Capstone 1: Health Insurance Marketplace

**Data Wrangling Techniques**

Generally we do not get the data in the formate best suited for our problems. We always need to do some data wrangling to bring the data to the desired format. Following are some common data problems we may come across when dealing with datasets.

* Inconsistent column names
* Missing data (NaN values)
* Outliers
* Duplicate rows

In my dataset, fortunately I have to deal with missing values(NaN values). And hope to see some outliers as well.

**To check for missing data points I have used the following method.**

1. Read the csv files into pandas dataframe.
2. use [dataframe.info](http://dataframe.info)() method to find if there is any missing data. In some cases the data[frame.info](http://frame.info)() method doesn’t show the desire output for these kind of dataframes, I have used isnull() method to find if there is any missing values in the columns.

**We have following options to deal with the missing values:**

1. fillna() method: There are lot of options available with the fillna() method to deal with the missing values. For example:

fillna(0) will replace all the NaN values with 0(zero)

fillna(method = ‘ffill’) will replace NaN with the preceding value in the column. There are option to control this ffill option.

fillna(method=‘bfill’) will replace NaN with the following value in the column.

2. dropna() method: to drop the entire data point if there are missing values in all the columns.

**Outliers:**

Yet to find any outliers, but if I get any outliers while working on the project will surely be considered for further investigation.