Visualizing U.S. Immigration and Economy
CSCE 679 - Final Project Report

Millennium Bismay - 633002191
Sai Ramana Reddy - 332009997

1 Goal

In an era characterized by globalization and interconnectedness, immigration plays a pivotal role in shaping the socio-economic landscape of nations. The United States, renowned for its cultural diversity and economic prowess, stands as a testament to the profound influence of immigration on national development. The United States has long been a magnet for skilled and unskilled labor, entrepreneurs, and professionals from across the globe. This report seeks to delve into the intricate web of U.S. immigration, focusing specifically on some key visa categories such as H-1B, H-2B, H-2A, E-3, H-1B1, etc. which contributes significantly to the large number of workers over the period from 2010 to 2020. By examining the multifaceted dimensions of these visa programs, we aim to unravel the diverse implications they have on the U.S. economy.

a. Why did you choose the topic?

Immigration and Economy are two closely related topics we found interesting. We chose this topic because we thought it would be interesting to collect data from multiple government sources about multiple visa-types (immigration), GDP and contribution of each important socio-economic sector to GDP, and narrate a story using visualizations generated from that data.

b. Who is the audience?

The audience for our project is anyone who is eager to understand the socio-economic dynamics of immigration and economy. A basic knowledge of the U.S foreign worker system and the U.S. industrial sectors is really helpful, however we aim to keep our project as easy to understand for anyone and everyone. Also, researchers in academia as well as policymakers can use this to understand the influence of immigration on the country's economy.

c. How do you expect this to be used (how frequently? In what environment?)?

We expect this project to be used in settings such as a conference or a research study. For example, policymakers might use our project to understand the contributions of immigration and its multifaceted impact on the country's economy. Researchers might use it to understand what factors contribute to immigration. Our project is intended to be used in a one-time interactive session in which the audience make observations from the visualizations.

d. What do you want people to learn/understand/be able to do with this visualization?

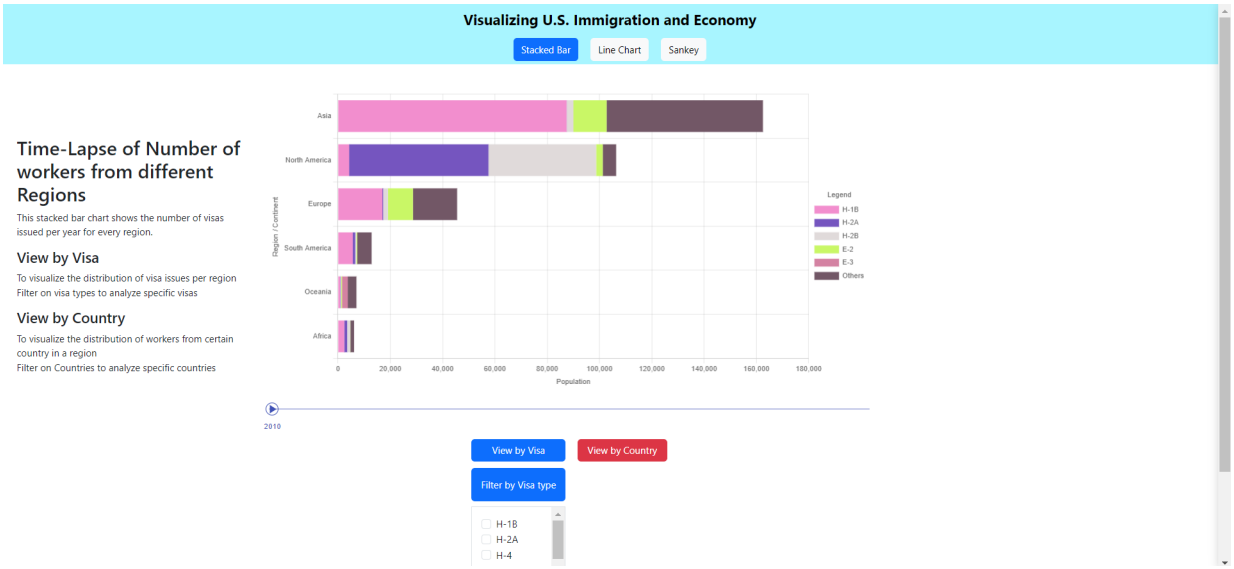
We provide a guided environment in our website for people to interact with the visualizations. The visualizations narrates a story which starts with the number of visas that each country and region has been allotted by the government of the United States. Then they can understand how the mean wage has changed over the last 10 years, from 2010 to 2020 for various states, sectors, and visa-types. And finally, they will see the broader picture of the composition of foreign workers in each sector and each sector's contribution to the GDP. In every part of our three-fold visualization story-telling, users will be able to interact and analyze the impact and effect of immigration to the economy.

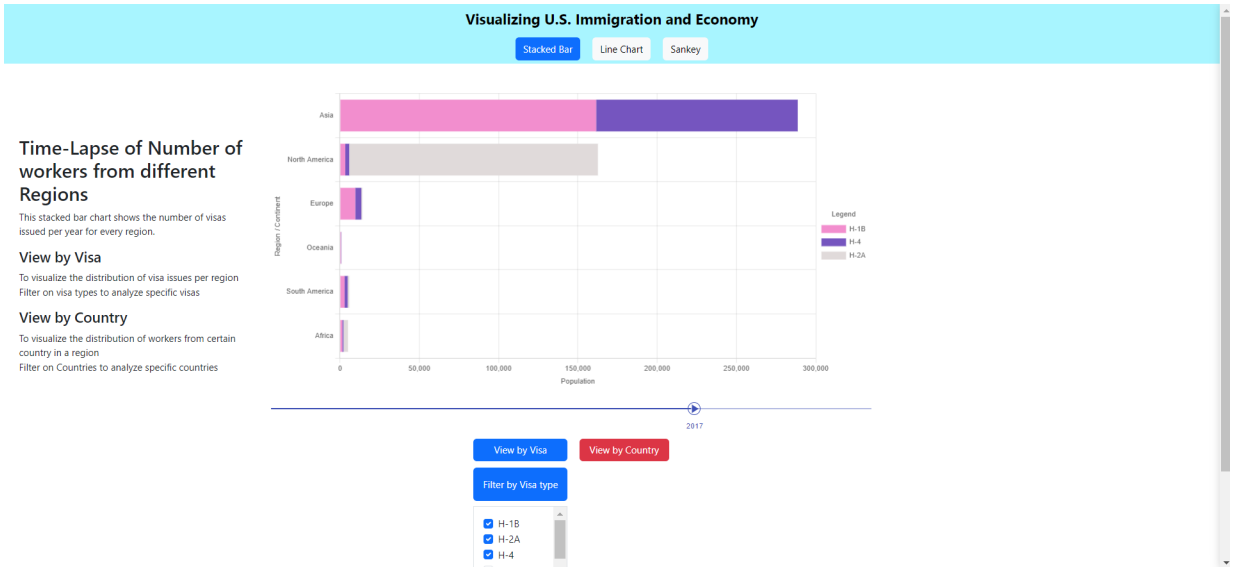
2 Visualizations

2.1 Time Lapse using Horizontal Stacked Bar Chart

The data visualization is a time lapse of number of visas (type of workers) coming into the United States from different countries and regions from the year 2010 to 2020. We have broadly divided the regions into 6 parts - Asia, North America, Europe, Africa, Oceania, and South America. The NIV Data has been used for this visualization and users will be able to interact with it by filtering specific country or visa types.

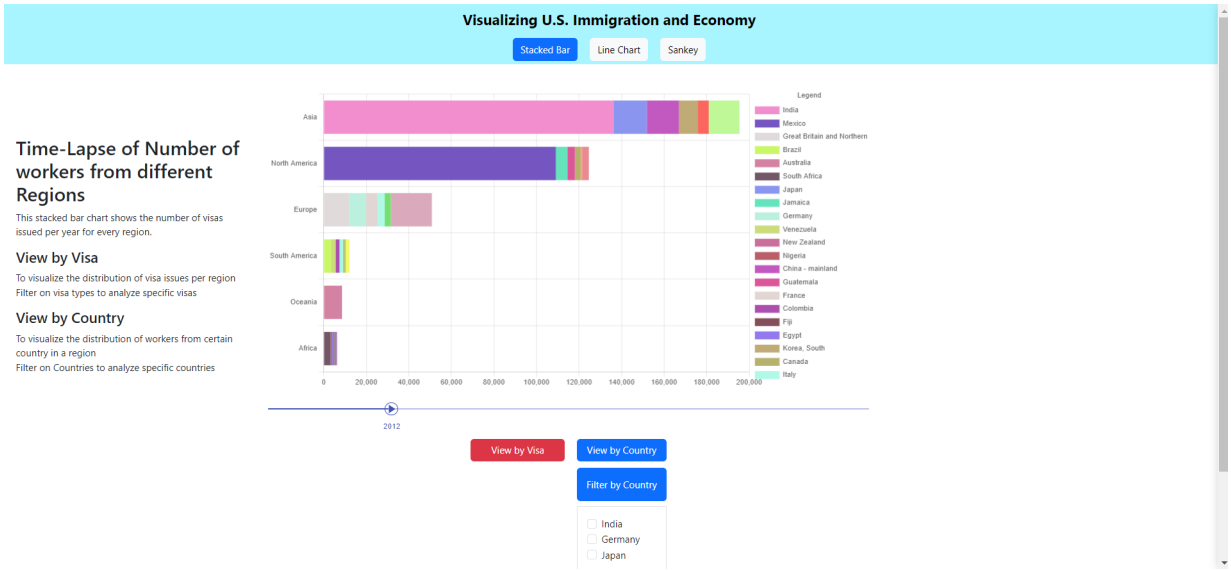
- User will land on the Stacked Bar tab of our website and the visualization will show case the 'View by Visa' by default. They will be able to see the paused version of the time-lapse for the year of 2010. They can click on the play button to start the time-lapse and if they can also pause the time-lapse at any point they want for further analysis.
- The 'View by Visa' will have options to filter by 'Visa Types' which has a long list of visa types in which we are interested in. They can filter multiple visa types to understand the dynamics of those visa types.
- They can hover upon each stack to see the exact number of people comping in each visa type from each region.

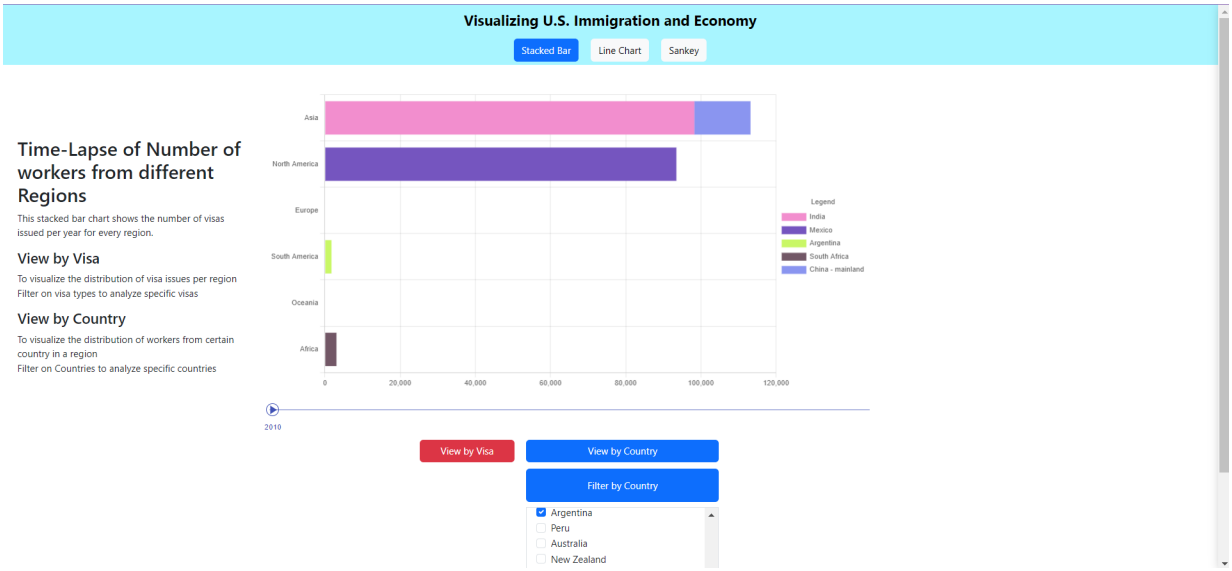




- User can change the view type to 'View by Country' at any time during the time-lapse and the view would continue from the same point. They can also restart if they want to do so. This view visualizes the proportion of people coming to the United States from each country from a region. For the clear verbosity, we have kept the top 3 countries from each region and have given the option to filter from any country as the user wishes.

- This can give a clear idea as to which countries contributes the most number of workers to the States. This in turn can help in analyze the sentiment of certain nationalities towards the United States which can also help in establishing certain economic relations.

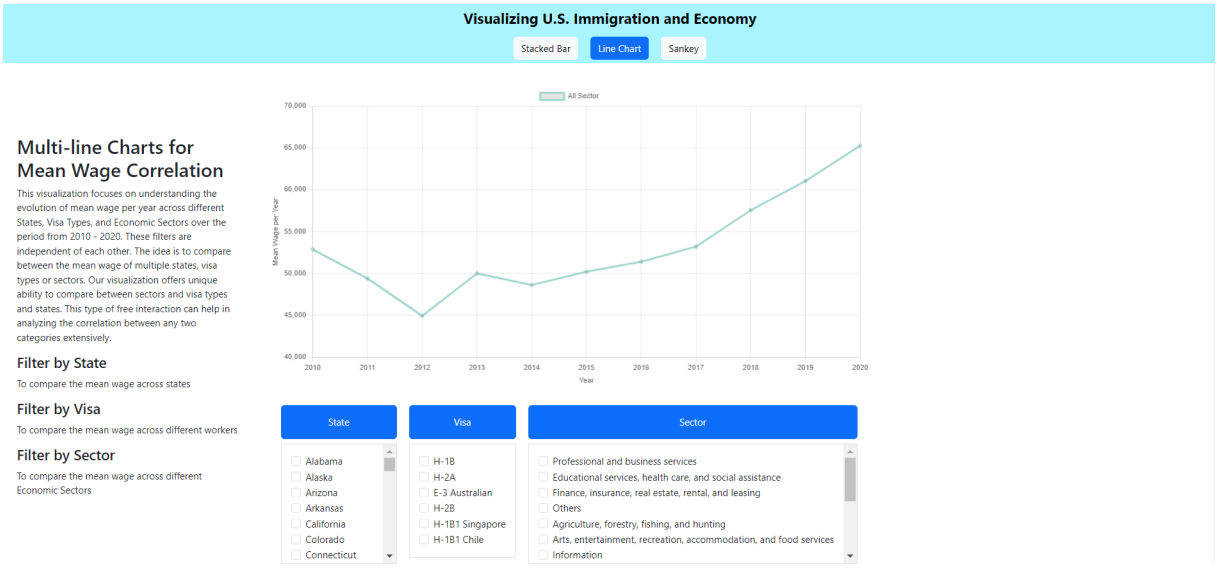




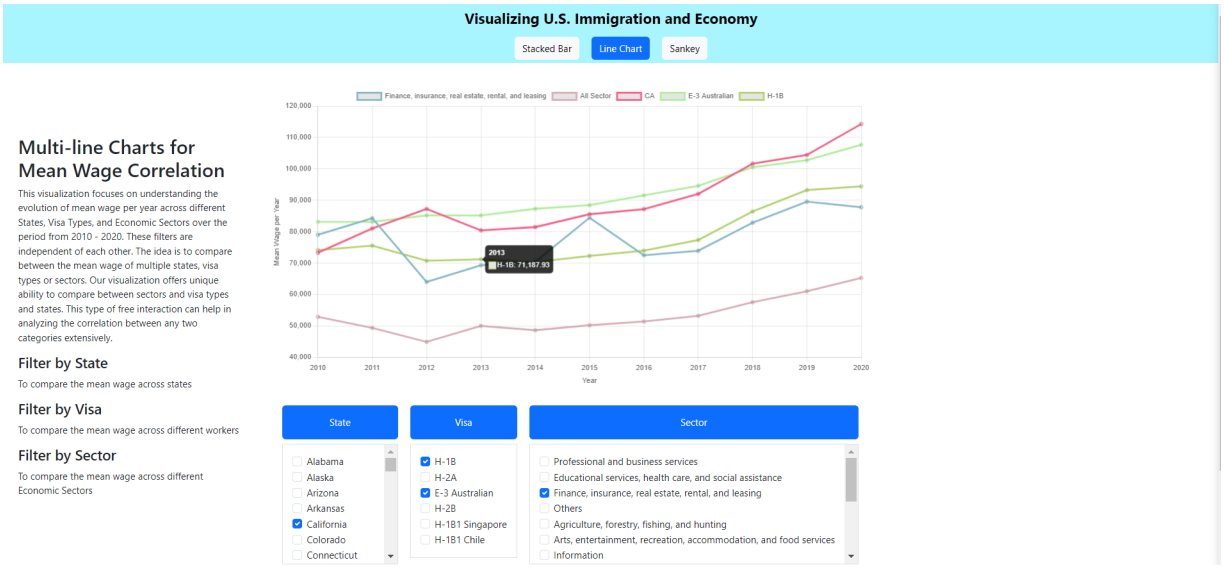
2.2 Multi-line Charts for Mean Wage Correlation

This visualization focuses on understanding the evolution of mean wage per year across different States, Visa Types, and Economic Sectors over the period from 2010 - 2020. We used the LCA Data for this visualization.

- User will be presented with the mean wage line chart for all sectors over the period from 2010 to 2020 on selecting the Line Chart tab on our website.
- There will be options for selecting multiple States, Sectors, and Visa Types. These filters are independent of each other. The idea is to compare between the mean wage of multiple states, visa types or sectors. Our visualization offers unique ability to compare between sectors and visa types and states. This type of free interaction can help in analyzing the correlation between any two categories extensively.



- The visualization helps to understand the trend of any sector or state or visa with the mean wage of all sectors over the years This also narrates the story of how a particular sector or state or visa has performed with respect to the mean wage trend.

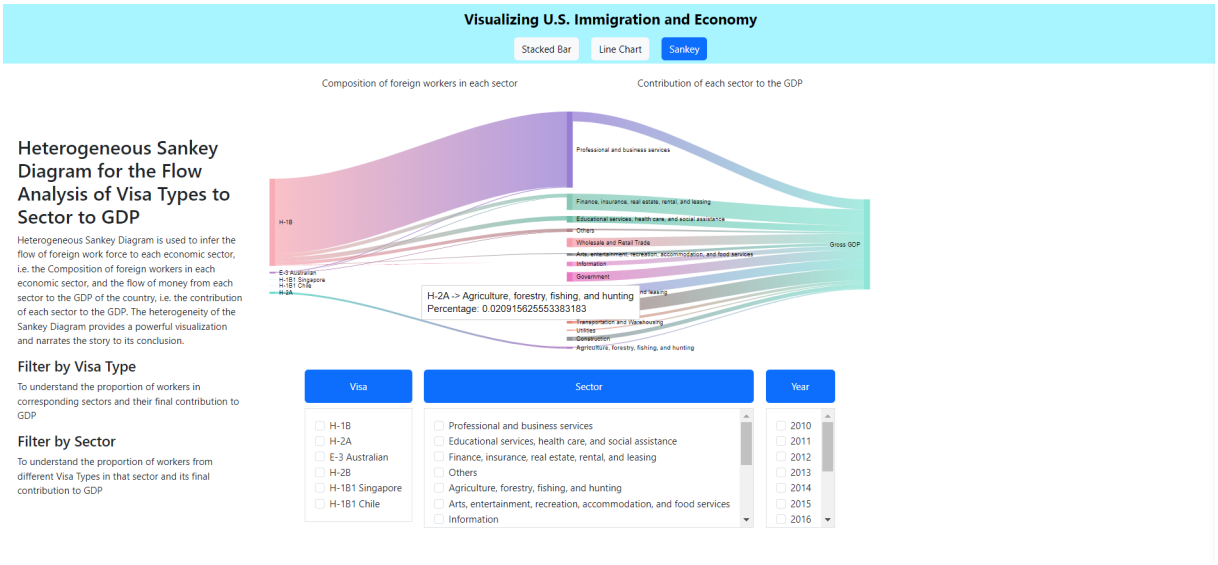


2.3 Heterogeneous Sankey Diagram for the Flow Analysis of Visa Types to Sector to GDP

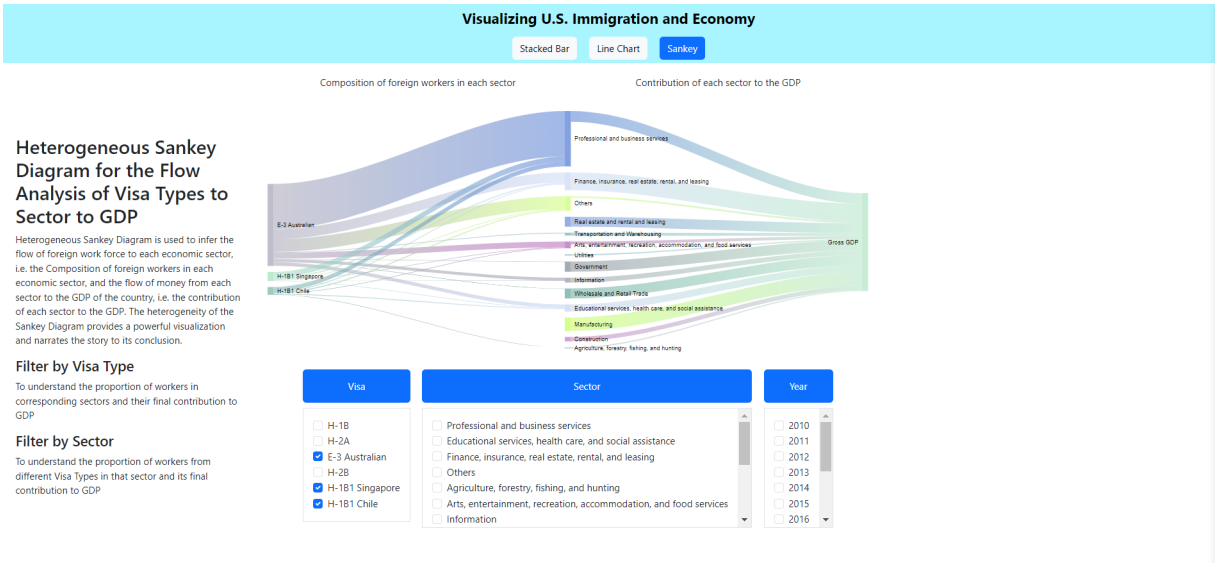
A Sankey Diagram generally represents a flow of resources, energy, or information within a system. It is a type of flow diagram that depicts the magnitude of flows between nodes, typically illustrating the transfer or transformation of one type of data into another. The width of the arrows or lines in a Sankey diagram is proportional to the quantity of the flow, making it a powerful tool for visualizing complex processes and understanding the distribution or transformation of resources.

In our visualization, we have used a heterogeneous Sankey Diagram to infer the flow of foreign work force to each economic sector, i.e. the Composition of foreign workers in each economic sector, and the flow of money from each sector to the GDP of the country, i.e. the contribution of each sector to the GDP. The heterogeneity of the Sankey Diagram provides a powerful visualization and narrates the story to its conclusion. As of now, we have understood the different inflow of foreign workers from different regions depending on their country of origin and the visa types over the period of 2010 - 2020. We also analyzed the trend of mean wage for these foreign workers over the same period of time. Now, in this visualization, we will understand the dynamics of this immigration to the GDP of the United States. For this visualization, both LCA Data and GDP Data were used.

- Users will land on the default Sankey Diagram on choosing the Sankey Tab on our website to view the flow for the year 2020. On the left most side, we visualize the major categories of visas by which the foreign workers travel to the States. In the middle, we see the major economical sectors such as Professional and business services, Educational services, health care, and social assistance, Finance, insurance, real estate, rental, and leasing, etc.
- User will be able to filter on Visa Type and Sectors for a certain year. Users will also be able to change the year and analyze the entire flow for the corresponding year.



- User will be able to filter multiple visa types and visualize all the sectors where workers from the corresponding visa types work and at the same time the contribution of those relevant sectors to the GDP. This in a broad sense gives a complete idea as to which visa types gives the best return to the GDP as compared to the composition in a sector.



- User will also be able to filter multiple sector types and visualize all the workers from different visa type who contribute to the sector and the sector's contribution to the GDP for a specific year.
- Users will also be able to move through different years to analysze any specific year's complete flow.

3 Design Decisions

We aimed the project to be of grand scale to capture every details of the Labour Market, Economic Sectors, and the Economy. Keeping in mind the large amount of data and the effectiveness of the visualizations, we made multiple design choices which can be broadly categorized into two parts -

3.1 Design Decisions for Data Processing

We had majorly 3 sources of data.

- Data regarding the number of foreign workers from each region (Used for Time-lapse Stacked Bar Chart Visualization) - NIV Data
- Data representing every foreign worker and their corresponding visa type, sector of work, and wage (Used for both Multi Line Plot and Heterogeneous Sankey Diagram) - LCA Data
- GDP Data (Used for Heterogeneous Sankey Diagram) - GDP Data

Now we will discuss the design decisions for each of these dataset.

3.1.1 NIV Data

The data consisted of all countries from major 6 regions - Asia, North America, Europe, Oceania, Africa, and South America. It consisted of all Visa types. As, we are concentrating on major work force, we finalized the following types of visas to be considered -

- H-1B

- H-2A
- H-4
- H-2B
- E-2
- O-1
- E-1
- O-2
- E-3
- O-3

We had 10 years of data which was a total of 24MB and the data was cleaned to remove redundancies and inconsistencies in the visa names. We manually cleaned the data for every year to make the visa names and region names consistent and then it was usable.

3.1.2 GDP Data

This data encompasses all sectors contribution to the GDP over the years. For our project, we kept the data from 2010 to 2020. We cleaned out many sub sectors and kept the major 16 sectors -

- Professional and business services
- Finance, insurance, real estate, rental, and leasing
- Real estate and rental and leasing
- Educational services, health care, and social assistance
- Arts, entertainment, recreation, accommodation, and food services
- Information
- Government
- Utilities
- Wholesale and Retail Trade
- Agriculture, forestry, fishing, and hunting
- Construction
- Mining

- Manufacturing
- Transportation and Warehousing
- Others

3.1.3 LCA Data

This data was the holy grail of this project, representing every foreign worker and their corresponding visa type, sector of work, and wage for every year. As discussed above, we had the above mentioned visa types for the period of 10 years, from 2010 to 2020. The total data was of 2.4GB with the H-1B data being 1.6GB in itself. There were multiple inconsistencies and redundancies in the data. Some of the most important design decisions were as follows.

- **Standardizing Visa Names:** We had to standardize the visa names by maintaining a standard - <Character - rest numbers/characters>
- **Standardizing names of fields:** While cleaning and processing the data, each data had to be manually checked and verified and then process because of different field names such as 'PW_UNIT_1' was mentioned as 'PW_UNIT' or 'pw_unit' in some datasets. A lot of fields had inconsistent names similar to this. We had to manually remove these inconsistencies and redundancies. This had to be performed for 4 large datasets for 10 years of data.
- **Mapping SOC Code to Economic Sectors:** We were provided with SOC codes in these datasets which we mapped manually into 16 major sectors as per the government website. We researched on plausible options to merge multiple sectors to make it coherent for our project. The final mapping is as below -
  - 11-2000: Professional and business services
  - 11-3030: Finance, insurance, real estate, rental, and leasing
  - 11-9140: Real estate and rental and leasing
  - 13-1000: Professional and business services
  - 13-2000: Finance, insurance, real estate, rental, and leasing
  - 15-0000: Professional and business services
  - 17-0000: Professional and business services
  - 19-0000: Professional and business services
  - 21-0000: Educational services, health care, and social assistance
  - 23-0000: Professional and business services
  - 25-0000: Educational services, health care, and social assistance
  - 27-1000: Arts, entertainment, recreation, accommodation, and food services

  - 27-2000: Arts, entertainment, recreation, accommodation, and food services
  - 27-3000: Information
  - 27-4000: Information
  - 29-0000: Educational services, health care, and social assistance
  - 31-0000: Educational services, health care, and social assistance
  - 33-0000: Government
  - 35-0000: Arts, entertainment, recreation, accommodation, and food services
  - 37-0000: Utilities
  - 39-3000: Arts, entertainment, recreation, accommodation, and food services
  - 41-3020: Finance, insurance, real estate, rental, and leasing
  - 41-3030: Finance, insurance, real estate, rental, and leasing
  - 41-4000: Wholesale and Retail Trade
  - 41-9020: Real estate and rental and leasing
  - 43-0000: Government
  - 43-9020: Information
  - 43-9040: Finance, insurance, real estate, rental, and leasing
  - 45-0000: Agriculture, forestry, fishing, and hunting
  - 47-1000: Construction
  - 47-2000: Construction
  - 47-3000: Construction
  - 47-4000: Construction
  - 47-5000: Mining
  - 49-0000: Manufacturing
  - 51-0000: Manufacturing
  - 53-0000: Transportation and Warehousing

- **Standardizing Wage:** To calculate the wages, we converted all wages to yearly. The value of the columns also were different, hence we manually cleaned the data and converted all type of units such as hourly, weekly, biweekly, monthly to yearly as a standard unit. All analysis has been done on yearly wage information thereafter.

- **Data Pre-processing for faster visualization:** With all the manual data cleaning and standardization, we finally pre-process the data into smaller datasets to efficiently utilize them during real-time visualization.

3.2 Design Choices for Visualization

Given our complex dataset, there are a plethora of other intuitive visualizations possible. However, owing to the lack of time, we chose to use simple yet effective visualizations like the following. We chose to make all of our visualizations interactive for prolonged user engagement, and the range of possible interactions differ as explained below. (Cognitive principle of element interactivity).

3.2.1 Stacked Bar Chart Time-Lapse

Our Stacked bar chart time-lapse has 3 components of interaction: 2 filters and one time-lapse player/selector. Our bar chart operates in 2 exclusive modes: filtering by country and filtering by visa. This limits the dimensionality of the visualization and also limits the choices available to the user (Cognitive principle of range of choices).

When the country filter is toggled, the stacked bar chart shows the population of people who immigrated in a particular year shown by the timelapse. As the timelapse progresses, the bar chart shows the immigration trend from each region over the years. We chose to show the regions on the Y-Axis in sorted order according to their total population (Cognitive principle of Hierarchy and Organization)

When the visa filter is toggled, the stacked bar chart similarly shows the population of each type of workers originating from a particular region in that particular year. We again chose to show the regions on the Y-Axis in sorted order according to their total population (Cognitive principle of Hierarchy and Organization)

3.2.2 Line Chart

Our line chart employs simple design choices. We chose diverse bright colors to illustrate the various plots (Gestalt's law of Similarity). Our plotting connects the discrete dots across the chart (Gestalt's law of Connectedness). Below the plot, we show various interactive filters, which can be used to select a specific State, Visa or a Sector for which the data is to be shown. Our design choice here is to apply these filters independently (without boolean composition) so that the user can easily visualize the correlation between two or more variables (Gestalt's law of Simplicity).

3.2.3 Heterogenous Sankey Diagram

Our Sankey Diagram mainly shows 2 aspects: how foreign workers of each type contributed to each sector, and how each sector has contributed to the GDP. More width in the former implies a greater contribution by workers of that particular type.(Cognitive principle of Visual Cues). We choose to visualize these 2 aspects together because economically this is an intuitive way to study the impact of immigration. (Cognitive principle of reasoning). Again, in the case of Sankey, we allow filtering on subsets of the data. Here, we choose to limit the choices available to the user for the "year" variable in order to minimize cognitive load (Cognitive principle of range of choices)

4 Evaluation

We evaluated our project by asking some peers to visit our interactive website and collecting subjective feedback on their ease of interpretation and usage. After this phase, we made a couple of improvements:

- In our stacked bar chart time-lapse, initially our stacked bars were ordered in a sorted order. This meant that over the years, if one category surpassed the other, it would be moved to the left of the stacked bar. However, we received feedback that this made it hard to interpret what was happenning. So instead, we heuristically determined an order of categories, and fixed that order for all years. This made it easier for users to understand which categories were increasing or decreasing.
- In our Sankey diagram, initially we allowed users to select multiple years in the filter. However, owing to the complexity of our data, we received feedback that this made it complex to interpret the visualization. Therefore we limited this filter so that the user can only select one year at a time.
- We also made UI tweaks across the website based on peer feedback as well as our own experiments.

5 Improvements

Given enough time, we would like to make the following improvements:

- Our data sources are limited to those reported by the Department of Labor (H-1B, H-2A, H-2B etc.) and State Department. However, there are other types of workers, although very limited in numbers in the U.S. Furthermore, illegal immigrant population is also significant and they contribute a non-trivial percentage to the economy. Given enough time, we could collect data on these sources to paint a more accurate picture of the U.S. economy.
- Replace randomized coloring with a more suitable color generation mechanism. Currently, we use randomly generated colors with a given brightness value. However, colors generated in this way are not ideal for effective visualizations. Perhaps a deterministic color generation based on a palette could be more suitable for our application.

6 Tools and Libraries

We used the following tools and libraries in our project:

- React Framework: For creating the frontend template of our website.
- Bootstrap library: For styling the various elements on our website.

- ChartJS: For visualizing the stack bar chart timelapse.
- React Chart: For visualizing the Line chart
- Google charts: For visualizing the Sankey Diagram.
- FastAPI (Python library): For creating the backend interface.
- Pandas: For all preprocessing, storing, parsing and retrieving stored datasets in the backend.
- Numpy: For numerical computations in Python