Real Estate analysis for State of Connecticut

1. Introduction:

This project aimed to leverage advanced data processing and visualization technologies to analyze a real estate dataset. By employing Azure Container, Snowflake as a data warehousing solution, and Power BI for data visualization, the project aimed to uncover insights from the dataset, create meaningful visualizations, and enhance decision-making within the real estate domain.

2. Dataset Description:

The real estate dataset consists of property transaction records encompassing attributes such as Serial Number, List Year, Date Recorded, Town, Address, Assessed Value, Sale Amount, Sales Ratio, Property Type, and Residential Type, Non-Use Code, Assessor Remarks, OPM remarks and Location. These features provide a comprehensive view of property transactions and valuation dynamics.

3. Methodology:

[1] Data Ingestion and Preprocessing:

- The real estate dataset was securely stored within an Azure Blob Storage using an Azure Container. This facilitated streamlined data management and access.

- Data preprocessing involved meticulous cleaning to address missing values, outliers, and data inconsistencies.

[2] Integration of Snowflake:

- Snowflake, a scalable and performant data warehousing solution, was chosen to host and manage the processed data. The dataset was organized into tables within the "project_data" schema, optimized for analysis.

- Transformations and manipulations were performed within Snowflake using SQL queries. These included aggregations and calculations to derive significant metrics for analysis.

[3] Data Loading and Transformation:

- A file format was created to handle CSV data, specifying delimiters and handling NULL values.

- A stage object, linked to the Azure Storage Integration "azure_object," was established to facilitate data loading.

- The dataset was loaded into the "real_estate_data" table using the COPY INTO command. The data underwent necessary type conversions and transformations and removal of some attributes to match the table schema and remove the unnecessary information.

[4] Data Visualization using Power BI:

- Power BI was employed to create an insightful dashboard for data visualization. The dashboard showcased essential key performance indicators (KPIs), property distribution, trends, and insights derived from the real estate dataset.

- A variety of visualizations, such as bar charts, line charts, maps, and KPI cards, were carefully designed to effectively communicate property types, sales trends, geographical distribution, and pricing dynamics.

- Some of the columns which had wrong calculations were corrected and new group column was introduced along with removal of outliers using M language. Necessary type conversions were performed.

[5] Data Visualization:

The first dashboard comprises a collection of visualizations designed to provide a comprehensive understanding of key trends and metrics within the real estate dataset.

1. Slicers for Year and Sales Ratio Group:
   The slicers enable dynamic filtering by year (ranging from 2006 to 2020) and sales ratio groups (categorized as “Loss”, "Bad," "Good," and "Beyond Expectation"). This empowers users to tailor their analysis based on specific time periods and sales ratio thresholds.

2. Property Type vs Count Bar Chart:
   A bar chart displays the distribution of property types based on their count. This visualization provides a clear overview of the prevalence of different property types within the dataset.

3. Assessed Value Sum TreeMap:
   The TreeMap visualizes the sum of assessed values across different towns. The intensity of size represents the magnitude of assessed values, while the labels provide insight into individual towns' contributions.

4. Property Type, Town, and Sales Group Table:
   A table showcases property type, town, and sales group information. This tabulated format facilitates a detailed examination of property distribution, town associations, and sales performance.

The second dashboard delves into more advanced visualizations to uncover intricate relationships and trends within the dataset.

1. Scatter Plot of Sale Amount vs Assessed Value with Linear Trend Line:
   The scatter plot illustrates the relationship between sale amount and assessed value, complemented by a linear trend line.

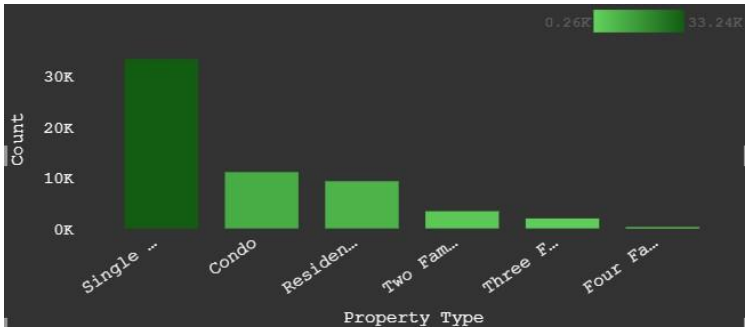
2. Shape Map with Color Intensity by Sale Amount:
   The shape map utilizes the "Town" column to depict geographic locations. The varying color intensity indicates sale amounts, with darker shades representing higher values. This visualization offers a geographical perspective of towns with varying sales amounts.

3. Area Chart of Sales Ratio vs Count of Town with Emphasis on Single Family Property Type: The area chart maps the relationship between sales ratio and the count of towns.

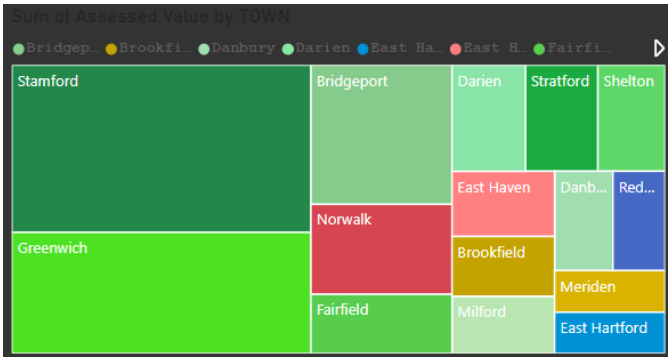
4. Sales Ratio KPI and Town Insights:
   A KPI (Key Performance Indicator) highlights sales ratio fluctuations.

[6] Insights

1. Property Type Distribution: The bar chart underscores the prevalence of property types, helping identify dominant categories within the dataset. In this case over the years the prevalent property type has been Single family homes and hence can be expected that it would keep going up.



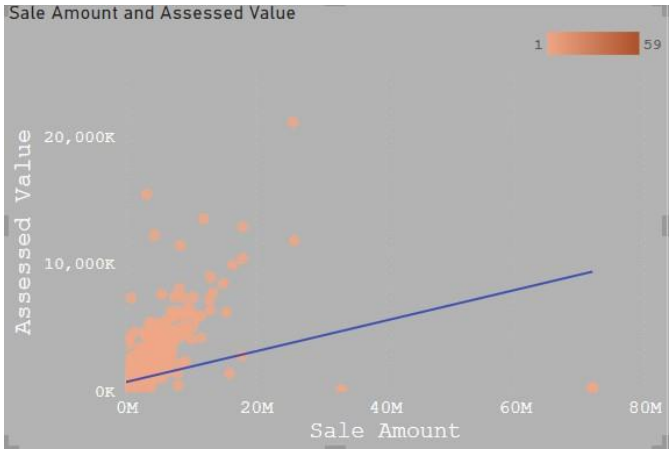
2. Assessed Value Sum TreeMap: This visualization unveils towns contributing significantly to assessed values, aiding in the identification of lucrative real estate markets. One of the most expensive houses with assessed values were found to be in town of Stamford and Greenwich along with BridgePort.



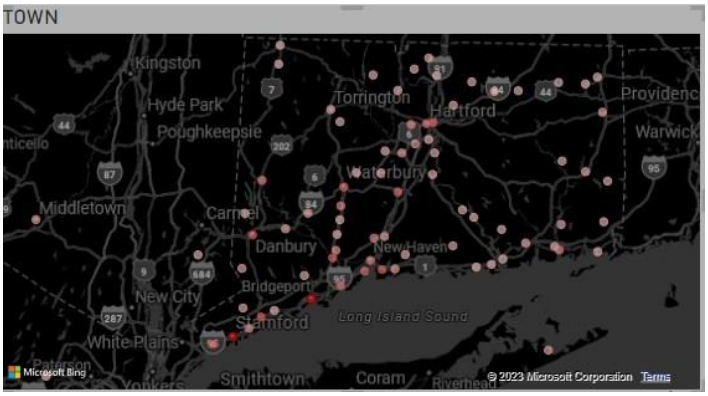
3. Based on the data presented in the table, it is evident that the property type "Condos" has exhibited a consistently lower sales ratio over the years, indicating a declining market performance for this particular property type.

Property Type	TOWN	Sales Group
7	Greenwich	Bad
Condo	Ansonia	Bad
Condo	Ashford	Bad
Condo	Avon	Bad
Condo	Beacon Falls	Bad
Condo	Berlin	Bad
Condo	Bethel	Bad
Condo	Bloomfield	Bad
Condo	Branford	Bad
Condo	Bridgeport	Bad
Condo	Bristol	Bad
Condo	Brookfield	Bad
Condo	Brooklyn	Bad

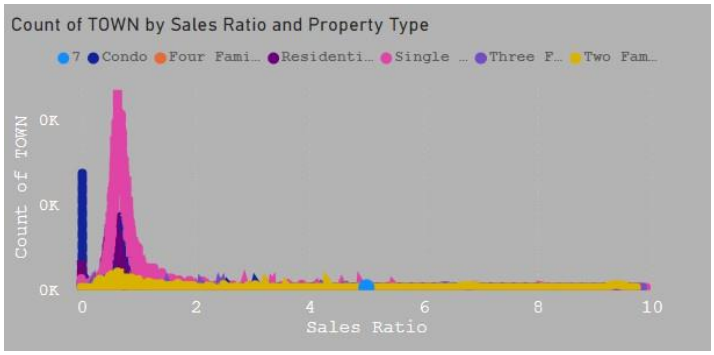
4. Scatter Plot and Trend Line: The scatter plot and trend line accentuate the common trend where sale amounts are lower than assessed values, suggesting potential decline in property values. The blue trend line is upward sloping, indicating a positive correlation. Most of the data points are clustered near the lower end of both axes hence Majority of properties have sale amounts and assessed values under 20M. There seems to be outliers which should have been removed before analysis or needs to be verified since if assessed value is low then how is the sale amount so high.



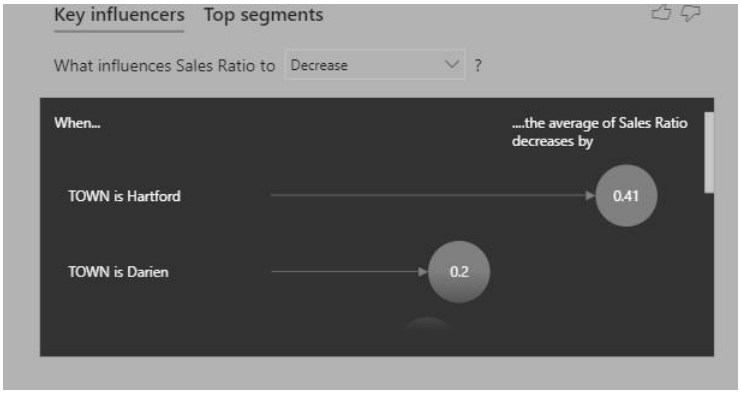
5. Geographical Sale Amount Representation: The shape map highlights towns with varying sale amounts, assisting in pinpointing high-value regions. Towns of Stamford, Norwalk and Shelton have high sale amount of houses. Meaning these are expensive neighborhoods.



6. Sales Ratio Trends and Single Family Property Type: The area chart sheds light on sales ratio distribution. Notably, the area corresponding to the "Single Family" property type is prominently elevated, indicating higher occurrences in the dataset compared to other property types.



7. Sales Ratio KPI and Town Insights: The KPI and town insights spotlight noteworthy sales ratio changes, allowing stakeholders to focus on areas of interest. For increasing sales ratios, the town of "Bridgeport" stands out with a value of 0.26. Conversely, decreasing sales ratios feature multiple towns, with "Hartford" having the highest value of 0.41. These insights spotlight towns with significant changes in sales ratios.



4. Conclusion and Future Enhancements:

Potential enhancements include integrating an Azure Pipeline for seamless data updates, while leveraging machine learning models offers predictive insights into property valuations and sales ratios. Expanding data sources to encompass economic indicators and demographics provides a comprehensive market view. These advancements collectively amplify the project's impact by ensuring real-time relevance, predictive analysis, interactive exploration, comprehensive insights, and seamless scalability for robust real estate data analysis.

The implementation of Azure Container, Snowflake, and Power BI showcased their prowess in data management, analysis, and visualization, significantly contributing to the informed decision-making process within the real estate domain.