The Mode

No exercício passado aprendemos sobre a média ponderada e mediana. Aqui vamos nos aprofundar na medida estatística conhecida como Moda.

Base Casas Vendidas em Ames entre 2006 e 2010

Vamos usar essa base com 2930 linhas com 82 colunas contendo informações de características de casas vendidas entre 2006 e 2010 na cidade Ames (estado de Iowa nos EUA).

Esse foi um trabalho feito pelo professor Dean DeCock, publicado neste artigo e os detalhes sobre as informações presentes na base estão neste link

O separador da base são tabs, é um arquivo do tipo TSV (tab-separated value), são basicamente espaços. Poderíamos usar a função read.csv e informar o parâmetro sep= "\t" que funcionaria da mesma forma.

```
base <- read_tsv("https://s3.amazonaws.com/dq-content/444/AmesHousing.txt")
```

```
## Rows: 2930 Columns: 82
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (45): PID, MS SubClass, MS Zoning, Street, Alley, Lot Shape, Land Contou...
## dbl (37): Order, Lot Frontage, Lot Area, Overall Qual, Overall Cond, Year Bu...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(base)
```

```
## Rows: 2,930
## Columns: 82
## $ Order             <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1~
## $ PID               <chr> "0526301100", "0526350040", "0526351010", "052635303~
```

```
## $ `MS SubClass`     <chr> "020", "020", "020", "020", "060", "060", "120", "12~
## $ `MS Zoning`       <chr> "RL", "RH", "RL", "RL", "RL", "RL", "RL", "RL", "RL"~
## $ `Lot Frontage`    <dbl> 141, 80, 81, 93, 74, 78, 41, 43, 39, 60, 75, NA, 63,~
## $ `Lot Area`        <dbl> 31770, 11622, 14267, 11160, 13830, 9978, 4920, 5005,~
## $ Street            <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pav~
## $ Alley             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ `Lot Shape`       <chr> "IR1", "Reg", "IR1", "Reg", "IR1", "IR1", "Reg", "IR~
## $ `Land Contour`    <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "HL~
## $ Utilities         <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "A~
## $ `Lot Config`      <chr> "Corner", "Inside", "Corner", "Corner", "Inside", "I~
## $ `Land Slope`      <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gt~
## $ Neighborhood      <chr> "NAmes", "NAmes", "NAmes", "NAmes", "Gilbert", "Gilb~
## $ `Condition 1`     <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "No~
## $ `Condition 2`     <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "Nor~
## $ `Bldg Type`       <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "Twn~
## $ `House Style`     <chr> "1Story", "1Story", "1Story", "1Story", "2Story", "2~
## $ `Overall Qual`    <dbl> 6, 5, 6, 7, 5, 6, 8, 8, 8, 7, 6, 6, 6, 7, 8, 8, 8, 9~
## $ `Overall Cond`    <dbl> 5, 6, 6, 5, 5, 6, 5, 5, 5, 5, 5, 7, 5, 5, 5, 5, 7, 2~
## $ `Year Built`      <dbl> 1960, 1961, 1958, 1968, 1997, 1998, 2001, 1992, 1995~
## $ `Year Remod/Add`  <dbl> 1960, 1961, 1958, 1968, 1998, 1998, 2001, 1992, 1996~
## $ `Roof Style`      <chr> "Hip", "Gable", "Hip", "Hip", "Gable", "Gable", "Gab~
## $ `Roof Matl`       <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompShg~
## $ `Exterior 1st`    <chr> "BrkFace", "VinylSd", "Wd Sdng", "BrkFace", "VinylSd~
## $ `Exterior 2nd`    <chr> "Plywood", "VinylSd", "Wd Sdng", "BrkFace", "VinylSd~
## $ `Mas Vnr Type`    <chr> "Stone", "None", "BrkFace", "None", "None", "BrkFace~
## $ `Mas Vnr Area`    <dbl> 112, 0, 108, 0, 0, 20, 0, 0, 0, 0, 0, 0, 0, 0, 0, 60~
## $ `Exter Qual`      <chr> "TA", "TA", "TA", "Gd", "TA", "TA", "Gd", "Gd", "Gd"~
## $ `Exter Cond`      <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA"~
## $ Foundation        <chr> "CBlock", "CBlock", "CBlock", "CBlock", "PConc", "PC~
## $ `Bsmt Qual`       <chr> "TA", "TA", "TA", "TA", "Gd", "TA", "Gd", "Gd", "Gd"~
## $ `Bsmt Cond`       <chr> "Gd", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA"~
## $ `Bsmt Exposure`   <chr> "Gd", "No", "No", "No", "No", "No", "Mn", "No", "No"~
## $ `BsmtFin Type 1`  <chr> "BLQ", "Rec", "ALQ", "ALQ", "GLQ", "GLQ", "GLQ", "AL~
## $ `BsmtFin SF 1`    <dbl> 639, 468, 923, 1065, 791, 602, 616, 263, 1180, 0, 0,~
## $ `BsmtFin Type 2`  <chr> "Unf", "LwQ", "Unf", "Unf", "Unf", "Unf", "Unf", "Un~
## $ `BsmtFin SF 2`    <dbl> 0, 144, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1120, 0,~
## $ `Bsmt Unf SF`     <dbl> 441, 270, 406, 1045, 137, 324, 722, 1017, 415, 994, ~
## $ `Total Bsmt SF`   <dbl> 1080, 882, 1329, 2110, 928, 926, 1338, 1280, 1595, 9~
## $ Heating           <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", "Gas~
## $ `Heating QC`      <chr> "Fa", "TA", "TA", "Ex", "Gd", "Ex", "Ex", "Ex", "Ex"~
## $ `Central Air`     <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y~
## $ Electrical        <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr"~
## $ `1st Flr SF`      <dbl> 1656, 896, 1329, 2110, 928, 926, 1338, 1280, 1616, 1~
## $ `2nd Flr SF`      <dbl> 0, 0, 0, 0, 701, 678, 0, 0, 0, 776, 892, 0, 676, 0, ~
```

```
## $ `Low Qual Fin SF` <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ `Gr Liv Area`     <dbl> 1656, 896, 1329, 2110, 1629, 1604, 1338, 1280, 1616,~
## $ `Bsmt Full Bath`  <dbl> 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1~
## $ `Bsmt Half Bath`  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ `Full Bath`       <dbl> 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 3, 2, 1~
## $ `Half Bath`       <dbl> 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1~
## $ `Bedroom AbvGr`   <dbl> 3, 2, 3, 3, 3, 3, 2, 2, 2, 3, 3, 3, 3, 2, 1, 4, 4, 1~
## $ `Kitchen AbvGr`   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ `Kitchen Qual`    <chr> "TA", "TA", "Gd", "Ex", "TA", "Gd", "Gd", "Gd", "Gd"~
## $ `TotRms AbvGrd`   <dbl> 7, 5, 6, 8, 6, 7, 6, 5, 5, 7, 7, 6, 7, 5, 4, 12, 8, ~
## $ Functional        <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Ty~
## $ Fireplaces        <dbl> 2, 0, 0, 2, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1~
## $ `Fireplace Qu`    <chr> "Gd", NA, NA, "TA", "TA", "Gd", NA, NA, "TA", "TA", ~
## $ `Garage Type`     <chr> "Attchd", "Attchd", "Attchd", "Attchd", "Attchd", "A~
## $ `Garage Yr Blt`   <dbl> 1960, 1961, 1958, 1968, 1997, 1998, 2001, 1992, 1995~
## $ `Garage Finish`   <chr> "Fin", "Unf", "Unf", "Fin", "Fin", "Fin", "Fin", "RF~
## $ `Garage Cars`     <dbl> 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3~
## $ `Garage Area`     <dbl> 528, 730, 312, 522, 482, 470, 582, 506, 608, 442, 44~
## $ `Garage Qual`     <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA"~
## $ `Garage Cond`     <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA"~
## $ `Paved Drive`     <chr> "P", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y~
## $ `Wood Deck SF`    <dbl> 210, 140, 393, 0, 212, 360, 0, 0, 237, 140, 157, 483~
## $ `Open Porch SF`   <dbl> 62, 0, 36, 0, 34, 36, 0, 82, 152, 60, 84, 21, 75, 0,~
## $ `Enclosed Porch`  <dbl> 0, 0, 0, 0, 0, 0, 170, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ `3Ssn Porch`      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ `Screen Porch`    <dbl> 0, 120, 0, 0, 0, 0, 0, 144, 0, 0, 0, 0, 0, 0, 140, 2~
## $ `Pool Area`       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ `Pool QC`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Fence             <chr> NA, "MnPrv", NA, NA, "MnPrv", NA, NA, NA, NA, NA, NA~
## $ `Misc Feature`    <chr> NA, NA, "Gar2", NA, NA, NA, NA, NA, NA, NA, NA, "She~
## $ `Misc Val`        <dbl> 0, 0, 12500, 0, 0, 0, 0, 0, 0, 0, 0, 500, 0, 0, 0, 0~
## $ `Mo Sold`         <dbl> 5, 6, 6, 4, 3, 6, 4, 1, 3, 6, 4, 3, 5, 2, 6, 6, 6, 6~
## $ `Yr Sold`         <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010~
## $ `Sale Type`       <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD"~
## $ `Sale Condition`  <chr> "Normal", "Normal", "Normal", "Normal", "Normal", "N~
## $ SalePrice         <dbl> 215000, 105000, 172000, 244000, 189900, 195500, 2135~
```

Em que situações vamos querer computar a Moda, ou até mesmo em que situações a Média e Mediana não poderão ser calculadas.

Vamos olhar a informação abaixo. Ela classifica o declive do terreno do imóvel entre baixo, moderado e severo.

```
base$`Land Slope` %>% unique()
```

```
## [1] "Gtl" "Mod" "Sev"
```

```
base %>% nrow()
```

```
## [1] 2930
```

Estamos lidando com uma informação que é matematicamente impossível de calcular a média. Ainda que a informação passe uma ordem de grandeza, a medida se dá em palavras e não em números, não podendo calcular a mediana. Se a base tivesse uma quantidade ímpar de linhas, poderíamos ordenar do declive mais baixo até o mais severo e selecionar a linha do meio. Mas sendo par deveríamos tirar a média das duas posições do meio, e novamente não daria para calcular média visto que não temos números.

Nesse caso a Moda é bastante útil, pois através dela podemos descobrir o valor que mais se repete na base. Aqui seria "Gtl", o nível de declive mais baixo.

```
base$`Land Slope` %>% table()
```

```
## .
##  Gtl  Mod  Sev
## 2789  125   16
```

O R infelizmente não tem em suas funções base nenhuma função que calcule a moda e portanto vamos criar nossa própria função. E com a função chegamos no mesmo resultado acima que manualmente olhamos e percebemos o valor que mais se repetia na base.

```
calcular_moda <- function(vetor){
    vetor <- tibble(vetor)

    vetor %>%
    group_by(vetor) %>%
    summarise(registros = n()) %>%
    arrange(desc(registros)) %>%
    filter(row_number() == 1) %>%
    pull(vetor)
}

calcular_moda(base$`Land Slope`)
```

```
## [1] "Gtl"
```

A moda foi uma medida estatística que nos ajudou a analisar uma informação ordinal porém em formato de texto. Se tratava de uma informação que apesar de ser texto representava uma ordem de grandeza. Indo um pouco além, a moda pode contribuir em análises de uma informação em texto que não represente uma grandeza, onde cada valor tem simplesmente um signifcado diferente, nem melhor ou pior, nem maior ou menor que o outro.

Essa informação determina o estilo do telhado do imóvel.

```
base %>%
    group_by(`Roof Style`) %>%
    summarise(registros = n())%>%
    arrange(desc(registros))
```

```
## # A tibble: 6 x 2
##   `Roof Style` registros
##   <chr>            <int>
## 1 Gable             2321
## 2 Hip                551
## 3 Gambrel             22
## 4 Flat                20
## 5 Mansard             11
## 6 Shed                 5
```

```
calcular_moda(base$`Roof Style`)
```

```
## [1] "Gable"
```

Outra situação é quando a informação permite matematicamente falando calcular a média ou a mediana, estamos lidando com números, no entanto não deveríamos efetuar esses cálculos. Há situações onde os números representam categorias, e a média entre uma categoria entre 1 e 2 ser 1,5 não diz muita coisa.

A informação abaixo informa quantas cozinhas a casa tem, imagine que gostaríamos de saber quantas cozinhas é mais comum encontrar numa casa típica, então 1,5 cozinha seria uma informação sem sentido.

Lembrando aqui que num primeiro momento parece ser possível calcular a mediana, mas caímos de novo no problema de termos registros pares e precisar tirar a média dos 2 valores do meio, podendo novamente resultar num número com casas decimais.

Sendo assim a moda acaba sendo a escolha mais sensata.

```r
base %>%
    group_by(`Kitchen AbvGr`) %>%
    summarise(registros = n())
```

```
## # A tibble: 4 x 2
##   `Kitchen AbvGr` registros
##             <dbl>     <int>
## 1               0         3
## 2               1      2796
## 3               2       129
## 4               3         2
```

```r
calcular_moda(base$`Kitchen AbvGr`)
```

```
## [1] 1
```

Pode acontecer de 2 resultados terem a mesma frequência de aparições na base, e nesse caso a moda seria bimodal, ou seja duas modas Pode acontecer casos com mais de 2 e a informação seria multimodal. Agora se todos os valores da base se repetem na mesma frequência, aí nesse caso não teríamos nenhuma moda, pois nenhum valor aparece mais vezes que os demais.

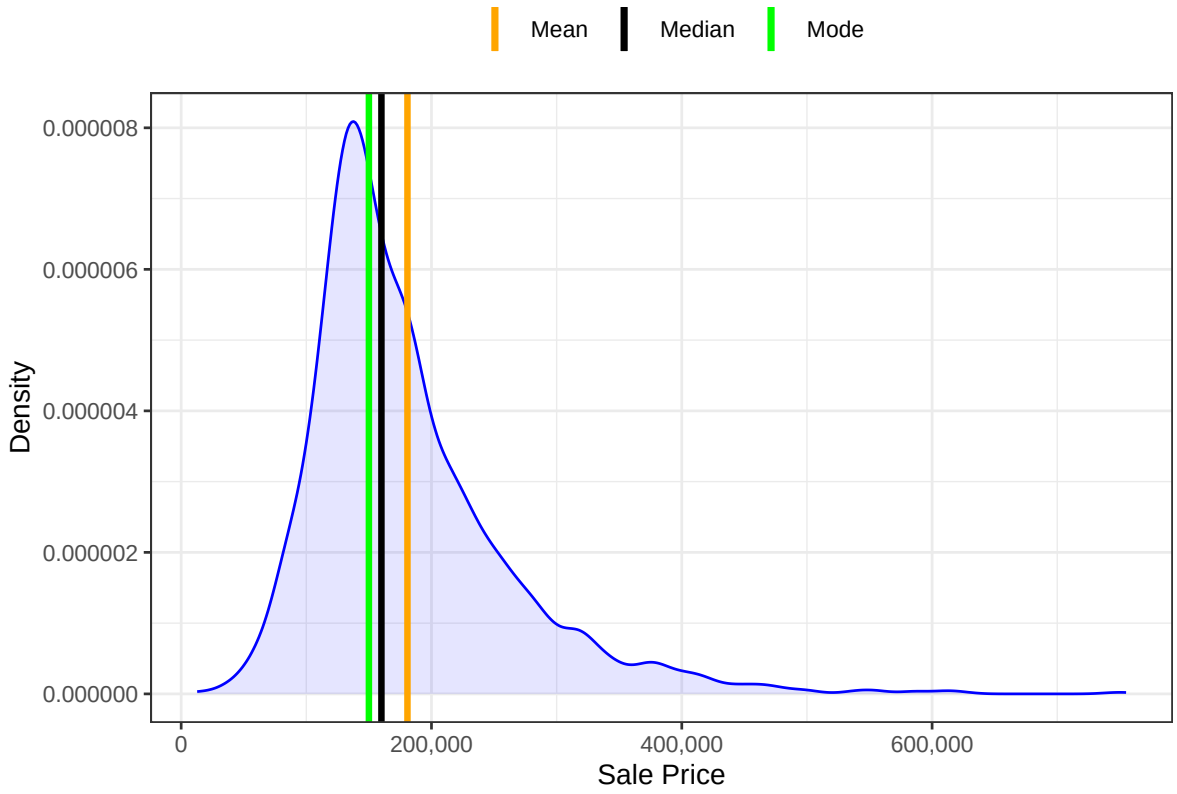
No caso de uma informação contínua, principalmente com valores decimais, pode ser mais difícil ocorrer essa repetição, e mesmo se ocorrer pode não ter uma quantidade significante para considerarmos essa moda.

Agora num histograma, usando um valor contínuo como SalePrice, em que posição ficaria a moda, média e mediana? Numa curva assimétrica notamos que a média sempre será puxada para o lado da cauda demonstrando o quando ela é afetada por valores muito discrepantes. A moda sempre fica no pico e a mediana no meio das duas medidas.

```r
base %>%
ggplot(aes(x = SalePrice)) +
    geom_density(alpha = 0.1,
                 color='blue',
                 fill='blue') +
    geom_vline(aes(xintercept = 150000,
                   color = 'Mode'),
               size = 1.2 ) +
    geom_vline(aes(xintercept = median(SalePrice),
                   color = 'Median'),
```

```
            size = 1.2 ) +
  geom_vline(aes(xintercept = mean(SalePrice),
                 color = 'Mean'),
             size = 1.2 ) +
  scale_y_continuous(labels = scales::comma) +
  scale_x_continuous(labels = scales::comma,
                     lim = c(min(base$SalePrice),
                     max(base$SalePrice))) +
  scale_colour_manual(values = c("Mode" = "green",
                                 "Median" = "black",
                                 "Mean" = "orange"),
                      name = "") +
  theme_bw() +
  theme(legend.position='top') +
  xlab("Sale Price") +
  ylab("Density")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Vale mencionar, que para uma distribuição normal, ou seja um histograma que é simétrico e centralizado, a média, moda e mediana vão estar centralizadas. A não ser que o caso seja uma curva simétrica com 2 picos, nesse caso a moda vai estar localizada no meio de cada um dos 2 picos, enquanto que a média e mediana estarão centralizadas no gráfico. Para distribuições uniformes, não teremos moda, uma vez que não existe picos e todos valores se repetem na mesma frequência.