Urban Accidents in the City of Porto Alegre

Lizeth Andrea Castellanos Beltran

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.

```r
library(readr)
URL <- "http://www.opendatapoa.com.br/storage/f/2013-11-06T16%3A52%3A35.356Z/acidentes-2009.csv"
df <- read_delim(URL, delim=";")
```

```
## `curl` package not installed, falling back to using `url()`
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   LOG1 = col_character(),
##   LOG2 = col_character(),
##   LOCAL = col_character(),
##   TIPO_ACID = col_character(),
##   LOCAL_VIA = 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.
```

```
## Warning: 1 parsing failure.
## row # A tibble: 1 x 5 col     row       col   expected         actual expected   <int>      <ch
```

```
df
```

```
## # A tibble: 22,127 x 37
##        ID                          LOG1                             LOG2
##     <int>                         <chr>                            <chr>
##  1 460836               AV ASSIS BRASIL                 R ALBERTO SILVA
##  2 460838             R JOSE DE ALENCAR            AV DR CARLOS BARBOSA
##  3 460730 R EUSTAQUIO INACIO DA SILVEIRA R JOAO VIEIRA DE AGUIAR SOBRINHO
##  4 460837                   AV FARRAPOS          R CONDE DE PORTO ALEGRE
##  5 460834             AV PROTASIO ALVES                             <NA>
##  6 460805                AV JOAO PESSOA                             <NA>
##  7 460732          AV PROF OSCAR PEREIRA                            <NA>
##  8 460835             AV VENANCIO AIRES                   AV JOAO PESSOA
##  9 460845       R GEN COUTO DE MAGALHAES            R CARLOS VON KOSERITZ
## 10 460839         R JOSE FERREIRA JARDIM                            <NA>
## # ... with 22,117 more rows, and 34 more variables: PREDIAL1 <int>,
## #   LOCAL <chr>, TIPO_ACID <chr>, LOCAL_VIA <chr>, DATA_HORA <dttm>,
## #   DIA_SEM <chr>, FERIDOS <int>, MORTES <int>, MORTE_POST <int>,
## #   FATAIS <int>, AUTO <int>, TAXI <int>, LOTACAO <int>, ONIBUS_URB <int>,
## #   ONIBUS_INT <int>, CAMINHAO <int>, MOTO <int>, CARROCA <int>,
## #   BICICLETA <int>, OUTRO <int>, TEMPO <chr>, NOITE_DIA <chr>,
## #   FONTE <chr>, BOLETIM <chr>, REGIAO <chr>, DIA <int>, MES <int>,
## #   ANO <int>, FX_HORA <int>, CONT_ACID <int>, CONT_VIT <int>, UPS <int>,
## #   LATITUDE <dbl>, LONGITUDE <dbl>
```

5 Critical analysis of the data

The website provides a dataset dictionary with the list and the description of each variable. There are 43 variables in the dataset to describe every event (accident). The data is not present in a tidy form. There are dupplicated information. For example the variables DATA, DIA, MES, ANO, FX_HORA contain the same information about the date and the time. There are missing data in some boxes. Is not clear which approach has been used to collet the data. We have seasons in Porto Alegre so the time change in the summer. I did not see information about that and I think that this is relevant, to aswer the first question, for example.

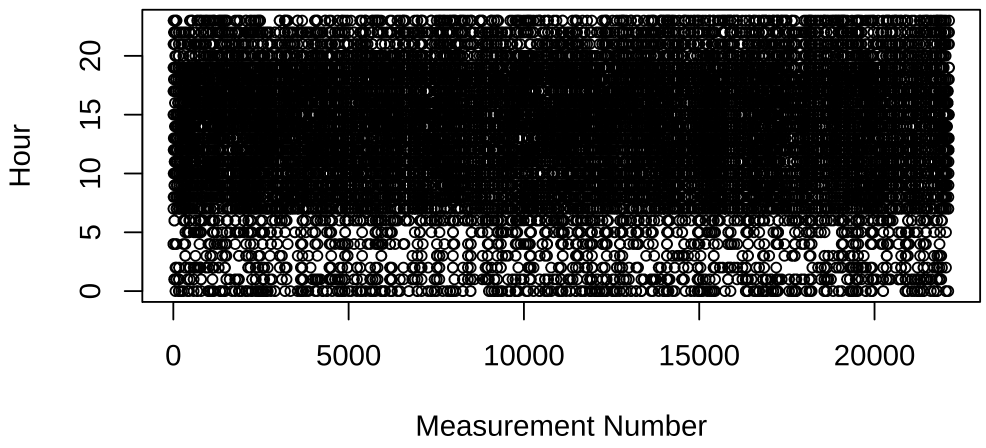
6 Question 1

I will describe the analysis for the accidents in Porto Alegre in 2009. I will answer the first question:

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

- To answer this question the relevant variable is FX_HORA. The first thing i will do is plot the values of FX_HORA to find meaningful data.

```
plot(df$FX_HORA, ylab="Hour", xlab="Measurement Number", cex=0.75);
```



This plot does not provide much help to answer the question, because there is a lot of values that need to be grouped.

To group the date I will use *dplyr*.

```
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
```

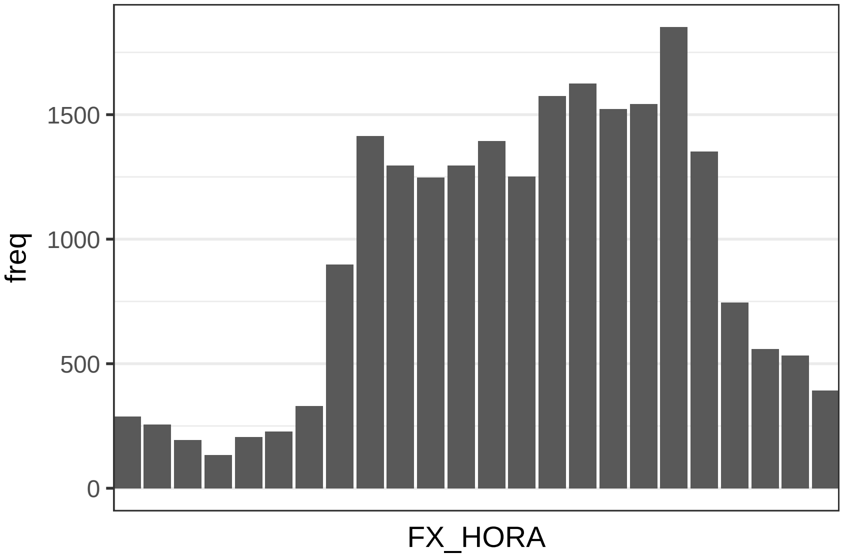
```
df %>% group_by(FX_HORA) %>% summarise(freq=sum(CONT_ACID))
```

```
## # A tibble: 24 x 2
##    FX_HORA  freq
##      <int> <int>
##  1       0   289
##  2       1   256
##  3       2   193
##  4       3   133
##  5       4   205
##  6       5   228
##  7       6   330
##  8       7   899
##  9       8  1414
## 10       9  1296
## # ... with 14 more rows
```

Now I get the two columns with the tidy data I want to plot.

The result of that *summarise* pipe will be passed to ggplot to visualize the number of accidents per hour.

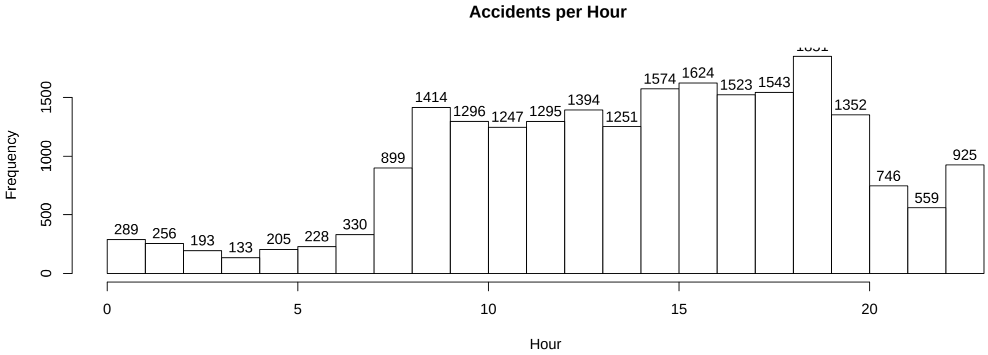
```
library(ggplot2);
df %>% group_by(FX_HORA) %>% summarise(freq=sum(CONT_ACID)) %>%
   ggplot(aes(x=FX_HORA, y=freq)) + geom_bar(stat="identity") + ylim(0,NA)+ theme_bw()  + scale_
```



Now I see a trend in the amount of accidents by hour of the day.

To compare the result and to have exact hours I will create an histogram for the FX_HORA column.

```
df$FX_HORA %>% hist(breaks=23, xlab="Hour",  main=" Accidents per Hour", labels=TRUE, right= FALS
```



Using the result of the two figures, I can answer the question.

7 Answer for the Question 1

The time of the day with most accidents in 2009 is 19h with 1367 accidents.