Task 6.1: Project Detail

Section I: Data Source

Q3. Criteria: The FY 2022 Federal Real Property Profile (FRPP) Public Dataset meets the project brief's criteria. It is open source, from an authoritative source, and includes non-anonymized column names. The cleaned subset dataset contains 10,000 rows, well above the required 1,500, and includes both continuous variables (such as latitude, longitude, number of federal employees, operations cost, maintenance cost, and square feet) and categorical variables (such as real property type, real property use, utilization, and asset status). It also includes a geographical component with latitude and longitude values. The subset is from FY2022 data set verified on data.gov to ensure it falls within the last 3 to 10 years.

Q5. Source Summary: The FY 2022 Federal Real Property Profile (FRPP) Public Dataset, updated in October 2023, is the U.S. federal government's centralized inventory of real property under the custody and control of executive branch agencies. The dataset provides detailed information on various federal properties, including their location, usage, and status, while excluding some data due to national security and other authorized reasons. This data is authoritative and meets the criteria for open-source datasets, offering a wealth of information for analyzing office space utilization. It includes both continuous variables (e.g., square footage, number of employees) and categorical variables (e.g., property type, utilization status), as well as geographical components such as latitude and longitude, making it suitable for comprehensive analysis and visualization. The dataset can be accessed through Data.gov.

Explanation of Interest: I chose the FY 2022 Federal Real Property Profile (FRPP) Public Dataset for my analysis because of my interest in analyzing how the U.S. General Services Administration (GSA) captures and measures office space utilization for federal agencies. This dataset provides detailed records on various aspects of these properties, including their location, usage, status, and physical characteristics, ensuring accurate and up-to-date information crucial for meaningful analysis.

The FRPP dataset is particularly suitable due to its extensive coverage, detailed attributes, and geographical component. It includes a wide range of properties, allowing for a thorough examination of office space utilization across different federal agencies. The dataset's attributes, such as property type, utilization status, number of federal employees, and building size, are essential for understanding office space management. Additionally, the inclusion of geographical data enables spatial analysis and visualization of office space distribution, valuable for identifying patterns and trends. By leveraging this dataset, I aim to gain insights into the GSA's strategies for office space utilization, helping to identify opportunities for optimization and efficiency improvements within federal agencies.

Section II: Data Profile

Subsetting: A subset of the FY 2022 Federal Real Property Profile (FRPP) dataset was created to focus on specific columns relevant to analyzing office space utilization. This step was necessary to streamline the data for analysis by selecting only the columns that provide critical information, such as utilization status, property type, number of federal employees, and geographical data. Working with a smaller, relevant dataset enhances the efficiency of the analysis and makes it easier to manage and visualize the data.

Cleaning and Consistency Checks: Cleaning the data involved addressing missing values, correcting data types, and removing duplicates. Missing values in categorical columns (e.g., Utilization, Utilization Code) were filled with the mode (most frequent value), while missing values in numerical columns (e.g., Latitude, Longitude, Building Age, Number of Federal Employees, Annual Operations Cost, Annual Maintenance Cost) were filled with the median. The Building Age column, initially of object type, was converted to a numeric type to ensure consistency in data analysis. Duplicate rows were identified and removed to ensure that each record in the dataset is unique.

Consistency checks included verifying data types and summarizing outliers to understand the distribution and identify any extreme values that might need further attention. These steps ensured the dataset was prepared for accurate and efficient analysis, maintaining the reliability of insights drawn from the data.

Q7. Understanding the Data

Summary of EDA on Sampled Data: The initial exploratory data analysis (EDA) encountered performance issues due to the large dataset size and plot generation time. To address this, key steps and findings were summarized. The dataset contains 10,000 rows and 14 columns, with a mix of numerical and categorical data types. Descriptive statistics provided a summary of central tendency, dispersion, and distribution shape. Histograms visualized the distribution of numerical variables like square footage, number of federal employees, operations cost, maintenance cost, and building age. Count plots displayed the frequency of categories within variables like utilization, real property type, real property use, and asset status. Key findings included several columns with missing values, particularly in Utilization and Utilization Code, a wide range of values with some outliers in numerical variables, and observable correlations between certain numerical variables, such as operations cost and maintenance cost.

I also used a sample dataset containing 1,000 entries and 14 columns, with a mix of numerical and categorical data types. There are missing values in columns such as Utilization, Utilization Code, Latitude, Longitude, Building Age, Number of Federal Employees, Owned and Otherwise Managed Annual Operations Cost, and Owned and Otherwise Managed Annual Maintenance Cost. This indicates a need for data cleaning to address these gaps. The dataset's basic structure includes columns that provide critical information for analyzing office space utilization, such as real property type, utilization status, and geographical coordinates.

Descriptive statistics reveal a range of values for each numerical variable, including minimum, maximum, mean, and standard deviation. Notable observations include Square Feet (Buildings) ranging from 10 to 269,541 square feet, with a mean of 2,545.65 square feet, and Number of Federal Employees ranging from 1 to 738, with a mean of 29.83. Both Owned and Otherwise Managed Annual Operations Cost and Maintenance Cost show significant variation, suggesting potential outliers. The Year of Construction ranges from 0 to 9999, indicating some erroneous entries that need correction. Overall, the dataset provides valuable insights but requires further cleaning and validation to ensure accuracy and reliability for analysis.

Q8. Limitations: The dataset presents several limitations that should be addressed to ensure accurate analysis. Missing values are prevalent in columns like Utilization and Utilization Code, which could lead to incomplete results and may require imputation or exclusion, potentially introducing bias. There are also erroneous entries, particularly in the Year of Construction column, which need correction to maintain data integrity. The presence of significant outliers in numerical variables, such as Square Feet (Buildings) and cost columns, could skew the analysis, necessitating proper outlier treatment. Additionally, while the dataset is relatively recent (FY 2022), property details may have changed since data collection, affecting the current relevance of the analysis. Lastly, the dataset's geographical scope includes properties across various regions, and without a clear understanding of regional differences and local contexts, some insights may be misinterpreted.

Ethical Considerations: Data privacy and security should be maintained, ensuring that the analysis respects privacy and does not expose sensitive information, especially in aggregated or visualized results. Some data exclusions due to national security reasons may lead to an incomplete analysis, affecting comprehensiveness and accuracy. Inherent biases in data collection and reporting by different federal agencies need to be acknowledged to avoid misleading conclusions. The analysis involves assessing the use of public funds, so findings and recommendations must be used responsibly to improve public resource management. Furthermore, recommendations based on the analysis could lead to policy changes affecting federal employees and property management practices, necessitating a consideration of equitable and beneficial impacts.

Section III: Questions to Explore

- **Space Optimization:** What opportunities exist for optimizing office space utilization? Can underutilized spaces be repurposed or consolidated to improve efficiency?
- **Utilization Efficiency:** How effectively are federal office spaces utilized across different agencies? Are there significant differences in utilization rates between different property types (e.g., office buildings, warehouses)?
- **Occupancy Trends:** What are the occupancy trends across various regions and agencies? How do the number of federal employees correlate with the size of the office space in different regions?

- **Cost Analysis:** What are the annual operations and maintenance costs for federal office properties? How do these costs vary by region, property type, and utilization status?
- **Policy and Compliance:** How well do federal properties comply with space utilization policies and guidelines? Are there specific agencies or regions that consistently meet or fail to meet utilization standards?