Team 7: Q(u)ac(ks)
Carlo Medina, Frederick Corpuz, Kelly Jamrog, Tiffany Coons

Datafest 2018

Intro
We focused on trying to help a job seeker have a better experience on Indeed.com through various visualization techniques. For a new college grad entering the workforce, it can be overwhelming looking at all the opportunities around the country for post-grad life. We have developed an app to help this type of user narrow down their search by location and tell them the optimal time to apply.

Interactive Map & Dynamic Table
To help a new job seeker determine where they want to look for jobs, we created a shiny application that allows the user to select characteristics of both the town/city they want to live in and the job they want to have. The user can zoom in and out of the graph, created with the leaflet package, to better visualize where in the country the city's indicated are and the table below are the top 10 best cities based on the above calculations. Users can specify if they want to live in a small town or large city, if they want a below average, average, or above average percentage of residents between the ages of 21 and 30 and percentage of residents over the age of 65. They can also limit the search by job industry, experience level, and education requirement. Using these criterias, the app calculates the the top ten locations around the country that has the best 'standard of living'. This standard of living index is created by looking at the the mean estimated salary based on the selected job criteria and mean average home price in a city. This ratio of average annual salary to average home price helps show where in the country a new grad is most likely make money (after taking cost of living into account).

Time Series graphs
To give the job seeker an idea of when he/she should start preparing for applications for jobs, we have a tab on the Shiny app to visualize the 'flux of job postings' and 'mean salaries' across one year that can be customized to a specific job title, education attainment and work experience.
On the calendar and the upper time series graph, the user can see inflow and outflow (flux) of job postings. Green indicates that more new jobs are posted/aggregated by Indeed.com on that day. Red indicates that more jobs were removed from Indeed.com on that day. We calculated the flux from *jobAgeDays* grouped by *jobId*. The date of the *jobAgeDays* with the least value was taken to be the post date and the date with the greatest value was taken to be the removal date. The data used to generate the graphs were normalized within each *jobId*. The user looking at the 'flux' charts can observe when in the year employers are posting new job positions (green) which indicates when he/she needs to start preparing. Additionally, the user can see whether the job the he/she is interested has new job postings consistently all year (i.e. stability) by observing the runs of green and red throughout the year.
On the lower time series graph, the user can see the mean *estimatedSalary* of job positions that he/she is interested in and qualified for. We generated the 25 distinct job position by truncating *normTitleCategory* after the first comma. We calculated the mean of *estimatedSalary* grouped by job position. The data used to generate the time series graph was scaled within each job position. The user looking at the 'mean salaries' graph can observe when in the year jobs he/she is most interested in are being offered at higher starting salaries than usual.

Data Munging
In the given dataset, the industry variable was 89.6% missing. Since this is an important variable that can help narrow down a person's job search, we decided to try and predict the industry using other known information in the dataset. Using the 57 distinct values of *normTitleCategory,* we were able to impute the missing industry classification with 44% confidence.

External data
Several external datasets were brought in to create this analysis. Average home prices were brought in from readily available Zillow datasets and matched in by city and state. Population data and age group breakdown was found through the US Census and matched by zip code and then aggregated by city, state. To help match all the different datasets, we geocoded the city, state locations to obtain both zip codes and latitude and longitude values.

Takeaways
Given the limited data on job posting patterns, we created an interactive visual application that allows the user to explore the best places to live for a the user's desired job specifications and living location. The uses of such an app are many: life planning, trend identification, and business modelling. With more accurate industry/job type data and additional years, this application would be extremely useful in many settings.