Group 11 Final Presentation

Tom Tribe, Ken MacIver, Jundi Yang, Mei Huang

2022-10-05

Group 11: Diamonds Dataset

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 of Numeric Variables

	carat	depth	table	price	x	y	z
sample size	53940	53940	53940	53940	53940	53940	53940
minimum	0.20	43.00	43.00	326.00	0.00	0.00	0.00
first quartile	0.40	61.00	56.00	950.00	4.71	4.72	2.91
median	0.70	61.80	57.00	2401.00	5.70	5.71	3.53
mean	0.80	61.75	57.46	3932.80	5.73	5.73	3.54
third quartile	1.04	62.50	59.00	5324.25	6.54	6.54	4.04
maximum	5.01	79.00	95.00	18823.00	10.74	58.90	31.80
IQR	0.64	1.50	3.00	4374.25	1.83	1.82	1.13
standard deviation	0.47	1.43	2.23	3989.44	1.12	1.14	0.71
skewness	1.12	-0.08	0.80	1.62	0.38	2.43	1.52
kurtosis	4.26	8.74	5.80	5.18	2.38	94.21	50.08

Cateogrical Summary

Cut	Fair	Good	Very Good	Ideal	Premium
Count	1610	4960	12082	21551	13791

Color	D	E	F	G	H	I	J
Count	6775	9797	9542	11292	8304	5422	2808

Clarity	I1	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
Count	741	1790	13065	9194	8171	12258	3655	5066

Pairs Plot

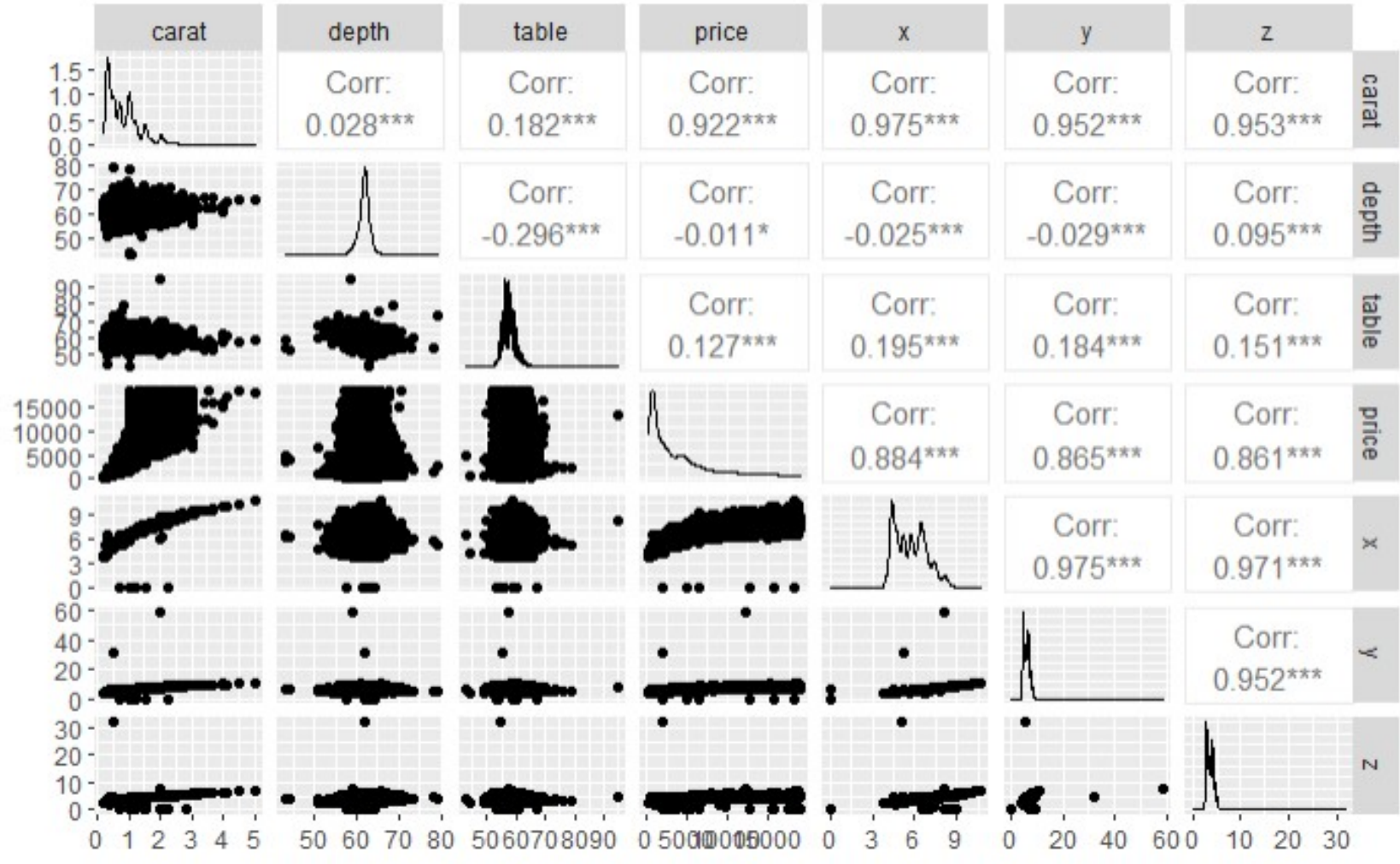
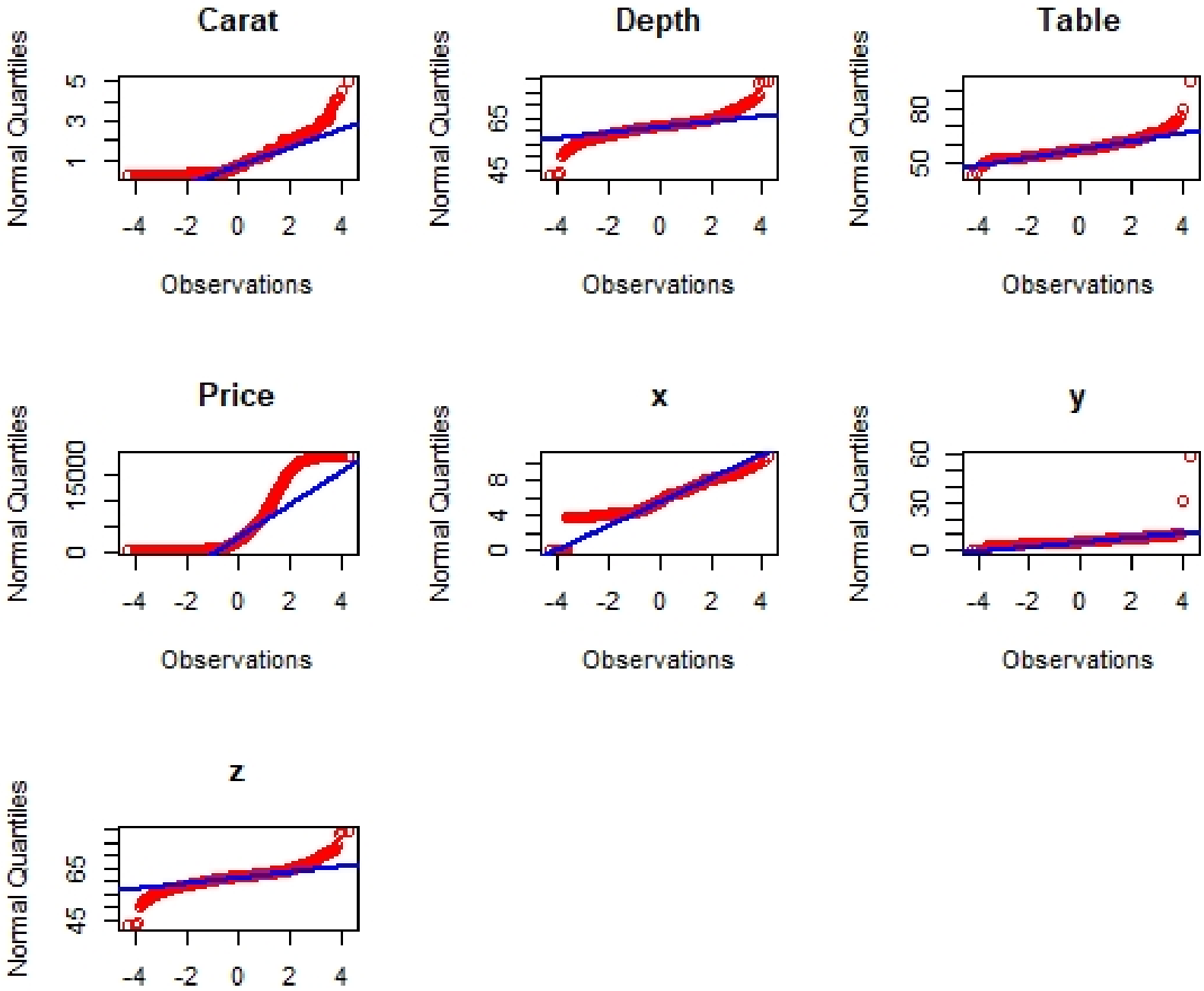
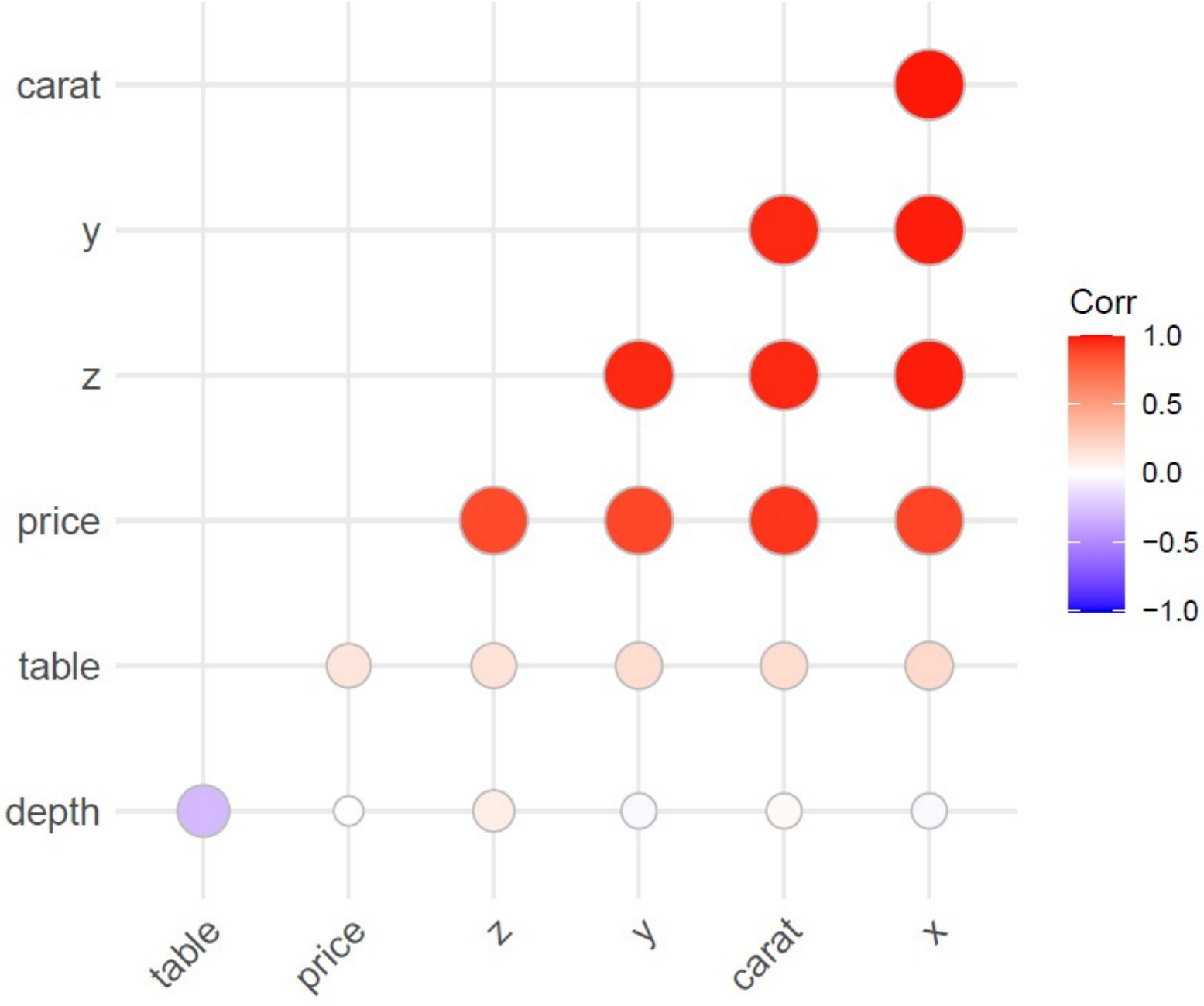


Figure 1: Pairs plot

Normal QQ Plots



Correlation Plot



Price by Cateogrical

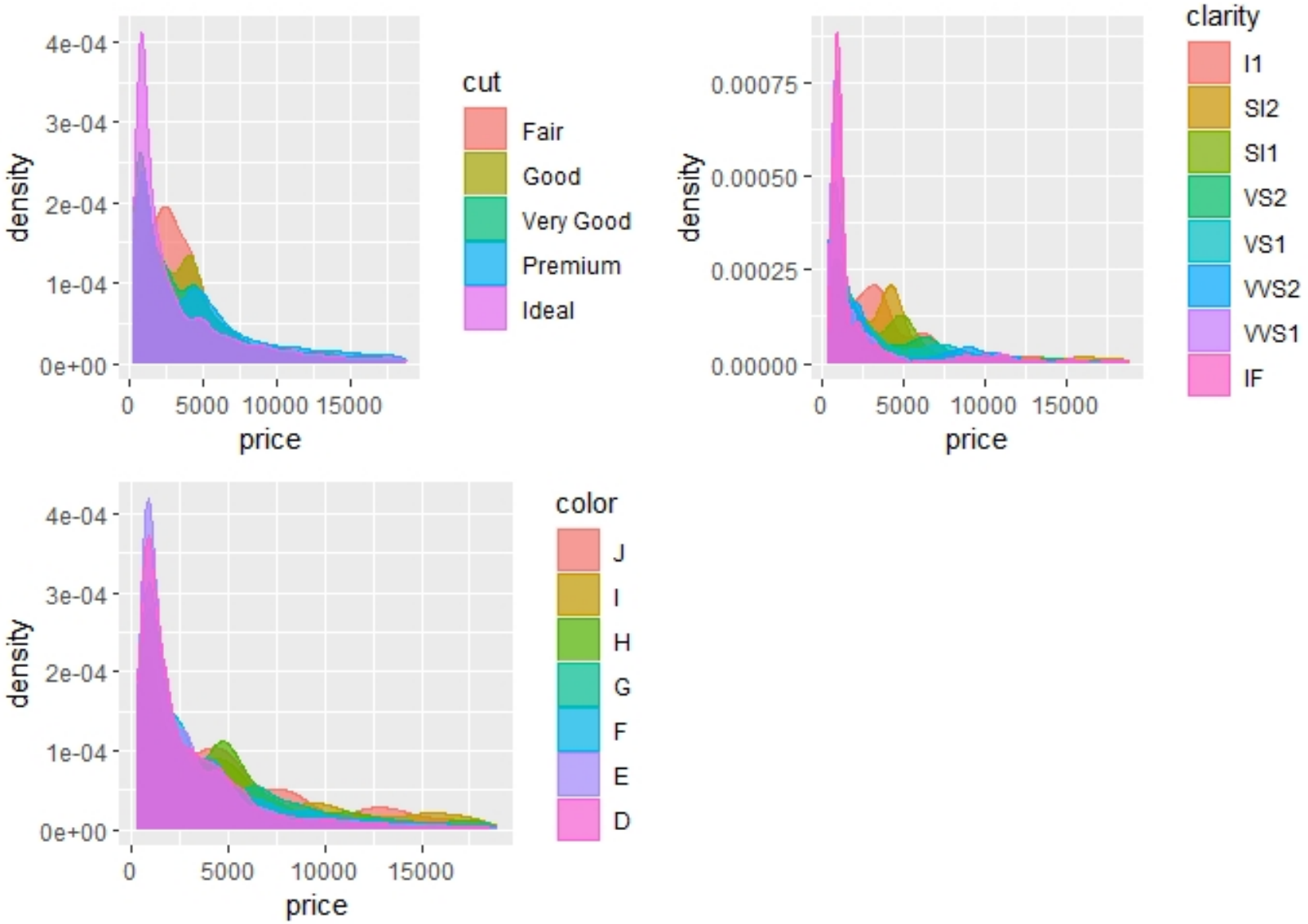


Figure 4: Price by Categorical

Leading Question 1

- How can we best predict diamond price?
- We intend to use the following techniques to investigate this question:
- Stepwise Regression, Principal Components Analysis, Principal Components Regression and an Exploratory Factor Analysis

Multiple Regression

- Starting with the full model we used a stepwise regression procedure to find the best model for predicting diamond price.
- According to AIC the best model was:
- price $\sim$ carat + cut + color + clarity + depth + table + x
- All variables excluding y and z are significant in the model
- The 'best' model had an Adjusted R^2 of 91.98%

Regression Assumptions

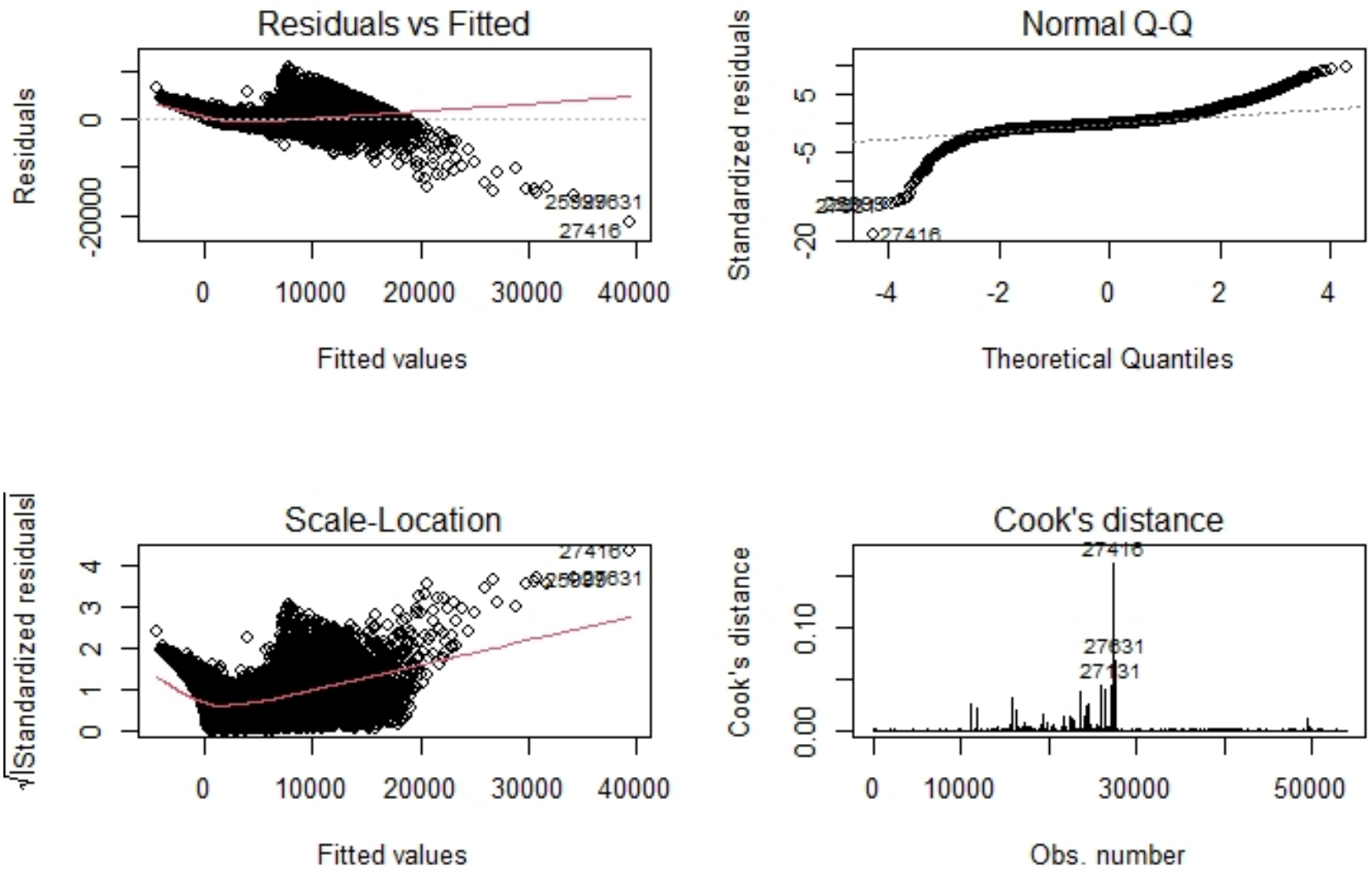


Figure 5: Regression Diagnostics

Principal Components Analysis: Screeplot

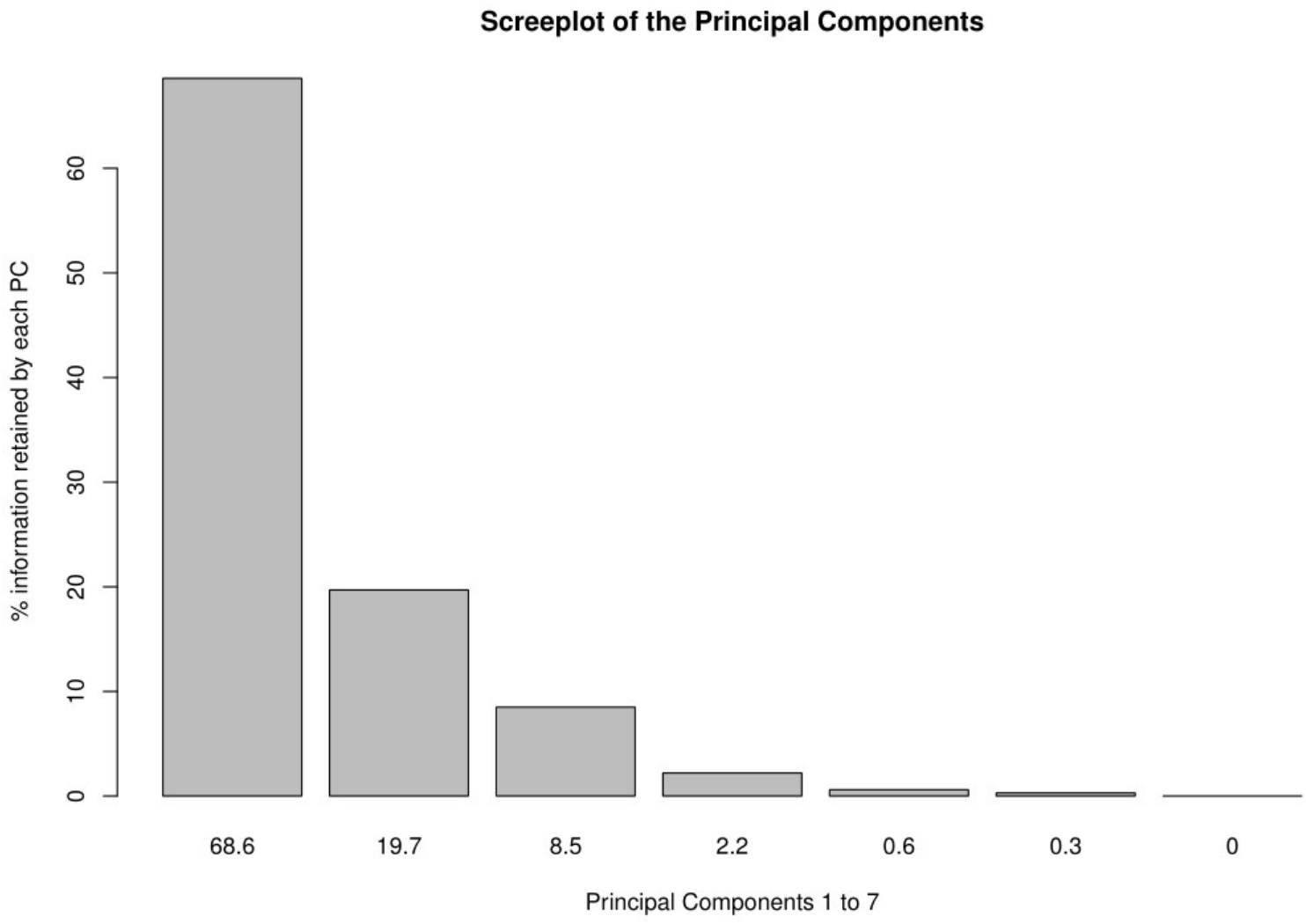


Figure 6: PCA Screeplot

Principal Components Analysis: Values v1

Can decide which of these two to keep

```
##                PC1          PC2
## carat   0.45032907 -0.049484287
## depth  -0.03619535 -0.722832026
## table   0.10580678  0.678011337
## price   0.42566082 -0.052989386
## x       0.45214098 -0.003202394
## y       0.45214716 -0.004959419
## z       0.44118352 -0.111906516
```

Principal Components Analysis: Values v2

Can decide which of these two to keep

Table 2: Principal Components 1 and 2

	PC1	PC2
carat	0.4503	-0.04948
depth	-0.0362	-0.7228
table	0.1058	0.678
price	0.4257	-0.05299
x	0.4521	-0.003202
y	0.4521	-0.004959
z	0.4412	-0.1119

Biplot

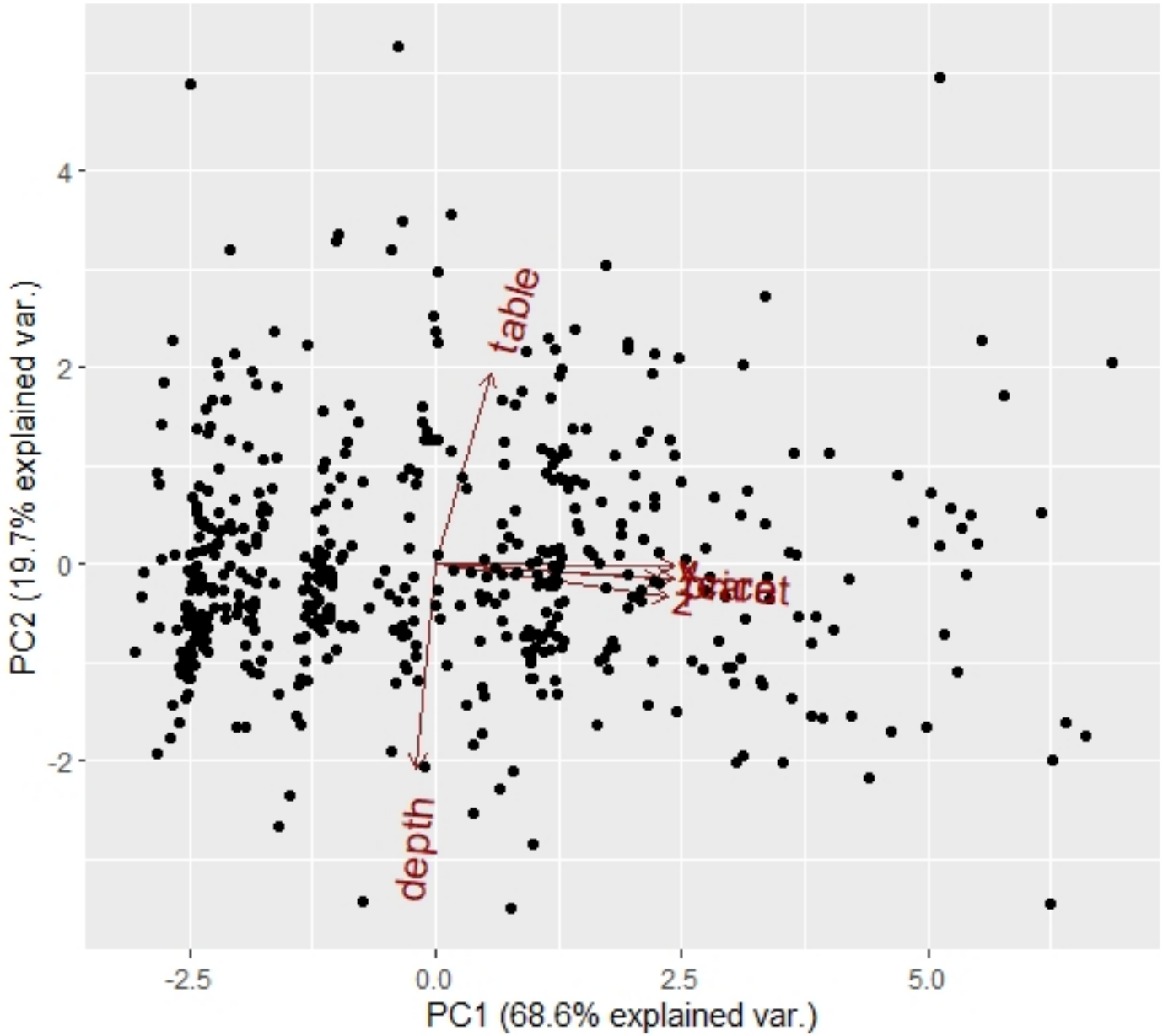


Figure 7: PCA Biplot

Principal Components Regression

- We conducted a Principal Components Regression with diamond price as the response variable
- We found that all six principal components were significant in the model for predicting price
- However when we only used the first two principal components we were still able to explain over 80% of the variation in price

Factor Analysis

- We hypothesized that the variation in the data might be able to be explained by two factors
- These were: “Dimension x Price” and “Light Conductance”
- We tested this hypothesis with a Factor Analysis