MINISTÉRIO DA EDUCAÇÃO
Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas
IFSULDEMINAS - Câmpus Poços de Caldas
Avenida Dirce Pereira Rosa, 300. Poços de Caldas/MG. CEP 37713-100
Fone: (35) 3713-5120

Inteligência Artificial - Trabalho Prático 2

Pré-Processamento e Análise de Dados

Prof. Douglas Castilho

Disponível Desde: 11 de maio de 2023

Data de Entrega: 08 de junho de 2023

Valor: 1.5 pontos

Adrian Damião

Objetivo

O objetivo deste trabalho é analisar uma base de dados escolhida pelo professor e aplicar técnicas de pré-processamento para que seja possível utilizar um algoritmo de aprendizado de máquina de maneira mais eficiente para predizer algum atributo de saída.

Base de dados escolhida

A base de dados que será utilizada para realizar as análises se trata de uma extração razoavelmente limpa feita por Barry Becker do banco de dados do Censo de 1994. O objetivo é utilizá-la na classificação de pessoas que ganham mais de 50 mil dólares por ano.

Atributos da base de dados

Age
Workclass
Fnlwgt
Education
Education-num
Marital-status
Occupation
Relationship
Race
Sex
Capital-gain
Capital-loss
Hours-per-week
Native-country
Above-Limit → Nome escolhido para o atributo de saída

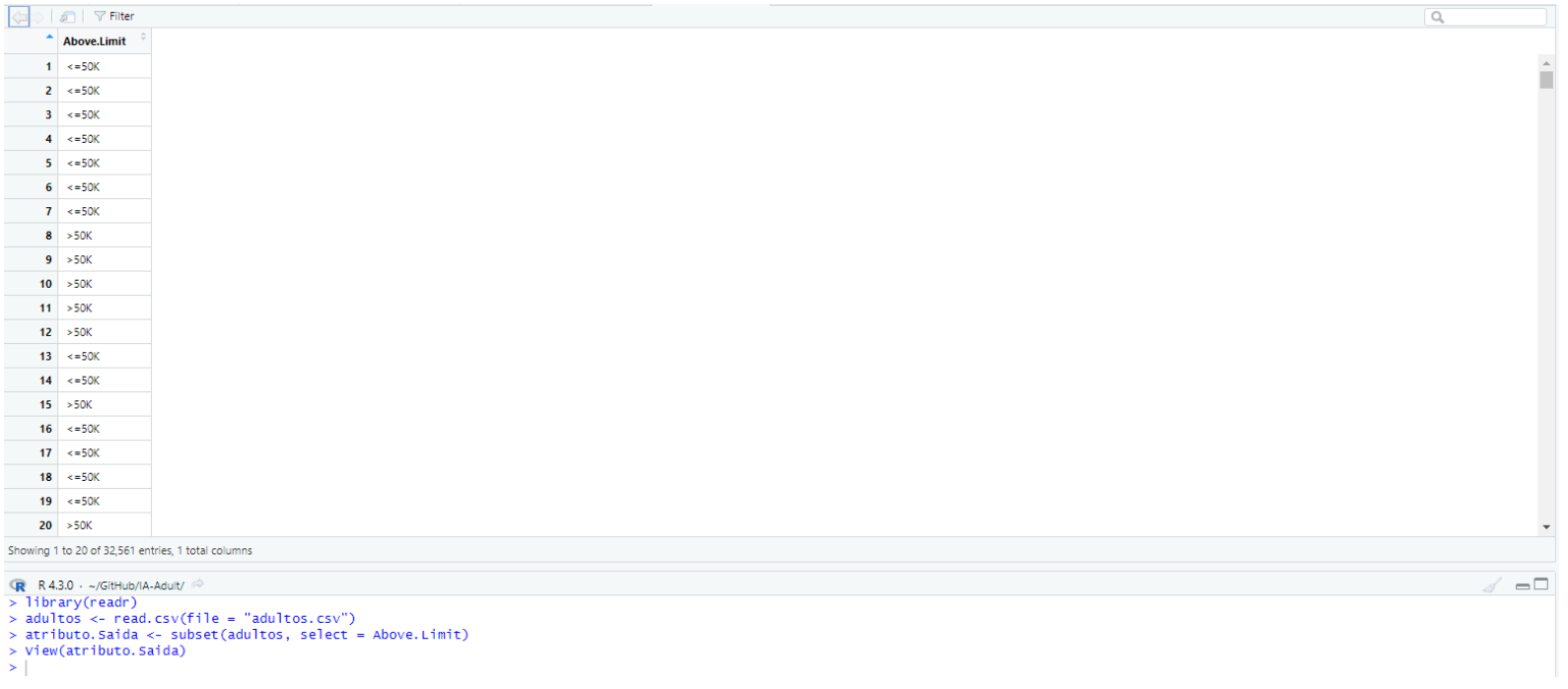
Abaixo, podemos ver como visualizar o dataset e o comando utilizado para exibi-lo:

	Age	Workclass	Fnlwgt	Education	Education.num	Marital.status	Occupation	Relationship	Race	Sex	Capital.gain	Capital.loss	Hours.per.week	Native.country	Above.Limit
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
11	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
13	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
15	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
17	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K

```
R 4.3.0 · ~/GitHub/IA-Adult/
> library(readr)
> adultos <- read.csv(file = "adultos.csv")
> view(adultos)
>
```

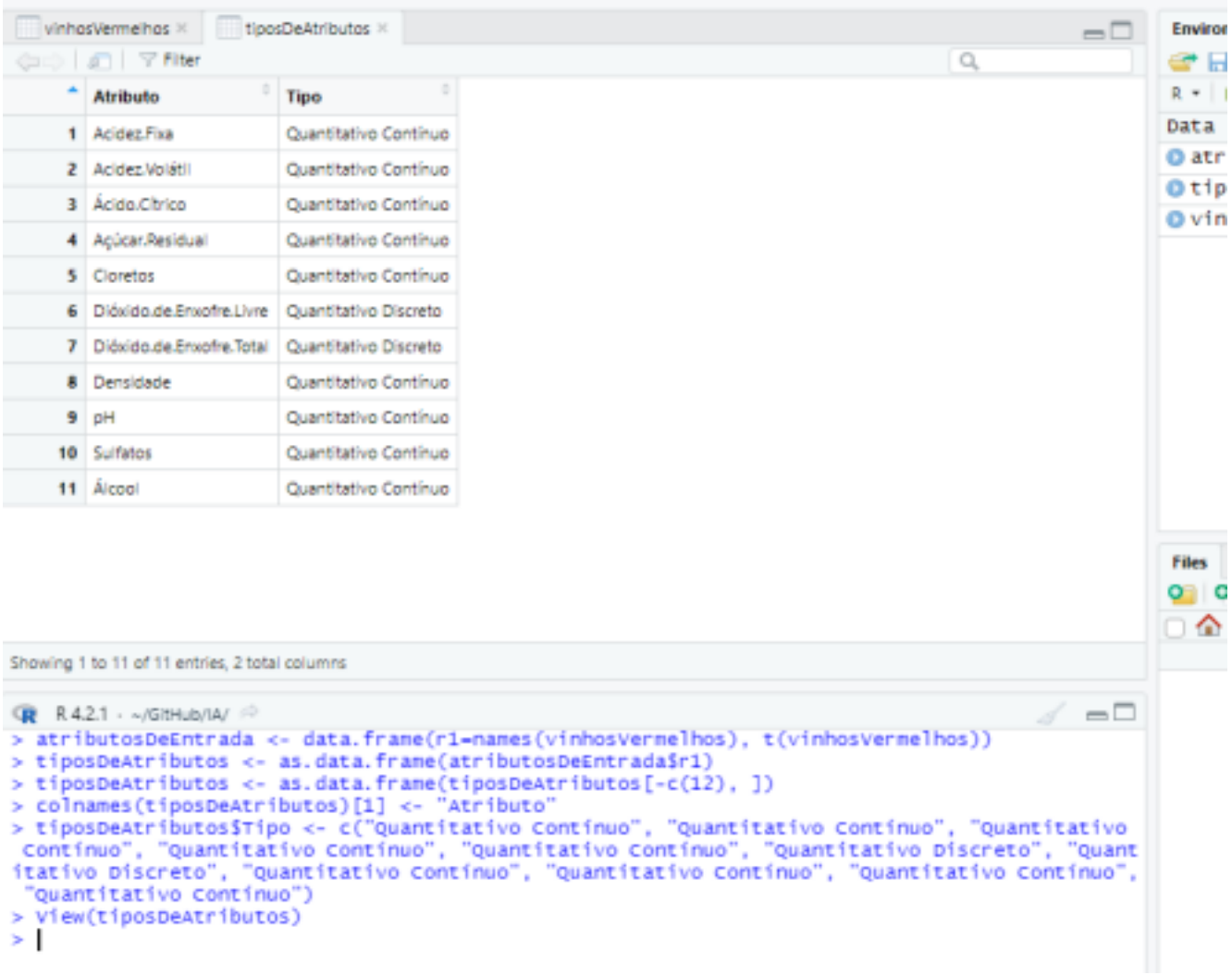
1. Identificação do atributo alvo (saída)

Como atributo de saída, será utilizado o atributo intitulado **Above-Limit**, onde por meio dos outros atributos socioeconômicos, descobriremos se o indivíduo ganha acima de 50 mil dólares por ano ou não. Para exibi-lo podemos usar os seguintes comandos:



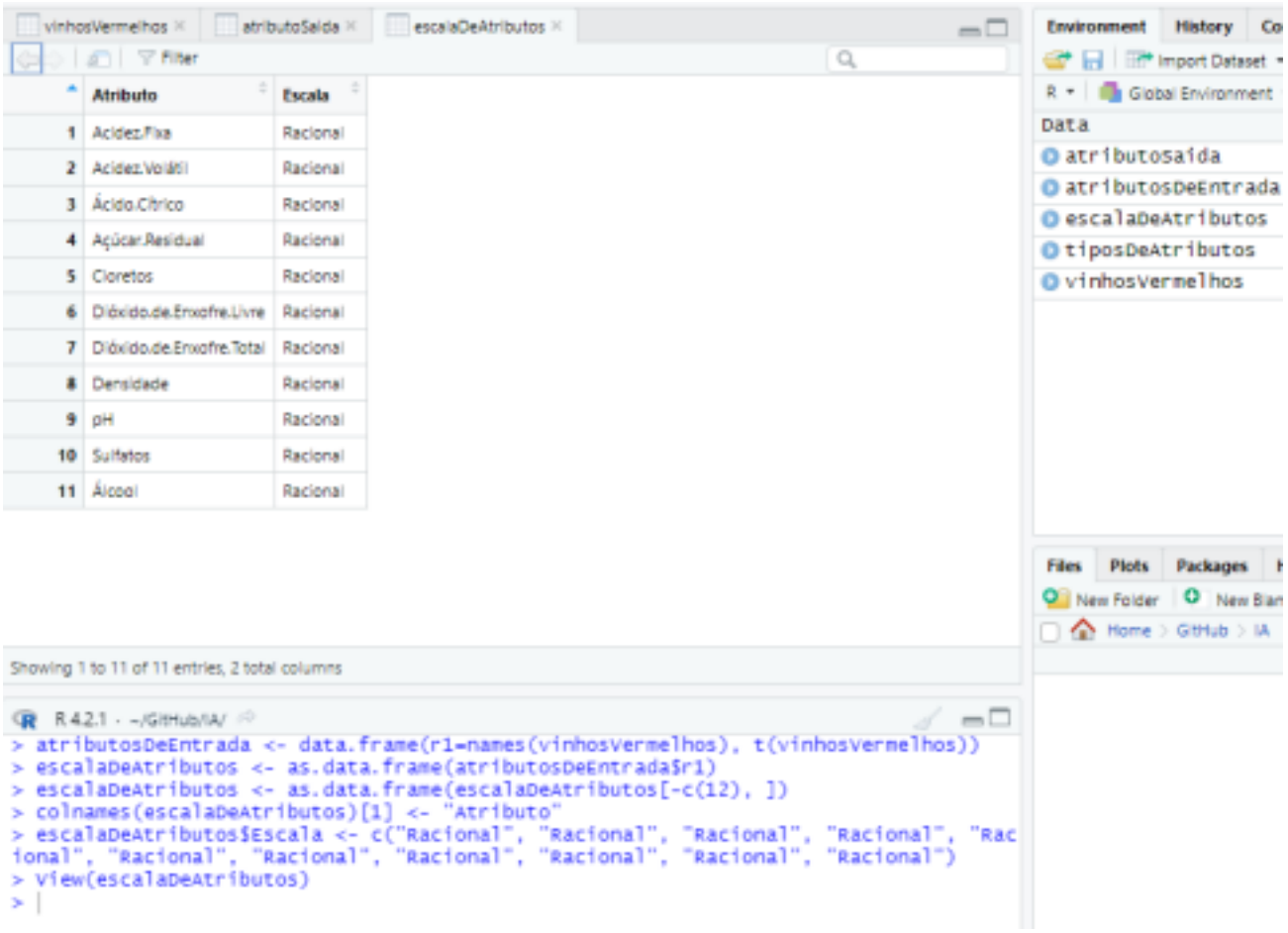
2. Identificação dos tipos de dados dos atributos de entrada (quantitativo, qualitativo)

Para identificar os tipos de dados dos atributos de entrada, podemos utilizar os seguintes comandos:



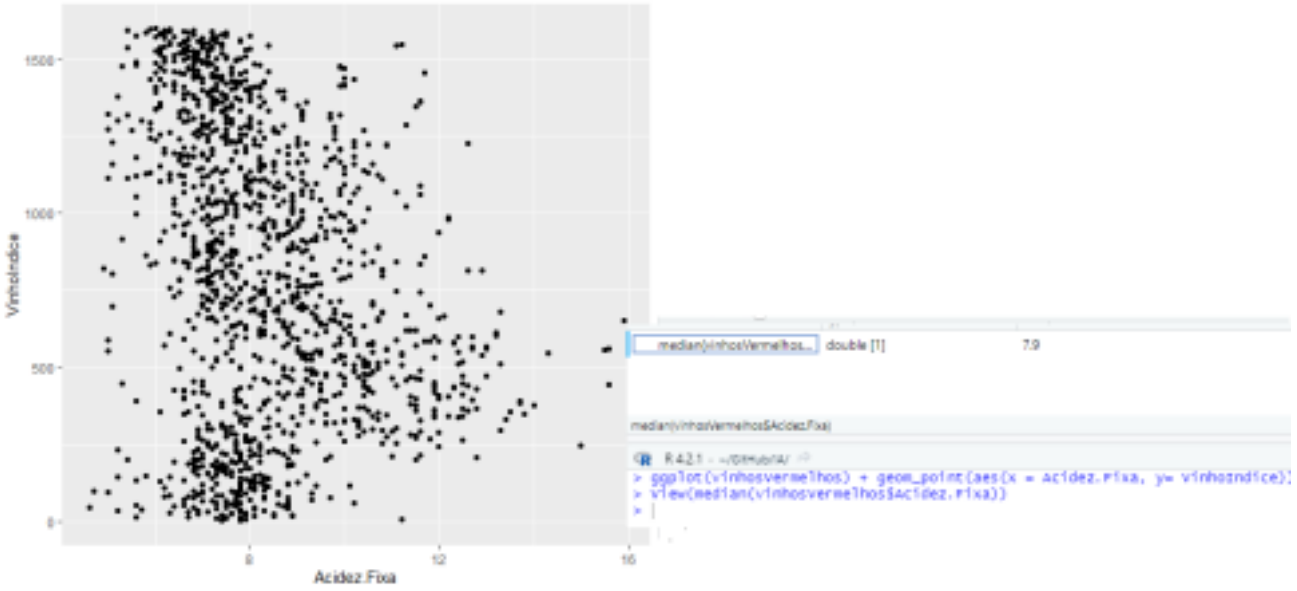
3. Identificação da escala de dados dos atributos de entrada (nominal, ordinal, intervalar, racional)

Para visualizar a escala de dados dos atributos de entrada podemos utilizar os seguintes comandos abaixo:



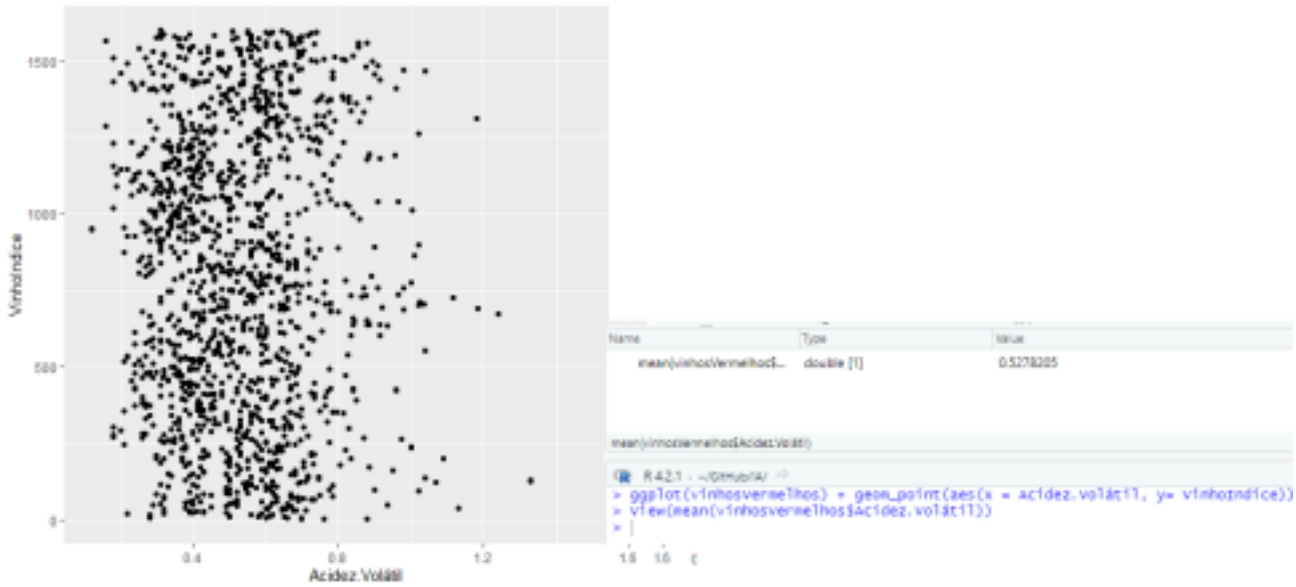
4. Exploração dos dados através de medidas de localidade

Acidez Fixa



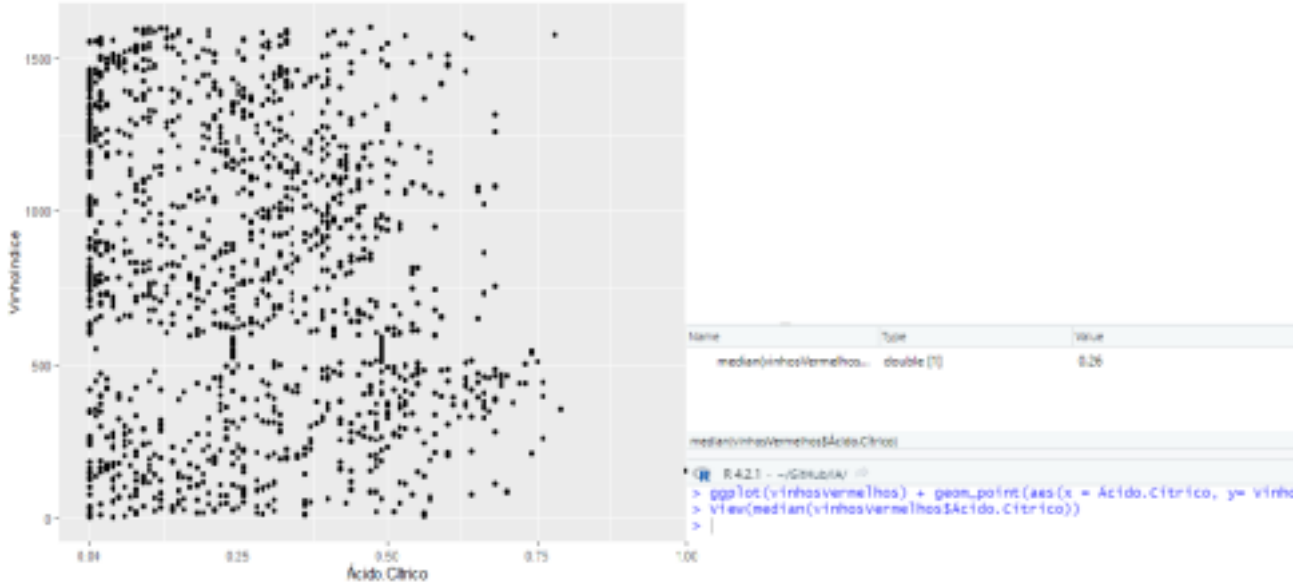
Podemos ver que a Acidez Fixa possui muitos *outliers*, logo optou-se por fazer uma mediana.

Acidez Volátil



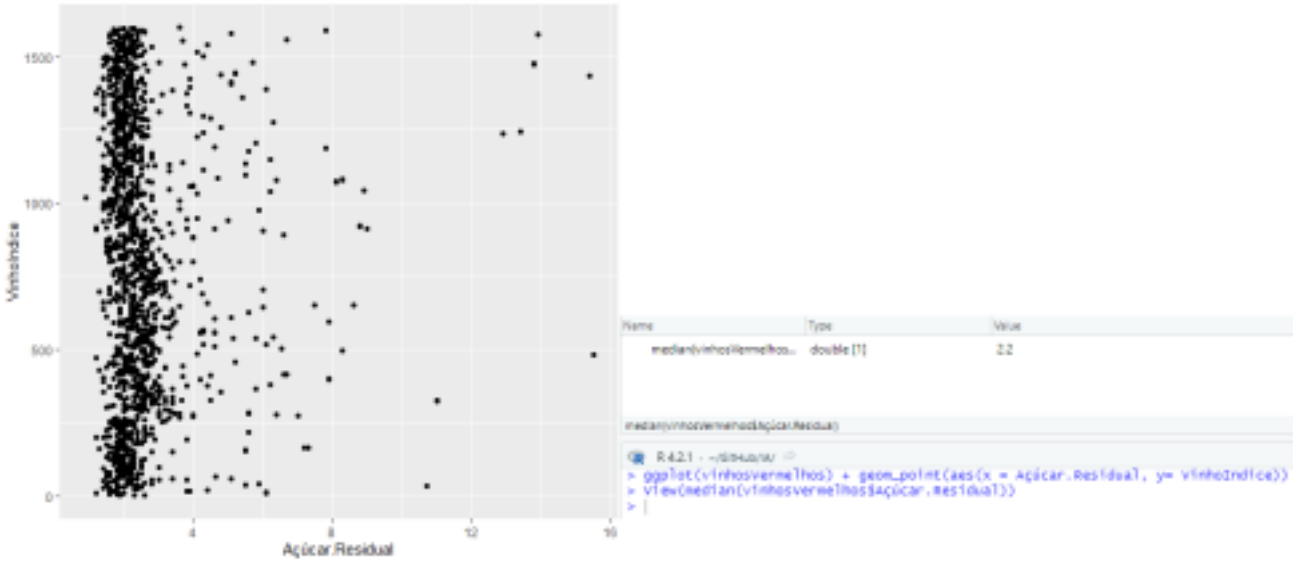
Com a Acidez Volátil foi usada a média pois existem poucas ocorrências de *outlier*.

Ácido Cítrico



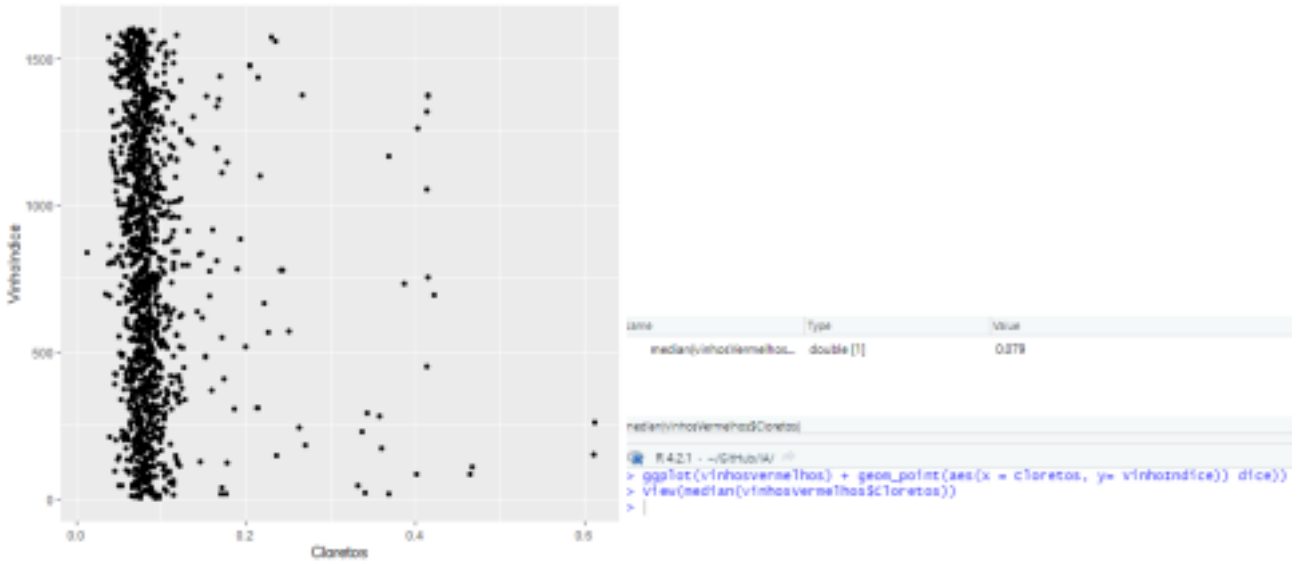
No caso do Ácido Cítrico, os valores estão muito dispersos, por isso, foi optado por fazer uma mediana.

Açúcar Residual



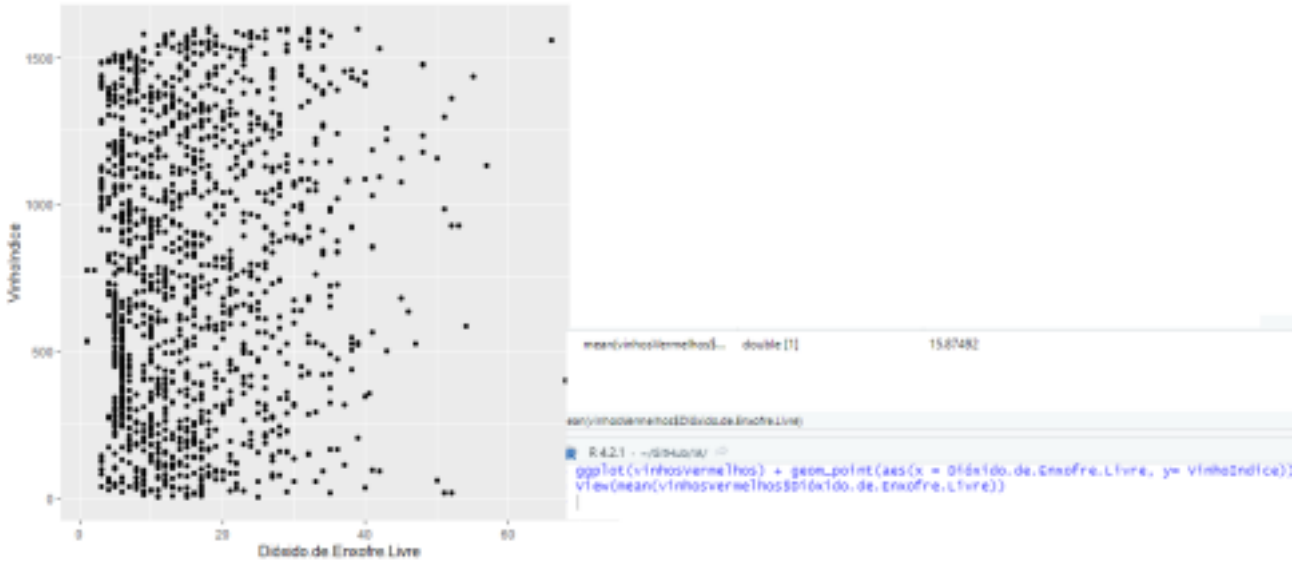
Devido ao fato da maioria dos valores estarem localizados juntos a alguns exemplares estarem distantes dessa linha principal, estes foram considerados como *outliers*, portanto, usou-se uma mediana.

Cloretos



Pelo mesmo motivo do atributo anterior, foi utilizada uma mediana.

Dióxido de Enxofre Livre

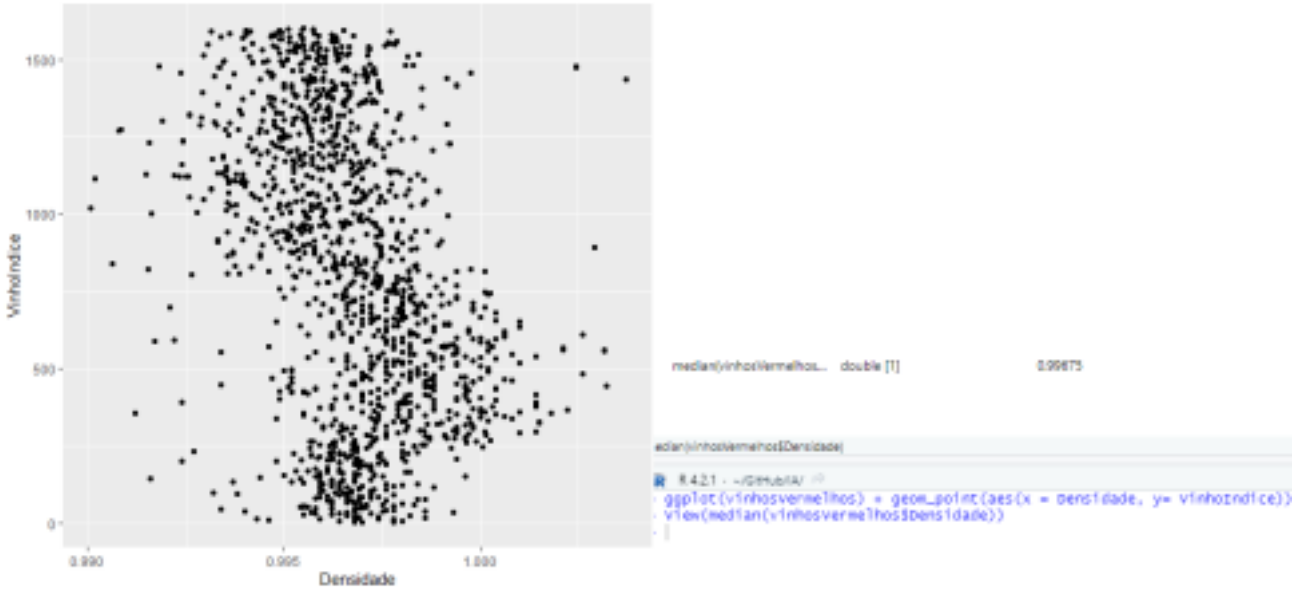


Nesse caso, apenas alguns dos exemplares estão localizados distantes da maioria, nesse caso pode-se utilizar uma média.

Dióxido de Enxofre Total

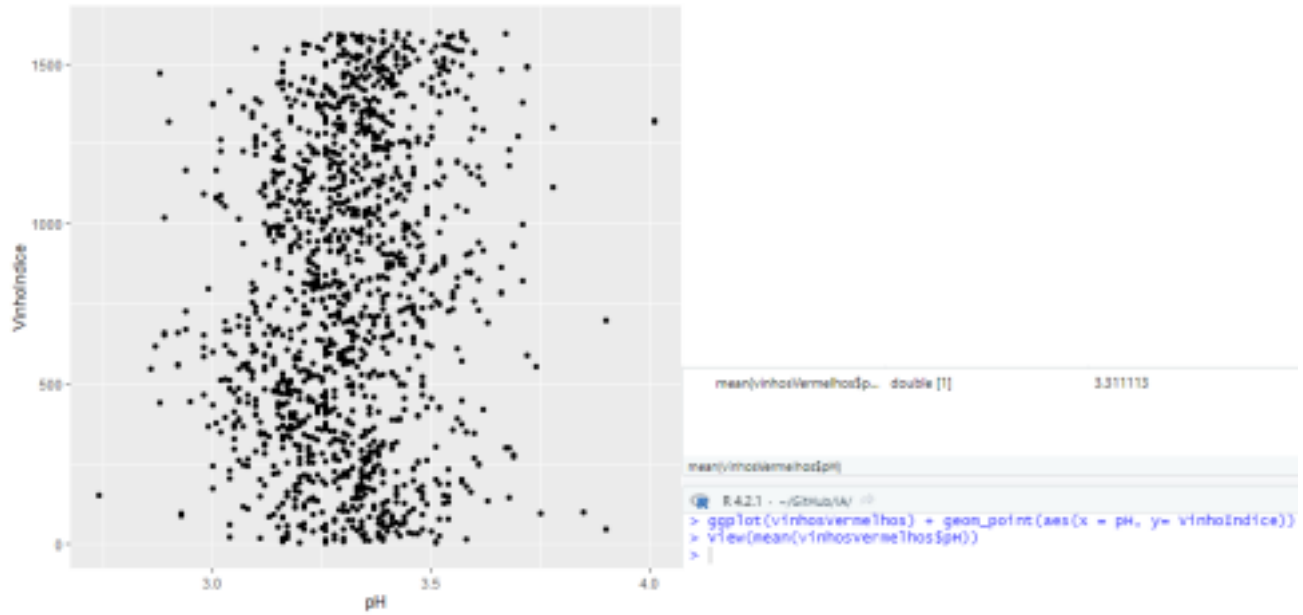
Pode-se fazer uma média pois existem muito poucos *outliers.*

Densidade



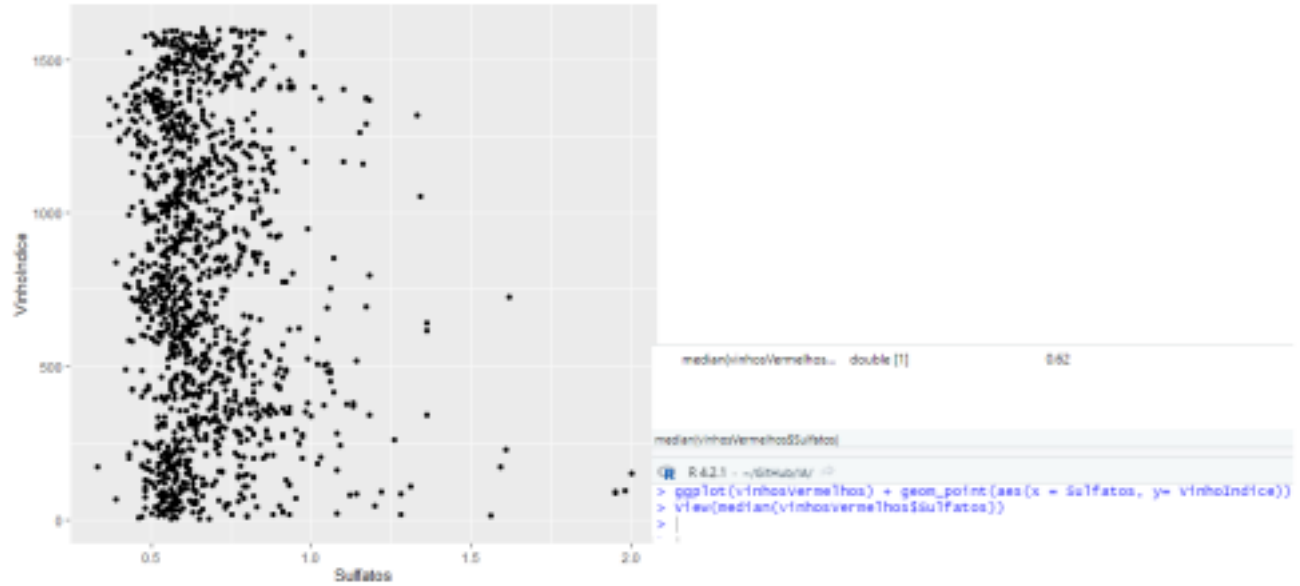
Os exemplares estão localizados muito distantes, por isso é melhor usar uma mediana.

pH



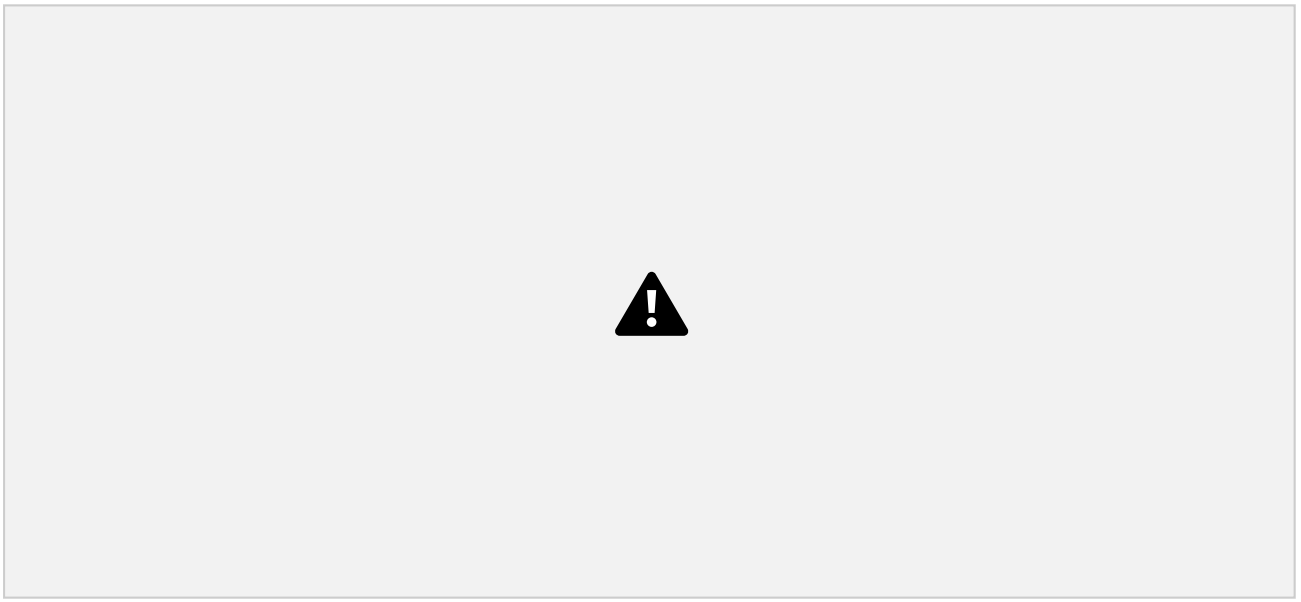
Foi optado utilizar uma mediana pois existem valores fora da curva.

Sulfatos



Pode-se ver vários *outliers*, portanto, mediana.

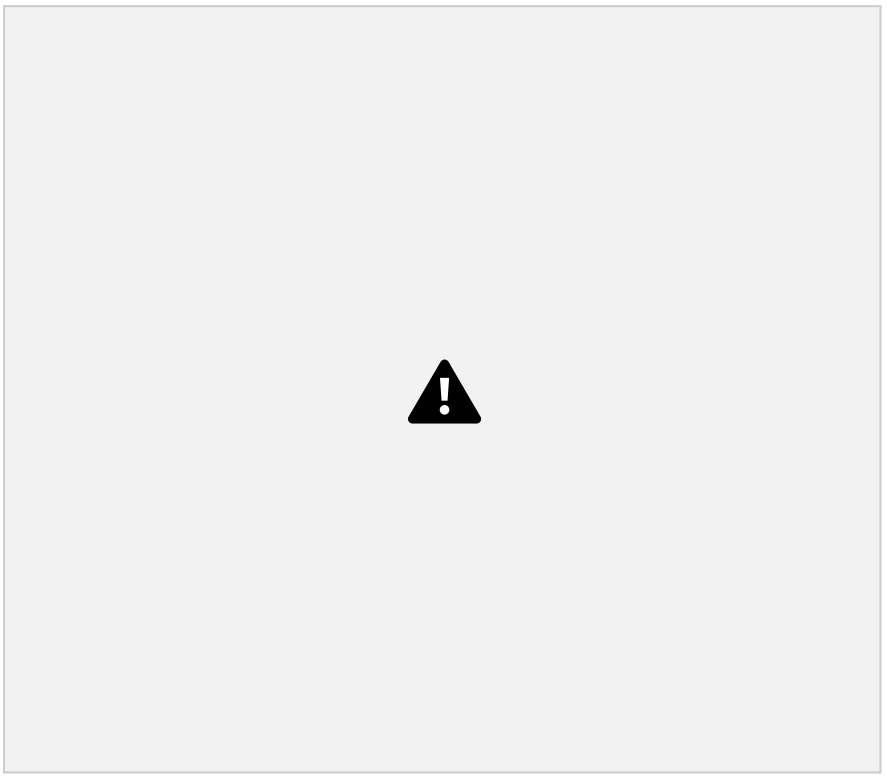
Álcool



Os valores estão localizados de maneira dispersa, por isso, mediana.

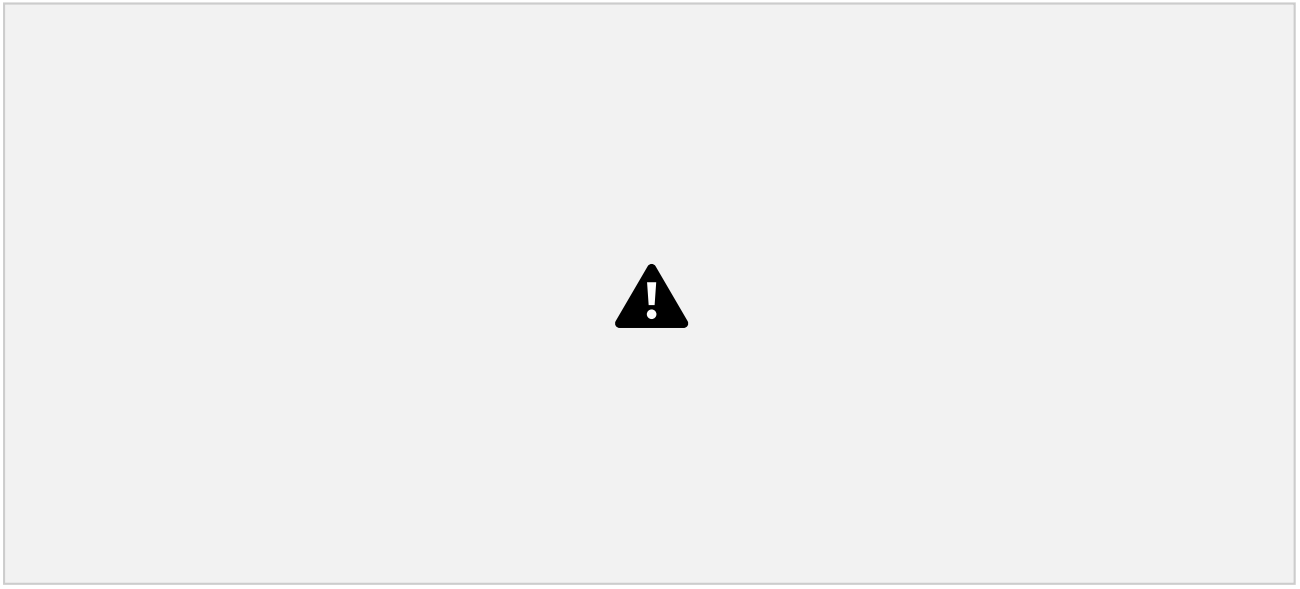
5. Exploração dos dados através de medidas de espalhamento

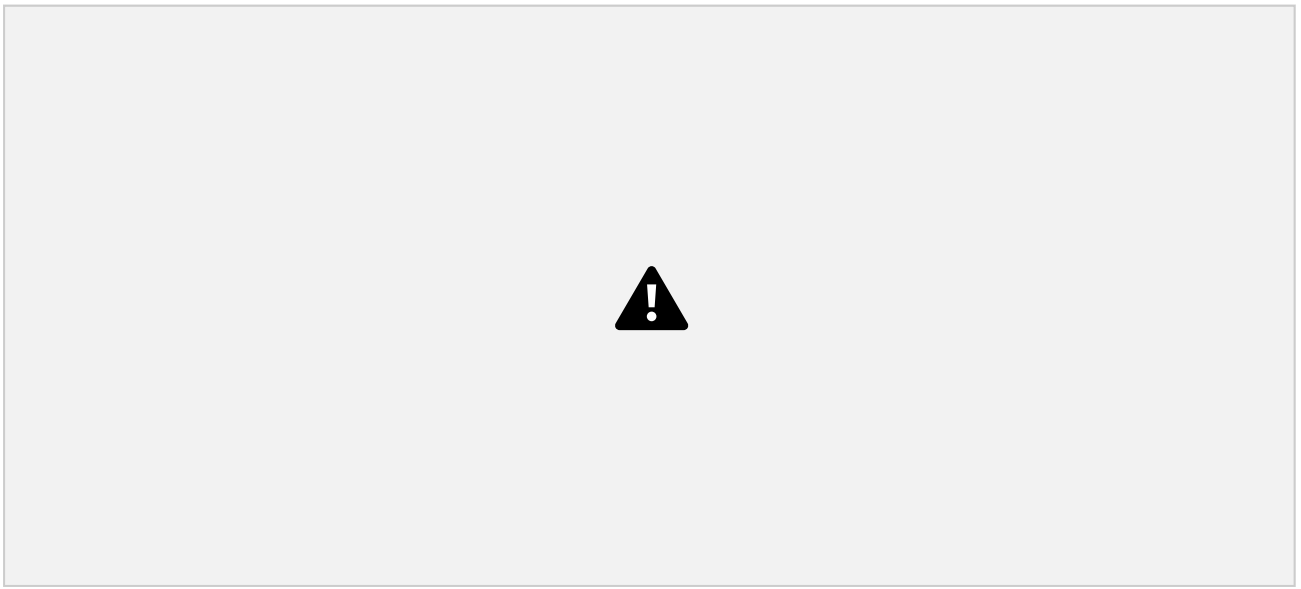
Para explorar as medidas de espalhamento, nos atributos Acidez Volátil, Dióxido de Enxofre Livre e Total foi possível utilizar apenas uma variância, devido a presença de poucos *outliers*, já nos demais atributos foi utilizado o desvio padrão para tentar minimizar os danos dos *outliers*.



6. Exploração dos dados através de medidas de distribuição

Para se explorar os dados utilizando medidas de distribuição, é possível utilizar os histogramas. Pode-se ver nos histogramas abaixo, que a maioria dos atributos possuem exemplares que se distribuem em faixas de valores próximos, com exceção do Ácido Cítrico, o Dióxido de Enxofre e o Álcool, que são mais dispersos.





7. Identificação e separação do conjunto de teste, que será utilizado para testar o desempenho dos modelos

– o conjunto de testes deve ser representativo e ter as características da população completa. Caso sua base de dados já tenha o conjunto de teste definido, analisar se este segue as características do conjunto de treinamento;

8. Identificação e eliminação de atributos não necessários

Ao pesquisar sobre o assunto de análise da qualidade de vinhos, podemos encontrar em diversos lugares que todos os atributos presentes no dataset são utilizados para determinar a qualidade de um vinho, com exceção da densidade. Outro motivo para não se utilizar a densidade como parâmetro preditivo é porque os valores dela não variam muito entre os exemplares.

9. Identificação e eliminação de exemplos não necessários

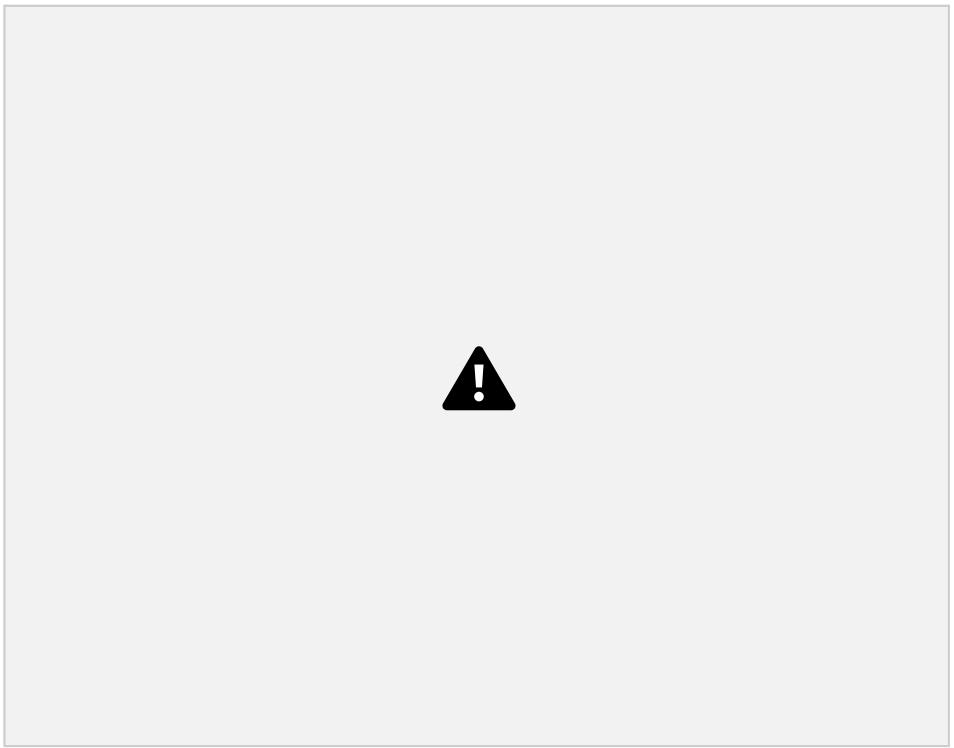
10. Análise e aplicação de técnicas de amostragem de dados (caso não seja necessário, analisar o porquê)

No caso deste dataset, não será necessário o uso de técnicas de amostragem muito complexas, pois o número de exemplos pode ser considerado baixo, sendo possível fazer a análise com eficiência e sem muito custo computacional, portanto, foi feita uma amostragem aleatória, pegando aleatoriamente 217 elementos da classe majoritária e 217 elementos da classe minoritária(para manter a base balanceada) considerando a seguinte classificação:

- Exemplos com qualidade igual ou superior a 7 foram considerados bons.
- Exemplos com qualidade abaixo de 6 foram considerados ruins.

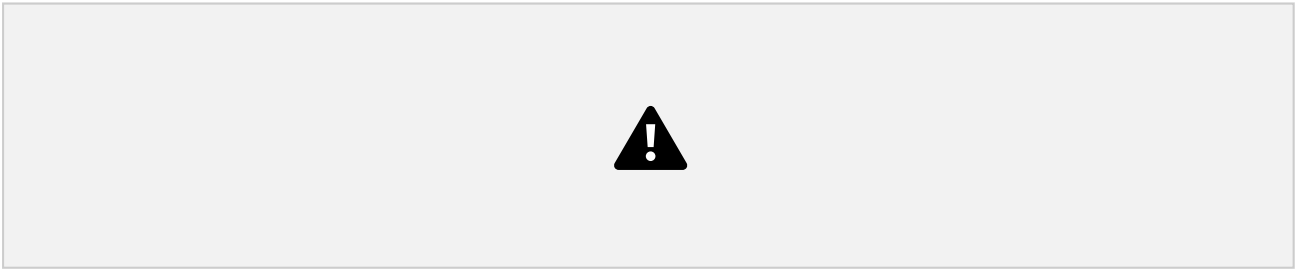
11. Identificação e aplicação de técnicas para minimizar problemas de desbalanceamento (caso não seja necessário, analisar o porquê)

Como podemos ver nas duas imagens abaixo, esse dataset possui mais exemplos ruins do que exemplos bons, fazendo com que ela seja desbalanceada.



Devido a esse fato, optou-se por reduzir as classes majoritárias a fim de deixá-la

equilibrada(under-sampling), portanto a estratégia utilizada foi igualar a quantidade de classes eliminando elementos aleatórios da classe majoritária como podemos ver abaixo:



O dataset gerado chamado **datasetBalanceado** agora possui o mesmo número de classes majoritárias e minoritárias.

12. Limpeza de dados:

a. Identificação e eliminação de ruídos ou outliers

Não foram utilizadas técnicas para eliminação de ruídos e outliers.

b. Identificação e eliminação de dados inconsistentes

Não foram encontrados dados inconsistentes no dataset.

c. Identificação e eliminação de dados redundantes

Para remover os exemplos redundantes podemos utilizar o seguinte comando para remover a coluna de índice para não nos atrapalhar:

Como podemos ver, temos 42 dados duplicados, para removê-los, utilizamos o seguinte comando:

d. Identificação e resolução de dados incompletos (ausentes) – utilização de alguma técnica de preenchimento e justificar

Para identificar os valores faltantes podemos utilizar a seguinte função:

Como pôde-se perceber, todos os resultados foram FALSE, isso significa que o dataset não possui nenhum valor nulo ou ausente.

13. Identificação e conversão dos tipos de dados (caso não seja necessário, analisar o porquê):

Foi considerado que neste dataset não seria necessário o uso de reescala, pois os limites dos

valores dos atributos não são muito discrepantes entre si e nem mesmo uma conversão, pois todos os valores são muito úteis para serem utilizados da forma que estão.

14. Análise e aplicação de alguma técnica para redução de dimensionalidade:

Dentre as técnicas de dimensionalidade que poderiam ser aplicadas a esse dataset, podemos citar a **Agregação**, onde os atributos Acidez Fixa e Acidez Volátil poderiam ser transformados em um novo atributo que classificaria uma Acidez única, podendo chamar apenas Acidez. O mesmo poderia ser feito com o Dióxido de Enxofre Livre e Dióxido de Enxofre Total