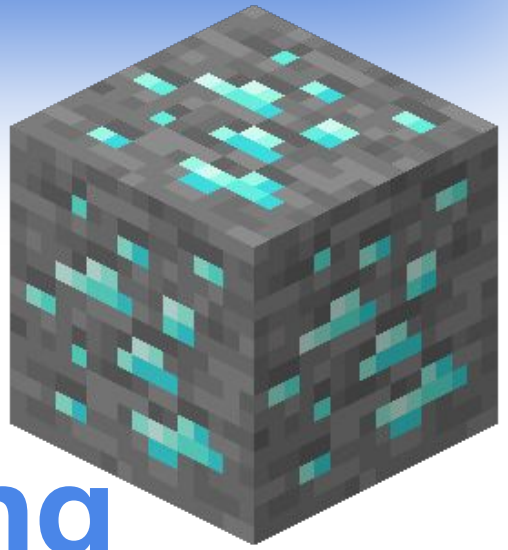
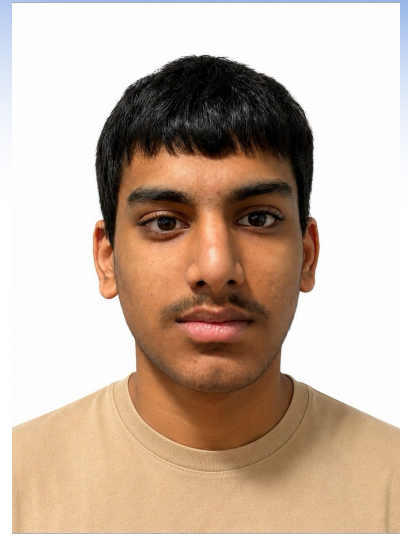


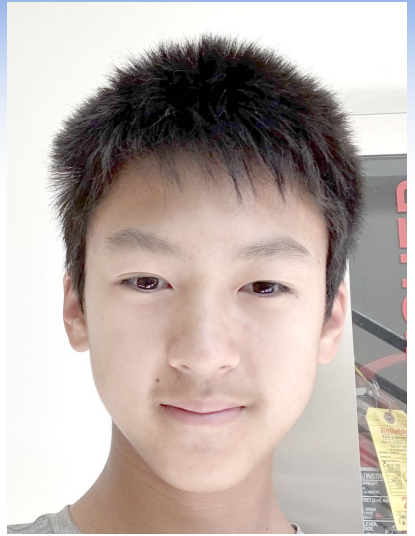
# Predictive Diamond Pricing

By Dheeraj Garg, William Lu, Lucas Su, Eric Yan

# The Team



Dheeraj Garg



William Lu

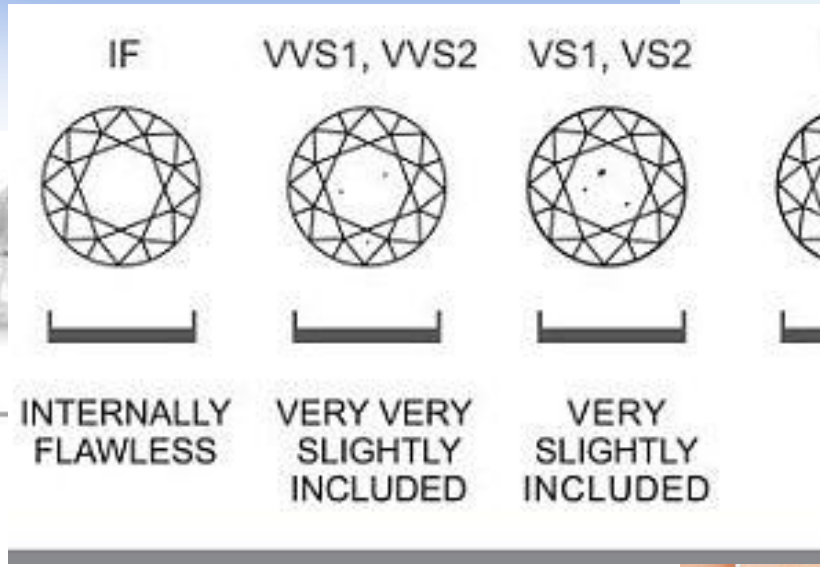
Lucas Su


Eric Yan

# Project Statement

- Diamonds are expensive
- Different factors of diamonds affect price
- Question: “Which of the 4 C’s (carat, cut, color, and clarity) affects diamond retail prices the most?”
- Rank the 4 C’s by importance

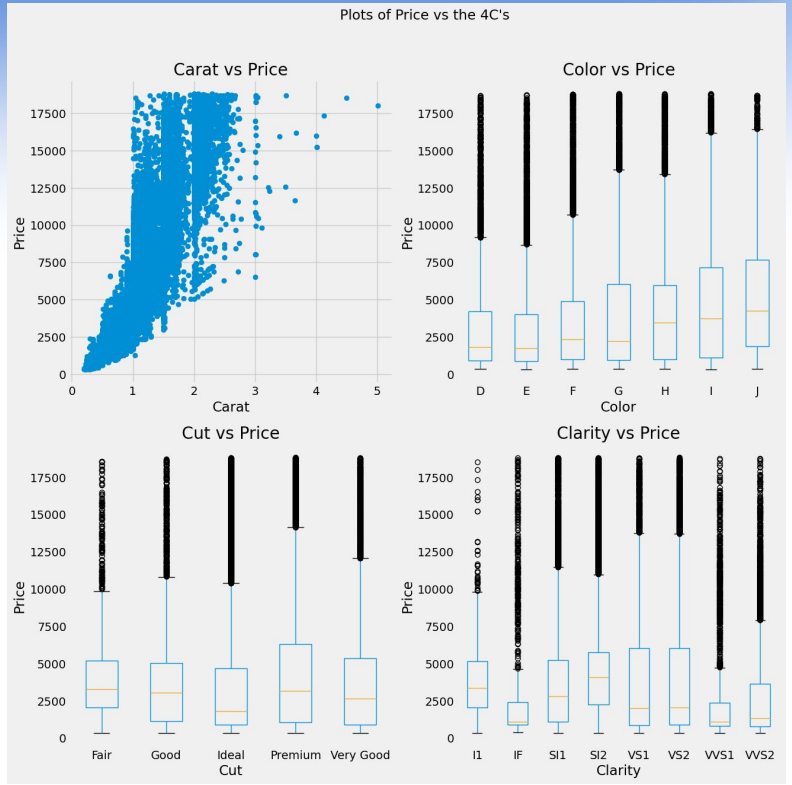
# 4 C’s


IF
VVS1, VVS2
VS1, VS2
INTERNALLY FLAWLESS
VERY VERY SLIGHTLY INCLUDED
VERY SLIGHTLY INCLUDED


Ideal
Premium
Very Good
Good
Fair

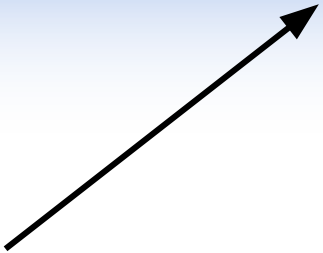
# The Data

- Original: 53940 rows, 11 columns
- Dataset from Kaggle, originated from ggplot2
- Differences in categorical data unevenly distributed
  - Used intervals to make categorical C's numerical
- Other information: what information it contains, etc.
  - Carat, cut, color, clarity
  - Price



# The Analytics

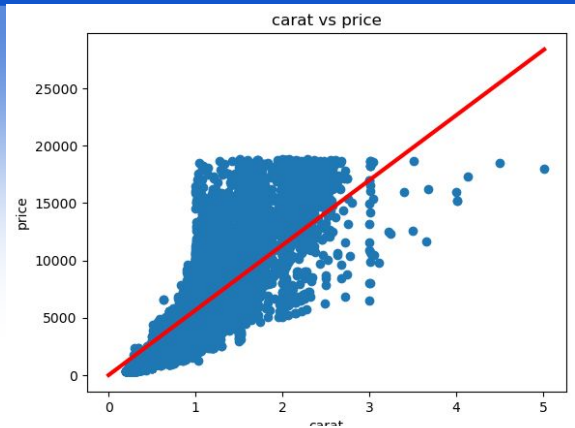
- Considered:
  - ~~kNN~~
  - Linear Regression
  - Random Forest
  - Gradient Boosting



- Linear Regression Chosen:
  - Exact coefficients for each feature
  - Easy comparisons
  - Graph has a clear trend.
- Evaluation metrics
  - MSE (Mean Squared Error)
  - $R^2$ (Variability explanation)

# Polynomial Linear Regression

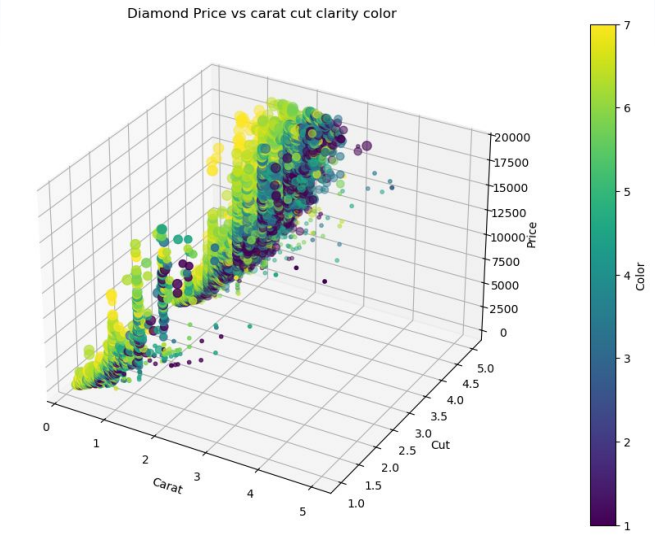
- Observed that graphs curved upwards, used polynomial linear regression
- Improved the correlation, suggesting an accelerated relationship




# Multivariable Linear Regression

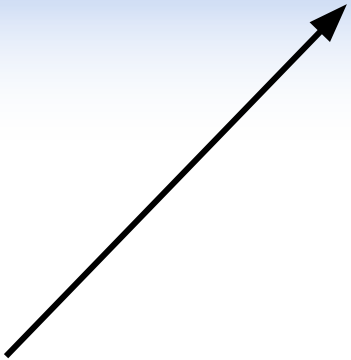
$$\text{Price} = -5707.877 + 3773.190(\text{carat}^2) + 145.228(\text{cut}) + 358.693(\text{color}) + 613.223(\text{clarity})$$

- Fits a line to training data
- 80/20 train/test split
- Train $R^2$ score = 0.841
- Test $R^2$ score = 0.828286



# The Analytics

- Considered:
  - ~~kNN~~
  - Linear Regression
  - Random Forest
  - Gradient Boosting



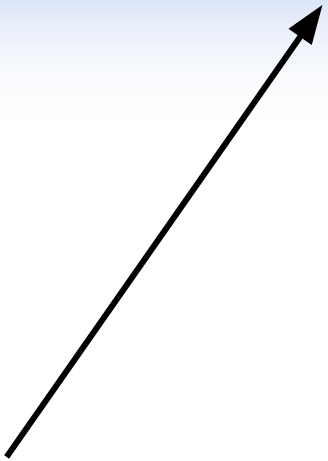
- Random Forest Chosen:
  - Capture nonlinearities
  - Interactions between features
  - Resists overfitting
- Evaluation metrics
  - MSE
  - $R^2$

# Random Forest

- Method
  - Ba
  - Tr
  - Av
- Train R
- Test $R^2$

# The Analytics

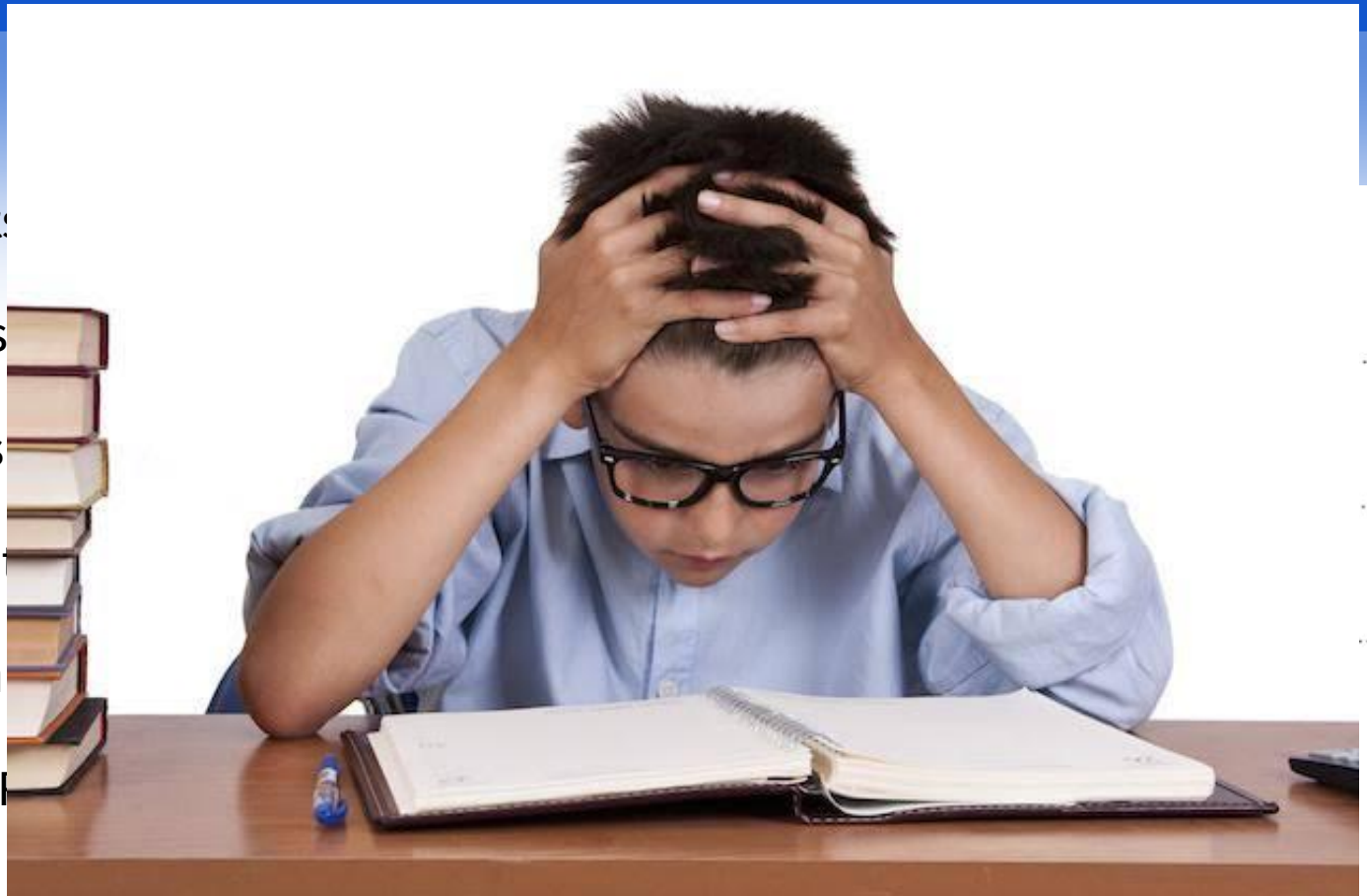
- Considered:
  - ~~kNN~~
  - Linear Regression
  - Random Forest
  - Gradient Boosting

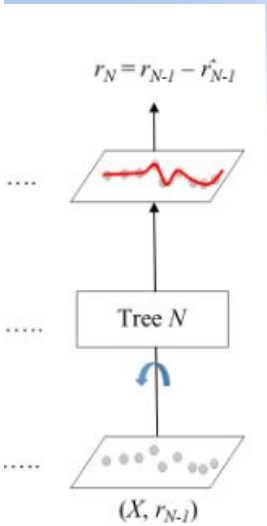


- Gradient Boosting Chosen:
  - Same perks as Random Forests
  - Corrects previous mistakes by trees
- Evaluation metrics
  - MSE
  - $R^2$

# Gradient Boosting

- Starts
- Meas
- Adds
- 500 i
- Train
- Test





# Metrics

| | Model | Train MSE | Test MSE | Train $R^2$ | Test $R^2$ |
|---|---|---|---|---|---|
| **0** | Linear Regression | 2.533e+06 | 2.714e+06 | 0.841 | 0.828 |
| **1** | Gradient Boosting | 1.866e+05 | 3.021e+05 | 0.988 | 0.981 |
| **2** | Random Forest | 2.586e+05 | 3.114e+05 | 0.984 | 0.980 |

# Conclusion/Results

- Gradient Boosting Model
- Difficulties:
  - Standardizing categorical values
- Importance: Carat, Clarity, Color, Cut
- Potential future work
  - Include region / other factors
  - Improve model accuracy

# Thank You For Listening
# Any Questions?