2022

NYC City Wide Payroll

Data Set:

We are analyzing the **NYC Citywide Payroll data** from 2015 to 2022. The dataset was obtained from NYC Open Data, free public data to help new Yorkers use, learn, and engage with information published by the city government.

NYC Citywide Payroll data is collected due to public interest in how the City's budget was utilized regarding salary and overtime pay for all municipal workers such as Firefighters, Police force, Mayor's office, etc. The respective agency has its Personnel Management system into which they would input their payroll data which is then compiled and published by the Office of Payroll Administration.

NYC Citywide Payroll data is a very popular and exciting dataset. It was viewed more than 245 thousand times and downloaded 13.5 thousand times. Each row of the payroll data represents a New York City employee.

Below is the URL for our NYC Citywide Payroll data:
https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e

Data Set Description:

Our NYC Citywide Payroll dataset has **17 columns** and **5.11 million rows**. See **exhibit 1** for the name, description, and data types (plain text, date, or number) for our 17 columns. However, all the columns were not required to focus on our research question better. As a result, we deleted the unnecessary columns and centered our attention on **eight (8) independent variables** related to our analysis.

Below is a table with our eight independent variables, their description, and the classification as either qualitative or quantitative data.

Columns	Description	Type
Agency Name	The Payroll agency that the employee works for	Qualitative
Agency Start Date	Date which employee began working for their current agency	Quantitative
Work Location Borough	Borough of employee's primary work location	Qualitative
Title Description	Civil service title description of the employee	Qualitative
Regular Gross Paid	The amount paid to the employee for base salary during the fiscal year	Quantitative
OT Hours	Overtime Hours worked by employee in the fiscal year	Quantitative
Total OT Paid	Total overtime pay paid to the employee in the fiscal year	Quantitative
Total Other Pay	Includes any compensation in addition to gross salary and overtime pay, ie Differentials, lump sums, uniform allowance, meal allowance, retroactive pay increases, settlement amounts, and bonus pay, if applicable.	Quantitative

As shown in the table, five (5) of our variables are quantitative, while the other three (3) are qualitative.

In our analysis, we will construct our dependent variable and label it "**Total Compensation.**" For example, Total Compensation is created by adding the **regular gross pay** (base salary), **total OT paid**, and **total other pay** (bonus or retroactive pay).

We were motivated by this dataset because of the opportunity within it. Many of the City's municipal jobs offer competitive salaries and commensurate benefits. Being able to see that value and weigh it against other options can create a multitude of pathways.

Research Question:

Do **Location** and **Agency** have a statistically significant relationship on **Total Compensation** across New York City civil workers?

Cleaning Data:

Working with a dataset collected by many governmental agencies can be challenging. A few of the data were incomplete or incorrect. As a result, we clean our data using many different techniques.

First, we dropped columns such as names of the city employees because it was not helpful for our analysis. Second, we remove the blank or missing data from our dataset using the function "na.omit."
Third, we regroup variables with similar names and categories. Again, it was a big part of cleaning our data. For instance, the agency name for "district attorney" was abbreviated, so it shows up as individual dummy variables. Similarly, some variables had numbers in its name, creating separate dummy variables. In these instances, we group the variables as one name each. As a result, it lowered the number of dummy variables which was helpful later in making the decision tree because **R gives an error message for more than 32 variables in a decision** tree.

Next, our columns names had spaces between them. We rename the variables and remove the spaces so R can interpret them better. Also, we converted our categorical variables into factors. The advantage of factors is that they can be used in statistical modeling; for example, they will be allocated the correct amount of degrees of freedom compared to string/character variables. Factor variables are also convenient in many various forms of graphics.

Lastly, we removed all the negatives from our total compensation columns because it is unlikely that a city employee's pay is negative.

Data Summary Statistics:

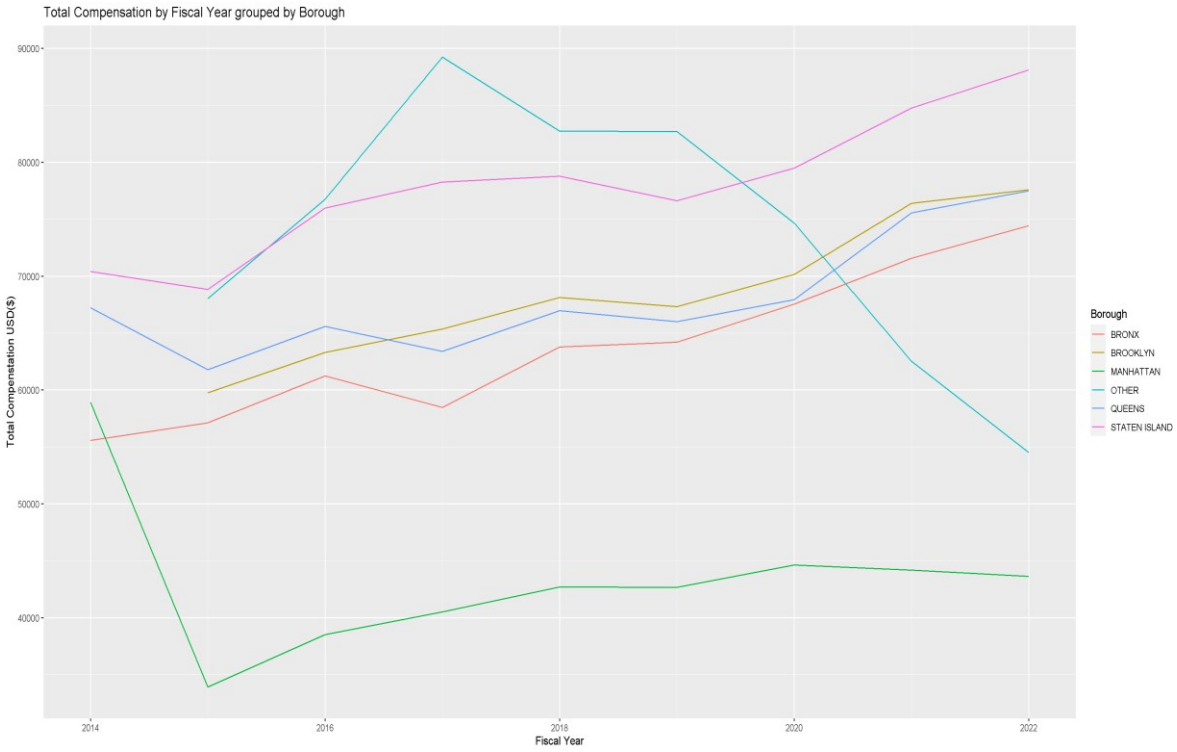
Summary Statistics: Payroll

Statistic	N	Mean	St. Dev.	Min	Max
Fisc_Year	4,592,597	2,019	2	2,014	2,022
Base_Salary	4,592,597	44,921	43,467	0	414,707
Reg_Hours	4,592,597	658	885	-1,260	4,160
Reg_Gross_Paid	4,592,597	42,890	40,787	-205,452	672,309
OT_Hours	4,592,597	64	167	-209	3,693
Total OT Paid	4,592,597	3,379	9,570	-26,494	256,000
Total_Other_Pay	4,592,597	3,002	6,078	-205,816	650,000
TotalComp	4,592,597	49,270	48,123	0	672,731

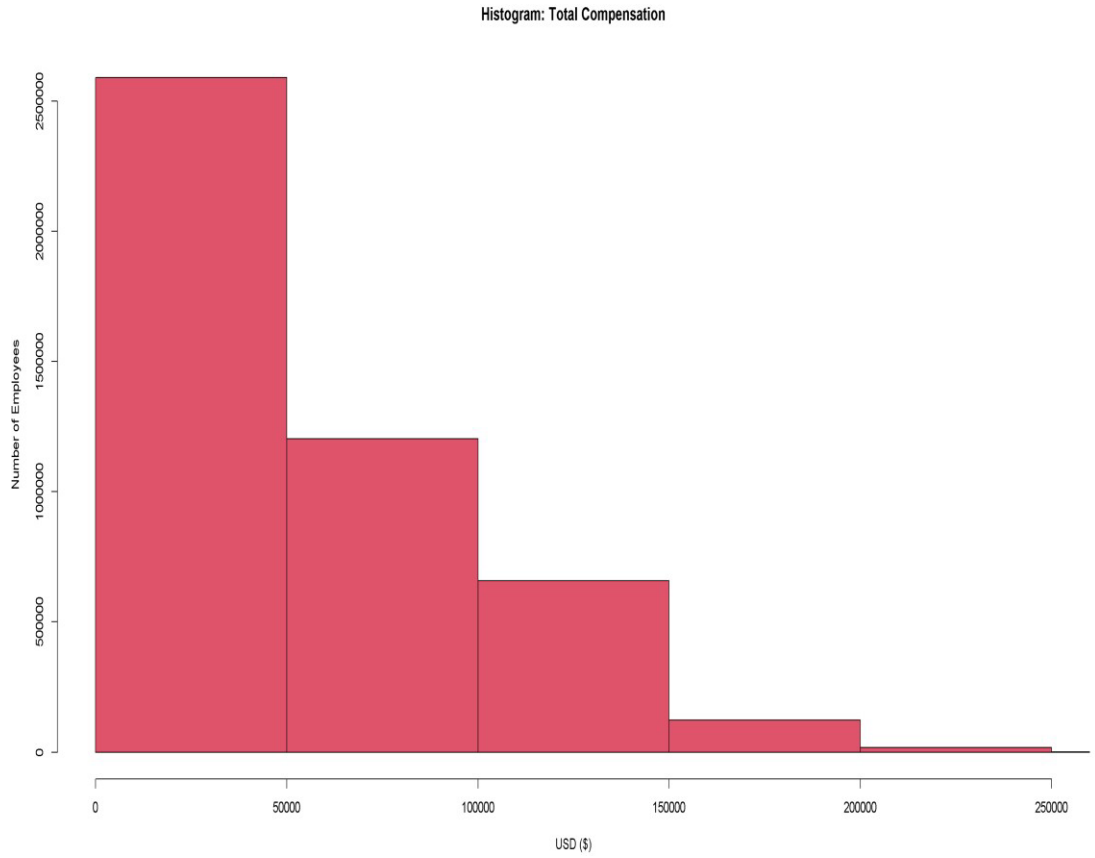
Looking at the data summary statistics, we can see some fundamental changes. Firstly, after cleaning up the data, we went down to 4.5 million rows from 5 million. We can see the basic variables like the Fiscal year, which ranges from 2014 to 2022. Other variables like Base Salary, Regular hours, Regular Gross Paid, Overtime Hours, Total Overtime Paid, Total Other Pay, and then total compensation.

Our mean total compensation ranges from a minimum of **zero (0)** to a maximum value of **$672,731**. The mean total wage of city employees is **$49,270**, with a standard deviation of **$48,123.** With a range like that, we can get a broader picture of most municipal workers are being paid from a lower to middle-class salary.

Now going into some visuals:

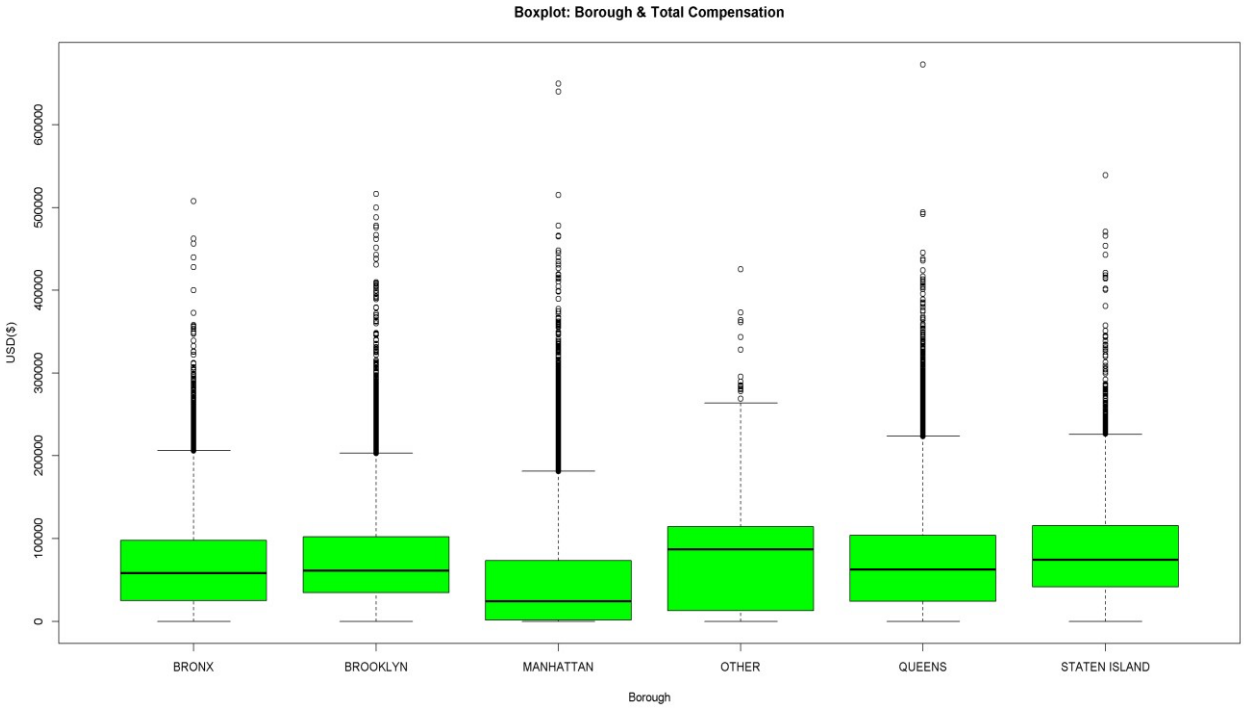


The first visual is a line chart with multiple lines, with each line being a borough with an addition of another—"Other" counties in New York, like Putnam county, Albany, Westchester, etc. The topmost line is Staten Island, and that's due to Staten Island just having many municipal workers, and most of those workers are higher rankings bringing in a higher salary. The following lines are Queens and Brooklyn, which both meet at the same endpoint in 2022. Then the next line would be the Bronx, and then it would be Manhattan. This came to us as a shock since we thought Manhattan would be higher in total compensation. We established this is due to many municipal workers skewing the numbers, most of which range in low to moderate total compensation.



In the histogram, we can see the different frequencies of total compensation. In the first bin, we have 0 - $50,000. That is where most of our total compensation falls the majority of them being from Manhattan. Then the frequencies go down at a steady pace, with the bins going $50,000 - $100,000, $100,000 - $150,000, $150,000 - $200,000, and $200,000 - $250,000, respectively.

As a result, the distribution is skewed to the right in the histogram above, and the right tail (larger values) is much longer than the left tail (small values). In a skewed right distribution, the bulk of the observations is small/medium, with a few observations that are much larger than the rest. This is because most people earn in the low/medium range of salaries, with a few exceptions (Chairman, Chancellors, Mayor, and Executive Directors) that are distributed along a large range (long “tail”) of higher values.



Now onto our last visual, the box plot. The plot follows the trend we saw in the line plot. Other has the biggest interquartile range than the boroughs. Then it follows Staten Island, Brooklyn, Queens, Bronx, and Manhattan. The most significant outlier in this data set is in Queens. The title for that outlier is "Administrative Engineer" working in a "Transportation" Agency making about $672,000. Also, the **median** differs tremendously between boroughs and the whiskers which are the upper and lower whiskers, represent scores outside the middle 50%. Whiskers often (but not always) stretch over a wider range of scores than the middle quartile groups.

Analysis Method:

We used **linear regression** as our first model because our datasets contain 5 million rows. The main advantage of linear regression is to makes the procedure of inference simple, and, most importantly, these linear equations have an easy-to-understand interpretation on a modular level. On the other hand, our project goal is to study the relationship between total compensation and the work location borough and agency in the New York City area, and linear regression shows the relationship between the independent variable and the dependent variable. It directly shows the result of our goal. Therefore, linear regression fits our project perfectly,

Our second goal is to predict the total compensation according to work location and agency name, and the **decision tree** is the method of choice for prediction. The advantage of a decision tree is that compared to other models, decision trees require less data preparation during pre-processing. In addition, since we have a huge dataset and decision trees can predict our results accurately and directly, they are relatively **easy to understand** and very effective.

Linear Regression-Results & Findings:

As mentioned in the last section, we already converted the categorical variables Agency Name and Borough into dummy variables using the as.factor function. From this conversion, we have that the baselines for our linear regression model will be Bronx and Benefit and Support, as we can see in the following matrixes:

➢ Borough variable:

	BROOKLYN	MANHATTAN	OTHER	QUEENS	STATEN ISLAND
BRONX	0	0	0	0	0
BROOKLYN	1	0	0	0	0
MANHATTAN	0	1	0	0	0
OTHER	0	0	1	0	0
QUEENS	0	0	0	1	0
STATEN ISLAND	0	0	0	0	1

➢ Agency variable:

	Courst and Law	Culture and Recreation	Education	Environment	Finance	Government and Elections	Health	Housing and Building	Public Safety	Transportation
Benefit and Support	0	0	0	0	0	0	0	0	0	0
Courst and Law	1	0	0	0	0	0	0	0	0	0
Culture and Recreation	0	1	0	0	0	0	0	0	0	0
Education	0	0	1	0	0	0	0	0	0	0
Environment	0	0	0	1	0	0	0	0	0	0
Finance	0	0	0	0	1	0	0	0	0	0
Government and Elections	0	0	0	0	0	1	0	0	0	0
Health	0	0	0	0	0	0	1	0	0	0
Housing and Building	0	0	0	0	0	0	0	1	0	0
Public Safety	0	0	0	0	0	0	0	0	1	0
Transportation	0	0	0	0	0	0	0	0	0	1

Linear Regression:

We run a multilinear regression model having 'Agency Name' and 'Borough' as predictor variables and 'Total Compensation' as our predicted variable.

The results are as follows:

Regression Model: Payroll

	Dependent variable:
	TotalComp
AgencyNameCourst and Law	-39,301.37***
	(118.28)
AgencyNameCulture and Recreation	7,687.05***
	(1,745.08)
AgencyNameEducation	-17,942.88***
	(99.97)
AgencyNameEnvironment	18,434.18***
	(143.88)
AgencyNameFinance	17,558.56***
	(234.06)
AgencyNameGovernment and Elections	11,138.28***
	(258.16)
AgencyNameHealth	4,758.05***
	(196.82)
AgencyNameHousing and Building	-11,535.84***
	(123.00)
AgencyNamePublic Safety	28,168.99***
	(105.89)
AgencyNameTransportation	10,386.11***
	(181.65)
BoroughBROOKLYN	164.22
	(111.67)
BoroughMANHATTAN	-3,966.01***
	(97.52)
BoroughOTHER	28,799.59***
	(159.33)
BoroughQUEENS	-924.04***
	(109.38)
BoroughSTATEN ISLAND	8,903.83***
	(197.27)
Constant	59,589.40***
	(123.78)
Observations	4,592,579
R2	0.20
Adjusted R2	0.20
Residual Std. Error	43,175.73 (df = 4592563)
F Statistic	74,178.84*** (df = 15; 4592563)
Note:	*p<0.1; **p<0.05; ***p<0.01

From this report we can conclude that there is a significant relationship between the predictor variables' Agency Name' and 'Borough,' and 'Total Compensation.'

Almost every predictor variable has a significant effect on 'Total Compensation.' The only one not significant is 'BoroughBROOKLYN' (which has a large p-value).

Since we are using two categorical variables as predictors that were converted into dummy variables, we need to take into consideration the baseline for each variable in order to interpret these results. The baseline for 'Borough' is the *Bronx*, and the baseline for 'Agency Name' is *Benefit and Support*.

R-squared **(20 percent)** is a statistical measure that represents the proportion of the variance of total compensation that's explained by our independent variables **(agency and borough)** in a regression model.

Interpretations:

The Constant value is 59,589. meaning that for employees working for the *Benefit and Support* agency located in a *Bronx* office, their Total Compensation will be $ 59,589.

Variable interpretation:

I. Agency Name:

- For Agency *Courts and Law*, the total compensation will be $ 59,589 - $39,301 = $20,588.
- For Agency *Culture and Recreation*, the total compensation will be $ 59,589 + $7,687 = $67,276.
- For Agency *Education*, the total compensation will be $ 59,589 - $17,942 = $41,647.
- For Agency *Environment*, the total compensation will be $ 59,589 + $18,434 = $78,023.
- For Agency *Finance*, the total compensation will be $ 59,589 + $17,558 = $77,147.
- For Agency *Government and Election*, the total compensation will be $ 59,589 + $11,138 = $70,727.
- For Agency *Health*, the total compensation will be $ 59,589 + $4,758 = $64,347.
- For Agency *Housing and Buildings*, the total compensation will be $ 59,589 - $11,535 = $48,054.
- For Agency *Public Safety*, the total compensation will be $ 59,589 + $28,168 = $87,757.
- For Agency *Transportation*, the total compensation will be $ 59,589 + $10,386 = $69,975.

II. Borough:

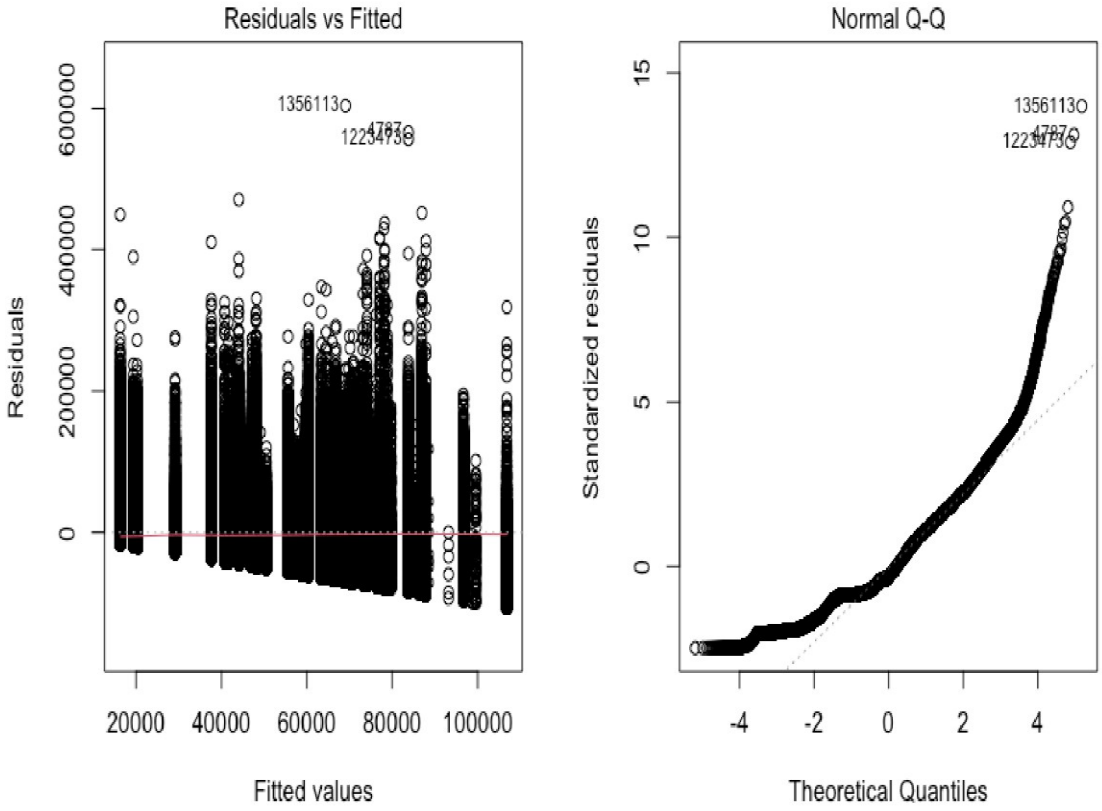
- For Borough *Brooklyn*, the total compensation will be $ 59,589 + $164 = $59,753.
- For Borough *Manhattan*, the total compensation will be $ 59,589 - $3,966 = $55,623.
- For Borough *Other*, the total compensation will be $ 59,589 + $28,799 = $88,388.
- For Borough *Queens*, the total compensation will be $ 59,589 - $924 = $58,665.
- For Borough *Staten Island*, the total compensation will be $ 59,589 + $8,903 = $68,492.

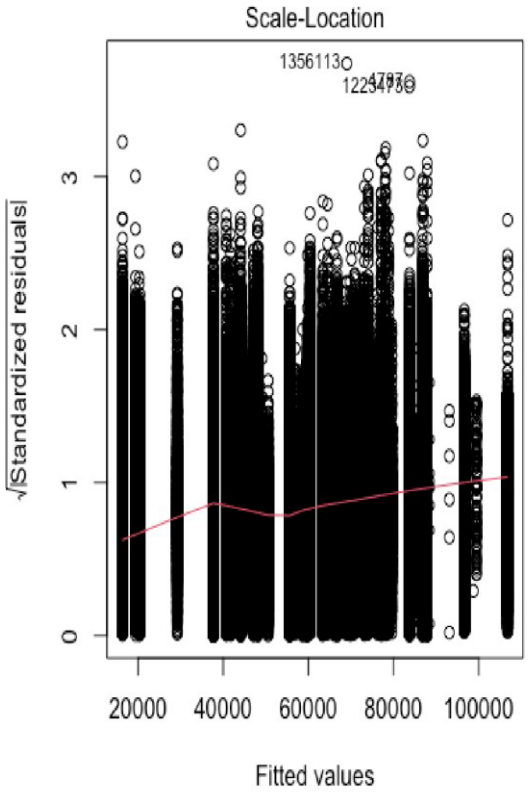
Some factors that may affect lower or higher Total Compensation are:

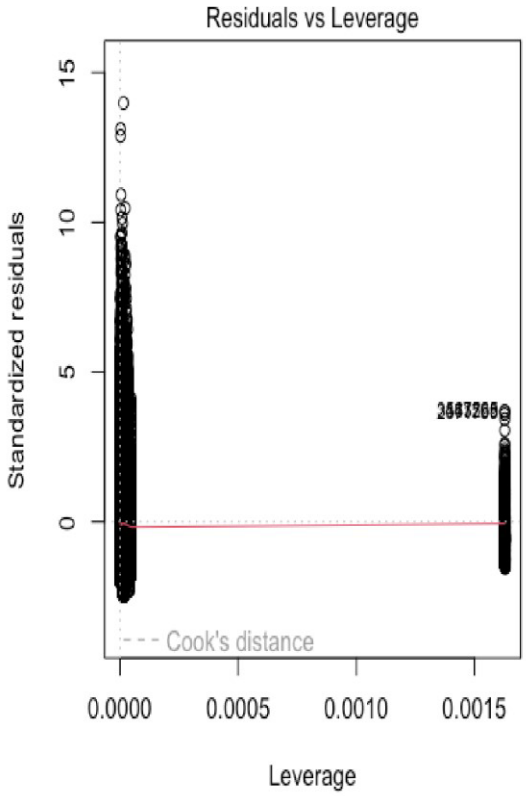
- ➢ Having a higher or lower base salary and having employees hired as full or part-time.
- ➢ Some agencies require employees to do more overtime than others.
- ➢ The borough *Others* includes Albany, which is the state capital, has a higher concentration of public offices.
- ➢ Some agencies have employees with more years of employment than others which adds to higher compensation.

Diagnostic Plots

Then we got four diagnostic plots of the least squares regression fit:







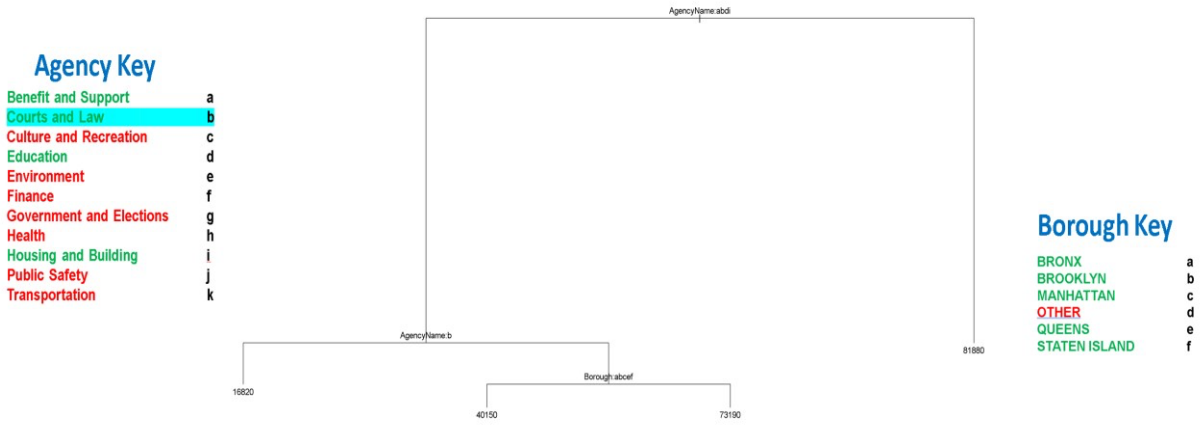
- Plot 1 (Residuals vs Fitted). We can conclude that there is no linear relationship between residuals and the fitted value. There are non-linear associations in the data.

- Plot 2 (Normal Q-Q). In this plot, we can appreciate that the Residuals follow a straight line most of the time and in the end, they deviate a little bit.

- Plot 3 (Scale location) We can see a horizontal line but not a funnel shape graph meaning that we can't assume homoscedasticity.

- Plot 4 (Residuals vs Leverage): We can identify some outliers that can influence the regression model.

Decision Tree-Results & Findings:




```
Regression tree:
tree(formula = TotalComp ~ Borough + AgencyName, data = Payroll)
Number of terminal nodes:  4
Residual mean deviance:  1898000000 = 8719000000000000 / 4593000
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -81880  -35840  -13960       0   31730  590800
```

Above is our decision tree when generated when using borough and agency names as variables. Our decision tree has **four(4) terminal nodes** and **five(5) branches**. The decision tree starts with the variables with the highest informational value. In our model, **agency names** have the highest informational value, better known as the **root node**. Also, Graphically shown above, a decision tree can handle qualitative predictors without requiring dummy variables.

Prediction:

Given the simplicity of our decision tree model, it is easy to explain. In addition, they closely resemble human decision-making.

For example, suppose we use the decision tree to predict the total compensation of a city employee that works for the "Benefit and Support" agency in Manhattan. In that case, **our prediction will be around $81,880 based on our decision tree model**.

Our Decision tree is not as strong as our linear regression model; it does not have the same level of predictive accuracy because it only has four possible outcomes, and a slight change in the data can result in a significant difference in the final projected tree

Practical Implications:

Recently, New York City's **Salary Transparency law** has taken effect. The new law states that an employer with at least four (4) employees is now required to list a "good faith" **salary range** on all posted job ads, as well as for promotions & job transfer opportunities.
To add, Colorado, Connecticut, and Maryland follow the popular trend in enacted salary transparency laws like New York.

The new law will include fines of up to $250,000 after one offense. As a result, understanding the **NYC Citywide Payroll data** and its implications can be helpful for both job seekers and employers in New York City. We present the impact into two categories: negotiation and career planning.

First, our research can aid in **negotiation**. For instance, potential or new employees can gain a perspective on which agencies have the highest starting salaries and know where they should expect their salaries to begin.

Furthermore, it can be worthwhile for current employees to determine if they are underpaid relative to others in similar positions. Employees, especially women and minorities, can ask, "Why am I paid differently than my coworker?"

Finally, **NYC Citywide Payroll data** will provide insight into how total compensation can influence your career planning decisions regarding the borough you wish to reside in or work.

Conclusion:

In conclusion, we would like to reiterate that agency and borough have statistical significance on total compensation for city employees, as proven in our regression model and decision tree. And lastly, we hope our research and analysis will advance the conversation about salary transparency and lead to actionable steps when negotiating salary and planning your career.