Mini Project 1: Structured Data Report(Real Estate Analysis)

Submitted by: Kruti Kotadia

Date: 15th October 2023

**The Data and its Source**

This report presents the analysis of a structured dataset using Python. The dataset used for this analysis is on New York City Real Estate Sales dataset, which was retrieved from Kaggle. The dataset contains detailed records of property sales in New York City. It includes information on sale prices, property types, boroughs, sale dates, and more. The dataset is a valuable resource for understanding the real estate market in New York City. The analysis in this report is based on this real estate dataset, and we will describe the data exploration, cleaning, and answer specific comparison questions using this dataset.

**Data Exploration and Cleaning**

In this section, I will describe the steps taken to explore and clean the dataset.

Data Source

* The dataset utilized in this analysis is sourced from Kaggle's dataset repository. It is a comprehensive real estate data dataset, providing valuable insights into property sales in New York City. The dataset comprises around 8 to 9k rows and 21 columns.

Data Cleaning

* The data cleaning process involved these steps:
* Dropped irrelevant columns such as 'TAX CLASS AT PRESENT,' 'BLOCK,' 'LOT,' and others.
* Removed duplicate rows to maintain data accuracy.
* Dealt with null values in key columns, like 'SALE PRICE,' 'YEAR BUILT,' 'TOTAL UNITS,' 'LAND SQUARE FEET,' and 'GROSS SQUARE FEET,' by replacing them with NaN for consistency.

**Comparison Questions and Analysis**

How does the total number of units (commercial + residential) vary across boroughs?

* Unit of Analysis: total number of units (both commercial and residential) in each borough
* Comparison: The values being compared are the total numbers of units across different boroughs. This comparison helps us understand how the availability of commercial and residential real estate varies from one borough to another.
* How it's computed: The comparison is computed by first grouping the data by borough, then summing up the number of commercial and residential units within each group. This gives us the total number of units for each borough.

Is there any co-relation between the average gross ft. area and total units for each borough?

* Unit of Analysis: Gross Square feet area and total units
* Comparison: From the graph, it appears that as the number of commercial and residential units increases, the average gross square feet also tend to increase. This suggests a positive correlation between the number of units and the property size. However, there are some exceptions to this trend, indicating that other factors may also influence the property size.
* How they are computed: - This observation was computed by calculating the average gross square feet within groups of properties based on their number of commercial and residential units and then plotting the results.

Program Description

The analysis was performed using a Python program. The following steps were taken within the program:

* Data was read from a CSV file.
* Data cleaning procedures were applied to remove irrelevant columns, duplicates, and inconsistent values.
* Specific columns, such as 'SALE PRICE,' 'YEAR BUILT,' 'TOTAL UNITS,' 'LAND SQUARE FEET,' and 'GROSS SQUARE FEET,' were converted to numeric types for analysis.
* The 'SALE PRICE' column was reduced by filtering the dataset to include values between $1,000,000 and $100,000,000.
* The 'YEAR BUILT' column was sorted to get a clear view of the data distribution.
* Visualizations, including line plots and bar charts, were generated to visualize and analyze the data.

**Conclusion**

In conclusion, the cleaning and exploration steps ensured the data's quality and reliability. By answering two specific comparison questions, we gained valuable insights into the dataset, which can be useful for various purposes.