**Capstone Project I – Data Wrangling**

**Data Mining:**

Data is obtained from United States Cancer Statistics, 1999-2013 provided at the Centers for Disease Control and Prevention.

<https://wonder.cdc.gov/CancerMort-v2013.html>

The data included various features like cancer sites, states, year, sex, age, population and race , however, due to the constraints on quantity of data that could be downloaded from the website at once, the whole data was downloaded into 60 text files. Also, due to constraints on number of features that could be selected to appear in text file, the female and male data was downloaded in separately. So, in total 30 files were downloaded for females and 30 files for males and were kept in separate folders on desktop.

The text files from these folders were uploaded onto Jupyter using read\_csv command line and concatenated to form two pandas dataframes namely female and male. A new column named ‘Sex’ was added to these dataframes. Female and male values were added into respective female and male dataframe Sex column. Further, these dataframes were concatenated to form final panda dataframe.

**Data Wrangling:**

* **Dealing with redundant columns**: Upon checking the information about the dataframe it was found that many columns like ‘Notes’, Codes for various demographics were present, these columns didn’t add any extra information to the dataset and therefore these columns were dropped from the dataframe.
* **Dealing with Null Values:** The dataset was search for rows that contain Null/NaN values and it was found that 3720 rows had null values for all the columns except ‘Sex’ column. Therefore it meant that probably ‘Notes’ column was adding these rows to the dataset and therefore had value only in ‘Sex’ column and no value in any other column. Therefore, these rows were dropped. And again dataset was checked for Null values and no more Null values were found.
* **Dealing with Not Applicable Values**: However, while screening the dataset it was found that there were values called Not Applicable were present in ‘Deaths’ , ‘Population’, Crude Rate’ columns. Around 116280 observations have it. To find out what these Not applicable values are information document from Centers for Disease Control and Prevention (CDC) was checked and it was found that this data was either not recorded or not present with the CDC, therefore it said Not Applicable in the dataframe. Since this information was not present, therefore these rows were dropped from the dataset.
* **Dealing with Suppressed Values:** There was another type of value called suppressed that was present in ‘Deaths’ and Crude Rate’ columns. It was found that these suppressed values are actually the death between 1 and 16. So if the death < 16 and >= 1, then instead of reporting the actual value the dataset shows suppressed. These dataset were also dropped for exploratory data analysis. These were around 325534 observations that were dropped.
* **Dealing with different data-types:** Next, the datatype for different columns were checked and it was found that some of the data where computational analysis will be performed were in object datatype, therefore these datatypes were converted into numeric format such that various computational analysis can be performed.

**Future Data Analysis**:

This cleaned data-frame will be used further for exploratory data analysis, where various features of the data like max, min, average, geographic/ demographic distributions and many others would be analyzed. This will provide a better understanding of the data and its caveats. Data will be wrangled again to remove any caveats that can skew further analysis. Finally, predictive data analysis will be performed and deeper insights into cancer-site mortality predictions will be achieved.