Proposal

Group member: Yimeng Cai (yc3577), Zheshu Jiang (zj2379), Ze Li (zl2746), Qianying Wu (qw2418)

1. Project title: Regression Analysis on Socio-economic factors and Bioindicators of Type II Diabetes by using R

2. Project Motivation Type II Diabetes is a global health concern that affects more than 37 million Americans and places a significant burden on healthcare systems. As a multifaceted disease, Type II diabetes was influenced by both socio-economic factors and biological indicators. By analyzing a comprehensive dataset from the Behavioral Risk Factor Surveillance System, we aim to uncover the correlations between diabetes prevalence and factors such as sex, education, income, lifestyle habits, and bioindicators like blood glucose and cholesterol levels. Through visualizations and statistical analysis in R, we aspire to highlight patterns that could inform better prevention and management strategies, thereby contributing to the broader dialogue on public health and socio-economic policies related to diabetes etc..

3. Data resources:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

- Data Description: The dataset was collected by the Behavioral Risk Factor Surveillance System (BRFSS) in 2015. This original dataset contains responses from 441,455 individuals and has 330 features. For this project, we use the one with 253,680 responses and a three-class diabetes target variable indicating no diabetes, prediabetes, or diabetes. The other potential factors for diabetes are blood pressure, cholesterol, cholesterol check, BMI, stroke, heart disease, and habits of smoking, physical exercise, eating fruit.

4. Intended Final Products

- Demographics Visualization: From scratch, to understand the basic demographics information around type II diabetes, we decide to perform data visualization of type II diabetes occurrences among population in different biological sex, educational level, and income level.
- Socio-economic Factors Analysis: In order to study socio - economic factors related to type II diabetes occurrences, we aim to distinguish the following three factors - smoking status, alcohol consumption, and body mass index (BMI). By implementing data analysis, statistical tests, along with data visualizations, we are able to study the correlation between each single factor and frequencies of type II diabetes.

- Bioindicators Association: With the understanding of factors correlated with the disease, we plan to study the two bioindicators: blood pressure and cholesterol level associated with type II diabetes in order to decide which biological indicator is able to perform as a significant biomarker through statistical tests.

5. Data Analysis

- In order to make demographics visualization, we plan to use filter and group samples according to different factors of sex, education, and income and then visualize the frequencies and distributions for each factors through histograms, box plots etc.
- To analyse the three socio-economic factors (sex, education, income), we are going to do sample t test, anova test, linear regression model, and model fitting to test for the associations and significance by looking at P-value, test statistics, or confidence intervals.
- To analyse the two bioindicators, we will perform a hypothesis test along with statistical test like sample t test, multiple linear regression and comparing the P-value or test statistics to ensure the significance.

6. Visualizations

- Use a stacked bar chart to show the proportion of individuals with heart disease within diabetic, pre-diabetic, and non-diabetic groups using ggplot.
- Visualize with a histogram or density plot, overlaying the distribution of blood glucose levels for diabetic, pre-diabetic, and non-diabetic groups using ggplot. We will also create a scatter plot with a trend line to observe the relationship between blood glucose levels and diabetes by ggplot.
- To visualize how diabetes prevalence varies across different demographic groups—such as sex, income, age, smoking status, and alcohol consumption—we will create a series of density plots. Each plot will display the distribution of diabetes cases within each demographic category, providing a clear visual comparison of how the disease's occurrence differs among various segments of the population.
- Display side-by-side boxplots to show the spread of BMI across the dataset to give BMI Level Distribution by ggplot.

7. Coding challenges

- Plotting: including handling missing values, outliers, and categorization.
- Ensuring accurate representation of data while constructing bar charts, dot plots, and density plots, ensuring accurate representation of data, for example, picking the right variables.
- Customizing aesthetics of the graphs by mastering ggplot2's aesthetic parameters to differentiate categories visually by color, shape, or size.
- Embedding statistical test results, like p-values, into plots and plotting dimensions and scales for clarity and to accommodate various factor levels could be difficult.

8. Planned Timeline

Timeline	Requirement	Predicted Timeline
11/13-17	Project review meeting	Nov 15th or 16th
12/9	Report	Nov 24
12/9	webpage and screencast	Dec 1-7
12/9	peer assessment	Dec 7-8
12/14	In class discussion	