Team: Emeka Mbazor & Mohammed Abdul Khaliq
Data Analysis/Management
CMP 464
11/30/19

NYC Citywide Payroll

I. Introduction

The Citywide Payroll dataset is a dataset containing the payroll information for NYC municipal employees in the years 2016-2019. We were interested in analyzing this dataset because New York City is the most populous city in the United States and as such has one of the largest, most expansive city governments in the world. It is imperative that its payroll funds are allocated and distributed properly in order to adequately address the concerns and needs of its diverse set of many inhabitants. The dataset consists of various features like fiscal year, work location borough, leave status, base salary, regular gross paid amount, pay basis, overtime pay, and more.

We decided to analyze this dataset to answer several questions: How does New York City allocate it payroll budget? How does the payroll differ from agency-to-agency and from borough-to-borough? What are the main factors that lead to a municipal employee being paid as much as they are?

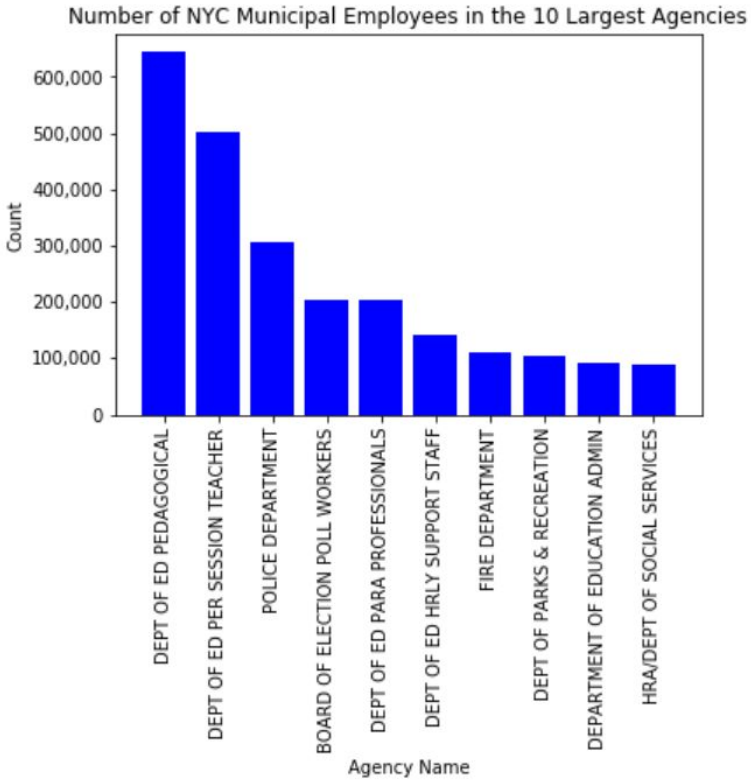
II. Data Transformation

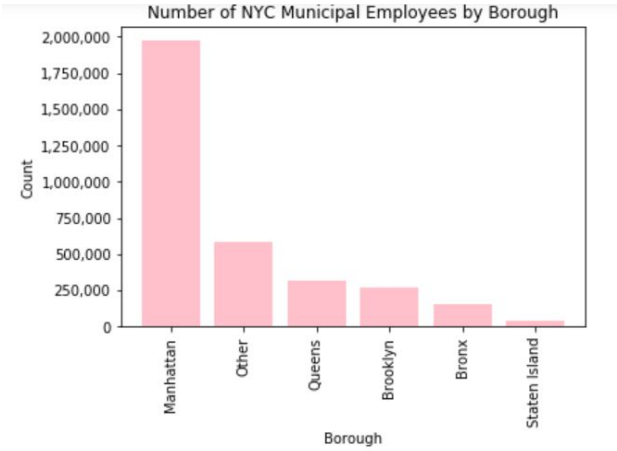
The dataset consists of 17 features and over 3,000,000 observations. For much of the data analysis and data visualization, a smaller dataset that was randomly sampled from the payroll dataset was utilized. This was done with the assumption that the smaller dataset was large enough (at 166,654 observations) to be a proper representative sample of the original dataset.

The “Agency Start Date” and “Fiscal Year” features were converted into datetime objects. Since the features “Payroll Number” and “Mid Init” missing values make up 52% and 40% of their composition respectively, they were dropped from both datasets. First and last names were also removed since they were deemed unimportant.

III. Data Visualization

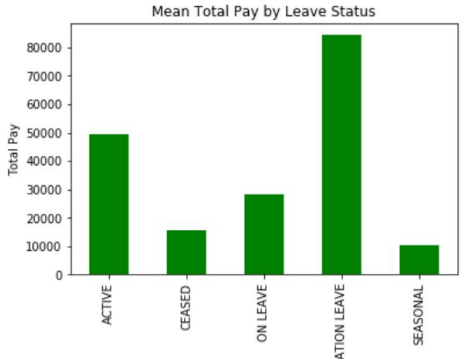
The biggest agencies by employee count are within the Department of Education.



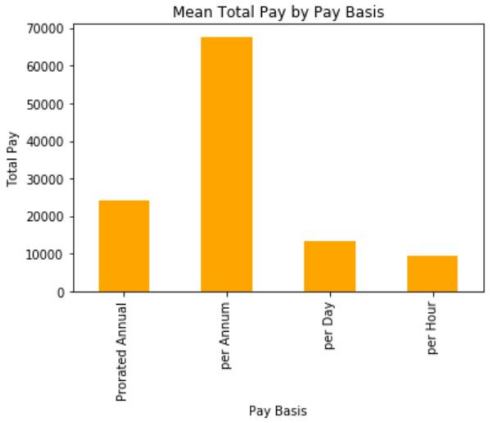


The borough with the highest number of municipal employees is Manhattan. There are a considerable amount of NYC municipal employees that work outside of New York City with more employees working outside the city than any single borough excluding Manhattan. The borough with the least number of municipal employees is Staten Island.

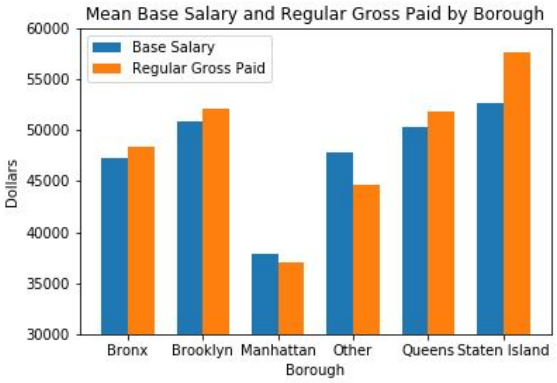
When comparing leave status, employees on separation leave have the highest mean total pay and employees who are seasonal have the lowest mean total pay.

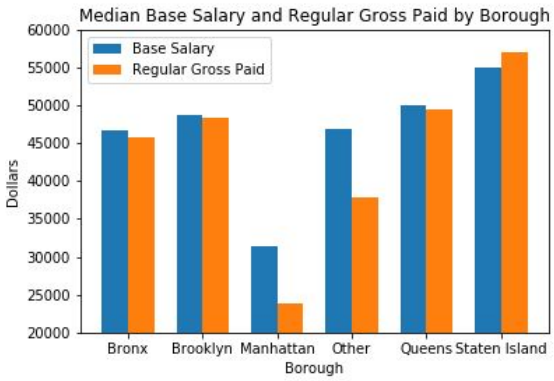


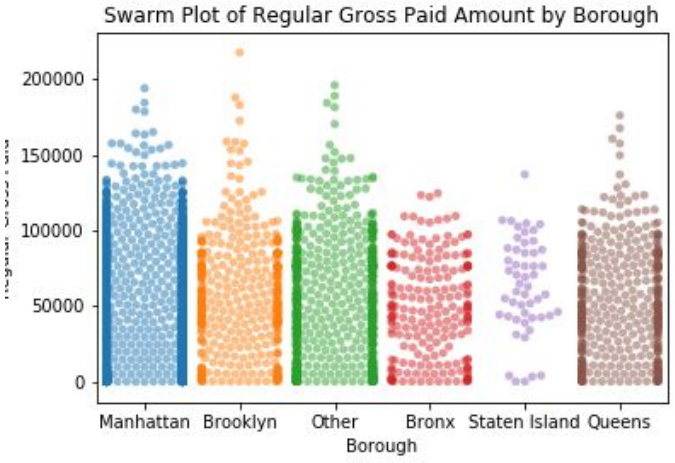
When comparing pay basis, employees who are paid on a per hour basis are the lowest paid by mean and employees who are paid on a per annum basis are paid the most overwhelmingly.



Manhattan has the lowest mean and median base salaries and non-overtime pay and Staten Island has the highest mean and median base salaries and non-overtime pay.



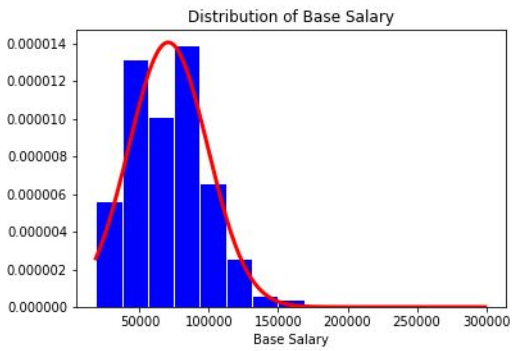


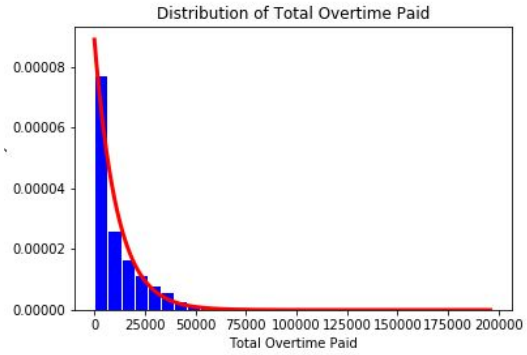


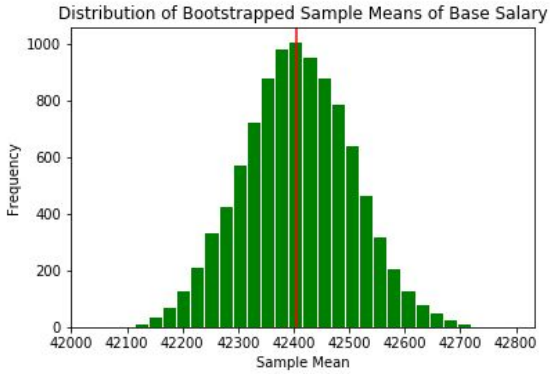
Upon further investigation and utilization of swarm plots to visualize the base salary and non-overtime pay distributions by borough, it seems that this could be because of Manhattan's abundance of low-paid municipal employees and Staten Island's lack thereof. This is despite Manhattan having many of the most compensated employees in New York City and the highest paid Staten Island employees not really being paid as much as those of other boroughs.

IV. Data Analysis

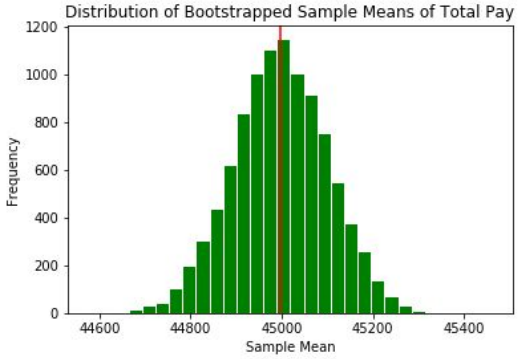
After you filter out the employees with lower salaries, the distribution of base salary roughly resembles the normal distribution with mean = $70,636 and variance = 806,943,731. Base salaries are centered around the $40,000 - $80,000 range. There are many employees with no to little overtime pay but there are quite a few outliers that make upwards to $200,000 in overtime pay. The distribution of total overtime paid to employees that have worked overtime closely resemble the exponential distribution with sigma = 8.889e-05.







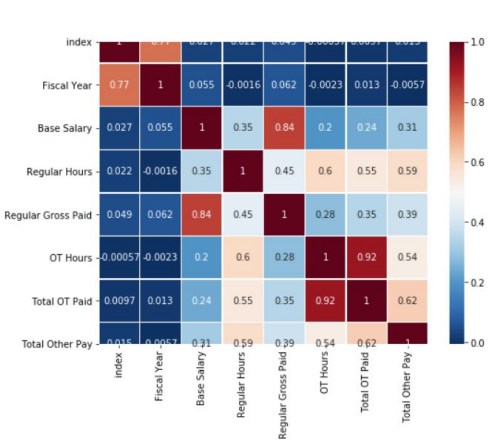
The 95% confidence interval of base salary is $42,211.00 - $42,602.44. With such a small confidence interval, one can conclude that the average NYC municipal employee has a base salary of ~$42,000. The median base salary is really close at $41,069.



The 95% confidence interval of total pay is $44,789.61 - $45,198.51. With such a small confidence interval, one can conclude that the average NYC municipal employee has a total pay of ~$45,000. However, the median total pay is much lower at $37,171.02. A large number of high-income earners could be skewing the mean of total pay.

We can conclude that base salary and total pay originate from two unique, distinct distributions with two different means since their 95% confidence intervals do not overlap.

There appears to be a positive correlation between total overtime worked and regular gross paid amount (Pearson's R = 0.596).



V. Statistical Modeling

One-Way ANOVA

A random effects one-way ANOVA model was created to see if the base salary means differ across boroughs. Strong evidence of heterogeneity was present among the borough base salary means ($F(5, 3{,}333{,}090)$, $p < 0.01$, ANOVA). We can reject the null hypothesis that the base salary means across boroughs have no difference. However, it's important to note the low R-Squared = 1.87%. The one-way ANOVA model only explains 1.87% of the variation in base pay. Due to the large population size, it's not necessary to check the one-way ANOVA assumptions of normality and equal variation.

	df	SS	MS	F	p
Treatments	5	1.043020e+14	2.08604e+13	12713.04	<0.01
Residuals	3,333,090	5.469156e+15	1,640,866,583	--	--
Total	3,333,095	5.573458e+15	--	--	--

Another random effects one-way ANOVA model was created to see if the total pay means differ across boroughs. Strong evidence of heterogeneity was present among the borough total pay means ($F(5, 3{,}333{,}090)$, $p < 0.01$, ANOVA). We can reject the null hypothesis that the base salary means across boroughs have no difference. However, it's important to note the low R-Squared = 4.21%. The one-way ANOVA model only explains 4.21% of the variation in total pay.

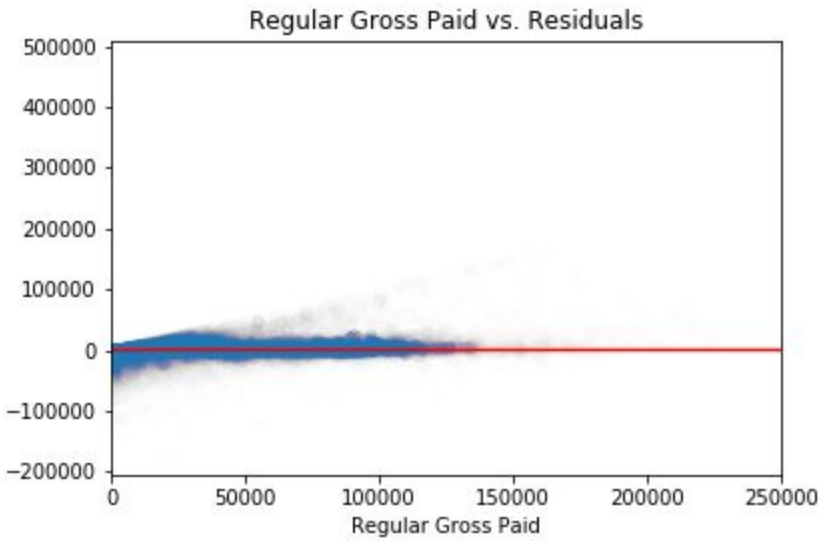
	df	SS	MS	F	p
Treatments	5	2.543818e+14	5.087636e+13	29310.28	<0.01
Residuals	3,333,090	5.785528e+15	1,735,785,112	--	--
Total	3,333,095	6.039909e+15	--	--	--

K-Nearest Neighbor Regression

A k-nearest neighbor regressor of k =5 was created in order to see if what appears to be the biggest factors on regular gross amount paid in our exploratory data analysis and data visualization efforts can be used to predict regular gross amount paid.

The response variable was regular gross paid amount and the predictors used were fiscal year, base salary, borough, leave status, and pay basis. These are what we perceive as the most important factors on regular gross amount paid.

Since borough, leave status, and pay basis are in the form of categorical data, they were converted into dummy variables in order to be used as input.



The data was split into 80% training data and 20% testing data. Training the model on the training data and validating the model on the testing data produced a relatively low RMSE of $14,145.89.

Graphing the actual regular gross paid amounts against the residuals reveals that the model

tends to perform well for both low paid employees and high paid employees alike.