***Data Activity 1.1**

Download the **Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset** from its catalogue page. It is an open access dataset which the data are available to download without any registration with the UK Data Service.

***Data Activity 1.2**

Using the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset, assess the level of anti-social behaviour that the survey respondents experience in their neighbourhood by creating a summary statistic, using the ‘antisocx’ variable.

```
library(haven)

csew1314teachingopen <- read_sav("C:/Users/Owner/Desktop/DataScience/R/Data Activities 1/csew1314teachingopen.sav")

View(csew1314teachingopen)

# load the .sav file

data <- read_spss("csew1314teachingopen.sav")

# calculate summary statistics for antisocx variable

summary(data$antisocx)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's

 -1.215  -0.788  -0.185  -0.007   0.528   4.015    6694
```

The mean, standard deviation, minimum, and maximum values of the variable were calculated for this purpose.

Learning Outcome

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.

***Data Activity 2**

Using the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset (see Unit 1), perform the following activities:

1. Explore whether survey respondents experienced any crime in the 12 months prior to the survey using the variable *bcsvictim*.

```
# check if any respondent experienced any crime in the previous 12
months

any_crime <- any(!is.na(data$bcsvictim))

# print the result

if (any_crime) {

  cat("Some respondents experienced crime in the previous 12 months.\n")

} else {

  cat("No respondents experienced crime in the previous 12 months.\n")

}

Some respondents experienced crime in the previous 12 months.
```

2. Create a frequency table to count if the survey respondents experienced any crime in the previous 12 months. Use the *table()* command.

```
# create a frequency table for bcsvictim

freq_bcsvictim <- table(data$bcsvictim, useNA = "always")

# print the frequency table

print(freq_bcsvictim)

   0    1 <NA>

7460 1383    0
```

The frequency table shows that the variable bcsvictim has more than two distinct values.

3. Assess the results and decide if you need to convert this variable into a factor variable. Use *as_factor*.

```
# convert bcsvictim into a factor variable

data$bcsvictim <- as_factor(data$bcsvictim)
```

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.

***Data Activity 3**

Using the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset (see Unit 1), perform the following activity:

Create a subset of individuals who belong to the '75+' age group and who were a 'victim of crime' that occurred in the previous 12 months. Save this dataset under a new name 'crime_75victim'.

```
# Convert bcsvictim into a numeric variable

csew1314teachingopen$bcsvictim <- as.numeric(csew1314teachingopen$bcsvictim)
# Create subset of individuals who belong to the '75+' age group == 7 and who
were a 'victim of crime' == 1 that occurred in the previous 12 months

crime_75victim <- subset(csew1314teachingopen, agegrp7 == 7 & bcsvictim == 1)

# save the dataset under a new name 'crime_75victim'

library(foreign)
write.foreign(crime_75victim, "crime_75victim.sav", "crime_75victim", package =
"SPSS")
```

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.

***Data Activity 4**

Using the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset (see Unit 1), perform the following activities:

1. Create a boxplot for the variable 'antisocx'

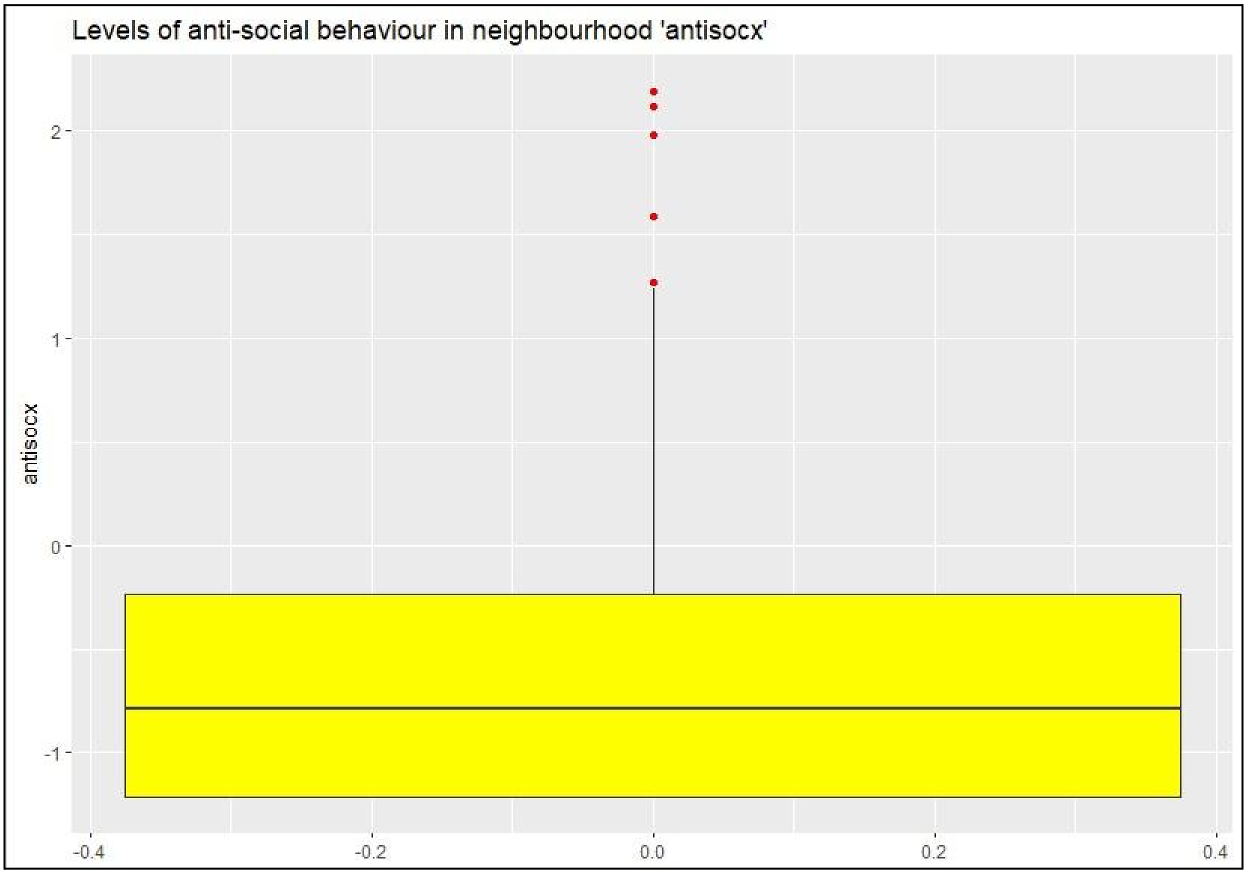
Follow the instructions below to create a boxplot for assessing levels of anti-social behaviour that the survey respondents experience in their neighbourhood (use the variable: antisocx).

If you're using 'graphics': Add "Levels of anti-social behaviour in neighbourhood 'antisocx'" as a title and colour the plot in purple and colour the outliers in blue.

```
# Create boxplot for antisocx

boxplot(crime_75victim$antisocx, main = "Levels of anti-social behaviour in
neighbourhood 'antisocx'", col = "purple", outlier.col = "blue")
```

If you're using 'ggplot2': Add "Levels of anti-social behaviour in neighbourhood 'antisocx' as a title, colour the plot in yellow and colour the outliers in red.

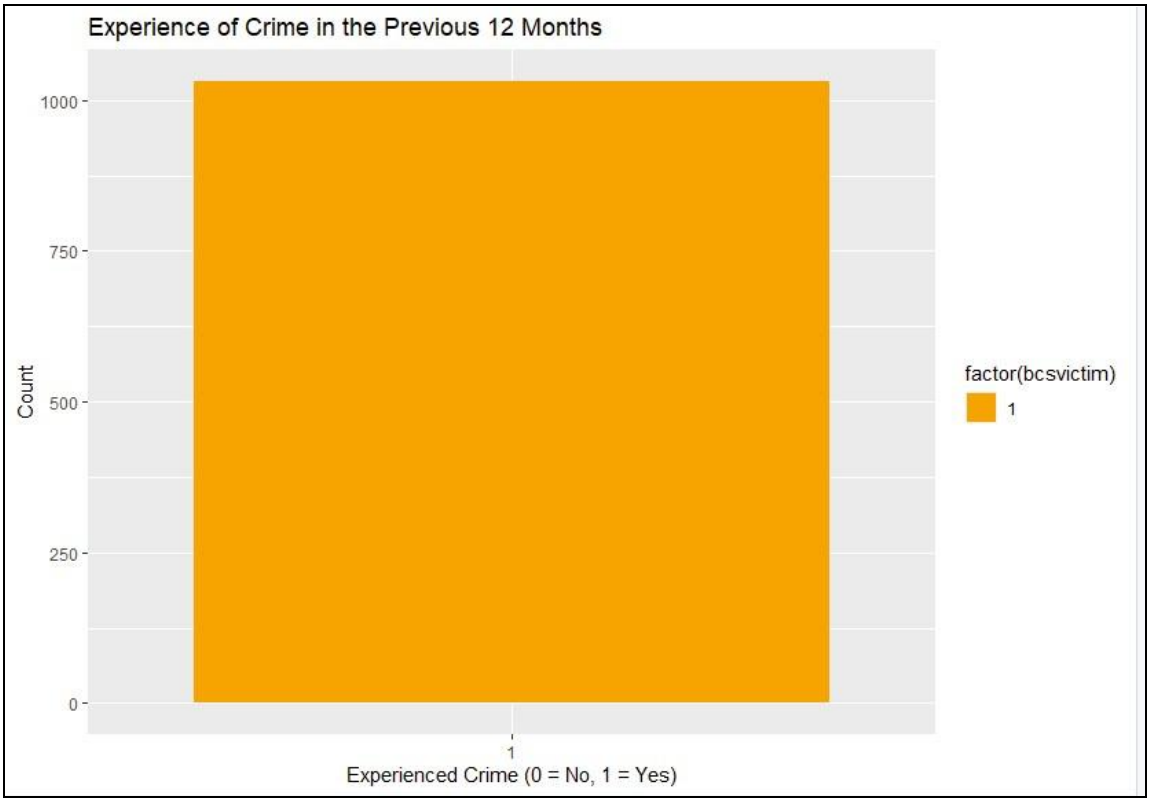
```
# Create barplot for bcsvictim

barplot(table(crime_75victim$bcsvictim), main = "Experience of crime in the
previous 12 months", col = "orange")
```



2. Create a bar plot using either the barplot() function or the ggplot() function to assess whether or not the survey respondents experienced crime in the 12 months prior to the survey (use the variable 'bcsvictim'). Give the graph a suitable title and choose a colour for the bars (e.g., orange).

```
# Plot the bar chart

ggplot(crime_75victim, aes(x = factor(bcsvictim), fill =
factor(bcsvictim))) +

geom_bar() +

scale_fill_manual(values = "orange") +

ggtitle("Experience of Crime in the Previous 12 Months") +

xlab("Experienced Crime (0 = No, 1 = Yes)") +
```

```
ylab("Count")
```



Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.

***Data Activity 5**

Using the **Health_Data**, please perform the following functions in R:

- Find out mean, median and mode of variables *sbp, dbp and income.*

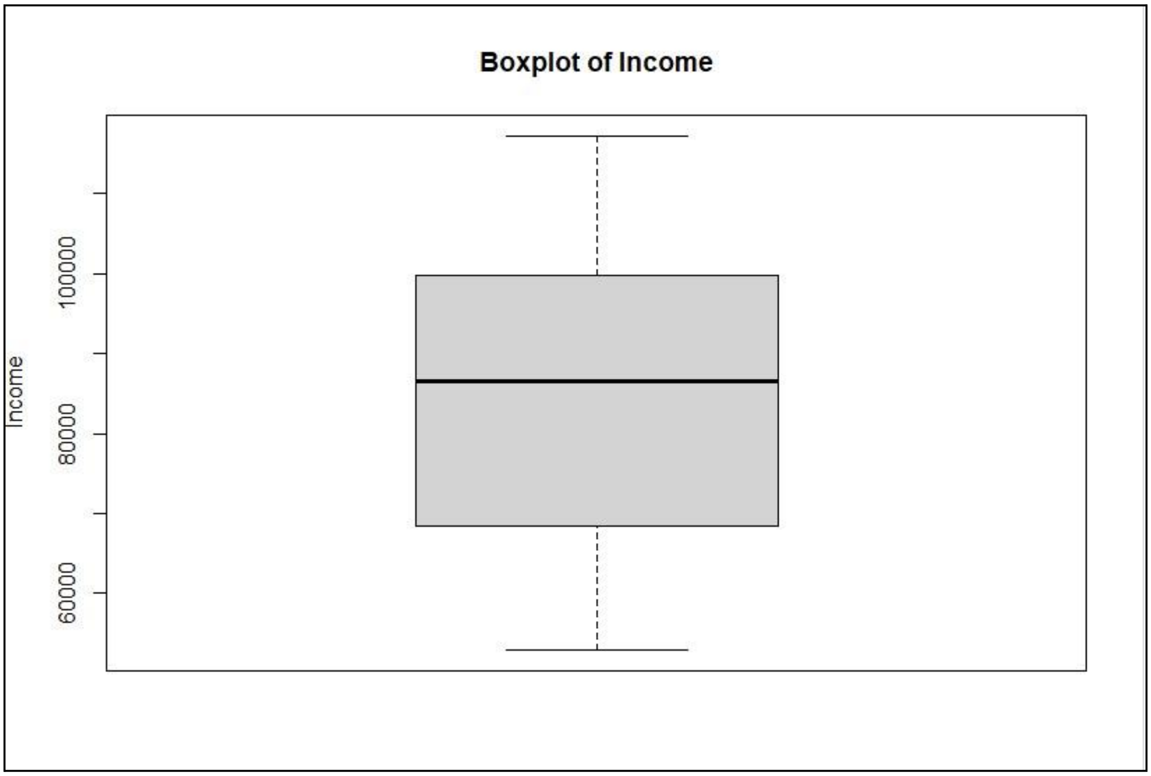
```
library(haven)

Health_Data <- read_sav("Health Data.sav")
```

```r
View(Health_Data)


# mean

mean(Health_Data $sbp)

[1] 127.7333

mean(Health_Data $dbp)

[1] 82.76667

mean(Health_Data $income)

[1] 85194.49


# median

median(Health_Data $sbp)

[1] 123

median(Health_Data $dbp)

[1] 82

median(Health_Data $income)

[1] 86560.5


# mode (using the 'Mode' function defined below)

Mode <- function(x) {

    ux <- unique(x)

    ux[which.max(tabulate(match(x, ux)))]

}
```

```
Mode(Health_Data $sbp)

[1] 120

Mode(Health_Data $dbp)

[1] 80

Mode(Health_Data $income)

[1] 79774
```

- Find out the five-figure summary of *income* variables and present it using a Boxplot.

```
summary(Health_Data $income)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  52933   68637   86561   85194   99696  117210

boxplot(Health_Data $income, main="Boxplot of Income", ylab="Income")
```



- Run a suitable hypothesis test to see if there is any association between systolic blood pressure and presence and absence of peptic ulcer.

```
# Create a contingency table

table <- table(Health_Data$sbp , Health_Data$pepticulcer )


# print table

table



       1  2

  91   0  1

  95   1  0

  97   1  0

  100  2  3
```

101	1	3
102	0	9
103	0	2
104	1	1
105	2	1
106	0	3
107	0	5
108	0	3
109	0	1
110	1	6
111	0	2
112	0	1
113	0	2
114	0	3
115	0	6
116	0	5
117	0	2
118	0	4
119	1	6
120	2	10
121	0	3
122	2	7
123	1	3
124	2	6
125	0	1
126	0	3

127	0	2
128	1	1
129	0	3
130	1	2
131	0	1
132	1	4
133	1	0
134	1	1
135	0	1
136	0	1
137	0	1
138	0	3
139	2	0
140	2	7
141	2	0
142	0	3
143	2	0
144	1	2
145	1	4
146	1	1
147	2	1
148	0	4
149	0	2
150	0	1
151	1	2
152	1	5

```
  153  0  1
  154  1  1
  155  1  0
  157  0  1
  158  0  1
  163  0  1
  164  0  1
  167  0  2
  172  0  2
  174  1  1
  176  0  2
  186  0  1
  190  0  1
  195  1  0
```

```r
# Calculate the percentage of each cell
prop.table(table) * 100
```

```
              1         2
  91  0.0000000 0.4761905
  95  0.4761905 0.0000000
  97  0.4761905 0.0000000
  100 0.9523810 1.4285714
  101 0.4761905 1.4285714
  102 0.0000000 4.2857143
  103 0.0000000 0.9523810
```

104	0.4761905	0.4761905
105	0.9523810	0.4761905
106	0.0000000	1.4285714
107	0.0000000	2.3809524
108	0.0000000	1.4285714
109	0.0000000	0.4761905
110	0.4761905	2.8571429
111	0.0000000	0.9523810
112	0.0000000	0.4761905
113	0.0000000	0.9523810
114	0.0000000	1.4285714
115	0.0000000	2.8571429
116	0.0000000	2.3809524
117	0.0000000	0.9523810
118	0.0000000	1.9047619
119	0.4761905	2.8571429
120	0.9523810	4.7619048
121	0.0000000	1.4285714
122	0.9523810	3.3333333
123	0.4761905	1.4285714
124	0.9523810	2.8571429
125	0.0000000	0.4761905
126	0.0000000	1.4285714
127	0.0000000	0.9523810
128	0.4761905	0.4761905
129	0.0000000	1.4285714

130	0.4761905	0.9523810
131	0.0000000	0.4761905
132	0.4761905	1.9047619
133	0.4761905	0.0000000
134	0.4761905	0.4761905
135	0.0000000	0.4761905
136	0.0000000	0.4761905
137	0.0000000	0.4761905
138	0.0000000	1.4285714
139	0.9523810	0.0000000
140	0.9523810	3.3333333
141	0.9523810	0.0000000
142	0.0000000	1.4285714
143	0.9523810	0.0000000
144	0.4761905	0.9523810
145	0.4761905	1.9047619
146	0.4761905	0.4761905
147	0.9523810	0.4761905
148	0.0000000	1.9047619
149	0.0000000	0.9523810
150	0.0000000	0.4761905
151	0.4761905	0.9523810
152	0.4761905	2.3809524
153	0.0000000	0.4761905
154	0.4761905	0.4761905
155	0.4761905	0.0000000

```
  157 0.0000000 0.4761905
  158 0.0000000 0.4761905
  163 0.0000000 0.4761905
  164 0.0000000 0.4761905
  167 0.0000000 0.9523810
  172 0.0000000 0.9523810
  174 0.4761905 0.4761905
  176 0.0000000 0.9523810
  186 0.0000000 0.4761905
  190 0.0000000 0.4761905
  195 0.4761905 0.0000000
```

```
# Create a contingency table
cont_table <- table(Health_Data$sbp , Health_Data$pepticulcer )
# Run a chi-squared test
chisq.test(cont_table)
```

```
	Pearson's Chi-squared test

data:  cont_table
X-squared = 86.154, df = 69, p-value = 0.07928
```

The output suggests that a Pearson's chi-squared test was conducted on a contingency table, which has 69 degrees of freedom. The test statistic is X-squared = 86.154, and the p-value is 0.07928.

Assuming a significance level of 0.05, since the p-value (0.07928) is greater than the significance level, we fail to reject the null hypothesis. This means that there is not enough evidence to conclude that there is a significant association between the variables represented in the contingency table.

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.

***Data Activity 6**

Before carrying out this data activity, review the Unit 8 notes on Nonparametric Tests. This will provide further insights on how to utilise R for these tests.

Using the **Health_Data**, please perform the following functions in R:

1. Find out the mean, median and mode of 'age' variable.

```
mean <- mean(Health_Data$age, na.rm = TRUE) # mean

median <- median(Health_Data$age, na.rm = TRUE) # median

mode <- names(sort(table(Health_Data$age), decreasing = TRUE)[1]) # mode


# Print the results

cat(paste0("Mean: ", mean, "\n"))

Mean: 26.5142857142857

cat(paste0("Median: ", median, "\n"))

Median: 27

cat(paste0("Mode: ", mode, "\n"))

Mode: 26
```

2. Find out whether median diastolic blood pressure is the same among diabetic and non-diabetic participants.

```
# Subset data to only include diabetic and non-diabetic participants

dbp_data <- subset(Health_Data, !is.na(diabetes), select = c("dbp",
"diabetes"))
```

```r
# Split the data into two groups: diabetic and non-diabetic

dbp_diabetic <- dbp_data[dbp_data$diabetes == 1, ]

dbp_non_diabetic <- dbp_data[dbp_data$diabetes == 2, ]

# Calculate the median diastolic blood pressure for each group

median_dbp_diabetic <- median(dbp_diabetic$dbp)

median_dbp_non_diabetic <- median(dbp_non_diabetic$dbp)


# Perform a two-sample Wilcoxon rank sum test to compare the medians

wilcox.test(dbp_diabetic$dbp, dbp_non_diabetic$dbp)


      Wilcoxon rank sum test with continuity correction


data:  dbp_diabetic$dbp and dbp_non_diabetic$dbp

W = 3804.5, p-value = 0.7999

alternative hypothesis: true location shift is not equal to 0
```

The test result shows a test statistic (W) of 3804.5 and a p-value of 0.7999. The null hypothesis for the Wilcoxon rank sum test is that there is no difference in the location (median) between the two groups. Since the p-value is greater than the significance level (usually set to 0.05), we fail to reject the null hypothesis. This means that there is not enough evidence to conclude that the median diastolic blood pressure is different between diabetic and non-diabetic participants.

3. Find out whether systolic BP is different across occupational groups.

```r
# Subset data to only include relevant variables

bp_occupation <- subset(Health_Data, !is.na(occupation), select = c("sbp",
"occupation"))
```

```r
# Run ANOVA

bp_lm <- lm(sbp ~ occupation, data = bp_occupation)

bp_anova <- anova(bp_lm)


# Print ANOVA table

print(bp_anova)

Analysis of Variance Table


Response: sbp

            Df Sum Sq Mean Sq F value Pr(>F)

occupation   1    101  101.36   0.251 0.6169

Residuals  208  83984  403.77
```

The table shows the degrees of freedom (Df), the sum of squares (Sum Sq), the mean squares (Mean Sq), the F-value, and the p-value for the occupation factor and the residual (error) term. The F-value for the occupation factor is 0.251 with a p-value of 0.6169, which suggests that there is no significant difference in systolic blood pressure across occupational groups.

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.

***Data Activity 7**

Using the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset, perform the following activities:

1. Create a crosstab to assess how individuals' experience of any crime in the previous 12 months bcsvictim vary by age group agegrp7. Create the crosstab with bcsvictim in the rows and agegrp7 in the columns, and produce row percentages, rounded to 2 decimal places.
2. Looking at the crosstab you have produced, which age groups were the most likely, and least likely, to be victims of crime?

```
# load the required package

library(dplyr)



# create the crosstab

mytable <- crime_75victim %>%

  group_by(bcsvictim, agegrp7) %>%

  summarise(n = n()) %>%

  mutate(pct = round(n/sum(n)*100, 2))



# pivot the table to have bcsvictim in rows and agegrp7 in columns

mytable <- pivot_wider(mytable, names_from = agegrp7, values_from = pct)



# rename the columns

colnames(mytable)[2:8] <- c("18-24", "25-34", "35-44", "45-54", "55-64",
"65-74", "75+")



# print the table

mytable
```

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.
- Critically evaluate the use of statistical analysis and the numeric interpretation of results as aids in the decision-making process.

***Data Activity 8**

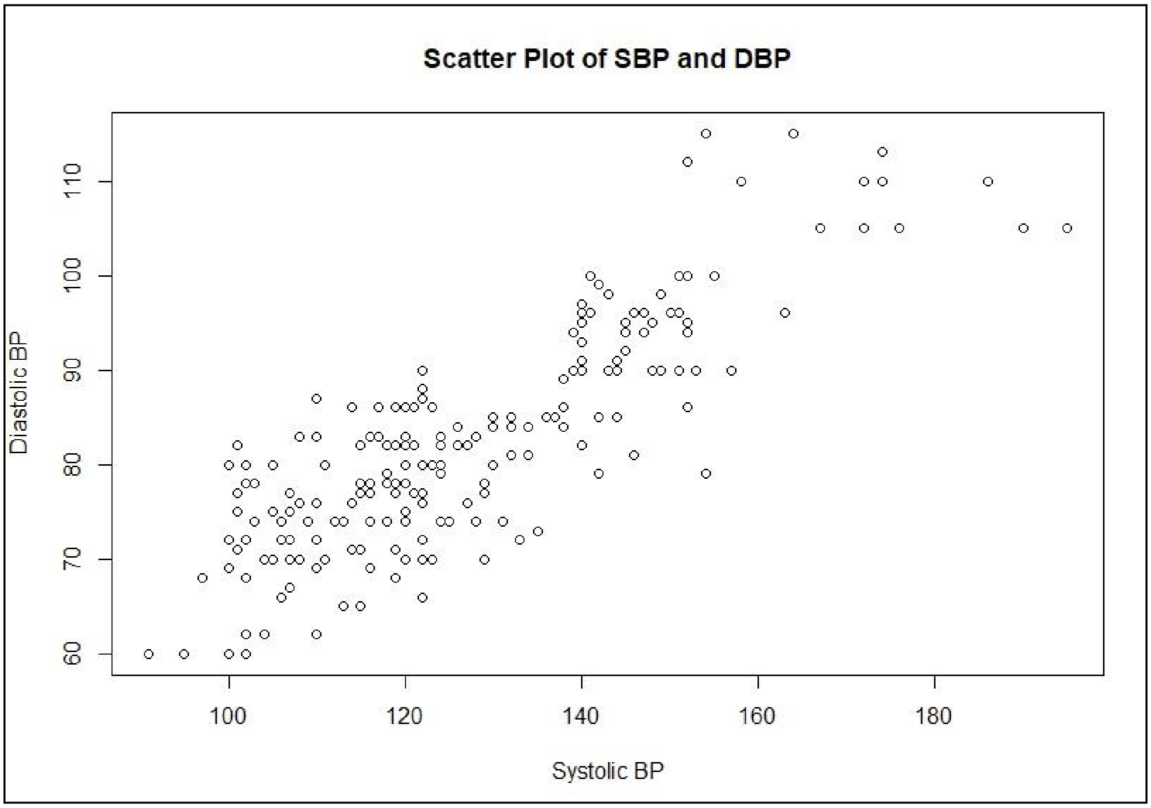
Using the **Health_Data**, please perform the following functions in R:

1. Find out correlation between systolic and diastolic BP.

```
# Find the correlation between systolic and diastolic BP

correlation <- cor(Health_Data$sbp, Health_Data$dbp)

correlation

[1] 0.846808
```

2. Produce a scatter plot between systolic and diastolic BP.

```
# Create a scatter plot between systolic and diastolic BP

plot(Health_Data$sbp, Health_Data$dbp, main = "Scatter Plot of SBP and
DBP", xlab = "Systolic BP", ylab = "Diastolic BP")
```



Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.
- Critically evaluate the use of statistical analysis and the numeric interpretation of results as aids in the decision-making process.

***Data Activity 9**

Before carrying out this data activity, review the Unit 11 notes on Regression Analysis. This will provide further insights on how to utilise R for these tests.

Using the **Health_Data**, please perform the following functions in R:

1. Perform simple linear regression analysis to find the population regression equation to predict the diastolic BP by systolic BP.

```r
# Create a data frame with the variables of interest

data <- data.frame(Health_Data$sbp, Health_Data$dbp)

# Perform linear regression analysis

model <- lm(dbp ~ sbp, data = Health_Data)

# Print the model summary

summary(model)
```

```
Call:

lm(formula = dbp ~ sbp, data = Health_Data)


Residuals:

     Min        1Q    Median        3Q       Max

-16.7958  -3.9366    0.1804    3.6685   19.2042


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  19.4068     2.7931   6.948 4.67e-11 ***

sbp           0.4960     0.0216  22.961  < 2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 6.264 on 208 degrees of freedom

Multiple R-squared:  0.7171,   Adjusted R-squared:  0.7157

F-statistic: 527.2 on 1 and 208 DF,  p-value: < 2.2e-16
```

2. Interpret the findings of regression analysis at 5% level of significance.

The findings of the regression analysis show that there is a statistically significant positive relationship between diastolic blood pressure (dbp) and systolic blood pressure (sbp) at the 5% level of significance. The intercept of the regression equation is 19.4068, which means that when sbp is zero, dbp is expected to be 19.4068. The slope of the regression equation is 0.4960, which means that for every unit increase in sbp, dbp is expected to increase by 0.4960 units, on average.

The regression model has a multiple R-squared value of 0.7171, which indicates that approximately 71.71% of the variability in dbp can be explained by the linear relationship with sbp. The adjusted R-squared value is 0.7157, which takes into account the number of variables in the model, and is slightly lower than the multiple R-squared value. The F-statistic has a value of 527.2 and a p-value less than 2.2e-16, indicating that the overall model is significant.

The residuals of the model appear to be normally distributed, with no obvious patterns or trends in the residual plot. The residual standard error is 6.264, which is the average distance of the observed dbp values from the predicted values in the regression equation.

Overall, these findings suggest that sbp is a significant predictor of dbp, and the regression equation can be used to predict dbp values based on sbp values.

Learning Outcomes

- Systematic understanding of the key mathematical and statistical concepts and techniques which underpin mechanisms in Data Science and AI.
- Apply mathematical and statistical methods in these fields to help in the decision-making process.
- Critically evaluate the use of statistical analysis and the numeric interpretation of results as aids in the decision-making process.

References:

https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8011#!/access-data