MONASH University

Data Exploration Project

FIT5147 Data Exploration and Visualization

Semester 1 2023

Name: Nazmus Sakib

Student ID: 33361991

Tutor: Niranjan Nanjunda

Introduction

Problem Description

The global refugee crisis has highlighted the need for countries to provide asylum to refugees, not only for their protection but also to offer them opportunities for employment and resettlement. However, it is important to note that the burden of accepting refugees should not fall solely on one country, as this could lead to an overburdening of resources and negatively impact the well-being of both refugees and host communities.

To gain a better understanding of the refugee situation, this project aims to visualize the flow of refugees in a country and compare it with job demand. By collecting data on the number of refugees entering the country and the availability of job opportunities in different regions or industries, we can analyze the distribution of refugees and the potential economic impact of their resettlement.

Questions

1. What percentage of the refugee population is employed, and how does this vary by country of origin, age, gender, and level of education?
2. What percentage of employment demand from the asylum countries can refugees fulfill?

The first question seeks to understand the employment status of refugees and how it varies based on different factors such as country of origin, age, gender, and education level. This question is important because employment is crucial for refugees to rebuild their lives and become self-sufficient in their host country. It also helps in assessing the economic contribution of refugees and their potential impact on the labor market. By analyzing the percentage of employed refugees, we can identify any gaps in employment opportunities and potential barriers to employment that refugees may face based on their background.

The second question aims to determine the percentage of job demand from asylum countries that refugees can fulfill. This question is essential to assess the potential economic impact of refugee resettlement on the host country. If refugees can fulfill a significant portion of employment demand, it can positively impact the labor market, the overall economy, and reduce any potential burden on the host country's social welfare system. It can also help identify any skill gaps among the refugee population and guide targeted training programs to enhance their employability.

Motivation

Gaining insights into the employment status of refugees can help dispel common misconceptions about refugees being a burden on the host country's economy. Instead, evidence-based research can highlight the significant economic contributions that refugees can make to the host country's labor market and economy.

By answering the two questions, we can also promote more inclusive and sustainable refugee resettlement practices, which prioritize the economic and social well-being of both refugees and host communities. Therefore, this research is crucial for policymakers, researchers, and advocates who seek to support refugees' successful integration into their host communities and facilitate their long-term economic self-sufficiency.

Data Wrangling and Data Checking

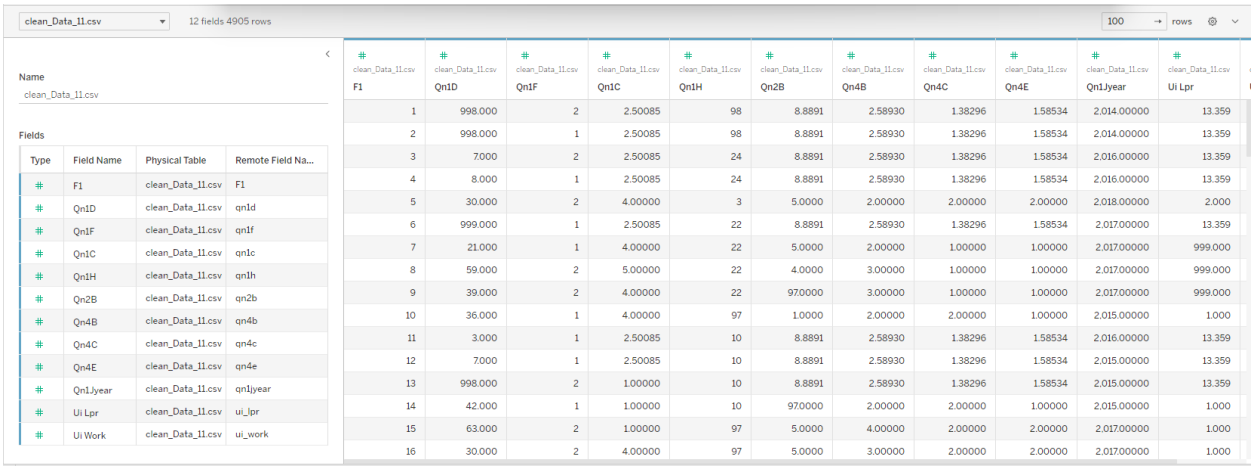
Description of Data Sources

There are three main datasets used along with some supplementary datasets.

The main dataset is the Annual Survey of Refugees 2019 (ASR 2019) dataset which is got UNHCR microdata unit, Central Data Catalog. Though the dataset is named to be of the year 2019 but it is updated every year. The last modification was made on Jun 21, 2022. Thus, the dataset is categorized as a reliable dataset. This dataset has almost 5000 cases and more than 250 variables, but the main objective of this project is to work with unemployment of the refugees. So for this I have selected the variables that are relevant to unemployment. A total of 12 variables are selected, these variables are a code but here the description is given as well, they are listed as follows:

- qn1c - is the person married or not?
- qn1d - what is the person's age at the last birthday?
- qn1f - Male or female?
- qn1h – Citizenship of which country?
- qn1jyear - what month and year did this person enter in US to stay?
- qn2b - what is the highest degree or certificate that this person obtained?
- qn3b – what kind of work did this person perform before coming to US?
- qn4b - how well does this person speak English?
- qn4c - after coming to us did this person have any English instruction?
- qn4e – within the past 12 months, has the person attended an English language training?
- ui_lpr - legal permanent residency status
- ui_work - work status

When the dataset is loaded in tableau it looks like this.



clean_Data_11.csv | 12 fields 4905 rows | 100 rows

Name: clean_Data_11.csv

Fields

Type	Field Name	Physical Table	Remote Field Na...
#	F1	clean_Data_11.csv	F1
#	Qn1D	clean_Data_11.csv	qn1d
#	Qn1F	clean_Data_11.csv	qn1f
#	Qn1C	clean_Data_11.csv	qn1c
#	Qn1H	clean_Data_11.csv	qn1h
#	Qn2B	clean_Data_11.csv	qn2b
#	Qn4B	clean_Data_11.csv	qn4b
#	Qn4C	clean_Data_11.csv	qn4c
#	Qn4E	clean_Data_11.csv	qn4e
#	Qn1Jyear	clean_Data_11.csv	qn1jyear
#	Ui Lpr	clean_Data_11.csv	ui_lpr
#	Ui Work	clean_Data_11.csv	ui_work

F1	Qn1D	Qn1F	Qn1C	Qn1H	Qn2B	Qn4B	Qn4C	Qn4E	Qn1Jyear	Ui Lpr
1	998.000	2	2.50085	98	8.8891	2.58930	1.38296	1.58534	2,014.00000	13.359
2	998.000	1	2.50085	98	8.8891	2.58930	1.38296	1.58534	2,014.00000	13.359
3	7.000	2	2.50085	24	8.8891	2.58930	1.38296	1.58534	2,016.00000	13.359
4	8.000	1	2.50085	24	8.8891	2.58930	1.38296	1.58534	2,016.00000	13.359
5	30.000	2	4.00000	3	5.0000	2.00000	2.00000	2.00000	2,018.00000	2.000
6	999.000	1	2.50085	22	8.8891	2.58930	1.38296	1.58534	2,017.00000	13.359
7	21.000	1	4.00000	22	5.0000	2.00000	1.00000	1.00000	2,017.00000	999.000
8	59.000	2	5.00000	22	4.0000	3.00000	1.00000	1.00000	2,017.00000	999.000
9	39.000	2	4.00000	22	97.0000	3.00000	1.00000	1.00000	2,017.00000	999.000
10	36.000	1	4.00000	97	1.0000	2.00000	2.00000	1.00000	2,015.00000	1.000
11	3.000	1	2.50085	10	8.8891	2.58930	1.38296	1.58534	2,016.00000	13.359
12	7.000	1	2.50085	10	8.8891	2.58930	1.38296	1.58534	2,015.00000	13.359
13	998.000	2	1.00000	10	8.8891	2.58930	1.38296	1.58534	2,015.00000	13.359
14	42.000	1	1.00000	10	97.0000	2.00000	2.00000	1.00000	2,015.00000	1.000
15	63.000	2	1.00000	97	5.0000	4.00000	2.00000	2.00000	2,017.00000	1.000
16	30.000	2	4.00000	97	5.0000	3.00000	2.00000	2.00000	2,017.00000	1.000

Figure 1

The second Dataset is from the US Bureau of Labor Statistics which contains the labor force of US region from the year 2000 to 2023. The data can be collected by year, quarterly or monthly. As for the thoroughness of the project the monthly data of some variables are collected.

When the data is loaded it looks like this.

Name
unemployement_rate.csv

Fields

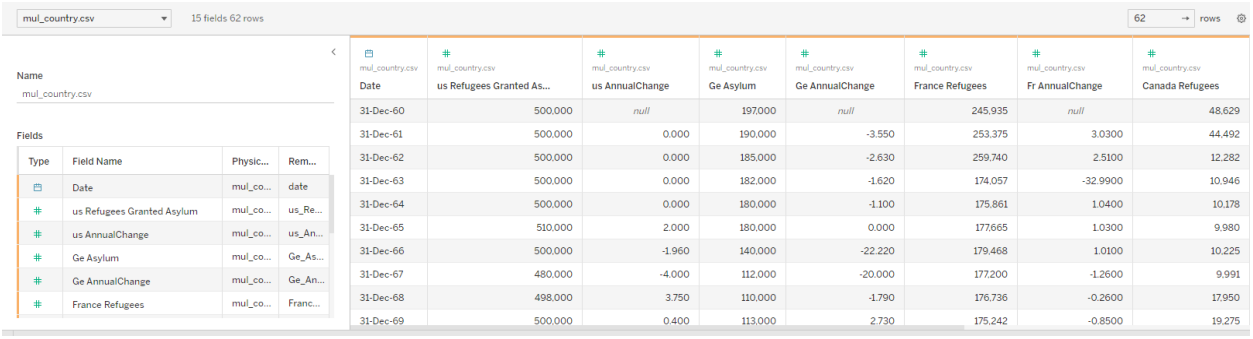
Type	Field Name	Physi...	Rem...
#	Year	unemp...	Year
#	Jan	unemp...	Jan
#	Feb	unemp...	Feb
#	Mar	unemp...	Mar
#	Apr	unemp...	Apr
#	May	unemp...	May
#	Jun	unemp...	Jun

unemployement_rate.csv Year	unemployement_rate.csv Jan	unemployement_rate.csv Feb	unemployement_rate.csv Mar	unemployement_rate.csv Apr	unemployement May
2013	8.00000	7.70000	7.50000	7.6000	7
2014	6.60000	6.70000	6.70000	6.2000	6
2015	5.70000	5.50000	5.40000	5.4000	5
2016	4.80000	4.90000	5.00000	5.1000	4
2017	4.70000	4.60000	4.40000	4.4000	4
2018	4.00000	4.10000	4.00000	4.0000	3
2019	4.00000	3.80000	3.80000	3.6000	3
2020	3.50000	3.50000	4.40000	14.7000	13
2021	6.30000	6.20000	6.10000	6.1000	5
2022	4.00000	3.80000	3.60000	3.6000	3
2023	3.40000	3.60000	3.50000	null	null

Figure 2

A third Dataset is used which gives information of Asylum countries with USA it gives information of Germany, Canada, France and USA.

When the dataset is loaded in the it looks like this



mul_country.csv 15 fields 62 rows

62 rows

Name
mul_country.csv

Fields

Type	Field Name	Physic...	Rem...
📅	Date	mul_co...	date
#	us Refugees Granted Asylum	mul_co...	us_Re...
#	us AnnualChange	mul_co...	us_An...
#	Ge Asylum	mul_co...	Ge_As...
#	Ge AnnualChange	mul_co...	Ge_An...
#	France Refugees	mul_co...	Franc...

mul_country.csv Date	mul_country.csv us Refugees Granted As...	mul_country.csv us AnnualChange	mul_country.csv Ge Asylum	mul_country.csv Ge AnnualChange	mul_country.csv France Refugees	mul_country.csv Fr AnnualChange	mul_country.csv Canada Refugees
31-Dec-60	500,000	null	197,000	null	245,935	null	48,629
31-Dec-61	500,000	0.000	190,000	-3.550	253,375	3.0300	44,492
31-Dec-62	500,000	0.000	185,000	-2.630	259,740	2.5100	12,282
31-Dec-63	500,000	0.000	182,000	-1.620	174,057	-32.9900	10,946
31-Dec-64	500,000	0.000	180,000	-1.100	175,861	1.0400	10,178
31-Dec-65	510,000	2.000	180,000	0.000	177,665	1.0300	9,980
31-Dec-66	500,000	-1.960	140,000	-22.220	179,468	1.0100	10,225
31-Dec-67	480,000	-4.000	112,000	-20.000	177,200	-1.2600	9,991
31-Dec-68	498,000	3.750	110,000	-1.790	176,736	-0.2600	17,950
31-Dec-69	500,000	0.400	113,000	2.730	175,242	-0.8500	19,275

Figure 3

Data Cleaning

The ASR dataset has a lot of has more than 250 variables, so after going through all the variables the 12 variables are selected. From those 12 variables a new dataset that is named Clean Dataset is taken. The cleaning process includes thoroughly checking all the variables from the data description to select the important variables of the dataset. The dataset had a lot of missing values and the arbitrary values. The missing values are handled by checking the distribution of the values and filling it with required values. If the missing values are more than 40% of the data, then the column is omitted and if the missing values are less than 40% then it is filled with the majority values and if the missing values are less than 20% then it is filled with the average value. The outliers are omitted as there are values which are way beyond the range of the values. The cleaned data is then imported in Tableau for the visualizations as shown above. The tool used for this Dataset is Python for cleaning.

	Variable	hhid	personid	respondent	cohort	qn1a	numppl	qn1b	qn1c	qn1d	...	qn34b_05	ui_qn8a_annual	ui_qn10a_annual	ui_cashassist	ui_lfp	ui_emprate	ui_
0	NaN	2	22	0	3	2	5	11	NaN	998.0	...	NaN	NaN	NaN	2	NaN	NaN	
1	NaN	2	23	0	3	3	5	11	NaN	998.0	...	NaN	NaN	NaN	2	NaN	NaN	
2	NaN	2	25	0	3	5	5	11	NaN	7.0	...	NaN	NaN	NaN	2	NaN	NaN	
3	NaN	2	24	0	3	4	5	11	NaN	8.0	...	NaN	NaN	NaN	2	NaN	NaN	
4	NaN	2	21	1	3	1	5	1	4.0	30.0	...	NaN	NaN	NaN	2	1.0	1.0	

Figure 4

The second Dataset is the US Bureau of Labor Statistics dataset and this dataset had only few missing values which of the year 2023 and the year 2023 is not included as it had missing values. As the data is for public display and can only be edited by the US Bureau of Labor Statistics and it is being updated so the data is assumed to be correct. Python is also used for cleaning this dataset.

The third dataset was also clean, and it had data from the year 1960 to 2023. It just had a few missing values which are filled with average values of that column because the number of missing values were very less in percentage. Python is used for cleaning the dataset.

Data Exploration

ASR Dataset

As the main dataset, the ASR dataset is used so the cleaned dataset is imported in the Tableau and some test are carried out. As most of the variables has a range of numbers the arbitrary values are filtered out for a good visualization.

This Shows the total participants of the ASR Dataset which are divided in male and female but note that some participants (2%) actually refused to give away their gender and 1% of participants were confused with their gender. Except this the dataset is evenly distributed between male and female as 50% is female and 49.7% is male in this dataset.

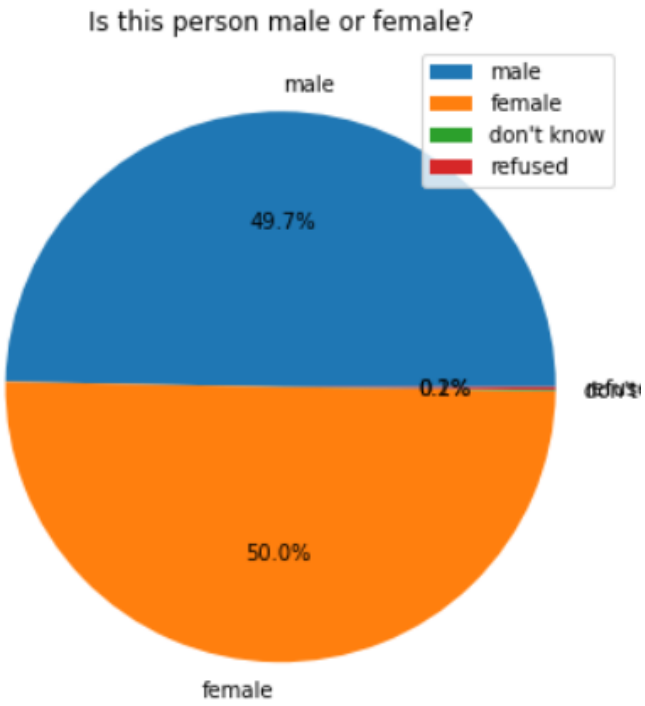


Figure 5

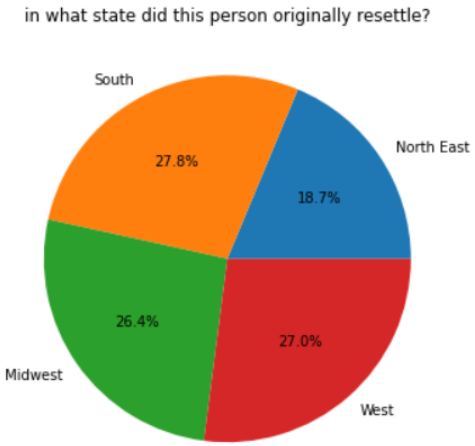


Figure 6

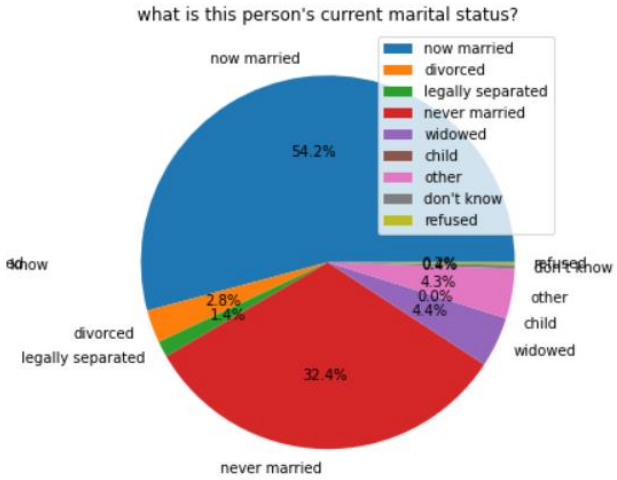


Figure 7

From the above diagram on the left, It can be seen that most refugees settled in the in South, Midwest and West region of US but on Northeast we can not say that they didn't settle there but North East has a bit less settlement compared to the other region of US.

The second diagram on the right shows us that more than half of the participants were recently married, and their partner does not actually stay with them and 32.4% were single. The next largest portion is the widowed portion which comes to 4.4%, very few (0.2%) actually refused to say anything about their marital status and 0.4% did not know about their status while they took shelter.

This graph shows how they spoke English while entering the country. The amount of well spoken participants are only 23% of the population and very well were only 11% of populations, 14% did not even know English while being granted in the asylum country.

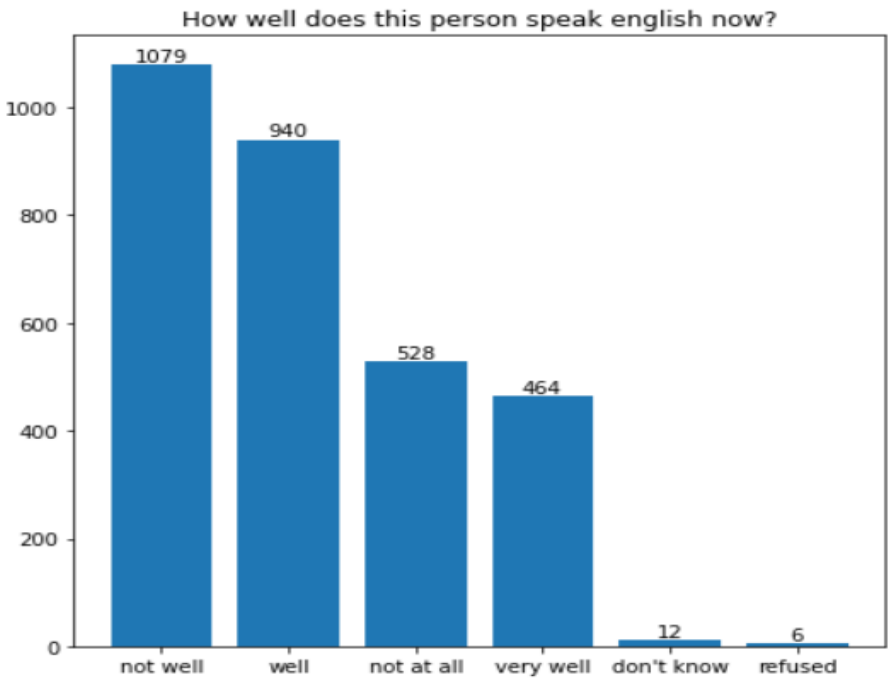


Figure 8

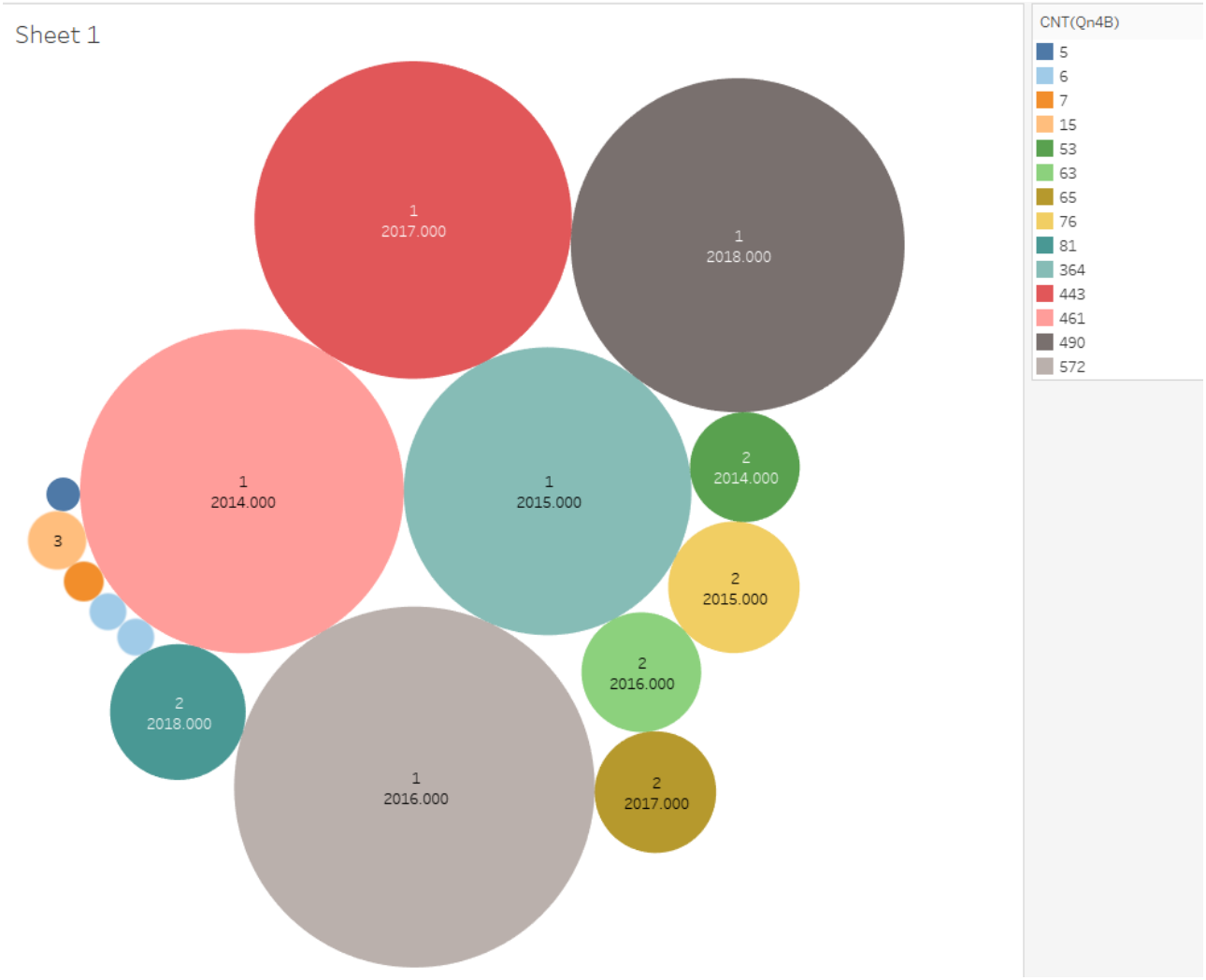


Figure 9

The above visualizations show how many entered in which year with their pr status when the survey was being conducted. 1 denotes they have already adjusted the permanent residency (PR), 2 denotes they are planning to adjust PR status in future, 3 denotes they have not applied to adjust PR yet or may not. The size denotes the number of participants, and the year is displayed in which they were granted asylum.

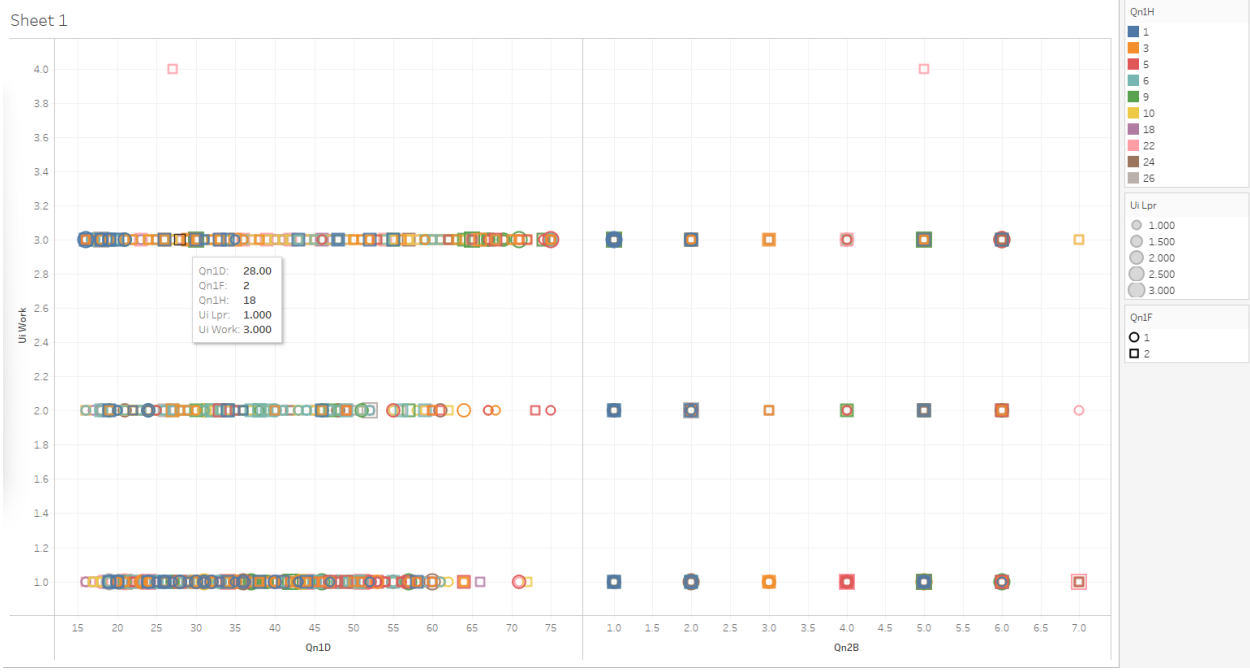


Figure 10

This image shows the Age (Qn1D), Education Level (Qn2B), Gender (Qn1f), Origin (Qn1H) and PR Status (Ui Lpr) against Work Status (Ui Work). For Work Status working now = 1, not working now but previously worked in us = 2, not working now and never worked in us = 3, not working now and unsure about previous work in = 4. For Gender 1 is Male and 2 is Female. Qn1d denotes the age of the refugee and for Qn2B 1 = none, 2 = primary, 3 = training in refugee camp, 4 = technical school, 5 = secondary, 6 = university, 7 = medical degree. The origin has 10 different war zone countries.

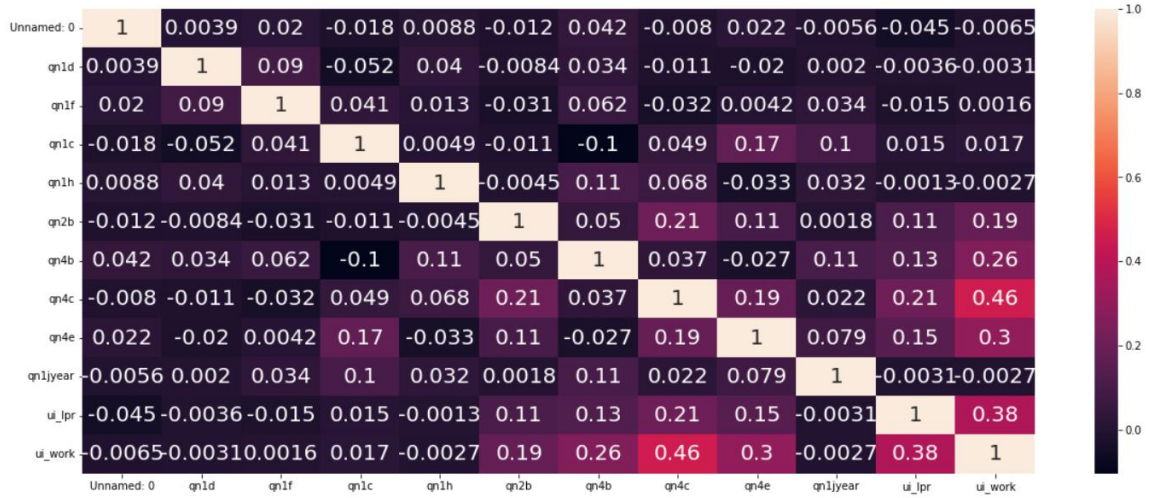


Figure 11

The correlation matrix shows that none of the variables are strongly related to any of the other variables, as 1 means very strong positive correlation and -1 means strong negative correlation with 0 being no relation.

US Bureau of Labor Statistics Dataset

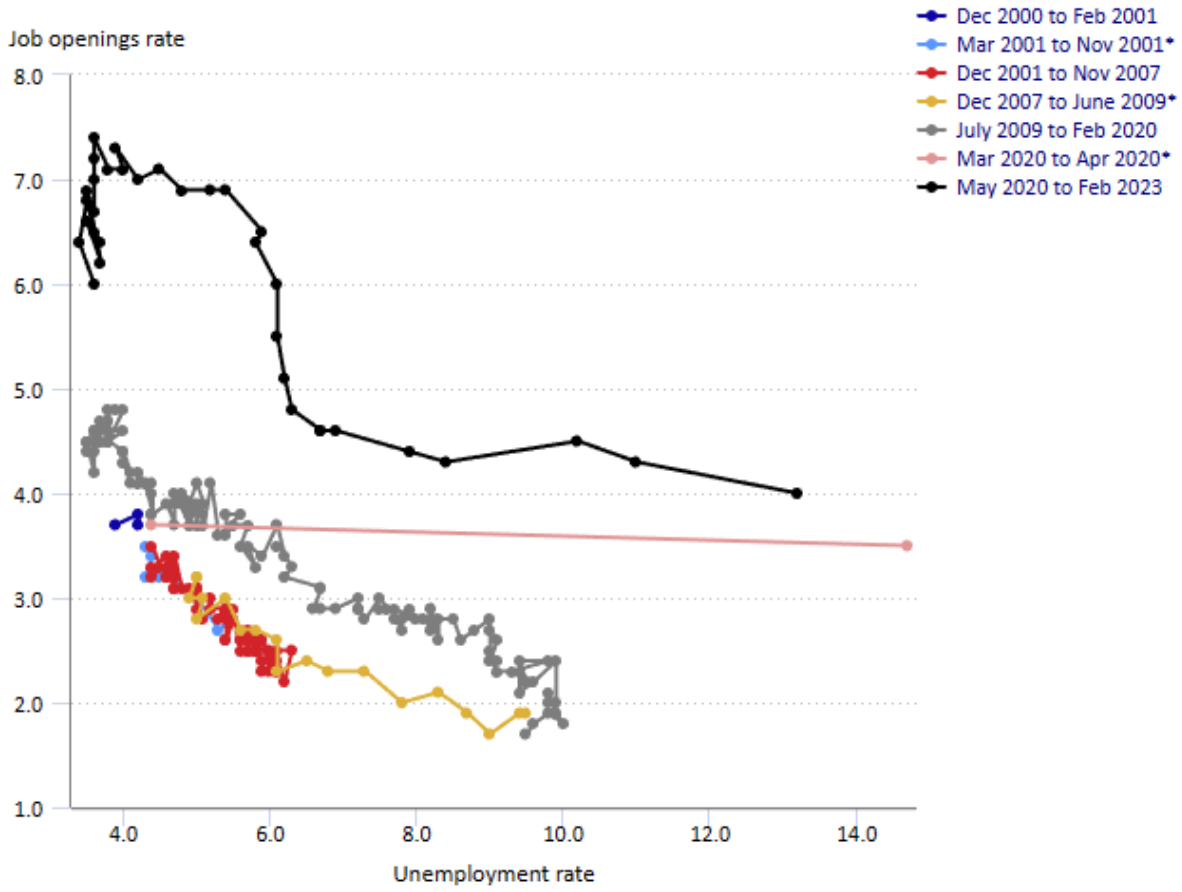
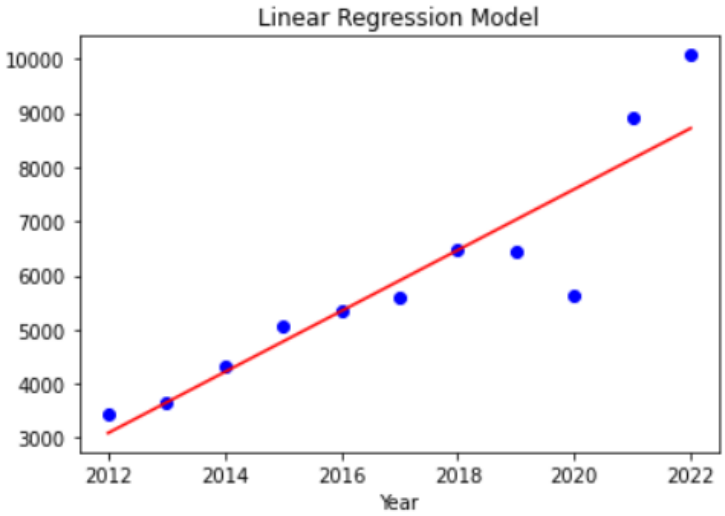


Figure 12

This graph displays the Job Opening Rate vs the Unemployment Rate. It can be observed that the most unemployment rate was from March 2020 to April 2020 with a bit decreased rate for new jobs, which was the start of the pandemic. And only in 2021 the rate was that high unemployment and low job opening rate is seen.



The Linear regression with number of jobs opening against year can be seen with uniform increment but only on the year 2020 the number decreased.

This is Figure 13

Asylum Countries Dataset

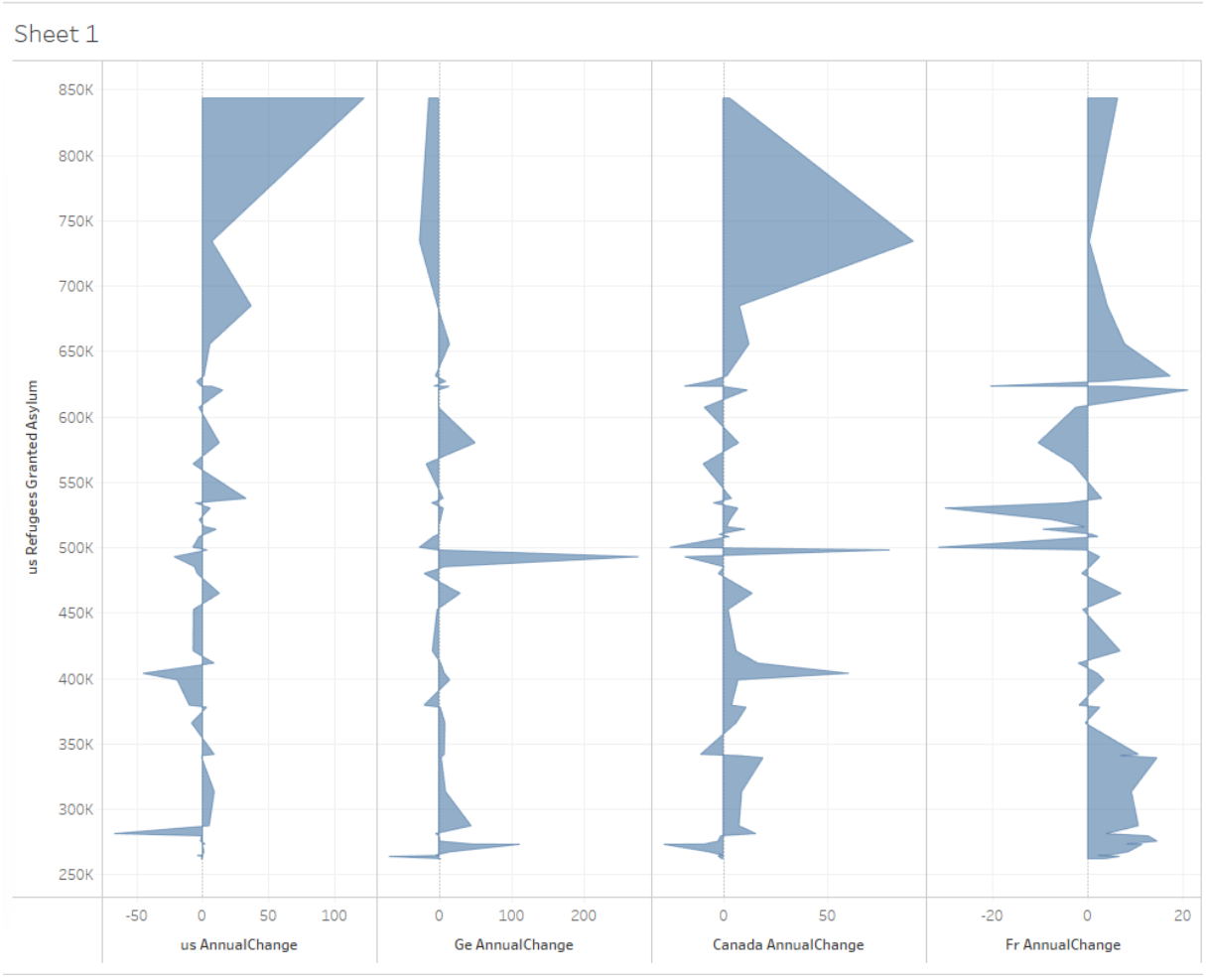


Figure 14

In this Graph the number of US Refugee granted against the US Annual Change is plotted and compared to the Annual Change of Germany Canada and France. We can see that on the high number Canada and US are the ones that were taking in refugees with a mass number. France and Germany we only interested in giving asylum to comparatively less refugees.

Conclusion

The main two questions were:

1. What percentage of the refugee population is employed, and how does this vary by country of origin, age, gender, and level of education?

2. What percentage of employment demand from the asylum countries can refugees fulfill?

To find the answer of the first question we can conclude that almost closed to 50% of the refugees are working now as from the Figure 10 we can see that the Ui Work is mostly on level 1 which means they

are currently working while the other 50% are not working now, or never worked and not sure if they will work or not. It can be observed that war prone countries are contributing most of the refugees and almost 60% of the working refugees are of age between 20 to 40, The refugees that are not working are not that educated and are mostly women and elderly people, though it includes some you and male people also. It can also be observed that only women are unsure of they want to work or not. Most of the working refugees, almost 50% are of either primary, secondary or university graduate and they are male. Though the percentage of employed refugees and the unemployed refugees is almost the same, it heavily depends on education and age.

For answer of the second question, a conclusion can not be made as more data is required for that analysis. We could only see that refugees granted asylum and Annual Change of US, the correct number of total refugees and the total number of refugees that are working now is yet to be determined. The job opening rate and the unemployment rate is discovered but that does not say anything about refugees, so those graphs are for refugees as well as citizens. A refugee employment dataset is needed for the analysis of the second question.

Reflection

What I learned about the project is that the datasets that I chose do not actually give enough information to come to a rigid conclusion for either of the questions, but it gives insights on the first question only. I need to find out more datasets and combine some of the datasets from the US Bureau of Labor Statistics together to find out the answer to the second question. The website has a lot of different datasets, and those datasets need to be combined to a common pivot point to come to any conclusion.

References

- UNHCR Microdata Library. (n.d.). Refugee and Returnee Data [Data set]. UNHCR. https://microdata.unhcr.org/index.php/catalog/681/study-description
- Inter-university Consortium for Political and Social Research. (n.d.). Eurobarometer 67.2: Citizenship in the European Union, October-November 2007 [Data set]. OpenICPSR. https://www.openicpsr.org/openicpsr/project/168641/version/V2/view;jsessionid=C85B31ADCB9FE7D38A5B2BB055ACD1D1
- Inter-university Consortium for Political and Social Research. (2019). 2019 American Society of Reproductive Medicine (ASRM) Public Use File [Data set]. OpenICPSR. https://www.openicpsr.org/openicpsr/project/168641/version/V2/view;jsessionid=C85B31ADCB9FE7D38A5B2BB055ACD1D1?path=/openicpsr/168641/fcr:versions/V2/2019-ASR_Public_Use_File.dta&type=file
- U.S. Bureau of Labor Statistics. (n.d.). Job Openings and Labor Turnover: Total Nonfarm [Graph]. Bureau of Labor Statistics. https://www.bls.gov/charts/job-openings-and-labor-turnover/unemp-per-job-opening.htm
- U.S. Bureau of Labor Statistics. (n.d.). Labor Force Statistics from the Current Population Survey [Data set]. Bureau of Labor Statistics. https://data.bls.gov/PDQWeb/ln

- U.S. Bureau of Labor Statistics. (n.d.). Survey Output [Data set]. Bureau of Labor Statistics. https://data.bls.gov/pdq/SurveyOutputServlet
- Macrotrends. (n.d.). U.S. Refugee Statistics - Historical Data [Data set]. MacroTrends. https://www.macrotrends.net/countries/USA/united-states/refugee-statistics#:~:text=U.S.%20Refugee%20Statistics%20-%20Historical%20Data%20%20,%20%209.09%25%20%2058%20more%20rows%20
- Hauser, C. (2023, January 20). US State Department creates new refugee program called "Welcome Corps". The Guardian. https://www.theguardian.com/us-news/2023/jan/20/us-state-department-new-refugee-program-welcome-corps
- The New York Times. (n.d.). Refugees and Displaced People [Webpage]. The New York Times. https://www.nytimes.com/topic/subject/refugees-and-displaced-people