Urban Accidents in the City of Porto Alegre

Fabio Gabriel de Magalhaes

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 work we are interested in the analysis of data for the years 2012-2016. 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 using statistics to verify the difference is significant without trying to do causal analysis.

The data set is composed of five data frames, each data frame holding accident occurences for a particular year. 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.

Data is already in a tidy format where each column is a variable and each row an observation. However, some preprocessing is needed to generate more informative graphics and facilitate our analysis. 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) when plotting, and not in lexicographical order. As the time of the accident is not relevant we igore it when creating the new column.

```
addWeekDays <- function(dataf) {
  weekDays <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",
                "Friday", "Saturday")

  dataf %>% mutate(weekDay=weekdays(as.Date(DATA_HORA))) %>%
    mutate(weekDay=factor(weekDay, weekDays), date=as.Date(DATA_HORA))
}

yearlyData <- map(yearlyData, addWeekDays)
```

As we would like to know if there are more accidents on weekends we create a new binary variable to indicate whether or not an observation was made on weekends.

```
addWeekendIndicator <- function(dataf) {
  dataf %>% mutate(isWeekend=weekDay %in% c('Saturday', 'Sunday'))
}

yearlyData <- map(yearlyData, addWeekendIndicator)
```

2 Exploratory Data Analysis

Having data classified by group we can compute the average number of accidents by weekends by year, and the same for weekdays. Doing that we have to groups of five elements each: the group of averages by weekend and the group of averages by weekday.

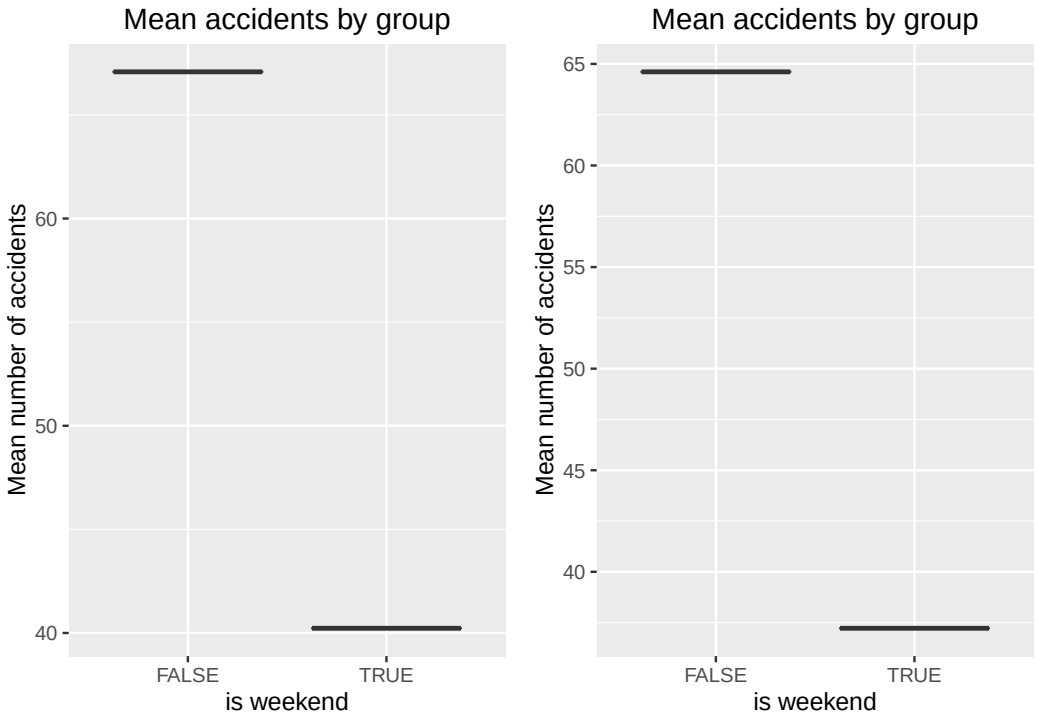
```
groupings <- map(yearlyData, function(dataf)
  dataf %>% group_by(isWeekend) %>%
  summarize(totalAccidents=n(), meanAccidents=n() / n_distinct(date))
  )
```

```r
weekendGroup <- unlist(map(groupings, function(dataf)
  as.numeric(dataf[dataf$isWeekend == T, 'meanAccidents'])
  ))

weekdayGroup <- unlist(map(groupings, function(dataf)
  as.numeric(dataf[dataf$isWeekend == F, 'meanAccidents'])
  ))
```

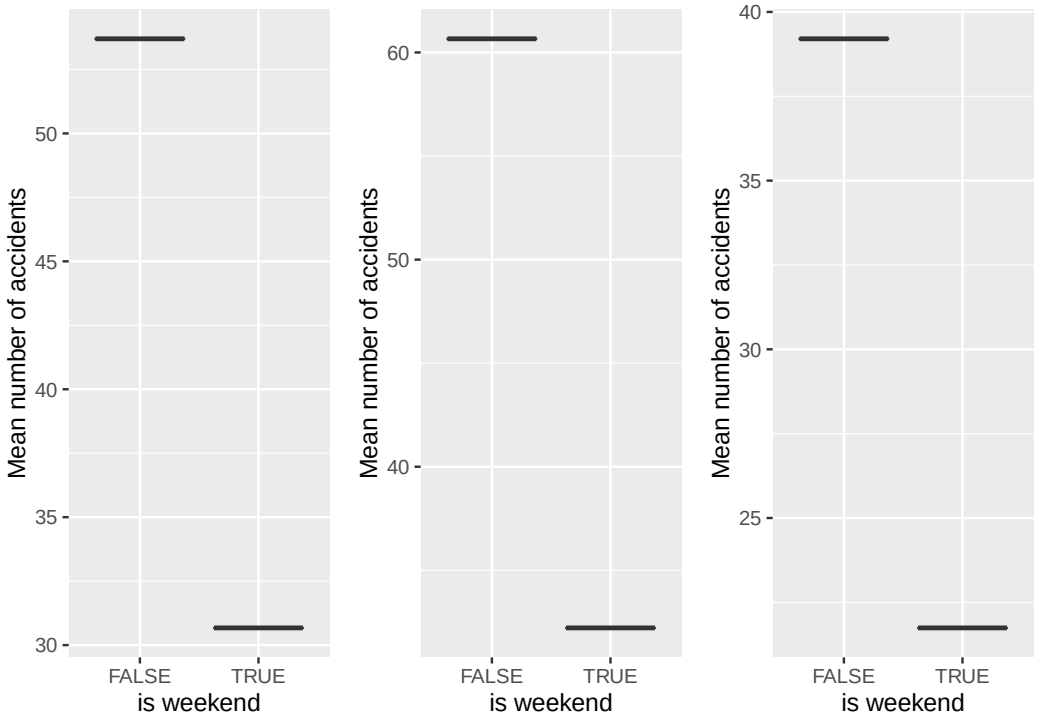
Plotting groups to have a visual indication of the difference between means of weekdays and weekends.

```r
# Routine from W.Chang "R Graphics Cookbook"
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

 if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}

plot1 <- ggplot(groupings[[1]], aes(isWeekend, meanAccidents)) +
  geom_boxplot() +
  labs(title="Mean accidents by group", x="is weekend",
       y="Mean number of accidents")

plot2 <- ggplot(groupings[[2]], aes(isWeekend, meanAccidents)) +
  geom_boxplot() +
  labs(title="Mean accidents by group", x="is weekend",
       y="Mean number of accidents")

plot3 <- ggplot(groupings[[3]], aes(isWeekend, meanAccidents)) +
  geom_boxplot() +
  labs(x="is weekend",
       y="Mean number of accidents")
```

```
plot4 <- ggplot(groupings[[4]], aes(isWeekend, meanAccidents)) +
  geom_boxplot() +
  labs(x="is weekend",
       y="Mean number of accidents")

plot5 <- ggplot(groupings[[5]], aes(isWeekend, meanAccidents)) +
  geom_boxplot() +
  labs(x="is weekend",
       y="Mean number of accidents")

multiplot(plot1, plot2, cols=2)
```




```
multiplot(plot3, plot4, plot5, cols=3)
```



3 Statistical Analysis

We can test if the difference between means of two groups is significant by using Welch's T-test. The null hypothesis for this test is that there are no difference between the two groups means.

```
t.test(weekendGroup, weekdayGroup)
```

```
##
##  Welch Two Sample t-test
##
## data:  weekendGroup and weekdayGroup
## t = -4.1589, df = 6.7675, p-value = 0.004574
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -38.73605 -10.52924
## sample estimates:
## mean of x mean of y
##  32.41701  57.04965
```

The p-value obtained is less than 0.05, that is, our result has a probability less than 5% of being due to chance, and as so, by following standard statistical practice we can reject the null hypothesis. This indicates the difference of means is indeed significant.

4 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 performed the Welch's T-test to the two groups, verifying that both means are indeed significantly different.

In order to assert with total confidence that we have less accidents on weekends we would neet to perform some steps. First, we would have to adjust data for the fact that are less cars in circulation during weekend (the author could not find such data for Porto Alegre). Second, we should also weight yearly means by a factor accounting for the yearly increase in the number of cars.

Lastly, after correcting data for the aforementioned issues it is probably more rigorous to fit a Generalized Linear Model to data instead of comparing means. The appropriate model for couting data like these is the Poisson model. With Poisson model we could estimate how many accidents we would expect per day on average and verify if those numbers differ from the numbers obtained for weekends and weekdays.

5 References

[1] "Acidentes De Transito." 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.

[4] W. Chang, *"R Graphics Cookbook"*. O'Reilly, 2012.