Mini Project 2

INSTITUTE OF DATA WK 5

Diamonds are a Data Scientist's Best Friend…?

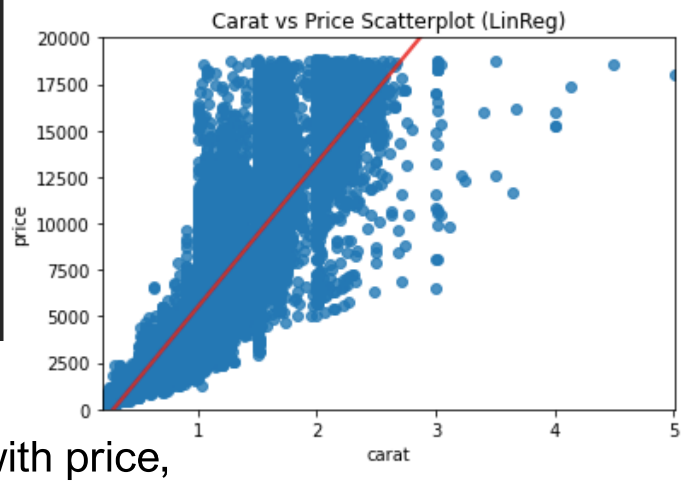
CONTEXT: A dataset containing just under 54 000 diamonds their relevant attributes.

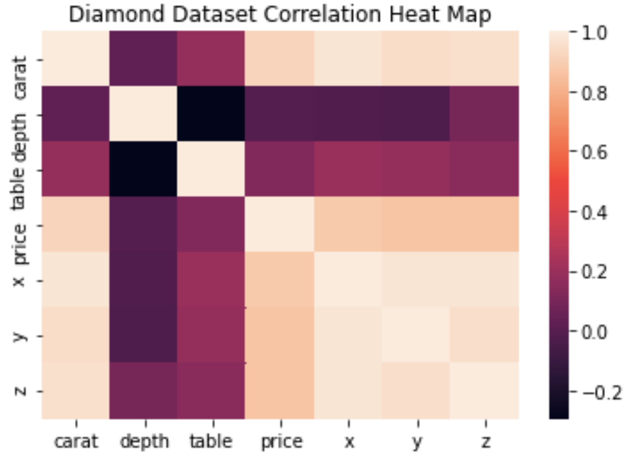
- Categorical variables:
  - Cut: describes cut quality (Fair, Good, Very Good, Premium, Ideal)
  - Color: diamond colour (D being the best, J is the worst)
  - Clarity: how clear the diamond is, or absence of inclusions/ blemishes - I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
- Numerical variables
  - Carat: weight of diamond (0.2-5.01)
  - Depth: The Height of a Diamond, measured from the Culet to the table, divided by its average Girdle Diameter.
  - Table: width of top of diamond relative to widest point (The Width of the Diamond's Table expressed as a Percentage of its Average Diameter: 43-95)
  - Price: price of diamond (USD326-USD18,823)
  - X: length in mm (0-10.74), Y: width in mm (0-58.9), Z: depth in mm (0-31.8)

PROBLEM:

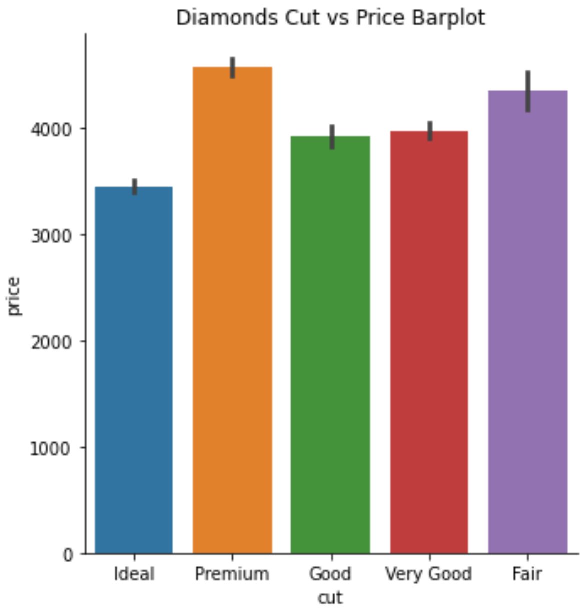
- Which features are related to price?
- Can there be a predictive model established?

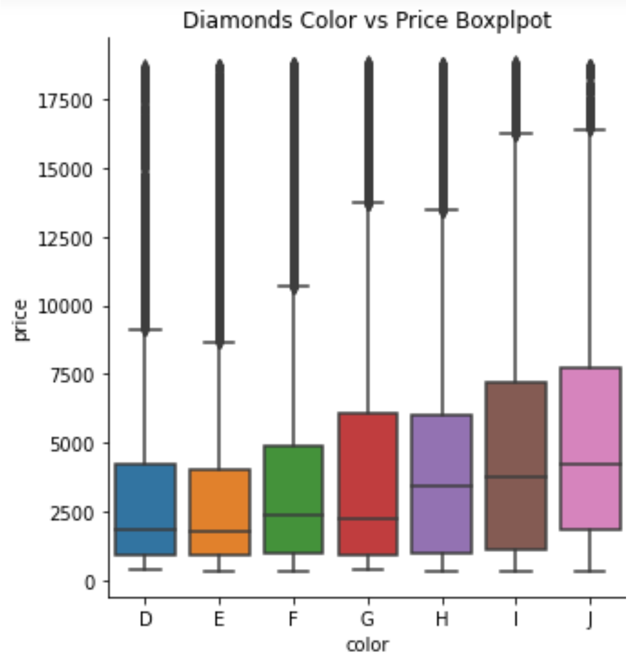
EDA Results





1. As a variable, Carat had the strongest correlation with price, as computed in a correlation matrix and constructed scatterplot
2. Dimensions X (length), Y (width) and Z (depth) were alsc highly correlated to price and, subsequently, each other. mutual correlations leads to the prospect of merging the into a singular feature, volume, in upcoming stages of the model.
3. Unexpectedly, categorical variables had not only a weak also inverse relationship with price. As the quality increa in each of these factors, price decreased. My potential hypothesis for this would be that as carat (weight) increa along with the equivalent price, the quality of the cut, col and clarity could decrease.





Explanation of DataBase Type

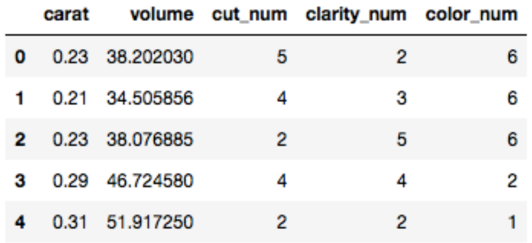
DATABASE: SQLite

- Definition: a standard programming language that commonly used in relational database management systems.
  - SQL: Structured Query Language
  - Enables users to access and manipulate databases.
  - It doesn’t function on a separate server
  - SQL is saved in a portable, single file
  - Highly efficient in how it can query, manipulate and aggregate data into useable information
  - Used here for further data exploration which will assist later in modelling and feature engineering

Feature Engineering and Machine Learning Model

FEATURE ENGINEERING: aggregated dimension variables, converted categorical to numerical.

- Output (target variable):
  - Price of diamond in USD
- Inputs (Feature vector): based on correlations and factors of interest
  - I chose the features which are strongly correlated with the target variable (price), namely, carat and volume.
  - Out of interest (although they seem inversely and weakly correlated) I will also be exploring the categorical attributes (color, clarity and cut).
  - The resulting feature matrix is as follows (head):



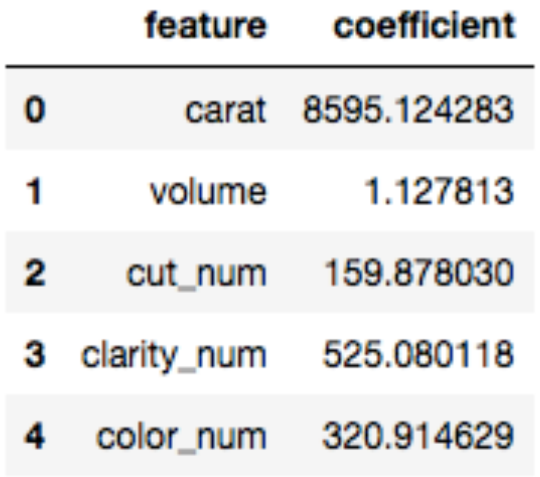
	carat	volume	cut_num	clarity_num	color_num
0	0.23	38.202030	5	2	6
1	0.21	34.505856	4	3	6
2	0.23	38.076885	2	5	6
3	0.29	46.724580	4	4	2
4	0.31	51.917250	2	2	1

ML Model: Supervised learning through Multiple Linear Regression

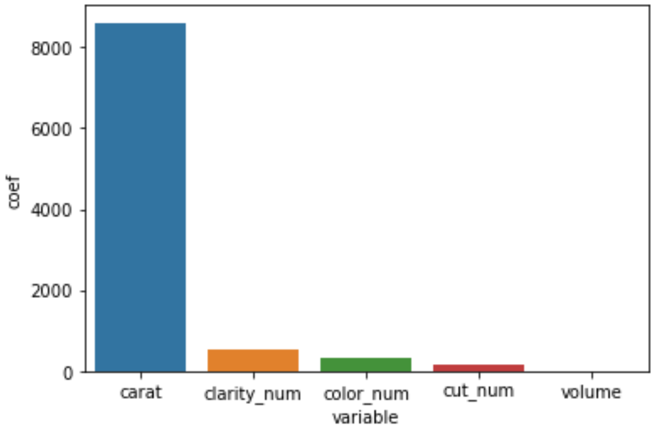
- Regression is a statistical method that aids in quantifying the relationship between the correlated variables. It involves estimating the coefficient of the independent variable (or multiple, as is in this feature vector) and then measuring the reliability of the estimated coefficient.

Results from training and testing phases of ML model

1. LINEAR REGRESSION: Resulting coefficients of predictor variables

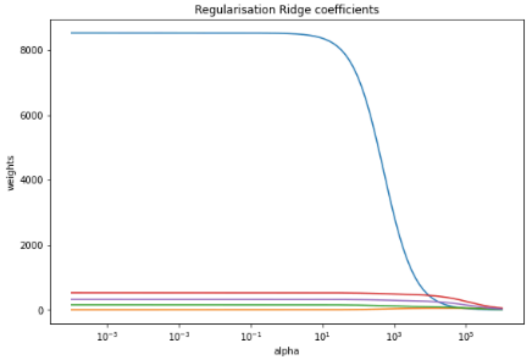


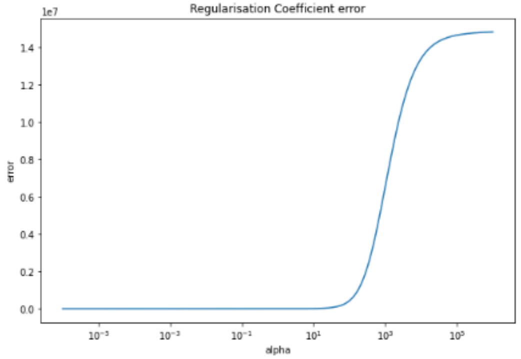
	feature	coefficient
0	carat	8595.124283
1	volume	1.127813
2	cut_num	159.878030
3	clarity_num	525.080118
4	color_num	320.914629

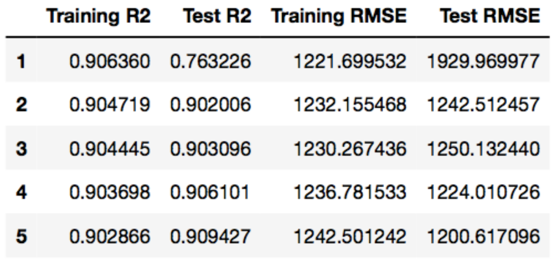


- Evident that carat had the highest coefficient in relation to the target variable, price. This was hypothesized through initial EDA and and seen in the correlation matrix. Yes, I am also aware that the aggregated volume variable has failed in a technical sense. It is expected to have a much more significant correlation and equivalent coefficient. Categorical variables were expected to have much lower coefficients.

2. RIDGE REGRESSION + Cross Validation: Visual representation displaying:
Ridge coefficients as a function of the regularisation and
Coefficient error as a function of the regularisation







	Training R2	Test R2	Training RMSE	Test RMSE
1	0.906360	0.763226	1221.699532	1929.969977
2	0.904719	0.902006	1232.155468	1242.512457
3	0.904445	0.903096	1230.267436	1250.132440
4	0.903698	0.906101	1236.781533	1224.010726
5	0.902866	0.909427	1242.501242	1200.617096