Student Name: Tevin Achong
Student ID: 816000026
Student Name: Name 2
Student ID: ID2
Student Name: Name 3
Student ID: ID3
Course Code: COMP3608
Course Title: Intelligent Systems
Assignment: 2

March 21st, 2020

Part 3 - Encoding Features

Since most machine learning algorithms (e.g. Linear Regression) require that input data be numerical, we would want to represent all our data in the dataset numerically so that we could apply an appropriate algorithm to it.

The following features are quantitative, i.e. regular numerical data. As such, they will be represented as decimal values in our feature vector:

- **price** - price in US dollars
- **carat** - weight of the diamond
- **x** - length in mm
- **y** - width in mm
- **z** - depth in mm
- **depth** - total depth percentage
- **table** - width of top of diamond relative to widest point

The following features are categorical, i.e. label data. Furthermore, they are each ordinal, meaning that the ordering of the labels is significant and cannot be ignored. As such, we will use **ordinal encoding** to represent them:

- **cut** - quality of the cut
  - Fair - 1
  - Good - 2
  - Very Good - 3
  - Premium - 4
  - Ideal - 5
- **color** - diamond color
  - J - 1
  - I - 2
  - H - 3
  - G - 4
  - F - 5
  - E - 6
  - D - 7
- **clarity** - a measurement of how clear the diamond is
  - I1 - 1
  - SI2 - 2

  - SI1 - 3
  - VS2 - 4
  - VS1 - 5
  - VVS2 - 6
  - VVS1 - 7
  - IF - 8

Each possible value for each ordinal feature above is given a value relative to the other possible values for that specific feature.

For example, a diamond

- **carat** - 0.23
- **cut** - Ideal
- **color** - E
- **clarity** - SI2
- **depth** - 61.5
- **table** - 55
- **price** - 326
- **x** - 3.95
- **y** - 3.98
- **z** - 2.43

will have feature vector $[0.23, 5, 6, 2, 61.5, 55, 326, 3.95, 3.98, 2.43]$