Proposal

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

1. Project title: Regression Analysis on Multiple Factors and Bioindicators of Type II Diabetes

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 is influenced by both social factors, biological factors, and bioindicators. By analyzing a comprehensive dataset from the Behavioral Risk Factor Surveillance System (BRFSS), we aim to uncover the correlations between diabetes prevalence and factors such as sex, education, income, bmi, smoking status 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 policies related to diabetes.

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 potential factors for diabetes are biological factors such as sex, blood pressure, cholesterol, BMI, heart disease, social factors such as education and income, and habits of smoking, physical exercise, eating fruit etc..

4. Intended Final Products

- Factors Visualization: We decide to perform data visualization of diabetes distributions among population in different biological factors(sex, bmi) and socio-economic factors(income, education, smoking).
- Factors Analysis: We aim to analyze the significance of correlation between biological and socio-economic factors and diabetes. By implementing data analysis, we are able to study the correlation between each factor and frequencies of diabetes.
- Bioindicators Association: We plan to study the bioindicators: blood pressure and cholesterol level associated with diabetes to decide which biological indicator is able to perform as a significant biomarker.

5. Data Analysis

- To make visualization, we plan to use filter, group by, and related functions according to different factors and then visualize the frequencies for each factor through histograms, box plots etc.
- To analyse factors significance, we perform sample t test, anova test, and linear regression model to test for the associations and significance by looking at test statistics, or confidence intervals.
- To analyse the two bioindicators, we perform a hypothesis test along with statistical test like sample t test, linear regression and comparing the P-value 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 sample groups using ggplot.
- Visualize with a histogram or density plot, overlaying the distribution biology factors blood pressure and cholesterol levels using ggplot.

- To visualize how diabetes prevalence varies across different demographic factor groups such as income, sex, bmi, and socio-economic factors like education and income, we create a series of density plots. Each plot displays the distribution of diabetes within each factor category, providing a clear visual comparison of how the disease's occurrence differs among various segments of the population.

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, 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 scales for clarity and to accommodate various factor levels can 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