

## New York City Payroll - Project Proposal - Group 6:

### Updated Progress 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

#### 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, they 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.

*Update: Given the large number of job titles and departments and 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.*

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/>

### **Proposed Work**

We plan to address the questions raised in our Motivation Statement. To answer each question, we will perform the following calculations:

#### **1. Allocation**

- a. calculate total budget for each year by adding total salary
- b. sum total compensation, group by borough
- c. sum total compensation, group by department

#### **2. Pay**

- a. sort unique job titles and pick titles of interest (Police Officer, Firefighter, etc.)
- b. calculate average salary per job title
- c. calculate average salary per job title and group by borough
- d. line chart showing average salary vs. time for chosen job titles
- e. bar chart showing average salary per borough for chosen job titles

#### **3. Overtime**

- a. Determine average number of OT hours worked by job over the years
- b. Graph OT hours worked by job title.
- c. Determine if a relationship exists between OT hours worked over the years in the dataset.

#### **4. Duration of Employment**

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

Missing values: If missing salary values come up, we will impute by using an average of other employees in the same department with the same title in the same year. Most other missing values can be ignored - if we don't know the job

title or department, there's not much we can do about that besides remove it from our calculations. We do anticipate many, if any, missing values.

Compensation Calculations: Not all pay information is presented as an annual salary. We plan to analyze all pay as annual salaries. This will require calculating salary for hourly or per diem employees using pay basis, regular pay, regular hours, and additional compensation attributes.

*Update:* As we went through and calculated the normalized pay rate, we realized that there were a lot of issues with this data. It appears that some job titles did not collect any hourly information, while other had negative values in hourly and salary columns which does not make sense. We worked around this issue and were able to get the code to run by replacing any weird values with a 0 normalized pay rate. Considering that there are thousands of job titles, we will only end up picking about 5-10 job titles to focus on for questions 2 and 4 which require accurate pay rate information. This allows us to choose recognizing job titles (Police Officer, Accountant, etc.) and ensure we are working with accurate data.

Reduce Redundancy: There is currently some redundancy among department names, where certain departments are listed with sub departments or slight differences in names. We aim to rectify these by standardizing names across departments. We also plan to drop the Payroll Number column since this is another way of identifying departments.

### **Data Set:**

Available online at:

[https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e/about\\_data](https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e/about_data)

Additionally, both group members have the dataset downloaded.

Some data is omitted in certain departments due to confidentiality policies.

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

### **Evaluation Methods**

We plan to calculate correlation in order to quantify the relationships between the variables in our four questions. We can also use visual analysis of figures to evaluate trends at a high level. Since many of our questions concern change over time, we will be using line graphs. We can use scatter plots for our questions regarding correlation between two variables.

These figures will allow us to draw conclusions and compare our results to expected results obtained in previous studies.

## Tools

We plan to utilize the online data viewing tool to help with quick filtering and viewing tasks. Cleaning, filtering, and some summary statistics will be done in Prolog. We will also use Python (pandas, numpy, scipystats, matplotlib) to do calculations and generate results.

*Update* - We have updated our strategy and plan to only use python moving forward. One group member tested Prolog and decided it was not worth pursuing. The dataset loads within a few seconds as a pandas dataframe and it is quick and easy to manipulate. We will also likely end up using scipy to calculate some metrics (correlation, etc.) and matplotlib or seaborn for visualizations.

## Milestones

### Milestones Completed

Week 8: Cleaning data, deriving new columns, looking for inconsistencies in data

We renamed departments to reduce redundancy. For example in situations where we had Department A #1, Department A #2, Department A #3, etc. we renamed all to simply be Department A. This makes sense since we are concerned with a high level department analysis.

We also added a new column "Calculated Pay Rates" to standardize pay rates since they are presented in various formats (per annum, per hours, per diem, etc.). Now we can compare simple base pay rate without worrying about hours worked, pay basis, overtime, etc. for our proposed questions #2 and #4. This did raise a couple issues with infinities and missing values that we had to deal with.

Week 9: EDA + summary statistics + progress report

We familiarized ourselves with the data by counting and exploring the unique values in our more important columns that we will use to answer questions - borough, department, year, etc. We also did some basic budget calculations to gain an understanding of the scale of our data across departments and years. We include these below in our "Results so far".

We also had a meeting to check in on progress, troubleshoot some coding issues, discuss data inconsistencies we found, put together this progress report, and confirm our plan moving forward. We also started to commit code to our repository so everything is localized and accessible to both group members.

### Milestones To-Do

Most of our progress thus far focused on cleaning and preparing the data and making sure we could do the kinds of grouping and aggregating that will be necessary to answer our questions. Moving forward, we are going to start producing data, figures, and statistics to answer our questions.

Week 10: basic visualizations, start looking into connections and trends that will help us answer our questions

spring break: catch up as needed

Week 11: work on questions 1/2

Week 12: work on questions 3/4

Week 13: Put together final project report

## Results So Far

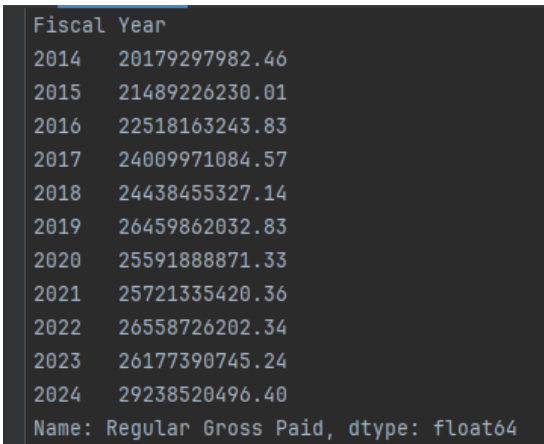
Most of our efforts so far have focused on cleaning the data and addressing missing values and unexpected negative values. We have started gathering some summary statistics in

order to build a robust understanding of our data set. We looked at # of years (2014 - 2024), department names, and job titles. Below we have done some calculations to start to understand budgets across years, departments, and boroughs, and normalized pay rate across job titles.

Regular payroll budget by year:

After some initial runs using Pandas, we were able to aggregate data for total budget expenditures for each of the fiscal years in the dataset (2014 - 2024). These numbers are a sum of the Regular Gross Paid column, so does not include overtime. We did this because we wanted to get a sense for how a normal payroll budget operates and we can later look at overtime information when we specifically ask our overtime question.

Looking at this initial data in this format, we can tell already that the budget is increasing year to year at an incremental rate. When we plot this as a figure, we should see a positively sloped line.



Fiscal Year	Regular Gross Paid
2014	20179297982.46
2015	21489226230.01
2016	22518163243.83
2017	24009971084.57
2018	24438455327.14
2019	26459862032.83
2020	25591888871.33
2021	25721335420.36
2022	26558726202.34
2023	26177390745.24
2024	29238520496.40

Name: Regular Gross Paid, dtype: float64

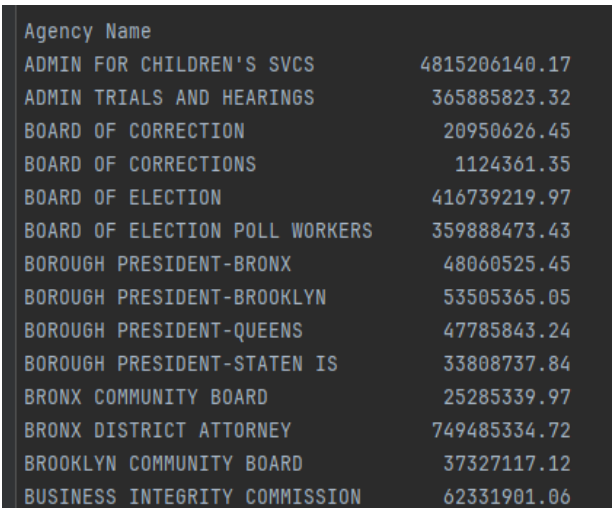
Screenshot showing regular payroll budget by year

Regular Payroll Budget by Agency:

We were also able to consolidate and remove a lot duplicate data (such as BRONX COMMUNITY BOARD #1, #2, #3, etc. and

similarly named community boards across all boroughs) and merge them into one entity for a more unified and compact view of the entities. After some initial data cleansing we then collected total budget expenditures by each city department using the Regular Gross Paid column.

Going forward, we can break this up by year and also use this information to focus on certain departments. For example, we can choose some top spending departments that are responsible for spending notable amounts of money over the past 10 years. It would be impractical to focus on every department since there are so many, so this will help us choose relevant ones.



Agency Name	Regular Gross Paid
ADMIN FOR CHILDREN'S SVCS	4815206140.17
ADMIN TRIALS AND HEARINGS	365885823.32
BOARD OF CORRECTION	20950626.45
BOARD OF CORRECTIONS	1124361.35
BOARD OF ELECTION	416739219.97
BOARD OF ELECTION POLL WORKERS	359888473.43
BOROUGH PRESIDENT-BRONX	48060525.45
BOROUGH PRESIDENT-BROOKLYN	53505365.05
BOROUGH PRESIDENT-QUEENS	47785843.24
BOROUGH PRESIDENT-STATEN IS	33808737.84
BRONX COMMUNITY BOARD	25285339.97
BRONX DISTRICT ATTORNEY	749485334.72
BROOKLYN COMMUNITY BOARD	37327117.12
BUSINESS INTEGRITY COMMISSION	62331901.06

Screenshot showing regular payroll budget by agency

Average Pay Rate by Job Title:

We collected average pay rate data by job title. This pay rate was obtained by dividing Total Regular Pay by hours (because not everyone works full time so we wanted an unbiased pay rate). After some more data cleansing pertaining to data objects that indicated working 0 hours (causing divide-by-zero errors), we then flattened these

values to a pay rate of 0 dollars per hour. This was accomplished by applying a lambda to the 'regular gross paid' dataframe, which transformed a 0 to a numpy infinity value. That way, when dividing by 0, we get 0. The presence of a 0 basically indicates that this data is not reliable so we will likely not use any job titles that have a 0. This is totally fine and in line with our goals because we just want to focus in on a few notable and recognizable job titles in order to answer our proposed questions.

As you can see from the screenshot, we also noticed a lot of encoding errors in this data-axis. Bad/questionable characters were present in many of these job titles descriptions. These should not theoretically pose a problem because we probably will not end up doing anything with those titles.

Title Description	
* ATTENDING DENTIST	0.00
*ADM DIR FLEET MAINT-MGRL ASGMT	33.89
*ADM DIR FLEET MAINTENANCE - NM	41.87
*ADM SCHOOL SECURITY MANAGER-U	25.24
*ADMIN SCHL SECUR MGR-MGL	69.02
*ADMINISTRATIVE ATTORNEY	89.62
*AGENCY ATTORNEY	64.72
*ASIST SYSTMS ANALYST	36.56
*ASSIST COORDINATING MANAGER	28.17
*ASSISTANT ADVOCATE-PD	57.37
*ASSOCIATE DIRECTOR HEALTH PROGRAM	67.63
*ASSOCIATE EDUCATION OFFICER	41.78
*ASSOCIATE EXECUTIVE DIRECTOR	60.64
*ATTORNEY AT LAW	51.53
*CERTIFIED APPLICATIONS DEVELOPER	59.80
*CERTIFIED DATABASE ADMINISTRATOR	61.25

Screenshot showing average pay rate across job titles

### Regular Payroll Budget by Borough:

We were also able to collect budget information by borough. This is the total budget across all boroughs across all years in the dataset. This gives us an idea of which boroughs have used the most money - notably Manhattan and Queens being some of the highest. Moving forward, we will want to break this data up and group by borough and then year to see how total

spending changes from year to year within each borough. This directly addresses one of our proposal questions so we are well set up to start working on that in our next stage of the project.

We also encountered problems here, where lower-cased versions of the boroughs were showing up as duplicates, and constrained as belonging only to the year 2014. To deal with this duplicate data, we simply uppercased all boroughs across all rows, which dealt with unnecessary splintering of the data, and disjoint information between 2014 and the rest of the decade.

Work Location Borough	
ALBANY	12210120.66
BRONX	14989735354.39
BROOKLYN	29081050165.21
DELAWARE	56861441.10
DUTCHESS	23669535.20
GREENE	7571756.33
MANHATTAN	159911431957.06
NASSAU	13024676.26
ORANGE	1136977.27
OTHER	8184416279.44
PUTNAM	25865742.54
QUEENS	34652404434.01
RICHMOND	4696210205.45
SCHOHARIE	19541880.54
SULLIVAN	93668833.63
ULSTER	231042961.39
WASHINGTON DC	5372282.51
WESTCHESTER	390738861.27
Name: Regular Gross Paid, dtype: float64	

Screenshot showing total payroll by borough

