Laboratory Notebook

Paulo Souza

Introduction

This laboratory book describes the methodology adopted in order to analyze different aspects regarding transit accidents recorded by a public company responsible for managing the public transport in the city of Porto Alegre (Brazil) in 2012.

Environment Setup

Firstly, I import the libraries that will be used throughout the data analysis.

```
library(ggplot2);
library(dplyr);
library(readr)
```

Analysis of Dataset Semantics

Once the base libraries are loaded in the R session, I download the CSV file containing the transit accidents from 2012 in Porto Alegre and store it into the `df` variable.

```
URL <- "http://www.opendatapoa.com.br/storage/f/2013-11-08T12%3A32%3A00.175Z/acidentes-2012.csv";
df <- read_delim(URL, delim=";");
```

As soon as I got the dataset, I decided to take a look on it in order to understand how the data was organized.

```
df
```

```
## # A tibble: 20,202 x 37
##        ID LOG1  LOG2  PREDIAL1 LOCAL TIPO_ACID LOCAL_VIA
##     <int> <chr> <chr>    <int> <chr> <chr>     <chr>
##  1 536036 ESTR~ <NA>      5788 Logr~ ABALROAM~ 5788 EST~
##  2 535378 R CO~ <NA>       162 Logr~ ATROPELA~ 162 R CO~
##  3 535547 AV A~ R VI~        0 Cruz~ ABALROAM~ AV ALBER~
##  4 535549 AV I~ AV C~        0 Cruz~ ABALROAM~ AV CEL L~
##  5 535564 AV P~ <NA>       867 Logr~ COLISAO   867 AV P~
##  6 535384 AV D~ <NA>      2347 Logr~ CAPOTAGEM 2347 AV ~
##  7 535581 AV P~ <NA>      1932 Logr~ ABALROAM~ 1932 AV ~
##  8 535388 TRAV~ <NA>       170 Logr~ CHOQUE    170 TRAV~
##  9 535390 R CR~ <NA>      2037 Logr~ ABALROAM~ 2037 R C~
## 10 535393 R RA~ <NA>      1938 Logr~ CHOQUE    1938 R R~
## # ... with 20,192 more rows, and 30 more variables: 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>
```

Then, I realized that the columns were not so self-explaining as I would like. So, I decided to take a look at the website in order to catch the real meaning of the columns. As a result, I found a file that explains each of the columns presented in the dataset. Thus, I would recommend read this file before reading the remaining of this laboratory book.

Question 1

As soon as I realized the meaning of each of the dataset columns, I decided to analyze **what the time of the day with most transit accidents** to identify if there all rows were filled correctly. To do so, I checked if all rows of the column `NOITE_DIA` (NOITE_DIA stands for NIGHT_DAY but in portuguese) were filled.

```
count(df) == df %>% filter(!is.na(NOITE_DIA) &&
!is.null(NOITE_DIA) && !is.nan(NOITE_DIA)) %>% summarise(N = n())
```

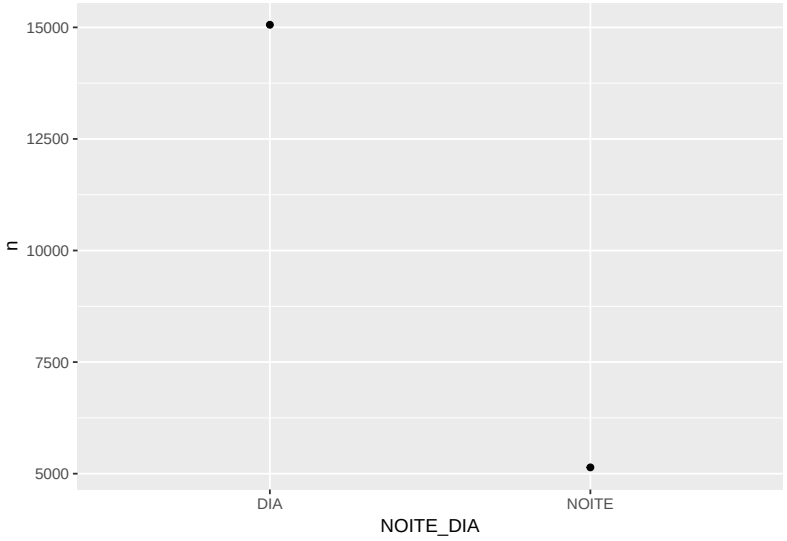
```
##         n
## [1,] TRUE
```

After noticing that all rows were filled out, I decided to analyze how many accidents occurred in each of the shifts of the day. To do so, I first used the `select()` function to get only the shift column (`NOITE_DIA`), then I grouped the data using the `group_by()` function. Then, In order to get a count of the occurrences of each shift I used the `tally()` function.

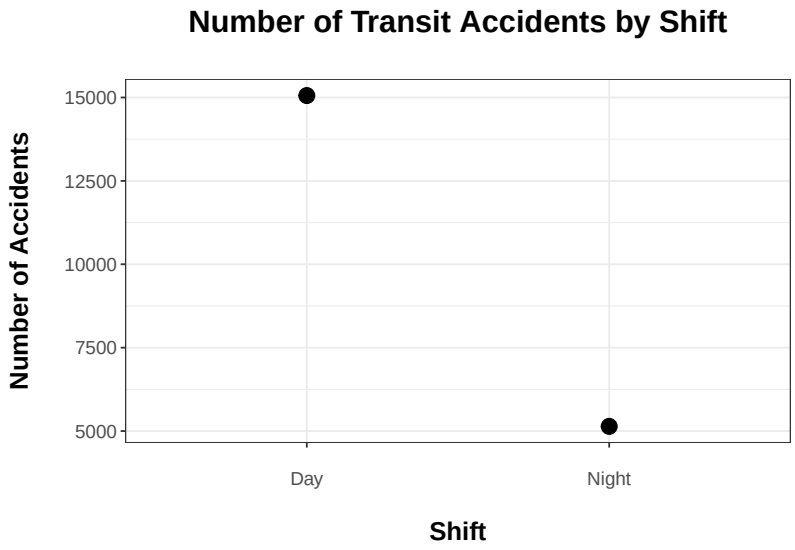
```
 df %>% select(NOITE_DIA) %>% group_by(NOITE_DIA) %>% tally()
```

```
## # A tibble: 3 x 2
##   NOITE_DIA     n
##   <chr>     <int>
## 1 DIA       15058
## 2 NOITE      5141
## 3 <NA>          3
```

After having the number of accidents per shift of the day, I just sent this data as a parameter to `ggplot()` in order to get the chart.



As soon as I realized that I got the chart I wanted, I customized some of the visual parameters of `ggplot()` in order to get a more sophisticated representation of the data.



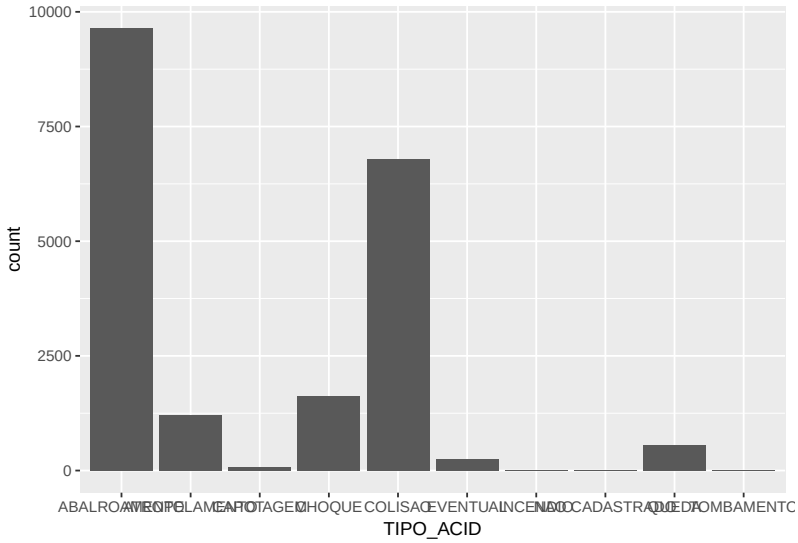
Question 2

Then, I moved to another question: **"What types of accidents are more common?"**. In order to answer it, I first used the `filter()` function to get only the rows in which the column regarding the type of accident (`TIPO_ACID`) was correctly filled. Then, in order to get a easier visualization of the results I called the `select()` function to get rid of the other columns since I was just interested about the kind accident of each occurrence.

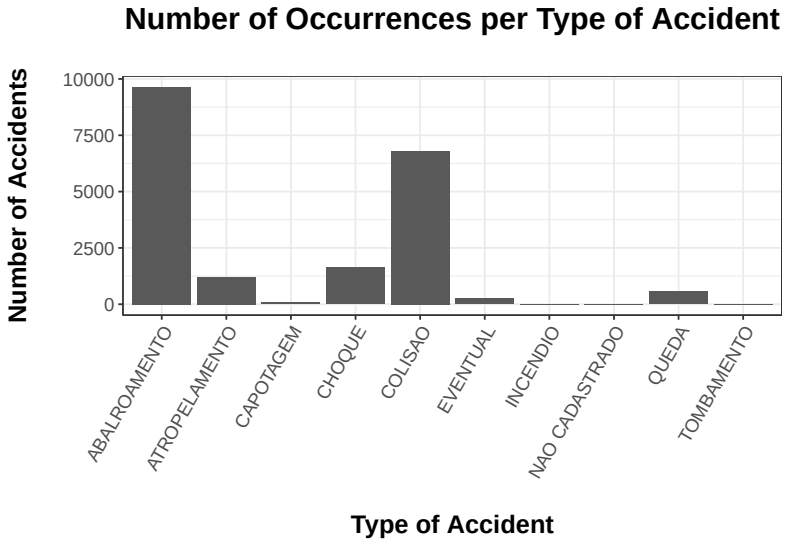
```
df %>% filter(!is.na(TIPO_ACID)) %>% select(TIPO_ACID)
```

```
## # A tibble: 20,202 x 1
##    TIPO_ACID
##    <chr>
##  1 ABALROAMENTO
##  2 ATROPELAMENTO
##  3 ABALROAMENTO
##  4 ABALROAMENTO
##  5 COLISAO
##  6 CAPOTAGEM
##  7 ABALROAMENTO
##  8 CHOQUE
##  9 ABALROAMENTO
## 10 CHOQUE
## # ... with 20,192 more rows
```

Then, I grouped the data according to the type of accident with the `group_by()` function, and passed it as parameter to `ggplot()` to get the chart.



After checking if the data was plotted in the way I wanted, I changed some parameters of `ggplot()` to get a fancier chart.



Question 3

The third question I wanted to answer about the transit accidents was if **the number of deaths increased or decreased throughout the year**. Thus, I first selected the columns regarding number of deaths (`FATAIS`) and the date of the accident (`DATA_HORA`). Then, I look foward to a way to parse the timestamp data provided in `DATA_HORA` in order to get only the month of each of the accidents. Thus, I found in the Internet that I could do this by using the `format()` function. Then, I used the `mutate()` function to create a new column (called `MES`) in the dataset corresponding to the month in which each of the accidents happened.

```
  df %>% select(FATAIS, DATA_HORA) %>% mutate(MES = format(DATA_HORA, "%m"))
```

```
## # A tibble: 20,202 x 3
##    FATAIS DATA_HORA           MES
##     <int> <dttm>              <chr>
##  1      0 2012-01-01 01:00:00 01
##  2      0 2012-01-01 02:00:00 01
```

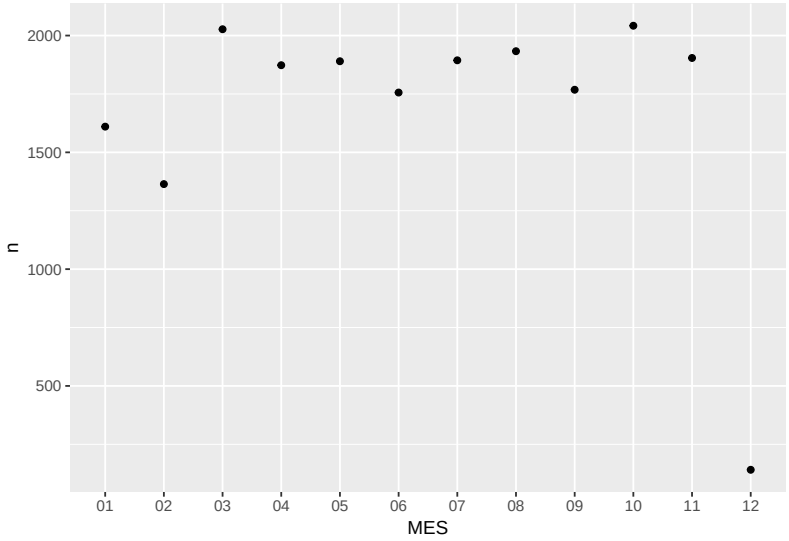
```
##  3      0 2012-01-01 02:30:00 01
##  4      0 2012-01-01 02:50:00 01
##  5      0 2012-01-01 03:00:00 01
##  6      0 2012-01-01 03:20:00 01
##  7      0 2012-01-01 03:20:00 01
##  8      0 2012-01-01 07:00:00 01
##  9      0 2012-01-01 07:20:00 01
## 10      0 2012-01-01 09:10:00 01
## # ... with 20,192 more rows
```

Then, I just had to use the `group_by()` function to group the accidents according to the month in which they happened, and call the `tally()` function to get the count of accidents per month.

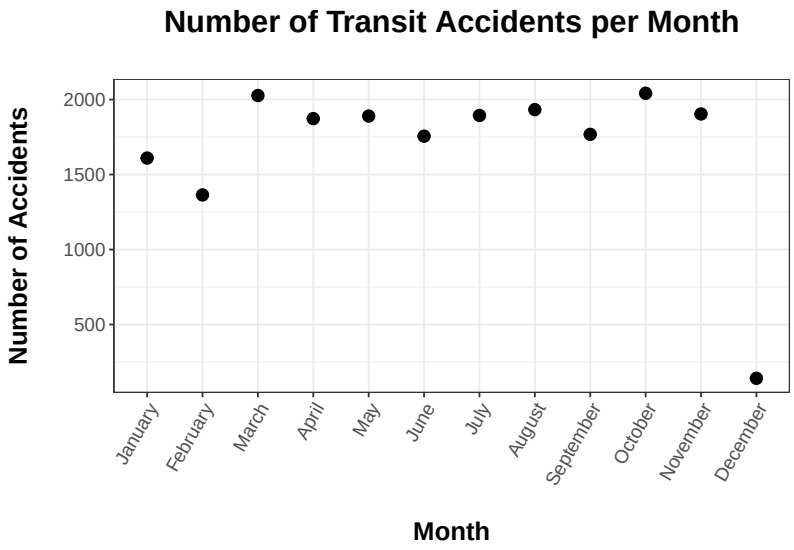
```
df %>% select(FATAIS, DATA_HORA) %>% mutate(MES = format(DATA_HORA, "%m")) %>%
group_by(MES) %>% tally()
```

```
## # A tibble: 12 x 2
##    MES       n
##    <chr> <int>
##  1 01     1610
##  2 02     1364
##  3 03     2027
##  4 04     1873
##  5 05     1890
##  6 06     1756
##  7 07     1894
##  8 08     1933
##  9 09     1768
## 10 10     2042
## 11 11     1904
## 12 12      141
```

After having summarised the accidents per month, I passed it as a parameter to ggplot2 in order to get the chart.



Finally, I just customized some visual properties of `ggplot2()` in order to have a more attractive chart.


Number of Transit Accidents per Month
Number of Accidents
2000
1500
1000
500
January
February
March
April
May
June
July
August
September
October
November
December
Month