Crime Analysis of car theft in Sao Paulo City

Saulo Galdino Alves Guedes

16/06/2020

INTRODUCTION

In this project we are going to analyze the crime of car theft in the city of Sao Paulo, from 2016 to April of 2020, and we will also try to build a model to predict crimes. To this project, we will use the data from the state public security agency page, which you can visit here: http://www.ssp.sp.gov.br/transparenciassp/Consulta.aspx

Analysis

First of all, it's necessary import the data from the state public security agency page. The bad news is that is not easy to export information from the page, as it only allows you to extract it month by month. But I have a good news: I suffered to get the data and you can find it on my github https://github.com/Saulgal/CrimePrediction/tree/BaseDadosExcel These are the data like in the page of the agency, e.g. with the crimes from across the state, with some inconsistencies on the date or location. Our intent is build a model only to the city of Sao Paulo, so I created a new base from these data, cleared of unnecessary data. The data is imported directly when you run the script, but if you want to get it for other purposes, you can find it here: https://github.com/Saulgal/CrimePrediction/tree/BaseCrimesFinal

OK, after import the information, we can see how is the data (just the head to get the idea):

```
## [1] TRUE
```

```
## [1] TRUE
```

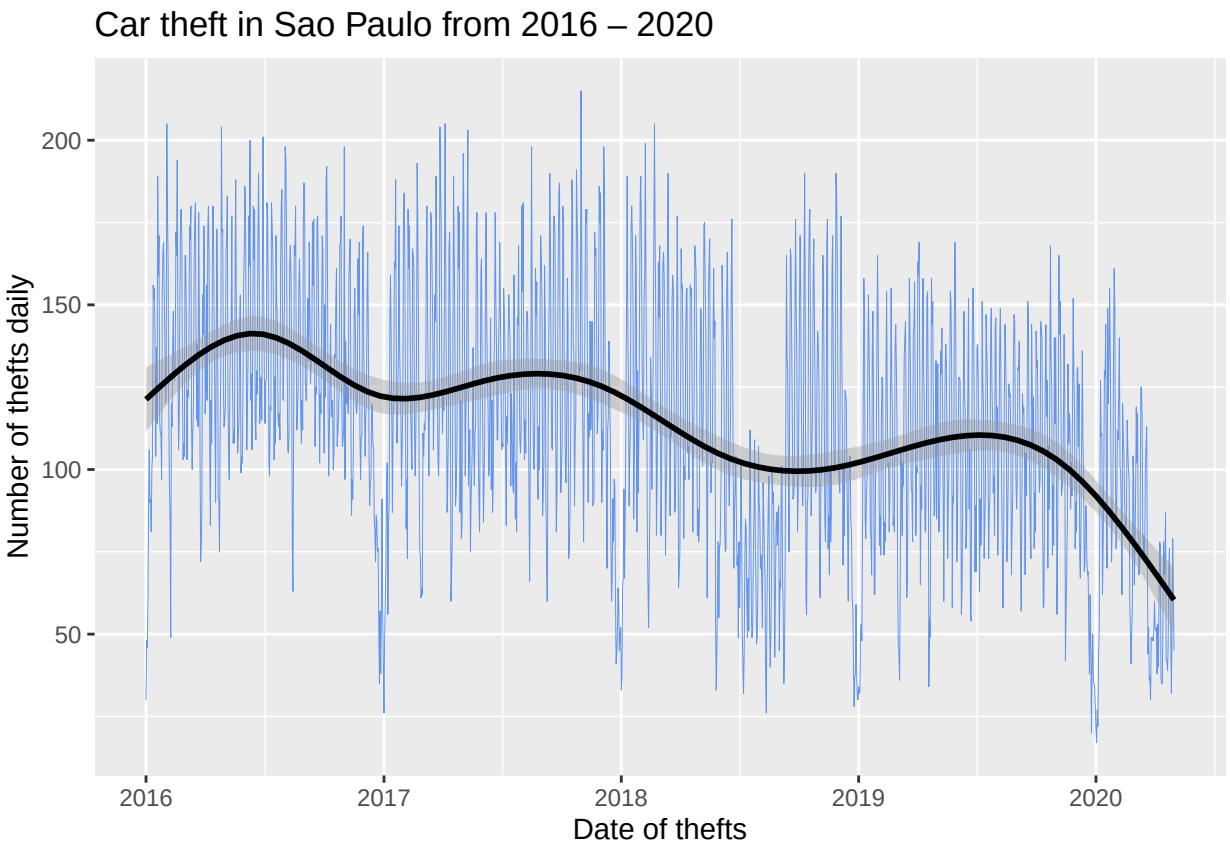
```
## # A tibble: 6 x 57
##   ANO_BO NUM_BO NUMERO_BOLETIM BO_INICIADO         BO_EMITIDO
##    <dbl>  <dbl> <chr>          <dttm>              <dttm>
## 1   2019     47 47/2019        2019-01-04 11:15:56 2019-01-04 11:55:55
## 2   2019     78 78/2019        2019-01-04 13:55:21 2019-01-04 15:28:29
## 3   2019  22388 22388/2019     2019-01-06 23:13:58 2019-01-06 23:14:07
## 4   2019  24508 24508/2019     2019-01-07 12:31:12 2019-01-07 12:31:45
## 5   2019  24725 24725/2019     2019-01-07 14:22:31 2019-01-07 14:22:34
## 6   2019  31458 31458/2019     2019-01-08 15:42:54 2019-01-08 15:42:52
## # ... with 52 more variables: DATAOCORRENCIA <dttm>, HORAOCORRENCIA <dttm>,
## #   PERIDOOCORRENCIA <chr>, DATACOMUNICACAO <dttm>, DATAELABORACAO <dttm>,
## #   BO_AUTORIA <chr>, FLAGRANTE <chr>, NUMERO_BOLETIM_PRINCIPAL <chr>,
## #   LOGRADOURO <chr>, NUMERO <dbl>, BAIRRO <chr>, CIDADE <chr>, UF <chr>,
## #   LATITUDE <dbl>, LONGITUDE <dbl>, DESCRICAOLOCAL <chr>, EXAME <chr>,
## #   SOLUCAO <chr>, DELEGACIA_NOME <chr>, DELEGACIA_CIRCUNSCRICAO <chr>,
## #   ESPECIE <chr>, RUBRICA <chr>, DESDOBRAMENTO <chr>, STATUS <chr>,
```

```
## #   NOMEPESSOA <lgl>, TIPOPESSOA <lgl>, VITIMAFATAL <lgl>, RG_UF <lgl>,
## #   NATURALIDADE <lgl>, NACIONALIDADE <lgl>, SEXO <lgl>, DATANASCIMENTO <lgl>,
## #   ESTADOCIVIL <lgl>, PROFISSAO <lgl>, GRAUINSTRUCAO <lgl>, CORCUTIS <lgl>,
## #   NATUREZAVINCULADA <lgl>, TIPOVINCULO <lgl>, RELACIONAMENTO <lgl>,
## #   PARENTESCO <lgl>, PLACA_VEICULO <chr>, UF_VEICULO <chr>,
## #   CIDADE_VEICULO <chr>, DESCR_COR_VEICULO <chr>, DESCR_MARCA_VEICULO <chr>,
## #   ANO_FABRICACAO <dbl>, ANO_MODELO <dbl>, DESCR_TIPO_VEICULO <chr>,
## #   QUANT_CELULAR <lgl>, MARCA_CELULAR <lgl>, DEL_NUM <chr>, DateCrime <date>
```

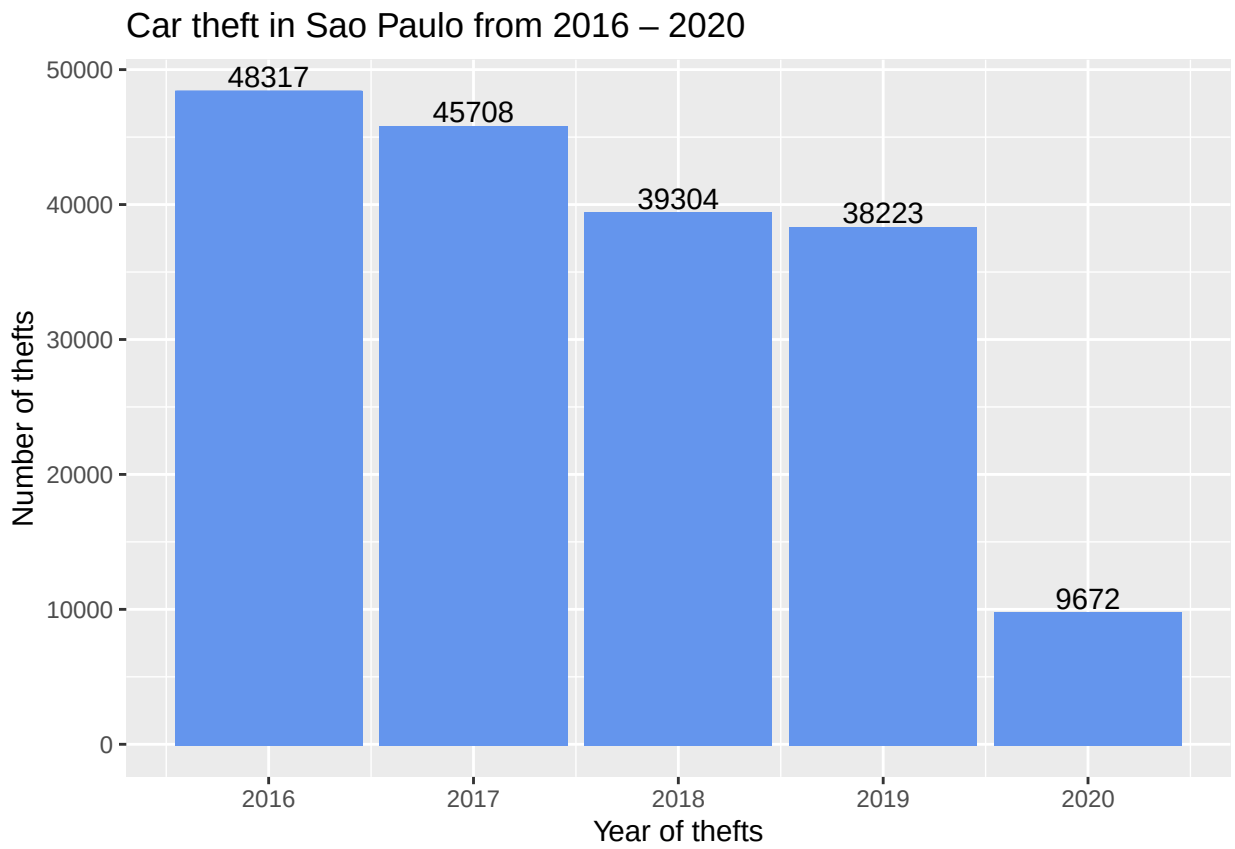
Ok. We have a lot of variables, including when the crime occurred (column: DATAOCORRENCIA), where (LATITUDE, LONGITUDE, LOGRADOURO (means street or avenue, in Portuguese)) and for some thefts, we know the hours when it occurs. With it, we can answer some questions

Temporal analysis

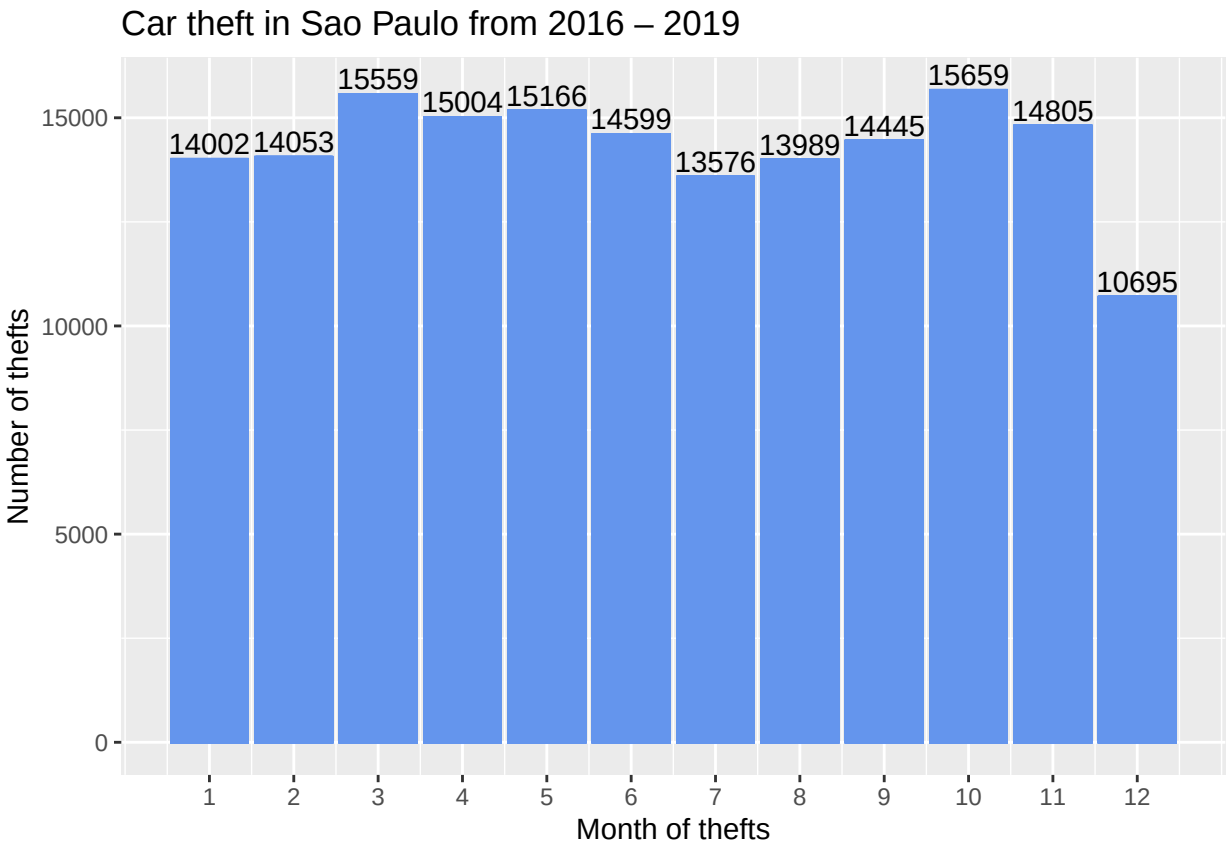
Maybe the first question that appears in our mind when we think about crimes is: They are increasing or decreasing? For answer it, we can plot a time line with the number of thefts:



We can see that are a slight decreasing from 2016 to 2019 and a very declined curve in 2020. It is, probably, due the COVID-19 quarantine, started in the state in March. We believe that if there is less cars in the streets, the thief has less windows of opportunities to do the thefts. We can see the total by years to verify if there is a year with much more crimes:

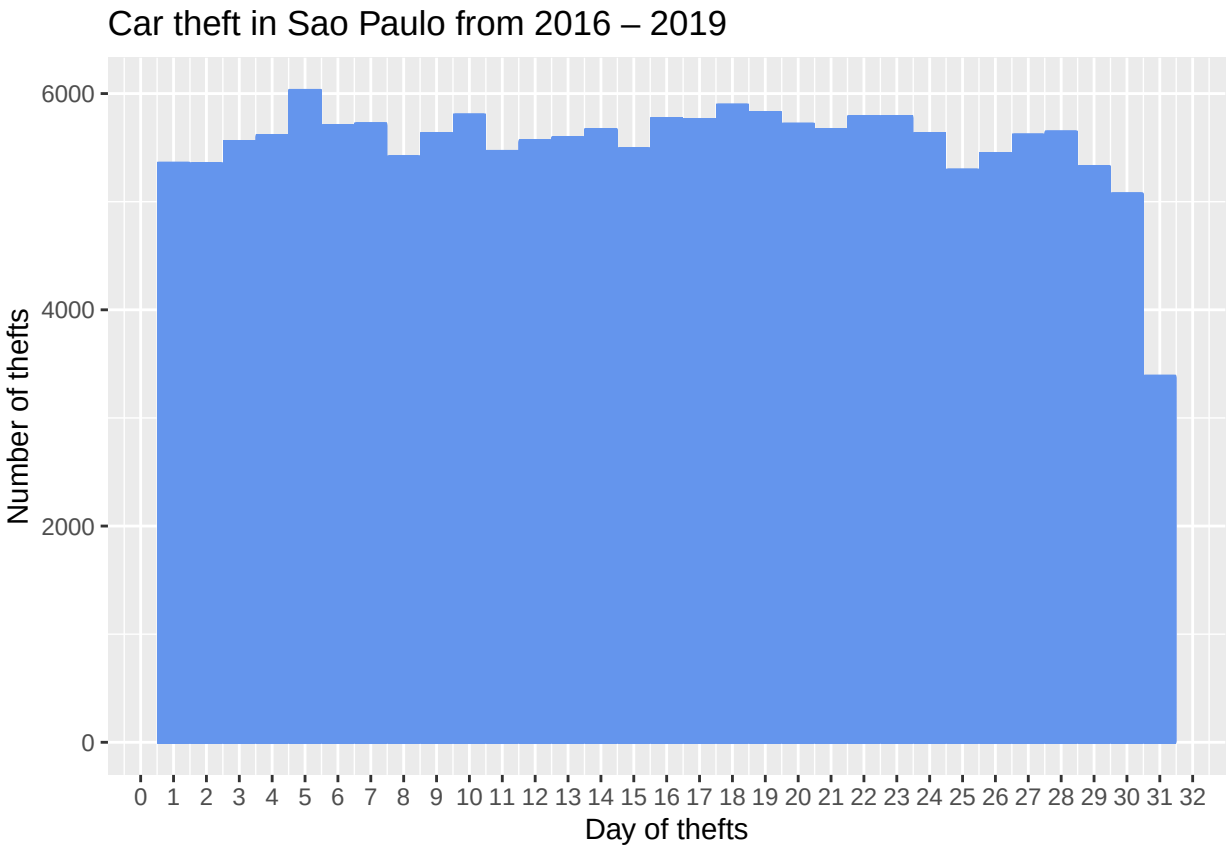


Another important question is if the month has a implication in the numbers (we drop the year of 2020 because it does not have all months).

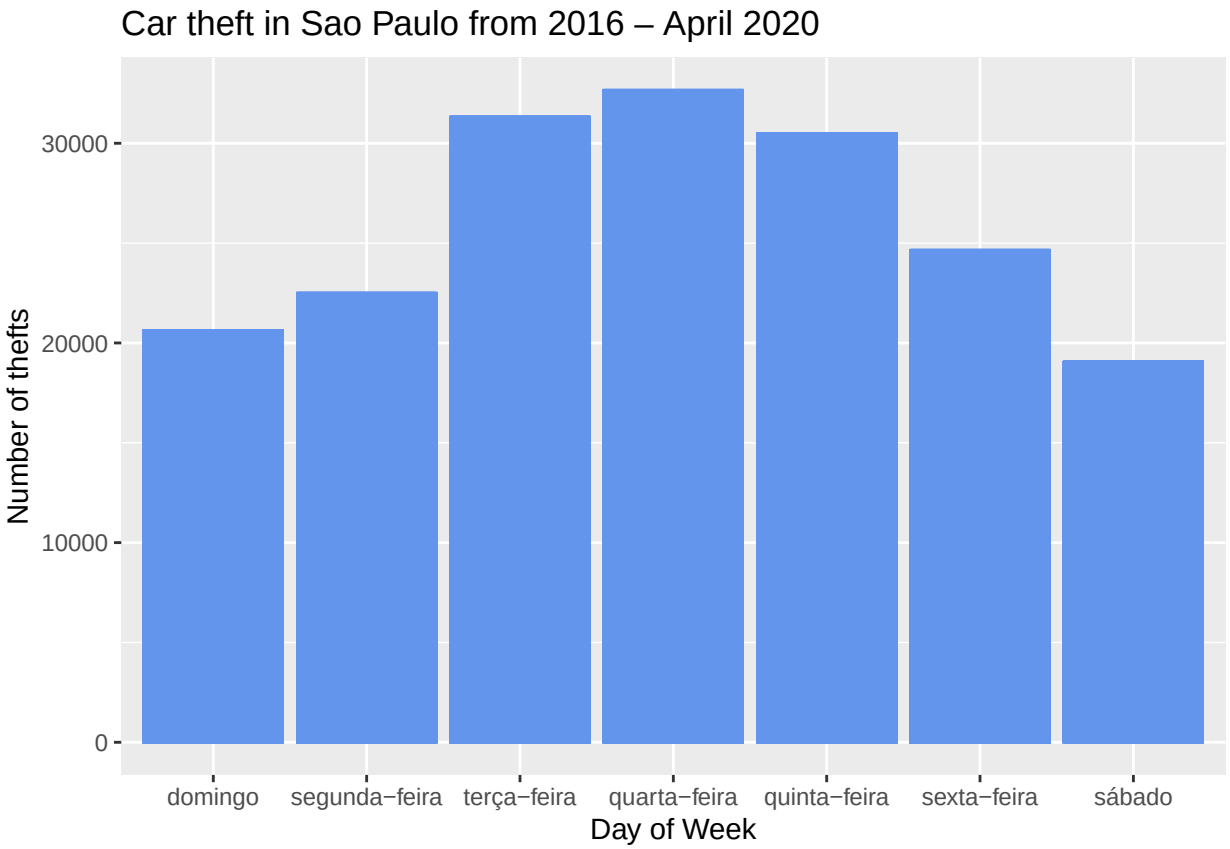


We see that there is less crimes in December than any others months. Traditionally, this is a month of vacations, when the people keep the cars in garage and, again,the thief has less windows of opportunities to do the thefts. The months of January, February and July have a slight fall if compared with another months. The explanation is similar to that occurs in December: months of vacations and holidays.

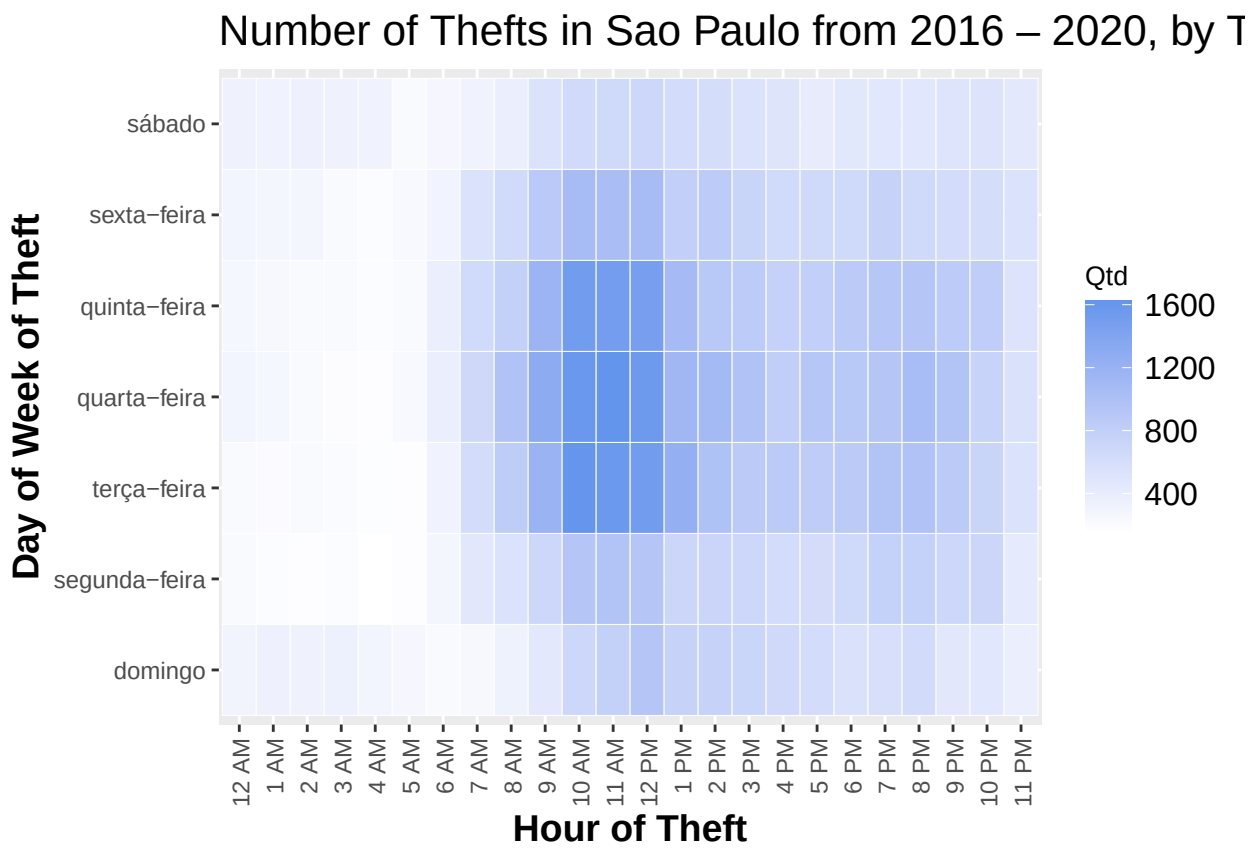
We can also verify if the day of the month has a important role in the theft curve:



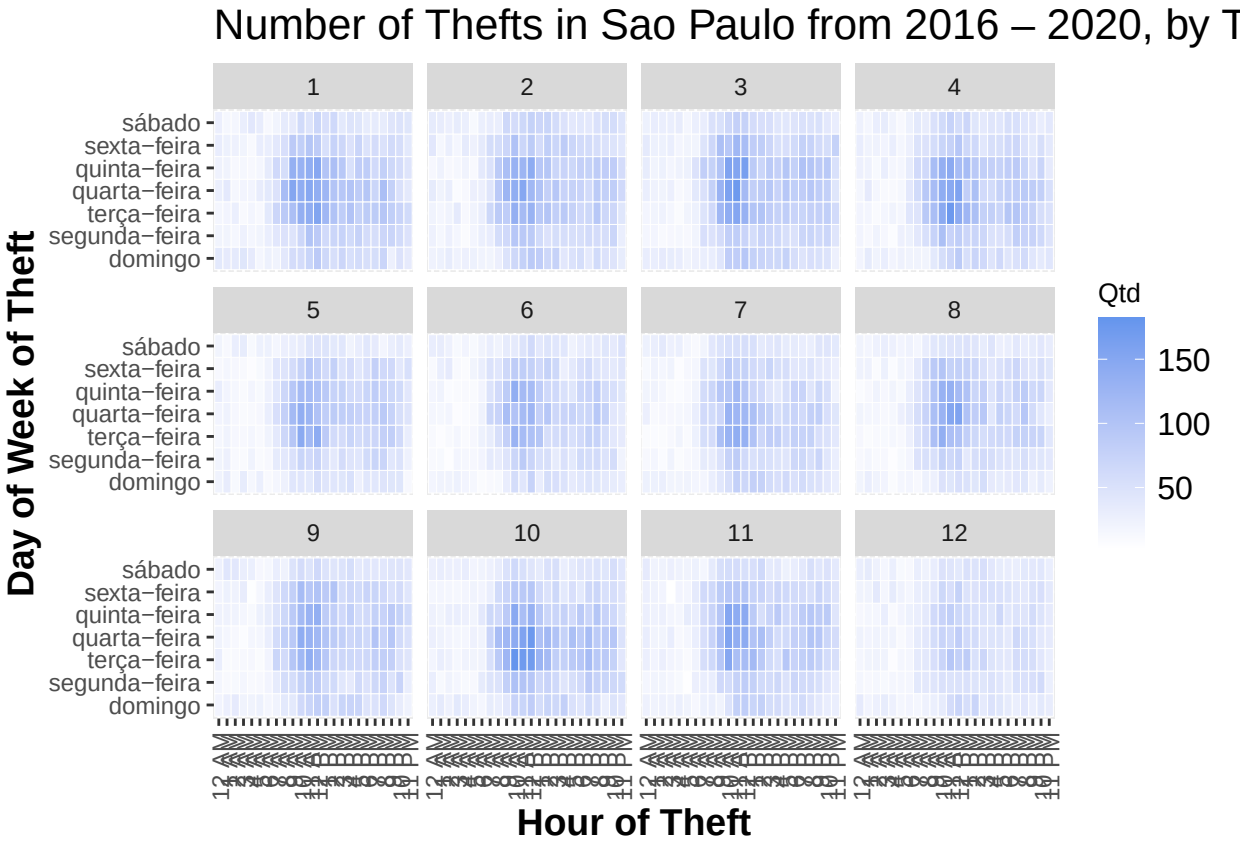
Obviously, the 31th day has less crimes due only half of months in the year has 31 days. Anyway, we can see that some days have more crime than others, like the 5th. Why? Maybe, just maybe, there is a correlation with the payment day, that traditionally occurs int 5th for many companies. I don't know, but we can use this information to build our model. Another important visualization is whether the day of the week has some impact



Here we can see that the day of the week has a huge impact. On weekends, Fridays and Mondays, the numbers of thefts is less than other days. (Domingo means Sunday in Portuguese, Segunda-feira means Mondays and so on) And, finally, we can see whether hour has a important role. Notwithstanding, our data has a lot of not informed hour for theft, but we can try visualize it anyway:

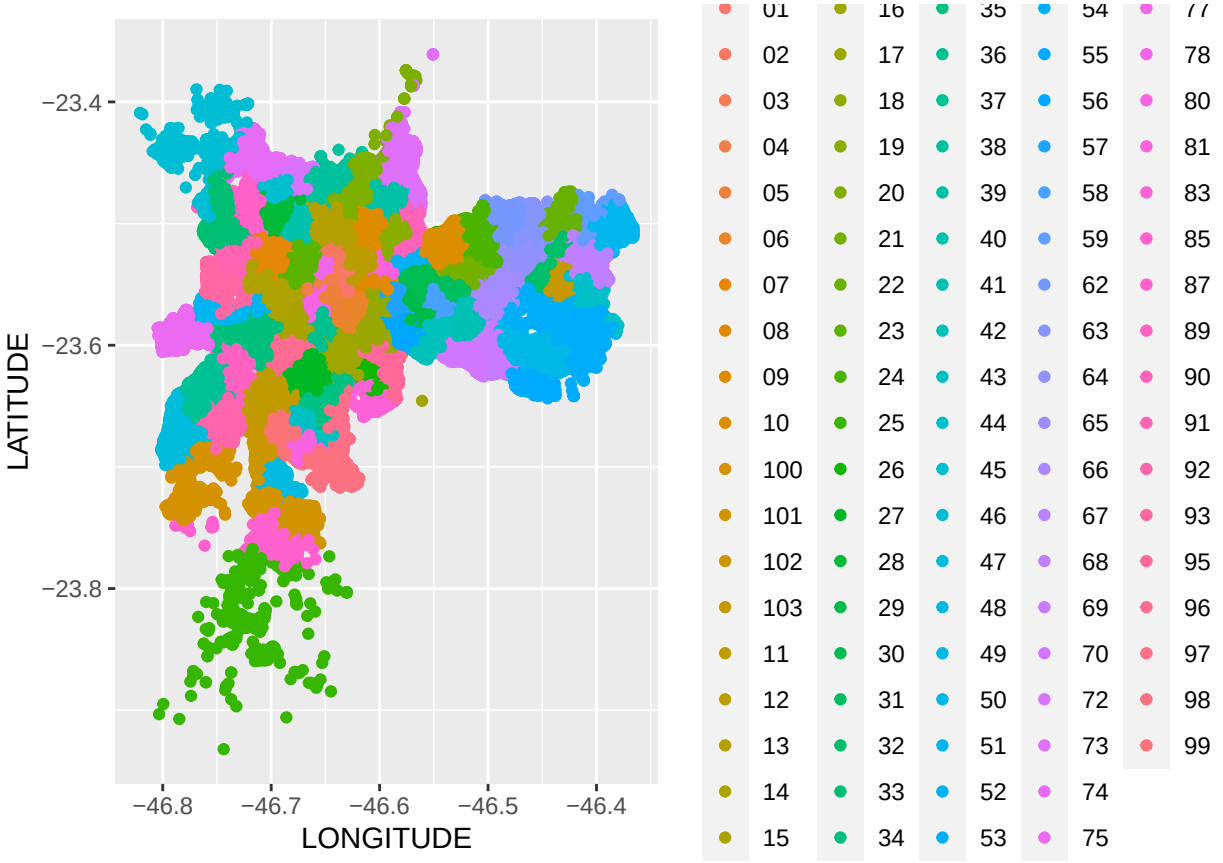


There is a concentration of informed thefts from 10 AM TO 12 PM. And, we can see bellow if the hours of crime change across the months

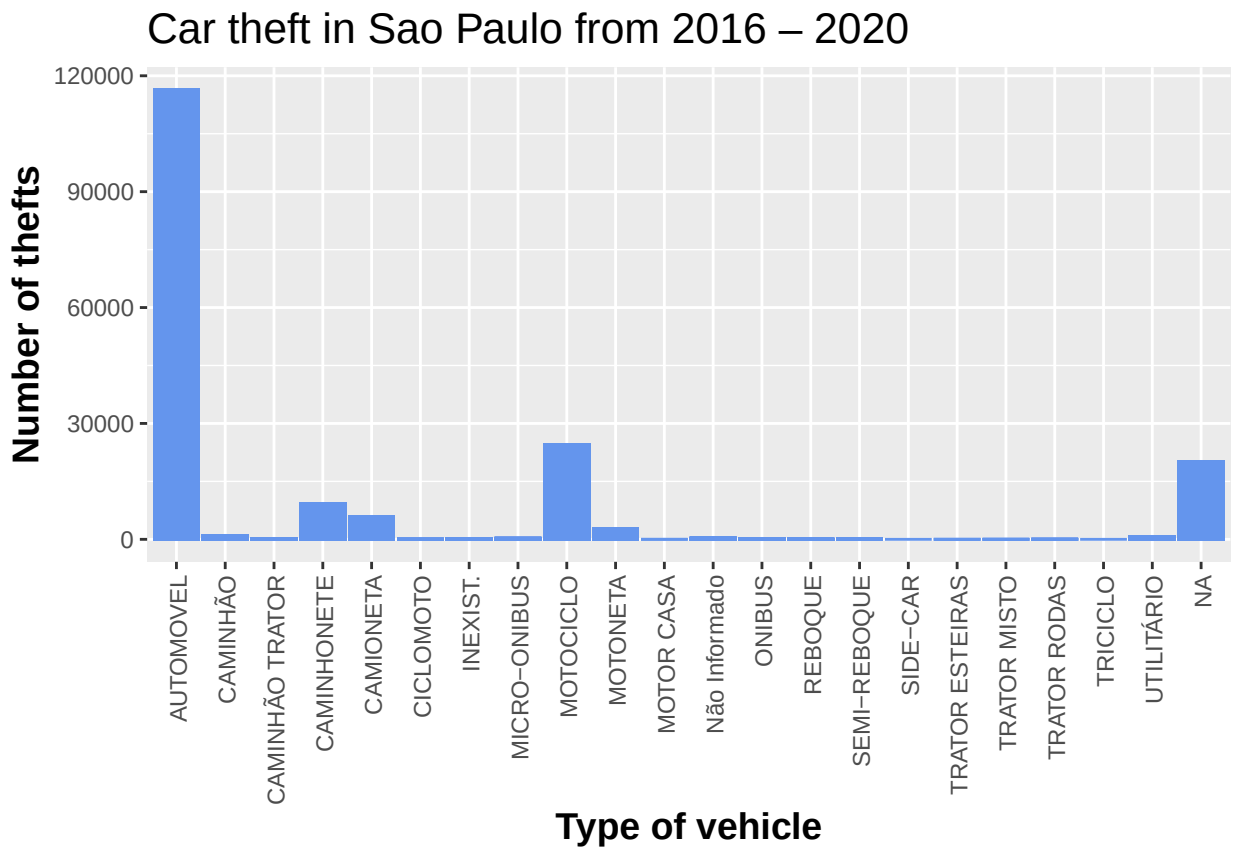


Spatial analysis

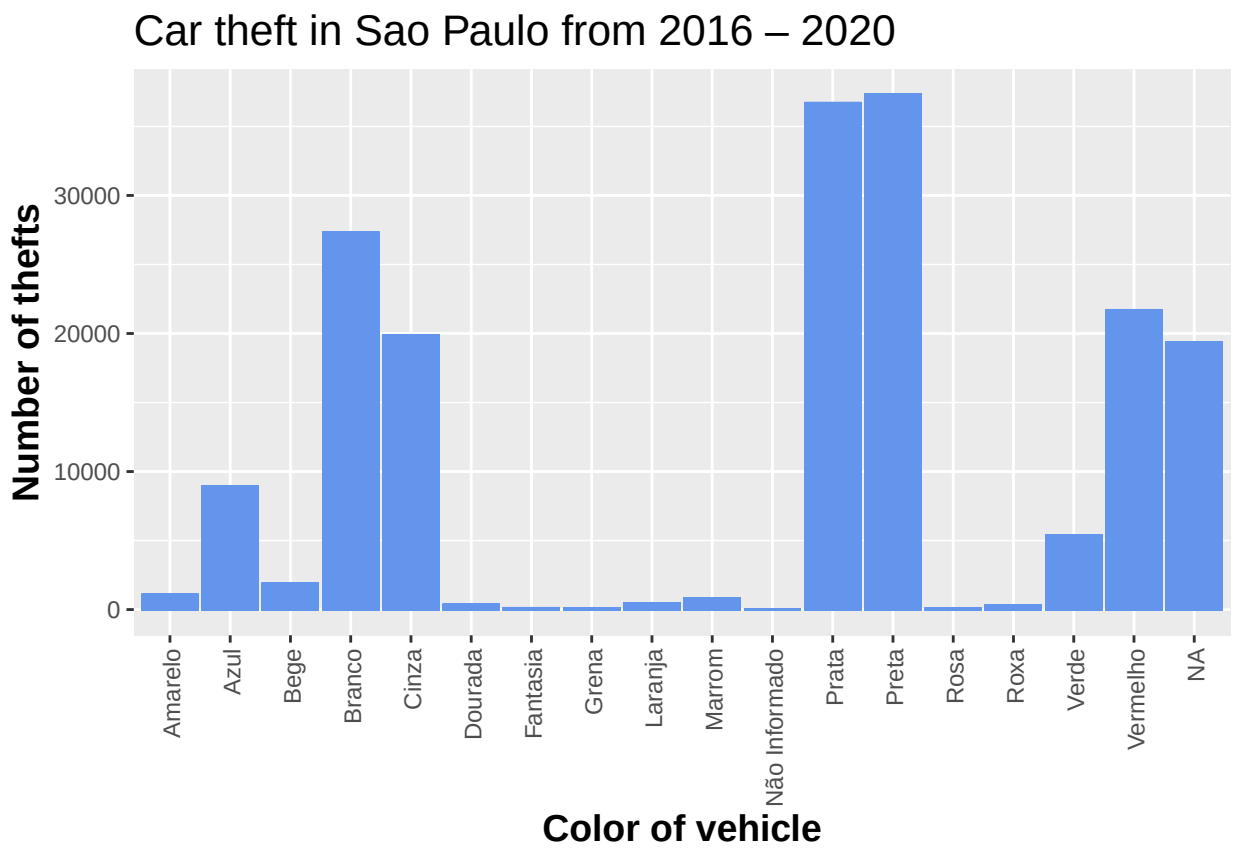
The most important information from the spatial analysis that we are going to keep is how is the crime across the delegacys, that we can see here (DEL_NUM is the number that indicate which DELEGACY is responsible for that area. So, 01 is the 01º Police Officer of the city (called 1º D.P. SÉ on the column DELEGACIA_CIRCUNSCRICAO) and so on) :



Other visualizations In the graphs bellow, we can see some interesting things, like the most thefts kind of cars:



In this, we can see the color that is most theft:



Modeling

Now, we are going to build a model that will try predict the crimes. Until now, we don't show the code, to because we were only showing the information. But, from now on, we will show, because it is important to show how we build our model.
Our model will use two variables to start: DEL_NUM that describe which DELEGACIA (delegacy in Portuguese) is responsible for that area (like you can see in the second graph on **spatial analysis**) and DiaCrime that indicate the day when the crime occurred.

```
DelNum <- sort(unique(BaseCrimeTotal$DEL_NUM))
DiaCrime <- sort(as.character(unique(BaseCrimeTotal$DateCrime)))
```

After that, we will create a table with all combinations possible for this two variables and then create another table with the data aggregate from BaseCrimeTotal (our first table of data) for this two variables plus the count of occurs of crimes in this combination (called Qtd):

```
##table with all combinations of del_num and dates
##Tabela com todas as combinações de delegacia e data
temp <- expand.grid(DelNum, DiaCrime)
names(temp) <- c("DELEGACIA_NUM", 'DATA_CRIME')
temp <- temp[order(temp$DELEGACIA_NUM),]

##table with the aggregated data
##tabela com os dados espaço temporais quantificados
```

```
modelo <- BaseCrimeTotal %>% dplyr::group_by(DEL_NUM,as.character(DateCrime)) %>% dplyr::summarise(Qtd=n
names(modelo) <- c("DELEGACIA_NUM", 'DATA_CRIME','Qtd')
```

Then, we merge the two tables. How not always there is crime in all PO all day, we need to put 0 when the value in Qtd is null:

```
##Merge data
##Combinação
modelo <- merge(temp, modelo, by= c('DELEGACIA_NUM', 'DATA_CRIME'), all.x = TRUE)
modelo$Qtd <- ifelse(is.na(modelo$Qtd),0,modelo$Qtd)
```

How we saw, the day of the week and the month are important to build a model, so lets add this to our model:

```
modelo$DIA_SEMANA <- weekdays(as.Date(modelo$DATA_CRIME),abbreviate=TRUE)
modelo$MES <- months(as.Date(modelo$DATA_CRIME),abbreviate=TRUE)
```

An important indicator is a criminal activity in that area. Areas with high criminal activity tend to have high criminal activity in the future. So, we will get the average of theft in the area, in a pre-determined time. For this pre-determined time, we will create a function, like that:

```
#Function that takes a vector started in zero followed by ones
#Função que cria um vetor iniciado com zero seguindo de uns, que usaremos para determinar quantos dias
pastDays <- function(x) {
  c(0, rep(1, x))
}
```

And apply it in the ave() function, for one day, seven days and thirty days:

```
#Crimes médio por data por intervalo
modelo$CrimePast1 <- ave(modelo$Qtd, modelo$DELEGACIA_NUM, FUN= function(x) stats::filter(x, pastDays(1)
modelo$CrimePast7 <- ave(modelo$Qtd, modelo$DELEGACIA_NUM, FUN= function(x) stats::filter(x, pastDays(7)
modelo$CrimePast30 <- ave(modelo$Qtd, modelo$DELEGACIA_NUM, FUN= function(x) stats::filter(x, pastDays(3
```

Again, we get some null values in this columns, because the function apply values only the time space is completed (with seven day, e.g., we don't get value in the days one to six). So, we will replace this nulls with the average:

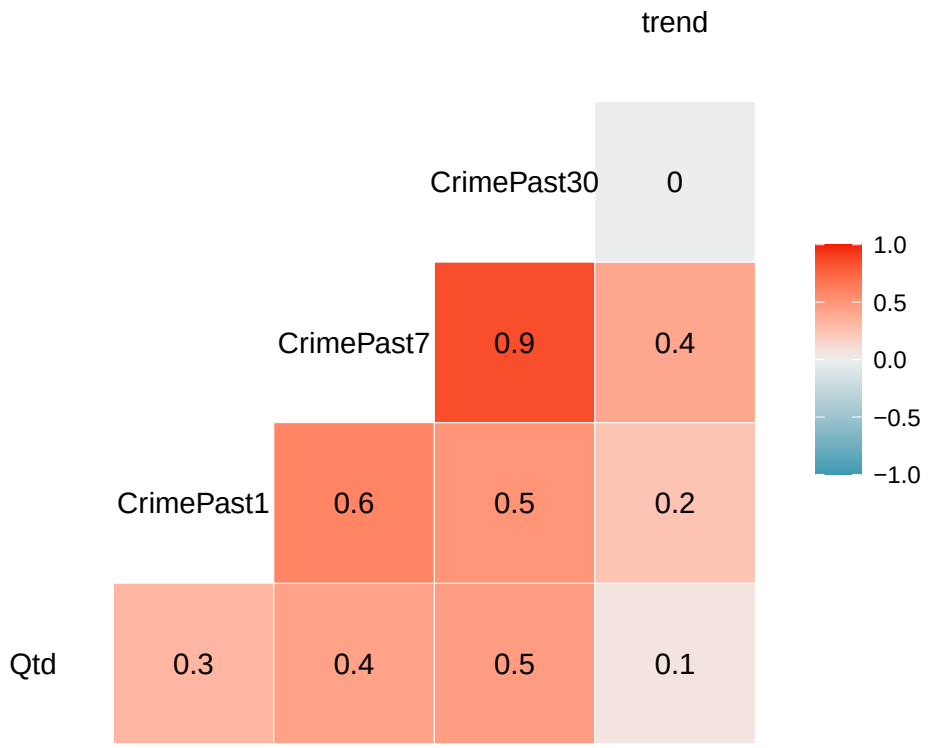
```
meanNA <- function(x){
  mean(x, na.rm = TRUE)
}

modelo$CrimePast1 <- ifelse(is.na(modelo$CrimePast1), meanNA(modelo$CrimePast1), modelo$CrimePast1)
modelo$CrimePast7 <- ifelse(is.na(modelo$CrimePast7), meanNA(modelo$CrimePast7), modelo$CrimePast7)
modelo$CrimePast30 <- ifelse(is.na(modelo$CrimePast30), meanNA(modelo$CrimePast30), modelo$CrimePast30)
```

Other useful indicator could be the trend. If the curve seems to increase in some area, it can be applied in the model

```
modelo$trend <- ifelse(modelo$CrimePast30 == 0, 0,modelo$CrimePast7/modelo$CrimePast30)
```

Finally, we can investigate if there is a correlation in this information:



Just to test, we can see how this all information perform in a generalized linear model:

```
CrimeModelo <- glm(Qtd ~ CrimePast1 + CrimePast7 + CrimePast30 + trend + factor(DIA_SEMANA)+MES, data =
summary(CrimeModelo)
```

```
##
## Call:
## glm(formula = Qtd ~ CrimePast1 + CrimePast7 + CrimePast30 + trend +
##     factor(DIA_SEMANA) + MES, data = modelo)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -8.337  -0.844  -0.273   0.607  58.733
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.0658493  0.0201458  -3.269 0.001081 **
## CrimePast1             0.0755654  0.0028607  26.415  < 2e-16 ***
## CrimePast7             0.0294505  0.0017689  16.649  < 2e-16 ***
## CrimePast30            0.0205503  0.0004268  48.151  < 2e-16 ***
## trend                 -0.2053194  0.0557006  -3.686 0.000228 ***
## factor(DIA_SEMANA)qua  0.5155264  0.0134992  38.190  < 2e-16 ***
```

```
## factor(DIA_SEMANA)qui  0.3990999  0.0135246  29.509  < 2e-16 ***
## factor(DIA_SEMANA)sáb -0.1062832  0.0134143  -7.923 2.33e-15 ***
## factor(DIA_SEMANA)seg  0.0872149  0.0133937   6.512 7.46e-11 ***
## factor(DIA_SEMANA)sex  0.1285061  0.0134872   9.528  < 2e-16 ***
## factor(DIA_SEMANA)ter  0.4948985  0.0134017  36.928  < 2e-16 ***
## MESago                 0.0175527  0.0172809   1.016 0.309760
## MESdez                -0.2202629  0.0173139 -12.722  < 2e-16 ***
## MESfev                -0.0113644  0.0166731  -0.682 0.495494
## MESjan                 0.1805170  0.0163975  11.009  < 2e-16 ***
## MESjul                -0.0146545  0.0172792  -0.848 0.396382
## MESjun                 0.0411792  0.0174392   2.361 0.018212 *
## MESmai                 0.0047426  0.0172864   0.274 0.783814
## MESmar                 0.0225266  0.0163053   1.382 0.167113
## MESnov                 0.0240983  0.0174418   1.382 0.167085
## MESout                 0.0441373  0.0172877   2.553 0.010678 *
## MESset                 0.0813027  0.0174481   4.660 3.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.884445)
##
##     Null deviance: 368649  on 147125  degrees of freedom
## Residual deviance: 277209  on 147104  degrees of freedom
## AIC: 510773
##
## Number of Fisher Scoring iterations: 2
```

And for a Poisson distribution:

```
CrimeModeloPoison <- glm(Qtd ~ CrimePast1 + CrimePast7 + CrimePast30 + trend + factor(DIA_SEMANA)+MES, 
summary(CrimeModeloPoison)
```

```
##
## Call:
## glm(formula = Qtd ~ CrimePast1 + CrimePast7 + CrimePast30 + trend +
##     factor(DIA_SEMANA) + MES, family = poisson(), data = modelo)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.4057  -1.2541  -0.3200   0.5091  19.2887
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -9.322e-01  1.440e-02 -64.747  < 2e-16 ***
## CrimePast1             2.687e-02  1.330e-03  20.206  < 2e-16 ***
## CrimePast7             5.718e-03  9.280e-04   6.162 7.18e-10 ***
## CrimePast30            1.607e-02  2.263e-04  71.005  < 2e-16 ***
## trend                  4.796e-01  3.974e-02  12.070  < 2e-16 ***
## factor(DIA_SEMANA)qua  4.189e-01  9.003e-03  46.529  < 2e-16 ***
## factor(DIA_SEMANA)qui  3.393e-01  9.141e-03  37.119  < 2e-16 ***
## factor(DIA_SEMANA)sáb -1.041e-01  1.008e-02 -10.331  < 2e-16 ***
## factor(DIA_SEMANA)seg  8.651e-02  9.646e-03   8.969  < 2e-16 ***
## factor(DIA_SEMANA)sex  1.316e-01  9.527e-03  13.813  < 2e-16 ***
```

```
## factor(DIA_SEMANA)ter  4.080e-01  8.980e-03  45.439  < 2e-16 ***
## MESago                 2.081e-05  1.147e-02   0.002 0.998552
## MESdez                -2.065e-01  1.241e-02 -16.637  < 2e-16 ***
## MESfev                -7.522e-03  1.096e-02  -0.686 0.492452
## MESjan                 1.697e-01  1.104e-02  15.371  < 2e-16 ***
## MESjul                -1.294e-02  1.156e-02  -1.119 0.263024
## MESjun                 4.534e-02  1.134e-02   3.999 6.36e-05 ***
## MESmai                 1.313e-02  1.123e-02   1.170 0.242190
## MESmar                 2.816e-02  1.076e-02   2.617 0.008866 **
## MESnov                 3.950e-02  1.130e-02   3.497 0.000471 ***
## MESout                 5.447e-02  1.113e-02   4.893 9.94e-07 ***
## MESset                 7.167e-02  1.138e-02   6.300 2.97e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 273918  on 147125  degrees of freedom
## Residual deviance: 212099  on 147104  degrees of freedom
## AIC: 425511
##
## Number of Fisher Scoring iterations: 5
```

Validation and RMSE

To validate our models, we need to split our data in two: test and training:

```
#Model Validation
#Validação Modelo

# Validation set will be 20% of modelo
set.seed(1, sample.kind="Rounding")
# if using R 3.5 or earlier, use `set.seed(1)` instead
test_index <- createDataPartition(y = modelo$DELEGACIA_NUM, times = 1, p = 0.2, list = FALSE)

train_set <- modelo[-test_index,]
test_set <- modelo[c(test_index),]
```

Now, we train our data in the

```
train_set <- modelo[-test_index,]
test_set <- modelo[c(test_index),]
```

And then, train in the two types (Linear Regression and Poisson)

```
  CrimeModeloLM <- glm(Qtd ~ CrimePast1 + CrimePast7 + CrimePast30 + trend + factor(DIA_SEMANA)+MES, da
  CrimeModeloPoISSON <- glm(Qtd ~ CrimePast1 + CrimePast7 + CrimePast30 + trend + factor(DIA_SEMANA)+MES
```

And after the train, apply it on the test_set

```
CrimeModeloPredLM <- predict(CrimeModeloLM, test_set, type= 'response')
CrimeModeloPredPoisson <- predict(CrimeModeloPoISSON, test_set, type= 'response')
```

Now, to determine how good are our models, we need a loss function. We will use the residual mean squared error (RMSE). The $RMSE$ is given by the formula: $ RMSE$ = $\sqrt{\frac{1}{N}\sum_{u,i}\left(\hat{y}_{u,i} - y_{u,i}\right)^2}$

```
  rmse_results_validation <- data_frame(method="Linear Regression",  RMSE = sqrt(mean((test_set$Qtd - C
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

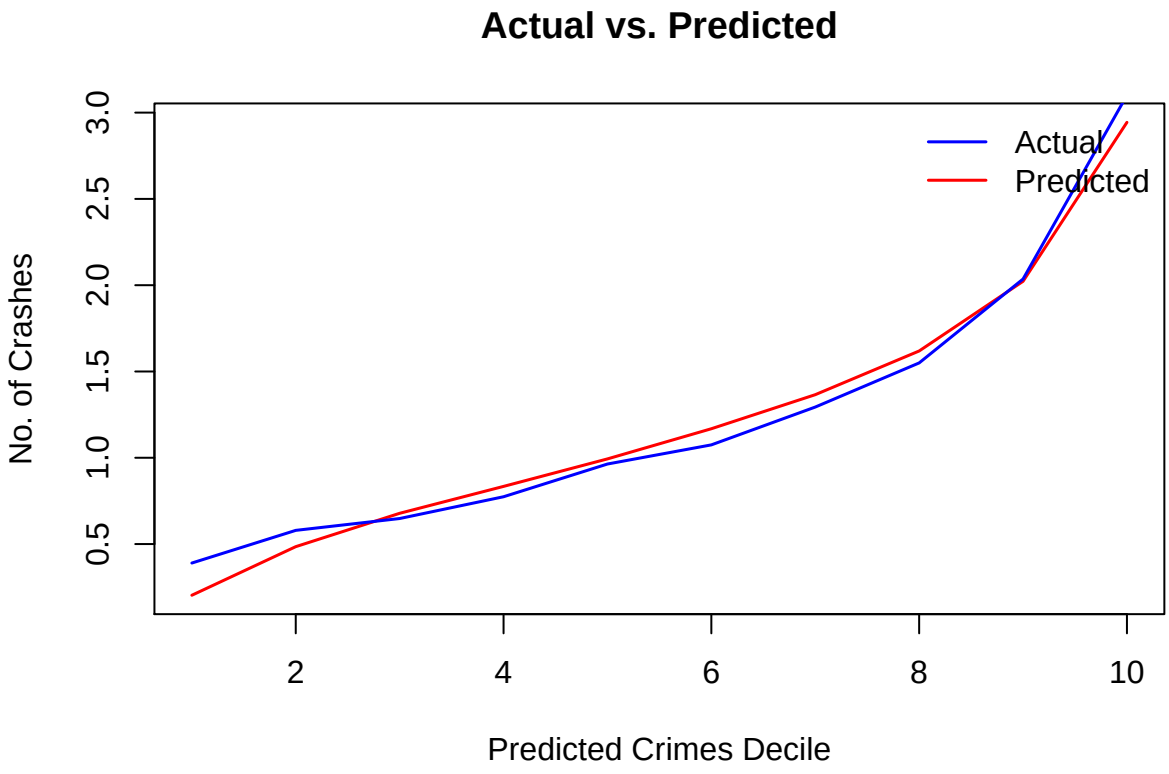
```
  rmse_results_validation %>% knitr::kable()
```

method	RMSE
Linear Regression	1.430847

```
  rmse_results_validation <- data_frame(method="Poisson",  RMSE = sqrt(mean((test_set$Qtd - CrimeModelo
  rmse_results_validation %>% knitr::kable()
```

method	RMSE
Poisson	1.448198

For that, we can see that the Linear Regression does better. In the graph bellow, we split our data in ten parts and compare our average forecast to actual



CONCLUSION

An approximately RMSE of one imply that we have one error of one crime error per day. OK, in a huge area like a P.O. we predicted well, but it is difficult to say where the crime will occur with it. In futures works, we can try to optimize this model with clusters inside the P.O. and with that makes the more accurated politicals of work

Inspirations

https://wetlands.io/maps/Crime-Analysis-Using-R.html#analyze_crime_hot_spots https://rpubs.com/djkpandian/CrimePrediction https://irgn452.files.wordpress.com/2016/03/3-s2-0-b9780124115118000141-main.pdf