Group 11, diamonds dataset

Tom Tribe, Ken MacIver, Jundi Yang, Mei Huang

2022-09-15

Group Members (photos)

Jundi

Tom

Mei

Ken

Group Members (name, email, ORCID)

Tom Tribe

- tom.tribe2016@gmail.com
- 0000-0002-5002-8066

Ken MacIver

- ken.maciver68@gmail.com
- 0000-0001-8999-4598

Jundi Yang

- ivyli112358@gmail.com
- 0000-0003-0888-9564

Mei Huang

- huangmei139@gmail.com
- 0000-0003-2401-0679

The Diamonds dataset

- This large dataset has 53940 rows (diamonds) of ten variables (approx 540,000 values)
- Slow to process!
- Nine of the variables are various measures of diamond size and quality, while the tenth is the price
- We selected diamonds because it was simple to understand what each variable was measuring, and to have the opportunity to work with a large dataset
- Particularly interested in which variables are most predictive of diamond price

The Variables

red font = categorical variable

- carat: the diamond's weight
- cut: a measure of quality
- color: a measure of colour quality
- clarity: a measure of clearness
- x: length in mm
- y: width in mm
- z: depth in mm
- depth: total depth percentage
- table: width of top of diamond relative to widest point
- price: the price of the diamond in US dollars

(List adapted from list at kaggle.com).

The Response Variable

'Price' seemed to us to be the obvious response variable.

Data Visualization (the dataset)

```
##    carat       cut color clarity depth table price    x    y    z
## 1   0.23     Ideal     E     SI2  61.5    55   326 3.95 3.98 2.43
## 2   0.21   Premium     E     SI1  59.8    61   326 3.89 3.84 2.31
## 3   0.23      Good     E     VS1  56.9    65   327 4.05 4.07 2.31
## 4   0.29   Premium     I     VS2  62.4    58   334 4.20 4.23 2.63
## 5   0.31      Good     J     SI2  63.3    58   335 4.34 4.35 2.75
## 6   0.24 Very Good     J    VVS2  62.8    57   336 3.94 3.96 2.48
## 7   0.24 Very Good     I    VVS1  62.3    57   336 3.95 3.98 2.47
## 8   0.26 Very Good     H     SI1  61.9    55   337 4.07 4.11 2.53
## 9   0.22      Fair     E     VS2  65.1    61   337 3.87 3.78 2.49
## 10  0.23 Very Good     H     VS1  59.4    61   338 4.00 4.05 2.39
## 11  0.30      Good     J     SI1  64.0    55   339 4.25 4.28 2.73
## 12  0.23     Ideal     J     VS1  62.8    56   340 3.93 3.90 2.46
## 13  0.22   Premium     F     SI1  60.4    61   342 3.88 3.84 2.33
## 14  0.31     Ideal     J     SI2  62.2    54   344 4.35 4.37 2.71
## 15  0.20   Premium     E     SI2  60.2    62   345 3.79 3.75 2.27
```

Data Visualisation (pairs plot)

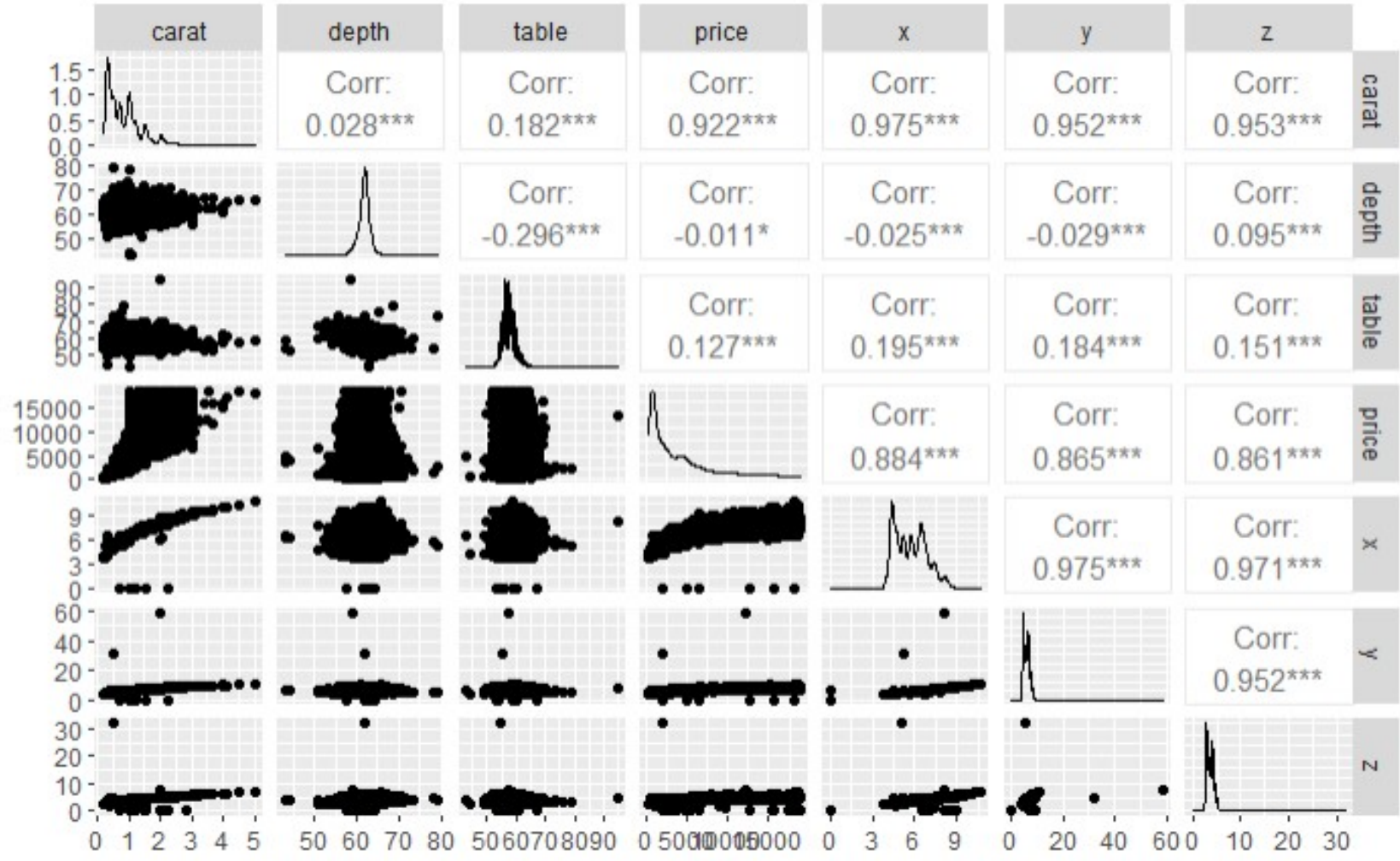

Figure 1: Pairs plot

Other things of interest

The EDA revealed the following:

- some variables not Normally distributed
- long right tail for 'price' due to a few very expensive diamonds
- some zero values
- 'price' probably follows a beta distribution (from the Cullen-Frey plot)

Next Steps

- Principal Component Analysis
- Regression using the Principal Components
- Find best predictor variable for price