Individual Reflection – Module: Numerical Analysis

1.0 Introduction

I considered the module Numerical Analysis a very important part of my learning process, as I understand it as a core component in machine learning. I had invested much time in this module and this reflection details about my learning journey.

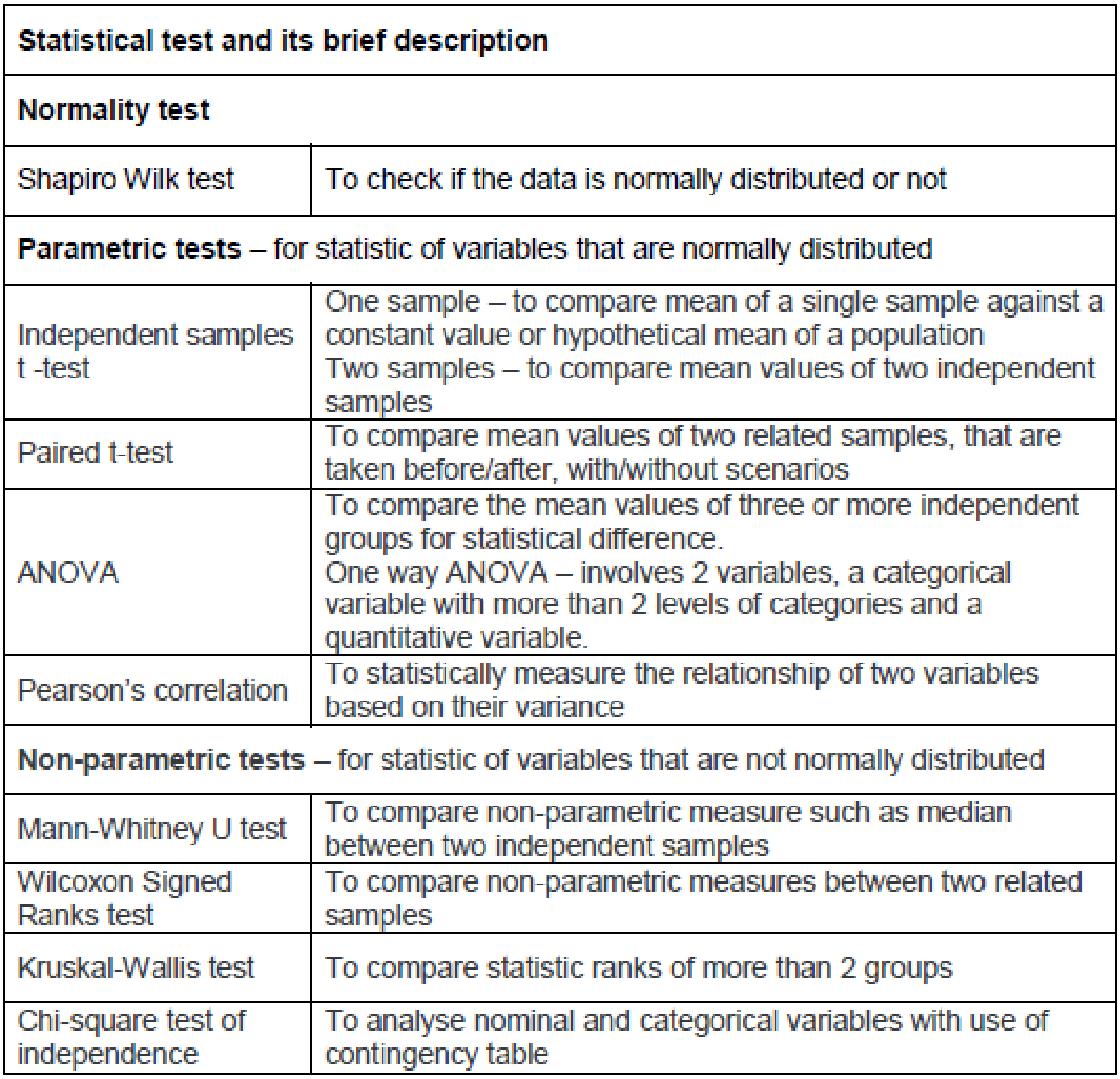
2.0 Knowledge gained throughout this module

Throughout this module, I have acquired the following statistical analysis knowledge and hand-on experience on performing the analysis with R Studio software.

- Good understanding of the terminologies, Population, Parameter, Sample, Statistic, independent and dependent variables.
- Experience in installing and running R studio software, importing and exporting datasets.
- Learned about different variable types: Quantitative data such as discrete and continuous (ratio, interval) and Qualitative or categorical data such as binary, nominal and ordinal, types of data structures in R such as Vectors. Lists, Matrices, Arrays, Data frames and Factors, converting variable types and managing missing values in the dataset with na.rm commands
- Knowledge of sampling methods such as simple random, stratified random, systematic and cluster, discrete probability distributions such as Binomial Simulation, Poisson Simulation Hypergeometric Simulation, discrete Gaussian and continuous probability distributions such as Normal, Uniform, Exponential and Gaussian
- Clear understanding of descriptive statistic measure of location such as mean, median and mode, measure of dispersion/ variability of the data such as variance, standard deviation, range, minimum, maximum, quartiles Q1 and

Q3, five figure statistical summary, percentiles and execution in R using functions such as Data(), names(), print(), str(), view(), class(), head(), table(), summary(), mean(), median(), var(), sd(), quantile(), range(), min(), max(), describe()

- Graphical data representation such as bar chart, histogram, Pie chart, scatter plot, box plot, line graph and producing them using R functions
- Extracting subset of data from the dataset using the function subset() and writing as a separate dataset.
- Clear-cut understanding of statistical inference, hypothesis testing, defining null and alternative hypothesis, type 1 and 2 errors, confidence interval, p-value and level of significance
- Performing normality tests, parametric test and non-parametric tests in R and interpreting the displayed results, creating cross-tabulations, which are detailed below.



Statistical test and its brief description	
Normality test	
Shapiro Wilk test	To check if the data is normally distributed or not
Parametric tests – for statistic of variables that are normally distributed	
Independent samples t -test	One sample – to compare mean of a single sample against a constant value or hypothetical mean of a population Two samples – to compare mean values of two independent samples
Paired t-test	To compare mean values of two related samples, that are taken before/after, with/without scenarios
ANOVA	To compare the mean values of three or more independent groups for statistical difference. One way ANOVA – involves 2 variables, a categorical variable with more than 2 levels of categories and a quantitative variable.
Pearson's correlation	To statistically measure the relationship of two variables based on their variance
Non-parametric tests – for statistic of variables that are not normally distributed	
Mann-Whitney U test	To compare non-parametric measure such as median between two independent samples
Wilcoxon Signed Ranks test	To compare non-parametric measures between two related samples
Kruskal-Wallis test	To compare statistic ranks of more than 2 groups
Chi-square test of independence	To analyse nominal and categorical variables with use of contingency table

- Understanding the relationship between variables using correlation and linear regression analysis, predicting value of dependent variable and interpreting results in R
- Knowledge about Bayesian data analysis and its strengths and limitations

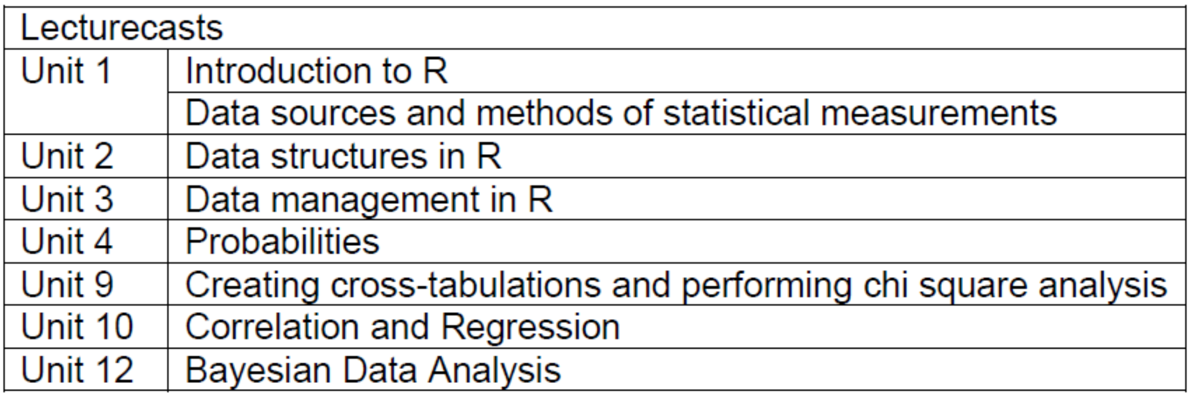
3.0 Activities carried out

In the course of this module, I carried out these activities independently to acquire the knowledge detailed in section 2.0.

Individual Reflection – Module: Numerical Analysis

Reading: I read the relevant chapters in the three core books listed in the module (Berenson et al., 2015; Bruce et al., 2020; Holmes et al., 2017), reading list and notes section in each unit, which provided the basis of my learning.

Lectures: I actively listened to the following lecture casts which were very useful in understanding principal points in each unit.



Lecturecasts	
Unit 1	Introduction to R
	Data sources and methods of statistical measurements
Unit 2	Data structures in R
Unit 3	Data management in R
Unit 4	Probabilities
Unit 9	Creating cross-tabulations and performing chi square analysis
Unit 10	Correlation and Regression
Unit 12	Bayesian Data Analysis

Seminars: I attended Tutor-led seminars offline which were very useful in all aspects, especially, in the preparation of module assessments such as the Mathematics test, Statistical analysis presentation and also this reflective writing.

Data activities: I independently performed the data activities in each unit using R Studio. These activities provided me confidence in applying statistical analysis tools in R to address research problems in a very practical way. Screenshots of activities performed attached in Appendix section

Module assessment: I carried out activities required for the statistical analysis presentation and Mathematics test.

Working out exercises in books: I worked out the solutions to the problems in the probability chapter of Holmes et al. (2017), which helped me to understand more about probability theory.

Tutorials: I watched several Khan Academy YouTube tutorials on statistics topics such as hypothesis testing, p values, probability and completed R tutorials in w3schools.

Discussion with others: I had verbal discussion with work colleagues about misinterpretations in using statistical tests on the basis of the guidelines reported by Greenland et al. (2017).

4.0 Emotional response and analysis

I had a very positive learning experience throughout this module. Progressing through each activity that I carried out, gave me more confidence, increased my interest and helped in more engagement on the subsequent activities.

Before this module, I had a good knowledge about the descriptive statistics, but inferential statistics was a grey area for me. Throughout this module, I had gained understanding about more practical ways to apply the acquired knowledge like what type of statistical test to perform for a given research problem, how to get the results with R and interpretation of results. I acquired more insights about applying my learned skills in my professional front. For example, at work I encountered a research question on comparing the dosimetric efficiency between two radiotherapy treatment techniques. I could immediately work out the type of data, samples required, ways of graphical data representation, appropriate statistical test to be used (paired t-test), defining the hypothesis and result interpretation using p-value.

I was a beginner to use R Studio, initially was challenging using it to import SPSS dataset, as the system couldn’t identify the location of library files and finally was able to fix it after a week of researching about R, through trouble shooting in R community blogs.

5.0 Learning and changed actions

This module involves learning of definitions, terminologies, concepts and number of statistical testing and interpretation methods that are required for gaining high level of knowledge and skills for statistical analysis. All the way through this module, I gained knowledge and practical skills through self-led learning process as reported by Hooshyar et al. (2020). I made notes while reading, attending seminars, lectures and tutorials and processed my thoughts during quite times to build a descriptive picture of all components that I learned, evaluated myself on my capability to use them fluently, prepared a personal learning plan with set targets and deadlines to meet, assessed my learning power (Deakin Crick, 2007) and regularly reviewed it to overcome the challenges that I faced. This helped in my learning process so that I am confident and I will not forget the skills that I learned in a long run.

6.0 Conclusion

In conclusion, I can confidently say that I feel equipped with knowledge and skills to perform statistical analysis, using R for any given research problem.

7.0 References

Berenson, L., Levine, D. & Szabat, K. (2015) Basic Business Statistics: Concepts and Applications. 13th ed. Pearson

Bruce, P., Bruce, A. & Gedeck, P. (2020) Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media.

Deakin Crick, R. (2007) Learning how to learn: the dynamic assessment of learning power. *The Curriculum Journal* 18(2): 135-153. DOI: https://doi.org/10.1080/09585170701445947

Greenland, S., Senn, SJ., Rothman, K. J., Carlin, B., Poole, C., Goodman, N., & Altman, G. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology* 31(4): 337–350.

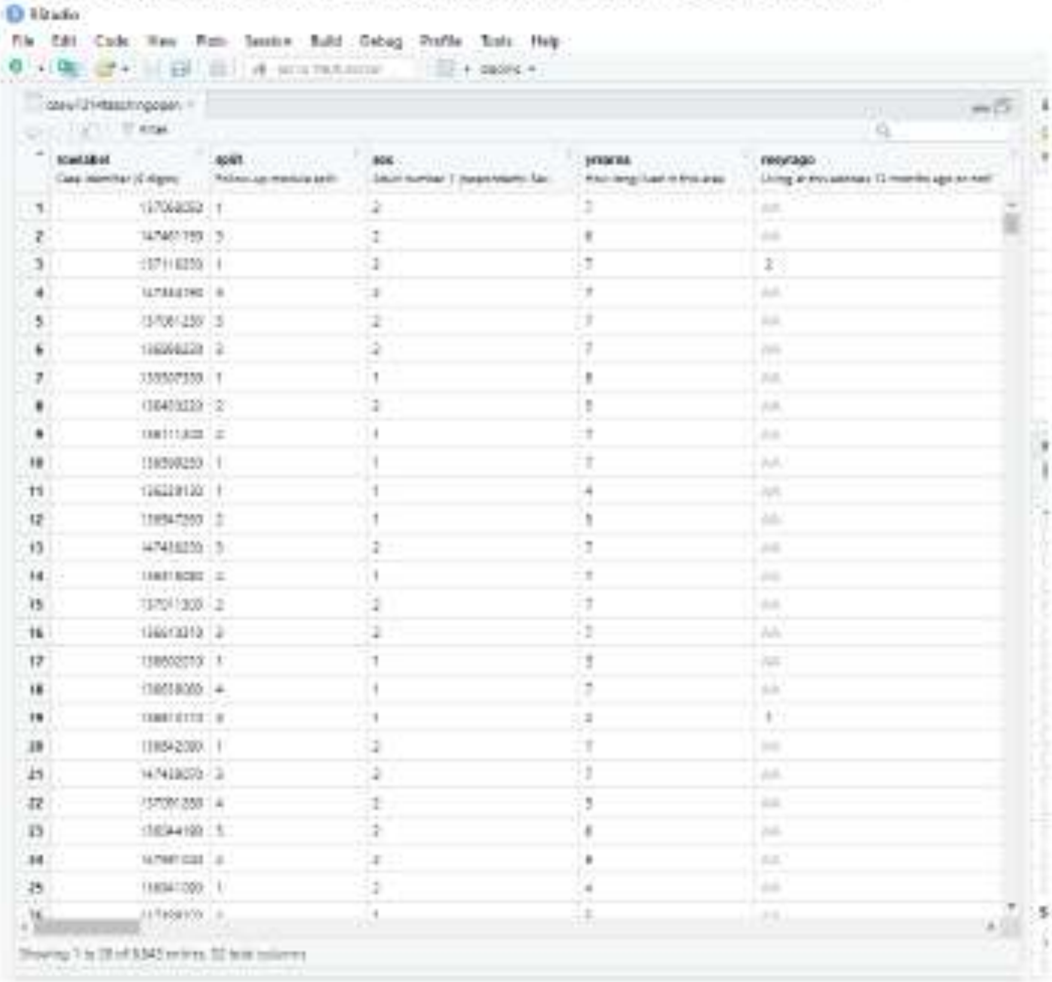
Holmes, A., Illowsky, B., & Dean, S. (2017) Introductory Business Statistics. OpenStax

Hooshyar, D., Pedaste, M, Saks, K., Leijen, A., Bardone, E., Wang, M. (2020) Open learner models in supporting self-regulated learning in higher education: A systematic literature review. *Computers & Education* 154(9): DOI: https://doi.org/10.1016/j.compedu.2020.103878.

8.0 Appendix: Evidence of activities done

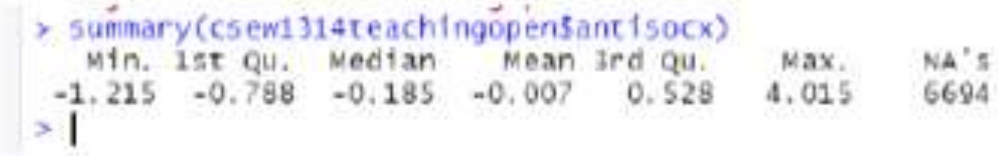
Data Activity 1

Download the Crime Survey for England and Wales, 2013-2014 dataset in R


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
csew1314teachingopen
Filter

	rowlabel Case identifier (6 digits)	split Follow-up module split	sex Adult number 1 (respondent): Sex	yrsarea How long lived in this area	resyrago Living at this address 12 months ago or not
1	[illegible]	1	2	7	NA
2	[illegible]	3	2	8	NA
3	[illegible]	1	2	7	2
4	[illegible]	4	2	7	NA
5	[illegible]	3	2	7	NA
6	[illegible]	2	2	7	NA
7	[illegible]	1	1	8	NA
8	[illegible]	2	2	5	NA
9	[illegible]	2	1	7	NA
10	[illegible]	1	1	7	NA
11	[illegible]	1	1	4	NA
12	[illegible]	2	1	5	NA
13	[illegible]	3	2	7	NA
14	[illegible]	2	1	7	NA
15	[illegible]	2	2	7	NA
16	[illegible]	2	2	7	NA
17	[illegible]	1	1	5	NA
18	[illegible]	4	1	7	NA
19	[illegible]	4	1	3	1
20	[illegible]	1	2	7	NA
21	[illegible]	3	2	7	NA
22	[illegible]	4	2	5	NA
23	[illegible]	5	2	8	NA
24	[illegible]	2	2	8	NA
25	[illegible]	1	2	4	NA

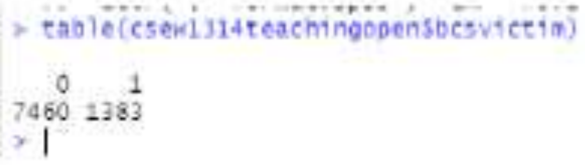
Creating a summary statistic, using the 'antisocx' variable.



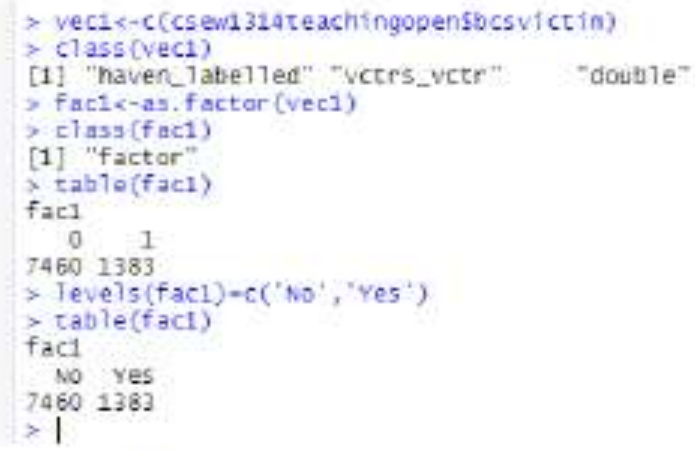
```
> summary(csew1314teachingopen$antisocx)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 -1.215  -0.788  -0.185  -0.007   0.528   4.015    6694 
> |
```

Data activity 2

Creating a frequency table to count if the survey respondents experienced any crime in the previous 12 months, using the table() command. Converting this variable into a factor variable using as_factor.



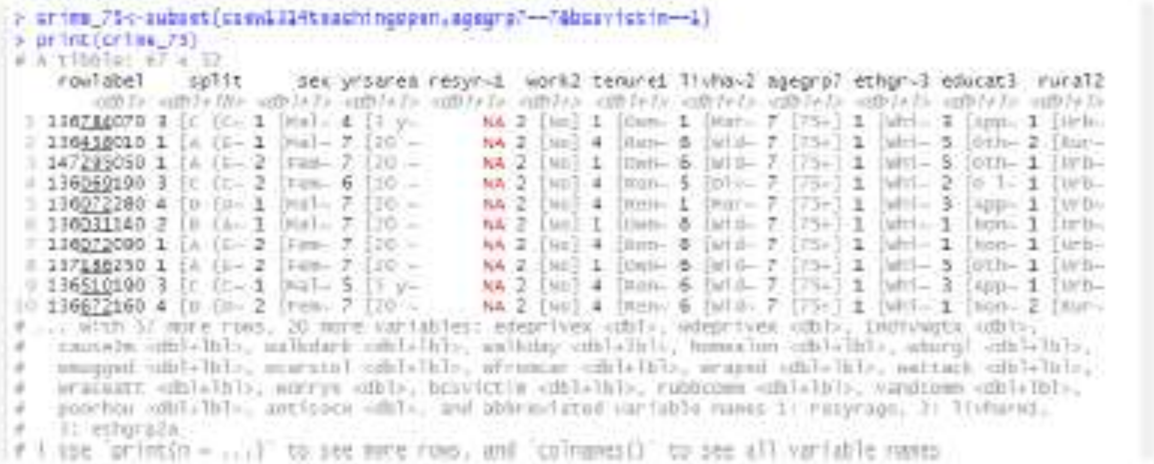
```
> table(csew1314teachingopen$bcsvictim)

   0    1 
7460 1383 
> |
```



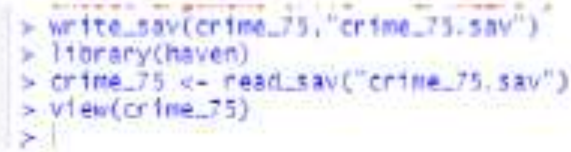
```
> vec1<-c(csew1314teachingopen$bcsvictim)
> class(vec1)
[1] "haven_labelled" "vctrs_vctr"     "double"
> fac1<-as.factor(vec1)
> class(fac1)
[1] "factor"
> table(fac1)
fac1
   0    1
7460 1383
> levels(fac1)=c('No','Yes')
> table(fac1)
fac1
  No  Yes
7460 1383
> |
```

Data activity 3

Creating 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. Saved this dataset under a new name 'crime_75'.



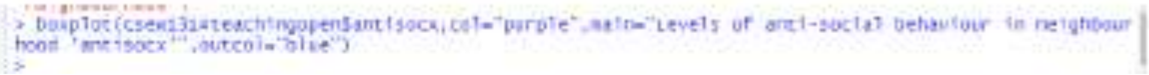
```
> crime_75<-subset(csew1314teachingopen,agegrp7==7&bcsvictim==1)
> print(crime_75)
# A tibble: 67 × 32
    rowlabel    split     sex yrsarea resyr~1  work2 tenure1 livha~2 agegrp7 ethgr~3 educat3  rural2
       <dbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>
 1 136784070 3 [C (C~ 1 [Mal~ 4 [3 y~      NA 2 [No] 1 [Own~ 1 [Mar~ 7 [75+] 1 [Whi~ 3 [App~ 1 [Urb~
 2 136418010 1 [A (E~ 1 [Mal~ 7 [20 ~      NA 2 [No] 4 [Ren~ 6 [Wid~ 7 [75+] 1 [Whi~ 5 [Oth~ 2 [Rur~
 3 147299050 1 [A (E~ 2 [Fem~ 7 [20 ~      NA 2 [No] 1 [Own~ 6 [Wid~ 7 [75+] 1 [Whi~ 5 [Oth~ 1 [Urb~
 4 136069190 3 [C (C~ 2 [Fem~ 6 [10 ~      NA 2 [No] 4 [Ren~ 5 [Div~ 7 [75+] 1 [Whi~ 2 [O l~ 1 [Urb~
 5 136072280 4 [D (D~ 1 [Mal~ 7 [20 ~      NA 2 [No] 4 [Ren~ 1 [Mar~ 7 [75+] 1 [Whi~ 3 [App~ 1 [Urb~
 6 136031140 2 [B (A~ 1 [Mal~ 7 [20 ~      NA 2 [No] 1 [Own~ 6 [Wid~ 7 [75+] 1 [Whi~ 1 [Non~ 1 [Urb~
 7 136072090 1 [A (E~ 2 [Fem~ 7 [20 ~      NA 2 [No] 4 [Ren~ 6 [Wid~ 7 [75+] 1 [Whi~ 1 [Non~ 1 [Urb~
 8 137186250 1 [A (E~ 2 [Fem~ 7 [20 ~      NA 2 [No] 1 [Own~ 6 [Wid~ 7 [75+] 1 [Whi~ 5 [Oth~ 1 [Urb~
 9 136510190 3 [C (C~ 1 [Mal~ 5 [3 y~      NA 2 [No] 4 [Ren~ 6 [Wid~ 7 [75+] 1 [Whi~ 3 [App~ 1 [Urb~
10 136672160 4 [D (D~ 2 [Fem~ 7 [20 ~      NA 2 [No] 4 [Ren~ 6 [Wid~ 7 [75+] 1 [Whi~ 1 [Non~ 2 [Rur~
# … with 57 more rows, 20 more variables: edeprivex <dbl>, wdeprivex <dbl>, indivwgtx <dbl>,
#   causeIm <dbl+lbl>, walkdark <dbl+lbl>, walkday <dbl+lbl>, homealon <dbl+lbl>, wburgl <dbl+lbl>,
#   wmugged <dbl+lbl>, wcarstol <dbl+lbl>, wfromcar <dbl+lbl>, wraped <dbl+lbl>, wattack <dbl+lbl>,
#   wraceatt <dbl+lbl>, worryx <dbl>, bcsvictim <dbl+lbl>, rubbcomm <dbl+lbl>, vandcomm <dbl+lbl>,
#   poorhou <dbl+lbl>, antisocx <dbl>, and abbreviated variable names 1: resyrago, 2: livharm1,
#   3: ethgrp2a
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```



```
> write_sav(crime_75,"crime_75.sav")
> library(haven)
> crime_75 <- read_sav("crime_75.sav")
> view(crime_75)
> |
```

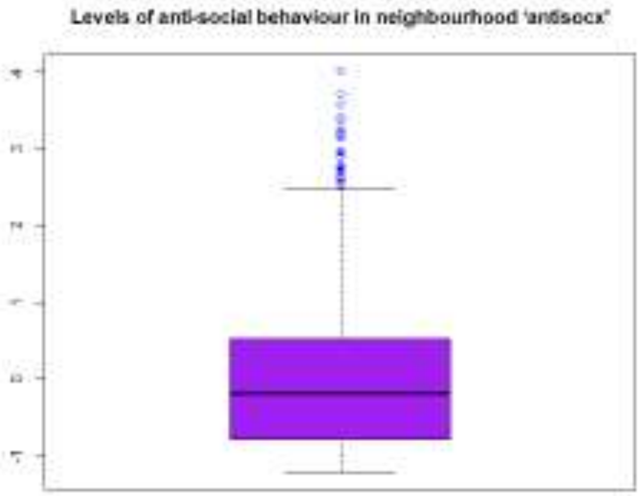
Data activity 4

Creating a boxplot for assessing levels of anti-social behaviour that the survey respondents experience in their neighbourhood using the variable: antisocx.

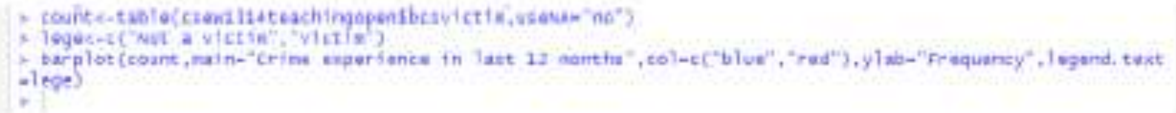


```
> boxplot(csew1314teachingopen$antisocx,col="purple",main="Levels of anti-social behaviour in neighbour
hood 'antisocx'",outcol="blue")
>
```

Boxplot

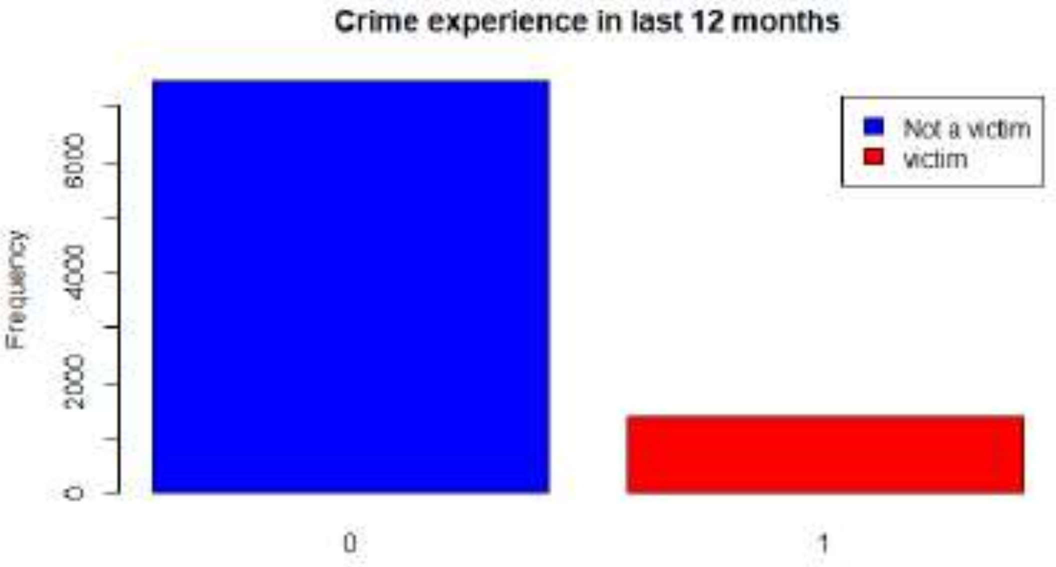

Levels of anti-social behaviour in neighbourhood 'antisocx'

Creating a bar plot using either the barplot() function to assess whether or not the survey respondents experienced crime in the 12 months prior to the survey using the variable 'bcsvictim'



```
> count<-table(csew1114teachingopen$bcsvictim,useNA="no")
> lege<-c("Not a victim","Victim")
> barplot(count,main="Crime experience in last 12 months",col=c("blue","red"),ylab="Frequency",legend.text
=lege)
> |
```

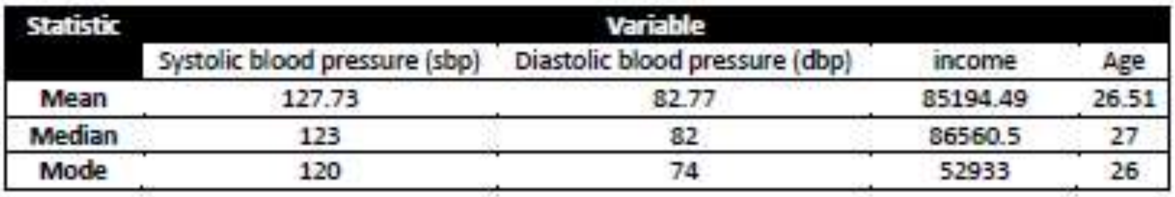
Barplot


Crime experience in last 12 months

Individual Reflection – Module: Numerical Analysis

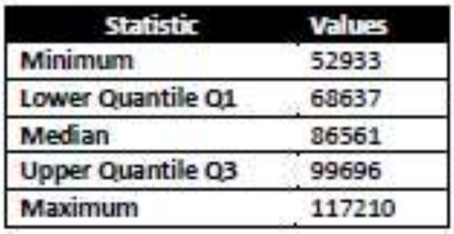
Data Activity 5 and 6

Health_Data: Mean, Median, mode of variables sbp,dbp, income and Age

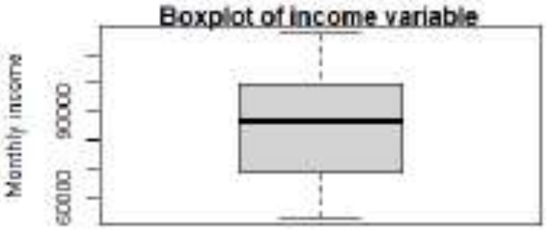


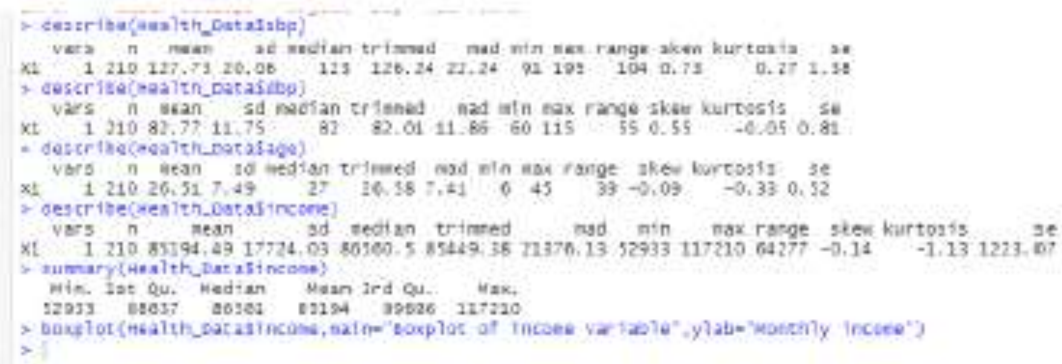
Statistic	Variable			
	Systolic blood pressure (sbp)	Diastolic blood pressure (dbp)	income	Age
Mean	127.73	82.77	85194.49	26.51
Median	123	82	86560.5	27
Mode	120	74	52933	26

Five-figure summary of income variable and present it using a Boxplot.



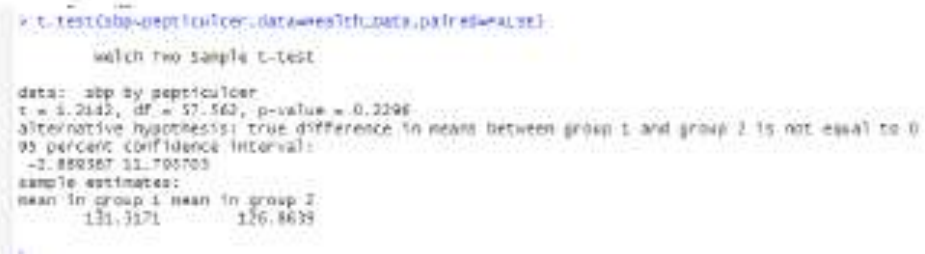
Statistic	Values
Minimum	52933
Lower Quantile Q1	68637
Median	86561
Upper Quantile Q3	99696
Maximum	117210


Boxplot of income variable



```
> describe(Health_Data$sbp)
   vars   n   mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 210 127.73 20.06    123  126.24 22.24  91 195   104 0.73     0.27 1.38
> describe(Health_Data$dbp)
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 210 82.77 11.75     82   82.01 11.86  60 115    55 0.55    -0.05 0.81
> describe(Health_Data$age)
   vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 210 26.51 7.49     27   26.58 7.41   6  45    39 -0.09    -0.33 0.52
> describe(Health_Data$income)
   vars   n     mean       sd  median  trimmed      mad   min    max range  skew kurtosis      se
X1    1 210 85194.49 17724.03 86560.5 85449.38 21376.13 52933 117210 64277 -0.14    -1.13 1223.07
> 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 variable",ylab="Monthly income")
> |
```

Hypothesis test to see if there is any association between systolic blood pressure and presence and absence of peptic ulcer.



```
> t.test(sbp~pepticulcer,data=Health_Data,paired=FALSE)

        Welch Two Sample t-test

data:  sbp by pepticulcer
t = 1.2142, df = 57.562, p-value = 0.2296
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
 -2.888387 11.793703
sample estimates:
mean in group 1 mean in group 2 
       131.3171        126.8639 
```

Interpretation: **Two sample t-test**

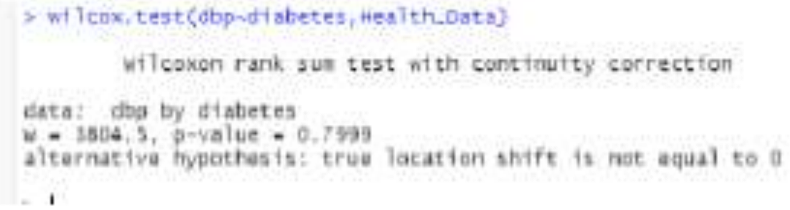
Null hypothesis: True difference in means of the sbp with presence and absence of peptic ulcer is equal to 0

Alternative hypothesis: true difference in means of the sbp with presence and absence of peptic ulcer is not equal to 0

- t is the t-Test statistic value (t=1.2142)
- df is the degrees of freedom (df=57.562)
- p-value is the significance level of the t-Test (p-value= 0.2296)
- Confidence interval of the mean at 95 percent (confidence interval: [-2.89 , 11.80]
- Sample estimates refers to the mean value of the two samples (Mean in men group = 131.32, Mean in women group = 126.86)

The p-value of the test is 0.2296 which is greater than the significance level alpha = 0.05. We conclude that the null hypothesis has failed to be rejected that the true difference in means of the sbp with presence and absence of peptic ulcer groups is equal to 0 and reject the alternative hypothesis.

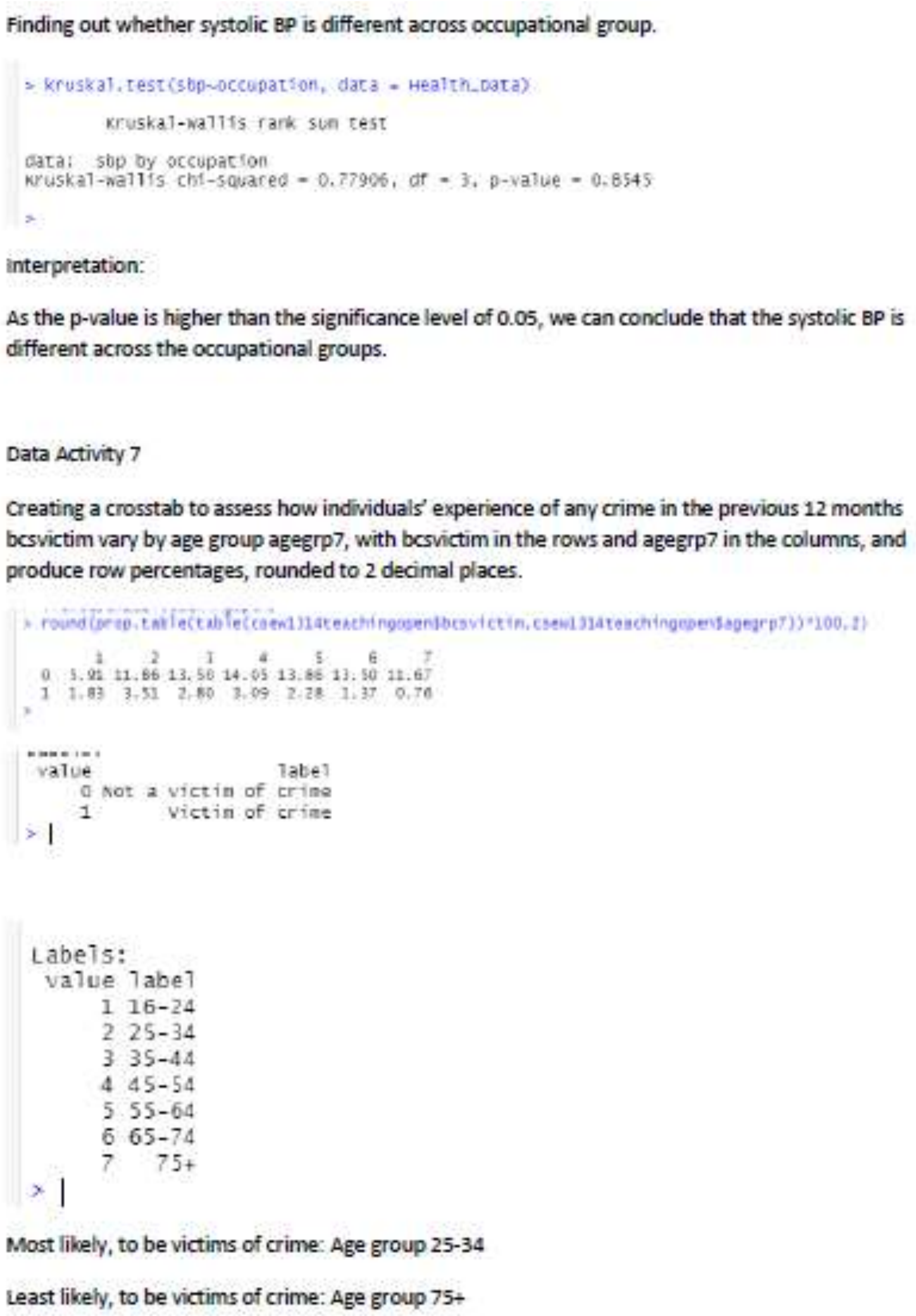
Finding out whether median diastolic blood pressure is same among diabetic and non-diabetic participants.



```
> wilcox.test(dbp~diabetes,Health_Data)

	Wilcoxon rank sum test with continuity correction

data:  dbp by diabetes
W = 3804.5, p-value = 0.7999
alternative hypothesis: true location shift is not equal to 0
```

Interpretation:

Here, p-value is 0.7999 turning out to be greater than 0.05. Hence, the null hypothesis will not be rejected, and this means the median diastolic blood pressure of diabetic and non-diabetic participants are the same.



Finding out whether systolic BP is different across occupational group.

```
> kruskal.test(sbp~occupation, data = Health_Data)

        Kruskal-Wallis rank sum test

data:  sbp by occupation
Kruskal-Wallis chi-squared = 0.77906, df = 3, p-value = 0.8545

>
```

Interpretation:

As the p-value is higher than the significance level of 0.05, we can conclude that the systolic BP is different across the occupational groups.

Data Activity 7

Creating a crosstab to assess how individuals' experience of any crime in the previous 12 months bcsvictim vary by age group agegrp7, with bcsvictim in the rows and agegrp7 in the columns, and produce row percentages, rounded to 2 decimal places.

```
> round(prop.table(table(csew1314teachingopen$bcsvictim,csew1314teachingopen$agegrp7))*100,2)

        1     2     3     4     5     6     7
  0  5.91 11.86 13.50 14.05 13.86 13.50 11.67
  1  1.83  3.51  2.80  3.09  2.28  1.37  0.76
>
```

```
Labels:
 value                label
     0 Not a victim of crime
     1       Victim of crime
> |
```

```
Labels:
 value label
     1 16-24
     2 25-34
     3 35-44
     4 45-54
     5 55-64
     6 65-74
     7   75+
> |
```

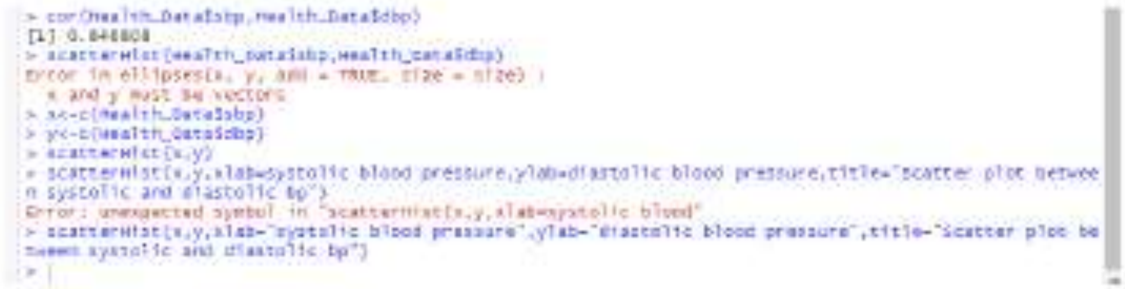
Most likely, to be victims of crime: Age group 25-34

Least likely, to be victims of crime: Age group 75+

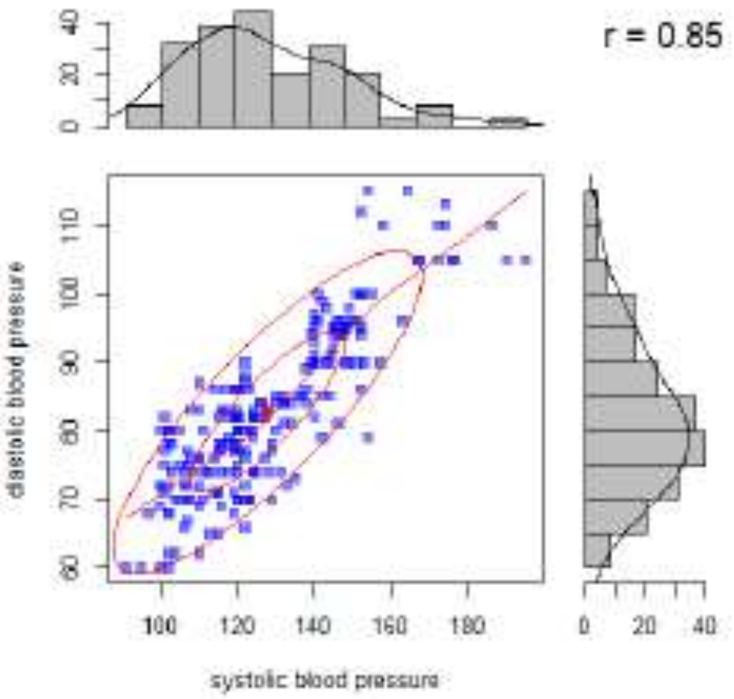
Data Activity 8

Health_Data: correlation between systolic and diastolic BP. r=0.85 Strong positive correlation

Scatter plot between systolic and diastolic BP.



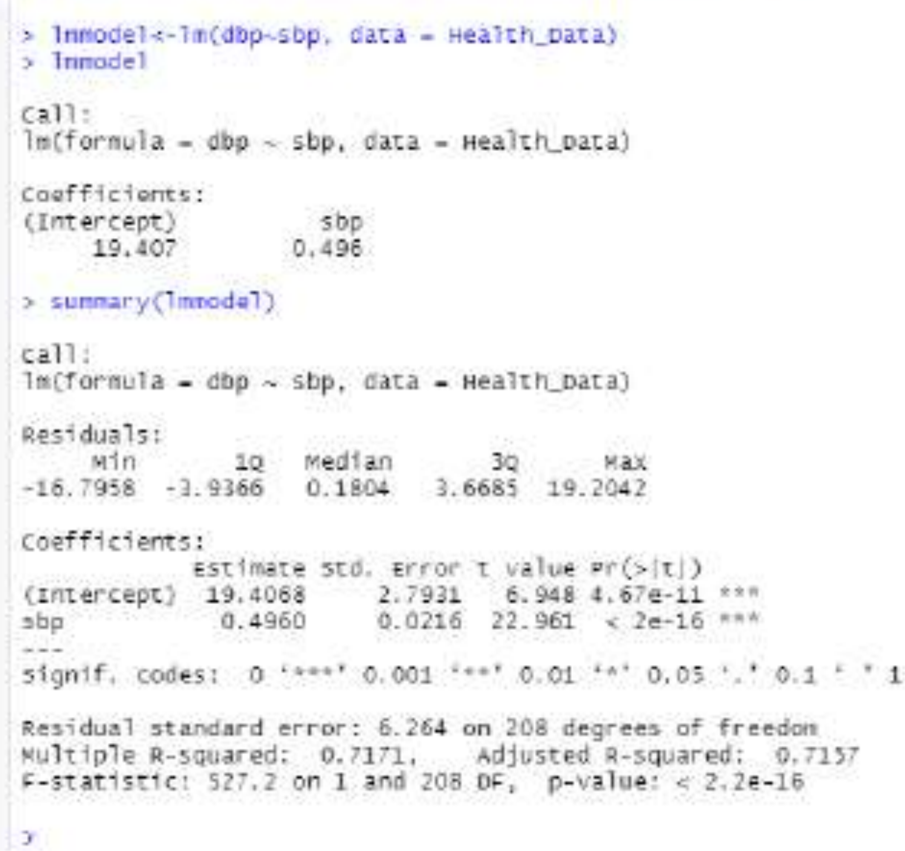
```
> cor(Health_Data$sbp,Health_Data$dbp)
[1] 0.846808
> scatterHist(Health_Data$sbp,Health_Data$dbp)
Error in ellipses(x, y, add = TRUE, size = size) :
  x and y must be vectors
> x<-c(Health_Data$sbp)
> y<-c(Health_Data$dbp)
> scatterHist(x,y)
> scatterHist(x,y,xlab=systolic blood pressure,ylab=diastolic blood pressure,title="scatter plot betwee
n systolic and diastolic bp")
Error: unexpected symbol in "scatterHist(x,y,xlab=systolic blood"
> scatterHist(x,y,xlab="systolic blood pressure",ylab="diastolic blood pressure",title="scatter plot be
tween systolic and diastolic bp")
> |
```



Data Activity 9

Health_Data: Simple linear regression analysis to find the population regression equation to predict the diastolic BP by systolic BP.

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



```
> lnmodel<-lm(dbp~sbp, data = Health_Data)
> lnmodel

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

Coefficients:
(Intercept)          sbp
     19.407        0.496

> summary(lnmodel)

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

>
```

The above results show the intercept and the beta coefficient for the sbp variable and the estimated regression line equation can be written as follow:

dbp =19.407 + 0.496*(sbp)

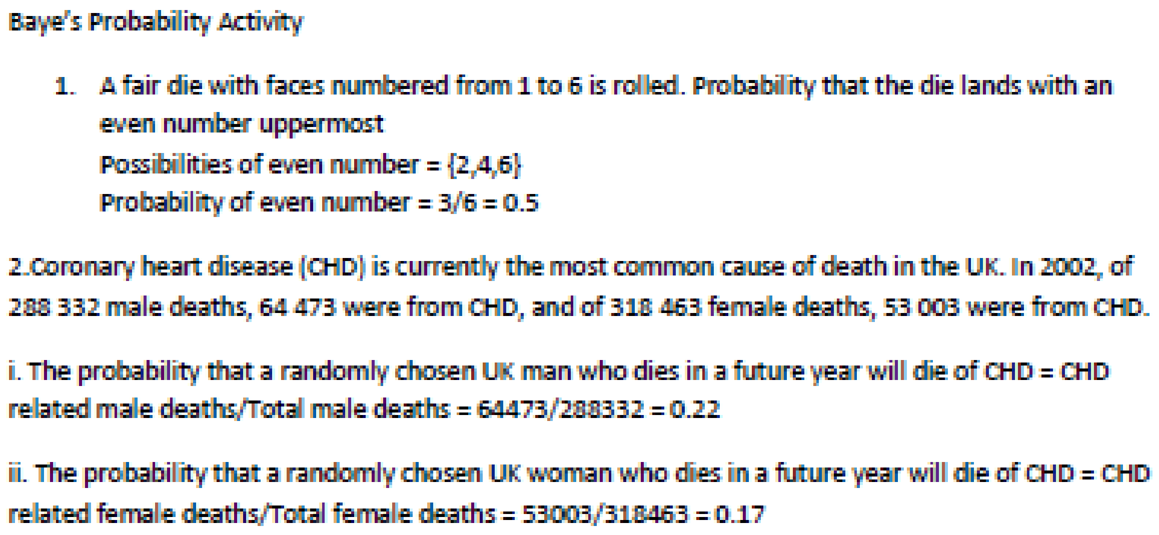
This can be interpretated as, for a sbp equal to 120, the dbp will be

19.407+0.496*120 = 78.93

Interpretation

In the above example, both the p-values for the intercept and the predictor variable are highly significant. So, we can reject the null hypothesis and accept the alternative hypothesis. This means that there is a significant association between the predictor (sbp) and the outcome (dbp) variables.

Individual Reflection – Module: Numerical Analysis



Baye's Probability Activity

1. A fair die with faces numbered from 1 to 6 is rolled. Probability that the die lands with an even number uppermost
   Possibilities of even number = {2,4,6}
   Probability of even number = 3/6 = 0.5

2.Coronary heart disease (CHD) is currently the most common cause of death in the UK. In 2002, of 288 332 male deaths, 64 473 were from CHD, and of 318 463 female deaths, 53 003 were from CHD.

i. The probability that a randomly chosen UK man who dies in a future year will die of CHD = CHD related male deaths/Total male deaths = 64473/288332 = 0.22

ii. The probability that a randomly chosen UK woman who dies in a future year will die of CHD = CHD related female deaths/Total female deaths = 53003/318463 = 0.17