



Understanding Employment Trends in NYC Job Postings

NIKHIL PATIL

SHIMIL SHIJO

SRINIVASA REDDY GURRAM

Introduction

Complex Job Market:

- NYC's diverse job landscape presents challenges in understanding salary trends.
- Inaccurate salary prediction hinders fair compensation and efficient hiring in NYC.

Data-Rich Opportunity:

- A large dataset of job postings provides a unique opportunity for advanced analysis.

Actionable Insights:

- Developing predictive models to inform better hiring practices and career choices.
- Uncovering the factors driving salary differences and improving hiring practices.

Impact:

- Data-driven solutions for a more equitable and efficient NYC job market.

Dataset

- The dataset can be accessed [here](#)

Data Source:

- NYC Open Data, a publicly accessible resource
- Maintained by the City of New York.
- Ensures transparency and allows for independent verification of the data.

Data Authenticity:

- Officially published by the City of New York, high credibility to its source.

Variables:

- The dataset includes variables such as Job ID, Agency, Posting Type, Number of Positions, Business Title, Civil Service Title, Salary Range, Work Location, Job Description, Minimum Qualifications, and Posting Dates.

Data Types:

- The data includes a mix of categorical
 - *(nominal: Agency, Posting Type; ordinal: Career Level)*
 - *numerical (continuous: Salary Range; time series: Posting Date) variables.*

Data Size:

- 5411 Records and 30 Columns

Data Preprocessing & Feature Engineering

Data Loading:

- Loaded NYC job postings data into a Pandas DataFrame.

Data Cleaning:

- Renamed columns for consistency (replaced spaces and / with _).

Missing Data Handling:

- Identified missing values but chose not to impute or remove them due to irrelevance to the analysis.

Feature Engineering:

- Created a binary 'Competitive' feature from 'Title_Classification'.

Encoding:

- Converted categorical features to numerical using Label Encoding.
 - 'Posting_Type'
 - 'Agency'
 - 'Career_Level'
 - 'Full_Time_Part-Time_indicator'
 - 'Civil_Service_Title'

Unnecessary Columns Removed:

- Dropped the 'Job_ID' column.

Research Questions

Research Question	Expected Outcome
Do different populations within similar job titles have observable salary differences?	How the salary of a job is determined by Posting_Type, Agency, Career_Level and Full-Time_Part-Time_indicator. Evaluate the correlation/influence of each of these factors in determining the average salary.
What are the main wage distinctions between competitive job postings and those that are not?	Determination of key salary differences between competitive and non-competitive job postings, potentially identifying factors contributing to these distinctions.
Do residency requirements make some organizations more likely to have open positions?	Assessment of the correlation between residency requirements and the number of open positions across different organizations.
Does the demography that is drawn to external postings differ from that of internal postings?	Identification of any significant differences in the demographic profiles (e.g., career level, work location) of applicants for external versus internal job postings.

Methodology and Expected Outcomes

Methodology Type	Detailed Steps	Expected Outcome/Results (How the Question Will Be Answered)
Linear Regression	<ul style="list-style-type: none">Regression (OLS): Feature selection, data encoding, model training, evaluation.Clustering (K-Means): Feature selection, data scaling, clustering, optimal cluster determination (Elbow & Silhouette), visualization (scatter plots, box plots).	<ul style="list-style-type: none">Regression analysis will quantify the impact of demographic factors on salary.Clustering will identify groups of job postings with similar salary and demographic profiles, highlighting potential disparities.
Classification	<ul style="list-style-type: none">Data Preparation: Feature engineering, data encoding, train-test split (80/20).Classification (Random Forest):Hyperparameter tuning , model evaluation (Accuracy, Precision, Recall, F1-score),Confusion Matrix.Feature Importance: Identify key features driving salary differences.	<ul style="list-style-type: none">Classification models will predict the competitiveness of a posting based on its characteristics.Feature importance analysis will reveal the most influential factors in determining competitiveness.
Regression	<ul style="list-style-type: none">Residency Classification: Used the Logistic Regression, model to classify job postings by residency requirement (No Residency, Residency Required, Unclear).Used GridSearchCV, Kfold cross validation and L2 regularization	<ul style="list-style-type: none">Regression models will predict the number of open positions based on residency requirements and other factors.
Descriptive Statistics	<ul style="list-style-type: none">Bar charts (posting type distribution),box plots (salary distribution by posting type),stacked bar charts (career levels by posting type),grouped bar charts (work locations by posting type).	<ul style="list-style-type: none">Visualizations (bar charts, box plots, etc.) will compare demographic distributions for external and internal postings.

Research Question

1 - Regression

Do different populations with similar job titles have observable salary differences?

- Identify key variables: salary (dependent variable), agency, career level, and full-time/part-time indicator (independent variables).
- Collect and Preprocess Data
- Use descriptive statistics to summarize key variables. Visualize relationships using plots.
- Choose a statistical model for examining relationships between salary and multiple independent variables.
- Fit the Model to analyze the results.
- Interpret results by Summarizing the findings between salaries based on agency, career level, and employment type.

Research Question 1 - Result

Model:

- The model explains 20% of the variation in average salary ($R^2 = 0.20$), suggesting several other factors influence pay.

Key Salary Predictors:

- **Career Level:** A significant positive impact on salary. Higher career levels are associated with substantially higher average salaries (+11,200).
- **Agency:** The agency employing the worker significantly influences salary (-200). This suggests that some agencies pay more than others on average.
- **Full-Time/Part-Time Status:** Part-time positions (coded 1) have significantly lower average salaries (-71,883) compared to full-time positions (coded 0).

Non-Significant Predictors:

- The 'Posting Type' variable was not statistically significant in predicting salary.
- This suggests that whether a posting is internal or external does not significantly affect the average salary.

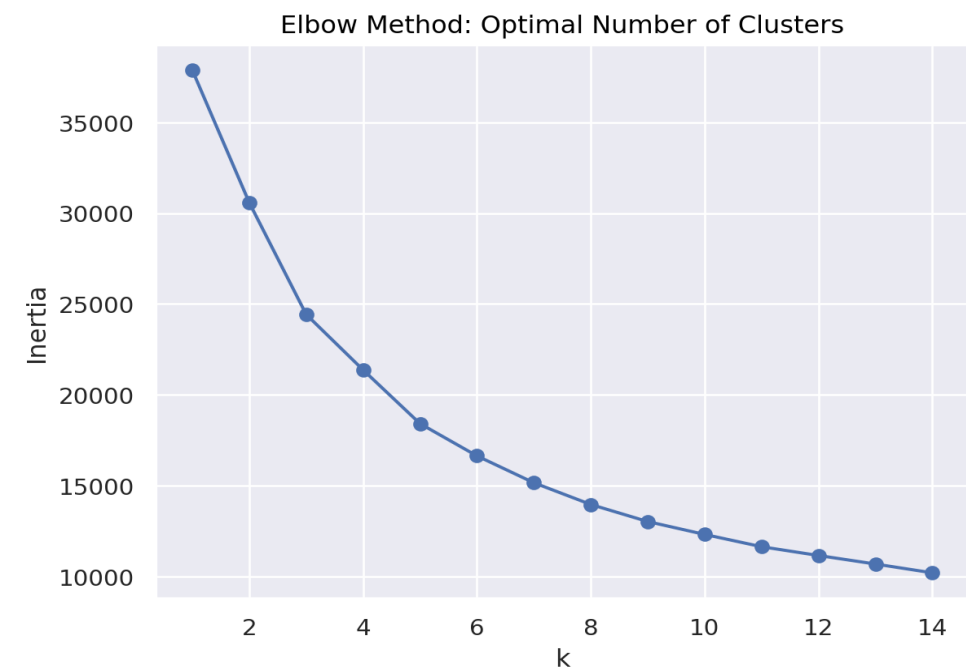
Further Investigation:

- Highlights the significant impact of career level and employment type
- Suggesting further analysis is needed to understand these relationships in more detail.

Research Question 1 - Result

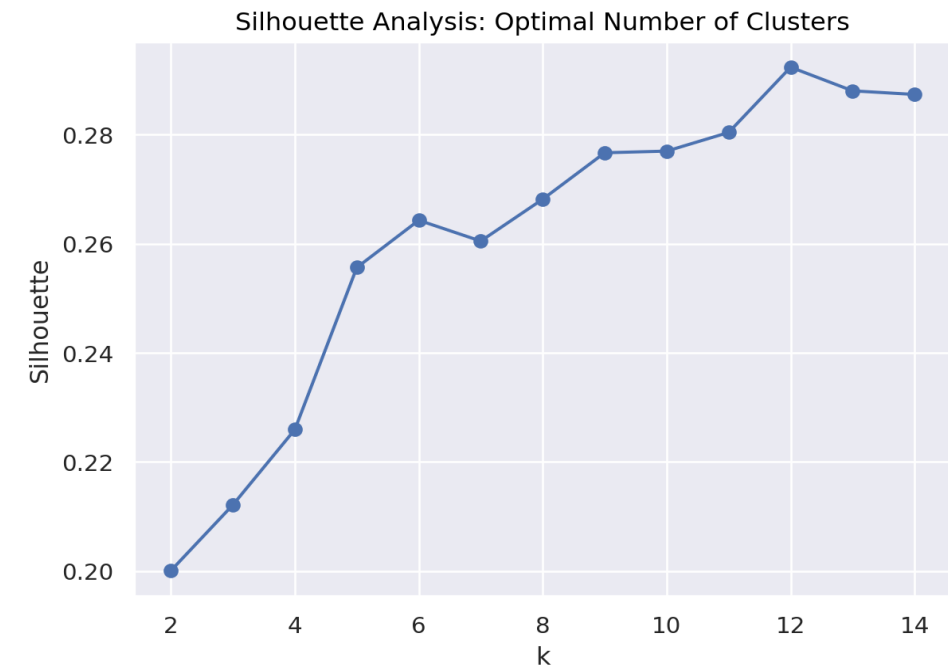
Elbow Method:

- The **Elbow Method** is a technique to find the optimal number of clusters by analyzing the inertia.
- A for loop iterates over (k), the number of clusters, ranging from 1 to 14. Inside the loop it initialized the K-Means clustering model with k clusters, with 20 different centroid seeds and ensures reproducibility of results using random_state.
- This elbow indicates the optimal kkk, balancing clustering performance and simplicity.



Silhouet Analysis:

- The **Silhouette Analysis** is used to determine the optimal number of clusters (k) for a K-Means clustering algorithm. By calculating how well each data point fits into its cluster and how different clusters are from one another, it analyzes the quality of clustering.
- For visualization the silhouette list was taken into a Pandas DataFrame with columns (k) and Silhouette scores.



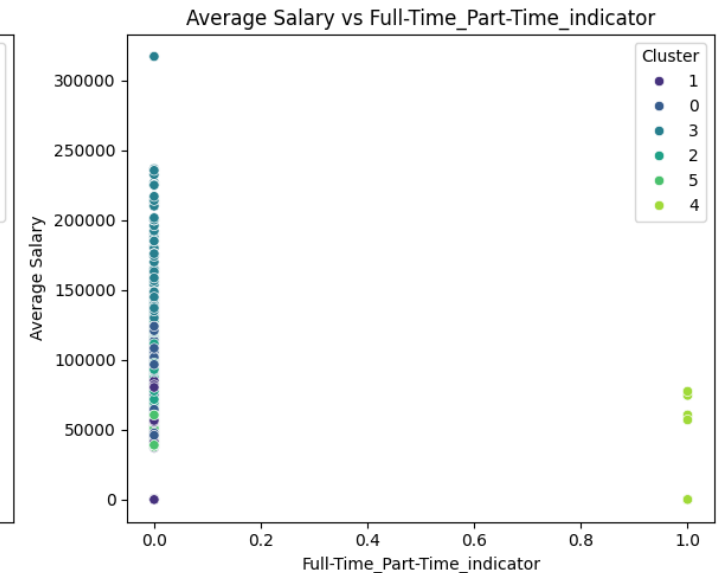
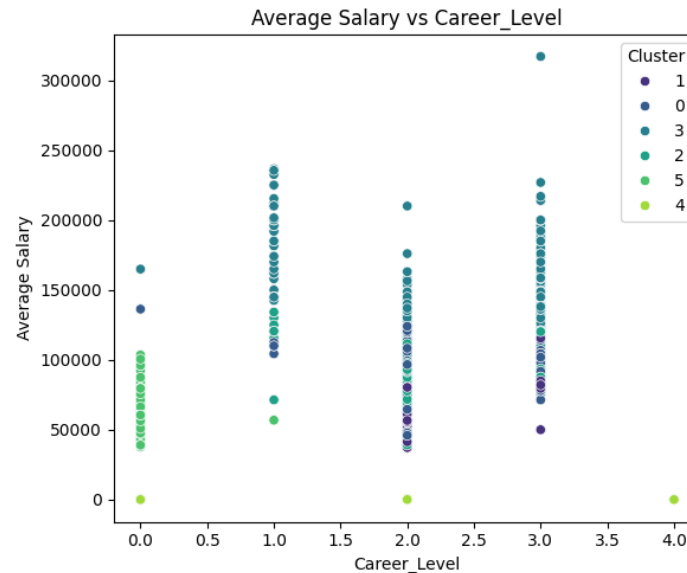
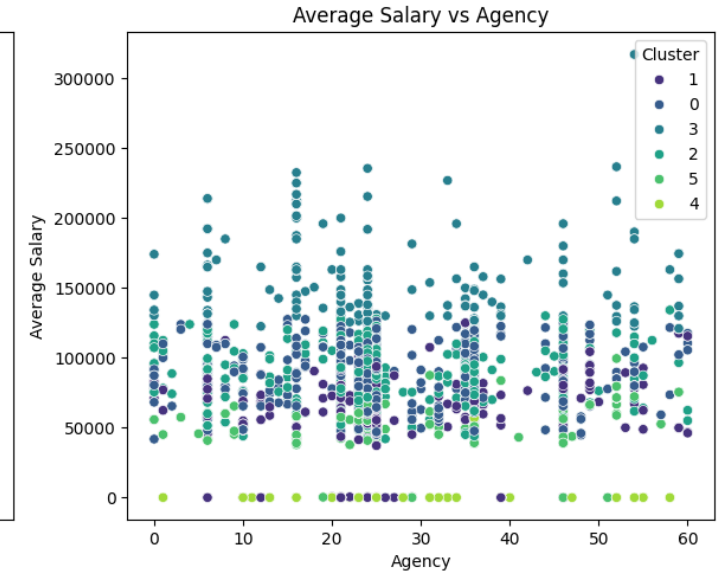
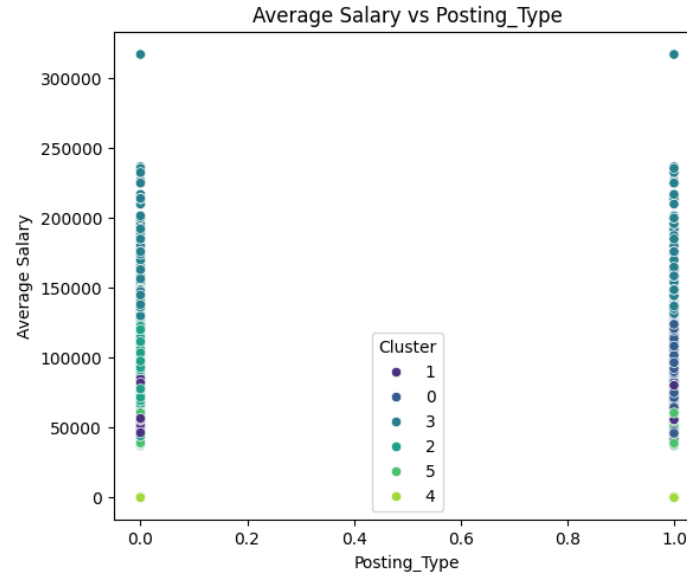
Cluster analysis with k=6

- **Distinct Salary Groups:**

- The six clusters exhibit significant salary variation, with some clusters concentrated at higher salary levels than others. This indicates that the clustering effectively captured salary-based distinctions in the job postings.

- **Feature Relationships:**

- The clustering analysis struggles to form well-defined clusters using the features Posting_Type, Agency, Career_Level, Full-Time_Part-Time_indicator in relation to Average_Salary



Research Question 2 – Classification

What are the main wage distinctions between competitive job postings and those that are not?

- Identify whether the metrics balance business needs, such as prioritizing precision, recall, or F1-score for competitive/non-competitive labels.
- Collect and Preprocess Data
- Validate that the ROC curve captures meaningful trade-offs between the true positive rate (TPR) and false positive rate (FPR).
- Investigate if high precision and recall persist across all thresholds or are influenced by data imbalances between the two labels.
- Confirm if the sample sizes for competitive and non-competitive salaries are comparable.
- Verify why median salaries are similar despite wider variability in non-competitive roles.

Research Question 2 - Result

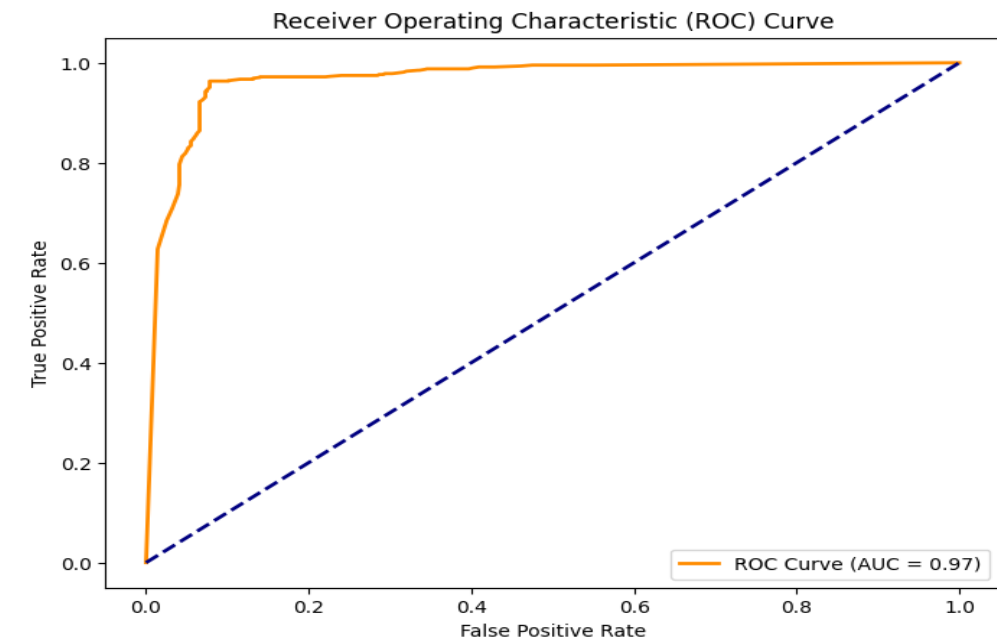
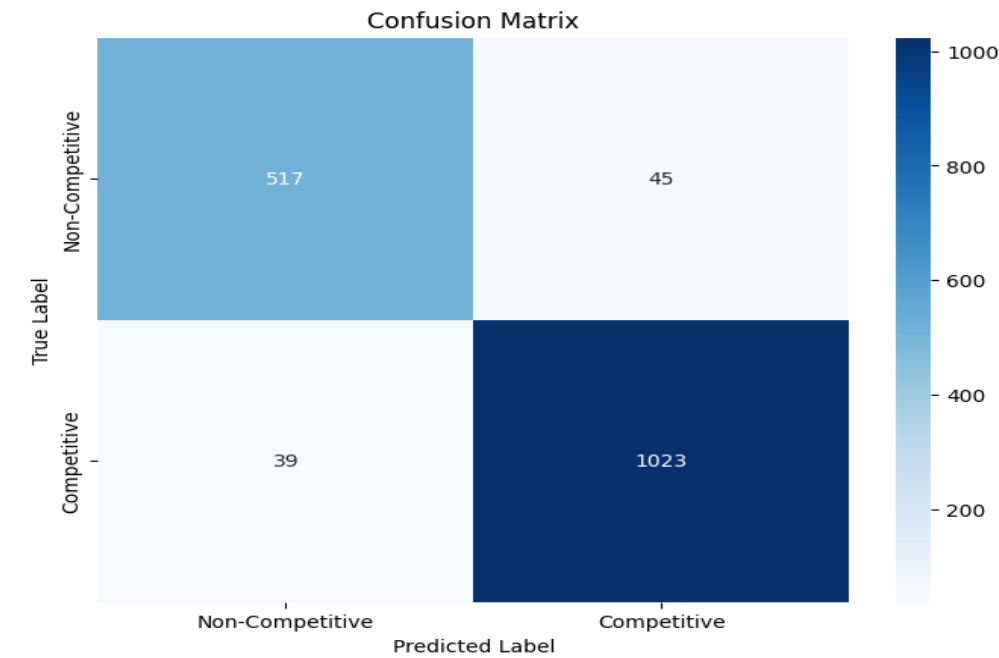
Clas

Classification Metrics and Confusion Matrix

- Accuracy: 95%
- Precision & Recall:
- Non-Competitive (Label 1): Precision (96%), Recall (96%), F1-Score (96%)
- Competitive (Label 0): Precision (93%), Recall (92%), F1-Score (92%)
- Confusion Matrix:
 - Correctly identified:
 - 517 non-competitive postings
 - 1023 competitive postings
 - Misclassifications:
 - 45 competitive as non-competitive
 - 39 non-competitive as competitive

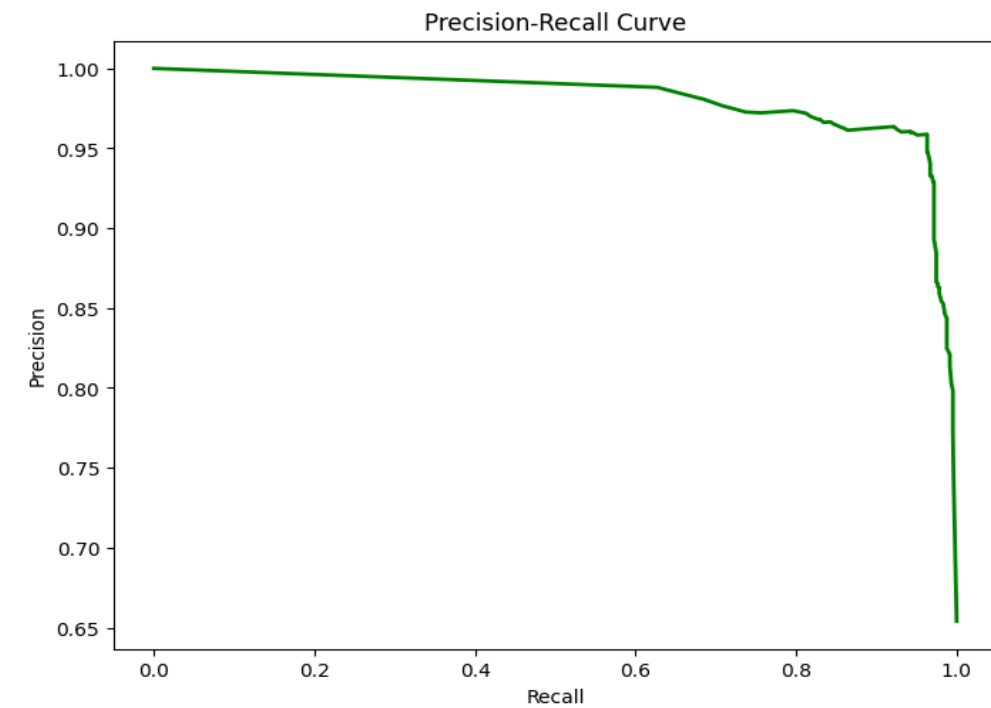
ROC Curve

- AUC: 0.97, indicating strong discriminatory power.

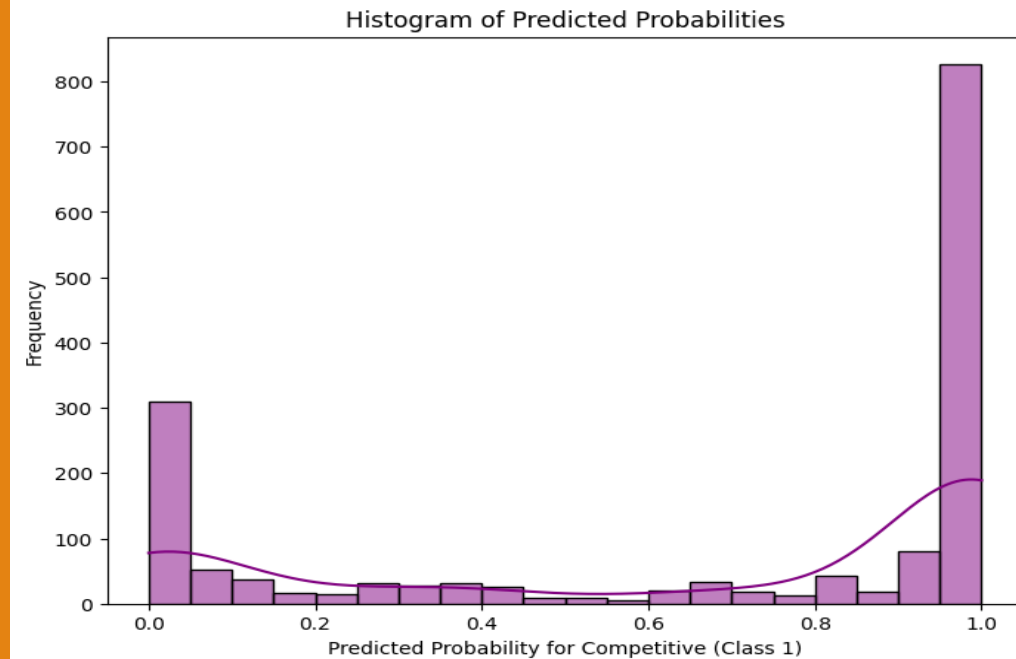


Research Question 2 - Result

Precision-Recall Curve: High precision and recall across thresholds, confirming reliable predictions, particularly for competitive postings.



Histogram of Predicted Probabilities - Predominantly extreme values (near 0 or 1), indicating high model confidence. - Few predictions with moderate probabilities (0.4 to 0.6), suggesting minimal ambiguity.



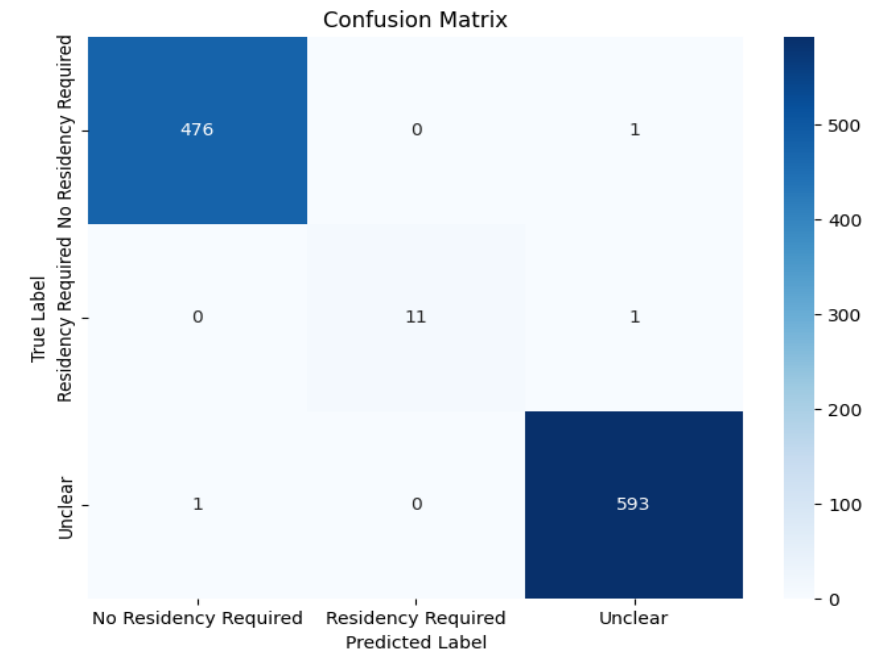
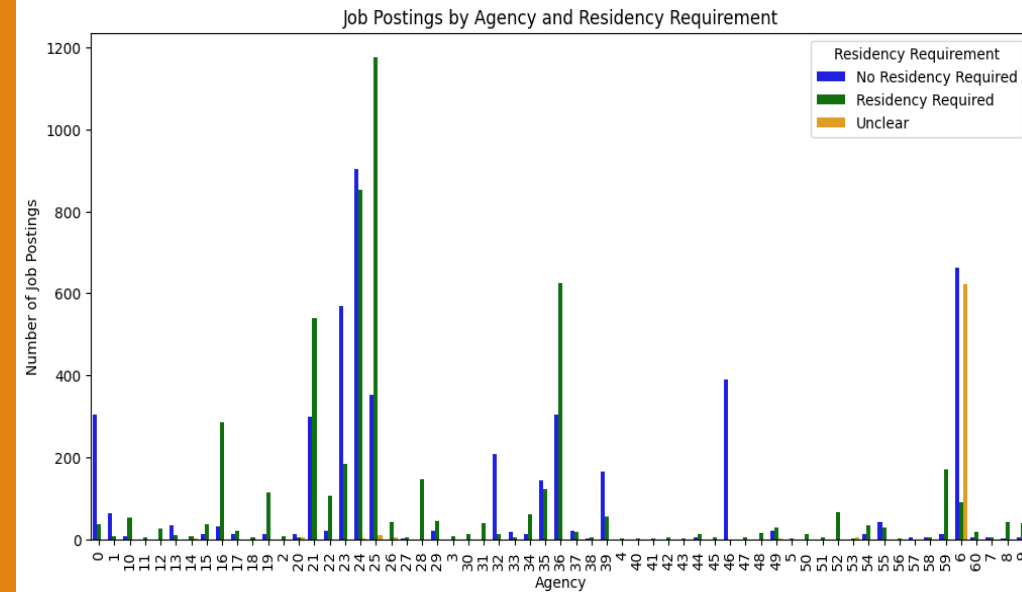
Research Question 3 - Methodology

"Do residency requirements make some organizations more likely to have open positions?"

- Identify the fields and transform 'residency requirement' field to apply classification.
- *Data preparation & Splitting*
- *Feature Engineering & Pipeline*: Use TF-IDF for text vectorization and build a Logistic Regression pipeline with L2 regularization.
- *Hyperparameter Tuning*: Optimize Logistic Regression using GridSearchCV.
- *Model Evaluation*: Evaluate with a classification report and confusion matrix.
- *Cross-Validation*: Perform 5-fold cross-validation for robust accuracy.

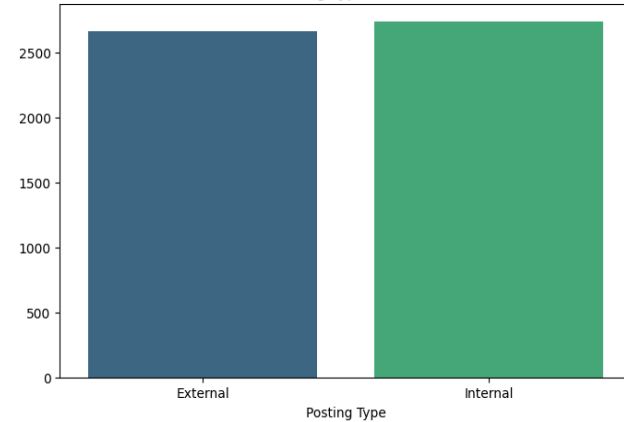
Research Question 3 - Results

- **Highly Accurate Model:** The Logistic Regression model, optimized using GridSearchCV, achieved near-perfect classification accuracy (99.57% based on k-fold cross-validation), demonstrating excellent predictive power in identifying residency requirements from job postings.
- **Near-Perfect Prediction for Most Categories:** The model showed exceptional performance in predicting postings with "No Residency Required" and "Unclear" requirements.
- **Slight Underperformance for "Residency Required":** While precision remained high (100%) for the "Residency Required" category, recall was slightly lower (92%), indicating a small number of false negatives (the model missed some postings that actually required residency).
- **Minimal Misclassifications:** The confusion matrix confirmed the model's high accuracy, showing very few misclassifications across all categories.
- **High Model Consistency:** The high mean cross-validation accuracy (99.57%) demonstrates that the model generalizes well to unseen data and is not overfitting.
- **Agency-Level Variation:** The analysis revealed substantial variation in residency requirements across different agencies. Some agencies predominantly posted jobs with clear residency requirements, while others showed a mixed pattern or lacked clear requirements.

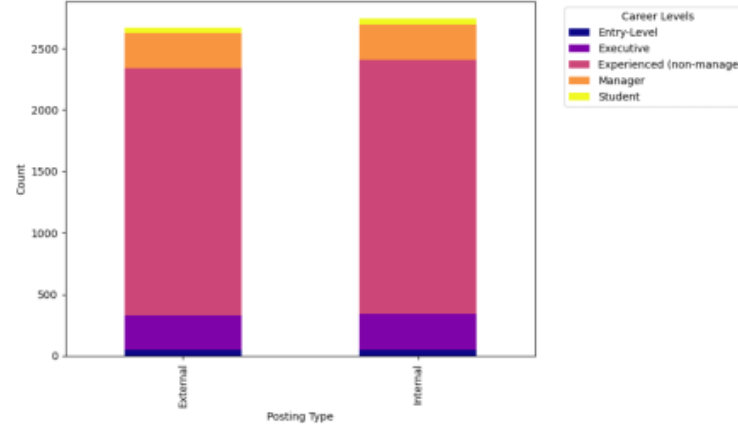


Research Question 4 - EDA

Distribution of Posting Types (External vs Internal)

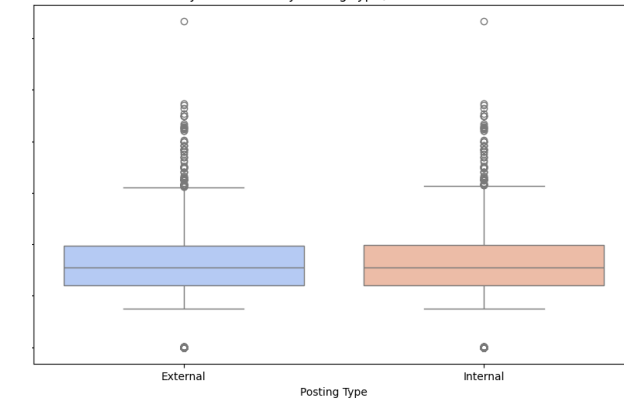


Career Levels by Posting Type (External vs Internal)

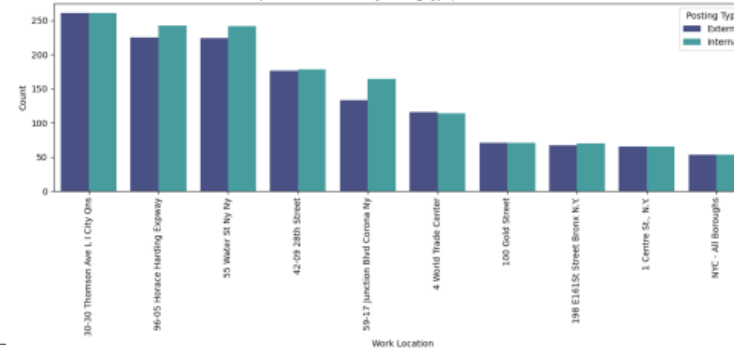


- **Balanced Posting Types:** The dataset contains a roughly even distribution of internal and external job postings, with a slight majority of internal postings.
- **Slightly Higher Median Salary for Internal Postings:** Internal postings tend to offer slightly higher median salaries than external postings, suggesting potential pay differences based on posting type. Variability (IQR) in salaries was similar.
- **Similar Career Level Distributions:** Both internal and external postings predominantly target experienced (non-managerial) professionals. The distribution of other career levels (manager, student, entry-level, executive) was also consistent across both posting types.
- **Consistent Work Location Distribution:** Most work locations showed a similar distribution of internal and external postings. While a few locations exhibited a slight preference for either internal or external postings, this was not a widespread pattern.

Salary Distribution by Posting Type(External vs Internal)



Top 10 Work Locations by Posting Type(External vs Internal)




Key Findings: External vs. Internal Job Postings


Near-Equal Posting Numbers: The number of internal and external job postings is nearly balanced, with a slightly higher count of internal postings (2741 vs. 2670).



Similar Average Salaries: Average salaries for internal and external postings are very similar (approximately \$79,747 and \$79,169, respectively), suggesting that posting type is not a major determinant of salary.



Consistent Salary Distributions: The standard deviations of salaries are also very similar for both posting types, indicating comparable salary variability.



Inconsistent Career Level Distribution: The distribution of career levels is not identical. The vast majority of postings target experienced professionals.

Conclusion and Future Scope

This study successfully analyzed a large dataset of NYC job postings to explore salary trends and hiring patterns.

Competitive job postings generally exhibited more consistent salaries than non-competitive postings (Research Question 2).

While significant variation in residency requirements exists across agencies, the relationship between residency requirements and the number of open positions requires further investigation (Research Question 3).

The demographic profiles of candidates applying for internal vs. external postings showed remarkable similarities (Research Question 4).

Future Scope :

- Focus on incorporating additional variables (e.g., skills, experience) into the predictive models and conducting more in-depth analyses of the identified clusters.
 - *Integration with External Labor Market Data:* Combining the NYC job postings dataset with external datasets, such as national labor market statistics or industry-specific benchmarks, to provide a broader context and validate findings.
 - *Development of Interactive Dashboards:* Creating user-friendly visualizations and dashboards to allow policymakers and organizations to explore salary trends, residency impacts, and job classifications dynamically.
-

References

[1] Eric Kober. “New York City Jobs Data Shine Light on NYC’s Economic Strengths and Weaknesses”. In: Analytics Vidhya (2024). url: <https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/>.

[2] Shubham Lipare. “Kmeans clustering using Elbow and Silhouette”. In: Kaggle (2023). url: <https://www.kaggle.com/code/shubhamlipare/kmeans-clustering-using-elbow-and-silhouette>.

[3] Amer Alnuaimi and Tasnim Albaldawi. “An overview of machine learning classification techniques”. In: BIO Web of Conferences 97 (Apr. 2024), p. 00133. doi: 10.1051/bioconf/20249700133.