Practical 4: The Three Project Types

In this course you have to produce a report based on one out of three different kinds of research question:

- a correlation,
- a comparison of means
- comparison of proportions.

To help you get an idea of what is required for each, we will use the R Scripts used in the last practical but this time we will look in much greater depth:
https://github.com/stivb/practical3_n_4.git

The files we particularly we want you to look at are:

- Comparison_of_means.R
- Comparison_of_proportions.R
- Correlation.R

These try to analyse the data from a survey last year made of students on this course.

If you havent already done this year's survey of Masters' students, do it now at:
https://forms.office.com/e/UCWxNfT11s

There are more questions and also the data will be fresher. So if you want to get an understanding of how much your peers spend on food and transport, how much homework they do, what are their favorite music genres etc, then use the other dataset which will be released dependin on how many students fill it in.

By now you should have chosen (or at least have some idea of) your research question. Choose one of the R files above, which is the most relevant to your own research question. Then open it, and run them – not quite line by line – but at least the parts between comments. Select those parts, and then press the "Run" button. Each time you do that, try to work out, by yourself, or speaking to your neighbours, what is actually happening in that line. You might also want to look in the environment window to see the value of certain variables, and you might also wish to do that by typing the variables into the console. Then read the relevant section below

Preparing Your Data in Your R Script

Your script will often start by importing data, and then renaming the columns in order to make it easier to query the data. The following lines appear in all the 3 scripts above.

Code	Explanation

library(readr)	This allows you to bring in libraries with specific functions. Usually you just need one line: library(tidyverse) since that has all the important libraries in it. In order to use any library, you need to have imported it into your system first. To do this use: `install.packages("tidyverse").`
survey_data <- read_csv("survey_data.csv") df<-survey_data	The first line here reads in a dataset The second line just renames it to df. Once you have read items into a dataset variable you can query them easily. For instance in the console window just type
names(df)[7] <- "award" names(df)[8] <- "gender" names(df)[9] <- "height" names(df)[10] <- "continent" names(df)[11] <- "pe_minutes" names(df)[12] <- "miles" names(df)[13] <- "gbptransport" names(df)[14] <- "get2campusminutes" names(df)[15] <- "gbpfood"	*This is an example of what is called "wrangling" - namely changing the data to make it more easy to use. The actual survey data has column headers which are very long and typing them each time we wished to use them would be tedious and error prone.* **Which Award Are You Studying For in UH (choose nearest one if yours is not here)** | **What is your gender** | **What is your height in cm (just put a number)** *You'll notice we can now refer to "What is your height..." With just a single word "height"*

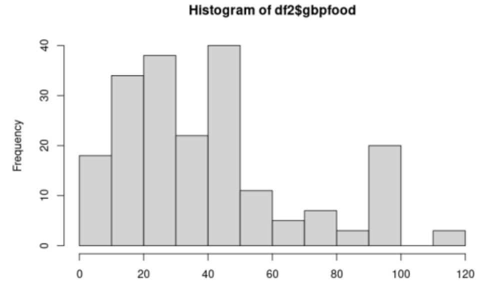
To see the value of this, after running these lines just go into the console and type
`mean(df$height)` - this will give us the average
To get a feel for any dataset, the best thing to do is run the "summary" command over the whole dataset
`summary(df)`
If you do that, you will get a description of all the columns in the dataset and also information (such as mean) about any of the numerical columns.

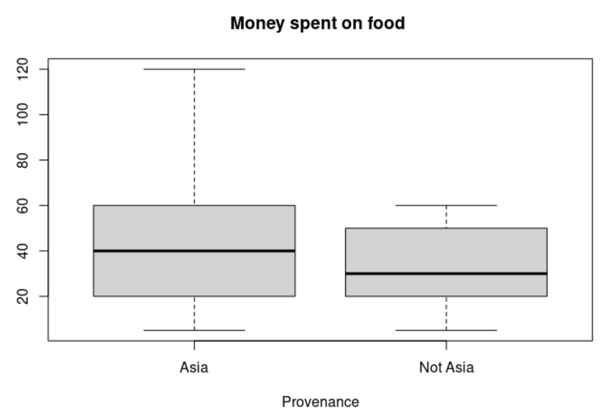
Research Question About Comparison of Means (Comparison_of_means.R)

Lets start with a simple question: *do Asian students spend more on food each week than non-Asian students?*

df$isAsian<-ifelse(df$continent=="Asia","Asia","Not Asia")	Firstly we do some wrangling – because the number of African and European students on their own are too small to make a meaningful comparison – lets create a field call **isAsian** which will have just two values "Asia" or "Not Asia"
df2<-subset(df,gbpfood<150)	Because some respondents have said they spend over £200 per week on food,(implausible) lets

<table>
<tr><td></td><td>remove them from the dataset and create a second dataset called df2</td></tr>
<tr><td>hist(df2$gbpfood)

#Having put the limit at 150 we get something
#that is a bit normal but not convincingly. (If we
#had put the limit at 80 it does appear more
#normal but still not very good.</td><td>
</td></tr>
<tr><td>t.test(df2$gbpfood ~ df2$isAsian)</td><td>This is the t test. Its for normal data so not really advisable. But it is included here since the case is arguable.</td></tr>
<tr><td><pre>data: df2$gbpfood by df2$isAsian
t = 2.4931, df = 30.309, p-value = 0.01835
alternative hypothesis: true difference in
means between group Asia and group Not Asia is
not equal to 0
95 percent confidence interval:
  2.091904 21.002018
sample estimates:
    mean in group Asia mean in group Not Asia
              45.54696   34.00000</pre></td><td>According to the data, Asian students spent £45 a week on food, and the non asians £34. The p-value is 0.018 meaning this could happen by chance 1.8 times every 100 times this number of students were asked this question. Since this is quite low (below 0.05) we can say this is a significant result and we can discard the null hypothesis.</td></tr>
<tr><td colspan="2">However, this is not completely convincing since the data does not look very normal (although it should be if people filled in the form responsibly). So lets do a wilcox test which compares the rankings of the non-Asians against Asians. Are most of the Asians in the top food spenders and most of the non-Asians in the more thrifty rankings?</td></tr>
<tr><td><pre>wilcox.test(df2$gbpfood ~ df2$isAsian)</pre></td><td>This is the test to compare the rankings of non-Asians and Asians regarding food expenditure</td></tr>
<tr><td><pre>data: df2$gbpfood by df2$isAsian
W = 2123.5, p-value = 0.2016
alternative hypothesis: true location shift is
not equal to 0</pre></td><td>The wilcox test does not give us actual values for the two distributions (Asian vs Non-Asian) but it does give us a number 0.2016 which the probability of the null hypothesis being true. (In other words that it was a fluke). And here its quite high (greater than 0.05) - and therefore we have to conclude that we can't actually reject the null hypothesis. If you want to see the actual median values of the two groups of consumers use the aggregate function like this:
aggregate(gbpfood ~ isAsian, df2, median)</td></tr>
</table>

`#The diagram used for comparison of means is a`
`#box plot`
 	
`boxplot(df2$gbpfood ~ df2$isAsian,  xlab = "Provenance", ylab = "£s weekly spent on food", main = "Money spent on food ")`
	

So concluding, if we decide the data is not normal, and we use a Wilcox test, then based on the data, even though there is a difference in the median values for money spent on food, it is not a significant one and we would not be able to quote it in a paper.	

Research Question About Comparison of Proportions (Correlation.R)

Next lets look at the number of students who choose different awards. Are the proportions of awards chosen by women different to those chosen by men? The test statistic we use to answer this question is the Chi Squared Test which totals up the numbers for the whole population – the choices of men and women. Then compares it with the choices made by just men, and then the choices made by just women. And then it comes up with a number to say whether gender really does influence the choice of course.

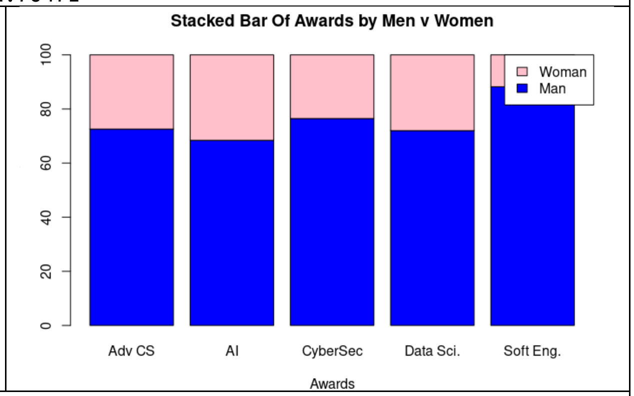
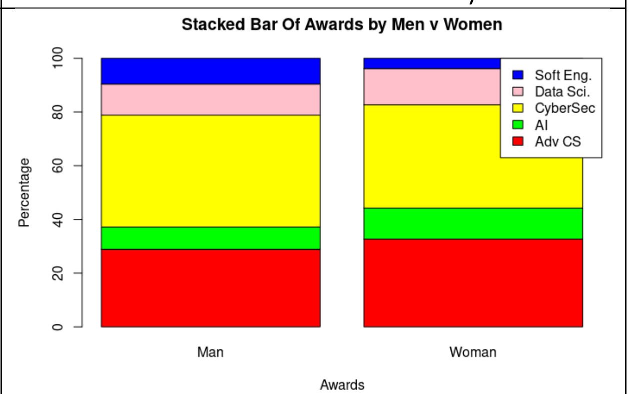
df2<-subset(df,gender=="Man" | gender=="Woman")	Here a wrangling to start. The survey itself allowed *non-binary* and *prefer not to say* as an answer. However, the numbers in those groups are too small to make a meaningful comparison. So we reduce our dataset to just those who declared themselves to be "Man" or "Woman" (this is a purely mathematical rationale – and should not be taken as any kind of traditionalism on the part of the author).
pt <- table(df2$gender,df2$award)	Whenever you do chi squared you first have to turn the data into a table of categories and counts. The two categories are gender (m/w) and awards (data science/AI etc). The table will look like below

	Advanced Computer Science	Artificial Intelligence	Cybersecurity	Data Science	Software Engineering
Man	45	13	65	18	15
Woman	17	6	20	7	2

chisq.test(pt) `X-squared = 2.4384, df = 4, p-value = 0.6557`	The result we have has a p value which is quite high (it says 65.5 times per hundred, the result we obtained would satisfy the null hypothesis) -

<table>
<tr><td></td><td>which means we definitely cannot reject the null hypothesis (that gender makes no difference)</td></tr>
<tr><td>colnames(pt) = c("Adv CS", "AI", "CyberSec", "Data Sci.", "Soft Eng.")</td><td>Just some wrangling to make the groups easier to display when we put them on a graph (renaming columns to make them shorter)</td></tr>
<tr><td>percentages <- prop.table(pt, margin=2) * 100</td><td>In order to make the stacked barchart – the typical way to display comparisons of proportion, we us the prop table function which turns all the numbers in the table above into percentages</td></tr>
<tr><td colspan="2">

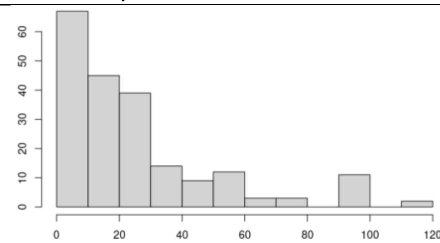
```
       Adv CS       AI CyberSec Data Sci. Soft Eng.
  Man   72.58065 68.42105 76.47059  72.00000  88.23529
  Woman 27.41935 31.57895 23.52941  28.00000  11.76471
```

</td></tr>
<tr><td>barplot(percentages, col = c("blue", "pink"), xlab = "Awards", ylab = "Percentage",
main = "Stacked Bar Of Awards by Men v Women", ylim = c(0, 100),
legend.text = c("Man", "Woman"), args.legend = list(x = "topright"))</td><td>
</td></tr>
<tr><td colspan="2">But is this the best way to display the data? Instead of having gender by color and awards by bars, we could invert it to have awards by color and gender by bars</td></tr>
<tr><td>tpercentages<-prop.table(t(pt), margin=2) * 100</td><td>The t() function transposes the data (changes columns to rows and rows to columns).</td></tr>
<tr><td>barplot(tpercentages, col = c("red", "green", "yellow", "pink", "blue"), xlab = "Awards", ylab = "Percentage",
main = "Stacked Bar Of Awards by Men v Women", ylim = c(0, 100),
legend.text = rownames(tpercentages), args.legend = list(x = "topright"))</td><td>
</td></tr>
<tr><td colspan="2">To be honest – it could be displayed either way – because the reality is – that the preferences for award titles between women and men are not very significant at all. Cybersecurity and Advanced Data Science are the more popular award titles for both men and women. But being a man or a woman does not really influence this very much.</td></tr>
</table>

Research Questions about Correlation (Correlation.R)

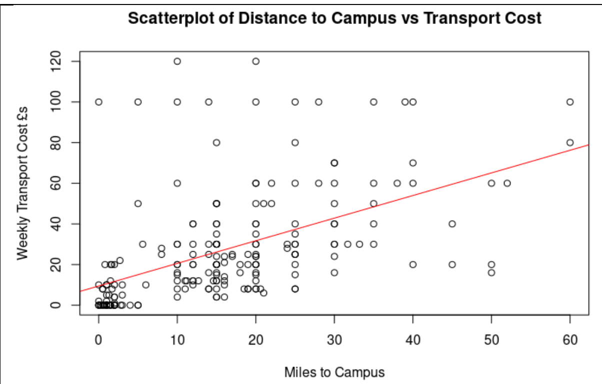
Here is our correlation question. Is there a relationship between the distance from university in miles and the amount spent by students on transport.

Code	Explanation

<table>
<tr><td>df2<-subset(df,miles<100 & gbptransport<150)</td><td>Here we use the subset function to filter the data. Because (in most surveys) there will always be people who answer flippantly. So we have to remove data. So if someone said they lived more than 100 miles away then its likely that this is either mistaken or frivolous. And probably if they say they spend greater than 150 pounds on transport that is probably untrue (or the person is extremely rich and takes taxis all the time)</td></tr>
<tr><td>hist(df2$gbptransport, main = "Histogram of Transport Costs")</td><td>

A histogram can tell us by sight – how "normal" the data is. This is clearly not normal. >60 people spend less than £10 per week, just over 40 people spend £10-20 and £20-30, only 11 people spend £30-40</td></tr>
<tr><td colspan="2">Why is this important? Because if the data is normal, we correlate the values (amount travelled in miles and amount spent in pounds), but if it not normal (which is the case here) - we correlate the *rankings*. E.g does ones position in the rankings of distance covered, agree with ones position in the rankings of amount spent?</td></tr>
<tr><td>cor.test(df2$miles, df2$gbptransport, method="spearman")</td><td>Because the data is not normal, we use a spearman correlation (if it was normal, we would use a pearson correlation)</td></tr>
<tr><td>data: df2$miles and df2$gbptransport
S = 522932, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6357956</td><td>When you get this result there are two really important values. The actual correlation 0.635 is quite high. Indicating that there really is a relationship between distance and spending. But can we be confident that this was not just a fluke? The way we measure the robustness of the result is through the p. value which we wish to be as low as possible. Here it is 2.2e-16 (which basically means 22 preceeded by 16 zeros) i.e. 0.0000000000000000022
Which means there is a very low probability the null hypothesis is valid (it would come up 22 times every quadrillian trials)</td></tr>
<tr><td>plot(df2$miles, df2$gbptransport, xlab = "Miles to Campus", ylab = "Weekly Transport Cost £s", main = "Scatterplot of Distance to Campus vs Transport Cost") # adds the dots</td><td>These two lines produce a scatterplot – the best graph for cases of correlation.</td></tr>
</table>

```
abline(lm(df2$gbptransport ~
df2$miles), col = "red") # add
the trendline
```



Each dot is one respondent in the survey. The first high dot is the unlikely person who spends £100 a week while living 0 miles from campus.

Once you have had a look, try to answer a corresponding quiz form:

- Quiz about Comparison_of_means.R
- Quiz about Comparison_of_proportions.R
- Quiz about Correlation.R

Now you have looked through these – look at the results instead of data from this year's survey. What interesting correlations, or comparisons of means or proportions can you find?