Data Analysis and Visualization in R (IN2339)

Case Study

Mujtaba Shahid Faizi; Amélie Claus; Kalina Damyanova; Philipp Zent

2022-01-22

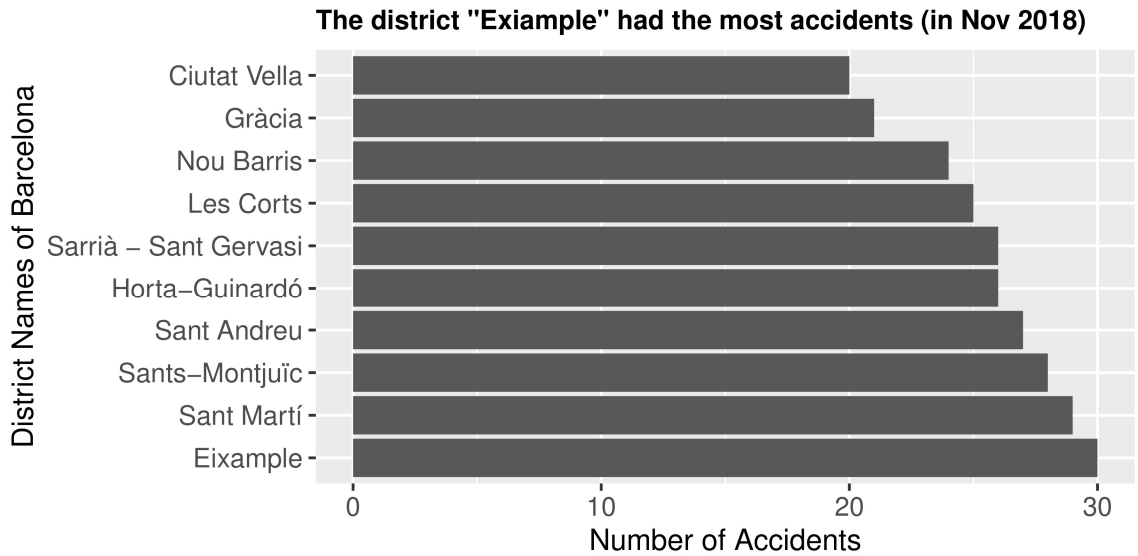
MOTIVATION

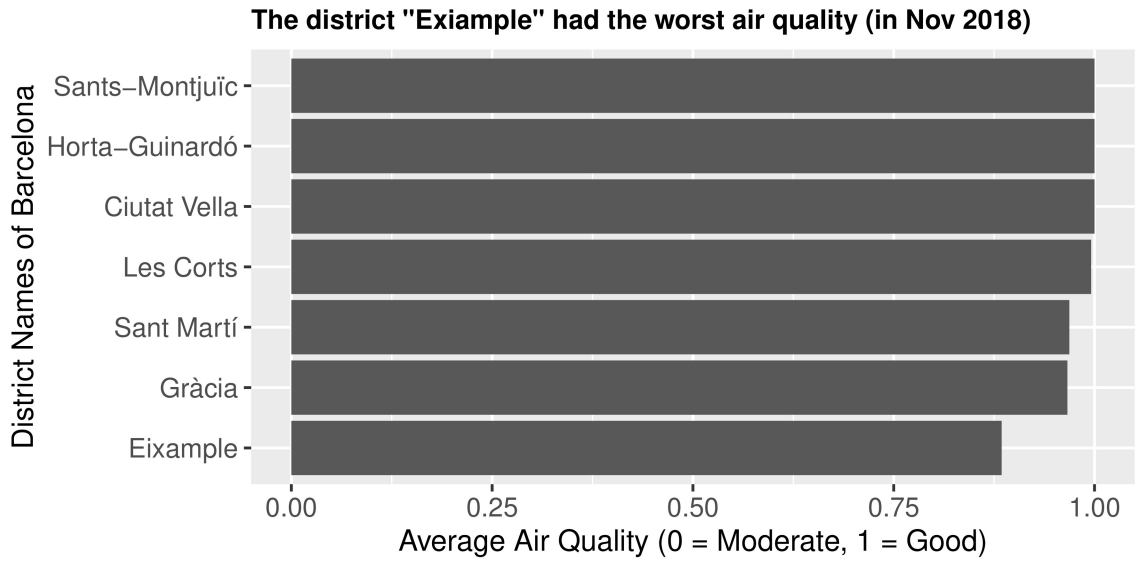
According to WHO, car accidents are the leading cause of death, especially for children and young adults aged 5-29 years. The goal of this case study is to have a better understanding of the accidents that occurred during November 2018 in Barcelona. More precisely, we want to see if there is an association between the number of accidents and air quality through careful analysis and by taking other variables into account. We also aim to predict the air quality based on measurements of compounds detected in the air of Barcelona.

DATA PREPERATION

We utilized these datasets for the data preparation steps: air_quality_Nov2017.csv, air_stations_Nov2017, accidents_2017, and transports.csv (although the file names say 2017, the actual date column inside the accidents' dataset say 2018 instead. So, we assumed the datasets to be taken from 2018). Unnecessary chunk codes are ommited in this compiled pdf-file.

We prepared the final accidents, air_quality, and transports datasets to be merged/joined with each other for the data analysis section. This preparation included correcting the district names and discarding missing data. Then, we corrected the dates in the right format. We converted categories of air quality: good, moderate into 1,0 respectively. The number of accidents and mean of air quality per date (days of the month November 2018) and per district were calculated. Similarly, the number of transports in each district was also calculated. Finally, the three datasets were joined/merged on the base of district for the analysis. They were also joined/merged on the base of date. These final datasets were then used for our data analysis.





DATA ANALYSIS

Our Claim: there exists an association between accidents and air quality.

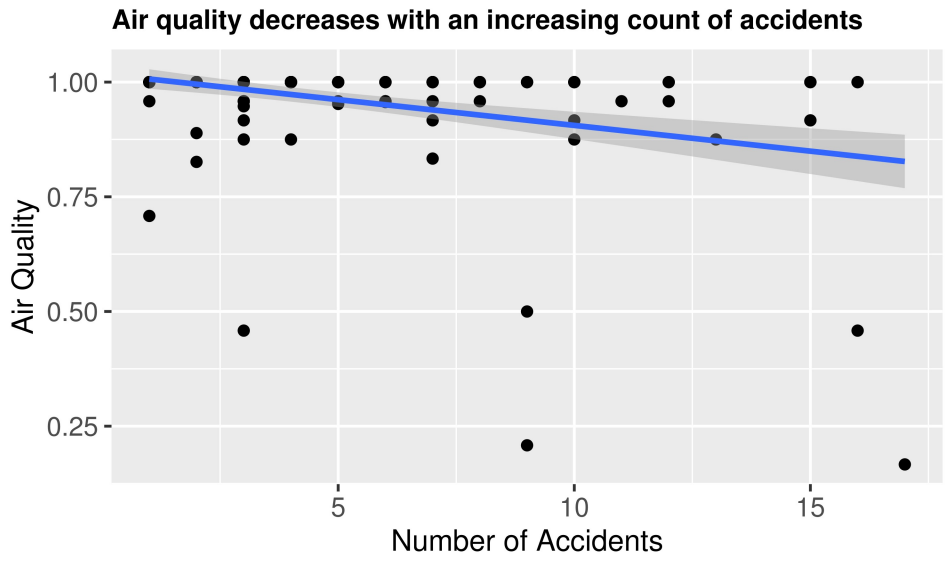
Statistical test: test for correlation between two quantitative variables i.e. no. of accidents and air quality. No assumption is made for the distribution, so we chose the spearman test.

```
cor_value = cor.test(accidents_with_air_quality_dt$Accidents,
                     accidents_with_air_quality_dt$Air_Quality,
                     method = "spearman")
cor_value
```

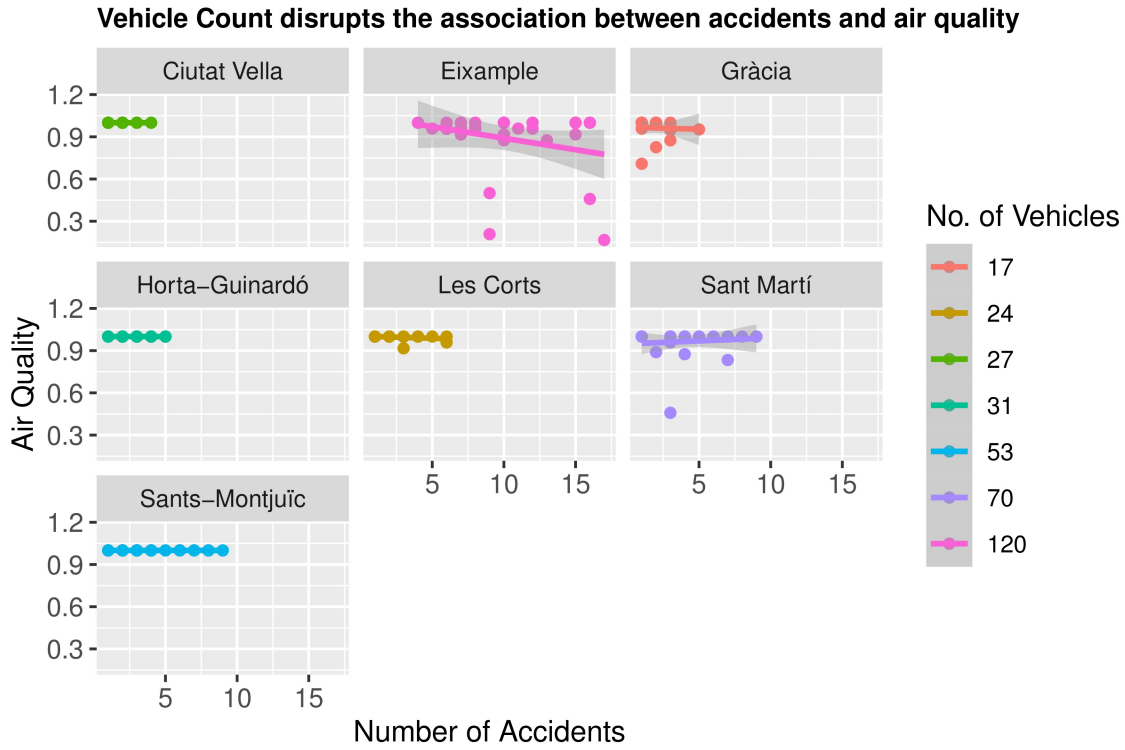
```
##
##  Spearman's rank correlation rho
##
## data:  accidents_with_air_quality_dt$Accidents and accidents_with_air_quality_dt$Air_Quality
## S = 1245113, p-value = 3.835e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.3026098
```

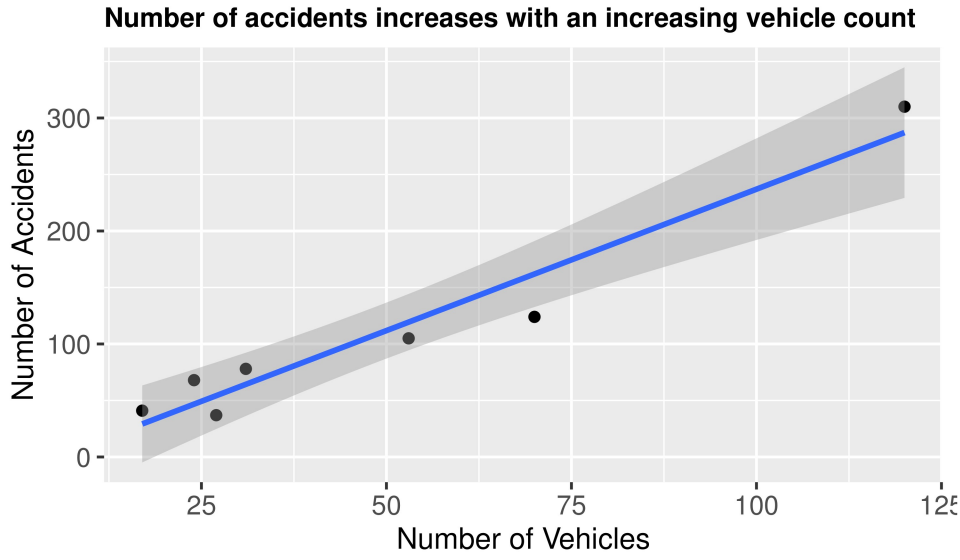
The p value is less than 5%, which implies that there is an association between the two variables (which statistically supports our claim)

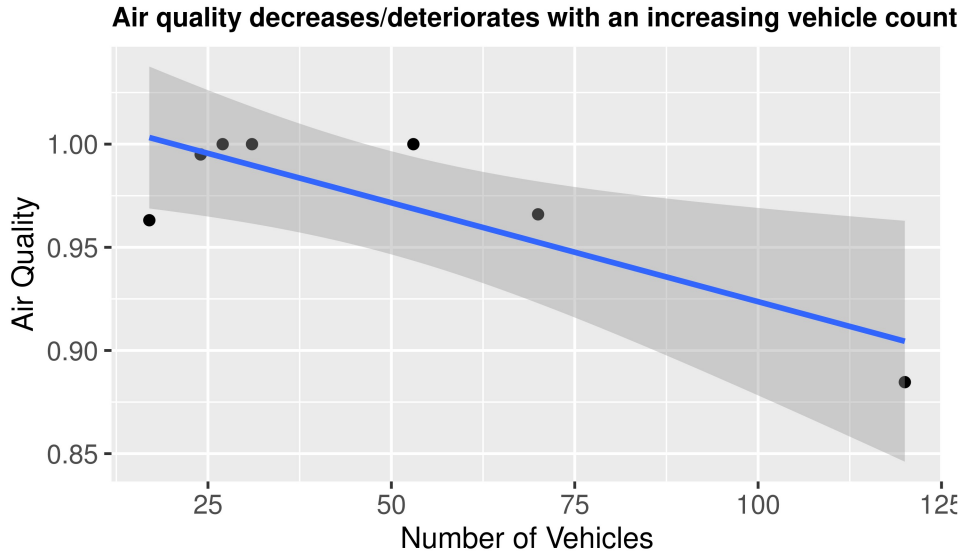
The graph below visualizes our statistically supported claim:



Now, we introduce a third variable into the situation i.e. "number of vehicles in each district".
Diagnostic plot: we construct a linear graph between accidents and air quality, facetted by vehicle count





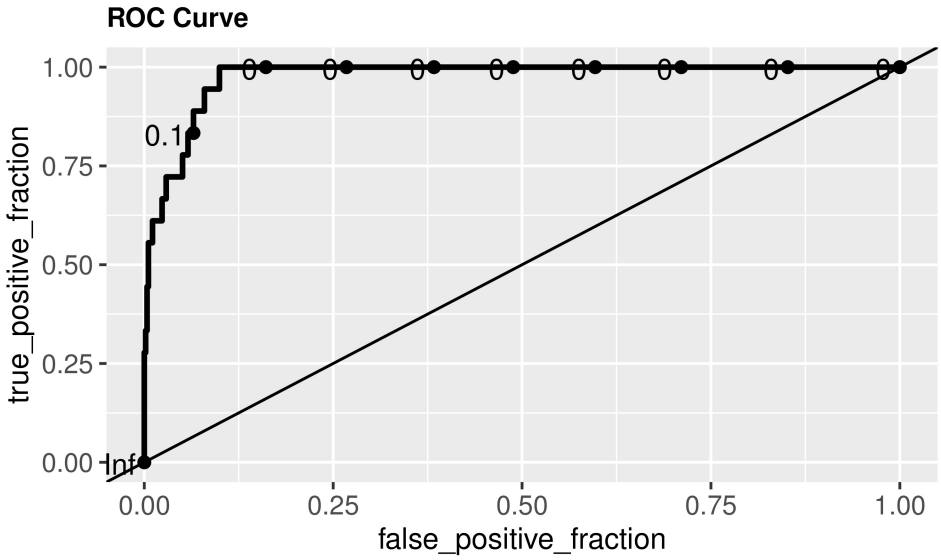


Our new claims are: no. of vehicles is associated with no. of accidents and air_quality. The above demonstrative plots support these claims. This suggests a common cause relationship between the three variables i.e. no. of vehicles causes a positive effect on no. of accidents, and no. of vehicles causes a negative effect on air quality. These new associations seem logical because we can assume that the higher the transport density, the higher the population density, the more car traffic there is, which increases the number of accidents and degrades the air quality.

In addition to controlling for confounding factors, we do a prediction of the quality of air based on some measured air compounds.

```
summary(glm(air_quality_value ~ O3 + NO2 + PM10,data = train,family = binomial))
##In reality, the air quality status (0 i.e. "good" or 1 i.e. "moderate") is established by
#measuring the quantity of PM10, O3 and NO2. However, according to the
#p-values, only NO2 and PM10 significantly associates (p<5%) with the response
#variable and thus, we used only these 2 variables for our prediction task.
glm.fit <- glm(air_quality_value ~ NO2 + PM10,data = train,family = binomial)
glm.probs <- predict(glm.fit, newdata = test,type = "response")
```

```
glm.pred <- ifelse(glm.probs > 0.5, 1, 0) # cutoff = 0.5
```




```
##    air_quality_value         V1
## 1:                 0 0.96847636
## 2:                 1 0.03152364
```

We checked the imbalance between the classes in our dataset, and as there is a huge imbalance in our dataset (~ 97% of class 0), we use the AUC to capture the overall performance of our model. This metric is not affected by an imbalance in classes.

```
calc_auc(ggroc)
```

```
##   PANEL group       AUC
## 1     1    -1 0.9757886
```

AUC = 0.9757886. Higher the AUC, better the model is at predicting 0 classes as 0 and 1 classes as 1. With an AUC of ~98%, we can conclude that the model can very well distinguish between the two classes with certainty.

CONCLUSION

According to the above analysis, it is clear that we see a pattern in the accidents - air quality relationship when controlling for the number of vehicles i.e. higher accident counts and lower air quality counts represent snapshots with high traffic intensity. We can speak of a common cause relationship for Z (number of vehicles) which causes both A (car accidents) and B (air quality). This has invalidated our first claim, that there exists an association or causation relationship between accidents and air quality.

In addition, we performed a prediction task to predict the air quality status. We get good prediction performance (i.e. AUC of ~98%) basing the Logistic Regression on PM10 and NO2 quantities.