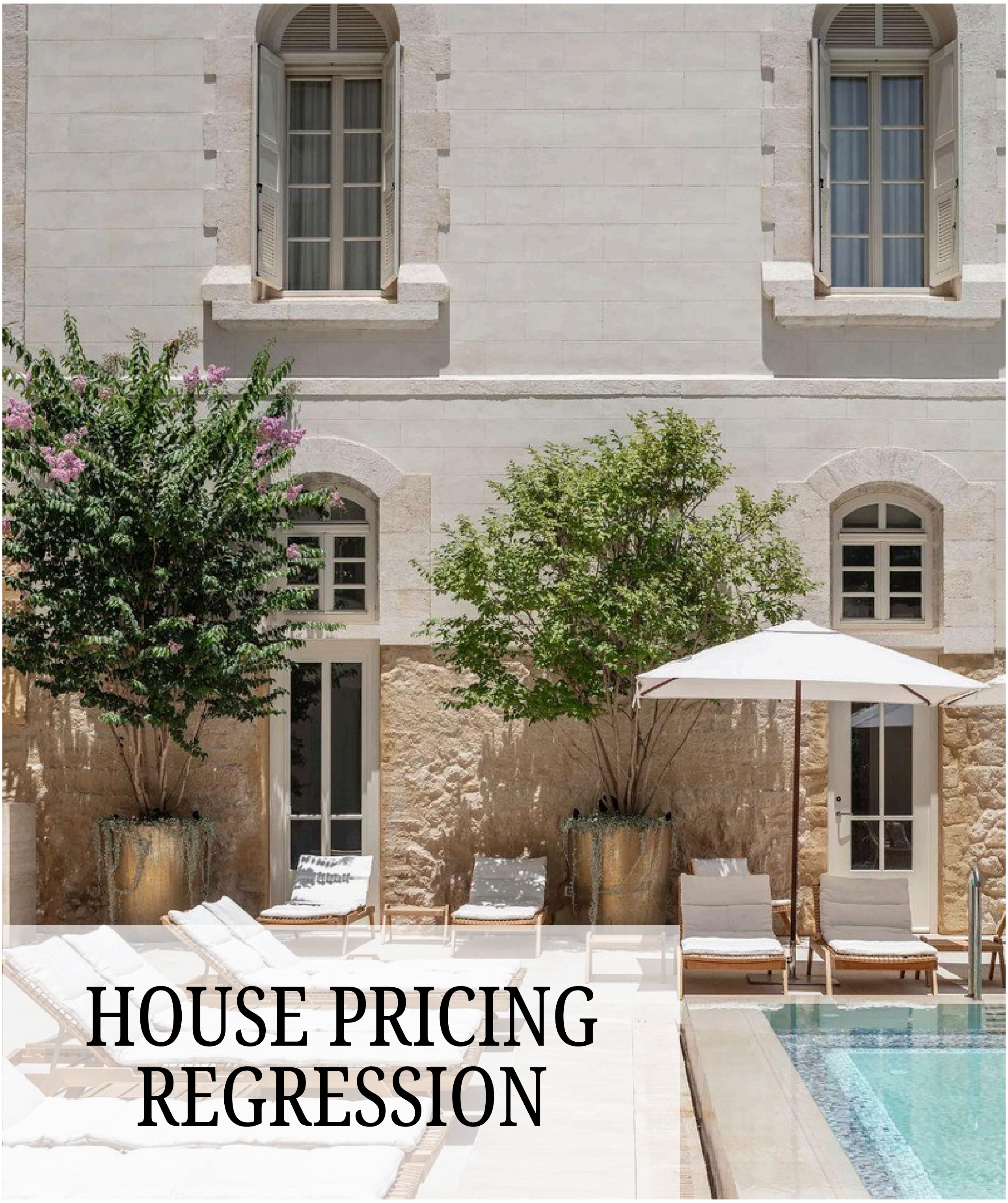


HOUSE PRICING REGRESSION

NOVA IMS Information Management School

CUICKKC CRIIBIS CLB 184 CORRPORATTION

DATA-DRIVEN DECISION-MAKING
JUNE 2024

GROUP 10:
Eduardo Xavier, 20230355
Joana Oliveira, 20230384
Joana Franco, 20230365
João Chang, 20201620
Rúben Machado, 20230367

PROFESSORS
Fernando Bação
Farina Pontejos
Ivo Bernardo

TABLE OF CONTENTS

1. Executive Summary.. 1

2. Available Data 1

3. Model Details and Performance 3

4. Business Outcome 3

5. Deployment 5

6. Monitoring and Maintenance 6

7. Improvements and Next Steps 7

8. Appendices 8

1.EXECUTIVE SUMMARY

The principal aim of this project is to illustrate the benefits of data-driven decision-making in the pricing strategy of Cuckoo Cribs Corporation. The utilization of a predictive model will facilitate enhanced pricing accuracy, operational efficiency and superior business outcomes. Currently, property pricing is based on subjective judgments, which can lead to inconsistencies and errors.

The predictive model will utilize historical data and regression techniques. During the testing phase, the model demonstrated enhanced precision and reliability in pricing accuracy.

Two suggestions were made regarding the deployment of the model into the company's operations, and three suggestions were made regarding how real estate agents across different divisions could use the model in real life.

In the absence of detailed information regarding the specific type of real estate business conducted by Cuckoo Cribs Corporation, we assume that they operate across various areas of the real estate industry. These include investment, property acquisition and management, house flipping, asset management, and traditional brokerage services, where they act as intermediaries between buyers and sellers.

2.AVAILABLE DATA

The dataset comprises 1460 rows and 80 variables regarding house pricing. The initial phase of this project involved an analysis of the data and the subsequent steps to be followed. This was achieved through the use of scatter plots and bar charts (Appendices 2 and 3), which were employed to analyze the descriptive statistics and structure of the dataset.

Upon examination of the statistics view node (Appendix 1), it was verified the presence of string-type and integer-type variables. One key finding from our analysis was that 19 variables contained missing values, recorded as "NA". However, since KNIME did not recognize "NA" as missing values, we used the Rule Engine node to convert all "NA" entries in each affected column to empty values.

Additionally, we converted the data types of LotFrontage and MasVnrArea from string to number (integer) using the “String to Number” node.

In the Data Preparation metanode, the variable ID was excluded in the train and test data. In order to ensure the model’s efficacy and to prevent overfitting, the partitioning node was employed on the training data, with a 70/30 split. Given that the objective is to apply the model to test data in order to obtain predictions, it is unnecessary to split the House test data. For the missing values in both datasets, the mean was applied to the numeric values and the most frequent value was applied to the strings. To treat the outliers we used the replacement strategy of closest permitted value for both datasets.

The Real Estate Data and Test Data datasets include variables with different ranges, which can negatively impact the model and predictions. To address this issue, Min-Max Normalization was applied to the numeric values, ensuring that all variables have a range of 0 to 1. This normalization process is crucial for improving the learning process and overall model accuracy, as it ensures that all variables contribute equally. Normalizing the House test data helps us to make sure that the training and testing conditions are the same, so that we get accurate predictions.

Upon examination of the Correlation Matrix (Appendix 4) and checking the Correlation Filter node, we decided to remove the variables 1stFlrSF, TotRmsAbvGrd and GarageCars due to high correlation, since in this regression project we defined the threshold as 0.8.

According to this, the Correlation Filter node removed the GarageArea instead of the GarageCars. However, as the parking area depends on the size of the car the customer has, we considered GarageArea more indicative of their needs than the number of garage cars.

Furthermore, the Correlation Filter node identified a high correlation between OverallQual and the target feature, SalesPrice (0.8164). However, we decided to keep both variables as it is important to preserve features that are good predictors of the target.

3.MODEL DETAILS AND PERFORMANCE

In an effort to find the “intrinsic value” of a property, three distinct regression models were trained on the train set: Random Forest, Gradient Boosted Trees and Polynomial Regression.

The Gradient Boosted Trees model was selected as the optimal choice, exhibiting the highest R-squared (R^2) score (0,89), the lowest root mean squared error (RMSE) (22 566,295) and the lowest mean absolute percentage error (MAPE) (0,09), visible in Appendix 5. This model demonstrated the ability to handle complex non-linear relationships and some noise in the data. The R^2 metric represents the proportion of variance in the dependent variable (SalesPrice) that is explained by the independent variables (the variables on which the prediction is based). It ranges from 0 to 1, where 1 represents a perfect fit. The RMSE metric is the standard deviation of the residuals (prediction errors). A lower RMSE indicates that the model's predictions are, on average, closer to the actual target values. Lastly, the MAPE metric measure how accurate predictions are, especially in forecasting. In this case, on average, the forecasts were 9% off the actual price.

In simple terms, this model demonstrated the greatest capacity to explain the influence of the features (such as house size) on sales price, while also providing a comprehensive understanding of the data and accurate predictions.

4.BUSINESS OUTCOME

Our innovative “pricing machine” will transform the way Cuckoo Cribs Corporation predicts house prices, revolutionizing the organization’s business and decision-making process. By employing this machine learning model, the company can transition from traditional subjective methods to a more objective, accurate, and, most importantly, lucrative strategy. The following scenarios will present a series of estimates regarding the potential financial gains that this model could bring to the business. It should be noted that these calculations are an approximation and do not account for taxes and other costs.

In the **first scenario**, the identification of undervalued properties is facilitated by the utilization our “pricing machine”. This enables Cuckoo Cribs Corporation to identify and purchase such properties with greater accuracy than would otherwise be possible, eliminating the need to rely on subjective judgment.

To illustrate, consider the following example: If the model permits Cuckoo Cribs to purchase properties at a 10% discount on average and the company purchases 50 properties per year, each valued at €250 000, this could result in savings of €1.25 million. Such savings could then be reinvested in additional properties.

In the **second scenario**, the brokerage services offered by Cuckoo Cribs are enhanced. The increased accuracy of property valuations provided by our model lends greater reliability to the Cuckoo Cribs' brokerage services, which in turn has the potential to attract a larger client base. This also enables us to select clients in a more informed manner. If a client wishes to sell a house but has a price expectation that exceeds the suggested figure provided by our pricing machine, it is preferable to decline the client's request.

To illustrate, consider the following example: If we attract 60 clients and each client generates a commission of €10 000, this results in an additional €600 000 in annual revenue for the company's brokerage services division.

In the **third scenario**, operational efficiency is enhanced by the implementation of our "pricing machine." This technology optimizes operational processes at Cuckoo Cribs by reducing the time and human resources required for manual property evaluations and decision-making. If Cuckoo Cribs currently spends €300,000 annually on evaluations and human resources related to these processes, and our model allows the company to reduce these costs by 50%, the company can save €150 000 annually.

Total Impact for the 3 scenarios: €2 million annually

The aforementioned improvements have the dual benefit of enhancing operational efficiency and contributing to revenue growth and cost savings at Cuckoo Cribs Corporation. While the financial gains have not yet been calculated, there is also the potential to license the model to other real estate companies. If this is an option Cuckoo Cribs Corporation is interested in exploring, they should license the model to organizations operating in markets that are not currently served by the company, or that the company wishes to expand into. In order to protect the company's business interests, non-competitive clauses would be included in any such licensing agreements.

5. DEPLOYMENT

It is of the utmost importance to engage the entire company in the deployment of the "pricing machine," integrating the predictive model into the daily activities of Cuckoo Cribs' real estate agents. Given that the objective is to enhance accuracy and efficiency in property pricing and decision-making, it is recommended that training sessions be conducted for all real estate agents, as they are the primary users of the model. This will facilitate comprehension of the model's functionality and prevent resistance to change. It is imperative that employees are made aware that the introduction of the model will not result in the replacement of their roles. In the initial stages of implementation, it is also recommended that there be constant support available in case of any doubts. The following section presents two potential deployment strategies for the model and three potential real-world applications for real estate agents across different divisions of the company.

User-Friendly Dashboard

Creation of an interactive dashboard where agents can input property details and receive price predictions with confidence intervals. This dashboard would be created using Business Intelligence tools like Power BI or Tableau. The dashboard simplifies data access and visualization, which is relevant for non-experts in machine learning, making it easier for real estate agents to make informed decisions.

Mobile Application

Creation of a mobile application where agents can input property details. This app would integrate our model and provide price valuations and alerts for undervalued properties. It would stimulate flexibility and responsiveness, allowing agents to work from anywhere and providing clients with instant insights, making it easier for agents to make informed decisions on the go.

Identifying Investment Opportunities

Real estate agents can use the model's ability to predict prices to identify undervalued properties currently in the market. By scanning the market and analyzing data, they can identify properties with growth potential, facilitating investment recommendations. This not only diversifies Cuckoo Cribs and their clients' portfolios but also supports their house flipping division by enabling agents to verify how specific house improvements could increase property prices.

This data-driven approach allows to increase returns on investments across the different divisions of the company.

Accurate Price Determination

The "pricing machine" helps real estate agents determine the market value of properties before listing them. This ensures competitive pricing, attracting buyers faster and increasing turnover. By providing precise valuations based on data analysis, agents can also enhance client trust and satisfaction, establishing a solid foundation for successful transactions. This benefits the sales, listing, investment, and asset management divisions.

Market Segmentation and Targeting

The model can be used to segment the market based on various criteria such as demographics, property types, locations, and other general property characteristics. Real estate agents can then refine their marketing and customer engagement strategies, ensuring personalized communication and targeted property recommendations that appeal to specific market segments. This idea involves technical work related to marketing, data mining, and business intelligence, particularly in understanding the model, creating clusters using unsupervised learning techniques, and extracting meaningful insights for marketing strategies. If this concept is something Cuckoo Cribs wants to explore, they should provide training to current real estate agents or hire new ones with technical experience in these areas.

6.MONITORING AND RETRAINING

To guarantee the continued accuracy of the house price model, it is vital to conduct regular monitoring and retraining.

Firstly, it is important to maintain the quality of the model by regularly cleaning and updating the data. This is crucial to prevent data quality degradation and ensure that the model continues to make accurate predictions. This process involves the identification and rectification of any missing values, outliers, or anomalies that may arise during the collection of new data.

It is similarly important to conduct regular analysis of the relationship between variables and the target in order to ensure that these remain relevant. For instance, if the significance of a feature increases over time, the model should be updated to reflect this.

Secondly, it is also important to confirm the performance of the model in real-world scenarios. This could involve implementing a continuous evaluation process, whereby actual property sales prices are compared with the model's predictions. Significant pricing deviations would then be examined in order to ascertain the root cause of the problem.

By updating and retraining the model on a regular basis with the most recent data, Cuckoo Cribs Corporation will be able to maintain its competitive position in the property market through the implementation of accurate and current data-driven pricing strategies. This will ensure that the pricing strategy remains reliable and provides valuable data-driven insights.

7. IMPROVEMENTS AND NEXT STEPS

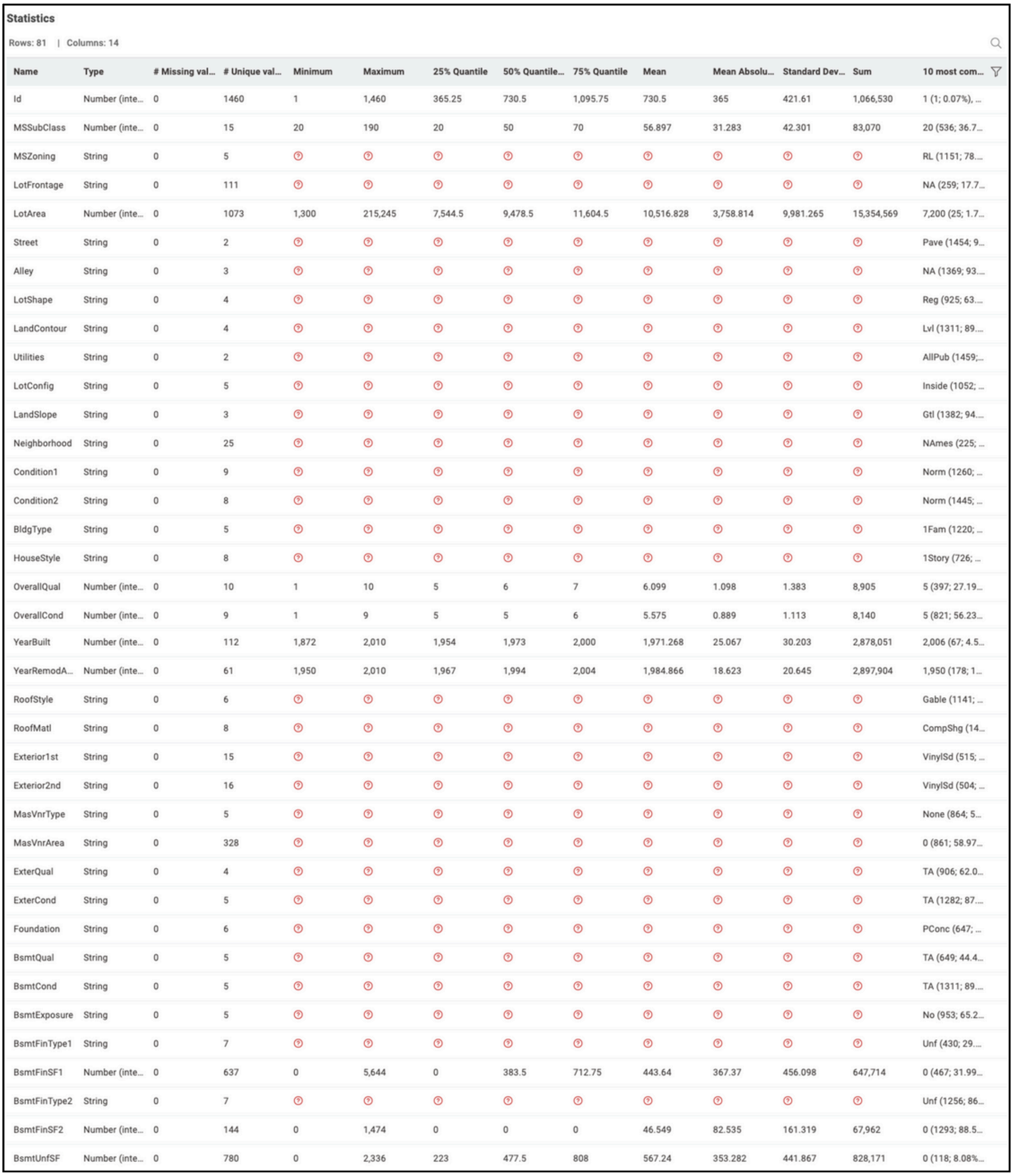
There is always room for improvement, and our predictive model is no exception.

A notable enhancement would be to equip the model with the capacity to discern time-based trend projections, such as seasonal variations or long-term market trends. Furthermore, in order for the company to become more data-driven, it is recommended that real-time data be integrated.

Additionally, the property management operations of Cuckoo Cribs would be enhanced by the implementation of a model that predicts rent prices.

The next stage of the process involves the regular updating of the model in order to reflect any changes in the data set, the expansion of the data sources, the tuning of the model's parameters, and the establishment of a feedback loop for real estate agents to provide suggestions based on their experience.

8.APPENDICES

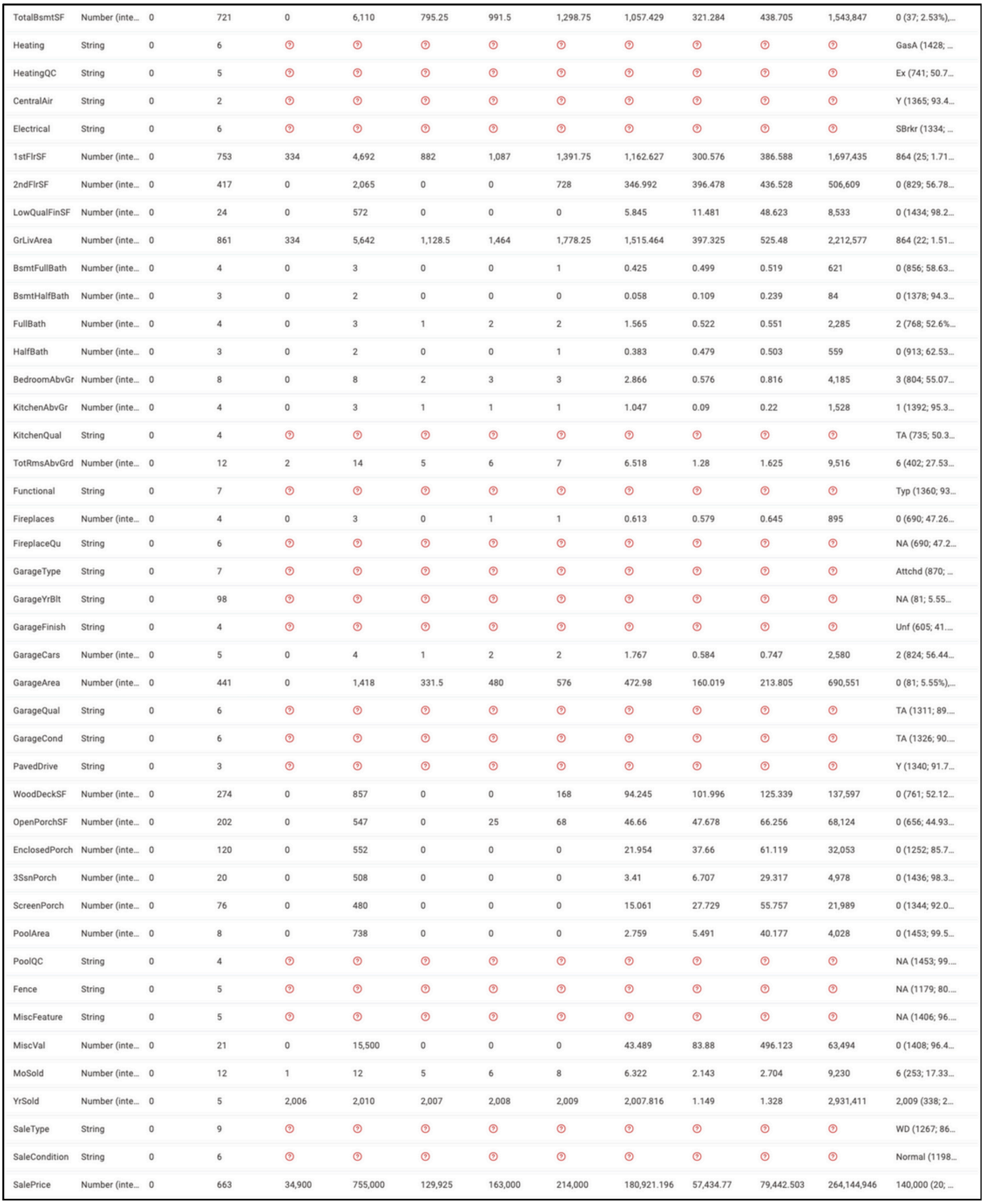


Statistics

Rows: 81 | Columns: 14

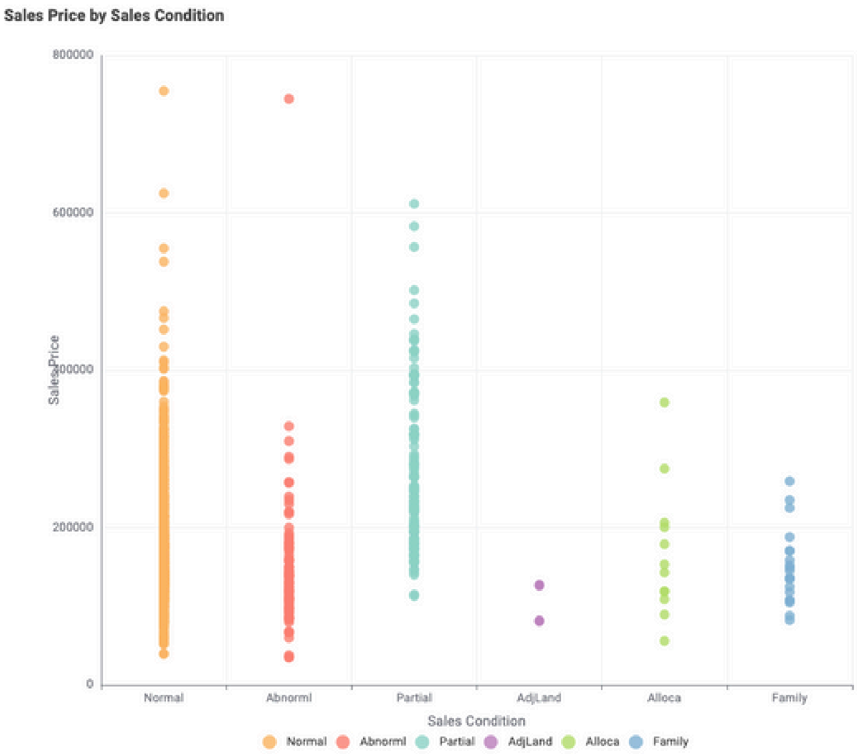
Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Mean	Mean Absolu...	Standard Dev...	Sum	10 most com...
Id	Number (inte...	0	1460	1	1,460	365.25	730.5	1,095.75	730.5	365	421.61	1,066,530	1 (1; 0.07%), ...
MSSubClass	Number (inte...	0	15	20	190	20	50	70	56.897	31.283	42.301	83,070	20 (536; 36.7...
MSZoning	String	0	5	?	?	?	?	?	?	?	?	?	RL (1151; 78....
LotFrontage	String	0	111	?	?	?	?	?	?	?	?	?	NA (259; 17.7...
LotArea	Number (inte...	0	1073	1,300	215,245	7,544.5	9,478.5	11,604.5	10,516.828	3,758.814	9,981.265	15,354,569	7,200 (25; 1.7...
Street	String	0	2	?	?	?	?	?	?	?	?	?	Pave (1454; 9...
Alley	String	0	3	?	?	?	?	?	?	?	?	?	NA (1369; 93....
LotShape	String	0	4	?	?	?	?	?	?	?	?	?	Reg (925; 63....
LandContour	String	0	4	?	?	?	?	?	?	?	?	?	Lvl (1311; 89....
Utilities	String	0	2	?	?	?	?	?	?	?	?	?	AllPub (1459;...
LotConfig	String	0	5	?	?	?	?	?	?	?	?	?	Inside (1052; ...
LandSlope	String	0	3	?	?	?	?	?	?	?	?	?	Gtl (1382; 94....
Neighborhood	String	0	25	?	?	?	?	?	?	?	?	?	NAmes (225; ...
Condition1	String	0	9	?	?	?	?	?	?	?	?	?	Norm (1260; ...
Condition2	String	0	8	?	?	?	?	?	?	?	?	?	Norm (1445; ...
BldgType	String	0	5	?	?	?	?	?	?	?	?	?	1Fam (1220; ...
HouseStyle	String	0	8	?	?	?	?	?	?	?	?	?	1Story (726; ...
OverallQual	Number (inte...	0	10	1	10	5	6	7	6.099	1.098	1.383	8,905	5 (397; 27.19...
OverallCond	Number (inte...	0	9	1	9	5	5	6	5.575	0.889	1.113	8,140	5 (821; 56.23...
YearBuilt	Number (inte...	0	112	1,872	2,010	1,954	1,973	2,000	1,971.268	25.067	30.203	2,878,051	2,006 (67; 4.5...
YearRemodA...	Number (inte...	0	61	1,950	2,010	1,967	1,994	2,004	1,984.866	18.623	20.645	2,897,904	1,950 (178; 1...
RoofStyle	String	0	6	?	?	?	?	?	?	?	?	?	Gable (1141; ...
RoofMatl	String	0	8	?	?	?	?	?	?	?	?	?	CompShg (14...
Exterior1st	String	0	15	?	?	?	?	?	?	?	?	?	VinylSd (515; ...
Exterior2nd	String	0	16	?	?	?	?	?	?	?	?	?	VinylSd (504; ...
MasVnrType	String	0	5	?	?	?	?	?	?	?	?	?	None (864; 5...
MasVnrArea	String	0	328	?	?	?	?	?	?	?	?	?	0 (861; 58.97...
ExterQual	String	0	4	?	?	?	?	?	?	?	?	?	TA (906; 62.0...
ExterCond	String	0	5	?	?	?	?	?	?	?	?	?	TA (1282; 87....
Foundation	String	0	6	?	?	?	?	?	?	?	?	?	PConc (647; ...
BsmtQual	String	0	5	?	?	?	?	?	?	?	?	?	TA (649; 44.4...
BsmtCond	String	0	5	?	?	?	?	?	?	?	?	?	TA (1311; 89....
BsmtExposure	String	0	5	?	?	?	?	?	?	?	?	?	No (953; 65.2...
BsmtFinType1	String	0	7	?	?	?	?	?	?	?	?	?	Unf (430; 29....
BsmtFinSF1	Number (inte...	0	637	0	5,644	0	383.5	712.75	443.64	367.37	456.098	647,714	0 (467; 31.99...
BsmtFinType2	String	0	7	?	?	?	?	?	?	?	?	?	Unf (1256; 86...
BsmtFinSF2	Number (inte...	0	144	0	1,474	0	0	0	46.549	82.535	161.319	67,962	0 (1293; 88.5...
BsmtUnfSF	Number (inte...	0	780	0	2,336	223	477.5	808	567.24	353.282	441.867	828,171	0 (118; 8.08%...

Appendix 1. Statistics View Table

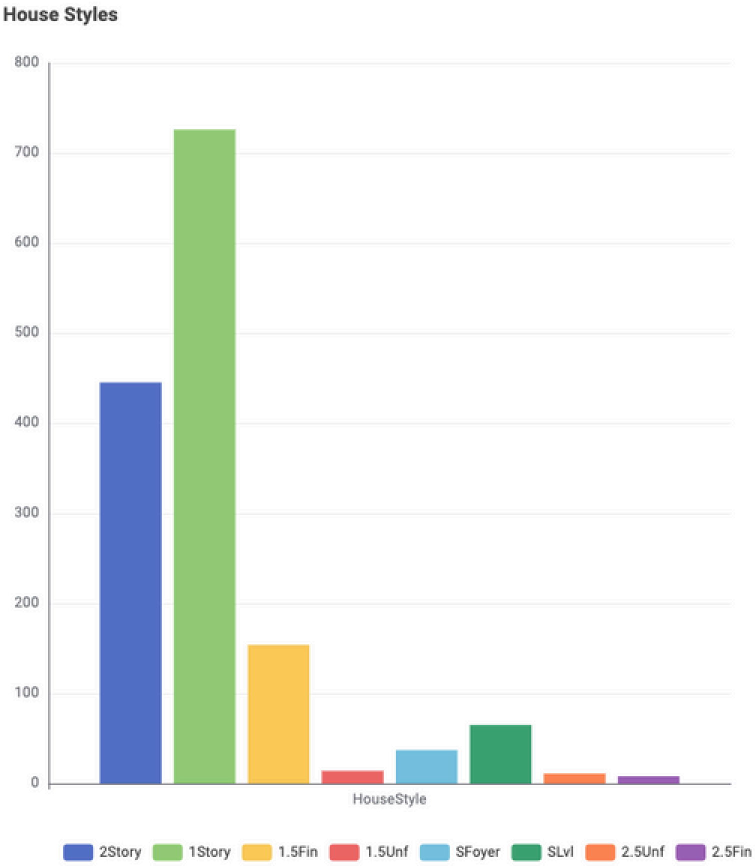


TotalBsmtSF	Number (inte…	0	721	0	6,110	795.25	991.5	1,298.75	1,057.429	321.284	438.705	1,543,847	0 (37; 2.53%),…
Heating	String	0	6	?	?	?	?	?	?	?	?	?	GasA (1428; …
HeatingQC	String	0	5	?	?	?	?	?	?	?	?	?	Ex (741; 50.7…
CentralAir	String	0	2	?	?	?	?	?	?	?	?	?	Y (1365; 93.4…
Electrical	String	0	6	?	?	?	?	?	?	?	?	?	SBrkr (1334; …
1stFlrSF	Number (inte…	0	753	334	4,692	882	1,087	1,391.75	1,162.627	300.576	386.588	1,697,435	864 (25; 1.71…
2ndFlrSF	Number (inte…	0	417	0	2,065	0	0	728	346.992	396.478	436.528	506,609	0 (829; 56.78…
LowQualFinSF	Number (inte…	0	24	0	572	0	0	0	5.845	11.481	48.623	8,533	0 (1434; 98.2…
GrLivArea	Number (inte…	0	861	334	5,642	1,128.5	1,464	1,778.25	1,515.464	397.325	525.48	2,212,577	864 (22; 1.51…
BsmtFullBath	Number (inte…	0	4	0	3	0	0	1	0.425	0.499	0.519	621	0 (856; 58.63…
BsmtHalfBath	Number (inte…	0	3	0	2	0	0	0	0.058	0.109	0.239	84	0 (1378; 94.3…
FullBath	Number (inte…	0	4	0	3	1	2	2	1.565	0.522	0.551	2,285	2 (768; 52.6%…
HalfBath	Number (inte…	0	3	0	2	0	0	1	0.383	0.479	0.503	559	0 (913; 62.53…
BedroomAbvGr	Number (inte…	0	8	0	8	2	3	3	2.866	0.576	0.816	4,185	3 (804; 55.07…
KitchenAbvGr	Number (inte…	0	4	0	3	1	1	1	1.047	0.09	0.22	1,528	1 (1392; 95.3…
KitchenQual	String	0	4	?	?	?	?	?	?	?	?	?	TA (735; 50.3…
TotRmsAbvGrd	Number (inte…	0	12	2	14	5	6	7	6.518	1.28	1.625	9,516	6 (402; 27.53…
Functional	String	0	7	?	?	?	?	?	?	?	?	?	Typ (1360; 93…
Fireplaces	Number (inte…	0	4	0	3	0	1	1	0.613	0.579	0.645	895	0 (690; 47.26…
FireplaceQu	String	0	6	?	?	?	?	?	?	?	?	?	NA (690; 47.2…
GarageType	String	0	7	?	?	?	?	?	?	?	?	?	Attchd (870; …
GarageYrBlt	String	0	98	?	?	?	?	?	?	?	?	?	NA (81; 5.55…
GarageFinish	String	0	4	?	?	?	?	?	?	?	?	?	Unf (605; 41.…
GarageCars	Number (inte…	0	5	0	4	1	2	2	1.767	0.584	0.747	2,580	2 (824; 56.44…
GarageArea	Number (inte…	0	441	0	1,418	331.5	480	576	472.98	160.019	213.805	690,551	0 (81; 5.55%),…
GarageQual	String	0	6	?	?	?	?	?	?	?	?	?	TA (1311; 89.…
GarageCond	String	0	6	?	?	?	?	?	?	?	?	?	TA (1326; 90.…
PavedDrive	String	0	3	?	?	?	?	?	?	?	?	?	Y (1340; 91.7…
WoodDeckSF	Number (inte…	0	274	0	857	0	0	168	94.245	101.996	125.339	137,597	0 (761; 52.12…
OpenPorchSF	Number (inte…	0	202	0	547	0	25	68	46.66	47.678	66.256	68,124	0 (656; 44.93…
EnclosedPorch	Number (inte…	0	120	0	552	0	0	0	21.954	37.66	61.119	32,053	0 (1252; 85.7…
3SsnPorch	Number (inte…	0	20	0	508	0	0	0	3.41	6.707	29.317	4,978	0 (1436; 98.3…
ScreenPorch	Number (inte…	0	76	0	480	0	0	0	15.061	27.729	55.757	21,989	0 (1344; 92.0…
PoolArea	Number (inte…	0	8	0	738	0	0	0	2.759	5.491	40.177	4,028	0 (1453; 99.5…
PoolQC	String	0	4	?	?	?	?	?	?	?	?	?	NA (1453; 99.…
Fence	String	0	5	?	?	?	?	?	?	?	?	?	NA (1179; 80.…
MiscFeature	String	0	5	?	?	?	?	?	?	?	?	?	NA (1406; 96.…
MiscVal	Number (inte…	0	21	0	15,500	0	0	0	43.489	83.88	496.123	63,494	0 (1408; 96.4…
MoSold	Number (inte…	0	12	1	12	5	6	8	6.322	2.143	2.704	9,230	6 (253; 17.33…
YrSold	Number (inte…	0	5	2,006	2,010	2,007	2,008	2,009	2,007.816	1.149	1.328	2,931,411	2,009 (338; 2…
SaleType	String	0	9	?	?	?	?	?	?	?	?	?	WD (1267; 86…
SaleCondition	String	0	6	?	?	?	?	?	?	?	?	?	Normal (1198…
SalePrice	Number (inte…	0	663	34,900	755,000	129,925	163,000	214,000	180,921.196	57,434.77	79,442.503	264,144,946	140,000 (20; …

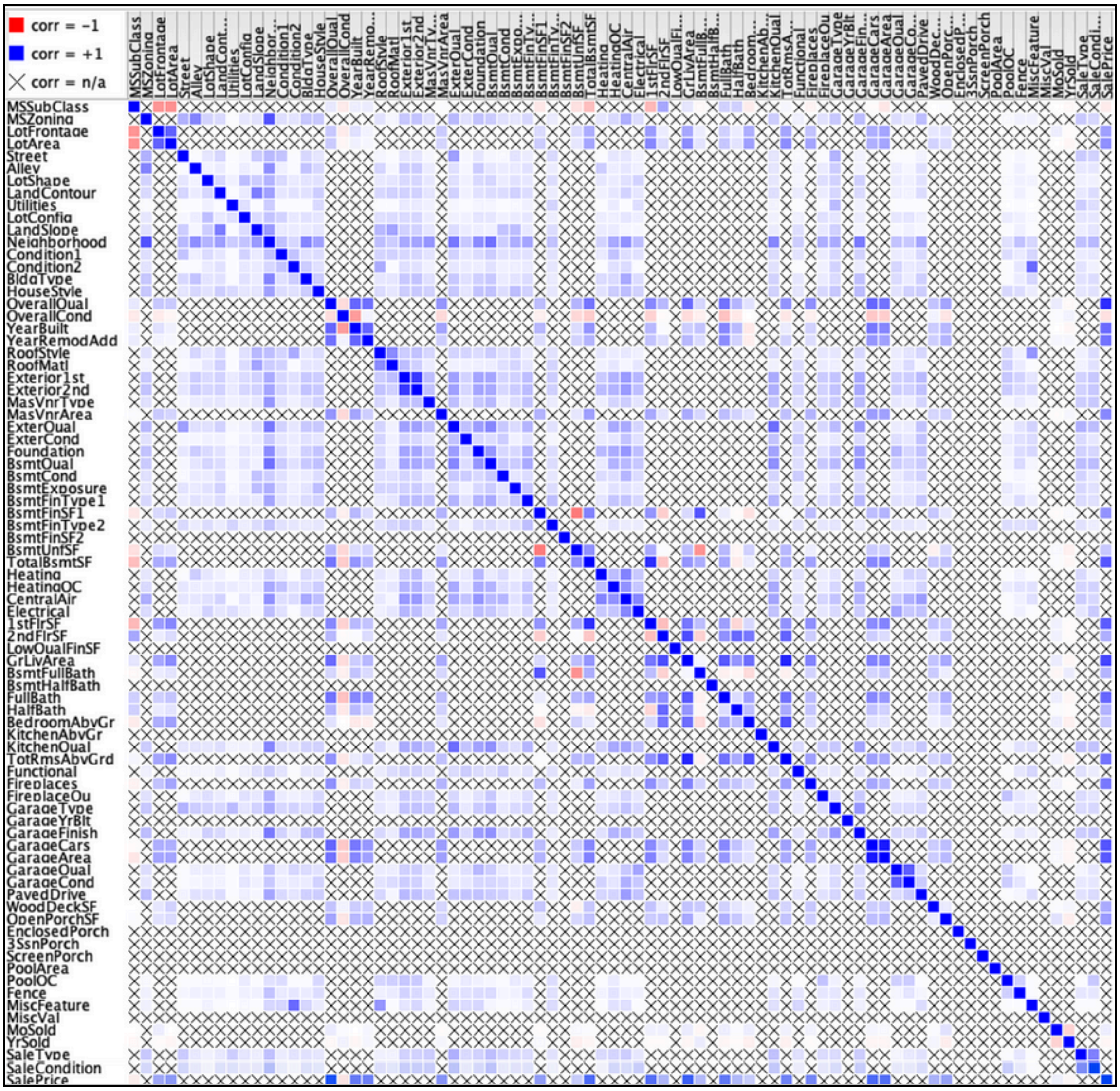
Appendix 1. (cont.) Statistics View Table



Appendix 2. Scatter Plot - Sales Price by Sales Condition



Appendix 3. Bar Chart - House Styles



Appendix 4. Correlation Matrix

	R^2	RMSE	MAPE
Random Forest	0.737	34 928,358	0,174
Gradient Boosted Trees	0.89	22 566,295	0,09
Polynomial Regression	0.825	28 500,619	0,101

Appendix 5. Scorers