



Project title: H1B Non-Immigrant Labor Visa Data Analysis on Tableau

Course: CIS5270

Authors: Navyasree Sriramoju, Sushmitha Dandu, Suman Chauhan

Instructor: Shilpa Balan

A. Introduction:

The H-1B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply for an H1-B visa, a US employer must offer a job and petition for an H-1B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/ higher education (Masters, Ph.D.) and work in a full-time position. Under the H1B visa, any company can employ a foreign worker for up to six years. Filing of an H1B visa is not in the hands of the individuals, only the employer is allowed to file the petition for the respective employee.

The H-1B visa program has been a highly debated topic in the United States, with concerns about its impact on the American job market, wages, and the economy. The program is intended to allow U.S. employers to hire highly skilled foreign workers to fill specific jobs, but some critics argue that it is being used to replace American workers with lower-cost foreign labor.

Motivation

As an international student ourselves, our curiosity drove us to choose this data set so as to gain some insights on H1B visa system. The purpose of this project is to analyze the H-1B Non-Immigrant Labour Visa petitions dataset to gain insights into the trends and patterns in the visa application process. By analyzing this dataset, we can gain a better understanding of

the demand for skilled foreign workers in the U.S. job market and the ways in which the program is being used by employers.

The motivation for this project is twofold. Firstly, it is important to gain a better understanding of the H-1B visa program and its impact on the American job market. As the U.S. job market becomes increasingly competitive, it is important to understand how the visa program is being used by employers to fill positions and to what extent it is displacing American workers. Additionally, the program may be affecting wages, which has an impact on the overall economy. Analyzing the dataset can provide valuable insights into these issues, which can be used to inform policy decisions and improve the U.S. job market. Secondly, the project can help us identify the areas where the U.S. job market lacks skilled workers and may need improvement. The dataset contains information on the job titles and skills required by employers to fill specific positions. By analyzing this information, we can gain insights into the skills that are in demand in the U.S. job market and identify areas where there may be a shortage of skilled workers. This information can be used to inform education and training programs that can help workers acquire the skills needed to fill these positions.

Moreover, the project can also provide insights into the companies that are applying for H-1B visas and the job titles that are being filled. This information can be used to identify the sectors that are most heavily reliant on foreign workers and the skills that are in demand in those sectors. It can also provide insights into the geographic distribution of H-1B visa holders, which can inform policies aimed at attracting skilled workers to specific regions.

Overall, the project's motivation is to provide valuable insights into the H-1B visa program and its impact on the American job market. By analyzing the dataset, we can gain a better understanding of the program's usage and its impact on the economy. Additionally, the project can help identify the skills and sectors that are in demand in the U.S. job market, which can inform policies aimed at improving the job market and attracting skilled workers. Ultimately, the project aims to provide policymakers with valuable information that can be used to make informed decisions about the H-1B visa program and its impact on the American job market.

B. Data Description

Source: This dataset is taken from Kaggle data repository. Kaggle has a wide range of datasets that we can use for Visualizations.

Dataset URL: <https://www.kaggle.com/datasets/thedevastator/h-1b-non-immigrant-labour-visa>

The dataset contains over 3 million H-1B Non-Immigrant Labour Visa petitions filed with the US Department of Labour between 2011 and 2017. The dataset includes information such as the employer's name, job title, job location, wages, and the number of certified and denied petitions. The dataset also includes information on the prevailing wage for each job title and the wage offered by the employer. The dataset is available in CSV format.

Field Name	Data Description
CASE_SUBMITTED	Date on which the application was submitted.
DECISION_DATE	Date on which the last significant event or decision was recorded by the Chicago National Processing Center
EMP_NAME	Name of employer submitting labor condition application

EMP_STATE	Contact information of the Employer requesting temporary labor certification.
EMP_COUNTRY	
SOC_NAME	Occupational name associated with the SOC_CODE
FULL_TIME_POSITION	Y = Full Time Position; N = Part Time Position
PREVAILING_WAGE	Prevailing Wage for the job being requested for temporary labor conditions.
PW_UNIT	Unit of Pay. Valid values include “Hourly (H),” “Bi-weekly (BW),” “Weekly (W),” “Monthly (M),” and “Yearly (Y)”.
PW_LEVEL	Level of Prevailing Wage
WAGE	Employer’s proposed wage rate.
H-1B_DEPENDENT	Y = Employer is H-1B Dependent; N = Employer is not H-1B Dependent.
WILLFUL_VIOLATOR	Y = Employer has been previously found to be a Willful Violator; N = Employer has not been considered a Willful Violator
WORKSITE_STATE	State information of the foreign worker's intended area of employment.
WORKSITE_CITY	City information of the foreign worker's intended area of employment.
WORKSITE_POSTAL_CODE	Zip Code information of the foreign worker's intended area of employment.
CASE_STATUS*	Status associated with the last significant event or decision. Valid values include “C=Certified,” “CW=Certified-Withdrawn,” “D=Denied,” and “W=Withdrawn”.

C. Data Cleaning

Data cleaning is a critical step in the data analysis and visualization process. It ensures that the data used for analysis is accurate, complete, and consistent, leading to better insights, improved efficiency, enhanced data quality, and increased credibility.

Removing Duplicates

Pre-cleaning:

case_year	case_status	case_submitted	decision_date	emp_name	emp_city	emp_state	emp_zip	emp_country	job_title	soc_code
2017	C	3/6/2017	3/10/2017	LAKELANDS NEPHROLOGY, PA	GREENWOOD	SC	29646	USA	NEPHROLOGIST	29-1063
2017	C	3/6/2017	3/10/2017	LAKELANDS NEPHROLOGY, PA	GREENWOOD	SC	29646	USA	NEPHROLOGIST	29-1063
2017	C	3/11/2017	3/11/2017	UNIVERSITY OF IDAHO	MOSCOW	ID	83844	USA	POST DOCTORAL FELLOW	19-1013
2017	C	3/17/2017	3/13/2017	XPO SUPPLY CHAIN, INC.	HIGH POINT	NC	27265	USA	OPERATION ANALYST	15-2031
2017	C	3/10/2017	3/16/2017	C AND S WHOLESALE GROCERS, INC.	KEENE	NH	3431	USA	SR. INDUSTRIAL ENGINEER	17-2112
2017	C	8/4/2017	8/10/2017	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	Hematologist/Oncologist	29-1069
2017	C	8/4/2017	8/10/2017	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	PEDIATRICIAN	29-1065
2017	C	3/5/2017	3/9/2017	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	FAMILY MEDICINE PHYSICIAN	29-1062
2017	C	7/10/2017	7/14/2017	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	PEDIATRICIAN	29-1065
2017	C	2/22/2017	2/28/2017	NORTHERN STATE UNIVERSITY	ABERDEEN	SD	57401	USA	INSTRUCTOR OF BUSINESS ACCOUNTING	25-1011
2017	C	10/4/2016	10/11/2016	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PEDIATRICIAN	29-1065
2017	C	10/4/2016	10/11/2016	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PEDIATRICIAN	29-1065
2017	C	3/7/2017	3/13/2017	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)	29-1063
2017	C	7/10/2017	7/14/2017	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	PEDIATRICIAN	29-1065
2017	CW	12/18/2015	2/9/2017	SANFORD CLINIC	SIOUX FALLS	SD	57105	USA	NEPHROLOGIST	29-1069
2017	C	3/7/2017	3/13/2017	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)	29-1063
2017	C	6/15/2017	6/21/2017	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PODIATRIST	29-1081
2017	CW	7/17/2015	2/23/2017	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	INTERVENTIONAL CARDIOLOGIST	29-1069
2017	C	3/29/2017	4/6/2017	SAFEWAY INC.	PLEASANTON	CA	94588	USA	PHARMACY MANAGER	29-1051
2017	C	10/5/2016	10/11/2016	HEALTH CAROUSEL, LLC	CINCINNATI	OH	45206	USA	PHYSICAL THERAPIST	29-1123
2017	C	1/9/2017	1/13/2017	GRAY'S HARBOR COLLEGE	ABERDEEN	WA	98520	USA	TECHNICAL DESIGN PROGRAM INSTRUCTOR	17-2071
2017	CW	8/24/2015	3/21/2017	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK	TX	79430	USA	RESEARCH ASSOCIATE	19-4021
2017	C	12/1/2016	12/28/2016	MANAGEMENT HEALTH SYSTEMS, INC.	SUNRISE	FL	33323	USA	PHYSICAL THERAPIST	29-1123
2017	C	3/1/2017	3/7/2017	OPHTHALMOLOGY SPECIALISTS OF TEXAS PLLC	ABILENE	TX	79606	USA	SOFTWARE DEVELOPER	15-1132
2017	D	3/7/2017	2/8/2017	SIMMONS UNIVERSITY	ABILENE	TX	79606	USA	ASSISTANT PROFESSOR OF NURSING	25-1072
2017	C	6/6/2017	6/12/2017	MCMURRY UNIVERSITY	ABILENE	TX	79697	USA	ESL INSTRUCTOR	25-1199
2017	C	2/28/2017	3/6/2017	HEALTHSOUTH REHABILITATION HOSPITAL OF ABILENE, LLC	ABILENE	TX	79606	USA	PHYSICAL THERAPIST	29-1123
2017	C	11/9/2016	12/15/2016	HENDRICK MEDICAL CENTER	ABILENE	TX	79602	USA	CASE MANAGER	29-1141

Post-cleaning:

case_year	case_status	case_submitted	decision_date	emp_name	emp_city	emp_state	emp_zip	emp_country	job_title	soc_cod
2017	C	42772	42776	LAKELANDS NEPHROLOGY, PA	GREENWOOD	SC	29646	USA	NEPHROLOGIST	29-1063
2017	C	42815	42821	UNIVERSITY OF IDAHO	MOSCOW	ID	83844	USA	POST DOCTORAL FELLOW	19-1013
2017	C	42811	42817	XPO SUPPLY CHAIN, INC.	HIGH POINT	NC	27265	USA	OPERATION ANALYST	15-2031
2017	C	42804	42810	C AND S WHOLESALE GROCERS, INC.	KEENE	NH	3431	USA	SR. INDUSTRIAL ENGINEER	17-2112
2017	C	42951	42957	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	Hematologist/Oncologist	29-1069
2017	C	42799	42803	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	FAMILY MEDICINE PHYSICIAN	29-1062
2017	C	42926	42930	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	PEDIATRICIAN	29-1065
2017	C	42788	42794	NORTHERN STATE UNIVERSITY	ABERDEEN	SD	57401	USA	INSTRUCTOR OF BUSINESS ACCOUNTING	25-1011
2017	C	42647	42654	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PEDIATRICIAN	29-1065
2017	C	42801	42807	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)	29-1063
2017	CW	42356	42775	SANFORD CLINIC	SIOUX FALLS	SD	57105	USA	NEPHROLOGIST	29-1069
2017	C	42901	42907	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PODIATRIST	29-1081
2017	CW	42202	42789	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	INTERVENTIONAL CARDIOLOGIST	29-1069
2017	C	42823	42829	SAFEWAY INC.	PLEASANTON	CA	94588	USA	PHARMACY MANAGER	29-1051
2017	C	42648	42655	HEALTH CAROUSEL, LLC	CINCINNATI	OH	45206	USA	PHYSICAL THERAPIST	29-1123
2017	C	42744	42748	GRAY'S HARBOR COLLEGE	ABERDEEN	WA	98520	USA	TECHNICAL DESIGN PROGRAM INSTRUCTOR	17-2071
2017	CW	42240	42787	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK	TX	79430	USA	RESEARCH ASSOCIATE	19-4021
2017	C	42725	42732	MANAGEMENT HEALTH SYSTEMS, INC.	SUNRISE	FL	33323	USA	PHYSICAL THERAPIST	29-1123
2017	C	42795	42801	OPHTHALMOLOGY SPECIALISTS OF TEXAS PLLC	ABILENE	TX	79606	USA	SOFTWARE DEVELOPER	15-1132
2017	D	42768	42774	HARDIN - SIMMONS UNIVERSITY	ABILENE	TX	79698	USA	ASSISTANT PROFESSOR OF NURSING	25-1072
2017	C	42892	42908	MCMURRY UNIVERSITY	ABILENE	TX	79697	USA	ESL INSTRUCTOR	25-1199
2017	C	42794	42800	HEALTHSOUTH REHABILITATION HOSPITAL OF ABILENE, LLC	ABILENE	TX	79606	USA	PHYSICAL THERAPIST	29-1123
2017	C	42713	42719	HENDRICK MEDICAL CENTER	ABILENE	TX	79602	USA	CASE MANAGER	29-1141
2017	C	42801	42807	HEALTH CAROUSEL, LLC	CINCINNATI	OH	45209	USA	PHYSICAL THERAPIST	29-1123
2017	C	42942	42948	MCH INVESTMENT CORPORATION	ABILENE	TX	79605	NA	FINANCE MANAGER	411
2017	C	42979	42986	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK	TX	79430	NA	ASSISTANT PROFESSOR (PHARMACEUTICAL SCIENCES)	25-1071
2017	C	42649	42656	THE UNIVERSITY OF IOWA	IOWA CITY	IA	52242	USA	SENIOR APPLICATION DEVELOPER	15-1034

As seen in screenshot above, there are few duplicate rows in the dataset so thus we decided to remove the duplicates from the data set using excel so that we get better visualizations and more accurate results. Thus, removing the duplicates was an essential step in our data cleaning-process.

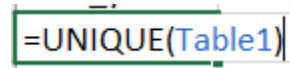
Step 1: Select the whole data which you want to clean and press CTRL+T , and create “table1”

Step 2: Click on ok

Step 3: Copy and paste all the headers wherever data cleaning is required

Step 4: Use the unique function as shown below

=UNIQUE(Select the complete data which you want to clean),

A screenshot of an Excel formula bar. The formula bar is highlighted with a green border and contains the text "=UNIQUE(Table1)". The word "Table1" is in blue, indicating it is a named range or table reference. The formula bar is part of the Excel ribbon interface, with the "Formulas" tab visible in the background.

=UNIQUE(Table1)

Removing Null Values:

Pre-cleaning:

	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	soc_name	full_time_positi	prevailing_wage	pw_unit	pw_level	wage_from	wage_to	wage_unit	work_city	work_stat	emp_h1b	emp_willf	lat
2	INTERNISTS, GENERAL	Y	\$187,200.00	Y		\$190,000.00	\$0.00	Y	ABBEVILLE SC	N	N		34.178
3	SOIL AND PLANT SCIENTISTS	Y	\$39,957.00	Y	Level I	\$47,507.00	\$0.00	Y	ABERDEEN ID	N	N		42.944
4	OPERATIONS RESEARCH ANALYSTS	Y	\$59,966.00	Y	Level I	\$65,000.00	\$0.00	Y	ABERDEEN MD	N	N		39.509
5	INDUSTRIAL ENGINEERS	Y	\$78,832.00	Y	Level II	\$86,988.15	\$0.00	Y	ABERDEEN MD	N	N		39.509
6	PHYSICIANS AND SURGEONS, ALL OTHER	Y	\$169,645.00	Y	NA	\$450,000.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
7	FAMILY AND GENERAL PRACTITIONERS	Y	\$131,581.00	Y	Level I	\$131,581.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
8	PEDIATRICIANS, GENERAL	Y	\$187,200.00	Y	NA	\$187,200.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
9	BUSINESS TEACHERS, POSTSECONDARY	Y	\$45,010.00	Y	Level I	\$51,487.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
10	PEDIATRICIANS, GENERAL	Y	\$166,982.00	Y	Level I	\$190,000.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
11	INTERNISTS, GENERAL	Y	\$41,725.00	Y	Level I	\$240,000.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
12	PHYSICIANS AND SURGEONS, ALL OTHER	Y	\$187,199.00	Y		\$187,199.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
13	PODIATRISTS	Y	\$54,330.00	Y	Level II	\$215,000.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
14	PHYSICIANS AND SURGEONS, ALL OTHER	Y	\$326,105.00	Y		\$326,105.00	\$0.00	Y	ABERDEEN SD	N	N		45.46
15	PHARMACISTS	Y	\$120,349.00	Y	Level III	\$67.50	\$0.00	H	ABERDEEN WA	N	N		46.975
16	PHYSICAL THERAPISTS	Y	\$28.66	H	Level I	\$28.66	\$28.66	H	ABERDEEN WA	N	N		46.975
17	ELECTRICAL ENGINEERS	Y	\$75,277.00	Y		\$75,277.00	\$0.00	Y	ABERDEEN WA	N	N		46.975
18	BIOLOGICAL TECHNICIANS	Y	\$29,078.00	Y	Level II	\$31,212.00	\$0.00	Y	ABILENE TX	N	N		32.448
19	PHYSICAL THERAPISTS	Y	\$32.02	H	Level I	\$40.00	\$0.00	H	ABILENE TX	Y	N		32.448
20	SOFTWARE DEVELOPERS, APPLICATIONS	Y	\$48,901.00	Y	Level I	\$60,000.00	\$0.00	Y	ABILENE TX	N	N		32.448
21	NURSING INSTRUCTORS AND TEACHERS, POSTSECOY	Y	\$60,307.00	Y	Level III	\$90,000.00	\$0.00	Y	ABILENE TX	N	N		32.448

Post-cleaning:

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Help

Power Pivot

Paste

Clipboard

Calibri

11

A

A²

B

I

U

Font

Alignment

Number

Date

\$

%

0.00

00

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Insert

Delete

Format

Cells

Editing

Analysis

Sensitivity

Comments

Share

Formulas

Data

Review

View

Help

Power Pivot

Clipboard

Font

Alignment

Number

Conditional Formatting

We used this data cleaning technique to remove Null values/blank spaces from our data set so that we are left with accurate data which will give better visualization results. The following steps were performed on several columns that had null values. For example, from our data set “pw_level” (as shown in screenshot) had a lot of null values:

Step 1: Select the column that contains Null values.

Step 2: Go to the Home Tab and click on Find & Select.

Step 3: Then click Go to Special, a new window will appear, Select Blanks and click on OK.

Step 4: The blank cells are highlighted, Right-click on any of the highlighted blank cells and select Delete.

Changing the Currency

Pre-cleaning:

	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
	job_title	soc_code	soc_name	full_time	prevailing_wage	pw_unit	pw_level	wage_from	wage_to	wage_unit	work_city	work_stat	emp_h1b	emp_willf	lat	lng
1	NEPHROL(29-1063		INTERNIST Y		187200	Y	N/A	190000		0 Y	ABBEVILLE SC	N	N		34.17817	-82.379
2	POST DOC 19-1013		SOIL AND Y		39957	Y	Level I	47507		0 Y	ABERDEEN MD	N	N		42.94408	-112.838
3	OPERATIO15-2031		OPERATIO Y		59966	Y	Level I	65000		0 Y	ABERDEEN MD	N	N		39.50956	-76.1641
4	SR. INDUS 17-2112		INDUSTRI Y		78832	Y	Level II	86988.15		0 Y	ABERDEEN MD	N	N		39.50956	-76.1641
5	HEMATOL 29-1069		PHYSICIAN Y		169645	Y	NA	450000		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
6	FAMILY M 29-1062		FAMILY A Y		131581	Y	Level I	131581		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
7	PEDIATRIC 29-1065		PEDIATRIC Y		187200	Y	NA	187200		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
8	INSTRUCT 25-1011		BUSINESS Y		45010	Y	Level I	51487		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
9	PEDIATRIC 29-1065		PEDIATRIC Y		166982	Y	Level I	190000		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
10	HOSPITAL 29-1063		INTERNIST Y		41725	Y	Level I	240000		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
11	PEDIATRIC 29-1065		PEDIATRIC Y		187200	Y	NA	187200		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
12	NEPHROL 29-1069		PHYSICIAN Y		187199	Y	N/A	187199		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
13	HOSPITAL 29-1063		INTERNIST Y		41725	Y	Level I	240000		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
14	PODIATRI 29-1081		PODIATRI Y		54330	Y	Level II	215000		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
15	INTERVEN 29-1069		PHYSICIAN Y		326105	Y	N/A	326105		0 Y	ABERDEEN SD	N	N		45.4647	-98.4865
16	PHARMAC 29-1051		PHARMACY Y		120349	Y	Level III	67.5		0 H	ABERDEEN WA	N	N		46.97537	-123.816
17	PHYSICAL 29-1123		PHYSICAL Y		28.66	H	Level I	28.66		28.66 H	ABERDEEN WA	N	N		46.97537	-123.816
18	TECHNICA 17-2071		ELECTRIC Y		75277	Y	N/A	75277		0 Y	ABERDEEN WA	N	N		46.97537	-123.816
19	RESEARCH 19-4021		BIOLOGIC Y		29078	Y	Level II	31212		0 Y	ABILENE TX	N	N		32.44874	-99.7331
20	PHYSICAL 29-1123		PHYSICAL Y		32.02	H	Level I	40		0 H	ABILENE TX	Y	N		32.44874	-99.7331

Prevailing Wage column doesn't have any currency, In order to add the currency type

Go to Home Page --> Select the Column to add the currency---> click on General ----> Select the currency you want to add -->click on OK

You get the required output

Post-cleaning:

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Help

Power Pivot

Calibri

11

</

Removing the Extra Spaces

Pre-cleaning:

1	2	3	4	5	6	7	8
emp_name	emp_city	emp_state	emp_zip	emp_coun	job_title	soc_code	soc_name
LAKELANDS NEPHROLOGY, PA	GREENWOOD SC	29646 USA			NEPHROLOGIST	29-1063	INTERNISTS, GENE
UNIVERSITY OF IDAHO	MOSCOW ID	83844 USA			POST DOCTORAL FELLOW	19-1013	SOIL AND PLANT S
XPO SUPPLY CHAIN, INC.	HIGH POINT NC	27265 USA			OPERATION ANALYST	15-2031	OPERATIONS RESE
C AND S WHOLESALE GROCERS, INC.	KEENE NH	3431 USA			SR. INDUSTRIAL ENGINEER	17-2112	INDUSTRIAL ENGI
SANFORD CLINIC	SIOUX FALLS SD	57117 NA			HEMATOLOGIST/ONCOLOGIST	29-1069	PHYSICIANS AND S
SANFORD CLINIC	SIOUX FALLS SD	57117 USA			FAMILY MEDICINE PHYSICIAN	29-1062	FAMILY AND GENE
SANFORD CLINIC	SIOUX FALLS SD	57117 NA			PEDIATRICIAN	29-1065	PEDIATRICIANS, GI
NORTHERN STATE UNIVERSITY	ABERDEEN SD	57401 USA			INSTRUCTOR OF BUSINESS ACCOUNTING	25-1011	BUSINESS TEACHE
AVERA ST. LUKE'S HOSPITAL	ABERDEEN SD	57401 USA			PEDIATRICIAN	29-1065	PEDIATRICIANS, GI
AVERA ST. LUKE'S HOSPITAL	ABERDEEN SD	57401 USA			HOSPITALIST (INTERNIST)	29-1063	INTERNISTS, GENE
SANFORD CLINIC	SIOUX FALLS SD	57117 NA			PEDIATRICIAN	29-1065	PEDIATRICIANS, GI
SANFORD CLINIC	SIOUX FALLS SD	57105 USA			NEPHROLOGIST	29-1069	PHYSICIANS AND S
AVERA ST. LUKE'S HOSPITAL	ABERDEEN SD	57401 USA			HOSPITALIST (INTERNIST)	29-1063	INTERNISTS, GENE
AVERA ST. LUKE'S HOSPITAL	ABERDEEN SD	57401 USA			PODIATRIST	29-1081	PODIATRISTS
SANFORD CLINIC	SIOUX FALLS SD	57117 USA			INTERVENTIONAL RADIOLOGIST	29-1069	PHYSICIANS AND S
SAFEWAY INC.	PLEASANTON CA	94588 USA			PHARMACY MANAGER	29-1051	PHARMACISTS
HEALTH CAROUSEL, LLC	CINCINNATI OH	45206 USA			PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPI
GRAY'S HARBOR COLLEGE	ABERDEEN WA	98520 USA			TECHNICAL DESIGN PROGRAM INSTRUCTOR	17-2071	ELECTRICAL ENGI
TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK TX	79430 USA			RESEARCH ASSOCIATE	19-4021	BIOLOGICAL TECH
MANAGEMENT HEALTH SYSTEMS, INC.	SUNRISE FL	33323 USA			PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPI
OPHTHALMOLOGY SPECIALISTS OF TEXAS PLLC	ABILENE TX	79606 USA			SOFTWARE DEVELOPER	15-1132	SOFTWARE DEVELO
HARDIN - SIMMONS UNIVERSITY	ABILENE TX	79698 USA			ASSISTANT PROFESSOR OF NURSING	25-1072	NURSING INSTRU
MCMURRY UNIVERSITY	ABILENE TX	79697 USA			ESL INSTRUCTOR	25-1199	POSTSECONDARY
HEALTHSOUTH REHABILITATION HOSPITAL OF ABILENE, LLC	ABILENE TX	79606 USA			PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPI
HENDRICK MEDICAL CENTER	ABILENE TX	79602 USA			CASE MANAGER	29-1141	REGISTERED NURS
HEALTH CAROUSEL, LLC	CINCINNATI OH	45209 USA			PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPI
MCR INVESTMENT CORPORATION	ABILENE TX	79605 NA			FINANCE MANAGER	Nov-31	FINANCIAL MANA
TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK TX	79430 NA			ASSISTANT PROFESSOR (PHARMACEUTICAL SCIENCES)	25-1071	HEALTH SPECIALTI

To remove the extra spaces from our data set, the column “employee name” has extra spaces in the rows. We used “TRIM” Function in excel. We followed the below steps to remove the spaces.

Step 1: Select the column that has extra space. In this case “emp_name”

Step 2: Add a new column next to emp_name and name that newly added column as clean_emp_name.

Step 3: By using the TRIM (CELL NUMBER) function, we can remove spaces.

and apply it to all the rows Then click Go to Special, a new window will appear, Select Blanks and click on OK.

Step 4: Select all the rows from that column and apply the trim function.

emp_name	Clean emp_name
LAKELANDS NEPHROLOGY, PA	=TRIM(E2)

Post-Cleaning:

File Home Insert Page Layout Formulas Data Review View Help Power Pivot										Comments Share	
<div>Clipboard Font Alignment Number Styles Cells Editing Analysis Sensitivity</div>											
F1 Clean_emp_name											
	A	B	C	D	E	F	G	H	I	J	
1	case_year	case_status	case_submitted	decision_date	emp_name	Clean_emp_name	emp_city	emp_state	emp_zip	emp_coun	job_title
2	2017	C	2/6/2017	2/10/2017	LAKELANDS NEPHROLOGY, PA	LAKELANDS NEPHROLOGY, PA	GREENWOOD	SC	29646	USA	NEPHROLOGIST
3	2017	C	3/21/2017	3/27/2017	UNIVERSITY OF IDAHO	UNIVERSITY OF IDAHO	MOSCOW	ID	83844	USA	POST DOCTORAL FELLOW
4	2017	C	3/17/2017	3/23/2017	XPO SUPPLY CHAIN, INC.	XPO SUPPLY CHAIN, INC.	HIGH POINT	NC	27265	USA	OPERATION ANALYST
5	2017	C	3/10/2017	3/16/2017	C AND S WHOLESALE GROCERS, INC.	C AND S WHOLESALE GROCERS, INC.	KEENE	NH	3431	USA	SR. INDUSTRIAL ENGINEER
6	2017	C	8/4/2017	8/10/2017	SANFORD CLINIC	SANFORD CLINIC	SIOUX FALLS	SD	57117	NA	HEMATOLOGIST/ONCOLOGIST
7	2017	C	3/5/2017	3/9/2017	SANFORD CLINIC	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	FAMILY MEDICINE PHYSICIAN
8	2017	C	7/10/2017	7/14/2017	SANFORD CLINIC	SANFORD CLINIC	SIOUX FALLS	SD	57117	NA	PEDIATRICIAN
9	2017	C	2/22/2017	2/28/2017	NORTHERN STATE UNIVERSITY	NORTHERN STATE UNIVERSITY	ABERDEEN	SD	57401	USA	INSTRUCTOR OF BUSINESS
10	2017	C	10/4/2016	10/11/2016	AVERA ST. LUKE'S HOSPITAL	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PEDIATRICIAN
11	2017	C	3/7/2017	3/13/2017	AVERA ST. LUKE'S HOSPITAL	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)
12	2017	C	7/10/2017	7/14/2017	SANFORD CLINIC	SANFORD CLINIC	SIOUX FALLS	SD	57117	NA	PEDIATRICIAN
13	2017	CW	12/18/2015	2/9/2017	SANFORD CLINIC	SANFORD CLINIC	SIOUX FALLS	SD	57105	USA	NEPHROLOGIST
14	2017	C	3/7/2017	3/13/2017	AVERA ST. LUKE'S HOSPITAL	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)
15	2017	C	6/15/2017	6/21/2017	AVERA ST. LUKE'S HOSPITAL	AVERA ST. LUKE'S HOSPITAL	ABERDEEN	SD	57401	USA	PODIATRIST
16	2017	CW	7/17/2015	2/23/2017	SANFORD CLINIC	SANFORD CLINIC	SIOUX FALLS	SD	57117	USA	INTERVENTIONAL CARDIOLOGIST
17	2017	C	3/29/2017	4/4/2017	SAFEWAY INC.	SAFEWAY INC.	PLEASANTON	CA	94588	USA	PHARMACY MANAGER
18	2017	C	10/5/2016	10/12/2016	HEALTH CAROUSEL, LLC	HEALTH CAROUSEL, LLC	CINCINNATI	OH	45206	USA	PHYSICAL THERAPIST
19	2017	C	1/9/2017	1/13/2017	GRAY'S HARBOR COLLEGE	GRAY'S HARBOR COLLEGE	ABERDEEN	WA	98520	USA	TECHNICAL DESIGN PROJECT MANAGER
20	2017	CW	8/24/2015	2/21/2017	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK	TX	79430	USA	RESEARCH ASSOCIATE
21	2017	C	12/21/2016	12/28/2016	MANAGEMENT HEALTH SYSTEMS, INC.	MANAGEMENT HEALTH SYSTEMS, INC.	SUNRISE	FL	33323	USA	PHYSICAL THERAPIST
22	2017	C	3/1/2017	3/7/2017	OPHTHALMOLOGY SPECIALISTS OF TEXAS PLLC	OPHTHALMOLOGY SPECIALISTS OF TEXAS PLLC	ABILENE	TX	79606	USA	SOFTWARE DEVELOPER
23	2017	D	2/2/2017	2/8/2017	HARDIN - SIMMONS UNIVERSITY	HARDIN - SIMMONS UNIVERSITY	ABILENE	TX	79698	USA	ASSISTANT PROFESSOR
24	2017	C	6/6/2017	6/12/2017	MCMURRY UNIVERSITY	MCMURRY UNIVERSITY	ABILENE	TX	79697	USA	ESL INSTRUCTOR
25	2017	C	2/28/2017	3/6/2017	HEALTHSOUTH REHABILITATION HOSPITAL OF ABILENE, LLC	HEALTHSOUTH REHABILITATION HOSPITAL OF ABILENE, LLC	ABILENE	TX	79606	USA	PHYSICAL THERAPIST
26	2017	C	12/9/2016	12/15/2016	HENDRICK MEDICAL CENTER	HENDRICK MEDICAL CENTER	ABILENE	TX	79602	USA	CASE MANAGER
27	2017	C	3/7/2017	3/13/2017	HEALTH CAROUSEL, LLC	HEALTH CAROUSEL, LLC	CINCINNATI	OH	45209	USA	PHYSICAL THERAPIST
28	2017	C	7/26/2017	8/1/2017	MCR INVESTMENT CORPORATION	MCR INVESTMENT CORPORATION	ABILENE	TX	79605	NA	FINANCE MANAGER
29	2017	C	9/1/2017	9/8/2017	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER	LUBBOCK	TX	79430	NA	ASSISTANT PROFESSOR

Splitting the column

Pre-Cleaning:

FileHomeInsertPage LayoutFormulasDataReviewViewHelpPower Pivot														CommentsShare			
Clipboard		Font		Alignment		Number		Styles		Cells		Editing		Analysis		Sensitivity	
CutCopyPasteFormat Painter		Calibri11A+AB I U		Wrap TextMerge & Center		General\$ % & #		Conditional FormattingFormat as TableCell Styles		InsertDelete Format		AutoSumFillSort & FilterFind & SelectClear		Analyze Data		Sensitivity	
L1																	
	G	H	I	J	K	L	M	N	O	P	Q	R	S	T			
	emp_city	emp_state	emp_zip	emp_coun	job_title	soc_code	soc_name	full_time	prevailing_pw_unit	pw_level	wage_fror	wage_to	wage_unit				
1	emp_city	emp_state	emp_zip	emp_coun	job_title	soc_code	soc_name	full_time	prevailing_pw_unit	pw_level	wage_fror	wage_to	wage_unit				
2	GREENWOOD	SC	29646	USA	NEPHROLOGIST	29-1063	INTERNS, GENERAL	Y	187200 Y	N/A	190000	0	Y				
3	MOSCOW	ID	83844	USA	POST DOCTORAL FELLOW	19-1013	SOIL AND PLANT SCIENTISTS	Y	39957 Y	Level I	47507	0	Y				
4	HIGH POINT	NC	27265	USA	OPERATION ANALYST	15-2031	OPERATIONS RESEARCH ANALYSTS	Y	59966 Y	Level I	65000	0	Y				
5	KEENE	NH	3431	USA	SR. INDUSTRIAL ENGINEER	17-2112	INDUSTRIAL ENGINEERS	Y	78832 Y	Level II	86988.15	0	Y				
6	SIoux FALLS	SD	57117	NA	HEMATOLOGIST/ONCOLOGIST	29-1069	PHYSICIANS AND SURGEONS, ALL OTHER	Y	169645 Y	NA	450000	0	Y				
7	SIoux FALLS	SD	57117	USA	FAMILY MEDICINE PHYSICIAN	29-1062	FAMILY AND GENERAL PRACTITIONERS	Y	131581 Y	Level I	131581	0	Y				
8	SIoux FALLS	SD	57117	NA	PEDIATRICIAN	29-1065	PEDIATRICIANS, GENERAL	Y	187200 Y	NA	187200	0	Y				
9	ABERDEEN	SD	57401	USA	INSTRUCTOR OF BUSINESS ACCOUNTING	25-1011	BUSINESS TEACHERS, POSTSECONDARY	Y	45010 Y	Level I	51487	0	Y				
10	ABERDEEN	SD	57401	USA	PEDIATRICIAN	29-1065	PEDIATRICIANS, GENERAL	Y	166982 Y	Level I	190000	0	Y				
11	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)	29-1063	INTERNS, GENERAL	Y	41725 Y	Level I	240000	0	Y				
12	SIoux FALLS	SD	57117	NA	PEDIATRICIAN	29-1065	PEDIATRICIANS, GENERAL	Y	187200 Y	NA	187200	0	Y				
13	SIoux FALLS	SD	57105	USA	NEPHROLOGIST	29-1069	PHYSICIANS AND SURGEONS, ALL OTHER	Y	187199 Y	N/A	187199	0	Y				
14	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNIST)	25-1063	INTERNS, GENERAL	Y	41725 Y	Level I	240000	0	Y				
15	ABERDEEN	SD	57401	USA	PODIATRIST	29-1081	PODIATRISTS	Y	54330 Y	Level II	215000	0	Y				
16	SIoux FALLS	SD	57117	USA	INTERVENTIONAL CARDIOLOGIST	29-1069	PHYSICIANS AND SURGEONS, ALL OTHER	Y	326105 Y	N/A	326105	0	Y				
17	PLEASANTON	CA	94588	USA	PHARMACY MANAGER	29-1051	PHARMACISTS	Y	120349 Y	Level III	67.5	0	H				
18	CINCINNATI	OH	45206	USA	PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPISTS	Y	28.66 H	Level I	28.66	28.66	H				
19	ABERDEEN	WA	98520	USA	TECHNICAL DESIGN PROGRAM INSTRUCTOR	17-2071	ELECTRICAL ENGINEERS	Y	75277 Y	N/A	75277	0	Y				
20	LUBBOCK	TX	79430	USA	RESEARCH ASSOCIATE	19-4021	BIOLOGICAL TECHNICIANS	Y	29078 Y	Level II	31212	0	Y				
21	SUNRISE	FL	33323	USA	PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPISTS	Y	32.02 H	Level I	40	0	H				
22	ABILENE	TX	79606	USA	SOFTWARE DEVELOPER	15-1132	SOFTWARE DEVELOPERS, APPLICATIONS	Y	48901 Y	Level I	60000	0	Y				
23	ABILENE	TX	79698	USA	ASSISTANT PROFESSOR OF NURSING	25-1072	NURSING INSTRUCTORS AND TEACHERS, POSTSECONDARY	Y	60307 Y	Level III	90000	0	Y				
24	ABILENE	TX	79697	USA	ESL INSTRUCTOR	25-1199	POSTSECONDARY TEACHERS, ALL OTHER	Y	24340 Y	Level I	25064	0	Y				
25	ABILENE	TX	79606	USA	PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPIST	Y	66602 Y	Level I	81000	0	Y				
26	ABILENE	TX	79602	USA	CASE MANAGER	29-1141	REGISTERED NURSES	Y	29.21 H	Level III	29.9	0	H				
27	CINCINNATI	OH	45209	USA	PHYSICAL THERAPIST	29-1123	PHYSICAL THERAPISTS	Y	32.02 H	Level I	32.02	32.02	H				

The column from our dataset `soc_code` has two parts, the first part defines the main category and the second part defines the sub-category. For instance, the code 29-1051, in this code, the first part '29' defines the main category which is the health care industry and 1051 specifies the subcategory, here it is the pharmacy. Thus, we decided to split the main and subcategory. We were able to achieve that by using the function "Convert Text to Columns" we achieved the below results.

Step 1: Select the Column to Split, in this case "`soc_code`"

Step 2: Click on DATA

Step 3: Select Text to Columns

Step 4: Select the Delimiter as "-" (since our code is delimited by character '-')

Step 5: click on Finish.

In our data we have column Soc_Code with Specific Format, the first two 2 letters with main category and remaining numbers denotes sub category, both of these codes are separated by character '-'

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☐ Tab
☐ Semicolon
☐ Comma
☐ Space
☒ Other: -

☐ Treat consecutive delimiters as one

Text qualifier: "

Data preview

soc_code
29 1063
19 1013
15 2031
17 2112
29 1069

Cancel < Back Next > Finish

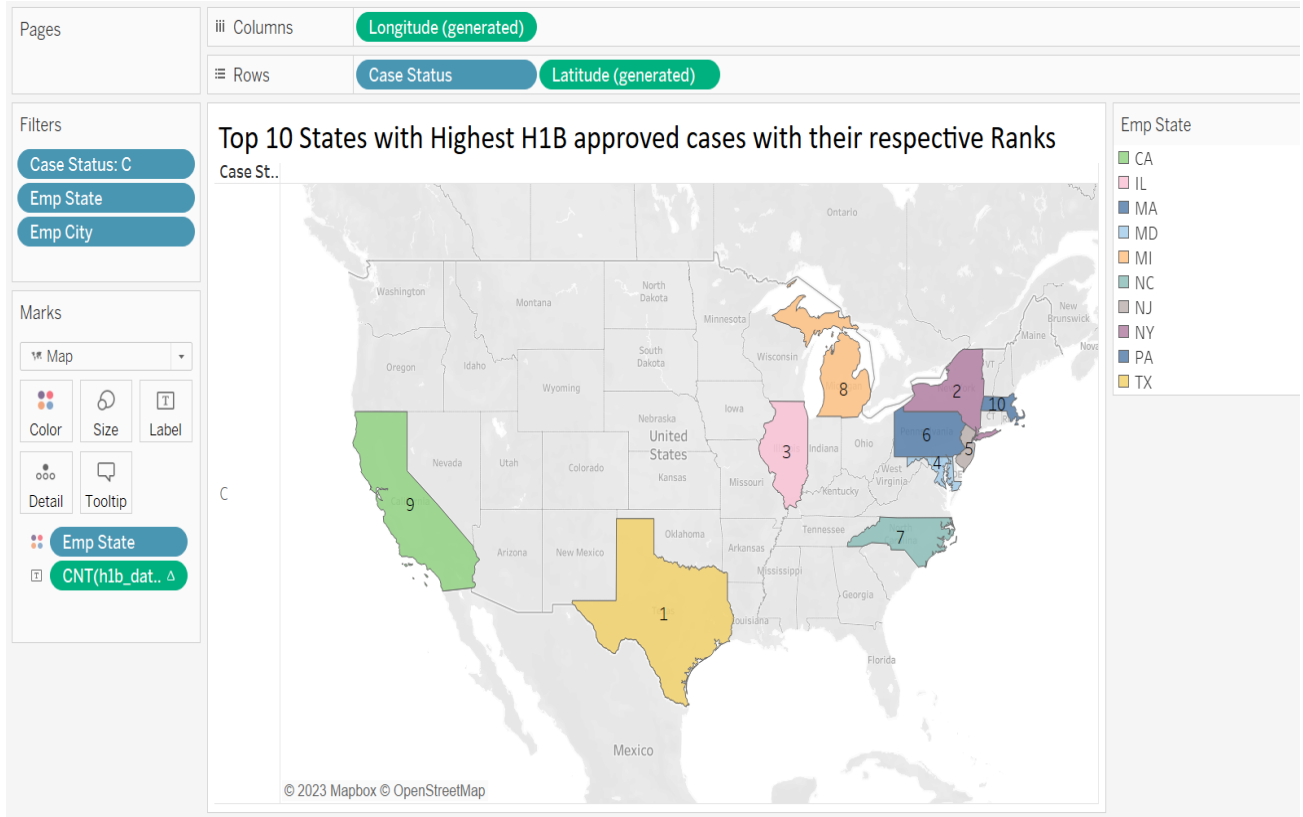
Post-Cleaning:

FileHomeInsertPage LayoutFormulasDataReviewViewHelpPower Pivot														CommentsShare	
Get Data				Queries & Connections				Data Types				Sort & Filter		Data Tools	
From Text/CSV				Recent Sources				OrganizationStocks				Filter		Text to Columns	
From Web				Existing Connections								Clear		What-If Analysis	
From Table/Range				Refresh All								Reapply		Forecast Sheet	
				Properties										Group	
				Edit Links										Ungroup	
				Queries & Connections				Data Types				Sort & Filter		Advanced	
												Data Tools		Forecast	
														Outline	
L1															
soc_code															
	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	emp_city	emp_state	emp_zip	emp_coun	job_title	soc_code		full_time	prevailing_pw_unit	pw_level	wage_fr	wage_to	wage_unit		
2	GREENWOOD	SC	29646	USA	NEPHROLOGIST	29		1063	Y	187200	Y	N/A	190000	0	Y
3	MOSCOW	ID	83844	USA	POST DOCTORAL FELLOW	19		1013	Y	39957	Y	Level I	47507	0	Y
4	HIGH POINT	NC	27265	USA	OPERATION ANALYST	15		2031	Y	59966	Y	Level I	65000	0	Y
5	KEENE	NH	3431	USA	SR. INDUSTRIAL ENGINEER	17		2112	Y	78832	Y	Level II	86988.15	0	Y
6	SIOUX FALLS	SD	57117	NA	HEMATOLOGIST/ONCOLOGIST	29		1069	Y	169645	Y	NA	450000	0	Y
7	SIOUX FALLS	SD	57117	USA	FAMILY MEDICINE PHYSICIAN	29		1062	Y	131581	Y	Level I	131581	0	Y
8	SIOUX FALLS	SD	57117	NA	PEDIATRICIAN	29		1065	Y	187200	Y	NA	187200	0	Y
9	ABERDEEN	SD	57401	USA	INSTRUCTOR OF BUSINESS ACCOUNTING	25		1011	Y	45010	Y	Level I	51487	0	Y
10	ABERDEEN	SD	57401	USA	PEDIATRICIAN	29		1065	Y	166982	Y	Level I	190000	0	Y
11	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNSIST)	29		1063	Y	41725	Y	Level I	240000	0	Y
12	SIOUX FALLS	SD	57117	NA	PEDIATRICIAN	29		1065	Y	187200	Y	NA	187200	0	Y
13	SIOUX FALLS	SD	57105	USA	NEPHROLOGIST	29		1069	Y	187199	Y	N/A	187199	0	Y
14	ABERDEEN	SD	57401	USA	HOSPITALIST (INTERNSIST)	29		1063	Y	41725	Y	Level I	240000	0	Y
15	ABERDEEN	SD	57401	USA	PODIATRIST	29		1081	Y	54330	Y	Level II	215000	0	Y
16	SIOUX FALLS	SD	57117	USA	INTERVENTIONAL CARDIOLOGIST	29		1069	Y	326105	Y	N/A	326105	0	Y
17	PLEASANTON	CA	94588	USA	PHARMACY MANAGER	29		1051	Y	120349	Y	Level III	67.5	0	H
18	CINCINNATI	OH	45206	USA	PHYSICAL THERAPIST	29		1123	Y	28.66	H	Level I	28.66	28.66	H
19	ABERDEEN	WA	98520	USA	TECHNICAL DESIGN PROGRAM INSTRUCTOR	17		2071	Y	75277	Y	N/A	75277	0	Y
20	LUBBOCK	TX	79430	USA	RESEARCH ASSOCIATE	19		4021	Y	29078	Y	Level II	31212	0	Y
21	SUNRISE	FL	33323	USA	PHYSICAL THERAPIST	29		1123	Y	32.02	H	Level I	40	0	H
22	ABILENE	TX	79606	USA	SOFTWARE DEVELOPER	15		1132	Y	48901	Y	Level I	60000	0	Y
23	ABILENE	TX	79698	USA	ASSISTANT PROFESSOR OF NURSING	25		1072	Y	60307	Y	Level III	90000	0	Y
24	ABILENE	TX	79697	USA	ESL INSTRUCTOR	25		1199	Y	24340	Y	Level I	25064	0	Y
25	ABILENE	TX	79606	USA	PHYSICAL THERAPIST	29		1123	Y	66602	Y	Level I	81000	0	Y
26	ABILENE	TX	79602	USA	CASE MANAGER	29		1141	Y	29.21	H	Level III	29.9	0	H
27	CINCINNATI	OH	45209	USA	PHYSICAL THERAPIST	29		1123	Y	32.02	H	Level I	32.02	32.02	H

D. Data Visualizations

1. Analysis question – What are the Top 10 States with H1B approved cases and their respective Ranks

Visualization



Category - Geographical map with Ranking

Geographical maps have been used to show the data of the states that have highest number of H1-B approved cases. We have also included the rank for each state with which we can say what's the state with most number of visa approved cases etc. Texas, New York, Illinois, Maryland, New

Jersey, Pennsylvania, North Carolina, Michigan, California, Massachusetts are the top 10 states with highest number of H1-B approved cases.

This visualization shows the geographical distribution of H1 B-approved cases across the United States. The color indicates the rank of the state based on the number of approved cases.

The insights from the data visualization are as follows:

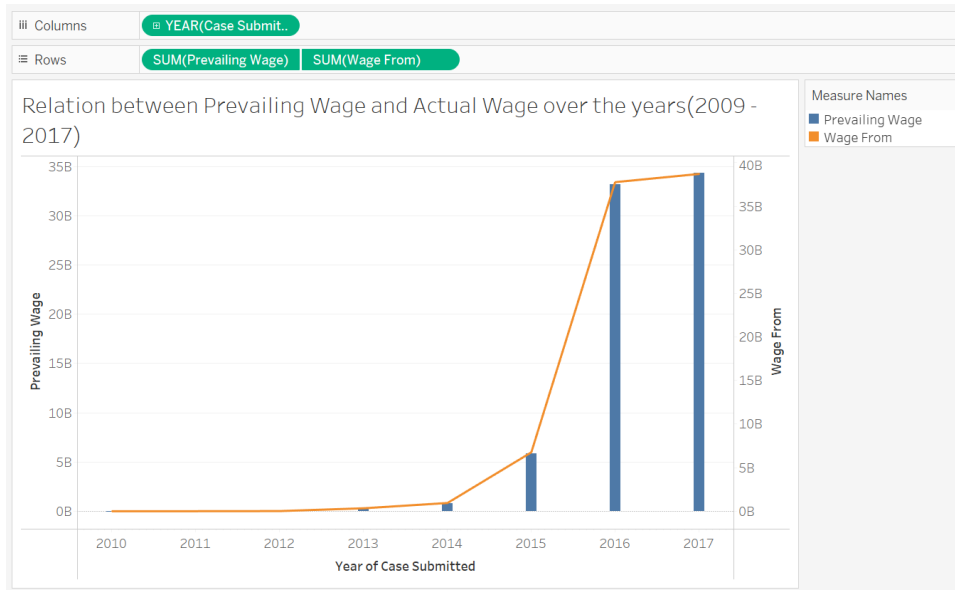
Texas has the highest number of H1 B-approved cases in 1st Place. It is followed by New Jersey and Illinois, standing in 2nd and 3rd places respectively.

The majority of the Top 10 States are in the East, with only Texas and Illinois representing the Midwest. There is a significant difference in the number of approved cases between the top three states and the rest of the states in the Top 10.

Overall, the dashboard visualization provides an effective way to understand the geographical distribution and concentration of H1B approved cases in the United States and the states with the highest number of approved cases. It also highlights the dominance of Texas in terms of the number of H1 B-approved cases.

2. Analysis question – Show the Relation between Prevailing Wage and Actual Wage over the years(2009 - 2017)

Visualization



Category – Dual Axis

This visualization shows the relationship between prevailing wage and actual wage over the years 2009-2017 using a dual axis graph. The prevailing wage is represented by the blue line, while the actual wage is represented by the orange line.

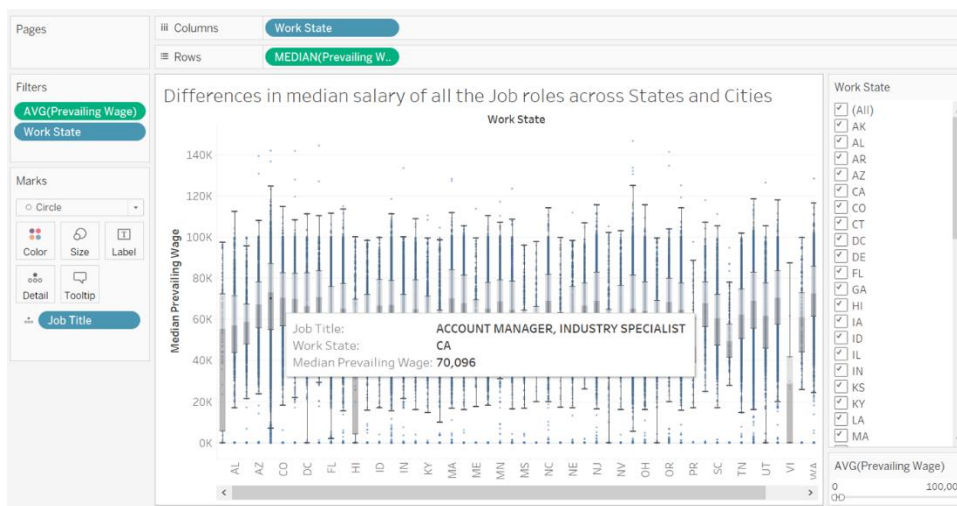
The first insight from this visualization is that the prevailing wage is consistently higher than the actual wage, which suggests that there is a gap between what employers are legally required to pay and what employees actually earn.

The second insight is that both wages have generally increased over time, but the prevailing wage has increased at a faster rate. This could be due to changes in minimum wage laws or other factors that have influenced the prevailing wage.

Overall, this visualization highlights the gap between prevailing wage and actual wage, and shows how both wages have changed over time. It can be useful for policymakers and researchers who are interested in understanding trends in wages and how they relate to minimum wage laws and other labor market factors.

3. Analysis question - Show the differences in median salary of all the Job roles across States and Cities

Visualization



Category – Box and whisker

This visualization represents the differences in the median salary of all the job roles across states, using a box and whiskers plot.

The box and whiskers plot show the distribution of median salaries for each job title in different states. The X-axis represents the States, while the Y-axis represents the Prevailing Wage. The

boxes represent the interquartile range (IQR) of the data, with the whiskers representing the minimum and maximum values. The horizontal line inside the box represents the median value.

The insights from the data visualization are as follows:

The box and whiskers plot show the differences in median prevailing salaries across different job titles, and states. For instance, some job titles have a higher median salary than others, while some states have higher median salaries than others.

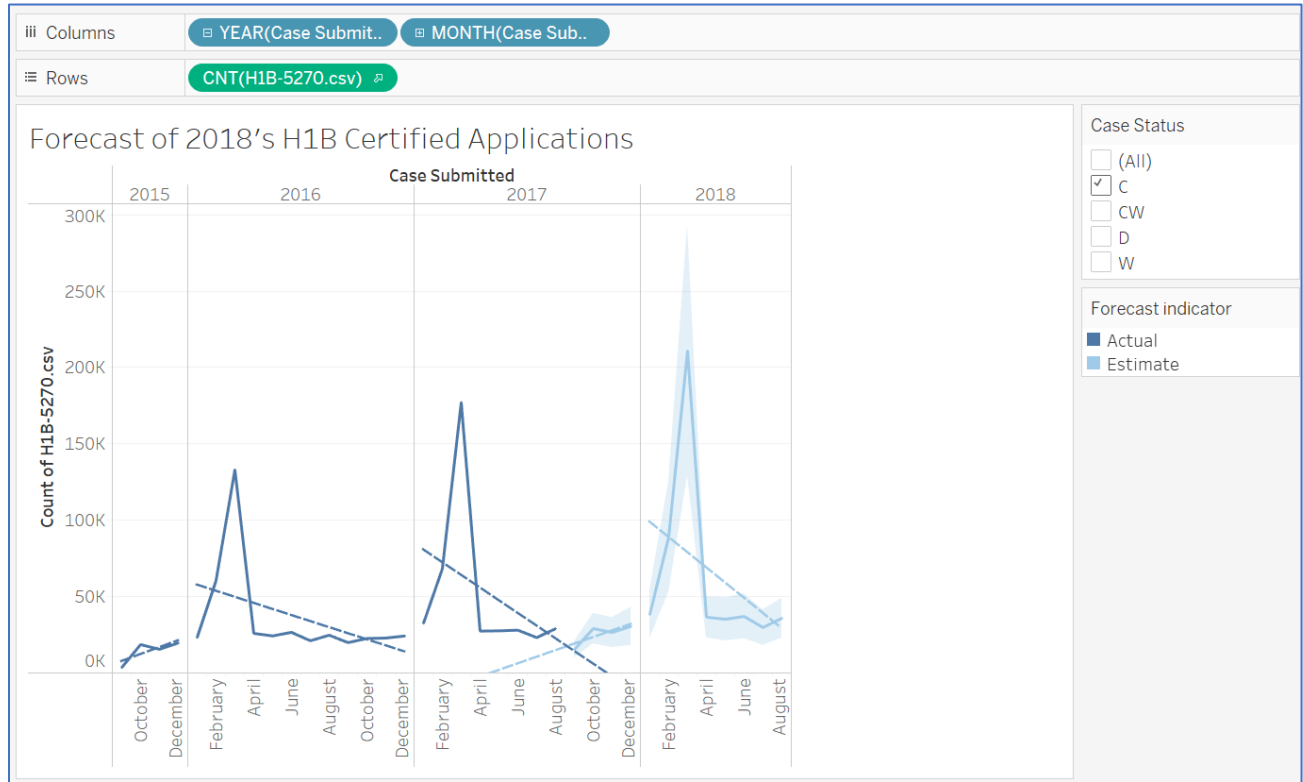
The visualization also highlights the distribution of salaries for each job title, with some job titles having a wider distribution than others. For instance, some job titles have a larger IQR, indicating a wider range of salaries for that job title.

The box and whiskers plot also provides insights into the minimum and maximum salaries for each job title in different states. This information can be used to identify the job titles that have the highest or lowest salaries in a particular state.

Overall, the Tableau visualization provides an effective way to understand the differences in median salaries across different job titles and states. The box and whiskers plot highlights the distribution of salaries for each job title and provides insights into the minimum and maximum salaries. This information can be used to identify the job titles that have the highest or lowest salaries in a particular state.

4. Analysis question – Visualize the Forecast of 2018's H1B Certified Applications

Visualization



Category – Forecast Trend Lines

This visualization uses forecast trend lines to visualize the forecast of 2018's H1 B-certified applications. The data used in the visualization includes H1 B-certified applications from the years 2015 to 2017, which were used to predict the number of applications in 2018.

The visualization shows a line chart with the X-axis representing the case submitted years and months from 2015 to 2017, and the Y-axis representing the number of Certified H1B applications.

The line chart includes trend lines that show the predicted number of applications in 2018.

The insights from the data visualization are as follows:

The forecast trend lines provide a clear visualization of the predicted trend of H1 B-certified applications for the year 2018. The trend lines are based on the previous years' data and can help stakeholders plan their activities accordingly.

The forecast trend lines also show the potential growth or decline in H1 B-certified applications. If the trend line shows an upward slope, it indicates potential growth, while a downward slope indicates a decline.

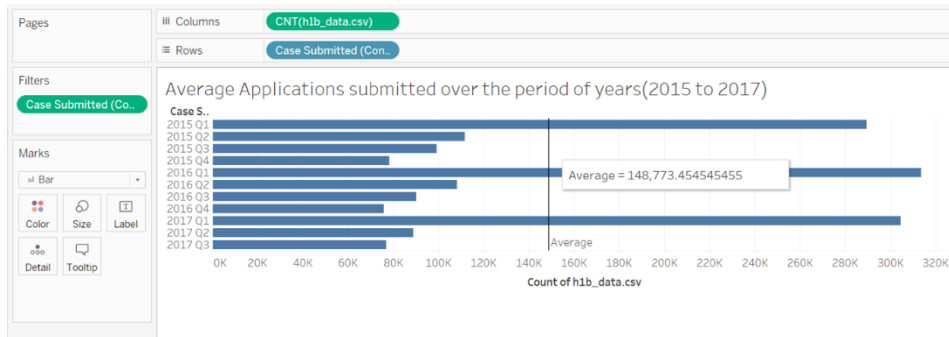
The visualization also highlights the accuracy of the forecast. As shown the trend line closely matches the actual number of H1 B-certified applications for the year 2018, which indicates a high level of accuracy in the forecast.

The visualization also allows stakeholders to compare the number of certified H1B applications between different years and identify any potential trends or patterns.

Overall, this visualization provides an effective way to visualize the forecast of H1 B-certified applications for the year 2018. The forecast trend lines provide valuable insights into the potential growth or decline of H1 B-certified applications and allow stakeholders to plan their activities accordingly.

5. Analysis question – What is the Average number of H1B Applications submitted over the period of years (2015 to 2017)

Visualization



Category — Dates and Reference Line

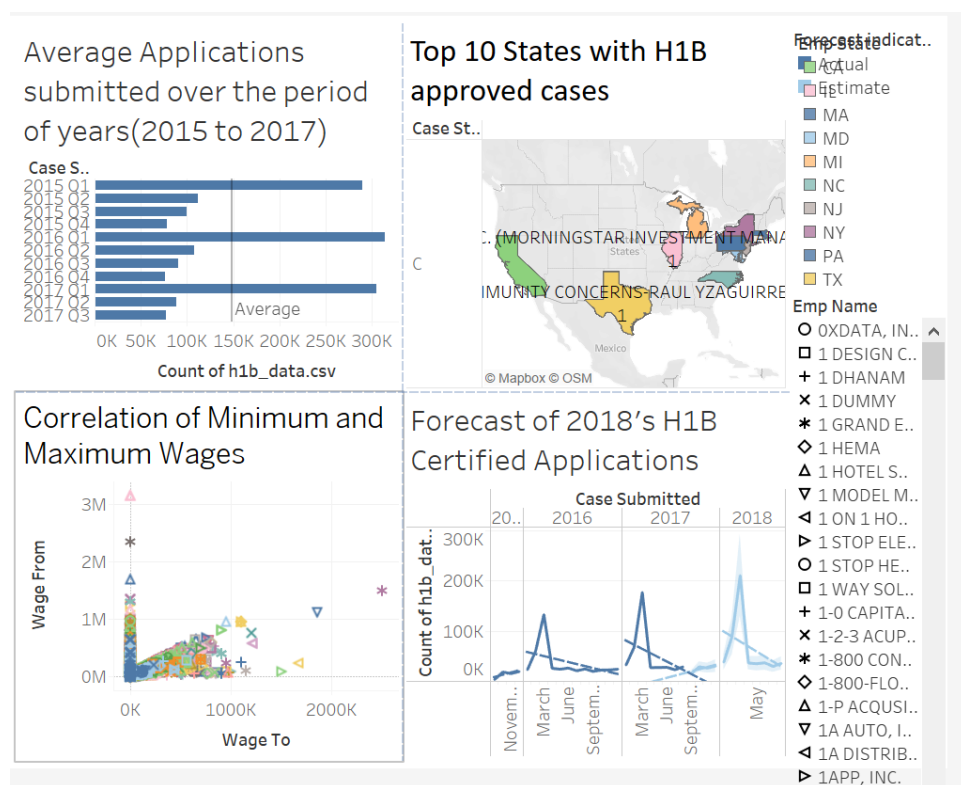
The dashboard visualization depicts the average number of H1B visa applications submitted over the period of years from 2015 to 2017. It utilizes both Dates graph and Reference Line graph to present the data.

The Dates graph represents the number of applications submitted each year. It shows that the number of H1B visa applications submitted has increased steadily from 2015 to 2017. This insight indicates that there is an increasing demand for foreign skilled workers in the US job market, and companies are seeking to fill these positions through the H1B visa program.

The Reference Line graph, on the other hand, provides a comparison of the average number of applications submitted over the three-year period. The reference line shows the average number of applications submitted, and the graph shows how the number of applications submitted in each year compares to this average. This visualization shows that the average number of applications were around 148,773.

Overall, the dashboard visualization provides valuable insights into the trends and patterns of H1B visa applications submitted over the period of years from 2015 to 2017, highlighting the increasing demand for foreign skilled workers in the US job market and potential factors that impact the number of applications submitted.

E. Dashboards



The other visualization added in the Dashboard other than the 5 Analysis Questions depicts the average number of H1B applications submitted over the period of years 2015-2017 across the United States using a Dates graph and a Reference Line graph. The Dates graph displays the

average number of applications submitted per month, while the Reference Line graph shows the overall average number of applications submitted over the entire period.

The insight from this visualization is that the Reference Line graph provides a useful comparison for the Dates graph, allowing viewers to see how the average number of applications submitted in each month compares to the overall average. This can help to identify which months are particularly high or low in terms of application volume.

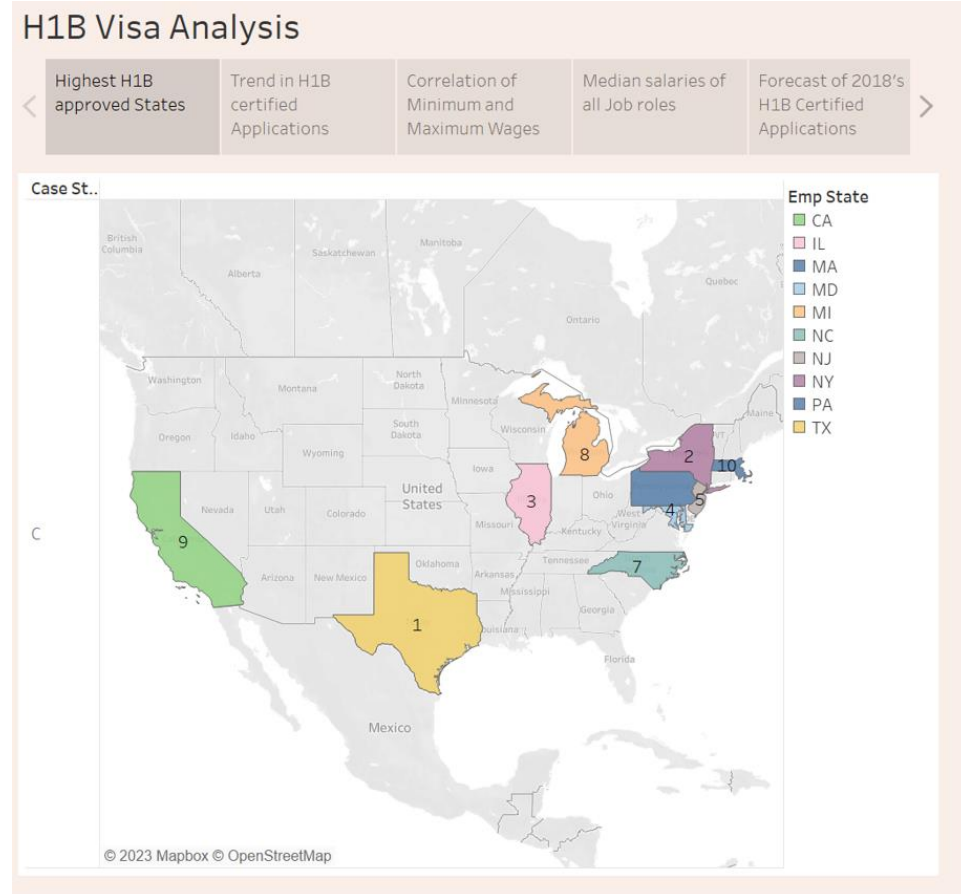
This visualization provides insights into trends in the number of job applications submitted over time. It can be useful for recruiters and hiring managers who are interested in understanding changes in the job market and how they may affect the number and quality of applicants.

Overall, the dashboard visualization presents an analysis of H1B data across different states, companies, and cities in the United States using a variety of visualization tools. The Dates and Reference Line graph provides insights into trends in H1B applications over time, while the Maps and Rank Graphs show which states have the highest numbers of approved H1B cases and their respective ranks. The Scatter Plot Graph allows for an analysis of the correlation between minimum and maximum H1B wages. Additionally, the Forecast Trend Line Graph visualizes the forecast for 2018's H1B certified applications. These visualizations help to reveal patterns and trends in H1B data, such as which states have the highest numbers of approved cases, how wages are correlated, and how H1B applications are expected to change in the future. This information can be useful for employers, employees, and policymakers who are interested in understanding the H1B visa program and its impact on the labor market.

F. Story Telling

We as a group of Information Systems students who were tasked with visualizing data related to H-1B visas and prevailing wage levels. We wanted to create visualizations that would help policymakers and researchers better understand trends in wages and how they relate to minimum wage laws and other labor market factors.

The first visualization we created was a geographical map that showed the number of H-1 B-approved cases in each state. we used color-coding to indicate the rank of each state and found that Texas had the highest number of approved cases. The data also showed that the majority of the top 10 states were in the East, with Texas and Illinois being the only states in the Midwest.



Next, we created Dates and Reference Lines. The dates graph displays the number of applications submitted over time, while the reference line graph shows the average number of applications submitted.

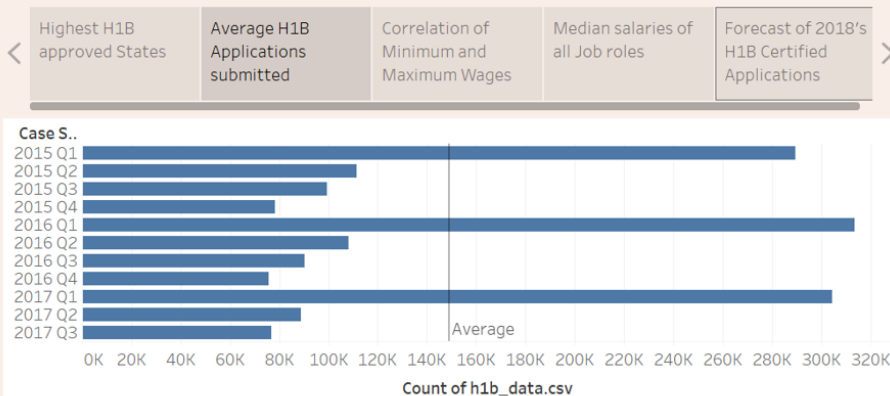
The article "H1B Visa Analysis. VisaGuide.World" provides an analysis of the H-1B visa program over the last 10 years, looking at the number of visas issued, the industries that use the program the most, and the top employers of H-1B visa holders. The analysis shows that the H-1B program has seen significant growth over the past decade, with the number of visas issued increasing from around 130,000 in 2009 to nearly 190,000 in 2018. Additionally, the technology industry continues to be the largest user of the H-1B program, with over 65% of visas issued going to workers in computer-related occupations.

The article also provides a list of the top 10 H-1B visa employers over the past decade, which includes several major technology companies such as Infosys, Tata Consultancy Services, and Wipro. The top employer, however, is Cognizant Technology Solutions, which has filed over 300,000 H-1B visa applications since 2009. The analysis shows that the H-1B program continues to be an important source of highly skilled workers for the U.S. economy, but also raises concerns about the high number of visas being issued to a few large employers, and the potential for abuse of the program by some employers.

This article supports our visualization that there was a significant increase in h1 b applications.

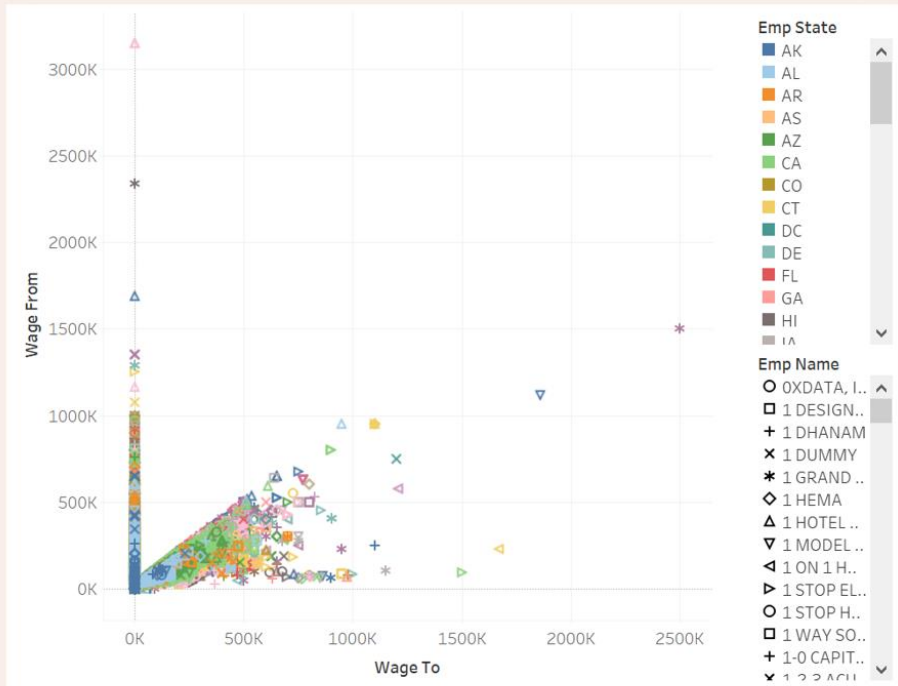
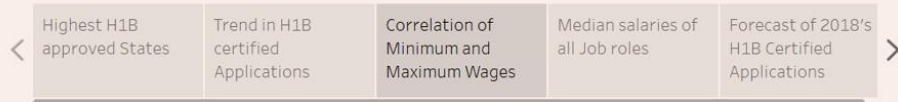
as we can see the average h1b applications ranges from 140k to 160 k

H1B Visa Analysis



We observed the correlation of minimum and maximum wages. The dots and shapes represented companies and states. The color coding of the dots based on the case status provided a way to identify the variation in the correlation between the minimum and maximum wages.

H1B Visa Analysis

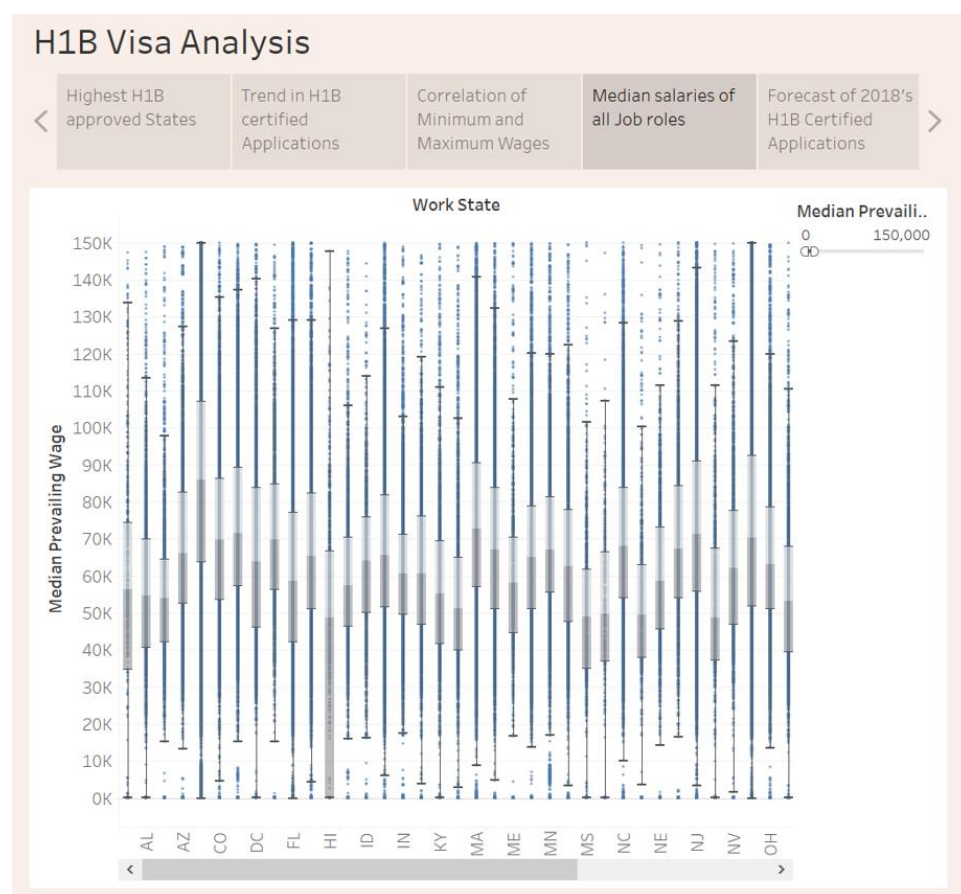


We then created a box and whisker plot that showed the differences in median salary for different job titles across states. The plot showed that some job titles had a higher median salary than others, and some states had higher median salaries than others. The plot also showed the distribution of salaries for each job title, with some having a wider range than others.

The below Visualization, as described in the context, is related to the article as it shows the predicted trend. This prediction can provide insight into the potential demand for H-1B workers and how this demand might impact the prevailing wage levels for foreign workers. By providing a visual representation of the trend lines, stakeholders can plan their activities accordingly and make informed decisions about hiring foreign workers under the H-1B visa program.

The authors in the article titled "H-1B visas and prevailing wage levels" also found that the H-1B

wage system is heavily influenced by the wage surveys conducted by the Department of Labor, which often do not accurately reflect the wages paid to U.S. workers in the same occupation and location. Furthermore, the authors found that the H-1B wage system has a negative impact on U.S. workers, who are often forced to compete with foreign workers who are willing to work for lower wages.



The visualization consisted of a line chart, with the X-axis representing the case submitted years and months from 2015 to 2017, and the Y-axis representing the number of Certified H1B

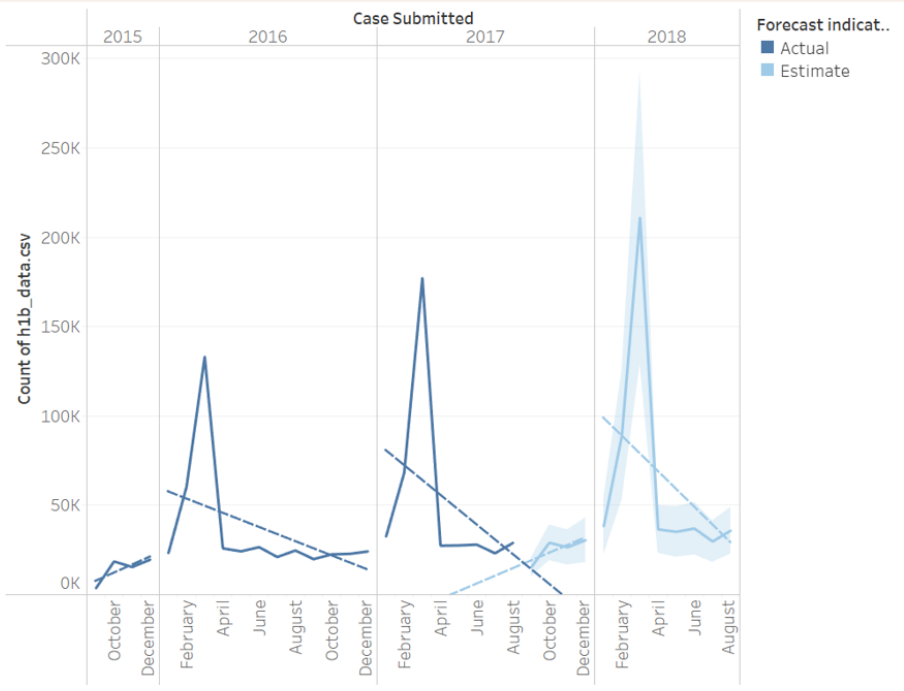
applications. The chart included trend lines that showed the predicted number of applications in 2018.

The insights from the visualization were fascinating. The trend lines provided a clear visualization of the predicted trend of H1 B-certified applications for the year 2018, and stakeholders could plan their activities accordingly. The trend lines also showed the potential growth or decline in H1 B-certified applications, with an upward slope indicating potential growth and a downward slope indicating a decline.

The article "H1B Visa Top Sponsors 2015 - Overall" presents a list of the top H-1B visa sponsors in 2015, based on data obtained from the U.S. Department of Labor. The list includes several major technology companies, such as Google, Microsoft, and Apple, as well as consulting and outsourcing firms like Cognizant, Infosys, and Tata Consultancy Services. The article notes that these companies are known to be heavy users of the H-1B visa program, and that they often bring in large numbers of foreign workers to fill highly skilled positions in the U.S. job market. However, the article also raises concerns about the potential for abuse of the H-1B program by some employers, who may use it to bring in cheap labor and displace American workers.

H1B Visa Analysis

Highest H1B approved States	Trend in H1B certified Applications	Correlation of Minimum and Maximum Wages	Median salaries of all Job roles	Forecast of 2018's H1B Certified Applications
---	---	--	--	---



Citation

Daniel Costa and Ron Hira. H-1B visas and prevailing wage levels: How employers use loopholes to pay below-market wages to U.S. information technology workers. Economic Policy Institute.

<https://www.epi.org/publication/h-1b-visas-and-prevailing-wage-levels/>

Radoiu, L. (2019, April 10). 10 Years Challenge: H1B Visa Analysis. Visa Guide World

<https://visaguide.world/news/us/h1b-visa/10-years-challenge-h1b-visa-analysis/>

Immihelp.com. (2015). H1B Visa Top Sponsors 2015 - Overall. [https://www.immihelp.com/h1b-](https://www.immihelp.com/h1b-visa-top-sponsors-2015-overall/)

[visa-top-sponsors-2015-overall/](https://www.immihelp.com/h1b-visa-top-sponsors-2015-overall/)