353 Project Report

Khushwant Singh Parmar <ksparmar>
Course Info: CMPT 353 - Computational Data Science, School of Computing Science, SFU
Instructor: Professor Greg Baker

Problem statement:

Vancouver's real estate market has been a hot topic in recent years. Through this project I aim to do an analysis of the market trends, gauge the impact of covid-19 pandemic on the property prices and possibly conclude the most favourable property types for investment purposes. Moreover, what better than a tool that can accurately predict property prices based on neighbourhood and old prices, this is yet another aspect I wish to cover.

Questions:

- Using the location,previous prices, year built .etc. can we predict current prices?
- Using the prices,year built, tax values can we predict the neighbourhood?
- What neighbourhoods experience the most increase in prices?
- What is the overall trend of price change among houses of different zoning types?
- Did one-family and two-family zoning properties experience a similar change in prices?

Data

Acquire

The data used in the project was acquired from the City of Vancouver open data portal. I specifically chose property tax reports for the years 2018,19 and 20. These were downloaded as excel files. Different versions of the original data are used in different files of the project.

Cleaning

The data though fair clean needed some modifications for the purpose of this project. The whole cleaning process was done using Pandas.
Firstly, the columns required were chosen, then a threshold value for property prices $30mil was imposed. The PIDs for the properties were string to begin with, they were converted to integers and zoning classification values were encoded using numbers 1-5 Price increase from the last year was also calculated for each property.
For the regression models and ML prediction models only Single-Family Dwelling, Two-Family Dwelling and Multiple Dwelling zoning type properties were used.
Based on the zipcode of the properties their neighbourhoods were decided and added to the data.

Loading

The modified data was saved using .csv format. Records from 2018,19 and 20 were concatenated to be used for regression and classifications models.
The records from the 3 years were merged to be used for statistical analysis.

Regression

Can we predict prices?

Using the following features about a property
'LAND_COORDINATE','ZONING_NUMR','CURRENT_IMPROVEMENT_VALUE'
,'PREVIOUS_IMPROVEMENT_VALUE','PREVIOUS_LAND_VALUE',
'YEAR_BUILT','TAX_LEVY','NEIGHBOURHOOD_CODE'
Can we predict the current land value of the property?

Data used

The concatenated data file of the property tax report was used to train the regression models.
The data was split using train_test_split into 80,20 sizes.
The values were scaled using Standard Scaler and I compared 4 regression models for price prediction comparison.
8 features were used to predict the Current Land Value of property.
The models used and their training and validation scores are as follows:

Regression Model	Validation Score	Training Score
KNeighbors Regressor	0.9693	0.9852
Random Forest Regressor	0.9638	0.9714
Gradient Boosting Regressor	0.9679	0.9779
MLP Regressor	0.9653	0.9708

MLPrediction

Can we predict neighbourhoods?

Using the following features about a property
'LAND_COORDINATE','ZONING_NUMR','CURRENT_IMPROVEMENT_VALUE'
,'PREVIOUS_IMPROVEMENT_VALUE','PREVIOUS_LAND_VALUE','CURRENT_LAND
_VALUE',
'YEAR_BUILT','TAX_LEVY','NEIGHBOURHOOD_CODE'
Can we predict what neighbourhood the property is located in ?

Data used

The models were trained on the property tax report data for 3 years, and the neighbourhood were identified based on property zipcode. The values were scaled using Standard Scaler and I compared 4 regression models for price prediction comparison.
8 features were used to predict the neighbourhood of the property.

The models used and their training and validation scores are as follows:

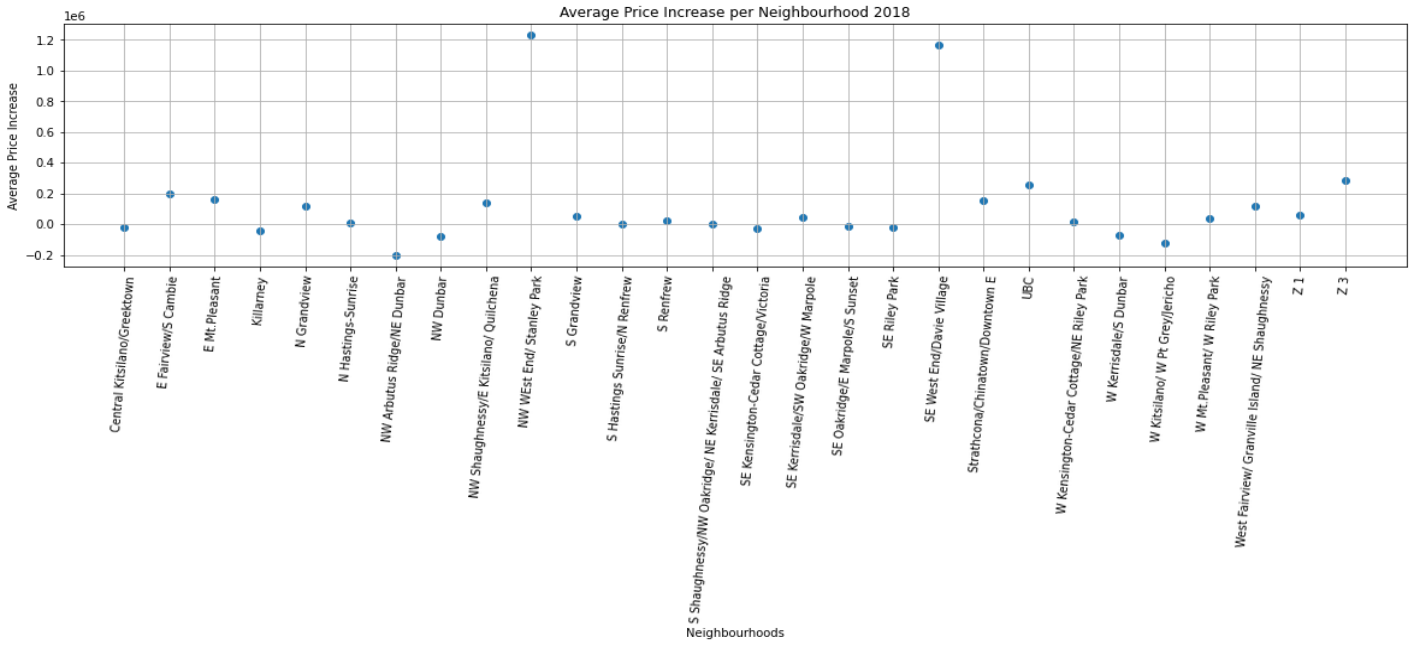
Classification Model	Validation Score	Training Score
GuassianNB Classifier	0.4766	0.4734
KNeighbours Classifier	0.8295	0.8632
Random Forest Classifier	0.8772	0.8826
Self Training Classifier	0.8472	0.8464

Statistics and Findings

What neighbourhoods experience the most increase in prices?

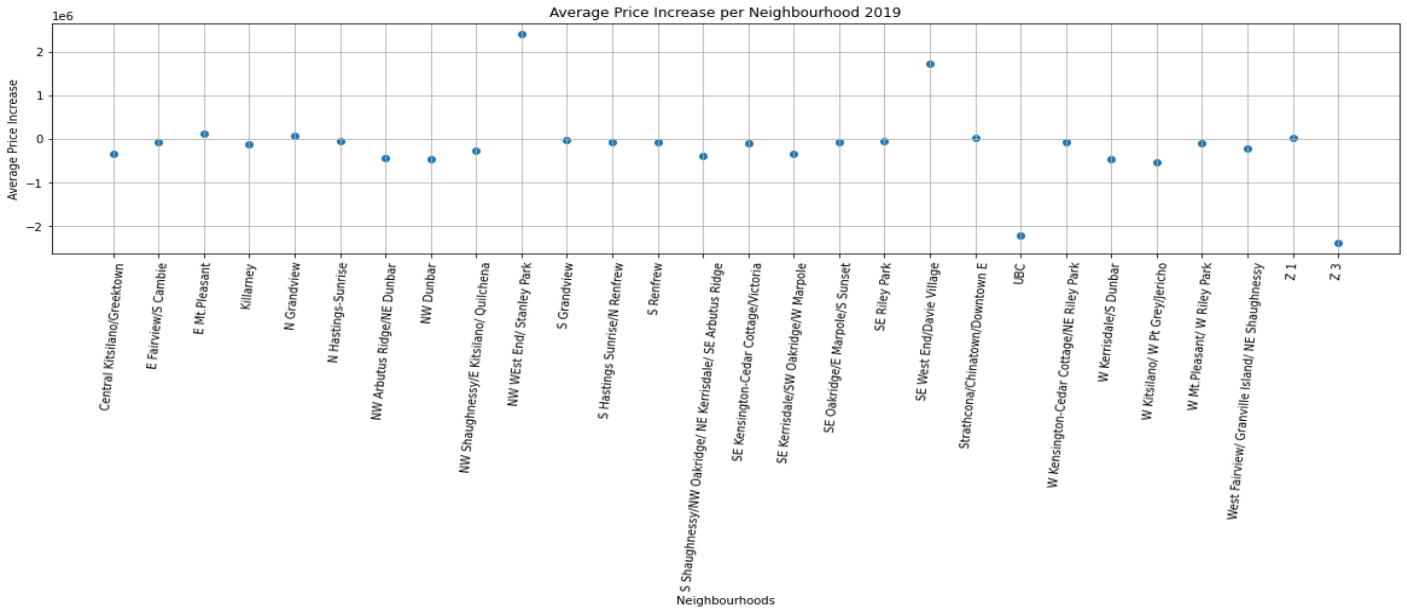
2018: NE West End/ Stanley Park, SE West End/ Davie Village
2019: NE West End/ Stanley Park
2020: Strathcona/ Chinatown/ Downtown E
Overall trend over the period of 3 years in different areas: the property prices went down from 2018-2020. 2019 didn't see a significant drop but 2020 saw a negative increase in the prices of the properties all over the city of Vancouver.

2018 Trend

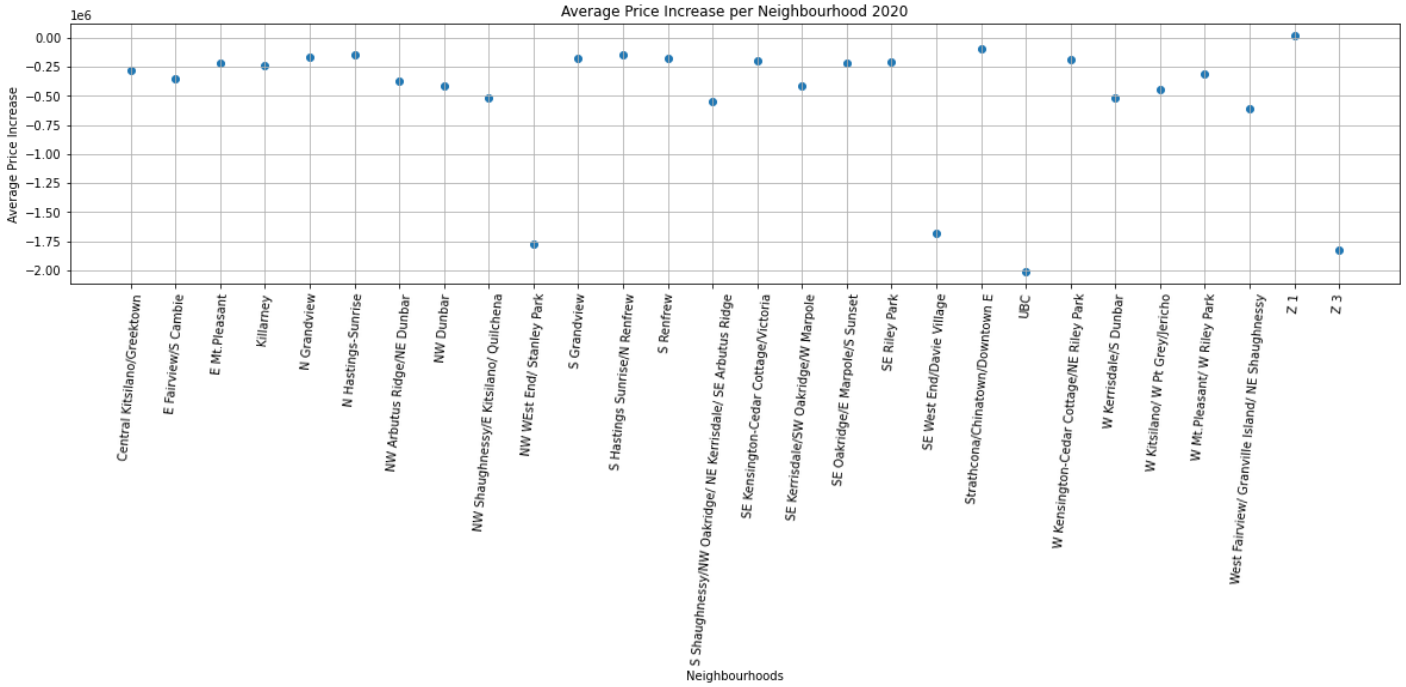


.

2019 Trend



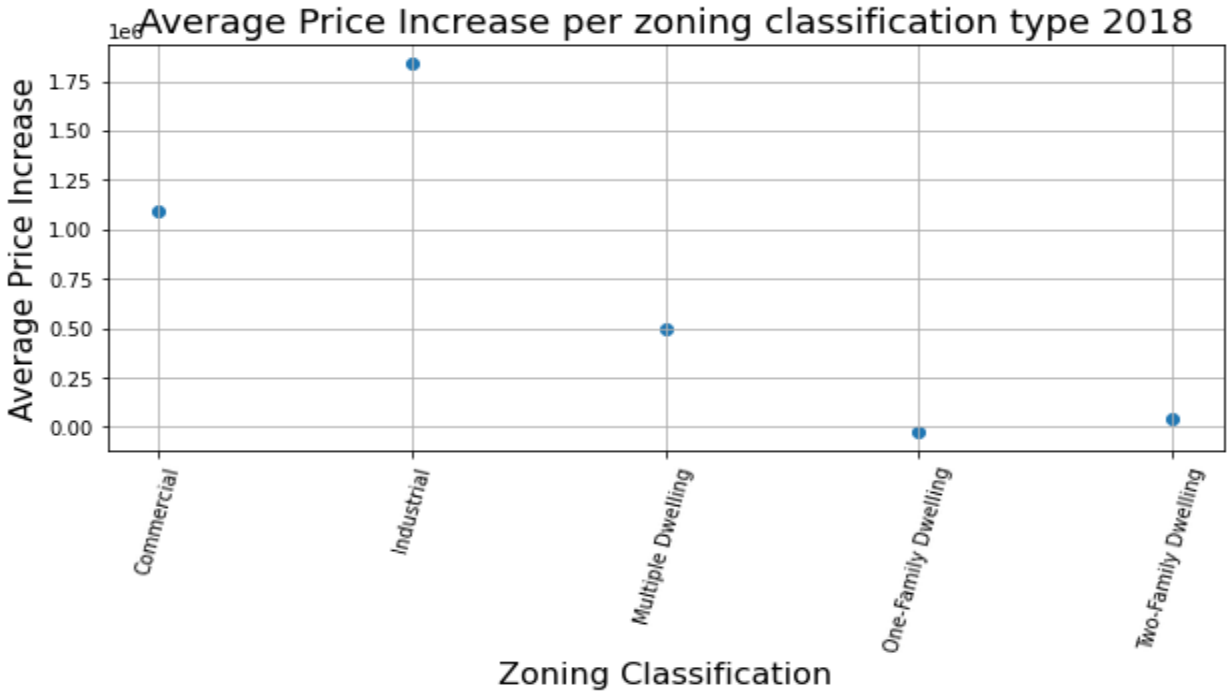
2020 Trend

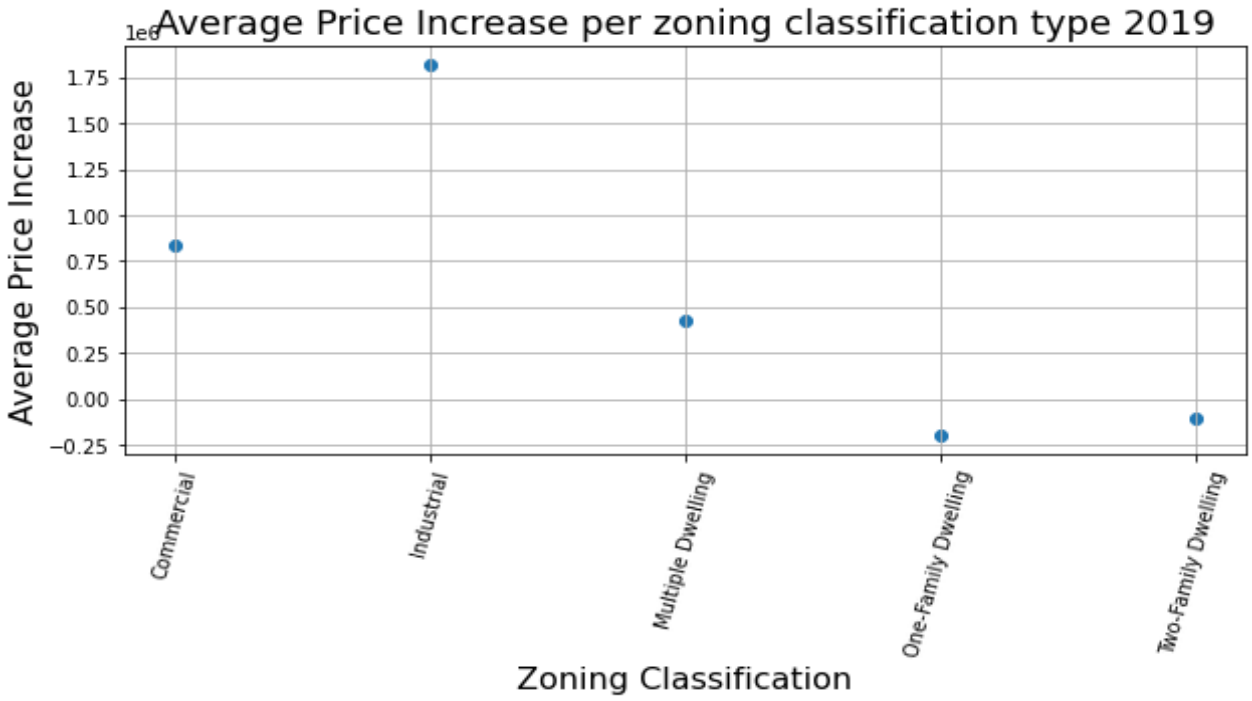


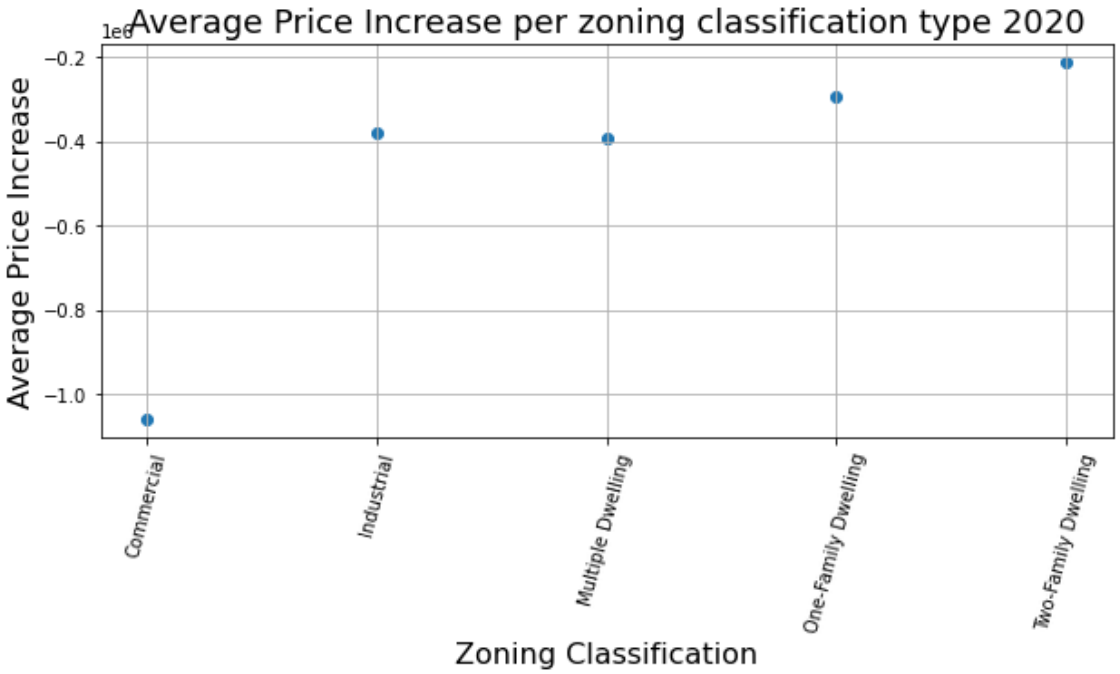
Note the heavy negative price change in all regions during the covid 19 pandemic.

What is the overall trend of price change among houses of different zoning types?

Years 2018-19 saw strikingly similar price increases for different types of properties, with the maximum increase being for industrial properties and minimum being for one-family dwellings. All the types of properties saw a positive increase in their prices. In the year 2020 the trends reversed, prices dropped rapidly with commercial properties taking the most hit and two-family dwellings taking the least. For all types of properties prices dropped.







Businesses being closed due to pandemic resulted in huge drops in prices of commercial properties.

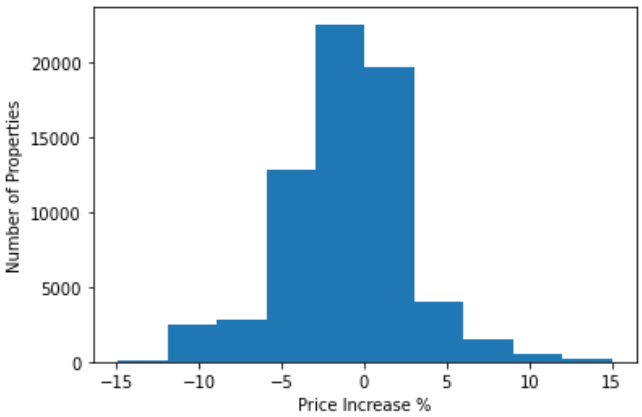
Did one family and two family zoning properties experience a similar change in price?

Mann whitney test was performed to see if the prices of one family and two family dwelling had similar changes in price or not.

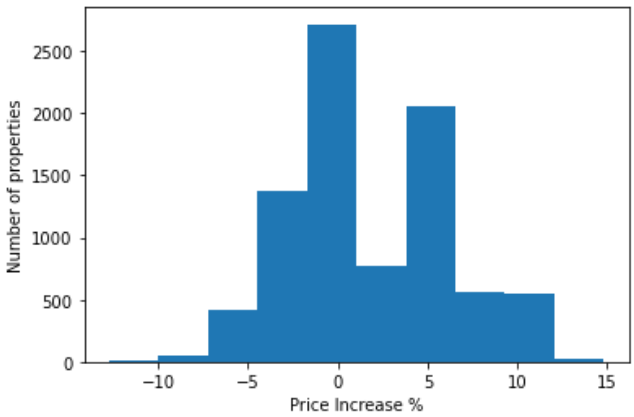
For all 3 years the u-test p value turned out very small showing that the values are evenly matched.

From the histograms it is clear that the spread of values is different but the majority of values overlap in the same price increase percentage ranges.

For the Year 2018



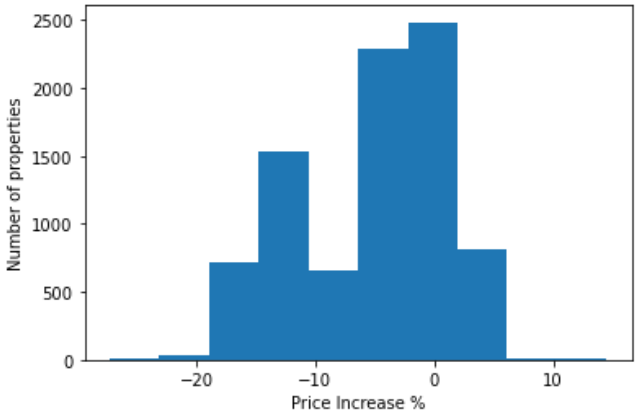
One Family Dwelling



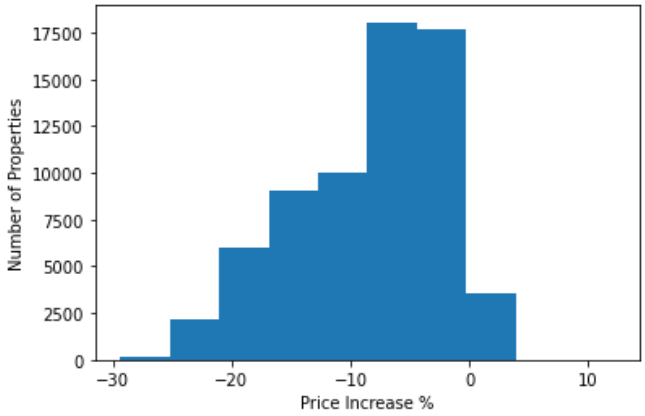
Two Family Dwelling

MannwhitneyuResult(statistic=53791.0, pvalue=1.1198017692142477e-15)

For the year 2019



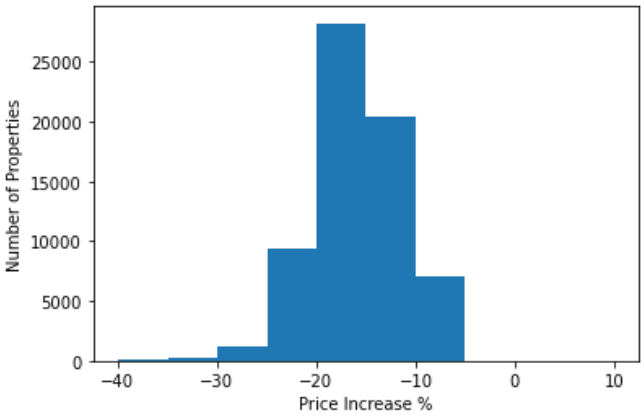
One Family Dwelling



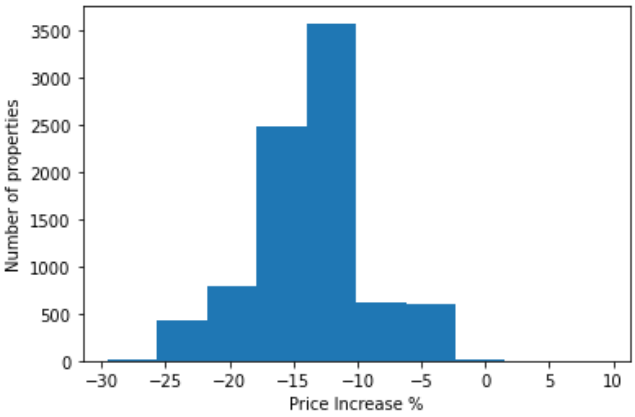
Two Family Dwelling

MannwhitneyuResult(statistic=56637.5, pvalue=8.760900304381514e-13)

For the year 2020



One Family Dwelling



Two Family Dwelling

MannwhitneyuResult(statistic=54256.5, pvalue=3.506441677843912e-15)

Conclusion:

- Prices of properties can be predicted using data about age of property, its location and its age.
- Neighbourhood of a property can be predicted using data about prices of property, age of property, tax percentages.etc.
- Industrial and commercial properties were most sensitive to price changes. They had the most increase during 2018-2019 and the greatest drops during the pandemic of 2020.
- Similarly NE West End/Stanley park and SE West End/Davie Village neighbourhoods were the most affected by the changing trends.

Limitations:

- Problems:
    - The project uses data with a great number of entries making it hard to train more complex models, thus, I have used simple models.
    - The data couldn't be filtered using Kalman or LOWESS filtering algorithms due to static data entries.

Project Experience Summary

- Acquired data and performed ETL steps on it using Pandas and Numpy libraries.
- Built regression models(KNN, Random Forest, Gradient Boosting, MLP) to predict price of property listing based on a number of features.
- Built classification models(GaussianNB, KNN, Random Forest, Self Training Classifier) to predict property listing neighbourhood based on a number of features.
- Performed statistical analysis on data and built visualization using matplotlib.pyplot:
    - Calculated and plotted price increase for each neighbourhood over 2018,2019 and 2020
    - Calculated and plotted price increase for each zoning classification over 2018,19 and 20
    - Compared price increase percentage of one-family and two-family type dwellings by using Mann Whitney U-test on population samples