Austin Lewis

Fall 2017

BQA 4000/7000

12/01/2017

Golden Triangle Residential Housing Market Analysis

Introduction

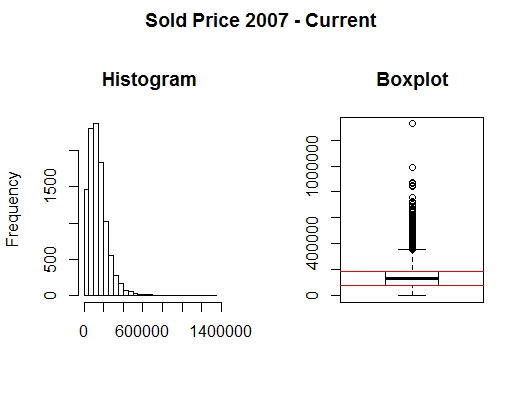
This is an analysis of the “Golden -Triangle” residential housing market. The “Golden -Triangle” refers to a geographic region in eastern Mississippi formed by the cities of Starkville, West Point and Columbus and encompasses their respective counties. Although the majority of this analysis will focus on homes located in the three point cities, I do intend to briefly look at the sold price of all homes in this dataset. My goal in conducting this analysis is to be able to leverage data about homes in this market in a way that could be useful in business application, examples of this could be creating a linear model to aid in more accurate pricing or predicting the length of time a home stays on the market.

Having access to quality and consistent data was paramount for this analysis and I was fortunate enough to obtain it through accessing the Golden Triangle Multiple Listing Service. Multiple listing services are designed for use by professionals in the real estate industry, they are managed by a board of realtors and are one of the most accurate sources for information about homes within their market. Realtors must collect and communicate a significant amount of data about homes that they list for sale and the data they contribute to MLS databases is useful for a variety of business purposes ranging from sales to appraisal. Each observation in the dataset provides quantitative and qualitative data about a home’s features, location, and sales status.

I have analyzed a housing dataset previously for a Business Statistics II course, and it proved to be a very interesting and challenging project. While the scope and mission of that project were very simple by comparison, it did allow me to experiment with this type of data and explore the relationships between basic features such as square footage of a home and the number of bedrooms. Through this analysis I will move beyond simply identifying relationships and correlations to developing useful predictive models. Additionally, I hope to create a couple of visualizations that may give an idea of the markets status. I believe applying statistical methods and techniques to this data will yield valuable insights about the trends of the housing market and lead to useful models. I think it will be interesting to see how Starkville compares to the other Golden Triangle housing markets, I do expect to find that Starkville’s housing prices are relatively inflated.

Data Exploration And Visualization

This dataset contains 36,867 observations of 74 variables, these variables include information about the homes price history, size, day on market, location, style, and listing agency. This dataset has a large amount of unstructured data that qualitatively describes some of the homes features. Unfortunately, while it would be great to use this information it’s preparation would delay this analysis beyond its time period and also, I feel this data would be best implemented after initial testing such as I will conduct in this project. The date range stretches over the past 24 years with the first observation being recorded in August 1993. I understand that inflation does play a role in housing prices and as such I will decrease the time frame of my sample to focus on the previous 10 years. This transformation will provide a better picture of the current market. Additionally, for my analysis I am only going to focus on homes that have been sold. The reason behind this decision is that two variables I may predict are sales price and days on market, both of which require that a home has been sold. Once I filtered my data down I wanted to look at the distribution of the sold price variable.



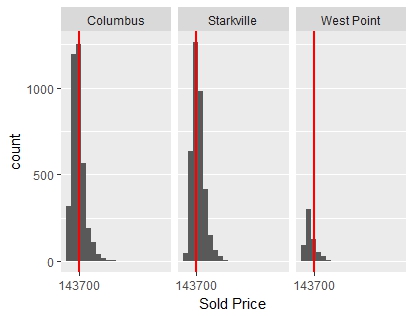
Min. 1st Qu. Median Mean 3rd Qu. Max.

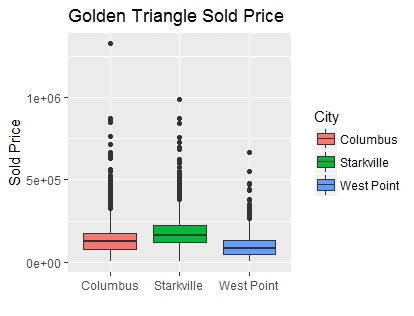
3000 77500 129900 143700 189900 1325000

As you can tell from the histogram the data appears to be unimodal and is slightly skewed to the right, as confirmed by the boxplot it appears that a majority of homes sold fall in a relatively narrow range even with a significant number of outliers outside the upper bound. As evidenced by the five-number summary above the majority of homes sold fall between $77,500 and $189,900, some as low as $3,000 and as high as $1,325,000. It would seem that homes located within this price region are most popular with consumers, although there is quite a demand for homes priced above the $200,000 they aren’t sold as frequently. While I don’t believe this is a ground-breaking inference this distribution and it’s outliers will be important factors to consider when I begin modeling my data.

While this dataset focuses on homes located in the three Golden Triangle cities it also contains homes located within smaller towns in their counties, I intend on trimming out all homes that are not located in one of the three main cities.

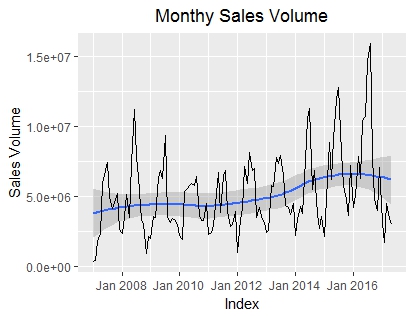
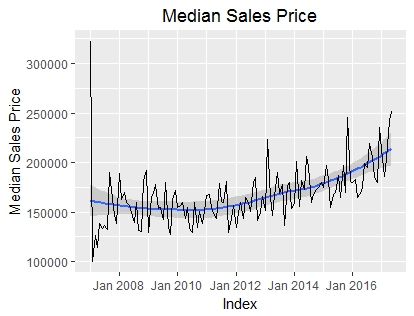
Now that we have seen the distribution of sold prices for the overall market I would like to see how the housing markets for three “Golden Triangle” cities compare to one another. As stated in the introduction my hypothesis is that Starkville’s prices will be relatively higher, I based this hypothesis on the fact that Mississippi State University drives demand for housing higher and in turn will inflate prices.





In the figure above, you will notice that all three cities have similarly shaped distributions, each has a fairly obvious mean and a skew slightly to the right of that mean. Starkville and Columbus’ markets are more comparable from a demand perspective than West Point, this is to be expected as their populations differ from around 15,000 people. The red line visible in each chart denotes the median home price for the overall market, while most of the homes sold in Columbus and Starkville are around this median it is evident that Starkville does indeed have more sold homes above this price. Conversely, it appears this median does not accurately represent the average selling price of homes in West Point. The boxplots will confirm my hypothesis and show that Starkville’s housing prices are the highest on average in the region, it’s not a drastic difference in prices but it is noticeable. It is pretty common sense to assume that the city your home is located in factors into its price, but I did feel this relationship was worth visualizing. I created some additional boxplots with the remaining numeric columns in my dataset and it appears that these three markets are similar in many other aspects.

Because this dataset contained date information I was able to aggregate data and plot it as a time series.



From the time series plots above, it would seem that there is some seasonality throughout each of the variables with sales volume appearing to be the most seasonal. Median sales price tends to be trending upward since around 2012. I intend on fitting a forecasting model to one of these variables in the next section.

Proposed Techniques

I wanted to begin by attempting to create a linear model predicting a homes’ sold price. While I understand a homes’ price is heavily dependent on attributes such as its neighborhood and condition that I am unable to fully account for I was interested to see if I could get decent range estimate based on the variables I do have. I started by implementing a simple linear model as a foundation. I have cut down my dataset to include only 28 variables which are listed below. The variables nrooms was created because the original total room variable contained a large amount of N/A’s and really only was completed for homes with an unusually large number of rooms. To remedy this, for observations that contained N/A I simply added the total number of bedrooms, bathrooms and half baths. Similarly many homes didn’t have information about their acreage so I replaced N/A with zero for the newacre column.

[1] "solddate" "originallist" "listprice" "soldprice" "area"

[6] "city" "totsqft" "yearbuilt" "style" "totbedroom"

[11] "totbathrooms" "tothalfbath" "garagetype" "taxes" "taxyear"

[16] "homestead" "vacant" "lockbox" "sign" "stories"

[21] "Schools" "floodzone" "watefront" "foreclosure" "dom"

[26] "nrooms" "newacre"

In my initial model I plan on using all variables except for the original and current list price as well as the taxes on the home. I will split my data 80/20 for training and testing respectively. As you will see by the results below my training model has a decent Mulitple R-Squared of 0.7347 and the large F-statistic indicates a strong relationship between our response and predictor variables. The residual standard error shows that this model is off by around $50,000 which isn’t great but if you notice the residual summary you will see that this model can be reasonably accurate although prone to serious error. **Training Model**

Residual standard error: 49160 on 6305 degrees of freedom

Multiple R-squared: 0.7347, Adjusted R-squared: 0.7325

F-statistic: 329.4 on 53 and 6305 DF, p-value: < 2.2e-16

Residuals:

Min 1Q Median 3Q Max

-346877 -25645 -554 23710 891258

**Test Results**

Mean Absolute Error (MAE ) = 33417.00

Root Mean Squared Error (RMSE) =47919.35

Mean Absolute Percentage Error = 0.4258538

Summary of Residuals

Min. 1st Qu. Median Mean 3rd Qu. Max.

-527500 -22640 2324 3466 27980 383900

Judging by this models Mean Absolute Error it tends to be off by around $35,000 total on average, the RMSE will show that this average is significantly higher when adding more weight to the extremely wrong predictions of which this model has the tendency to make. As noted in the visualization portion of this analysis sold price is subject to outliers and these seem to be heavily influencing the results of this model as for the most part it can be fairly accurate for the information it is given. For my next model I will transform my soldprice variable with a Box-Cox Transformation and implement a Backwards Step-Wise Regression to aid in optimization. I chose to use a Box-Cox Transformation because all my y-values(Sold Price) were positive. The results of my Box-Cox transformation are below.

Box-Cox Transformation

Input data summary:

Min. 1st Qu. Median Mean 3rd Qu. Max.

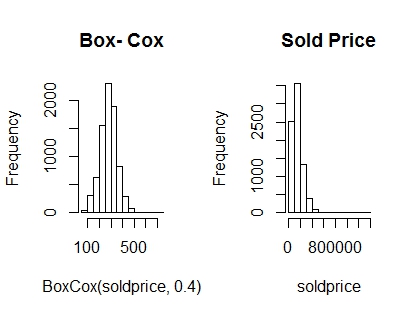
3000 88000 138500 154000 198000 1325000

Largest/Smallest: 442

Sample Skewness: 1.78

Estimated Lambda: 0.4

You can see the impact this transformation had on the distribution of the sold price variable. Hopefully this normalization coupled with the added optimization of the stepwise model will produce better results.



The stepwise regression method determined that our model was best fit with 4 variables which are acres, total half-baths, sign, and tax year. This a random assortment of variables but again this model is still very much an experiment at this point and its potential for use yet to be determined. Our model’s accuracy statistics are displayed below.

**Test Results**

Mean Absolute Error = 55.95988

Root Mean Squared Error = 71.47255

Mean Absolute Percentage Error = 0.2412739

Residual Summary

Min. 1st Qu. Median Mean 3rd Qu. Max.

-214.1000 -45.5700 4.3650 0.2976 47.5400 323.4000

You will see that this model did not fare much better than the original, while the values are altered by the power transformation of the Box-Cox its accuracy is essentially the same. The only indicator that has been improved in this model is the Mean Absolute Percentage Error which is about half of that of original model indicating that the accuracy actually has been improved slightly, you will notice in the residual summary that this model’s minimum residual error has been improved with regards to the original.

After looking at the results of these models I do not believe that they will be sufficient to estimate a useful range of the price a home will sell for. There are many variables that influence home price and while this model can give a very large ballpark estimate I don’t believe it is accurate enough to useful.

As mentioned in the previous section I wanted to briefly attempt to fit a few timeseries forecasting models to predict the future median sales price of homes in the market. I created a timeseries of the sold price, aggregated it to obtain monthly medians, and then split the data for cross-validation. To split the data I simply cut off my timeseries with 4 months remaining to use as my test set. I then trained and tested a variety of forecasting models of which you will see the results of below.

Seasonal Naïve Method

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 3151.264 18992.31 15121.28 1.337602 9.983550 1.000000 -0.08337547 NA

Test set 6174.446 19420.86 14968.57 2.880603 8.621891 0.989901 0.37789652 0.861863

|  |  |
| --- | --- |
| |  | | --- | |  | |

Average Method

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set -7.462398e-12 16749.45 12990.25 -1.335783 8.925739 0.8590704 0.2524494 NA

|  |
| --- |
| Test set 2.274444e+04 25499.57 22744.44 12.722703 12.722703 1.5041345 0.1054060 1.309615 |
|  |
| |  | | --- | |  | |

Holt-Winters Seasonal Multiplicative Method

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set -1.067698 13809.933 10602.103 -0.8864136 7.310119 0.7011378 -0.10861902 NA

|  |
| --- |
| Test set 1311.493729 9048.354 7634.194 0.4083466 4.554735 0.5048641 -0.05886097 0.565225 |
|  |
| |  | | --- | |  | |

Exponential Smoothing w/ Box-Cox

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 2896.123 15136.60 12193.69 0.8766868 8.259213 0.8063923 -0.04738255 NA

Test set 5796.230 12904.06 11756.29 2.8901747 6.798749 0.7774667 0.10540597 0.6411749

ARIMA

ME RMSE MAE MPE MAPE MASE ACF1 Theil's U

Training set 1786.969 14727.28 11774.98 0.1904901 8.092225 0.7787023 -0.02523192 NA

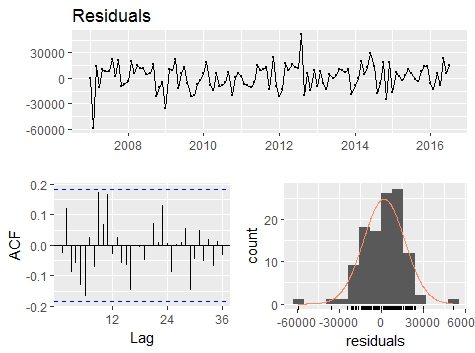
|  |
| --- |
| Test set 5307.414 13039.92 10312.10 2.6212307 5.902618 0.6819597 0.35989771 0.5688422 |

The Holt-Winters model seems to have performed this best with the lowest MAE and RMSE. This model is great at capturing seasonality which I assume resulted in its low error.

The ARIMA model seems to have been the second best fitting model, I assumed that it would be the

most optimal as ARIMA’s tend to be one of the more powerful forecasting techniques. The residuals of

this model are displayed below.



Ljung-Box test

data: residuals

Q\* = 25.224, df = 20, p-value = 0.193

Model df: 4. Total lags used: 24

It seems that this ARIMA ordering was sufficient and the residuals are behaving as white noise. The

Ljung-Box test will confirm that the correlations are small enough that we can rule out significant

autocorrelation between observations.

Conclusion

In conclusion, this analysis has yielded some useful and interesting insights into the Golden Triangle residential housing market. Although my linear model to predict sales price wasn’t as accurate as I had initially hoped I do think that it has the potential to be improved significantly if some of the unstructured data about a homes’ additional features could be prepared and implemented into the model. I was pleased with the results of my forecasting models, I believe that using seasonal forecasting methods on more variables in this dataset such as days on market or sales volume could provide useful indicators to compare against actual movement. I feel that my conditioned boxplots and histograms of the three cities provided a straightforward way to understand how their markets compare to one another, these visualizations could be utilized by investors or developers not familiar with the region. It is important to understand that this was a primary analysis essentially skimming the surface of this dataset. I believe that continued preparation, visualization, and modeling of different variables could yield promising results. One variable I’d be interested in utilizing would be the selling agency column, I think it would be interesting to visualize how agencies have performed relative to each other in terms of average sales volume, price, and days on market. There are many directions you can go when analyzing this dataset depending on what your goal is, from economic insights to agency performance this data creates a foundation for exploration. As with many projects in data science you are only limited by your imagination when deciding what to experiment on next!