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

The dataset we want to utilize in our project is the Citywide Payroll dataset. It is a database containing payroll information regarding the municipal employees within New York City. This dataset is publicly available because of public interest and concern over how New York City allocates and spends its budget. We expect to extract information related to the counts of municipal employees across the city, the distributions of base pay, the distributions of actual pay, the distributions of the difference between base pay and actual pay, how the pay of active and ceased employees compare, how well hours worked correlate with pay, differences in pay across agencies, differences in pay across boroughs, and more.

We start by taking care of the missing values (standard missing values - empty cells, NA values and non-standard missing values - n/a, na) to work on a clean dataset. We also change data types when needed from object to datetime for example. We plan to extensively use density histograms to compare the distributions of different statistics. To examine information like the differences between base pay and actual pay, feature engineering is going to need to be employed to create new features. The groupby() function will be employed to compare municipal employees across different boroughs, agencies, and leave status. Utilizing filtering methods, overtime pay will be compared across different agencies. The data will be split into training and testing sets and an elementary K-Nearest Neighbor regression will be created in order to attempt to predict payroll data for municipal employees.

There are three main questions we aim to answer:

- How does New York City allocate its 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?

This project is important because New York City is the most populous city in the United States and as such has one of the biggest, 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.