In-memory Aggregators
st4324@nyu.edu
sys8910@nyu.edu
vg2237@nyu.edu

# Data Cleaning Decisions

1. Our dataset was provided with 17 columns. Of these, we found that 4 - **'Payroll Number', 'First Name', 'Last Name',** and **‘Mid Init’** - were columns that are specific to each employee. We will not be using them for our analysis as they don't offer the ability to aggregate upon. Hence they were dropped from our final dataset.
2. The original dataset was found to have 23 distinct values for ‘**Work Borough Location**’. This included the 5 boroughs, ‘Others’, Blanks and certain counties that are not in NYC. Because our analysis is limited to that of NYC, we decided to drop the ‘Others’ values. We marked blanks as ‘UNSPECIFIED’. We also discovered duplicates for some boroughs. Uppercase and lowercase spellings were being distinguished as separate entities. To solve this, we converted all values to Uppercase.
3. We noticed ‘Staten Island’ is missing from the values of ‘**Work Location Borough**’. On checking the whereabouts of Richmond on the internet, we found that Richmond is coextensive with Staten Island. Hence for sake of clarity we decided to rename ‘Richmond’ to ‘Staten Island’ as that is the most common name of the given borough.
4. After having a look at the distinct values in the ‘**Title Description'** column, we first noticed that some names were similar. We decided to have a further look into clustering them using edit distance. As we suspected, there were many duplicate titles with extra trailing/leading special characters. Thus we fed the values into a special character filter and reduced a number of duplicates in the final dataset.
5. A big discussion among the team was about what to do for the **‘Regular Hours’** column. After the initial profiling steps, it was found that about 2 million of the rows in our dataset had values of Zero for ‘Regular Hours’ even though there were many employees that received pay for that fiscal year. We speculated that there may have been many reasons for this inconsistency (negligence, typo, technical glitch) but what had to ultimately be done. Although we considered backfilling the data based on employees with similar Pay Basis and Gross Paid, we came to the conclusion that such an assumption could be misplaced and lead to non credible results.