Northeastern University
Khoury College of Computer and Information Sciences

Intro to Programming for DS DS5200

Spring 2025

Project Report

Immigration Data Integration & Analysis

Layashree Adepu

📊 Immigration Data Integration & Analysis

1. Introduction

In an effort to understand global migration trends through historical data, this project aims to merge and analyse a large collection of immigration statistics from multiple countries spanning several decades. These datasets, published as Excel files by a government agency, contain detailed demographic breakdowns but are riddled with inconsistencies, naming irregularities, and non-uniform formatting. This required a combination of automation, cleaning, and transformation techniques to turn a fragmented archive into an integrated, insightful dataset.

The final goal was to build a high-quality dataset that could support demographic trend analysis across dimensions like Age, Occupation, and Admission Category, using tools from Python's data science ecosystem.

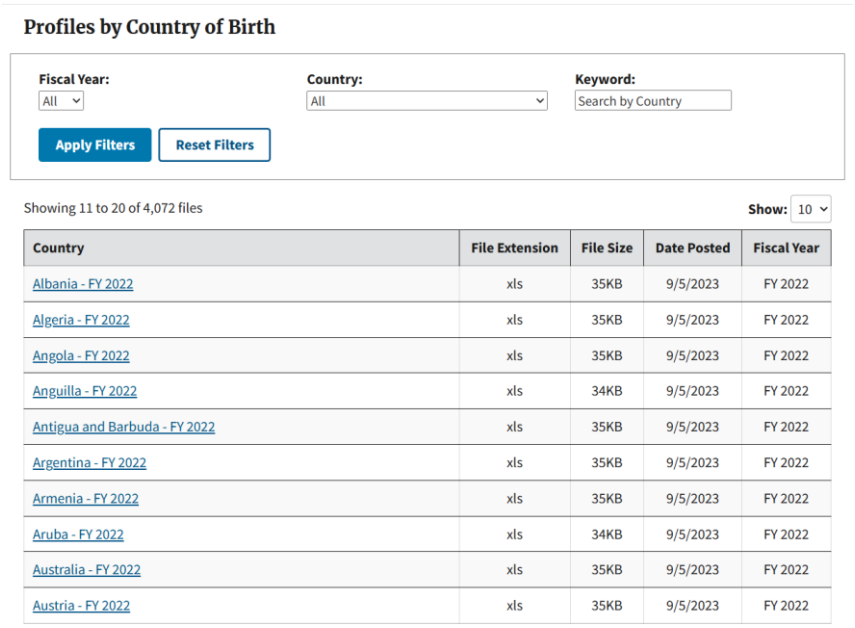
---

2. Data Analysis

Part 1: Acquisition via Web Scraping

Getting our hands on all the immigration data was quite a challenge. The information wasn't available as a simple download - it was scattered across more than 4,000 Excel files buried in a government website that wasn't designed for bulk downloading.

Rather than clicking through thousands of pages manually (which would have taken weeks and probably caused some serious hand cramps!), I wrote a script using Selenium to do the heavy lifting for me. Think of it as teaching the computer to browse the website like a human would, but much faster.



Profiles by Country of Birth

Fiscal Year: All | Country: All | Keyword: Search by Country

Apply Filters | Reset Filters

Showing 11 to 20 of 4,072 files — Show: 10

Country	File Extension	File Size	Date Posted	Fiscal Year
Albania - FY 2022	xls	35KB	9/5/2023	FY 2022
Algeria - FY 2022	xls	35KB	9/5/2023	FY 2022
Angola - FY 2022	xls	35KB	9/5/2023	FY 2022
Anguilla - FY 2022	xls	34KB	9/5/2023	FY 2022
Antigua and Barbuda - FY 2022	xls	35KB	9/5/2023	FY 2022
Argentina - FY 2022	xls	35KB	9/5/2023	FY 2022
Armenia - FY 2022	xls	35KB	9/5/2023	FY 2022
Aruba - FY 2022	xls	34KB	9/5/2023	FY 2022
Australia - FY 2022	xls	35KB	9/5/2023	FY 2022
Austria - FY 2022	xls	35KB	9/5/2023	FY 2022

The script navigated to the Homeland Security website, clicked the dropdown menu to show all entries at once, and then grabbed all the download links for the Excel files we needed:

I added some clever bits to handle duplicate filenames and keep track of what was downloaded. The script even checked if each download was successful and logged any problems for me to fix later.

By the end, we had a complete collection of country-of-birth profiles covering 18 years and 208 countries - a treasure trove of data that became the foundation for all our analysis and visualisations.

Part 2: Merging of files

Cleaning and merging the data turned out to be one of the most challenging parts of the project. After downloading thousands of Excel files, I needed to combine them into a single, clean dataset that I could analyze.

The biggest challenge was handling inconsistencies across files. Many countries appeared under different names over the years. For example, "Swaziland" officially changed its name to "Eswatini" mid-way through our study period. "North Macedonia" was previously listed as just "Macedonia". To solve this, I created a comprehensive mapping dictionary:

```
def standardize_region_name(region):
    """
    Standardize region names to handle inconsistencies across different years
    """
    region_mapping = {
    # Variations for countries with "The" in their name
    "Bahamas": "Bahamas",
    "Bahamas, The": "Bahamas",
    "Gambia": "Gambia",
    "Gambia, The": "Gambia",

    # Countries that changed names
    "Czechia": "Czech Republic",
    "Czech Republic": "Czech Republic",
    "Czechoslovakia (former)": "Czech Republic",

    "Eswatini": "Eswatini",
    "Eswatini (formerly Swaziland)": "Eswatini",
    "Swaziland": "Eswatini",
```

I also had to handle special data codes in the files. The code 'D' meant data that was suppressed to protect privacy, while '-' represented zero values. Instead of just removing entries with 'D', I replaced them with the mean values for that specific occupation category and year:

```
    target_mask = filtered_df["Subgroup"].isin(target_subgroups)

    subgroup_means = filtered_df[target_mask].groupby(
        ["Year", "Subgroup"]
    )[col].transform("mean")

    # Fill only NaNs in those subgroups with the precomputed means
    filtered_df.loc[target_mask & filtered_df[col].isna(), col] = subgroup_means
```

The structure of the Excel files was inconsistent too. Some files had different section headers or extra footer notes that needed to be removed. I had to identify where each data section began and ended: carefully

```
# Explicitly filter to keep only the sections we want
filtered_df = merged_df[merged_df["Group"].isin(["Age", "Occupation", "Broad Class of Admission"])]
```

I also needed to exclude continental totals and certain categories that would skew our analysis:

```
    regions_to_exclude = [
        "Total",
        "frica",
        "Asia",
        "Caribbean",
        "Central America",
        "Europe",
        "North America (Includes Caribbean and Central America)",
        "Oceania",
        "South America"
    ]
```

Duplicates were another check I had to do to make sure no data is being duplicated

```
# Remove duplicates by keeping the first occurrence
filtered_df = filtered_df.drop_duplicates(subset=['Year', 'Region', 'Group', 'Subgroup'])
print(f"Removed {len(duplicates)} duplicate entries. Final dataset has {len(filtered_df)} rows.")
```

After all this cleaning, I finally had a usable dataset containing over 81,000 rows spanning 18 years and 177 countries. The data was organized into three main categories: Age demographics, Occupation types, and Broad Class of Admission.

We had a file that looked like this ⬇

Persons Obtaining Legal Permanent Resident Status During Fiscal Year 2013
by Region/Country of Birth and Selected Characteristics

Region/Country: Total

Characteristic	Total	Male	Female	Unknown
Total	9,90,553	4,34,284	5,13,736	42,533
New arrivals	4,59,751	1,89,366	2,27,867	42,518
Adjustments of status	5,30,802	2,44,918	2,85,869	15
Age				
Under 18 years	1,80,247	90,053	86,925	3,269
18 to 24 years	1,22,587	55,052	65,934	1,601
25 to 34 years	2,34,690	1,03,640	1,29,495	1,555
35 to 44 years	1,86,102	88,369	96,635	1,098
45 to 54 years	1,13,819	48,775	64,153	891
55 to 64 years	71,724	28,680	42,193	851
65 years and over	48,875	19,707	28,396	772
Unknown	32,509	8	5	32,496
Marital status				
Single	3,55,199	1,77,516	1,63,544	14,139
Married	5,79,295	2,44,100	3,09,961	25,234
Other	51,671	10,486	38,030	3,155
Unknown	4,388	2,182	2,201	5
Occupation				
Management, professional, and related occupations	1,17,974	71,952	41,529	4,493
Service occupations	46,841	25,137	19,091	2,613
Sales and office occupations	29,767	12,301	15,148	2,318
Farming, fishing, and forestry occupations	11,589	8,672	2,071	846
Construction, extraction, maintenance and repair occupations	6,538	6,321	135	82
Production, transportation, and material moving occupations	40,153	29,563	9,140	1,450
Military	46	38	8	-
No occupation/not working outside home	4,71,041	1,62,159	2,83,733	25,149
Homemakers	1,30,162	4,735	1,14,407	11,020
Students or children	2,49,910	1,20,224	1,19,489	10,197
Retirees	9,330	3,980	5,277	73
Unemployed	81,639	33,220	44,560	3,859
Unknown	2,66,604	1,18,141	1,42,881	5,582
Broad class of admission				
Family-sponsored preferences	2,10,303	92,857	1,01,165	16,281
Employment-based preferences	1,61,110	81,638	78,003	1,469
Immediate relatives of U.S. citizens	4,39,460	1,67,980	2,51,004	20,476
Diversity	45,618	23,099	18,220	4,299
Refugees and asylees	1,19,630	61,693	57,937	-
Other	14,432	7,017	7,407	8
Leading states of residence				
Arizona	16,097	6,798	8,665	634
California	1,91,806	81,630	1,01,304	8,872
Colorado	11,108	4,928	5,742	438
Connecticut	10,985	4,827	5,627	531
Florida	1,02,939	45,289	54,518	3,132
Georgia	24,387	11,030	12,545	812
Illinois	35,988	15,953	18,268	1,767
Maryland	25,361	11,158	13,103	1,100
Massachusetts	29,482	13,161	14,932	1,389

Michigan	16,952	7,682	8,483	787
Minnesota	12,781	5,711	6,616	454
Nevada	9,886	4,037	5,345	504
New Jersey	53,082	23,333	26,883	2,866
New York	1,33,601	59,727	67,436	6,438
North Carolina	16,798	7,424	8,740	634
Ohio	13,819	6,285	7,022	512
Pennsylvania	24,720	11,041	12,538	1,141
Texas	92,674	40,313	48,738	3,623
Virginia	27,861	12,179	14,561	1,121
Washington	22,994	9,995	12,114	885
Other	1,17,232	51,783	60,556	4,893

D Data withheld to limit disclosure.

- Represents zero.

Source: U.S. Department of Homeland Security

From which we converted it to this ⬇

Total	Male	Female	Unknown	Year	Region	Group	Subgroup
4321	2193	2128	0	2014	Jamaica	Age	Under 18 years
2638	1324	1314	0	2014	Jamaica	Age	18 to 24 years
3517	1748	1769	0	2014	Jamaica	Age	25 to 34 years
3507	1529	1978	0	2014	Jamaica	Age	35 to 44 years
2887	1053	1834	0	2014	Jamaica	Age	45 to 54 years
1477	514	963	0	2014	Jamaica	Age	55 to 64 years
679	232	447	0	2014	Jamaica	Age	65 years and over
0	0	0	0	2014	Jamaica	Age	Unknown
1157	445	712	0	2014	Jamaica	Occupation	Management, professional, and related occupations
2136	985	1151	0	2014	Jamaica	Occupation	Service occupations
440	80.68571429	102.4680851	0	2014	Jamaica	Occupation	Sales and office occupations
297	210	87	0	2014	Jamaica	Occupation	Farming, fishing, and forestry occupations
328	319	9	0	2014	Jamaica	Occupation	Construction, extraction, maintenance and repair occupations
296	254	42	0	2014	Jamaica	Occupation	Production, transportation, and material moving occupations
4	0.075581395	0	0	2014	Jamaica	Occupation	Military
8401	3675	4726	0	2014	Jamaica	Occupation	No occupation/not working outside home
675	23	652	0	2014	Jamaica	Occupation	Homemakers
5744	2847	2897	0	2014	Jamaica	Occupation	Students or children
100	40	60	0	2014	Jamaica	Occupation	Retirees
1882	765	1117	0	2014	Jamaica	Occupation	Unemployed
5967	2566	3401	0	2014	Jamaica	Occupation	Unknown
6379	3078	3301	0	2014	Jamaica	Broad Class of Admission	Family-sponsored preferences
629	260	369	0	2014	Jamaica	Broad Class of Admission	Employment-based preferences
11917	5204	6713	0	2014	Jamaica	Broad Class of Admission	Immediate relatives of U.S. citizens
4	351.2848788	434.2058045	0	2014	Jamaica	Broad Class of Admission	Diversity
44	351.2848788	434.2058045	0	2014	Jamaica	Broad Class of Admission	Refugees and asylees

Part 3: Initial Data Exploration and Cleaning

After assembling our master dataset from thousands of files, I needed to carefully examine it to understand what we were working with. This step was crucial for ensuring our analysis would be both valid and meaningful.

First, I loaded the merged dataset and did a basic inspection:

```
df = pd.read_excel("C:/Users/layas/OneDrive/Desktop/Layashree documents/NEU Cources/Intro to Programming in DS/output merged file/all_countries_merged.xlsx")
df.info()
df.isnull().sum()
df.head(10)
```

✓ 7.9s

I discovered right away that the 'Occupation' category classifications had changed significantly before 2006, which would make trend analysis difficult. To maintain consistency, I filtered out the pre-2006 data:

```
occupation_years = df[df['Group'] == 'Occupation'].groupby('Subgroup')['Year'].unique()
print(occupation_years)
```

✓ 0.1s

```
Subgroup
Administrative support                                                    [2005]
Construction, extraction, maintenance and repair occupations    [2014, 2013, 2012, 2011, 2010, 2009, 2008, 200...
Executive and managerial                                                  [2005]
Farming, fishing, and forestry occupations                       [2014, 2013, 2012, 2011, 2010, 2009, 2008, 200...
Farming, forestry, fisheries                                              [2005]
Homemakers                                                       [2014, 2013, 2012, 2011, 2010, 2009, 2008, 200...
Management, professional, and related occupations                [2014, 2013, 2012, 2011, 2010, 2009, 2008, 200...
Military                                                         [2014, 2013, 2012, 2011, 2010, 2009, 2008, 200...
No occupation                                                             [2005]
```

Next came one of the most important steps - checking for continuity in our dataset. For trend analysis to be reliable, we needed countries with data across the entire time period.

```
country_year_counts = Counter()
for year_set in countries_by_year.values():
    for country in year_set:
        country_year_counts[country] += 1
```

I created a visualization to identify gaps The heatmap revealed a patchwork of missing data. Some countries appeared sporadically throughout the years, while others had consistent records. I decided to focus on countries that were almost complete - those missing at most 2 years of data:

This filtering left us with 177 countries that had consistent data across the 18-year period - a solid foundation for our trend analysis. I created a nicely formatted grid display to show these countries:

```
Number of countries present in all years or missing at most two years: 177
Selected countries:
Afghanistan          Bosnia and Herzegovina     Cuba                 Guatemala      Laos          Nethe
Albania              Botswana                   Cyprus               Guinea         Latvia        New Z
Algeria              Brazil                     Czech Republic       Guinea-Bissau  Lebanon       Nicar
Angola               British Virgin Islands     Denmark              Guyana         Liberia       Niger
Anguilla             Bulgaria                   Djibouti             Haiti          Libya         Niger
Antigua and Barbuda  Burkina Faso               Dominica             Honduras       Lithuania     North
Argentina            Burma                      Dominican Republic   Hong Kong      Luxembourg    Norwa
Armenia              Burundi                    Ecuador              Hungary        Macau         Oman
Aruba                Cabo Verde                 Egypt                Iceland        Madagascar    Pakis
Australia            Cambodia                   El Salvador          India          Malawi        Panam
```

Finally, I got the dataset that I would be working with for exploring the data

Save Filtered Dataset

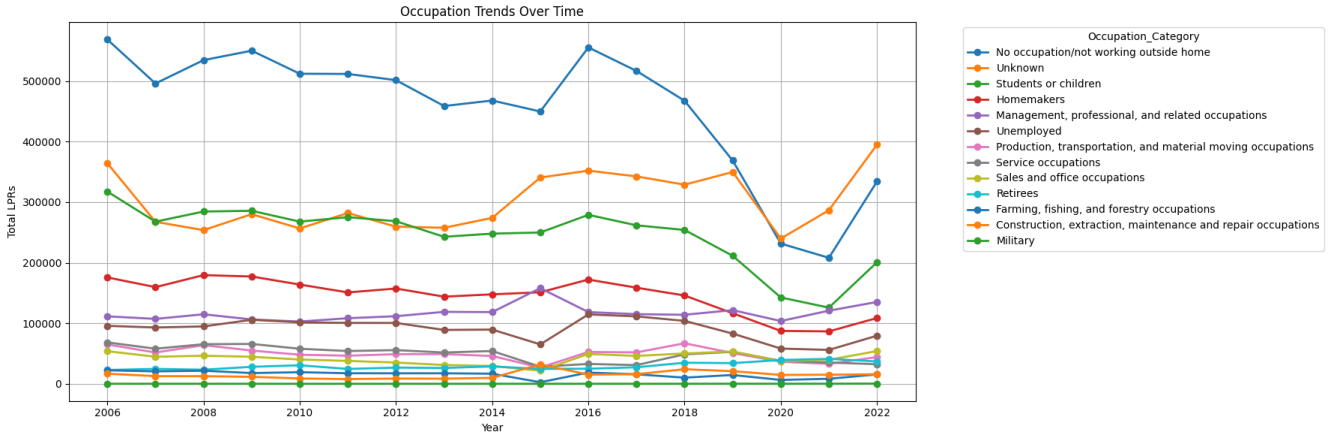
Save this new filtered dataset (with consistent countries only) for trend forecasting.

```
output_path = "C:/Users/layas/OneDrive/Desktop/Layashree documents/NEU Cources/Intro to Programming in DS/output merged file/common_countries_dataset.xlsx"
df_common.to_excel(output_path, index=False)
print(f"Filtered dataset saved to {output_path}")
```

Part 4: Occupational Trend Analysis

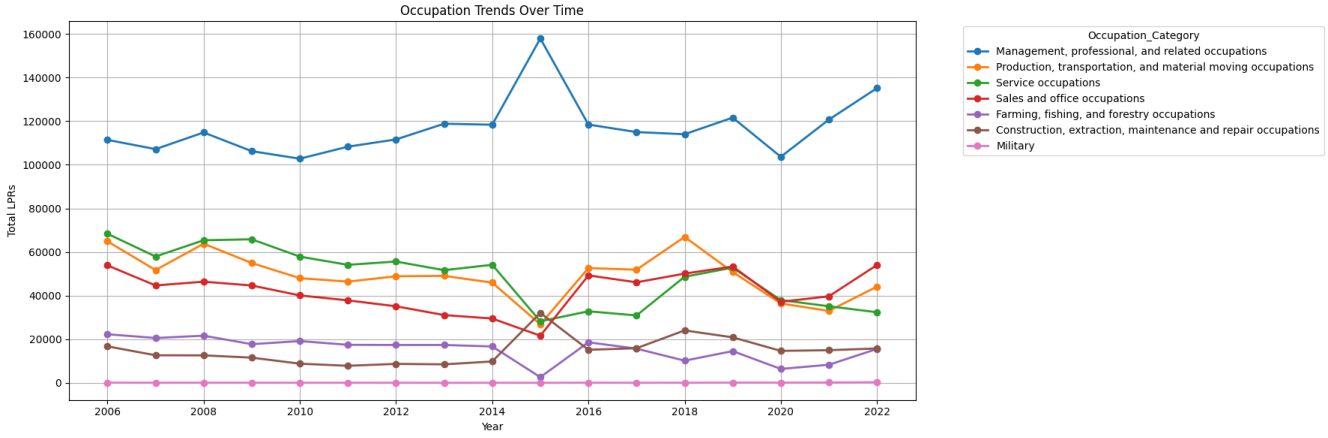
After cleaning and filtering our dataset to include only countries with consistent data, I turned my attention to analyzing occupation trends - the heart of our project.

I started by isolating just the occupation-related data from our cleaned dataset:

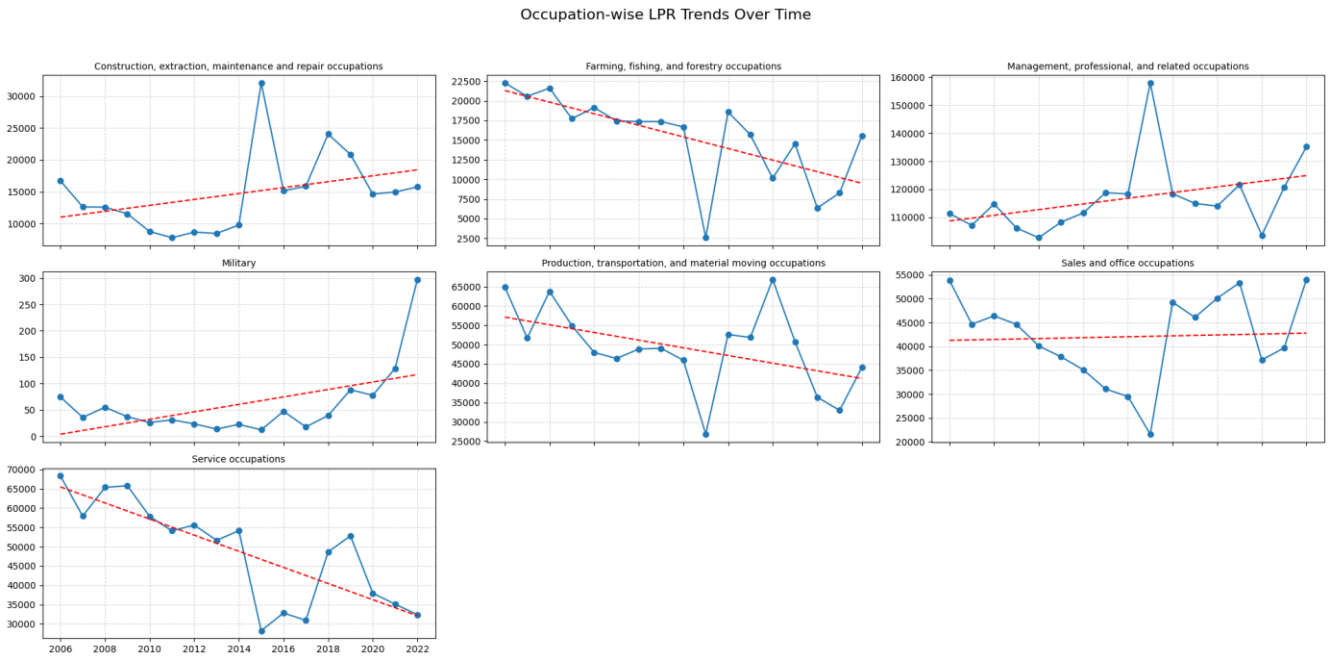


My first visualization showed all occupation categories, but I immediately noticed a problem - many categories weren't actual occupations. Categories like "Unknown," "Homemakers," and "Not working outside the home" would skew our analysis of employment trends. To focus on actual jobs, I filtered down to seven key occupation categories:

- **Management, professional, and related occupations,**
- **Construction, extraction, maintenance and repair occupations,**
- **Farming, fishing, and forestry occupations,**
- **Military,**
- **Production, transportation and material moving occupations,**
- **Sales and office occupations,**
- **Service occupations**



This filtered view gave us a much clearer picture of employment trends. But to really understand each occupation category, I created individual trend charts with regression lines:



These individual charts revealed fascinating patterns in each occupation. Management and professional jobs showed a clear upward trend despite volatility. Service occupations displayed a dramatic decline from 2006 to 2022. Construction jobs showed a slight overall decline with periodic spikes.

The most striking pattern was the volatility - every occupation category showed significant year-to-year fluctuations, suggesting external factors like policy changes, economic conditions, or global events heavily influence immigration patterns.

One particularly interesting observation was the 2016 spike in management and professional occupations, which could relate to policy changes that year. Meanwhile, the post-2016 data appeared more stable across several categories, potentially offering more reliable inputs for our forecasting models.

This initial trend analysis gave us crucial insights for our later forecasting work and highlighted the importance of considering external factors when interpreting immigration patterns. By focusing on actual occupation categories rather than non-employment statuses, we ensured our analysis would provide meaningful insights into labour market trends.

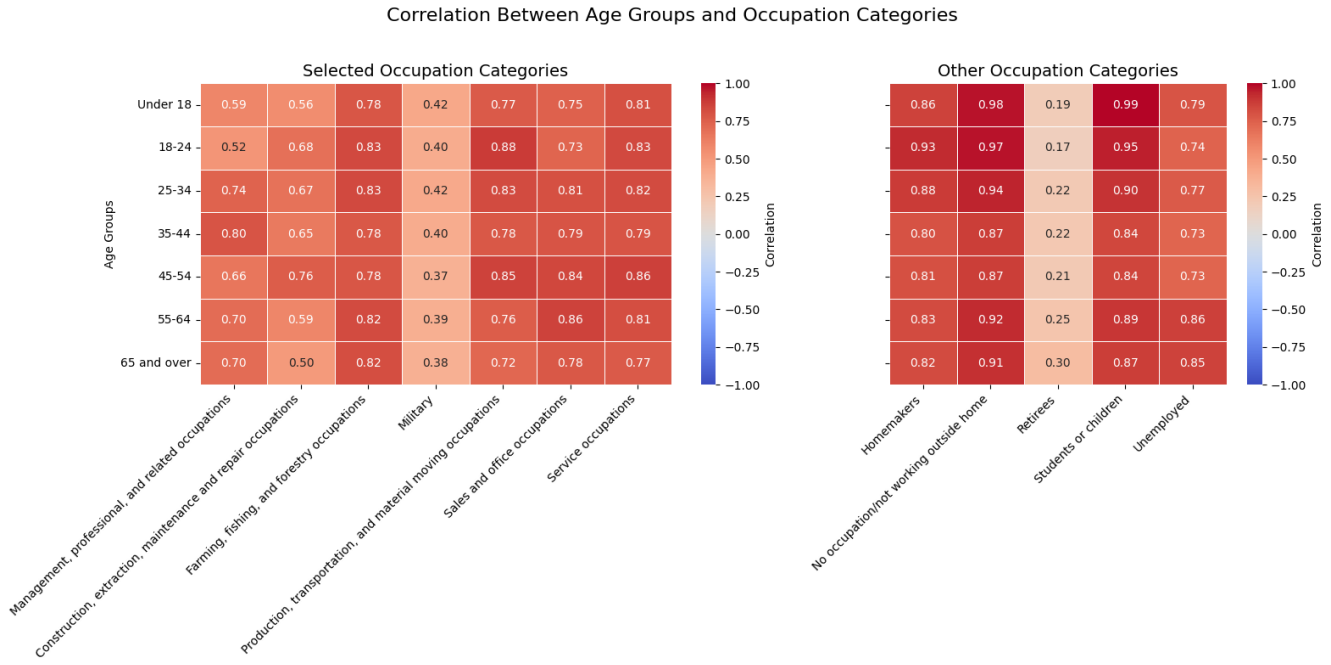
Part 5: Demographic-Occupation Correlation Analysis

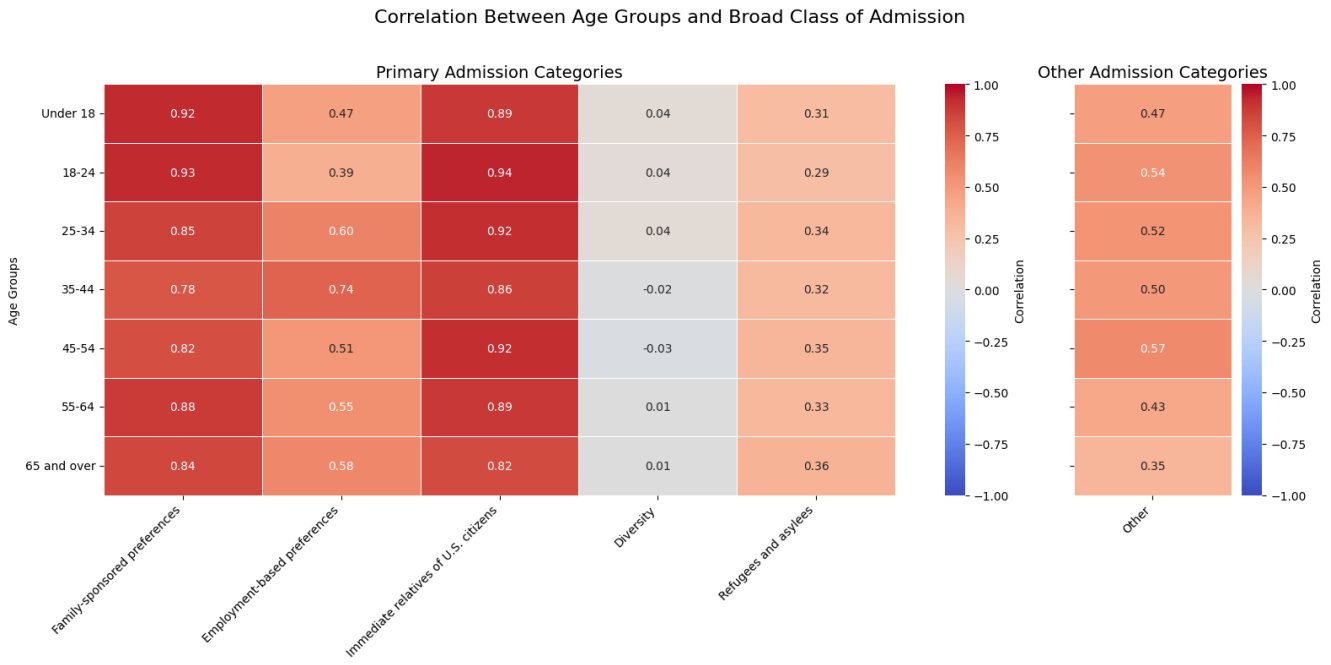
Once I had a clear picture of occupation trends, I wanted to dig deeper to understand who was taking these jobs. Who were the people behind the numbers? To find out, I explored the connections between age groups and occupation choices.

What the Correlations Revealed

Looking at the heatmaps, some clear patterns jumped out:

- Young adults (18-24) gravitated strongly toward production/transportation jobs (0.89) and service work (0.84). These positions often serve as entry points for newcomers still building skills.
- Mid-career immigrants (35-44) showed the strongest connection to management and professional roles (0.80), which makes sense as these jobs typically require more experience and established credentials.
- Military immigration followed its own unique pattern, with weak correlations across all age groups (around 0.50), suggesting different recruitment factors at play.



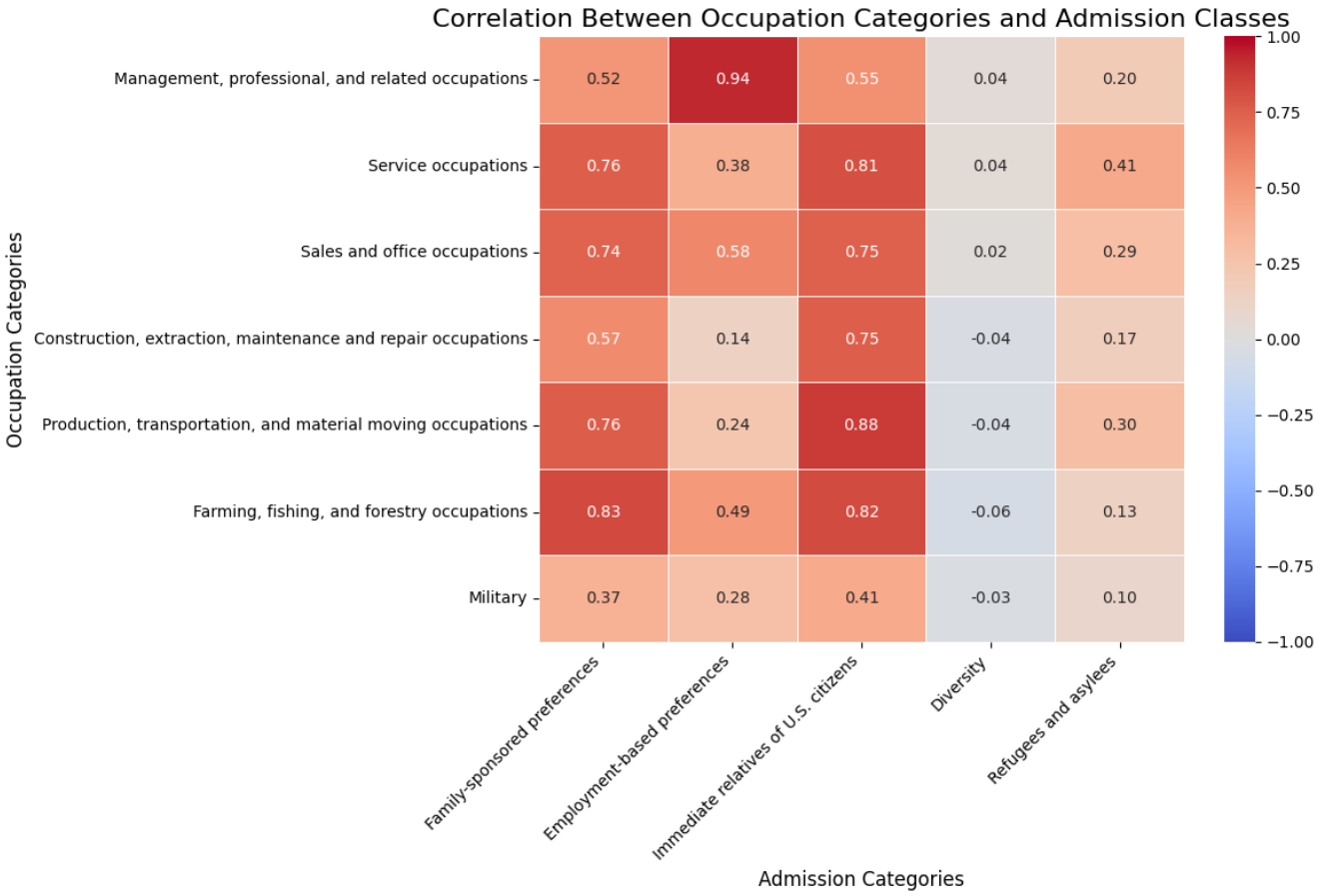


Following the Immigration Pathways

When I connected the dots between immigration categories and occupations, I discovered two main routes into the American workforce:

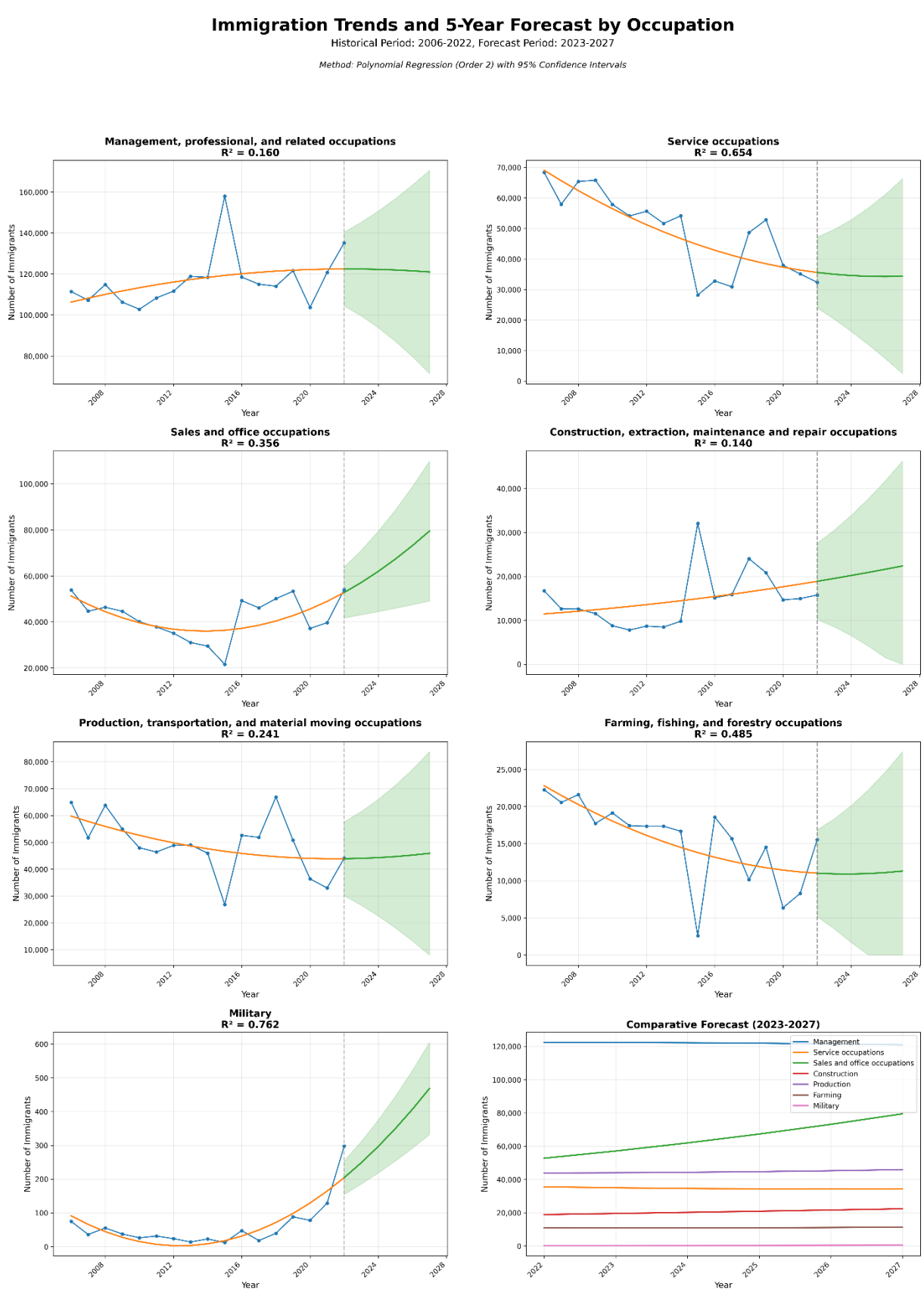
- **The Skilled Professional Route**: Management and professional workers overwhelmingly arrived through employment-based visas (0.94 correlation) - a direct pipeline from skill-based immigration to professional roles.
- **The Family Network Route**: Service workers, production employees, and sales staff primarily came through family connections, with strong correlations to family-sponsored preferences (0.76-0.77) and immediate relatives of citizens (0.75-0.88).

The Diversity Visa program was fascinating in a different way - it showed weak correlations with all occupation types, suggesting these immigrants spread across various job sectors without forming strong patterns.



These insights helped me understand not just the "what" of immigration trends, but the "who" and "why" - essential context for building meaningful forecasts about the future of immigrant occupations.

Part 5: Forecasting Future Occupation Trends



After exploring historical patterns and demographic connections, I turned to the future: What will immigration occupation trends look like in the coming years?

For each key occupation category, I built polynomial regression models that could capture the non-linear patterns in our historical data. The resulting forecast dashboard revealed some fascinating potential futures:

What the Forecasts Tell Us

Looking at the graphs, several clear patterns emerge:

- **Sales and Office Occupations** show the strongest projected growth, with numbers expected to increase dramatically from about 50,000 to nearly 80,000 immigrants by 2027. This surprising turnaround after years of fluctuation suggests possible structural changes in the economy driving demand.
- **Management and Professional Occupations** are projected to maintain relatively steady numbers around 120,000 annually, continuing their role as the largest occupation category for immigrants.
- **Service Occupations** show stabilization after years of decline, leveling off around 35,000-40,000 annually rather than continuing their previous downward trajectory.
- **Construction and Farming** both show modest growth projections despite their historical volatility, suggesting potential recovery in these sectors.
- **Military** immigration shows the most dramatic proportional increase, nearly tripling from recent lows, though absolute numbers remain small relative to other categories.

How Reliable Are These Forecasts?

The R^2 values vary significantly across occupation categories:

- Military forecasts show the strongest fit ($R^2 = 0.762$), suggesting more predictable patterns
- Service occupations also show reasonable predictability ($R^2 = 0.654$)
- Management, Construction, and Production show weaker fits ($R^2 < 0.3$), reflecting their historical volatility

The widening confidence intervals (green shaded areas) appropriately show increasing uncertainty the further we project, with some occupations showing particularly wide ranges by 2027.

To be more specific about the jobs that come under these categories, I did more research and found:

Management, Professional, and Related Occupations

- **Management**: CEOs, executives, operations managers, healthcare administrators
- **Business/Financial**: Accountants, financial analysts, HR specialists, market researchers
- **Computer/Mathematical**: Software developers, data scientists, systems analysts

- **Engineering/Architecture**: All types of engineers, architects, surveyors
- **Science**: Physicists, chemists, biologists, environmental scientists
- **Social Services**: Social workers, counselors, clergy
- **Legal**: Lawyers, paralegals, judges
- **Education**: Teachers, professors, education administrators
- **Arts/Media**: Artists, designers, writers, media specialists

Construction, Extraction, Maintenance and Repair Occupations

- **Construction**: Carpenters, electricians, plumbers, masons, roofers
- **Extraction**: Oil rig workers, miners, drilling operations
- **Installation/Maintenance**: HVAC technicians, industrial machinery mechanics
- **Repair**: Automotive mechanics, electronic equipment repairers

Farming, Fishing, and Forestry Occupations

- Agricultural workers, farm managers, crop harvesters
- Fishers, fishing vessel operators, Animal breeders, ranchers
- Logging workers, forest and conservation workers

Military

- All branches of armed forces personnel
- Both enlisted and officer positions

Production, Transportation, and Material Moving Occupations

- **Production**: Assembly line workers, machinists, food processing workers
- **Transportation**: Truck drivers, bus drivers, pilots, ship captains
- **Material Moving**: Crane operators, warehouse workers, freight handlers

Sales and Office Occupations

- **Sales**: Retail salespeople, real estate agents, insurance agents
- **Office/Administrative**: Secretaries, receptionists, clerks, customer service reps

Service Occupations

- Healthcare support workers (nursing assistants, home health aides)
- Protective service workers (police, security guards, firefighters)

- Food preparation and serving workers
- Building and grounds cleaning/maintenance workers
- Personal care workers (childcare workers, hairdressers)

Conclusion

This comprehensive analysis of lawful permanent resident occupation trends from 2006-2022 reveals a dynamic landscape shaped by policy changes, economic forces, and demographic patterns. Through our exploration of occupation-demographic correlations and polynomial regression forecasting, we've identified both persistent patterns and emerging shifts in immigrant workforce participation.

Our forecasts suggest that while Management/Professional occupations will remain the largest category, Sales and Office roles may see the strongest growth through 2027. Service occupations appear to be stabilizing after years of decline, while Construction and Military categories show potential rebounds from historical lows.

These insights can inform evidence-based policy decisions, workforce development strategies, and immigrant integration programs. The correlation between immigration pathways and occupation outcomes—particularly the strong connection between family-based immigration and service/production sectors—highlights the multifaceted economic contributions of immigrants arriving through various channels.

Further research could explore how external factors like policy changes, economic cycles, and global events influence these patterns, potentially improving forecast accuracy and providing deeper contextual understanding.

References

- U.S. Department of Homeland Security. (2022). Yearbook of Immigration Statistics. https://www.dhs.gov/immigration-statistics/yearbook
- Department of Labor - Bureau of Labor Statistics Standard Occupational Classification System (which DHS uses as a basis for their categories): https://www.bls.gov/soc/
- Anthropic. (2023). Claude: AI assistant developed by Anthropic to be helpful, harmless, and honest. https://www.anthropic.com/claude
- Regression Analysis | Full Course 2025 by DataTab. https://www.youtube.com/watch?v=T5AoqxQFkzY&t=503s