Milestone 1 Code

Erik Rye

2023-09-23

Milestone 1

1a Research Questions

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:

- 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, 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.

- 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 forego applying when a salary isn't listed on the advertisement.

- 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 De 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?

- 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.

1c. Data sets

Job Industry Stats

Our data set includes a CSV called `company_industries.csv` in the `company_details/` that contains a mapping between `company_id`, which is the primary key for each company, and the industry that company is a part of. We have industry data for 6,003 `company_id`s, 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 was “Staffing & Recruiting”, followed by “IT”, “Health Care”, and “Retail.”

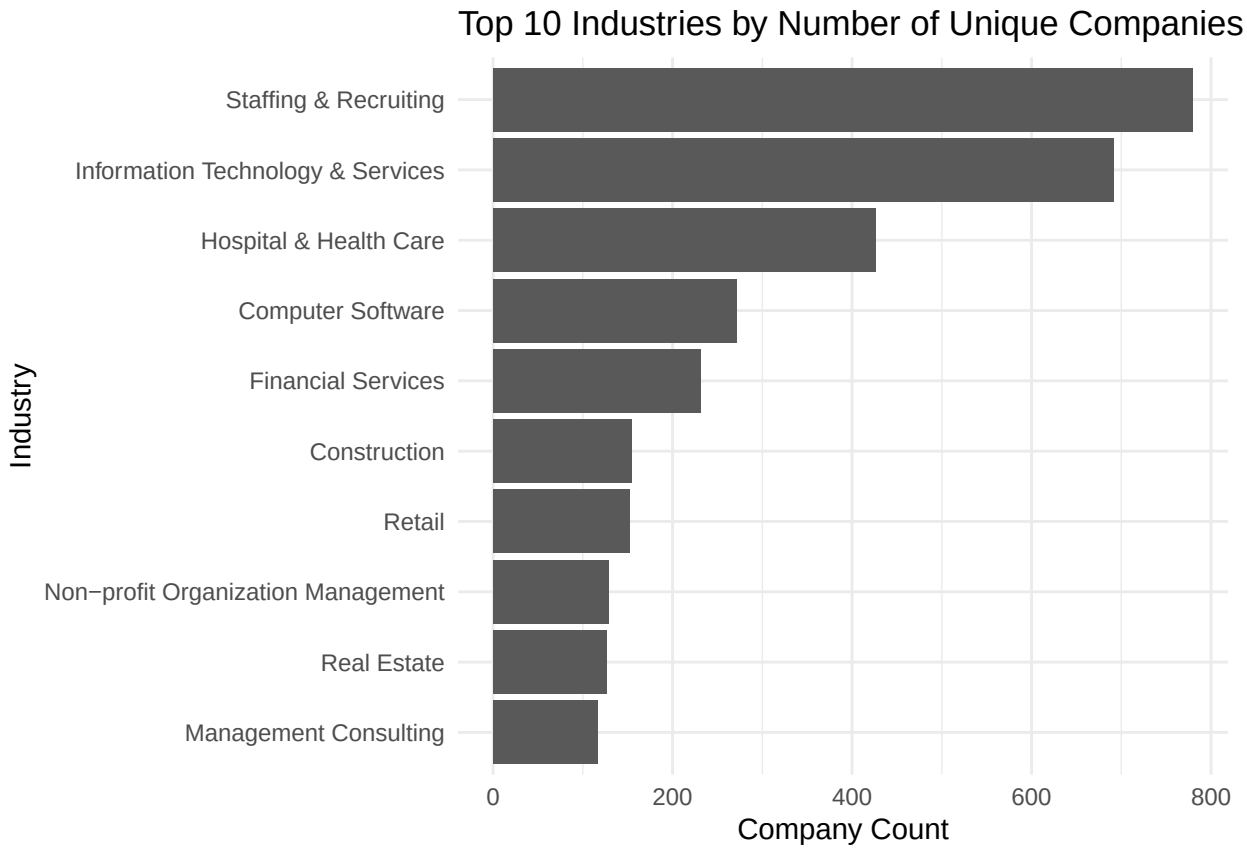
```
################################################################################
# Job industry code
################################################################################

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(stats)

setwd('/home/erik/repos/gigaryte/INST737-Team-8')
# Load the company industry info into a dataframe
industries <- read.csv('company_details/company_industries.csv')

# Create a dataframe of the industry-counts of unique company IDs.
# Order them descending by the count
industry_counts <- industries %>%
  group_by(industry) %>%
  summarize(company_count = n_distinct(company_id)) %>%
  arrange(desc(company_count))

# Select the top 10 industries
top_10_industries <- head(industry_counts, 10)

# Create a bar plot
ggplot(top_10_industries, aes(x = reorder(industry, company_count), y = company_count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x = "Industry", y = "Company Count", title = "Top 10 Industries by Number of Unique Companies") +
  theme_minimal()
```




```
# Summarize the number of total companies
counts <- sapply(industries, function(x) length(unique(x)))
show(counts)
```

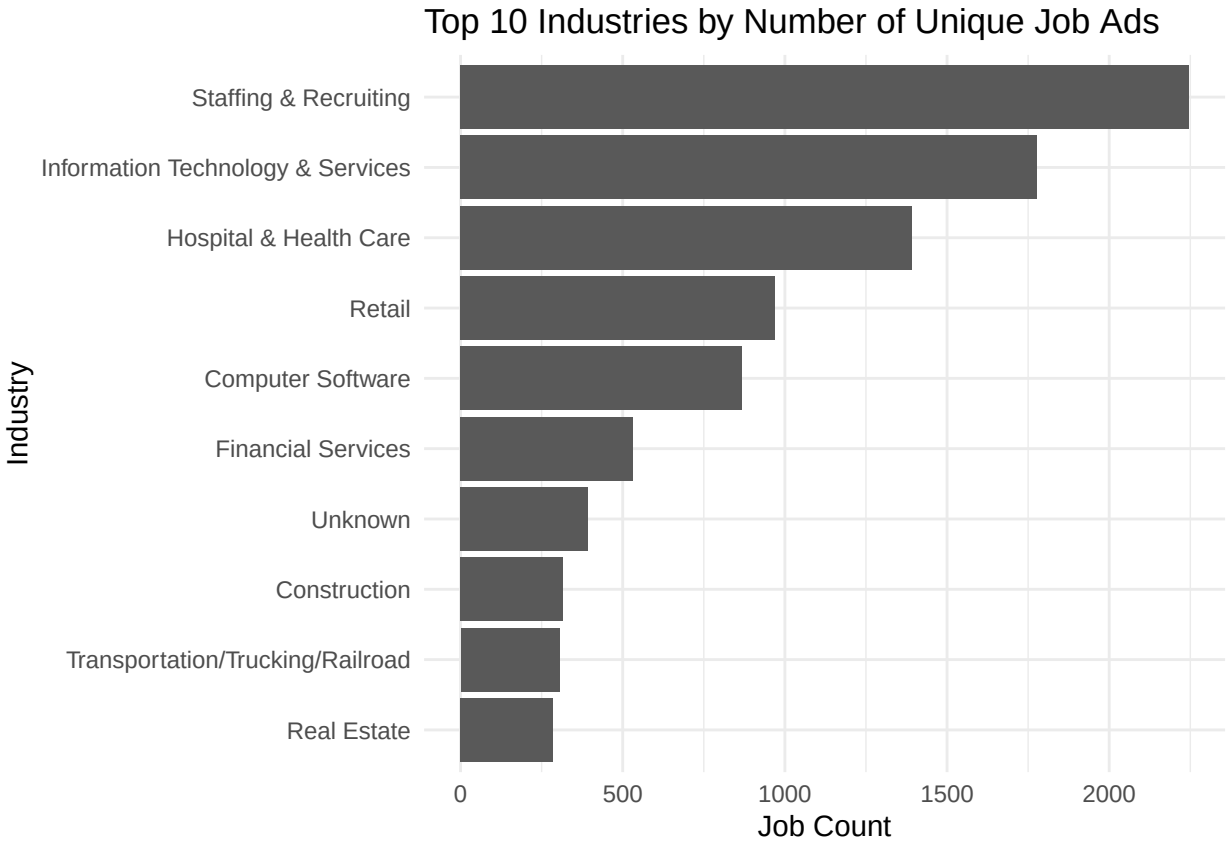
```
## company_id   industry
##       6003        141
```

```
# Load the jobs CSV into a dataframe
jobs <- read.csv('job_postings.csv')


# Merge the two on company_id
job_industries <- left_join(jobs, industries, by = "company_id")
```

```
## Warning in left_join(jobs, industries, by = "company_id"): Detected an unexpected many-to-many relati
## i Row 9 of `x` matches multiple rows in `y`.
## i Row 9248 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
# Clean up entries with no industry
job_industries$industry[is.na(job_industries$industry)] <- "Unknown"

# Summarize jobs by industry now

# Create a dataframe of the industry-counts of unique job IDs
# Order them descending by the count
job_counts <- job_industries %>%
  group_by(industry) %>%
```

```
  summarize(job_count = n_distinct(job_id)) %>%
  arrange(desc(job_count))

# Select the top 10 job industries
top_10_job_industries <- head(job_counts, 10)

# Create a bar plot
ggplot(top_10_job_industries, aes(x = reorder(industry, job_count), y = job_count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x = "Industry", y = "Job Count", title = "Top 10 Industries by Number of Unique Job Ads") +
  theme_minimal()
```



Job Location Data

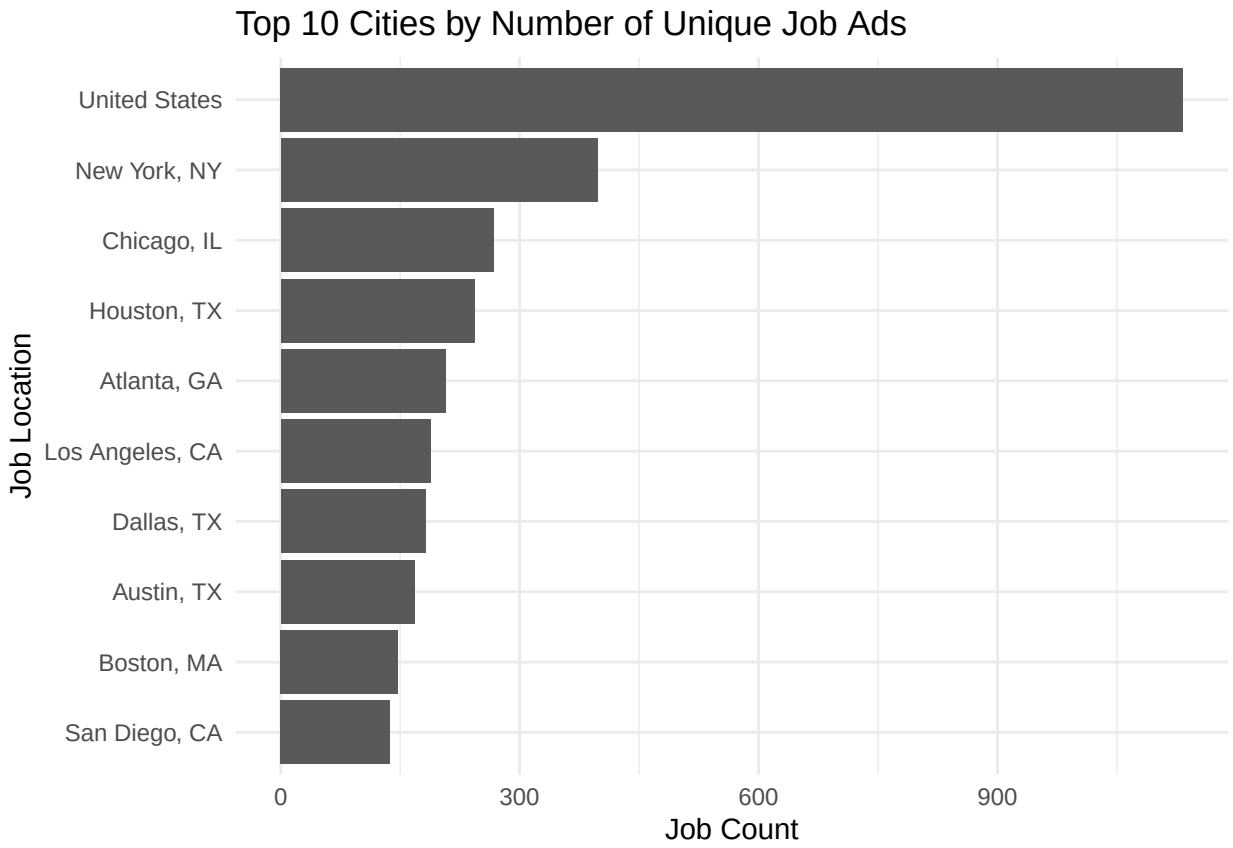
Next, we turned our attention to the locations that the job advertisements listed. "location" is a field in our primary data table, "job_postings.csv", so we analyzed this column of the data after reading it into a data table. There are 3,010 distinct locations listed for the 15,886 jobs listed in our 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.

```
library(dplyr)
library(ggplot2)
library(stats)
################################################################################
```

```r
# Job location data
################################################################################

setwd('/home/erik/repos/gigaryte/INST737-Team-8')
# Load the jobs CSV into a dataframe
jobs <- read.csv('job_postings.csv')

# Create a dataframe of the counts of job_ids by locaion
location_counts <- jobs %>%
  group_by(location) %>%
  summarize(location_count = n_distinct(job_id)) %>%
  arrange(desc(location_count))

# Print some stats
summary(location_counts$location_count)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.000    1.000    1.500    5.278    3.000 1133.000
```

```r
# Get the top 10 again
top_10_locations = head(location_counts, 10)

# Create a bar plot
ggplot(top_10_locations, aes(x = reorder(location, location_count), y = location_count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(x = "Job Location", y = "Job Count", title = "Top 10 Cities by Number of Unique Job Ads") +
  theme_minimal()
```



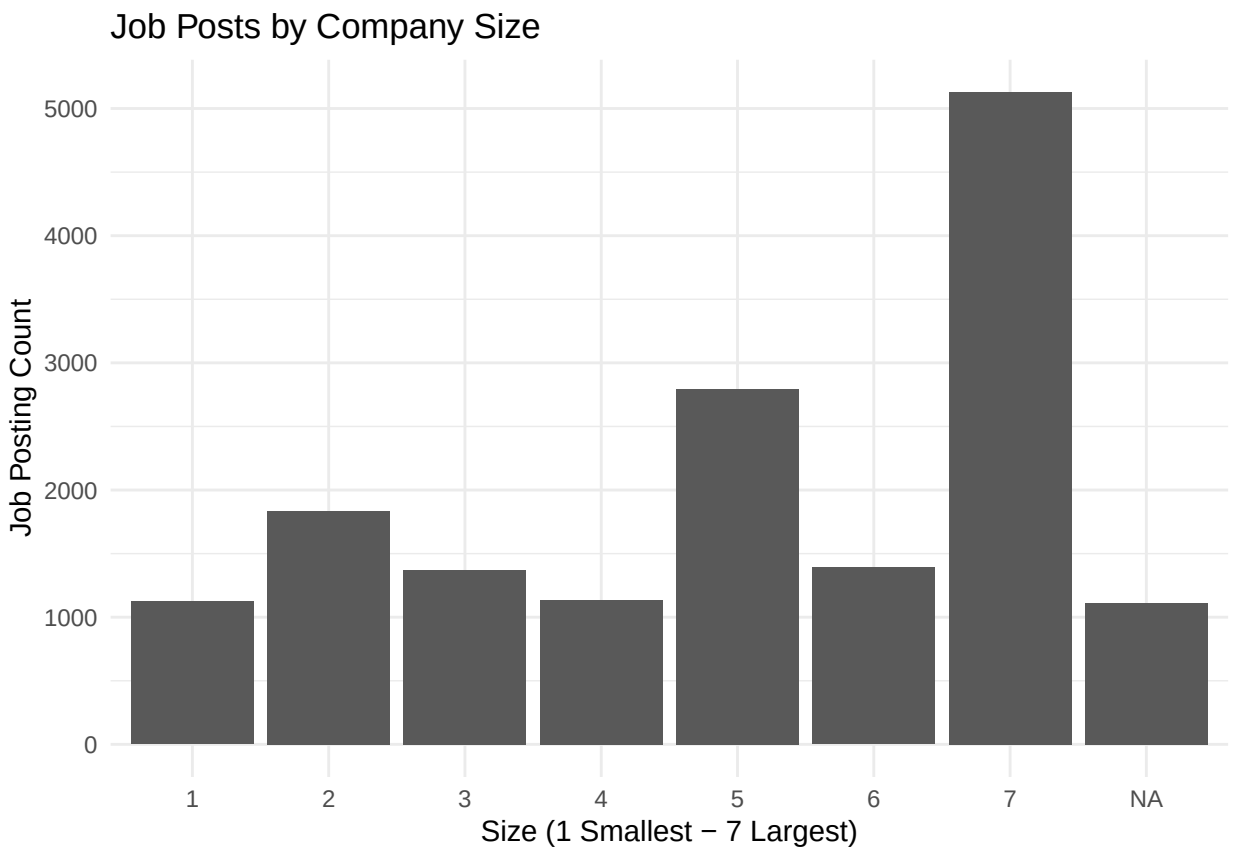

```
# Cities that only have 1 job post :(
ones <- jobs %>%
  group_by(location) %>%
  summarize(location_count = n_distinct(job_id)) %>%
  filter(location_count == 1)
```

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.

```
library(dplyr)
library(ggplot2)
library(stats)
##############################################################################
# Company size  data
##############################################################################

setwd('/home/erik/repos/gigaryte/INST737-Team-8')
# Load the companies CSV into a dataframe
companies <- read.csv('company_details/companies.csv')
```

```r
# Load the jobs CSV into a dataframe
jobs <- read.csv('job_postings.csv')

# Merge the two on company_id
job_companies <- left_join(jobs, companies, by = "company_id")

# Create a dataframe of the counts of job_ids by location
size_counts <- job_companies %>%
  group_by(company_size) %>%
  summarize(size_count = n_distinct(job_id)) %>%
  arrange(desc(size_count))

#Plot the results
ggplot(size_counts, aes(x = factor(company_size), y = size_count)) +
  geom_bar(stat = "identity") +
  labs(x = "Size (1 Smallest - 7 Largest)", y = "Job Posting Count", title = "Job Posts by Company Size
  theme_minimal()
```




```r
summary(size_counts)
```

```
##   company_size   size_count
##  Min.   :1.0   Min.   :1111
##  1st Qu.:2.5   1st Qu.:1132
##  Median :4.0   Median :1381
##  Mean   :4.0   Mean   :1986
##  3rd Qu.:5.5   3rd Qu.:2073
##  Max.   :7.0   Max.   :5127
```

```
##  NA's   :1
```