A
Seminar Report
On

ANALYZING DATASET OF THE IMMIGRATION OF PEOPLE TO CANADA FROM WORLDWIDE

By

CHETAN DANGE

Roll No: - 19

DEPARTMENT OF COMPUTER ENGINEERING,

SINHGAD INSTITUTE OF TECHNOLOGY,

LONAVALA

C E R T I F I C A T E

This is to Certify that seminar work entitled **"ANALYZING DATASET OF THE IMMIGRATION OF THE PEOPLE TO CANADA FROM WORLDWIDE"** is a bonafide work carried out in the sixth semester by **"CHETAN DANGE** having **Roll No- 19"** in partial fulfillment for the award of Bachelor of Engineering in Computer Engineering from Sinhgad Institute of Technology, Lonavala during the academic year 2019- 2020.

SIGNATURE	**SIGNATURE**	**SIGNATURE**
Name of Guide	**Name of Seminar Coordinator**	**HOD**

A C K N O W L E D G M E N T

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude I give to my guide, [Prof. V.S. KADAM], whose contribution to stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of the Faculties of Computer Department, who permitted me to use all required resources and the necessary materials to complete the task "*ANALYZING DATASET OF IMMIGRATION OF PEOPLE TO CANADA FROM WORLDWIDE*". Also, I appreciate the guidance given by other supervisors as well as the panels especially in our project presentation that has improved our presentation skills thanks to their comments and advice. At last, I would like to say that, it would not have been possible without the help of these people, so thank you to all those who contributed to this project.

P A G E I N D E X

F I G U R E I N D E X

Analyzing Dataset of the Immigration of People to Canada from Worldwide

CHETAN DANGE
Department of Computer Science and Engineering
Savitribai Phule Pune University
Lonavala, India
dangechetan09@gmail.com

PROF. V.S. KADAM
Department of Computer Science and Engineering
Savitribai Phule Pune University
Lonavala, India

***Abstract*— From long before the computer age was started, we are collecting data in almost every field, after years of collecting data, now we have a large set of data to work on and it will continue growing exponentially in every coming year. We have to make something useful from this data. So, we use Data Science to explore this data and extract valuable information from it. In this report, we are going to work on the dataset which has the data of the Immigration of People to Canada from Worldwide.**

1. INTRODUCTION

The intension of this report is to deal with the real-world data, which we have collected throughout the years till now and will use data science and some machine learning algorithms to deal with these datasets.

1.1 What is Data Science?

Data Science is the field of Computer Science in which we deal with the real-time datasets to extract meaningful information from the raw data and use this information in the real-world Projects.

This is very helpful in the ongoing real-world projects because it helps us to predict the valuable information needed by extracting a valuable understanding from such a large dataset, which is almost impossible for a human to do it manually.

In this report, we are also going to deal with such a real-world dataset of the Immigration of People to Canada from Worldwide.

To deal with these datasets we will require a computer language. One of the most popular and widely used language is Python. Python is considered the most powerful and suitable language for Data Analyzing.
Some of the Python libraries which we will use is - pandas, Numpy, Matplotlib, sklearn (Sci-kit Learn).

For Data Science, Jupyter Notebook provides the best user-interface environment. It has an already installed python interpreter and there is no need to install it manually. Here, Jupyter Notebook provided by IBM Watson Studio is used to extract meaningful insights from the dataset.

1.2 Source of Data

We got this dataset from the - 'United Nations Population Division Department of Economic and Social Affairs'.

Link:-
https://www.un.org/en/development/desa/population/index.asp

FEATURES OF THE DATA SET

a) Countries
b) Continent
c) Region
d) Developing Regions or Not
e) Years containing the no. of Immigrants from 1980 to 2013.

This dataset contains annual data on the flow of international immigrants to Canada. This Dataset contains both inflows and outflows of the people according to their place of birth, citizenship, and previous/next residence for both foreigners and nationals.

1.3 Importing Dataset into Jupyter Notebook

The Dataset which is provided by the United Nations Population Division Department is in excel format (xlrd format). So, we will be importing this Dataset into the Jupyter Lab environment and will be performing various Data analysis on the Dataset and will try to extract some meaningful insights which will give is a deeper understanding of the facts.

The imported data looks something like the figure shown in:-
Fig- 1.1*(Table of Immigration of People to Canada).*

	Country	Continent	Region	DevName	1980	1981	1982	1983	1984	1985	...	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
0	Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	...	2978	3436	3009	2652	2111	1746	1758	2203	2635	2004
1	Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	...	1450	1223	856	702	560	716	561	539	620	603
2	Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	...	3616	3626	4807	3623	4005	5393	4752	4325	3774	4331
3	American Samoa	Oceania	Polynesia	Developing regions	0	1	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
4	Andorra	Europe	Southern Europe	Developed regions	0	0	0	0	0	0	...	0	0	1	1	0	0	0	0	1	1

Fig- 1.1 *Immigration of People to Canada.*

2. DATA ANALYSIS

We will now try to understand the imported data before performing operations on them. Understanding the data includes, knowing the size, shape, and some information about the statistical data of the dataset.

To know the statistical information of the dataset python provides a function called `describe()`.
This function will give statistical information like min_count, max_count, mean, median, and information on quartile data of the dataset. The output of this function will somewhat look like the figure shown in ***Fig-2.1(****Statistical information of the dataset****)***

	AREA	REG	DEV	1980
count	195.000000	195.000000	195.000000	195.000000
mean	912.764103	1249.015385	901.753846	508.394872
std	13.082835	1185.526885	0.431878	1949.588546
min	903.000000	905.000000	901.000000	0.000000
25%	903.000000	914.000000	902.000000	0.000000
50%	908.000000	922.000000	902.000000	13.000000
75%	922.000000	925.500000	902.000000	251.500000
max	935.000000	5501.000000	902.000000	22045.000000

Fig-2.1 (*Statistical information of the dataset****)***

Now let's try to know the countries from where people immigrated to Canada. We can do so by filtering the country name and the years of which we have to know the data.
Here's an example is shown to filter the distinct country for the specific years, e.g.

```
df_can[['Country', 1980, 1981, 1982, 1983, 1984, 1985]]
```

The output of the above line of code is shown in ***Fig-2.2(****Filtered countries and their data of the given years****).***

	Country	1980	1981	1982	1983	1984	1985
0	Afghanistan	16	39	39	47	71	340
1	Albania	1	0	0	0	0	0
2	Algeria	80	67	71	69	63	44
3	American Samoa	0	1	0	0	0	0
4	Andorra	0	0	0	0	0	0
...	...	...	...	...	...	...	...
190	Viet Nam	1191	1829	2162	3404	7583	5907
191	Western Sahara	0	0	0	0	0	0
192	Yemen	1	2	1	6	0	18
193	Zambia	11	17	11	7	16	9
194	Zimbabwe	72	114	102	44	32	29

Fig-2.2(*Filtered countries and their data of the given years****)***

The above figure shows the list of the countries and the people immigrated from that country to Canada in the year 1980 to 1985.

Since we found this let us also try to find the top 5 countries from which people immigrated to Canada in all these years in total. To do this we will sort the data of the list in descending order of the total immigrants to Canada and then display its top 5 rows. We will insert a 'Total' column into the list and store it with the total value of the immigrants in all these years and then perform the sort operation. E.g.

```
df_can['Total'] = df_can.sum(axis=1)

df_can.sort_values(['Total'], ascending=False, axis=0, inplace=True)

df_can.head(5)
```

The output of these lines of code is shown in ***Fig-2.3(****List of top 5 countries immigrated to Canada****)***

Country	India	China	United Kingdom	Philippines	Pakistan
1980	8880	5123	22045	6051	978
1981	8670	6682	24796	5921	972
1982	8147	3308	20620	5249	1201
1983	7338	1863	10015	4562	900
1984	5704	1527	10170	3801	668

Fig-2.3(*List of top 5 countries immigrated to Canada****)***

The figure shows the top 5 countries along with the count of peoples from which people are most involved in immigration to Canada are India, China, United Kingdom, Philippines, and Pakistan.

These are some operation of Data analysis which we can perform on the Datasets.
Now, let's get on to Data Visualization techniques which are used to visualize the information from the Datasets.

3. DATA VISUALIZATION

Data visualization is the process of representation of the data in graphical/visual form. Some method of representing it in visual form is to represent it using various types of graphical charts like

a) Line Plots
b) Area Chart
c) Bar Charts
d) Pie Charts
e) Box Plots
f) Folium Maps
g) Regression Plots

All these plots are very important and useful to understand the dataset and it helps to display the information extracted from it in a simple and easily understandable way.
Note: *For all these plots you have to import a python library called matplotlib as*

```
import matplotlib as mpl
import matplotlib.pyplot as plt
```

3.1 Line Plots

A line chart or line plot is a type of plot which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. We use Line Plots when we have a continuous data set. These are best suited for trend-based visualizations of data over a period of time.

Case Study*: - In 2010, Haiti suffered a catastrophic magnitude 7.0 earthquake. The quake caused widespread devastation and loss of life and about three million people were affected by this natural disaster. As part of Canada's humanitarian effort, the Government of Canada stepped up its effort in accepting refugees from Haiti. So, let's quickly visualize the immigration of people from Haiti to Canada.*

```
haiti = df_can.loc['Haiti', years]

haiti.index = haiti.index.map(int)
haiti.plot(kind='line')


plt.title('Immigration from Haiti')
plt.ylabel('Number of immigrants')
plt.xlabel('Years')

plt.text(2000, 6000, '2010 Earthquake')

plt.show()
```

The above code will take the dataset as a reference and plot the Line Chart of the Immigration of Haiti population to Canada. The Line Plot of the same is shown in ***Fig-3.1(****Line Plot of Immigration of Haiti people to Canada****).***

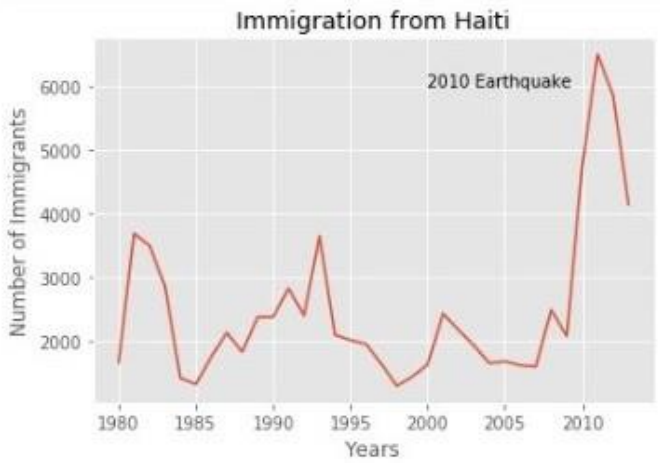

Fig-3.1(*Line Plot of Immigration of Haiti people to Canada****)***

In the above figure, we can see that the sudden increase in the no. of immigrants in 2010, So, we can depict that the sudden increase is due to the earthquake in Haiti which leads to the immigration of a large amount of population to Canada.

Line Plot of India's vs China's population Immigration to Canada

As we plot the line plot for Haiti, similarly we can also plot it for India and China, the process for plotting is shown below:

```
df_CI=df_can.loc[['India','China'],year]
df_CI = df_CI.transpose()
df_CI.plot(kind='line')

plt.title("The graph of INDIANS and
CHINESE migrated to CANADA")
plt.xlabel("YEARS")
plt.ylabel("No. of Migrants")
```

The plot between India and China will be depicted by two lines showing the trend of population Immigration of both the countries separately. The plot will automatically assign each line a different color and will automatically assign the legend/label for it.
The plot for the above line of code is shown in ***Fig-3.2(****Line*

*plot of India's and China's immigration to Canada**).***

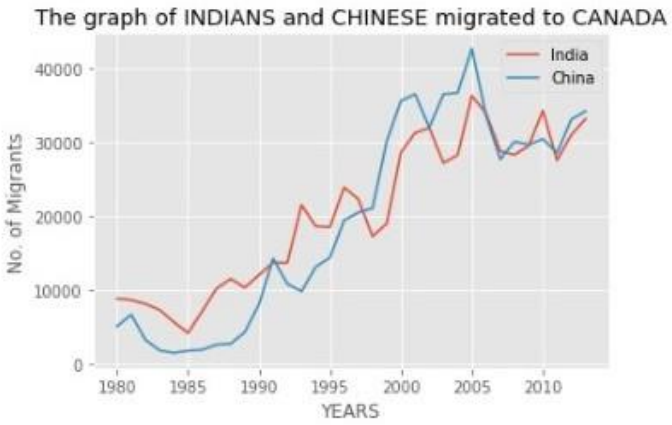

Fig-3.2(*Line plot of India's and China's immigration to Canada**)***

3.2 Area Plots

Area plot is the visualization of the data by covering the area in the plot which is directly proportional to the increase in the dependent variable, in our case the dependent variable is the number of people immigrated to Canada.

Now, we are going to plot the Immigration of the people to Canada of the top 5 countries. To do so we will sort the dataset in descending order and store its top 5 rows to another table and will use that to plot the Area Chart. The code for the same is given below:

```
df_can.sort_values(['Total'],
      ascending=False,axis=0,
      inplace=True)

df_top5 = df_top5[years].transpose()

df_top5.index = df_top5.index.map(int)

df_top5.plot(
             kind='area',
             stacked=True,
             figsize=(20, 10),
            )

plt.title('Immigration Trend of Top 5
Countries')

plt.ylabel('Number of Immigrants')

plt.xlabel('Years')

plt.show()
```

The plot for the above line of code is given in ***Fig-3.3 (****Area Chart of Immigration of people to Canada of the top 5 countries**).***

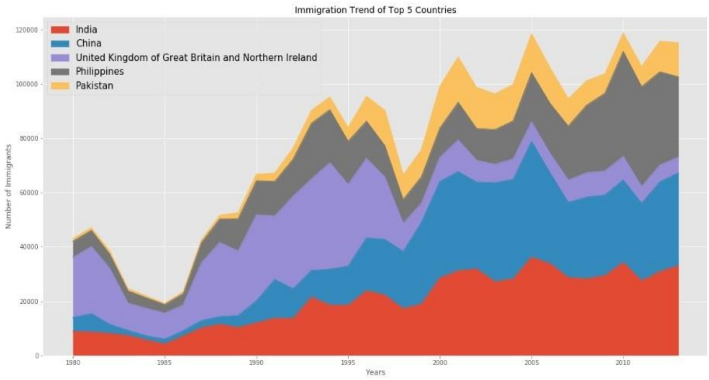

Fig-3.3 (*Area Chart of Immigration of people to Canada of top 5 countries**)***

3.3 Bar Charts

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

Here we will plot the bar chart of the immigration of People to Canada from Iceland in the year 1980 to 2013. In the Bar chart, the height of the bar will represent the number of people immigrated to Canada.

Here the code:

```
df_iceland = df_can.loc['Iceland', years]

df_iceland.plot(kind='bar', figsize=(10,
6), rot=90)

plt.xlabel('Year')

plt.ylabel('Number of Immigrants')

plt.title('Icelandic Immigrants to Canada
from 1980 to 2013')

plt.annotate(
                    '',
                     xy=(32, 70),
                     xytext=(28, 20),
                     xycoords='data',

                    arrowprops=dict(arrows
                    tyle='->',
                    connectionstyle='arc3'
                    , color='blue', lw=2)
                )

plt.show()
```

The output of the following lines of code is given in ***Fig-3.4 (****Bar Chart of the Immigration of people to Canada from Iceland**).***

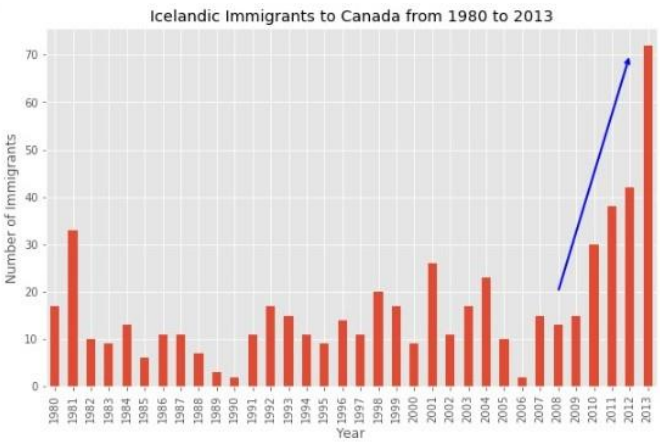

Fig-3.4 (*Bar chart of Immigration of people to Canada from Iceland****)***

The arrow shows the sudden increase in the population immigration which occurred due to the Financial Crisis in Iceland in the year 2008-2011.

3.4 Pie Charts

A pie chart is a circular graphic that displays numeric proportions by dividing a circle (or pie) into proportional slices. We can create pie charts in Matplotlib by passing in the kind=pie keyword.

Now, we will plot the pie chart of the population Immigration to Canada by Continent.
We will plot it for various continents and show its percentage. The code for the same is :

```
colors_list = ['gold', 'yellowgreen',
'lightcoral', 'lightskyblue',
'lightgreen', 'pink']
explode_list = [0.1, 0, 0, 0, 0.1, 0.1]

df_continents['Total'].plot
      (
            kind='pie',
            figsize=(15, 6),
            autopct='%1.1f%%',
            startangle=90,
            shadow=True,
            labels=None,
            pctdistance=1.12,
            colors=colors_list,
            explode=explode_list
      )


plt.title('Immigration to Canada by
Continent [1980 - 2013]', y=1.12)
plt.axis('equal')

plt.legend(labels=df_continents.index,
loc='upper left')

plt.show()
```

The plot for the above line of code is shown in ***Fig- 3.5 (****Pie Chart of population Immigration to Canada by Continent****).***

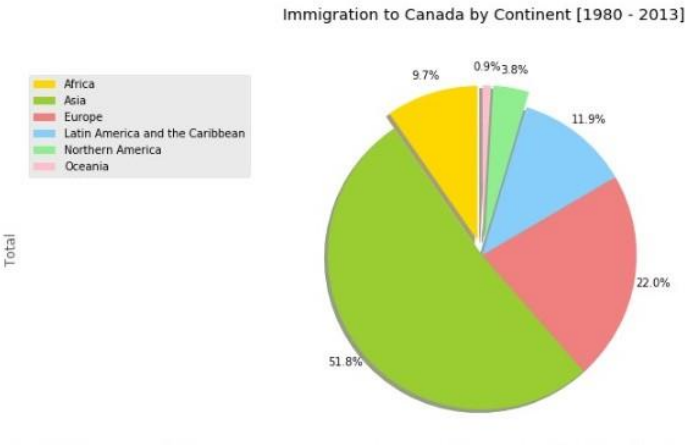

Fig- 3.5(*Pie Chart of population Immigration to Canada by Continent****)***

3.5 Box Plots

A box plot is a way of statistically representing the distribution *of the data through five main dimensions:*

- ***a) Minimum:*** *Smallest number in the Dataset*
- ***b) First Quartile:*** *Middle no. between minimum and median.*
- ***c) Second Quartile (Median):*** *Middle no. of the sorted Dataset.*
- ***d) Third Quartile:*** *Middle no. between median and maximum.*
- ***e) Maximum:*** *Highest no. in the Dataset.*

It is difficult to understand the statistical information in the table format, but the Box Plot provides the best way to visualize those statistical data and makes the understanding of the Dataset a lot better.

Now we will visualize the statistical data of Japanese immigrants from the year 1980-2013. Here's the code for the same:

```
df_japan = df_can.loc[['Japan'],
years].transpose()

df_japan.plot(kind='box', figsize=(8,
6))

plt.title('Box plot of Japanese
Immigrants from 1980 - 2013')

plt.ylabel('Number of Immigrants')

plt.show()
```

The visualization of the statistical data of Japanese Immigration to Canada is shown in ***Fig- 3.6(****Box Plot of Japanese Immigration to Canada****).***

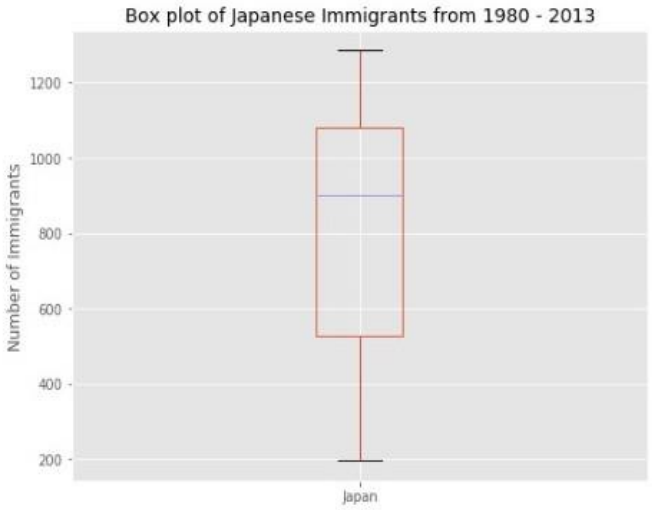

Fig- 3.6(*Box Plot of Japanese Immigration to Canada****)***

Similarly, if we want to compare the statistical data of Immigration to Canada of the countries India and China we can do that by including India and China on the x-axis and no. of immigrants in the y-axis.
Here's the code for the same

```
df_CI=df_can.loc[['India','China'],years
].transpose()

df_CI.plot(kind='box',    figsize=(8,6),
color='darkred')

plt.title('Number of Immigrants from
India and China to Canada in year 1980 to
2013')
plt.ylabel('Number of Immigrants')
plt.xlabel('Countries')
plt.show()
```

The Box Plot visualization of Indian and Chinese Immigrants to Canada is shown in ***Fig- 3.7 (****Box Plot Visualization of the Indian and Chinese immigrants to Canada****).***

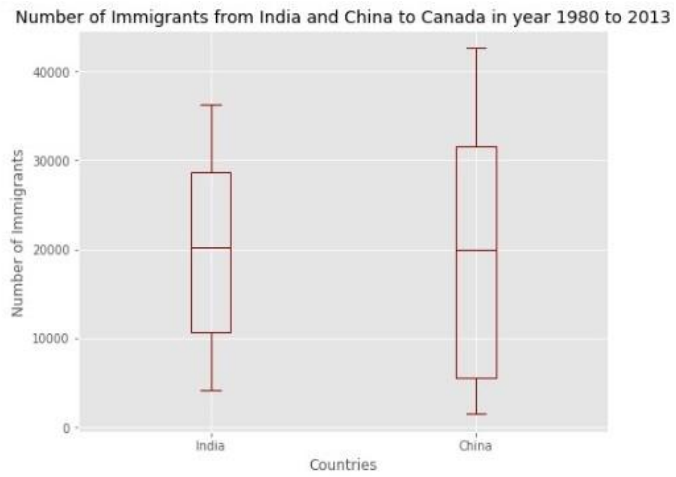

Fig- 3.7 (*Box Plot visualization of Indian and Chinese Immigrants to Canada****)***

3.6 Folium Maps

Folium Maps are the method of visualization of the data to a map using choropleth visualization meaning if the larger the number of observations is observed in a particular area the darker will be its shade.
In our case, if the immigrants count to Canada are larger in some particular country the darker will be its shade.

Folium maps are the best for representing geographical data like weather maps, population maps, river maps, Terrain maps, etc.
To use it you have to import a folium library (import folium).
To use folium, we have to import the world.json file too.
Here's the code:

```
world_geo = r'world_countries.json'

threshold_scale =
np.linspace(df_can['Total'].min(),

df_can['Total'].max(), 6, dtype=int)

threshold_scale =
threshold_scale.tolist()

threshold_scale[-1] = threshold_scale[-
1] + 1

world_map = folium.Map(location=[0, 0],
zoom_start=2, tiles='Mapbox Bright')

world_map.choropleth

(
    geo_data=world_geo,

    data=df_can,

    columns=['Country', 'Total'],

    key_on='feature.properties.name',

    threshold_scale=threshold_scale,

    fill_color='YlOrRd',

    fill_opacity=0.7,

    line_opacity=0.2,

    legend_name='Immigration to Canada',

    reset=True
)

World_map
```

This code will generate the folium choropleth map of the dataset and will shade all countries according to the dataset. It is a good way of visualizing the dataset. As per our choropleth map legend the darker the color of the country and the closer the color to red, the higher the number of immigrants from that country.

Fig- 3.8 (*Folium choropleth map of the immigrants to Canada from worldwide****)*** *shows the Folium Choropleth representation of the Dataset.*

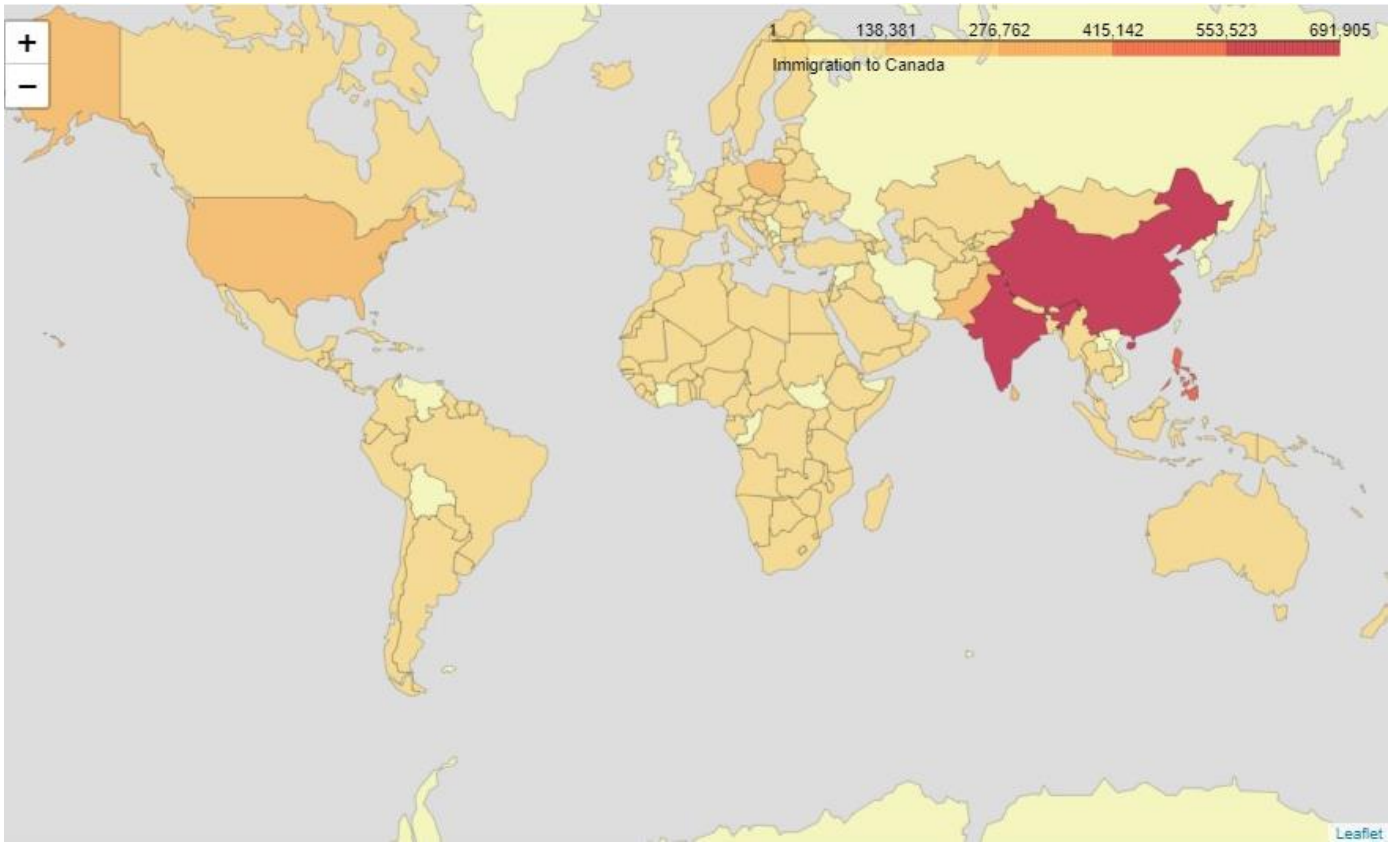

Fig- 3.8 (*Folium Choropleth map of the Immigrants to Canada from worldwide****)***

3.7 Regression Plots

There is more than one technique to do this operation. The data can be branched into two groups 1. The unsupervised technique, 2. Supervised technique. The Unsupervised technique is used for clustering, dimension reduction, density estimation, market basket analysis. The data in unsupervised are all unlabeled and the model works on its own to unrevealed new information. The supervised technique is used for classification and regression models, there are used to predict the future data by analysis of the vast amount of previous data. The work uses a regression model to predict the data instead of the classification model because in our data it is not classifying anything like whether the developing countries or developed countries are immigrating. The work focuses only on predicting the increased number of immigrants to Canada for the top 5 countries.

The model of multi-variable linear regression can be represented by the following:

Y = a_1*X_1 + a_2*X_2 + a_3*X_3 ……. a_n*X_n + b

The work used the multiple regression model because it is using multiple independent variables to predict one dependent variable.

India – Canada Immigration

India is a developing country with a vast population. There are 1.1 billion people and it will increase in the upcoming years. The huge growth of the population has led to being the main reason for the increase in immigrants. The prediction is made for the number of immigrants to Canada from India. The prediction results as can out it look very good.

The ***Fig- 3.9 (****Percentile vs No. of Immigrants from India to Canada****)*** *shows the regression plot prediction of the increase in the number of Immigrants moving from India to Canada.*

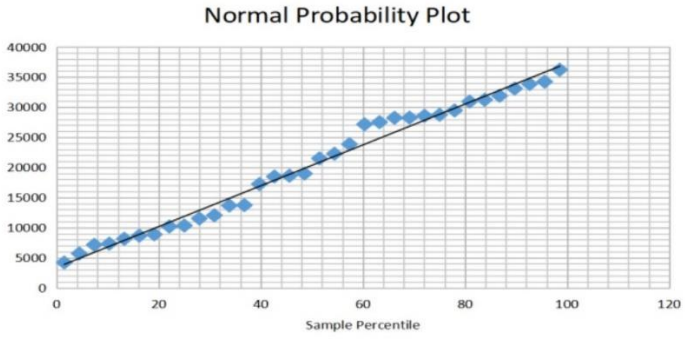

Fig- 3.9 (*Percentile vs No. of Immigrants from India to Canada****)***

Hence, through the regression analysis, we can say that the No. of Immigrants from India to Canada is going to increase with accuracy and error of:

Accuracy (r) *- 94.8%*
Error *– 3175(Less)*

4. CONCLUSION

Hence, from all the above analysis techniques we can conclude that the top five countries which are involved in immigration to Canada are India, China, United Kingdom, Philippines, and Pakistan.

And the plot of India vs China immigration depicts that both the countries are following the same curve and the population immigration from India and China to Canada is almost the same.

We also got the percentage of Immigrants from various Continent to Canada, the percentage of immigration from various Continents are:

Africa: - 9.7%
Asia: - 51.8%
Europe: - 22%
Latin America and the Caribbean: - 11.9%
Northern America: - 3.8%
Oceania: - 0.9%

And, from using the box plot we came to know the statistical data of Japan, India, and China, and thus we can compare them with this technique.

Using the Folium Choropleth technique, we depicted the immigration of people to Canada form worldwide using maps.

From Regression Analysis of Indian Population Immigration to Canada, we concluded that the No. of Immigrants from India to Canada is going to increase
in years to come with an Accuracy of 94.8% and an Error of 3175 (Less).

B I B L I O G R A P H Y

1. Data Analyzing Immigration to Canada using Predictive Analysis (Multiple Linear and Non-Linear Regression).
https://www.ijeat.org/wp-content/uploads/papers/v9i2/B2281129219.pdf

2 . Analysis of the Entrepreneurial Immigrant Profile in Spain by Genoveva Millán, Virginia Navajas, and Ricardo Hernández in April 2018.
https://www.researchgate.net/publication/330881048_Analysis_of_the_Entrepreneurial_Immigrant_Profile_in_Spain/fulltext/5c598de3299bf1d14cadb404/Analysis-of-the-Entrepreneurial-Immigrant-Profile-in-Spain.pdf

3. International Migration: A Panel Data Analysis Of Economic and NonEconomic determinants by Anna Maria Mayda in May 2005.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=72544

4. Quality of Work Experience and Economic Development: Estimates Using Canadian Immigrant Data by Serge Coulombe, Gilles Grenier, Serge Nadeau on_January2011.
https://www.researchgate.net/publication/254439974_Quality_of_Work_Experience_and_Economic_Development_Estimates_Using_Canadian_Immigrant_Data

5. Student Flow and Migration: An Empirical Analysis by Axel Dreher and Panu Poutvaara in March 2006.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=731765

6. United Nations Population Division Department of Economic and Social Affairs.
https://www.un.org/en/development/desa/population/index.asp