**Capstone Project Data Wrangling**

This document prepared to explain briefly how the data is collected, cleaned and prepared for the further analysis.

**Data collection and cleaning steps performed:**

**Data**

**Data Cleaning and Transformation**

Although two different data sets are used in the project, one general summary of data wrangling steps will be described here. First dataset Healthy Aging Data is collected from CDC website in form of csv file. The second dataset downloaded from Kaggle website which was in csv format and can be reached from this link. Only cognitive impairment type of disease is extracted from data as the scope of the project will cover dementia and its dependencies.

* Healthy Aging dataset had more than 29 columns and 20,000 rows. After **filtering** for cognitive diseases, the dataset had about 3696 rows. If the column(s) was no entries at all, they were eliminated from the dataset.
* Besides the column(s) that are fully empty, **dropping threshold** was set to 50 %. What that means is that if 50 % of a column(feature) is empty, we drop them to preserve consistency in the data.
* For the remaining null(empty) observations in columns(features); if the feature consists of continuous variables, the average value of the feature used to **fill in missing** values by taking groups taking into account such as the state and the age-group that a patient belongs to. Duplicate reports are removed as well.
* The Oasis MRI dataset which is used to determine if the subject has any dementia has 373 rows and 15 columns. As the dataset obtained from a research facility it is almost tidy. There were only two columns with missing data probably because they were not able to collect the information from subjects. **MMSE** feature had only two missing values and this feature was left skewed so to fill in missing values I preferred using **median value** to reduce the effect of skewness. I wanted to use **MICE (Multiple Imputation Chain Equations)** however my system did not support the package. Secondly, to fill in missing values in **SES** column I preferred to use **mode value (most frequent value)** as this data was an **ordinal** data, we may use most frequent value to replace the missing values.