MQM Stats Final Project

Serena Rong, Ruben Valentin, Eric Weng, Siyu Zhu

Business Understanding

Our team is interested in understanding the performance of residential real estate market in the United States. According to Realtor.com, the median number of days property listing spend on the market is one of the most important market metrics used to assess the real estate market within a specific geographical region[1]. Days on market, defined by Realtor.com, is the time spent between the initial listing of a property and either its closing date or date it is taken off the market. When a specific geographical region has a median Days on Market lower than national average, it generally means the houses are selling fast in that region, thus indicating a popular market. Based on this information, our team decided to assess the market performance in terms of the days on the market.

Predicting the Days on Market in specific geographical regions provides business value for real estate firms to use existing market data to understand the performance of the real estate market in certain locations, make informed investment decisions, and modify their portfolios accordingly. While we recognize that certain external factors such as supply and demand would also impact the market, we developed our model using only internal market attributes that most real estate firms can obtain without performing costly market research.

Data Understanding

The specific dataset we have chosen is the inventory data for all residential listings in April 2019 by zip code. The data source is the 'Realtor.com residential listings database' provided by News Corp., which owns Realtor.com, the listing website of the National Association of Realtors. With data from nearly 800 Multiple Listing Services (MLS) across the U.S., Realtor.com's listings database is the most comprehensive and up-to-date source of active inventory.

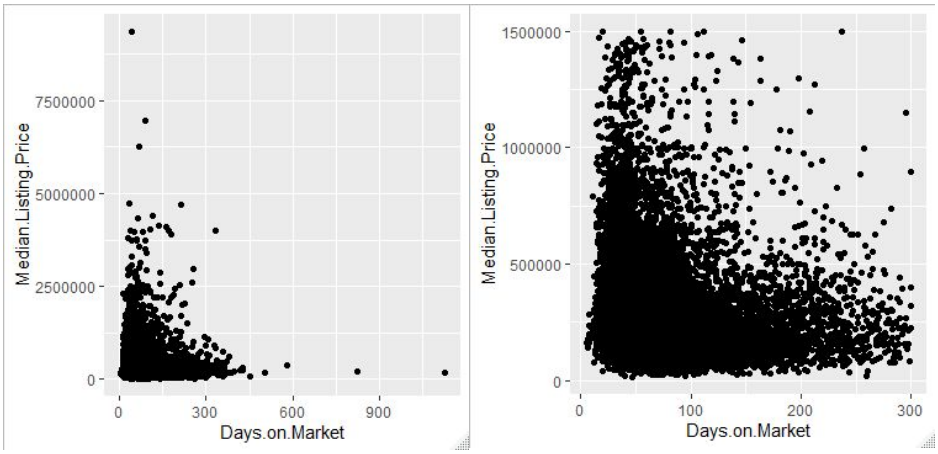
The dataset contains 15,030 rows and 34 columns, and describes key attributes for each real estate market by zip code in the United States. All of the variables are numerical, except for the geographical indicators such as zip codes and zip names. We selected our target variable to be the Days on Market, which represents the median number of days property listings spend on the market within the specified geography. Some of the other

[1] Realtor.com National Market Outlook, retrieved from https://www.realtor.com/research/market-outlook/.

variables include listing price, active listing count, new listing count, price increase count, pending listing count, and many others.

Data Preparation

As the first step to prepare the dataset for analysis, our team focused on eliminating outliers in key market metrics. We decided to eliminate the records that have less representative inputs for Days on Market, total listings, and median listing price. As shown below, the plot on the left shows the original data and its potential outliers. After filtering the records mentioned above, we effectively "zoomed in" on the data as shown on the plot on the right. We first filtered out all Median Listing Prices above $2,500,00 but found that we could have a more focused dataset by filtering above $1,500,000. Specifically, we eliminated records with more than 300 average Days on Market, or more than 1,500 total listings, or a median listing price of more than $1,500,000.

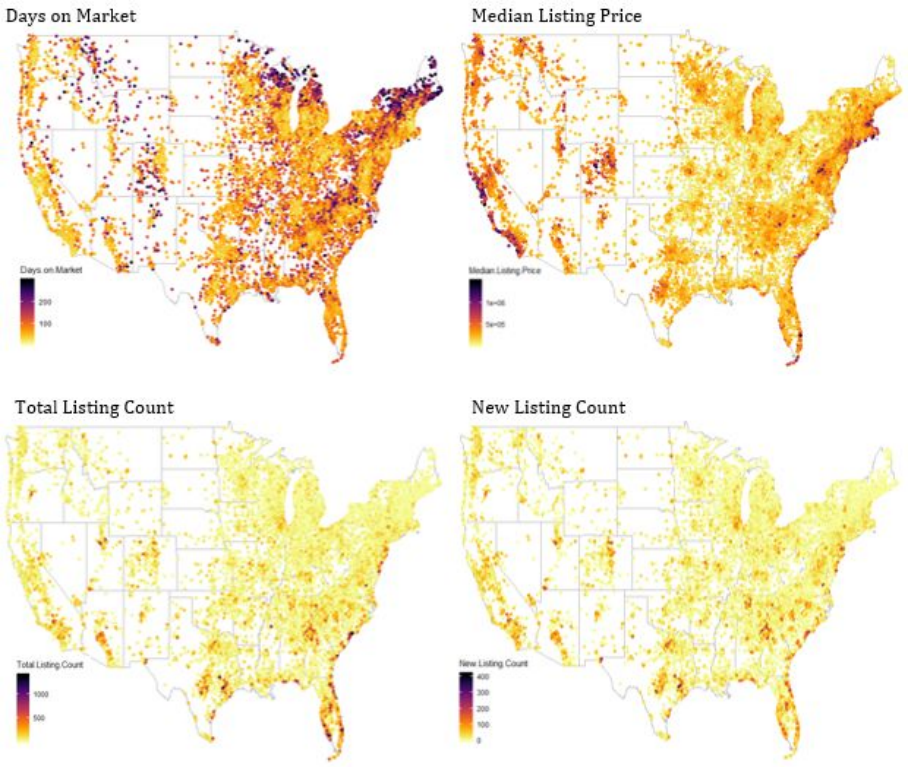


Our second step to prepare the data was transforming the zip names into the geographical regions. This is a necessary step to help generate overall performance of larger geographical areas instead of specific zip codes. To do this, we first extracted the state abbreviation from the zip names column, which consists of city name and state abbreviation of each zip code. Then we used the geographical regions described by the United States Census to assign each record into the corresponding region (Midwest, Northeast, South, and West).

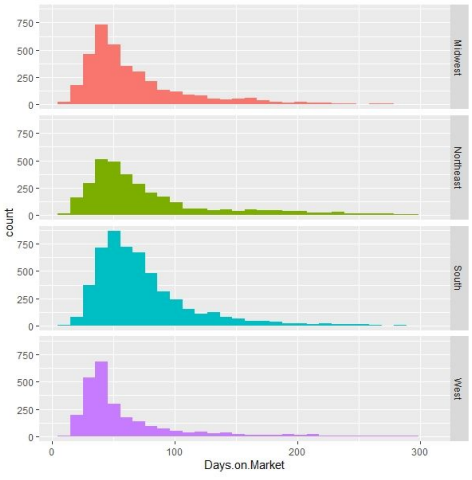
Exploratory Data Analysis

As the first part of our exploratory data analysis, we created heat maps on different variables to get the big picture on how our target variable and some other variables are geographically distributed, utilizing the unique geographical indicators in our dataset. As shown below, there are clear differences in regional distributions of Days on Market and

median listing price. For total listing count and new listing count, the difference in regional distribution is less significant, but still exists to a certain extent.



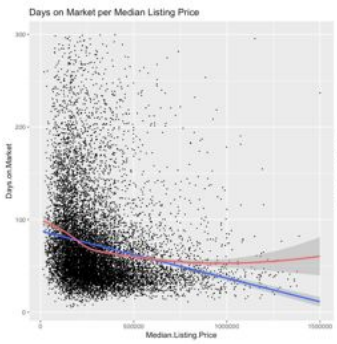
We then developed a histogram of our target variable, Days on Market, by region. As shown to the right, most zip codes in all four major regions have listings on the market for 30 to 50 days, while the South region has more zip codes with listings on the market for 50 days. According to the summary report of the dataset, the mean Days on Market is 72 days. The gap between the mean and the median is relatively huge, and implies the potential impacts from outliers.



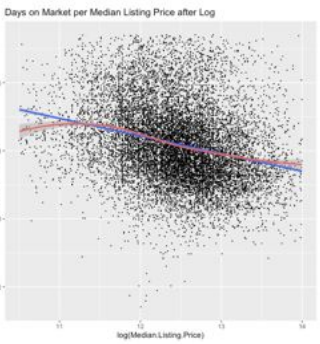
We also noticed that the distributions of Days on Market in all four regions are left-skewed, and confirmed this finding in the histogram of overall Days on Market as shown below. To assist in modeling, we want to reduce the skew and rescale the data to make it more symmetric. We chose to perform a log transformation on the variable, and after the transformation, the Days on Market variable appears to be normally distributed, and the outliers seem to be less extreme. We repeated this process for a few other variables that we found to be heavily

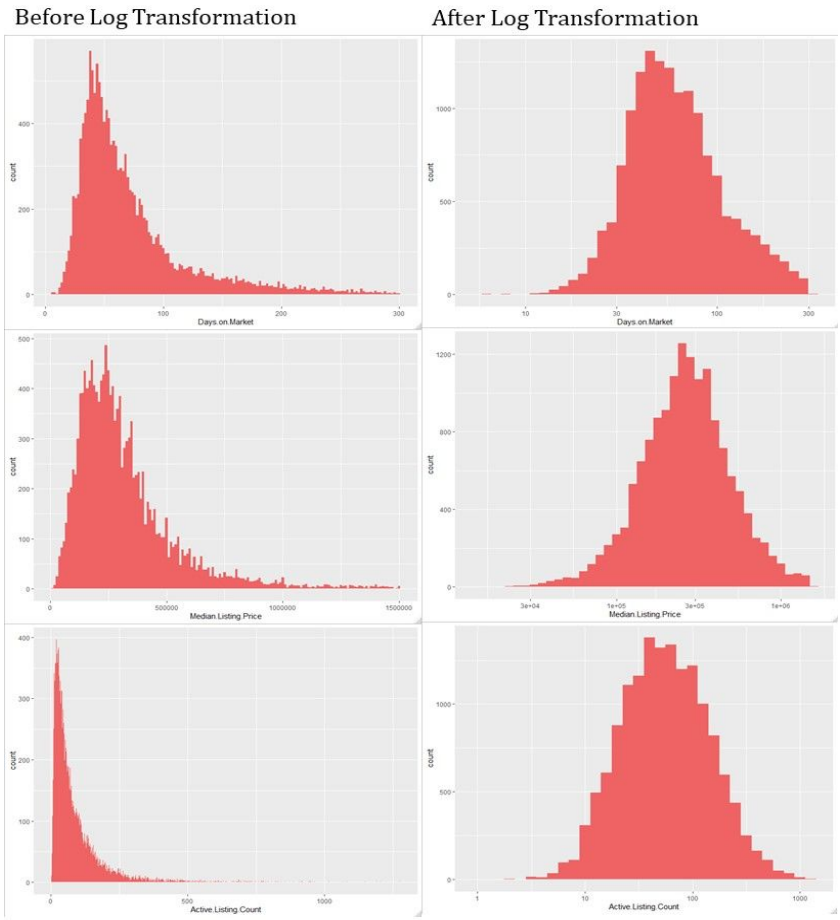
left-skewed as well, and the distribution before and after log transformation are shown below.

Before Log Transformation



After Log Transformation





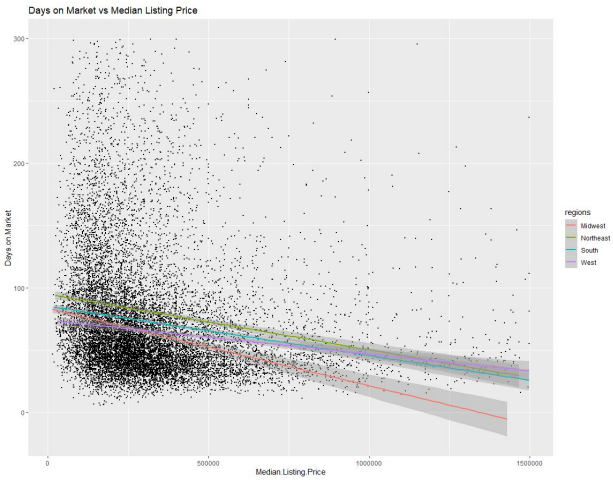
The second part of our exploratory analysis was to investigate the relationship between different variables in relation to the Days on Market. We first plotted median listing price against Days on Market as shown below. In addition to linear regression, we also created a more fitted curve using Loess Regression, a non-parametric method where least squares regression is performed in localized subsets. Before log transformation, the red Loess regression line was further apart from the blue linear regression line as the median listing price got bigger. After the log transformation, the blue loess regression line appeared to be more linear. This further confirmed the necessity of performing log transformation for analyzing the linear relationship between these variables.

Interaction Exploration

We expanded our exploratory analysis with more plots exploring the interaction between different variables that correlates with our target variable.

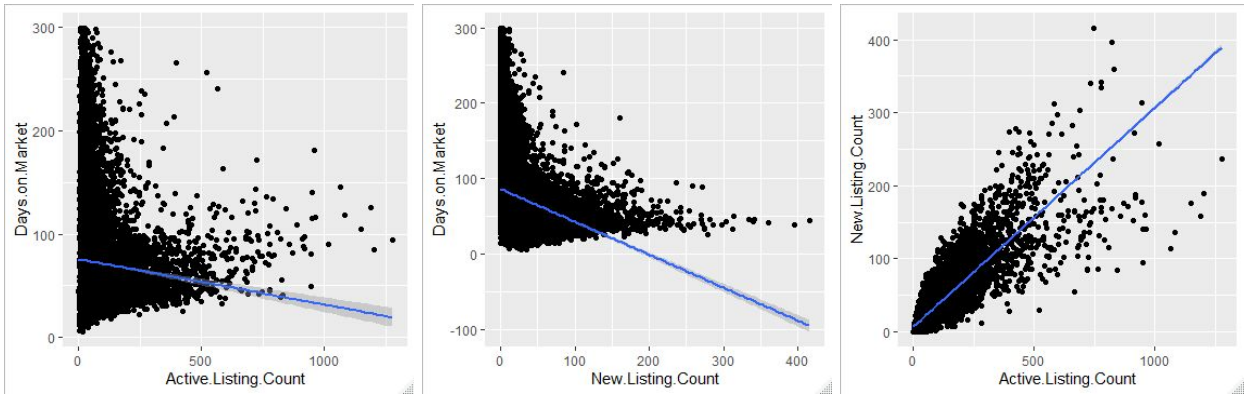
Region Vs. Median Listing Price

The plot to the right shows the interaction between regions and median listing price. While the median listing price is negatively correlated with Days on Market in all four regions, the plot clearly shows that for real estate markets in the midwest region, the median listing price, on average, has a greater effect on Days on Market.



New Listing Count Vs. Active Listing Count

We also explored the interaction between new listing count and active listing count. We first plotted each variable against our target variable, Days on Market, and both plots show a negative correlation between the variables and Days on Market. We then plotted active listing count against new listing count, which shows a positive correlation between these two variables. By analyzing all three plots together, it appears that there is interaction between new listing count and active listing count.



Modeling

Basic Intuitive Model

$$\widehat{Days\ on\ Market} = Regions + Median\ Listing\ Price + Active\ Listing\ Count + New\ Listing\ Count$$

We built our first linear regression model based on four variables that we intuitively selected, which are the region, median listing price, active listing count, and new listing count. According to the regression output below, all four variables are statistically significant in predicting the Days on Market. On average, listings in Northeast region spend more days on the market compared to the Midwest. Surprisingly, this model implies that more active listings in an area would increase the number of days listings are on the market for, and that more new listings, on the other hand, would decrease the number of day listings are on the market. Intuitively both of active listings and new listings should decrease the Days on Market, as it would indicate a market with more demand. However, our base model does not deliver what we expected. To address this, we added the interaction that we found in our exploratory analysis in our log transformed model, to account for the overlap effects of these two variables on predicting Days on Market.

Variable	Coefficient.	P. Value.	sig.
(Intercept)	8.64E+01	<2e-16	***
as.factor(regions)Northeast	9.31E+00	<2e-16	***
as.factor(regions)South	4.53E+00	9.23E-07	***
as.factor(regions)West	6.77E+00	6.97E-09	***
Median.Listing.Price	-2.77E-05	<2e-16	***
Active.Listing.Count	3.03E-01	<2e-16	***
New.Listing.Count	-1.12E+00	<2e-16	***

This model delivers a R-square of 0.2335, meaning the model explains 23.35% of the total sample variation. Because our goal is for this model to predict the Days on Market using only internal factors of the market itself, it makes sense that the model can explain a very limited proportion of the total sample variation.

Log Transformed Intuitive Model

$$\log(Days\ on\ Market) = Regions + \log(Median\ Listing\ Price) + Regions * \log(Median\ Listing\ Price) + log(Active\ Listing\ Count) + New\ Listing\ Count + log(Active\ Listing\ Count) * New\ Listing\ Count$$

After digging deeper in our exploratory analysis, we found that applying Log transformations to certain columns in our model, we could see a better distribution. With this knowledge, we applied Log transformation to Days on Market, Median Listing Price, and Active Listing Count. We also added the interaction between Regions and Median Listing Price, as well as the interaction between Active Listing Count and New Listing Count. Coefficients are shown below:

Variable	Coefficient.	P. Value.	sig.
(Intercept)	4.71E+00	< 2E-16	***
as.factor(regions)Northeast	9.89E-01	6.20E-07	***
as.factor(regions)South	5.25E-03	9.78E-01	
as.factor(regions)West	4.26E-01	0.092	.
log(Median.Listing.Price)	-8.54E-02	4.00E-13	***
log(Active.Listing.Count)	2.83E-01	< 2E-16	***
New.Listing.Count	-7.41E-02	< 2E-16	***
as.factor(regions)South:log(Median.Listing.Price)	0.004154	0.792	
as.factor(regions)West:log(Median.Listing.Price)	-0.026226	0.187	
log(Active.Listing.Count):New.Listing.Count	0.0108297	< 2E-16	***

Compared to our first model, the most notable difference is on the Estimate per regions. On average, listings in Northeast region spend more days on the market compared to the Midwest, but is second to the South. The other coefficients remain acting as before. Differently from our previous model, this model has an R-square of .4865. While still low, it's more than before.

Automatic Backwards Selection

We also performed automatic backwards selection in selecting our variables. Starting with a model with all the variables in our dataset, we took out Price Decrease Count and Total Listing Count because those two variables were statistically insignificant in the result. We then repeated the automatic backwards selection process with the rest of the variables, building the model below:

$$log(Days\ on\ Market) = as.factor(regions) + log(Median\ Listing\ Price) + Active\ Listing\ Count + New\ Listing\ Count + Active\ Listing\ Count * New\ Listing\ Count + Price\ Decrease\ Count + Pending\ Listing\ Count + Avg\ Listing\ Price + Pending\ Ratio$$

This model delivers an R-square of .3816, with ALC lowered to 19931. Coefficients are shown below:

Variable	Coefficient.	P. Value.	sig.
(Intercept)	7.18E+00	< 2e-16	***
as.factor(regions)Northeast	7.38E-02	8.43E-11	***
as.factor(regions)South	1.14E-01	< 2e-16	***
as.factor(regions)West	9.52E-02	6.31E-14	***
log(Median.Listing.Price)	-2.44E-01	< 2e-16	***
Active.Listing.Count	4.19E-03	< 2e-16	***
New.Listing.Count	-1.53E-02	< 2e-16	***
Pending.Listing.Count	4.34E-03	< 2e-16	***
Avg.Listing.Price	2.38E-07	< 2e-16	***
Pending.Ratio	-4.71E-01	< 2e-16	***

This process is helpful in selecting the variables that most significantly explain the data, and provides a logical reference to compare with the selected variables in our intuitive model. This model would make the most accurate prediction of Days on Market for real estate firms that have access to all the market information needed. However, we recognize that some of the variables, such as Pending Listings and Pending Ratio, are not commonly analyzed by real estate firms, and thus not readily available for use. We also hesitate to include Average Listing Price in the model, as it is easily affected by outliers.

Evaluation

Among all the independent variables in the dataset, region plays the biggest role in Days on Market. It was necessary to consider the interaction between variables to help explain the contradicting results for Active Listings and New Listings. The log transformed intuitive model is a more thought out version of the basic intuitive model with additional interactions between factors taken into consideration. The automatic backwards selection provides an accurate model using all possible variables based on the significance.

To compare the three models, we decided to look at their R^2 and AIC:

Model	R^2	AIC
Basic Intuitive Model	0.2335	110308.8
Log Transformed Intuitive Model	0.4865	-26461.3
Automatic Backwards Selection	0.3816	-19931

In terms of R^2, the modified log transformed intuitive model has the highest R^2, meaning it explains most of the total sample variation among all three models. In terms of AIC, the automotive backwards selection model has the smallest AIC because the model is achieved by calculating the significance of each variable. However, it is not ideal compared to log transformed intuitive model because it is very costly to collect all the information needed in this model.

In conclusion, from a business perspective, log transformed intuitive model is the best fit for our dataset and goal, because it uses the least information to get the most possible accurate prediction.

Deployment

Our model provides useful information for real estate firms to understand the performance of the real estate market in certain locations using existing market data. Using the model we built, real estate investment companies will be able to gain insights on how popular the markets in the four regions in the United States are and what are the expected days take to sell those houses. When using our model, real estate companies should be aware that this model does not consider any external factors such as economic performance, or any individual characteristics such as house square foot, condition, year built or the sales agency. It is essential for real estate companies to keep in mind the factors that were neglected by our model, as those could be highly important factors of projecting the time it takes to sell a particular property in reality. One potential drawback of the model we built is that real estate agents might find it difficult to combine our model with other datasets that contain internal housing factors since they do not know the accurate weight for both internal and external factors when predicting the housing market.

Appendix - Contribution by Team Member

Serena Rong:

- Found dataset.
- Compared and evaluated each model
- Summarized deployment of our model on benefits for real estate companies and some potential drawbacks.

Eric Weng:

- Found dataset.
- Tried different plots to analyze patterns(scatter plots, box plots, histograms)
- Visualized regression graphs
- Cleaned data, found outliers
- Found which columns of data should be log in order to provide meaningful insights
- Helped the regression models.

Ruben Valentin:

- Assisted in finding dataset.
- Assisted in cleaning the dataset.
- Assisted in creating plots such as boxplots, histograms, and scatter plots as part of the exploratory analysis.
- Explored a way to be able to group up states into regions and explored ways to visualize our data by using the map of the US.
- Ran Linear Regression Models

Siyu Zhu:

- Defined business problem
- Found dataset
- Compiled and cleaned up R codes
- Composed Report
- Interpreted model results to evaluate models