DATA CLEANING COURSEWORK

PROBLEM 2

When starting to merge the columns I decided to merge each year one at a time: 2020 and 2021 first then 2019. The first realisation is that not all the data frames have the same dimensions and not all the questions are the same in all three surveys. Not only that but some questions are phrased in a different way, or the part/options given are not the same for all. Therefore, we would need to match the columns in a particular way before merging. For the merging for 2020 and 2021 we looked at the column names. Our first initial observation is that up until Q8 all the questions and answers are written the same, so when merging these columns together there is no problem. However, when we reach Q9_Part_1/Q9_Part_11 we can see that in 2020 there is one less part than 2021, as for question 9 in 2021 they gave two different options for Jupyter. Therefore, using this example, the 'None' column in 2020 is Part 11 whereas it would be Part 12 for 2021. To solve this issue, we used a dictionary to map the values that we want to change from 2020 to 2021.

Before doing this, we need to cut down the data frame and make sure to drop any columns not in both data sets. By observation when looking at 2019, I could see that part B was not included like in 2020/2021 so I automatically dropped these columns. We then produced a list of the columns we want to change in 2020 and a list of the columns we want to map to in 2021 and looped through a dictionary function. We also appended empty columns for 2020 for columns that were only in 2021 instead of dropping them for reasons explained in cleaning. At the end we add a year column to each new data frame for each year and finally we concat them.

When looking at 2019, we can see it is quite different from the previous circumstance as the parts are not as similar as the previous data frames. 2019 includes other text answers and different questions from the other 2 data frames. When working with this data frame we decided to drop any columns that had other text and different questions as these were irrelevant to us. Therefore, for this, we manually rearranged the columns according to the order of the 2020-2021 data frame and dropped all other_text columns from the 2019 data frame. Again, we also included empty columns within our list when an option is not available for 2019 but it is for 2020-2021.

The reason why we decided to produce empty columns for the question parts is so that if we have an option available in only one year but not in another this would be merged with the "Other" column as in another year the individual would have selected "Other" if that option was not available. This helps reduce the loss of data between the years. One interesting observation when looking at the 2019 data set is that 32_Part_A questions in 2021 could be considered a combination of parts of Question 34 and 31 from 2019 so I included a unique merge of these in my dataset.

Therefore, after adding the year column for 2019 we then concat it with the 2020-2021 data frame and save this as a csv file. Before cleaning we finish off merging by combining the column parts that aren't an option in all 3 data frames with the "Other" column of the question for reasons explained above aided by my function "other".

PROBLEM 3

When cleaning the dataset, we first want to drop the rows where an individual missed out on less than 5 single questions and drop any duplicates. We also want to replace any empty answers with "missing" and drop any rows where an individual spent less than a minute. Considering the number of questions, we can conclude that some people may just put random answers out of boredom or lack of care and so we can say that the data on these rows would not be reliable. When merging the parts that were not options in all 3 years, we get an answer of a combination of the other row answers. Therefore, we created a function that returns "Other" if that slot is not empty/nan. We also found that some of the single answer columns have the same answers, however written in a different way. This can cause issues when trying to call values and visualise graphs. For example, when looking at a column such as compensation we can see the ranges and signs are different for the 3 years. Therefore, as part of

cleaning I replaced and regrouped the values to match each other. We also do the same with other numerical values with ranges such as experience. Similarly for categorical single columns such as jobs or experience have different names for the same jobs, therefore time was taken to replace these with new names. In the same way with the multipart questions/answers, each column represents a single answer, however, the names do not all match such as Google Colab vs Colab Notebooks. For this we were able to create a function "similar" that replaced all answered slots with the same name and replaced the question (second row) with the new name. Finally, an issue with all columns is that some values/names have gaps or missing letters while others do not, therefore, these were also replaced. Now that we have cleaned, we are able to save the data to a csv file.

PROBLEM 4

Using our new data, we would like to evaluate the top 5 programming languages and top 5 visualisation tools for senior data scientists. When analysing the data, we needed to involve 5 different attributes, year ,experience ,job ,visualisation libraries and programming languages. We first need select only those who selected "Data Scientist" for question 5, and also select only those who put "5-10 years","10-20 years" and "20+ years" for Question 6. After narrowing down the data frame to only include these answers we then split the data frame into 3 different parts: 2019,2020,2021. After producing these we then proceeded to create a function to plot the graphs.
Our function takes in the data frame/year and the question being analysed (when inputting the Question column parts, we did not include the None and Other columns as this is not very useful for our conclusions) and loops through the different column parts to produce 2 arrays: value and count. The function then finds and minimises both arrays to only the top 5 values. This is then used to produce a bar plot of the data using seaborn. Using this, we input Question 7(programming languages) and Question 14(visualisation libraries) into our function along with the combinations of the 3 different years to produce 6 different bar charts as demonstrated below.

Figure 1: Bar Chart of Top 5 programming languages for 2019

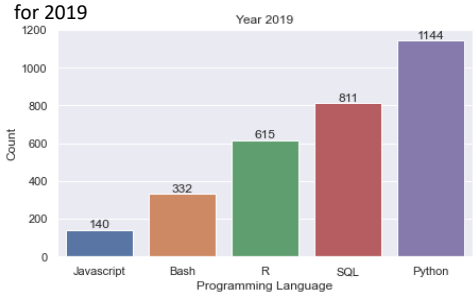


Figure 2: Bar Chart of Top 5 programming languages for 2020

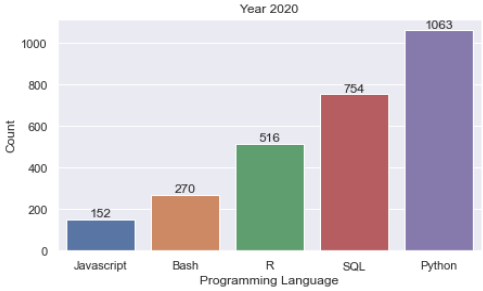


Figure 3: : Bar Chart of Top 5 programming languages for 2021

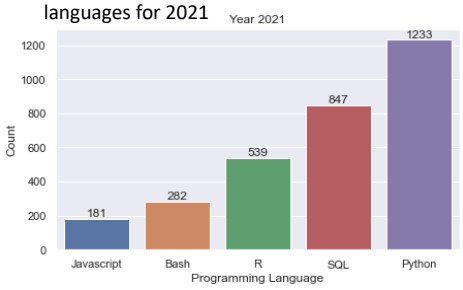


Figure 4: Bar Chart of Top 5 libraries for 2019

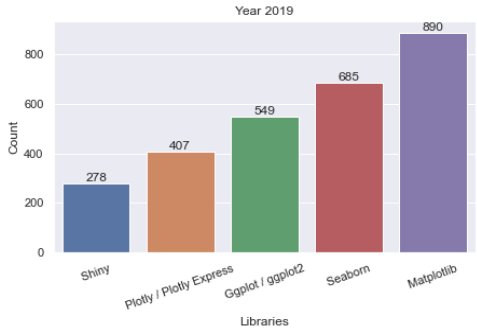


Figure 5: Bar Chart of Top 5 libraries for 2020

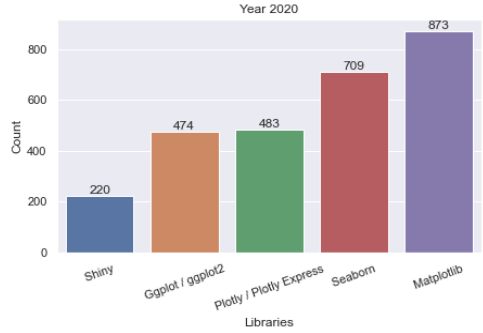
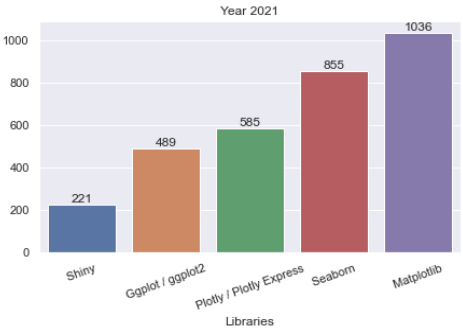


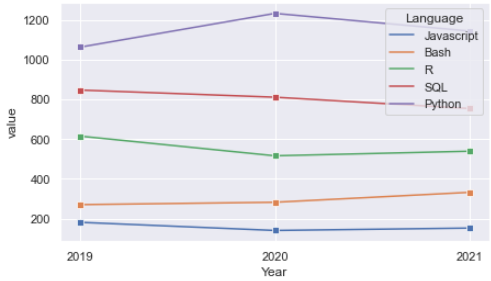
Figure 6: Bar Chart of Top 5 libraries for 2021



From our figures above we can conclude a few things. Firstly, in terms of programming languages, we have found that our top 5 programming languages initially in 2019 are in descending order: Python, SQL, R, Bash, and JavaScript. As the years go on, we do not see this order change however we do see that the number of responses decreases for the top 4 languages in 2020 before increasing again in 2021. When looking at the libraries we see

the order stay the same 2020 onwards but we see Plotly take over Ggplot and increase rapidly in comparison in 2021. We can further demonstrate the progression of languages over the years with a line graph as shown above.

Figure 7: Line graph of Top 5 libraries for 2019-2021



PROBLEM 5

We can also use our data to explain the world-wide situation of Woman in Data Science. To analyse this, we would like most visualisation data to be based on both Gender and Country. In our graphs we would also like to analyse our main variables against other data such as Year of Data, Age, Education, Professional Experience and Salary.

GENDER OVER THE YEARS

Figure 8: Bar Chart of Gender counts 2019-2021

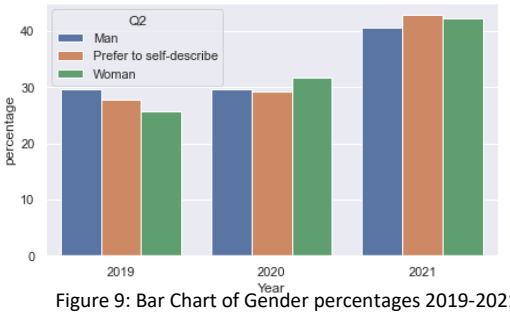


Figure 9: Bar Chart of Gender percentages 2019-2021

Above, we have 2 bar charts, both with an x -axis for year and a key for gender, but different y labels (count and percentage). Our first clear observation is that our total responses are increasing every year regardless of gender, however we can also see the marginal gap from the number of male responses. This can help us get to our first conclusion, that women do not feature in Data Science as much as men, however , from our percentage bar chart, we can also say that women have had the greatest increase in responses over the years.

COUNTRY VS GENDER

Our first initial method for demonstrating women in data science over the world is to find the top and bottom responses from all genders and compare them to the responses from just women. From the data frames created above we can see that the top countries for women does not correlate with the top countries overall. In both cases the majority are from India and the USA, however when only analysing women, we can see that the UK has a better proportion of Women in Data Science, whereas Brazil has a worse proportion. When looking at the bottom few countries and compare the overall data to just the data with women we can see that Romania has a better proportion of women whereas Chile moves to the bottom of the data for women. Now that we have found the top and bottom countries for women in data science, we are able to analyse and draw conclusions.

All Country Counts

	Q3
India	17512
United States of America	7606
Brazil	2118
Japan	2103
Russia	1882
China	1766
Nigeria	1569
...	
Sweden	244
Switzerland	233
Ireland	212
Saudi Arabia	211
Belgium	186
Romania	174
Belarus	163

Women Country Counts

	Q3
India	3599
United States of America	1624
United Kingdom of Great Britain and Northern Ireland	282
Russia	251
Nigeria	249
Brazil	245
...	
Greece	42
Romania	41
Switzerland	40
Sweden	36
Peru	27
Chile	27
Belarus	27
Belgium	22

Figure 10: Chart of all country value counts vs women country counts

As we can see here from the diagram on the left, India has the largest proportion of applicants and women overall, where China is also in second place for number of applicants, however, has a lower proportion of women in comparison. In the diagram on the right we can look a bit closer at the other countries and we can prove our previous point that Romania and Saudi Arabia have a higher proportion of women overall compared to Chile.

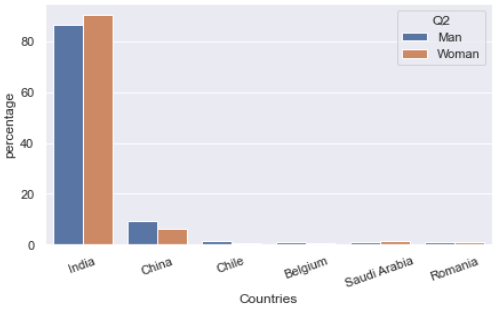


Figure 11: Bar Chart of Country counts Men vs Women

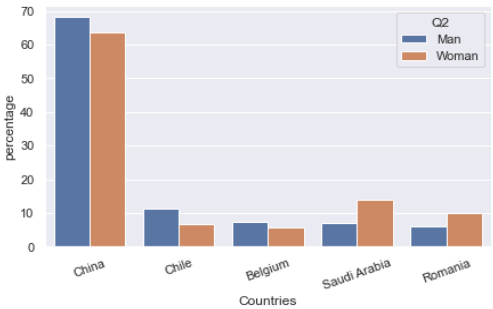
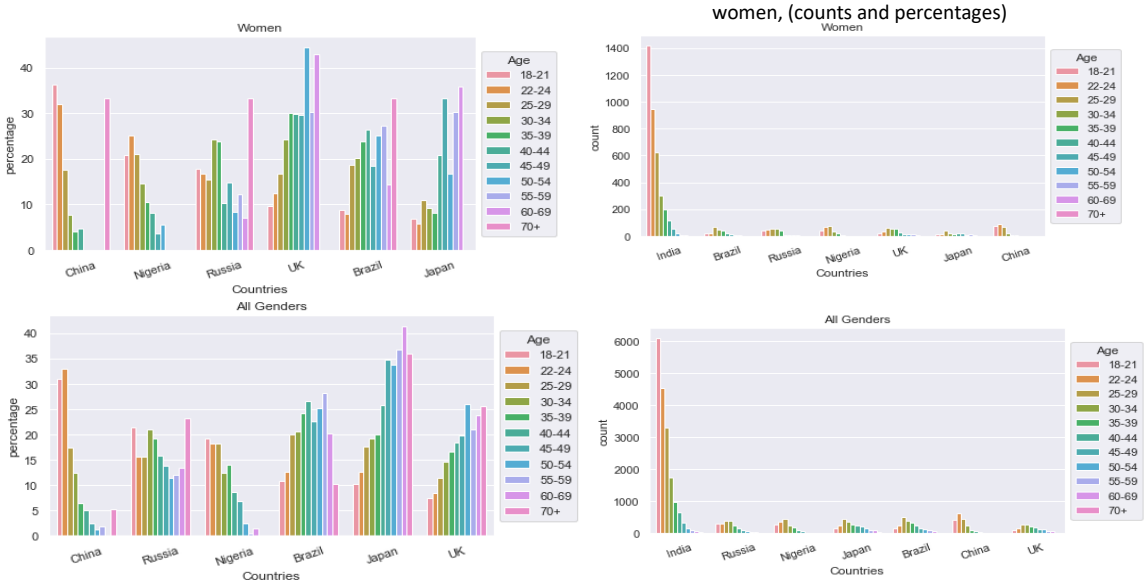


Figure 12: Bar Chart of Country percentages Men vs Women

AGES OF DATA SCIENTISTS OF TOP COUNTRIES

Figure 13: Bar Chart of Country vs Ages for overall and women, (counts and percentages)



From the graphs above we can see that for both women and overall, we still have India as our top country where we can see that as the age range increases, the number of responses decreases. From the original graphs we can see that 18–21-year-olds, supposedly students, contribute to most data scientists/ female data scientists in India. When looking at a closer look between the next top countries, Japanese citizens of ages 60-69 seem to have the highest percentage overall whereas for Women, the age range 50-54 has the highest percentage.

EDUCATION AROUND THE WORLD

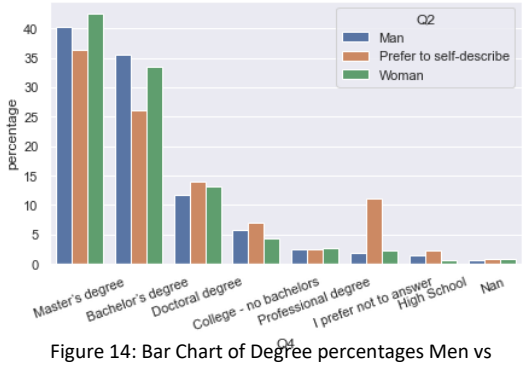


Figure 14: Bar Chart of Degree percentages Men vs Women

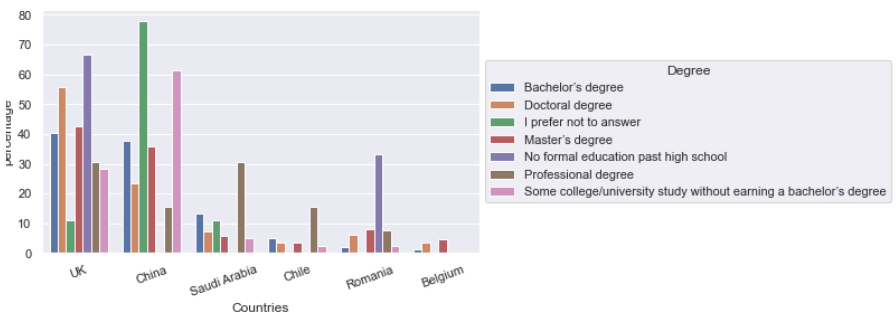


Figure 15: Bar Chart of Country percentages based on different degrees.

From the diagram on the left, though the proportions are similar across genders, we can see that Women hold the highest proportion of Master's Degrees and Doctoral Degrees, which can tell us that Women may be getting the same job as Men even with higher qualifications. From the table of the right we can see that Chile and Saudi Arabia have the highest proportion of Professional degrees for Women, whereas, the UK, Romania and China have highest proportion of lower level qualifications, which may demonstrate the comparison of opportunities for Women in different countries.

Figure 16: Bar Chart of years of experience percentages Men vs Women

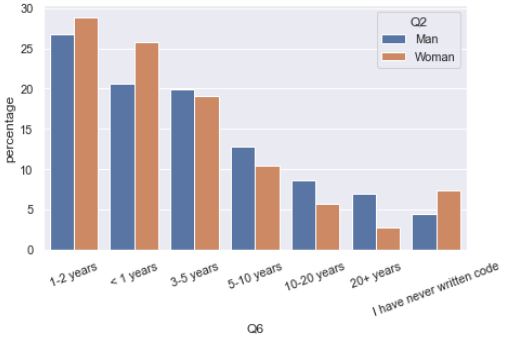
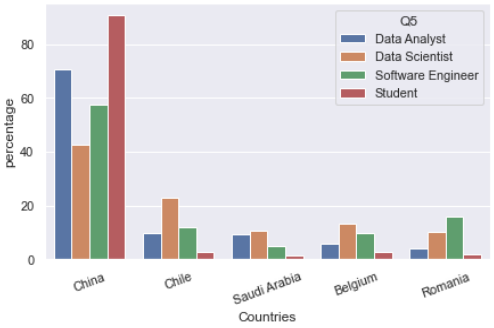


Figure 17: Bar Chart of work experience percentages by Country for Women



EXPERIENCE AROUND THE WORLD

From the table on the left, we can see that Men generally have the most years of experience. When looking at job experience overall for women around the world, we can see in the diagram on the left, that China has the highest proportion of students, whereas Chile, Saudi Arabia and Belgium have the highest proportion of Data Scientists.

SALARIES

Figure 18: Violet plot of salaries by country Men vs Women

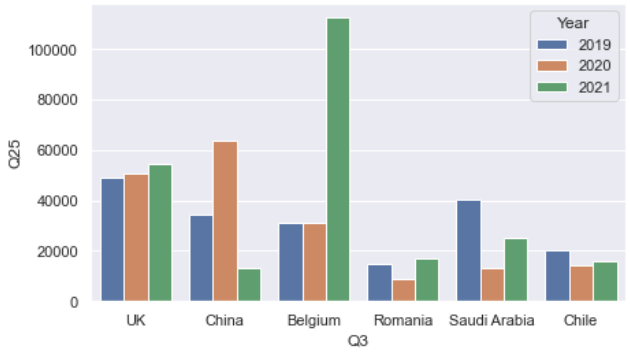


Figure 19: Bar chart of salaries by country for women shown by year

When looking at Salary, as expected, Men and Women have a large pay gap, where China seems to have the smallest pay gap. We can also infer that the highest paying salaries for women and overall comes from European countries. In terms of progression of salary of women over the years, we can see that Belgium has the biggest improvement on salary for Women, whereas China is the opposite, and the UK is virtually the same over all three years.

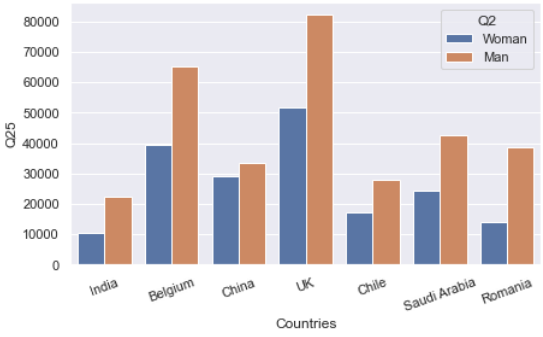


Figure 20: Bar chart of salaries by country Men vs Women

CONCLUSION

We can conclude that there is a male dominance in Data Science from looking into pay gap, differences in qualifications and experience. However, looking through the progression throughout the years we can say that the numbers are improving. Worldwide, there are some countries like the UK and China have a higher proportion of Women, and the qualifications needed are not to the extremes. However, in countries such as Saudi Arabia, we can see that though there are a larger proportion of women than other countries, they generally need higher qualifications, and they receive lower pay.