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.

```r
library(readr)
data <- read_delim("/home/gauss/lllemos/lps/tasks/acidentes-2006.csv", delim=";")
```

```
## 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_character(),
##   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.
```

```r
data
```

```
## # A tibble: 20,333 x 37
##        ID                             LOG1                    LOG2
##     <int>                            <chr>                   <chr>
##  1 390318   AV BALTAZAR DE OLIVEIRA GARCIA                    <NA>
##  2 390319                   R DOM PEDRO II                    <NA>
##  3 390256 AV BERNARDINO SILVEIRA DE AMORIM                    <NA>
##  4 390280                  AV ASSIS BRASIL                    <NA>
##  5 390281               AV DR NILO PECANHA                    <NA>
##  6 390283               AV BENTO GONCALVES                    <NA>
##  7 390783             AV CRISTOVAO COLOMBO AV PLINIO BRASIL MILANO
##  8 390284                   AV MOAB CALDAS                    <NA>
##  9 390344                  ESTR VILA MARIA         AV DA CAVALHADA
## 10 390282                         AV ROCIO                    <NA>
## # ... with 20,323 more rows, and 34 more variables: PREDIAL1 <int>,
## #   LOCAL <chr>, TIPO_ACID <chr>, LOCAL_VIA <chr>, DATA_HORA <chr>,
## #   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>
```

Load the necessary packages:

```r
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
```

```r
library(magrittr);
```

5 Answer

The data used to develop the work is the urban accidents of 2006 from Porto Alegre. The dataset has specific information about the accidents.

The question choseen to be answered is “Q2. How many vehicles are involved in the accidents?”

First, it's selected the variables related to vehicles and the rows of these variables represent the amount of vehicles involved in the accidents.

```
data %>% select(AUTO, TAXI, LOTACAO, ONIBUS_URB, ONIBUS_INT, CAMINHAO, MOTO, CARROCA, BICICLETA,
```

```
## # A tibble: 20,333 x 10
##     AUTO  TAXI LOTACAO ONIBUS_URB ONIBUS_INT CAMINHAO  MOTO CARROCA
##    <int> <int>   <int>      <int>      <int>    <int> <int>   <int>
##  1     3     0       0          0          0        0     0       0
##  2     2     0       0          0          0        0     0       0
##  3     1     0       0          0          0        0     0       0
##  4     0     0       0          0          1        0     0       0
##  5     1     0       0          0          0        0     0       0
##  6     2     0       0          0          0        0     0       0
##  7     1     1       0          0          0        0     0       0
##  8     1     0       0          0          0        0     0       0
##  9     2     0       0          0          0        0     0       0
## 10     0     1       0          0          0        0     0       0
## # ... with 20,323 more rows, and 2 more variables: BICICLETA <int>,
## #   OUTRO <int>
```

Second, the following plot shows during the year of 2006 how many vehicles of each type are involved in accidents. We can notice that automobiles (variable 'AUTO') are the ones that are most involved in accidents. The x axis represents the classes of vehicles and the y axis the amount of these vehicles involved in accidents during the whole year.

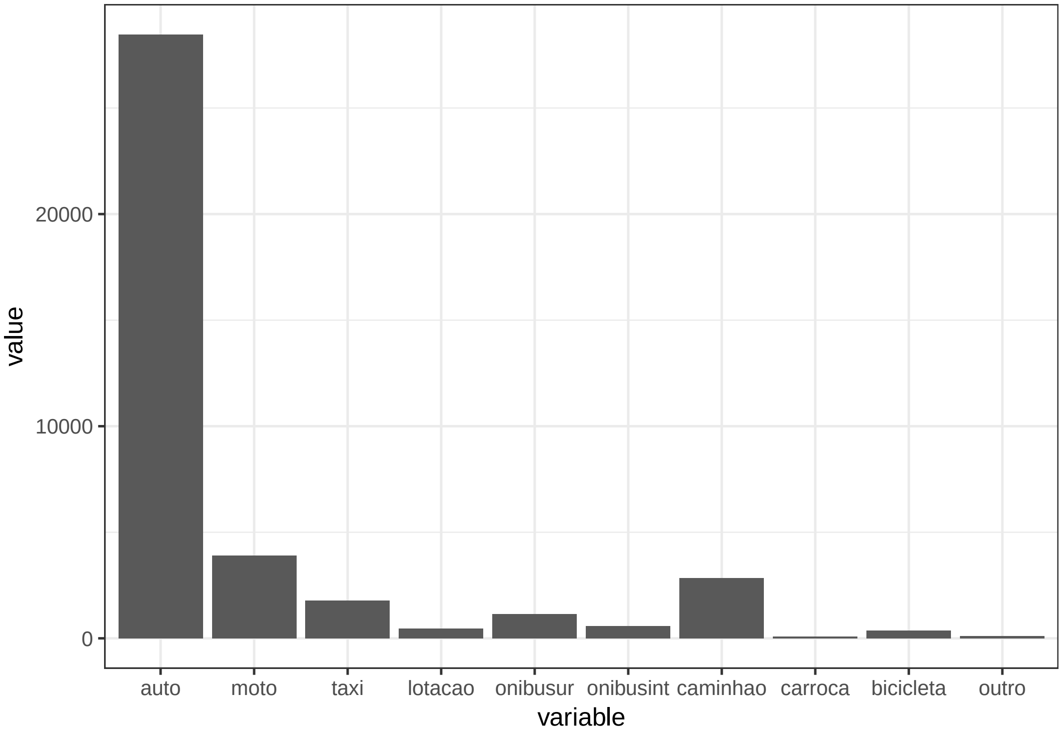
```
library(ggplot2);
library(reshape2);

new_data <- data %>% summarize(auto=sum(AUTO),  moto=sum(MOTO), taxi=sum(TAXI), lotacao=sum(LOTAC

new_data <- melt(new_data[,c('auto', 'moto', 'taxi', 'lotacao', 'onibusur', 'onibusint', 'caminha
```

```
## No id variables; using all as measure variables
```

```
ggplot(new_data, aes(x=variable, y=value)) + geom_bar(stat = "identity") + ylim(0,NA) + theme_bw
```



Third, it's used the verb mutate to create a new column(variable) with the sum of all types of vehicles involved in the accidents. After that, it's created a bar plot to see which interval has the highest incidence of accidents. Also, the mean and standart deviation are showed.

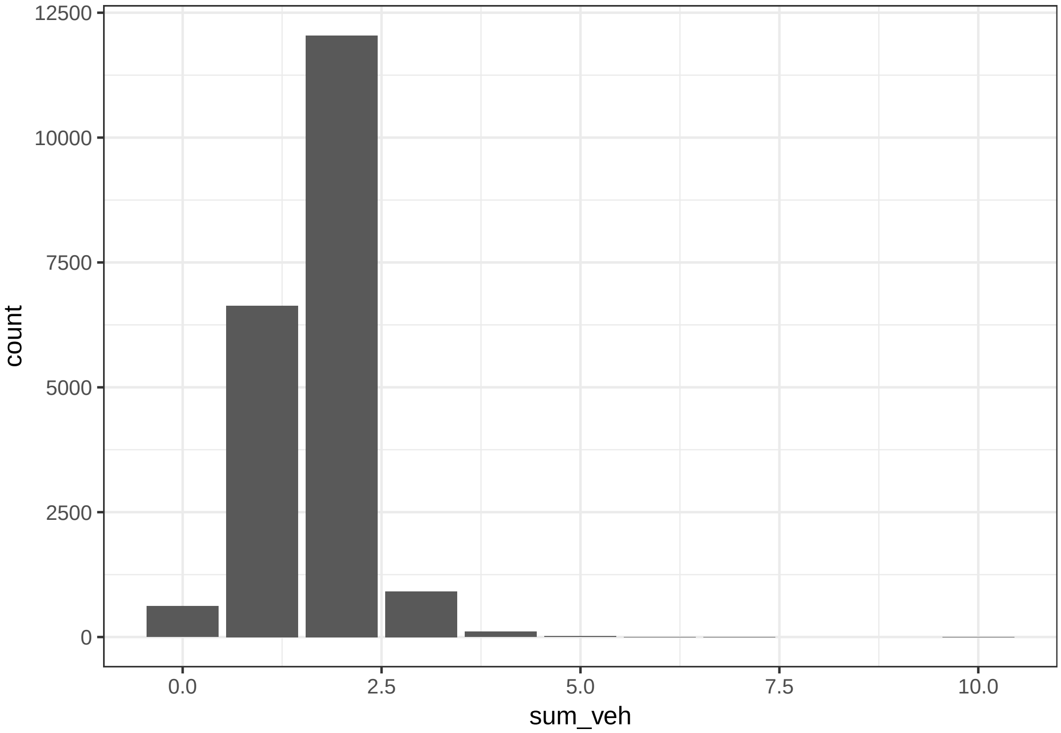
In the plot the x axis represents the new variable of how many vehicles are involved in each accidents and the y axis represents the amount of vehicles involved.

```
library(ggplot2);

data %>%  mutate(sum_veh = AUTO+MOTO+LOTACAO+ONIBUS_URB, ONIBUS_INT, CAMINHAO, MOTO, CARROCA, BIC
```

```
## # A tibble: 1 x 2
##   mean_sum_veh       std
##          <dbl>     <dbl>
## 1      1.67088 0.6458728
```

```
library(ggplot2);
data %>%  mutate(sum_veh = AUTO+MOTO+LOTACAO+ONIBUS_URB, ONIBUS_INT, CAMINHAO, MOTO, CARROCA, BIC
```


count
sum_veh