Stat 6021 Project 2 Proposal: Exploring Cardiovascular Risks

Haley Egan (vkb6bn), JD Pinto (jp5ph), Seth Galluzzi (vzw6yk), Sydney Masterson (bmp6tz)

Introduction

Cardiovascular disease is the number one cause of death in the United States. According to the CDC, one in every four deaths in the United States can be attributed to cardiovascular disease. Thus, it is important to understand the different factors that increase the risk of cardiovascular disease. Among these factors are obesity, diabetes, and risk of coronary heart disease. During this exploration, we will observe how different predictor variables such as weight, total cholesterol, and age impact response variables associated with obesity, diabetes, and risk of coronary heart disease. We will also explore the different relationships between these variables, and analyze the prevalence of obesity, diabetes, and risk of coronary heart disease within communities in central Virginia.

The Data

The data set we are exploring is from the faraway package in R, and is named *diabetes*. The data was obtained in a study where 403 African Americans from central Virginia were interviewed. During the interviews, information such as total cholesterol, weight, age, and gender were collected. We chose this data set because it gave us the ability to make comparisons between different variables that impact cardiovascular disease. By making these comparisons, we can gain a better understanding of cardiovascular disease among communities in central Virginia. To gain further insight of the data set we will be exploring, an explanation of the variables is below:

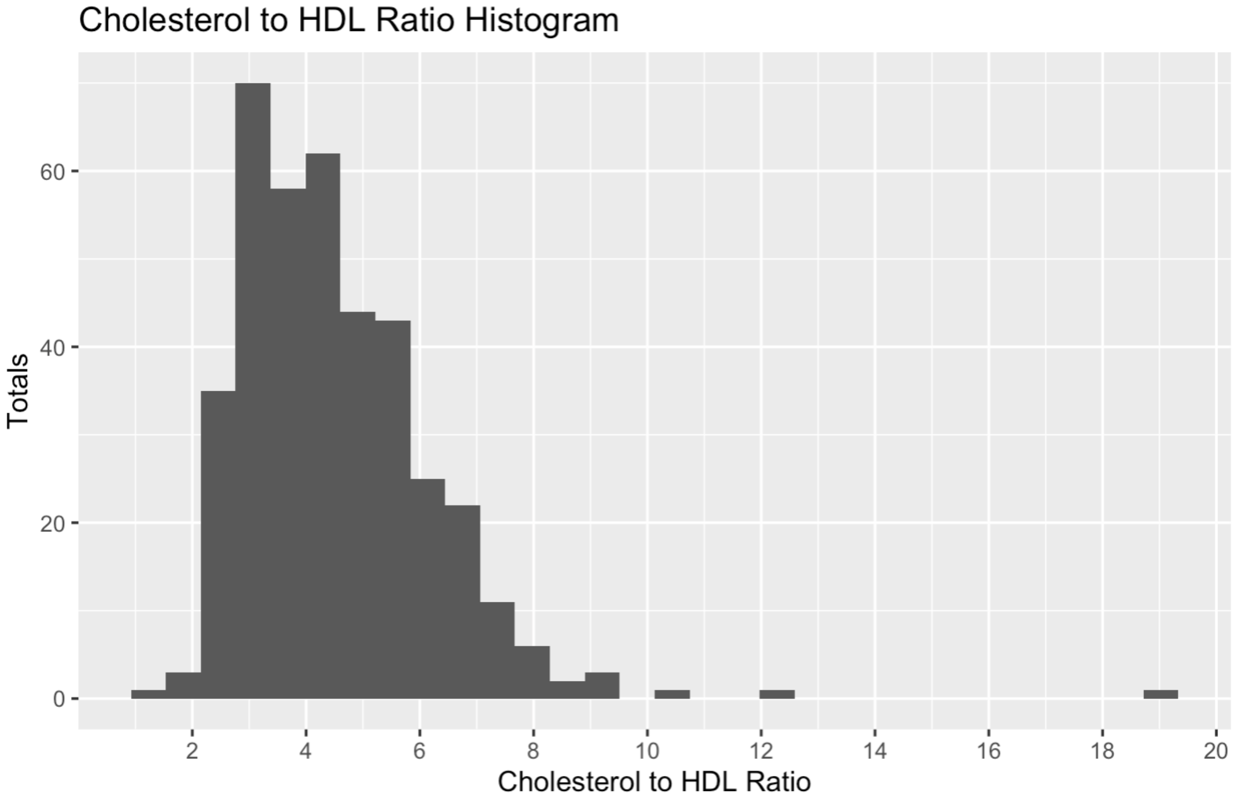
id	The subject ID of the observation. This column will not be needed for our exploration, and will be eliminated from the data set.
chol	The total cholesterol of the observation. Cholesterol is measured in milligrams per deciliter of blood. A healthy level of total cholesterol is between 125 to 200 mg/dL. The minimum cholesterol of our data set is 78, the maximum is 443, and the median is 204 mg/dL.
stab.glu	The stabilized glucose of the observation. Stabilized glucose is measured in milligrams per deciliter of blood.
hdl	The high density lipoprotein of the observation. HDL aides in the removal of cholesterol from the arteries. HDL is also measured in milligrams per deciliter of blood. A healthy level of HDL in males and females is above 40 mg/dL and 50 mg/dL respectively. The minimum HDL in our data set is 12, the maximum is 120, and the median is 46.
ratio	The ratio between total cholesterol and HDL. This ratio is often used to identify coronary heart disease. For perspective, a ratio of 3.5 is considered very good, while a ratio of 4.5 is considered at risk of coronary heart disease. Furthermore, most healthcare providers prefer a ratio less than 5. The minimum ratio in our data is 1.5, the maximum is 19.3, and the median is 4.2.

glyhb	The glycosylated hemoglobin value of the observation. This measures the amount of sugar in your blood. A glycosylated hemoglobin value greater than 7 is usually taken as a positive diabetes test. The minimum glycosylated hemoglobin value is 2.68, the maximum is 16.11, and the median is 4.84. There are also 13 missing values.
location	The location value of the observation. For this study, location was categorical with values Buckingham and Louisa with 200 and 203 total observations respectively.
age	The age of the observation. For this study, the minimum age was 19, the maximum age was 92, and the median was 44.
gender	The gender of the observation. For this study, gender is categorical with values male or female. There are 160 males and 228 females.
height	The height of the observation. Height was measured in inches. The minimum was 52 inches, the maximum was 76 inches, and the median was 66 inches.
weight	The weight of the observation. Weight was measured in pounds. The minimum was 99 lbs, the maximum was 325 lbs, and the median was 173.
frame	The frame of the observation. Frame is a categorical variable with values small, medium, and large. There are 101 small, 177 medium, and 99 large observations. There are also 11 missing values.
bp.1s	The first systolic blood pressure of the observation, measured in millimeters of mercury (mmHg).
bp.1d	The first diastolic blood pressure of the observation, measured in millimeters of mercury (mmHg).
bp.2s	The second systolic blood pressure of the observation. This variable contains a multitude of missing values.
bp.2d	The second diastolic blood pressure of the observation. This variable contains a multitude of missing values.
waist	The waist measurement of the observation. Waist was measured in inches. The minimum was 26 inches, the maximum was 56 inches and the median was 37 inches.
hip	The hip length of the observation. Hip was measured in inches. The minimum was 30 inches, the maximum was 64 inches, and the median was 42 inches.
time.ppn	The postprandial time when the observation was conducted. This variable describes the amount of time since the last meal of the observation and is measured in minutes. The minimum of our data set is 5, the maximum is 1560, and the median is 240 minutes.

*To further our understanding of the data, other variables may be added to the data frame such as body mass index (BMI).

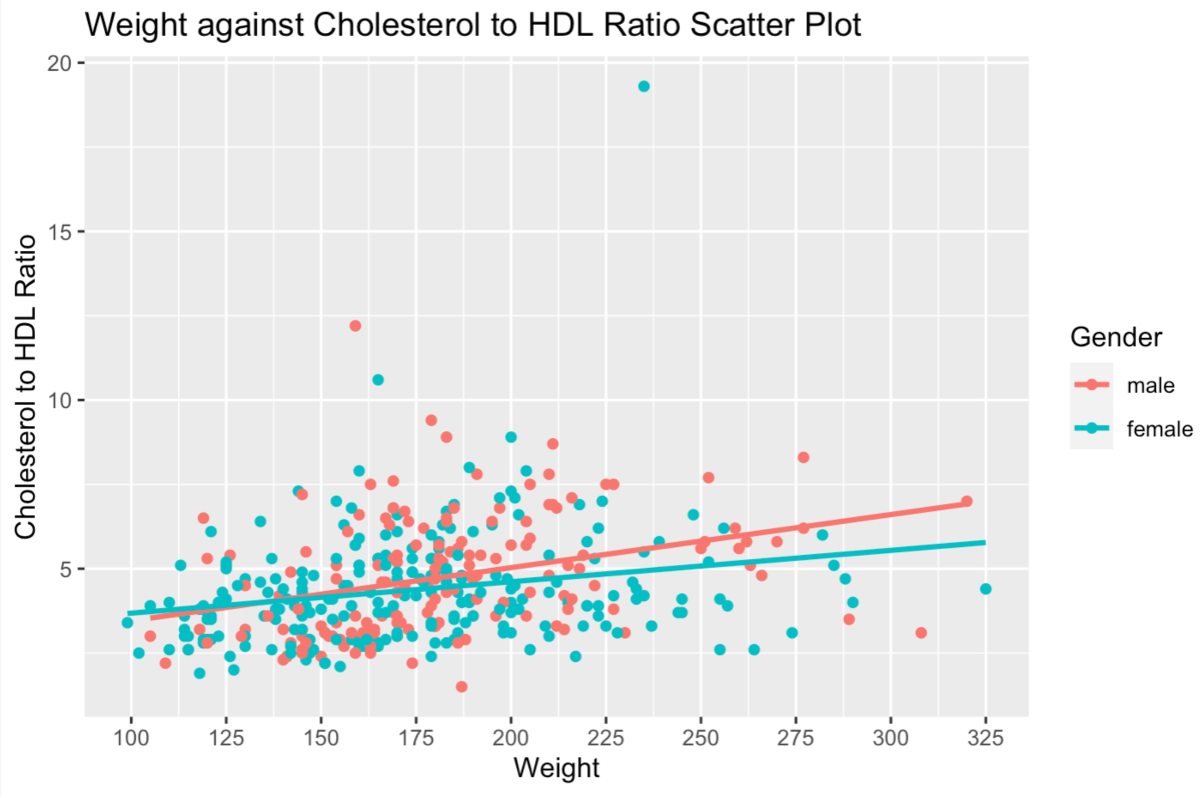
Preliminary Visualizations and Analysis

To explore the different predictors that attribute to a heightened risk of cardiovascular disease, and to further our understanding of the prevalence of cardiovascular disease in central Virginia communities, we will create a number of visualizations from the data. These visualizations will contain response variables such as cholesterol to HDL ratio (ratio) and predictor variables such as weight. To start, let's look at a histogram depicting the cholesterol to HDL ratio.



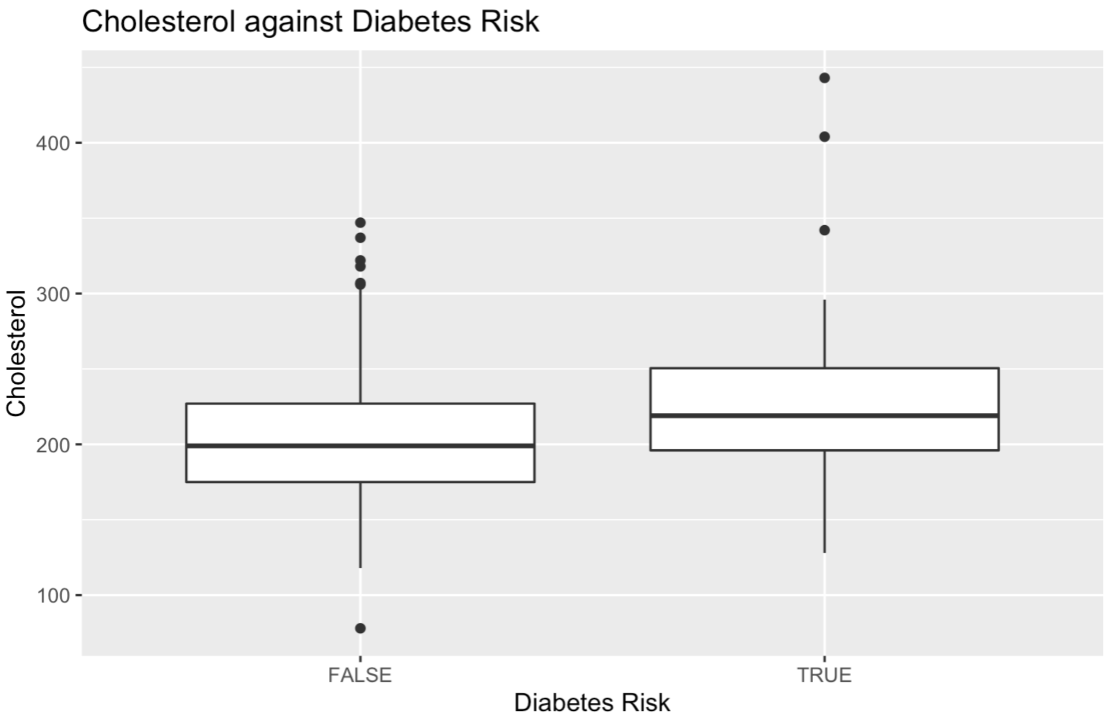
The histogram above indicates a large number of observations with ratio values between three and six. This is already disheartening news! A ratio of ~3.5 is considered good, while a value greater than ~4.5 is considered high risk of coronary heart disease. As seen above, there are many observations with values of 5 or more! To further this exploration within our project, we will view the ratio based on gender and location. We can also compare the ratio to variables such as weight, age, and cholesterol. This will allow us to better understand the impact different variables have on our cholesterol to HDL ratio.

To further the exploration of our data, below is a scatter plot comparing cholesterol to HDL ratio to the weight of the observation between genders.

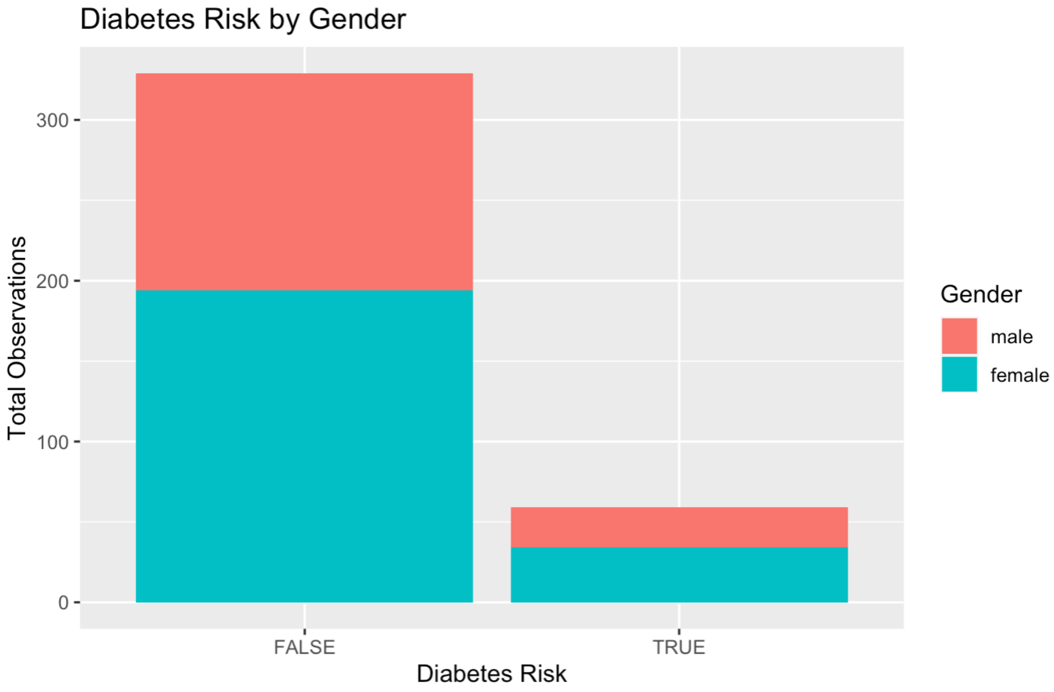


The scatter plot suggests a slight positive linear relationship, indicating that weight may have some impact on the cholesterol to HDL ratio. The slope of the regression line appears to be only marginally steeper between genders. This indicates that weight has little impact on the cholesterol to HDL ratio between genders. Since there is not a strong relationship, we need to look at other variables to determine better indicators of the cholesterol to HDL ratio.

Within the details of our data frame is information regarding glycosylated hemoglobin values. The information states that a glycosylated hemoglobin value greater than 7 is usually taken as a positive diagnosis of diabetes. To gain a better understanding of the data, we decided to create a new variable within the data frame. This categorical variable is called *diabetes risk,* and has a value of TRUE when the observation's glycosylated hemoglobin is above 7 and FALSE otherwise. After creating the variable, we created a box plot comparing the new categorical variable to the total cholesterol of each observation.



This box plot indicates that the observations with a glycosylated hemoglobin value higher than 7 have higher cholesterol levels than the observations that do not have a glycosylated hemoglobin value higher than 7. Further analysis in R shows that 59 observations are in the true subset while 329 are in the false subset. Thus, nearly 18% of observations are considered at risk of diabetes. This information is further illustrated in the box plot below.



Similarly to diabetes, we can explore other causes of cardiovascular disease such as obesity by making comparisons to the observations’ weight and bmi. After exploring the visualizations, we will analyze the regression lines that best predict our response variables. Furthermore, we will explore the percentage of observations with high risk of cardiovascular disease in the community.

Outcome

Results from this exploration will be discussed in three ways: a visual representation of the main results, statistical summaries, and written explanations. These three methods have standalone advantages and complement each other to explain the results comprehensively. During this project, we hope to gain a better understanding of how different variables impact the risks of cardiovascular disease. We also hope to create clear visualizations that inform others about the prevalence of those variables within our communities. Finally, we hope to strengthen our understanding of R as a tool that helps us gather and present data efficiently and effectively.

To accomplish analysis, we will compare predictor variables, such as weight and height, to response variables, such as glycosylated hemoglobin. We will also look at the percentage of diabetes, coronary heart disease, and obesity within the central Virginian communities. By completing this project, we will learn more about how cardiovascular disease impacts our communities, we will inform others about different variables that help indicate the risk of cardiovascular disease, and we will learn more about R as a tool to help us present our new found knowledge.

References

Centers for Disease Control and Prevention. (2020, January 31). *LDL & HDL: Good & Bad Cholesterol*. Centers for Disease Control and Prevention. Retrieved November 15, 2021, from https://www.cdc.gov/cholesterol/ldl_hdl.htm.

Centers for Disease Control and Prevention. (2021, August 10). *All about your A1C*. Centers for Disease Control and Prevention. Retrieved November 15, 2021, from https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html.

Centers for Disease Control and Prevention. (2021, September 27). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved November 15, 2021, from https://www.cdc.gov/heartdisease/facts.htm.

Diabetes: Diabetes and obesity, cardiovascular risk factors. RDocumentation. (n.d.). Retrieved November 15, 2021, from https://www.rdocumentation.org/packages/faraway/versions/1.0.7/topics/diabetes.

Lipid panel with total cholesterol: HDL ratio. Lipid Panel with Total Cholesterol: HDL Ratio - Health Encyclopedia - University of Rochester Medical Center. (n.d.). Retrieved November 15, 2021, from https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=167&ContentID=lipid_panel_hdl_ratio.

Mayo Foundation for Medical Education and Research. (2020, January 29). *Cholesterol ratio or non-HDL cholesterol: Which is most important?* Mayo Clinic. Retrieved November 15, 2021, from https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/expert-answers/cholesterol-ratio/faq-20058006.

U.S. National Library of Medicine. (2020, October 2). *Cholesterol levels: What you need to know*. MedlinePlus. Retrieved November 15, 2021, from https://medlineplus.gov/cholesterollevelswhatyouneedtoknow.html.

WebMD. (n.d.). *Cardiovascular (heart) diseases: Types and treatments*. WebMD. Retrieved November 15, 2021, from https://www.webmd.com/heart-disease/guide/diseases-cardiovascular.