House Price Analysis

Final Group Project - Team 10

Quique De La Cruz (U87436598), Tina Hoang (U09660658), Marshall Naegelin (U64517278), Ania Shaheed (U27925140), Julius Yang (U58049158)

BA222 E1

Professor Feng

December 7, 2024

I. Data Description

The *HousePrices.csv* dataset consists of 1,460 observations detailing residential property sales in Ames, Iowa, between 2006 and 2010. The dataset includes 81 variables, categorized as follows: a property identifier variable (*Id*), the dependent variable (*SalePrice*) serving as the target variable for analysis and modeling, and 79 independent variables. We classified the independent variables based on their data type and characteristics. This classification influences how each variable is analyzed, transformed, and included in regression. These classifications are detailed in *Table 1*.

The dataset had a notable amount of null values, described in *Table 2*. These were largely due to the absence of a certain feature in some properties (e.g., no garage leading to a missing *GarageYrBlt*), We replaced these null values with 0 to indicate the absence of the feature. However, the variable *LotFrontage* had genuine missing values, so we opted to replace these nulls with the mean. This was based on the assumption that the average lot frontage was a reasonable representation of the true missing value.

To understand the distribution of *SalePrice*, we calculated its summary statistics (*Table 3*) and plotted its histogram (*Fig. 1*). Visually, it is skewed to the right, most likely reflecting luxury homes and their impact on the overall distribution. Although the typical range for a home is roughly \$130K to \$214K, the prices go as far as \$720k and the standard deviation is about \$79K. From this analysis (as well visualizations in *Fig. 2*), it is clear that the distribution would benefit from a standard logarithmic transformation to normalize the data and reduce variance. Post-transformation, the distribution exhibited a more symmetric shape, with reduced skewness, as shown in *Fig. 2*. It is important to note that in regression, predicted values must be exponentiated to return to the original scale of *SalePrice*, which allows us to interpret the predictions in dollars.

To explore the statistical associations between *log_SalePrice* and the independent variables, we used correlation analysis and bivariate regression. This enables us to distinguish variables with strong predictive potential from those with weaker or negligible relationships. We used bivariate regression models between all independent variables and *log_SalePrice* to calculate R^2 values. A threshold of 0.3 was was used to highlight variables with moderate to strong associations with *log_SalePrice*. All statically associated variables are summarized in *Table 4* and *Table 5*.

Independently of R^2 values, we also analyzed the scatterplots of numerical variables and the boxplots of categorical variables to determine which variables seem like strong predictors. The most noteworthy visualizations are detailed in *Fig. 3*. Our key findings are that area (*GrLivArea* and related variables), garage size (*GarageCars*), type (*GarageType*), and finish (*GarageFinish*), amount of bathrooms (*FullBath* and related variables), year built (*YearBuilt* and related variables), quality (*OverallQual* and other quality variables), neighborhood (*Neighborhood*), type of dwelling (*MSSubClass*), and central air (*CentralAir*) seem to be the most important independent variables. Of course, these are logically important features in determining the valuation of a property.

Table 1 - Variable Classifications

Variable Type	Variables
Ordinal (23)	LotShape, Utilities, LandSlope, OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, HeatingQC, Electrical, KitchenQual, Functional, FireplaceQu, GarageFinish, GarageQual, GarageCond, PavedDrive, PoolQC, Fence
Nominal (24)	MSSubClass, MSZoning, Street, Alley, LandContour, LotConfig, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, Foundation, Heating, CentralAir, GarageType, MiscFeature, SaleType, SaleCondition, MoSold
Discrete (13)	YearBuilt, YearRemodAdd, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageYrBlt, GarageCars, YrSold
Continuous (19)	LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, FirstFlrSF, SecondFlrSF, LowQualFinSF, GrLivArea, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, ThreeSsnPorch, ScreenPorch, PoolArea, MiscVal

Table 2 - Nulls Counts

Variable	Null Counts
LotFrontage	259
Alley	1369
MasVnrType	872
MasVnrArea	8
BsmtQual	37
BsmtCond	37
BsmtExposure	38
BsmtFinType1	37
BsmtFinType2	38
Electrical	1
FireplaceQu	690
GarageType	81
GarageYrBlt	81
GarageFinish	81
GarageQual	81
GarageCond	81
PoolQC	1453
Fence	1179
MiscFeature	1406

Table 3 - SalePrice statistics

Statistic	Value
count	1460
mean	180921.20
std	79442.50
min	34900.00
25%	129975.00
50%	163000.00
75%	214000.00
max	755000.00
IQR	84025.00
range	720100.00

Table 4 - Associations

Numerical Var	R^2
GrLivArea	0.49
GarageCars	0.46
GarageArea	0.42
TotalBsmtSF	0.37
FirstFlrSF	0.36
FullBath	0.35
YearBuilt	0.34
YearRemodAdd	0.32

Table 5 - Associations

Categorical Var	R^2
OverallQual	0.67
Neighborhood	0.56
ExterQual	0.46
BsmtQual	0.45
KitchenQual	0.45
GarageFinish	0.38
GarageType	0.33
MSSubClass	0.32
FireplaceQu	0.31
Foundation	0.30

Figure 1 - SalePrice Histogram

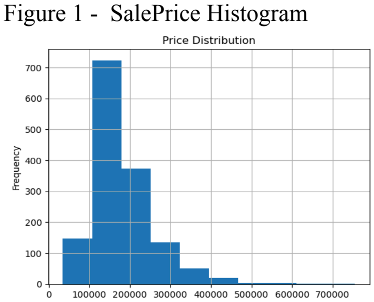


Figure 2 - Transformations on SalePrice

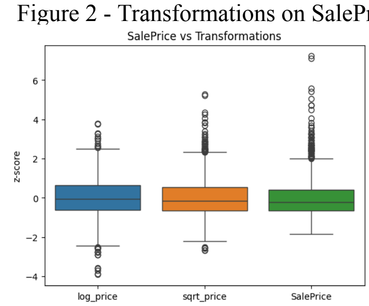
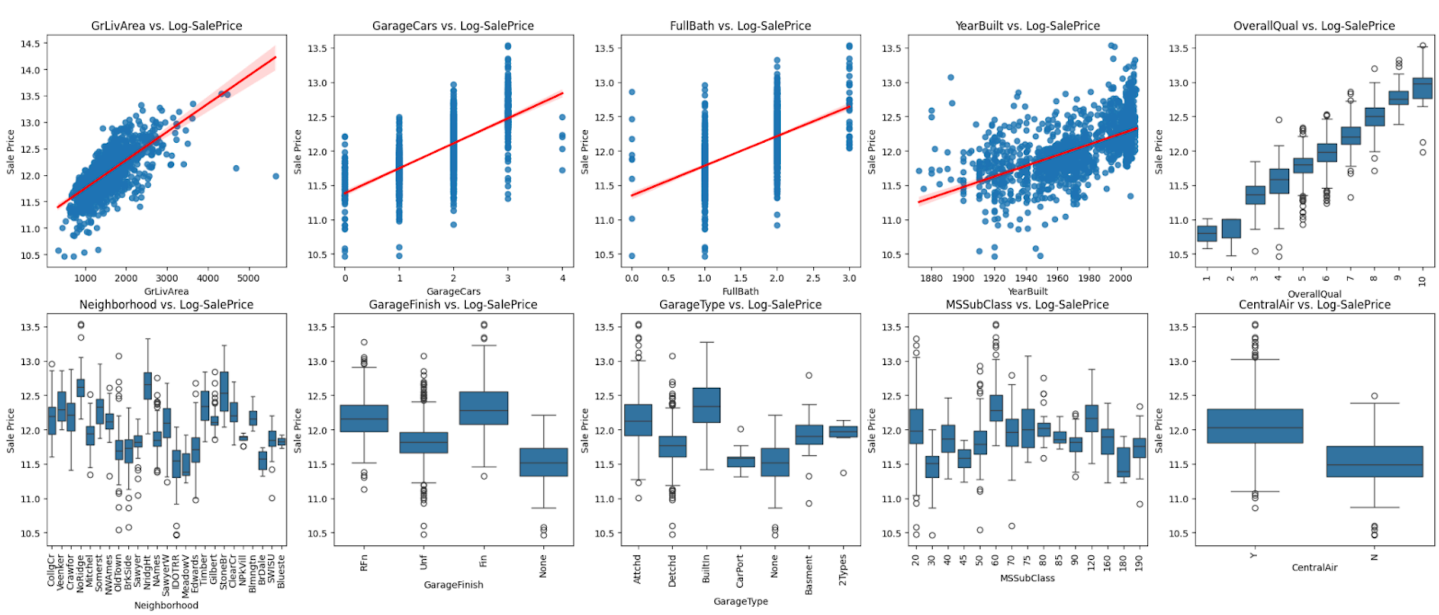


Figure 3 - Noteworthy Variables and their Distributions



II. Causal Analysis of the Effect of Number of Bedrooms on Sale Prices

This analysis seeks to estimate the causal effect of adding one bedroom on the sale price by using a multivariate regression model. We initiated our process by running a univariate regression model for reference, using *SalePrice* as the dependent variable and *BedroomAbvGr* as the independent variable, where we found a statistically significant and large relationship between the two variables. The table, labeled *Table 7*, illustrates that for every unit increase in bedrooms, the average sales price increases by $13,969.31. However, this does not take into account for any controlling variables, meaning we cannot determine if this regression alone displays a casual effect of bedrooms on sales price.

Logically, the number of bedrooms should be correlated with the size of the home and quality, since they are variables that independently influence the sale prices. To precisely control for confounding variables, we ran univariate regressions for every variable on SalePrice to identify which variables are significant to the dependent variable. We then repeated the process on BedroomAbvGr, our independent causal variable. Given the results, we concluded that most confounding variables fell into the categories of home size indicators, quality metrics, neighborhood attributes, and structural features.

The multivariate regression model, incorporating the confounding variables, reveals a more nuanced consequence of adding an additional bedroom has on SalePrice. In fact, as *Table 7* shows, the beta coefficient for BedroomAbvGr is now negative while remaining statistically significant, implying an almost-counterintuitive causal effect on sales price. After accounting for these confounding variables in the model, the interpretation of the effect of bedrooms changes. Based on our updated multivariate model, an increase in bedroom by one reduces the average sale price by $8164.12, holding all other variables constant. However, we still cannot determine the causal effect of the number of bedrooms on sales price. There could be other variables in this data set that affect both sales price and bedrooms not accounted for due to human error, as well as other factors outside of this dataset.

III. Predictions

To prepare the dataset for multivariate regression, each subgroup of independent variables required different methods. Ordinal variables were assigned a numerical scale corresponding with their inherent rank, often starting at 0 to represent the lack of a feature. Nominal variables required “dummy variables,” where each category within the nominal variable became a new binary variable representing the presence or absence of that category. Numerical variables are suitable for direct use in regression, so no recoding was necessary.

To reduce dimensionality and create more meaningful features, we aggregated certain variables. Several square footage variables were combined into a single feature, *TotalSF*, representing the total living area or space in the house. Similarly, bathroom-related variables were combined into a new feature, *TotalBaths*, representing the total number of bathrooms across the entire property. Components of these aggregated variables were not included in regression to avoid redundancy.

Outliers were addressed through visual inspection using scatterplots. We identified extreme values that disproportionately affected bivariate relationships between continuous variables and *log_SalePrice*. Data points deemed anomalous based on visual patterns were selectively removed to better train the regression model. This approach sufficiently improved model performance, although it is limited in comprehensively addressing all outliers.

In this analysis, a benchmark regression was built using *TotalSF* as the sole predictor variable, given that it is the strongest individual predictor of *log_SalePrice*. This simple regression model achieved an adjusted R-squared value of 0.75, indicating that *TotalSF* alone explains a substantial portion of the variance in house prices.

To optimize feature selection for the predictive model, we implemented a bidirectional elimination algorithm, a hybrid of forward and backward selection methods. This algorithm iteratively refines the set of predictor variables by evaluating their individual contributions to the model's performance. The process begins with the strongest predictor, *TotalSF*, as the starting point. In each iteration, the algorithm performs two steps: (1) forward selection, in which the feature that most improves the model's adjusted R-squared value is added, and (2) backward elimination, in which the least statistically significant feature (with a p-value greater than 0.05) is removed. These steps are alternated until no further improvements to the model's adjusted R-squared value can be achieved. This systematically eliminates less significant predictors.

After running the bidirectional elimination process, the list of predictors was further refined based on knowledge about relationships in the dataset as well as logical reasoning. Ultimately, this led to the selection of 30 predictors for the final model (not including “dummy” variables). This final feature set demonstrated strong predictive power with an adjusted R-squared value of 0.93, highlighting the effectiveness of the bidirectional elimination process in identifying the most relevant and impactful features for the regression model.

While TotalSF alone captured much of the variance in *log_SalePrice*, other features—such as overall quality, the number of bathrooms, neighborhood effects, and additional property attributes—provided additional predictive power. The bidirectional elimination process accounted for interaction effects between predictors, improving the model's ability to capture complex relationships in the data. The final feature selection process simplified the model while retaining predictive accuracy, ensuring it remained interpretable and efficient. These results are summarized in *Table 6*.

Table 6 - Predictive Model Results

Variable	Benchmark β	Benchmark St Err	Final Model β	Final Model St Err
Intercept	10.9450	0.0173	7.7000	0.3652
TotalSF	0.0002	0.0000	0.0002	0.0000
Alley_Pave			0.0521	0.0191
CentralAir_Y			0.0488	0.0135
Condition1_Norm			0.0427	0.0083
Exterior1st_BrkFace			0.0818	0.0158
Exterior1st_HdBoard			-0.0216	0.0084
Exterior1st_WdSdng			-0.0171	0.0089
Foundation_PConc			0.0237	0.0086
GarageType_BuiltIn			0.0313	0.0123
GarageType_CarPort			-0.0912	0.0348
HouseStyle_2Story			0.0462	0.0076
LandContour_Low			-0.0607	0.0204
LotConfig_CulDSac			0.0368	0.0116
MSSubClass_160			-0.1068	0.0168
MSSubClass_30			-0.0723	0.0145
MSZoning_FV			0.0472	0.0163
MSZoning_RM			-0.0397	0.0095
MasVnrType_Stone			0.0346	0.0110
Neighborhood_BrkSide			0.0679	0.0149
Neighborhood_ClearCr			0.0570	0.0227
Neighborhood_Crawfor			0.1493	0.0159
Neighborhood_MeadowV			-0.1250	0.0274
Neighborhood_NoRidge			0.0682	0.0184
Neighborhood_NridgHt			0.0659	0.0146
Neighborhood_StoneBr			0.1319	0.0220
Neighborhood_Veenker			0.0521	0.0317
RoofStyle_Mansard			0.0844	0.0392
SaleCondition_Alloca			0.0663	0.0344
SaleCondition_Normal			0.0606	0.0099
SaleType_New			0.1216	0.0147
BsmtCond			-0.0143	0.0057
BsmtExposure			0.0113	0.0031
BsmtFinType1			0.0087	0.0017
ExterCond			-0.0194	0.0085
Fireplaces			0.0318	0.0051
Functional			-0.0248	0.0044
GarageQual			0.0149	0.0043
HeatingQC			0.0133	0.0037
KitchenQual			0.0300	0.0062
LotArea			0.0000	0.0000
OverallCond			0.0399	0.0032
OverallQual			0.0552	0.0037
TotalBaths			0.0343	0.0057
YearBuilt			0.0014	0.0002
R-squared		0.7479		0.9319
R-squared Adj.		0.7477		0.9297

Table 7 - Causal Analysis Regression Results

Variable	Univariate β	Univariate St Err	Multivariate β	Multivariate St Err
Intercept	140281.3511	7217.7282	-720342.4944	111675.9245
BedroomAbvGr	**13969.3089**	2421.8347	**-8164.1231**	1344.0226
Neighborhood_SWISU			-3595.6061	6701.0804
MSZoning_RL			30417.6682	13662.2006
GarageType_CarPort			-23466.2813	10574.2941
Condition1_PosN			9179.0858	7841.4568
Neighborhood_NoRidge			27001.3463	5410.5543
MSSubClass_90			-2171.5151	3522.3603
HouseStyle_15Unf			9436.5668	23072.6607
HouseStyle_SFoyer			5.7820	5708.6441
MSSubClass_120			-10448.0635	4273.0331
MSSubClass_160			-44678.9040	7522.2649
Condition1_Feedr			2755.6962	4668.1464
MSZoning_RM			29524.2376	13386.6652
BldgType_Twnhs			5058.5708	6392.9912
Exterior2nd_HdBoard			-562.4989	4984.1839
GarageType_Detchd			-170.4960	2464.7193
Neighborhood_StoneBr			51134.8662	6453.8166
Heating_Grav			-7717.8481	14424.8932
MSZoning_FV			36821.2619	15765.1619
Neighborhood_BrkSide			11127.9343	4603.7400
MSSubClass_30			448.8851	4802.2098
MSZoning_RH			36065.2295	15680.4909
MSSubClass_45			5914.4412	24288.7977
Neighborhood_Somerst			105.6595	7119.9057
GarageType_BuiltIn			18771.7266	3625.3885
MasVnrType_None			-6831.4693	3492.1391
HouseStyle_2Story			27211.6740	4325.2644
MSSubClass_60			-20872.2273	5024.7047
Exterior1st_HdBoard			-5522.1932	4898.3331
MSSubClass_190			25888.5743	30860.9782
Neighborhood_MeadowV			-12370.1099	9484.6354
Exterior2nd_MetalSd			8525.1241	9998.2559
MSSubClass_180			-1318.4727	12858.2602
SaleType_New			46039.8422	17717.6116
SaleCondition_Partial			-17042.1029	17524.4005
MasVnrType_BrkFace			-16381.7974	3180.8616
Neighborhood_IDOTRR			4908.1512	6316.4789
BldgType_2fmCon			-24483.0380	30125.9530
Exterior1st_MetalSd			-6873.3857	9894.4985
Condition1_Norm			12339.0661	3267.1658
BldgType_Duplex			-2171.5151	3522.3603
Neighborhood_Veenker			23822.0309	9299.6437
MSSubClass_50			6973.7145	3504.2798
Fireplaces			7515.2681	1491.0440
LotArea			0.7312	0.1835
TotalSF			43.6043	1.5941
GarageCars			501.4745	1713.7223
MasVnrArea			52.9098	6.7347
TotalBaths			10718.4445	1577.4999
KitchenAbvGr			-26017.2731	6171.5858
YearBuilt			367.8683	56.4014
LotFrontage			97.4594	53.1633
R-squared		0.0226		0.8502
R-squared Adj.		0.0219		0.8447