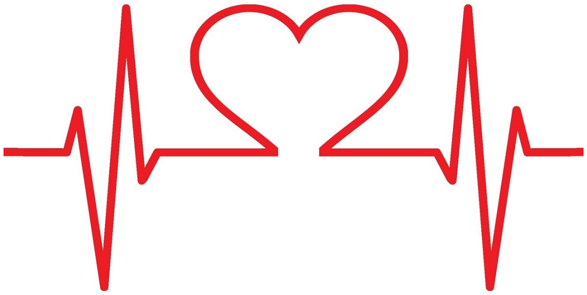
SDS 363 Final Project

Group name: Heartbeat
Names: Aorta, Atrium, Ventricle



Introduction

Throughout this course, we used multiple different datasets that each had their own characteristics in variables and were each analyzed with a different multivariate test. For the final project, our group was interested in analyzing a dataset that contained variables that pertained to a very important issue of health.

In 2020, coronary heart disease was the leading cause (41.2%) of deaths in cardiovascular disease in just the United States alone. About 697,000 people in the United States died from heart disease. In fact, it’s jolting to see how one person dies every 34 seconds in the United States due to cardiovascular disease. Every 40 seconds, someone in the US has a heart attack. And the formidable aspect of these is that ⅕ of these attacks are silent, meaning that the damage is done but the person is not aware of it. The cost to maintain and fight back against heart disease costs the US $229 billion each year from 2017-2018. We know that heart disease varies by sex, race, ethnicity, etc. However, we wanted to study at a deeper level, the strings of relation these variables play into coronary health. Because of the relevancy of this issue, we wanted to study what sort of variables could be causing this and what possible relations among each of the variables existed.

Design and Primary Questions

The primary question we want to answer is: **How can certain demographic factors, health behaviors, and biological markers affect the development of heart disease?**

In this experiment, we will be using the following tests:

1. Principal Components Analysis
2. Cluster analysis
3. MANOVA

Data

The dataset that we chose is called “Risk Factors for Cardiovascular Heart Disease”, where it examines risk factors in age, gender, height, weight, and other health metrics. The aim of what we want to discover is: to what extent do the variables impact and weigh on the effect of heart health and likelihood of heart disease.

I. Description of Variables

A.

Variable name	Description
Age	Age of participant (integer)
Gender	Sex of participant (male/female).
Height	Height measured in centimeters (integer)
Weight	Weight measured in kilograms (integer)
Ap_hi	Systolic blood pressure reading taken from patient (integer)
Ap_lo	Diastolic blood pressure reading taken from patient (integer)
Cholesterol	Total cholesterol level read as mg/dl grouped into low, medium, high (1, 2, 3 respectively)
Gluc	Glucose level read as mmol/l grouped into low, medium, high (1, 2, 3 respectively)
Smoke	Whether person smokes or not(binary; 0= No , 1=Yes)
Alco	Whether person drinks alcohol or not (binary; 0 =No ,1 =Yes)
Active	Whether person physically active or not(Binary ;0 =No,1 = Yes)
Cardio	Whether person suffers from cardiovascular diseases or not(Binary ;0 – no , 1 -yes)

II. How the data was collected

A. Data was collected by the author, Kuzak Dempsy (2021)

B. Dataset available on Kaggle.

C. https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

III. Sources of error

A. A few of the variables were not normally distributed, and most of the categorical variables had only two levels (yes or no), which could make finding a meaningful interpretation more difficult. Regardless, we push forth.

IV. Questionable points

A. The description of the data on Kaggle was suspect. For example, the author says that cholesterol levels were “read as mg/dl on a scale 0 - 5+ units(integer). Each unit denoting increase/decrease by 20 mg/dL respectively.” However, the maximum value of cholesterol was 3, so we assumed that 1, 2, 3, were low, medium, and high levels of cholesterol. Same for glucose.

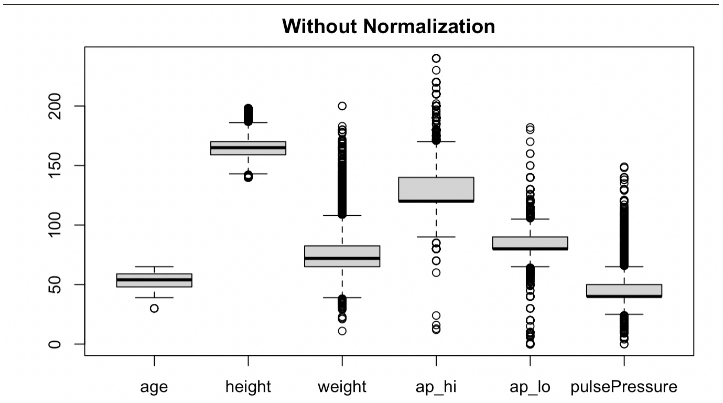
B. Additionally, after normalization, as you can see in one of our plots, some of the variance still maintained a high degree. This happened for the variable ap_lo.

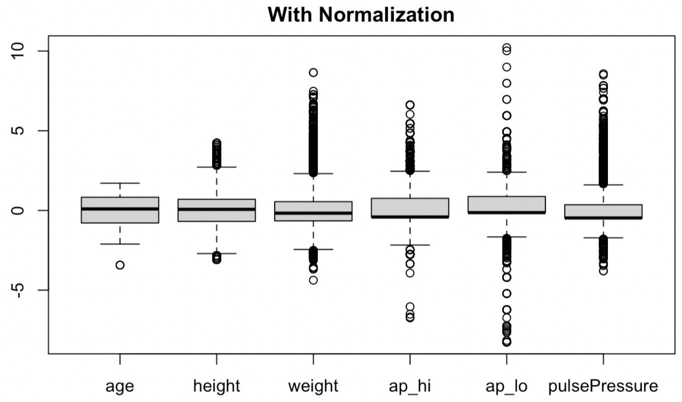
Plots / Summary Statistics

PCA and Cluster Analysis

In cluster analysis, we don't want to standardize our data if the variables have meaningful differences in units or scales. An example would be: clustering customers based on their purchase behavior and you have variables such as "total purchase amount" and "number of purchases", standardizing the data would remove the original differences in scale between the two variables.

In this case, we want to standardize. The variables are on different scales (ie. "height" "weight" "age") and have different units, so standardizing the data can help to ensure that each variable contributes equally to the analysis. By doing so, we ensure that each variable has an equal influence on the resulting clusters, and the clusters are not biased towards any particular variable.





Based on this, we chose the Manhattan distance. This is because Euclidean distance, despite being one of the most commonly used for continuous variables, is often best used for datasets without outliers since they can influence the Euclidean distance. However, upon looking at the outliers in the visualized dataframe, we notice that even after standardizing, there are still some outliers in the dataset (ie. in the weight).

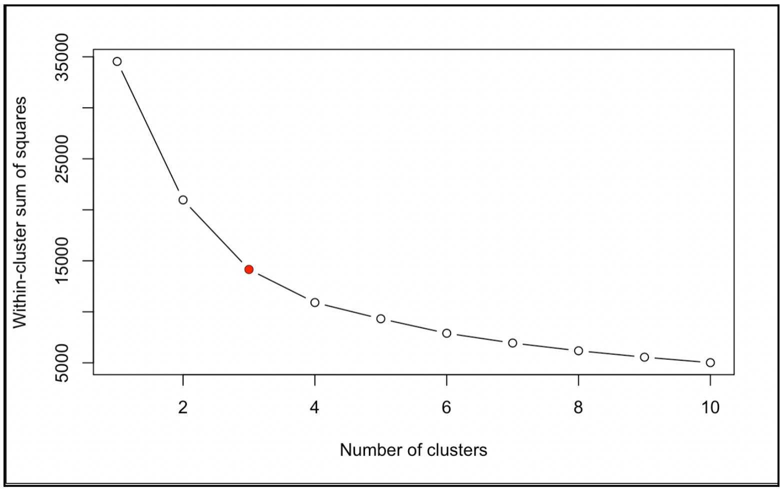
The main difference between these two distance metrics is how they account for the direction of the differences between corresponding coordinates. Euclidean distance considers both the magnitude and direction of the differences, whereas Manhattan distance only considers the magnitude. As a result, Euclidean distance tends to be more sensitive to differences in magnitude and is useful when the variables have the same scale, while Manhattan distance is more useful when variables are measured in different scales or units.The Manhattan distance would be best because it measures the distance between two points by summing the absolute differences of their coordinates.

Before we begin, we notice that since the dimension of "heart_norm" is 65179 rows x 9 columns. This can lead to computational difficulties and make it difficult to visualize the resulting clusters.

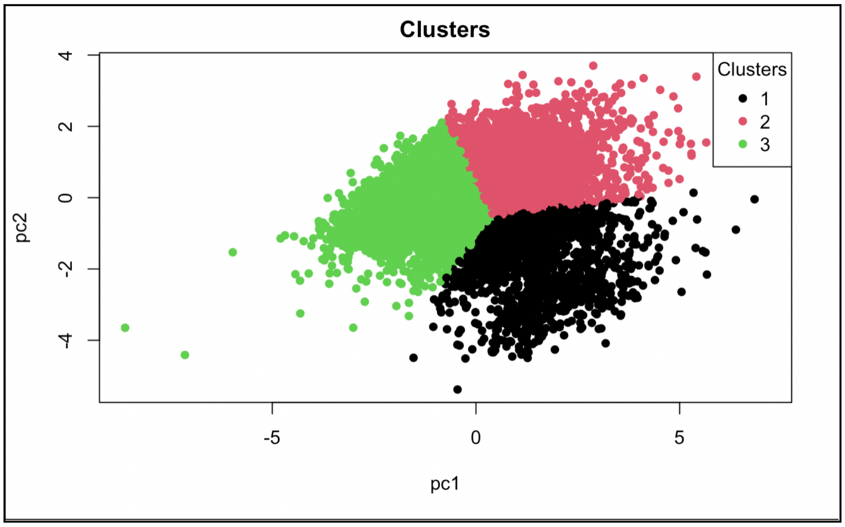
To solve for this, we perform clustering on a subset of the data. We chose a random sample of 10000 data points.

We use the elbow technique below to find how to find the optimal number of clusters to use.

Based on this, the optimal number of clusters is 3.

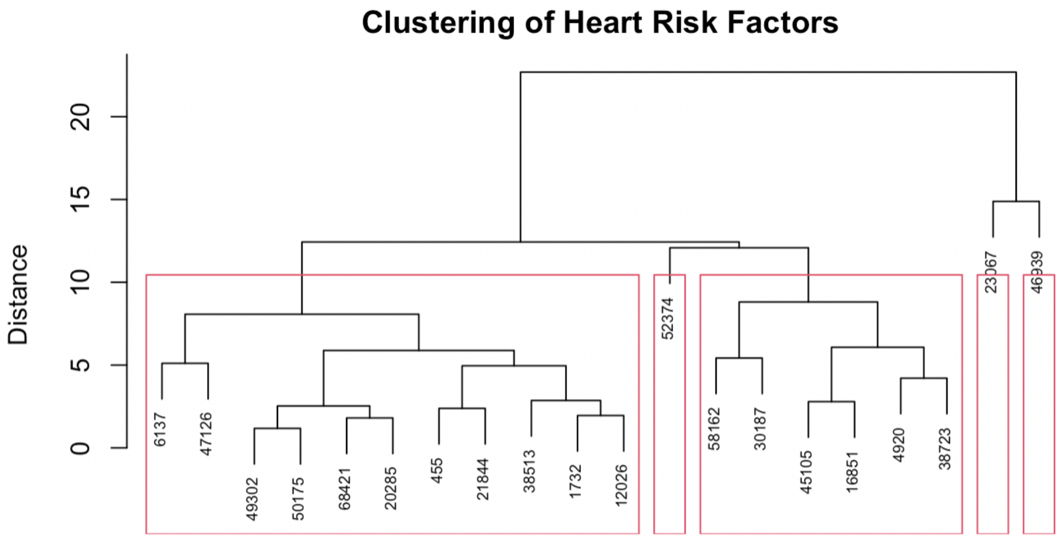


Below is our cluster plot. We see a good split of the three clusters. Based off that, we decided to do a test to see the correlations of the original variables in the principal components in order to identify which variables most strongly associate with each principal component. For PC1 the highest correlation were systolic blood pressure and diastolic blood pressure, and cholesterol level and not so much with age and weight. This means that PC1 is most likely to be related to general cardiovascular health. The highest correlation with PC2 was weight, followed by height, which is most likely related to the general physique of the patient. All of the other variables were negatively correlated to a lesser degree.



Cluster Analysis Sampling

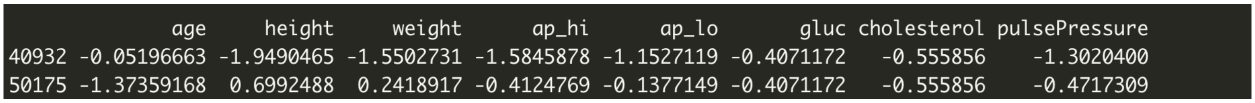
In this case because there are 10000 rows, we would not want to do cluster analysis on all of these rows. Because of that, we would ideally want to find a subset of that that we can then perform closer analysis on. From there we can then deduce commonalities and differences among the participants in terms of the variables.



From the dendrogram above, we notice that there are two primary groups. In the group on the left, since the first branch is a bit longer, that means that the differences on the left are a bit closer than the differences on the right. With this randomly sampled cluster, we can also deduce the similarities among the groups of participants.

We can note the similarities among some groups that are much lower on the y-axis, including:

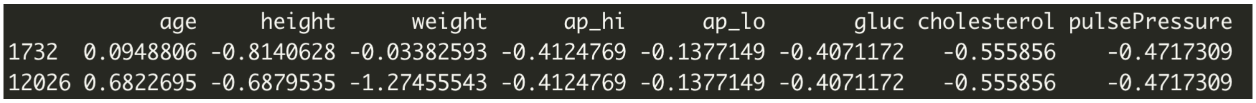
(1). 40932 and 50175



```
              age     height     weight      ap_hi      ap_lo       gluc cholesterol pulsePressure
40932 -0.05196663 -1.9490465 -1.5502731 -1.5845878 -1.1527119 -0.4071172   -0.555856    -1.3020400
50175 -1.37359168  0.6992488  0.2418917 -0.4124769 -0.1377149 -0.4071172   -0.555856    -0.4717309
```

From the above, we notice a difference in age, height, weight, ap_hi, ap_lo, and pulsePressure. The glucose and cholesterol reading taken from the patient are identical, which could signify that cholesterol and glucose are large predictors of cardiovascular disease.

(2). 1732 and 12026



```
            age     height      weight      ap_hi      ap_lo       gluc cholesterol pulsePressure
1732  0.0948806 -0.8140628 -0.03382593 -0.4124769 -0.1377149 -0.4071172   -0.555856    -0.4717309
12026 0.6822695 -0.6879535 -1.27455543 -0.4124769 -0.1377149 -0.4071172   -0.555856    -0.4717309
```

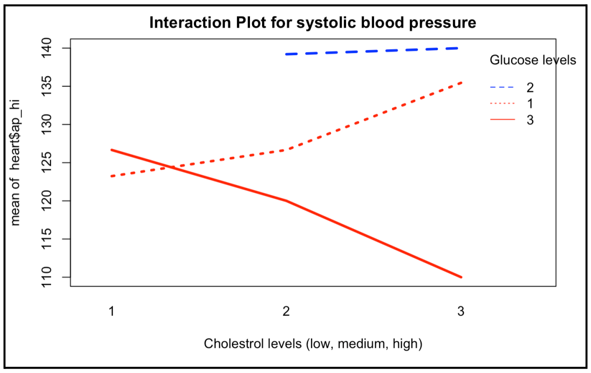
From the above, we notice a difference in age, height, weight, ap_hi, and ap_lo. The glucose and cholesterol and pulsePressure reading taken from the patient are identical.

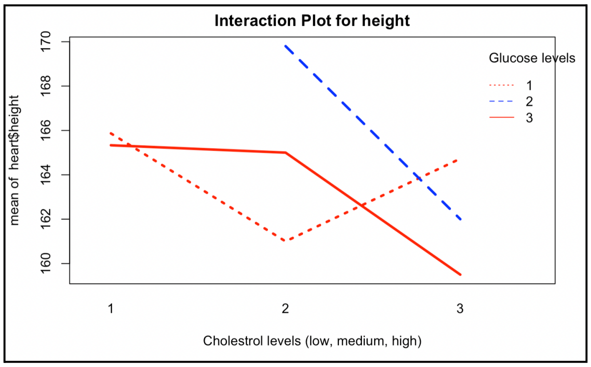
Even though the dendrogram cannot be used to visualize all 10,000 (and 65,000 from the entire dataset) observed participants, it can definitely help with observing minute differences and similarities at the individual level.

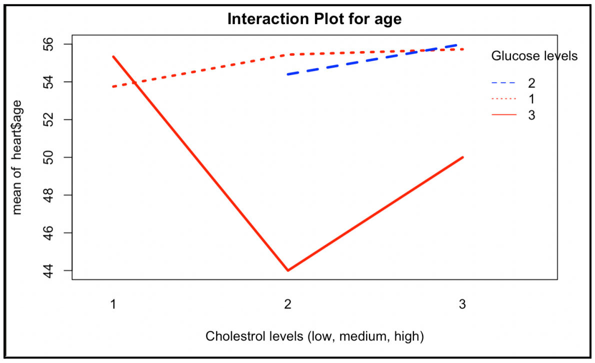
MANOVA

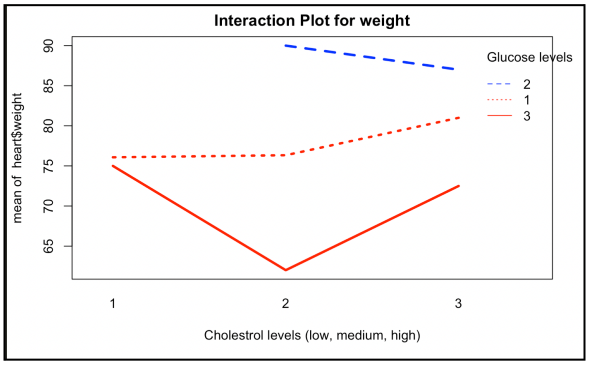
One-Way MANOVA

For our MANOVA analysis, we used categorical variables glucose levels and cholesterol levels [low (1), medium (2), high (3)], and continuous variables systolic blood pressure, weight, height, and age.



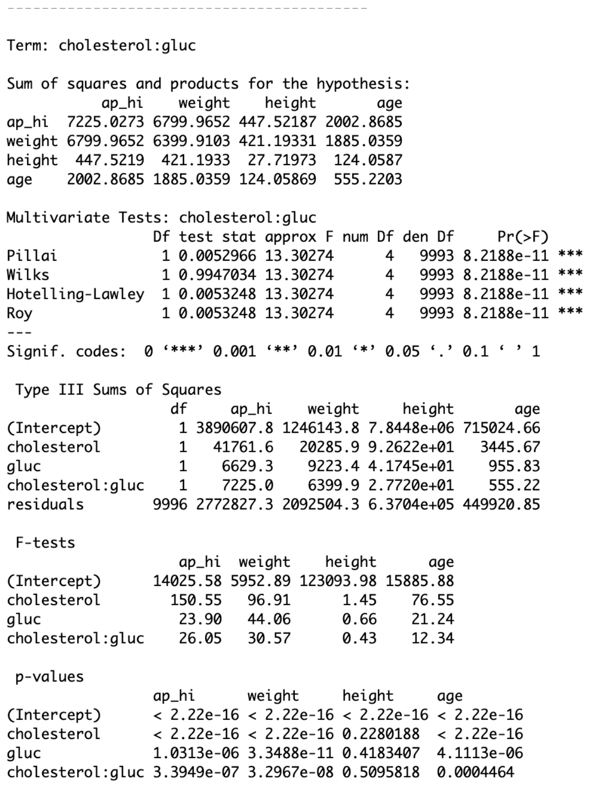




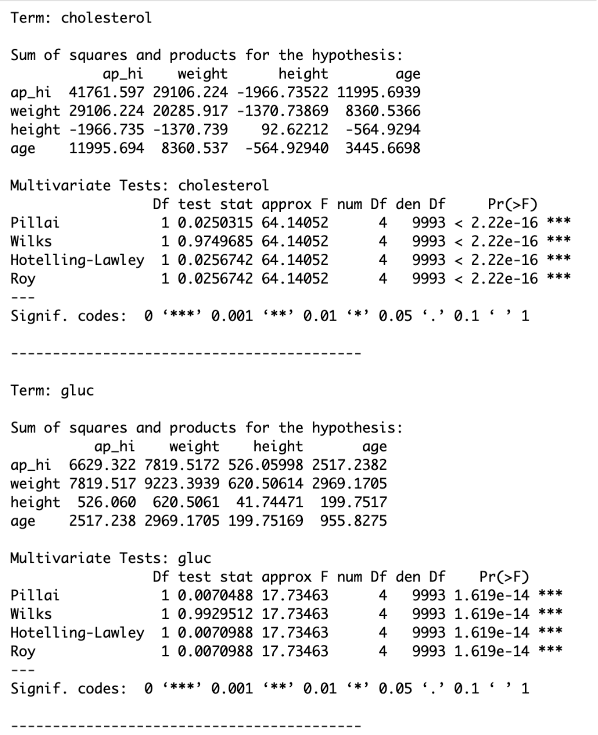


The interaction plots show many intersecting lines and few parallel lines, so there appears to be possible interaction effects between the response variables of systolic blood pressure, weight, height, and age. The first plot shows that regardless of glucose level, the general trend is that as cholesterol levels increased, so did systolic blood pressure, but at different rates. The second plot shows that for glucose levels of 2 and 3, the height reaches its maximum at cholesterol level of 2 and its minimum at cholesterol level of 3, while glucose level of 1 cholesterol level increases we see a decrease in height. The third plot shows that for glucose levels of 1 and 2, as cholesterol levels increase we see age increase, while for glucose level of 3, age dips at cholesterol level of 2. The fourth plot shows glucose level of 3 reaching peak weight at cholesterol level of 2, while glucose levels of 1 and 2 increase in weight as cholesterol levels increase.

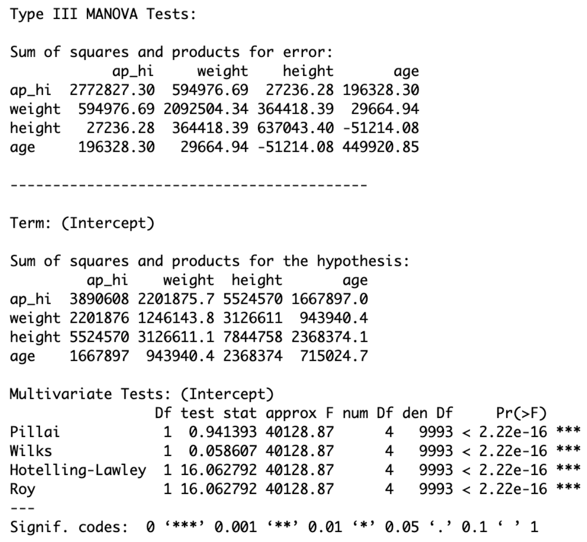
Two-Way MANOVA



```
------------------------------------------

Term: cholesterol:gluc

Sum of squares and products for the hypothesis:
          ap_hi    weight    height       age
ap_hi  7225.0273 6799.9652 447.52187 2002.8685
weight 6799.9652 6399.9103 421.19331 1885.0359
height  447.5219  421.1933  27.71973  124.0587
age    2002.8685 1885.0359 124.05869  555.2203

Multivariate Tests: cholesterol:gluc
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1 0.0052966 13.30274      4   9993 8.2188e-11 ***
Wilks             1 0.9947034 13.30274      4   9993 8.2188e-11 ***
Hotelling-Lawley  1 0.0053248 13.30274      4   9993 8.2188e-11 ***
Roy               1 0.0053248 13.30274      4   9993 8.2188e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Type III Sums of Squares
                   df     ap_hi    weight     height       age
(Intercept)         1 3890607.8 1246143.8 7.8448e+06 715024.66
cholesterol         1   41761.6   20285.9 9.2622e+01   3445.67
gluc                1    6629.3    9223.4 4.1745e+01    955.83
cholesterol:gluc    1    7225.0    6399.9 2.7720e+01    555.22
residuals        9996 2772827.3 2092504.3 6.3704e+05 449920.85

 F-tests
                   ap_hi  weight    height      age
(Intercept)     14025.58 5952.89 123093.98 15885.88
cholesterol       150.55   96.91      1.45    76.55
gluc               23.90   44.06      0.66    21.24
cholesterol:gluc   26.05   30.57      0.43    12.34

 p-values
                 ap_hi      weight     height     age
(Intercept)      < 2.22e-16 < 2.22e-16 < 2.22e-16 < 2.22e-16
cholesterol      < 2.22e-16 < 2.22e-16 0.2280188  < 2.22e-16
gluc             1.0313e-06 3.3488e-11 0.4183407  4.1113e-06
cholesterol:gluc 3.3949e-07 3.2967e-08 0.5095818  0.0004464
```



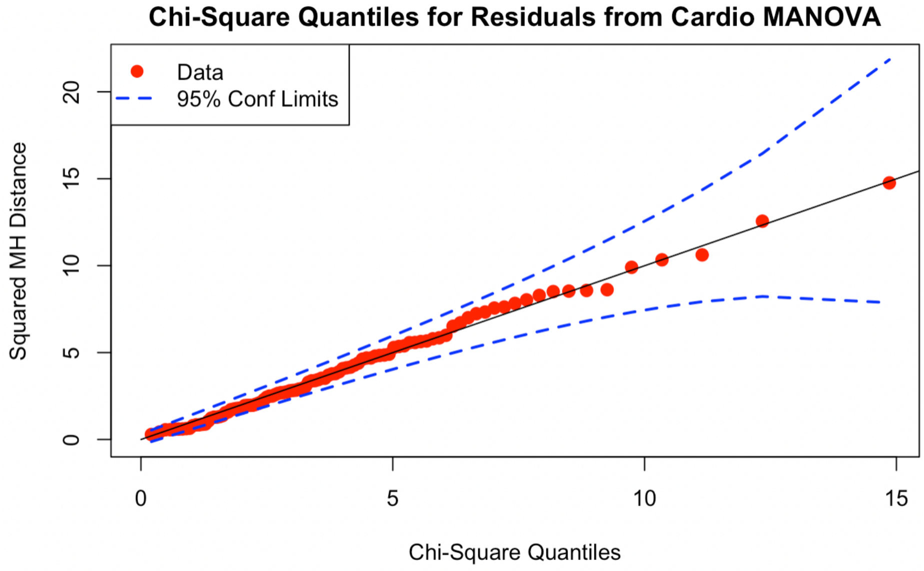
```
Term: cholesterol

Sum of squares and products for the hypothesis:
           ap_hi    weight      height        age
ap_hi  41761.597 29106.224 -1966.73522 11995.6939
weight 29106.224 20285.917 -1370.73869  8360.5366
height -1966.735 -1370.739    92.62212  -564.9294
age    11995.694  8360.537  -564.92940  3445.6698

Multivariate Tests: cholesterol
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1 0.0250315 64.14052      4   9993 < 2.22e-16 ***
Wilks             1 0.9749685 64.14052      4   9993 < 2.22e-16 ***
Hotelling-Lawley  1 0.0256742 64.14052      4   9993 < 2.22e-16 ***
Roy               1 0.0256742 64.14052      4   9993 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------

Term: gluc

Sum of squares and products for the hypothesis:
          ap_hi    weight    height       age
ap_hi  6629.322 7819.5172 526.05998 2517.2382
weight 7819.517 9223.3939 620.50614 2969.1705
height  526.060  620.5061  41.74471  199.7517
age    2517.238 2969.1705 199.75169  955.8275

Multivariate Tests: gluc
                 Df test stat approx F num Df den Df    Pr(>F)
Pillai            1 0.0070488 17.73463      4   9993 1.619e-14 ***
Wilks             1 0.9929512 17.73463      4   9993 1.619e-14 ***
Hotelling-Lawley  1 0.0070988 17.73463      4   9993 1.619e-14 ***
Roy               1 0.0070988 17.73463      4   9993 1.619e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------
```



```
Type III MANOVA Tests:

Sum of squares and products for error:
            ap_hi     weight    height       age
ap_hi  2772827.30  594976.69  27236.28 196328.30
weight  594976.69 2092504.34 364418.39  29664.94
height   27236.28  364418.39 637043.40 -51214.08
age     196328.30   29664.94 -51214.08 449920.85

------------------------------------------

Term: (Intercept)

Sum of squares and products for the hypothesis:
         ap_hi    weight  height       age
ap_hi  3890608 2201875.7 5524570 1667897.0
weight 2201876 1246143.8 3126611  943940.4
height 5524570 3126611.1 7844758 2368374.1
age    1667897  943940.4 2368374  715024.7

Multivariate Tests: (Intercept)
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1  0.941393 40128.87      4   9993 < 2.22e-16 ***
Wilks             1  0.058607 40128.87      4   9993 < 2.22e-16 ***
Hotelling-Lawley  1 16.062792 40128.87      4   9993 < 2.22e-16 ***
Roy               1 16.062792 40128.87      4   9993 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the multivariate tests, we reject the null hypothesis for each of our multivariate methods, since all of our p-values are significantly lower than the alpha level of 0.05. A very small p-value (e.g., less than 0.05 or 0.01) indicates strong evidence against the null hypothesis and suggests that there are statistically significant differences among the groups in terms of the dependent variables included in the analysis. We do see interactions.

From the univariate tests, we see similar results of low p-values, indicating interactions.



Our Chi-Squre Quantiles for Residuals appears to be roughly linear, which means our data has a multivariate normal distribution, suggesting that MANOVA is appropriate for our dataset.

Conclusion

In this study, we analyzed factors related to heart disease collected by Kuzak Dempsy (2021), and we used statistical methods to determine relationships between the variables and risk factors. Through PCA, we were able identify the direction and strength of the correlation between variables, and clustering aided in the visualization of these patterns and structures within the data.

Another interesting tool we applied is that for our clustering analysis, we chose to sample 20 of the participants so that instead of having to analyze all observations, which were in the tens of thousands, we chose a subset to look at the individual differences. Even though we looked at only a few observations in the grand scheme of things, if we were to keep sampling, we would probably be able to find some adumbrate pattern amongst individuals. Furthermore, we looked at PCA to see how the data can be clustered into three sections. After looking at analyzing at the correlation, we were then able to see which ones were most highly correlated with each PC axis.

We found that MANOVA determined that there was statistically significant difference between groups. We used both one-way and two-way MANOVA and checked on normality of residuals to prove the validity of our findings.

Overall, PCA, cluster analysis, and MANOVA are able to help us decipher information from our very large data set of multiple variables. Despite the intimidating amount of data that is presented, we were able to draw conclusions, important variables, as well as, correlations amongst variables, and all-in-all have a greater understanding of the risk factors for heart disease.