Project 2, STAT/CS 190

Data and Scope Proposal

**Data:**

The data I chose is called Diamonds. Source from Kaggle.com, <https://www.kaggle.com/shivam2503/diamonds>

The dataset has 53940 rows and 11 columns with 1 column being the index number.

The data have different variable for each diamond and the price of it.

Content

**“price**” Price in US dollars (numeric)

**“carat”** weight of the diamond (numeric)

**“cut”** quality of the cut (Fair, Good, Very Good, Premium, Ideal) (Categorical)

**“color”** diamond color, from J (worst) to D (best) (Categorical)

**“clarity”** a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)) (Categorical)

**“x”** length in mm (numeric)

**“y”** width in mm (numeric)

**“z”** depth in mm (numeric)

**Scope:**

My main goal in this dataset is to predict the price of the diamond based on all variables.

For jewelry company trading, evaluating the diamond. Personally I care about the diamond as I love to get one in the future and get the best with the best price.

The methods I have in mind while looking at the dataset, will be using modeling inverse gaussian and gamma. Using random forest for machine learning. The dataset will be a good visualization dataset.

**Deliverable:**

For video, I plan to work on an interactive dashboard, where I can sub in different variable to predict the price. Have a couple visualization graphs shown. Backup plan is slideshow.