# **Data Cleaning Project: Data Analyst Job Roles in Canada (Indeed Data)**

**Overview:**

This is a data cleaning project done solely on Excel, to showcase my excel skills. Please note I have not done much analysis on this one, rather I focused on using various Excel functions to prepare the data for further analysis.

**Data Set:**

**Data Analyst Job Roles in Canada**

**Source**:

<https://www.kaggle.com/datasets/amanbhattarai695/data-analyst-job-roles-in-canada?select=Raw_Dataset.csv>

**About the Data set:**

With 8 Columns, 1797 rows this data set represents the job listing in Indeed Canada.

The columns are as follows before cleaning:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Job ID** | **Job Title** | **Company Name** | **Language and Tools** | **Job Salary** | **City** | **Province** | **Job Link** |

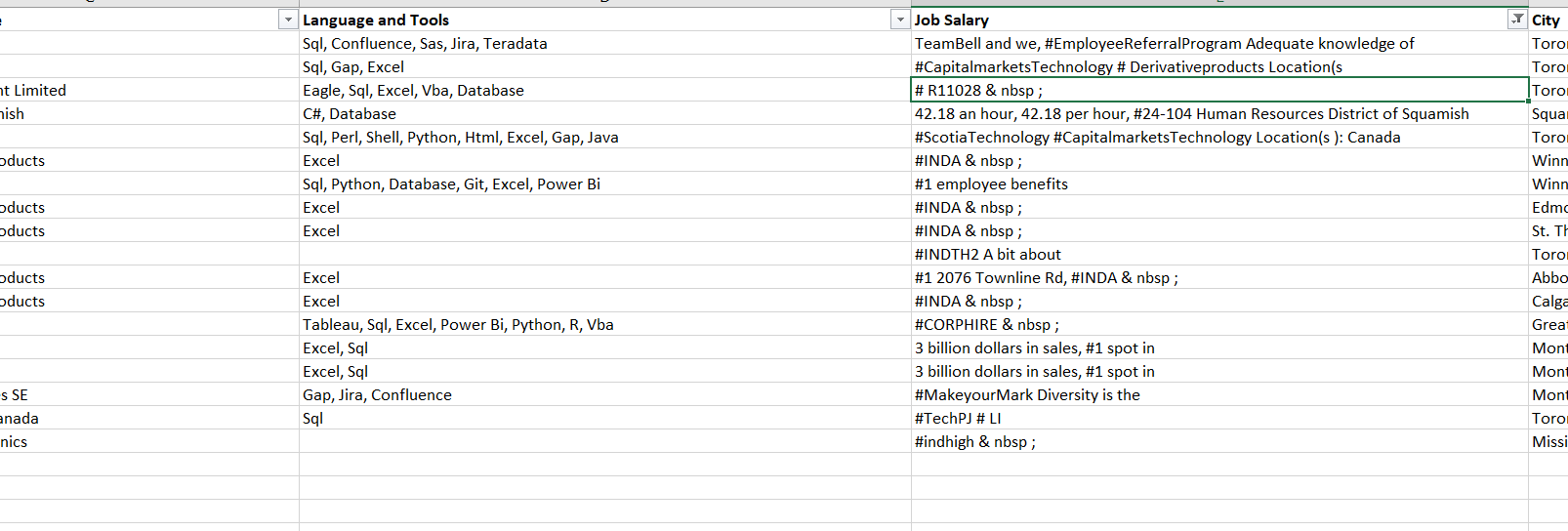
After cleaning up, The final dataset has columns and 1150 rows.

**Steps I followed for the Cleaning:**

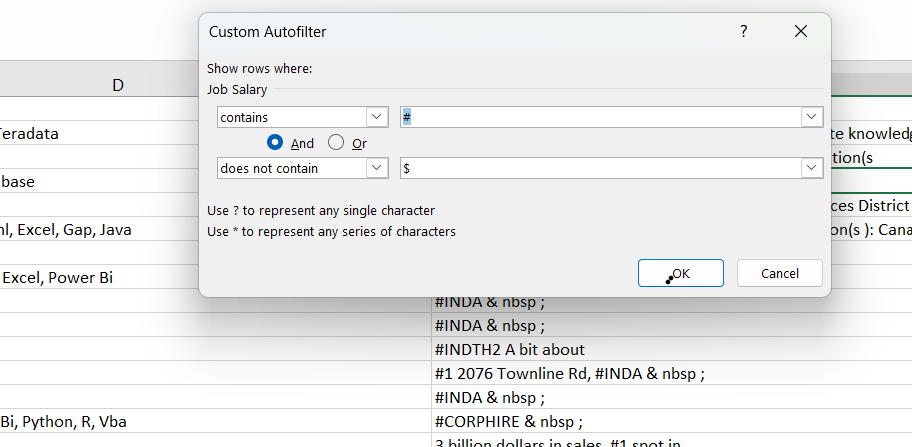
1.Applied “**Remove Duplicate**” feature in excel, to delete any duplicate data, which in this case were none.

2.Since, this dataset is about job listings, salary is a key information. I encountered several listings with blank values in the **Job Salary** Column. I decided to remove those rows as well. There were 558 numbers of blank values!

3.Now even in the non-blank columns of **Job Salary** Column, there were records with nothing regarding salaries.



I filtered these with “Custom text filter” as

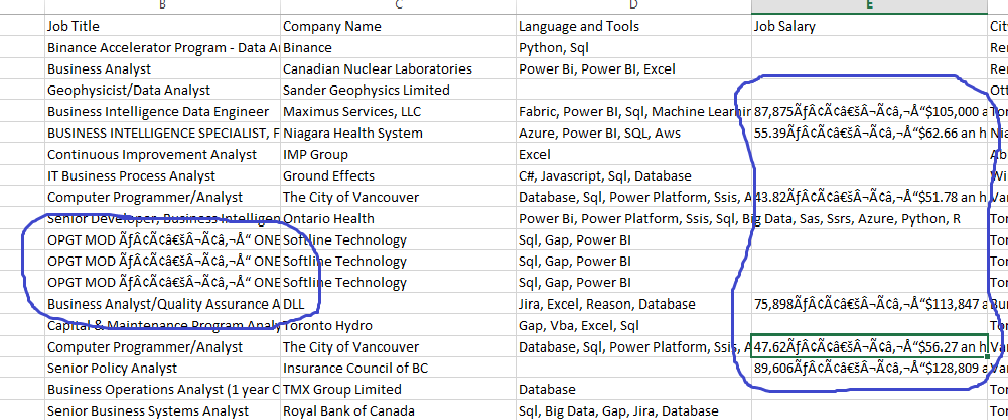


Since I have seen that most of them contained “#”

Then I deleted those rows as well except the one that contained valid data.

Again, I applied text filter for those records that do not contain “$”. The result were 138 rows and mostly they had valid data. Since the numbers of rows were fewer, I manually deleted some rows that did not have salary information.

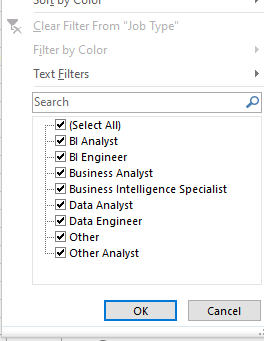
4.There are several special characters in Job Title and **Job Salary** Column like



Since the special characters are quite consistent, they were cleaned by simply using the Find and Replace functionality.

5. Now I started with reviewing each column and cleaning them one by one. The reason I deleted some records first is that I do not want to put effort into records that are going to get deleted anyway.

5. Since I am focusing on Data Job roles here, and this list is to provide that information, I wanted to extract the basic Data Job Categories from the Job Roles. I have categorized the roles as below



For that, I added a new Column as “Job Type” and wrote a formula to categorize the Job Titles into those Job Types.

IF(ISNUMBER((SEARCH("Data",A2))\*AND(SEARCH("Analyst",A2))),"Data Analyst",

IF(ISNUMBER((SEARCH("Business Intelligence Analyst",A2))),"BI Analyst",

IF(ISNUMBER((SEARCH("Business",A2))\*AND((SEARCH("Analyst",A2)))),"Business Analyst",

IF(ISNUMBER((SEARCH("Data",A2))\*AND((SEARCH("Engineer",A2)))),"Data Engineer",

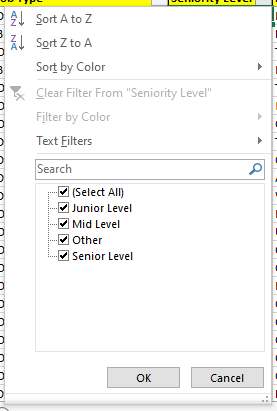
IF(ISNUMBER((SEARCH("Data",A2))\*AND((SEARCH("Scientist",A2)))),"Data Scientist",

IF(ISNUMBER((SEARCH("BI Engineer",A2))),"BI Engineer",IF(ISNUMBER(SEARCH("Analyst",A2)),"Other Analyst",

IF(ISNUMBER((SEARCH("Business",A2))\*AND((SEARCH("Intelligence",A2)))),"Business Intelligence Specialist",

"Other"))))))))

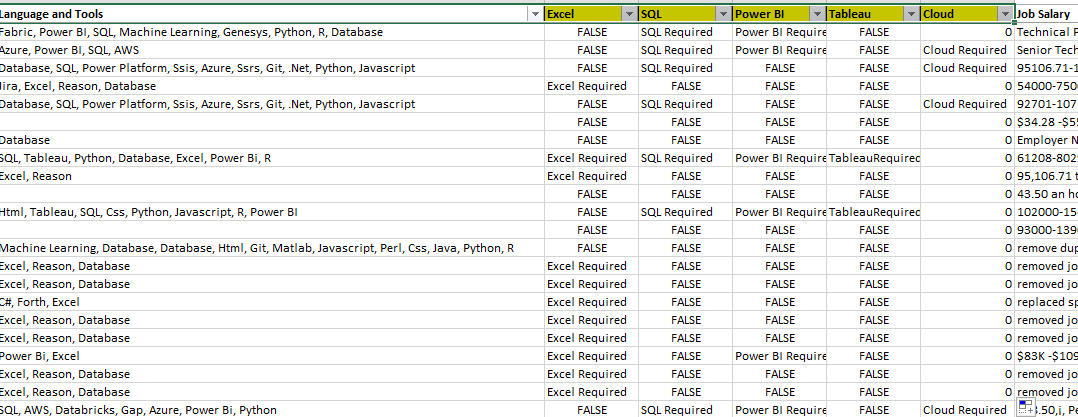
6. I then decided to add a new column as “Seniority Level”, where vase on the Job Title I determined the Seniority Level of the employees. I wrote the below formula to achieve this:



=IF(ISNUMBER(SEARCH("Junior",B2)),"Junior Level",IF(ISNUMBER((SEARCH("Senior",B2))),"Senior Level",IF(ISNUMBER((SEARCH("Manager",B2))),"Mid Level",IF(ISNUMBER((SEARCH("Lead",B2))),"Lead

","Other"))))

7. The column language and tools have information on the tools or skills required for the job and have multiple skills separated by comma. I decided to determine the most sought-after skills in the industry! So I chose the common skills for the data jobs, that are Excel, SQl, Power BI, Tableau and Cloud (AWS, Azure or Google Cloud).

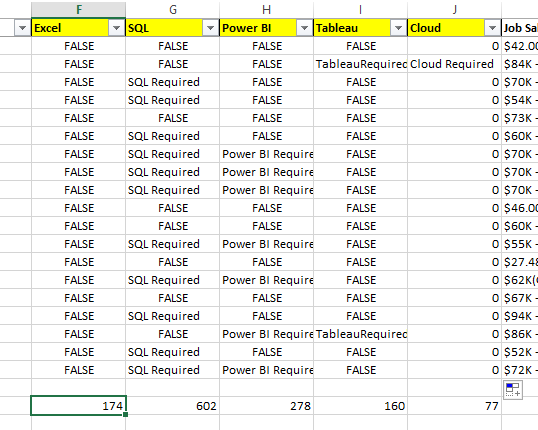


For this I wrote formulas like:

IF(ISNUMBER(SEARCH("Excel",$E2)),"Excel Required") like this for all the skills.

After that I counted the number of times Those specific Skills appeared in the data set. I did that by writing

COUNTIF(F2:F1150,"Excel Required")



Thus, further conclusions can be made regarding the in-demand skills.

8. Now coming to the most time consuming and complex cleaning process, which is for the column, **Job Salary.**

The data in this column was very inconsistent so it was impossible to clean it using just one or two formulas. For example, the data looked like

|  |
| --- |
| 92701-107184 |
| $34.28 -$55.07‚¡‚ Per hour(Employer Est.) |

|  |
| --- |
| 95,106.71 to $127,429.57 Hourly Pay, $52.26 to $70.02 Benefits : |

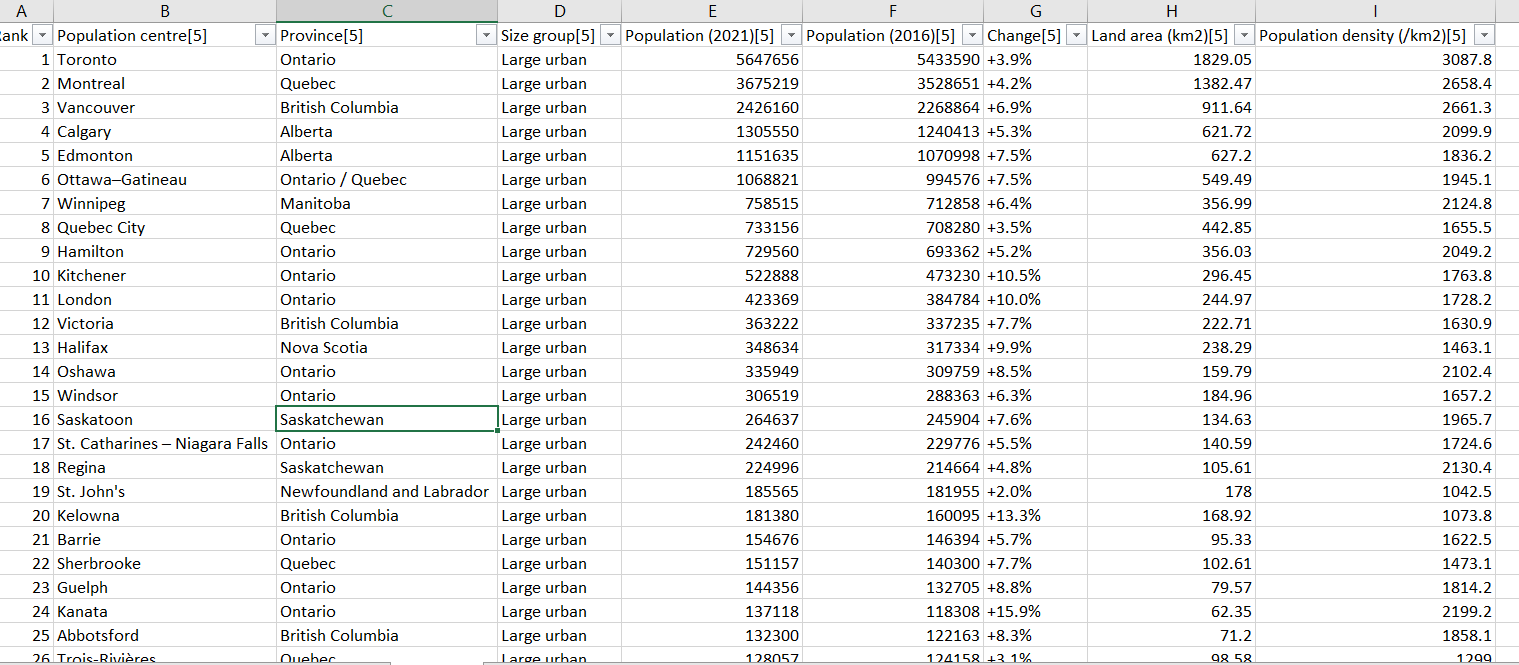
|  |
| --- |
| $80.00 -$100.00‚¡‚ Per hour(Employer Est.) |

|  |
| --- |
| 11 -$75,223 ., 75,223 . with, 96,196 . Pay  Etc. and so on.  I have created some custom columns to analyze and extract the salary data. Let me show you,  **Column 1.First ‘$**’: in this column I have extracted the numerical value followed by “ $ “ sign. For this I wrote  MID(K2, FIND("$",K2)+1, FIND(" ",K2, FIND( "$",K2)+1) - FIND("$",K2)-1)  **Column 2.First Comma:** in this column I have extracted the numerical value preceded by “,”. For this I wrote  **=LEFT(K2,FIND("‚",K2)-1)**  **Column 3.Second $:** since some records has range of values, I searched for the second occurrence of $ and extracted the value followed by that. For that I applied:  =MID(K101, FIND("$", K101)+1, FIND(" ", K101, FIND("$", K101)+1) - FIND("$", K101)-1)  Those three columns produce results like    **Column 4,5.max,min:**  Clearly the minimum and maximum salary in the range were caught but not uniformly throughout the columns.  So, in order to get the minimum and the maximum values among them I inserted 2 columns “**max**” and “**min**”.  The formulas I wrote are:  =MIN(IFERROR(INT(M2),0),IFERROR(INT(N2),0),IFERROR(INT(O2),0))  =MAX(IFERROR(INT(M2),0),IFERROR(INT(N2),0),IFERROR(INT(O2),0))  **Column 6: “hourly or yearly”** Since some of the jobs have hourly rates and some even monthly, so I checked what is the basis of pay: hourly,monthly or yearly  IF(ISNUMBER(SEARCH("hour",K2)),"Hourly",IF(ISNUMBER(SEARCH("month",K2)),"Monthly","Yearly"))  **Column 7: currency:** This column checks if the currency is USD or CAD  =IF(ISNUMBER(SEARCH("CAD",K2)),"CA$","$")  **Column 8:**  Now I created a column “**manual**” to address what was working and what was not.This one is as the name suggest, has manual values. There were quite a few rows in the “**Column 1.First ‘$**’,**Column 2.First Space:** and **Column 3.Second $** that did not produce any values (#VALUE!) . I filtered those rows by selecting the error it produced. Then I tried to see if there are any patterns within those data. And there actually was!  I started by filtering blanks or errors in the “First $”, “First Comma”, “Second $” .  Here I could see because there were spaces after $, some values were not getting populated. So I replaced “$ “ with “$”.  Then I filtered “value errors for “first $”. I could see a clear pattern for those Salary values.    So in the manual column I applied a formula  IF(ISNUMBER(SEARCH("hour",L11)),LEFT(L11,SEARCH("an hour",L11)-1),IF(ISNUMBER(SEARCH("year",L11)),LEFT(L11,SEARCH("a year",L11)-1)))  Thus, for the #VALUE errors in first $ column I can populate the value.  Also, at this stage I filtered some odd values for “First $”, and encountered non salary data in Job salary column. I deleted those records. Example      Then I encountered some Blank errors appearing in “First $”.    For these records, I either populated them by hard coding the values in the “manual column”, or I wrote some formulas:    I encountered another pattern and wrote another formula in the manual column to extract the data.    I found a pattern in the **Job Salary** column and wrote a formula to extract it. This one produced 518 no of rows!    Then I encountered a lot of errors in the values, like “$, some number” and hence the data was not extracted. I replaced “$, “ with “$ “. And so on .  Now since the manual column is now being populated with values, which are mostly from erroneous values from the custom columns I created. So, I am proceeding with “**first comma**”, “second $”, “**max**” and “**min**” column. I am filtering out the blank/#VALUE errors of those columns and trying to figure out if there are certain patterns in the **Job Salary** Column and populating the values in the manual column. If there is not pattern then I am manually inserting values. In this process I am also finding non salary data in the Salary Column and I am eliminating those rows too. So as this process is progressing I am basically populating the desired salary figures in the manual column.  After all the steps above now it is time to populate the non blank values of the “**first comma**”, “**second $**”, “**max**” and “**min**” columns, for which the manual column is still blank. Since the “**max**” and “**min**” value is derived from the “first comma”, “second comma” columns, I only need to consider the “max” and “min” values. So I wrote:  =CONCAT(P2," - ",Q2)  And with that all the salary values are now populated in the manual column, using the calculated columns I prepared. Now it is time to assemble all the values for a single column summarizing the salary figures.  Just before that I noticed there are few values in the manual column that have “$” and some do not. So, I replaced $ signs with blank to write a uniform formula to get a clean looking salary figure.  =CONCAT(R2, " ",U2,S2) |
|  |

**9. City Size:** I added this column for further data analysis. For this I added a new sheet and imported city and size data from Wikipedia, by import data from web functionality in Excel.

I used the data from this Wikipedia page:

<https://en.wikipedia.org/wiki/List_of_the_largest_population_centres_in_Canada>



Then I applied this formula to extract the “size group” of the cities

=IFERROR(IF(LEFT(W2,6)="Remote","Remote",IF(W2="Ottawa","Large Urban",VLOOKUP(W2,Table\_1[@[Population centre'[5']]:[Size group'[5']]],3))),"Other City")

The city Ottawa needed to have its own formula because the value in the wikipedia page had “Ottawa–Gatineau” and hence keeping no check for Ottawa was fetching wrong value.

**Conclusion:**

This project demonstrates advanced Excel-based data cleaning techniques, including:

- Removing duplicates and handling missing data

- Text filtering and categorization

- Extracting structured salary data from varied formats.

- Identifying in-demand skills

- Integrating external datasets for enhanced analysis

The cleaned dataset is now ready for further insights and visualizations.