SMU®

CLUSTER ANALYSIS OF H1B VISAS IN THE UNITED STATES

Project 2 in EMIS 7331 - Data Mining.

Nikhila Byreddy & Samarth Suresh Kumar
CSE/EMIS 7331, Project 2

Table of Contents

Executive Summary

Data Mining methods can be broadly classified as Supervised and Unsupervised learning. If there is a purpose or a target associated it is referred to as supervised learning. Otherwise, it is called Unsupervised learning. Clustering is an unsupervised learning method which attempts to group individual data together by their similarity – but not driven by any purpose. Clustering is a mechanism to evaluate if the data points naturally fall into different groups.

In this report, using data analytics, we use the publicly available immigration data for the available work visas in the United states to analyze trends in the visa applications over time.

For a foreign national to work in the United States, an employer must offer the job and file a petition for an H1B visa with the US immigration department. The primary focus of this project is to analyze about the attributes in the dataset that contribute to the final visa status.

For each dataset, before clustering the data, cleaning is performed. All the duplicate data is removed, appropriate methods are implemented in dealing with the missing data and outliers. Some attributes are selected, and the analysis is performed individually.

This report shows the distribution of top employers, top job positions, Salary trends, Employers and top states. This gives us the understanding of how these variables are distributed in the dataset. Using clustering, we then group applications based on how their locations data, prevailing wages and Job Titles. The goal of this analysis is to obtain insights on the H1b Data applications, salary patterns and location trends to make it easier for the stakeholders: Company Immigration analysts, or job seekers to target specific jobs, locations, employers or salaries.

We compared both the clustering algorithms by using internal validation techniques and found out the best algorithm for clustering for each dataset.

.

List of Figures:

List of Tables:

Data Understanding:

H1b Visa-Dataset:

H1B visas are a category of employment-based, non-immigrant visas for temporary foreign workers in the United States. Labor Condition Application(LCA) is a mandatory document that H1B Sponsor/employer needs to file with US Department of Labor before they submit an H1B petition with USCIS for any non-immigrant worker. As foreign workers are new to America, certain employers can take advantage and mistreat them regarding wages, benefits, etc. In this context, US Dept. of Labor had mandated LCA, and it is essential for a foreign worker as it protects their fundamental rights.

LCA form has essential information about the offered job position for the foreign worker. The fields in LCA form include Job title of the post provided, duration of job position if the position offered is full time or not, salary submitted for the position, location of job position, the prevailing wage for the same position in that area.

The primary goal of the analyst is to analyze the fields of LCA form that could influence the LCA status (Case status). One can interpret the statistics such as

Which employers send most number of H1B visa applications.

Is the number of petitions with a specified Job title increasing over time.

The Location which has the most number of Data Engineers.

1. Are the jobs concentrated in few specific regions?
2. The employers who file the most petitions per year.

The given dataset has 3,002,458 records. Important variables of data file and type of data is provided in the table below:

Table 1: H1b Visa- Dataset Variable info

Column Names	Type	Description	Statistics
X	Id	ID	
CASE_STATUS	Nominal	Case status of application	
EMPLOYER_NAME	Nominal	Name of the employer	
SOC_NAME	Nominal	Job title	
JOB_TITLE	Nominal	Title of the job	
FULL_TIME_POSITION	Nominal	True or false	
PREVAILING_WAGE	RATIO	Salary of the applicant	MIN: 0 MAX: 6.998e+09 Median: 6.502e+04 Mean: 1.470e+05

YEAR	Interval	Year of the application	
WORKSITE	Nominal	Location of the employer	
lon	Interval	Location - longitude	
lat	Interval	Location - latitude	

EMPLOYER_NAME can be used in comparing the prevailing wages, number of applications for different employers and finding the employers who have status as "Certified."

JOB_TITLE - We can find out Specific job position (E.g. Software Engineer) based on which we can compare the case status for a given job title.

PREVAILING_WAGE - We can explore the relationship between prevailing wage, job title.

WORKSITE - Using which we can compare prevailing wage for a given position (software Engineer) for various work locations.

CASE_STATUS - This helps us to analyze how many H1B Visas are certified by different employers.

Data Quality:

H1b Visa-Dataset:

The first column of H1B visa dataset is just a row count. Below is the analysis on the important variables in our dataset:

CASE_STATUS:

There are 13 NA s, one blank (") for CASE_STATUS. However, same CASE_STATUS occurs more than once. We should check if there is a necessity to remove them. Checking the status reveals that the possible status of any employee should be "CERTIFIED," "CERTIFIED-WITHDRAWN," "WITHDRAWN," "DENIED," INVALIDATED," REJECTED," UNASSIGNED." Checking the status shows that no mistakes occurred because different employees may have the same case status.

EMPLOYER_NAME:

There are missing values for EMPLOYER_NAME. The same EMPLOYER_NAME existed in the dataset. There is no mistake in the repetition of EMPLOYER_NAME because the employer may be filing LCA form for different job titles.

SOC_NAME:

There are eleven missing values in the dataset we chose (one "N/A" and ten "NA") for SOC_NAME. Duplicate data is not a problem here because different employers may file a petition for the same case. As SOC_NAME is nominal variable, all the missing values are replaced by "NA" for consistency. These missing values may affect our analysis.

JOB_TITLE:

There is missing data for JOB_TITLE. It is a nominal variable and replaces it with "NA." Also, there are duplicate data for the job title. But this should not be a problem because an employer may need to file a petition for different employees with different job titles.

FULL_TIME:

There is one missing value for the employer "Four seasons heating and air conditioning," for which the CASE_STATUS is "Denied." As there is only one lost value in the entire dataset, this missing value should not make much difference in our analysis. The ratio of missing values and the existing values are minimal. Also, there are duplicate values but that should not be a problem because the possible answers if the employee is full time or not is Y(Yes) or N(No) or a missing value.

PREVAILING_WAGE:

There are some missing values in the prevailing wage. There are some wages which are in billions. One may replace all the salaries which are greater than one million(1e6) with NA. Also, the minimum wage is zero which is again exceptional. Prevailing wage cannot be zero because a foreign professional must receive at least some fixed prevailing wage to be eligible to file for an H1B visa. One may replace this with the median. Also, duplicate salaries exist. But this should not be a problem because different employers may have same prevailing wages.

YEAR:

There are no missing values in YEAR. Duplicate values do exist, but this should not be a problem because the given dataset consists of all the petitions for H1B by the employers in the years 2011 to 2016.

WORKSITE:

There are some missing locations in the dataset. Although data about the state or country is present, the exact location is missing. Duplicate locations exist but that should not be a problem because there may be different employers in the same area.

LON:

There are many of missing values in longitudes. Duplicate longitudes exist but that should not be a problem because there may be different employers in the same location.

LAT:

There are a lot of missing values in latitudes. Duplicate latitudes exist but that should not be a problem because there may be different employers in the same location.

Here the number of missing values of both latitudes and longitudes are same.

After cleaning and removing the duplicate entries, we have 2068225 records.

Data Preparation

After cleaning the data and removing the duplicates as in project 1, cluster analysis is performed.

Specific features in H1B dataset do not give much information, so we concentrated on some features and separately made into different datasets.

Important features for our dataset is:

- Prevailing Wage : Easily categorize salary ranges across application. One can easily pick up a range for deeper analysis.
- Latitude Longitude : Numeric values and provides additional insights on the location of applications.
- Aggregate of SOC_Name , Employer name, Full-time position , job title, and Prevailing wage : Comparing all these features by State.

Prevailing wages cluster has all the current wages. Location cluster has long, lat values.

We found out the optimal number of clusters for each technique using both Internal and External Validation (using ground truth)

We may need to scale the data as well to perform the analysis. This prevents one feature with a broad range to dominate the others for the distance calculation.

In this project, we are performing k-means and hierarchical clustering.

K-means is performed for the features PREVAILING_WAGE, FULL_TIME_POSITION, and SOC_NAME. Euclidean distance will not work for full-time position and SOC_NAME. Here we have mixed data, so convert nominal variables into dummy variables and then perform k-means.

Also, k-means is performed for the cluster which has prevailing wage. This makes it easier as one can easily differentiate the salaries and perform the analysis on that cluster whose range of salary Stakeholders are looking for.

We performed hierarchical clustering as well on the prevailing wage subset and the performance is compared.

Scale of Measurement of the features :

Table 2: Final features, scale and their distance measurements.

Feature	Scale	Distance measure
Prevailing wage	Numeric	Euclidean for K-means
Full time position	Numeric	Euclidean (dummy encoding, k-means), Gower for hierarchical
SOC name	Numeric	Euclidean (dummy encoding for k-means), Gower for hierarchical
Employer name	Numeric	Gower for hierarchical
Job title	Numeric	Gower for hierarchical

State	Factor	Gower for hierarchical
Latitude	Numeric	Euclidean(K-means)
Longitude	Numeric	Euclidean(K-means)

k-means clustering uses Euclidean distance by default. In our project, prevailing wage is numerical. Convert all other features to numeric from factors. We dummy encode them for using them in k-means clustering algorithm.

For hierarchical clustering we have used Gower distance as the dataset we used for clustering has mixed data.

Clustering for Removing Outliers.

We implemented K-means clustering to find groups of prevailing wages.

We started off by taking a subset of data of complete cases of Prevailing Wages. We later perform K-means clustering to find 4 clusters within the dataset.Upon analyzing the four clusters, we found 2 clusters with abnormally high salaries. Further analysis revealed that they are were rejected/denied candidates of H1b.

We can establish that these are outliers which disturb k-means clustering because outliers throw off the centers of clusters. We can remove them for further analysis of the data.

We again perform K-means Clustering on the Pre-Processed Prevailing wage data to further analyze the following:

Table 3 : Clustering using Prevailing Wage

Cluster	Description	Statistics
Medium to High Salaries	Salaries ranging from 80k to 285K	Min. 1st Qu. Median Mean 3rd Qu. Max. 83025 89461 98342 106470 113340 2850016
Low to Medium Salaries	Salaries randing from 0 – 83k	Min. 1st Qu. Median Mean 3rd Qu. Max. 0 51210 60133 59580 68994 83025
High to Very High	Salaries from 285k-1Million	Min. 1st Qu. Median Mean 3rd Qu. Max. 2897400 4180800 4991324 5637535 6594610 16566514
Greater than very high	Salaries greater than 3 Million	Min. 1st Qu. Median Mean 3rd Qu. Max. 33820800 50013600 56987840 55623005 64355200 66410240

Below are our findings for Top Employers within each Clusters:

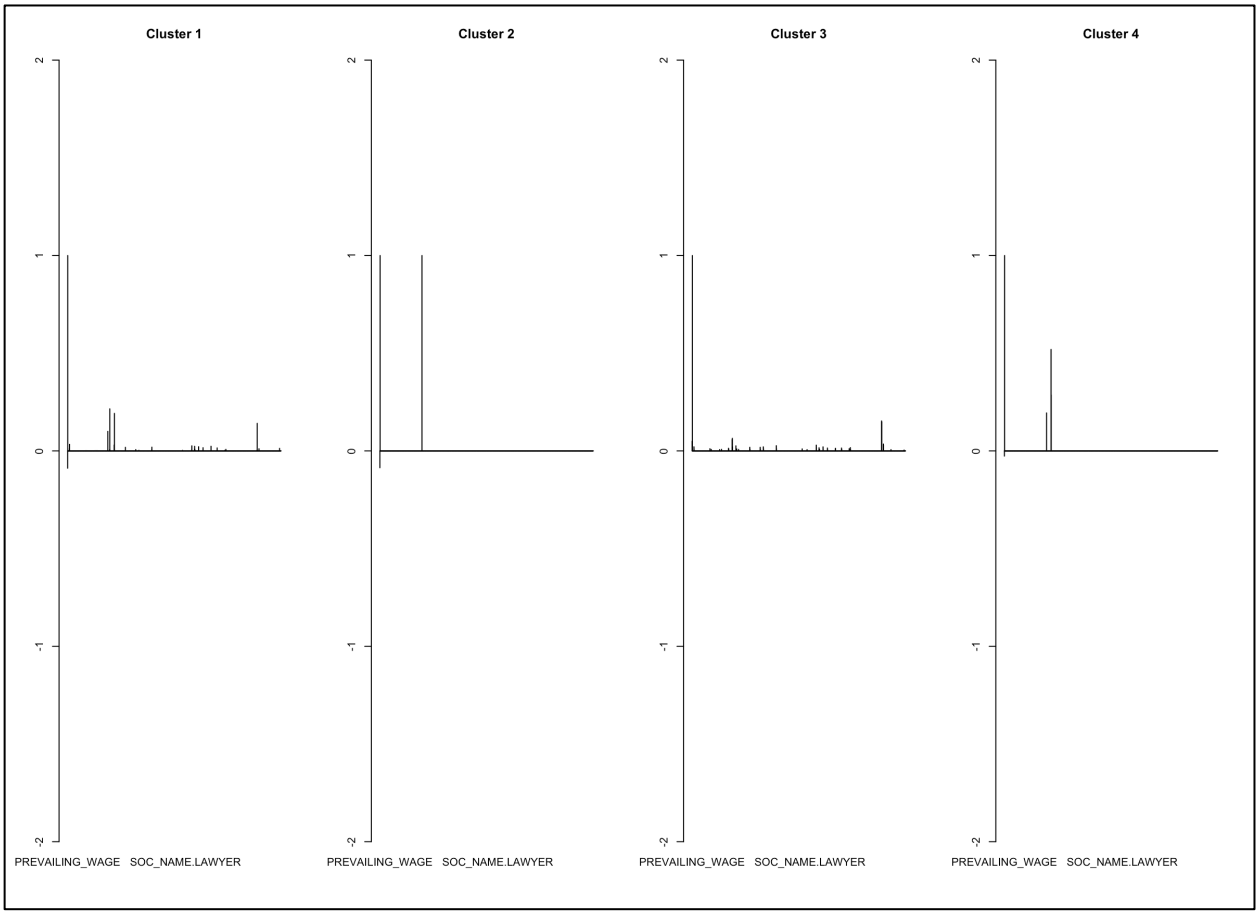


Figure 1: Cluster Visualization

Table 4: Employers with Low to Medium Salaries

Top Employers having low to Medium Salaries	**Frequency**
INFOSYS LIMITED	28295
TATA CONSULTANCY SERVICES LIMITED	17020
DELOITTE CONSULTING LLP	15045
WIPRO LIMITED	13793
IBM INDIA PRIVATE LIMITED	11120
ACCENTURE LLP	9857
LARSEN & TOUBRO INFOTECH LIMITED	8637

ERNST & YOUNG U.S. LLP	8000
HCL AMERICA, INC.	6187
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	6014

Table 5: Employers with Medium to high Salaries

Top Employers having Medium to High Salaries	**Frequency**
INFOSYS LIMITED	41649
MICROSOFT CORPORATION	24055
GOOGLE INC.	14340
WIPRO LIMITED	11786
INTEL CORPORATION	10576
TATA CONSULTANCY SERVICES LIMITED	9428
AMAZON CORPORATE LLC	8556
APPLE INC.	6959
HCL AMERICA, INC.	6667
ACCENTURE LLP	5942

Table 6: Employers with High to Very Salaries

Top Employers having High to Very High Salaries	**Frequency**
IBM INDIA PRIVATE LIMITED	5
ATLAS DEVELOPMENT CORPORATION	4
IBM CORPORATION	4
ABACUS TECHNICAL SERVICES, INC., A DIVISION OF PAR	3
FOUND ORGANIC USA, LLC	3
SATYAM COMPUTER SERVICES LTD	3
TATA CONSULTANCY SERVICES LIMITED	3
AMDOCS INC.	2
ANIMEX LATIN PRODUCTIONS INC.	2
BIO-OXFORD USA, INC.	2

Table 7 : Employers with the highest Salaries

Top Employers having the highest salaries	**Frequency**
REDIKER SOFTWARE, INC.	2
THE UNIVERSITY OF TENNESSEE HEALTH SCIENCE CENTER	2
ADVANCED PROFESSIONAL SERVICES INC	1
ALBERT EINSTEIN COLLEGE OF MEDICINE OF YESHIVA UNI	1
ASCENSUS	1
AUBURN UNIVERSITY	1
BEAUTY PLUS TRADING CO., INC.	1

BUCKEYE ELEMENTARY SCHOOL DISTRICT #33	1
BYTEWARE, INC.	1
CHALLENGER SPORTS CORP.	1

We also performed Hierarchical Clustering to compare clustering performance with K-means.

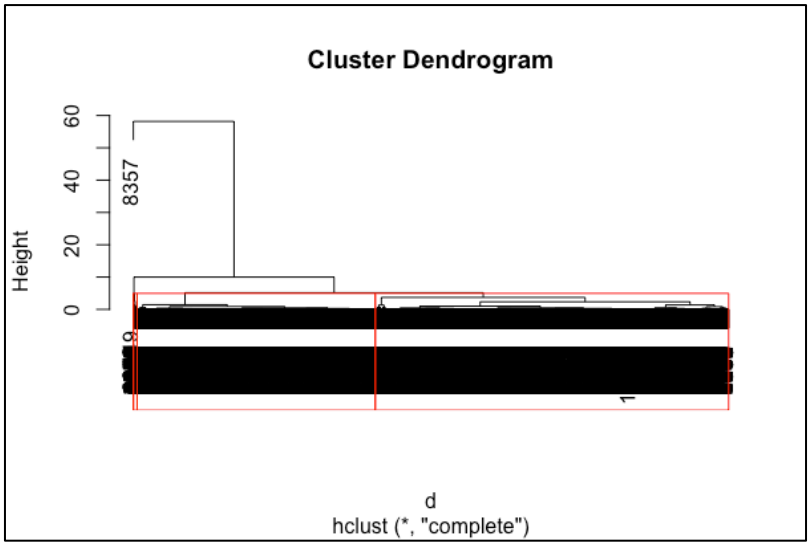


Figure 2 : Hierachical Clustering - 4 Clusters

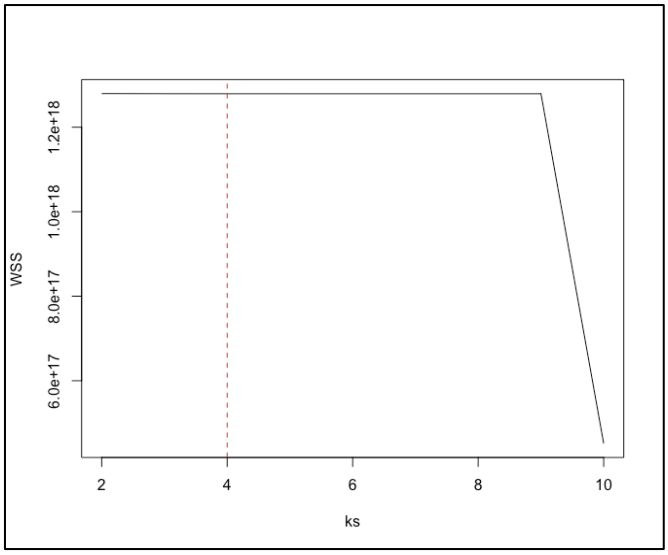


Figure 3 : . WSS for prevailing wages cluster

From the above figure, we can conclude that the optimal number of clusters for prevailing wages cluster are 4.

Now we perform internal validation and check the quality of the clusters. We use the Silhouette plot to do that. The following figure shows the Silhouette plot for our K-means clusters.

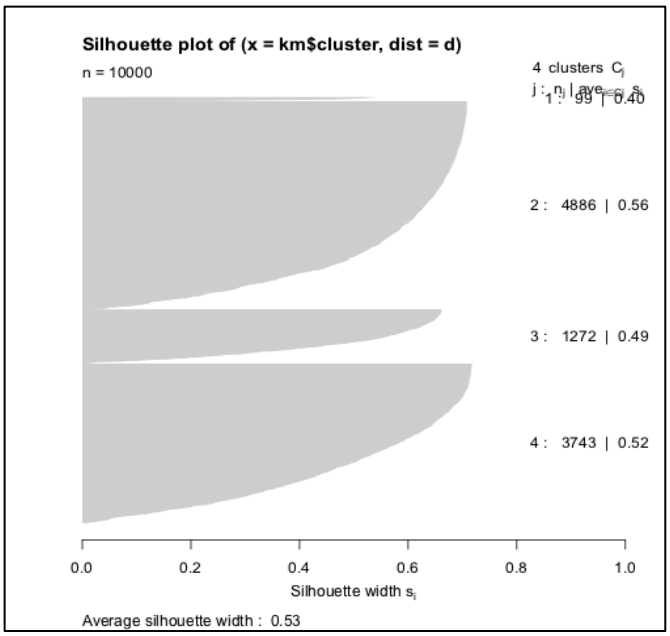


Figure 4 : Silhouette for kmeans for prevailing wage cluster

Similarly, we found out the silhouette width for hierarchical clusters as well,

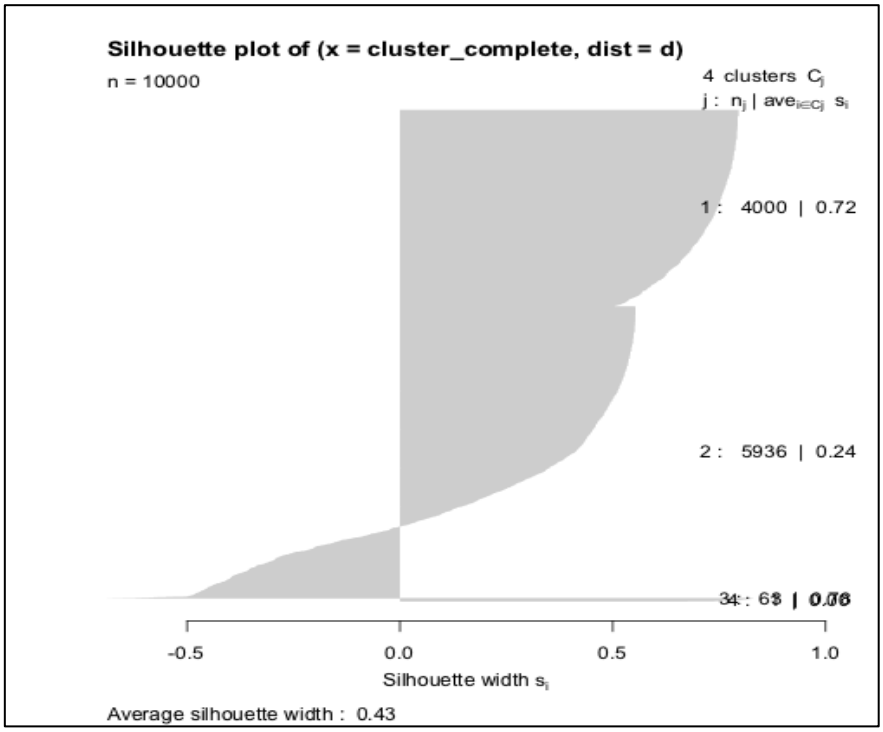


Figure 5 : Silhouette for complete Hierarchical Clustering for prevailing wage cluster

Average silhouette width for K-means is 0.53, which shows us that the value of silhouette co-efficient is closer to one. That is they are closer to records in their own cluster rather than to records in neighboring clusters. Fewer negative observations which means few are placed in the wrong cluster. Hence, we can say that our records are well clustered, i.e. the quality of clustering is good.

We also do hierarchical clustering for this data. We use the same features as we used for kmeans. We also keep the value of optimal clusters same. For hierarchical, average silhouette width is 0.43.

When compared to hierarchical, K-means has silhouette, which is closer to one.

Therefore, we can conclude that performing K-means clustering gives us much effective results.

Clustering for Location Co-ordinates:

Using the above pre-processed data, we cluster the location co-ordinates. K-means is the best way to cluster geo-coordinates given their nature.

We initially began dividing dataset into >8 parts given the clusters located outside of the USA states, but to avoid overfitting, we divide the region into four parts. We conducted K-means clustering using 4 cluster with the help of external knowledge about dividing the US Region in to four parts being:

1. Northwest Region
2. Southwest Region
3. Northeast Region
4. Middle-east Region (including parts of south)

Similarly, for the dataset Location cluster , to find the optimal number of cluster We plot the WSS chart which indicates that the optimal number of cluster is 4 :

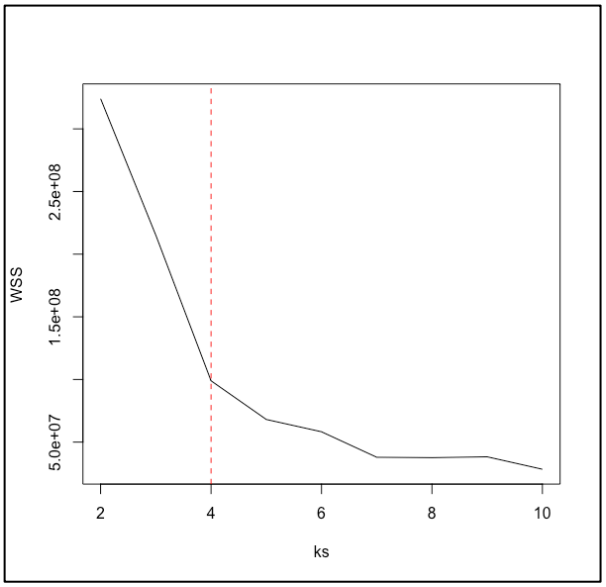


Figure 6 : WSS for Location cluster

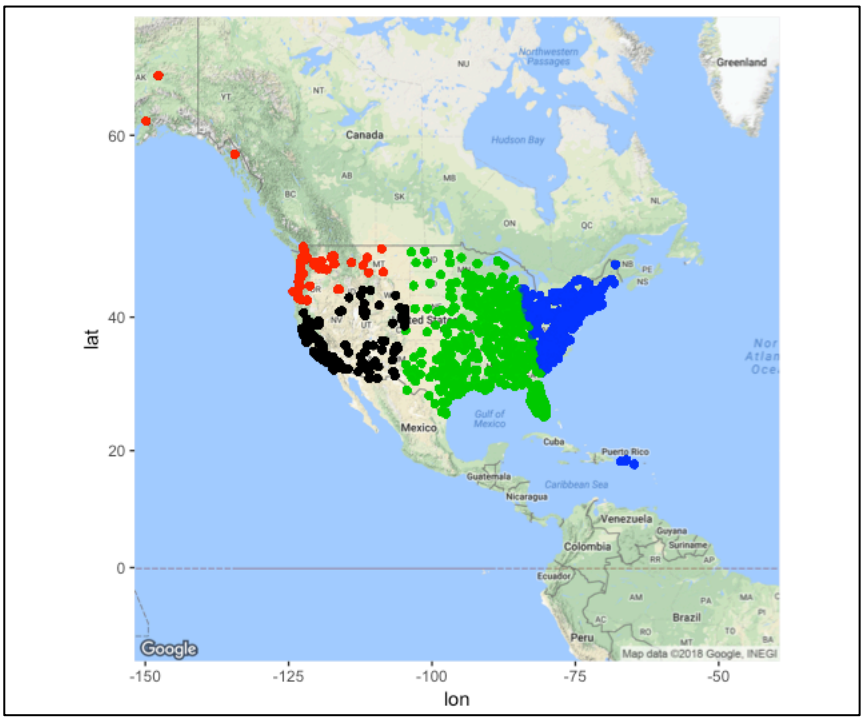


Figure 7: Location Clustering

Distinct cluster (with some external validations) were generated. We were able to divide the dataset into 4 parts by location with the help of k-means clustering. Below are our findings for the each Location Cluster :

North East Region Analysis:

Table 8 : Top Jobs in North-east Region

Top Jobs in the North - East Regions	Frequency
PROGRAMMER ANALYST	96368
SOFTWARE ENGINEER	26786
COMPUTER PROGRAMMER	23149
BUSINESS ANALYST	17245
SYSTEMS ANALYST	16010
SOFTWARE DEVELOPER	13364
COMPUTER SYSTEMS ANALYST	9582
SENIOR CONSULTANT	9349
CONSULTANT	9020
ASSOCIATE	8324

Table 9: Top Employers in North-east Region

Top Employers in the North - East Regions	Frequency
INFOSYS LIMITED	28295
TATA CONSULTANCY SERVICES LIMITED	17020
DELOITTE CONSULTING LLP	15045
WIPRO LIMITED	13793
IBM INDIA PRIVATE LIMITED	11120
ACCENTURE LLP	9857
LARSEN & TOUBRO INFOTECH LIMITED	8637
ERNST & YOUNG U.S. LLP	8000
HCL AMERICA, INC.	6187
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	6014

Table 10 : Median Salaries in North-east Region

STATE	PREVAILING_WAGE
CONNECTICUT	67413.00
DELAWARE	63232.00
DISTRICT OF COLUMBIA	66810.00
MAINE	60133.00
MARYLAND	63648.00
MASSACHUSETTS	69306.00
NEW HAMPSHIRE	62358.00
NEW JERSEY	63877.00
NEW YORK	65437.00
NORTH CAROLINA	54662.00

PENNSYLVANIA	60653.00
PUERTO RICO	42432.00
RHODE ISLAND	63149.00
VERMONT	58760.00
VIRGINIA	66622.00
WEST VIRGINIA	60174.00

West Region Analysis:

Table 11: Top Jobs in West Region

Job Title (West)	Frequency
SOFTWARE ENGINEER	51829
PROGRAMMER ANALYST	41427
SENIOR SOFTWARE ENGINEER	13942
SYSTEMS ANALYST	12819
COMPUTER PROGRAMMER	10124
TECHNOLOGY LEAD - US	9797
TECHNOLOGY ANALYST - US	8592
SOFTWARE DEVELOPER	8014
COMPUTER SYSTEMS ANALYST	7578
BUSINESS ANALYST	6263

Table 12 : Top Employers in West Region

Employers (West)	Frequency
INFOSYS LIMITED	41649
MICROSOFT CORPORATION	24055
GOOGLE INC.	14340
WIPRO LIMITED	11786
INTEL CORPORATION	10576
TATA CONSULTANCY SERVICES LIMITED	9428
AMAZON CORPORATE LLC	8556
APPLE INC.	6959
HCL AMERICA, INC.	6667
ACCENTURE LLP	5942
DELOITTE CONSULTING LLP	5910
QUALCOMM TECHNOLOGIES, INC.	5509
IBM INDIA PRIVATE LIMITED	4961
ORACLE AMERICA, INC.	4028
FACEBOOK, INC.	3846
QUALCOMM INCORPORATED	3690
LARSEN & TOUBRO INFOTECH LIMITED	3332
EBAY INC.	3328
YAHOO! INC.	3064
ERNST & YOUNG U.S. LLP	2744

Table 13 : Median Salaries in West Region

STATE	PREVAILING_WAGE
ALASKA	64802.50
ARIZONA	62275.00
CALIFORNIA	77563.00
HAWAII	54766.00
IDAHO	63502.00
MONTANA	49302.50
NEVADA	60757.00
NEW MEXICO	35000.00
OREGON	71536.00
UTAH	59654.00
WASHINGTON	81432.00
WYOMING	55983.50

Middle- East Region Analysis:

Table 14: Top Jobs in Middle-east Region

Job Title (Middle-East)	Frequency
PROGRAMMER ANALYST	67143
SOFTWARE ENGINEER	24613
COMPUTER PROGRAMMER	22769
SYSTEMS ANALYST	19059
SOFTWARE DEVELOPER	13077
COMPUTER SYSTEMS ANALYST	11580
BUSINESS ANALYST	9781
ASSISTANT PROFESSOR	9492
TECHNOLOGY ANALYST - US	7959
TECHNOLOGY LEAD - US	7874
SENIOR CONSULTANT	7344
CONSULTANT	6118
PROJECT MANAGER	5547
PHYSICAL THERAPIST	5377
DEVELOPER	5374
DATABASE ADMINISTRATOR	4675
SENIOR SOFTWARE ENGINEER	4343
COMPUTER PROGRAMMER ANALYST	4070
RESEARCH ASSOCIATE	3674
LEAD ENGINEER	3225

Table 15 : Employers in Middle East

Employers (Middle-East)	Frequency
INFOSYS LIMITED	37931
TATA CONSULTANCY SERVICES LIMITED	25678
WIPRO LIMITED	13393
IBM INDIA PRIVATE LIMITED	11127
ACCENTURE LLP	11012
DELOITTE CONSULTING LLP	9509
CAPGEMINI AMERICA INC	8291
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	5941
HCL AMERICA, INC.	5850
ERNST & YOUNG U.S. LLP	5231

Table 16 : Median Salaries in Middle East

STATE	PREVAILING_WAGE
ALABAMA	55028.50
ARKANSAS	54340.00
FLORIDA	57928.00
GEORGIA	61714.00
ILLINOIS	61464.00
INDIANA	57637.00
IOWA	58448.00
KENTUCKY	55848.00
MICHIGAN	61901.00
MINNESOTA	55804.00
MISSISSIPPI	45333.60
MISSOURI	63003.00
NA	38327.50
NORTH CAROLINA	65270.00
OHIO	60278.00
PENNSYLVANIA	57325.00
SOUTH CAROLINA	57117.00
TENNESSEE	58448.00
VIRGINIA	51979.00
WASHINGTON	64074.50
WEST VIRGINIA	52083.00
WISCONSIN	58968.00

South-East Region Analysis:

Table 17 : Top jobs in South East Region

Job Title (South-East)	Frequency
PROGRAMMER ANALYST	39589
SOFTWARE ENGINEER	16327
COMPUTER PROGRAMMER	13219
SYSTEMS ANALYST	12990
SOFTWARE DEVELOPER	7487
BUSINESS ANALYST	5585
COMPUTER SYSTEMS ANALYST	5517
ASSISTANT PROFESSOR	5480
TECHNOLOGY LEAD - US	4666
TECHNOLOGY ANALYST - US	4141

Table 18 : Top Employers in South East Region

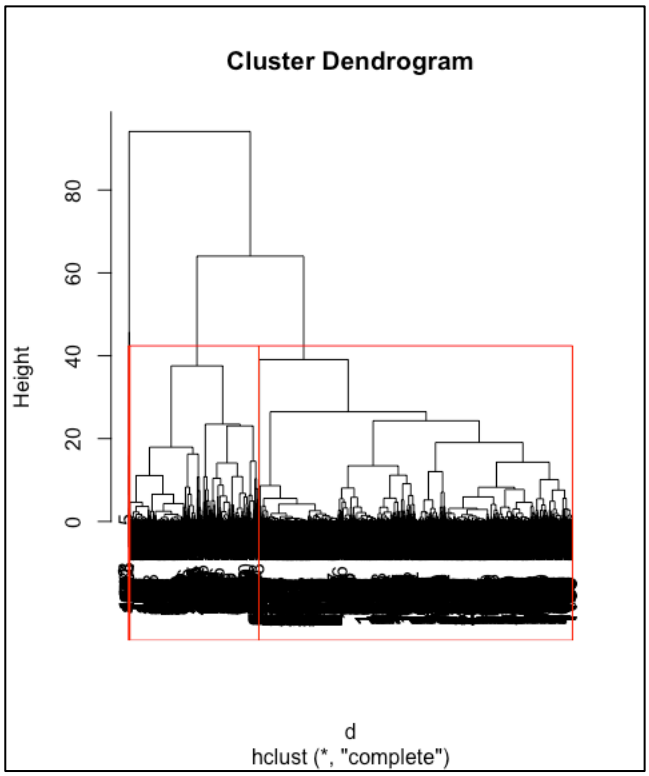
Employers(South-East)	Frequency
INFOSYS LIMITED	22358
TATA CONSULTANCY SERVICES LIMITED	12141
WIPRO LIMITED	8722
IBM INDIA PRIVATE LIMITED	6730
ACCENTURE LLP	6414
DELOITTE CONSULTING LLP	5532
HCL AMERICA, INC.	3784
COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	2690
TECH MAHINDRA (AMERICAS),INC.	2625
LARSEN & TOUBRO INFOTECH LIMITED	2339

Table 19 : Median Salaries

STATE	PREVAILING_WAGE
ALABAMA	55028.50
ARKANSAS	54340.00
FLORIDA	57928.00
GEORGIA	61714.00
ILLINOIS	61464.00
INDIANA	57637.00
IOWA	58448.00
KENTUCKY	55848.00
MICHIGAN	61901.00
MINNESOTA	55804.00
MISSISSIPPI	45333.60
MISSOURI	63003.00
NORTH CAROLINA	65270.00

OHIO	60278.00
PENNSYLVANIA	57325.00
SOUTH CAROLINA	57117.00
TENNESSEE	58448.00
VIRGINIA	51979.00
WASHINGTON	64074.50
WEST VIRGINIA	52083.00
WISCONSIN	58968.00

We also performed hierarchical clustering on the above subset:



Aggregate Subset:

The subset agg_cluster has the information about each state and their corresponding employer_name, soc_name, job_title, prevailing_wage, full_time_position.

Now, to calculate the optimal number of clusters for the dataset agg_clust for hierarchical clustering ASW method is used. Below plot shows ASW for agg_cluster dataset.

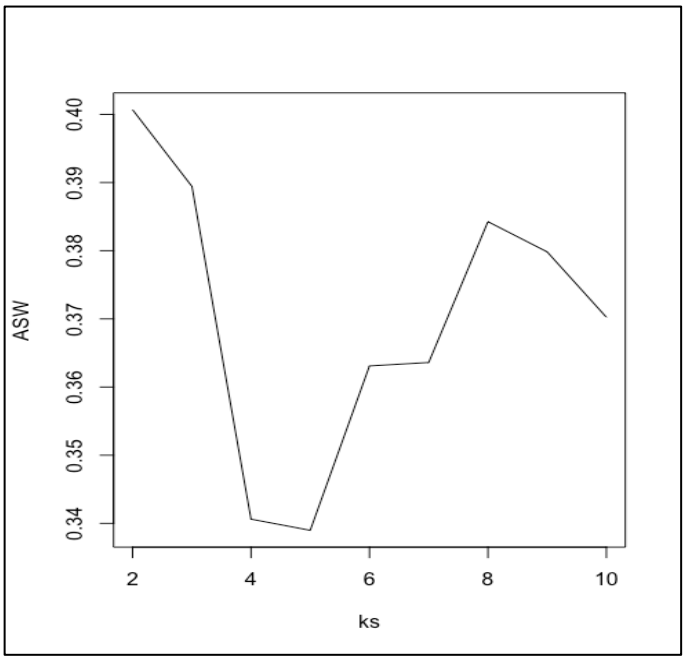


Figure 8 : ASW for agg_cluster.

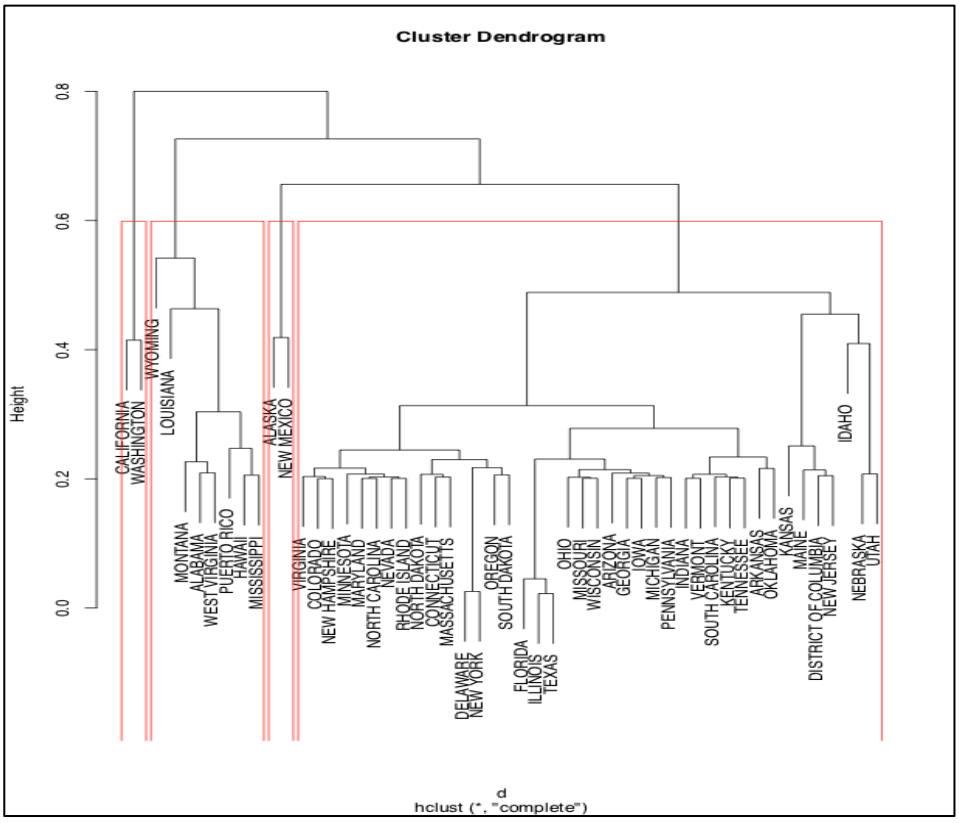


Figure 9 : hClust for agg_cluster

Here for hierarchical clustering, we can find the similarity between the states based on the features in agg_cluster.

From the above figure, we can find that the states which are grouped at very less height are much similar, whereas the states which are grouped at more considerable height are dissimilar. Consider the states Delaware and Newyork from the above plot, these are grouped at a smaller height(approximately 0.004) from which we can conclude that the states Delaware and Newyork has similar wages, most of the employers in these states are same, job_title of the applicants is similar to these states, full-time position status is same for the majority of the applicants in these states.

Also, we divided the entire plot into four clusters. The first cluster has the states Washington and California at a height 0.4.

Both these states have common soc_name, similar wages, and same full-time position.

The second cluster has eight states. For all the eight states, job title is an assistant professor. In the second cluster, we have at least two features in common. At height 0.2, we have Mississipi and Hawaii, both these states have similar wages, and rest of the features apart from employer name are exactly same for both the countries. As these two states are grouped at smaller heights, we may assume they have similar features.

Now for the same cluster consider the states Mississippi and Wyoming, these two states are grouped at largest height in the cluster. So, these states have only two features in common, job title and status of full time.

From the above discussion, we can conclude that in the second cluster all the states have at least two features in common and maximum of 4 elements in common.

The third cluster has Alaska and Mexico. Both these states have precisely two features soc name; full-time status is same. Job title and prevailing wages are very similar to each other. Both these states are grouped at the height 0.4.

The fourth cluster has the rest 40 states. Illinois and Texas are grouped at small height, so looking at the behavior of these states we found that prevailing wages are similar and the rest of the four features are exactly same.

Now, look at the behavior of those states which are grouped at more considerable heights. Consider Texas and Idaho; both these states have two features in common, job title, full-time position status, and similar wages.

From the above discussion, we can conclude that this cluster of 40 states has at least two features in common and a maximum of 4 elements in common. All the states in this cluster have the same job title and full-time status.

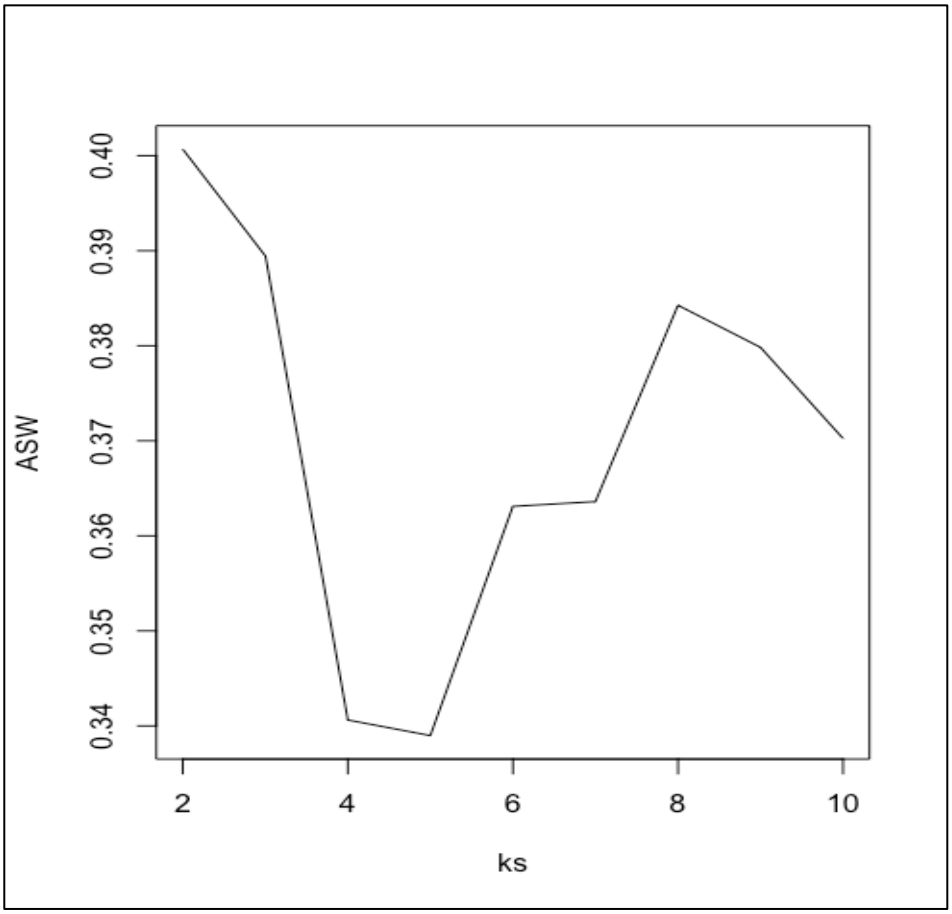


Figure 10 : ASW for agg_cluster.

External Validations:

For performing external validation, we use sampled data of 10000 records from the entire dataset with features prevailing wage, soc name, full time position, employer name, job title. We use state as the ground truth.

We then find optimal number of clusters which is 4 using ASW, then we perform k-means and Hierarchical clustering of the dataset. Then we perform external validation for our data.

	Entropy	Purity
K-means	1.025	0.08
Hierarchical	0.046	0.31

From above, we can find that entropy for K-means is greater than Hierarchical. Entropy is a negative measure. So, based on entropy we may say that Hierarchical clustering is the better way for clustering.

Also, consider the purity, it is higher for hierarchical. From above discussion we may conclude that for the dataset agg_cluster, hierarchical clustering is better approach.

Conclusion

The data of the H1B had useful variables and clusters which influence the final decisions.

The following conclusion is made from mining the data in this project:

- The results from our clusters support the findings of previous project.
- For every data set, we compared the clustering algorithms and found out the best clustering algorithm to be used.
- For every clustering algorithm, we found optimal number of clusters using ASW and WSS methods.
- Using internal validations, such as silhouette we found that for the prevailing wage cluster, it is better option to use K-means clustering.
- Top Employers with ideal Salary Ranges include Infosys, Microsoft and IBM.
- For Location cluster, it is better choice to use K-means clustering algorithm.

- Program Analysts are top jobs applying for H1b across most of the US regions followed by Software Engineers.
- For agg_cluster hierarchical clustering is best algorithm.
- For agg_cluster, when we performed hierarchical clustering, in each cluster at least two features are exactly same, also each cluster has same Job title.

References

1. https://www.foreignlaborcert.doleta.gov/docs/Performance_Data/Disclosure/FY16Q2/PERM_FY16_Record_Layout.pdf
2. https://www.uscis.gov/working-united-states/permanent-workers
3. http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/
4. https://www.kaggle.com/nsharan/h-1b-visa
5. https://stackoverflow.com/questions/36568070/extract-year-from-date
6. https://www.kaggle.com/jboysen/us-perm-visas
7. https://amunategui.github.io/ggmap-example/
8. https://dabblingwithdata.wordpress.com/2016/10/10/clustering-categorical-data-with-r/
9. https://rpubs.com/gabrielmartos/ClusterAnalysis
10. http://www.kimberlycoffey.com/blog/2016/8/k-means-clustering-for-customer-segmentation