main | Go to file | Add file | Code

This branch is 11 commits ahead of learn-co-curriculum:main. Contribute | Fetch upstream

sealaurel final updates ... 41 seconds ago 20

data	final updates	1 hour ago
images	final updates	1 hour ago
.canvas	Create .canvas	7 months ago
.gitignore	pushing it upstream	17 days ago
LICENSE.md	Initial commit	8 months ago
Phase2_project_presentation....	final updates	41 seconds ago
README.md	pushing it upstream	17 days ago
README.pdf	pushing it upstream	17 days ago
project2_OSEMN_plus_mvp_0...	final updates	1 hour ago
project2_OSEMN_plus_mvp_0...	final updates	1 hour ago
visualization_appendix - Jupyt...	final updates	1 hour ago
visualization_appendix.ipynb	final updates	41 seconds ago

README.md

About

No description, website, or topics provided.

Readme

View license

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 100.0%

Increase Your Property Value and Stretch Your House Buying Budget Further

Phase 2 Final Project

- Student name: **Elena Kazakova**
- Student pace: **full time**
- Cohort: **DS02222021**
- Scheduled project review date/time: **TBD**
- Instructor name: **James Irving**
- Blog post URL: **TBD**

Table of Contents

Click to jump to matching Markdown Header.

Introduction

This Project analyses the data on the properties sold in King County, WA (not including the Seattle inner city areas) over the period of one year, from May 2014 till MAy 2015. The resulting model provides insight what features of a property increases its' value and what variables outside of property owners control affect the price of a house.

Business Problem

This project is the Inference Analysis project of King County, WA house prices, and various factors that might affect the sales price.
This study aims to build a model(s) of house sale prices depending on the features of the property in the dataset provided. This information can be helpful for house owners, house buyers, and real estate agents in the county.

Obtain

Data Understanding

The dataset used in this project has been downloaded from KAGGLE site. The dataset includes the information about properties sold in King County of Washington State between May 2014 and May 2015. The area consists of Seattle city area but does not include the inner city. The dataset consists of 21 dependent and independent variables and 21597 records.

Description of the fields

The file has 21597 records with 21 columns, out of which 11 columns have integer values, 8 are real numbers, and 2 are strings.
The annotation to the fields and associated data
(link to the definitions here)

- **id:** Unique ID for each home sold

    i. no NULL values
    2.integer numbers
    3.176 duplicate records

> **353 rows to be dropped**

- **date:** Date of the sale

> no NULL values
> string

> **Convert to DateTime type**

- **price** - Price of each home sold

> no NULL values
> Real numbers
> Minimum price: 78000
> Maximum price: 7700000

- **bedrooms** - Number of bedrooms

> no NULL values
> Integer numbers, between 1 and 33

- **bathrooms** - Number of bathrooms, where .5 accounts for a room with a toilet but no shower

> no NULL values
> Real numbers, between 0.5 and 8.0

- **sqft_living** - Square footage of the house interior living space

> No NULL values
> Integer numbers
> Minimum value: 370
> Maximum value: 13540

- **sqft_lot** - Square footage of the land lot

> No NULL values
> Integer numbers
> Minimum value: 520
> Maximum value: 1651359

- **floors** - Number of floors

> no NULL values
> Real numbers, between 1.0 and 3.5

- **waterfront** - A categorical variable for whether the house was overlooking the waterfront or not

> **2376** NULL values
> Real numbers, only two values 1.0 and 0.0

> **Convert to a categorical variable**
> **Waterfront, not Waterfront**
>
> **Replace NULL values with "Missing" category**

- **view** - A categorical variable describing how good the view of the property was

> **63** NULL values
> Real numbers: 1.0, 2.0, 3.0, 4.0

> **Convert to a categorical variable**
> **Poor, Fair, Good Excellent**

> **Replace NULL values with "Missing" category**

- **condition** - A categorical variable describing the condition of the house

> no NULL values
> Integer numbers, between 1 and 5

> **Convert to a categorical variable**
> **Poor, Fair, Good, Very Good, Excellent**

- **grade** - A categorical variable describing the quality of construction, from 1 to 13; 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

> no NULL values
> Integer numbers, between 3 and 13

- **sqft_above** - The square footage of the interior housing space that is above ground level

> No NULL values
> Integer numbers
> Minimum value: 370
> Maximum value: 9410

- **sqft_basement** - The square footage of the interior housing space that is below ground level

> No NULL values
> String

> **Convert to integer**

- **yr_built** - The year the house was initially built

> No NULL values
> Integer numbers
> Minimum value: 1900
> Maximum value: 2015

- **yr_renovated** - The year of the last house renovation

> **3842** NULL values
> Real numbers, between 0.0 and 2015.0

> **Convert to integer**

- **zipcode** - What zipcode area the house is in

> no NULL values
> Integer numbers, 70 unique values

> **Convert to categorical variable or drop**

- **lat** - Lattitude

> no NULL values
> Real numbers

- **long** - Longitude

> no NULL values
> Real numbers

- **sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors

> No NULL values
> Integer numbers
> Minimum value: 399
> Maximum value: 6210

- **sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors

> No NULL values
> Integer numbers
> Minimum value: 651
> Maximum value: 871200

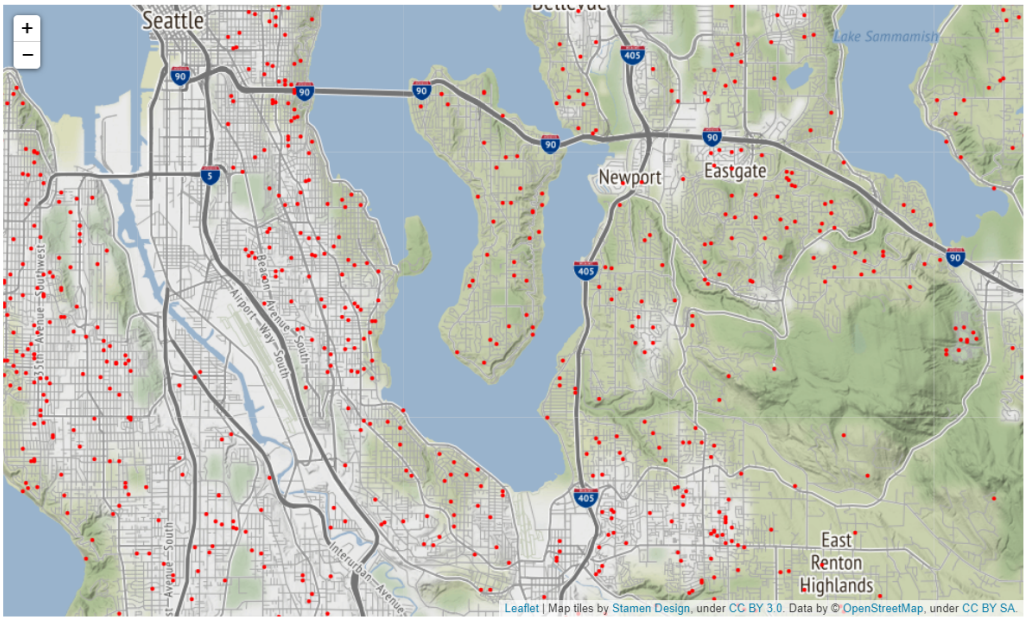
Initial cleaning of the data

- i. Dropping duplicate rows for houses sold twice in the timeframe of the dataset
- ii. Converting the 'date' field to DateTime formate and making sure it worked
- iii. Locations of the houses with missing 'waterfront' values

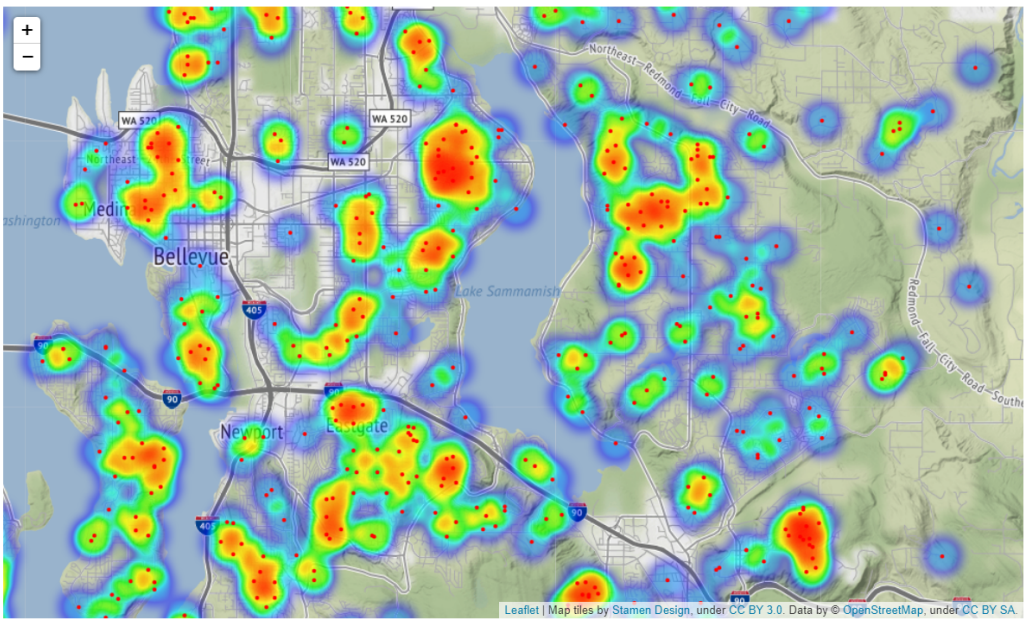
> Possible strategies:

- Check if there are waterfront properties among neighbors within a certain distance range
- Make a map and place properties with missing values on it visually
- What is the longitude of the bay shore? Any house with a missing value too far away from it should have their waterfront value set to 0. Hopefully, it will eliminate most of the missing values in this field

The choice is option 2 as the simplest of the three: ##### 3.1 Visual assesment of the houses with Null value in the 'waterfront column



]:



It is self-evident from the visuals above that the vast majority of the houses are located inland. Simple zooming in the maps allows a rough counting of alleged waterfront properties. The estimate is approximately 20 waterfront houses. It is 0.85% of all properties with no value in 'waterfront column (2353). In the primary dataset, the percentage of waterfront properties out of the total number of properties is 0.68%. The numbers above indicate that replacing the NaN values with 0 would introduce a systemic error of 0.01% to the whole system.

Conclusion: The NULL values in the 'waterfront' column will be replaced with 0.

3.2 NULL values in 'waterfront' column replaced with 0 and the column converted to the integer datatype to make it categorical

3.3 Replacing NULL values in waterfront and view field using IterativeImputer (testing the approach)

Conclusion: Based on the results of the IterativeImputer application to the data the original approach of replacing missing values in 'waterfront', 'yr_renovated', and 'view' with 0 will be taken

4. NULL values in 'yr_renovated column' replaced with 0 and the type changed to integer

5. '?' values in 'sqft_basement' column replaced with 0 and the type changed to integer

6. NULL values in 'view' column replaced with 0 and the type changed to integer

Scrub and Explore

Additional data cleaning

Dropping non-needed fields

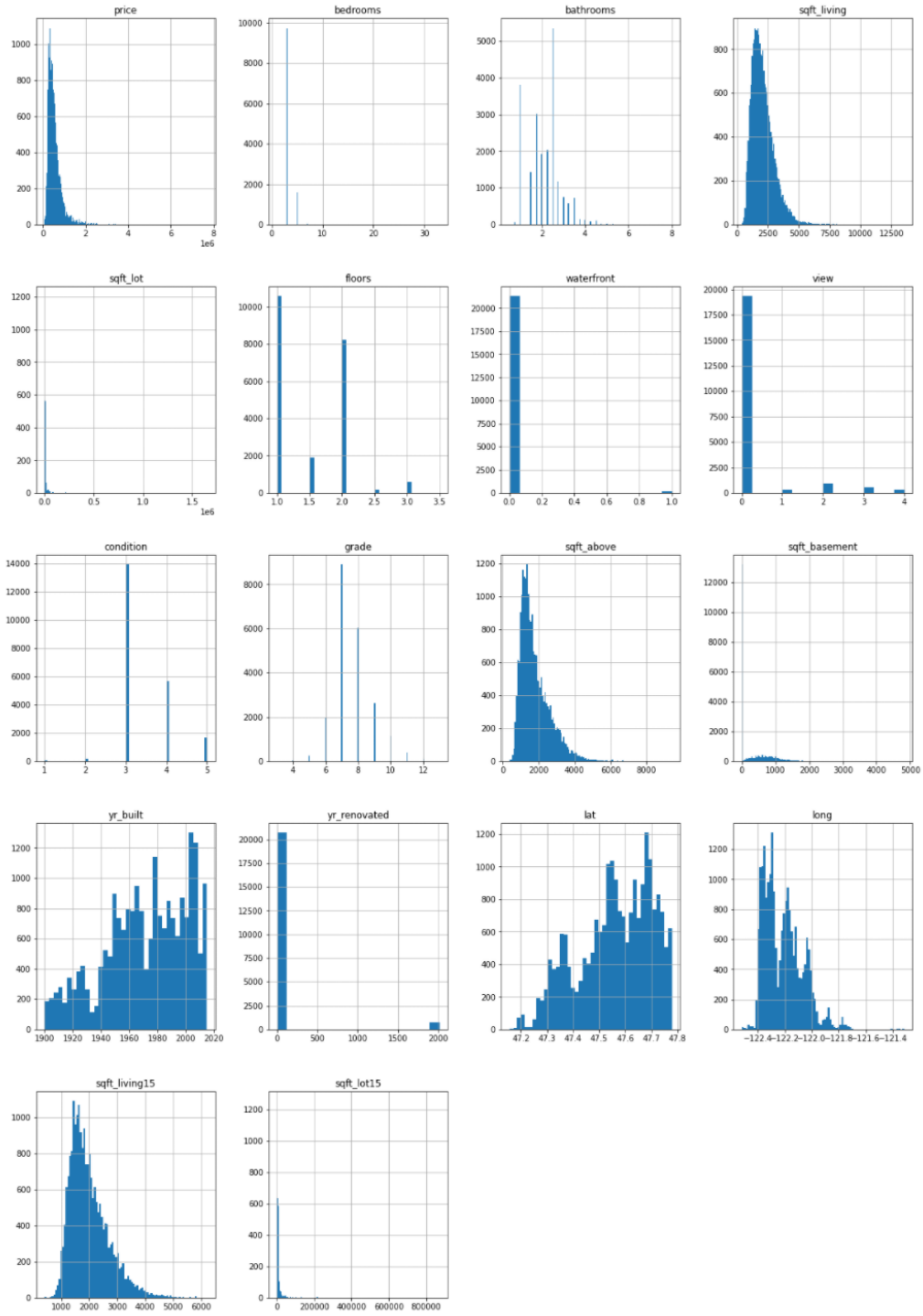
> Dropping 'zipcode' variable because there are better indicators of a location. Zipcodes boundaries are usually drawn out of convenience for postal services or other more formal reasons than geographic location

> Dropping 'id' field

Exploring distributions and correlation of original variables

Numerical variables: Investigating distributions and correlations between the original, minimally processed predictors and the target (price)

Histograms of the original predictor variables



Based on the histograms above

1. The following variables should be considered categorical:

> Waterfront
> Condition
> View

2. sqft_basement, sqft_lot, sqft_lot15, and yr_renovated have a large number of zeros and are strong candidates for removal of outliers and/or engineered variables
3. Latitudes and Longitudes can be used as descriptors of a geographic location of a property. However, there is a better variable to describe the location of a property, a distance from the center of the city, which can be calculated from geocoordinates.
   >4. The target variable, the price of the property, has a strong positive skew attributed to

> outliers in the higher price bracket. The strategy is to **remove the outliers and to transform the variable** to make it more normally distributed

A new categorical variable (integer datatype) renovation_done with values [0,1,2,3,4] is created

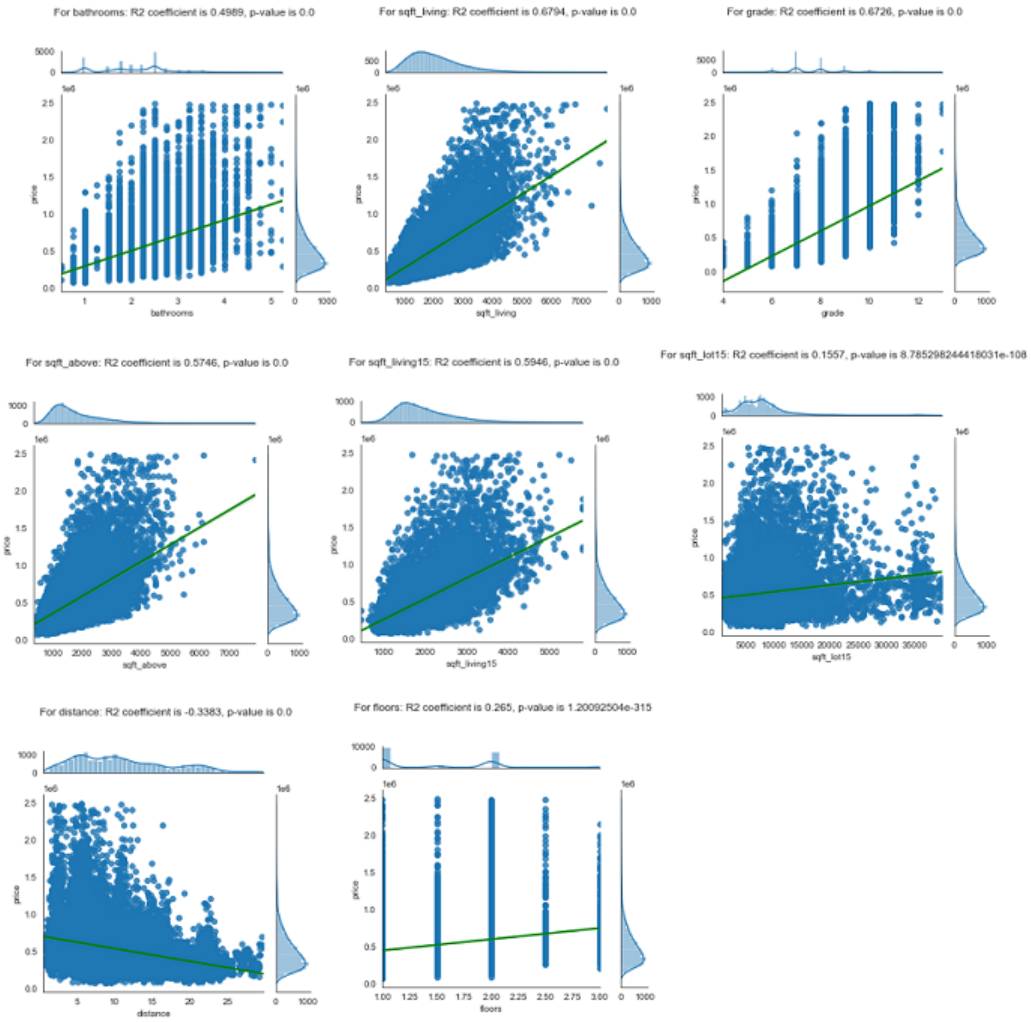
- 0 representing renovation never done on houses more than 9 years old (yr_built between 2015 and 2006)
- 1 representing renovation done more than or equal 50 years ago
- 2 representing renovation done between 30 and 49 years ago
- 3 representing renovation done between 29 and 10 years ago
- 4 representing renovation done between 9 and 1 year ago OR houses built less or equal 9 years ago (yr_built between 2015 and 2006)

sqft_basement, date, latitude & longtitude variables are dropped

Coefficients of Determination and p-values for the remaining variables

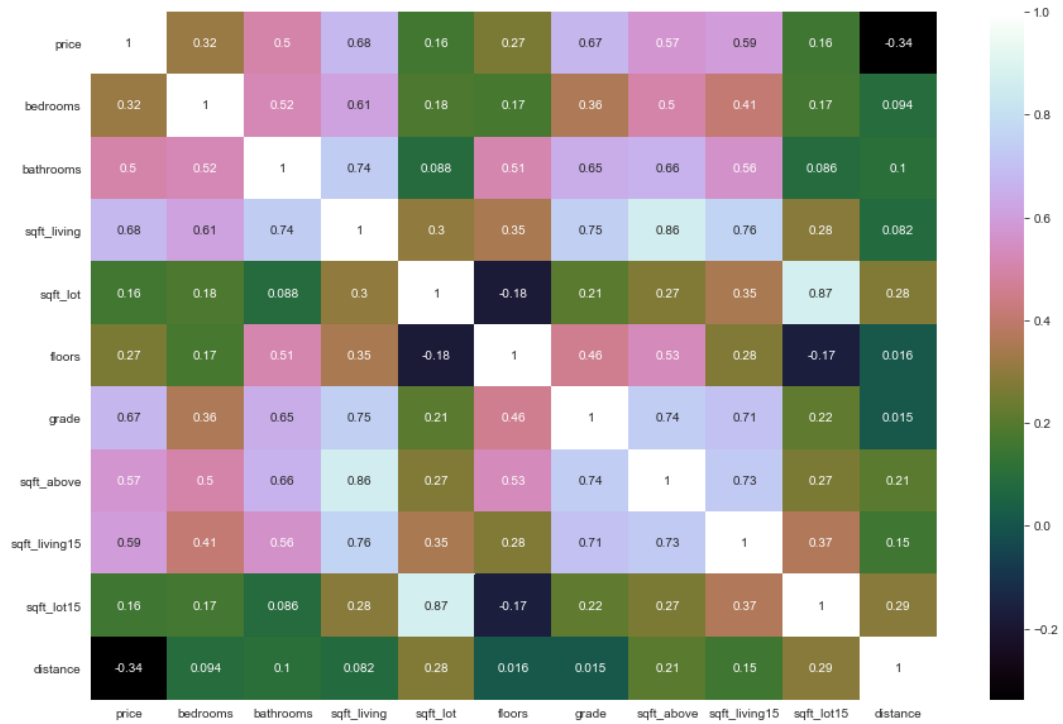
- For price: R2 coefficient is 1.0, p-value is 0.0
- For bedrooms: R2 coefficient is 0.3195, p-value is 0.0
- For bathrooms: R2 coefficient is 0.4999, p-value is 0.0
- For sqft_living: R2 coefficient is 0.6795, p-value is 0.0
- For sqft_lot: R2 coefficient is 0.1565, p-value is 7.279024950162686e-110
- For floors: R2 coefficient is 0.2657, p-value is 1.2233e-320
- For grade: R2 coefficient is 0.6726, p-value is 0.0
- For sqft_above: R2 coefficient is 0.5746, p-value is 0.0
- For yr_built: R2 coefficient is 0.0348, p-value is 8.471205832359682e-07
- For sqft_living15: R2 coefficient is 0.5951, p-value is 0.0
- For sqft_lot15: R2 coefficient is 0.1512, p-value is 1.1950868785722046e-102
- For distance: R2 coefficient is -0.3432, p-value is 0.0

Plotting numerical variables against the target variable



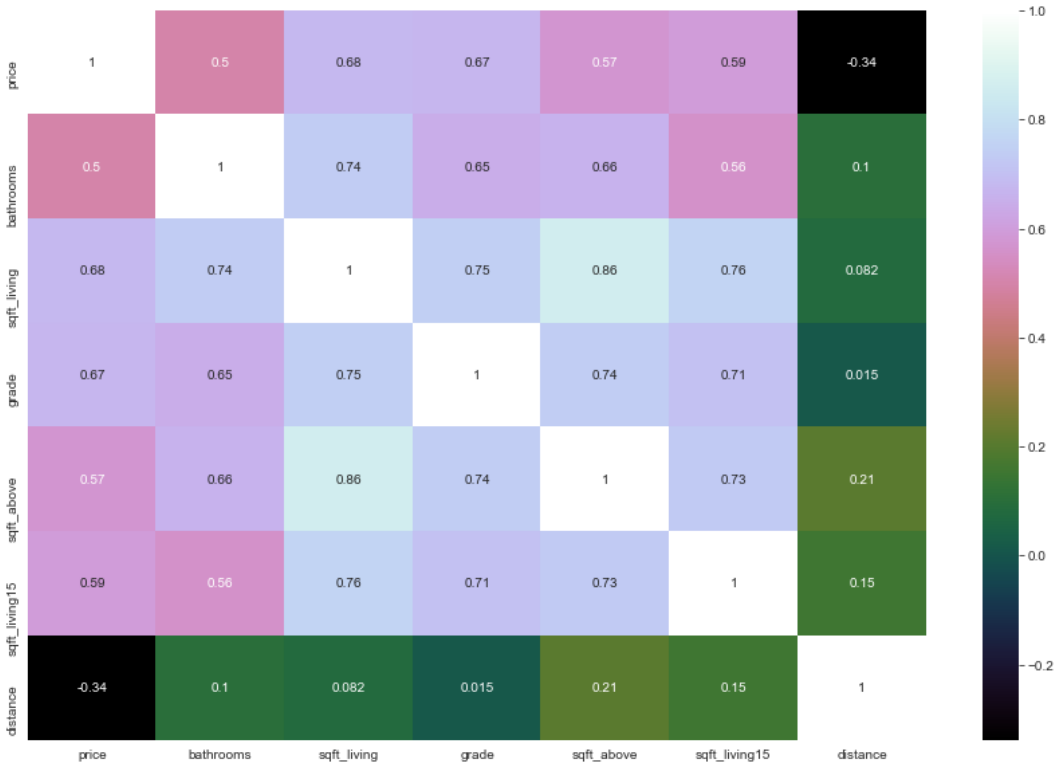
It is evident from the visualization that 'floors' and 'sqft_lot' do not display a strong correlation with the target

Numerical predictors correlation with the price and with each other presentes as a heatmap

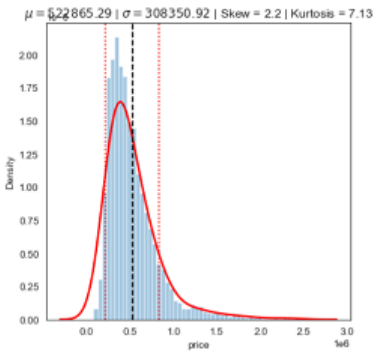


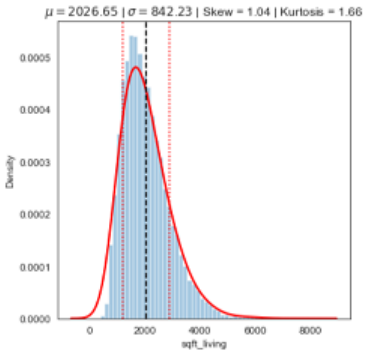
Bedrooms is relatively highly correlated with bathrooms and sqft_living; correlations of floors, sqft_lot and sqft_lot15 with prices are low. Dropping these variables

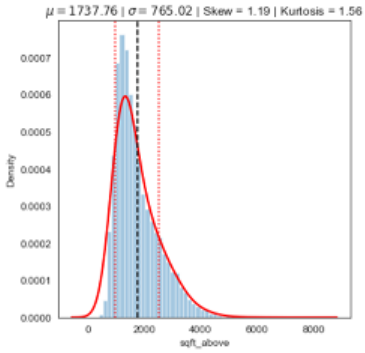
Heatmap of the remaining variables

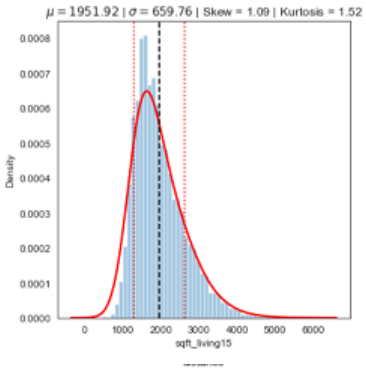


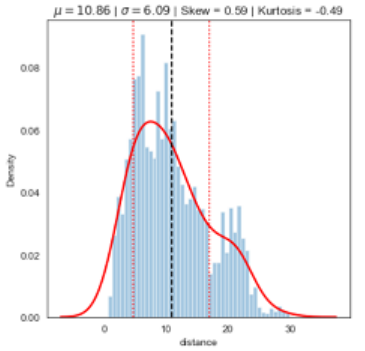
KDE and distribution plots of the remaining variables

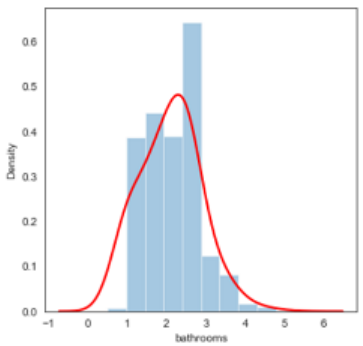


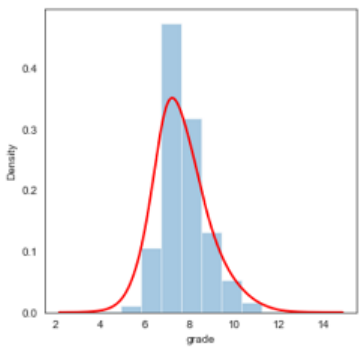








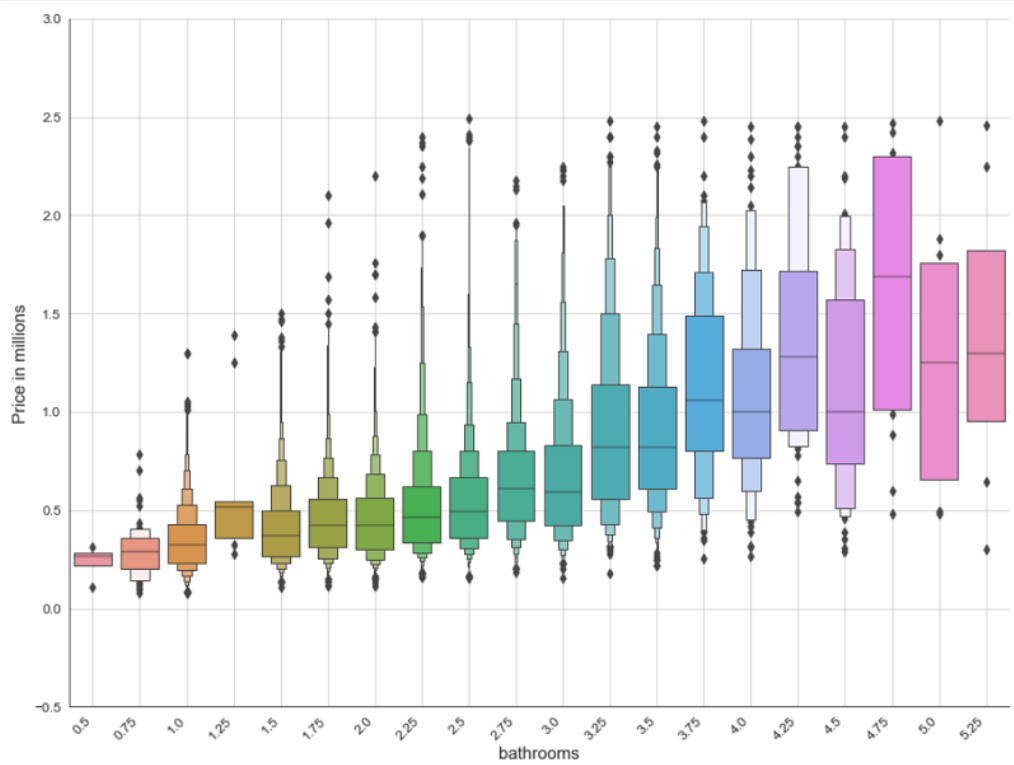


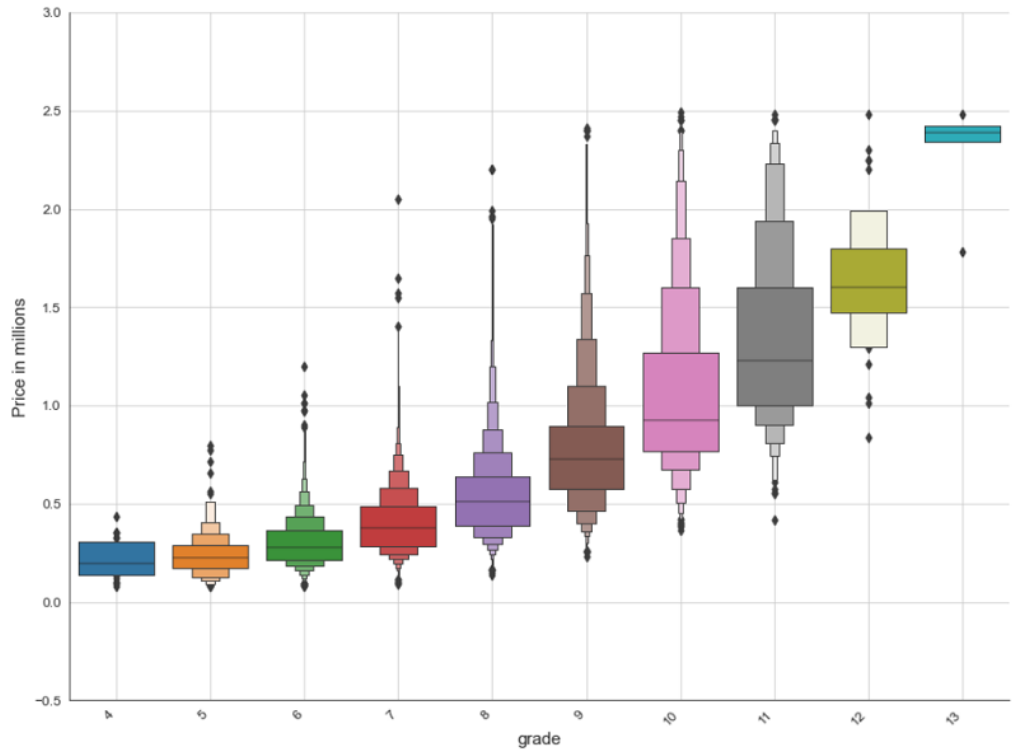


All but one variable (distance) are left shifted, which is indicated by the skewness values >1, with price has the most skewed distribution. Because the skewness it to the left (positive values), the log transformation might be needed to normalize the variables. Kurtosis values for all variables are different from 0 (Pearson's definition of kurtosis of a normal distribution).

Price distribution is highly **Leptokurtic**, other variables are slightly **Leptokurtic** (sqft_living, sqft_above, sqft_living15, grade), slightly **Platykurtic** (distance) or almost **Mesokurtic** (bathrooms)

Diamond box plots for descrete numerical variables





Based on the distribution plots, variable 'bathroom' has a symmetrical distribution (Skewness is 0.28), with a very low kurtosis (0.09) indicative of a **Mesokurtic** curve (Gaussian distribution has a kurtosis of 0 by Pearson's definition used by scipy.stats.kurtosis method)
Variable 'grade' is slightly skewed to the right (0.73) and relatively low kurtosis, slightly above 1.
The pronounced correlation of these variables with the price is identifiable in the box plots above.
The plots show that the numbers of outliers in the distribution of the variables are reasonable. Both variables have a wider range of values in the higher price brackets.

Numerical variables: Exploring Mutual Correlation Coefficients and Variance Inflation Factor

Using VIF as an indicator of collinearity between independent variables

	variables	VIF
0	price	11.713924
1	bathrooms	21.024526
2	sqft_living	42.036909
3	grade	27.796857
4	sqft_above	26.850680
5	sqft_living15	28.037076
6	distance	6.670397

- Dropping the columns that have high VIF. However, it is more logical to leave sqft_living versus sqft_living15 in because it is easier to interpret in a model and it is a feature under control of a property owner (versus sqft_living15 which is a charecteristic of a neighborhood)

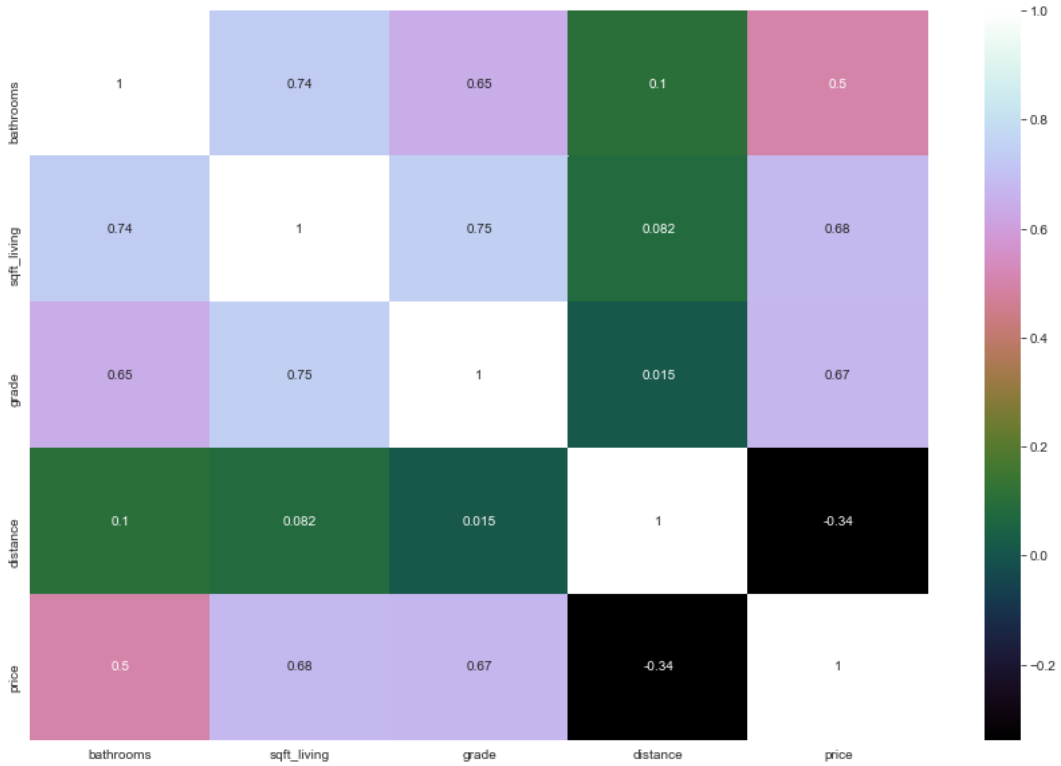
After dropping the fields

	variables	VIF
0	bathrooms	20.829518
1	sqft_living	16.749281
2	grade	18.971120
3	distance	3.971751

Pearson coefficients analysis of the remaining independent variables

- For bathrooms: R2 coefficient is 0.4989, p-value is 0.0
- For sqft_living: R2 coefficient is 0.6794, p-value is 0.0
- For grade: R2 coefficient is 0.6726, p-value is 0.0
- For distance: R2 coefficient is -0.3383, p-value is 0.0
- For price: R2 coefficient is 1.0, p-value is 0.0

Heatmap of the remaining predictor and the target



Mutual correlation coefficients between the remaining independent variables are slightly higher or below 0.7. I am leaving sqft_living, grade, and bathroom variables in because of their logical connection with a property price despite their multicollinearity (0.74 & 0.73 are above the 0.7 threshold).

Therefore the remaining numerical variables for modeling are

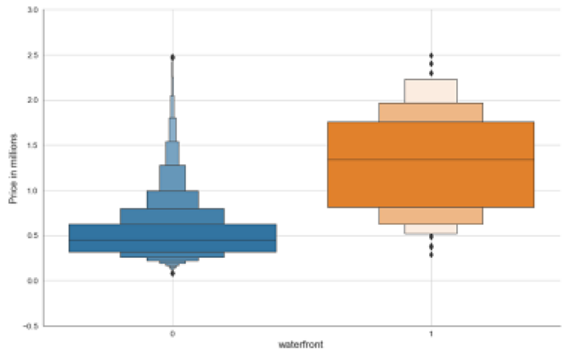
1. grade
2. bathrooms
3. sqft_living
4. distance

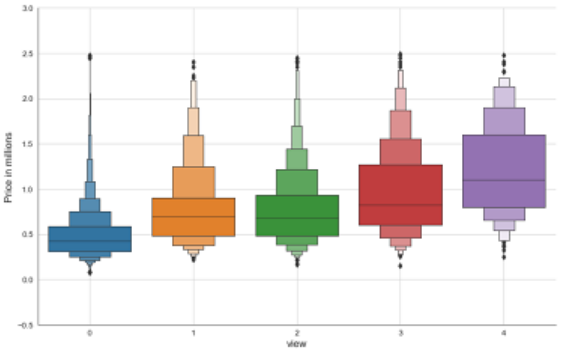
Categorical variables: Investigating distributions and the raw correlations between the original, minimally processed predictor and the target (price)

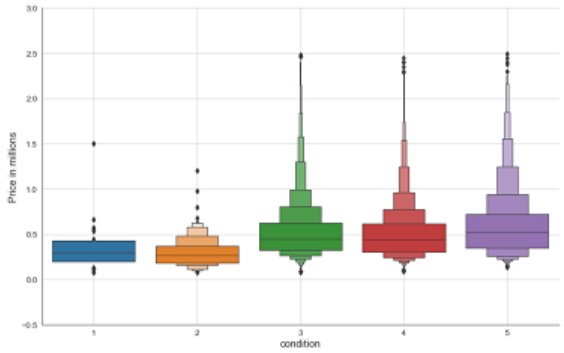
Original categorical variables:

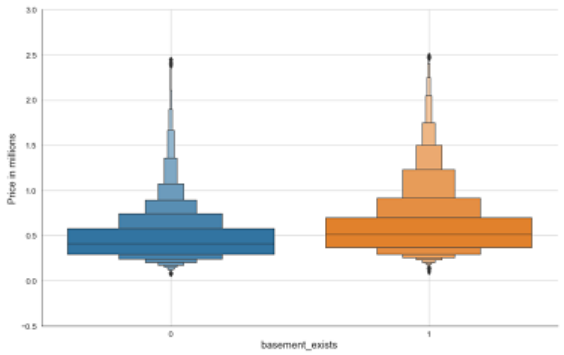
- waterfront
- view
- condition
- basement_exists
- renovation_done
- month

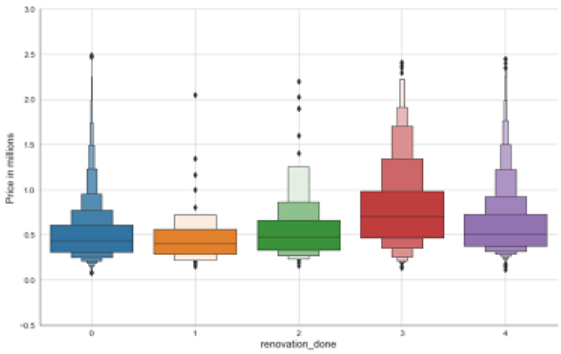
Visual investigation of the box plots

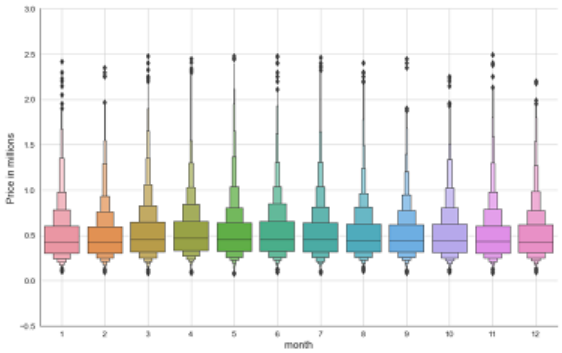

Price in millions
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0
1
waterfront


Price in millions
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0
1
2
3
4
view


Price in millions
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
1
2
3
4
5
condition


Price in millions
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
0
1
basement_exists





Based on the plots, it is self-evident that 'month', 'condition' and 'basement_exists' variables do not significantly affect the price of the properties and can be dropped from the categorical variables. Variable 'waterfront' is also dropped becasue waterfront properties represent a tiny portion of all properties in the dataset

The remaining categorical variables, grade, view, and renovation_done are dummed out with OheHotEncoder

Model

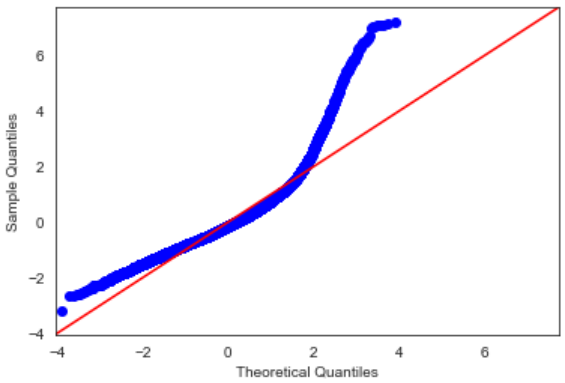
Data Modeling

Baseline model

> ***"Everything should be made as simple as possible, but no simpler."***
>
> Albert Einstein

The chosen baseline model is a model with only one numerical variable, grade.

Creating a model



The baseline model has a coefficient of determination of 0.452, indicating that roughly 45% of the observations fit the model. F-statistics is very high that indicates that the baseline model is a significant improvement of the "intercept only model"
The Skewness and the Kurtosis values indicate non-normal distribution of the target variable
QQ plot is also indicative of the abnormal distribution of the residuals, especially in the upper Quantile

Model 1 (all numerical variables considered significant, see Explore section)

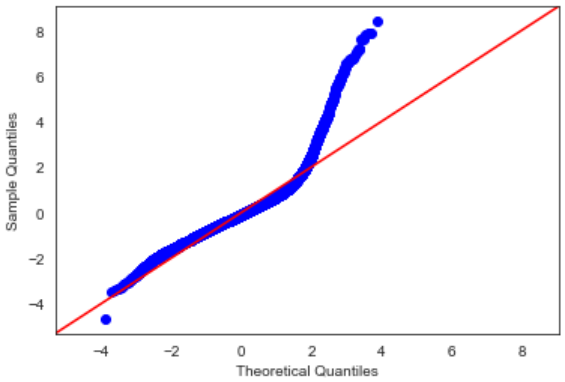
Model summary

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.668
Model:	OLS	**Adj. R-squared:**	0.668
Method:	Least Squares	**F-statistic:**	9947.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:31	**Log-Likelihood:**	-2.6759e+05
No. Observations:	19811	**AIC:**	5.352e+05
Df Residuals:	19806	**BIC:**	5.352e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.142e+05	1.04e+04	-30.071	0.000	-3.35e+05	-2.94e+05
bathrooms	-1.725e+04	2599.797	-6.634	0.000	-2.23e+04	-1.22e+04
sqft_living	177.5465	2.610	68.031	0.000	172.431	182.662
grade	9.475e+04	1748.262	54.195	0.000	9.13e+04	9.82e+04
distance	-1.917e+04	209.218	-91.613	0.000	-1.96e+04	-1.88e+04

Omnibus:	8098.307	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	65266.147
Skew:	1.762	**Prob(JB):**	0.00
Kurtosis:	11.163	**Cond. No.**	1.84e+04

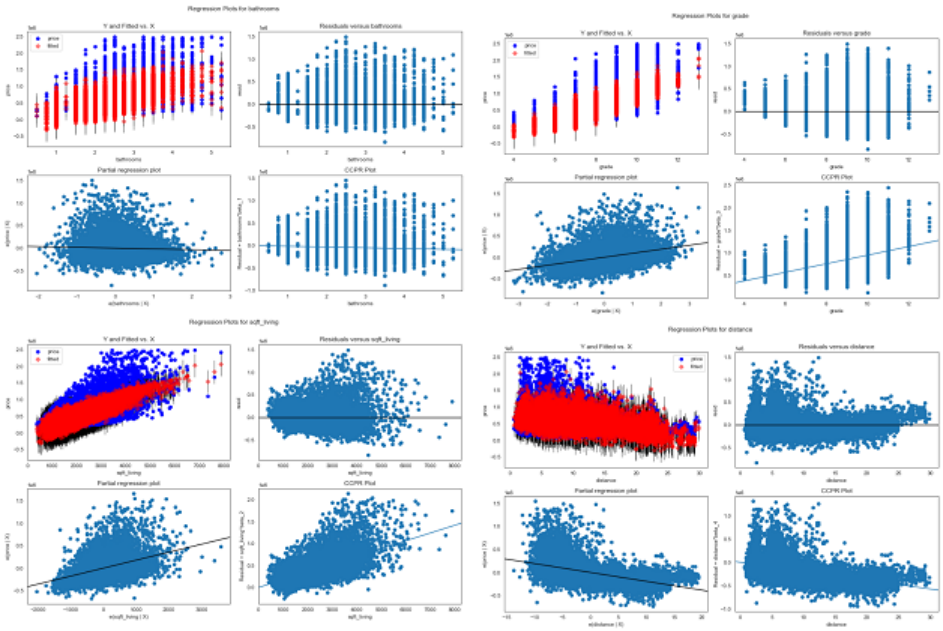


The summary of the model above indicates that
1. All the independent variables coefficients and the intercept value are significant (p-values< 0.05)
2. The coefficient of determination (R2) is not very high, but it is significantly higher than R2 of the baseline model. It indicates that about 66.8 percent of the observations fall within the regression line
3. The skew and the Kurtosis values indicate the highly non-normal distribution of the target variable
4. The high value of JB coefficient also indicates that the data is highly non-normal

From the model's QQ plot, it is also quite obvious that the 'price' target variable is not normally distributed. A steep swing up indicates that the higher-priced houses are less likely to fit the baseline model and are more spread out. One possible reason might be **an unusually large number of outliers in the dataset**

There are two potential approaches that can be taken
1. Normalization by either log or square root transformation
2. Removal of outliers

Accessing linearity of model predictors' relationship with the target and their homoscedasticity



The results indicate that all of the predictors display linear relationship with the target. The distance, sqft_living and bathrooms variables display less heteroscedasticity than the grade variable does.

There might be several appropriate ways to address this issue
1. Log transformation of the target and/or independent variables

2. Using either Generalized Least Squares or Weighted Least squares
3. Bootstrapping

Model 2 (adding categorical variables)

Model Summary

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.669
Model:	OLS	**Adj. R-squared:**	0.668
Method:	Least Squares	**F-statistic:**	3328.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:39	**Log-Likelihood:**	-2.6756e+05
No. Observations:	19811	**AIC:**	5.352e+05
Df Residuals:	19798	**BIC:**	5.353e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.189e+05	1.05e+04	-30.473	0.000	-3.39e+05	-2.98e+05
bathrooms	-1.502e+04	2616.457	-5.740	0.000	-2.01e+04	-9888.816
sqft_living	176.1368	2.615	67.355	0.000	171.011	181.263
grade	9.553e+04	1750.159	54.586	0.000	9.21e+04	9.9e+04
distance	-1.917e+04	208.998	-91.705	0.000	-1.96e+04	-1.88e+04
view_1	5556.1173	1.02e+04	0.547	0.584	-1.43e+04	2.55e+04
view_2	-1185.4496	6242.542	-0.190	0.849	-1.34e+04	1.11e+04
view_3	4061.7300	8636.371	0.470	0.638	-1.29e+04	2.1e+04
view_4	1.425e+04	1.14e+04	1.249	0.212	-8114.035	3.66e+04
renovation_done_1	1.333e+04	2.89e+04	0.462	0.644	-4.32e+04	6.99e+04
renovation_done_2	1.11e+04	1.72e+04	0.644	0.520	-2.27e+04	4.49e+04
renovation_done_3	3143.7672	1.03e+04	0.305	0.761	-1.71e+04	2.34e+04
renovation_done_4	-2.634e+04	3771.177	-6.985	0.000	-3.37e+04	-1.9e+04

Omnibus:	8082.328	**Durbin-Watson:**	1.989
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	64840.177
Skew:	1.760	**Prob(JB):**	0.00
Kurtosis:	11.134	**Cond. No.**	5.02e+04

The summary indicates a very slight improvement over the previous model, 66.9% versus 66.8% of all of the observations fall within the results of the line formed by the regression equation.

It is also evident that p-values of most of the categorical values are very high, indicating their insignificance in the model. However, because they describe the same feature, I am leaving them in for now

The residuals normality did not improve

Model 3 (preprocessing and removal of outliers)

Outlier removal by IQR method

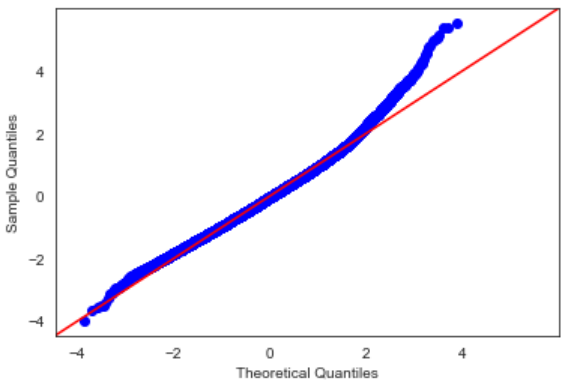
Model Summary

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.633
Model:	OLS	**Adj. R-squared:**	0.633
Method:	Least Squares	**F-statistic:**	2672.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:39	**Log-Likelihood:**	-2.4463e+05
No. Observations:	18619	**AIC:**	4.893e+05
Df Residuals:	18606	**BIC:**	4.894e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.397e+05	7752.045	-18.019	0.000	-1.55e+05	-1.24e+05
sqft_living	129.1653	2.028	63.697	0.000	125.191	133.140
distance	-1.527e+04	152.634	-100.023	0.000	-1.56e+04	-1.5e+04
grade	7.293e+04	1291.290	56.478	0.000	7.04e+04	7.55e+04
bathrooms	-7626.9730	1901.298	-4.011	0.000	-1.14e+04	-3900.255
view_1	1295.3996	7317.902	0.177	0.859	-1.3e+04	1.56e+04
view_2	-1419.9738	4457.993	-0.319	0.750	-1.02e+04	7318.099
view_3	454.9216	6141.592	0.074	0.941	-1.16e+04	1.25e+04
view_4	1.163e+04	8208.818	1.417	0.156	-4457.709	2.77e+04
renovation_done_1	3.485e+04	2.08e+04	1.673	0.094	-5972.705	7.57e+04
renovation_done_2	1.589e+04	1.22e+04	1.299	0.194	-8087.988	3.99e+04
renovation_done_3	-907.0686	7331.656	-0.124	0.902	-1.53e+04	1.35e+04
renovation_done_4	-1.562e+04	2714.396	-5.753	0.000	-2.09e+04	-1.03e+04

Omnibus:	849.136	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1199.719
Skew:	0.438	**Prob(JB):**	3.05e-261
Kurtosis:	3.882	**Cond. No.**	4.74e+04



While the IQR removal of outliers decreased the R squared of the model, it made the distribution more normal (Skew and Kurtosis values are almost within the normality ranges). This fact is also reflected by the QQ plot of the model residuals. Unfortunately, the Coefficient of Determination dropped.

Outlier removal by Z-scores method

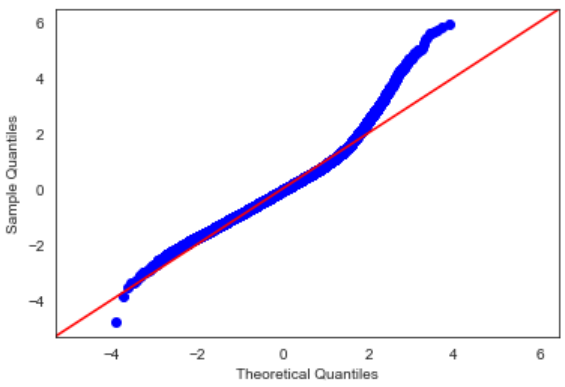
Model Summary

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.654
Model:	OLS	**Adj. R-squared:**	0.653
Method:	Least Squares	**F-statistic:**	3027.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:39	**Log-Likelihood:**	-2.5551e+05
No. Observations:	19261	**AIC:**	5.110e+05
Df Residuals:	19248	**BIC:**	5.111e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.074e+05	8498.928	-24.399	0.000	-2.24e+05	-1.91e+05
sqft_living	141.6977	2.203	64.311	0.000	137.379	146.016
distance	-1.666e+04	167.636	-99.371	0.000	-1.7e+04	-1.63e+04
grade	8.286e+04	1416.839	58.485	0.000	8.01e+04	8.56e+04
bathrooms	-1.001e+04	2109.297	-4.746	0.000	-1.41e+04	-5875.419
view_1	1.247e+04	8077.275	1.544	0.123	-3357.516	2.83e+04
view_2	-1948.1291	4983.950	-0.391	0.696	-1.17e+04	7820.847
view_3	-3331.9562	6896.935	-0.483	0.629	-1.69e+04	1.02e+04
view_4	1.641e+04	9104.994	1.802	0.072	-1441.323	3.43e+04
renovation_done_1	3.577e+04	2.27e+04	1.576	0.115	-8719.452	8.03e+04
renovation_done_2	1.284e+04	1.37e+04	0.938	0.348	-1.4e+04	3.97e+04
renovation_done_3	-4898.3754	8252.665	-0.594	0.553	-2.11e+04	1.13e+04
renovation_done_4	-2.12e+04	3022.668	-7.014	0.000	-2.71e+04	-1.53e+04

Omnibus:	2690.071	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6382.550
Skew:	0.811	**Prob(JB):**	0.00
Kurtosis:	5.307	**Cond. No.**	4.75e+04

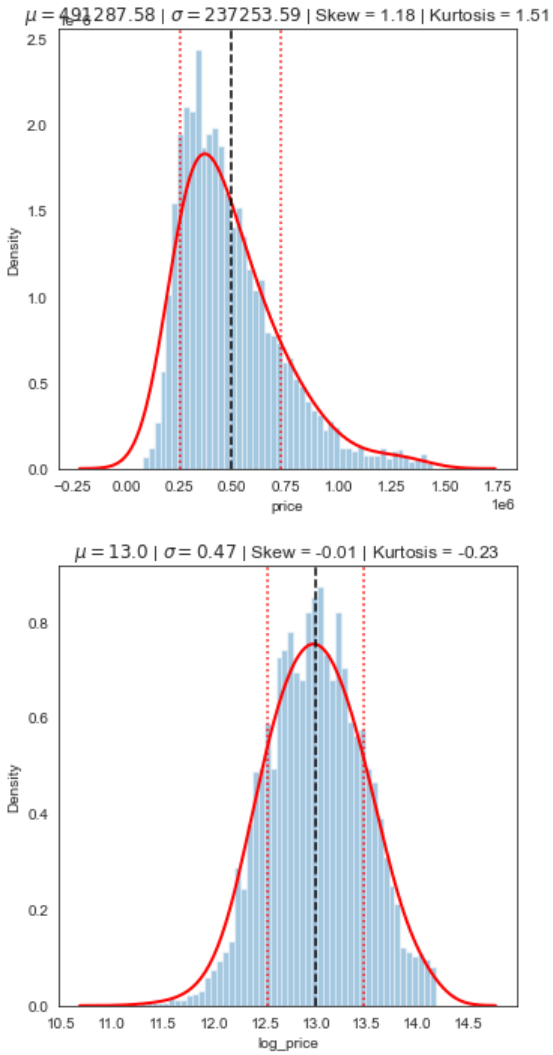


The R squared of the model is 0.654 and F-statistics is higher than for the previous model
The IQR method of outliers removal made the residual distribution more normal than Z-score method due to the former having more strict criteria. The decision is to use the dataset compiled after Z-score outlier removal.

The next step is functional transformation of the target variable

Model 4 (Using log and square root transformations on the target variable)

Log Transformation

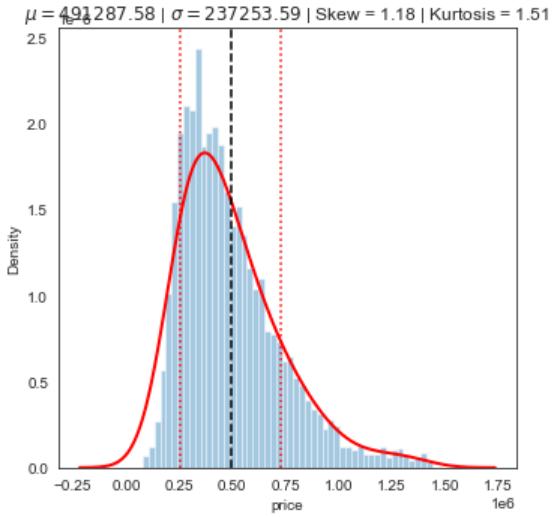


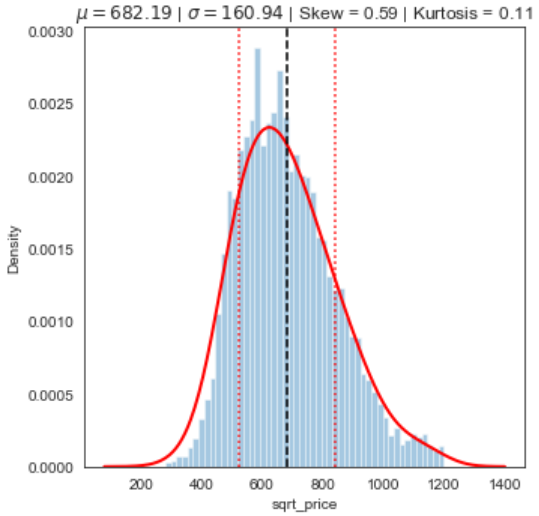
The transformation worked well, improving the normality of the 'price' variable. Log_price distribution looks more symmetrical. Skewness improved dramatically (from 1.18 to -0.01, 0 being perfectly symmetrical)

Kurtosis value decreased, making the curve more Mesokurtic (close to a Gaussian curve). It is an expected effect of log transform.

The next step is test a square root transformation.

Square Root Transformation





The square root transformation also worked well in improving the normality of the 'price' variable. sqrt_price distribution looks more symmetrical. Skewness improved dramatically (from 1.18 to -0.59, 0 being perfectly symmetrical)
However, both of the parameters are worse that the parameters of log_price distribution.

The next step is create two separate models and to see if the transformations made a difference

Model using log transformed target variable
Model Summary

OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.664
Model:	OLS	**Adj. R-squared:**	0.663
Method:	Least Squares	**F-statistic:**	3164.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:41	**Log-Likelihood:**	-2224.4
No. Observations:	19261	**AIC:**	4475.
Df Residuals:	19248	**BIC:**	4577.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.6414	0.017	704.222	0.000	11.609	11.674
sqft_living	0.0003	4.29e-06	60.161	0.000	0.000	0.000
distance	-0.0348	0.000	-106.862	0.000	-0.035	-0.034
grade	0.1599	0.003	58.035	0.000	0.155	0.165
bathrooms	0.0125	0.004	3.052	0.002	0.004	0.021
view_1	0.0120	0.016	0.764	0.445	-0.019	0.043
view_2	-0.0027	0.010	-0.281	0.779	-0.022	0.016
view_3	0.0017	0.013	0.128	0.898	-0.025	0.028
view_4	0.0253	0.018	1.427	0.154	-0.009	0.060
renovation_done_1	0.0788	0.044	1.785	0.074	-0.008	0.165
renovation_done_2	0.0449	0.027	1.688	0.091	-0.007	0.097
renovation_done_3	-0.0161	0.016	-1.002	0.316	-0.048	0.015
renovation_done_4	-0.0340	0.006	-5.782	0.000	-0.046	-0.022

Omnibus:	310.255	**Durbin-Watson:**	2.003
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	482.347
Skew:	-0.167	**Prob(JB):**	1.82e-105
Kurtosis:	3.700	**Cond. No.**	4.75e+04

Model using square root transformed target variable

Model Summary

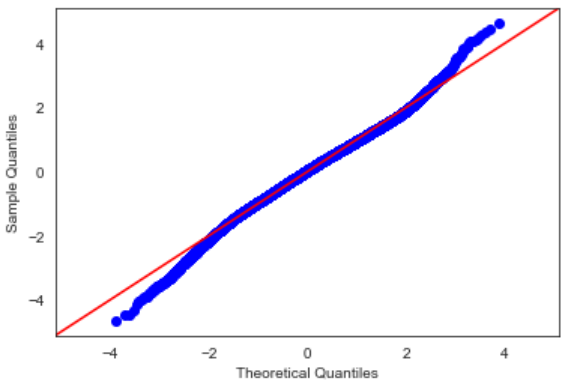
OLS Regression Results

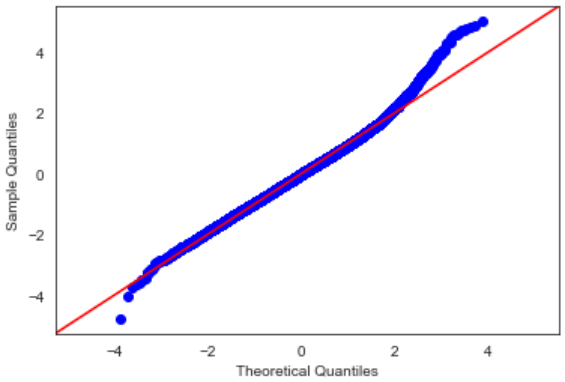
Dep. Variable:	sqrt_price	**R-squared:**	0.673
Model:	OLS	**Adj. R-squared:**	0.673
Method:	Least Squares	**F-statistic:**	3307.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:41	**Log-Likelihood:**	-1.1442e+05
No. Observations:	19261	**AIC:**	2.289e+05
Df Residuals:	19248	**BIC:**	2.290e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	207.9501	5.598	37.145	0.000	196.977	218.923
sqft_living	0.0938	0.001	64.625	0.000	0.091	0.097
distance	-11.7839	0.110	-106.715	0.000	-12.000	-11.567
grade	56.1638	0.933	60.178	0.000	54.334	57.993
bathrooms	-1.7413	1.389	-1.253	0.210	-4.465	0.982
view_1	6.2009	5.321	1.165	0.244	-4.228	16.630
view_2	-1.1299	3.283	-0.344	0.731	-7.565	5.305
view_3	-0.9735	4.543	-0.214	0.830	-9.878	7.931
view_4	9.9751	5.998	1.663	0.096	-1.781	21.731
renovation_done_1	25.8342	14.951	1.728	0.084	-3.471	55.140
renovation_done_2	12.2778	9.018	1.362	0.173	-5.397	29.953
renovation_done_3	-4.3359	5.436	-0.798	0.425	-14.991	6.319
renovation_done_4	-13.1475	1.991	-6.603	0.000	-17.050	-9.245

Omnibus:	571.397	**Durbin-Watson:**	1.997
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	863.310
Skew:	0.300	**Prob(JB):**	3.42e-188
Kurtosis:	3.846	**Cond. No.**	4.75e+04

QQ plots for the models





Both models improved the R squared and the F-statistics of the previous models. The residuals of both models display a close-to-normal distribution. Log transformation helped improve the upper part of the distribution, while square root transformation worked better in the lower part of the distribution.

R squared of the square root transformation-based model is slightly higher, while its kurtosis value is slightly worse than the kurtosis value of the log-transformed price model. The decision is to use the log-transformed target variable.

The next step is to remove unnecessary categorical variables, scale the remaining variables, and built the last model with coefficients in the regression model, which are easy to compare and interpret

Final Model

Model Summary

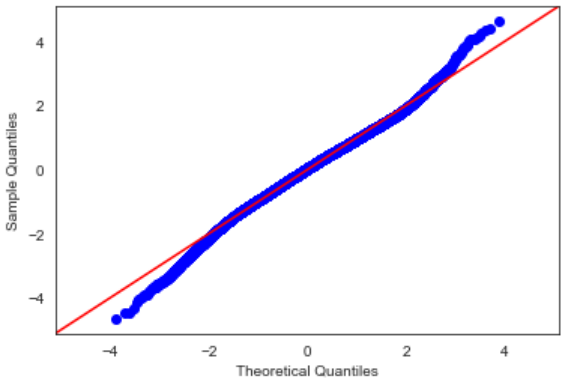
OLS Regression Results

Dep. Variable:	log_price	**R-squared:**	0.663
Model:	OLS	**Adj. R-squared:**	0.663
Method:	Least Squares	**F-statistic:**	7593.
Date:	Sat, 24 Apr 2021	**Prob (F-statistic):**	0.00
Time:	18:37:43	**Log-Likelihood:**	-2228.2
No. Observations:	19261	**AIC:**	4468.
Df Residuals:	19255	**BIC:**	4516.
Df Model:	5		
Covariance Type:	nonrobust		

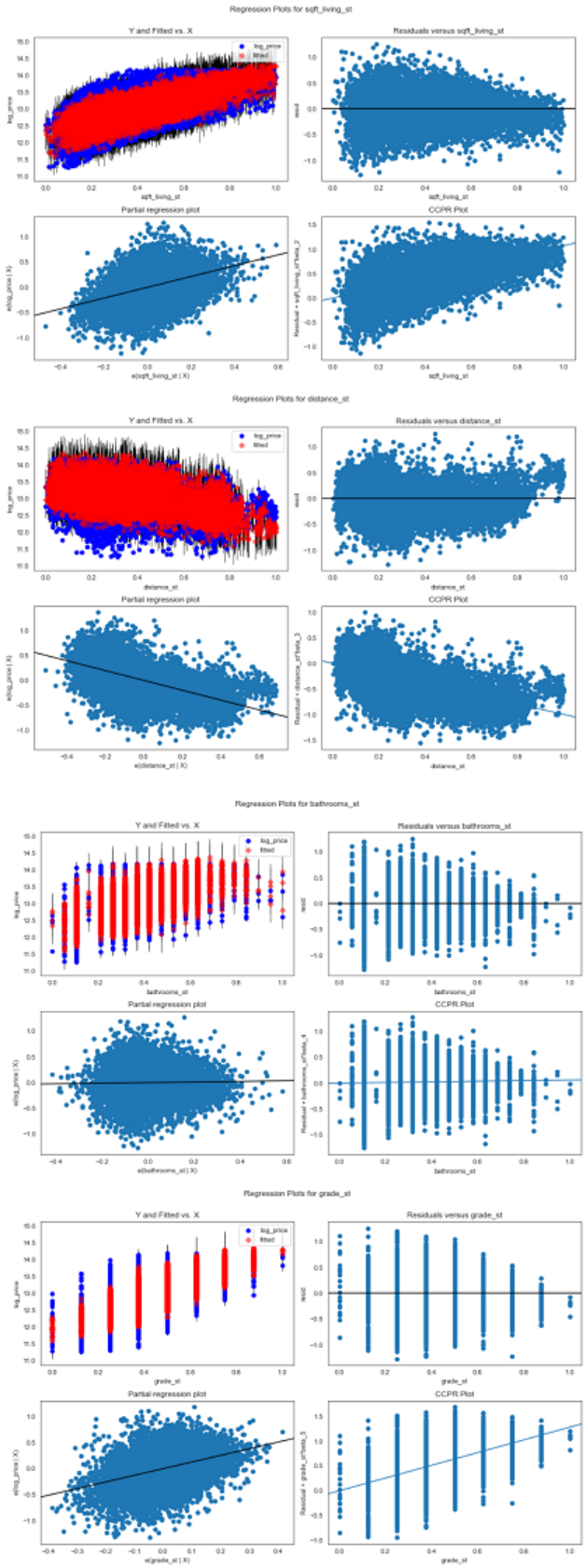
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.3621	0.008	1602.897	0.000	12.347	12.377
recent_renovation_new	-0.0345	0.006	-5.875	0.000	-0.046	-0.023
sqft_living_st	1.0747	0.018	60.141	0.000	1.040	1.110
distance_st	-0.9931	0.009	-106.878	0.000	-1.011	-0.975
bathrooms_st	0.0598	0.019	3.072	0.002	0.022	0.098
grade_st	1.2805	0.022	58.089	0.000	1.237	1.324

Omnibus:	310.179	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	482.427
Skew:	-0.166	**Prob(JB):**	1.75e-105
Kurtosis:	3.700	**Cond. No.**	16.5

QQ plot of the residuals



Regression plots and visualization of homoscedasticity



The final linear regression model of a log of the price variable versus grade, bathrooms, distance from the center of the city, sqft_living space, and the indicator if a house has been renovated recently or a newer house has a Coefficient of Determination of 0.663. It is indicative of the fact that 66.3% of the sold properties fall within the results of the line formed by the regression equation. F-statistics displays the high value and the overall p-value much lower than the confidence interval

The independent variables used in the equation display a clear linear relationship with the target and homoscedasticity.

The next step is to validate the model by using training and test datasets

Train the model

The model was trained by using SciKit Learn train_test_split model resulting in the following parameters:

$$\ln(Price) = 12.366 + 1.265 \cdot (grade) + 1.080 \cdot (sqft\_living) - 0.989 \cdot (distance) + 0.060 \cdot (bathrooms) - 0.035 \cdot (recent\_renovation\_new)$$

Validation

> Train MSE: 0.272
> Test MSE: 0.272
> Are equal down to the third decimal digit indicating a good agreement between the training and the test sets
>
> **R** squared for for the prediction on the full dataset is the same as in model_4_3: 0.663
> **Mean Absolute Error** for the prediction on the full dataset is 0.213 which is not great but acceptable. The best possible theoretical value is 0.
> **Mean Squared Error** for the prediction on the full dataset is 0.074. The >best possible theoretical value is 0.

iNterpret

The final model has a reasonable predictive ability tested in the final step of the model validation. MSE, MAE, and R2 score along with the model p-values for all predictors indicate a good fit.

The most influential predictor is a **building grade**, following by a **living space footage**. Both factors are **positively correlated** with the price of the property. Both factors are within property owners' control when they are renovating their houses.
A **distance** from the center of the city is **negatively correlated** with the price of property, meaning the further away a property is, the lower is the price. It is not a controllable variable but is helpful for home buyers if the living space and the number of bedrooms/bathrooms are important.
A **number of bathrooms** has a **positive effect** on the price of a property, but it is not as strong as the first two factors. This fact indicates that the convenience of having multiple bathrooms is essential for potential buyers and should be taken into account when owners are planning a renovation.
The last predictor in the model is an indicator of whether a property **has been renovated recently or a new construction**. It is very **weakly negatively correlated** with the price variable. The negative correlation (reduction of the price) might be related to the following factors: newer properties, on the average, are of less building quality.
The intercept of the model is a **bias** of the model and can be interpreted as an offset of the model due to other factors not taken into account for various reasons.

Conclusions and Recommendation

Recommendations to property owners planning a renovation to their properties:

- Increase the living space of your property
- Do the renovation with higher building quality
- Consider adding a bathroom

Recommendations to potential buyers:

- Look for properties further away from the city center to make the best out of your property buying budget
- Properties in some zipcodes of the city are more affordable than others at the same distance from the city center
- Properties in some zipcodes of the city are more affordable than others with a better view, more considerable property lots, and with older houses of better quality construction if these factors are essential to a buyer

Limitations of the model:

- The original dataset does not include other important factors, and therefore the model is biased
- Multiple linear regression models, while easily interpretable, are limited in their predictive ability
- Some variables in the dataset are strongly correlated with each other, and that affect the predictive power of the model

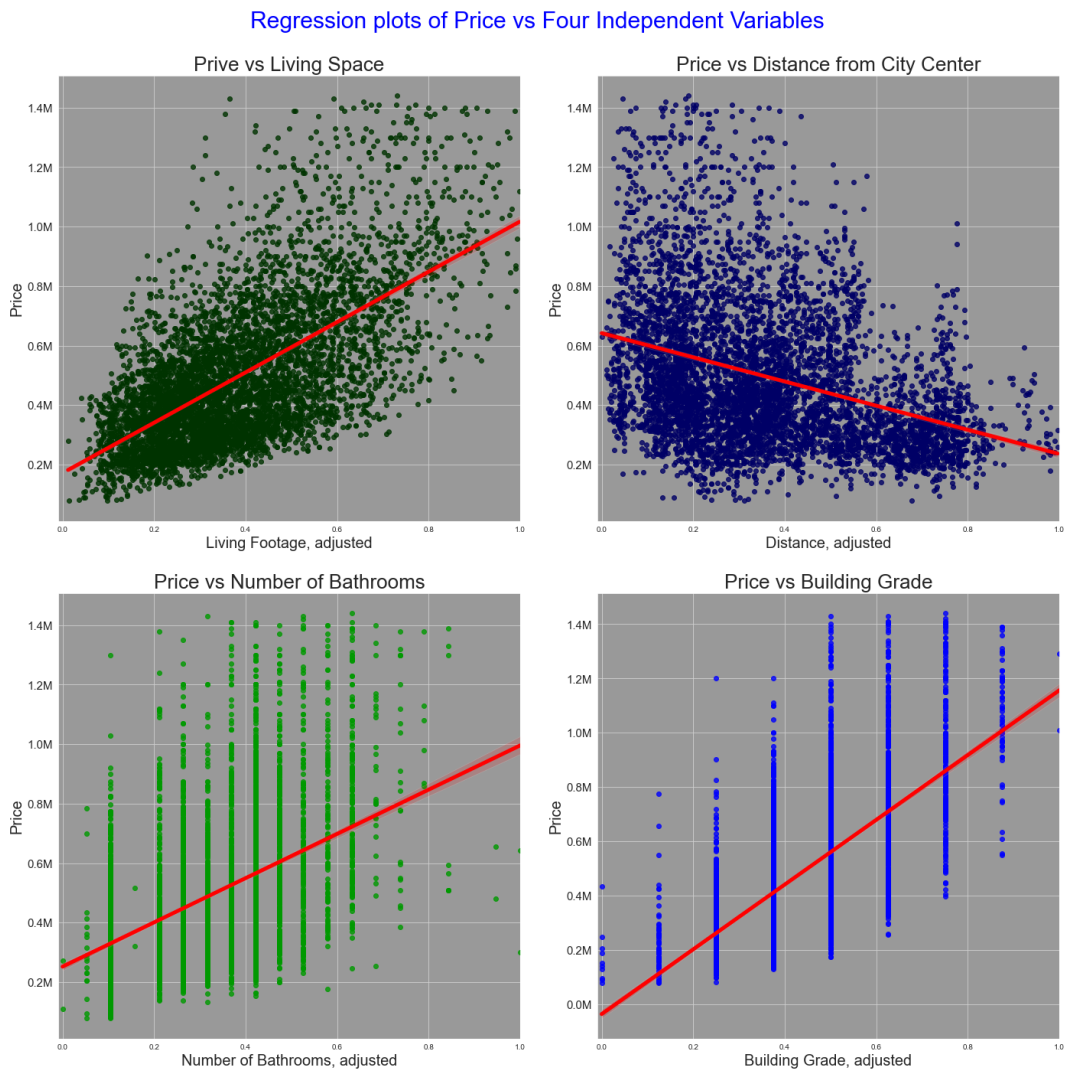
Suggestion for future improvements:

- Add variables to the original dataset like kitchen renovation, average commute time, crime index, average nearby public school quality, etc.
- Update the dataset with more current data

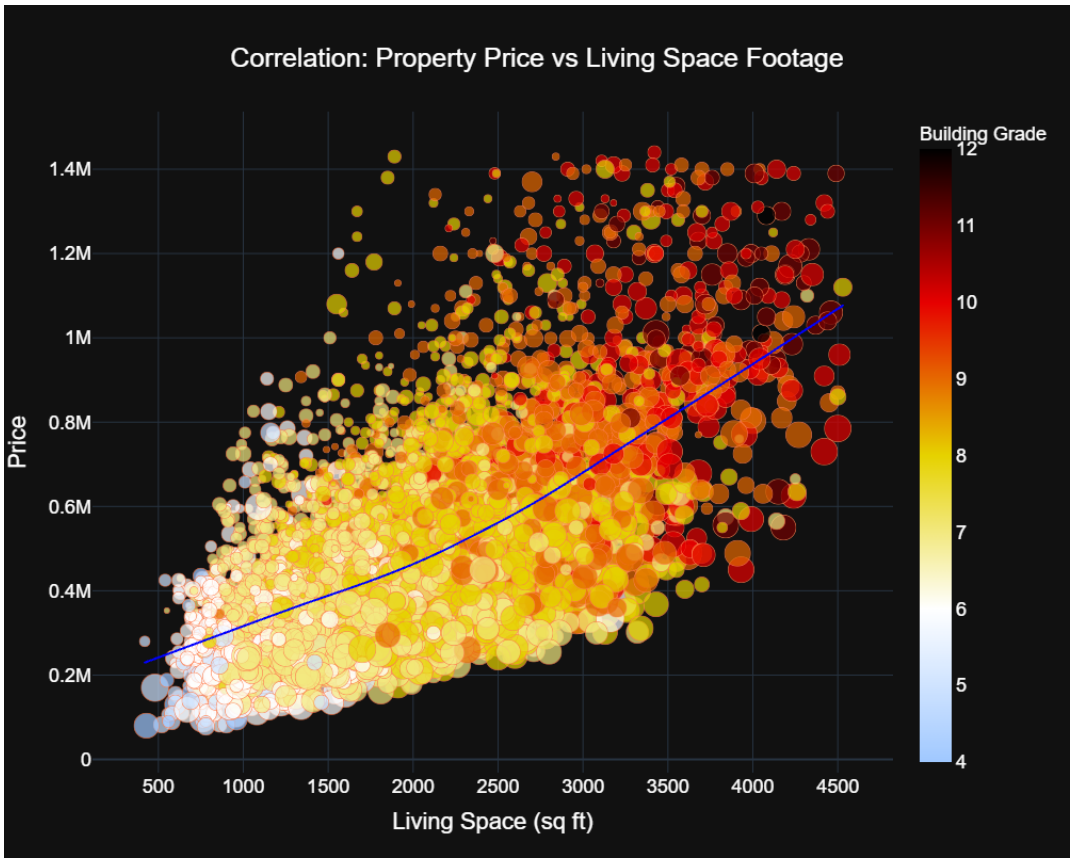
Appendix

Visualization

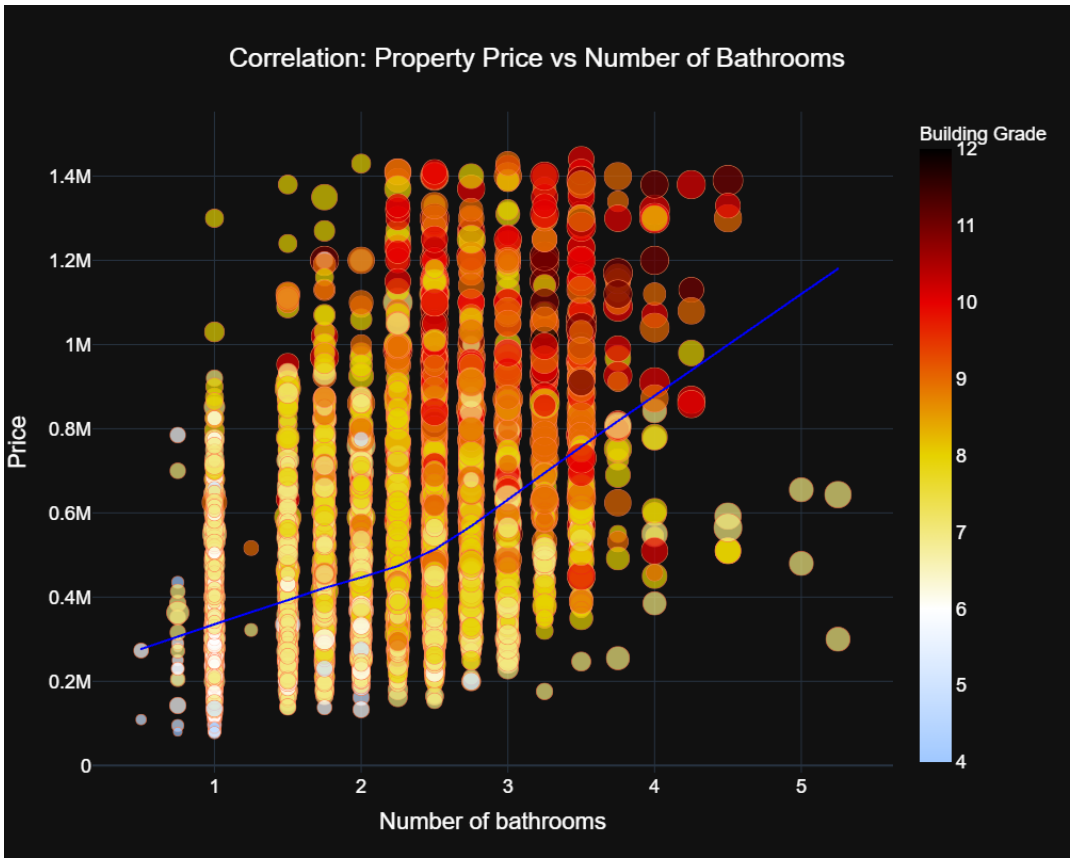
Regplots for all four variables



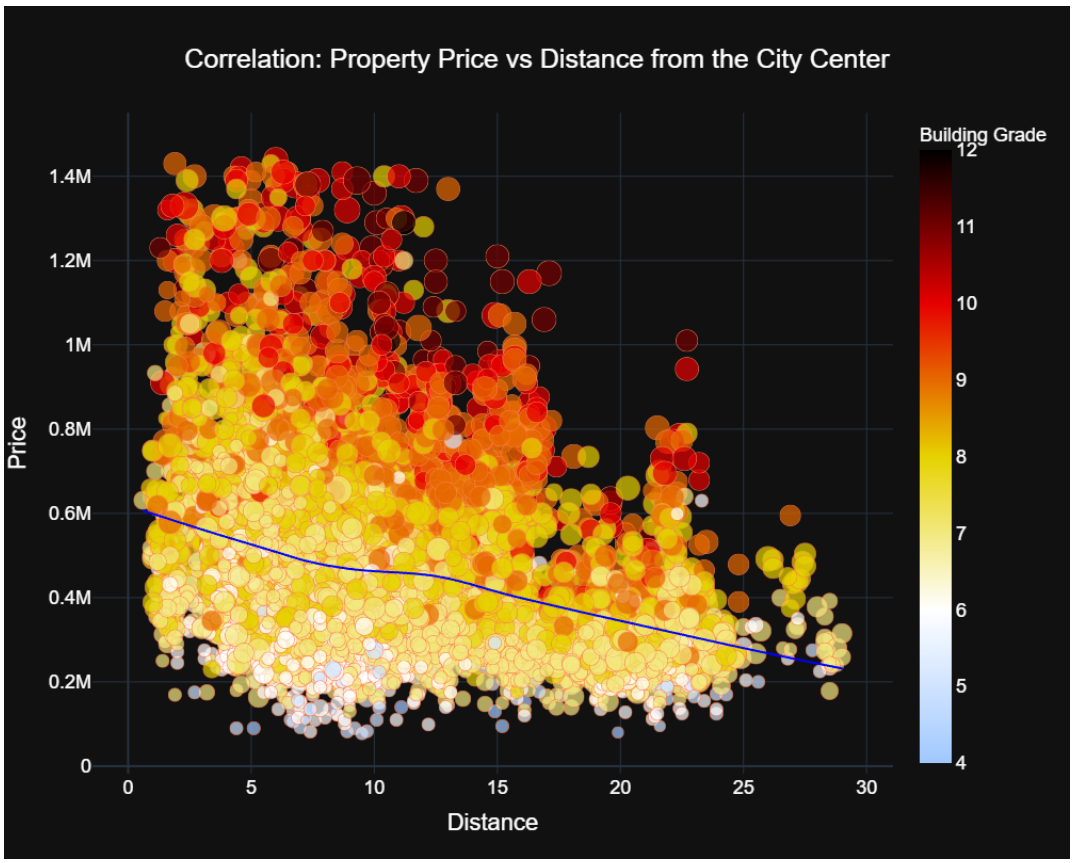
Correlation: Property Price vs Living Space Footage



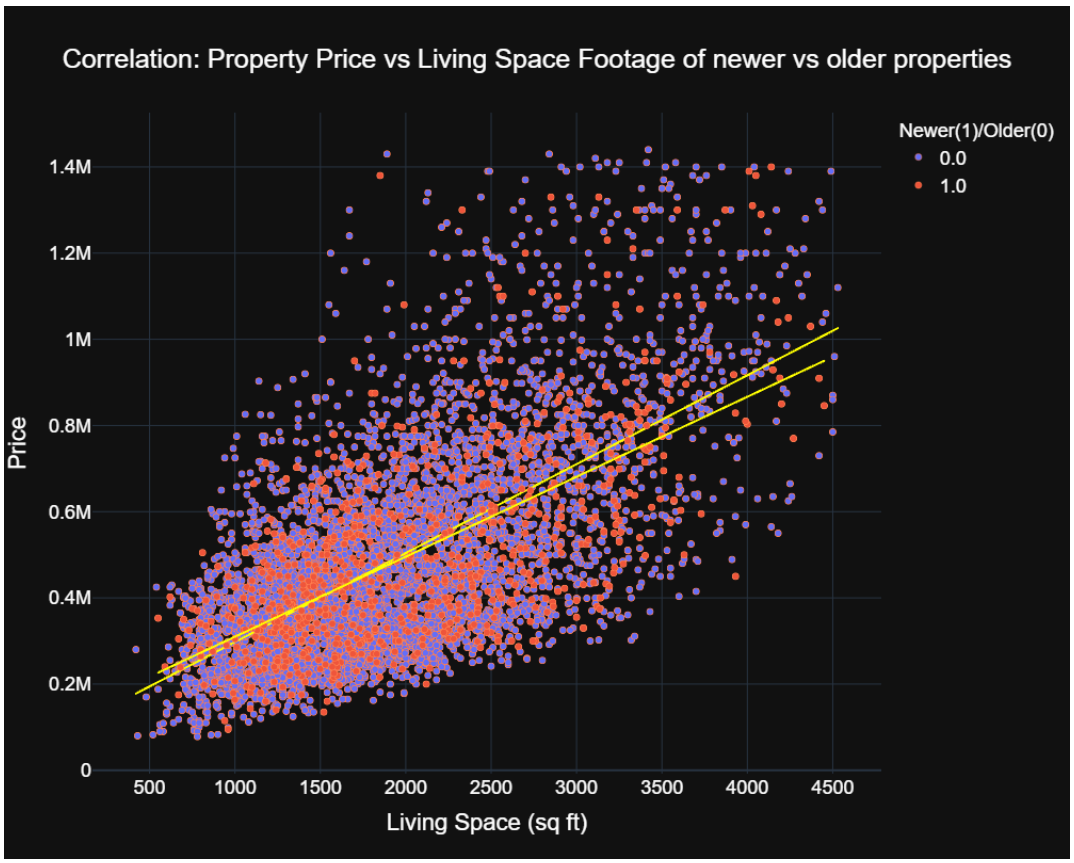
Correlation: Property Price vs Number of Bathrooms



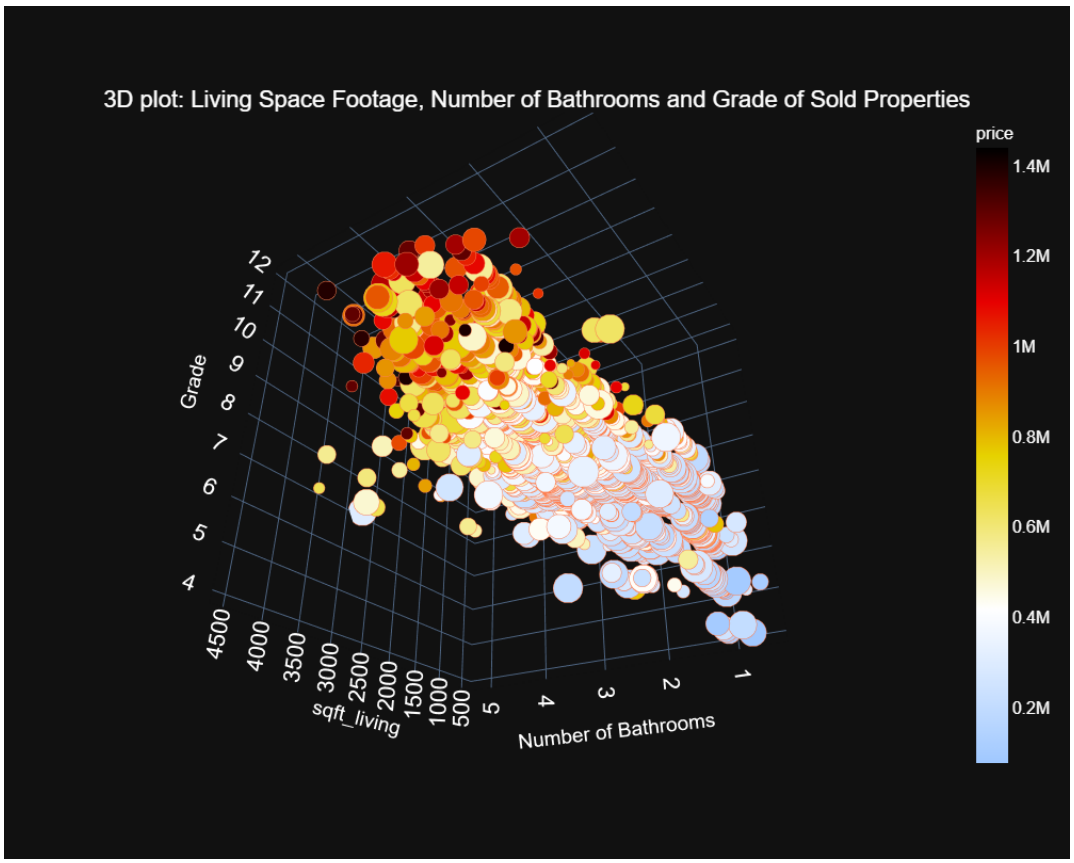
Correlation: Property Price vs Distance from the City Center



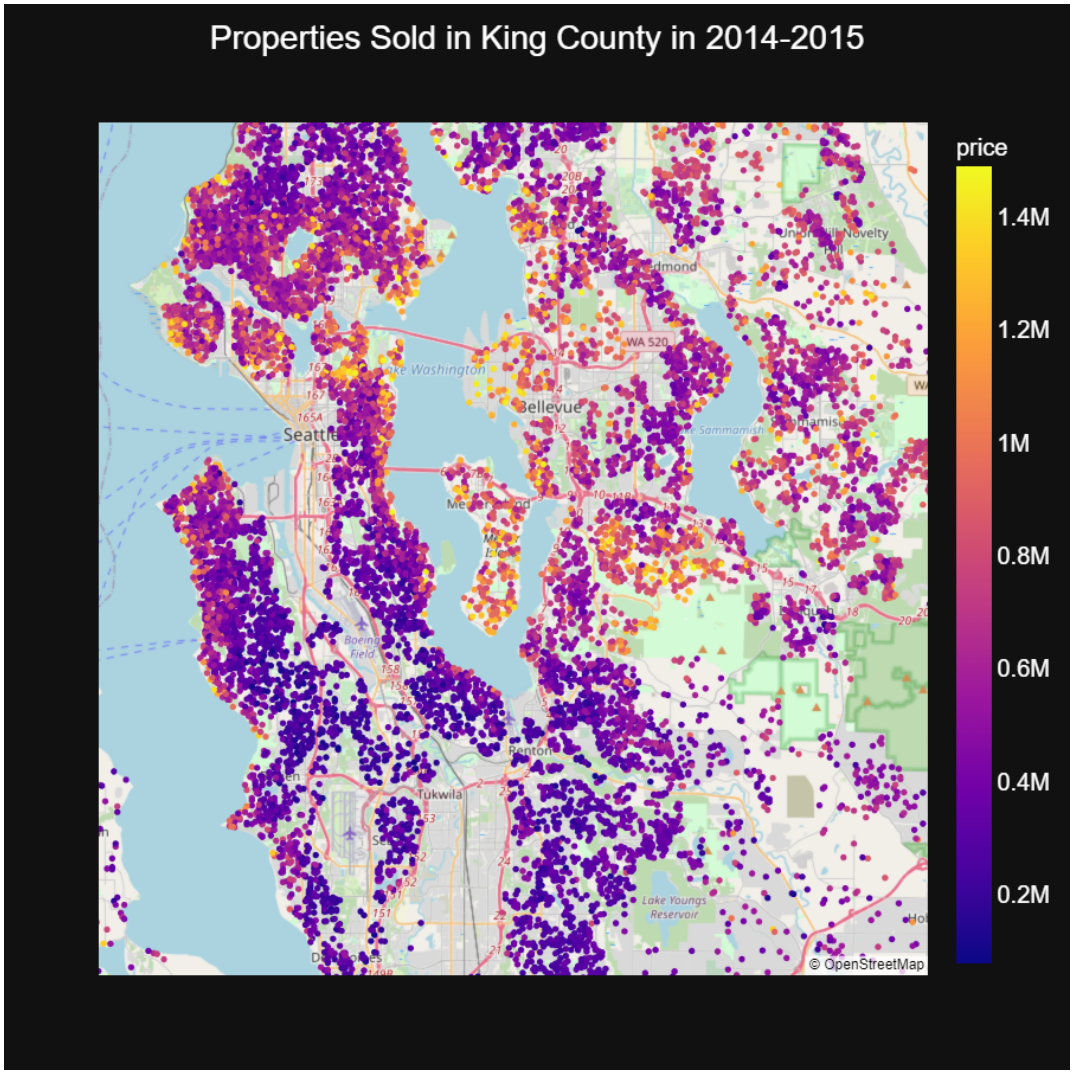
Correlation: Property Price vs Living Space Footage of newer vs older properties



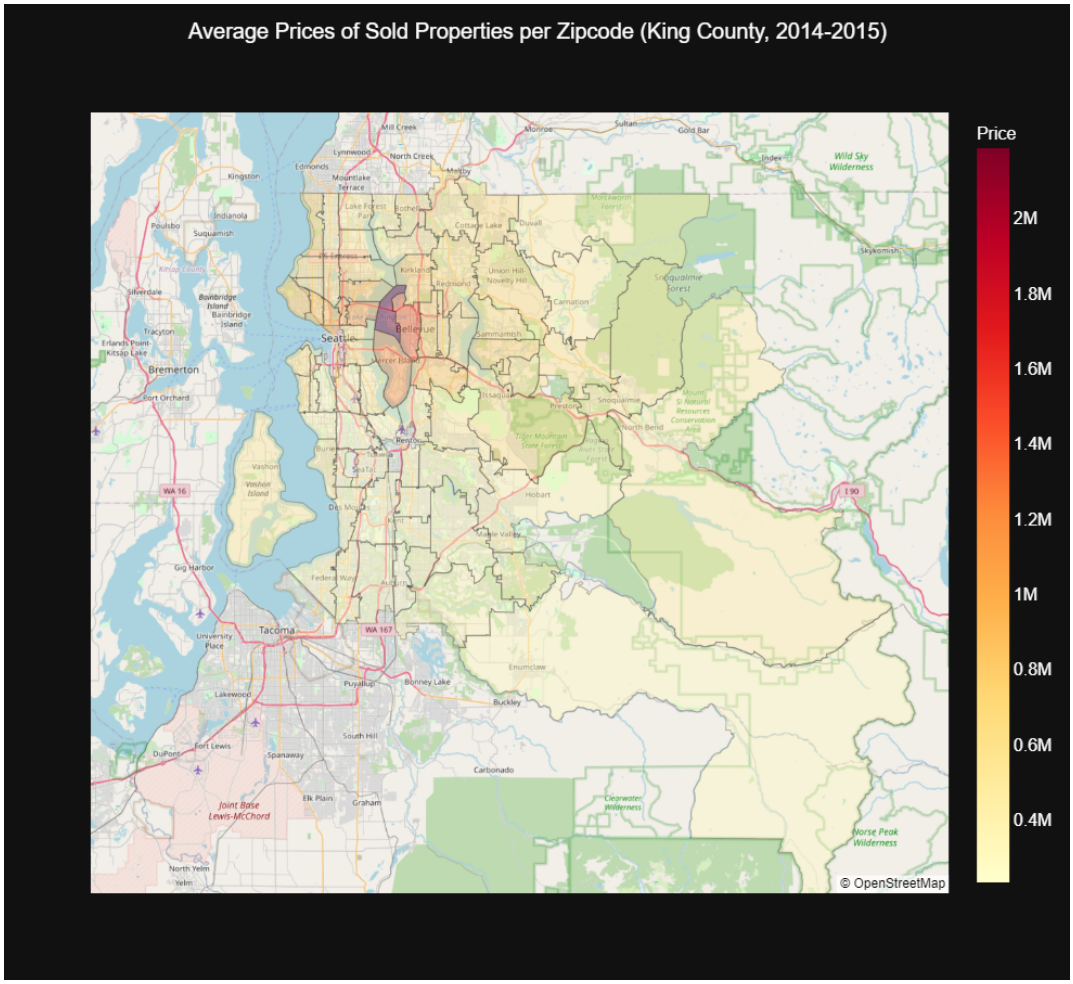
3D plot: Living Space Footage, Number of Bathrooms and Grade of Sold Properties



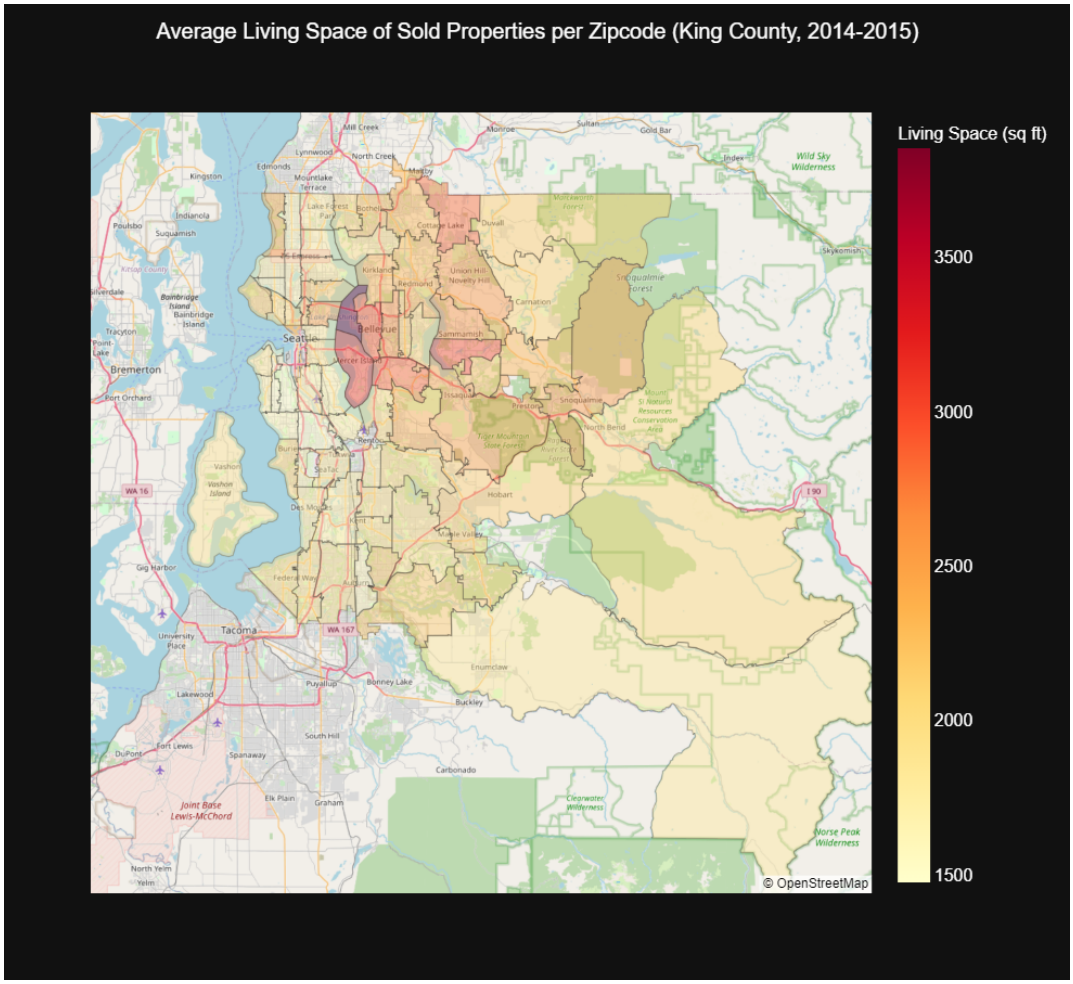
Map of Properties Sold in King County in 2014-2015



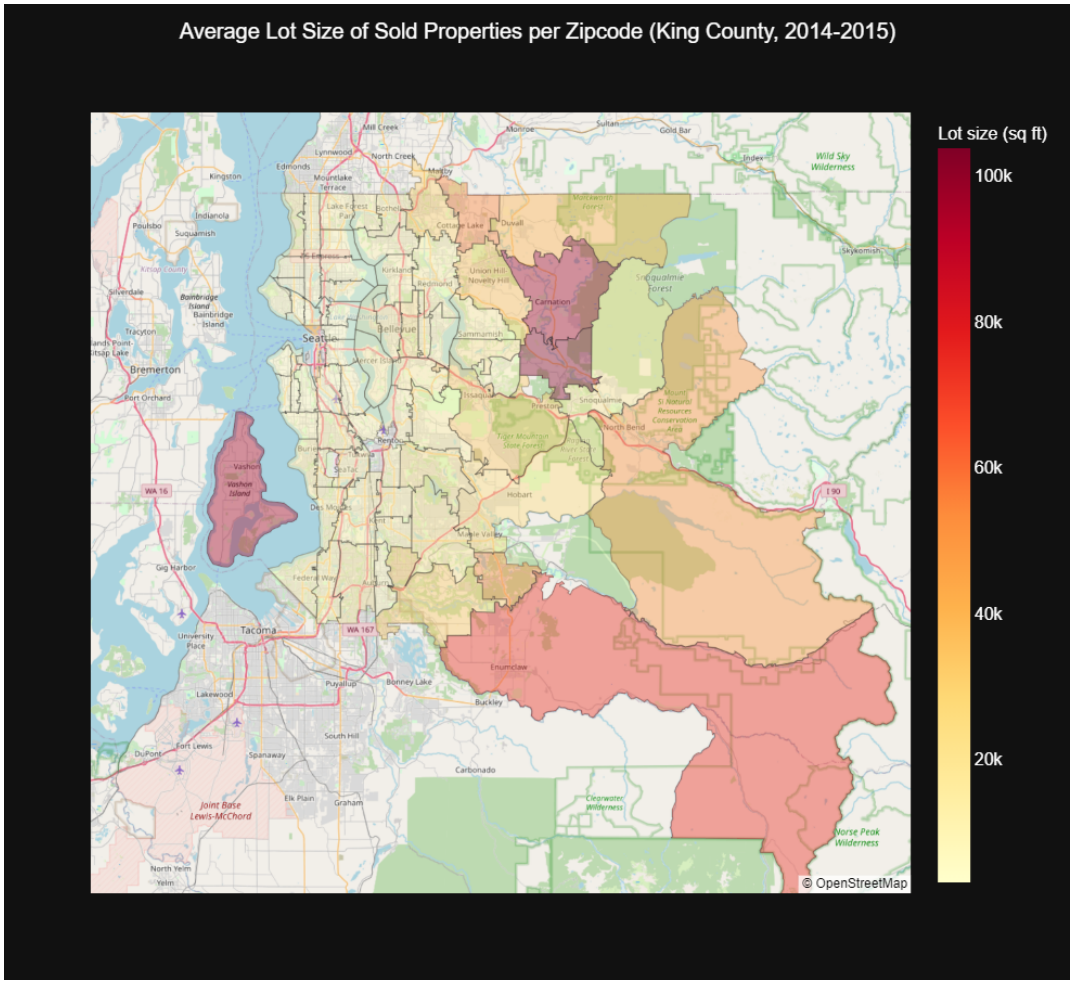
Average Prices of Sold Properties per Zipcode (King County, 2014-2015)



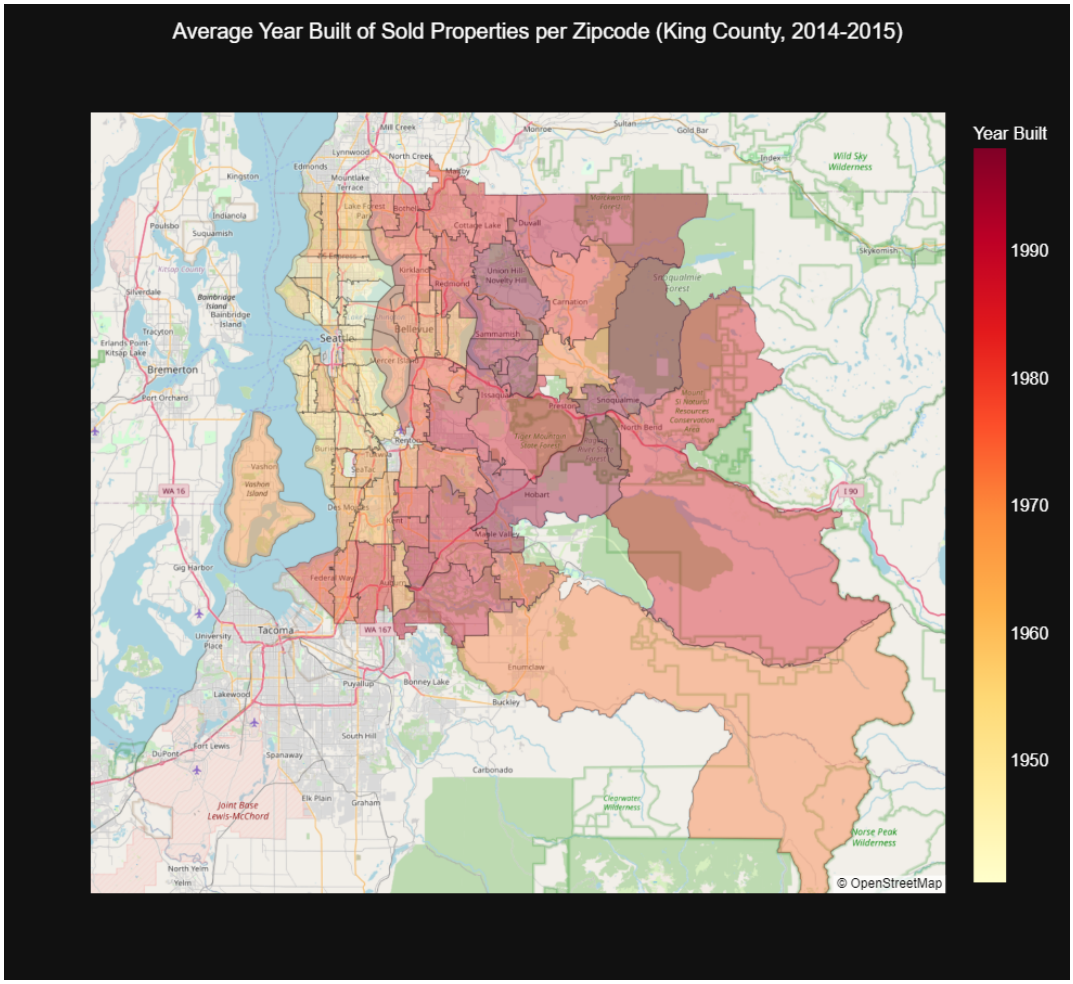
Average Living Space of Sold Properties per Zipcode (King County, 2014-2015)



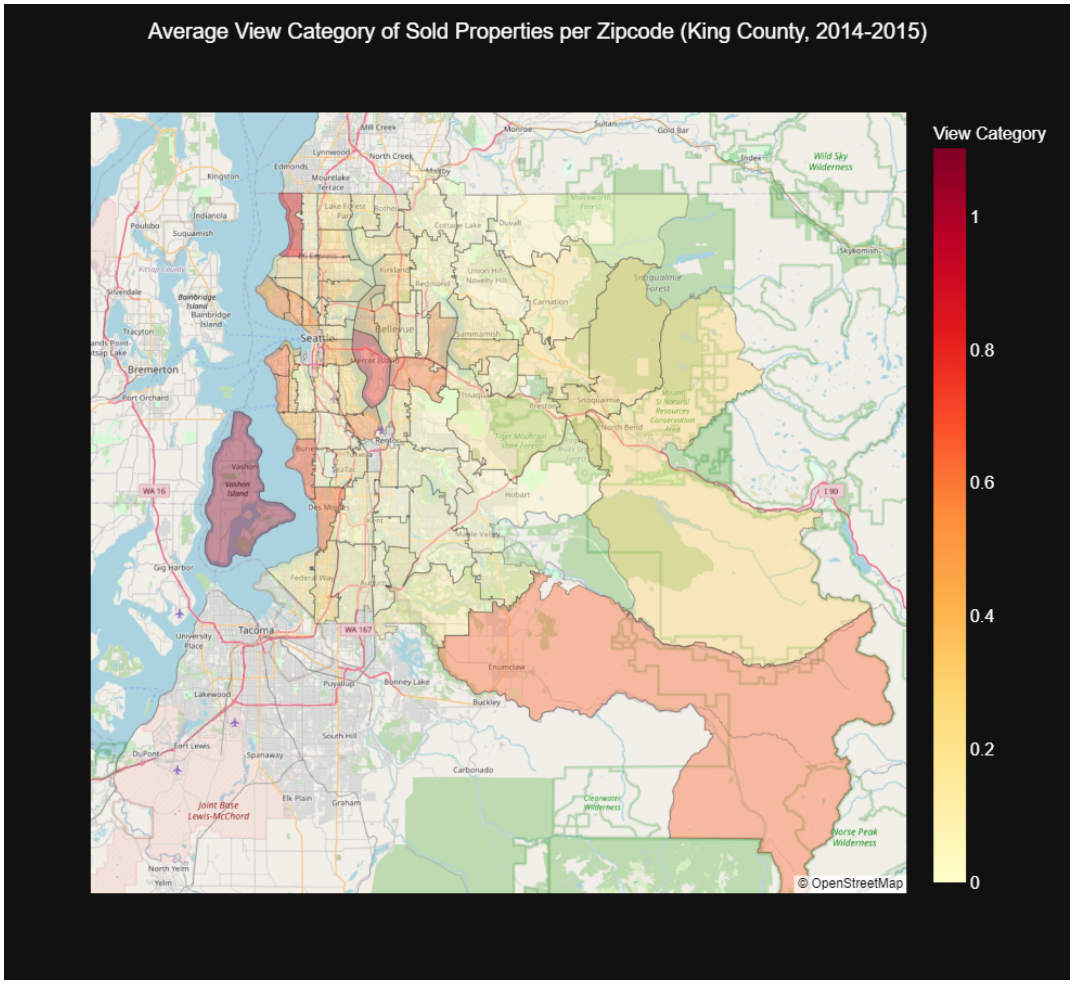
Average Lot Size of Sold Properties per Zipcode (King County, 2014-2015)



Average Year Built of Sold Properties per Zipcode (King County, 2014-2015)



Average View Category of Sold Properties per Zipcode (King County, 2014-2015)



For More Information

Please review our full analysis in our Jupyter Notebook or our presentation.

For any additional questions, please contact **Elena Kazakova @ e.v.kazakova@gmail.com**

Repository Structure

```
├── LICENSE.md                                  <- Learn.co Educational Content
License
├── README.md                                   <- The top-level README for
reviewers of this project
├── README.pdf                                  <- The top-level README for
reviewers of this project in PDF format
├── project2_OSEMN_plus_mvp_final.pdf           <- Final Project Jupyter
notebook as a PDF file
├── project2_OSEMN_plus_mvp_final.ipynb         <- Final Project Jupyter
notebook
├── project2_OSEMN_plus_mvp_final_backup.ipynb  <- The latest backup of the
notebook
├── DS_Phase2_Project_Presentation.pdf          <- PDF version of project
presentation
├── directory structure.txt                     <- Text fiel with this
directory tree
├── repo.pdf                                    <- Github repo structure image
├── old_files                                   <- Files not in use
├── data                                        <- original data files and
GeoJSON file
└── images                                      <- Both sourced externally and
generated from code
```