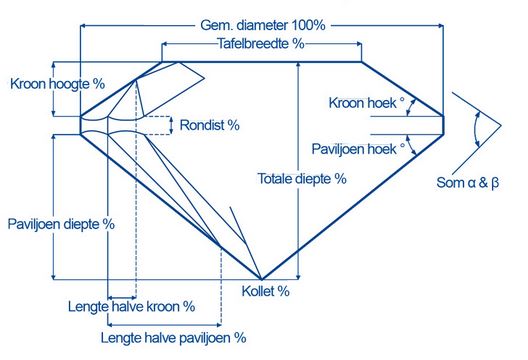
Data Science Demo ‘Diamonds Are Forever’

Henk op den Akker

14-7-2020

Na een eerste kennismaking met R in een Codecademy training is dit mijn eerste presentatie van de mogelijkheden om via R uit een verzameling data interessante en nuttige informatie te destilleren.

Met een knipoog naar Bond’s film Diamonds Are Forever koos ik voor een dataset met bijna 54000 diamanten.

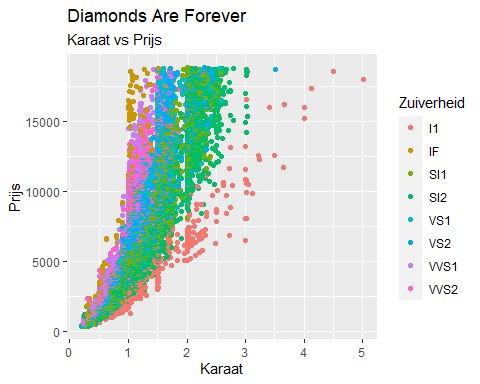


Schematische weergave diamant

## Aanpak

1. in RStudio RScripts gemaakt en vrijwel complete Codecademy lesstof inclusief gebruikte datasets geprogrammeerd.
2. in RStudio een RScript gemaakt met operaties op een diamonds.csv dataset. Een aantal aspecten uit het geleerde hier toegepast. Waaronder enkele basis statistische berekeningen.
   * variantie (steekproef) en standaard afwijking moet nog een keer
3. een RMarkdown file aangemaakt met de bedoeling om relevante RScripts te laden. Dit lukte mij alleen met de import van de libraries en dataset. Dat de rest niet ging zal ongetwijfeld te maken hebben met beginners manco’s …..
4. na run van alle code gekozen voor rendering naar HTML.

## Een scatter plot voorbeeld

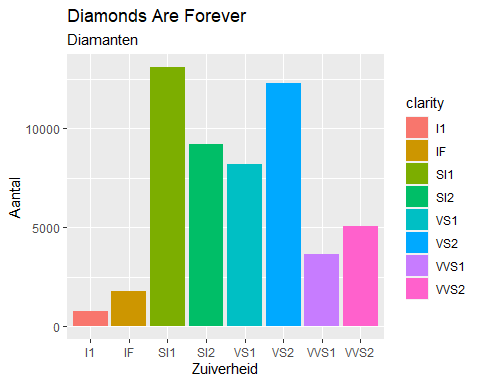
Deze scatter plot toont aan dat, volgens verwachting, bij een hogere karaatwaarde een hogere prijs hoort: 

## Voorbeeld tabelpresentatie

Aantal diamanten per zuiverheids categorie

|  |  |
| --- | --- |
| clarity | count |
| I1 | 741 |
| IF | 1790 |
| SI1 | 13065 |
| SI2 | 9194 |
| VS1 | 8171 |
| VS2 | 12258 |
| VVS1 | 3655 |
| VVS2 | 5066 |

Idem maar in histogram (ggplot):



Hoogste prijs per zuiverheids categorie

|  |  |
| --- | --- |
| clarity | max\_price |
| I1 | 18531 |
| IF | 18806 |
| SI1 | 18818 |
| SI2 | 18804 |
| VS1 | 18795 |
| VS2 | 18823 |
| VVS1 | 18777 |
| VVS2 | 18768 |

Gemiddelde prijs per zuiverheids categorie maar voor slijpsel Premium en karaatwaarde tussen 1 en 2

|  |  |
| --- | --- |
| clarity | mean(price, na.rm = TRUE) |
| I1 | 3790.036 |
| IF | 11032.264 |
| SI1 | 6949.606 |
| SI2 | 5655.201 |
| VS1 | 9226.879 |
| VS2 | 8430.114 |
| VVS1 | 10863.481 |
| VVS2 | 9958.858 |

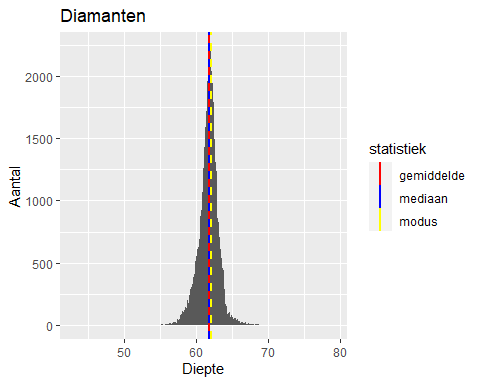
Alleen eerste 20 records

Prijsverschil met gemiddelde prijs per zuiverheids categorie

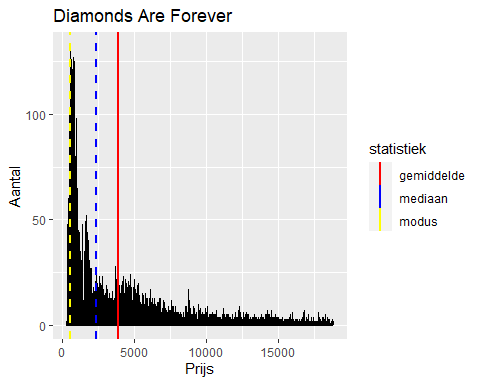
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| carat | cut | color | clarity | depth | table | price | x | y | z | prijsverschil\_met\_gemiddelde |
| 0.23 | Ideal | E | SI2 | 61.5 | 55 | 326 | 3.95 | 3.98 | 2.43 | -4737.029 |
| 0.21 | Premium | E | SI1 | 59.8 | 61 | 326 | 3.89 | 3.84 | 2.31 | -3670.001 |
| 0.23 | Good | E | VS1 | 56.9 | 65 | 327 | 4.05 | 4.07 | 2.31 | -3512.455 |
| 0.29 | Premium | I | VS2 | 62.4 | 58 | 334 | 4.20 | 4.23 | 2.63 | -3590.989 |
| 0.31 | Good | J | SI2 | 63.3 | 58 | 335 | 4.34 | 4.35 | 2.75 | -4728.029 |
| 0.24 | Very Good | J | VVS2 | 62.8 | 57 | 336 | 3.94 | 3.96 | 2.48 | -2947.737 |
| 0.24 | Very Good | I | VVS1 | 62.3 | 57 | 336 | 3.95 | 3.98 | 2.47 | -2187.115 |
| 0.26 | Very Good | H | SI1 | 61.9 | 55 | 337 | 4.07 | 4.11 | 2.53 | -3659.001 |
| 0.22 | Fair | E | VS2 | 65.1 | 61 | 337 | 3.87 | 3.78 | 2.49 | -3587.989 |
| 0.23 | Very Good | H | VS1 | 59.4 | 61 | 338 | 4.00 | 4.05 | 2.39 | -3501.455 |
| 0.30 | Good | J | SI1 | 64.0 | 55 | 339 | 4.25 | 4.28 | 2.73 | -3657.001 |
| 0.23 | Ideal | J | VS1 | 62.8 | 56 | 340 | 3.93 | 3.90 | 2.46 | -3499.455 |
| 0.22 | Premium | F | SI1 | 60.4 | 61 | 342 | 3.88 | 3.84 | 2.33 | -3654.001 |
| 0.31 | Ideal | J | SI2 | 62.2 | 54 | 344 | 4.35 | 4.37 | 2.71 | -4719.029 |
| 0.20 | Premium | E | SI2 | 60.2 | 62 | 345 | 3.79 | 3.75 | 2.27 | -4718.029 |
| 0.32 | Premium | E | I1 | 60.9 | 58 | 345 | 4.38 | 4.42 | 2.68 | -3579.169 |
| 0.30 | Ideal | I | SI2 | 62.0 | 54 | 348 | 4.31 | 4.34 | 2.68 | -4715.029 |
| 0.30 | Good | J | SI1 | 63.4 | 54 | 351 | 4.23 | 4.29 | 2.70 | -3645.001 |
| 0.30 | Good | J | SI1 | 63.8 | 56 | 351 | 4.23 | 4.26 | 2.71 | -3645.001 |
| 0.30 | Very Good | J | SI1 | 62.7 | 59 | 351 | 4.21 | 4.27 | 2.66 | -3645.001 |

## Met enkele waardes uit beschrijvende statistiek

Histogram diepte diamant met gemiddelde, mediaan en modus (ggplot):



Histogram prijs diamant met gemiddelde, mediaan en modus (met qplot):



Aantal diamanten geprijsd onder 1000,00

|  |
| --- |
| count |
| 14524 |