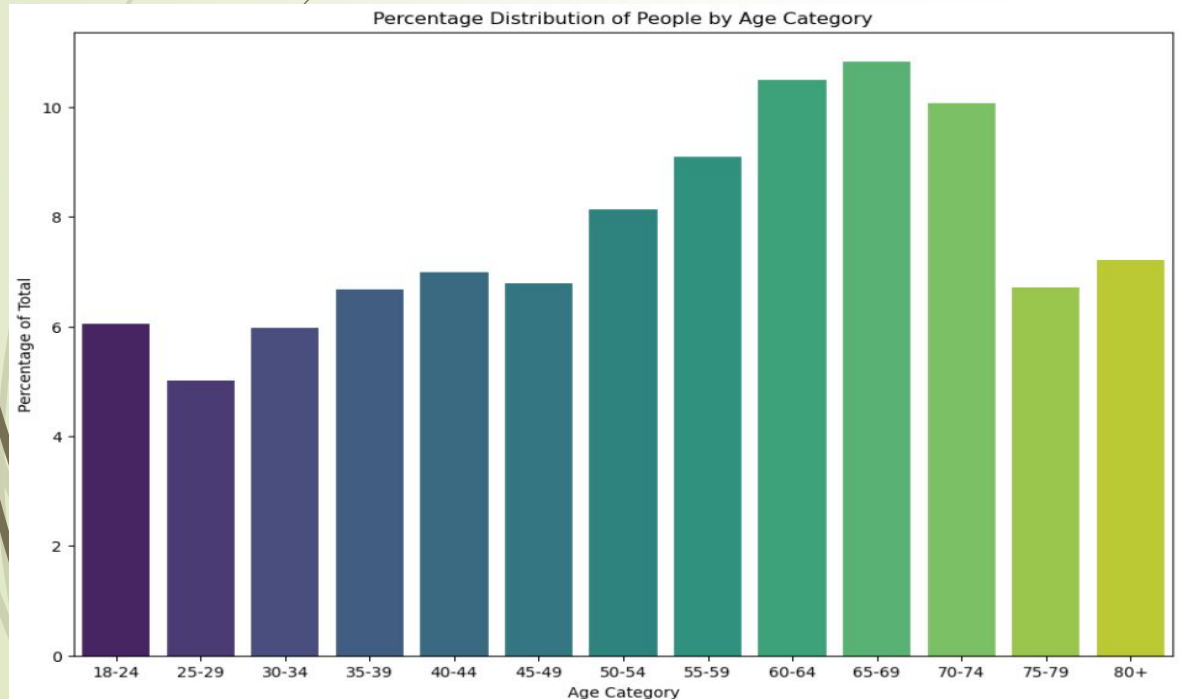




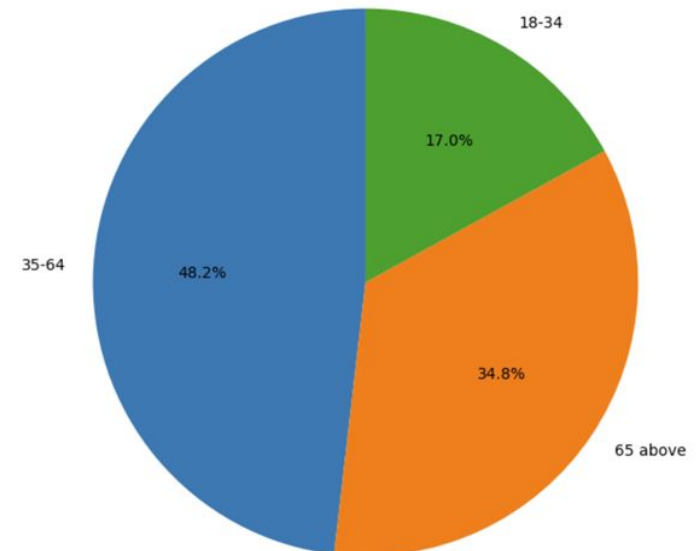
Analysis of Selected Health Issues

Sample Population

- There were **308,854** respondents in total across 13 age groups.
- The age category 65-69 has the highest representation in the dataset, followed closely by 60-64 and 70-74, while the 25-29 age group has the lowest count. The dataset exhibits a discernible age distribution pattern, with a concentration of individuals in the middle and early senior years.



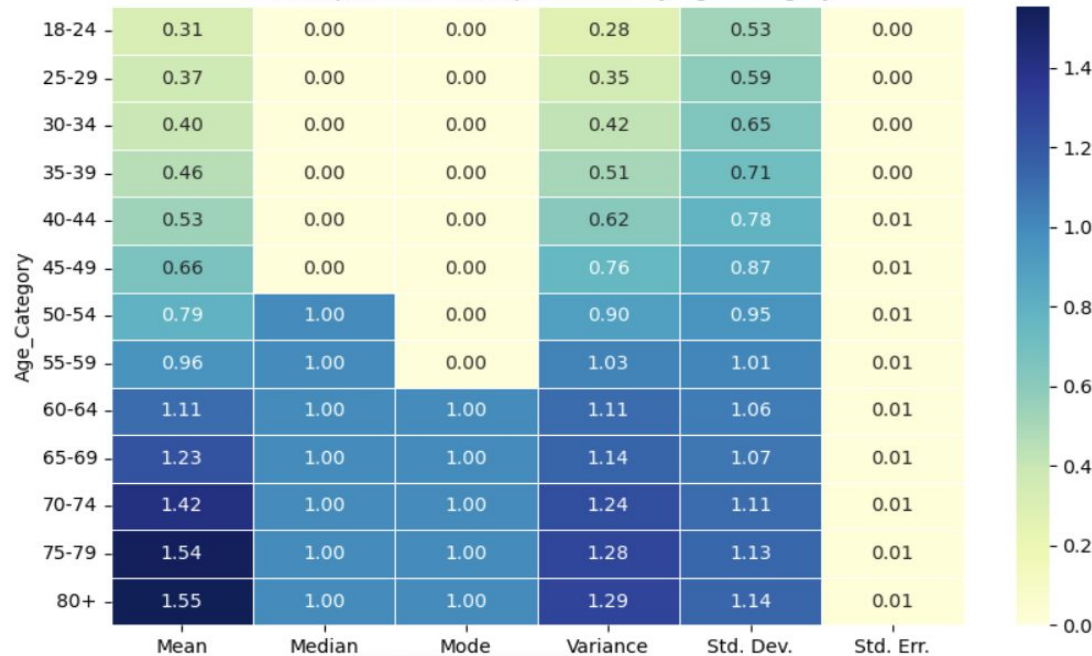
Distribution of Age Groups 18-34 (Youths), 35-64 (Adults), 65 above (Senior citizens))



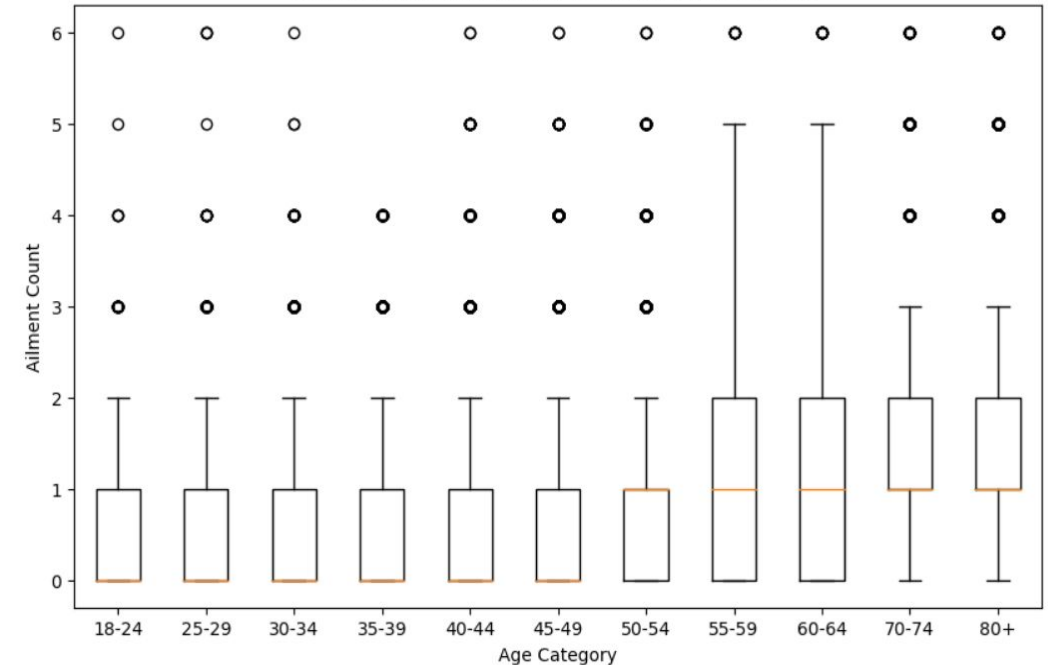
Summary Statistics- Health issue per Person

- The average number of health issues per person in our sample is around 1 (more precisely, 0.93); this gradually increases across age groups, with a notable acceleration in the 70-74 and 75-79 age categories. Median and Mode are both 1.
- Both the median and mode consistently shifted from 0 to 1, indicating a shift in the central tendency towards higher ailment counts in the older age groups.
- There is a moderate level of variability (as indicated by the variance and standard deviation) with a relatively small standard error, which suggests that sample means are likely to be close to the population mean.

Heat Map: Health issue per Person by Age Category



Box Plot: Health issue per Person by Age Category



Statistical Analysis

Exploring the Relationship between Alcohol Consumption and Health issue Per Person across all Reported Diseases and Age Groups: A Correlation Analysis

Null Hypothesis (H0):

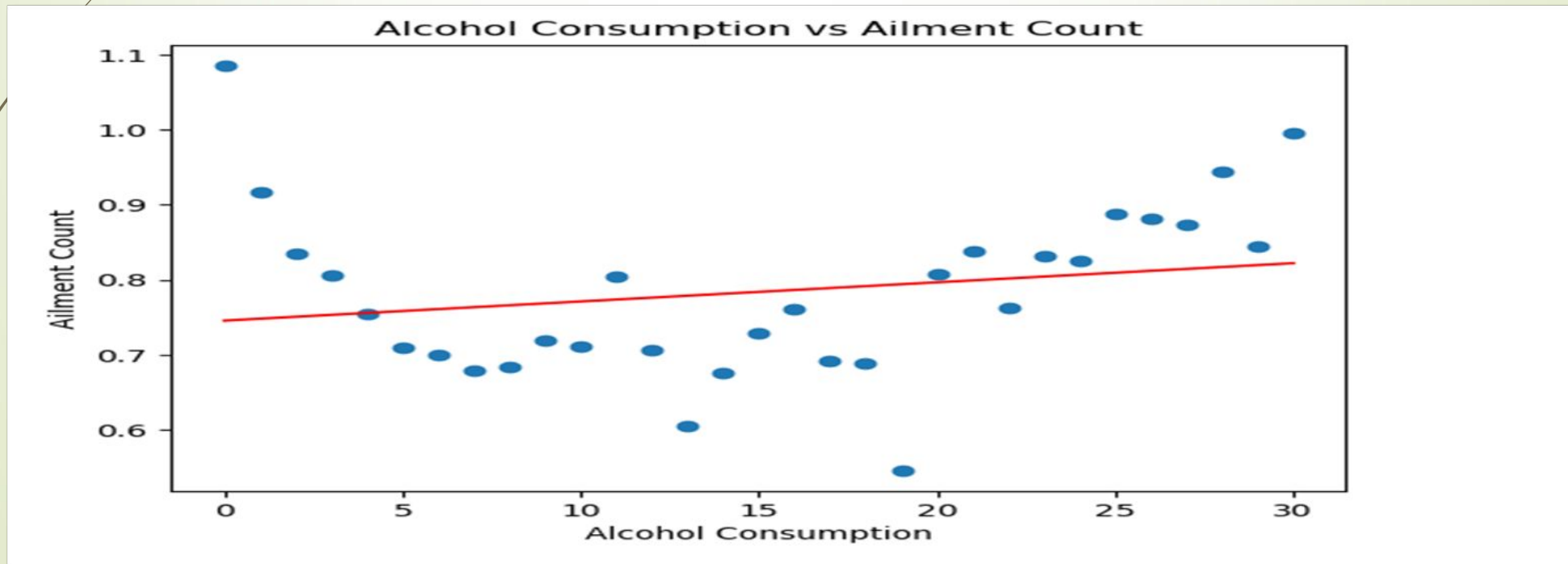
No significant linear relationship exists between Alcohol Consumption and Ailment count (correlation coefficient = 0).

Alternate Hypothesis (H1):

A statistically significant positive linear relationship exists between Alcohol Consumption and 'Ailment Count' (correlation coefficient $\neq 0$).

Conclusion:

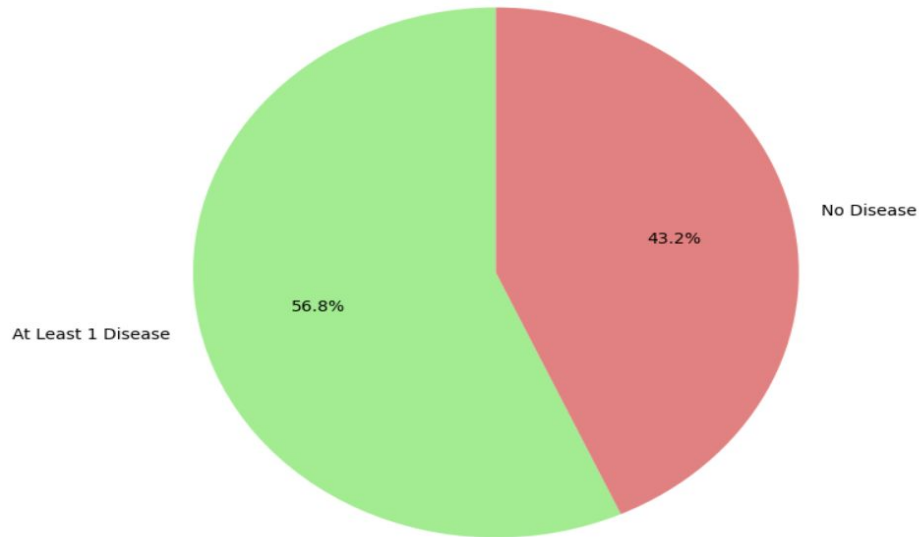
The obtained correlation coefficient **0.20** suggests a **positive but weak linear association** between Alcohol Consumption and Ailment count. Therefore, while a positive trend exists, the impact of Alcohol Consumption on Ailment count per person is relatively weak in this dataset. Further context about the domain would be valuable in making an informed decision.



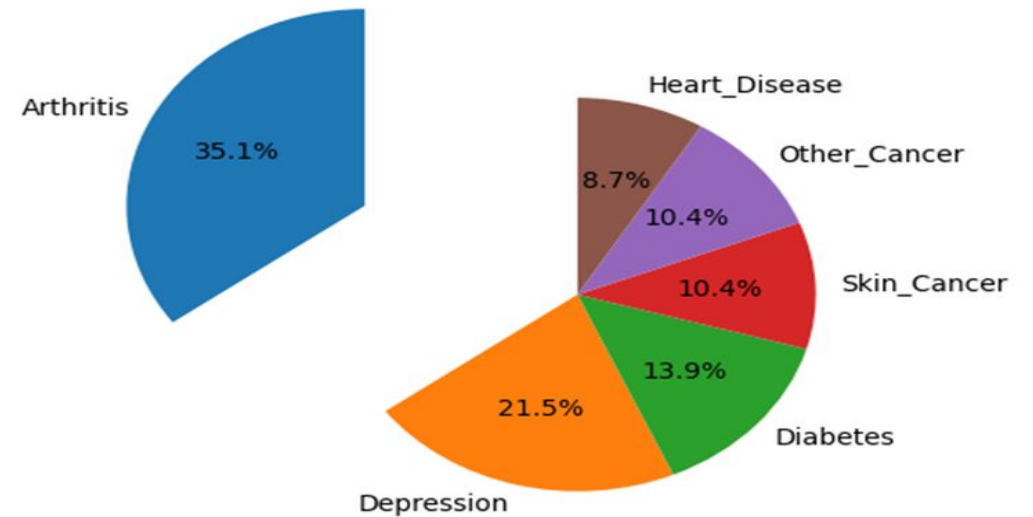
Summary of Health issues

- 57% (175,331 individuals) had reported at least one health issue while 43% (133,523 individuals) reported none in the dataset.
- Arthritis is the most prevalent health condition, followed by depression, diabetes, and heart disease has the lowest prevalence.

Distribution of Health Status



Distribution of Ailments

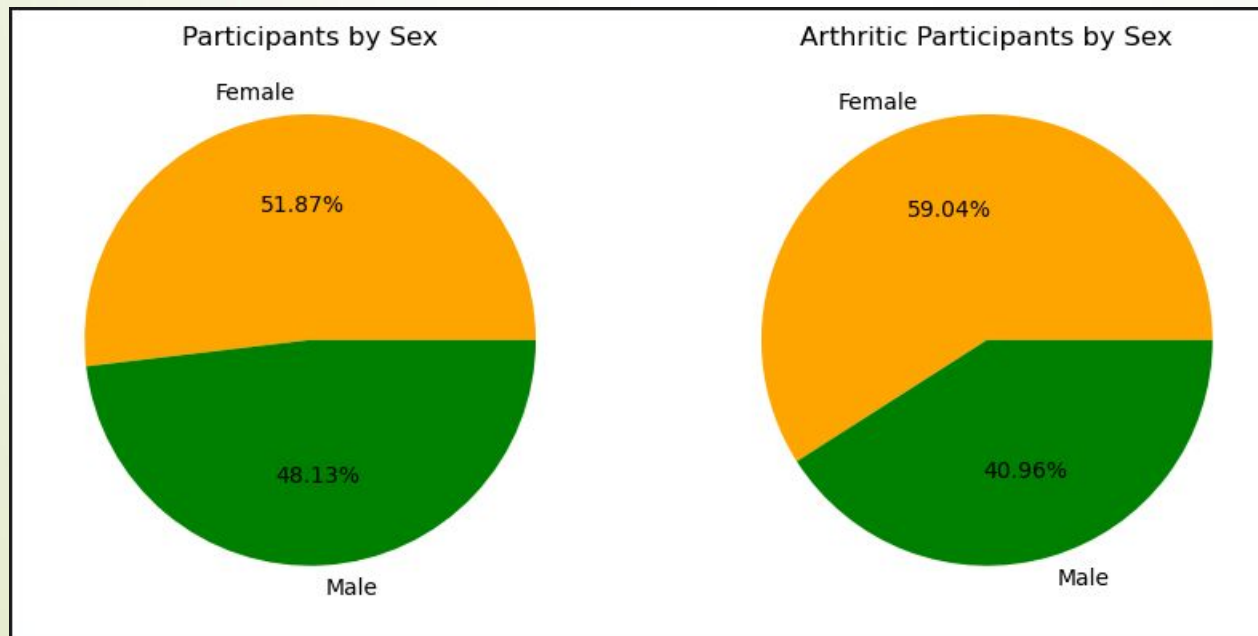




Deep dive into selected Health issues

Arthritis

- Arthritis occurs the most frequently with 32.7% of the participants reporting this health issue. When we looked into the data, we found the incidence of arthritis was higher for female participants than male participants. Given this we wanted to determine if the difference could be explained by randomness or whether the difference was statistically significant



Chi-Square Contingency Table Analysis

Arthritis by Sex

Count of Candidates	No Arthritis	With Arthritis
Female	100528	59668
Male	107255	41403

We chose to put the incidence of Arthritis by Sex through a Chi-Square as follows:

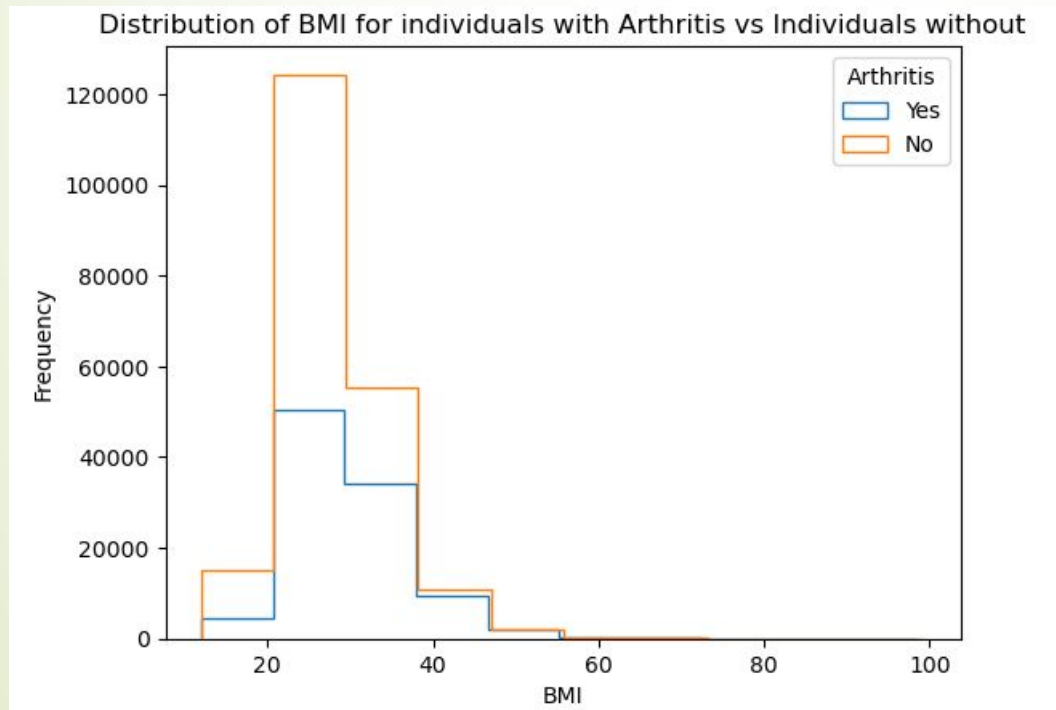
Alternative Hypothesis: The occurrence of Arthritis varies significantly by Sex.

Null Hypothesis: The variance in the occurrence of Arthritis is not significant and can be explained by randomness.

Outcome: Using the distribution of individuals who do not have arthritis by sex and comparing that to the observed distribution of individuals who have Arthritis by sex, the chi-square contingency analysis results in a p-value of 0.0. This suggests there is no statistical relationship between the two groups and the observed variation we see is not happening by chance and we must reject the null hypothesis and assume the changes in the incidence of arthritis by sex is significant.

Can a Candidates BMI Impact the incidence of Arthritis?

- Investigating further we wanted to determine if the BMI score of a Candidate could impact the incidence of Arthritis. We chose a T-Test



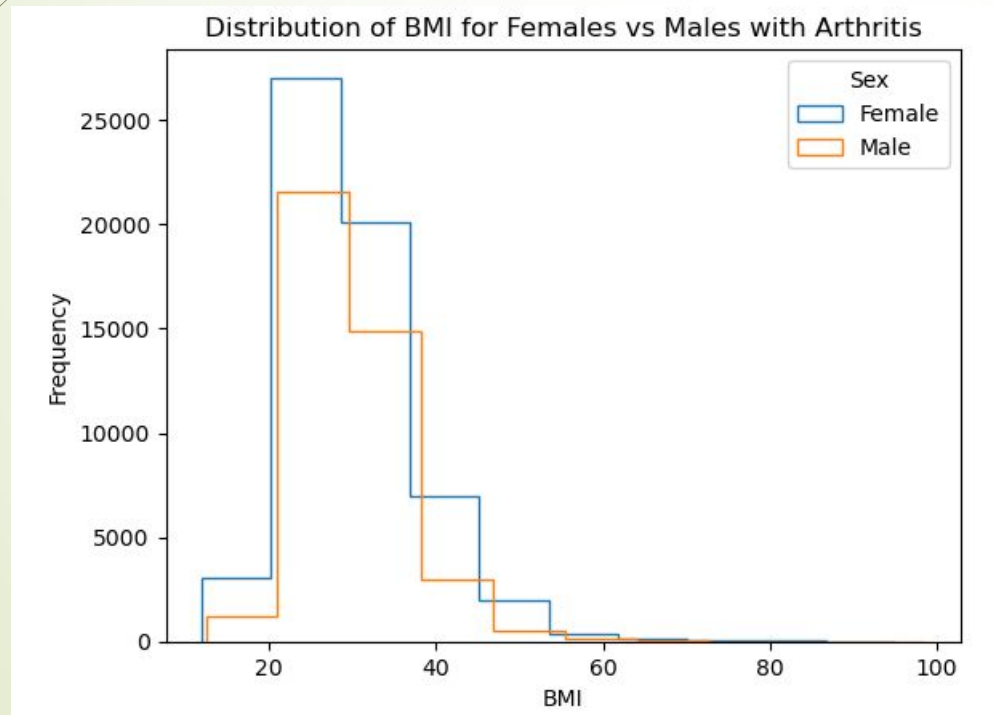



T-Test Arthritis vs BMI for Individuals with Arthritis vs Individuals without Arthritis

- Alternative Hypothesis: The BMI score for candidates with Arthritis is statistically different from the BMI scores for candidates without Arthritis.
- Null Hypothesis: The variance in the BMI scores is not significant and can be explained by randomness.
- Outcome: the p-value of 0.0 indicates the variance in the means of people with Arthritis vs people without Arthritis is statistically significant and not happening by chance and therefore incidence of Arthritis is influenced by a person's BMI

Can the Variance in the Occurrence of Arthritis by Sex be explained by differences in BMI by Sex?

- There appears to be variance in the BMI scores for Females vs Males for the candidates with Arthritis. Can this explain the variance in the incidence of Arthritis between the two groups?



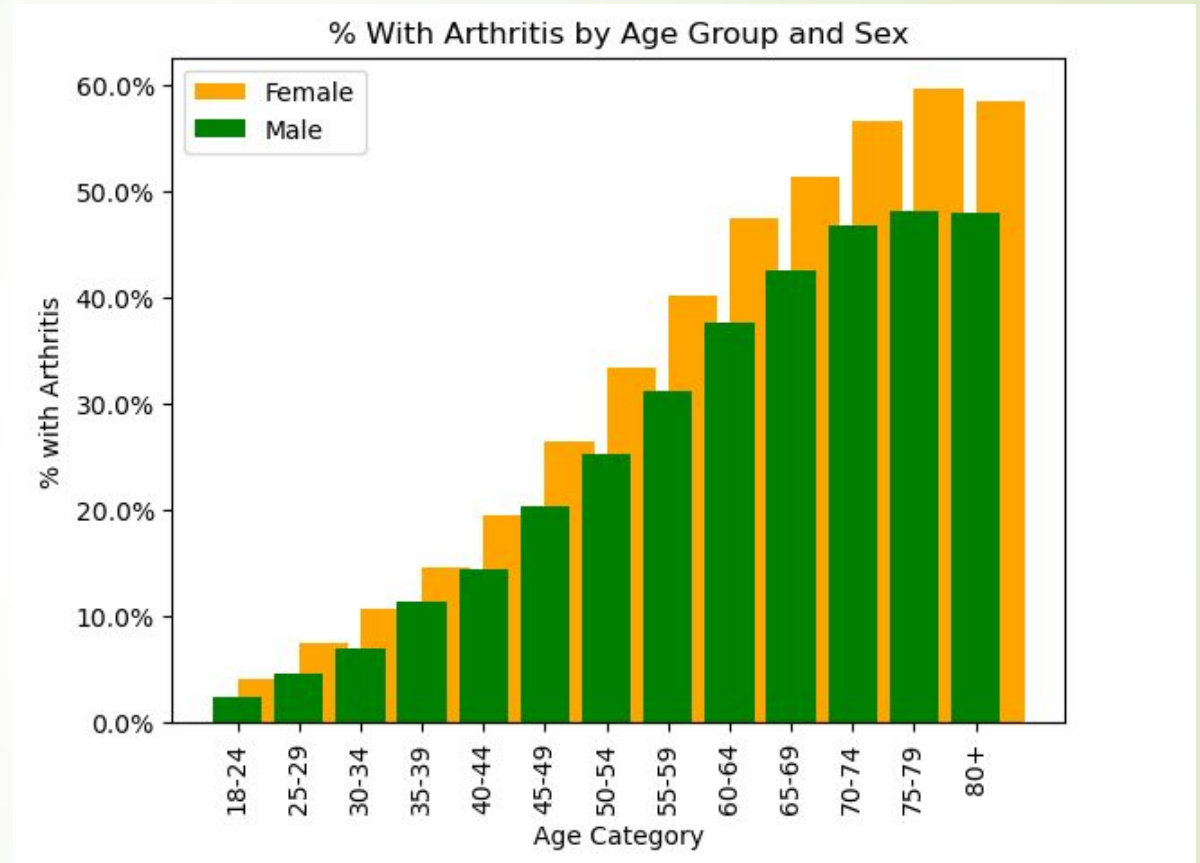


T-Test of BMI for Females with Arthritis vs Males with Arthritis

- Alternative Hypothesis: The incidence of arthritis is related to the difference in BMI score for each sex.
- Null Hypothesis: The incidence of arthritis in individuals is unrelated to the BMI scores for each sex and the difference in the mean BMI between the two groups we are seeing is to be expected by chance.
- Outcome: a T-test yielded a p-value of 0.08 when assessed at a confidence level of 95% indicates the variance in the mean we are seeing may be happening by chance and the incidence of Arthritis appears unrelated to the BMI of each sex. Therefore, the difference in the incidence of Arthritis by sex can NOT be explained by the difference in BMI scores.

The Incidence of Arthritis by Age

- The incidence of Arthritis increases with age, could the difference in the incidence of arthritis be related to the difference in the relative ages of female candidate's vs male?



Chi Square Contingency Table Analysis

- Age

	Female	Male
Age_Category		
18-24	8215	10466
25-29	7118	8376
30-34	8963	9465
35-39	10367	10239
40-44	11203	10392
45-49	11000	9968
50-54	12968	12129
55-59	14660	13394
60-64	16969	15449
65-69	17427	16007
70-74	16739	14364
75-79	11400	9305
80+	13167	9104

We put all candidates into a table by age category and sex in an effort to determine if the variance we see in the distribution of candidates can explain the difference in the incidence of Arthritis.

Alternative Hypothesis: The distribution of candidates by age category and sex contributes to the variance we are seeing in the variance in the incidence of Arthritis we see by sex Alone.

Null Hypothesis: The variance we see in the distribution in candidates by Sex and Age Category is not significant and can be explained by randomness.

Outcome: The Chi-Square Contingency Table yields a p-value of nearly 0.0 indicating the variance in the incidence of Arthritis by sex can be related to the variance in Age by Females vs Males.



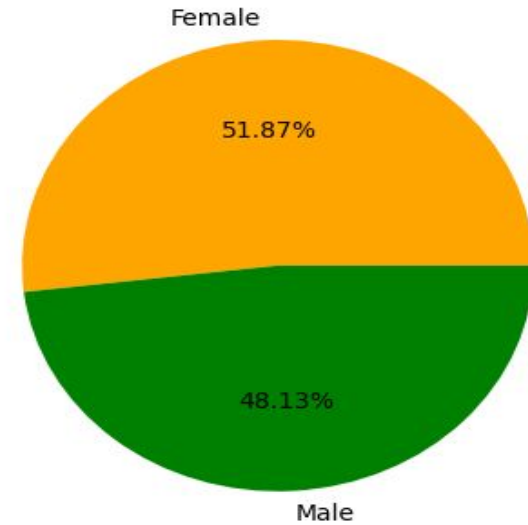
Final Conclusion

- There is a strong relationship detected between the incidence of arthritis and sex, where women appear to have a higher incidence of arthritis when compared to men.
- Analysis shows the incidence of arthritis is related to an individual's BMI score and age but neither of these factors appear to explain the delta in the frequency of arthritis by sex within the studied population.

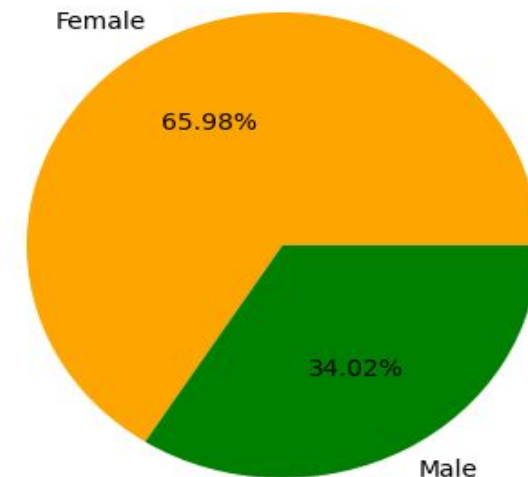
Analysis of depression

- In the data set **depression** is diagnosed in **20%** of the total participants.
- The **sex** of the participants was nearly **even**; (F)**52%** : (M)**48%**
- Our preliminary analysis discovered that depression was **32% more prevalent in the female participants.**

Participants by Sex



Depressed Participants by Sex



Further Analysis

- In order to determine if there was **evidence** of a statistically significant association between depression and sex we conducted a **chi-square contingency table** analysis.

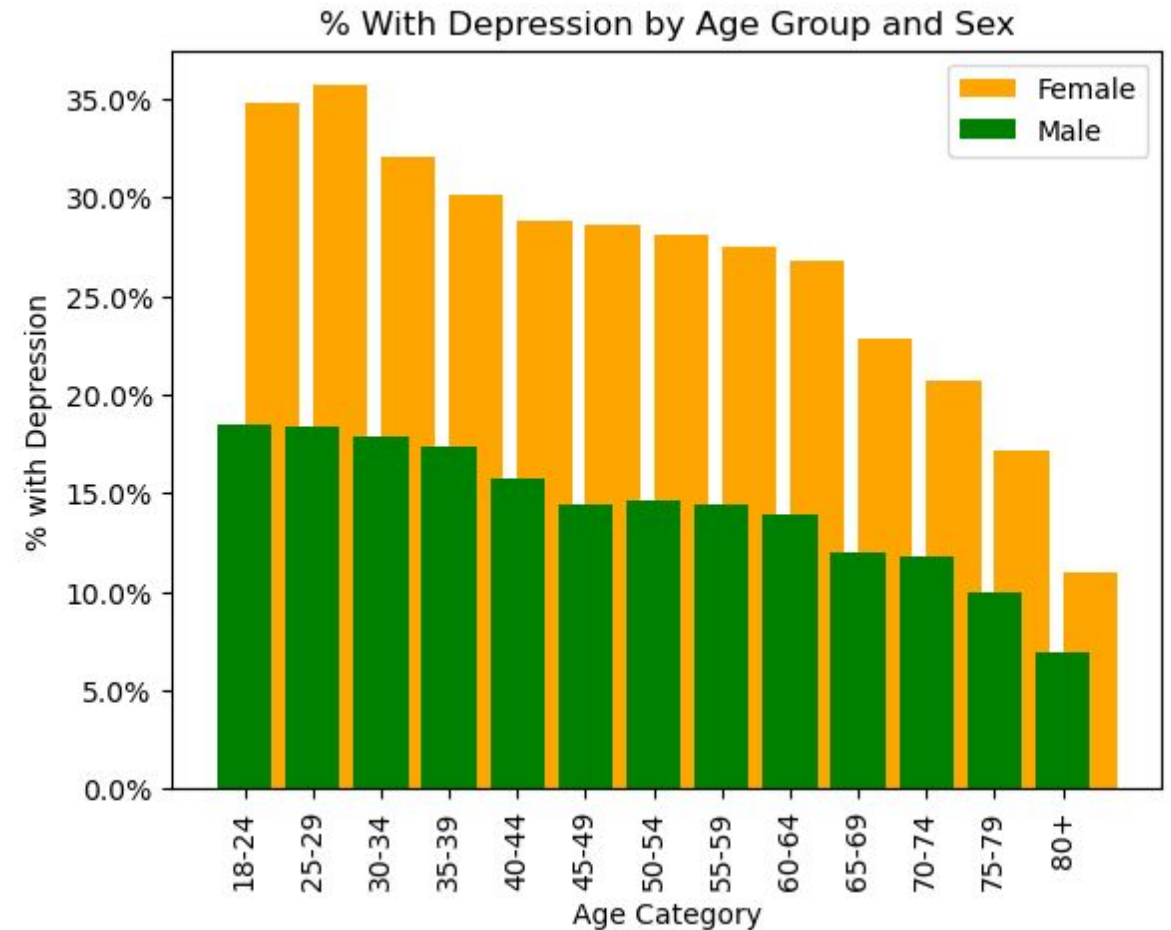
	No Depression	With Depression
Female	119351	40845
Male	127602	21056

- **Alternative Hypothesis (H1)**: The incidence of depression is **related** to the sex of an individual.
- **Null Hypothesis (H0)**: The distribution of individuals with depression is **unrelated** to the sex of the individual, & any observed variance is due to **chance**.
- **Outcome**: The chi-square contingency analysis yields a **p-value of 0.0**, which is **less than 0.05**, and so we must **reject** the null hypothesis in **favour** of the alternative.
- We thus **conclude** that there is a **statistically significant** relationship between the incidence of **depression** and the **sex** of the individual.

Analysis by Age

- In another analysis we also found that the incidence of depression **decreases** with age.
- This indicates that the prevalence of depression in females may be influenced by the variance in the distribution of males and females in different **age groups**.

(If the incidence of depression decreases with age, the higher percentage of older males in the data set may influence the lower count of depression in men)



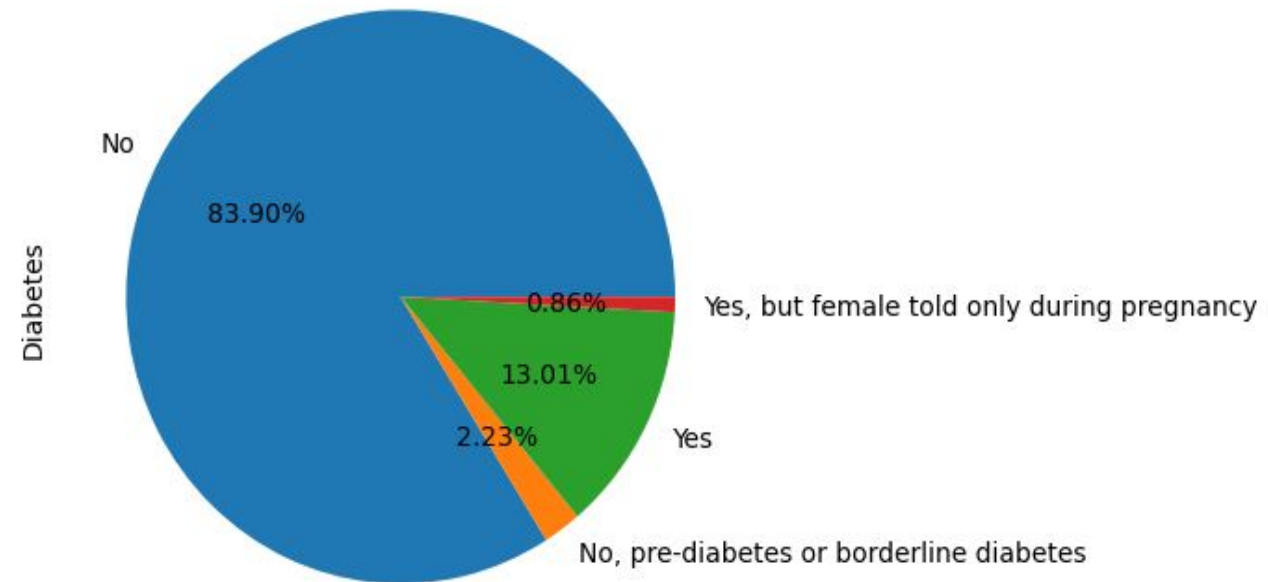
In Conclusion

- 1) We detected a **strong relationship** between the incidence of **depression and sex**; where **women** experience depression at a **higher rate** than **men**.
- 2) We then used a **chi-square contingency table analysis** to determine if the results were **statistically significant** and concluded that the **results** were **significant**.
- 3) We then analyzed **depression** with **age** and found that the incidence of **depression decreases** with **age**
(this may influence the higher numbers of depression in women)

Diabetes

- This visual representation illustrates the overall prevalence of Diabetes and non-Diabetes among a group of individuals.
- The chart categorizes the Diabetes dataset into four groups: "Yes," "No," "Yes, but reported only during pregnancy for females," and "No, pre-diabetes or borderline diabetes."
- According to the chart, the majority of individuals in the dataset are not diagnosed with Diabetes.

Distribution of individuals having Diabetes vs Not vs Prediabetes vs Females during pregnancy





One Way ANOVA for Weight by Age Category for Individuals with Diabetes

Alternative Hypothesis:

There is a statistically significant difference in Weight scores based on Age Category for people with Diabetes.

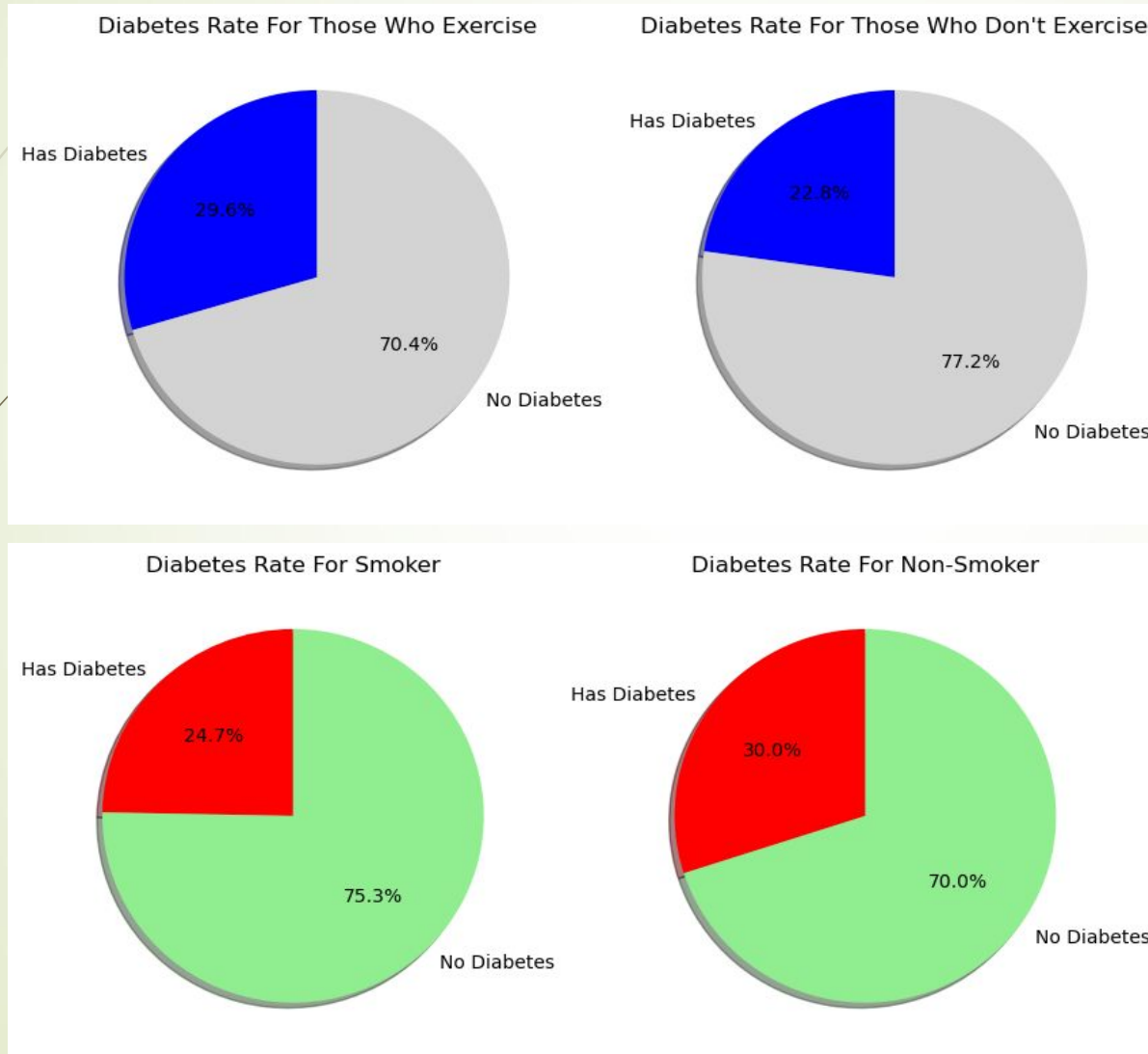
Null Hypothesis:

The mean Weight by age category are statistically the same.

Analysis

The One-Way ANOVA test yields a statistic of 176.3 and a p-value of 0.0, therefore we must reject the null hypothesis in favor of the Alternative hypothesis and conclude the means of Weight are statistically influenced by the age category.

Comparing Behaviour Pattern of individuals from the data set



- The initial visualization compares individuals with diabetes who engage in exercise **29.6%** versus those who do not **22.8%**.
- The second visualization contrasts individuals with diabetes who smoke **24%** with those who do not smoke **30%**.
- When considering both behaviors based on the dataset, the visualizations indicate that neither engaging in exercise nor smoking appears to be associated with the presence of diabetes in individuals.



Conclusion:

Diabetes:

The analysis reveals a minimal correlation between diabetes and individuals who maintain consistent smoking or exercise habits.

The data suggests that there is little connection between an individual's weight and the likelihood of having diabetes, as observed in the dataset.



Heart Disease

What are we looking for?

Correlations between heart disease and the following factors:

1. # of Alcohol Consumption
2. # of Fruit Consumption
3. # of Green Vegetables Consumption
4. # of Fried Potato Consumption
5. Smoking History (Yes/No)
6. Exercise (Yes/No for the past few months)



Heart Disease

Null Hypotheses

1. There is no significant difference in alcohol consumption between those who has heart disease and doesn't.
2. There is no significant difference in fruit consumption does between those who has heart disease and doesn't.
3. There is no significant difference in green vegetables consumption does between those who has heart disease and doesn't.
4. There is no significant difference in fried potato consumption does between those who has heart disease and doesn't.
5. There is no significant difference in heart disease rate between smokers and non-smokers.
6. There is no significant difference in heart disease rate between those who exercise and don't.

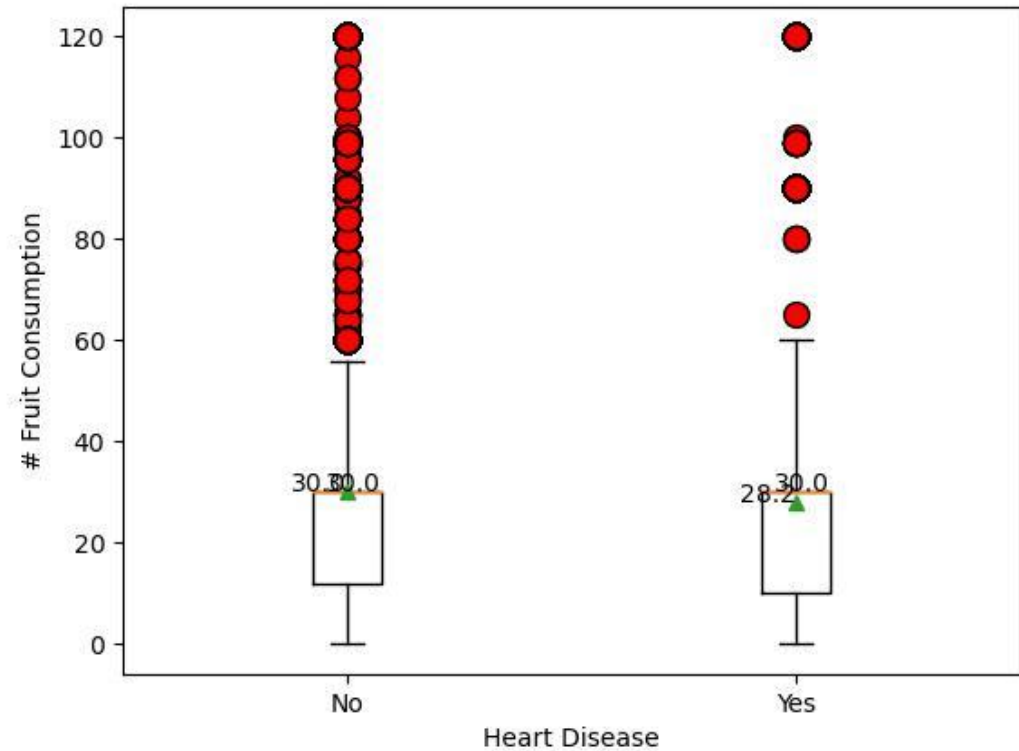
Heart Disease: Alcohol Consumption

	No	Yes
Population	283883	24971
Mean	5.2	4.1
Median	1.0	0
T-test p-value	2.56e-06	



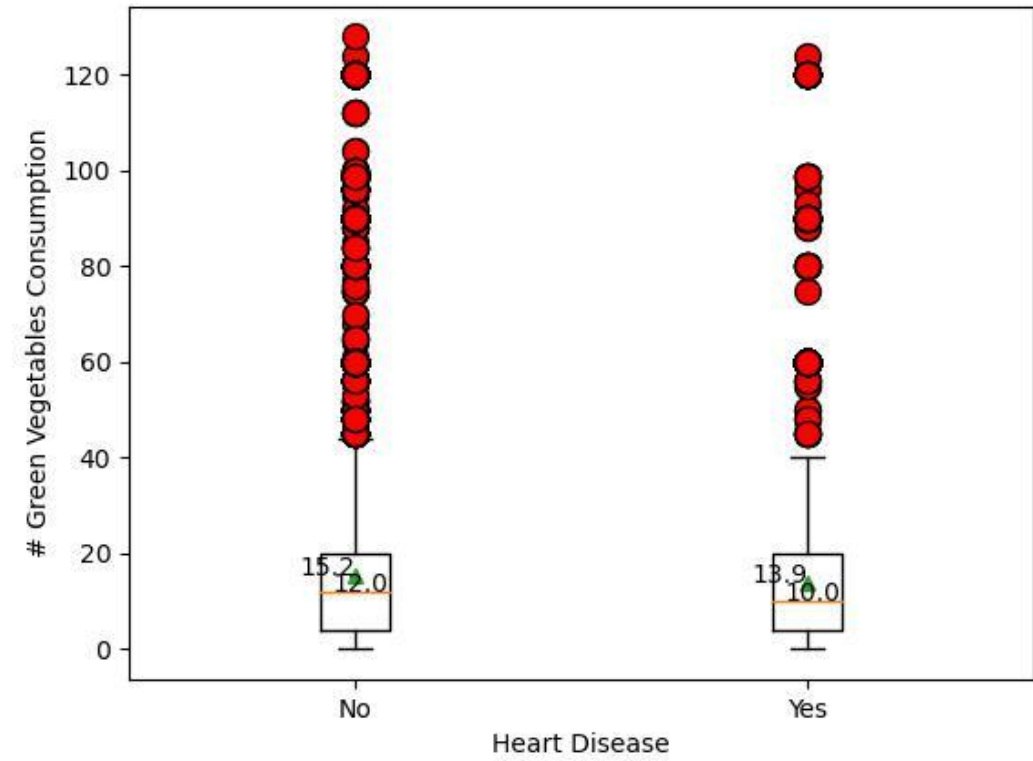
Heart Disease: Fruit Consumption

	No	Yes
Population	283883	24971
Mean	30.0	28.2
Median	30.0	30.0
T-test p-value	1.12e-29	



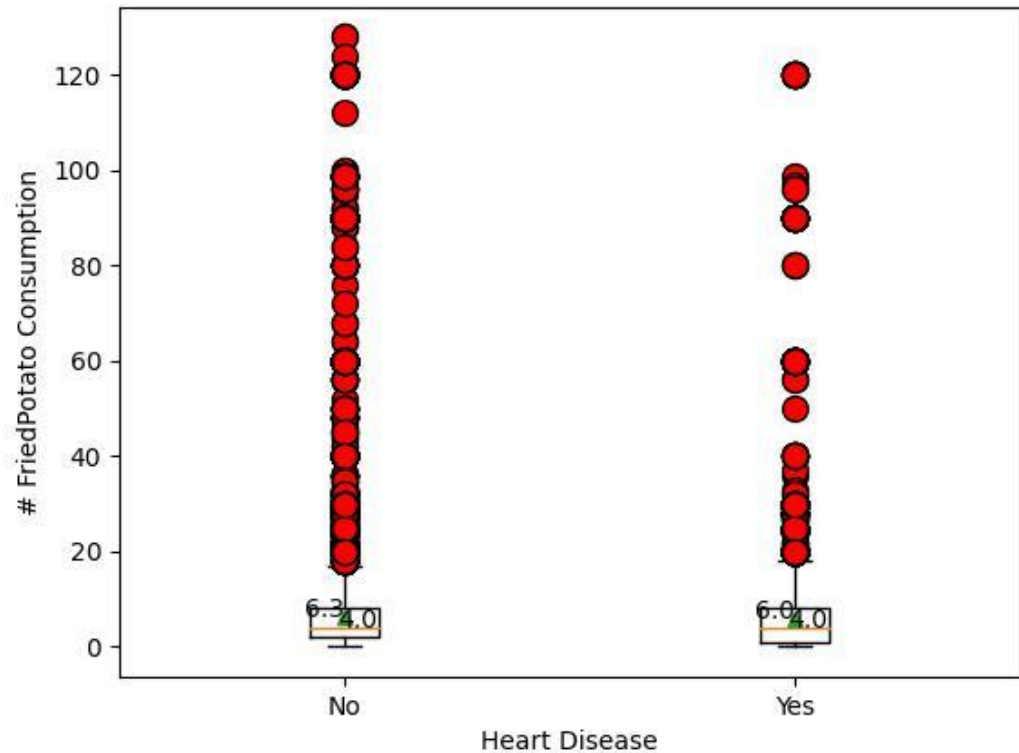
Heart Disease: Green Vegetables Consumption

	No	Yes
Population	283883	24971
Mean	15.2	13.9
Median	12.0	12.0
T-test p-value	1.35e-45	



Heart Disease: Fried Potato Consumption

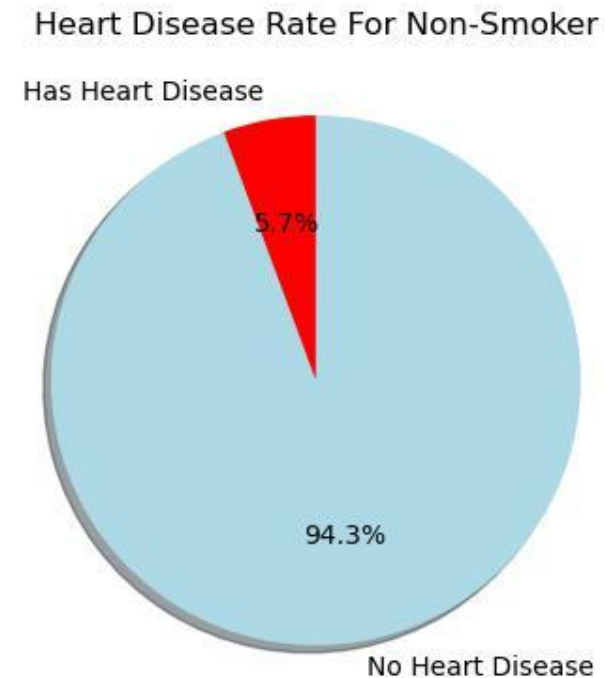
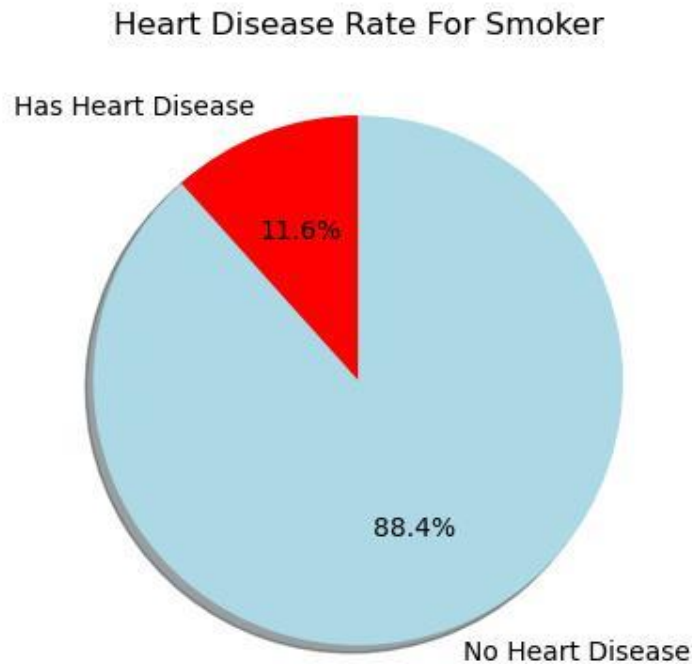
	No	Yes
Population	283883	24971
Mean	6.3	6.0
Median	4.0	4.0
T-test p-value	2.36e-07	



Heart Disease: Smoking History

Contingency Chi-Square	Has Heart Disease	No Heart Disease
Smoker	14584	110680
Non Smoker	10387	173203

A p-value of 0 indicates **rejection** of null hypothesis.

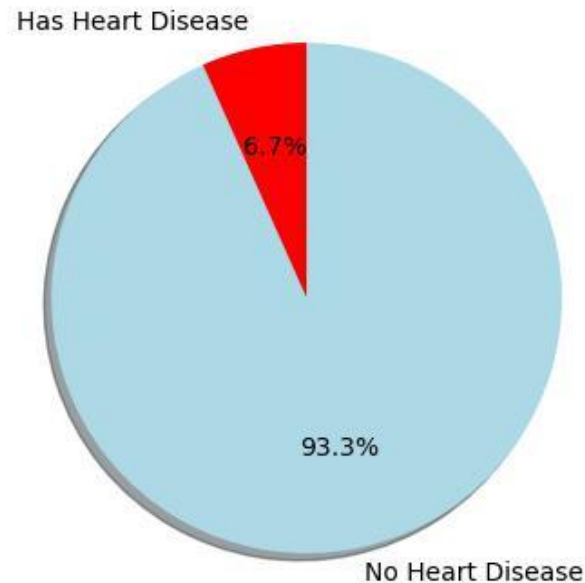


Heart Disease: Exercise

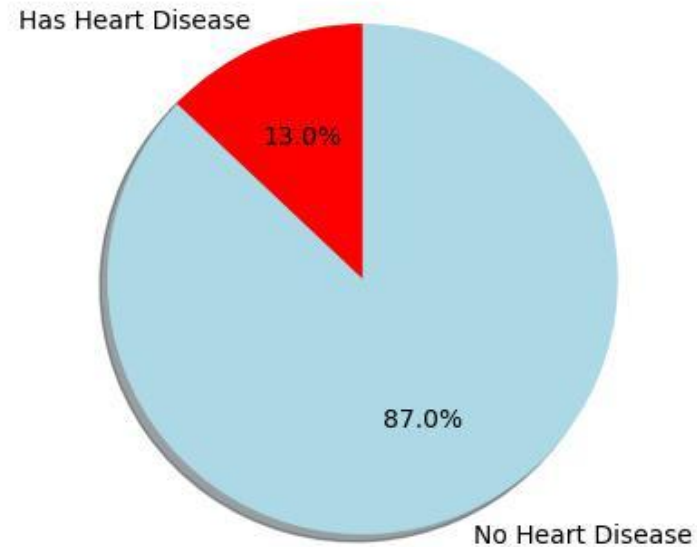
Contingency Chi-Square	Has Heart Disease	No Heart Disease
Do Exercise	15967	223414
Don't Exercise	9004	60469

A p-value of 0 indicates **rejection** of null hypothesis.

Heart Disease Rate For Those Who Exercise



Heart Disease Rate For Those Who Don't Exercise





Heart Disease

Results

1. There IS a significant difference in alcohol consumption between those who has heart disease and doesn't.
2. There IS a significant difference in fruit consumption does between those who has heart disease and doesn't.
3. There IS a significant difference in green vegetables consumption does between those who has heart disease and doesn't.
4. There IS a significant difference in fried potato consumption does between those who has heart disease and doesn't.
5. There IS a significant difference heart disease rate between smokers and non-smokers.
6. There IS a significant difference heart disease rate between those who exercise and don't.



Heart Disease

Conclusion

1. The alcohol consumption is higher on people [without heart disease](#).
2. The fruit consumption is higher on people [without heart disease](#).
3. The green vegetables consumption is higher on people [without heart disease](#).
4. The fried potato consumption is higher on people [without heart disease](#).
5. [Smokers](#) has a higher rate of heart disease.
6. People who [don't exercise](#) has a higher rate of heart disease.