Data Analysis Report

- ## Introduction

This project predicts the values/prices of residential housing properties in Burlington, Vermont. An unbiased and objective evaluation of home prices based on big data is beneficial to both buyers and sellers of houses. While the sale prices/current values of housing properties are given in the "City of Burlington Property Details" data, it is not clear how various aspects of a house is related to its sale price/current value. We attempt to establish a predictive model to demonstrate the effect of various features of a housing property on its value. The "CurrentValue" and "SalePrice" information in the "City of Burlington Property Details" data set instead will be used to test the accuracy of our predictive model. The goal of this project is to give an estimated value or a range of estimated values based on the information on a housing property. Interested home buyers/sellers can use our established models of a house price evaluation system as a reference before making a decision on a transaction.

- ## Data Source

The main data set to be used is the "City of Burlington Property Details" data, which is open to public in https://data-burlingtonvt.opendata.arcgis.com/datasets/276ccff527454496ac940b60a2641dda_0

This data set contains 10832 records on taxable house property details including "Number of Units", "Number of Rooms", "Heat Type", "Building Type", "Current Building Value" etc.

- ## Data Preparation

While the data set contains sufficiently many columns which describe details of housing properties, empty records, mistyped data etc. need to be dealt with before the conduct of data analysis and machine learning. The data set also contains irrelevant information of housing properties such as the ID number of the property, owners' names etc. which be discarded for the purpose of better data analysis. What have been done to the original data set are 1. preliminary selection of relevant columns 2. treatments to empty data; 3. encoding of non-numeric data and 4. treatments to suspicious data and outliers. Details for each of these data preprocessing steps will be explained below.

1.Relevant Columns

The original data set contains columns such as 'AccountNumber', 'ParcelID', 'SpanNumber', which serve as identifiers of housing properties, and columns consisting of owners' names. Most of these identifier/name columns will be deleted. The data set that is used for data analysis contains the columns from the original data set which the author believes to be relevant for analysis of sale prices and current values. The columns that are preserved from the original data set are FID(at least one identifier should be kept), StreetNumber, StreetName, LandUse, CurrentAcres, TotalGrossArea, FinishedArea, CurrentValue, CurrentLandValue, CurrentYardItemsValue, CurrentBuildingValue, BuildingType, HeatFuel, HeatType, Grade, YearBlt, SalePrice, NumofRooms, NumofBedrooms, NumofUnits, ZoningCode, Foundation, Depreciation, PropertyCenterPoint. Most of the column names suggest by their meaning what information is stored. StreetNumber and StreetName together indicate locations of housing properties. PropertyCenterPoint stores the tuples consisting of latitudes and longitudes of housing properties.

2.Empty data

Since no information will be obtained from empty data, and empty data is of no use to machine learning algorithms, we will have to treat the empty data either by deleting or filling with reasonable estimated values. The percentage of missing data for all columns are listed in the table below.

Columns	Missing Percentage
StreetNumber	0.0462%
StreetName	0%
Unit	80.8346%
LandUse	0%
CurrentAcres	0%
TotalGrossArea	0%
FinishedArea	0%
CurrentValue	0%
CurrentLandValue	0%
CurrentYardItemsValue	0%
CurrentBuildingValue	0%
BuildingType	3.7020%
HeatFuel	4.5513%
HeatType	3.9051%
Grade	3.7020%
YearBlt	0%
SalePrice	0%
NumofRooms	0%
NumofBedrooms	0%
NumofUnits	0%

Columns	Missing Percentage
ZoningCode	0.5078%
Foundation	4.1544%
Depreciation	0%
PropertyCenterPoint	2.0033%

Based on the missing percentage for each column, we will delete the entire column named ‘Unit’ since this column lost over 80% of data, it will have little use in analysis. After the deletion of the ‘Unit’ column, we will delete rows which contain missing data since other columns do not contain large amounts of missing values.

3.Non-numeric data

We will convert categorical data to numeric data by transforming a category to the number of times the category appears in its corresponding column. This frequency method values the importance of categories by their appearing frequencies. More analytically reasonable encoding methods such as OneHotEncoding could be performed when machine learning comes into play. The data in columns ‘StreetName’, ‘LandUse’, ‘BuildingType’, ’HeatFuel’, ‘HeatType’, ‘ZoningCode’ and ‘Foundation’ are converted to numerical values by this frequency method. The column ‘Grade’ is converted to 0-19 based on what the author believes to be a reasonable rating system (Poor<Fair<Average<Good<VeryGood<Custom<Excellent). The column ‘PropertyCenterPoint’ is separated into two columns ‘PropertyCenterPoint_x’ and ‘PropertyCenterPoint_y’ which store latitudes and longitudes.

4.Suspicious data

The data pertained to each column is checked. Rows that contain exceptional values are deleted. The existence of exceptionally high or low values in numeric-typed columns and infrequent categories in categorical-typed columns could be due to data input errors or if not, due to exceptional factors which cannot be measured easily. We will avoid to use these unusual data in the development of machine learning models.

Since the goal of the project is to estimate values/prices of RESIDENTIAL housing properties, rows with non-residential land use purpose and building types are deleted.

The relation CurrentValue = CurrentLandValue + CurrentYardItemsValue +CurrentBuildingValue has been found. One record(row) violates this equation and hence be deleted.

FinishedArea is proportional to TotalGrossArea. No record has an unusual high/low FinishedArea/TotalGrossArea quotient. No record is dropped based on this criterion.

SalePrice is proportional to CurrentValue. 13 records have unusual high/low sale prices, to be more specific, unusual high/low SalePrice/CurrentValue quotients, and are hence deleted from the data set. These records are dropped since an unusual high/low sale price is very likely due to buyers/sellers' unusual financial situation and the unusual housing market which we have little or no measure of based on the original data set. We will not take this unusual data points to build our predictive models.

After the data cleaning, the data to be used contains 4308 rows.

- # Exploratory Data Analysis

The relationships among features(columns) are investigated. A preliminary data analysis based on visualizations and statistical inferences is conducted to answer the following questions. 1.How do CurrentValue and SalePrice depend on other features? 2. Are there any strong relations between pairs of features? The first question is basically the main question throughout this project. A preliminary answer is given. The question itself will be revisited and answered in more depth as the project develops. The second question is also important for the simplicity of predictive models.

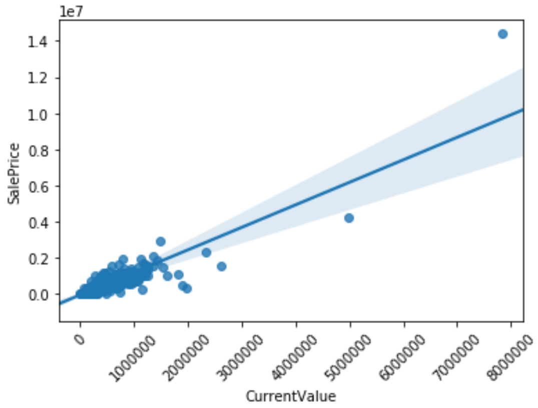
Some important or interesting findings are described below.

1.SalePrice is proportional to CurrentValue. They have a strong linear relationship.

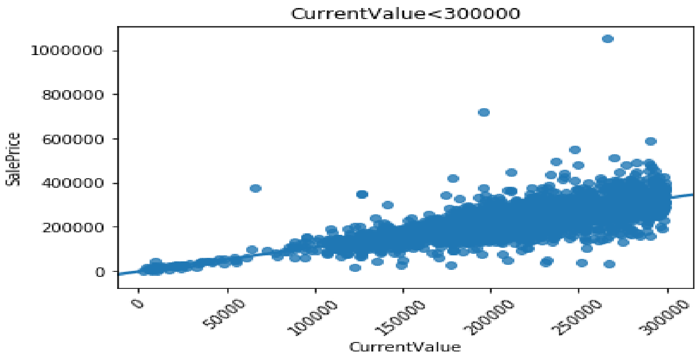
Statistics:

A simple linear regression (OLS) model with SalePrice being the dependent variable and CurrentValue being the independent variable gives R-squared: 0.809 and a zero p-value for the t-test.

Visualizations:



We will truncate the data set to preserve only the data with CurrentValue less than 300000, and fit the linear model again. The linear relationship is easier to visualize for the truncated data set as the following plot shows.

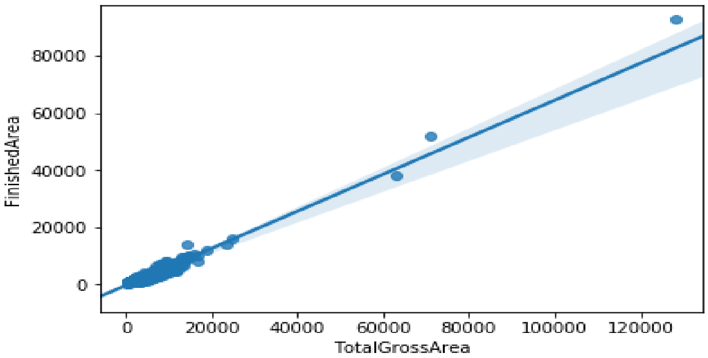


2.TotalGrossArea is proportional to FinishedArea. They have a strong linear relationship.

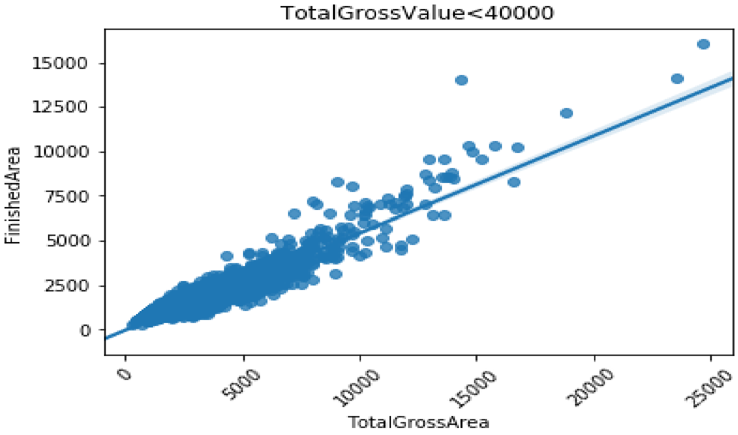
Statistics:

A simple linear regression (OLS) model with FinishedArea being the dependent variable and TotalGrossArea being the independent variable gives R-squared: 0.942 and a zero p-value for the t-test.

Visualizations:



We will truncate the data set to preserve only the data with TotalGrossValue less than 40000, and fit the linear model again. The linear relationship is easier to visualize for the truncated data set as the following plot shows.

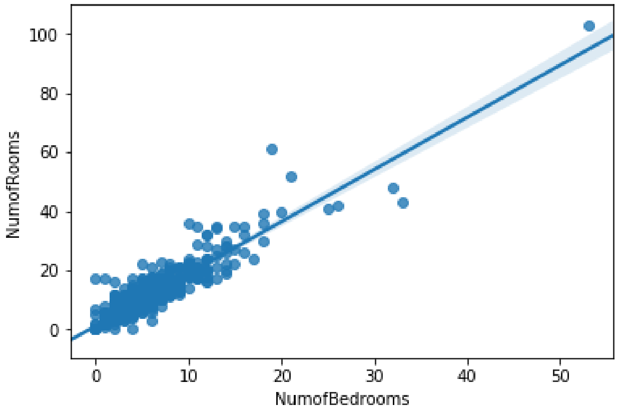


3.NumofRooms is proportional to NumofBedrooms. They have a strong linear relationship. This could mean most of rooms are bedrooms in a large portion of housing properties.

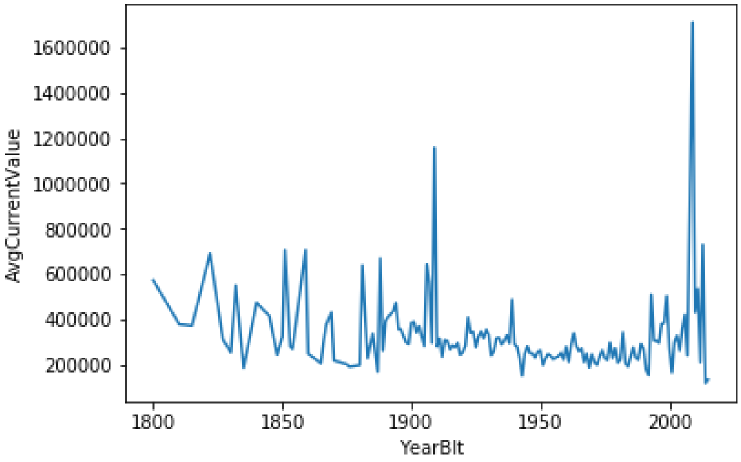
Statistics:

A simple linear regression (OLS) model with NumofRooms being the dependent variable and NumofBedrooms being the independent variable gives R-squared: 0.827 and a zero p-value for the t-test.

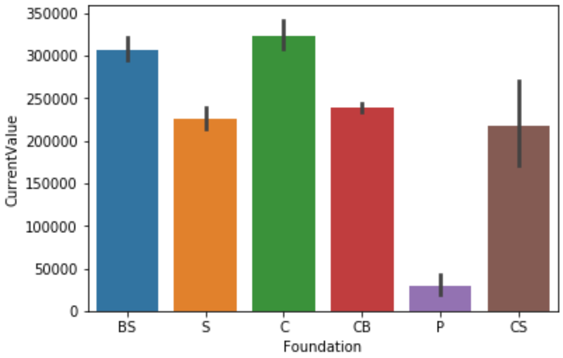
Visualizations:



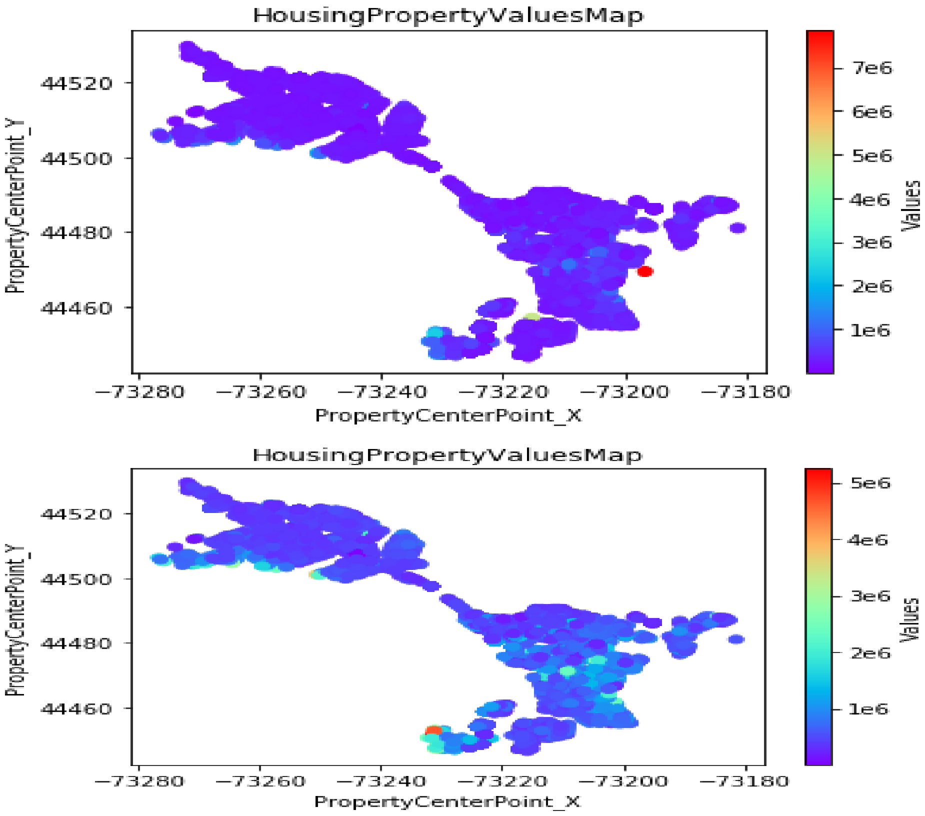
4.The average current values do not have large fluctuations for the housing properties built in the 20th century. However, properties built in the early 20th century or early 21st century might have high values.



5.The average current value of housing properties with the ‘P’ foundation tend to be lower than the average current values of housing properties with other foundations.



6.The current values of housing properties in the South part of the city tend to be lower than the average current values of housing properties in the North part of the city. The highest value has been found in the South part of the city though. The first plot is based on the full cleaned data set. The second plot truncates the cleaned data set to only preserve the data with CurrentValue<4000000.


HousingPropertyValuesMap
PropertyCenterPoint_Y
PropertyCenterPoint_X
7e6
6e6
5e6
4e6
3e6
2e6
1e6
Values
HousingPropertyValuesMap
PropertyCenterPoint_Y
PropertyCenterPoint_X
5e6
4e6
3e6
2e6
1e6
Values