Data Collection and Data Cleaning Report

Reporter: knpv42

Problem 2

Final Goal: 1. Merge data from three years together
2. Add in each year's dataframe a new column called "year of the answer"
3. Save the dataset to a csv file called "Kaggle_survey 2019-2021.csv"

1. Data merging
Before concatenating data from three different datasets vertically, we need to make sure that the column headers of each dataframe are the same. Instead of the original header, I set the column headers to be the question text so I can compare the questions between three datasets more easily. [cell 5] Survey question part and options for multiple-answer question part are the two main parts for column integration.

1.1 Survey question integration

Questions observations were done after extracting all questions out from all the column headers. [cell 6] Three issues were detected for survey questions.

The three issues for survey question integration to deal with were:

- Issue 1: Questions which have the same meaning but with different wording across the three years.
- Issue 2: Questions in 2020 and 2021 which are separated by part A and part B are the same question with different wordings.
- Issue 3: Some of the questions don't exist in all three years' datasets.

Issues 1 and 2 were solved together. To get enough data from all three years' datasets, instead of only preserving all questions that have the same wording, first I looked through the "kaggle_survey_2020_answer_choices.pdf" and "kaggle_survey_2021_answer_choices.pdf" files which were extracted in problem 1 to rename all the part A and B questions together. The reason why I merged part A and B together in 2020 survey and 2021 survey is that: first, 2019 survey doesn't have two parts questions; secondly, grouping part A and part B won't cause data collision because respondents won't fill in both parts at the same time since they would only be given one according to the schema; thirdly, we can use data science methods to filter out respondents with different experience levels later, so the division of the questions is not necessary.

After combining questions with different divisions, I filtered out all questions that are different between years [cell 6 – 8; cell 12-14] and found the questions that have the same meaning. I also used the pdf files mentioned before for better observation, comparison, and validation, because I can compare the options under the questions to validate whether they are the same questions. For 2019 survey, I can only look through the excel file for comparison and validation. After that, I mapped the questions that have the same meaning and renamed them to the same wording [cell 9-10; cell 15]. As for the renaming choices, I went with the 2021 wording because they are more accurate and considerate.

For issue 3, I dropped the questions which don't exist in all three years' datasets because having only one- or two-year's data for a problem is not enough for a 3-year-based data analysis. [cell 11; cell16]

After all the data merging, I checked the disjunctive union of the three year's question sets to make sure there aren't still questions in all three surveys that are different [cell 18]. After the cleaning of survey questions, there are 34 distinguished questions remaining in the survey [cell 40].

1.2 Options under multiple-answer questions integration

There are mainly three observed issues for option integration for multiple-answer questions.

- Issue 1: Options that doesn't exist in all three datasets.
- Issue 2: Options that expressed the same thing but with different wordings.
- Issue 3: Questions that have misspellings or weird formatting.

For Issue 1, I filtered for all the column headers in each year which share the same question, compared them, and found the options that only exist in one or two years. After that, I renamed those columns to the ‘Other’ option for better data preservation. For different multiple-answer questions, I used different methods. For those questions which don’t have many options that need to be moved to ‘Other’, I renamed them manually [e.g.: cell 22]. For questions that have a lot of options to change, I put all the option names that need to be changed in a list and used a for loop to select and rename them [e.g.: cell 28, cell 32], or I used set theory to remove the difference between one set of questions and the other two [cell 33].

For Issue 2, I filtered for all the column headers in each year with the same question, compared them and found options with different wordings but the same content in all three years and renamed them. If the options have a slight change between different years, then, like in question integration, I usually go for the wording in 2021 because I tend to think that researchers changed them for a reason.

For Issue 3, typical misspelling errors, such as misuse of ‘O’ and ‘0’, were fixed by renaming [cell 34]. Apart from that, while investigating the options I realised some of the questions have abundant whitespace, so I fixed those at the start [cell 21].

2. Addition of “year of the answer” column

After integrating the questions, I added a new column in each year’s dataframe filled with the current year using a basic pandas method. [cell 19]

3. Save merged dataset to csv file

After data merging, I used the “groupby” method to combine questions with the same column header together so that all three years’ datasets shared the same column headers [cell 39], and then I concatenated the dataframes vertically [cell 40]. Next, I stored the dataframe to 'Kaggle_survey 2019-2021.csv' [cell 41].

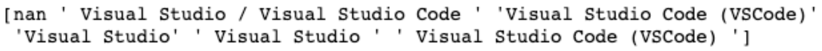
Problem 3

Final Goal: 1. Clean the dataframe’s formatting
2. Clean individual invalid data
3. Save the dataset to a csv file called “Kaggle_survey 2019-2021_cleaned.csv”

1. Clean the dataframe’s formatting

Certain data formatting issues need to be handled by observing all unique values in the dataframe resulting from problem 2 [cell 44] and the csv file generated by problem 2.

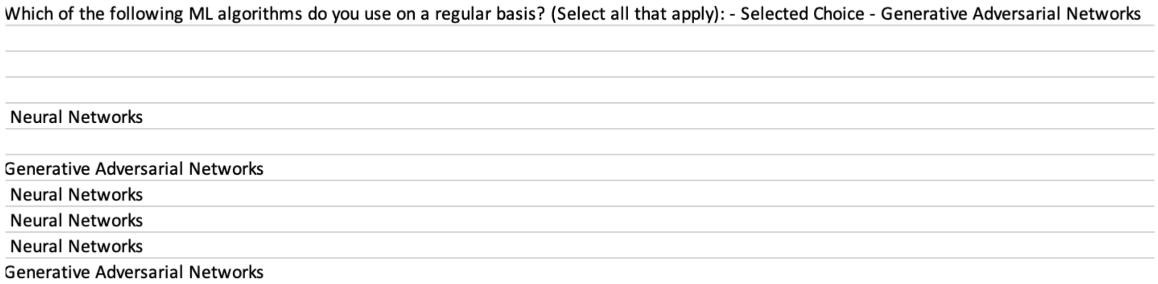
- First, trailing whitespace in some values needed cleaning [e.g.: Image 1].



```
[nan ' Visual Studio / Visual Studio Code ' 'Visual Studio Code (VSCode)'
 'Visual Studio' ' Visual Studio ' ' Visual Studio Code (VSCode) ']
```

Image 1: Values with trailing whitespace

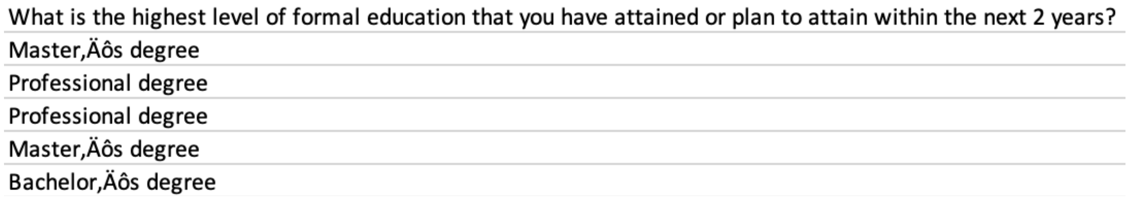
- Second, a lot of values under multiple-answer questions do not correspond to the column after pandas “groupby” method in problem 2 [e.g.: Image 2].



```
Which of the following ML algorithms do you use on a regular basis? (Select all that apply): - Selected Choice - Generative Adversarial Networks



 Neural Networks

Generative Adversarial Networks
 Neural Networks
 Neural Networks
 Neural Networks
Generative Adversarial Networks
```

Image 2: Incorrect values under multiple selected choice questions in 'Kaggle_survey 2019-2021.csv’ file

- Third, some spellings of values are incorrect, which will cause unreadability [e.g.: Image 3].



What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
Master‚Äôs degree
Professional degree
Professional degree
Master‚Äôs degree
Bachelor‚Äôs degree

Image 3: Unreadable values in 'Kaggle_survey 2019-2021.csv' file

1.1 Data cleaning for trailing whitespace

The method for handling this problem was learned from the recommended material from the lecture [1]. Notice that because there are a lot of 'nan' values in the dataframe, when we apply the clean_string function we need to skip those values [cell 45].

1.2 Renaming incorrect values under multiple-answer questions

To select all the multiple-answer questions, I used a filter function from pandas to find all the corresponding columns. Apart from that, I stored the option names (such as 'Python', 'C' etc.) from the columns I found as a set called 'options'. Next, when iterating through all multiple-answer questions, all responses to the relevant question were replaced by the correct value [cell 46].

1.3 Correcting misspelling issues

After observing the csv file generated by problem 2, I found two targeted columns to deal with and replaced the incorrect punctuation with the correct one. [cell 47]

2. Clean individual invalid data

There were several issues to deal with in this section.

- First, move options under single-answer questions which don't exist in all three years to 'Other'
- Second, correct phrasing differences between years under single-answer questions
- Third, rename some options with lengthy wording
- Fourth, drop respondents that didn't answer most of the questions

2.1 Single-answer question values cleaning

The first three issues I found are handled together. The reason is that the method I used for problem 3 part 1 would not work for single-answer questions, so they required more manual work.

I looped through the columns to find the unique values [cell 44] and picked out the lists matching single-answer questions, then found the question using the excel searching tool. For each question, I checked all the answers per year within the dataframe and found out the disjunctive union of all three years' answer sets. For answers that are the same meaning but with different expression, I rename them to be the same [e.g.: cell 53]. For answers that had lengthy wording, I shortened them [e.g.: cell 57]. For answers that did not exist in all three years, I changed them to 'Other' [cell 49-50].

2.2 Drop respondents that didn't answer most of the questions

After observation, to target these respondents, I found an indicator column "For how many years have you been writing code and/or programming?" which indicates the respondent didn't finish most of the survey if the question is left empty and dropped all rows if the answer for this question is left empty [cell 62].

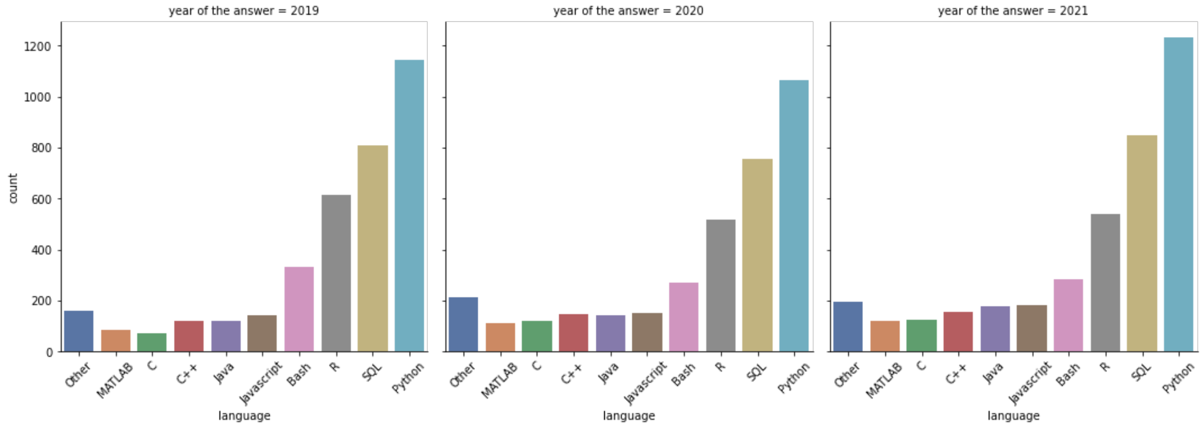
3. Save cleaned dataset to csv file

After all data cleaning, I saved the cleaned dataset to 'Kaggle_survey 2019-2021_cleaned.csv' [cell 63].

Problem 4

Based on the data in 'Kaggle_survey 2019-2021_cleaned.csv' generated by problem 3, Graph 1 shows the top programming languages used by senior data scientists in the survey each year [cell 66]. There are three subgraphs, which represent the situation from 2019(left), 2020(middle) and 2021(right) respectively. The following conclusions are made according to the graph:

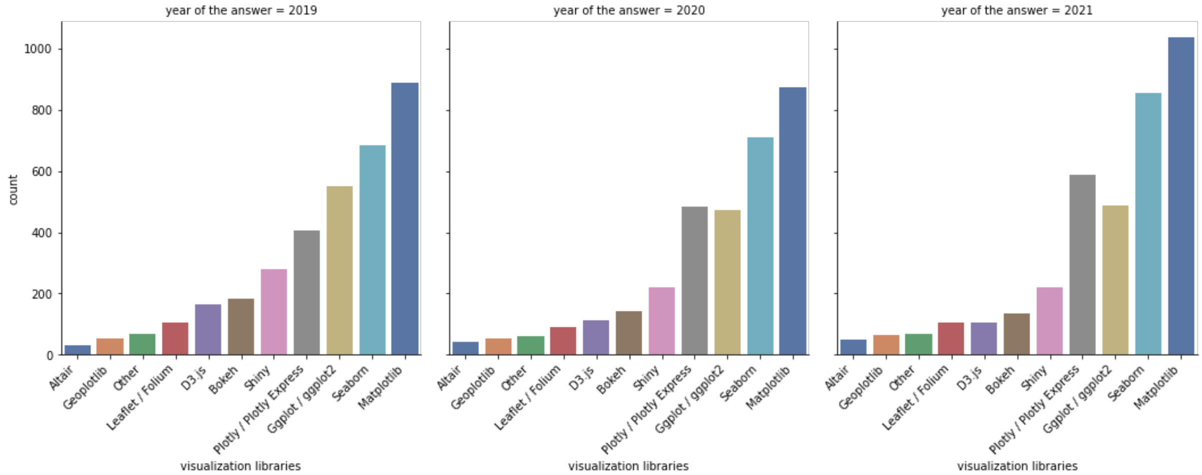
1. The top five programming languages used by senior data scientists in the survey are the same from 2019 to 2021. They are, in order: Python, SQL, R, Bash, and JavaScript.
2. Python continuously is the most commonly used language among all three years.



Graph 1: Top 5 programming languages used by senior data scientists from 2019 to 2021

Graph 2 indicates the top visualisation libraries and tools used by senior data scientists in the survey each year [cell 67]. There are three subgraphs, which represent the situation from 2019(left), 2020(middle) and 2021(right) respectively. The following conclusions are made according to the graph:

1. Even though the sequence has slightly changed in different years, the top five visualisation libraries and tools remain the same. They are, in order by 2019: Matplotlib, Seaborn, Ggplot/ggplot2, Plotly/Plotly Express, and Shiny.
2. Matlpotlib and Seaborn are always the top 2 most popular choices for senior data scientists in the survey every year.



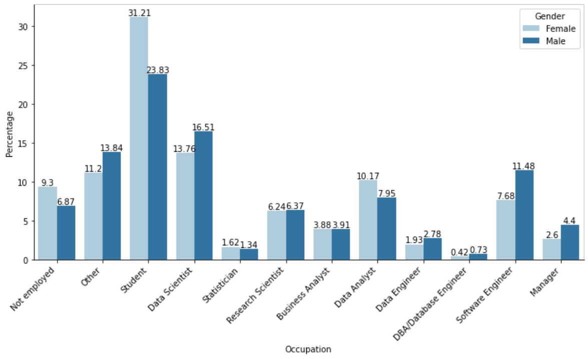
Graph 2: Top 5 visualisation libraries/tools used by senior data scientists from 2019 to 2021

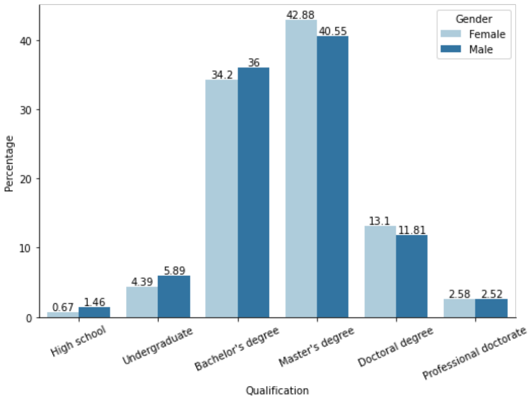
Problem 5

The number of respondents to this survey varies considerably in terms of gender; the amount for other gender options (e.g.: 'Nonbinary', 'Prefer not to say' etc.) apart from male and female were too little, so I didn't take them into consideration. Moreover, there were four times as many male respondents as there were female respondents [cell 70]. Therefore, it is more meaningful to take data on a particular characteristic for some gender as a proportion of total respondents of that gender, rather than just displaying a count.

The characteristics of the world-wide situation of women in data science can be concluded by the following aspects:

1. In terms of occupation, compared to male data scientists, there are relatively fewer female data scientists in senior roles. According to Graph 3, the proportion of engineers and managers is greater in the male data scientists' population than in the female data scientists' population, whereas the proportion of unemployed or students is greater in women [cell 71].

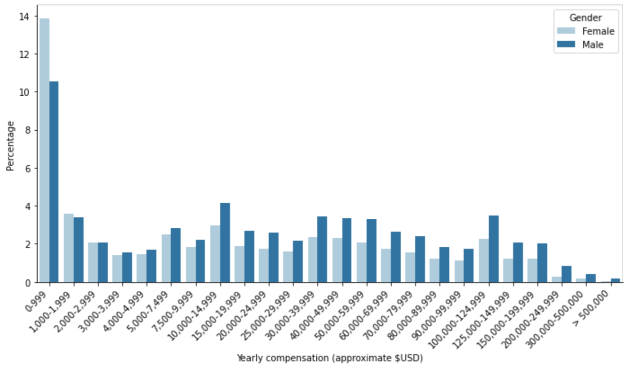


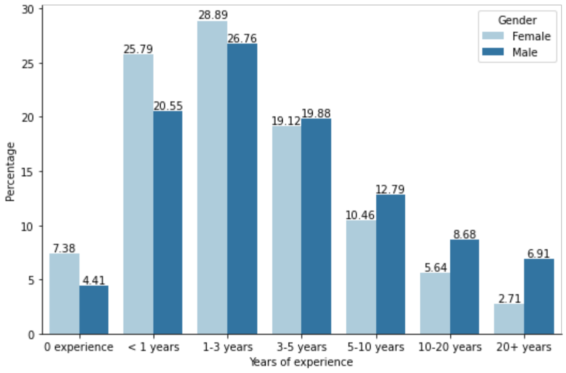


Graph 3(Left): The proportion in different occupations of women in data science compared to men's
Graph 4(Right): The proportion in different qualifications of women in data science compared to men's

2. In terms of qualifications, there is little difference between the educational levels of men and women. This, as shown by Graph 4, means that there is no inherent advantage for men and women in this sector. Although there may be more male data scientists in terms of quantity around the world, they are not necessarily better than female data scientists in terms of quality [cell 72].

3. In terms of yearly compensation, the proportion of female data scientists who have high salaries is much smaller than that of male data scientists. For example, in Graph 5, the percentage of women in the two lowest salary categories is higher than the percentage of men, and the percentage of female data scientists with higher salaries is lower men's [cell 73].





Graph 5: The proportion in different yearly compensation of women in data science compared to men's
Graph 6: The proportion in years of experience of women in data science compared to men's

4. In terms of years of experience, according to Graph 6, more than one in five female data scientists have less than one year of experience, while a larger proportion of male data scientists have more experience. From this we can also deduce that the rate of growth in the number of new entrants to the industry is higher for women compared to men [cell 74].

Overall, male data scientists currently have a greater overall advantage in terms of salary and position in the world. However, we can also see that more and more female data scientists are entering the industry and are comparable to men in terms of qualification. Hopefully, in the near future, working in data science will not be different because of one's gender.

Reference:
[1] https://www.kaggle.com/code/docxian?scriptVersionId=53917973&cellId=25