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("http://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)
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(),
##   NOITE_DIA = col_character(),
##   FONTE = col_character(),
##   BOLETIM = col_character(),
##   REGIAO = col_character(),
##   CONSORCIO = col_character()
## )
```

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

```
ac_data
```

ID <int>	LONGITU... <dbl>	LATITUDE <dbl>	LOG1 <chr>	LOG2 <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

We need to get a grasp for what is in our dataset

```
summary(ac_data)
```

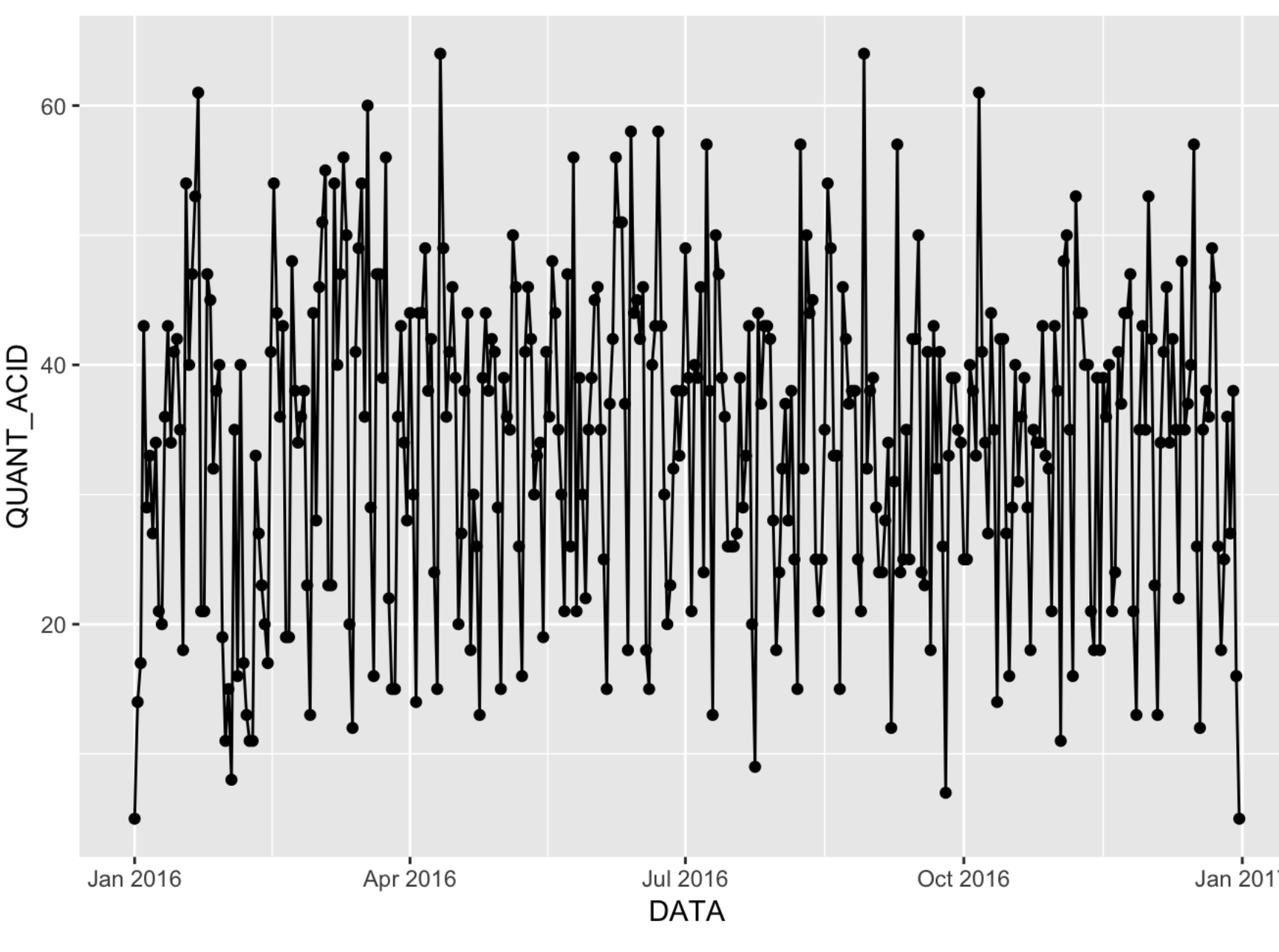
```
##        ID            LONGITUDE         LATITUDE            LOG1          
##  Min.   :622181   Min.   :-51.27   Min.   :-29999977   Length:12515      
##  1st Qu.:625918   1st Qu.:-51.22   1st Qu.:       -30   Class :character  
##  Median :629367   Median :-51.19   Median :       -30   Mode  :character  
##  Mean   :629344   Mean   :-51.17   Mean   : -3012386                     
##  3rd Qu.:632774   3rd Qu.:-51.16   3rd Qu.:       -30                     
##  Max.   :637678   Max.   :-30.05   Max.   :       -30                     
##                                                                           
##      LOG2             PREDIAL1        LOCAL            TIPO_ACID        
##  Length:12515       Min.   :    0   Length:12515       Length:12515      
##  Class :character   1st Qu.:    0   Class :character   Class :character  
##  Mode  :character   Median :  391   Mode  :character   Mode  :character  
##                     Mean   : 1267                                        
##                     3rd Qu.: 1563                                        
##                     Max.   :15555                                        
##                                                                          
##   LOCAL_VIA           QUEDA_ARR              DATA           
##  Length:12515       Min.   :0.0000000   Min.   :2016-01-01  
##  Class :character   1st Qu.:0.0000000   1st Qu.:2016-04-04  
##  Mode  :character   Median :0.0000000   Median :2016-06-30  
##                     Mean   :0.0001598   Mean   :2016-07-01  
##                     3rd Qu.:0.0000000   3rd Qu.:2016-09-30  
##                     Max.   :1.0000000   Max.   :2016-12-31  
##                                                             
##    DATA_HORA                     DIA_SEM              HORA         
##  Min.   :2016-01-01 05:45:00   Length:12515       Length:12515      
##  1st Qu.:2016-04-04 17:27:30   Class :character   Class1:hms        
##  Median :2016-06-30 13:30:00   Mode  :character   Class2:difftime   
##  Mean   :2016-07-02 08:12:41                      Mode  :numeric    
##  3rd Qu.:2016-09-30 13:35:30                                        
##  Max.   :2016-12-31 21:13:00                                        
##                                                                     
##     FERIDOS         FERIDOS_GR          MORTES           MORTE_POST      
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000  
##  Median :0.0000   Median :0.00000   Median :0.000000   Median :0.000000  
##  Mean   :0.4048   Mean   :0.03052   Mean   :0.003756   Mean   :0.003596  
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000  
##  Max.   :9.0000   Max.   :2.00000   Max.   :2.000000   Max.   :1.000000  
##                                                                          
##      FATAIS               AUTO            TAXI            LOTACAO       
##  Min.   :0.000000   Min.   :0.000   Min.   :0.00000   Min.   :0.00000  
##  1st Qu.:0.000000   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000  
##  Median :0.000000   Median :1.000   Median :0.00000   Median :0.00000  
##  Mean   :0.007351   Mean   :1.399   Mean   :0.09061   Mean   :0.02197  
##  3rd Qu.:0.000000   3rd Qu.:2.000   3rd Qu.:0.00000   3rd Qu.:0.00000  
##  Max.   :2.000000   Max.   :7.000   Max.   :4.00000   Max.   :2.00000  
##                                                                         
##   ONIBUS_URB       ONIBUS_MET        ONIBUS_INT          CAMINHAO     
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.00000   Median :0.000000   Median :0.0000  
##  Mean   :0.0628   Mean   :0.01231   Mean   :0.009109   Mean   :0.1134  
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.0000  
##  Max.   :3.0000   Max.   :2.00000   Max.   :2.000000   Max.   :2.0000  
##                                                                       
##       MOTO           CARROCA    BICICLETA           OUTRO         
##  Min.   :0.0000   Min.   :0   Min.   :0.00000   Min.   :0.000000  
##  1st Qu.:0.0000   1st Qu.:0   1st Qu.:0.00000   1st Qu.:0.000000  
##  Median :0.0000   Median :0   Median :0.00000   Median :0.000000  
##  Mean   :0.2363   Mean   :0   Mean   :0.01159   Mean   :0.003196  
##  3rd Qu.:0.0000   3rd Qu.:0   3rd Qu.:0.00000   3rd Qu.:0.000000  
##  Max.   :2.0000   Max.   :0   Max.   :2.00000   Max.   :1.000000  
##                                                                   
##     TEMPO            NOITE_DIA            FONTE          
##  Length:12515       Length:12515       Length:12515      
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##                                                          
##    BOLETIM             REGIAO               DIA             MES        
##  Length:12515       Length:12515       Min.   : 1.00   Min.   : 1.000  
##  Class :character   Class :character   1st Qu.: 8.00   1st Qu.: 4.000  
##  Mode  :character   Mode  :character   Median :16.00   Median : 6.000  
##                                        Mean   :15.69   Mean   : 6.526  
##                                        3rd Qu.:23.00   3rd Qu.: 9.000  
##                                        Max.   :31.00   Max.   :12.000  
##                                                                        
##       ANO          FX_HORA        CONT_ACID    CONT_VIT     
##  Min.   :2016   Min.   : 0.00   Min.   :1   Min.   :0.0000  
##  1st Qu.:2016   1st Qu.: 9.00   1st Qu.:1   1st Qu.:0.0000  
##  Median :2016   Median :13.00   Median :1   Median :0.0000  
##  Mean   :2016   Mean   :12.81   Mean   :1   Mean   :0.3394  
##  3rd Qu.:2016   3rd Qu.:16.00   3rd Qu.:1   3rd Qu.:1.0000  
##  Max.   :2016   Max.   :23.00   Max.   :1   Max.   :1.0000  
##                 NA's   :3                                   
##       UPS           CONSORCIO            CORREDOR       
##  Min.   : 1.000   Length:12515       Min.   :0.000000  
##  1st Qu.: 1.000   Class :character   1st Qu.:0.000000  
##  Median : 1.000   Mode  :character   Median :0.000000  
##  Mean   : 2.414                      Mean   :0.001039  
##  3rd Qu.: 5.000                      3rd Qu.:0.000000  
##  Max.   :13.000                      Max.   :1.000000  
## 
```

As we see, there is a lot of information here.To work with this data we need to do some cleaning.

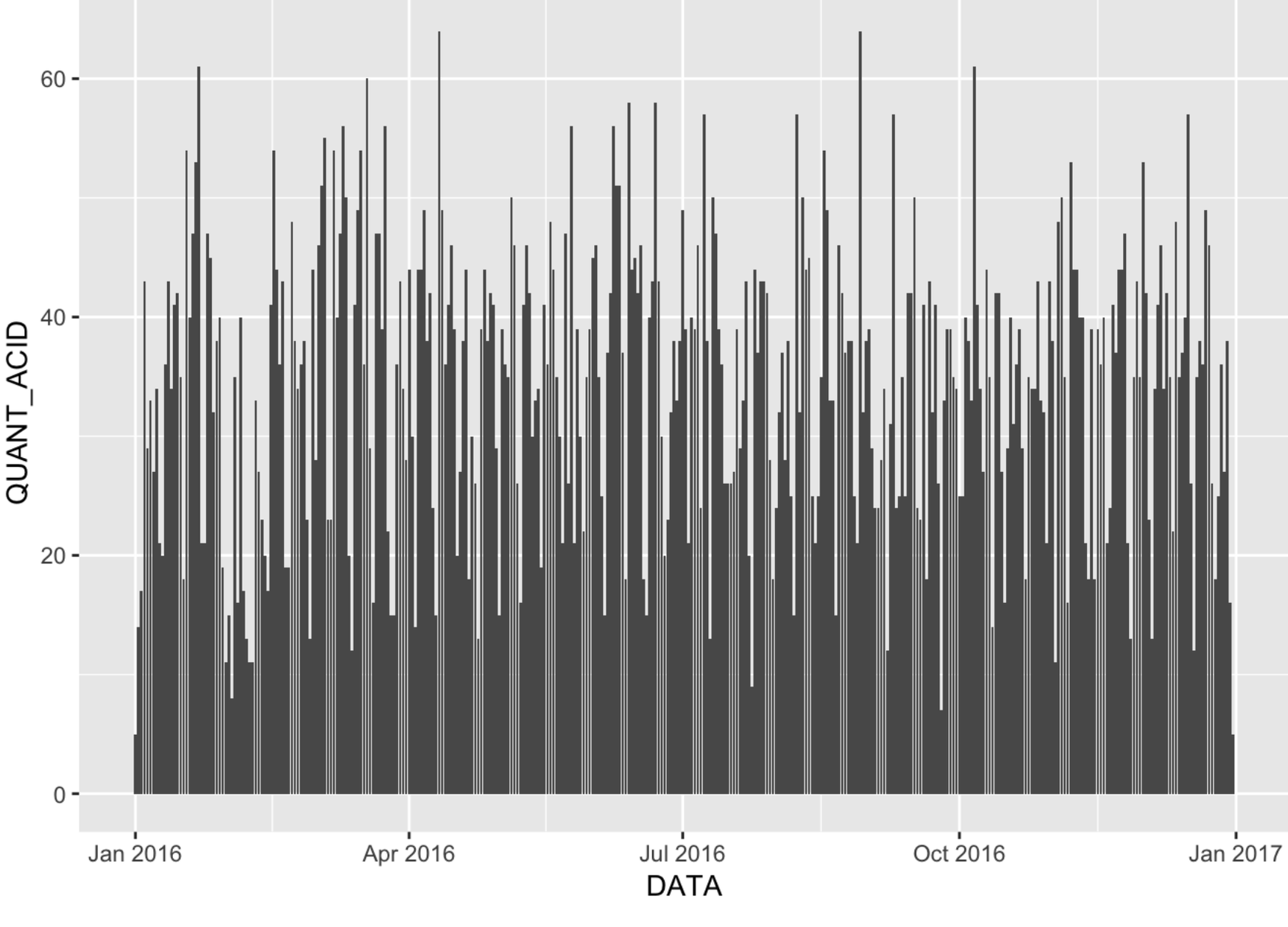
Since for this first analysis we'll be trying to find out if there is a time of the year with more accidents, we'll limit this dataset for this pourpose.

Here we group our data by date and summarise it to a dataframe with only two columns. Then, we plot it to see the number of occurences by date.

```
ac_data %>%
  group_by(DATA) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=DATA, y =QUANT_ACID))+geom_line() + geom_point()
```



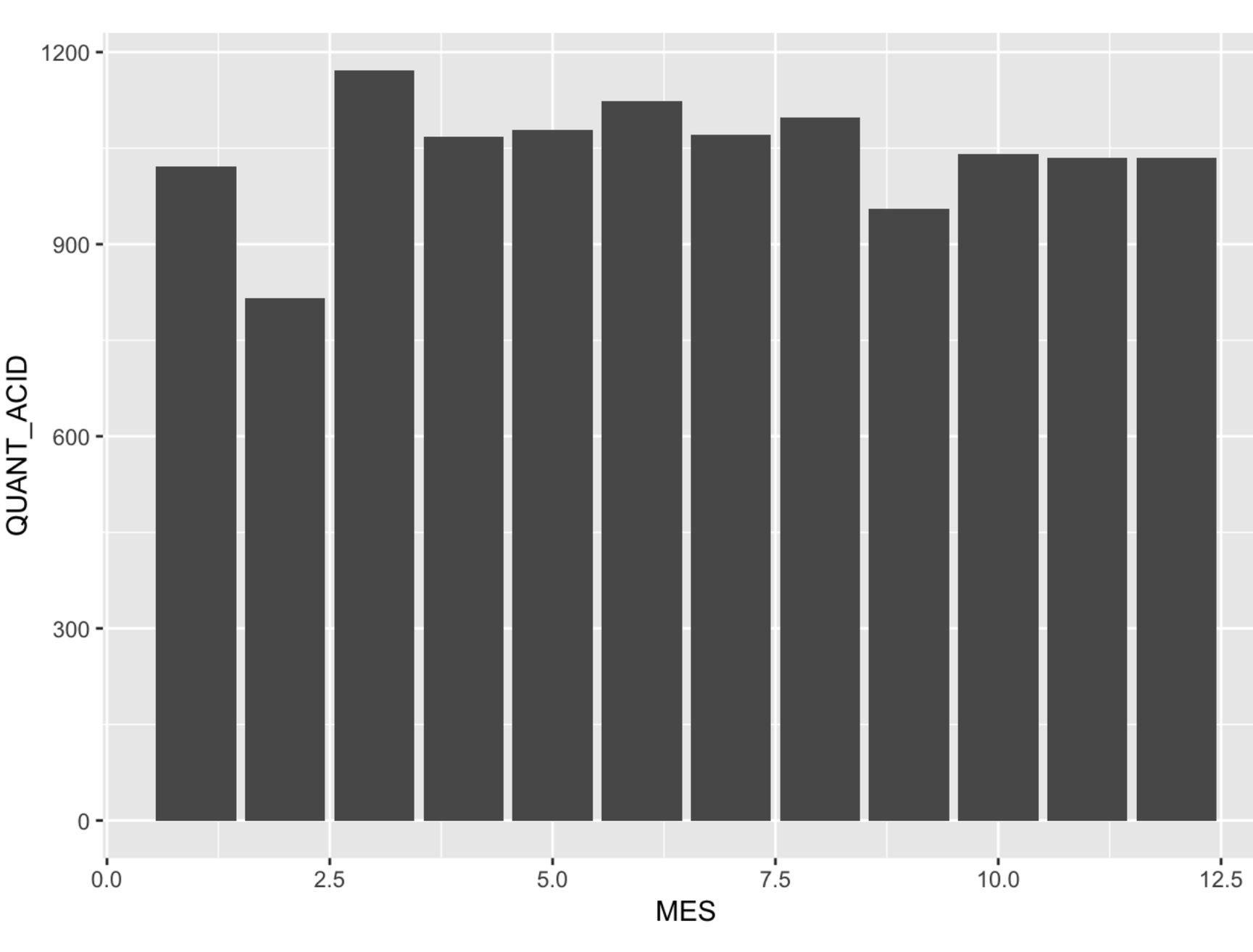

```
ac_data %>%
  group_by(DATA) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=DATA, y =QUANT_ACID))+geom_col()
```



As we can see, there seems to be no correlation between the time of the year and the accidents.

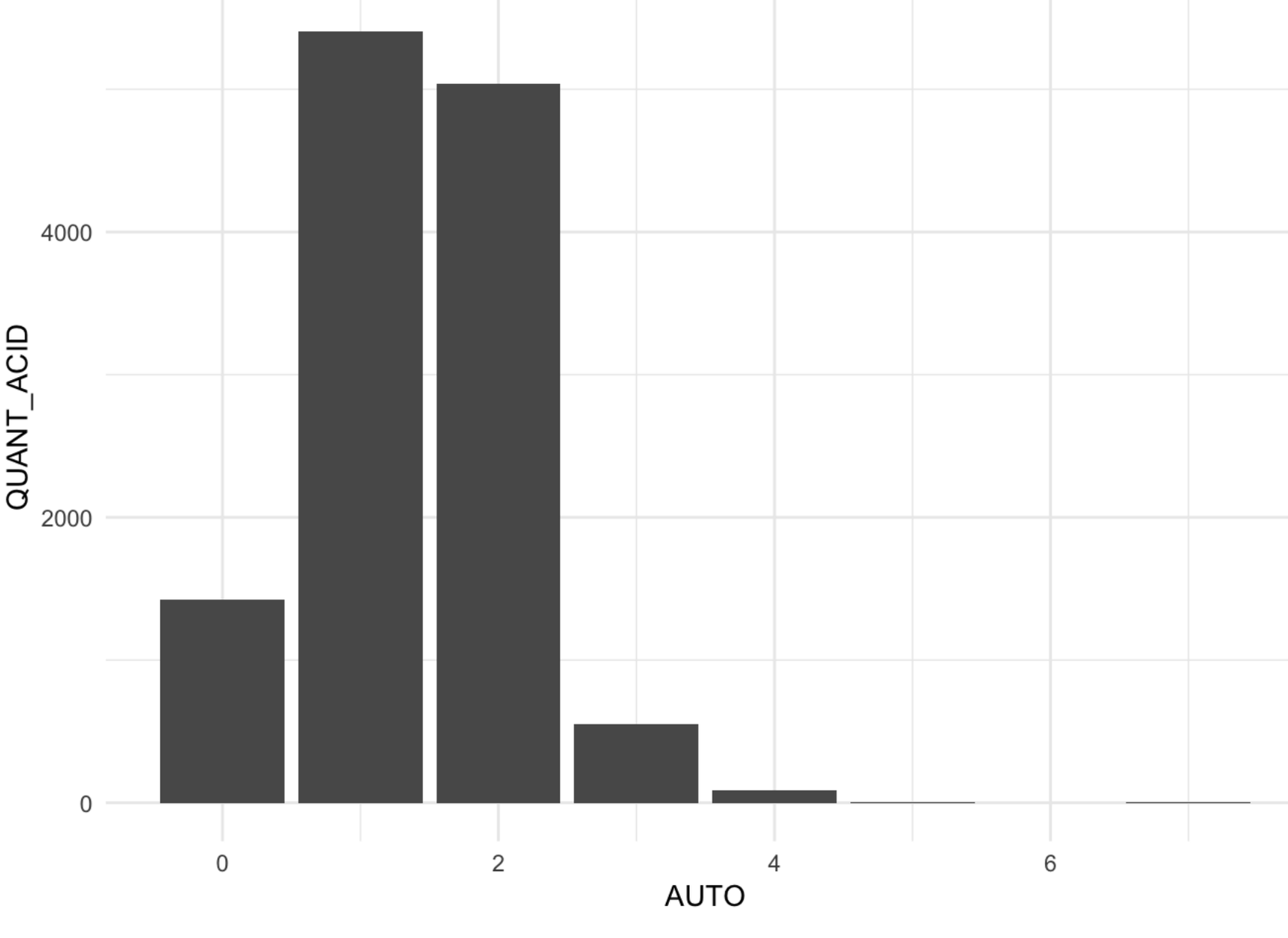
```
ac_data %>%
  group_by(MES) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=MES, y =QUANT_ACID))+geom_col(count=12,binwidth = 1)
```

```
## Warning: Ignoring unknown parameters: count, binwidth
```



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

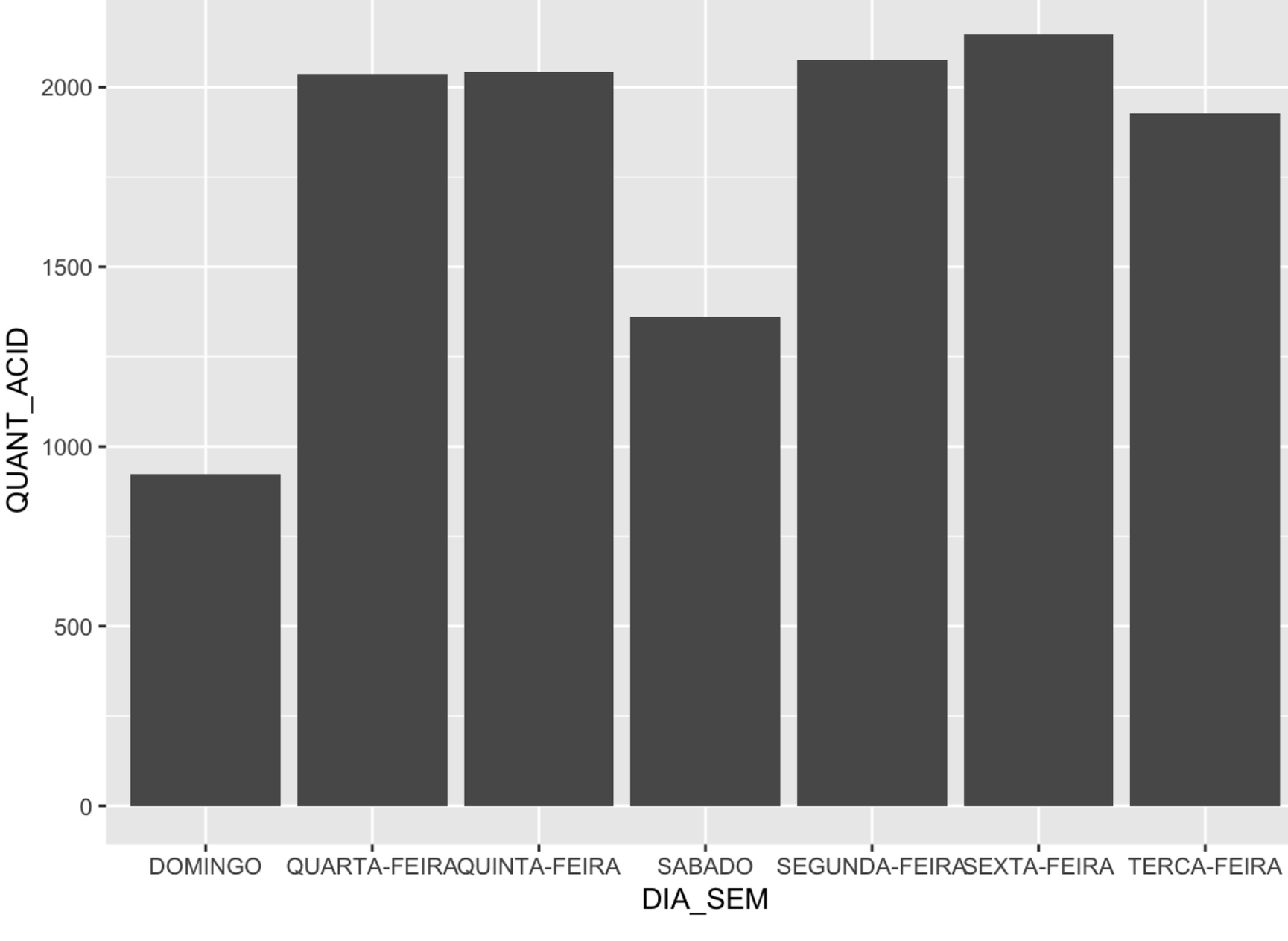
```
ac_data %>%
  group_by(AUTO) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=AUTO, y =QUANT_ACID))+geom_col()+theme_minimal()
```



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

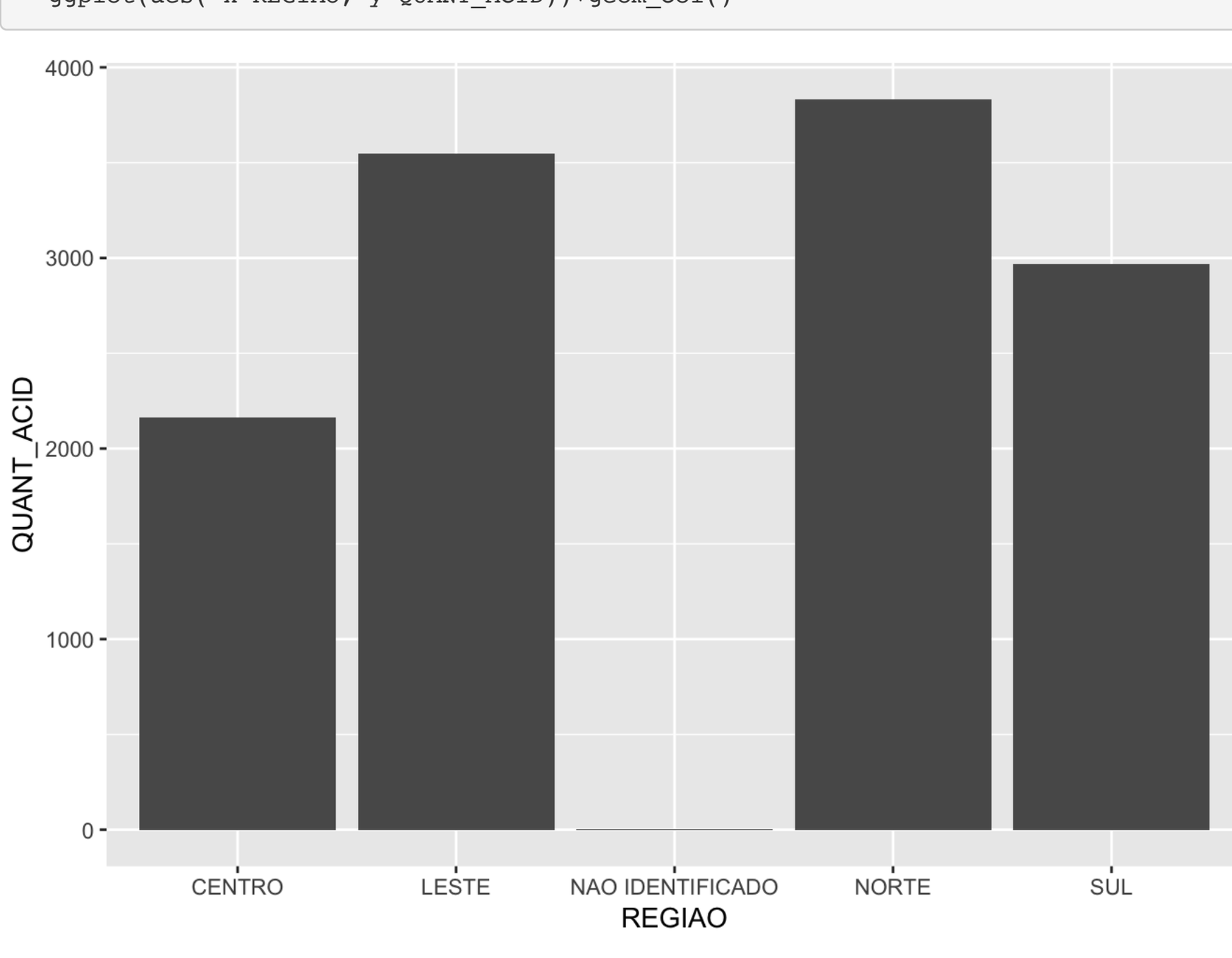
```
ac_data %>%
  group_by(DIA_SEM) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=DIA_SEM, y=QUANT_ACID))+geom_col()
```



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.

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 %>%
  group_by(REGIAO) %>%
  summarise(QUANT_ACID = sum(CONT_ACID)) %>%
  ggplot(aes( x=REGIAO, y=QUANT_ACID))+geom_col()
```



Now, this seems to be good enough. One thing i'm concerned, though,it's to see the percentages of the total accidents by region. This might be more interesting.

```
ac_data %>%
  group_by(REGIAO) %>%
  summarise(PERCNT_ACID = sum(CONT_ACID)/nrow(ac_data) * 100) %>%
  ggplot(aes( x=REGIAO, y=PERCNT_ACID))+geom_col()
```