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

**Parte 1: Entendendo os dados (entrega da parte 1 até 27/08)**

Usando o conjunto de dados disponibilizado na competição KDD-BR 2019 (<https://www.kaggle.com/c/kddbr-2019>), faça o que é pedido nos exercícios a seguir.

**Atenção:** discutam os resultados, a apresentação de tabelas e gráficos sem discussão não é uma análise de dados.

1. **Caracterização básica dos dados**

*Apresentar:*

* *Dimensão dos dados*
* *Número de exemplos em cada conjunto (treinamento e teste)*
* *Tipo e escala de cada atributo (segundo visto em aula)*

Primeiramente, o conjunto de dados disponibilizado em formato 'csv' e possui 3 conjuntos: 'test\_dataset' para os testes; 'training\_dataset' para o treinamento; e 'training\_data\_labels' para os rótulos (função target "Y" para a regressão através do atributo "score"") .

Como faz parte do desafio KDD BR, o estágio inicial deste pré-processamento podemos testar hipóteses e identificar quais são os atributos quantitativos e qualitativos deste cenário. Além da criação e relação de objetos (linhas) e atributos (colunas).

A partir da análise exploratória de dados, constatou-se inicialmente as dimensões do conjunto de dados dispostas da seguinte forma para objetos e atributos, respectivamente:

`6933922` e `7` para o `test\_dataset`;

`27663395` e `7` para o `training\_dataset`;

`80000` e `2` para o `training\_data\_label`.

Diante disso, é necessário atentar-se para a quantidade extensa de dados, sobretudo, para a base de treinamento.

É válido ressaltar também que uma porção desses objetos representa apenas uma regressão, ou seja, grupos de até 384 objetos (datapointID) por gráfico de dispersão (scatterplotID). Sendo o conjunto de dados de rótulo (training\_data\_label) associa-se ao resultado de uma regressão avaliado pela métrica /atributo `score`, cujo objetivo do desafio KDD BR será a criação de um outro modelo que se aproxime deste modelo especialista ‘score’.

As classificações são avaliadas pelo `silhouette` (training\_dataset e test\_dataset) . Para o conjunto de dados de teste, foram notados `20000` exemplos de regressão; enquanto para o conjunto de dados de treinamento, `80000` exemplos após a agregação com o conjunto de dados de rótulo (scatterplotID), tal como para os dados de qualificação do algoritmo de clusterização.

Diante do resumo do formato dos dados, tem-se para cada atributo de entrada:

\* `scatterplotID`: identifica o gráfico de dispersão sendo, pois, qualitativo e escala nominal;

\* `datapointID`: identifica cada ponto dentro de cada gráfico, sendo atributo qualitativo e escala nominal;

\* `sampletype`: categoriza os tipos de amostras, é qualitativo e escala nominal;

\* `signalX`: são valores quantitativos contínuos e escala racional;

\* `signalY`: são valores quantitativos e escala racional;

\* `cluster`: classifica em grupos, é qualitativo e escala nominal;

\* `silhouette`: é uma métrica para a dispersão dos pontos a partir da sua classificação, é quantitativo contínuo limitado e escala racional;

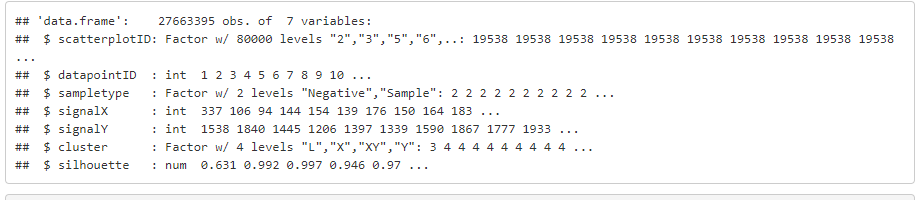
\* `score`: é uma métrica para as regressões realizadas, é quantitativo e escala racional;

A estrutura de cada conjunto de dados:

* Treinamento:

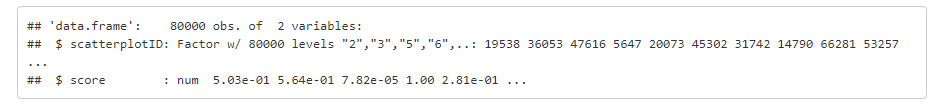
|  |  |
| --- | --- |
| Atributo | Tipo |
| ScatterplotID | Factor |
| DatapointIF | Int |
| Sampletype | Factor |
| SignalX | Int |
| SignalY | Int |
| Cluster | Factor |
| Silhouette | Num |

comando str



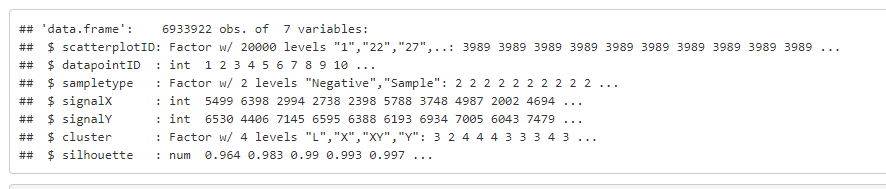
* Label:

|  |  |
| --- | --- |
| Atributo | Tipo |
| ScatterplotID | Factor |
| Score | Num |



* Teste:

|  |  |
| --- | --- |
| Atributo | Tipo |
| ScatterplotID | Factor |
| DatapointID | Int |
| Sampletype | Factor |
| SignalX | Int |
| SignalY | Int |
| Cluster | Factor |
| Silhouette | Num |



1. **Estatísticas descritivas básicas**

*Apresentar estatísticas básicas de cada atributo, para o conjunto de treinamento e o de teste. Comparar as características dos dois conjuntos.*

Na geração de hipóteses (significado e correlação entre as colunas) foram criados alguns atributos contendo a estatística descritiva, cuja intenção é a tentativa de avaliar alguns padrões / características para a validação e associação dos objetos.

Para esta avaliação foram criadas as seguintes informações:

\* `scatterplotID`: identifica o gráfico de dispersão;

\* `qtde.ponto`: evidencia a quantidade de pontos em cada gráfico de dispersão (cada ‘scatterplotID)’;

\* `Negative` : identifica a quantidade de pontos de amostra negativa;

\* `Sample` : identifica a quantidade de pontos de amostra sample;

\* `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;

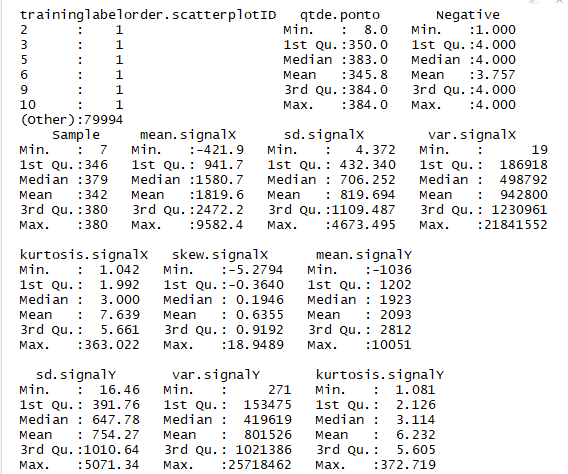
\* `mean.silhouette` : evidencia a média da silhouette;

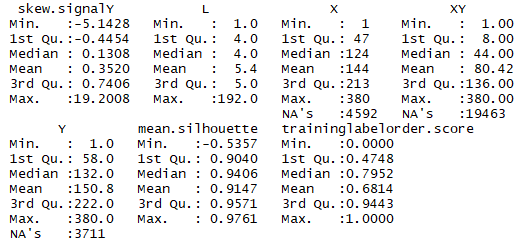
\* `score` : evidencia a pontuação/qualidade das classificações \*\*\* . Cujo atributo será a coluna target de um modelo futuro para a regressão.

É válido ressaltar também que uma porção desses objetos representam apenas uma regressão, ou seja, grupos de pontos para cada scatterplot com até 384 objetos. A empresa Corteva possui um modelo de regressão avaliado pela métrica/ coluna score e as classificações são avaliadas pelo silhouette. Para o conjunto de dados de teste, foram notados 20000 exemplos de regressão; enquanto para o conjunto de dados de treinamento, 80000 exemplos, tal como para os dados de qualificação do algoritmo de clusterização.

Abaixo um resumo estatístico com este novos atributos (commando summary)

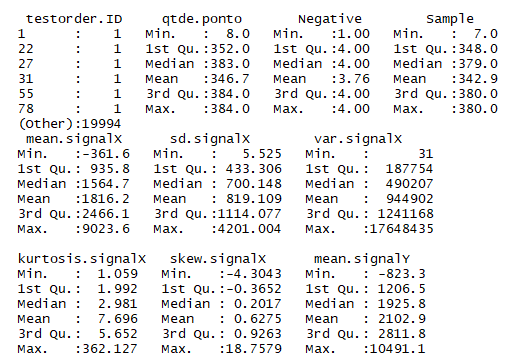
* Treinamento

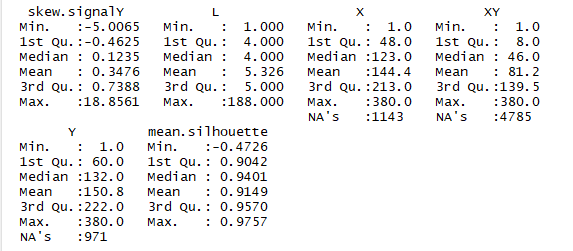




Através deste comando (summary) avaliado o range para valores mínimos e máximos de cada atributo, bem como outras características tais como mediana, primeiro quartil e outras. Porém neste quadro resumo enfatiza-se que alguns clusters possuem valores ausentes para determinada classificação (“NA’s) para os X, XY, e Y, além de visualizarmos poucos pontos para cluster L. E em quase a totalidade de objetos classificados como “sample”.

* Teste



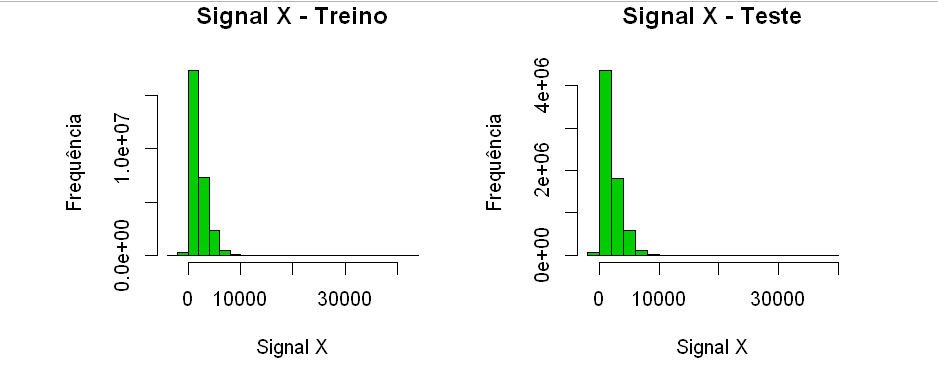


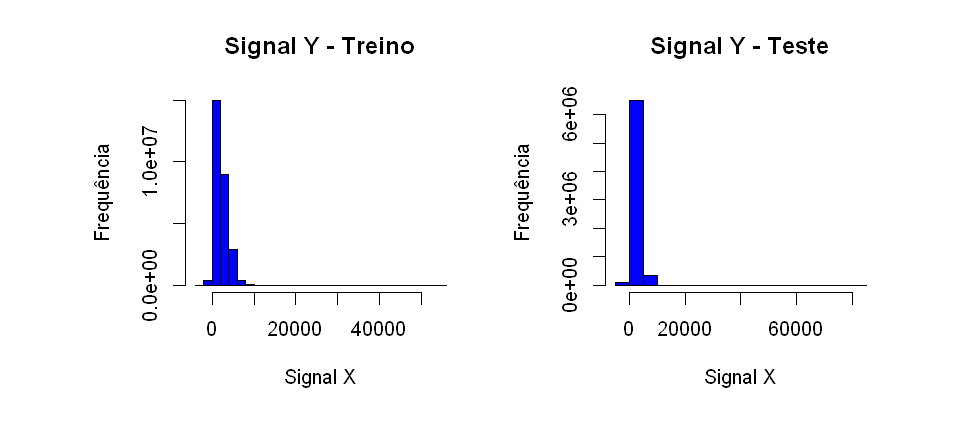
Possui característica bem similar ao conjunto de treinamento, com valores ausentes para determinada classificação (“NA’s) para os clusters X, XY, e Y, além de visualizarmos poucos pontos para cluster L. E em quase a totalidade de objetos classificados como “sample”.

1. **Visualização dos dados**

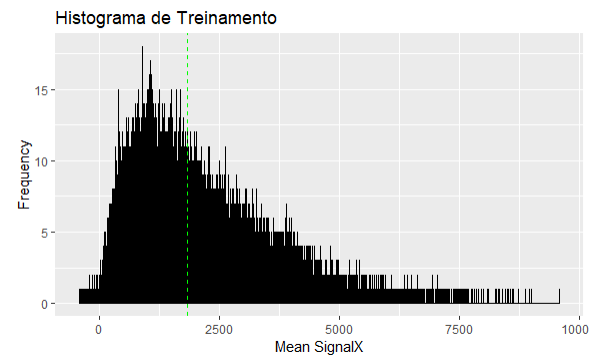
*Apresentar gráficos para visualizar a distribuição dos atributos (dos conjuntos de treinamento e teste, comparando-os).*

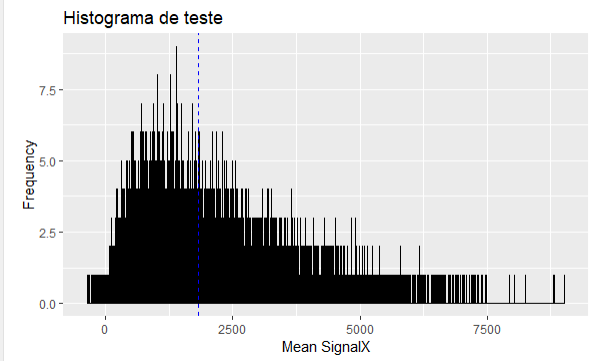
Inicialmente, sem quase nenhum pré-processamento (criação de atributos) alguns gráficos foram elaborados para identificar a frequência e a visualização de um agrupamento estatístico dos clusters.

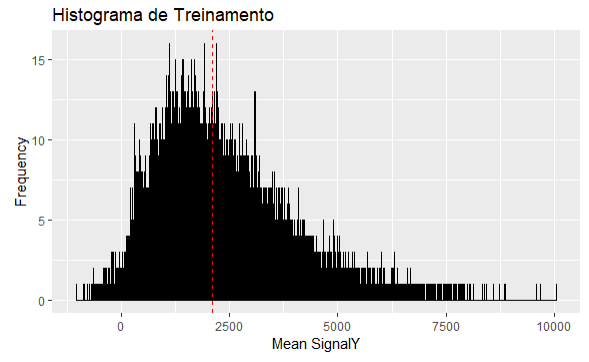


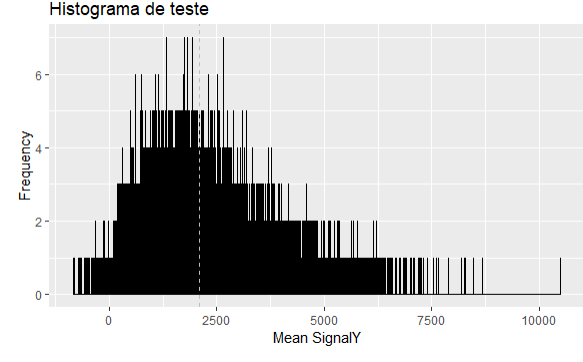


Em comparação do conjunto de dados de treinamento e teste, avaliado o histograma das frequências para os valores médios de signalX e signalY (com a criação de atributos de média, etc por scatterplotID, e a linha tracejada indicado o valor médio geral de cada atributo.

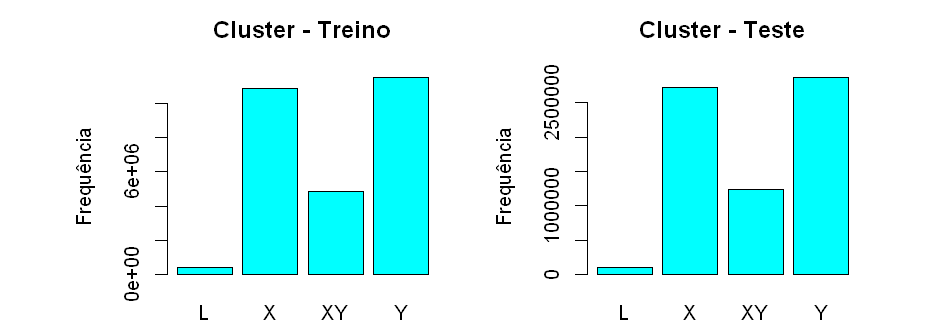


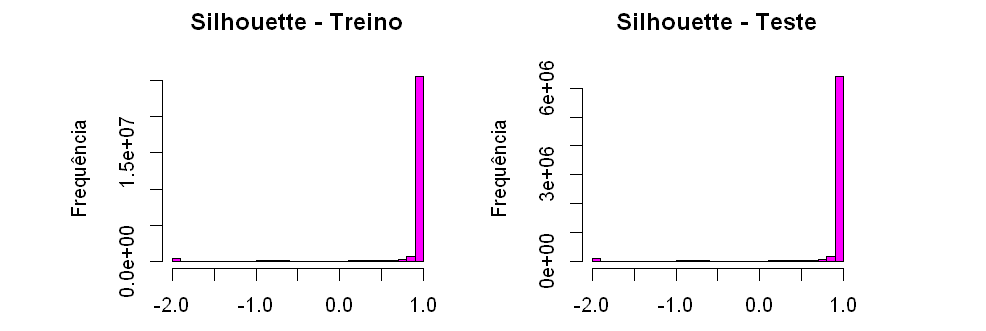




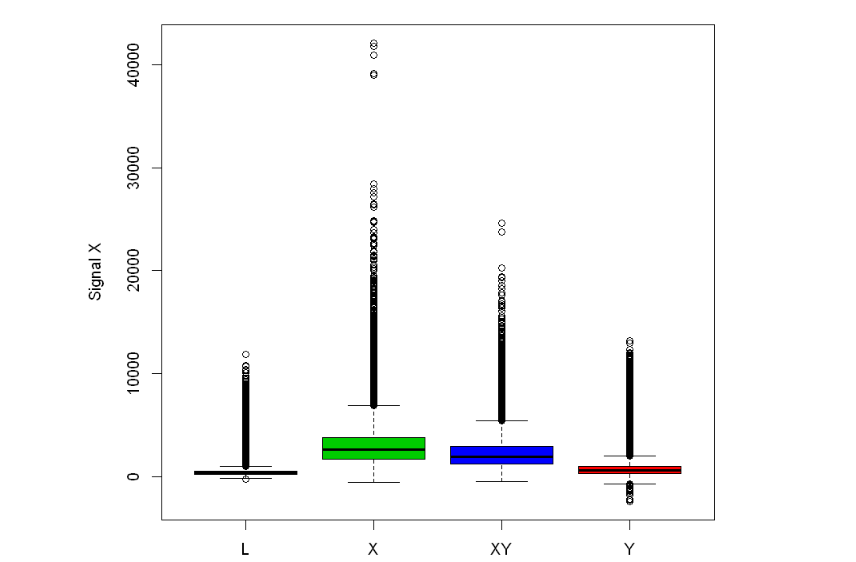


Os histogramas abaixo foram feitos para identificar o conjunto de clusters existentes e comparação de frequência dos mesmos. E também o silhouette.





Boxplot:

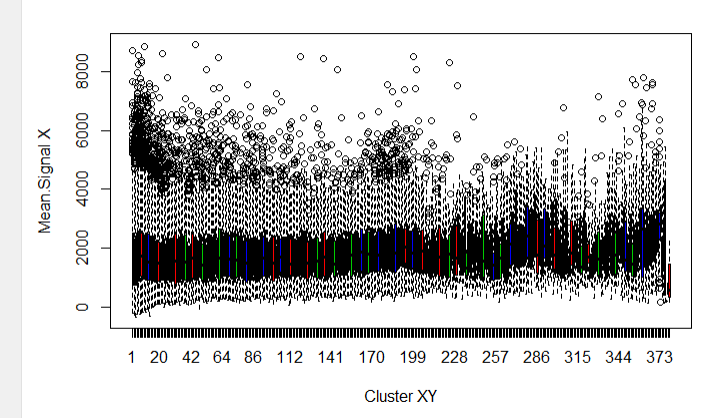


No desafio enfatizado que existe um outro cluster “E” o que não foi visualizado, talvez com a criação de atributos será identificado a criação de mais um cluster.

Além de que o cluster “L” está associado com valores -2, range de silhoutte varia de [-1: 1]

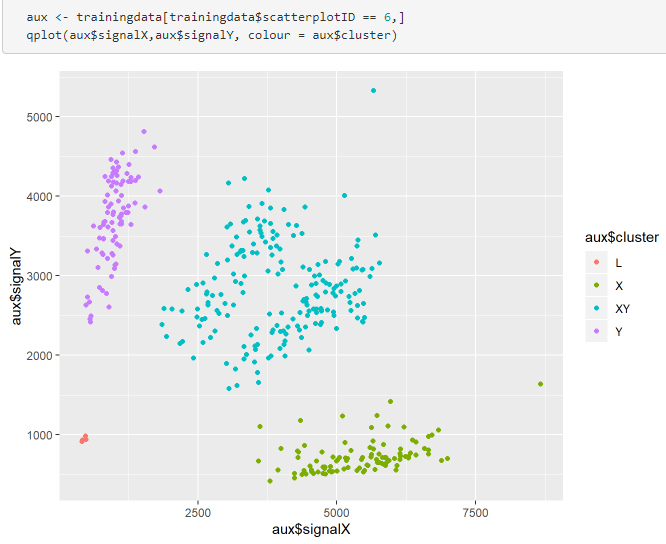
Com o silhoutte podemos tentar associar as similaridades e dissimilaridades destes clusters artificiais versus o que realmente existem, validados pelo “sistema” especialista.

* Avaliando os valores médios de signalX por o cluster XY

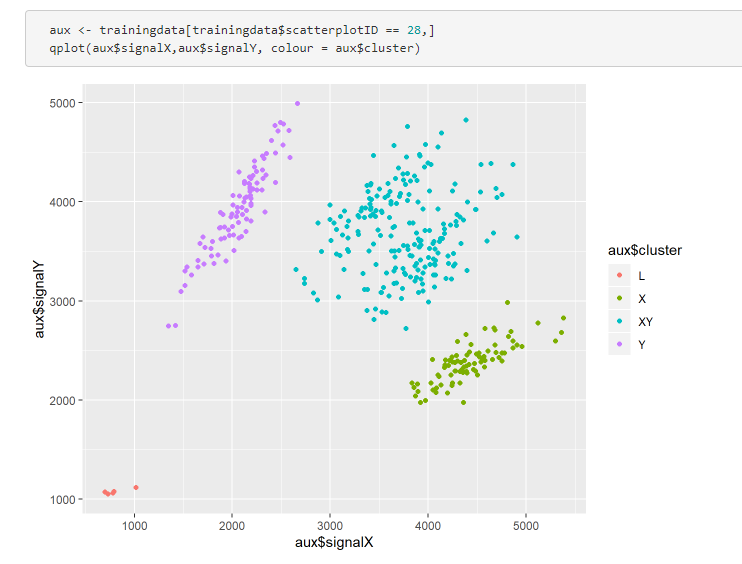


* Abaixo alguns gráficos de dispersão com exemplos de alguns scatterplotIDs.

sctterplotID == 6



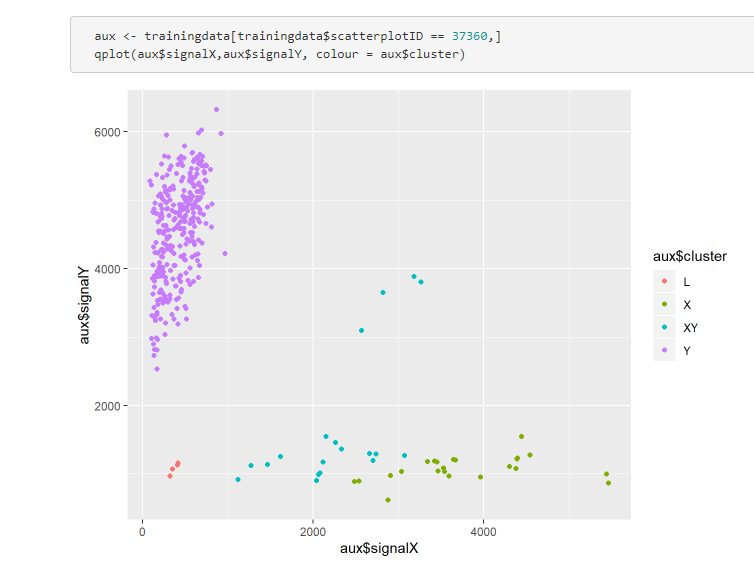
Nota-se esta amostra possui um aglomerado no cluster XY bem definido em comparação com os cluster X e Y, sendo que o L possui poucos datapointID e seus valores em comparação ao SignalY é menor do que 1000 e para Signal X menor que 2500.



Nesta amostra de scatterplotID = 28 possui maior concentração ao cluster XY, porém a dispersão dos clusters parece similar ao do gráfico anterior, onde o cluster L com pontos pouco acima de 1000 para SignalY e SignalX. Clusters Y e X em menor quantidade de datapointIDs.



Nesta amostra de scatterplotID = 37333 possui maior concentração ao cluster X, porém a dispersão dos clusters parece similar ao do gráfico anterior, onde o cluster L com pontos abaixo de 1000 para SignalY e SignalX. Clusters Y e XY em menor quantidade de datapointIDs.



Nesta amostra, scatterplotID == 37360 percebe-se que o cluster XY possui datapointIDs bastante dispersos e sobrepõem-se com o cluster X. Apenas o cluster Y está bem definido.

O valores do cluster Y concentrados acima de 2000 em comparação ao SignalX. O cluster L possui valores similares e com cluster similar em comparação com outros exemplos.