Group 11 Final Presentation

Tom Tribe, Ken MacIver, Jundi Yang, Mei Huang

2022-10-11

Group 11: Diamonds Dataset
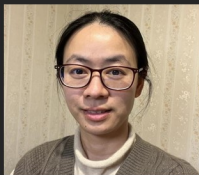
Group Members (photos)

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

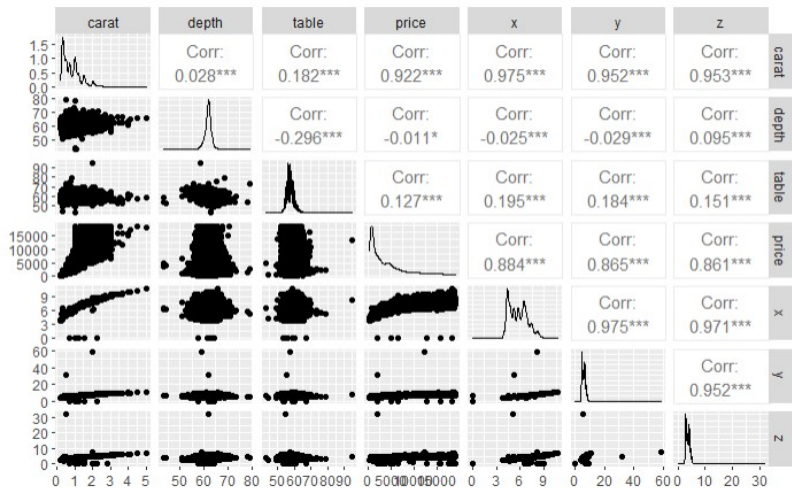(List adapted from list at kaggle.com).

Pairs Plot

Figure 1: Pairs plot
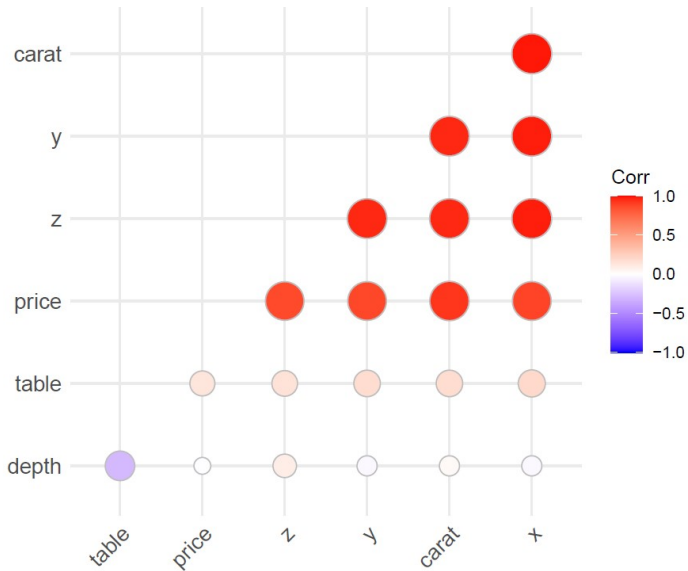
Correlation Plot

Figure 2: Correlation Plot
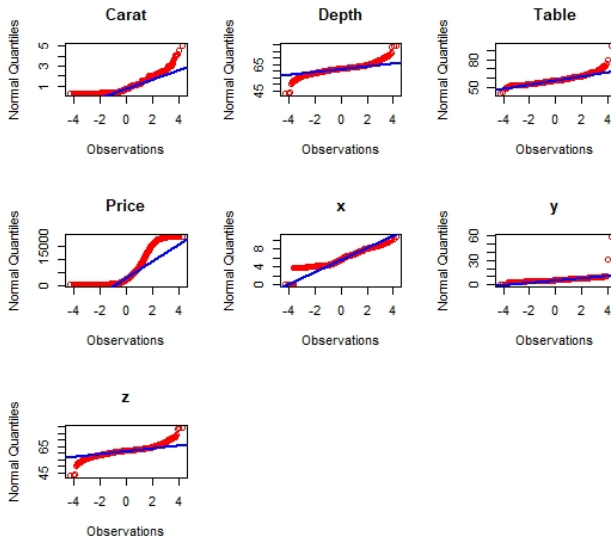
Normal QQ Plots
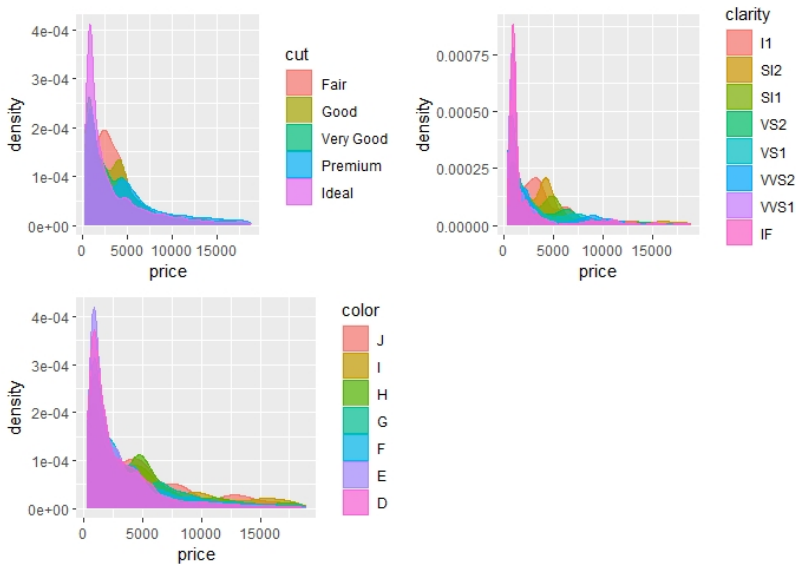

Figure 3: Normal QQ Plots

Price by Cateogrical

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 $\sim$ 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%
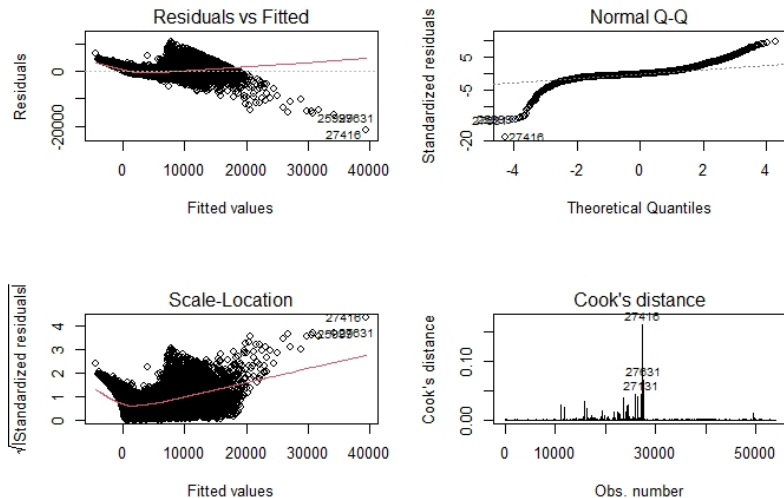
Regression Assumptions

Figure 5: Regression Diagnostics

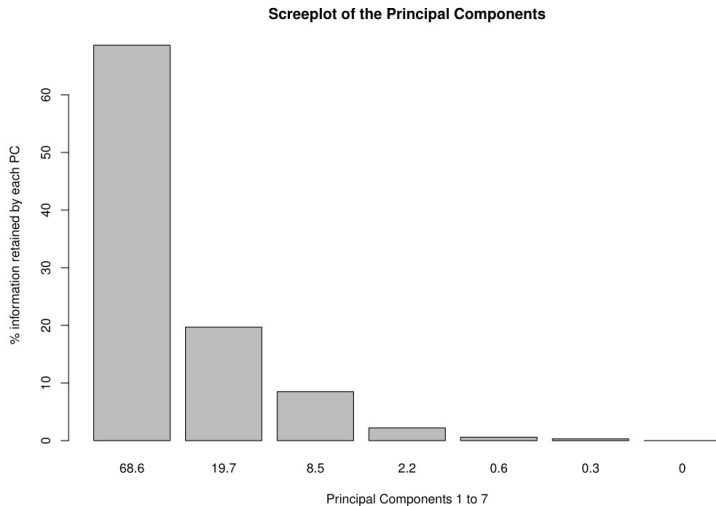# Principal Components Analysis: Screeplot

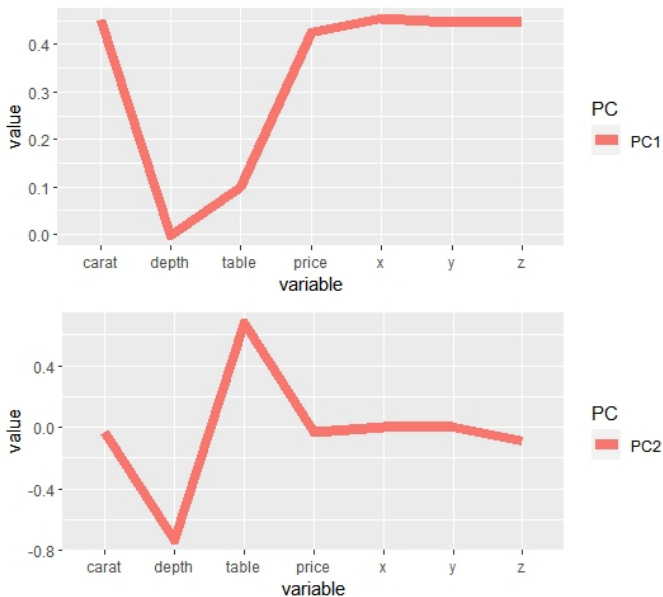Figure 6: PCA Screeplot

Principal Components Analysis: Eigenvectors
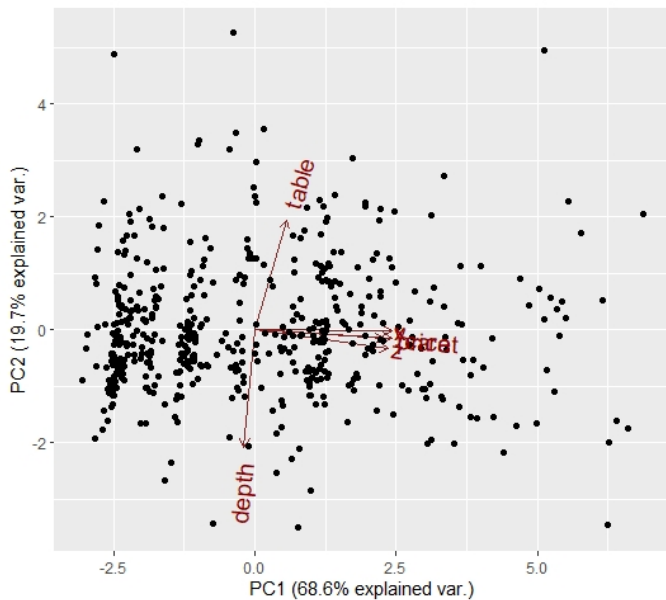

Figure 7: Plot of EigenVectors

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

▶ 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

LDA

Ken's Investigation (Change title - just didn't know what to call it)

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

Conclusion