ALY6015: Module 3 Multi-linear Regression

Linear regression and residue analysis



Hang Wu

Jan 30th, 2022

[Introduction 3](#_Toc1943619719)

[Assumption 3](#_Toc1622941432)

[Methodology 3](#_Toc2077540766)

[Assumption Test 3](#_Toc1372050742)

[Proof of linearity 4](#_Toc1904012267)

[Proof of Normality 4](#_Toc1527241743)

[Proof of Independence 4](#_Toc102338325)

[Technical difficulties 4](#_Toc1953004122)

[Chosen Explanatory Variables: 4](#_Toc1855484565)

[Coefficient Analysis 6](#_Toc1007670985)

[Hypothesis test for Regression Coefficient 6](#_Toc864280252)

[R2 and Adjusted R2 6](#_Toc696891852)

[F statistic as the Join Significance 6](#_Toc378544110)

[Adequacy Evaluation 6](#_Toc519968850)

[Residue Normality Test 6](#_Toc37911638)

[Analysis 10](#_Toc1222312439)

[Multicollinearity 11](#_Toc1612518333)

[Conclusion/Summary: 11](#_Toc1641357825)

[Appendix 11](#_Toc1303264173)

## Introduction

The dataset analyzed today has 2930 data points across 82 variables.

75% of the data is split into the sample and the test data (since it does not require a training set). The test data is to verify the correlation analysis.

The sample size is 2392, 75% of the original dataset, 25% are the test dataset.

In the third week, Linear Regression model with explicatory variables is built and tested along with residue analysis.

To mention that there are 8 csv files automatically built from functions using R in week 1, 10 from week 2 and 12 from week 3. These set up the statistical reasons for choosing variables

### Assumption

* all the explanatory variables correlated linearly with the response variable. (“Assumptions of linear models and what to ... - Cross Validated”) (This was the case)
* there was no collinearity among the explanatory variables. (There was little collinearity).
* The Cook's distances of the datapoints of my model are below 1 (this is the case, all distances are below 0.4, so no influence points).
* the residuals are normally distributed. (However, the y might not be normally distributed)

## Methodology

* Choose explanatory variables: variables with the 5 variables with highest correlation against SalePrice from each of the Ordinal and Numerical subset.
* Response variable: Sales Price.
* Estimate a linear regression model to predict **“Sale Price” of house** in Ames, IA as a function of **relevant explanatory variables** that you think might influence sale price. Write a brief report on the findings in terms of regression coefficient estimates (including intercept) and their statistical significance. Additionally, write remarks on the goodness of fitness statistics (such as R2, Adjusted R2, and F-statistics reported by the post estimation regression command.

### Assumption Test

Before we can perform linear regression, we need to do an assumption test:

* The data y are independently distributed, i.e., cases are independent. (“Generalized Linear Models - yanfei.site”)
* The dependent variable *￼*does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g., binomial, Poisson, multinomial, normal, etc.). (“6.1 - Introduction to GLMs | STAT 504”)
* Linearity between response variable and the explanatory variable.
* Curvilinear relationship is also fine.

Assumption 2 isn’t met as Sales Price doesn’t pass the Shapiro Normality test, but it’s not crucial.

Assumption 1 is done by a customized function that allow a chi square matrix to be built, and output the only p-value, all the field with p-value<0.05 can be eliminated due to dependency unless the dependency is with the response variable Sales Price, and p-value<0.05 with the Sales Price will be ranked based on the p-value.

### Proof of linearity

Scatter matrix performed in Appendix can visualize any linearity between the predictors and the response variable, all variables that pass the linearity test are used next.

### Proof of Normality

The explanatory variables should satisfy the normality test, good to fit test or Shapiro can be done to ensure that. This assumption shouldn't affect the prediction accuracy.

### Proof of Independence

P-value of chisqr.test(var1, var2) with H0 being that var1 and var2 are independent categorical values., if p-value<0.05, proof of dependency between var1 and var2, if not, then H0 not rejected, and we can’t prove for dependency.

Not surprisingly, monthly sold and yearly sold are the 2 columns that are mostly independent (or we can’t prove for dependency) of the other variables (Appendix)

### Technical difficulties

since recursion can’t be built in R (at least to my knowledge, and it doesn't matter it’s too slow). and for loop takes more than 5 minutes to run and my MacBook gets frozen, and there will be an out of memory error popped while trying to perform Chi Square P-Value Matrix on a large scale of data, therefore a frequency table is not setup for every variable and perform the chi-squared good to fit test. However, they can be done, and the p-values will be recorded in a table and saved as a csv file.

So, it’s likely that p-value matrix will run separately on numeric attributes and ordinal attributes.

## Chosen Explanatory Variables:

They are chosen with R-value > 0.5 and being in top 5 rank (in terms of correlation against SalePrice), and they satisfy the 3 assumptions above. Although other categorical values also influence the sales price, but they are not considered if they are nominal attributes. Only ordinal (match to specified levels and converted to int) and numeric data will be considered.

Numeric Attributes considered:

Overall.qual

gr.liv.area

Garage.cars

Garage.Area

Ordinal Attributes considered:

pool.QC

Exter.Qual

Kitchen.Qual

Bsmt.Qual

Garage.Finish

We further use intuition to find the more interesting attributes we want to know about, this includes the Overall quality, the above grade living area, and the garage area (because garage cars is limited by a low upper bound, the spread is not good enough for studying residue) for the numeric attributes,

And the ordinal attributes we shall use **Garage Finish Quality and Pool Quality** because garage seems to be deeply correlated with the response variable, and we might want to draw that conclusion (using garage alone we can predict the saleprice), and the other 3 qualities are interchangeable. Plot 2b demonstrates the deep correlation between these variables, and if we sort the exterior quality in descending order in the correlation matrix (saved as correlationmatrixRR.csv), the top 5 columns are

|  |
| --- |
| Exter.Qual |
| Kitchen.Qual |
| SalePrice |
| Bsmt.Qual |
| Garage.Finish |

Therefore, proving that we can replace the rest of the quality with the Garage Finish. Also, these variables will cause the **Multicollinearity problems once they are included in the multiple linear regression model.**

Now we have 5 attributes to consider, and we shall build a table to list their intercept, R, R adjust, F-statistic and p-value

Reminder: The SalePrice (y value) standard deviation is 79886, meanwhile all the ordinal attributes have standard deviation closed to 1, and the largest numeric attributes’ Sd is the ground living are which is 47, therefore the beta coefficient (slope) is expected to be large for row 1,4 and 5.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Intercept (expected mean of y when x=0)** | **Beta Coefficient(slope)** | **R2** | **Adjusted R2** | **F Statistic** | **P-value coefficient** | **P-value Intercept** | **P, n = 300, intercept** |
| Overall Qual | -103649.8 | 46901.1 | 0.6492 | 0.649 | 4423 on 1 and 2390 DF | 0.0000 | 0.00 | 0.00 |
| Living | 9700.625 | 115.408 | 0.51 | 0.5098 | 2487 on 1 and 2390 DF | 0.0000 | 0.00 | 0.0704 |
| Garage Size | 64681.217 | 251.393 | 0.434 | 0.4337 | 1651 on 1 and 2390 DF | 0.0000 | 0.00 | 0.00 |
| Pool Quality | -330568 | 113389 | 0.4689 | 0.4149 | 8.829 on 1 and 10 DF | 0.0140 | 0.1480 | 0.00 |
| Garage Finish | 48800 | 29732 | 0.2725 | 0.2722 | 717.1 on 1 and 1914 DF | 0.0000 | 0.00 | 0.00 |
|  |  |  |  |  |  |  |  |  |

Please notice that there are only 10 out of the 2392 samples (sample is 75% of the population) has pool, and 135 of them doesn’t have garage. Ones which don’t have garage isn’t shown in the garage size variable either, which makes it more interesting as garage size in the scatter matrix does show a linear relationship. Although the lowest intercept should be 0, as price can’t go lower than that, the LM model still supplies interesting insights. The qualitative predictors all share negative intercept, Garage Finish has the least impact while Overall Qual has the strongest impact. Since both living area (above ground) and garage size are continuous variables that have larger mean and std, they are expected to have less coefficient.

### Coefficient Analysis

larger coefficients are expected for ordinal attributes, and they would be a lot smaller if Sales Price is categorized as percentiles. However, that diminishes the comparison of coefficients. Since coefficients function as the standard deviation or variance, one can use that to decide the volatility of the predictors. Pool Quality has the largest volatility meanwhile living area has the least.

### Hypothesis test for Regression Coefficient

* “Null hypothesis (H0): the coefficients are equal to zero (i.e., no relationship between x and y)” (“Simple Linear Regression in R - Articles - STHDA”)
* “Alternative Hypothesis (Ha): the coefficients are not equal to zero (i.e., there is some relationship between x and y)” (“Simple Linear Regression in R - Articles - STHDA”)

H0 is rejected for all beta coefficients of all variables, showing that it’s statistically significant that sale Price and the predictors have relationship. Pool quality doesn't reject H0 for its intercept, while others’ intercept rejects H0 as well, shows that their intercept is statistically significant enough to relate to the sale price.

### R2 and Adjusted R2

If we use adjusted R2 to form the rank of all predictors, overall quality score is still the best predictor (R2 adjust = 0.65) as it has the best model fit, living area above ground has the second highest (R2 adjust=0.51).

### F statistic as the Join Significance

Since the larger the F stat the better the significance, Overall Qual (ordinal) and living area(continuous) still have the best statistical significance.

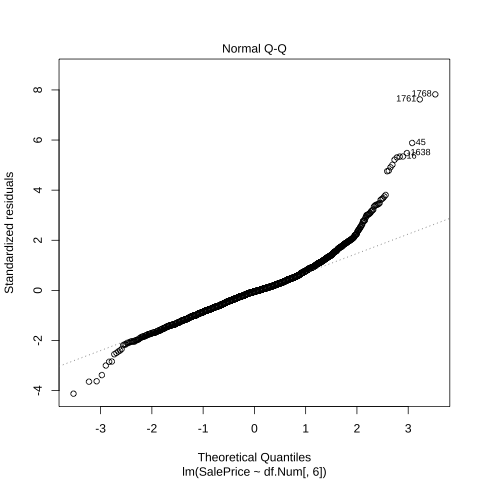
However, volatility wise (coefficient) the living area is 99.75% less than the overall quality, besides, OC still requires a person to personally rate the house, meanwhile living area is decided once the house is designed. This makes the ground level living area the best predictor of the price.

## Adequacy Evaluation

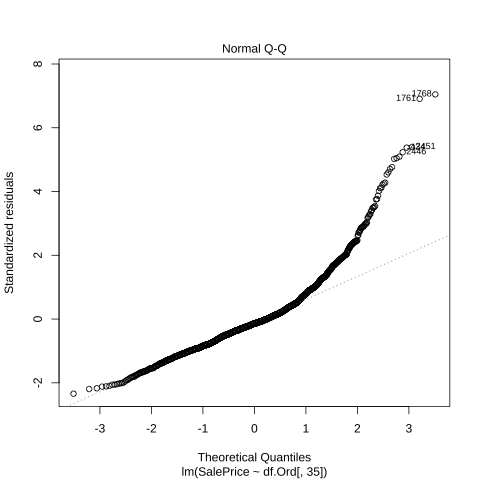
Since the highest t-stat and the highest R2 all referenced to the Overall Quality plot, we will plot the second highest R2 (aka, Living Area) as the second plot for comparison.

### Residue Normality Test

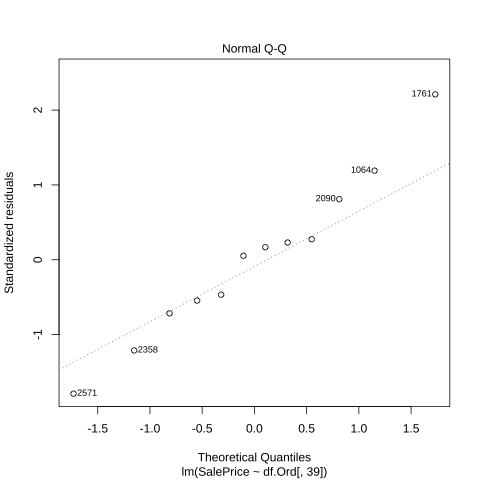
QQ Plot for Garage Finish



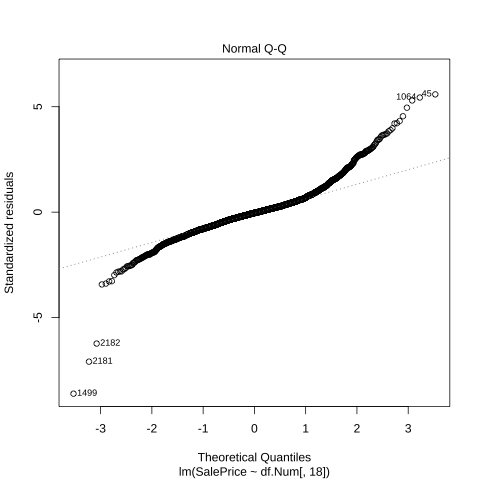
QQ Plot for Overall Quality



QQ Plot for Pool Quality (NA omitted)



QQ Plot for Living Area



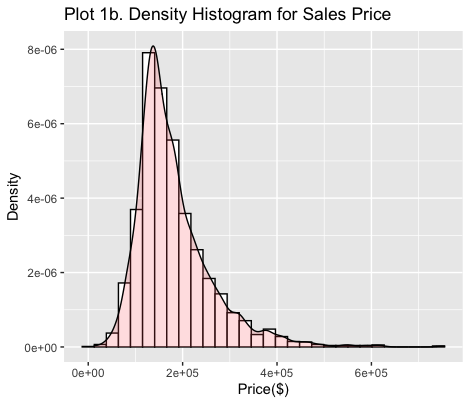
## Analysis

In plot 1b we indicate the **right skewness** of the Sales Price by itself. Not surprisingly the QQ plot with the residue of the Linear Regression model of Sales Price against explanatory variables also shows **right skewnes**s. This is mostly indicated by the Garage Finish and the Overall Quality QQ plot. It also makes sense as the residues are the differences between predicted and observed values, if the dependent variable is right skewed, then the residues would be right skewed as well against its linear prediction.

Pool Qual will also be omitted in future analysis as predictors because of the lack of data.

All plots show different magnitudes of **over-dispersion. Garage Finish** shows the most dispersion and it’s expected to have **positive excess kurtosis (kurtosis>3),** and it’s confirmed that it has a kurtosis of **8.60** and a right skew of **1.64. Extreme outliers are expected**

**QQ plot of Living Area also shows the 5 extreme outliers shown in the data description.**



### Multicollinearity

To deal with the multicollinear problem we must use VIF, tolerance and alias. Where VIF and tolerance are used to measure multicollinearity, where tolerance > 0.25 and VIF < 5 should suffice, otherwise collinearity exists. Alias is used to measure dependency of the explanatory variables, where only full alias is marked. Red highlighted columns (see Appendix) are the explanatory variables that don’t meet the VIF and tolerance requirements. Further investigation on alias shows that full multicollinearity occurred between Total.Bsmt.SF and BsmtFin.SF.1 BsmtFin.SF.2 Bsmt.Unf.SF, as well as Gr.Liv.Area and X1st.Flr.SF X2nd.Flr.SF.Since we would still use gr.liv.area, we will drop the columns that are alias to it. This does favor Overall Quality as one of the primary predictors.

### Conclusion/Summary:

As we are moving into the Multi Linear Regression territory, more should be considered, normality of residues, residues independency, collinearity between independent variables, coefficient(slope), R2 and variance of the residues. Overfit also needs to be considered while evaluating three performances of the model. Right now, we will settle with a 5-variable model with

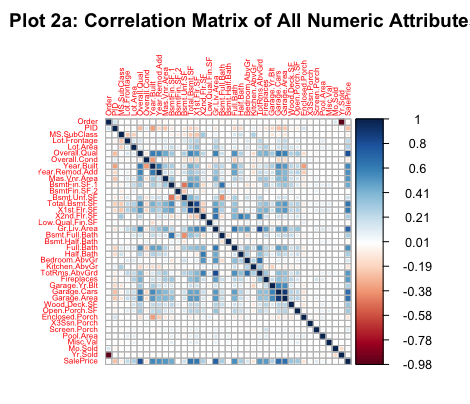
## Appendix

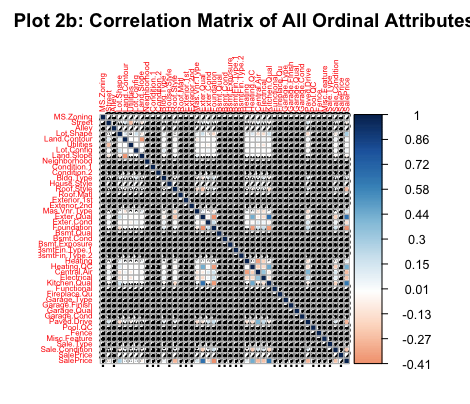
**Table for chi-square Proof of Independency Test: P-Value**

|  |  |  |
| --- | --- | --- |
| Variables | **Mo.Sold** | **Yr.Sold** |
| Exterior.1st | NA | NA |
| Exterior.2nd | NA | NA |
| Electrical | NA | NA |
| Sale.Type | NA | NA |
| SalePrice | NA | NA |
| Overall.Qual | 0.007793521 | 0.378645882 |
| Year.Built | 0.083565096 | 5.39E-15 |
| Year.Remod.Add | 0.005786749 | 1.34E-33 |
| Roof.Matl | 0.488711909 | 0.646997 |
| BsmtFin.SF.2 | 0.881437215 | 0.505239359 |
| Low.Qual.Fin.SF | 0.494725136 | 0.551433848 |
| Kitchen.AbvGr | 0.297455918 | 0.263050898 |
| Garage.Yr.Blt | 0.142451928 | 2.47E-18 |
| Enclosed.Porch | 0.978988121 | 0.571948688 |
| Screen.Porch | 0.592686244 | 0.705452271 |
| Pool.Area | 0.029445233 | 0.688420276 |
| Pool.QC | 0.083793666 | 0.086077676 |
| Fence | 0.891365266 | 0.69018543 |
| Misc.Val | 0.874953392 | 0.536933158 |
| Roof.Style | 0.319271594 | 0.611881557 |
| BsmtFin.Type.2 | 0.601739049 | 0.459339237 |
| Condition.1 | 0.446246141 | 0.334653689 |
| X3Ssn.Porch | 0.002432402 | 0.569905441 |
| BsmtFin.SF.1 | 0.284099675 | 0.448561501 |
| House.Style | 0.470637464 | 0.870498909 |
| Bedroom.AbvGr | 0.194850472 | 0.240504186 |
| Bsmt.Half.Bath | 0.100252847 | 0.307329799 |
| Fireplaces | 0.180174203 | 0.044051135 |
| Bldg.Type | 0.417435117 | 0.151768966 |
| BsmtFin.Type.1 | 0.241839385 | 0.346823091 |
| Alley | 0.304581995 | 0.225265144 |
| Paved.Drive | 0.787863114 | 0.834956392 |
| Land.Slope | 0.087859246 | 0.877155642 |
| Bsmt.Full.Bath | 0.798339662 | 0.530833696 |
| Lot.Config | 0.113333732 | 0.7155294 |
| Order | 0.479718007 | 0.486543071 |
| PID | 0.479718007 | 0.486543071 |
| Misc.Feature | 0.859075232 | 0.392691273 |
| Mo.Sold | NA | 1.58E-35 |
| Yr.Sold | NA | NA |
| Land.Contour | 0.248194888 | 0.496477025 |
| MS.SubClass | 0.708072179 | 0.266459403 |
| Mas.Vnr.Type | 0.133220174 | 0.204300564 |
| Bsmt.Cond | 0.925951636 | 0.059058415 |
| Neighborhood | 0.032427444 | 0.248867649 |
| Lot.Shape | 0.758316309 | 0.616305148 |
| Heating.QC | 0.039172173 | 0.529420243 |
| Exter.Cond | 0.03676758 | 0.05334271 |
| Half.Bath | 0.172873002 | 0.340921704 |
| Garage.Type | 0.190656429 | 0.485399628 |
| Foundation | 0.599090916 | 0.077156039 |
| Central.Air | 0.433437342 | 0.701230905 |
| Wood.Deck.SF | 0.938120956 | 0.025181007 |
| Fireplace.Qu | 0.866641143 | 0.737649575 |
| Garage.Cond | 0.031656515 | 0.693162484 |
| Heating | 0.542119569 | 0.320742097 |
| Utilities | 0.066252053 | 0.536812573 |
| Overall.Cond | 0.077649304 | 0.000752045 |
| Garage.Finish | 0.366170701 | 0.687830281 |
| Bsmt.Exposure | 0.323821356 | 0.122798889 |
| Kitchen.Qual | 0.009161799 | 0.17791284 |
| Sale.Condition | NA | NA |
| Street | 0.793498154 | 0.388837737 |
| Condition.2 | 0.555875108 | 0.336007348 |
| Bsmt.Unf.SF | 0.055972402 | 0.130342585 |
| Garage.Qual | 0.338992542 | 0.202707669 |
| Garage.Cars | 0.011885336 | 0.361502865 |
| Functional | 0.059040366 | 0.536857449 |
| Bsmt.Qual | 0.362566637 | 0.23070424 |
| TotRms.AbvGrd | 0.001237069 | 0.69465598 |
| Full.Bath | 0.000500372 | 0.077354043 |
| Exter.Qual | 0.131709299 | 0.189715478 |
| MS.Zoning | 0.090274985 | 0.129521789 |
| Lot.Frontage | 0.134668433 | 0.082096964 |
| Lot.Area | 0.110635838 | 0.230387904 |
| X1st.Flr.SF | 0.093953788 | 0.149351861 |
| X2nd.Flr.SF | 0.674104482 | 0.23586659 |
| Mas.Vnr.Area | 0.023894899 | 0.685665334 |
| Total.Bsmt.SF | 0.014595286 | 0.149517934 |
| Gr.Liv.Area | 0.384446698 | 0.615621342 |
| Garage.Area | 0.339995762 | 0.788625921 |
| Open.Porch.SF | 0.023932135 | 0.968855442 |

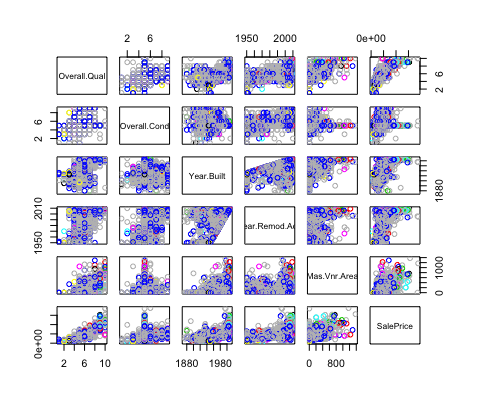
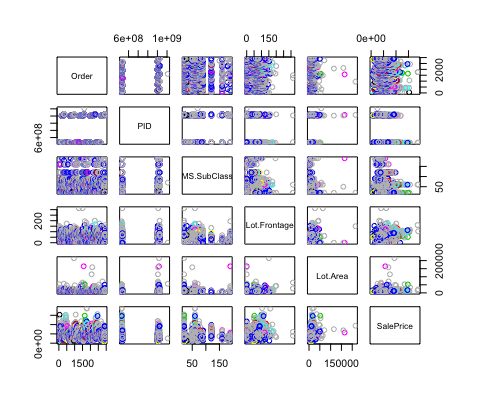
**Table of Tolerance and VIF for all predictors**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Tolerance** | **VIF** | **Variables** | **Tolerance** | **VIF** |
| BsmtFin.SF.1 | 0 | Inf | Fence | 0.094398188 | 10.59342364 |
| BsmtFin.SF.2 | 0 | Inf | MS.Zoning | 0.094601087 | 10.57070306 |
| Bsmt.Unf.SF | 0 | Inf | Garage.Finish | 0.315191566 | 3.172673725 |
| Total.Bsmt.SF | 0 | Inf | Exterior.1st | 0.380665552 | 2.626977918 |
| X1st.Flr.SF | 0 | Inf | Exterior.2nd | 0.38438779 | 2.601539449 |
| X2nd.Flr.SF | 0 | Inf | Exter.Qual | 0.423807507 | 2.359561793 |
| Low.Qual.Fin.SF | 0 | Inf | Garage.Cond | 0.452846058 | 2.208255945 |
| **Gr.Liv.Area** | **0** | **Inf** | Bsmt.Qual | 0.460176334 | 2.173080028 |
| Order | 0.012494809 | 80.03323333 | Garage.Type | 0.466051973 | 2.145683437 |
| Yr.Sold | 0.012789795 | 78.18733733 | Kitchen.Qual | 0.480382257 | 2.081675551 |
| Garage.Cars | 0.150031042 | 6.665287293 | Garage.Qual | 0.487197918 | 2.052553929 |
| Garage.Area | 0.172914864 | 5.783192821 | Foundation | 0.624125718 | 1.602241298 |
| TotRms.AbvGrd | 0.220162702 | 4.542095417 | Central.Air | 0.63017729 | 1.586855025 |
| Year.Built | 0.24818974 | 4.029175415 | Heating.QC | 0.666299921 | 1.500825632 |
| PID | 0.256962078 | 3.891624813 | Bsmt.Exposure | 0.670831077 | 1.490688243 |
| Overall.Qual | 0.292488156 | 3.418941858 | Paved.Drive | 0.67262397 | 1.486714784 |
| Full.Bath | 0.342962038 | 2.915774603 | Fireplace.Qu | 0.756538155 | 1.321810399 |
| Year.Remod.Add | 0.413694148 | 2.417244735 | Land.Slope | 0.780599911 | 1.281065993 |
| Bedroom.AbvGr | 0.424453948 | 2.355968193 | BsmtFin.Type.1 | 0.796391889 | 1.255663215 |
| Half.Bath | 0.449501731 | 2.224685538 | Land.Contour | 0.804319123 | 1.243287611 |
| Bsmt.Full.Bath | 0.449765683 | 2.223379947 | Electrical | 0.805891478 | 1.240861863 |
| Garage.Yr.Blt | 0.549921683 | 1.818440755 | Bsmt.Cond | 0.820936509 | 1.218120999 |
| Kitchen.AbvGr | 0.61434589 | 1.627747522 | Lot.Shape | 0.831839841 | 1.20215449 |
| Fireplaces | 0.625738569 | 1.598111495 | Heating | 0.837594156 | 1.193895626 |
| Overall.Cond | 0.645278476 | 1.549718513 | Alley | 0.838481846 | 1.192631665 |
| MS.SubClass | 0.662611875 | 1.509179111 | Bldg.Type | 0.852449041 | 1.173090651 |
| Mas.Vnr.Area | 0.667389341 | 1.498375743 | Sale.Condition | 0.857936019 | 1.165588083 |
| Lot.Area | 0.798548083 | 1.252272744 | House.Style | 0.864640681 | 1.156549792 |
| Enclosed.Porch | 0.806309883 | 1.240217962 | Roof.Style | 0.868754362 | 1.151073357 |
| Wood.Deck.SF | 0.809716497 | 1.235000156 | Street | 0.869232869 | 1.150439699 |
| Open.Porch.SF | 0.823009709 | 1.215052494 | Neighborhood | 0.870988721 | 1.148120493 |
| Lot.Frontage | 0.836172408 | 1.195925613 | Mas.Vnr.Type | 0.88585036 | 1.128858829 |
| Bsmt.Half.Bath | 0.866476307 | 1.154099646 | Utilities | 0.886661004 | 1.127826752 |
| Pool.Area | 0.911115959 | 1.097555136 | Exter.Cond | 0.893237584 | 1.119522977 |
| Screen.Porch | 0.919275706 | 1.087812931 | BsmtFin.Type.2 | 0.895764673 | 1.116364633 |
| Misc.Val | 0.942496862 | 1.06101149 | Sale.Type | 0.914207855 | 1.093843151 |
| Mo.Sold | 0.943590209 | 1.059782086 | Condition.1 | 0.921088746 | 1.085671716 |
| X3Ssn.Porch | 0.977741244 | 1.022765487 | Functional | 0.929432849 | 1.07592496 |
|  | | | Lot.Config | 0.931741155 | 1.073259451 |
| Roof.Matl | 0.950616979 | 1.05194839 |
| Pool.QC | 0.961820541 | 1.039694992 |
| Misc.Feature | 0.966335109 | 1.034837699 |
| Condition.2 | 0.966704718 | 1.03444204 |

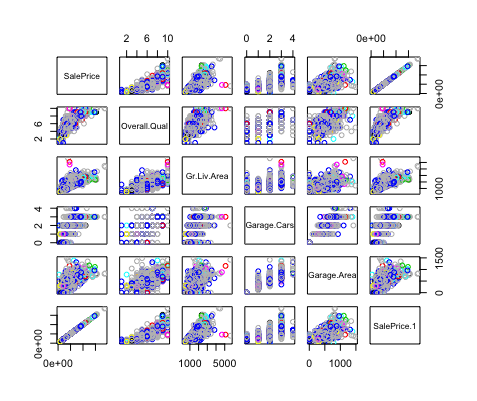
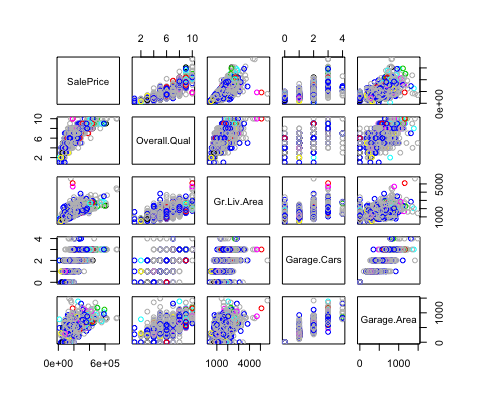




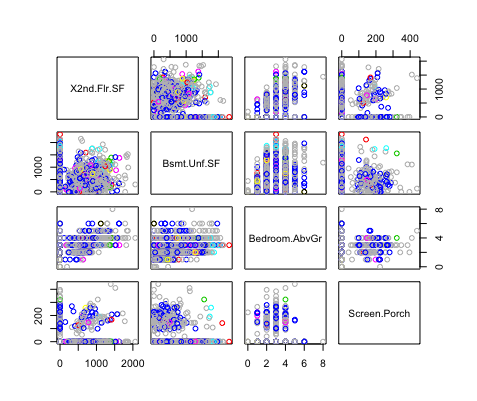
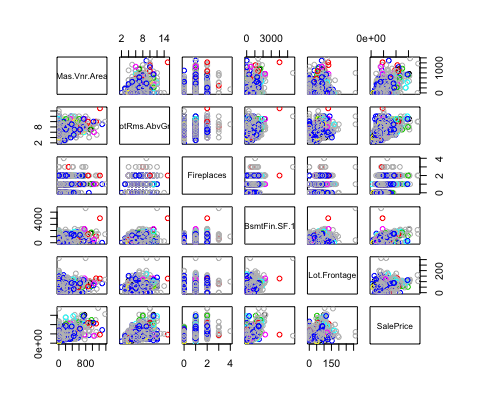
Scatter Matrix for Numeric Value against Sales

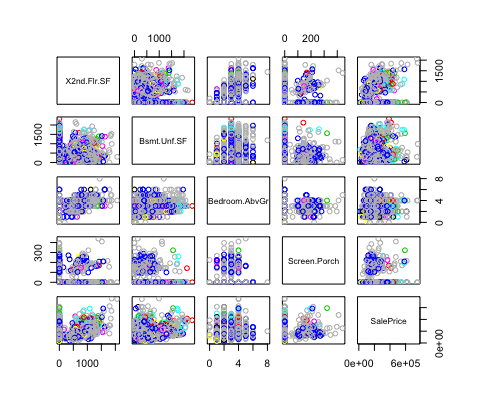
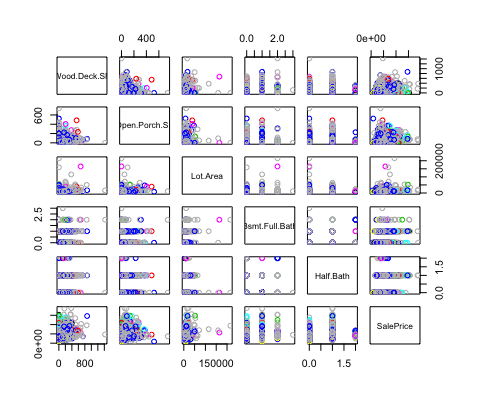


Scatter Matrix for Numeric Value Against SalePrice

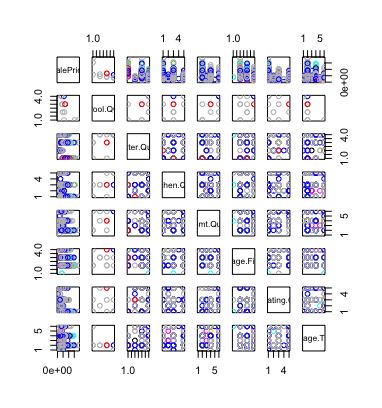


Scatter Matrix for Ordinal Value (convert to levels) against SalePrice





39 records are extracted from the correlation matrix for further analysis.



For ordinal attributes (that needed conversion), the top 5 are Pool.QC, Exter Qual, Kitchen Qual and Basement Qual are among the best r value against SalePrice, p value<0.05 which rejects H0 properly. The red highlighted in the table have p-value > 0.05, therefore not statistically significant enough to draw a conclusion.

**Table of Ordinal Attribute against SalePrice**

|  |  |  |
| --- | --- | --- |
| **Ordinal Attributes** | **SalePrice r value** | **SalePrice P Value** |
| **Pool.QC** | 0.684766241 | 0.014013045 |
| **Exter.Qual** | 0.659936504 | 0 |
| **Kitchen.Qual** | 0.629060687 | 0 |
| **Bsmt.Qual** | 0.616658872 | 0 |
| **Garage.Finish** | 0.519069447 | 0 |
| **Heating** | 0.431876883 | 0 |
| **Garage.Type** | 0.363983198 | 0 |
| **Lot.Shape** | 0.284207172 | 0 |
| **Fireplace.Qu** | 0.195910887 | 4.01E-12 |
| **Fence** | 0.172508893 | 0.000224966 |
| **Mas.Vnr.Type** | 0.138424949 | 1.05E-11 |
| **BsmtFin.Type.1** | 0.106338693 | 2.76E-07 |
| **MS.Zoning** | 0.103869574 | 0.027062624 |
| **Heating** | 0.097138072 | 1.94E-06 |
| **Bldg.Type** | 0.057331709 | 0.005034389 |
| **Functional** | 0.031354189 | 0.698540065 |
| **Misc.Feature** | -0.026909029 | 0.813895909 |
| **Roof.Matl** | -0.032849075 | 0.108752766 |
| **BsmtFin.Type.2** | -0.03338078 | 0.107584476 |
| **Condition.2** | -0.091247793 | 7.90E-06 |

Quality, Gr Liv Area, Garage Cars and Garage Area have the most r value against SalePrice.

**Table of Numeric Attributes Correlation against SalePrice**

|  |  |  |  |
| --- | --- | --- | --- |
| **Numeric Attributes** | **SalePrice r value** | **SalePrice P Value** | **Note** |
| **Pool.QC** | 1 | -0.529414302 | NA |
| **SalePrice** | 1 | NA |  |
| **Overall.Qual** | 0.803944231 | 0 |  |
| **Gr.Liv.Area** | 0.714170744 | 0 |  |
|  |  |  |  |
| **Garage.Cars** | 0.661287719 | 0 | Garage capacity is related to house prices as it attracts family-based home buyer, owner of multiple cars also shows buying power |
| **Garage.Area** | 0.647847134 | 0 |  |
| **Total.Bsmt.SF** | 0.645714746 | 0 |  |
| **X1st.Flr.SF** | 0.638761685 | 0 |  |
| **Kitchen.Qual** | 0.634219707 | 0.077949664 | 0.124722272 |
| **Full.Bath** | 0.559935411 | 0 | Importance: Garage->1st floor->Bath->Lot,deck porch->2nd floor |
| **Year.Built** | 0.559622315 | 0 | recent home shows better pricing |
| **Garage.Yr.Blt** | 0.541859948 | 0 |  |
| **Year.Remod.Add** | 0.538239673 | 0 |  |
| **Mas.Vnr.Area** | 0.527794157 | 0 |  |
| **TotRms.AbvGrd** | 0.52327295 | 0 |  |
| **BsmtFin.SF.1** | 0.435782879 | 0 |  |

**Bibliography**

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Chopik, W. J., & Lucas, R. E. (2019). Actor, partner, and similarity effects of personality on global and experienced well-being. *Journal of Research in Personality*, *78*, 249–261. <https://doi.org/10.1016/j.jrp.2018.12.008>

*Bluman, A. G. (2009). Elementary statistics: A step by step approach*. New York: McGraw- Hill Higher Education.

*John Blischak, Daniel Chen, Harriet Dashnow, and Denis Haine (eds): "Software"*  *Carpentry: Programming with R*." Version 2016.06, June 2016, <https://github.com/swcarpentry/r-> novice-inflammation, 10.5281/zenodo.57541.

*Kabacoff, R. (2015). R in action: Data analysis and graphics with R*. <http://www.cs.uni.edu/~jacobson/4772/week11/R_in_Action.pdf>

*Venables, W.N., Smith, D.M, R Core Team (2020) An Introduction to R*: A Programming Environment for Data Analysis and Graphics Version 4.0.3 <https://cran.rproject.org/doc/manuals/r-release/R-intro.pdf>

*R errors explained: Incomplete final line found by rattleheaded*. Programming. (n.d.). Retrieved October 2, 2021, from [https://www.programmingr.com/r-errormessages/incomplete-finalline-found-by-readtableheader/.](https://www.programmingr.com/r-error-messages/incomplete-final-line-found-by-readtableheader/)

Dolenc. (2021, January 25). *Bulltroutrml2: Ages and lengths of Bull Trout from two Rocky Mountain Lakes... in draglink/Sadaat: Data to support Fish Stock Assessment ('FSA') package*. BullTroutRML2: Ages and lengths of Bull Trout from two Rocky Mountain lakes... in draglink/Sadaat: Data to Support Fish Stock Assessment ('FSA') Package. Retrieved October 6, 2021, from <https://rdrr.io/github/droglenc/FSAdata/man/BullTroutRML2.html>.

Dugar, D. (2020, July 18). *Skew and kurtosis: 2 important statistics terms you need to know in Data Science*. Medium. Retrieved October 6, 2021, from <https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa>.

*Skew and Kurtosis: 2 Important Statistics terms you need ...*. (n.d.). Retrieved from <https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa>

*Applied Time Series and forecasting*. Andrea Peralto. (n.d.). Retrieved January 28, 2022, from <https://www.andreaperlato.com/tspost/applied-time-series-and-forecasting/>

*Are tests of normality basically useless? - ResearchGate*. (n.d.). Retrieved January 30, 2022, from <https://www.researchgate.net/post/Are_tests_of_normality_basically_useless>

Heckman, E. (n.d.). *Correlation: What it shows you (and what it doesn't)*. Minitab Blog. Retrieved January 30, 2022, from <https://blog.minitab.com/en/starting-out-with-statistical-software/correlation-what-it-shows-you-and-what-it-doesnt#:~:text=Correlation%20can%20tell%20if%20two,the%20strength%20of%20that%20relationship>.

Radeck, D. (2021, July 15). *Time series from scratch - decomposing Time Series Data*. Medium. Retrieved January 28, 2022, from <https://towardsdatascience.com/time-series-from-scratch-decomposing-time-series-data-7b7ad0c30fe7>

These worksheets are part of the modules BIOL30030, B. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. I. O. L. and E. N. V. B. (n.d.). Quantile-quantile plots. Retrieved January 31, 2022, from <https://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html#left-skewed-data>

Collinearity Diagnostics, Model Fit & Variable Contribution. (n.d.). Retrieved January 31, 2022, from <https://cran.r-project.org/web/packages/olsrr/vignettes/regression_diagnostics.html>

Kassam bara, soya, R., Vivid diagnostics, Eva, Visitor, & Mann, T. (2018, March 11). *Linear regression assumptions and diagnostics in R: Essentials*. STHDA. Retrieved January 31, 2022, from <http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/>

*Qualtrics*. Qualtrics XM. (2021, April 12). Retrieved January 31, 2022, from <https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/>

*Assumptions of linear models and what to ... - Cross Validated*. (n.d.). Retrieved from <https://stats.stackexchange.com/questions/100214/assumptions-of-linear-models-and-what-to-do-if-the-residuals-are-not-normally-di>