07

Mãos na massa: Explorando os dados

Chegou a hora de você executar o que foi visto na aula! Para isso, execute os passos listados abaixo.

Será utilizado o **RStudio** como interface gráfica, então todos os comandos aqui mostrados deverão ser executados no seu console.

1) Ainda no script **Coleta_e_Explora.R**, caso você tenha fechado-o, não se preocupe, não será necessário realizar a coleta de dados novamente, basta ler o CSV gerado na última aula, atribuindo o *data frame* à variável `df_OVNI` :

```
df_OVNI <- read.csv("OVNIS.csv",stringsAsFactors = FALSE)
```

Explorando os dados com SQL

2) Para explorar dados, o **R** te dá muitos recursos. Por exemplo, para descobrir quantas observações (linhas) e quantas variáveis (colunas) um *data frame* possui, basta executar:

```
dim(df_OVNI)
```

O primeiro valor retornado é a quantidade de observações e o segundo valor retornado é a quantidade de variáveis.

3) Para investigar melhor o *data frame*, você utilizará SQL, então instale o *package* **sqldf**:

```
install.packages('sqldf')
```

4) E para que os *package* esteja na memória, você pode executar o comando `require` :

```
require(sqldf)
```

5) A ideia agora é criar *data frames* mais específicos, baseados no *data frame* que você já tem, que serão resultados de *queries*. Primeiramente, conte quantos relatos houveram por estado, ordenando em ordem decrescente pela coluna `Views` :

```
OVNI_EUA_por_Estado = sqldf("select State, count(*) Views
                            from df_OVNI
                            group by state
                            order by 2 desc")
```

Assim que o SQL é executado, um novo *data frame* é criado. Repare que nele, há um estado vazio e mais estados que 51, número de estados dos Estados Unidos.

6) Agora, conte quantos relatos há por cidade, desta vez eliminando o estado vazio e estados não-americanos. Considere também somente cidades com mais relatos:

```
OVNI_EUA_por_Cidade =
```

```
sqldf("select state as state, city as city,count(*) Views
  from df_OVNI
  where state <> ''
  and city not like '%Canada%'
  group by state, city


  having count(*) >= 10
  order by 3 desc")
```

7) Como Califórnia é o estado com mais relatos, filtre suas cidades com 10 ou mais relatos, incluindo também os tipos de OVNIs:

```
OVNI_CA = sqldf("select Shape, City, count(*) Views
                from df_OVNI
                where State = 'CA'
                group by Shape, City
                having count(*) > 10
                order by 3 desc")
```

Mostrando os dados em forma de gráfico

8) Agora, mostre os dados em forma de gráfico, primeiramente instalando o *package* **ggplot2**:

```
install.packages('ggplot2')
```

E colocando-o em memória:

```
library(ggplot2)
```

9) Você já sabe quais são os estados com relatos mais frequentes e os tipos de OVNIs mais populares, faça uma *query* com estados e tipos de OVNIs específicos:

```
OVNI_EUA_por_Tipo = sqldf("select State, Shape, count(*) Views
                from df_OVNI
                where state in ('CA', 'FL', 'WA', 'TX')
                and shape in ('Light', 'Circle', 'Fireball', 'Sphere')
                group by state, shape
                order by 3 desc")
```

10) E para gerar o gráfico, execute:

```
ggplot(OVNI_EUA_por_Tipo, aes(x = State, y = Views)) +
  geom_col(aes(fill = Shape))
```

Fazendo mapas

10) Para fazer mapas, você precisa de latitude e longitude. No *package* **zipcode** há todos os CEPs de cidades dos Estados Unidos, então instale-o:

```
install.packages('sqldf')
```

Coloque-o em memória:

```
library(zipcode)
```

E crie um *data frame* com os CEPs, latitudes e longitudes das cidades dos Estados Unidos:

```
data(zipcode)
```

Há cidades com mais de um CEP por cidade, logo haverá diferentes valores para latitude e longitude.

11) Com esse *data frame* e com a *query* com OVNIs por cidade, faça um *merge*, colocando como critério em comum o estado e cidade:

```
d <- merge(OVNI_EUA_por_Cidade, zipcode, by=c("state","city"))
```

12) Agora, instale o *package* para mapas, o **ggmap**, e coloque-o em memória:

```
install.packages("ggmap")
library(ggmap)
```

13) Crie o mapa do país inteiro:

```
us<-map_data('state')
```

14) E depois faça o `ggplot` no mapa que você acabou de criar:

```
ggplot(d,aes(longitude,latitude)) +
  geom_polygon(data=us,aes(x=long,y=lat,group=group),color='gray',fill=NA,alpha=.35)+
  geom_point(aes(color = Views),size=.15,alpha=.25) +
  xlim(-125,-65)+ylim(20,50)
```

Como resultado, o mapa será mostrada, e os relatos são representados nele. Uma cor mais forte representa uma incidência de relatos maior.

15) Como a Califórnia é o estado com maior incidência, crie um mapa somente dele:

```
ca <- map_data('state', 'california')
d = d[d$state == 'CA' ,]

ggplot(d,aes(longitude,latitude)) +
  geom_polygon(data=ca,aes(x=long,y=lat,group=group),color='gray',fill=NA,alpha=.35)+
```

```
geom_point(aes(color = Views),size=.15,alpha=.25) +
xlim(-125,-110)+ylim(30,45)
```