Análise de Componentes Principais e de Clusters

PEDRO LEITE - 201906697

PEDRO CARVALHO - 201906291

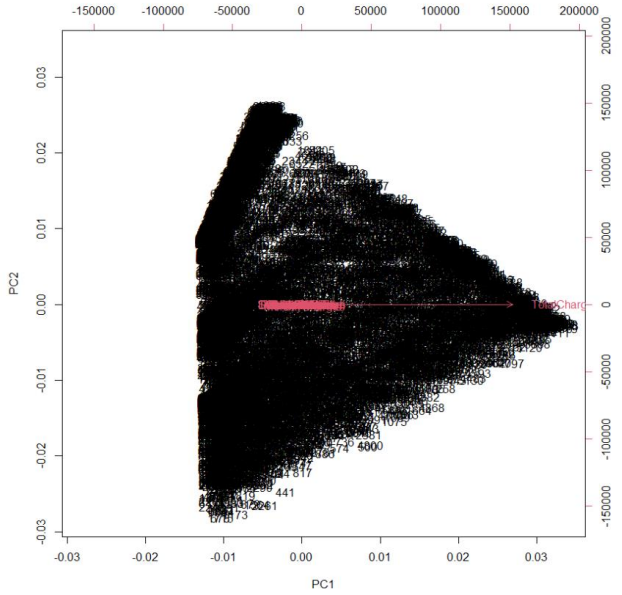
Telco Customer Churn

Focused customer retention programs

7043 clientes e 21 atributos

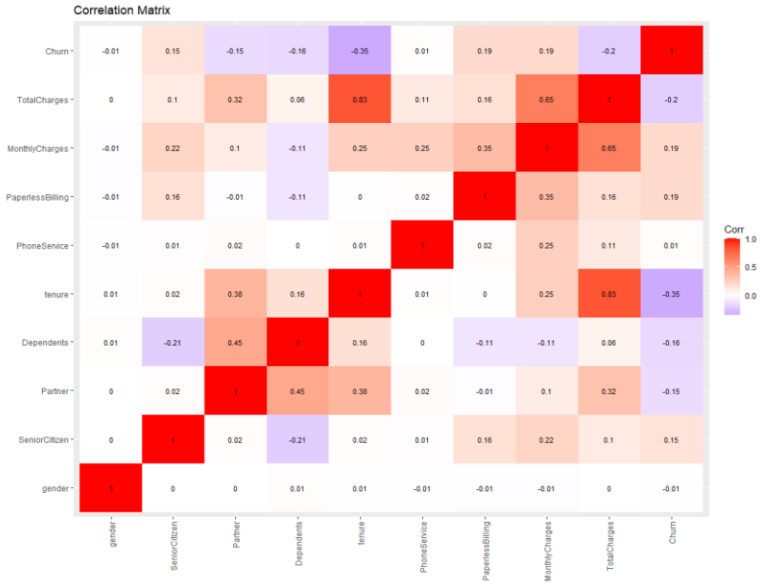
Análise de Componentes Principais (ACP)

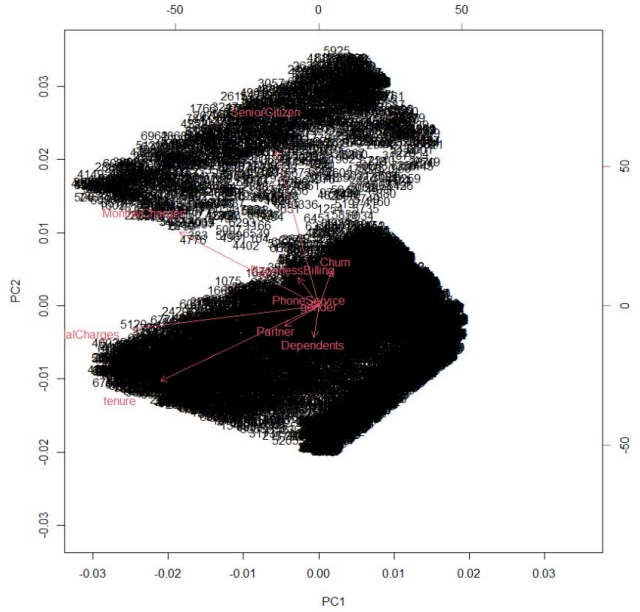
```
> summary(pca)
Importance of components:
                             PC1      PC2     PC3   PC4   PC5    PC6    PC7    PC8 PC9   PC10   PC11   PC12   PC13  PC14
Standard deviation     2266.9469 25.07315 9.25969 1.159 1.064 0.6101 0.5275 0.5077 0.5 0.4475 0.4075 0.3894 0.3741 0.372
Proportion of Variance    0.9999  0.00012 0.00002 0.000 0.000 0.0000 0.0000 0.0000 0.0 0.0000 0.0000 0.0000 0.0000 0.000
Cumulative Proportion     0.9999  0.99998 1.00000 1.000 1.000 1.0000 1.0000 1.0000 1.0 1.0000 1.0000 1.0000 1.0000 1.000
                          PC15   PC16   PC17   PC18   PC19   PC20
Standard deviation     0.3627 0.3522 0.3318 0.3118 0.1929 0.1048
Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
Cumulative Proportion  1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
```



One Hot Encoding a todas os atributos do dataset

Análise de Componentes Principais (ACP)

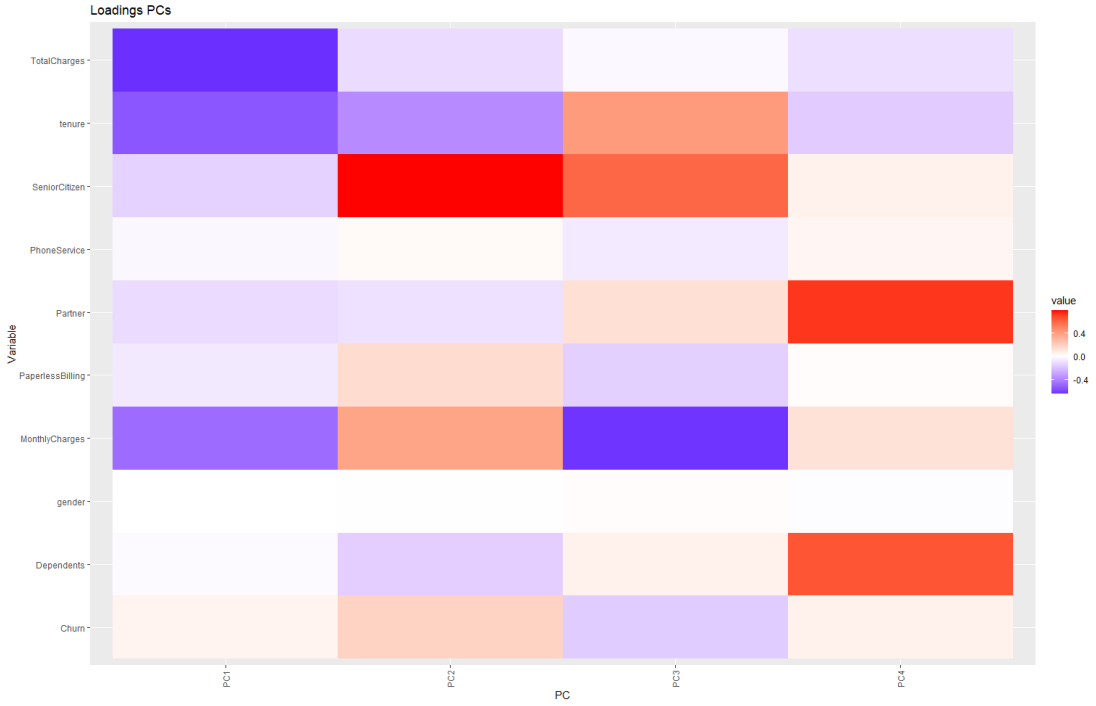




One Hot Encoding a apenas aos atributos categóricos binários do dataset

Análise de Componentes Principais (ACP)

```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     1.5009 1.0630 0.8522 0.52916 0.49996 0.45138 0.37742 0.33272 0.28316 0.23847
Proportion of Variance 0.4305 0.2159 0.1388 0.05351 0.04777 0.03893 0.02722 0.02116 0.01532 0.01087
Cumulative Proportion  0.4305 0.6464 0.7852 0.83874 0.88650 0.92544 0.95266 0.97381 0.98913 1.00000
```



Análise de Clusters

- Utilizamos como medida da distância entre os clusters, a distância euclidiana:

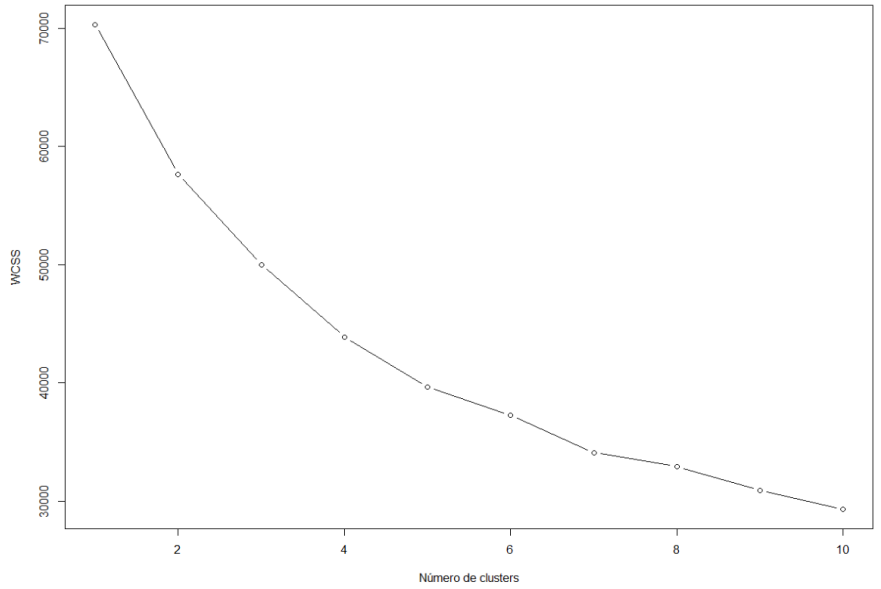
$$\boldsymbol{d(p,q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2}$$

- O método para a escolha da melhor distância para agrupar 2 grupos, foi o método de Ward:

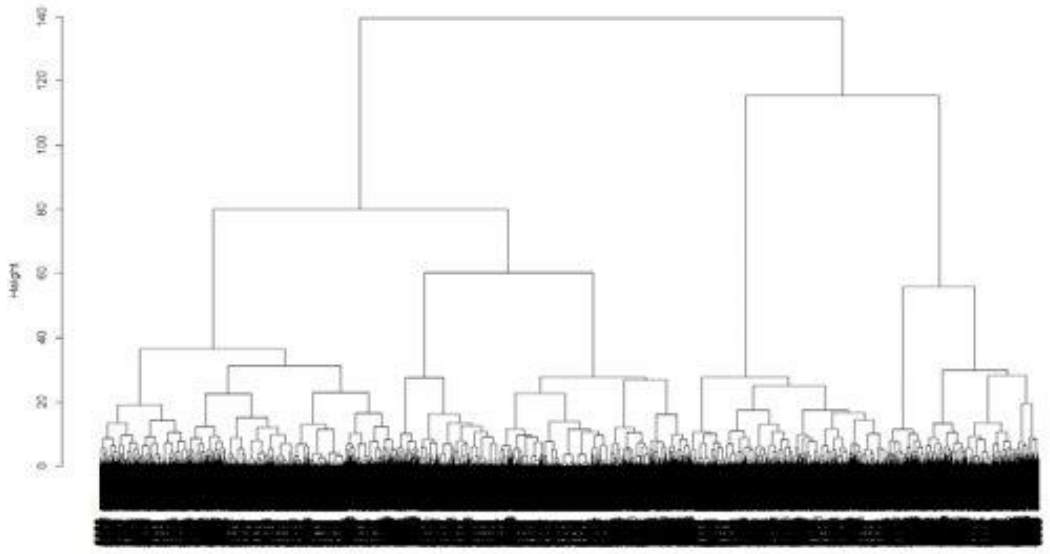
$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B} \|\vec{x}_i - \vec{m}_{A\cup B}\|^2 - \sum_{i\in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i\in B} \|\vec{x}_i - \vec{m}_B\|^2 \\ &= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2\end{aligned}$$

Análise de Clusters

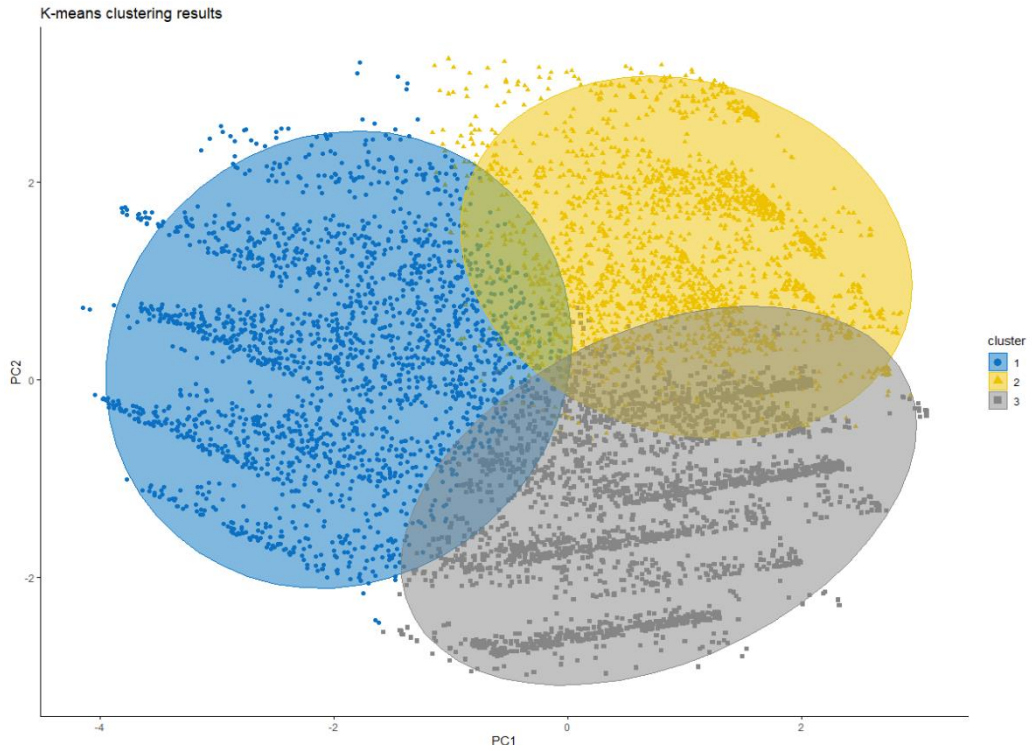
- Fizemos um gráfico que apresenta a relação entre o WCSS (Within Cluster Sum of Squares) e o número de clusters, para determinar o número de clusters ideal:



Análise de Clusters



Análise de Clusters



- CP1 baixo -> Constituído por clientes com o valor elevado de TotalCharges, MonthlyCharges e Tenure.
- CP1 alto e CP2 baixo -> Constituído por poucos SeniorCitiziens com valor baixo de TotalCharges e MonthlyCharges.
- CP1 e CP2 altos -> Constituído por maioritariamente SeniorCitiziens com valor baixo de TotalCharges e MonthlyCharges.

Internal Measure

- Avaliar os clusters.
- Os valores de cada observação do Silhouette Coefficient:
  - -1 quando está no cluster errado;
  - 0 quando está em 2 clusters ao mesmo tempo;
  - 1 quando está no cluster certo.
- Fizemos a média:
  - No K-Means obtemos 0.1762239;
  - No primeiro scatterplot do cluster hierárquico obtemos 0.6489913.