Data Collection and Cleaning Report

[redacted]

Date: 30/04/2022

[redacted]

[redacted]

1 Introduction

IN this report, I will simply be outlining the method in which I have solved problems 2-5, related to the data collection and data cleaning of three Kaggle surveys.

2 Problem Solutions

2.1 Problem 2

To begin with, I first read in the four surveys that I had extracted through Problem 1, making sure to assign a dataframe to each of them.

To merge the questions from the three surveys, I analysed the questions that were shared by going through the survey schemas and question titles. During this process, I was able to compile together a preliminary mapping, in order to detail the questions shared across the three surveys, for computational processes.

From my findings, I was able to disregard many questions. This involved the exclusion of the 2019 "other_text_responses.csv" file, since the "other" responses were already included in the "multiple_choice_responses.csv" file, and the text responses were not recorded in a readable format for the 2020 and 2021 surveys.

Once I had analysed the datasets, I created a function which added questions from the datasets, one by one, and all their parts to an array for each subsequent year. To correctly merge the three array sets together, I had to ensure that the questions were placed in the correct order, along with each of their parts. I devised a system to cumulatively calculate how many columns our final dataset required, through iterating through two datasets at a time, and finding the maximum number of parts per question. Additionally, this left an array which indicated the order in which to merge all the questions. I used this to sort the three datasets into their correct order, however I had to manually tweak the order of some of the subquestion parts.

Once this was complete, I simply created the "year of the answer" column and removed the headers from the 2020 and 2021 surveys. At this point, I was finally able to merge the three datasets together, which occurred successfully.

I further merged columns which answered the same question, if they hadn't been merged thus far, before saving this to a csv file.

2.2 Problem 3

To carry out the data cleaning process, I began by removing any irrelevant data – the first of which included extra columns I had created when solving the previous problem. I followed this up by removing the text input questions, which I mentioned previously, since their data was stored in an unreadable format.

I decided to reformat the data question by question, as I felt this was the best approach to ensure that nothing would be missed. Also, whilst analysing the columns within our dataset, it did appear to me that many questions were already clean, such as all the uses of quantile binning found throughout.

There were many differences between the choices available for each of the surveys. I began by ensuring that the references to the same gender were consistent; for example, I replaced all instances of "Man" to "Male" and "Woman" to "Female". I also changed any instances of "or more" to ">".

Another issues I had to correct were the character errors within the corrupted "highest qualification" answers, to ensure that the data was readable as "Bachelor's degree" and "Master's degree".

Instead of maintaining long sentences, I thought it was important to keep detail yet ensure that data was concise, readable and in a good usable format – unlike the text-based answers that I previously removed. Hence, I reword some data pieces in a consistent manner, and removed some notions to money and time, which are stated in the question.

A large part of the data cleaning I carried out was to change the format of the questions that were asked that allowed many answers to be selected. Their original format simply repeated the choice that was made, in a particular cell. Instead I turned this into a binary table, where a 1 indicates that a choice was made, and a 0 indicates that the variable was not chosen. This change would significantly help in data analysis.

Since there were many incomplete cells of the data frame, I removed all the unfinished surveys, or all the rows which contained empty values.

The last part of my cleaning of the data was to change some of the headers, for consistency, and correctly align the "year of answer" column. This was done right before exporting the newly cleaned table to another csv file.

2.3 Problem 4

To visualize the data, I reduced the survey to only include data from senior data scientists.

I then filtered this further by year and printed the sum of each column, in order to produce the graphs below.

The first set of graphs demonstrates the top 5 **programming languages** used by senior data scientists.

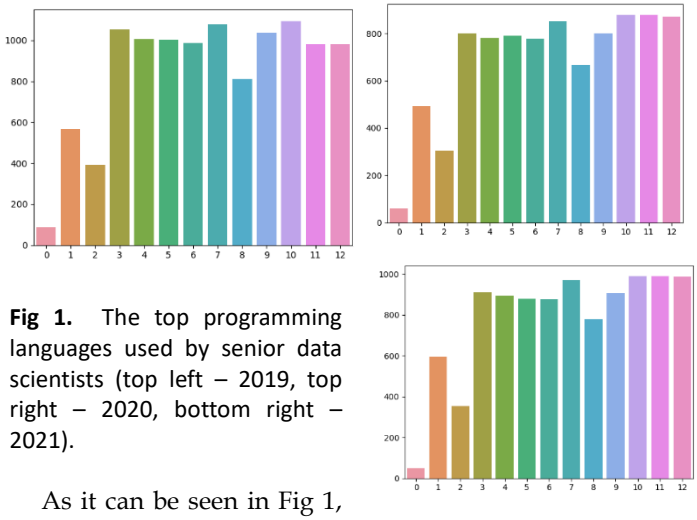

Fig 1. The top programming languages used by senior data scientists (top left – 2019, top right – 2020, bottom right – 2021).

As it can be seen in Fig 1, the top programming languages include 3, 4, 7, 10 and 11 which refer to C, C++, TypeScript, other languages and no languages.