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 (4 levels)
- color: a measure of colour quality (7 levels)
- clarity: a measure of clearness (6 levels)
- 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).

Data Visualisation (pairs plot)

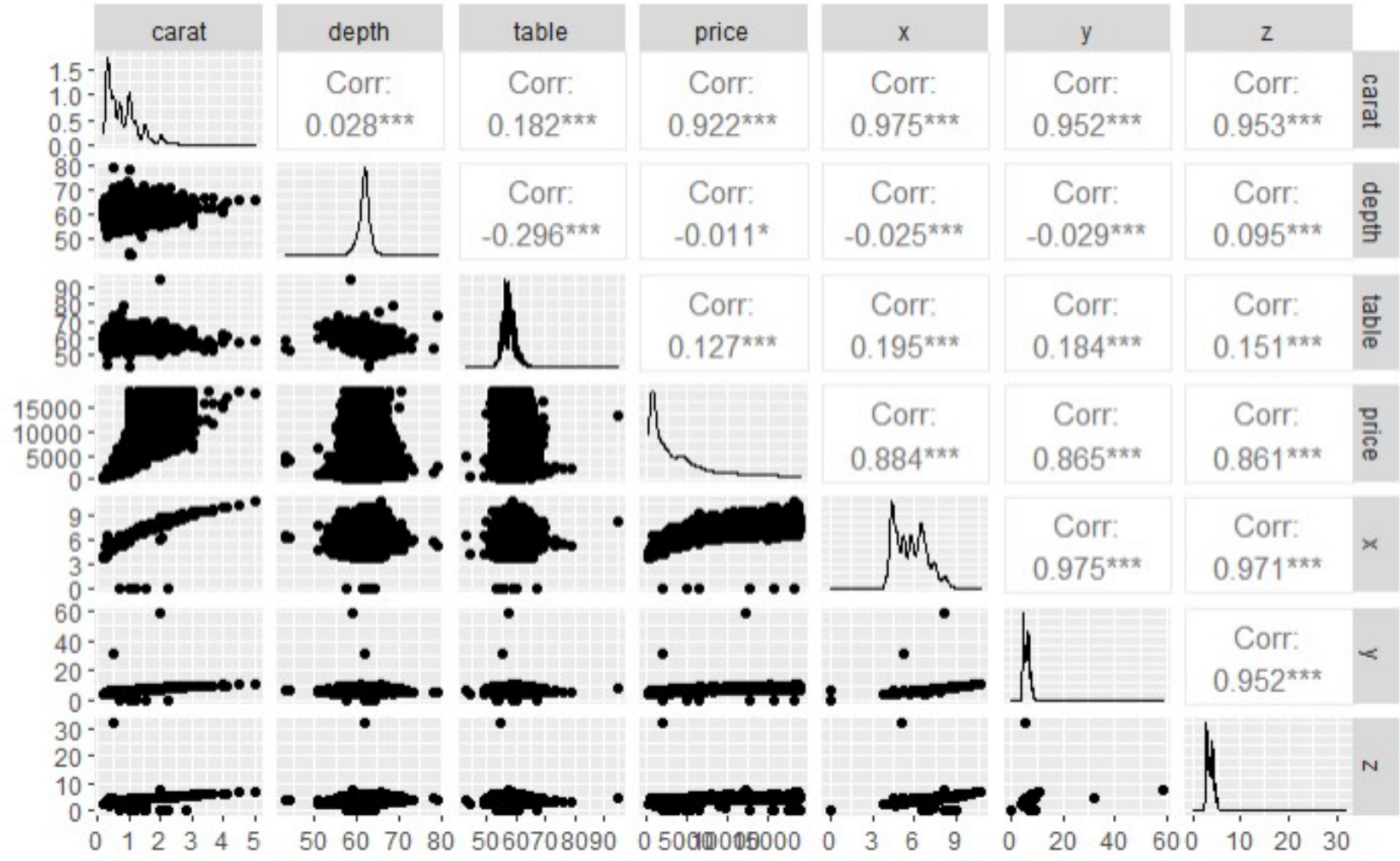
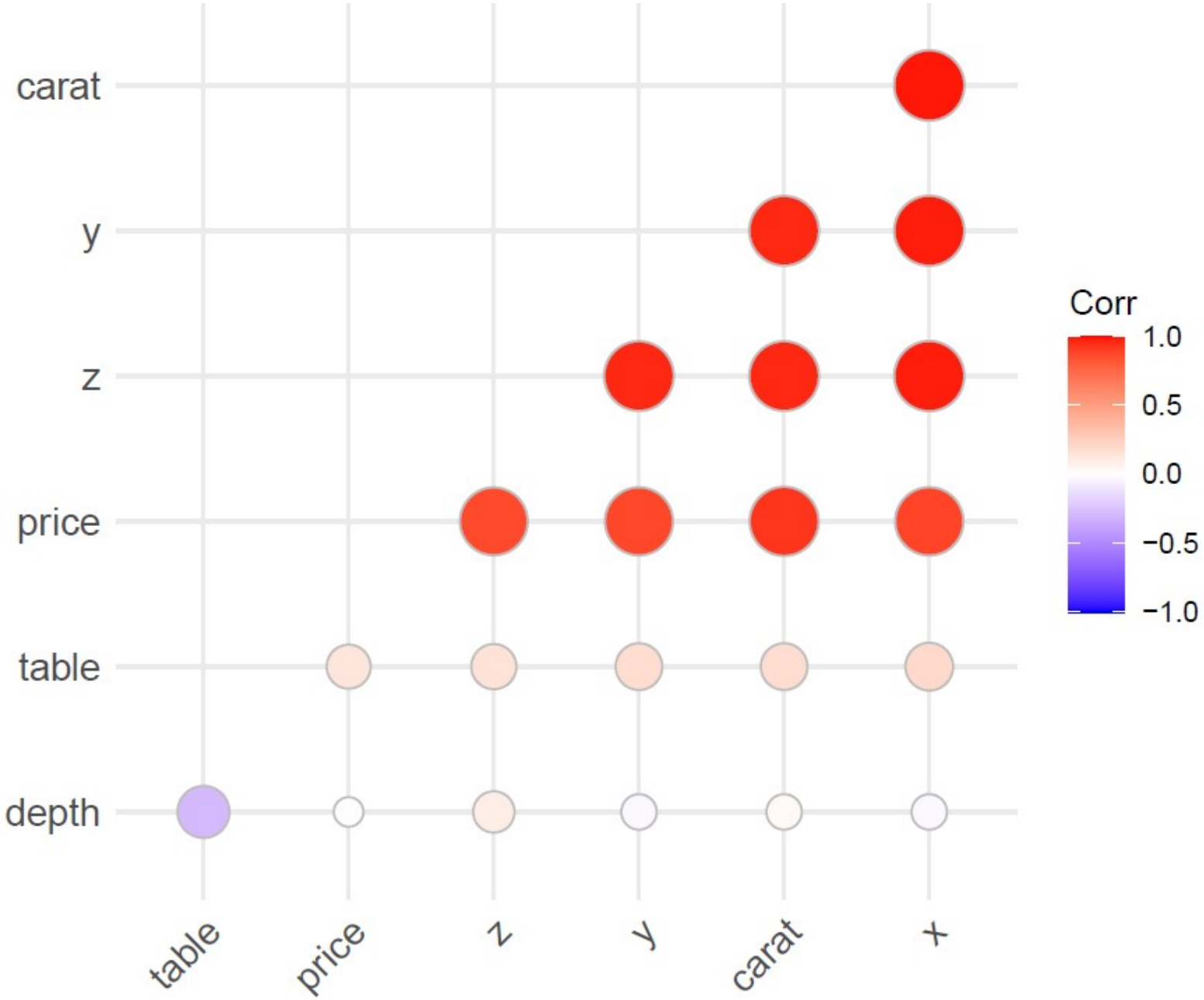


Figure 1: Pairs plot

Data Visualization (correlation 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