Course: **CIS 5357 Computing for Data Analytics**

Writing assignment

The [final exam](#) for this course will be a take home project and the goal is to think creatively. You will explore a given dataset and use the skills you've learned in the course to process the data and do something interesting with it.

For this week's writing assignment, you will start to devise your final exam plan in a short writeup (equivalent to 1-2 pages in a Word/PDF document). Select one of the provided datasets and give an overview of what you plan to explore (note that you do not need to do the analysis just yet but start thinking about how you will approach the final exam project). Do you expect any problems with the data? If so, describe how you will deal with any problems with the data.

Exploring the heart_dataset and Planning for the Final Exam Project

Introduction

The final exam for this course will involve a take-home project that challenges us to creatively explore and analyze a dataset related to heart attack risk factors. The dataset provides information on various factors, including patient demographics, lifestyle habits, and medical history, and the goal is to utilize our data analysis skills to uncover meaningful insights and potentially develop predictive models for heart attack risk assessment.

The dataset contains a rich collection of variables that can be explored to gain a deeper understanding of heart attack risk factors. Some of the key aspects I plan to investigate include:

1) Understanding the distribution of risk factors:

Examining the distribution of variables like age, cholesterol levels, blood pressure, heart rate, obesity, diabetes, smoking, and alcohol Consumption habits can reveal patterns and relationships between these factors and heart attack risk.

2) Exploring the impact of lifestyle habits:

Investigating the influence of factors like exercise, diet, physical activity days per week, sleep hours, stress levels on heart attack risk can provide valuable insights into lifestyle modifications that can reduce risk.

3) Assessing the role of medical history:

Examining the relationship between previous heart problems, family history, medication use, and heart attack risk can shed light on the importance of preventive measures and early intervention.

4) Identifying correlations:

Correlational analysis is a statistical technique that measures the strength and direction of the relationship between two quantitative variables. It helps us understand how closely two variables are linked and whether they tend to change together in a predictable way.

Potential Data Issues and Mitigation Strategies

While the dataset offers a wealth of information, it is important to consider potential data quality issues that may affect the analysis. Some common data issues include:

1) Missing values:

Missing data can introduce bias and reduce the statistical power of the analysis. Techniques like imputation or exclusion of incomplete records can be employed to address missing values.

2) Outliers:

Outliers can distort the distribution of variables and affect the interpretation of results. Techniques like outlier detection and removal or assigning appropriate weights can be used to handle outliers.

3) Data inconsistencies:

Inconsistencies in data formats, coding errors, or incorrect measurements can compromise the validity of the analysis. Careful data cleaning and validation can help identify and rectify these issues.

Project Plan

To effectively analyze the dataset and address potential data issues, I plan to follow a structured approach:

1) Data Cleaning and Preparation:

This initial step involves identifying and addressing missing values, outliers, and data inconsistencies to ensure data quality.

2) Exploratory Data Analysis:

This involves summarizing and visualizing the data to gain a preliminary understanding of the distribution of variables, relationships between variables, and potential patterns related to heart attack risk.

3) Correlational Analysis:

This involves quantifying the strength and direction of associations between variables to identify significant correlations that may indicate underlying relationships.

4) Visualization and Communication:

This involves creating clear and informative visualizations to effectively communicate the findings and insights derived from the analysis.

5) Addressing Data Issues:

Throughout the analysis, I will be vigilant in identifying and addressing any data quality issues that may arise, ensuring the integrity of the results.

Conclusion

The heart attack dataset presents a valuable opportunity to apply data analysis skills to explore and understand the complex factors that contribute to heart attack risk. By carefully examining the data, addressing potential issues, and employing appropriate statistical techniques, we can gain valuable insights that may aid in developing preventive strategies and improving heart health outcomes.