UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SEMESTER 2/20212022

SECI 2143 - Probability & Statistical Data Analysis

SECTION 02

LECTURER:

Dr. Nor Azizah Ali

GROUP NO: 02

PROJECT 2: Personal Key Indicators of Heart Disease

Name	**Matric ID**
ANG YI QIN	A21EC0163
FONG KHAH KHEH	A21EC0052
SOO WAN YING	A21EC0227
KEE SHIN PEARL	A21EC0190

Table of Contents

1) Introduction or background

The world is going through rapid changes, and the only constant thing that is in our hands is our health. Health is a state of physical, mental, and social well-being and is very important for success and survival. Heart disease is one of the leading causes of death in people. There are many key indicators that lead to heart disease such as smoking, alcohol drinking, mental health, physical health, sleeping time and so on.

The main purpose of this study is to show the key data of personal key indicators of heart disease such as BMI, sleeping time and mental health. They will apply the uses of statistical analysis skills in the dataset. We choose this title as our project topic in order to prove whether the sleep time of females is more than the sleep time of males. We also want to estimate the relationship between sleep time and BMI. Other than that, we want to determine the risk of heart disease depending on the BMI values. Moreover, we need to prove that there is a difference between the observed frequency and expected frequency of the people who have smoked. Last, we want to test whether gender and heart disease has relationship.

The dataset regarding the Personal Key Indicators of Heart Disease is a secondary data source retrieved from the Kaggle website, this data was collected by Kamil Pytlak who is a Bioinformatics Student at Wrocław University of Environmental and Lifescience. The data shows 400,000 adults related to their health status.

2) Dataset

Dataset: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Population: 2020 annual CDC survey data of 400,000 adults related to their health status

Sample: 199 adults

No.	Selected variables	Objectives	Test analysis and expected outcome
a.	Gender, Sleeping time	To test whether the mean sleeping time of woman is larger than the mean sleeping time of man at 95% confidence level, assuming unequal variances.	**Analysis:** 2 Sample Hypothesis Testing (Test on Mean, Variance Unknown) **Expected Outcome:** The mean sleeping time of woman is larger than the mean sleeping time of man at 95% confidence level, assuming unequal variances.
b.	Sleeping time, BMI	To test whether linear relationship exists between the sleeping time and BMI using Pearson's Product Moment Correlation Coefficient, at 95% confidence level.	**Analysis:** Correlation Analysis **Expected Outcome:** Linear relationship exists between the sleeping time and BMI using Pearson's Product Moment Correlation Coefficient, at 95% confidence level.
c.	Physical health, BMI	To test whether the rate of heart disease depends on BMI value, using the BMI values as the independent variable(x) and risk of facing physical health problems as the dependent variable(y).	**Analysis:** Regression Analysis **Expected Outcome:** The rate of heart disease depends on the BMI values. The larger the BMI values, the larger the risk of facing physical health problems.
d.	Smoking	To test whether there is difference between the observed frequency and expected frequency of the people who have smoked or	**Analysis:** Goodness of Fit Test (One Way Contingency Table) **Expected Outcome:**

		not, at 95% confidence level.	There is a difference between the observed frequency and expected frequency of the people who have smoked or not at 95% confidence level. The observed frequency is not a good fit for the assumed distribution.
e.	Gender, Heart Disease	To test whether the gender and heart disease are related using Two Way Contingency Table, at 95% confidence level.	**Analysis:** Chi-Square Test of Independence **Expected Outcome:** The gender and heart disease are related using Two Way Contingency Table, at a 95% confidence level.

3) Data Analysis

a. Two Sample Hypothesis Testing (Test on Mean, Variance Unknown)

For this hypothesis testing, our group decided to use variable **gender** and **sleeping time**. Our objective of this analysis is to test if the mean sleeping time of females is larger than the sleeping time of males at 95% confidence level, $\alpha = 0.05$ and assuming unequal variances. From the data, frequency (n), mean ($\bar{x}$) and standard deviation(s) are calculated.

Calculated and grouped frequency(n), mean($\bar{x}$) and standard deviation(s)

Group$_1$ is for sleep time of female
Group$_2$ is for sleep time of male

$n_1 = 138$	$n_2 = 61$
$\bar{x}_1 = 7.297101$	$\bar{x}_2 = 7.327869$
$s_1 = 1.511076$	$s_2 = 1.795562$

```
> # Mean(xbar)
> xbar1
[1] 7.297101
> xbar2
[1] 7.327869
> # Standard Deviation(s)
> s1
[1] 1.511076
> s2
[1] 1.795562
```

1. **Hypothesis Statement**

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

$\mu_1 = mean\ of\ sleep\ time\ of\ female,\ \mu_2 = mean\ of\ sleep\ time\ of\ male$

2. **Test Statistics, t_0 (Given 95% confidence level, $\alpha = 0.05$)**

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}}$$

```
> t0 = (xbar1-xbar2-0)/(sqrt((s1^2/n1)+(s2^2/n2)))
> t0
[1] -0.1167922
```

Test statistics can be calculated using the formula above.
By using RStudio, test statistics, $t_0 = -0.11679$.

3. **Degree of Freedom, v**

$$v = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1} + \frac{\left(S_2^2/n2\right)^2}{n2-1}}.$$

```
> v=((s1^2/n1)+(s2^2/n2))^2/((((s1^2/n1)^2)/(n1-1))+(((s2^2/n2)^2)/(n2-1)))
> v
[1] 99.18961
> alpha=0.05
> t.alpha=qt(alpha,floor(v))
> t.alpha
[1] -1.660391
```

The formula given is used to calculate the degree of freedom, v.
By using RStudio, degree of freedom, v = 99.19.

4. **Conclusion**

```
> t.test(female_data$SleepTime, male_data$SleepTime)

        Welch Two Sample t-test

data:  female_data$SleepTime and male_data$SleepTime
t = -0.11679, df = 99.19, p-value = 0.9073
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5534717  0.4919369
sample estimates:
mean of x mean of y
 7.297101  7.327869
```

Hence, using 95% confidence level, $\alpha = 0.05$, H_0 will be rejected if $t_{0.025,99.19} = -1.66$.

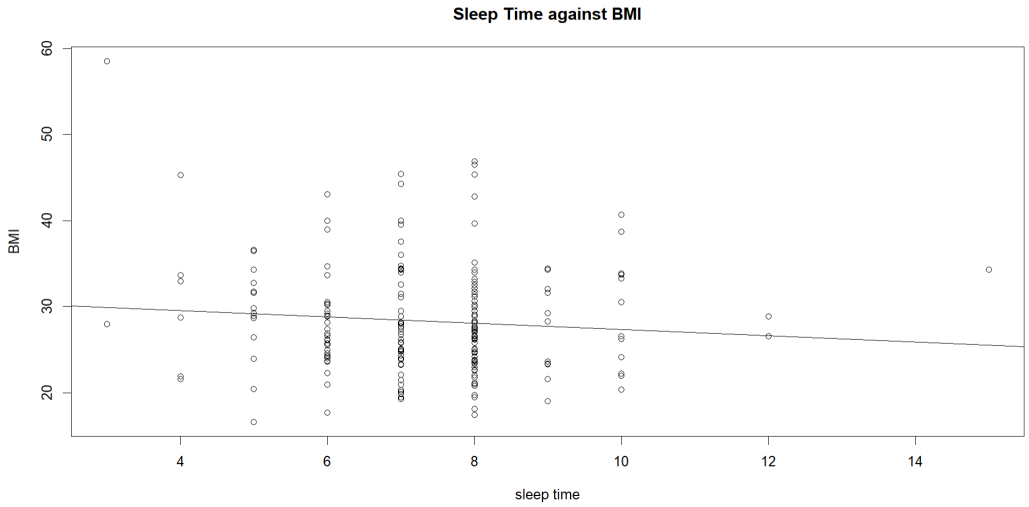
Critical value, $t_{0.025,99.19} = -1.66,\ p - value = 0.9073$

Since the test statistic, t=-0.11679 $> t_{0.025,99.19} = -1.66$. We reject the null hypothesis.

There is sufficient evidence to prove that the sleep time of females is more than the sleep time of males.

b. Correlation test

In this analysis, we are using the variables sleeping time and Body Mass Index (BMI), where we will test whether there is linear relationship between sleeping time and BMI using Pearson's Product Moment Correlation Coefficient. Assume the confidence level to be 95%, significant level, $\propto = 0.05$. Correlation analysis measures the strength of association (linear relationship) between two variables. For the correlation coefficient, we use Pearson's Product Moment Correlation Coefficient since variables are sleep time and BMI are ratio-type variables.



Visualize data in scatter plot using RStudio

From the scatter plot, it indicates that there is a weak correlation relationship between sleeping time and BMI.

1. **Calculate the sample correlation coefficient using Pearson's method by:**

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

where:
- r = Sample correlation coefficient
- n = Sample size
- x = Value of the independent variable
- y = Value of the dependent variable

```
> cor(x,y)
[1] -0.090696
```

By using RStudio, we get sample correlation coefficient, r = -0.090696, which indicates that there is a weak negative correlation between x and y.

2. **Significant Test for Correlation**

- Hypothesis Statement

  H_0: ρ = 0 (*no linear correlation*)

  H_1: ρ ≠ 0 (*linear correlation exists*)

- **Calculate test statistic by:**

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

```
> n <- 200
> r <- cor(x,y)
> t <- r/(sqrt((1-(r^2))/(n-2)))
```

```
t = -1.2782
```

  By using RStudio, test statistic t = -1.2782

- **Find critical value, using** $\alpha = 0.05, df = n - 2 = 197$

  From t-table, since this is a two-tailed test, there are two critical values:

  Lower tail critical value $-t_{\alpha/2=0.025, df=197} = -1.9721$

  Upper tail critical value $t_{\alpha/2=0.025, df=197} = 1.9721$

  From RStudio, we also get p-value = 0.2027
  Hence, if test statistic > 1.9721 or test statistic < -1.9721, reject H_0. Otherwise, fail to reject H_0.

- **State the decision:**
  Since the test statistic, $t = -1.2782 >$ lower tail critical value $-t_{\alpha/2=0.025, df=197} = -1.9721$ .It does not fall within the rejection region. Hence we failed to reject the null hypothesis.
  Since the correlation coefficient, r = -0.090696 which is negative and falls between -0.5 and 0.5, hence it has a weak linear relationship between sleep time and BMI. There is sufficient evidence to prove that there is no linear correlation between sleep time and BMI.

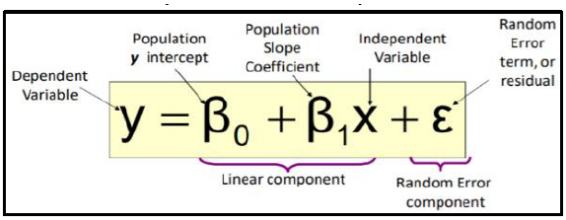
```
> x <- c(Health$SleepTime)
> y <- c(Health$BMI)
> # visualize data
> plot(x, y, xlab="sleep time", ylab="BMI", main = "Sleep Time against BMI")
> # calculate correlation coefficient
> cor(x,y)
[1] -0.090696
> # calculate test statistic t
> n <- 200
> r <- cor(x,y)
> t <- r/(sqrt((1-(r^2))/(n-2)))
> # significance test for correlation
> # H0: no linear correlation
> # H1: linear correlation exists
> cor.test(Health$SleepTime,Health$BMI, method="pearson")

        Pearson's product-moment correlation

data:  Health$SleepTime and Health$BMI
t = -1.2782, df = 197, p-value = 0.2027
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.22692331  0.04901221
sample estimates:
      cor
-0.090696
```

Performing significance test for correlation using RStudio

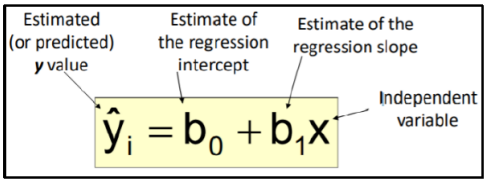
c. Regression Analysis

In this analysis, we are using variable **physical health** and **BMI,** where we will test whether the value of BMI depends on the sleep time, by using BMI values as the independent variable (x) and physical health as the dependent variable (y).

Equation for Population Linear Regression:



1. **Estimated Regression Model:**



From the equation, b_0 is the estimated average value of y ($physical\ health$) when the value of x (BMI) is zero. Whereas, b_1 is the estimated change in the average value of y ($physical\ health$) due to one-unit change in x (BMI).

- **Find least squares criterion:**

From the equation above, we can find the values of b_0 and b_1 by using formula:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

```
> b1 <- (sum(x*y)-(sum(x)*sum(y)/n))/
(sum(x^2)-((sum(x)^2)/n))
> b1
[1] 0.3546155
```

```
> b0 <- mean(y)-(b1*mean(x))
> b0
[1] -4.807351
```

By using RStudio, we get $b_1 = 0.3546$ and $b_0 = -4.8074$

Substitute the variable of b_0 and b_1 into the regression model equation:

$$\hat{y}_i = -4.8074 + 0.3546x$$

- **Explained and Unexplained Variation:**

By using RStudio, we get:

$SSR = 1023.874$

$SST = 16823.44$

$SSE = 15799.56$

$$SST = SSE + SSR$$

Total sum of Squares: $SST = \sum(y - \bar{y})^2$

Sum of Squares Error: $SSE = \sum(y - \hat{y})^2$

Sum of Squares Regression: $SSR = \sum(\hat{y} - \bar{y})^2$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

```
> SSR <- sum((yhat-mean(y))^2)
> SSR
[1] 1023.874
> SST <- sum((y-mean(y))^2)
> SST
[1] 16823.44
> SSE <- SST-SSR
> SSE
[1] 15799.56
```

$$R^2 = \frac{SSR}{SST}$$

By using RStudio, we get

$Coefficient\ of\ Determination,\ R^2 = 0.0608$

∴ Hence, we can interpret it as 6.08% of the variation in physical health is explained by variation in BMI.

- **Find Standard Error of Estimate by:**

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}}$$

By using RStudio, we get Standard Error of Estimate, $s_\varepsilon = 8.9555$

- **Find Standard Deviation of Regression Slope by:**

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum(x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

By using RStudio, we get Standard Deviation of Regression Slope, $s_{b_1} = 0.0992$

2. **Inference about the Slope: t-Test**
   - **Hypothesis Statement:**

     $H_0: \beta_1 = 0\ (no\ linear\ relationship)$

     $H_1: \beta_1 \neq 0\ (linear\ relationship\ does\ exist)$

   - **Find critical value, using** $\alpha = 0.05,\ df = n - 2 = 197$

     From t-table, since this is a two-tailed test, there are two critical values:

     Lower tail critical value $-t_{\alpha/2=0.025, df=197} = -1.9721$

     Upper tail critical value $t_{\alpha/2=0.025, df=197} = 1.9721$

From RStudio, we also get $p - value = 0.2027$

Hence, we reject H_0 if test statistic > 1.9721 or test statistic < -1.9721.

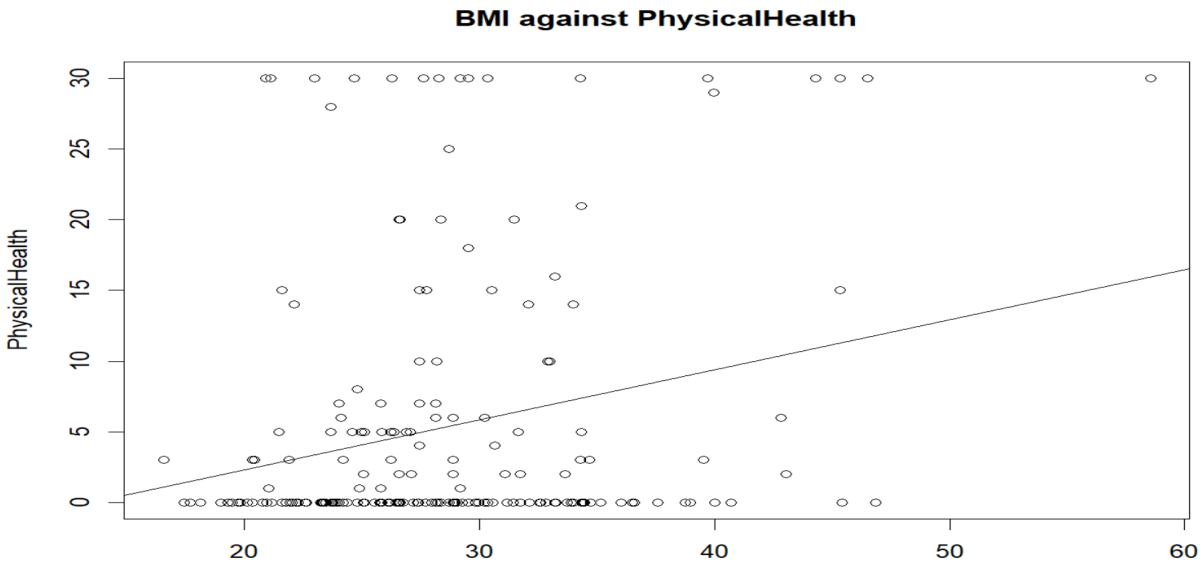
- **Calculate test statistic by:**

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

By using RStudio, we get test statistic $t = 3.5730$

- **State the decision:**

∴ Since test statistic $t = 3.5730 > t_{\alpha/2=0.025, df=197} = 1.9721$, we reject the null hypothesis, H_0. There is sufficient evidence that BMI affects physical health, at $\alpha = 0.05$.



Visualize data in scatter plot using RStudio

d. Goodness of fit test (One Way Contingency Table)

In this Goodness of fit test, also known as the chi-square test with a one-way contingency table, the variable used is **smoking.** Generally, the goodness of fit test is used to test the hypothesis that an observed frequency distribution fits some claimed distribution. Thus, we will test whether there are differences between the observed frequency and expected frequency of a total number of people smoking, at a confidence level of 95%. For the smoking status, it has two choices, which are yes or no. Other than that, the people who smoke are claimed to be distributed with 40% and the people who do not smoke are claimed to be 60%.

<u>Step 1</u>: Statement of test hypothesis:

H_0: $\rho_{smoke} = 0.40$, $\rho_{do\ not\ smoke} = 0.60$

H_1: $At\ least\ one\ of\ the\ proportions\ is\ different\ from\ the\ claimed\ value$

<u>Step 2</u>: Calculate the expected frequency:

	Smoke	Do not smoke	Total
Observed frequency, O	83	116	199
Expected Frequency, E	$np = 199 \times 0.4 = 79.6$	$np = 199 \times 0.6 = 119.4$	199

<u>Step 3</u>: Calculate the test statistic @ chi-square value:

Formula: $\chi^2 = \sum \frac{(O-E)^2}{E}$

```
data:  smoking
X-squared = 0.24204, df = 1, p-value = 0.6227
```

By using RStudio, test statistics value $x^2 = 0.6227$.

<u>Step 4</u>: Find the critical value from chi-square table/ RStudio:

$Degree\ of\ freedom = n - 1 = 2 - 1 = 1,\ \alpha = 0.05$

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09

From the Chi-square distribution table, critical value, $x^2 = 3.84$

```
[1] 0.05
> #Finding critical value
> x2.alpha <- qchisq(alpha,df=1, lower.tail = FALSE)
> x2.alpha
[1] 3.841459
```

From RStudio, the critical value, $x^2 = 3.841459$

<u>Step 5</u>: State the decision:

Test statistic value ($x^2 = 0.6227$) < critical value ($x^2_{k=1,\alpha=0.05} = 3.84$), we **do not reject** H_0 as it falls within the critical region. There is sufficient evidence to warrant rejection of the claim that the total number who smokes is distributed with the given percentages. Thus, it shows that there are differences between the observed frequency and expected frequency who smokes and who do not smoke.

e. Chi-Square Test of Independence

In this analysis, we are using variables **sex** and **heart disease**, where we will test whether the sex and heart disease are related using the Two Way Contingency Table, at a 95% confidence level.

1. **State the test hypothesis:**

   H_0 : *There is no relationship between variables gender and heart disease*

   H_1 : *Variables gender and heart disease are related and dependent*

2. **Find the critical value:**

   Critical value, x^2=3.841459

   (with $df = (2-1)(2-1) = 1, \alpha = 0.05$))

```
> # find critical value
> alpha <- 0.05
> x2.alpha <- qchisq(alpha, df=1, lower.tail=FALSE)
```

3. **Calculate the expected counts:**

Sex	**Heart Disease**				**Total**
	No		**Yes**		
	Obs	**Exp**	**Obs**	**Exp**	
Female	124	$\frac{138\times177}{199} = 122.744$	14	$\frac{138\times22}{199} = 15.256$	138
Male	53	$\frac{61\times177}{199} = 54.256$	8	$\frac{61\times22}{199} = 6.744$	61
Total	177	177	22	22	199

4. Calculate the test statistic value using RStudio:

```
> # perform chi-square test
> tbl = table(Health$Sex, Health$HeartDisease)
> chisq.test(tbl, correct=FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 0.37943, df = 1, p-value = 0.5379
```

When we calculate the statistic using RStudio, we get $x^2 = 0.37943$, with $p - value = 0.5379$.

5. State the decision:

∴ Since the test statistic value ($x^2 = 0.37943$) < critical value ($x^2_{k=1,\alpha=0.05} = 3.841$), hence we fail to reject H_0. There is sufficient evidence to conclude that there is no relationship between sex and heart disease, at $\alpha = 0.05$.

4) Conclusion

In conclusion, our group have done the whole project, from choosing a suitable dataset, pre-processing and analysis process. Although it was quite a challenging project for us to complete, we enjoyed it when we are discussing, communicating, and solving problems that we faced when we are doing this project. At first, on the project, we find datasets and do surveys for numerous types of datasets, luckily we have found a dataset which is suitable for our project. After completing this project, we have a deeper knowledge of test analysis. For example, Hypothesis 1 or 2 sample test, Correlation test, Regression test, Goodness of fit test and Chi-Square test of Independence. After having this knowledge of test analysis, I believe that it will be very useful and helpful for our further study and internship. All in all, we are thankful that our lecturer, Dr Nor Azizah Ali for guiding and helping us when we face problems throughout the project.

5) Appendix (Sample of your original/raw dataset and your processed dataset (if applicable)

Type and measurement level for variables:

Variable	Type of variable	Measurement level
Heart Disease (Yes/ No)	Qualitative	Nominal
BMI (Metric Number)	Quantitative	Ratio
Smoking (Yes/ No)	Qualitative	Nominal
Alcohol Drinking (Yes/ No)	Qualitative	Nominal
Stroke (Yes/ No)	Qualitative	Nominal
Physical Health (Metric Number)	Quantitative	Ratio
Mental Health (Metric Number)	Quantitative	Ratio
Diff Walking (Yes /No)	Qualitative	Nominal
Gender (Female/Male)	Qualitative	Nominal
Age Category (Metric Number)	Quantitative	Ordinal
Race (White/ Black/ Asian/ American/ Indian)	Qualitative	Nominal
Diabetic (Yes/ No)	Qualitative	Nominal
Physical Activity (Yes/ No)	Qualitative	Nominal
Gen Health (Very Good/ Good/ Fair/ Poor)	Qualitative	Ordinal
Sleep Time (Metric Number)	Quantitative	Ratio
Asthma (Yes/ No)	Qualitative	Nominal
Kidney Disease (Yes/ No)	Qualitative	Nominal
Skin Cancer (Yes/ No)	Qualitative	Nominal

Meeting date: 9th June 2022 (2 pm - 5.30 pm)