View license

0 stars 0 forks 1 watching Activity

Public repository

master

Branches Tags

bennettandrewm Update README.md … 2 hours ago 19

View code

README.md

Kings County Real Estate Analysis

Kings County - A beautiful place to live!

1. Project Overview

A Seattle real estate brokerage wants to expand their services to developers. They're offering "state of the art" linear regression data analysis to new developers in the area on where and what to build, as it relates to price. This repository includes, at a minimum, a multi-variable linear regression model to examine the relationship between square footage and zip code on sales price - the target variable. They're are also additional property factors included in the analysis.

The Data

Source Data

This project uses the King County House Sales dataset, which can be found in `kc_house_data.csv` in the `data` folder in this GitHub repository. The description of the column names can be found in `column_names.md` in the same folder. The data set consists of characterists about houses that have been sold in Kings County for fiscal year 2021-2022. They contain about 25 data columns including

- `price`
- `sqft_living`
- `addess`
- `grade`
- `condition`
- `date`
- `view`

A complete description of all the variables can be found in `column_names.md` in the `data` folder

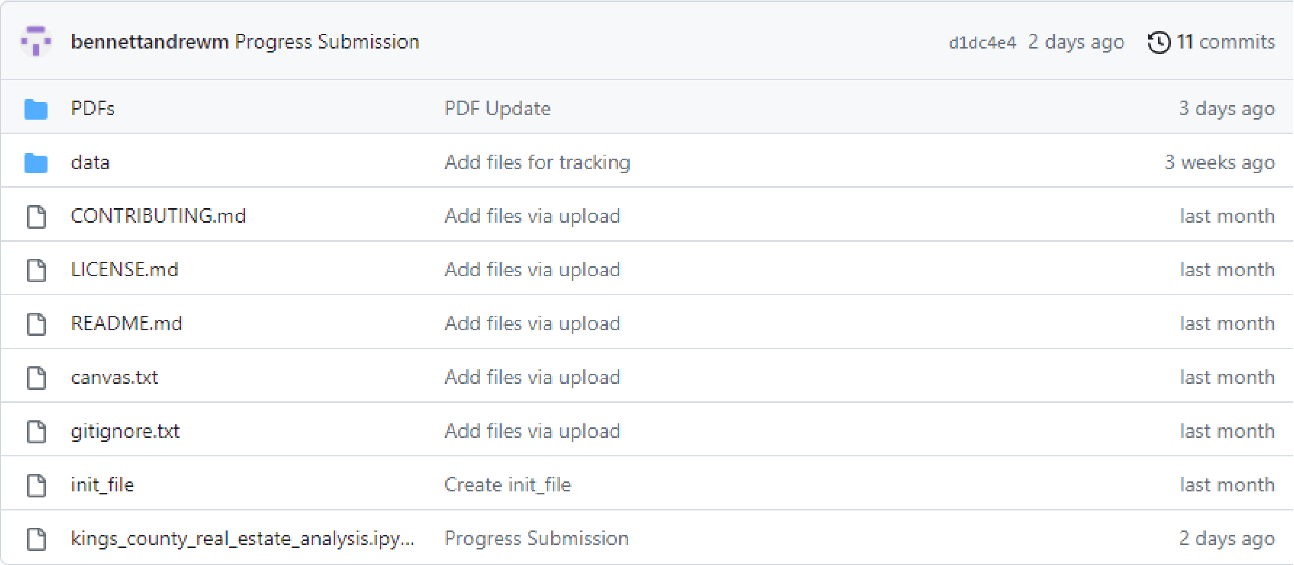
Data Modeling Approach

In order to aid real estate developers in their quest to know what and where to build, the first step is to focus on the size and location of the home. Anecodotally, we know that builders and buyers often think about homes in terms of price per square foot, a simple linear expression. We also know that this increases or decreases by neighborhood. A linear regression model incorporating both of these factors is appropriate for the analysis we're providing.

We will achieve this with our linear regression models using an iterative approach. We will start with a simple, single variable model and build more complexity with additional variables. The sales price would be our target, or depedent variable. From there, we quantify the impact of other features (bedrooms, baths, or condition) on the price.

Github Repository

To execute this project, a github repository is utilized for public viewing and collaboration



bennettandrewm Progress Submission d1dc4e4 2 days ago 11 commits

PDFs	PDF Update	3 days ago
data	Add files for tracking	3 weeks ago
CONTRIBUTING.md	Add files via upload	last month
LICENSE.md	Add files via upload	last month
README.md	Add files via upload	last month
canvas.txt	Add files via upload	last month
gitignore.txt	Add files via upload	last month
init_file	Create init_file	last month
kings_county_real_estate_analysis.ipy...	Progress Submission	2 days ago

You can see the following files stored in the github repository.

PDFs - Folder containing pdf files below * `Kings County Real Estate Presentation` - Non-technical presentation of the Analysis * `Kings County Real Estate Analysis` - complete Jupyter Python Notebook in pdf form * `README` - README file converted to PDF

Images - Folder containing the image files used in the Notebook, Presentation, and README file

data - Folder containg the source data files * `kc_house_data.csv` - Kings County Housing Data Set * `column_names.md` - file explaining the column names

Contributing.md - Folder containg instruction for how to contribute to Learn.co

README.md - the currently file you're reading with descriptions about the coding file

canvas.txt - a Canvas text file

gitignore.txt - git ignore file

kings_county_real_estate_analysis.ipynb - Notebook with Python analysis

2. Data Import, Cleaning, and Engineering

Data Import

To get started we import the CSV file in a pandas dataframe. It's got 30,155 entries with 25 columns. We see both numeric continuous, numeric discrete, and categorical data.

Data Cleaning

Next we removed or clean a variety of data points. This includes handling null values as well as data the doesn't seem right. We consider square footage, lot square footage, above ground square footage, bedroom and bathroom combos, and addresses that do not appear to have a Kings County Zipcode.

Drop Null Values (removed 44 entries)

Of the 30,155 entries, there were 44 entries with NUll values in any of the columns. Because it's such a small percentage (< .01%) we'll discard these rows.

Small Square Footages (removed 1 entry)

Of the now 30,111 entries, we have 1 entry where `sqft_living` is less than 100 sq. ft. According to the listing, it's a 4 bed, 4 bath house of 3 sq.ft. house. This isn't right. We can delete that.

Samll Lot Square Footages (removed 0 entries)

We inspected homes with `sqft_lot` less than 500 sq. ft. It turns out there are quite a few homes on small lots but many are multistory. These entries look real. They could be small row houses, condos, or something multi-story not requiring large lots. We'll leave these entries.

Small Above Square Footages (removed 0 entries)

We inspected homes with `sqft_above` less than 300 sq. ft. There are 4 entries and and they look okay. Probably a small studio apartment or something. We removed none.

Bedrooms and Bathrooms - Less than 1 (removed 21 entries)

We inspected homes with less than 1 bed and 1 bath to see if these looked suspicious. There were 21 entries. Based on the `column_names` document. To make sure these weren't cabins, we also checked there `grade` cateogry and realized they were mostely all average to good (7-8). This seemed suspicious so they were deleted.

Addresses with "Washington" (removed 902 entries)

Because we do care about zipcodes and neighborhoods, we want to be sure that our data is based on Kings County info. To do this, we created a function to check to see if the word "Washington" appeared in the address. And, there are apparently 902 entries that do not contain that word. A spot check on these reveals data from out of state. So... we deleted this as well.

So, we are now down to 29187 from 30155. We deleted 968 entries, the majority of those from non-Washington state entries.

Feature Engineering

ZipCode

Because we do care about zip codes, we want to extract the zipcodes from the address and make a new column with justt those 5 digits as integers (later, we will use one-hot encoding to convert these to columns). So we made a function that will extract the five digits of whereever the numbers of the zipcode are located in the address. We then created a new column called `zipcode` . We have 77 different zipcodes.

Grade

The grade column will be considered later. This is currently a string containing a number 1-12 and a qualititative assessment, ("7 average", for ex). We will convert this to just an integer of the number. So from "7 average" to "7". This column is called `intgrade`

Year

It's interesting to understand how the age of the house effects price. For our clients, they may want to know if there's an extra boost for a new or recent construction. To make a potential linear relationship more clear, we will subtract this number by 1900. So the range will be 0-120 in terms of recency. A new `yr_built_transform` column will be created for this.

Sales Month

Sometimes we here that spring is a popular time to sell homes, but does that translate to price? To check this, let's add a column called `sales_month`. Before we do that, we check the number of years and realize... ah. This is just 2021-222 data.

But anyhoo, we add a column called `sales_month` which includes just a number for the month.

Outliers Reduction

Now that we have clean data, let's start looking at some of the outlier information.

Mansions - Large Homes (288 entries removed)

A quick inspection of the `sqft_living` data reveals a slight left_skew, meaning a tail extending to the larger square foot homes. The 99th percentile home is 5190 sq. ft. The 75th percentil home is 2640 sq.ft. This is a big difference, so we removed them from our set. This reduces our entries from 29,187 to 28,899, or 288 entries.

Pricey Homes (173 entries removed)

A quick inspection of the `price` data reveals a slight left_skew, meaning a tail extending to the more expensive homes. The 99th percentile home is 4.3M. The 75th percentil home is 1.3M. We'll go ahead and get rid of the 1% that's higher than 4.3M. This reduces our entries from 28,899 to 28,726, or 173 entries.

3. Data Modeling

With what we believe to be clean data, we can begin to model.

The approach is to begin with a simple, basic linear regression model to describe sq.ft vs price. From here, we will then consider zipcode. The zipcodes will be one-hot encoded, which means all 77 zipcodes will be given a column in the dataframe. 1 of 77 entry will be given a 1 to represent a house in that zipcode. The other zipcodes will be 0. After that model, we included additional features to quantify their effect.

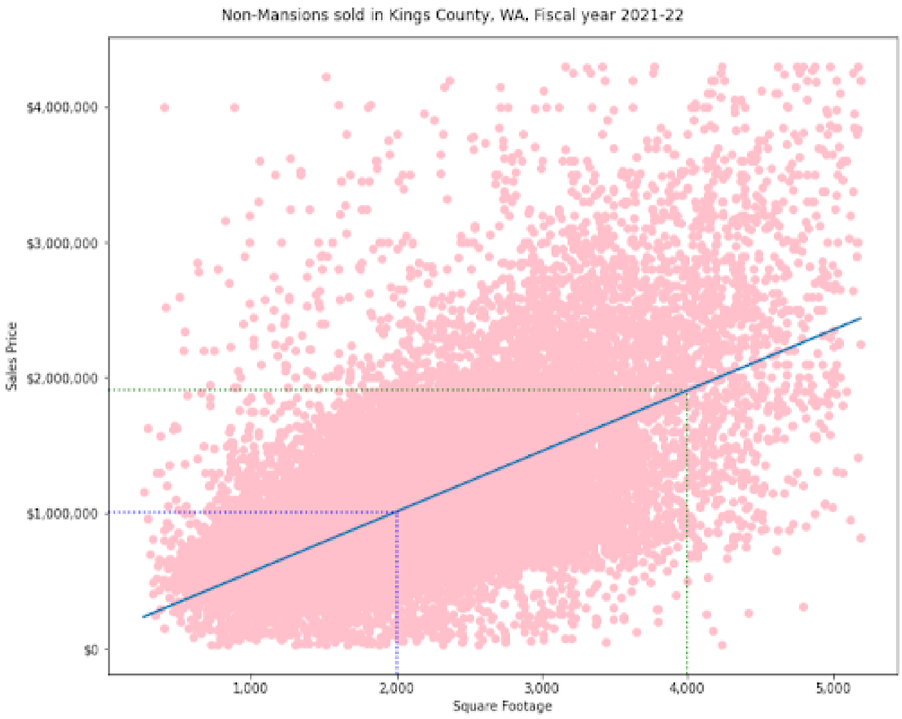
Baseline Model

Technique

So, first we're going to create a model using the Ordinary Linear Squares (OLS) method. The endogenous variable will be `price` and the exogenous, or independent variable, will be `sqft_living` .

Results

When we do this, we get a model that explains variance of 39.3% of the variance, but with high confidence that are coefficients are accurate. We get $447/ sq. ft. + $115,000. When we plot this line as a best_fit against all of the listings on a graph of `price` vs `sqft_living` we get the following graph



What we see here is interesting. Homes sell for almost $500/sq. ft. So a 2,000 sq. ft home sales for nearly 1M USD. A 4,000 sq.ft. home sales for nearly 2M USD. At a minimum, there seems to be a $115,000 premium for having a home in Kings County.

Zipcodes

Now that we have a simple model with a straightforward linear relationship, let's begin to make it more complex.

Technique

We're still going to use the OLS method, but we're going to add zipcodes to the set with one-hot encoding for zipcode data. We now have an exogenous set with roughly 78 variables. We would have more but we dropped zipcode 98070 from our analysis because it had the least impact and was thereby closer to the averge.

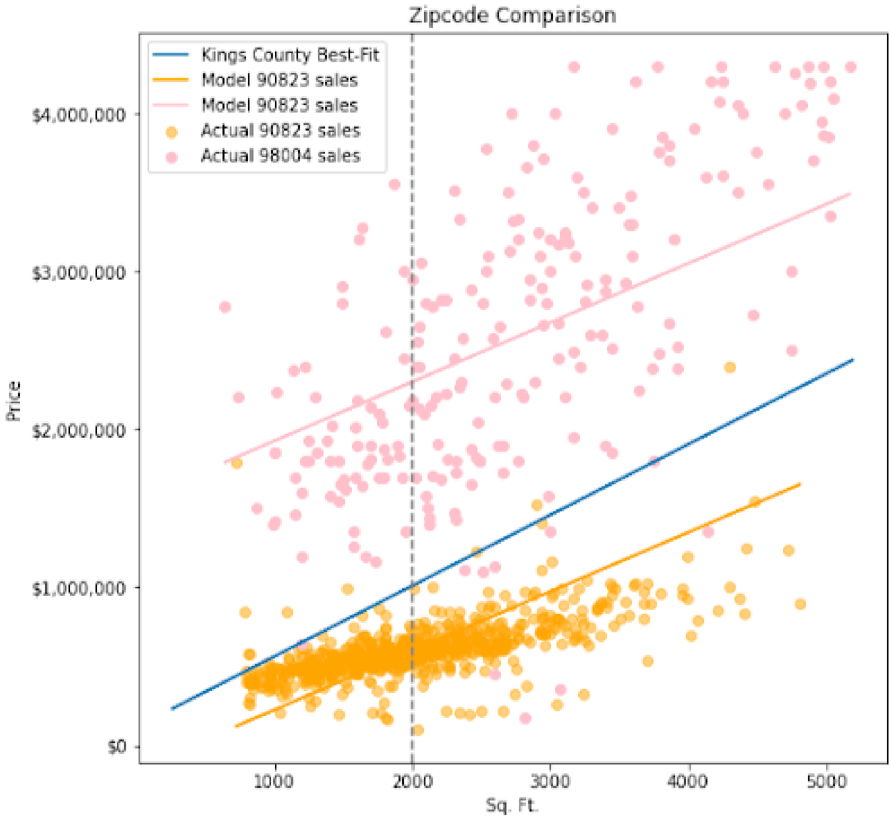
Results

What we found is a model accounting for 68.4% of all variance.

Using the standard alpha of 0.05 to evaluate statistical significance, Coefficients for sqft_living and most of our zipcodes are statistically significant. Our baseline zipcode is 98070. It seems that relative to our baseline, the zipcodes do have a statistically significant effect on price, except for zipcodes 98224, 98113, 98118, 98027, and 98011). So...

According to the model, houses are selling at approximately $374/sq. ft. with a coefficients for the intercept is $295,200. That means that, when not accounting for square footage or zipcode, you could assume a house will sell for $295,200.

But once you factor in the zipcodes, they do have an impact on price. In fact, our model shows noticeable impact on price. To illustrate this point we've graphed two high impact zipcodes.



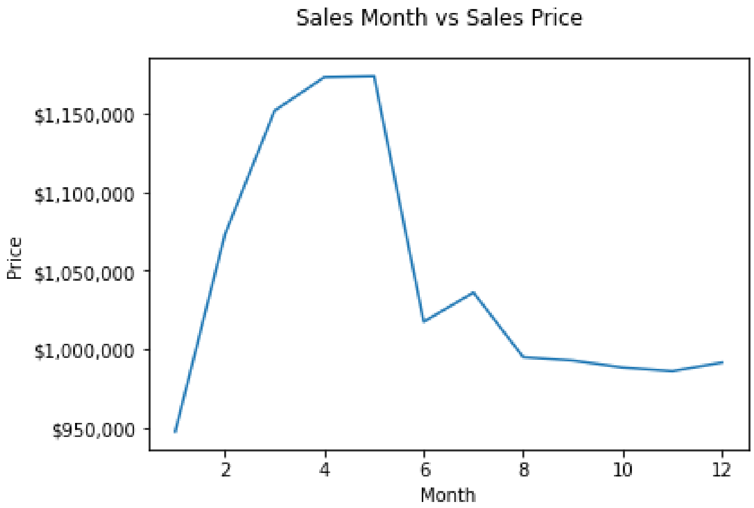
You can see that there is a significant difference between zip codes. 98004 homes are approximately 1.3M more than a typical zip code (in this case 98070). Zip code 98023 homes are approximately 445k less than the homes in 98070. We've also seen the price per square foot decrease. As we include the neigborhood, we see other things effect price. This makes sense

Additional Features

Now that we have a model that achieves what we set out to do. We're going to look at other factors like bathroom and bedroom to see any effect on price. We'll also look at other factors like views, conditions, grade, condition, and other perks to see if we can tease out any additional effects.

Closing Date

Realtors tell us there is a "season", usually in the spring, that's the best time to list a home. Let's see if that data plays out and if theres a linear relationship. When we plot the closing month vs the sales price, this is what we get.



So, the realtors are on to something... prices peak in the spring, when the sales close in April and May. This is not a linear relationship but fairly compelling data - nearly 200K difference between April and January. We can't include it as linear but I will include it categorically.

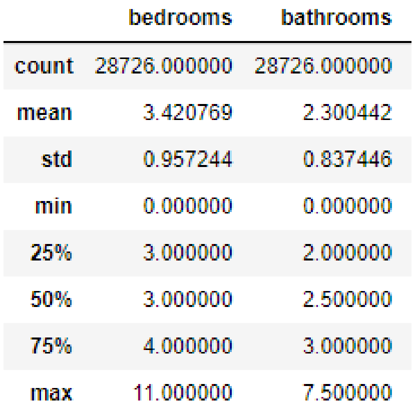
To include the sales data as categorical, each month is one hot encoded, and then included in the set for analysis.

Additional Numeric Variables - Not Considered in Model

To determine if other numerical variables should be added, the variables were plotted against eachother using an iterative loop, and really no linear relationship was found with potentially finding collinearity with `sqft_living` . The only variable that shows possible linearity, and it's faint, is the `yr_built` . This will be included in the model later.

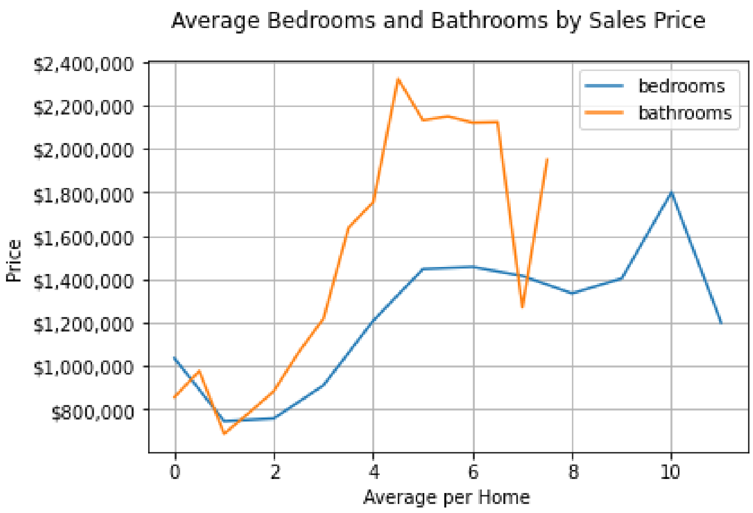
Bedrooms & Bathrooms - Not Considered in Model

To examine bedrooms and bathrooms, let's first look isolate bedrooms and bedrooms and their effect on price. We will do that with two charts. First, let's look at a few stats for the number of bedrooms and bathrooms built.



	bedrooms	bathrooms
count	28726.000000	28726.000000
mean	3.420769	2.300442
std	0.957244	0.837446
min	0.000000	0.000000
25%	3.000000	2.000000
50%	3.000000	2.500000
75%	4.000000	3.000000
max	11.000000	7.500000

We can see here that the median home has 3 beds and 2.5 baths, which is interesting. Let's look at the overall relationship between bedrooms and bathrooms and sales price.



Interesting. It looks like there isn't a linear relationship. In fact, the price seems to peak around 5 bedrooms and 4.5 baths. After this, the price either drops or flattens out. This doesn't look like a great candidate for a linear relationship but something to note - building more than 5 bedrooms or 4.5 baths may not result in a price increase.

Grade - Not Considered in Model

When examining the `grade` variable, which is essentially a rating 1-12 given by a potential buyer, it seemed to be linear with `sqft_living`. So we did not include it in the model.

Year Built - Added to Model

Based on the previous observation of the `yr_built` variable, the `yr_built_transfrom` was included in the model.

Condition - Added to Model

The condition of the house is based on the overall condition of the house, from a maintenance perspective. So, a small house could be in good condition. There are five ratings - "Poor", "Fair", "Average", "Good", and "Very Good". These categories were one-hot encoded, with "Average" Condition dropped from consideration.

View - Added to Model

The view associated with the house is self-explanatory. The categories are is based on the overall condition of the house, from a maintenance perspective. There are five ratings - "None", "Fair", "Average", "Good", and "Excellent". These categories were one-hot encoded, with "None" view dropped from consideration because this is the typical, or neutral case (ie, the majority of homes don't have a view specified).

Waterfront - Added to Model

The Waterfront variable specifies whether the home has any waterfront footage. The two options are "Yes" or "No". We will one-hot encode the variable and drop the `Waterfront_NO` , which is the typical condition.

Greenbelt - Added to Model

Greenbelt, just means undeveloped property adjacent to the property. The two options are "Yes" or "No". We will one-hot encode it and drop the `greenbelt_NO` , which is the typical condition.

4. Preliminary Results Summary

Now that we have completed the model, let's review some of the statistics.

Adjusted R-Square

Our Adjusted R-Square is now showing a value of 72.7% of all variance accounted for. This is better, but one wonders if 5% increase when accounting for all of these models is really worth it.

Statistical Significance

Using the standard alpha of 0.05 to evaluate statistical significance:

Coefficients for sqft_living and most of our zipcodes are statistically significant. Our baseline zipcode is 98070. It seems that relative to our baseline, the zipcodes do have a statistically significant effect on price, except for zipcodes 98014, 98019, 98045, 98050, 98056, 98059, 98108, 98118, 98126, 98133, 982242). So...

Intercepts and Variables

Our coefficient for the intercept is significanty significant for an alpha of .05.

According to the model, houses are selling at approximately $355/sq. ft.

The coefficients for the intercept is 51,930. That means that, when not accounting for square footage or zipcode, you could assume a house will sell for 51,930.

Impact

The zipcode with the largest effect is zipcode 98004, 98005, 98039, and 98040. Zipcode 98010, 98001, 98003, 98023, have the most negative effect on pricing.

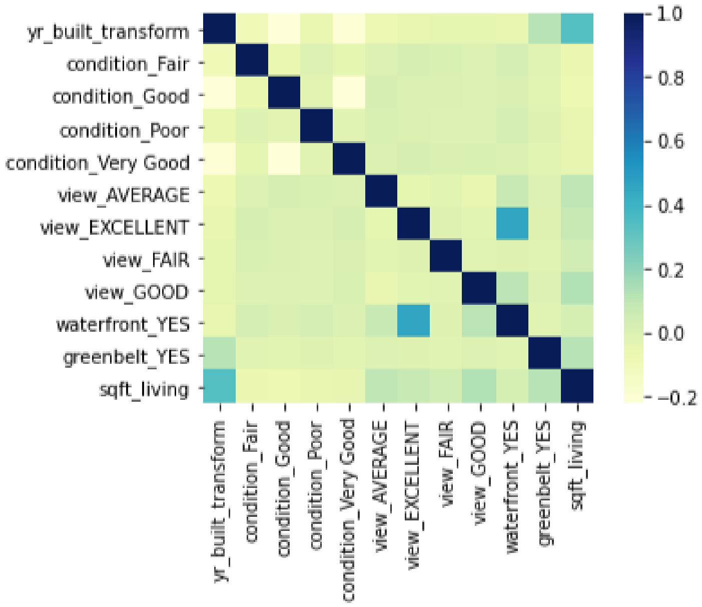
Now that we have this model, let's double check for errors.

5. Error Testing

Now that we have this amazing model, we're going to perform some error testing to see if the results we get are accurate and also, well, still linear. So, we're going to examine collinearity, Mean Absolute Error (MAE) Testing vs Root Mean Squared Error (RMSE), homoscedasticity, and Goldfeld Quandt testing.

Collinearity

To examine collinearity, we're going to look at a seagram heatmap of a correlation matrix between a number of the additional features we looked at.



From here we can see the majority of our features show very little correlation. The two areas that do show it are `view_EXCELLENT` correlates slightly with `waterfront` . This makes sense because it's intuitive that if you live on the water, chances are you have an excellent view. So, based on that logic, I will remove the `waterfront` variable to avoid this.

The other correlated variables are `yr_built_transform` with `sqft_living` . It seems as though houses that were built more recently are larger than older houses. This makes sense, and we will remove `yr_built_transform` because, well, we can't really remove `sqft_living` .

Mean Absolute Error (MAE) Testing vs Root Mean Squared Error (RMSE)

The next item to examine is MAE vs RMSE. This analysis will tell us the error in our model.

Mean Absolute Error (MAE)

Mean Absolute Error is a means to evaluate how "dar off" or inaccurate a model may be. It sums the absolute value of all of the errors and takes the mean.

When we calculate the MAE for our model, we get $212,000. In this case we say the our model has an error of $212,000.

Mean Absolute Error (RMSE)

The other method of determining error is to take the square root of all of "the average squared difference between the estimated values and the actual values" (according to Wikipedia). This is another method to analyze residuals of a model and determine how much error exists.

In our model, the RMSE is 323,000, which is larger than the MAE.

Discussion

Because the RMSE > MAE, there may be more outliers in our data even though we eliminated the top 1% for square footage and price.

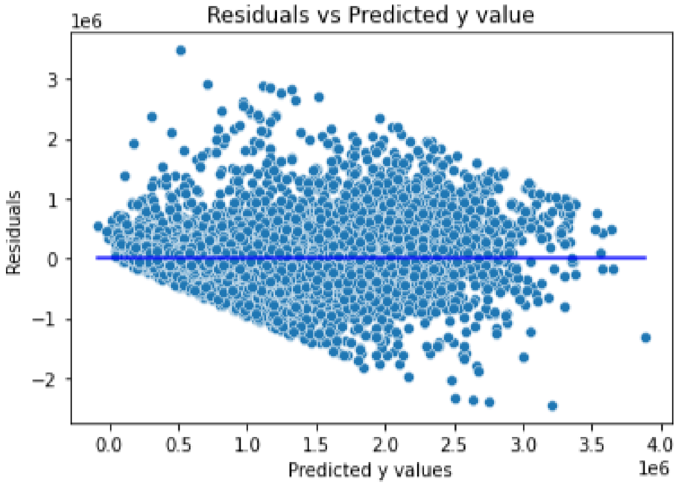
Linearity

This was analyzed previously. We believe that only those variable with a linear relationship have been included.

Homoscedasticity

Homoscedasticity is the observation that the magnitude of the errors (or residuals) is the same no matter what the input, or independent variable is. The most effective way to observe this is by plotting the residuals against the predicted values. Homoscedasticity will result in a straight line around the max residuals. Heteroscedasticity will result in a curved line around the max residuals.

The plot below shows our residuals vs the predicted.



We do not see great homoscedasticity. This would improve possibly if we removed some outliers. Let's do an additional test to

Results Discussion

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Languages

● **Jupyter Notebook** 100.0%