

**PREDICTIVE ANALYTICS**

FINAL PROJECT – RECOMMENDATION OF TOP 20 LUXURIOUS HOMES

Team A

Fall 2020

**Project advisor:**Dr. Xiao Ma

**Team Members:**  
Sean M Adams   
Toby John  
Stacey Terrell Hood   
Kevin Davis

**Table of Contents**

**I. Phase I: Business/Research Understanding Phase 2**

* Objective 2
* Background and goal 2

**Il. Phase 2: Data Understanding Phase 4**

* Variable Description 4
* Data file description 6
* Master Data Preparation: Importing Data into SAS EM 8
* Graphical representation of data base 11
* Master Data Preparation and EDA 12
* Score Data Preparation and EDA 23

**Ill. Phase 3: Data Preparation 26**

* Assignment of Roles 26
* Data Partitioning 26
* Data Transformation 27

**IV. Phase 4: Modeling Phase 28**

* Decision Tree 28
* Neural Network 30
* *K*-Nearest Neighbor 34
* Regression 36

**V. Phase 5: Evaluation Phase 39**

* Model Results and Best Model Selection 39
* Variables Included in Final Sales Price Model 39

**VI. Phase 6: Deployment Phase 40**

* Results: Top 20 Luxury Homes 40
* Explanation of Results 40

**VI. Final Model Diagram 42**

**1. Business/Research Understanding Phase (Phase 1)**

**1.1 Objective**

The main goal is to recommend for investment the purchase of the twenty (20) most luxurious (i.e. priciest) homes out of 100 candidate properties for which the sales prices are not known.

**1.2 Background and the Goal**

Our group is acting as a set of investors who plan to purchase 20 residential properties in the city where the raw data files were collected. The broker of the properties gave us a list of 100 properties and their information as your Score data set, consisting of many assessed values for them. However, this “Score” data set does not contain the final sale prices or a couple of other variables. We will develop models and appropriate procedures to predict and rank the sale price and recommend the 20 most luxurious houses. The reason for investing in the top luxury houses is that their potential investment growth is believed to be the greatest.

**2. Data Understanding Phase (Phase 2)**

**2.1 Variables Description**

1. PID (Property Identification) – a unique number to identify each property.

2. Lot Area – The size of the lot, measured in square feet, on which the house is located.

3. Lot Shape – The general shape of the lot. A lot with a regular shape has a value of 1, and another with not a regular shape has a value of 0.

4. Bldg Type (i.e., Building Type) – This describes the type of home in terms of its footprint. A single-family detached type of home is indicated by a value of 1, and a townhouse type of home is indicated by a value of 0.

5. Overall Quality – This is a rating of the overall material and finish of the house. The numeric scale of this rating is as follows.  
10 - Very Excellent  
9 – Excellent  
8 - Very Good  
7 -Good  
6 - Above Average  
5 – Average  
4 - Below Average  
3 – Fair  
2 – Poor  
1 - Very Poor

6. Overall Condition: This is a rating of the overall condition of the house. The numeric scale of this rating is as follows.

10 - Very Excellent  
9 – Excellent  
8 - Very Good  
7 – Good  
6 - Above Average  
5 – Average  
4 - Below Average  
3 – Fair  
2 – Poor  
1 - Very Poor

7. Exterior Quality – This is a rating of the quality of the material on the exterior. A good quality is indicated by a 1, and an average quality is indicated by a 0.

8. Foundation – This describes the type of foundation upon which the house is built. A concrete foundation is indicated by a value of 2; a cinder-block foundation by a value of 1; and brick foundation by a value of 0.

9. Year Built – This describes the year when the house was constructed.

10. Year Remodel – This describes the year when the house was remodeled. If the house was never remodeled, then the “year remodel” is the same as the “year built.”

11. Veneer Area of Exterior Wall – This describes the area in square feet of the exterior wall that is veneer.

12. Bsmt Finish Type (Basement Finished Type ) – This indicates whether a home’s basement is finished or not in the sense that it can be lived in or not. When it is finished, it has a value of 1, and a value of 0 otherwise.

13. Basement Finished Sqr ft – This is the measure of the area of a finished basement.

14. Basement Unfinished Sqr ft – This is the measure of the area of an unfinished basement.

15. Total Bsmt Sqr ft – This is the measure of the total basement area.

16. Heating QC (Heating Quality Condition) – This is a measure of the rating of how well the heating unit is for a house. The rating scale is as follows.

3 – Excellent  
2 – Good  
1 – Average  
0 - Fair

17. 1st Flr Sqr ft (First floor Sqt ft) – This is a measure of the living space on the first floor of a house.

18. 2nd Flr Sqr ft (Second floor Sqt ft) – This is a measure of the living space on the second floor of a house.

19. Above Ground Living Area – This is a measure of the living space of the entire house, excluding the basement.

20. Number Full Bath Bsmt - This indicates the number of full bathrooms in the basement of a house. A value of 1 indicates that there is a full bathroom and a value of 0 indicates that there is not a full bathroom in the basement.

21. Half Bath House - This indicates whether there is a half bathroom in the house (excluding the basement). A value of 1 indicates that there is a half bathroom and a value of 0 indicates that there is not a half bathroom in the house.

22. Number Full Bath House - This indicates the number of full bathrooms there are in the house, not including bathroom in the basement.

23. Bedroom Above Ground - This indicates the number of bedrooms there are in the house, not including the basement.

24. Number Room Above Ground - This indicates the number of rooms there are in the house, not including the basement.

25. Fireplaces – This indicates the number of fireplaces there are in the house, not including the basement.

26. Garage Type – Whether there is a garage of a given type is described and indicated as follows.

3 - Attached to house  
2 - Built-In (Garage part of house - typically has room above garage)  
1 - Detached from home  
0 - No garage

27. Garage Cars – This indicates the number of cars that can be accommodated in the garage of the house.

28. Garage Area – This is the size of garage in square feet.

29. Wood Deck Sqr ft – This is the size of the wood deck area in square feet for a house.

30. Open Porch Sqr ft - This is the size of the open porch area in square feet for a house.

31. Sale Price – This is the sale price of a house (not included in the Score data set).

**2.2 Data files Description**

1. Property Survey – 1: Contains 600 rows with the following variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type

2. Property Survey – 2: Contains 1770 rows with the following variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type

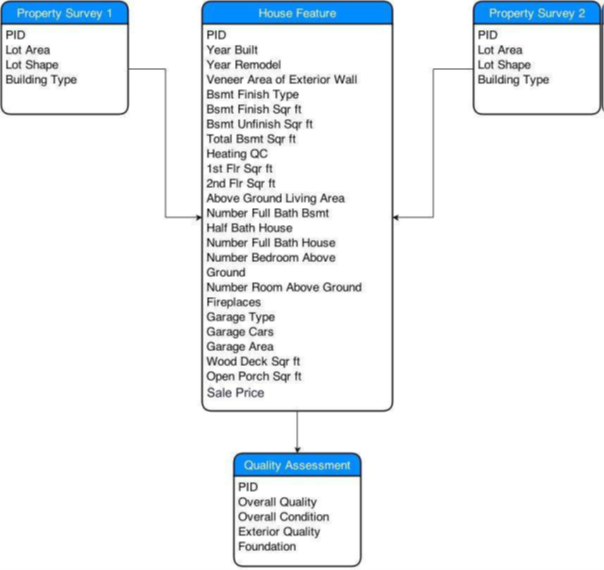
3. Quality Assessment Variables: PID (Property Identification), Overall Quality, Overall Condition, Exterior Quality, and Foundation

4. House Feature

Variables: PID (Property Identification), Year Built, Year Remodel, Veneer Area of Exterior Wall , Bsmt Finish Type, Bsmt Finish Sqr ft, Bsmt Unfinish Sqr ft, Total Bsmt Sqr ft, Heating QC, 1st Flr Sqr ft, 2nd Flr Sqr ft, Above Ground Living Area, Number Full Bath Bsmt, Half Bath House, Number Full Bath House, Number Bedroom Above Ground, Number Room Above Ground, Fireplaces, Garage Type, Garage Cars, Garage Area, Wood Deck Sqr ft, Open Porch Sqr ft, Sale Price

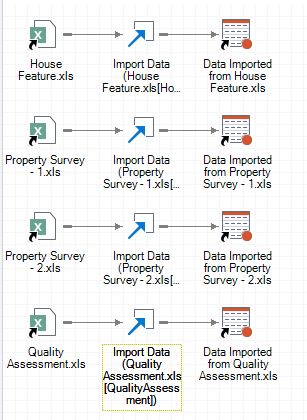
5. Score Data - No Sale Price

Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type, Overall Quality, Overall Condition, Year Built, Year Remodel, Veneer Area of Exterior Wall , Bsmt Finish Type, Bsmt Finish Sqr ft, Bsmt Unfinish Sqr ft, Heating QC, 1st Flr Sqr ft, 2nd Flr Sqr ft, Above Ground Living Area, Number Full Bath Bsmt, Half Bath House, Number Full Bath House, Number Bedroom Above Ground, Number Room Above Ground, Fireplaces, Garage Type, Garage Cars, Garage Area, Wood Deck Sqr ft, Open Porch Sqr ft.

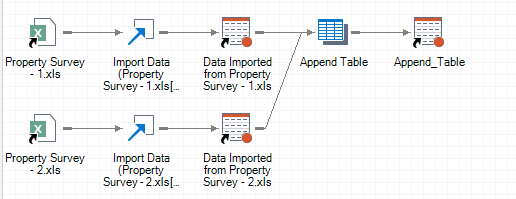


**2.3 Importing the data into SAS EM**

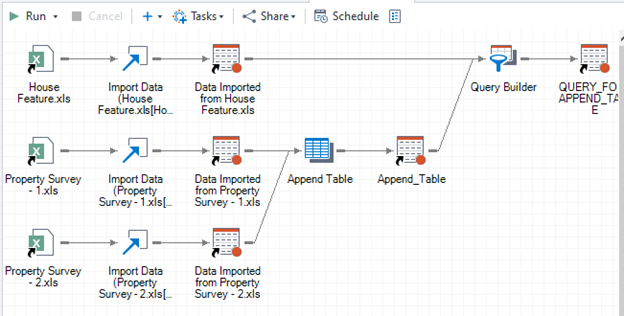
First, the different original data sets for the training data must be imported into SAS EG and then the separate tables need to be joined. Below, we have shown the first step in the process, importing the tables into SAS EG.



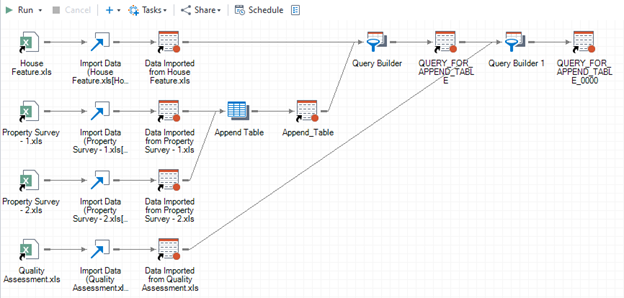
Next, we must start to append the tables to each other in order to build the final data set for training. Below I have shown the process to do this.



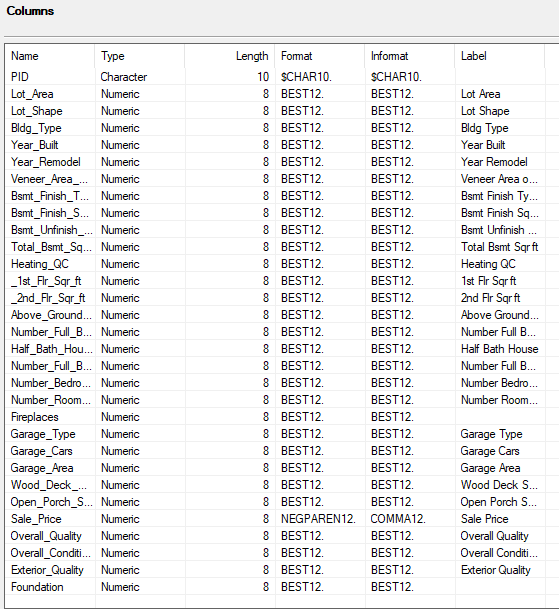
First, we joint the two Property Survey files using the built-in functions.



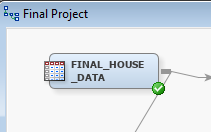
Then we join the newly created table with the House Feature table, making sure that we only keep one of the ID columns.



Finally, we merge this larger House Feature table with the Quality Assessment table to create our final data table to use for training. Again, we make sure there are no duplicate columns. Below is a list of all the variables from the final data table.

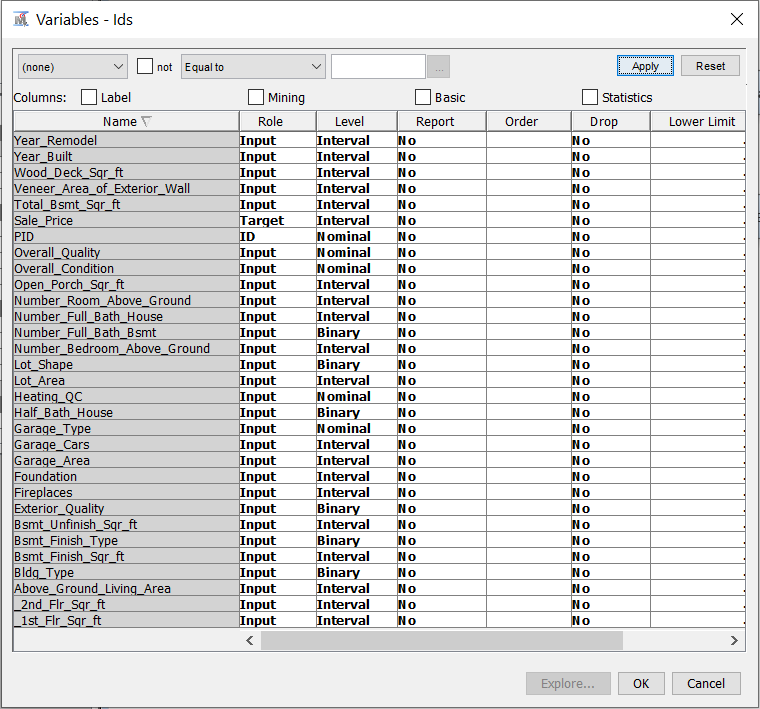


After this has been done, we Exported the resulting final table to an SAS file so that it can be imported into SAS EM. So, we open SAS EM, create a new project, assign the proper library, create a diagram for our project tasks, and then create the data source.

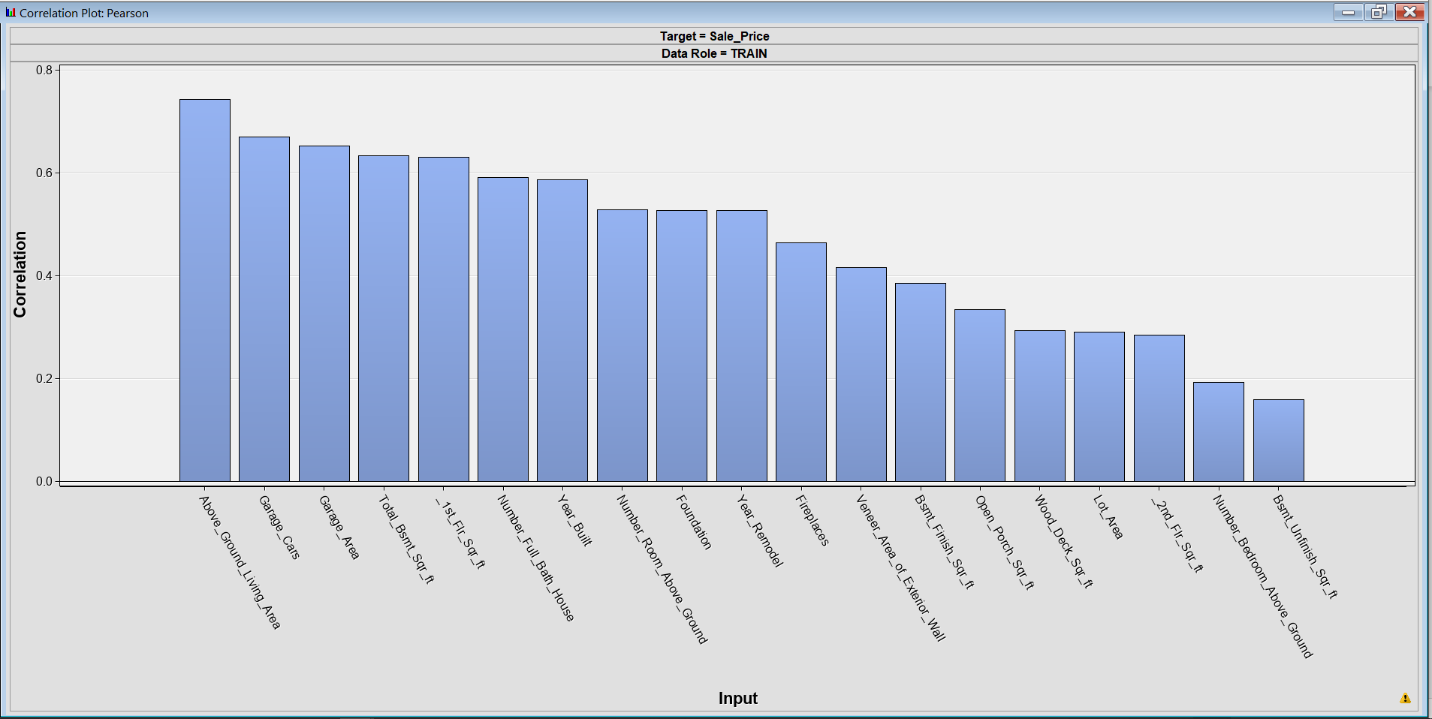


**2.4 Graphical representation of the data base**

Below is an image showing all the roles and levels chosen for the training and validation data sets. This shows the variables we have chosen to be binary, nominal, and interval to start this process.



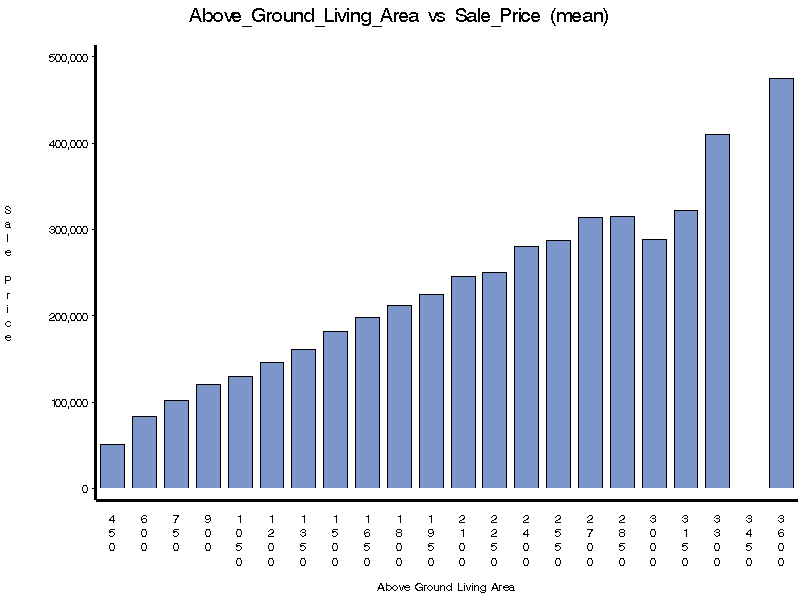
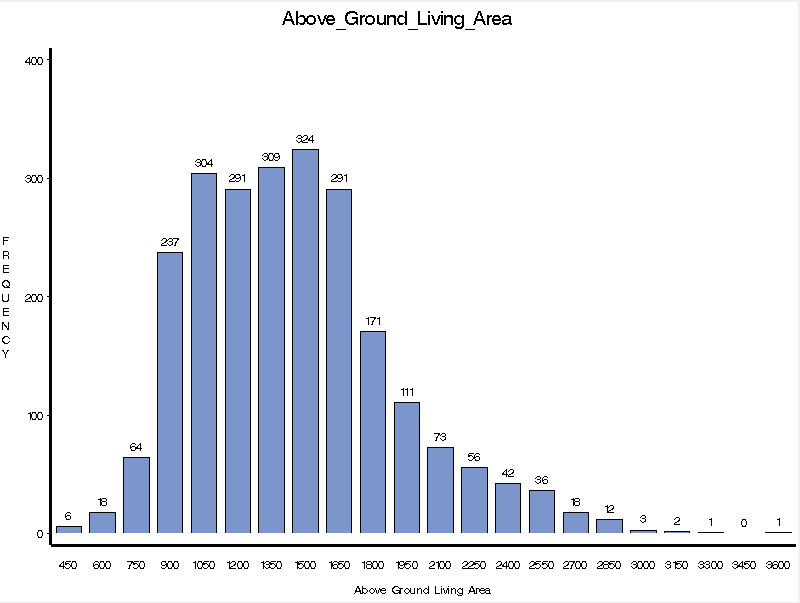
And here is a quick representation of the correlation between the different input variables and the target variable of Sale Price. As we can see, the variable with the highest correlation here is Above Ground Living Area.



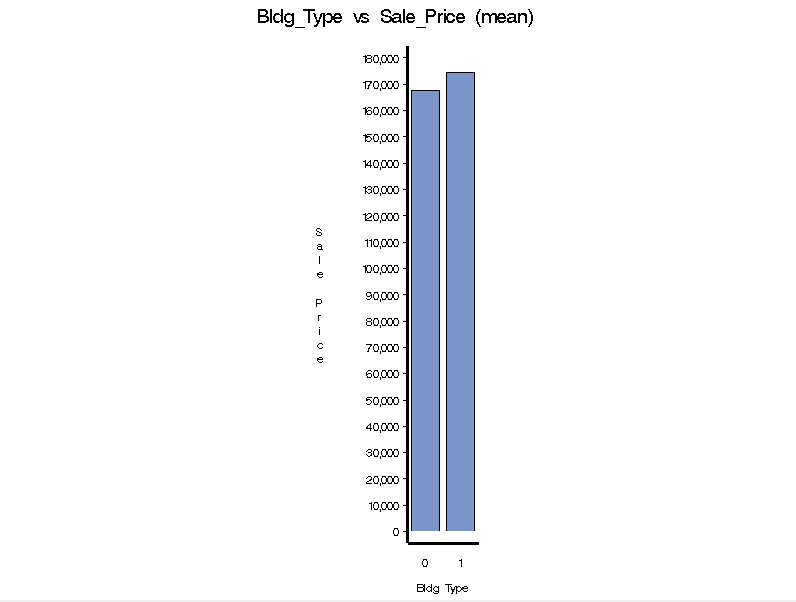
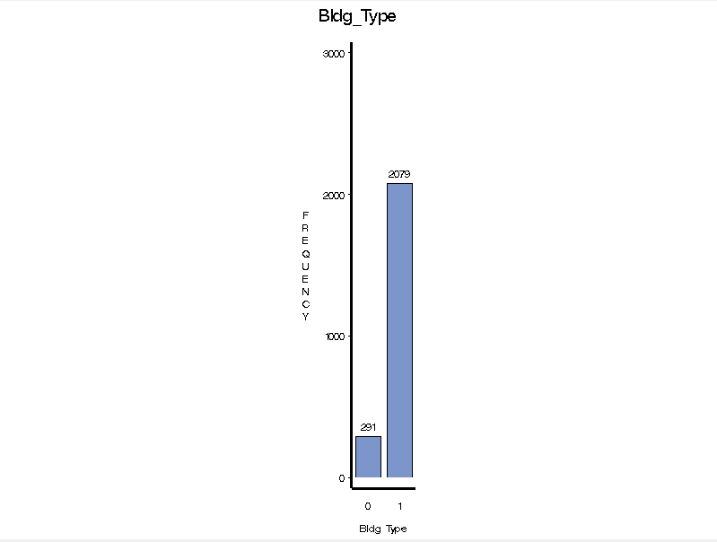
**2.5 Master Data Preparation and EDA**

Below, we have graphs showing our exploration of all the different variables from the training data set.

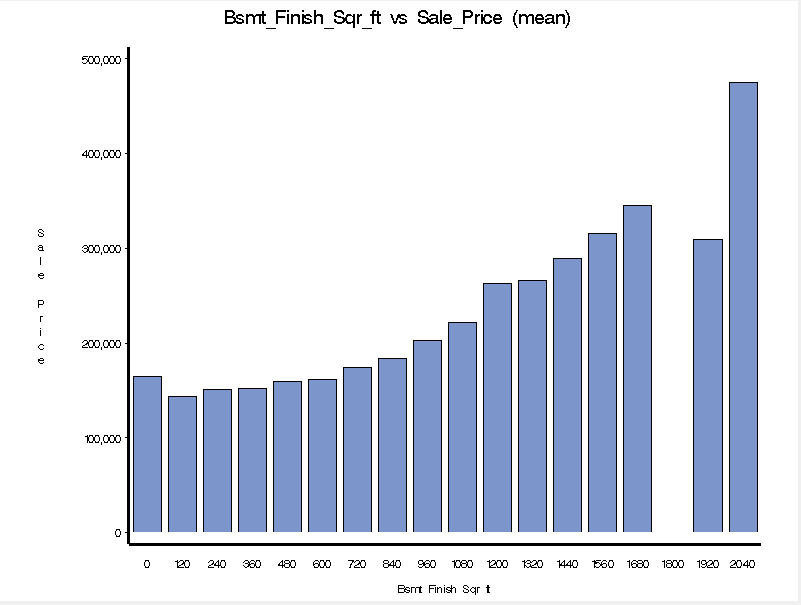
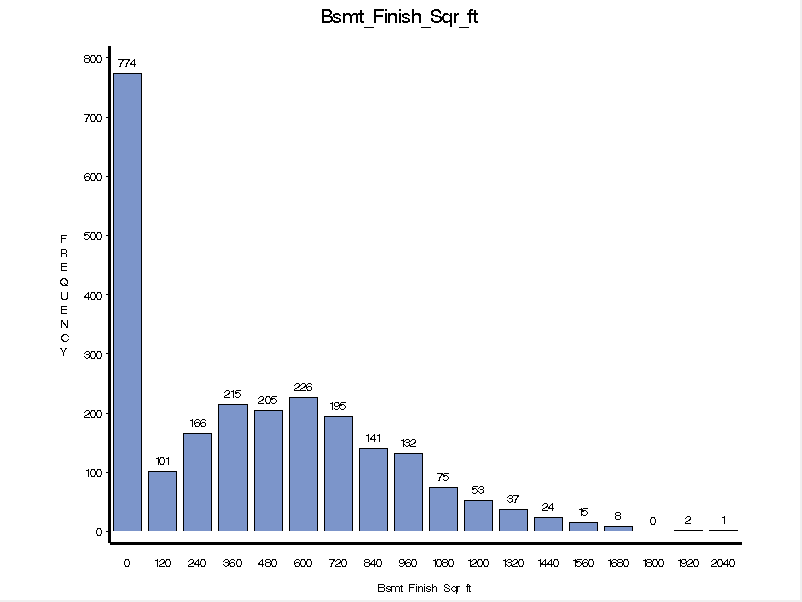
Living area (above ground)

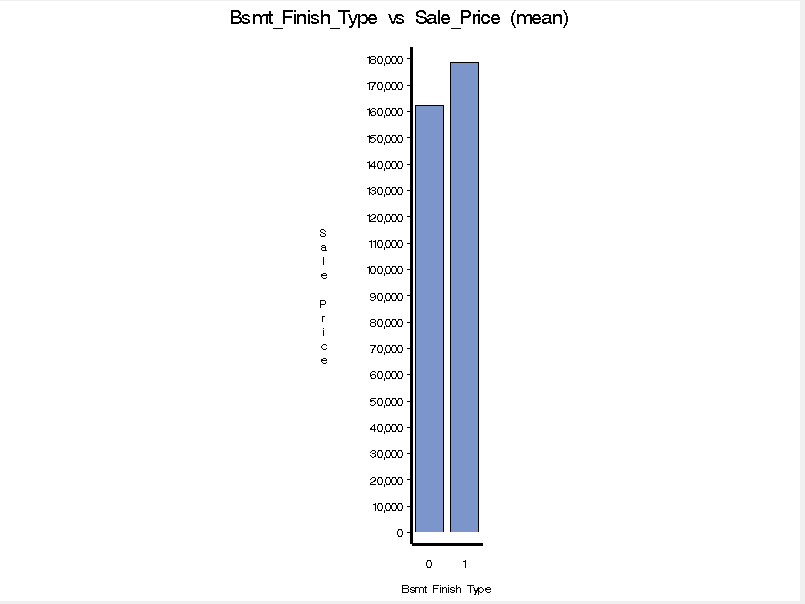
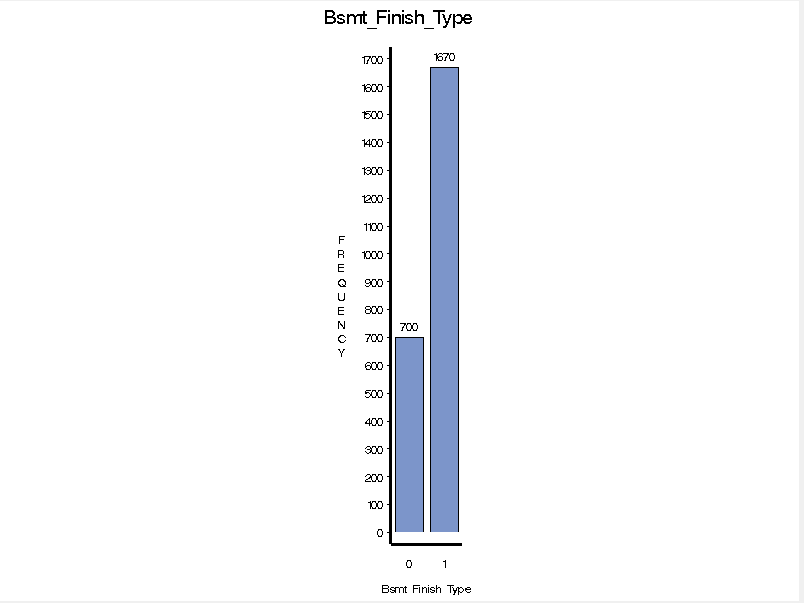
Building type

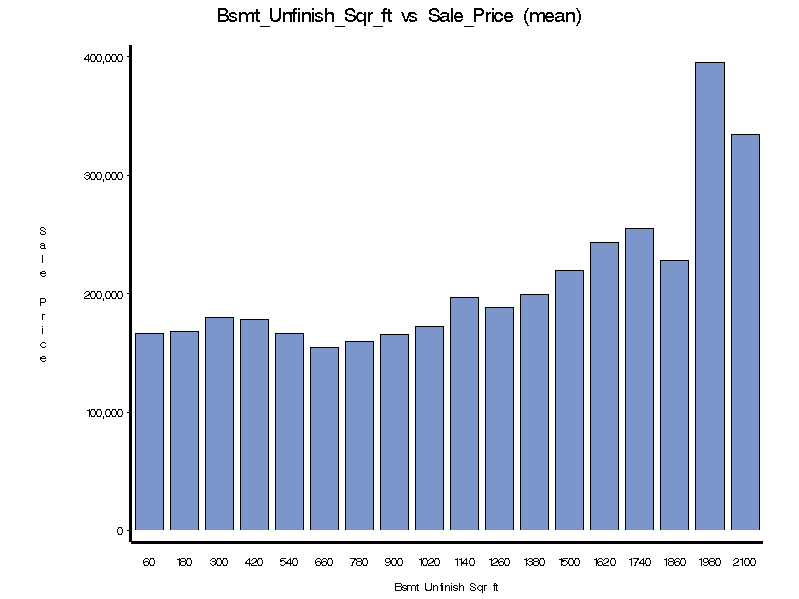
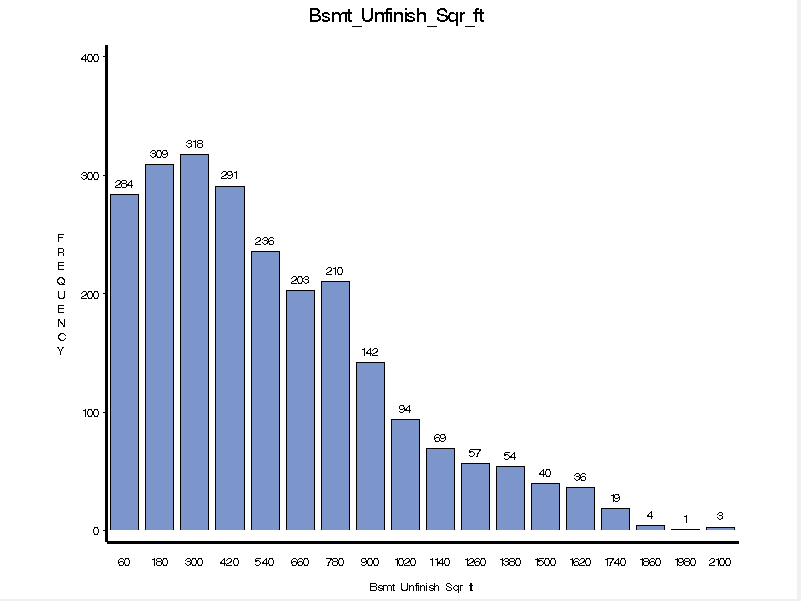
Basement area finished (sq. ft.)

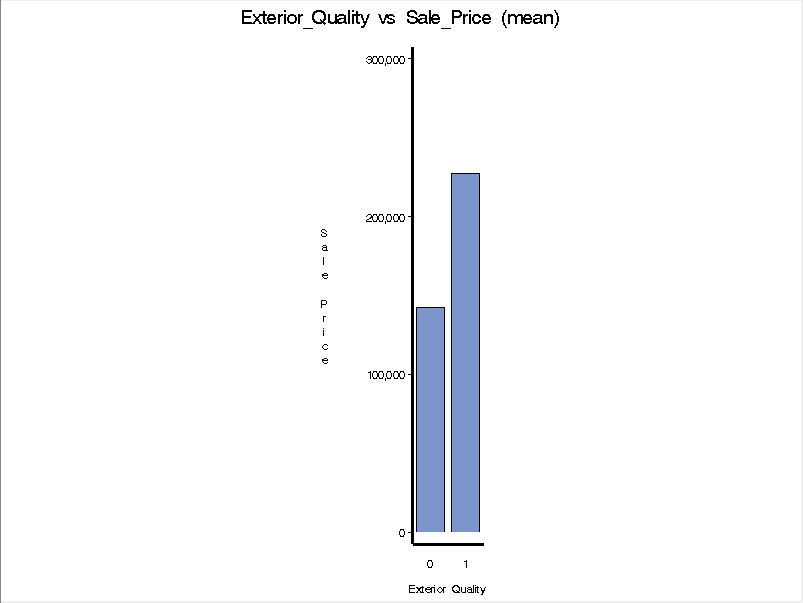
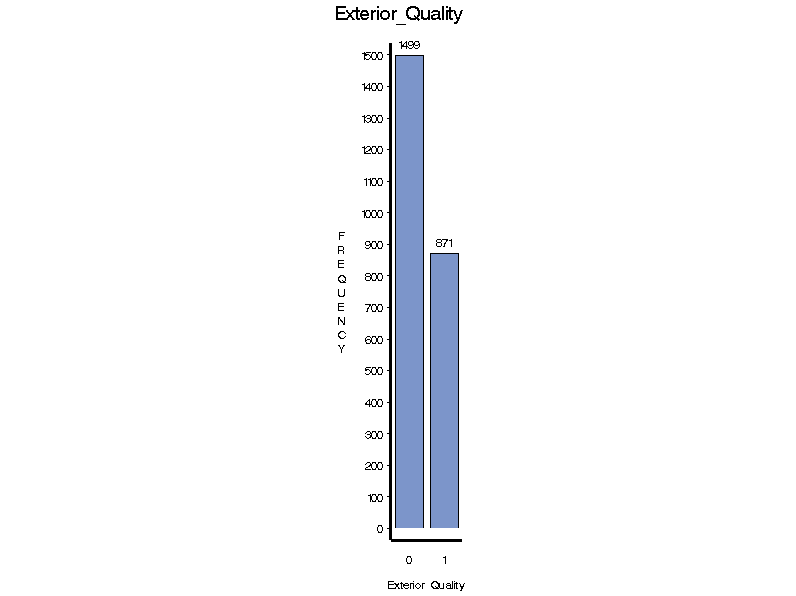
Basement finish type

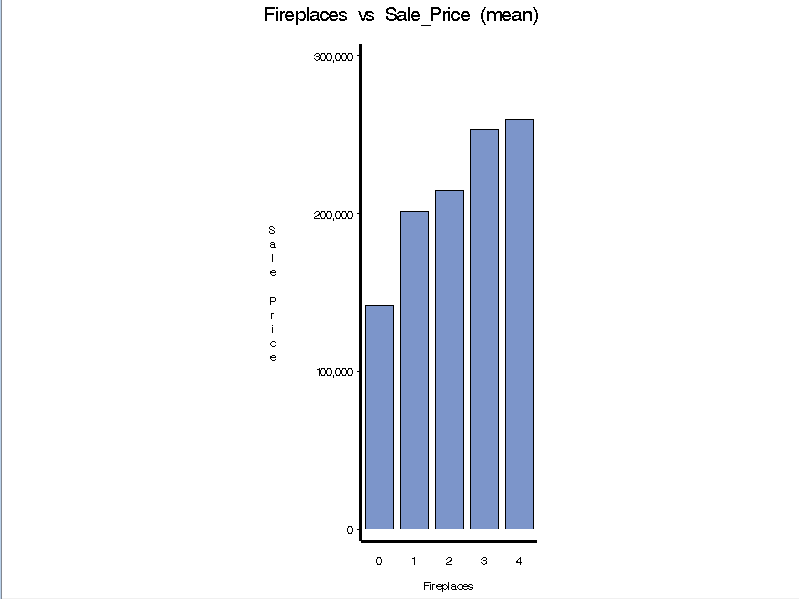
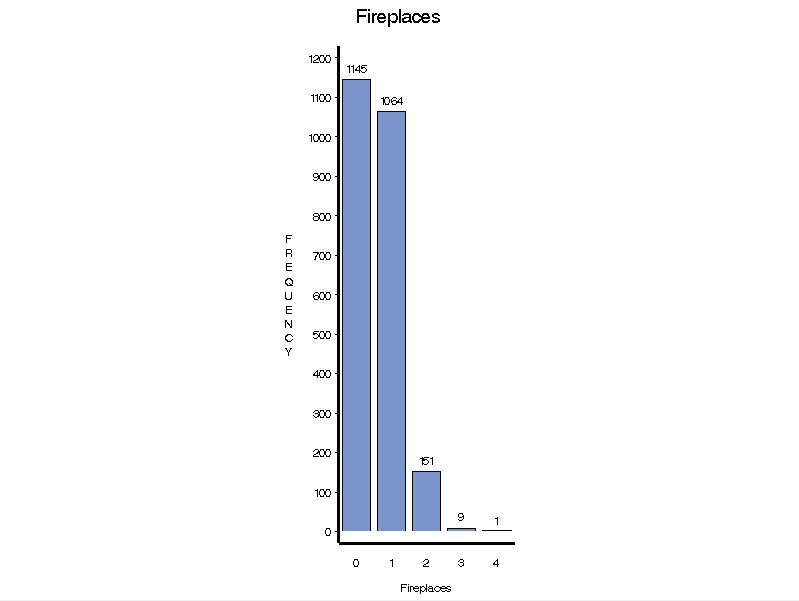
Basement area unfinished (sq. ft.)

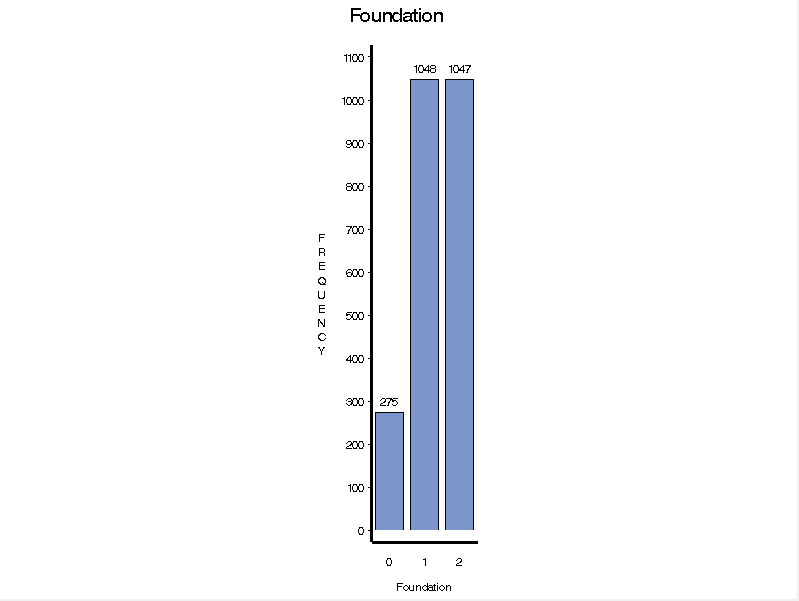
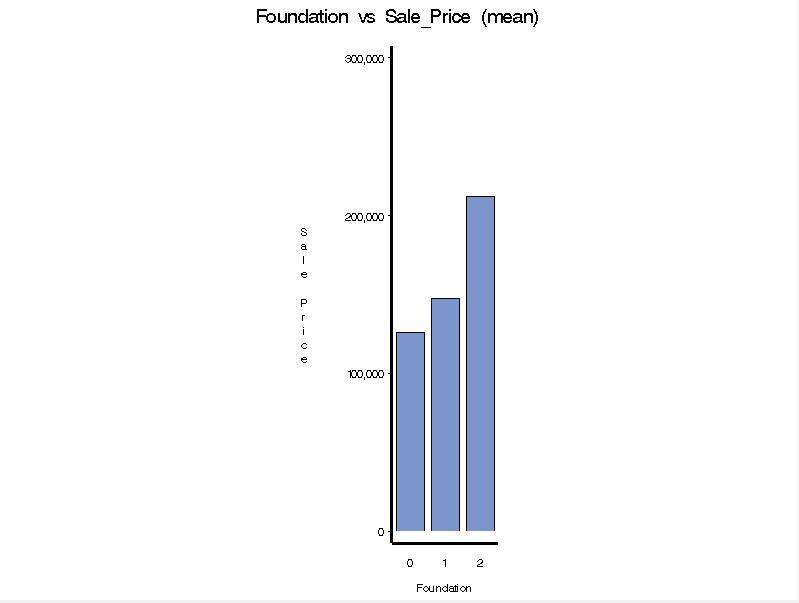
Exterior quality

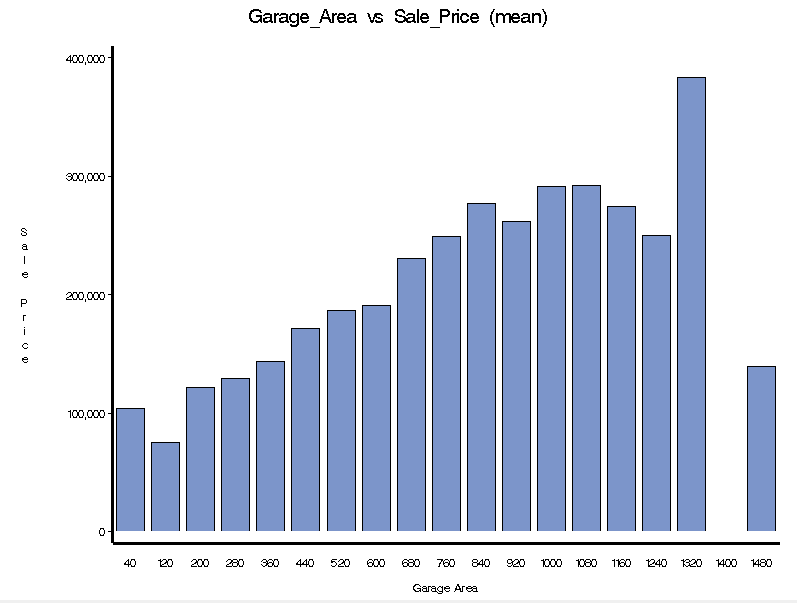
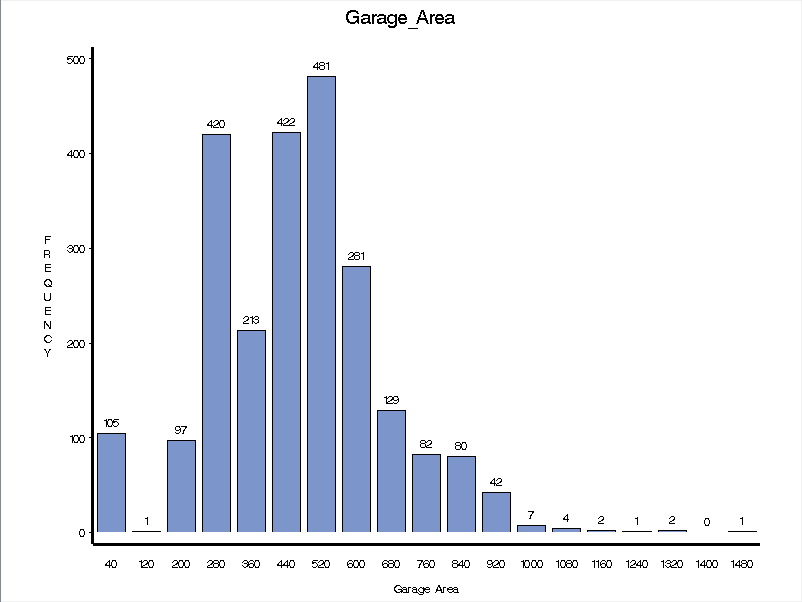
Fireplace

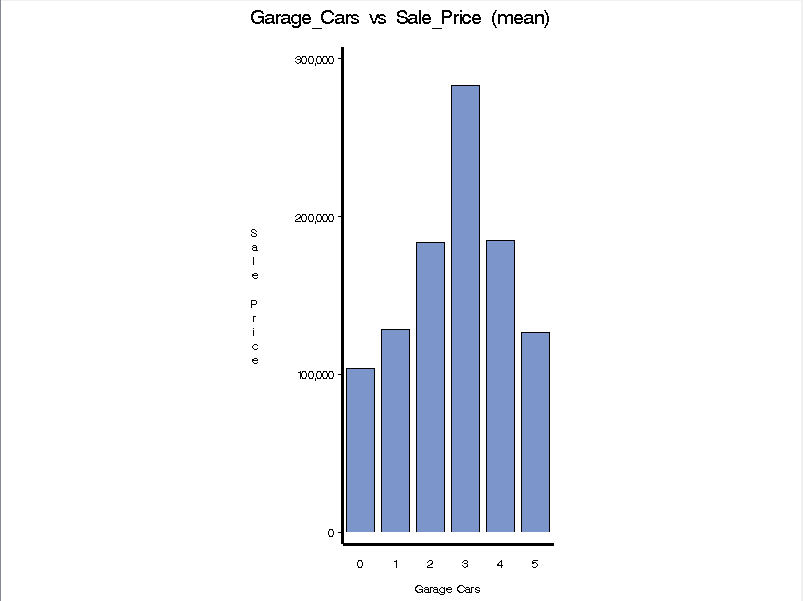
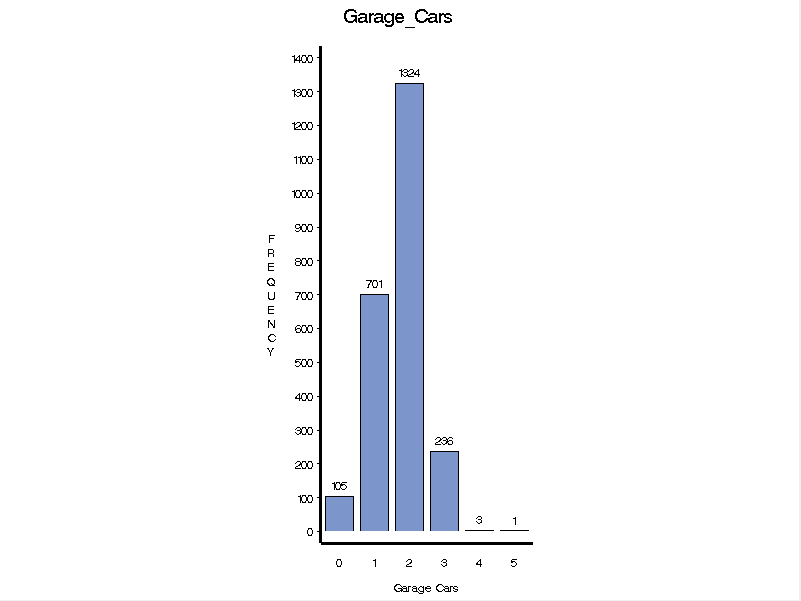
Foundation



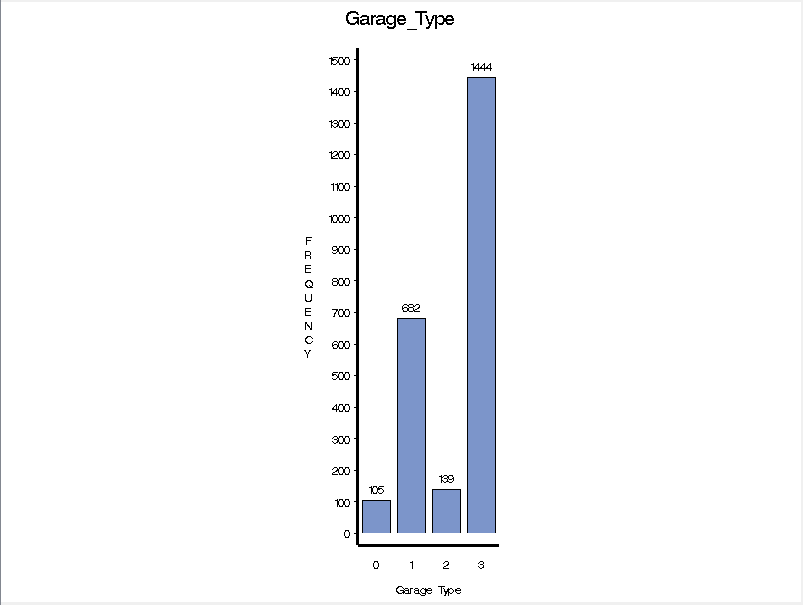
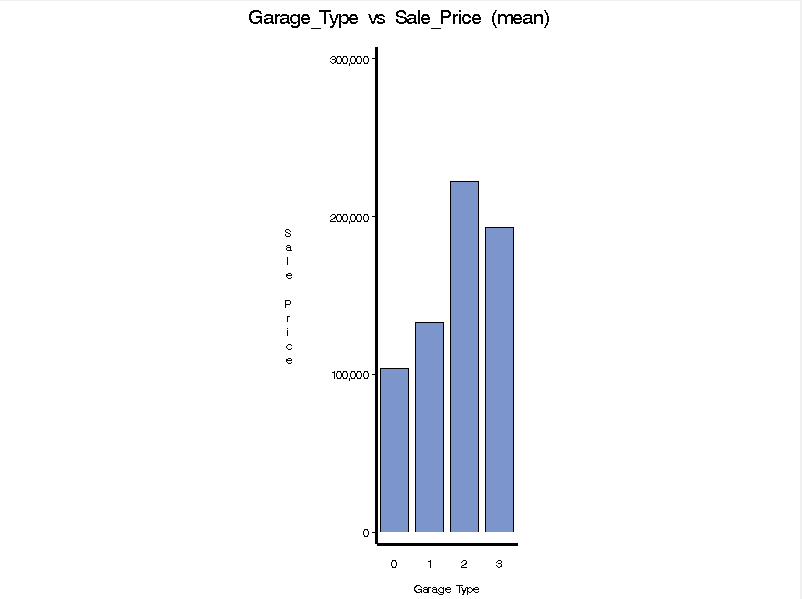
Garage area

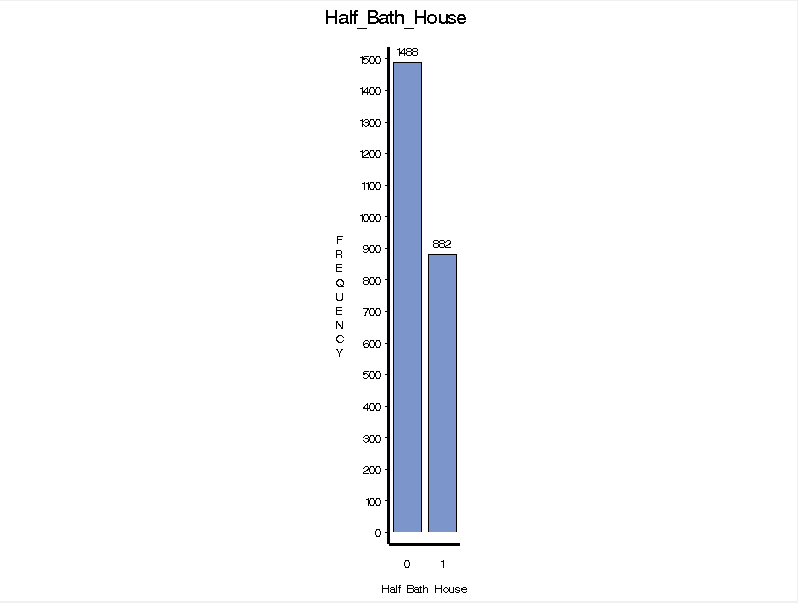
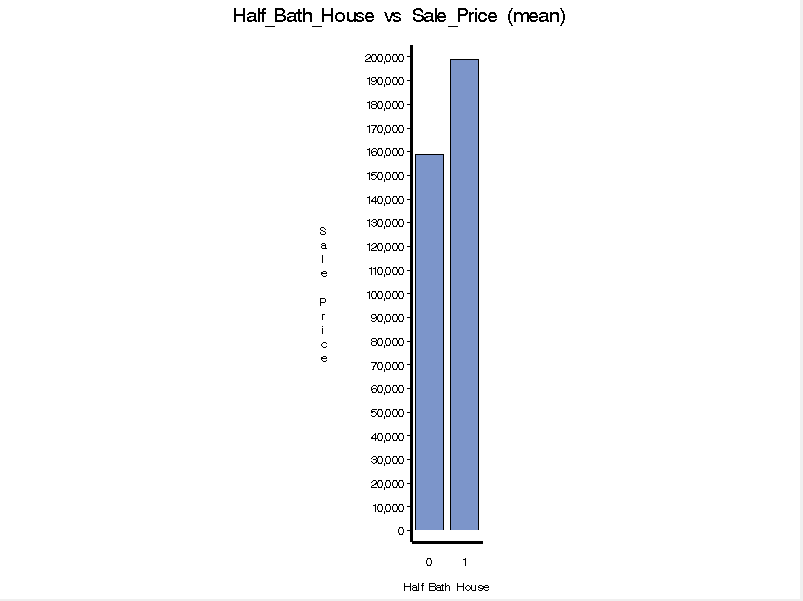
Garage cars

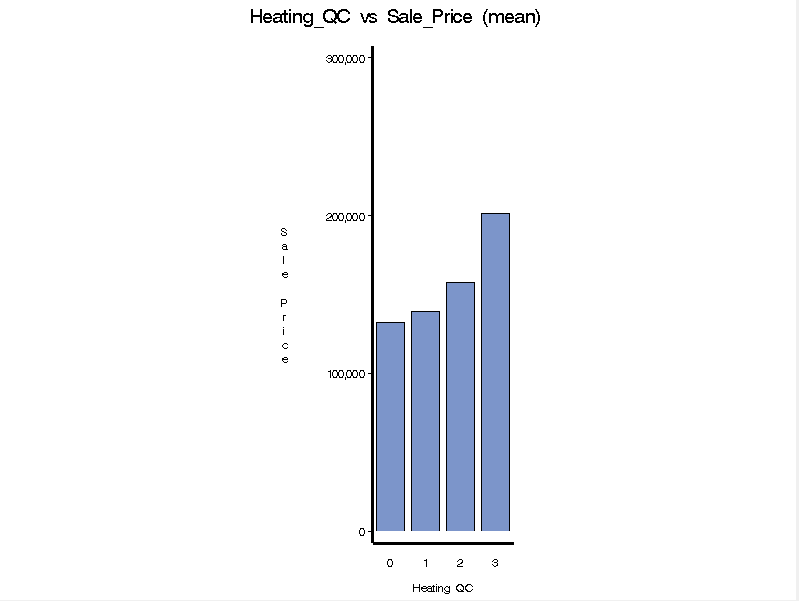
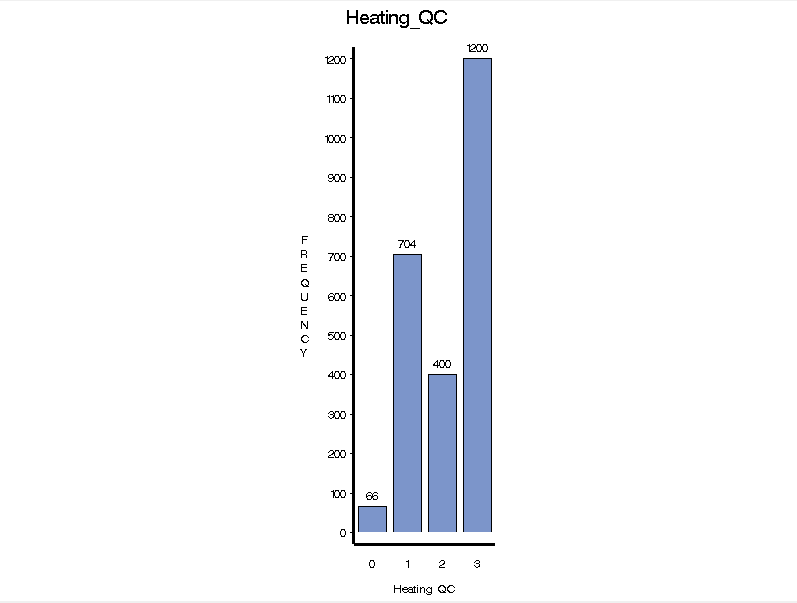
Garage type



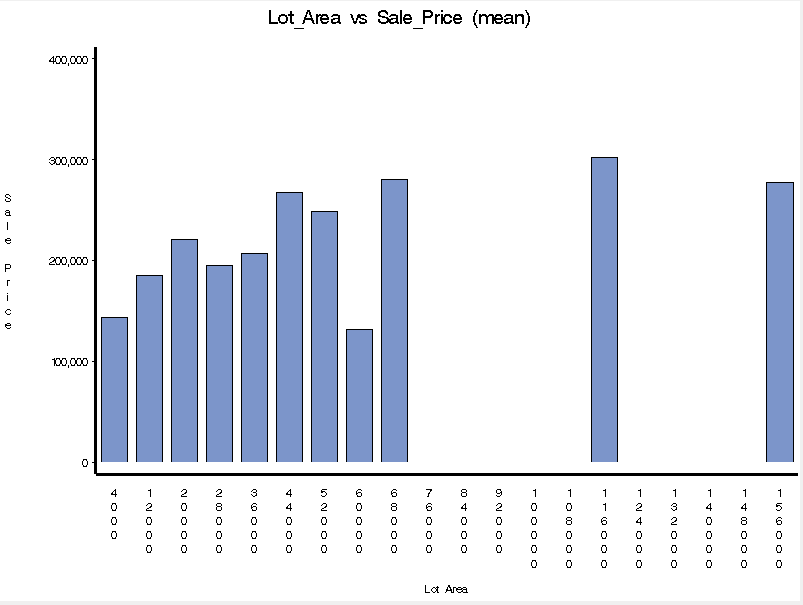
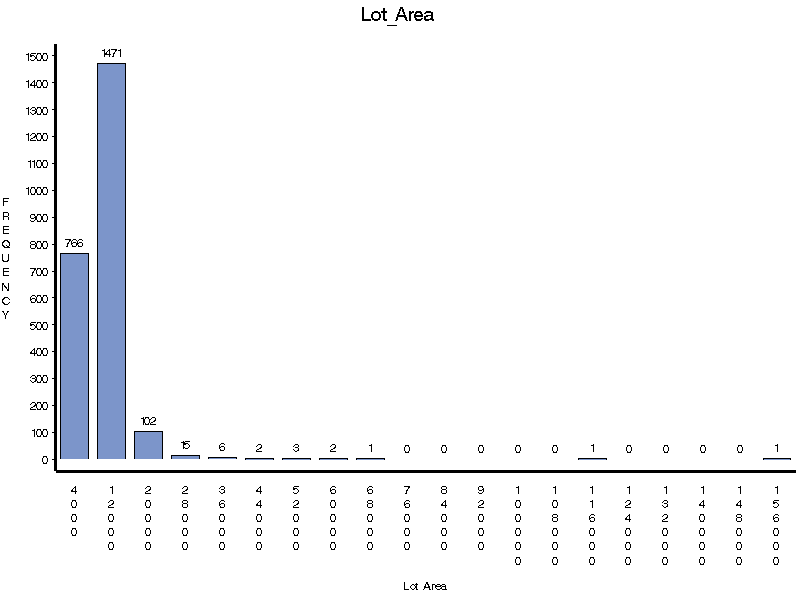
Half bath in house (Yes/No)



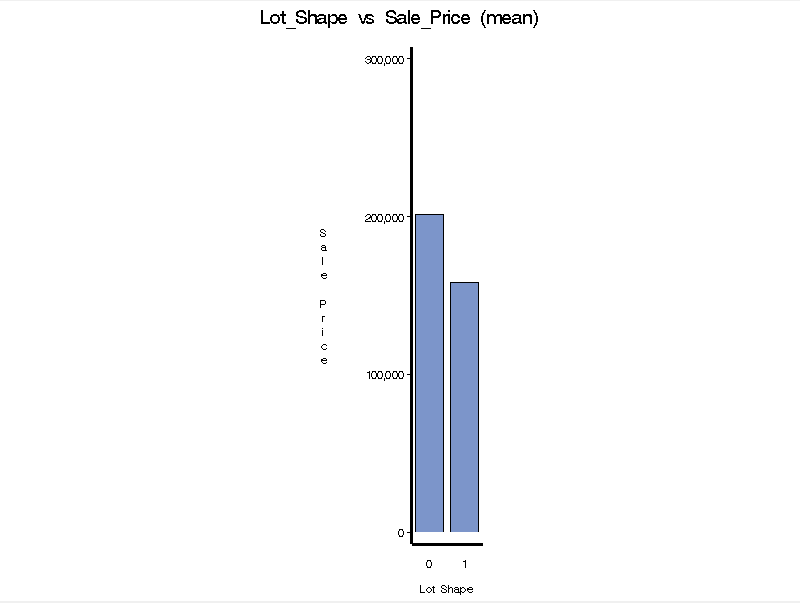
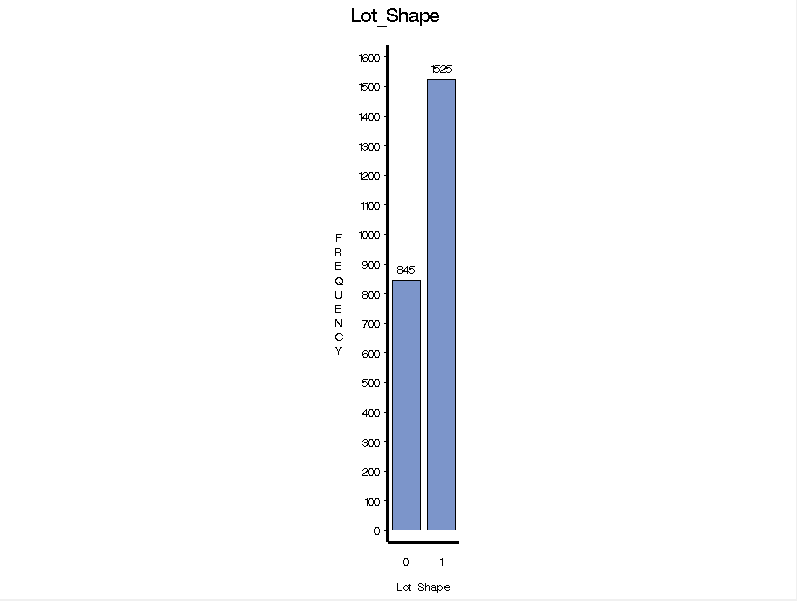
Heating QC

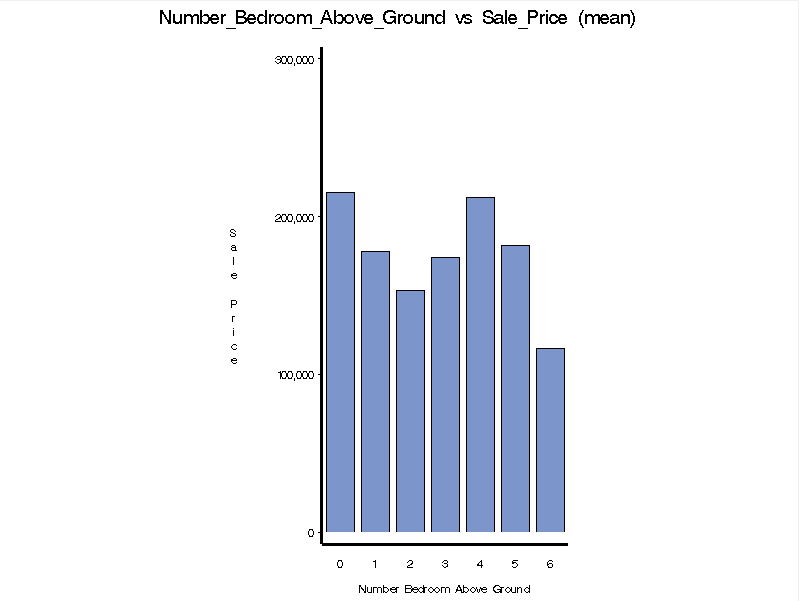
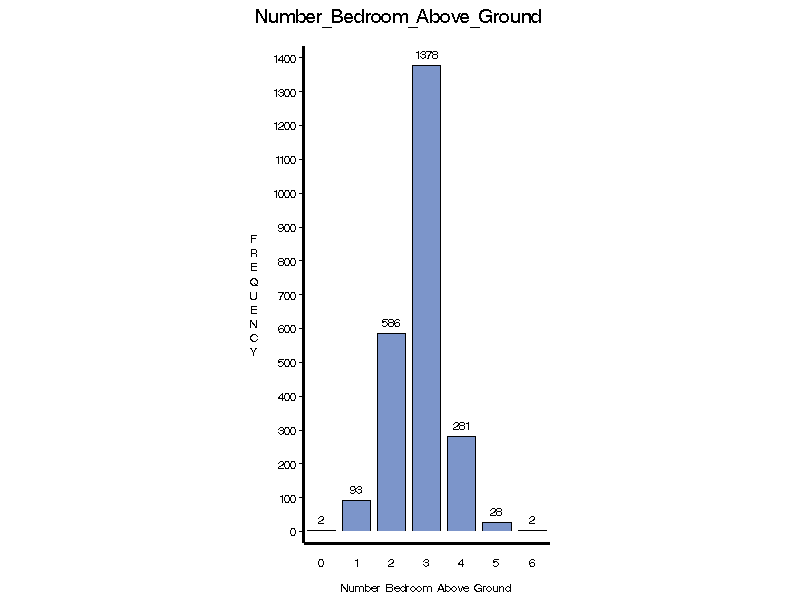
Lot area

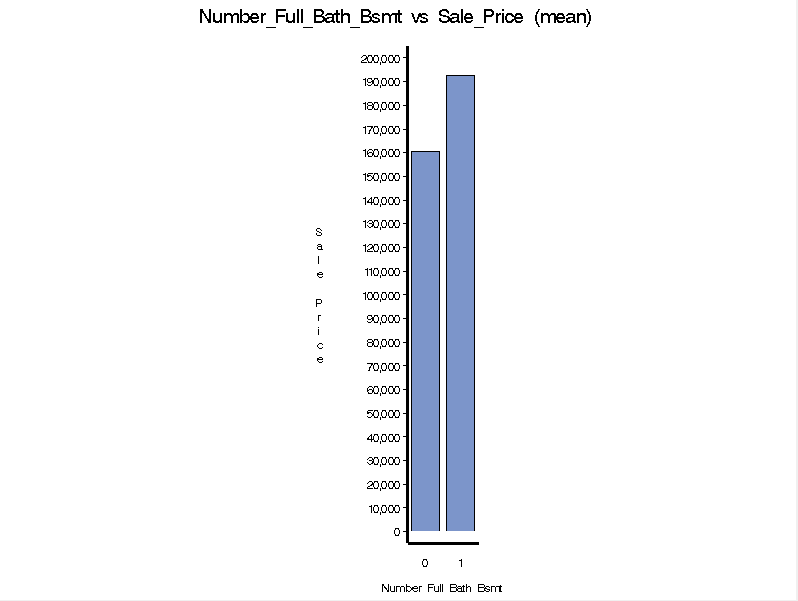
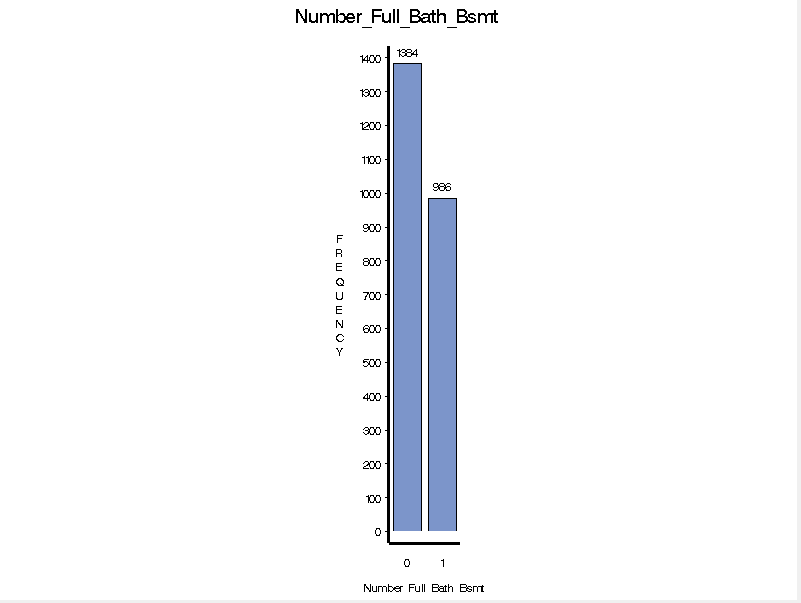
Lot shape

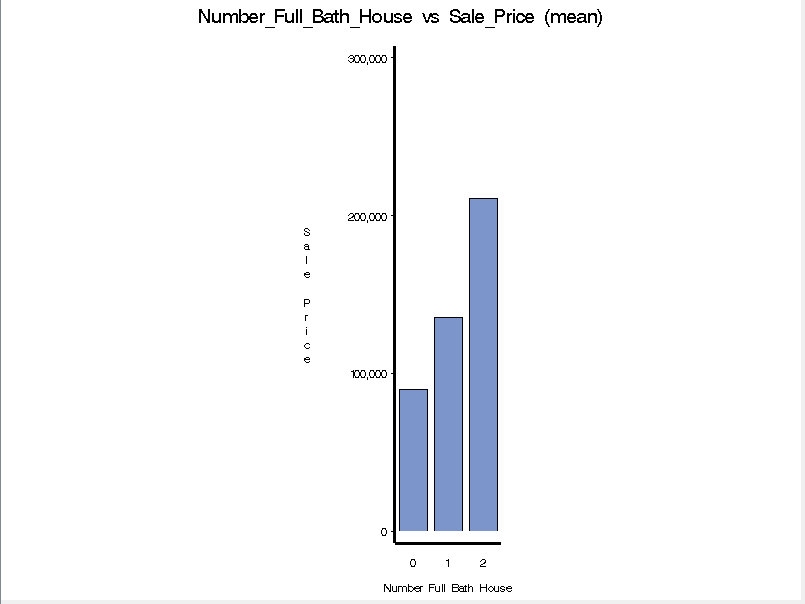
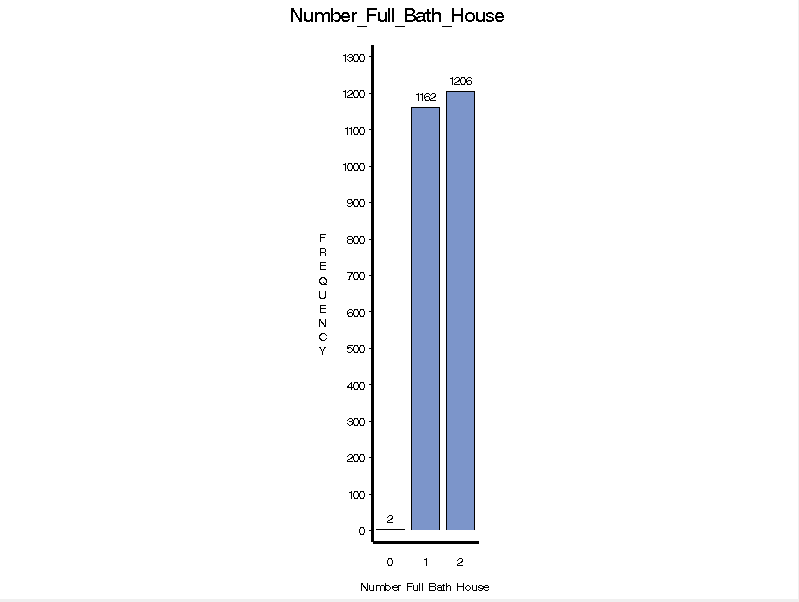
Number of bedrooms above ground

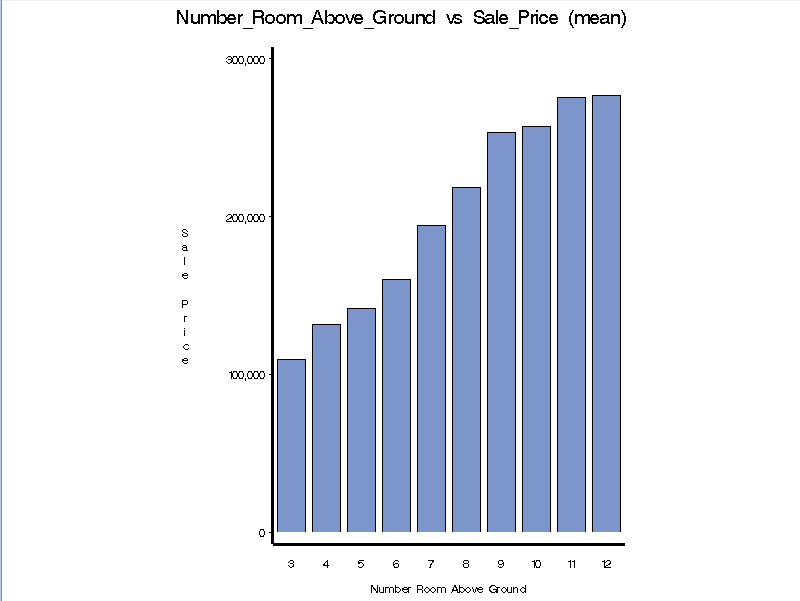
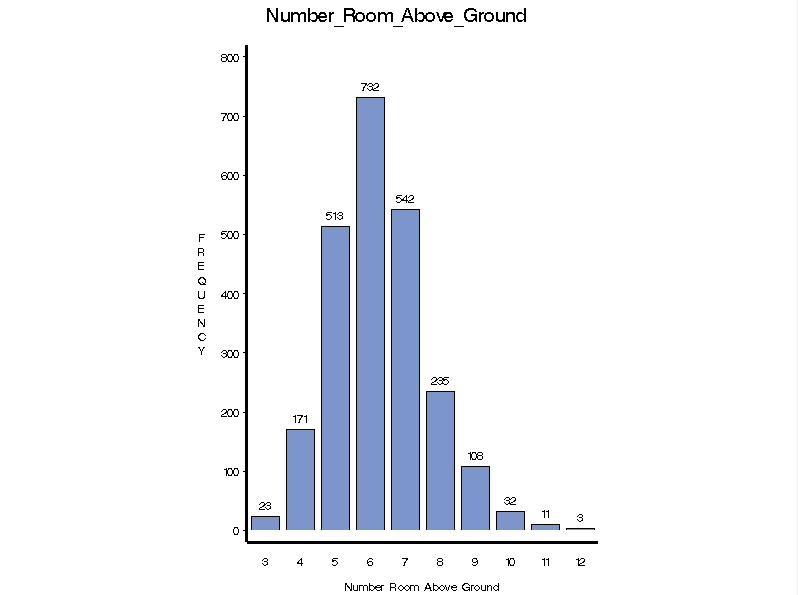
Number of full bathrooms in basement

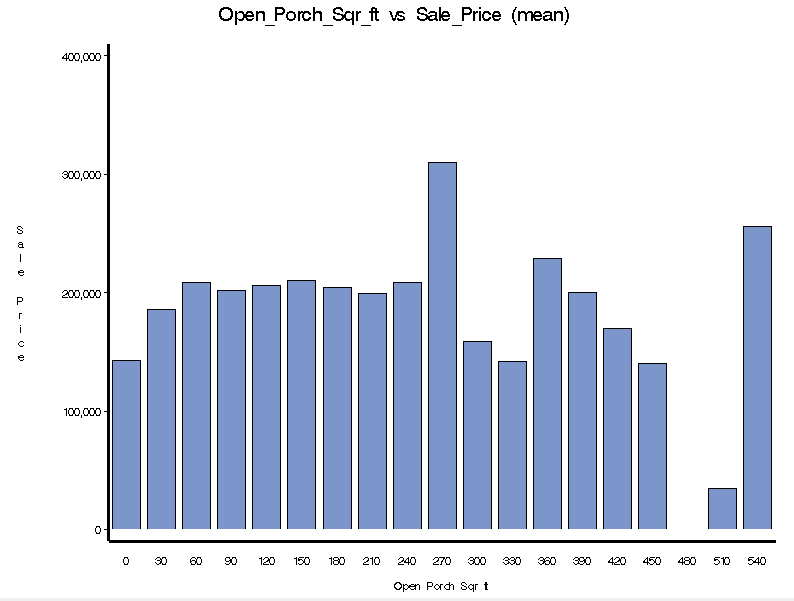
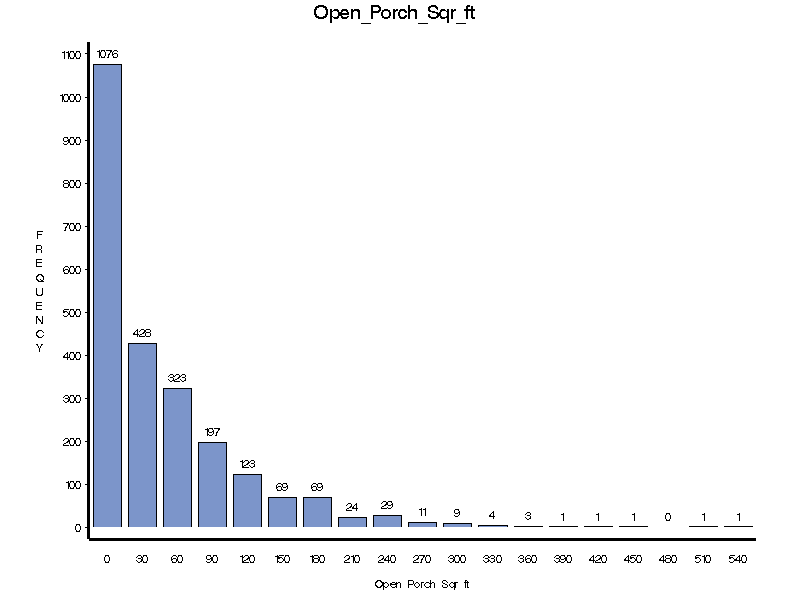
Number of full bathrooms

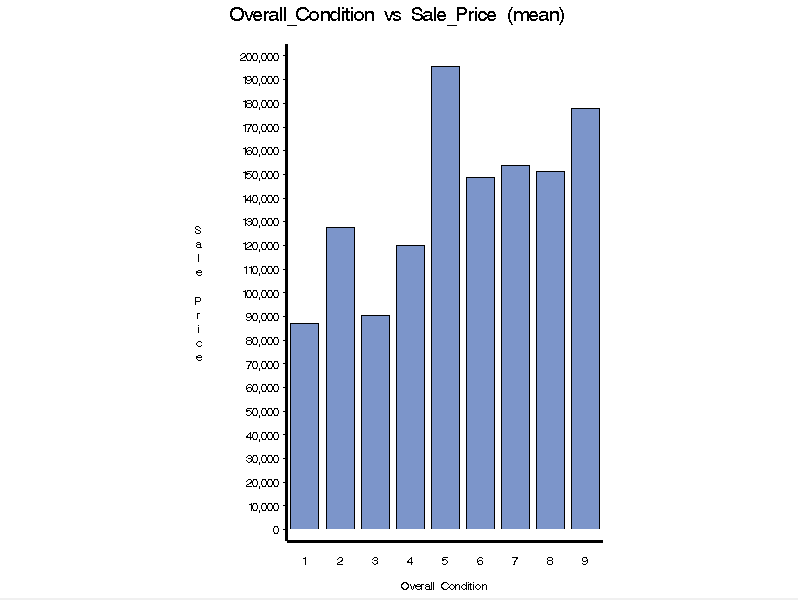
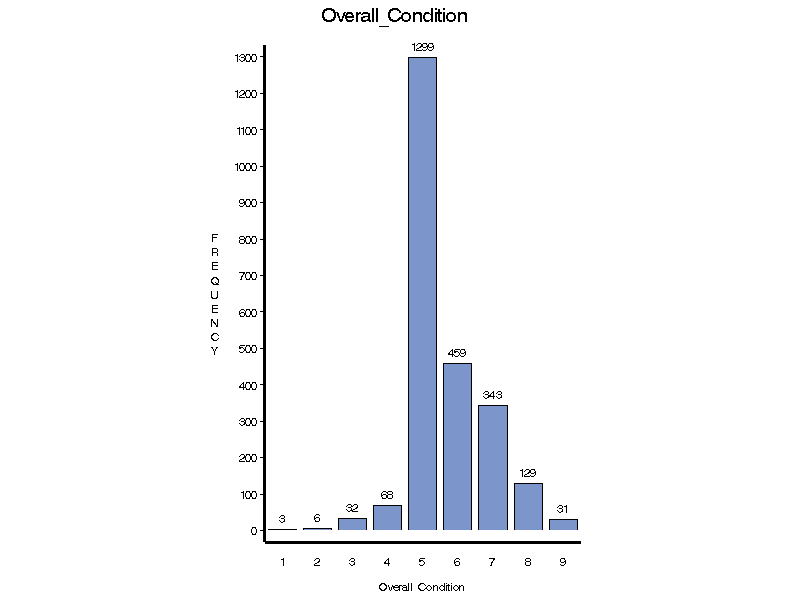
Number of rooms above ground

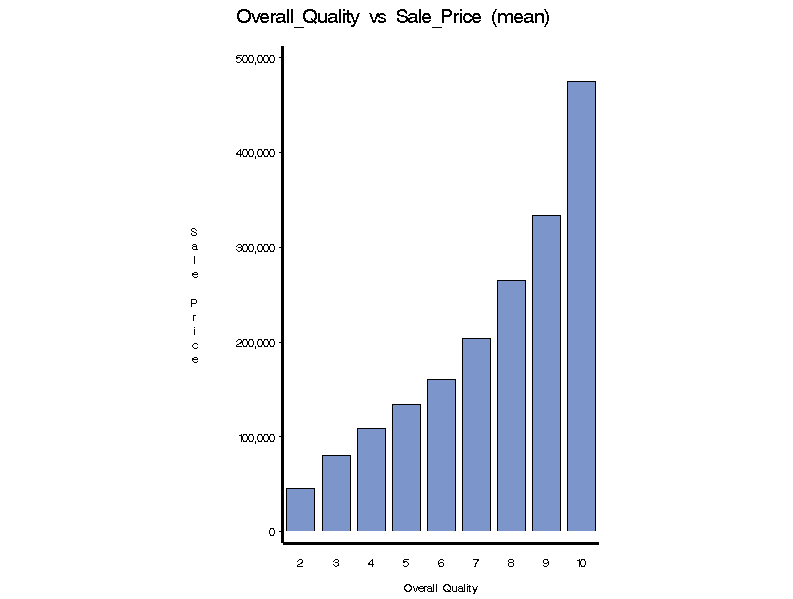
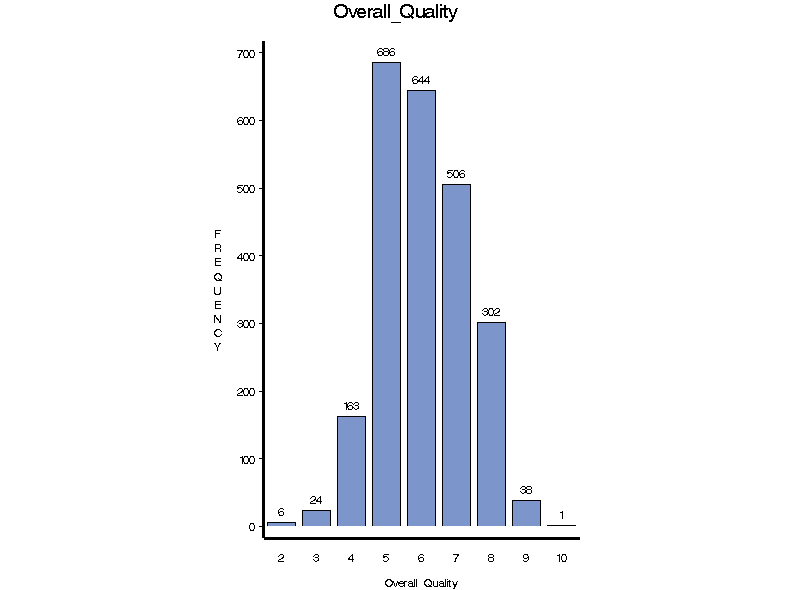
Open porch area (sq. ft.)

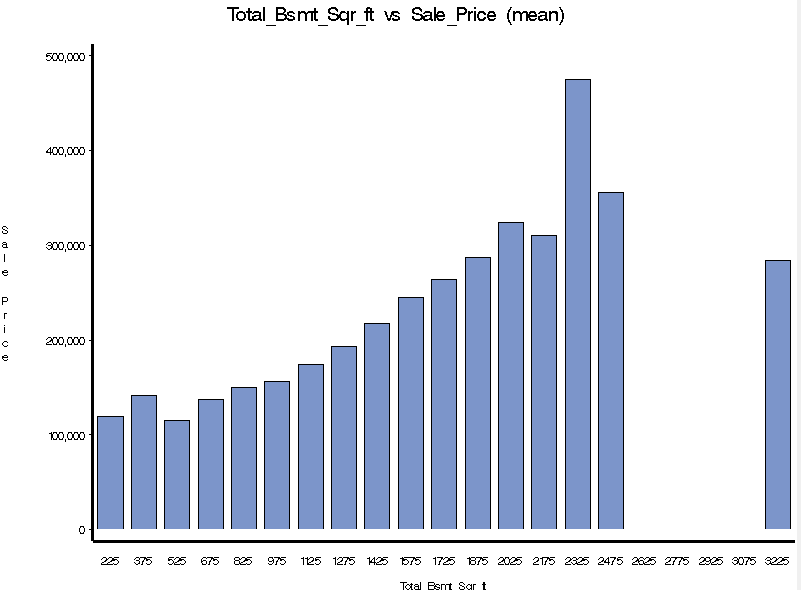
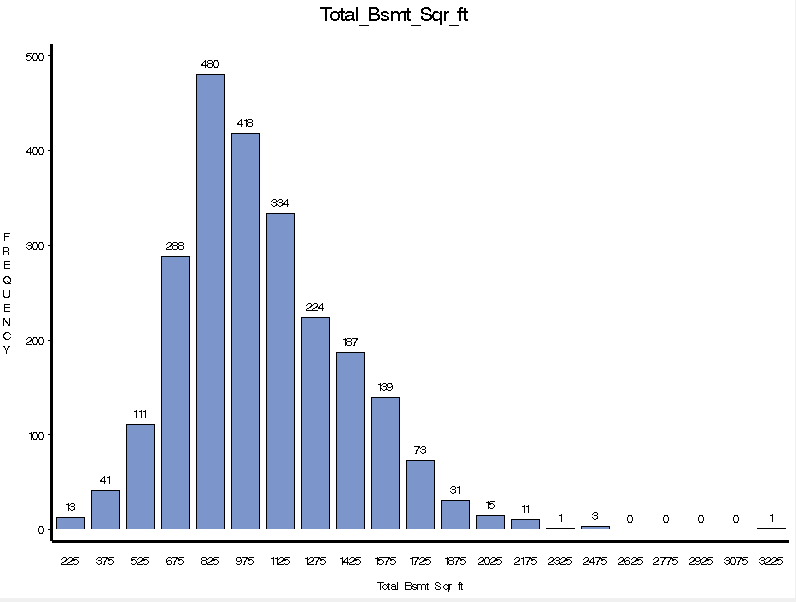
Overall condition

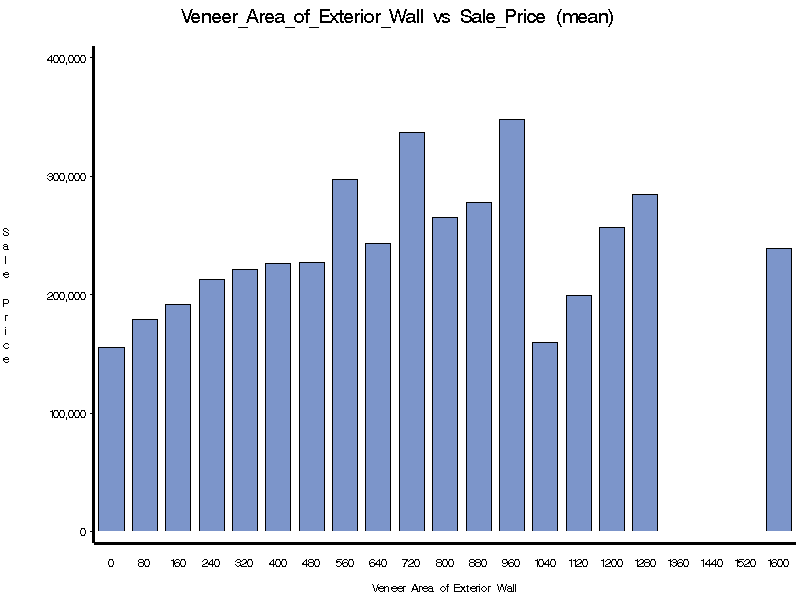
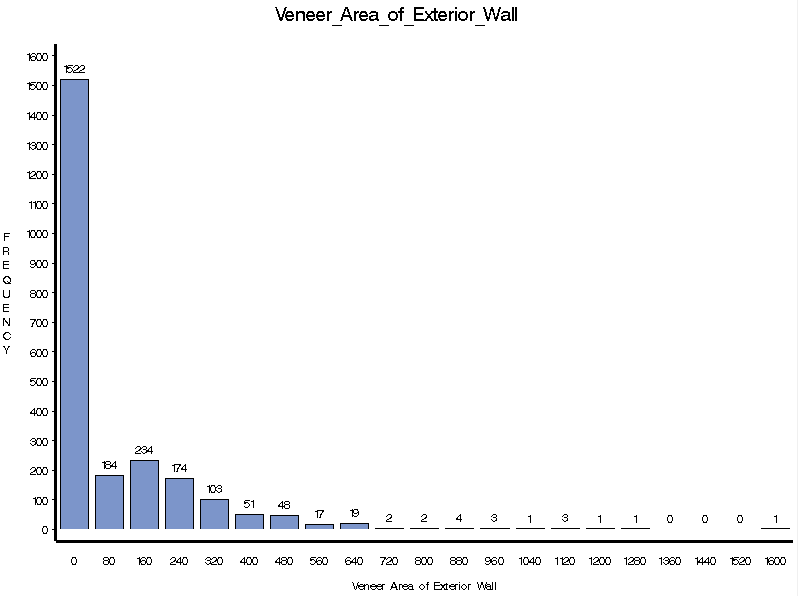
Overall quality

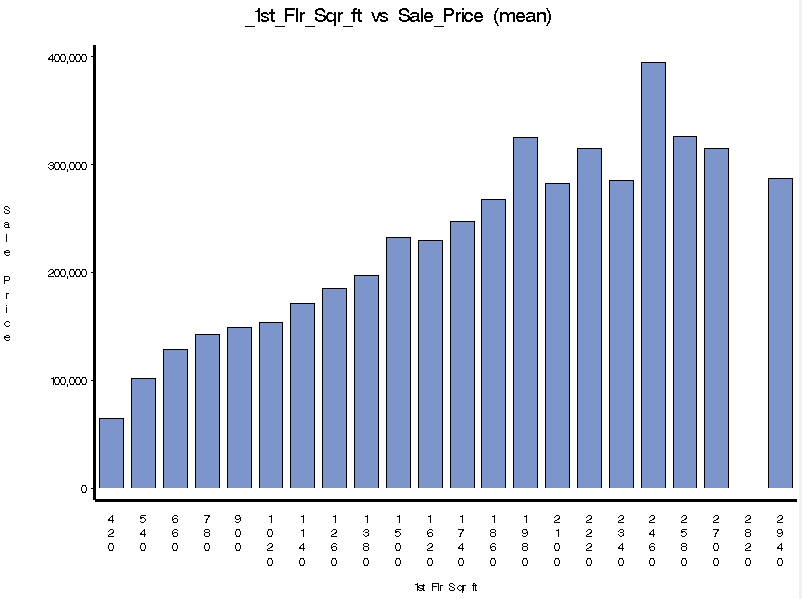
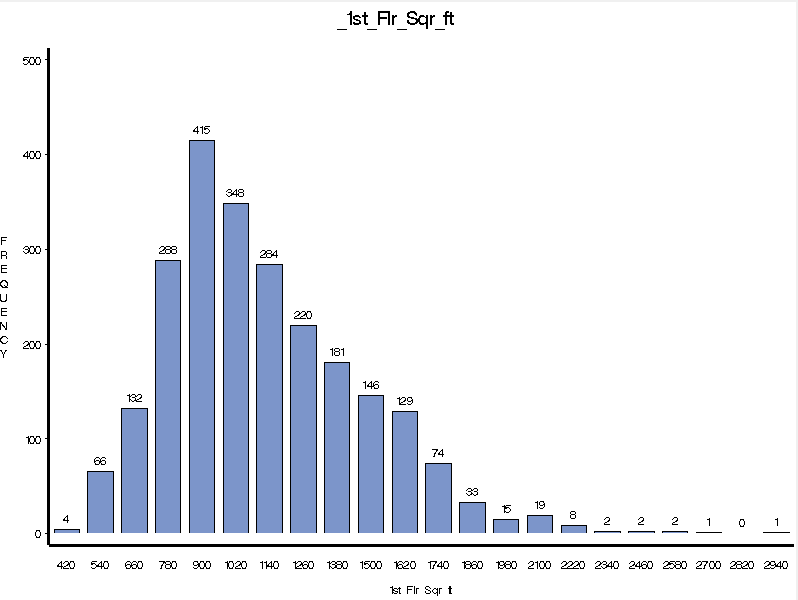
Total basement area (sq. ft.)

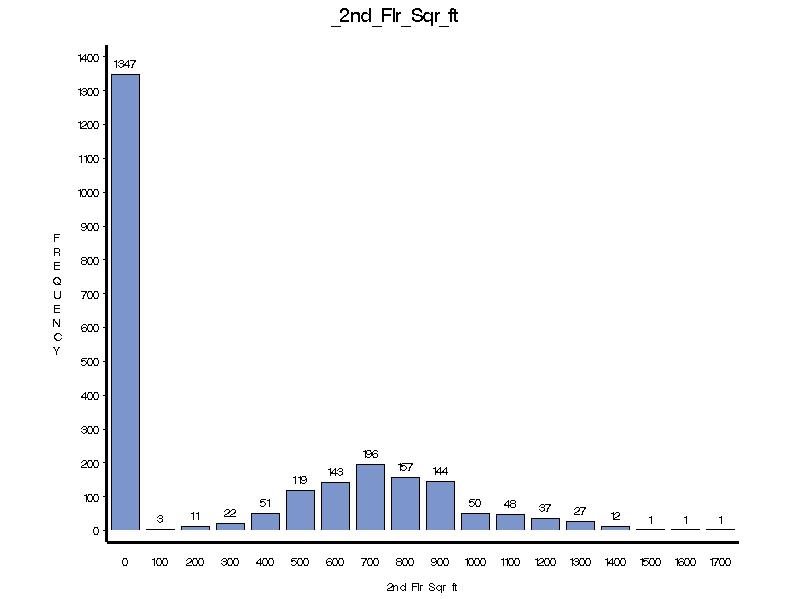
Veneer Area of Exterior Wall

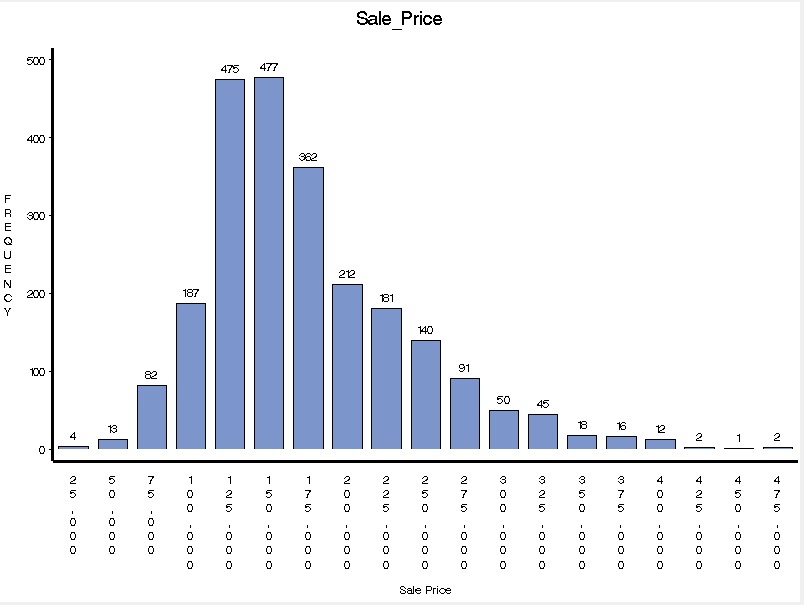
1st Floor Square Feet

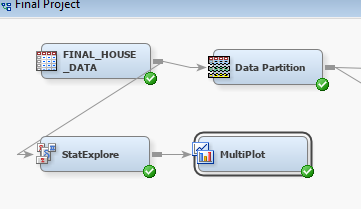
2nd Floor Square Feet

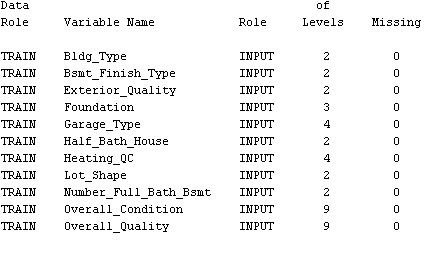
 

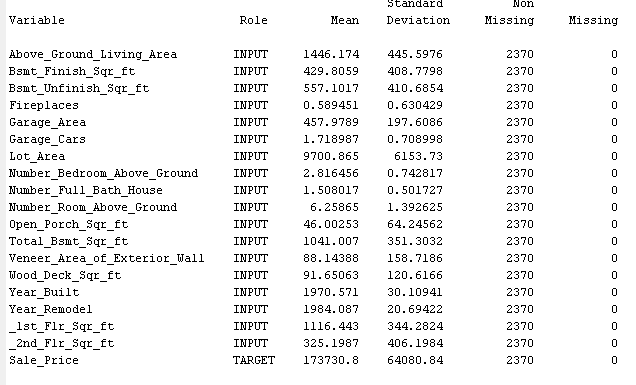
And Below is a graph representing the distribution of the target variable in the training and validation data sets.



Another part of the EDA process is using Stat Explore to see if there are any missing fields that need to be addressed or if any variables have stats that don’t make sense. We found that none of the variables contained any missing fields, and it appeared that all of the variables had stats that seemed to make sense.

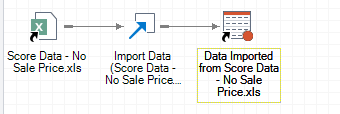






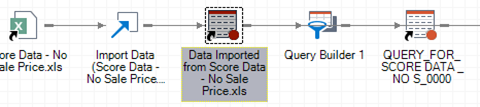
**2.6 Score Data Preparation and EDA**

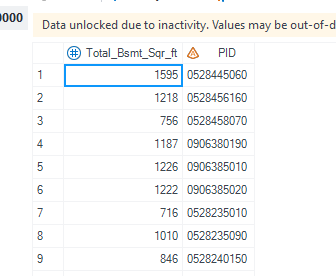
The last step in the data understanding phase is the creation and exploration of the Score data set. So, first we needed to import the score data into SAS EG in order to convert it into the correct format for SAS EM.



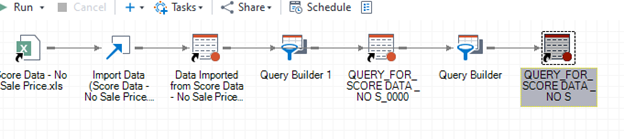
The next step is to look at the table and see what variables we have in this set, in order to compare with what is available in the training and validation sets. We discovered that we are missing the Exterior Quality, Foundation, and Total Basement Square Feet variables in this set. Since the Total Basement Square Feet variable was highly correlated with the target variable in the other data sets, we decided to create this variable to use. Luckily, we discovered that this variable is the sum of the Finished and Unfinished Basement Square Feet variables. So, we used SAS EG to create this column in the score data set.

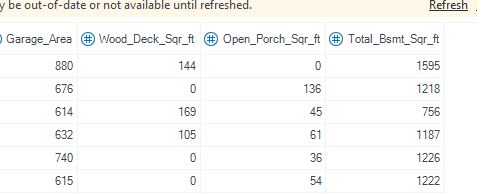
To create this new column, we first created a new table using the query builder containing the computed column Total Bsmt Sqr ft and the Property ID.



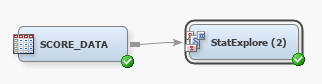


Then we merged this table with the original score data table to create the final data table with our new computed column.

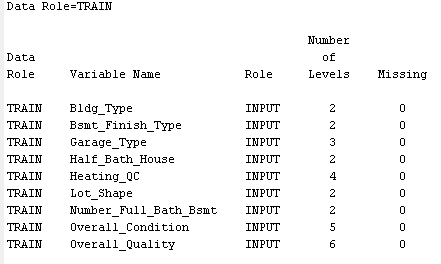


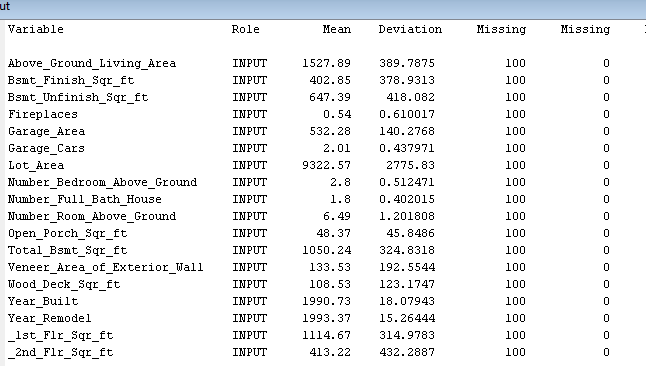


We then exported this data file as a SAS file to be used in SAS EM. After this was done, we created this new data source in our final project file and diagram in SAS EM. After doing this, we assigned the role of Score to this data set and used StatExplore to get a quick rundown on the variables in this data set.

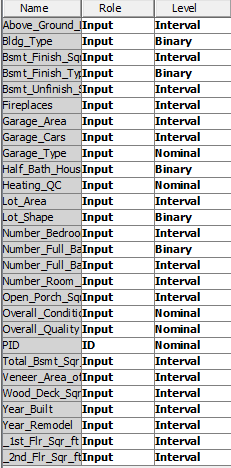


Using StatExplore, we discovered that there were no missing fields in any of the variables in this set.





We also made sure that all of the variables in this set were assigned the correct levels and roles to correspond with the training and validation sets.

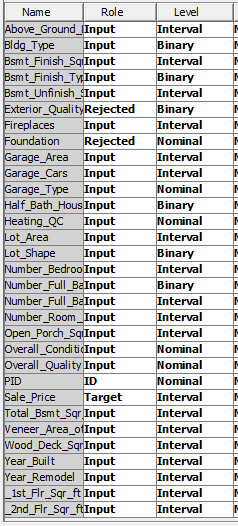


We then left this node alone until it is needed later in the process.

**3. Data Preparation Phase (Phase 3)**

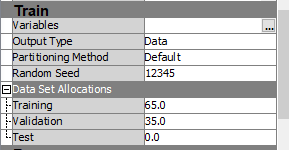
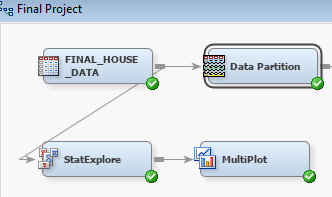
**3.1 Assignment of Roles**

Our next step in the process is to prepare the data so that we can create and test our models. The first step in this phase is to make sure we have assigned the proper roles to our variables. After performing our EDA, we decided that we would reject the variables Exterior Quality and Foundation, as they do not appear in the Score data set and do not appear to be very correlated with the target variable Sale Price.



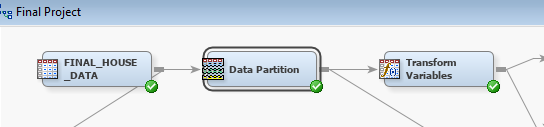
**3.2 Data Partitioning**

The next step in the data preparation phase is to partition our data set into a training and validation set for model creation and testing purposes. We decided that the best partition for our data would be 65% in the training set and 35% in the validation set. This ensures that we have enough data to be able to create a robust model, while still retaining enough data to sufficiently test our models with the validation data.

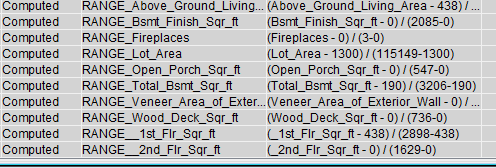


**3.3 Data Transformation**

The final step in the data preparation phase is to transform any of the variables that may need to be transformed for certain models to work better. While there may be additional transformations needed for specific models, these transformations could be used for all the models.



For our transformations, we decided to use range transformations, and have shown the results below.



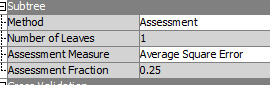
Any other transformations needed for specific models will be covered in the sections for those models.

**4. Modeling Phase (Phase 4)**

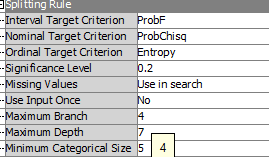
In this phase, we are now creating our models that we will evaluate against each other to determine the model that will make the best predictions of Sale Price in the Score data set. Below, we will discuss the four different kinds of models that will eventually be evaluated against each other. We will start with the Decision Tree model.

**4.1 Decision Trees**

The first type of model that we explored was decision trees. This type of model will create decision rules that, once applied, will split up our data into discreet groups that are the most similar. It applies these rules to the different input variables in order to accomplish this partitioning so that we can apply the mean sales price of that group to the score data in order to predict sales prices. We tested three different configurations for this model, all using average square error as the assessment measure. All three models also have a maximum number of splits set to 4.

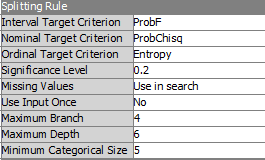


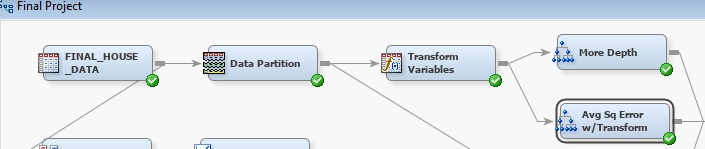
The first model will be using the transformed variables from earlier and a maximum depth of 7. This model will be named More Depth.



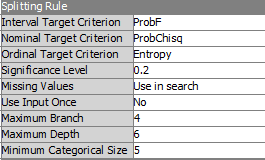


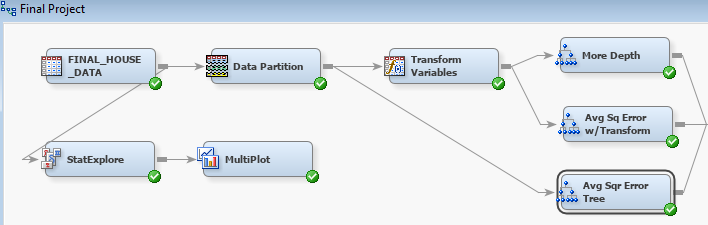
The next model will be created using transformed variables again but will have a maximum depth of 6 instead of 7. This model is titled Avg Sq Error w/Transform.



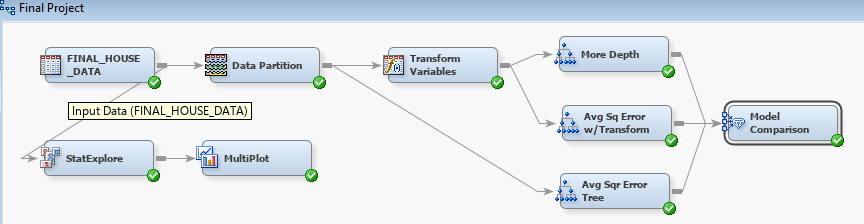


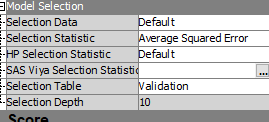
The last model that we created under the decision tree style will not be using the transformed version of the variables that the other two use. It will have a maximum depth of 6 and will be titled Avg Sqr Error Tree.



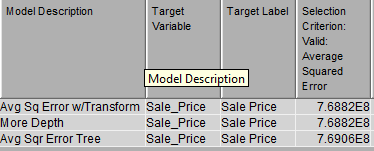


The next step was to evaluate and compare these three models using the model compare utility in SAS. For the comparison, we decided to evaluate the models based on the average square error using the validation set.



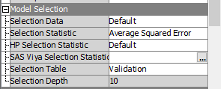


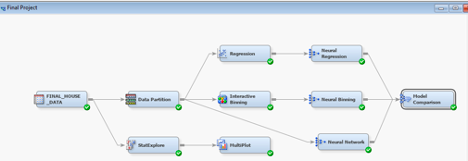
After running the comparison, we discovered that the Avg Sq Error w/Transform was the best model of this type for our purposes, based on the selection statistic of average square error. And In fact, it would have been the best fitting model based on most of the available selection criteria.



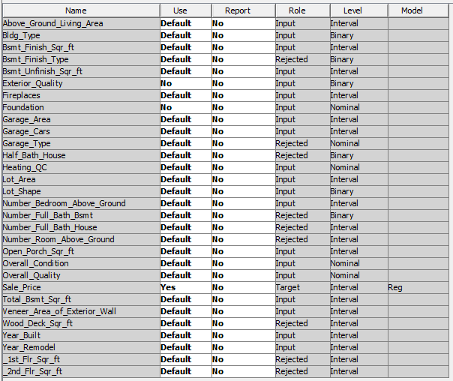
**4.2 Neural Network**

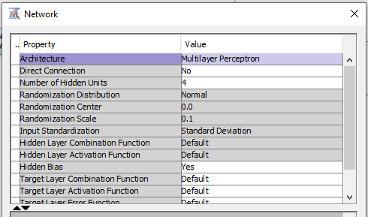
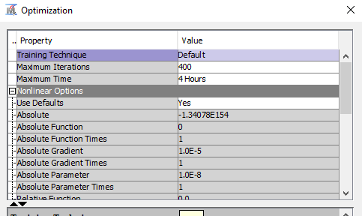
The second model we created is a Neural Network. This type of model will create a set of interconnected nodes in three layers (input, hidden, output) and use multiple connected regressions between the nodes to predict sales prices. We tested three different models with different variations of inputs and hidden units and used the average squared error to assess the best model.



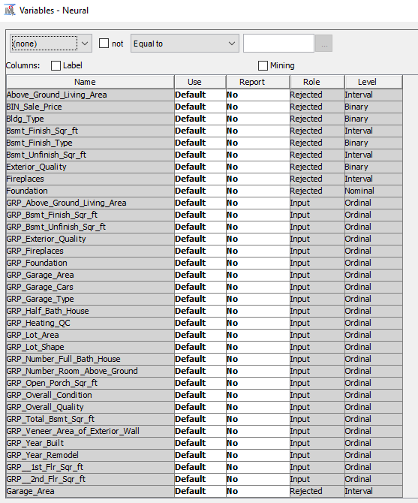


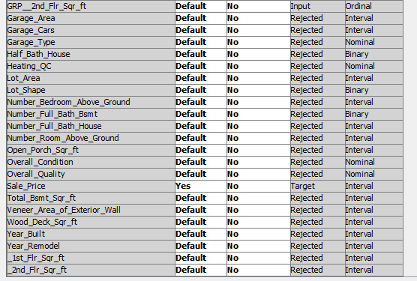
The first model used regression to determine the best inputs for the model. In addition to the variables rejected by the Regression node, Exterior Quality and Foundation were manually excluded because these variables are not included in the Score data. This model used maximum iterations of 400 and 4 hidden units. This model is titled Neural Regression.

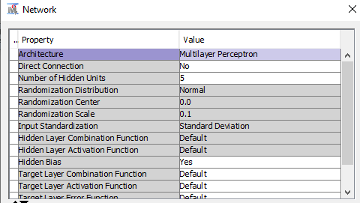
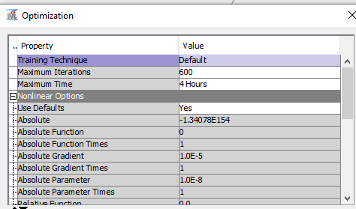




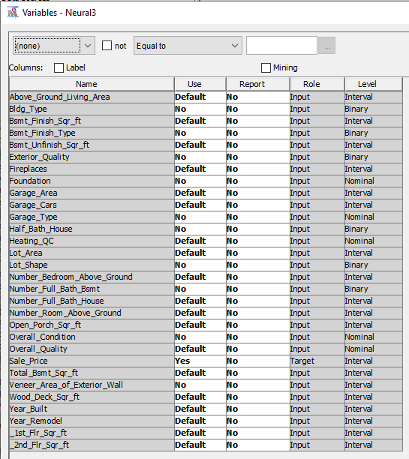
The next model used interactive binning to determine the best inputs for the model. This model used 600 maximum iterations and 5 hidden units and is titled Neural Binning.

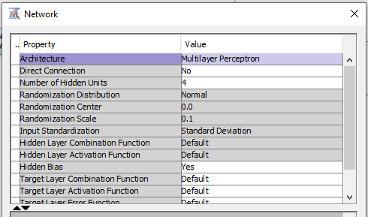
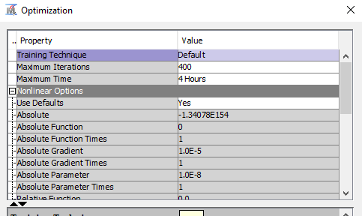




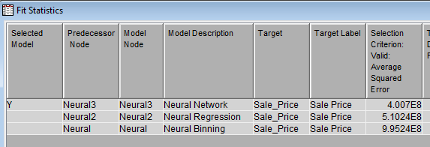


The third model used manual selection of input variables based on the EDA performed. The variables Bldg\_Type, Bsmt\_Finish\_Type, Garage\_Type, Half\_Bath\_House, Lot\_Shape, Number\_Full\_Bath\_Bsmt, and Overall Condition were excluded because they were not highly correlated with Sales Price. Veneer\_Area\_of\_Exterior\_Wall was somewhat correlated with Sales Price, but it’s distribution was extremely skewed, with over half of the houses having no Veneer Area at all, so we felt this was not a useful indicator. Exterior Quality and Foundation were excluded because these variables are not included in the Score data and were not highly correlated with Sales Price in the EDA. This model used 400 maximum iterations and 4 hidden units and is titled Neural Network.



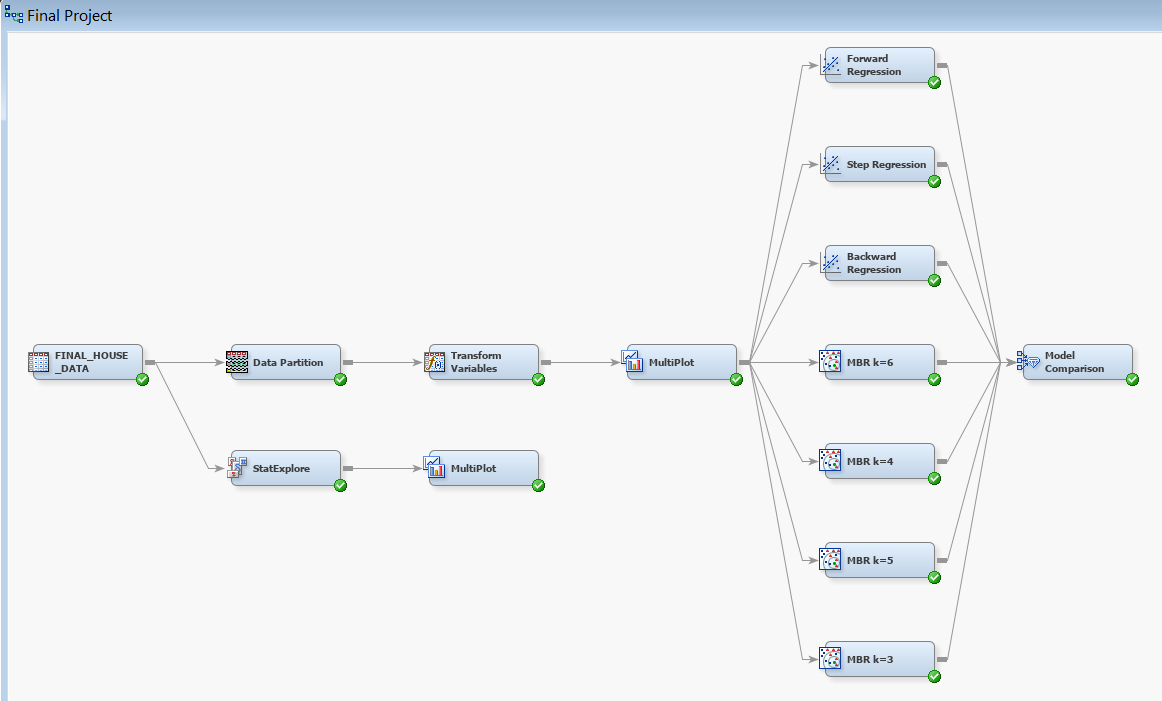


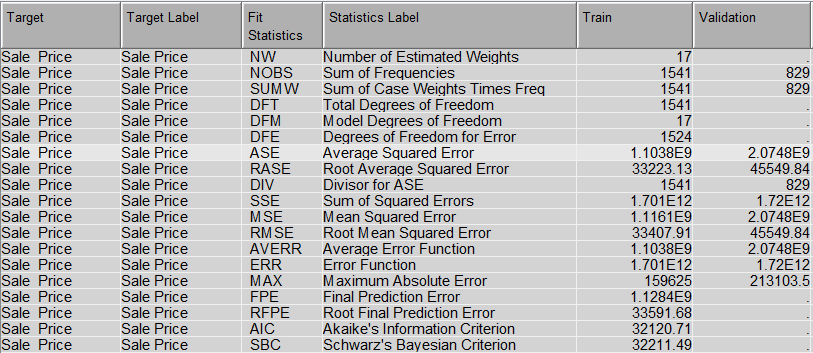
After comparing the three models, we determined the best model was the Neural Network using manually selected inputs and 4 hidden units.

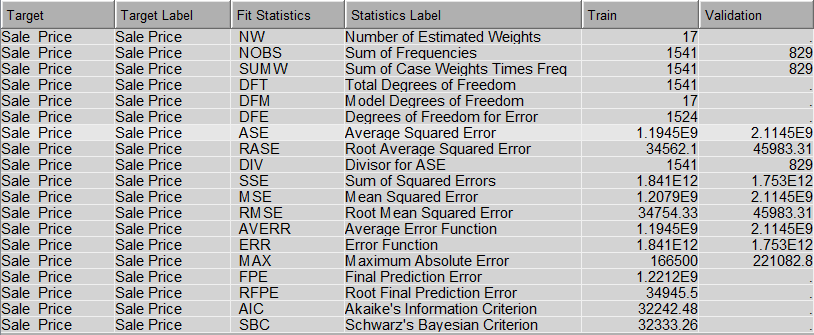


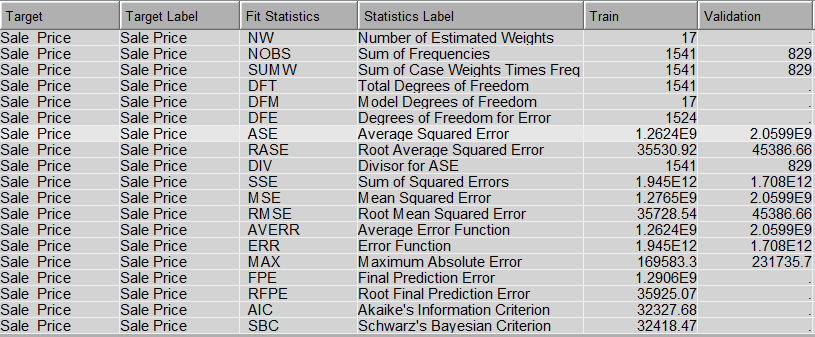
**4.3 *k-*Nearest Neighbor**

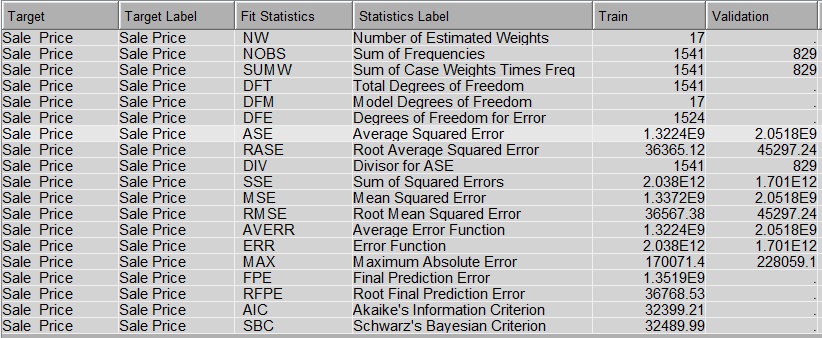
The k-nearest neighbor algorithm is an easy to use supervised machine learning algorithm. Here to find the optimum k-nearest neighbor value, we compared the average squared error value for k = 3,4,5 and 6. We considered the ASE value as our performance matric since Sale price (the target variable) is interval.



K=3  


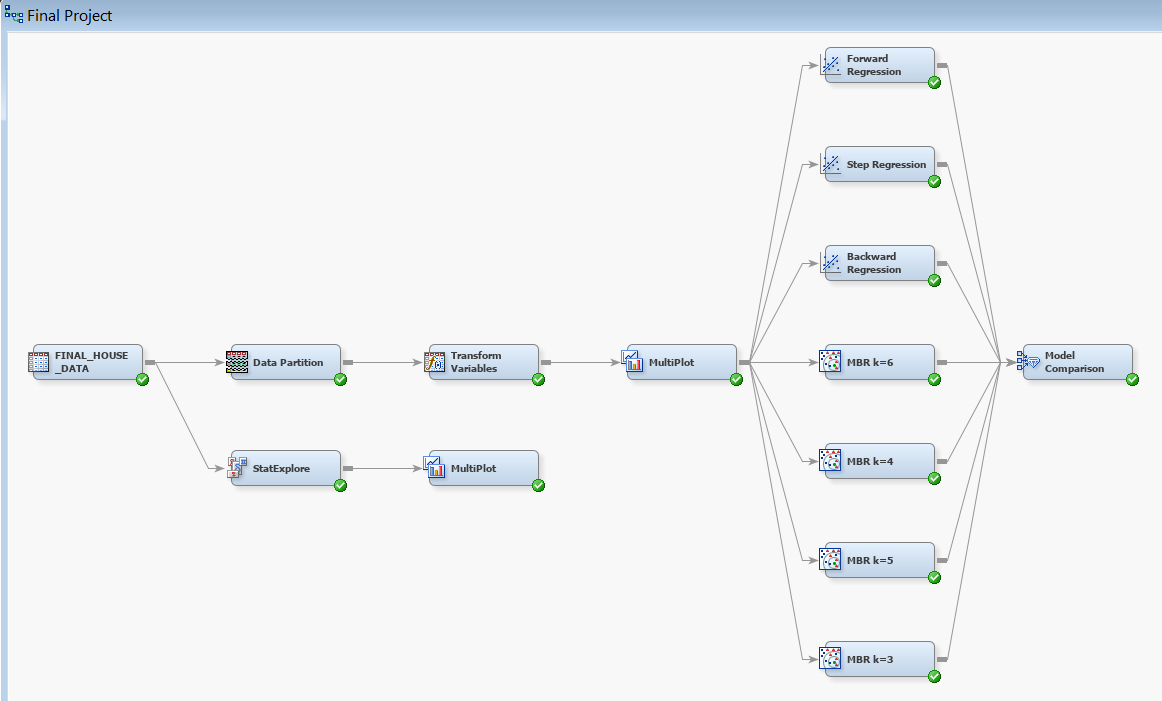
K=4  


K=5  


K=6  


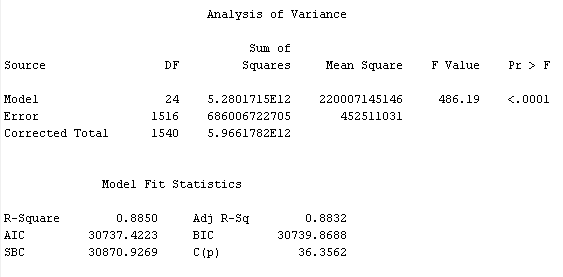
On comparison, we can see that k=3 has the lowest ASE value of 1.1038E9

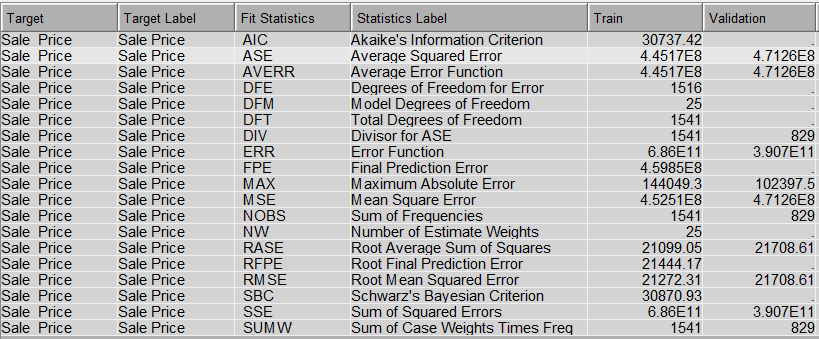
**4.4 Regression**



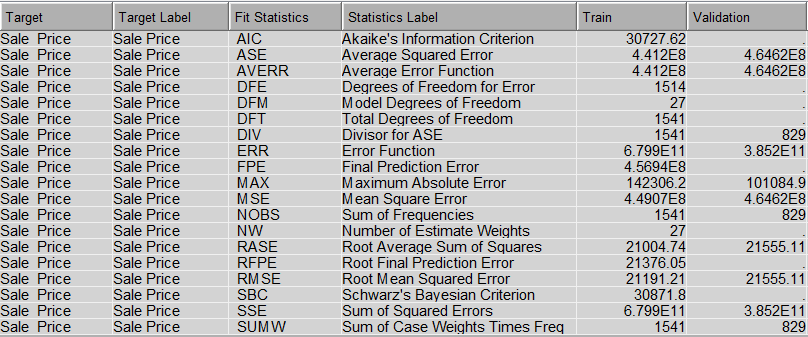
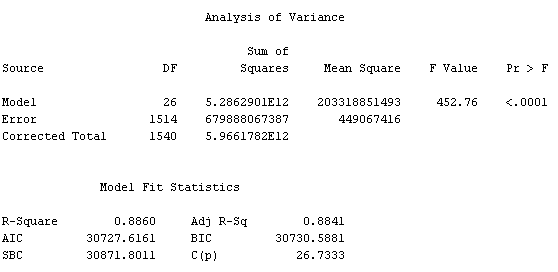
Linear regression is an easy to use and less complex model that assumes a linear relationship between the input variables. Here we have performed three types of regression analysis: forward, backward and stepwise regression to compare which regression model is the best.

Forward regression:

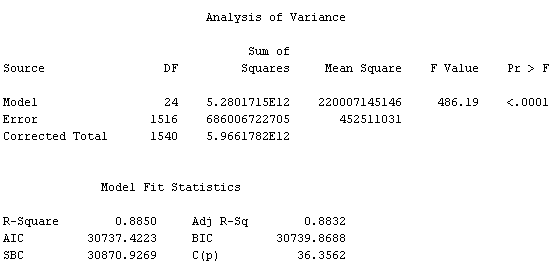
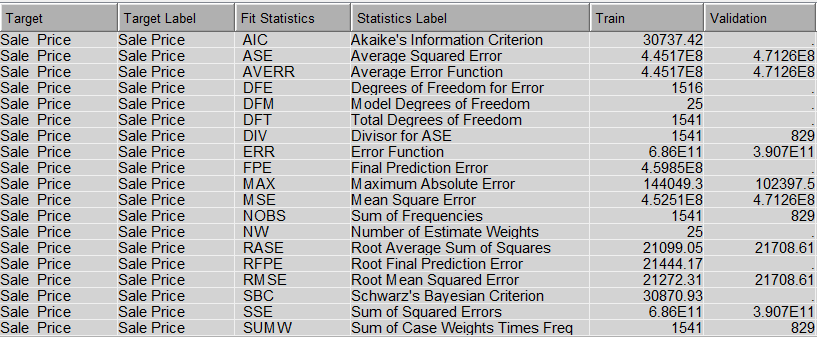




Backward regression:



Stepwise regression:

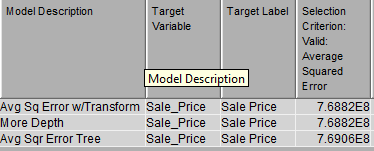
  


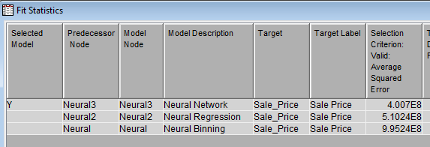
On comparison, we can see that the backward regression has the least ASE value.

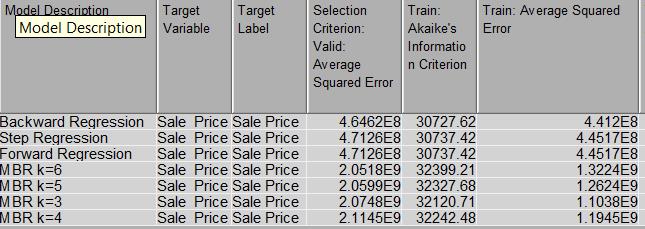
**5. Evaluation Phase (Phase 5)**

**5.1 Model Results and Best Model Selection**

We compared each of the best models and determined that the Neural Network model with manual selection of input variables was the best model because it had the lowest Average Squared Error of 4.007E8. This model appears to have the best predictive qualities of all the models created.







**5.2 Variables Included in Final Sales Price Model**

The variables included in the Final Sales Price Model are:

|  |  |
| --- | --- |
| Above Ground Living Area | Number of Rooms Above Ground |
| Basement Finish Sq. Ft. | Open Porch Sq. Ft. |
| Basement Unfinished Sq. Ft | Overall Quality |
| Fireplaces | Total Basement Sq. Ft. |
| Garage Area | Wood Deck Sq. Ft. |
| Garage Cars | Year Built |
| Heating QC | Year Remodel |
| Lot Area | 1st Floor Sq. Ft. |
| Number of Bedrooms Above Ground | 2nd Floor Sq. Ft. |
| Number of Full Bath House |  |

**VI. Phase 6: Deployment Phase**

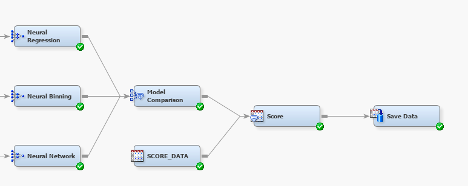
**6.1 Results: Top 20 Luxury Homes**

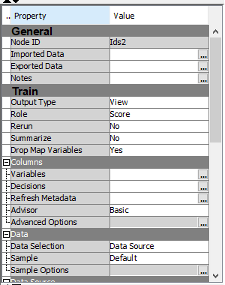
The top 20 luxury homes we recommend based on our model are:

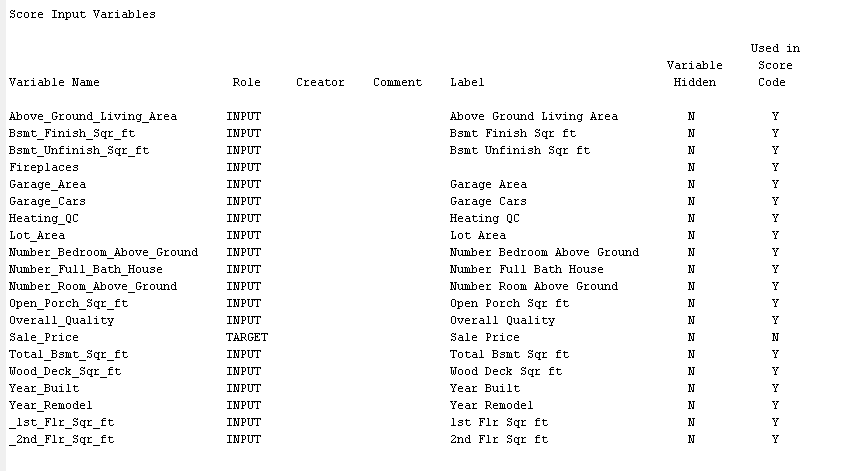
|  |  |
| --- | --- |
| **PID** | Prediction Sale Price |
| 528166120 | $ 385,711.39 |
| 528118040 | $ 359,158.29 |
| 528178070 | $ 344,255.17 |
| 528114050 | $ 335,631.51 |
| 528172080 | $ 275,254.62 |
| 528170070 | $ 272,048.95 |
| 907254020 | $ 264,715.14 |
| 528445060 | $ 247,535.75 |
| 907253110 | $ 237,056.48 |
| 907254090 | $ 229,682.23 |
| 907285020 | $ 228,756.08 |
| 907192130 | $ 223,175.66 |
| 528458040 | $ 222,267.21 |
| 528280180 | $ 218,812.20 |
| 534128010 | $ 218,328.79 |
| 528253020 | $ 215,922.48 |
| 907192090 | $ 214,612.24 |
| 534128020 | $ 214,241.76 |
| 528490040 | $ 211,507.53 |
| 907405020 | $ 211,458.31 |

**6.2 Explanation of Results**

Using the Score Data Set that was imported and reviewed in Section 2.6, we proceeded to use the Neural Network model to score and predict the sales price of the properties. Using the model created, we were able to reasonably predict a sale price for the properties contained in the score data set. The above chart shows the top 20 homes in the score set based on the predicted sale price.







We tried to use all the variables that we could in order to create the best model without overfitting. And since the target variable here is an interval variable, there is not the same kind of worries about false positives and negatives that may have been present if the target had been nominal or binary. Also, there is no weight to the decisions that needed to be made, such as a cost for sending out mail versus the expected benefit from receiving a donation. This means that we could focus our attention on mainly the average square error to make our decisions. This led us to the neural network model as the best fit, and thus our results from applying that to the score set.

**VI. Final Model Diagram**

Diagram

Description automatically generated