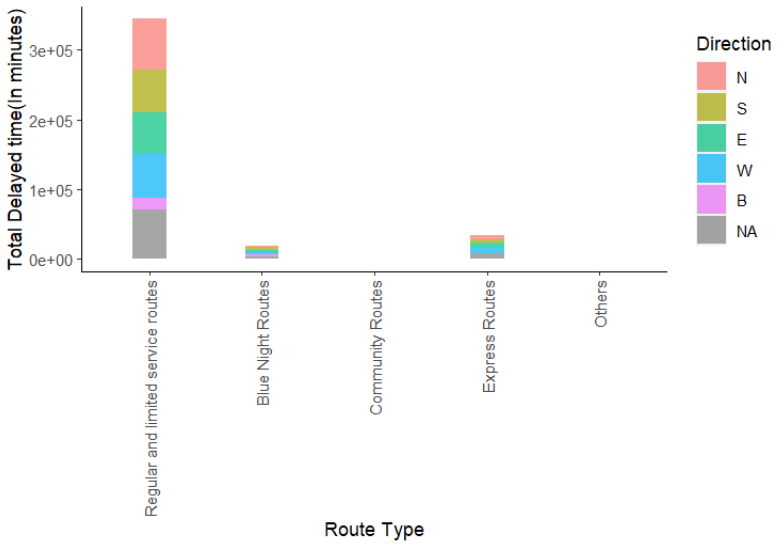


Figure 5.3

Moving further in the analysis, as shown in the figure 5.2, number of incidents reported in each route type were represented in a bar plot and it was categorized according to the direction of the route as well. Most of the cases were reported in the regular and limited services routes. The number of records exceeds 15,000 whereas the total number of records is 19,083 in the training dataset. This result is acceptable as regular and limited service routes function within the maximum portion of a day, both in weekdays and weekends according to TTC. The second highest number of incidents were recorded in the Express Routes. The counts on the rest of the routes are negligible when compared to these 2 route types. Considering direction wise, in each route type, the buses that goes on both directions in those identified routes, records the minimum number of incidents which is negligible compared to the other directions. Buses which go to the northern direction records most of the incidents.

In order to identify how the delay time is distributed among these routes, in figure 5.3 the same bar plot is plotted considering the total delayed time in the y axis.

Similar to the previous plot, the total delayed time of the buses that travel in the regular and limited service routes record the highest with a dominant number of 3,44,921minutes when compared to other routes.

Following that, the Express routes record the second highest total delayed time and Blue Night routes are next in line as expected.

Here we see that when considering both plots in figure 5.2 and figure 5.3, the missing values of direction variable play a significant role such that most of them belong to the regular and limited service routes and it would have been misleading to impute them using the mode in direction variable(N), since it will make the cases relevant to North become significantly huge comparatively as they are approximately equally distributed among the categories*(N-North, S-South, E-East, W-West, B-Both ways of a route)* except "B" without imputation.

Here in figure 5.4, a significant result was observed. According to TTC bus schedule, the Regular and Limited Service buses operate from 6 am to 1 am and 8 am to 1 am on weekdays and weekends respectively, while Blue Night network buses operate from 1 am to 6 am and 1am to 8am on weekdays and weekends respectively. Here the reason for the higher mean delay times of the hours between midnight to 6am in both weekdays and weekends may be the presence of outliers with larger values which we identified in figure 5.1 while having small number of data points relevant to those hours as we identified in the counts of incidents in blue night routes which function in this period. As a whole, the mean delay times of the weekends are larger than that of weekdays.

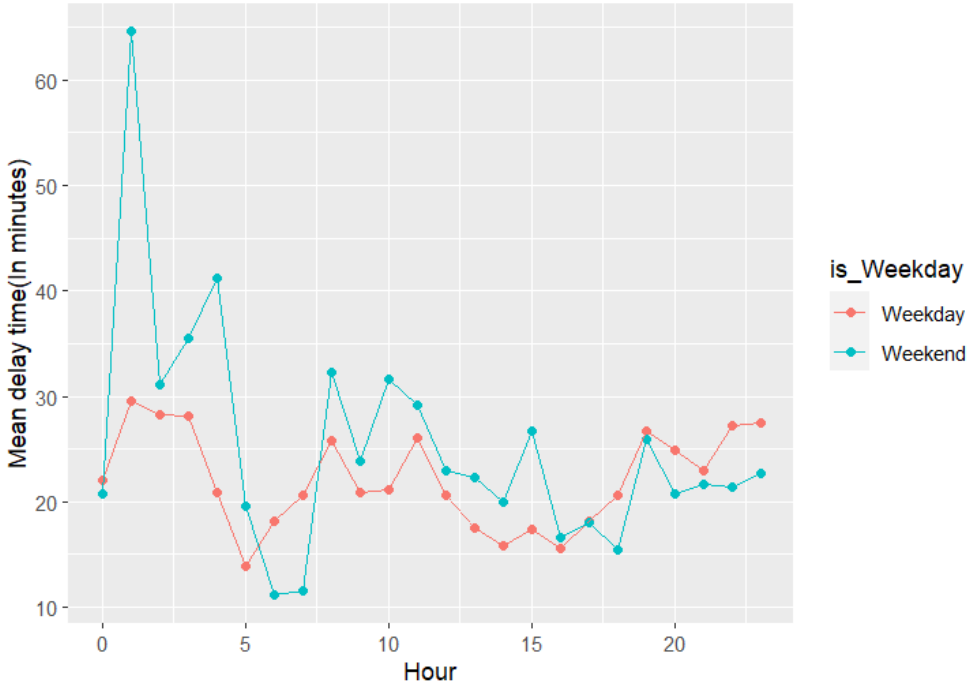


Figure 5.4

In order to clarify these observations, we also plotted the distribution of the delay data by hours of a day considering weekdays and weekends.

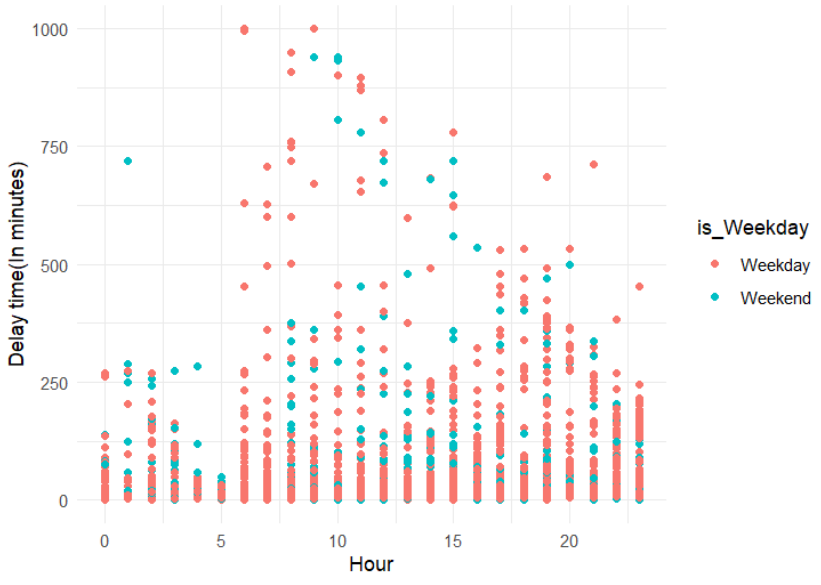


Figure 5.5

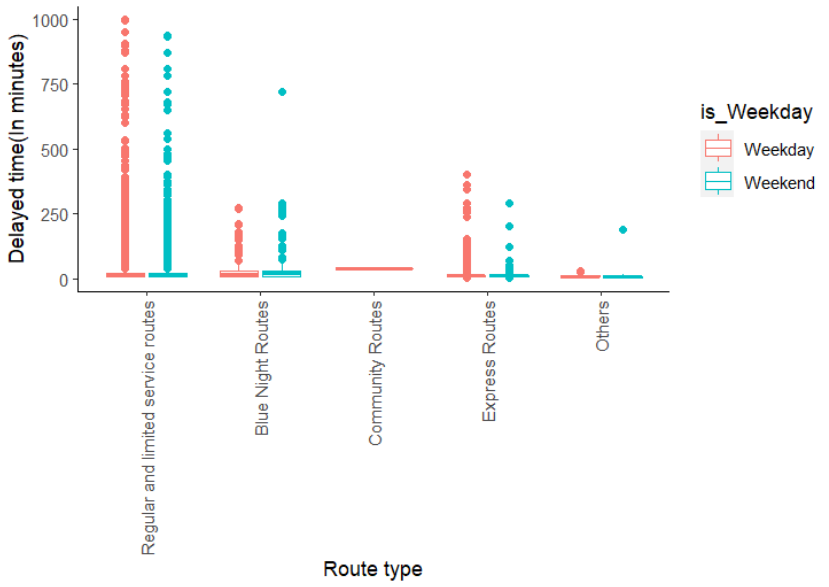


Figure 5.6

As observed in figure 5.5, clear majority of the data points belong to weekdays which shows that the higher mean delay times of the weekends in figure 5.4 are due to the less number of data points corresponding to the hours in weekends than in weekdays.

According to figure 5.1, most of the delays do not exceed the 100 min limit which is verified by figure 5.5 observations, as we can see that majority of the data points belong to the 6am-1am range which is the common range for regular and limited services routes in both weekdays and weekends. Also, we see that majority of the outliers also belong to the same range but they have not resulted higher mean delay times of those hours verifying the observations in figure 5.4 as the number of data points in this period is considerably very large as shown in figure 5.5.

By observing figure 5.6 which represent the delay distributions of weekdays and weekends by the 5 route types, the wider distributions of regular and limited service routes, say that the shape of figure 5.1 is affected mostly due to the cases in regular and limited service routes while the others stay way behind considerably.

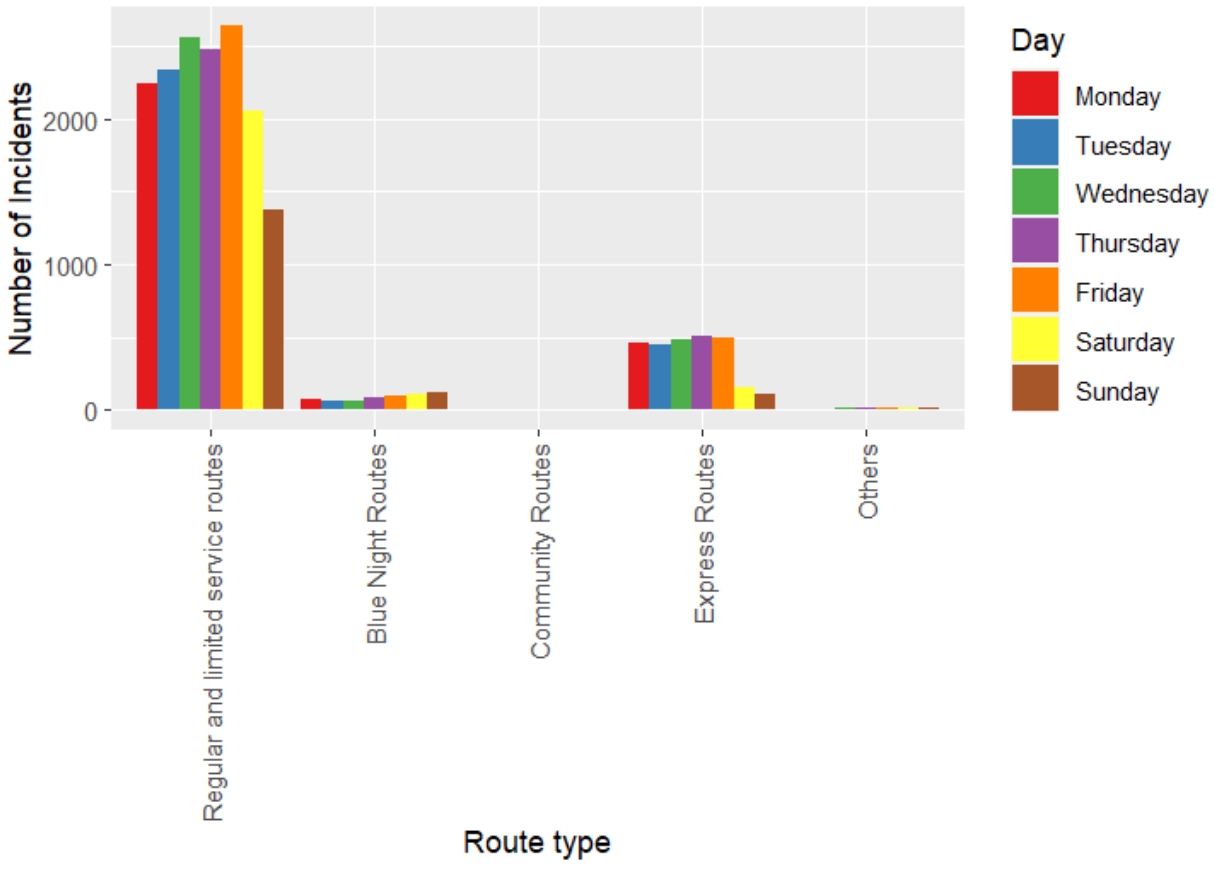


Figure 5.7

Even though we identified that the weekdays have reported the majority of the cases, we wanted to examine how the delay cases have been distributed among the days of the week in order to identify which days have become the worst for the Toronto bus riders in the first 6 months of 2022.

In figure 5.7, we see that Fridays have reported the highest number of cases as a whole, while regular and limited service routes and express routes capture the majority of the cases .

Since the reasons for the delay is one of the most important factors in this delay analysis, figure 5.8 shows how the delay time is distributed among the reason types.

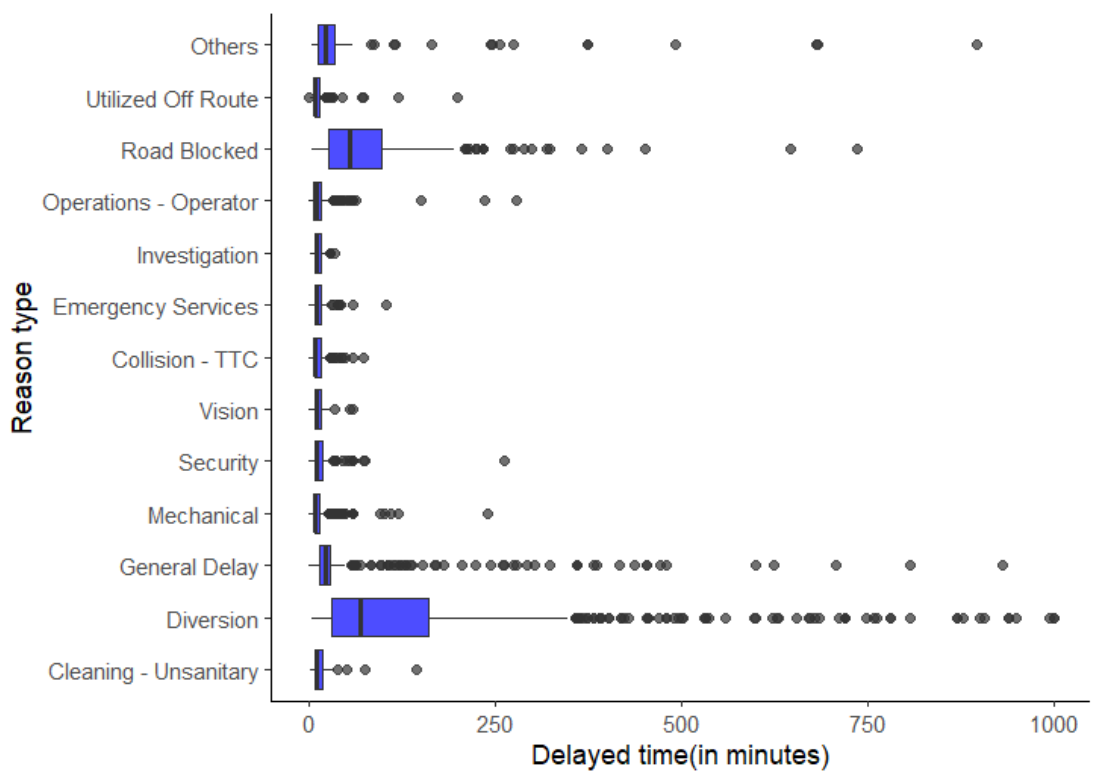


Figure 5.8

Here diversion and blocked roads report wider distributions with outliers of larger values, while all the other reasons significantly disperse in a narrow range depicting the larger outliers we observed in the above plots are mainly due to blocked roads and diversion of routes. Since we identified that the weekdays have reported more delay cases in the first 6 months of 2022, we wanted to see the distribution of the counts of cases among these reasons by weekdays and weekends.

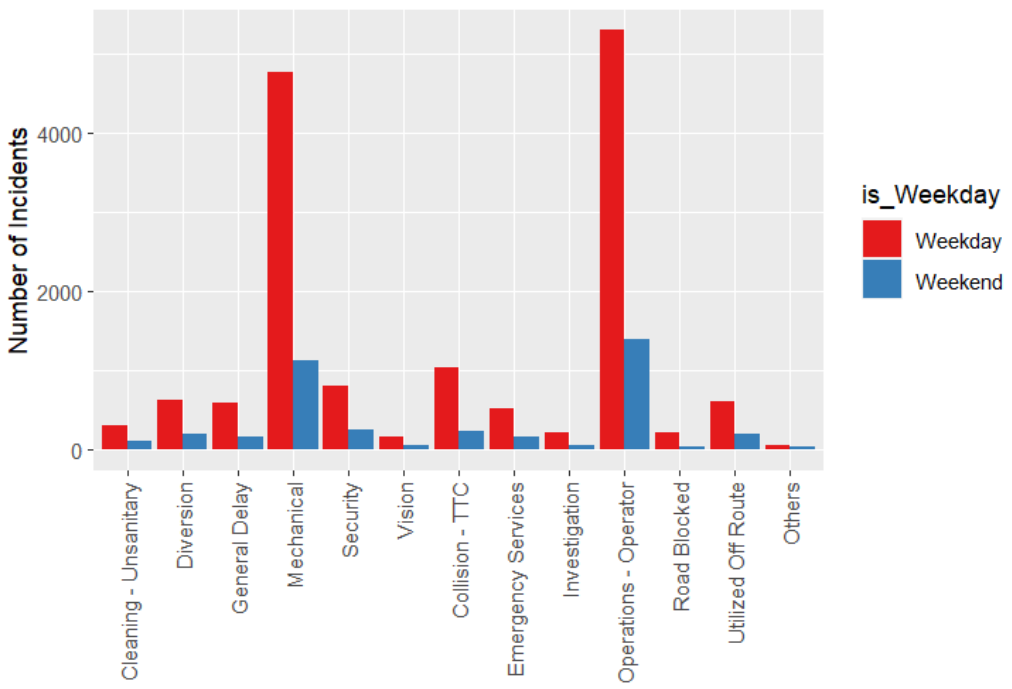


Figure 5.9

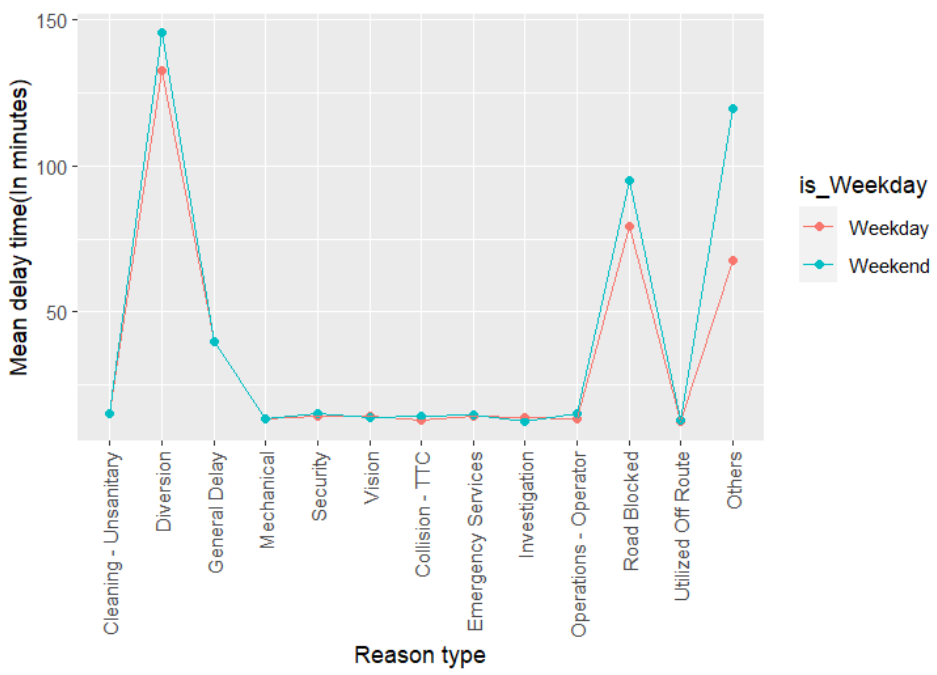


Figure 5.10

In the multiple bar plot in figure 5.9, delays due to operations - operator and mechanical problems have reported the majority of the cases in both weekdays and weekends which further shows that the reasons such as blocked roads and diversions which we identified to have wider distributions in figure 5.8 own a very less number of cases considerably proving that ,the extreme outliers that we identified in figure 5.1 , belong to them. When considering figure 5.9 we can see that the majority of the delay cases which belong to 0-50 time limit are resulted mainly due to delays of operations operator and mechanical problems.

In order to get an idea about how much of a mean delay time is distributed among these identified reasons, using figure 5.10 it is visible that due to extreme outliers observed in figure 5.1 with the less number of incidents corresponding to them, diversion and blocked roads report the higher mean delay times. As expected, the incidents which have caused majority of the cases report less mean delay times in both weekdays and weekends due to the small range of value distributions corresponding to them as observed in figure 5.8.

According to the below figure 5.11 , it shows that the delay cases in regular and limited service routes which function in a common range of 6am to 1am in both weekends and weekdays are due to the reasons namely delays due to operations operator and mechanical problems while other reasons are negligible comparatively as expected.

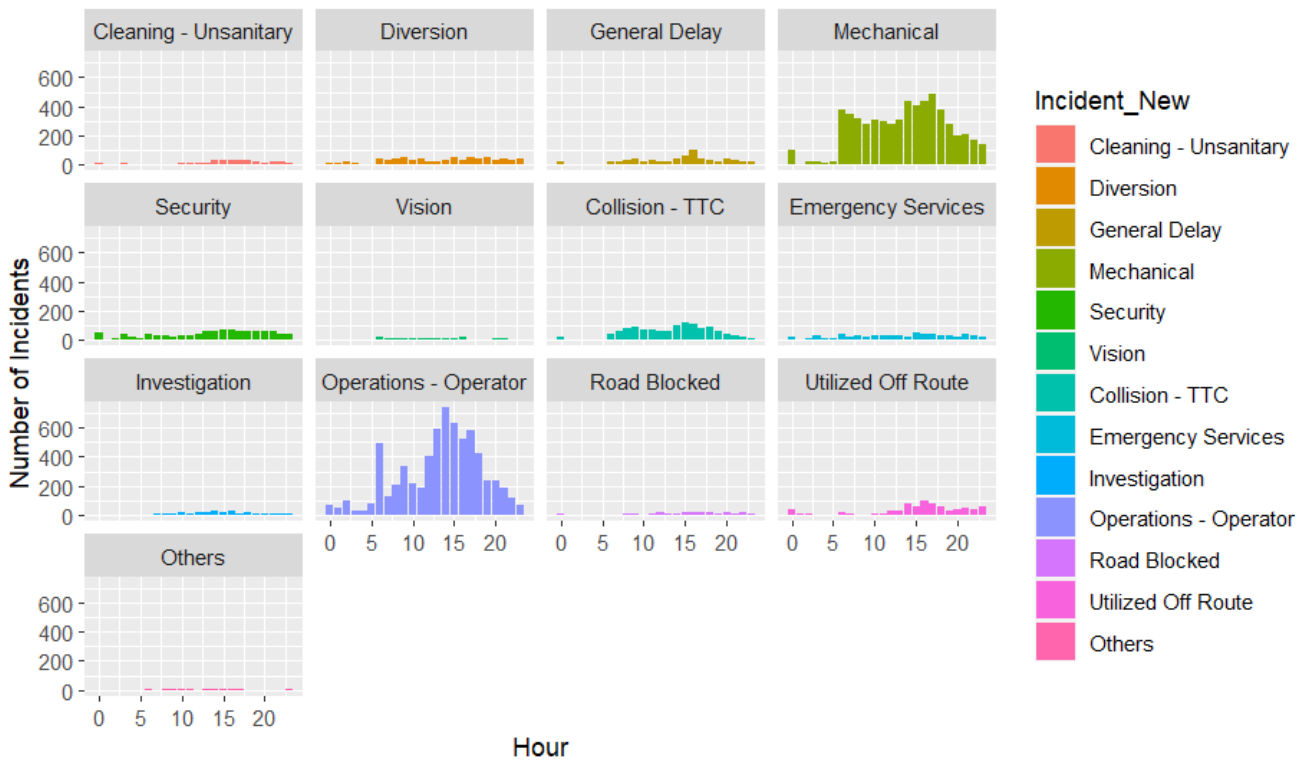


Figure 5.11

As a whole when we observe the distribution of the cases from January to June dominant numbers of cases are reported during the regular and limited service bus hours in every month verifying the big picture we obtained from the above plots further, as shown in figure 5.12.

Figure 5.12

Finally, the scatter plot shown in figure 5.13, which is the delay time plotted against time gap shows a perfect positive linear relationship between those 2 variables. Therefore, we can say that the time gap in minutes of a bus with the next bus scheduled shares a significant relationship to the delayed time of a bus in minutes, and those 2 variables are strongly correlated with a value of r=0.96. There might be other factors causing this perfect linear relationship such as the route type etc which are to be examined further.

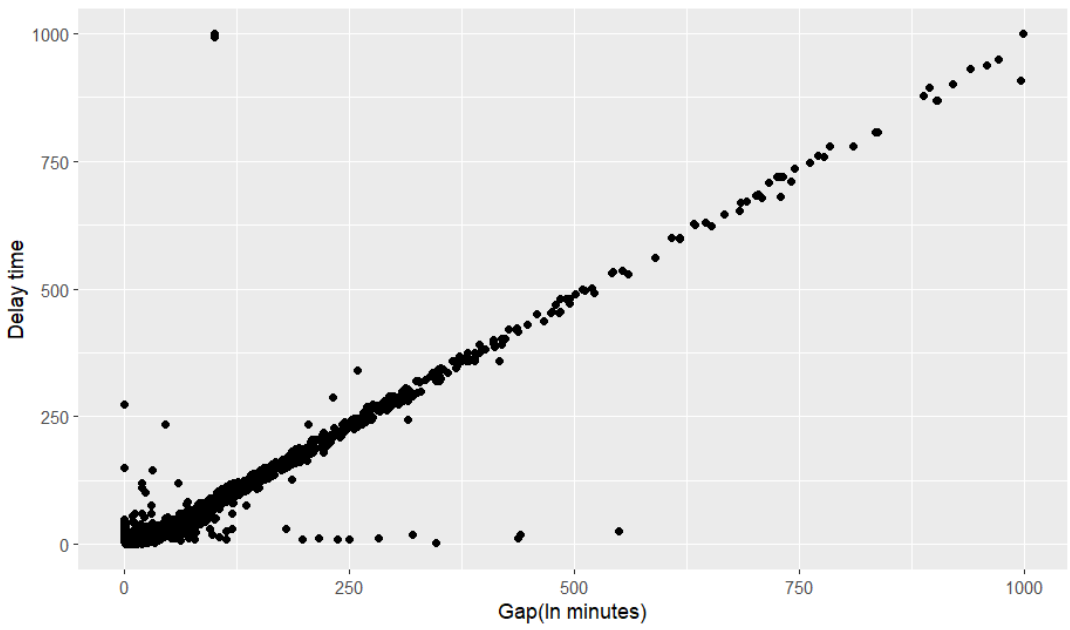


Figure 5.13

Multiple Correspondence Analysis

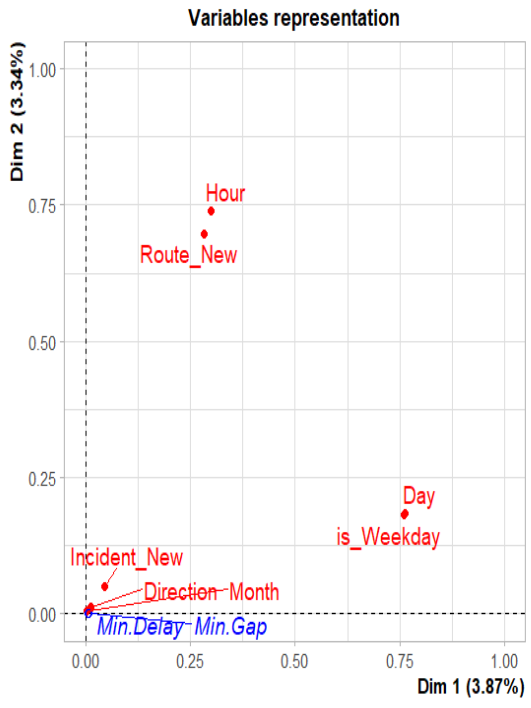


Figure 5.14

In order to get an understanding about how all the categorical variables are related with the response variable *"Min.Delay"* , we performed multiple correspondence analysis considering *"Min.Delay"* and *"Min.Gap"* as the two supplementary quantitative variables. As shown in figure 5.14 the overall result is not satisfactory as the most important first two dimensions only describe a proportion of nearly 7% out of the total variance of the categorical data. Furthermore, only *"Incident_New", "Direction", "Month"* are related with the response variable while *"Hour", "Route_New"* and *"Day", "is_Weekday"* pairs share distant relationships with the response being highly related to each other making these relationships contradictory to the practical scenario.

So we decided to perform One Way ANOVA to clarify these findings. But the normality of *"Min.Delay"* by the categories of the categorical variables is clearly violated as shown below in the Normal Q-Q plots in figure 5.15 .

Checking the normality of the response by the categories of the categorical variables

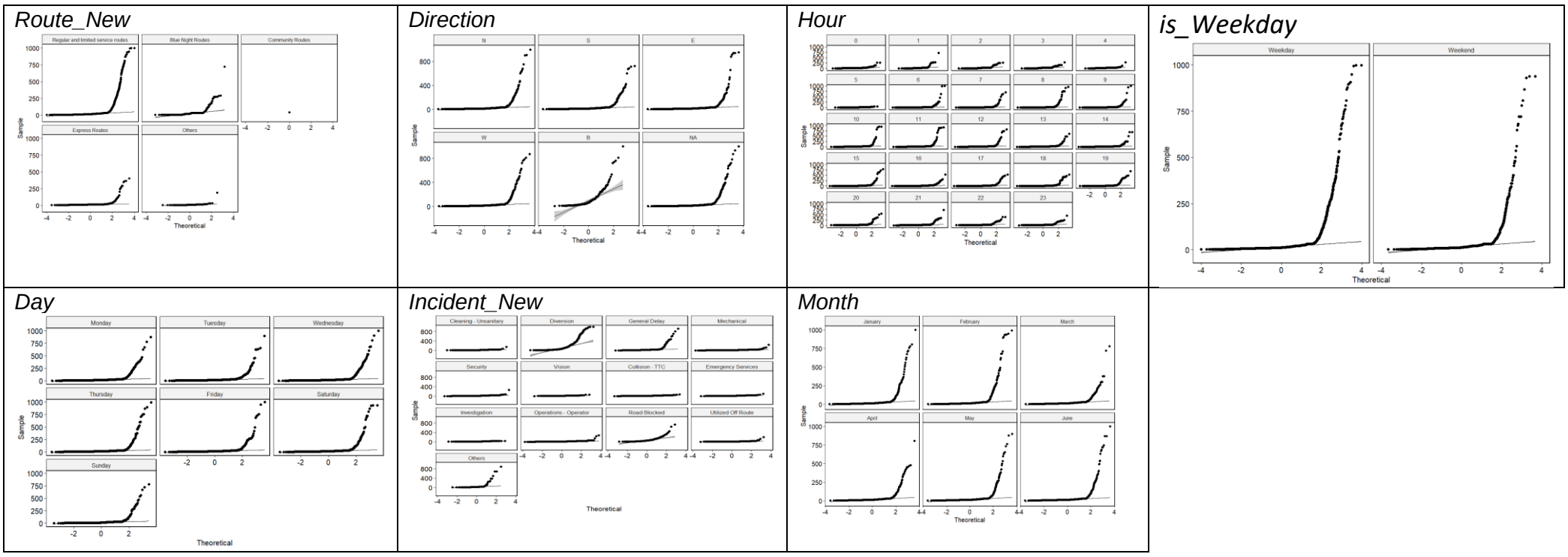


Figure 5.15

With the violation of assumptions of the One Way ANOVA we shifted to the Kruskal Wallis test.The test results are given below in table 5.1.

Variable	P- value
Route_New	<2.2e-16
Direction	<2.2e-16
Hour	< 2.2e-16
Is_Weekday	4.498e-06
Day	3.314e-05
Incident_New	< 2.2e-16
Month	0.2723

Table 5.1

Since all the p values of the Kruskal Wallis tests except that of Month are clearly significant at 5% significance level suggesting that there is a significant effect on the median of the response variable *"Min.Delay"* by the categories of all the categorical variables except *"Month"* . Here we have got some justifiable results than that of the multiple correspondence analysis as the categories of the variables of *"Route_New", "Hour", "Day"* and *"is_Weekday"* are affecting the median of the response while *"Month"* shares a distant association with the delay time as expected from the real world scenarios and the results of the previously conducted analysis .

6. Suggestions for a quality advanced analysis

- In the beginning of the descriptive analysis we identified that the delay time has a skewed distribution with 994 outliers. Therefore, transformations can be used in order to reduce the effect of outliers and the variance in data. There are different methods to transform the variables to Gaussian distribution. Some of the methods are: "square root function", "logarithmic function" and "Box Cox transformation".
- Since *"Min.Gap"* is the only one continuous predictor variable in the data set which we consider for further analysis while time variable is kept aside, we decided to perform tests to check whether there are associations between the continuous predictor and the categorical predictors and the associations between the categorical predictors. Firstly we identified Time gap distribution to be positively skewed with 843 outliers as shown below in figure 6.1.

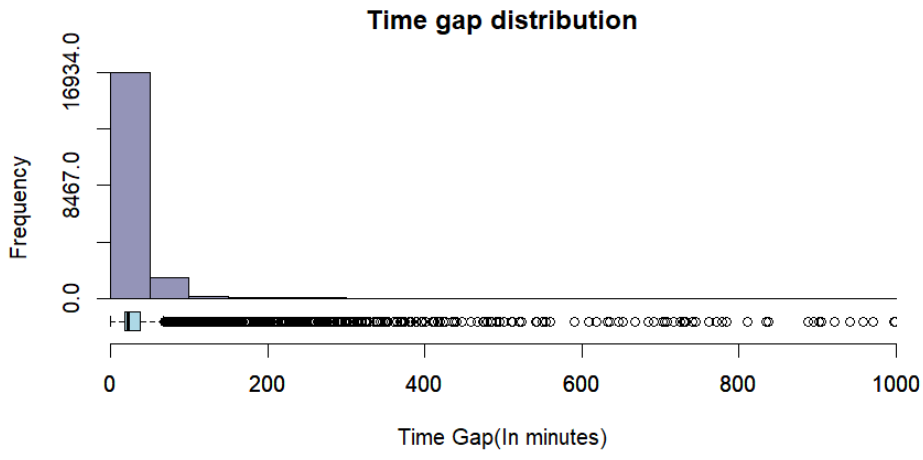


Figure 6.1

1)Route_New and Min.Gap

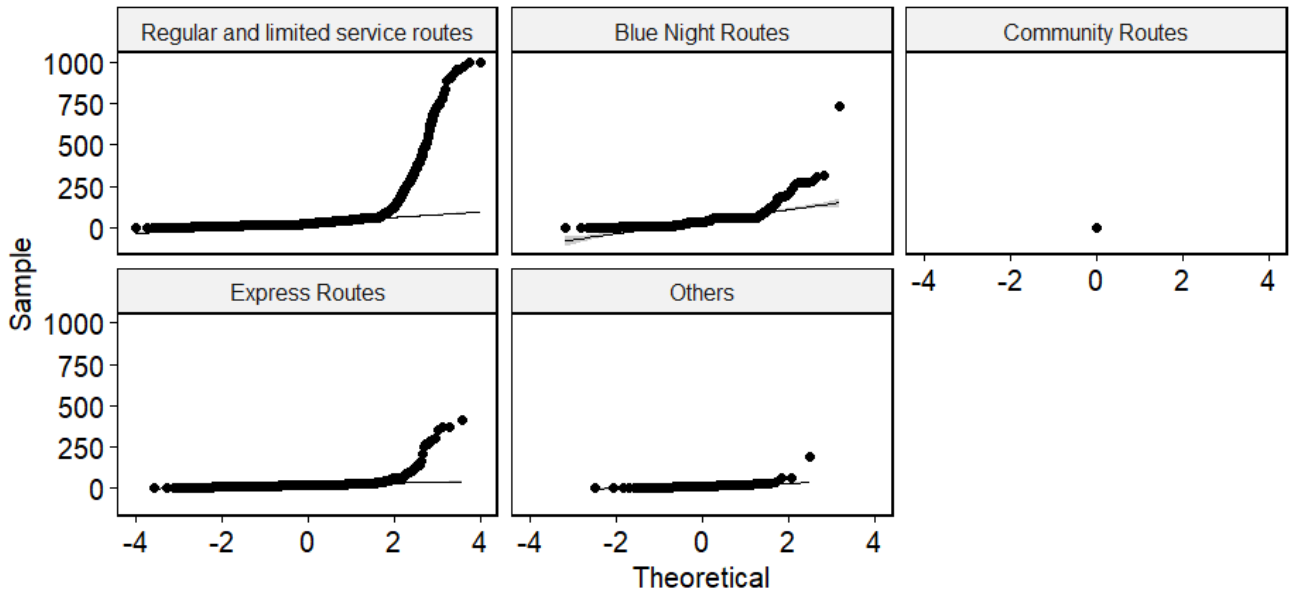


Figure 6.2

In order to perform an One Way ANOVA test to find whether there is an effect on the means of time gap grouped by "Route_New" ,we identified that the normality of the time gap variable by the categories of Route_New variable is not satisfied using the Normal Q-Q plots, which leads to the violation of the ANOVA assumption of normality of the continuous variable over the different groups, as shown in figure 6.2.

So, we decided to shift for the Kruskal Wallis test which is non-parametric and requires only the continuous variable observations to be independent over the categories of the categorical variable.

```
> result = kruskal.test(Min.Gap ~ Route_New,
+                       data = trainset)
> print(result)

        Kruskal-Wallis rank sum test

data:  Min.Gap by Route_New
Kruskal-Wallis chi-squared = 693.49, df = 4, p-value < 2.2e-16
```

From the above p - value we can conclude that at 5% significance level, there is a significant difference in medians of the "Min.Gap" categorized by "Route_New" which shows that there is an association between the two variables.

Also, we performed chi squared test between the Route_New and the Incident_New variables which resulted the following output.

```
> #chisquared test between Route-New and Incident-New
> chisq.test(trainset$Route_New, trainset$Incident_New, correct=FALSE)

        Pearson's Chi-squared test

data:  trainset$Route_New and trainset$Incident_New
X-squared = 591.24, df = 48, p-value < 2.2e-16
```

From the above output we can conclude that at 5% significance level, there is a significant association between these 2 categorical variables.

By the above existing relationships between those specific variables, we can suggest Ridge, Lasso and Elastic Net regression as the better approaches in fitting the model than the ordinary least squares method as it will result high variances in the parameter estimates. Random forest regression can also be used in this since it produces good predictions that can be understood easily. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm. Furthermore, there's another technique called "XGBoost" which is faster than Random Forest and can handle multiple issues simultaneously.

7. Appendix including R code and technical details

1.

```
data=read.csv("D:/3RD YEAR/SEMESTER II/ST3082/Project I Materials//ttc-bus-delay-data-2022.csv")
View(data)
head(data)
summary(data)
str(data)
library(dplyr)
library(stringr)
library(mgsub)
library(ggplot2)
library(corrplot)
library(psych)
library(packHV)
library(plyr)
library(moments)
library(tidyverse)
```

2.

```
##############################  Pre- processing and Feature Engineering  ###################################

#Removing duplicates
sum(duplicated(data))
data=distinct(data)

#Replacing "" with NA
data = replace(data, data=='', NA)
colSums(is.na(data))

#Removing records with delay=0
data=data[data$Min.Delay != 0, ]
nrow(data)

#Create Month column
data = data %>%
  mutate(Month = str_extract(Date, "[:alpha:]+")
         )%>%

  mutate(
    Month = ifelse( str_detect(Date, "Jan"),"January", Month)
  ) %>%

  mutate(
    Month = ifelse( str_detect(Date, "Feb"),"February", Month)
  ) %>%
  mutate(
    Month = ifelse( str_detect(Date, "Mar"),"March", Month)
  ) %>%
  mutate(
    Month = ifelse( str_detect(Date, "Apr"),"April", Month)
  ) %>%
  mutate(
    Month = ifelse( str_detect(Date, "May"),"May", Month)
  ) %>%
  mutate(
    Month = ifelse( str_detect(Date, "Jun"),"June", Month)
  )
```

3.

```
#Cleaning Direction column
table(data$Direction)
which(data$Direction=="/")
which(data$Direction=="2")
which(data$Direction=="6")
which(data$Direction=="O")
which(data$Direction=="I")
which(data$Direction=="J")
which(data$Direction=="Q")
data=data %>%  filter(!row_number() %in% c(78,19277,21705,6757,11606,154,4310,10421,14092))
nrow(data)

#Create Route_New column
x=c(table(data$Route))
which(data$Route=="RAD")
which(data$Route=="OTC")
data$Route[5703]=1000
data$Route[7435]=1000
data$Route[19824]=1000
data$Route[13156]=1001
Route_New=c()
for(i in 1:nrow(data)){
  if((!is.na(data$Route[i]))&& (as.numeric(substr(data$Route[i],1,4))>=7)&(as.numeric(substr(data$Route[i],1,4))<=189)){
    Route_New[i]="Regular and limited service routes"

  }else if((!is.na(data$Route[i]))&&(as.numeric(substr(data$Route[i],1,4))>299)&(as.numeric(substr(data$Route[i],1,4))<400)){
    Route_New[i]="Blue Night Routes"

  }else if((!is.na(data$Route[i]))&&(as.numeric(substr(data$Route[i],1,4))>399)&(as.numeric(substr(data$Route[i],1,4))<500)){
    Route_New[i]="Community Routes"

  }else if((!is.na(data$Route[i]))&&(as.numeric(substr(data$Route[i],1,4))>899)&(as.numeric(substr(data$Route[i],1,4))<1000)){
    Route_New[i]="Express Routes"

  }else if(!is.na(data$Route[i])){
    Route_New[i]="Others"
  }else{
    Route_New[i]=NA
  }
}
data=cbind(data,Route_New)
```

4.

```
#Create Hour column
Hour=c()
for(i in 1:length(data$Time)){
    Hour[i]=as.numeric(substr(data$Time[i],1,regexpr(":",data$Time[i])-1))
}
data=cbind(data,Hour)

#Create Incident_New column
table(data$Incident)
Incident_New=c()
for(i in 1:length(data$Incident)){
  if((data$Incident[i]=="Cleaning - Disinfection")|(data$Incident[i]=="Held By")|
     (data$Incident[i]=="Late Entering Service")){
      Incident_New[i]="Others"
  }else if(data$Incident[i]=="Road Blocked - NON-TTC Collision"){
      Incident_New[i]="Road Blocked"
  }else{
    Incident_New[i]=data$Incident[i]
}

}
data=cbind(data,Incident_New)

#Create is_Weekday column
is_Weekday=c()
for(i in 1:nrow(data)){
  if((data$Day[i]=="Saturday") | (data$Day[i]=="Sunday") ) {
    is_Weekday[i]="weekend"
    }else{
    is_Weekday[i]="Weekday"
  }
}
data=cbind(data,is_Weekday)
```

5.

```
#Cleaning Data
y=c()

for(i in 1:nrow(data)){
  if((!is.na(data$Route_New[i]))&&( (data$Route_New[i]=="Regular and limited service routes")&
                                    (data$is_Weekday[i]=="Weekday")&((data$Hour[i]>=1)&(data$Hour[i]<6)))){
    y=append(y,i)

  }else if((!is.na(data$Route_New[i]))&&( (data$Route_New[i]=="Regular and limited service routes")&
                                         (data$is_Weekday[i]=="weekend")&((data$Hour[i]>=1)&(data$Hour[i]<8)))){
    y=append(y,i)

  }else if((!is.na(data$Route_New[i]))&& ((data$Route_New[i]=="Blue Night Routes")&
                                         (data$is_Weekday[i]=="Weekday")&(data$Hour[i]<1)&(data$Hour[i]>=6))){
    y=append(y,i)

  }else if((!is.na(data$Route_New[i]))&& ((data$Route_New[i]=="Blue Night Routes")&
                                         (data$is_Weekday[i]=="weekend")&(data$Hour[i]<1)&(data$Hour[i]>=8))){
    y=append(y,i)

  }else if((!is.na(data$Route_New[i]))&& ((data$Route_New[i]=="Community Routes")&
                                         (data$is_Weekday[i]=="Weekday")& !(((data$Hour[i]>=6)&
                                                                             (data$Hour[i]<10))|((data$Hour[i]>=15)&(data$Hour[i]<19))))){
    y=append(y,i)

  }else if((!is.na(data$Route_New[i]))&& ((data$Route_New[i]=="Community Routes")&
                                         (data$is_Weekday[i]=="weekend"))){
    y=append(y,i)
  }else{
    y=append(y,"")
  }
}
y
y=y[y != ""]
data=data %>%  filter(!row_number() %in% as.numeric((y)))
nrow(data)
```

6.

```
#Factoring
data$Route_New = factor(data$Route_New,level=c("Regular and limited service routes","Blue Night Routes","Community Routes",
                                               "Express Routes","Others"))
data$Day = factor(data$Day,level=c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))
data$Month = factor(data$Month,level=c("January","February","March","April","May","June"))
data$Incident_New = factor(data$Incident_New,level=c("Cleaning - Unsanitary","Diversion","General Delay","Mechanical","Security","Vision",
                                                    "Collision - TTC","Emergency Services","Investigation","Operations - Operator",
                                                    "Road Blocked","Utilized Off Route","Others"))
data$Direction=factor(data$Direction,level=c("N","S","E","W","B"))
data$is_Weekday=factor(data$is_Weekday,level=c("Weekday","weekend"))

#Removing Vehicle Number column
data=subset(data,select=-Vehicle)

#Splitting the data in to training and testing sets
set.seed(100)
indexes=sample(1:nrow(data),0.2*nrow(data))
testset=data[indexes,]
trainset=data[-indexes,]
View(trainset)
View(testset)
head(trainset)
head(testset)
nrow(trainset)
nrow(testset)
#Imputing missing values
colSums(is.na(trainset))
#Mode function.
getmode = function(v) {
  uniqv = unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Calculate the mode of Route variable.
result1 = getmode(v=trainset$Route_New)
print(result1)
# Calculate the mode of Direction variable.
result2 = getmode(v=trainset$Direction)
print(result2)
```

7.

```
table(trainset$Direction)


z1=which(is.na(trainset$Route_New))

h1=c()
w1=c()
for(i in 1:length(z1)){
  h1[i]=trainset$Hour[z1[i]]
  w1[i]=trainset$is_Weekday[z1[i]]
}

z2=which(is.na(testset$Route_New))


h2=c()
w2=c()
for(i in 1:length(z2)){
  h2[i]=testset$Hour[z2[i]]
  w2[i]=testset$is_Weekday[z2[i]]
}
```

8.

```
#Descriptive Analysis

# 1) Distribution of Delay

#Histogram and Boxplot
options(repr.plot.width=12,repr.plot.height=7)
hist_boxplot(trainset$Min.Delay,main="Delay time distribution",col="#9494b8",xlab="Dealy time(In minutes)")
boxplot.stats(trainset$Min.Delay)$stats

#Numerical Summaries
#Mean of Delay time
mean(trainset$Min.Delay)

#Skewness and kurtosis
skewness(trainset$Min.Delay)#>1 positively skewed
kurtosis(trainset$Min.Delay)
x=which(trainset$Min.Delay %in% boxplot.stats(trainset$Min.Delay)$out)
length(x)


#Stacked bar chart with Direction Vs Number of Incidents
ggplot(trainset,aes(x=Route_New,fill=Direction))+
  geom_bar(stat='count',width=0.7,alpha=0.7)+
  scale_x_discrete(guide = guide_axis(angle = 90))+
  labs(x="Route Type",y="Number of Incidents")+
  theme(
    panel.grid.major=element_blank(),panel.grid.minor=element_blank(),
    panel.background=element_blank(),axis.line=element_line(colour="black")
  )
#Stacked bar chart Direction by total delay time
tg1=ddply(trainset,c("Route_New","Direction"),summarise,delay=sum(Min.Delay))
ggplot(tg1,aes(x=Route_New,y=delay,fill=Direction,label=delay))+
  geom_bar(stat="identity",width=0.3,alpha=0.7)+
    scale_x_discrete(guide = guide_axis(angle = 90))+
  labs(x="Route Type",y="Total Delayed time(In minutes)")+
  theme(
    panel.grid.major=element_blank(),panel.grid.minor=element_blank(),
    panel.background=element_blank(),axis.line=element_line(colour="black")
  )
```

9.

```
aggregate(Min.Delay~ Route_New, data = trainset, sum)


#Line plot of Mean delay time by hour and is_Weekday
tg2=ddply(trainset,c("Hour","is_Weekday"),summarise,delay=mean(Min.Delay))
ggplot(tg2,aes(x=Hour,y=delay,colour=is_Weekday,group=is_Weekday))+
  geom_point()+
  geom_line()+
  labs(x="Hour",y="Mean delay time(In minutes)")


#Dot plot of delay time by hour and is_Weekday
ggplot(data=trainset,aes(x=Hour,y=Min.Delay,fill=is_Weekday,color=is_Weekday))+
  geom_point()+
  theme_minimal()+
  labs(x="Hour",y="Delay time(In minutes)")

#Boxplot of delay times by Route and is_Weekday
ggplot(trainset, aes(x=factor(Route_New), y=Min.Delay, color = is_Weekday))+
  geom_boxplot()+
  scale_x_discrete(guide = guide_axis(angle = 90))+
  labs(x="Route type",y="Delayed time(In minutes)")+
  theme(
    panel.grid.major=element_blank(),panel.grid.minor=element_blank(),
    panel.background=element_blank(),axis.line=element_line(colour="black")
  )

#Multiple bar chart of Number of incidents by route and day
ggplot(trainset, aes(Route_New, fill = Day)) +
  geom_bar(stat="count",position = "dodge") +
  scale_x_discrete(guide = guide_axis(angle = 90))+
  scale_fill_brewer(palette = "Set1")+
  labs(x="Route type",y="Number of Incidents")
```

10.

```
#Boxplot of delay time by incidents
ggplot(trainset,aes(x=Incident_New,y=Min.Delay))+
  geom_boxplot(alpha=0.7,fill="blue")+
  coord_flip()+
  labs(x="Reason type",y="Delayed time(in minutes)")+
  theme(
    panel.grid.major=element_blank(),panel.grid.minor=element_blank(),
    panel.background=element_blank(),axis.line=element_line(colour="black")
  )
#Multiple bar chart of Number of incidents by Incidents and is_Weekday
ggplot(trainset, aes(Incident_New, fill = is_Weekday)) +
  geom_bar(stat="count", position = "dodge") +
  scale_x_discrete(guide = guide_axis(angle = 90))+
  scale_fill_brewer(palette = "Set1")+
  labs(x="Reason type",y="Number of Incidents")


#Line plot of mean delay time by incident type and is_Weekday
tg3=ddply(trainset,c("Incident_New","is_Weekday"),summarise,delay=mean(Min.Delay))
ggplot(tg3,aes(x=Incident_New,y=delay,colour=is_Weekday,group=is_Weekday))+
  geom_point()+
  geom_line()+
  scale_x_discrete(guide = guide_axis(angle = 90))+
  labs(x="Reason type",y="Mean delay time(In minutes)")

#Summary plot
ggplot(trainset, aes(x= Hour,fill=Incident_New))+
  geom_bar(stat="count")+
  labs(x="Hour",y="Number of Incidents")+
  facet_wrap(~Incident_New)
```

11.

```
#Multiple bar chart of Numbere of incidents by month
RBH=c()
for(i in 1:nrow(trainset)){
  if(trainset$Route_New[i]=="Regular and limited service routes"){
    RBH[i]="Regular and limited service Hours"
  }else{
    RBH[i]="Others"
  }

}
trainset=cbind(trainset,RBH)
trainset$RBH=factor(trainset$RBH,level=c("Regular and limited service Hours","Others"))
ggplot(trainset, aes(Month, fill = RBH)) +
  geom_bar(stat="count", position = "dodge") +

  scale_x_discrete(guide = guide_axis(angle = 90))+
  scale_fill_brewer(palette = "Set2")+
  labs(x="Month",y="Number of incidents")


#continuous variables
ggplot(trainset, aes(x=Min.Gap, y=Min.Delay)) +
  geom_point() +
  labs(x="Gap(In minutes)",y="Delay time")
#correlation of gap and the delay time variables
corr.test(trainset$Min.Delay,trainset$Min.Gap)
```

12.

```
##################################### MULTIPLE CORRESPONDANCE ANALYSIS #################

library(FactoMineR)
trainset2=subset(trainset,select=-c(Date,Time,Route,Location,Incident))
View(trainset2)
nrow(trainset2)
cats = apply(trainset2, 2, function(x) nlevels(as.factor(x)))
res=MCA(trainset2,quanti.sup =c(2,3))
```

13.

```
#Checking the normality assumption
ggqqplot(trainset, "Min.Gap", facet.by = "Route_New")
ggqqplot(trainset, "Min.Delay", facet.by = "Route_New")
ggqqplot(trainset, "Min.Delay", facet.by = "Direction")
ggqqplot(trainset, "Min.Delay", facet.by = "Hour")
ggqqplot(trainset, "Min.Delay", facet.by = "is_Weekday")
ggqqplot(trainset, "Min.Delay", facet.by = "Day")
ggqqplot(trainset, "Min.Delay", facet.by = "Incident_New")
ggqqplot(trainset, "Min.Delay", facet.by = "Month")

#Since ANOVA normality assumptions are violated we are using Kruskal wallis test
# Performing Kruskal-Wallis test
result = kruskal.test(Min.Gap ~ Route_New,
                      data = trainset)
print(result)
kruskal.test(Min.Delay ~ Route_New,
             data = trainset)
kruskal.test(Min.Delay ~ Direction,
             data = trainset)
kruskal.test(Min.Delay ~ Hour,
             data = trainset)
kruskal.test(Min.Delay ~ Incident_New,
             data = trainset)
kruskal.test(Min.Delay ~ is_Weekday,
             data = trainset)
kruskal.test(Min.Delay ~ Day,
             data = trainset)
kruskal.test(Min.Delay ~ Month,
             data = trainset)
```

14.

```
#Suggestions for Advanced Analysis
#Examining the relationship of Route_New on Gap variable
#Checking the ANOVA assumptions
library(ggpubr)
library(rstatix)
#Time Gap distribution and Outlier detection
options(repr.plot.width=12,repr.plot.height=7)
hist_boxplot(trainset$Min.Gap,main="Time gap distribution",col="#9494b8",xlab="Time Gap(In minutes)")
x=which(trainset$Min.Gap %in% boxplot.stats(trainset$Min.Gap)$out)
length(x)
#Checking the normality assumption
ggqqplot(trainset, "Min.Gap", facet.by = "Route_New")

#Since ANOVA normality assumptions are violated we are using Kruskal wallis test
# Performing Kruskal-Wallis test
result = kruskal.test(Min.Gap ~ Route_New,
                      data = trainset)
print(result)
#chisquared test between Route-New and Incident-New
chisq.test(trainset$Route_New, trainset$Incident_New, correct=FALSE)
```