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:

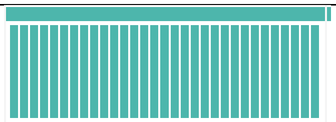
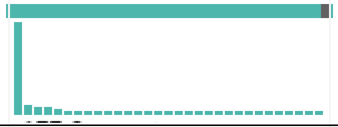
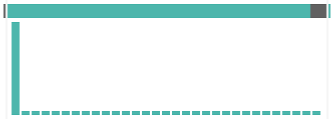
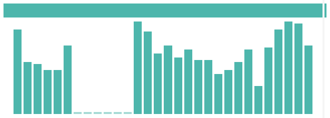
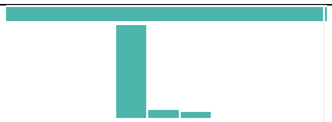
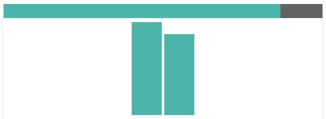
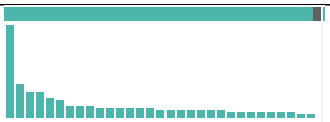
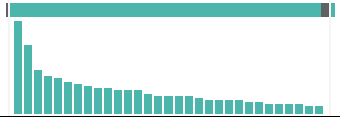
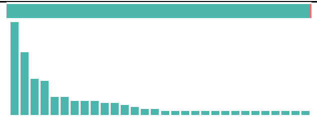
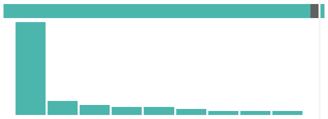
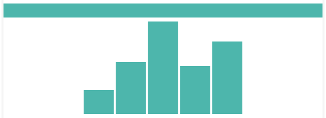
The dataset that we retrieved had many issues. A majority of the feature columns were populated with very little, if any data, while some were paired with misspelled columns, and others were redundant! The case_no and case_number were perfectly overlapping, as if the variable name was changed after a 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:

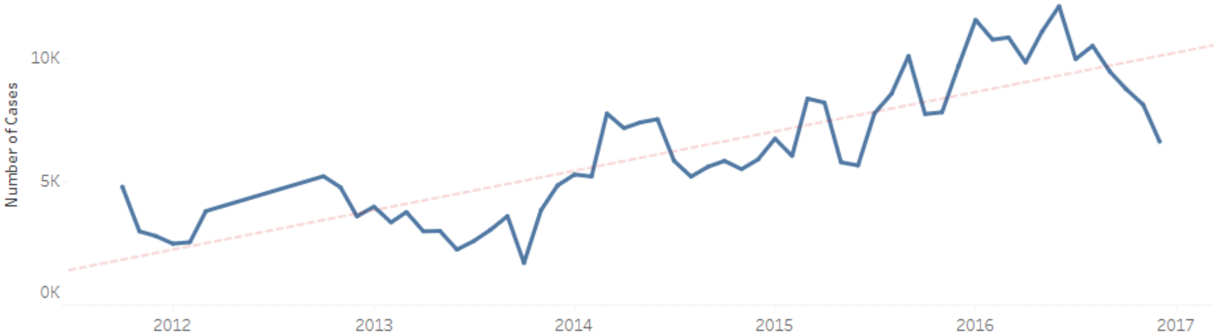
Column Name	Distribution
case_no	
citizenship	
class_of_admission	
decision_date	
case_status	
expired	
employer_name	
employer_city	
employer_state	
economic_sector	
Job_level	

salary	

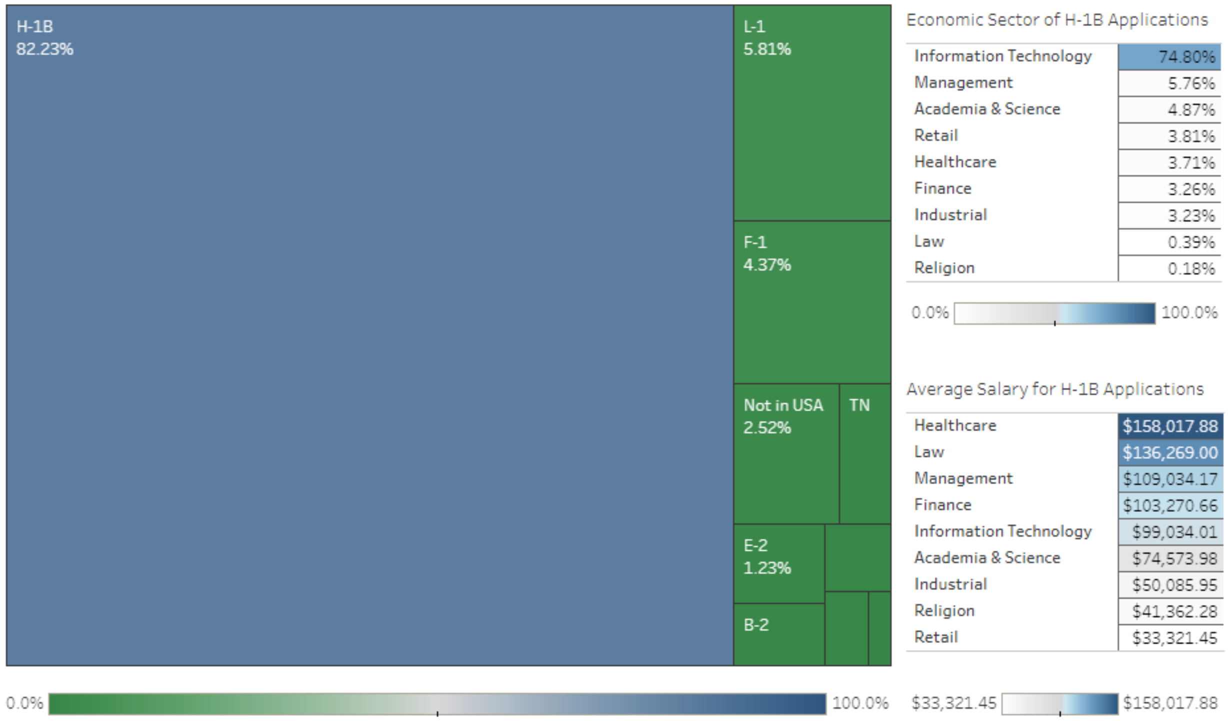
Visualizations:

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.



Economic Sector of H-1B Applications

Information Technology	74.80%
Management	5.76%
Academia & Science	4.87%
Retail	3.81%
Healthcare	3.71%
Finance	3.26%
Industrial	3.23%
Law	0.39%
Religion	0.18%

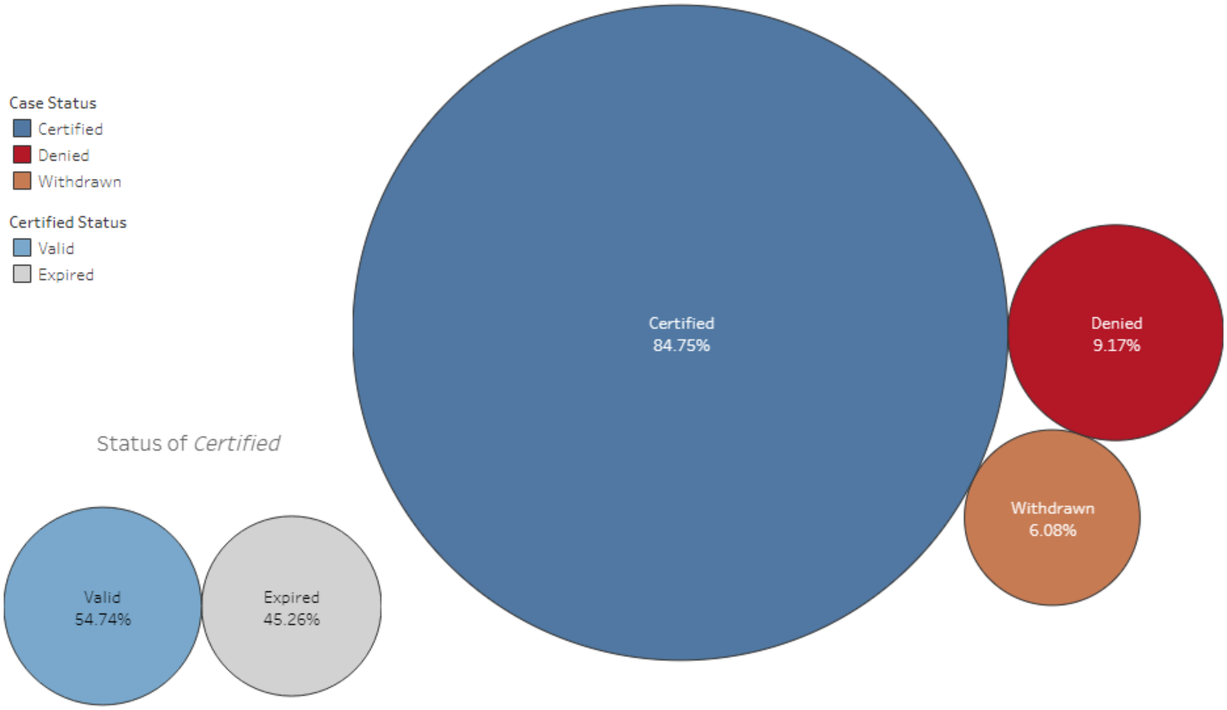
0.0% 100.0%

Average Salary for H-1B Applications

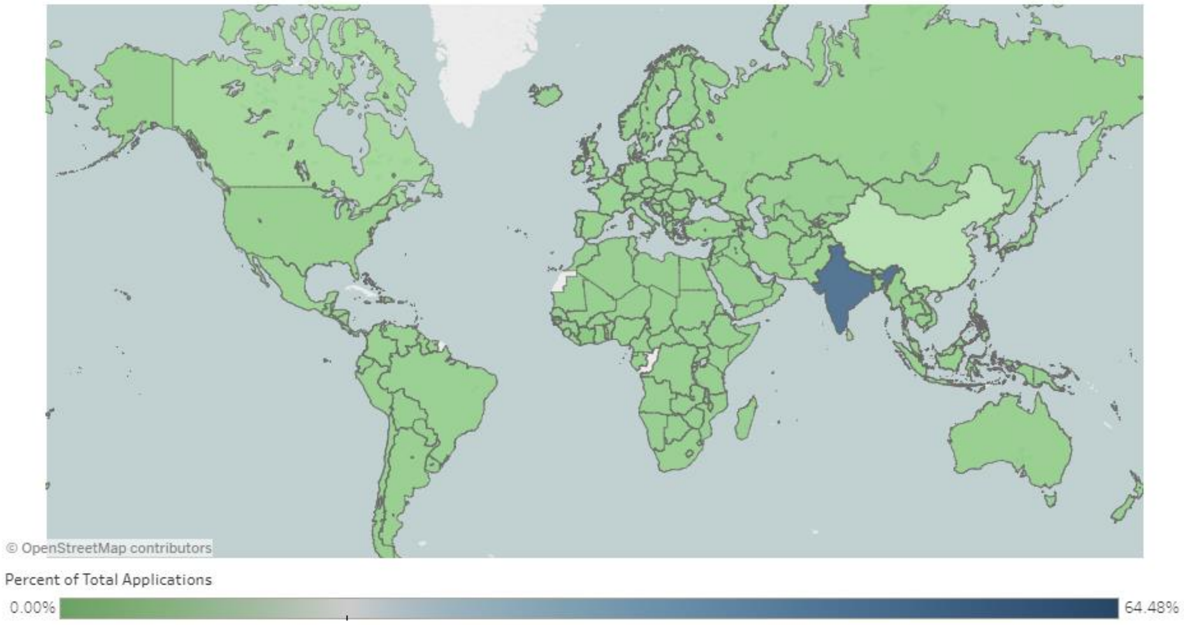
Healthcare	$158,017.88
Law	$136,269.00
Management	$109,034.17
Finance	$103,270.66
Information Technology	$99,034.01
Academia & Science	$74,573.98
Industrial	$50,085.95
Religion	$41,362.28
Retail	$33,321.45

$33,321.45 $158,017.88

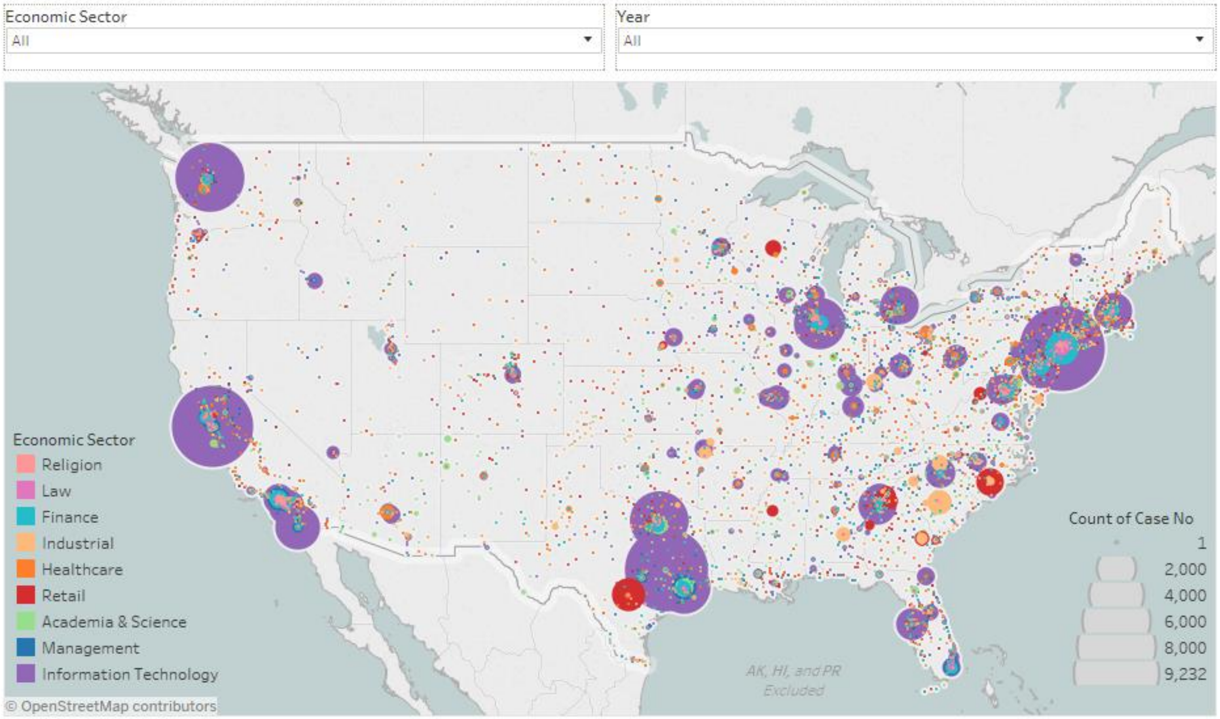
We also felt it was important for the viewer to see information about the frequency with which the H-1B applications were Certified, Denied, or Withdrawn. Knowing that pie charts are despicable, and that we had already used an area chart, we debated the choice between a bar chart and a bubble chart. 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.

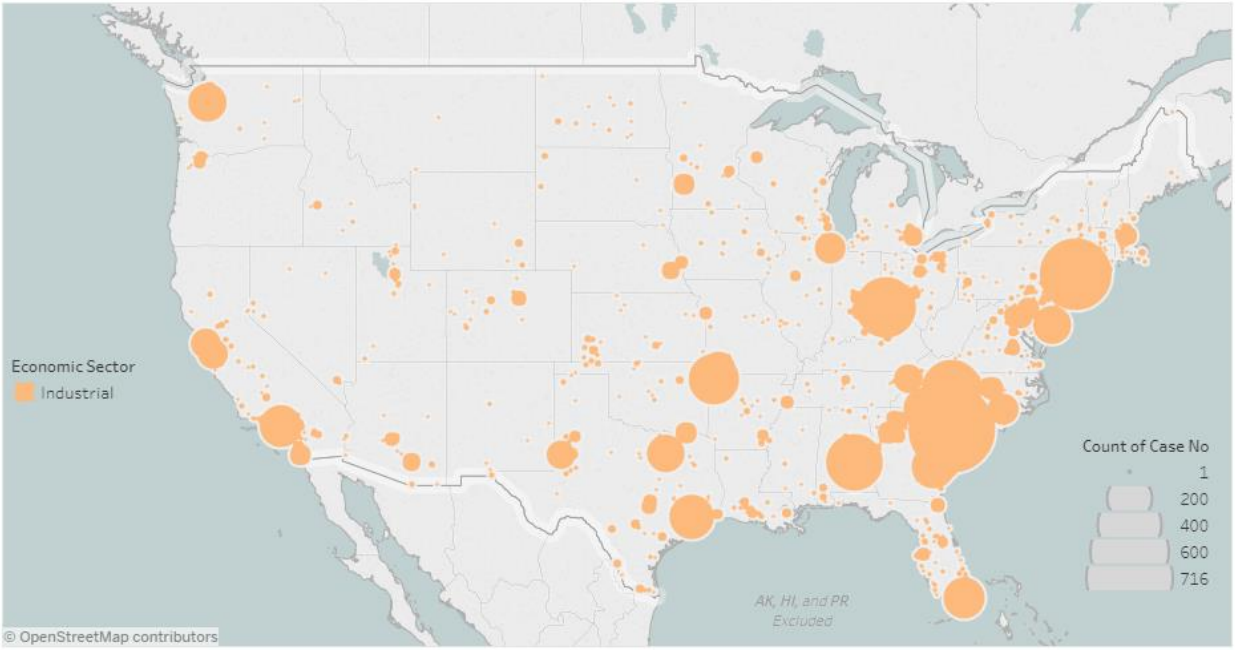


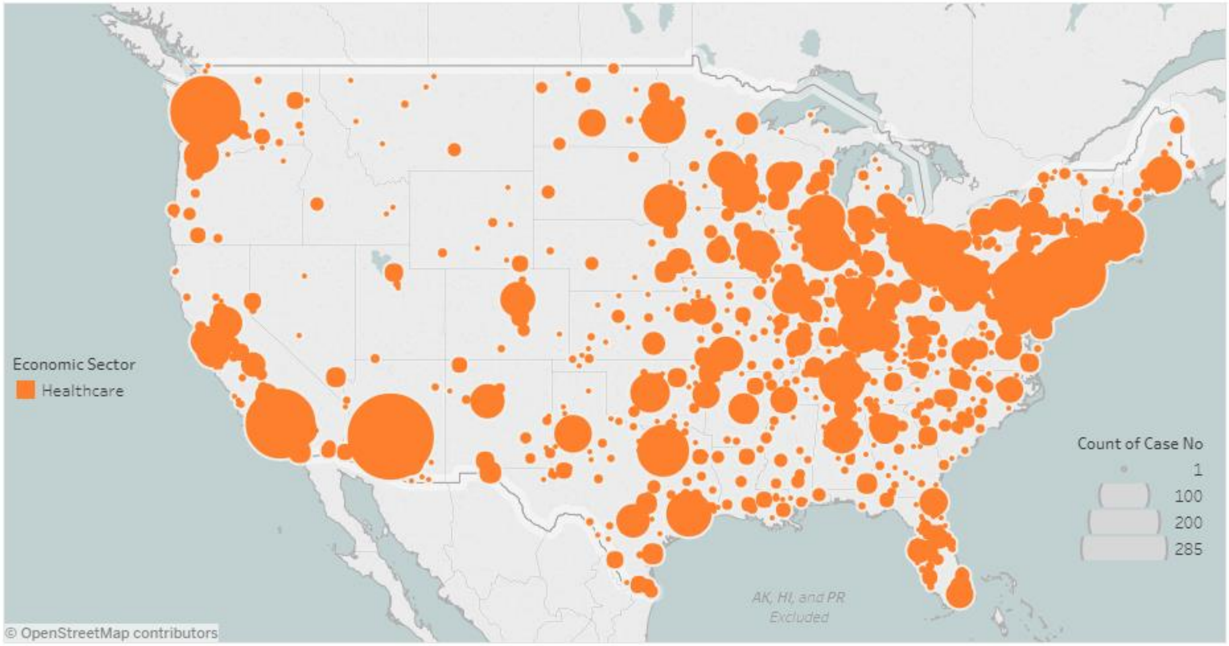
The US map was sorted in an ascending fashion so to keep the largest bubbles on the bottom, and as much of the variation can be seen. The Economic Sectors are parameterized, allowing different industries to be analyzed in more detail. We opted to prevent the user from manipulating the maps, which resulted 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.



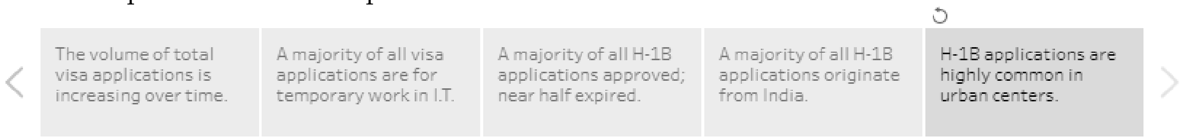
Economic Sector: All

Year: All





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.



The volume of total visa applications is increasing over time.

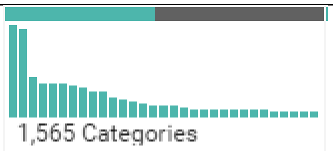
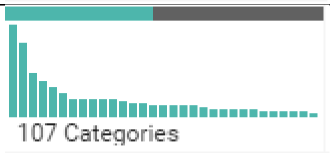
A majority of all visa applications are for temporary work in I.T.

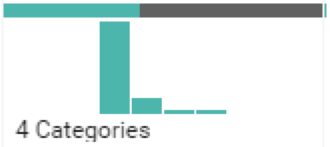
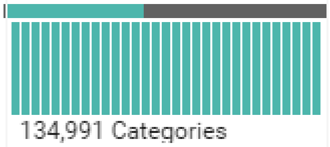
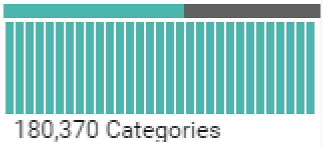
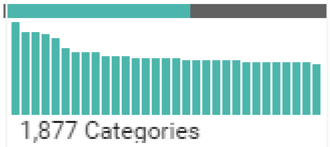
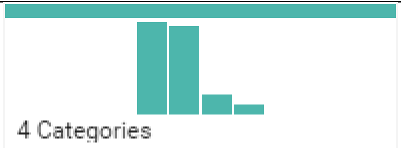
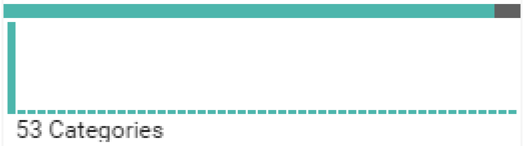
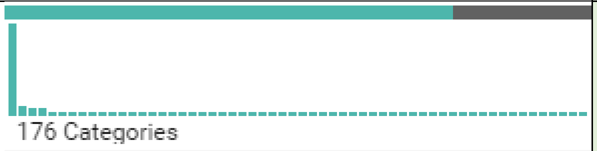
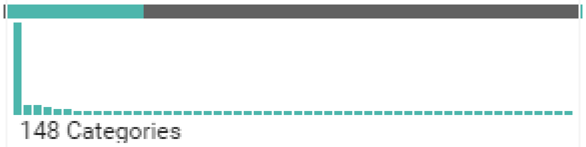
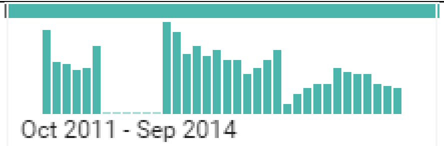
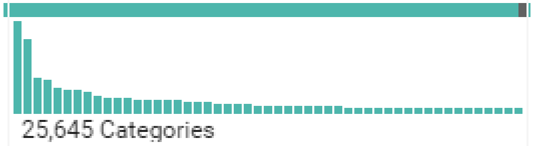
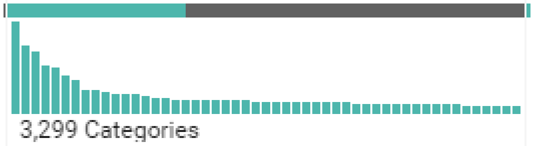
A majority of all H-1B applications approved; near half expired.

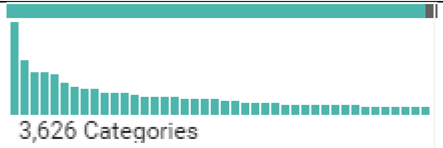
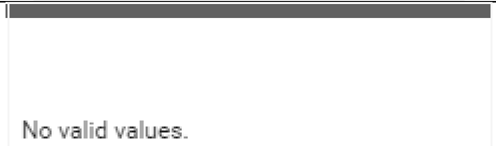
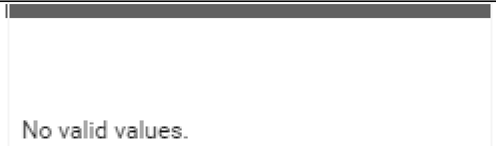
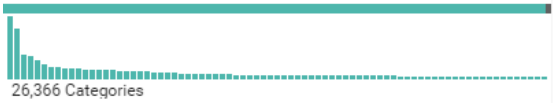
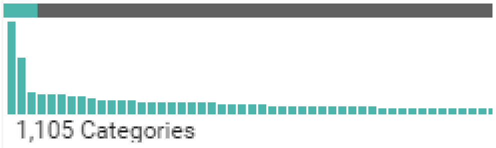
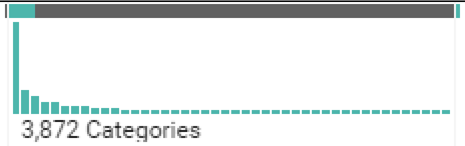
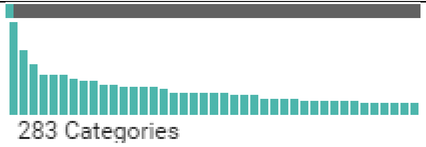
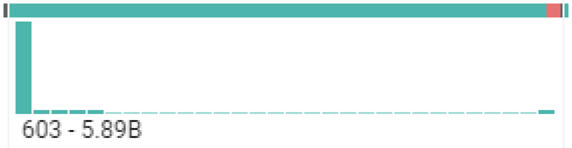
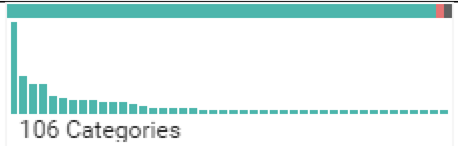
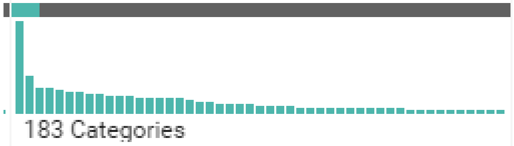
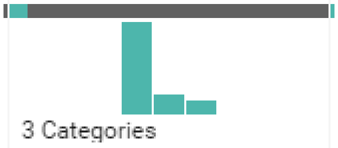
A majority of all H-1B applications originate from India.

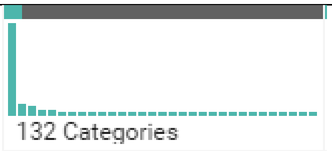
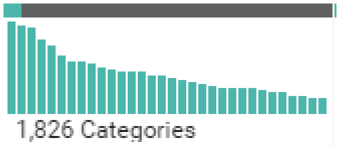
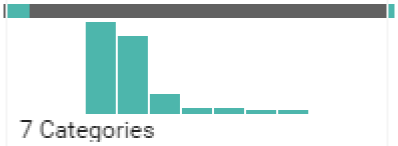
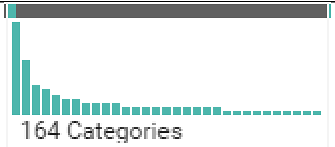
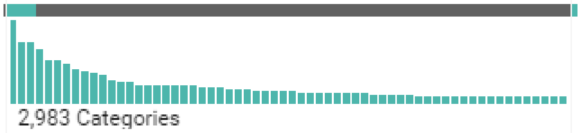
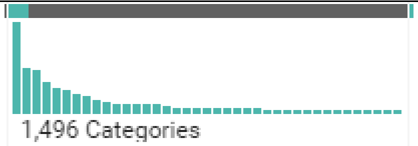
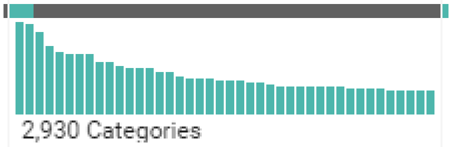
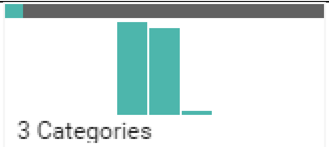
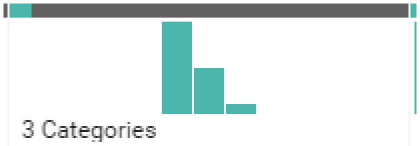
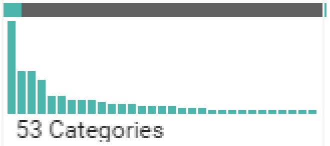
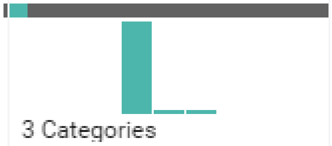
H-1B applications are highly common in urban centers.

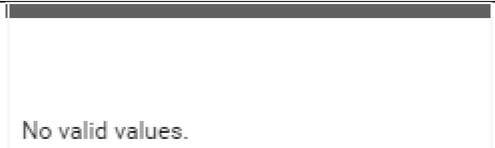
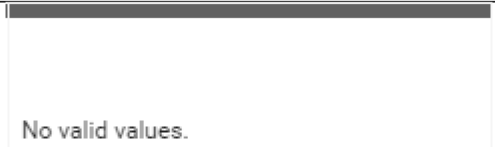
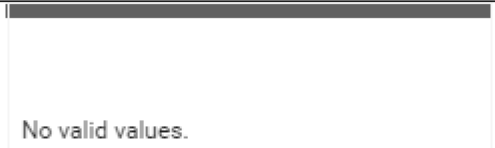
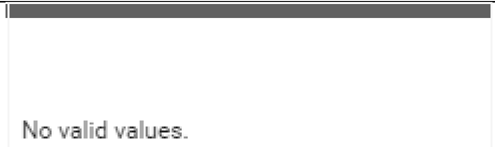
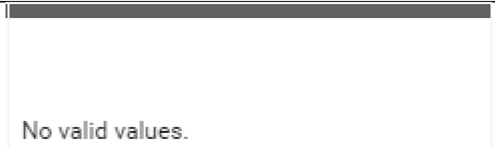
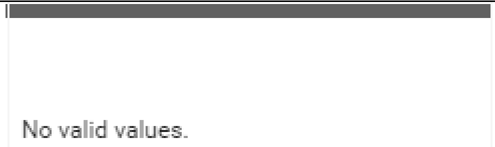
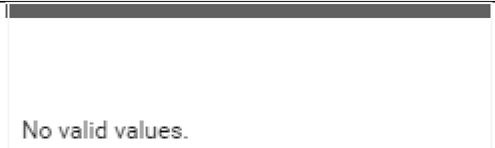
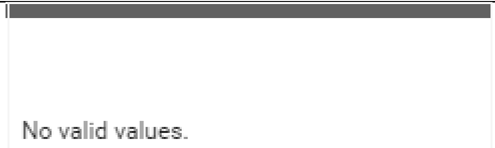
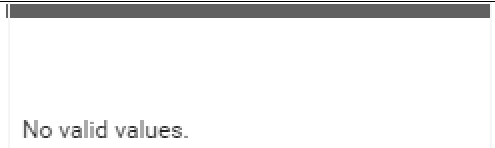
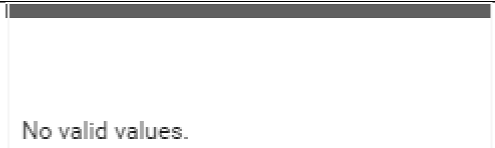
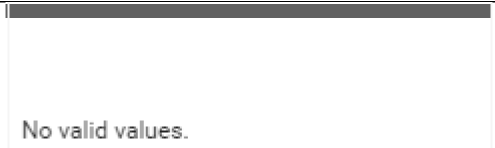
Data cleaning:

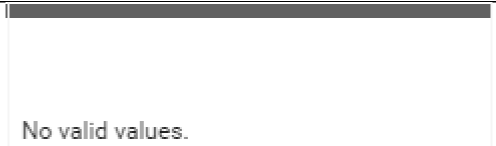
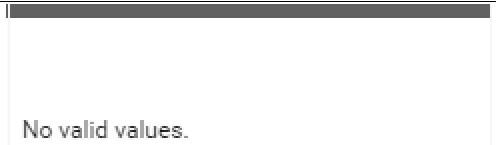
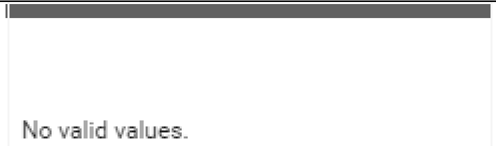
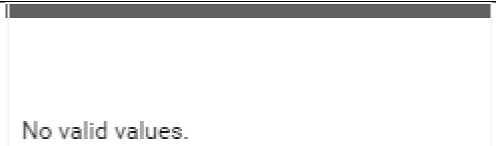
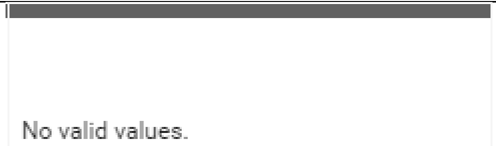
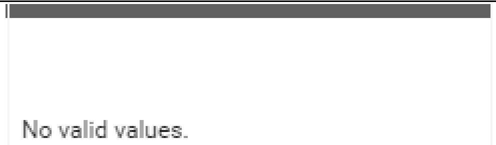
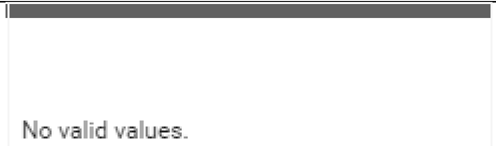
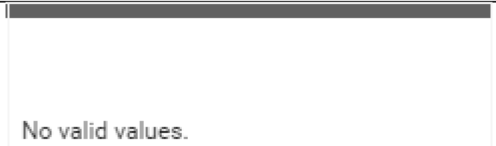
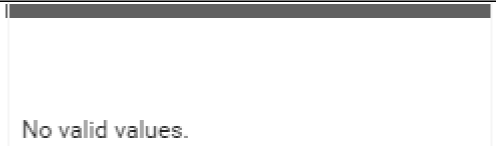
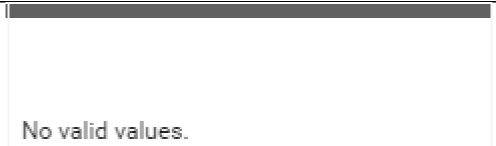
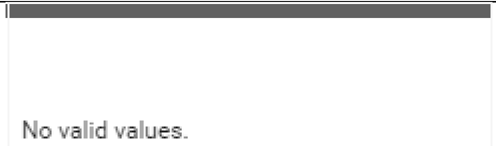
Column Name	Distribution	Changes
add_these_pw_job_title_9089	4,277 Categories	Drop – Insufficient Data
agent_city		Drop – Insufficient Data
agent_firm_name	7,820 Categories	Drop – Insufficient Data
agent_state		Drop – Insufficient Data

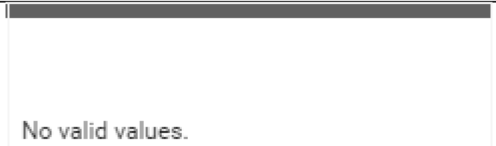
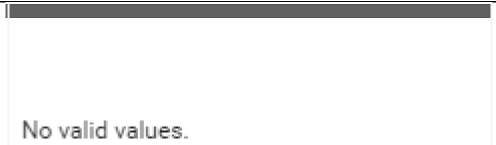
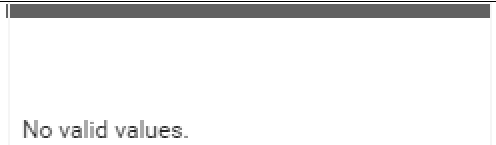
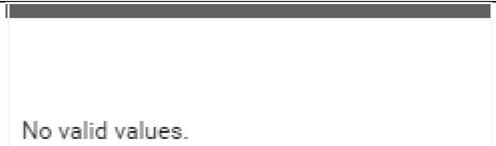
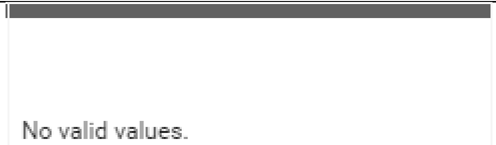
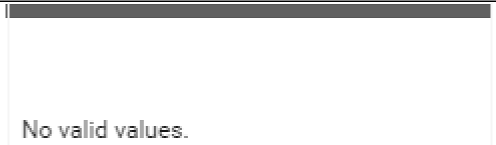
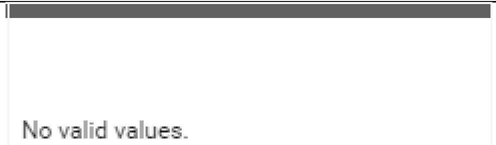
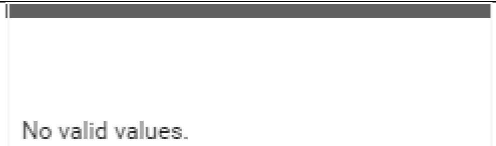
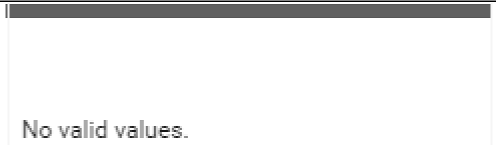
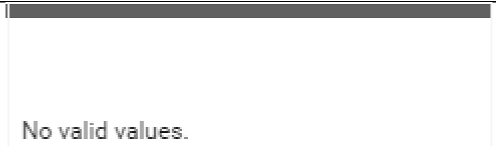
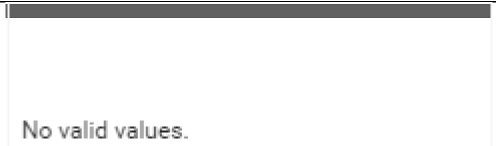
application_type		Drop – Insufficient Data
case_no		Combine with case_number
case_number		Combine with case_no
case_received_date		Drop – Insufficient Data
case_status		Keep - Wrangle
class_of_admission		Keep - Wrangle
country_of_citizenship		Combine with country_of_citzenship
country_of_citzenship		Combine with country_of_citizenship
decision_date		Keep
employer_address_1		Keep - Wrangle
employer_address_2		Drop – Insufficient Data

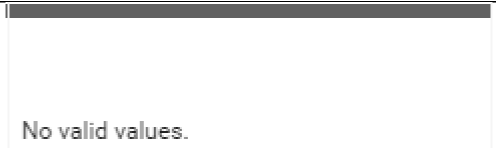
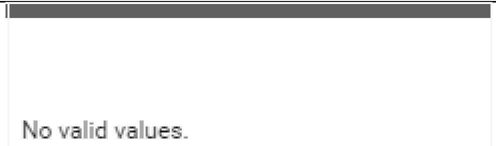
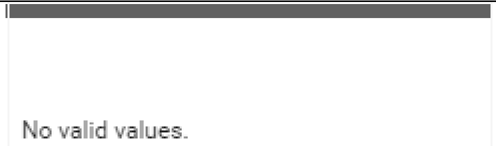
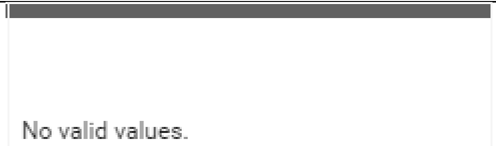
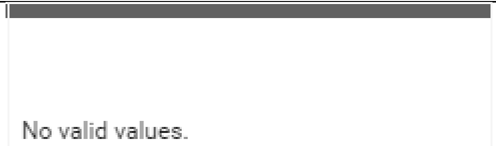
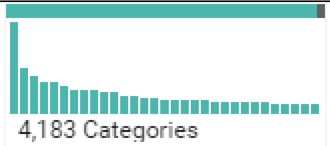
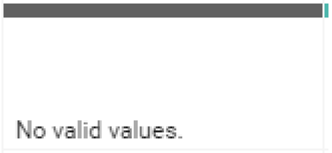
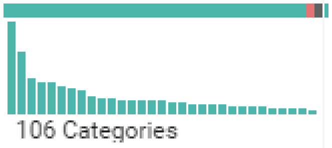
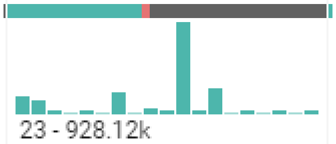
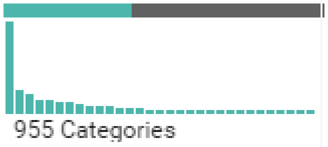
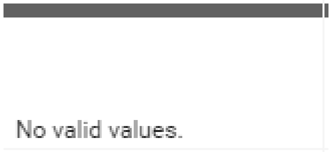
employer_city		Keep - Wrangle
employer_country		Drop – Insufficient Data
employer_decl_info_title		Drop – Insufficient Data
employer_name		Keep - Wrangle
employer_num_employees		Drop – Insufficient Data
employer_phone		Drop – Insufficient Data
employer_phone_ext		Drop – Insufficient Data
employer_postal_code		Keep
employer_state		Keep - Wrangle
employer_yr_estab		Drop – Insufficient Data
foreign_worker_info_alt_edu_experience		Drop – Insufficient Data

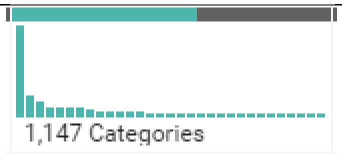
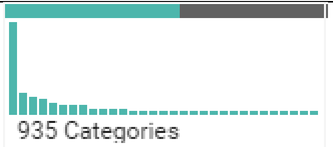
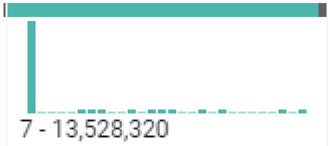
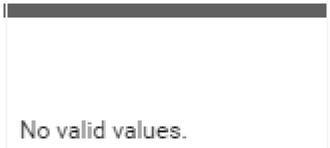
foreign_worker_info_birth_country		Drop – Insufficient Data
foreign_worker_info_city		Drop – Insufficient Data
foreign_worker_info_education		Drop – Insufficient Data
foreign_worker_info_education_other		Drop – Insufficient Data
foreign_worker_info_inst		Drop – Insufficient Data
foreign_worker_info_major		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

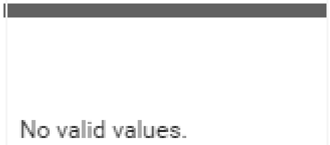
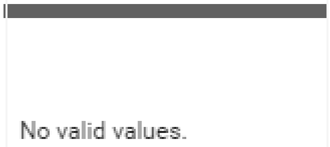
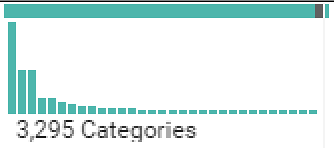
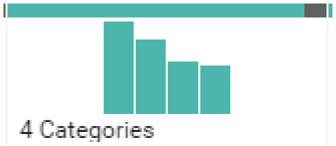
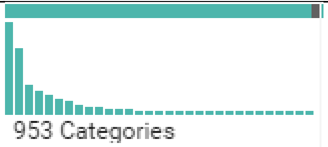
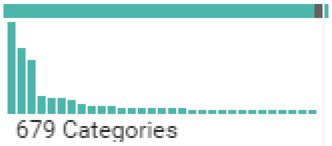
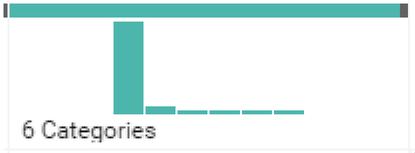
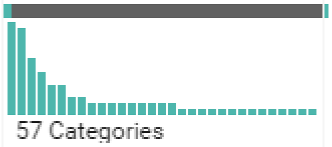
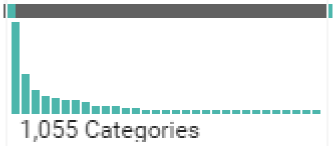
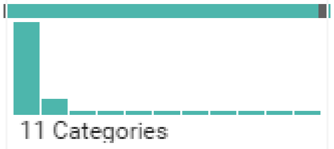
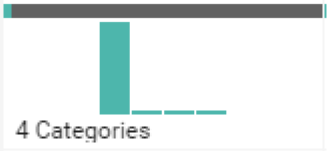
foreign_worker_ownership_interest	 No valid values.	Drop – Insufficient Data
foreign_worker_yr_rel_edu_completed	 No valid values.	Drop – Insufficient Data
fw_info_alt_edu_experience	 No valid values.	Drop – Insufficient Data
fw_info_birth_country	 No valid values.	Drop – Insufficient Data
fw_info_education_other	 No valid values.	Drop – Insufficient Data
fw_info_postal_code	 No valid values.	Drop – Insufficient Data
fw_info_rel_occup_exp	 No valid values.	Drop – Insufficient Data
fw_info_req_experience	 No valid values.	Drop – Insufficient Data
fw_info_training_comp	 No valid values.	Drop – Insufficient Data
fw_info_yr_rel_edu_completed	 No valid values.	Drop – Insufficient Data
fw_ownership_interest	 No valid values.	Drop – Insufficient Data

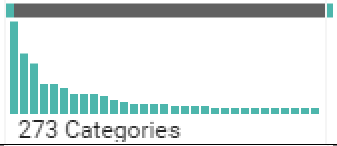
ji_foreign_worker_live_on_premises	 No valid values.	Drop – Insufficient Data
ji_fw_live_on_premises	 No valid values.	Drop – Insufficient Data
ji_live_in_dom_svc_contract	 No valid values.	Drop – Insufficient Data
ji_live_in_domestic_service	 No valid values.	Drop – Insufficient Data
ji_offered_to_sec_j_foreign_worker	 No valid values.	Drop – Insufficient Data
ji_offered_to_sec_j_fw	 No valid values.	Drop – Insufficient Data
job_info_alt_cmb_ed_oth_yrs	 No valid values.	Drop – Insufficient Data
job_info_alt_combo_ed	 No valid values.	Drop – Insufficient Data
job_info_alt_combo_ed_exp	 No valid values.	Drop – Insufficient Data
job_info_alt_combo_ed_other	 No valid values.	Drop – Insufficient Data
job_info_alt_field	 No valid values.	Drop – Insufficient Data

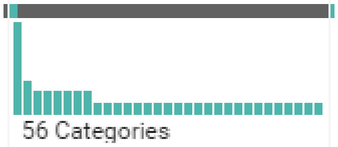
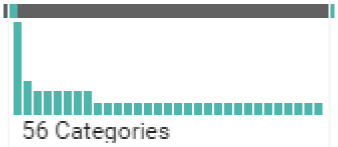
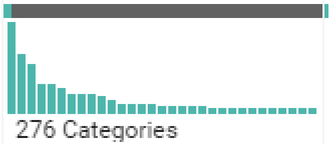
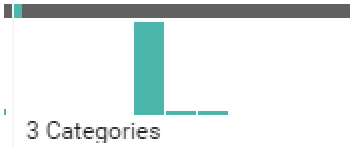
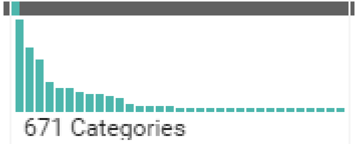
job_info_alt_field_name	 No valid values.	Drop – Insufficient Data
job_info_alt_occ	 No valid values.	Drop – Insufficient Data
job_info_alt_occ_job_title	 No valid values.	Drop – Insufficient Data
job_info_alt_occ_num_months	 No valid values.	Drop – Insufficient Data
job_info_education	 No valid values.	Drop – Insufficient Data
job_info_education_other	 No valid values.	Drop – Insufficient Data
job_info_experience	 No valid values.	Drop – Insufficient Data
job_info_experience_num_months	 No valid values.	Drop – Insufficient Data
job_info_foreign_ed	 No valid values.	Drop – Insufficient Data
job_info_foreign_lang_req	 No valid values.	Drop – Insufficient Data
job_info_job_req_normal	 No valid values.	Drop – Insufficient Data

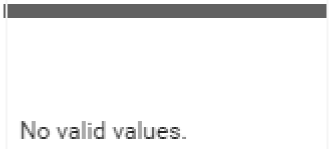
job_info_job_title		Drop – Insufficient Data
job_info_major		Drop – Insufficient Data
job_info_training		Drop – Insufficient Data
job_info_training_field		Drop – Insufficient Data
job_info_training_num_months		Drop – Insufficient Data
job_info_work_city		Drop - Redundant
job_info_work_postal_code		Drop – 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

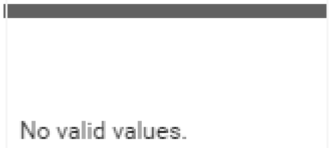
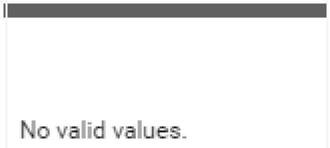
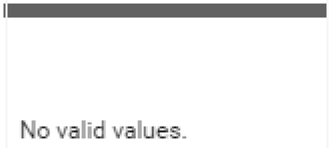
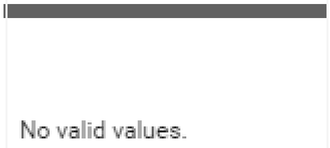
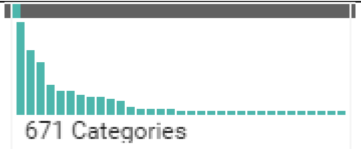
naics_title		Drop – Insufficient Data
naics_us_code		Drop – Insufficient Data
naics_us_code_2007		Drop – Insufficient Data
naics_us_title		Drop – Insufficient Data
naics_us_title_2007		Drop – Insufficient Data
orig_case_no		Drop – Insufficient Data
orig_file_date		Drop – Insufficient Data
preparer_info_emp_completed		Drop – Insufficient Data
preparer_info_title		Drop – Insufficient Data
pw_amount_9809		Drop - Redundant
pw_determ_date		Drop – Insufficient Data

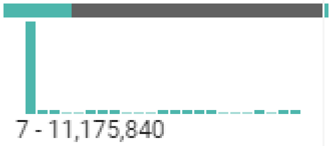
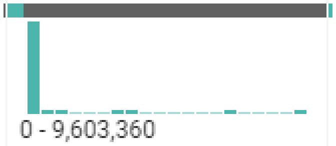
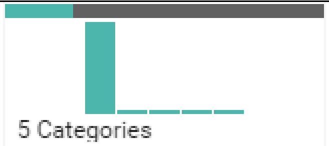
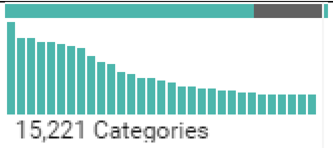
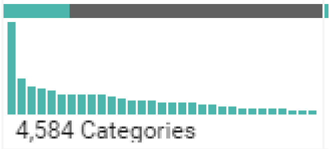
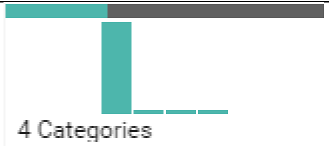
pw_expire_date		Drop – Insufficient Data
pw_job_title_908		Drop – Insufficient Data
pw_job_title_9089		Keep – Wrangle : Economic Sector
pw_level_9089		Keep - Wrangle
pw_soc_code		Drop – Unrelated Data
pw_soc_title		Drop – Unrelated Data
pw_source_name_9089		Drop – Unrelated Data
pw_source_name_other_9089		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	3 Categories	Drop – Insufficient Data
recr_info_coll_univ_teacher	3 Categories	Drop – Insufficient Data
recr_info_employer_rec_payment		Drop – Insufficient Data
recr_info_first_ad_start		Drop – Insufficient Data
recr_info_job_fair_from	21 Categories	Drop – Insufficient Data
recr_info_job_fair_to	21 Categories	Drop – Insufficient Data
recr_info_on_campus_recr_from	21 Categories	Drop – Insufficient Data
recr_info_on_campus_recr_to	21 Categories	Drop – Insufficient Data
recr_info_pro_org_advert_from	88 Categories	Drop – Insufficient Data
recr_info_pro_org_advert_to	95 Categories	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	 No valid values.	Drop – Insufficient Data
ri_job_search_website_to	 No valid values.	Drop – Insufficient Data
ri_layoff_in_past_six_months	 No valid values.	Drop – Insufficient Data
ri_local_ethnic_paper_from	 No valid values.	Drop – Insufficient Data
ri_local_ethnic_paper_to	 No valid values.	Drop – Insufficient Data
ri_posted_notice_at_worksite	 No valid values.	Drop – Insufficient Data
ri_pvt_employment_firm_from	 No valid values.	Drop – Insufficient Data
ri_pvt_employment_firm_to	 No valid values.	Drop – Insufficient Data
ri_us_workers_considered	 No valid values.	Drop – Insufficient Data
schd_a_sheepherder	 No valid values.	Drop – Insufficient Data
us_economic_sector	 671 Categories	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