Data cleaning

Problem 1: Data Merging:

Before we jump into the merging data process, we first check the general information of data. So generally, we have three data about Machine Learning & Data Science Survey from Kaggle in 2019,2020 and 2021. And to make process efficient and organised, we made few steps to merge the data:

1. Analysis each dataset and find the connection between each of them.

- **Kaggle survey 2019**:
  The dataset contains four parts:
  - multiple_choice_responses.csv: Contains main data information but lacks details for the OTHER_TEXT column and all Q14_Part_TEXT. For example: Q2_OTHER_TEXT, Q14_Part_1_TEXT.
  - other_text_responses.csv: Contains the details in OTHER list text and all Q14_Part_TEXT.
  - questions_only.csv: Contains questions name and the number of each question.
  - survey_schema.csv: the number respondents in each question.

  So the way to merge 2019 dataset, we need to check the following:

  - ***Find the connection between other_text_responses.csv and multiple_choice_responses.csv and find a way to merge them together.***
    Since there is no more detail explanation of this dataset, so one assumption made is: The number appearing in the 'multiple_choice_responses' indicates the times valid entry is in 'other_text_responses.csv'. And for better understanding here is an example in Figure 1. As 'cplx' is the second value appear in OTHER_TEXT, so it response as 1 in 'multiple_choice_responses', need to mention that the first value appearance response to 0.

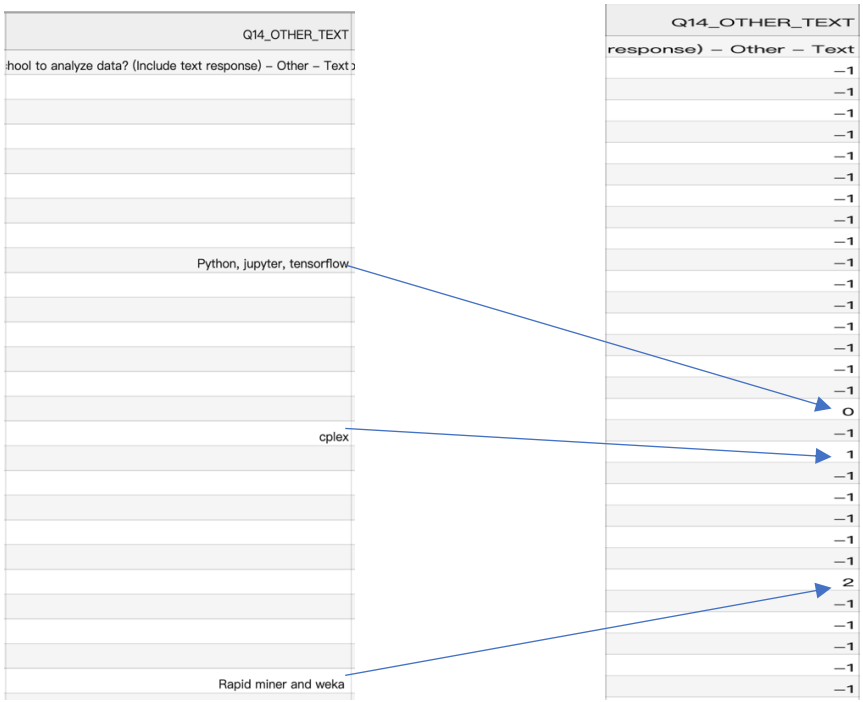

Figure 1: Example about the assumption

- **Kaggle survey 2020 & 2021**:
  This table only contains one file. And we notice that there is no valid or useful information in OTHER_TEXT column.
- **General:**
  Before we merge tables together, we checked the name of each question and found that there is one question has different expression, shown as in Figure 2.

2019	2020	2022
Q14: 'For how many years have you been writing code and/or programming?'	Q6: 'For how many years have you been writing code and/or programming?'	Q6: 'For how many years have you been writing code and/or programming?'

Figure 2: Similar question

2. Merge each dataset separately according to same part of question.

After we have good understanding on the dataset and find the connection for each of them, it's time to merge them together and here are few steps we made.

- **Kaggle survey 2019**:
  a) Change the format of 'other_text_responses.csv' so the row line number indicates to the times valid entry.
  b) For survey in 2019, we merge different part question together into one column but except Q14 as lack of details in some parts, will do it after we have full details.
  c) Use 'other_text_responses.csv' to fill in the main table's column.
  d) Merge different parts of Q14 as one column.
- **Kaggle survey 2020&2021**:
  Remove the OTHER_TEXT columns. Then do the merge as in 2019.
- **General**

Replace the name of the question shown in Figure 2.

3. *Merge three dataset together.*

After merging table separately, it is time to merge them together, and here is the steps we made.

- Find same question name through three tables, removed other questions.
- Merge three dataset together

Problem 2: Data Cleaning

In the data cleaning, this is what steps we made:

- According to the 'kaggle_survey_2021_methodology', we cleaned data as followings:
    - Remove the row with 'Time from Start to Finish (seconds)' less than 2 minutes (120 seconds).
    - Remove the row with over 10 'NAN' values. (Similar percentage as methodology made)
    - Group the number of appearances for each country less than 50 as 'Other'.
- According to the 'kaggle_survey_2021_answer_choices', we cleaned data as followings:
    - Group answers not in the range in the 'kaggle_survey_2021_answer_choices'' as 'Other'.
- Group duplicate answers together.
- Group answers with the same meaning but different expressions together.
- Group the answer appears less than 5 as 'Other'.

Problem 4:

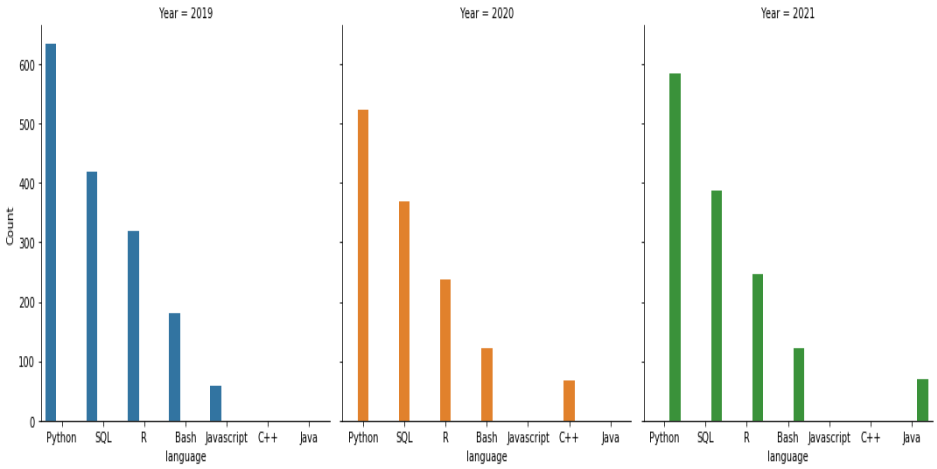

Figure 3:

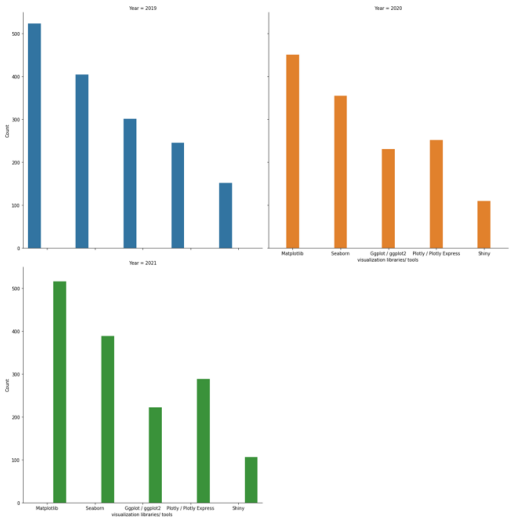

Figure 4:

Conclusion:

1. Python, SQL, R and Bash is top 5 languages used by the senior (more than 5-year programming experience) Data Scientists during 2019 to 2022. But for 5th place it is changing all the time, 2019 is JavaScript, 2020 is C++ and 2021 is JAVA. Shown as in Figure 3.
2. Matplotlib, Seaborn, Ggplot/ggplot2, Ploty/Ploty Express and Shiny is the top 5 visualization libraries/ tools used by the senior (more than 5-year programming experience) Data Scientists. Shown as in Figure 4.

Problem 5:

We applied background and yearly salary to analysis the situation in Women in Data Science:

- Background: age distribution and how many years programming language experience
- Top 5 Yearly salary

1. Age distribution:
   It is clear to see from the Figure 5 that from 2019 to 2021, the number of women aged 18 to 24 will grow gradually, and the number of women in 2021 has increased significantly in 2021. It means more and more women are willing to join the data science relent career.

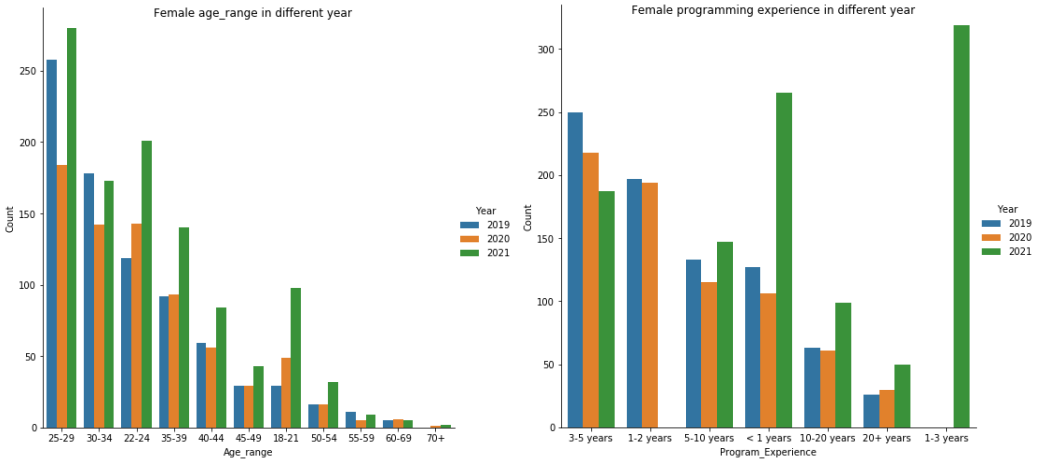

Figure 5: Female age range in different years

2. Programming experience:
   It can be demonstrated from the Figure 5 that more women have more than 5 years' experience compared with 2019. Also, it is need to mention that there is a significant progress from 2020 to 2021 about more people has less than one year experience. One possible assumption is more and more women bee accepted and more and more women willing to join the data science related career.

3. Top 5 yearly salary:

   Figure 6 shows the top five salary from 2019 to 2021, and it demonstrates that from 2019 to 2020, more women got basic salary which is range 0 to 999, which could be the reason why from 2019 to 2020, the number is growing but not significant. But from 2020 to 2021, there is new salary range came in top5 which is 7500 to 9999, in the meantime, more women joins data science career. But unfortunately, the salary in total is not an ideal number, if wants to attracts more women joins, increasing salary could be a possible solution.

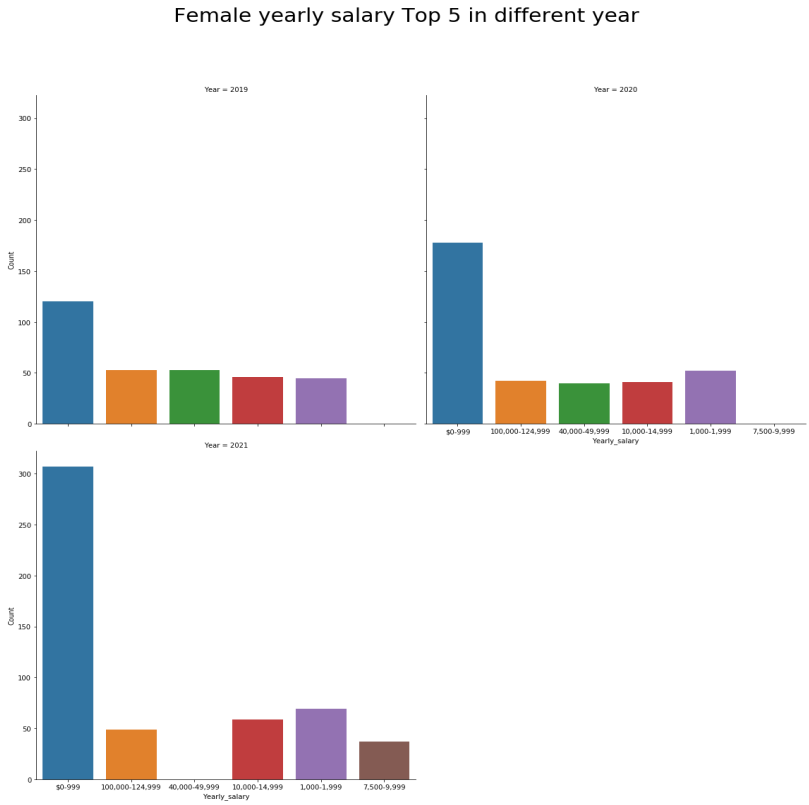

Figure 6: Female Top 5 salary in different years

In general, more and more women joining and having interests in data science, and the salary is going better, but still lots of women got low payment. In future should increase the salary to attract more women to join.