Milestone 2 - Data, Methods, and Analysis Reporting Update

The question we previously proposed is, *is there a practical way for individuals to reduce how much their insurance company spends on them, and thereby reduce how much the individual spends on insurance?* To answer this question, we selected to use the Medical Insurance Cost Prediction dataset from Kaggle. This data was downloaded and has a CC0: Public Domain license, so we are able to use this data without repercussion.

Some hiccups arose as we started analyzing the data. Our intention was to analyze how lifestyle and habits affect health insurance payouts with the assumption that insurance companies could in turn incentivize people to conduct healthier lives. The intention being that the individual would live a healthier life and pay less in insurance and the company would then save money with less being paid out in costs.

However, the author of this dataset claimed that 54 columns were present, but only 52 were. The two missing were **sleep_hours** and **exercise_frequency;** two variables of great interest to our study. To replace these variables, we decided to include the **urban_rural** (does the person live in an urban, suburban, or rural environment), **income,** and **employment_status.** While these aspects of life aren't very malleable, they are able to be changed by the individual to some degree. The other variables included are **bmi, smoker,** and **alcohol_freq.** These variables are predicting Y or **total_claims_paid**.

Not all of these variables were ready to be used. The **alcohol_freq** variable had roughly 30,000 missing values, despite all the other variables having no missing values. In order to decide what to properly do with these missing values, t-tests for difference in means and chi square tests were conducted. The t-tests were conducted for each individual numerical variable. The sample's mean between the missing values and the non-missing values was taken. The chi square tests were conducted to compare the proportions of each individual categorical variable between the missing values of alcohol frequency and where a value was present.

None of the t-tests nor the chi square tests were statistically significant. As such, it seemed that the data was likely missing completely at random as it was unrelated to any other variable being considered in the dataset. The options then were to either impute a value for each or remove them from the dataset. Even though 30,000 is a lot of data, I decided to remove them from the dataset. I did this as there is still 70,000 left, which indicates that there is enough power in any model we generate to perform inferences.

After removing the missing values, each variable was analyzed individually. A graphic was produced for each to analyze its distribution. Categorical variables had histograms of the proportions of each category created and numerical variables had distribution density plots generated. Categorical variables were transformed so that the modal level was deemed the base line and dummy variables were created for the other levels.

The data is ready to be fit to a model. After which, we will see if our model assumptions hold. If the assumptions can not be properly held, a transformation of total claims paid will be applied, as suggested by the box-cox method