**7PAM2000-0901-2023 APPLIED DATA SCIENCE 1**

**ASSIGNMENT 1: VISUALIZATION**

***MOHIT AGARWAL* | *STUDENT ID – 22031257***

* Link to the Dataset can be found [here](https://www.kaggle.com/datasets/reenapinto/real-estate-sales-2001-2020), taken from Kaggle website.
* Link to the GitHub Repo can be found here.

**Dataset**

**Real Estate Dataset Maintained by the Office of Policy and Management**

The dataset holds the list of all the property sold between October 1 and September 30 of each year, from 2001 to 2020, with prices more than or equal to 2000 USD. Each recorded entry holds the following entries:

* Town
* Property Address
* Date of Sales
* Property Type
  + Residential Type
  + Apartment
  + Commercial
  + Industrial
  + Vacant Land
* Sales Price
* Property Assessment

The total number of entries in the dataset is 997213 x 14.

**Methodology**

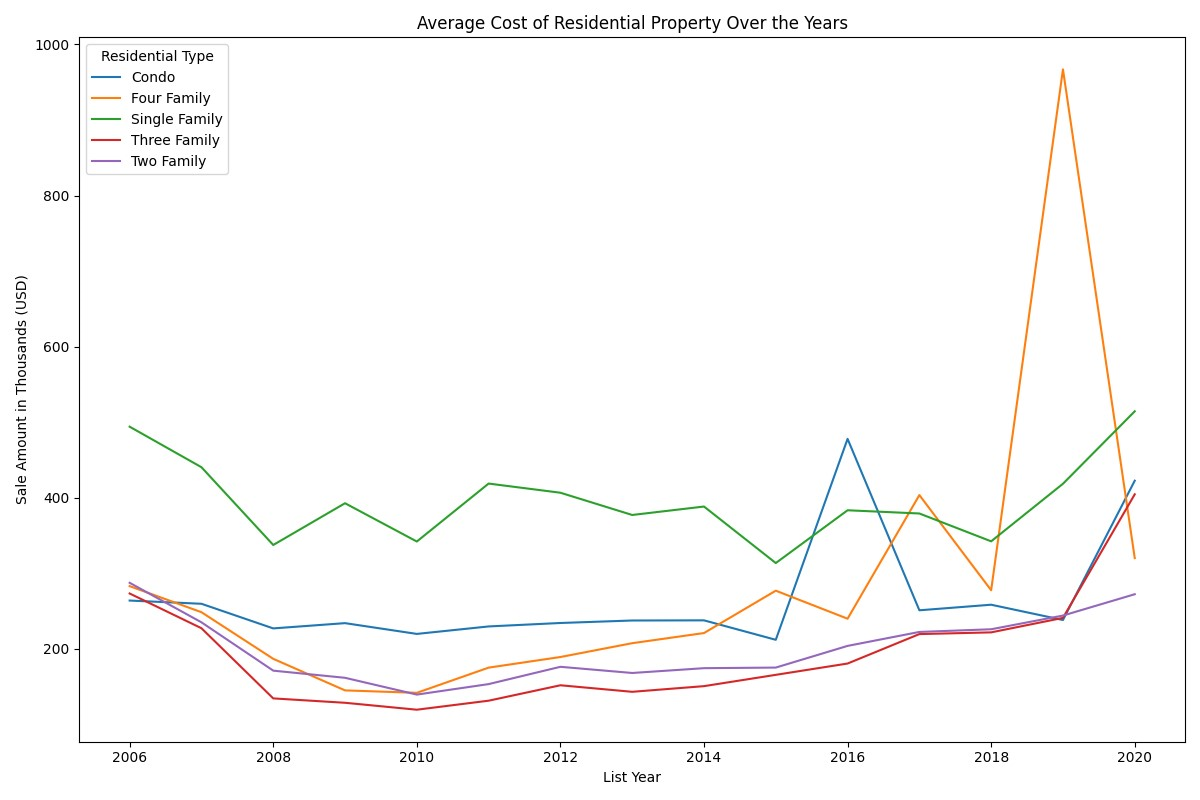
First the data was explored using info(), describe(), and several meaningless columns, at least not useful for this assignment, were dropped, by method of subsets. The missing data in the dataset were left alone, as the available non-null entries were more than enough to process with further analysis.

Second, the Property Type seen in the dataset were mainly divided into residential, apartment, commercial, industrial, and vacant land. The residential property type was further grouped into condo, four family, three family, two family and single family, which formed most of the dataset. So, I decided to explore/visualize upon this type of property.

Finally, three plot types were chosen for the data visualization, namely, line plot for time series problem, bar plot for categorical representation of number sales per town, and a pie chart for market capture by various types of property sold. Each plot has a size of 12x8.

**Visualizations**

**1. Line Plot Visualization**: **Average Cost of Residential Property Over the Years.**



**Reason**: Average Cost of each Residential Property type over the years, makes it a time-series problem, which can be shown easily through a line plot. Line Plot are good with tracking the changes over the time.

**Analysis:** Overall, the cost of Single-Family properties has remained high compared to other properties sold, but around the year 2015, Condo prices did below up, above the selling price of the single-family homes and recovering the following years. In 2006 the average selling prices of Condo, Four Family, Three Family, Two Family, were almost the same, sitting around 300,000 US dollars. The significant changes can be seen in the prices of four family houses which quickly rose from 2014 with a downfall around 2017 and rapidly climbed from 2018 to be the costliest type of property for almost a year and half.

**2. Bar Plot Visualization: Total Number of Residential Property Sold per Town.**

A graph of different colored bars

Description automatically generated

**Reason:** Bar Plot is good for representing categorical data, in this case, the number of residential properties sold in each town is shown, and using the stacked property of the bar plot, groups of residential properties has been visualized.

**Analysis:** From the stacked bar plot, it quickly established that Bridgeport is the highest residential selling Town, sold above 17500 over 20 years, closely followed by Waterbury and Stamford. The majority of sold properties in each town over the span of 20 years are single family homes, followed by Condo and then Two-Family type. However, some towns have also sold large number of three family homes, with Waterbury in lead, then Bridgeport and then New Haven. Finally, many towns have sold any four family homes over the past 20 years, like West Hartford and Fairfield.

**3. Pie Chart Visualization: Market Captured by Property Type.**

A close-up of a graph

Description automatically generated

**Reason:** Looking at the change of market captured by the property type from the first five years over to the last five years, can swiftly be done by a pie graph.

**Analysis:** In the first five years, i.e., from 2001 to 2006, the market is solely occupied by the residential property type, with single family homes having highest market value at 71.3 percent, followed by Condo at 21.46 percent and the four family homes at the lowest market value at only 0.26 percent. However, the last five years, i.e., 2015 to 2020, the market has been shared by non-residential property types, due to which share of single-family home shrink by almost 22 percent and Condo’s share also decreases by nearly 6 percent.

**Conclusion**

1. The single-family homes are the costliest among the residential properties with average cost around 400,000 USD, despite which the number of single-family homes sold are very high, occupying the highest market share.
2. Second choice is condo, which are bought in each town, where single-family homes are also sold, and the two-family comes at third option, which are sold for nearly half the prices of single-family homes.
3. Three-family homes and four-family homes are the least expensive and least preferred options in that order, with many towns, like Newton, Darien, not even selling a single unit.

**Note:** I found it bit strange that single-family homes are expensive than two or three or four -family homes.

**Dataset Sort Comings**

1. The dataset has plenty of entries but lacks the proper categorizing of property types, above 250,000 of categories were missing.
2. The dataset didn’t have any information about the square footage of the property been sold.
3. The dataset also didn't provide the information of surrounding areas, factors that lead to the rise in selling prices.