Milestone 1 Report

Sandy Straub, [Sharvil Shastri](mailto:sharvil@umd.edu), Erik Rye

**1a.Research Questions** **[10pts]**

Our team is examining the interplay between various features of job postings and how they affect the overall goal of a job advertisement -- hiring a new employee. Toward that end, we obtained a public LinkedIn job advertisement data set from Kaggle from which we will attempt to answer some of the following research questions:

1. Can we predict the number of applicants a job post will receive given the number of views that it receives?

This question attempts to understand how the visibility of a job post affects the number of people that end up actually applying for the job. While it may seem intuitive that the more views a job advertisement has, the more applications are submitted – perhaps companies pay for particularly hard-to-fill roles to be boosted. These ads might therefore be seen by more people than popular jobs that qualified applicants know to seek out.

1. Can we predict the number of applicants based on the job's salary information, including both salary range and whether the salary is listed vs unlisted?

Whether job advertisements include salary information is a hot topic – California and Colorado have both recently passed laws mandating that employers list a salary range for the roles they're hiring for. This question aims to understand whether applicants act on this information, or decide to forgo applying when a salary isn't listed on the advertisement.

1. Can we predict a job's salary based on the industry the job is from and where the job is located?

Salary is often an applicant's most pressing concern -- but a dollar earned in Des Moines takes you farther than the same dollar earned in San Francisco. Given an industry and job location, can we predict the salary for a job posting, or does a doctor get paid about the same regardless of where she works?

1. Can we predict the number of views a job advertisement will receive based on the size of the company posting the ad?

Competition for qualified labor, particularly in high-demand fields like tech and medicine, is frequently fierce. Do large companies dominate their competitors when it comes to getting job postings in front of potential new employees? Or do their smaller rivals stand a chance at getting their ads in front of new recruits? This question attempts to understand power dynamics between larger, more well-established companies, and their smaller challengers.

**1b. State of the Art [5pts]**

**Košťálová, Z., Lyócsa, Š., Štefánik , M., 2022.**

[**“Online job vacancy attractiveness: Increasing views, reactions and conversions”**](https://www.sciencedirect.com/science/article/pii/S156742232200076X)

In their research regarding Online Job Vacancy (OJV) attractiveness, Košťálová, Lyócsa, and Štefánik set out to answer the following questions:

1. “Can we improve the predictability of OJV attractiveness?” and
2. “Which variables are useful in doing so?”

In their study, they defined the attractiveness of an OJV based on the number of views, the number of reactions (such as the number of completed applications), and the ratio between them. Using a sample size of 32482 OJVs, and comparing up to 175 job features, they determined that there is a significant importance in including information in the OJV regarding job classification and benefits, job description and title.

**Brothwell, Pat. “**[**2023 Job Application Statistics.**](https://blog.hiringthing.com/job-application-statistics#:~:text=Glassdoor%20found%20that%2C%20on%20average,job%20from%20those%20who%20apply)**” HiringThing.**

This article summarizes some relevant statistics regarding 2023 Job Applications. It highlights the importance of the candidate experience and the interview experience. The most important negative experiences, according to this article, are:

1. Initial application takes longer than 20 minutes. (60% will quit)
2. Not disclosing how long the application will take before they start it. (76%)
3. Lack of information about pay (50%)
4. Scheduling issues .(50%)
5. Technical difficulties (60%)
6. Not hearing back after application (76%), longer than 2 weeks (66%)
7. Not explaining the culture (72%)

Even with a good interview experience, 35% will turn down a job because they could not do remote/flexible schedules. In summary, LinkedIn users rank the 3 most important considerations when accepting a new job:

1. Compensation (49%)
2. Professional Development (33%)
3. Better work/life balance (29%)

**Shumway, Emilie. “**[**4 in 5 workers say they’re unlikely to apply for a job without salary range.**](https://www.hrdive.com/news/workers-demand-salary-transparency/641305/#:~:text=According%20to%20a%20recent%20survey,explain%20how%20pay%20is%20determined)**” HRDive. Jan. 26, 2023**

Pay transparency is a contentious subject between employers and job seekers. While there is an increase in workers demanding it, at least ⅓ of employers are not ready for it. Most recently, California, Rhode Island, and Washington have enacted laws around pay transparency, bringing the total to 8 states in the USA..

On the employer side, there are a variety of benefits to pay transparency, including the increase of pay parity, which can promote DEI goals for many companies. Additionally, disclosing pay ranges can reduce recruiting costs, since more applicants are willing to apply. Unfortunately, some drawbacks of pay transparency revolve around worker envy and having applicants demand the top of the range.

**Mayer, Kathryn. “**[**Want to Attract Job Seekers? Make Sure to Post a Salary**](https://www.shrm.org/resourcesandtools/hr-topics/compensation/pages/workers-unlikely-to-apply-to-a-job-without-a-pay-range.aspx)**.” SHRM.org. March 6, 2023**

Job seekers are increasingly prioritizing pay transparency. A report from ResumeLab revealed that 80% of workers are unlikely to apply for a job that doesn't specify a salary range. Moreover, 77% believe it should be legally mandated to include a salary range in job postings, and 80% think employers should clarify how pay is determined. The importance of salary disclosure varied based on education level, with 89% of those with master's degrees emphasizing its necessity compared to 66% without a college degree.

Companies not providing salary ranges may face reduced applicant pools, while those embracing transparency gain a competitive edge in attracting quality candidates. The push for pay transparency has grown due to legislation and social pressure from platforms like TikTok and Gen Z advocating for openness around pay. Despite concerns about administrative burdens and pricing out candidates, companies that disclose salaries are seen as more trustworthy and gain an advantage in the job market.

**Umoh, Ruth. “**[**Why you’re more likely to get hired at companies with fewer than 500 employees.”**](https://www.cnbc.com/2018/05/25/why-job-seekers-should-apply-to-companies-with-fewer-than-500.html) **CNBC, May 25, 2019**

According to a survey by TalentWorks, small and mid-size companies present better hiring opportunities compared to larger corporations like Amazon and Google. Analyzing over 6,900 applications across various industries and U.S. cities, the study found that applicants to companies with fewer than 500 employees had a 192% higher interview rate. This trend is especially beneficial for candidates with resume blemishes or resume gaps.

The research also highlighted two key reasons for this trend: larger companies have stricter application filters and more bureaucratic HR processes, while also facing higher competition due to a larger applicant pool. Working at smaller businesses offers added benefits such as hands-on experience, flexibility, and easier advancement. As a result, more young professionals are opting for startups, which can provide similar perks as larger corporations.

**1c. Datasets [10pts]**

We used a dataset from Kaggle called “[LinkedIn Job Postings - 2023](https://www.kaggle.com/datasets/arshkon/linkedin-job-postings?resource=download)”. As the name suggests, it consists of over 15,000 unique job postings drawn from a 2-day period on LinkedIn during the summer of 2023.

The dataset consists of 8 unique csv files. The primary file, “job\_postings.csv”, is a CSV that is comprised of the 15k job postings scraped from LinkedIn. The additional files are separated into two subdirectories – company\_details, and job\_details. company\_details consists of four CSVs – companies.csvs, which maps a company ID to the name, description, and other company-specific data, company\_industries.csv, which maps a company ID to the industry the company is in, company\_specialities, which maps the company ID to one of more specialities the company has, and employee count, which maps a company ID to the number of employees and followers it has (according to LinkedIn). Within the job\_details subdirectory, benefits.csv maps a job to one or more benefits that the job provides, job\_industries maps a job to an industry ID, and job\_skills maps a job ID to one or more abbreviations for skills (e.g. ACCT for accounting) that the job requires.

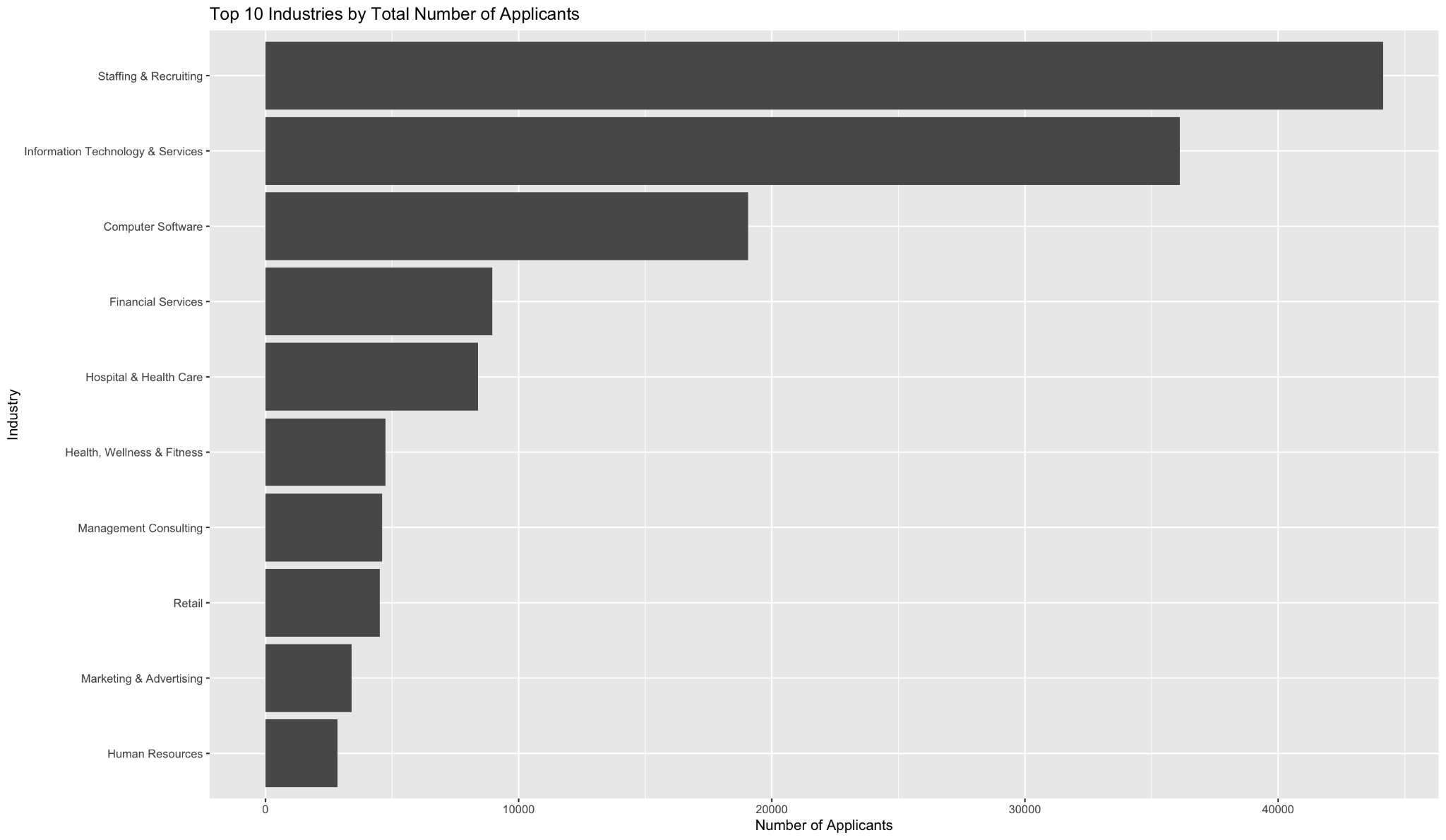
Of these 8 CSVs, we will primarily focus on job\_postings.csv, but enrich it with data from the other ancillary datasets when necessary. For example, a job to an industry requires mapping the job\_id to a company’s industry through the company ID column which exists in both (the job\_industries.csv mapping between job IDs and industry IDs lacks an explanation of what the industry IDs mean).

In order to answer our research questions, we will focus on 6 variables from our dataset. We describe each of these variables next in turn:

**Number of Applicants Data:**

In our Job Postings data, there are more than 150,000 total applicants. The ‘applies’ column in the job\_postings.csv file provides the data of the number of applicants for each job posting. There are more than 150,000 total applicants. Mean number of applicants is 23.04 and max value is 1615

We can observe that the Staffing & Recruiting Industry had the most number of applicants close to around 45000, followed by Information Technology & Services. However, there is a significant decline in the number of applicants for the subsequent industries.

****

For the data collection process for this bar plot, the job\_postings.csv and company\_industries.csv files were merged using merge() function since they contain a common column (or a primary key). Then, the number of applicants by each industry was aggregated using the aggregate() function in R and then plotted on a bar graph using ggplot() and geom\_bar(), while simultaneously sorting it in a descending order and providing the appropriate labels with labs().

**Job Salary Data**

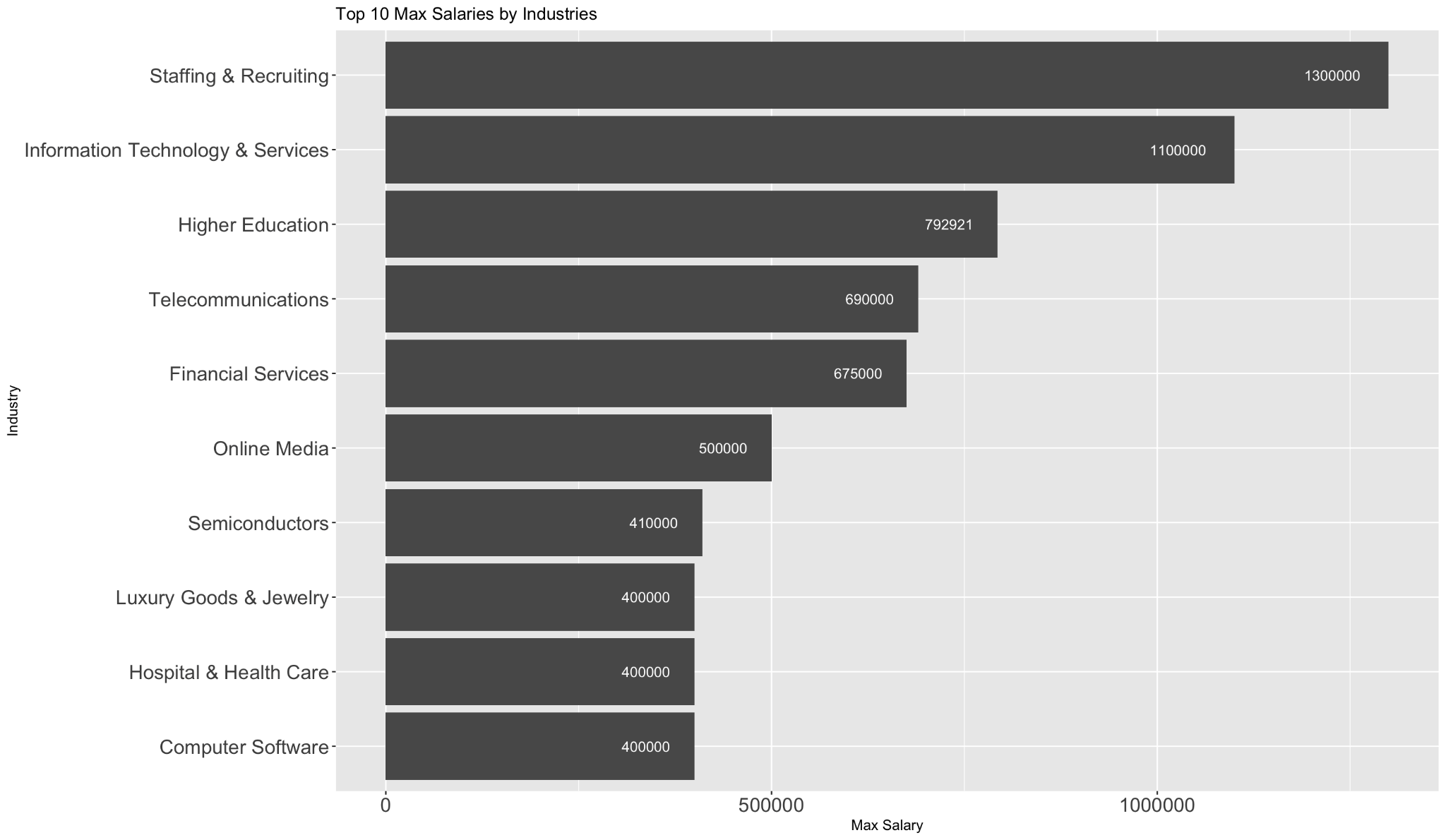
The salary is one of the most important aspects while applying for a job. In our dataset, there are 3 separate columns for salaries.

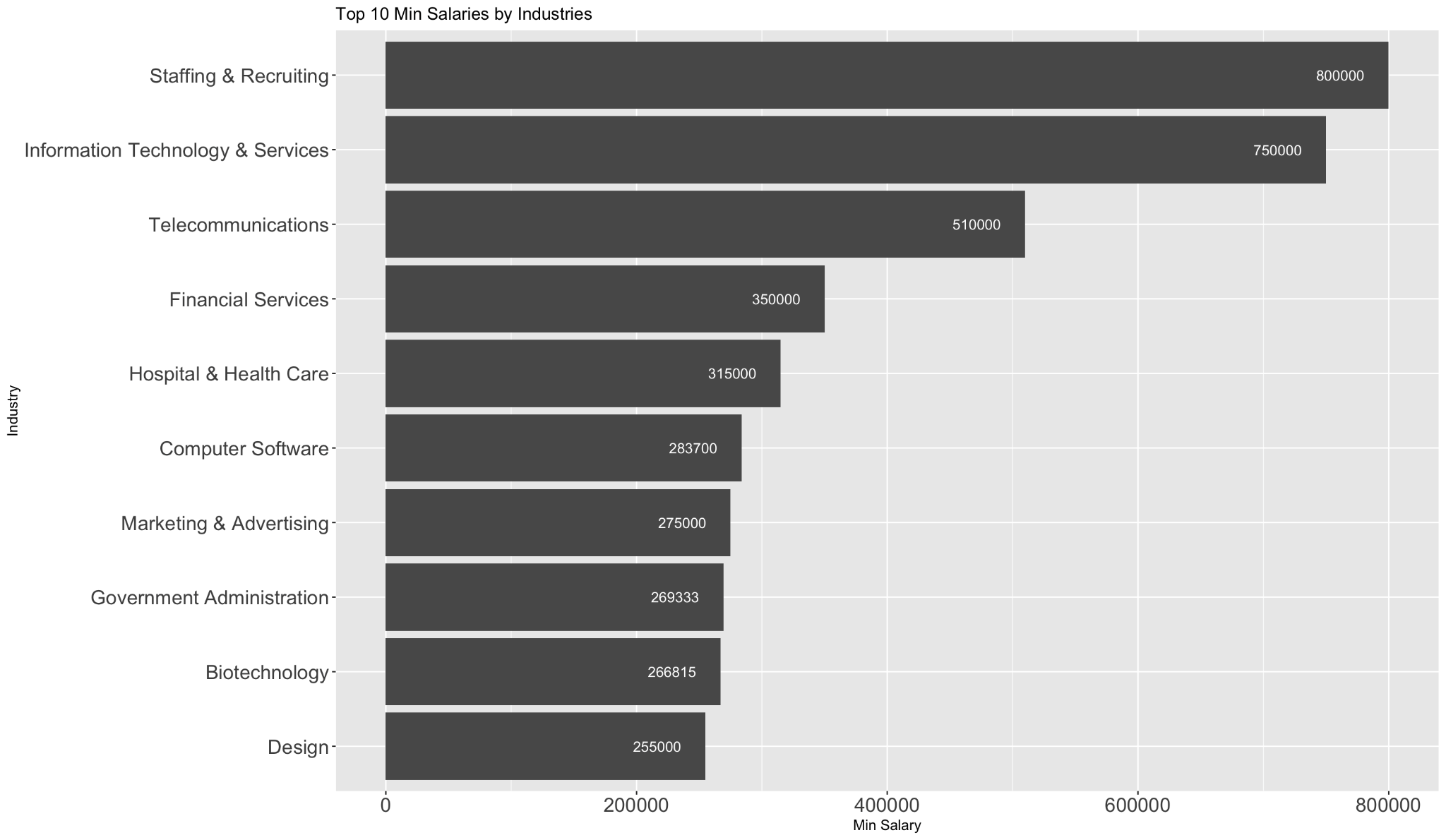
* min\_salary
* max\_salary
* med\_salary

General Statistics for Job Salary Data:

Mean of the max\_salary column is 88836 while the Max is 1,300,000. Mean Maximum Salary for yearly pay period is 133505 whereas the highest salary for a job in our dataset is 130000 for yearly pay period. Around 300 job postings have not provided the max\_salary in this case. For the hourly pay period, Mean max\_salary is 37.26, maximum max\_salary is 250 and ~ 650 hourly pay period applications did not provide a max salary.

Mean of the min\_salary column is 62352 whereas the Max is 800000. For the hourly pay period, Mean min\_salary is 30.34, maximum value is 200 and approximately 650 job applications have not provided the minimum salary. For the yearly pay period, Mean min\_salary is 94244 Max is 800000 and 301 job postings have not provided a minimum salary.

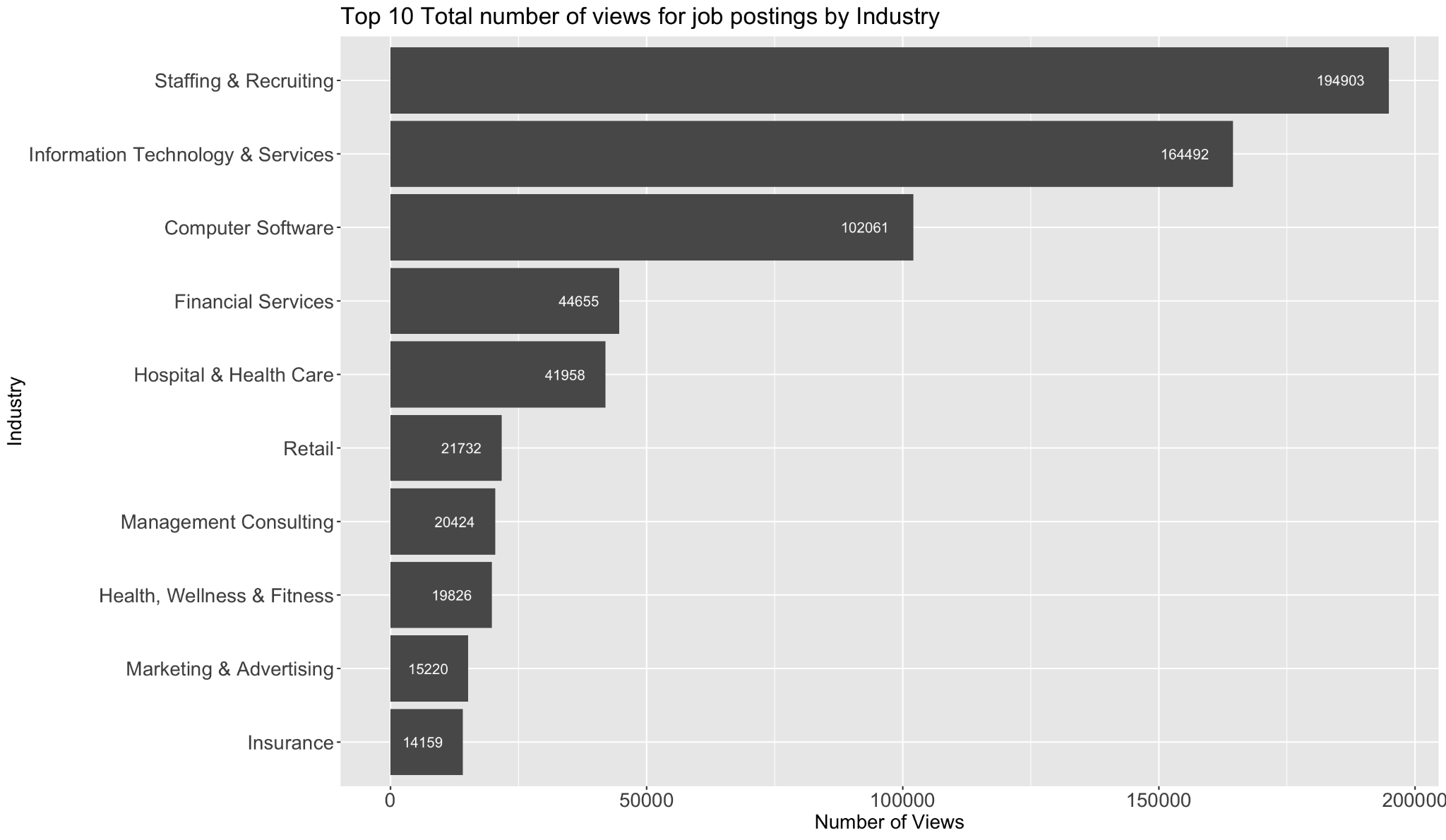




We also had to separate out the yearly salaries and hourly salaries while generating the general statistics summaries and also creating these representations. For that, we used the filter() function of the ‘dplyr’ library, and specified pay\_period column as either “YEARLY” or “HOURLY”

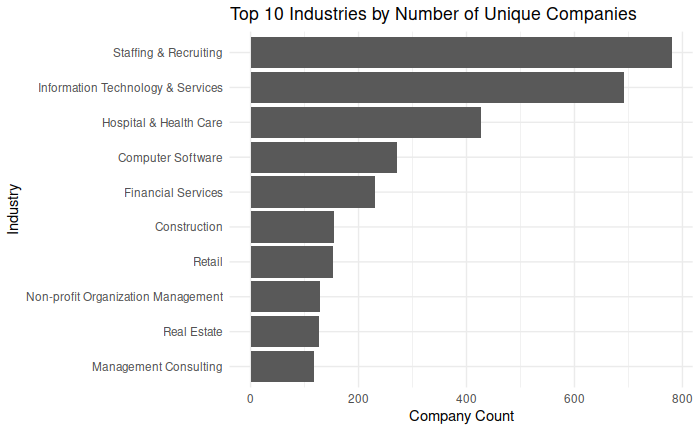
**Number of Job Posting Views**

We then turn our attention to the number of times a job posting was viewed. This is provided by the column “views” in our dataset. Mean is 77.49, Maximum value is 5656 and around 2700 job postings do not have any views. We again aggregated the data and have represented the Total number of views for job postings by specific industries just to get a quick overview.

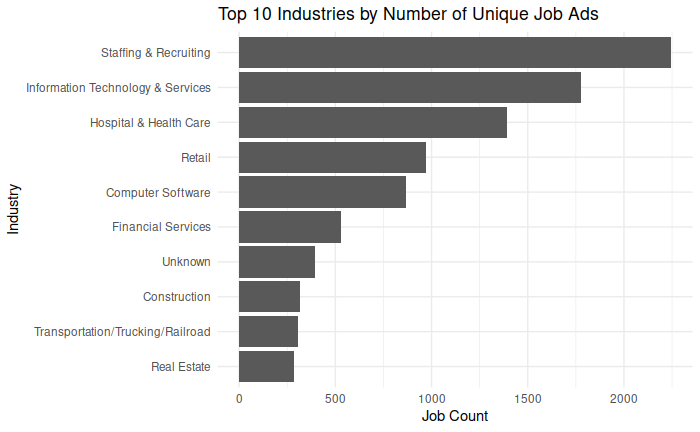


**Job Industry Data**

Our data set includes a CSV called company\_industries.csv in the company\_details/ subdirectory that contains a mapping between company\_id, which is the primary key for each company, and the string industry that company is a part of. We have industry data for 6,003 company\_ids, with a total of 141 different industries those 6,003 companies map to. After joining the company\_industries.csv with our job\_postings.csv data, we were able to determine that the most common industry advertising jobs on LinkedIn when our data was collected was "Staffing & Recruiting", followed by "IT", "Health Care", and "Retail."



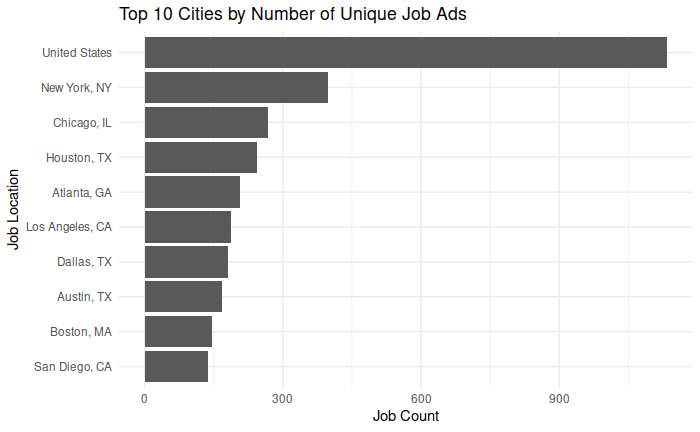
The number of unique company\_ids per industry in our company industry data set.



The number of unique job\_ids per industry in our job posting data set.

**Location Data**

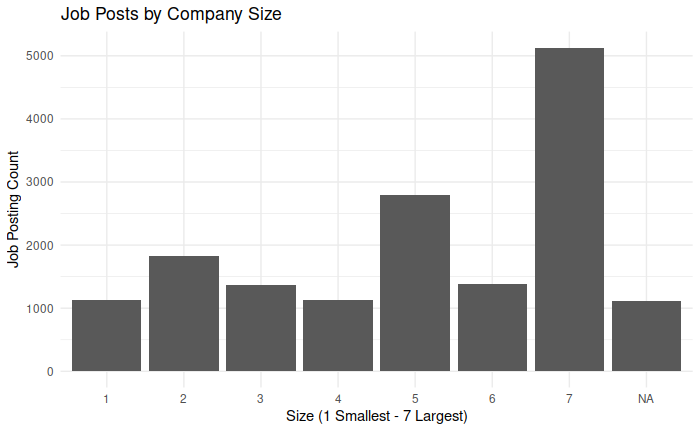
Next, we turned our attention to the locations that the job advertisements listed. "location" is a field in our primary data table, "job\_postings.csv", so we analyzed this column of the data after reading it into a data table. There are 3,010 distinct locations listed for the 15,886 jobs listed in our data set. The general "United States" is the most common location, with 1,133 different job\_ids listing it as the location. Most jobs list a city and US state, however, with New York, NY being the city with the largest number of jobs listed at 398. 1,502 cities only appear in one job listing. The median number of jobs listed per city is 1.5 and the mean is 5.28.



The number of distinct jobs listed by LinkedIn job posting location

**Company Size Data**

Last, we approached our job posting data set by examining the number of jobs that companies post according to the relative size of the company. Our data set consists of 7 different labelled sizes (1-7) which correspond to relative company size, where 1 is the smallest and 7 is the largest. We counted the number of unique job\_ids that correspond to each company size to get a sense of what types of companies are posting the fewest and most LinkedIn ads. Our results show that the largest company size (7) posts nearly five times as many LinkedIn ads as companies in the smallest bracket. This is perhaps unsurprising; the largest companies likely have the largest budgets for recruiting and also have a large number of positions to fill, necessitating more recruiting.



Number of job advertisements posted by company by size of the company

**1d. Data Cleaning Efforts [5pts]**

Some companies do not have an identifier that maps to an industry in the company-industry mapping. Therefore, to properly display the number of unique job\_ids per industry, we had to use a default value (“Unknown”) when no industry data was available for job postings attributed to those companies.

Additionally, there are a lot of duplicates in the company\_industries.csv file which caused issues while joining the two files. So, the duplicates had to be removed using the unique() function.

Several times, when aggregating data, if the fields had a NA value, it would create miscalculations for the computations. So, for the rows which contain NA values, whenever we had to group them and aggregate a particular function on it, we rectify it by using the attribute na.rm = T.

We also made use of the Pipeline Operator, %>% while using the functions to make the programming part a bit more convenient.

Another issue we encountered is that while plotting some of the charts, the number were represented in scientific notation, i.e they were represented in exponential forms instead of the direct value. To rectify this, we used the format() function.

format([dataframe name], scientific = F)

**1e. Other Software Engineering Efforts [2pts]**

We had initially intended to use similar job posting data from the Smith Business School, where Sandy works. This data would have allowed us to make predictions about job placement relevant to UMD students, using anonymized data to understand how academic achievement and test scores affect the types and kinds of jobs that UMD students and alumni are interested in and hired for.

Unfortunately, despite expressed interest in the project by administrators with the Smith Business School, the stewards of the data we were interested in ultimately rejected our application to use it. Encountering this problem highlighted for us a real-world issue with data analysis – obtaining relevant and interesting data!

Finally, we created a git repository to maintain and version-control our source code.

**2. *Code (18 pts)***

The code needs to have comments that explain what each routine does. Please indent and comment your code and be well organized. Also, add a README file explaining how to run your code, and remember to add all the necessary code and data files you have used to the zipped submission file.

**This comes about naturally from looking at 1c.**

1. ***Presentation (10pts)***

**You are expected to record a video with your project presentation and to upload it to Youtube. The duration of the presentation should be at most 10 minutes. Please make sure you share the Youtube link with all the class by adding it to [this](https://docs.google.com/document/d/1_4IE87b7E2CdLNr732LMTFyoi-Ynu-_NN2PbRPqtXuk/edit?usp=sharing)**

**[Links to an external site.](https://docs.google.com/document/d/1_4IE87b7E2CdLNr732LMTFyoi-Ynu-_NN2PbRPqtXuk/edit?usp=sharing) shared file.**

**Contributions:**

1a: All members helped formulate the research questions

1b: Sandra researched from different journals and articles

1c. Erik and Sharvil formulated and wrote about the dataset and different variables used in detail

1d: Sharvil outlined the different steps taken in data preparation and cleaning operations

1e: Sandy worked to obtain initial data set, which was then replaced with Kaggle data

2: Erik and Sharvil wrote the code for general statistics and visual representations that can be seen for different prominent variables.

3: All three members recorded their respective parts and then Sharvil merged the video and it was then uploaded to a Youtube Channel.

All in all, all group members contributed in a meaningful and equitable way.