Data Collection and Cleaning

By Owen Shaw

Introduction

Throughout this report, references shall be made to lines of code in the Python file app.py within the zip folder. This file is split into several functions, named relevant to the question that they are trying to solve. The file is properly commented and a step by step procedure is documented in it.

Q2

The 3 files with the respondents' answers to the surveys had questions that were very similar, but not identical, in each. This was due to many reasons, but most common was due to the change in the data science world over the 3 years. This can be seen in questions such as updating the available Reddit pages relevant to data science (2019 includes the example (r/machinelearning, etc), whereas 2021 includes the examples (r/machinelearning, r/datascience, etc). This is clearly asking the same question, just with slightly different wording. Similarities like this needed to be addressed before merging the datasets.

To do this I wrote an algorithm which generated a list of possible similar questions, which I then manually approved or disapproved. The algorithm compiled all the questions and alphabetically sorted them, as most similar questions were next to each other in the alphabet. The algorithm then compared adjacent elements in the list of questions for a similarity rating, which was the proportion of characters in the same place in each string. If the similarity rating was above a specified value, then those questions were potential similar questions and added to the list of them. The manually approved similar questions were then stored permanently in a .txt file, so they could be later retrieved, and the manual steps did not need to be repeated. For each pair of questions deemed similar, one was assigned to be kept and the other replaced.

Next, I replaced the relevant questions in all 3 data sets. Initially, the datasets had the question numbers and parts (Q34partb) as column names, which were not useful and did not match up very well between the datasets, so these were dropped and the actual question was made the column name, so the merging of the data sets can then be done. In addition to this, a new column was created to denote the year of survey for each row, to determine the year each data entry was collected once the data sets had been merged. The data sets were then merged, the result of which was exported to a .csv file.

Q3

The cleaning of the data was done in 2 stages: cleaning of the whole data set; cleaning of individual attributes relevant to the analysis being done. The reason for this was that there were many parts of the data set that were not going to be used for the analysis, so to maximise efficiency, I only cleaned the data that would be used. However, there was still some universal cleaning that could be done, so that it did not have to be repeated.

To clean the whole data set I first removed all rows that were wholly empty and any that were duplicates, due to the number of questions and entries possible it would be extremely improbable that anyone had an identical response. I also removed any columns that were not going to be useful and may be preventative to the analysis. For example, this included columns that stated that the respondent had written in their own text option instead of selecting from the list of options, which was not going to be useful for our analysis but could have been accidentally included if not removed.

The holistically cleaned data frame is now exported to a new .csv file.

As stated previously, most of the cleaning occurred at a later point and in isolated containers, relevant to the present analysis. When investigating the top 5 programming languages and visualisation libraries, which shall be explained further later, the values for the number of years that the respondent had been writing code or programming were changed, as there was an overlap of different answers from different years' survey, which needed to be fixed. For example: the answers "1-3 years" and "1-2 years" were available in different years, and these had to be changed to fit in a common set of distinct answers.

Q4

To investigate the top 5 programming languages used in 2019, 2020 and 2021 by Data Scientists with more than 5 years of programming experience we had to do the cleaning explained previously and then filter the survey by how many years respondents had been programming for. Next, we had to return the frequency for each language and store them in a dictionary. Since the data was arranged in Yes/No style for each language, we had to count the number of "Yes"s in each column corresponding to programming languages for each particular year, using the "year of survey" column created earlier. The dictionary was then sorted, so that the order on the bar charts that shall be produced is in descending order, for ease of viewing. I then plotted the top 5 data from these dictionaries on to separate figures and added the relevant captions and axes labels. A custom palette was created, matching each language to a colour, so that each language would have the same colour in each year's bar chart. Since the number of respondents varied by year, it would be useful to see the percentage of that year's respondents who selected each language, so these values were added to each bar on the chart. The bar charts for top programming languages for 2021, 2020 and 2019 is displayed below.

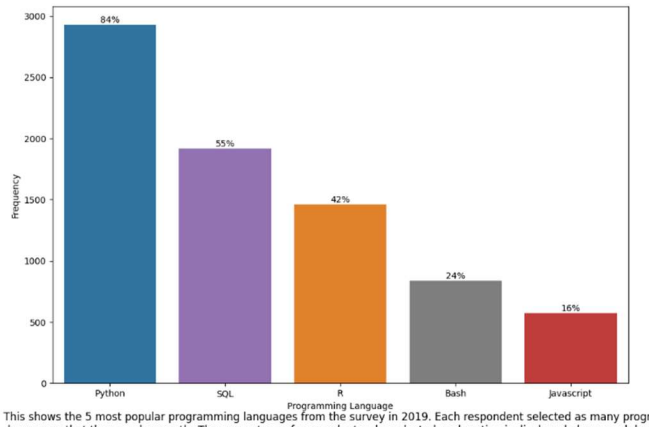

Fig. 1: This shows the 5 most popular programming languages from the survey in 2019. Each respondent selected as many programming languages that they used currently. The percentage of respondents who selected each option is displayed above each bar.

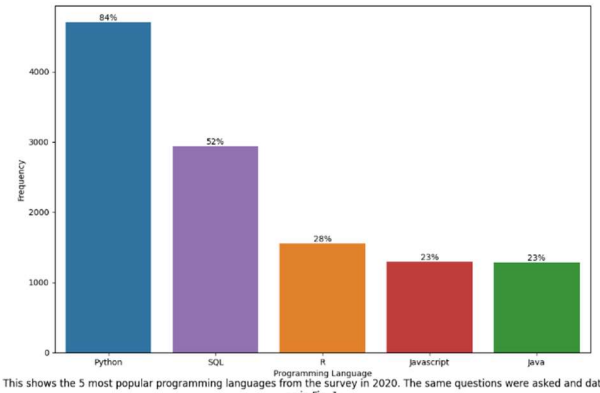

Fig. 2: This shows the 5 most popular programming languages from the survey in 2020. The same questions were asked and data displayed as in Fig. 1.

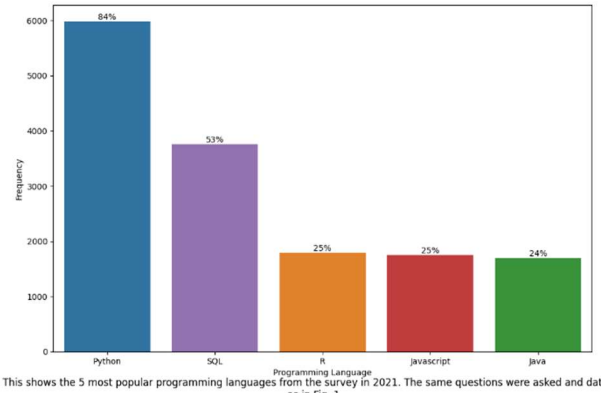

Fig. 3: This shows the 5 most popular programming languages from the survey in 2021. The same questions were asked and data displayed as in Fig. 1

Using the charts shown above, we can conclude several things. Generally, the distribution is extremely similar between 2020 and 2021, with the same order being kept and minor percentage changes being observed. Python and SQL is used by data scientists consistently throughout the 3 years, with Python being used by nearly all of respondents. Secondly, R's usage dropped a significant percentage from 2019 to 2020 and 2021. In addition to this, Bash's popularity decreased significantly, whilst JavaScript and Java increased by a similar amount.

To investigate the top 5 visualization libraries and tools we did the same as the previous task but counted the values in the visualization tools columns instead. The bar charts produced can be seen below.

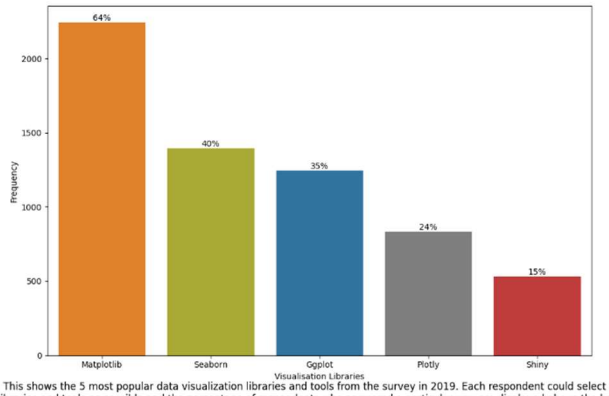

Fig. 4: This shows the 5 most popular data visualization libraries and tools from the survey in 2019. Each respondent could select as many libraries and tools as possible and the percentage of respondents who answered a particular way are displayed above the bars.

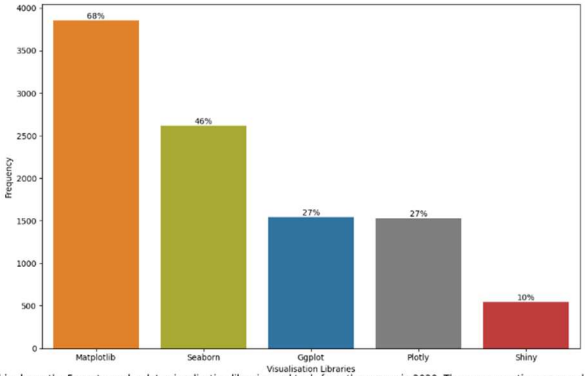

Fig. 5: This shows the 5 most popular data visualization libraries and tools from the survey in 2020. The same questions were asked and data displayed as in Fig. 4.

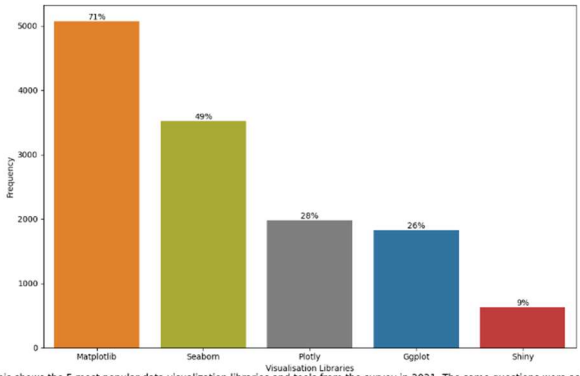

Fig. 6: This shows the 5 most popular data visualization libraries and tools from the survey in 2021. The same questions were asked and data displayed as in Fig. 4.

From the charts above, we can see that the top 5 visualization libraries and tools were consistent throughout the 3 years. Matplotlib is the clear favourite, with its percentage of users increasing by a small amount each year. In second, Seaborn behaved similarly, increasing by a small amount each year. The additional users gained by Matplotlib and Seaborn seems to be due to users transferring from Ggplot and Shiny to the former, as the % of respondents who used these libraries decreased by a small amount each, similar to the opposite change for Matplotlib and Seaborn. Plotly was consistent in usage percentage throughout.

Q5

To explain the world-wide situation of Women in Data Science, we are going to investigate yearly compensation and the highest level of education achieved by men and women.

Compensation Analysis

For the compensation analysis some answers for gender had to be removed and other had to be mapped, as some surveys had “Male” as an answer and others had “Man. To use the yearly compensation data, which was provided categorically, in bins, the answers had to be transformed to numerical data. To do this I replaced the string-type bin with the mean of the upper and lower bound of the bin. This means that further analysis will not be affected by the loss of the bin data, as the mean of the bin can still be used effectively for histograms, overall mean and standard deviation among other statistics. Similar methods were used when using the data on respondents’ highest education level achieved, such as mapping answers to reduce the number of possible answers to make visualization easier.

To visualize the data, we used a histogram which showed both the male and female compensation distribution, overlayed on top of one another. Due to the disparity in the number of respondents that were male and female, the proportion of the respondents is shown instead of the frequency. The histogram of all respondents can be seen below.

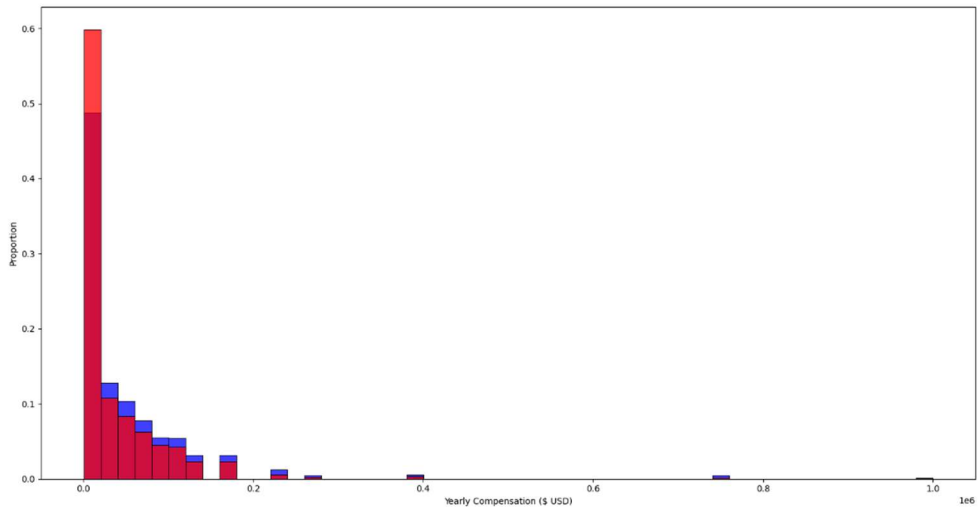

Fig. 7: The histogram above shows the distribution of males (Blue) and females (Red), where the height of each bar is the proportion of each group in that bar's bin range. The orange parts of the histogram indicate bins where the females' proportion is that large, and as such the line between red and orange is the height of the males' bar.

The histogram above [Fig. 7] demonstrates a skew which is even more positive for females than for males. The only bin where there are there is a higher proportion of females than males is the lowest bin, in all others above this, males made up a larger proportion. Males have a mean compensation of $49,500, median of $22,500 with a standard deviation of $84,300; by contrast, females have a mean compensation of $35,300, median of $8,750 with a standard deviation of $62,800. This shows that on average males have 40% higher compensation than their female counterparts and the average male data scientist earns nearly 3 times the average female data scientist. To dive further into this, below are two histograms that split the respondents into high and low earners.

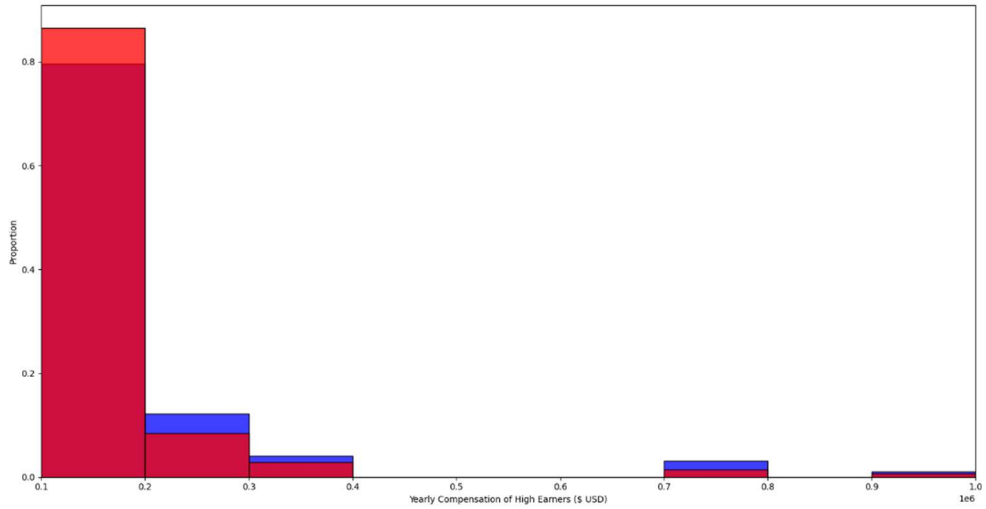

Fig. 8: The histogram above shows the distribution of high earners, which is defined as those earning more than $100,000. The height of each bar is the proportion of high earners in each group in that bar's bin range.

Figure 8 shows the distribution of male and female high earners. As in the holistic view previously, there is a stronger positive skew for females than there is for males, with the only bin that females out number males is the lowest category. Of high earners, males have a mean compensation of $187,000, median of $137,000 and standard deviation of $146,000. By contrast, females have a mean compensation of $166,600, median of $137,000 and standard deviation of $114,000.

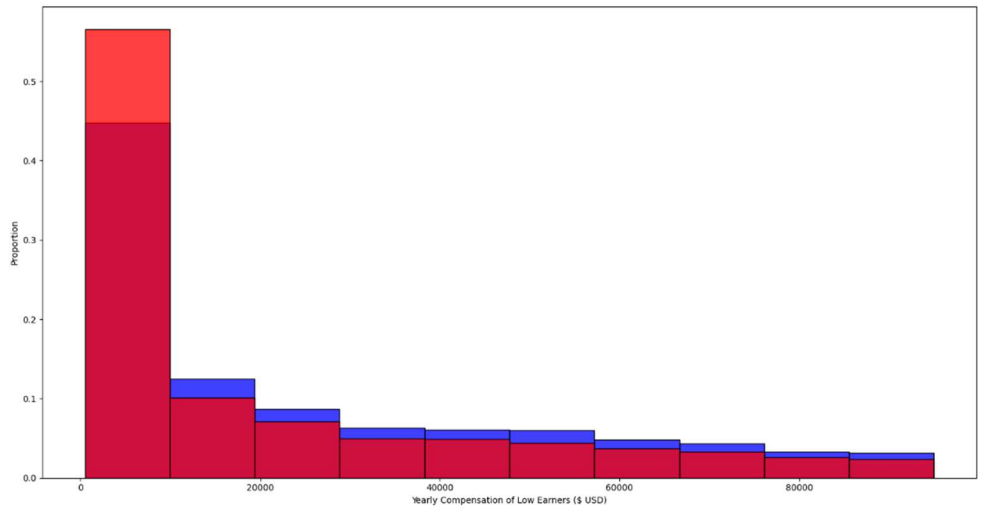

Fig. 9: The histogram above shows the distribution of low earners, which is defined as those earning less than or equal to $100,000. The height of each bar is the proportion of low earners in each group in that bar's bin range.

The above histogram [Fig. 9] shows the distribution of low earners, which is defined as less than or equal to $100,000. As before, females are more positively skewed than their male counterparts and only out number males in the first category. Of low earners, males have a mean of $25,600, median of $12,500 and standard deviation of $28,000. Females have a mean of $20,200, median of $6,250 and standard deviation of $26,500.

From all the previous analyses, it is clear that females are paid less than their male counterparts at all levels.

Education Analysis

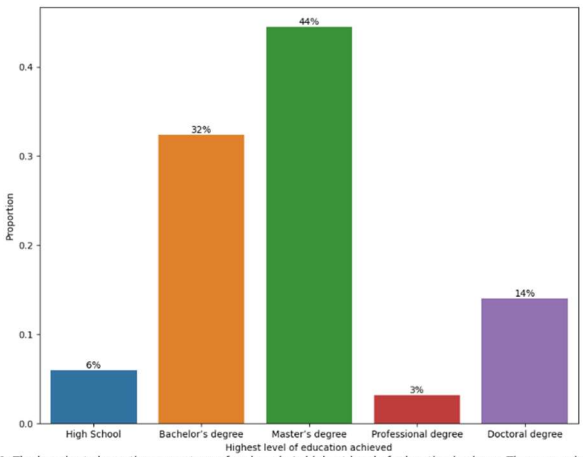

Fig. 10: The bar chart shows the percentage of males who's highest level of education is above. They are ordered from lowest to highest, apart from the professional and doctoral degree bars, which are comparable to each other.

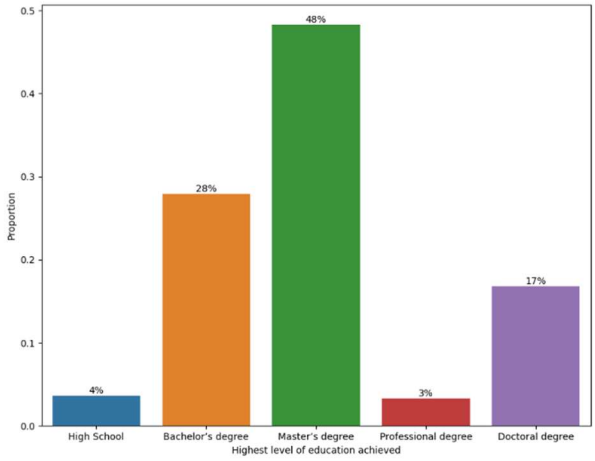

Fig. 11: The bar chart shows the percentage of females who's highest level of education is above.

Above [Fig.10 and Fig.11] show that the distribution of education is very similar between males and females. This was done using similar filtering and mapping techniques as for plotting top programming languages and visualization libraries.

Conclusion

Combining the analyses for compensation and education shows that despite the fact that in data science women are as educated as males, they are compensated significantly less across all levels of skill. This is the current situation for women in data science and action needs to be taken to address this.