EDA do Data Set Wine Quality por Ericsson Graciolli para a Disciplina de TCC da Pós-graduação de Ciência de Dados e Big Data da PuC Minas. ==========

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
##  Median :0.07900   Median :14.00       Median : 38.00
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

Conhecendo os conjuntos de dados

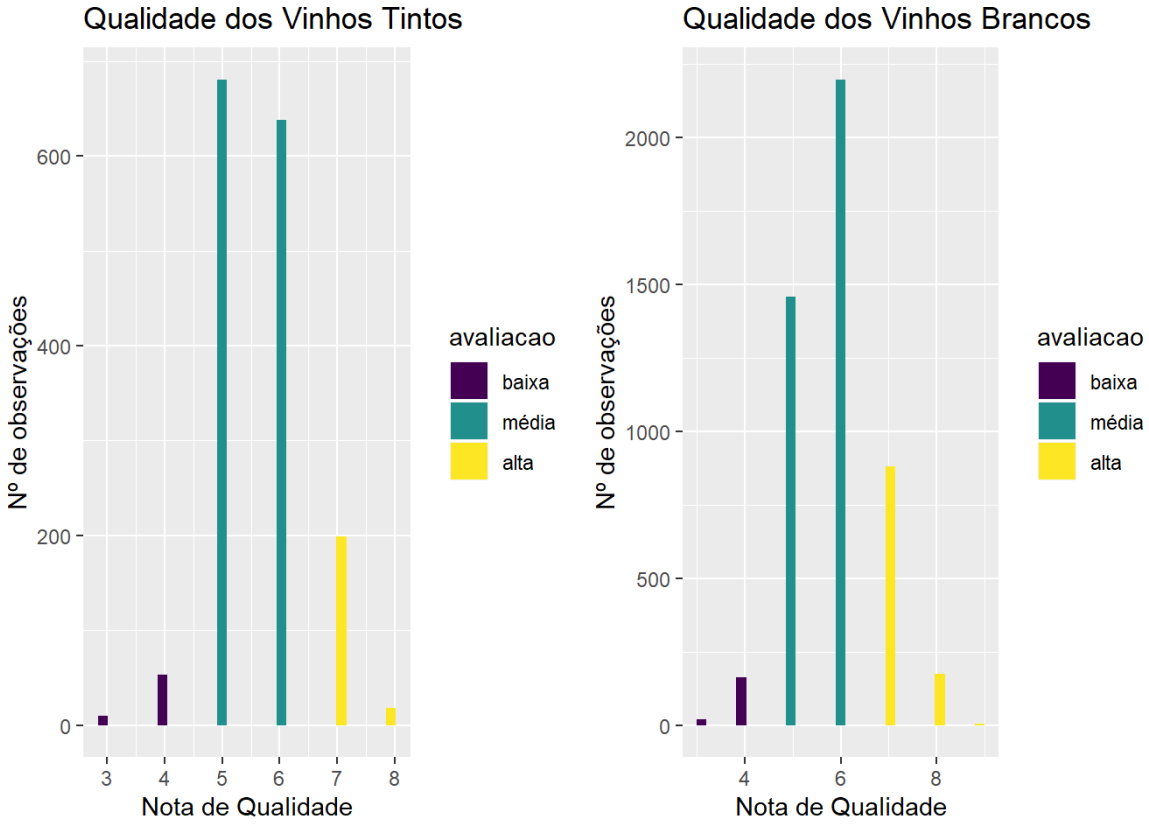
O conjunto de dados winequality-red possui 1599 observações e 12 variaveis relacionadas. Sendo que para facilitar as análises posteriores foi criada uma nova variável chamada de 'avaliacao'.

O conjunto de dados winequality-white possui 4898 observações e 12 variaveis relacionadas. Sendo que para facilitar as análises posteriores foi criada uma nova variável chamada de 'avaliacao'.

Nos histogramas abaixo distribuiremos as obserções de acordo do os valores atribuídos à variável quality.

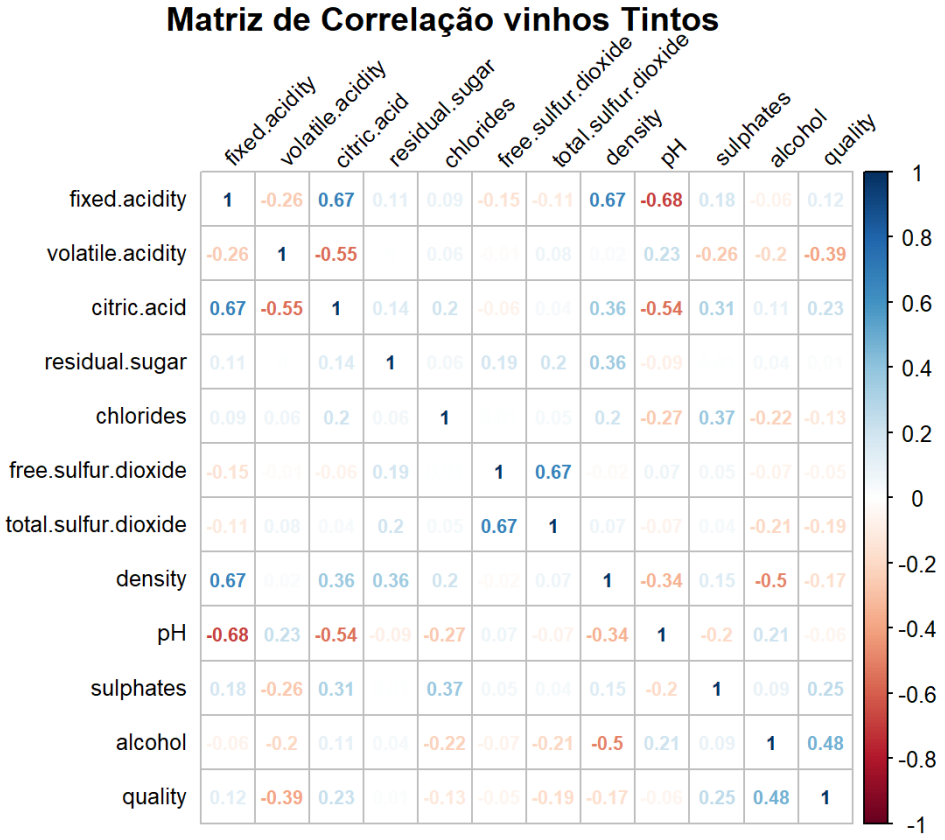
Histograma da variável qualidade segmentada por avaliação para cada tipo de vinho

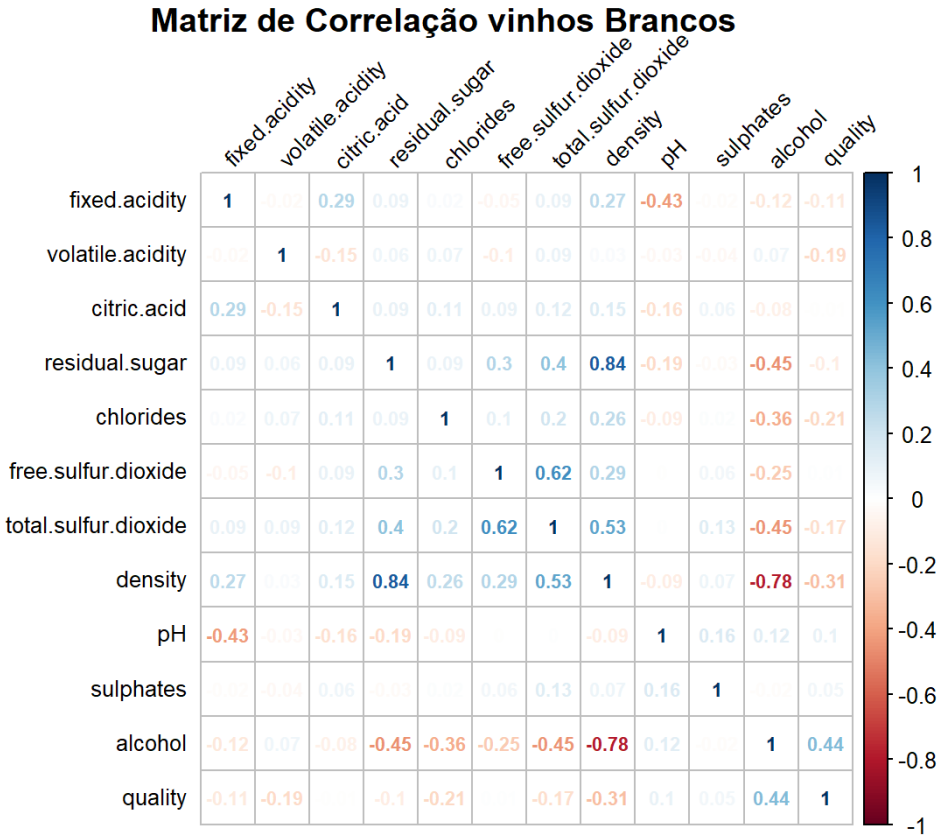
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Matriz de correlação entre as variáveis do conjunto de dados

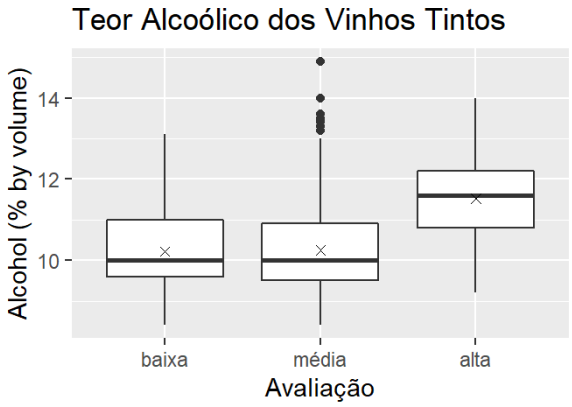
Por meio da matriz de correlação podemos encontrar os relacionamentos mais significativos entre as variáveis do conjunto de dados.

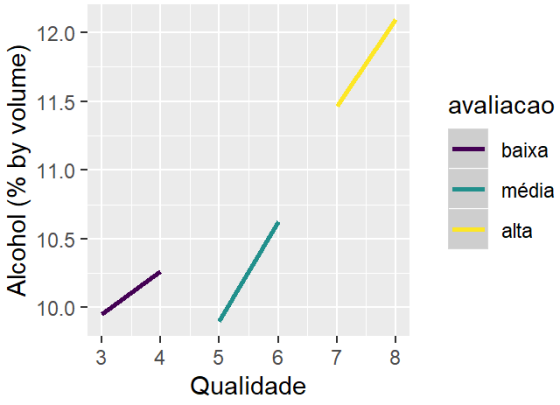


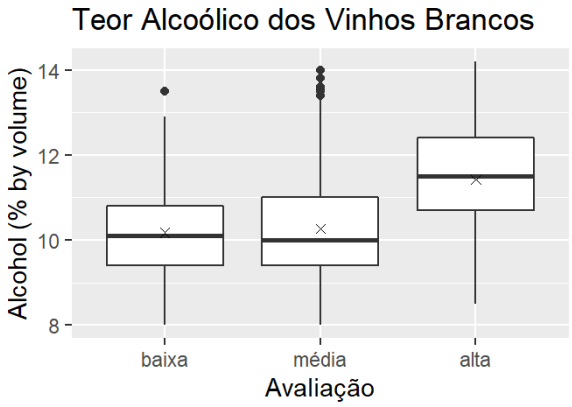


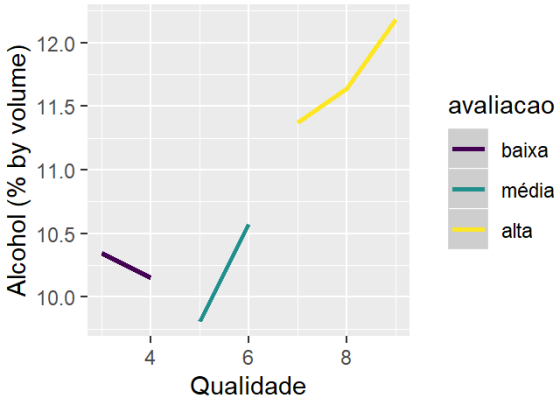
Box plot entre Teor Alcoólico e Qualidade

Como existe uma correlação significativa entre as variáveis alcohol e quality para ambos os tipos de vinho, podemos analisar adistribuição das observações entre as categorias de Avaliação. E verificamos que existe uma diferenciação do teor alcoólico dos vinhos com uma avaliação mais alta.





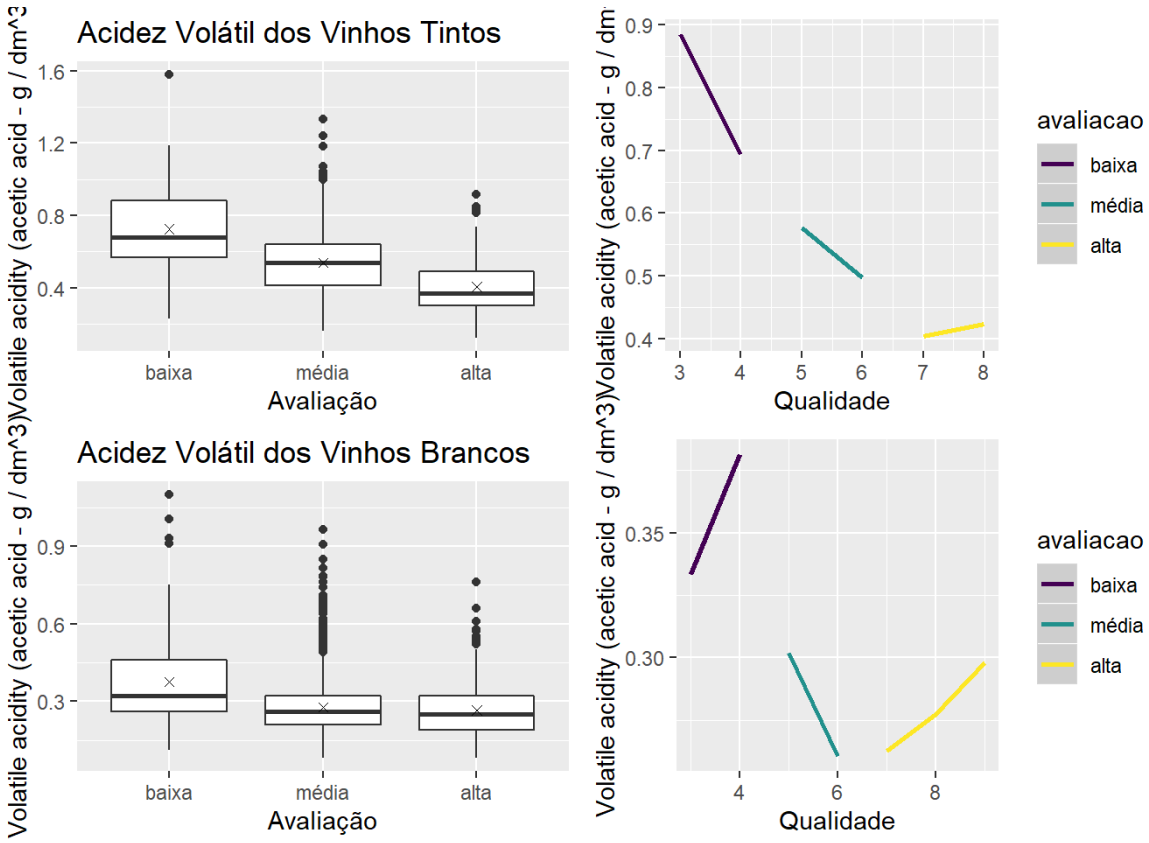




Box plot entre Volatile.acidity e Quality

Também existe uma correlação significativa entre as variáveis volatie.acidity e quality para os vinhos tintos e que não se repete para os vinhos brancos. Para os vinhos tintos, observamos que a volatile.acidity cai de acordo com que as notas de qualidade aumentam.

```
#Box plot - Volatile.acidity x quality, segmentado por Avaliacao
#ggplot(aes(x=quality, y=volatile.acidity, color=avaliacao), data=vinho_tinto) +
#  geom_boxplot() +
#  geom_jitter(alpha=1/8) +
#  facet_wrap(~avaliacao) +
#  scale_color_brewer(type='qual')

p1 <- ggplot(aes(x=avaliacao, y=volatile.acidity), data=vinho_tinto) +
  #geom_jitter(alpha=1/8) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=4) +
  xlab("Avaliação") +
  ylab("Volatile acidity (acetic acid - g / dm^3)") +
  ggtitle("Acidez Volátil dos Vinhos Tintos")

p2 <- ggplot(aes(x = quality, y = volatile.acidity), data = vinho_tinto) +
  geom_smooth(aes(color=avaliacao), stat = "summary", fun.y = mean) +
  xlab("Qualidade") +
  ylab("Volatile acidity (acetic acid - g / dm^3)")

p3 <- ggplot(aes(x=avaliacao, y=volatile.acidity), data=vinho_branco) +
  #geom_jitter(alpha=1/8) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=4) +
  xlab("Avaliação") +
  ylab("Volatile acidity (acetic acid - g / dm^3)") +
  ggtitle("Acidez Volátil dos Vinhos Brancos")

p4 <- ggplot(aes(x = quality, y = volatile.acidity), data = vinho_branco) +
  geom_smooth(aes(color=avaliacao), stat = "summary", fun.y = mean) +
  xlab("Qualidade") +
  ylab("Volatile acidity (acetic acid - g / dm^3)")

grid.arrange(p1, p2, p3, p4, ncol=2)
```



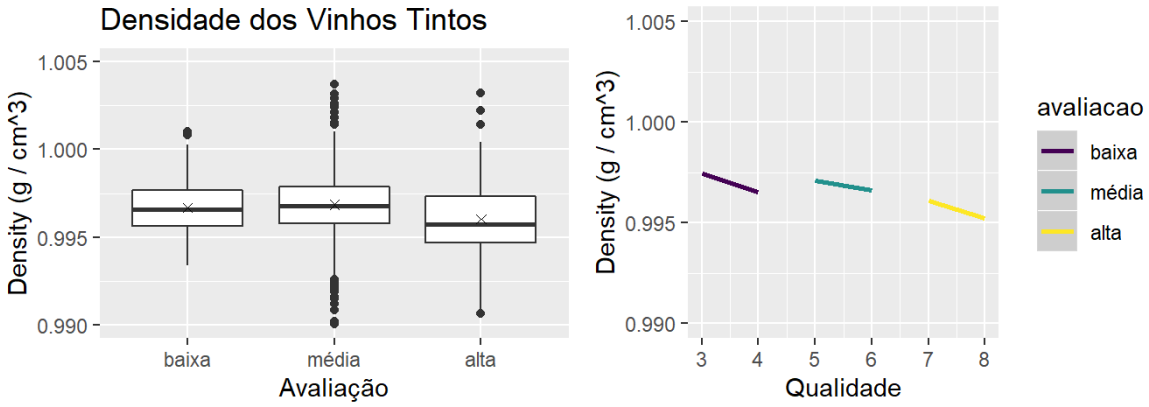
Box plot entre Density e Quality

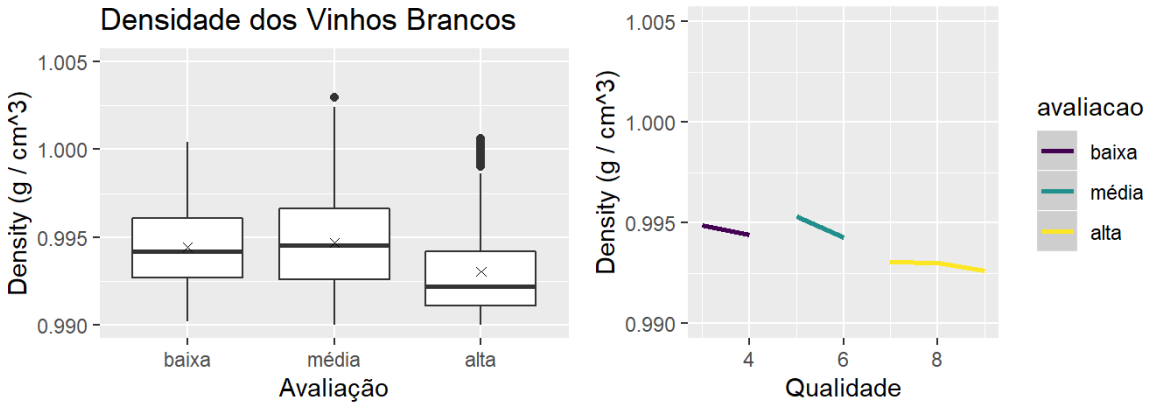
Já para o conjunto de dados de vinho branco, há uma correlação identificável entre as variáveis density e quality. Assim, podemos analisar a distribuição das observações entre as categorias de Avaliação. Infelizmente como a correlação entre as variáveis não é tão elavada, não foi possível identificar com ênfase no gráfico a segmentação entre classes de avaliação.

```
p1 <- ggplot(aes(x=avaliacao, y=density), data=vinho_tinto) +
  #geom_jitter(alpha=1/8) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=4) +
  scale_y_continuous(limits = c(0.990,1.005))  +
  xlab("Avaliação") +
  ylab("Density (g / cm^3)") +
  ggtitle("Densidade dos Vinhos Tintos")

p2 <- ggplot(aes(x = quality, y =density), data = vinho_tinto) +
  #geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_smooth(aes(), stat = "summary", fun.y = mean) +
  geom_smooth(aes(color=avaliacao), stat = "summary", fun.y = mean)+
  scale_y_continuous(limits = c(0.990,1.005)) +
  xlab("Qualidade") +
  ylab("Density (g / cm^3)")

p3 <- ggplot(aes(x=avaliacao, y=density), data=vinho_branco) +
  #geom_jitter(alpha=1/8) +
  geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=4) +
  scale_y_continuous(limits = c(0.990,1.005))  +
  xlab("Avaliação") +
  ylab("Density (g / cm^3)") +
  ggtitle("Densidade dos Vinhos Brancos")

p4 <- ggplot(aes(x = quality, y =density), data = vinho_branco) +
  #geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_smooth(aes(), stat = "summary", fun.y = mean) +
  geom_smooth(aes(color=avaliacao), stat = "summary", fun.y = mean) +
  scale_y_continuous(limits = c(0.990,1.005)) +
  xlab("Qualidade") +
  ylab("Density (g / cm^3)")

grid.arrange(p1, p2, p3, p4, ncol=2)
```

```
## Warning: Removed 348 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 348 rows containing non-finite values (stat_summary).

## Warning: Removed 348 rows containing non-finite values (stat_summary).
```



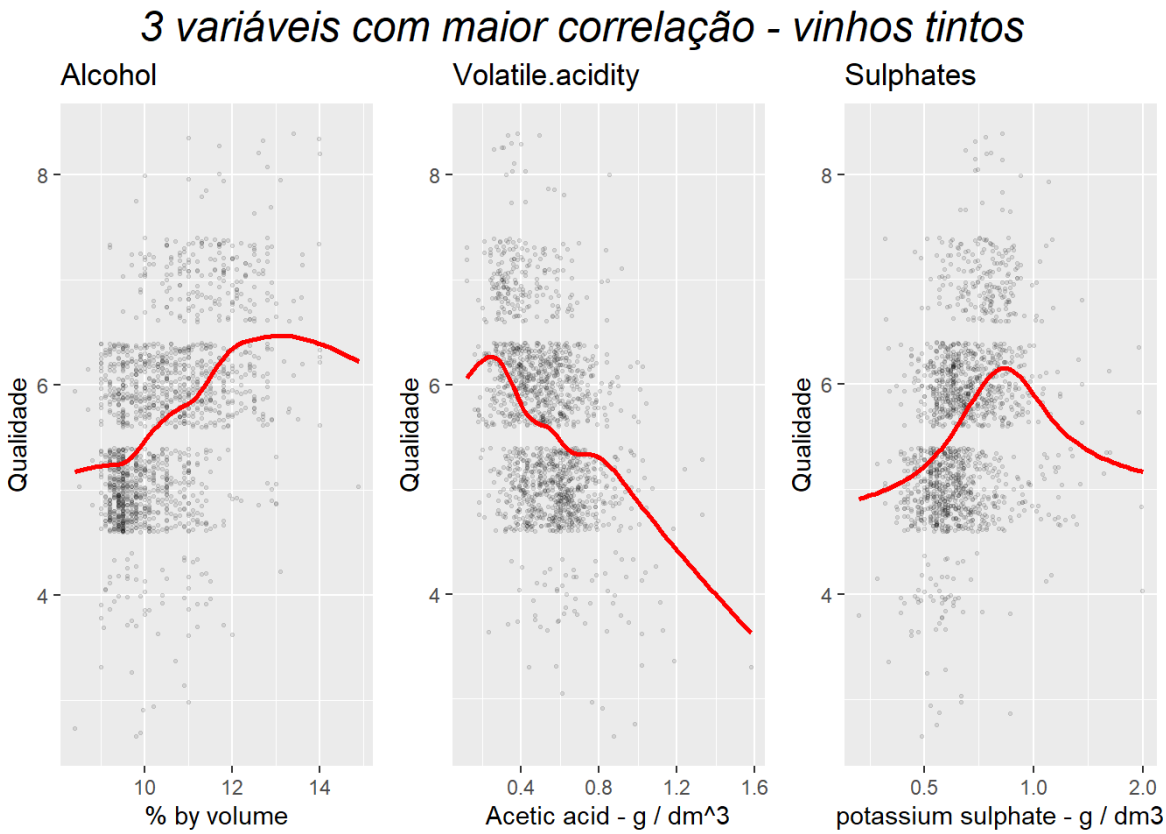


Relação entre Qualidade e as três variáveis com maior correlação para cada tipo de vinho

Após analisarmos cada uma das uma das correlações de forma individualizada, podemos construir uma visão consolidada entre elas. E verificamos o comportamento de cada uma delas com o aumento das notas de qualidade, onde alcohol tem uma variação positiva, volatile.acidity negativa e sulphates segue uma distribuição normal, no caso dos vinhos tintos.

```
# Relacao Qualidade e tres variaveis com mais correlacao para os vinhos tinto

p1 <- ggplot(aes(x = alcohol, y =quality), data = vinho_tinto) +
  geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_line(aes(), stat = "summary", fun.y = mean)
  geom_smooth(se = FALSE,color='red') +
  xlab("% by volume") +
  ylab("Qualidade") +
  ggtitle("Alcohol")

p2 <- ggplot(aes(x = volatile.acidity, y =quality), data = vinho_tinto) +
  geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_line(aes(), stat = "summary", fun.y = mean)
  geom_smooth(se = FALSE,color='red') +
  xlab("Acetic acid - g / dm^3") +
  ylab("Qualidade") +
  ggtitle("Volatile.acidity")

p3 <- ggplot(aes(x = sulphates, y =quality), data = vinho_tinto) +
  geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_line(aes(), stat = "summary", fun.y = mean)
  geom_smooth(se = FALSE,color='red') +
  scale_x_log10() +
  xlab("potassium sulphate - g / dm3") +
  ylab("Qualidade") +
  ggtitle("Sulphates")

grid.arrange(p1, p2, p3, ncol=3, top = textGrob("3 variáveis com maior correlação - vinhos tintos",gp=gpar
(fontsize=18,font=3)))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



No caso dos vinhos brancos, observamos que o aumento das notas de qualidade, acompanha uma variação positiva para o teor alcoólico, negativa para densidade e cloretos.

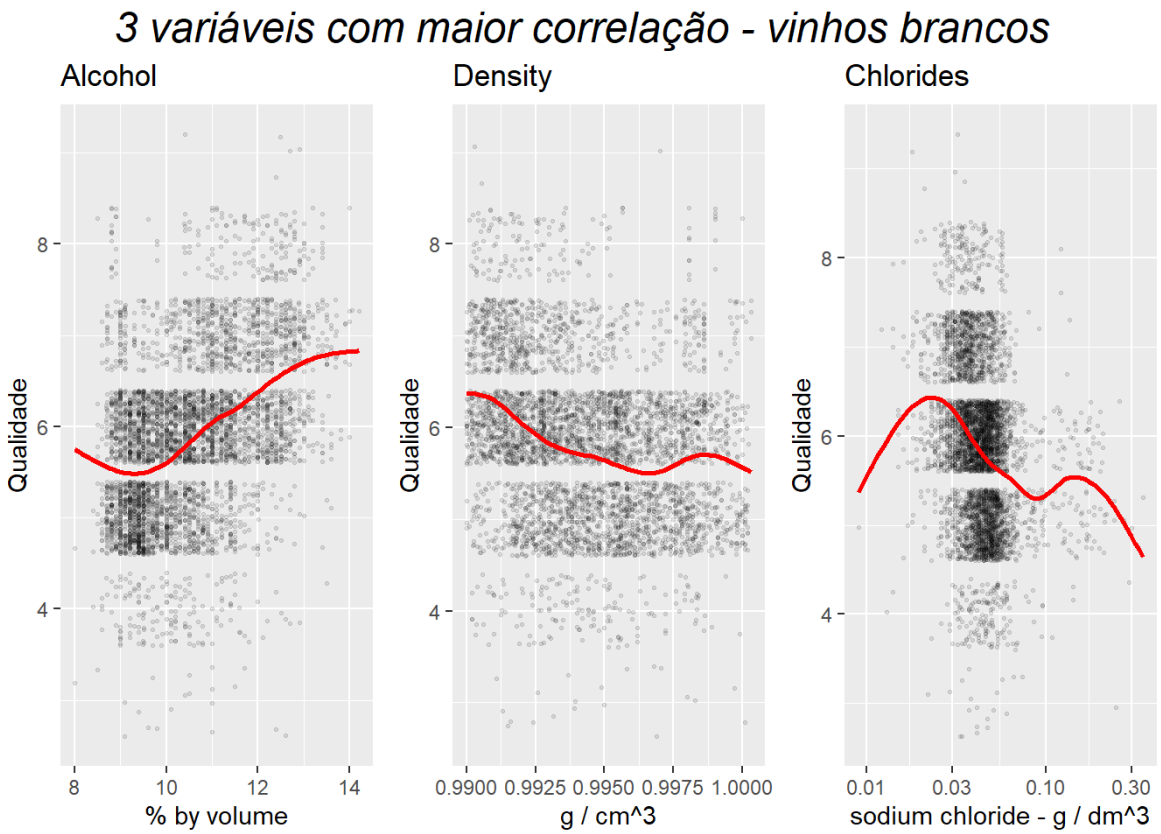
```
# Relacao Qualidade e tres variaveis com mais correlacao para os vinhos brancos

p1 <- ggplot(aes(x = alcohol, y =quality), data = vinho_branco) +
  geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_line(aes(), stat = "summary", fun.y = mean)
  geom_smooth(se = FALSE,color='red') +
  xlab("% by volume") +
  ylab("Qualidade") +
  ggtitle("Alcohol")

p2 <- ggplot(aes(x = density, y =quality), data = vinho_branco) +
  geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_line(aes(), stat = "summary", fun.y = mean)
  geom_smooth(se = FALSE,color='red')  +
  scale_x_continuous(limits = c(0.990,quantile(vinho_branco$density, 0.99))) +
  xlab("g / cm^3") +
  ylab("Qualidade") +
  ggtitle("Density")

p3 <- ggplot(aes(x = chlorides, y =quality), data = vinho_branco) +
  geom_point(alpha = 1/10, size = 1/2, position = 'jitter') +
  #geom_line(aes(), stat = "summary", fun.y = mean)
  geom_smooth(se = FALSE,color='red') +
  scale_x_log10() +
  xlab("sodium chloride - g / dm^3") +
  ylab("Qualidade") +
  ggtitle("Chlorides")

grid.arrange(p1, p2, p3, ncol=3, top = textGrob("3 variáveis com maior correlação - vinhos brancos",gp=gpar
(fontsize=18,font=3)))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 394 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 407 rows containing missing values (geom_point).
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

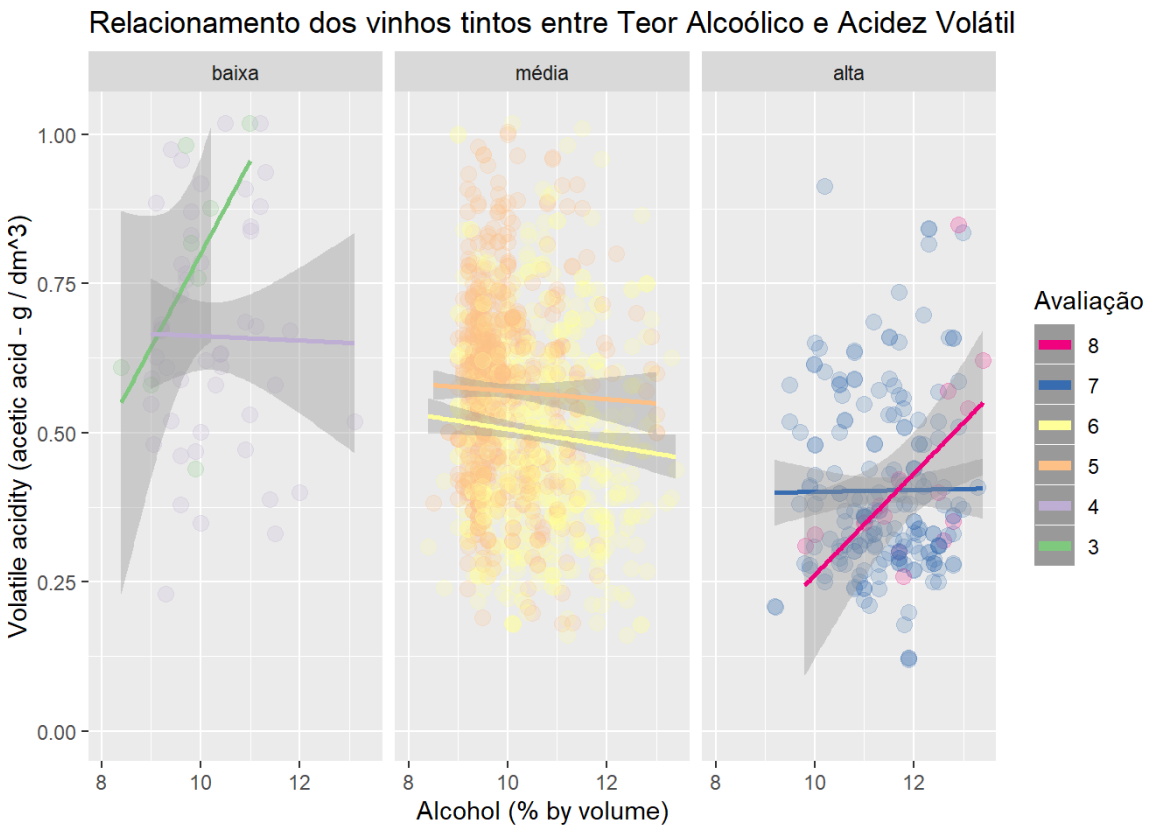


O relacionamento entre Alcohol, Volatile.acidity e Quality nos vinhos tintos

Como os maiores índices de correlação da variável quality são com as variáveis alcohol e volatile.acidity para o conjunto de dados dos vinhos tintos, iremos representá-las em um gráfico segmentado pelo grau de avaliação.

```
## Warning: Removed 29 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 30 rows containing missing values (geom_point).
```



O relacionamento entre Alcohol, Density e Quality nos vinhos brancos

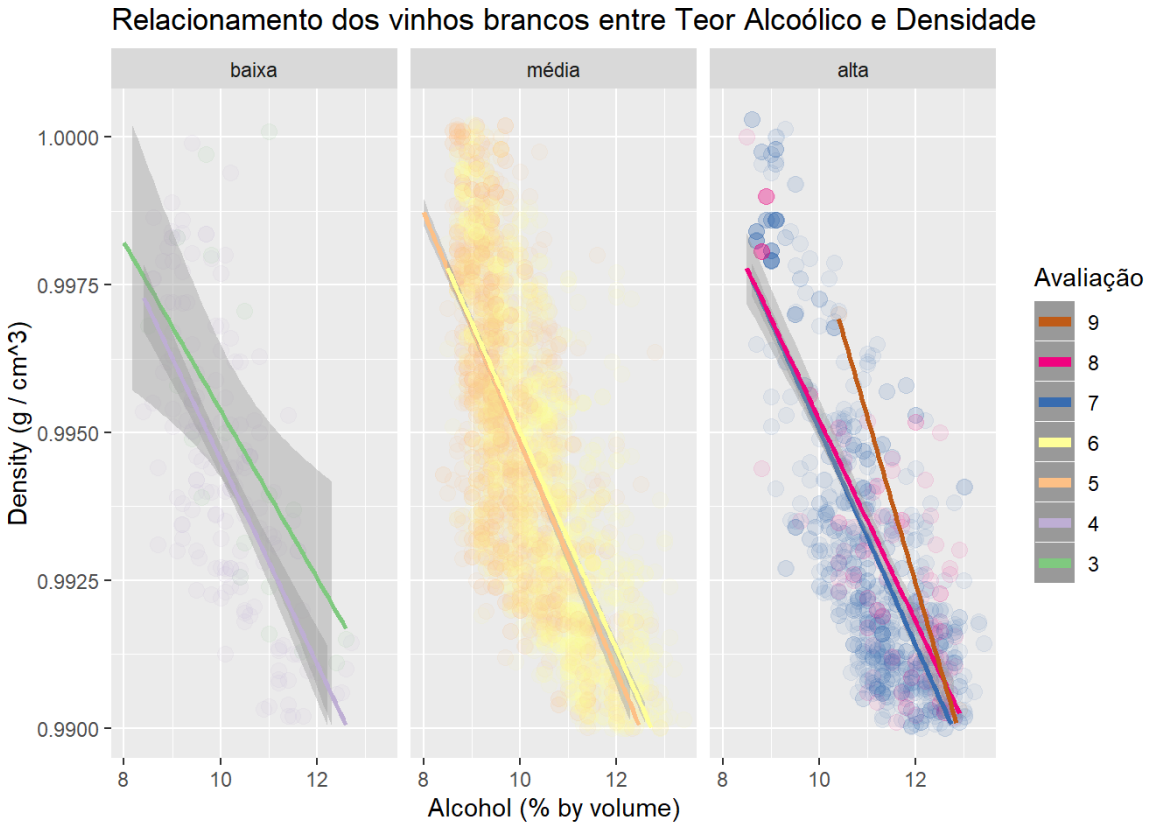
Como os maiores índices de correlação da variável quality são com as variáveis alcohol e volatile.acidity para o conjunto de dados dos vinhos brancos, iremos representá-las em um gráfico segmentado pelo grau de avaliação.

```
#
#Quality possui as maiores correlações com as variaveis: alcohol e volatile.acidity. No grafico abaixo rela
cionamos alcohol e volatile.acidity, destacando a variavel quality e segmentando a visualização por Avaliaç
ão.
ggplot(data = vinho_branco, aes(x = alcohol, y = density, color=as.factor(quality))) +
    facet_wrap(~avaliacao) +
    scale_x_continuous(lim = c(8, quantile(vinho_branco$alcohol, 0.99))) +
    scale_y_continuous(lim = c(0.990, quantile(vinho_branco$density, 0.99))) +
    geom_point(alpha = 0.075, size = 3, position = 'jitter') +
  scale_color_brewer(type = 'qual',
    guide = guide_legend(title = 'Avaliação', reverse = T,
    override.aes = list(alpha = 1, size = 2)), palette = 1) +
    stat_smooth(method = 'lm') +
  xlab("Alcohol (% by volume)") +
  ylab("Density (g / cm^3)") +
  ggtitle("Relacionamento dos vinhos brancos entre Teor Alcoólico e Densidade")
```

```
## Warning: Removed 406 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 427 rows containing missing values (geom_point).
```

```
## Warning: Removed 33 rows containing missing values (geom_smooth).
```



#Referências utilizadas:

- https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html (https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html)
- http://rstudio-pubs-static.s3.amazonaws.com/198466_b17daa66ce6748a6a91cd27017608720.html (http://rstudio-pubs-static.s3.amazonaws.com/198466_b17daa66ce6748a6a91cd27017608720.html)
- http://rstudio-pubs-static.s3.amazonaws.com/53416_83b9685bc8c54afebcbb1e65a7c688fc.html (http://rstudio-pubs-static.s3.amazonaws.com/53416_83b9685bc8c54afebcbb1e65a7c688fc.html)
- https://rpubs.com/inageorgescu/whitewine2 (https://rpubs.com/inageorgescu/whitewine2)
- http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram (http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram)
- https://www.wardsci.com/www.wardsci.com/images/Chemistry_of_Wine.pdf (https://www.wardsci.com/www.wardsci.com/images/Chemistry_of_Wine.pdf)
- https://revistaadega.uol.com.br/artigo/o-alcool-e-acidez-dos-vinhos_6055.html (https://revistaadega.uol.com.br/artigo/o-alcool-e-acidez-dos-vinhos_6055.html)
- https://rpubs.com/szon0111/P4#targetText=Wine%20seems%20to%20have%20better,is%20between%208%20and%2010 (https://rpubs.com/szon0111/P4#targetText=Wine%20seems%20to%20have%20better,is%20between%208%20and%2010)
- http://periodicos.ses.sp.bvs.br/scielo.php?script=sci_arttext&pid=S0073-98552011000200009&lng=pt&nrm=iso (http://periodicos.ses.sp.bvs.br/scielo.php?script=sci_arttext&pid=S0073-98552011000200009&lng=pt&nrm=iso)