City of Boston: City Budget

by
Julissa Mijares

Analysis Paper
Continuation of CS506 Fall 2023 at Boston University
BU Spark!

Data Cleaning:

FY24 RECOMMENDED OPERATING BUDGET[1]: This dataset should include the detailed breakdown of the operating budget for the City of Boston. It should provide information on the budget allocations for different departments, services, and personnel expenses such as teachers, police officers, and firefighters.

The raw data frame consisted of 8 columns and 900 rows. The first step of cleaning the data included removing any extra whitespaces from all columns and then replacing *#Missing* values with *Nan* (null values). Rows containing any null values were dropped, leaving 656 rows or 72.89% of the original data. Lastly, all FY datatypes were changed from *object* to *float64*.

FY24 RECOMMENDED CAPITAL BUDGET PLAN[2]: This dataset should include the capital budget for the City of Boston, specifying the funding sources, project descriptions, and budget allocations for acquiring or improving physical assets owned by the city.

The raw data frame consisted of 22 columns and 433 rows. The first step of cleaning the data included cleaning the column *Total_Project_Budget* type from *object* to *int64*. All column names were then stripped of access whitespaces. Next, columns that would not be analyzed were dropped from the data frame, including *Scope_Of_Work* and *Project_Name*. All *int64* columns replace the value *0* with *Nan* (null values). Each column that contained less than 100 non-null values were also dropped from the dataset. This included *Authorization_Future*, *External_Funds*, and all *Grant* columns, as they did not have a sufficient size to make inferences from.

The original dataset included 17 unique *Department* values and 20 unique *PM_Department* values. After reviewing the various categories, it was determined that the best method would be to combine similar departments so they would still hold value rather than being combined into a "catch-all" *other* group. Here is a list of the *Departments* and *PM_Departments* that were changed:

Department Column:

- "Fire Department" and "Police Department" combined into "Police/Fire Departments"
- "Environment Department" and "Parks and Recreation Department" combined into "Environmental Departments"
- "Property Management Department" and "Public Health Commission" combined into "Repairs and Renovation Departments"
- "Office of New Urban Mechanics" falls into "Department of Innovation and Technology"

[1] https://data.boston.gov/dataset/operating-budget/resource/8f2971f0-7a0d-401d-8376-0289e3b810ba
[2] https://data.boston.gov/dataset/capital-budget/resource/c62d666e-27ea-4c03-9cb1-d3a81a1fb641

PM_Department Column:

- “Boston Fire Department”, “Boston Police Department”, and “Police Department” combined into “Police/Fire Departments”
- “Environment Department” and “Parks and Recreation Department” combined into “Environmental Departments”
- “Boston Centers for Youth and Families” and “Youth Engagement and Employment” combined into “Youth Departments”
- “Property Management Department”, “Boston Public Health Commission”, and “Boston Public Library” fall into “Public Facilities Department”
- “Mayor's Office of Housing” falls into “Boston Housing Authority”
- “Office of New Urban Mechanics” falls into “Department of Innovation and Technology”

The cleaned dataset was left with a total of 13 unique *Department* values and 11 unique *PM_Department* values. None of the rows were deleted, leaving the cleaned data frame consisting of 11 columns and 433 rows.