Urban Accidents in the City of Porto Alegre

Fábio Gabriel de Magalhães

October 2017

1 Introduction

Under the transparancy law the city of Porto Alegre has provided a data set with all the urban accidents within its city limits since 2000[1]. In this short note we are interested in the analysis of data for the year of 2001. More specifically we are interested in knowing if weekends impact the number of accidents.

As on weekends there are much less cars in the streets our hypothesis is that there are less accidents on weekends. We will be providing an exploratory data analysis. More rigorous analysis is reserved for future works.

The data set is composed of 21138 observations of 37 variables, where each row represents an accident. Among the variables there are the date of the accident, location, number of deaths occasioned, number of injured people and types of vehicles involved. In this analysis we will be using just the count of accidents, implicitly defined by the number of rows for a particular date or week day.

2 Exploratory Data Analysis

Data is already in a tidy format but some preprocessing is needed to generate more informative graphics. First we create a new column for holding English weekdays and reorder its factor levels to show them in natural order (i.e. starting from Sunday up to Saturday) and not in lexicographical order. As the time of the accident is not relevant we create a new column holding the date, what will be useful to our analysis.

```
weekDays <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",
              "Friday", "Saturday")
df <- df %>% mutate(weekDay=weekdays(DATA_HORA)) %>%
  mutate(weekDay=factor(weekDay, weekDays), date=as.Date(DATA_HORA))
```

Data seems representative as the earliest date recorded is the first day of the year and the latest date recorded is the last day of the year.

```
(earliestDate <- as.Date(min(df$DATA_HORA)))
```

```
## [1] "2001-01-01"
```

```
(latestDate <- as.Date(max(df$DATA_HORA)))
```

```
## [1] "2001-12-31"
```

```
(nDays <- as.numeric(latestDate - earliestDate + 1))
```

```
## [1] 365
```

Procceeding with the plot we notice that there is a lower number of total accidents happening on weekends

```
ggplot(data=df) + geom_bar(mapping = aes(x=weekDay)) +
  labs(title="Number of accidents by day of the week",
       x="Day of the week", y="Number of accidents")
```

Inspecting Figure 1 it is possible to see an aparent decrease in the total number of accidents on weekends. We now inspect the difference from the mean number of accidents 57.9123288.

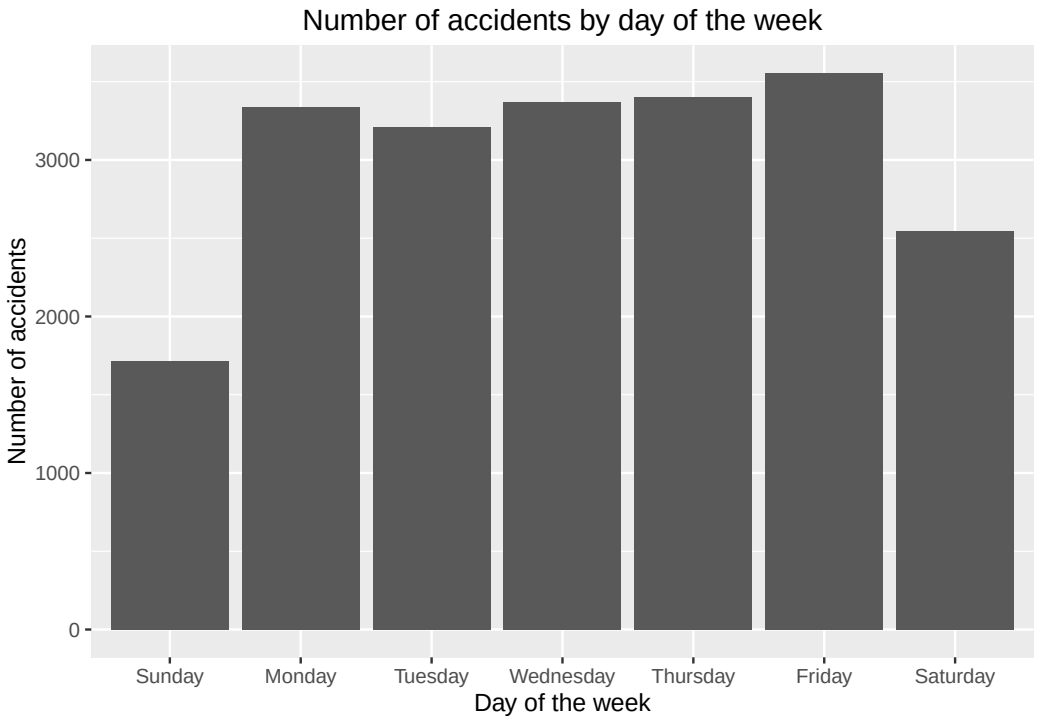

Figure 1: Histogram of the number of accidents

```
(dfStats <- df %>%
   group_by(weekDay) %>%
   summarize(totalAccidents=n(), meanAccidents=n() / 52))
```

```
## # A tibble: 7 x 3
##     weekDay totalAccidents meanAccidents
##      <fctr>          <int>         <dbl>
## 1    Sunday           1718      33.03846
## 2    Monday           3341      64.25000
## 3   Tuesday           3210      61.73077
## 4 Wednesday           3367      64.75000
## 5  Thursday           3404      65.46154
## 6    Friday           3555      68.36538
## 7  Saturday           2543      48.90385
```

The table above shows the mean number of accidents per weekday. It is interesting to note how those differ from the entire sample mean number of accidents, 57.9123288. Note that in the computation above we are ignoring the fact that there are 53 Mondays in the recorded span, while there are 52 of the remaining week days. That is because the recording starts and ends on a Monday.

```
dfStats2 <- df %>%
   group_by(date, weekDay) %>%
   summarize(totalAccidents=n())

ggplot(dfStats2, aes(weekDay, totalAccidents)) +
  geom_boxplot() +
  labs(title="Number of daily accidents by day of the week",
       x="Day of the week", y="Number of accidents")
```

Once more we have a visual indication that less accidents happen on Saturdays and Sundays. In Figure 2 it is possible to notice that the mean values for Saturdays and Sundays are lower than all other mean values, as well as the quartiles for those two days are lower.

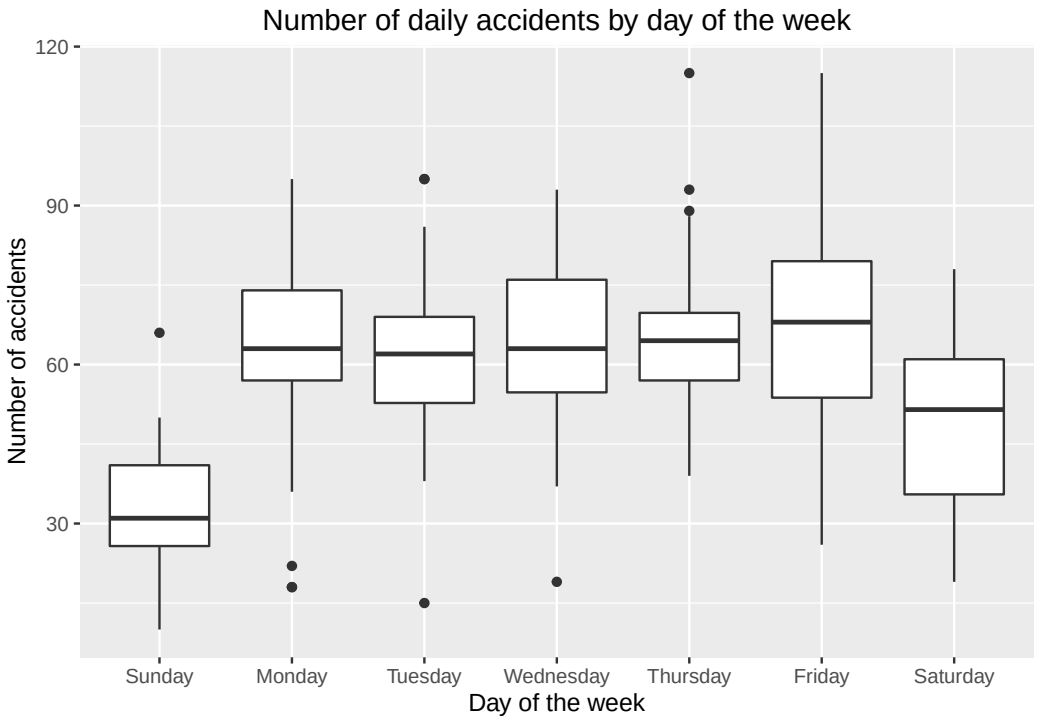

Figure 2: Number of accidents per day

3 Conclusions and Future Work

We observed that there are less accidents on weekends. In order to establish whether this difference is significant or not we should use rigorous statistics, testing the difference in means and possibly modeling data with Poisson models as those are suitable for count data[2].

The more data we analyze the more confident we can be in our hypothesis. Hence we shoudn't restrict our analysis to one particular year, but analyze the whole data set instead. Having more years also allows us to check the representativity of the samples (year worth of data).

4 References

[1] "Acidentes De Trânsito." Portal de Dados Abertos da Cidade de Porto Alegre, www.datapoa.com.br/dataset/acidentes-de-transito.

[2] P. Dalgaard, *"Introductory Statistics with R"*. Springer, 2008.

[3] H. Wickham and G. Grolemund, *"R for Data Science"*. O'Reilly, 2016.