By: Alfredo Martinez
Date: 2017

The Weather, Home Price Predictions and a Random Forest

The real estate market plays an essential role in our economy, in some way or another at some point we are touch by its behavior. People that mostly care about the trending values of this market are buyers, sellers, renters, investors and many others, and for most of us, when we first consider purchasing a home, we relate its price to the number of rooms, kitchen style, yard space, living room area, etc. a property has. However, we will find out through this dataset, that other factors can influence the negotiations and prices when it comes to purchasing a real estate property.

The goal of this project is to create a regression model that can more accurately predict the price of residential homes given the different homes' features and weather data available to us.

Client

The analysis is for business client Keller Williams Realty, Inc.; the model serves to better estimate the prices of real estate properties through several homes' features and yearly weather. The practical value come as sharper report analyses for agents, investors, business owners and general practitioners of the real estate business, providing stronger negotiations in the sales, purchases and strategic agreements in real estate properties.

Data

Housing

The data collected constitute of the "The Ames Housing Dataset" and was originally compiled by Dean De Cock. This data is publicly available and constitutes of major characteristics, year and month sale price for each home in Ames Iowa between the year 2006-2009 and part of 2010.

Weather

The weather data was obtained from the National Center for Environmental Information website, it includes daily, monthly and yearly records from different weather stations in the city of Ames, Iowa.

Data Cleaning

During data analysis, one of the early steps is cleaning our data. This process can help us avoid working with erroneous data or outliers, which can influence and cover actual or truth values within your model and analysis. Furthermore, our data needs to take a format in which the predictive model can accept the data.

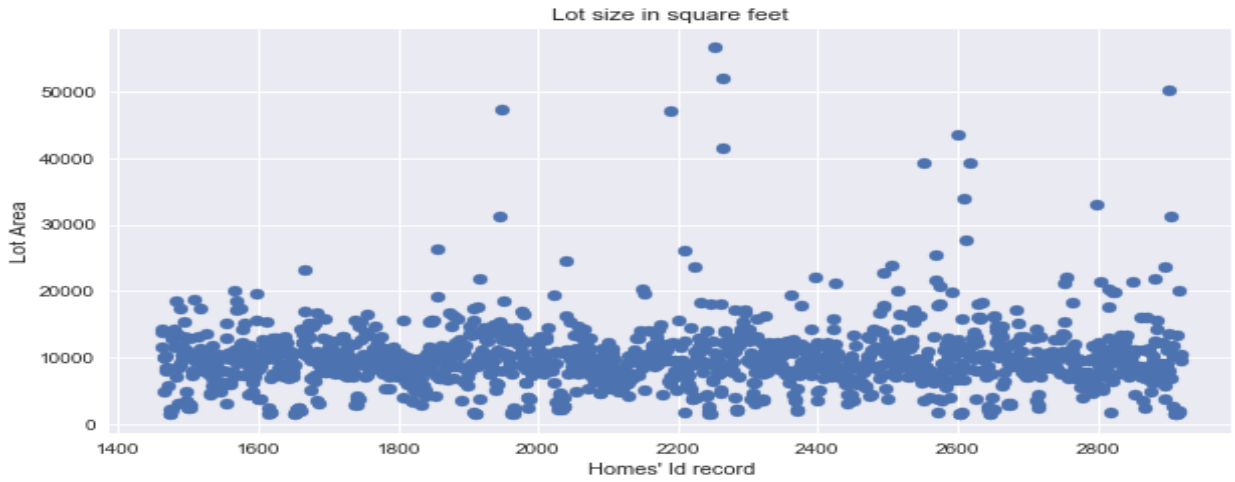
The cleaning process steps for this projects outlined as follow:

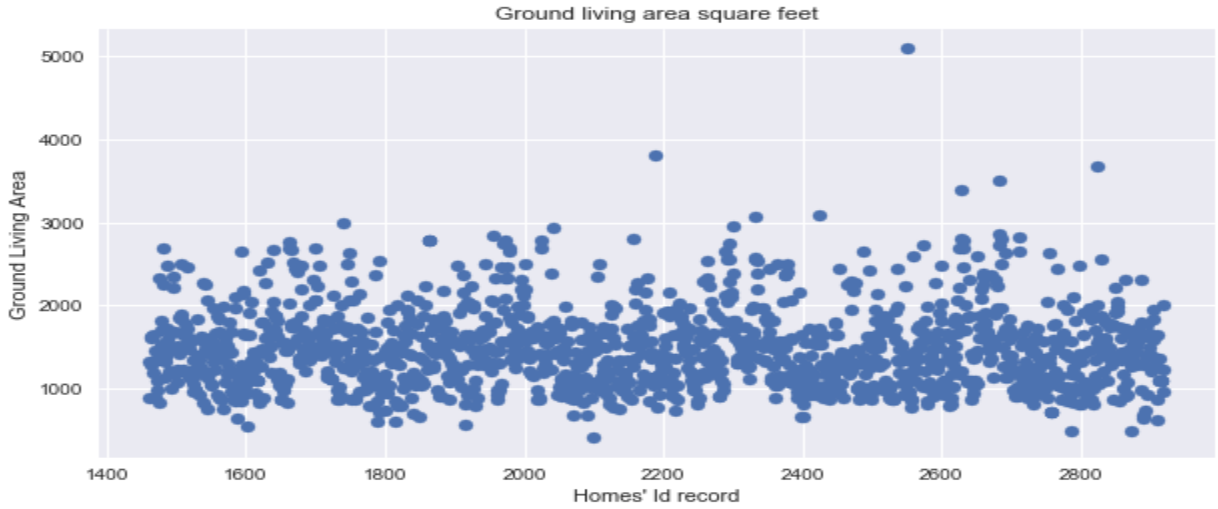
Data Preparation:

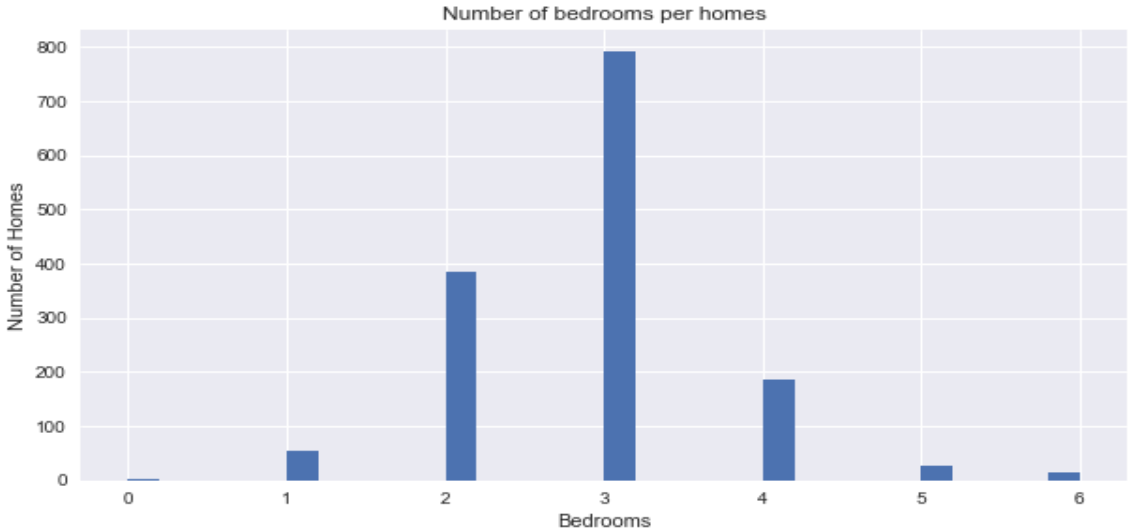
First, we opened libraries and packages needed, then preceded to open data files into Data Frames.

Bad Data and Outliers

This project's implementation is through Random Forest and Decision Trees which are not very sensitive to outliers. However, I still wanted to do a quick Exploratory Data Analysis to see if there was anything out there that appeared to be error type data.







After analysis of a couple key variables, we concluded that nothing was outstanding in regards of erroneous data, so we moved on.

Preprocessing Housing data:

We had two different data files for housing, one consisted of the home's features, and the other contained the prices. We first attached these two data frames to their corresponding values, then filtered cells to dates between 2006 and 2009; I decided on this range by the most available weather dates in my weather data. And lastly, renamed sale dates columns and built a DateTime index for the homes.

Preprocessing Weather Data:

Our weather data files consisted of records from different weather stations.

We filtered and subtracted the weather station which provided the best recordings available. Next, the dates were filtered to match our housing dates, and significant weather variables were subtracted.

Corrupted values showed as a digit (-999), we changed them through numpy library as NaN(not a number values), then we filled them with the average values of their corresponding variables.

After cleaning and building the data frames, these were joint to form a single data frame file.

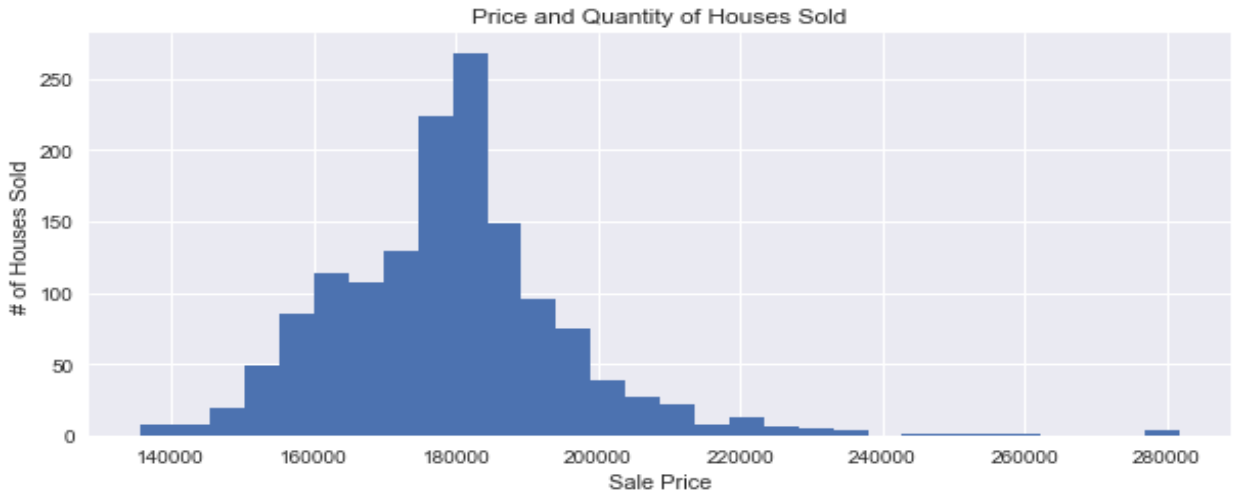
Data Conversion to Fit Model Building:

We split the new data frame into our train and test objects. From this point, we filled our train variables' NaN values with means and modes by either being an integer or floating value. Also, for categorical variables, we built dummy columns since our model works only with numerical points, but not data types such as objects or strings.

For our test data, we preceded to use the mean and mode values from our training data for any NaN value. The reason for this was the fact that we don't know if there will be something different about the new data (test) that can change the basis of our model. And finally, we reindex(transfer) the dummy variables from our training data to capture all the possible categorical columns values into our test set.

Exploratory Data Analysis

The upcoming work and charts show some of the most exciting findings while going throughout exploratory data analysis, it covers data for both housing and weather files.

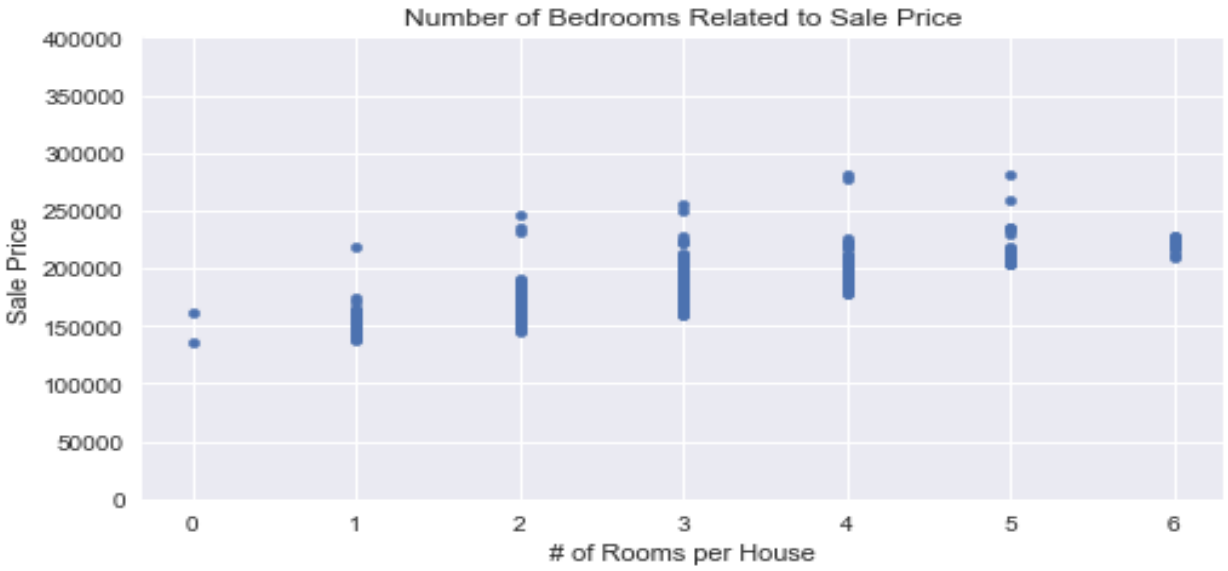


Through this first chart, we can see how the majority of these homes were in a close range price of $170,000 and $190,000. Also, just a few houses had an estimated cost of $240,000 or more. These details give insight into how the cost of living is probably not that high for this particular city.

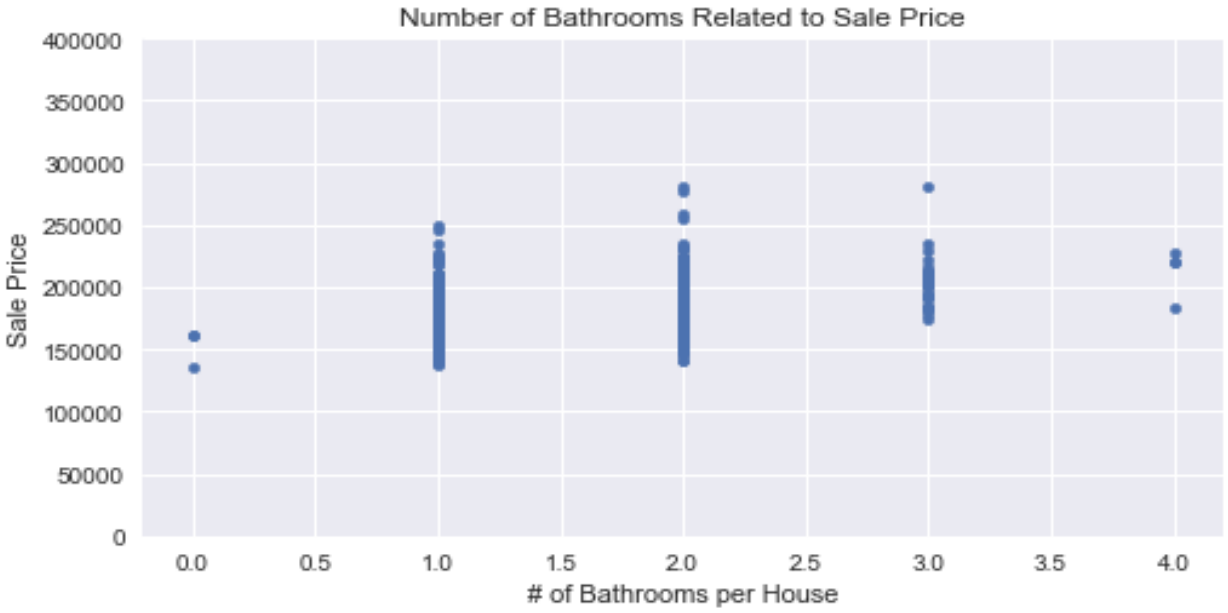
For this second histogram, we see how people make more home purchases during the summer season in May, June, and July and how the sales of homes start declining through November, December, January, and February. The difference in sales between these two seasons is quite outstanding to be twice as much.

- Winter to Summer months have roughly a 150% increase in sales
- Summer(50)−Winter(20)/Winter(20) x 100 =150%

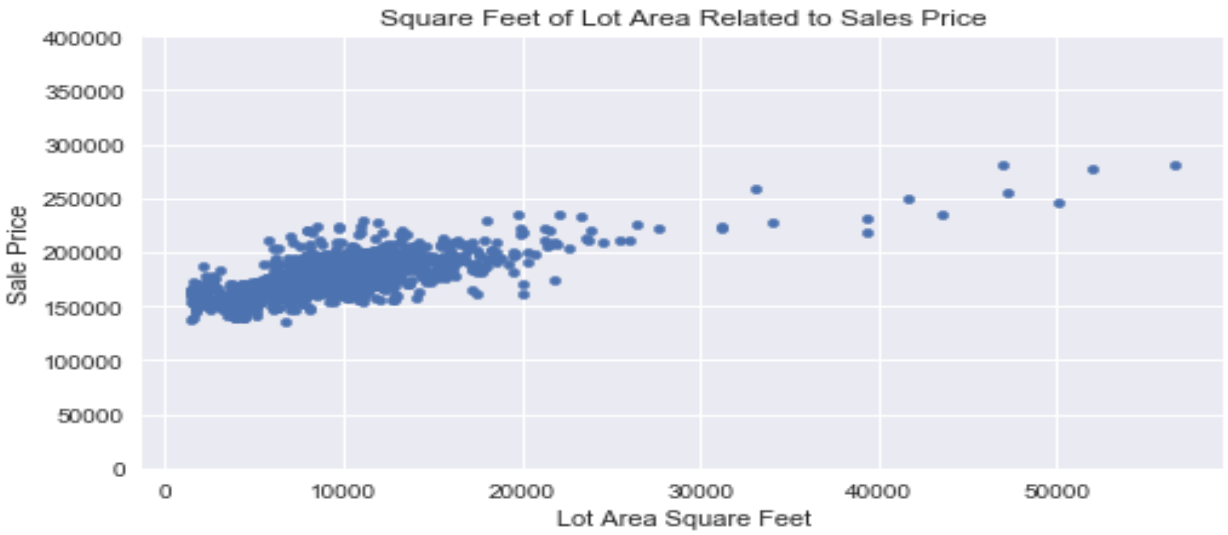
The findings on this chart show a strong distribution throughout the months for every year; there could be different factors causing this placement for the sale of these homes. One factor causing the drastic hiked during summer could be the warmers weather and would be further study in future analysis. Other possible causes for this lifted in sales could be the increase and availability of finish home projects from builders or the tax return funds available to many people just before these months allowing for extra fundings in the purchase of homes.



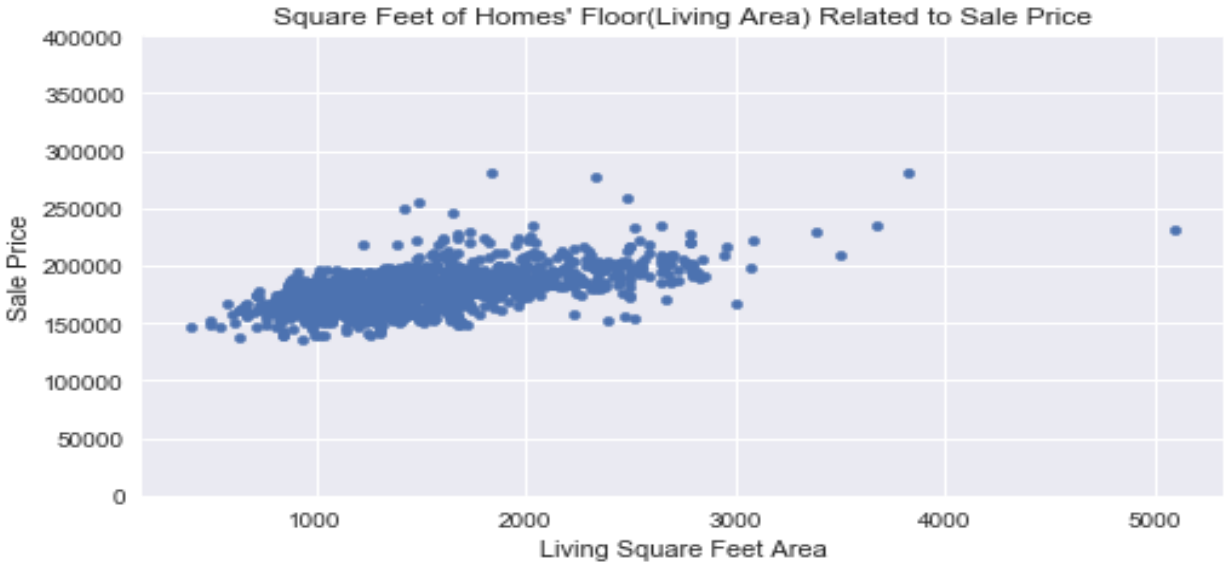
Next, we analyzed the number of bedrooms in the homes. We found how there is a positive relationship between the number of bedrooms and sale price of homes. Most of the homes have between one and five bedrooms, but people seem to prefer the three bedrooms homes as we see a slightly more significant amount of purchases in comparison to others, and the estimate range price of these three bedrooms houses was between $150,000 and $250,000.



After analyzing the number of bathrooms against the homes' prices we found there was no relationship between the ascending numbers of bathrooms to price. Although, we can see that the majority homes have between one to two bathrooms and their corresponding estimated cost was between $140,000 and $280,000.



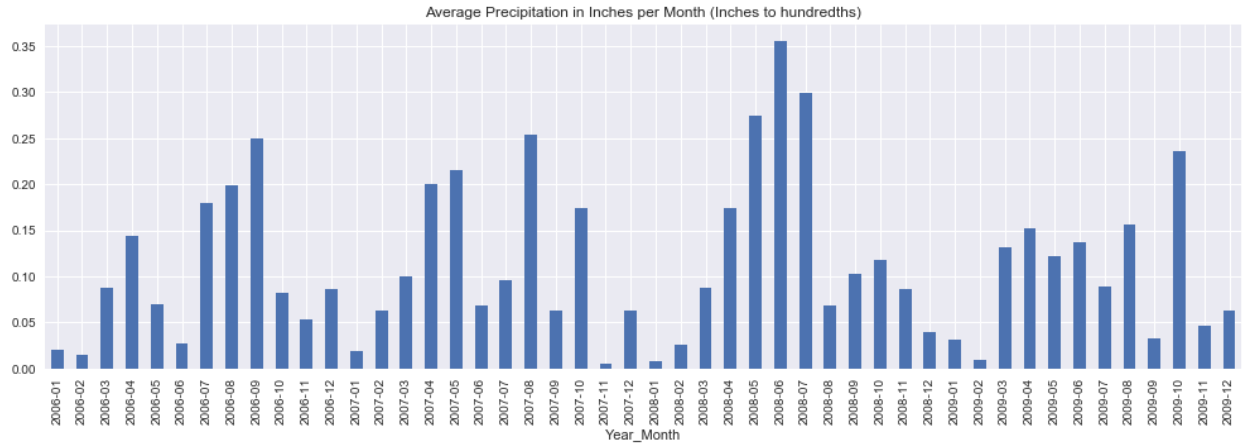
The plot above shows a positive relationship between the homes Lot Square Footage and price, which is expected as the land is an essential asset of real estate properties. Most houses sold had an estimated Lot Square Footage range of 1,000 to 28,000 and a Sale Price range of 140,000 to 240,000. Also, most homes have a Lot Square Footage range of 1,000 to 28,000 and a Sale Price range of 140,000 to 240,000.



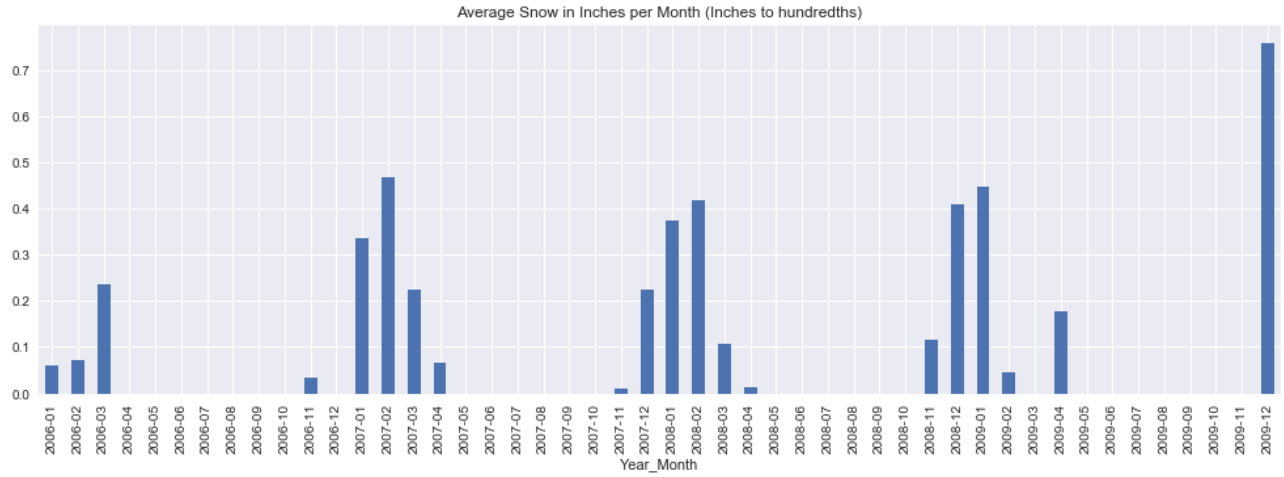
In the case of the Living Area Square Footage and home prices, there no indeed a good relationship. Other insights that we can notice is how most homes sold had an estimated range of Living Square Feet Area of 400 to 3000, and Sale Price range of $140,000 to $240,000.

So far we have notice data points which could be outliers in our data, but as we have mentioned before they will be left alone as we are going to be working on a random forest algorithms which are fortunately not very sensitive to this type of data points.

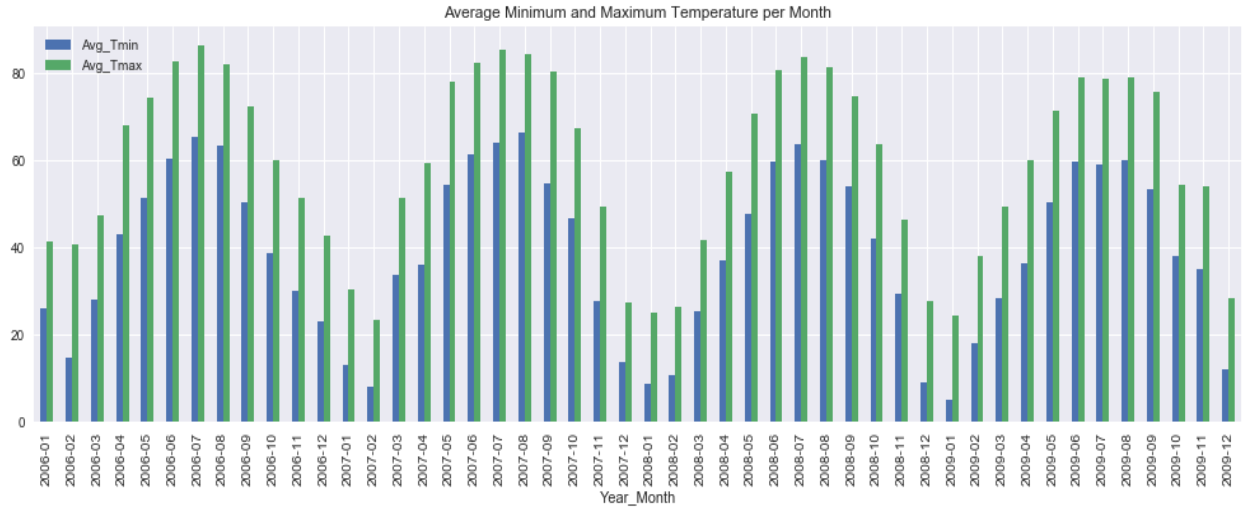
Up next we are going to be analyzing our weather data and how is distributed through the months. This data was exciting to explore since will be used for future analysis and add to our predictive model to see how important weather and timing of the year affects the value of homes.



On this first weather figure, we can see how some winter months show less precipitation than the rest of the months, March through May has an estimated consistent amount of Precipitation of .10 inches throughout the years and the precipitation during summer months highly variated through this four year period.



In regards of snow in our second graph, we see how snow always shows present during winter months, while May through October shows no presence of snow at all.



This last graph shows the average temperatures through the year; we see how May through October had the highest temperatures, with estimated average maximum temperatures range of 70 to 85 degrees. Also, December through February had the lowest temperatures, with estimated average Minimum Temperatures range of 5 to 25 degrees.

Summaries

Surprisingly the visualization of the data on hand served as a significant step for us to find relevant information about these homes, their market, and weather through the year. We can now see what the average price of homes is in the city, how the cost of homes increase as the number of rooms increase, giving us insight on what a person should expect to pay for the homes in the area. We also see how people tend to buy more houses during summer when there is no snow and temperature are higher, probably turning into a seller's market from the increase in homes' demand, and switching to a more difficult negotiation to the buyer in home prices.

All these are valuable trends of the market and weather and can serve as a good starting point when analyzing the value of a home.

Inferential Analysis

We tried to infer specific ideas and try to find answers through our sample data which can hopefully give us insights of the actual population of homes.

Up next is some of the studies done in features such as air units in the homes, One-story vs. Two -story homes, and whether there is a significant difference in price between houses with front fences or not fence. Let’s now go over and check each of these studies.

1. **Finding a significant difference in prices between houses with and without a central air unit**

Ho: Average price of houses with central air unit = average price of houses without central air unit
Ha: Average price of houses with central air unit ≠ average price of houses without central air unit

Average prices between house with and without central air conditioning

House Average Price with Central Air Conditioning: $180,233.19

House Average Price without Central Air Conditioning: $175,487.14

Average difference in prices is: **$4,746**

Is the difference statistically significant within a significant level of .05?

P-value: 0.010090621217186681

Our probability value of 0.010090621217186681(1%) is less than our set significance level 0f .05(5%), so we can reject the null hypothesis and accept the alternative theory. So, we can conclude that there is a statistical difference in prices between houses with central air units and homes without central air units. Also, there is a practical difference since we have almost a 5,000 dollars difference in average between this two different groups.

2. **Finding a significant difference in prices between One-story houses and Two-story houses**

Ho: Average price of 1Story houses = Average price of 2Story houses
Ha: Average price 1Story houses ≠ average price of 2Story houses

Average Price for 1Story Houses: $177,241.89

Average Price for 2Story Houses: $184,862.87

Average difference in prices is: **$7,621**

P-value: 3.5535259636604621e-13

Is the difference statistically significant within a significant level of .05?

Our probability value of 3.5535259636604621e-13 is far less than our set significance level 0f .05(5%), so we can reject the null hypothesis and accept the alternative hypothesis.

We can conclude that there is a statistical difference in prices between 1Story houses and 2Story houses. Having a practical difference of $7,621

3. **Finding a difference in price between houses with a good privacy fence and houses with a minimum privacy fence**

Ho: Average price of houses with a good privacy fence = Average price of houses without a good privacy fence
Ha: Average price of houses with a good privacy fence ≠ Average price of houses without a good privacy fence

Average price of houses with good privacy fence $180536.3744333333

Average price of houses with minimum privacy fence $178473.1263473333

Average difference in prices is: **$2,063**

P-value: 0.37003494954970051

Is the difference statistically significant within a significance level of .05?

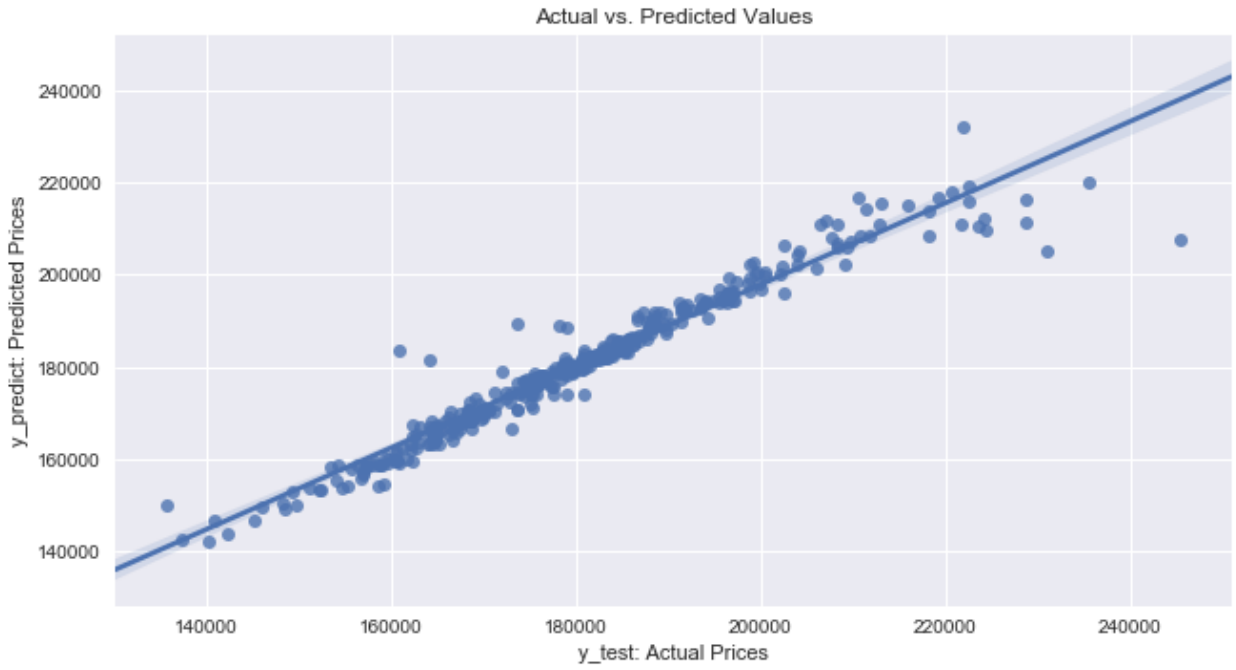
Our p-value of 0.37003494954970051 is greater than our set significance level 0f .05(5%), so we cannot reject the null hypothesis. Further investigation would be needed to determine this.

Machine Learning Modeling and Analysis

The goal of our model is to predict home prices with the data on hand. As mention earlier we will be using Random Forest Predictive algorithm to find these values. For our data, the target variable is continuous, so it best served to use a regression function.

The process done to build our model was as followed; Through the usage of python ski-learn library, we created a random forest of trees regression model object and found the best estimators(parameters) to generate the model, after this, our training data set was fitted into our model. The best parameters for our data consisted of the usage of bootstrapping for our trees, tree depth of eight (edges from the tree's root node to the node), square number of features, and number of estimators (trees) of two hundred and fifty. To obtain our predicted values, we built a price predicted object (y_predict) by scaling out libraries, random forest regressor, best estimators and predict libraries.

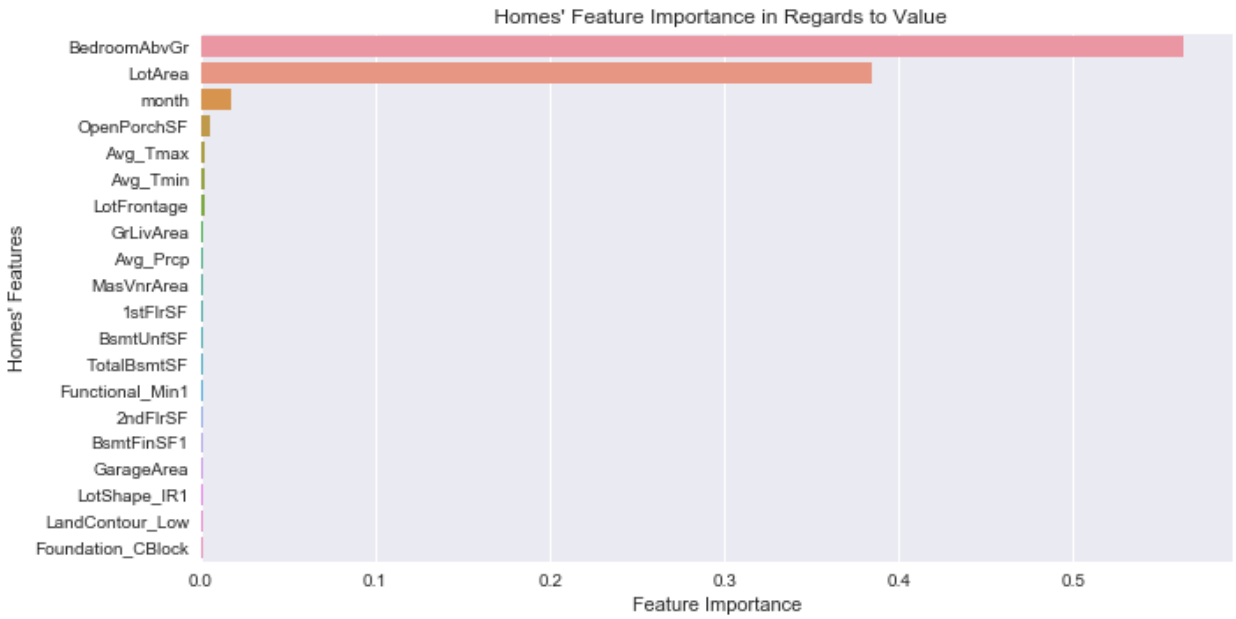
The following graphs show the comparing actual vs. predicted prices. We see a strong positive linear relationship.



The method used to calculate the performance of our model was through the Root Mean Square Error(RMSE), which gave us an error of $4213. This value was analyzed and determined as a small error difference.

The following charts shows the feature importance for our model:

	Features	Importance
20	BedroomAbvGr	5.67e-01
2	LotArea	3.81e-01
34	month	1.72e-02
38	Avg_Tmax	2.60e-03
28	OpenPorchSF	2.22e-03
39	Avg_Tmin	1.96e-03
1	LotFrontage	1.77e-03
12	1stFlrSF	1.55e-03
215	Functional_Min1	1.42e-03
36	Avg_Prcp	1.35e-03



We see how Homes' bedrooms, Lot Square Feet Area, the time of the year when the sale of a home happens, an Open Porch, Average Maximum and Minimum Temperature are some of the critical determinants of prices.

Conclusion

After cleaning and analyzing our data, building our model and subtracting the importance of features, we can determine how Random Forest is a great alternative to the use of linear regression and the determinants of coefficients when analyzing continues target values. We determined by practice and agreed in how this is one of the most useful algorithms in the market; which shows as the outcome accuracy value of our model. Furthermore, through the use of this model, we see how much more important weather and time of the year in regards to determining houses' values are in contrast to other homes features.