Introduction

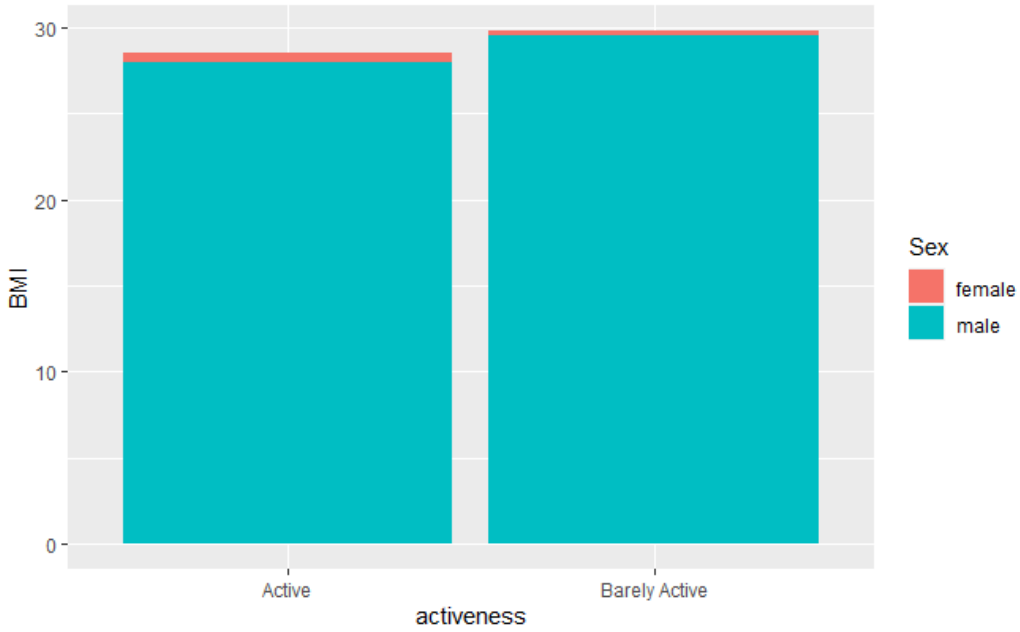
Author: Y3926947

I installed R and Rstudio, loaded the necessary library needed for my analysis including quarto. The Tidyverse library is a perfect choice having packages like *ggplot2*, *readr* and others. I loaded my data which was in a CSV file on my desktop and assigned it to a variable ‘data’. I then began my analysis

Exploratory analysis is crucial in any analysis as it helps to make known the kind of data being worked on. I ran Exploratory Analysis observing the total number of surveys taken (1,109). I wondered if physical activeness had an effect on BMI. For a more accurate analysis, I replaced the NA values with the mean of the 'Physical Activeness’ and ‘BMI’ columns. I observed the total number of surveys taken by each Gender (Male: 530, Female: 589).

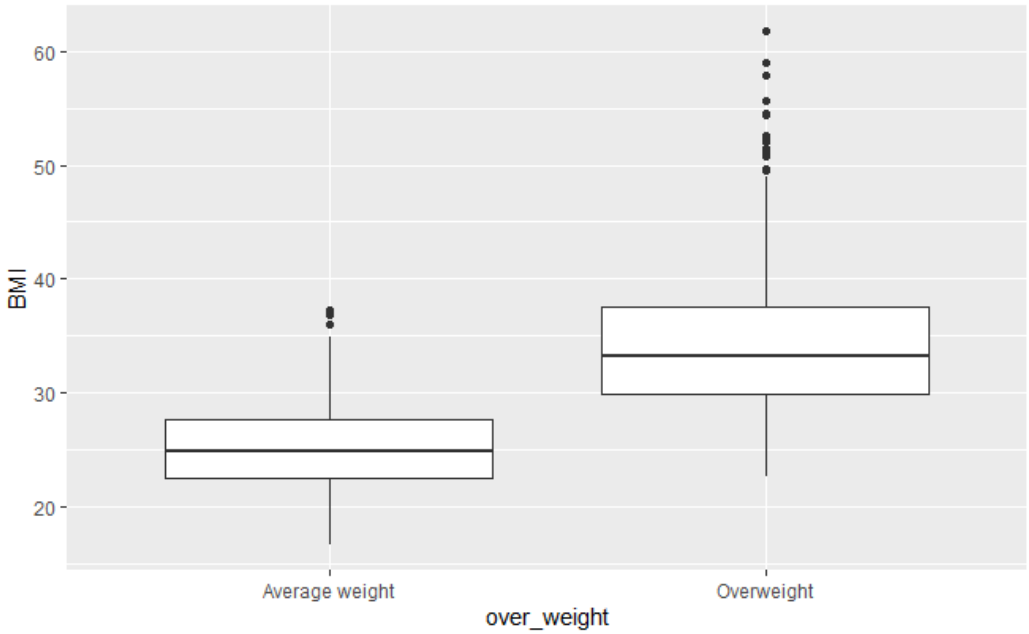
Exploratory Analysis 1

I created a new column conditioned less or greater than the mean of the ‘Physical Active Days’ using the ‘Physical Active Days’ column to create categories ‘Barely Active’ and ‘Active’ under the new column ‘Activeness’. I plotted the ‘Physical Activeness’ values against ‘BMI’ filtering with ‘Sex’. See Figure 1. below.



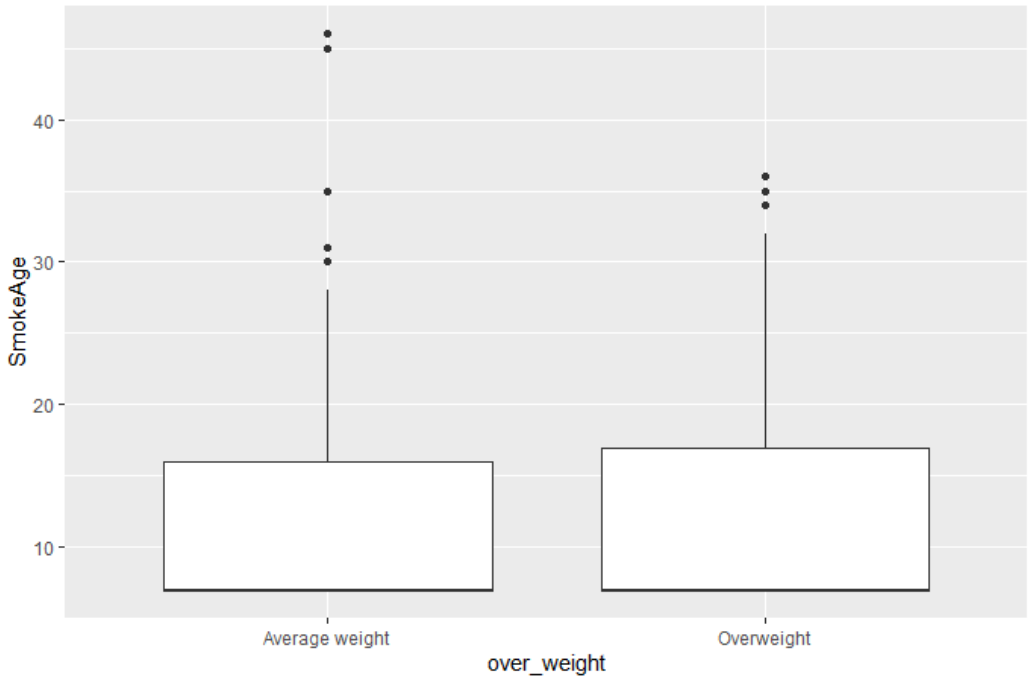
Exploratory Analysis 2

I created a new column conditioned less or greater than the mean value of ‘Weight’ categorizing the new column ‘Over_weight’ into ‘Overweight’ and ‘Average Weight’.
The Exploratory analysis was to observe the relationship between ‘Weight’ and ‘BMI’. I plotted ‘Over-weight’ against ‘BMI’



Exploratory Analysis 3

My third exploratory analysis was to observe the weight of smokers. Most especially if smoking at an early age increases the tendency to be an averagely weighed person. I replaced the NA value in the ‘SmokeAge’ column with the mean value of the column. Then visualized ‘Over_weight’ against ‘SmokeAge’. See Figure 3. Below.



Methods

Hypothesis Test 1

Two-Way Anova: testing for the group column ‘Sex’[Male & Female]’ against ‘Pulse pressure’. This is an hypothesis stating that as Age increases so does the risk of having a heart attack from increased pulse pressure.

Hypothesis Test 2

Chi-test: testing for the column ‘SmokeNow’ and ‘Diabetes’. This is an hypothesis stating that smokers are liable to have Diabetes.

Hypothesis Test 3

Linear regression (two-sample test): testing the column ‘Age’ and ‘Weight’ values but filtered showing data only for females. This is an hypothesis that women older than 45 tend to be overweight than younger women.

Results

Hypothesis Test 1: Two-Way Anova test.
See table 1.

	A	B	C	D	E	F	G
		term	df	sumsq	meansq	statistic	p.value
1		Sex	1	4615.941956	4615.941956	31.72908973	2.24E-08
2		Residuals	1117	162500.9481	145.4798103	NA	NA

Hypothesis Test 2: Chi-test

	A	B	C	D	E	F
		statistic	p.value	parameter	method	
1		3.028071338	0.08183523914	1	Pearson's Chi-squared test	

Hypothesis Test 3: Two-sample test

	term	estimate	std.error	statistic	p.value
1	(Intercept)	58.70708901	2.119572848	27.69760383	3.43E-83
2	Weight	0.000833561516	0.02634321648	0.03164235914	0.9747791797

Discussion

Exploratory Analysis 1:
Summary:

- From Figure 1. A histogram is showing the relationship between ‘Physical activeness’ and ‘BMI’ filter with ‘Sex’.

  Barely active people tend to have higher BMI.
  The amount of days Men are barely active is higher than the number of days.
  The percentage of Women who are barely active are less than the percentage of active women

Exploratory Analysis 2:
Summary

- From Figure 2. A box plot is showing the relationship between ‘Weight and BMI’

- People that are overweight i.e have a weight greater than 83, have a higher tendency to have high BMI.
- It is noticed that the maximum BMI for averagely weighed people is around 38, with no outliers.

Exploratory Analysis 3:
Summary

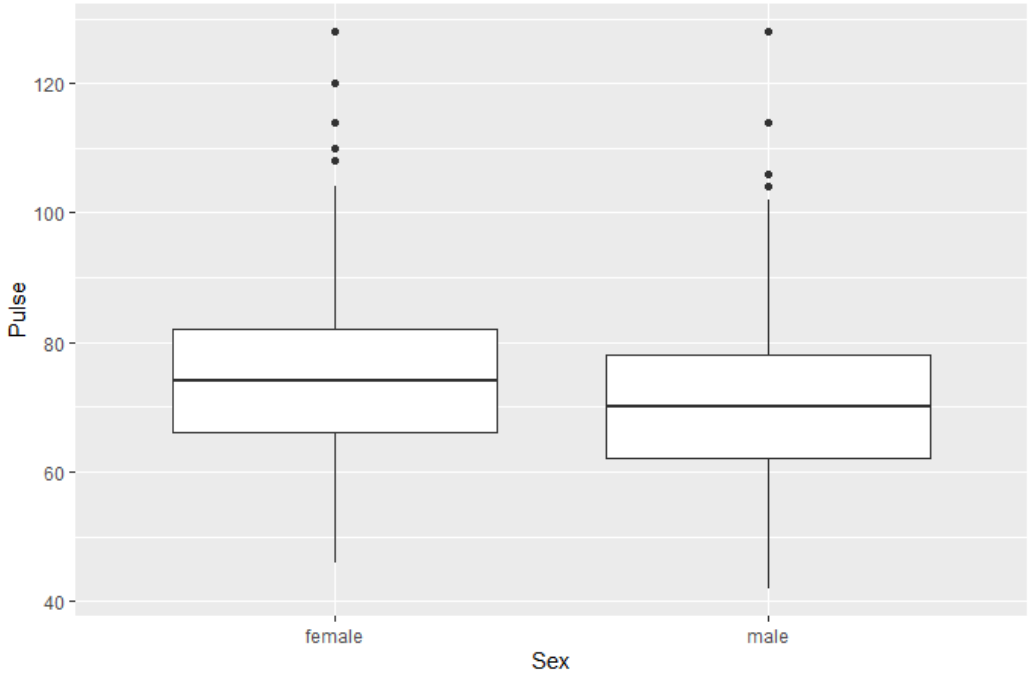
- From Figure 3. A box plot is showing the relationship between ‘Weight and SmokeAge’

- Majority of people that are averagely weighed started smoking at a younger age. People that maintain an average weight are clearly smokers with the outliers above 40 years.

Hypothesis Test 1:
Summary

- From Table 1: Two-Way Anova test testing the relationship between ‘Sex’ and ‘Pulse pressure’
- The p-value has a value of 2.24 meaning that we can accept the hypothesis that the older people get the higher tendency to have a heart attack due to increased pulse pressure.

Visualize the model. See Figure 4. Below

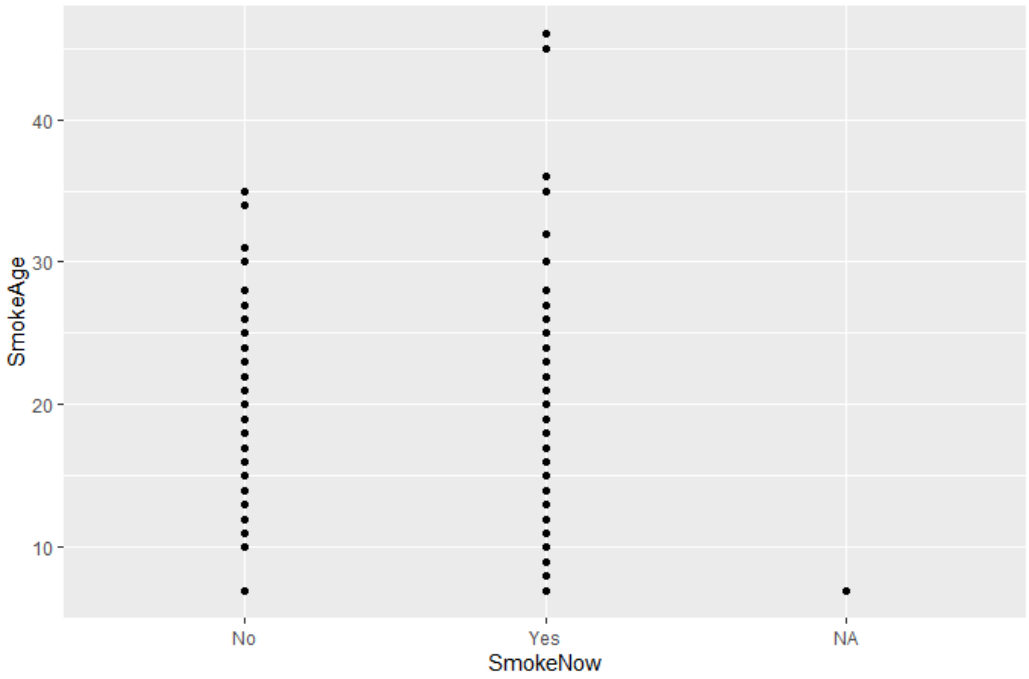


Conclusion: We can see that females have a higher tendency to have a heart attack over men.

Hypothesis Test 2:
Summary

- From Table 2: Chi-test, testing the relationship between people that smoke and chance of being diabetic.
- The p-value has a value of 0.08 which means we can accept the hypothesis that people that currently smoke have a high tendency to be diabetic.

Visualize the model. See Figure 5.



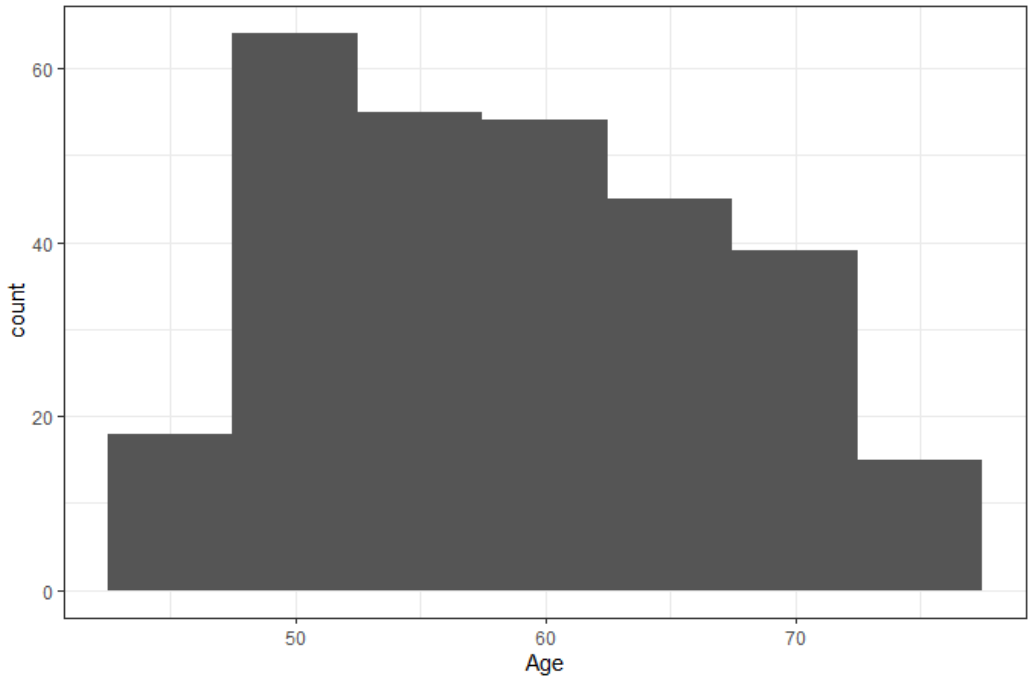
We can see regardless of the assumption, elderly people of age 40+ still smoke

Hypothesis Test 3:

Summary

- From Table 3: Two-sample test testing the relationship between ‘Age’ and ‘Weight’ in women over the age of 45.
- The p-value has a value of 3.43 meaning we can accept the hypothesis that the older women get, the more carefree they are about their weight.

Visualize. See Figure 6. Below.



Conclusion: We can see women in the age group 50 - 70 are mostly over weighed.

Words count: 781 words
Slides: 9