Data Science and UX H-1B Petition (2011-2016)

Group 18: Xianyue Li, Yingjin Qi, Muhan Yuan, Yize Zang

Executive Summary

1) Briefly describe your proposed idea (paragraph or two). What will you be communicating?

   Our core idea is to present H-1B petition trends from different aspects within time span of 2011 to 2016, specifically targeting on Data Science and UX track related jobs, which fits in interests of UMSI students. We would express number of H-1B petition submission, H-1B petition status, companies submitting most H1b petitions and job titles appearing most in H-1B petitions in each U.S. state over 6 years regarding to Data Science and UX. We would also provide detailed interactive visualization on companies submitting most H-1B petitions per year and number of certified cases.

2) Who is your audience for this?

   Our audience is:
   - International students majoring in Data Science or UX and looking for full-time positions in U.S.
   - People interested in H-1B petition trends from 2011 to 2016

Questionnaire

1) What is your data?

Our data is H-1B Visa Petitions dataset on Kaggle[1]. The original dataset is .csv file and includes around 3 million records. Important variables are:

- case_status: Status associated with the most recent decision from USCIS
- employer_name: Name of employer submitting H-1B petition
- SOC_name: Occupational name encoded according to Standard Occupational Classification (SOC) System
- prevailing_wage: Average wage paid in the area of intended employment based on the employer's minimum requirements for the position.
- job_title, year, worksite, longitude and latitude

From the original dataset, we filter out jobs related with Data Science and UX. Data Science related jobs are defined as job titles include string "data", while UX related jobs are defined as job titles include one of the following strings: "UX", "UI", "experience design", "interaction design", "information architect", "visual design", "digital design", "product design", "content strategy" and "usability". There are 85850 records in total, with 19703 UX related and 66147 data science related.

We also create two variables:

- state, which is extracted from worksite variable and
- salary, which is transformed from prevailing wage. We set ⅓ and ⅔ quantiles as boundaries of low and medium, medium and high salary of Data Science and UX related jobs.

After aggregating by year, track and state, we count number and calculate percent of H-1B petitions regarding to different case status, employers, SOC names and salary range. We have three different .json files for each visualization based on needs.

2) What are the tasks or learning goals you want to support? What should someone be able to understand after seeing / using your visualization?

Our visualizations are aimed to answer the following questions:

- How does the **total number of H-1B application** with different status change over time?
- What's the **geographical distribution** pattern of the number of H-1B application with different status?
- How does the number of H-1B applications with **different job titles** (pass/deny/withdraw etc.) change over time?

[1] https://www.kaggle.com/nsharan/h-1b-visa/data

- What is the application status in certain conditions(job title, salary, passing rate, etc.) in **different companies** in the past 6 years?

By interacting with our visualizations, people should have an overview of the H-1B application changing trend in the past 6 years as well as application attributes at states, company and specific job title level.

3) How are you encoding the data visually?

To uncover more insights, we created three different visualizations.

In the first visualization, we used a dot distribution map as an introduction to our visualization project. Two colors of dots represent two **types** of H-1B petition **job titles** (data science and UX design). The **latitude** and **longitude** of the petition is encoded as the location of the dot and the **year** of each petition is encoded as the **time** in the animation.

In the second visualization, users can toggle between data for the two **types** of H-1B petition **job-titles** (data science and UX design). For both scenarios, we created 5 series of U.S. map pictograms to visualize demographics information for each state. In each serie, rectangles are used to represent **states** and distributed according to the **geographical locations** of the states. The **application number** and **category ratio** are encoded as the rectangle size and sub-rectangle size respectively. Also encoded are data for different **categories** by color hues, **application intensity** by color brightness and **year** of application by slidebar. Moreover, users are able to view **detailed statistics** by hovering on the rectangles to bring out tooltips.

In the third visualization, we adopt **bar chart style** to demonstrate the difference between UX and Data Science tracks in the top 50 companies of each year. The **colored bar** represent total **passed submission** of each company while **grey bar** represents **total submission** of each company. The **contrasting colors** convey a strong comparison and the **ratio** of the bright color and grey color of each bar shows the **passing percentage** of each company.

4) Why is your solution effective? (you may argue this relative to some other solution but it needs to be using perception/cognition/semantic justification)

In our first visualization, we use square to represent petitions from each state is relatively easy for user to interpret the data and to find the specific state they interested in, since we embedded each square based on the real location of the state. And we splitted each square into several rectangular based on salary range or status, so that the user would be able to figure out the absolute quantity and relative proportion at the same time. In our third visualization, we encoded the number of petition from each company with the length of the bar and using contrasting color make strong comparison.

5) How are you using text to support your visualization? Do you have any narrative structure in mind?

The way we present our visualization is telling a story about the H-1B application trend in the past six years from a big picture to specific data entries. We start our visualization with an introduction, explaining the background, goal and concepts related to the story.

At the end, we list out key conclusions identified from all three visualizations, which facilitates people to take away the helpful insights.

There are implied narrative structures hidden in our three visualizations in a row. The 0 visualization (click on when clicking on left-top header on nav bar) provides users with intuitive impression on that how number of H-1B petitions changes over 7 years and in which way the allocation between Data Science and UX trach related jobs is different. The first visualization offers users geographical distribution based interactions with different categories, as an extension of the first U.S. map animation. From this visualization, international students could develop their location preference based on H-1B visa petition opportunities. Followed is the second visualization, a even further exploration on hot employers, which could be an organization side guidance for international students' full-time job search.

6) How are you using interactivity (if at all)? Why does it support your task? (use the language from the 7 categories we described in class)

One of our main interactions is encoding, using color in hover-effect to highlight the element selected by the user. As we have rich data, and they are geographically dense, the color encoding makes it easy for users to distinguish a specific data entry.

Another interaction is the filter, allowing users to see the attributes they are mostly interested in. We include multiple filter types, including tab, slidebar, drop down list and flip card, which add fun to users experience.

7) What are the limitations of your solution?

Since the status of most petitions are “Certified”, user might not notice the difference between certified petition and total submission, even though we have already used different colors.

In the second visualization, we have to ensure enough size for each rectangle in order to clearly show ratios encoded by sub-rectangles. However, the states in northeast U.S. are intensively distributed, rendering the corresponding rectangles being crowded together. To solve this problem, we may have need to add manual noises to the geographical data instead of using the real longitudes and latitudes of these states.

Another limitation in the second visualization is about data source. In the application intensity serie, we use the proportion of application count over the population for each state to represent application intensity. We derive the application counts for each year through 2011 to 2016 from our Kaggle H-1B data source, but we use an “static” external source for population which only contains data for 2015.

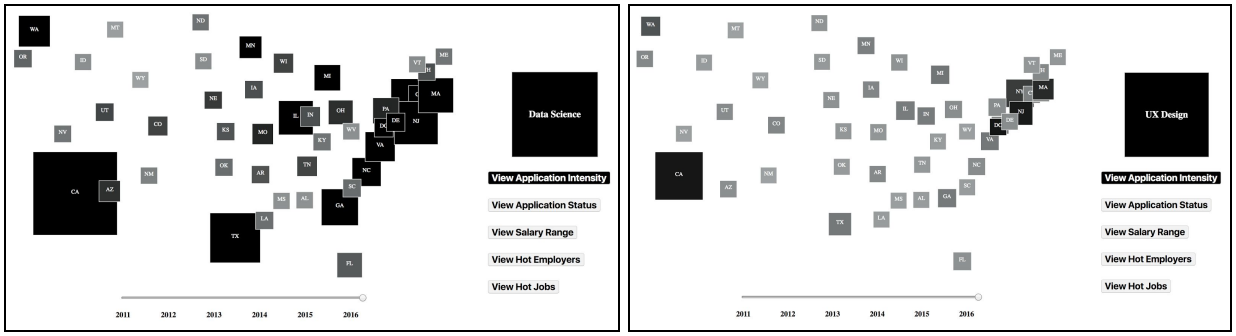
Screenshots

Visualization 0

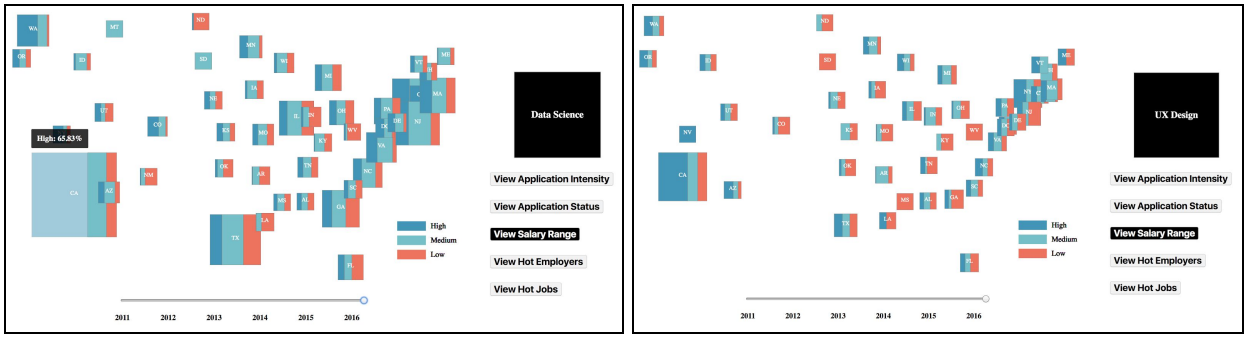
Description

- Two colors of dots represent two types of H-1B petition job titles (data science and UX design) and the location of each dot matches the real latitude and longitude of the petition. Clicking on the top-left header on navigation bar,, the map will automatically changes through 2011 to 2016, and show data science and UX design petition alternately from 2011 to 2016.
- We do not expect the user to learn much details from the map. Instead, we would like use the map to arouse user’s interest and give them a general idea of our project.

Visualization 1



(a)



(b)

Our second visualization contains multiple views for different job types (data science and UX design), years and series. This visualization is highly interactive and users can draw a lot of interesting conclusions as they explore multiple views by flipping the card (data science/UX design), clicking on the serie buttons, draggin the time slidebar or hovering mouse on the graphs. We will take two series out of the five series as examples.

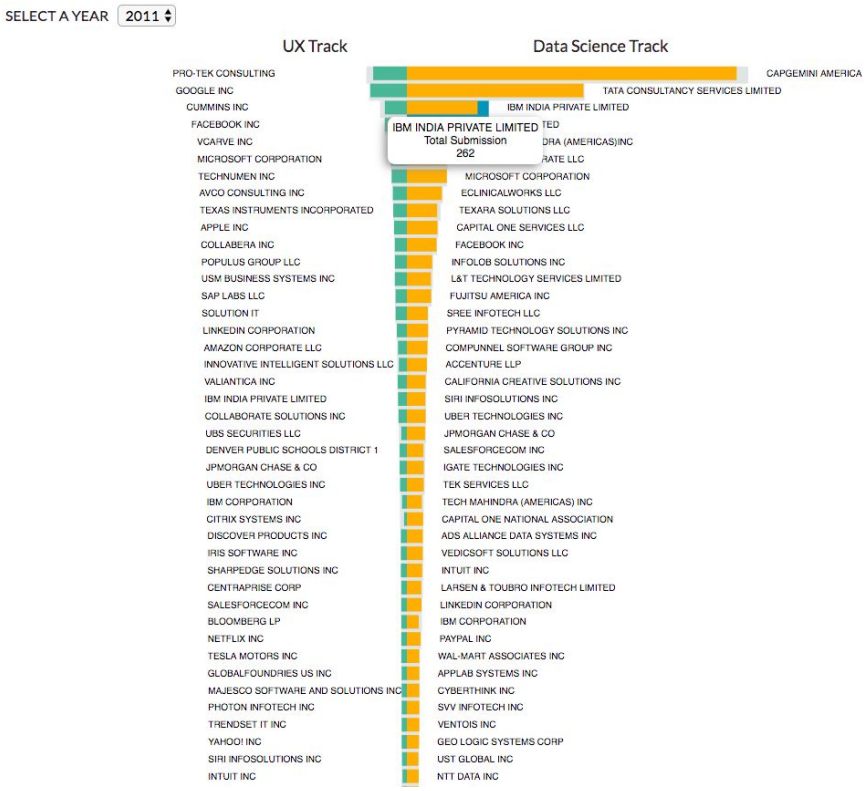
Description (a):

- The two views show application intensity of year 2016 for data science and UX design related positions respectively. Application intensity for each state is calculated by dividing the number of application by the state population.
- By comparing the two views, we can instantly find that application for data science related positions is more intense than UX design related positions in 2016. In fact, this is also the case from 2011 to 2015, and the data science field has an especially dramatic increase after 2013.
- The application intensity for both fields are relatively high in California and some northeastern states like New York. For data science, some central U.S. states like Texas and Illinois also show great increase trend.

Description (b):

- The two views demonstrate salary level distribution of year 2016 for data science and UX design positions respectively. We consider the salary for data science and UX design positions together and separate them into high, medium and low levels by quantiles.
- This serie shows vividly that California, Washington and New York offer more positions in H-1B application with high ranged salary through the years 2011 to 2016. And for both fields, higher ranged salary tends to take more and more ratios in the states overall from 2011 to 2016.

Visualization 2



Description

- Visualizing the top 50 companies' H-1B submissions in descending order.
- Using half-half layout to stress comparison between UX track and Data Science track.
- Users can require specific data through interacting with each bar and explore the hiring needs of a certain company by comparing its ranking on the two sides.