Urban Accidents in the City of Porto Alegre

Rodrigo Moni

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(readr);
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);
URL <- "http://www.opendatapoa.com.br/storage/f/2013-11-06T17%3A26%3A29.293Z/acidentes-2000.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.
```

5 Task

- Semantics of the Data

In this analysis was used the following variables:

1. NOITE_DIA: Variable of type "chr" that identifies the period of the day when happened the accident. It assumes three values in the dataset: "DIA" means "Day", "NOITE" means "Night" and "NA" means "No Value".
2. FX_HORA: Variable of type "int" that identifies the hour of the day when happend the accident. The value means the hour of the day in brazilian format.
3. Obs: There are some inconsistencies in the dataset, for example are some records with the NOITE_DIA = Day and FX_HORA = 23 which is clearly wrong, because 23h (11:00pm) is in the period of night.

- The selected question "What is the time of the day with most accidents?"
- Question answer

The time of the day with most accidents in 2000 is in period of the day at 15 hours (3:00pm)

First we wanted to see which period of the day with most accidents then we plotted a histogram with the count of accidents by period (day and night). To achieve this we wrote this code below:

```
histogram <- df %>% group_by(NOITE_DIA) %>% filter(NOITE_DIA %in% c("DIA", "NOITE"))
count(histogram)
```

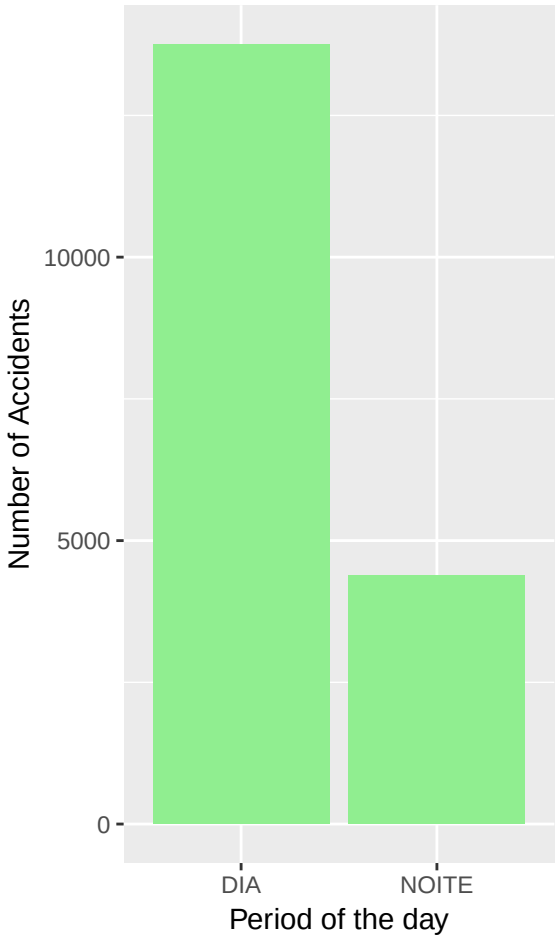
```
## # A tibble: 2 x 2
## # Groups:   NOITE_DIA [2]
##   NOITE_DIA     n
##       <chr> <int>
## 1       DIA 13757
```

```
## 2     NOITE  4391
```

We did a "group_by" NOITE_DIA and then we applied a filter to get only the records that had value "DIA" or "NIGHT", excluding the records with the Value "NA".

See the plot below:

```
library(ggplot2)
ggp <- ggplot(histogram,aes(x=NOITE_DIA))
# counts
ggp + geom_bar(fill="lightgreen") + ylab("Number of Accidents") + xlab("Period of the day")
```



In this image we can see that the period with the most accidents is the period of the Day ("DIA"" in the image). In the period of the day in this year happened more than 13000 accidents.

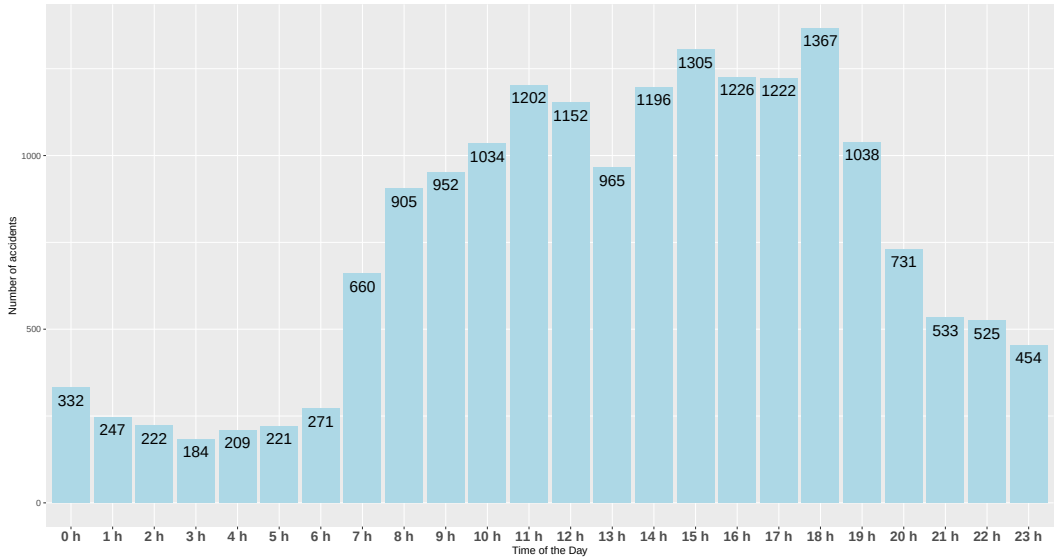
After that we wanted to know the exactly time of the day with most accidents, so we wrote the following code:

```
histogramDia <- df %>% group_by(FX_HORA) %>% summarise(N=n())
```

We did a "group_by" FX_HORA (hour) to verify the exactly hour with most accidents. The result is 18h (6pm). We can see better in the histogram plotted below:

```
sequence <- 0:23
result <- paste(sequence,"h")
ggp2 <- ggplot(histogramDia,aes(x=factor(FX_HORA, levels= as.character(seq(0,23,1)), labels = res

# counts
ggp2 + geom_bar(fill="lightblue", stat = "identity") + xlab("Time of the Day") + ylab("Number of
```



In this image we can clearly see that the hour with most accitents is the 18h and in the second place the 15h.

To generate the x axis as categorical data it was necessary to read the hours as “char” instead of “int”.