09

Mãos na massa: Analisando com gráficos

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

1) Baixe aqui (https://s3.amazonaws.com/caelum-online-public/799-data-science-preparacao-e-exploracao-dos-dados/03/Arquivo_Apoio_Aula_03_graficos.R) o arquivo **Arquivo_Apoio_Aula_03_graficos.R** e abra-o no **RStudio**.

2) Seguindo passo-a-passo o programa, importe o pacote **dplyr**:

```
library(dplyr)
```

Dispersão

3) Parta do *data frame* **df_clima_consolidado**, fazendo um filtro de linha, selecionando apenas as observações com os meses de verão dos EUA (junho a setembro, ou mês 6 ao mês 9). O resultado desse filtro servirá como entrada para o agrupamento por **cidade**. Então, por cidade, visualize a média das temperaturas. Em seguida, conte quantas observações têm, ordene por quantidade, em ordem decrescente, e exiba os 10 primeiros dados. Salve tudo isso no *data frame* **df_temp_vs_visualiz**:

```
df_temp_vs_visualiz = df_clima_consolidado %>%
  filter(mes %in% c("06", "07", "08", "09")) %>%
  group_by (cidade) %>%
  summarise(media_temperaturas =
              mean(as.numeric(history.observations.tempm)),
            quantas = n()) %>%
  arrange (desc(quantas)) %>%
  head (10)
```

4) Para tornar a criação do gráfico mais limpa, crie duas variáveis, uma para a quantidade de observações e outra para a média de temperaturas do *data frame* **df_temp_vs_visualiz**:

```
quantidades =     df_temp_vs_visualiz$quantas
temperaturas = df_temp_vs_visualiz$media_temperaturas
```

5) Com o `plot`, faça o gráfico, passando para ele os dados a serem trabalhados, o título, os títulos dos eixos X e Y, e o tamanho do ponto:

```
plot(quantidades, temperaturas,
     main="Relação Visualizações com Temperatura Média no Verão",
     xlab="Visualizações ",
     ylab="Temperatura Média no Verão",
     pch=19)
```

6) Com o gráfico criado, extraia as cidades do *data frame* **df_temp_vs_visualiz** para uma variável e passe esse valor para a função `text`:

```
cidades=df_temp_vs_visualiz$cidade
```

```
cidades=df_temp_vs_visualiz$cidade
text(quantidades, temperaturas, labels=cidades, cex= 0.7, pos=2)
```

Gráfico de linhas

7) Para saber como evoluiu a quantidade de visualizações ao longo dos anos, crie um novo *dataframe*:

```
df_Phoenix_vs_visualiz = df_clima_consolidado %>%
  filter(cidade == "Phoenix") %>%
  group_by (ano) %>%
  summarise(quantas = n()) %>%
  arrange (ano)
```

8) Instale e coloque em memória a biblioteca ggplot2:

```
install.packages("ggplot2")
library(ggplot2)
```

9) Com o `ggpplot` , faça o gráfico, passando para ele os dados a serem lidos, como você quer construir o gráfico, e o seu tipo, por exemplo:

```
ggplot(data=df_Phoenix_vs_visualiz, aes(x=ano, y=quantas, group=1)) + geom_line()+ geom_point()
```

Gráfico de barras verticais e horizontais

10) Compare a quantidade de visualizações do Arizona e na Califórnia. Primeiramente, crie o *dataframe*, selecione somente as observações dos estados citados, filtre por anos maiores ou iguais a 2012, agrupe por estado e ano, conte as visualizações e por fim ordene por ano:

```
df_Compara_Arizona_California = df_clima_consolidado %>%
  filter(estado %in% c("CA", "AZ")) %>%
  filter (as.numeric(ano) >= 2012) %>%
  group_by (estado, ano) %>%
  summarise(quantas = n()) %>%
  arrange (ano)
```

11) Do mesmo jeito que você criou o gráfico de linhas, com o `ggpplot` , faça o gráfico de barras, passando para ele os dados a serem lidos, como você quer construir o gráfico, e o seu tipo, por exemplo:

```
ggplot(df_Compara_Arizona_California,
       aes(estado, quantas, colour=ano))+
    geom_bar(aes(fill = ano),
    position = "dodge", stat="identity")
```

12) Você também pode criar um gráfico com barras horizontais:

```
ggplot(df_Compara_Arizona_California,
       aes(ano, quantas, colour=estado))+
  geom_bar(aes(fill = estado),
  position = "dodge", stat="identity")+
  coord_flip()
```

Gráfico de barras empilhadas

13) Lembrando do que foi feito no primeiro curso, instale e coloque em memória a biblioteca **sqldf**:

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

14) Selecione os OVNIs dos Estados Unidos por tipo, dessa vez utilizando SQL:

```
OVNI_EUA_por_Tipo =
    sqldf("select estado, formato, count(*) Quantas
          from df_clima_consolidado
          where estado in ('CA', 'FL', 'WA', 'TX')
          and formato in ('Light', 'Circle', 'Fireball', 'Sphere')
          group by estado, formato
          order by 3 desc")
```

15) E mais uma vez com o `ggpplot` , faça o gráfico de barras, só que dessa vez com o `geom_col` , por exemplo:

```
ggplot(OVNI_EUA_por_Tipo, aes(x = estado, y = Quantas)) +
  geom_col(aes(fill = formato))
```