Handson analysis on the POA accidents dataset

Joao Pedro Oliveira

October 31 2018

This is my hands on analysis of the POA accidents dataset

First, download the dataset

```
file = "acidentes-2016.csv"
if(!file.exists(file)){
  download.file("https://datapoa.com.br/storage/f/2017-08-03T13%3A19%3A45.538Z/acidentes-2016.csv", destfile=file)
}
```

Now, read the CSV file to a Dataframe using readr

```
library(readr)
library(RColorBrewer)
ac_data <- read_delim(file, ";")
```

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

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

```
ac_data
```

ID	LONGITU...	LATITUDE	LOG1	LOG2
<dbl>	<dbl>	<dbl>	<chr>	<chr>
623243	-51.23386	-3.008521e+01	R ARAPEI	R COMANDAI
622413	-51.23195	-3.010831e+01	R PADRE JOAO BATISTA REUS	R JOAO MORA
622460	-51.21203	-3.004587e+01	AV DO LAMI	*NA*
622540	-51.18561	-3.003446e+01	AV DR NILO PECANHA	R CARLOS TREIN FILHO
622181	-51.09736	-3.013143e+01	ESTR JOAO DE OLIVEIRA REMIAO	*NA*
622232	-51.22502	-3.004690e+01	AV IPIRANGA	*NA*
622414	-51.22152	-3.005982e+01	R JOSE DE ALENCAR	*NA*
622186	-51.21841	-3.004594e+01	AV ERICO VERISSIMO	*NA*
622235	-51.21583	-3.004363e+01	R GEN LIMA E SILVA	*NA*
622185	-51.20063	-3.000445e+01	AV EDVALDO PEREIRA PAIVA	*NA*

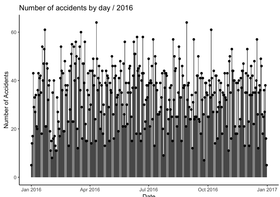
1-10 of 10,000 rows | 1-5 of 44 columns

Previous 1 2 3 4 5 6 .. 1000 Next

As we see, there is a lot of information here.Though at my first look, I can't seem to find any relevant missing data.

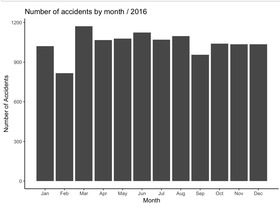
Since for this first analysis we'll be trying to find out if there is a time of the year with more accidents, we'll limit this dataset for this purpose.

```
ac_data %>%
  group_by(DATA) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=DATA, y =QUANT_ACID))+geom_col() +
  geom_point() +
  ggtitle("Number of accidents by day / 2016") +
  xlab("Date") + ylab("Number of Accidents") +
  scale_fill_gradient(low="yellow", high="red") +
  theme_classic()
```



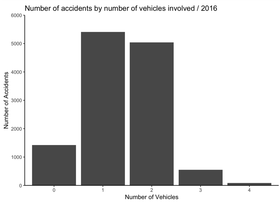
Although this graph shows a lot, it's better for us to analyse and understand the relations between time of the year and accidents if we look at the number of accidents per month.

```
ac_data %>%
  group_by(MES) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=MES, y =QUANT_ACID))+geom_col() +
  ggtitle("Number of accidents by month / 2016") +
  xlab('Month') + ylab("Number of Accidents") +
  scale_x_discrete(limit = c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'))
  +
  theme_classic()
```



Now, let's analyse to learn how many vehicles are usually involved in the accidents.

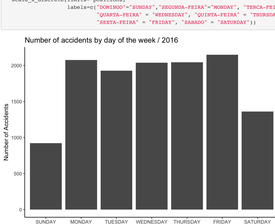
```
ac_data %>%
  group_by(AUTO) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=AUTO, y =QUANT_ACID))+geom_col()+
  theme_classic() +
  ggtitle("Number of accidents by number of vehicles involved / 2016") +
  xlab('Number of Vehicles') + ylab("Number of Accidents") +
  scale_x_continuous(breaks = c(1,2,3,4, 5), expand = c(0,0.6)) +
  scale_y_continuous(breaks = c(0,1000,2000,3000,4000,5000, 6000), expand = c(0,0)) +
  coord_cartesian(xlim = c(0, 4), ylim=c(0,6000))
```



So we can see from here that most accidents happen envolving 1 or 2 vehicles.

Now let's see if there is a certain weekday that has more accidents than others.

```
positions <- c('DOMINGO', 'SEGUNDA-FEIRA', 'TERCA-FEIRA', 'QUARTA-FEIRA', 'QUINTA-FEIRA', 'SEXTA-FEIRA', 'SABADO'
)
ac_data %>%
  group_by(DIA_SEM) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=DIA_SEM, y=QUANT_ACID))+geom_col()+
  theme_classic() +
  ggtitle("Number of accidents by day of the week / 2016") +
  xlab('Day of the week') + ylab('Number of Accidents') +
  scale_x_discrete(limits= positions,
                   labels=c('DOMINGO'='SUNDAY','SEGUNDA-FEIRA'='MONDAY', 'TERCA-FEIRA'='TUESDAY',
                            'QUARTA-FEIRA' = 'WEDNESDAY', 'QUINTA-FEIRA' = 'THURSDAY',
                            'SEXTA-FEIRA' = 'FRIDAY', 'SABADO' = 'SATURDAY'))
```

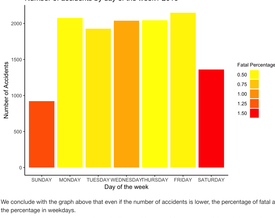


From this graph we can certainly observe some interesting things. The first thing that comes to mind is that there are more accidents on Fridays, usually when people go out to party. And the number of accidents on Saturdays and Sundays are low, maybe because people tend to stay at home during those days.

Another interesting thing to look at is in what days the percentage of fatal accidents is higher.

In this dataset, the fatal accidents are separated into 2 rows: "MORTES" and "MORTE_POST", but the row "FATAIS" shows us the sum of these two rows, with the total number of fatal accidents.

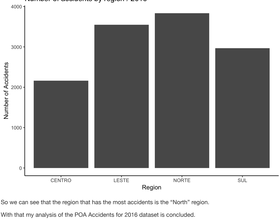
```
positions <- c('DOMINGO', 'SEGUNDA-FEIRA', 'TERCA-FEIRA', 'QUARTA-FEIRA', 'QUINTA-FEIRA', 'SEXTA-FEIRA', 'SABADO'
)
ac_data %>%
  group_by(DIA_SEM) %>%
  summarise(QUANT_ACID = sum(CONT_ACID), Prcnt_fatal = sum(FATAIS)/sum(CONT_ACID)*100) %>%
  ggplot(aes( x=DIA_SEM, y=QUANT_ACID, fill=Prcnt_fatal))+geom_col()+
  scale_fill_gradient(low="yellow", high="red") +
  theme_classic() +
  ggtitle("Number of accidents by day of the week / 2016") +
  xlab('Day of the week') + ylab('Number of Accidents') +
  scale_x_discrete(limits= positions,
                   labels=c('DOMINGO'='SUNDAY','SEGUNDA-FEIRA'='MONDAY', 'TERCA-FEIRA'='TUESDAY',
                            'QUARTA-FEIRA' = 'WEDNESDAY', 'QUINTA-FEIRA' = 'THURSDAY',
                            'SEXTA-FEIRA' = 'FRIDAY', 'SABADO' = 'SATURDAY')) +
  guides(fill = guide_legend(title = "Fatal Percentage", label.position = "left", title.theme=element_text(size=9
)))
```



We conclude with the graph above that even if the number of accidents is lower, the percentage of fatal accidents on weekends is far higher than the percentage in weekdays.

So, lets see if there are regions in Porto Alegre with more accidents than others. For this, I define "Region" as the column 'REGIAO'.

```
ac_data %>%
  subset(REGIAO != "NAO IDENTIFICADO") %>%
  group_by(REGIAO) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=REGIAO, y=QUANT_ACID))+geom_col()+
  theme_classic() +
  ggtitle("Number of accidents by region / 2016") +
  xlab("Region") + ylab("Number of Accidents")
```



So we can see that the region that has the most accidents is the "North" region.

With that my analysis of the POA Accidents for 2016 dataset is concluded.