Pmweb Report

Paulo Souza Junior

January 23, 2018

Introdução

Este documento apresenta um relatório analisando os dados abertos da cidade de Porto Alegre. O mesmo analisa dados relativos aos anos 2000 até 2016. Este relatório é gerado automaticamente através da linguagem R, pelo R Markdown e R Knit.

Inicialmente os arquivos a serem analisados são carregados, prearados e estruturados de forma equivalente para garantir uma analise global destes dados. Isto é necessário, visto que os dados disponibilizados em datapoa[1] nãoo possuem a mesma distribuição de colunas e formato de datas.

```
path = "~/poa_opendata/dataset/"

fNames <- list.files(path, pattern = "acidentes-2*")

acidentes <- lapply(paste(path, fNames, sep=""), function(fNames){
  data.frame(read.csv(fNames, header=TRUE, sep=";"))
})

aux <- data.frame()
year <- 2000

for(i in 1:17){
  acidentes[[i]]$year <- year
  if(i > 15){
    acidentes[[i]]$DATA_HORA <- format(as.POSIXlt(
      acidentes[[i]]$DATA_HORA, format="%Y-%m-%dT%H:%M"), "%Y%m%d %H:%M")
  }
  aux <- rbind(aux, select(acidentes[[i]], DATA_HORA, TEMPO, UPS, year))
  year = year + 1
}

aux$freq <- 1
```

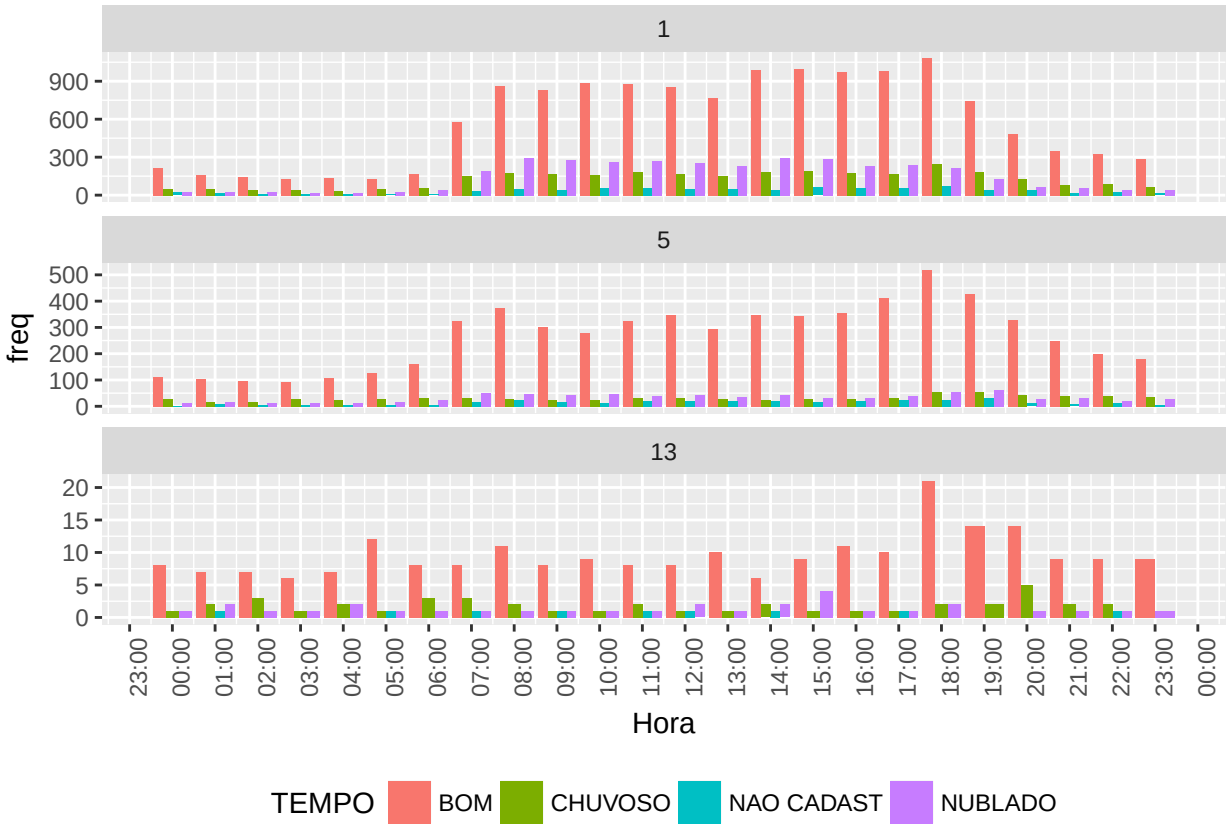
Condições atmosféricas, severidade e horário de acidentes

As relações das condições atmosféricas, severidade dos acindentes e horário das ocorrências são apresentadas nesta seção. O gráfico a seguir apresenta a frequencia de ocorrência a cada hora, em um período de 24 horas. As severidades das ocorrências são classificas em 1, 5 e 15, sendo (1 acidente com danos materiais, 5 acidente com ferido, 13 acidente com morte), estes são apresentados em um grid de gráficos. O clima é representado pela variável TEMPO, sendos eles BOM, CHUVOSO, NÃO CADASTRADO e NUBLADO, que estão representados por cores no gráfico.

```
plot <- aux %>% group_by(year, UPS, TEMPO, time =
                           floor_date(as.POSIXct(format(strptime(DATA_HORA,
                          "%Y%m%d %H:%M"), format='%H:%M'), format="%H:%M"), "1 hours")) %>%
```

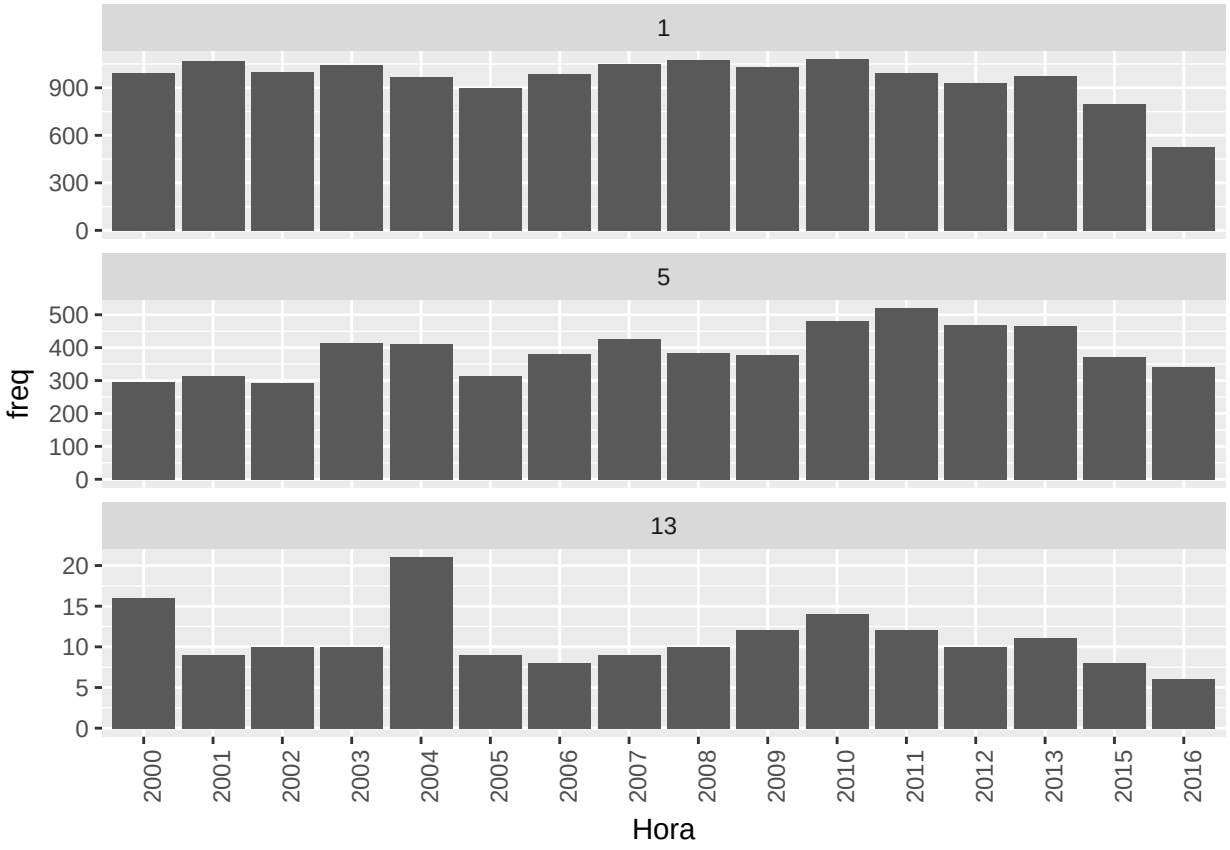
[1] http://datapoa.com.br

```
  summarise(freq = n())

ggplot(plot[complete.cases(plot),], aes(x = as.POSIXct(time, format="%H:%M"),
                y = freq, fill= TEMPO)) + geom_bar(stat = 'identity', position = 'dodge')  +
  labs(x="Hora") + facet_wrap(~UPS, nrow=3, scales="free_y") +
  scale_x_datetime(date_breaks = "1 hour", date_labels = "%H:%M") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")
```



Em relação ao passar dos anos e a severidade dos acidentes. O gráfico abaixo apresenta cada barra representando o número de ocorrência para cada severidade. É possível indicar que o número de acidentes aumento consideravelmente para severidades igual a 1 e 5. E teve uma breve redução em 2005, que logo foi aumentando o número de ocorrências, ao passar dos anos. O ano de 2010 apresentou um alto indice de ocorrências, principalmente para severidades 1 e 13, onde ao passar dos anos, até 2016, teve uma alta redução de ocorrências. É importante destacar o alto indice de acidentes de severidade 13 no ano de 2014, que não é visivel o aumento em casos de severidade 1, porém em casos de severidade 5 há um leve aumento. Isto pode ter sido causado por alguma variavel além, que pode ou não ser encontrada neste dataset (mera especulação).

```
ggplot(plot[complete.cases(plot),], aes(x = as.factor(year),
                y = freq, )) + geom_bar(stat = 'identity', position = 'dodge')  +
  labs(x="Hora") + facet_wrap(~UPS, nrow=3, scales="free_y") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")
```



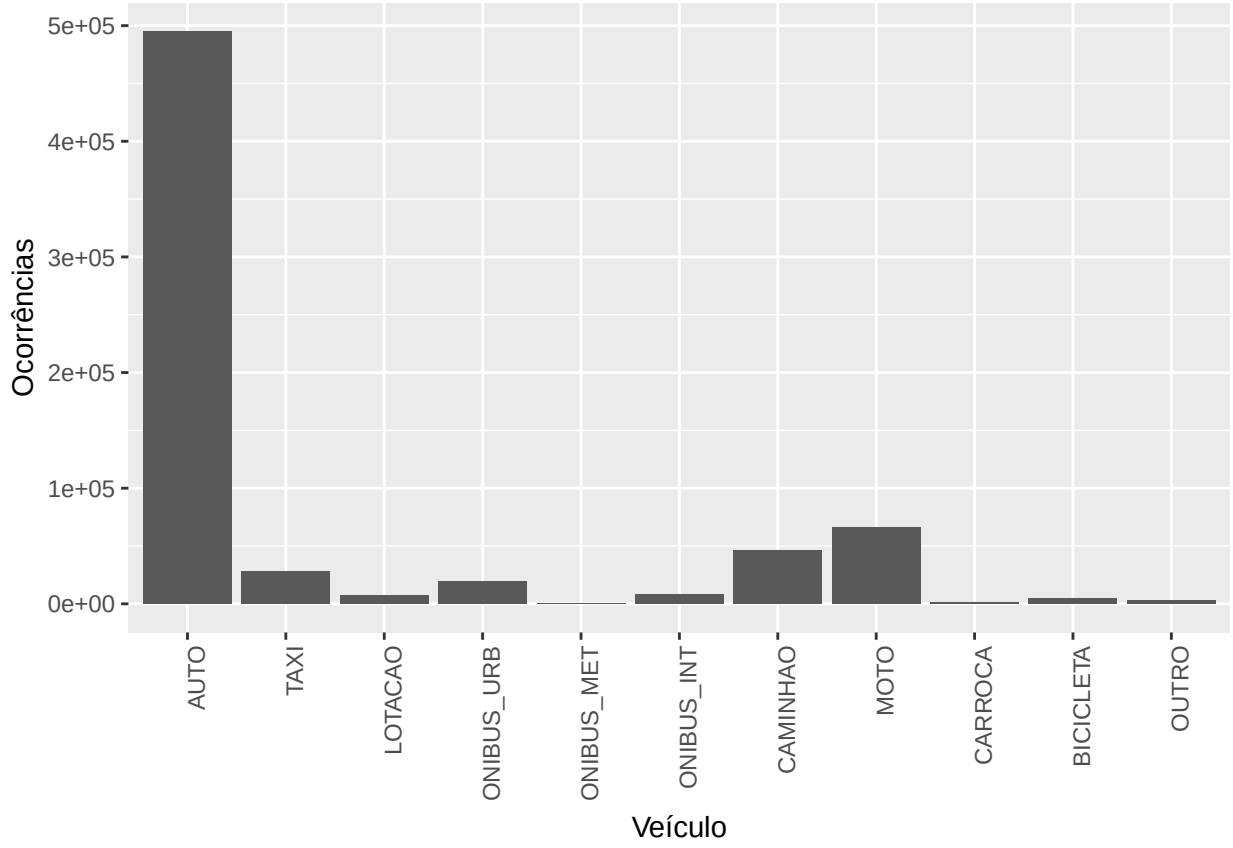
Danos causados em acidentes

Nesta seção é apresentado os veículos que estão envolvidos em acidentes. São apontandos os veículos que mais causaram acidentes e também os que mais causaram baixas.

No gráfico abaixo são apresentados a quantidade de acidentes envolvendo cada veículo. É clara a predominancia de acidentes envolvendo automóveis.

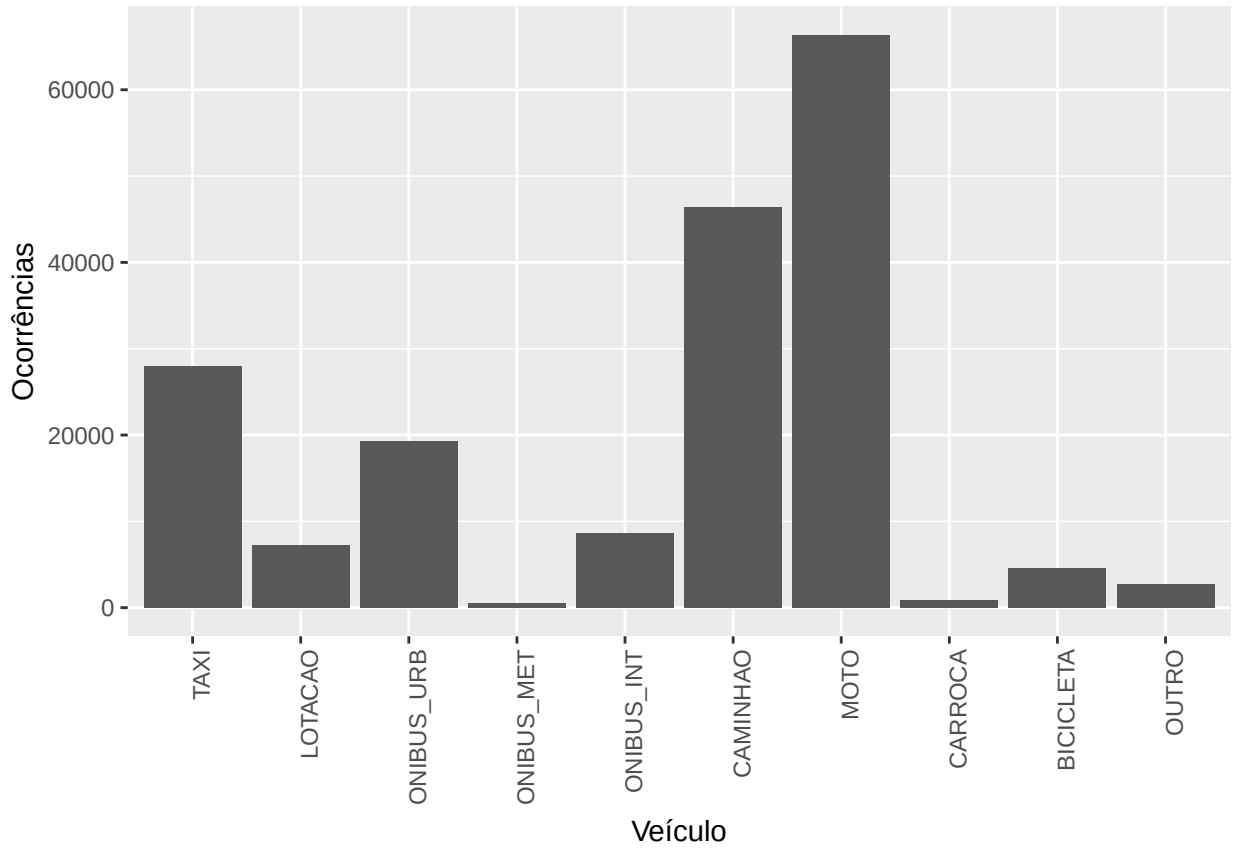
```
perdas <- data.frame()

for(i in 1:17){
  acidentes[[i]]$year <- year
  if(i < 15){
    acidentes[[i]]$ONIBUS_MET <- 0
    acidentes[[i]]$FERIDOS_GR <- 0
  }
  perdas <- rbind(perdas, select(acidentes[[i]], ID, AUTO, TAXI, LOTACAO,
                                 ONIBUS_URB, ONIBUS_MET, ONIBUS_INT, CAMINHAO,
                                 MOTO, CARROCA, BICICLETA, OUTRO, FERIDOS, FATAIS))
  year = year + 1
}

perdas <- as.data.frame(lapply(perdas, function(x) as.numeric(as.character(x))))
materiais <- perdas %>% summarize_all(funs(sum(., na.rm=TRUE))) %>% melt()

## No id variables; using all as measure variables
```

```
ggplot(materiais[2:12,], aes(x = variable,
                y = value)) + geom_bar(stat = 'identity', position = 'dodge')  +
  labs(x="Veículo", y = "Ocorrências") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")
```



O gráfico abaixo apresenta os outros veículos, além de automóveis. Visto que sua incidência é elevada o mesmo foi removido do gráfico para uma melhor visualização dos veículos. Logo atrás dos automóveis, os veículos que mais participaram de acidentes são: Motos, caminhões, taxis e onibus urbanos. Ônibus metropolitanos não faziam parte dos dados até 2014, a partir deste ano a coluna onibus_met passou a existir. Além disso, o gráfico abaixo destaca outros tipos de veículos, tornando mais visivel a participação de veículos pesados

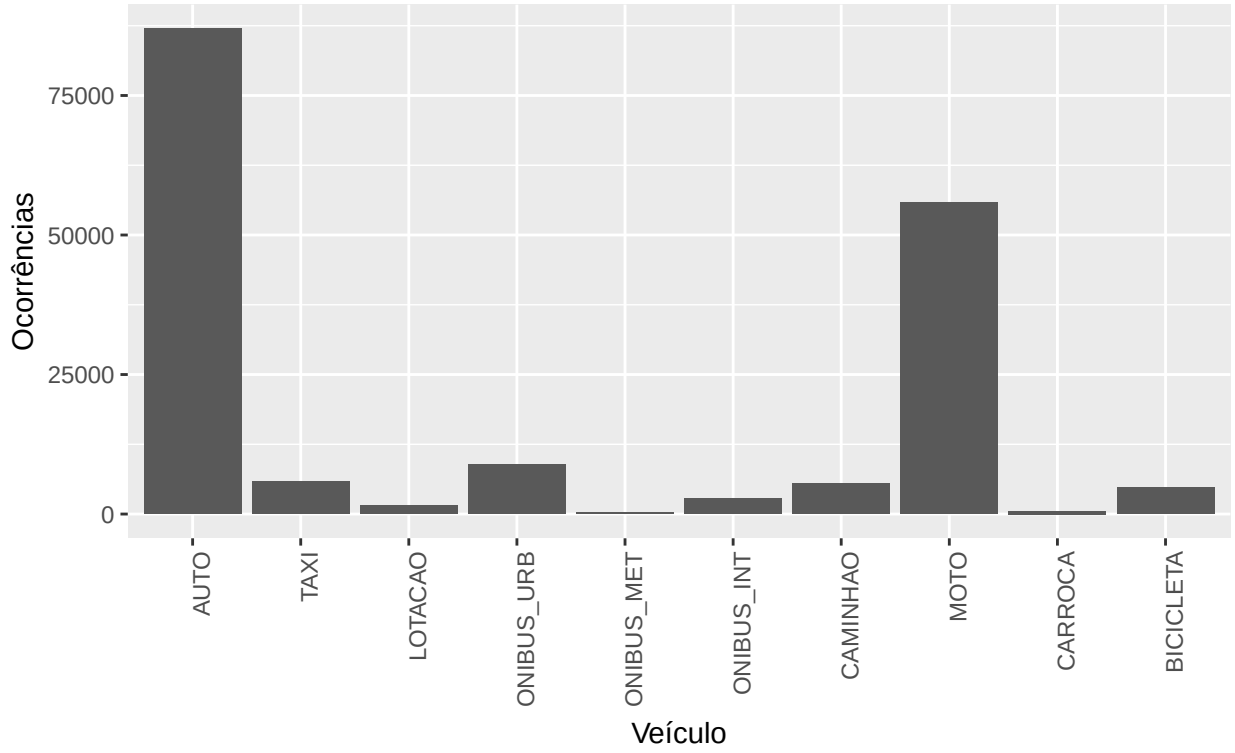
```
ggplot(materiais[3:12,], aes(x = variable,
                y = value)) + geom_bar(stat = 'identity', position = 'dodge')  +
  labs(x="Veículo", y = "Ocorrências") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")
```



Os próximos gráficos descatacam a participação dos veículos que causaram mais feridos e mortes ao longo dos anos, respecitivamente. Os acidentes que mais causaram ferimentos foram os de automoveis, logo após os acidentes de motocicletas, sendo possível ver claramente os seus altos índices de feridos.

```
feridos <- data.frame(sum(perdas[perdas$AUTO >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$TAXI >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$LOTACAO >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$ONIBUS_URB >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$ONIBUS_MET >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$ONIBUS_INT >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$CAMINHAO >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$MOTO >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$CARROCA >= 1, ]$FERIDOS, na.rm=TRUE),
                      sum(perdas[perdas$BICICLETA >= 1, ]$FERIDOS, na.rm=TRUE))

colnames(feridos) <- c("AUTO", "TAXI", "LOTACAO", "ONIBUS_URB", "ONIBUS_MET", "ONIBUS_INT", "CAMINHAO",
                       "MOTO", "CARROCA", "BICICLETA")

ggplot(melt(feridos), aes(x = variable,
                          y = value)) + geom_bar(stat = 'identity', position = 'dodge')  +
  labs(x="Veículo", y = "Ocorrências") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")
```

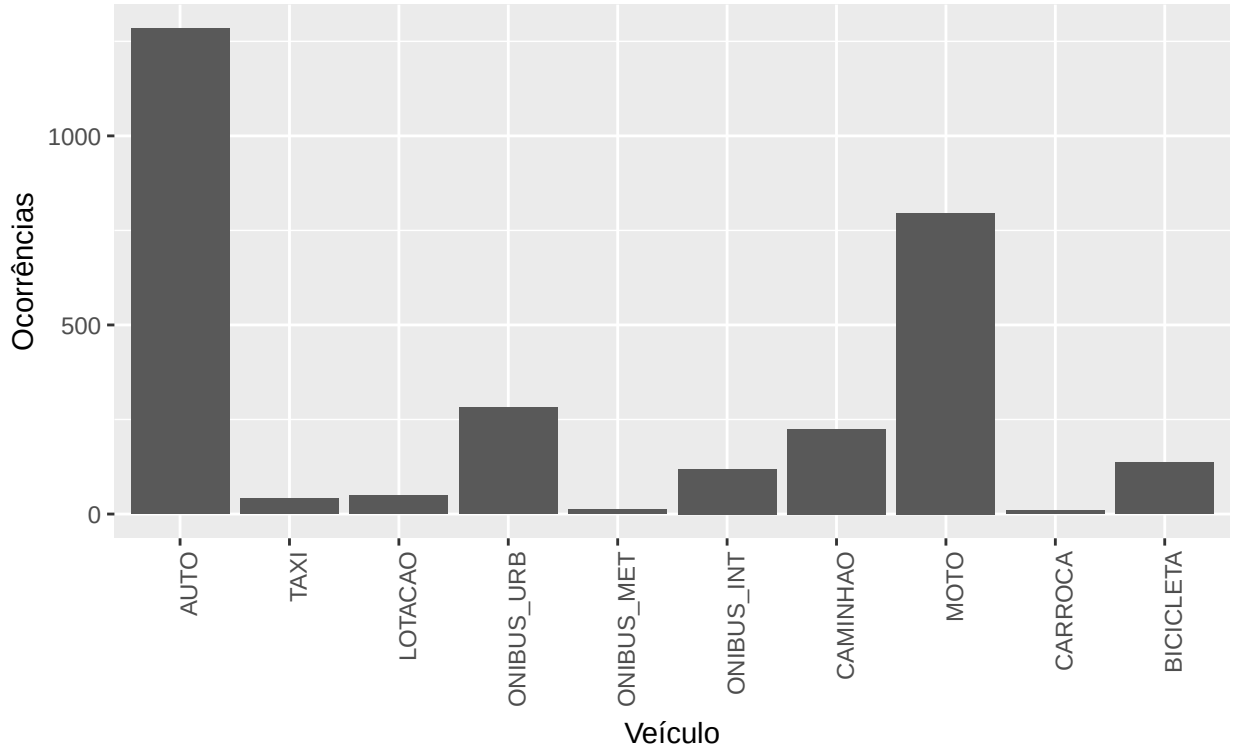
```
## No id variables; using all as measure variables
```



Já o índice de ocorrências com morte possui um comportamento parecido com os de feridos, apesar de houver um crescimento no indice de ocorrências em ônibus urbanos, caminhões e bicicletas. É necessário salientar, que esta analise não demonstra de modo proporcional o índice de acidentes, visto que é acidentes de automóveis são muito mais frequentes que acidentes de outros veículos. Neste caso, acidentes de motos apresentam maiores números de pessoas feridas e mortes.

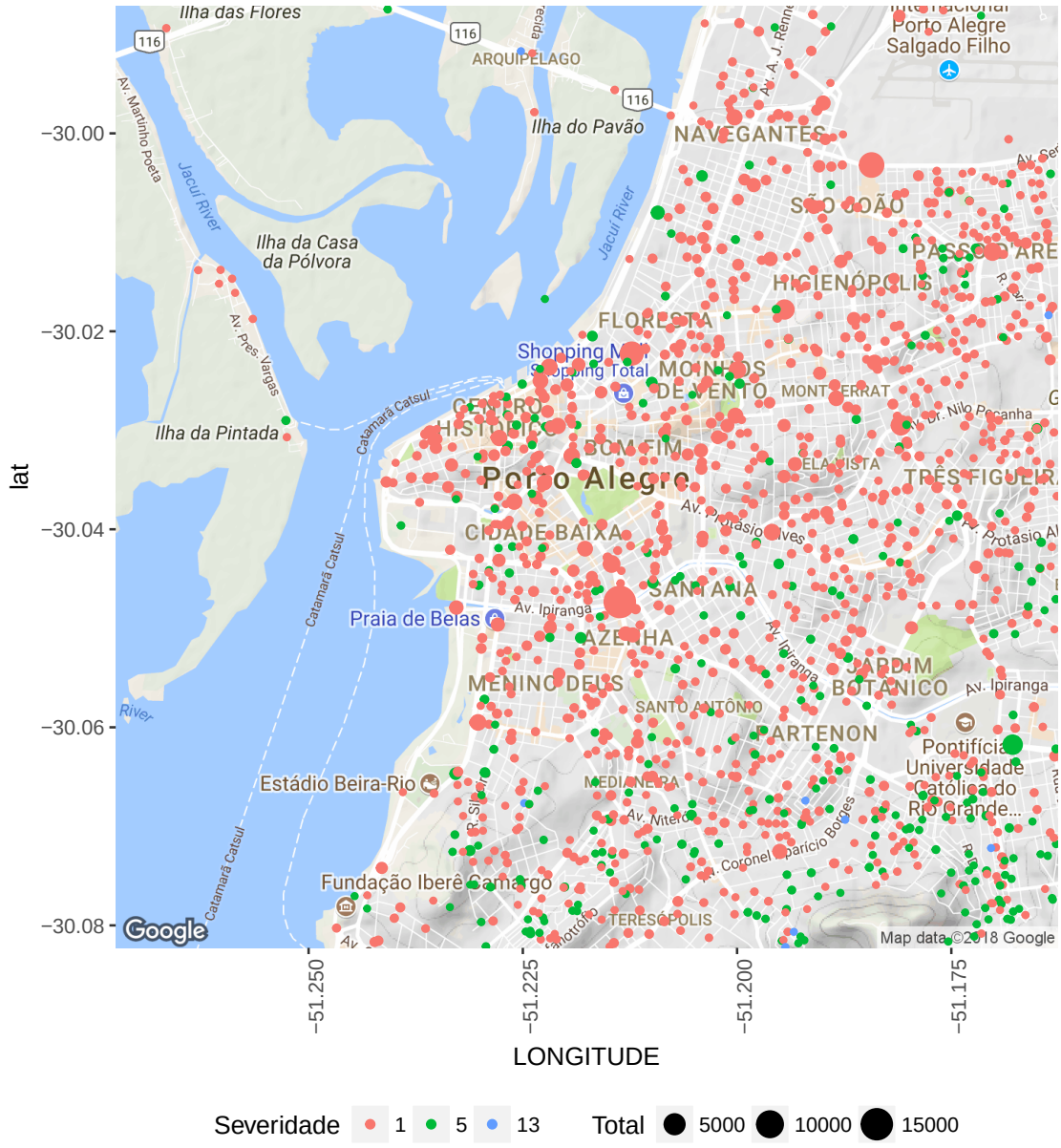
```
FATAIS <- data.frame(sum(perdas[perdas$AUTO >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$TAXI >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$LOTACAO >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$ONIBUS_URB >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$ONIBUS_MET >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$ONIBUS_INT >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$CAMINHAO >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$MOTO >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$CARROCA >= 1, ]$FATAIS, na.rm=TRUE),
                     sum(perdas[perdas$BICICLETA >= 1, ]$FATAIS, na.rm=TRUE))

colnames(FATAIS) <- c("AUTO", "TAXI", "LOTACAO", "ONIBUS_URB", "ONIBUS_MET", "ONIBUS_INT", "CAMINHAO",
                      "MOTO", "CARROCA", "BICICLETA")

ggplot(melt(FATAIS), aes(x = variable,
                         y = value)) + geom_bar(stat = 'identity', position = 'dodge')  +
  labs(x="Veículo", y = "Ocorrências") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1), legend.position = "bottom")
```

```
## No id variables; using all as measure variables
```



O gráfico abaixo apresenta a cidade de Porto Alegre, onde os pontos representão as ocorrências de acidentes, para as três possíveis severidades. São muito pontos, que precisam ser agrupadas para melhor entendimento. A ideia inicial seria também adicionar os pardais e semaforos no gráfico, para tornar visual se o índice de ocorrência próximas a essas vias diminui ou aumento. Isto não foi possível até então, como está sendo discutido na próxima seção.

```
##
## Attaching package: 'ggmap'

## The following object is masked from 'package:plotly':
##
##     wind

## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=PortoAlegre&zoom=13&size=640x640

## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=PortoAlegre&sensor=fa
```


lat
−30.00
−30.02
−30.04
−30.06
−30.08
−51.250
−51.225
−51.200
−51.175
LONGITUDE
Severidade
1
5
13
Total
5000
10000
15000
Porto Alegre
Ilha das Flores
Ilha do Pavão
Ilha da Casa da Pólvora
Ilha da Pintada
ARQUIPÉLAGO
NAVEGANTES
SÃO JOÃO
HIGIENÓPOLIS
FLORESTA
MOINHOS DE VENTO
MONTSERRAT
CENTRO HISTÓRICO
BOM FIM
ELA VISTA
TRÊS FIGUEIR
CIDADE BAIXA
SANTANA
AZENHA
MENINO DEUS
JARDIM BOTÂNICO
SANTO ANTÔNIO
PARTENON
MEDIANEIRA
Praia de Belas
Estádio Beira-Rio
Fundação Iberê Camargo
TERESÓPOLIS
Pontifícia Universidade Católica do Rio Grande...
Salgado Filho
Jacuí River
Catamará Catsul
Av. Ipiranga
Av. Nilo Peçanha
Av. Coronel Marcírio Borges
Av. Martinho Poeta
Av. Pres. Vargas
Google
Map data ©2018 Google

Problemas encontrados

Foram encontrados problemas relacionados a estrutura das coluna nos CSVs, onde algumas colunas não estavam presentes para os anos menores que 2014. Também, foram encontrados problemas em alguns datasets (mais especificamente na linha 11084 do ano de 2014), que resultou em valores incorretos para a leitura dos acidentes. Além disso, não foi possível conciliar a posição das semaforos e pardais, visto que suas colunas não apresentam o mesmo nome/título, porém isso é algo que pode ser trabalho com expressões regulares.

Metodologia

A linguagem R é um projeto open source, além de ser uma linguagem de programação, um ambiente para computação estatística, modelagem e visualização de dados. A linguagem R apresenta diversos pacotes que possibilitam a sua paralelização. Uma lista atual destes pacotes pode ser encontrada na página do CRAN Task View: High-Performance and Parallel Computing with R, que é a página da entidade que disponibiliza o R e seus pacotes oficiais. Há diversos pacotes que utilizam de funções que aproveitam e fazem uso de BigData. Alguns destes pacotes são geralmente implementados em C, Java e outros. O R é um ambiente estatistico para analise de dados, o mesmo auxilia na coleta, inferencia e tratamento de dados afim de determinar comportamentos e extrair informações. Alguns pacotes em R estão disponíveis para trabalhar com BigData, por exemplo o RHadoop.

Os dados foram analisados de forma rapida, há muito mais a ser explorado deste dataset. E cruzando dados além do que está disponibilizado aqui, por exemplo do clima real de cada dia do ano, é possível inferir em resultados mais precisos e valiosos.

Insights

É possível extrair ainda mais destes dados, uma analise seria os pontos de ocorrência que causaram maiores casos de acidentes com motos, podendo determinar o fator que maximizou o mesmo. Condiçoes de trânsito, asfalto, clima, afim de evitar esta categoria de ocorrência que possui altos indices de mortos e feridos. Também, uma ideia seria determinar os pontos que mais causaram acidentes envolvendo bicicletas buscando prover uma infra estrutura de trânsito que se preocupe com o ciclista e possibilite mais segurança.