**Data Cleaning and Transformation of Data Science Job Posting on Glassdoor Dataset in Microsoft Excel**

**Introduction**

Imagine you’re working in a data analytics firm and have been handed over an uncleaned dataset containing information on job listings on an online job portal. You have been assigned to clean and transform the dataset concerning “Data Scientists” job listings to make the dataset more useful for analysis.

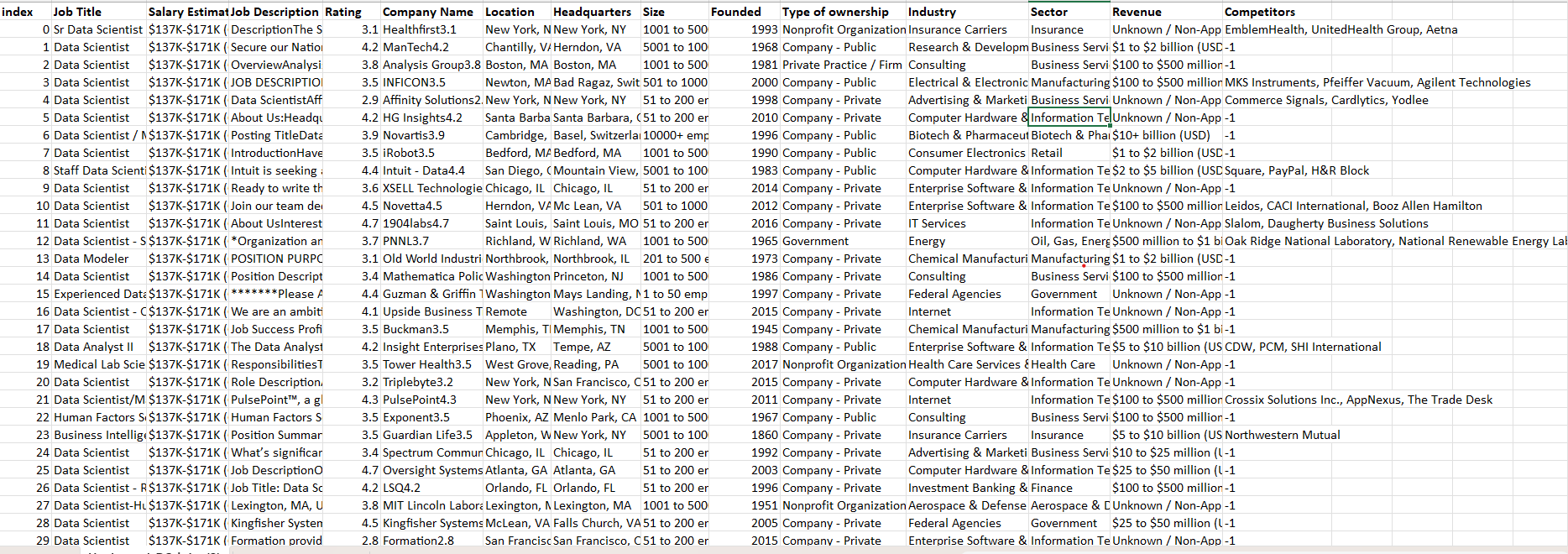
Here, I will walk you through the step-by-step process I followed in cleaning and transforming the [dataset](https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor?select=Uncleaned_DS_jobs.csv) on the data science job posting on Glassdoor in Microsoft Excel.

The data on the job posting is available in [Kaggle](https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor?select=Uncleaned_DS_jobs.csv) in a CSV file.

**Understanding the Data**

I loaded the data from the CSV file into the Excel workbook.

While exploring the data, it was established that it has 14 variables/columns and 672 observations/records.



SS 1: Preview of Dirty Data

I named the sheet containing uncleaned data as raw\_data. Created another sheet and renamed it as “working\_sheet\_1”. Then, I copied the dirty data to the newly created sheet to clean and transform it. I follow this procedure of having untouched dirty data in one sheet in case something goes wrong during the cleaning process, and then I will have a backup.

I then went through all the columns to structure, logically break down the whole process and specify the direction I wanted to follow to make the data-cleaning process as efficient and accurate as possible.

I followed the following steps to clean the data while keeping in mind at each stage that I want to use it to gain insights into the "Data Scientist" job listings.

While going through the data cleaning and transformation process, I will explain the process as much as possible for better understanding and replication.

**Step 1: Index Column**

I have no use for the index column; therefore, I deleted it.

**Step 2: Job Listing**

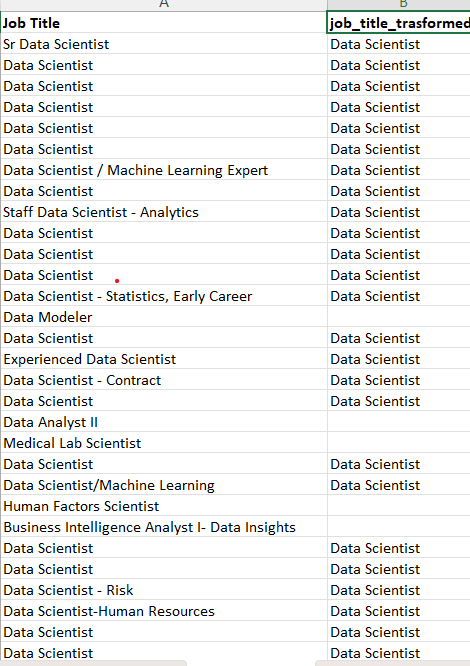
The job listing column contains information about many types of data-related job listings. As I explained earlier, I wanted to clean the data to only gain insights about the "Data Scientist" listings.

Therefore, I decided to clean/transform the "Job Listing" column into a new one that only shows information on Data Scientists.

To do so, I used the Excel formula:

**=IF(ISNUMBER(SEARCH("data scientist", Cell)), "Data Scientist", " ")**

The output was saved in a new column named “job\_title\_transfomed” as shown by the screenshot of the workbook:

****

**SS 3: Job Title and job\_titile\_transformed**

Then I filtered the “job\_titile\_transformed” by blanks and deleted all the records where job\_title\_transfromed is blank.

**Step 3: Salary Estimate**

The “Salary Estimate” has information on the maximum and minimum salary for each listing of data scientist jobs. I want to calculate the average salary, for which I need information on the maximum and minimum salaries in separate columns.

The first number in the cells in the columns is the minimum salary, while the second is the maximum salary. For extracting the information and loading it into separate columns, I used the below Excel formulas:

**For minimum salary:**

**“=VALUE((MID(C2,2,SEARCH("K",Cell)-2)))”**

**For maximum salary:**

**“=VALUE(MID(Cell,SEARCH("-",Cell)+1,SEARCH("K",Cell,SEARCH("-",Cell))-SEARCH("-",Cell)-1))”**

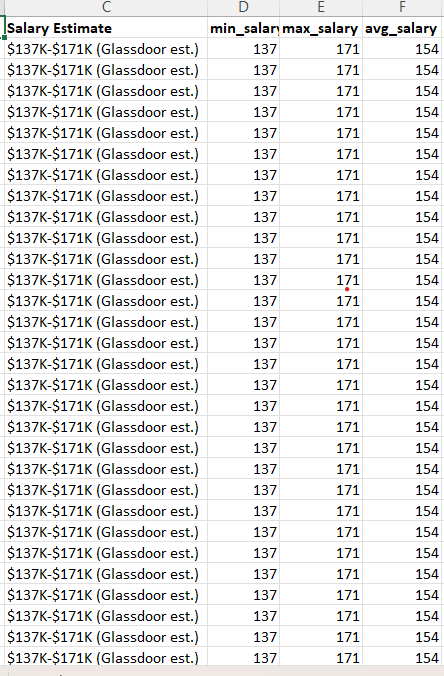
The column containing information on minimum salary was named "min\_salary”, while the column containing information on maximum salary was named as max\_salary.

My next objective is to calculate the average salary for each job listing. To do so, I used the below Excel formula

**Average Salary:**

**“=(min\_salary + max\_salary)/2**

The calculated columns were named avg\_salary. The “Salary Estimate” and min\_salary, max\_salary and the avg\_salary are shown:



**SS 3: Salary Estimate, min\_salary, max\_salary, and avg\_salary**

**Step 4: Job Description**

The "Job Description" column contains information about the job description and the qualifications and skills required for a job listing. I want to know about the skills that are being demanded/preferred by the companies.

I am interested in extracting information about skills such as Python, SQL, AWS, Hadoop, Spark, Tableau, and Power BI to gain insight about, let's say, its relation to salary.

I used the following Excel formula to extract and load the information on each skill in separate columns.

**Python:**

**=IF(ISNUMBER(SEARCH("python", Job Description)), "1", "0")**

**SQL: =**

**IF(ISNUMBER(SEARCH("sql", Job Description)), "1", "0")**

**AWS:**

**=IF(ISNUMBER(SEARCH("aws", Job Description)), "1", "0")**

**Hadoop:**

**=IF(ISNUMBER(SEARCH("hadoop", Job Description)), "1", "0")**

**Spark:**

**=IF(ISNUMBER(SEARCH("spark", Job Description)), "1", "0")**

**Tableau:**

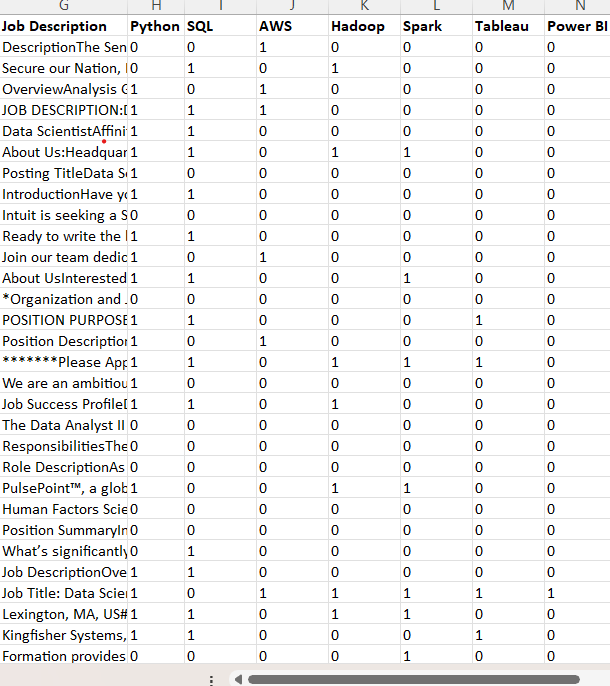
**=IF(ISNUMBER(SEARCH("tableau", Job Description)), "1", "0")**

**Power BI:**

**=IF(ISNUMBER(SEARCH("power bi", Job Description)), "1", "0")**

Here, 1 means the skill is required, while 0 means the skill is not mentioned in the job description.

The job description and the newly constructed columns are shown in the screenshot of the workbook:



**SS 4: Salary Estimate, python, SQL, AWS, Hadoop, Spark, Tableau, Power BI**

**Step 5: Ratings**

While exploring the data in the ratings columns, I came across "-1" values. However, the range of the rating values is from 1 to 5. Therefore, the rating cannot be -1.

It should be further investigated whether the data is incorrectly entered or whether it is missing data. Since no further information is available, I will assume that the data has been incorrectly entered as -1 instead of 1.

I filtered the column and replaced -1 with 1.

**Step 5: Company Name**

Looking at the column “Company Name” it was noted that at the end of each observation, there is a rating value which should be removed.

The rating at the end of the company name column is removed using the split-to-column function in the Data tab, choosing delimiter, choosing other and entering Ctrl + J.

The original column, along with the output, is shown by the screenshot of the workbook as:

A screenshot of a computer

Description automatically generated

**SS 5: Company Name, company\_name\_transformed, to\_drop**

**Step 6: Location**

The location column in the dataset has information on the job location. However, the column contains information on the job city and state/country, delimited by a comma. The next objective is to separate the job city and job state/country information into separate columns.

For this purpose, I used the text-to-column function in the Data tab and comma as the delimiter.

The screenshot of the original location column and the new columns is shown by the screenshot below:

A screenshot of a table with names

Description automatically generated

**SS 6: Location, job\_city, job\_state/country**

**Step 7: Headquarters**

The headquarters column contains information on the city and state/country of the company's headquarters.

While exploring the column "-1," values were observed, which seemed counterintuitive since a city or state cannot be -1. I filtered the column by "-1" deleted the minus values, and left the cells blank.

Next, I want the city and state/country information in separate columns. I again used the Text to Columns function in the Data tab. The comma was used as the delimiter.

The original column and the new columns are shown as:



**SS 7: Headquarter, headqtrs\_city, headqtrs\_state/country**

**Step 8: Size**

The “size” column contains information on the minimum and maximum number of employees in a company. While exploring the column, I came across -1 values. Since the size of a company cannot be -1, I filtered the column by “-1”, deleted “-1” and left the cells blank.

I want information on the minimum and maximum number of employees in separate columns. To do so, I use the following Excel functions.

**Minimum\_employee**

**“=IFERROR(VALUE(LEFT(Size, FIND(" ",Size) - 1)), "")”**

**Maximum\_employee**

**“=IF(ISNUMBER(FIND("to", Size)), MID(Size, FIND("to", Size) + 3, FIND("employees", Size) - FIND("to", Size) - 3), "")”**

Then, I compared data in the “Size” column and the newly created columns by first filtering the max\_employees by “blanks” and then looking at the data in the “Size” column. I noted that there were values such as "10000+ employees" and "unknown" while the corresponding cells in the max\_employees were blank, as shown in the workbook screenshot below.



**SS 7: Size, min\_employees, max\_employees**

**Solution:**

**“10000+ employees”:** In cells having 10000+ employees, I entered 10000. To do so, I filtered the “Size” column by “10000+ employees”, then entered 10000 in the first cell of the max\_employee, and populated all the cells below.

**“unknown”:** Since the value is unknown, I left the corresponding cells in max\_employee blank.

A screenshot of a computer

Description automatically generated

**SS 9: Size, min\_employees, max\_employees**

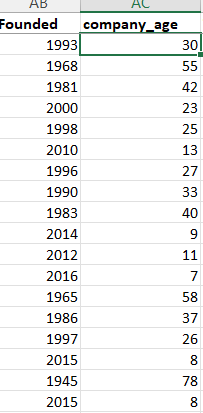
**Step 9: Founded**

The Founded column contains information on the year a company was founded. Furthermore, while exploring the column, I observed “-1” values. Since a year cannot be -1, therefore it seems that the data needs to be corrected. I filtered the column by “-1” and deleted it, leaving the cells blank.

I want to transform the "Founded" column and get information about the company's age in a new column named company\_age. To do so, I used the following Excel formula.

“=IF(LEN(Founded)=4, YEAR(TODAY())-Founded, "")”

The “Founded” column and the newly created “company\_age” columns are shown by the screenshot(SS) as:



**SS 10: Founded, company\_age**

**Step 10: Type of Ownership**

The "Type of Ownership" column contains information about the company's ownership, i.e., whether the company is privately owned or government, etc.

While exploring the column, we came across a -1 value, which is odd since the type of ownership cannot be -1.

The cells containing -1 were made blank, and the output column was named as ownership by using the Excel conditional formula

**=IF(Type of Ownership="-1", "",Type of Ownership)**

**Step 11: Industry**

The “Industry” column contains information on the industry in which the company is operating.

While exploring the column, I again came across -1 values. Since the name of an industry cannot be -1, it should be removed.

The cells containing -1 were made blank, and the output column was named industry\_cleaned again using an Excel conditional formula

**=IF(Industry="-1", "",Industry)**

**Step 12: Sector**

The “Sector” column contains information on the sector of the economy in which the company operates.

While exploring the column, I again noted -1 values. Since the sector's name cannot be -1, it should be removed.

The cells containing -1 were made blank, and the output column was named sector\_cleaned once again using an Excel conditional formula:

**=IF(Sector="-1", "",Sector)**

**STEP 13: Copying and Paste the data from worksheet\_1 into a new worksheet**

I copied all the columns (ctrl + a) from worksheet\_1 and pasted them as values in a newly created worksheet\_2 for further processing.

The idea behind creating and further processing the data in worksheet\_2 is to have a record and a kind of benchmark. I don't have to start from scratch if I make any mistakes.

**Step 14: Dropping Unwanted Columns**

Next, I deleted unwanted columns: **Job Title, Salary Estimate, Job Description, Company Name, to\_drop, Location, Headquarters, Size, Founded, Type of ownership, Industry, Sector.**

**Step 15: Renaming Columns:**

I renamed the columns as:

1. **job\_title\_trasformed to job\_title,**
2. **company\_name\_transformed to company\_name**
3. **industry\_cleaned to industry**
4. **sector\_cleaned to sector**

**Step 16: Rearranging Columns**

I rearranged the column to make the table more presentable.

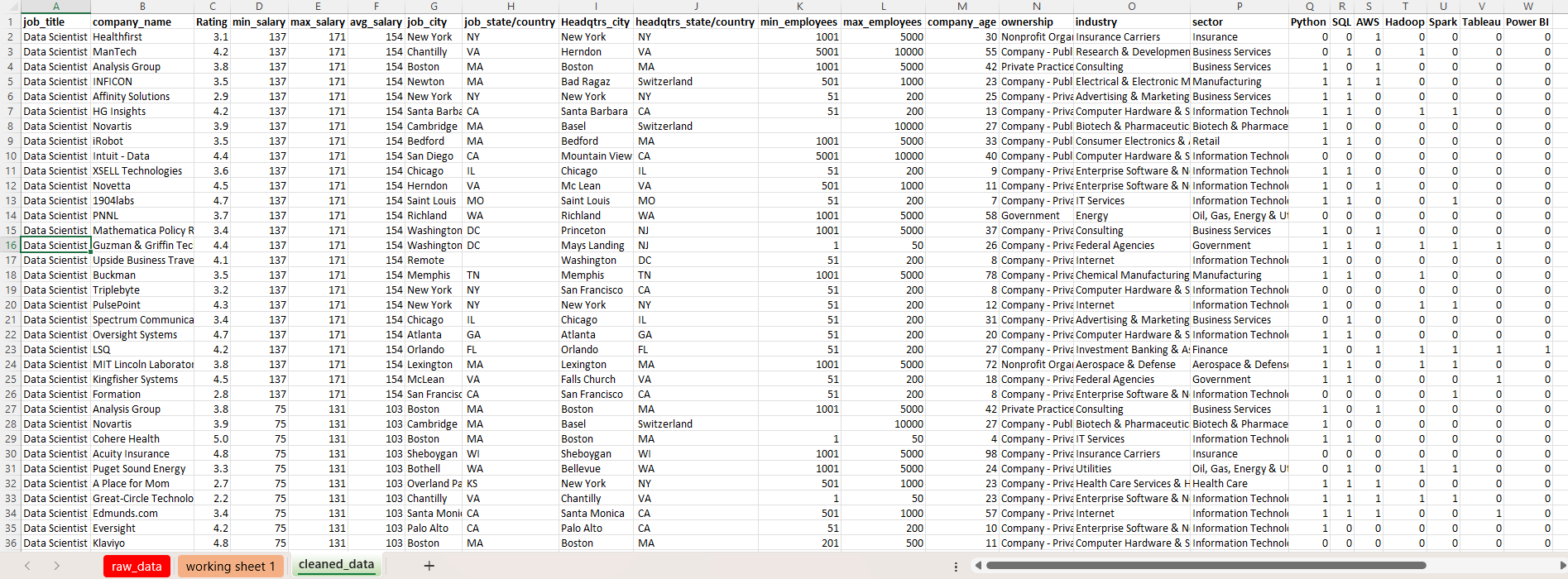
**Step 17: Formatting Columns**

In this step, I formatted all, chose the appropriate format and applied it to the columns.

**Step 18: Checking of Duplicates**

Lastly, 12 duplicates were found and deleted using the Delete Duplicates function in the Data tab.

The data is now ready for analysis.



SS 11: Cleaned Data Set

This is how we cleaned and transformed dirty data. If you have a comment or suggestion, please reach me.