Project Proposal

JRLK - Jess Ringness, Rebekah Kim, Laura Cai, Karen Dong

```
library(tidyverse)
library(tidymodels)
job_postings <-read.csv("data/job_postings.csv")
benefits <-read.csv("data/benefits.csv")
employee <-read.csv("data/employee_counts.csv")
```

```
benefits <- benefits |>
  select(-inferred) |>
  mutate(count = 1) |>
  pivot_wider(names_from = "type", values_from = "count")
```

```
jobs_employee <- job_postings |>
  left_join(employee, by = join_by("company_id"))
```

```
linkedin <- jobs_employee |>
  left_join(benefits, by = join_by("job_id"))

linkedin <- linkedin |> distinct(job_id, .keep_all = TRUE)
```

Introduction

LinkedIn is a platform that connects companies, big and small, and professionals spanning various levels of experience. As a result of its growing popularity, there are thousands of active job postings available on LinkedIn, with more consistently added. Numerous variables may impact the popularity of a job posting – an employer's goal is likely to attract more outreach on their job posting to attract more applicants, which could potentially lead to more qualified candidates to apply for their company's roles.

We are interested in exploring this topic because online job search services and platforms are now considered equally important for employees and employers accessing a wide variety of opportunities across the country compared to in-person job postings. However, this can mean that applicants are overwhelmed by the amount of options they have access to, and are less likely to come across or apply to individual postings. This problem creates a demand for services like Linkedin to monetize priority for some job postings over others, where it claims employers will "get 3x the amount of qualified applicants when [they] add a budget and promote [their] job post."

Our primary research question is - what factors about job postings increase popularity among applicants? Our hypothesis is that job postings associated with companies who are more established, invest more into hiring, and offer more benefits for employees will have greater popularity. These factors can be characterized by having followers and employees, monetizing the posting, allowing for flexibility in the job (measured by whether remote work is allowed or not), adding more benefits - such as higher salary, health benefits, etc., which will positively correlate with both the job's view count and the number of applicants.

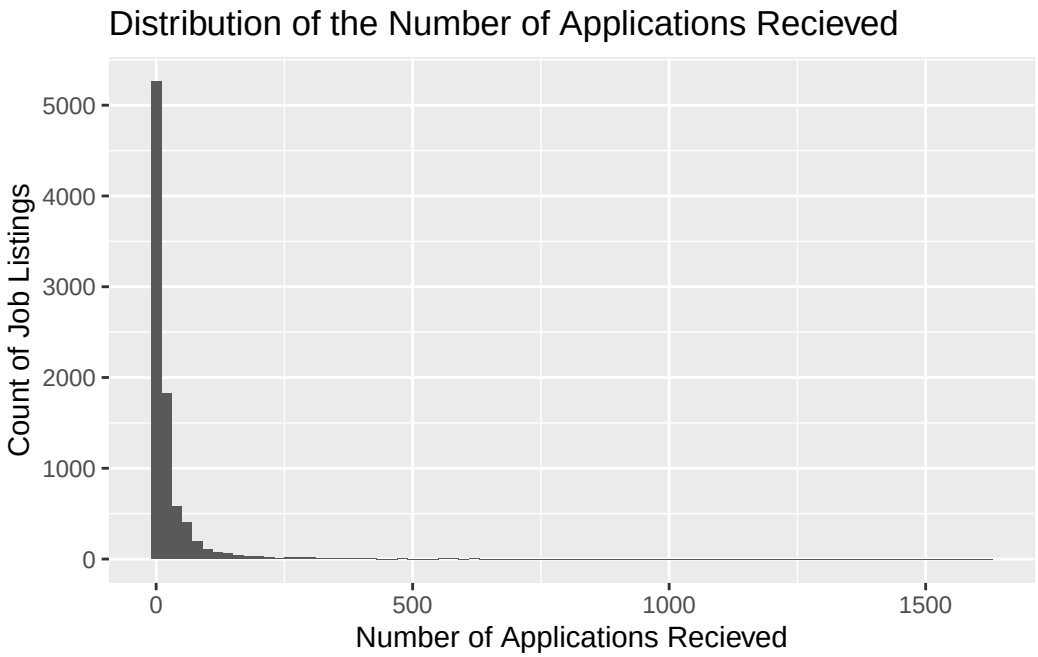
Data description

This data included in this set is sourced from linkedin.com, the website for LinkedIn. The creator of this data set, Arsh Koneru-Ansari, used Python to scrape data directly from linkedin.com and the scraper code is published in their GitHub (https://github.com/ArshKA/LinkedIn-Job-Scraper#jobs). The observations and general characteristics being measured through this data set are the number of applications for each job listing, the number of views for each job listing, the maximum salary rate, whether the job is remote or in-person, whether or not the listing is sponsored, the company's follower count on LinkedIn, the number of employees in the company, and the benefits associated with each job.

Initial exploratory data analysis

```
linkedin <- linkedin |>
  drop_na(applies)
```

```
linkedin |>
  ggplot(aes(x = applies)) +
  geom_histogram(binwidth = 20) +
  labs(x = "Number of Applications Recieved",
       y = "Count of Job Listings",
       title = "Distribution of the Number of Applications Recieved")
```



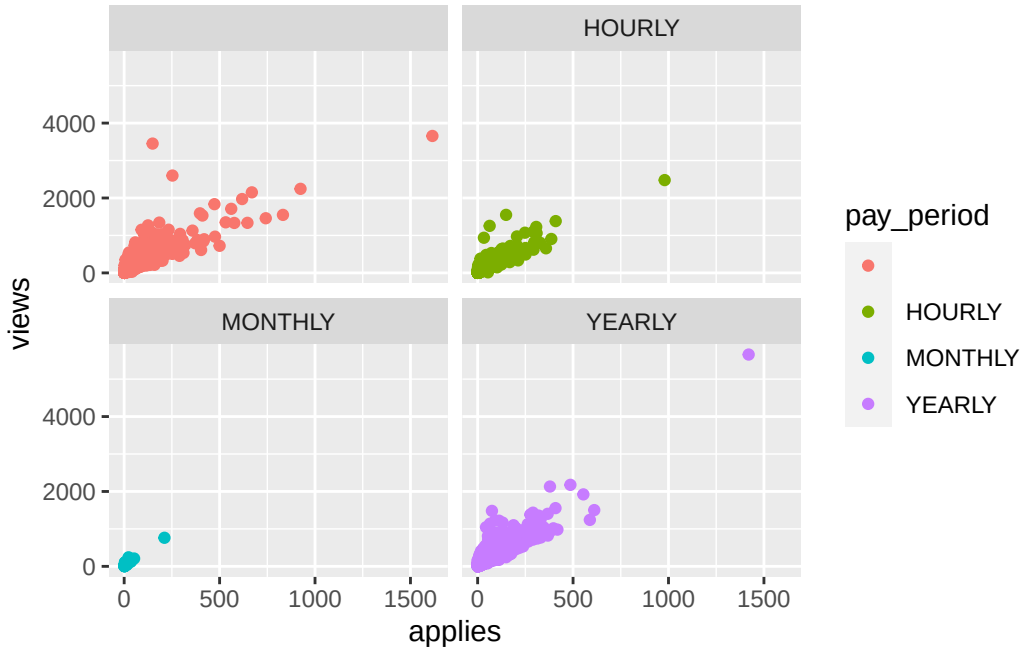

```
summary(linkedin$applies)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    6.00   22.83   21.00 1615.00
```

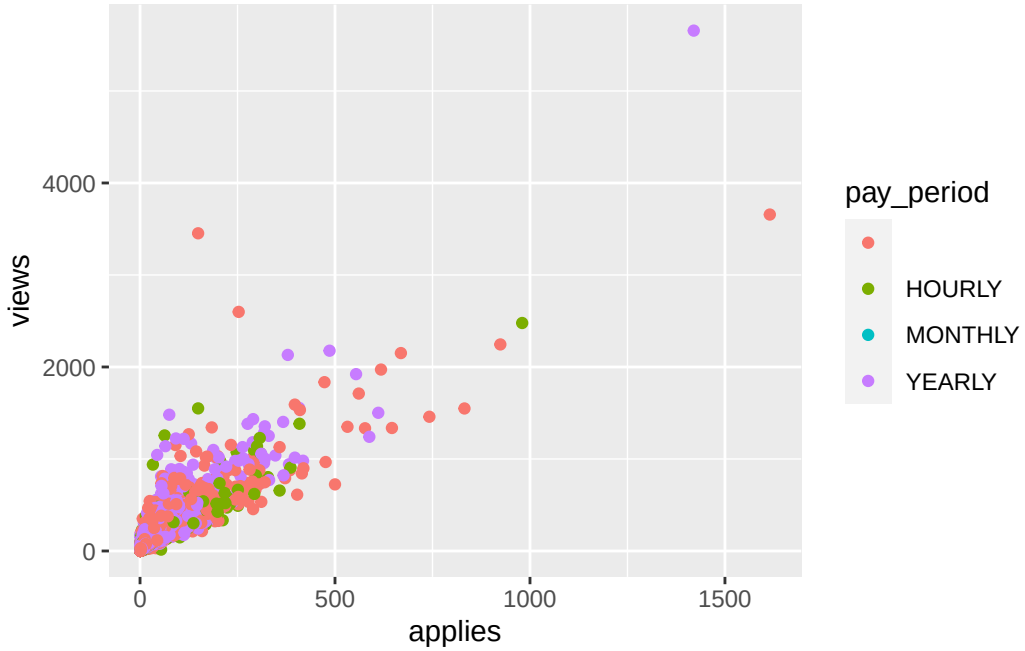
The distribution of the number of applications for a job listing on LinkedIn is right-skewed and uni-modal, with fewer applications for a job listing most prevalent. Given that the distribution is skewed, the center is 4.00 applications, as estimated by the median. The IQR describing the spread of the middle 50% of data is 12 applications (13 - 1).

There are 2 major outliers with a number of applications greater than 1,000. One popular position is the Junior Software Engineer job listing at Brooksource, with 1615 total listings. Another job listing is for Customer Success Manager at Noom with 1420 applications.

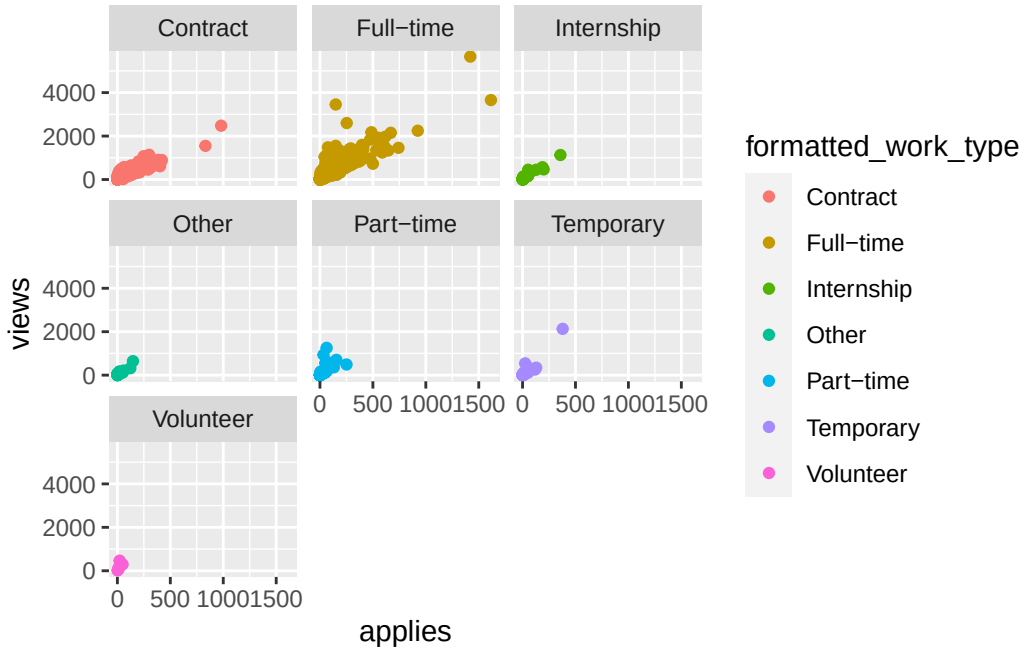
```
job_postings |>
  ggplot(aes(x = applies, y = views, color = pay_period)) +
  geom_point() +
  facet_wrap(~pay_period)
```




```
job_postings |>
  ggplot(aes(x = applies, y = views, color = pay_period)) +
  geom_point()
```

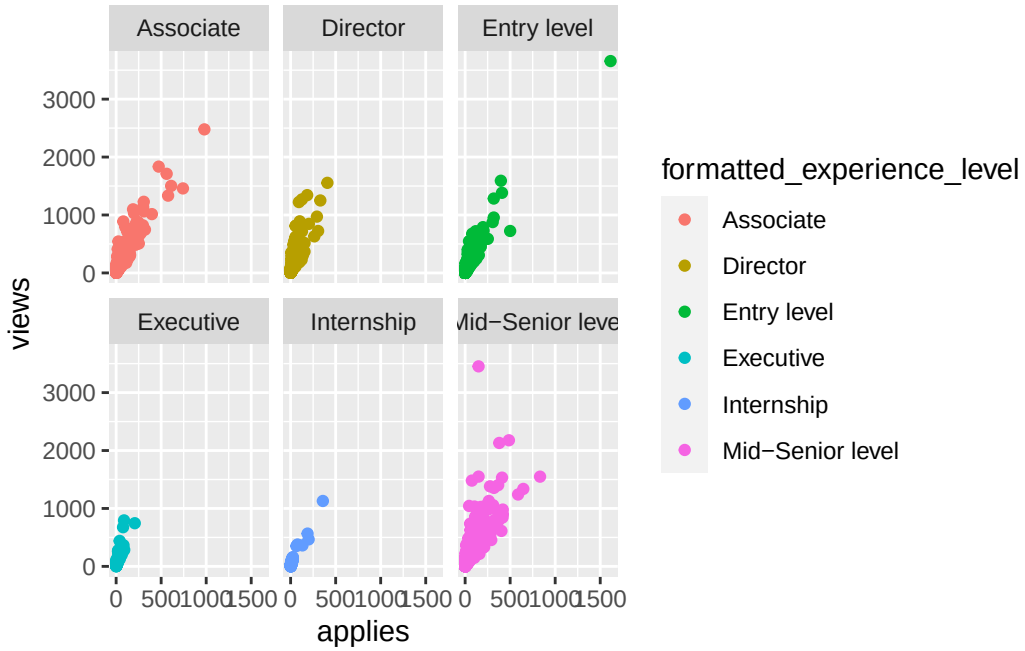



```
job_postings |>
  ggplot(aes(x = applies, y = views, color = formatted_work_type)) +
  geom_point() +
  facet_wrap(~formatted_work_type)
```

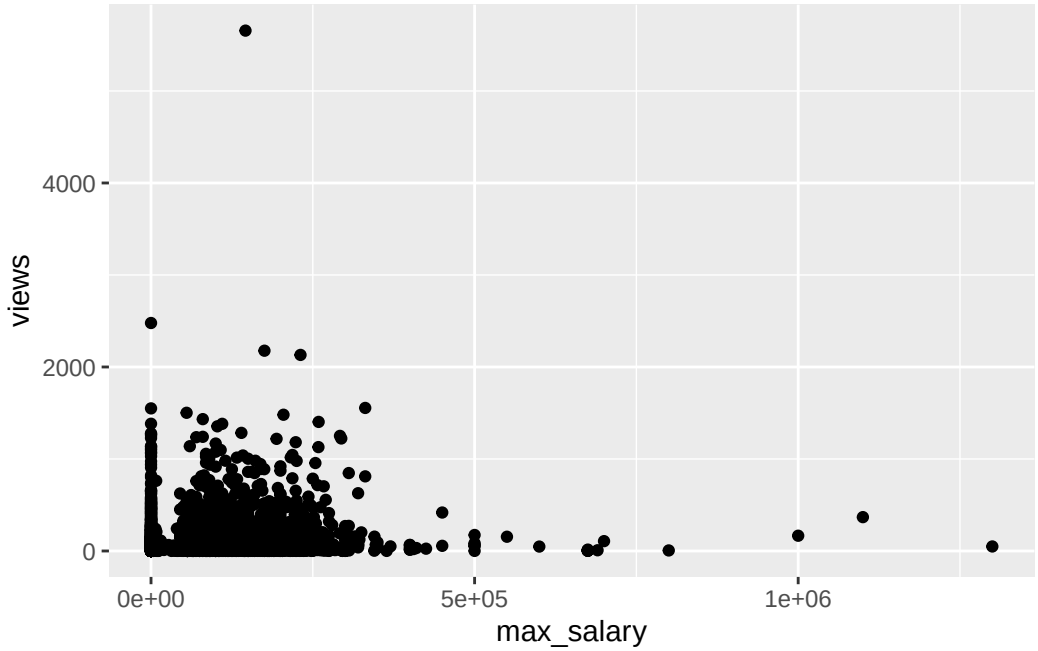



```
job_postings_exp_na <- job_postings |>
  drop_na(formatted_experience_level) |>
  filter(formatted_experience_level != "")
```

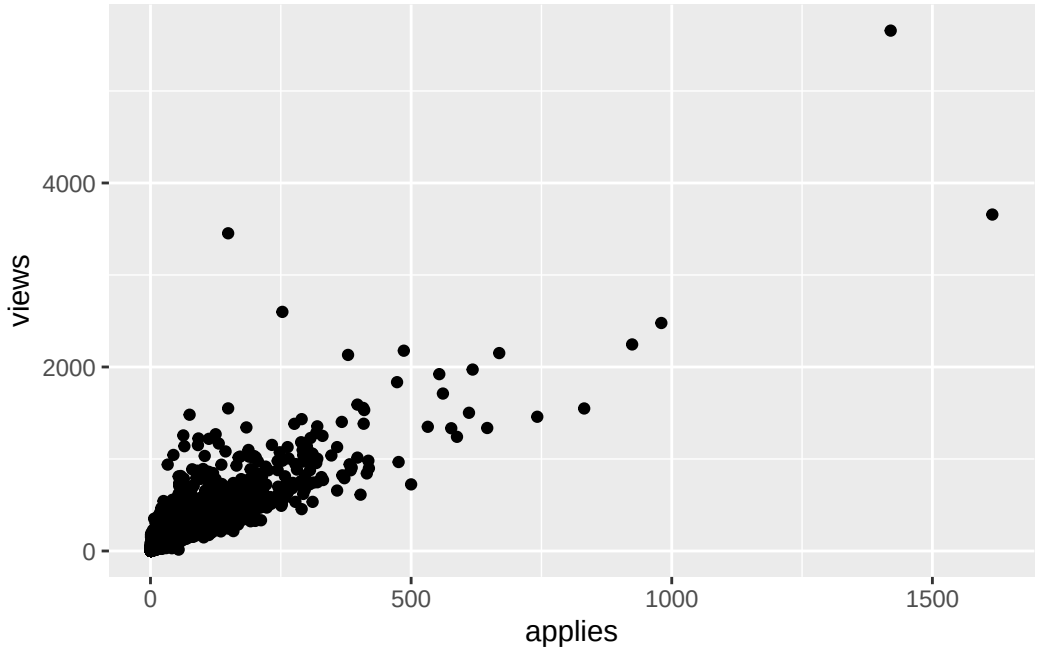
```
job_postings_exp_na |>
  ggplot(aes(x = applies, y = views, color = formatted_experience_level)) +
  geom_point() +
  facet_wrap(~formatted_experience_level)
```



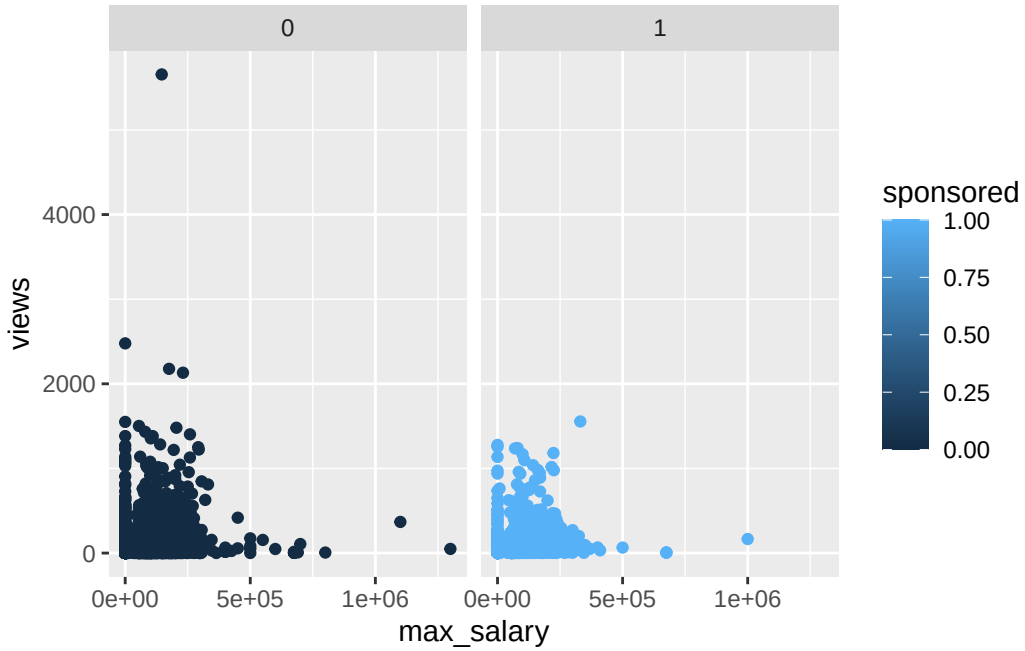

```
job_postings |>
  ggplot(aes(x = max_salary, y = views)) +
  geom_point()
```




```
job_postings |>
  ggplot(aes(x = applies, y = views)) +
  geom_point()
```




```
job_postings |>
  ggplot(aes(x = max_salary, y = views, color = sponsored)) +
  geom_point() +
  facet_wrap(~sponsored)
```




```
unique(job_postings$sponsored)
```

```
[1] 1 0
```

Analysis approach

A linear regression model is fitting for this prediction, as the number of applications for a listing is a numerical variable. The number of views the post received, the maximum salary of the position, the number of benefits provided by the job, the number of followers the company has, the number of employees the company had at the time of listing, if remote working is permitted, and the location of the job as metropolitan or rural could be used as potential useful predictors of the amount of applications for a job on a LinkedIn job listing. These variables include views, employee_count, remote_allowed, follower_count, max_salary multiplied by the period in pay_period to find annual pay, benefits manipulated to be a categorical variable, and location manipulated to be a categorical variable.

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]