Python Module Project:

<u>House Sales Project</u>

Spencer Fairbairn, Allison Bruhn, Portia Gifford,

Chidi Nzerem, Jacob Pulitzer, William Tardif

MSBC 5070, Section 203

Micah McGee

UNIVERSITY OF COLORADO
LET YOUR LIGHT SHINE
1876

Business Problem Statement

The housing market in Washington has become fiercely competitive, and to remain competitive, our wholesale real estate investment firm is seeking to understand how home features relate to the adjusted sales price in order to outsell the competition and maximize our profits. Using descriptive and predictive analytics we set out to understand how home features are correlated to the overall sales price. By understanding how different features bring value to Washington homes, we are able to maximize the returns on our investment properties.

Research Question

How do we improve home and land features of an investment property to increase the returns on our investment.

Target Feature

We want to be able to predict the sale price of future houses based on each feature and be able to assess what features have the greatest impact.

Descriptive Analysis:

The dependent variable used in this data set was the adjusted price. After performing descriptive analytics on each of the following questions below, our findings include:

1. ***What percentage of houses in our data set sold over a million dollars? What is the average, sq ft living, bathrooms, bedrooms and building grade of those houses?***
    a. The percent of houses sold over a million was 7.16%.
    b. The average square footage to living was 3802 ft.
    c. The average bedroom was 4 and the average bathroom was 3.
2. ***What was the average Market Predictor Scale(MPS)? What percent of the MPS was negative?***
    a. The average housing MPS is 44433 dollars.
    b. The percentage of houses with a negative MPS 23.80.
3. ***What are the top 3 Zip Codes with the highest average adjusted price***
    a. The zip codes with the top 3 highest average adjusted prices are:
        i. Zip Code: 98039
        ii. Zip Code: 98004
        iii. Zip Code: 98040
4. ***What is the difference in average adj prices in newly renovated and non-renovated properties?***
    a. When looking at properties that have been recently renovated we can see that there is an average increase in property value of $76,864.98 from houses that have not been recently renovated.

Based on our descriptive analysis, we were able to pinpoint from the data the houses within zip code areas that would provide our clients with the most profitable sales prices. The highest selling area code is 98039. We were also able to give the projected averages of square footage for houses that were sold for over a million dollars which was 3802 sqft. as well as the number of beds and bathrooms which were 4 and 3. The data also allowed us to find MPS of the house, which had an average of $44,433 dollars. It also allowed us to find houses with negative MPS values, also showing that 24% of the houses had negative MPS values.

Predictive Analytics:

How can we best predict the adjusted sale price of a house, taking into account all important features?

In order to answer this question we ran a regression with adjusted sales price as our dependent variable and all other features as independent variables. These variables included; year built, number of bathrooms, year renovated etc. By doing this we can understand if these features truly had an impact on the selling price of a house.

After running this regression, our results showed:

R squared: 0.8529

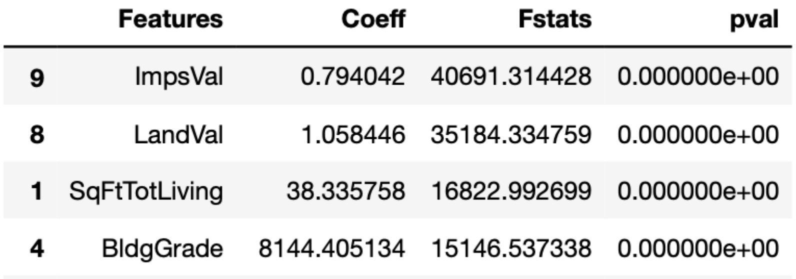
The R-squared value for this model is 0.8529, meaning that approximately 85% of the variance in adjusted sales price can be explained by the variance in all features observed by our model.

Within our regression model, along with R squared, we looked at three different metrics. These metrics included the mean absolute error (MAE), mean squared error (MSE), and the root mean squared error (RMSE). The main metric we wanted to focus on was the **RMSE** because it informs us how much our model is off numerically from the actual values. The result from our regression was:

RMSE: 147601.33

What this is saying is that based on our regression model, we will be off by about $147,601 when predicting the selling price of a home. To put that number into a percentage, by taking the RMSE value then dividing that number by the average selling price of all of the houses within the dataset, our estimates for pricing are off by ~25% on average.

Below is a list of the four most important features to the target variable which was determined by the F-statistics. The Fstat shows how much of a correlation there is to the dependent variable which means the higher the Fstat, the more of a correlation there is. Essentially, the four features listed below have the highest impact on our adjusted sales price. The coefficients represent how each feature will increase for every $1 increase in adjusted sale price. The p-value is another way to determine how confident we can be about our model. A lower p-value means there is a higher correlation and therefore we can have higher confidence in our model.



	Features	Coeff	Fstats	pval
9	ImpsVal	0.794042	40691.314428	0.000000e+00
8	LandVal	1.058446	35184.334759	0.000000e+00
1	SqFtTotLiving	38.335758	16822.992699	0.000000e+00
4	BldgGrade	8144.405134	15146.537338	0.000000e+00

Recommended Business Actions

Based on our analysis we can now see certain aspects and features can have a significant impact on the selling price of home. From this analysis we are able to buy and sell smarter and faster as well as flip homes more efficiently.

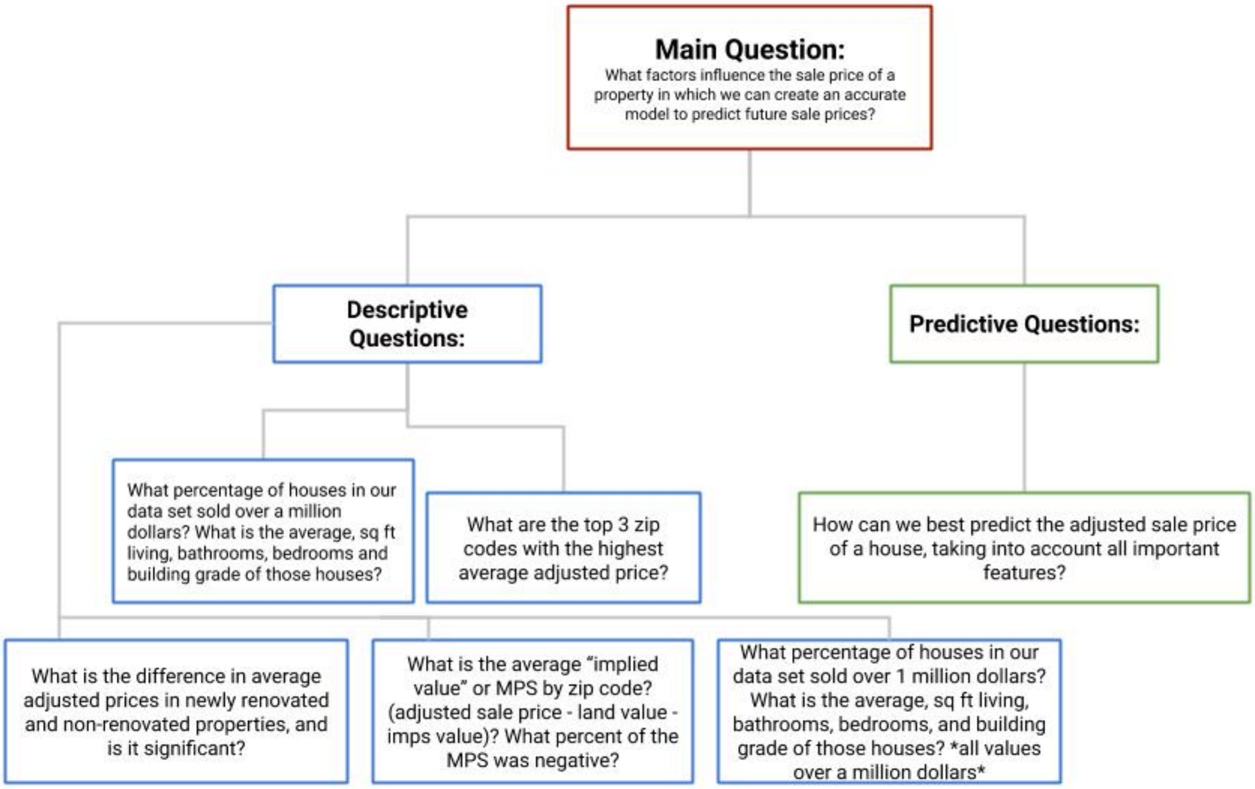
Some specific business actions that we can take would be:

- Focus our buy/selling/ flipping on zip codes and locations with the highest adj sales price.
- Continue to invest our profits into renovating homes to add more desirable features.

Next steps

We want to increase our RMSE so that our predictions are only off by an average of ~10% meaning we would be 90% accurate in our predictions. To do this we would need to take into account more features to include what renovations were done. From this we could accurately predict the sales price based on specific renovations. This will in turn create more profits and an increase in return on investment.

Question Tree:



Main Question:

What factors influence the sale price of a property in which we can create an accurate model to predict future sale prices?

Descriptive Questions:

What percentage of houses in our data set sold over a million dollars? What is the average, sq ft living, bathrooms, bedrooms and building grade of those houses?

What are the top 3 zip codes with the highest average adjusted price?

What is the difference in average adjusted prices in newly renovated and non-renovated properties, and is it significant?

What is the average "implied value" or MPS by zip code? (adjusted sale price - land value - imps value)? What percent of the MPS was negative?

What percentage of houses in our data set sold over 1 million dollars? What is the average, sq ft living, bathrooms, bedrooms, and building grade of those houses? *all values over a million dollars*

Predictive Questions:

How can we best predict the adjusted sale price of a house, taking into account all important features?