Week 11, Exercise 11.3

DSC520-T303: Statistics for Data Science (Dr. Chase Denton)

Javier Corpus

March 1, 2025

Table of Contents

Analysis of Job Postings and H-1B Visas approvals in United States

Introduction

In an increasingly globalized economy, the dynamics of the labor market are evolving, particularly in sectors that rely on specialized skills and expertise. One significant aspect of this evolution is the role of H-1B visas in the workforce landscape in the United States. The H-1B visa program allows U.S. employers to temporarily employ foreign workers in specialty occupations, which often require advanced degrees and specialized knowledge. As the demand for skilled labor continues to rise, understanding the relationship between job postings and H-1B visa sponsorship becomes crucial for policymakers, employers, and job seekers alike.

This research aims to analyze job postings across the tech industry to quantify the prevalence of H-1B visa sponsorship and to identify trends and patterns that may inform future workforce strategies. By leveraging data science techniques, we can extract meaningful insights from large datasets of job postings, enabling us to answer critical questions: What percentage of job postings are associated with H-1B sponsorship? Which employer is most reliant on H-1B workers? How do these trends vary by geographic region?

The importance of this research goes beyond mere statistics; it has implications for labor market policies, immigration reform, and the strategic planning of organizations. As businesses navigate the complexities of hiring in a competitive environment, understanding the role of H-1B visa holders can help them make informed decisions about talent acquisition and workforce development.

Problem statement

One key problem is predicting the likelihood of a job posting offering H-1B visa sponsorship based on factors such as job description, salary, state, and employer name. This analysis can help job seekers identify which companies are more likely to sponsor visas, enabling them to make informed decisions in their job search. Additionally, we can explore trends in salary across different states and industries, providing insights into regional job markets and compensation practices. It will also be helpful to have a better understanding of the jobs with most visas approved and the ones with most visas denied. Overall, these datasets can facilitate a better understanding of employment opportunities for international candidates.

This research addresses a data science problem by utilizing analytical methods to explore the intersection of job postings and H-1B visa sponsorship and uncover trends and insights.

Research questions

1. How do the rates of H-1B visa sponsorship vary across different geographic regions in the United States?
2. Which industries or sectors show the highest demand for a particular job?
3. What job roles are most associated with H-1B visas?
4. What trends can be observed over time in the number of H-1B visas petitioned?
5. What demographic characteristics (e.g., gender, nationality) are most common among H-1B visa holders?
6. What are the salary ranges offered in different geographic regions in the United States?
7. What role do large tech companies play in the overall landscape of H-1B visa sponsorship?
8. Which employers submit the larger number of H-1B visas requests?

Methodology

Initial review of the datasets

For this research, all three datasets were reviewed to determine the structure they have, the kind of data, names of columns, which columns and rows are relevant or irrelevant for this purpose.

Data Cleaning and Preprocessing.

After the initial review, the datasets need to be cleaned and preprocessed. Job titles need to be standardized because they might be different, like “software developer”, “software engineer”, or “.NET developer”. All the geography information needs to be normalized, because the name of the states could be long, like “California”, or just use the abbreviation, like “CA”. The name of the employers needs to be normalized as well because they could be written in a different way (e.g. “Walmart” vs “Walmart, Inc”).

Missing values need to be handled. If it is not relevant for this research, those values can be ignored. Otherwise, we need to define the best way to fill those values (e.g. If a salary is missing, we can use an average from all salaries for the same job title and the same state).

Data Integration

After all the datasets are clean, they need to be integrated using a common key, like a standardized job title and state. If helpful, new columns will be derived based on the existing ones.

Exploratory Data Analysis (EDA)

The data needs to be explored to identify patters, like which job titles are more likely to receive approval for an H-1B visa, or what is the average salary for certain position in a particular state. Visualizations can be used to better understand trends, like the number of jobs posting per employer or per job title.

Modeling and Analysis

A statistical analysis can be conducted to find significant differences between approved or rejected H-1B visas, and a predictive model can be built to forecast the outcome of a visa petition based on the employer and the job title.

Interpretation of the results

Results of the model need to be summarized and interpreted in order to give recommendations to job applicants, employers, or policymakers alike.

How will this approach address the problem.

With the information of these datasets, we will get:

- A better understanding of the job market. This will help identify which jobs are most likely to be associated with H-1B visas.
- Trends in H-1B visa approvals, showing which jobs are more likely to have approved visa requests, as well as demographic information about the applicants.
- By comparing similar jobs, this research could show disparities in salaries based on the same region.
- Determine which geographical areas are more likely to offer visa sponsorships.
- By creating visualizations, it will be easier to analyze data and make informed decisions.

Data

Summary of datasets.

Dataset name	Link
Salary Prediction	https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor
H1B LCA Disclosure Data (2020-2024)	https://www.kaggle.com/datasets/zongaobian/h1b-lca-disclosure-data-2020-2024
H1B visa statistics	https://www.kaggle.com/datasets/konradb/h1b-visa-statistics

Salary Prediction
The dataset contains job postings from Glassdoor.com from 2017-2018 with the following features: job title, salary estimate, job description, rating, company name, location, headquarters, size, founded, type of ownership, industry, sector revenue and competitors.

This dataset contains 742 records, 27 columns.

By "The Devastator", retrieved on February 2025 from https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor.

H1B LCA Disclosure Data (2020-2024)
This dataset provides a comprehensive record of Labor Condition Application (LCA) disclosures for H1B visa petitions filed with the U.S. Department of Labor (DOL) from 2020 to 2024.

This dataset contains 2,760,563 records, 96 columns.

By Zongao Bian, retrieved on February 2025 from https://www.kaggle.com/datasets/zongaobian/h1b-lca-disclosure-data-2020-2024.

H1B visa statistics
This dataset includes data on H-1B lottery registrations, selections and petitions from 2021 to 2024.

This dataset contains 1.8M records distributed in five files with 56 columns each.

By Konrad Banachewicz, retrieved on February 2025 from https://www.kaggle.com/datasets/konradb/h1b-visa-statistics.

Required Packages

- **dplyr** – Provides a set of functions for data manipulation, like selecting, filtering and summarizing data. It also provides functions to join different datasets.
- **ggplot** – This package allows the creation of visuals, making it easier to understand complex data.
- **stats** – Includes functions for statistical tests and modeling.
- **scales** – Used to format amounts as currency.
- **lubridate** – Used to process data fields.

Plots and Tables Needs

Plots

1. Bar Charts to show the percentage of job postings with H-1B sponsorship by employer, the H-1B sponsorship by geographic region, and different demographic characteristics on the visa applicants.
2. Line Graphs to show trends over time to depict the number of job postings offering H-1B sponsorship, highlighting any significant trends or fluctuations.
3. Box Plots to show the distribution of salaries for H-1B sponsored jobs, highlighting medians, quartiles, and outliers.
4. Pie Charts to show the percentage of visa petition statuses.

Tables

1. Summary Statistics: A table summarizing key statistics (mean, median, mode, standard deviation) for salaries for H-1B sponsored job postings.
2. Industry Breakdown: A table listing the number of job postings, percentage of H-1B sponsorship, and average salary for each industry.
3. Geographic Distribution: A table showing the number of H-1B sponsored job postings by state or region, along with corresponding average salaries.
4. Job Role Frequency: A table detailing the most common job roles associated with H-1B sponsorship, including the number of postings and average salaries for each role.

```r
# Importing the required libraries
library(dplyr, warn.conflicts = FALSE)
library(scales)
library(ggplot2)
library(lubridate, warn.conflicts = FALSE)
```

Analysis

All the required data is imported from CSV files into data frames.

Importing the data

```r
# List of salaries
df_salaries_full <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/salary_data_cleaned.csv")

# Status of visa requests, employer name
df_LCA_full <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/Combined_LCA_Disclosure_Data_FY2020_to_FY2024.csv")

# Demographic/Geographic data
df_TRK_2021 <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/TRK_13139_FY2021.csv")
df_TRK_2022 <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/TRK_13139_FY2022.csv")
df_TRK_2023 <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/TRK_13139_FY2023.csv")
df_TRK_2024_multi <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/TRK_13139_FY2024_multi_reg.csv")
df_TRK_2024_single <- read.csv("/Users/javier/Desktop/Datasets/All
Datasets/TRK_13139_FY2024_single_reg.csv")
```

Combining all the demographic/geographic data frames into a single one.

```r
df_demographic_full <- rbind(df_TRK_2021, df_TRK_2022, df_TRK_2023,
df_TRK_2024_multi, df_TRK_2024_single)
```

Cleaning up the data

Since not all the columns are relevant for this research, only a subset of this data will be used.

Selecting only the relevant columns from the list of salaries, ignoring invalid values and hourly salaries.

```
df_salaries <- df_salaries_full %>%
  filter(Type.of.ownership != -1 & hourly == 0) %>%
  select(Job.Title, Type.of.ownership,
         Industry, min_salary, max_salary, avg_salary,
         company_txt, job_state)
```

Selecting only the relevant columns from the combined LCA disclosure data, ignoring data from other countries and visas different than H-1B.

```
df_LCA <- df_LCA_full %>%
  select(CASE_STATUS, RECEIVED_DATE, VISA_CLASS, JOB_TITLE,
         SOC_TITLE, EMPLOYER_NAME, EMPLOYER_STATE, EMPLOYER_COUNTRY,
         WAGE_RATE_OF_PAY_FROM, WAGE_RATE_OF_PAY_TO,
         WAGE_UNIT_OF_PAY, PREVAILING_WAGE) %>%
  filter(EMPLOYER_COUNTRY == "UNITED STATES OF AMERICA" & VISA_CLASS == "H-
1B")
```

Selecting only the relevant columns from the demographic data, ignoring invalid data.

```
df_demographic <- df_demographic_full %>%
  select(country_of_birth, ben_year_of_birth,
         gender, employer_name, state, lottery_year,
         FIRST_DECISION, WORKSITE_STATE,
         ED_LEVEL_DEFINITION, BEN_PFIELD_OF_STUDY,
         BEN_COMP_PAID) %>%
  filter(country_of_birth != "(b)(3) (b)(6) (b)(7)(c)")
```

Removing data frames that are no longer required.

```
rm(df_salaries_full, df_LCA_full, df_TRK_2021, df_TRK_2022, df_TRK_2023,
   df_TRK_2024_multi, df_TRK_2024_single, df_demographic_full)
```

Normalizing job titles in the salaries data frame.

```r
df_salaries <- df_salaries %>%
  mutate(Job.Title = case_when(
    grepl("Data Scientist", Job.Title, ignore.case = TRUE) ~ "Data Scientist",
    grepl("Machine Learning", Job.Title, ignore.case = TRUE) ~ "Machine Learning Engineer",
    grepl("Scientist", Job.Title, ignore.case = TRUE) ~ "Research Scientist",
    grepl("Data Engineer", Job.Title, ignore.case = TRUE) ~ "Data Engineer",
    grepl("Data Modeler", Job.Title, ignore.case = TRUE) ~ "Data Modeler",
    grepl("Data Analyst", Job.Title, ignore.case = TRUE) ~ "Data Analyst",
    grepl("Data Analytics", Job.Title, ignore.case = TRUE) ~ "Data Science Analyst",
    grepl("Product Engineer", Job.Title, ignore.case = TRUE) ~ "Product Engineer",
    grepl("Director", Job.Title, ignore.case = TRUE) ~ "Director",
    grepl("Project Manager", Job.Title, ignore.case = TRUE) ~ "Project Manager",
    grepl("Analytics Manager", Job.Title, ignore.case = TRUE) ~ "Analytics Manager",
    grepl("Consultant", Job.Title, ignore.case = TRUE) ~ "Consultant - Analytics",
    grepl("Data Engineer", Job.Title, ignore.case = TRUE) ~ "Data Science Engineer",
    grepl("Senior Spark", Job.Title, ignore.case = TRUE) ~ "Senior Data Science Systems Engineer",
    TRUE ~ Job.Title
  ))
```

Normalizing job titles in the LCA Disclosure data frame.

```r
df_LCA <- df_LCA %>%
  mutate(JOB_TITLE = case_when(
    grepl("Software Engineer", JOB_TITLE, ignore.case = TRUE) ~ "Software Engineer",
    grepl("Software Developer", JOB_TITLE, ignore.case = TRUE) ~ "Software Engineer",
    grepl("Assistant Professor", JOB_TITLE, ignore.case = TRUE) ~ "Assistant Professor",
    grepl("Manager", JOB_TITLE, ignore.case = TRUE) ~ "Manager",
    grepl("System Analyst", JOB_TITLE, ignore.case = TRUE) ~ "System Analyst",
    grepl("Architect", JOB_TITLE, ignore.case = TRUE) ~ "Architect",
    grepl("Accountant", JOB_TITLE, ignore.case = TRUE) ~ "Accountant",
    grepl("Data Scientist", JOB_TITLE, ignore.case = TRUE) ~ "Data Scientist",
    grepl("Data Analyst", JOB_TITLE, ignore.case = TRUE) ~ "Data Analyst",
    grepl("Data Analytics", JOB_TITLE, ignore.case = TRUE) ~ "Data Science
Analyst",
    grepl("Product Engineer", JOB_TITLE, ignore.case = TRUE) ~ "Product
Engineer",
    grepl("Director", JOB_TITLE, ignore.case = TRUE) ~ "Director",
    grepl("Analytics Manager", JOB_TITLE, ignore.case = TRUE) ~ "Analytics
Manager",
    grepl("Consultant", JOB_TITLE, ignore.case = TRUE) ~ "Consultant -
Analytics",
    grepl("Data Engineer", JOB_TITLE, ignore.case = TRUE) ~ "Data Science
Engineer",
    grepl("Senior Spark", JOB_TITLE, ignore.case = TRUE) ~ "Senior Data
Science Systems Engineer",
    TRUE ~ JOB_TITLE
  ))
```

Converting the dates in the LCA data frame to date format (YYYY-MM-DD)

```
df_LCA$RECEIVED_DATE <- ymd(df_LCA$RECEIVED_DATE)
```

Final Datasets Configuration

Salaries:

```
str(df_salaries)

## 'data.frame':    717 obs. of  8 variables:
##  $ Job.Title        : chr  "Data Scientist" "Data Scientist" "Data
Scientist" "Data Scientist" ...
##  $ Type.of.ownership: chr  "Company - Private" "Other Organization"
"Company - Private" "Government" ...
##  $ Industry         : chr  "Aerospace & Defense" "Health Care Services &
Hospitals" "Security Services" "Energy" ...
##  $ min_salary       : int  53 63 80 56 86 71 54 86 38 120 ...
##  $ max_salary       : int  91 112 90 97 143 119 93 142 84 160 ...
##  $ avg_salary       : num  72 87.5 85 76.5 114.5 ...
##  $ company_txt      : chr  "Tecolote Research\n" "University of Maryland
Medical System\n" "KnowBe4\n" "PNNL\n" ...
##  $ job_state        : chr  " NM" " MD" " FL" " WA" ...
```

LCA Disclosure:

```
str(df_LCA)

## 'data.frame':    3470809 obs. of  12 variables:
##  $ CASE_STATUS          : chr  "Certified" "Certified" "Certified"
"Certified" ...
```

```
##  $ RECEIVED_DATE        : Date, format: "2019-09-25" "2019-09-25" ...
##  $ VISA_CLASS           : chr  "H-1B" "H-1B" "H-1B" "H-1B" ...
##  $ JOB_TITLE            : chr  "APPLICATION ENGINEER, OMS [15-1199.02]"
"BI DEVELOPER II" "QUALITY ENGINEER" "Software Engineer" ...
##  $ SOC_TITLE            : chr  "COMPUTER OCCUPATIONS, ALL OTHER" "SOFTWARE
DEVELOPERS, APPLICATIONS" "MECHANICAL ENGINEERS" "SOFTWARE DEVELOPERS,
APPLICATIONS" ...
##  $ EMPLOYER_NAME        : chr  "JO-ANN STORES, INC." "DENKEN SOLUTIONS
INC." "EPITEC, INC." "SYSTEMS TECHNOLOGY GROUP, INC." ...
##  $ EMPLOYER_STATE       : chr  "OH" "CA" "MI" "MI" ...
##  $ EMPLOYER_COUNTRY     : chr  "UNITED STATES OF AMERICA" "UNITED STATES
OF AMERICA" "UNITED STATES OF AMERICA" "UNITED STATES OF AMERICA" ...
##  $ WAGE_RATE_OF_PAY_FROM: num  100000 38.6 43.5 57.7 75000 ...
##  $ WAGE_RATE_OF_PAY_TO  : num  NA 38.6 NA 57.7 NA ...
##  $ WAGE_UNIT_OF_PAY     : chr  "Year" "Hour" "Hour" "Hour" ...
##  $ PREVAILING_WAGE      : num  95118 39 39 53 65333 ...
```

Demographic:

```
str(df_demographic)
```

```
## 'data.frame':    1804147 obs. of  11 variables:
##  $ country_of_birth   : chr  "CHN" "IND" "CAN" "PAK" ...
##  $ ben_year_of_birth  : chr  "1981" "1994" "1988" "1993" ...
##  $ gender             : chr  "male" "male" "male" "male" ...
##  $ employer_name      : chr  "D&R I.P. Law Firm" "ITTECHNICA INC" "Tesla,
Inc." "Crorama Inc." ...
##  $ state              : chr  "CA" "TX" "CA" "CA" ...
##  $ lottery_year       : chr  "2021" "2021" "2021" "2021" ...
##  $ FIRST_DECISION     : chr  "" "" "Approved" "" ...
##  $ WORKSITE_STATE     : chr  "" "" "CA" "" ...
##  $ ED_LEVEL_DEFINITION: chr  "" "" "MASTER'S DEGREE" "" ...
##  $ BEN_PFIELD_OF_STUDY: chr  "" "" "COMPUTER ENGINEERING" "" ...
##  $ BEN_COMP_PAID      : chr  "" "" "125000" "" ...
```

Handling Non-Self-Evident Information

One of the mains aspects that is not self-evident, is how many of the jobs posted do not sponsor H-1B visas. This has to be inferred based on the list of jobs posted in Glassdoor (df_salaries), and see which ones of those are missing from the LCA Disclose dataset (df_LCA). However, there is no simple way to do this because the job descriptions could be different on each company, despite having a similar description.

For example, one employer could post a job for a “Software Engineer”, other for a “Software Developer”, or “Application Developer”. There are close to 500,000 jobs listed that would have to be normalized.

Exploratory and Analytical Methods

Descriptive and demographic analyses will be used for preliminary insights. For a more in-depth exploration, time series analysis and predictive analysis techniques can be applied to these datasets.

Segmenting and Organizing Data

The data will be segmented or grouped in different categories:

Jobs
This will help getting better answers to questions like which are the jobs with the highest salaries, and which are the jobs with most visas denied.

Employers
This will help us to understand which employers have the highest demand of H-1B visas, and how many of these visas requests get approved or denied.

Demographic
This will show which country has the most petitioners for H-1B visas.

Summarizations

Showing the top 10 jobs published by Glassdoor

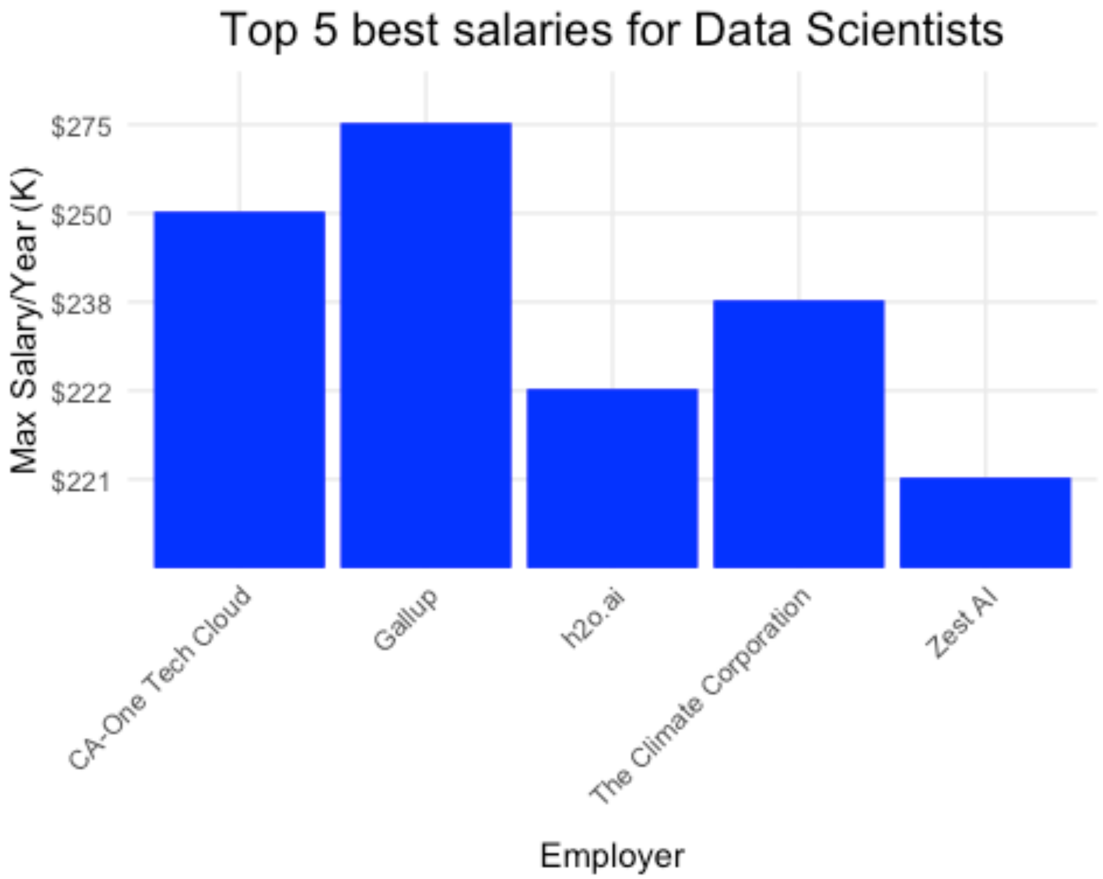
```
job_counts_Glassdoor <- df_salaries %>%
  group_by(Job.Title) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

top_10_jobs_posts <- head(job_counts_Glassdoor, 10)
print(top_10_jobs_posts)

## # A tibble: 10 × 2
##    Job.Title                 count
##    <chr>                     <int>
##  1 Data Scientist              279
##  2 Research Scientist          126
##  3 Data Engineer               119
##  4 Data Analyst                 95
##  5 Machine Learning Engineer    22
##  6 Analytics Manager            11
##  7 Director                     11
##  8 Consultant - Analytics        7
##  9 Data Modeler                  5
## 10 Data Science Analyst          5
```

Getting the top 5 salaries for the Data Scientist positions from the Glassdoor data se to be shown in a bar graph.

```r
df_data_scientist_salary <- df_salaries %>%
  select(Job.Title, company_txt, max_salary) %>%
  filter(Job.Title == "Data Scientist") %>%
  group_by(.) %>%
  arrange(desc(max_salary))

df_data_scientist_salary <- df_data_scientist_salary %>%
  distinct(Job.Title, company_txt, max_salary)

df_data_scientist_salary$max_salary <-
dollar(df_data_scientist_salary$max_salary)

top_5_data_scientist_salaries <- head(df_data_scientist_salary, 5)
```

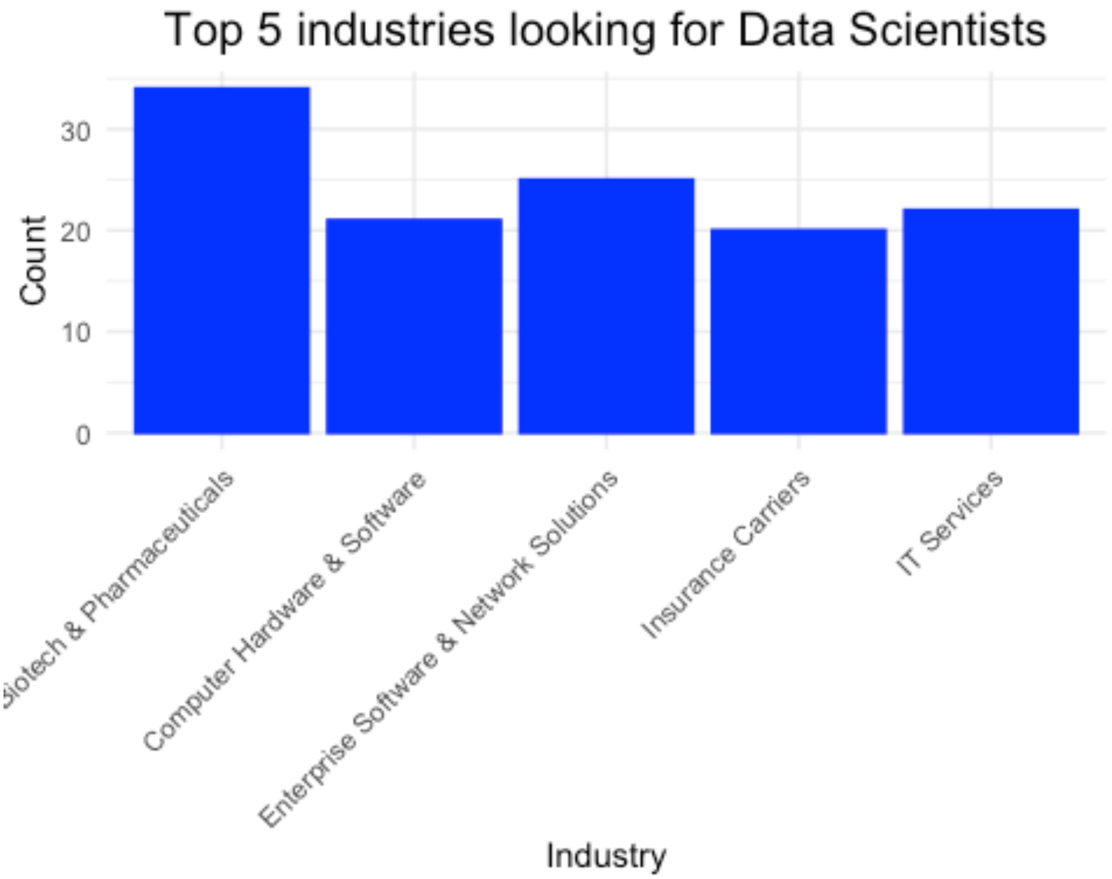
Top 5 best salaries for Data Scientists.

```r
ggplot(top_5_data_scientist_salaries, aes(x = company_txt, y = max_salary)) +
  geom_bar(stat = "identity", color = "blue", fill = "blue") +
  labs(title = "Top 5 best salaries for Data Scientists",
       x = "Employer",
       y = "Max Salary/Year (K)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size = 15, hjust = 0.5)
        )
```

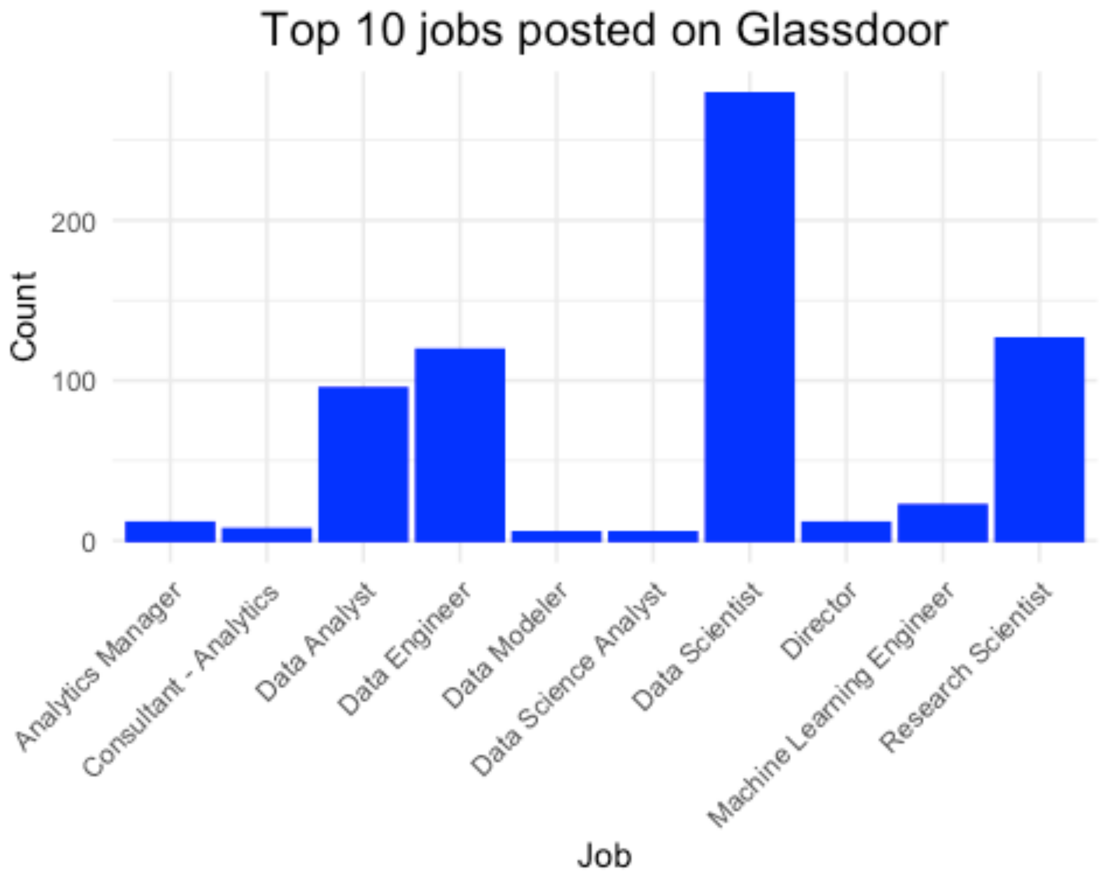


Industries with the highest demand for Data Scientists:

```r
df_industries <- df_salaries %>%
  select(Job.Title, Industry) %>%
  filter(Job.Title == "Data Scientist") %>%
  group_by(Job.Title, Industry) %>%
  summarize(count = n(), .groups = "drop") %>%
  arrange(desc(count))

top_5_industries <- head(df_industries, 5)

ggplot(top_5_industries, aes(x = Industry, y = count)) +
  geom_bar(stat = "identity", color = "blue", fill = "blue") +
  labs(title = "Top 5 industries looking for Data Scientists",
       x = "Industry",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size = 15, hjust = 0.5)
        )
```



```
ggplot(top_10_jobs_posts, aes(x = Job.Title, y = count)) +
  geom_bar(stat = "identity", color = "blue", fill = "blue") +
  labs(title = "Top 10 jobs posted on Glassdoor",
       x = "Job",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size = 15, hjust = 0.5)
        )
```



Showing the top 10 jobs in the LCA Disclosure data frame.

```
job_counts_LCA <- df_LCA %>%
  group_by(JOB_TITLE) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

head(job_counts_LCA, 10)

## # A tibble: 10 × 2
##    JOB_TITLE                 count
##    <chr>                     <int>
##  1 Software Engineer        689629
##  2 Manager                  353510
##  3 Architect                127740
##  4 Consultant - Analytics   118907
##  5 Assistant Professor       59798
##  6 Director                  55786
##  7 Data Science Engineer     47390
##  8 Data Scientist            39827
##  9 Accountant                29616
## 10 Data Analyst              26099
```

Showing the top 10 employers requesting H-1B visas, and the status of the petitions.

```
petitions_summary_table <- df_LCA %>%
  group_by(EMPLOYER_NAME) %>%
  summarize(
    total = n(),
    certified = sum(CASE_STATUS == "Certified"),
    denied = sum(CASE_STATUS == "Denied"),
    cert_withdrawn = sum(CASE_STATUS == "Certified - Withdrawn"),
    withdrawn = sum(CASE_STATUS == "Withdrawn")
  )

employers <- petitions_summary_table %>%
  group_by(EMPLOYER_NAME) %>%
  arrange(desc(total))

head(employers, 10)

## # A tibble: 10 × 6
## # Groups:   EMPLOYER_NAME [10]
##    EMPLOYER_NAME                     total certified denied cert_withdrawn
withdrawn
##    <chr>                             <int>     <int>  <int>          <int>
<int>
##  1 "COGNIZANT TECHNOLOGY SOLUTI… 94223     92361      2            563
1297
##  2 "Ernst & Young U.S. LLP"      61155     60205     74            117
759
##  3 "Google LLC"                  61020     59131    115            972
802
##  4 "Amazon.com Services LLC"     58437     56561     31           1251
594
##  5 "Microsoft Corporation"       54382     54379      3              0
0
##  6 "INFOSYS LIMITED"             42865     42643     10             15
197
##  7 "TATA CONSULTANCY SERVICES L… 38985     38804      9             62
110
##  8 "Apple Inc."                  24822     23967     29            586
240
##  9 "Accenture LLP"               23961     23545     19             63
334
## 10 "Intel Corporation "          22602     22570     19              0
13
```

Salaries

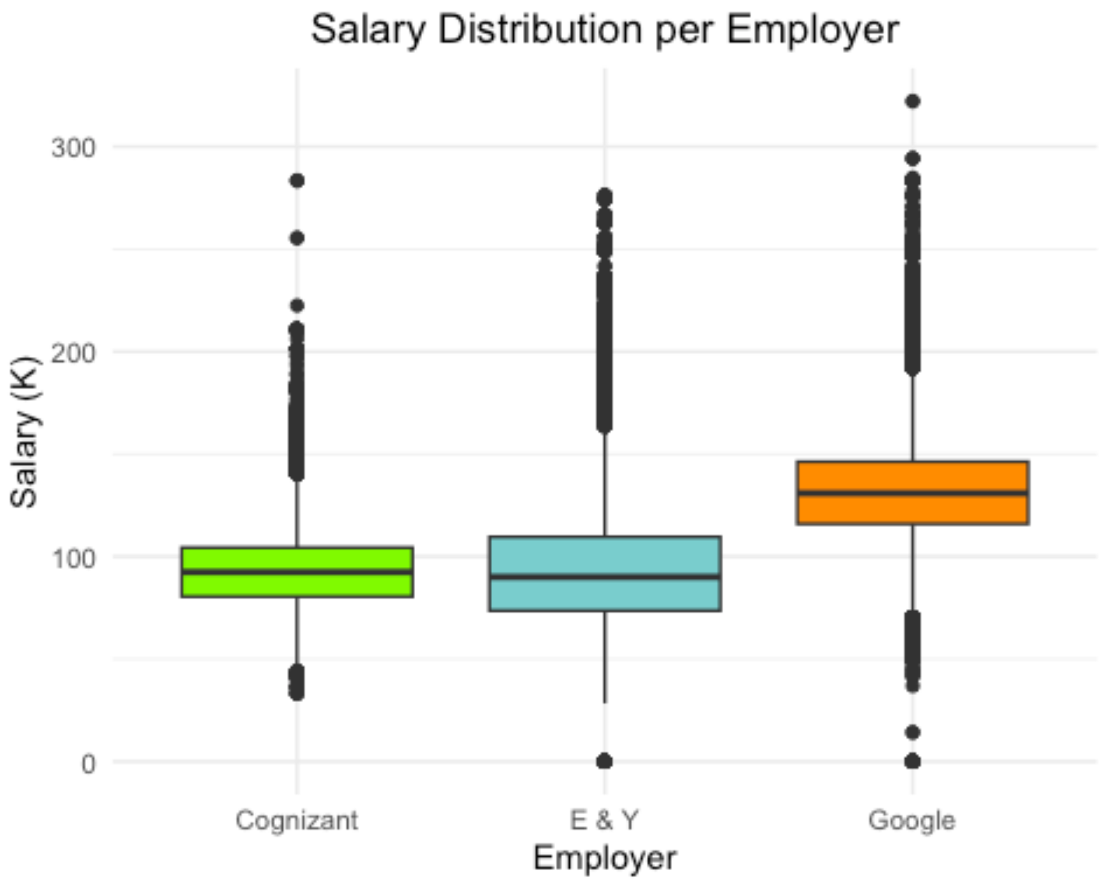
Showing the top 10 jobs listed in **Glassdoor** with the highest average annual salary.

```
highest_salaries <- df_salaries %>%
  group_by(Job.Title) %>%
  select(Job.Title, max_salary, min_salary, avg_salary) %>%
  summarize(
    max_salary = max(max_salary),
    min_salary = min(min_salary),
    avg_salary = mean(avg_salary)) %>%
  arrange(desc(max_salary))

# Formatting the salary as currency
highest_salaries$max_salary <- dollar(highest_salaries$max_salary * 1000)
highest_salaries$min_salary <- dollar(highest_salaries$min_salary * 1000)
highest_salaries$avg_salary <- dollar(highest_salaries$avg_salary * 1000)

top_10_highest_salaries <- head(highest_salaries, 10)
print(top_10_highest_salaries)

## # A tibble: 10 × 4
##    Job.Title                      max_salary min_salary avg_salary
##    <chr>                          <chr>      <chr>      <chr>
##  1 Director                       $306,000   $39,000    $173,500
##  2 Machine Learning Engineer      $289,000   $61,000    $126,432
##  3 Data Scientist                 $275,000   $15,000    $117,565
##  4 Data Science Manager           $272,000   $95,000    $158,833
##  5 Research Scientist             $231,000   $29,000    $95,516
##  6 Data Engineer                  $228,000   $42,000    $105,403
##  7 Senior Quantitative Analyst    $228,000   $118,000   $173,000
##  8 Data Science Engineer - Mobile $208,000   $116,000   $162,000
##  9 Data Analyst                   $178,000   $20,000    $65,016
## 10 VP, Data Science               $166,000   $83,000    $124,500
```

Preparing the data of the top 3 employers to do a box plot.

```
df_salaries_boxplot <- df_LCA %>%
  select(EMPLOYER_NAME, PREVAILING_WAGE) %>%
  filter(EMPLOYER_NAME == "COGNIZANT TECHNOLOGY SOLUTIONS US CORP" |
           EMPLOYER_NAME == "Ernst & Young U.S. LLP" |
           EMPLOYER_NAME == "Google LLC") %>%

  # Renaming employers to they fit better in the plot
  mutate(EMPLOYER_NAME =
           ifelse(EMPLOYER_NAME == "COGNIZANT TECHNOLOGY SOLUTIONS US CORP",
"Cognizant",
              ifelse(EMPLOYER_NAME == "Ernst & Young U.S. LLP", "E & Y",
```

```r
                 ifelse(EMPLOYER_NAME == "Google LLC",
"Google",EMPLOYER_NAME))))

# Diving the salary by 1,000 to make it easier to plot
df_salaries_boxplot$PREVAILING_WAGE <- df_salaries_boxplot$PREVAILING_WAGE /
1000

ggplot(df_salaries_boxplot, aes(x = EMPLOYER_NAME, y = PREVAILING_WAGE)) +
  geom_boxplot(fill = c("chartreuse", "darkslategray3", "darkorange")) +
  labs(title = "Salary Distribution per Employer",
       x = "Employer",
       y = "Salary (K)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```



Demographics

Top 10 countries with most beneficiaries of H-1B visas, and the genders of the beneficiaries.

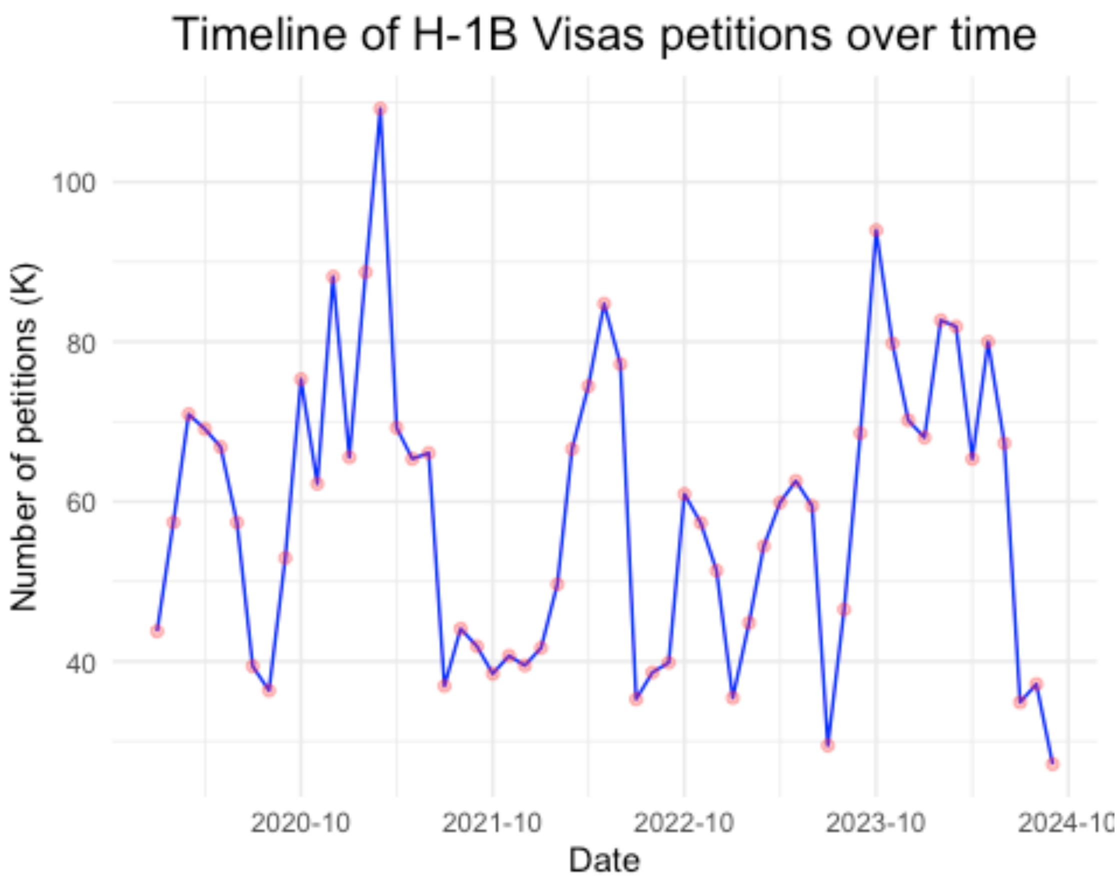
```
beneficiaries <- df_demographic %>%
  group_by(country_of_birth) %>%
  summarize(
    count = n(),
    males = sum(gender == "male"),
    females = sum(gender == "female")
    ) %>%
  arrange(desc(count))

head(beneficiaries, 10)
```

```
## # A tibble: 10 × 4
##    country_of_birth   count  males females
##    <chr>              <int>  <int>   <int>
##  1 IND              1388598 992311  396287
##  2 CHN               147953  75525   72428
##  3 CAN                16241  11176    5065
##  4 MEX                15087  11749    3338
##  5 PHL                14843   6287    8556
##  6 KOR                14795   8530    6265
##  7 PAK                14210  12136    2074
##  8 TWN                12668   7146    5522
##  9 NPL                12662   9571    3091
## 10 BRA                10665   7377    3288
```

Visualization and Reporting

Number of H-1B visas requested from 2020 to 2024.

```
visas_requested <- df_LCA %>%
  filter(RECEIVED_DATE > "2020-01-01") %>%
  group_by(month_year = floor_date(RECEIVED_DATE, "month")) %>%
  summarize(total_petitions = n())

visas_requested$total_petitions <- visas_requested$total_petitions / 1000

ggplot(visas_requested, aes(x = month_year, y = total_petitions)) +
  geom_line(color = "blue") +
  geom_point(color = "red", alpha = 0.3) +
  labs(title = "Timeline of H-1B Visas petitions over time",
       x = "Date",
       y = "Number of petitions (K)") +
  theme_minimal() +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "12 months") +
    theme(
          plot.title = element_text(size = 15, hjust = 0.5),
      )
```

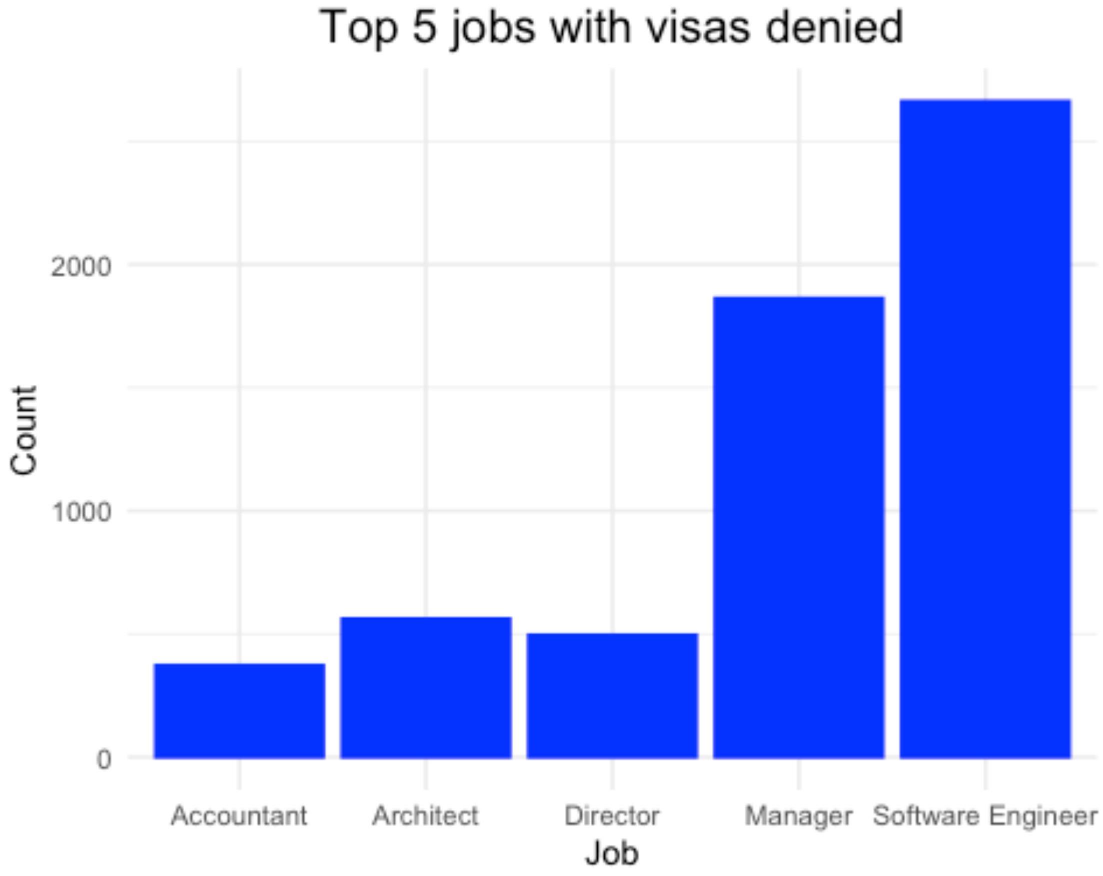


Jobs with the highest number of denied visas.

Getting the data to be used in the bar graph.

```
df_denied_visas <- df_LCA %>%
  select(CASE_STATUS, JOB_TITLE) %>%
  filter(CASE_STATUS == "Denied") %>%
  group_by(JOB_TITLE) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

top_5_denied_visas <- head(df_denied_visas, 5)
```

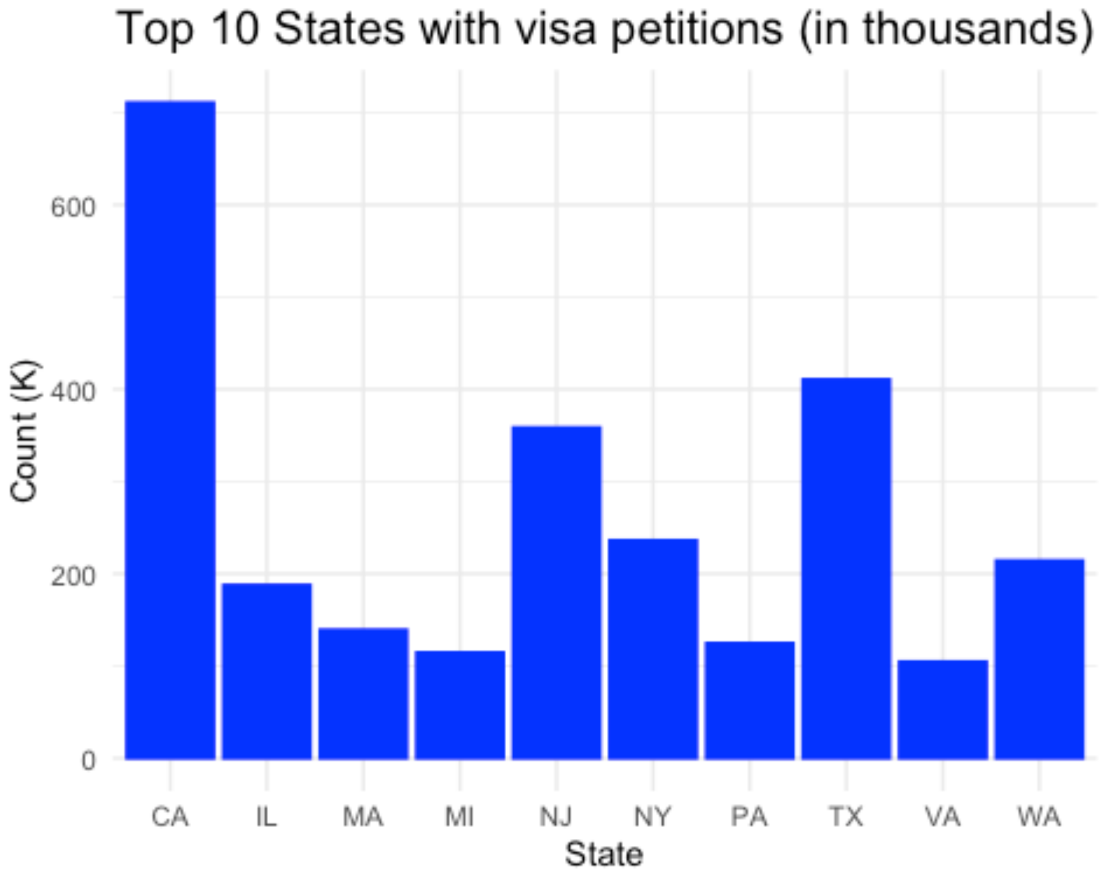
```r
ggplot(top_5_denied_visas, aes(x = JOB_TITLE, y = count)) +
  geom_bar(stat = "identity", color = "blue", fill = "blue") +
  labs(title = "Top 5 jobs with visas denied",
       x = "Job",
       y = "Count") +
  theme_minimal() +
  theme(
        plot.title = element_text(size = 15, hjust = 0.5)
        )
```



States with most H-1B visa petitions

```r
petitions_by_state <- df_LCA %>%
  select(EMPLOYER_STATE) %>%
  group_by(EMPLOYER_STATE) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

top_10_petitions_by_state <- head(petitions_by_state, 10)
```

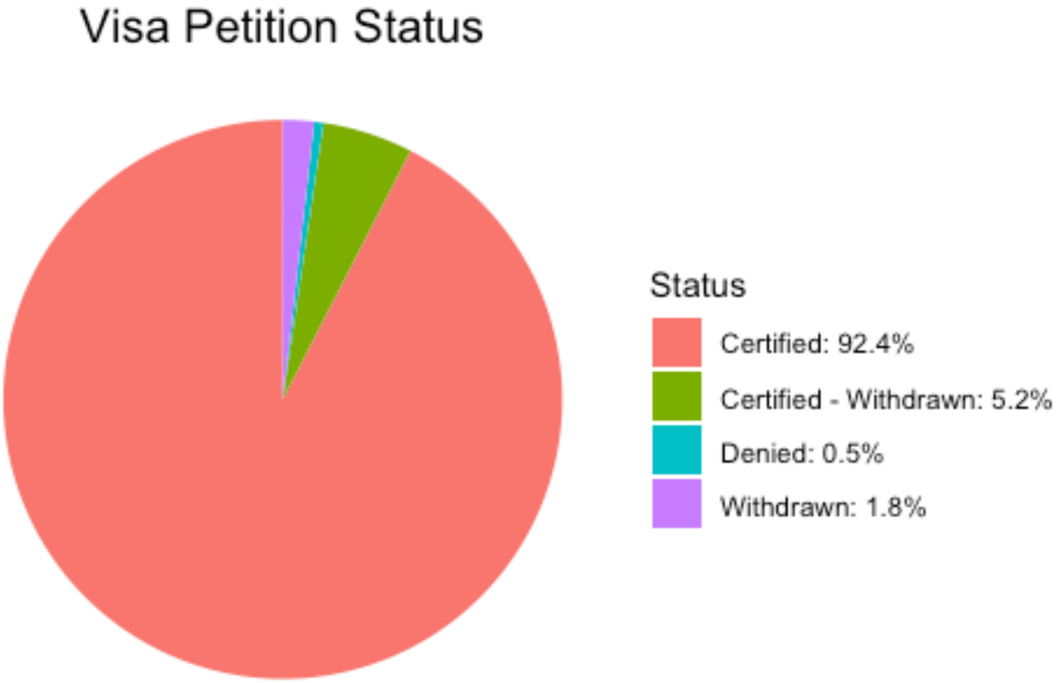
```
#Divided by 1000 to have a better scale
top_10_petitions_by_state$count <- top_10_petitions_by_state$count / 1000

ggplot(top_10_petitions_by_state, aes(x = EMPLOYER_STATE, y = count)) +
  geom_bar(stat = "identity", color = "blue", fill = "blue") +
  labs(title = "Top 10 States with visa petitions (in thousands)",
       x = "State",
       y = "Count (K)") +
  theme_minimal() +
  theme(
        plot.title = element_text(size = 15, hjust = 0.5)
        )
```



Visa petitions: Status

The following chart is a visual representation of the different visa petition statutes from 2020 to 2024, according to the LCA Disclosure data.

```r
df_visa_status <- df_LCA %>%
  select(CASE_STATUS)

# Calculating counts and percentages
status_summary <- df_visa_status %>%
  group_by(CASE_STATUS) %>%
  summarize(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

# Creating the pie chart
pie_chart <- ggplot(status_summary,
                    aes(x = "", y = percentage, fill = CASE_STATUS)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Visa Petition Status",
       fill = "Status") +
  theme_void() +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 15, hjust = 0.5)
    )

# Creating labels with percentages
legend_labels <- paste0(status_summary$CASE_STATUS, ": ",
                        round(status_summary$percentage, 1), "%")

# Adding legends to the plot
pie_chart + scale_fill_discrete(labels = legend_labels)
```



Incorporating Advanced Techniques

A linear regression model can be used to predict the outcome of visa petitions (certified or denied) by analyzing the relationship between the job title and the employer. By treating the visa petition outcome as a `dependent variable` and incorporating job titles and employers as `independent variables`, the model could identify patterns and trends that influence the likelihood of approval or denial.

This statistical approach can provide valuable insights for applicants and immigration professionals. This model would be useful as a predictive tool that can help employers and job applicants to make informed decisions based on historical data.

Further Analysis and Iterative Exploration

Further analysis and iterative exploration are needed for a comprehensive understanding of all the datasets related to jobs, employers, and visa petitions.

Given the complexity and variability in these datasets, it is necessary to employ different analytical techniques to uncover patterns and relationships. This includes conducting exploratory data analysis, creating and refining models, and testing different hypotheses to compare against findings.

Additionally, iterative exploration will allow for more accurate insights and informed decision-making regarding employment and visa processes. By continuously refining the analysis, stakeholders can be better informed about the job market and visa landscapes.

Summary

Problem Statement

The labor market is continuously evolving, especially in sectors that require specialized skills and experience. The H-1B visa plays a an important role in this, as it allows people with these skills and experience to work in United States, if they were born in other countries. However, this process is complex and costly. It is fundamental to keep up with the ever-evolving process of requesting this kind of visas, and it is extremely helpful to understand different aspects of it.

Companies benefit from this information by knowing beforehand if a job role is a suitable for an H-1B visa, and they can compare salaries of similar roles of different employers to gargantee they remain competitive.

Job seekers also benefit from this information, knowing if they current skills are useful to request an H-1B visa. They can also compare salaries among different employers and positions to make an informed decision.

Addressing the Problem Statement

Using different datasets we can get useful insights. Combining data about jobs posting, salaries, employers, locations, visa requests status, and demographic characteristics, along with statistical methods, we will have a better understanding of the trends and the probability of getting a visa request approved based on a combination of characteristics.

Analysis

After the initial Exploratory Data Analysis (EDA), the following insights were found:

Top 10 jobs posted in Glassdoor

Job	Count
Data Scientist	279
Research Scientist	126
Data Engineer	119
Data Analyst	95
Machine Learning Engineer	22
Analytics Manager	11
Director	11
Consultant - Analytics	7
Data Modeler	5
Data Science Analyst	5

Looking at the salaries offered by different employers for a **Data Scientist** position, we found these are the top 5 best salaries, and the employers offering it:

Employer	Salary (K/year)
Gallup	$275
CA-One Tech Cloud	$250
The Climate Corporation	$238
h2o ai	$222
Zest AI	$221

The following are the top 5 industries looking for Data Scientists:

Industry	Jobs Posted
Biotech & Pharmaceuticals	34
Enterprise Software & Network Solutions	25
IT Services	22
Computer Hardware & Software	21
Insurance Carriers	20

These are the top 10 jobs requesting H-1B visas

Job	H1-B requests
Software Engineer	689629
Manager	353510
Architect	127740
Consultant - Analytics	118907
Assistant Professor	59798
Director	55786
Data Science Engineer	47390
Data Scientist	39827
Accountant	29616
Data Analyst	26099

These are the top 5 jobs with the highest number of visas denied.

Job	Visas Denied
Software Engineer	2663
Manager	1863
Architect	562
Director	496
Accountant	373

The following table summarizes the top 10 employers requesting H-1B visas, and the status of those requests. `Cognizant Technology Solutions` is the employer with most visas requested, and it is also the employer with the lowest number of visas denied (only 0.0021% of visas denied).

Not shown in the table, but `HCL America Solutions` is the employer with most visas denied (19.07% of requests got denied).

Employer	Total	Certified	Denied	Cert. Withdrawn	Withdrawn
COGNIZANT TECHNOLOGY SOLUTIONS US CORP	94223	92361	2	563	1297
Ernst & Young U.S. LLP	61155	60205	74	117	759
Google LLC	61020	59131	115	972	802
Amazon.com Services LLC	58437	56561	31	1251	594
Microsoft Corporation	54382	54379	3	0	0
INFOSYS LIMITED	42865	42643	10	15	197
TATA CONSULTANCY SERVICES LIMITED	38985	38804	9	62	110
Apple Inc.	24822	23967	29	586	240
Accenture LLP	23961	23545	19	63	334
Intel Corporation	22602	22570	19	0	13

Although `Cognizant` is the top employer and `Google` comes in third, when it comes to salaries, Google has better salaries (on average) than Cognizant.

These are the top 5 countries with more beneficiaries of H-1B visas, and the gender of the beneficiaries.

Country	Total Visas	Males	Females
IND	1388598	992311	396287
CHN	147953	75525	72428
CAN	16241	11176	5065
MEX	15087	11749	3338
PHL	14843	6287	8556

These are the top 5 annual salaries with their corresponding job title:

Job	Max Salary	Min Salary	Average Salary
Director	$306,000	$39,000	$173,500
Machine Learning Engineer	$289,000	$61,000	$126,432
Data Scientist	$275,000	$15,000	$117,565
Data Science Manager	$272,000	$95,000	$158,833
Research Scientist	$231,000	$29,000	$95,516

These are the top 10 States with the highest number of visa requests.

State	Count (K)
CA	710.752
TX	410.539
NJ	358.395
NY	236.162
WA	214.249
IL	187.61
MA	139.151
PA	124.592
MI	114.497
VA	104.626

Implications and Limitations

These datasets provides valuable insights into job market trends, salary distributions, industries, and employers, as well as the demographics of applicants. It can help identify patterns in visa request statuses and inform policy decisions related to employment and immigration.

However, there are limitations to consider. The dataset may not be representative of the entire job market, as it could be biased towards certain industries or regions. Additionally, demographic information may be incomplete or inaccurate, affecting the reliability of any conclusions drawn. Finally, the dataset reflects a specific time frame, which may limit its relevance for future analysis. This is specially true with the ever-changing migration policies of the US. Laws and requeriments are contantly changing. Even under the same administration, there are policy revisions year after year. When the administration change, the policy changes become more drastic.

It is important to be informed of the latest policies, because the requirements existing today could be different in the near future.

Concluding Remarks

These datasets and the corresponding analysis serve as a valuable resource for analyzing various aspects of the job market, including salary trends, top employers sponsoring visas, and the demographics of applicants. By examining the relationships between these factors, stakeholders can obtain insights about hiring strategies, compensation ranges, and skills with a higher probability of getting the visa request approved. Additionally, understanding visa request statuses can help organizations navigate immigration processes and address potential barriers for international talent.

However, it is crucial to recognize the limitations inherent in these datasets. The sample may not fully represent the broader job market, as it could be skewed towards specific industries or types of employment. Furthermore, the demographic information provided may be incomplete or ir could be innaccurate.

To enhance a future analyses, it is recommended that these datasets be supplemented with additional data sources that capture a wider range of employment scenarios and demographic profiles. By acknowledging these limitations, researchers and policymakers can better address the complexities of employment trends and their implications for the market.