

Introduction

This project explores trends in international student demographics, aiming to determine if there has been a shift in preference for studying in Canada instead of the USA among students from various countries. By analyzing data on international students in Canada and comparing it with broader international student demographics. This analysis will involve constructing ETL pipelines to clean, transform, and analyze the data.

Main Question

Do students from certain countries prefer Canada over the USA as a study destination in recent years?

Data sources

Datasource 1: International Students in Canada

- **Data URL:** [International Students in Canada Dataset](#)
- **Data Type:** CSV
- **License:** CC0: Public Domain
- **Author Name:** IRCC
- **Origin:** IRCC,
<https://open.canada.ca/data/en/dataset/90115b00-f9b8-49e8-afa3-b4cff8facaee>
- **Description:** This dataset provides information on international students in Canada, including enrollment numbers by country and year. It is useful for analyzing trends in international education within Canada.

	origin	2015	2016	2017	2018	2019	2020	2021	2022	2023
0	Afghanistan	95	115	95	80	95	90	80	170	140
1	Albania	115	165	185	245	375	250	305	345	545
2	Algeria	1060	845	1020	1490	2690	2170	3165	5360	7180
3	Andorra	0	0	0	0	0	0	10	5	0
4	Angola	65	80	40	25	120	30	50	75	65

Datasource 2: International Student Demographics

- **Data URL:** [International Student Demographics Dataset](#)
- **Data Type:** CSV
- **Author:** Syed Abdul Shameer
- **License:** CC0: Public Domain
- **Origin:** OpenDoorsData.org

- **Description:** This dataset contains demographic details of international students globally. It includes data on students studying in the USA and other popular destinations, enabling comparative analysis across countries.

	year	origin_region	origin	academic_type	students
0	2000/01	Africa, Sub-Saharan	Africa, Sub-Saharan, Unspecified	Graduate	2
1	2000/01	Africa, Sub-Saharan	Africa, Sub-Saharan, Unspecified	Other	0
2	2000/01	Africa, Sub-Saharan	Africa, Sub-Saharan, Unspecified	Undergraduate	6
3	2000/01	Asia	Asia, Unspecified	Graduate	0
4	2000/01	Asia	Asia, Unspecified	Other	6

Licenses & Obligations

Both datasets are CC0 which means you can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

Data Pipeline

1. Download Dataset:

Using the Kaggle API and pandas, we download the datasets and save them as CSV files in the data directory. This process automates the retrieval of datasets from Kaggle and ensures that the data is ready for further processing.

2. Transformation & Cleaning:

Canada Dataset: The Canada dataset is almost in the desired format; only minor adjustments are required. We will:

- Rename columns to unify them with US table.
- Check data types (dtypes) to ensure consistency.

USA Dataset: The USA dataset presents a challenge because the number of students per year is not available. The data is split in a way that requires grouping and aggregation to create a table format similar to the Canada dataset. We will:

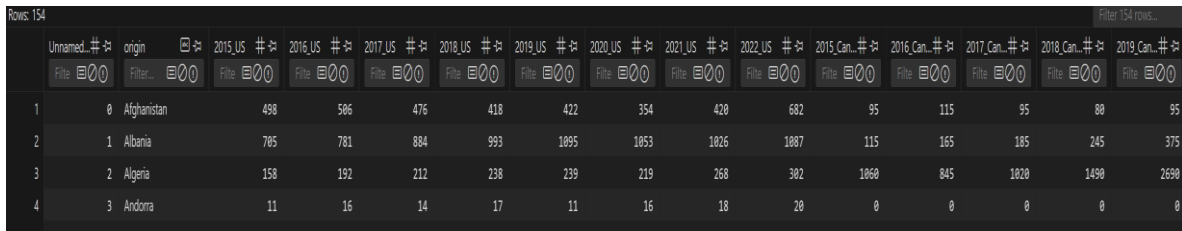
- Group the data by country and aggregate the student numbers per year.
- Check for any NaN and Null values, dropping or filling them with 0 as necessary.

3. Merging:

Once both datasets are cleaned and transformed, we merge them based on the students' country of origin. To avoid column name conflicts, we use prefixes to differentiate between the Canadian and USA international students. After merging the datasets, we save the combined table as a CSV file for further inspection and use.

4. Output:

After merging, we load the resulting table and convert it into an SQLite3 database table for Future analysis. The result is 157 countries' data from 2015-2022 in sqlite3 format.



	Unnamed: #	origin	2015_US #	2016_US #	2017_US #	2018_US #	2019_US #	2020_US #	2021_US #	2022_US #	2015_Can... #	2016_Can... #	2017_Can... #	2018_Can... #	2019_Can... #
1	0	Afghanistan	498	506	476	418	422	354	420	682	95	115	95	80	95
2	1	Albania	785	781	884	993	1095	1053	1026	1007	115	165	185	245	375
3	2	Algeria	158	192	212	238	239	219	268	302	1060	845	1020	1490	2690
4	3	Andorra	11	16	14	17	11	16	18	20	0	0	0	0	0

Dataset Future changes

If new countries or data for new year will be added to the dataset this pipeline can automatically takes this changes into account and wont break.

Quality

Both datasets are from government statistics so they are reliable, but they could be not complete in terms of years and countries.

Limitation

Removing rows with null values could introduce bias and data loss. The data is only available until 2022 which makes it hard to generalize the results to 2024.