

 jbloewencolon / Phase-2-Project---The-Garage-Problem

Public

This project is the phase 2 data analysis and presentation for the Flatiron Data Science program

 0 stars

 0 forks


 Star


 Unwatch ▾


 Code


 Issues

 Pull requests

 Actions

 Projects

 Wiki


 Security


 Insights

 Settings

 main ▾



 jbloewencolon Add files via upload ...

now  40

[View code](#)

 README.md 

The Garage Problem: Real Estate Recommendations for Kings County, Washington



By Jordan Loewen-Colón October 14th 2022

The Business Problem

King's County Realtors are interested in whether or not they should renovate homes before trying to sell. Specifically, they'd like to know how much adding a garage might affect price, and if so, what size of garage.

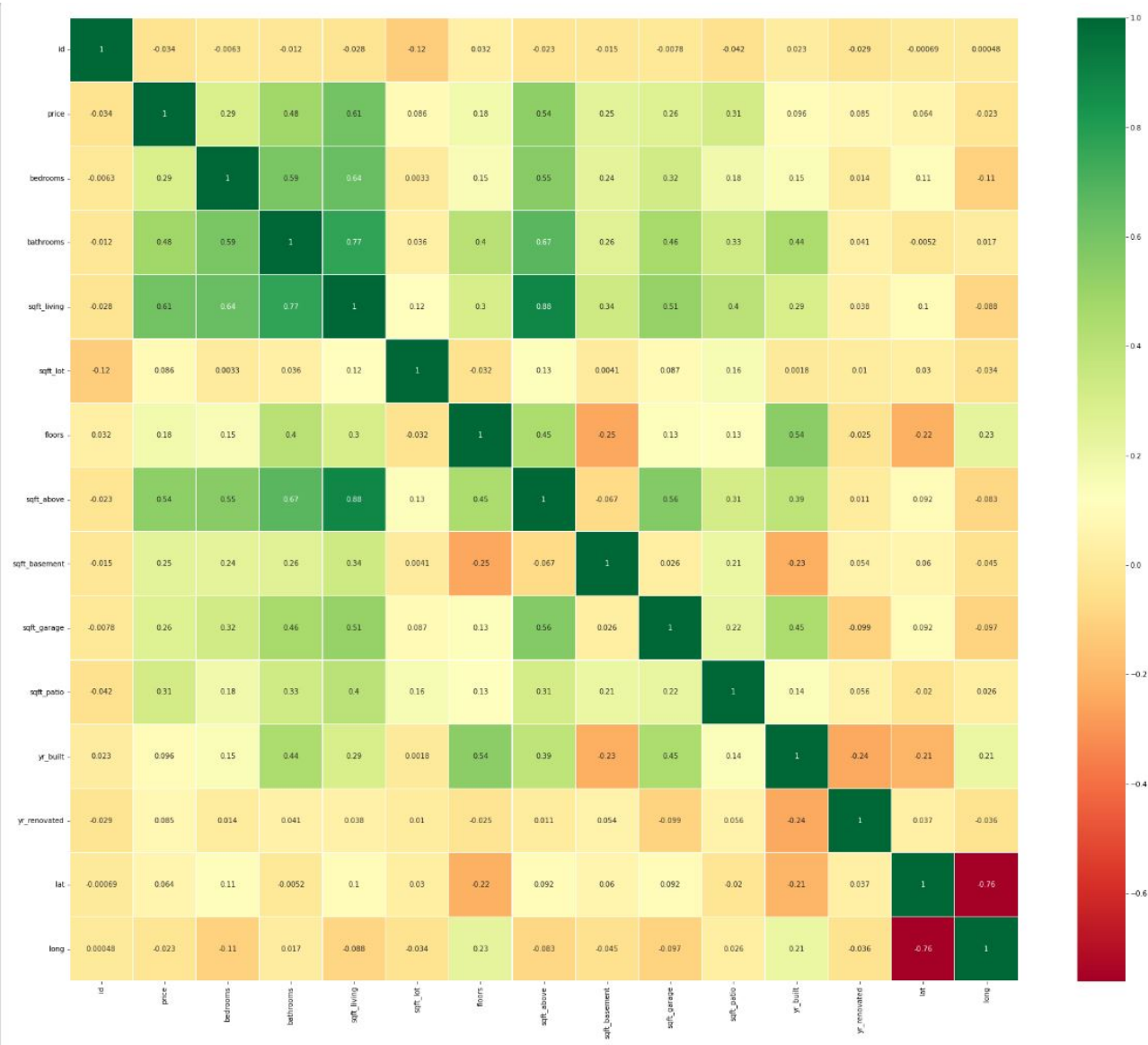
Recommendations:

Based on our models and analysis, we recommend that if renovations are going to occur, it's best to target square footage of living space, but if renovations are going to include the garage, it is probably worth focusing on homes that have no garage and adding a 1-car sized garage, rather than increasing the size of an existing garage.

Step 1: Data Understanding

To make our recommendations, we analyzed the 2022 data from King's County.

The dataset has 30155 entries and 25 columns with a mix of string values, floats, and integers. Bathrooms as float makes sense, but "floors" as float seems odd. It was also not clear what elements were contained in some of the object categories like "grade" or "nuisance." There were also some missing entries for "sewer_system" and "heat_source." Some odd things we noticed initially is that there were houses without bedrooms or bathrooms. There were also outliers on the larger end as well, with houses containing 13 bedrooms and 10.5 bathrooms. There was even a house listed with only 3sqft of living space. It also became clear that we were going to need to seperate out the houses that already have garages from those that do not.

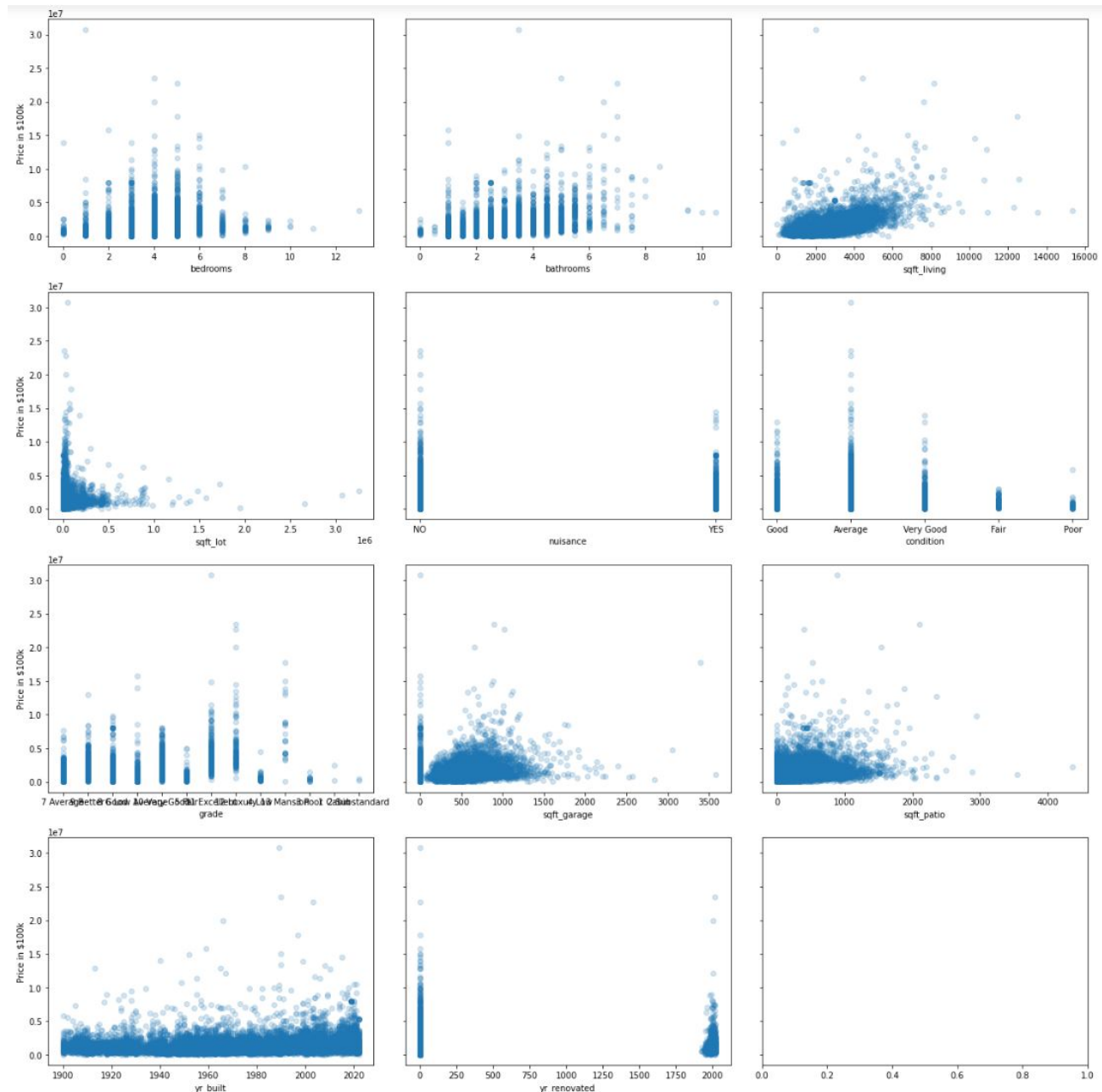


Given that are focus is on garages, it's important to note that there are both houses with no garages or which have never been renovated. In narrowing down, we noticed that columns like "waterfront" and "greenbelt" had a small number of entries, so they probably would not add much to our analysis. Additionally, given time constraints, we wouldn't be able to spend time on categories like "date, view, sqft_above, sqft_basement, address, lat, and long." We ended up focusing primarily on "yr_renovated," "condition," "grade," and, especially, "sqft_living" as categories in order to build the most accurate model.

Step 2: Data Preperation

In preparing the data, we focused on elements we thought would affect our numbers involving renovations broadly, and garage additions more specifically. Since we were dealing with both continuous numbers and integers, experimented with log transformations, but didn't find their adjustments useful and getting a more accurate model. To check the accuracy of our model, we focused primarily on the correlation between price per sqft of house. According to according to the website www.redfin.com the average price per sqft is \$453. So we thought that would be a good target for checking the accuracy of our data prep.

With price as our major target, we set it as our Y and looked for correlations with the other columns as our x's.



Since scatterplots are more useful at visualizing the relationship with continuous data, we adjusted columns like "grade" and "condition" by turning them from strings to integers which allowed our model to see the relationship between the numbers more clearly and hopefully give us more accuracy. We then dealt with outliers which filtered out 2513 rows from our set of 30155.

We then created new columns checking for which houses had garages, and which ones didn't, in addition to a new column for garages based on size (either 1-car, 2-car, or 3-car).

Data Modeling

Now with our data prepared and in hand, we created some models. First, we checked to see the accuracy of our data preparation by checking price per sqft.

```

=====
                        OLS Regression Results
=====
Dep. Variable:          price    R-squared:                0.397
Model:                  OLS      Adj. R-squared:            0.397
Method:                 Least Squares    F-statistic:          1.818e+04
Date:                  Thu, 12 Jan 2023    Prob (F-statistic):    0.00
Time:                  18:11:14    Log-Likelihood:       -4.0038e+05
No. Observations:      27642    AIC:                  8.008e+05
Df Residuals:          27640    BIC:                  8.008e+05
Df Model:               1
Covariance Type:       nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	1.16e+05	7431.206	15.607	0.000	1.01e+05	1.31e+05
sqft_living	455.7727	3.380	134.841	0.000	449.148	462.398

```

=====
Omnibus:                 9877.343    Durbin-Watson:           1.971
Prob(Omnibus):           0.000    Jarque-Bera (JB):        59218.875
Skew:                    1.598    Prob(JB):                 0.00
Kurtosis:                 9.419    Cond. No.                 5.75e+03
=====

```

So it looks like the coefficient is within the right range, but the R-squared is very low. However, it seems that transforming the number logarithmically actually lowers the R-value. So we kept it as it was and then added our other variables until we achieved the highest R-value given our selected data:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.534
Model:	OLS	Adj. R-squared:	0.533
Method:	Least Squares	F-statistic:	1374.
Date:	Thu, 12 Jan 2023	Prob (F-statistic):	0.00
Time:	18:11:14	Log-Likelihood:	-3.9682e+05
No. Observations:	27642	AIC:	7.937e+05
Df Residuals:	27618	BIC:	7.939e+05
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	8.985e+06	2.27e+05	39.664	0.000	8.54e+06	9.43e+06
sqft_living	263.9585	5.930	44.511	0.000	252.335	275.582
bedrooms	-3.928e+04	3795.404	-10.350	0.000	-4.67e+04	-3.18e+04
bathrooms	8.194e+04	5161.863	15.874	0.000	7.18e+04	9.21e+04
yr_built	-4302.6286	114.421	-37.604	0.000	-4526.899	-4078.358
renovated_1	6.951e+04	1.3e+04	5.352	0.000	4.41e+04	9.5e+04
garage_size_1	3887.3469	9216.170	0.422	0.673	-1.42e+04	2.2e+04
garage_size_2	-3.84e+04	6517.193	-5.892	0.000	-5.12e+04	-2.56e+04
garage_size_3	-1.124e+05	1.49e+04	-7.547	0.000	-1.42e+05	-8.32e+04
grade_1	-1.641e+05	4.21e+05	-0.390	0.696	-9.89e+05	6.61e+05
grade_2	3.783e+05	4.21e+05	0.899	0.368	-4.46e+05	1.2e+06
grade_3	-8.124e+04	8.91e+04	-0.912	0.362	-2.56e+05	9.34e+04
grade_4	-2.992e+05	2.59e+04	-11.535	0.000	-3.5e+05	-2.48e+05
grade_5	-3.159e+05	1.1e+04	-28.616	0.000	-3.38e+05	-2.94e+05
grade_6	-2.006e+05	6724.100	-29.838	0.000	-2.14e+05	-1.87e+05
grade_8	3.565e+05	8902.655	40.046	0.000	3.39e+05	3.74e+05
grade_9	8.616e+05	1.57e+04	54.849	0.000	8.31e+05	8.92e+05
grade_10	1.293e+06	3.38e+04	38.234	0.000	1.23e+06	1.36e+06
grade_11	1.262e+06	8.1e+04	15.590	0.000	1.1e+06	1.42e+06
grade_12	1.704e+06	4.16e+05	4.093	0.000	8.88e+05	2.52e+06
condition_0	-1.319e+05	6.53e+04	-2.021	0.043	-2.6e+05	-3992.735
condition_1	-1.023e+04	3.02e+04	-0.338	0.735	-6.95e+04	4.9e+04
condition_3	2.275e+04	6245.103	3.642	0.000	1.05e+04	3.5e+04
condition_4	6.813e+04	8699.732	7.831	0.000	5.11e+04	8.52e+04

Omnibus:	10622.554	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	90383.905
Skew:	1.618	Prob(JB):	0.00
Kurtosis:	11.246	Cond. No.	4.94e+05

Data Understanding

Results of model: This model explains about 53.3% of the variance in our data This models F-statistic is statistically significant compared to our alpha of 0.05 Most of the coefficients are statistically significant when compared to our alpha of .05

Interpretations: For a house with no garage, of average grade and condition, and with no renovations we would expect the house to about 70k less than a home that is renovated. We expect that same house to sell for about 4k more with a 1-car garage For each additional 1 square foot in living space size and all other features remaining the same, we would expect the house to gain about \$263

Conclusion

According to our models, renovating a home, specifically targeting square footage of living space, will have a significant increase on the value of the home. While it looks like adding a garage can increase the value of a home in some scenarios, it is dependent on the size of garage and other factors. We would be hard pressed to recommend adding a 1-car garage to a home that does not have one, given that the price difference is only about 4k increase.

To that end, based on our models and analysis, we recommend that if renovations are going to occur, its best to target square footage of living space, but if renovations are going to include the garage, it is probably worth focusing on homes that have no garage and adding a 1-car sized garage, rather than increasing the size of an existing garage.

Next Steps

With more time, we could include other factors, like zip code, that might increase the accuracy of our model and thus update our recommendations.

Questions?

For a full analysis please check the Jupyter Notebook or slide presentation. Further questions? Contact Jordan Loewen-Colón @ jbloewen@syr.edu

Repository Structure

- |— data : data used for modeling
- |— images : images used in PPT and README
- |— Microsoft Movie Studio.ipynb : notebook used to pull from API
- |— README.md : project information and repository structure
- |— presentation.pdf : the powerpoint presentation used to present data analysis

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

- Jupyter Notebook 100.0%