Regression Modeling on House Prices in Miami

Completed By: Quinn He

University of Massachusetts at Amherst

Introduction

Housing prices are always difficult to predict and can fluctuate randomly due to many variables or economic events. At the core, location is an extremely important factor in dictating the price of houses, and always will be. The purpose of this project is to implement multiple regression techniques in determining how important home distance to the ocean is for sale price of single-family houses in Miami. I am particularly interested in Miami due to its proximity to the ocean and because Florida is a flat, low elevation state with a lot of beach front property, but susceptible to hurricanes. This research lays the groundwork for potential research in housing prices related to climate change.

Data set was pulled from Kaggle and includes information on 13,932 single family homes sold in Miami, Florida. Variables include distance to ocean, sale price, distance to subway, structure quality, home age, distance to rail station, floor square footage, land square footage, and special home features

Research Question & Hypothesis

- Do homes closer to the ocean have a higher sale price than homes that are farther away?

- Alternative: Homes closer to ocean will have a higher sale price
- Null: Distance to ocean has no impact on sale price

Model Comparison

- Model 1 contains distance to ocean as the explanatory variable
- Model 2 has both distance to ocean and floor square footage
- Model 3 contains every variable in the data set

I use floor square footage of the home as a control variable because it's an important variable for any type of home, can be referenced easily in other research, and is a constant variable for every home.

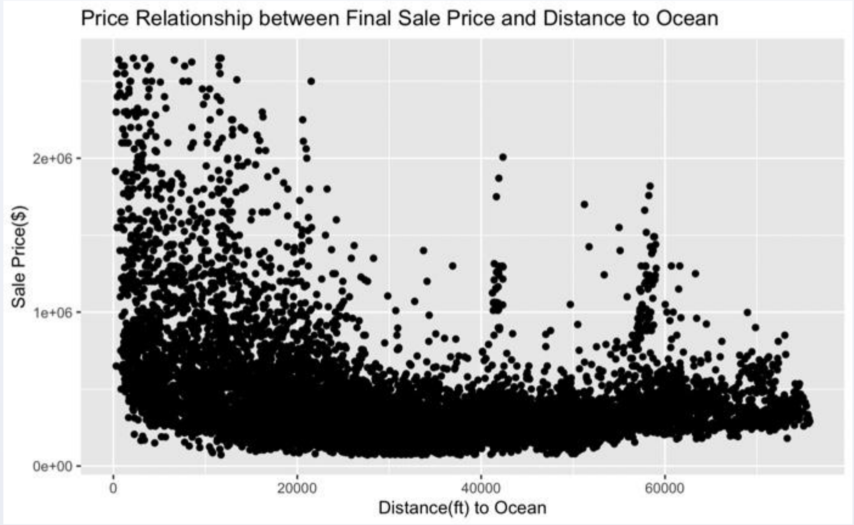
- Model 2 fits the best with an adjusted R squared of 0.5 and AIC of 382,748
- Model 1 is a poor fit with adjusted R squared of 0.07 and AIC of 391,412
- Model 3 has an adjusted R squared of 0.71 and an AIC of 375,102.

Model 2 is chosen as the best fit because it includes the primary variable of interest and floor sq ft with a high adjusted R squared and one of the lower AICs, indicating good fit.

	Model 1	Model 2	Model 3
floor_sqfoot		255.466*** (2.331)	190.761*** (2.437)
special_features			2.974*** (0.125)
RAIL_DIST			4.929*** (0.275)
Constant	556,875.800*** (5,323.914)	12,346.350* (6,316.385)	125,032,873.000*** (10,202,394.000)
Observations	13,932	13,932	13,932
R2	0.075	0.504	0.714
Adjusted R2	0.075	0.504	0.714
Residual Std. Error	305,024.700 (df = 13930)	223,506.700 (df = 13929)	169,785.000 (df = 13915)
F Statistic	1,136.729*** (df = 1; 13930)	7,066.145*** (df = 2; 13929)	2,169.588*** (df = 16; 13915)
Note:			*p<0.1; **p<0.05; ***p<0.01

Model 1 Model 2 Model 3

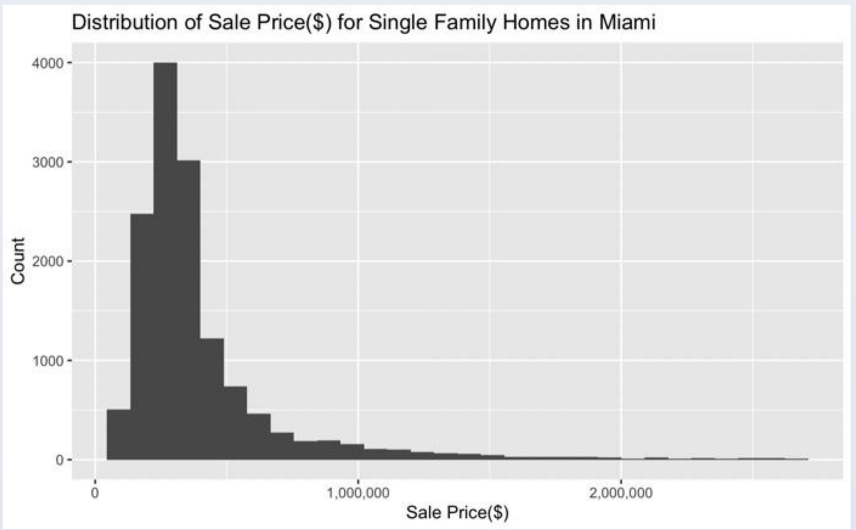
Results



- Graph indicates homes that are closer to the ocean tend to have a higher sale price than homes that are farther away
- At the 40,000 and 60,000ft mark there is a general spike in house prices
- Graph follows an "L" curve, values with a lower X have a higher Y

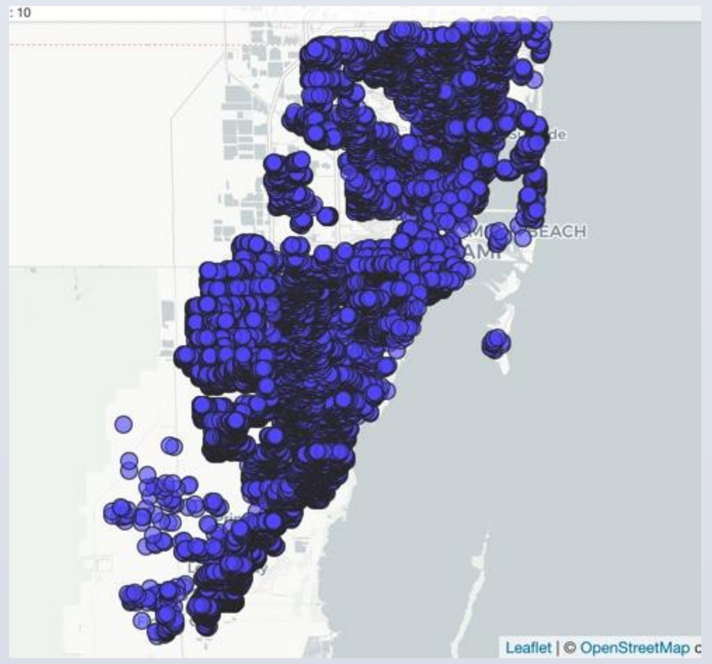
This visualization depicts the distribution of homes based on their sale price

- Most of the prices of homes fall slightly below the $500,000 mark at around $400,000
- A few outlier homes are well past the 1 million and 2 million dollar mark

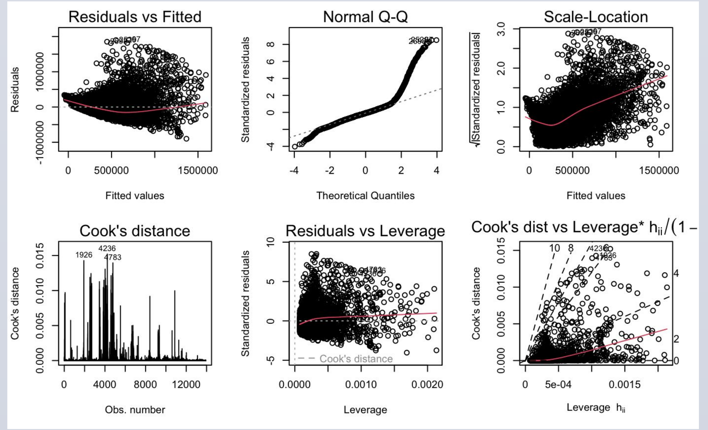


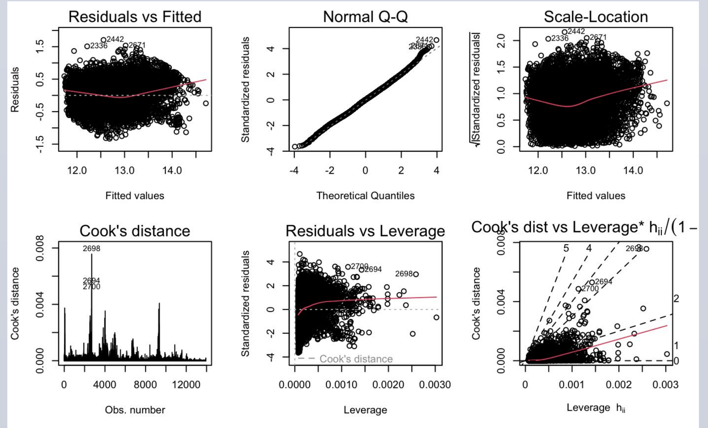
The map view is very cluttered, but it is an interactive map that allows me to click on the data point and look at all the information associated with that house

- Data points in Key Biscayne range from about $1.6 to $2.6 million
- Most of the homes are only ~2000 feet away from the ocean
- The case is the same for houses in Miami Beach, Surf Side, and Sunny Isles Beach.
- Clicking on homes farther from the ocean are generally cheaper



Diagnostics





This diagnostic plot is based off the final model I chose (model 2) which contains OCEAN_DIST and floor_sqfoot as the explanatory variables and sale_price as the outcome.

- The residuals vs fitted plot indicates some heteroskedasticity with the fanning out of the residuals as the values increase.
- Normal QQ indicates the data is a bit skewed and shows residuals are not normally distributed
- Scale-Location violates assumption of linearity. Residuals are not spread equally among predictors
- There are no extreme outliers that violate Cook's distance
- Residuals v. Leverage does not show any significant outliers

Transformation: I took the logarithm of all variables in the regression equation and it normalized all the diagnostic plots, turning them into desirable outcomes

Analysis

- This research is useful for anyone who wishes to study climate change's impact on Miami, a major coastal city
- It lays the groundwork for establishing the impact of ocean distance and sale price of homes for a city that is continually bombarded by hurricanes and rising sea level
- The adjusted R squared of the regression model with solely distance to ocean as the explanatory is only 0.07, I did not think that was significant enough so I chose Model 2 w/ floor sq ft as a control due to its better fit
- I am able to see ocean distance is significant based on its p value.
- Separate visualizations and graphs support that a lower ocean distance value results in a higher sale price

I created a table to see how distance to ocean and price are related to other variables like home age and structure quality.

- Homes with the lowest quality of structure are on average the oldest and cheapest, opposite is true for homes with better quality
- Structure 3 is interesting because homes are the youngest and also significantly more expensive
- Structure 3 homes are also the closest to the ocean, which may indicate a substantial impact on price.

structure_quality <dbl>	mean(home_age) <dbl>	mean(sale_price) <dbl>	mean(OCEAN_DIST) <dbl>
1	66.10056	162639.7	23426.392
2	27.56326	269672.3	25563.743
3	15.06250	1847250.0	7103.031
4	32.03777	382571.0	35087.479
5	28.79021	743189.3	32269.203

Previous Research

- Previous research uses OLS regression to predict prices taking into account the correlation between prices of neighboring homes
- Uses kriging because it allows spatial correlations to be incorporated into predictions
- Another study uses quantile regression accounting for spatial autocorrelation
- A study in Australia examines housing price change in Australia from 1970 to 2003 using a long run equilibrium model

Conclusion

I am able to conclude homes that are closer to the ocean have a higher sale price than homes farther away. The selected model has an adjusted R squared of 0.5. Multiple analyses show distance to ocean is an important factor in price. Even though a model with only distance to ocean has an adjusted R squared of 0.07, its p value is highly significant and multiple visualizations show homes with a lower distance to ocean value have a higher sale price.

For future research, I would like to track house prices over time as they relate to climate change and rising sea levels. Do coastal counties that vote republican tend to have or maintain high sale prices compared to coastal counties that vote democrat? With an increase hurricanes that hit Florida, do people still want to buy homes there and is the housing market sustainable?

Future research should implement multiple regression techniques on other coastal cities not only in the United States to see how distance to ocean impacts prices in other countries. Similar research on states within the U.S. could also show how Florida's housing economy compares to other states.

References

Clapham, D. (2017) Housing theory, housing research and housing policy. Housing, Theory and Society, 35(2), pp. 163-177. (doi: 10.1080/14036096.2017.1366937)

Dubin, R.A. Predicting House Prices Using Multiple Listings Data. *The Journal of Real Estate Finance and Economics* 17, 35–59 (1998). https://doi.org/10.1023/A:1007751112669

Explaining house prices in Australia: 1970–2003 - Wiley Online Library. (n.d.). Retrieved December 11, 2022, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2005.00243.x

Zietz, J., Zietz, E.N. & Sirmans, G.S. Determinants of House Prices: A Quantile Regression Approach. J Real Estate Finance Econ 37, 317–333 (2008). https://doi.org/10.1007/s11146-007-9053-7