Practical Demonstration in Cleaning a Large Dataset [Ward Dib – 20/05/2021]

This report will show an examples of cleaning different types of fields in a large dataset. The attached dataset contains results of a jobs and income survey conducted in 2019. [1]

We start by dealing with the data frame as a whole first, then move onto cleaning column by column. A methodical approach ensures the data is clean without much data loss.

There are multiple techniques for cleaning text fields to prepare for further analysis. Figure 1 shows a snippet of what the raw data looks like.

	If you want to add your salary to this spreadsheet, go here: https://www.askamanager.org/2019/04/how-much-money-do-you-make-3.html	Unnamed: 1	Unnamed: 2
0	Timestamp	How old are you?	What industry do you work in?
1	4/24/2019 11:43:21	35-44	Government
2	4/24/2019 11:43:26	25-34	Environmental nonprofit

Figure 1

To start, the header needs to be dropped. And the column names are the questions in the survey, which is inconvenient for analysis. So we will rename them. We will drop the last column containing job descriptions, as it won't be useful for statistical analysis.

Then, we start cleaning column by column. From the "Timestamp" column, we will only keep the year, as that's most relevant to our analysis.

We also need to remove or replace null values. Missing data in the salary, occupation, or location columns need to be dropped, as these are the core focus of the analysis. Others could be replaced with manageable values in the analysis such as "other" or "none".

Next is the industry and occupation columns, which we will remove special characters from. The salary column contains many special characters such as currency symbols, brackets, and commas. These will be removed to facilitate numerical analysis.

Using summary functions, we found 125 entries of hourly salary. Since the survey asked for annual salaries, we will drop these, as there's no way of knowing how many hours a person worked, or how many days a year.

Cleaning the location column depends on what the required analysis is. In this report, we are interested in salaries in the 50 American states.

industry	occupation	salary_annum	currency
Government	Talent Management Asst. Director	75000	USD
Environmental nonprofit	Operations Director	65000	USD
Market Research	Market Research Assistant	36330	USD
Biotechnology	Senior Scientist	34600	GBP
Healthcare	Social worker	55000	USD

Figure 2

Thus, we will create a copy dataset with just USA entries. We will include state capitals to minimise any data loss, then split it between state and city.

After these processes, 6632 entries were deleted. Now we may proceed with analysing the areas we're interested in.

Using the "value_counts" function, we determined the most common industries and occupations. The top 3 being technology, healthcare, and higher education.

We can find the average salary in these fields by creating a pivot table around a selected value. For example, the average annual salary in the American technology industry is $107,731. We can also use "groupby" to find out the average salary in certain industries depending on state or city.

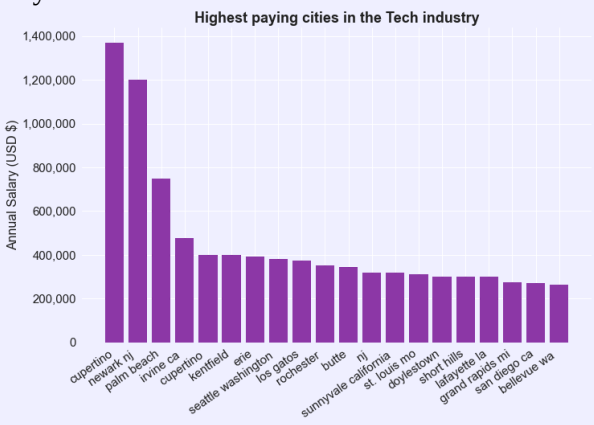

Figure 3

These are examples of the kind of analysis that can be performed on this kind of dataset. As it contains a wealth of information, countless analysis can be performed to reach various results.

---

1. Ask A Manager Salary Survey 2019 (Responses) [Internet]. [cited 2021 May 20]. Available from: shorturl.at/iDJ58