PUBLIC HEALTH:NATIONAL NUTRITIONAL HEALTH PROJECT REPORT

INTRODUCTION

NHANES is a program run by the CDC to assess the health and nutritional status of adults and children in the US. It combines survey questions and physical examinations, including medical and physiological measurements and laboratory tests, and examines a representative sample of about 5,000 people each year. The data is used to determine the prevalence of diseases and risk factors, establish national standards, and support epidemiology studies and health sciences research. This information helps to develop public health policy, design health programs and services, and expand the nation's health knowledge.

DESCRIPTION OF DATASET

The dataset is a data-frame containing 5000 observations of 32 variables amounting to a total of 160,000 observations. Amongst the 160,000 observations, 33,931 were missing, signified by 'N/A'. The variables in the observations included but were not limited to, gender, age, race, weight etc. The dataset was sourced from: Hackbio.
Dataset here

METHODS

1. Importing and Preprocessing the Dataset

To import the dataset from the source link into the R environment, the read.csv() function was used as the dataset was a .csv file. Before analysis began, all the missing data identified in the dataset were converted to '0' and assigned to the variable name 'cleaned_nhanes'.

2. Creating Histograms to Visualise Select Distributions

2.1 Setting The Dimensions for The Plot

The task required the creation of four histograms to visualise the distribution of Age, BMI and Weight across the dataset. The task also required that the distribution of weight be shown in both kilograms (the default) and in pounds (lbs).
To start this visualisation, a new column was created in the data-frame to store the values of weight in pounds. This was done by multiplying the initial values of Weight in kilograms by 2.2 and storing in the result under the variable name 'WeightInPounds'.
To begin visualisation. The par(mfrow) function was used and the parameters were set to ensure all the plots could be properly viewed in the plotting window.

```
#using par(mfrow) to create a grid layout for the four plots
par (mfrow=c(2,2), mar= c(4.2,4.2,2.5,1.0), mgp= c(2.8,0.7,0))
```

- mfrow= c(2,2) divides the plotting window into a 2 by 2 matrix (2 rows and 2 columns), to allow four plots be displayed in one window.
- mar=c(4.2,4.2,2.5,1.0) set the margins of individual plots in the order bottom, left, top, and right margins (in lines of text). So bottom and left margins are 4.2 lines, top margin is 2.5 lines, and right margin 1 line.
- mgp=c(2.8,0.7,0) controls the margin line for axis title, labels, and line. The first value (2.8) is the distance of the axis title from the plot edge, the second (0.7) is for the axis labels, and the third (0) is for the axis line.

2.2 Creating the Histograms

The function used for creating the histograms for the plots was the hist() function. To use the function, it is necessary to understand the basic syntax of hist(). To take the code for Histogram #1 as an example,

```
#Histogram 1: BMI
hist(cleaned_nhanes$BMI,
     breaks= "Sturges",
     main="BMI Distribution",
     xlab="BMI",
     ylab='Frequency',
     col="steelblue",
     border= “black")
```

- hist() function plots the histogram, which displays the frequency distribution of BMI values in ranges (bins).
- breaks = "Sturges" uses the Sturges method to determine the number of bins. This is a default method that generally provides a good balance by computing the number of bins based on the log of the sample size, avoiding too many or too few bins.
- main = "BMI Distribution" sets the title of the histogram plot.
- xlab = "BMI" labels the x-axis with BMI.
- ylab = “Frequency" labels the y-axis with the frequency count of observations in each bin.
- col = “steelblue" colors the bars of the histogram in steel blue for better visual appeal.
- border = “black" sets the color of the bar borders to black to define the bar edges clearly.

This was applied and adjusted as needed for each of the histograms.

3. Mean, Range, Variance and Standard Deviation of Select Variables

3.1 Mean

The task required finding the mean pulse for all the participants in the dataset.

To determine this, the function mean() was applied to the Pulse column under the cleaned_nhanes data frame.

The code used to determine this was:

```
mean_pulse <- mean(cleaned_nhanes$Pulse)
```

3.2 Range

The task required to determine the range of the diastolic blood pressure among the participants in the dataset. This was done by implementing the range () function on the diastolic blood pressure column denoted by ‘BPDia’. The code used to determine the range was:

```
range(cleaned_nhanes$BPDia)
```

3.3 Variance and Standard Deviation

The task required determining the variance and standard deviation that existed in the income made by the participants. This was done by applying the var () function and sd() function respectively on the column containing the values of Income from the cleaned_nhanes dataframe. The code used to determine this was:

```
income_variance <- var(cleaned_nhanes$Income)
income_sd <- sd(cleaned_nhanes$Income)
```

4. Scatterplot for Data Visualisation

4.1 Installing ggplot2 Into the R Environment

The task required creating a scatterplot to visualise the relationship between weight and height while also coloring the points according to: gender, diabetes & smoking status. To avoid extensive code that would have been needed with the basic R plot() function, ggplot was installed into the R enviroment on the console using,

install.packages("ggplot2").

After which, library(ggplot2) was run to begin developing the scatterplot.

4.2 Scatterplot Using ggplot

After running library (ggplot2) in the console, the plotting parameters were put in place.

```
ggplot(cleaned_nhanes,
       aes(x=Height,
           y=Weight,
           colour=Gender,
           shape=Diabetes))
+ geom_point(size=3,alpha=0.7)
+ facet_wrap(~SmokingStatus)
+ labs(title='Weight vs Height Colored by Diabetes Status',
       x='Height', y='Weight')
+ theme_minimal()
```

- ggplot(cleaned_nhanes, aes(x=Height, y=Weight, colour=Gender, shape=Diabetes)) initialises a plot by referencing the data frame cleaned_nhanes and sets up a mapping:
- x-axis: Height
- y-axis: Weight
- colour: Gender
- shape: Diabetes

The aesthetic mappings from the initial ggplot call are applied here, with the colour and shape of the plot points determined by Gender and Diabetes, respectively [geom_point details].

- facet_wrap (Smoking Status): serves to create multiple subplots (facets) of the same scatterplot, one for each unique variable within the Smoking Status variable.
- labs(title='Weight vs Height Colored by Diabetes Status', x='Height', y='Weight') sets labels to improve readability and gives a clear plot description and axis names:
- theme_minimal() applies a clean, minimal theme that reduces background clutter and emphasizes the data. This affects gridlines, background, axis lines, and text styling to achieve a modern look without the default theme's heavy gridlines.

5. Conducting T-test for Select Variables

5.1 Age and Gender

To conduct t-test for the values in the Age and those in Gender category. This is to determine if the average age is different between the two groups namely 'Females' and 'Males', the function t.test() was applied in the format,

```
t.test(Age ~ Gender, data = cleaned_nhanes)
```

5.2 BMI and Diabetes

To conduct t-test for the values in the BMI and Diabetes in Gender category. This is to determine if the average mean of the participant's BMI was different between the two categories of individuals, those with diabetes and those without diabetes. The function t.test() was applied in the format,

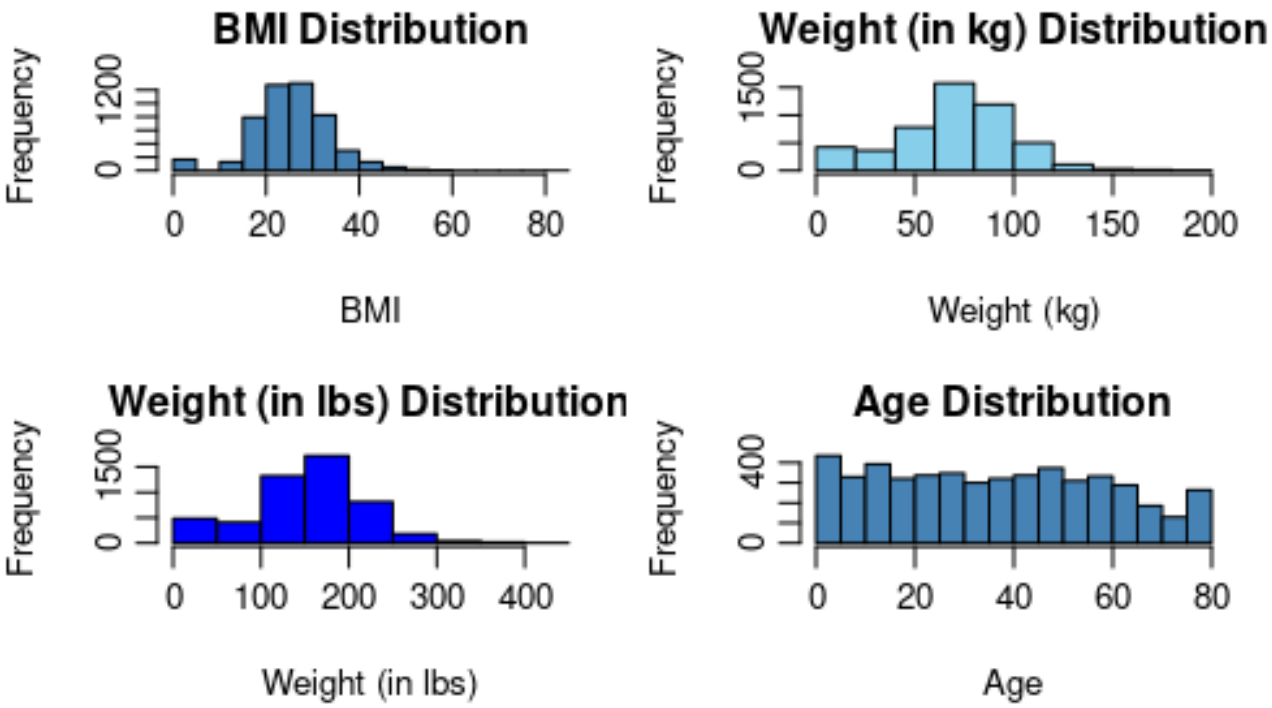
```
t.test(BMI ~ Diabetes, data = cleaned_nhanes)
```

5.3 Alcohol Year and Relationship Status

To conduct t-test for the values in the Alcohol Year and those in Relationship Status category. This is to determine if the average mean of alcohol use in the last year was different between the two groups namely 'Committed' and 'Single'. The function t.test() was applied in the format,

```
t.test(AlcoholYear ~ RelationshipYear, data = cleaned_nhanes)
```

RESULTS
Histogram Grid Results



(Pictured above: the grid of the histograms visualising the frequency of the values of BMI, Age and Weight in kilograms and pounds across the individuals in the data set)

Mean, Range, Variance and Standard Deviation Results
The result for the mean pulse across the individuals in the dataset was 63.06.

```
> mean_pulse
[1] 63.06
>
```

The range of values for diastolic blood pressure in all participants was 0 - 116.

```
> range(cleaned_nhanes$BPDia)
[1]   0 116
```

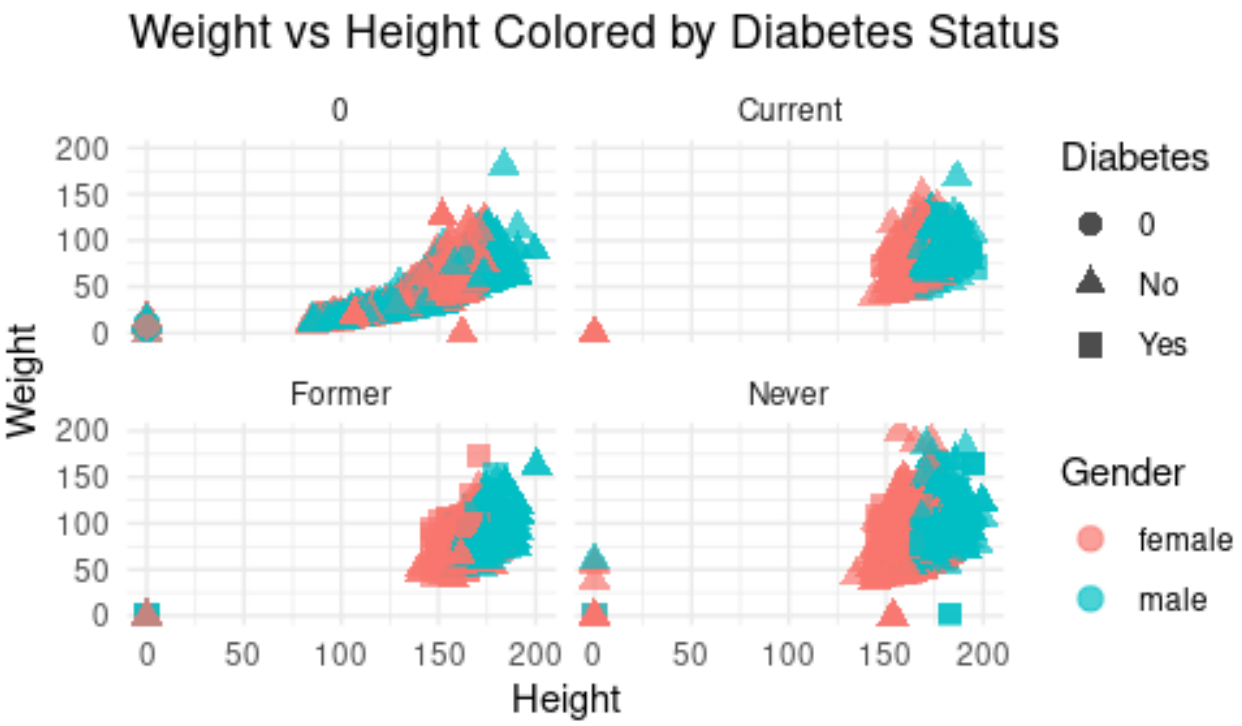
The variance for income across the participants was 1,264,147,754.

```
> income_variance
[1] 1264147754
```

The standard deviation for income across the participants was 35,554.86

```
> income_sd
[1] 35554.86
```

Scatterplot for Data Visualisation



(Pictured above: scatterplot visualising the relationship between Weight and Height coloured by Gender, shaped by Diabetes Status and across different Smoking Status groups: Current, Former and Never. As well as those whose smoking statuses are unknown)

T-test Results

The t-test results for the relationship between Age and Gender

```
> t.test(Age ~ Gender, data = cleaned_nhanes)

        Welch Two Sample t-test

data:  Age by Gender
t = 1.7498, df = 4992.5, p-value = 0.08022
alternative hypothesis: true difference in means between group female and group
male is not equal to 0
95 percent confidence interval:
 -0.1344235  2.3672964
sample estimates:
mean in group female   mean in group male
            37.26733             36.15090
```

The t-test results for the relationship between BMI and Diabetes

The t-test results for the relationship between Alcohol Year and Relationship Status

DISCUSSION

The analysis of the NHANES dataset involving 5000 participants provides valuable insights into the health and nutritional status of the representative population. The distribution of key variables—BMI, Age, and Weight—was effectively visualised using histograms, revealing population and

```
> t.test (AlcoholYear ~ RelationshipStatus, data= cleaned_nhanes)

        Welch Two Sample t-test

data:  AlcoholYear by RelationshipStatus
t = 16.379, df = 3481.8, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Committed and gr
oup Single is not equal to 0
95 percent confidence interval:
 36.70108 46.68283
sample estimates:
mean in group Committed    mean in group Single
               68.10548                26.41352
```

variability trends. The BMI distribution, hints at a prevalence of overweight or obese participants, and could be a signal for public health interventions.

Descriptive statistics showed the mean pulse rate of participants as 63.06 bpm, which is within the normal resting heart rate range, indicating generally good cardiovascular health in the sample. The diastolic blood pressure ranged broadly from 0 to 116 mmHg, suggesting the inclusion of individuals with varied blood pressure statuses, potentially including hypertensive cases. The financial income variation, with a very high variance and standard deviation, highlights substantial economic diversity, which may correlate with health outcomes and access to care.

The scatterplot linking weight and height, differentiated by gender, diabetes status, and smoking status, visually depicted complex interactions between these health determinants. This plot illustrates the relationship between height and weight, coloured by gender and shaped by diabetes status, across different smoking status groups (Current, Former, Never, Unknown (denoted by '0').

Height and Weight Distribution:

Both height and weight values show distinct clustering by smoking status. Current smokers have a wide spread in height and weight, but generally height ranges around 150-190 cm. Former and Never smokers show tighter clusters with higher concentration around common adult heights and weights.

Gender Differences:

Male data points (blue) tend to cluster at higher weight ranges compared to females (red), consistent across all smoking statuses. Females generally appear at slightly lower weights for similar heights.

Diabetes Status:

The shapes represent diabetes status: circles for no diabetes, and squares for diabetes. Diabetes cases (square shapes) predominate at higher weights in both genders, suggesting a positive association between higher weight and diabetes. Diabetic individuals are present across all smoking statuses but appear more frequently in former and never smokers clusters.

Smoking Status Facets:

The distribution of diabetic versus non-diabetic differs by smoking status: Current smokers include both diabetic and non-diabetic with relatively less spread in weight. Former smokers show more diabetic points at higher weights. Never smokers show a tight grouping of both diabetic and non-diabetic, corresponding with moderate weight ranges.

T-test results indicated statistically insignificant differences in age by gender, but significant differences in BMI by diabetes status, and alcohol year by relationship status.

(Note: Alcohol Year is an estimate of the number of days over the past year that participants drank alcoholic beverages).

In BMI by Diabetes status, the mean BMI for the “No diabetes” group is 25.05, while for the “Yes diabetes” group it is 31.92. This indicates a clear, statistically significant higher average BMI in the diabetic group compared to the non-diabetic group. These findings are consistent with established epidemiological understanding that higher BMI is associated with diabetes risk.
For the Alcohol Year* by Relationship Status, the p-value is extremely small (< 2.2e-16), indicating a highly statistically significant difference in mean estimated drinking days between the “Committed” and “Single” groups. Mean estimated drinking days over the past year are about 68.1 for those in a committed relationship and 26.4 for singles. This indicates that individuals in committed relationships report drinking alcohol on many more days per year on average compared to single individuals. Reasons for this could be due to an increase in social gatherings attended as couples or perhaps a marker for underlying coping mechanisms among individuals in unhealthy relationships, it is difficult to conclude with the current data available.
Overall, these findings emphasise demographic and lifestyle factors as critical determinants of health and behaviour patterns, which can inform targeted health promotion and disease prevention programs.

LIMITATIONS:
The dataset had a notable portion (~21%) of missing values, which were replaced with zeroes during preprocessing; this method could bias results and understate variability. Additionally, the cross-sectional nature of the survey limits causal inference, making it difficult to establish temporal relationships between variables. Self-reported data elements and measurement errors may also affect data accuracy.

IMPLICATIONS:
These findings reinforce the utility of large-scale health surveys like NHANES to monitor population health and inform policy. Public health efforts should focus on addressing the demographic disparities observed, particularly in managing chronic conditions linked to BMI and diabetes. Moreover, integrated prevention strategies considering lifestyle factors, such as smoking and alcohol use, and their relationship to demographic traits will be beneficial.

CONCLUSION
In conclusion, this analysis contributes to understanding key health indicators and their distributions within a population, while highlighting important considerations for methodological approaches and public health interventions.