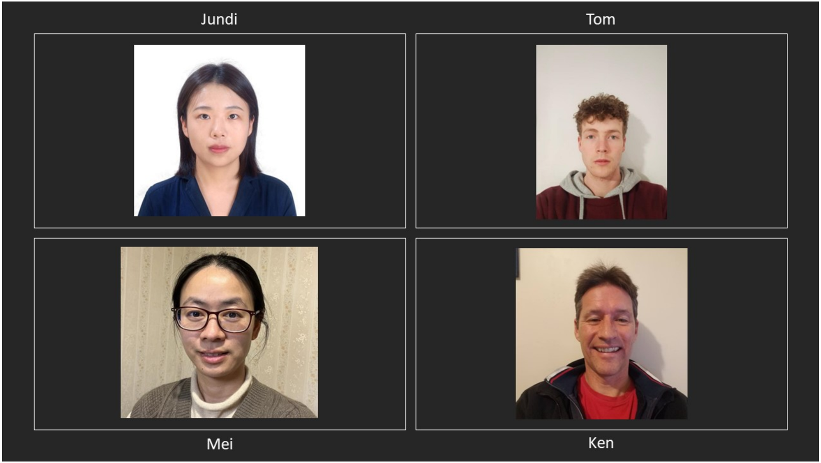
Group 11 Final Presentation

Tom Tribe, Ken MacIver, Jundi Yang, Mei Huang

2022-10-11

Group 11: Diamonds Dataset

Group Members (photos)



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!
- There are seven numeric variables and three categorical variables
- We selected diamonds because it was conceptually simple to understand what each variable was measuring, and to have the opportunity to use the analytical techniques taught in STAT394 with a large dataset

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	Premium	Ideal
Count	1610	4960	12082	13791	21551

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

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

Pairs Plot

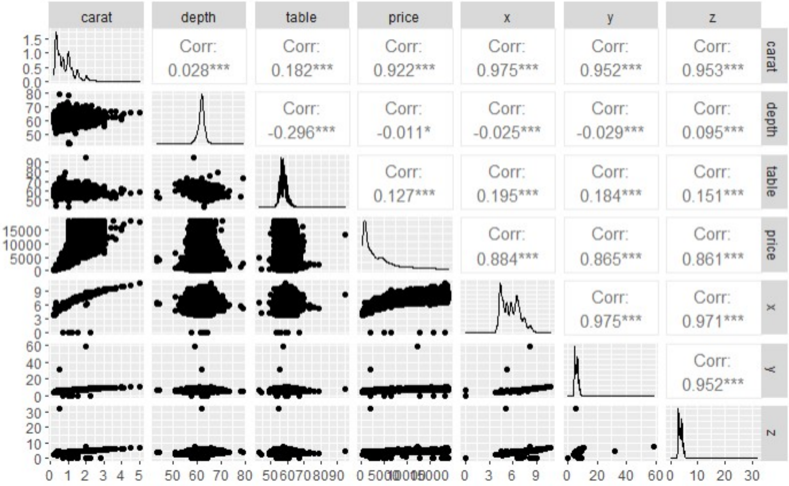


Figure 1: Pairs plot

Normal QQ Plots

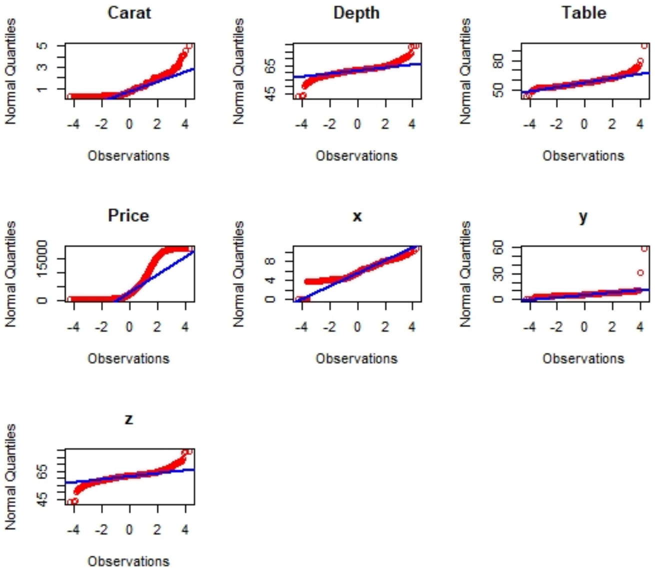


Figure 2: Normal QQ Plots

Correlation Plot

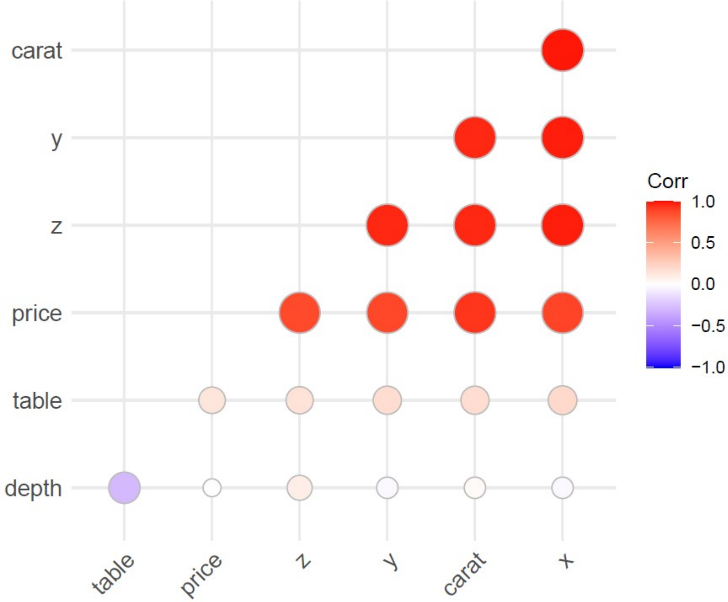


Figure 3: Correlation Plot

Price by Categorical

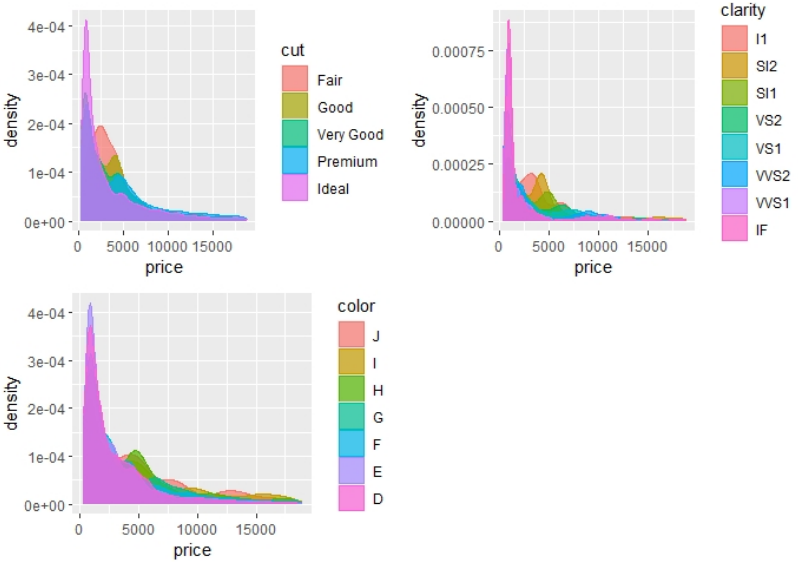


Figure 4: Price by Categorical

Leading Question 1

- How can we best predict diamond price using the other variables?
- We intend to use the following techniques to investigate this question:
- Stepwise Regression, Principal Components Analysis, Principal Components Regression

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 ~ 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

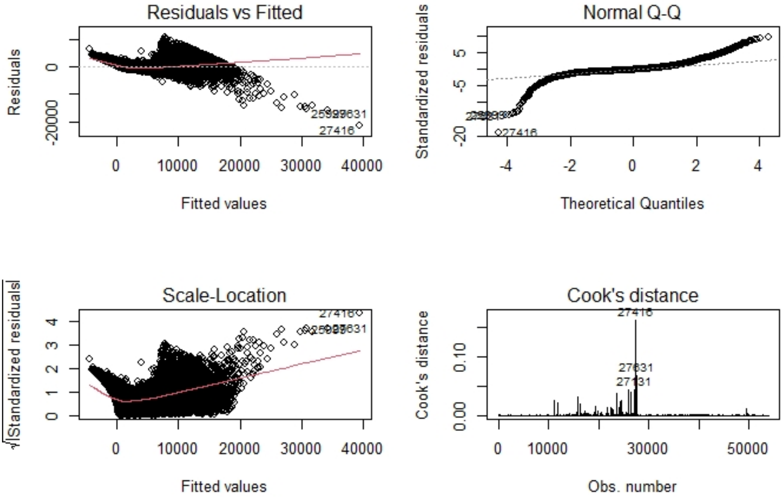


Figure 5: Regression Diagnostics

Principal Components Analysis: Screeplot

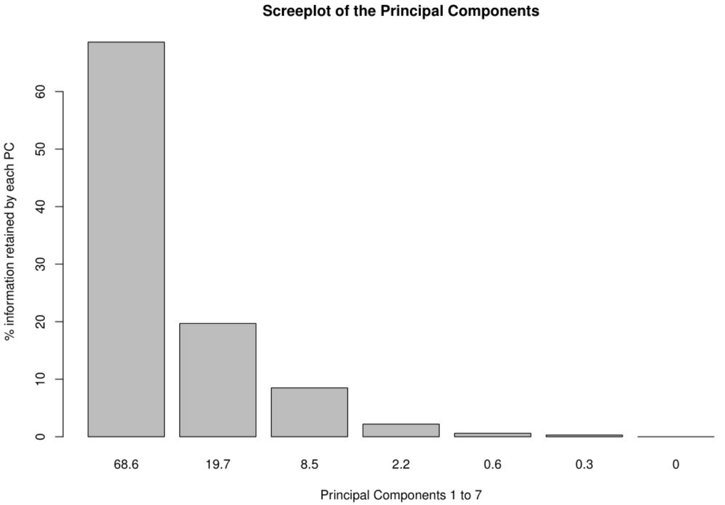


Figure 6: PCA Screeplot

Principal Components Analysis: Eigenvectors

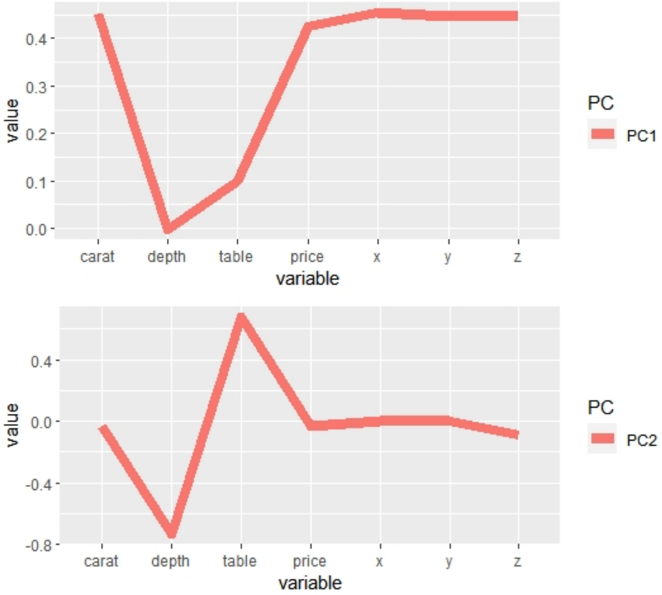


Figure 7: Plot of EigenVectors

Biplot

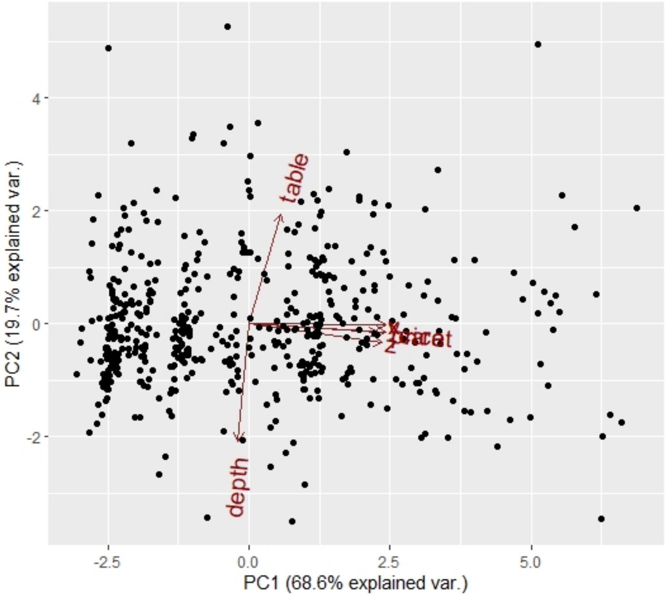


Figure 8: PCA Biplot

Principal Components Regression

- We conducted a Principal Components Regression with diamond price as the response variable
- The PCA excluding price was almost identical to the original PCA
- We were able to explain over 80% of the variation in price using just the first two principal components as predictors
- A more parsimonious model!

Summary of Models Predicting Diamond Price

Model	No. of Predictors	Adjusted R^2
Full Model	9	0.9198
Best Model	7	0.9198
Numeric Model	7	0.8592
Two PC	2	0.8092
All PC	6	0.8695

Leading Question 2 ... and the issues we encountered...

- The diamonds dataset includes 280 interactions between different levels of the categorical variables
- Our second leading question was to investigate if we could classify the diamonds data more simply using analytical techniques such as LDA and CA

Problems encountered

- Despite a correlation of 0.9216, ‘carat’ was not a great predictor of ‘price’
- LDA did not work with the full dataset