**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***Parte 2: Pré-processando os dados***

1. ***Eliminação inicial de atributos***

*Discorrer se fez eliminação de alguns atributos (evidentemente não relevantes), quais foram e por quê*.

Para o conjunto de dados, no estágio de pré-processamento é imprescindível avaliar na amostragem quanto aos dados faltantes/ ausentes e fazer a substituição pela média ou mediana, quando se refere à valores numéricos, com o objetivo de não perder informação relevante ao modelo.Também podemos calcular a moda, quando por exemplo não temos valores numéricos (strings), ou seja, para identificarmos termos de maior frequência/ incidência e incluirmos ou agregarmos aos atributos (features).

Fazendo a análise qualitativa para após elaborar análise quantitativa do conjunto de dados são os estágios de pré-processamento, para termos um correto balanceamento dos objetos, filtrar corretamente, excluir ruídos e a discretização.

Em aprendizado de máquina se ocorrer “super ajustes” de atributos e objetos, poderá acarretar em perda de informação.

Em específico para os dados do desafio KDD BR, a escolha de atributos para análise qualitativa foi executada por julgamento de engenharia, ou seja, foram excluídas as variáveis descritivas e selecionado as variáveis estatísticas.

Sendo estas:

\* `mean.signalX` : evidencia a média aritmética do signalX;

\* `sd.signalX` : evidencia o desvio padrão do signalX;

\* `var.signalX` : evidencia a variância do signalX;

\* `kurtosis.signalX` : evidencia a kurtosis do signalX;

\* `skew.signalX` : evidencia a skewness do signalX;

\* `mean.signalY` : evidencia a média aritmética do signalY;

\* `sd.signalY` : evidencia o desvio padrão do signalY;

\* `var.signalY` : evidencia a variância do signalY;

\* `kurtosis.signalY` : evidencia a kurtosis do signalY;

\* `skew.signalY` : evidencia a skewness do signalY;

\* `L` : revela a quantidade de cluster L;

\* `X` : revela a quantidade de cluster X;

\* `XY` : revela a quantidade de cluster XY;

\* `Y` : revela a quantidade de cluster Y;

\* `mean.silhouette` : evidencia a média da silhouette;

1. ***Limpeza dos dados***

*Verifique a presença de alguma das seguintes características:*

* *Redundância de dados*

<https://stackoverflow.com/questions/38309729/count-unique-values-with-pandas-per-groups/38309823>

# [**Count unique values with pandas per groups [duplicate]**](https://stackoverflow.com/questions/38309729/count-unique-values-with-pandas-per-groups)

df = df.groupby('domain')['ID'].nunique()

print (df)

train = train.groupby(‘ID’)[‘’].nunique()

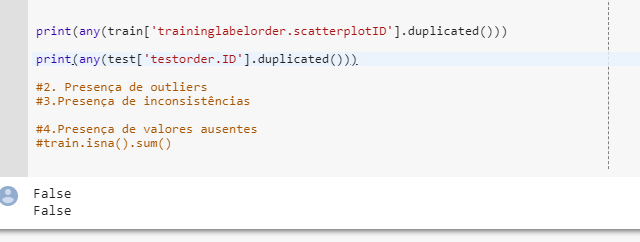
print(train)

df = train.groupby(by='traininglabelorder.scatterplotID', as\_index=False).agg({'Unnamed: 0': pd.Series.nunique})

print(df)

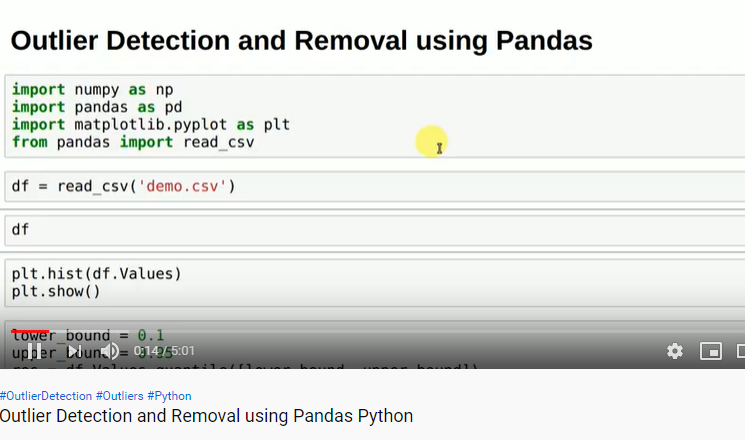
print(any(train['traininglabelorder.scatterplotID'].duplicated()))

## False





* *Presença de outliers*



*Qtde-pontos & count*

*train['qtde.ponto'].value\_counts()*

384 36720

383 4884

382 2973

381 1332

380 1220

192 1124

379 1110

376 1091

375 925

378 914

374 830

377 816

288 655

352 481

362 467

371 463

96 460

373 448

370 426

360 424

368 419

369 408

372 381

356 380

344 356

336 340

269 315

366 308

319 302

359 293

...

171 5

29 5

25 4

45 4

43 4

126 4

30 3

38 3

81 3

84 3

102 3

218 3

37 2

11 2

57 2

50 2

61 2

106 2

174 2

119 2

41 1

152 1

42 1

56 1

70 1

145 1

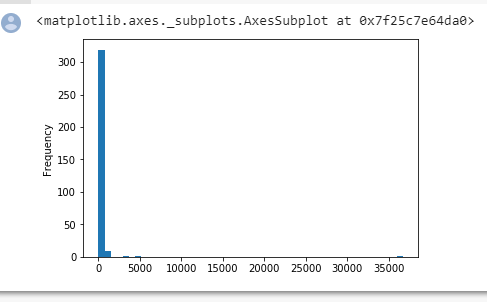
54 1

72 1

14 1

64 1

train['qtde.ponto'].value\_counts().plot.hist(bins=50)



<https://www.kaggle.com/residentmario/indexing-selecting-assigning>

*## reviews.loc[(reviews.country == 'Italy') & (reviews.points >= 90)]*

* *Presença de inconsistências*

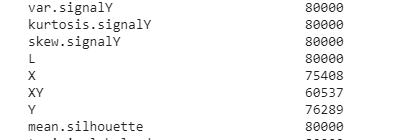
<https://www.kaggle.com/residentmario/indexing-selecting-assigning>

reviews.loc[reviews.price.notnull()]

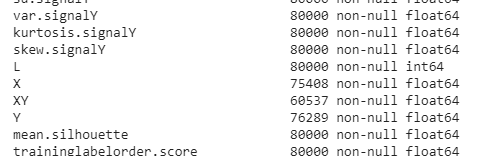
*#3.Presença de inconsistências*

*#Contagem de dados não-nulos*

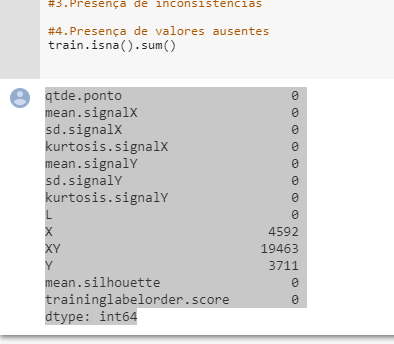
*train.count()*



*train.info()*



* *Presença de valores ausentes*



*train.isna().sum()*

qtde.ponto 0

mean.signalX 0

sd.signalX 0

kurtosis.signalX 0

mean.signalY 0

sd.signalY 0

kurtosis.signalY 0

L 0

X 4592

XY 19463

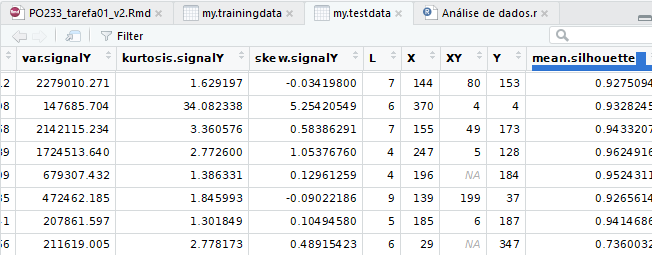
Y 3711

mean.silhouette 0

traininglabelorder.score 0

dtype: int64

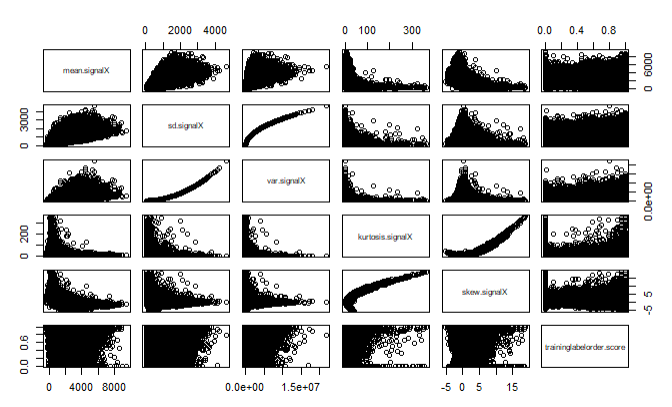
A limpeza de dados foi feita pelos meios sugeridos no enunciado. A redundância de dados não havia de fazer limpeza pois não existia redundância de dados. Já os valores que estavam ausentes existiam pois não havia dados para determinados scatterplot\_ID em alguma das colunas de clusters, podendo ser X, Y, XY ou L. Esses valores ausentes foram substituídos por 0, uma vez que indicam a ausência do grupo ao invés de dado ausente. A imagem a seguir, evidencia a solução do problema.

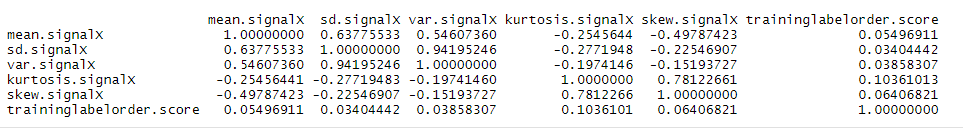
**

1. ***Redução de dimensionalidade***

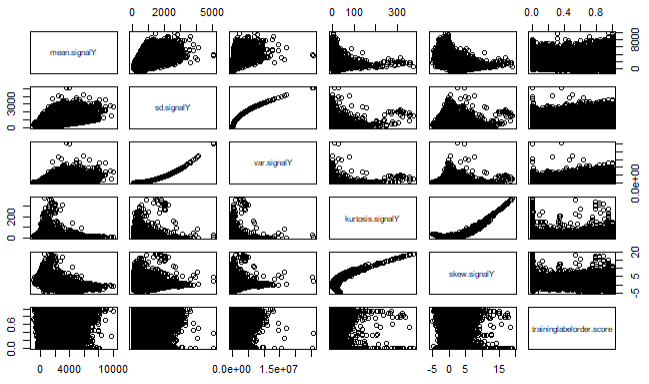
*Aplique alguma técnica de redução de dimensionalidade e comente sobre os resultados obtidos (alguns exemplos de técnicas que podem ser usadas: PCA, seleção de atributos wrapper com algum regressor ou algum método embutido, como LASSO, ou correlação entre variáveis contínuas, etc.).*

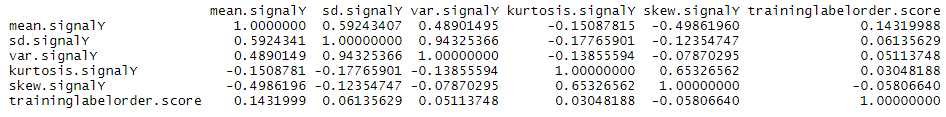
Para redução de dimensionalidade, iniciou-se pela correlação entre variáveis contínuas. Para análise, foram gerados os gráficos de dispersão entre as variáveis segmentadas da seguinte forma: variáveis relacionadas a X, variáveis relacionadas a Y e as demais variáveis



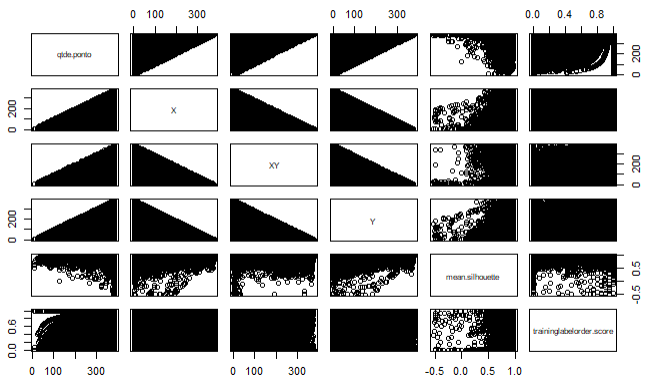


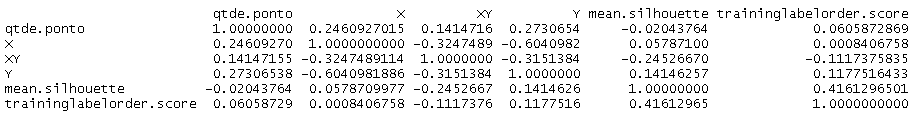
Para as variáveis originadas de X, percebeu-se que existe um alto índice de correlação entre a var.signalx e a sd.signalx de 0,9419, podendo ser excluída uma dessas variáveis. Também existe um alto índice de correlação entre a kurtosis.signalX e Skew.sigmalX de 0,7812, podendo ser excluída alguma dessas variáveis.





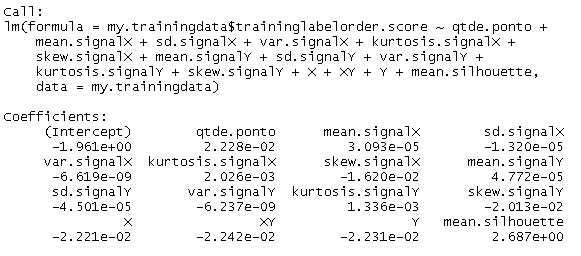
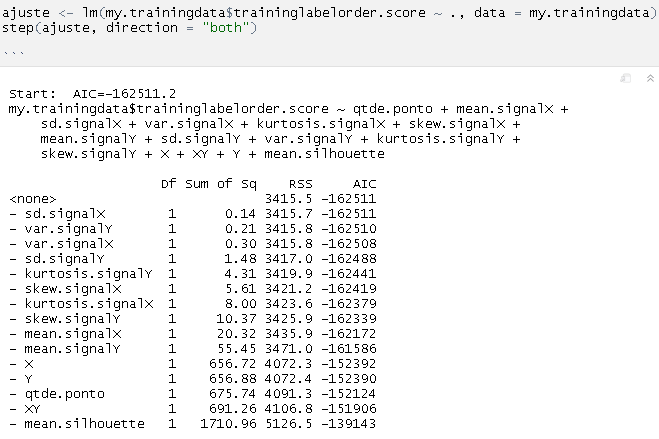
Para a análise das variáveis de Y, notou-se percebeu-se que existe um alto índice de correlação entre a var.signaly e a sd.signaly de 0,9432, podendo ser excluída uma dessas variáveis. Também existe um alto índice de correlação entre a kurtosis.signalY e Skew.sigmalY de 0,6532, podendo ser excluída alguma dessas variáveis.



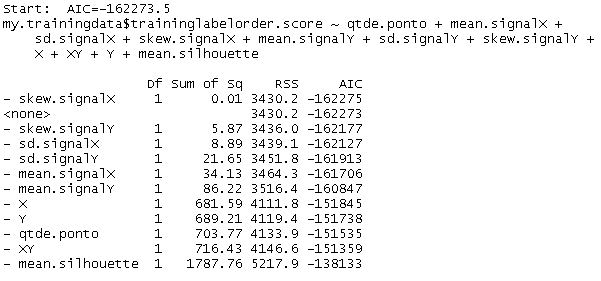


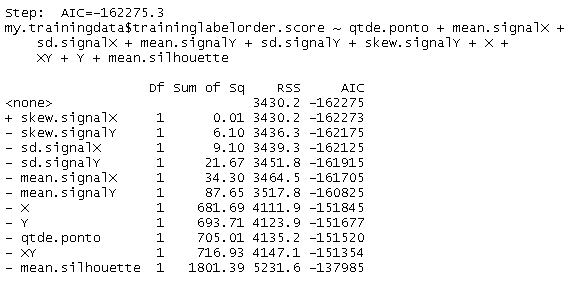
Para as demais variáveis, foi percebido que existe uma dispersão grande das amostras em relação às variáveis.

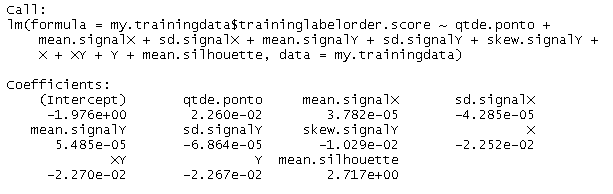
Em seguida, fez-se a seleção de atributo a partir da metodologia filtro e do conceito de busca em ambas as direções (forward e backward). Realizou-se um ajuste linear de todas as variáveis/atributos em relação ao `score`, a fim de perceber a relevância dessa associação entre as variáveis explicativas e variável valor. Obteve-se os resultados seguintes:



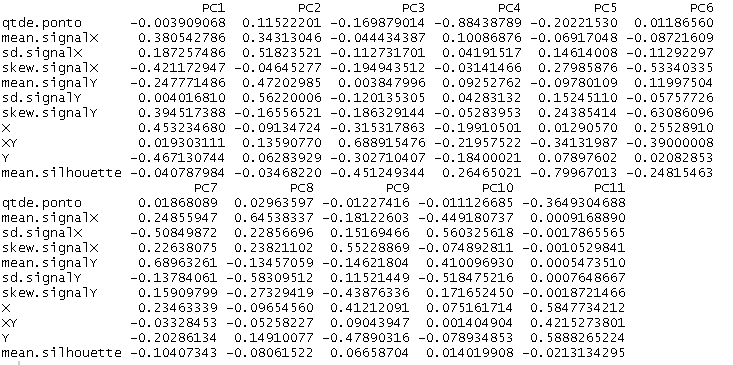
A partir dos resultados, pode-se desprezar as variáveis explicativas `var.signalX`, `kurtosis.signalX`, `var.signalY` e `kurtosis;signalY`. Refez-se a seleção de atributos, sem essas variáveis, para caracterização.

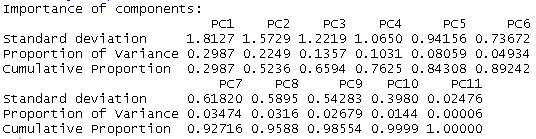


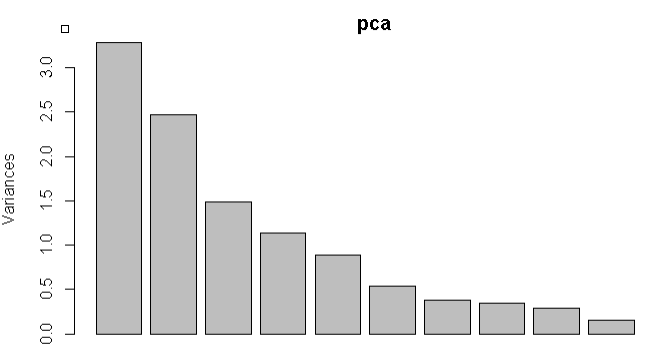




A partir dos atributos restantes, é possível fazer PCA para agregação de atributos e para a redução de atributos.







A partir do PCA, pode-se reduzir a dimensão e trabalhar com a matriz de rotação cedida a partir da função prcomp, além disso, checar as variáveis que apresentam maior peso na caracterização da variável valor.