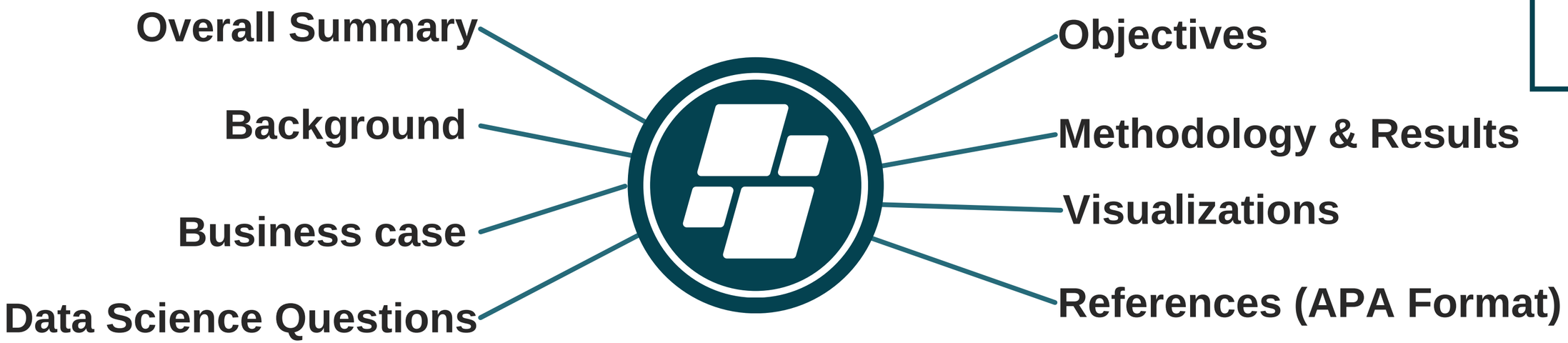
BDA TEST 2

Fadwa Ramadan Ali Hassan *2024334*

Table of Contents



Overall Summary

The report focus on utilizing big data analytics techniques using Apache Spark and related libraries. It involved the analysis of four datasets from the LinkedIn Digital Data dataset. The data underwent a comprehensive deep-cleaning process, including checking for null values, duplicated rows, invalid values, and outliers. Additionally, feature selection was applied to enhance model performance. The report aims Analysis of the Jobs Market in the Middle East & North African countries.

Background

Big data analytics involves analyzing large volumes of data to uncover trends, patterns, and correlations. steps such as data inspection, statistical analysis, data visualization, grouping techniques, and interpretation help extract meaningful information and make data-informed decisions.

The datasets introduced here are 4 data sets from the LinkedIn Digital Data dataset:

- public_use-industry-employment-growth
- public_use-talent-migration (Country)
- public_use-talent-migration (Industry)
- public_use-talent-migration(Skill)

These datasets from LinkedIn's Digital Data dataset provide valuable information on various aspects. The "public_use-industry-employment-growth" dataset focuses on employment growth in different industries. The "public_use-talent-migration (Country)" dataset examines talent migration trends across countries. The "public_use-talent-migration (Industry)" dataset analyzes talent migration within specific industries. Lastly, the "public_use-talent-migration (Skill)" dataset explores skill-related migration patterns.

Business case

Big data analytics uncovers trends and patterns in large datasets for data-driven decision-making. The data is based on LinkedIn's professional network of over 774 million members, by using The LinkedIn Digital Data datasets, I was able to perform an Analysis of the Jobs Market in the Middle East & North African countries. this analysis aims to delve into industry trends, identify patterns, and gain a deeper understanding of the region's workforce dynamics.

Data Science Questions

1. what are the countries classified in the Middle East & North Africa?
2. what are the most comment industries in the Middle East & North African countries?
3. Which countries in the Middle East & North Africa have the highest and lowest job growth rates through the years (2015-2019)?
4. Which skill categories have the most significant demand in the Middle East & North Africa region for each income category?
5. what are the skills that have the highest and lowest NetFlow through the years (2015-2019) in the Middle East & North African countries?
6. which countries in the Middle East & North African countries have the highest gain & loss of members each year?
7. what are the top countries that users emigrate from in the Middle East & North African countries?
8. predict the growth rate for each country in the Middle East & North African in 2020.

Objectives

1. Identify countries within the Middle East & North Africa region to establish the geographical scope of the analysis.
2. Determine the most common industries in the region.
3. Find countries with the highest and lowest job growth rates (2015-2019).
4. Analyze skill categories with high demand for each income level, aiding in understanding the specific skills desired after by employers in different income brackets.
5. Identify skills with the highest and lowest NetFlow (2015-2019), providing insights into talent migration patterns.
6. Determine countries with the highest gain and loss of members annually, offering insights into the dynamics of migration trends.
7. Identify the top countries of user emigration from the region, providing insights into the destinations preferred by users leaving the region.
8. Predict growth rates for each country in 2020.

Methodology & results

Methodology (Data cleaning)

- install spark and needed libraries.
- upload datasets into df2, df3, df4, df5.
- apply deep cleaning including the following :
  - check the null values.
  - check the duplicated rows.
  - check the invalid values.
  - check the outliers.
- apply feature selection to filter the most significant columns that will be needed, and also to remove the % char from growth_rate values.

```
        print("Null values in column '{}': {}".format(colum

    ##there r no null values in df1
```

```
Null values in column 'country_code': 0
Null values in column 'country_name': 0
Null values in column 'wb_region': 0
Null values in column 'wb_income': 0
Null values in column 'isic_section_index': 0
Null values in column 'isic_section_name': 0
Null values in column 'industry_id': 0
Null values in column 'industry_name': 0
Null values in column 'growth_rate_2015': 0
Null values in column 'growth_rate_2016': 0
Null values in column 'growth_rate_2017': 0
Null values in column 'growth_rate_2018': 0
Null values in column 'growth_rate_2019': 0
```

```
No duplicates found in df2
No duplicates found in df3
No duplicates found in df4
No duplicates found in df5
```

Methodology & Results

1- what are the countries classified in the Middle East & North Africa?
Those are all the Middle East & North African countries from the dataset public_use-industry-employment-growth.

2- what are the most comment industries in the Middle East & North African countries?
Financial Services | |Accounting | |Internet | |Oil & Energy as stated in the table.

industry_name
Financial Services
Accounting
Internet
Oil & Energy
Electrical & Electronic Manufacturing
Computer Software
Automotive
Telecommunications
Design
Information Technology & Services
Marketing & Advertising
Banking
Pharmaceuticals
Management Consulting

country_name
Iraq
Jordan
Algeria
Iran, Islamic Rep.
Qatar
Kuwait
Malta
Morocco
Israel
Oman
Egypt, Arab Rep.
Libya
Tunisia
Saudi Arabia
United Arab Emirates
West Bank and Gaza
Lebanon
Yemen, Rep.
Bahrain
Syrian Arab Republic

Methodology & Results

3- Which countries in the Middle East & North Africa have the highest job growth rates through the years (2015-2019)?

```
Country with the highest net job growth rate:
Row(country_name='Saudi Arabia', industry_name='Railroad Manufacture', avg_growth_rate=24.189999389648438)
Country with the lowest net job growth rate:
Row(country_name='Tunisia', industry_name='Semiconductors', avg_growth_rate=-19.070001220703126)
```

4- Which skill categories have the most significant demand in the Middle East & North Africa region for each income category? (for different income types)

skill_group_name	count
Oil & Gas	6
Public Safety	6
International Law	6
Cybersecurity	6
Manufacturing Ope...	6

skill_group_name	count
International Law	4
Manufacturing Ope...	4
Cybersecurity	4
Digital Marketing	4
Corporate Communi...	4

skill_group_name	count
International Law	2
Teaching	2
Manufacturing Ope...	2
Public Safety	2
Oil & Gas	2

skill_group_name	count
Travel Management	8
Administrative As...	8
Information Manag...	8
Investment Banking	8
Entrepreneurship	8

Methodology & Results

5- what are the skills that have the highest and lowest NetFlow through the years (2015-2019) in the Middle East & North African countries?

```
# Sort by average net_per_10K in descending order
sorted_netflow_df = mena_df5 .orderBy(desc("avg_net_per_10K"))

# Select the skill group with the highest and lowest average net_per_10K
highest_netflow = sorted_netflow_df.first()
lowest_netflow = sorted_netflow_df.orderBy("avg_net_per_10K").first()

# Display the results
print("Skill with the highest NetFlow:")
print(highest_netflow["skill_group_name"], highest_netflow["avg_net_per_10K"], "in Middle East & North Africa")

print("Skill with the lowest NetFlow:")
print(lowest_netflow["skill_group_name"], lowest_netflow["avg_net_per_10K"], "in Middle East & North Africa")
```

```
Skill with the highest NetFlow:
Artificial Intelligence (AI) 970.2719999999999 in Middle East & North Africa
Skill with the lowest NetFlow:
Mobile Application Development -1401.6560000000002 in Middle East & North Africa
```

Methodology & Results

6- which countries in the Middle East & North African countries have the highest gain & loss of members each year?

```
For net_per_10K_2015:
Country with the highest net_per_10K gain:
United Arab Emirates Middle East & North Africa 2280.6

For net_per_10K_2016:
Country with the highest net_per_10K gain:
United Arab Emirates Middle East & North Africa 1616.52

For net_per_10K_2017:
Country with the highest net_per_10K gain:
United Arab Emirates Middle East & North Africa 1362.36

For net_per_10K_2018:
Country with the highest net_per_10K gain:
United Arab Emirates Middle East & North Africa 1187.61

For net_per_10K_2019:
Country with the highest net_per_10K gain:
Qatar Middle East & North Africa 955.84
```

```
For net_per_10K_2015:
Country with the lowest net_per_10K gain:
Bahrain Middle East & North Africa -998.98

For net_per_10K_2016:
Country with the lowest net_per_10K gain:
Kuwait Middle East & North Africa -1876.78

For net_per_10K_2017:
Country with the lowest net_per_10K gain:
Kuwait Middle East & North Africa -1495.33

For net_per_10K_2018:
Country with the lowest net_per_10K gain:
Bahrain Middle East & North Africa -1358.05

For net_per_10K_2019:
Country with the lowest net_per_10K gain:
Bahrain Middle East & North Africa -1111.57
```

Methodology & Results

7- what are the top countries that users emigrate from in the Middle East & North African countries?

base_country_wb_income	base_country_name	count
High Income	United Arab Emirates	93
Low Income	Syrian Arab Republic	14
Lower Middle Income	Egypt, Arab Rep.	42
Upper Middle Income	Iran, Islamic Rep.	29

Methodology & Results

8- predict the growth rate for each country in the Middle East & North African in 2020.

```
predict the growth rate for each Middle East & North African country in 2020? user mena_countries_df2 and use the columns
growth_rate_2015|growth_rate_2016|growth_rate_2017|growth_rate_2018|growth_rate_2019| avg_growth_rate to predict growth_rate_2020

tain be 80% and testing 20%
```

it appears that all the predictions are currently set to 0.0. This could indicate that the model is not able to capture meaningful relationships between the selected features and the target variable.

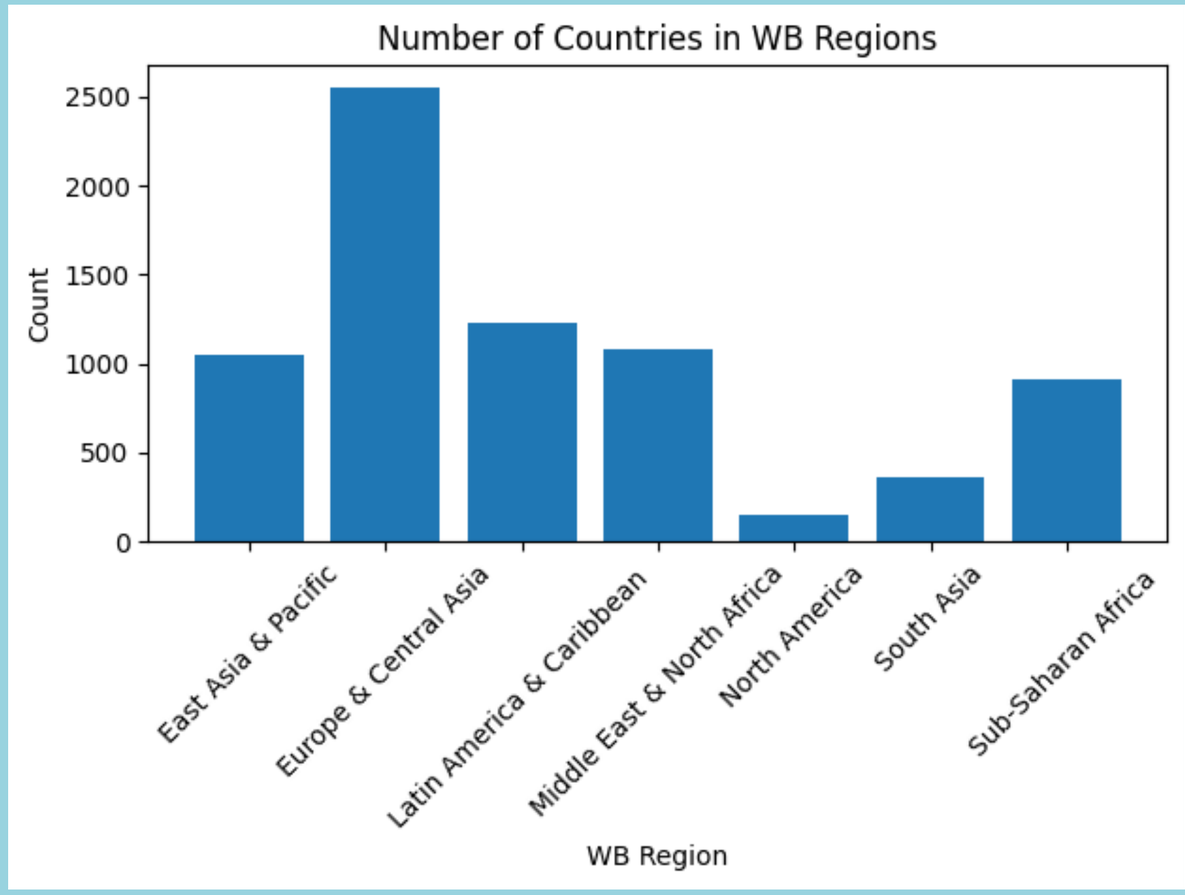
```
+--------------------+---------------+----------+
|            features|growth_rate_2020|prediction|
+--------------------+---------------+----------+
|           (2,[],[])|            0.0|       0.0|
|           (2,[],[])|            0.0|       0.0|
|[-17.040000915527...|            0.0|       0.0|
|[-15.479999542236...|            0.0|       0.0|
|[-12.0,-5.7100000...|            0.0|       0.0|
|[-11.829999923706...|            0.0|       0.0|
|[-11.760000228881...|            0.0|       0.0|
|[-10.260000228881...|            0.0|       0.0|
|[-9.9600000381469...|            0.0|       0.0|
|[-9.5200004577636...|            0.0|       0.0|
```

Visualizations

Visualizations

bar chart for the number of countries in each World Bank (WB) region.

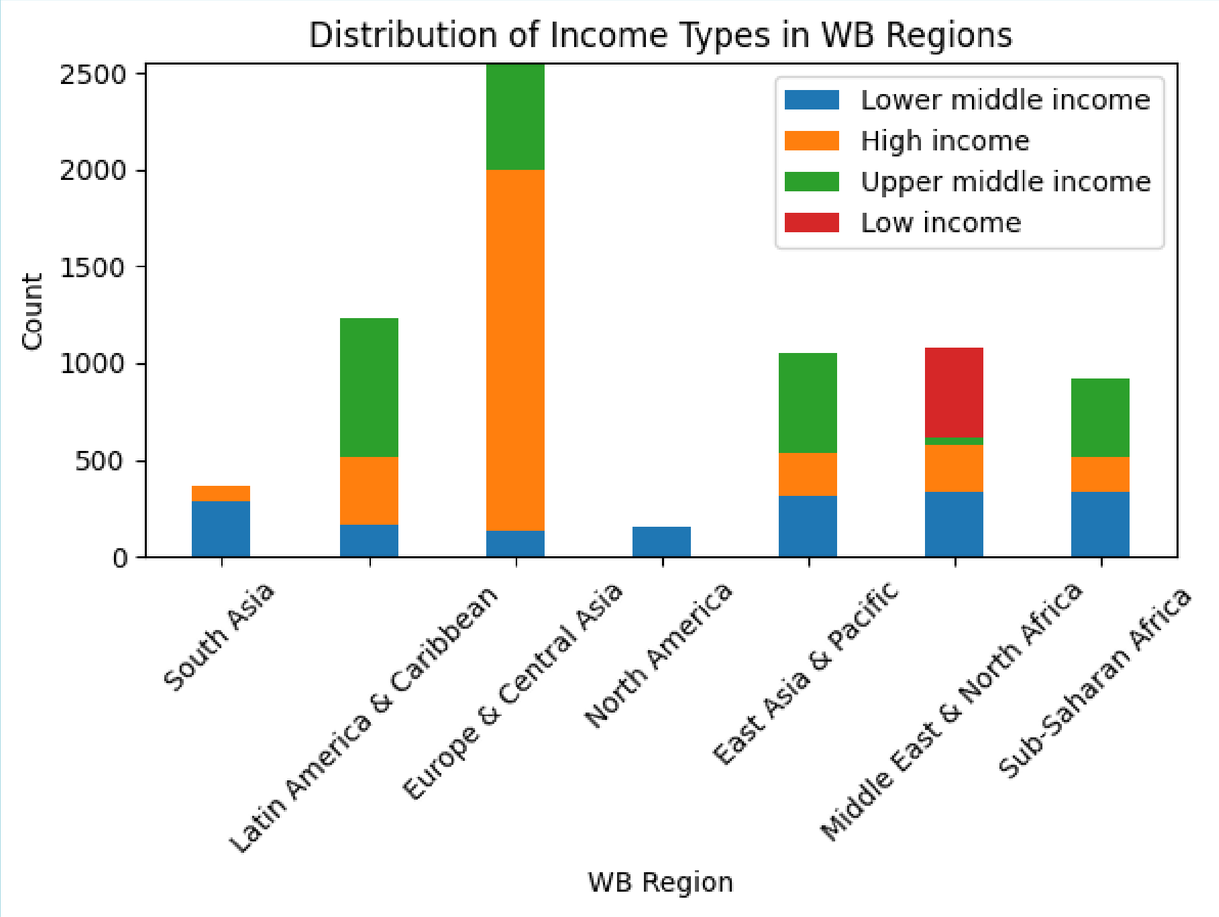
Europe & Central Asia region have the highest number of countries, on the other hand, North America has the lowest number of countries and that is logical considering that there are 2 countries there Canada and the United States



Visualizations

Distribution of Income Types in WB Regions

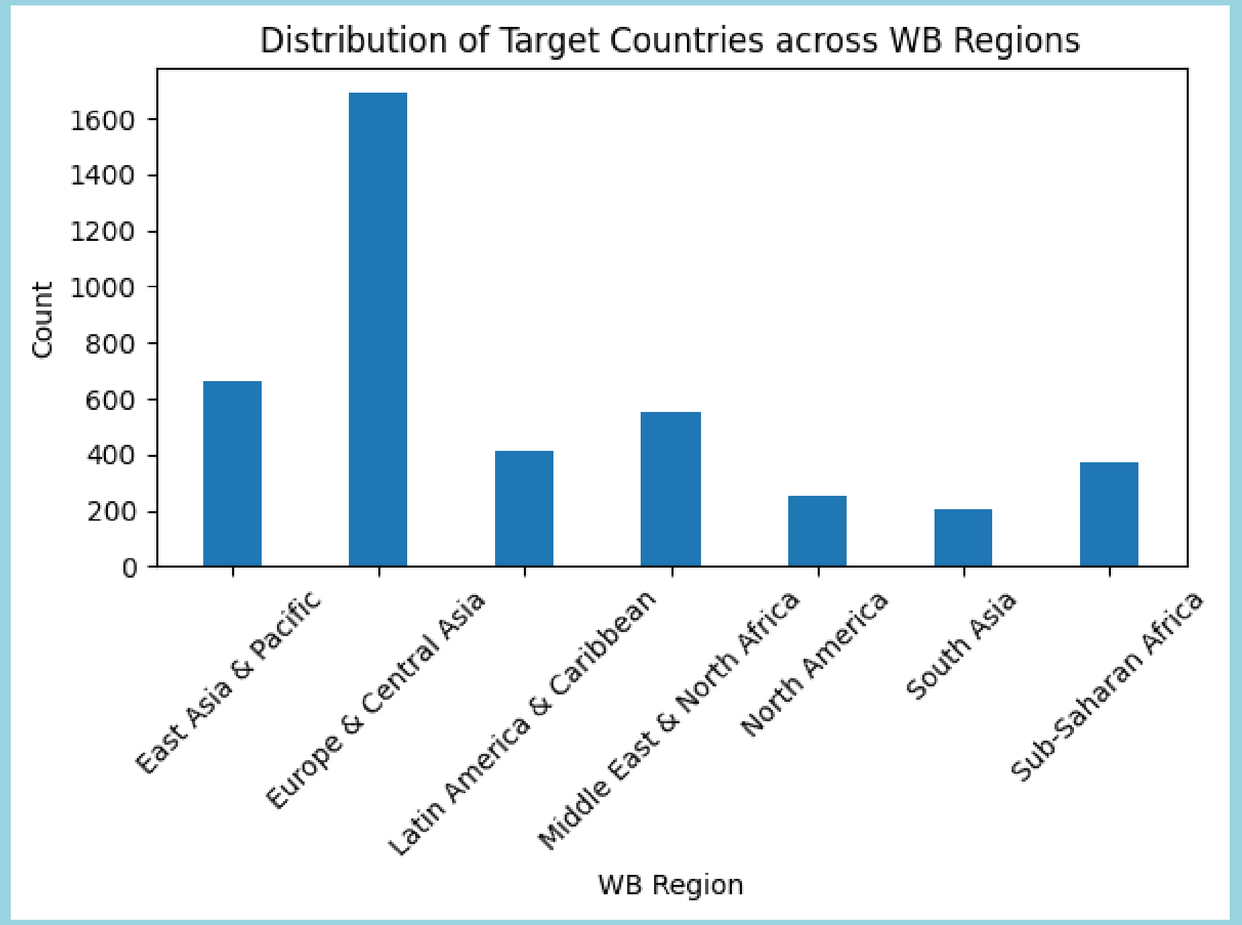
Europe & Central Asia region has the highest Distribution for high income, while Middle East & North Africa have the lowest income distribution.



Visualizations

Distribution of Target Countries across WB Regions

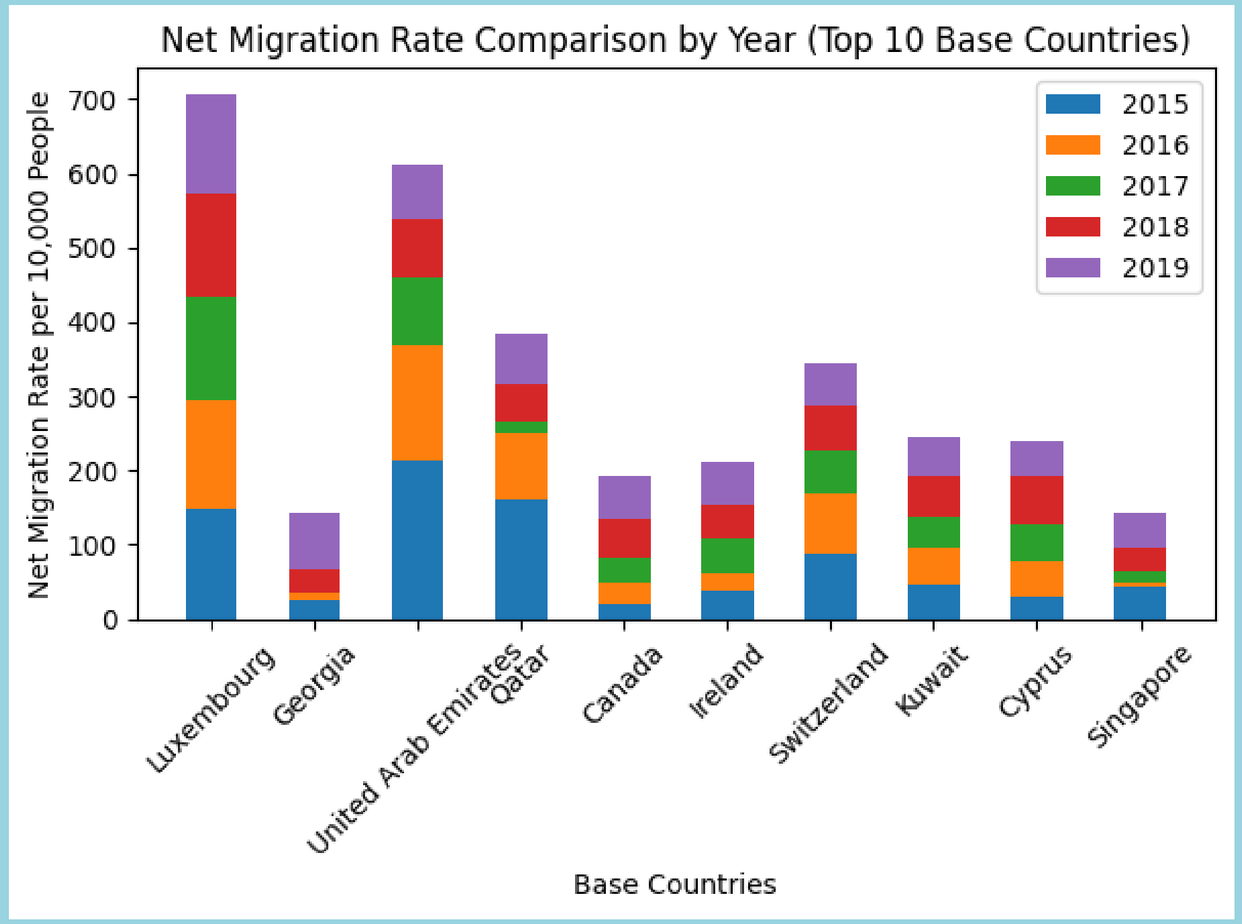
the most targeted region where users usually emigrate to are countries in Europe & Central Asia region, however, South Asia is the least region where used immigrate to.



Visualizations

Net Migration Rate Comparison by Year (Top 10 Base Countries)

UAE and Qatar are 2 of the countries that have a high net migration rate among all countries throughout the year from 2015 - 2019.



References (APA Format)

- sites that helped in analysing the datasets:
- Chai, W., & Labbe, M. (2021, December). What is Big Data Analytics and Why is it Important? (C. Stedman, Ed.). SearchBusinessAnalytics. https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics
- Google Colaboratory. (n.d.). Colab.research.google.com. Retrieved June 16, 2023, from https://colab.research.google.com/github/datacommonsorg/api-python/blob/master/notebooks/intro_data_science/Regression_Basics_and_Prediction.ipynb

Thank you!

Fadwa Ramadan Ali Hassan *2024334*