New York City Payroll - Group 6:
Final Report

Examining trends in payroll funds allocation for employees of New York City

Nicole Cosmany
University of Colorado
Boulder
Boulder, CO USA
nico3601@colorado.edu

Ebrahim Azarisooreh
University of Colorado
Boulder
Boulder, CO USA
ebaz7868@colorado.edu

Abstract

Problem State/Motivation

Taxpayers in the United States fund a wide range of government services ranging from payroll to defense to infrastructure and more. Naturally, taxpayers want to know how and where their money is being spent. To ensure transparency and accountability, government organizations make this data available online for public viewing and use.

In this project, we are focusing specifically on New York City's payroll data. Our goal is to explore how the city allocates its payroll budget across various job titles and departments. By analyzing this data, we hope to gain a deeper understanding of budget allocation trends, identify patterns over time, and uncover insights into the relative financial priorities of various departments and roles within the city.

Given the large number of job titles and departments in addition to the presence of some inconsistencies in what data is available for each, we are going to choose certain popular departments and job titles and focus on those for our investigation - for example Fire Department, Police Department, Department of Education, etc.. These departments are recognizable by the general public and also tend to be big spenders.

We propose the following questions:

1. **Allocation:** How is the total NYC payroll budget allocated among departments? Among boroughs?
2. **Pay**: How does pay for a certain job title vary from year to year? Does the average pay for a certain job title vary noticeably across boroughs?
3. **Overtime**: How much of the total payroll spending is for overtime? Are certain departments consistently spending more on overtime? Is there a relationship between the number of employees in a department and overtime spending? Are certain job titles working overtime more often than others?
4. **Duration of Employment:**

   How does pay rate/salary relate to employment duration? Do employees who are paid more/less have longer/shorter tenures than others? How does this vary across boroughs, departments, and job title?

Ultimately, this analysis will reveal key insights into how government resources are prioritized, identify how pay rate is affected by time and place, and the impact of overtime spending on the overall budget. Our results could be relevant to a wide range of stakeholders including New York City taxpayers concerned with allocation, budget committees planning for future years,

hiring managers looking to address overtime concerns, or even city employees looking to understand what factors contribute to their pay rates.

Literature Survey

Overtime

New York City has been seeing consistent increases in overtime spending over the last decade but no adjustments have been made to the budget to account for this issue, thus the city is consistently surpassing its proposed budget year and after (1). This is problematic because it defeats the purpose of having a budget if it is constantly overspent. Police are generally the biggest overspenders. They had a period from 2015 - 2019 when overtime spending in the police department stabilized, but it has since continued to grow again(1), according to the article published in 2023. With our dataset ranging from 2014-2024 we expect to be able to confirm these previously observed trends. In an attempt to mediate overspending, overtime budgets were reduced but this has been shown to be ineffective as budgets continue to be surpassed(1).

(1)https://comptroller.nyc.gov/reports/overtime-overview/#:~:text=The%20FY%202022%20actual%20overtime,uniformed%20overtime%20in%20FY%202022

Wage Trends

According to Forbes, employees tend to be rewarded more often for changing jobs rather than remaining loyal. On the other hand, many employees have the nagging concern that changing jobs too often can reflect negatively on their resumes and perception of work history. We'd like to know if there's a relationship between base pay rate and employment duration.

(2)https://www.forbes.com/sites/cameronkeng/2014/06/22/employees-that-stay-in-companies-longer-than-2-years-get-paid-50-less/

Data Set:

We use a publicly available dataset found online on the City of New York website. The data can be viewed in their online portal or downloaded. The link is here: https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e/about_data

Note: Some data is omitted in certain departments due to confidentiality policies. We are working with what is available.

The dataset has 6.22 million entries, where each entry corresponds to a single employee's pay information for the given fiscal year. Each entry has 17 attributes, which we define below:

Fiscal Year - fiscal year - interval

Payroll Number - int - ordinal

Agency Name - string- nominal

Last Name - string - nominal

First Name - string - nominal

Middle Init - string - nominal

Agency Start Date - date - interval

Work Location Borough - string - nominal

Title Description - string - nominal

Leave Status as of June 30 - string - nominal

Base Salary - int - interval

Pay Basis - string - nominal

Regular Hours - int - interval

Regular Gross Paid - int - interval

OT Hours - int - interval

Total OT Paid - int - interval

Total Other Pay - int - interval

Main Techniques Applied

Data Cleaning and Preprocessing

Our first task was to explore the data and look for inconsistencies or irregularities that would need to be fixed in order to ensure a proper analysis.

Compensation Calculations: Not all pay information is presented as an annual salary in the original data set. We needed to be able to compare pay rates across employees so we derived an additional column to calculate salary for hourly or per diem employees using pay basis, regular pay, and regular pay hours. Additionally, some entries contained erroneous data such as negative hours or not tracking hours at all. In those cases, we replaced the pay rate values with 0 and essentially ignored them. Considering the large volume of data, this small percentage will not greatly affect our overall analysis.

Reduce Redundancy: The original dataset contained some redundancy among department names, where certain departments are listed with sub departments or slight differences in names, such as all capitalized vs all lowercase. We aimed to rectify these by standardizing names across departments. This allowed us to both reduce inaccuracies caused by redundancies and avoid overly fine-detailed analysis so that we can capture the larger trends.

Analysis

1. Allocation

- calculate total spending across time period by year and borough to gain general understanding of data and magnitude
- sum total compensation, group by borough => identify top spending boroughs
- sum total compensation, group by department => identify top spending departments

2. Pay

- Pick a diversified set of job titles out of the larger set: Certified Applications Developer, Physical Therapist, Administrative Assistant, Assistant Professor, Police Officer, Executive Director, Construction Project Manager, Attorney at Law, Research Scientist, and School Lunch Helper
- calculate average salary per job title
- calculate average salary per job title and group by borough
- line chart showing average salary vs. time for chosen job titles
- bar chart showing average salary per borough for chosen job titles
- Attempt to build a k-means clustering model based on different pay-grades using work location, years of experience, and hourly pay-rate as vector dimensions.
- Attempt to use the model not only to hypothesize 3 or 4 different classes of a job, but use it to impute salary when missing.

3. Overtime

- Determine average number of OT hours worked by job over the years, compare to regular with a bar chart
- use mean + standard deviation to calculate coefficient of variation for departments overtime hours
- Graph percent OT hours worked by department
- Use linear regression model to determine relationship between number of employees and overtime spending

- add additional variable (department) to see how accurately linear regression model can predict overtime worked
- list top job titles working overtime hours

4. Duration of Employment

- Calculate average employment duration by job
- Use calculated salary average for job titles and graph them by employment duration
- Graph employment duration by job-title and salary and determine if a correlation exists on any of these item-pairs.

Visualization

To visualize our results, we used built in python libraries like matplotlib and seaborn. We used line graphs to represent our time series data and bar charts to compare values. We also output sorted tables to show ranked results when applicable.

Evaluation Methods

We use a combination of visual and statistical analysis. Since many of our questions concern change over time, we use line graphs. We use bar charts to look at aggregated sums of yearly spending, department spending, etc. and show comparisons. These figures will allow us to draw conclusions and compare our results to expected results obtained in previous studies.

For statistical measures, we use the coefficient of variation to understand how the top overtime spending departments are varying over the time period. We also use a regression model and the R squared value to quantify the model's prediction results. We use the regression coefficients to understand the contribution of different variables to the prediction.

Key Results

1. Allocation

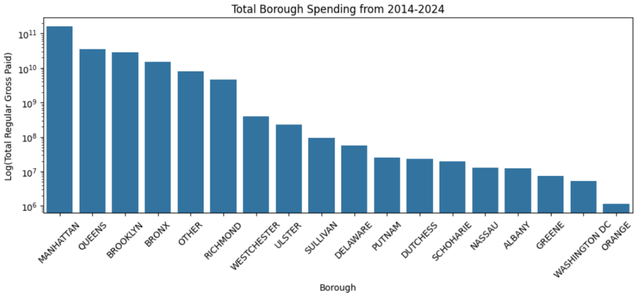


Figure 1. Total borough spending on regular gross pay from 2014 - 2024, including all departments

Figure 1 shows the total spending by each borough aggregated over the 10 year time period. We can see from these results that Manhattan, Queens, and Brooklyn take the lead as the top three spenders. Given the scale of the differences between boroughs, we use a log transformed y axis to maintain the relationships between boroughs while still being able to visualize them all on one graph.

This gives us an initial idea of the kind of relationships that exist among boroughs. There is a large difference between the top spenders and the lowest spenders, and even a noticeable difference between the highest spender, Manhattan, and the second highest, Queens.

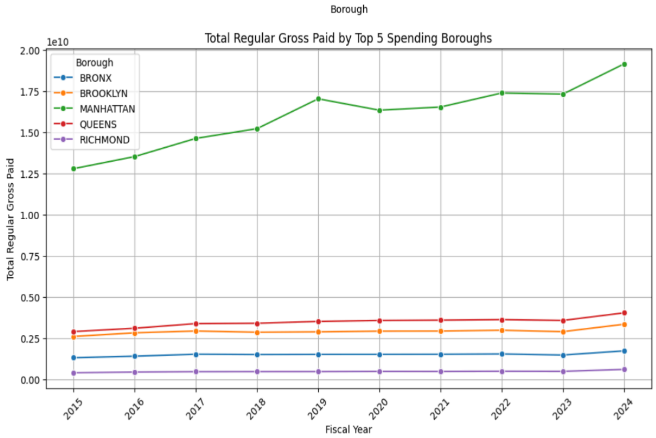


Figure 2. Time series graph depicting the change in the top 5 boroughs' spending on regular gross pay over the 10 year period

In Figure 2, we get a deeper understanding of the trends in borough spending by looking at spending across the full 10 year time period. This figure focuses on the top 5 spenders over the 10 year period - Bronx, Brooklyn, Manhattan, Queens, and Richmond. We can see from the graph that spending is slowly growing across all boroughs, though Manhattan far exceeds the others in magnitude of dollars being spent. It is both spending more dollars every year and its spending is growing at a faster rate than any other borough.

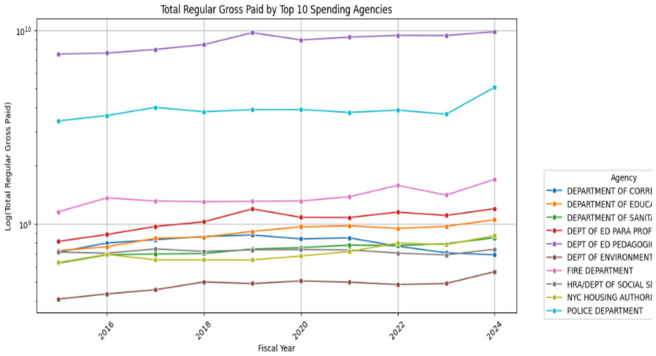


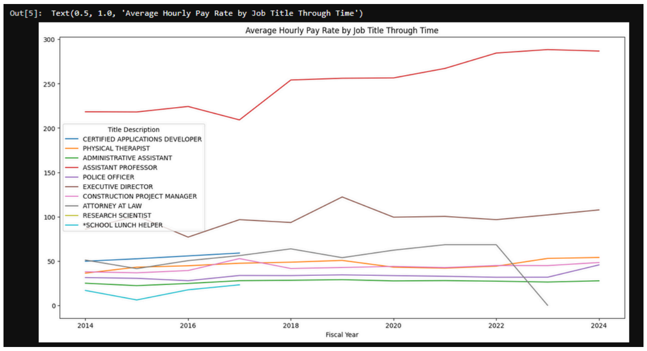
Figure 3. Time series graph depicting the change in the top 10 agencies spending on regular gross pay over the 10 year time period

In Figure 3, we now look at trends in department spending over the 10 year time period. Since there are far too many agencies to show them all, we identified the top 10 biggest spenders and focused on those here. We see that the top spender by a large margin is the Department of Ed Pedagogical. This is followed by the Police Department and the Fire Department. The fourth highest spender is another education related department - Department of Education ParaProfessional. This makes education-related spending by far the top user of payroll dollars.

Most departments are showing consistent growth over the ten year period, with small variations here and there. This makes sense given the overall increase in spending observed in Figure 2. Out of this group, no department seems to be growing noticeably faster than the others. From this, we can conclude that the top spending departments do not see dramatic changes in payroll budget from year to year.

Overall, these three figures provide an introduction to understanding our dataset and how payroll funds are allocated across boroughs and departments. We now have an idea of some of the top spenders and the magnitude of money that they are spending compared to others. One main trend observed across these graphs is the increase in spending over the last ten years. This increase is consistent though and there do not appear to be any sharp changes. A direction for further research could be to investigate what is causing this increase - likely due to inflation but potentially other factors like a growing workforce to serve growing populations in NYC.

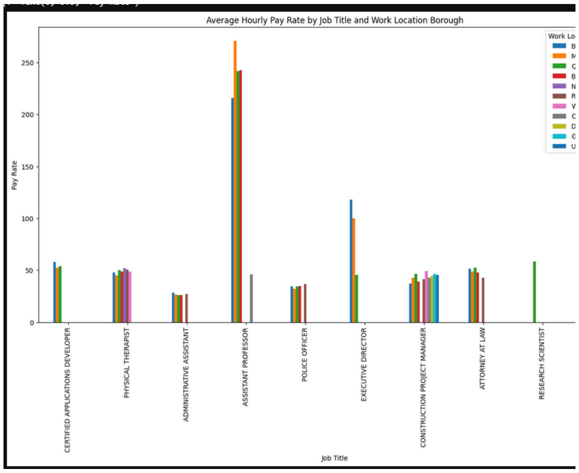
2.



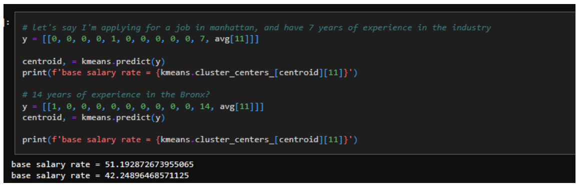
```
Out[5]:  Text(0.5, 1.0, 'Average Hourly Pay Rate by Job Title Through Time')
```

An overview of compensation across 10 diverse government allocated jobs in NYC, shows salaries mostly bundled in the 30 - 60 an hour range, with a noticeable departure from that grouping at above $100/hr for the Executive Director position.The Attorney position has a steep drop off to 0 around 2023, but appears to be an anomaly where there was actually no pay to anybody in this position during that fiscal year.The most notable outlier in this selected group of titles is that of "Assistant Professor", which ranges from 230 to 280 an hour, which is a clear outlier for the entire selected group.

The overall trend throughout the years is trending generally upwards for all positions. Most are inline with expected US inflation rates of ~2.5 - 3.5% per annum. The Certified Applications Developer was notably lower at around 1.6% percent increase, while the School Lunch Helper was showing ~11% increase per year, though was the lowest paying job by far in this group. All pay increase rates were measured using the geometric means to account for compounding rates.



Another view of the data is presented above, where each of the selected jobs are grouped by borough. There is some noticeable variation in pay-rates, some jobs exhibit more than others. This led to the hypothesis that perhaps work location borough, years of experience, and salary/pay-rate can be used to learn clusters of different pay-grades.



```
# let's say I'm applying for a job in manhattan, and have 7 years of experience in the industry
y = [[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 7, avg[11]]]

centroid, = kmeans.predict(y)
print(f'base salary rate = {kmeans.cluster_centers_[centroid][11]}')

# 14 years of experience in the Bronx?
y = [[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 14, avg[11]]]
centroid, = kmeans.predict(y)

print(f'base salary rate = {kmeans.cluster_centers_[centroid][11]}')
```

```
base salary rate = 51.192872673955065
base salary rate = 42.24896468571125
```

As can be seen in the figure above, an initial attempt was made to build a k-means model to classify 4 different pay grades for the Construction Project Manager job. k = 3 clusters was assumed on the basis of a junior/mid-level/senior stratification level of pay clusters, plus an additional to account for a second layer of stratification in the mid-level. The final cluster assumption was k=4. As can be seen from a trial run, using an average method on the salary component of the data vector, we attempt to aggregate the existing values on the 2-D plane, while using the imputed missing salary component and calculate the closest euclidean distance to one of 4 clusters. In terms of results, while the strategy could be viable in general, the selected variables don't appear to be producing very consistent results. More exploration on sub-satisfactory results are discussed in part 4.

3. Overtime Spending

Overtime spending is a topic that gets a lot of attention because it can quickly blow payroll budgets over their limit. We saw in our literature investigation that the police department has a history of significant overtime spending. We will now present the results of our overtime spending analysis to compare to these previous studies about the police department and understand trends within the larger scope of overtime spending across NYC departments.

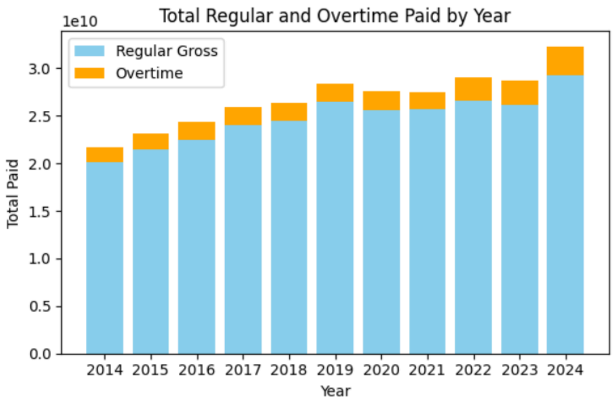
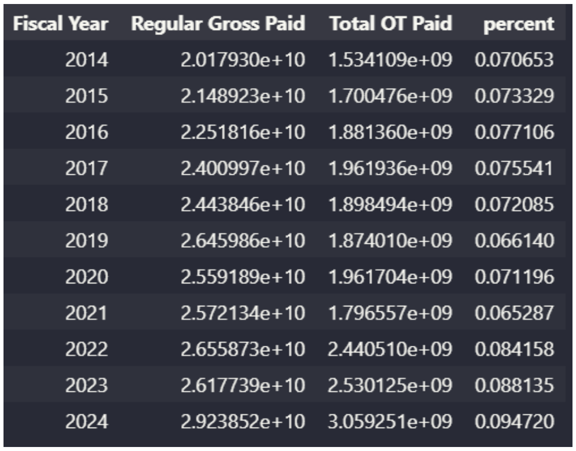


Figure 4. Bar graph showing the total regular and overtime pay across all boroughs and departments from 2014 - 2014

The first question we must answer - how much of the total NYC payroll budget is being spent on overtime? In Figure 4 we can see that the ratio of overtime to total pay stays relatively consistent across the 10 year time period. Additionally, we observe the same trend as before - increasing overall spending as time progresses. We can see that trend across both regular pay and overtime earnings.

Below in Figure 5 we see some concrete numbers explaining the difference in regular and total overtime. Overtime spending is hovering around 6-7% of the total budget until 2021. Then in 2022 onwards it increases to 8-9% of the budget. This is a considerable chunk of taxpayer dollars and it makes sense that overtime spending is a concern for the public.



Fiscal Year	Regular Gross Paid	Total OT Paid	percent
2014	2.017930e+10	1.534109e+09	0.070653
2015	2.148923e+10	1.700476e+09	0.073329
2016	2.251816e+10	1.881360e+09	0.077106
2017	2.400997e+10	1.961936e+09	0.075541
2018	2.443846e+10	1.898494e+09	0.072085
2019	2.645986e+10	1.874010e+09	0.066140
2020	2.559189e+10	1.961704e+09	0.071196
2021	2.572134e+10	1.796557e+09	0.065287
2022	2.655873e+10	2.440510e+09	0.084158
2023	2.617739e+10	2.530125e+09	0.088135
2024	2.923852e+10	3.059251e+09	0.094720

Figure 5. Table showing regular gross pay, overtime spending, and overtime/ total for 2014 - 2024

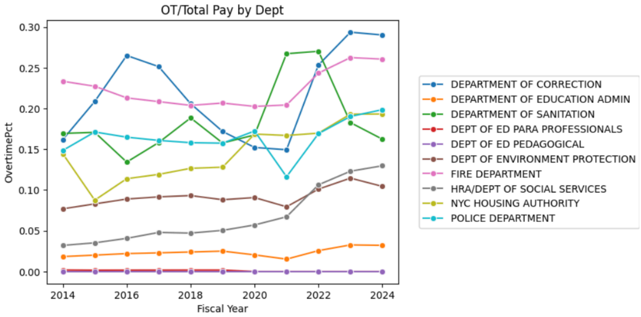


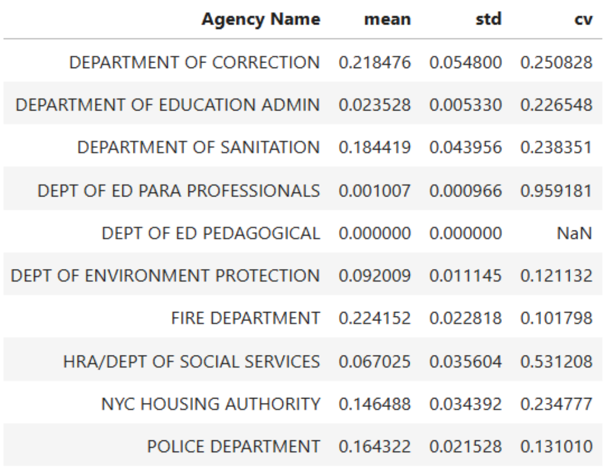
Figure 6. Percent overtime spending by the top 10 spending departments from 2014 - 2024

Now we look at overtime spending by department, focusing on the top 10 spenders from Part 1. Are certain departments consistently spending more than others?

Figure 6 uses percent overtime out of total (regular pay + overtime). We can see that the top overtime spenders - Fire Department, Department of Sanitation, Police Department, Department of Correction, and NYC Housing Authority have a decent amount of variation in their spending. In contrast, the lower spenders are showing slow growth year to year.

Out of the top spenders, we see some interesting patterns around 2020, the year of the COVID-19 pandemic. First, we see a spike in overtime hours worked by the Department of Sanitation. This makes a lot of sense given their role in public health makes them likely candidates for increased work during a pandemic. Second, we see a drop in Police Department overtime. This could be due to the fact that people were confined to their homes and there were naturally less incidents for the police to respond to. Third, the Department of Correction has a huge spike in overtime in 2021, the year following the pandemic. This could be an interesting point for further investigation - did the pandemic lead to increased crime rates in NYC?

To understand more about the variation in overtime spending within departments, we can look at Figure 7. This is another view of the information presented in Figure 6. The coefficient of variation column shows how much variation is within each department. We see noticeably high values in the Department of Ed ParaProfessional and the HRA/Dept of Social Services. These trends are obscured in Figure 6 due to scale, but further investigation revealed that these departments did see a fair amount of variation. This shows that overtime spending can vary greatly across spending levels and it is important to consider for all departments.



Agency Name	mean	std	cv
DEPARTMENT OF CORRECTION	0.218476	0.054800	0.250828
DEPARTMENT OF EDUCATION ADMIN	0.023528	0.005330	0.226548
DEPARTMENT OF SANITATION	0.184419	0.043956	0.238351
DEPT OF ED PARA PROFESSIONALS	0.001007	0.000966	0.959181
DEPT OF ED PEDAGOGICAL	0.000000	0.000000	NaN
DEPT OF ENVIRONMENT PROTECTION	0.092009	0.011145	0.121132
FIRE DEPARTMENT	0.224152	0.022818	0.101798
HRA/DEPT OF SOCIAL SERVICES	0.067025	0.035604	0.531208
NYC HOUSING AUTHORITY	0.146488	0.034392	0.234777
POLICE DEPARTMENT	0.164322	0.021528	0.131010

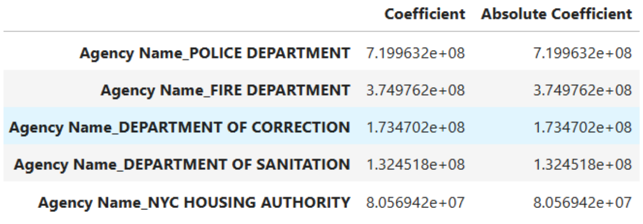
Figure 7. Chart showing mean, standard deviation, and coefficient of variation for department overtime spending

Given these previous findings, we conclude that overtime spending is an important part of the payroll budget. To better understand what was causing overtime spending, we split our data into train and test sets and then created three linear regression models to see how well we could predict over time spending and what factors were the main contributors. Our initial hypothesis was that the number of employees in a department would be a big factor.

In Model 1, we used number of employees as the sole predictor and the resulting R squared value was a low 0.17, suggesting that number of employees is not a strong predictor of overtime spending.

In Model 2, we one hot encoded the department attribute and used department name as the sole predictor. This resulted in a R squared value of 0.97, which is highly accurate and suggests that department is a very strong predictor of overtime spending.

In Model 3, we used both department name and number of employees as our predictor variables. This model yielded the same result as Model 2, 0.97. This suggested that the number of employees was not an effective predictor. Upon analyzing the coefficient values of the model, we confirmed this by determining that the number of employees was listed as 112 out of 115 variables.

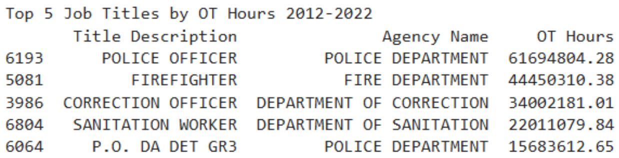


	Coefficient	Absolute Coefficient
Agency Name_POLICE DEPARTMENT	7.199632e+08	7.199632e+08
Agency Name_FIRE DEPARTMENT	3.749762e+08	3.749762e+08
Agency Name_DEPARTMENT OF CORRECTION	1.734702e+08	1.734702e+08
Agency Name_DEPARTMENT OF SANITATION	1.324518e+08	1.324518e+08
Agency Name_NYC HOUSING AUTHORITY	8.056942e+07	8.056942e+07

Figure 8. Chart showing the top 5 predictor variables in Model 3

In Figure 8, we can see the top 5 variables contributing to the model are the Police Department, Fire Department, Department of Correction, Department of Sanitation, and NYC Housing Authority. This tells us that these departments are predictable spenders.

Given this information and our previous findings, we confirm the previous study asserting that the police are notorious overtime spenders. The article mentioned that despite the consistent overtime, no significant budget alteration has been made to account for overtime spending, so the department constantly overspends its budget. This model confirms that the spending is predictable and the budget could benefit from an increased overtime budget.



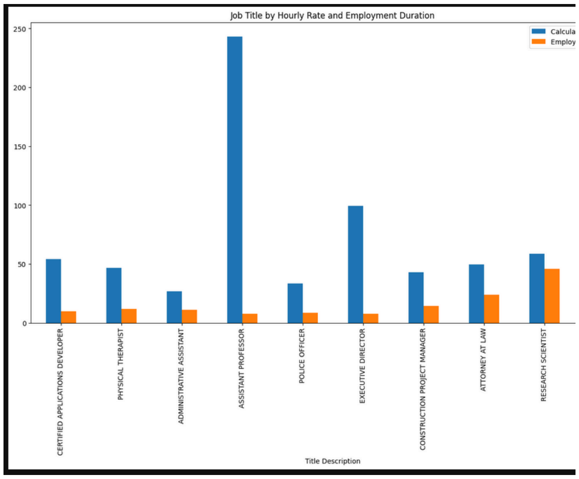
```
Top 5 Job Titles by OT Hours 2012-2022
       Title Description                 Agency Name     OT Hours
6193      POLICE OFFICER           POLICE DEPARTMENT  61694804.28
5081         FIREFIGHTER             FIRE DEPARTMENT  44450310.38
3986   CORRECTION OFFICER  DEPARTMENT OF CORRECTION  34002181.01
6804    SANITATION WORKER  DEPARTMENT OF SANITATION  22011079.84
6064       P.O. DA DET GR3          POLICE DEPARTMENT  15683612.65
```

Figure 9. Chart showing top job titles by number of overtime hours worked

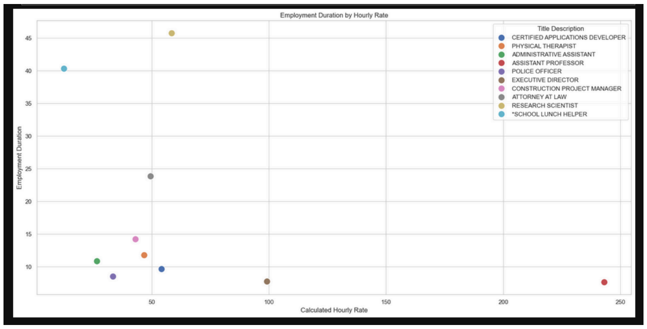
Finally, we complete our overtime investigation by identifying which job titles are working the most overtime hours. This tells us who, within the popular departments, is working more than their standard hours. The results are not very surprising, as the job titles are all belonging to the previously identified top overtime spenders. We see the top job title most likely to work overtime hours in the Police Officer, followed by Firefighter, and Correction Officer. Given the nature of these jobs and previously identified spending trends in these departments, this makes sense.

In summary, the overtime spending by NYC departments accounts for roughly 6-9% of total payroll spending over the past decade. It is heavily linked to certain job titles within key departments whose employees are consistently working overtime, with very little relation to the number of employees.

4.



When examining the compensation rates of each job juxtaposed with employment duration, there doesn't seem to be an obvious relationship here. For example, the lowest paying job (School Lunch Helper), had the longest tenure, and in second place was Research Scientist which was one of the higher paying jobs, but also had the second longest tenure of the group. Executive Director and Assistant Professor were the top 2 highest paying jobs in this group, but had fairly low tenures. Overall, It looks like this slice of the data (which is well diversified across sectors) doesn't show any promising trends in terms of a relationship between pay and employment duration.



A scatter plot of the same data, seems to reinforce the notion that there is no clear relationship between these two variables. The results here seem to indicate why my chosen representation of the k-means model ultimately did not work out.

Applications

Our Question 1 findings create a picture of the government spending on employee pay. We start by showing a general breakdown of how funds are allocated among boroughs and departments. This is informative for taxpayers who might be curious about where their dollars are being spent geographically and among what services. This could also be useful for budget forecasting.

In Question 2, an attempt was made to determine the interplay between not only compensation and employment duration, but a mediating variable in terms of work location.

Some variance was observed in the latter, but ultimately a k-means imputation method to determine a reasonable baseline for salary from these variables proved to be inconsistent, uninsightful, and somewhat mismatching in terms of the data. The strategy of using a k-means model to impute salary from other known variables can be effective in terms of knowing a reasonable cutoff value for a hopeful job applicant, thereby adding a data-to-decision negotiation model to help remove uncertainty and emotion from salary discussions. Nonetheless, the chosen vectors for this model proved to be unsatisfactory.

In Question 3, we explored overtime spending. This is useful information for budget planners who need to be aware that additional funds may need to be allocated for departments that are constantly spending on overtime. Similarly, hiring managers might want to know which positions are most commonly working overtime hours and use that information to make staffing changes.

In Question 4, we attempted a deeper dive on why the clustering model didn’t work effectively in terms of the partitions themselves, and ultimately using an imputation method to derive mean salary cutoffs in reverse. The scatterplot and grouping histograms presented in our analysis show that there isn’t a clear correlation between these variables, and that there is probably a deeper interplay determining how long one stays at a particular job. Ultimately, compensation does not seem to show any type of meaningful relationship with how long one stays employed at a position. As a matter of fact, one of the lowest paying jobs had the highest employment duration. Conversely, one of the higher paying jobs exhibited the same trait. The results of the analysis shows that a deeper dive on what anchors someone to a job can be helpful for employers in minimizing employee turnover.

In summary, our results are useful for a wide range of audiences - from government planners to curious taxpayers.