DATA JOB ANALYSIS & MODELING

Anuja Panthari & Sai Velicheti

PROJECT BACKGROUND & MOTIVATION

As students enter the workforce in the next year, it is important to know the landscape of data-related jobs in the market, as job titles are fluid and technologies are ever-changing. Our initial motivation for the project was to understand what skills are most needed for data-related jobs (e.g. relevant programming languages, applications, database management skills, certifications, etc.). Some further project motivations include:

PROJECTED JOB GROWTH

Demand for data scientists is skyrocketing, with a projected 35% jump in job openings between 2022 and 2032

CURRENT MARKET FRENZY

Job openings are growing at a sluggish pace (7%) compared to the recent surge in applications (31%)

REMOTE WORK RESHAPING HIRING TRENDS

The pandemic-era shift to remote work has solidified as a norm, with over 60% of data science roles now offering flexible or fully remote options.

DATA DESCRIPTION

DATA SOURCE: JOB FUNNEL, scraped using a YAML file.

JobFunnel

Key Relevant Features (after cleaning):

- **JOB TITLE**
- **COMPANY**
- **BLURB**
  - The first couple of sentences could be scraped from the job description without being stopped by AI bot detector
- **WAGE**
  - converted into annual salary
- **REMOTENESS**
  - Full or Not Remote
- **STATE**
  - CA, TX, NY, NC

PROBLEM STATEMENT

We are interested in acquiring the job level metric for each job. Our interest in this feature comes from our motivation to get a picture of how we can segment the jobs by some metric. We chose job level because this metric is relevant to our motivation of getting a better idea of what types of jobs are in the market relevant to students.

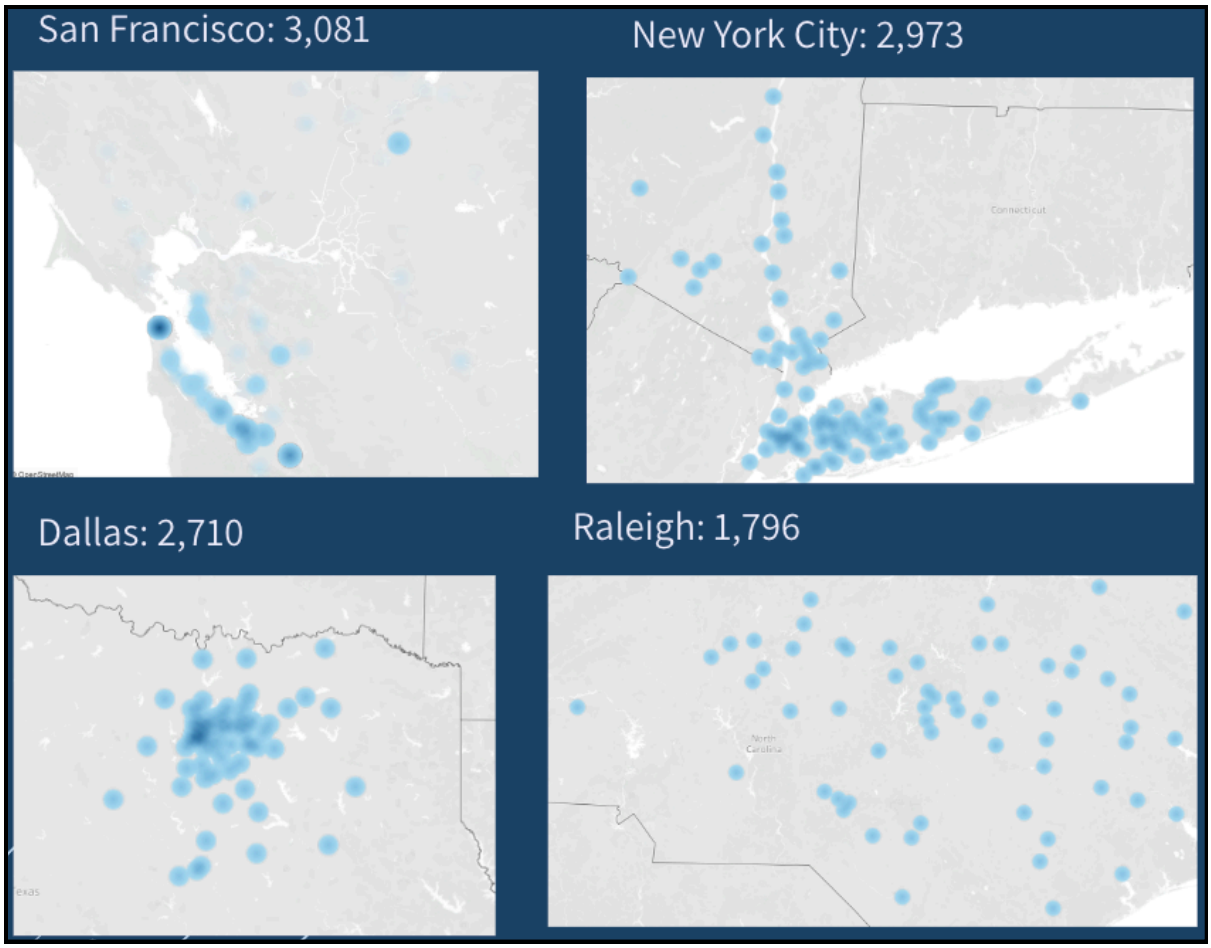
HIGH LEVEL ANALYTICS

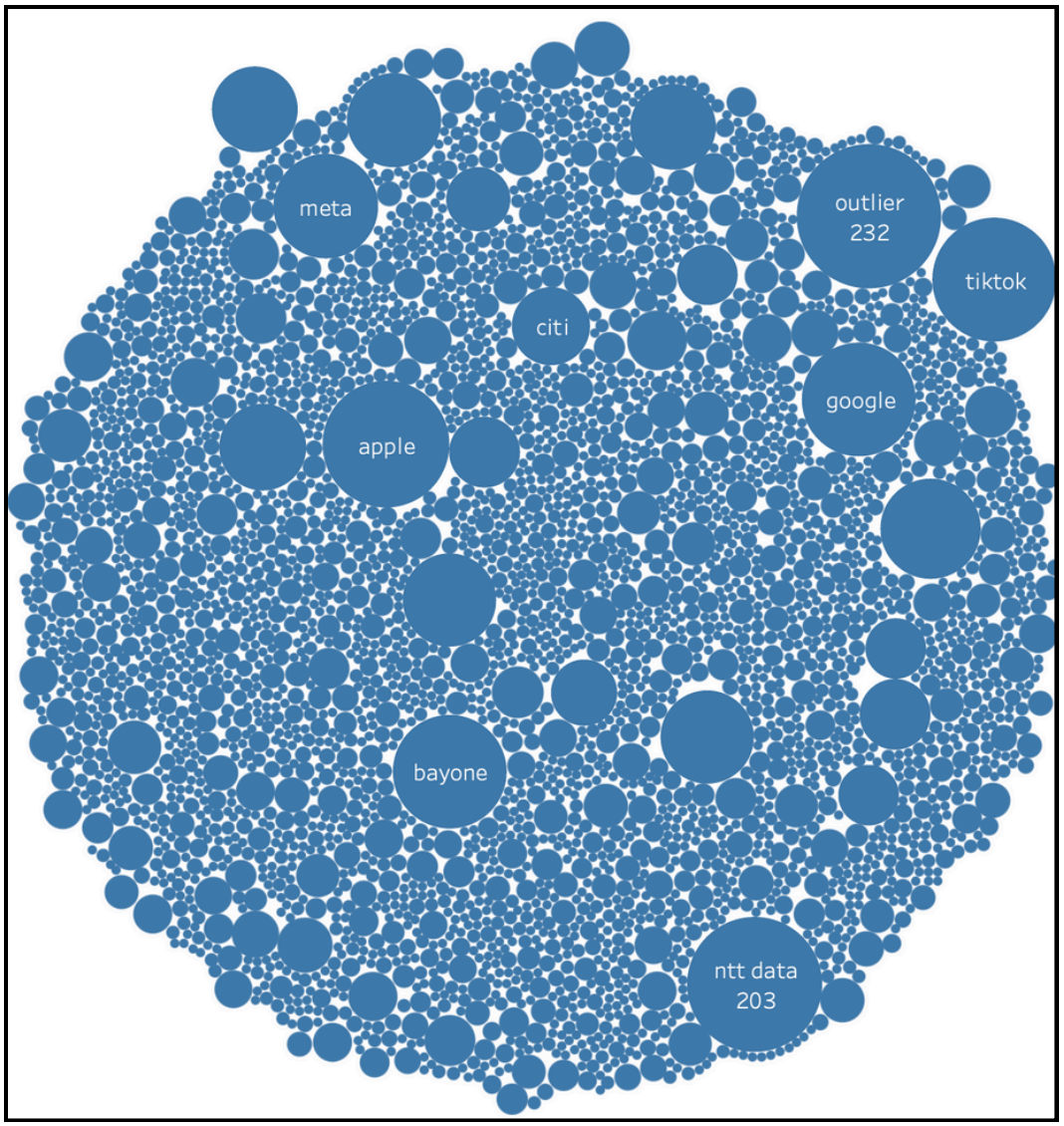
TOTAL JOBS: 10,604

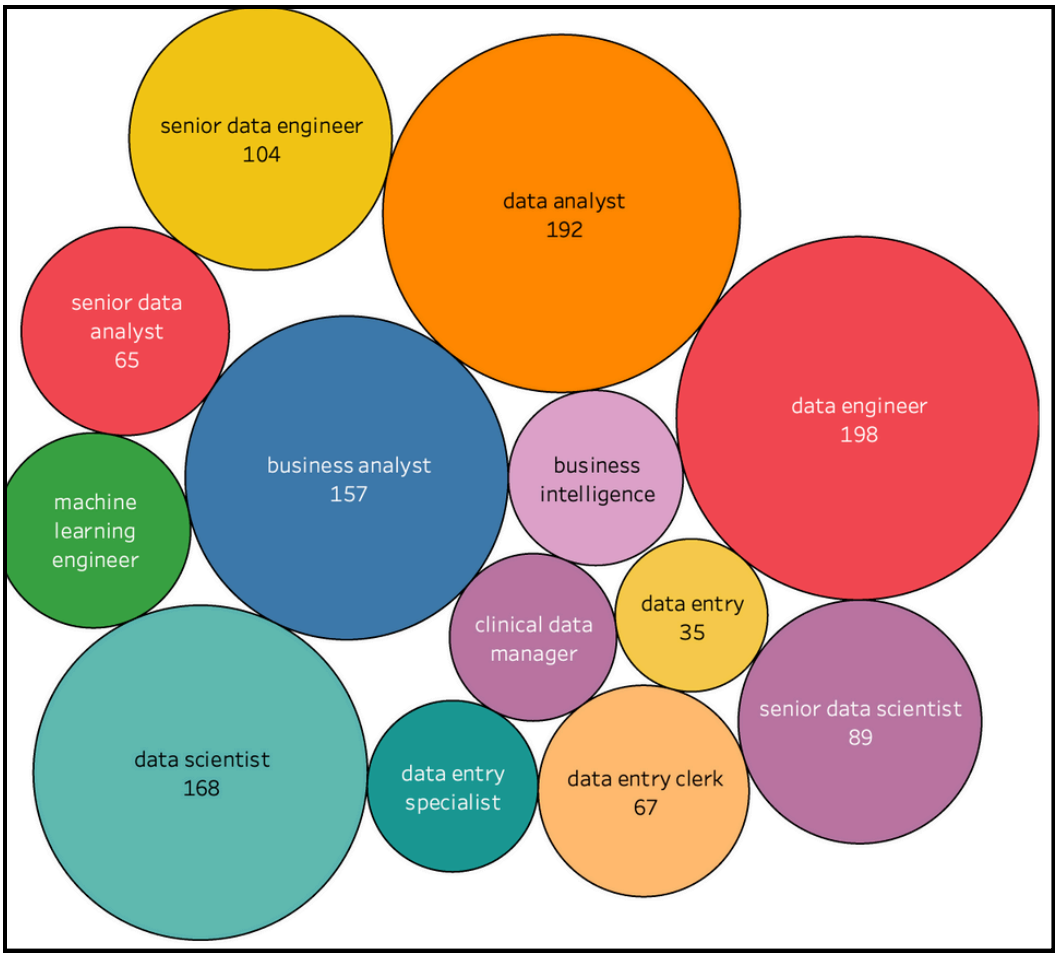
Time frame: The query was limited to postings within the last month (11/2 to 11/27), ensuring the data reflects the most recent job trends and labor market conditions.

Keyword filter: The query centered around the keyword “data," ensuring relevance to roles involving data-related responsibilities, such as data analysis, engineering, or science.

Geographical scope: The search was constrained to four major cities, focusing on prominent metropolitan areas to capture roles in diverse but economically significant regions. Additionally, an 80 km radius around each city was considered to include surrounding suburbs and satellite towns.


San Francisco: 3,081
New York City: 2,973
Connecticut
Dallas: 2,710
Raleigh: 1,796
North Carolina
Texas

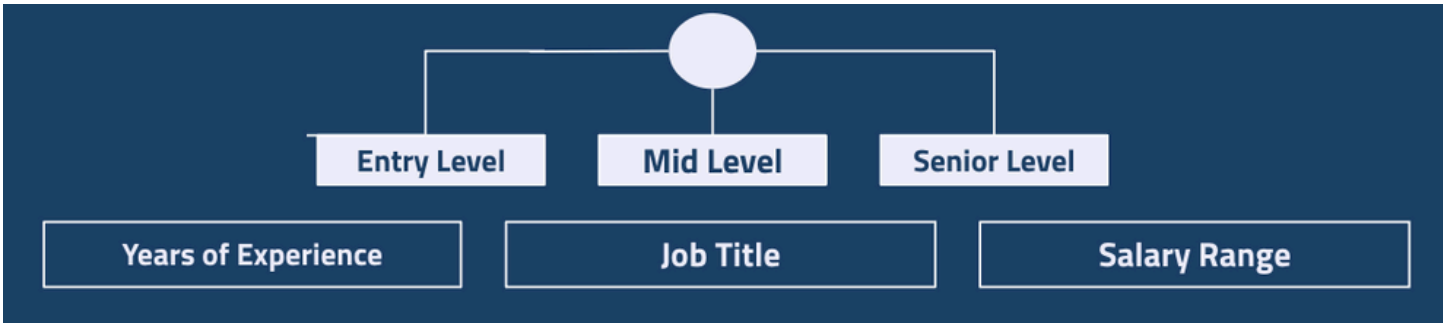

meta
outlier
232
tiktok
citi
google
apple
bayone
ntt data
203



METHODS

METHOD 1: SUPERVISED LEARNING

We are interested in acquiring job-level features in the dataset, but we don't have them! Can we predict job levels using the dataset metrics? We first asked ChatGPT to segment a section of the dataset to be our training set. The model used keywords to segment jobs into 3 levels (entry, mid, and senior). We wanted to do a quality control measure for this metric and compare it to a "gold standard." This idea is similar to retrieval augmented generation (RAG model), used to check the efficacy of Gen AI models like ChatGPT. We manually labeled 600 data points by clicking through each link and checking the years of experience needed. We also used job titles and salary ranges, but this got dicey, as explained in the next section. Once they were labeled, we could check how close the ChatGPT prediction of job level was to our "gold standard."



We trained the ChatGPT model on a multinomial logistic regression after vectorizing on job title and blurb. We chose a simple regression model because we didn't want to overcomplicate the

metrics used - after all, segmenting (if it is straight forward) should be an easy task. We used a chi-square test to select which n-grams were the most relevant for our model and selected the top 6,000 values. Sadly, our model had a very low accuracy (35%) when checked on our manually labeled test set. What does this mean? It means that ChatGPT's way of labeling does not align with our method of labeling. Why? The metrics used to define seniority differed between training & test sets. So maybe defining job level isn't that easy after all! More on this in the next section.

```
▾ LogisticRegression
LogisticRegression(multi_class='multinomial')
```

```
logistic regression, accuracy on test set: 0.34223706176961605
```

DEFINING JOB LEVEL

These results lead to a larger discussion about how to define job level. Should it be by years of experience? Company? Title? Salary? A combination? The main issue is that there is no clear way of defining job level that all companies abide by. Here is a thought experiment:

How would you segment these jobs?

- "Senior Software Engineer" requires 6 years of experience
- "Database Manager" has salary of $75,000 and, requires 15+ years of experience
- Amazon "Data Scientist" in CA that has a $250,000 salary requiring 2-3 years of experience

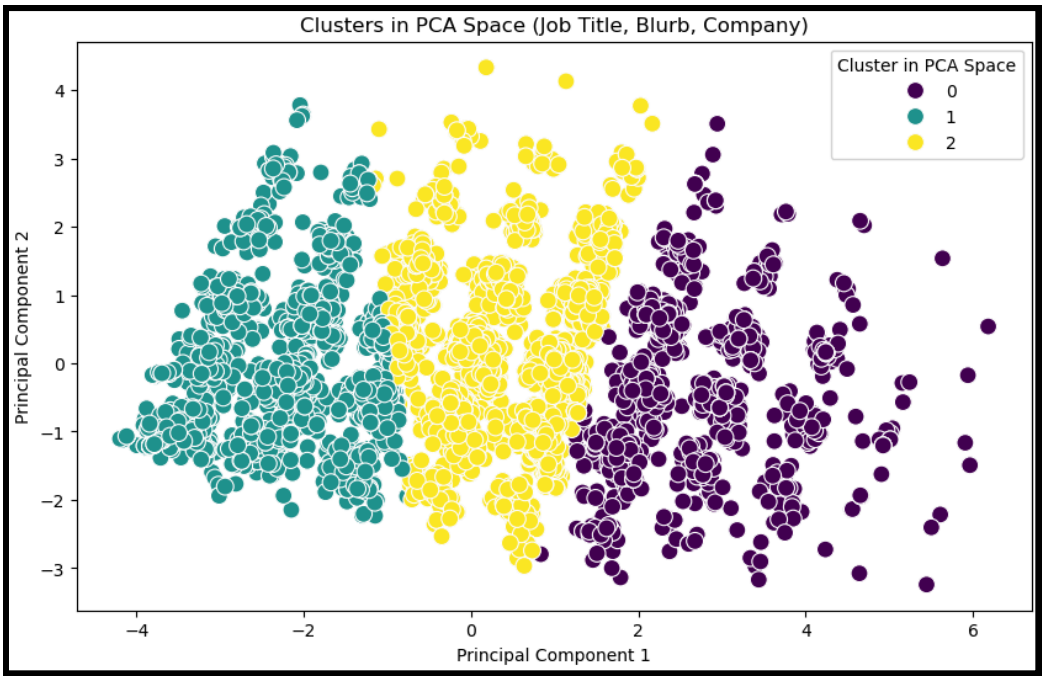
The answer isn't very clear, and each of these jobs could arguably be labeled differently depending on your definition. Since "years of experience" was how we decided to manually label out the dataset, and that metric was not in our original dataset, the training and test sets would never align! That is, unless years of experience are mapped to keywords, which we now know it doesn't because of how astonishingly low our model accuracy was.

METHOD 2: UNSUPERVISED LEARNING

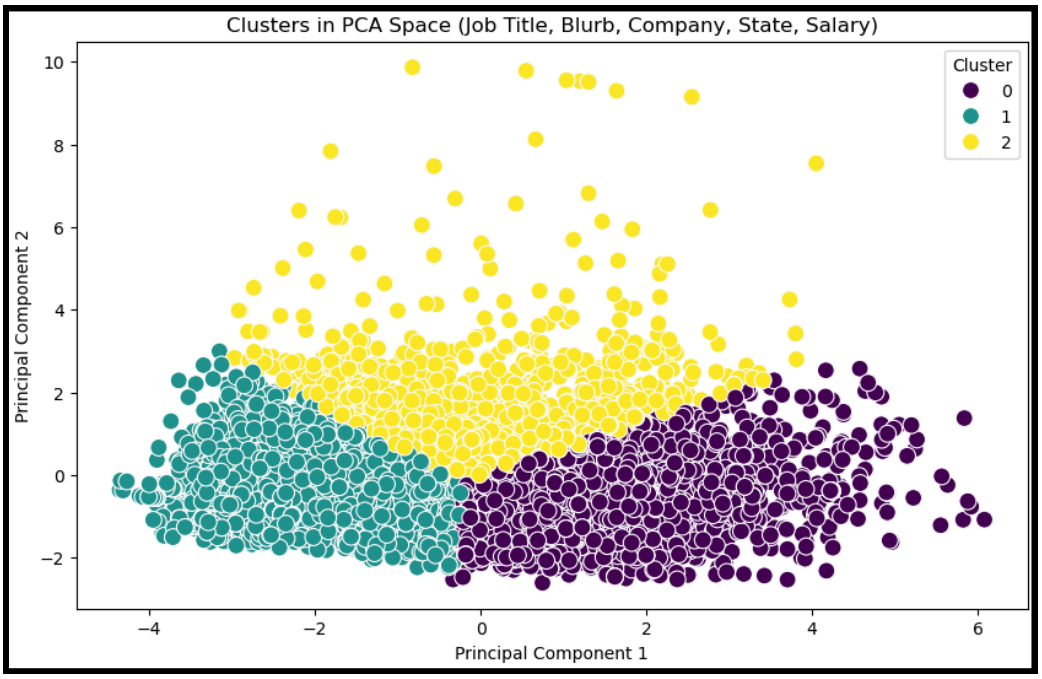
So, if segmenting by job level is proving to be an impossible task, can we generally check if there are any clear patterns found in the data? We vectorized the data to convert textual features such as job titles, blurbs, and company descriptions into a machine-readable format. A chi-square test was then used to identify the most relevant features within these fields, ensuring that the analysis focused on the variables that have the most significant impact. We encoded the 4 states into dummy variables and used PCA to reduce the data into a sparse matrix. We plotted the PCA components below and clustered using K-Means.

Cluster 1: Focused on textual features, including job title and blurb. We added in company, but this did not add or detract information. So, below we have shown the PCA components of just the latter 1 variables. This is the most interesting clustering map because there seems to be a linear

pattern from up to down and then also clusters within each linear bunch. The data is evenly spread across the x-axis. Very low silhouette accuracy score.



Cluster 2: Included both textual features and additional variables such as state and salary data. This dual approach helped with the identification of distinct patterns, highlighting clusters based on the interplay of job descriptions, company attributes, geographical factors, and compensation levels. Comparatively, this PCA plot didn't have clear patterns. Again, very low silhouette accuracy score.



KEY OBSERVATIONS & FINDINGS

- Defining job levels proved to be a complex task due to the lack of uniformity in how roles are described across companies and various industries. Data standardization is needed to build a more effective model that can then more accurately interpret and group rules based on those criteria, leading to improved uniform patterned clustering and analysis outcomes.
- There was an interesting pattern that was found in PCA values:

- When using the title, blurb, clear and distinct patterns were observed in the PCA plot
- However, these patterns were changed to a more abstract form when state and salary were added.

CONCLUSION & FUTURE DIRECTION

- **Exploring clusters:** Observing how titles, blurbs, and company data group together and identifying any functional consistencies.
- **Impact of State and Salary:** Assessing why these variables disrupt the patterns and exploring whether different encoding methods or weighing schemes could change it.
- **Cross-validating results:** Comparing the PCA patterns with other clustering techniques to validate whether the observed trends are stable.

INDIVIDUAL CONTRIBUTIONS

Anuja: data querying, data cleaning, data modeling

Sai: data labeling, data visualization