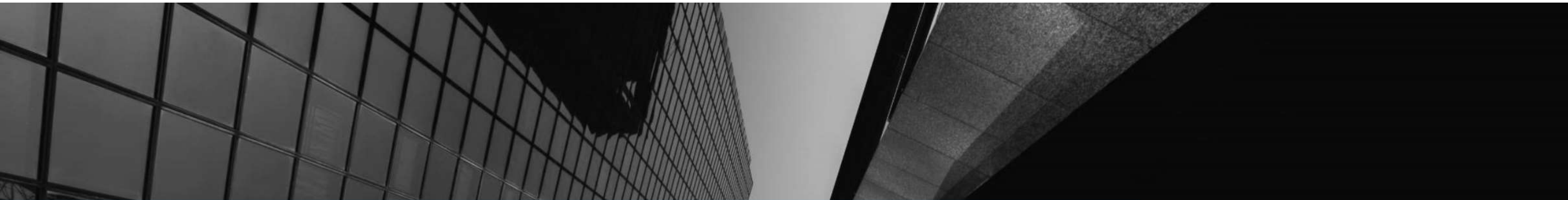
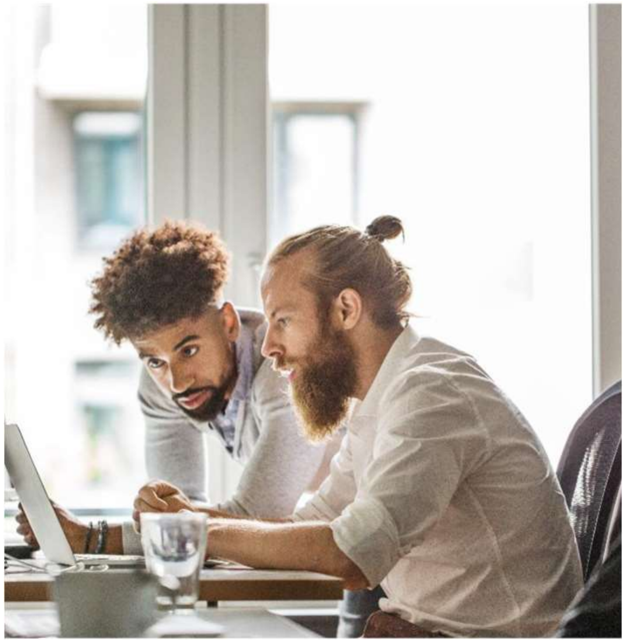


Analysis of Property Sales

Ryan Chung

Outline



Introduction

Business Problem

Data & Methods Used

Results & Conclusion

Final Thoughts

Project Introduction

Business Problem

A local real estate agency is looking to identify features of house listings that act as accurate predictors of their selling prices to take into consideration when purchasing and selling residential property.

Data Used

For this project, the real estate agency provided a data set which was comprised of residential property listings with various features.

Some of the features in the data set include housing aspects such as number of bedrooms, bathrooms, and floors.

Methods

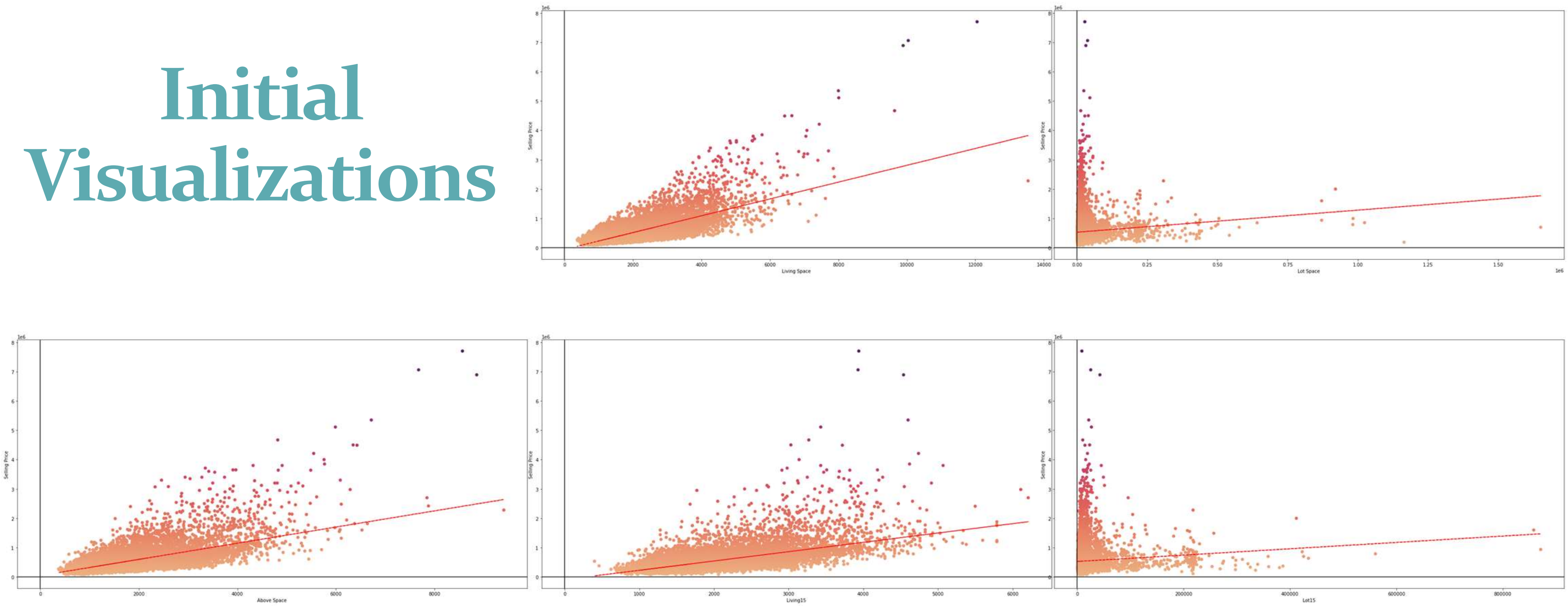
Initial Steps

- Upon receiving the data set, it was cleaned and checked for null values.
- After cleaning the data, basic visualizations were created for several housing features.
- Three simple linear regression models were created based upon the insights from the basic visualizations.
- After looking at various housings features and their impact on their selling prices, the decision to make a multiple linear regression was made.

Final Model

- The multiple regression model encompassed all housing features and was created using an iterative approach.
- A train test split approach was utilized, features were scaled, and the model was then checked for multicollinearity along with its distribution of error terms.
- Upon the completion of the finalized model, R2 values were compared and the RMSE was evaluated.
- Several housing features were identified to be accurate predictors of a house's selling price.

Initial Visualizations



Initial MLR OLS Summary + VIF Values

```
LR_1.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.610
Model:	OLS	**Adj. R-squared:**	0.610
Method:	Least Squares	**F-statistic:**	1407.
Date:	Mon, 09 Aug 2021	**Prob (F-statistic):**	0.00
Time:	20:40:09	**Log-Likelihood:**	22072.
No. Observations:	10800	**AIC:**	-4.412e+04
Df Residuals:	10787	**BIC:**	-4.402e+04
Df Model:	12		
Covariance Type:	nonrobust		

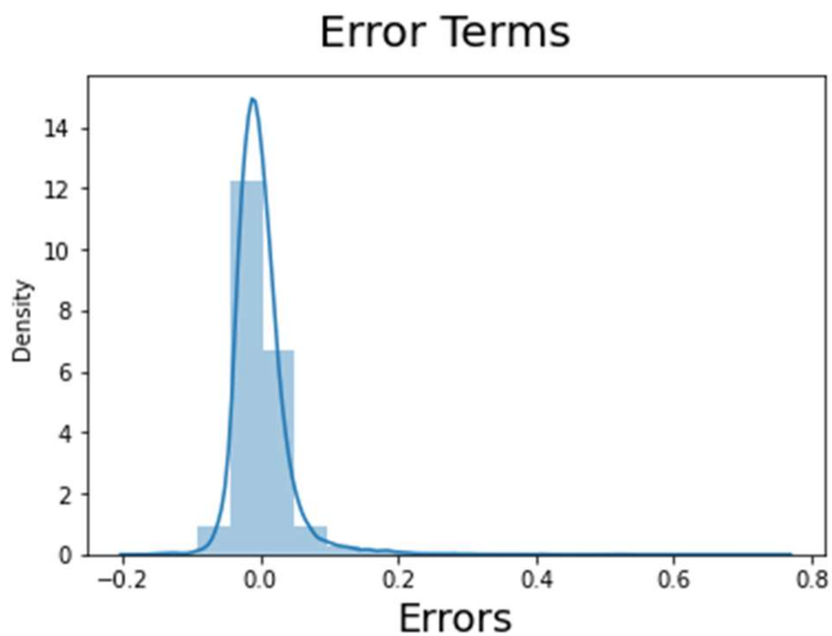
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0495	0.002	-24.282	0.000	-0.054	-0.046
bedrooms	-0.1571	0.013	-12.159	0.000	-0.182	-0.132
bathrooms	-0.0098	0.005	-1.941	0.052	-0.020	9.6e-05
sqft_living	0.1354	0.003	38.861	0.000	0.129	0.142
sqft_lot	0.0107	0.017	0.615	0.538	-0.023	0.045
floors	-0.0038	0.002	-1.972	0.049	-0.008	-2.2e-05
waterfront	0.0838	0.004	21.758	0.000	0.076	0.091
view	0.0337	0.002	18.119	0.000	0.030	0.037
condition	0.0279	0.002	14.345	0.000	0.024	0.032
grade	0.1250	0.005	27.239	0.000	0.116	0.134
sqft_above	0.1613	0.005	34.035	0.000	0.152	0.171
sqft_basement	0.0786	0.003	24.002	0.000	0.072	0.085
sqft_living15	-0.0024	0.004	-0.576	0.564	-0.011	0.006
sqft_lot15	-0.0887	0.014	-6.516	0.000	-0.115	-0.062

Omnibus:	8500.808	**Durbin-Watson:**	2.006
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	661410.477
Skew:	3.207	**Prob(JB):**	0.00
Kurtosis:	40.798	**Cond. No.**	2.42e+16

```
vif
```

	Features	VIF
2	sqft_living	inf
9	sqft_above	inf
10	sqft_basement	inf
8	grade	31.62
11	sqft_living15	16.67
1	bathrooms	15.90
0	bedrooms	10.62
7	condition	9.73
4	floors	3.45
12	sqft_lot15	2.70
3	sqft_lot	2.53
6	view	1.48
5	waterfront	1.20

Final MLR OLS Summary + VIF Values



	Features	VIF
6	sqft_lot15	2.65
0	sqft_lot	2.52
4	condition	2.37
5	sqft_basement	1.66
1	floors	1.63
3	view	1.42
2	waterfront	1.20

Final MLR OLS Summary

```
print(LR_7.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.330
Model:                            OLS   Adj. R-squared:                  0.329
Method:                 Least Squares   F-statistic:                     758.3
Date:                Mon, 09 Aug 2021   Prob (F-statistic):               0.00
Time:                        21:11:11   Log-Likelihood:                 19146.
No. Observations:               10800   AIC:                        -3.828e+04
Df Residuals:                   10792   BIC:                        -3.822e+04
Df Model:                           7
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.0193      0.002     11.162      0.000       0.016       0.023
sqft_lot          0.0741      0.023      3.266      0.001       0.030       0.119
floors            0.0786      0.002     40.197      0.000       0.075       0.082
waterfront        0.0850      0.005     16.871      0.000       0.075       0.095
view              0.0640      0.002     26.988      0.000       0.059       0.069
condition         0.0165      0.003      6.509      0.000       0.012       0.021
sqft_basement     0.1438      0.004     36.260      0.000       0.136       0.152
sqft_lot15        0.0497      0.018      2.813      0.005       0.015       0.084
==============================================================================
Omnibus:                     9602.168   Durbin-Watson:                   2.039
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           793601.084
Skew:                           3.911   Prob(JB):                         0.00
Kurtosis:                      44.260   Cond. No.                         81.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Confirmation of R2 Values and RMSE Value

Images from JYNB

Check R2 values to confirm if final predicted model is best fitted or not.

```
from sklearn.metrics import r2_score
```

```
r2_score(y_true = y_test, y_pred = y_pred_M7)
```

```
0.33375961922542563
```

```
# With R2 values almost equal the model is best fitted!!!
```

Check RMSE

```
import sklearn
import math
```

```
mse = sklearn.metrics.mean_squared_error(y_test, y_pred_M7)
```

```
rmse = math.sqrt(mse)
```

```
print(rmse)
```

```
0.03750663037047103
```

When comparing the two computed R2 values, the final model was confirmed to be best fitted.

Using the sklearn library, the RMSE value was found to be 0.0375.

Results & Conclusions

Reviewing the final multiple regression OLS summary, the features indicated to be accurate predictors of a house's selling price include the following: Square footage of property's lot, number of floors, waterfront, view, condition of property, square footage of basement, & square footage of nearest 15 neighbor's lot.

Returning to the initial business problem, it can be said that the local real estate agency should look at these features of residential housing listings when purchasing and selling property.

- Data Cleaned + Visualized (EDA)
- Initial Model Checked for Multicollinearity + Error Term Distribution
- R2 Values of Model Compared + RMSE Calculated
- Features Identified as Accurate Predictors for Business Recommendations.

Thank You!