

# STAT394 Group Project: Group 11

Tom Tribe, Ken MacIver, Jundi Yang, Mei Huang

2022-09-14

## Group Members

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

- ▶ Dataset: Diamonds
- ▶ 53940 rows (diamonds)
- ▶ ten variables
- ▶ very large dataset (approx 540,000 values)
- ▶ slow to process!

# The Variables

red font = categorical variable

- ▶ carat: the diamond's weight
- ▶ cut: a measure of quality
- ▶ color: a measure of colour quality
- ▶ clarity: a measure of clearness
- ▶ 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).

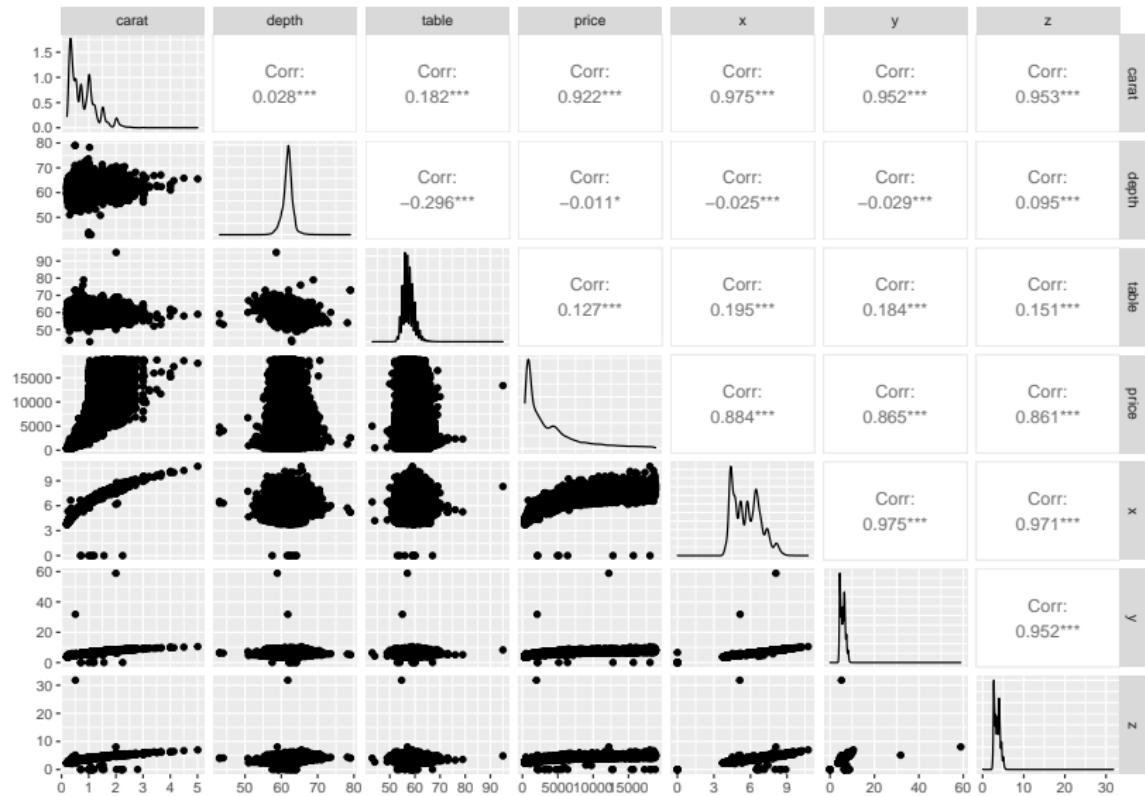
## The Response Variable

'Price' seemed to us to be the obvious response variable.

## Data Visualization (the dataset)

```
##   carat      cut color clarity depth table price     x
## 1  0.23    Ideal     E    SI2  61.5    55   326 3.95 3.9
## 2  0.21  Premium     E    SI1  59.8    61   326 3.89 3.8
## 3  0.23      Good     E    VS1  56.9    65   327 4.05 4.0
## 4  0.29  Premium     I    VS2  62.4    58   334 4.20 4.2
## 5  0.31      Good     J    SI2  63.3    58   335 4.34 4.3
```

# Data Visualisation (pairs plot)



## Other things of interest

The EDA revealed the following:

- ▶ some variables not Normally distributed
- ▶ long right tails due to a few very expensive diamonds
- ▶ some zero values

## Next Steps

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