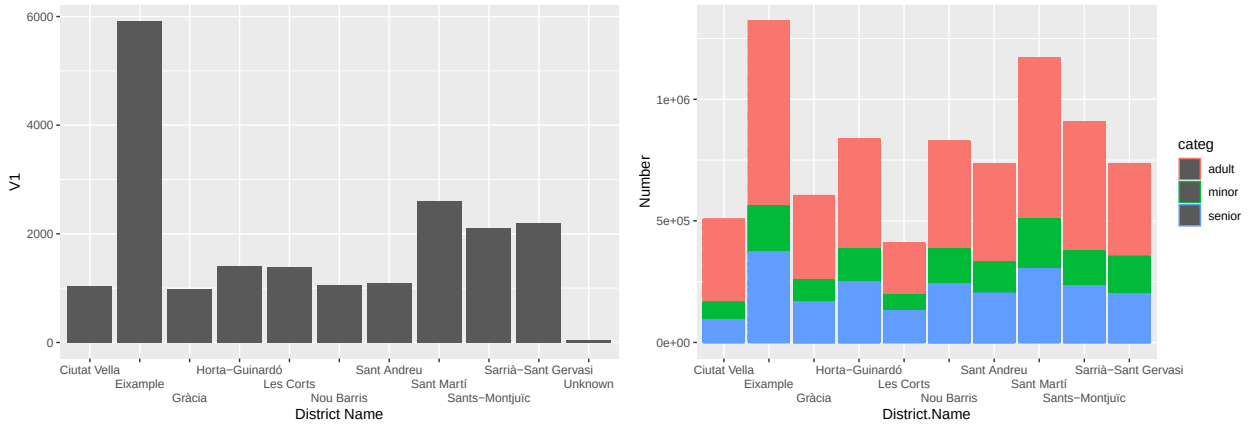
Motivation

We want to investigate the accidents in Barcelona. Our goal is to determine some key factors that are affecting accidents, the frequency and the severity of them, and we will examine that factors to see if there are some hidden confounding effects. Later, we will come up with some hypotheses and try to test them statistically. We will support our findings with plots to make them more meaningful.

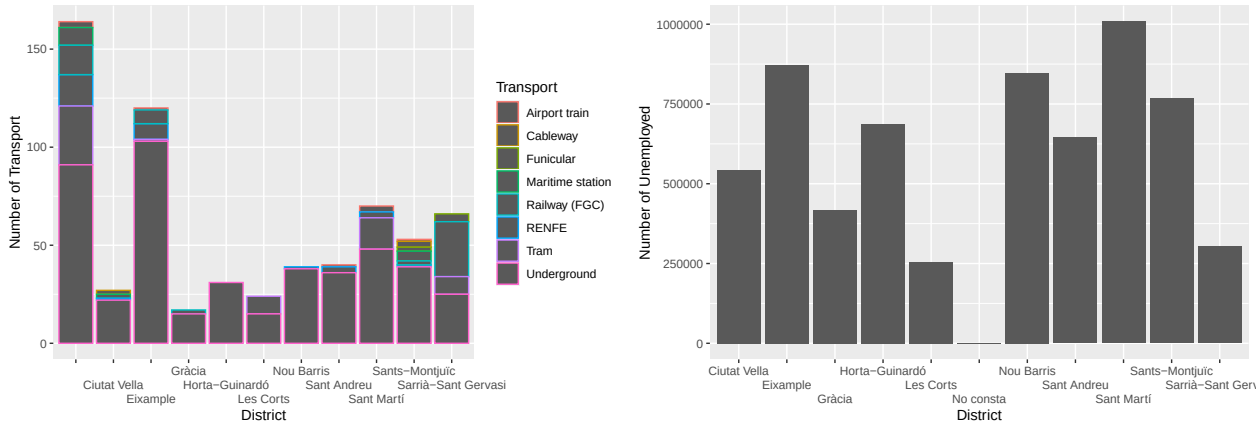
Data Analysis

We first plot the accidents in Barcelona by District. Next, we separate the population into minors, adults, and seniors to check if there is a larger population of driving age in Eixample.



As far as we can see, the difference between the population of Eixample and any other districts is not that high, but Eixample has a significantly higher number of accidents compared to others. Meanwhile, there is not much difference in the proportions of the working age population.

We then analyzed the transport types and unemployment of each district to see if there is any clear reason.



From the graphs we see that although Eixample has the highest number of accidents,the number of transports is not the highest in this district. From the transport types and unemployment categorized by district, there seems to be no clear reason for either to affect accidents number.

```
##     Hour Number of accidents
## 1:  TRUE                6957
## 2: FALSE                3382
```

We can also see that more accidents happen between 7am and 7pm, when early or late jobs would have commutes. The fewest number of accidents take place at Night

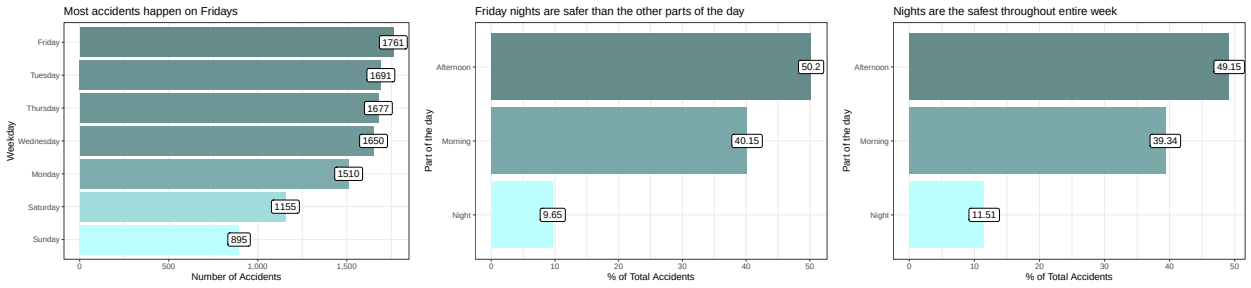
District Name	mild	serious
Unknown	39	1
Gràcia	576	12
Ciutat Vella	623	11
Nou Barris	647	13
Sant Andreu	693	11
Horta-Guinardó	827	10
Les Corts	865	27
Sarrià-Sant Gervasi	1266	24
Sants-Montjuïc	1305	21
Sant Martí	1593	51
Eixample	3499	60

Weekday	District Name	mild	serious
Sunday	Eixample	293	16
Saturday	Eixample	392	7
Monday	Eixample	481	3
Wednesday	Eixample	564	5
Friday	Eixample	571	10
Thursday	Eixample	580	8
Tuesday	Eixample	618	11

The above tables show 1) Accidents by District. 2) Accidents in Eixample by day We see that i) Eixample has more accidents than anywhere else and mild accidents in particular. ii) Eixample has more mild accidents on weekdays when it looks like traffic is slower and more crowded. iii) There isn't a great disparity but it looks like there are more serious accidents on Sundays when roads are less crowded. As Eixample is centrally located, any commute between districts would likely pass through here, explaining its high number of accidents. We will now examine the time distribution of the accidents.

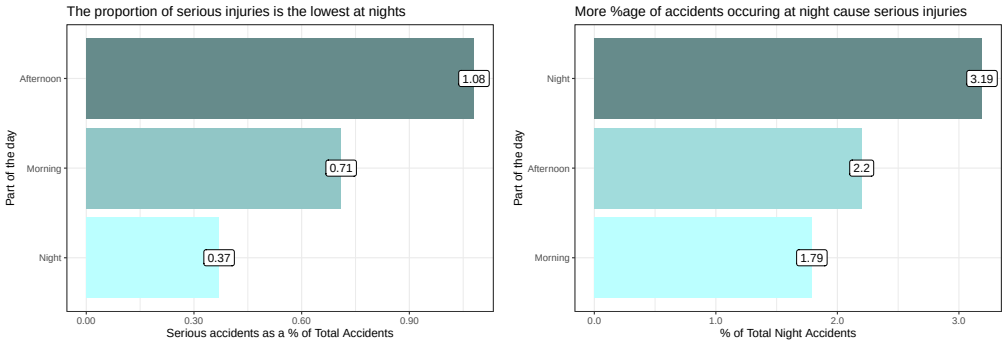
Example of Case where confounding factors are necessary to support claim or invalidate hypothesis Let's analyse the accidents in Barcelona with better visualization. Open questions: which days and/or parts of days are more dangerous in terms of the number of accidents? Let's visualize on which weekdays more accidents happen. As we can see the most number of accidents happen on Fridays, almost 0.87(STD) above the average accidents per weak. On Saturday and Sunday the least amount of accidents happen.

Now let's look at the part of the day when accidents occur the most.On the most accident prone day over 50% of the accidents occur during afternoons. About 41% occur during Morning and less than 10% at nights. This trend is persistent over the entire week.



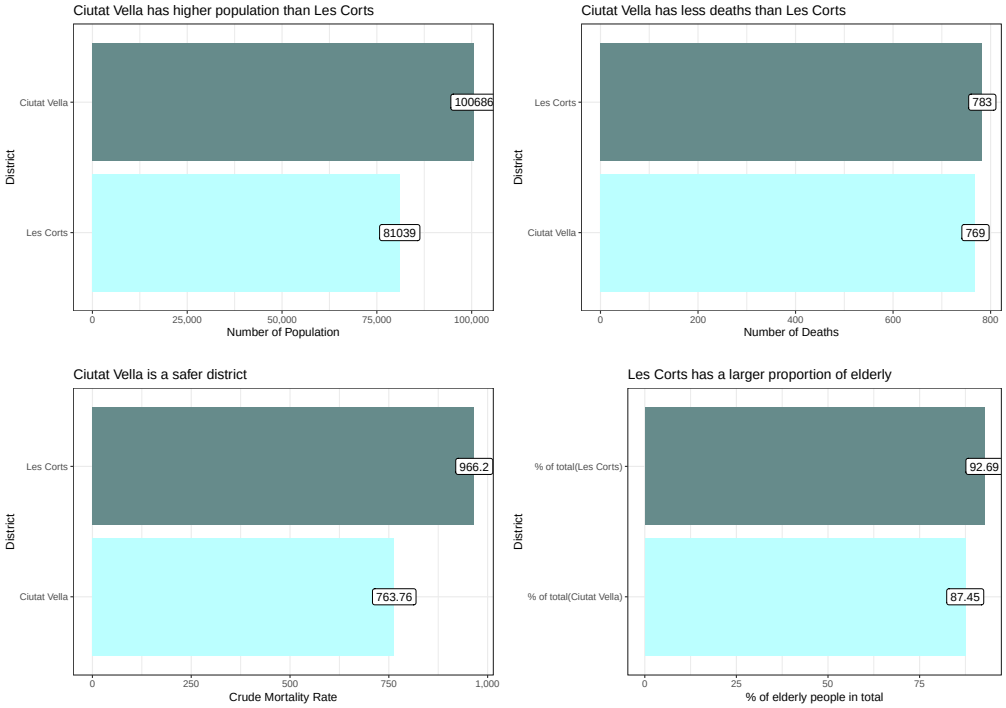
So as we saw Fridays are the most dangerous, Sundays the safest. Friday Afternoons are the most dangerous in terms of accidents number.Furthermore, the accident proportion over the part of days remains similar for the entire week.

Let's have a look on what proportion of total accidents account for serious injuries. And, on the other side, also let us see the proportion of serious injuries in all accidents over the parts of days.



As we can see 3,2% of accidents occuring during nights are serious, which is the highest proportion. While mornings have the least proportion of serious injusries. When we only look at the proportion of serious injuries in total accidents we can see that nights are safer and assume that it's true However, by analyzing further, if we consider the number of accidents at each part of the day(serious accidents at a part of day/accidents at that part of day), we can see that more proportion of total accidents occurring at night are serious. This is a paradox. So, blindly looking at a single data reveals wrong results.

We are interested in the safeness of two districts with different age proportions. We consider a comparison of mortality rates in Les Corts and Ciutat Vella.



The crude mortality ratio is 966,22/763,76= 1,26. Does this mean it is riskier to live in Les Corts? Since older age is clearly a risk factor for death, we introduced a third variable(differences in age) to see if this might be a confounding factor. It is noteworthy, besides the obvious difference in total population size, there is a difference in the age distribution of the two districts. Cortes has a larger proportion of older people(which contribute to higher number of deaths) and Ciutat Vella has a greater proportion of younger people. Hence, one will be mistaken to claim that Les Corts is more riskier only by considering difference in total deaths.

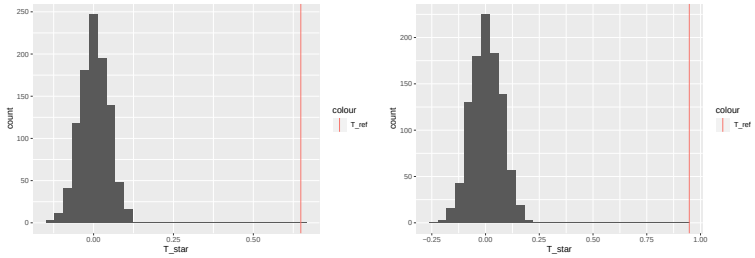
Hypothesis Testing

1) Are there more accidents during the day 2) Are there more accidents on weekdays than weekends?

(1)The null hypothesis is that the average number of accidents is the same during night and day. (2)The null hypothesis is that the average number of accidents is the same during Weekends and Weekdays.

```
## p-value= 0.001998002
```

```
## p-value= 0.001998002
```



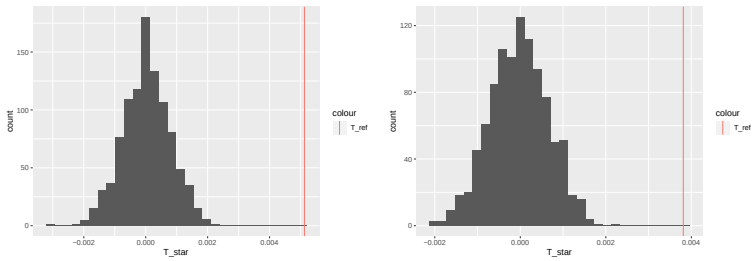
With these p value we can easily reject the null hypotheses and say that there are more accidents in the day on average, and there are definitely more accidents on weekdays during the daytime/working hours.

3) Are there more severe accidents on the weekends? 4) Are there more serious accidents at night?

(3)The null hypothesis is that the average number of serious accidents is the same for weekends and weekdays. (4)The null hypothesis is that the average number of serious accidents is the same for Day and Night.

```
## p-value= 0.001998002
```

```
## p-value= 0.001998002
```



Looking at the p-values,we can reject the null hypotheses and state that there are more serious accidents on Weekends, and that there are more serious accidents at night This could be attributed to the fact that there are less vehicles on the road and less active traffic signals, hence there is more speeding.

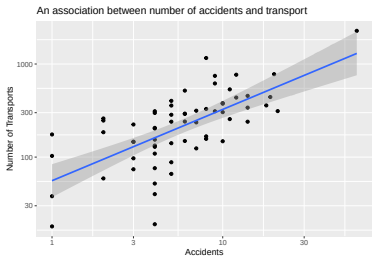
These conclusions corroborate with our claim that more accidents occur in Eixample due to its central location.

Statistically supported claim and visualization

The aim of this part of the analysis to see whether there is an association between the number of transportation modes in the neighbourhoods and the accidents happening.

Motivation: To see the neighbourhood with higher number of transportation tend to have more accidents due to more traffic. For this purpose we need the accidents and transport datasets merged according to the neighbourhoods.

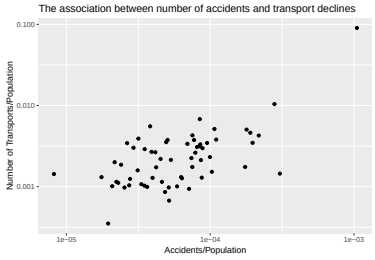
The data contained outliers and we cannot assume normality, that is the reason we use the Spearman Ranking Correlation test instead of Pearson correlation to check the association between two quantitative variables.




```
##
##  Spearman's rank correlation rho
##
## data:  dt$N and dt$Accidents
## S = 12458, p-value = 1.584e-10
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7009916
```

Our initial assumption about the positive correlation is right. Positive correlation of around 0,7 was detected with signifant P value.

However, there might be a third variable inducing this association. That is the reason we add the population to look at the proportional variables such as the number of transport per person and the ratio of accidents to population. Again as the data involves outliers, we cannot assume normality, therefore we use the Spearman's correlation test.




```
##
##  Spearman's rank correlation rho
##
## data:  dt_m$V1 and dt_m$V2
## S = 19034, p-value = 1.921e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.520687
```

As we can see the correlation decreased to around 0.52 after adding the third variable, which can mean we have an example of common cause.