Navigating NYC's Job Market:
Insights from Exploratory Data Analysis to Predictive Modeling

Team Members

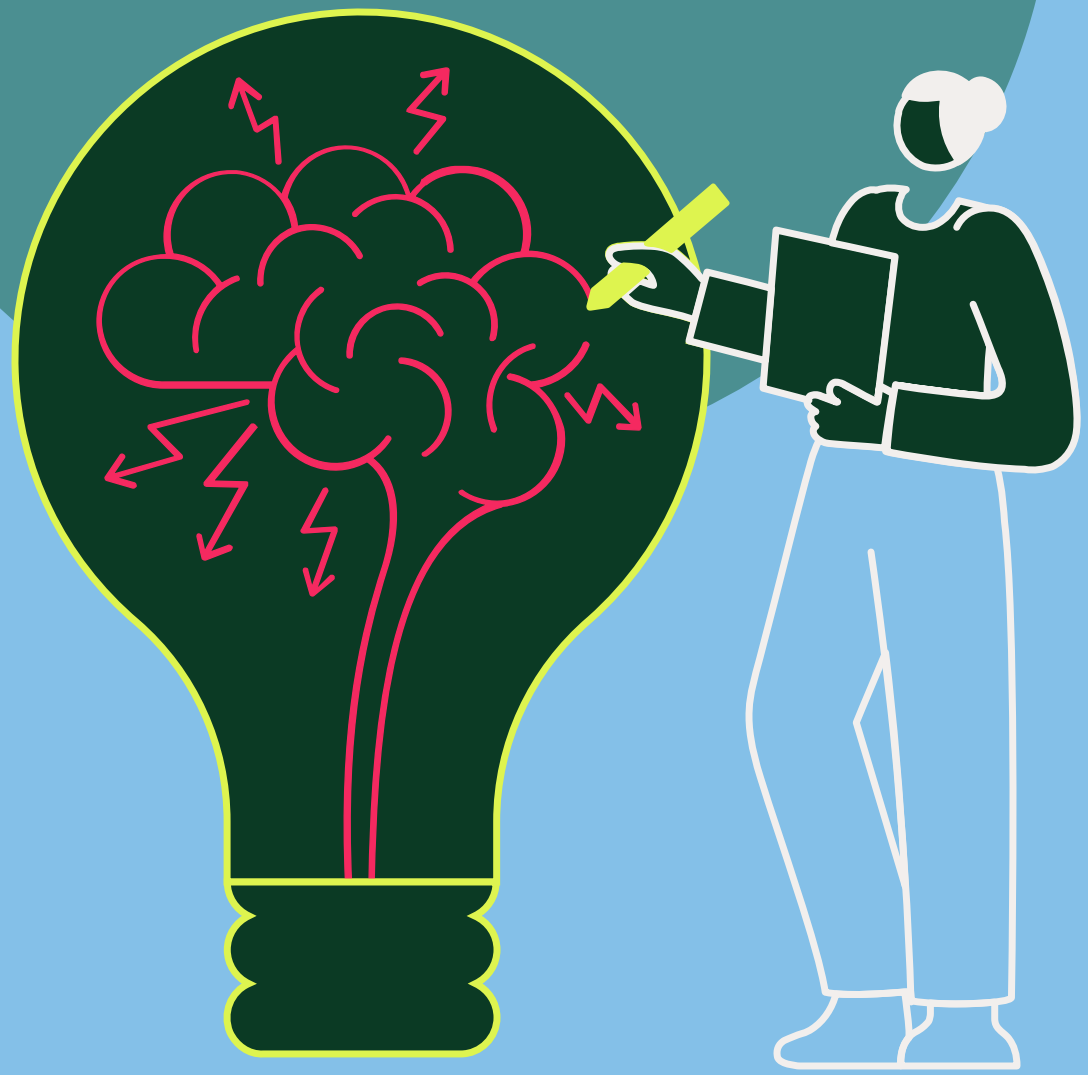
- Arib Mirza
- Kunal Sanghvi
- Prachi Holkar
- Priyam Sheth
- Sejal Arora
- Shreeya Desai

Agenda

- Introduction
- Exploratory Data Analysis
- Prediction Model
- Conclusion

INTRODUCTION

- This project uses the Jobs NYC Postings dataset available on Data.gov.
- The dataset contains current job postings available on the City of New York's official jobs site. Internal postings available to city employees and external postings available to the public are included.
- There are nine numerical variables and twenty-one categorical variables in this dataset.
- We aim to analyze this comprehensive dataset and gain insights into the factors influencing the salary ranges in NYC.



DATA CLEANING

Load Data

Load: Created a new Data Frame with specific columns relevant to the analysis.

Removed Columns like jobID, Title code no, Work Location, Division/ work unit, additional information, To Apply, Hours, recruitment contact, post until, posting updated, processed date

Usual Drill

Summarization: summarizing the main characteristics of the dataset, such as its size, dimensions, and data types.

Descriptive Statistics: Describe the central tendency, dispersion, and shape of the data distribution.

Handling Missing Values

Dropped rows with missing values in the 'Full-Time/Part-Time indicator' and 'Minimum Qual Requirements' columns.

Replaced blanks in the 'Work Location 1' and 'Preferred Skills' columns with 'NA'.

Mapping Attributes

Defined a function to map the attribute 'residency requirements' to broader categories and stored it in a new attribute 'Residency'.

Defined a function to map original 'job categories' to broader categories and created a new attribute 'Domain'.

Non-English characters

Identified and removed non-English characters from the attributes 'Business Title', 'Job Description', 'Preferred Skills', 'Work Location' and 'Residency'.

Adding Custom Attributes

Created a new column 'Avg Salary' as the average of 'Salary Range From' and 'Salary Range To'.

DATA CLEANING

Before

Agency	Posting Type	# Of Positions	Business Title	Civil Service Title	Title Classification	Level	Job Category	Full-Time/Part-Time indicator	Career Level	Salary Range From	Salary Range To	Salary Frequency	Job Description	Minimum Qual Requirements	Preferred Skills	Work Location 1	Residency Requirement	Posting Date
DEPT OF ALTH/MENTAL HYGIENE	External	1	Senior Director Healthcare and Community Readi...	AGENCY ATTORNEY	Non-Competitive-5	4	Constituent Services & Community Programs Heal...	F	Experienced (non-manager)	90870.0	135000.0	Annual	Established in 1805, the New York City Departm...	1. Admission to the New York State Bar; and ei...	At least 8 years combined experience in public...	NaN	New York City Residency is not required for th...	12/8/2023
OFFICE OF CRIMINAL JUSTICE	External	1	Information Technology Infrastructure Manager	RESEARCH PROJECTS COOR(MA)-MGR	Pending Classification-2	M1	Technology, Data & Innovation	F	Experienced (non-manager)	58700.0	115000.0	Annual	The Mayorâ□□s Office of Criminal Justice (MOCJ...	1. Do you have a baccalaureate degree from an ...	â□¢ \tExcellent technical knowledge of Network ...	NaN	New York City residency is generally required ...	2/1/2024

```
['No Residency Requirement' 'NYC Residency Required' 'NYCHA Residency'
 'NY State Residency Required' 'Other Residency Requirement']
```

```
['Community Programs' 'Technology and Data' 'Social Services'
 'Finance and Procurement' 'Building Operations'
 'Administration and Human Resources' 'Policy, Research & Analysis'
 'Engineering and Construction' 'Legal Affairs'
 'Public Safety and Enforcement' 'Intergovernmental Affairs' 'Health'
 'Multi Category']
```

After

```
new_df
```

# Of tions	Business Title	Civil Service Title	Title Classification	Level	Job Category	Full-Time/Part-Time indicator	Career Level	...	Salary Frequency	Job Description	Minimum Qual Requirements	Preferred Skills	Work Location 1	Residency Requirement	Posting Date	Residency	Domain	Avg Salary
1	Senior Director Healthcare and Community Readi...	AGENCY ATTORNEY	Non-Competitive-5	4	Constituent Services & Community Programs Heal...	F	Experienced (non-manager)	...	Annual	Established in 1805, the New York City Departm...	1. Admission to the New York State Bar; and ei...	At least 8 years combined experience in public...	NA	New York City Residency is not required for th...	12/8/2023	No Residency Requirement	Community Programs	112935.0
1	Information Technology Infrastructure Manager	RESEARCH PROJECTS COOR(MA)-MGR	Pending Classification-2	M1	Technology, Data & Innovation	F	Experienced (non-manager)	...	Annual	The Mayors Office of Criminal Justice (MOCJ) a...	1. Do you have a baccalaureate degree from an ...	\tExcellent technical knowledge of Network har...	NA	New York City residency is generally required ...	2/1/2024	NYC Residency Required	Technology and Data	86850.0

Let's Explore Data

Salary Analysis

Requirements and Qualifications

Q1

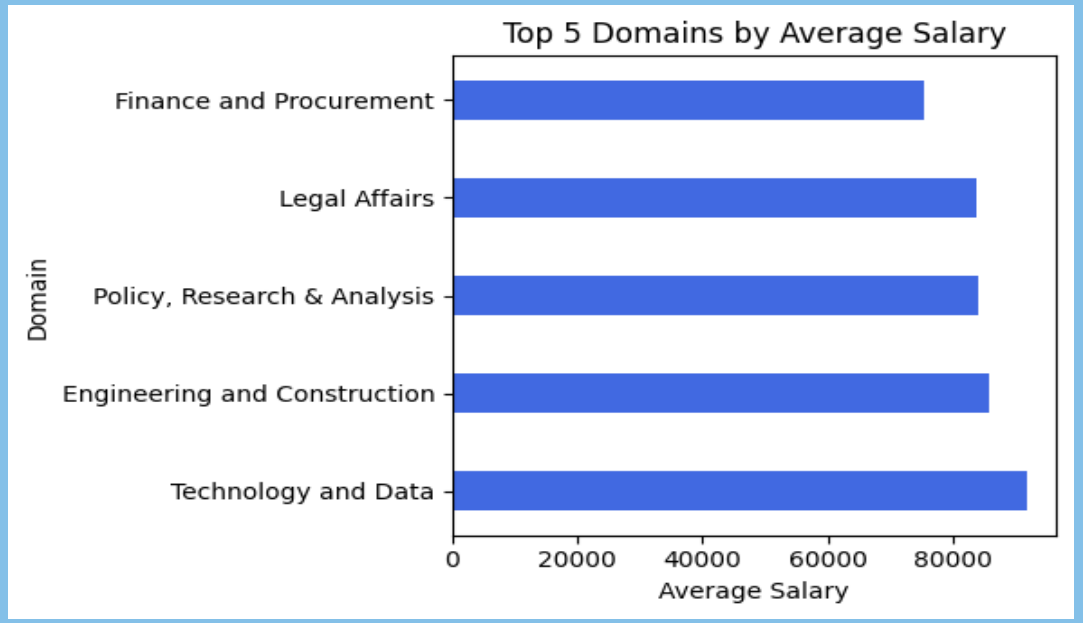
Q2

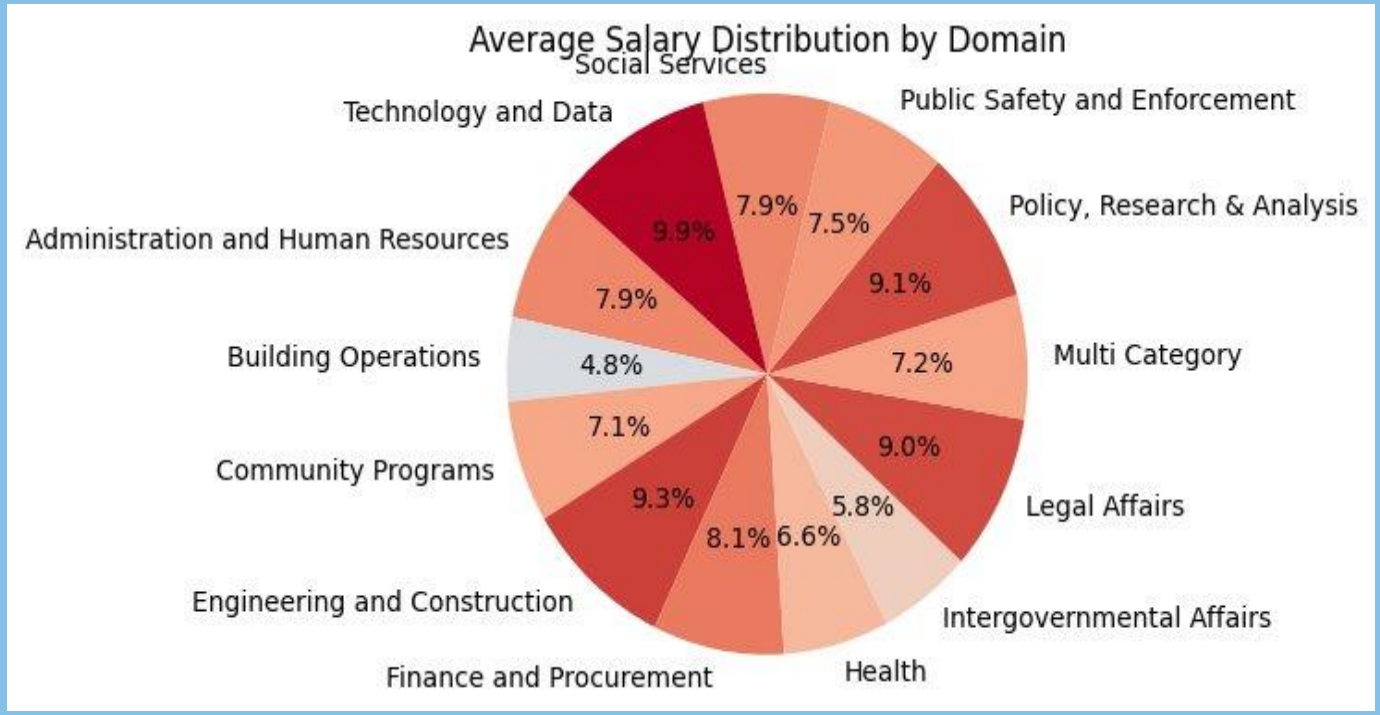
Q3

Job Distribution and Trends

Salary Analysis

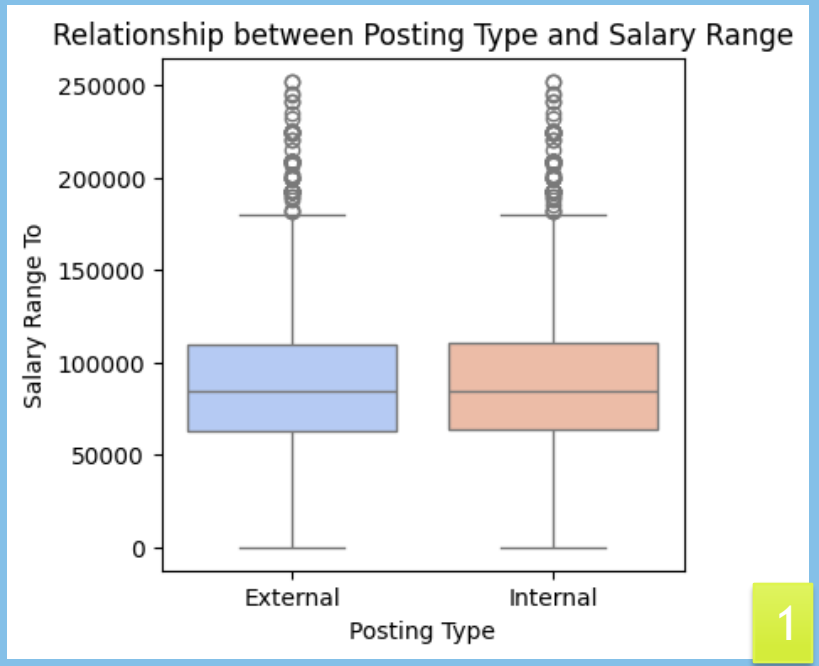
- The bar chart shows the top highest-paying job domains based on average salary in our data.
- Domains with the longest bars offer the highest average salaries.
- The pie-chart gives a complete overview of how the salary is distributed throughout all the domains.
- These domain are clubbed based on the Job Category attribute.
- The color keeps fading as the salary lowers.
- Here we can see how the salaries of Admin, HR, Finance, and public safety fall into middle tire while the building operations and intergovernmental affairs fall in lower tire.

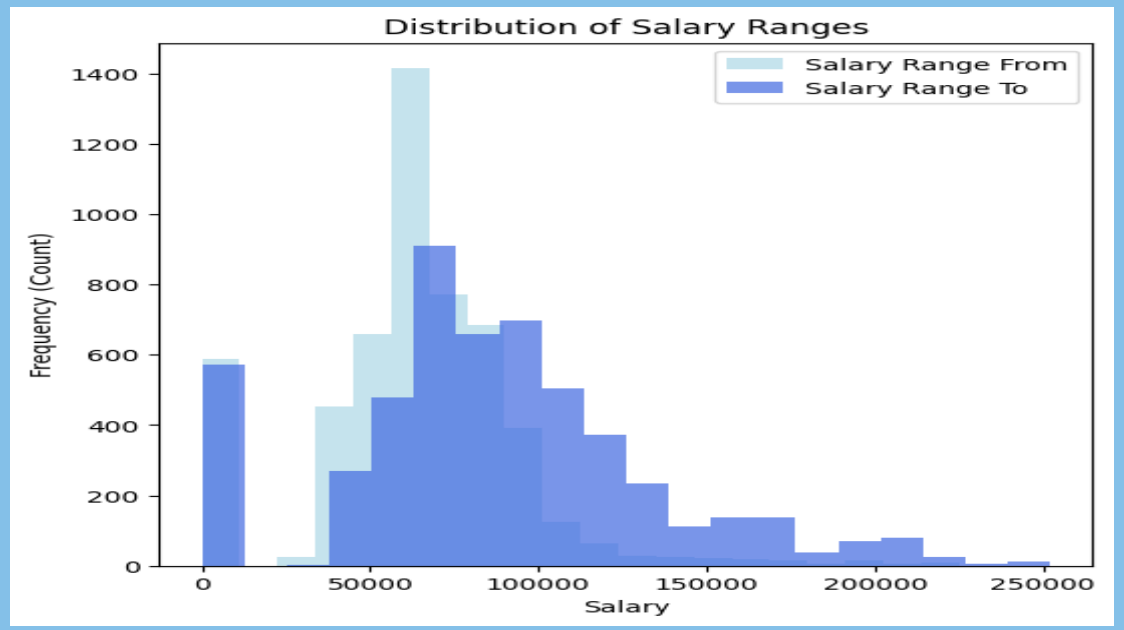


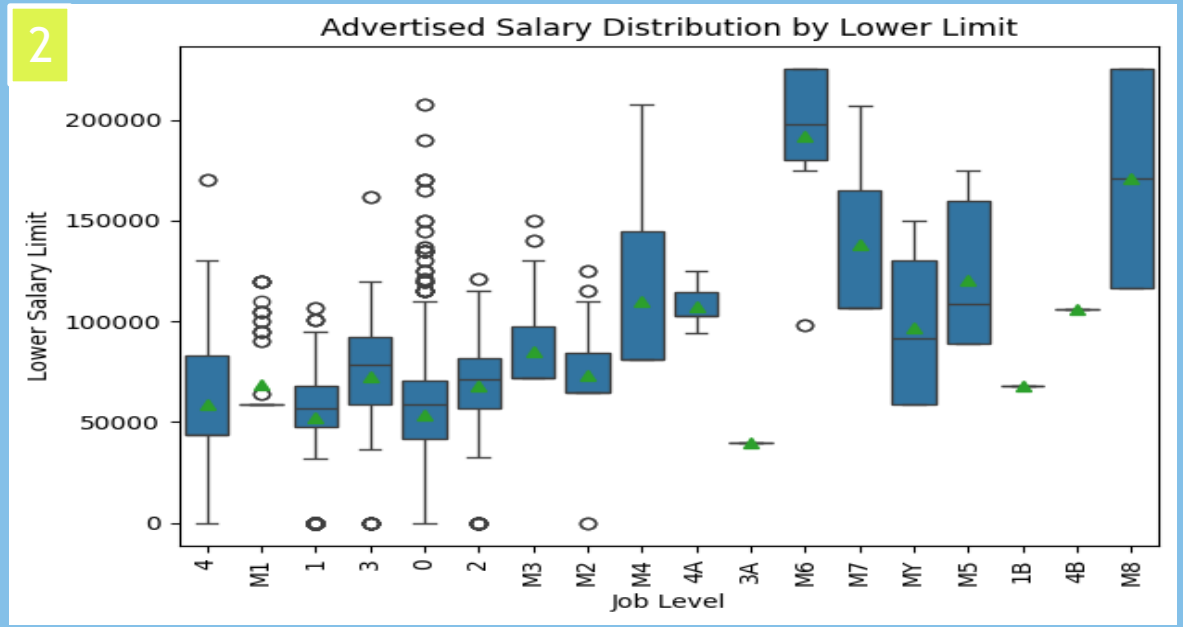


Salary Analysis

- The 1st box plot depicts the relationship between job posting type and salary range. The horizontal line in the middle of each box represents the median salary within that posting type.
- The box itself depicts the interquartile range (IQR), which represents the middle 50% of the salary data for that posting type. The bottom and top edges of the box extend to 1.5 times the IQR from the median.
- The whiskers extend from the top and bottom of the box to the most extreme data points within 1.5 times the IQR. Any data points beyond these whiskers are considered outliers and are plotted as individual circles. The plot shows similarity between the external and internal job postings in terms of upper limit of the salary range.
- The 2nd plot visualizes the distribution of advertised lower salary ranges by job level. The small diamonds scattered across the boxes represent the mean salary (average) for each job level. Higher Managerial job levels tend to have higher median salaries.
- The histogram visualizes the distribution of salary ranges in the dataset. The data suggests that a salary range between $80,000 and $100,000 is most frequently advertised across the job postings we analyzed.



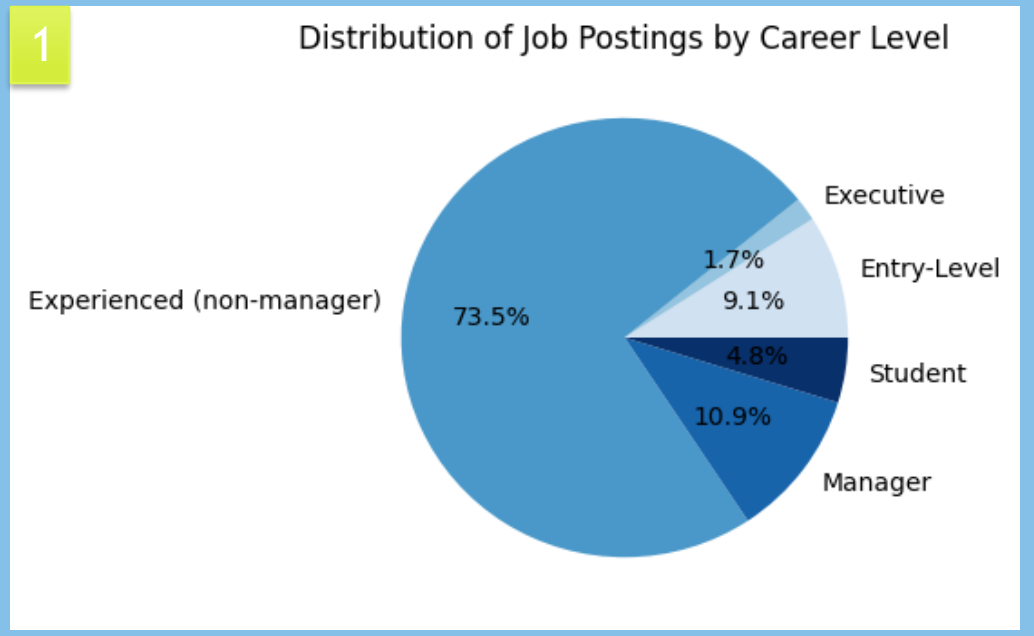




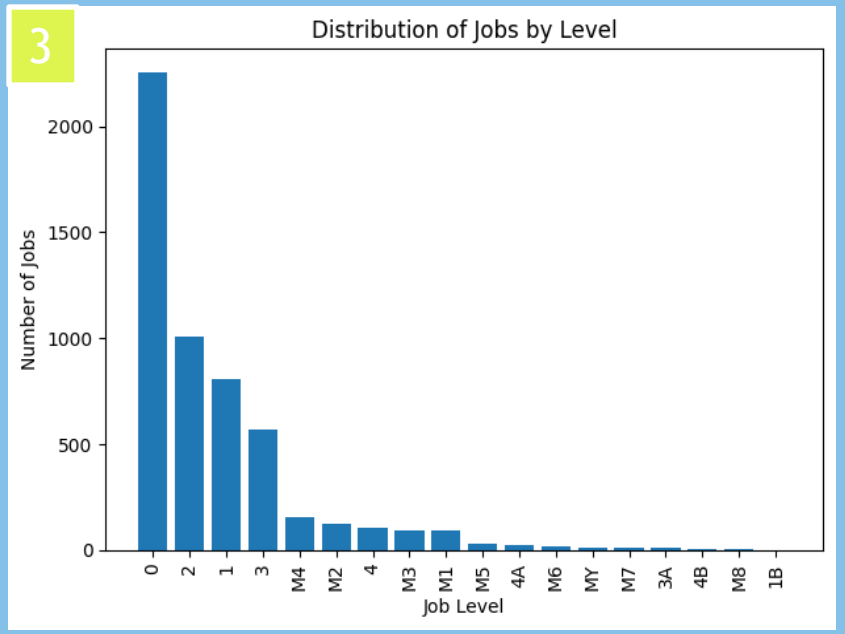
Job Distribution and Trends

- The 1st pie chart shows the distribution of job postings across various career levels in our data.
- Each slice represents a career level, and its size corresponds to the proportion of job postings in that category.
- The data shows that Experienced (non-manager) level positions are the most common type of job posting in our dataset.
- The 2nd pie chart shows the distribution of job postings across residency requirements in our data.
- The data shows that most job postings in our dataset require "NYC residency".
- The bar chart depicts how many job postings exist for each job level in the data set.
- The level 0 with the tallest bar has the most job postings.

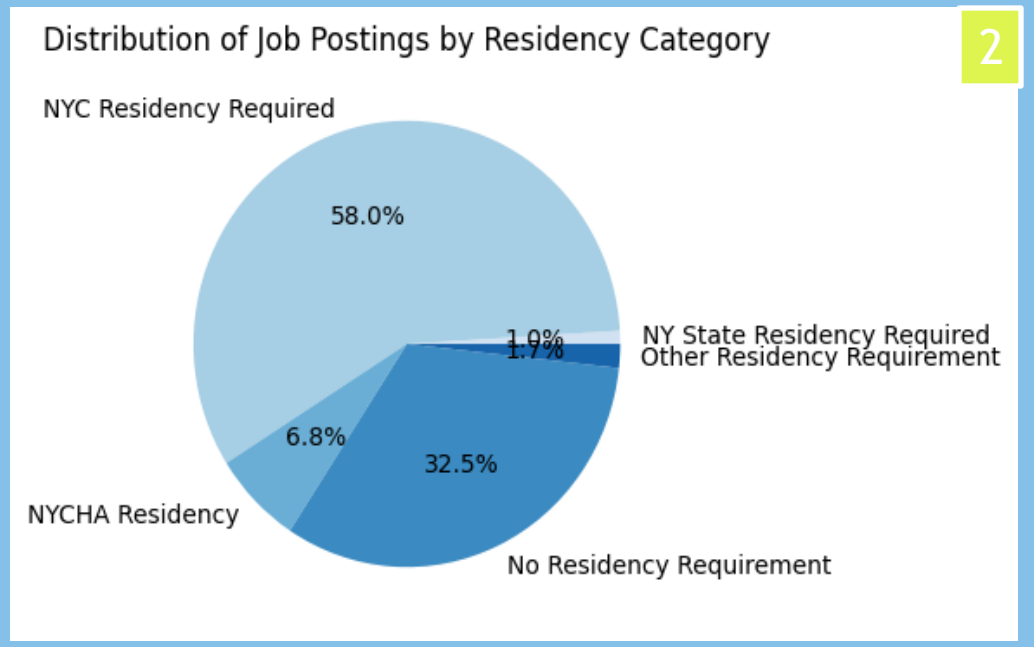
1



3

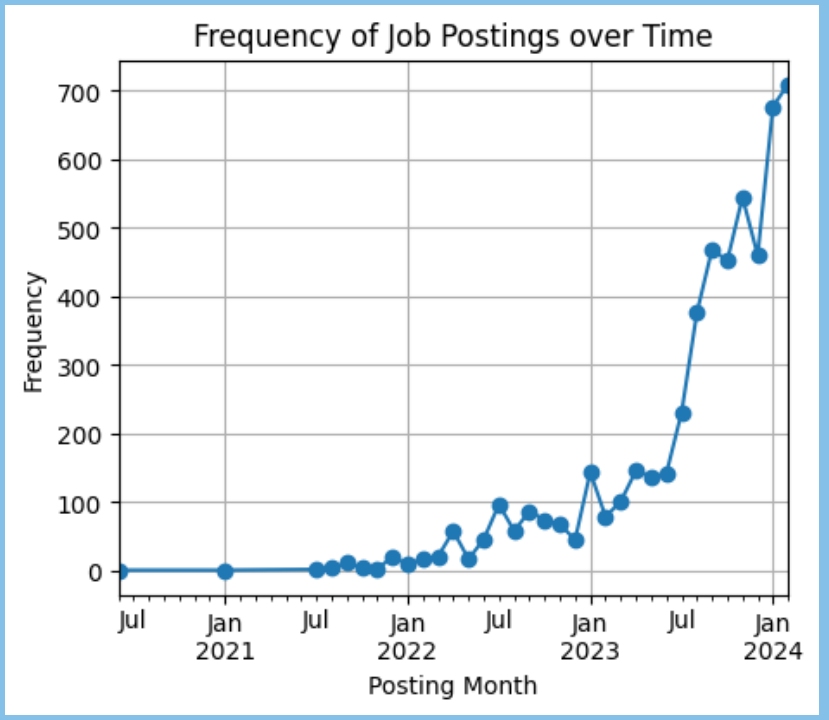


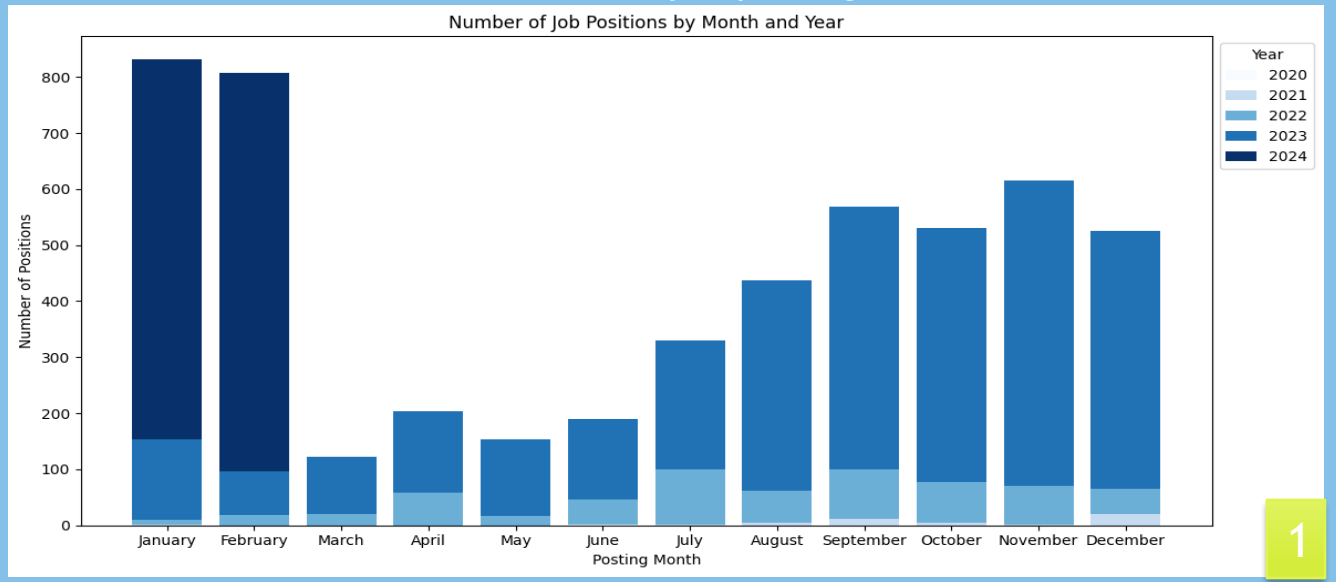
2



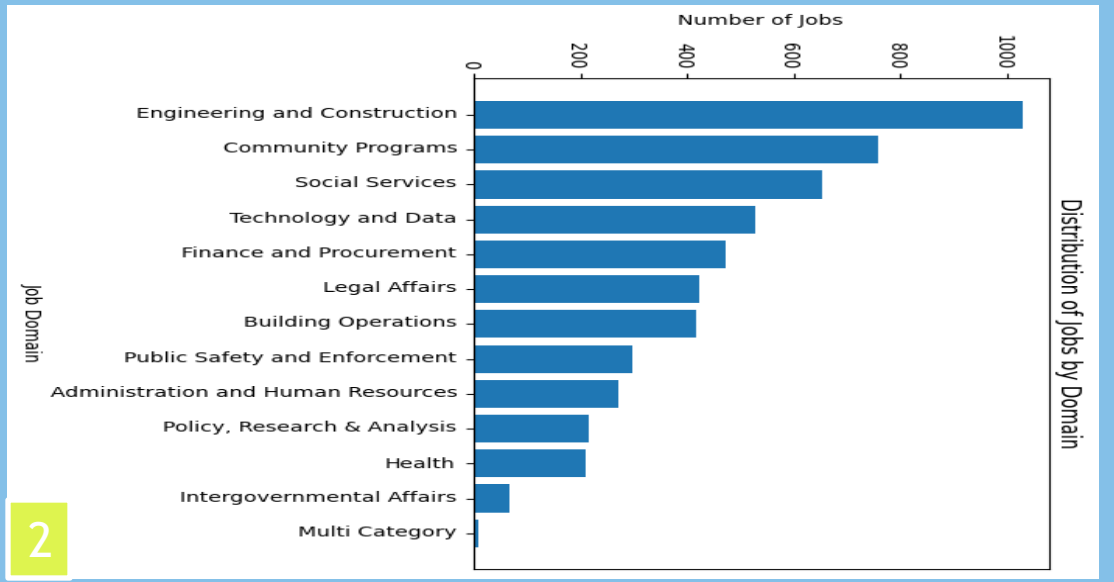
Job Distribution and Trends

- The line plot depicts the frequency of job postings over time (by month) in your dataset.
- The circular markers connected by a line represent the trend in the number of postings over time.
- The 1st bar graph represents the number of job postings added each month across various years in the data set.
- We see a surge in postings towards the end of the year in 2023.
- The 2nd bar chart depicts how many job postings exist for each job domain in your data set.
- The "Engineering & Construction", followed by "Community Programs" and "Social Services" domains have the most job postings.





1

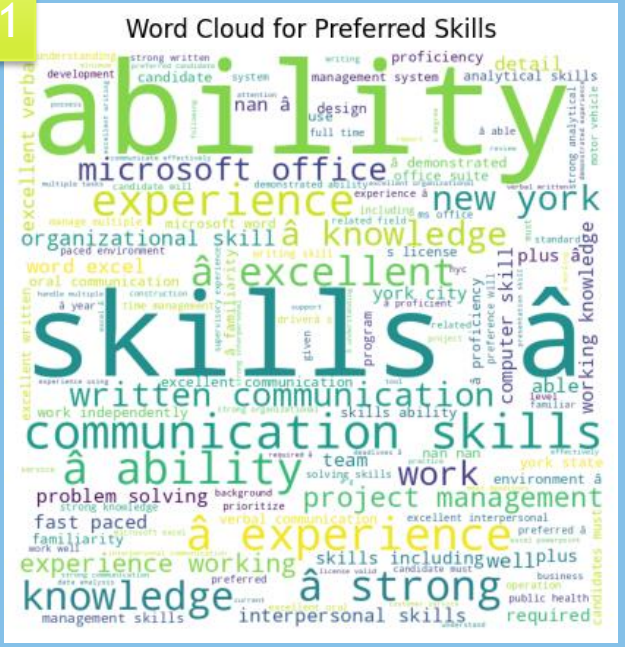


2

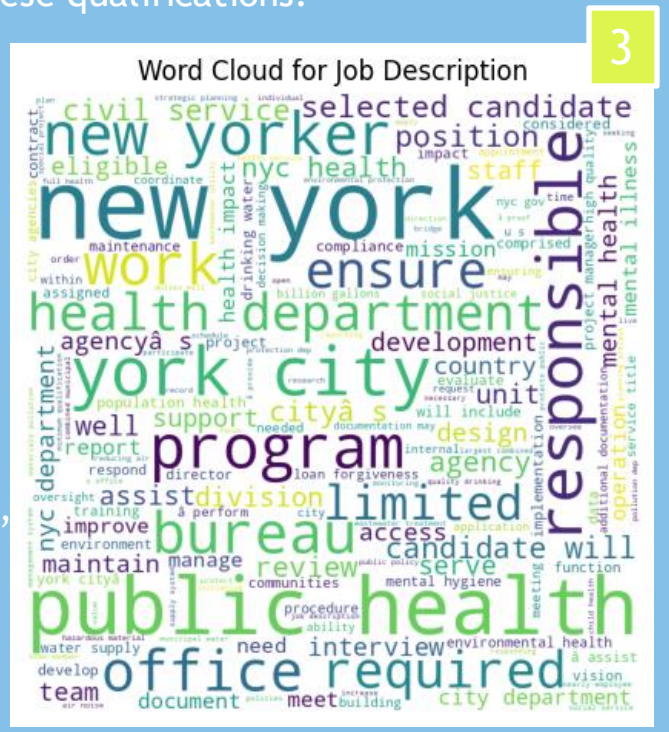
Job Requirements and Qualifications

- The 1st word cloud depicts the most sought-after skills mentioned in the **Preferred Skills** within our data set. The larger a word appears, the more frequently it is mentioned across the job postings.
- Some of the most prominent skills include Microsoft office, communication skills and written communication highlighting the importance of these skills for the jobs we analyzed.
- The 2nd word cloud depicts the most sought-after **Minimum Qualifications** mentioned in the Min Qualifications Requirement within our data set.
- Some of the most prominent qualifications include college, full time, four-year, school diploma and university highlighting the importance of these qualifications.
- The 3rd word cloud visualizes the most frequently used words listed in the **Job Description**.
- We can see trend of 'new yorker' , 'new york' , 'york city' being a big part of the job descriptions, even though we have a separate attribute/field for it as 'residency req'
- Apart from that we have domain specific words as bureau, health, civil services etc.
- Skills/ keywords that could be identified are responsible, assist etc. One would have to check the minimum qualifications or preferred skills

1



3



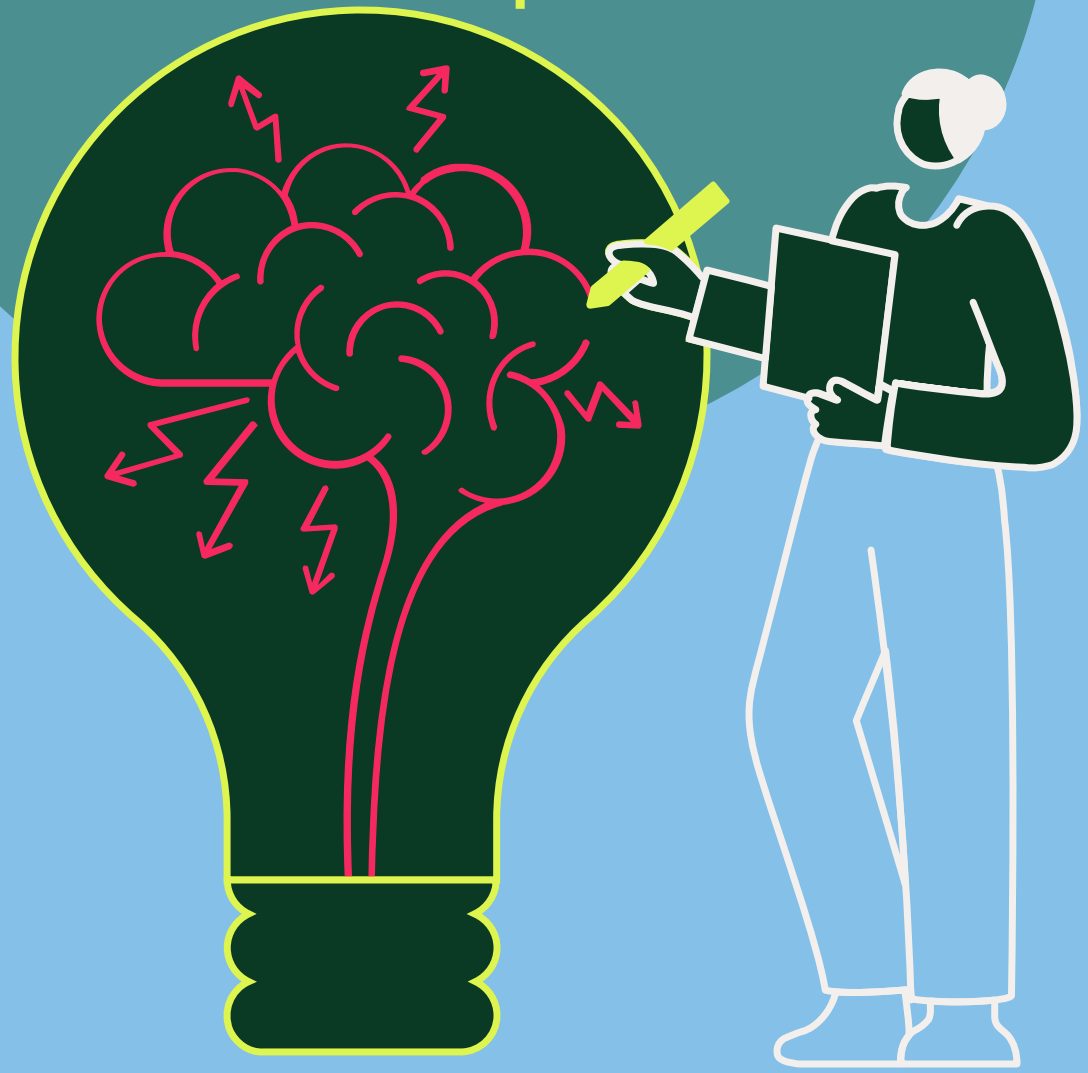
2

Prediction Model with Random Forest Regression

- **Feature Selection:** We identified relevant features such as 'Business Title', 'Title Classification', 'Job Category', and 'Career Level' from the dataset.
- **Data Preprocessing:** Rows with missing values in the selected features were dropped to ensure data integrity.
- **Encoding Categorical Variables:** Categorical variables were encoded using one-hot encoding to convert them into numerical form at suitable for the model.

- **Model Training:** Random Forest Regression was initialized with 100 trees and trained on the preprocessed data.
- **Model Evaluation:** The trained model was used to make predictions on the test set to predict Average Salary, and the accuracy was 89%.
- **Model Evaluation Metrics:**
  R-squared (R2) Score: 0.89
  Root Mean Squared Error (RMSE): 12,502.16
  Accuracy Percentage: 89.16%

Conclusion & Future Scope



Current Usage:

Decision Support: The analysis of average salary prediction using Random Forest Regression can serve as a valuable decision support tool for human resource departments, policymakers, and job seekers.

Resource Allocation: Organizations can utilize the insights gained from this analysis to optimize resource allocation, budget planning, and talent acquisition strategies.

Future Scope:

Enhanced Feature Engineering: Exploring additional features such as education level, years of experience, and geographical location could further enhance the predictive power of the model.

Integration with HR Systems: Integrating the predictive model with existing HR systems can streamline the recruitment process and improve workforce planning.

Real-time Monitoring: Implementing a real-time monitoring system to track changes in salary trends and market dynamics can provide timely insights for decision-making.

Potential Benefits:

Cost Savings: Optimized resource allocation based on accurate salary predictions can lead to cost savings for organizations.

Talent Retention: Insights gained from the analysis can help identify factors influencing employee satisfaction and retention, leading to improved talent management strategies.

Competitive Advantage: Leveraging advanced predictive analytics can provide a competitive edge in talent acquisition and workforce planning.

Thank you