**707 Project Proposal**

Group Members: Yvonne Feng, Jingya Cheng, Gloria Gu

**Background:**

Cardiovascular diseases (CVDs) refers to a group of diseases that cause problems in heart and blood vessel conditions. The World Health Organization states that CVDs are the leading cause of mortality worldwide, killing more people each year than any other cause combined. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths1. From the information above, it is clear that cardiovascular disease (CVD) is the leading cause of increased hospitalization, mortality, medical expenses, and productivity losses in the United States1.

Premature deaths can be avoided by identifying those who are most at risk for CVDs and ensuring they receive the proper care. The previous study indicated the limitations with moderate discrimination and poorly calibrated utilizing the predicted models such as pooled Cohort Equations (PCE) and the Framingham CV risk equation (FRS)2. Therefore, further analysis on the predicted model is required.

In order to ensure that more people may live healthy lives, we will apply effective data-driven methods for predicting cardiac problems that can enhance the overall prevention process. The variety of Machine learning algorithms will be implemented with two types of predicted models – Classification models and regression models, including logistic regression, support vector machine, K-Nearest Neighbors, decision tree and random forest. Our goal is to find the most accurate algorithms to predict the presence of cardiovascular diseases.

In addition, we will most focus on the following risk factors: cholesterol, Systolic blood pressure, Diastolic blood pressure and glucose. We treat those four factors as features to our prediction models. Since the datasets are collected from the medical examination, the output and the models can help to identify the abnormal values as the potential factors for the diagnosis of CVDs. The glucose levels in the diagnosis of diabetes and cardiovascular events has not been adequately studied, despite the fact that diabetes increases the risk of cardiovascular disease (CVD) and mortality3. Therefore, we hypothesize that the glucose level is associated with the probability of having CVDs.

**Data Description:**

The dataset for this project is cardiovascular disease dataset, which can be downloaded from Kaggle. The datasets are collected in medical examinations and contain information about patients. The datasets include information about patient demographics, encounters, lab Results and vital Signs, which can be used to predict whether the disease exists or not. Those variables are important features for model training and validation. There are 70,000 observations in the datasets and 12 variables. The table below is a description of the data:

| **Variables** | **Data type** |
| --- | --- |
| Age | int (days) |
| Height | int (cm) |
| Weight | float (kg) |
| Gender | categorical |
| Systolic blood pressure | int |
| Diastolic blood pressure | int |
| Cholesterol | 1: normal, 2: above normal, 3: well above normal |
| Glucose | 1: normal, 2: above normal, 3: well above normal |
| Smoking | binary |
| Alcohol intake | binary |
| Physical activity | binary |
| Presence or absence of cardiovascular disease | binary |

Since there is no missing value in the datasets, we are not sure whether there can be any effect on our results, because we do not see the original raw data. Though there is no missing value in the datasets, there are some incorrect values. For example, for the data for blood pressure, there are negative records and values greater than 1000, which are impossible in the real world. We will discuss this situation in the reports about how we will deal with these values.

**Exploratory Analysis (Data Visualization):**

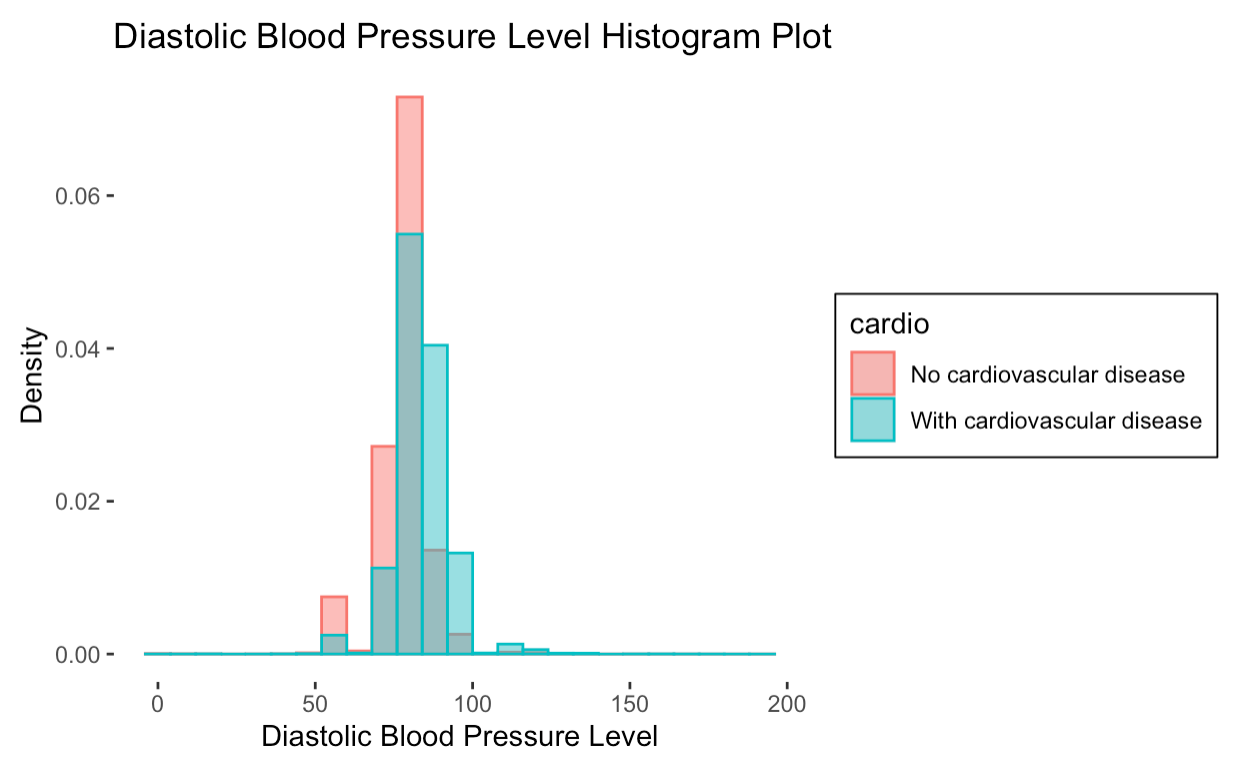
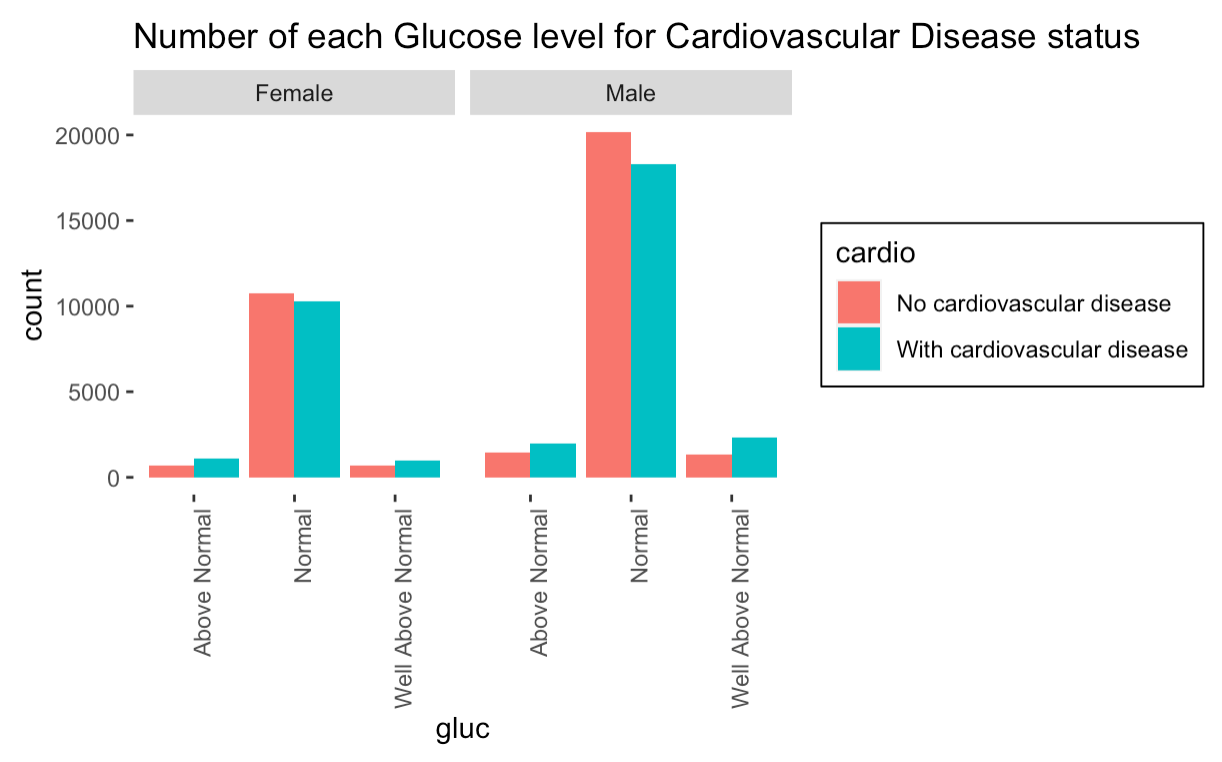
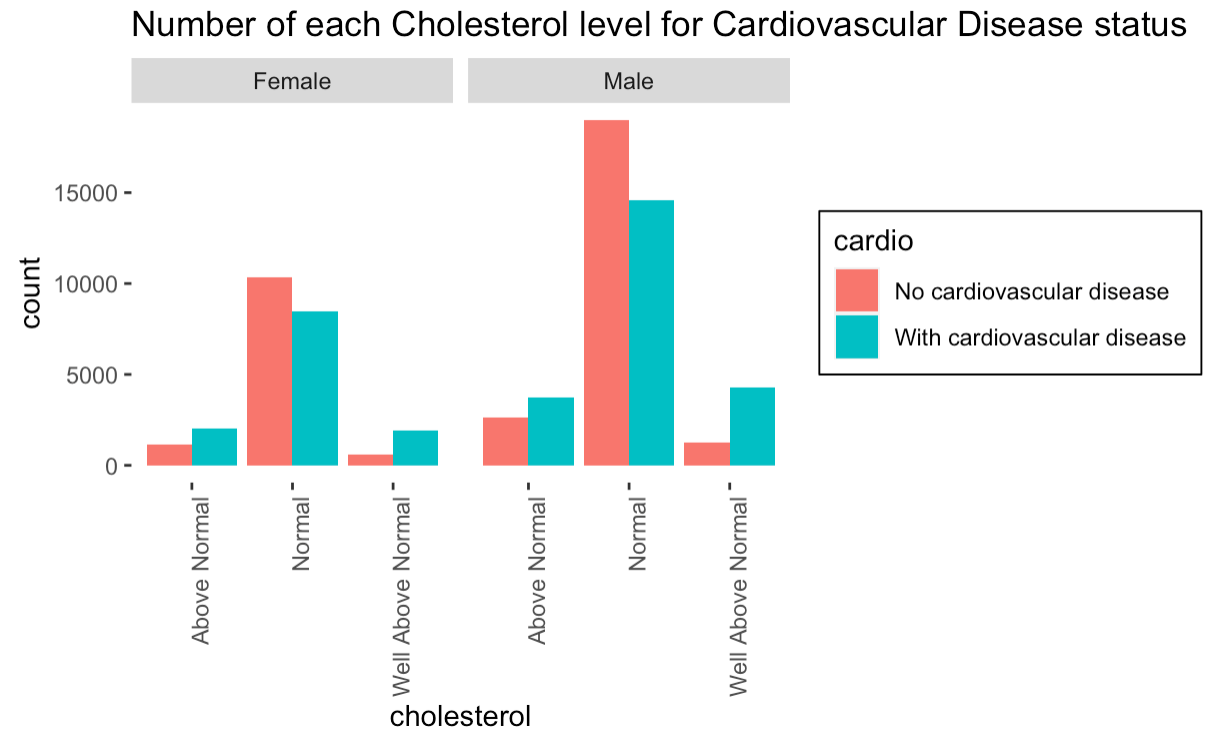
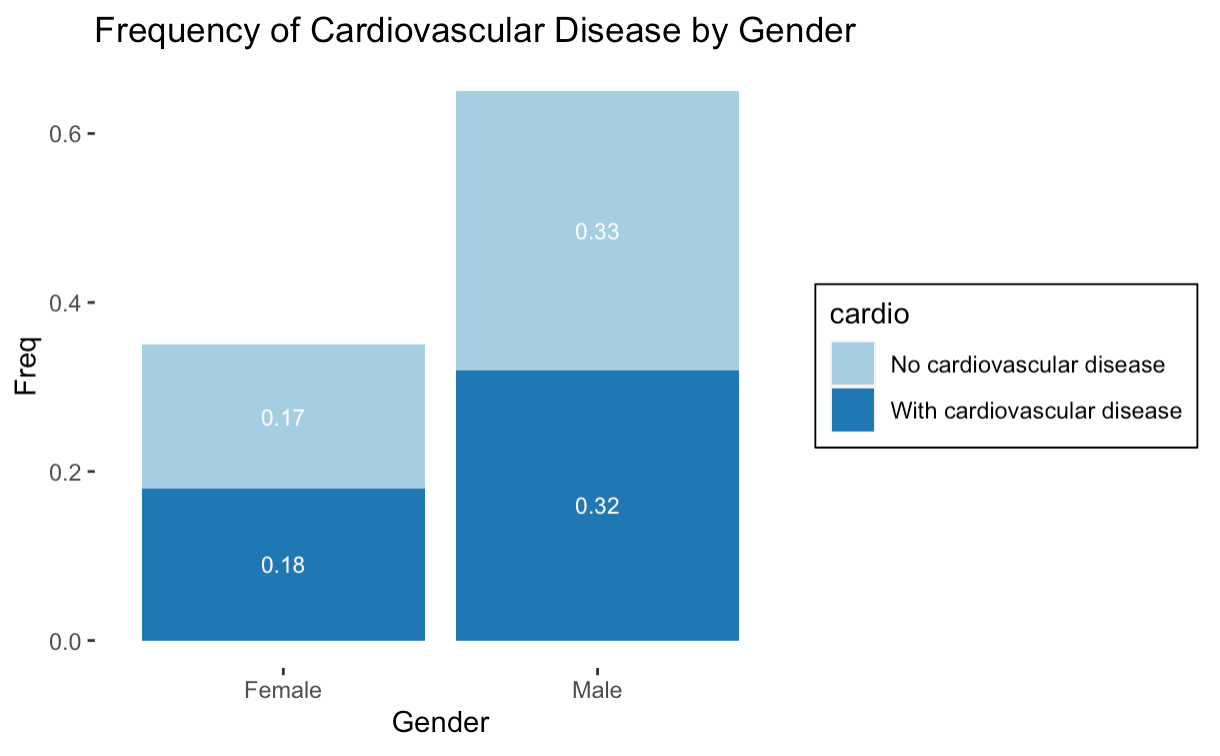
Figure 1: **Frequency of Cardiovascular Disease by Gender** The figure shows that more male had cardiovascular disease than female and approximately half of the individuals in each gender diagnosed with it

Figure 2: **Histogram of Diastolic Blood Pressure** Individuals with cardiovascular disease have significantly higher diastolic blood pressure

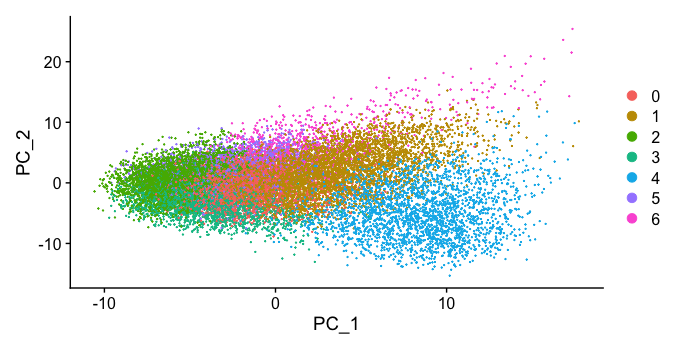
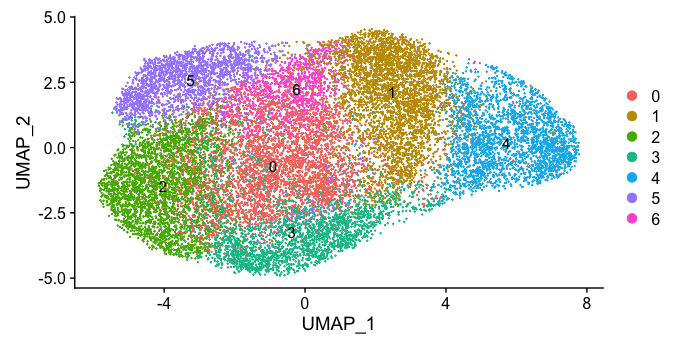
(For diastolic blood pressure, individuals with negative value and extremely high values over 200 are considered as outliers which may recorded due to mistake)

Figure 3 and 4: **Number of cholesterol and glucose level by cardiovascular disease and separated by gender** More individuals in both gender with cardiovascular disease have abnormal cholesterol and glucose level (either above or well above normal level)

Figure 6 and 7 indicate that by PCA and UMAP for dimension reduction, there are 7 clusters identified

 **Figure 1 Figure 2**

**Figure 3 Figure 4**



**Figure 5 Figure 6**

**References**

1. Sanz, M., D'Aiuto, F., Deanfield, J., & Fernandez-Avilés, F. (2010). European workshop in periodontal health and cardiovascular disease—scientific evidence on the association between periodontal and cardiovascular diseases: a review of the literature. *European heart journal supplements*, *12*(suppl\_B), B3-B12.
2. Jiang, Y., Zhang, X., Ma, R., Wang, X., Liu, J., Keerman, M., ... & Guo, H. (2021). Cardiovascular disease prediction by machine learning algorithms based on cytokines in Kazakhs of China. *Clinical epidemiology*, *13*, 417
3. Narloch, J. A., & Brandstater, M. E. (1995). Influence of breathing technique on arterial blood pressure during heavy weight lifting. *Archives of physical medicine*
4. Github: <https://github.com/Gloriazg92/BIOSTAT707_Project>