Group 11 Final Presentation

Tom Tribe

2022-09-28

Title Page

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).

Summary

Table 1: Summary statistics for 'Diamonds' (3 s.f.)

	carat	depth	table	price	x	y	z
sample size	53940.000	53940.000	53940.000	53940.000	53940.000	53940.000	53940.000
minimum	0.200	43.000	43.000	326.000	0.000	0.000	0.000
first quartile	0.400	61.000	56.000	950.000	4.710	4.720	2.910
median	0.700	61.800	57.000	2401.000	5.700	5.710	3.530
mean	0.798	61.749	57.457	3932.800	5.731	5.735	3.539
third quartile	1.040	62.500	59.000	5324.250	6.540	6.540	4.040
maximum	5.010	79.000	95.000	18823.000	10.740	58.900	31.800
IQR	0.640	1.500	3.000	4374.250	1.830	1.820	1.130
standard deviation	0.474	1.433	2.234	3989.440	1.122	1.142	0.706
skewness	1.117	-0.082	0.797	1.618	0.379	2.434	1.522

Data Visualisation (pairs plot)

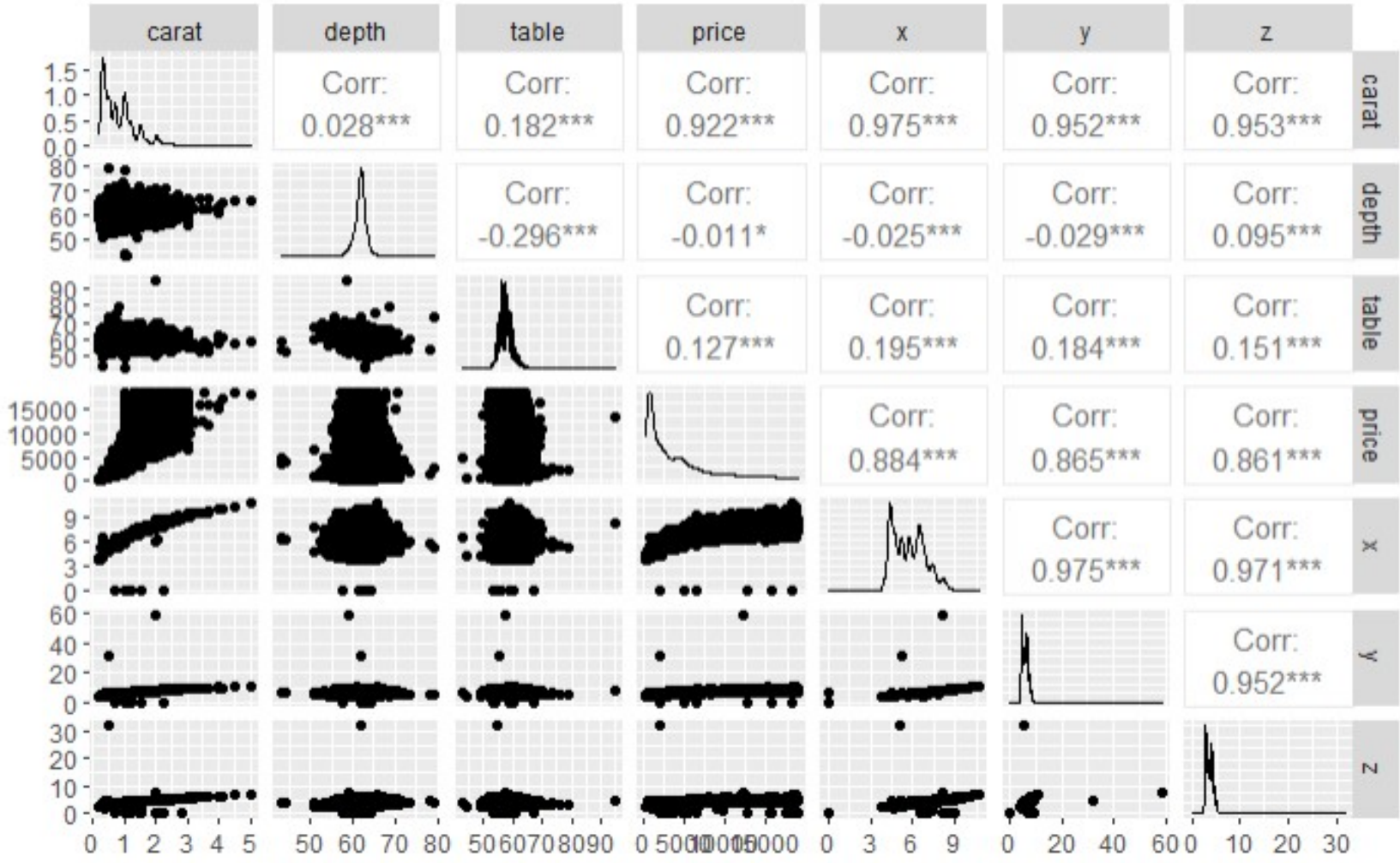
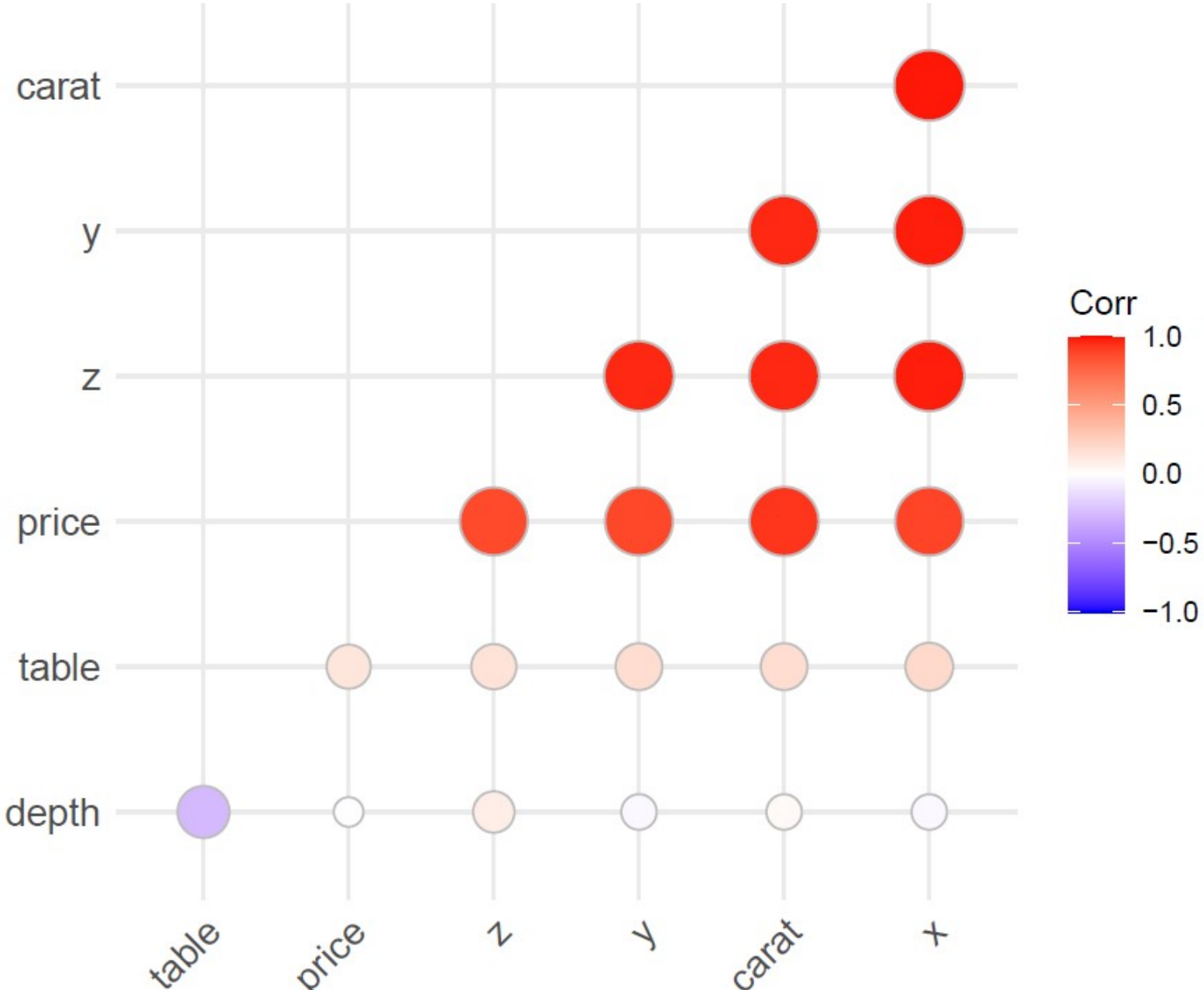


Figure 1: Pairs plot

Data Visualization (correlation plot)



Other things of interest

The EDA revealed the following:

- long right tail for 'price' due to a few very expensive diamonds
- 'price' probably follows a beta distribution (from the Cullen-Frey plot)
- some zero values

Next Steps

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