DATA ANALYTICS FOR BUSINESS,
St. CLAIR COLLEGE

Interim Report

GLOBAL MIGRATION INSIGHTS

SUBMITTED BY
Group_12

FOR
DAB322 – CAPSTONE PROJECT 1 W 2024 - 002

Sahith Valluripally - W0825415
Deepak Pattem - W0758417
Sudharshni Radhakrishnan - W0833471
Iyanthika Basnayake - W0824786
Dhruva Ray - W082653

GLOBAL MIGRATION INSIGHTS

Project Overview

Migration, a multifaceted phenomenon, encapsulates the deliberate movement of individuals, families, or large groups from one geographical location to another with the intent of establishing either temporary or permanent residence. This global occurrence spans vast distances and often involves crossing national borders. According to the World Economic Forum, approximately 200 million international migrants currently exist worldwide, constituting 3.5 percent of the global population. Remarkably, this figure has already surpassed certain projections for the year 2050, showcasing the increasing prevalence of human migration. Since 1970, the number of individuals residing in a country other than their country of birth has tripled.

The multifaceted motivations driving people to embark on migratory journeys are encapsulated in four overarching categories, as succinctly delineated by the BBC. Economic migration, propelled by the pursuit of employment opportunities and career advancements, interlaces with social migration, where the quest for an enhanced quality of life or proximity to family and friends becomes a guiding force. Simultaneously, political migration emerges as a poignant driver, with individuals seeking refuge from political persecution or fleeing regions marred by conflict and war. Environmental factors, including the aftermath of natural disasters such as fires and floods, further contribute to the intricate tapestry of reasons underlying human migration.

Understanding and anticipating future migration trends assume paramount importance in various spheres, including urban planning, international trade, public health, conservation, and policymaking. A nuanced approach to predicting these patterns involves three key methodologies: the Early Warning System, Forecasting, and Foresight. The Early Warning System employs a combination of quantitative and qualitative data to monitor real-time population movements, setting predefined thresholds to trigger proactive responses. Forecasting relies on statistical modeling techniques, leveraging historical quantitative migration data to predict future trends with a medium to long-term horizon. In contrast, Foresight employs qualitative scenario methods, weaving narratives about the future of migration that explore potential structural changes and their repercussions.

Problem Statement

In this expansive context, a program emerges as aiming to predict future human migrations between countries through the convergence of machine learning and Python. The roadmap delineates a meticulous journey, starting with the importation of requisite software libraries and dataset exploration. This analytical odyssey delves into data analysis, partitioning the dataset into test and training sets, and proceeding to train the model on the latter. The subsequent steps involve making predictions on the test data, evaluating the model's performance, and drawing insightful conclusions from these evaluations.

Motivations for migration are diverse, falling into four principal categories outlined:
Economic Migration: Individuals move in pursuit of employment opportunities or to navigate specific career trajectories. Social Migration: Relocation is driven by a desire for an improved quality of life, proximity to family or friends, or both.Political Migration: Individuals migrate to escape political persecution or the ravages of war.Environmental Migration: Migration is prompted by natural disasters such as fires and floods.Anticipating future migration patterns is crucial for urban planning, international trade, disease control, conservation, and policy formulation. The three main predictive approaches include,

- Utilizing quantitative and qualitative data to monitor potential drivers and movements of populations in real-time, providing short-term estimations in rapidly changing contexts.
- Predicting future migration flows and trends using quantitative modeling methods based on historical data.
- Employing qualitative scenario methods to envision potential future migration scenarios, considering structural changes and their consequences.

Analysis

The subsequent sections of this article outline a program designed to predict future human migrations between countries using machine learning and Python. The roadmap includes the following steps:

1. Import the Required Software Libraries: Gather the necessary tools for the predictive modeling task.
2. Access and Import the Dataset: Acquire and load the relevant migration data for analysis.
3. Data Analysis and Exploration: Examine and explore the dataset to gain insights and identify patterns.
4. Split the Data: Divide the dataset into training and test sets for model training and evaluation.
5. Model Training: Train the machine learning model using the training data
6. Make Predictions: Utilize the trained model to predict migrations on the test data.
7. Evaluate Performance: Assess the model's accuracy and effectiveness in predicting future migrations.
8. Draw Conclusions: Derive meaningful insights and conclusions from the evaluation results.

Ultimately, the overarching goal of this program is to enhance our ability to forecast and understand the complexities of human migrations between different countries in forthcoming time periods.

Data Exploration & Visualization

Data Overview

Columns as of now, namely
'STRUCTURE', 'STRUCTURE_ID', 'STRUCTURE_NAME', 'ACTION', 'FREQ',
'Frequency of observation', 'TERRITORIAL_LEVEL', 'Territorial level',
'TERRITORIAL_TYPE', 'Territorial typology', 'REF_AREA',
'Reference area', 'MEASURE', 'Measure2', 'AGE', 'Age3', 'SEX', 'Sex4',
'UNIT_MEASURE', 'Unit of measure', 'TIME_PERIOD', 'Time period',
'OBS_VALUE', 'Observation value', 'COUNTRY', 'Country5', 'OBS_STATUS',
'Observation status', 'UNIT_MULT', 'Unit multiplier', 'DECIMALS'

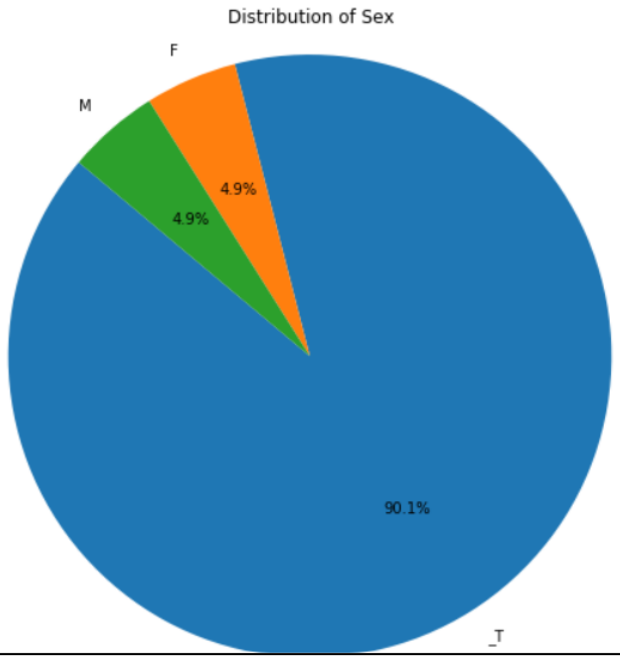
```
[3]: # Display the first five rows of the dataset
     print(data.head())
```

```
  STRUCTURE                                STRUCTURE_ID  \
0  DATAFLOW  OECD.CFE.EDS:DSD_REG_DEMO@DF_MIGR_FLOW(1.0)
1  DATAFLOW  OECD.CFE.EDS:DSD_REG_DEMO@DF_MIGR_FLOW(1.0)
2  DATAFLOW  OECD.CFE.EDS:DSD_REG_DEMO@DF_MIGR_FLOW(1.0)
3  DATAFLOW  OECD.CFE.EDS:DSD_REG_DEMO@DF_MIGR_FLOW(1.0)
4  DATAFLOW  OECD.CFE.EDS:DSD_REG_DEMO@DF_MIGR_FLOW(1.0)

                          STRUCTURE_NAME ACTION FREQ  \
0  International migration flows - Regions      I    A
1  International migration flows - Regions      I    A
2  International migration flows - Regions      I    A
3  International migration flows - Regions      I    A
4  International migration flows - Regions      I    A

  Frequency of observation TERRITORIAL_LEVEL Territorial level  \
0                   Annual               TL3               TL3
1                   Annual               TL3               TL3
2                   Annual               TL3               TL3
3                   Annual               TL3               TL3
4                   Annual               TL3               TL3

  TERRITORIAL_TYPE Territorial typology  ... OBS_VALUE Observation value  \
0               _Z       Not applicable  ...      0.00               NaN
1               _Z       Not applicable  ...      0.03               NaN
2               _Z       Not applicable  ...      0.02               NaN
3               _Z       Not applicable  ...      0.06               NaN
4               _Z       Not applicable  ...      0.02               NaN

  COUNTRY Country5 OBS_STATUS Observation status UNIT_MULT Unit multiplier  \
0     CAN   Canada          A       Normal value         0           Units
1     CAN   Canada          A       Normal value         0           Units
```

Exploratory Data Analysis

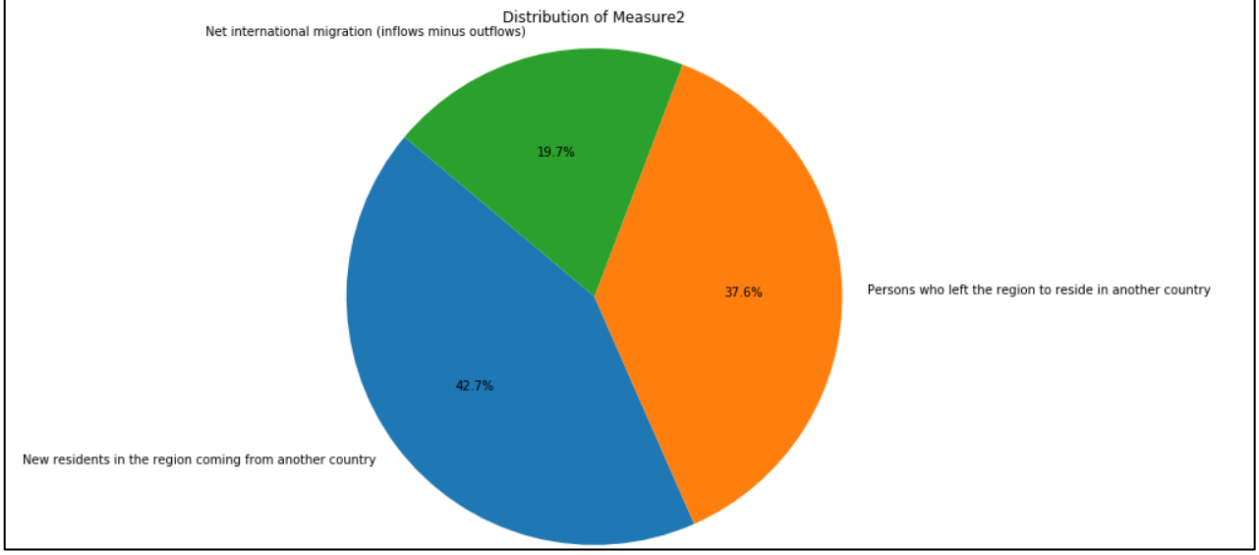
Counts pf the frequency of each gender

```
# Count the frequency of each sex
sex_counts = data['SEX'].value_counts()

# Plotting a pie chart
plt.figure(figsize=(8, 8))
plt.pie(sex_counts, labels=sex_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Sex')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```



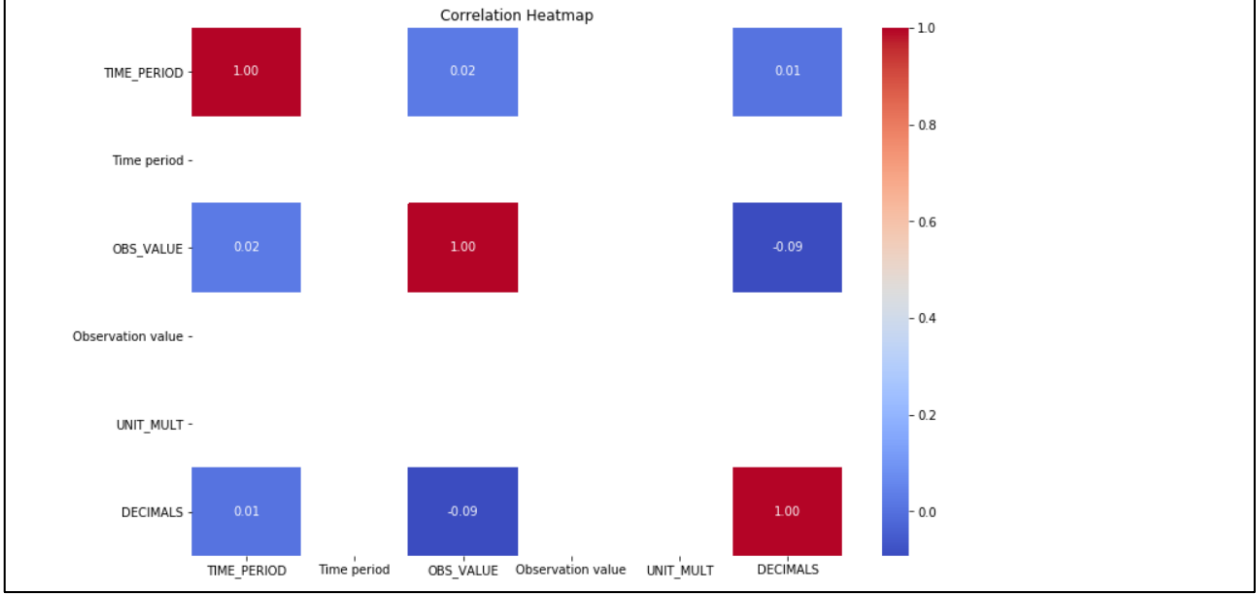
Let’s check the overview of the immigrations. Here we can observe that the people who have migrated to other countries.

```
# Count the frequency of each value in 'Measure2' column
measure2_counts = data['Measure2'].value_counts()

# Plotting a pie chart for 'Measure2'
plt.figure(figsize=(8, 8))
plt.pie(measure2_counts, labels=measure2_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Measure2')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```



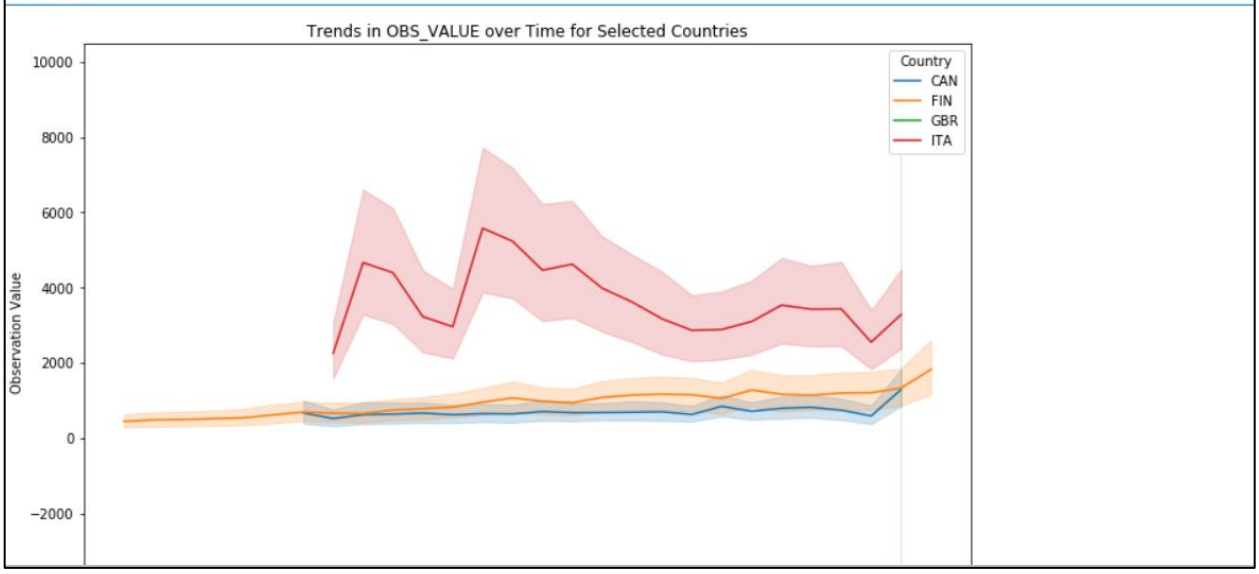
overview by using .corr() method

```
[13]: # Correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```

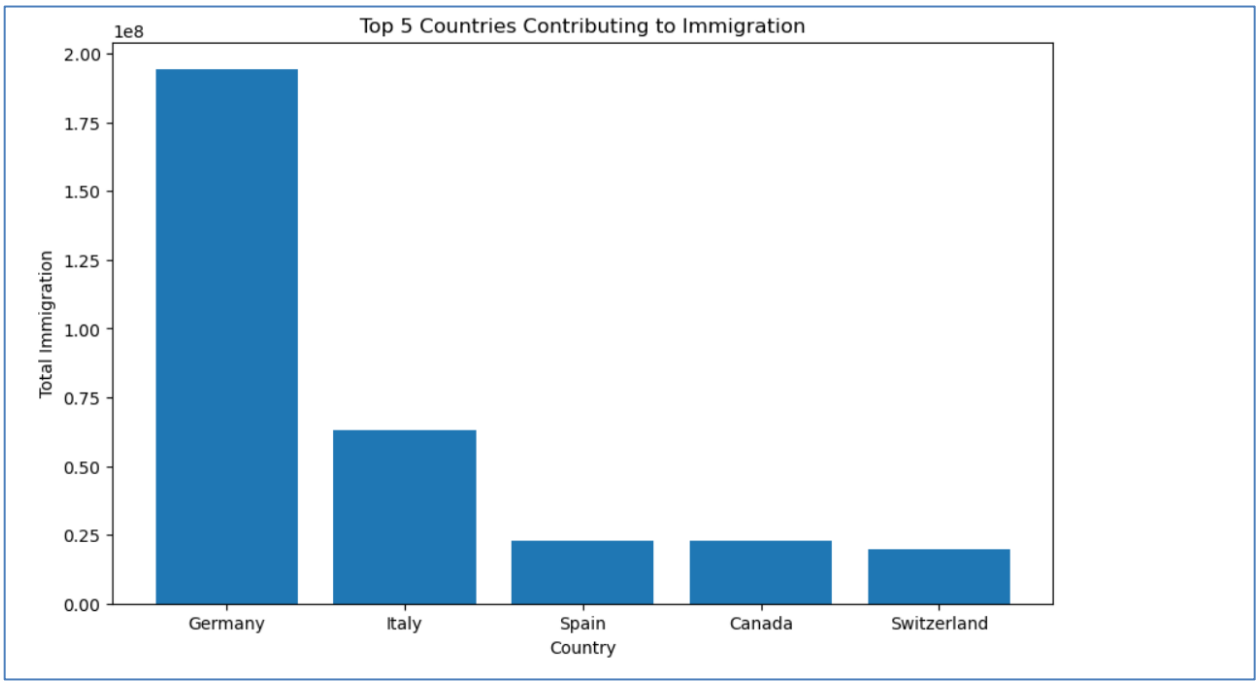


Examining trends in OBS_VALUE over time for specific countries

```
countries_to_plot = ['CAN', 'FIN', 'GBR', 'ITA']  # Choose specific countries for analysis
plt.figure(figsize=(12, 8))
for country in countries_to_plot:
    country_data = data[data['COUNTRY'] == country]
    sns.lineplot(x='TIME_PERIOD', y='OBS_VALUE', data=country_data, label=country)
plt.title('Trends in OBS_VALUE over Time for Selected Countries')
plt.xlabel('Time Period')
plt.ylabel('Observation Value')
plt.legend(title='Country')
plt.xticks(rotation=45)
plt.show()
```

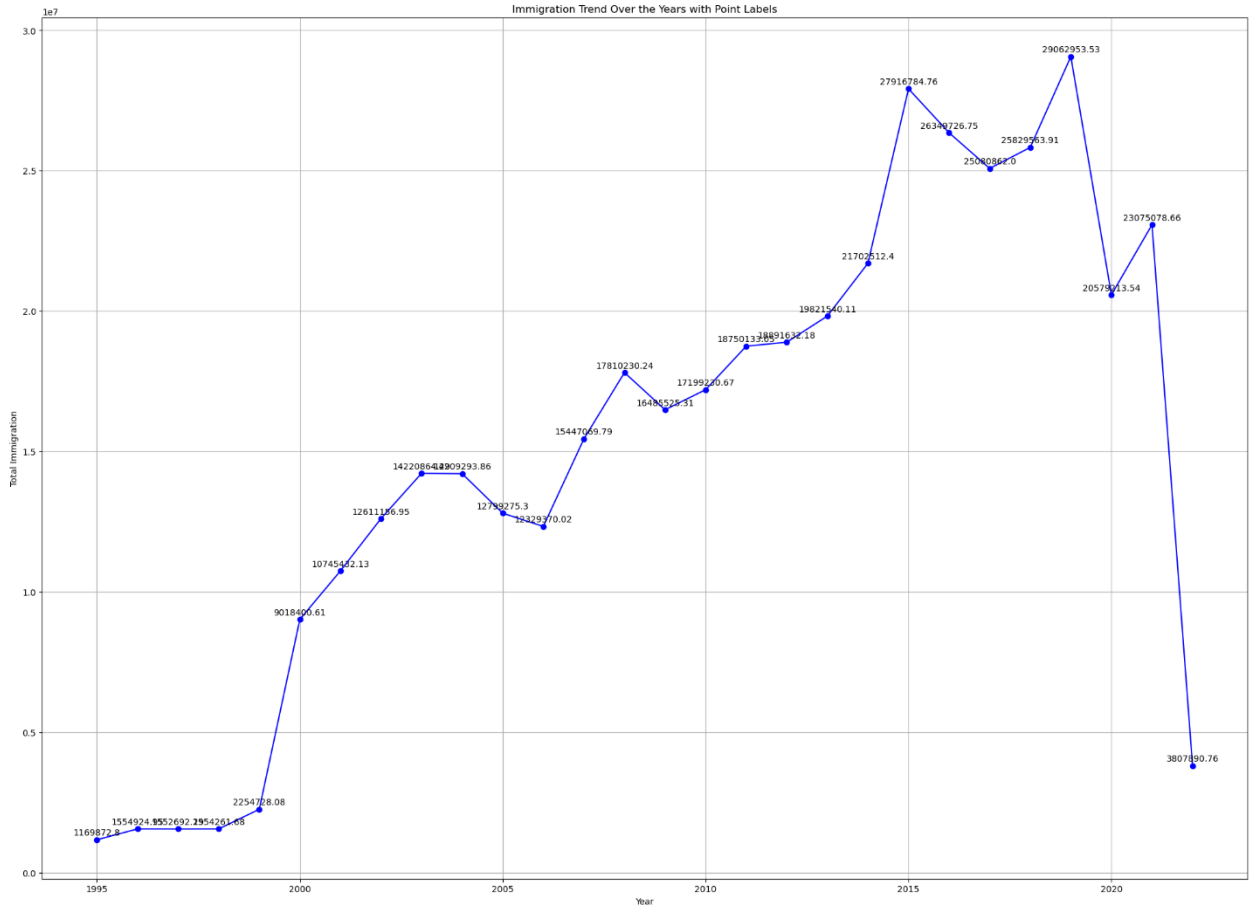


Top 5 countries that contributed the most to immigration

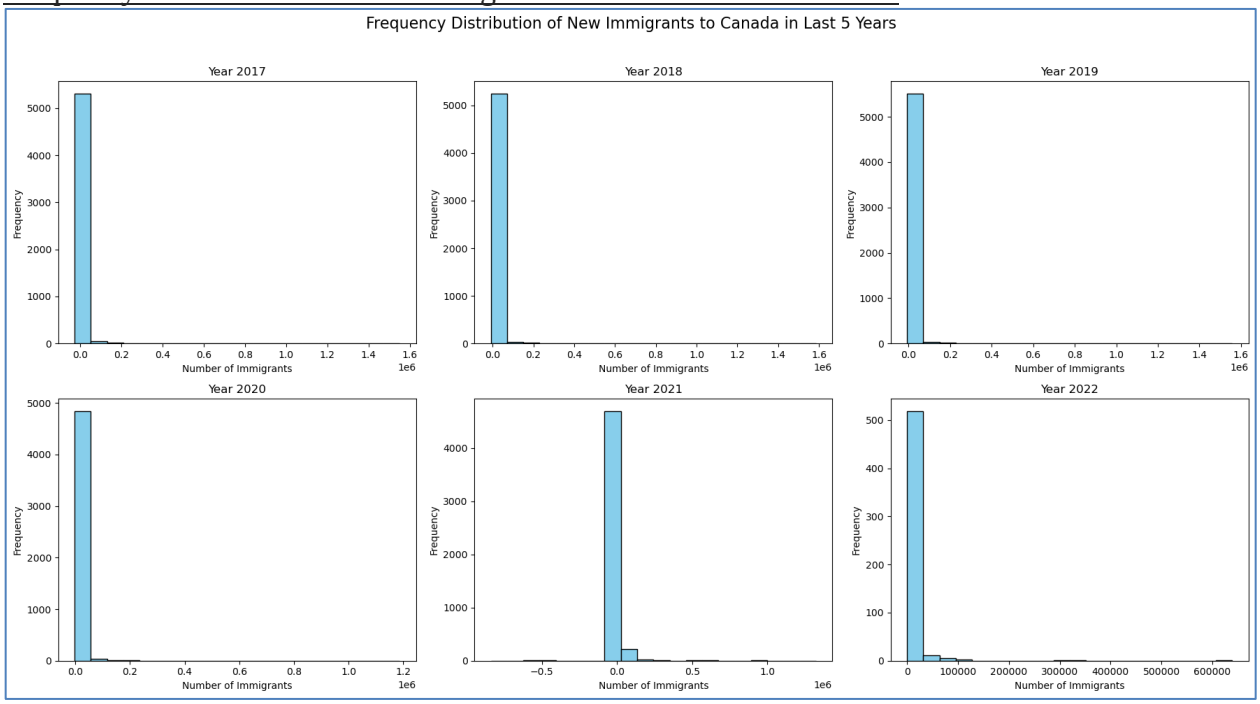


Immigration trends over years from 1995-2020

```
# Assuming 'TIME_PERIOD' column represents the years and 'OBS_VALUE' represents the immigration values
trend_data = df.groupby('TIME_PERIOD')['OBS_VALUE'].sum().reset_index()

# Plot the trend with point labels
plt.figure(figsize=(25, 18))
plt.plot(trend_data['TIME_PERIOD'], trend_data['OBS_VALUE'], marker='o', linestyle='-', color='b')

# Annotate points with their immigration values
for i, txt in enumerate(trend_data['OBS_VALUE']):
    plt.annotate(txt, (trend_data['TIME_PERIOD'].iloc[i], trend_data['OBS_VALUE'].iloc[i]), textcoords="offset points", xytext=(0,5), ha='center'

plt.title('Immigration Trend Over the Years with Point Labels')
plt.xlabel('Year')
plt.ylabel('Total Immigration')
plt.grid(True)
plt.show()
```



Frequency Distribution of New Immigrants to Canada in Last 5 Years



Mapping the real-world problem to an ML problem

As part of our project proposal update, we have conducted exploratory data analysis (EDA) on our dataset, which comprises five variables: Country of birth/nationality, Variable, Country of immigration, Immigration Year, and No of immigrants/Value. Among these variables, we aim to select one as the target variable (Y) and treat the rest as features (X) for our predictive modeling task.

To proceed with our analysis, we have converted the "No of immigrants/Value" variable into categorical bins. This transformation allows us to leverage decision tree algorithms for predicting the bin category of the "No of immigrants/Value" variable. We have discretized the numerical values into 10 bins using the percentile method, ensuring an equal distribution of data points across the bins.

However, during our project review, our coordinator suggested an alternative approach. They proposed incorporating an additional feature variable representing the census population of each country. By including this census population data as an additional feature, we can enhance our predictive model's accuracy and provide more meaningful insights. Specifically, the new feature can aid in predicting the target country by working on the sample data of the huge 460K+ records.

References

1. Activity Stream – Monthly updates – IRCC -

https://open.canada.ca/data/en/dataset/activity/f7e5498e-0ad8-4417-85c9-9b8aff9b9eda

2. Open Government portal – Canada – IRCC – https://open.canada.ca/data/dataset?license_id=ca-ogllgo&_keywords_limit=0&keywords=PR&frequency=P1M

3. Operational processing - Open Government – Monthly IRCC Updates –

https://open.canada.ca/data/en/dataset/9b34e712-513f-44e9-babf-9df4f7256550%204

4. Open Government > Organizations > Immigration, refugees, and citizenship Canada –

https://open.canada.ca/data/en/dataset/f7e5498e-0ad8-4417-85c9-9b8aff9b9eda

5. Immigration, refugees, and citizenship Canada > Corporate information > Publications and Manuals > operational instructions and guidelines > Updates –

https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/operational-bulletins-manuals/updates.html

Guy J. Abel and Nikola Sander (2014).

Quantifying Global International Migration Flows. Science, 343 (6178).

World Data Bank.