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"))
```

Introduction

...

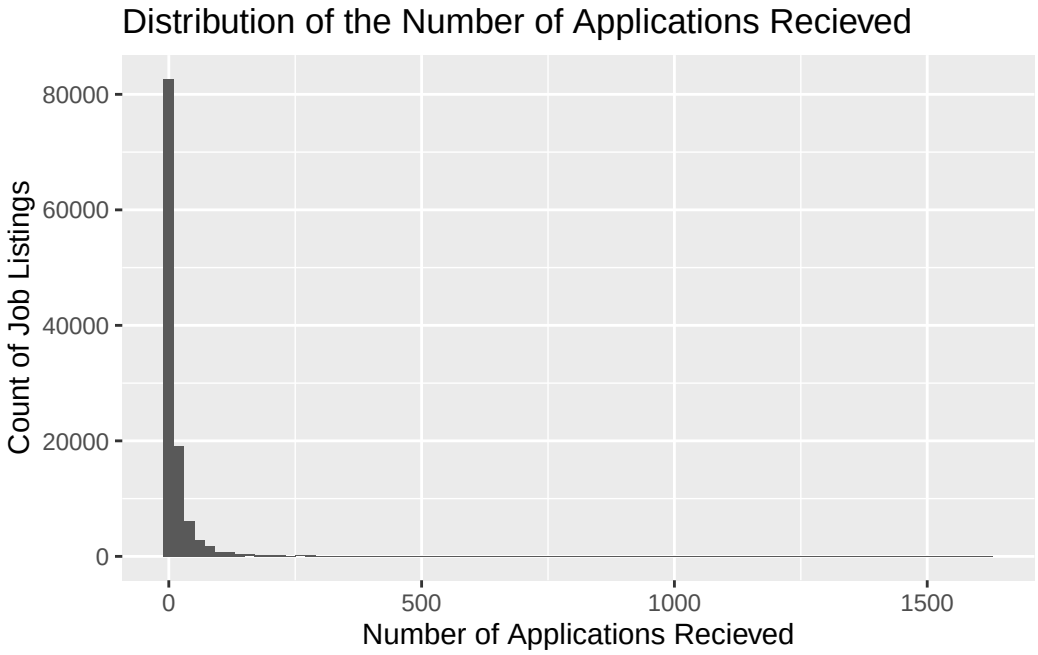
Data description

...

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    1.00    4.00   15.25   13.00 1615.00
```

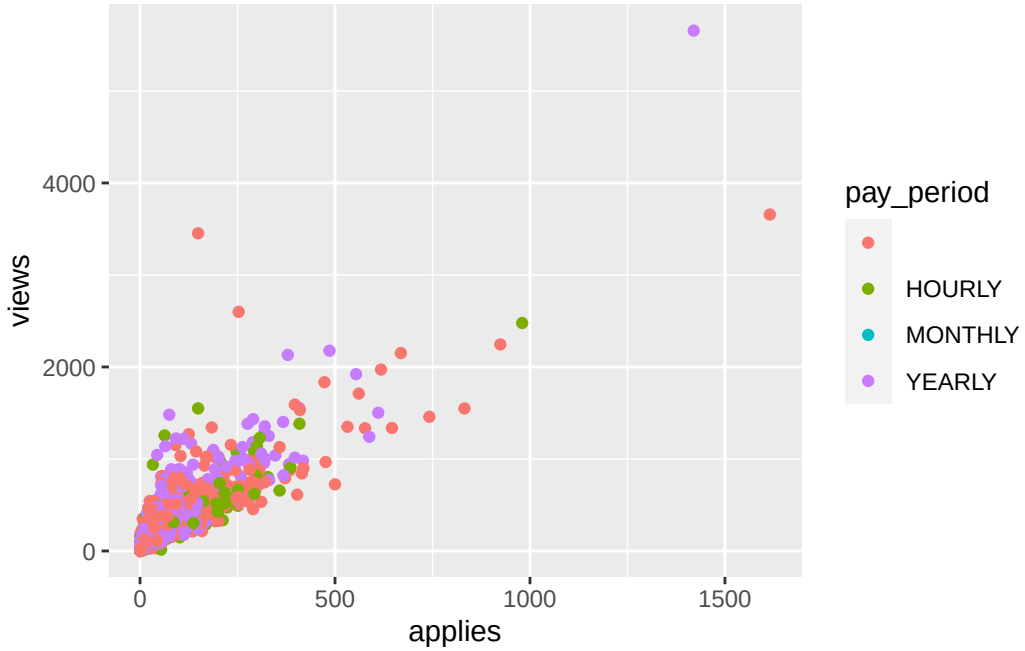
The distribution of the number of applications for a job listing on LinkedIn is right-skewed and uni-modal, with fewer applications for a job listing most prevalent. Given that the distribution

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).

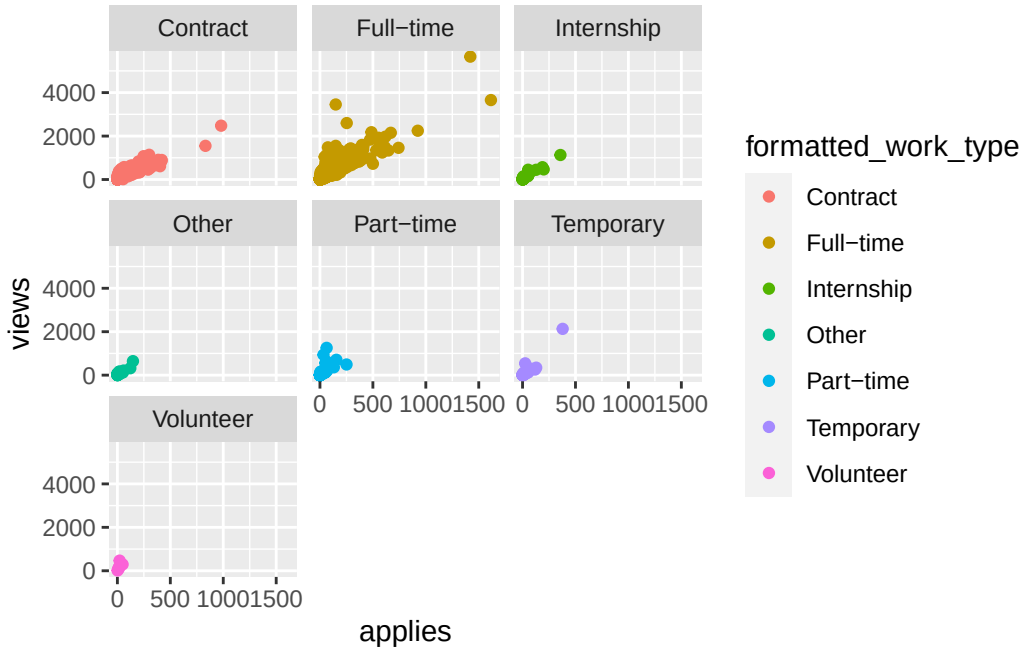
There are a couple of outliers with a high number of applications.

```
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()
```

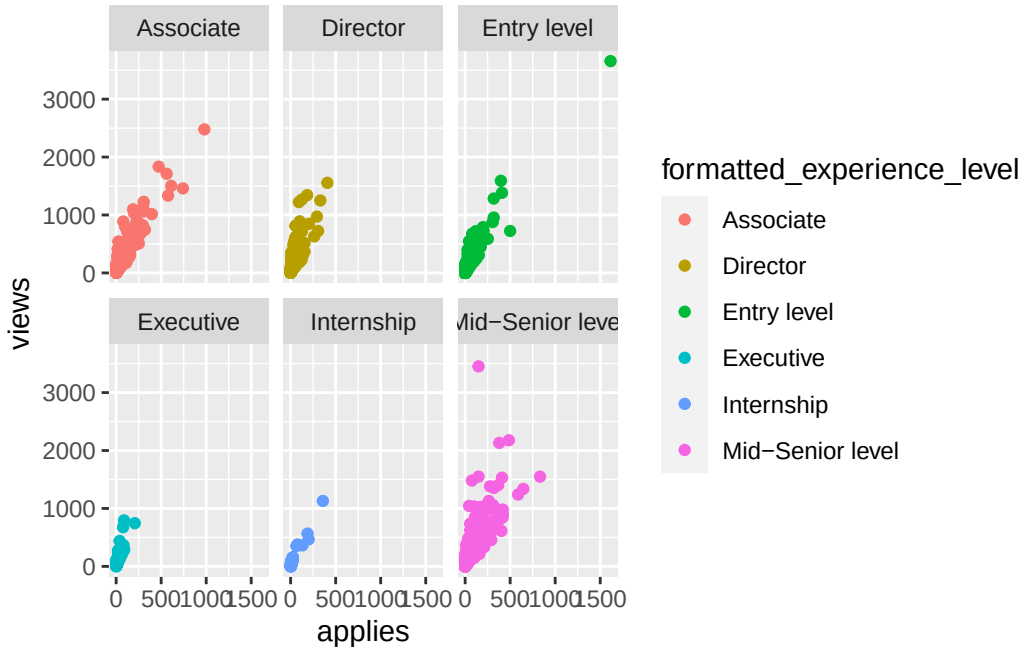



```
job_postings |>
  ggplot(aes(x = applies, y = views, color = formatted_work_type)) +
  geom_point() +
  facet_wrap(~formatted_work_type)
```

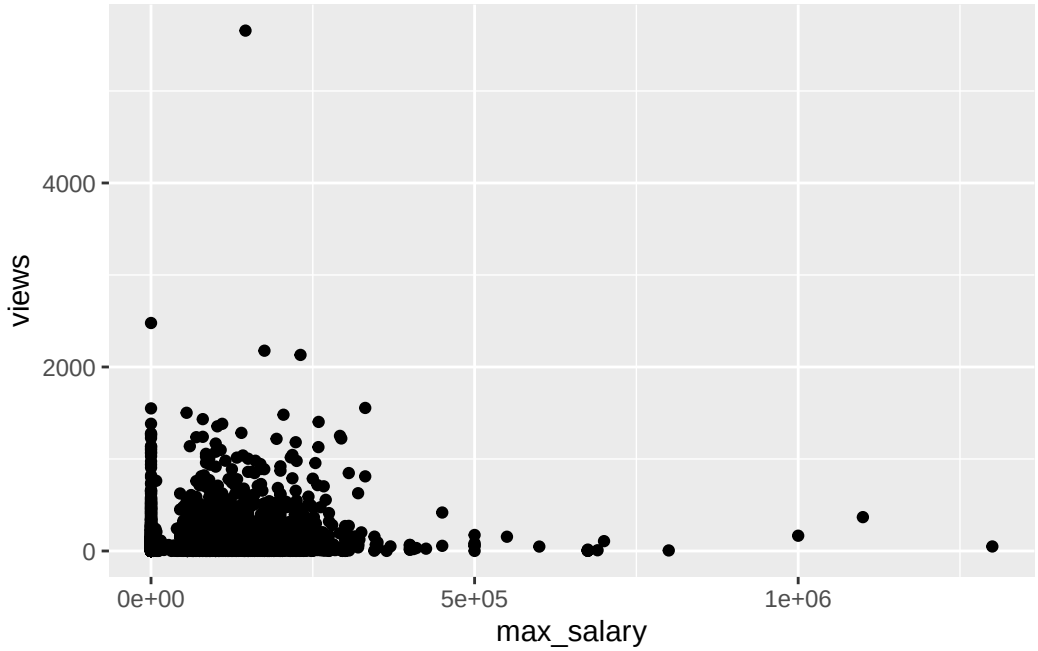



```
job_postings_exp_na <- job_postings |>
  drop_na(formatted_experience_level) |>
  filter(formatted_experience_level != "")
```

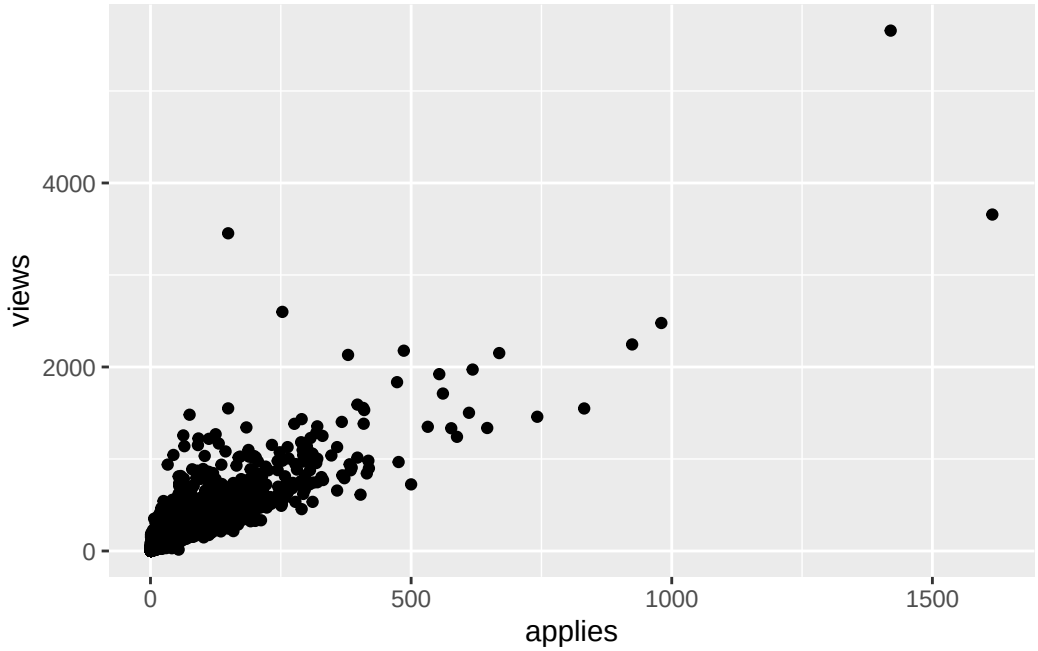
```
job_postings_exp_na |>
  ggplot(aes(x = applies, y = views, color = formatted_experience_level)) +
  geom_point() +
  facet_wrap(~formatted_experience_level)
```



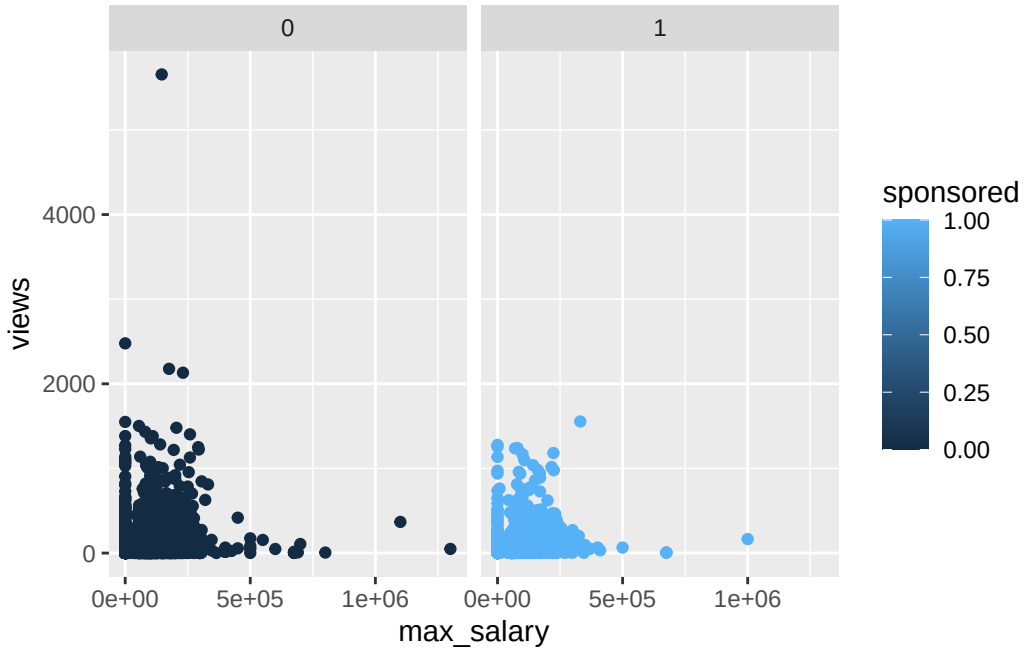

```
job_postings |>
  ggplot(aes(x = max_salary, y = views)) +
  geom_point()
```




```
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

...

Data dictionary

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