# Project 2: Linear Regression

### 36-600

### Fall 2022

The goal of this project is to use linear regression to model the price of diamonds as a function of other descriptive variables.

Your `HTML` file (generated by knitting an `R Markdown` source file) should be uploaded to Canvas by Tuesday, November 1st, at 11:59 PM. As was the case for Project 1, there is no limit on the length here, but concision is a virtue.

# Data

You will examine the dataset `diamonds.csv`, which you will find in the `DATA` directory on `Canvas`. (You have seen this before, when in Week 5 you input these data into the $K$-means algorithm. You will *not* be doing unsupervised learning here, though.)

The response variable is `price`. Your goal is prediction, and not inference.

The predictor variables are a mix of quantitative and factor variables:

| name | description |
| --- | --- |
| carat | diamond weight (1 carat ~ 200 milligrams) |
| cut | graded quality (Fair, Good, Very Good, Premium, Ideal) |
| color | graded color (J is worst, to D, which is best) |
| clarity | graded measurement of clarity (I1, SI1, SI2, VS1, VS2, VVS1, VVS2, IF, in that order from worst to best) |
| x | length of diamond (millimeters) |
| y | width of diamond (millimeters) |
| z | depth/height of diamond (millimeters) |
| table | width of top part of diamond relative to widest point (percentage) |
| depth | depth of top part of diamond from the widest point, relative to total depth (percentage) |

The file contains an unimportant variable that you will need to remove.

# Expectations

Your report should include the following elements:

- A description of the data (sample size, number of variables).

- Concise EDA, including the identification and removal of proposed outliers (if there are any) and, potentially, the transformations of the response variable and/or (some) predictor variables that are highly skew. (Note: always try transformations first before identifying and removing outliers…that lonely data point in a skew distribution may look fine after a transformation is performed.) Also, create a correlation plot and comment on the possibility of multicollinearity in the predictor variables (while mentally noting that your goal here is prediction, so multicollinearity is OK). (One last note: the number of data is large, so if you create scatter plots, randomly sample some much smaller number of points to plot.)
- A description of how you split the data into training and test sets.
- An analysis of the full dataset with linear regression, with comments on the output (does the fit appear to be good?). Provide the mean-squared error and the predicted response vs. observed response diagnostic plot. Comment on the value of adjusted $R^2$: is the linear model useful, or might other models perform better? Also, realize that the response variable needs to be transformed if and only if (a) the *residuals* of the fit are not normally distributed *and* (b) you wish to do inference using the hypothesis test output from `lm()`. To plot the residuals, find the difference between the observed test-set response values and the predicted test-set response values, and make a histogram: does this plot look normal? Here, the goal is prediction of price, not inference, but you may still find that a transformation may (or may not!) make for better model predictions.
- A best-subset-selection analysis of the dataset. Which predictor variables are important for predicting diamond prices? If predictor variables are removed from the full set, compute the new mean-squared error and compare it to the MSE for the full set of predictors.

Done. Remember, again, you are not publishing a journal paper about this analysis, nor are you charged with finding some new or interesting result here. It's a simple straight-up analysis—treat it as such!

For those desiring extra credit (which won't actually be given, but still):

- Do a simple PCA analysis on the predictors with the goal of determining the "true" dimensionality of the predictor variables. No need to map original variables to PCs or do PC regression; I'd just be curious to know if the data lay on, e.g., a subsurface within the native space. This is extra credit because we would not normally go down this path unless we were performing an inferential analysis and our data has a high level of multicollinearity that needs to be mitigated.