Final Report

Employment in non-profit sector

By: Sangjin (Eric) Lee

Submitted by November 27th, 2023

Page of content

Abstract Introduction Page 3

Research Question Page 6

Link for Github Page and details Page 13

Descriptive Detail Page 16
- Explanation on splitting for indicators Page 22
- Explanation on splitting for characteristics Page 31
- Explanation on splitting for Provinces Page 39

Result of Analysis Page 42
- Annual Hour worked based Page 45
- Average Annual Wages based Page 49
- Average hourly wages based Page 52
- Average weekly hour worked based Page 54
- Hours worked based Page 57
- Wages and Salaries based Page 60

Conclusion Page 63

Source Cited Page 67

Abstract Introduction

The database that I have choose are for the employment in non-profit sector by demographic characteristics in Canada. Those are easily accessible from “Open Government Portal” which managed the Canadian government datasets. The database explains how many workers are working in the non-profit sectors. The database explains by number of the jobs, hours worked, and average salaries by not only whole country but by province and territory. The database also classified into sex, age, level of education, immigration status, indigenous identity, and visible minority status. It also divided how many people work at non-profit for government, community, and/or businesses.

The theme for this dataset that I will choose is data mining. Data mining is the process of sorting through large data sets to identify patterns and relationships that can improve. (Kumar) Since they are the largest dataset, I will extract portions of the data to analysis. Inside the dataset, they are classified in Geographical, Sectors, Characteristics, and Indicators. By using data mining, the result shows more accurately and easier for me to analysis.

The analysis of this problem will be how did working for non-profit organizations have improved throughout the years and by provinces. For my analysis, I will use to split years of “2010-2012”, “2013-2015”, “2016-2018”, and “2019-2021”. This dataset has data from 2010 to 2021. For this purpose, I will use “2010-2012” as references. I will use “2013-2015”, “2016-2018” as training set and “2019-2021” as testing set. If I use entire datasets, I will have way much more to analysis which leads to inaccurate results. There are seven Indicators that I need to analysis and split. Furthermore, there’s four more characteristics that I need to analysis. I also need more time and effort to analysis every individual dataset. Also, it will result the longer technical report (sometimes extra pages). As a result, I may skip or may do more duplication. Plus, I do not have unlimited amounts of time to analysis.

The data that I will be using from the dataset is the following. First, as mentioned, I will use data from “2013-2015”, “2016-2018” as training set, and “2019-2021” as test set. Second, I will use all Indicators and will be split by each indicator. There’s are seven indicators in total. The dataset or the file itself has all the indicators inside. Then, I will split into “Gender group”, “Age group”, “Education group”, and “Immigration status” from ‘Characteristics’ indicators. Last, out of thirteen provinces inside the datasets, I will extract only five provinces to analysis in the given dataset. Unlike the last two division of datasets I did, I will use one hot encoding to convert the five provinces. The dataset is found here is in csv files from “The Canada Government Datasets”. Last for characteristics items that I feel like it can be convert into numerical, least to greater, I will likely convert those indicators into numerical. Those are ‘Age group’ and ‘Education group’. Furthermore, for the script, I have included both training and testing set. However, I will only use the testing set for the result.

The techniques that I will use are classification and clustering. Classification is reorganized and grouping ideas into categories. (Monkeylearn) I will use classification to reorganize the data into suitable data that I can use to recognized. The next technique that I will use is clustering. Clustering is a technique that explores occurring groups within a data set. (Tibco) I will use clustering techniques to select the portions of the data to analysis the dataset.

In terms of the tools, I will use Python to analysis the data. Although Python may require additions tool to analysis each data, I will be enough to conduct the result from this analysis. For this experience, I have used “Pandas” to use loc features, “Numpy” to use aggregation, and finally use “Panda-profiling” also known as “ydata-profiling” to generate for analysis data for my next direction. In addition, I have use package called ‘fitter’ to describe best fit for normally distributed item.

Overall conclusion, the dataset that I choose are employment in non-profit sector by demographic characteristics in Canada founded publicly in the Canada government datasets. The dataset is organized by number of jobs, hours worked and average salary by province and all provinces. The theme for this dataset will be data mining to select portions of the data. The problem I will try to investigate is how did working for the non-profit has improved overall. I will use data from “Gender”, “Age”, “Education level” and “Immigration status” each province by training set (2013-2015, 2016-2018), and testing set (2019-2021). The technique that I will use in this analysis will be done in classification and clustering. Finally, the tool that I will be using in Python using “Numpy”, “Pandas”, “Panda-profiling”

My Research Questions

What are significant predictors for classifying non-profit employment datasets?

Predictive analytics is the use of data to predict future trends and events. It uses historical data to forecast potential scenarios. One of the predictive analytics tools is regression analysis. (Harvard Business School Online).

The significant predictors in this classification on non-profit employment dataset is to see which categories of employees work more, paid more, and the number of the job available on the market currently. My main purpose of analysis this dataset is to see if there's any demand for working for non-profit organization.

In addition, my focus aimed to see if there's any improvement throughout the year in the non-profit organizations. The dataset that I was given is based on the year. However, almost all of them are not numerical but categorical data. Fortunately, there's some categorical data, primarily in characteristics that can be converted into range. I will use the year and the indicator as linear regression to see if there's potential inside non-profit organizations. This means that I will go through to see who's working less and/or who's paying more. In addition, I will see if there's any new hires available throughout the organization, especially in the five provinces.

Another focus for this analysis is to see if the working conditions in non-profit organizations have improved and who should be working for these organizations at least in Ontario and compare to all five provinces.

Which categories of the employees work more, paid more, number of the job available on the market?

For the next step, as I mentioned above, I am going to examine whose employees work more, paid more, number of the jobs available on the market. Right now, in modern society generally, we are in recession season. After the covid, the economy doesn't seem like getting any better. This also affects our job markets as well. Since we are having recession, the job markets are also not doing so great. This means non-profit organizations are also having more difficulty than the profit organizations. So, this is the great way to see if the non-profit organizations are indeed having difficulty and see whether they hire less, hire more, and even number of hours that they work.

According to Cause Leadership, in 2023, most non-profit organizations, mostly NGOs are tight on budgets. So, most of these organizations are working with below-average salaries and some work as volunteers as well. The site also mentions that most nonprofits also adapted remote hybrid work, good well-being, diversity, and equality, known in-dept financial knowledge, top talents, and salary transparency. Therefore, stuff mentioned in this site will likely influence the result of this analysis. To analysis further, Statistic Canada release the data say, there's rising number of immigrants in the non-profits organizations as well. Also, most of the graduates are from college and university. Also, there's a lot of workers aged 55 or older.

How I am going to do this analysis is categorical by each section. First by age group, then by gender, then by education and finally by immigrant status. Then apply this data to the five of the provinces. After that I will see which group pays more, works more, and number of the job demands on the market.

How does Decision Tree Classifier perform better for the classification?

Decision tree is a flowchart-like tree structure where an internal code represents a feature. (Navlani) Decision Tree is one of the algorithms to build a predictive model. Decision Tree is hierarchical in the structures that include root nodes, branches, internal nodes, and leaf nodes. (analytixilabs) Decision tree is a good method to do as it allows to classify into specific set of the data which make this analysis much easier and analysis more accurately.

Decision Tree can solve both classifications and regression models. It can predict categorical variables and predict continuous variables. However, I don't think that Decision Tree is about long-term and is 'short-sighted' solutions.

In my analysis, decision trees also play some roles. I am going through this by splitting into a lot of datasets. At first, I opened the CSV dataset file, there's a lot of data and I was probably unable to compare the values at all. So, to divide into different datasets, I have examined each column individually to see if there's value worth dividing. I chose 'Indicators' columns to divide into seven datasets. Then I use 'GEO' also 'years' columns to divide into different datasets, in my case, three dataset, 2010-2012, training (2013-2018), testing (2019-2021). That's around 21 datasets. Then I applied 'Characteristics' indicators to divide into four categorical types. That's now around 84 datasets. If I were to split into 'provinces' as well, then it will reach 1092 datasets (13 provinces in this dataset).

As there are more datasets or trees to divide, there's a lot of work being done. Those are mostly dealing with a lot of complex coding or having to repeat so many times. Furthermore, this will create a lot of debugging and perhaps remove some of the important data. This is one of the challenges that I need to go through and maybe perhaps this is the downside of using decision tree classifier.

Is data mining/selection allowing to perform more accurate result than without it?

Data mining/selection allows us to filter the data and to show clearer and better results. However, data mining also or cannot show accurate results. Data mining is the process of sorting through large data sets to identify patterns and relationships. Data mining involves data gathering, preparation, mining the data, and finally data analysis and interpretation.

Data mining is required to do this analysis. When I investigate the datasets, it is a large set of datasets. For me to analysis the result, I must sort out the data and I had to separate it into more than one dataset. This eventually ends up with more than 20 datasets perhaps could lead to more datasets by the end. However, there's pro and cons related to data mining.

First with data mining, I can filter the result as what I was intended to. Using data mining, I can show and filter the final dataset as the result I want to show. In my experiences, by splitting the 'Indicators' columns, I can filter the result that I want to analysis such as 'Number of the Job' on the non-profit organizations. Data mining is mainly to show their opinions and persuade people to follow their way. However, Data Mining does help me to reorganize the data to analysis much easier, there's also con with this too.

Second, it can be both pro and con, but data mining helps me to show the bias results. How this can be done is by people who see the database can purposely analysis differently than what it is intended for. They are mainly there to show bias result is that the dataset provided already contains the bias. If I was to use the dataset to analysis the result, the result may be dramatically different whether that is used by excel, R, even Python.

Thirdly, if I don't fully consider the dataset, there's a possibility that we might see different results. If I have misunderstood the datasets, I will probably filter the wrong set of datasets. Just filtering one wrong set of datasets shows completely different results. When I was analysis the dataset inside this project, I had to remove and add columns because when I first looked at it, there were a lot of columns that I had difficulty understanding. To prevent that I need to make sure I understand the datasets, especially the purpose of these datasets and its associated columns.

Third, there may be numerous amounts of repeated processing. When I was analyzing the data, I had to do multiple repeating of the code because there were a lot of datasets that I needed to divide. Some of these data inside the dataset are very confused and unable to be analyzed. In addition, as I split the dataset and repeat the same code over again, I have found some of the concerns I need to address. First, since the page gets too long, I use methods and classes to condense the number of the repeated code. Second, since the code gets too long, I need to deal with a lot of code errors, debugging, and required the time to understand my code. For this analysis, when I got into splitting into ‘Characteristics’ columns, I am finding myself spend a lot of time debugging or time to understand my code. For my case, I had to use a lot of variables too.

How accurate is this information is?

The information that is given inside the dataset is nearly accurate. However, the structure of the dataset is very confusing and very complex to analysis. Also, from the beginning of the research, I have noticed that they have filtered some of the information. From the beginning, they have gathered a certain number of samples to analysis. For example, they have gathered 4000 – 5000 samples from each province or indicators. Although this dataset indeed filtered some of the missing samples, I found out that removing even some of the data will give inaccurate results.

In addition, how this is structured gives me more inaccurate information as soon as I look at it, the VALUES and INDICATORS are located all in one column. This not only makes me confused on analyzing but the values inside cannot be compared. So, the division of the columns is required and is really time consuming. Perhaps, the result may not be even more accurate than I anticipated it. Even removing all null values. There's may be possible that valid values might have been removed due to this. By the time I reached split the 'REF_DATE', years, I am started to notice a lot more work being code and work involved. Then by the reached the 'Characteristics' and 'Provinces', I am started to easily exhaust to do code and work involved.

How did information have gathered from?

As I was analyzing the information from this dataset, I started to have questions about it. As I looked through the datasets, I found so many irrelevant columns that I have no idea what this means or is this even related to my analysis.

Based on my analysis on the dataset, I feel like other columns are meant for referencing where the information comes from. This includes reference code, version, and geographic location. This may be helpful for those who are making this dataset. In my analysis, the dataset come from the Government of Canada.

In addition to this, I found that all the samples that they gathered from are already filtered. This means out of all the data they collected they have filtered a certain number of samples. For example, I have a sample dataset where all the samples have equally divided data observations. This means that there's at least 3000 observations for each province. Some have null value as I mention. This could make my research much easier, however it may give me more inaccurate information. Therefore, how they collected data and how they shared the information, and the resources will help me to analysis much more details.

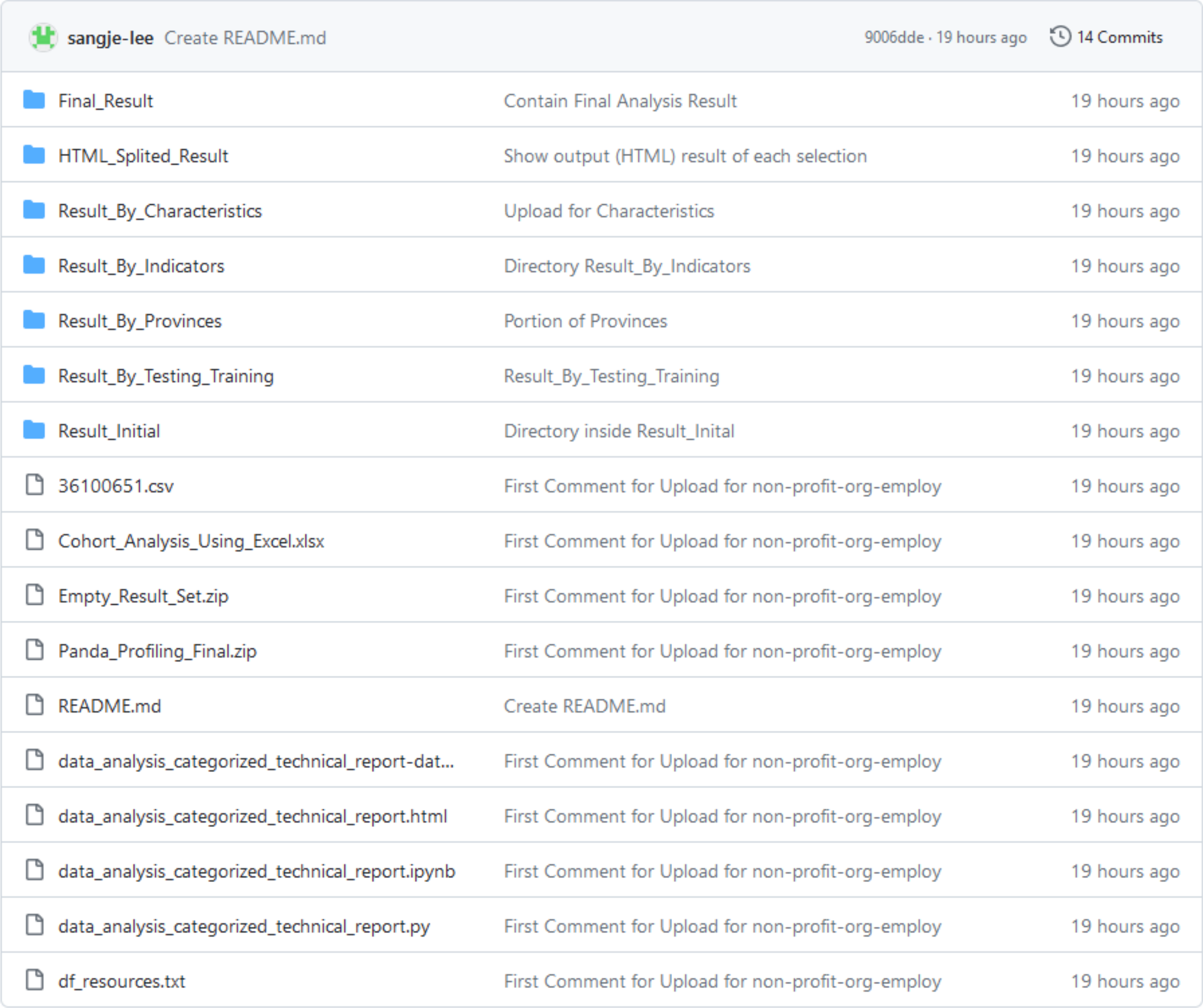
A link to a repository on GitHub.

https://github.com/sangje-lee/non-profit-org-employment

Here is the Github website where codes and results are uploaded. I have uploaded not only my progress but as well as results of each code. Since the analysis took a lot of time and effort, I could not be able to paste all the results in this report. The result itself is more than 50 pages. This includes Jupiter Notebook file, Panda Profiling, and outputting all the results in the report.

In there, I have uploaded the result that includes python code, my analysis, and the result that I haven’t put in the report. In addition, I placed the project in the main directory. It will be the same file as I am submitted with this report.

Here is the detail of my Github where I put all the results below,



sangje-lee Create README.md 9006dde · 19 hours ago 14 Commits

Final_Result	Contain Final Analysis Result	19 hours ago
HTML_Splited_Result	Show output (HTML) result of each selection	19 hours ago
Result_By_Characteristics	Upload for Characteristics	19 hours ago
Result_By_Indicators	Directory Result_By_Indicators	19 hours ago
Result_By_Provinces	Portion of Provinces	19 hours ago
Result_By_Testing_Training	Result_By_Testing_Training	19 hours ago
Result_Initial	Directory inside Result_Inital	19 hours ago
36100651.csv	First Comment for Upload for non-profit-org-employ	19 hours ago
Cohort_Analysis_Using_Excel.xlsx	First Comment for Upload for non-profit-org-employ	19 hours ago
Empty_Result_Set.zip	First Comment for Upload for non-profit-org-employ	19 hours ago
Panda_Profiling_Final.zip	First Comment for Upload for non-profit-org-employ	19 hours ago
README.md	Create README.md	19 hours ago
data_analysis_categorized_technical_report-dat...	First Comment for Upload for non-profit-org-employ	19 hours ago
data_analysis_categorized_technical_report.html	First Comment for Upload for non-profit-org-employ	19 hours ago
data_analysis_categorized_technical_report.ipynb	First Comment for Upload for non-profit-org-employ	19 hours ago
data_analysis_categorized_technical_report.py	First Comment for Upload for non-profit-org-employ	19 hours ago
df_resources.txt	First Comment for Upload for non-profit-org-employ	19 hours ago

First of all, “data_analysis_categorized_technical_report.ipynb” is where I did all of my code. It is saved in Jupiter notebook format. I used Anaconda virtual environment using python 3.8.18 installing Numpy, Pandas, Fitter, and Panda-profiling.

Next “data_analysis_categorized_technical_report.html” is basically in html/PDF file from the Jupiter Notebook file. “data_analysis_categorized_technical_report.py” is the python file of Jupiter Notebook file but saved as python script. There’s “data_analysis_…_report-data-…-only.py” file, where it’s storing only the data processing of the report.

Next, “Panda Profiling Result” directory stored all the html that was generated from the Jupiter Notebook files. The html files are the main source of analysis of this report. It displays much more details than the two PDF files I mentioned.

There are two files, “df_resources.txt” and the CSV files. The txt file is just showing all the variables that I used for this analysis and from Jupiter Notebook. The CSV file is the original dataset from the website.

Here are the remaining file descriptions,

“36100651.csv”	Original dataset contain employment in non-profit organizations.
“36100651-eng.zip”	Zip file for original dataset and some description
Empty_Result_Set.zip	Contain all requirement folder with Jupiter notebook script inside. No CSV files, when extract to main directory and run the script, it will automatically add CSV files to the dedicated directory.
README.md	Read me for the Github
HTML_Splited_Result	Output that are already executed and saved in html format. The final html is also saved in there as well.
Result_By_Characteristics	Part contains about split by “Characteristics” and its portion script.
Result_By_Indicators	Part contains about split by “Indicators” and its portion script.
Result_By_Provinces	Part contains about added five provinces + modified characteristics
Result_Inital	Part contains before the split by ‘Indicators’ and removal of nonessential files. It’s the beginning of the code.
Final_Result	Script contains result of this Analysis. Important for final analysis.

For references, the Jupiter notebook file inside the directory needs main script to run before running the script inside the directory. Also, **my portion of final analysis script** is inside **Final_Result** directory. **Also, script inside the directory required to executed and have all csv file inside the directory.**

Descriptive details and steps to do analysis.

The selected dataset, “Employment in the non-profit organization” can be found at the Statistic Canada Website. This dataset alone is divided by non-profit sector by demographic characteristics. The dataset is saved in CSV file. The four columns/attributes that are important in this dataset are “Geography”, “Sector”, “Characteristics”, and “Indictor”. There are other columns that are presented here. However, they are not required to do the analysis. Out of these four most important columns, “Geography”, “Characteristics” and “Indicators” are the important columns that I need to do analysis. Here are the summarized details for the dataset,

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105840 entries, 0 to 105839
Data columns (total 17 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   REF_DATE         105840 non-null  int64  
 1   GEO              105840 non-null  object 
 2   DGUID            105840 non-null  object 
 3   Sector           105840 non-null  object 
 4   Characteristics  105840 non-null  object 
 5   Indicators       105840 non-null  object 
 6   UOM              105840 non-null  object 
 7   UOM_ID           105840 non-null  int64  
 8   SCALAR_FACTOR    105840 non-null  object 
 9   SCALAR_ID        105840 non-null  int64  
 10  VECTOR           105840 non-null  object 
 11  COORDINATE       105840 non-null  object 
 12  VALUE            102816 non-null  float64
 13  STATUS           3024 non-null    object 
 14  SYMBOL           0 non-null       float64
 15  TERMINATED       0 non-null       float64
 16  DECIMALS         105840 non-null  int64  
```

With sample output following:

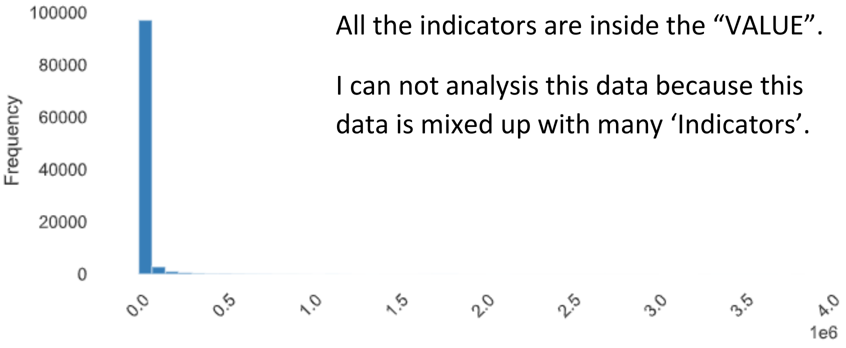
```
   REF_DATE     GEO           DGUID                         Sector  \
0      2010  Canada  2016A000011124  Total non-profit institutions
1      2010  Canada  2016A000011124  Total non-profit institutions

    Characteristics                         Indicators      UOM  UOM_ID  \
0    Male employees                     Number of jobs     Jobs     190
1    Male employees                       Hours worked    Hours     152

  SCALAR_FACTOR  SCALAR_ID       VECTOR COORDINATE       VALUE STATUS  SYMBOL  \
0         units          0  v1273033811    1.1.1.1   642584.00    NaN     NaN
1     thousands          3  v1273033812    1.1.1.2  1048516.00    NaN     NaN

   TERMINATED  DECIMALS
0         NaN         0
1         NaN         0
```

Detail outputs are in other documents with **(Output #01)** and Github

All the contents in the column are equally divided as mentioned above. With sample size of following:

	Size
'REF_DATE'	8820
'GEO'	7560
'Sector'	21168
'Characteristics'	5880
'Indicators'	15120

Histogram for "VALUE"



All the indicators are inside the "VALUE".

I can not analysis this data because this data is mixed up with many 'Indicators'.

Histogram with fixed size bins (bins=50)

Detail outputs are in other document with **(Output #02)** and Github

Based on the dataset given, there’s no way to compare and analysis this dataset especially if all the ‘Indicators’ are given in one big dataset. I have reviewed the dataset and carefully looked at “Indicators”.

However, even though I divided into “Indicators” columns, I need to remove some of the columns/attributes because some of the columns are not helpful to deal with my analysis. Some of the columns just contain ID, Vector, Coordinate that are repetitive to the problem. By removing some of the columns/attributes, I can simplify the result and try to analysis more accurately. After the modification, here is the result after the removing some of the columns.

```
print(df_sorted.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105840 entries, 0 to 105839
Data columns (total 8 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   REF_DATE         105840 non-null  int64  
 1   DGUID            105840 non-null  object 
 2   GEO              105840 non-null  object 
 3   Sector           105840 non-null  object 
 4   Characteristics  105840 non-null  object 
 5   Indicators       105840 non-null  object 
 6   UOM              105840 non-null  object 
 7   SCALAR_FACTOR    105840 non-null  object 
 8   VALUE            102816 non-null  float64
dtypes: float64(1), int64(1), object(6)
memory usage: 7.3+ MB
None

    REF_DATE     GEO                           Sector      Characteristics  \
0       2010  Canada  Total non-profit institutions       Male employees   
1       2010  Canada  Total non-profit institutions       Male employees   

                      Indicators      UOM SCALAR_FACTOR       VALUE  
0                 Number of jobs     Jobs         units   642584.00  
1                   Hours worked    Hours     thousands  1048516.00  
```

Detail outputs are in other document with (**Output #03**) and Github

These are the main reasons why I have decided to remove the following columns/attributes.

UOM_ID

UOM_ID inside this dataset is simply translate actual words into numeric codes. Since this look like having repetitive code, I have decided to remove it. Alternatively, I can use this instead of UOM.

	UOM	UOM_ID
0	Jobs	190
1	Hours	152
2	Dollars	81

Detail outputs are in other document
with (**Output #04**) and Github

Noticed that UOM_ID for Jobs is 190, Hours = 152, and Dollars = 81.

SCALAR_ID

For similar reason why, I removed UOM_ID, I think there's a lot of repetitive inside the SCALAR_ID to SCALAR_FACTOR. If I were to remove SCALAR_FACTOR, then SCALAR_ID can be used.

	SCALAR_ID	SCALAR_FACTOR
0	0	units
1	3	thousands
2	6	millions

Detail outputs are in other document
with **(Output #04)** and Github

Noticed that, Units = 0, thousands = 3, and million = 6

‘VECTOR’

I removed VECTOR column because I don’t believe that it refers to anything that I will analysis. VECTOR columns use 11 words that usually start with “v1273”. I believed that the Vector is code that is used to refer the data.

Value	Count	Frequency (%)
v1273033811	12	< 0.1%
v1273033912	12	< 0.1%
…	…	…
v1273033911	12	< 0.1%
v1273034698	12	< 0.1%
Other values (8810)	105720	99.9%

Detail outputs are in other document with (**Output #04**) and Github

COORDINATE

I have decided to remove ‘COORDINATE’ columns. I believed it doesn’t really need to analysis my result. From my perspective, I think Coordinate is being used as where this data is collected according to the GPS.

Value	Count	Frequency (%)
1.1.1.1	12	< 0.1%
1.1.2.4	12	< 0.1%
…	…	…
1.1.2.3	12	< 0.1%
1.1.10.6	12	< 0.1%
Other values (8810)	105720	99.9%

Detail outputs are in other document with (**Output #04**) and Github

STATUS

I removed “STATUS” columns because the column is not being used with my analysis. In addition, it doesn’t really contain anything meaningful to analysis anything. Most of these column rows contain null (or missing value) or marked “x”.

SYMBOL

Similar reason to STATUS column. It doesn’t contain anything meaningful and do not need to analysis this column. All column rows contain null (or missing value).

TERMINATED

Similar reason to TERMINATED column. It doesn’t contain anything meaningful and do not need to analysis this column. All column rows contain missing value.

DECIMAL

I removed ‘DECIMAL’ columns because I don’t need to use this column to analysis my result. It’s there for me to check whether this row contains any decimal values or not inside “VALUES” columns.

Value	Count	Frequency (%)
0	90720	85.7%
2	15120	14.3%

Now, I want to see if filtering characteristics or indicators are the right columns to filter. Here are two columns that I will need to modify next,

Characteristics columns (total sample of 5880 per each equally divided):

15 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years old and over
College diploma
Female employees
High school diploma and less
Immigrant employees
Indigenous identity employees
Male employees
Non-immigrant employees
Non-indigenous identity employees
Not a visible minority
Trade certificate
University degree and higher
Visible minority

Indicator columns (total sample of 15120 equally divided):

Average annual hours worked
Average annual wages and salaries
Average hourly wage
Average weekly hours worked
Hours worked
Number of jobs
Wages and salaries

Detail outputs are in other documents with (**Output #05**) and Github

Before I proceed with filtering, I need to investigate whether the data has some missing values. After analysis, I have noticed there's some missing values inside VALUE columns. according to the panda-profiling. According to this, out of 105840 observations, 3024 samples are found with missing values (2.9%) inside VALUE.

```
print(df_sorted.isnull().sum())
…
VALUE                  2.857143       3024        105840
```

Before I can modify using “Indicators” or “Characteristics”, I want to make sure, I would remove missing values so that the results are much more accurate. After filtering some of the missing value, the number of size inside ‘characteristics’ attribute/column has decreased. However, interesting enough ‘indicators’ attribute/column has decreased equally.

For Characteristics, I found out that age between ’15-24 years old’, ’65 years over’ has decreased from 5880 to 5376. In addition, ‘Immigrant’, ‘Non-immigrant’, ‘Indigenous identity’, ‘Non-indigenous’, ‘Trade certificate’, and ‘University degree’ has decreased from 5880 to 5544. Similar categories associated with ‘College diploma’, they haven’t decreased because of the missing value. This will make the analysis very interesting. For indicator attributes as mentioned decreased equally from 15120 to 14688. I have attached the output into another document (**Output #06**) and Github for this reference.

After this process, almost all the data that doesn’t have ‘VALUE’ has been removed. Although there is an uneven record inside Characteristics, interesting indicators have still been divided equally. After this process is completed, I have decided to use ‘indicators’ as the first one to be split. There’s no way of analyzing ‘VALUE’ attributes if there’s multiple indicators being there. Also, there’s way too extreme values to compare. I really can’t compare the result between the hours worked and wages. I am going to create new datasets for each indicator. After the splitting, this is the new result I will get.

For Average annual hours worked,

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**462**
median	**1593**
Maximum	**2500**
Range	2038

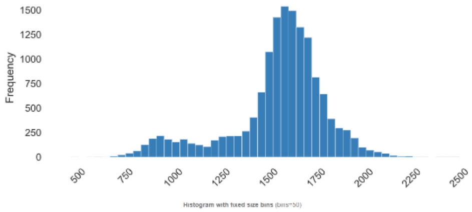
Descriptive statistics (VALUE)

Mean	**1551.4361**
Skewness	**-1.0026593**
Sum	22787494

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

Detail outputs are in other documents with (**Output #07a**) and Github



For Average annual wages and salaries,

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**8769**
median	**42186.5**
Maximum	**133071**
Range	124302

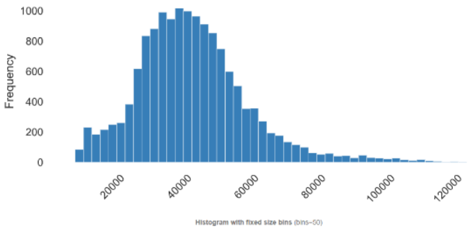
Descriptive statistics (VALUE)

Mean	**43804.783**
Skewness	**0.91651225**
Sum	6.4340465 × 10^8

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

Detail outputs are in other documents with (**Output #07b**) and Github



For Average hourly wages,

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**10.16**
median	**26.7**
Maximum	**75.37**
Range	65.21

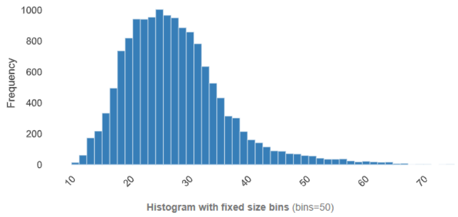
Descriptive statistics (VALUE)

Mean	**27.825611**
Skewness	**1.1323432**
Sum	408702.58

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

Detail outputs are in other documents with (**Output #07c**) and Github



Histogram with fixed size bins (bins=50)

Average weekly hours worked,

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**9**
median	**31**
Maximum	**48**
Range	39

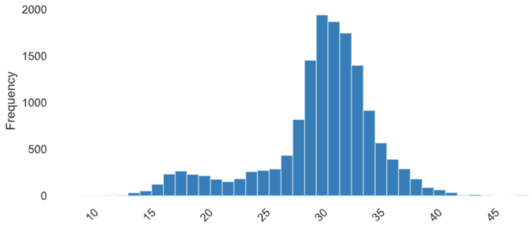
Descriptive statistics (VALUE)

Mean	**29.831767**
Skewness	**-0.9985335**
Sum	438169

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

Detail outputs are in other documents with (**Output #07d**) and Github



Histogram with fixed size bins (bins=38)

Hours worked

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**6**
median	**9586.5**
Maximum	**3857813**
Range	3857807

Descriptive statistics (VALUE)

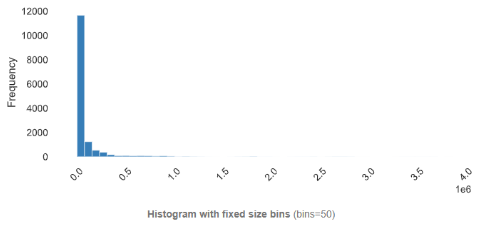
Mean	**83596.946**
Skewness	**7.1113237**
Sum	1.2278719×10^{9}

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

The data is skew toward left. I may not use this data afterward.

Detail outputs are in other documents with (**Output #07e**) and Github



Number of Jobs

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**11**
median	**6305.5**
Maximum	**2428289**
Range	2428278

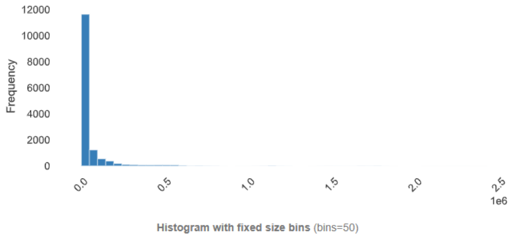
Descriptive statistics (VALUE)

Mean	**53441.062**
Skewness	**7.1394116**
Sum	7.8494232×10^{8}

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

Detail outputs are in other documents with (**Output #07f**) and Github



Wages and Salaries

Dataset statistics

Number of variables	**8**
Number of observations	**14688**

Variable types: 2 numeric and 6 categorical

Quantile statistics (VALUE)

Minimum	**0**
median	**224**
Maximum	**132601**
Range	132601

Descriptive statistics (VALUE)

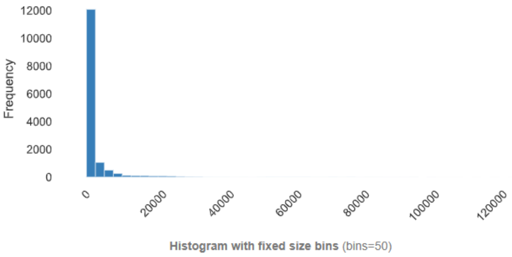
Mean	**2484.9182**
Skewness	**7.3257394**
Sum	36498479

Based on panda-profiling.

Right now, only "Indicators" columns have one value.

The data is skew toward left. I may not use this data afterward.

Detail outputs are in other documents with (**Output #07g**) and Github



Based on the result I analysis above, I noticed that they are all distributed except for those who are skewed toward left. I also noticed they are not evenly distributed. Proof is in Output #08a-g. It also shows that those there is analysis based on the average are skewed toward center, those that are calculated in raw, those data not collected based on the average are skewed toward left.

In addition, however, the data that are collected above, cannot be analyzed because they are just recognized based on "Indicators" and the remaining values are not filtered or divided equally. All the values are sorted by "Indicators" does not matter if the date is set in 2017, or 2019 or if they are 16 years old or granulated at university.

To analysis, to match my purpose of the analysis, I will split the data into 2010-2012, 2013-2015, 2016-2018, and 2019-2021. Originally, I was going to analysis 2017, 2019, and 2021 or 2016-2017, 2018-2019, and 2020-2021. Then I will use 2019-2021 as testing dataset and 2013-2018 as training dataset. There's a tool within Python where it helps to split between training and testing dataset because the data I have already matches with the standard, there's more training set than the testing set.

Once I successfully divide the dataset, I will not be using the dataset in 2010-2012. I believed that there's just too many data to analysis. Although, I will not use the dataset, I will however, keep the dataset in term of backup or use it for further analysis if necessary. For the result and the year that are splinted, you can refer to Output #09 in other documents.

For Average annual hours worked,

Only Testing set based on ‘Characteristics’ will be displayed here. Details will be on the other documents. (Output #10 and Github)

For Average Annual hour worked,

```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

For Average annual wages and salaries,

```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

For Average hourly wages,

```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

Average weekly hours worked,

```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

Hours Worked

```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

Number of Jobs

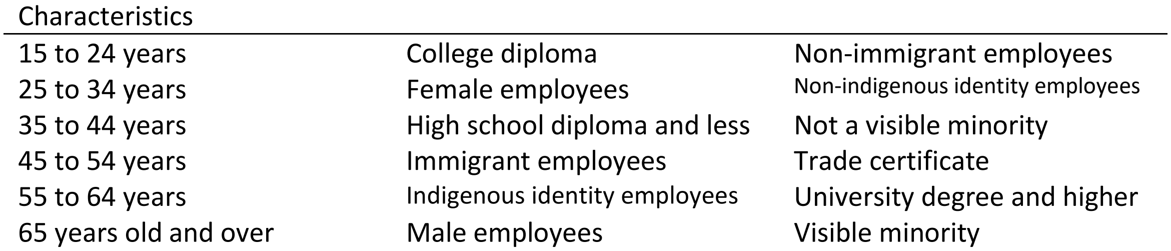
```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

Wages and Salaries

```
testing set
Overall,
Sum :  5669496.0
Mean :  1543.9803921568628
Min/median/max : 565.0 / 1583.0 / 2250.0
Standard Deviation :  242.19188733505
Skewnewss :  -0.9940967289246635
Total size :  3672
```

After numerous attempts to divide the datasets into training and testing datasets. I finally have two sets of sample datasets that I can use to test the results. Each dataset contains 7344 observations for training dataset and 3672 observations for testing dataset. This dataset is more than enough to proceed to the next steps.

However, I need to split the dataset much further. Net, I will use "characteristics" columns to divide much further. However, I will not use all the rows in the dataset. First, I will use based on the age group, six categorical variables. Second, I will use based on the gender group, two variables. Third, I will use based on the education level, three variables. Finally, I will use immigrant levels divide the dataset, two variables. For any other dataset available, I will be going to discard them.



Characteristics
15 to 24 years
25 to 34 years
35 to 44 years
45 to 54 years
55 to 64 years
65 years old and over
College diploma
Female employees
High school diploma and less
Immigrant employees
Indigenous identity employees
Male employees
Non-immigrant employees
Non-indigenous identity employees
Not a visible minority
Trade certificate
University degree and higher
Visible minority

I will discard following rows, "Indigenous identity employees", "non-indigenous identity employees", "Not a visible minority", and "Visible minority". I will not use these data because I am not analyzing any of Indigenous identity or Visible minority that are hired in non-profit organization.

This is done directly via panda, numpy, and aggregation.
Detail outputs are in other documents with (**Output #12a**) and Github

Average Annual Hours worked “testing dataset”:

```
testing set By Age
                            sum         mean    amin  median    amax  size
Characteristics
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                       sum         mean    amin  median    amax  size
Characteristics
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                  sum         mean    amin  median    amax  size
Characteristics
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                              sum         mean    amin  median    amax  size
Characteristics
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

Detail outputs are in other documents with (**Output #12b**) and Github

Average annual wages and salaries in ‘testing dataset’:

```
testing set By Age
                            sum         mean    amin  median    amax  size
Characteristics
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                       sum         mean    amin  median    amax  size
Characteristics
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                   sum         mean    amin  median    amax  size
Characteristics
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                              sum         mean    amin  median    amax  size
Characteristics
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

Detail outputs are in other documents with (**Output #12c**) and Github

Average hourly wage ‘testing dataset’:

```
testing set By Age
                             sum         mean    amin  median    amax  size
Characteristics
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                        sum         mean    amin  median    amax  size
Characteristics
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                    sum         mean    amin  median    amax  size
Characteristics
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                               sum         mean    amin  median    amax  size
Characteristics
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

Detail outputs are in other documents with (**Output #12d**) and Github

Average weekly hours worked ‘testing dataset’:

```
testing set By Age
                            sum         mean    amin  median    amax  size
Characteristics
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                       sum         mean    amin  median    amax  size
Characteristics
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                   sum         mean    amin  median    amax  size
Characteristics
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                              sum         mean    amin  median    amax  size
Characteristics
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

Detail outputs are in other documents with (**Output #12e**) and Github

Hours Worked in ‘testing dataset’:

```
testing set By Age
                            sum         mean    amin  median    amax  size
Characteristics
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                       sum         mean    amin  median    amax  size
Characteristics
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                   sum         mean    amin  median    amax  size
Characteristics
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                              sum         mean    amin  median    amax  size
Characteristics
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

Detail outputs are in other documents with (**Output #12f**) and Github

Number of Job in testing dataset:

```
testing set By Age
                            sum         mean    amin  median    amax  size
Characteristics                                                           
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                       sum         mean    amin  median    amax  size
Characteristics                                                      
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                   sum         mean    amin  median    amax  size
Characteristics                                                                  
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                              sum         mean    amin  median    amax  size
Characteristics                                                             
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

Detail outputs are in other documents with (**Output #12g**) and Github

Wages and Salaries in ‘training dataset’:

```
testing set By Age
                            sum         mean    amin  median    amax  size
Characteristics
15 to 24 years         179872.0   936.833333   713.0   927.5  1281.0   192
25 to 34 years         333662.0  1588.866667  1292.0  1576.0  1870.0   210
35 to 44 years         371386.0  1768.504762  1424.0  1757.0  2092.0   210
45 to 54 years         384987.0  1833.271429  1541.0  1826.5  2191.0   210
55 to 64 years         351843.0  1675.442857  1377.0  1675.0  2071.0   210
65 years old and over  206760.0  1076.875000   565.0  1076.5  1415.0   192
Overall,
Sum :  1828510.0
Mean :  1493.8807189542483
Min/median/max : 565.0 / 1633.0 / 2191.0
Standard Deviation :  357.1076307712267
Skewnewss :  -0.5829393561867995
Total size :  1224

testing set By Gender
                       sum         mean    amin  median    amax  size
Characteristics
Female employees  323755.0  1541.690476  1302.0  1540.0  1773.0   210
Male employees    343101.0  1633.814286  1373.0  1644.0  1821.0   210
Overall,
Sum :  666856.0
Mean :  1587.752380952381
Min/median/max : 1302.0 / 1596.0 / 1821.0
Standard Deviation :  93.82948373587371
Skewnewss :  -0.3050908032633471
Total size :  420

testing set By Education
                                   sum         mean    amin  median    amax  size
Characteristics
High school diploma and less  275636.0  1312.552381  1054.0  1307.5  1667.0   210
Trade certificate             306174.0  1546.333333   789.0  1547.5  1808.0   198
University degree and higher  341795.0  1726.237374  1536.0  1706.5  2043.0   198
Overall,
Sum :  923605.0
Mean :  1524.1006600660066
Min/median/max : 789.0 / 1545.0 / 2043.0
Standard Deviation :  204.02000670247472
Skewnewss :  -0.1568377438071896
Total size :  606

testing set By Immigrant
                              sum         mean    amin  median    amax  size
Characteristics
Immigrant employees      318818.0  1610.191919  1336.0  1589.0  2250.0   198
Non-immigrant employees  310245.0  1566.893939  1315.0  1570.0  1767.0   198
Overall,
Sum :  629063.0
Mean :  1588.5429292929293
Min/median/max : 1315.0 / 1580.0 / 2250.0
Standard Deviation :  112.27714481578558
Skewnewss :  1.6694520665533967
Total size :  396
```

As final steps, I have decided to divide it into province levels to show more accurate result that I did. Skipping this step does not easily able to analysis the data properly. Since, there’s 13 provinces too much to analysis, with advice from instructor, I have decided to analysis only five provinces. The next page will illustrate the result of my analysis.

All the indicator analysis is done in following format: (Output #13)

```
<class 'pandas.core.frame.DataFrame'>
Index: 450 entries, 85137 to 100796
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   REF_DATE             450 non-null    int64
 1   DGUID                450 non-null    object
 2   Sector               450 non-null    object
 3   Characteristics      450 non-null    object
 4   Indicators           450 non-null    object
 5   UOM                  450 non-null    object
 6   SCALAR_FACTOR        450 non-null    object
 7   VALUE                450 non-null    float64
 8   GEO_Alberta          450 non-null    bool  (“One hand encoding”)
 9   GEO_British Columbia  450 non-null    bool  (“One hand encoding”)
 10  GEO_Nova Scotia      450 non-null    bool  (“One hand encoding”)
 11  GEO_Ontario          450 non-null    bool  (“One hand encoding”)
 12  GEO_Quebec           450 non-null    bool  (“One hand encoding”)
 13  Age_group            450 non-null    int64 (“[20, 30, 40, 50, 60, 70]”)
dtypes: bool(5), float64(1), int64(2), object(6)
memory usage: 37.4+ KB
None
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 150 entries, 85053 to 100684
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   REF_DATE             150 non-null    int64
 1   DGUID                150 non-null    object
 2   Sector               150 non-null    object
 3   Characteristics      150 non-null    object
 4   Indicators           150 non-null    object
 5   UOM                  150 non-null    object
 6   SCALAR_FACTOR        150 non-null    object
 7   VALUE                150 non-null    float64
 8   GEO_Alberta          150 non-null    bool    (“One hand encoding”)
 9   GEO_British Columbia  150 non-null    bool    (“One hand encoding”)
 10  GEO_Nova Scotia      150 non-null    bool    (“One hand encoding”)
 11  GEO_Ontario          150 non-null    bool    (“One hand encoding”)
 12  GEO_Quebec           150 non-null    bool    (“One hand encoding”)
 13  Gender_group         150 non-null    int32   (“[1 0]”)
dtypes: bool(5), float64(1), int32(1), int64(1), object(6)
memory usage: 11.9+ KB
None
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 225 entries, 85109 to 100754
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   REF_DATE             225 non-null    int64
 1   DGUID                225 non-null    object
 2   Sector               225 non-null    object
 3   Characteristics      225 non-null    object
 4   Indicators           225 non-null    object
 5   UOM                  225 non-null    object
 6   SCALAR_FACTOR        225 non-null    object
 7   VALUE                225 non-null    float64
 8   GEO_Alberta          225 non-null    bool    (“One hand encoding”)
 9   GEO_British Columbia  225 non-null   bool    (“One hand encoding”)
 10  GEO_Nova Scotia      225 non-null    bool    (“One hand encoding”)
 11  GEO_Ontario          225 non-null    bool    (“One hand encoding”)
 12  GEO_Quebec           225 non-null    bool    (“One hand encoding”)
 13  Education_group      225 non-null    int64   (“[1, 2, 3]”)
dtypes: bool(5), float64(1), int64(2), object(6)
memory usage: 18.7+ KB
None
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 150 entries, 85067 to 100698
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   REF_DATE             150 non-null    int64
 1   DGUID                150 non-null    object
 2   Sector               150 non-null    object
 3   Characteristics      150 non-null    object
 4   Indicators           150 non-null    object
 5   UOM                  150 non-null    object
 6   SCALAR_FACTOR        150 non-null    object
 7   VALUE                150 non-null    float64
 8   GEO_Alberta          150 non-null    bool   (“One hand encoding”)
 9   GEO_British Columbia  150 non-null   bool   (“One hand encoding”)
 10  GEO_Nova Scotia      150 non-null    bool   (“One hand encoding”)
 11  GEO_Ontario          150 non-null    bool   (“One hand encoding”)
 12  GEO_Quebec           150 non-null    bool   (“One hand encoding”)
 13  Immigrant_status     150 non-null    int32  (“[0, 1]”)
dtypes: bool(5), float64(1), int32(1), int64(1), object(6)
memory usage: 11.9+ KB
None
```

After each division of the five provinces, I finally can be able to analysis the result much better than before. However, with a lot of data being divided into multiple small datasets, it is generally taking a long time to analyze the divided datasets. In addition, there are now long codes to execute. Due to that reason, it takes a lot of time to execute and is very time consuming. Also, because of the space that take in the report, I end up removing some of the analysis from this report.

Result of Analysis

Overall conclusion

Based on the results from all the Indicators and Characteristics, employees from non-profit organizations get older and get more educated, they get to work more and get paid more. However, this may be true only if they worked annually or weekly. Some of the results are a little bit different.

As for work hours by age group, employees in 20s and 70s get the lowest working hours. Employees in their 30s to 60s work most of the hours. There's also the biggest gap between 20s and 30s and 60s and 70s. I can assume that people in their 20s don't have enough experience or may be prepared to be ready for job experiences. People in their 70s are ready for their retirement and in the age group to get pension. The highest working hours that they received are the 50s. For annual and weekly worked hours, both minimum/median/maximum working hours for each age group are either increased or decreased. When measured not annually or weekly, minimum wages stay the same. It's the medium/maximum wages that get increased or decreased.

As for wages by age group, employees in their 20s and 70s get the lowest working hours. However, employees in their 20s get the lowest wages. 70s get much higher wages than 20s including higher wages groups in 20s match with lower wages group in 70s. There's also a gap between 20s and 30s and 60s and 70s, only in weekly wages. For average and weekly wages, it only applies to the 20s and 30s. Some 60s and 70s have duplicated observations in terms of wages. In terms of Wages not measured weekly or annually, minimum wages stay the same for all ages, its maximum earn wages that are different depend on the age group. They look like those in work hours not measured on weekly or annually.

As for working hours for education, people who have the highest education get more working hours than those in lowest education. The only difference is those who work not annually or weekly, minimum work hours stayed the same. For those measured annually or weekly, lowest education values change as they get more education. One interesting part I find is trade certification. By measuring annually and weekly, there's some employees who get the highest number of working hours than their fellow coworkers. By measure working hours itself, they may work similarly as their coworkers even less than those with higher or lower education.

As for wages based on education, people who have the highest education also get to pay more than those in lower education. I notice more when I measure those with Wages and Salaries themselves. When measured wages and salaries themselves, employees with highest education, get up 6 times more than those with lowest education. However, like hours worked, those with trade certification get lowest wages when compared to measure by annually or weekly. However, they aren't high as those in highest education.

As for gender based on working hours based, Male employees generally work more than the female employees do. There are several samples that prove that male work more than those in highest working hours for female employees. Also, there's several female employees who work less than the lowest male employees do. That's just annual based. For weekly based, there's male employees who work more than the highest female employees do. However, for lowest, I find both male and female employees work only minimum hours. However, there's more female employees who does than the male employees do. As for measure itself for gender group, female employees work way more than the male counterpart does. There are several samples showing that female employees work more than highest of the male employees does.

As for gender based on wages, like Hour worked for annual and weekly, male generally get to paid more than the female does. Some female employees paid less than lowest male employees paid annually. Some of male employees paid more than highest female employees paid annually. For weekly, some of male workers higher than the highest female workers do. For weekly, both female employees and male employees get lower payment. However, more female workers paid the lowest than the fellow male workers do. Finally, just wages alone, female make more than the male counterpart does. Like worked hours, females, work more and paid more. There are some who get more than the highest male earn.

As for Immigrant for work hours, based on annually and weekly, Immigrants work the most hours, but they also work the lower hours. Compared to annually, there’s less immigrant who worked lower than the those who are nonimmigrant. If I were to check the hour worked itself, non-immigrant work more than the those who immigrant to Canada. There’s a lot more non-immigrant employees who work more than the those in immigrant.

As for Immigrant for wages, based on annually and weekly, both immigrant and non-immigrant earn the highest however, some of the immigrant employees earn less than non-immigrant. For wages by looking at immigrant alone, non-immigrant make more than the immigrant employees. Some of the non-immigrant employees they make much higher than their following immigrant employees. However, both immigrant and non-immigrant look like earn the less as well.

Annual Hour Worked based on ‘Age group’ (Output #14a)

Based on the result given above from the five provinces, individuals who are older generally worked more. The linear regression lines in the graph represent that employees who are older work more hours. The regression line is true for some data but not for all the data.

Based on the result, those who are in their 20s and 70s work less than those who are in their 30s to 60s. Also, I have noticed that employees who are 50s and older are starting to work less. Based on this statistic, employees who are in their 20s may not have experience working at same times, people who are older than 50s are more likely to reach their retirement. In addition to this, people in their 20s worked 800 – 1100 hours/annually vs people in 30s worked in 1400 – 1800 hours/annually. Furthermore, people in 60s worked 1400-1800 hours/annually vs people in 70s worked 850 – 1300 hours/annually.

I have also noticed that people who are in their 50s and 70s have some dramatic gap between each other’s. For employees aged 50s we have someone who works more than their coworkers who are same age group. They are from Nova Scotia in 2019-2021. I also noticed, employees whose age 70s have someone who work less than their coworkers who are same age group. They are from Quebec in 2021.

Annual Hour Worked based on 'Education level' (Output #14b)

As for education for five provinces, people who get the highest education get to work more than those who got lowest education. The linear regression represented employees who get more education get to work more. This concept is indeed very true.

Based on the results given, those who get higher education get to work more than those who got lower education. Those with higher education worked approximately, 1500 hors/annually to 1800 hours/annually compared to lower education who work 1000 hours/annually to 1500 hours/annually. This shows that non-profit organizations do care about education. This shows that those who get the highest work hours at lower level have the lowest work hours at the highest level of education.

In addition, I also noticed that some employees that have a trade certification have a bigger gap of working hours than their following workers. The employees who have trade certification worked in Alberta in 2019. At same time and place, there's employees who graduated from university or higher but earn less than the one from trade certification. There's another group of employees who earn the same or more than employees from trade certification but have university degrees and from Nova Scotia or Ontario.

Those who have lowest educations worked between 1450-1550 hours/annually, worked from Alberta in 2019 and those who have highest educations worked from Quebec in 2020. The remaining amount of people who worked have trade certification from a variety of provinces.

Annual Hour Worked based on ‘Gender group’ (Output #14c)

As for the gender, more male employees are more likely to work more than fellow female employees. I also noticed that some of the female employees are more likely to work less than their coworkers and male employees do. There are several female employees who work less than the lowest male employee work (1500 hours/annually). At same time, there’s several male employees who work more than the highest female employees worked (1670 hours/annually). Lowest female employees work around (1450 hours/annually) while lowest male employees work around 1500 hours/annually.

According to the data given, there’s appropriately around 16 female employees who works less than the lowest male employees do. They are from variety of the provinces; however, the lowest female working is from Quebec. There are approximately 19 male employees who worked more than the highest working female employees. The highest male employee worked is from Alberta.

Annual Hour Worked based on ‘Immigrant group’ (Output #14d)

As for the immigrant from five provinces, the highest working employees are immigrants. Also at the same time, the lowest working employees are also immigrant. There’s variety of immigrant employees who works more than their fellow immigrant coworkers. This also apply same for lowest working immigrant employees.

For immigrant who worked more than highest non-immigrant employee is also the highest immigrant employee who worked in non-profit organization. They are from Ontario in 2021. Highest immigrant employee worked around 1697 hours/annually vs highest non-immigrant employee worked 1691 hours/annually. For lowest non-immigrant employee, there’s about three more employees who worked less than the non-immigrant employee. All of them are from Quebec between 2019-2021.

Average Annual Wages for ‘age group’ (Output #15a)

For the age group for ‘average annual wages’, employees who get older get more pay. The linear regression line represented in this graph shows that, the ages they get, the more salaries that they received. This trend is like one for annual hourly work.

Employees who are aged between 40s to 60s get the highest pay amount. Employees who are in their 20s get the lowest pay. There’s a big gap between 20s and 30s. However, the highest salaries between 60s and 70s, there are a lot of gaps in between. Highest salaries in age of 70s are $50445 dollars/annually. There are 58 employees in their 60s that are getting more wages than people in their 70s. The highest pay in their 60s is $77141 dollars/annually from Ontario in 2021. Some of the employees who are in their 50s get more wages than their coworkers. Those employees are all from Ontario during 2020-2021. Employees in their 50s earn the most but also work longer too. Also, some of the working group in 30s to 60s get similar wages group those are in medium or higher wages in 70s. Lowest ages group in 30s, 40s, 50s and 60s, $27239/annual, $36021/annual, $37873/annual, and $34578/annual. There are employees in their 70s who earn same as those in lowest group in those range.

Based on this analysis, like hourly worked, people in 30s-60s get most of the experiences and have not reached their retirement yet. People who are 20s or 70s may not have enough experiences or is retired.

Average Annual Wages for ‘education’ (Output #15b)

The linear regression lines tell me that those who have higher education get more wages than who have less education. This is similar result as annual hourly worked.

Those who have the highest wages and have a big gap are occurred in those who have trade certification. As for the remaining, there’s gap in between but it’s not high as those who have trade certification. As for those who take the lowest education, their highest wages are like the one who take almost lowest wages in highest education.

Those who are making the highest money in trade certification are from Alberta between 2019-2021. However, there’s still a lot of employees from highest educations who make the same wages as those in trade certification. As for the lowest education who make highest wages, the value is $41040 dollars/annually and they are from Ontario in 2021. As for the highest education who make lowest wages, they make $42833 dollars/annually from Nova Scotia in 2019. However, the highest education with lowest wages makes about $1000 dollars/annually from the lowest educations with highest wages.

Average Annual Wages for 'gender' (Output #15c)

Male employees get the more wages than the female employees. Some of the female employees get less wages while some of the male employees get more wages. In fact, male employees get much more wages than those the female employees.

The lowest male employees make around $36768 dollars/annually from Nova Scotia in 2019. Compared to this, there's nine female employees who make less than lowest male employees. They are from Quebec, Nova Scotia, and British Columbia.

However, the highest female employees make around $59616 dollars/annually from Ontario in 2021. There are approximately 26 male employees who make more than the highest female employees does mostly from Alberta, BC, and Ontario.

Average Annual Wages for 'immigrant' (Output #15c)

Both immigrant and non-immigrant employees make about similar annual wages. Unlike the number of hours they worked, both immigrant and non-immigrant employees make about the same salaries.

The highest annual wages between immigrant and non-immigrant employees are about $300 dollar/annually differences. The lowest annual wages differences between the two are $200 dollar/annually. The highest salaries for both immigrant and non-immigrants are $63265 and $63582 dollars/annually from Ontario in 2021. The lowest salaries for both immigrant and non-immigrant are $31977 and $34013 dollars/annually from Quebec in 2019 and Nova Scotia in 2019.

Average hourly wages for ‘age group’ (Output #16a)

Unlike the annual wages or hourly worked, people who are older than 50s have possibility to make more wages per hour. However, as they get older, the gap between lowest and highest ages they make get bigger as they get older. The linear line is like the annual wages or hourly work.

Employees in their 20s still get the lowest wages per hour and people in 50s and 60s still make similar salaries as annual and hourly. However, people in their 70s are different. Some make higher salaries per hour. Some make lower salaries per hour.

Also compared to annually, those who make higher annual salaries than highest salaries in 60s make around $41.60, $43.20, $43.62 dollars per hour. There are two more candidates that make more than those who make highest annually. They are all from Ontario. For the references, those who make highest salaries in 60s make $42.70 per hour. They are second highest.

As for people in their 70s, the highest hourly salaries are $46.07 per hour. For the references, they make more hourly wages than employees in 50s and 60s. As for average, they make about $33.25 per hour. As for the lowest wages per hour in their 70s, those ages consist of people in mostly in their 20s and 30s. This is the trend because they are closer to the retirement age.

Average hourly wages for ‘Gender’ (Output #16c)

Like the annual wages and hourly worked, Male employees still make more money per hours than the female employees do. Unlike the previous session, there isn’t really differences between lowest hourly wages between male and female employees does but highest hourly wages do reflect the results.

There are only two female employees who make less than male lowest hourly wages. They are all from Nova Scotia in 2019. This is different for the highest. Therea are more male employees who is making more than the female highest hourly wages. They are all from a variety of provinces and happens throughout the years.

Average hourly wages for ‘immigrant’ (Output #16c)

Both non-immigrant and immigrant population have split equally. In fact, both minimum and maximum only differences range between about $1 and 50 cents per hour. Similar for the median and mean as well. They are about $1 different. Most of the observations that are toward both immigrant and non-immigrants are in between $30-35 per hours.

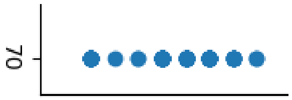
Lowest non-immigrant and immigrant employees earn about $21.23 per hours and $20.60 annually both from Nova Scotia in 2019. Highest non-immigrant and immigrant employees earn about $38.02 annually and $38.63 annually from Alberta in 2019 and Ontario in 2020.

Average weekly hours worked by ‘Age group’ (Output #17a)

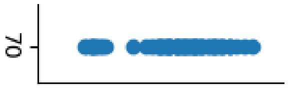
Like the annual worked hours, people who are in 20s and 70s worked less and people in 30s to 60s worked more. There’s a dramatic gap between 20s and 30s and 60s and 70s. The regression line represents as we get older, they work more.

I also noticed that unlike annually worked hours, I noticed that each age group there isn’t dramatically gap between highest and lowest. This means, for example, age of 50s and 70s from annual worked hours, there is not dramatic gap between highest earning and second highest earning.

The three highest earnings in 50s for annual hours worked about 39 hours/weekly, and 38 hours/weekly. For the 70s, the differences between maximum worked hours and minimum worked hours are around 7 hours/weekly and evenly distributed unless the annual hours worked.



Average weekly hours worked (70s)



Average annual hours worked (70s)

The highest weekly worked hours are from 50s with 39 hours per week. The lowest weekly worked hours are from 20s with 16 hours per week. The median and mean value for this age group is around 17 hours per week and 17.44 hours per week.

Average weekly hours worked based on ‘Education group’ (Output #17b)

For employees, like annual hours, employees who have the highest educations work more than the employees with lowest education. Also, the linear line represents this concept as well.

In annual hours, the gap between employees who work highest in lowest education and employees who work lowest in highest education is high. However, this is not the case for weekly hours worked. The only difference between those are about two hours. Unlike the previous statement, there’s more employees who work less than highest employees for lowest education.

Also, for trade certification, there’s still a gap between highest education working hours and their coworkers. The highest education with the highest working hours is 35 hours per week. The highest education with the second highest working hours is around 33 hours per week. Highest working hours for highest education and medium (trade) education are the same, 35 hours per week.

For the references those who work highest in lowest education worked around 28 hours. Meanwhile those who are in highest education in lowest working hours are 30 hours per week.

Average weekly work hours for on ‘Gender group’ (Output #17c)

As for the gender, like the annual, more male employees are more likely to work more than fellow female employees. However, I also noticed that both male employees and female employees also work the lowest. However, there’s more female works who work less than the male does. There are several female employees who work less (28 hours per week) than the lowest male employee work (also 28 hours per week). At same time, there’s several male employees who work more than the highest female employees worked (32 hours/weekly).

According to the data given, there’s appropriately around 16 female employees who works lowest and two male employees do the same too. Female employees are from mostly in BC and Quebec. Two employees who work less are from Quebec in 2020. Those female workers who work highest amount are from Nova Scotia and Ontario between 2020-2021. Approximately 14 Male workers who work higher than these two female workers do.

Average weekly Hour Worked based on ‘Immigrant group’ (Output #17c)

As for the immigrant from five provinces, the highest working employees are both immigrants and non-immigrant. However, the lowest working employees are immigrant.

Both immigrant and non-immigrant employees worked 33 hours per week. They are from Nova Scotia and Ontario. For immigrant who worked for 27 hours per week is the lowest from Quebec in 2019. For those who works for 28 hours per week, second lowest, there are more immigrant who work 28 per week than non-immigrant employee does, there’s about three employees from Quebec 2019-2021 who does. For immigrant who works 28 hours are from mostly Quebec except for one where from Ontario in 2020.

Hours worked in ‘Age group’ (Output #18a)

Like the weekly and annual worked hours, people in their 20s and 70s have the least number of working hours. People in their 30s to 60s work the most. Also, like annual and weekly worked hours, there’s big gap between 20s and 30s and 60s and 70s have most gap, especially with highest working hours between each these age group.

Unlike the weekly and annual worked hours, the linear represented here are not increasing. They stay straight. To add the references, the histogram represented also not distributed and skewed toward the left. Also, I noticed that people in the lowest of all ages group are approximately the same or similar. However, those in 30s to 60s are the highest working but also have big and multiple gaps in between.

The lowest working hours per each ages group are roughly from 290 hours to 1500 hours. The lowest age group is 293 hours in 70s and highest lowest working hours age group is 1494 hours in 50s. For highest annual working employees who earn more than their fellow coworkers worked about 1635 hours to 1808 hours in Nova Scotia. They earn about 38-39 hours/weekly.

There’s a gap in employees in 70s who worked more than their coworkers. Those in the 70s worked more than 50000 hours from Ontario from 2019-2021. The rest of the employees in 70s work less than 50000 hours. For references there’s gaps in annual working hours in employees in 70s who worked less than their coworkers. For weekly, the wages are distributed almost equally.

Hours Worked by 'Education level' (Output #18b)

Based on my observation, people who are in higher education worked more than people who are in lower education. By looking at Hours worked, this result makes me much clearer now. Unlike annual hourly worked and weekly worked, I am seeing more gaps as there's higher education. Compared to annual hourly worked or weekly hours worked, the lowest hours worked are all the same but more education, there's bigger and more gaps occurring. For annual and weekly, lowest hourly worked also increased.

For higher education, the highest working hours did increase but the exception applies to trade certification. In fact, the highest working hours for trade certification are less than 100000 hours, 99207 hours. Also, interesting enough, trade certification always has gaps when there's for annual and weekly worked hours. However, I find that there's an additional gap but bigger in those who worked less hours. Addition big gap occurred between 35706 hours in Ontario in 2019 to 114397 hours in Quebec also in 2019.

Hour worked for ‘By Gender’ (Output #18c)

Unlike the Annual and Weekly worked hours, female employees work more than their counter male part. I found it very interesting because when measuring annual and weekly worked hours, there’s more male tends to work more than their female counterpart. For annual working hours, some of the female work less than lowest male working annually. This applies same for weekly working hours.

Those female employees who work more than the highest male workers are from Ontario and Quebec. Lowest female employees working are 455862 hours that are not in annually or weekly. To add a point, those who worked lowest for annually or weekly are more likely to work higher than anticipated hours. I am assuming that the data that they collect probably the total number of hours or may be filtered based on weekly or annually.

Hour worked by ‘Immigrant status’ (Output #18c)

Based on the result, non-immigrant employees work more than their immigrant counterpart. This is little bit different than the annual and weekly, where either immigrant employees worked more than non-immigrant employees or both immigrant and non-immigrant works similar rate to each other. There are approximately 12 non-immigrant employees who works more than the highest work for those who immigrant works around 447657 hours. Most of these employees are from Quebec or Ontario in 2019-2021 and between 500000 hours to 900000 hours.

Wages and Salaries for Age group (Output #19a)

Like the annual wages and hours worked, people in their 20s and 70s get the lowest pay and 30s-60s get the highest pay. People in their 50s get the highest pay. There’s a big gap between 20s and 30s and between 60s and 70s.

People in their 50s get the highest salaries and wages. Then people in 40s and then 60s. People in 50s and 60s get the biggest gap in between one highest. People in 50s, there’s gap between $8986 million (2021, Quebec) and $11978 million (2019, Ontario) People in 60s, there’s gap between $7272 million (2021, Ontario) and $9885 million (2019, Ontario).

Wages and Salaries by Education (Output #19b)

Based on the result, the more education that they take, the more salaries they receive. The lowest salaries for all employees are between $10-91 million. What is very noticeable is that trade certification gets the lowest highest wages/salaries. This is different from annual or weekly wages/salaries. Also, those who take highest educations, we are seeing in different wages observations. They have the highest wages and lowest wages at same time. I saw same concept for investigation on immigrant vs non-immigrant for both annual and weekly.

Just look at highest education alone, there’s three observations where they make $25000 million or more. Another six samples between $15000 to 20000 million. Then other samples between maximum high school wages ($5300 million) and $15000 million. Finally, other samples who make same as trade certification or high school graduate do.

Employees who have trade certification, the highest wages that are not annually or weekly are the lowest for both lower and higher education. However, there’s some employees who also make lower than those in trade certification even with both the highest and lowest education available.

Wages and Salaries by Gender (Output #19c)

Ironically, compare to hours worked and annual or weekly wages, female get to earn more wages than the male employees part does. There's about several Female observations show that they make more than highest male employees who earn their wages. That's around 16000 million.

For example, the highest female salaries that she makes is approximately $59616 annually. However, she makes around $22094 million. However, male salaries that the make is approximately $36768 annually. However, he makes around $192.0 million. This shows that, by calculating individual salaries and salaries annually or even weekly, the result may be very different. This experience might relate to perhaps the how much hours they worked or perhaps the way of calculating but showing that female earn more but by looking into weekly or yearly, they make less or males just make more money. For references, there are ten female employees from Ontario and Quebec who earn more than the highest male earn not by annual or weekly.

Wages and Salaries by Immigrant status (Output #19d)

Non-immigrant employees make more wages than those who aren't immigrant. Some of the non-immigrant employees' same number of salaries as those who are immigrant. Also, to add the point, both of non-immigrant and immigrant employees also have gap in between fellow coworkers.

First, there are ten non-immigrant that are making more wages than those who are immigrant to Canada. Highest immigrant sample is $16030 million. Almost most of them are from Ontario or Quebec.

Second, for immigrant observations, there are top three immigrant employees who make more than $13000. They are all from Ontario from 2019-2021. Then there are second top three immigrant employees who make less than $13000 million but more than $7000 million. They too are also from Ontario. Finally, but last, there's a lot of samples occurred in below $6000 million.

I find those interesting because when I was observed annual or weekly, it was always immigrant who make more and sometimes less. However, if calculate not by weekly or annual, the opposite might happen.

Conclusion

Based on the data that I have collected, depend on the situation like age, education, gender, and immigrant status, working for non-profit organizations have future inside them.

Employees who are younger and have at least a trade certification had the potential to be working inside non-profit organizations. Although, at their young age, they will get lower wages, but as they get more experience and older, they would like to receive a much higher salary. Same for working hours, as they get older, they can be able to work more flexibly. Some can choose to work a lot and at the same time choose to work less. Plus, if there's more educational background, there's a chance you get paid more and probably work less. As a bonus, there's a lot of positions available throughout the age group and higher education background.

However, people who are older than 60s are not a good recommendation work for non-profit organizations, as their wages and working hours shrink dramatically. Although some of the elderly still get a good income and working hours, most of them won't. Therefore, if the employees are before 60s or/and have lower education degrees, there's no future for them available.

As for the gender group and immigrant group, I do not think it doesn't influence too much in working for non-profit organizations. As short term, non-immigrant, and female group more likely to get paid more and work more but as a long term especially weekly and annually, more immigrant and male group are more likely to benefit the most. As a result, I do not think checking based on gender or immigrant group won't make differences in term of working for non-profit organizations.

The significant predictors here in this research and analysis are the ‘Indicators’ and ‘Characteristics’. Most of the research that is done here is mostly based on ‘Indicators’ and ‘Characteristics’. The dataset is really divided based on the ‘Indicators’ and ‘Characteristics’. Although not all the ‘Characteristics’ data weren’t used, dividing dataset based on their characteristics type do help me to generate result and analysis more accurately. ‘Indicators’ columns do help a lot too. When it was in big dataset with all ‘Indicators’ there, there was difficult and unable compare between the two observations. By dividing it, it helps to easily compare the observations, some observations with different datasets.

As mentioned above, younger with more education they get, there’s potential to work for non-profit organizations. Younger employees with more education are working more and likely paying more and have a greater number of positions available. However, if employees are in their 20s, they will likely get lower working hours and low wages as well until they reach about 30s, (25 – 34 years old).

Decision Tree Classifier performs better for the classification. However, when it gets deeper with a lot of data to analysis, the story goes differently. Although it might be well classified, it will have difficult time to analysis and read the data. My main problem using this decision tree is that I suddenly have a lot of datasets to deal with. This not only required me to read the data and analysis it but also required me to code (aggregate, loc) as well. There is also use of class method or copy and paste same code required. Not only that, but it also took a lot of time dealing with technical difficulties with debugging and needed to spend a lot of time reading the result and referencing in the report. Decision tree classifier performs better when there is less data but at same time spends a lot of time if there’s more data to dealing with.

Data mining/selection does allow to perform more accurate results than without it. Without data mining, the result will get much worse and much more complex. The current state of analysis is still complex and dealing with a lot of issues including looking into the code, dataset, and a lot of scrolling in the long-written python script (or technical report). Data selection allows me to select appropriate data type and source of the data to analysis. (Javatpoint) Without it, it would be much worse position to be right now. There may be some bias inside dataset whether that’s intentional or not but without it, the result would be also significantly different perhaps more bias inside.

According to statistics Canada, it might say it is very accurate but depend on the complex of dataset and the coding it may not be that accurate. From the beginning, I had to deal with a lot of cleaning and splitting the dataset. During that process, I might intentionally or accidentally have removed some of the observations that are very important. Therefore, they might have less chance of being more accurate. Not only that, but the data also inside in that dataset might be unintentionally removed by wrong coding. Since there’s too much splitting and too many analyzing, I could have removed some of the data by accidentally typing the wrong code. During this analysis, I could have different results just with one tiny coding error, where I could have shown the different results. For hours wages section, I could not able to analysis the educations section, because there was code error occurring. I tried to fix the problem but ended up not fix the problems. It was due to the complexity of my coding and length of time to fix.

The information that I gather from is in Statistic Canada. They have gathered the data from various sources from their database. How they gather the database, and how they expressed is little confusing. Initially, when I did this presentation, it was very confusing. I had no idea what some of the indicators mean here. However, as I analysis much further, I started to get some of the ideas very clear. However, there’s still some confusion that I still need to figure out though. As I was doing the analysis, there was one component, where I got very confused, the indicators showing annual/weekly wages vs wages and salaries. The result inside the data show units while wages and salaries all sudden show millions. I still have confusion and understand why the data is structured like that. As a result, those components make me show different results. Therefore, after all indicators are analyzed, I have decided to skip some of the components to analysis. One of them is Number of Job because of the confusion of the results.

References

5 types of classification algorithms in machine learning. MonkeyLearn Blog. (2020, August 26). https://monkeylearn.com/blog/classification-algorithms/

Cause Leadership. (2023, February 15). *Hiring Trends in Nonprofit Organizations*. November 20, 2023, https://causeleadership.com/hiring-trends-in-nonprofit-organizations/

Cloud Software Group. (n.d.). *What is cluster analysis?*. TIBCO Software. https://www.tibco.com/reference-center/what-is-cluster-analysis

Cote, C. (2021, October 26). *What is predictive analytics? 5 examples: HBS Online*. Business Insights Blog. https://online.hbs.edu/blog/post/predictive-analytics

GeeksforGeeks. (2023, February 9). *Data mining in R*. GeeksforGeeks. https://www.geeksforgeeks.org/data-mining-in-r/

Gopalan, B. (2021, March 18). *Is decision tree a classification or regression model?*. Numpy Ninja. https://www.numpyninja.com/post/is-decision-tree-a-classification-or-regression-model

Government of Canada, S. C. (2021, April 30). *Non-profit organizations and Volunteering Satellite Account: Human resources module, 2010 to 2019*. The Daily — Non-Profit Organizations and Volunteering Satellite Account: Human Resources Module, 2010 to 2019. https://www150.statcan.gc.ca/n1/daily-quotidien/210430/dq210430d-eng.htm

Javatpoint. (n.d.). *Data Selection in Data Mining*. Data Selection in Data Mining - Javatpoint. https://www.javatpoint.com/data-selection-in-data-mining

Kapil, A. R. (2022, October 1). Decision Tree Algorithm in Machine Learning: Advantages, Disadvantages, and Limitations. *Advantages and disadvantages of decision tree in machine learning*. November 20, 2023, https://www.analytixlabs.co.in/blog/decision-tree-algorithm/

Kumar, A. (2023, September 11). *What is data mining?: Data Mining Tools for python*. Taazaa. https://www.taazaa.com/python-tools-for-data-mining/

Minitab Blog. (2016, September 7). *How to identify the most important predictor variables in regression models*. How to Identify the Most Important Predictor Variables in Regression Models. https://blog.minitab.com/en/adventures-in-statistics-2/how-to-identify-the-most-important-predictor-variables-in-regression-models

Navlani, A. (2023, February 23). *Python decision tree classification tutorial: Scikit-Learn Decisiontreeclassifier*. DataCamp. https://www.datacamp.com/tutorial/decision-tree-classification-python

Statistic Canada. (2023, September 18). *Employment in the non-profit sector by demographic characteristic* [Data set]. Open Government Portal. https://open.canada.ca/data/en/dataset/edc0fe3c-23a3-4ccf-929b-8a385b62f6c3

Stedman, C., & Hughes, A. (2021, September 7). *What is data mining?*. Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining#:~:text=Data%20mining%20is%20the%20process,make%20more%2Dinformed%20business%20decisions.

Stedman, C., & Hughes, A. (2021, September 7). *What is data mining?*. Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining