# MSDS 6372 Project 2
# House Prices: Advanced Regression Techniques

Evangelos Giakoumakis

**Introduction:**

The Boston Housing Data Set is a collection of data from approximately 500 US census reports in the Boston area. In order to attempt to predict home sale prices in various neighborhoods, multiple linear regression will be used determine which home features are related to sale price and whether the square footage of the home is related to the sale price.

The data used for this analysis is the Ames Housing dataset, compiled by Dean De Cock as an alternative to, and subset of, the standard Boston Housing Data Set. The Ames Housing dataset includes only residential sales from Ames between 2006 and 2010. Information collected about these residential sales includes dimensions of various parts of the home, lot size and dwelling square footage, and quantifying variables about the number of specific rooms in the homes.

The Century 21 office in Ames, IA, would like to determine whether the final sale price of homes in their sales area is related to the size of the living area of the home. To analyze the information, home sales in all neighborhoods of interest from 2006 to 2010 were gathered and various models where created.

**Data Description:**

Below is a list of all variables present in the dataset with a brief description of each. We will be using 2 datasets (a training and a testing) that hold 1460 and 1459 observations respectively. There is a total of 81 variables with 43 categorical and 38 quantitative. Our dependent variable is SalePrice.

**MSSubClass**: Identifies the type of dwelling involved in the sale.
**MSZoning**: Identifies the general zoning classification of the sale.
**LotFrontage**: Linear feet of street connected to property
**LotArea**: Lot size in square feet
**Street**: Type of road access to property
**Alley**: Type of alley access to property
**LotShape**: General shape of property
**LandContour**: Flatness of the property
**Utilities**: Type of utilities available
**LotConfig**: Lot configuration
**LandSlope**: Slope of property
**Neighborhood**: Physical locations within Ames city limits
**Condition1**: Proximity to various conditions

**Condition2**: Proximity to various conditions (if more than one is present)
**BldgType**: Type of dwelling
**HouseStyle**: Style of dwelling
**OverallQual**: Rates the overall material and finish of the house
**OverallCond**: Rates the overall condition of the house
**YearBuilt**: Original construction date
**YearRemodAdd**: Remodel date (same as construction date if no remodeling or additions)
**RoofStyle**: Type of roof
**RoofMatl**: Roof material
**Exterior1st**: Exterior covering on house
**Exterior2nd**: Exterior covering on house (if more than one material)
**MasVnrType**: Masonry veneer type
**MasVnrArea**: Masonry veneer area in square feet
**ExterQual**: Evaluates the quality of the material on the exterior
**ExterCond**: Evaluates the present condition of the material on the exterior
**Foundation**: Type of foundation
**BsmtQual**: Evaluates the height of the basement
**BsmtCond**: Evaluates the general condition of the basement
**BsmtExposure**: Refers to walkout or garden level walls
**BsmtFinType1**: Rating of basement finished area
**BsmtFinSF1**: Type 1 finished square feet
**BsmtFinType2**: Rating of basement finished area (if multiple types)
**BsmtFinSF2**: Type 2 finished square feet
**BsmtUnfSF**: Unfinished square feet of basement area
**TotalBsmtSF**: Total square feet of basement area
**Heating**: Type of heating
**HeatingQC**: Heating quality and condition
**CentralAir**: Central air conditioning
**Electrical**: Electrical system
**1stFlrSF**: First Floor square feet
**2ndFlrSF**: Second floor square feet
**LowQualFinSF**: Low quality finished square feet (all floors)
**GrLivArea**: Above grade (ground) living area square feet
**BsmtFullBath**: Basement full bathrooms
**BsmtHalfBath**: Basement half bathrooms
**FullBath**: Full bathrooms above grade
**HalfBath**: Half baths above grade
**Bedroom**: Bedrooms above grade (does NOT include basement bedrooms)
**Kitchen**: Kitchens above grade
**KitchenQual**: Kitchen quality
**TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
**Functional**: Home functionality (Assume typical unless deductions are warranted)
**Fireplaces**: Number of fireplaces
**FireplaceQu**: Fireplace quality

**GarageType**: Garage location
**GarageYrBlt**: Year garage was built
**GarageFinish**: Interior finish of the garage
**GarageCars**: Size of garage in car capacity
**GarageArea**: Size of garage in square feet
**GarageQual**: Garage quality
**GarageCond**: Garage condition
**PavedDrive**: Paved driveway
**WoodDeckSF**: Wood deck area in square feet
**OpenPorchSF**: Open porch area in square feet
**EnclosedPorch**: Enclosed porch area in square feet
**3SsnPorch**: Three season porch area in square feet
**ScreenPorch**: Screen porch area in square feet
**PoolArea**: Pool area in square feet
**PoolQC**: Pool quality
**Fence**: Fence quality
**MiscFeature**: Miscellaneous feature not covered in other categories
**MiscVal**: $Value of miscellaneous feature
**MoSold**: Month Sold (MM)
**YrSold**: Year Sold (YYYY)
**SaleType**: Type of sale
**SaleCondition**: Condition of sale
**SalePrice**: Final price the unit was sold

**<u>Exploratory Data Analysis:</u>**

Initially data was imported in SAS workspace and a visual check was performed. Nothing stood out so continued with missing value exploration. The following table shows missing values detected and applied solution.

| Variable Name | Missing Values | Solution |
|---|---|---|
| MasVnrArea | 23 | Impute with zero |
| BsmtFinSF1 | 1 | Impute with zero |
| BsmtFinSF2 | 1 | Impute with zero |
| BsmtUnfSF | 1 | Impute with zero |
| TotalBsmtSF | 1 | Impute with zero |
| BsmtFullBath | 2 | Impute with zero |
| BsmtHalfBath | 2 | Impute with zero |
| GarageYrBlt | 159 | Impute with (Min-1) |
| GarageCars | 1 | Impute with zero |
| GarageArea | 1 | Impute with zero |

After investigating it was decided to drop the following variables since they did not contribute greatly to our predictive model: BsmtFinSF2, LowQualFinSF, BsmtHalfBath, 3SnPorch, ScreenPorch, MiscVal. Before creating our model, all assumptions for multiple linear regression must be checked:

- Multivariate Normality (normal distribution of residuals)
- No Multicollinearity (no high correlation between variables)
- Homoscedasticity (variance of error terms is similar across variables)

Most of the data is in good shape; however, some parameters appear to be right skewed and some left so different transformations were applied:

**LotArea** - max limiting to $99^{th}$ percentile, min limiting to 0.01 and log transformation applied.
**MasVnrArea** – min limiting to 0.01 and various transformations applied but none helped.
**BsmtFinSF1** – min limiting to 0.01 and various transformations applied but none helped.
**TotalBsmtSF** – max limiting to $99^{th}$ percentile applied.
**1stFlrSF** – min limiting to 0.01 and log transformation applied.
**2ndFlrSF** – min limiting to 0.01 and log transformation applied.
**WoodDeckSF** – min limiting to 0.01 and various transformations applied but none helped.
**OpenPorchSF** – min limiting to 0.01 and various transformations applied but none helped.
**EnclosedPorch** – transformation to binary categorical variable applied.
**PoolArea** – transformation to binary categorical variable applied.

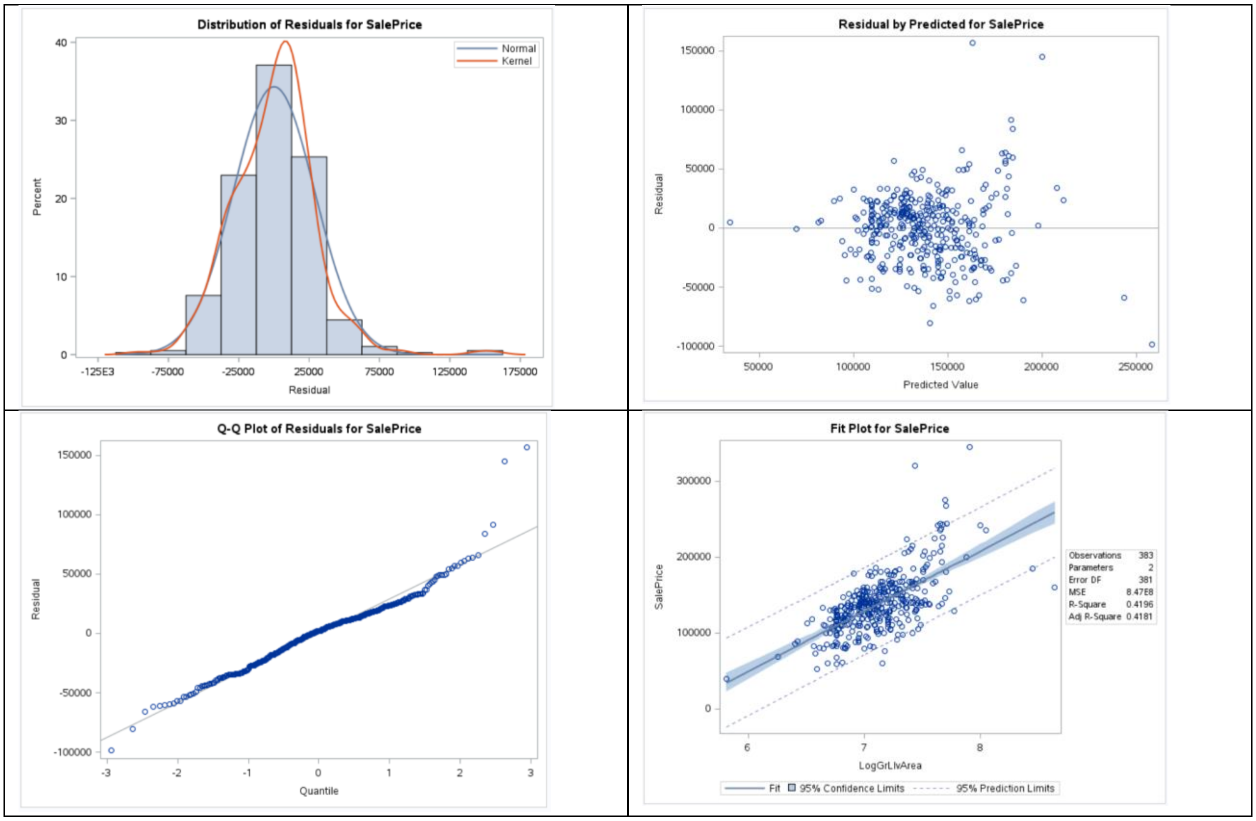
Detailed information and histograms found on appendix.

## Multiple Linear Regression:

What we are trying to accomplish is to improve our previous multiple linear regression analysis to predict house sale price for the company "Century 21" located in Ames, Iowa.

We decide to include models using forward selection, stepwise selection as well as LASSO. Criteria for selecting our best model were AIC, BIC, Adj R2 and CVPress.

The histogram seems to be normally distributed, and the scatterplot of residuals shows a normal variance with minimal clustering. The Q-Q plot shows that the data appears linear and fits the line well. From the Crooks D chart there appear to be a couple of outliers, however, since this data has already been filtered down to only residential sales the decision was made to keep these outliers and proceed.




For the final model using forward selection, we have an intercept of -425,922 for the sale price and parameter estimate of 79,219 for the log-transformed square footage of the homes:

$$sale\ price = -425{,}922 + 79{,}219\ln(square\ footage)$$

When the square footage is 1, the sale price would be -\$425,922. Multiplying the square footage by e will increase the sale price by \$79,219. More practically, a home with 1,000 square feet of living area is expected to have a sale price of \$121,303. The price would increase by approximately \$79,219 when the living area is multiplied by e (approximately 2.718).

In conclusion, it does appear that the living area of the home in square feet has a correlation to the ultimate sale price of the home. The larger area available for living space, the greater the sale price is expected to be.

To build a predictive model for sales prices of homes in *all* neighborhoods of Ames, IA, three regression models (forward selection, stepwise selection, and LASSO) were created. A brief summary of the three models can be seen below:

| Predictive Model | Adjusted $R^2$ | AIC | Kaggle Score |
|---|---|---|---|
| Forward selection | 0.9510 | 29933 | 0.15031 |
| Stepwise selection | 0.9427 | 28459 | 0.15423 |
| LASSO | 0.7141 | 32408 | 0.22931 |

The forward selection and stepwise selection models were created using all variables, two interactions maximum, single hierarchy, and random CVMETHOD. The LASSO selection used all variables, two interactions maximum with still single hierarchy and random SBC.

| | |
|---|---|
| Root MSE | 17554 |
| Dependent Mean | 180615 |
| R-Square | 0.9544 |
| Adj R-Sq | 0.9510 |
| AIC | 29933 |
| AICC | 29949 |
| PRESS | 5.075129E11 |
| SBC | 29018 |

| | |
|---|---|
| Root MSE | 18882 |
| Dependent Mean | 185182 |
| R-Square | 0.9465 |
| Adj R-Sq | 0.9427 |
| AIC | 28459 |
| AICC | 28472 |
| SBC | 27561 |

| | |
|---|---|
| Root MSE | 42480 |
| Dependent Mean | 180615 |
| R-Square | 0.7141 |
| Adj R-Sq | 0.7129 |
| AIC | 32408 |
| AICC | 32408 |
| SBC | 30991 |

Forward selection | Stepwise elimination | LASSO selection

Reviewing our three models revealed that the forward selection model seemed to be the most appropriate of the bunch, as it has the highest adjusted $R^2$ and the lowest CV press and SBC as well as AIC.

Certain variables were removed from the model that were found to be irrelevant to the sale price of the homes: Alley, Building Style, Exterior Quality, LowQualFinSF, BsmtHalfBath, Screen Porch, and Misc Feature.

Overall, the most predictive model of the three described above is the forward model. It granted us the following Kaggle score. This model can be used to predict future sale prices of homes in Ames, Iowa, based on the many varied features and sizes of the home. For more details on the analysis please refer to appendix.

| | | | | | |
|---|---|---|---|---|---|
| 1616 | ▲578 | Evangelos Giakoumakis | 0.15031 | 4 | now |

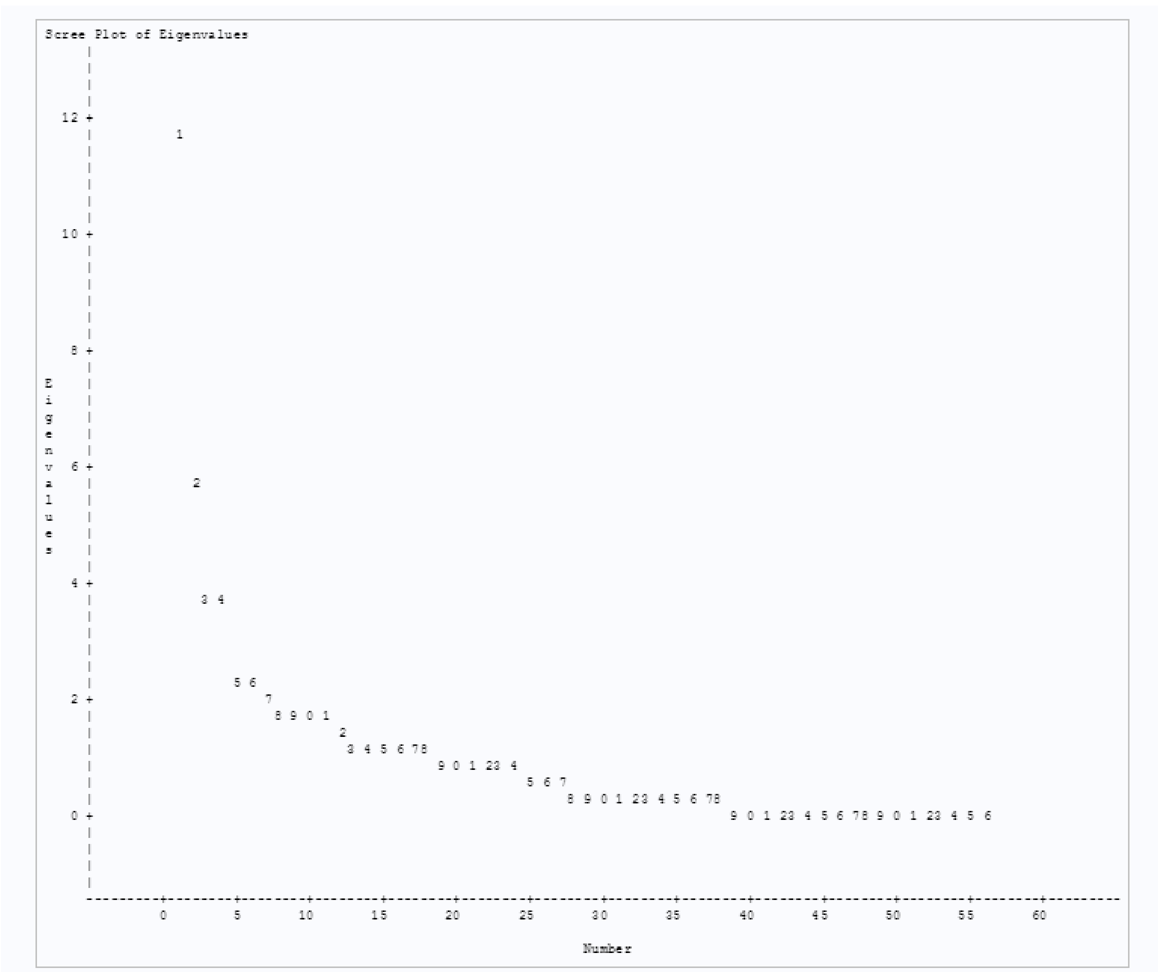
**Your Best Entry ↑**

Your submission scored 0.15031, which is not an improvement of your best score. Keep trying!

## **Principal Components Analysis:**

For PCA we will use quantitative continuous variables from the data set which have at least some correlation with sale price (Pearson Correlation Coefficient more than 0.1) because otherwise, these variables would add mostly a noise. Also, we should notice, those component loadings will not extend to sale price prediction beyond provided data sets because they are built on them. The same loadings on principle components, as produced on this data sets will never appear on subsequent data sets. If we want to extend our model to other sets, we would have to create linear combinations from the features identified by principal components.

We will use transformed variables because PCA is also sensitive to outliers. PCA is sensitive to variance, so for PCA calculation, we'll use standardized variables. We want to choose enough components to explain about 90% of data. According to scree-plots and eigenvalues of the correlation matrix, we need 18 principal components. Scree plot has an elbow at about 13 components after which variance explained declines very slowly, but these 13 components are not enough so we decided to include 18.



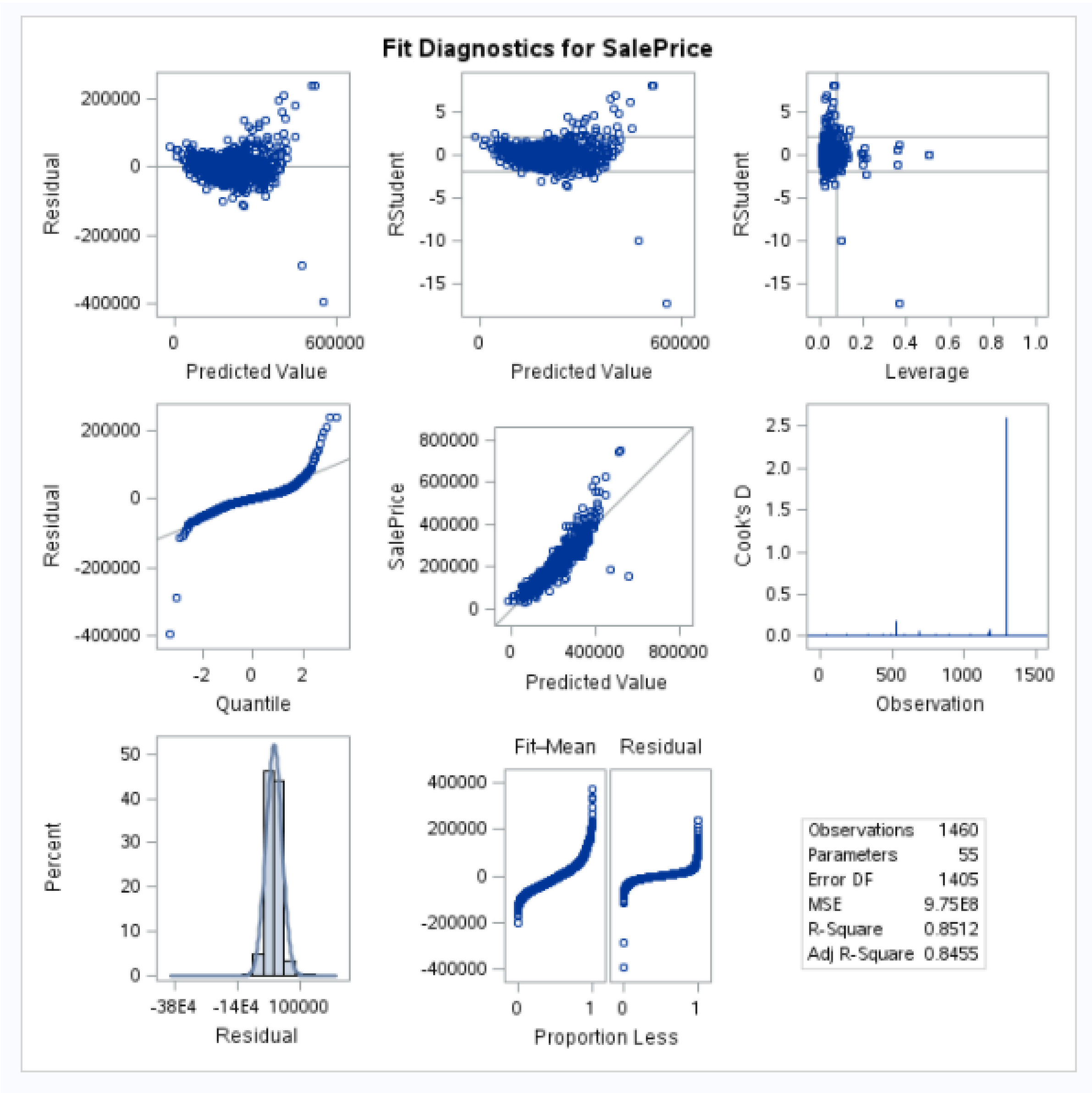
This is the variance explained by each of the 18 principle components.

| Variance Explained by Each Factor | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Factor1 | Factor2 | Factor3 | Factor4 | Factor5 | Factor6 | Factor7 | Factor8 | Factor9 | Factor10 | Factor11 | Factor12 | Factor13 | Factor14 | Factor15 | Factor16 | Factor17 | Factor18 |
| 11.847358 | 5.571824 | 3.841261 | 3.700826 | 2.290836 | 2.251848 | 1.953250 | 1.781466 | 1.769627 | 1.652061 | 1.592607 | 1.386874 | 1.185380 | 1.113579 | 1.065933 | 1.055635 | 1.034929 | 1.008118 |

| Standardized Scoring Coefficients | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 | Factor6 | Factor7 | Factor8 | Factor9 | Factor10 | Factor11 | Factor12 | Factor13 | Factor14 | Factor15 | Factor16 | Factor17 | Factor18 |
| Id | -0.00149 | -0.00130 | 0.00506 | 0.00146 | 0.02672 | -0.03847 | 0.04294 | -0.02717 | 0.04544 | -0.06726 | -0.04488 | 0.05355 | 0.04696 | -0.08212 | 0.51075 | -0.41550 | 0.16138 | 0.29898 |
| MSSubClass | -0.00540 | -0.04004 | 0.03907 | -0.12869 | -0.03868 | -0.02946 | 0.03232 | 0.04402 | 0.13881 | 0.02304 | -0.04770 | 0.32783 | 0.04708 | 0.21557 | -0.06477 | 0.05113 | -0.03525 | -0.12992 |
| LotArea | 0.02179 | 0.02964 | 0.06992 | 0.07463 | 0.05327 | 0.02958 | -0.04469 | -0.07370 | -0.09389 | -0.02637 | -0.08262 | -0.20372 | -0.12062 | -0.19526 | -0.08701 | 0.00179 | -0.10981 | -0.22003 |
| OverallQual | 0.06565 | -0.03254 | -0.00371 | -0.01802 | 0.01250 | 0.00035 | 0.01257 | 0.06848 | -0.03378 | 0.06134 | 0.14049 | 0.03481 | 0.09567 | 0.08416 | 0.08303 | 0.08889 | -0.09014 | -0.01116 |
| OverallCond | -0.02190 | 0.01120 | 0.05690 | 0.02819 | 0.00833 | 0.07357 | 0.00667 | -0.03096 | -0.13492 | 0.04822 | 0.12626 | -0.15213 | 0.24057 | 0.24596 | 0.13280 | 0.07664 | 0.43627 | 0.23367 |
| YearBuilt | 0.05568 | -0.00520 | -0.09851 | -0.11553 | 0.03914 | 0.04572 | 0.02625 | 0.05965 | 0.00712 | -0.01612 | 0.00038 | 0.02384 | 0.01799 | -0.19427 | 0.05499 | 0.03615 | -0.13275 | -0.08778 |
| YearRemodAdd | 0.04708 | -0.02398 | -0.06257 | -0.08083 | 0.08812 | 0.04851 | 0.03481 | 0.00004 | -0.05637 | 0.06621 | 0.13163 | -0.01149 | 0.20517 | -0.00885 | 0.09469 | 0.11278 | 0.12306 | 0.05619 |
| MasVnrArea | 0.09357 | -0.00755 | 0.05218 | -0.02946 | -0.54668 | 0.04270 | -0.12331 | 0.39509 | 0.18202 | 0.10166 | -0.06619 | -0.26196 | -0.16812 | -0.02926 | -0.08346 | -0.03172 | 0.24805 | 0.06662 |
| BsmtFinSF1 | 0.14051 | 0.32412 | 0.08746 | 0.02221 | -0.16082 | -0.08804 | 0.17951 | -0.06008 | -0.16219 | -0.01094 | 0.09896 | 0.25249 | 0.12683 | -0.05556 | 0.02054 | -0.02437 | -0.15163 | -0.08865 |
| BsmtFinSF2 | 0.01668 | 0.12202 | 0.06245 | 0.12860 | 0.42649 | 0.29023 | -0.44199 | 0.41865 | 0.47174 | -0.08750 | 0.02666 | 0.06942 | 0.03209 | -0.04239 | 0.02743 | 0.10907 | 0.02442 | 0.05146 |
| BsmtUnfSF | 0.11227 | -0.15091 | -0.27285 | 0.39034 | -0.02645 | -0.03131 | 0.10920 | 0.21999 | -0.13059 | 0.06654 | 0.07466 | 0.21682 | 0.12990 | 0.04547 | -0.03291 | -0.08887 | -0.13534 | -0.08420 |
| TotalBsmtSF | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1stFlrSF | 0.10481 | 0.00738 | 0.08937 | 0.16949 | -0.00198 | -0.02428 | 0.00580 | 0.01609 | -0.03350 | -0.01781 | -0.04909 | 0.19169 | 0.11109 | 0.04197 | -0.01282 | 0.04062 | -0.02528 | 0.05028 |
| 2ndFlrSF | 0.06758 | -0.15514 | 0.28874 | -0.03957 | 0.02695 | -0.03816 | -0.02925 | 0.01119 | -0.04531 | 0.00839 | 0.01357 | 0.07176 | 0.09481 | -0.00583 | 0.00829 | 0.04296 | -0.08165 | 0.02869 |
| LowQualFinSF | 0.00184 | -0.01455 | 0.04209 | 0.02279 | 0.05123 | -0.01393 | 0.05650 | -0.03213 | 0.02097 | -0.03225 | 0.00164 | -0.06853 | 0.02696 | 0.42830 | -0.47713 | 0.06626 | 0.00756 | 0.08856 |
| GrLivArea | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| BsmtFullBath | 0.04767 | 0.25632 | 0.07648 | -0.12276 | 0.11672 | -0.24796 | -0.10425 | 0.00069 | -0.02075 | 0.03652 | 0.00189 | 0.13208 | -0.01659 | 0.04002 | -0.08094 | -0.03280 | 0.01645 | 0.07847 |
| BsmtHalfBath | -0.00729 | 0.02520 | 0.07346 | 0.03740 | -0.18958 | 0.73002 | 0.19083 | -0.35352 | 0.16194 | 0.01106 | 0.23108 | 0.19560 | 0.04494 | -0.06502 | -0.06499 | -0.14619 | -0.08076 | -0.06652 |
| FullBath | 0.05362 | -0.06946 | 0.01825 | -0.01377 | 0.03834 | -0.00471 | 0.00140 | -0.01766 | 0.00347 | -0.00817 | -0.04780 | 0.16434 | 0.09099 | -0.12406 | 0.01459 | 0.09937 | -0.00250 | 0.01987 |
| HalfBath | 0.01983 | -0.06316 | 0.10632 | -0.10464 | 0.01376 | 0.02360 | -0.04045 | 0.07763 | -0.04276 | -0.03826 | 0.07472 | -0.09192 | -0.00770 | -0.03291 | 0.02045 | -0.05730 | -0.14312 | -0.07902 |
| BedroomAbvGr | 0.01572 | -0.06975 | 0.12940 | 0.05601 | -0.00268 | 0.01398 | -0.04159 | -0.03929 | -0.02554 | -0.06791 | -0.15431 | 0.05349 | 0.05539 | -0.18182 | 0.01803 | -0.03494 | 0.01616 | 0.12908 |
| KitchenAbvGr | -0.00459 | -0.02078 | 0.02748 | 0.03272 | -0.05556 | -0.06646 | -0.03089 | -0.14529 | 0.15466 | -0.10604 | -0.28688 | 0.31697 | 0.02400 | -0.00152 | -0.11393 | 0.04736 | 0.22026 | -0.08434 |
| TotRmsAbvGrd | 0.04566 | -0.07174 | 0.13075 | 0.04963 | 0.00220 | -0.02154 | -0.04627 | -0.04148 | -0.02365 | -0.03776 | -0.10885 | 0.10394 | 0.08705 | -0.03908 | -0.01336 | 0.00992 | 0.02888 | 0.05526 |
| Fireplaces | 0.03647 | 0.00995 | 0.08478 | 0.04609 | -0.00447 | 0.02397 | -0.01444 | 0.07157 | -0.11682 | 0.01425 | 0.05548 | -0.00654 | -0.02357 | 0.17443 | 0.12022 | 0.02478 | -0.23644 | -0.12393 |
| GarageYrBlt | 0.11163 | -0.04600 | -0.20659 | -0.23439 | 0.14924 | 0.04558 | 0.04375 | -0.05104 | 0.11892 | 0.03569 | 0.03423 | -0.03661 | 0.08353 | -0.19579 | 0.01959 | 0.07305 | -0.09409 | -0.09864 |
| GarageCars | 0.13332 | -0.05395 | -0.05623 | -0.04739 | 0.02866 | -0.05930 | -0.08657 | -0.35585 | 0.29393 | -0.04533 | -0.03077 | -0.21080 | -0.07290 | 0.22992 | 0.07578 | -0.02932 | 0.01094 | 0.02462 |
| GarageArea | 0.13194 | -0.01277 | -0.04380 | -0.00955 | 0.03367 | -0.09135 | -0.04765 | -0.37971 | 0.31678 | -0.06218 | 0.00920 | -0.31049 | -0.09591 | 0.22821 | -0.01772 | -0.08844 | 0.16123 | 0.06329 |
| WoodDeckSF | 0.03064 | 0.01723 | 0.02301 | -0.02482 | 0.10221 | 0.11397 | 0.00003 | 0.02056 | -0.13464 | 0.36017 | -0.21661 | -0.00508 | -0.12241 | 0.10784 | -0.05283 | -0.10045 | 0.11921 | 0.01516 |
| OpenPorchSF | 0.03396 | -0.02048 | 0.02433 | -0.01087 | 0.08434 | -0.00265 | 0.01987 | 0.06565 | -0.11126 | -0.18613 | 0.24177 | 0.11602 | -0.23807 | -0.03840 | -0.23571 | -0.20164 | 0.32601 | 0.02687 |
| EnclosedPorch | -0.01904 | -0.00770 | 0.06675 | 0.08855 | -0.00922 | -0.13491 | -0.02345 | -0.07307 | 0.17804 | 0.28990 | 0.29092 | 0.05598 | -0.03769 | -0.08261 | 0.00777 | 0.00241 | -0.01711 | -0.04835 |
| 3SsnPorch | 0.00355 | 0.00388 | -0.01046 | 0.00701 | -0.02618 | 0.02292 | 0.01450 | -0.03900 | -0.04361 | -0.04380 | 0.04318 | -0.05856 | 0.10744 | -0.19813 | -0.15398 | 0.58934 | -0.01496 | 0.41780 |
| ScreenPorch | 0.00506 | 0.00972 | 0.04039 | 0.02962 | -0.01149 | 0.03352 | -0.01115 | 0.09489 | -0.02347 | -0.22057 | 0.05452 | -0.09145 | 0.10061 | 0.42619 | 0.19928 | -0.03285 | -0.16256 | -0.18102 |
| PoolArea | 0.00952 | 0.01552 | 0.07280 | 0.03089 | 0.10300 | -0.02616 | 0.40223 | 0.14201 | 0.16976 | -0.00053 | -0.06367 | -0.10167 | -0.01907 | -0.04460 | 0.01309 | 0.02711 | 0.02680 | 0.03772 |
| MiscVal | -0.00264 | 0.00243 | 0.01636 | 0.01193 | 0.01173 | -0.00277 | 0.02051 | -0.00864 | -0.02618 | -0.01540 | -0.01142 | -0.04696 | 0.14018 | -0.06964 | 0.17422 | 0.29593 | 0.49836 | -0.64422 |
| MoSold | 0.00366 | -0.00990 | 0.00342 | 0.00816 | 0.00503 | 0.02729 | -0.01179 | -0.04177 | -0.06033 | -0.02173 | -0.00321 | 0.13938 | -0.40735 | 0.15501 | 0.37772 | 0.35853 | 0.03918 | 0.20293 |
| YrSold | -0.00240 | 0.01294 | -0.00607 | -0.01207 | 0.00096 | -0.01995 | -0.07197 | 0.02475 | -0.01086 | 0.06060 | -0.03431 | -0.06039 | 0.53409 | -0.06446 | -0.13118 | -0.21038 | 0.05061 | 0.10355 |
| SalePrice | 0.07263 | -0.00726 | 0.03921 | 0.00744 | 0.01125 | 0.00550 | -0.01647 | 0.01374 | -0.05536 | 0.03708 | 0.08262 | -0.03180 | 0.07324 | 0.05133 | 0.06665 | 0.05516 | -0.05476 | 0.00398 |
| Imp_MasVnrArea | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_GarageYrBlt | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_BsmtFinSF1 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_BsmtFinSF2 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_BsmtUnfSF | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_BsmtFullBath | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_BsmtHalfBath | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_GarageCars | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Imp_GarageArea | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Trcf_LotArea | 0.03174 | 0.02207 | 0.07569 | 0.12183 | 0.06476 | 0.01898 | -0.05551 | -0.13200 | -0.11711 | -0.05335 | -0.07226 | -0.27658 | -0.10840 | -0.20924 | -0.02860 | -0.03282 | -0.01940 | -0.06120 |
| Trcf_MasVnrArea | 0.04451 | 0.00379 | -0.01189 | -0.02928 | -0.23513 | 0.03706 | -0.04095 | 0.18265 | 0.08216 | 0.03530 | -0.04262 | -0.05902 | -0.07826 | -0.05182 | -0.01324 | 0.00020 | 0.13107 | 0.06001 |
| Trcf_BsmtFinSF1 | 0.01392 | 0.13156 | 0.06250 | -0.06698 | -0.05312 | 0.02137 | -0.01797 | -0.00642 | -0.02681 | -0.03919 | 0.05351 | 0.06045 | 0.05911 | -0.04929 | 0.01984 | 0.02717 | -0.00194 | 0.01951 |
| Trcf_TotalBsmtSF | 0.06278 | 0.05239 | -0.04729 | 0.10994 | -0.00936 | 0.00302 | 0.00249 | 0.06805 | -0.03669 | 0.00776 | 0.04870 | 0.12381 | 0.06336 | 0.00513 | 0.00581 | -0.01010 | -0.06865 | -0.03621 |
| Trcf_1stFlrSF | 0.05879 | 0.04888 | -0.01904 | 0.14293 | -0.00522 | 0.00016 | -0.00937 | -0.00971 | -0.02177 | -0.02411 | -0.05068 | 0.12132 | 0.04617 | 0.02585 | 0.00899 | 0.02317 | 0.01259 | 0.03562 |
| Trcf_2ndFlrSF | 0.00312 | -0.10950 | 0.16437 | -0.08579 | 0.01565 | -0.02153 | -0.02303 | 0.00995 | -0.03507 | 0.02117 | 0.03739 | 0.01507 | 0.03133 | -0.02579 | 0.02489 | 0.00405 | -0.07483 | -0.00470 |
| Trcf_WoodDeckSF | 0.03082 | 0.00686 | -0.00257 | -0.04510 | 0.11523 | 0.11689 | 0.00722 | 0.01688 | -0.12553 | 0.37256 | -0.20107 | 0.03967 | -0.10396 | 0.11111 | -0.02734 | -0.05014 | 0.07785 | 0.01036 |
| Trcf_OpenPorchSF | 0.04344 | -0.03168 | -0.00909 | -0.04066 | 0.09012 | 0.00797 | 0.02191 | 0.06403 | -0.11949 | -0.13191 | 0.22782 | 0.11327 | -0.19681 | -0.09313 | -0.19910 | -0.15447 | 0.25447 | 0.02357 |
| Trcf_EnclosedPorch | -0.02512 | -0.00945 | 0.05739 | 0.09092 | -0.01849 | -0.14220 | -0.04810 | -0.07070 | 0.15360 | 0.29190 | 0.28268 | 0.04552 | -0.03858 | -0.07817 | 0.02763 | 0.00668 | 0.00224 | -0.04929 |
| Trcf_PoolArea | 0.01057 | 0.01730 | 0.07606 | 0.03237 | 0.09963 | -0.02866 | 0.40242 | 0.13967 | 0.17176 | -0.00378 | -0.05680 | -0.09657 | -0.01088 | -0.04312 | 0.00726 | 0.02845 | 0.01727 | 0.02862 |

In our model selection, we will include the most useful categorical variables and principal components 1 through 18.

We will use the custom method to choose a model. At this model, we took principle components which have the correlation with the sale price at least of 0.05 and then deleted them according to the least statistical significance. This way we came to the principal components chosen by stepwise model. And in the custom model, we included variables, which were chosen the most by others selection methods.



The GLM Procedure
Dependent Variable: SalePrice

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 54 | 7.8379046E12 | 145146381829 | 148.85 | <.0001 |
| Error | 1405 | 1.3700067E12 | 975093747.94 | | |
| Corrected Total | 1459 | 9.2079113E12 | | | |

| R-Square | Coeff Var | Root MSE | SalePrice Mean |
|---|---|---|---|
| 0.851214 | 17.25972 | 31226.49 | 180921.2 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Prin1 | 1 | 6.6942533E12 | 6.6942533E12 | 6865.24 | <.0001 |
| Prin2 | 1 | 95061880085 | 95061880085 | 97.49 | <.0001 |
| Prin3 | 1 | 12861591699 | 12861591699 | 13.19 | 0.0003 |
| Prin4 | 1 | 141956836492 | 141956836492 | 145.58 | <.0001 |
| Prin5 | 1 | 134692051317 | 134692051317 | 138.13 | <.0001 |
| Prin6 | 1 | 10577708729 | 10577708729 | 10.85 | 0.0010 |
| Prin7 | 1 | 59292423032 | 59292423032 | 60.81 | <.0001 |
| Prin8 | 1 | 14039730616 | 14039730616 | 14.40 | 0.0002 |
| Prin9 | 1 | 7935911.3594 | 7935911.3594 | 0.01 | 0.9281 |
| Prin10 | 1 | 6160297343.4 | 6160297343.4 | 6.32 | 0.0121 |
| Prin11 | 1 | 108711238650 | 108711238650 | 111.49 | <.0001 |
| Prin12 | 1 | 3060076111 | 3060076111 | 3.14 | 0.0767 |
| Prin13 | 1 | 87845104584 | 87845104584 | 90.09 | <.0001 |
| Prin14 | 1 | 187681878.91 | 187681878.91 | 0.19 | 0.6609 |
| Prin15 | 1 | 37118804598 | 37118804598 | 38.07 | <.0001 |
| Prin16 | 1 | 2250244842.3 | 2250244842.3 | 2.31 | 0.1290 |
| Prin17 | 1 | 7211673.1619 | 7211673.1619 | 0.01 | 0.9315 |
| Prin18 | 1 | 5640870907.6 | 5640870907.6 | 5.78 | 0.0163 |
| Neighborhood | 24 | 296097005952 | 12337375248 | 12.65 | <.0001 |
| KitchenQual | 3 | 118197719816 | 39399239939 | 40.41 | <.0001 |
| Foundation | 5 | 5271181323 | 1054236264.6 | 1.08 | 0.3689 |
| HeatingQC | 4 | 4613702311.6 | 1153425577.9 | 1.18 | 0.3165 |

It appears that our model is valid. It doesn't have high influential points (Cook's D below 0.04) and there are no points that have high leverage and high influence. The residuals are normally distributed with heavy left tail and form random cloud against the regression line. If our purpose is the prediction of sale price of the houses in test data set, then our custom model has good predictive power, the loadings in this model have the correlation with the sale price and this model neither overfits nor underfits the data plus has good predictive power.

**Linear Discriminant Analysis:**

Here we will use a classification algorithm to predict foundation of the houses in our test data set using the training data set.

Assumptions that must be met:

- Multivariate normal distribution
- Outliers sensitivity

It is hard to check multivariate normality given such a huge number of variables in the data set. We created variance-covariance matrices to check this assumption, which is necessary for levels of foundation with a few observations. Those matrices can be found in appendix. We decided to use only variables which don't have a lot of zero values and which don't look suspicious on the variance-covariance matrices.

The final set of variables for classification is MSSubClass, OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1st1FlrSF, 2nd2FlrSF, GrLivArea, BedroomAbvGr, KitchenAbvGr, Fireplaces, GarageYrBlt, GarageCars, GarageArea with priors for types of foundation "BrkTil" is 0.1018, "CBlock" is 0.4416, "PConc" is 0.4337, "Slab" is 0.0165, "Stone" is 0.0043 and "Wood"is 0.0022, which we have taken from the percentage of certain type of foundation in the train data set.

The test of homogeneity of within covariance matrices show that our matrices have homogeneity, so we will proceed with LDA. Only stone and wood types of foundation don't have enough observation for CLT to work, so we expect to have higher misclassification rates for these types of foundations.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.PCAPRED
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Foundation

| From Foundation | BrkTil | CBlock | PConc | Slab | Stone | Wood | Total |
|---|---|---|---|---|---|---|---|
| BrkTil | 238<br>85.61 | 34<br>12.23 | 3<br>1.08 | 3<br>1.08 | 0<br>0.00 | 0<br>0.00 | 278<br>100.00 |
| CBlock | 102<br>8.81 | 910<br>78.58 | 128<br>11.05 | 18<br>1.55 | 0<br>0.00 | 0<br>0.00 | 1158<br>100.00 |
| PConc | 76<br>5.99 | 119<br>9.38 | 1069<br>84.31 | 2<br>0.16 | 0<br>0.00 | 2<br>0.16 | 1268<br>100.00 |
| Slab | 0<br>0.00 | 0<br>0.00 | 1<br>2.50 | 39<br>97.50 | 0<br>0.00 | 0<br>0.00 | 40<br>100.00 |
| Stone | 9<br>90.00 | 1<br>10.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 10<br>100.00 |
| Wood | 0<br>0.00 | 1<br>20.00 | 4<br>80.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 5<br>100.00 |
| Total | 425<br>15.40 | 1065<br>38.60 | 1205<br>43.68 | 62<br>2.25 | 0<br>0.00 | 2<br>0.07 | 2759<br>100.00 |
| Priors | 0.10179 | 0.44156 | 0.43386 | 0.0165 | 0.0043 | 0.0022 | |

Error Count Estimates for Foundation

| | BrkTil | CBlock | PConc | Slab | Stone | Wood | Total |
|---|---|---|---|---|---|---|---|
| Rate | 0.1439 | 0.2142 | 0.1569 | 0.0250 | 1.0000 | 1.0000 | 0.1842 |
| Priors | 0.1018 | 0.4416 | 0.4337 | 0.0165 | 0.0043 | 0.0022 | |

This confusion matrix shows the misclassification rate is 18.4%.

**Conclusion:**

In this paper we tried to predict the selling price of houses using Multiple Linear Regression and Principal Component Analysis (PCA). Multiple linear regression is a good choice when we know the underling structure of the data and when we have more observations than explanatory variables or when the explanatory variables are correlated with each other. Principal Components Analysis is great for this data set, because of its volume. It proved to have high predictive power, however, for it to be used beyond the scope of our test data set we will need to interpret the principle components and make linear combinations of the features.
With Linear Discriminant Analysis we were able to classify a foundation type in the test data set.

## Appendix

Analysis

```
/* Project 2 SAS Code */

/* Import train data-set */
FILENAME REFFILE '/home/egiakoumakis0/sasuser.v94/Stat 2/Project2/train data
project 2.csv';
```

```
PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV
      OUT=WORK.TRAINHOUSING;
      GETNAMES=YES;
RUN;

/* Import test data set */
FILENAME REFFILE '/home/egiakoumakis0/sasuser.v94/Stat 2/Project2/test data
project 2.csv';

PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV
      OUT=WORK.TESTHOUSING;
      GETNAMES=YES;
RUN;

/* Add empty SalePrice column */
data WORK.TESTHOUSING;
set WORK.TESTHOUSING;
SalePrice = .;
;

/* Combine Datasets */
data house;
set WORK.TRAINHOUSING WORK.TESTHOUSING;
run;

/* create a format to group missing and nonmissing */
proc format;
 value $missfmt ' '='Missing' other='Not Missing';
 value  missfmt  . ='Missing' other='Not Missing';
run;

 /* check columns for missing and nonmissing values */
proc freq data=house;
format _CHAR_ $missfmt.; /* apply format for the duration of this PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;

proc means data=house mean std min p25 median p75 max p5 p95 p99 maxdec=1;
run;

/* Data Imputes */
data imp_house;
set house;
/*Impute MasVnrArea with 0 if no value present */
Imp_MasVnrArea = MasVnrArea;
if Imp_MasVnrArea = . then Imp_MasVnrArea = 0;
/*Impute GarageYrBlt with (Min-1) if no value present */
Imp_GarageYrBlt = GarageYrBlt;
if Imp_GarageYrBlt = . then Imp_GarageYrBlt = 1899;
/*Impute LotFrontage with 0 if no value present */
Imp_LotFrontage = LotFrontage;
if Imp_LotFrontage EQ 'NA' THEN Imp_LotFrontage = 0;
```

```
/*Impute BsmtFinSF1 with 0 if no value present */
Imp_BsmtFinSF1 = BsmtFinSF1;
if Imp_BsmtFinSF1 EQ 'NA' THEN Imp_BsmtFinSF1 = 0;
/*Impute BsmtFinSF2 with 0 if no value present */
Imp_BsmtFinSF2 = BsmtFinSF2;
if Imp_BsmtFinSF2 EQ 'NA' THEN Imp_BsmtFinSF2 = 0;
/*Impute BsmtUnfSF with 0 if no value present */
Imp_BsmtUnfSF = BsmtUnfSF;
if Imp_BsmtUnfSF EQ 'NA' THEN Imp_BsmtUnfSF = 0;
/*Impute BsmtFullBath with 0 if no value present */
Imp_BsmtFullBath = BsmtFullBath;
if Imp_BsmtFullBath EQ 'NA' THEN Imp_BsmtFullBath = 0;
/*Impute BsmtHalfBath with 0 if no value present */
Imp_BsmtHalfBath = BsmtHalfBath;
if Imp_BsmtHalfBath EQ 'NA' THEN Imp_BsmtHalfBath = 0;
/*Impute GarageCars with 0 if no value present */
Imp_GarageCars = GarageCars;
if Imp_GarageCars EQ 'NA' THEN Imp_GarageCars = 0;
/*Impute GarageArea with 0 if no value present */
Imp_GarageArea = GarageArea;
if Imp_GarageArea EQ 'NA' THEN Imp_GarageArea = 0;

*if SalePrice  = . then SalePrice = 0;
run;

proc univariate data=imp_house noprint plots;
 histogram;
run;

/* Data Transformations */
data trsfm_imp_house;
set imp_house;
/*Transform LotArea limit to 99th percentile */
Trsf_LotArea = LotArea;
if Trsf_LotArea >= 40000 then Trsf_LotArea = 40000;
if Trsf_LotArea = 0 then Trsf_LotArea = 0.01;
Trsf_LotArea = log(Trsf_LotArea);

if Imp_MasVnrArea = 0 then Imp_MasVnrArea = 0.01;
Trsf_MasVnrArea = log(Imp_MasVnrArea); *not use;

if Imp_BsmtFinSF1 = 0 then Imp_BsmtFinSF1 = 0.01;
Trsf_BsmtFinSF1 = log(Imp_BsmtFinSF1); *not use;

Trsf_TotalBsmtSF = TotalBsmtSF;
if Trsf_TotalBsmtSF >= 3000 then Trsf_TotalBsmtSF = 3000;

if "1stFlrSF"n = 0 then "1stFlrSF"n = 0.01;
Trsf_1stFlrSF = log("1stFlrSF"n);

if "2ndFlrSF"n = 0 then "2ndFlrSF"n = 0.01;
Trsf_2ndFlrSF = log("2ndFlrSF"n); *not use;

if WoodDeckSF = 0 then WoodDeckSF = 0.01;
Trsf_WoodDeckSF = log(WoodDeckSF); *not use;

if OpenPorchSF = 0 then OpenPorchSF = 0.01;
```

```
Trsf_OpenPorchSF = log(OpenPorchSF); *not use;

Trsf_EnclosedPorch = EnclosedPorch;
if Trsf_EnclosedPorch > 0 then Trsf_EnclosedPorch = 1;

Trsf_PoolArea = PoolArea;
if Trsf_PoolArea > 0 then Trsf_PoolArea = 1;

run;

proc univariate data=trsfm_imp_house noprint plots;
 histogram;
run;

/* MDA*/

/* New selection */
proc glmselect data = trsfm_imp_house;
class MSZoning Imp_LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model SalePrice = MSSubClass | MSZoning | Imp_LotFrontage | Trsf_LotArea |
Street | LotShape | LandContour | Utilities | LotConfig
| LandSlope | Neighborhood | Condition1 | Condition2 | BldgType | HouseStyle
| OverallQual | OverallCond | YearBuilt | YearRemodAdd
| RoofStyle | RoofMatl | Exterior1st | Exterior2nd | MasVnrType |
Imp_MasVnrArea | ExterQual | ExterCond | Foundation | BsmtQual | BsmtCond
| BsmtExposure | BsmtFinType1 | Imp_BsmtFinSF1 | BsmtFinType2 |
Imp_BsmtFinSF2 | Imp_BsmtUnfSF | Trsf_TotalBsmtSF | Heating | HeatingQC |
CentralAir
| Electrical | Trsf_1stFlrSF | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
Imp_BsmtFullBath | Imp_BsmtHalfBath | FullBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageFinish
| Imp_GarageCars | Imp_GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | Trsf_EnclosedPorch | "3SsnPorch"n | ScreenPorch
| Trsf_PoolArea | PoolQC | Fence | MiscVal | MoSold | YrSold | SaleType |
SaleCondition | Imp_GarageYrBlt @2
/ selection=backward(choose=BIC) showpvalues;
output out = results_new p = Predict;
run;

/* Custom selection */
proc glmselect data = trsfm_imp_house;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
```

```
model SalePrice = MSSubClass | MSZoning | LotFrontage | LotArea | Street |
LotShape | LandContour | Utilities | LotConfig
| LandSlope | Neighborhood | Condition1 | Condition2 | BldgType | HouseStyle
| OverallQual | OverallCond | YearBuilt | YearRemodAdd
| RoofStyle | RoofMatl | Exterior1st | Exterior2nd | MasVnrType | MasVnrArea
| ExterQual | ExterCond | Foundation | BsmtQual | BsmtCond
| BsmtExposure | BsmtFinType1 | BsmtFinSF1 | BsmtFinType2 | BsmtFinSF2 |
BsmtUnfSF | TotalBsmtSF | Heating | HeatingQC | CentralAir
| Electrical | "1stFlrSF"n | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
BsmtFullBath | BsmtHalfBath | FullBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageFinish
| GarageCars | GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | EnclosedPorch | "3SsnPorch"n | ScreenPorch
| PoolArea | PoolQC | Fence | MiscVal | MoSold | YrSold | SaleType |
SaleCondition @2
/ selection=forward(choose=PRESS) showpvalues;
output out = results_cus p = Predict;
run;

/* Make results Kaggle friendly and fix prediction issues */
data results_final;
set results_cus;
if Predict > 0 then SalePrice = Round(Predict);
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

/* Stepwise selection using all variables */
proc glmselect data = trsfm_imp_house;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
ExterCond Foundation
BsmtQual BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir
Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model SalePrice = MSSubClass | MSZoning | LotFrontage | LotArea | Street |
Alley | LotShape | LandContour | Utilities | LotConfig
| LandSlope | Neighborhood | Condition1 | Condition2 | HouseStyle |
OverallQual | OverallCond | YearBuilt | YearRemodAdd
| RoofStyle | RoofMatl | Exterior1st | Exterior2nd | MasVnrType | MasVnrArea
| ExterCond | Foundation | BsmtQual
| BsmtExposure | BsmtFinType1 | BsmtFinSF1 | BsmtFinType2 | BsmtFinSF2 |
BsmtUnfSF | TotalBsmtSF | Heating | HeatingQC | CentralAir
| Electrical | "1stFlrSF"n | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
BsmtFullBath | BsmtHalfBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageYrBlt | GarageFinish
| GarageCars | GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | EnclosedPorch | "3SsnPorch"n | ScreenPorch
| PoolArea | PoolQC | Fence | MiscFeature | MiscVal | MoSold | YrSold |
SaleType | SaleCondition @2
/ selection=stepwise(choose=AIC) hierarchy=single showpvalues
cvmethod=random(2);
```

```
output out = results_step p = Predict;
run;

/* Make results Kaggle friendly and fix prediction issues */
data results_final;
set results_step;
if Predict > 0 then SalePrice = Round(Predict);
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

/* LASSO selection */
proc glmselect data = trsfm_imp_house;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model SalePrice = MSSubClass | MSZoning | LotFrontage | LotArea | Street |
LotShape | LandContour | Utilities | LotConfig
| LandSlope | Neighborhood | Condition1 | Condition2 | BldgType | HouseStyle
| OverallQual | OverallCond | YearBuilt | YearRemodAdd
| RoofStyle | RoofMatl | Exterior1st | Exterior2nd | MasVnrType | MasVnrArea
| ExterQual | ExterCond | Foundation | BsmtQual | BsmtCond
| BsmtExposure | BsmtFinType1 | BsmtFinSF1 | BsmtFinType2 | BsmtFinSF2 |
BsmtUnfSF | TotalBsmtSF | Heating | HeatingQC | CentralAir
| Electrical | "1stFlrSF"n | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
BsmtFullBath | BsmtHalfBath | FullBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageFinish
| GarageCars | GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | EnclosedPorch | "3SsnPorch"n | ScreenPorch
| PoolArea | PoolQC | Fence | MiscVal | MoSold | YrSold | SaleType |
SaleCondition @2
/ selection=lasso showpvalues;
output out = results_las p = Predict;
run;

/* Make results Kaggle friendly and fix prediction issues */
data results_final;
set results_las;
if Predict > 0 then SalePrice = Round(Predict);
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

/* PCA */

proc princomp data=trsfm_imp_house plots=all out=pcaPred;
var LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1
BsmtUnfSF TotalBsmtSF Trsf_1stFlrSF
```

```
"2ndFlrSF"n GrLivArea Fireplaces GarageCars BedroomAbvGr KitchenAbvGr
Imp_GarageYrBlt GarageArea Imp_MasVnrArea;
run;

proc corr data=pcaPred;
var SalePrice prin1-prin11;
run;

proc corr data=pcaPred;
var SalePrice prin11-prin18;
run;

proc glm data=pcaPred plots=diagnostics;
class Neighborhood KitchenQual Foundation HeatingQC;
model SalePrice=Prin1-Prin18 Neighborhood KitchenQual Foundation
HeatingQC /solution;
output out=results_ov p=Predict;
run;

data results_final;
set results_ov;
if Predict > 0 then SalePrice = Round(Predict);
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

/* LDA */

proc sort data=pcaPred;
by foundation;
run; proc
sgscatter data = pcaPred;
by Foundation;
matrix MSSubClass LotArea OverallQual OverallCond
YearBuilt/ ellipse=(type = mean alpha = .05);
run;
proc sgscatter data = pcaPred;
by Foundation; matrix YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2
BsmtUnfSF TotalBsmtSF/ ellipse=(type = mean alpha = .05);
run;
proc sgscatter data = pcaPred;
by Foundation; matrix "1stFlrSF"n "2ndFlrSF"n LowQualFinSF GrLivArea
BsmtFullBath
BsmtHalfBath/ ellipse=(type = mean alpha = .05);
run;
proc sgscatter data = pcaPred;
by Foundation; matrix FullBath HalfBath BedroomAbvGr KitchenAbvGr
TotRmsAbvGrd
Fireplaces/ ellipse=(type = mean alpha = .05);
run;
proc sgscatter data = pcaPred;
by Foundation; matrix GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF
EnclosedPorch/ ellipse=(type = mean alpha = .05);
run;
proc sgscatter data = pcaPred;
```

```
by Foundation; matrix Ssn3Porch ScreenPorch PoolArea MiscVal MoSold YrSold
SalePrice/ ellipse=(type = mean alpha = .05);
run;

proc discrim data=pcaPred pool=test;
class Foundation;
var MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea
BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF "1stFlrSF"n "2ndFlrSF"n
LowQualFinSF GrLivArea BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd
Fireplaces GarageYrBlt
GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch "3SsnPorch"n
ScreenPorch PoolArea
MiscVal MoSold YrSold SalePrice;
run;

data testFixLDA;
set test;
if TotalBsmtSF=. then TotalBsmtSF=0;
if BsmtFullBath=. then BsmtFullBath=0;
if BsmtHalfBath=. then BsmtHalfBath=0;
if GarageCars=. then GarageCars=0;
if GarageArea=. then GarageArea=0;
if MasVnrArea=. then MasVnrArea=0;
if BsmtFinSF1=. then BsmtFinSF1=0;
if BsmtFinSF2=. then BsmtFinSF2=0;
if BsmtUnfSF=. then BsmtUnfSF=0;
if LotFrontage=. then LotFrontage=0;
if GarageYrBlt=. then GarageYrBlt=YearBuilt;
if LotFrontage=0 then LotFrontage=0.01;
if MasVnrArea=0 then MasVnrArea=0.01;
if BsmtFinSF1=0 then BsmtFinSF1=0.01;
if BsmtFinSF2=0 then BsmtFinSF2=0.01;
if BsmtUnfSF=0 then BsmtUnfSF=0.01;
if TotalBsmtSF=0 then TotalBsmtSF=0.01;
if GarageArea=0 then GarageArea=0.01;
if OpenPorchSF=0 then OpenPorchSF=0.01;
if WoodDeckSF=0 then WoodDeckSF=0.01;
MSSubClass=MSSubClass**(-0.25);
LotArea=LotArea**0.5;
YearBuilt=YearBuilt**3;
MasVnrArea=MasVnrArea**(-0.25);
BsmtFinSF2=BsmtFinSF2**(-0.75);
BsmtUnfSF=BsmtUnfSF**0.5;
GrLivArea=GrLivArea**0.25;
GarageYrBlt=GarageYrBlt**3;
WoodDeckSF=log(WoodDeckSF);
OpenPorchSF=log(OpenPorchSF);
run;

proc discrim data=pcaPred pool=test crossvalidate;
class Foundation;var MSSubClass OverallQual OverallCond YearBuilt
YearRemodAdd
BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF "1stFlrSF"n "2ndFlrSF"n
BedroomAbvGr KitchenAbvGr Fireplaces GarageYrBlt GarageCars GarageArea;
```

```
priors "BrkTil"=.1018 "CBlock" = .4416 "PConc" = .4337 "Slab" = .0165
"Stone"=0.0043 "Wood"=0.0022;
run;

data predFound;
set pcaPred;
drop Foundation;
run;
```

The GLMSELECT Procedure

| Forward Selection Summary | | | | | |
|---|---|---|---|---|---|
| Step | Effect Entered | Number Effects In | Number Parms In | SBC | PRESS |
| 0 | Intercept | 1 | 1 | 32765.7557 | 9.13385E12 |
| 1 | GrLivArea*Neighborho | 2 | 26 | 30743.5053 | 2.12485E12 |
| 2 | OverallQu*BsmtExposu | 3 | 31 | 30304.8297 | 1.57253E12 |
| 3 | BsmtFinSF*LandContou | 4 | 35 | 30076.7664 | 1.34228E12 |
| 4 | GarageCar*KitchenQua | 5 | 39 | 29867.1899 | 1.14277E12 |
| 5 | OverallCo*TotalBsmtS | 6 | 40 | 29723.9471 | 1.05679E12 |
| 6 | LandConto*SaleCondit | 7 | 57 | 29593.8342 | 9.51937E11 |
| 7 | OverallQua*GrLivArea | 8 | 58 | 29373.8827 | 8.72002E11 |
| 8 | LotArea*BldgType | 9 | 63 | 29325.4189 | 8.41813E11 |
| 9 | PoolArea*BsmtExposur | 10 | 66 | 29279.5657 | 6.82493E11 |
| 10 | YearBuilt*BsmtFinSF1 | 11 | 67 | 29211.5184 | 6.48194E11 |
| 11 | BsmtUnfSF*ScreenPorc | 12 | 68 | 29181.2618 | 6.33672E11 |
| 12 | BsmtFinSF1*BsmtQual | 13 | 71 | 29148.9301 | 6.11586E11 |
| 13 | YearBuilt*YearRemodA | 14 | 72 | 29112.2476 | 5.95716E11 |
| 14 | OverallCo*Fireplaces | 15 | 73 | 29093.9224 | 5.85637E11 |
| 15 | LotArea*Fireplaces | 16 | 74 | 29085.4805 | 5.78295E11 |
| 16 | BsmtFinSF*BsmtFullBa | 17 | 75 | 29080.0862 | 5.74591E11 |
| 17 | MasVnrArea*2ndFlrSF | 18 | 76 | 29074.5561 | 5.72307E11 |
| 18 | FullBath*GarageArea | 19 | 77 | 29071.3956 | 5.69856E11 |
| 19 | GrLivArea*HalfBath | 20 | 78 | 29063.8876 | 5.63706E11 |
| 20 | KitchenAb*GarageCars | 21 | 79 | 29060.0212 | 5.60781E11 |
| 21 | BsmtFinSF*BedroomAbv | 22 | 80 | 29056.4349 | 5.5964E11 |
| 22 | BsmtFinSF*TotRmsAbvG | 23 | 81 | 29043.7121 | 5.57151E11 |
| 23 | GarageCars*PoolArea | 24 | 82 | 29041.9795 | 5.51303E11 |
| 24 | MasVnrAre*BsmtHalfBa | 25 | 83 | 29039.4669 | 5.51297E11 |
| 25 | MasVnrAre*MasVnrType | 26 | 87 | 29035.5456 | 5.41361E11 |
| 26 | BsmtUnfSF*OpenPorchS | 27 | 88 | 29032.7939 | 5.38607E11 |
| 27 | BsmtUnfSF*WoodDeckSF | 28 | 89 | 29031.5982 | 5.36374E11 |
| 28 | OverallCo*EnclosedPo | 29 | 90 | 29030.0573 | 5.34624E11 |
| 29 | OverallCon*BsmtUnfSF | 30 | 91 | 29028.5508 | 5.32628E11 |
| 30 | BsmtFullB*EnclosedPo | 31 | 92 | 29027.7754 | 5.29185E11 |
| 31 | MSZoning | 32 | 96 | 29027.5327 | 5.2094E11 |
| 32 | YearBuilt*GrLivArea | 33 | 97 | 29026.1748 | 5.19852E11 |
| 33 | OverallCo*KitchenAbv | 34 | 98 | 29026.1122 | 5.18371E11 |
| 34 | BsmtUnfSF*TotRmsAbvG | 35 | 99 | 29024.2629 | 5.16537E11 |
| 35 | BsmtUnfSF*KitchenAbv | 36 | 100 | 29023.7395 | 5.14908E11 |
| 36 | OverallQu*TotRmsAbvG | 37 | 101 | 29019.7024 | 5.11019E11 |
| 37 | LowQualFi*ScreenPorc | 38 | 102 | 29017.7439* | 5.07513E11* |
| * Optimal Value of Criterion | | | | | |

Selection stopped at a local minimum of the SBC criterion.

The GLMSELECT Procedure

| Stepwise Selection Summary | | | | | | |
|---|---|---|---|---|---|---|
| Step | Effect Entered | Effect Removed | Number Effects In | Number Parms In | AIC | SBC |
| 0 | Intercept | | 1 | 1 | 32291.6565 | 30923.8798 |
| 1 | OverallQual | | 2 | 2 | 30973.8333 | 29611.2799 |
| 2 | GrLivArea | | 3 | 3 | 30596.7313 | 29239.4012 |
| 3 | Neighborhood | | 4 | 27 | 30266.4067 | 29034.4357 |
| 4 | GrLivArea*Neighborho | | 5 | 51 | 29858.6027 | 28751.9908 |
| 5 | BsmtFinSF1 | | 6 | 52 | 29606.7962 | 28505.4076 |
| 6 | 2ndFlrSF | | 7 | 53 | 29495.6841 | 28399.5188 |
| 7 | KitchenQual | | 8 | 56 | 29406.4123 | 28325.9169 |
| 8 | BsmtExposure | | 9 | 60 | 29339.9316 | 28280.3293 |
| 9 | MSSubClass | | 10 | 61 | 29285.1652 | 28230.7862 |
| 10 | OverallCond | | 11 | 62 | 29244.9020 | 28195.7463 |
| 11 | YearBuilt | | 12 | 63 | 29192.0509 | 28148.1186 |
| 12 | RoofMatl | | 13 | 70 | 29119.9653 | 28112.5960 |
| 13 | LotArea | | 14 | 71 | 29081.3375 | 28079.1915 |
| 14 | OverallQu*BsmtFinSF1 | | 15 | 72 | 29044.4170 | 28047.4943 |
| 15 | OverallCon*GrLivArea | | 16 | 73 | 29010.4402 | 28018.7408 |
| 16 | SaleCondition | | 17 | 78 | 28954.1373 | 27988.5544 |
| 17 | RoofMatl*SaleConditi | | 18 | 82 | 28884.5082 | 27939.8184 |
| 18 | Condition2 | | 19 | 89 | 28821.8580 | 27913.7313 |
| 19 | BsmtFinSF*Condition2 | | 20 | 92 | 28765.3492 | 27872.8924 |
| 20 | OverallQua*GrLivArea | | 21 | 93 | 28668.9610 | 27781.7275 |
| 21 | | GrLivArea*Neighborho | 20 | 69 | 28737.3898 | 27724.7972 |
| 22 | | OverallCon*GrLivArea | 19 | 68 | 28735.9360 | 27718.1201 |
| 23 | TotalBsmtSF | | 20 | 69 | 28702.7161 | 27690.1235 |
| 24 | | 2ndFlrSF | 19 | 68 | 28700.7581 | 27682.9422 |
| 25 | TotalBsmt*SaleCondit | | 20 | 72 | 28644.1167 | 27647.1939 |
| 26 | GarageCars | | 21 | 73 | 28623.8244 | 27632.1250 |
| 27 | ScreenPorch | | 22 | 74 | 28607.2577 | 27620.7816 |
| 28 | OverallQu*TotalBsmtS | | 23 | 75 | 28591.7241 | 27610.4713 |
| 29 | GrLivArea*BsmtExposu | | 24 | 79 | 28563.8823 | 27603.5227 |
| 30 | KitchenAbvGr | | 25 | 80 | 28553.2578 | 27598.1215 |
| 31 | GrLivArea*GarageCars | | 26 | 81 | 28544.0360 | 27594.1230 |
| 32 | OpenPorchSF | | 27 | 82 | 28535.1265 | 27590.4368 |
| 33 | Fireplaces | | 28 | 83 | 28527.0847 | 27587.6183 |
| 34 | LotArea*Fireplaces | | 29 | 84 | 28512.8918 | 27578.6486 |
| 35 | YearBuilt*BsmtFinSF1 | | 30 | 85 | 28503.0382 | 27574.0183 |
| 36 | | OverallQu*BsmtFinSF1 | 29 | 84 | 28507.5614 | 27573.3182 |
| 37 | LotArea*OverallQual | | 30 | 85 | 28497.9168 | 27568.8969 |
| 38 | | MSSubClass | 29 | 84 | 28500.2340 | 27565.9908 |
| 39 | MSZoning | | 30 | 88 | 28476.4806 | 27563.1306 |
| 40 | OverallCo*Fireplaces | | 31 | 89 | 28470.2457 | 27562.1190 |
| 41 | CentralAir | | 32 | 90 | 28464.1387 | 27561.2353 |
| 42 | BsmtFullBath | | 33 | 91 | 28458.5386* | 27560.8585* |
| * Optimal Value of Criterion | | | | | | |

Selection stopped at a local minimum of the SBC criterion.

The GLMSELECT Procedure

| LASSO Selection Summary | | | | |
| --- | --- | --- | --- | --- |
| Step | Effect Entered | Effect Removed | Number Effects In | SBC |
| 0 | Intercept | | 1 | 32765.7557 |
| 1 | OverallQua*GrLivArea | | 2 | 32548.4104 |
| 2 | OverallQu*GarageCars | | 3 | 31794.6456 |
| 3 | OverallQual*1stFlrSF | | 4 | 31370.3923 |
| 4 | OverallQua*YearBuilt | | 5 | 31069.9778 |
| 5 | TotalBsmtSF*PoolQC_NA | | 6 | 31006.5948 |
| 6 | TotalBsmtSF*Condition1_Norm | | 7 | 30991.2351* |
| * Optimal Value of Criterion | | | | |

Selection stopped at a local minimum of the SBC criterion.

| Variable | Mean | Std Dev | Minimum | 25th Pctl | Median | 75th Pctl | Maximum | 5th Pctl | 95th Pctl | 99th Pctl |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 1460.0 | 842.8 | 1.0 | 730.0 | 1460.0 | 2190.0 | 2919.0 | 146.0 | 2774.0 | 2890.0 |
| MSSubClass | 57.1 | 42.5 | 20.0 | 20.0 | 50.0 | 70.0 | 190.0 | 20.0 | 160.0 | 190.0 |
| LotArea | 10168.1 | 7887.0 | 1300.0 | 7476.0 | 9453.0 | 11577.0 | 215245.0 | 3182.0 | 17169.0 | 33120.0 |
| OverallQual | 6.1 | 1.4 | 1.0 | 5.0 | 6.0 | 7.0 | 10.0 | 4.0 | 8.0 | 10.0 |
| OverallCond | 5.6 | 1.1 | 1.0 | 5.0 | 5.0 | 6.0 | 9.0 | 4.0 | 8.0 | 9.0 |
| YearBuilt | 1971.3 | 30.3 | 1872.0 | 1953.0 | 1973.0 | 2001.0 | 2010.0 | 1915.0 | 2007.0 | 2008.0 |
| YearRemodAdd | 1984.3 | 20.9 | 1950.0 | 1965.0 | 1993.0 | 2004.0 | 2010.0 | 1950.0 | 2007.0 | 2009.0 |
| MasVnrArea | 102.2 | 179.3 | 0.0 | 0.0 | 0.0 | 164.0 | 1600.0 | 0.0 | 468.0 | 772.0 |
| BsmtFinSF1 | 441.4 | 455.6 | 0.0 | 0.0 | 368.5 | 733.0 | 5644.0 | 0.0 | 1274.0 | 1636.0 |
| BsmtFinSF2 | 49.6 | 169.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1526.0 | 0.0 | 435.0 | 875.0 |
| BsmtUnfSF | 560.8 | 439.5 | 0.0 | 220.0 | 467.0 | 806.0 | 2336.0 | 0.0 | 1480.0 | 1777.0 |
| TotalBsmtSF | 1051.8 | 440.8 | 0.0 | 793.0 | 989.5 | 1302.0 | 6110.0 | 451.0 | 1777.0 | 2200.0 |
| 1stFlrSF | 1159.6 | 392.4 | 334.0 | 876.0 | 1082.0 | 1388.0 | 5095.0 | 665.0 | 1831.0 | 2290.0 |
| 2ndFlrSF | 336.5 | 428.7 | 0.0 | 0.0 | 0.0 | 704.0 | 2065.0 | 0.0 | 1133.0 | 1402.0 |
| LowQualFinSF | 4.7 | 46.4 | 0.0 | 0.0 | 0.0 | 0.0 | 1064.0 | 0.0 | 0.0 | 156.0 |
| GrLivArea | 1500.8 | 506.1 | 334.0 | 1126.0 | 1444.0 | 1744.0 | 5642.0 | 861.0 | 2466.0 | 2944.0 |
| BsmtFullBath | 0.4 | 0.5 | 0.0 | 0.0 | 0.0 | 1.0 | 3.0 | 0.0 | 1.0 | 2.0 |
| BsmtHalfBath | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 | 0.0 | 1.0 | 1.0 |
| FullBath | 1.6 | 0.6 | 0.0 | 1.0 | 2.0 | 2.0 | 4.0 | 1.0 | 2.0 | 3.0 |
| HalfBath | 0.4 | 0.5 | 0.0 | 0.0 | 0.0 | 1.0 | 2.0 | 0.0 | 1.0 | 1.0 |
| BedroomAbvGr | 2.9 | 0.8 | 0.0 | 2.0 | 3.0 | 3.0 | 8.0 | 2.0 | 4.0 | 5.0 |
| KitchenAbvGr | 1.0 | 0.2 | 0.0 | 1.0 | 1.0 | 1.0 | 3.0 | 1.0 | 1.0 | 2.0 |
| TotRmsAbvGrd | 6.5 | 1.6 | 2.0 | 5.0 | 6.0 | 7.0 | 15.0 | 4.0 | 9.0 | 11.0 |
| Fireplaces | 0.6 | 0.6 | 0.0 | 0.0 | 1.0 | 1.0 | 4.0 | 0.0 | 2.0 | 2.0 |
| GarageYrBlt | 1978.1 | 25.6 | 1895.0 | 1960.0 | 1979.0 | 2002.0 | 2207.0 | 1928.0 | 2007.0 | 2009.0 |
| GarageCars | 1.8 | 0.8 | 0.0 | 1.0 | 2.0 | 2.0 | 5.0 | 0.0 | 3.0 | 3.0 |
| GarageArea | 472.9 | 215.4 | 0.0 | 320.0 | 480.0 | 576.0 | 1488.0 | 0.0 | 857.0 | 1020.0 |
| WoodDeckSF | 93.7 | 126.5 | 0.0 | 0.0 | 0.0 | 168.0 | 1424.0 | 0.0 | 328.0 | 501.0 |
| OpenPorchSF | 47.5 | 67.6 | 0.0 | 0.0 | 26.0 | 70.0 | 742.0 | 0.0 | 184.0 | 285.0 |
| EnclosedPorch | 23.1 | 64.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1012.0 | 0.0 | 176.0 | 264.0 |
| 3SsnPorch | 2.6 | 25.2 | 0.0 | 0.0 | 0.0 | 0.0 | 508.0 | 0.0 | 0.0 | 144.0 |
| ScreenPorch | 16.1 | 56.2 | 0.0 | 0.0 | 0.0 | 0.0 | 576.0 | 0.0 | 161.0 | 260.0 |
| PoolArea | 2.3 | 35.7 | 0.0 | 0.0 | 0.0 | 0.0 | 800.0 | 0.0 | 0.0 | 0.0 |
| MiscVal | 50.8 | 567.4 | 0.0 | 0.0 | 0.0 | 0.0 | 17000.0 | 0.0 | 0.0 | 1000.0 |
| MoSold | 6.2 | 2.7 | 1.0 | 4.0 | 6.0 | 8.0 | 12.0 | 2.0 | 11.0 | 12.0 |
| YrSold | 2007.8 | 1.3 | 2006.0 | 2007.0 | 2008.0 | 2009.0 | 2010.0 | 2006.0 | 2010.0 | 2010.0 |
| SalePrice | 180921.2 | 79442.5 | 34900.0 | 129950.0 | 163000.0 | 214000.0 | 755000.0 | 88000.0 | 327000.0 | 446261.0 |

The DISCRIM Procedure

| Generalized Squared Distance to Foundation | | | | | | |
|---|---|---|---|---|---|---|
| From Foundation | BrkTil | CBlock | PConc | Slab | Stone | Wood |
| BrkTil | 79.28407 | 84.93631 | 85.54303 | 413650197 | 21789493 | 577875838 |
| CBlock | 88.99603 | 76.67014 | 77.79196 | 547706153 | 134508420 | 286605363 |
| PConc | 121.15666 | 86.29509 | 73.18215 | 824158816 | 392578031 | 165033092 |
| Slab | 126.91313 | 104.49402 | 166.70354 | 28.08952 | 132483119 | 697533184 |
| Stone | 81.94057 | 97.55344 | 97.05731 | 364445598 | -24.41060 | 655436936 |
| Wood | 127.06458 | 88.73381 | 79.10913 | 536089703 | 516319649 | -108.55165 |

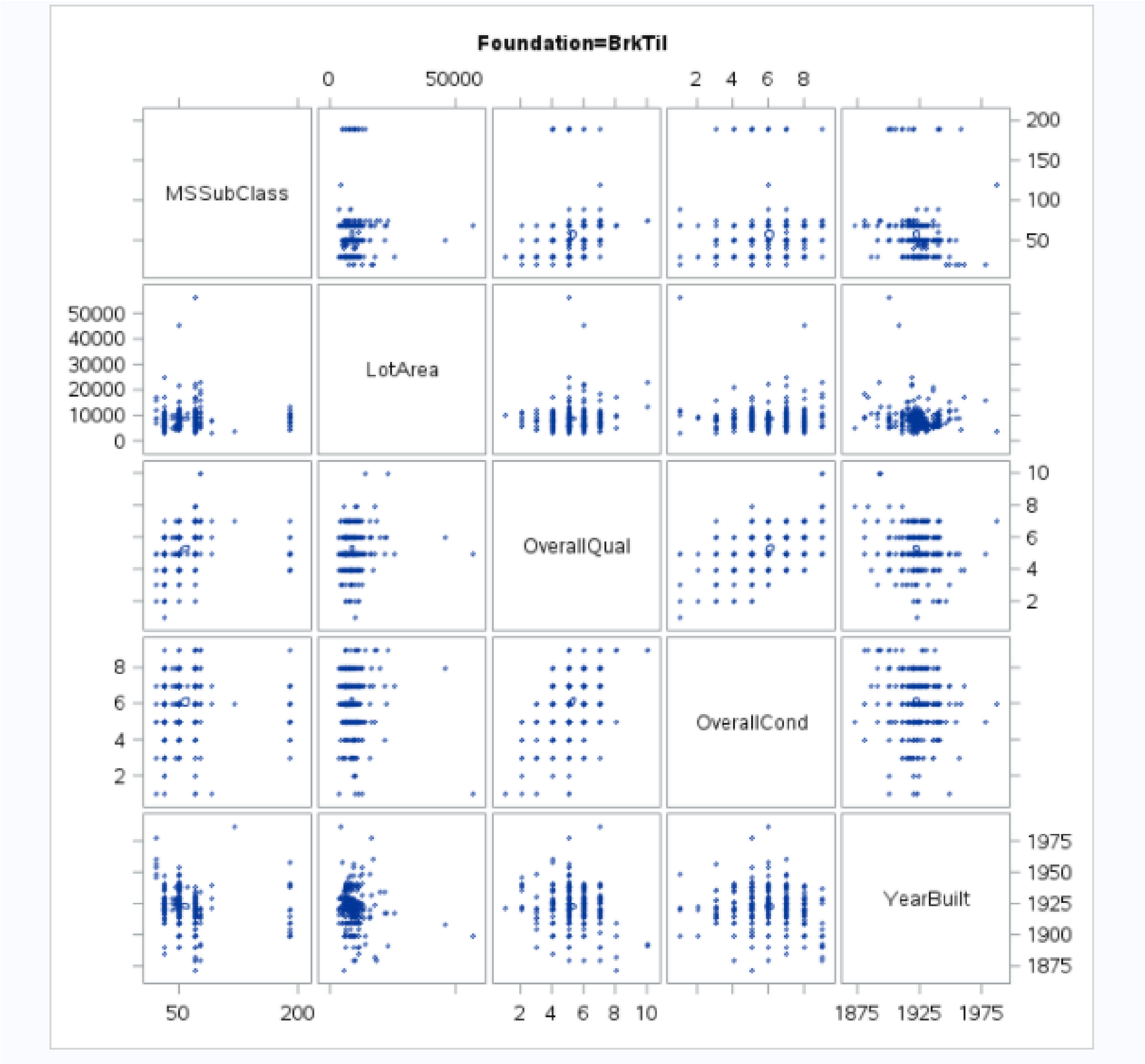
The DISCRIM Procedure

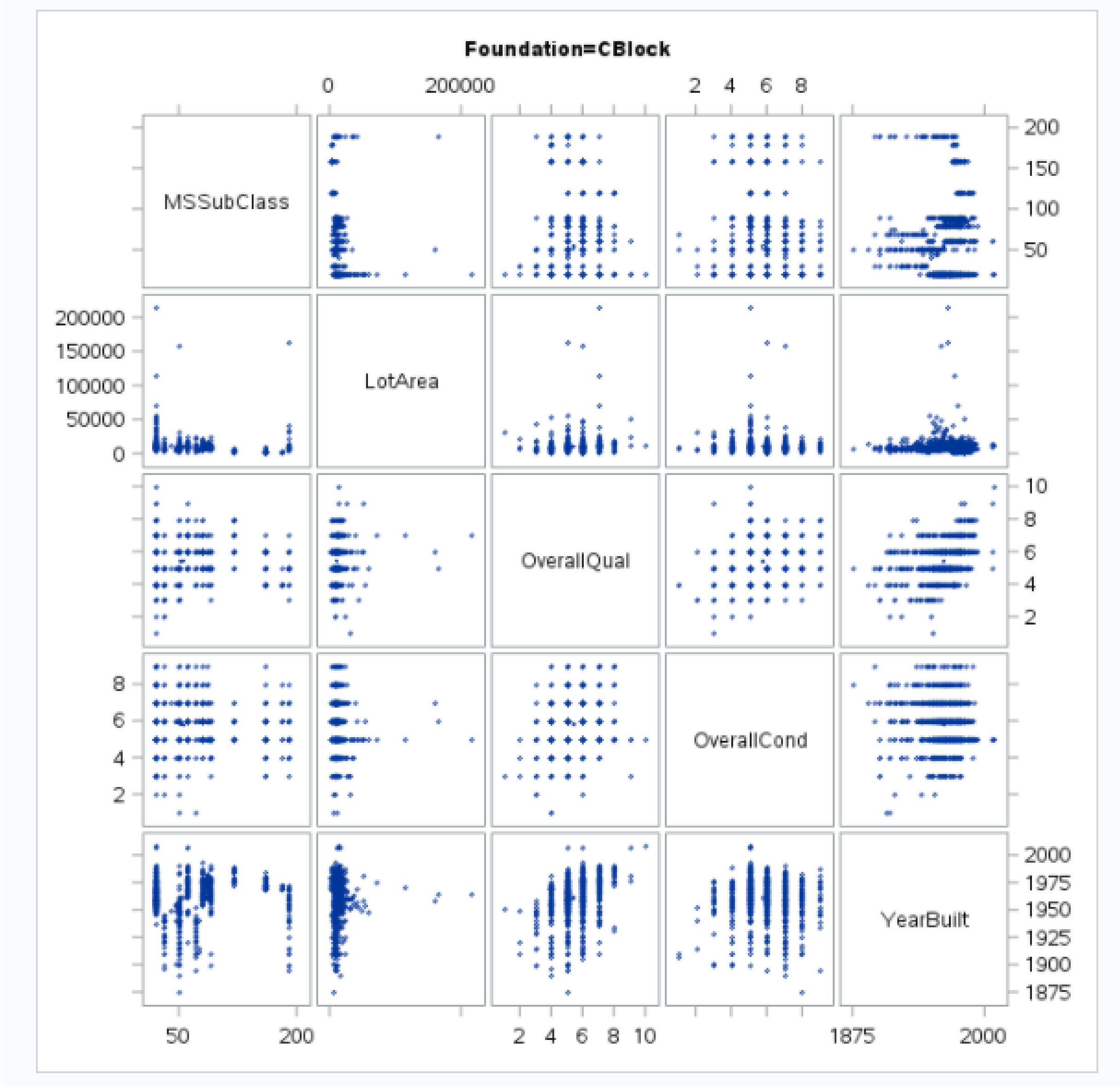
| | | | |
|---|---|---|---|
| Total Sample Size | 2759 | DF Total | 2758 |
| Variables | 17 | DF Within Classes | 2753 |
| Classes | 6 | DF Between Classes | 5 |

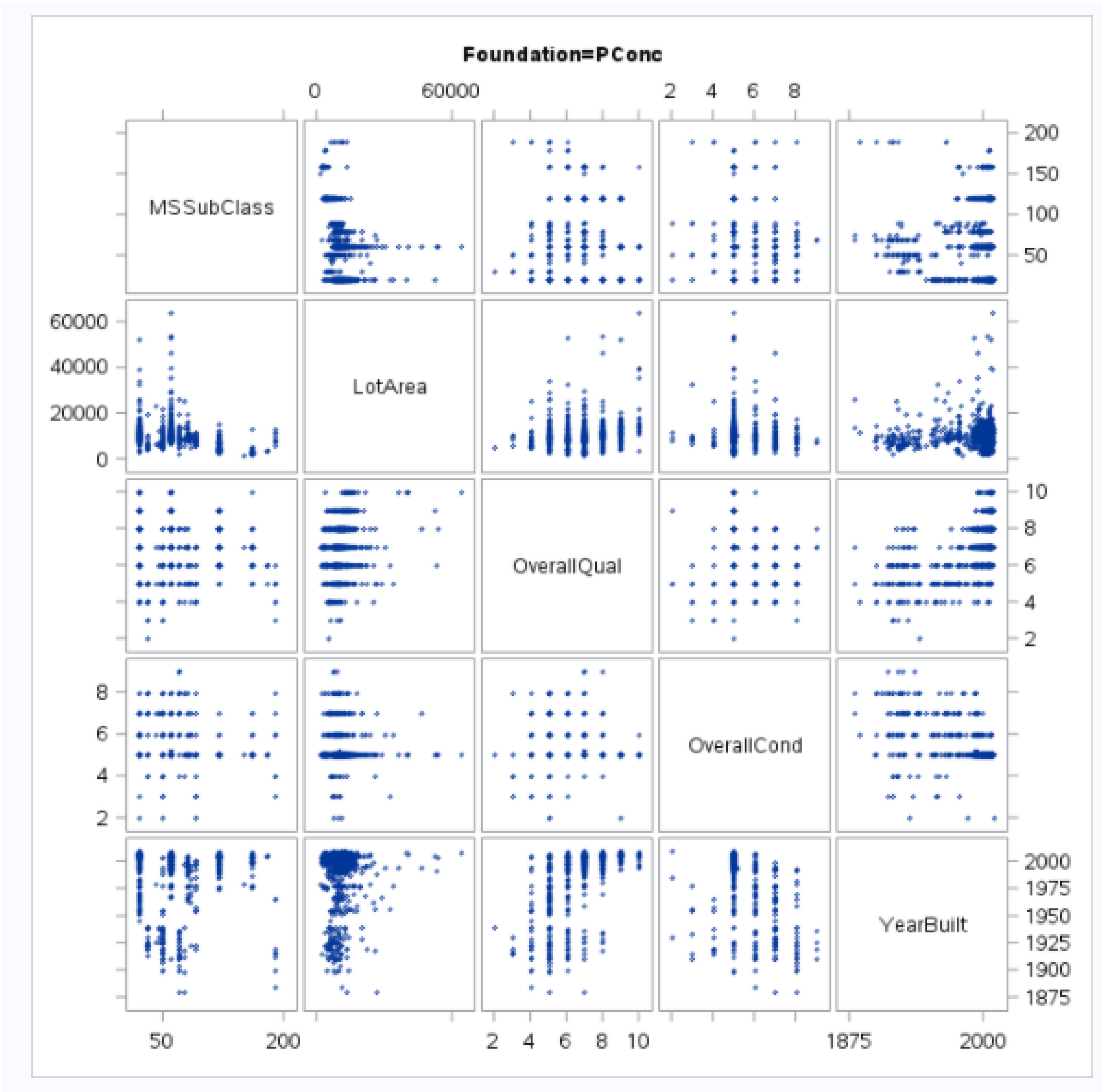
| | |
|---|---|
| Number of Observations Read | 2919 |
| Number of Observations Used | 2759 |

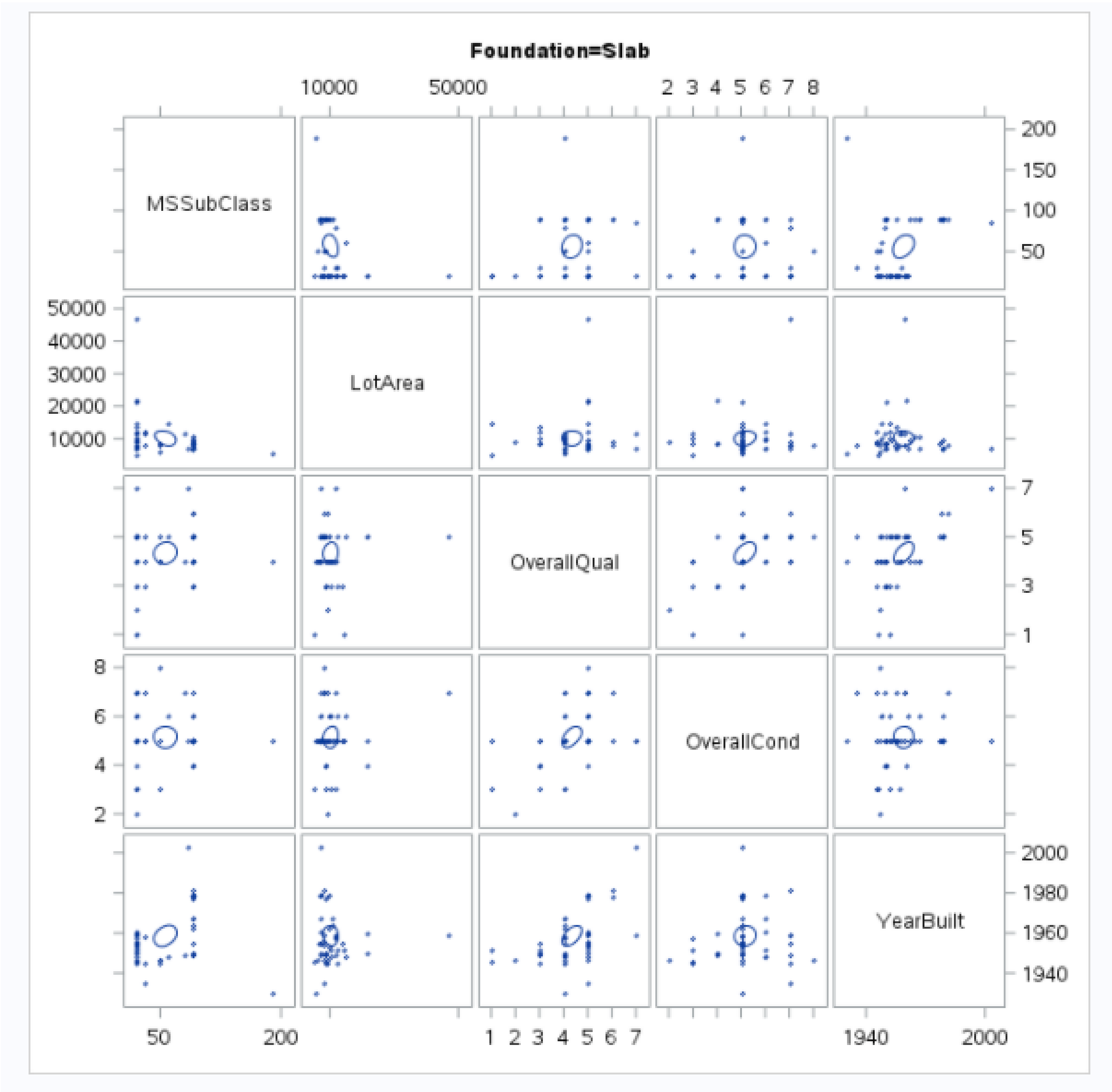
| Class Level Information | | | | | |
|---|---|---|---|---|---|
| Foundation | Variable Name | Frequency | Weight | Proportion | Prior Probability |
| BrkTil | BrkTil | 278 | 278.0000 | 0.100761 | 0.101790 |
| CBlock | CBlock | 1158 | 1158 | 0.419717 | 0.441556 |
| PConc | PConc | 1268 | 1268 | 0.459587 | 0.433657 |
| Slab | Slab | 40 | 40.0000 | 0.014498 | 0.016498 |
| Stone | Stone | 10 | 10.0000 | 0.003625 | 0.004300 |
| Wood | Wood | 5 | 5.0000 | 0.001812 | 0.002200 |

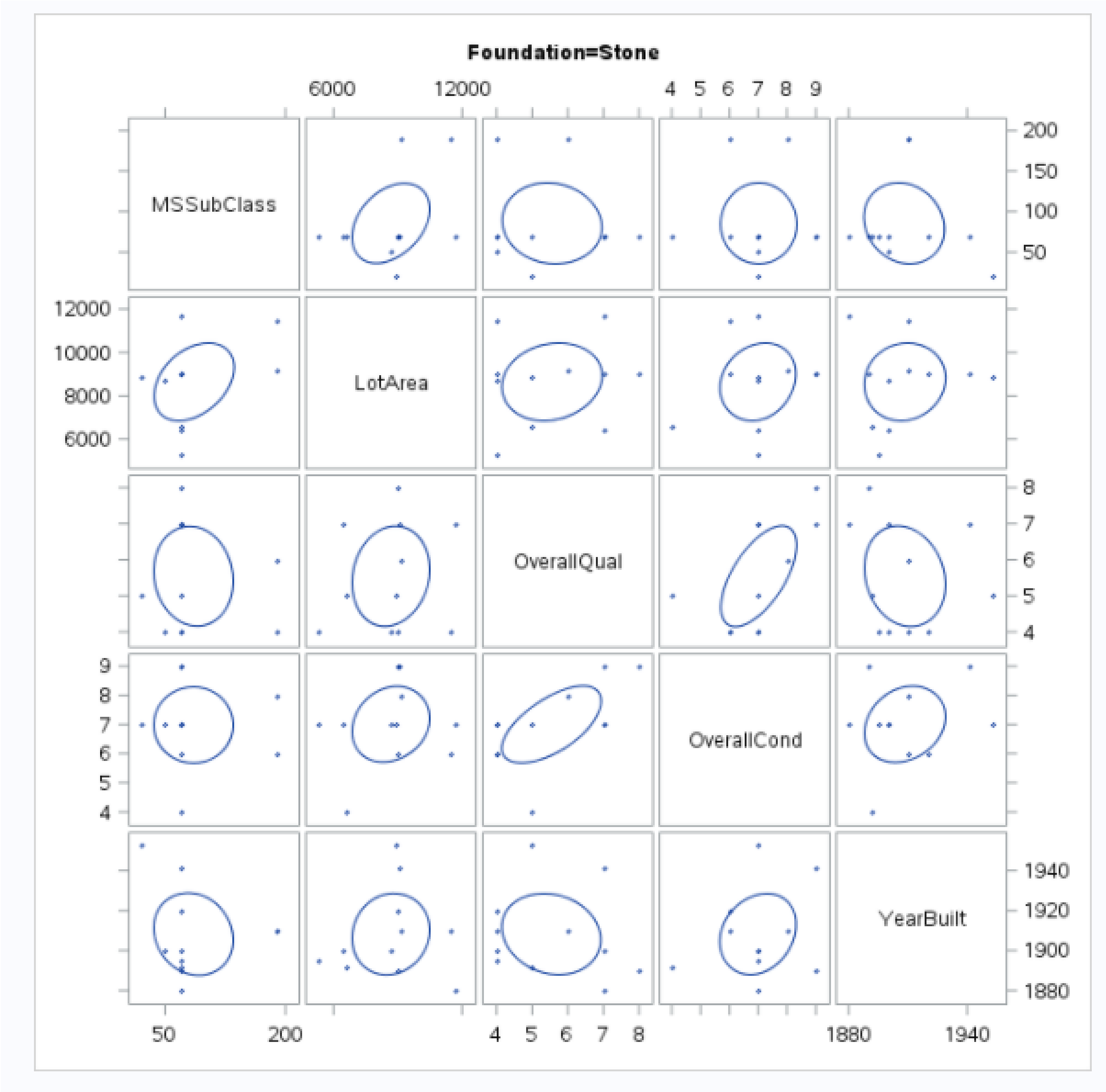
| Within Covariance Matrix Information | | |
|---|---|---|
| Foundation | Covariance Matrix Rank | Natural Log of the Determinant of the Covariance Matrix |
| BrkTil | 16 | 74.71438 |
| CBlock | 16 | 75.03524 |
| PConc | 16 | 71.51114 |
| Slab | 13 | 19.88053 |
| Stone | 9 | -35.30908 |
| Wood | 4 | -120.79045 |
| Pooled | 16 | 75.89413 |

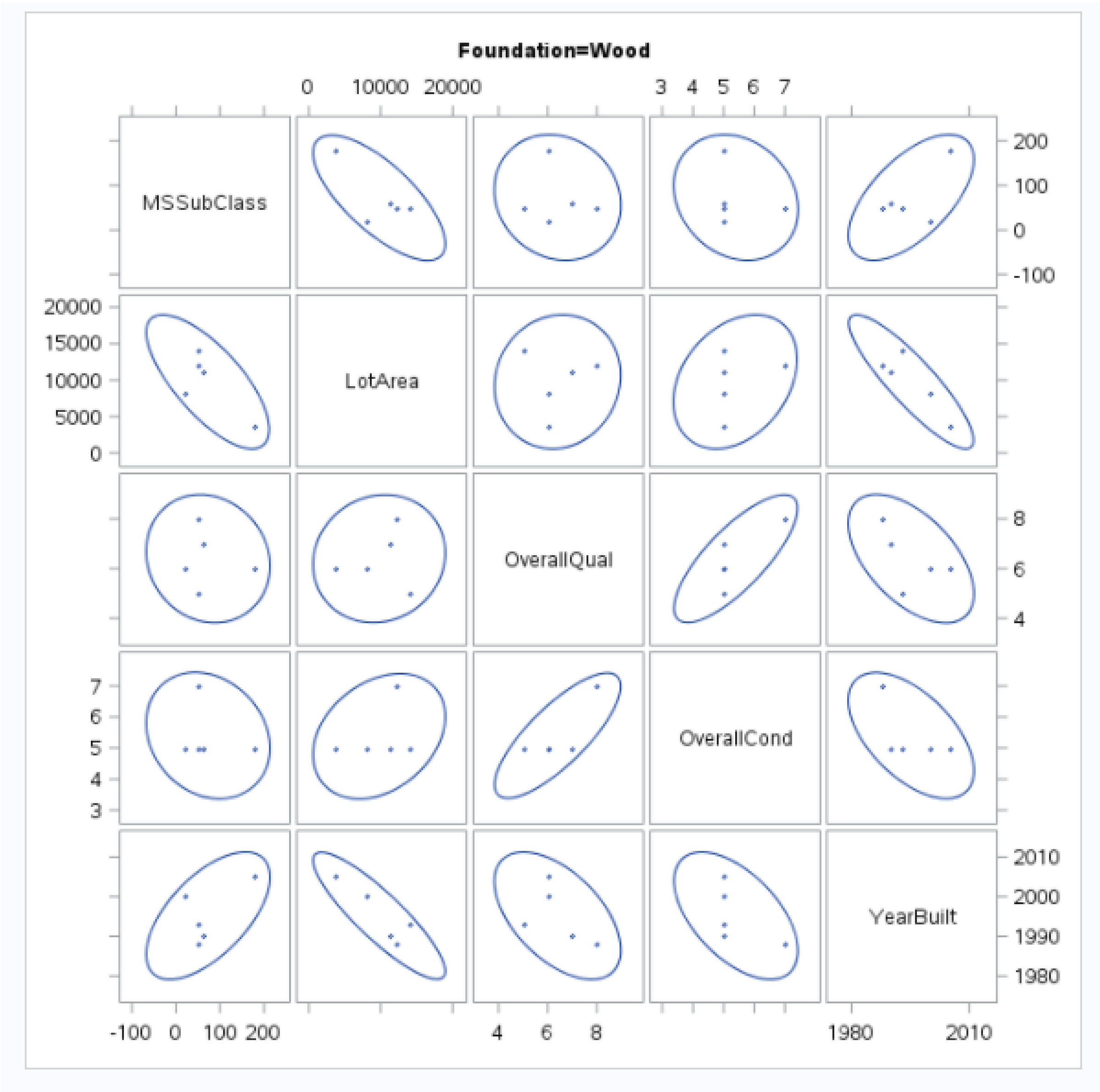

Foundation=BrkTil
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
0
50000
2
4
6
8
200
150
100
50
50000
40000
30000
20000
10000
0
10
8
6
4
2
8
6
4
2
1975
1950
1925
1900
1875
50
200
2
4
6
8
10
1875
1925
1975


Foundation=CBlock
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
0
200000
2
4
6
8
200
150
100
50
200000
150000
100000
50000
0
10
8
6
4
2
8
6
4
2
2000
1975
1950
1925
1900
1875
50
200
2
4
6
8
10
1875
2000


Foundation=PConc
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
0
60000
2
4
6
8
200
150
100
50
60000
40000
20000
0
10
8
6
4
2
8
6
4
2
2000
1975
1950
1925
1900
1875
50
200
2
4
6
8
10
1875
2000


Foundation=Slab
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
10000
50000
2 3 4 5 6 7 8
200
150
100
50
50000
40000
30000
20000
10000
7
5
3
1
8
6
4
2
2000
1980
1960
1940
50
200
1 2 3 4 5 6 7
1940
2000


Foundation=Stone
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
6000
12000
4 5 6 7 8 9
200
150
100
50
12000
10000
8000
6000
8
7
6
5
4
9
8
7
6
5
4
1940
1920
1900
1880
50
200
4 5 6 7 8
1880
1940


Foundation=Wood
MSSubClass
LotArea
OverallQual
OverallCond
YearBuilt
0
10000
20000
3
4
5
6
7
200
100
0
-100
15000
5000
8
6
4
2010
2000
1990
1980
-100
0
100
200
4
6
8
1980
2010