1.0 METHODS

The participants were a random sample of 500 from an annual telephone survey of 350,000 people in the United States.

Of the nine (9) variables, genhealth, and gender houses categorical values while exercise, healthplan, smoke100, height, weight, weightdsr, and age contains numerical values. Numerical operations are not expected on the exercise, healthplan and smoke100 columns, therefore the need to convert to categorical variables (strings). The as.character() method was used on the three (3) columns to make the type conversion.

To better understand if a respondent is overweight or underweight, there is a need to calculate their respective Body Mass Index (BMI) which is calculated using the respondents weight and height. The BMI for each respondents was obtained and were further categorized into underweight (BMI below 18.5), Health Weight (BMI between 18.5 and 24.9), Overweight (BMI between 25.0 and 29.9) and Obesity (BMI between 30.0 and Above) [1].

The need to also group the ages arose since questions arose by getting to know if there are age categories that smoke or exercise.

Since this research seeks to answer questions with values that are categorical in nature, the chi square statistical test is the best to ascertain our claim.

1.1 CHI SQUARE

Chi-square is a non-parametric statistical test used to examine the differences between categorical variables from a random sample in order to judge the goodness of fit between expected and observed results. It can also be said to be a test normally used to check if two categorical variables are related or independent.

It is important to ensure the data to be used meet up to the assumption. The assumptions of Chi Square test are [5]:

1) Both variables must be categorical
2) All observations are independent
3) Cells in the contingency table are mutually exclusive
4) The expected value of cells should be 5 or greater in at least 80% of cells

1.1.1 Test of Hypothesis

The Chi Square test of independence uses the following hypotheses to draw conclusion:

H_0 (null hypothesis): The two variables under study are independent

H_1 (alternative hypothesis): The two variables under study are not independent

1.1.2 Test Statistics

Chi Square has a test statistic which is obtained using the expression [4]:

$$\chi^2_{calc} = \sum_{i}^{n}\sum_{j}^{n} \frac{(o_{ij} - E_{ij})^2}{E_{ij}} \qquad (2.1)$$

Where,

O_{ij} is the observed value

E_{ij} is the expected value

The test statistic χ^2_{calc} follows a Chi Square distribution $\chi^2_{d.f,\alpha}$, where d.f stands for degree of freedom which is obtained using: (number of rows - 1) * (number of columns - 1)

1.1.3 Decision Rule

The conclusion can either be made using the chi square table or the p-value.

When conclusion is made using the chi square table, we reject H_0 if:

$\chi^2_{calc} > \chi^2_{d.f,\alpha}$ and accept H_0 if: $\chi^2_{calc} < \chi^2_{d.f,\alpha}$.

When statistical software is used for calculation, we resort to using the p-value and the significance level (α) to make our conclusion. Our decisions then becomes:

Accept H_0 if p-value > significance level (α), and

Reject H_0 if p-value < significance level (α).

1.1.4 Contingency Coefficient

Whenever we reject H_0 and conclude that there is association between the two variables, one might decide to know how strong is the relationship between the two variables, and that can be achieved using the contingency coefficient.

The contingency coefficient value can be obtained using the expression:

$$\sqrt{\frac{\chi^2}{\chi^2+n}} \qquad (2.2)$$

2.0 RESULTS

This section seeks to explain the summary of the analysis results alongside some visualizations to back up our claim.

2.1 Description of Participants Characteristics

Of the random sample of 500 respondents , it was observed that 51.4% were female and 48.6% were male.

Upon grouping of the respondents into three categories by age (i.e 18-44 years, 45-64 years and over 64 years), 53.4% of respondents fell within the 18-44 years category [3], 30.4% of respondents fell within the 45-64 years category, while 16.2% of respondents are over the age of 64. It is also important to note that none of the respondents were below the age of 18.

It was also observed that out of the 500 respondents, 11 fell into the Underweight category, 184 fell into the Healthy Weight category, 197 fell into the Overweight category, and the remaining108 fell into the Obesity category

It was also observed that the mean body mass index (BMI) of respondents who have smoked at least 100 cigarettes in their entire life is approximately equal with those who haven't smoked at 100 cigarettes in their entire life.

It was also evident from the visualization that 12.8% of respondents do not have any form of health coverage while 87.2% of respondents do have some form of health coverage.

The importance of exercise to healthy living cannot be over emphasized. From the visualization, it was observed that 29% of the respondents haven't exercised in the past month while 71% of the respondents exercised in the last month which could be responsible for the least value encountered in the poor category of respondents' general health status.

It was also evident that the number of respondents who exercised in the last month was greater than those who haven't exercised in the past month for all the health status categories except for the poor health status category.

It is evident from the visualization that 51.6% of respondents have smoked less than 100 cigarettes while 48.4% of the respondents have smoked at least 100 cigarettes in their entire life. Of the 51.6% who have smoked less than 100 cigarettes, a larger percentage are female and for the 48.4% who have smoked at least 100 cigarettes in their entire life, the majority are male.

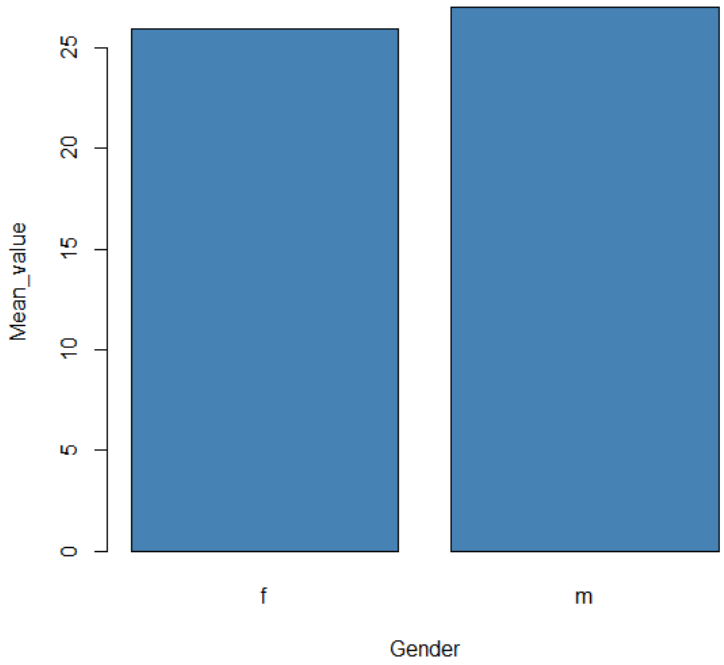


Fig 2.1: Bar Chart Plot of the Mean Body Mass Index (BMI) by Gender

The visual analysis above reveals that the average body mass index (BMI) for males exceeds that of females. Given the associated mean BMI value surpasses 25, it suggests that the mean BMI values for both genders are situated within the Overweight category.

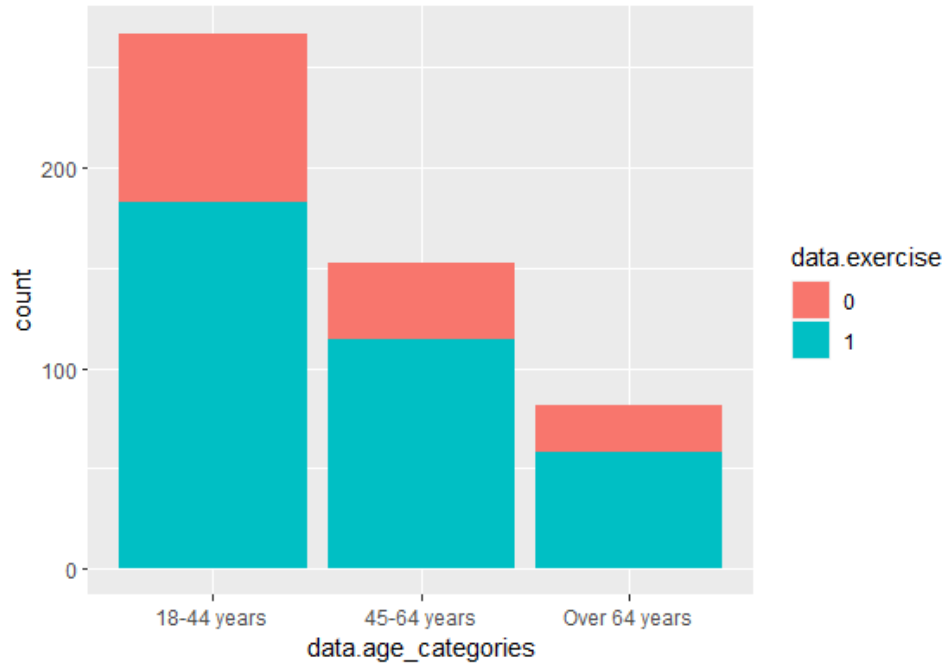


Fig 2.2: Stacked Bar Chart of the Age Categories and Respondent Exercise Level

The above visualization reveals that of the 53.4% of respondents who fell within the 18-44 years category, 30.4% of respondents who fell within the 45-64 years category, 16.2% of respondents who are over the age of 64, over 65% of them have exercised in the past month across each categories.

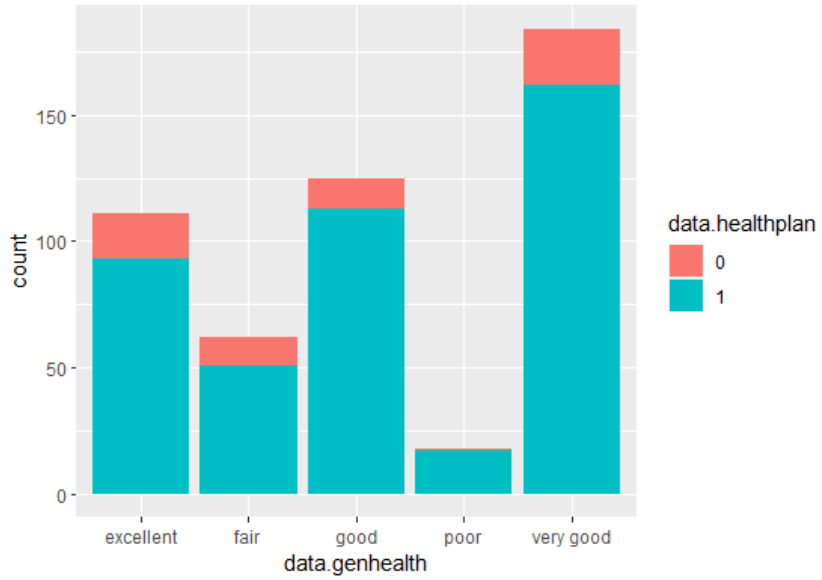


Fig 2.3: Stacked Bar Chart of Respondent General Health Status and their Health Plan

The above stacked bar chart explains that less than 10% of respondents in each category of health status don't have any form of health coverage.

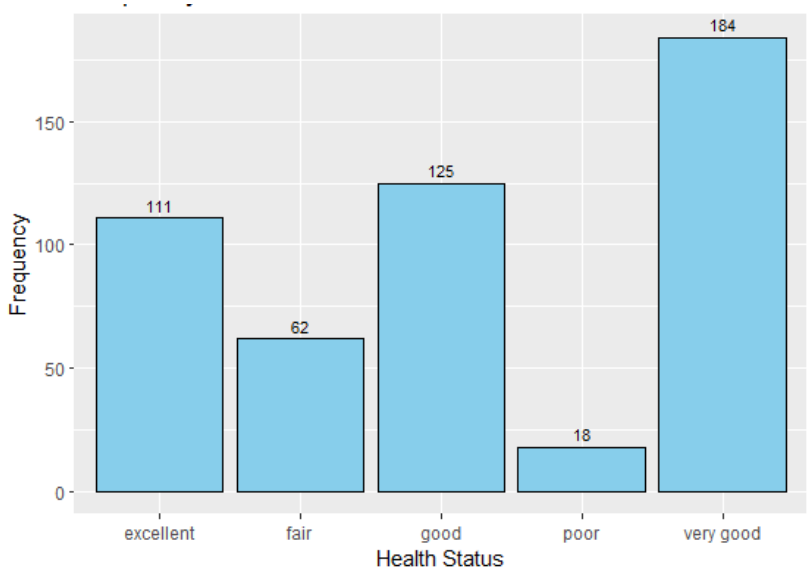


Fig 2.4: Frequency Visualization of Respondent's Health Status

The above visualization explains that of the 500 respondents, 22.2% health status is excellent, 36.8% health status is very good, 25% health status is good, 12.4% health status is fair and 3.6% health status is poor.

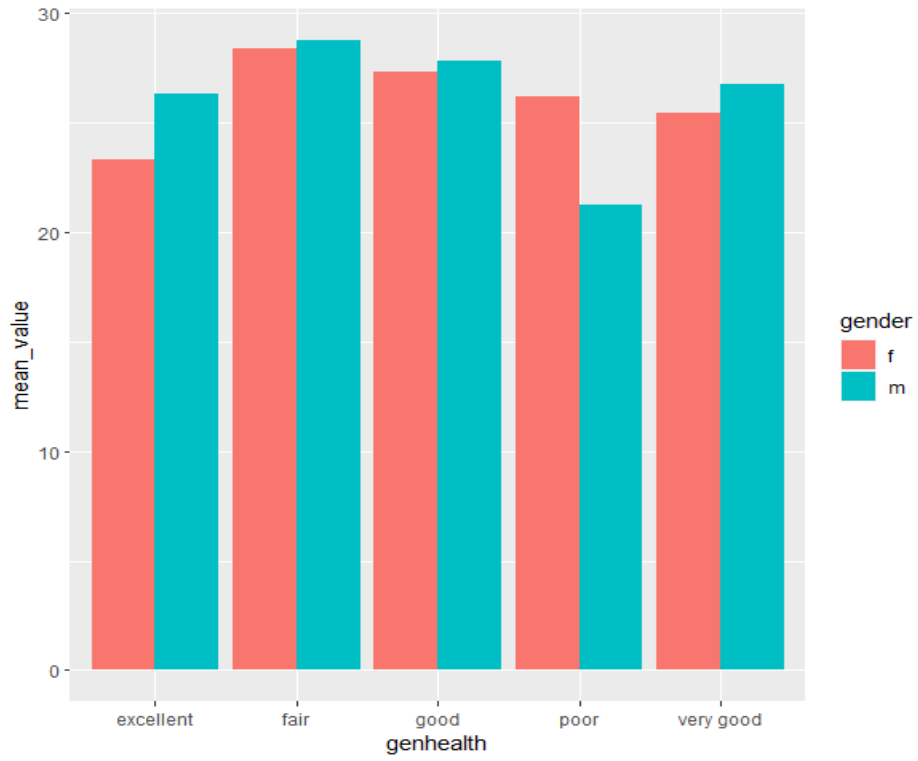


Fig 2.5: Clustered Bar Chart of Mean Body Mass Index (BMI) by General Health Status and Gender

The above clustered bar chart indicates that the mean body mass index of male is greater than that of females irrespective of the health status except for categories of respondents whose health status is poor.

Inferential Statistics Result

This section outlines the results of all inferential analysis done.

2.2 The Participant's Perspective on Whether they are Satisfied with their Current Weight

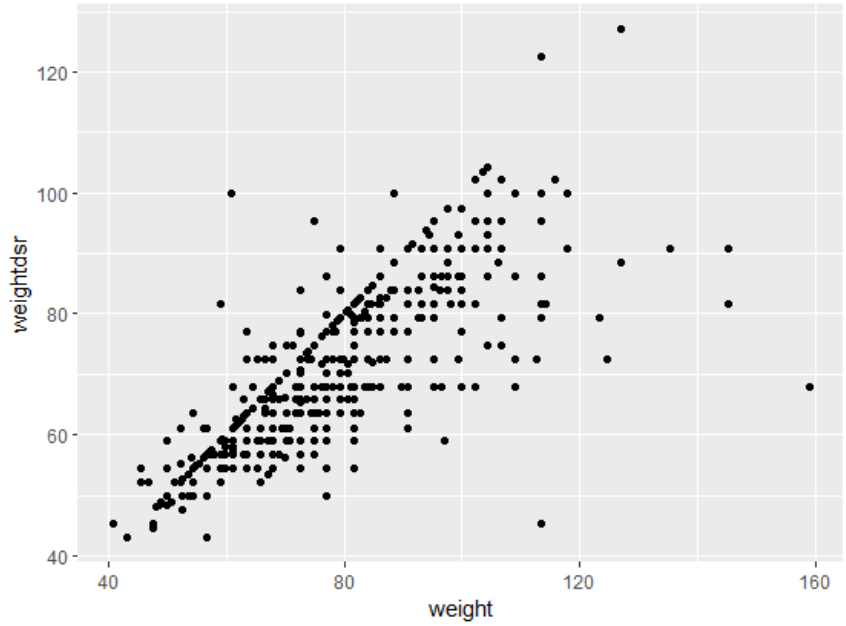


Fig 2.6: Scatterplot Visualizing the Correlation Between the Respondent Current Weight and their Desired Weight

From the above scatter plot visualization of respondents actual weight and their corresponding desired weight, we could see an upward trend which indicates the positive correlation between a respondents current weight and their actual weight.

The visualization corresponds with the correlation value of 0.7748008 (using the cor() method on the desired weight and current weight column) which indicates a strong positive correlation.

2.3 Investigating the Univariate and Multivariate Relationship Between Gender and Other Variables with Weight Self-Perception

2.3.1 Chi Square Analysis to Check if Smoking is Related to Respondent Health

H_0 (null hypothesis): General Health Status and Smoking are independent

H_1 (alternative hypothesis): General Health Status and Smoking are not independent

Table 2.1: Chi Square Test of Independence for General Health Status and Smoking

Pearson's Chi-squared test		
data: g_BMI		
X-squared = 9.9778	df = 4	p-value = 0.0408

Decision: Since the p-value (0.0408) of the test is less than 0.05, we reject the null hypothesis. This means that there is an association between general health status and smoking.

2.3.2 Chi Square Analysis to Check If General Health Status is Related to Exercise

H_0 (null hypothesis): General Health Status and Exercise are independent

H_1 (alternative hypothesis): General Health Status and Exercise are not independent

Table 2.2: Chi Square Test of Independence for General Health Status and Exercise

Pearson's Chi-squared test
data: g_exer

X-squared = 47.574	df = 4	p-value = 1.158e-09

Decision: Since the p-value (1.158e-09) of the test is less than 0.05, we reject the null hypothesis. This means that there is an association between exercising within the past month and general health status.

2.3.3 Chi Square Analysis to Check if Gender is Related to Respondent Health

H_0 (null hypothesis): General Health Status and Gender are independent

H_1 (alternative hypothesis): General Health Status and Gender are not independent

Table 2.3: Chi Square Test of Independence for Gender and General Health Status

Pearson's Chi-squared test		
data: g_health		
X-squared = 12.157	df = 4	p-value = 0.01622

Decision: Since the p-value (0.01622) of the test is less than 0.05, we reject the null hypothesis. This means that there is an association between gender and general health status.

2.3.4 Chi Square Analysis to Check if Gender is Related to Exercise

H_0 (null hypothesis): Gender and Exercise are independent

H_1 (alternative hypothesis): Gender and Exercise are not independent

Table 2.4: Chi Square Test of Independence for Gender and Exercise

Pearson's Chi-squared test with Yates' continuity correction		
data: gen_exer		
X-squared = 9.9171	df = 1	p-value = 0.001638

Decision: Since the p-value (0.001638) of the test is less than 0.05, we reject the null hypothesis. This means that there is an association between gender and exercise.

2.4 Some Results From the Appendix

From Table 1 (in Appendix), since the p-value (0.1115) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between gender and smoking.

In other words, gender and smoking are independent.

From Table 2 (in Appendix), since the p-value (0.0001835) of the test is less than 0.05, we reject the null hypothesis. This means that there is an association between general health status and body mass index categories.

In other words, general health status and body mass index categories are not independent.

From Table 3 (in Appendix), since the p-value (0.2507) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between age categories and smoking.

In other words, age categories and smoking are independent.

From Table 4 (in Appendix), since the p-value (1) of the test is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that there is an association between gender and health plan.

In other words, gender and health plan are independent.

From Table 5 (in Appendix), since the p-value (0.0002406) is less than 0.05, we reject the null hypothesis. This means that there is an association between gender and body mass index categories.

In other words, we conclude that gender and body mass index categories are not independent.

2.5 What is the confidence interval for the population mean of body mass index

Table 2.5: Confidence Interval for the Population Mean of Body Mass Index

25.98495	26.91861

Based on the 500 random sample data, the result [25.98495 26.91861] shows the body mass index mean's computed interval. We can say with 95% degree of certainty that the true population body mass index means is within the range 25.98495 to 26.91861

3.0 DISCUSSION

This section provides a clear conclusion based on the results in the result section.

From the sample data set used for the analysis, we have the following conclusion.

As a result of the association between a respondent's health status and smoking, we can tell a person's health status if we know the number of cigarettes the person has taken.

We can be right to say that a large percentage of respondents are happy with their current weight as a result of the positive correlation between their desired weight and their current weight.

The true population mean of the respondent's body mass index is between 25.98495 to 26.91861.

The majority of respondents in the sample data are within the age range 18-44 years.

We could deduce from the visualizations that men smoke the most.

It can also be said that having some form of health coverage determines the health status of a person

3.1 Limitation of the study

The sample size of 500 respondents used in this study is extremely small to make a representation of the 350,000 people who participated in the telephone survey. We can back up our claim by obtaining the sampling fraction which is a ratio of the sample size to the population size leaving us with a value of 0.0014.

We are limited by the number of features available to ascertain our claim on factors that contribute to the health and weight of individuals in the United States.

3.2 Recommendations

To make better conclusions, the sample size should be increased to accommodate as many attributes of people that would be a true representation of the population of people living in the United States.

Features like respondent's occupation, academic background, marital status, number of children and many more should be considered while conducting surveys so as to generally understand factors that truly contribute to health status and weight amongst gender.

To strengthen our claims on association of variable, obtaining the contingency coefficient is also advisable

REFERENCES

[1] *All About Adult BMI*. (2022, June 3). Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[2] *CDC - BRFSS*. (n.d.). https://www.cdc.gov/brfss/index.html

[3] *Demographics of the United States*. (2023, November 24). Wikipedia. https://en.wikipedia.org/wiki/Demographics_of_the_United_States

[4] Hayes, A. (2023, May 22). *Chi-Square (χ2) Statistic: What It Is, Examples, How and When to Use the Test*. Investopedia. https://www.investopedia.com/terms/c/chi-square-statistic.asp#:~:text=Chi%2Dsquare%20is%20a%20statistical,between%20expected%20and%20observed%20results.

[5] Z. (2021, August 14). *The Four Assumptions of a Chi-Square Test*. Statology. https://www.statology.org/chi-square-test-assumptions/

APPENDIX

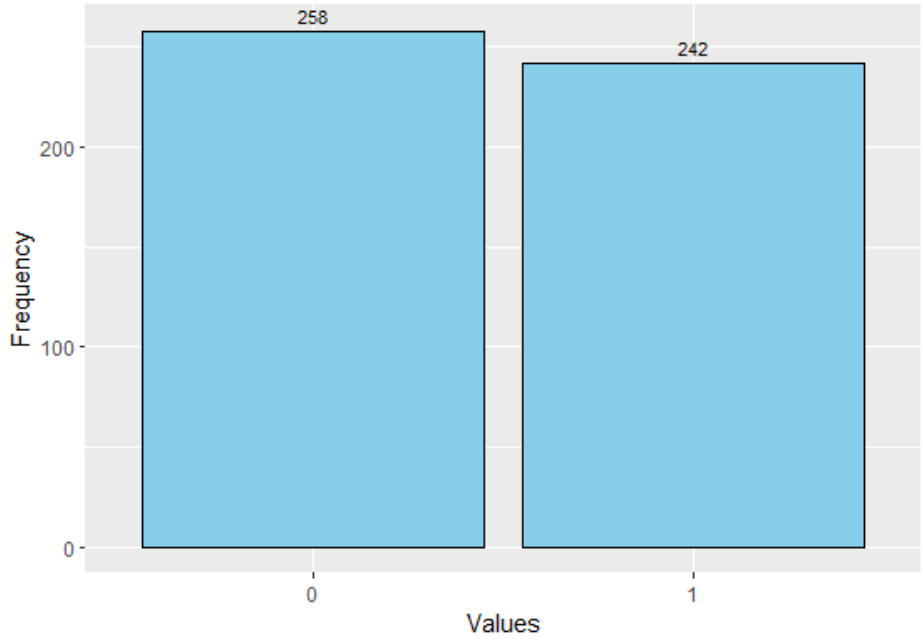


Fig 1: Frequency Plot of Respondent Smoking Level

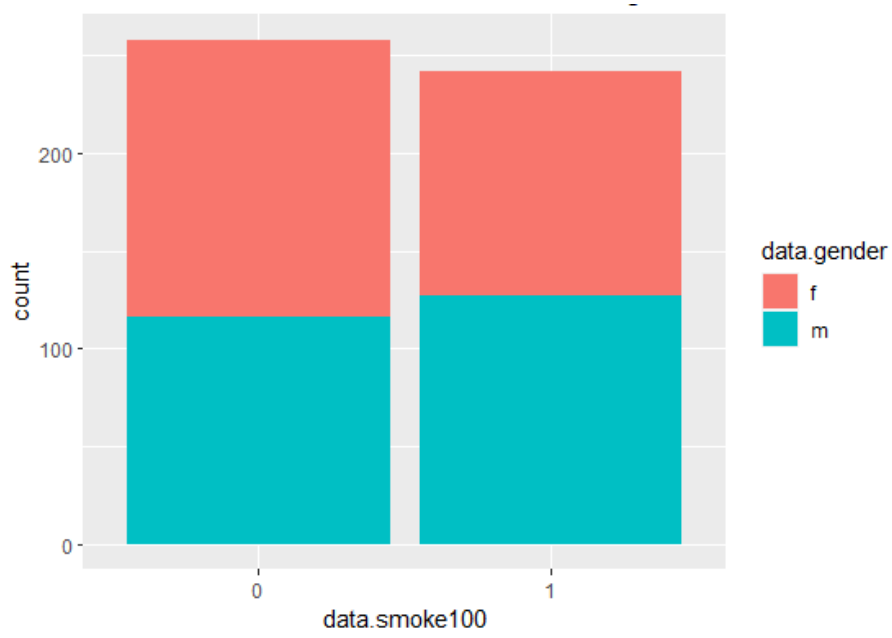


Fig 2: Stacked Bar Chart of Respondent Smoking Level by Gender

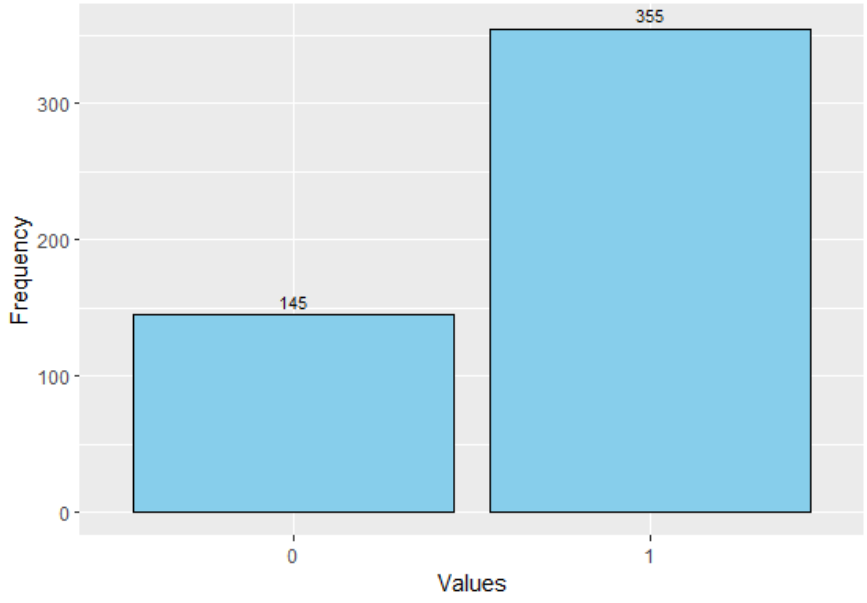


Fig 3: Frequency Plot of Respondent Level of Exercise

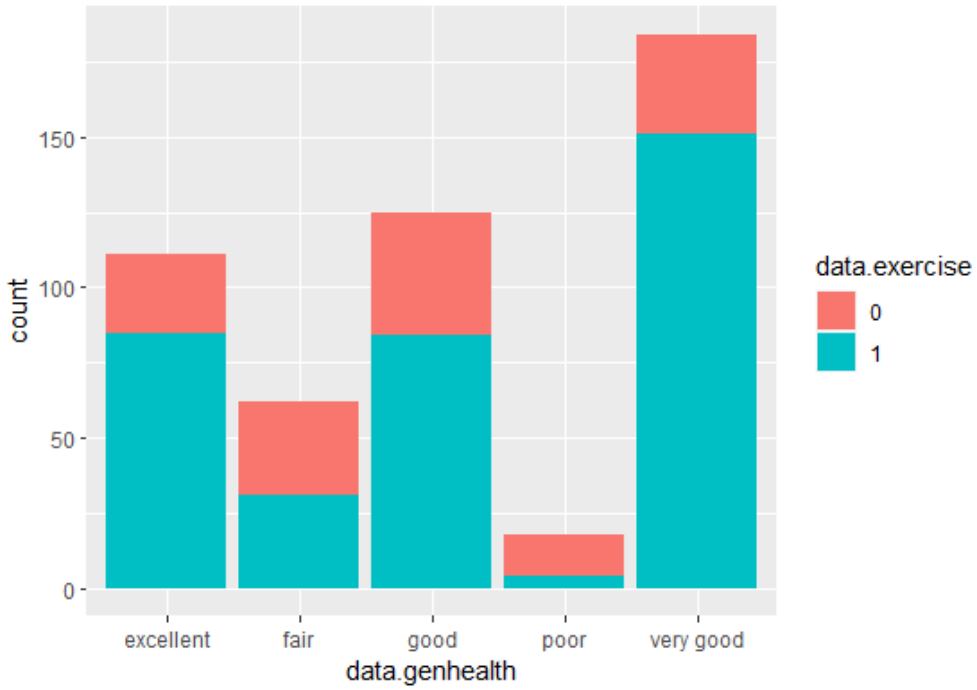


Fig 4: Stacked Bar Chart of Respondent General Health Status by Exercise Level

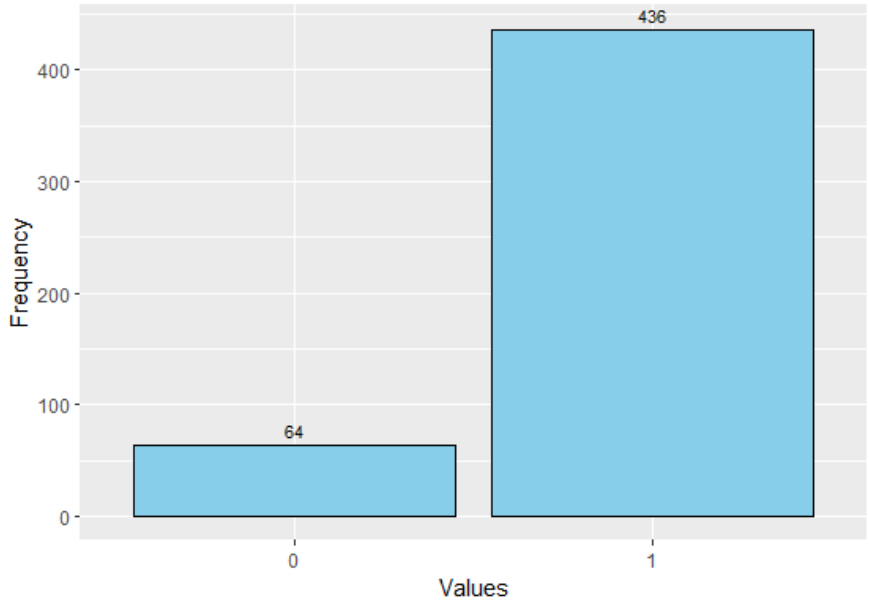


Fig 5: Frequency Plot of Respondent Health Plan

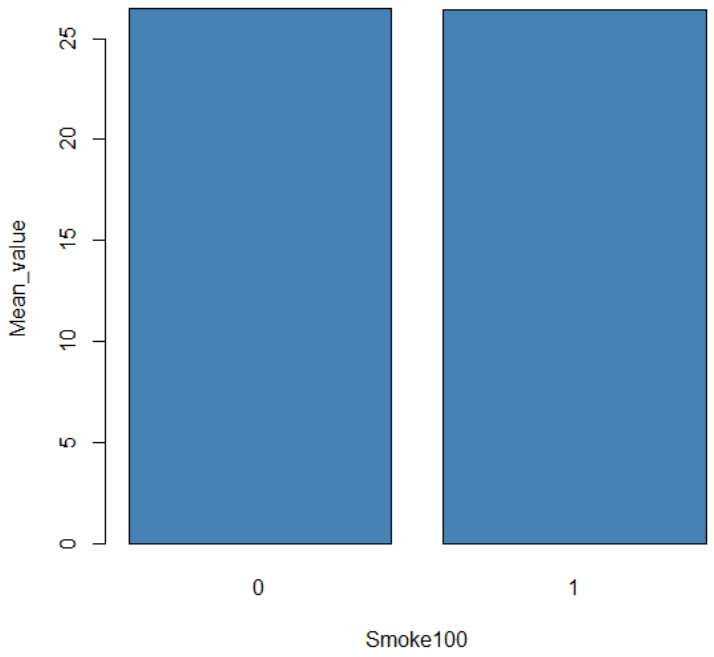
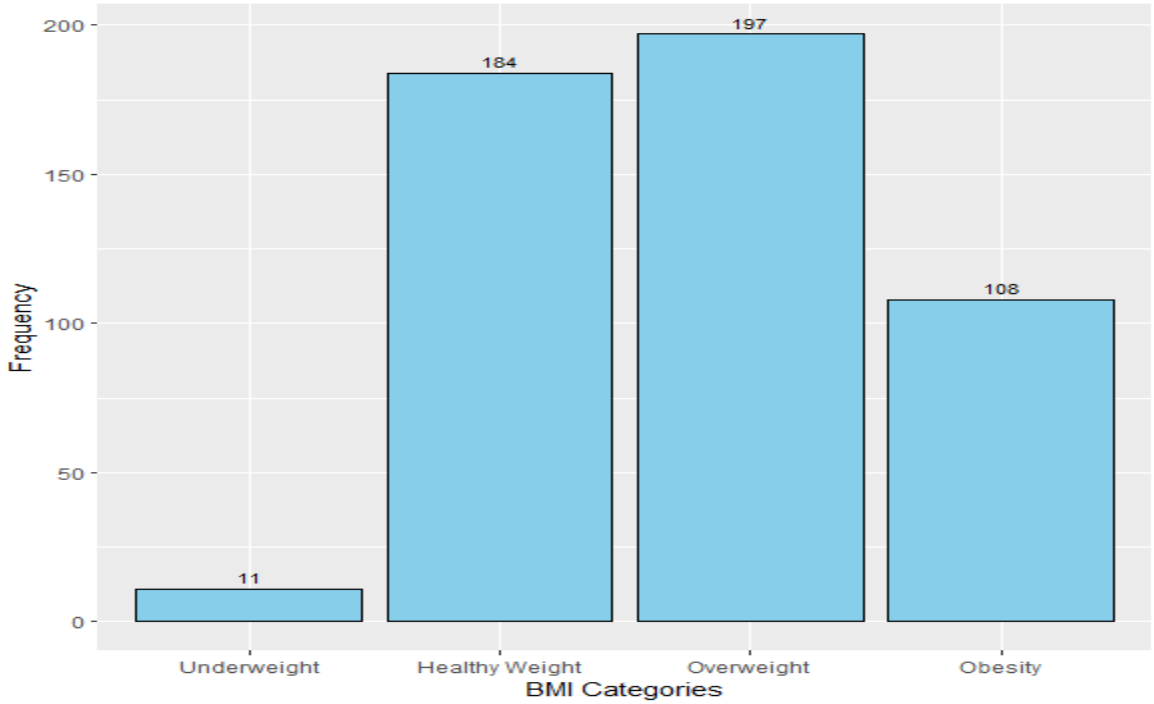


Fig 6: Bar Chart of Mean Body Mass Index (BMI) by Respondent Smoking Level



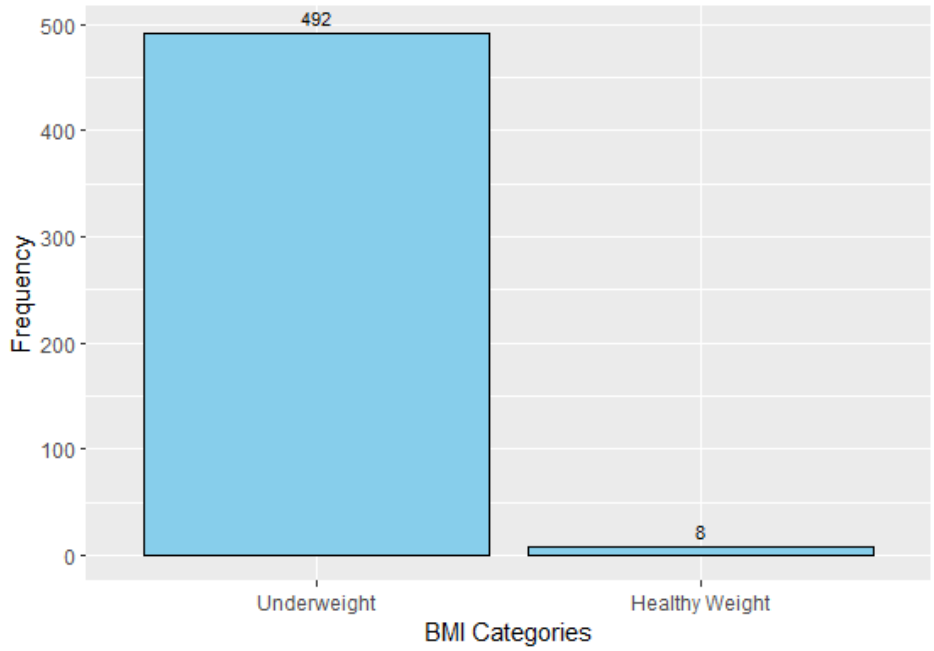


Fig 7: Frequency Plot of Respondent Body Mass Index (BMI) Categories

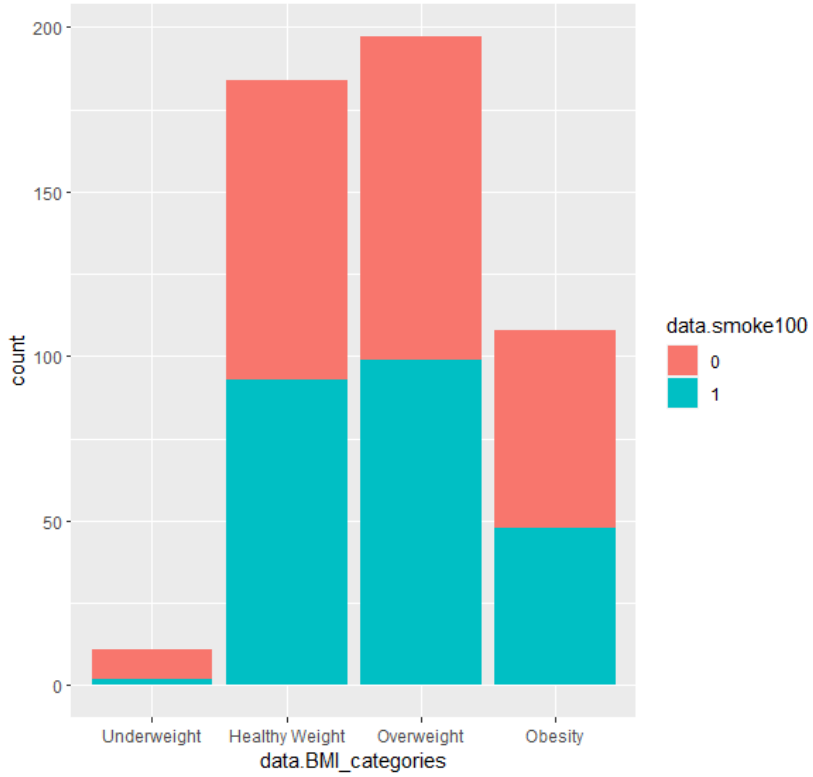


Fig 8: Stacked Bar Chart of Respondent Body Mass Index Categories by Smoking Level

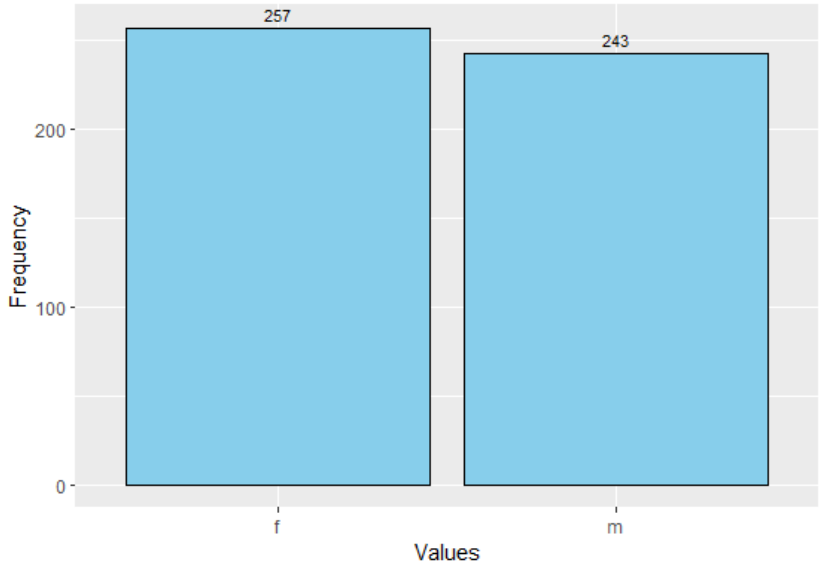


Fig 9: Frequency Plot of Respondent Gender

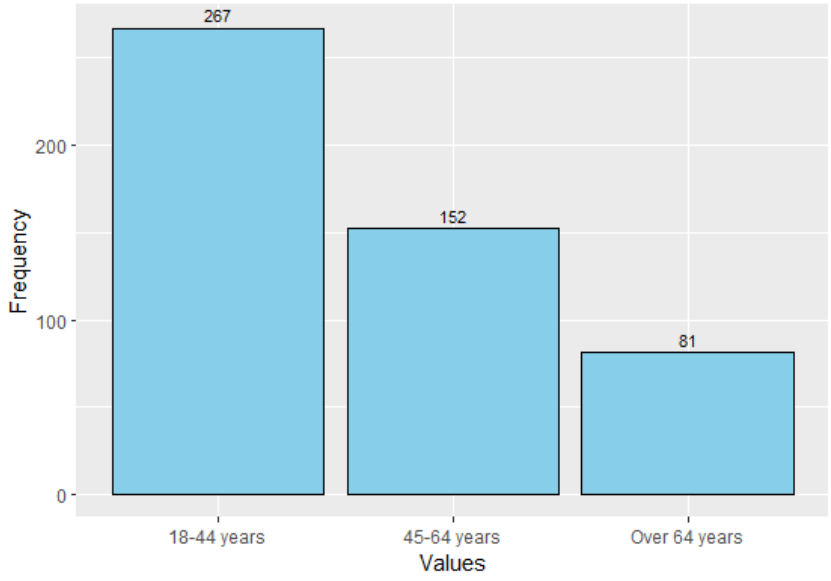


Fig 10: Frequency Plot of Respondent Age Category

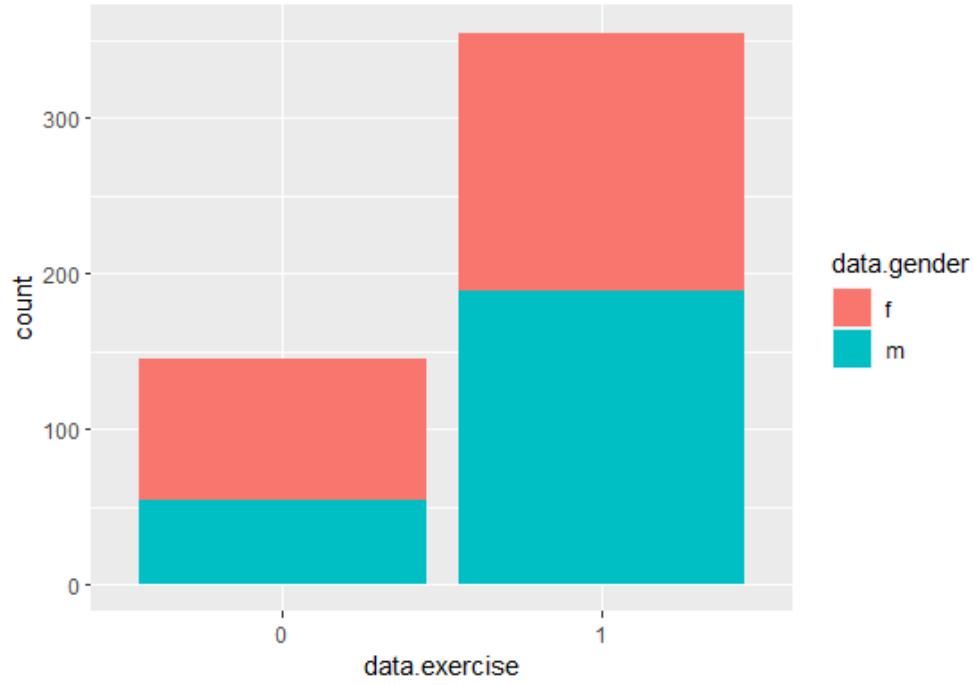


Fig 11: Stacked Bar Chart of Respondent Exercise Level by Gender

Table 1: Chi Square Test of Independence for Gender and Smoking

Pearson's Chi-squared test with Yates' continuity correction		
data: gender_smoke		
X-squared = 2.5325	df = 1	p-value = 0.1115

Table 2: Chi Square Test of Independence for General Health Status and Body Mass Index Categories

Pearson's Chi-squared test		
data: health_BMI		
X-squared = 37.526	df = 12	p-value = 0.0001835

Table 3: Chi Square Test of Independence for Age Categories and Smoking

Pearson's Chi-squared test with Yates' continuity correction		
data: a_smoke		
X-squared = 2.7667	df = 2	p-value = 0.2507

Table 4: Chi Square Test of Independence for Gender and Health Plan

Pearson's Chi-squared test with Yates' continuity correction		
data: g_plan		
X-squared = 1.8542e-30	df = 1	p-value = 1

Table 5: Chi Square Test of Independence for Gender and Body Mass Index Categories

Pearson's Chi-squared test		
data: gen_BMI		
X-squared = 19.268	df = 3	p-value = 0.0002406