**Milestone Report2 of Capstone Project 2**

**1. Problem Definition**

Energy consumption is growing due to tremendous changes in global and industrial world. Building energy consumption is related 36.6% of the total energy consumption in the United States by 2017. Many aspects of urban form influence energy consumption in commercial buildings, so city planners take in to consideration of those forms critically by understanding distributions of the energy related intensities.

The main problem in energy consumption is, how to control the habit of energy consumption in buildings. There are many different policies and researches proposed about this issue. From utilization to architectures of buildings, researchers focus on improving efficient and more reliable predictive models by using many data sources and machine learning algorithms.

**2. Client Identification**

This issue purposes to improve reliable solutions based on government concerns to make reforms for commercial buildings in cities of the nation. The U.S. Energy Information Administration (EIA) takes those kinds of responsibilities by collecting, analyzing energy information as a principal agency of the U.S. Federal Statistical System. EIA provide energy information for people from policy makers to public as a purpose of giving more inside of relations between energy and the environment with the economy.

By this study, people from public, business and academia will be understood how electricity consumption is related to any indicators from the information provided by the EIA.

**3. Describe your data set, and how you cleaned/wrangled it**

This study uses necessary information from EIA’s Commercial Buildings of Energy Consumption Survey (CBECS) to review trends in commercial buildings. The survey is consisting of information about 5000 to 7000 buildings conducted between years of 1983 and 1995. After than, the CBECS sample size was increased in 2012 with data that belongs to 6,720 buildings.

The CBECS survey data was provided from www.eia.gov. Two kinds of data were used in this study. The first data set is based on the information about the variables placed in the 2012 CBECS survey data and the second data is CBECS survey data which has 6720 rows and 1119 columns coded as either integer or floats.

Since the size of the data set is large, we follow many data wrangling steps before implementing a machine-learning model.

First of all, we needed to handle missing values. We found that there were many variable with the 99% of missing values, so we filtered our data by dropping variables with the 25% missing values. We obtained data with some values having 25% missing values and then we filled those values with the median of their values.

Having outliers was another issue that we handled during the data cleaning process. There were some outliers detected on the upper and lower bound of the target variable, so the scatter plot displayed them.

On the other hand, a look up function for finding variable names provided an efficiency without looking the excel sheets every time.

Additionally, we visualize the values of our target variable; ‘ELBTU’ (Electricity Consumption) and we found that it was not normally distributed, so logarithmic transformation was applied. Thus, we reached more normal looking target variable.

Lastly, many columns starting with letter ‘Z’, which corresponding with imputation of some variable as ‘0’ or ‘1’ and columns starting with ‘FINAL’ corresponding of number of buildings in the population that the observation represents were deleted.

**4. Explain your initial findings**

Histograms, bar plots, scatter plots and coefficient correlations plots provided a front view before starting of model implementation. Depending on our curiosity of how some values move on their path in which trend, some frequencies obtained based on different categorical values such as ‘ CENDIV’ (census of division), ‘PBPLUS’ (specific building activities), ‘SCFTC’(square foot of buildings categories) and any other categorical variables based on their high correlation with the target variable were detected. Another different style of energy consumption such as ‘MFBTU’ was related to the response variable by having a scatter plot and positive linear relationship explored. Also, the correlation plot proved that relationship. Some features related to the building architecture examined by box plots. We saw how electricity consumption changed in buildings among decades. Although there is not too much change after 1970s in electricity consumption, we see a decreasing at the last decades corresponding with our decade of now. Wall and roof construction material show some changing under some of their categories in terms of electricity consumption. Heating degree-days and cooling degree-days display similar trend throughout different buildings.

By constructing a correlation coefficient plot between the target and the predictive variables, we obtained that there are 23 variables having .5 or higher correlation with the target variable. For example, categorical variables such as ‘SQFTC’, ‘NWKERC’, ‘PCTRMC’, ‘LAPTPC’, ‘RFGVEN’ are highly correlated with the ‘log\_ELBTU’.