Capstone Project Final Report

New York Citywide Payroll Data (Fiscal Year 2014 - 2023)

Team 4: Xiaoting Zheng, Dhwani Patel

ALY6140 Python & Analytics Technology, Fall 2023

Professor Daya Rudhramoorthi

December 15th, 2023

Northeastern University

College of Professional Studies

I. Introduction

The realm of public sector employment and compensation is a complex and multifaceted area, often reflecting broader socio-economic dynamics. Meanwhile, in the era of rapid technological advancement, the role of technology departments in public administration has become increasingly pivotal. The database here has been sources from the 'Office of Payroll Administration' of NYC. Data has been annually updated each year beginning from 2014 till 2023. The data provides an input into the cities' 'Personal Management Systems'(PMS). Our capstone project centers on a detailed examination of the Citywide Payroll Data (Fiscal Year) for New York City, with a specific focus on the Technology Department vs other top labor-intensive departments. This project is an embodiment of applying theoretical knowledge acquired in our course to a real-world context, offering practical insights into the dynamics of employment and compensation in the technology sector of public service.

The primary objective of this analysis is to explore various facets of the payroll structure within the Technology Department and top 10 departments (number of employees) of New York City's administration.

Research Objectives:

- Investigating the factors that influence salary disparities among employees within the Technology Department.
- Ensuring the wage imparities between the technology and other top 10 departments within two of the most populated areas.
- Assessing the distribution and impact of overtime hours on compensation
- Evaluating the predictability of an employee's salary bracket based on their role and other relevant factors.
- Exploring the roles-based payroll within IT and Other departments if there are similarities or disparity for each role.

This analysis aims to provide an in-depth understanding of the compensation mechanisms in a critical sector of public administration, offering insights that could inform policy and operational strategies in the technology realm. To achieve these goals, we have delineated the following questions:

- What factors contribute to the variations in salaries within the Technology Department?

- How does the allocation of overtime hours vary among different roles within the Technology Department?
- What is the relationship between overtime pay and total compensation for various technology-focused job titles?
- Can we use job titles, departmental roles, and other variables to predict salary brackets for employees in the Technology Department?
- Is there any match between non-tech department's salary vs tech salary over the years?
- Factors contributing to payroll parities in top 10 highly employed departments of database for each fiscal year passing by.
- Do different seasons have different payroll for the same position in different areas of the same city?

Research Methodology

After going through the basic details mentioned in data dictionary, to understand the data batter for producing valuable insights it was analyzed using python frameworks. Methodology focused on understanding the data and processing it for generating valuable insights before and after cleaning the data.

- Data Pre-processing: Includes having the basic information and knowledge of what the data is about and contains for producing the insights.
- Exploratory Data Analysis (EDA): Conducting an initial exploration of the dataset to identify trends, patterns, and anomalies specific to the Technology Department.
- Data Cleaning: Removing anomalies and missing values along with irrelevant data which is not falling into research scope.
- Data Analysis: Creating data related insights on the research objectives and research questions.
- Predictive Modeling: Applying linear regression to predict salaries, classification models to categorize employees into different salary brackets, and time-series analysis to understand compensation trends over time within the Technology Department.

This methodological approach will ensure a detailed and nuanced exploration of the dataset, enabling us to derive substantive conclusions and recommendations pertinent to the technology sector in public administration.

II. Exploratory Data Analysis(EDA)

Overview and Descriptive Analysis of the Whole Dataset

The dataset at hand presented a formidable challenge with its extensive size, encompassing 5,662,713 rows and 17 columns. This vast pool of data offered a comprehensive snapshot of New York City's payroll, capturing an array of dimensions from job titles to salary details. Our initial steps involved gaining a macroscopic view of this dataset, particularly focusing on the distribution of employees across various agencies and departments. A significant part of this analysis was dedicated to understanding the geographical spread of the workforce, particularly identifying the top five work locations for the fiscal year 2023. This approach was instrumental in painting a broad picture of the city's employee distribution and setting the stage for more targeted analysis.

Data Filtering for the Technology Department

In our exploratory analysis, a bar plot revealed the Technology Department ranked 27th in employee count for NYC in 2023. Despite its modest size, we chose it for analysis, recognizing its significant role in modern public administration.

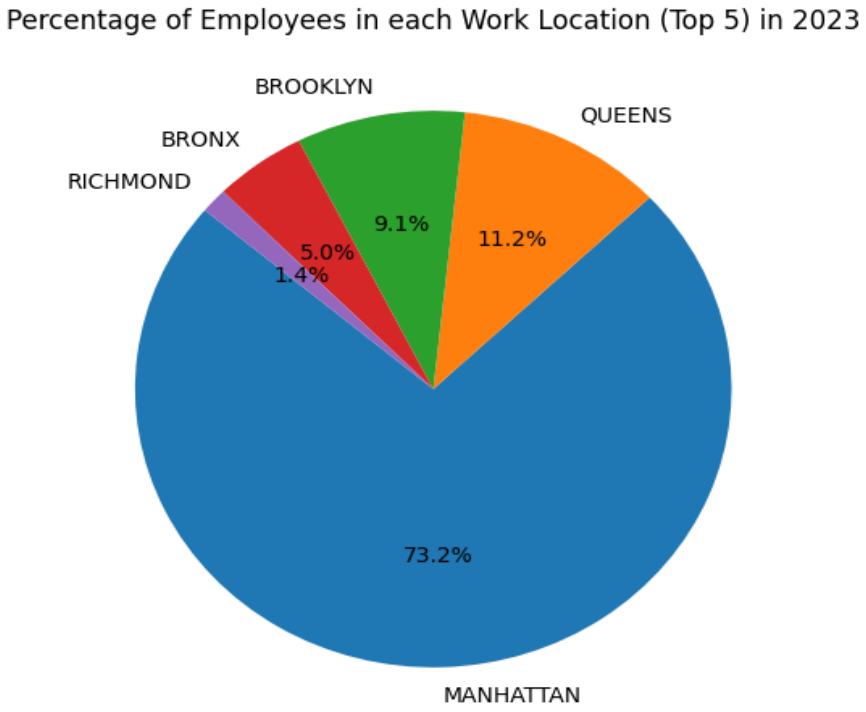
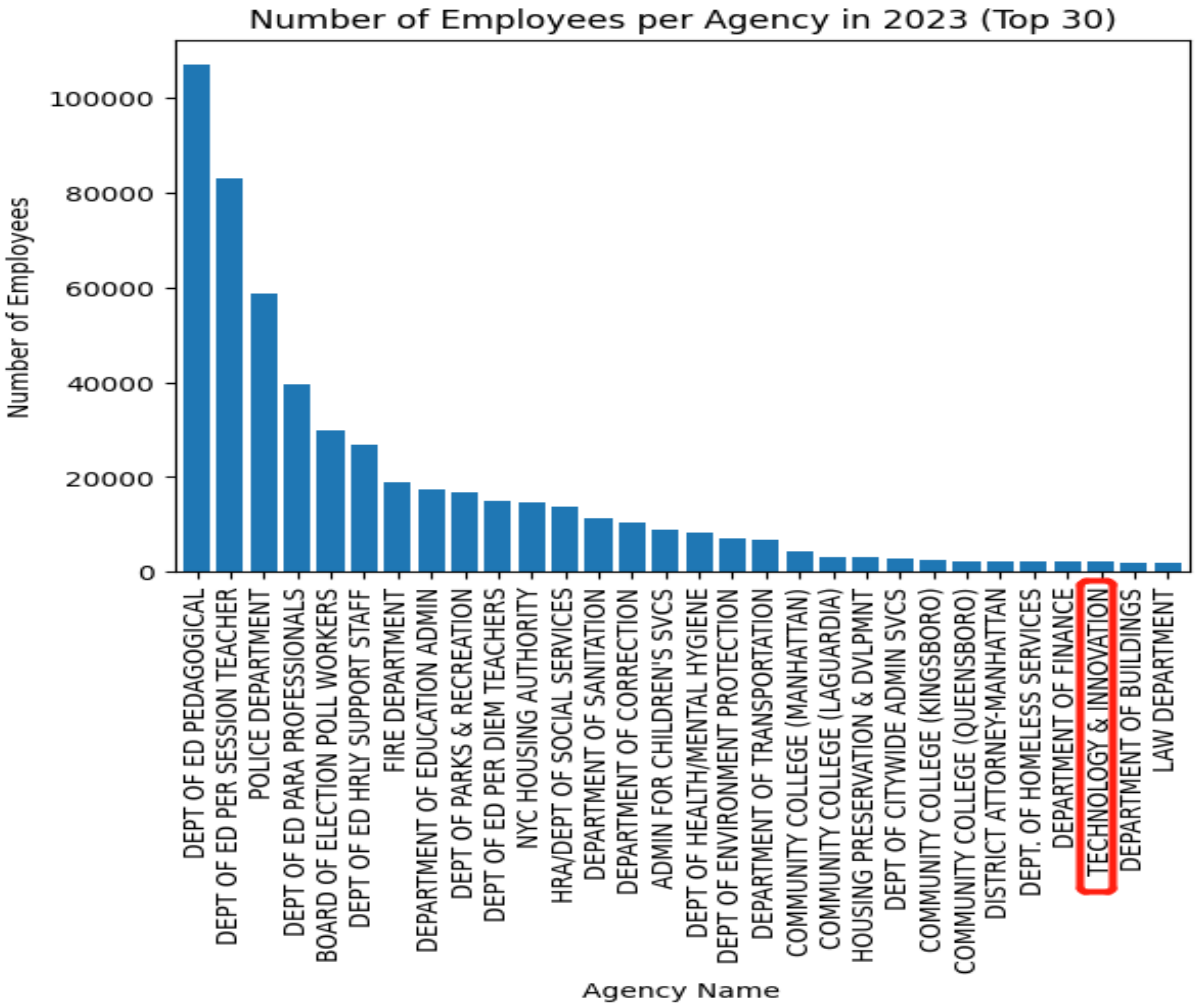


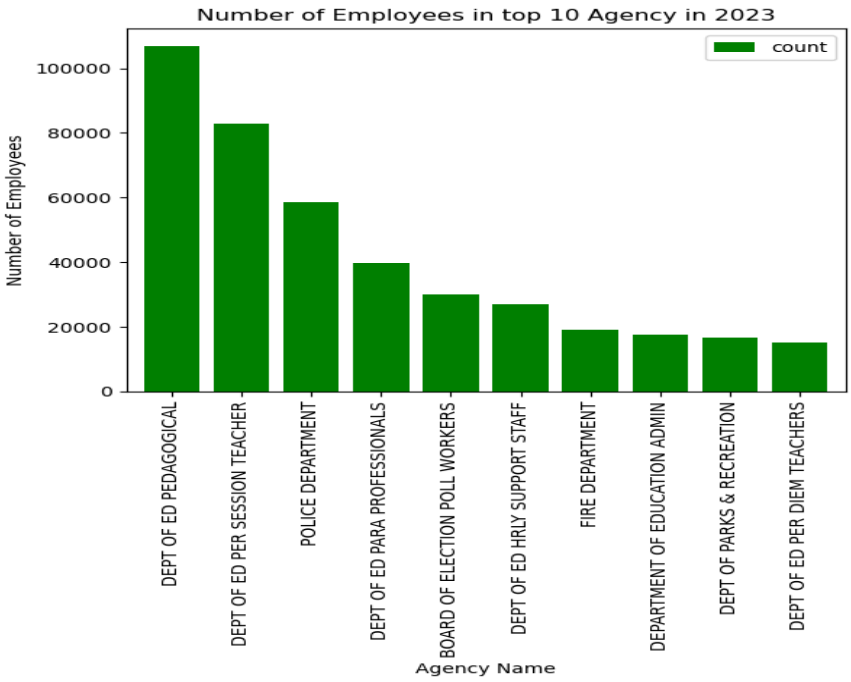
Exhibit 1 (Left)

The department's importance in governmental operations, especially in the digital era, presents an opportunity to glean insights into the evolving dynamics of technology roles in government, making it a focal point of our study despite not being the largest department. This strategic choice underscores our aim to explore areas of high relevance and potential impact.

Exhibit 2



In continuation to the EDA to dig deeper into the research questions, we narrowed down the scope by taking the first 10 topmost employed departments. Which is not inclusive of 'Technology & Innovation' but fields like 'Department of ED Pedagogical', 'Department of ED per session Teacher', 'Election poll workers', 'Fire department', 'Police department and more. This departments are named as agency names which includes specific departments listed by the NYC government. The top 10 ranking is based on the associated employees over the years from 2014 to 2023. Number of employees is inclusive of active and not active leave status. Looking at Exhibit 3, gives better idea about top 10 departments to be considered for further research work. As seen

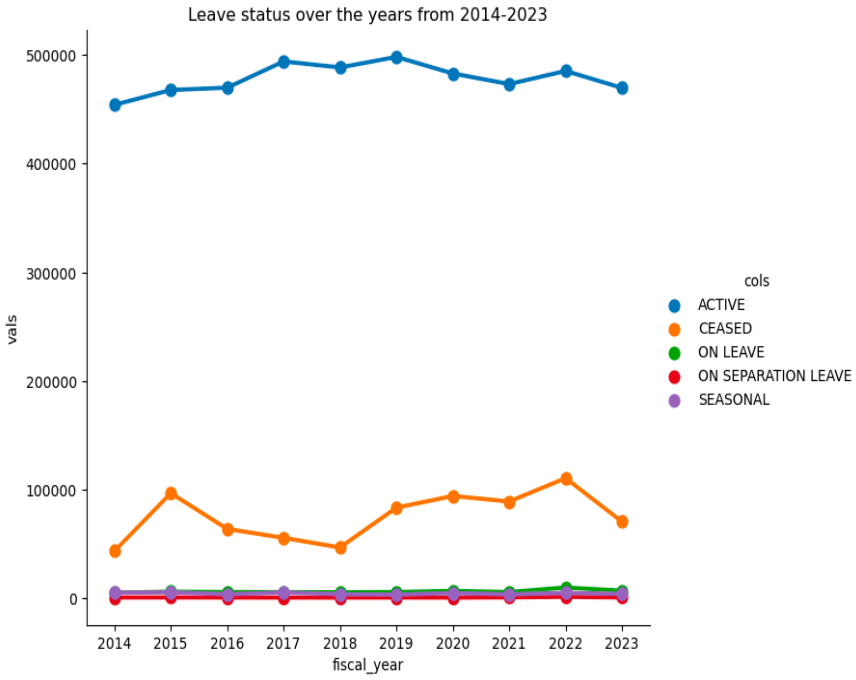


into exhibit 1, top 5 work locations include Manhattan, Queens, Brooklyn, Bronx, and Richmond. For the non-tech departments, the first two locations will be considered being Manhattan and Queens.

Exhibit 3 (Left)

Grouping the leave status with the yearly leaves below to see number of people active or on leave each year. Looking at Exhibit 4, it can be seen that above mentioned departments have 454,121 active employees which had raised up till 2019. After which, number of people on leave is also seems to be increasing. The 'On separation leave' and 'Seasonal' leaves seems to be constant. However, the 'Ceased' account seems to have increased from 2018 and seems to be going down from 2022.

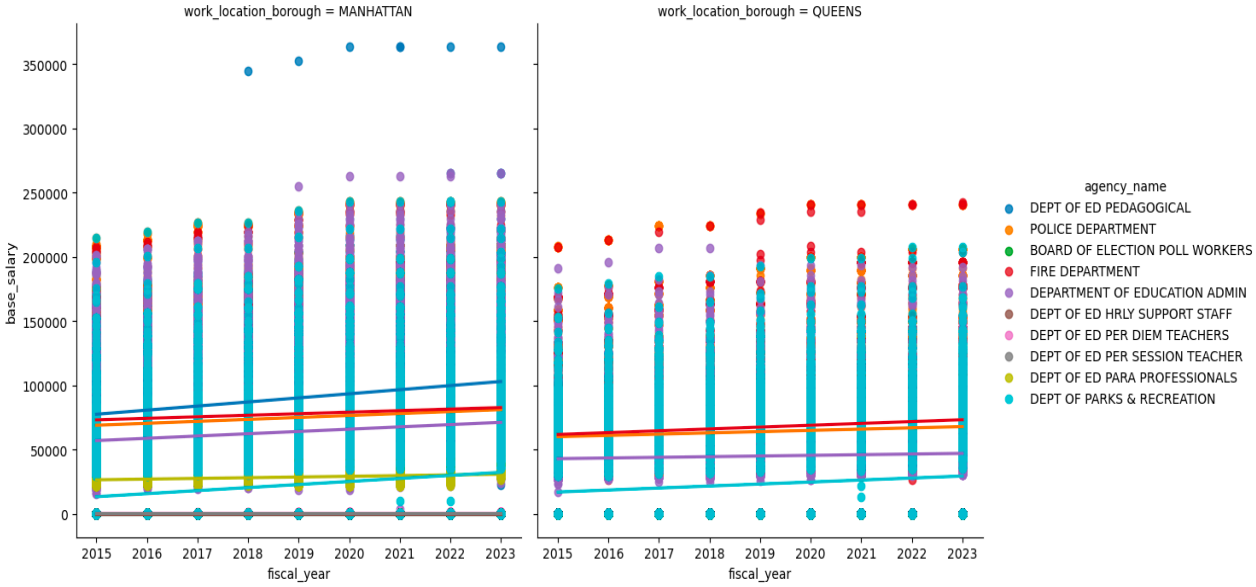
Exhibit 4



From the observations, it can also be seen that the fiscal years 2015 and 2022 had highest ceased accounts which could be referring to any natural big events to happen. The increase from 2018 up till 2022 could be an impact from Pandemic and organizations laying people home due to work-from home policies. In Exhibit 5 below, we can see the relation between fiscal

year, base salary, work location and departments. In the legend, different colors can be seen representing different departments in both the locations. Manhattan seems to be having more employees in the Department of ED Pedagogical from year 2018 onwards whereas in Queens, the scenario is not the same. In queens, agency names like ‘Fire Department’ and ‘Police Department’ seems to be leading. Followed by ‘Department of Education Admin’ which seems to be in top in both the locations.

Exhibit 5



III. Data Cleaning Process

Data Type Corrections:

We adjusted the data types of all columns to ensure consistency and accuracy in the dataset.

Handling Missing Values:

Approximately 50% of missing values in the ‘work_location_borough’ columns were filled with 'MANHATTAN' and the remaining with 'BROOKLYN', based on their equal distribution.

For the top 10 department, with the help of function ‘.isnull’ the null values in each column were found in which columns payroll_number, last_name, first_name, mid_init, work_location_borough and title_description seem to have missing values. After converting the data type of the relevant columns, the irrelevant columns were dropped. Which included detail like Employee name. For important columns like agency start date which was spliced into three columns for days, months and years, was replaced with mean values. Missing values in new columns created was also replaced with mean value.

Cleaning 'regular_hours' Column:

Abnormal data in the regular_hours column for the ‘Technology & Innovation’ department, including zero and negative values, were replaced with the average of other values to rectify inconsistencies.

Standardizing Base Salary:

To facilitate analysis, the base salary was standardized into two new columns: base_salary_hourly and base_salary_annual for ‘Technology & Innovation’ department as there were not daily pay provided. For the Other departments, the data way standardizes as annual from the hourly, daily, and annual categories mentioned in ‘pay_basis’ column.

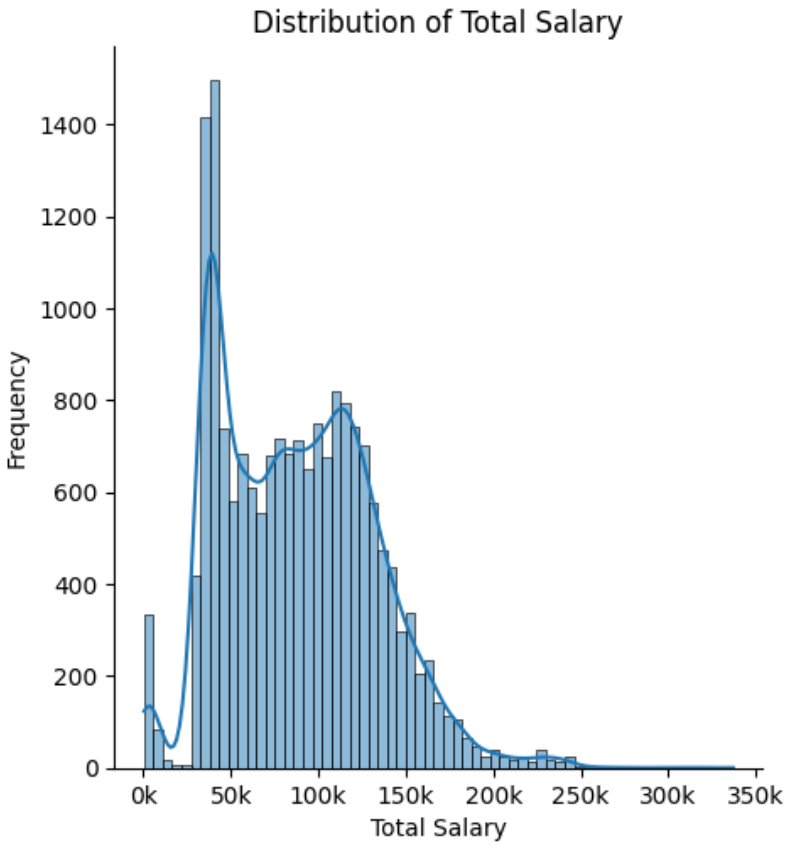
Creation of New Variables:

Three additional variables were created to enhance the depth of our analysis. ‘total_salary’, combining base salary with overtime and other pay for a complete picture of compensation; ‘work_year’, calculating employees' tenure to the fiscal year-end for insights into career duration; and ‘job_type’, categorizing job titles into groups like 'MANAGER', 'ENGINEER', ‘INTERN’, ‘ANALYST’, and counselor as ‘COUNSEL’, enabling detailed analysis of workforce segmentation and salary structures.

IV. Data Analysis & Insights

Following the meticulous data cleaning and filtering process, the New York City payroll dataset was distilled to a focused subset containing 18,157 records across 23 columns. This refined dataset served as the foundation for a series of insightful visualizations that illuminated the complexities of the technology department's compensation structure.

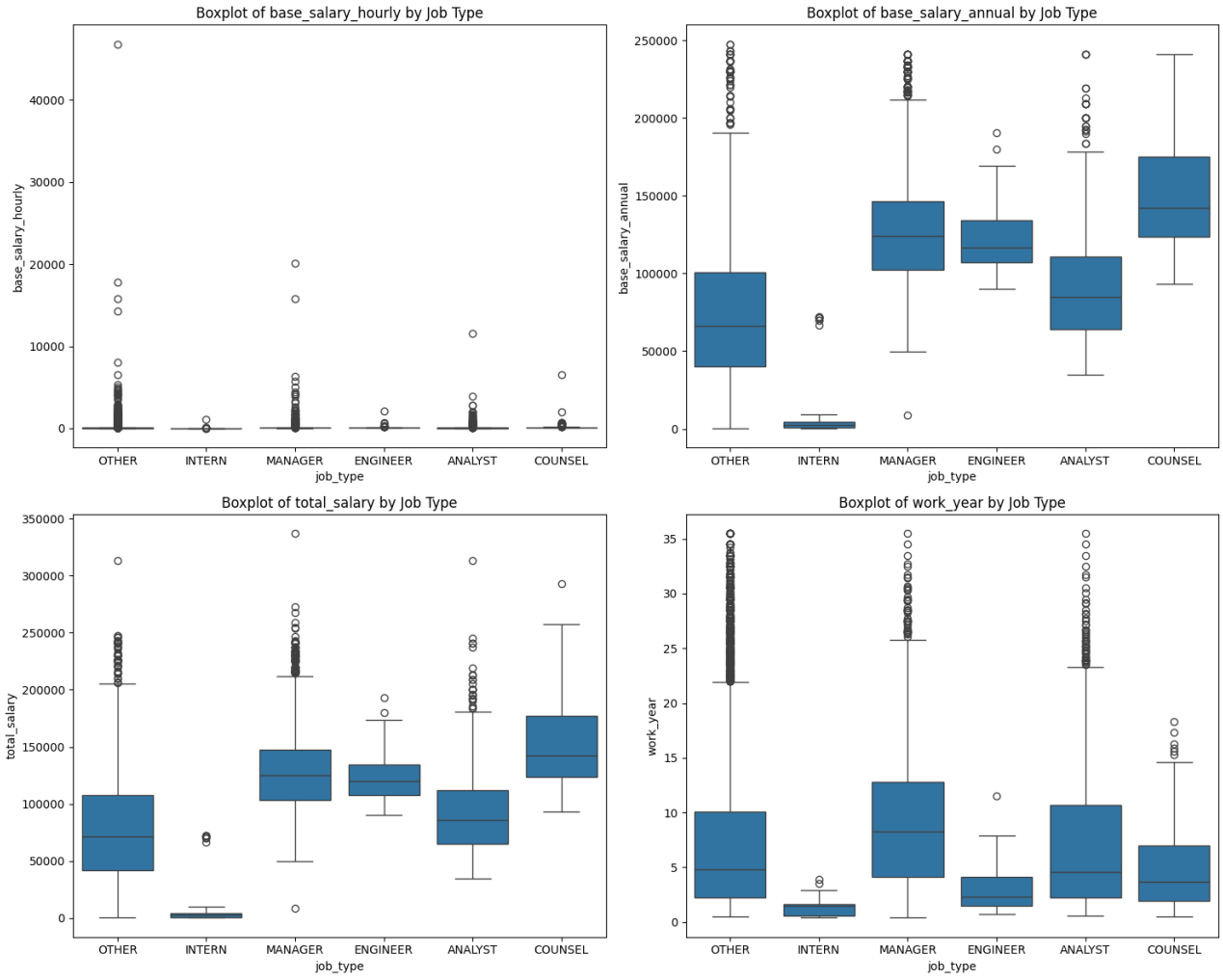
Exhibit 6



The histogram displaying the distribution of total salaries within the New York City Technology Department suggests a bimodal distribution, as indicated by the presence of two peaks. Such a distribution could imply that there are two prevalent salary groups within the department. The lower peak may represent entry-level or lower-paying positions, which are typically more abundant, while the second peak could correspond to a concentration of mid-range salaries often associated with mid-career professionals. The variance between these peaks could be indicative of distinct salary bands or clusters within the organizational structure, such as a separation between junior staff and more experienced or specialized roles.

The histogram also revealed a skewed distribution, with a high frequency of employees earning in the lower salary brackets and a long tail extending into higher earnings. This indicates a concentration of employees in lower salary ranges and fewer individuals earning at the upper end, which is typical in large organizations.

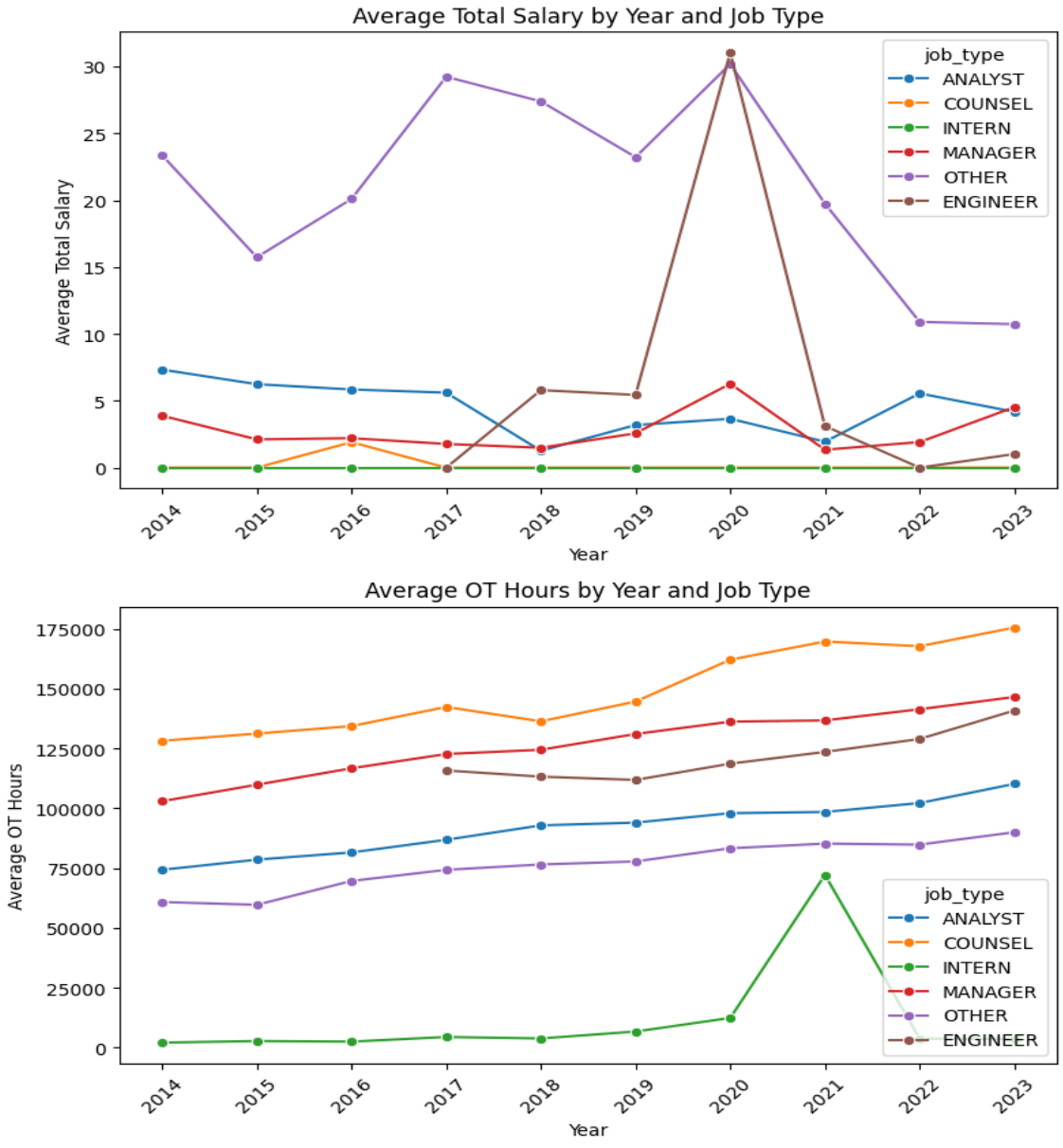
Exhibit 7



The box plots segmented by job type offered a clear depiction of salary dispersion and variance within job categories. Notably, roles such as 'MANAGER' and 'ANALYST' displayed a wider interquartile range, indicating a more significant spread in the salaries within these categories. Outliers, as visualized by points beyond the whiskers of the plots, suggest that certain job titles may include roles or responsibilities that command unusually high or low salaries compared to the median of the group.

Additionally, the box plots of base salary, both hourly and annual, contrasted with the total salary plot, gave insights into the composition of total earnings, distinguishing between regular pay and additional compensation such as overtime while the plot of 'work_year' by job type illuminated the distribution of tenure across different roles, which can be a proxy for experience levels within the department.

Exhibit 8



The line charts presented offer a compelling visual narrative of the evolving landscape of the New York City Technology Department's workforce over a near-decade span. The first chart displays the average overtime (OT) hours by year and job type, revealing significant variations

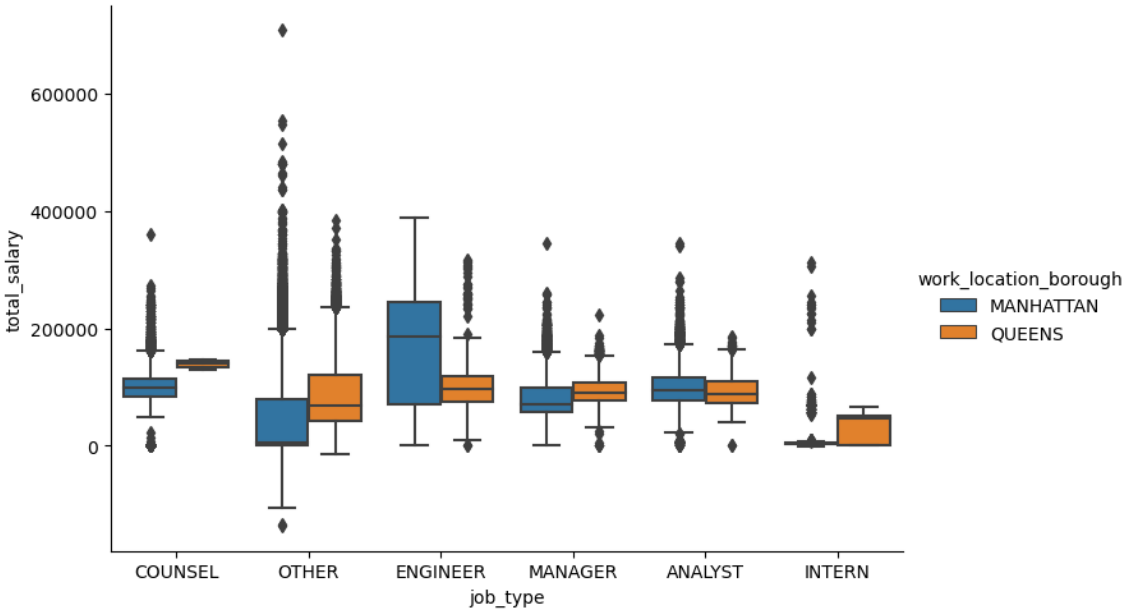
and trends in overtime allocation among different roles. A noticeable peak in OT hours for certain job types, such as 'ENGINEER' in 2021, may indicate specific departmental projects or deadlines that required extensive additional work hours. However, this spike is followed by a sharp decrease in 2022, which could suggest the completion of those projects or an adjustment in workload distribution.

The second line chart illustrating the average total salary by year and job type captured the trends over time, showing how average compensation has evolved from 2014 to 2023. Each job type followed a unique trajectory, with some showing steady growth while others, such as 'INTERN', depicted volatility, which may reflect changes in internship programs or data anomalies.

Together, all these visualizations paint a detailed picture of the Technology Department's payroll dynamics, providing a crucial backdrop for further analysis and modeling in the pursuit of understanding the factors influencing salary and employment within the department.

After finishing up the base analysis for the technology department, we wanted to check the similar analysis based on job types for the top 10 departments. To check if different category of departments also has the similar category of roles. To perform the same analysis, a catplot with box was created for the top 2 cities with most employees.

Exhibit 9



As shown above in Exhibit 9, the salary for counselor seems to be lesser in Queens over Manhattan which can also be seen in Exhibit 8 for the tech department. It can be seen that the total Salary for Engineers are also higher in the non-tech departments located at Manhattan over Queens. Interns in Manhattan seems to be paid based on each department and no clustering can be seen. However, for queens there seem to be more interns falling into a categorized pay slab. Salary for manager appears to be ranging around 70 to 100K plus.

Exhibit 10

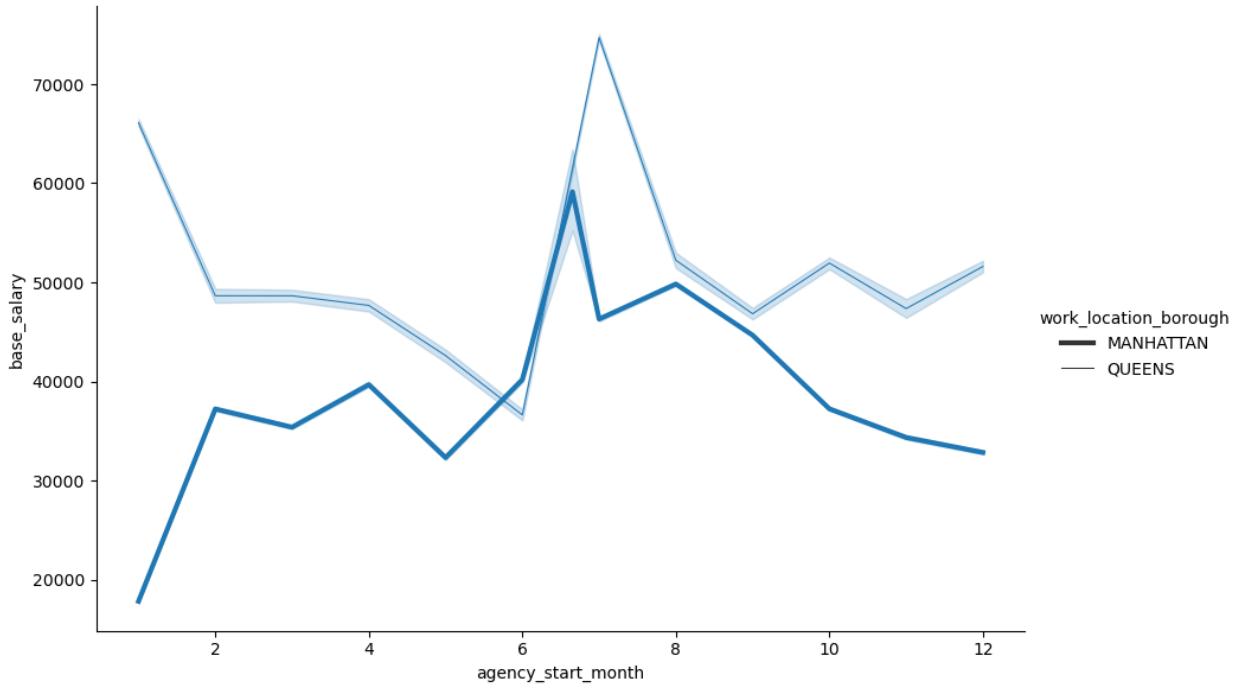
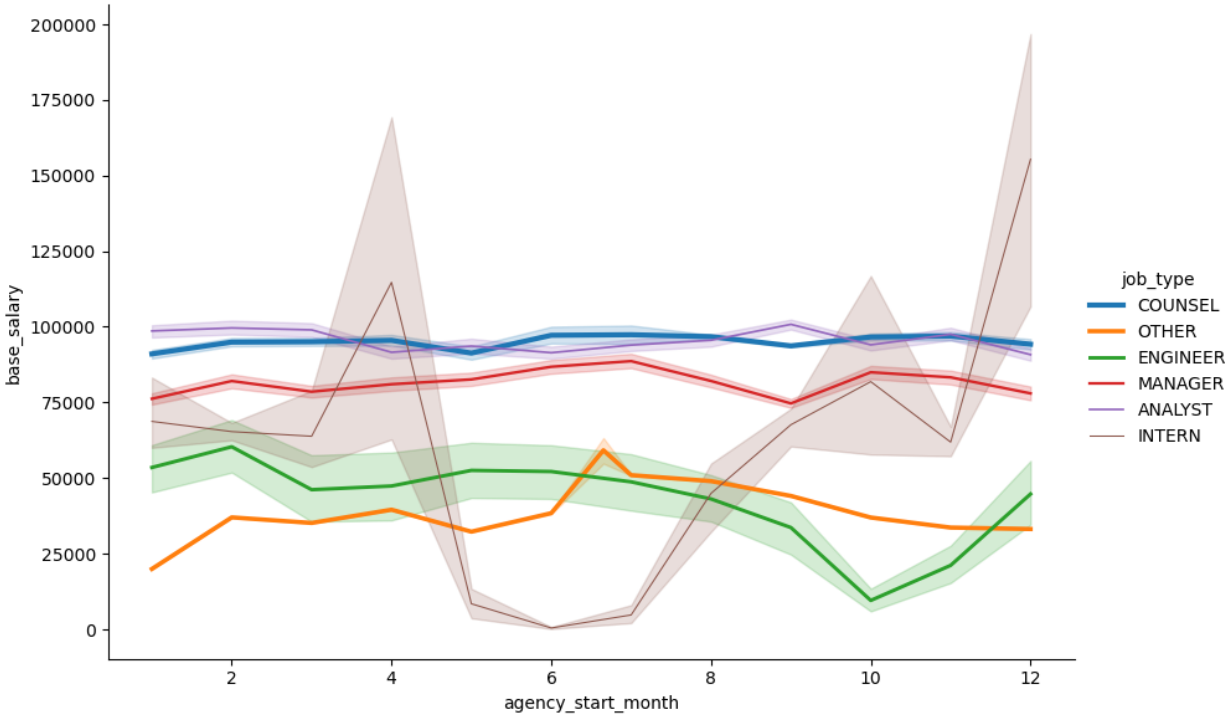


Exhibit 10 represents the co-relation between agency start month and base salary. We wanted to check if there is any correlation between the employee base salary and when they are joining on board. Here, from the graph we can see that there is a strong correlation between when the employee is being taken on board. Where we can see that an employee is being offered a better salary during the beginning of a fiscal year. From June to August the base salary is higher than March to June quarters. Ans so on an so forth, the base salary seems to be going down. Also, it can be observed that in terms of location-based payroll, Queens seems to be offering more than Manhattan payroll for the similar category of work.

Exhibit 11



The visualization in exhibit 11 represents various job type-based salary variation over the months. For the position of 'INTERN' we can see the most amount of variation. In the base pay salary during the end of fiscal year. By analyzing the graph, we can say that the role, industry, title, joining month have an impact on the base salary which could vary from quarterly report of the department.

V. Predictive Models

Linear Regression Model (Table 1, see Appendix)

In the predictive modeling phase of our analysis, the first model we employed is a linear regression which identified the determinants of total salary within the New York City Technology Department. Linear regression was deemed an appropriate method for this task due to its effectiveness in revealing relationships between a continuous target variable, such as salary, and multiple predictor variables. It also provides a clear interpretative framework to understand the impact of each feature on the target variable.

For instance, OT hours showed a positive coefficient of approximately 44.53, indicating that for each additional hour of overtime, we can expect an average increase of about $44.53 in total salary. This finding underscores the significant contribution of overtime to overall compensation, which may reflect policies or cultural practices within the department regarding overtime work.
Job type was another significant predictor. Compared to the baseline job type, Analyst, which is not shown due to one-hot encoding with "drop_first=True", roles such as Counsel were associated with the highest salary increase, while roles like Intern saw a negative association. This disparity illustrates the wide salary range across different job functions and possibly the varying levels of experience and responsibility associated with each role.

The negative coefficients for work location boroughs in Manhattan suggests that, relative to the baseline borough Brooklyn, Manhattan is associated with lower salaries, which could be attributed to cost-of-living adjustments or location-specific demands for tech talent.Furthermore, the pay_basis_per Hour variable revealed that hourly pay is associated with lower total salaries compared to annual pay bases, which might reflect the nature of employment (part-time or contractual work) typically associated with hourly wages.

The performance metrics, specifically the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE), were used to evaluate the model's accuracy. The RMSE for both the training and testing sets was around 31,600, providing a measure of the typical prediction error in dollars. The R-squared values indicated that approximately 48-49% of the variability in total salary was explained by the model, a moderate level of explanatory power which suggests that while the

model captures a significant proportion of the factors affecting salary, there is still unexplained variance that could be addressed with additional data or alternative modeling techniques.

In conclusion, the linear regression model provided valuable insights into salary determinants within the Technology Department, but also highlighted the complexity of salary dynamics, suggesting the potential utility of exploring more sophisticated models or feature engineering to improve the predictive power. The choice of linear regression as a starting point was justified by its interpretability and the straightforward manner in which it quantifies the relationship between salary and various predictors. However, the moderate R-squared values obtained from the model imply that further model refinement or the inclusion of more predictive features could enhance our understanding of salary determinants in future analyses.

Logistic Regression Model **(Table 2, see Appendix)**

For the predictive modeling of top 10 departments/agencies, we generated logistic regression model with an intension of performing binary classification. With a target variable as 'salary_category' which includes categorization of 'Low' and 'High' considering the median value of total salary. Since the total salary was a continuous data, we moved ahead to divide the data in to two categories to check if the model will be able to predict the over the bar salary. Or if the salary is low than the usual. In the research context, we can say that the model will help in defining if majority employees fall into which segmentation of the distribution.

The insights will generate an understanding of various job titles, having different roles under specific agency are being under paid or more than the general salary. Based on the model accuracy being 83.94%, the model will successful predict true positive and false negatives. Based on the F1-score represents the performance of classification model, that shows the true positive values. The score usually ranges between 0 to 1 which it is in this case as 0.84. Precision rate which is known as P would help in knowing if the model is successfully showing the false positive and not showing the positive as negative.

VI. Conclusion

We were able to get the insights for various variables having an impact on the total salary and the base salary of employees located at New York City. It can be states that the season of hiring also have an impact on employees from different department working at different locations. Manhattan and Queens have the greatest number of employees working across all the agency names. Comparing the Technology department which is not where the most employees work in NYC the wage disparities can also be seen. Looking at the various most popular job types with title as 'Manager', 'Engineer, 'Counselor' and 'Analyst' are comparatively have more 'Total salary' than the people with similar role in different departments. However, there are non-tech jobs which are highly paid but are not mainstream and considered into 'Other' category. It can be found that job type 'Intern' has the most volatile pay in both IT & non-tech agencies. Queens seems to be having a better pay for interns and non-tech job types. The base salary is higher for the roles 'Manager', 'Engineer' and 'Counselor' in NYC with top 5 locations (Exhibit 1).

Based on the chart in exhibit 6, the most frequent salary range for tech department is from 50 to 150K per annum. Employees working overtime seems to be having higher compensation with the total salary.

VII. Appendix

Table 1: Coefficients of Linear Regression Model

Regular Hours	**7.42**	**Job Type:**	
OT Hours	**44.53**	**COUNSEL**	**64393.13**
Work Year	**154.02**	**ENGINEER**	**11515.6**
Fiscal Year:		**INTERN**	**-40694.31**
2015	**1988.97**	**MANAGER**	**24631.87**
2016	**8484.51**	**OTHER**	**-16991.21**
2017	**12431.7**	**Work Location Borough:**	
2018	**14024.44**	**MANHATTAN**	**-29278.57**
2019	**18900.1**	**Pay Basis:**	
2020	**25098.84**	**per Hour**	**-42900.78**
2021	**24518.03**		
2022	**26393.41**		
2023	**30132.34**		

Table 2: Accuracy of Logistic Regression Model

Accuracy:	**0.8394000307862258**
Confusion Matrix:	**[[280525 , 50503]** **[55916 , 275690]]**
Classification Report:	**precision recall f1-score support** **High 0.83 0.85 0.84 331028** **Low 0.85 0.83 0.84 331606** **accuracy 0.84 662634** **macro avg 0.84 0.84 0.84 662634** **weighted avg 0.84 0.84 0 .84 662634**

VIII. References

NYC Open Data. (2023). Citywide payroll data (Fiscal Year). https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e