Samuel L. Peoples

STMATH 493

Visualization 4 : Five Years of American Opportunity

Description:

The U.S. Department of Labor issues certification which allows an employer to hire a foreign worker under permanent visa status. Before the employer may submit an immigration petition to the Department of Homeland Security U.S. Citizenship and Immigrations Services, the employer must obtain a certified labor certification application from the DOL Employment and Training Administration. The DOL must certify to the USCIS that there are not sufficient local workers able, willing, qualified, and available to accept the job opportunity in the area of intended employment and that the employment of the foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. Citizens. The dataset includes permanent visa application decisions from 2012 to 2017.

Question:

For which industries, and from which countries are the most permanent visa applications received?

Data Source:

<https://www.kaggle.com/jboysen/us-perm-visas>

Data Cleaning Process:

The dataset that we retrieved had many issues. A majority of the feature columns were populated with very little, if any data, while some were paired with misspelled columns, and others were redundant! The case\_no and case\_number were perfectly overlapping, as if the variable name was changed after a specific date and the tables were joined. The same process was for country\_of\_citizenship and country\_of\_citzenship. The class\_of\_admission needed no change. The decision\_date was left unchanged, but we did note missing data in APR-SEP 2012. The case\_status was separated from “Certified-Expired” and “Certified”, and created a new column “expired”, which is populated with TRUE or FALSE. The employer\_name was wrangled to proper-casing, as was the employer\_city, and employer\_state. The job\_level was left unchanged, and zeros were populated for null values, while the salary was left nearly unchanged, except for values less than 500, which were treated as hourly wages, and converted to yearly wages.

A large portion of the preprocessing included converting the job\_title to a manageable set of nine economic sectors. Information Technology, Finance, Academia & Sciences, Retail, Management, Industrial, Healthcare, Religion, Law. After filtering by keywords, roughly ten thousand entries were left unfilled, and these were manually selected, capturing between 200-500 entries at a time.

* Information Technology: All job titles including key words such as software, database, or computer. Specific technologies such as languages and platforms are searched for as well.
* Finance: Any keywords associated with money, accounting, finances, bookkeeping, or banking.
* Academia & Sciences: Any keywords associated with instruction, teaching, education, colleges & universities, specific hard sciences (physics, chemistry, biology), and research.
* Retail: Any keywords associated with clerical work, sales, secretaries, food work, or customer service.
* Management: Any keywords associated with business, management, leadership, or senior executive roles.
* Industrial: Any keywords associated with manufacturing, construction, civics, or planning.
* Healthcare: Any keywords associated with clinical, medical, surgical, nursing, doctors, health trades, and medical research.
* Religion: Any keywords associated with pastoral services of various denominations.
* Law: Any keywords associated with administration or practice in a law firm, or in public policy.

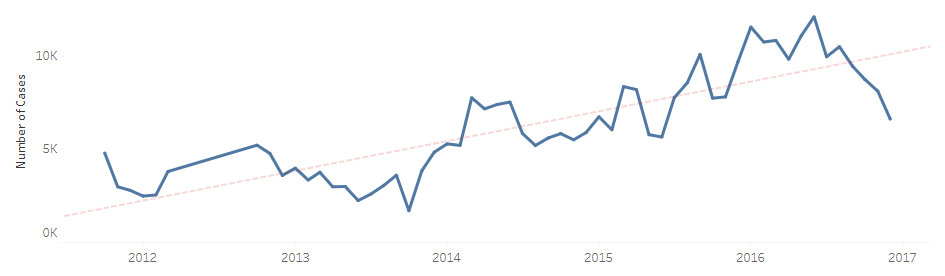
Wrangled:

|  |  |
| --- | --- |
| Column Name | Distribution |
| case\_no |  |
| citizenship |  |
| class\_of\_admission |  |
| decision\_date |  |
| case\_status |  |
| expired |  |
| employer\_name |  |
| employer\_city |  |
| employer\_state |  |
| economic\_sector |  |
| Job\_level |  |
| salary |  |

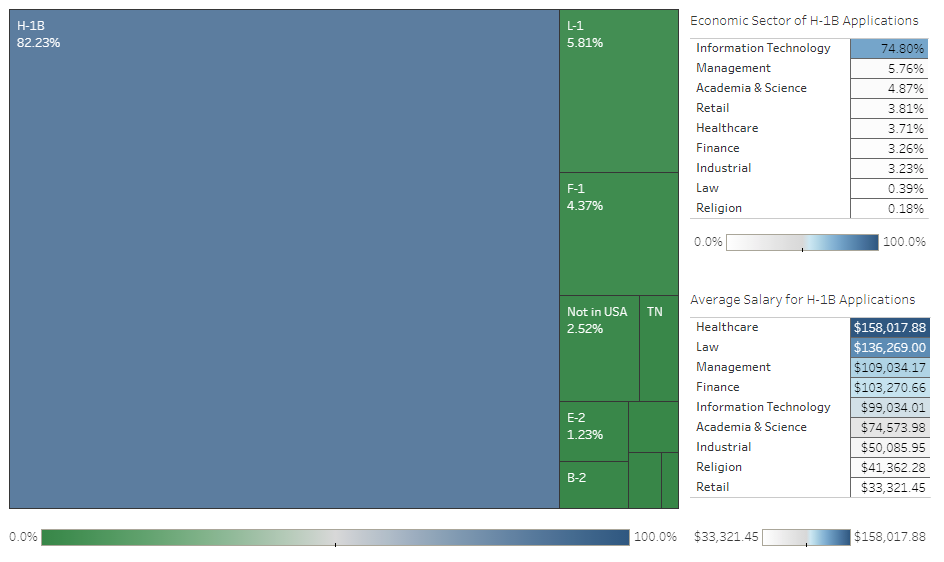
Visualizations:

Our group was able to come together and make very similar visualizations, ultimately deciding on a workflow that turned out to be very effective.

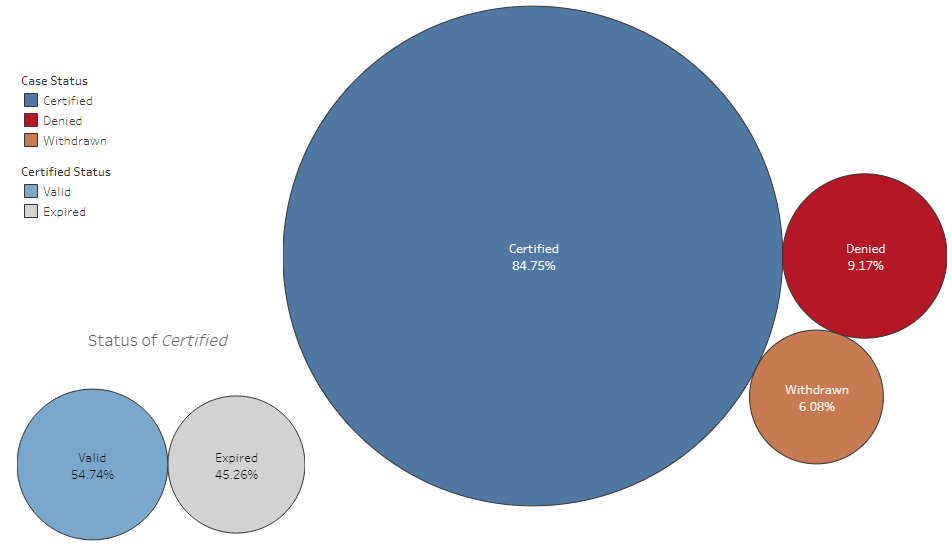
We first built the Number of Cases Over Time view, mostly to get a good feeling for the direction we wanted to go with the rest of the project. We wanted to keep things simple, so we made sure to remove any unnecessary information and lines, while also adding a trend line, letting the user know that the focus is on the upward trend over time. We noticed that there was missing data in 2012 between April and September, but ultimately decided to move forward with the analysis.



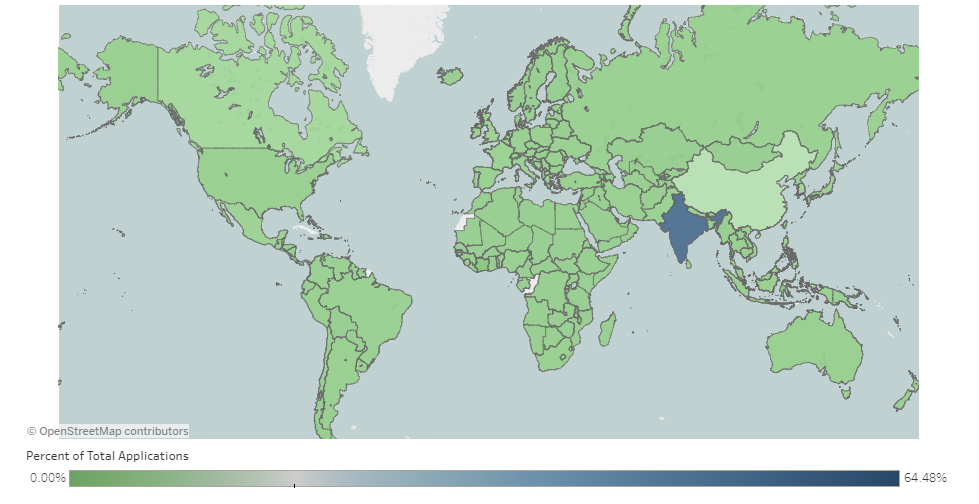
We toyed with different ideas for how to show the Distribution of Permanent Visa Applications, and felt that the area graph draws the viewer’s attention to the vast supermajority of H-1B visas in the dataset. We wanted to include further information about the Economic Sector and Average Salaries of these H-1B visa applicants, and had a hard time reducing the speeded classification. The colors we had chosen (green and blue) were forcing people to mis-associate different aspects of the visualization, and ultimately chose to use and white-to-blue gradient.



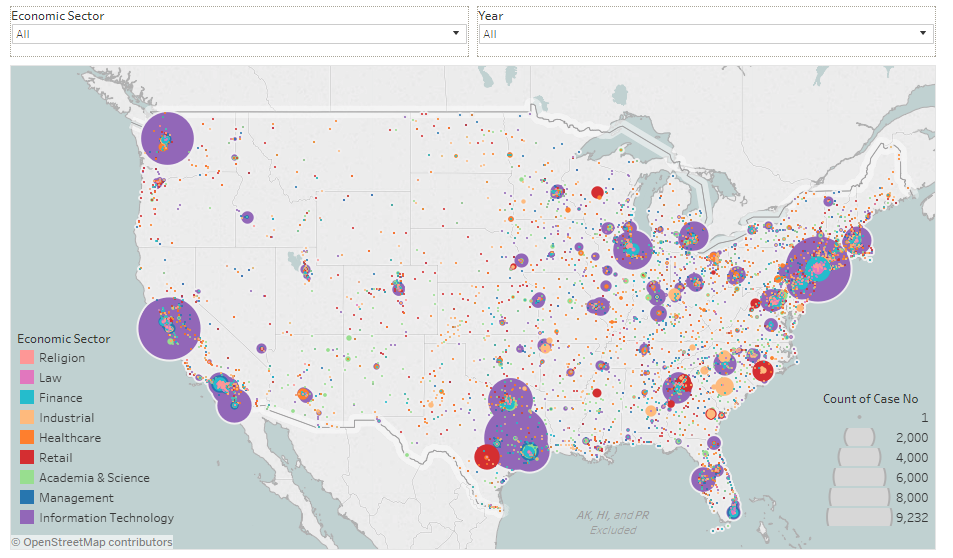
We also felt it was important for the viewer to see information about the frequency with which the H-1B applications were Certified, Denied, or Withdrawn. Knowing that pie charts are despicable, and that we had already used an area chart, we debated the choice between a bar chart and a bubble chart. We decided on the latter with a desire to “zoom” into a separate bubble chart of the Valid and Expired subsets of the Certified bubble, but couldn’t get the connecting lines (nifty little slashes) to remain dynamic with any changes in the view. So we opted for a similar coloring.

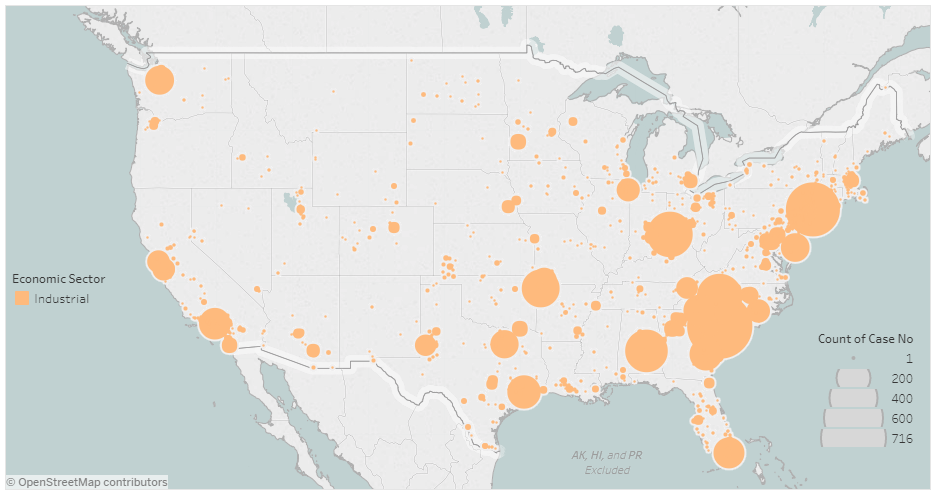


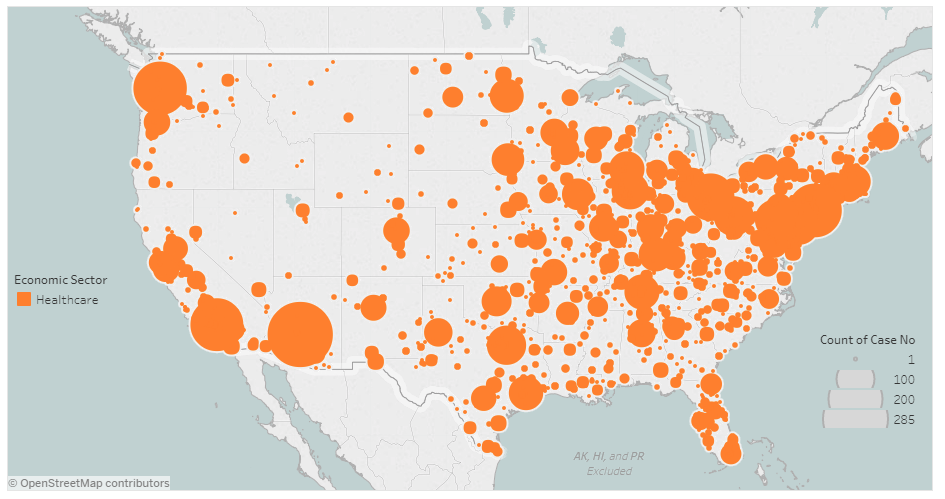
We finally wanted to create two different maps, displaying where H-1B applicants are coming from, and going to. We chose to modify the World Map to remove any markings, and to use the naturally-colored sea. We colored low-value areas in green, and drew the attention to darker-blue areas, trying to keep things simple.



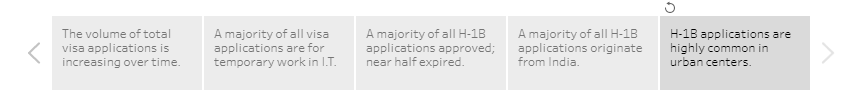
The US map was sorted in an ascending fashion so to keep the largest bubbles on the bottom, and as much of the variation can be seen. The Economic Sectors are parameterized, allowing different industries to be analyzed in more detail. We opted to prevent the user from manipulating the maps, which resulted in the exclusion of the visualization of Hawaii, Alaska, and Puerto Rico. There was very insignificant data for the regions, so we felt comfortable making this decision.







Following this parameterization, we decided to extend the capability for the year across all our sheets, and began discussing the questions, and slide labels for our storyboard. After bouncing ideas back and forth, we decided to label our slides with generalized conclusions which could direct the expectations of the specific slide.



Data cleaning:

|  |  |  |
| --- | --- | --- |
| Column Name | Distribution | Changes |
| add\_these\_pw\_job\_title\_9089 |  | Drop – Insufficient Data |
| agent\_city |  | Drop – Insufficient Data |
| agent\_firm\_name |  | Drop – Insufficient Data |
| agent\_state |  | Drop – Insufficient Data |
| application\_type |  | Drop – Insufficient Data |
| case\_no |  | Combine with  case\_number |
| case\_number |  | Combine with  case\_no |
| case\_received\_date |  | Drop – Insufficient Data |
| case\_status |  | Keep - Wrangle |
| class\_of\_admission |  | Keep - Wrangle |
| country\_of\_citizenship |  | Combine with  country\_of\_citzenship |
| country\_of\_citzenship |  | Combine with country\_of\_citizenship |
| decision\_date |  | Keep |
| employer\_address\_1 |  | Keep - Wrangle |
| employer\_address\_2 |  | Drop – Insufficient Data |
| employer\_city |  | Keep - Wrangle |
| employer\_country |  | Drop – Insufficient Data |
| employer\_decl\_info\_title |  | Drop – Insufficient Data |
| employer\_name |  | Keep - Wrangle |
| employer\_num\_employees |  | Drop – Insufficient Data |
| employer\_phone |  | Drop – Insufficient Data |
| employer\_phone\_ext |  | Drop – Insufficient Data |
| employer\_postal\_code |  | Keep |
| employer\_state |  | Keep - Wrangle |
| employer\_yr\_estab |  | Drop – Insufficient Data |
| foreign\_worker\_info\_alt\_edu\_experience |  | Drop – Insufficient Data |
| foreign\_worker\_info\_birth\_country |  | Drop – Insufficient Data |
| foreign\_worker\_info\_city |  | Drop – Insufficient Data |
| foreign\_worker\_info\_education |  | Drop – Insufficient Data |
| foreign\_worker\_info\_education\_other |  | Drop – Insufficient Data |
| foreign\_worker\_info\_inst |  | Drop – Insufficient Data |
| foreign\_worker\_info\_major |  | Drop – Insufficient Data |
| foreign\_worker\_info\_postal\_code |  | Drop – Insufficient Data |
| foreign\_worker\_info\_rel\_occup\_exp |  | Drop – Insufficient Data |
| foreign\_worker\_info\_req\_experience |  | Drop – Insufficient Data |
| foreign\_worker\_info\_state |  | Drop – Insufficient Data |
| foreign\_worker\_info\_trainin\_comp |  | Drop – Insufficient Data |
| foreign\_worker\_ownership\_interest |  | Drop – Insufficient Data |
| foreign\_worker\_yr\_rel\_edu\_completed |  | Drop – Insufficient Data |
| fw\_info\_alt\_edu\_experience |  | Drop – Insufficient Data |
| fw\_info\_birth\_country |  | Drop – Insufficient Data |
| fw\_info\_education\_other |  | Drop – Insufficient Data |
| fw\_info\_postal\_code |  | Drop – Insufficient Data |
| fw\_info\_rel\_occup\_exp |  | Drop – Insufficient Data |
| fw\_info\_req\_experience |  | Drop – Insufficient Data |
| fw\_info\_training\_comp |  | Drop – Insufficient Data |
| fw\_info\_yr\_rel\_edu\_completed |  | Drop – Insufficient Data |
| fw\_ownership\_interest |  | Drop – Insufficient Data |
| ji\_foreign\_worker\_live\_on\_premises |  | Drop – Insufficient Data |
| ji\_fw\_live\_on\_premises |  | Drop – Insufficient Data |
| ji\_live\_in\_dom\_svc\_contract |  | Drop – Insufficient Data |
| ji\_live\_in\_domestic\_service |  | Drop – Insufficient Data |
| ji\_offered\_to\_sec\_j\_foreign\_worker |  | Drop – Insufficient Data |
| ji\_offered\_to\_sec\_j\_fw |  | Drop – Insufficient Data |
| job\_info\_alt\_cmb\_ed\_oth\_yrs |  | Drop – Insufficient Data |
| job\_info\_alt\_combo\_ed |  | Drop – Insufficient Data |
| job\_info\_alt\_combo\_ed\_exp |  | Drop – Insufficient Data |
| job\_info\_alt\_combo\_ed\_ other |  | Drop – Insufficient Data |
| job\_info\_alt\_field |  | Drop – Insufficient Data |
| job\_info\_alt\_field\_name |  | Drop – Insufficient Data |
| job\_info\_alt\_occ |  | Drop – Insufficient Data |
| job\_info\_alt\_occ\_job\_title |  | Drop – Insufficient Data |
| job\_info\_alt\_occ\_num\_months |  | Drop – Insufficient Data |
| job\_info\_education |  | Drop – Insufficient Data |
| job\_info\_education\_other |  | Drop – Insufficient Data |
| job\_info\_experience |  | Drop – Insufficient Data |
| job\_info\_experience\_num\_months |  | Drop – Insufficient Data |
| job\_info\_foreign\_ed |  | Drop – Insufficient Data |
| job\_info\_foreign\_lang\_req |  | Drop – Insufficient Data |
| job\_info\_job\_req\_normal |  | Drop – Insufficient Data |
| job\_info\_job\_title |  | Drop – Insufficient Data |
| job\_info\_major |  | Drop – Insufficient Data |
| job\_info\_training |  | Drop – Insufficient Data |
| job\_info\_training\_field |  | Drop – Insufficient Data |
| job\_info\_training\_num\_months |  | Drop – Insufficient Data |
| job\_info\_work\_city |  | Drop - Redundant |
| job\_info\_work\_postal\_code |  | Drop – Insufficient Data |
| job\_info\_work\_state |  | Drop - Redundant |
| naics\_2007\_us\_code |  | Drop – Insufficient Data |
| naics\_2007\_us\_title |  | Drop – Insufficient Data |
| naics\_code |  | Drop – Insufficient Data |
| naics\_title |  | Drop – Insufficient Data |
| naics\_us\_code |  | Drop – Insufficient Data |
| naics\_us\_code\_2007 |  | Drop – Insufficient Data |
| naics\_us\_title |  | Drop – Insufficient Data |
| naics\_us\_title\_2007 |  | Drop – Insufficient Data |
| orig\_case\_no |  | Drop – Insufficient Data |
| orig\_file\_date |  | Drop – Insufficient Data |
| preparer\_info\_emp\_completed |  | Drop – Insufficient Data |
| preparer\_info\_title |  | Drop – Insufficient Data |
| pw\_amount\_9809 |  | Drop - Redundant |
| pw\_determ\_date |  | Drop – Insufficient Data |
| pw\_expire\_date |  | Drop – Insufficient Data |
| pw\_job\_title\_908 |  | Drop – Insufficient Data |
| pw\_job\_title\_9089 |  | Keep – Wrangle : Economic Sector |
| pw\_level\_9089 |  | Keep - Wrangle |
| pw\_soc\_code |  | Drop – Unrelated Data |
| pw\_soc\_title |  | Drop – Unrelated Data |
| pw\_source\_name\_9089 |  | Drop – Unrelated Data |
| pw\_source\_name\_other\_9089 |  | Drop – Insufficient Data |
| pw\_track\_num |  | Drop – Insufficient Data |
| pw\_unit\_of\_pay\_9089 |  | Drop – Unrelated Data |
| rec\_info\_barg\_rep\_notified |  | Drop – Insufficient Data |
| recr\_info\_barg\_rep\_notified |  | Drop – Insufficient Data |
| recr\_info\_coll\_teach\_comp\_proc |  | Drop – Insufficient Data |
| recr\_info\_coll\_univ\_teacher |  | Drop – Insufficient Data |
| recr\_info\_employer\_rec\_payment |  | Drop – Insufficient Data |
| recr\_info\_first\_ad\_start |  | Drop – Insufficient Data |
| recr\_info\_job\_fair\_from |  | Drop – Insufficient Data |
| recr\_info\_job\_fair\_to |  | Drop – Insufficient Data |
| recr\_info\_on\_campus\_recr\_from |  | Drop – Insufficient Data |
| recr\_info\_on\_campus\_recr\_to |  | Drop – Insufficient Data |
| recr\_info\_pro\_org\_advert\_from |  | Drop – Insufficient Data |
| recr\_info\_pro\_org\_advert\_to |  | Drop – Insufficient Data |
| recr\_info\_prof\_org\_advert\_from |  | Drop – Insufficient Data |
| recr\_info\_prof\_org\_advert\_to |  | Drop – Insufficient Data |
| recr\_info\_professional\_occ |  | Drop – Insufficient Data |
| recr\_info\_radio\_tv\_ad\_from |  | Drop – Insufficient Data |
| recr\_info\_radio\_tv\_ad\_to |  | Drop – Insufficient Data |
| recr\_info\_second\_ad\_start |  | Drop – Insufficient Data |
| recr\_info\_sunday\_newspaper |  | Drop – Insufficient Data |
| recr\_info\_swa\_job\_order\_end |  | Drop – Insufficient Data |
| recr\_info\_swa\_job\_order\_start |  | Drop – Insufficient Data |
| refile |  | Drop – Insufficient Data |
| ri\_1st\_ad\_newspaper\_name |  | Drop – Insufficient Data |
| ri\_2nd\_ad\_newspaper\_name |  | Drop – Insufficient Data |
| ri\_2nd\_ad\_newspaper\_or\_journal |  | Drop – Insufficient Data |
| ri\_campus\_placement\_from |  | Drop – Insufficient Data |
| ri\_campus\_placement\_to |  | Drop – Insufficient Data |
| ri\_coll\_tch\_basic\_process |  | Drop – Insufficient Data |
| ri\_coll\_teach\_pro\_jnl |  | Drop – Insufficient Data |
| ri\_coll\_teach\_select\_date |  | Drop – Insufficient Data |
| ri\_employee\_referral\_prog\_from |  | Drop – Insufficient Data |
| ri\_employee\_referral\_prog\_to |  | Drop – Insufficient Data |
| ri\_employer\_web\_post\_from |  | Drop – Insufficient Data |
| ri\_employer\_web\_post\_to |  | Drop – Insufficient Data |
| ri\_job\_search\_website\_from |  | Drop – Insufficient Data |
| ri\_job\_search\_website\_to |  | Drop – Insufficient Data |
| ri\_layoff\_in\_past\_six\_months |  | Drop – Insufficient Data |
| ri\_local\_ethnic\_paper\_from |  | Drop – Insufficient Data |
| ri\_local\_ethnic\_paper\_to |  | Drop – Insufficient Data |
| ri\_posted\_notice\_at\_worksite |  | Drop – Insufficient Data |
| ri\_pvt\_employment\_firm\_from |  | Drop – Insufficient Data |
| ri\_pvt\_employment\_firm\_to |  | Drop – Insufficient Data |
| ri\_us\_workers\_considered |  | Drop – Insufficient Data |
| schd\_a\_sheepherder |  | Drop – Insufficient Data |
| us\_economic\_sector |  | Drop – Insufficient Data |
| wage\_offer\_from\_9089 |  | Drop – Insufficient Data |
| wage\_offer\_to\_9089 |  | Drop – Insufficient Data |
| wage\_offer\_unit\_of\_pay\_9089 |  | Drop – Insufficient Data |
| wage\_offered\_from\_9089 |  | Keep – Wrangle |
| wage\_offered\_to\_9089 |  | Drop –  Unrelated Data |
| wage\_offered\_unit\_of\_pay\_9089 |  | Drop – Unrelated Data |