DATA SCIENCE

CASE STUDY ON

H1B VISA PETITION DATA SET

SARVESH PRAJAPATI
PRN 17030141070, MBA-IT SEM-3
DIV-A

H1-B VISA PETITION DATA SET

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 H1-B visa, a US employer must offer a job and petition for 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, PhD) and work in a full-time position.

1. Data Collection

Data collection involves gathering of data from various sources which could be flat files, legacy systems etc. The source data can come from 4 types of categories which are:

- Production Data: Data comes from financial system, manufacturing system, CRM systems etc.
- Internal Data: Includes private spreadsheets, documents, customer profiles and sometimes departmental databases also.
- Archived Data: In every operational system, data is stored into archived files. A data warehouse keeps historical snapshots of data which can be needed for analysis over time.
- External Data: Includes data provided by external entities. For instance, executives use the statistical data that may be provided by external agencies.

The initial data for our project **H1-B Visa Petition** has been collected from the following link:

https://www.kaggle.com

	A	B	C	D	E	F	G	H	I	J	K
1	SERIAL	CASE_STATUS	EMPLOYER_NAME	SOC_NAME	JOB_TITLE	FULL_TIME_POS	PREVAILING_W	YEA	WORKSITE	Longitude	Latitude
2	1	CERTIFIED-WITHDRA	UNIVERSITY OF MICHIGAN	BIOCHEMISTS AND BIOPHYSI	POSTDOCTORAL RESEARCH FELLOW	N	36067	###	ANN ARBOR, MICHIGAN	-83.743	42.2808
3	2	CERTIFIED-WITHDRA	GOODMAN NETWORKS, INC.	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	242674	###	PLANO, TEXAS	-96.6988	33.0198
4	3	CERTIFIED-WITHDRA	PORTS AMERICA GROUP, INC.	CHIEF EXECUTIVES	CHIEF PROCESS OFFICER	Y	193066	###	JERSEY CITY, NEW JERSE	-74.0776	40.7281
5	4	CERTIFIED-WITHDRA	GATES CORPORATION, A WHOLLY-OWNED SUBSIDIARY OF T	CHIEF EXECUTIVES	REGIONAL PRESIDEN, AMERICAS	Y	220314	###	DENVER, COLORADO	-104.9902	39.7392
6	5	WITHDRAWN	PEABODY INVESTMENTS CORP.	CHIEF EXECUTIVES	PRESIDENT MONGOLIA AND INDIA	Y	157518.4	###	ST. LOUIS, MISSOURI	-90.1994	38.6271
7	6	CERTIFIED-WITHDRA	BURGER KING CORPORATION	CHIEF EXECUTIVES	EXECUTIVE V P, GLOBAL DEVELOPMENT AND PRESIDENT, L	Y	225000	###	MIAMI, FLORIDA	-80.1917	25.7616
8	7	CERTIFIED-WITHDRA	BT AND MK ENERGY AND COMMODITIES	CHIEF EXECUTIVES	CHIEF OPERATING OFFICER	Y	91021	###	HOUSTON, TEXAS	-95.3698	29.7604
9	8	CERTIFIED-WITHDRA	GLOBO MOBILE TECHNOLOGIES, INC.	CHIEF EXECUTIVES	CHIEF OPERATIONS OFFICER	Y	150000	###	SAN JOSE, CALIFORNIA	-121.8863	37.3382
10	9	CERTIFIED-WITHDRA	ESI COMPANIES INC.	CHIEF EXECUTIVES	PRESIDENT	Y	127546	###	MEMPHIS, TEXAS	90.0491	35.1495
11	10	WITHDRAWN	LESSARD INTERNATIONAL LLC	CHIEF EXECUTIVES	PRESIDENT	Y	154648	###	VIENNA, VIRGINIA	-77.2652	38.9012
12	11	CERTIFIED-WITHDRA	H.J. HEINZ COMPANY	CHIEF EXECUTIVES	CHIEF INFORMATION OFFICER, HEINZ NORTH AMERICA	Y	182978	###	PITTSBURGH, PENNSYLV	-79.9958	40.4406
13	12	CERTIFIED-WITHDRA	DOW CORNING CORPORATION	CHIEF EXECUTIVES	VICE PRESIDENT AND CHIEF HUMAN RESOURCES OFFICER	Y	163717	###	MIDLAND, MICHIGAN	-84.2472	43.6155
14	13	CERTIFIED-WITHDRA	ACUSHNET COMPANY	CHIEF EXECUTIVES	TREASURER AND COO	Y	203860.8	###	FAIRHAVEN, MASSACHUS	70.9036	41.6378
15	14	CERTIFIED-WITHDRA	BIOCAIR, INC.	CHIEF EXECUTIVES	CHIEF COMMERCIAL OFFICER	Y	252637	###	MIAMI, FLORIDA	-80.1917	25.7616
16	15	CERTIFIED	ERNST & YOUNG U.S. LLP	COMPUTER SYSTEMS ANALY	ADVISORY STAFF	N	61797	###	NEW YORK, NEW YORK	-74.0059	40.7127
17	16	CERTIFIED	INFOSYS LIMITED	COMPUTER SYSTEMS ANALY	TECHNOLOGY LEAD - US	Y	86070	###	PEAPACK, NEW JERSEY	-74.6576	40.7151

Size of the data set is: 17 * 11 (i.e. number of rows are 17 and number of columns are 11).

Name of Columns are as follows:

1. Serial_No
2. Case_Status
3. Employer_Name
4. SOC_Name
5. Job_Title
6. Full_Time
7. Prevailing_Wage
8. Year
9. Worksite
10. Latitude
11. Longitude

2. Data Preprocessing

Data preprocessing involves readying the data for staging i.e. applying transformations on the source data to make it consistent, have standard and uniform values across various dimensions, removal of duplicates, assigning proper naming conventions and filling missing data etc.

2.1 Data Cleansing:

Data cleansing may be of different types. For instance, it may just be correction of misspellings of resolution of conflicts between state codes and zip codes in source data.

2.2 Data Staging:

The external data coming from several disparate sources needs to be converted and made ready in a format that is suitable to be stored for querying and analysis. The staging area involves major functions of extraction, transformation and preparation for loading. Data staging provides a place and an area with a set of functions to:

- Clean
- Change
- Combine
- Convert
- Deduplicate, and
- Prepare source data for storage and use in data warehouse.

2.3 Data Integrity:

Integrity in data means having uniform values for a dimension in the source data. For instance, if a dimension contains numeric data, then it must be ensured that all the values in that dimension for every record must have only numeric type of data.

- The latitude and longitude (geographic coordinates) of few employers were not consistent with other values. That was rectified.
- For instance, the coordinates of location Rhinebeck, New York, were found to be having values N.A., each for latitude and longitude. These values were replaced by actual coordinates of the place as -41.9318 and 73.9074.

2.4 Noise Removal:

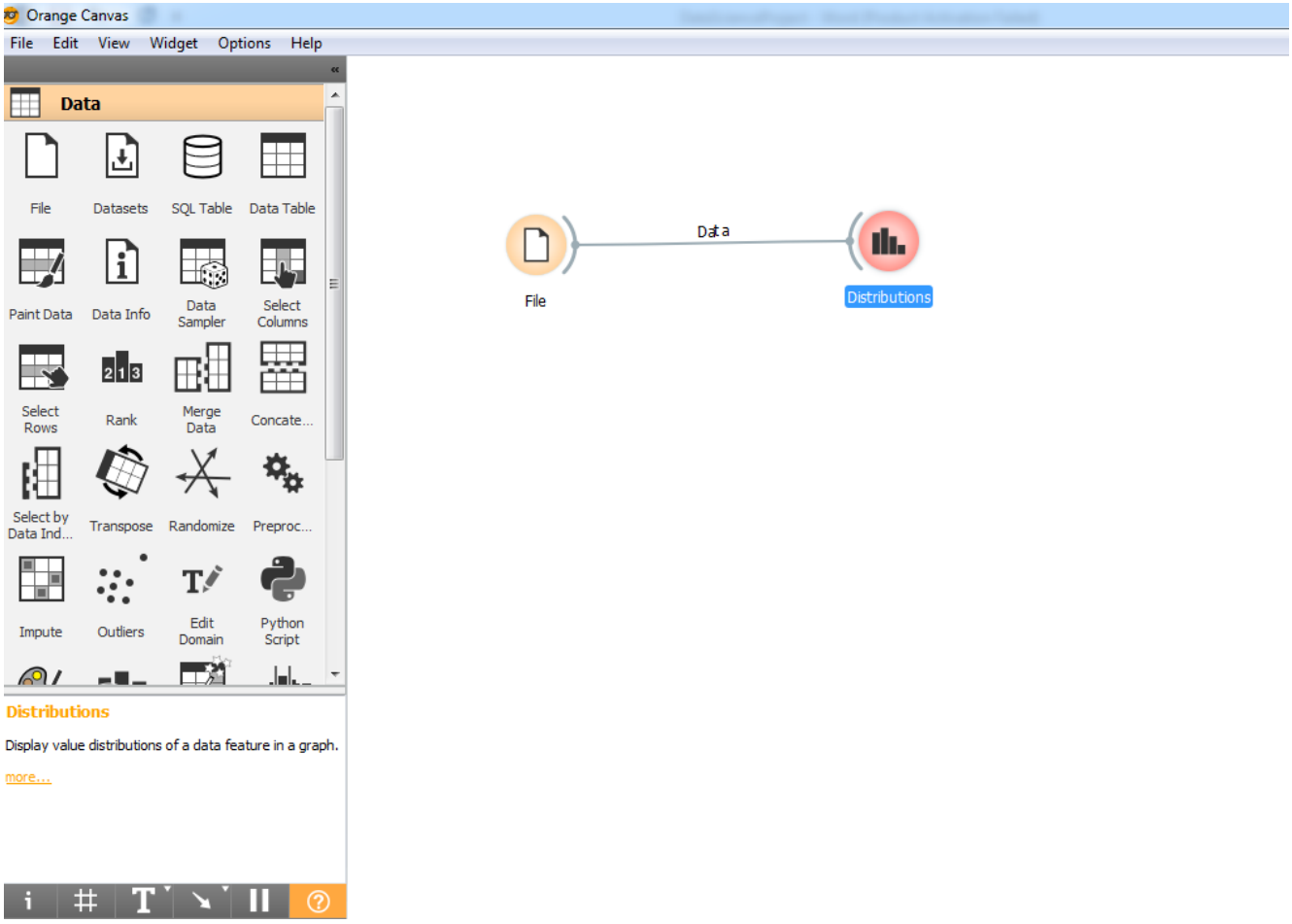
- There were discrepancies in dimension 'SOC_Name' values (somewhere it was COMPUTER SYSTEMS ANALYST and in other records it was COMPUTER SYSTEMS ANALYSTS. It was corrected to COMPUTER SYSTEMS ANALYST).

2.5 Data Formats:

- The source data file was been collected in **CSV** format. The source data has been converted to **XLSX.**

3. Tool Description

The tool utilized for visualization purposes is **Orange 3.**16 which provides exploration with limitless visual analytics. It is a good tool to perform ad hoc analyses in just a few clicks. It has inbuilt widgets like Data, Visualize, Model, Evaluate etc. Each of this widgets has a number of functionalities. The following figure shows the tool having various widgets under Data tab.

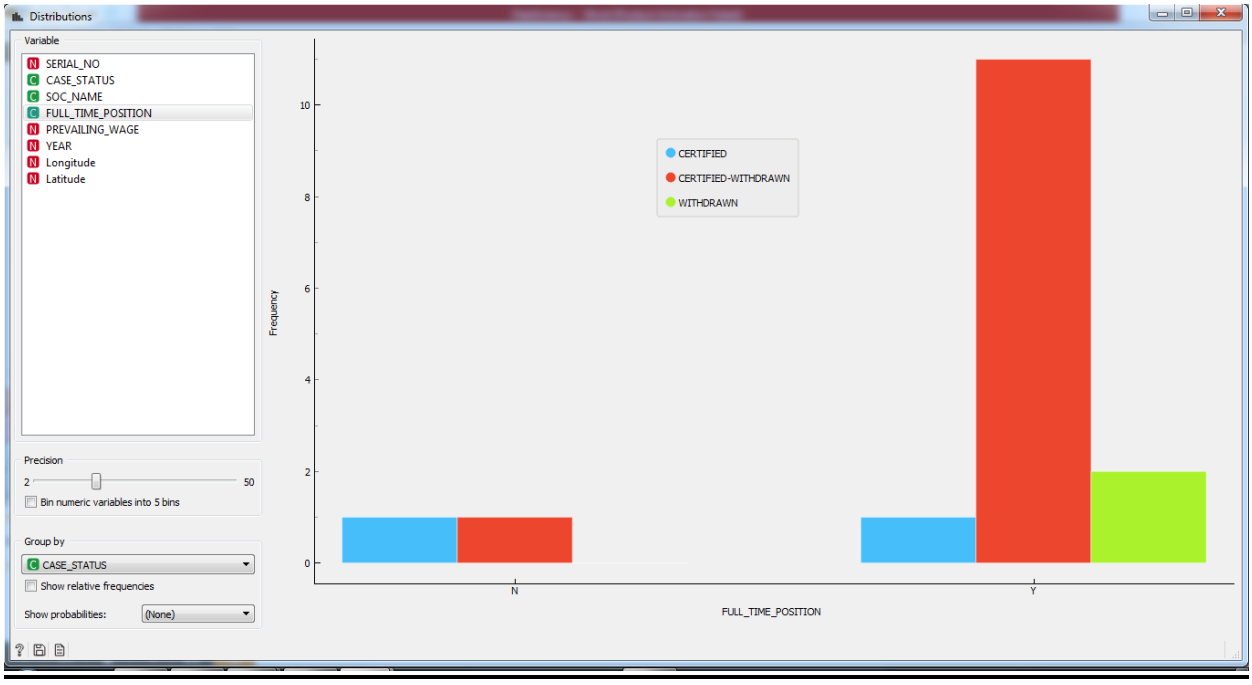


4. Visualization

For instance, for our data set, we had objective of finding out the trend of 'Full_Time' column for those visa petitions which had case status as CERTIFIED, CERTIFIED_WITHDRAWN or WITHDRAWN. We imported the file into Orange3.16 using the '**file**' widget given in the **Data** tab. Then we extended a link to a widget called **'Distributions'** (see above figure). Double clicking on **Distribution** widget gave us the visualization in the form of bar graph and also trend curve. Both these figures are shown ahead.

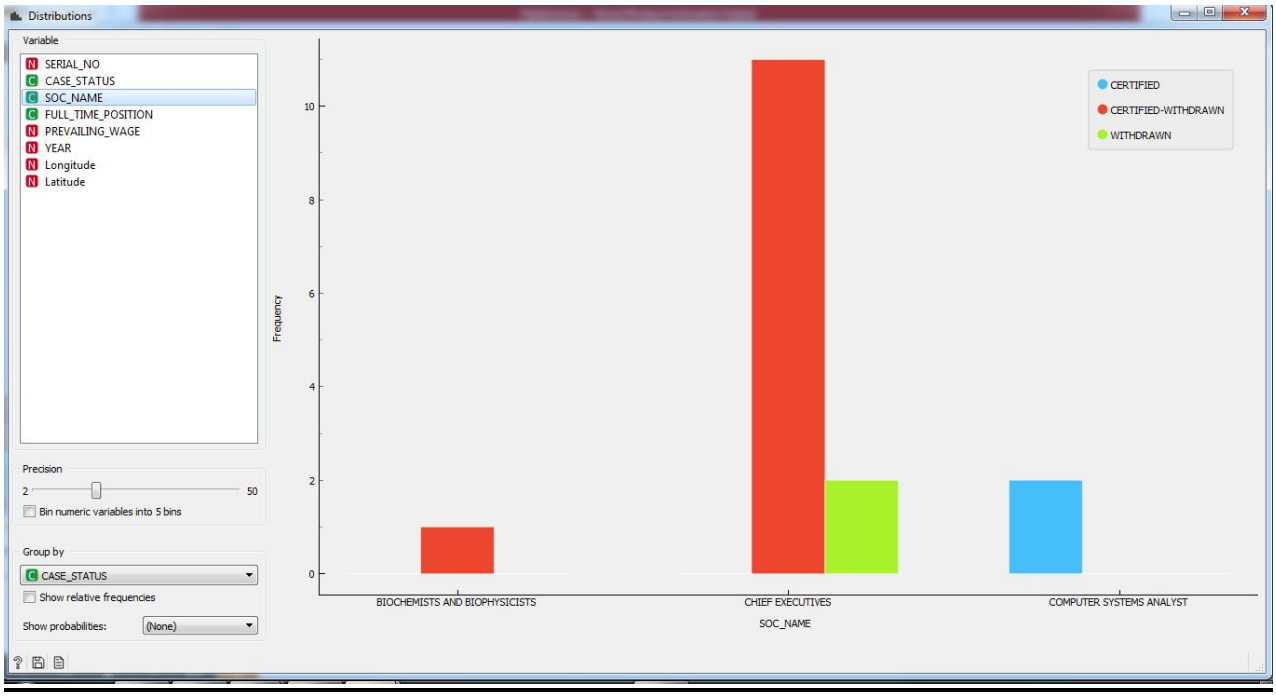
Objective 1: Identify which CASE_STATUS has maximum frequency for FULL_TIME_POSITION.

Output:



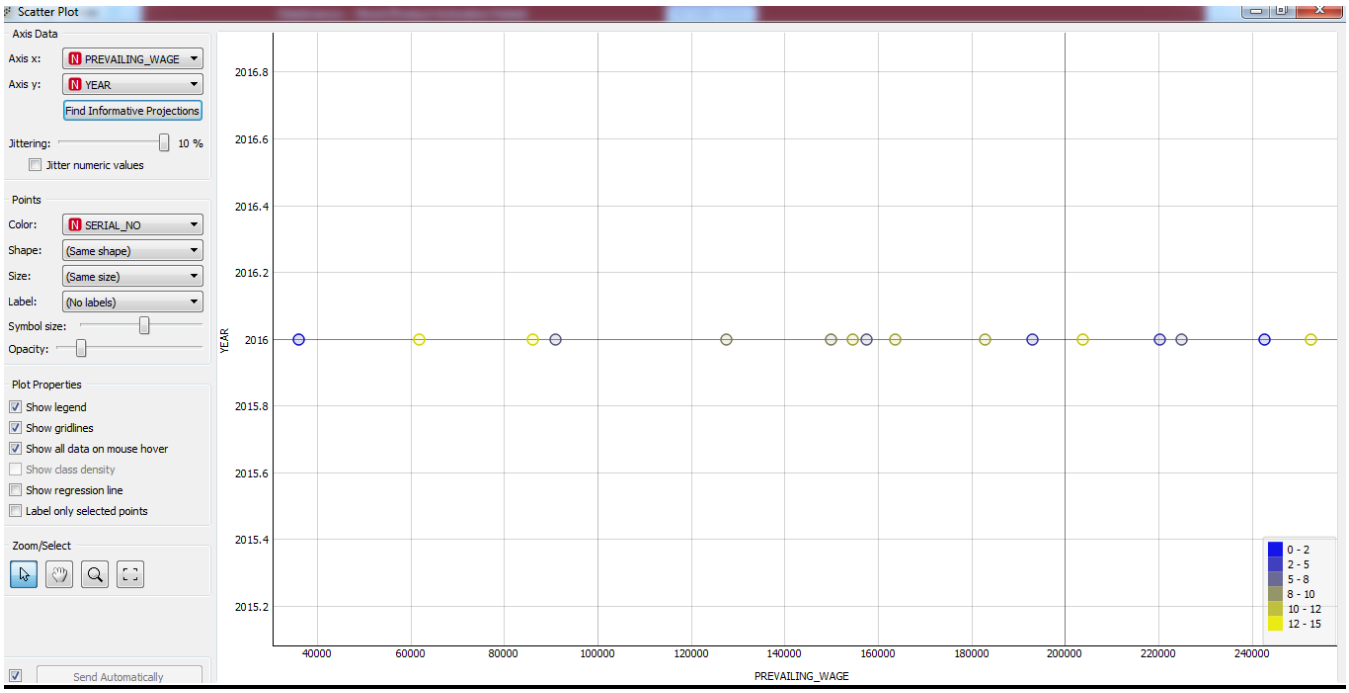
Objective 2: Identify which Case_Status has maximum frequency for attribute 'Soc_Name'.

Output:



Objective 3; How is the prevailing wage varying over the years?

Output: Scatter plot.



CONCLUSION

It is found that the maximum number of H1B Visa petitions have been filed for the job role of Executive profile.

Also, the prevailing age was maximum during the year of 2016 as it has shown steady incline over the years from 2013 onwards.