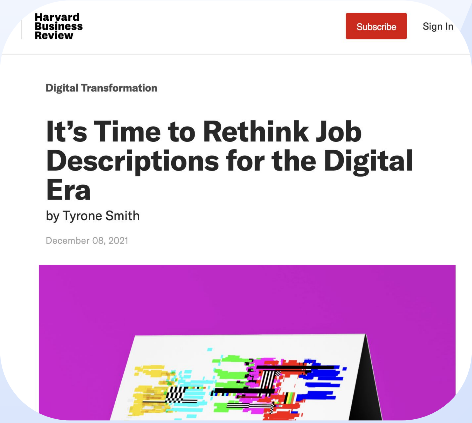
Linkedin Job Application Rates

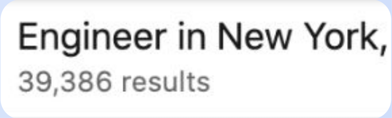
By: Laura Cai, Jessie Ringness, Karen Dong, Rebekah Kim

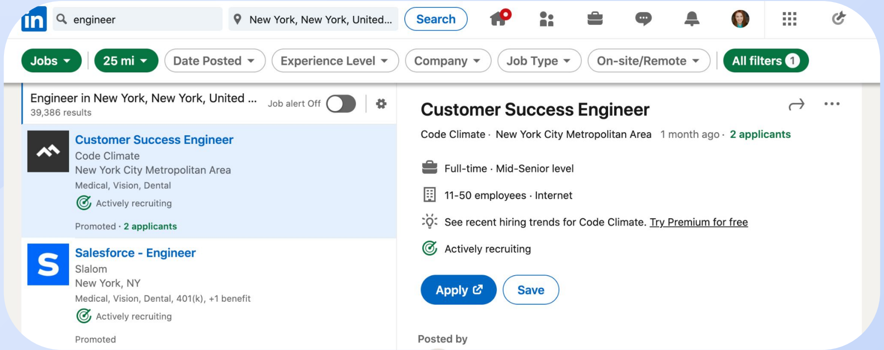
Why Linkedin Job Listings?

- **LinkedIn** connects companies and potential employees with thousands of active job listings

- What variables contribute to people **applying** to an online job posting **after viewing** it?
    - *Hourly salary*
    - *Company follower count*
    - *Remote work*
    - *Job experience level*
    - *Time posted*
    - *Benefits*







Data Introduction

Created by Arsh Koneru-Ansari in July 2023, who used Python to scrape data directly from **linkedin.com**

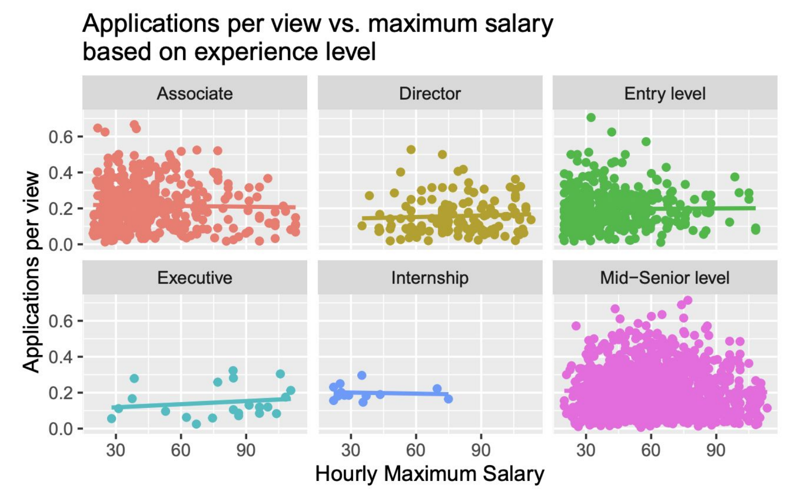
- Scraper code is published in their GitHub

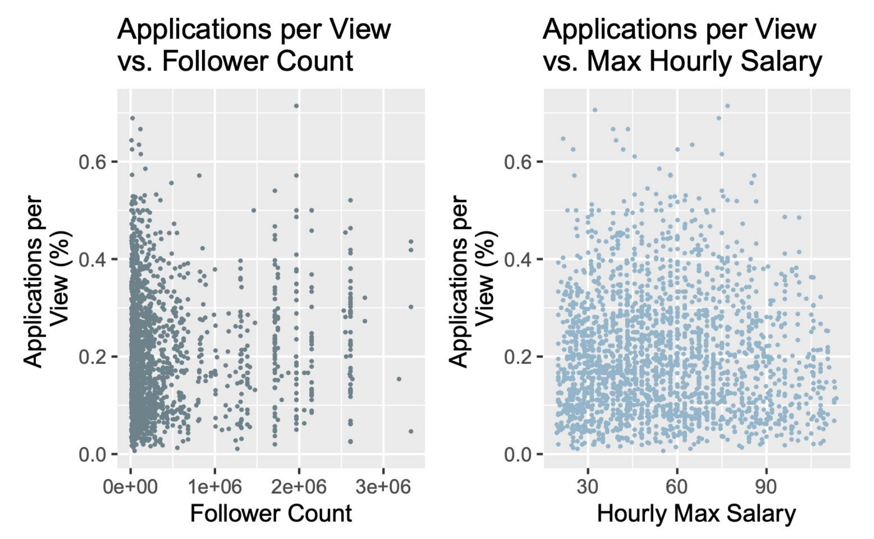
Variables we are observing:

- ***applies:*** # of applications
- ***views:*** # of views
- ***max_salary:*** max salary offered
- ***remote_allowed:*** (1 = yes)
- ***follower_count:*** # of company followers on LinkedIn
- ***formatted_experience_level:*** job experience level (entry level, associate, mid-senior level, etc.)
- ***original_listed_time:*** time and date when a listing was posted
- ***type:*** type of benefit provided (Medical insurance, 401(k), etc.)

- ***per_applies:*** % of viewers who apply
  - *applies* divided by *views*
- **hourly_max_salary:** maximum salary in hourly rate
  - Standardized *max_salary*
- **listed_time:** categorical variable with 4 levels about the time of day posted
  - *night* (0 am - 5 am), *morning* (6 am - 11 am), *afternoon* (12 pm - 5 pm), *evening* (6 pm - 11 pm)
- ***if_benefit:*** are benefits listed? (1 = yes)

EDA





- *no apparent direction or shape between the hourly maximum salary and applications per view for each of the experience levels*
  - *Correlation: -0.033 (moderately weak relationship)*
  - *concentrated to have less than $200k for adjusted hourly max salary*

- *slight linear correlation between follower count of a company and percentage of viewers who apply to the job*
  - *concentrated to have less than 1M followers*
- *slight positive linear correlation between a job's adjusted hourly maximum salary and percentage of viewers who apply to the job*
  - *concentrated to have less than $200k for adjusted hourly max salary*

Methodology

A **multiple linear regression model** is most appropriate, given that our response is numerical and we have more than one potential predictor variable

1) Manipulate data

2) Split data

- 75% of data in training, 25% in testing sets

3) Create recipes and workflows for models

Model 1: hourly max salary, follower count, remote allowed, formatted experience level, if benefits, time posted

- Full model containing all proposed variables

Model 2: remote allowed, formatted experience level, follower count

- Contains only statistically significant variables from the full mode, using a significance level of alpha = 0.05

Model 3: formatted experience level, follower count, remote allowed, hourly max salary

- Includes interaction effect between forformatted experience level and hourly max salary

4) Compare model fit stats using cross validation

Model 1:

mean_adj_rsq	mean_aic	mean_bic
0.023	-2095.867	-2021.562

Model 2:

mean_adj_rsq	mean_aic	mean_bic
0.024	-2101.431	-2053.664

Model 3:

mean_adj_rsq	mean_aic	mean_bic
0.022	-2092.22	-2012.607

5) Select model

We selected **model 2** as out final model as it is more parsimonious, has a higher adjusted R squared, and lower AIC and BIC

6) Evaluate model

- Tests for multicollinearity and overfit
- Tested conditions for inference

Final Model + Interesting Findings

$$per\_applies = 0.209 + 0.010(follower\_count(millions)) + 0.033(remote\_allowed)$$
$$-0.054(Director) - 0.016(Entry\_level) - 0.068(Executive) - 0.003(Internship) - 0.009(Mid\_Senior\_level)$$

Variables

Predictors:

- Follower count (millions)
- Remote allowed
- Formatted experience level

Response:

- Percent of viewers who applied

Results

No overfit:

- RMSE: **0.1189** (training) vs 0.1193 (testing)

No multicollinearity:

- VIF < 10

The **R-squared** value of **0.0280** suggests there is no significant relationship between our response variable and predictor variables.

Context

Our model suggests that the predictors we analyzed don't provide much insight into what makes a LinkedIn post have significant popularity, which means that job listing popularity is difficult to quantify.

R Squared	Adj. R Squared	AIC	BIC
0.028	0.024	-2293.407	-2244.857

Conclusions + Future Work

No Significant Statistical Relationship

between **% of viewers who applied** and whether it's **remote**, its **experience level,** its **hour posted**, if **benefits** were listed

Limitations

- **Assumption errors** about annual max salary
- **Filtering** and **threshold** for **outliers**
- **Lack of data** about how long jobs were listed

Future Work

- Improve methods for **filtering outliers** and deciding **thresholds** for views, follower count, salaries, etc.
- Improve data scraping consistency and discern **original listed time** from scraped time