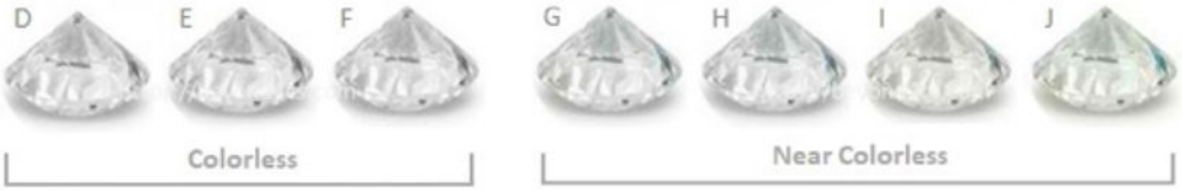
Diamonds Price Prediction and Feature Analysis

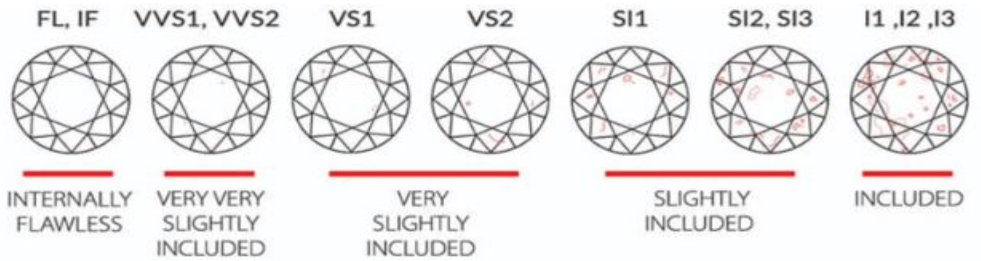
By Zhi Chen

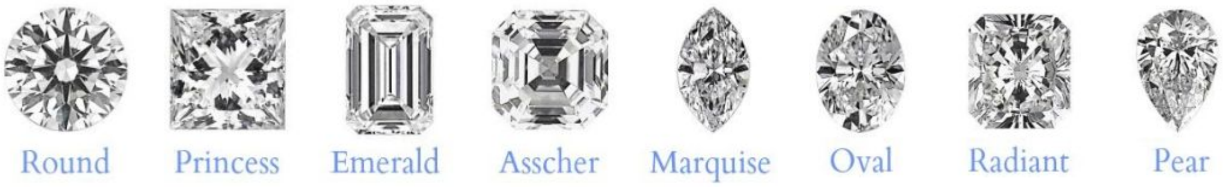
Various Features

- Carat Weight
- Clarity
- Cut
- Color
- Depth
- Table
- L/W Ratio
- Polish
- Symmetry
- Fluorescence

To establish a model using Kaggle dataset to predict prices



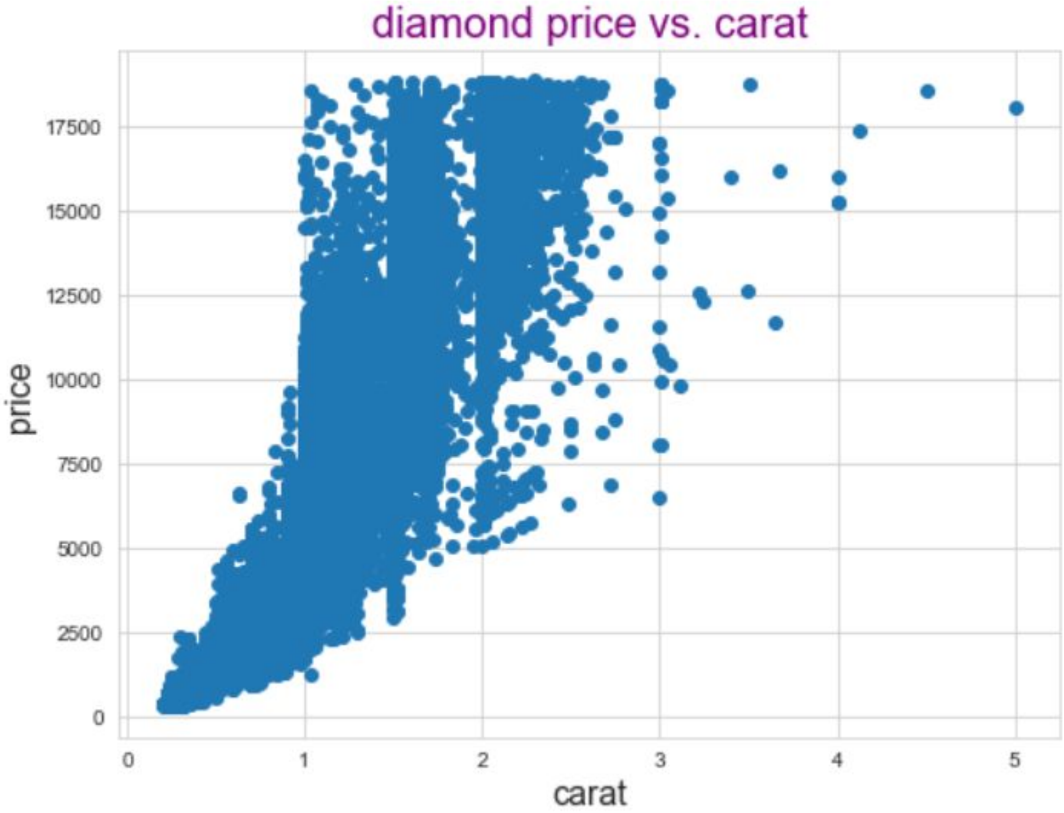




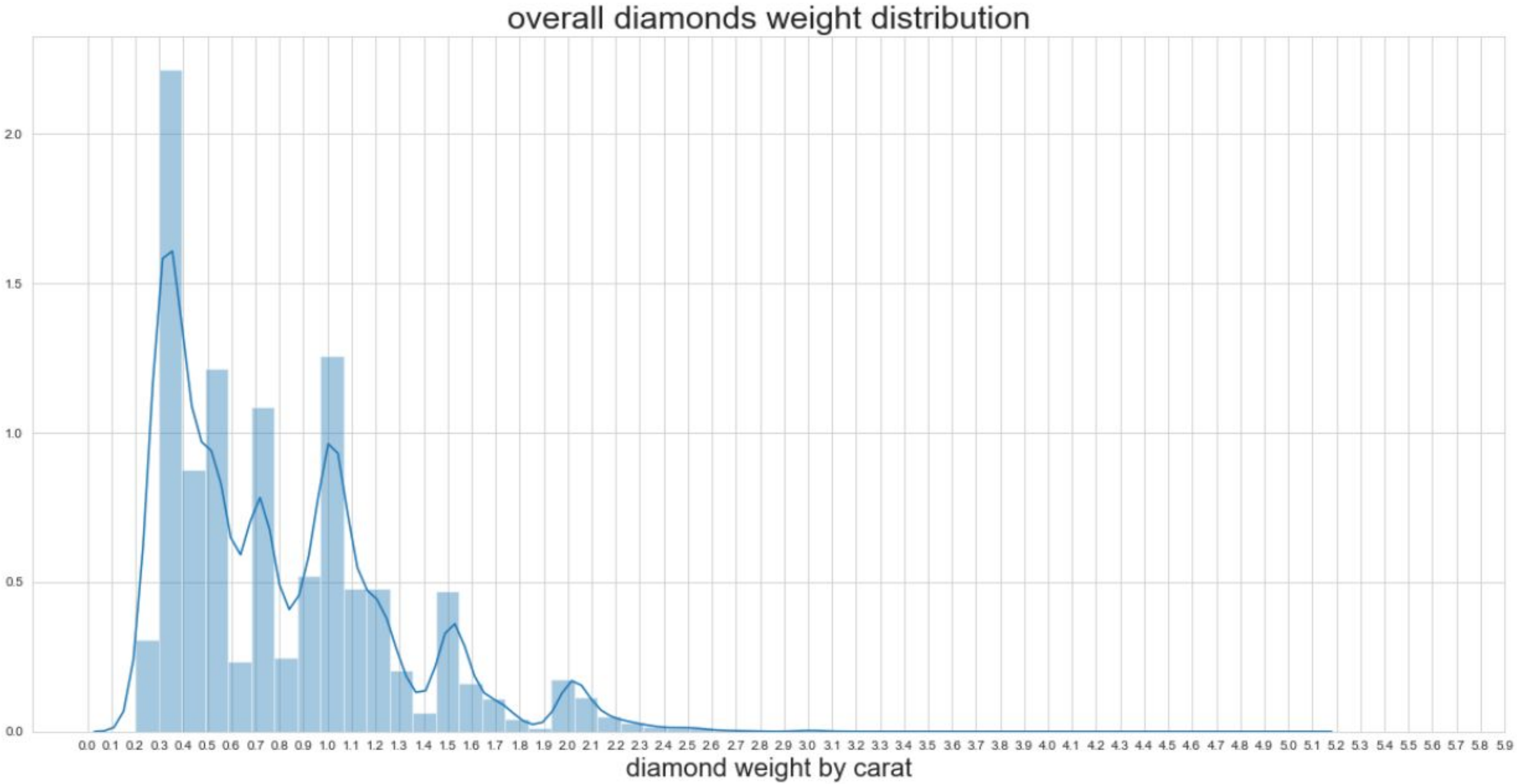
Price and Carat Weight

- Price is determined by various features

- But Generally, price increase exponentially with carat weight



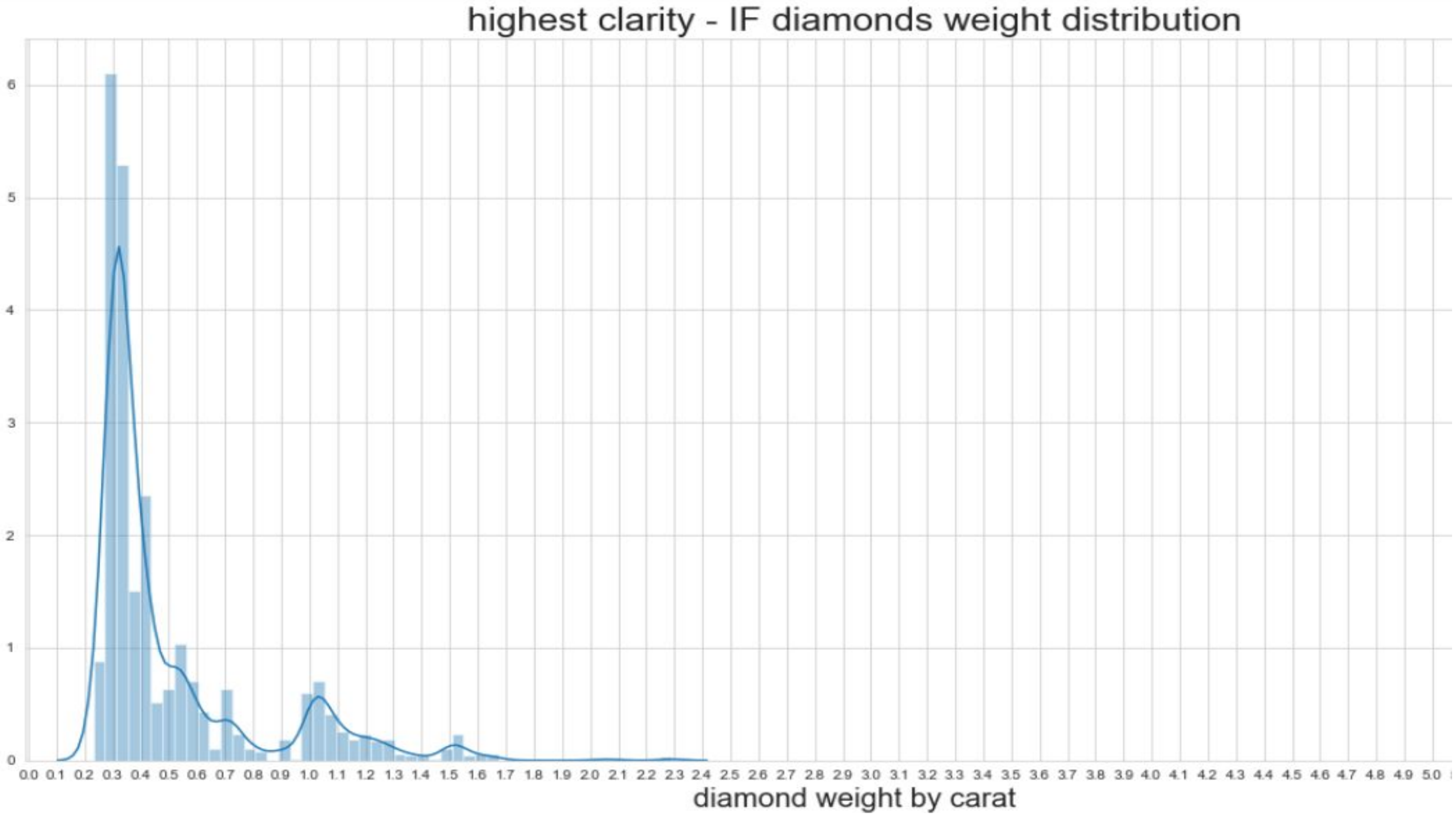
Carat Weight Distribution



Carat Weight Distribution

Highest clarity ones are in small value range

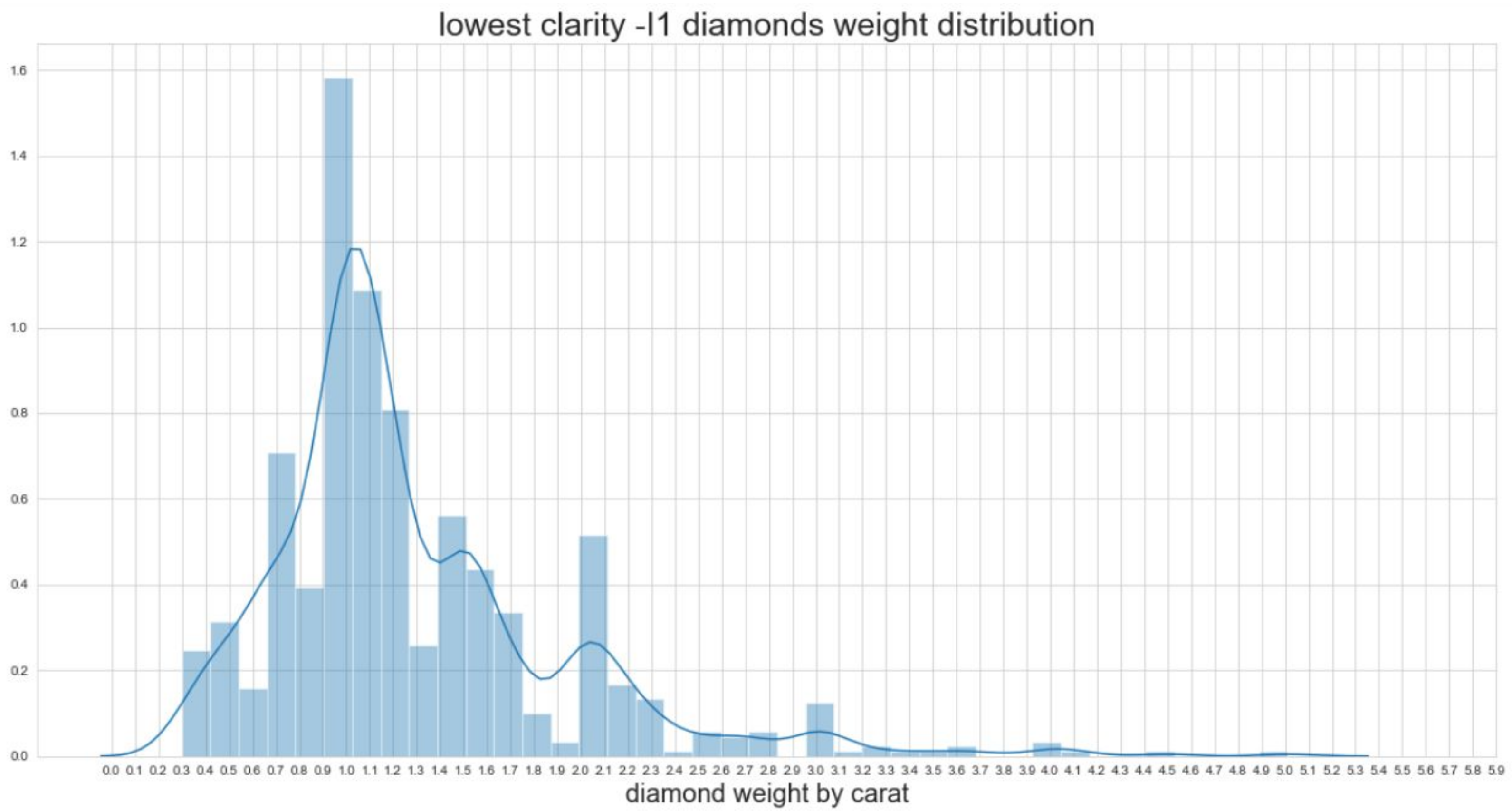
Almost no big ones in highest clarity region



Carat Weight Distribution

It is easier to get bigger diamonds in low clarity level

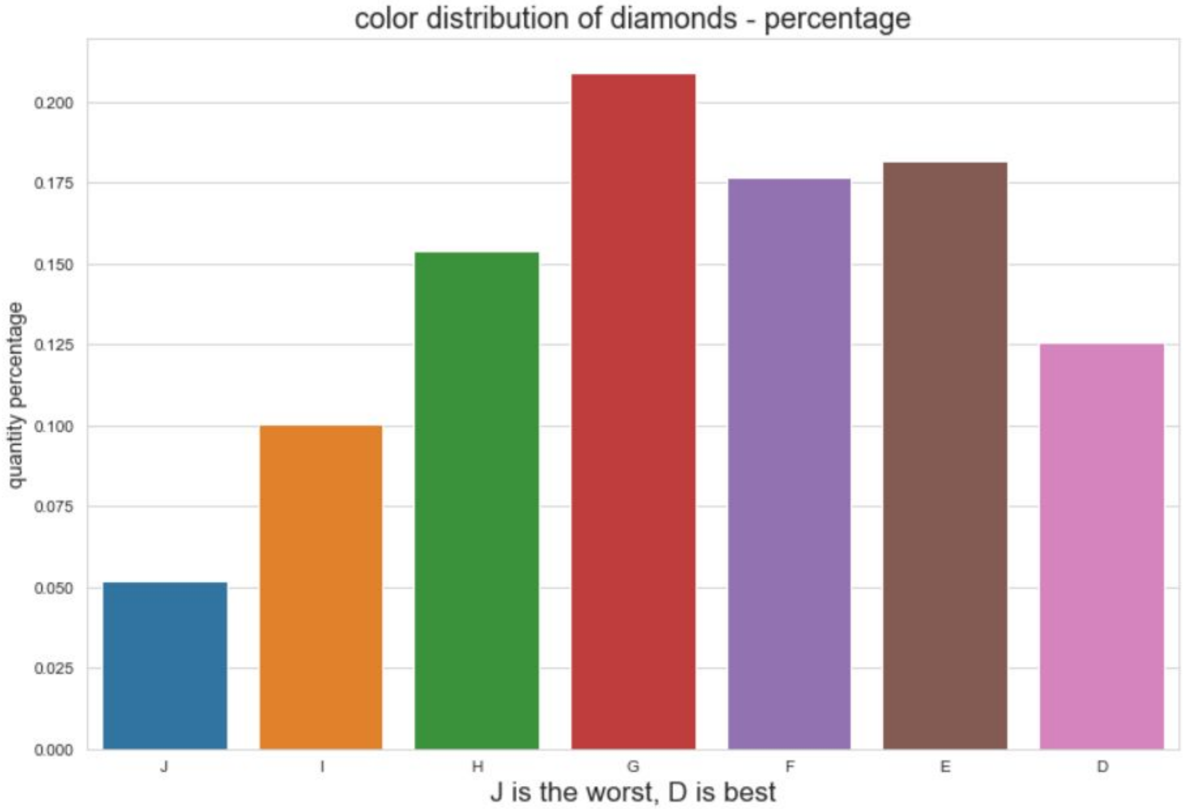
Explained by solid state material science



Color Distribution

- There are not many diamonds truly colorless

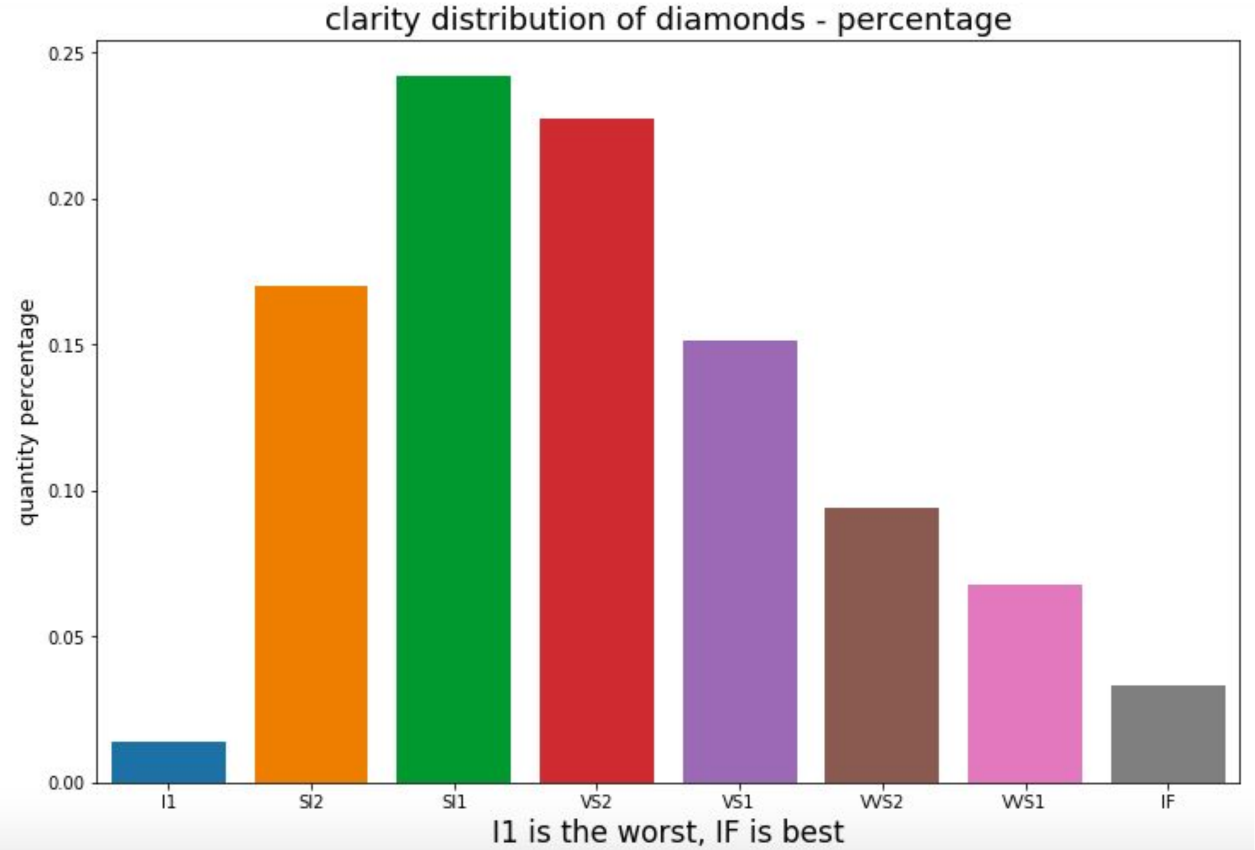
- Most of diamonds are in the middle color range, sales people will recommend you to purchase ‘G’, ‘F’ and ‘H’ if you don’t want to pay too much but you still want larger



Clarity Distribution

- There are not many diamonds at top clarity or bottom clarity (optically perfect)

- Similar to color distribution, the middle level consist most of diamonds, but leaning to lower end



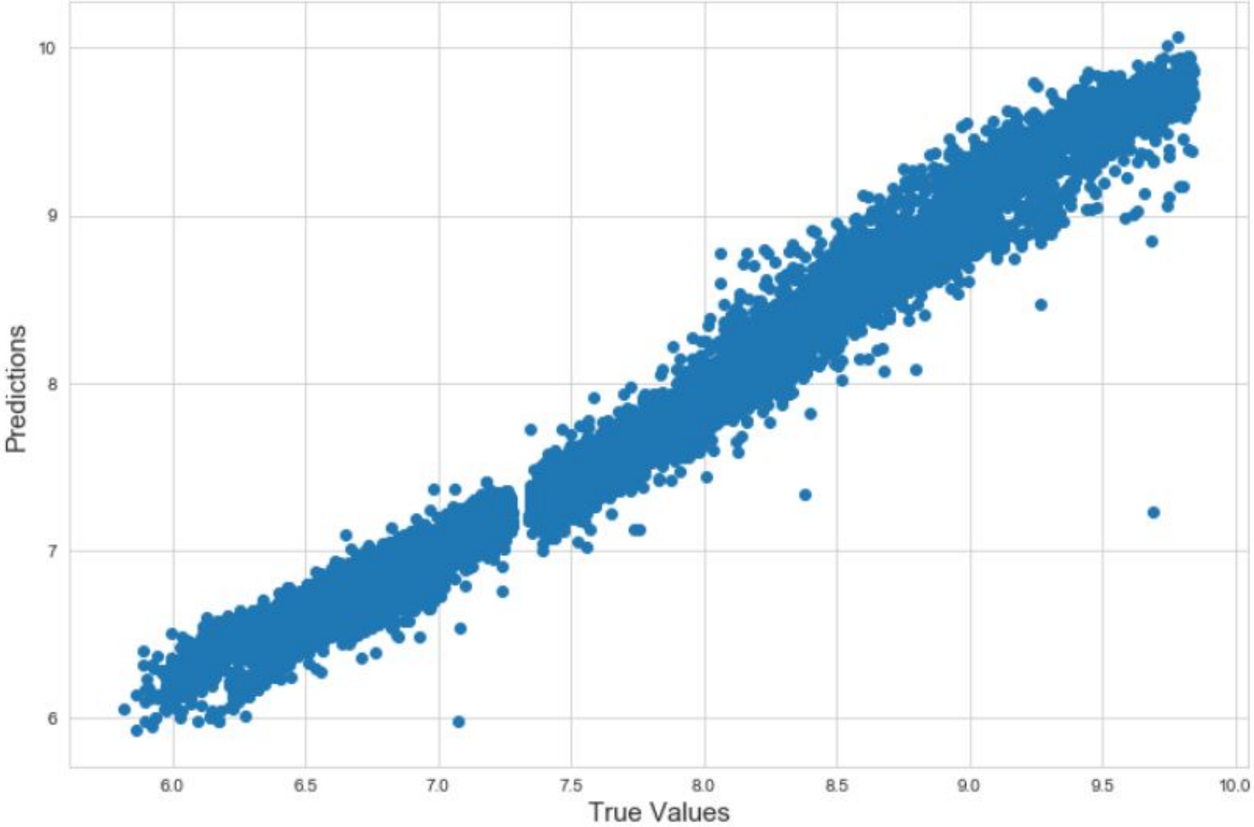
Hypothesis Testing

- ***Consistent with previous EDA***
- ***Chi2: there is association/relationship between the diamond clarity levels categories and the color levels categories (they are not independent)***
- ***Anova: there is statistically significant difference in carat weight of diamonds with different cut levels, clarity levels, color levels***

Prediction model with logarithmic target values

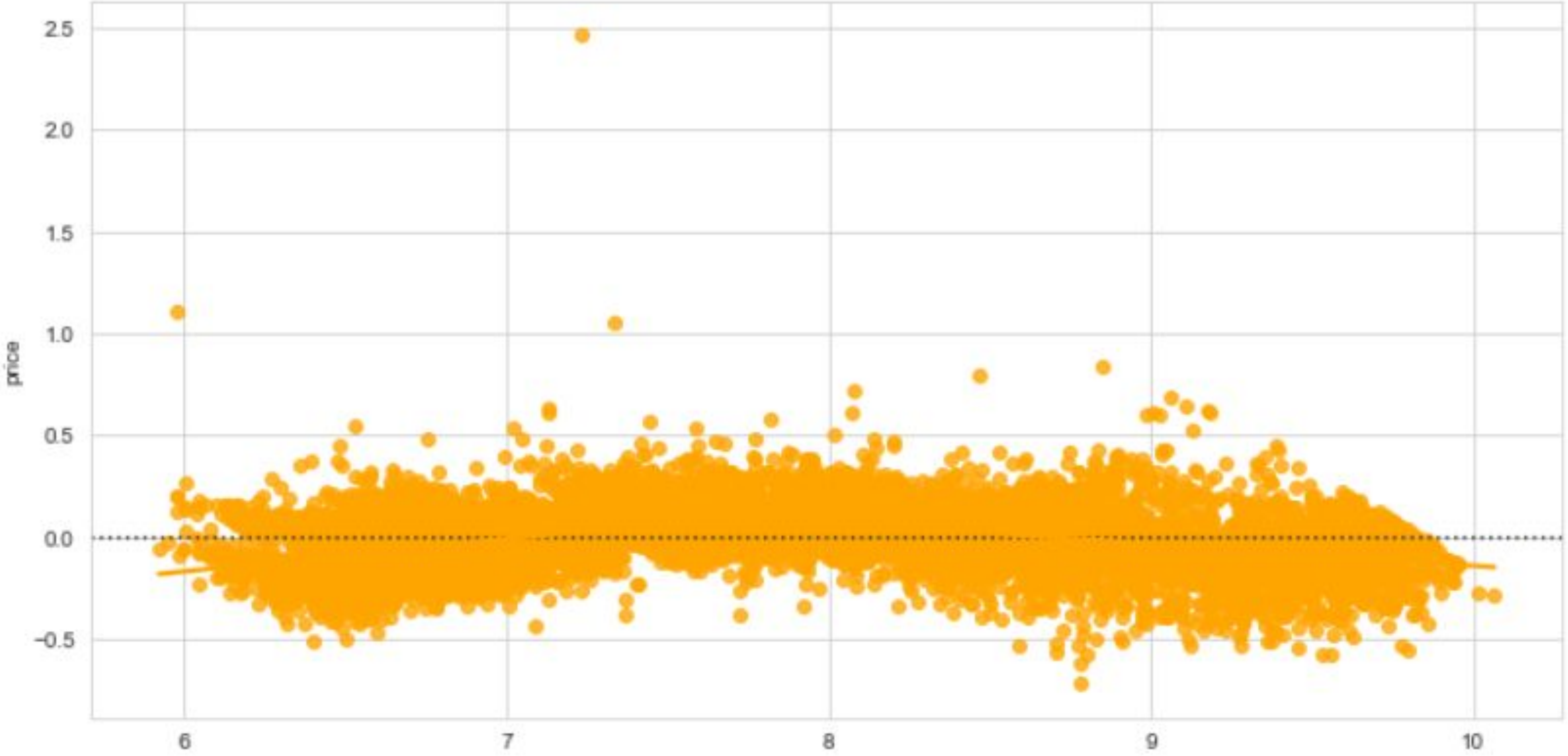
- Linear Regression
- R^2: 0.977
- RSME: 0.155
- Lasso model also gives good result, but not as good as linear

- With all the above features, diamonds prices can be well estimated, with little error



Prediction model with logarithmic target values

The residuals plot

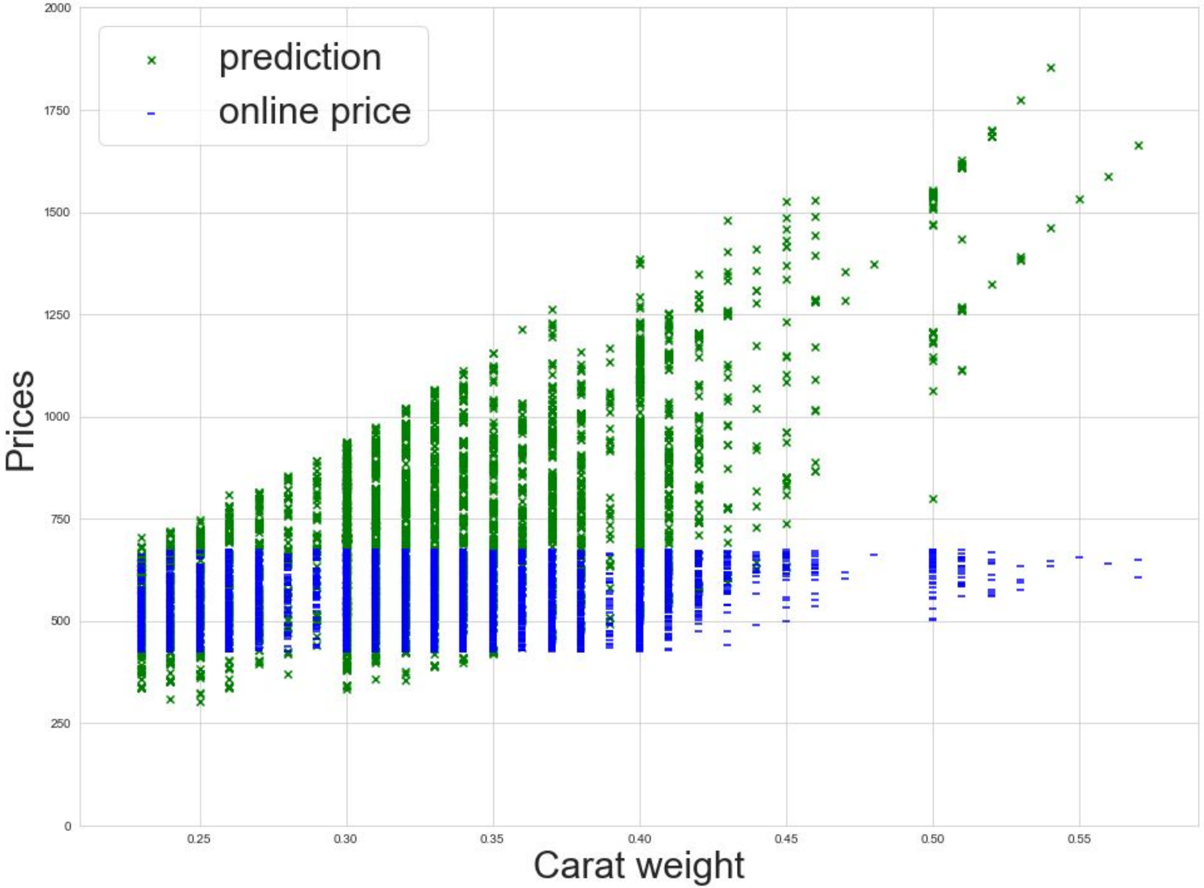


Price prediction and compare with online sales prices

- Use the model generated from Kaggle dataset to predict prices of diamonds on bluenile.com

- Online sales prices are averagely cheaper by 20%, gap is larger for big ones

- Different data resources may not reflect each other very well



Conclusions

- With all features, you can do accurate prediction of diamond prices

- Model established based on one resource may not predict other resource well

- Limitations: may need to further investigate different sources of data before prediction