Lab 2 Parte 1

```
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
library(ggplot2)
library(GGally)
```

```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

Perguntas criadas:

1. Quanto menor o tempo que o usuário passou no formulário, mais aleatório ele acha que foi? E qual a região do teclado mais utilizada por essa pessoa?
2. Qual a distância média dos caracteres da variável "agora uma digitação aleatória usando todo o teclado"?
3. Qual a região do teclado mais utilizado?

```
data <- read.csv("pessoas-aleatorias-digitando-coisas-aleatorias-report.csv")
```

Vamos selecionar a pergunta 1ª para responder.

Primeiro vamos ver quais as ocorrências de cada letra do alfabeto:

```
data.letters.occurence <- data
data.letters.occurence <- data.letters.occurence %>%
  mutate(all.letters = paste0(agora.uma.digitação.aleatória.usando.todo.o.teclado, e.uma.última..por.fa

data.letters.occurence$a <- str_count(data.letters.occurence$all.letters, "a")
data.letters.occurence$b <- str_count(data.letters.occurence$all.letters, "b")
data.letters.occurence$c <- str_count(data.letters.occurence$all.letters, "c")
data.letters.occurence$d <- str_count(data.letters.occurence$all.letters, "d")
data.letters.occurence$e <- str_count(data.letters.occurence$all.letters, "e")
data.letters.occurence$f <- str_count(data.letters.occurence$all.letters, "f")
data.letters.occurence$g <- str_count(data.letters.occurence$all.letters, "g")
data.letters.occurence$h <- str_count(data.letters.occurence$all.letters, "h")
data.letters.occurence$i <- str_count(data.letters.occurence$all.letters, "i")
data.letters.occurence$j <- str_count(data.letters.occurence$all.letters, "j")
data.letters.occurence$k <- str_count(data.letters.occurence$all.letters, "k")
data.letters.occurence$l <- str_count(data.letters.occurence$all.letters, "l")
data.letters.occurence$m <- str_count(data.letters.occurence$all.letters, "m")
data.letters.occurence$n <- str_count(data.letters.occurence$all.letters, "n")
data.letters.occurence$o <- str_count(data.letters.occurence$all.letters, "o")
data.letters.occurence$p <- str_count(data.letters.occurence$all.letters, "p")
data.letters.occurence$q <- str_count(data.letters.occurence$all.letters, "q")
data.letters.occurence$r <- str_count(data.letters.occurence$all.letters, "r")
data.letters.occurence$s <- str_count(data.letters.occurence$all.letters, "s")
data.letters.occurence$t <- str_count(data.letters.occurence$all.letters, "t")
data.letters.occurence$u <- str_count(data.letters.occurence$all.letters, "u")
data.letters.occurence$v <- str_count(data.letters.occurence$all.letters, "v")
data.letters.occurence$w <- str_count(data.letters.occurence$all.letters, "w")
data.letters.occurence$x <- str_count(data.letters.occurence$all.letters, "x")
data.letters.occurence$y <- str_count(data.letters.occurence$all.letters, "y")
data.letters.occurence$z <- str_count(data.letters.occurence$all.letters, "z")
```

Vamos calcular agora o tempo que a pessoa passou no formulário:

```
data.letters.occurence <- data.letters.occurence %>%
  mutate(
    diferenca.tempo = as.numeric(difftime(ymd_hms(data.letters.occurence$Submit.Date..UTC.), ymd_hms(da
    )

summary(data.letters.occurence$diferenca.tempo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    35.0    76.0   107.0   163.6   155.0  4503.0
```

A mediana de tempo de resposta foi de cerca de 2,5 minutos.

Agora vamos ver quanto cada região do teclado foi utilizada:

```
data.letters.occurence <- data.letters.occurence %>%
  mutate(
    regiao1 = q+a+z+x+s+w,
    regiao2 = e+d+c+v+f+r,
    regiao3 = t+g+b+n+h+y,
    regiao4 = u+j+m+k+i+o+l+p
  )
```

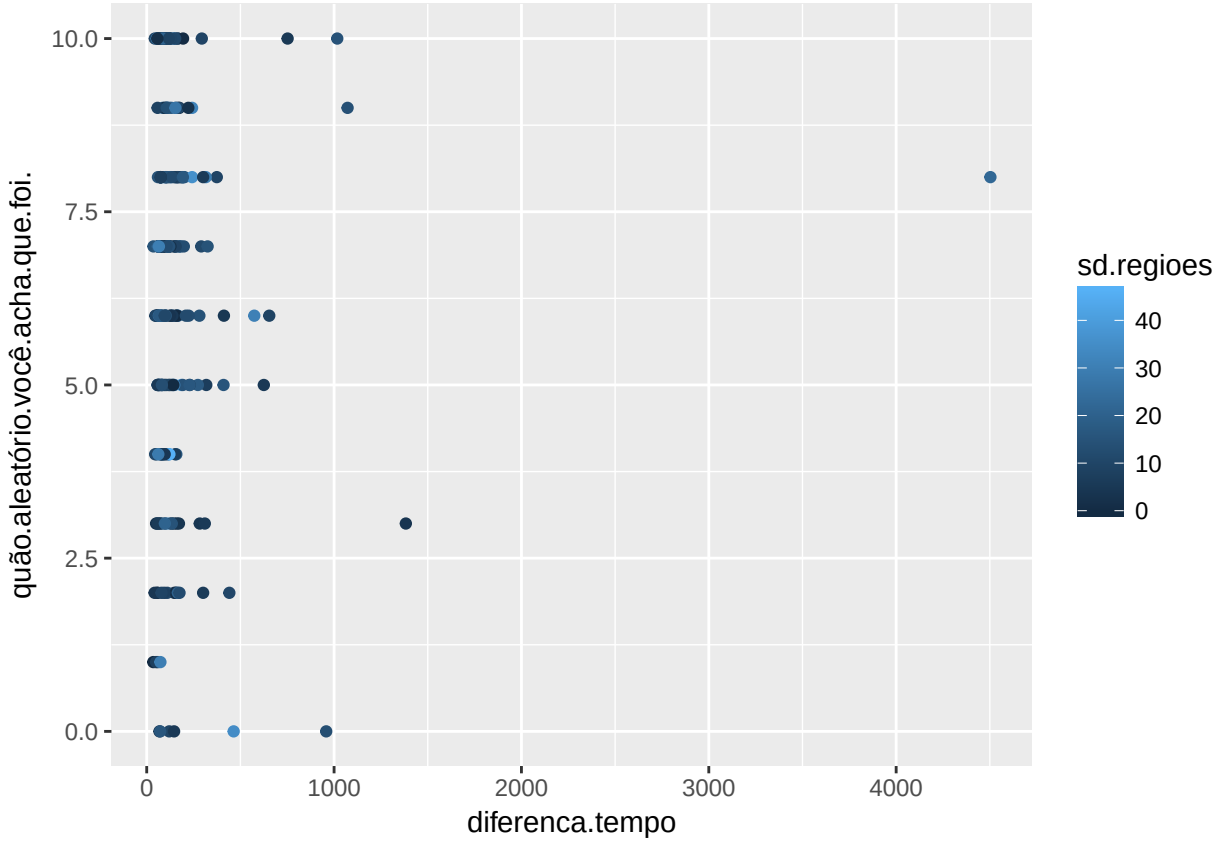
Agora vamos calcular o desvio padrão das regiões de cada usuário:

```
data.letters.occurence$sd.regioes <- apply(
    select(data.letters.occurence, regiao1, regiao2, regiao3, regiao4),
    1,
    sd,
    na.rm = TRUE
  )
```

Agora vamos responder a pergunta:

- Quanto menor o tempo que o usuário passou no formulário, mais aleatório ele acha que foi? E qual a região do teclado mais utilizada por essa pessoa?

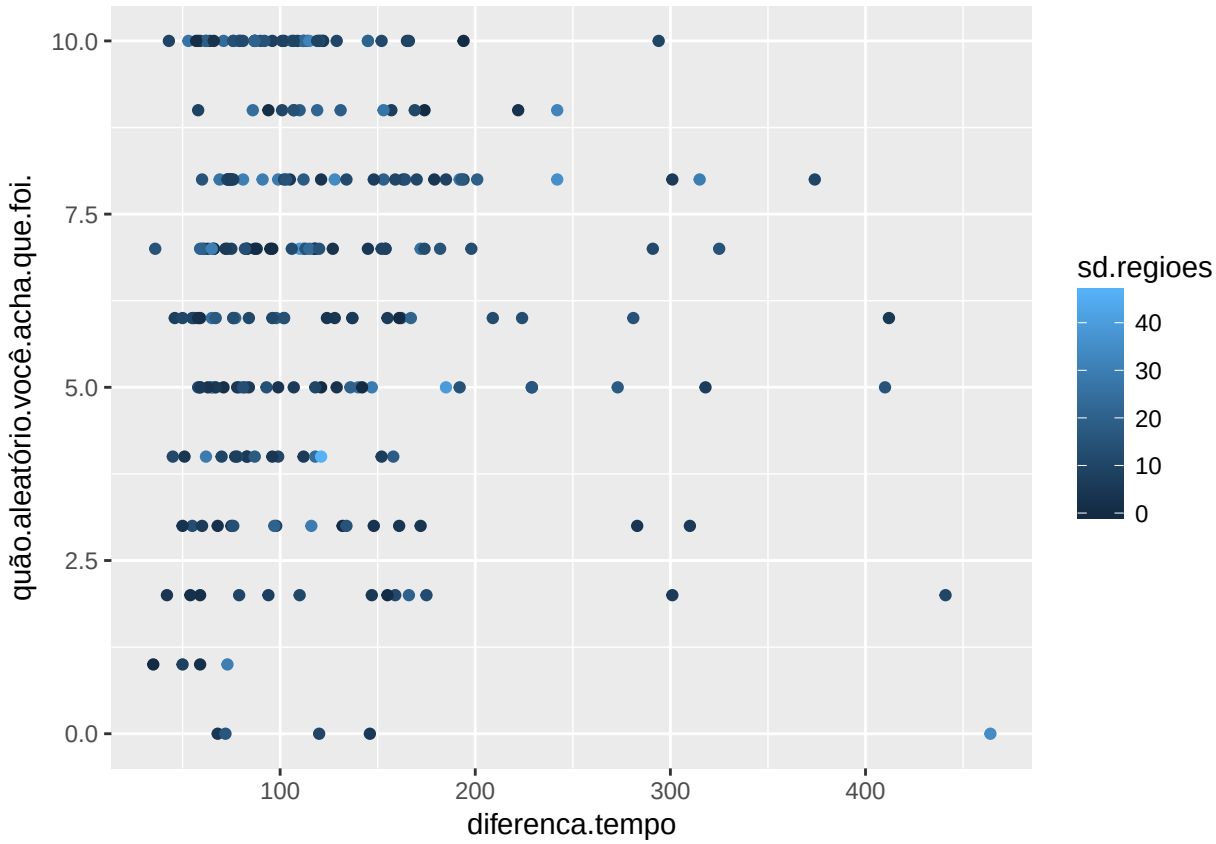
```
ggplot(data.letters.occurence, aes(x = diferenca.tempo, y = quão.aleatório.você.acha.que.foi.)) +
  geom_point(aes(colour = sd.regioes))
```



Para fins de visualização, vamos remover os que levaram mais de 500 segundos para responder o formulário:

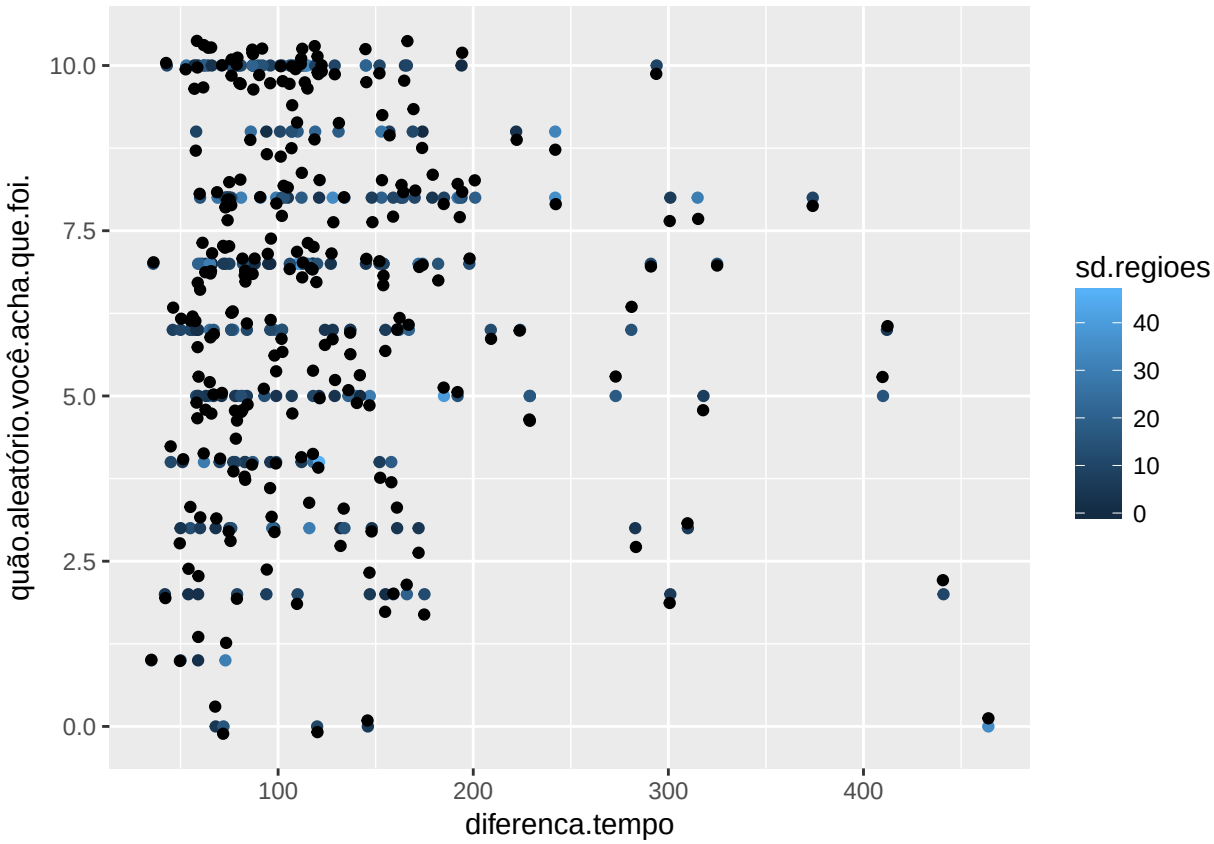
```
data.letters.occurence.filtered <- data.letters.occurence %>%
  filter(diferenca.tempo < 500)

ggplot(data.letters.occurence.filtered, aes(x = diferenca.tempo, y = quão.aleatório.você.acha.que.foi.))
  geom_point(aes(colour = sd.regioes))
```



Como ficou difícil ver as diferenças de cores, vamos fazer um espalhamento dos pontos, para ter uma maior visibilidade:

```
ggplot(data.letters.occurence.filtered, aes(x = diferenca.tempo, y = quão.aleatório.você.acha.que.foi.)
  geom_point(aes(colour = sd.regioes)) +
  geom_jitter()
```



Não parece existir uma tendência entre as variáveis "quão aleatório você acha que foi", "tempo que o usuário levou para preencher o formulário" e "desvio padrão das regiões do teclado".

Podemos verificar isso com as correlações:

```
cor(data.letters.occurence$quão.aleatório.você.acha.que.foi., data.letters.occurence$diferenca.tempo)
```

```
## [1] -2.537937e-05
```

```
cor(data.letters.occurence$quão.aleatório.você.acha.que.foi., data.letters.occurence$sd.regioes)
```

```
## [1] 0.1019627
```

```
cor(data.letters.occurence$diferenca.tempo, data.letters.occurence$sd.regioes)
```

```
## [1] 0.0696848
```

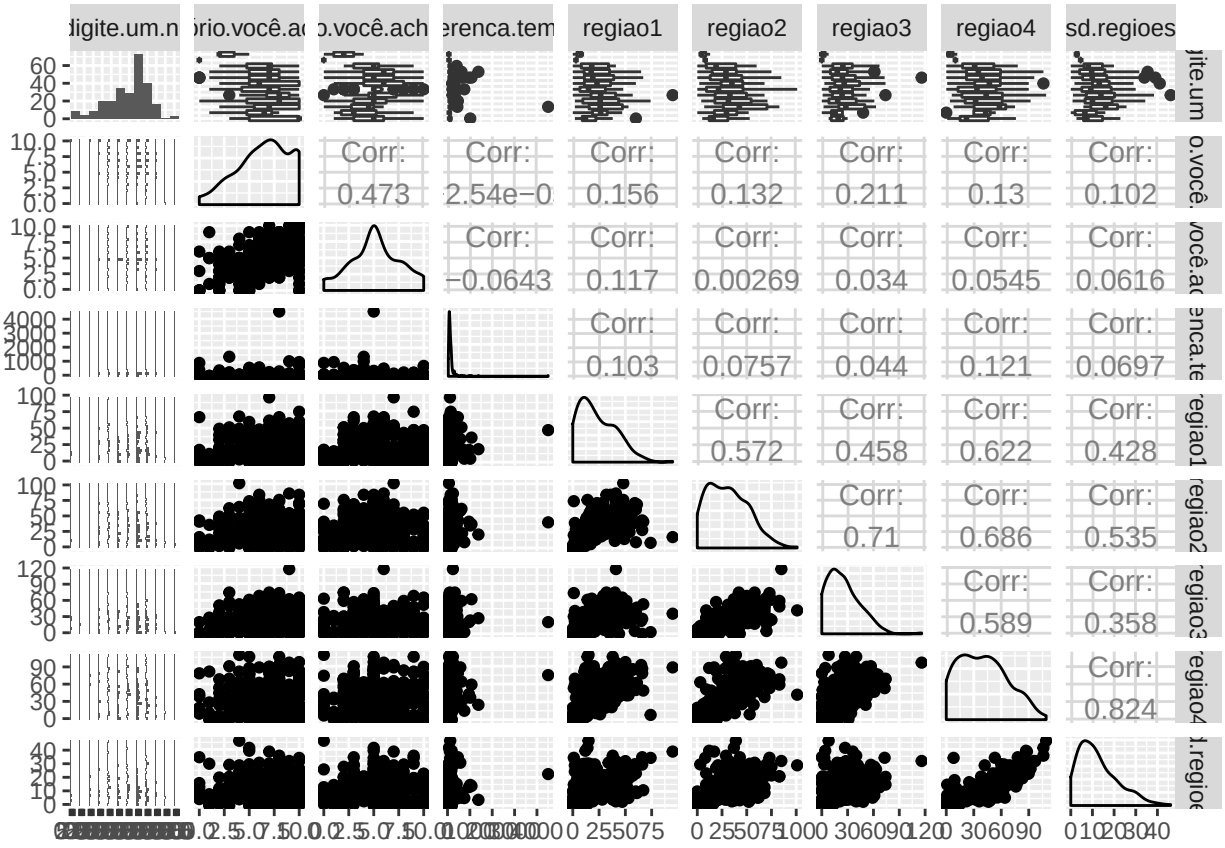
Onde nenhuma combinação tem uma correlação significativa.

Apenas a nível de curiosidade, vamos ver se existe alguma relação entre as variáveis:

```
ggpairs(select(data.letters.occurence, um.esquente..digite.um.número.de.0.a.9, quão.aleatório.você.acha
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

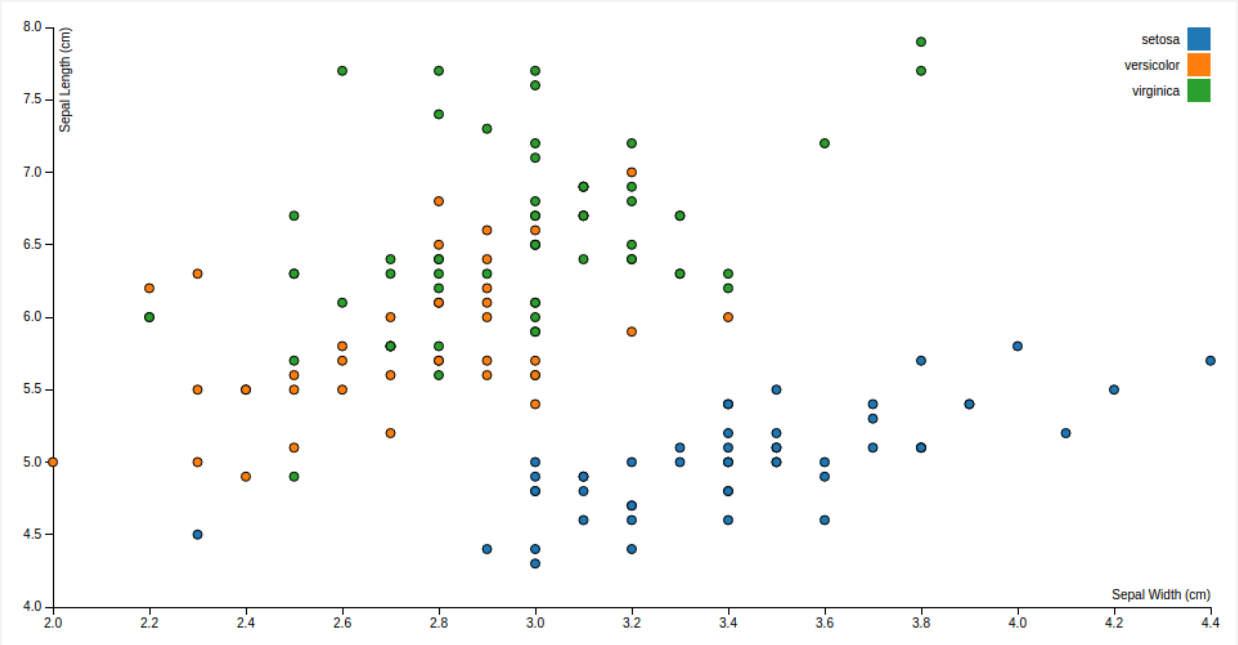
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Podemos verificar que existe uma alta correlação entre a região 4 do teclado e o desvio padrão das regiões, ou seja, quanto mais a pessoa digita nesta região, mais variação entre as regiões do teclado ela produziu, o que é um pouco surpreendente.

A visualização que será feita no d3, será parecida com esta:

Scatterplot



Que pode ser encontrada neste link