Urban Accidents in the City of Porto Alegre

Jean-Marc Vincent, Lucas Mello Schnorr

October 2017

Each student should provide a Rmd file with *two* to *four* plots, with text describing the semantics of the data, the question, how they have answered the question, and an explanation for each figure, showing how that particular figure helps the answering of the initial question. Fork the LPS repository in GitHub, push your Rmd solution there. Send us, by e-mail, the link for your GIT repository, indicating the PATH to the Rmd file. Check the LPS website for the deadline.

1 Introduction

The City of Porto Alegre, under the transparency law, has provided a data set with all the urban accidents (within the city limits) since 2000. The data set, including a description of each column in the PDF file format, is available in the following website:

http://www.datapoa.com.br/dataset/acidentes-de-transito

2 Goal

For a given year (defined by the LPS coordination for each student enrolled in the cursus), the goal is to answer one of the following questions. The solution must use the data import and manipulation verbs of the R programming language and the tidyverse metapackage (readr, tidyr, dplyr) using Literate Programming.

3 Questions

1. What is the time of the day with most accidents?
2. How many vehicles are involved in the accidents?
3. What types of accidents are more common?
4. Is the number of deaths increasing or decreasing?
5. Is there a street of the city with more accidents than others?
6. Do holidays impact in the number of accidents?

4 Download the data

Supposing you have the URL for the CSV file, you can read the data using the code below. You can also download it manually and commit it to your repository to avoid an internet connection every time you knit this file. If the URL changes, the second solution might even make your analysis be more portable in time.

```
library(dplyr);
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr);
library(ggplot2);
library(lubridate);
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(readr)
URL <- "http://www.opendatapoa.com.br/storage/f/2013-11-06T17%3A34%3A58.965Z/acidentes-2002.csv"
df <- read_delim(URL, delim=";");
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   LOCAL_VIA = col_character(),
##   LOG1 = col_character(),
##   LOG2 = col_character(),
##   LOCAL = col_character(),
##   TIPO_ACID = col_character(),
##   DATA_HORA = col_datetime(format = ""),
##   DIA_SEM = col_character(),
##   TEMPO = col_character(),
##   NOITE_DIA = col_character(),
##   FONTE = col_character(),
##   BOLETIM = col_character(),
##   REGIAO = col_character(),
##   LATITUDE = col_number(),
##   LONGITUDE = col_number()
## )
```

```
## See spec(...) for full column specifications.
```

4.1 1. What is the time of the day with most accidents?

To answer this question, we group elements by the FX_HORA column and count how many observations there are in each of the 24 groups. After sorting, we can see that 18h is when most accidents happen.

```
df %>%
  group_by(FX_HORA) %>%
  summarise(n=n()) %>%
  arrange(-n)
```

```
## # A tibble: 24 x 2
##    FX_HORA     n
##      <int> <int>
##  1      18  1755
##  2      15  1549
##  3      14  1520
##  4      17  1486
##  5      16  1473
##  6      12  1386
##  7      19  1350
##  8      11  1319
```

```
##  9      10  1261
## 10       8  1246
## # ... with 14 more rows
```

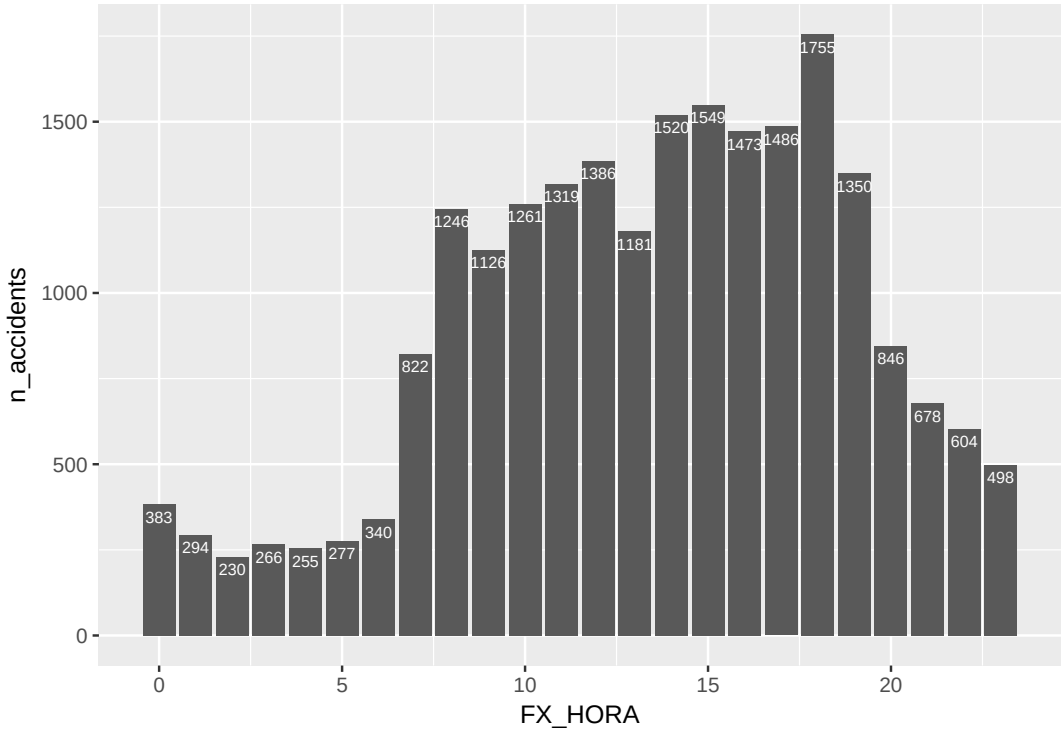
In the year of 2002, 7.9% of accidents happened in the interval between 18:00 and 18:59.

```
df %>%
  group_by(FX_HORA) %>%
  summarise(n=n()) %>%
  max() / nrow(df)
```

```
## [1] 0.0792504
```

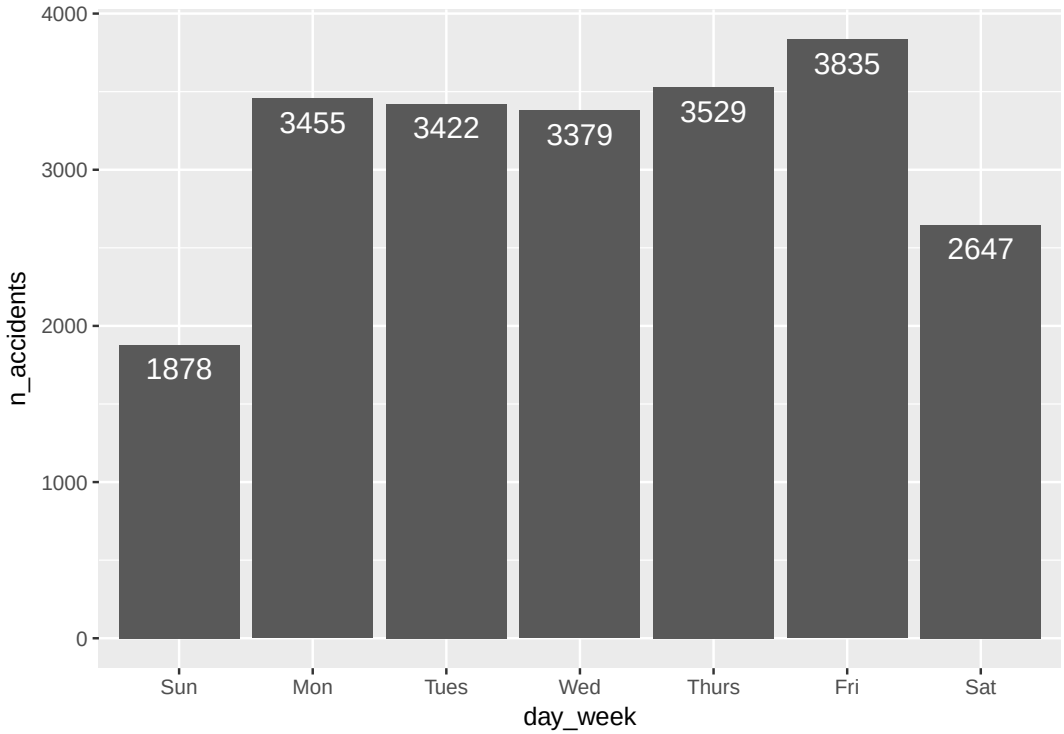
The distribution of number of accidents per hour can be seen in the histogram below.

```
df %>%
  group_by(FX_HORA) %>%
  summarise(n_accidents=n()) %>%
  ggplot(aes(x=FX_HORA, y=n_accidents)) +
    geom_bar(stat="identity") + ylim(0,NA) +
    geom_text(aes(label=n_accidents), vjust=1.6, color="white", size=2.5)
```



The day of the week also has a large impact on the number of accidents. Sundays tend to have less than half the number of accident seen on Fridays.

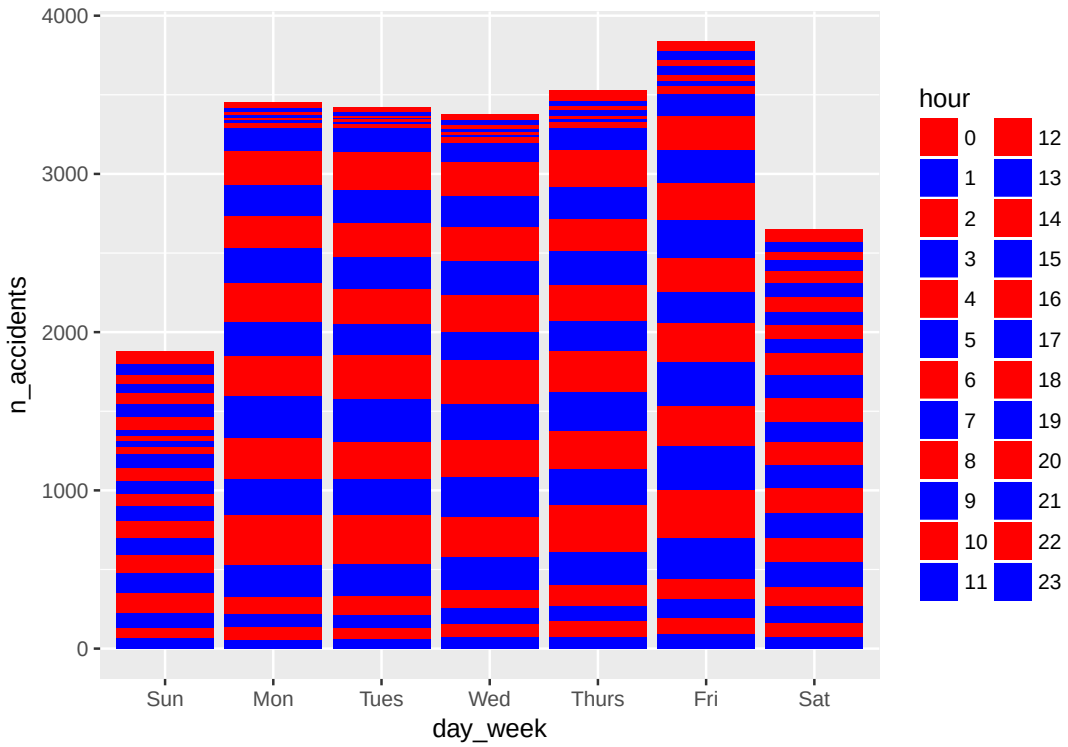
```
df %>%
  mutate(day_week = wday(DATA_HORA, label = TRUE)) %>%
  group_by(day_week) %>%
  summarise(n_accidents=n()) %>%
  ggplot(aes(x=day_week, y=n_accidents)) +
    geom_bar(stat="identity") +
    ylim(0,NA) +
    geom_text(aes(label=n_accidents), vjust=1.6, color="white", size=4.5)
```



We can combine these two factors in a single stacked bar chart. We can see that during weekdays, the amount of accidents in the first hours of the day (0 through 5) is very low (top of the stack), while on weekends, the variance between hours is much lower.

```
zebra_colormap <- rep(c("red", "blue"), 12)

df %>%
  mutate(day_week = wday(DATA_HORA, label = TRUE)) %>%
  group_by(day_week, FX_HORA) %>%
  summarise(n_accidents=n()) %>%
  mutate(hour = as.factor(FX_HORA)) %>%
  ggplot(aes(x=day_week, y=n_accidents, fill=hour)) +
    geom_bar(stat="identity") +
    scale_fill_manual(values=zebra_colormap)
```


4000
3000
2000
1000
0
n_accidents
Sun
Mon
Tues
Wed
Thurs
Fri
Sat
day_week
hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23