Individual Reflection

Activities within the module

Activities included lecturecasts, seminars, reading textbooks and academic papers, videos, and practical activities.

I initially found the lecturecasts to be simplistic and didn't add much value. However, when I encountered units without a lecturecast I realised that their value was providing structure. I came to value that structure in the remote learning environment. I learned to use lecturecasts to introduce topics whilst leaving the detailed learning to the reading and activities.

The practical activities were all performed using R, which I had not used before. They covered many elements of data analysis, starting with installing and running the R software, data structures, and importing and exporting data files. They moved on to probabilities, basic arithmetic operations and converting and managing variables before getting into more complex statistical analysis examining null and alternative hypotheses and visualising results with plots and charts.

I learned when to use different tests such as the parametric t-test and nonparametric Mann-Whitney U test and Chi-square depending upon whether the data are quantitative or categorical, whether the distribution is normal, and the sample size. Comparing the resultant p-value with a previously defined alpha value determines if there is a statistical difference in the data meaning the null hypothesis can be rejected.

I learned how to test for normality to select the correct test, including Shaprio-Wilk for small sample sizes and Kolmogorov-Smirnov for large sample sizes where a p-value greater than 0.05 indicated a normal distribution. Visual clues were also examined in histograms and Q-Q plots.

Correlations between variables were tested using Pearson's and Spearman's Rho, using a p-value to show the statistical significance of the correlation alongside an r-value showing the strength of the correlation.

Linear regression and Bayesian data analysis were also discussed and practiced.

Working autonomously

There were 9 data activities which provided ample opportunity to practice what was being learned. This helped to re-enforce the learning, consistent with goal-based scenarios where "students pursue a goal by practicing target skills and using relevant content knowledge to help them achieve their goal" (Schank et al, 1999: 165). The presentation assessment was another excellent example of learning in pursuit of a goal.

At the end of each week I published a summary of my thoughts on the unit plus my results of the data activities in my e-portfolio for others to learn from if needed. Reflecting for one hour after a challenge significantly increases learning (Daudelin, 1996) so updating my e-portfolio at the end of every week ensured that I embedded the knowledge through regular reflection.

Appendix A contains screenshots of my e-portfolio reflections and Appendix B contains screenshots of my e-portfolio data activities results.

Most of the activities were quite simple, but often needed extra research to complete. I found that ggplot in R was particularly useful, and gave significantly better visuals than standard R functions, so when an activity required a simple visual to be produced I usually repeated it using ggplot to practice.

There were also two collaborative discussions. I contributed to both, which was a good way to further learn, with Laal et al (2013) noting that collaborative learning is more productive than individual learning. Unfortunately, only one other student contributed to the discussion so whilst it was useful to pass ideas between us, it wasn't as good as it could have been if more students had contributed.

I am also in a WhatsApp group with some of the other students and I gave advice there, for example sharing ideas about which statistical tests to use for the statistical analysis presentation, without giving specific answers of course.

Emotional response

I was concerned about the exam so I used Gibbs' reflective cycle (Gibbs, 1998) to analyse my approach.

Description: The module had an exam and I had not performed complex maths for a long time.

Feelings: I was feeling nervous about my lack of recent experience and had some doubts about my ability to pass the exam.

Evaluation: I shared my nervousness with the tutor in the first seminar, with students in the WhatsApp group, and in my e-portfolio. The tutor reassured me that it would not be too difficult, and other students said we would all support each other, which was encouraging and made me more confident. I resolved to make time to stay ahead of all reading and data activities to ensure that I didn’t fall behind. I attended all except one seminar, which I watched afterwards.

Analysis: The proactive “get it done” strategy worked and I achieved an excellent mark on the exam. Leading up to the exam I even gave help and advice to other students. The emotional support from students at the beginning was important in giving me the confidence to study, even though I didn’t need any help in the end.

Action plan: Exposing my feelings and getting emotional support gave me the confidence to study hard and achieve an excellent result. When faced with a similar challenge of confidence I would do the same again.

Skills, knowledge, learning, and changed actions

I have developed new skills in statistical analysis, including how to apply and present it in R. The evidence of this learning is demonstrated in the data activities from my e-portfolio in Appendix B. I found the data activities were where I learned most, which is consistent with Schank et al (1999) who advocate “learning by doing”.

The collaborative discussions that I contributed to are further evidence of the practical application of what I’ve learned in this module. The first pushed my skills in producing visuals in R, with an interesting discussion between myself and another student about different ways to use a stacked chart to present a table of data. The second focused on the misuse of statistical tools such as p-value. It was relevant and interesting to reflect upon how we are taught to apply things like p-value and confidence intervals whilst considering what they really mean and the subtleties of not always taking them as binary “yes/no” but rather what they tell us about the data.

I feel confident applying statistical analysis using R in subsequent modules during this MSc, and also back at work when required.

There were opportunities to practice duplicating the R activities in Python. I elected to focus entirely on R. The risk with that approach is I’m probably going to have to re-learn some methods in Python in future modules, but I accepted that trade-off to ensure I had a deep understanding of the statistical methods and their application in a single application, in this case, R.

Word count: 1,095

References

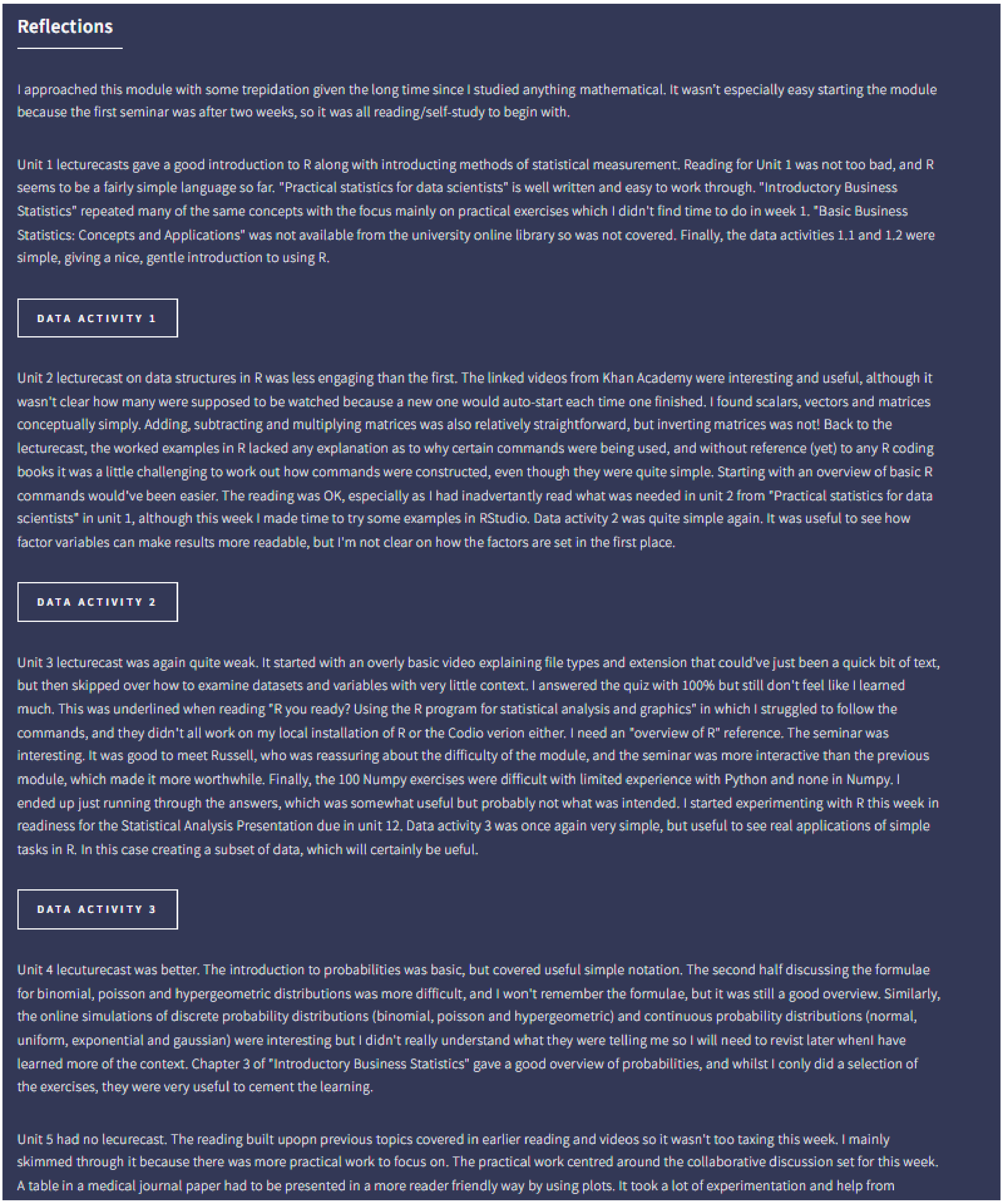
Daudelin, M.W. (1996) Learning from experience through reflection. *Organizational Dynamics*, 24(3): 36-48. https://doi.org/10.1016/S0090-2616(96)90004-2

Gibbs, G (1988) Learning by doing: a guide to teaching and learning methods. Oxford: Further Education Unit.

Laal, M., Naseri, A.S., Laal, M. & Khattami-Kermanshahi, Z. (2013) What do we achieve from learning in collaboration?. *Procedia-Social and Behavioral Sciences*, 93:1427-1432. https://doi.org/10.1016/j.sbspro.2013.10.057

Schank, R.C., Berman, T.R. and Macpherson, K.A. (1999) ‘Learning by doing’ In: Reigeluth, C.M. (ed). *Instructional-Design Theories and Models: A New Paradigm of Instructional Theory Volume 2*. New York: Routledge

Appendix A - Reflections



Reflections

I approached this module with some trepidation given the long time since I studied anything mathematical. It wasn't especially easy starting the module because the first seminar was after two weeks, so it was all reading/self-study to begin with.

Unit 1 lecturecasts gave a good introduction to R along with introducting methods of statistical measurement. Reading for Unit 1 was not too bad, and R seems to be a fairly simple language so far. "Practical statistics for data scientists" is well written and easy to work through. "Introductory Business Statistics" repeated many of the same concepts with the focus mainly on practical exercises which I didn't find time to do in week 1. "Basic Business Statistics: Concepts and Applications" was not available from the university online library so was not covered. Finally, the data activities 1.1 and 1.2 were simple, giving a nice, gentle introduction to using R.

DATA ACTIVITY 1

Unit 2 lecturecast on data structures in R was less engaging than the first. The linked videos from Khan Academy were interesting and useful, although it wasn't clear how many were supposed to be watched because a new one would auto-start each time one finished. I found scalars, vectors and matrices conceptually simply. Adding, subtracting and multiplying matrices was also relatively straightforward, but inverting matrices was not! Back to the lecturecast, the worked examples in R lacked any explanation as to why certain commands were being used, and without reference (yet) to any R coding books it was a little challenging to work out how commands were constructed, even though they were quite simple. Starting with an overview of basic R commands would've been easier. The reading was OK, especially as I had inadvertantly read what was needed in unit 2 from "Practical statistics for data scientists" in unit 1, although this week I made time to try some examples in RStudio. Data activity 2 was quite simple again. It was useful to see how factor variables can make results more readable, but I'm not clear on how the factors are set in the first place.

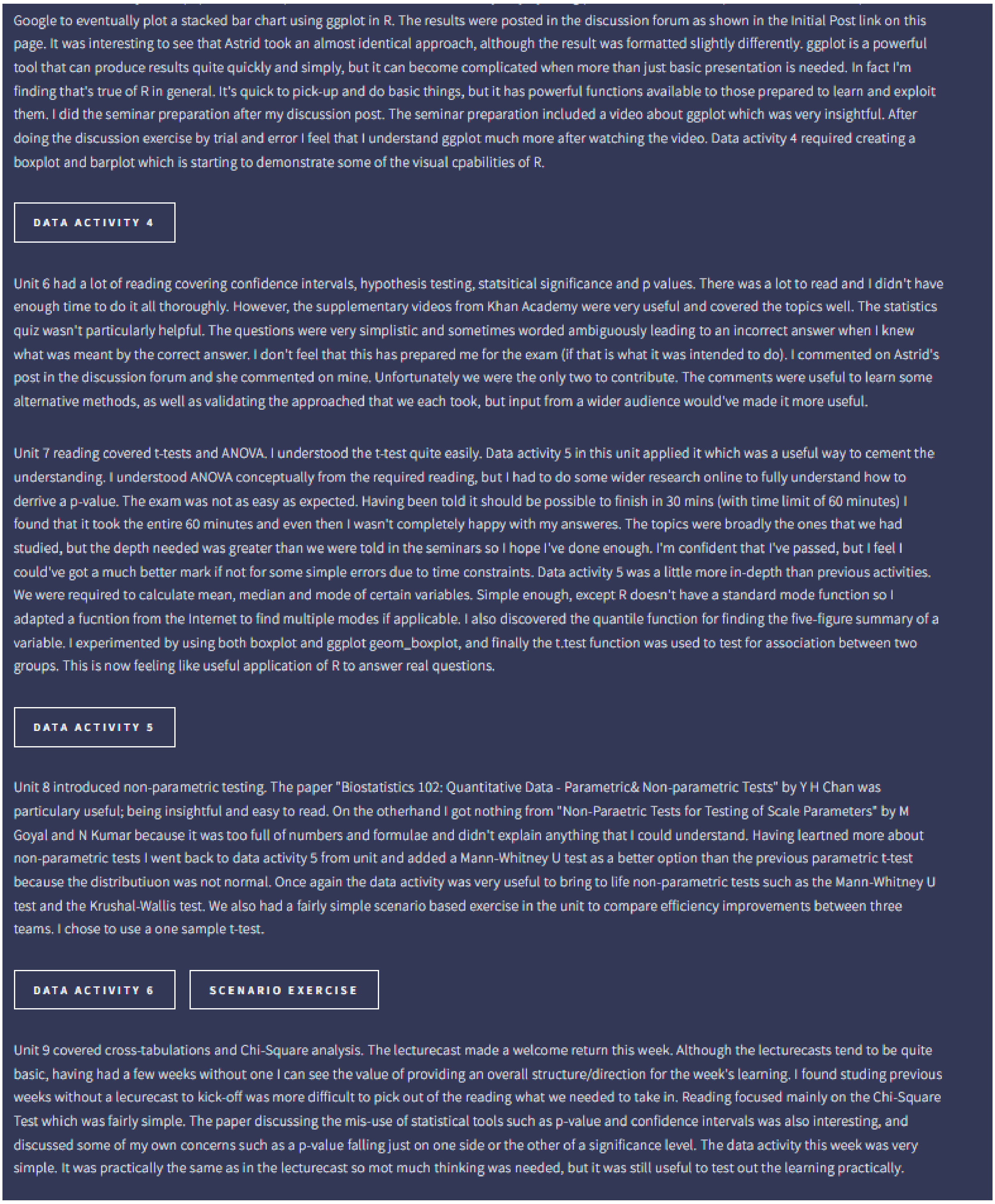
DATA ACTIVITY 2

Unit 3 lecturecast was again quite weak. It started with an overly basic video explaining file types and extension that could've just been a quick bit of text, but then skipped over how to examine datasets and variables with very little context. I answered the quiz with 100% but still don't feel like I learned much. This was underlined when reading "R you ready? Using the R program for statistical analysis and graphics" in which I struggled to follow the commands, and they didn't all work on my local installation of R or the Codio verion either. I need an "overview of R" reference. The seminar was interesting. It was good to meet Russell, who was reassuring about the difficulty of the module, and the seminar was more interactive than the previous module, which made it more worthwhile. Finally, the 100 Numpy exercises were difficult with limited experience with Python and none in Numpy. I ended up just running through the answers, which was somewhat useful but probably not what was intended. I started experimenting with R this week in readiness for the Statistical Analysis Presentation due in unit 12. Data activity 3 was once again very simple, but useful to see real applications of simple tasks in R. In this case creating a subset of data, which will certainly be ueful.

DATA ACTIVITY 3

Unit 4 lecuturecast was better. The introduction to probabilities was basic, but covered useful simple notation. The second half discussing the formulae for binomial, poisson and hypergeometric distributions was more difficult, and I won't remember the formulae, but it was still a good overview. Similarly, the online simulations of discrete probability distributions (binomial, poisson and hypergeometric) and continuous probability distributions (normal, uniform, exponential and gaussian) were interesting but I didn't really understand what they were telling me so I will need to revist later whenI have learned more of the context. Chapter 3 of "Introductory Business Statistics" gave a good overview of probabilities, and whilst I conly did a selection of the exercises, they were very useful to cement the learning.

Unit 5 had no lecurecast. The reading built upopn previous topics covered in earlier reading and videos so it wasn't too taxing this week. I mainly skimmed through it because there was more practical work to focus on. The practical work centred around the collaborative discussion set for this week. A table in a medical journal paper had to be presented in a more reader friendly way by using plots. It took a lot of experimentation and help from



Google to eventually plot a stacked bar chart using ggplot in R. The results were posted in the discussion forum as shown in the Initial Post link on this page. It was interesting to see that Astrid took an almost identical approach, although the result was formatted slightly differently. ggplot is a powerful tool that can produce results quite quickly and simply, but it can become complicated when more than just basic presentation is needed. In fact I'm finding that's true of R in general. It's quick to pick-up and do basic things, but it has powerful functions available to those prepared to learn and exploit them. I did the seminar preparation after my discussion post. The seminar preparation included a video about ggplot which was very insightful. After doing the discussion exercise by trial and error I feel that I understand ggplot much more after watching the video. Data activity 4 required creating a boxplot and barplot which is starting to demonstrate some of the visual cpabilities of R.

DATA ACTIVITY 4

Unit 6 had a lot of reading covering confidence intervals, hypothesis testing, statsitical significance and p values. There was a lot to read and I didn't have enough time to do it all thoroughly. However, the supplementary videos from Khan Academy were very useful and covered the topics well. The statistics quiz wasn't particularly helpful. The questions were very simplistic and sometimes worded ambiguously leading to an incorrect answer when I knew what was meant by the correct answer. I don't feel that this has prepared me for the exam (if that is what it was intended to do). I commented on Astrid's post in the discussion forum and she commented on mine. Unfortunately we were the only two to contribute. The comments were useful to learn some alternative methods, as well as validating the approached that we each took, but input from a wider audience would've made it more useful.

Unit 7 reading covered t-tests and ANOVA. I understood the t-test quite easily. Data activity 5 in this unit applied it which was a useful way to cement the understanding. I understood ANOVA conceptually from the required reading, but I had to do some wider research online to fully understand how to derrive a p-value. The exam was not as easy as expected. Having been told it should be possible to finish in 30 mins (with time limit of 60 minutes) I found that it took the entire 60 minutes and even then I wasn't completely happy with my answeres. The topics were broadly the ones that we had studied, but the depth needed was greater than we were told in the seminars so I hope I've done enough. I'm confident that I've passed, but I feel I could've got a much better mark if not for some simple errors due to time constraints. Data activity 5 was a little more in-depth than previous activities. We were required to calculate mean, median and mode of certain variables. Simple enough, except R doesn't have a standard mode function so I adapted a fucntion from the Internet to find multiple modes if applicable. I also discovered the quantile function for finding the five-figure summary of a variable. I experimented by using both boxplot and ggplot geom_boxplot, and finally the t.test function was used to test for association between two groups. This is now feeling like useful application of R to answer real questions.

DATA ACTIVITY 5

Unit 8 introduced non-parametric testing. The paper "Biostatistics 102: Quantitative Data - Parametric& Non-parametric Tests" by Y H Chan was particulary useful; being insightful and easy to read. On the otherhand I got nothing from "Non-Paraetric Tests for Testing of Scale Parameters" by M Goyal and N Kumar because it was too full of numbers and formulae and didn't explain anything that I could understand. Having leartned more about non-parametric tests I went back to data activity 5 from unit and added a Mann-Whitney U test as a better option than the previous parametric t-test because the distributiuon was not normal. Once again the data activity was very useful to bring to life non-parametric tests such as the Mann-Whitney U test and the Krushal-Wallis test. We also had a fairly simple scenario based exercise in the unit to compare efficiency improvements between three teams. I chose to use a one sample t-test.

DATA ACTIVITY 6 SCENARIO EXERCISE

Unit 9 covered cross-tabulations and Chi-Square analysis. The lecturecast made a welcome return this week. Although the lecturecasts tend to be quite basic, having had a few weeks without one I can see the value of providing an overall structure/direction for the week's learning. I found studing previous weeks without a lecurecast to kick-off was more difficult to pick out of the reading what we needed to take in. Reading focused mainly on the Chi-Square Test which was fairly simple. The paper discussing the mis-use of statistical tools such as p-value and confidence intervals was also interesting, and discussed some of my own concerns such as a p-value falling just on one side or the other of a significance level. The data activity this week was very simple. It was practically the same as in the lecturecast so mot much thinking was needed, but it was still useful to test out the learning practically.

DATA ACTIVITY 7

Unit 10 covered correlation. The lecturecast gave a useful overview. It was generally quite simple, with probably the easiest summary quiz so far. However, the example calculations were unecessary given they were not worked through. The demonstration of how to do regression analysis in Excel was a very nice inclusion though. The data activity was a simple correlation test that was then visualised using a scatterplot. It was simple, but once agail useful to experiment.

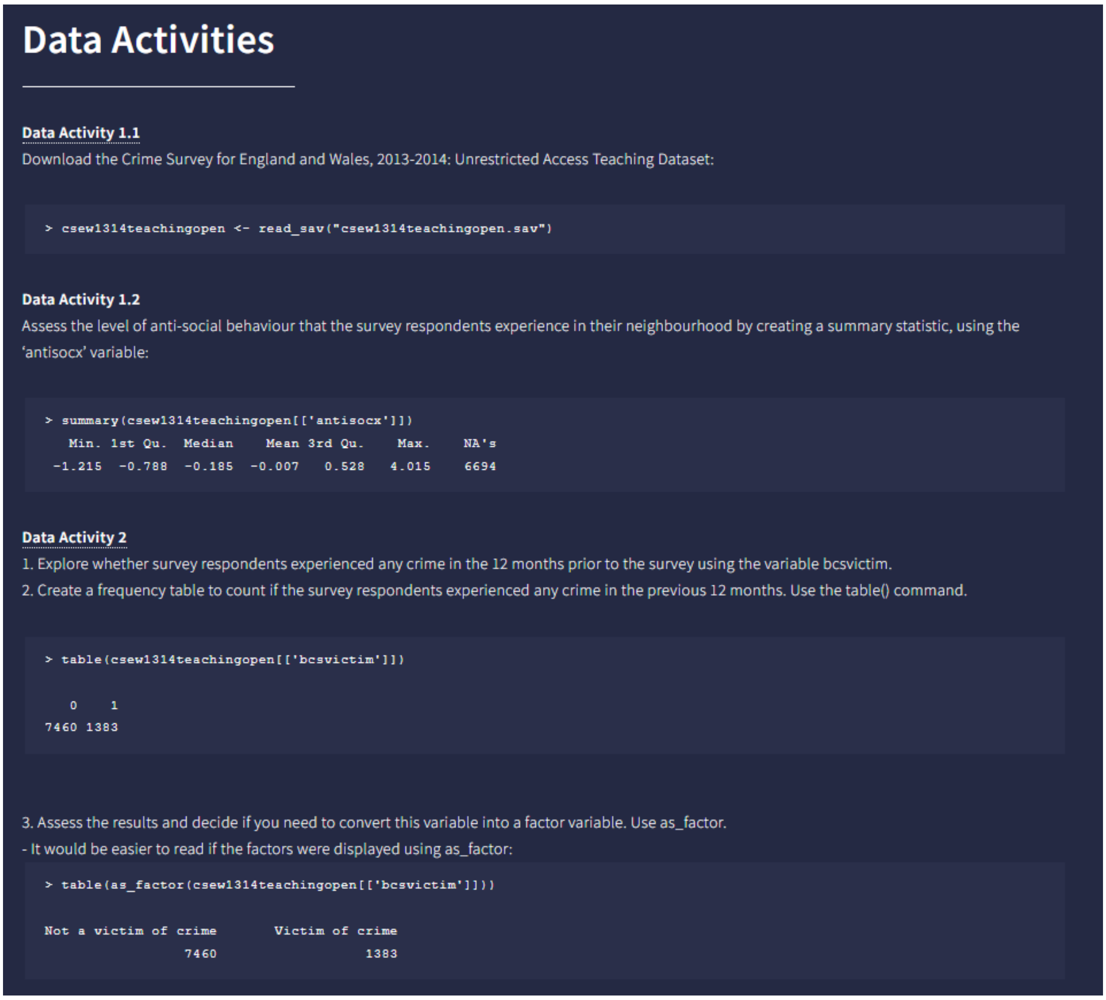
DATA ACTIVITY 8

Unit 11 covered regression analysis. There was no lecturecast this week. The reading was fairly straightforward, although the formulae were a little more complicated than previous units. As has been the case throughout the module, Practical Statistics for Data Scientists by Bruce et al has an excellent balance of theory and practical application. The data activity was a simple regression analysis and interpretation of the findings.

DATA ACTIVITY 9

Unit 12 covered Bayesian data analysis. The lecturecast was overly simplistic and didn't really describe the topic very well. The quiz was also broken because it required statements to be paired with either Bayesian or Frequentist, but it registered as incorrect if the correct instance of each wasn't selected. E.g. if a statement was matched with the second Frequentist label but it was expecting the first it was marked as incorrect, which is clearly not right. I ended up skipping the quiz because of this. The reading was a little complicated, but mainly because it felt out of context with the rest of the module. There was no data activity this week but there was a formative activity on Baye's probability, except it wasn't. The exercise just covered very simple probabilities and nothing to do with Baye's, which is about relative probabilities. This final unit was a dissapointing end to the module. I do, however, feel that I've learned a lot about statistical analysis and especially how to apply it in R during this module overall.

Appendix B – Data Activities



Data Activities

Data Activity 1.1

Download the Crime Survey for England and Wales, 2013-2014: Unrestricted Access Teaching Dataset:

```
> csew1314teachingopen <- read_sav("csew1314teachingopen.sav")
```

Data Activity 1.2

Assess the level of anti-social behaviour that the survey respondents experience in their neighbourhood by 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

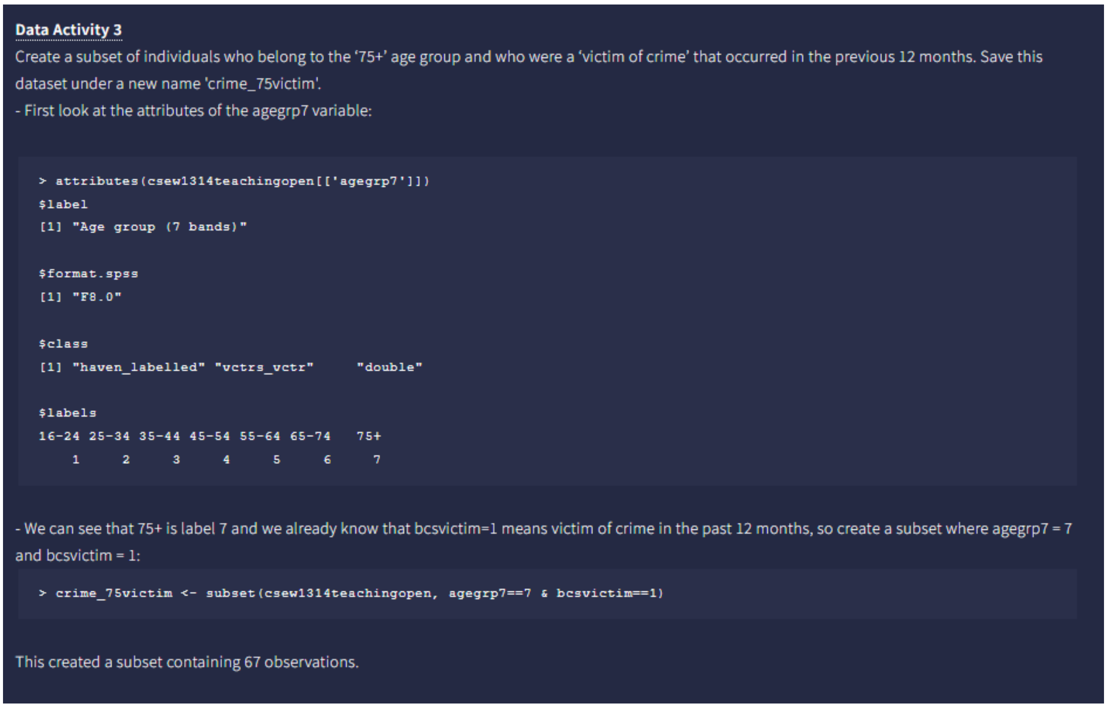
1. Explore whether survey respondents experienced any crime in the 12 months prior to the survey using the variable bcsvictim.
2. Create a frequency table to count if the survey respondents experienced any crime in the previous 12 months. Use the table() command.

```
> table(csew1314teachingopen[['bcsvictim']])

   0    1 
7460 1383 
```

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

- It would be easier to read if the factors were displayed using as_factor:

```
> table(as_factor(csew1314teachingopen[['bcsvictim']]))

Not a victim of crime       Victim of crime 
                 7460                  1383 
```



Data Activity 3

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'.

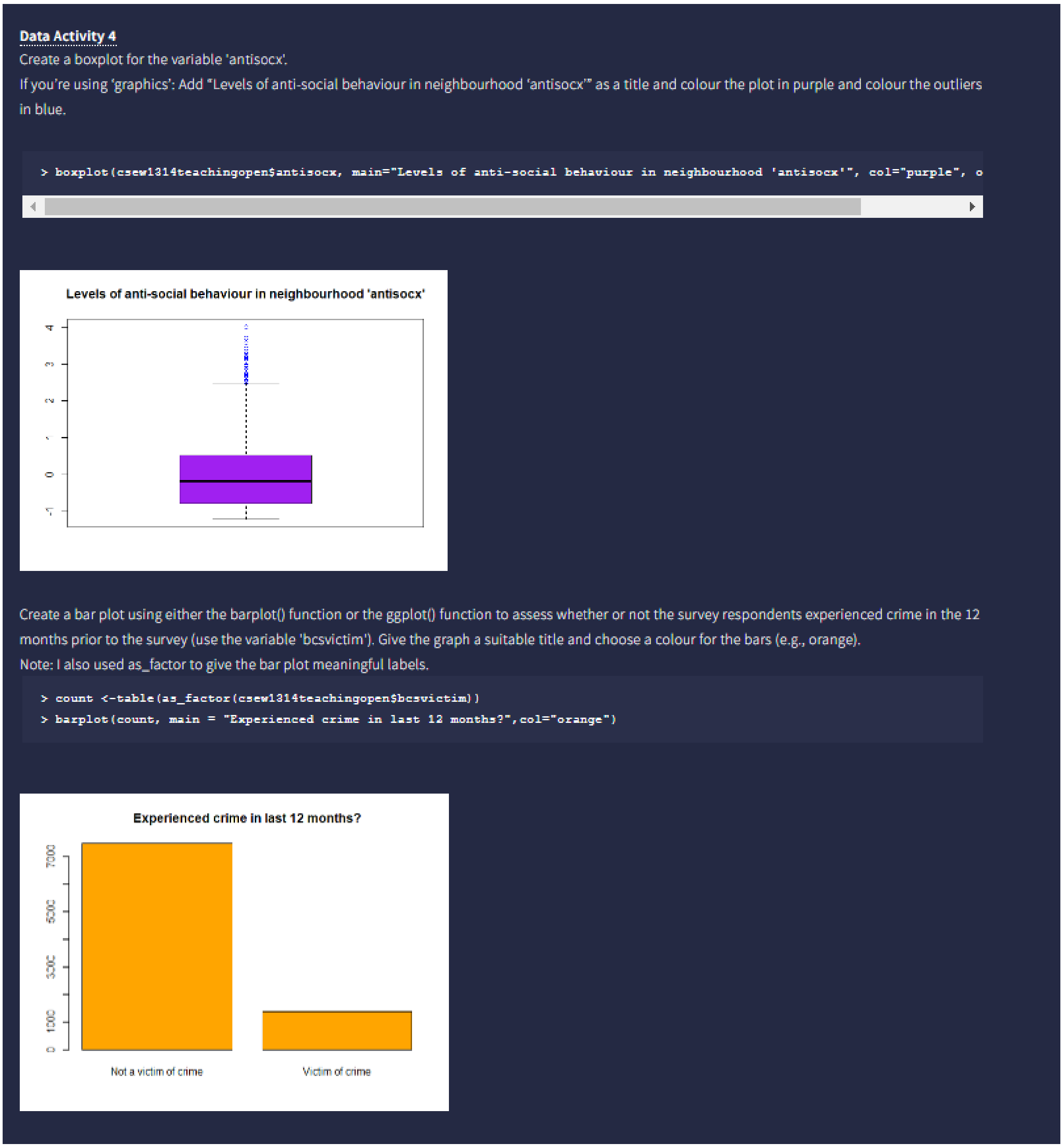
- First look at the attributes of the agegrp7 variable:

```
> attributes(csew1314teachingopen[['agegrp7']])
$label
[1] "Age group (7 bands)"

$format.spss
[1] "F8.0"

$class
[1] "haven_labelled" "vctrs_vctr"     "double"        

$labels
16-24 25-34 35-44 45-54 55-64 65-74   75+ 
    1     2     3     4     5     6     7 
```

- We can see that 75+ is label 7 and we already know that bcsvictim=1 means victim of crime in the past 12 months, so create a subset where agegrp7 = 7 and bcsvictim = 1:

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

This created a subset containing 67 observations.



Data Activity 4

Create a boxplot for the variable 'antisocx'.

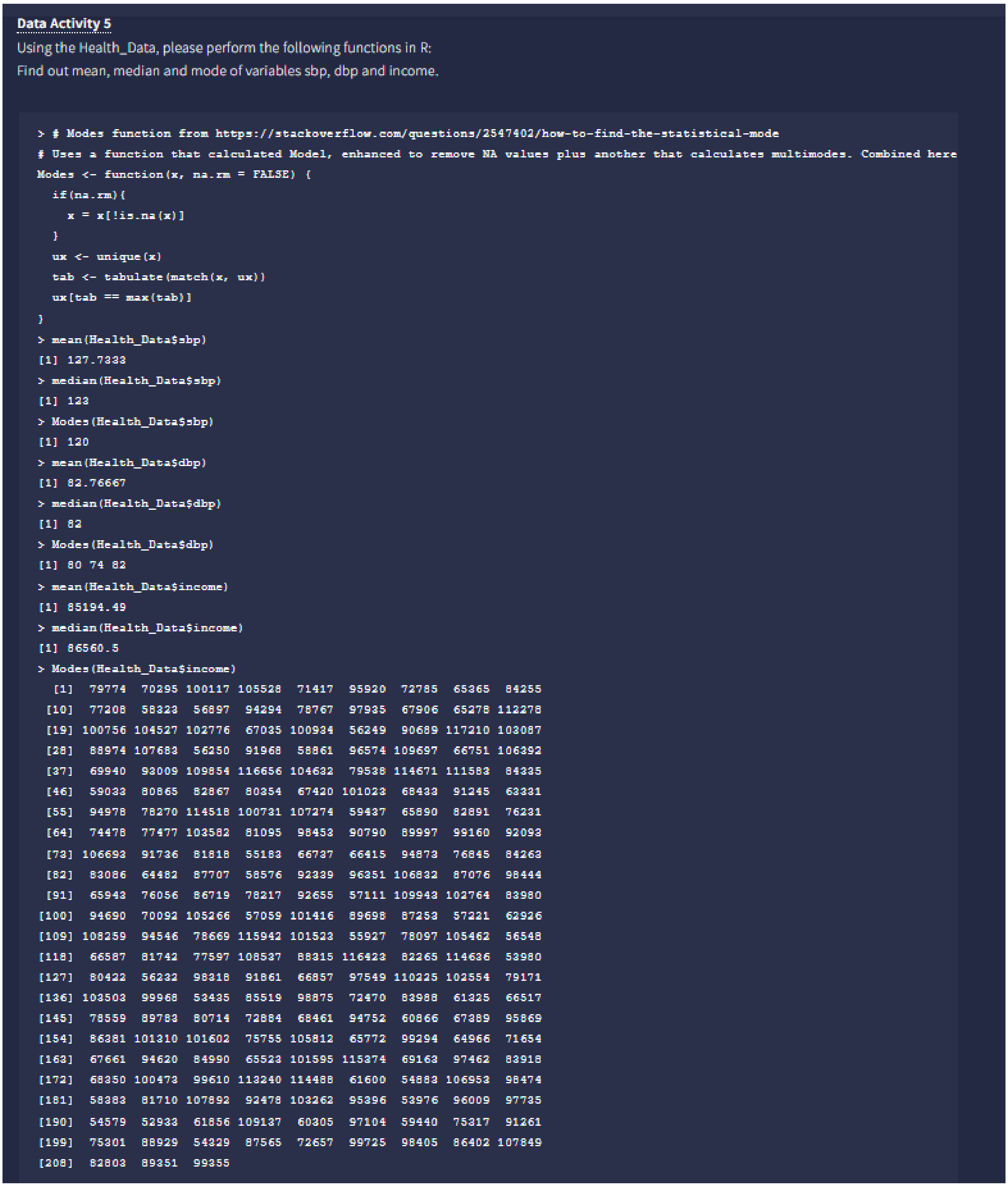
If you're using 'graphics': Add "Levels of anti-social behaviour in neighbourhood 'antisocx'" as a title and colour the plot in purple and colour the outliers in blue.

```
> boxplot(csew1314teachingopen$antisocx, main="Levels of anti-social behaviour in neighbourhood 'antisocx'", col="purple", o
```

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).

Note: I also used as_factor to give the bar plot meaningful labels.

```
> count <-table(as_factor(csew1314teachingopen$bcsvictim))
> barplot(count, main = "Experienced crime in last 12 months?",col="orange")
```

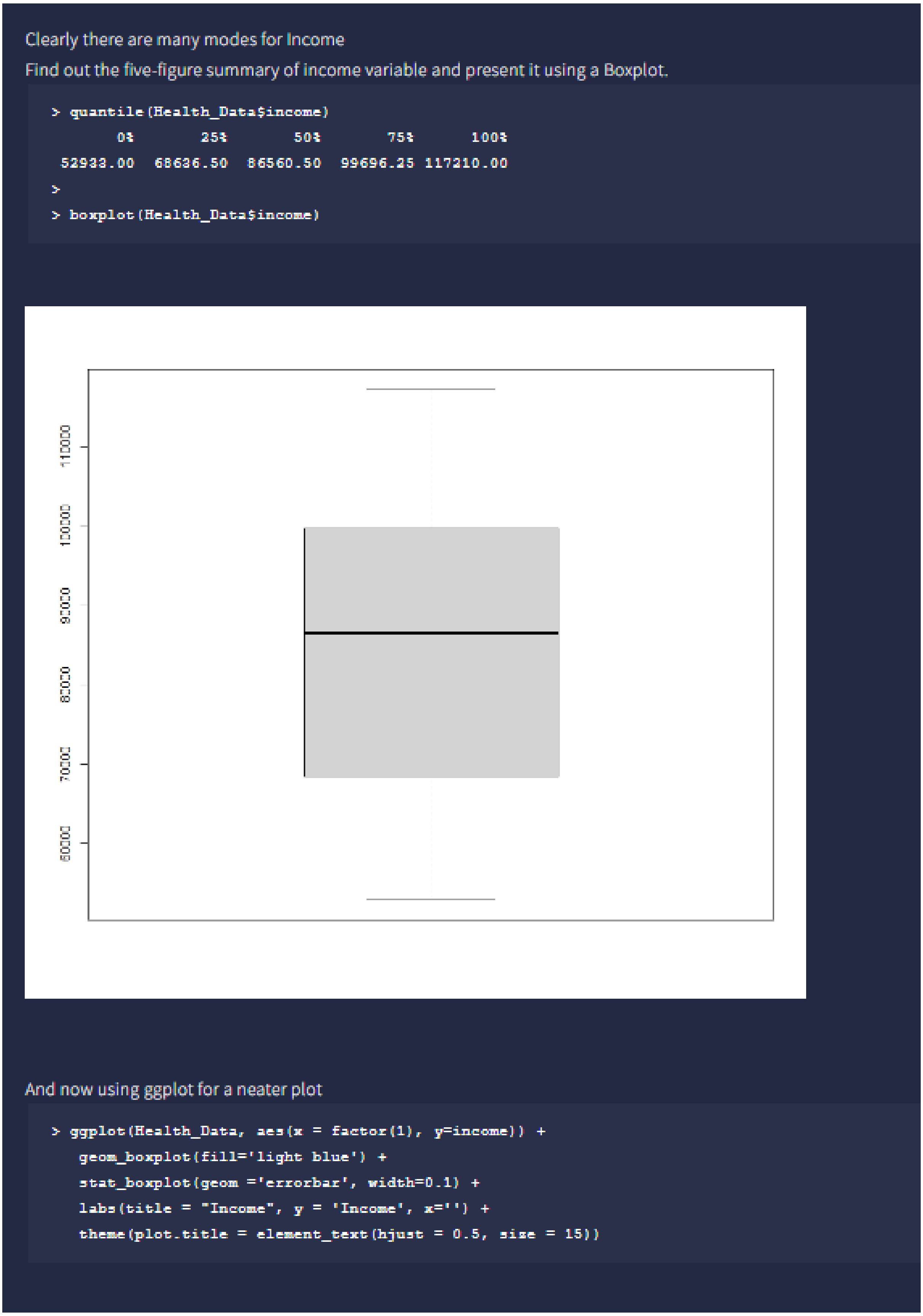


Data Activity 5

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

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

```
> # Modes function from https://stackoverflow.com/questions/2547402/how-to-find-the-statistical-mode
# Uses a function that calculated Model, enhanced to remove NA values plus another that calculates multimodes. Combined here
Modes <- function(x, na.rm = FALSE) {
  if(na.rm){
    x = x[!is.na(x)]
  }
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  ux[tab == max(tab)]
}
> mean(Health_Data$sbp)
[1] 127.7333
> median(Health_Data$sbp)
[1] 123
> Modes(Health_Data$sbp)
[1] 120
> mean(Health_Data$dbp)
[1] 82.76667
> median(Health_Data$dbp)
[1] 82
> Modes(Health_Data$dbp)
[1] 80 74 82
> mean(Health_Data$income)
[1] 85194.49
> median(Health_Data$income)
[1] 86560.5
> Modes(Health_Data$income)
  [1]  79774  70295 100117 105528  71417  95920  72785  65365  84255
 [10]  77208  58323  56897  94294  78767  97935  67906  65278 112278
 [19] 100756 104527 102776  67035 100934  56249  90689 117210 103087
 [28]  88974 107683  56250  91968  58861  96574 109697  66751 106392
 [37]  69940  93009 109854 116656 104632  79538 114671 111583  84335
 [46]  59033  80865  82867  80354  67420 101023  68433  91245  63331
 [55]  94978  78270 114518 100731 107274  59437  65890  82891  76231
 [64]  74478  77477 103582  81095  98453  90790  89997  99160  92093
 [73] 106693  91736  81818  55183  66737  66415  94873  76845  84263
 [82]  83086  64482  87707  58576  92339  96351 106832  87076  98444
 [91]  65943  76056  86719  78217  92655  57111 109943 102764  83980
[100]  94690  70092 105266  57059 101416  89698  87253  57221  62926
[109] 108259  94546  78669 115942 101523  55927  78097 105462  56548
[118]  66587  81742  77597 108537  88315 116423  82265 114636  53980
[127]  80422  56232  98318  91861  66857  97549 110225 102554  79171
[136] 103503  99968  53435  85519  98875  72470  83988  61325  66517
[145]  78559  89783  80714  72884  68461  94752  60866  67389  95869
[154]  86381 101310 101602  75755 105812  65772  99294  64966  71654
[163]  67661  94620  84990  65523 101595 115374  69163  97462  83918
[172]  68350 100473  99610 113240 114488  61600  54883 106953  98474
[181]  58383  81710 107892  92478 103262  95396  53976  96009  97735
[190]  54579  52933  61856 109137  60305  97104  59440  75317  91261
[199]  75301  88929  54329  87565  72657  99725  98405  86402 107849
[208]  82803  89351  99355
```



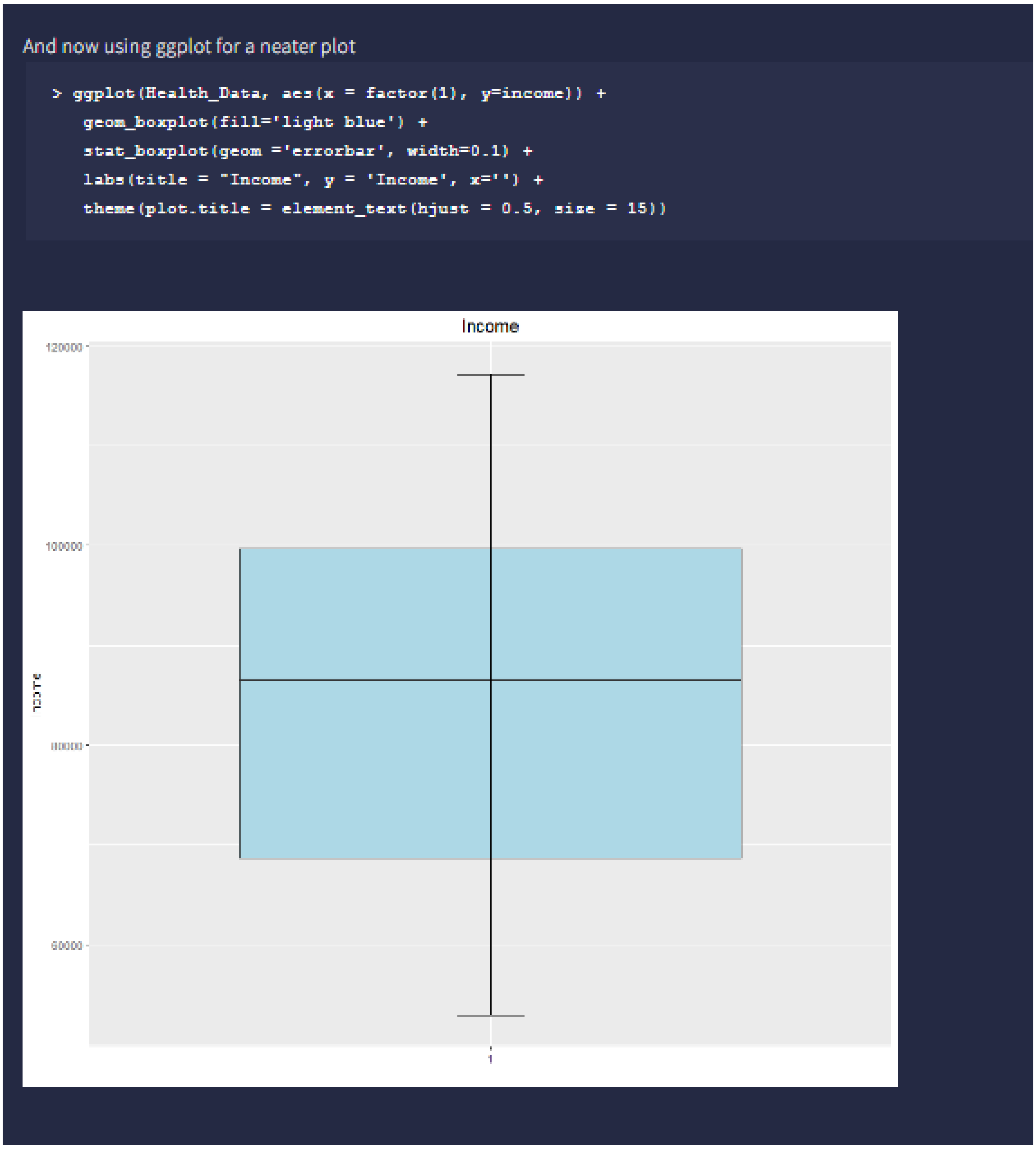
Clearly there are many modes for Income

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

```
> quantile(Health_Data$income)
       0%       25%       50%       75%      100% 
 52933.00  68636.50  86560.50  99696.25 117210.00 
>
> boxplot(Health_Data$income)
```

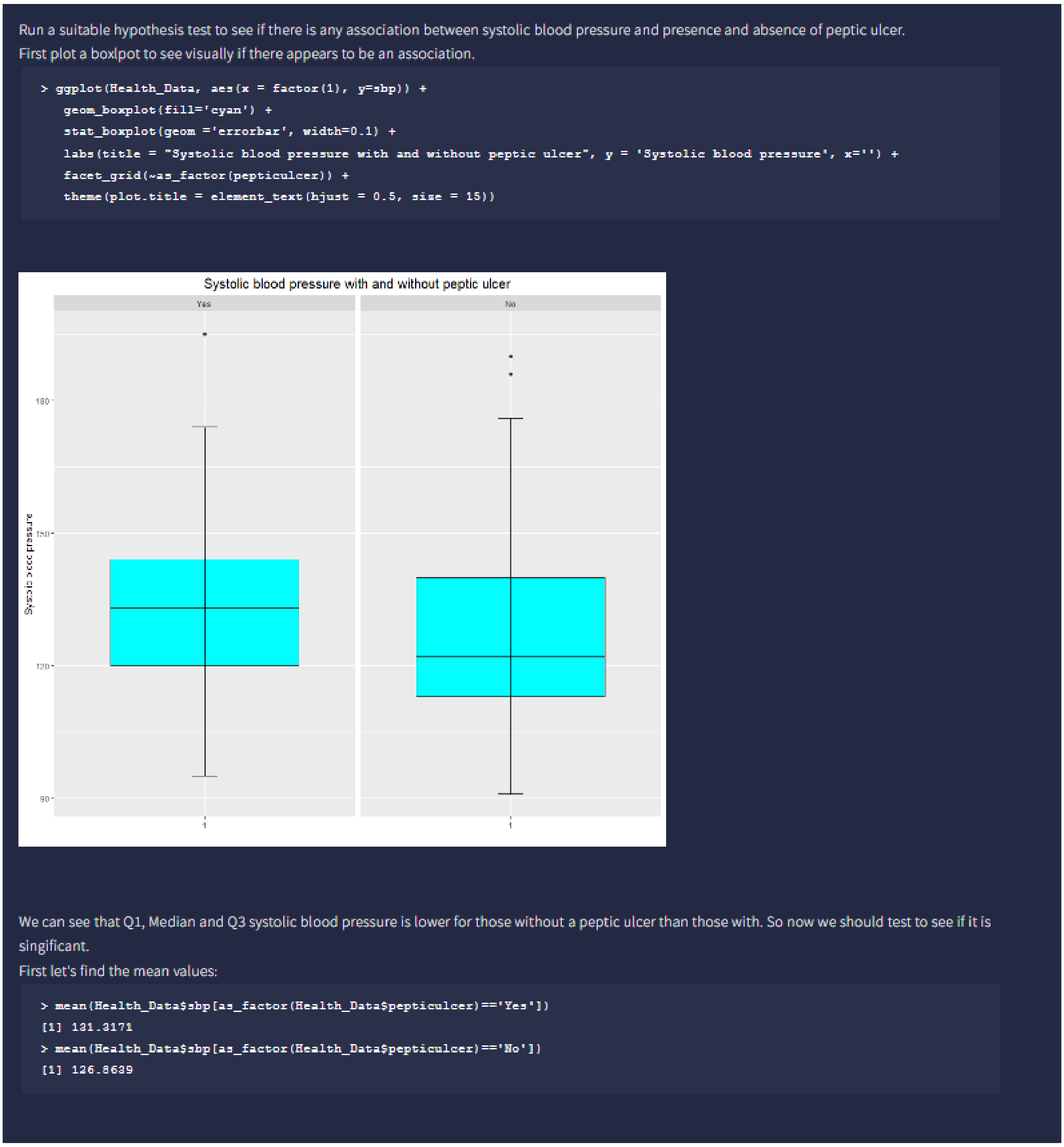
And now using ggplot for a neater plot

```
> ggplot(Health_Data, aes(x = factor(1), y=income)) +
   geom_boxplot(fill='light blue') +
   stat_boxplot(geom ='errorbar', width=0.1) +
   labs(title = "Income", y = 'Income', x='') +
   theme(plot.title = element_text(hjust = 0.5, size = 15))
```



And now using ggplot for a neater plot

```
> ggplot(Health_Data, aes(x = factor(1), y=income)) +
    geom_boxplot(fill='light blue') +
    stat_boxplot(geom ='errorbar', width=0.1) +
    labs(title = "Income", y = 'Income', x='') +
    theme(plot.title = element_text(hjust = 0.5, size = 15))
```

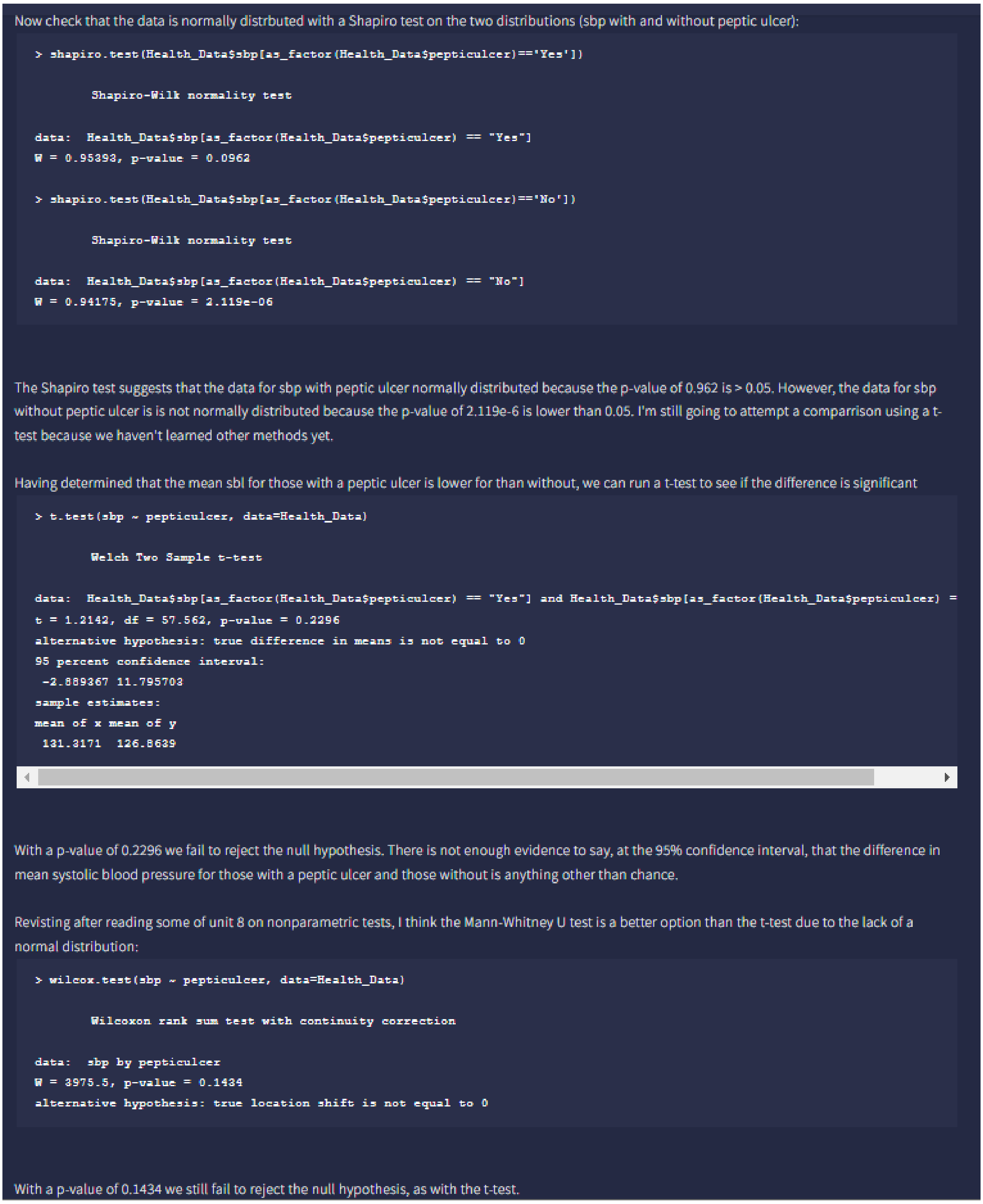


Run a suitable hypothesis test to see if there is any association between systolic blood pressure and presence and absence of peptic ulcer.
First plot a boxlpot to see visually if there appears to be an association.

```
> ggplot(Health_Data, aes(x = factor(1), y=sbp)) +
    geom_boxplot(fill='cyan') +
    stat_boxplot(geom ='errorbar', width=0.1) +
    labs(title = "Systolic blood pressure with and without peptic ulcer", y = 'Systolic blood pressure', x='') +
    facet_grid(~as_factor(pepticulcer)) +
    theme(plot.title = element_text(hjust = 0.5, size = 15))
```

We can see that Q1, Median and Q3 systolic blood pressure is lower for those without a peptic ulcer than those with. So now we should test to see if it is significant.
First let's find the mean values:

```
> mean(Health_Data$sbp[as_factor(Health_Data$pepticulcer)=='Yes'])
[1] 131.3171
> mean(Health_Data$sbp[as_factor(Health_Data$pepticulcer)=='No'])
[1] 126.8639
```



Now check that the data is normally distrbuted with a Shapiro test on the two distributions (sbp with and without peptic ulcer):

```
> shapiro.test(Health_Data$sbp[as_factor(Health_Data$pepticulcer)=='Yes'])

        Shapiro-Wilk normality test

data:  Health_Data$sbp[as_factor(Health_Data$pepticulcer) == "Yes"]
W = 0.95393, p-value = 0.0962

> shapiro.test(Health_Data$sbp[as_factor(Health_Data$pepticulcer)=='No'])

        Shapiro-Wilk normality test

data:  Health_Data$sbp[as_factor(Health_Data$pepticulcer) == "No"]
W = 0.94175, p-value = 2.119e-06
```

The Shapiro test suggests that the data for sbp with peptic ulcer normally distributed because the p-value of 0.962 is > 0.05. However, the data for sbp without peptic ulcer is is not normally distributed because the p-value of 2.119e-6 is lower than 0.05. I'm still going to attempt a comparrison using a t-test because we haven't learned other methods yet.

Having determined that the mean sbl for those with a peptic ulcer is lower for than without, we can run a t-test to see if the difference is significant

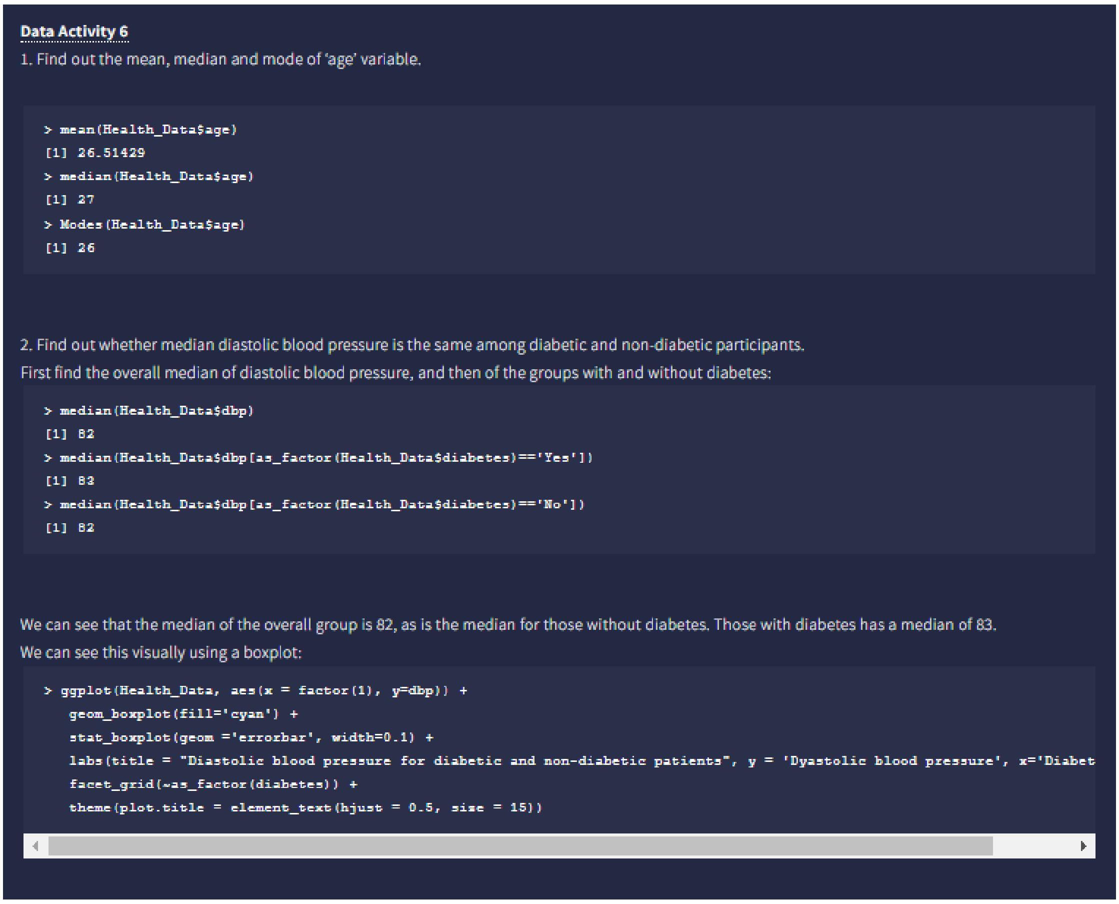
```
> t.test(sbp ~ pepticulcer, data=Health_Data)

        Welch Two Sample t-test

data:  Health_Data$sbp[as_factor(Health_Data$pepticulcer) == "Yes"] and Health_Data$sbp[as_factor(Health_Data$pepticulcer) =
t = 1.2142, df = 57.562, p-value = 0.2296
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.889367 11.795703
sample estimates:
mean of x mean of y 
 131.3171  126.8639 
```

With a p-value of 0.2296 we fail to reject the null hypothesis. There is not enough evidence to say, at the 95% confidence interval, that the difference in mean systolic blood pressure for those with a peptic ulcer and those without is anything other than chance.

Revisting after reading some of unit 8 on nonparametric tests, I think the Mann-Whitney U test is a better option than the t-test due to the lack of a normal distribution:

```
> wilcox.test(sbp ~ pepticulcer, data=Health_Data)

        Wilcoxon rank sum test with continuity correction

data:  sbp by pepticulcer
W = 3975.5, p-value = 0.1434
alternative hypothesis: true location shift is not equal to 0
```

With a p-value of 0.1434 we still fail to reject the null hypothesis, as with the t-test.



Data Activity 6

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

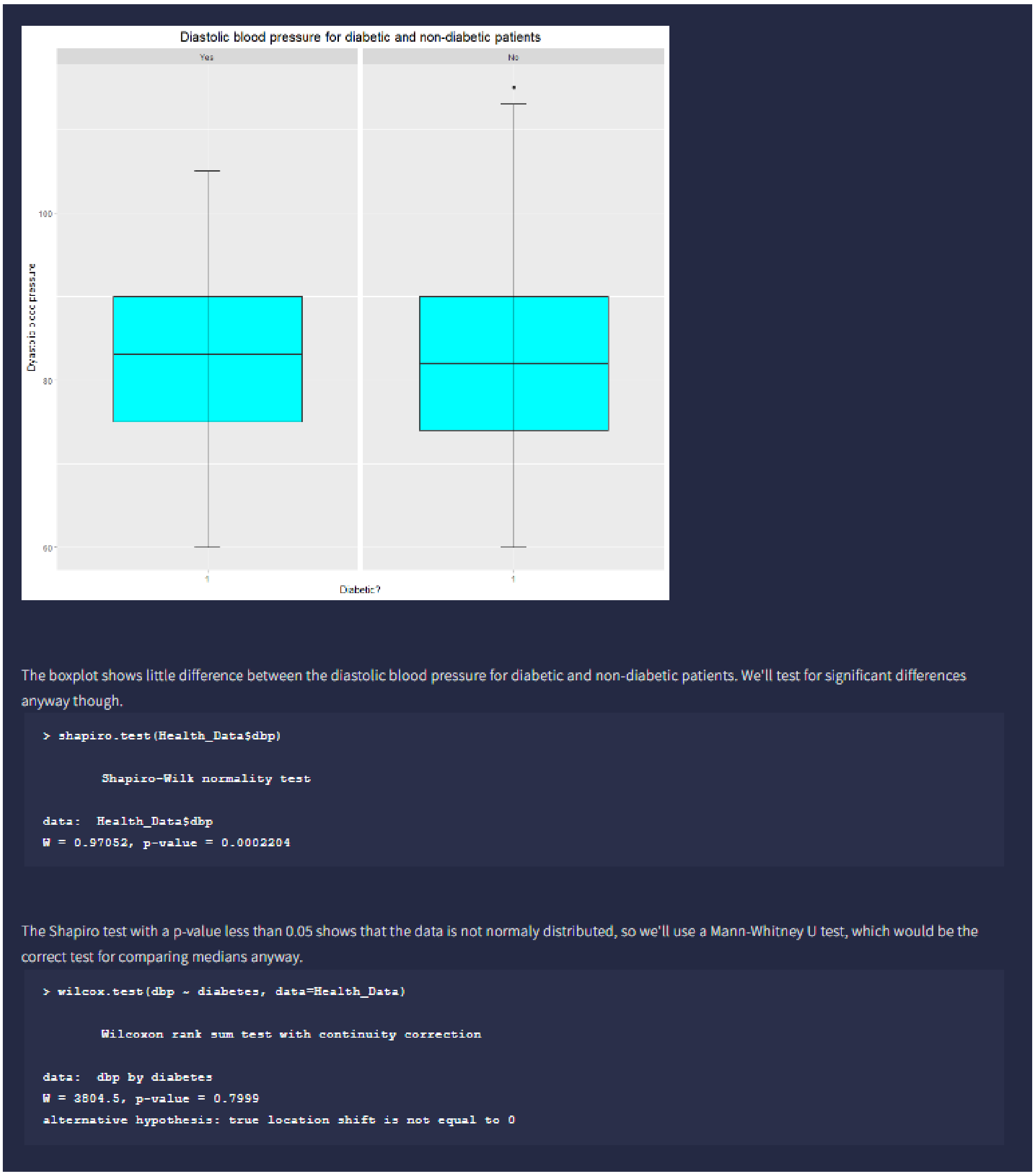
```
> mean(Health_Data$age)
[1] 26.51429
> median(Health_Data$age)
[1] 27
> Modes(Health_Data$age)
[1] 26
```

2. Find out whether median diastolic blood pressure is the same among diabetic and non-diabetic participants.
First find the overall median of diastolic blood pressure, and then of the groups with and without diabetes:

```
> median(Health_Data$dbp)
[1] 82
> median(Health_Data$dbp[as_factor(Health_Data$diabetes)=='Yes'])
[1] 83
> median(Health_Data$dbp[as_factor(Health_Data$diabetes)=='No'])
[1] 82
```

We can see that the median of the overall group is 82, as is the median for those without diabetes. Those with diabetes has a median of 83.
We can see this visually using a boxplot:

```
> ggplot(Health_Data, aes(x = factor(1), y=dbp)) +
   geom_boxplot(fill='cyan') +
   stat_boxplot(geom ='errorbar', width=0.1) +
   labs(title = "Diastolic blood pressure for diabetic and non-diabetic patients", y = 'Dyastolic blood pressure', x='Diabet
   facet_grid(~as_factor(diabetes)) +
   theme(plot.title = element_text(hjust = 0.5, size = 15))
```

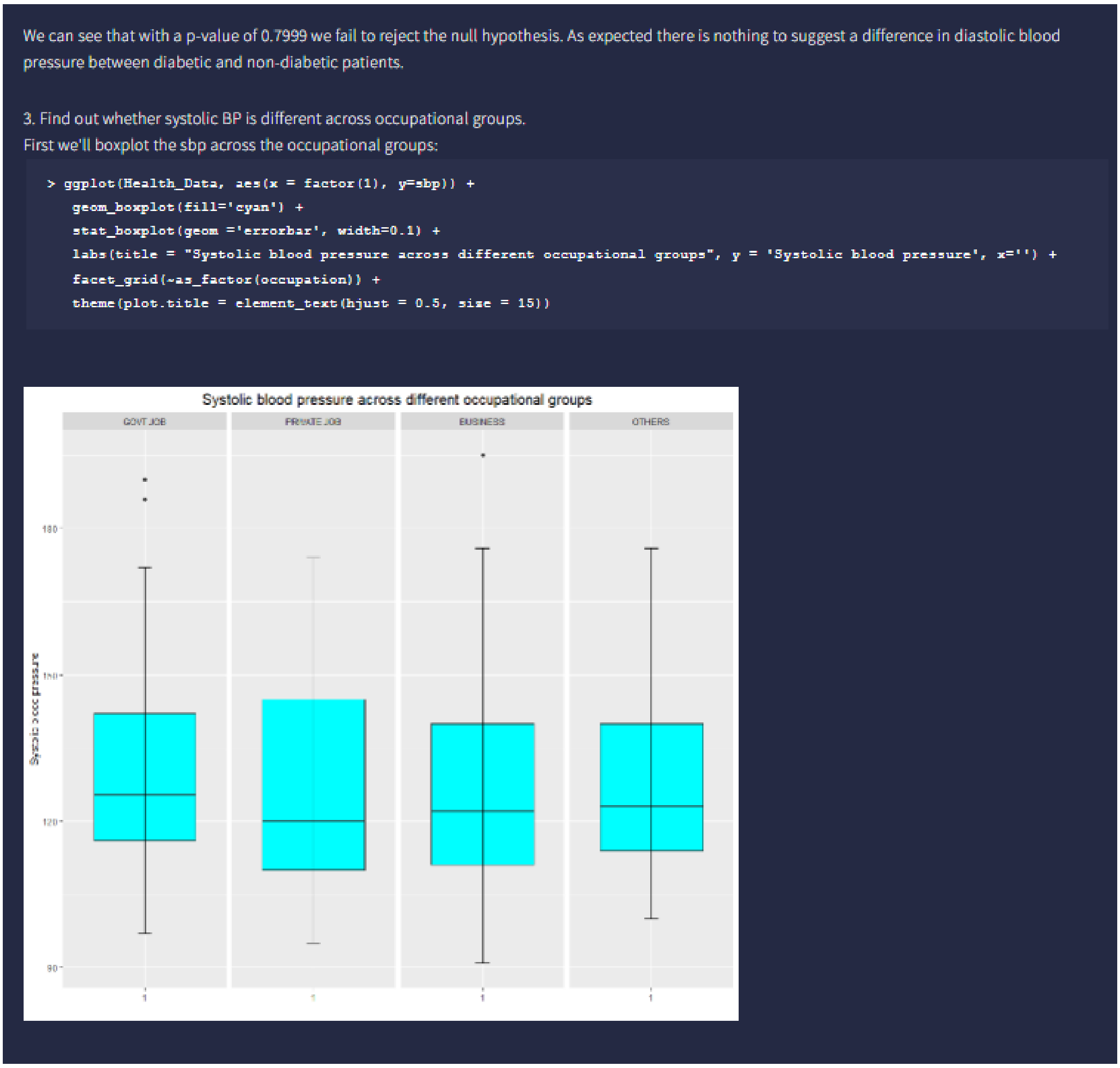

Diastolic blood pressure for diabetic and non-diabetic patients
Yes
No
Dyastolic blood pressure
100
80
60
1
1
Diabetic?

The boxplot shows little difference between the diastolic blood pressure for diabetic and non-diabetic patients. We'll test for significant differences anyway though.

```
> shapiro.test(Health_Data$dbp)

        Shapiro-Wilk normality test

data:  Health_Data$dbp
W = 0.97052, p-value = 0.0002204
```

The Shapiro test with a p-value less than 0.05 shows that the data is not normaly distributed, so we'll use a Mann-Whitney U test, which would be the correct test for comparing medians anyway.

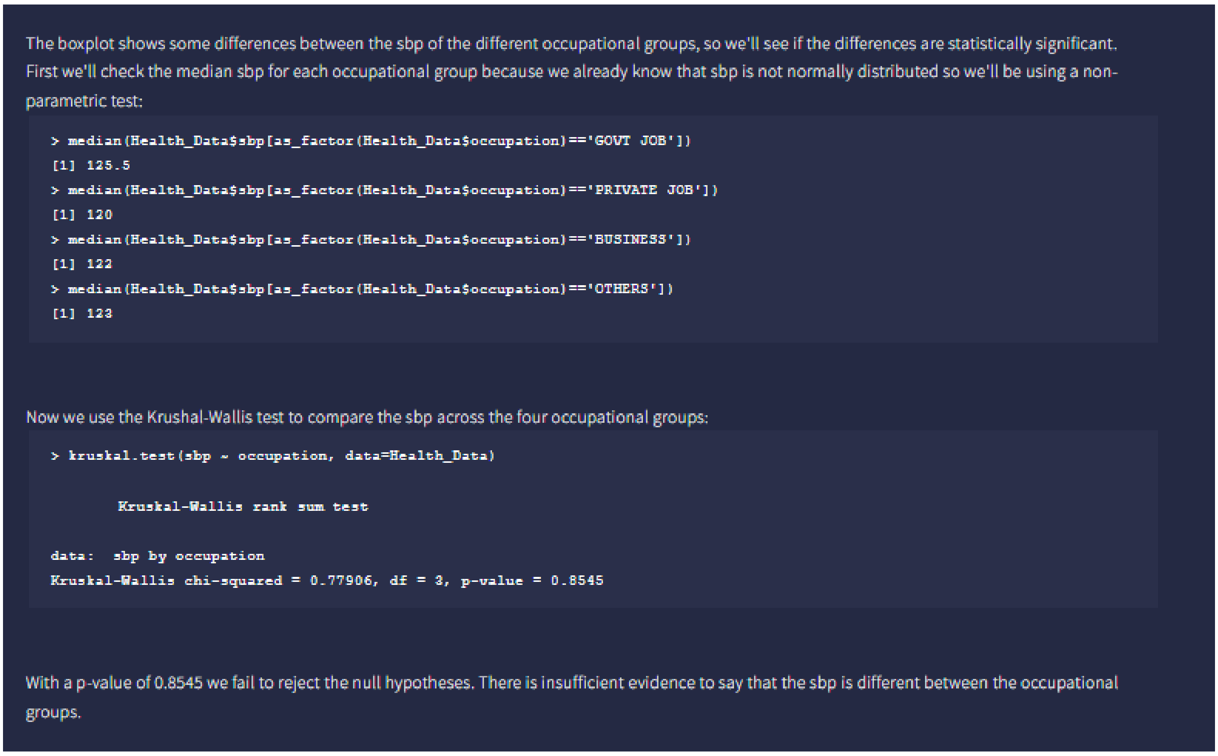
```
> wilcox.test(dbp ~ diabetes, data=Health_Data)

        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
```



We can see that with a p-value of 0.7999 we fail to reject the null hypothesis. As expected there is nothing to suggest a difference in diastolic blood pressure between diabetic and non-diabetic patients.

3. Find out whether systolic BP is different across occupational groups.
First we'll boxplot the sbp across the occupational groups:

```
> ggplot(Health_Data, aes(x = factor(1), y=sbp)) +
    geom_boxplot(fill='cyan') +
    stat_boxplot(geom ='errorbar', width=0.1) +
    labs(title = "Systolic blood pressure across different occupational groups", y = 'Systolic blood pressure', x='') +
    facet_grid(~as_factor(occupation)) +
    theme(plot.title = element_text(hjust = 0.5, size = 15))
```



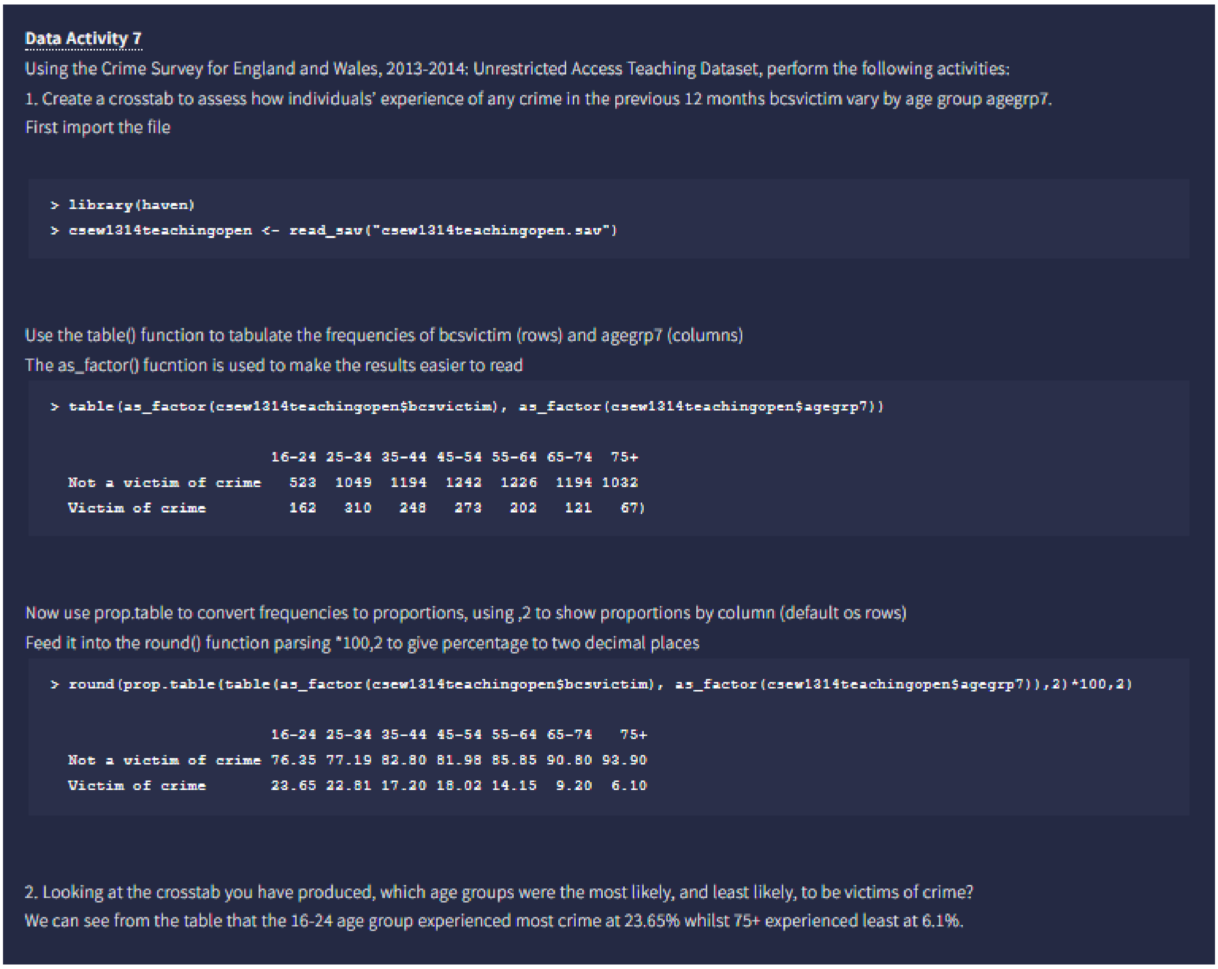
The boxplot shows some differences between the sbp of the different occupational groups, so we'll see if the differences are statistically significant. First we'll check the median sbp for each occupational group because we already know that sbp is not normally distributed so we'll be using a non-parametric test:

```
> median(Health_Data$sbp[as_factor(Health_Data$occupation)=='GOVT JOB'])
[1] 125.5
> median(Health_Data$sbp[as_factor(Health_Data$occupation)=='PRIVATE JOB'])
[1] 120
> median(Health_Data$sbp[as_factor(Health_Data$occupation)=='BUSINESS'])
[1] 122
> median(Health_Data$sbp[as_factor(Health_Data$occupation)=='OTHERS'])
[1] 123
```

Now we use the Krushal-Wallis test to compare the sbp across the four occupational groups:

```
> 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
```

With a p-value of 0.8545 we fail to reject the null hypotheses. There is insufficient evidence to say that the sbp is different between the occupational groups.



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.

First import the file

```
> library(haven)
> csew1314teachingopen <- read_sav("csew1314teachingopen.sav")
```

Use the table() function to tabulate the frequencies of bcsvictim (rows) and agegrp7 (columns)

The as_factor() fucntion is used to make the results easier to read

```
> table(as_factor(csew1314teachingopen$bcsvictim), as_factor(csew1314teachingopen$agegrp7))

                        16-24 25-34 35-44 45-54 55-64 65-74  75+
  Not a victim of crime   523  1049  1194  1242  1226  1194 1032
  Victim of crime         162   310   248   273   202   121   67)
```

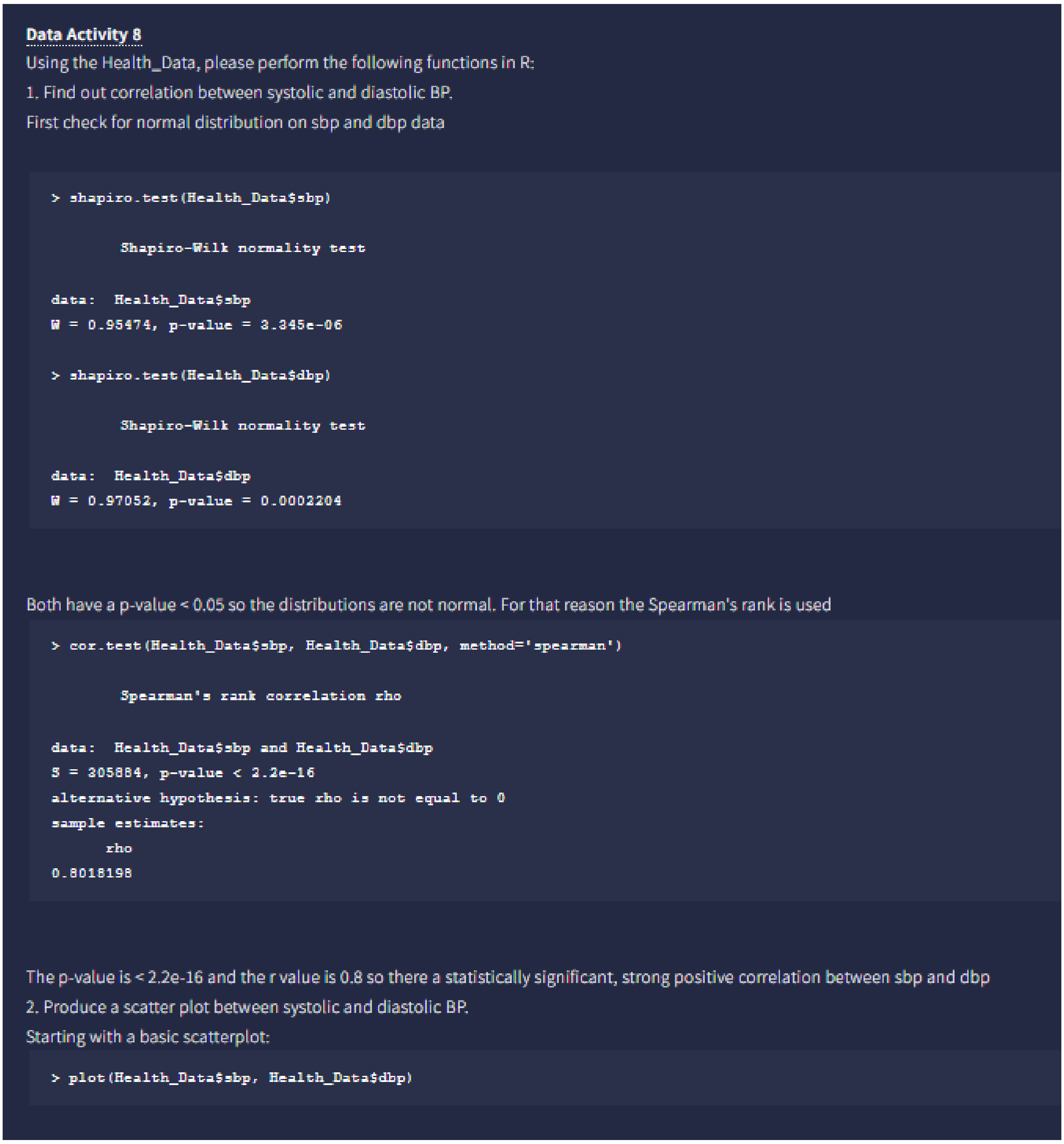
Now use prop.table to convert frequencies to proportions, using ,2 to show proportions by column (default os rows)

Feed it into the round() function parsing *100,2 to give percentage to two decimal places

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

                        16-24 25-34 35-44 45-54 55-64 65-74   75+
  Not a victim of crime 76.35 77.19 82.80 81.98 85.85 90.80 93.90
  Victim of crime       23.65 22.81 17.20 18.02 14.15  9.20  6.10
```

2. Looking at the crosstab you have produced, which age groups were the most likely, and least likely, to be victims of crime?

We can see from the table that the 16-24 age group experienced most crime at 23.65% whilst 75+ experienced least at 6.1%.



Data Activity 8

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

1. Find out correlation between systolic and diastolic BP.

First check for normal distribution on sbp and dbp data

```
> shapiro.test(Health_Data$sbp)

        Shapiro-Wilk normality test

data:  Health_Data$sbp
W = 0.95474, p-value = 3.345e-06

> shapiro.test(Health_Data$dbp)

        Shapiro-Wilk normality test

data:  Health_Data$dbp
W = 0.97052, p-value = 0.0002204
```

Both have a p-value < 0.05 so the distributions are not normal. For that reason the Spearman's rank is used

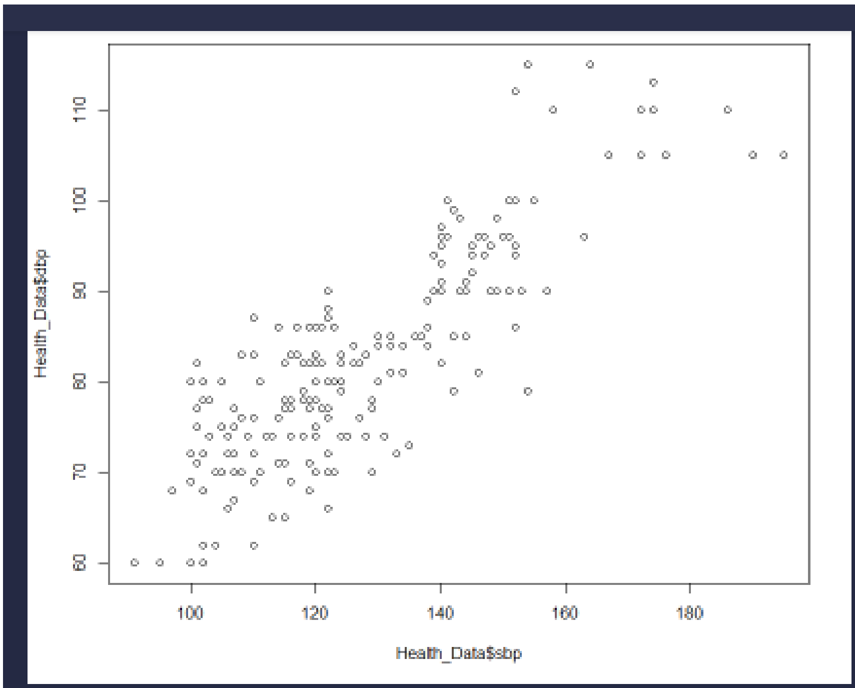
```
> cor.test(Health_Data$sbp, Health_Data$dbp, method='spearman')

        Spearman's rank correlation rho

data:  Health_Data$sbp and Health_Data$dbp
S = 305884, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8018198
```

The p-value is < 2.2e-16 and the r value is 0.8 so there a statistically significant, strong positive correlation between sbp and dbp

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

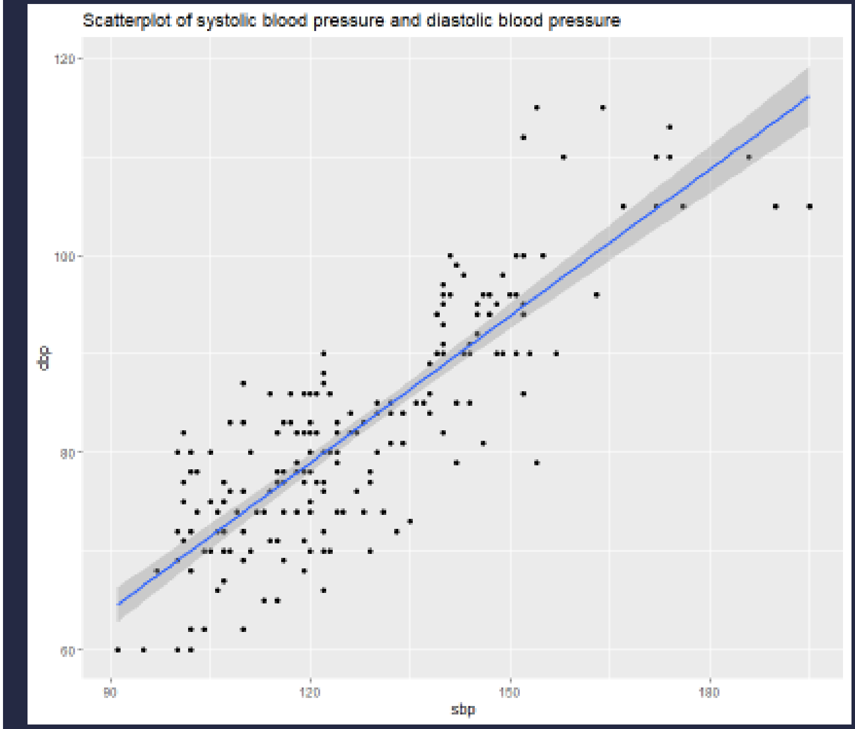
Starting with a basic scatterplot:

```
> plot(Health_Data$sbp, Health_Data$dbp)
```

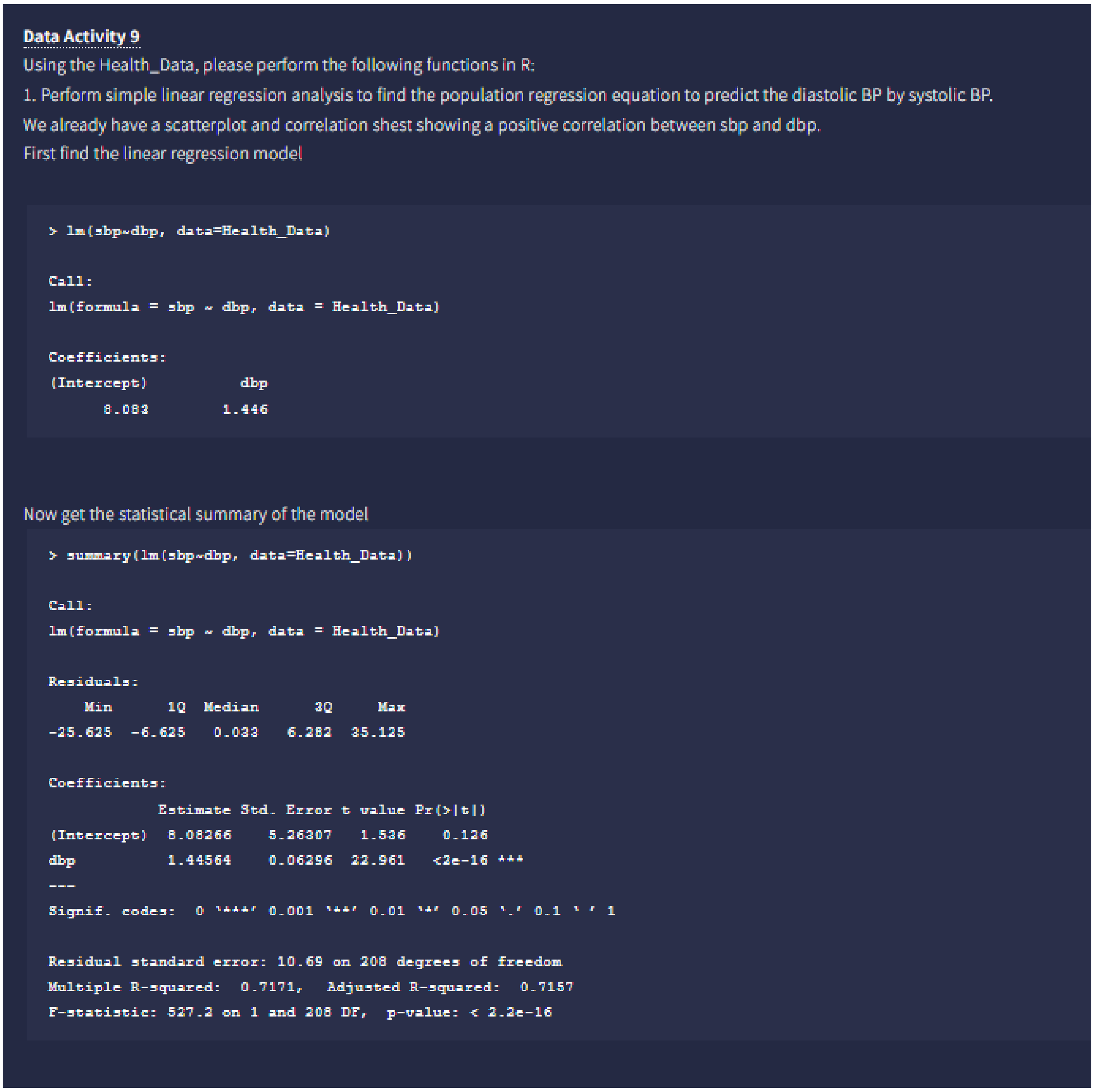


And now a ggplot scatterplot including a regression line:

```
> ggplot(Health_Data, aes(x=sbp, y=dbp)) +
    geom_point() +
    geom_smooth(method=lm) +
    labs(title = "Scatterplot of systolic blood pressure and diastolic blood pressure")
```



The scatterplots show the high positive correlation previously calculated



Data Activity 9

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.

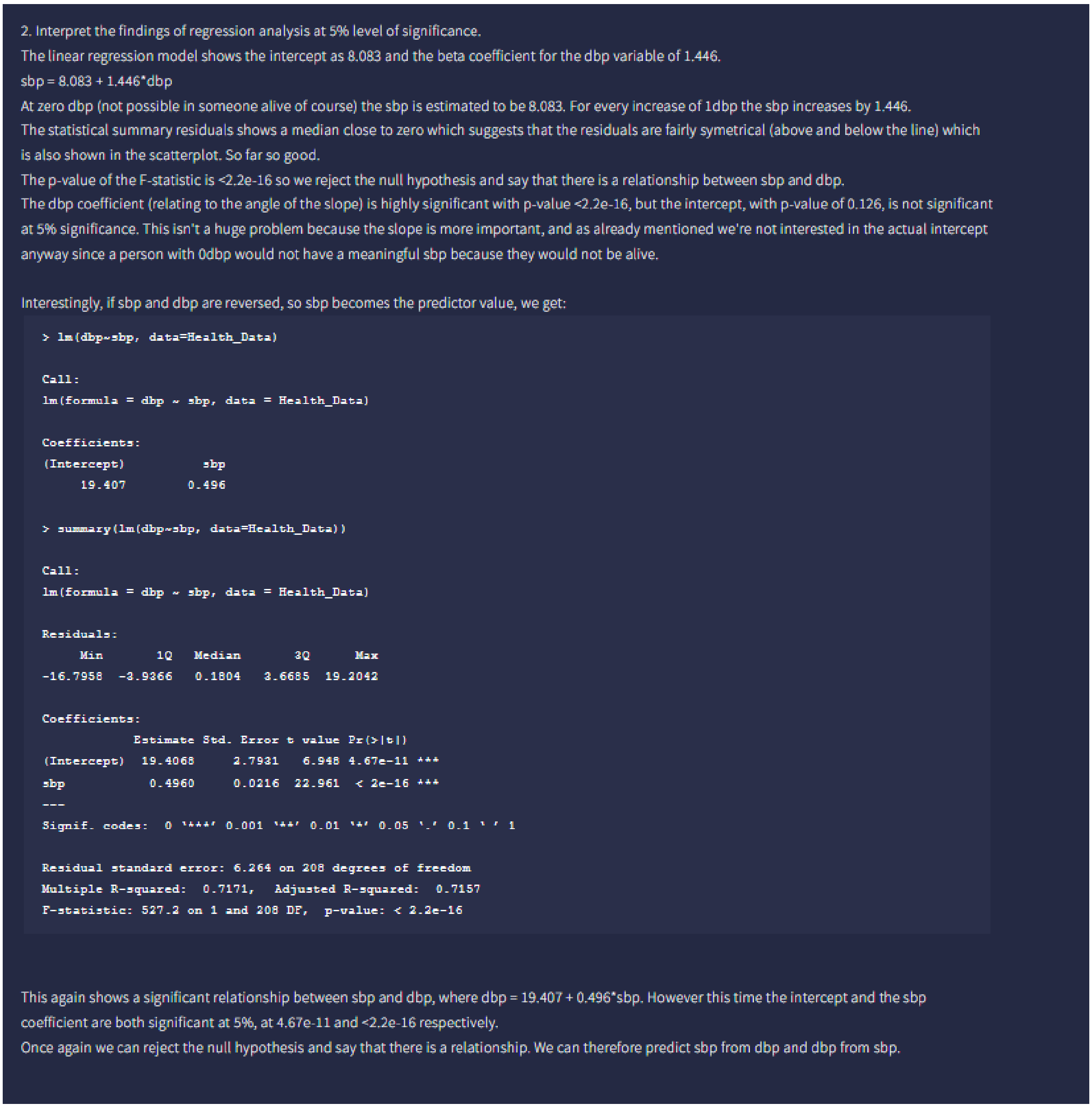
We already have a scatterplot and correlation shest showing a positive correlation between sbp and dbp.

First find the linear regression model

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

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

Coefficients:
(Intercept)          dbp
      8.083        1.446
```

Now get the statistical summary of the model

```
> summary(lm(sbp~dbp, data=Health_Data))

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

Residuals:
    Min      1Q  Median      3Q     Max
-25.625  -6.625   0.033   6.282  35.125

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.08266    5.26307   1.536    0.126
dbp          1.44564    0.06296  22.961   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.69 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 linear regression model shows the intercept as 8.083 and the beta coefficient for the dbp variable of 1.446.
sbp = 8.083 + 1.446*dbp
At zero dbp (not possible in someone alive of course) the sbp is estimated to be 8.083. For every increase of 1dbp the sbp increases by 1.446.
The statistical summary residuals shows a median close to zero which suggests that the residuals are fairly symetrical (above and below the line) which is also shown in the scatterplot. So far so good.
The p-value of the F-statistic is <2.2e-16 so we reject the null hypothesis and say that there is a relationship between sbp and dbp.
The dbp coefficient (relating to the angle of the slope) is highly significant with p-value <2.2e-16, but the intercept, with p-value of 0.126, is not significant at 5% significance. This isn't a huge problem because the slope is more important, and as already mentioned we're not interested in the actual intercept anyway since a person with 0dbp would not have a meaningful sbp because they would not be alive.

Interestingly, if sbp and dbp are reversed, so sbp becomes the predictor value, we get:

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

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

Coefficients:
(Intercept)          sbp
     19.407        0.496

> summary(lm(dbp~sbp, data=Health_Data))

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
```

This again shows a significant relationship between sbp and dbp, where dbp = 19.407 + 0.496*sbp. However this time the intercept and the sbp coefficient are both significant at 5%, at 4.67e-11 and <2.2e-16 respectively.
Once again we can reject the null hypothesis and say that there is a relationship. We can therefore predict sbp from dbp and dbp from sbp.