Projeto Final - Métodos Matriciais e Análise de Clusters

Rafael Rocha - FGV - T8 - A56660250

Maio/2021

Contents

Briefing da marca e definição do problema

Ao longos dos últimos 12 anos na cena gastronômica do Rio de Janeiro, a marca ¡Venga! com sua essência fundada na identificação do espírito carioca com a cultura das tapas, incorporou conceitos relacionados com a Espanha e hoje proporciona experiências em três restaurantes na cidade, em Copacabana, Ipanema e Leblon. Em comum entre todas as diferentes propostas por cada um deles, a alegria de servir e proporcionar bons momentos de uma maneira informal e com produtos da mais alta qualidade. Os restaurantes caíram no gosto do público e a refinação nos preparos, no entanto, ainda distinguem as casas, que também primam pelos projetos marcantes de decoração e ambientes condizentes com a proposta da marca. A rede é uma das líderes no setor, com faturamento anual em 2019 acima dos 18 milhões. Ela emprega diretamente no momento aproximadamente 150 pessoas.

A crise gerada pela pandemia do coronavírus e o impacto brutal na estrutura de funcionamento presencial nos restaurantes, colocou todos nós administradores em reflexão sobre possíveis manobras e a inserção e utilização da tecnologia para minimizar os danos. Como gerente operacional da rede diante desse desafio, e mais do que nunca por ser imprescindível fidelizar nossa clientela no delivery, quando esse formato era o único possível com as restrições impostas, sugeri a implantação de um ousado programa de fidelização com excelentes benefícios para os que aderissem, com objetivo principal de garantir regularidade.

Nesse relatório utilizarei dados reais extraídos no cadastro e em pesquisas sobre os nossos serviços prestados no ano passado. Vale lembrar que de julho de 2020 em diante, voltamos a atender presencialmente e ampliamos o programa em todas as casas.

Objetivo

Com o uso de clusterização, almejo segmentar os **312 clientes** que aderiram ao programa e separá-los em grupos com características parecidas. Com o resultado, espero gerar insights e aumentar a efetividade das campanhas de marketing e chances de conversão.

Este relatório seguirá os seguintes passos:

1. Importação do dataset Pesquisa_Clientes.csv para o ambiente R;
2. Exploração detalhada com uso de visualizações e descrições estatísticas para descobrir fatos e tendências sobre o nosso conjunto de dados;
3. Utilização dos seguintes recursos gráficos para geração de *insights*:
    - Boxplots agrupados por sexo e análises;
    - Histogramas para visualização das distribuições das variáveis;
4. Aplicação de técnicas de agrupamento de dados utilizando os algoritmos K-means e de clusterização hierárquica;
5. Ao final apresentarei a validação dos resultados, conclusão e sugestões;

Carregando os pacotes

```
library(dplyr)
library(tidyverse)
library(ggplot2)
library(caret)
library(gdata)
library(ggplot2)
library(cowplot)
library(dendextend)
library(NbClust)
library(factoextra)
library(hrbrthemes)
library(cluster)
library(DataExplorer)
library(psych)
library(proxy)
library(ClusterR)
```

Importando os dados da base: Pesquisa_Clientes.csv

```
baseclientes <- read_csv(
  file="/Users/rafaeldesouza/Desktop/MetodosMatriciaisAnaliseDeClusters/Pesquisa_Clientes.csv")
```

Aplicando ajustes na nomenclatura das variáveis

```
baseclientes <- rename(baseclientes,c('MediaSalarial'='Media Salarial Mensal (R$)'))
baseclientes <- rename(baseclientes,c('PontuacaoGasto'='Pontuacao gasto (1-100)'))
baseclientes <- rename(baseclientes,c('Frequencia'='Frequencia'))
baseclientes <- rename(baseclientes,c('NivelSatisfacao'='Nivel Satisfacao (1-5)'))
```

Descrição dos dados

Cliente ID - Id do cliente cadastrado no programa de fidelização implantado em março de 2020;

Sexo - Sexo;

Idade - Idade;

Media Salarial - Média salarial declarada no cadastro do programa de fidelização em reais (mil por mês);

Pontuacao Gasto - De 1 a 100, onde cada R$100 gastos em qualquer um dos três restaurantes da rede, durante o período de março a dezembro de 2020 (consumos no salão e delivery direto com as lojas) o cliente acumulou 1 ponto;

Frequencia - Número total de visitas em qualquer um dos restaurantes + quantidade de pedidos de entrega (direta) no período em 2020 de vigência do programa;

Nivel Satisfacao - Nível de satisfação com o serviço prestado (salão e delivery) em 2020 de 1 a 5 - esses dados foram coletados em dezembro de 2020, em uma pesquisa, e adicionados a base.

Análise exploratória

```
head(baseclientes)
```

```
## # A tibble: 6 x 7
##   `Cliente ID` Sexo  Idade MediaSalarial PontuacaoGasto Frequencia
##          <dbl> <chr> <dbl>         <dbl>          <dbl>      <dbl>
## 1            1 M        19             2              9          6
## 2            2 M        21             6             81         22
## 3            3 F        20             4              6          4
## 4            4 F        23             4             77         10
## 5            5 F        51            17             80         16
## 6            6 F        22             8             76         19
## # ... with 1 more variable: NivelSatisfacao <dbl>
```

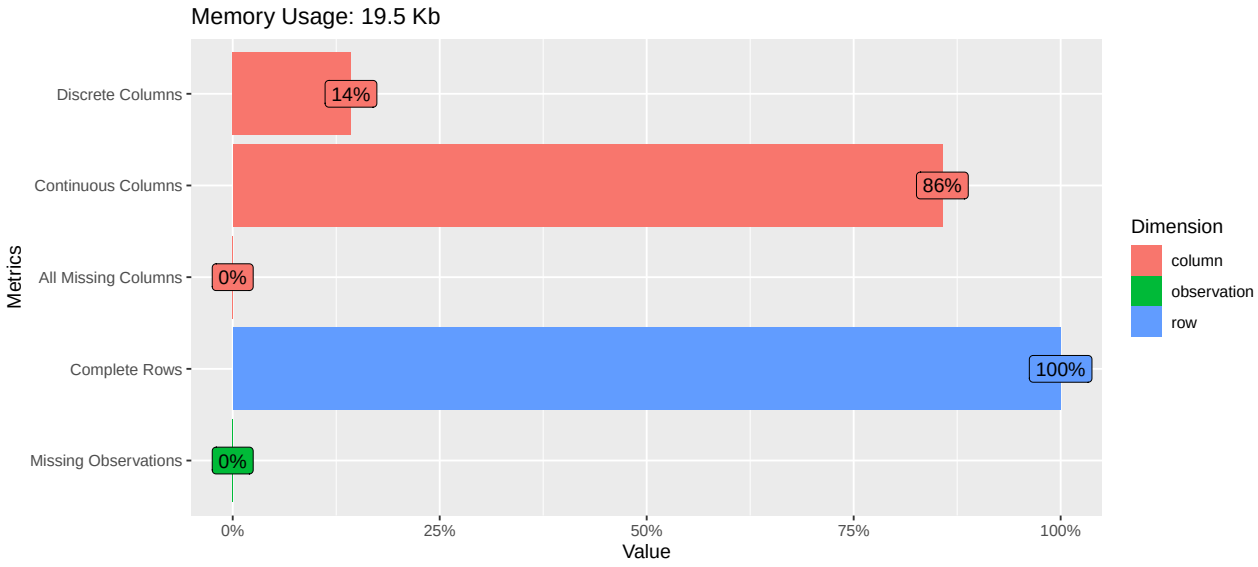
```
summary(baseclientes)
```

```
##    Cliente ID         Sexo                Idade       MediaSalarial
##  Min.   :  1.00   Length:312         Min.   :18.00   Min.   : 1.000
##  1st Qu.: 78.75   Class :character   1st Qu.:27.00   1st Qu.: 4.000
##  Median :156.50   Mode  :character   Median :35.00   Median : 6.000
##  Mean   :156.50                      Mean   :38.55   Mean   : 7.221
##  3rd Qu.:234.25                      3rd Qu.:49.00   3rd Qu.: 8.000
##  Max.   :312.00                      Max.   :83.00   Max.   :35.000
##  PontuacaoGasto     Frequencia     NivelSatisfacao
##  Min.   :  1.00   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 25.75   1st Qu.: 5.000   1st Qu.:3.000
##  Median : 50.00   Median : 7.000   Median :4.000
##  Mean   : 49.73   Mean   : 9.378   Mean   :3.766
##  3rd Qu.: 75.00   3rd Qu.:13.250   3rd Qu.:5.000
##  Max.   :100.00   Max.   :44.000   Max.   :5.000
```

```
introduce(baseclientes)
```

```
## # A tibble: 1 x 9
##    rows columns discrete_columns continuous_columns all_missing_columns
##   <int>   <int>            <int>              <int>               <int>
## 1   312       7                1                  6                   0
## # ... with 4 more variables: total_missing_values <int>, complete_rows <int>,
## #   total_observations <int>, memory_usage <dbl>
```

```
plot_intro(baseclientes)
```



Não foram identificados valores faltantes.

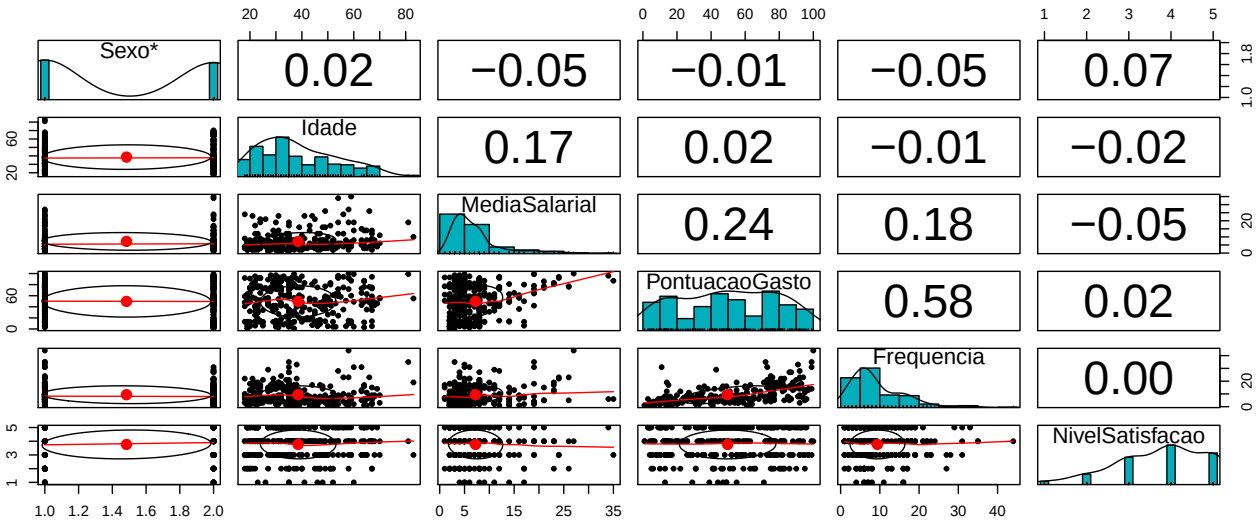
Visualizações e descrições estatísticas

```
plot_correlation(baseclientes_cleaned)
```

A função acima nos proporcinou uma visualização gráfica de todos os pares e a magnitude de suas correlações. Nesse caso podemos observar que as variáveis: Media Salarial, Pontuacao de Gasto e Frequencia são as que apresentam mais relações entre si.

```
pairs.panels(baseclientes_cleaned,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
```

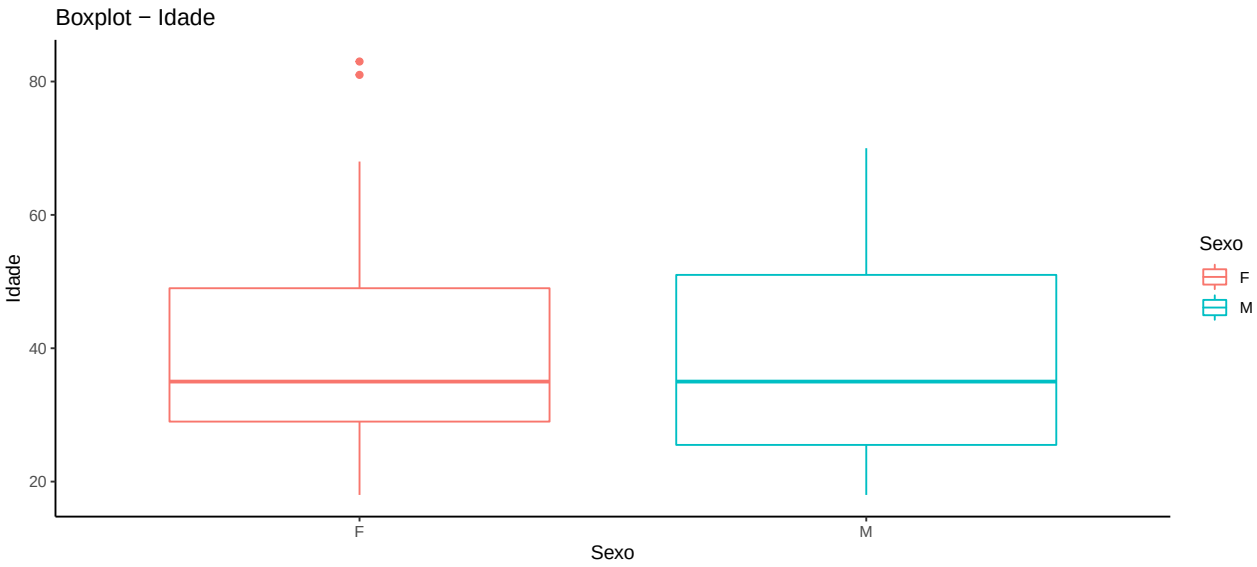
```
            density = TRUE,   # show density plots
            ellipses = TRUE # show correlation ellipses
)
```



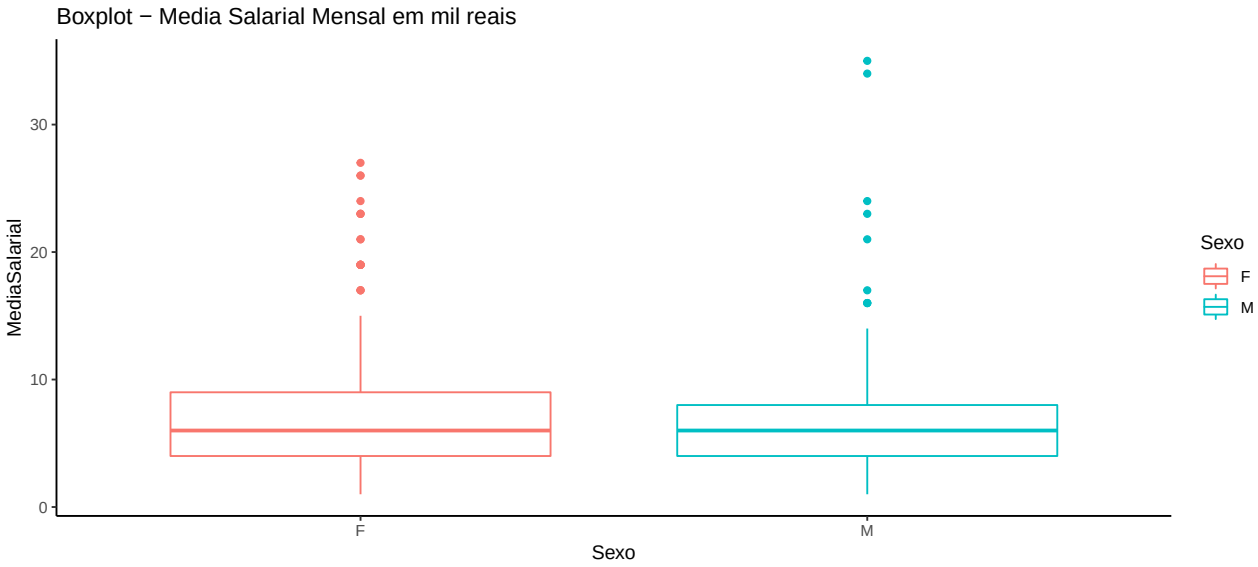
O uso da função 'pairs.panels' nos apresentou um *overview* inicial dos nossos dados. Com gráficos de dispersão, histogramas e a correlação nos confirmando as informações sugeridas no plot anterior de que as variáveis: Media Salarial, Pontuacao de Gasto e Frequencia estão diretamente relacionadas.

Boxplots divididos por sexo

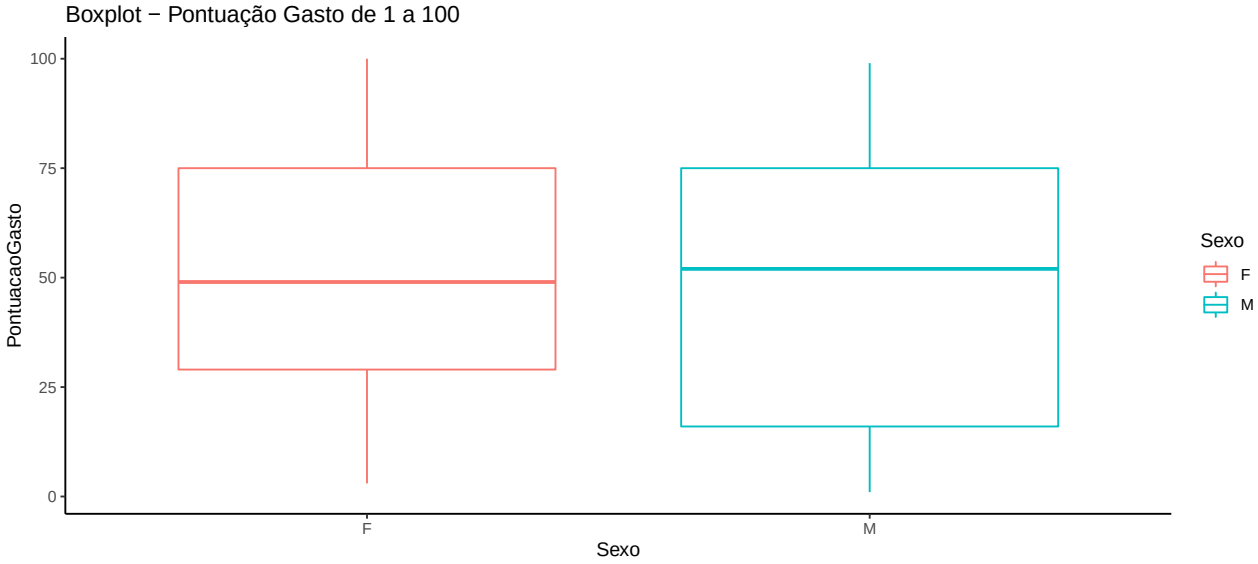
O boxplot nos fornecerá uma análise visual da posição, dispersão, simetria, caudas e valores discrepantes (outliers) do nosso conjunto de dados.



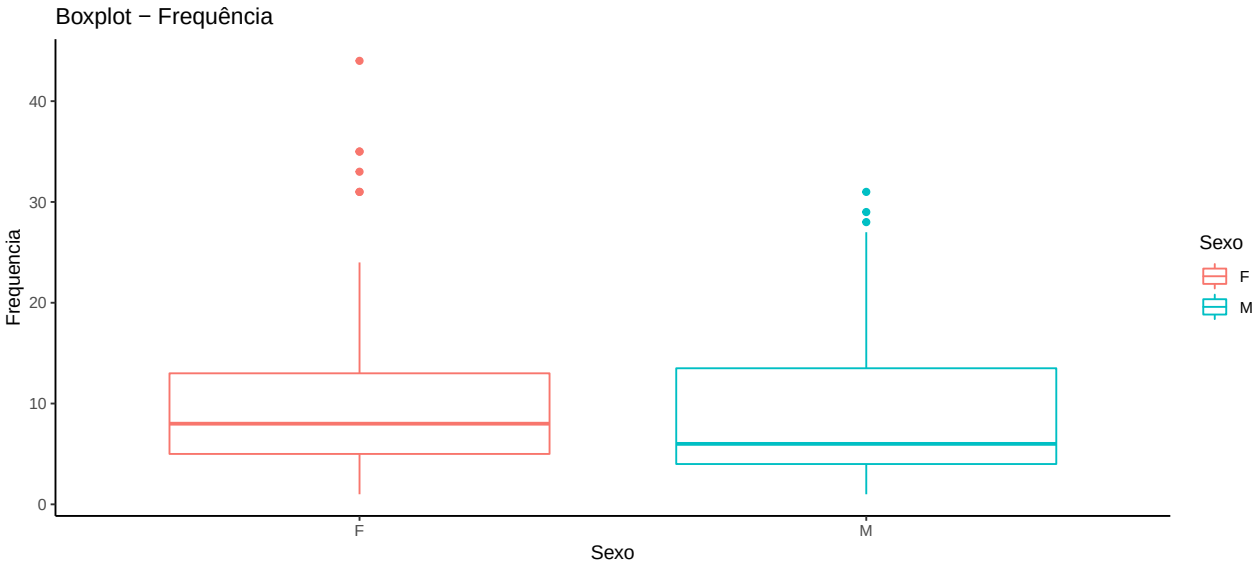
Notamos no primeiro quartil, homens mais novos de aproxidamamente 25 anos e uma maior amplitude entre o primeiro e o terceiro quartil, enquanto no caso das mulheres o primeiro quartil indica idades aproximadas aos 30 anos. As medianas, tanto do sexo feminino quanto do masculino, se encontram por volta dos 35 anos. Os pontos vermelhos sinalizam a ocorrência de outliers, mulheres com mais de 80 anos. O público masculino não passa dos 70 anos.



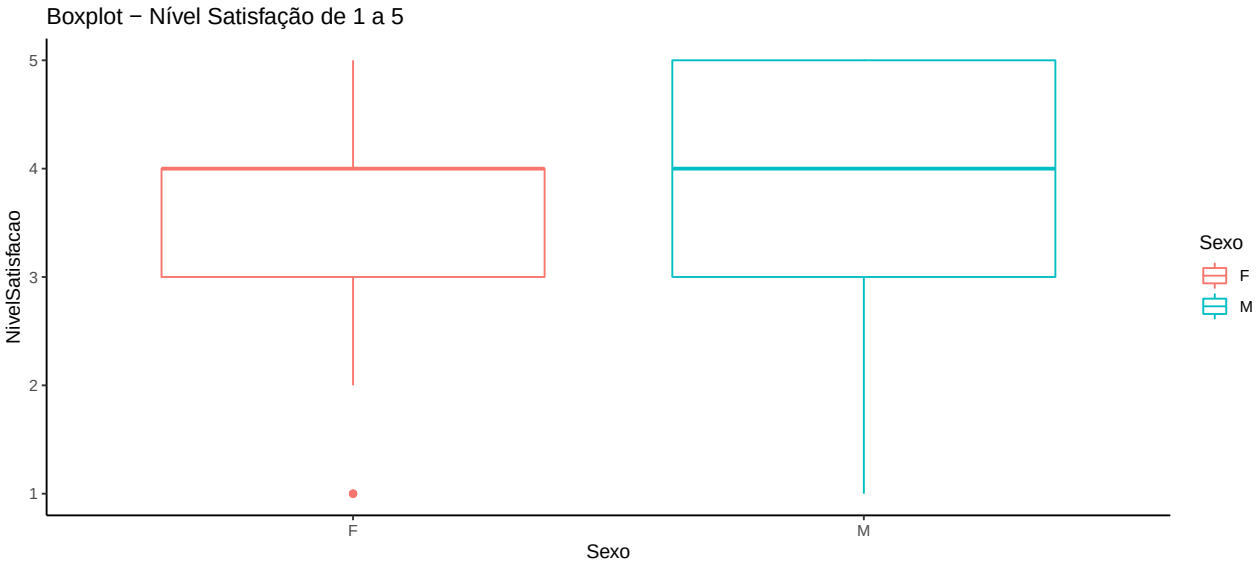
Os pontos indicam salários mensais de alguns clientes do sexo masculino que aderiram ao programa de até R$ 35 mil reais. As medianas nos dois casos está nos R$ 6 mil reais.



No boxplot acima podemos concluir que as pontuações de gastos do sexo masculino apresentam maior variabilidade que as do sexo feminino.



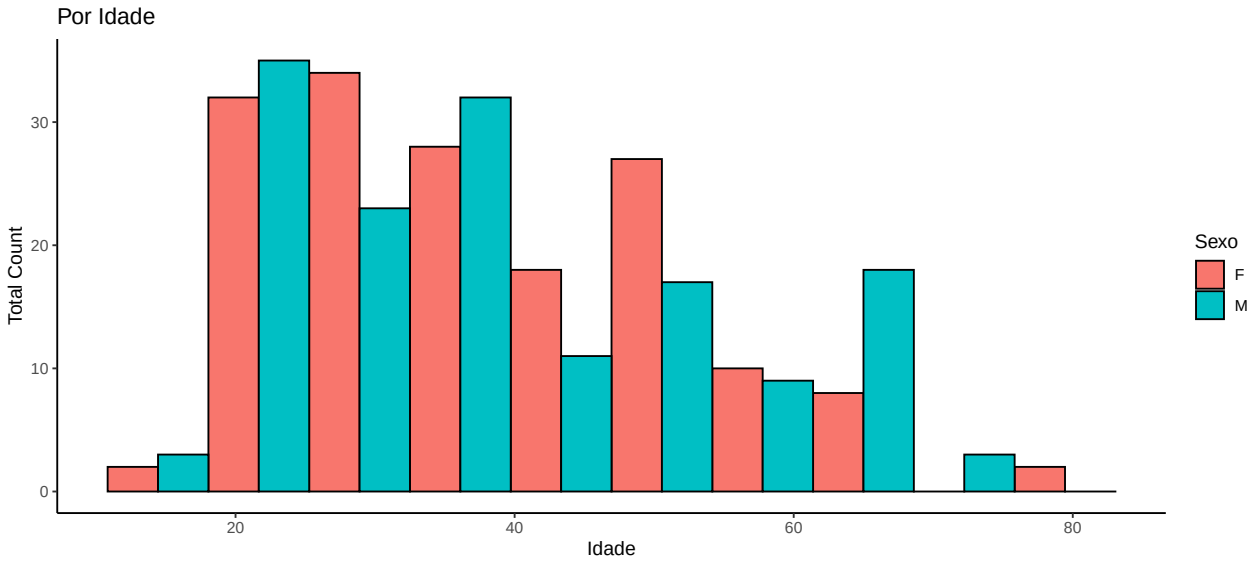
As frequências (visitas aos restaurantes + pedidos de delivery) das mulheres apresenta maior dispersão, com um pico de 44 ocorrências. As medianas também apresentam números bem diferentes, indicando um comportamento específico relacionado ao sexo dos clientes.

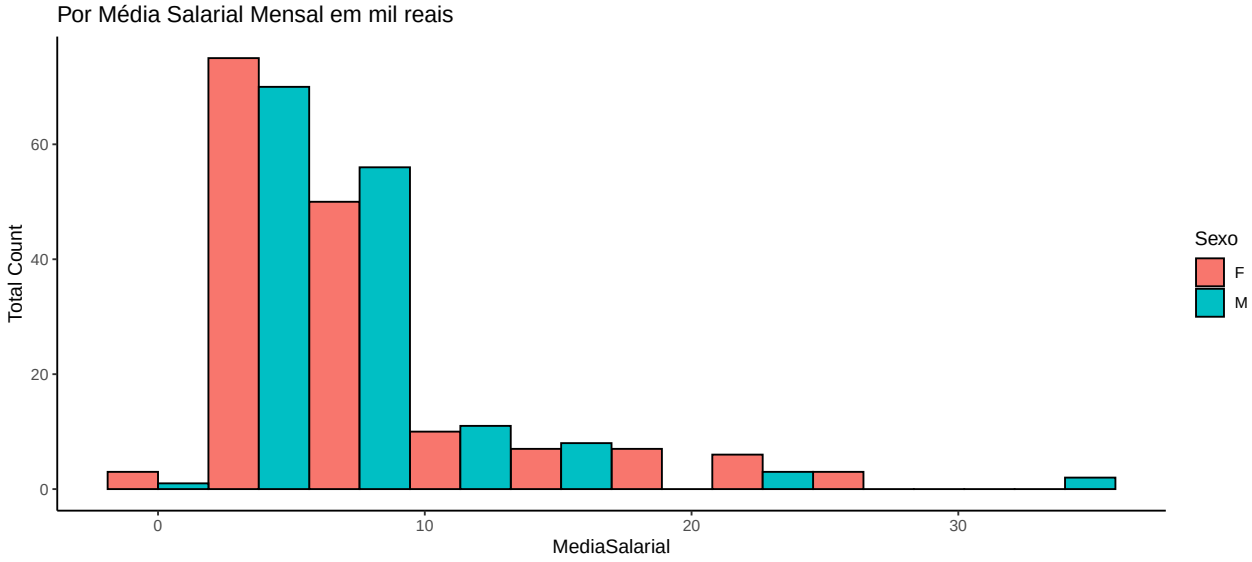


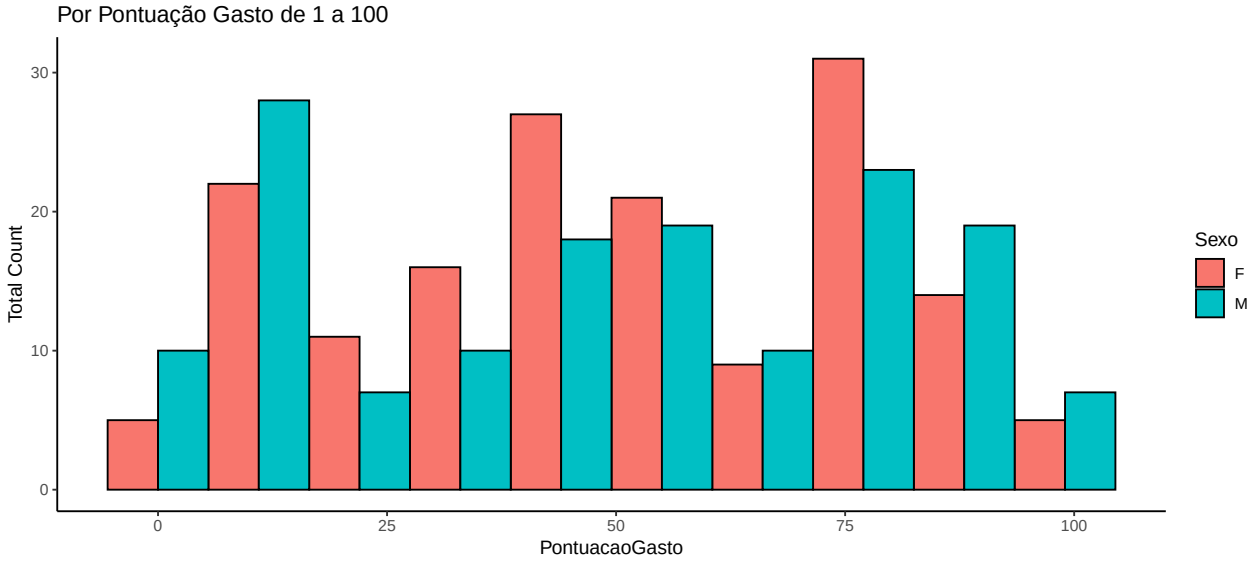
Mais uma vez podemos afirmar que os níveis de satisfação de clientes do sexo masculino apresentam maior variabilidade que as do sexo feminino. Em ambos os casos a mediana se encontra no nível 4.

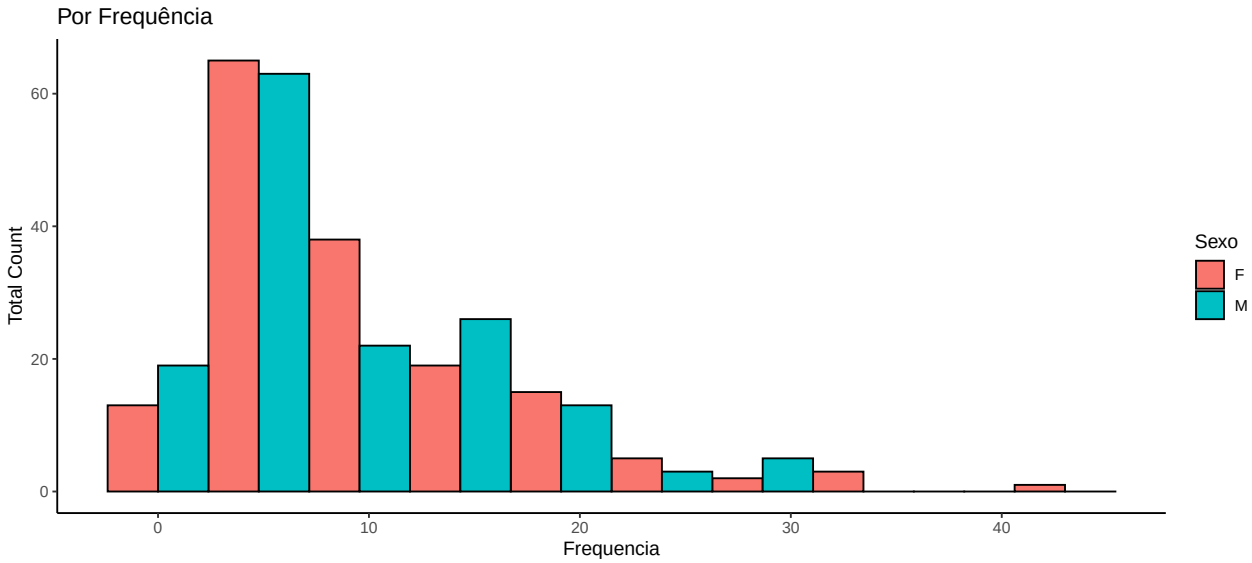
Visualisando as distribuições

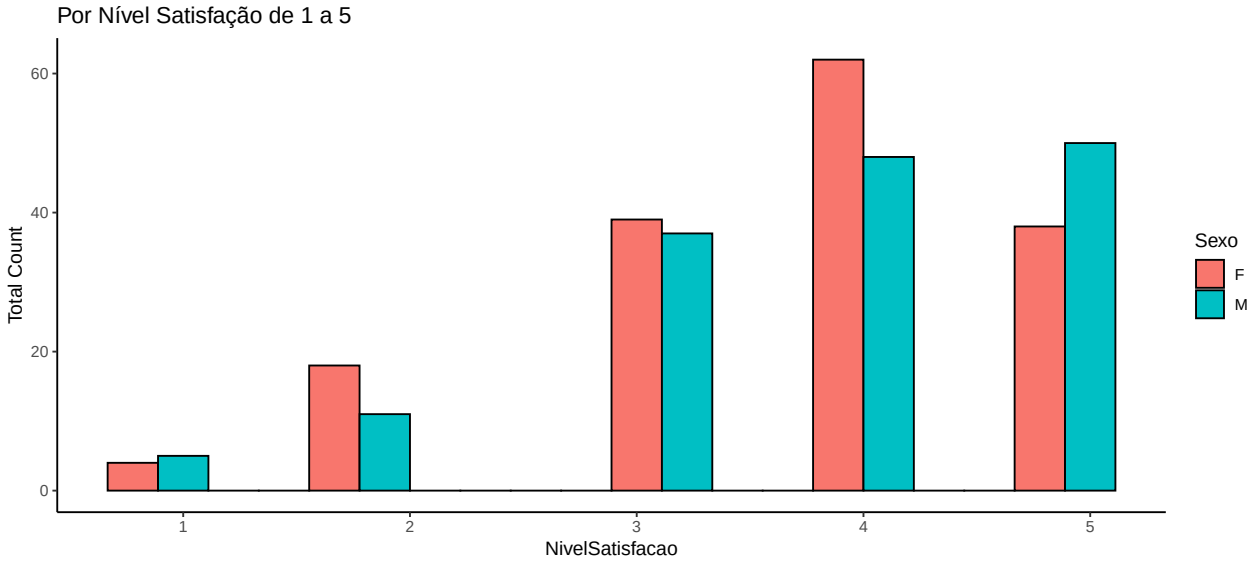
Abaixo apenas algumas visualisações alternativas aos boxplots demonstrados anteriormente.


Por Idade
Total Count
0
10
20
30
20
40
60
80
Idade
Sexo
F
M


Por Média Salarial Mensal em mil reais
Total Count
0
20
40
60
0
10
20
30
MediaSalarial
Sexo
F
M


Por Pontuação Gasto de 1 a 100
Total Count
0
10
20
30
0
25
50
75
100
PontuacaoGasto
Sexo
F
M





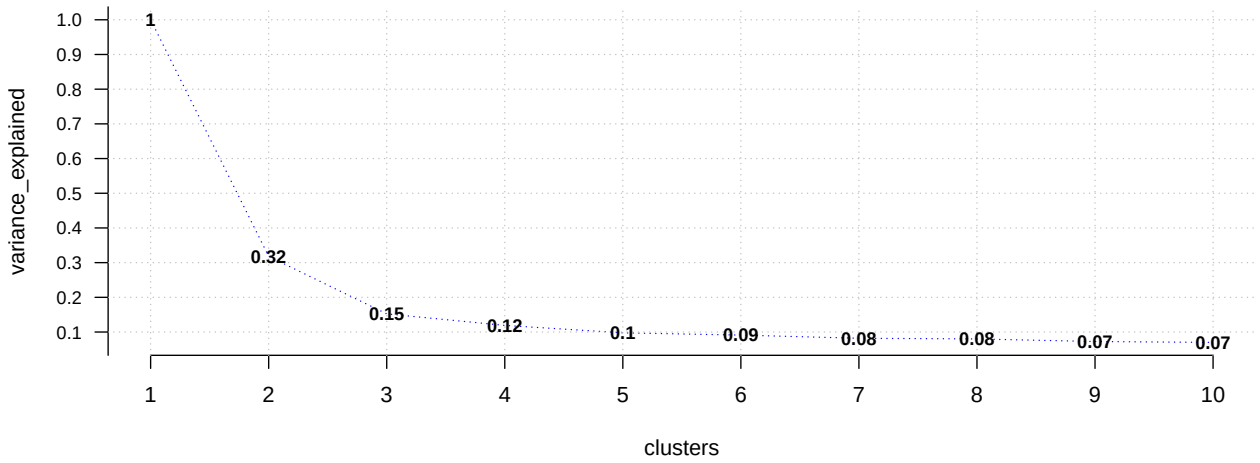
Agrupamento de dados utilizando K-means

Usarei o algoritmo k-means, com seu método que identifica o número 'k' de centros de um cluster, conhecidos como centróides, alocando as observações para o cluster mais próximo. Anteriormente neste relatório vimos que as variáveis: Media Salarial, Pontuacao de Gasto e Frequencia são as que mais influenciam o comportamento do cliente ¡Venga!, portando geraremos os clusters com base nessas variáveis.

Método do cotovelo

O método do cotovelo ou *Elbow Curve* é uma das técnicas mais usadas para sugerir a quantidade ideal de clusters no conjunto de dados.

```
# Uso das variáveis: Media Salarial, Pontuacao Gasto e Frequencia
opt <- Optimal_Clusters_KMeans(baseclientes[, 4:6], max_clusters = 10, plot_clusters = T)
```

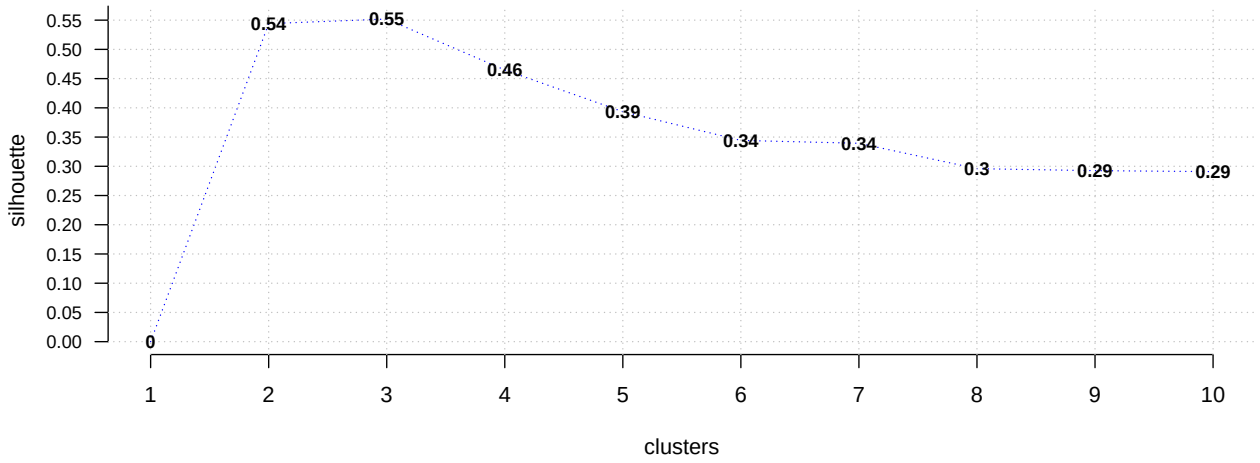


Com o aumento do número de clusters, a soma das distâncias quadráticas intra clusters diminui, quando a diferença entre a distância é quase insignificante temos o valor ótimo de 'k'. No nosso caso, como ilustrado acima, esse valor seria igual a 3.

Coeficiente de silhueta

Vamos aplicar outra técnica conhecida como silhueta ou *silhouette method* para identificar o número de clusters. Ela mede a semelhança de um objeto com seu próprio cluster e comparação com outros clusters. Vou representar graficamente o valor 'silhouette' médio para 'k', variando de 2 a 10. O valor mais alto deve nos ajudar a determinar o número ideal de clusters para melhor dividir nossa base de clientes ¡Venga! que participaram do programa.

```
# Uso das variáveis: Media Salarial, Pontuacao Gasto e Frequencia
opt <- Optimal_Clusters_KMeans(baseclientes[, 4:6], max_clusters = 10, plot_clusters = T, criterion = '
```



Na figura acima, se nota que o maior valor médio (0,55) está presente para k=3, portanto devemos optar por 3 clusters para agrupar os clientes que aderiram ao programa de fidelização. Na próxima etapa, adicionarei esse número 'ótimo' de clusters sugerido no *dataframe*, e em seguida plotarei as observações já agrupadas com ggplot.

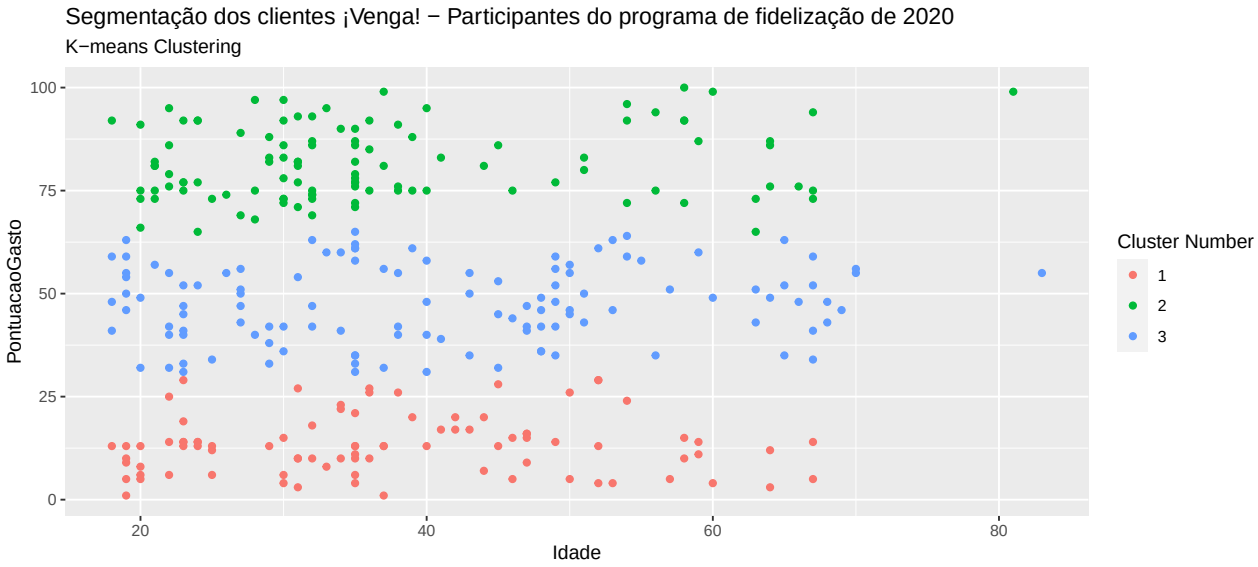
```
set.seed(22)
km <- kmeans(baseclientes[,4:6], 3, nstart=25)
print(km)
```

```
## K-means clustering with 3 clusters of sizes 87, 107, 118
##
## Cluster means:
##   MediaSalarial PontuacaoGasto Frequencia
## 1      5.735632       13.22989   5.080460
## 2      8.953271       81.75701  14.831776
## 3      6.745763       47.60169   7.601695
##
## Clustering vector:
##   [1] 1 2 1 2 2 2 1 1 1 2 1 2 2 2 1 2 3 2 1 2 3 2 1 2 3 2 3 3 3 3 1 2 1 2 1 2 1
##  [38] 2 1 2 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2
##  [75] 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 3 3 3
## [112] 3 3 3 3 3 3 3 3 3 3 3 3 3 1 2 3 2 1 2 1 2 3 2 1 2 1 2 1 2 1 2 3 2 1 2 3 2
## [149] 1 2 1 2 1 2 1 2 1 2 3 2 3 2 1 2 1 2 1 2 1 3 1 2 1 2 1 2 1 2 1 2 3 2 1 3 3
## [186] 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 3 2 1 3 1 2 1 2 1 3 1 2 1 1 2 1 2 3 2 1
## [223] 2 3 2 1 2 1 2 1 1 3 2 1 2 1 2 3 3 1 3 1 2 1 2 1 2 1 2 1 2 1 2 1 3 1 2 1 2
## [260] 1 2 1 2 1 2 1 3 3 3 2 1 2 3 2 1 3 3 3 3 2 1 2 1 2 1 2 1 3 3 3 1 3 1 2 1 2
## [297] 3 3 3 2 1 2 1 2 1 2 1 2 3 3 3 2
##
## Within cluster sum of squares by cluster:
## [1]  6900.759 20625.421 14164.932
##  (between_SS / total_SS =  84.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

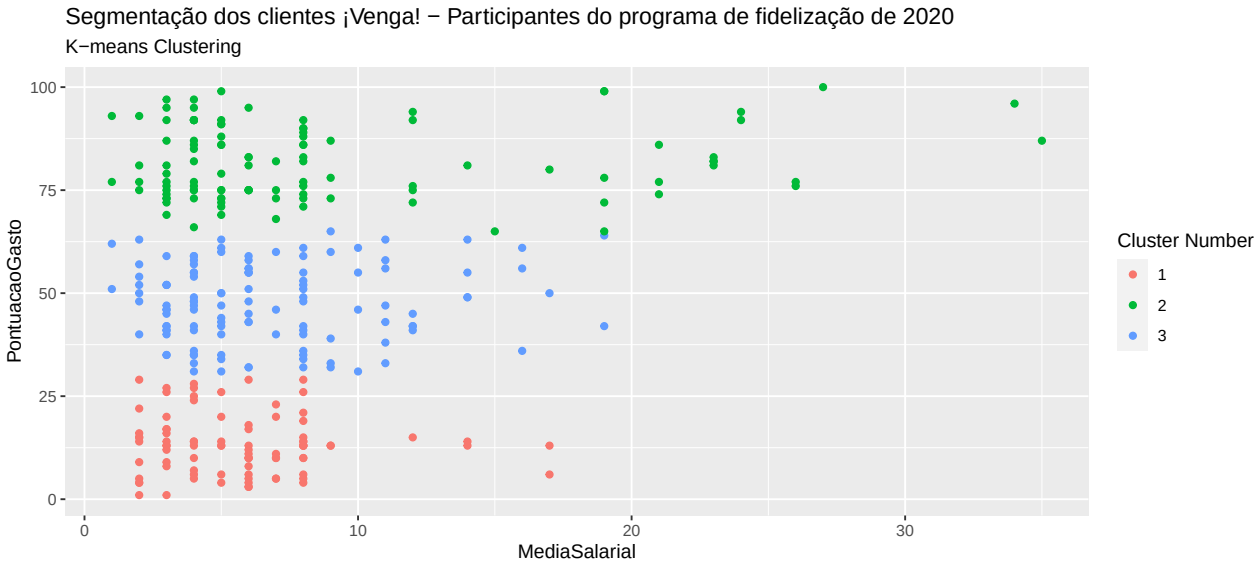
Na saída da nossa aplicação do algoritmo k-means, observamos uma lista com várias informações-chave. A partir daí, concluímos que as informações úteis são:

- *cluster* - Este é um vetor de vários inteiros que denotam o cluster que tem uma alocação de cada ponto;
- *centers* - Matriz composta de vários centros de agrupamento;
- *totss* - Isto representa a soma total dos quadrados;
- *withinss* - Este é um vetor que representa a soma intra-cluster dos quadrados tendo um componente por cluster;
- *tot.withinss* - Denota a soma total intra-cluster de quadrados;
- *betweenss* - Esta é a soma dos quadrados entre os quadrados de um cluster;
- *size* - O número total de pontos que cada aglomerado possui;
- *iter* - O número exterior (outer) de interações;
- *ifault* - Indicador de um possīvel problema com o algoritmo.

A seguir a plotagem da nossa base de clientes nos três conjuntos de agrupamentos sugeridos pelo algoritmo:



Podemos observar na figura acima que a variável idade está bem distribuída em todos os clusters, portanto não seria o melhor caminho para segmentar nossos clientes. Vamos tentar agora por média salarial.



Agora sim! Notamos claramente que o poder aquisitivo dos clientes interfere diretamente na frequência nas lojas e nos pedidos de delivery. Quanto maior a média salarial, maiores as pontuações de gasto e frequência.

Para confirmar a afirmação acima, extrairei as médias das variáveis dos três agrupamentos com o uso da função abaixo:

```
cluster_result <- aggregate(baseclientes[,4:6], by=list(cluster=km$cluster), mean)
cluster_result
```

```
##   cluster MediaSalarial PontuacaoGasto Frequencia
## 1       1      5.735632       13.22989   5.080460
## 2       2      8.953271       81.75701  14.831776
## 3       3      6.745763       47.60169   7.601695
```

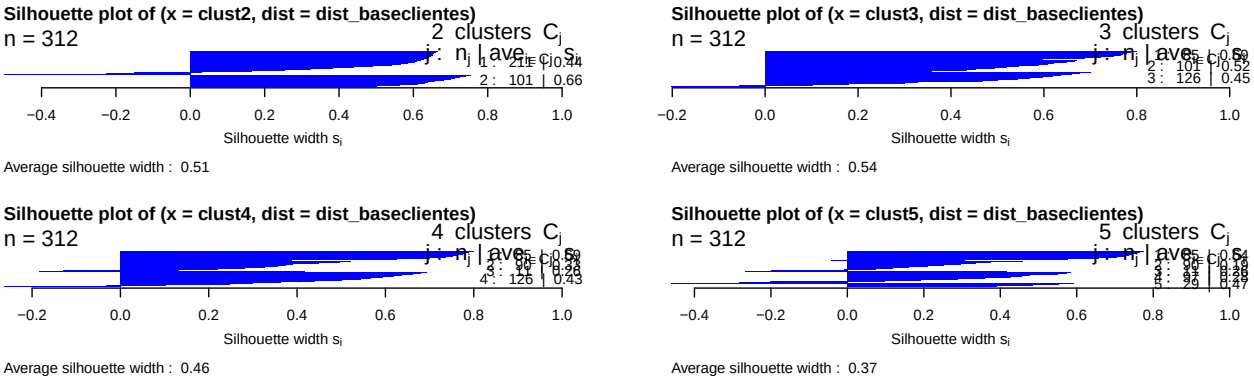
Clusterização hierárquica

Nesta parte da análise de segmentação da nossa base de clientes, utilizarei o agrupamento hierárquico aglomerativo (também conhecido como abordagem ascendente), sendo esse um método usado para agrupar objetos com base em sua similaridade. No início, cada observação começa em seu próprio agrupamento e, passo a passo, pares de agrupamentos são mesclados à medida que avançamos na hierarquia. Antes de implementar o algoritmo, compararei as distâncias entre os pontos de dados e, na próxima etapa, a função 'hclust' será usada para realizar a análise de agrupamento.

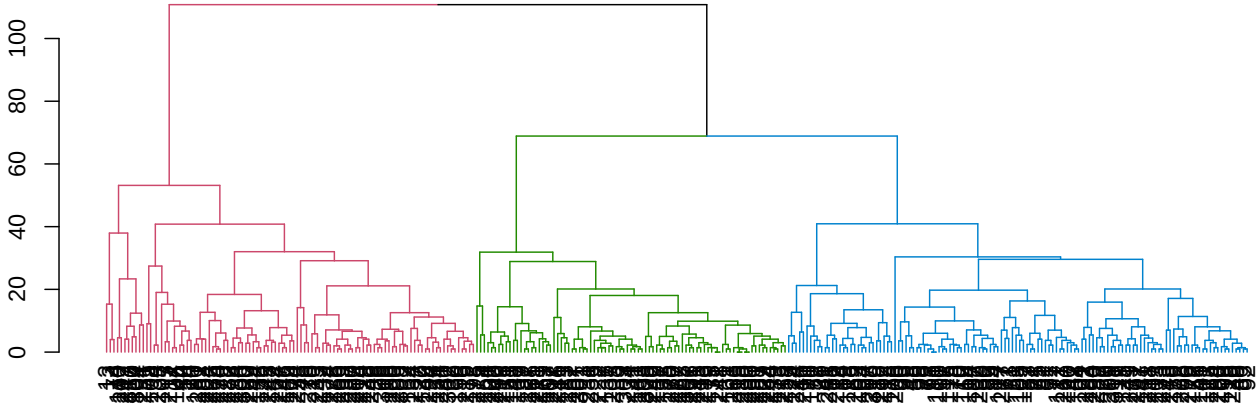
```
dist_baseclientes <- dist(baseclientes[,4:6])
hc_baseclientes <- hclust(dist_baseclientes, method = 'complete')
```

Nova consulta usando o método da silhueta

Agora é a hora de escolher o número ideal de clusters. Para fazer isso, da mesma forma que no método k-means, utilizarei o critério de silhueta.



O resultado das plotagens acima sugerem o uso de 3 clusters com a *average silhouette width* = 0.54 sendo o maior dos resultados. Com essa sugestão em mãos vamos a criação do dendograma.



O dendograma acima nos apresentou o processo de clusterização passo a passo, assim como analisou os níveis de distância dos três clusters formados. Um bom ponto de decisão da clusterização final é onde os valores de distância mudam consideravelmente.

Relacionando as variáveis

```r
segment_clientes <- mutate(baseclientes, cluster = clust_baseclientes)

segment_clientes = subset(segment_clientes, select = -c(ClusterNumber))

segment_clientes %>% group_by(cluster, Sexo) %>%
  summarise_all(list(mean)) %>% arrange(cluster)
```

```
## # A tibble: 6 x 8
## # Groups:   cluster [3]
##   cluster Sexo  `Cliente ID` Idade MediaSalarial PontuacaoGasto Frequencia
##     <int> <chr>        <dbl> <dbl>         <dbl>          <dbl>      <dbl>
## 1       1 F             185.  36.8          5.45           14.7       5.7
## 2       1 M             188.  37.8          5.93           11.2       4.51
## 3       2 F             148.  37.4         10.0            81.8      16.1
## 4       2 M             181.  37.2          7.88           83.5      14.1
## 5       3 F             127.  39.7          6.81           47.2       7.39
## 6       3 M             133.  41.3          6.89           49.5       8.07
## # ... with 1 more variable: NivelSatisfacao <dbl>
```

Podemos observar que as variáveis idade e sexo não trazem relevância quando agrupamos a base, pois se distribuem equilibradamente nos 3 clusters, não apresentando tendências.

Nosso foco deve estar na média salarial, já que suas variações influenciam diretamente nos agrupamentos e consequentemente impactam em cascata na pontuação de gastos e frequência nas visitas aos restaurantes e nos pedidos de delivery, que crescem proporcionalmente.

Considerações finais

A tabela acima nos mostra quantas informações importantes podemos obter com a análise de agrupamentos. Com a ajuda dos fatos apresentados neste projeto de ciência de dados, podemos compreender muito melhor o comportamento de consumo dos clientes da rede ¡Venga!, e o levará a tomadas de decisões muito mais cuidadosas e estratégicas.

A categorização das informações sobre esses consumidores da marca deve gerar segmentações relevantes para as próximas campanhas, visando esses clientes com base em vários parâmetros como renda, idade, poder aquisitivo, recorrência, preferências e etc. Além disso, padrões mais complexos, como revisões de produtos, engenharia de cardápio e preços.

A vantagem da clusterização é sua flexibilidade. Se ao acompanhar o desempenho das campanhas for percebido que a conversão está baixa, o custo de aquisição do cliente (CAC) alto e o retorno do investimento (ROI) baixo, é possível adaptar o cluster com novas segmentações a fim de melhorar o seu desempenho.