

Data Dames: When does a job post go stale?

We were interested in using a senior college student's insight to shape how employers could improve upon job listings. These variables were chosen based on an applicant's consideration of industry type, job location, job qualifications, position description, and age of job post. We found that this last variable of how long the job had been posted was a crucial variable for employers to consider, especially if they are paying Indeed to be sponsored.

Indeed differentiates "organic listings" from "sponsored jobs" by making the sponsored jobs remain high in the search results and appear prominently at the top and bottom of the page even as time passes while organic listings fall in the search results due to new jobs that are added at a rate of 8.2 per second (blog.indeed.com). While we did not have the data to determine which companies were sponsored, we were able to use this knowledge for later interpretations.

We tried to predict the number of applications with simple linear regression, multiple regression, and mixed models. We wanted to find which were the factors that lead a job post to having more application submissions in Indeed. After analyzing our regression models and seeing very low r-squared values, and a high sensitivity in the p-values, due to the size of the data, we went back to analyzing the data from a visualization point of view. We plotted multiple variables against the number of application and we noticed that a common denominator was time. Around the 4 week mark, no matter what the variable was, the number of applies for job posts start go stale. This is what lead up to looking at the number of applications per week, over the first six weeks a job posting has been on Indeed.

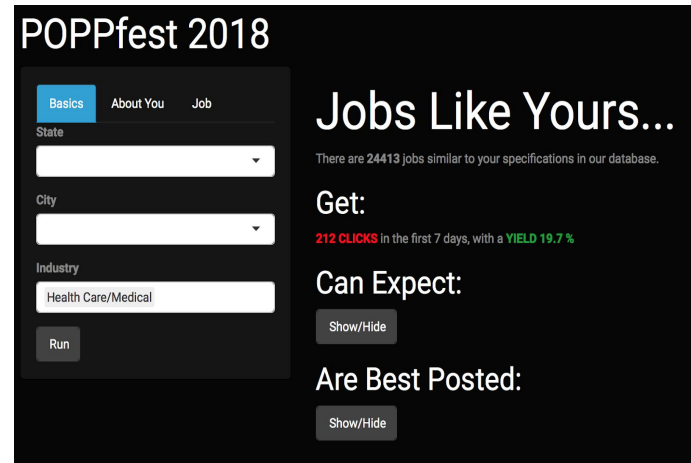
Taking a random sample of 10,000 jobs, we created six different boxplots, one for each week after a job posting has been created, to compare how the number of applications changed over time. These plots showed that a bulk of the number of applications happen in the first and third week age of a job posting. It also showed that after week three the number of applications per week starts to show no change and the biggest difference was the number of outliers for each box plot. Basically, the same companies consistently receive extremely high number of applications, within the first 6 weeks. However, as the number of weeks increase, the number of outliers decrease. This might explain why the total number of applications goes stale around the 4 week mark. One reason for these outliers might be due to certain companies having their job postings sponsored. Sponsored Jobs are based on a pay for performance model. Employers set a monthly budget and only pay when a candidate clicks to view your job. You decide how much to spend and we deliver relevant, high-quality candidate traffic. Some other confounding variables might include: new users on Indeed, company fame, and type of industry

ACOF: POPPfest 2018

Bodhi, Brendan, Jason, Kelly, Natalia

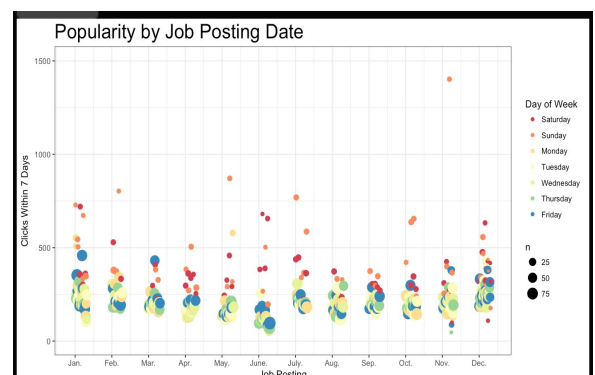
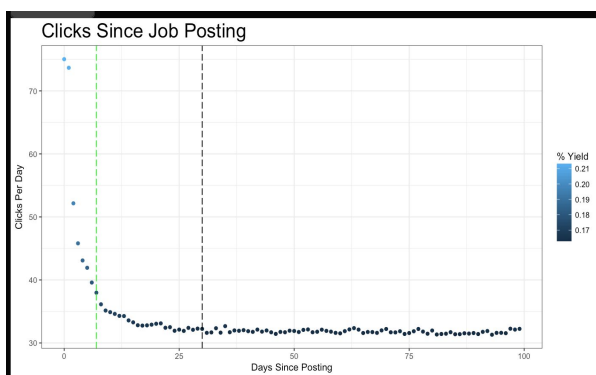
POPP (Personalized Occupation Posting Program):

POPP is our suggestion for a product or service that we believe indeed.com should offer to its clients. Our predictive models were thwarted by inconsistencies between industries and other factors, and we realized that employers care primarily about jobs similar to their own. Therefore, we created this tool to give them important and relevant information before posting a job. Variables included are: # of similar jobs, expected #of clicks and yield (applications/clicks) within 7 days, timeline of job activity after posting and best days in the calendar year to post. The benchmark of 7 days is especially important to employers as jobs tend to get more clicks, with a higher yield, in that timeframe. It is also important to note that all projections are based on historical data from 2016-17. **POPP** should serve as a marketing strategy for indeed.com to allow them to recruit more participants and increase their market share.



Insights: Beyond our service suggestion, we also gathered coordinates of approximately 38,000 cities in the United States from an external source and used them to map job applications. We found that out of the top ten largest cities in the country, San Diego and San Jose seemed to have fewer job postings than other top ten cities. San Francisco also had conspicuously few technology-related jobs, odd for such a large center of tech. We suggest indeed.com focus on expanding into these markets and will note that **POPP** makes it easy for indeed.com or prospective firms to make similar decisions.

Next Steps: The power of the **POPP** relies on how many jobs have been previously posted with similar criteria. As a result, in non-metropolitan cities, and in industries where posting on indeed.com is rare, the insights obtained from **POPP** are less useful. Thus, gathering more data in these areas is crucial to **POPP's** success as a useful application for indeed.com to offer. The end goal of **POPP** is to be a premium service that suggests the optimal posting times and provides the most useful business insights so that companies can maximize the number of clicks and applications on their job postings.



The purpose of this investigation was to give an insight to high school students about future careers. Our three graphs show information provided by indeed.com and outside sources such as the Bureau of Labor Statistics that demonstrates where in the United States are jobs, whether having a high school and college education played a role in qualifying for certain jobs, and which industries have high valued companies. This may influence their future choices concerning pursuing higher education, where they go to college and what they major in.

When introducing the idea about finding a job to high schoolers, our group tried to define what makes a good company. Some companies had a score of 0, so we developed a **new scoring system** (**new_score**). For nonzero `avgoverallcompanyrating`, we kept such score in `new_score`; otherwise we used our prediction model to calculate the value of the company (`new_score`) and filled in the blanks. Such prediction model is fed by nonzero `new_score`; it contains four major variables: salary, company size, number of applicants, and popularity (number of clicks of that company's web page).

However, when testing the model, we found out that the correlation between `clicks` and `applies` was one—which, in real world data, is unlikely. For fitting our multiple regression model we have to take one of them away. Thus, in our final model, we have the following three variables: **employeecount**, **applies**, and **Salary**.

We then further cleaned our data. Since the database doesn't include salary, we searched on "Bureau of Labor and Statistics" and added it into our table. By generalizing the 59 types of `normtitlecategory` into 11 larger categories we got the median wage for each category. With this method, we gave each individual company (each row) a simulated salary based on what type of job they are offering. We then divided the `employeecount` into three major groups, instead of eight. In this way, we reduced the chance that by having too specific categories, `employeecount` has too much influence on the regression model.

When thinking of the aimed audience, we wanted to aim this towards high schoolers who were looking at jobs and wanted to show the number of jobs correlated with the number of institutions in each state. By doing so, we used the variables pertaining to this idea, including `admin1`, `JobCount`, `state`, `long`, and `lat`. In addition to this, we imported other data regarding the number of [institutions](#) and center coordinates of each [state](#).

We first "summarised()" all of the jobs in each state, grouping by `admin1`. This way, we would have one job count per state, which was not the case prior to summarising. After doing so, we joined two tables; one of which included the total job count per state and one which included the coordinates to create the map of the United States (from `map_data("states")`). We also joined another table that combined the table that was previously described, and one that included the center coordinates of each state. By combining these two tables, we were able to create a data graphic that included the map of the United States, with color correlating to the number of jobs and size of points correlating to the number of institutions.

For high school students, pursuing a higher education is a decision to make. Does going to college give them a higher chance of getting a better job? Thus, we used an index that shows the company values and the education requirements that companies have for their job offers. The result in our box plot shows that companies that require high school education have the highest company value, which is slightly higher than that of companies that require a higher education.

In our final graph, we wanted to analyze the relationship between the careers and companies found in indeed.com. We decided to categorize eleven general industries based on the careers provided in the data frame to incorporate our company value index. Using these two variables, we were able to demonstrate the overall value of companies found in each industry.

User Application for Hiring Season Investigation

The Code Girls

Mount Holyoke College

Sumatra Dhimoyee, Amelia Johnson, Raeesa Mehjabeen, Rachel Bostick, Anisa Kabir

We used the data to build an application that can be useful for Indeed's job seekers who are interested in knowing more about the popular peak times for job postings in their industry of interest. We define peak times as monthly fluctuations in any industry for the data provided from December 2016 to January 2018.

We represent industry specific job fluctuations and their relative maps in the app that can be a visual friendly way for users to analyze the increase in log of the summation of the job counts in their industry of interest. The users basically get to click on their industry of interest to see the monthly fluctuations of that industry, and also have the option to see the map that highlights the specific cities that have the highest job posting counts for that industry

We used RStudio to create the app and the graphs. We created new variables two new variable year and month using the lubridate function year() and month() and we use these two variables to create new time series for 14 month. We found the unique categories in our employer industries. We ran a for-loop that ran through all the categories and created subsets of our large data for each industry. It counted the total employerJobCount for each month and plotted it against that month using ggplot().

We finally end the presentation by selecting six very interesting industries from the 29 industries that show sharp fluctuation patterns during specific months. We show that the financial industry for instance has a very sharp increase (almost 139%) from the month of September to October , which for instance may be very representative of the recruitment cycle of graduates going into their fall semester of senior year. This is a pattern that has become very popular amongst the financial and banking industries, and especially among entry level positions. Our research from the data shows that we were able to observe that frequent job titles that came up during the months of September – October were mostly for “assistant” level positions which are representative of entry level positions and may explain the recruitment cycle for undergraduates in the financial industry. However, some of the data did not explain why very different job titles showed up during those months which may mean that there may be confounding variables clouding our judgement from the data.

Thus there may be a lot of confounding variables that can affect our decisions about the industry monthly fluctuations such as a change in indeed use by companies in each industry

or industry growth/ decline, or even how some of the cycles might be representative of the fact that the job market in 2017 has been doing really well due to the low unemployment in the U.S. economy.

DataFest 2018_Smith College Team : Thicc Data (Janelle Lin, Crystal Zang, Maggie Wang, Ha Cao, Dardalie Brooks)

Indeed, the U.S. employment search engine is responsible for matching job seekers with employers. This daunting process may be simplified through an understanding of competitiveness across employer industry and job types, along with applicant skill level. Based on the International Standard Classification of Occupation, we categorized job into eight types major ones: Professional, Managers, Service and sales workers, Skilled workers, Elementary occupation, Technicians, Clerical support workers and Armed force. We measure competitiveness of a job post by number of applies over all the job openings (number of applications per job opening). Applicant skill level was defined as education level as required by the job posting (Higher education, high school diploma or no education requirement), along with license requirement.

Insights

It's important for the job seekers to understand what are the characteristics of recruiters on Indeed. Most companies posting on Indeed are small size, 0-49 people. Among all the methods of job postings, we compared the two majority categories: direct and third party hiring, and we found out except for Staffing Industry, most are from recruiters are from direct employers. Our insight for job seekers is that they are able to have direct contact with Indeed. Also, based on the trend from the recruiting season, which is very high in winter, it's important for them to apply early for the recruiting season.

Running with the theme and definition of competitiveness, we saw that number of applies per job opening is consistent over the eighth job types. Among the applicants, about 50% are reviewed by recruiters. We conclude that the competitiveness among all the job types is high. Along with the competitiveness, we explored the amount of candidates reviewed as a function of job post age. A plot of these data revealed a steady decline in candidates reviewed as job post age increased. Because the competitiveness, we suggest job seekers to apply early, especially to the jobs posted within two weeks.

Further insights on the education level and license. We compared three education levels' competitiveness with having a license. Having a license does not alter the competitiveness for high school education and higher education. But for applicants who do not have education, having a license increase the competitiveness which is defined by us: number of application per job posts. By investigating supervised jobs, among applicants who do not have education, there are significantly more of them have a license. The take away from these analyses would be getting a license is most beneficial for people applying to the jobs with no education requirements especially supervising jobs.

DataFest2018

Jared Nussbaum, Katrina Greene, David Welch, Robin Wu

We began to analyze the data by looking through the lense of each users. These include potential employees looking for jobs, companies posting their jobs trying to find the right employees, and Indeed.com themselves.

Companies

We used this data to find insights to help companies that post jobs and look for applicants on Indeed. Companies using Indeed want to find the perfect fit for their job opening; in order to help with this goal, we focused on the question of how companies can best maximize the amount of hires that result in their use of this website. In order to answer this question, we compared the amount of clicks per application post to the amount of applies per application post. They are linearly related– the more people that click on an application, the more people apply to the job. As such, a company would best use their resources by creating Indeed job postings that get as many people to click on them as possible.

Employees

As a generalization, we can split people up (relative to their careers) into two different categories: place people and position people. Position people are those who care exclusively (or at least mostly) about their job, not where it is. Place people are those who want a suitable job for them, but are more interested in taking whatever is available to them that allows them to stay in (or move to) their favorite place. Indeed does a fantastic job catering to position people, with a very effective feature that allows you to search for jobs based on job title or some job metrics. However, Indeed somewhat struggles catering to position people. The location radius feature sometimes stretches to include jobs outside the radius to fluff the number of matches, and it's hard to visualize where jobs are relative to the place your searching if you don't know every single suburb in a 50 mile radius. So, we created a tool in Tableau to do exactly that. You can visualize the jobs in the area and check the postings based on their proximity to your target location. As everyone knows: Location. Location. Location.

Indeed

Indeed wants to get as many people as possible to use their website. As such, they are invested in the success of both the companies and the employees. So, to maximize the usage, indeed.com wants a breadth of jobs in many different locations, a large depth of jobs in popular locations, and many clicks on every job posted.

Team 23333

Team members: Ziwei He, Hao Gao, Huan Wang, Fusheng Yang

Abstract:

Given the data, we want to find the correlations and the distribution of the jobs and labor forces and to find if any of the variables is correlated with the number of applies.

We did the regression on applies verses avgoverallcompanyrating, jobagedays, descriptionLengthChars, and clicks. We used alpha level of 0.05 and concluded that jobagedays and number of clicks are relatively significant to number of applies. We also found that jobagedays has negative correlation with clicks. It is probably because of the negative correlation between jobagedays and candStatusReviewedCount.

To get some insight of distribution of jobs and labor forces, we analyzed the amount of regional jobs and regional education level requirement. We use python to visualize the data on the U.S map.

Conclusion:

Base on our analysis, we claim that the employers should update or re-upload their open positions frequently. For employees, states with high ratio between number of opening jobs and number of applicants would offer them a better chance to get the job. Employees could also make the decision by checking the distribution of regional education level requirement, corresponding to their own education experience.

Indeed: Clicks, Applies, and Job Opportunity

Team Members: Munkh-Erdene Baatarsuren, Ryan Cox, Evan Moore, Ilina Shah, Kristina Yamkovoy

Indeed connects users of all backgrounds across the US to jobs that suit their skills, with average daily clicks per posting serving as a direct proxy for interest in a job due the completely linear relationship between clicks and applies. We perform descriptive statistics to determine what industries have the most (and least) potential for employment opportunity using a metric of availability over interest we call "job opportunity rating".

We find that, while the average reviewed over applied ratio (ROA) is roughly .4 across all locations and industries, meaning only about 40% of resumes sent through Indeed are ultimately seen by the company, certain types of jobs exhibit more potential for growth than others due to the difference in available job postings compared to the average clicks per day for jobs in that industry.

Methods and Results

Job opportunity rating is a metric that describes, for each state and industry combination, the percentage of unique job postings for that industry in that state divided by the percentage of average daily clicks for that industry in that state, each relative to the total state values. This provides a normalized estimate of industry availability vs. interest for each state, which we then aggregate by industry and state and normalize on a scale of 1 to 10 to assign a job opportunity rating.

A low job opportunity rating (JOR) indicates that the ratio of applicants to job posts is higher, meaning that more people are in competition applying for the same types of jobs. Instinctively, the jobs with the lowest JOR are common jobs with lower barriers of entry such as customer service, administration, and janitorial/sanitation, which our findings describe as having the lowest rating.

On the other hand, a high JOR indicates that there are many postings for jobs in that industry, but relatively less interest for those jobs, meaning there is more opportunity for a single person applying to them. Jobs in technology and healthcare such as analyst, software engineer, and doctor which require higher education rank at the top of our list, so looks like DataFest participants are in luck! While these types of jobs make up a mere 3.5% of unique job postings in the dataset, our findings indicate that it is these jobs which have the highest potential for employment based on availability compared to interest.

Conclusion

Future employees looking to enter the workforce should focus on jobs in healthcare and technology that are currently undersaturated on Indeed, and consider states like Nevada and Georgia that currently lack employer interest in these industries. Employers looking to expand into high technology jobs should focus on states like North/South Dakota and New Hampshire that display high interest for these industries relative to the amount of opportunities present there. For license or no education requirement jobs, states like Montana and Oregon are worth focusing on due to the unemployment rate relative to number of job postings and applications.

Additional Findings

- **Outside resources:** Bureau of Labor Statistics - The number of unemployed individuals in each state.
- **The purposes:** Mainly for individuals with just license, high school, or below education looking for jobs in their local areas (by states).
- **Intuition:** Jobs that require only licensed jobs, high school education or no education tend to hire locals.
- **Explanation of Colors in Map:**
 - (i) The more red (dark red) it becomes, it means the more jobs are available for state residents with just licenses, high school or no education.
 - (ii) The more blue (dark blue) it becomes, it means the more active local applicants with just licenses, high school or no education are applying for jobs in their states.

General

- There is 1 candidate review per ~ 25.5 clicks
- On average, companies review 41% of the applications they receive on a given day, consistent among industry, company size, and state.
- People are more likely to apply for jobs in the first half of the week, with Wednesday having the highest number of applies, and Saturday having the lowest.
- People apply significantly more in the months of December and January in comparison to other months. People apply least in June, with applies being less than 1/3 of those in December. We hypothesize that holiday seasons may affect this result.

Job description length analysis

- There is no correlation between applies/clicks and job description length.
- The average length of job postings in characters is ~1651 characters (or ~250 words)
- The average length by territory ranges from 1514 (NH) to 2180 (DC)
- Larger companies tend to have longer job descriptions
- Direct employer postings more likely to have longer descriptions.

Job posting length analysis

- The average length of job postings is roughly 33 days or 1 month.
- Job length has negative correlation with average applies, clicks, and candidates reviewed. Jobs that are posted for longer tend to get less reviews on average since most applies happen in first week or so.

Raeesa Alam, Meredith Pan, Qia Qia Ji,

Naila Arsky, Caroline Li

Where Should You Work?

A common dilemma most people will have is competition in the job market when trying to find a job. To measure this competition, we grouped by industry and used the variable “applies” to indicate competition, since it referenced number of people applying to a job on a given day per industry, and divided by the variable “employerJobCount”, since it indicated count of job per employers by industry. As we can see, lower ratio means lower applies of job in comparison to the supply of jobs available, so the lower the ratio the smaller the competition to the amount jobs demanded by the industry for a given employer. We mapped the ratio for each industry over time to see the general trends, and found that many of the sectors with higher ratios (media, legal, government, agriculture, and work-at-home) are generally much more volatile than industries with lower ratios.

Since the majority of our group members are international students interested in landing a job in the U.S post-graduation, we considered how those seeking H-1B visas could optimize their job opportunity in the U.S. In other words, what are the best states to for international students look and apply for jobs and get H-1B visas approved? So, we joined the Indeed data with an external H-1B visa dataset. We then aggregated the chance of getting visa ratio by each state and mapped it. The darker the state appears, the more chances of applicants getting H-1B visas in that state.

Data Champs 2

Review System Critique

Our group focused on finding data on the reviewing system used by indeed.com. What we concluded is that a five star system is overall ineffective in expressing information of interest to its users. Because five star rating systems are used largely for consumer items and to apply it to jobs is already a conflict of interest? But what data is available on the rating systems though? This is what our presentation will show?

Preprocessing Data

- Selected a number of random samples from the entire dataset
 - Reservoir sampling to uniformly and randomly select samples
 - Used sample sizes of 10,000 and 500,000 samples
- Select a number of columns used as features to be used for machine learning models
 - Choose data that are related by job details and popularity
 - Average overall company rating, Description Length Chars, Clicks, Applies, Education Requirements, Cities, States
- Longitude and Latitude references for Cities
 - Use Google Maps to read in longitude and latitude strings
 - Show education requirement trends in United States regions

Machine Learning Models

- Decision Trees
 - Non ensemble learner for random forest classification. Mainly splits off one feature of the data. The split for children nodes is not always binary(20% accuracy)
- Random Forest Classification
 - Ensemble learning method consisting of many classification trees on subsets of the entire data set(10% accuracy)
- Support Vector Machines(Classification)
 - Minimize the distance between each data point with the line. Only able to run for the 10,000 samples(10% accuracy)

Data Visualization

- We tried to make a data visualization through querying google maps API
- By querying google maps API, we are able to get the longitude and latitude points of each major city
- Preprocessing the data through grabbing certain columns with querying proved to be too much
 - Columns queried: noEducationRequirementsJob, highSchoolEducationRequirementsJob, higherEducationRequirementsJob

For the Future

- Try grabbing more data and features to be able to distinguish between different classes
- The accuracy of the model increased with more data but the processing time also increased
- Use cluster to speed up computation to balance workload versus wait time
- Supply company activity information to users
- Analyze job type and date posted in relation to current events