Regression Modeling in Predicting Optimal House Sales Price in Ames, Iowa

Dexter Nguyen · Follow
Published in Towards Data Science · 7 min read · Oct 20, 2020

This analysis is a part of my team's project in the Applied Probability and Statistics course at Duke's Fuqua School of Business, MQM Business Analytics Program. I want to send my special thank to Anika Abrahamson, Michael Ruch, Xinying (Silvia) Sun, and Yaqiong (Juno) Cao for their great work.

Photo by Phil Hearing on Unsplash

Business Understanding

Within the housing market, a Comparative Market Analysis, or CMA, is utilized by the broker to present the seller with a proposed sale price and a comprehensive justification for this price (Miller 2018). Although many brokers utilize software to complete a CMA, personal experience and intuition are also employed to decide the proposed price. Our analysis aims to create a regression model for pricing homes in the Ames, Iowa housing market. In theory, pricing homes closer to their “real value” (based on a concrete model) will result in lower resource use on the part of the broker/agency, and thus, a quicker (and more lucrative) sale. There are various models that websites, for example, Zillow.com, apply to provide estimates on the market value of a particular home (McDonald 2006). Our analysis will build a model specifically for Ames, Iowa, that real estate brokers can utilize for their CMA reports to be more confident in their proposed price.

Data Understanding

Our analysis will use the dataset compiled by Dean DeCock in 2011 and published on Kaggle by Mehdi in 2018. The dataset contains 80 variables recorded for 2930 properties in Ames, Iowa (DeCock 2011). This data will allow us to create a linear regression model to find how different independent variables affect our dependent variable, sales price. The knowledge of how each variable will impact the home's price will help real estate brokers better assess a proper sales price for a home in Ames, Iowa.

Our first step was to clean and prepare the data for analysis. We removed the extraneous columns of "Order" and "PID" because they were irrelevant to our research. We chose to change the subclass from numerical to categorical to simplify the computation and visualization of correlation. We removed all N/A values and used a sampling method to impute the missing values, where appropriate (Buuren & Groothuis-Oudshoorn 2011). We also removed columns with over 85% of the values missing. In some cases, we replaced all missing categorical values with the modal value. For our categorical variables, we created dummy variables to allow for the numerical calculation of correlations.

Data Exploration and Transformation

To see which variables are likely to affect the price of homes in Ames, IA the most, we ran a correlation analysis of our independent variables against our dependent variable, sale price. Once this was completed, we chose to keep the top 10 variables of interest, which had the highest price correlation.

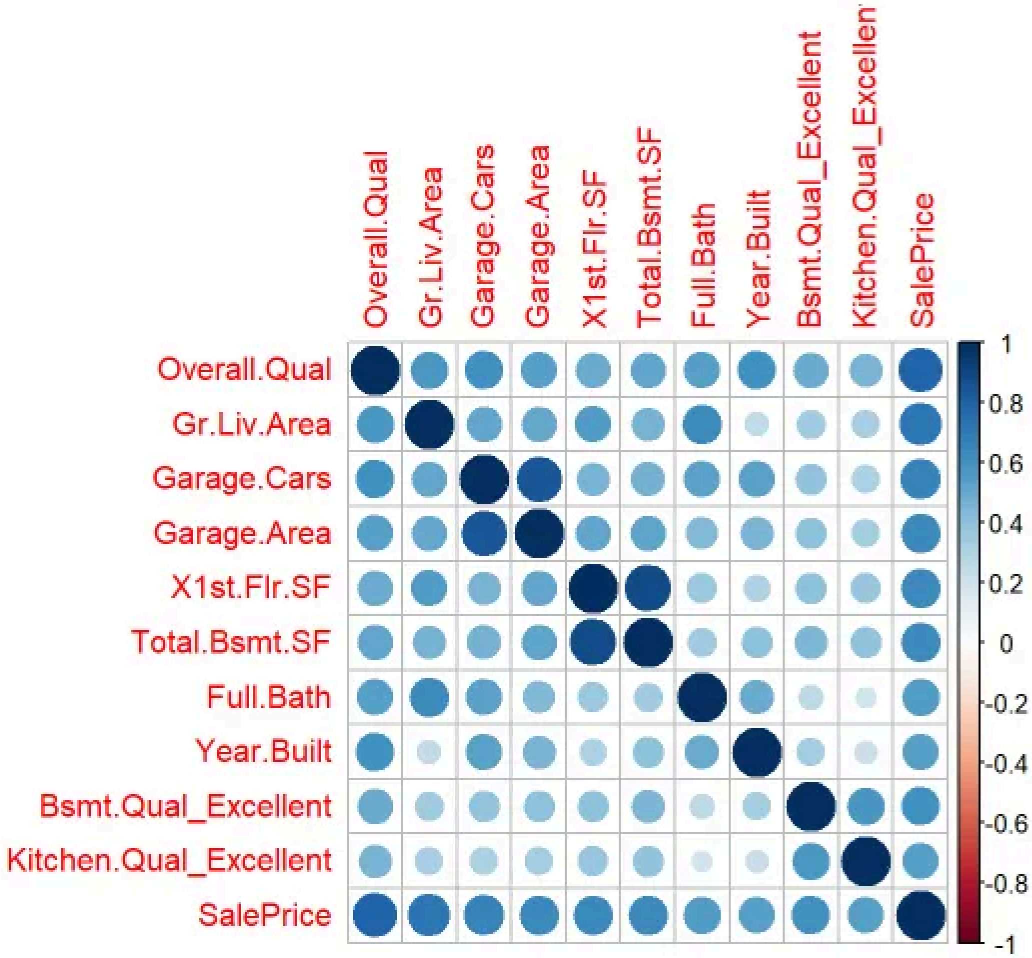


Image by Author

Looking at the distribution of our dependent variable SalePrice, we concluded that our data is not normally distributed. After running a QQ plot, it was clear that we needed to transform our data so it would be normally distributed. The new QQ plot demonstrates that our log(SalePrice) values are much more normally distributed, allowing us to move forward with our analysis.

Next, we plotted the marginal distributions of the key categorical variables of interest and displayed their relationship with price. Not surprisingly, we found clear positive correlations between kitchen quality and price and between basement quality and price. For numerical variables, the first step to further analyze the relationship with our dependent variable was to create density plots visualizing the spread of the data. After analyzing the density plots, we plotted the interaction between our numeric variables of interest and our dependent variable of price. The variables ground floor, living area, garage area, first-floor square footage, total basement square footage, and year built show a similar pattern to the overall quality graph.

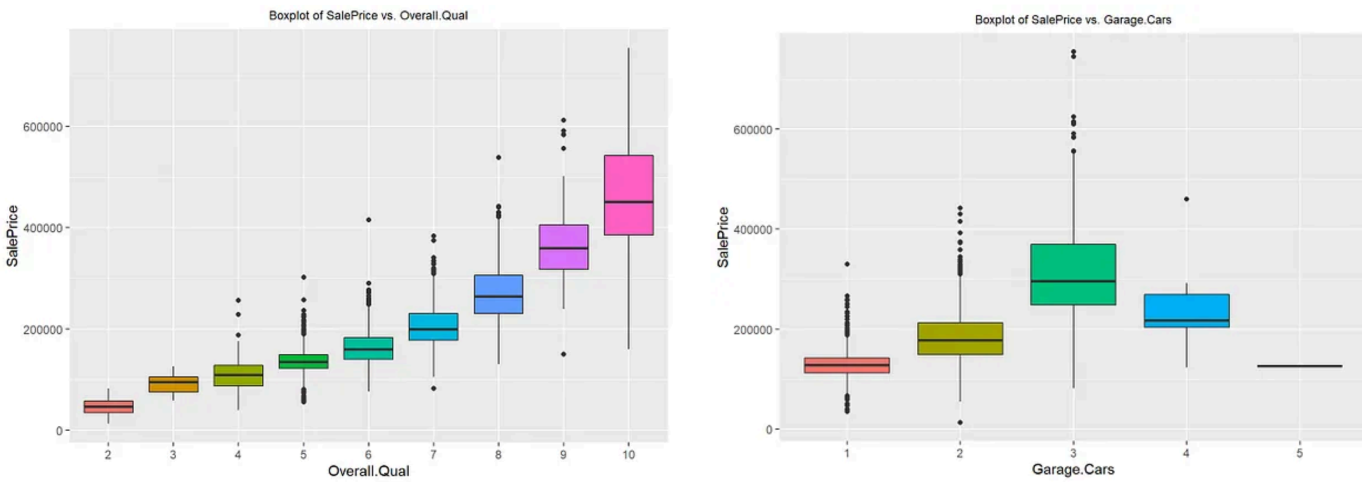


Image by Author

We found a few correlations that did not play out as expected. When inspecting the variables garage cars and full bath (and their respective relationships with price, we saw the expected positive correlation up to a point but then declined. Homes with four-car garages, for example, had generally lower prices than those with three-car garages. Similarly, homes with four full baths had generally lower prices than those with three full baths. We assume that fewer buyers are looking for these attributes and, as such, four-car garages simply command a lower price. It is also likely that there are fewer of these homes on the market and that other variables simply have a more significant impact on price across the relatively small sample of four-car garage / four full bath homes.

Last, we considered if the collinearity problem existed in our analysis. Calculating different correlations among our independent variables, we discovered the high correlation (0.8461) between two variables: Garage.Cars and Garage.Area. This, in addition to the output from our previous analysis,

From the exploratory data analysis, we know that sale price is highly correlated with a number of variables (our “top 9”). We will now employ linear regression to build an optimal pricing model for homes in this market that includes four logarithmically transformed: Overall.Qual, log(Gr.Liv.Area), log(Garage.Area), log(X1st.Flr.SF), log(Total.Bsmt.SF), Bsmt.Qual_Excellent, Full.Bath, Kitchen.Qual_Excellent, and Year.Built.

We created four multi-regression models. Our first model (“Model 1”) included our “top 9” explanatory variables (variables highly correlated with price). We developed our second model (“Model 2”) by removing the Full.Bath variable, which was not statistically significant with its p-value > 0.05. For comparison, we developed Model 3, which included all explanatory variables in the original dataset. In Model 4, we removed non-significant variables (p-values > 0.05) from this more comprehensive model.

Based on the p-value and R-squared performance (and not wanting to over-fit the model), we picked Model 2, which includes all the significant independent variables without the high associated p-values, as our final model.

```
##
## Call:
## lm(formula = log(SalePrice) ~ Overall.Qual + log(Gr.Liv.Area) +
##     log(Garage.Area) + log(X1st.Flr.SF) + log(Total.Bsmt.SF) +
##     Year.Built + Kitchen.Qual_Excellent + Bsmt.Qual_Excellent,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57490 -0.07235  0.00854  0.08602  0.52705
##
## Coefficients:
##                        Estimate Std. Error t value            Pr(>|t|)
## (Intercept)           1.7310802  0.3147271   5.500     0.00000004281046 ***
## Overall.Qual          0.0945891  0.0041915  22.567 < 0.0000000000000002 ***
## log(Gr.Liv.Area)      0.3870152  0.0152852  25.320 < 0.0000000000000002 ***
## log(Garage.Area)      0.0853321  0.0121971   6.996     0.00000000000358 ***
## log(X1st.Flr.SF)      0.1274433  0.0215072   5.926     0.00000000365704 ***
## log(Total.Bsmt.SF)    0.0683001  0.0178556   3.825             0.000135 ***
## Year.Built            0.0025450  0.0001583  16.082 < 0.0000000000000002 ***
## Kitchen.Qual_Excellent 0.0768435  0.0171594   4.478     0.00000795310902 ***
## Bsmt.Qual_Excellent   0.0586147  0.0159077   3.685             0.000235 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1551 on 1991 degrees of freedom
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8339
## F-statistic:  1255 on 8 and 1991 DF,  p-value: < 0.00000000000000022
```

Image by Author

In Model 2, all identified variables are highly correlated with our target variable (SalePrice) and show statistical significance. All the variables have a positive relationship with SalePrice. To be more specific, we expect an average increase of 0.4% in SalePrice for every 1% increase in Gr. Liv. Area, holding other variables constant. Another interpretation can be seen from two variables: Kitchen.Qual and Bsmt.Qual. These two only have positive impacts on the sale price if the value is 'Excellent'. To dive deeper into the regression analysis, we also tried to find interaction among independent variables but found no meaningful insights.

Evaluation

After running our two models: Model 1 and Model 2, we used R-squared and AIC to evaluate our model performance. As we expected, Model 2 is best suited for our business use case. We compared the R-Squared and AIC for Model 1 and Model 2. The evaluation factor is quite close in the two models. The model performance remains high after removing the high p-value variable in the first model. Since we used fewer variables to predict SalePrice and didn't hurt the model performance after removing said high-p-value-variable, we determine that Model 2, which includes highly correlated variables without large p-values, is the highest performance model we have so far.

Implications, limitations, and conclusion

By analyzing the data collected in the Ames, Iowa real estate market, we created a model that can help future sellers price their homes in the market to sell quickly while still generating a profit. The most important factors when determining the price, as determined by our analysis, are the year built, excellent kitchen and basement quality, the square footage of both the basement and first floor, the square footage of both above-grade living area and garage area, and the overall quality (as determined by material and finish) of the home. Because our model is based on these variables, we believe it to be a useful tool for real estate agents to utilize in the Ames, Iowa market.

However, since we used the 2011 dataset to build the model and the real estate market is constantly changing, our best model might not fit the current market. Going forward, we would recommend frequently recording housing specs and sales prices in the Ames area and maintaining a database with the relevant information to continually improve on the model's ability to predict sales price, even in the face of an ever-changing market landscape.

Aside from considering the time-efficiency of using our model, we have to consider a better way to deal with missing values. Besides, regression modeling has its limitations. As such, we would recommend exploring more comprehensive machine learning models and different evaluation methods in the future.

You can visit my GitHub to see more information about this analysis.

Regression Modeling   Housing Prices

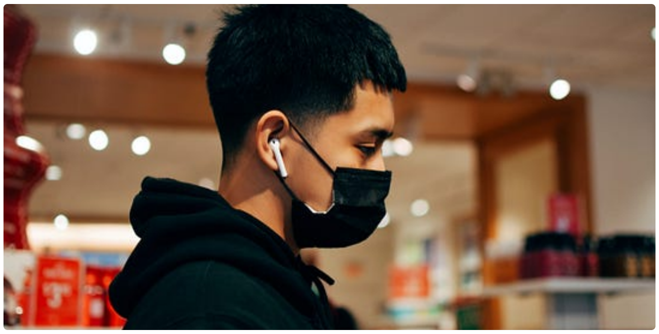
Written by Dexter Nguyen

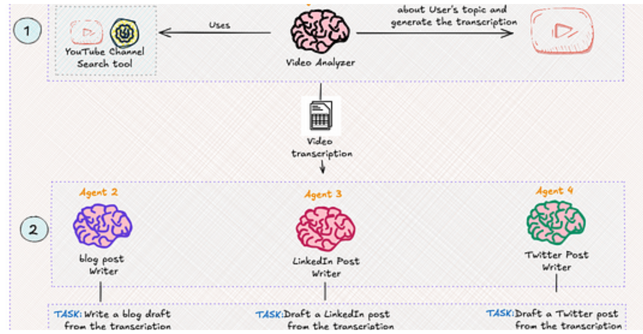
Follow

64 Followers · Writer for Towards Data Science

https://www.linkedin.com/in/dextertinhnguyen/

More from Dexter Nguyen and Towards Data Science





Dexter Nguyen in Towards Data Science

Analyzing Customer Satisfaction of Apple AirPods Using Exploratory...

In Part 1, I went through the statistics regarding the industry, Apple, and AirPods. I...

Dec 30, 2020 · 21

Zoumana Keita in Towards Data Science

AI Agents — From Concepts to Practical Implementation in Python

This will change the way you think about AI and its capabilities

Aug 12 · 1.1K · 13

Ahmed Besbes in Towards Data Science

What Nobody Tells You About RAGs

A deep dive into why RAG doesn't always work as expected: an overview of the...

Aug 23 · 1.5K · 22

Dexter Nguyen in Towards Data Science

Red Wine Quality Prediction Using Regression Modeling and Machin...

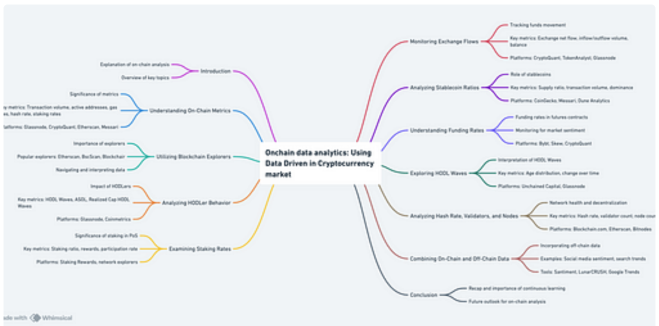
This is my personal project, a part of the Data Science course at Duke Fuqua School of...

Nov 22, 2020 · 28 · 2

See all from Dexter Nguyen

See all from Towards Data Science

Recommended from Medium

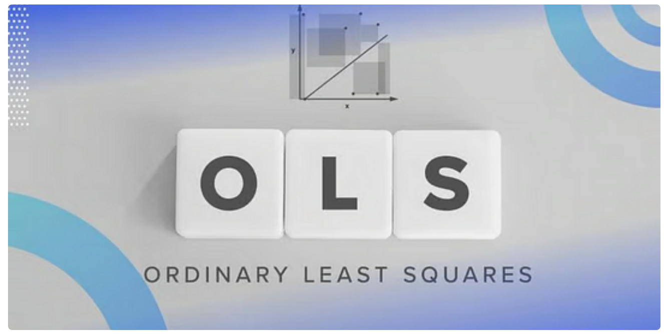


DaxTan

Onchain data analytics: Using Data- Driven in Cryptocurrency...

Introduction

Apr 2



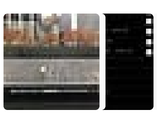
Younes Dahami

Understanding Ordinary Least Squares (OLS) and Its Application...

Ordinary Least Squares (OLS) is a cornerstone method in statistics and machin...

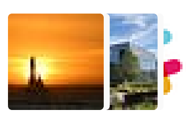
May 22 6

Lists



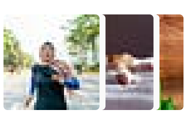
Staff Picks

730 stories · 1287 saves



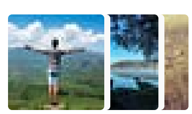
Stories to Help You Level-Up at Work

19 stories · 792 saves



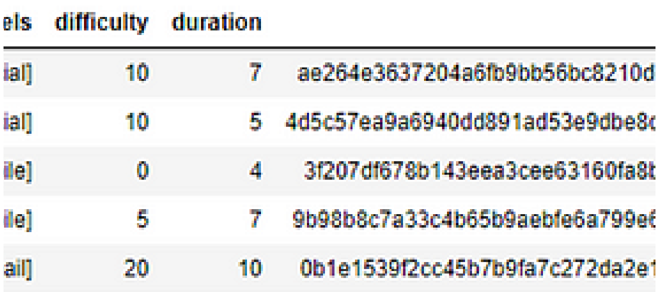
Self-Improvement 101

20 stories · 2716 saves



Productivity 101

20 stories · 2328 saves



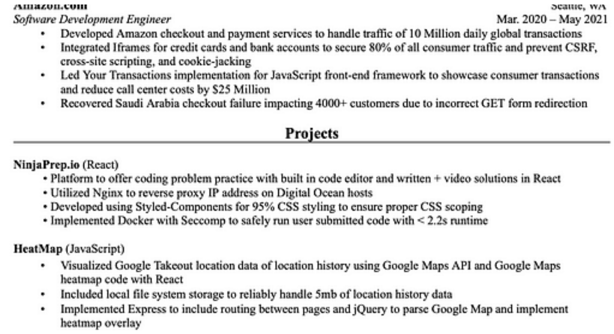
els	difficulty	duration	
ial]	10	7	ae264e3637204a6fb9bb56bc8210d
ial]	10	5	4d5c57ea9a6940dd891ad53e9dbe8c
ile]	0	4	3f207df678b143eea3cee63160fa8b
ile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e6
ail]	20	10	0b1e1539f2cc45b7b9fa7c272da2e1

Giga85

Udacity Data Scientist Capstone Project: Starbucks dataset

Udacity Data Scientist Capstone Project: Starbucks dataset

Jun 3 · 2



Software Development Engineer Mar. 2020 – May 2021

- Developed Amazon checkout and payment services to handle traffic of 10 Million daily global transactions
- Integrated Iframes for credit cards and bank accounts to secure 80% of all consumer traffic and prevent CSRF, cross-site scripting, and cookie-jacking
- Led Your Transactions implementation for JavaScript front-end framework to showcase consumer transactions and reduce call center costs by $25 Million
- Recovered Saudi Arabia checkout failure impacting 4000+ customers due to incorrect GET form redirection

Projects

NinjaPrep.io (React)
- Platform to offer coding problem practice with built in code editor and written + video solutions in React
- Utilized Nginx to reverse proxy IP address on Digital Ocean hosts
- Developed using Styled-Components for 95% CSS styling to ensure proper CSS scoping
- Implemented Docker with Seccomp to safely run user submitted code with < 2.2s runtime

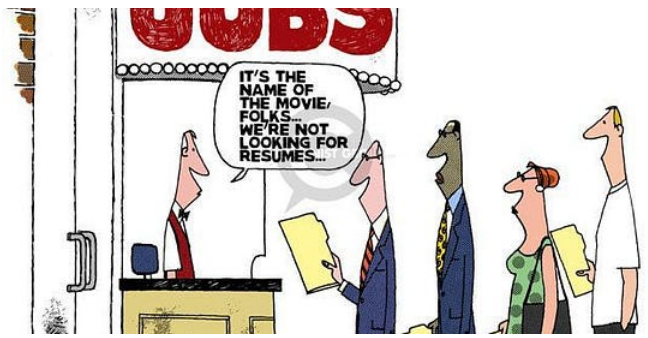
HeatMap (JavaScript)
- Visualized Google Takeout location data of location history using Google Maps API and Google Maps heatmap code with React
- Included local file system storage to reliably handle 5mb of location history data
- Implemented Express to include routing between pages and jQuery to parse Google Map and implement heatmap overlay

Alexander Nguyen in Level Up Coding

The resume that got a software engineer a $300,000 job at Google.

1-page. Well-formatted.

Jun 1 · 20K · 376

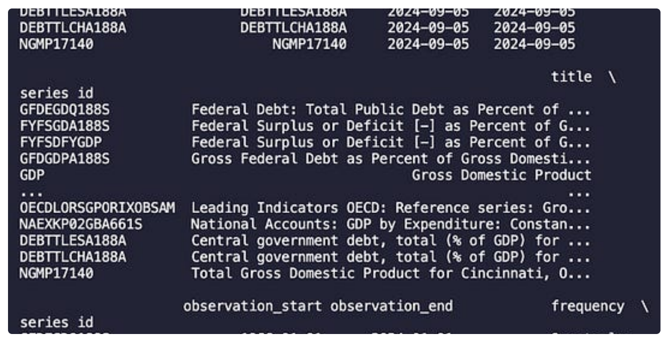


Ethan Duong

Stop aiming at "Data Analyst" position

You won't be a Data Analyst and that is ABSOLUTELY FINE!

Jul 14 · 352 · 17



```
DEBTTLCHA188A          DEBTTLCHA188A   2024-09-05  2024-09-05
NGMP17140                  NGMP17140   2024-09-05  2024-09-05

                                                          title  \
series id
GFDEGDQ188S        Federal Debt: Total Public Debt as Percent of ...
FYFSGDA188S        Federal Surplus or Deficit [-] as Percent of G...
FYFSDFYGDP         Federal Surplus or Deficit [-] as Percent of G...
GFDGDPA188S        Gross Federal Debt as Percent of Gross Domesti...
GDP                                      Gross Domestic Product
...                                                         ...
OECDLORSGPORIXOBSAM Leading Indicators OECD: Reference series: Gro...
NAEXKP02GBA661S    National Accounts: GDP by Expenditure: Constan...
DEBTTLESA188A      Central government debt, total (% of GDP) for ...
DEBTTLCHA188A      Central government debt, total (% of GDP) for ...
NGMP17140          Total Gross Domestic Product for Cincinnati, O...

                 observation_start observation_end          frequency  \
series id
```

Jakub Polec

Comprehensive Guide to Fetching and Analysing Macroeconomic…

Introduction

6d ago · 21

See more recommendations