LPS_poa_accidents_plots

Francis Moreira

November 8, 2018

This is a summary of the "2_TD_poa_car_accidents.Rmd" file, focused on plots.

First, let's load the libraries:

```
library("tidyverse")
```

```
## Warning: replacing previous import by 'tibble::as_tibble' when loading
## 'broom'
```

```
## Warning: replacing previous import by 'tibble::tibble' when loading 'broom'
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Now let's load the file

```
file = "acidentes-2003.csv"
if(!file.exists(file)){
  download.file("http://www.opendatapoa.com.br/storage/f/2013-11-06T17%3A38%3A06.476Z/acidentes-2003.csv",
    destfile=file)
}

df <- read_csv2("acidentes-2003.csv")
```

```
## Using ',' as decimal and '.' as grouping mark. Use read_delim() for more control.
```

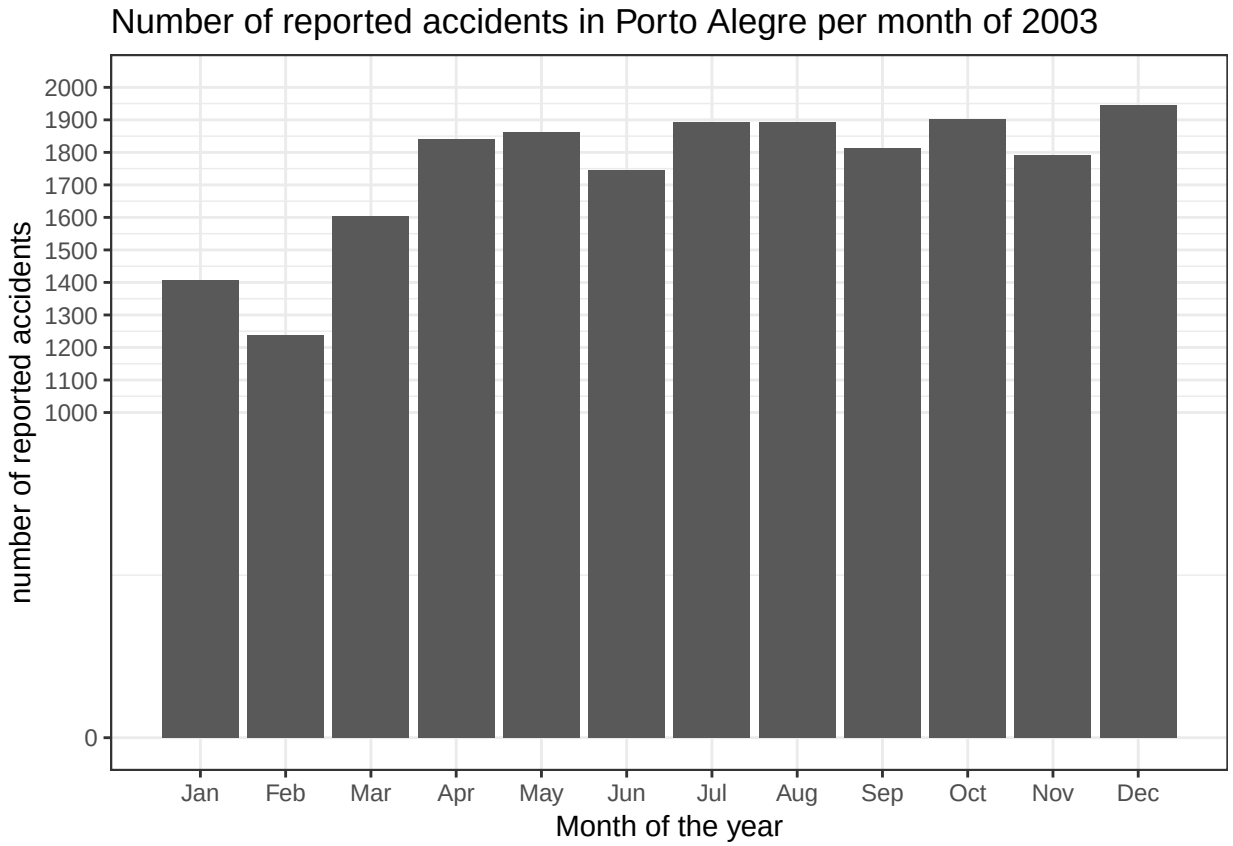
```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   LOCAL_VIA = col_character(),
##   LOG1 = col_character(),
##   LOG2 = col_character(),
##   LOCAL = col_character(),
##   TIPO_ACID = col_character(),
##   DATA_HORA = col_datetime(format = ""),
##   DIA_SEM = col_character(),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

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

Question 1: Is there a time of the year with more accidents? First, we provide a definition for "time" as a month, so the question becomes "Is there a month of the year with more accidents?".

To verify this, we will observe the sum of accidents in each month, then plot with x axis month and y axis count.

```
df %>% group_by(MES) %>%
  summarize(count = n()) %>%
  ggplot(aes(x=as.integer(MES), y=count)) +
  geom_bar(stat="identity") +
  labs(x = "Month of the year",
       y = "number of reported accidents",
       title = "Number of reported accidents in Porto Alegre per month of 2003") +
  scale_x_discrete(limits=c("Jan",
                            "Feb",
                            "Mar",
                            "Apr",
                            "May",
                            "Jun",
                            "Jul",
                            "Aug",
                            "Sep",
                            "Oct",
                            "Nov",
                            "Dec")) +
  theme_bw() +
  scale_y_continuous(breaks=c(0,1000,1100,1200,1300,1400,1500,1600,1700,1800,1900,2000),
                     limits=c(0,2000))
```



As we can see, December has the largest number of reported accidents. The difference is not very significant when compared to May, July, August, and October. February had significantly less accident reports than other months.

Question 2:

How many vehicles are usually involved?

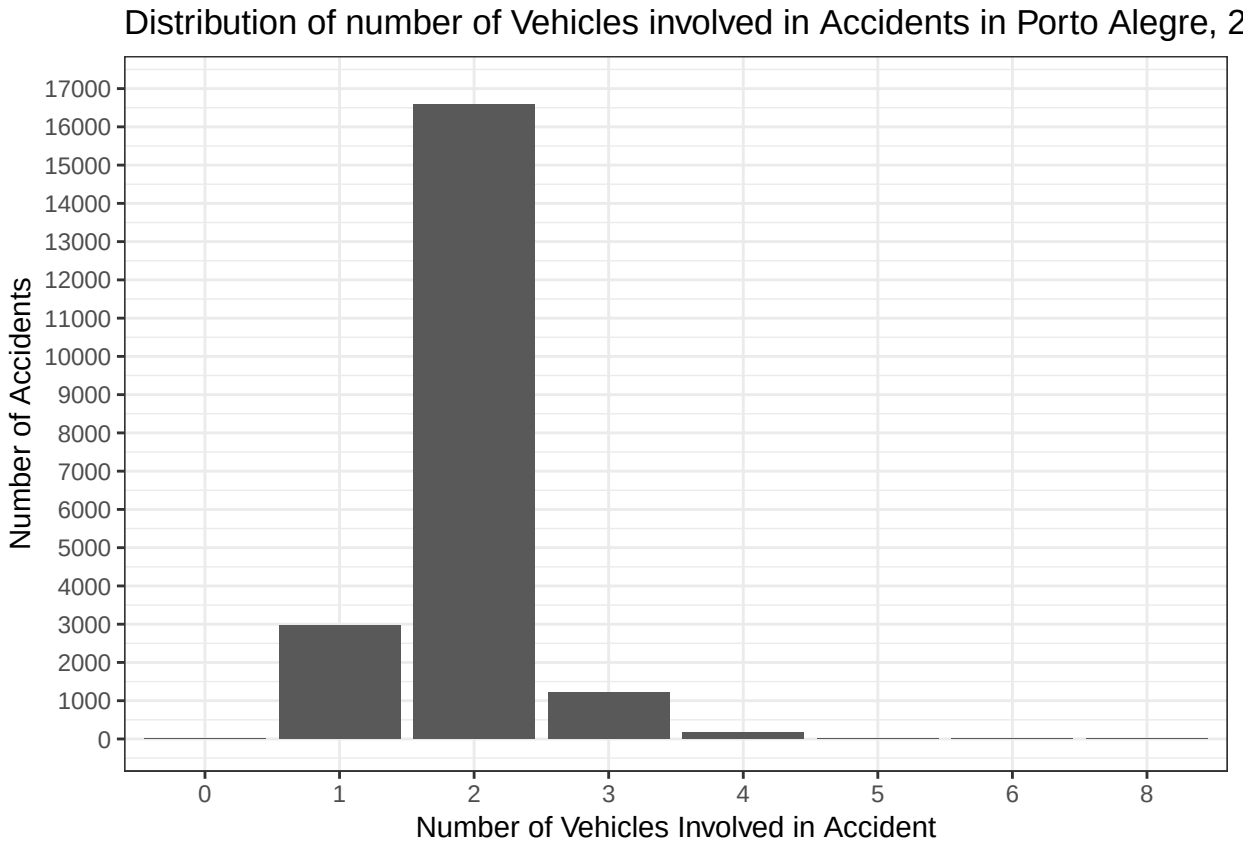
Instead of calculating the mode, we just tabled the vector with the sum of vehicles, and then sorted into descending order to find which is the most common number of vehicles in an accident (2).

```
dd <- df %>%
  mutate(vehicles = AUTO +
           TAXI +
           LOTACAO +
           ONIBUS_URB +
           ONIBUS_INT +
           CAMINHAO +
           MOTO +
           CARROCA +
           BICICLETA +
           OUTRO)
vec <- as.vector(dd['vehicles'])

dvec <- as.data.frame(table(vec))

dvec %>% ggplot(aes(x=vec,y=Freq)) +
  geom_bar(stat='identity') +
  labs(x = "Number of Vehicles Involved in Accident",
```

```
      y = "Number of Accidents",
      title = "Distribution of number of Vehicles involved in Accidents in Porto Alegre, 2003") +
theme_bw() +
scale_y_continuous(breaks=c(0,
                            1000,
                            2000,
                            3000,
                            4000,
                            5000,
                            6000,
                            7000,
                            8000,
                            9000,
                            10000,
                            11000,
                            12000,
                            13000,
                            14000,
                            15000,
                            16000,
                            17000),
                   limits=c(0,17000))
```



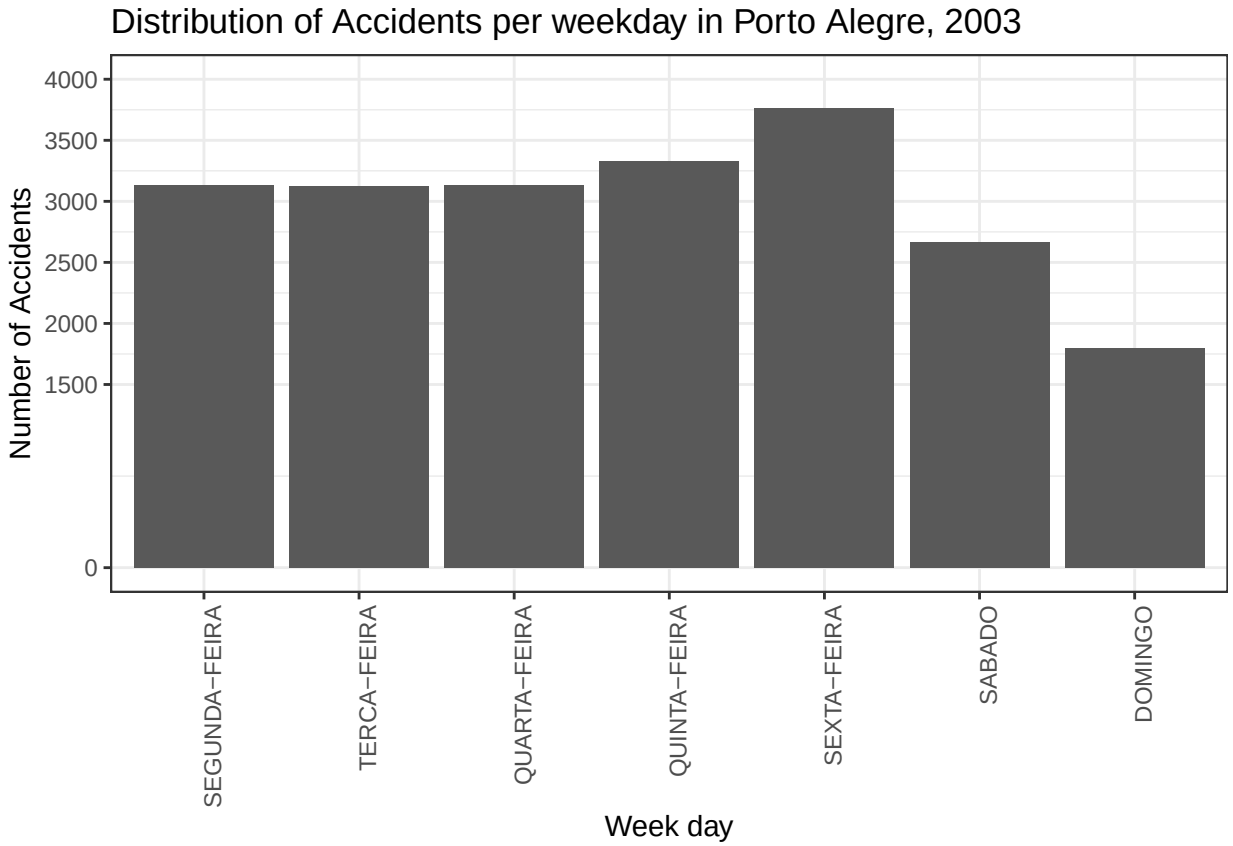
Here we can see that the mode of the distribution is "2", i.e. there were usually two vehicles involved in car accidents in Porto Alegre, 2003.

Question 3:

Is there a specific weekday with more accidents?

Group by weekday, sum instances. However, what is significantly "more"?

```
df %>% group_by(DIA_SEM) %>%
  summarize(count = n()) %>%
  ggplot(aes(
    x = fct_relevel(DIA_SEM,
                    "SEGUNDA-FEIRA",
                    "TERCA-FEIRA",
                    "QUARTA-FEIRA",
                    "QUINTA-FEIRA",
                    "SEXTA-FEIRA",
                    "SABADO",
                    "DOMINGO"),
    y = count)
  ) +
  geom_bar(stat='identity') +
  labs(x = "Week day",
       y = "Number of Accidents",
       title = "Distribution of Accidents per weekday in Porto Alegre, 2003") +
  theme_bw() +
  scale_y_continuous(breaks=c(0,
                              1500,
                              2000,
                              2500,
                              3000,
                              3500,
                              4000),
                     limits=c(0,4000)) +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1))
```



We can see that Friday ("SEXTA-FEIRA") is the day with the largest number of reported accidents.