Declaration on Plagiarism

This form must be filled in and completed by the student(s) submitting an assignment

Name:	Maureen Maguire
Student Number:	19213997
Programme:	Msc in Computing(DA)
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	13 Dec 2019
Module Coordinator:	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at
http://www.dcu.ie/info/regulations/plagiarism.shtml,
https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name:____Maureen Maguire__________________________ Date: ___14/12/2019______

Is there a gender imbalance in Irish educational institutions?

Abstract (max 200 words)

What is the question or story you are trying to tell?

The main theme of this visualisation is education. The goal is to compare the education levels of male and female students in Ireland and to illustrate the decrease in the number of students attending third level education. In terms of gender balance in education, It is widely known that in the past less women attended educational institutions than men. This visualisation aims to inform the audience on whether a gender imbalance still exists in our education systems today.The aim is to compare Ireland to Norway to see if there are any similarities in terms of the percentage of male versus female students.

What is the conclusion that you reached?

It is clear from the chart that Ireland does not suffer any form of gender imbalance in our education systems. There is very little difference in terms of the percentage of young male and female students attending education, at any level. It must be noted that the percentage of students, male and female, pursuing postgraduate studies drops significantly. Looking at Norway, we can see that gender imbalance is almost non-existent, but in some countries the imbalance is pushing in the opposite direction, with more female students than males.

1. Dataset [½ page]

Where/how did you retrieve it or them

The dataset used in this project was downloaded from 'The World Bank'. This is a public dataset and can be found at the following link: https://data.worldbank.org/indicator/sp.pop.totl as well as population data from the following link: https://population.un.org/wpp/Download/Standard/Population/

Describe the data - size (GB or attributes), number of rows, attributes, data types present

The dataset is 0.018718 GB with 64 columns and 41502 rows. The file mainly consists of string and float data types. Each column represents a year with the rows containing the statistics for that year. Each statistic is a representation of different world indicators, which were labeled 'Series', for example, 'Educational attainment, at least a Bachelor's or equivalent, population 25+, total (%) (cumulative)'. This dataset had quite a significant number of missing values.

What aspects (if any) of big data (volume, variety, velocity) are present in your data

The variety aspects tis present here because there are two population datasets being used.

2. Data Exploration, Processing, Cleaning and/or Integration [½ page]

What did you need to do to prepare the dataset(s) to create your graph/chart?

Data Exploration

Data exploration was carried out in Tableau and Excel. The initial exploration was carried out in Excel to get an idea of the type of data involved. The uncleaned dataset was then imported to Tableau and plotted in different ways in an attempt to get a basis for the visualisation.

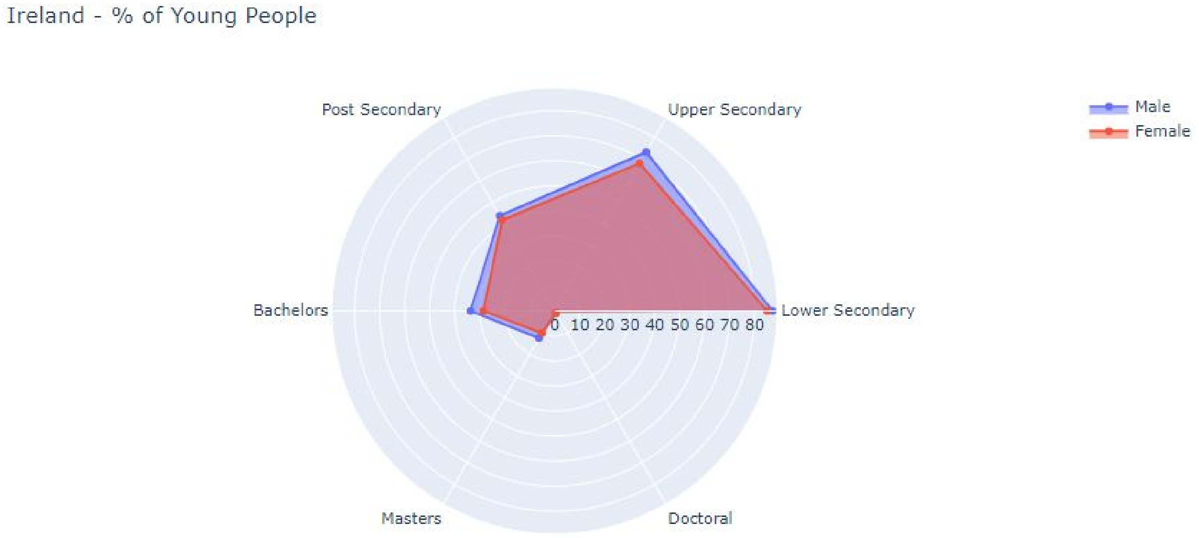
Cleaning
The cleaning of this dataset was completed using Python in Jupyter Notebooks. The column name formats made it difficult to return columns easily due to their length and the presence of special characters. For example, time domain columns were '1972 [YR1972]' which I renamed to '1972'. The series row values were also renamed from 'Educational attainment, at least a Bachelor's or equivalent, population 25+, male (%) (cumulative)' to 'Bachelors'. This allowed for smoother retrieval and clearer graph. This dataset had a lot of missing values resulting in a lot of the data being unusable. Where applicable, the empty cells were replaced with the mean value of the year before and after. All rows and columns that were not being used in the visualisation were removed. The data frame was then reindexed to account for this. Processing the data was made difficult due to the layout of this dataset. New columns were added to the dataset to allow for easier row/column retrieval and plotting. These columns were 'Gender' and 'Year'. The reason for this was because I wanted the series name and gender separated to allow for comparison on the chart.

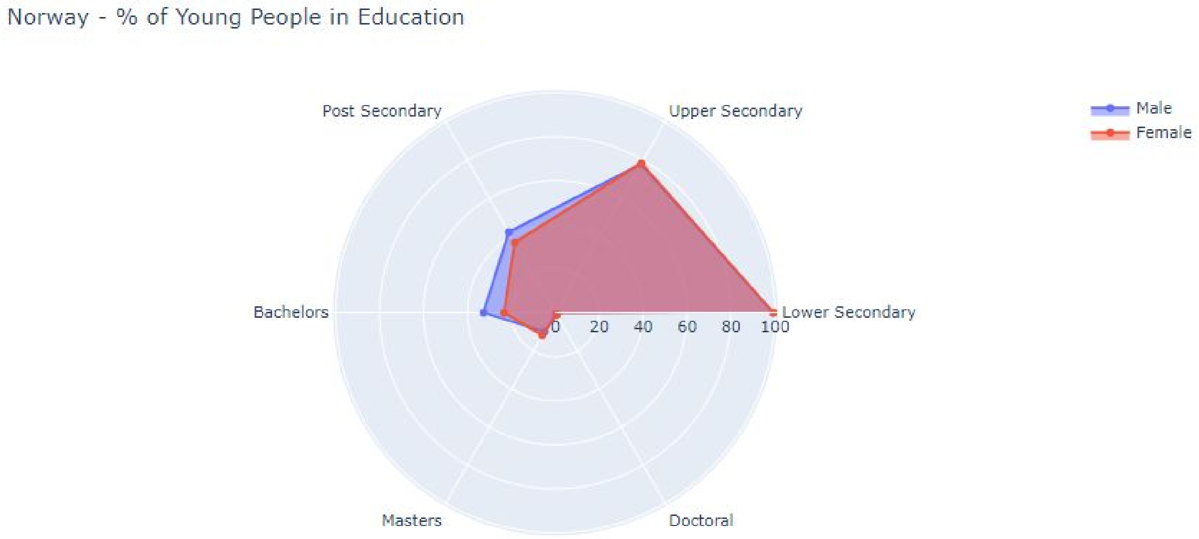
Integration
To make the integration as smooth as possible, I generated another dataframe from the existing one which consisted of r and theta values for the radar chart. These values represent the radial and angular coordinates and are passed as arguments for scatter polar. Ireland and Norway dataframes were separated as they were being plotted on separate charts.

How did you choose the attributes to visualise?
The attributes I choose to visualise were country, gender, sex, series name and percentage. I choose the country Norway for the main reason I wanted it to act as a contrast to Ireland. The values in Norway did not correlate exactly to Irelands. I excluded any attributes which I felt would not give an accurate or informative picture. For example, there were quite a number of columns which were set to 100% of the population. I also choose attributes which had values that could be seen clearly on the radar chart.

3. Visualisation [½-1 page]

Screenshot or image of visualisation


Ireland - % of Young People


Norway - % of Young People in Education

Explain your choice of chart or graph type - what relationship or data type are you showing?

My choice of graph for this visualisation project is the Plotly Radar Chart. Radar charts are normally used when wanting to compare quantitative measures of one or sometimes more categories. I choose it for three reasons. The first reason is that my goal was to compare population percentage (quantitative measure) of gender and education level (categories). The second reason I choose the radar chart is because my dataset suits the standard data type used in these charts. The category ‘Education Level/Series’ has a cyclical order starting from the lowest which is Lower Secondary School’ up to ‘Doctoral’. These categories are significant beside one another because I wanted to make it clear the difference in the percentage of students that attend each course. It is common to hear that one should not plot multiple categories on a radar chart. In this case, I wanted to plot both males and females on the same chart on top of one another to further emphasis the fact that we are very equal in terms of our student education levels.

Design choices - justify your use of colour, shapes, marks, layout, structure, font, labels

The color is the stereotypical color for male and females; blue and pink. The reason I chose these colors was so it would be immediately clear to the audience that the chart is a comparison of male and female. I choose to fill the shapes with transparent color. The reason I did this was to make it as clear as possible to the reader that male and female education levels are similar. I choose to include category labels. The axis range of 0 through to 100 can also be seen on the chart to aid in the readability of the chart.

Any interactivity or animation and how it helps answer your question

As this graph was plotted using the Plotly python module it is interactive. The audience has the option of selecting male, female or both categories and the chart will update accordingly. This interactivity aids in keeping the viewer engaged in the visualisation. Once again, the comparison between male and female students is further emphasised.

List of tools or libraries used

4. Conclusion [½ page]

Critically analyse the outcome of your visualisation.

Were there aspects that you think could be improved upon?

The data point values are not very easily seen in some cases. This can be seen when looking at ‘Doctoral’ on the chart. The values for Doctoral students were very low compared to the other values. When both ‘Male’ and ‘Female’ are selected it can be difficult to derive any precise values for doctoral. In addition, often a value is actually hidden. For example, if category ‘Male’ is in the foreground of the chart and ‘Female’ is in the background,it will not be immediately obvious that the female data point even exists. Another aspect I felt could be approved upon was the color of the gridlines . In both charts they are white and very difficult to see. This will require the viewer to work harder to derive the exact values of the data points.

Were there effects or functionality that you were technically unable to achieve?

I had planned on adding an interactive slider which would allow the user to see a range of years. This would have been a great tool as I feel the comparison of the 2017 dataset with a much older one from the 1900s would have been very interesting for the viewer. Would have been interesting to contrast the education level of women in the 1990s and 2000s.

References

Include any citation of the dataset

Include links to any tutorial or example that contributed significantly to your work

https://python-graph-gallery.com/392-use-faceting-for-radar-chart/
https://plot.ly/python/polar-chart/
https://plot.ly/python/radar-chart/
https://python-graph-gallery.com/390-basic-radar-chart/

Include any articles or web resources supporting your design choices

https://python-graph-gallery.com/392-use-faceting-for-radar-chart/
http://bl.ocks.org/nbremer/raw/21746a9668ffdf6d8242/
http://worldshap.in/#/IE/
https://www.cso.ie/en/interactivezone/visualisationtools/babynamesofireland/