Data Analysis and Visualization in R (IN2339)

Case Study

Mahima Arunkumar, Mariem Aboushawareb, Merve Kılıçarslan, Hoda Hamdy

2022-01-23

Motivation for Air Quality Dataset

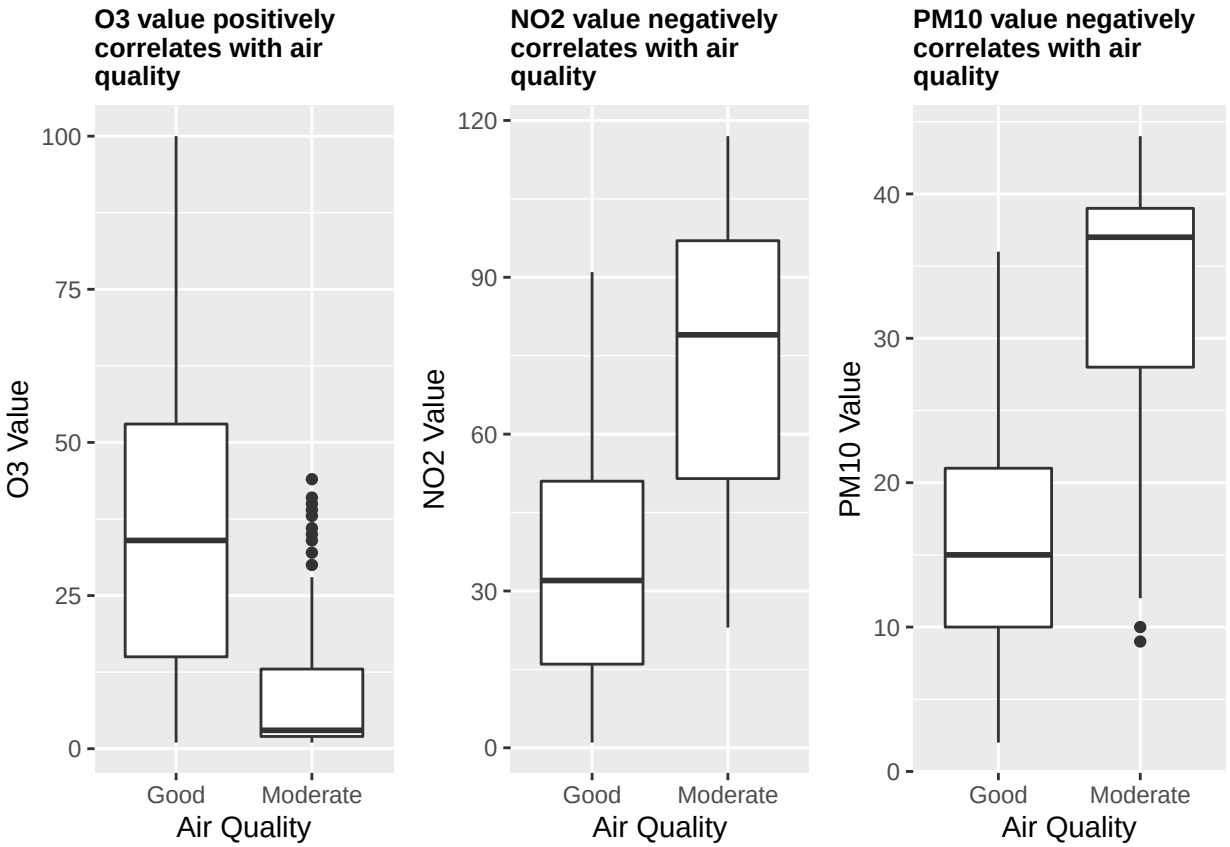
The goal of this analysis is to have a better understanding of contributing factors of the air quality in Barcelona in November 2017.

Data Preparation

We used "air_quality_Nov2017" dataset to examine the correlation between the O3, NO2 and PM10 values and air quality. We eliminate the data inputs which had undefined values for these parameters using "filter" function. We also removed NA values.

Hypothesis: Air quality is correlated with O3 Value. Air quality is correlated with NO2 Value. Air quality is correlated with PM10 Value.

```
#scatter plots to see the correlation between values and air quality
q_o <- ggplot(air, aes(`Air Quality`, `O3 Value`))+
  geom_boxplot()+ ggtitle(str_wrap("O3 value positively correlates with air quality", width = 25))+
  theme(plot.title = element_text(size = 10, face = "bold"))

n_o <- ggplot(air, aes(`Air Quality`, `NO2 Value`))+
  geom_boxplot()+ ggtitle(str_wrap("NO2 value negatively correlates with air quality", width = 25))+
  theme(plot.title = element_text(size = 10, face = "bold"))

p_o <- ggplot(air, aes(`Air Quality`, `PM10 Value`))+
  geom_boxplot()+ ggtitle(str_wrap("PM10 value negatively correlates with air quality", width = 25))+
  theme(plot.title = element_text(size = 10, face = "bold"))

grid.arrange(q_o,n_o, p_o,nrow=1)
```



From the plots we see no overlap between the medians. It also seems that NO2 and PM10 values are negatively correlated with the air quality where O3 is positively correlated. We verify these correlations with a statistical test.

```
#claim supported by the Wilcoxon test
w1 <-wilcox.test(`O3 Value` ~ `Air Quality`, data=air)
w2 <-wilcox.test(`NO2 Value` ~ `Air Quality`, data=air)
w3 <-wilcox.test(`PM10 Value` ~ `Air Quality`, data=air)
w1$p.value
```

```
## [1] 3.064943e-31
```

```
w2$p.value
```

```
## [1] 2.874347e-47
```

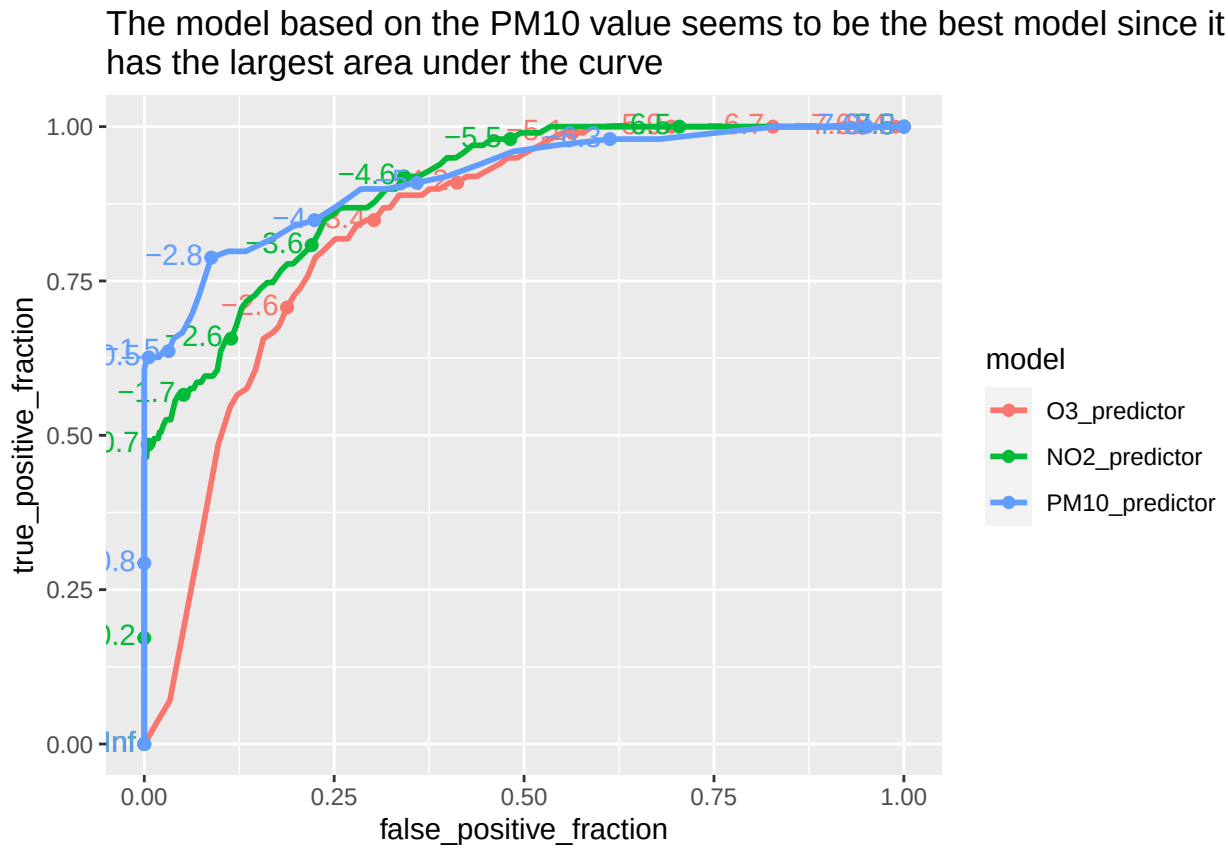
```
w3$p.value
```

```
## [1] 1.606485e-56
```

We chose a Wilcoxon test for the following reasons: 1. Testing a correlation between a binary and a continuous variable. 2. No assumption for the distributions of O3, NO2, PM10 values P-values generated by the Wilcoxon test are very small, hence correlations are significant (reject H0).

Based on the previous analysis we think that O3, NO2, PM10 values would be good variables to fit a logistic regression model to classify air quality based on.

```
air_quality_dt <- qua_2017['Air Quality' != '--']
air_quality_dt <- air_quality_dt[!is.na('O3 Value') & !is.na('NO2 Value') & !is.na('PM10 Value')]

air_quality_dt[, air_quality_modified:= (as.numeric(factor('Air Quality'))-1)]

model_1 <- glm('air_quality_modified' ~ 'O3 Value', data = air_quality_dt, family = "binomial")

air_quality_dt[,'O3_predictor' := predict(model_1)]

model_2 <- glm('air_quality_modified' ~ 'NO2 Value', data = air_quality_dt, family = "binomial")

air_quality_dt[,'NO2_predictor' := predict(model_2)]

model_3 <- glm('air_quality_modified' ~ 'PM10 Value', data = air_quality_dt, family = "binomial")

air_quality_dt[,'PM10_predictor' := predict(model_3)]

melted_air_quality <- melt(air_quality_dt, measure.vars = c('O3_predictor', 'NO2_predictor', 'PM10_predi

ggplot(melted_air_quality, aes(m=prediction, d=air_quality_modified, color = model)) + geom_roc() + ggt
```



Motivation for Accidents Dataset

In 2017, there were more than 230 thousand registered vehicles in the city of Barcelona. The goal of this analysis is to better understand the accidents during 2017. The main goal is to extract some insights that

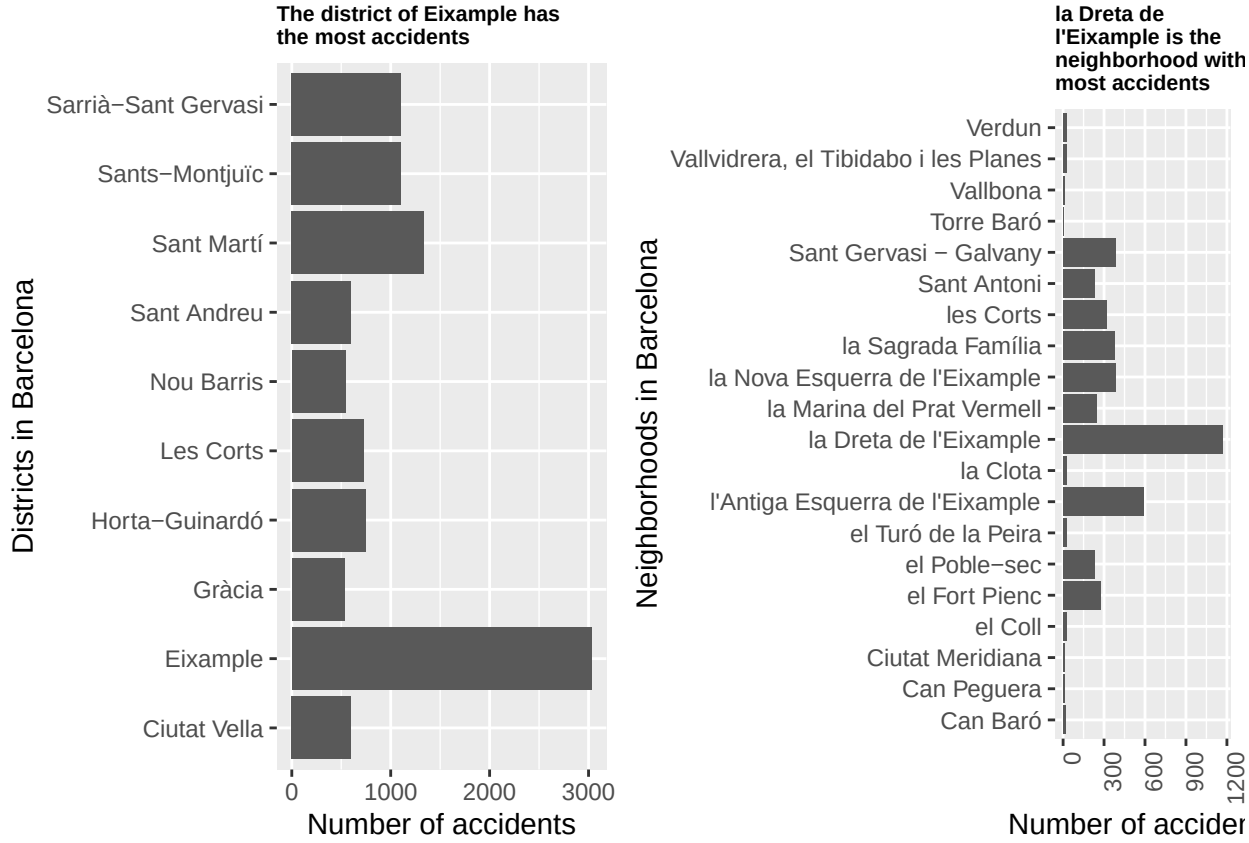
can be useful to reduce accidents in Barcelona.

#Data Preparation For this part of the case study "accidents_2017" dataset is used. To eliminate the unknown values for the Districts filter function is used. Moreover, neighborhood data is used for plot however there are too many neighborhood to consider hence only the hightest and lowest 10 neighborhoods in accidents count.

Claim: We expect the number of accidents to change according to the day of the week, part of the day and districts as a consequence of human behavior (work days, driving late, etc.).

We used bar plots to investigate if the number of accidents varies according to districts.

```
#Plot
districts_plot <- ggplot(data=districts, aes(x='Number of accidents', y = 'Districts in Barcelona')) +
  geom_histogram(stat='identity', breaks=10, bins = 10, cex.names=0.5)+ ggtitle(str_wrap("The district
  theme(plot.title = element_text(size = 8, face = "bold"))
#Neighborhood Plot
neighborhood_plot <- ggplot(data=neighborhood_top_ten_least_ten, aes(Accidents_Count, Neighborhood)) +
  geom_bar(stat='identity')+ylab("Neighborhoods in Barcelona") + xlab("Number of accidents")+  ggtitle(
  theme(plot.title = element_text(size = 8, face = "bold"), axis.text.x=element_text(angle=90, hjust=1))
grid.arrange(districts_plot, neighborhood_plot, ncol=2 )
```

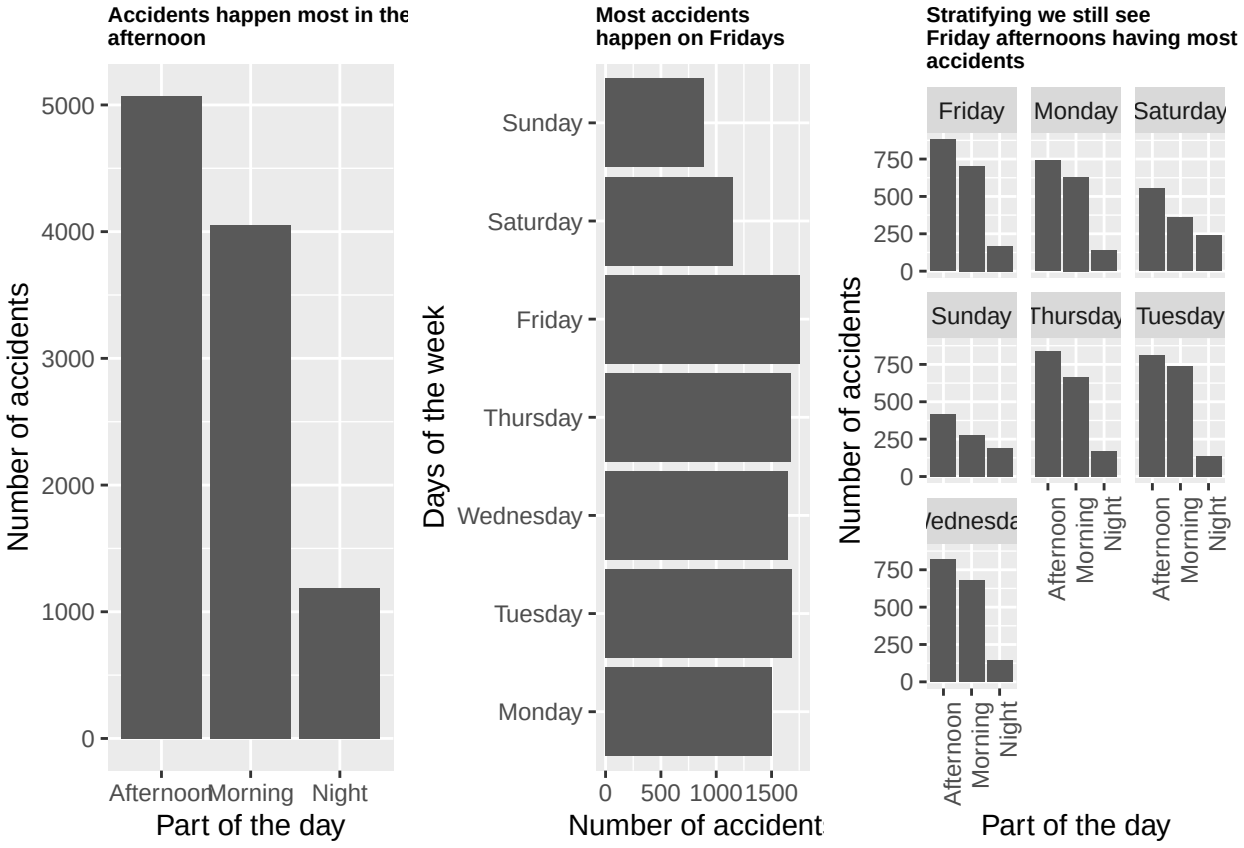


From this plot we can see that the district with significantly the largest number of accidents is Eixample. So what neighborhood had the most accidents? We look at the top 10 neighborhoods and least 10 neighborhoods with accidents. The least neighborhood with accidents being Torre Baró with only 7 accidents and the biggest one being la Dreta de l'Eixample with 1167 accidents.

Moving from space to time, when did these accidents happen the most? As a claim we expect to have more accidents in the weekdays, since more people drive frequently due to work. Secondly, we expect to have least

number of accidents at the night time when people sleep.

```
weekdays_plot <- ggplot(data=weekday, aes(y='Days of the week', x = 'Number of accidents')) +
  geom_bar(stat='identity', breaks=10, bins = 10, cex.names=0.5) + ggtitle(str_wrap("Most accidents happ

part3_plot<-ggplot(accidents_2017, aes('Part of the day')) + geom_bar() +
  labs(y='Number of accidents') + ggtitle(str_wrap("Accidents happen most in the afternoon", width = 28)

pat_facet_plot <- ggplot(accidents_2017, aes('Part of the day')) + geom_bar() + facet_wrap(~Weekday) +
  labs(y='Number of accidents') + ggtitle(str_wrap("Stratifying we still see Friday afternoons having mo
  theme(plot.title = element_text(size = 8, face = "bold"), axis.text.x=element_text(angle=90, hjust=1))
#part3_plot
grid.arrange(part3_plot,weekdays_plot, pat_facet_plot, nrow=1)
```



Weekdays have more accidents compared to weekends. Friday have the largest number of accidents and Sundays the least. We see that night time has the least number of accidents even if we facet by days.

Conclusion

From the air quality data we saw from our analysis that air quality is affected by values of O3, NO2 and PM10. We also saw that PM10 is the best estimator of air quality.

For the accidents data, we were able to pinpoint that Friday afternoons had the most accidents where Sundays had the least especially at night. However this might be because the number of people driving at night is significantly less than at other times. We also observed that Eixample had the most accidents, specifically in the neighborhood of la Dreta de l'Eixample.