

0

Department of Management Science and Technology (AUEB)

AMES IOWA HOUSING DATASET

Statistics for business analytics I

Eva Giannatou January 2017

Contents

1. [Abstract 4](#_bookmark0)
2. [Introduction 4](#_bookmark1)
3. [Data preparation 5](#_bookmark2)
   1. [Data cleansing 5](#_bookmark3)
   2. [Data transformation 6](#_bookmark5)
      1. [Numeric attributes 6](#_bookmark6)
      2. [Categorical attributes 7](#_bookmark7)
      3. [High correlated attributes 7](#_bookmark9)
4. [Exploratory analysis 8](#_bookmark11)
   1. [Numeric Variables 10](#_bookmark13)
5. [Attribute selection 11](#_bookmark15)
   1. [Lasso attribute selection 11](#_bookmark16)
   2. [Selected attributes 14](#_bookmark21)
6. [Regression models 15](#_bookmark24)

[6.1 M0 15](#_bookmark25)

* + 1. [Interpretation 16](#_bookmark27)
    2. [Linear model assumptions 18](#_bookmark31)
  1. [Prediction 23](#_bookmark40)
     1. [k-fold cross validation 23](#_bookmark41)
     2. [Test data prediction 23](#_bookmark43)
  2. [Model improvements 24](#_bookmark45)
     1. [Outliers 24](#_bookmark46)
     2. [Non-normal transformations 26](#_bookmark50)
     3. [Experimentation 27](#_bookmark54)

[6.4 M1 30](#_bookmark56)

[6.4.1 Interpretation 30](#_bookmark57)

* 1. [Prediction 34](#_bookmark64)
     1. [k-fold cross validation 34](#_bookmark65)
     2. [Test data prediction 34](#_bookmark67)

1. [Conclusions and future improvements 36](#_bookmark71)
2. [Παράρτημα κώδικα 36](#_bookmark72)
   1. [Data cleaning & null replacement 36](#_bookmark73)
   2. [Data transformation & new columns 40](#_bookmark74)
   3. [Visualize variables 40](#_bookmark75)
   4. [Split numeric and nominal variables into two different data frames 41](#_bookmark76)
   5. [Categorical variables to dummy 42](#_bookmark77)
   6. [Reduce number of columns 42](#_bookmark78)
   7. [Lasso 42](#_bookmark79)
   8. [Train regression models 44](#_bookmark80)
   9. [Evaluate outliers 44](#_bookmark81)
   10. [Plot predicted vs actual & residuals 45](#_bookmark82)
   11. [More plots 46](#_bookmark83)
   12. [Evaluate regression model assumptions 47](#_bookmark84)
   13. [Improve model 49](#_bookmark85)
   14. [Test model 50](#_bookmark86)
3. [Bibliography 52](#_bookmark87)

Figures

[*Figure 1 Number of null values per column 5*](#_bookmark4)

*Figure 2 plot of existing and imputed values of “Lot.Frontage” 6*

[*Figure 3 Converting categorical attribute "MS.Zoning" to binary dummy attributes 7*](#_bookmark8)

[*Figure 4 Barplots used to identify which attributes do not provide additional information 7*](#_bookmark10)

[*Figure 5 Bar Plot1: Categorical variables 8*](#_bookmark12)

*Figure 6 Bar Plot2: Categorical variables 9*

*Figure 7 Density plots: Numeric variables 10*

*Figure 8 Density plot 1: Numeric variables 10*

*Figure 9 Density Plot: Numeric variables 10*

[*Figure 10 Average home characteristics 10*](#_bookmark14)

[*Figure 11 Lars Lasso coefficients ………………………………………………………………………………………..…….10*](#_bookmark17)[*Figure 12 S corresponding to the optimal Cp equals to 0.8 11*](#_bookmark17)

[*Figure 13 Lars Lasso coefficient selection 12*](#_bookmark18)

[*Figure 14 Glmnet lasso coefficients 12*](#_bookmark19)

[*Figure 15 Glmnet lasso lamda min 13*](#_bookmark20)

[*Figure 16 Lasso selected attributes 14*](#_bookmark22)

[*Figure 17 Lasso selected attributes collinearity 14*](#_bookmark23)

[*Figure 18 m0: Stepwise regression full model 15*](#_bookmark26)

[*Figure 19 Model M0 evaluation 16*](#_bookmark28)

*Figure 20 Negative intersect cause 16*

[*Figure 21 Model M0 evaluation 17*](#_bookmark29)

[*Figure 22 m0: Plot linear full model 18*](#_bookmark30)

[*Figure 23 m0: Testing linearity of correlation between individual variable and price 20*](#_bookmark32)

[*Figure 24 Mo QQ Plot 21*](#_bookmark33)

[*Figure 25 M0 plot of studentised residuals…………………………………………………………………………………….20*](#_bookmark34)[*Figure 26 m0: residuals density plot……………………………………………………………………………………………21*](#_bookmark34)[*Figure 27 m0: Lilliefors test, testing normality of residuals 22*](#_bookmark35)

[*Figure 28 m0: Shaphiro-Wilk test, testing normality of residuals 22*](#_bookmark36)

[*Figure 29 mo: ncvTest, testing homoscedasticity 22*](#_bookmark37)

[*Figure 30 Vif testing collinearity 22*](#_bookmark38)

[*Figure 31 m0: Durbin Watson Test, testing autocorrelation 23*](#_bookmark39)

[*Figure 32 Leave one out 10 fold cross validation 23*](#_bookmark42)

[*Figure 33 M0 prediction on test set 24*](#_bookmark44)

[*Figure 34 Identifying outliers with Cook’s distance 25*](#_bookmark47)

[*Figure 35 Identifying outliers with the largest absolute studentised residual 25*](#_bookmark48)

[*Figure 36 Evaluating outliers by identifying influential rows 26*](#_bookmark49)

[*Figure 37 Sale price histogram………………………………………………………………………………………………….25*](#_bookmark51)[*Figure 38 Log(SalePrice) histogram 26*](#_bookmark51)

[*Figure 39 m0: Testing linearity of correlation between individual variable and price 27*](#_bookmark52)

[*Figure 40 M0, Gr.Liv.Area: Fitting polynomial degrees 26*](#_bookmark53)

[*Figure 41 M0, Total.Bsmt.SF: Fitting polynomial degrees ………………………………………………………………...27*](#_bookmark53) *Figure 42 Visualizing variables with problematic distribution 28*

[*Figure 43 Added value plots 29*](#_bookmark55)

[*Figure 44 m1: Stepwise regression full model 30*](#_bookmark58)

[*Figure 45 Model M1 evaluation 31*](#_bookmark59)

[*Figure 46 Model M1 evaluation 31*](#_bookmark60)

[*Figure 47 m1: Plot linear model 32*](#_bookmark61)

*Figure 48 m1: residuals density plot 33*

*Figure 49 M1 plot of studentised residuals 33*

[*Figure 50 m1: Lilliefors test, testing normality of residuals 33*](#_bookmark62)

[*Figure 51 m1: Shaphiro-Wilk test, testing normality of residuals 33*](#_bookmark63)

[*Figure 52 Leave one out 10 fold cross validation 34*](#_bookmark66)

[*Figure 53 M1 prediction on test set 35*](#_bookmark68)

[*Figure 54 dataset 22 both as train and test set 10-fold 35*](#_bookmark69)

[*Figure 55 dataset 22 as train and ames\_iowa\_housing\_test as test set 36*](#_bookmark70)

# Abstract

This paper presents a data set of 1500 residential property sales in Ames, Iowa between 2006 and 2012. The data set contains 82 explanatory variables describing every aspect of the home. The dataset is heterogeneous containing both ordinal, nominal, continuous and discrete attributes. Lasso was used in order to reduce the number of redundant variables and to define which ones will be used as input in the final multiple regression model. Objective of this paper is to compare the predictive performance of multiple regression models and to identify the best model for predicting the prices of the properties.

# Introduction

Objective of this paper is identify the most important variables and to define the best regression model for predicting the housing prices in Ames, Iowa. The data set used for this paper’s purposes, describes 1500 residential property sales in Ames, Iowa between 2006 and 2012. It contains 82 explanatory variables describing every aspect of the home. The dataset is heterogeneous and contains 23 ordinal, 23 nominal, 22 continuous and 14 discrete attributes. Continuous variables determine the various area dimensions such as the size of the living area, the basement and the porch while discrete variables quantify the number of rooms, baths, kitchens, parking spots etc. Nominal variables typically describe the various types or classes of dwellings, materials and locations such as the name of the neighborhood, the garage type, the sale type etc. Ordinal variables typically rate the quality and condition of different house parts and utilities.

The fact that the dataset was over parameterized and heterogeneous lead to the following hardships and increased the difficulty of the analysis. The first problem was the necessity of reducing the number of attributes used as input in the multiple regression model from 82 to 15. Glmnet Lasso was used for dealing with the first problem. The fact that nominal variables should be converted to numeric before they can be used as input in the regression model lead to the second problem. All nominal variables’ levels were transformed to individual dummy variables and were then treated as binary attributes. Finally, multicollinearity, the phenomenon in which predictor variables in a multiple regression model are highly correlated complicated the attribute selection procedure.

In this paper, multiple linear regression models are applied to the ANOVA enhancement and the model which performs best will be further analyzed, interpreted and evaluated. The experimental results presented in this paper (based on the Boston Housing data set) indicate the performance of the better fitted model relative to the other models.

Only 15 attributes were used as input in the best fitted model. According to this paper’s analysis the variables which foremost define the housing price are the overall quality of the house, the type of zone where it is located, the year when it was built and remodeled, the number of parking spots and the garage type, the size of living and basement area, whether it has fireplaces or not, the quality of the basement and the kitchen, and finally the total number of rooms above ground.

# Data preparation

The dataset needed to be cleaned and to be applied to variable transfromations before any further analysis.

## Data cleansing

Columns “Alley”, “Misc.Feature”, “Fence”, “Pool.Qu” and “Fireplace.Qu” were deleted because more than 80% of their values were missing.

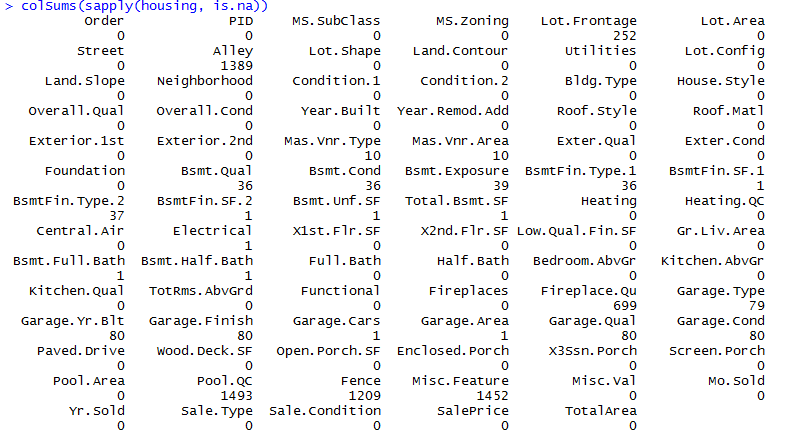


Figure 1 Number of null values per column

The rest of the columns which contained null values were split into categories. For each category different procedure was used for handling the missing values. Individual categories contained correlated attributes. Each category’s inner correlation was used for predicting missing values in different columns of the same category.

#### Garage Category

Garage related attributes contained "Garage.Type", "Garage.Yr.Blt", "Garage.Finish", "Garage.Qual","Garage.Cond","Year.Built", "Garage.Cars" and "Garage.Area”. There were 79 observations missing GarageType values, 1 observation with nulls in GarageArea and GarageCars, 159 observations with nulls in GarageYrBlt, GarageFinish, GarageQual and GarageCond. The rest of garage related attributes were used for predicting missing values in the Garage.Area, GarageCars, GarageYrBlt columns. Observations with nulls in all the parking related columns, were completed with the value “No parking” in case of nominal attribute or “0” in case of numeric.

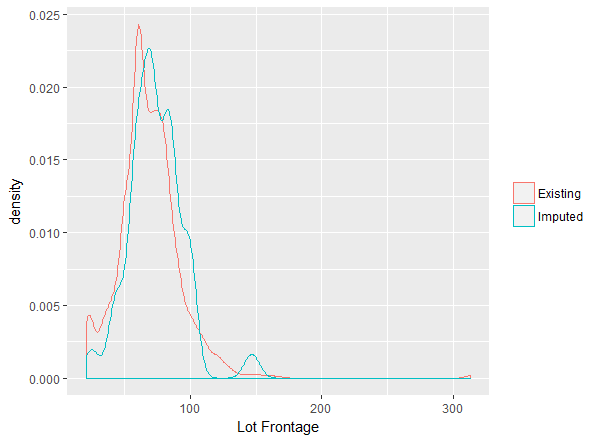
#### Basement related attributes

Columns "Total.Bsmt.SF", "Bsmt.Exposure", "Bsmt.Cond", "Bsmt.Qual","Bsmt.Fin.Type.1", "Bsmt.Fin.Type.2", "Bsmt.FinSF.1","Bsmt.Fin.SF.2", "Bsmt.Unf.SF" were used for predicting missing values in the rest basement related attributes.

#### MasVnrType

Contained 10 missing values. Since one square feet of masonry veneer area in square feet seem to be very unlikely, the 3 observations with “MasVnrArea” equal to 1 were replaced with the value “0”. Fields with missing values in both “MasVnrType” and “MasVnrArea”, were also replaced with “None” or “0” accordingly.

#### Lot.Frontage



*Figure 2 plot of existing and imputed values of “Lot.Frontage”*

Columns "MS.SubClass", "MS.Zoning", "Lot.Frontage", "Lot.Area", "Street", "Lot.Shape", "Land.Contour", "Lot.Config", "Land.Slope", "Bldg.Type", "House.Style", "Yr.Sold", "Sale.Type", "Sale.Condition" were used for predicting the missing values of the “Lot.Frontage” column.

Predicted values fit our data well enough and therefore they will be used for replacing the missing values.

#### Pool related attributes

Only 7 houses contained a swimming pool. A new binary column named “Has.Pool” was defined containing the value “1” in houses with swimming pool and “0” in those without swimming pool.

## Data transformation

### Numeric attributes

Six new attributes were calculated. The first one was named “TotalArea” and contained the sum of the living area and the total basement area. The second one was named “PriceTotalSQRM” and contained the price per square feet or the sale price divided by the total area. The third column was named “PriceLivSQRM” and contained the sale price divided by the living area. A binary column named “Has.Fireplace” was created containing the value “1” in houses with fireplace and “0” in those without fireplace. Another binary column named “Remodeled” was created containing the value “1” in houses that have been remodeled and “0” in those who haven’t. Finally, a new column named “DecadeBuilt” was created as an alternative to the column “Year.Built”.

### Categorical attributes

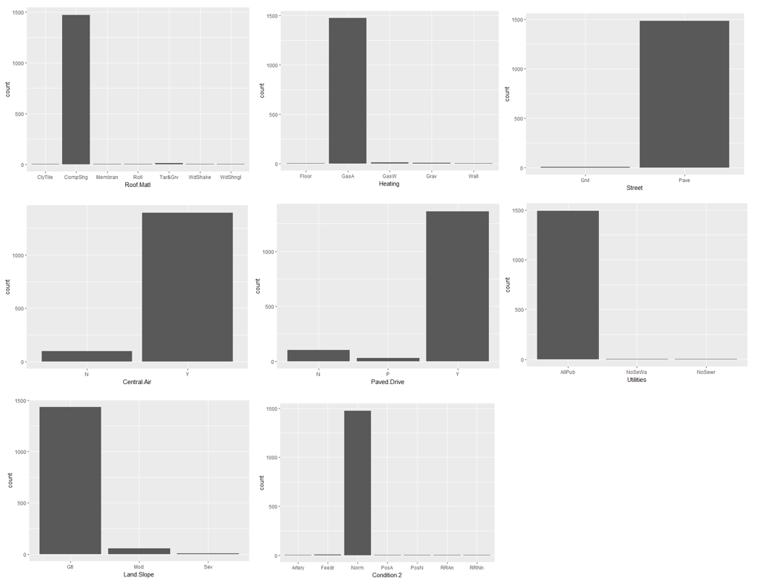
The dataset was then split into two parts, a part with all the numeric values and a part with all the categorical values. Categorical variables must be transformed to dummy variables. A new binary attribute will be created for every unique value/category of the categorical values.



Figure 3 Converting categorical attribute "MS.Zoning" to binary dummy attributes

### High correlated attributes

Highly correlated attributes were examined further. For example the correlation between “Gr.Liv.Area” and “TotalArea” was as high as 0.88, the correlation between “Fireplaces” and “HasFireplace” was 0.89, between “Garage.Area” and “Garage.Cars” was 0.88 and the correlation between “MS.SubClass” etc. In cases of high correlation between two columns, only one was kept, the one with the highest correlation to the sale price. Moreover, categorical variables were represented with bar plots for identifying which attributes do not provide additional information and should therefore be deleted. According to the following plots and the collinearity between attributes, columns “Fireplaces”, “Garage.Area”, ‘Street”, “Utilities”, “Land.Slope”, “Condition.2”, “Roof.Matl”, “Heating”, “Central.Air”, “Paved.Drive” were deleted.



*Figure 4 Barplots used to identify which attributes do not provide additional information. These attributes will be deleted.*

# Exploratory analysis

Categorical Variables

The following bar plots and density plots describe the most important categorical and numeric variables accordingly.

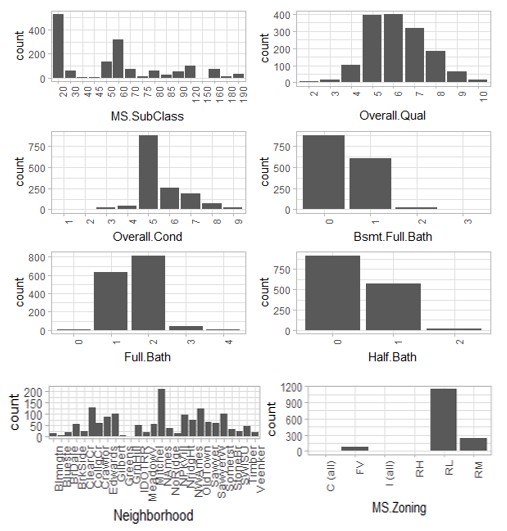
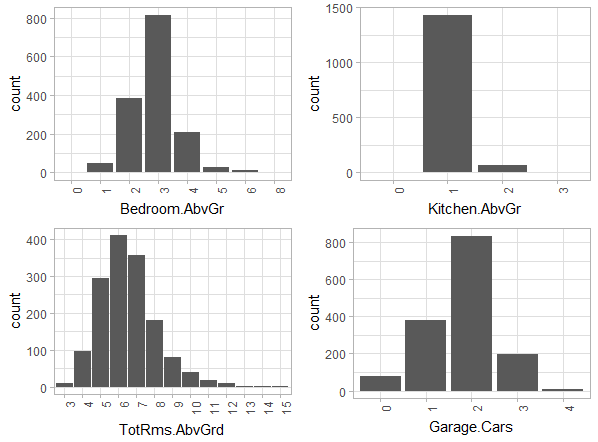
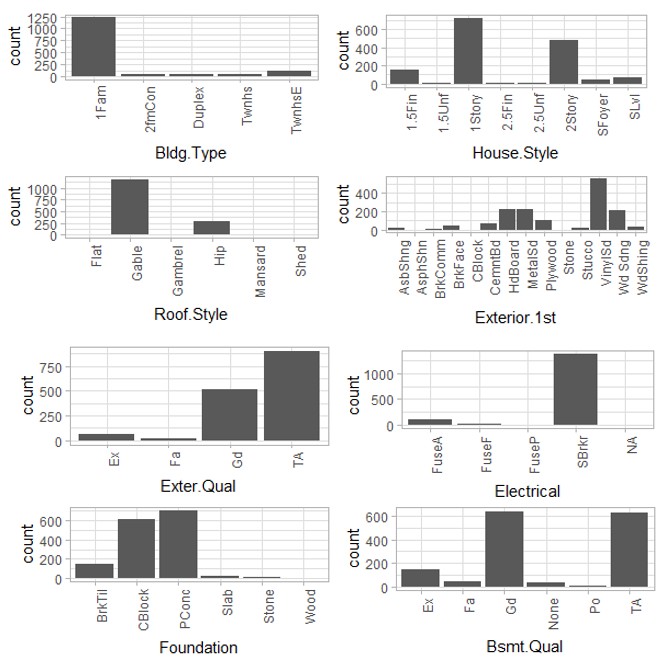


Figure 5 Bar Plot1: Categorical variables

A typical house is identified by the dwelling class “1-story 1946 & newer all styles”, it has “Above average” overall quality and “Average” overall condition. It is located in residential low density areas, contains two full baths, zero half baths, three bedrooms, one kitchen, six rooms above ground and two parking spots. Most of the listed houses are located at the “North Ames” neighborhood.



The most common building is “Single family detached” with one story. Most of the houses have gable roof, vinyl siding average exterior and basement quality. A typical building have poured concrete foundation and standard electrical circuit breakers & romex.



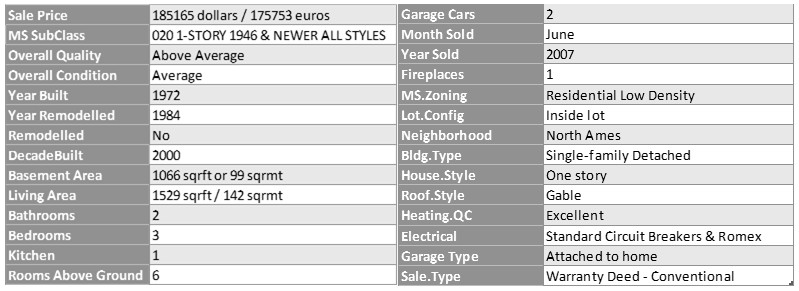
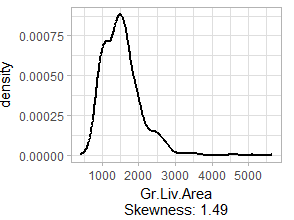
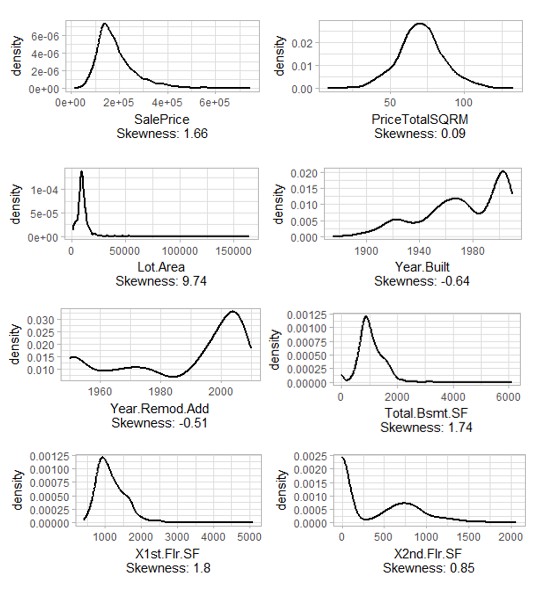
*Figure 6 Bar Plot2: Categorical variables*

## Numeric Variables

The average sale price is 185165 dollars or 175753 euros. Most of the dataset’s houses were built during the decade 2000-2010. Most of the houses haven’t been remodeled but most of those who did, they got remodeled in the period 2000-2012. An average house has living area size of 1529 square feet or 142 square meters and basement area of 1066 square feet or 99 square meters.

*Figure 8 Density plot 1: Numeric variables*

*Figure 9 Density Plot: Numeric variables*



*Figure 7 Density plots: Numeric variables*

Figure 10 Average home characteristics, computed as mean of numeric variables and mode of categorical variables

# Attribute selection

#### Merged dataset

Categorical and numeric attributes were cleaned and transformed separetely. Subsequently they were merged again in one numeric dataset. The new merged dataset contained 255 attributes. The number of atributes was significantly increased due to the transofrmarion of categorical values to dinary dummies. Only variables with correlation to price greater than 0.25 or less than -0.25 were kept, resulting a merged dataset of 57 attributes. Lasso was then used for further attribute reduction. Both “lars” and “glmnet” lasso packages were implemented. The intersection of the two different packages’ outputs resulted 20 attributes. These attributes are highly correlated to the sale price and they will be used as input of the multiple regression model.

## Lasso attribute selection

#### Lars package

Having a set of input measurements x1, x2 ...xp and an outcome measurement sale price, the lasso fits a linear model yhat=b0 + b1\*x1+ b2\*x2 + ... bp\*xp. Lars lasso aims to minimize sum( (y-yhat)^2 ) subject to the tuning parameter “s”. When "s" is large enough, the constraint has no effect and the solution is just the usual multiple linear least squares regression of y on x1, x2, ...xp. However for smaller values of s (s>=0) the solutions are reduced versions of the least squares estimates. Often, some of the coefficients bj are zero. S defines the number of predictors to use in a regression model. In this case the “S” minimizing the Cp statistic was used (figure 12). The Cp statistic is an estimate of the mean-square error in a model based on a selected subset of predictors, corrected for the number of predictors. Each colored line in figure 11 represents the value of a different coefficient in the model.

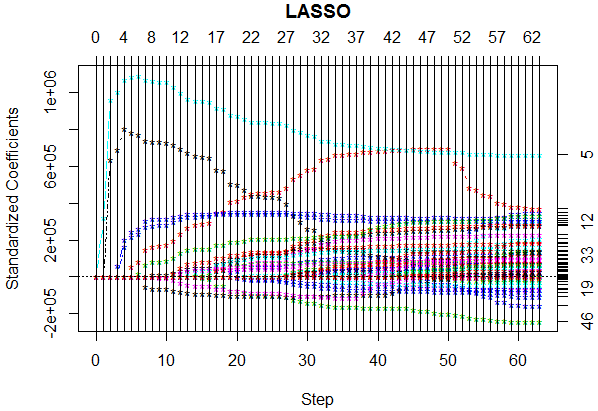
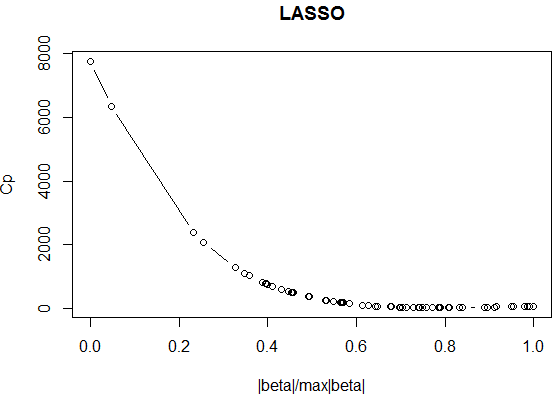
 

Figure 11 Lars Lasso coefficients Figure 12 S corresponding to the optimal Cp equals to 0.8.

According to “Lars lasso” the attributes followed by a zero coefficient, should be excluded from the model (Figure 13).

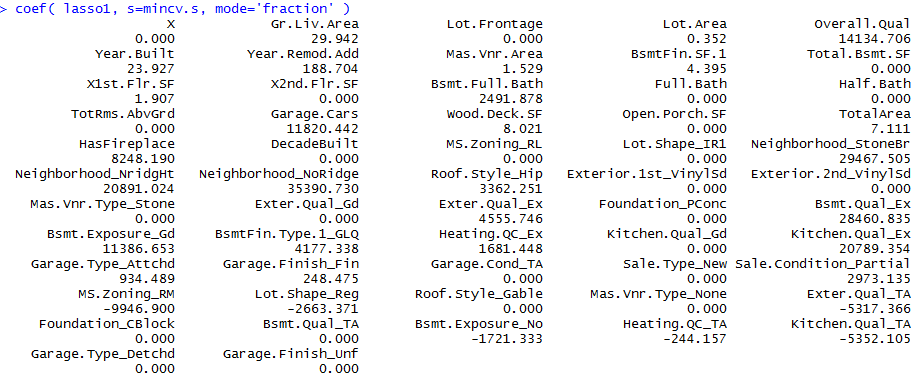
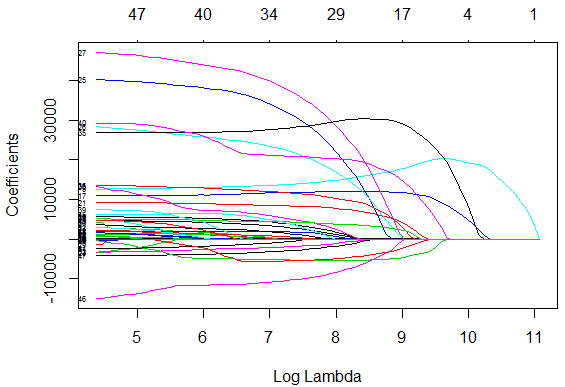


Figure 13 Lars Lasso coefficient selection

#### Glmnet package

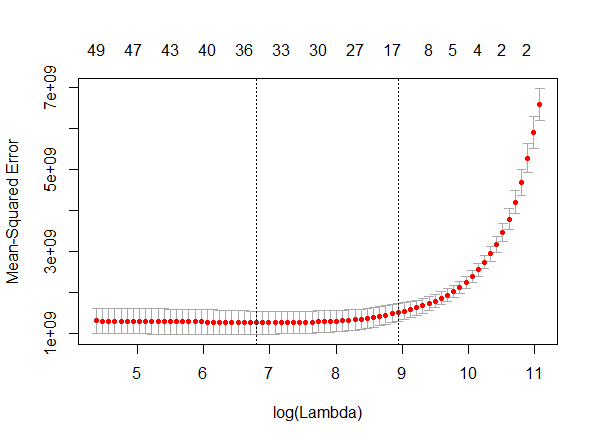


*Figure 14 Glmnet lasso coefficients*

Each colored line of figure 14 represents the value of a different coefficient in the model. Lambda (figure

16) is the weight given to the regularization term (the L1 norm). When lambda approaches zero, the loss function of the model approaches the ordinary least squares (OLS) loss function. OLS is the method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed and predicted values. Therefore, when lambda is very small, the LASSO solution should be very close to the OLS solution, and all of the coefficients will be included in the model. L1 norm is the regularization term for LASSO. So when L1 norm is small, the regularization is high. Therefore, an L1 norm of zero gives an empty model, and increased L1 norm, result

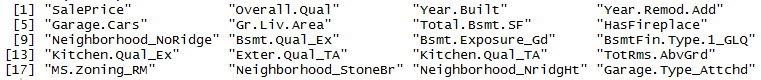
more variables to be characterized as significant. Lambda.min is the value of λ that gives minimum mean cross-validated error. However lamda.min was too complex and over fitted. lambda.1se, returns the most regularized model such that error is within one standard error of the minimum. The outputs for both lamda min and lamda 1se were taken under consideration during the final attribute selection.



*Figure 15 Glmnet lasso lamda min*

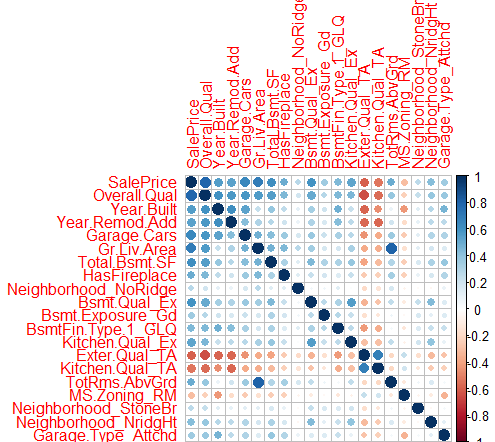
## Selected attributes

Combining the outputs of “lars” and “glmet” lasso the following attributes are considered to be the most important for predicting the sale price.



*Figure 16 Lasso selected attributes*

Figure 17 represents the correlations between the selected attributes. Existing collinearity issues will be further examined. With high positive correlation is represented with dark blue color while high negative correlation with red. Lightly colored cells represent low correlation and white cells no correlation at all.



*Figure 17 Lasso selected attributes collinearity*

# Regression models

We will start our analysis with both ways stepwise regression. We will then check whether the model meets the linear model’s assumptions.

## M0

We will perform stepwise regression. For that reason we need to fit a full linear model (figure 17) which will include all the variables selected in the previous chapter. Furthermore, we need to fit a null model which will only include the constant value and no variables.

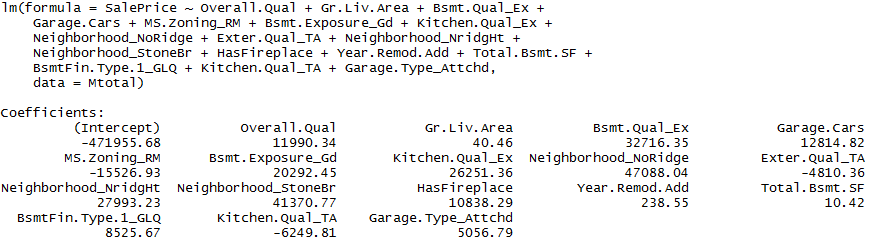


Figure 18 m0: Stepwise regression full model

The model consists of 17 variables.

<http://www.statsmakemecry.com/smmctheblog/confusing-stats-terms-explained-heteroscedasticity->

heteroske.html

<http://r-statistics.co/Outlier-Treatment-With-R.html>

M0: SalePrice = 1.199e+04\*Overall.Qual + 4.046e+01\*Gr.Liv.Area + 3.272e+04\*Bsmt.Qual\_Ex + 1.281e+04\*Garage.Cars - 1.553e+04\*MS.Zoning\_RM + 2.029e+04\*Bsmt.Exposure\_Gd + 2.625e+04\*Kitchen.Qual\_Ex + 4.709e+04\*Neighborhood\_NoRidge - 4.810e+03\*Exter.Qual\_TA + 4.137e+04\*Neighborhood\_StoneBr + 2.779e+04\*Neighborhood\_NridgHt + 1.084e+04\*HasFireplace + 2.385e+02\*Year.Remod.Add + 1.042e+01\*Total.Bsmt.SF + 8.526e+03\*ΒsmtFin.Type.1\_GLQ - 6.250e+03\* Kitchen.Qual\_TA + 5.057+03\*Garage.Type\_Attchd – 4.720e+05 + ε

### Interpretation

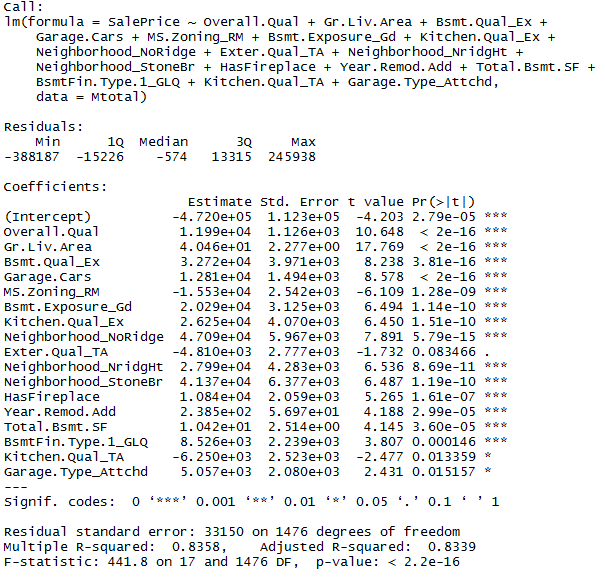
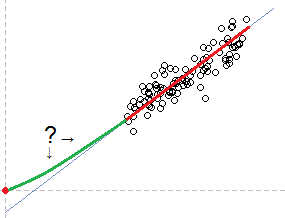


Figure 19 Model M0 evaluation

When all variables are zero, we have a cost of -4.720e+05 or - 472000$. This cost usually reflects the fixed amount you have to pay for every sale when all variables equal to zero. It can be thought as a fixed cost. In this case the intercept is negative. This may lead to two different assumptions. The model can predict a relationship identified on data over a limited range as shown on figure 20. Zero values on all variables are not included in this limited range and as a result the model does not perform well in this case. However this makes us question the assumption of linearity of relationship



*Figure 20 Negative intersect cause*

between price and variables. The assumption of linearity will be furthered examined.

Now concerning the estimate of the variables coefficients, increase by 1 causes sale price to change by b, when everything else stays the same. So for example if we increase the living area square by 1 square feet ceteris paribus, this will result an increase of 4.046e+01= **40.46**$ or 38€ in the sale price. Concerning dummy values, houses which have a fireplace cost 1.084e+04= **10840**$ or 10205€ more than houses that do not include a fireplaces.

T-values shows whether the coefficient is meaningful for the model and it is used to calculate the p-value. P-value is the probability that the variable is not relevant. The lower the better. The stars on the right indicate that it is unlikely that no relationship exists. More stars are better.

R-squared is the percent of the standard deviation described by the model.

Residual standard error (s) is an overall measure of how well the model fits the data. It represents the average distance that the observed and the predicted values fall. The lower the s is the more accurate the model prediction is. This number is the standard deviation of the residuals and it should be proportional to the quantiles of the residuals. For a normal distribution, the 1st and 3rd quantiles should be 1.5 +/- the std error.

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p- value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p- value is likely to be a meaningful addition to our model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

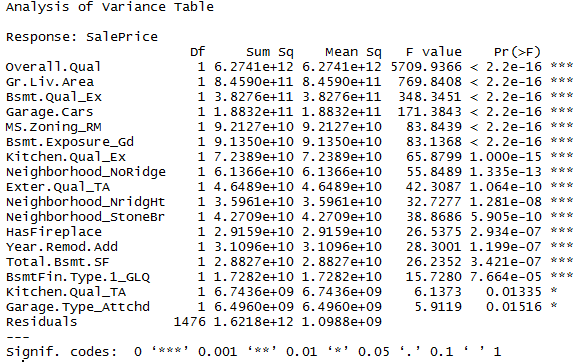


Figure 21 Model M0 evaluation

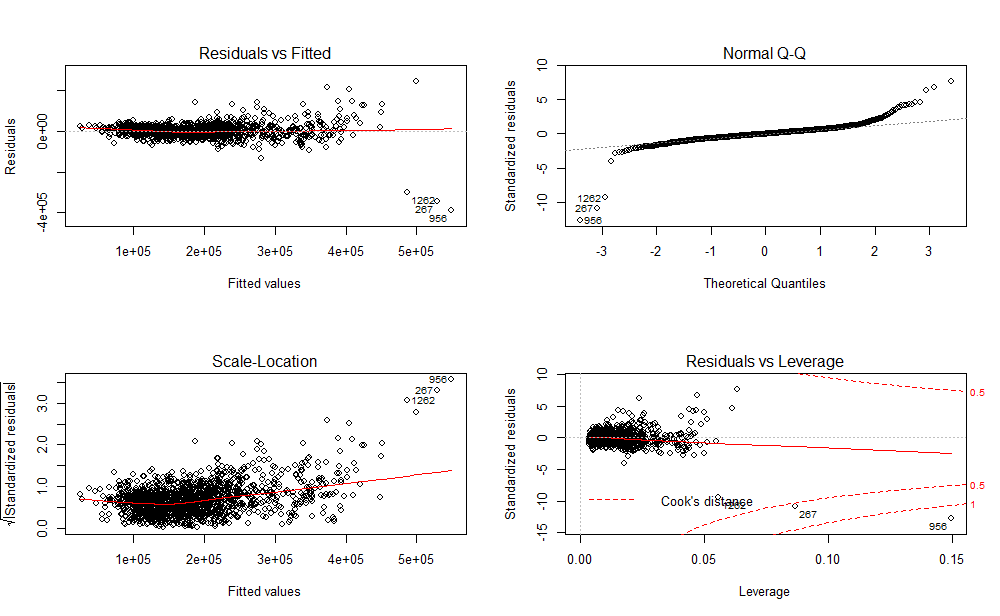


Figure 22 m0: Plot linear full model

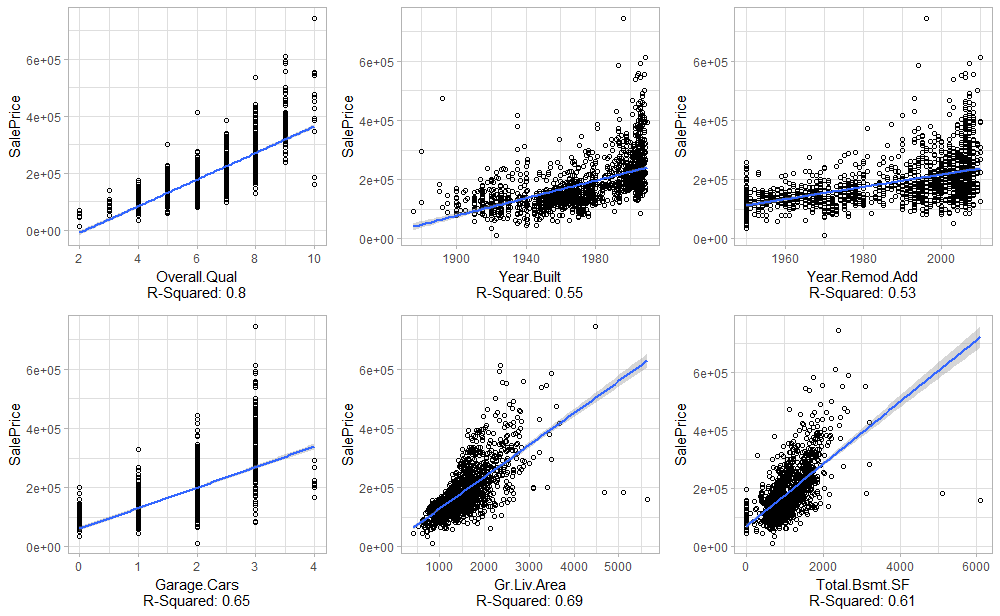
### Linear model assumptions

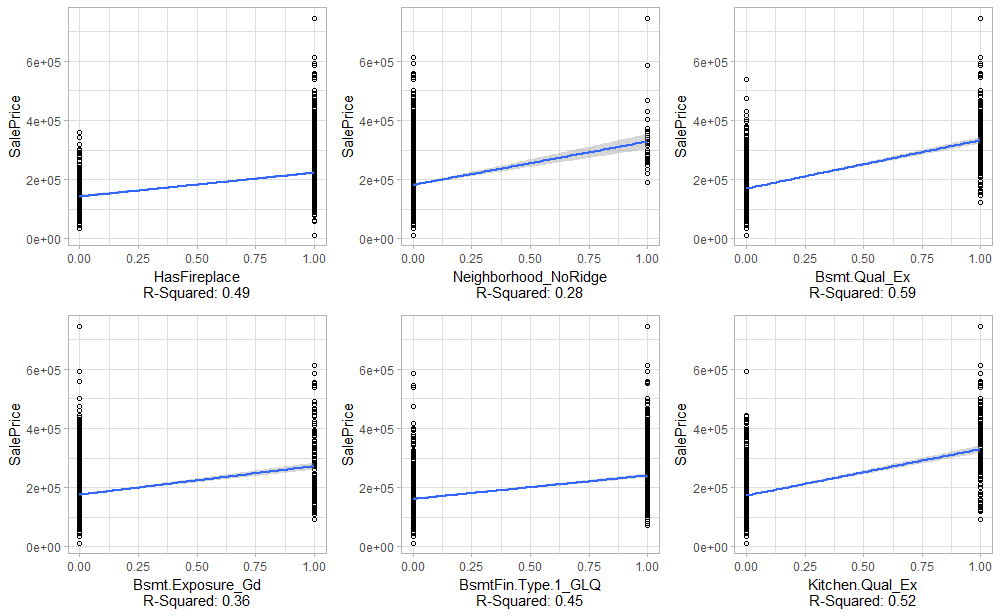
When using linear models we make four assumptions and therefore we should check whether model m0 meets these assumptions. Having Figure 22 as reference, we should check the following.

* + - 1. Linear relationship between the variables and the Sale Price

The first plot represents the residuals versus the fitted values. The red line is a smoothed curve that passes through the actual residuals (grey dashed line). The two lines should ultimately identify. Note that several points are numbered, and these are points that we need to pay special attention to. Residuals in the residuals vs fitted graph of plot 17 seems to be symmetrically distributed around a diagonal line and do not form a curve.

According to the following partial residual plots, we can see that all variables have linear correlation to sale price except maybe from Gr.Liv.Area and Total.Bsmt.SF. Therefore, log and polynomial transformation on the Gr.Liv.Area and Total.Bsmt.SF will be attempted to examine whether it will improve the model.





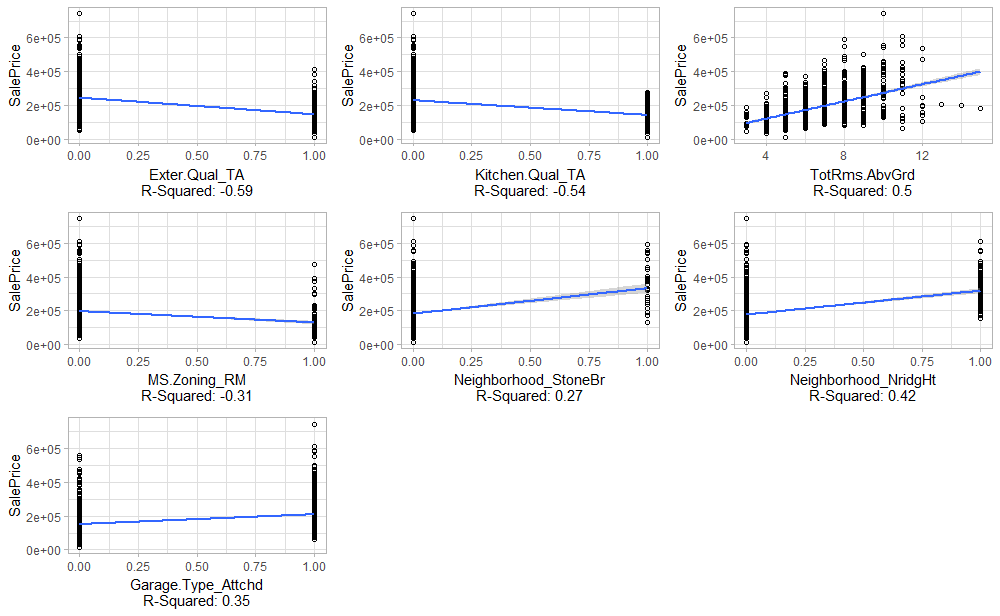


Figure 23 m0: Testing linearity of correlation between individual variable and price

* + - 1. Normality of residuals

This assumption will be checked with the Q-Q-Plot on figure 22 and 24 and with the goodness of fit test Kolmogorov-Smirnof test.

According to the linear model’s assumptions, the errors on the **qqnorm() plot of the residuals** should be normally distributed. This plot evaluates the normality of errors assumption. If the errors (residuals) were precisely normally distributed, they would lie exactly on this line. Some deviation is to be expected, particularly near the ends, but the deviations should be small in contradiction to this case.

The qqplot diagram is heavy tailed (figure 22 and 24). We can investigate further this problem by examining the plot of the studentized residuals (figure 23) and the residual’s density plot (figure 21). Moreover the bottom right plot on figure 22 shows the standardized residuals against leverage. The normally distributed standardized residuals are expected to be centered around zero and reach 2 – 3 standard deviations away from zero. In this case the residuals are not normally distributed. Finally, the p-value of both Lilliefors (observations > 50, figure 27) and Saphiro-Wilk tests (figure 28) is less than .05 which means that the residuals do not come from a normally distributed population. A non-linear transformation

such log or polynomial transformation might fix this issue, however might affect the multicollinearity.

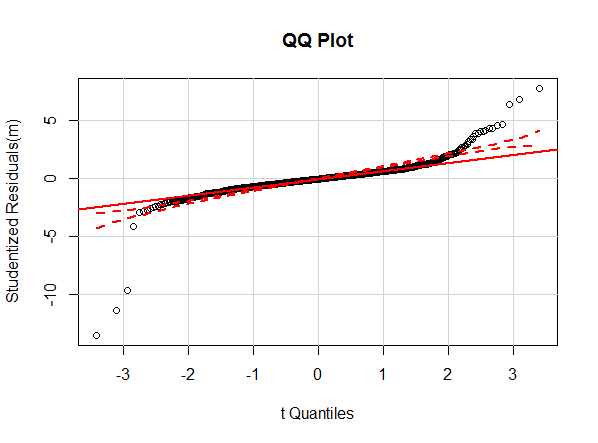


Figure 24 Mo QQ Plot

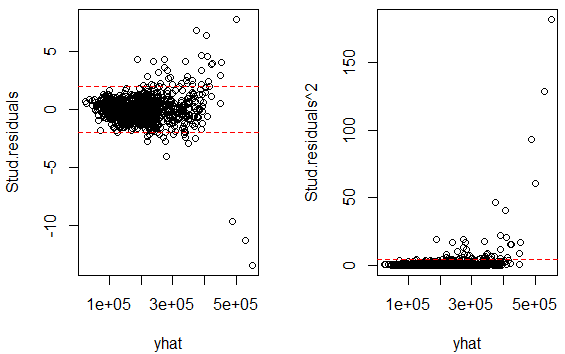
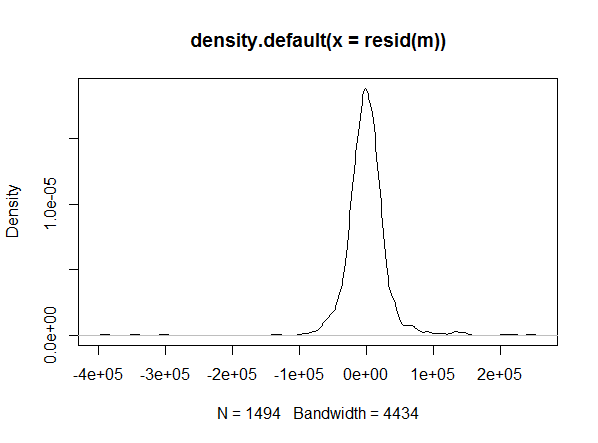
 

Figure 25 M0 plot of studentised residuals Figure 26 m0: residuals density plot

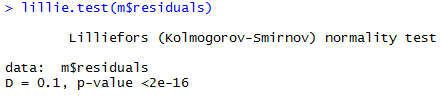


Figure 27 m0: Lilliefors test, testing normality of residuals

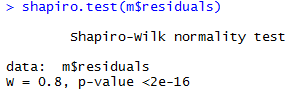


Figure 28 m0: Shaphiro-Wilk test, testing normality of residuals

* + - 1. Homoscedasticity of residuals

The third plot (figure 22) represents the square root of the standardized residuals, which are residuals rescaled so that they have a mean of zero and a variance of one. This plot has large residuals (both positive and negative) plotting at the top and smaller residuals plotting at the bottom. Completely random points would mean absolute homoscedastic but in the above diagram this is not the case. In this case the residuals seem to increase as the fitted Y values increase which means existence of heteroscedasticity. We will also test this assumption with the No-constant variance score test (ncvTest) which is used for testing homoscedasticity. P-value is less than 0.05. Which means that we reject the null hypothesis that the variance of the residuals is constant and infer that heteroscedasticity is present.



Figure 29 mo: ncvTest, testing homoscedasticity

* + - 1. Collinearity

Collinearity is a linear relationship between two explanatory variables and should not be present.

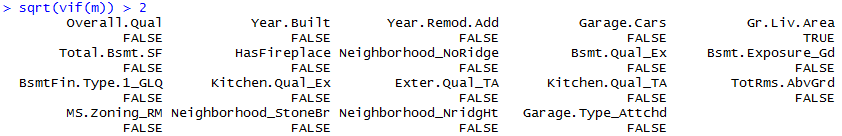


Figure 30 Vif testing collinearity

* + - 1. Autocorrelation

Autocorrelation is a characteristic of data in which the [correlation](http://www.statisticssolutions.com/resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman) between the values of the same variables is based on related objects and violates the assumption of instance independence. It generally exists in data- sets in which the data, instead of being randomly selected, is from the same source.

Durbin Watson test results p-value greater than 0.05 and therefore we do not reject the null hypothesis that there is no correlation among the residuals.



Figure 31 m0: Durbin Watson Test, testing autocorrelation

Concluding, non of the linear model’s assumptions are met in the m0 model.

## Prediction

### k-fold cross validation

When evaluating models, we often want to assess how well it performs in predicting the target variable on different subsets of the data. One such technique for doing this is k-fold cross-validation, which partitions the data into k equally sized segments (called ‘folds’). One fold is held out for validation while the other k- 1 folds are used to train the model and then used to predict the target variable in our testing data. This process is repeated k times, with the performance of each model in predicting the hold-out set being tracked using a performance metric such as accuracy. We will use the most common variation of cross validation, the 10-fold cross-validation. The dataset was split in 10 equal parts (folds) of size 149. The first nine folds were used for training the model and the last one for testing the model. The diagram represented below represents the performance. The prediction of the 10-fold-cross-validation was not accurate and the mean square error (mse) was high, equal to 2.83+11.

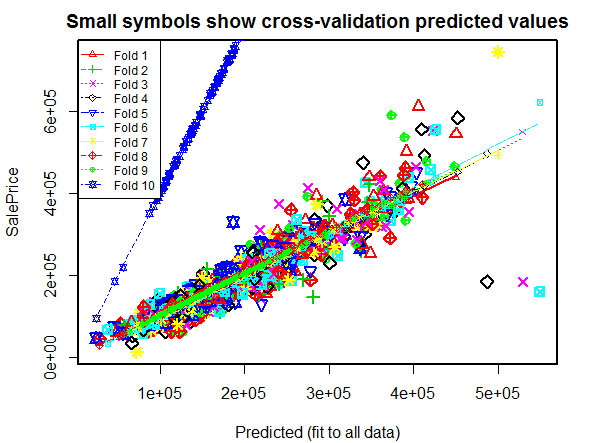


Figure 32 Leave one out 10 fold cross validation



### Test data prediction

We will now check how our model performs when the test dataset is used as a test set. The mse now is significantly lower and therefore the prediction is much better. The mse now equals to 1.27e+09 against

2.83+11 which was resulted when the test set came from the same dataset as the training set. This makes us worried because it is either caused by the existence of outliers in our training set or some significant error emerged during the replacement of null values in the data preparation section.

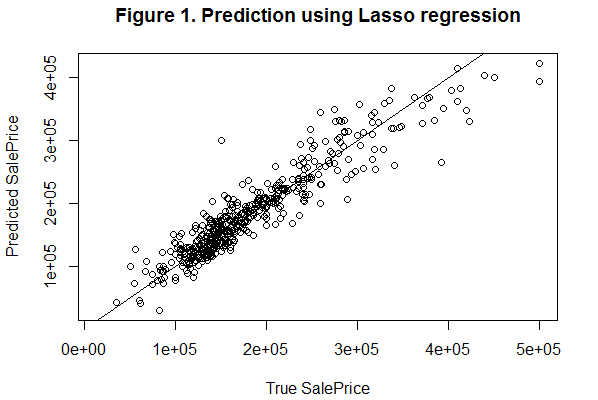


Figure 33 M0 prediction on test set



## Model improvements

### Outliers

* + - 1. Identifying outliers

In the residual vs leverage plot (figure 22), a point far from the centroid with a large residual can severely distort the regression. In this plot the red line should be close to the horizontal gray dashed line and no points should have a large Cook’s distance (i.e, >0.5). Point 956, the one far below the line at the bottom right of the data plot is flagged as having a large Cook’s distance and it has really high negative residual (figure 22, upper right plot). ) Note that the numbered points appear in all the graphs. They have large negative residuals because they plot below the regression line.

Cook’s distance is a measure computed with respect to a given regression model and therefore is impacted only by the X variables included in the model. It computes the influence exerted by each data point (row) on the predicted outcome.

The cook’s distance for each observation i measures the change in Y^Y^ (fitted Y) for all observations with and without the presence of observation i, so we know how much the

observation i impacted the fitted values. In general use, those observations that have a cook’s distance greater than 4 times the mean may be classified as influential.

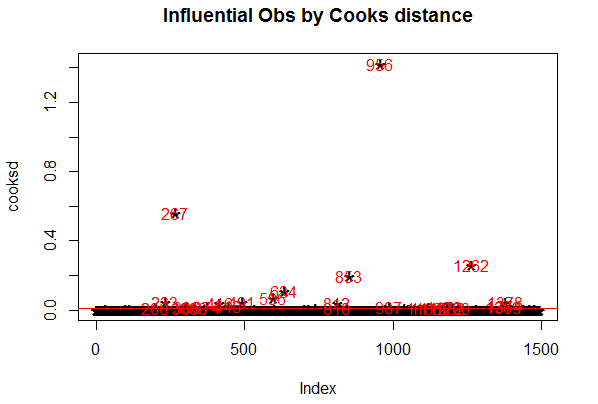


Figure 34 Identifying outliers with Cook’s distance

We will next check the rows with the largest absolute studentised residual.

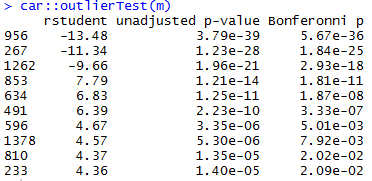


Figure 35 Identifying outliers with the largest absolute studentised residual

* + - 1. Evaluating outliers

Now lets find out the influential rows from the original data. If you extract and examine each influential row 1-by-1 (from below output). It is likely that one of the X variables included in the model had extreme or false values. For example row 267 has perfect overall quality equal to 10, has extremely large living and basement area which are both equal to 5095 square feet (this is probably a misspelling during the data

entry) and it has 15 rooms above ground when an average house has 6 rooms. It is clear that there is some misspelling error in the rows 956 and 267 and therefore we will delete them.

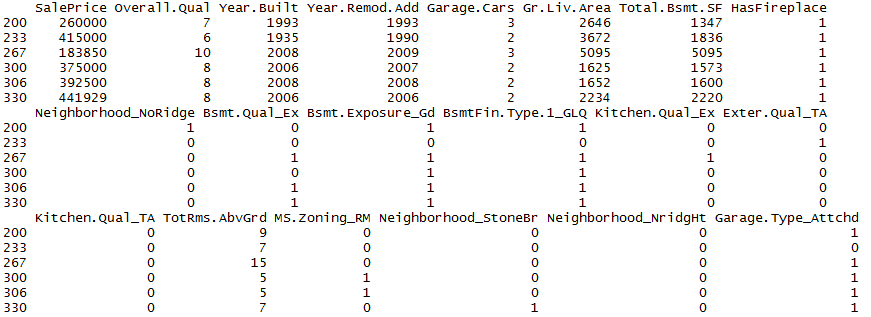


Figure 36 Evaluating outliers by identifying influential rows

### Non-normal transformations

When the residuals are not normally distributed it is suggested that we use with polynomial or logarithmic transformations.

* + - 1. Logarithmic transformations

When applying a logarithmic transformation to the sale price, the sale price distributions approaches the normal distribution. This can be seen in figure 38. Therefore in our next model we will use logarithmic transformation to the price.

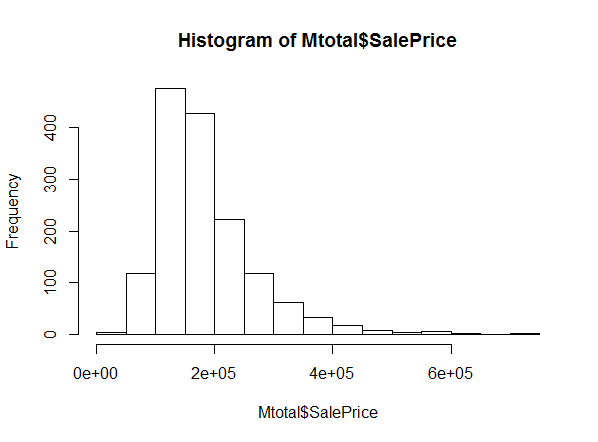
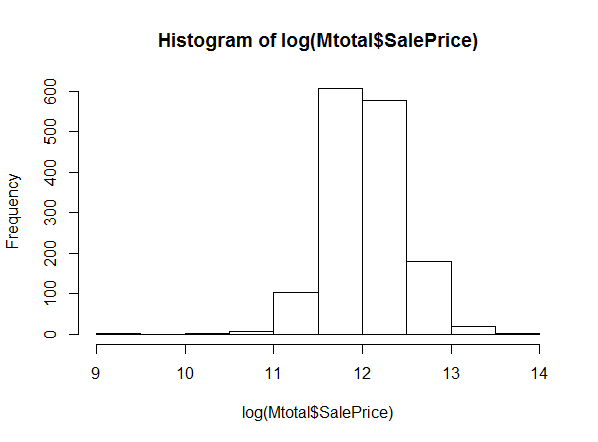
 

Figure 37 Sale price histogram Figure 38 Log(SalePrice) histogram

* + - 1. Polynomial transformation

As it was mentioned earlier, when checking the linearity assumption, variables “Ge.Liv.Area” and “Total.Bsmt.SF” insinuate a nonlinear relationship to price. We will now test whether a polynomial transformation in these two variables will improve the model.

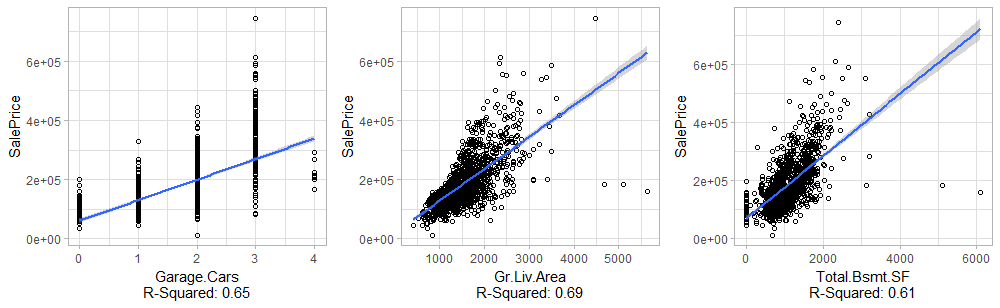


Figure 39 m0: Testing linearity of correlation between individual variable and price

Leave one out cross validation was used to measure the errors for polynomials fit with different degrees. Each observation used in the model development is removed in turn and then the model is refitted with the remaining observations. Then the out-of-sample prediction for the refitted model is calculated with the removed observation one by one to assemble the LOO, e.g. leave-one-out predicted values for the whole model development sample. Figure 40 implies that a third degree polynomial should be used in Gr.Liv.Area variable and Figure 41 implies that a second degree polynomial should be used in Gr.Liv.Area in order to minimize the error

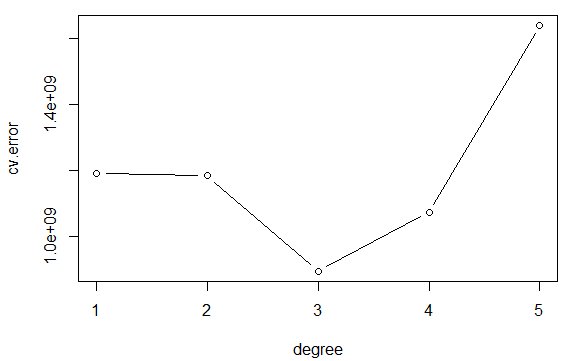
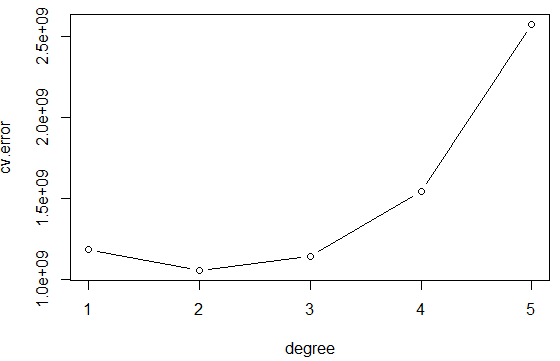
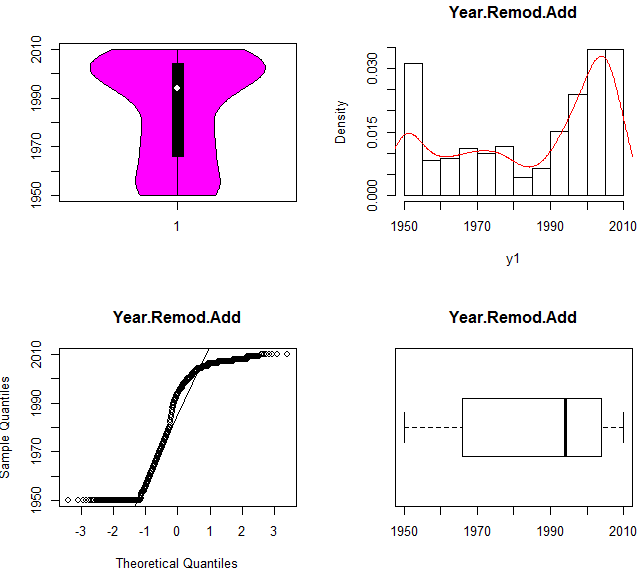
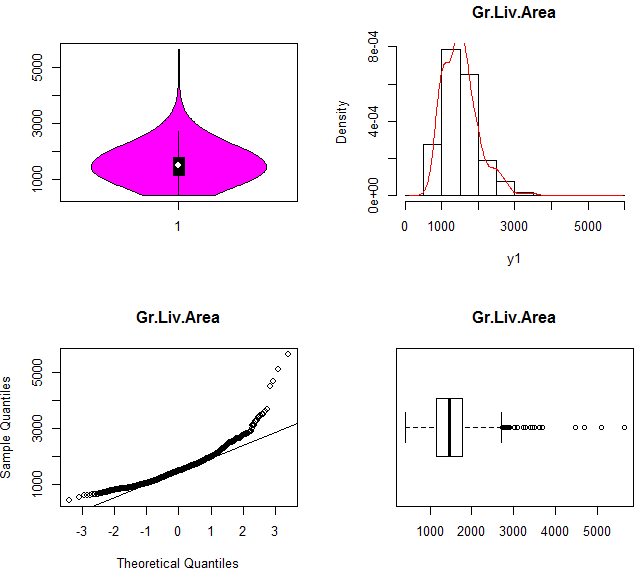
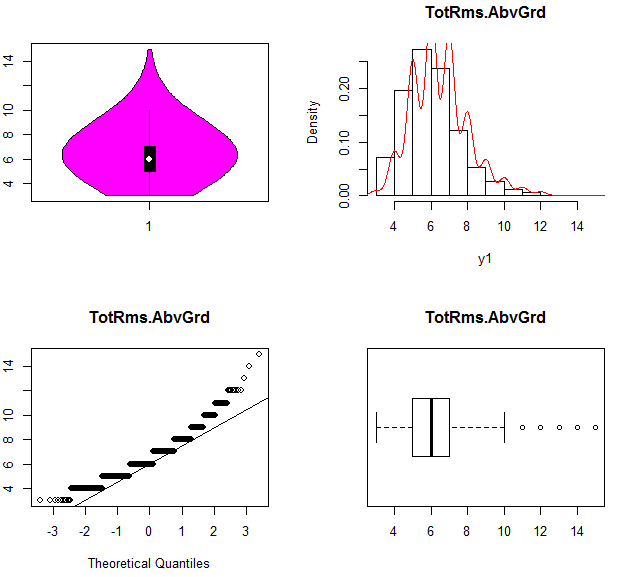
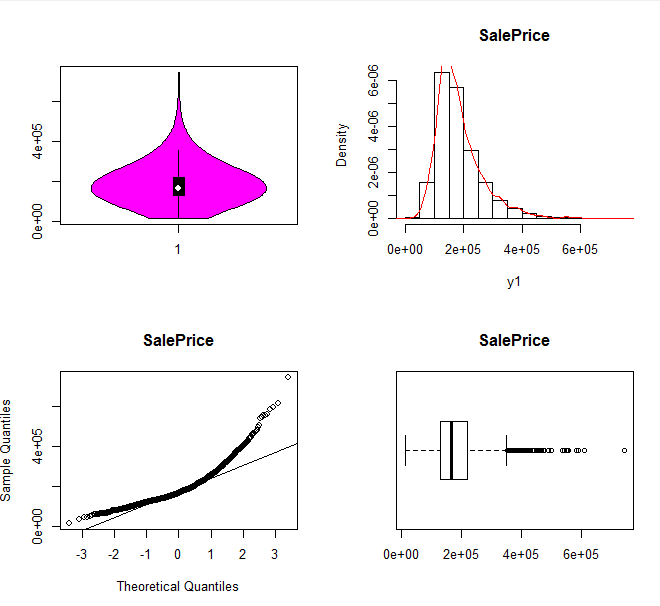
 

Figure 40 M0, Gr.Liv.Area: Fitting polynomial degrees Figure 41 M0, Total.Bsmt.SF: Fitting polynomial degrees

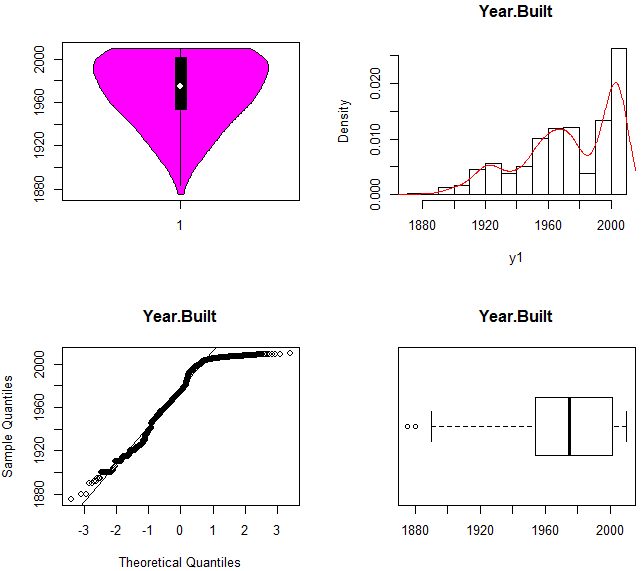
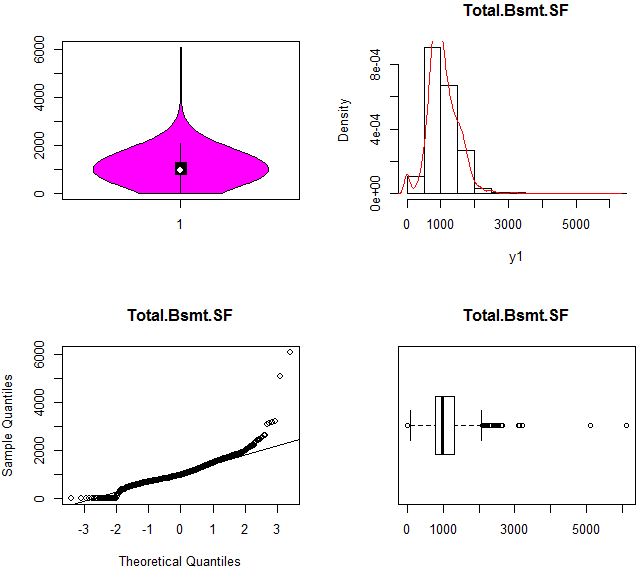
### Experimentation

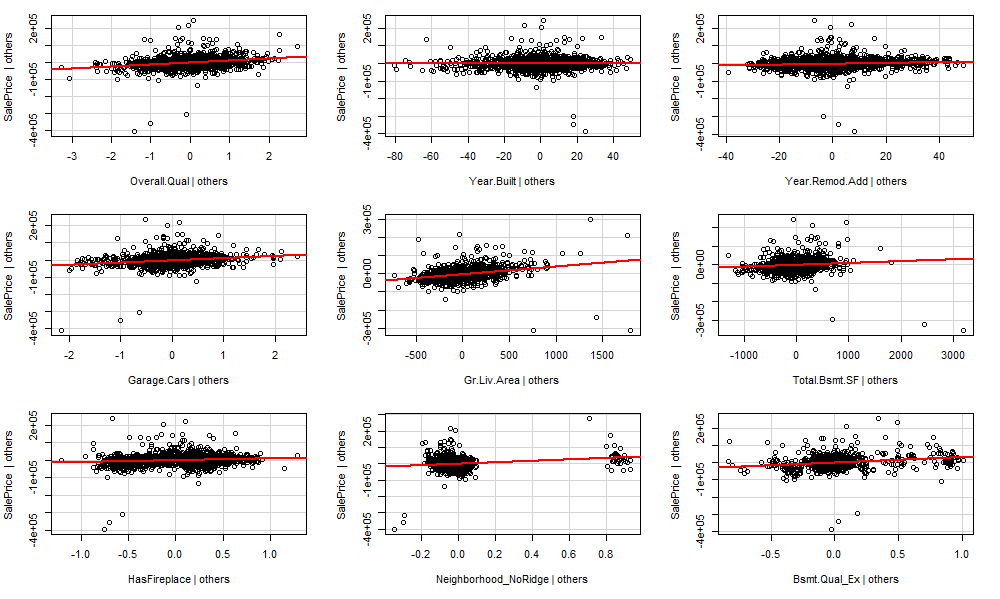
Experimentation is important in this part. Different combination of attributes were used in the models and we also experimented with polynomial and logarithmic transformations applied to several variables.

Representation of problematic variables helped to better interpret the problem (figure 42). Moreover added value plots helped as interpret how much does each variable influence the model (figure 43).



*Figure 42 Visualizing variables with problematic distribution*





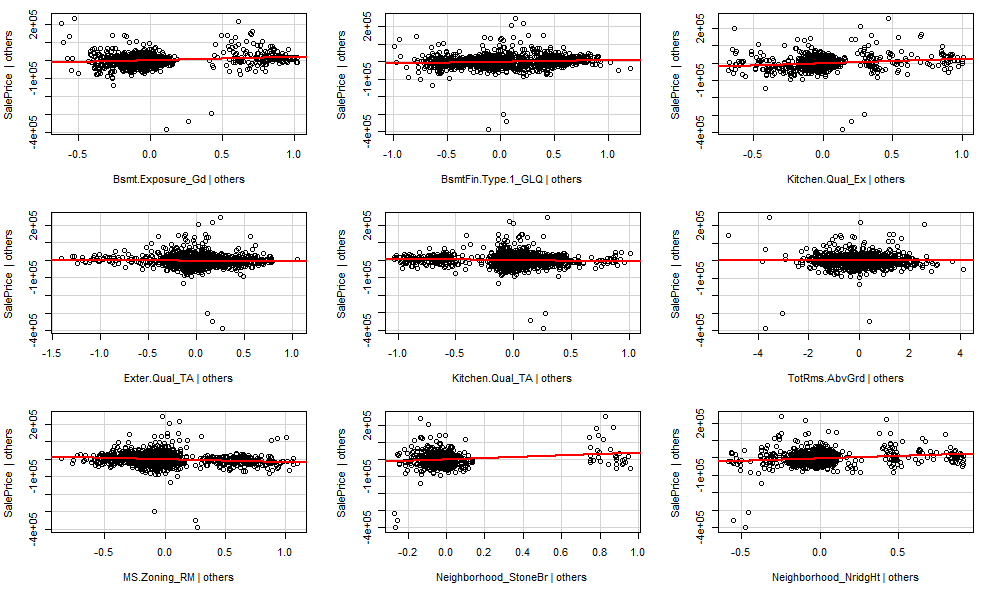




Figure 43 Added value plots

## M1

### Interpretation

After comparing different models we came to the conclusion that the attributes which lead to better models and should therefore be included in the model are the following:

Overall.Qual, Year.Built, Year.Remod.Add, Garage.Cars , Gr.Liv.Area, Total.Bsmt.SF, HasFireplace, Bsmt.Qual\_Ex , Bsmt.Exposure\_Gd , Kitchen.Qual\_Ex , BsmtFin.Type.1\_GLQ, TotRms.AbvGrd, MS.Zoning\_RM, Garage.Type\_Attchd

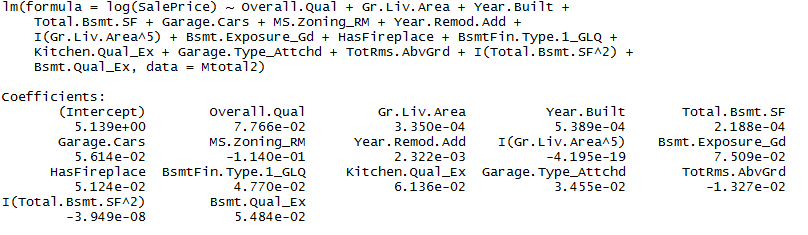


Figure 44 m1: Stepwise regression full model

The model consists of 14 variables.

<http://www.statsmakemecry.com/smmctheblog/confusing-stats-terms-explained-heteroscedasticity->

heteroske.html

<http://r-statistics.co/Outlier>

M1: SalePrice = 7.766e-02\*Overall.Qual + 3.350e-04\*Gr.Liv.Area + 5.389e-04\*Year.Built + 2.188e- 04\*Total.Bsmt.SF + 5.614e-02\*Garage.Cars - 1.140e-01\*MS.Zoning\_RM 2.322e-03\*Year.Remod.Add - 4.195e-19\*I(Gr.Liv.Area^5) + 7.509e-02\*Bsmt.Exposure\_Gd + 5.124e-02\*HasFireplace + 4.770e-02\* ΒsmtFin.Type.1\_GLQ + 6.136e-02\*Kitchen.Qual\_Ex + 3.455e-02\*Garage.Type\_Attchd - 1.327e- 02\*TotRms.AbvGrd - 3.949e-08\*I(Total.Bsmt.SF^2) + 5.484e-02\*Bsmt.Qual\_Ex + ε

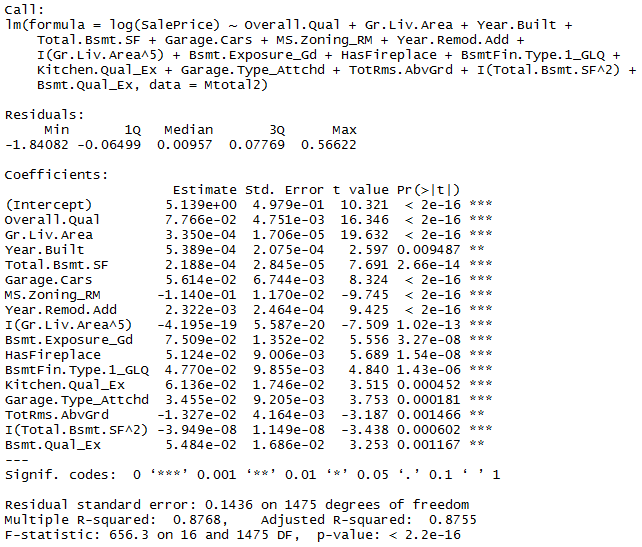


Figure 45 Model M1 evaluation

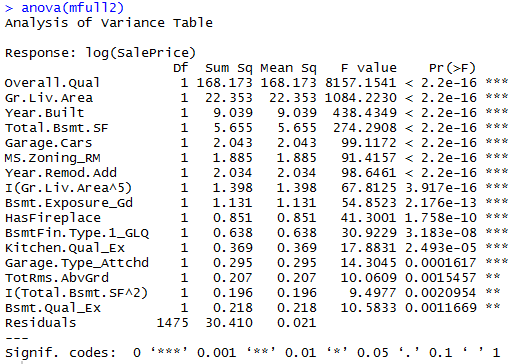


Figure 46 Model M1 evaluation

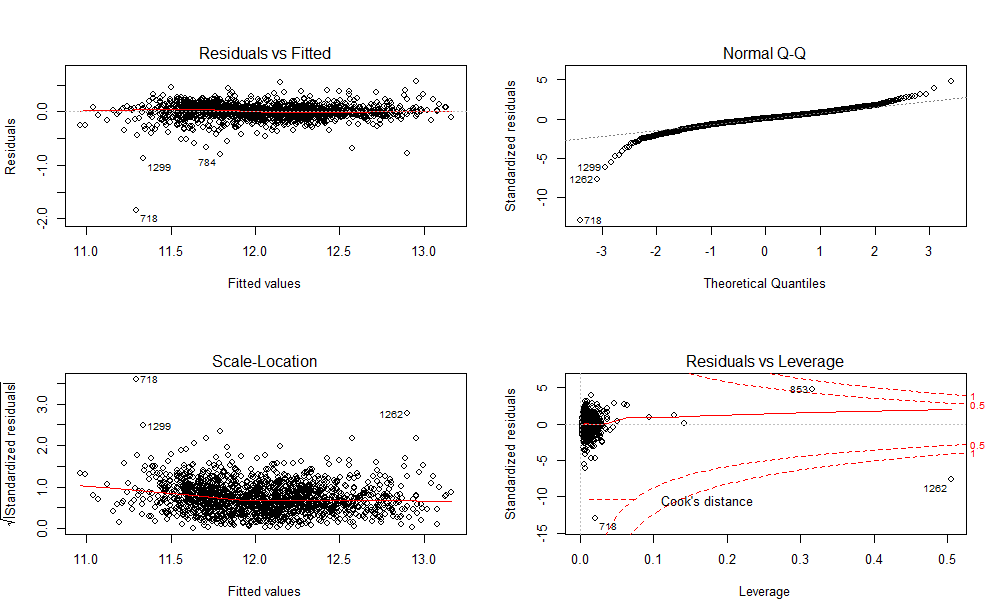


Figure 47 m1: Plot linear model

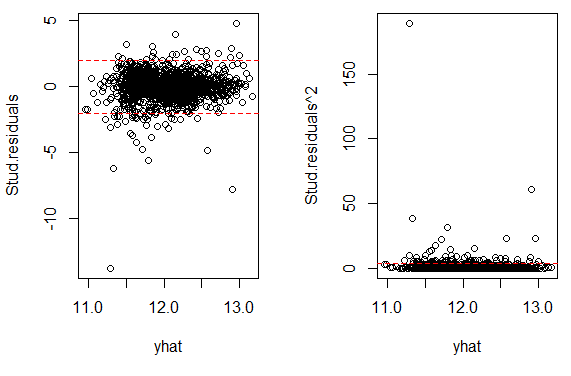
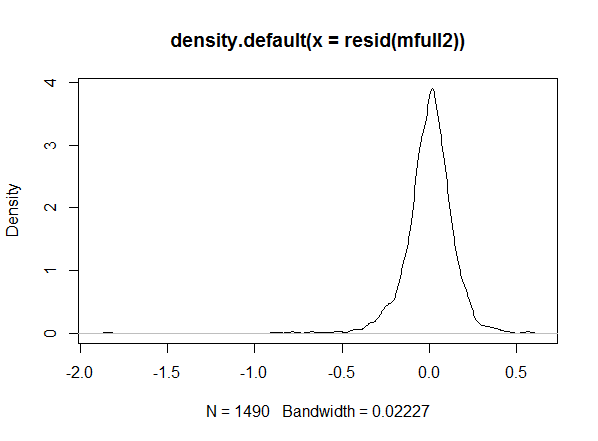
* + - 1. Linear relationship between the variables and the Sale Price

The first plot represents the residuals versus the fitted values. The red line is a smoothed curve that passes through the actual residuals (grey dashed line). The two lines should ultimately identify. Residuals in the residuals vs fitted graph of plot 45 seems to be symmetrically distributed around a diagonal line and do not form a curve. This plot is better than it was in m0.

* + - 1. Normality of residuals

This assumption will be checked with the Q-Q-Plot on figure 50 and 51 and with the goodness of fit test Kolmogorov-Smirnof test.

The qqplot diagram is still tailed (figure 49). We can investigate further this problem by examining the plot of the studentized residuals (figure 49) and the residual’s density plot (figure 48). Moreover the bottom right plot on figure 49 shows the standardized residuals against leverage. Residuals are still not normally distributed but this plot is also improved compared to the m0 model. Finally, the p-value of both Lilliefors and Saphiro-Wilk tests (figure 51) is less than .05 which means that the residuals do not come from a normally distributed population.

*Figure 49 M1 plot of studentised residuals Figure 48 m1: residuals density plot*

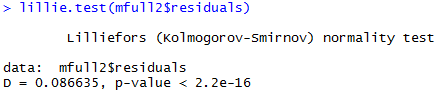


Figure 50 m1: Lilliefors test, testing normality of residuals

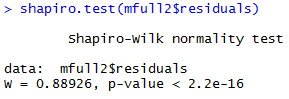
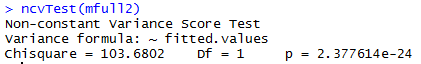


Figure 51 m1: Shaphiro-Wilk test, testing normality of residuals

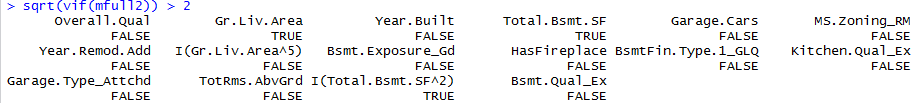
* + - 1. Homoscedasticity of residuals

The third plot (figure 47) represents the square root of the standardized residuals, which are residuals rescaled so that they have a mean of zero and a variance of one. This plot has large residuals (both positive and negative) plotting at the top and smaller residuals plotting at the bottom. Completely random points would mean absolute homoscedastic but in the above diagram this is not the case. We will also test this assumption with the No-constant variance score test (ncvTest) which is used for testing homoscedasticity. P-value is less than 0.05. Which means that we reject the null hypothesis that the variance of the residuals is constant and infer that heteroscedasticity is present. Homoscedasticity of residuals assumption is not met in m1.



* + - 1. Collinearity

Collinearity is a linear relationship between two explanatory variables and should not be present.



Concluding, non of the linear model’s assumptions are met in the m1 model.

## Prediction

### k-fold cross validation

We will use the most common variation of cross validation, the 10-fold cross-validation. The dataset was split in 10 equal parts (folds) of size 149. The first nine folds were used for training the model and the last one for testing the model. The diagram represented below represents the performance. The prediction of the 10-fold-cross-validation was really accurate and the mean square error (mse) really small, equal to 1.26e-30.

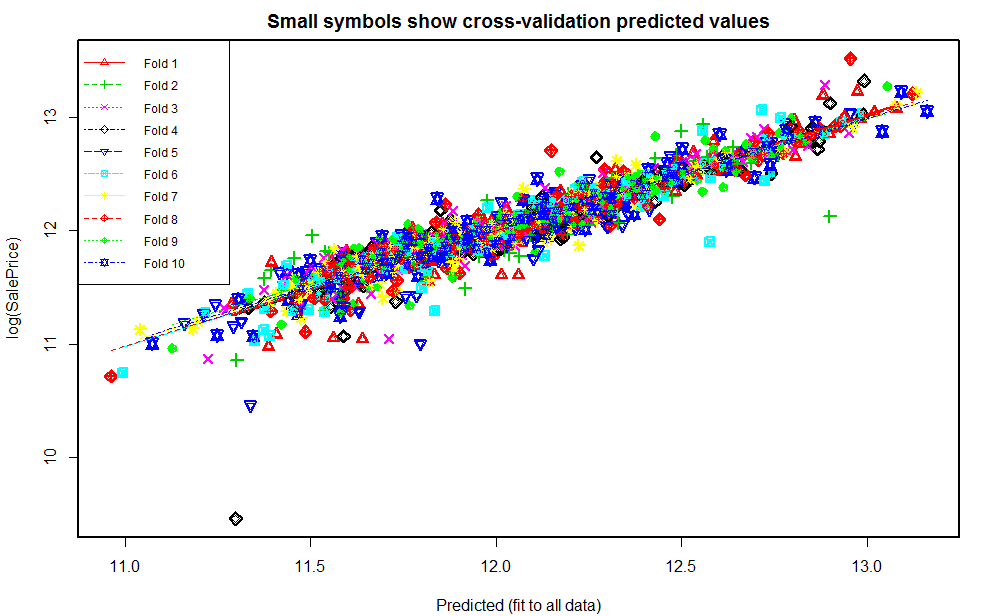
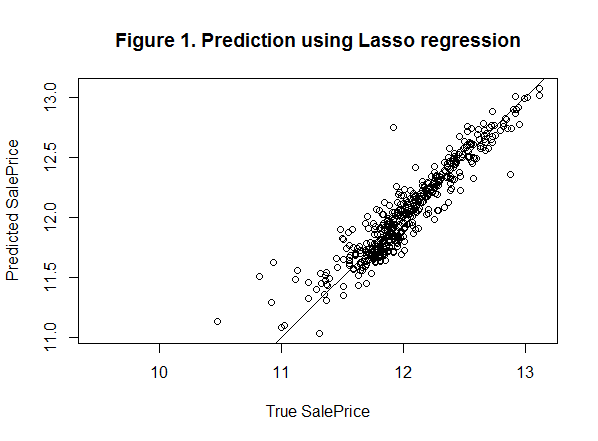


Figure 52 Leave one out 10 fold cross validation



### Test data prediction

We will now check how our model performs when the test dataset is used as a test set. The mse of model m1 is significantly lower and therefore the prediction is much better comparing to m0. The mse now equals to 0.06.



*Figure 53 M1 prediction on test set*



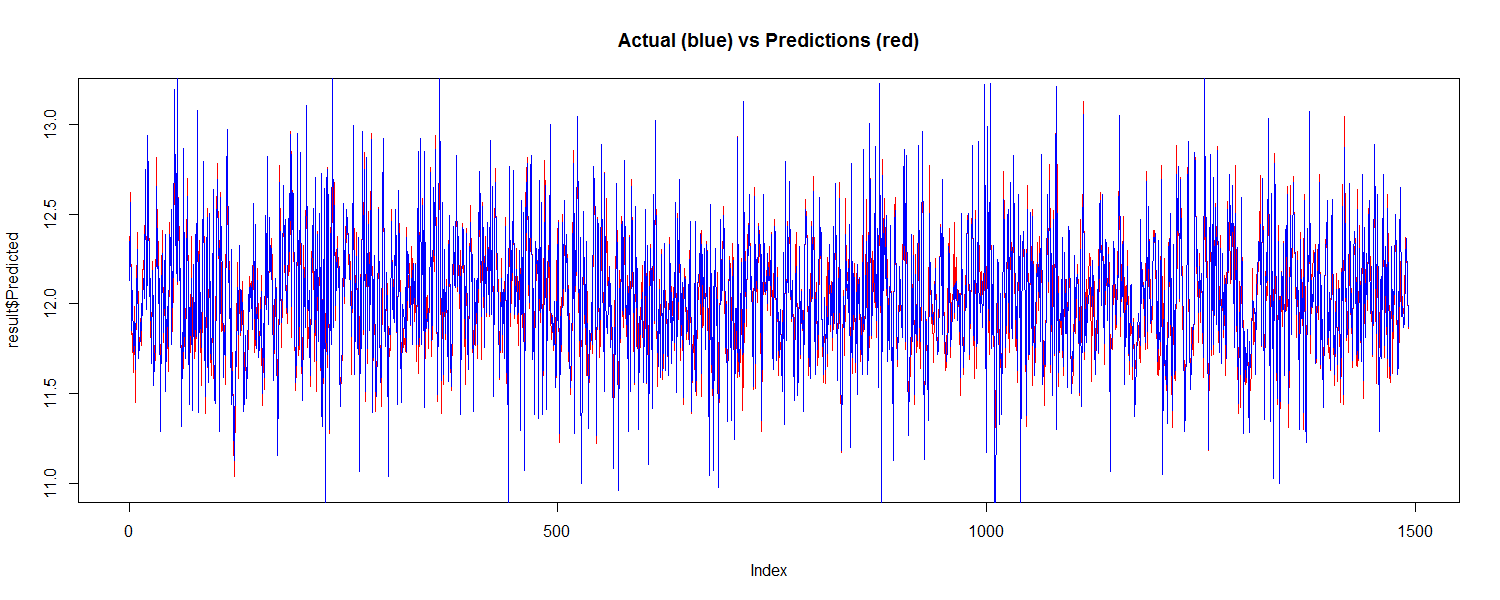


Figure 54 dataset 22 both as train and test set 10-fold

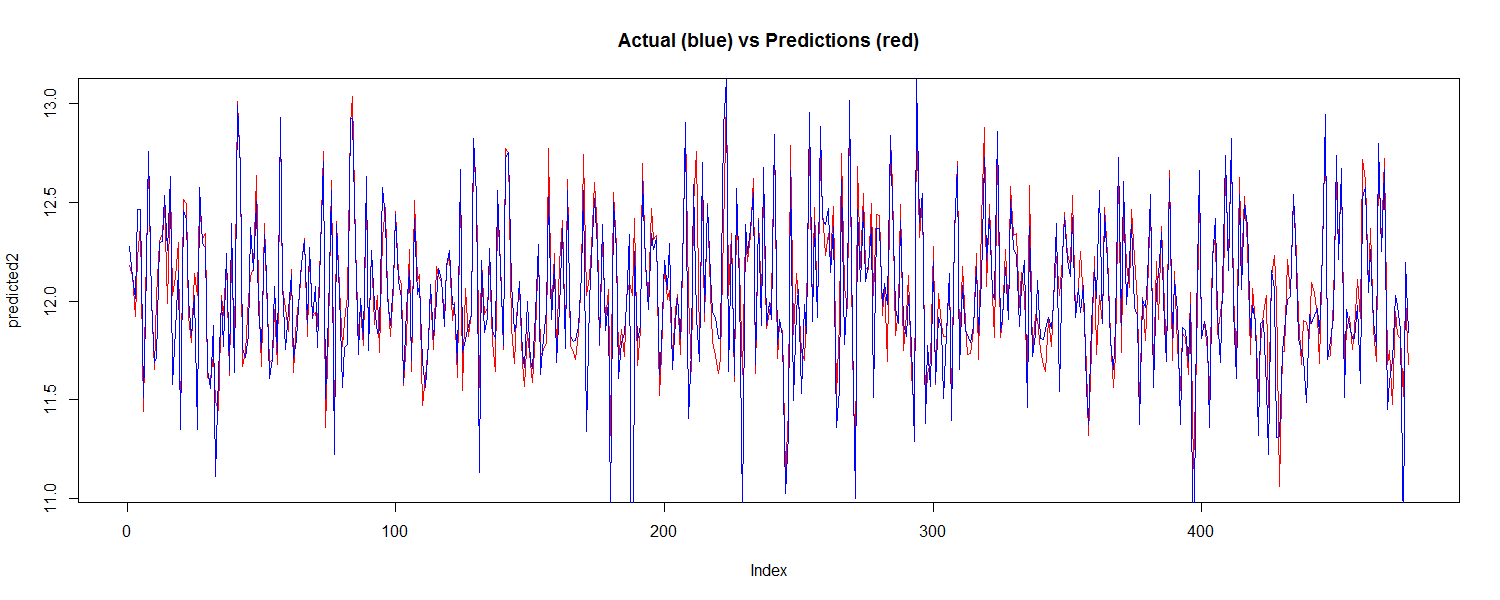


Figure 55 dataset 22 as train and ames\_iowa\_housing\_test as test set

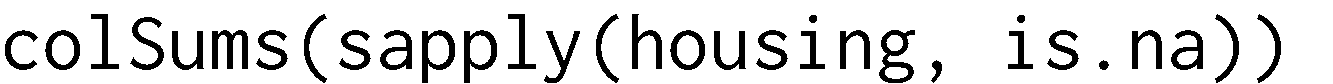
# Conclusions and future improvements

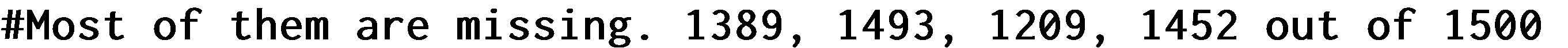
It was not possible to meet the linear models assumptions. This was probably caused by the data cleaning and transformation. The prediction of null attributes in the beginning was successful and didn’t correspond to the test dataset and this lead to a miscalculated model. Moreover, categorical values were manually transformed to dummies which also lead to errors, because this is what “strings as factors” does when passed to function requiring numeric data. Re calculating the null and dummy values will probably lead to a model which meets the linear model’s assumptions.

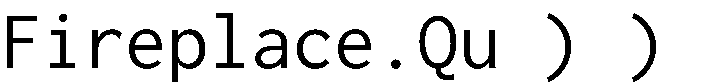
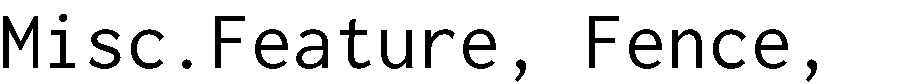
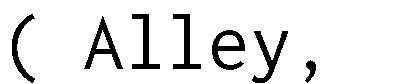
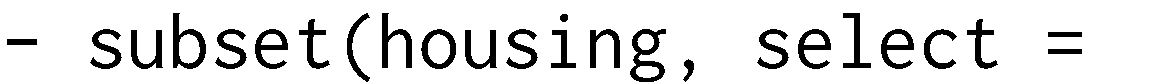
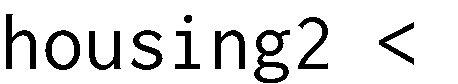
However m1 perform much better comparing to m0 and has significantly lower standard errors.

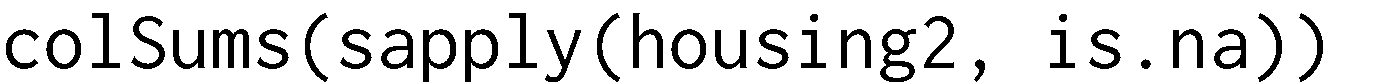
# Παράρτημα κώδικα

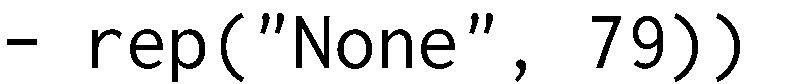
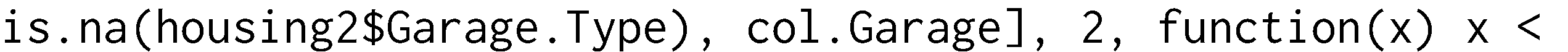
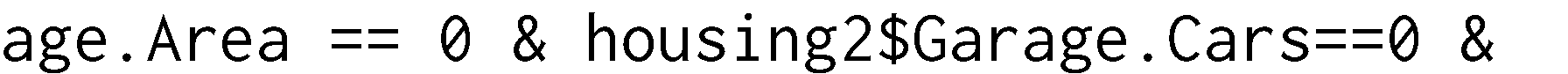
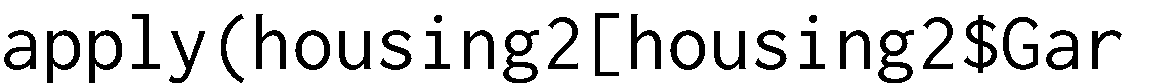
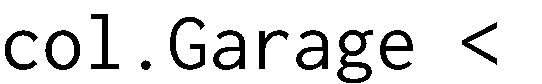
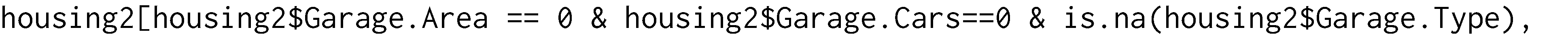
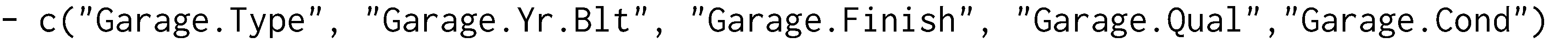
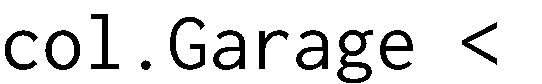
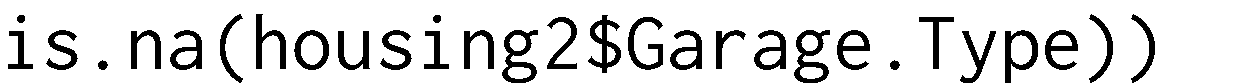
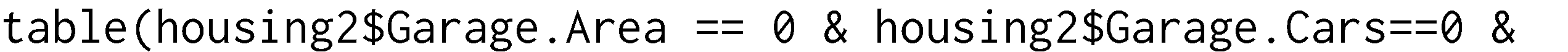
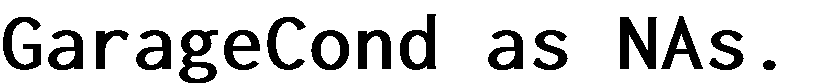
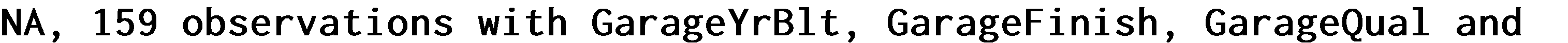
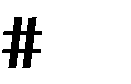
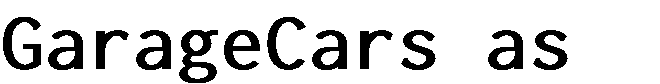
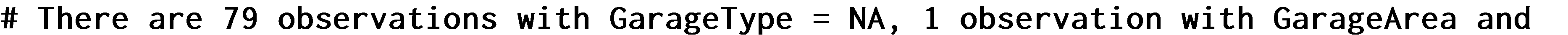
## Data cleaning & null replacement

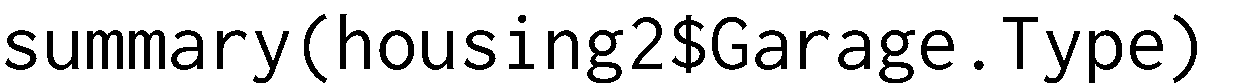


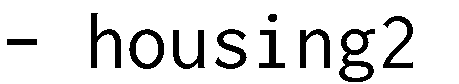
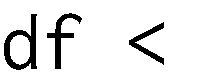
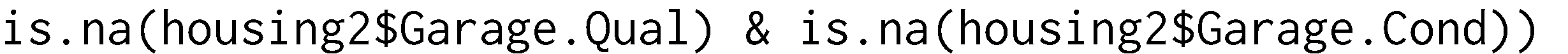
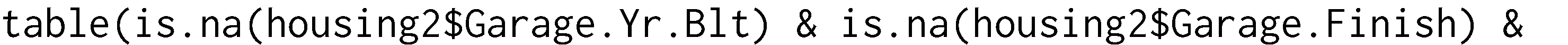
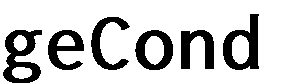
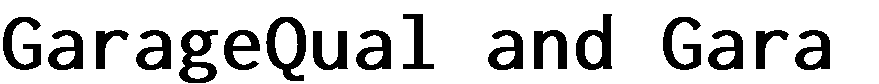
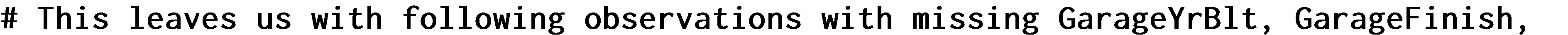


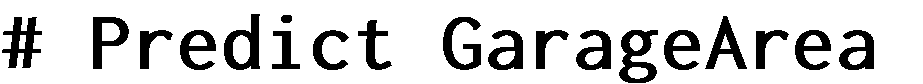


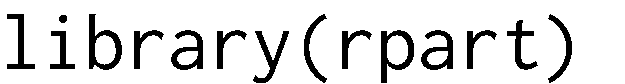


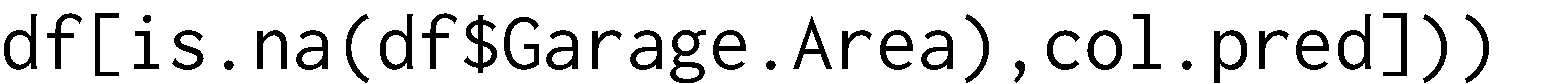
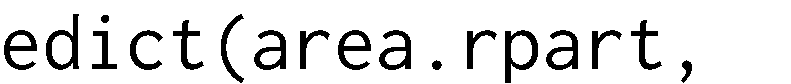
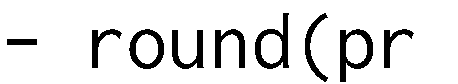
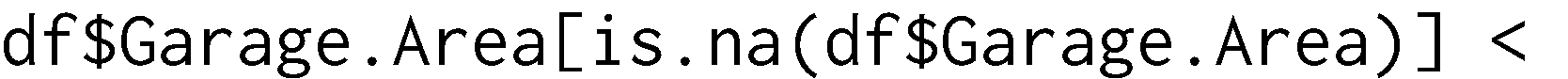
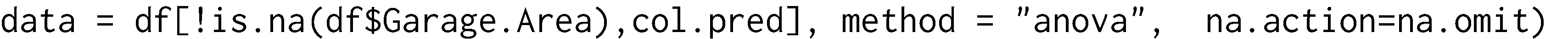
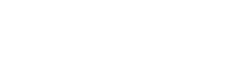
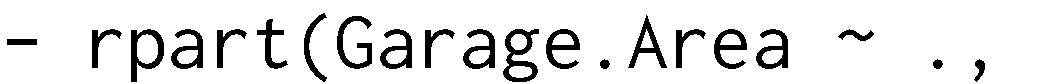
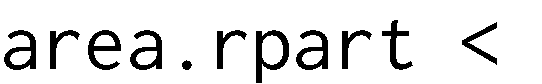
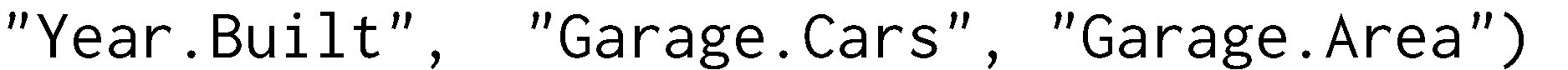
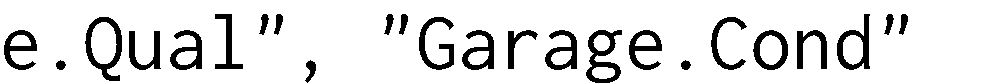
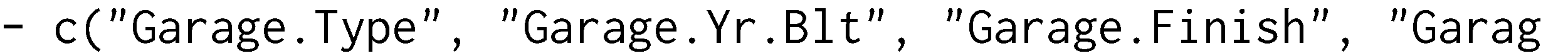
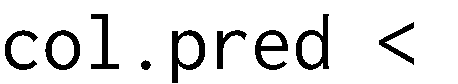


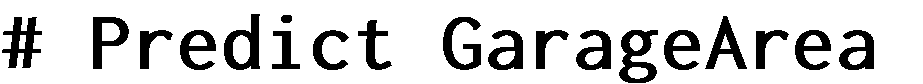


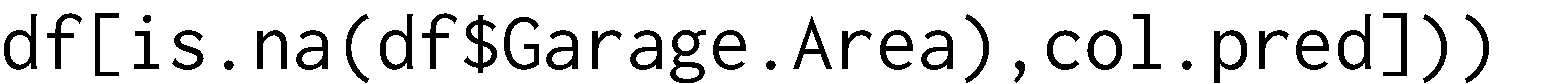
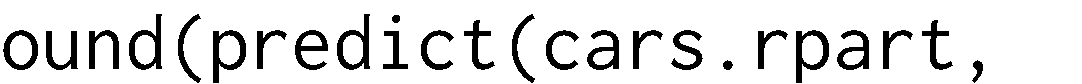
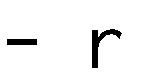
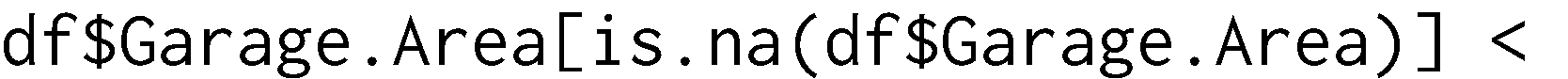
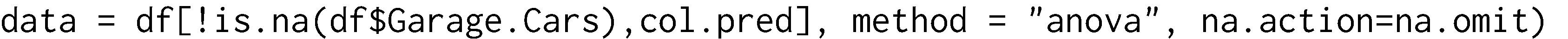
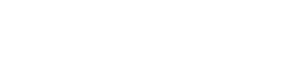
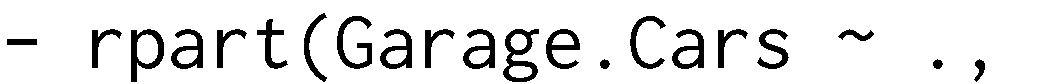
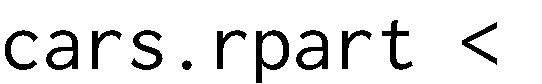


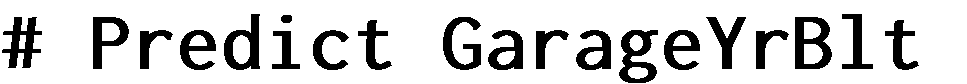


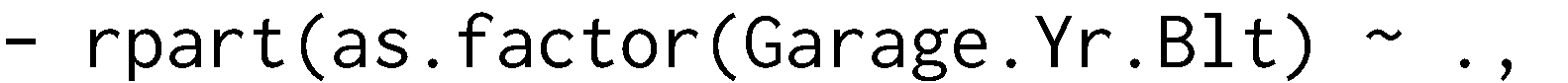
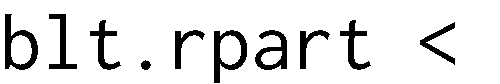


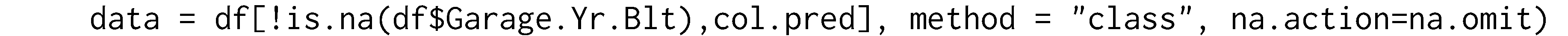


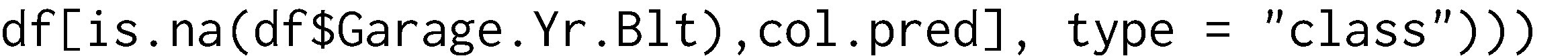
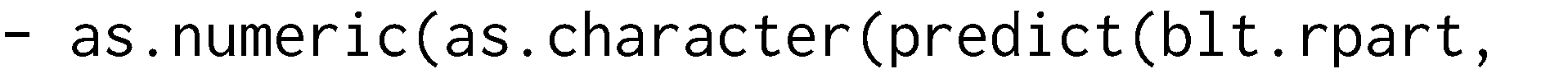
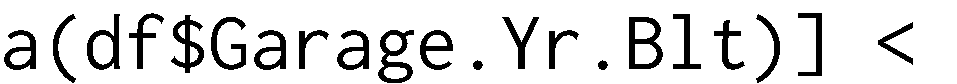
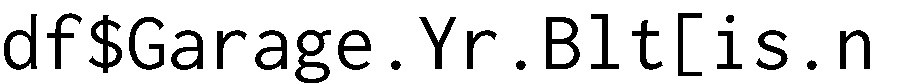


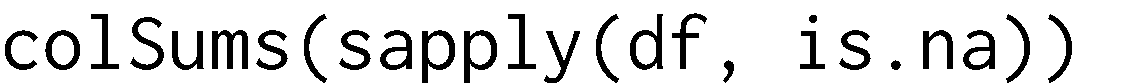


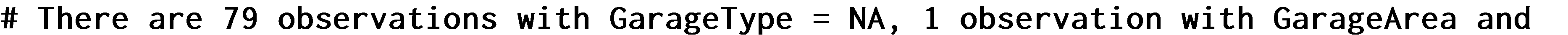


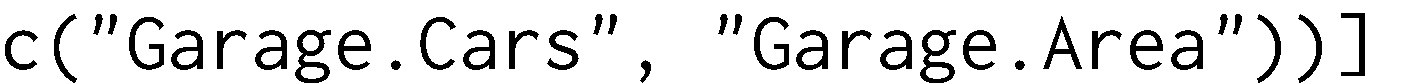
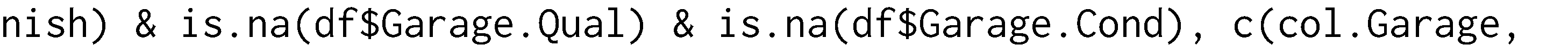
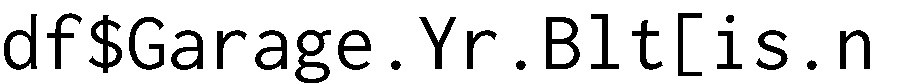


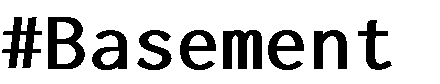


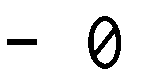
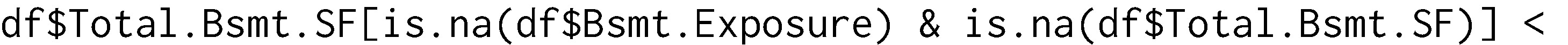
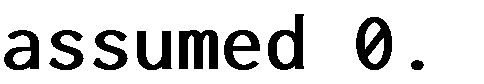
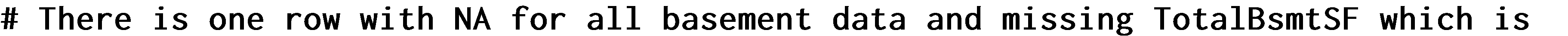
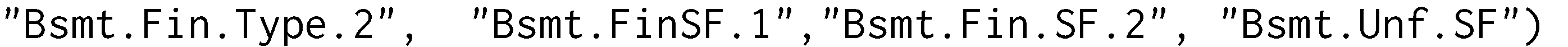
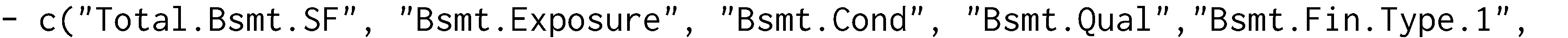
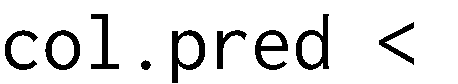


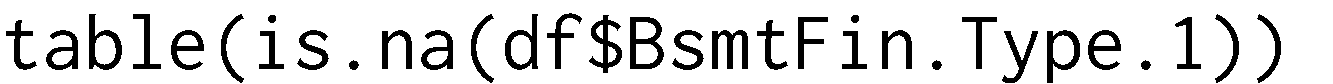


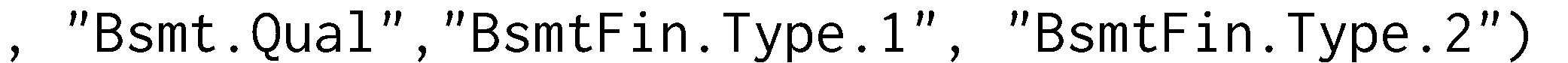
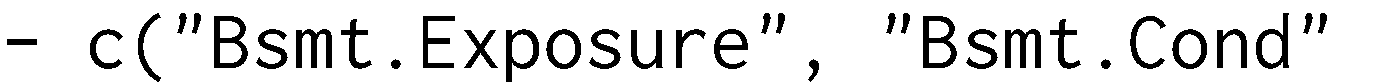
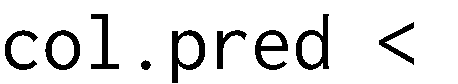


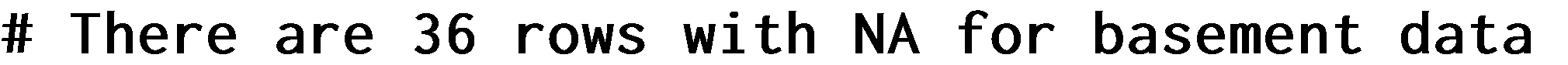


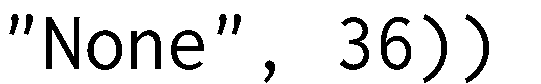
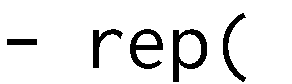
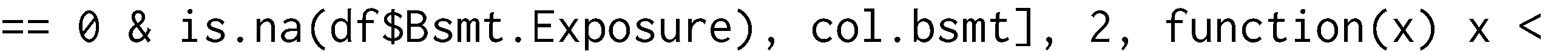
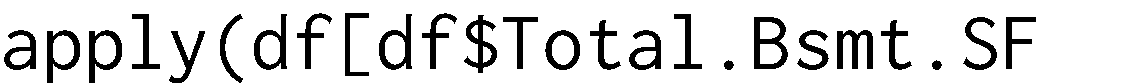
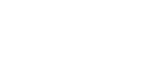
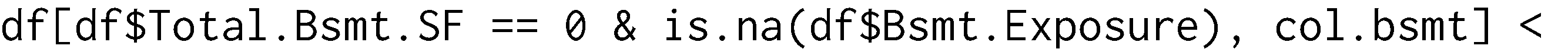


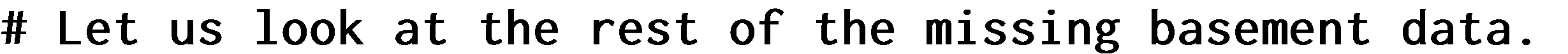


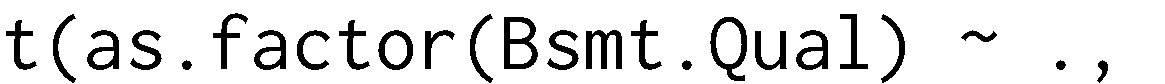
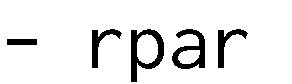
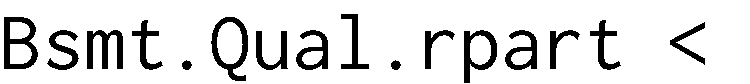
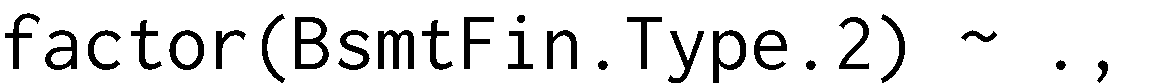
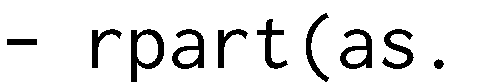
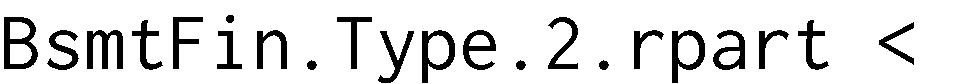
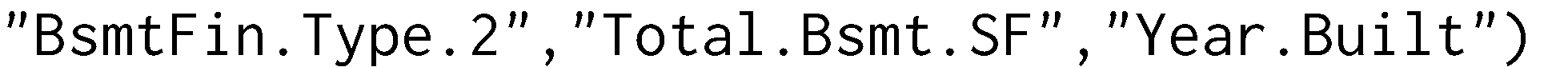
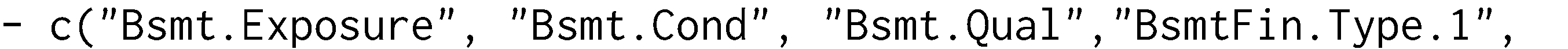
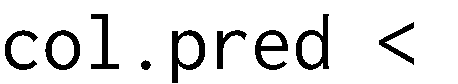
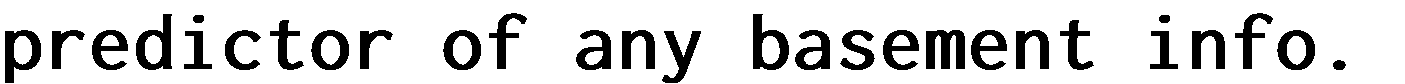
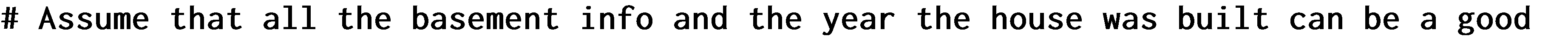
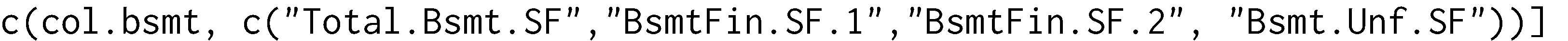
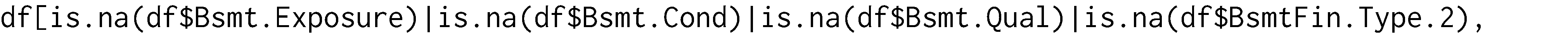
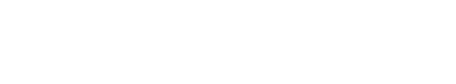
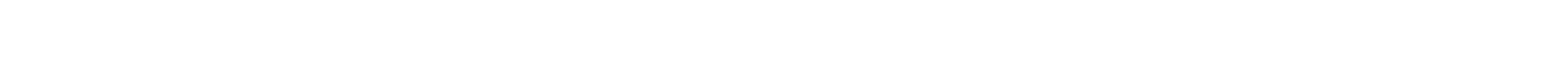
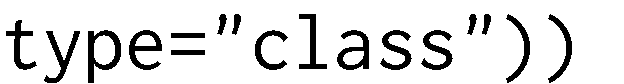
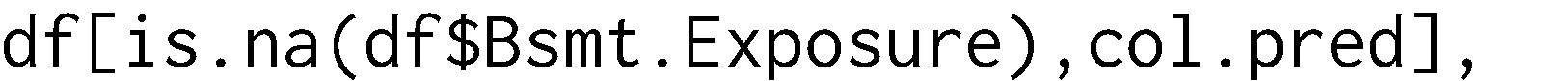
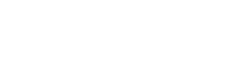
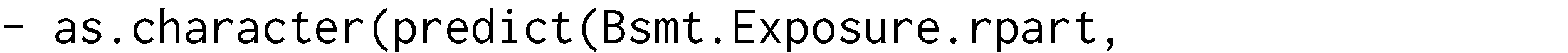
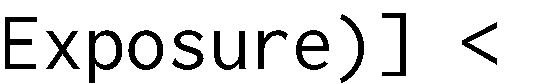
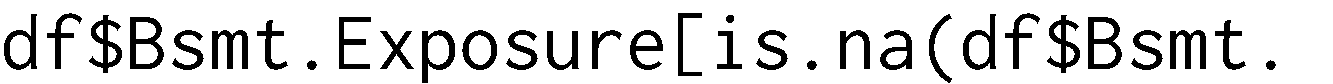
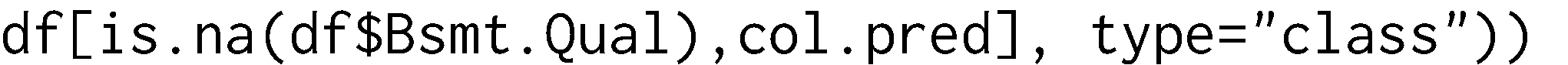
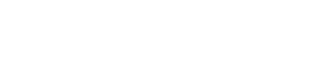
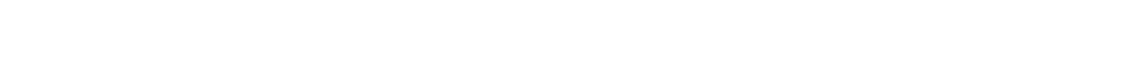
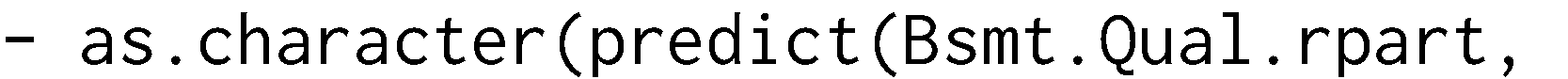
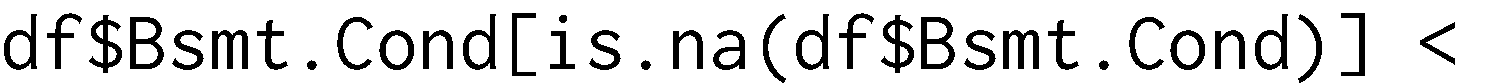
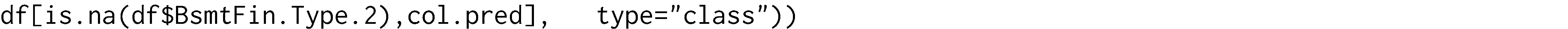
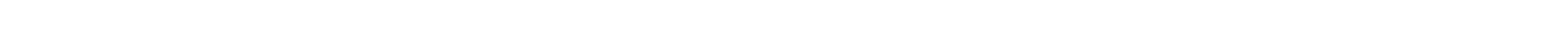
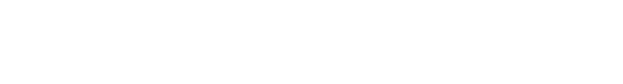
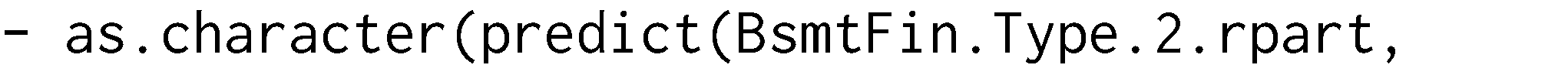
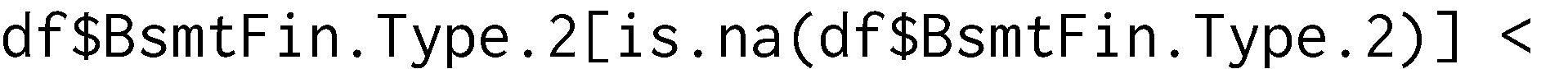
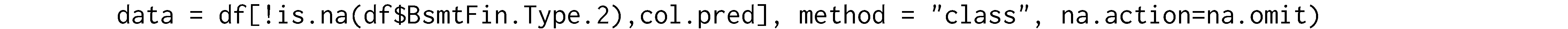


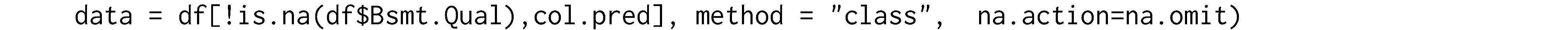


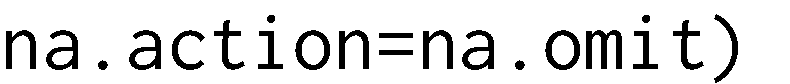
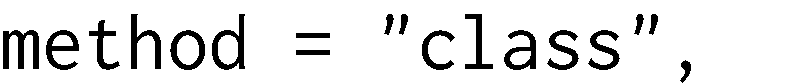
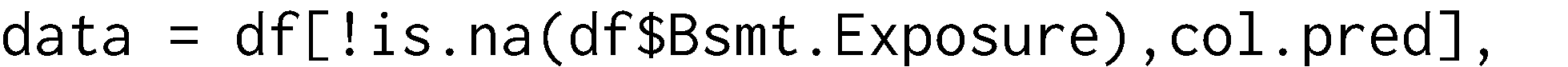
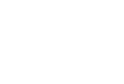
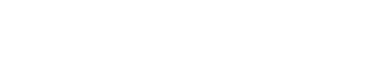
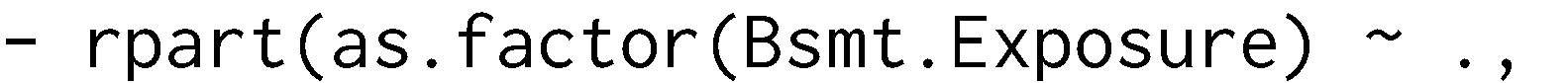
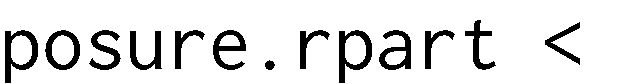
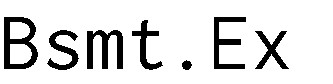
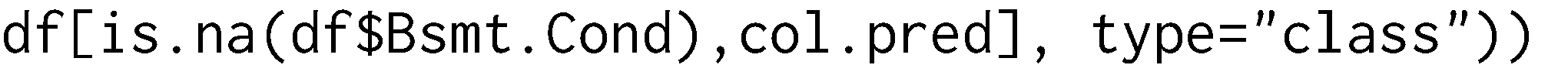
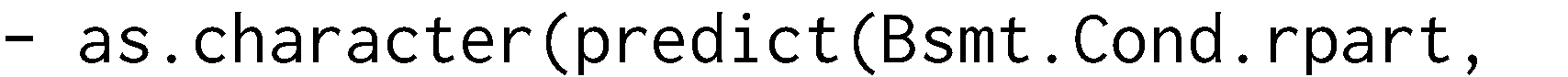
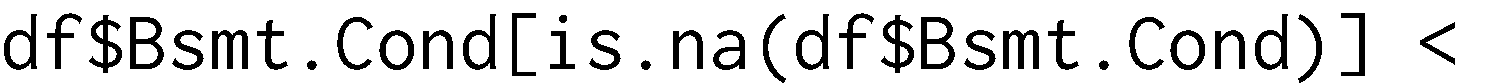
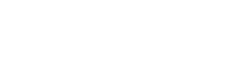
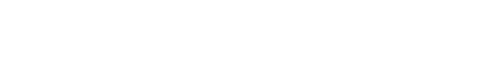
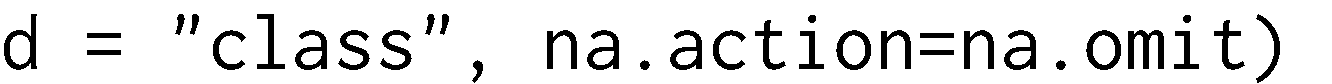
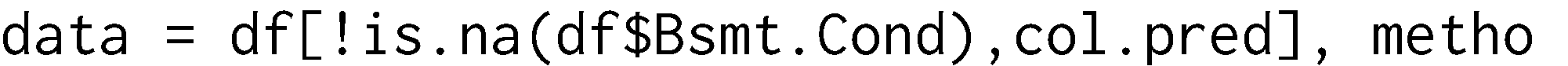
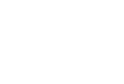
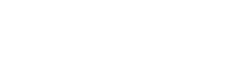
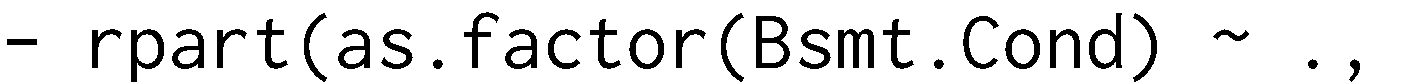
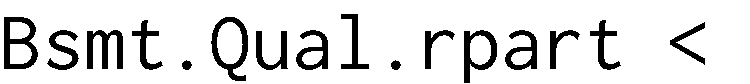


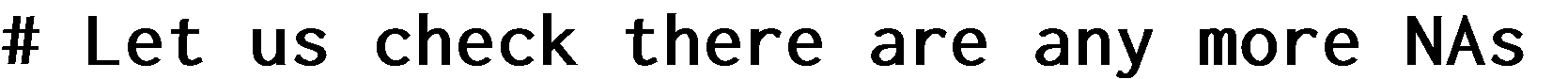


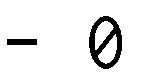
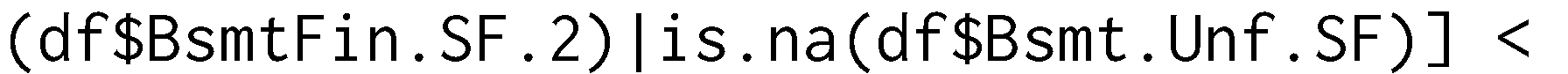
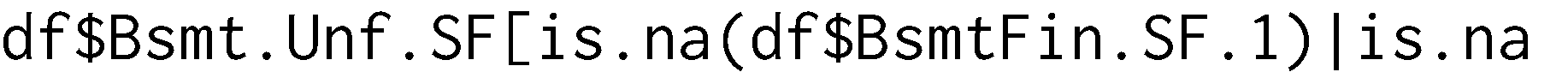
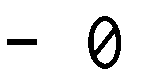
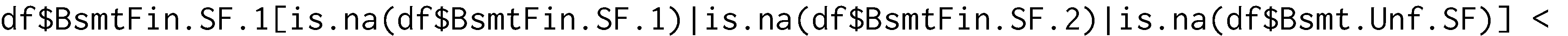
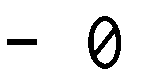
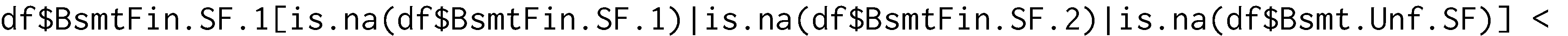
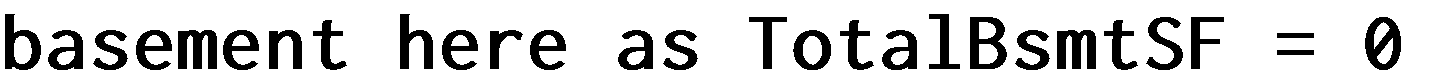
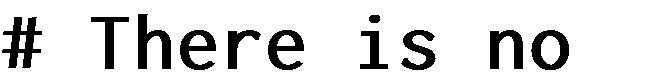
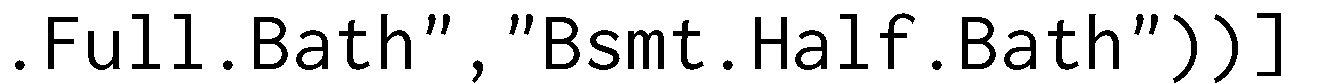
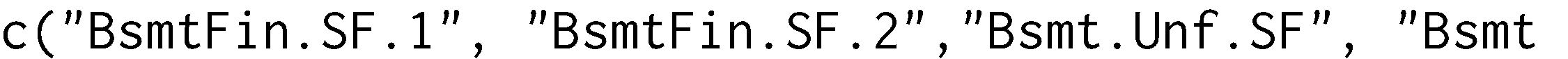
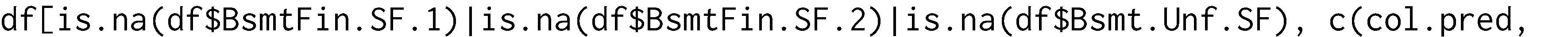


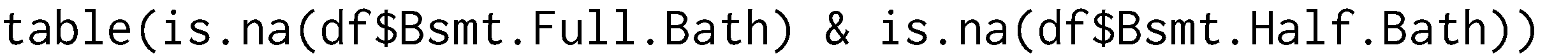


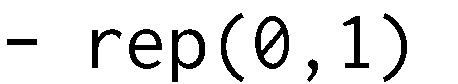
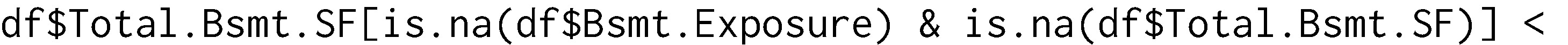
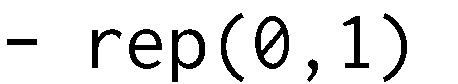
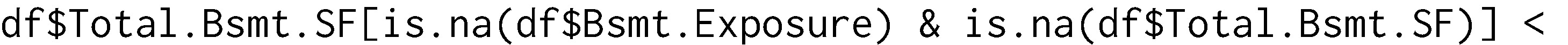


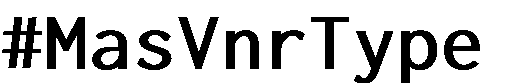


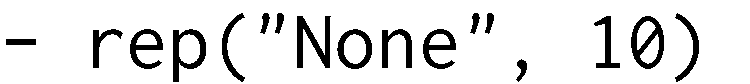
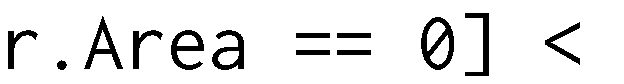
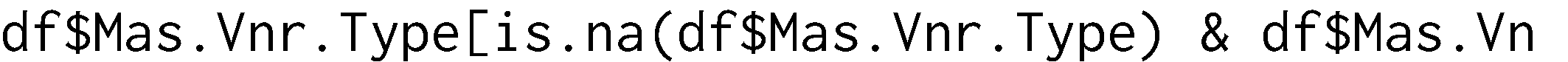
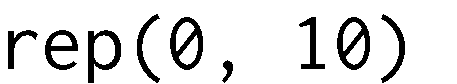
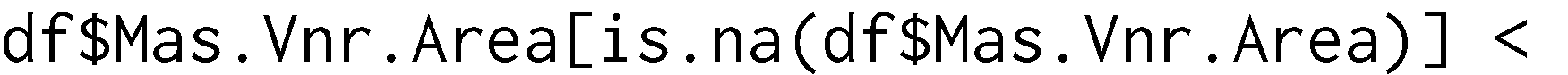
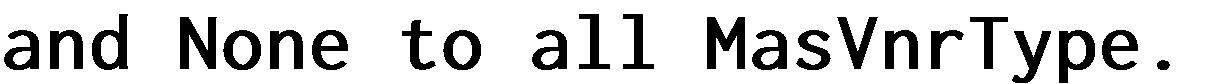
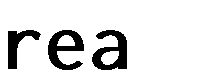
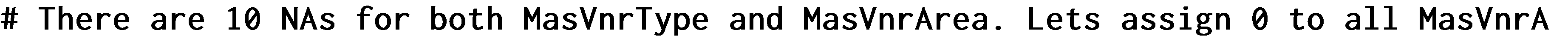
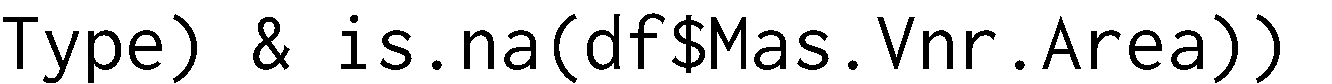
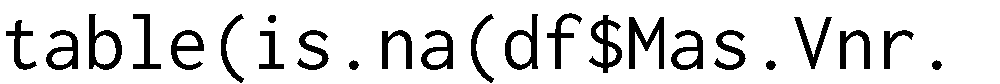
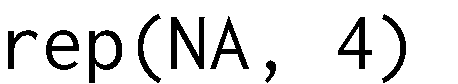
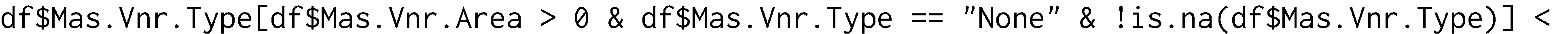
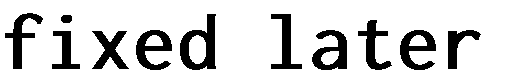
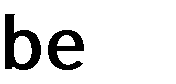
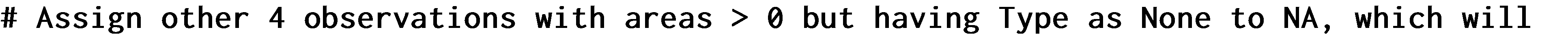
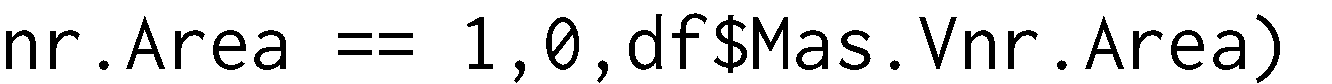
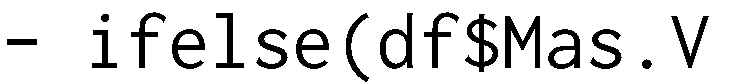
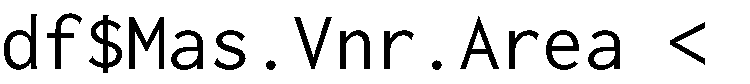
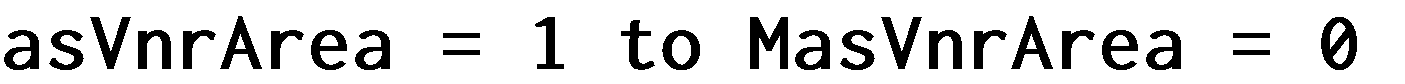
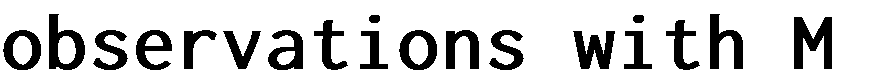
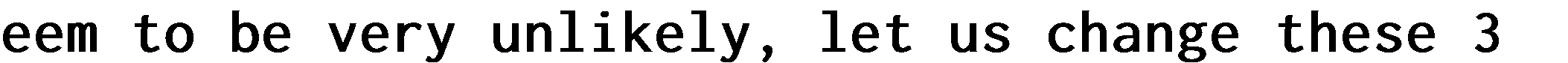
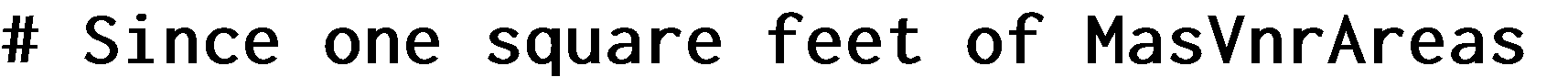
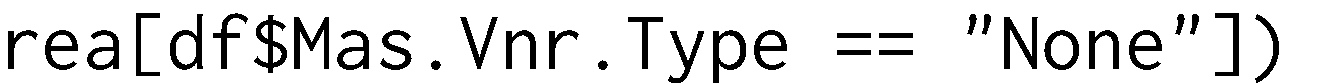
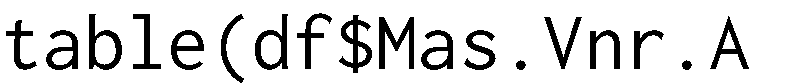
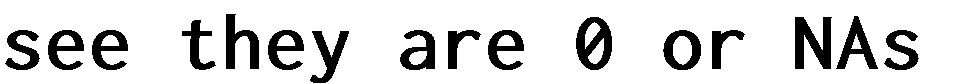
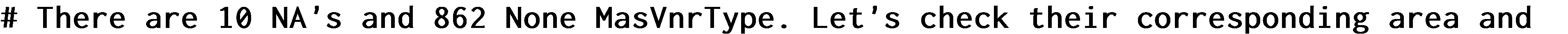
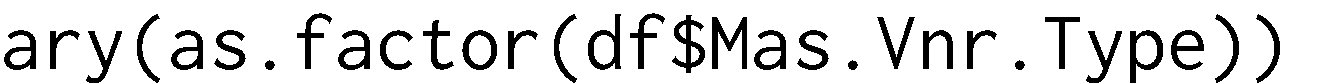
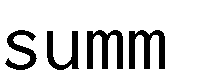


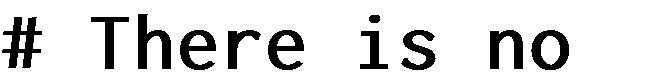


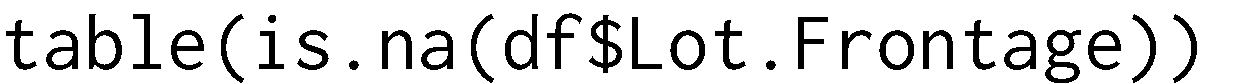


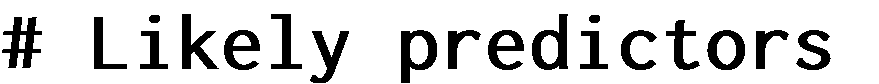


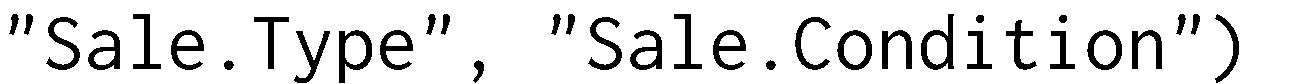
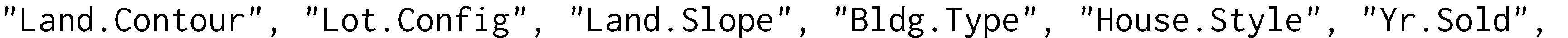
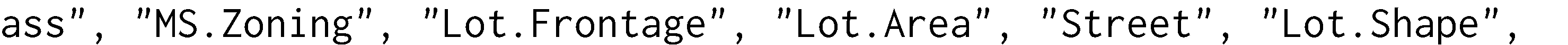
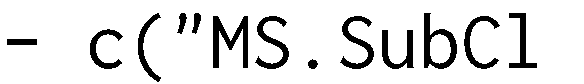
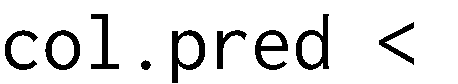


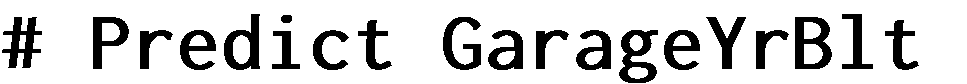


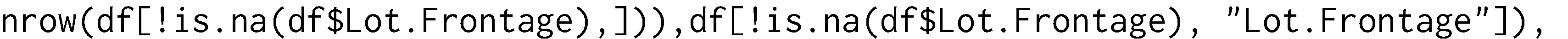
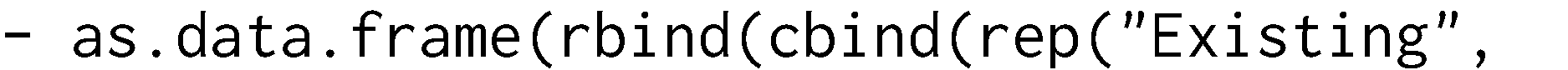
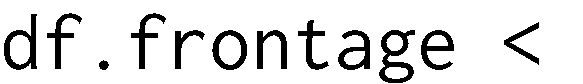
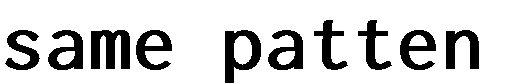
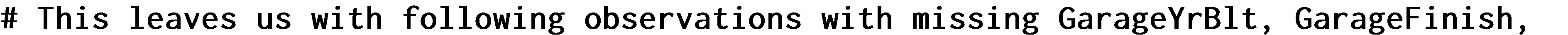
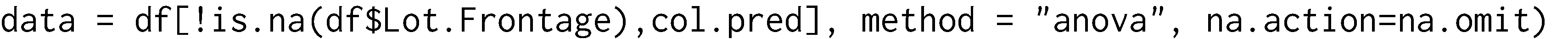
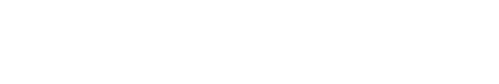
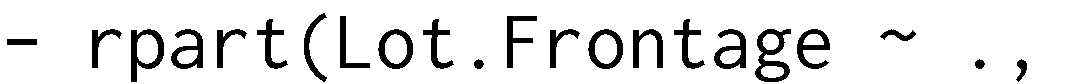
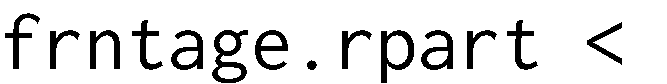


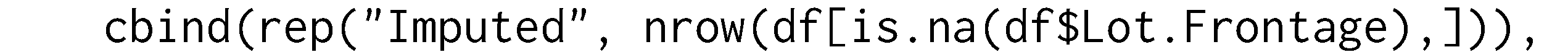


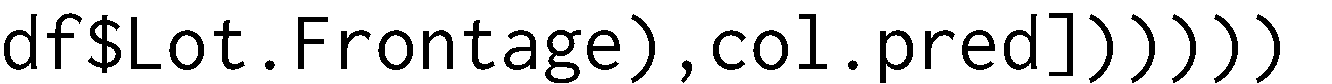
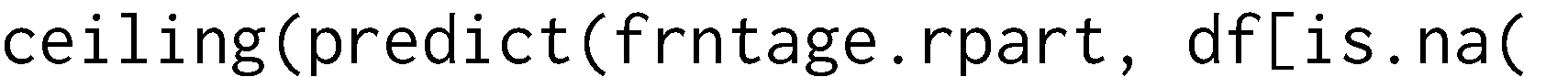
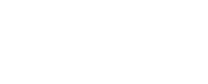


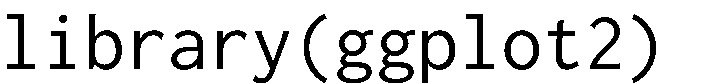


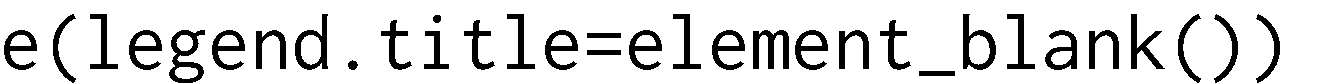
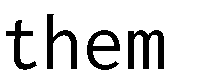
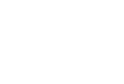
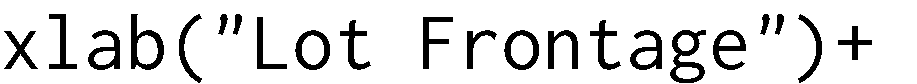
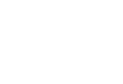
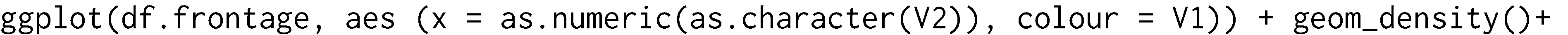


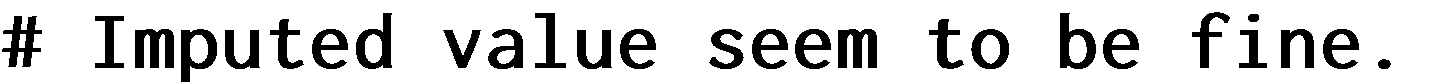


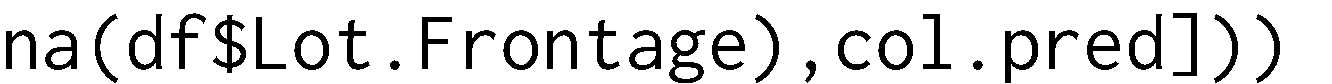
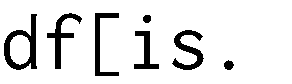
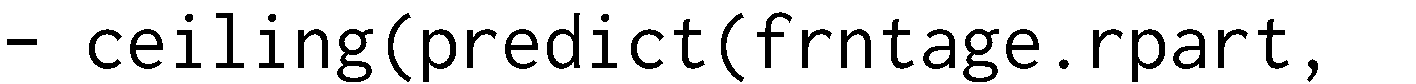
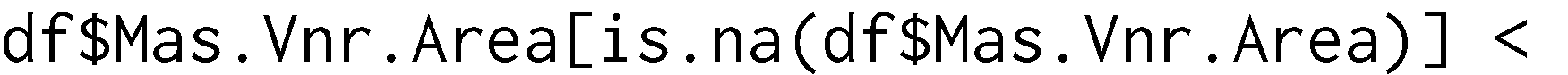


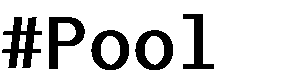


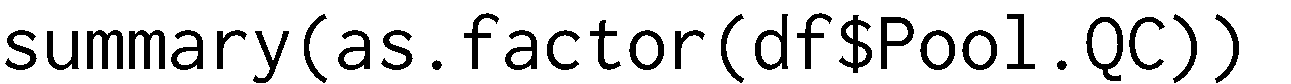


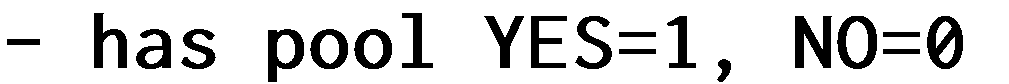
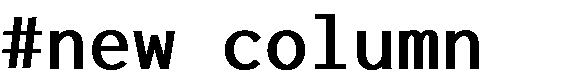
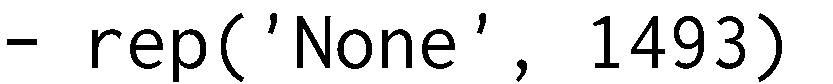
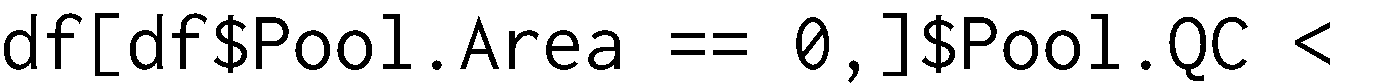
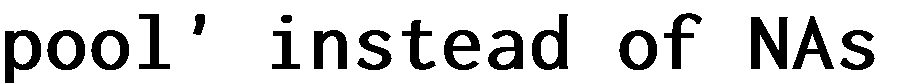
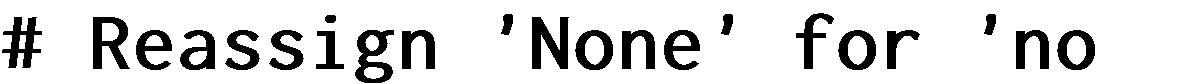
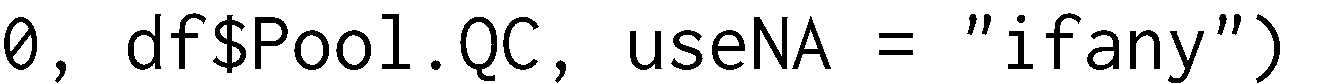
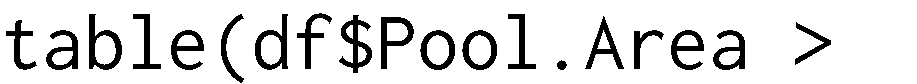




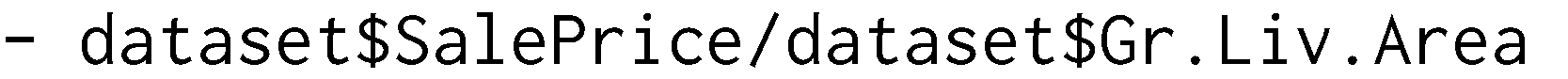
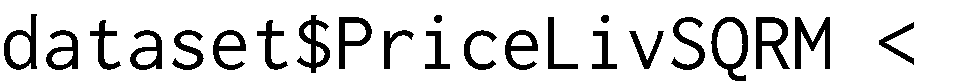
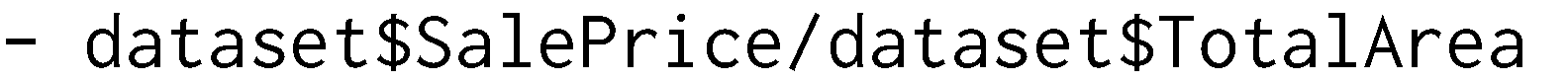
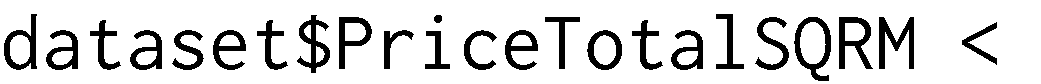
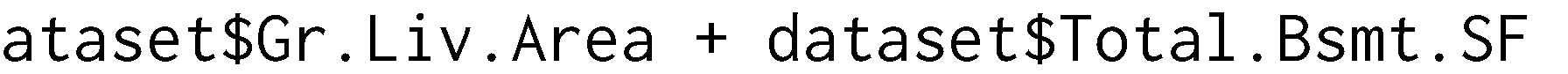
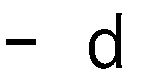
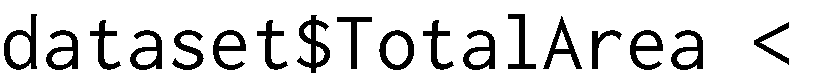


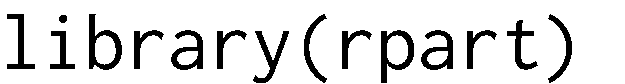


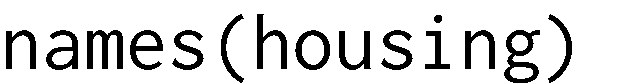


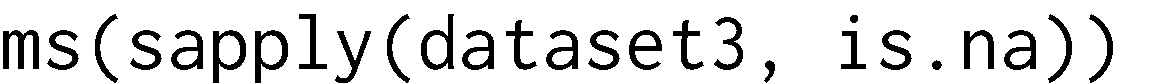
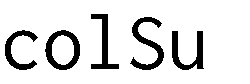
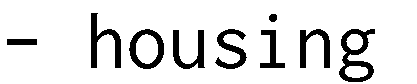
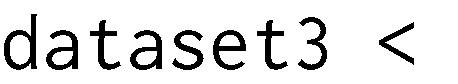


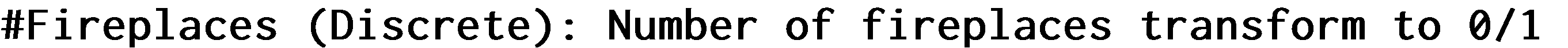
## Data transformation & new columns

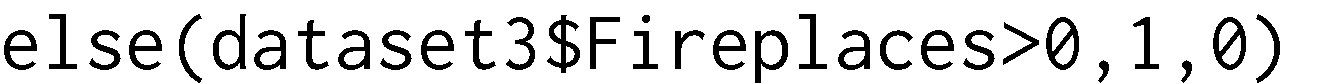
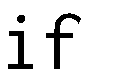
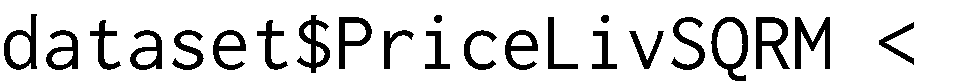


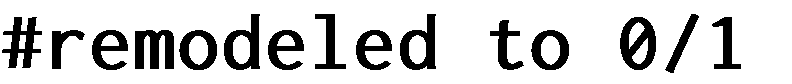


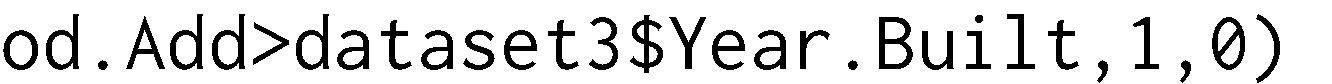
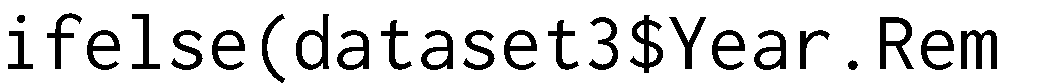
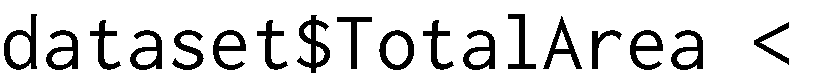


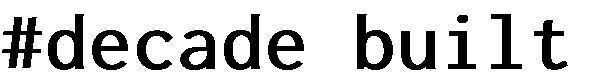


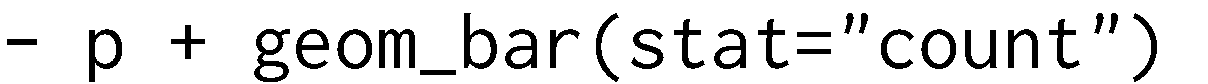
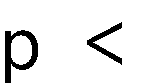
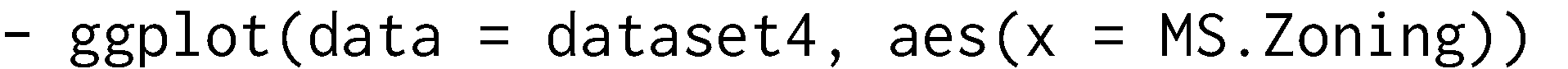
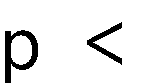
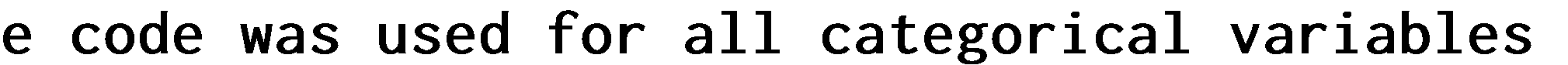
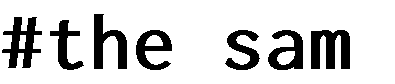
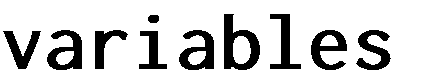
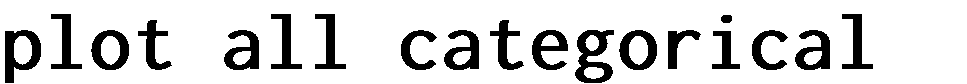
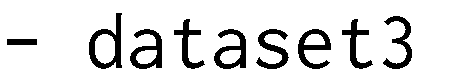
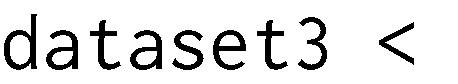
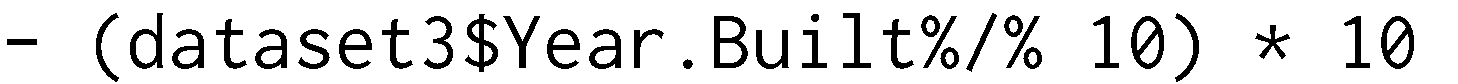
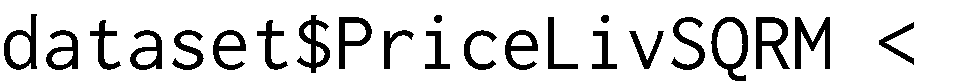




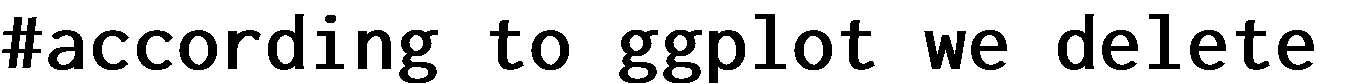


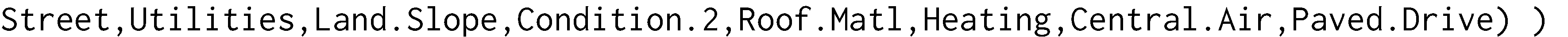
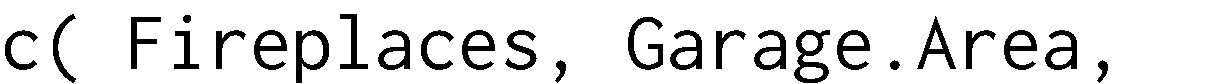
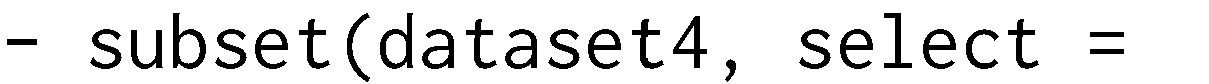
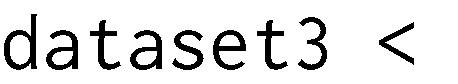




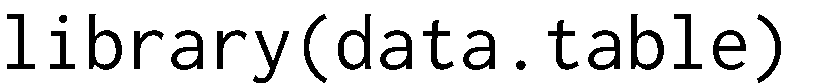


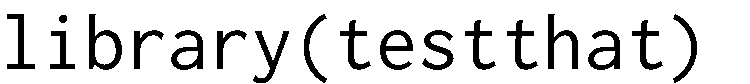


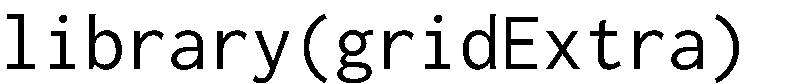


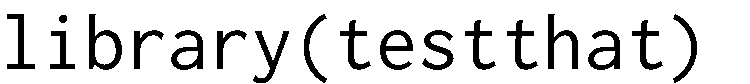


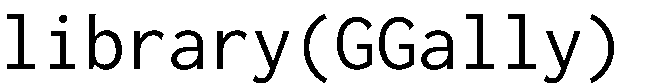
## Visualize variables

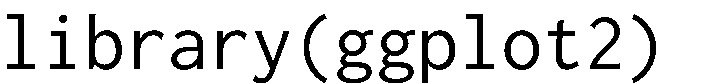


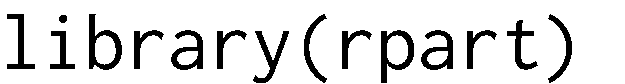


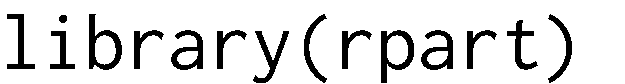


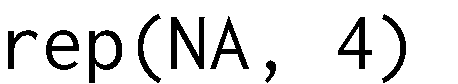
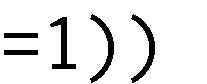
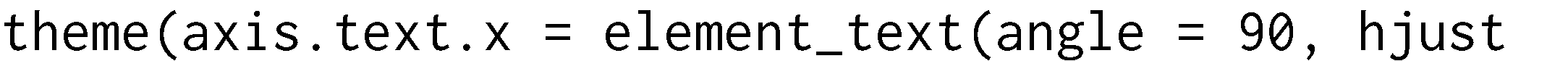
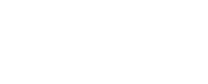
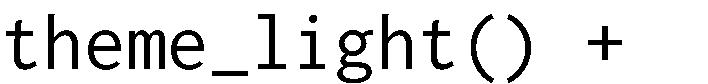
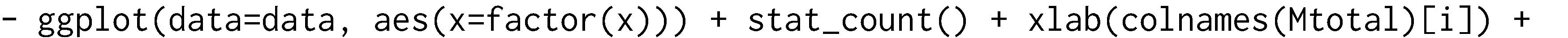
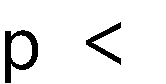
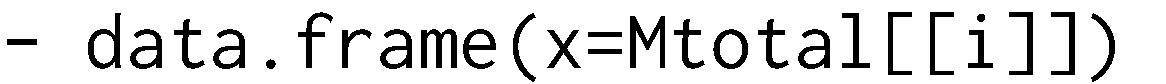
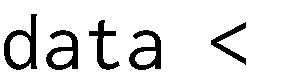
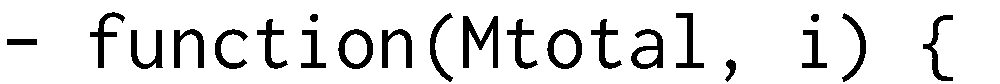
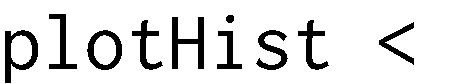
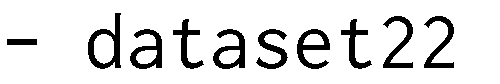
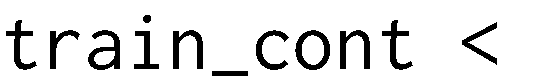
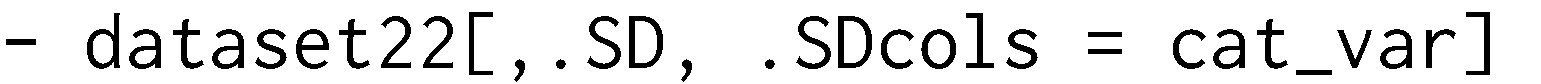
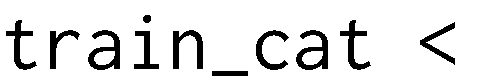




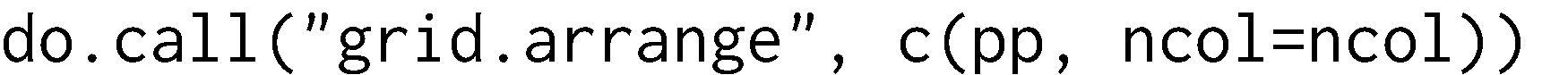
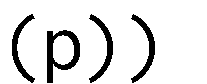
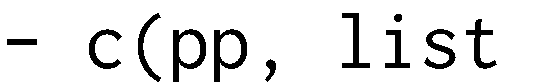
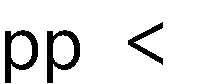
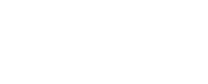
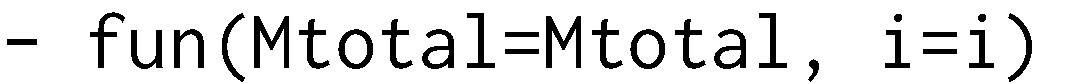
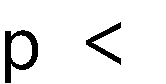
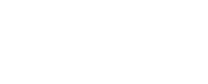
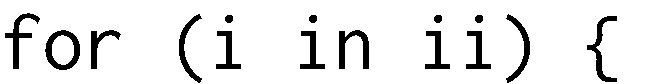
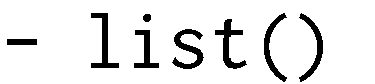
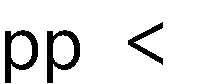
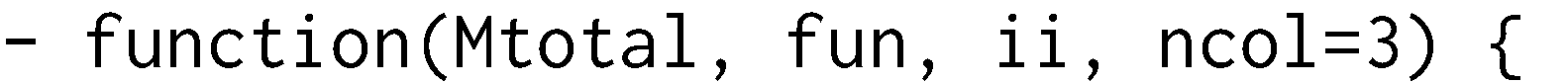
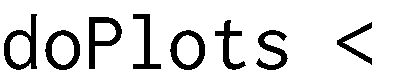
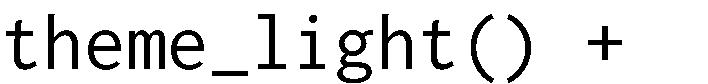
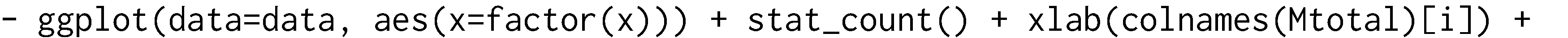
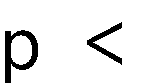




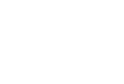
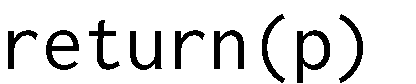
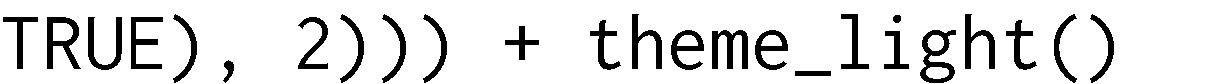
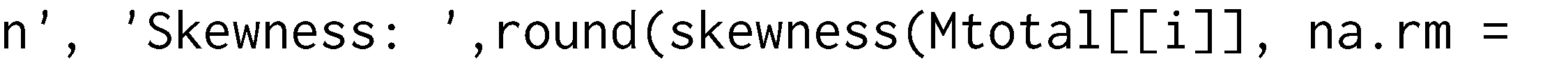
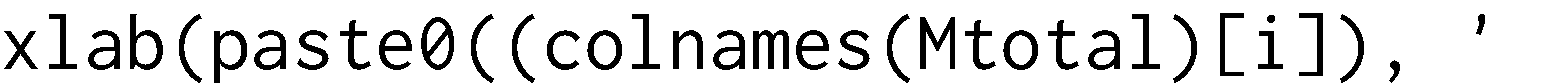
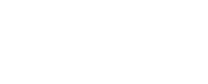
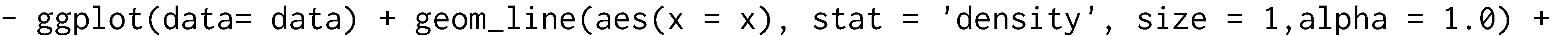
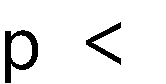
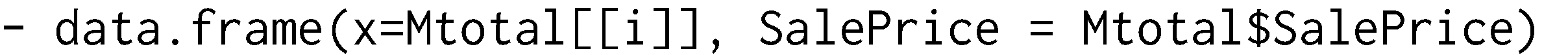
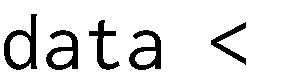
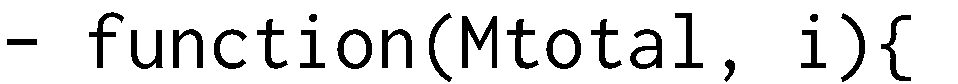
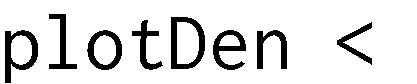




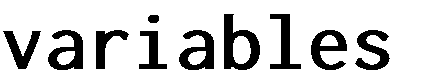
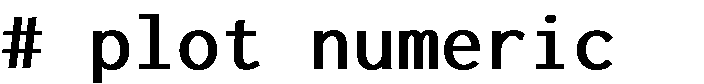


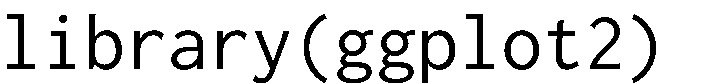


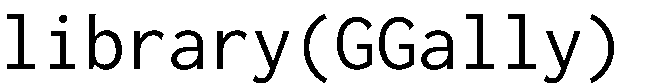


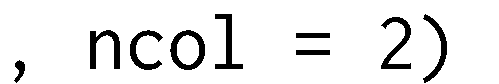
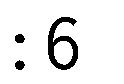
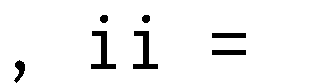
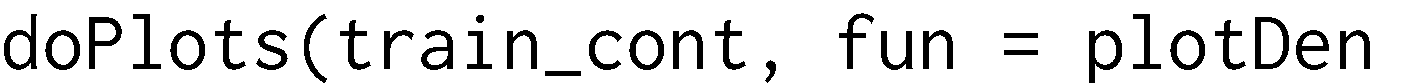


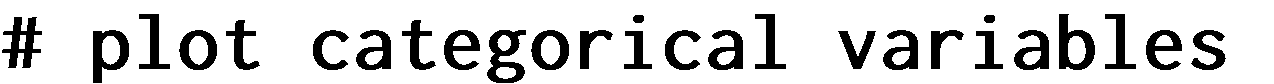


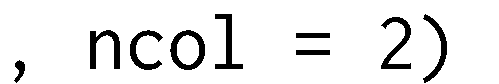
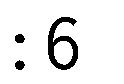
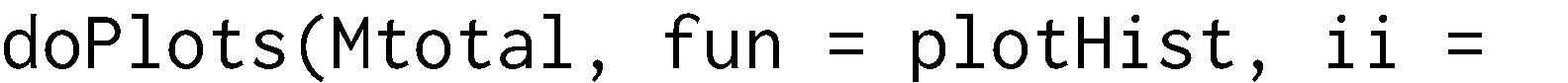




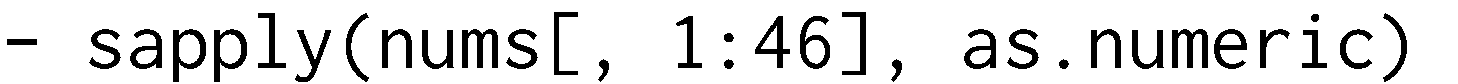
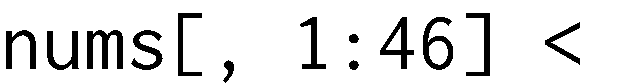
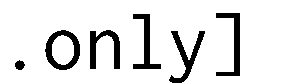
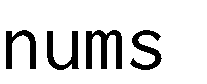
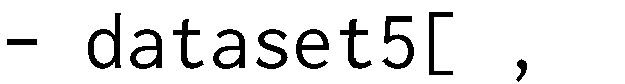
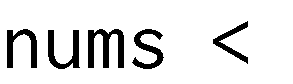
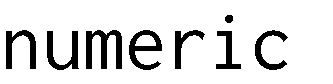
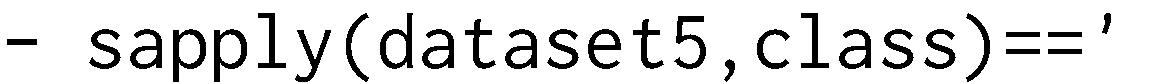
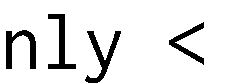
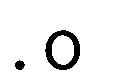
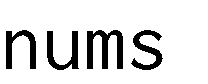
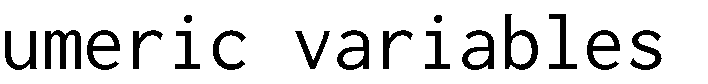
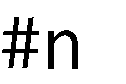
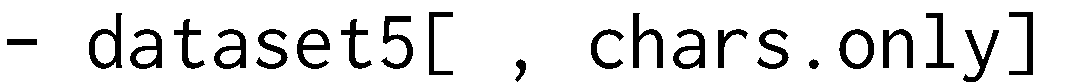
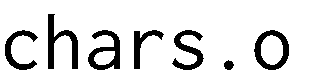
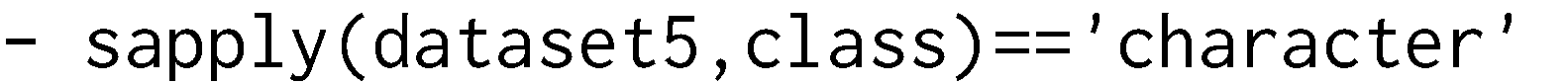
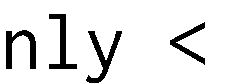
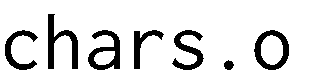
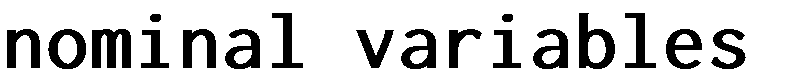




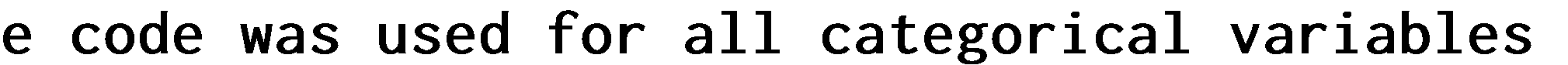
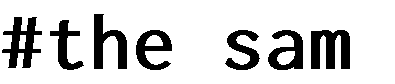
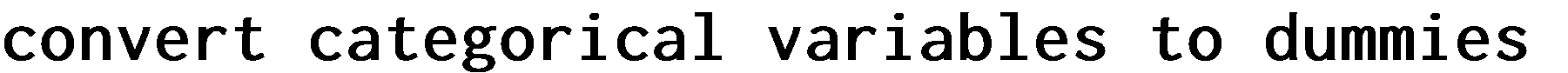


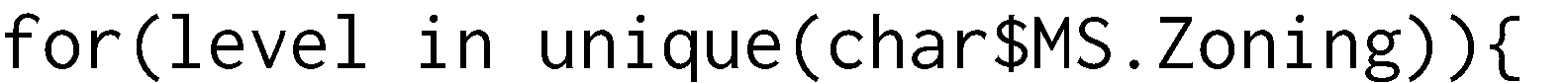


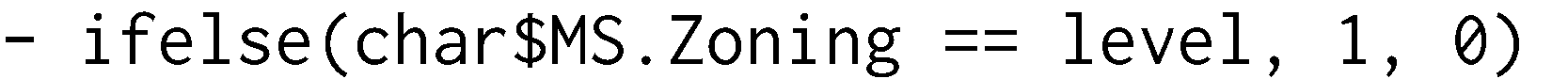
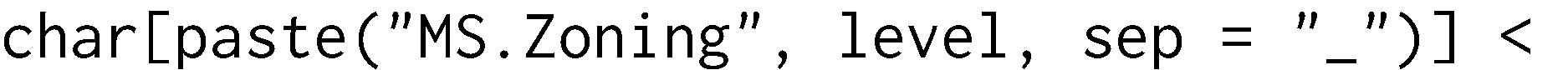
## Split numeric and nominal variables into two different data frames



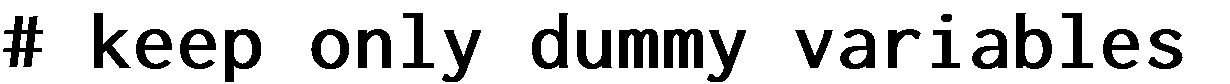
## Categorical variables to dummy

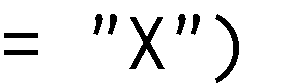
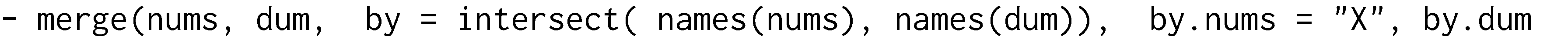
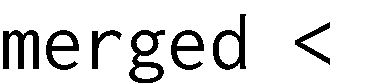
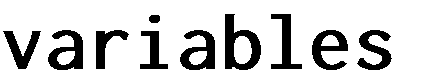
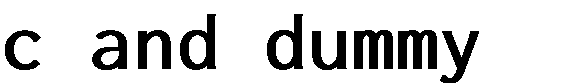
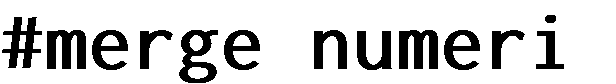
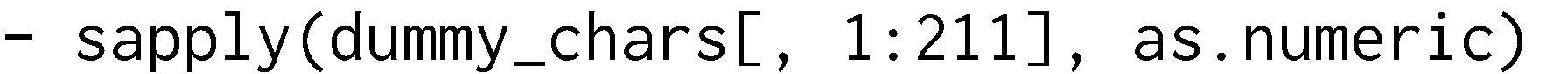
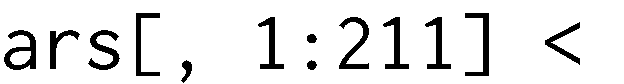
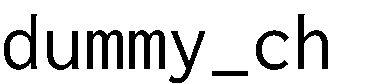
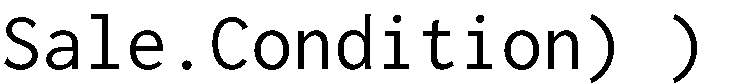
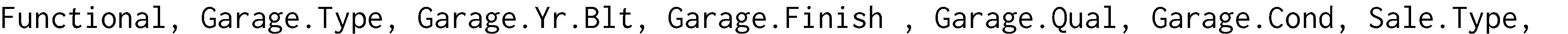
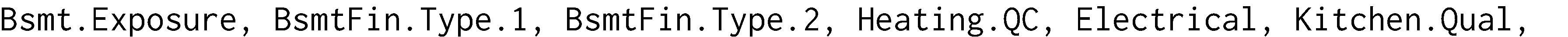
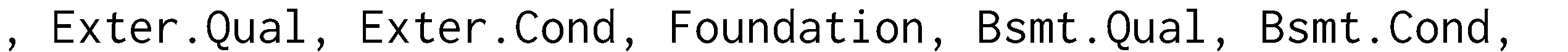
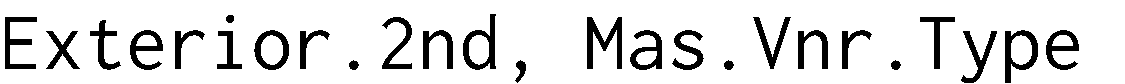
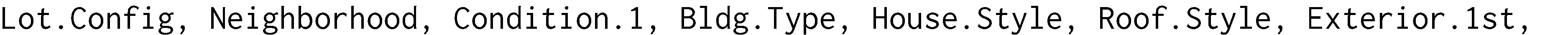
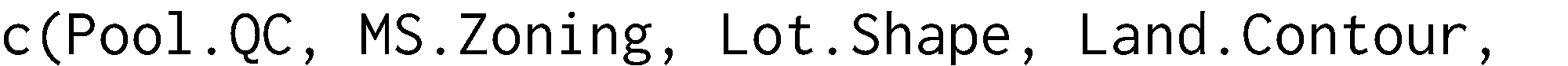
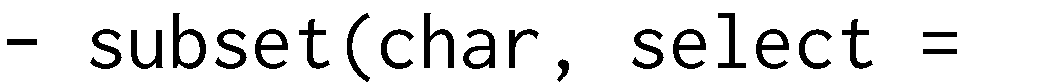
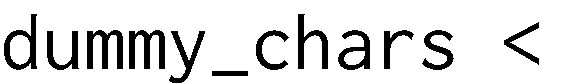




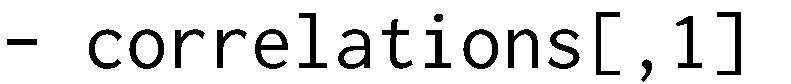
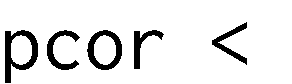
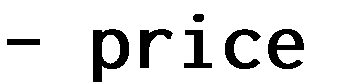
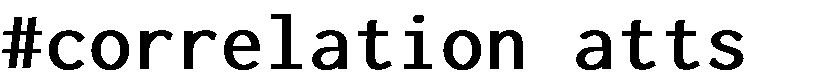
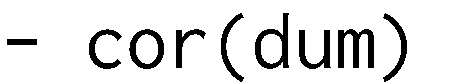
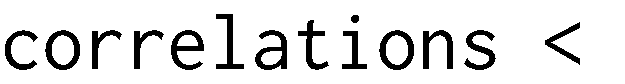
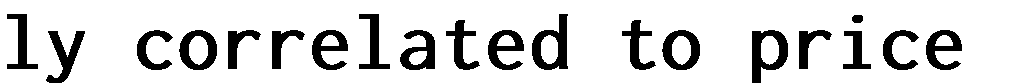
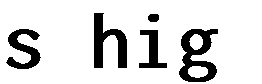
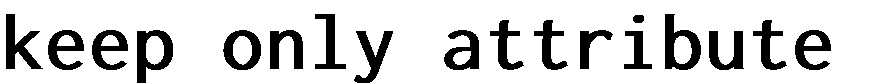


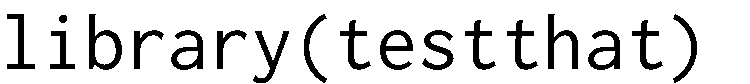


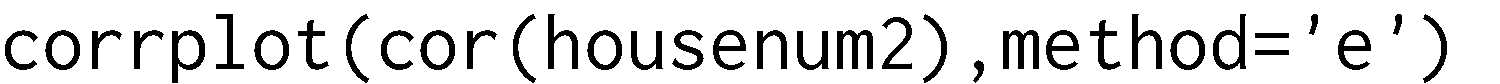


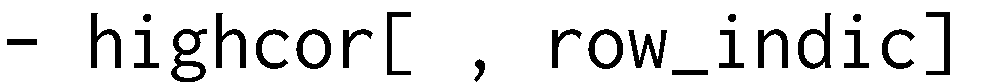
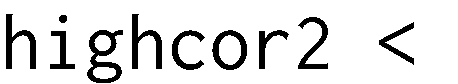
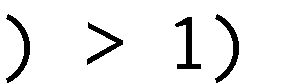
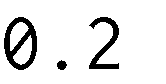
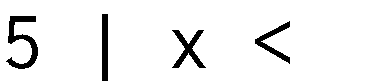
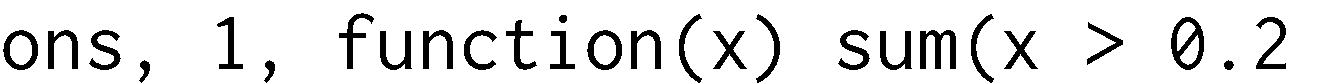
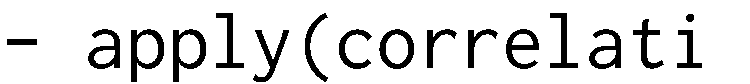
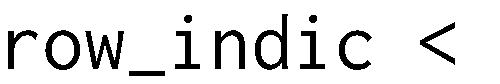
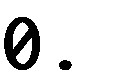
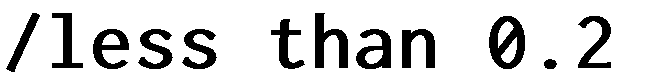
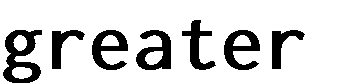
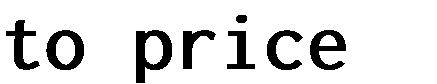
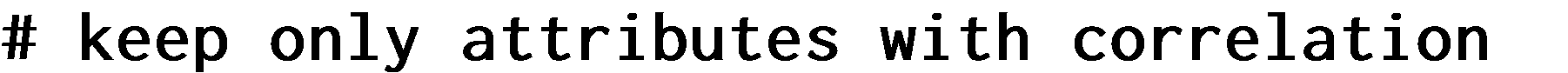


## Reduce number of columns



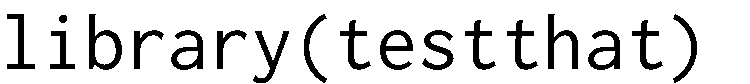


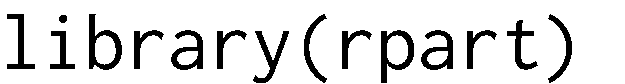


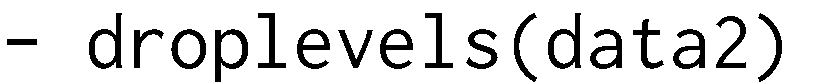
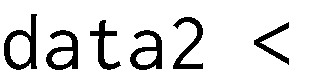
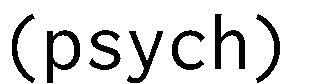
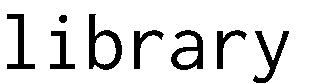


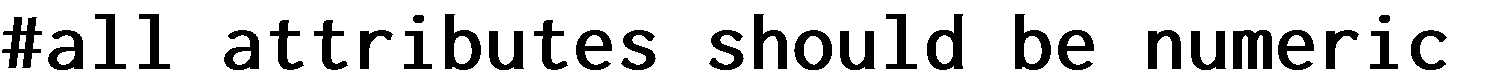
## Lasso

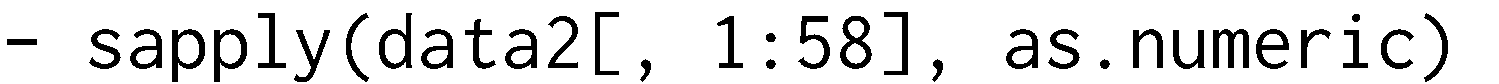
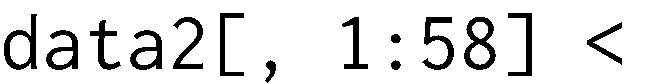
#lasso - reduce number of attributes

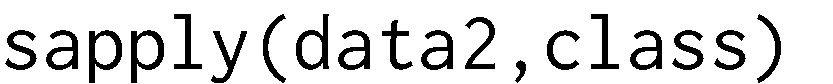


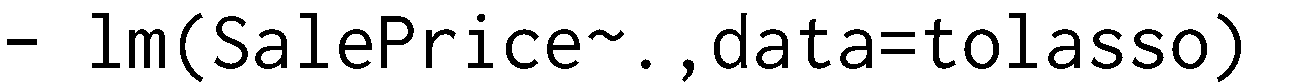
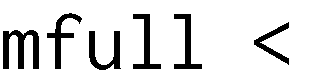
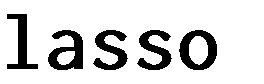
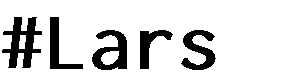
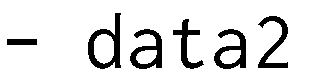
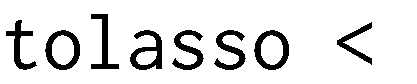


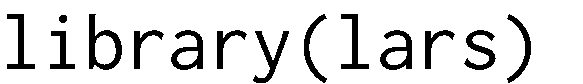


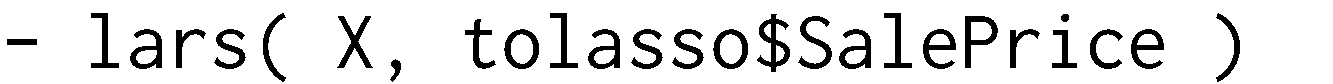
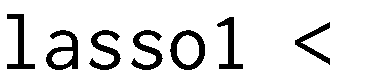
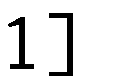
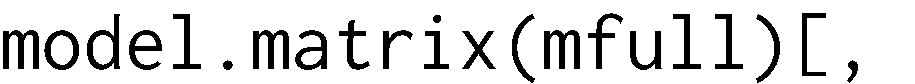
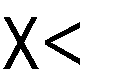


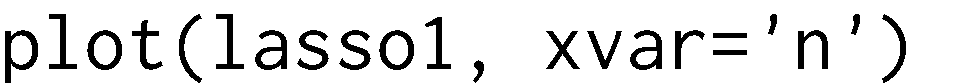


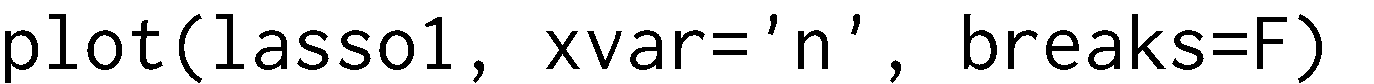


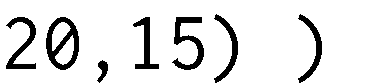
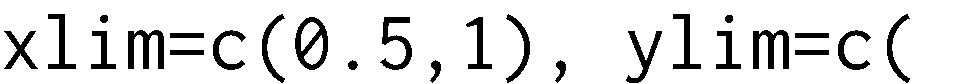
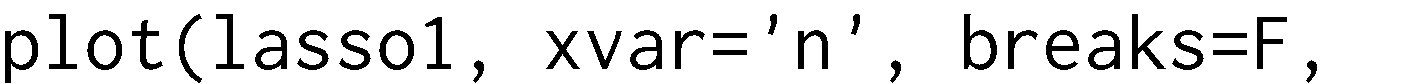


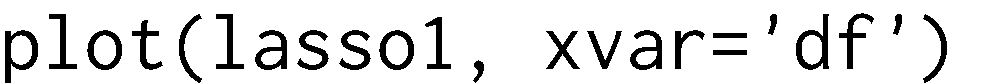


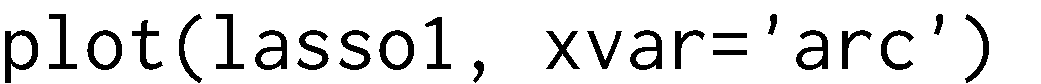


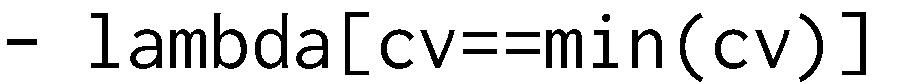
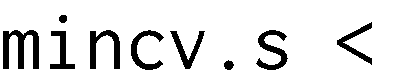
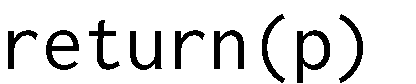
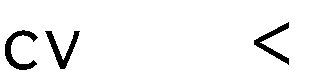
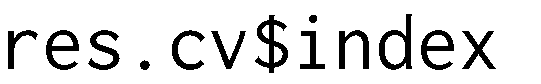
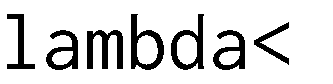
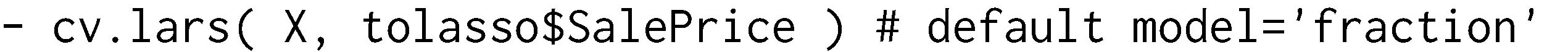
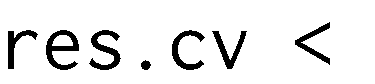
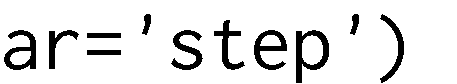
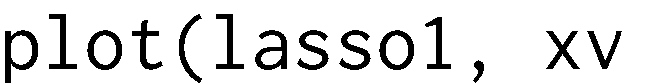


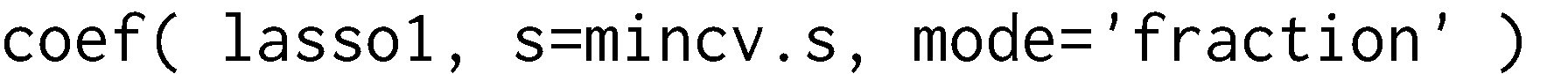


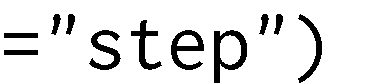
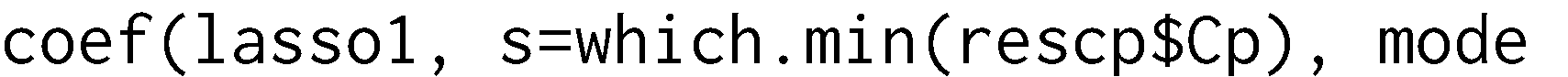
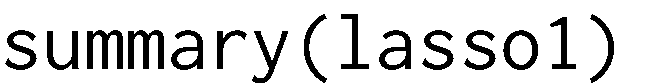
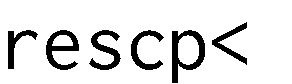


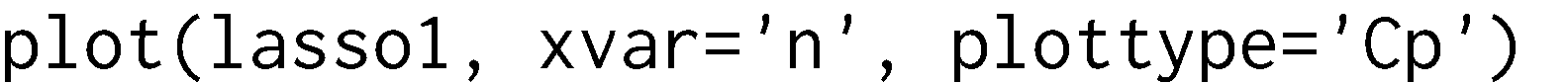


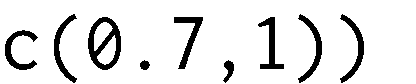
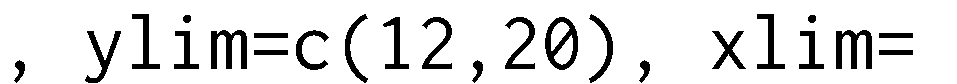
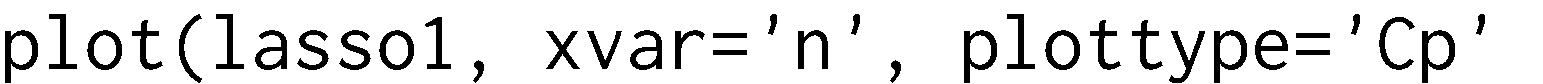


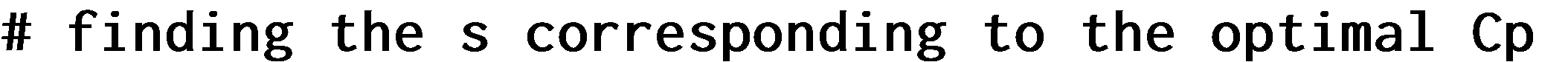


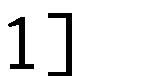
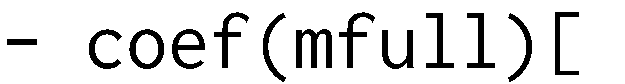
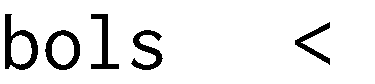
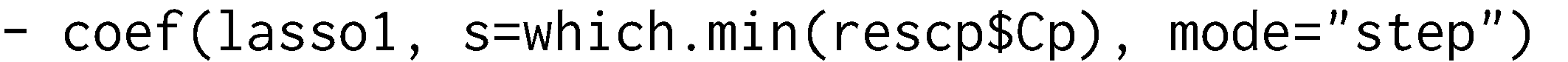
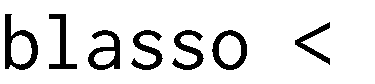


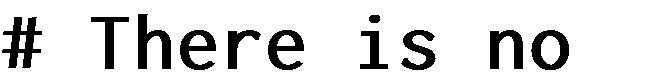


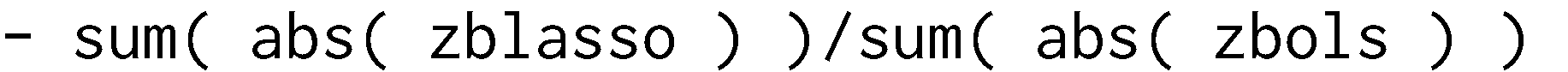
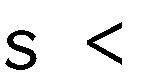
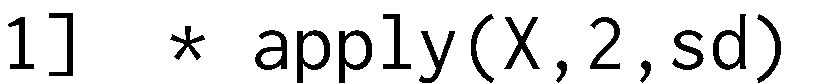
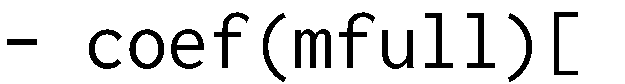
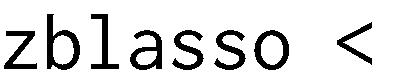
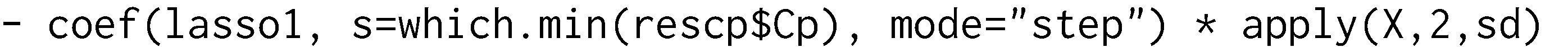
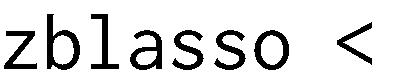




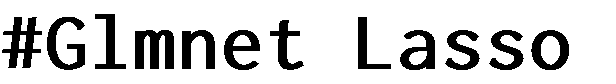


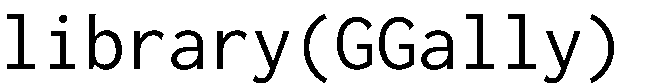


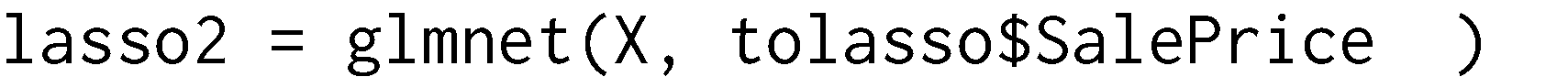


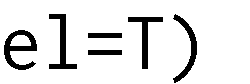
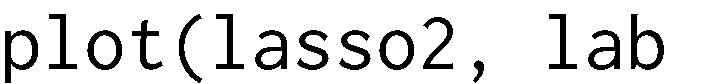


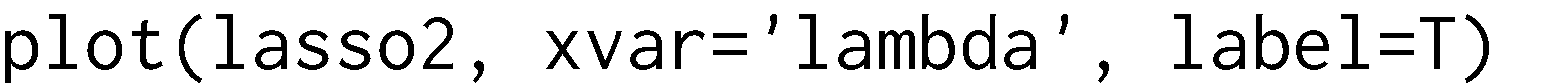


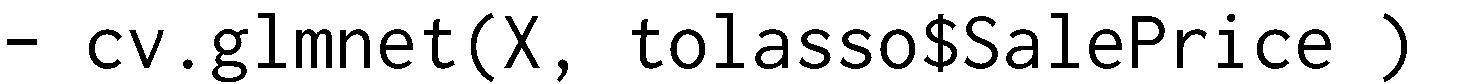
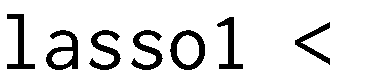
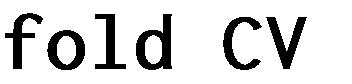
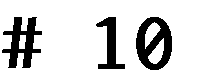


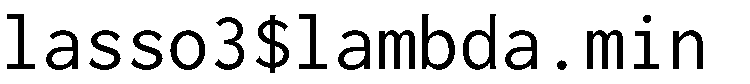


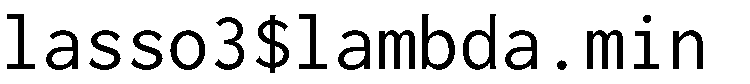


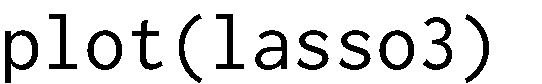


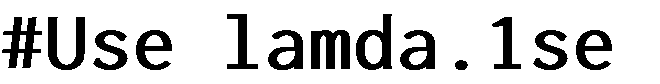


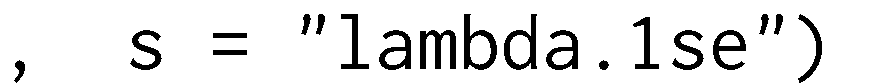
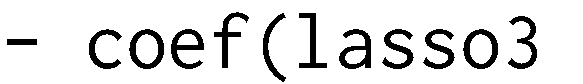
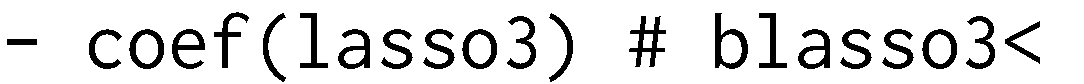
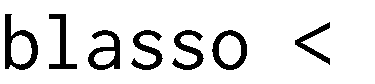


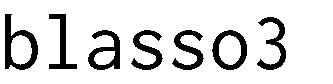


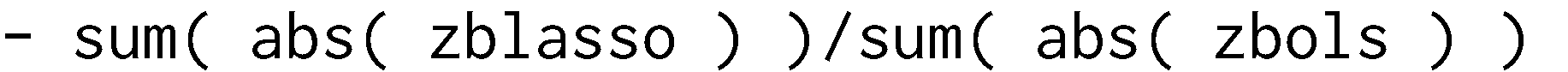
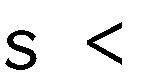
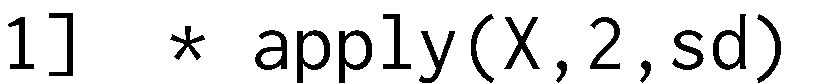
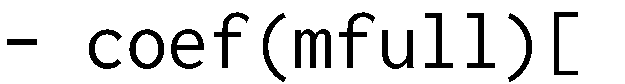
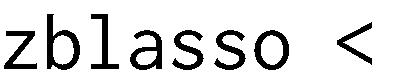
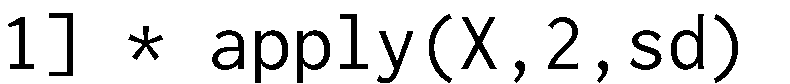
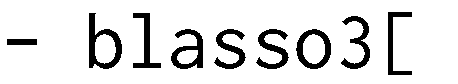
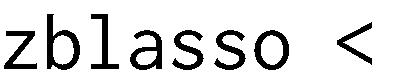






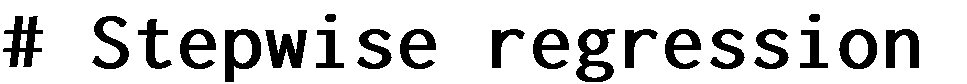


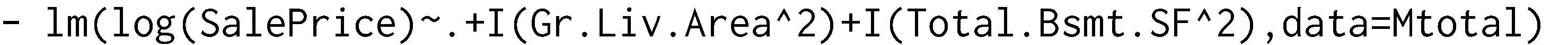
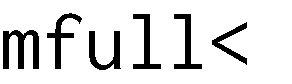


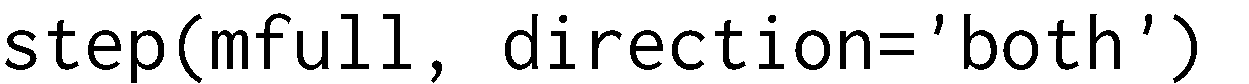


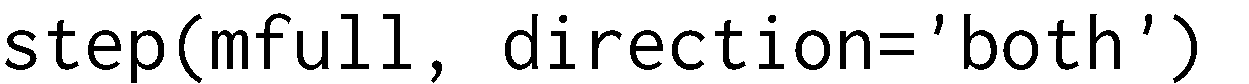


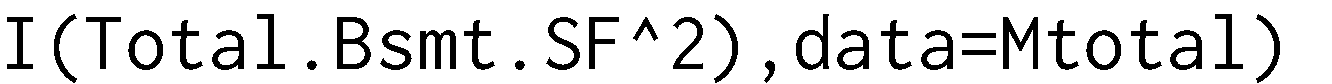
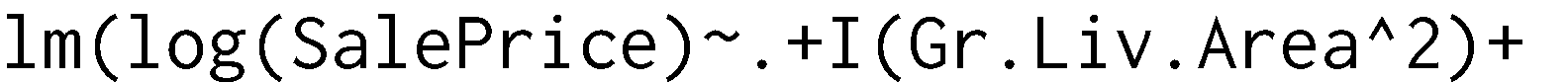
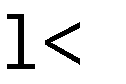
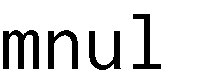
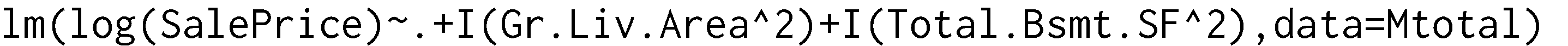
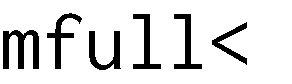
## Train regression models

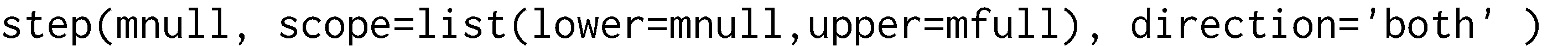


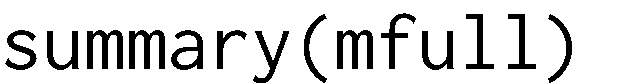


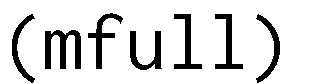
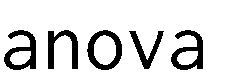


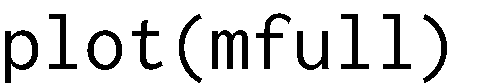




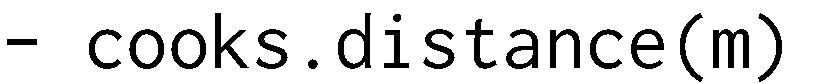
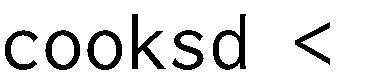


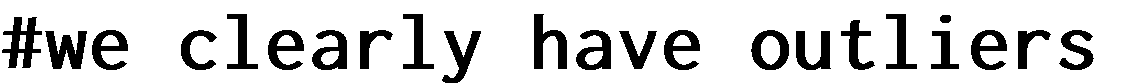


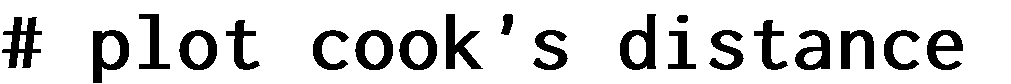


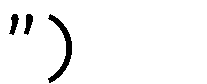
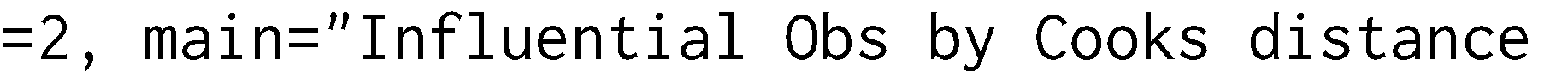
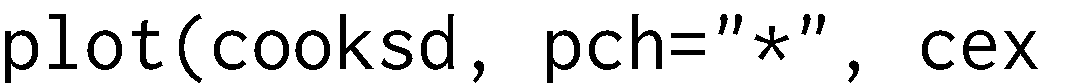


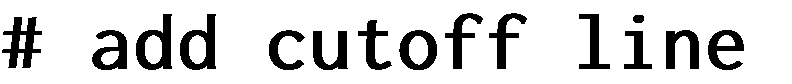
## Evaluate outliers

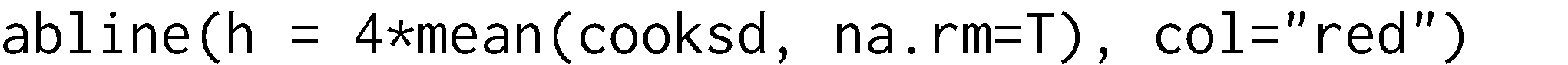


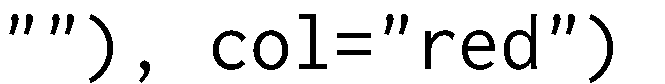
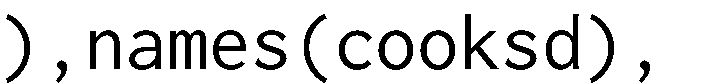
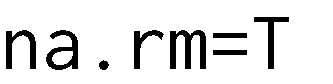
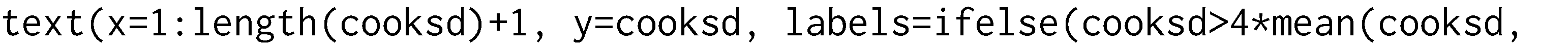


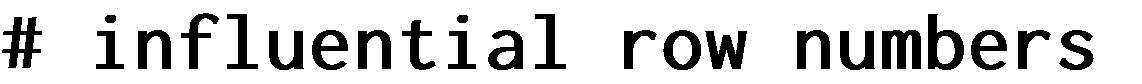


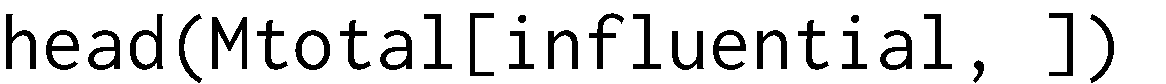
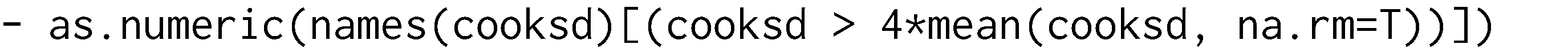
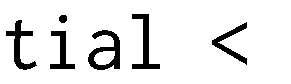
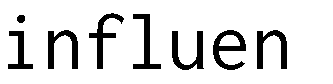


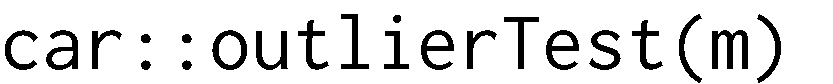


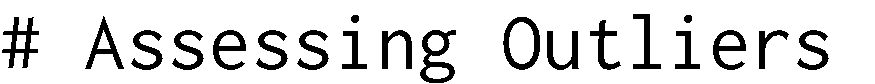


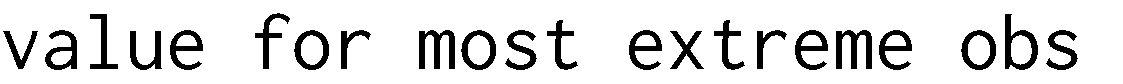
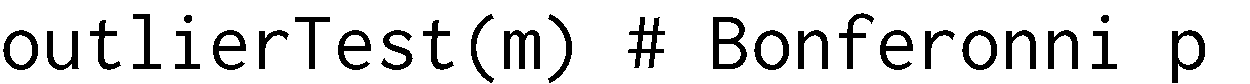


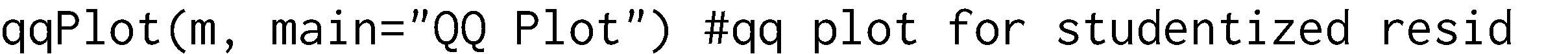


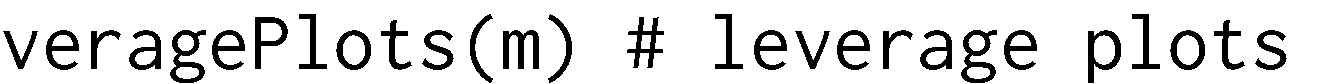
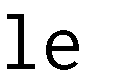


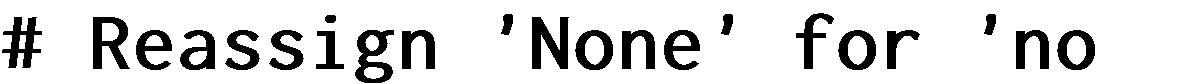


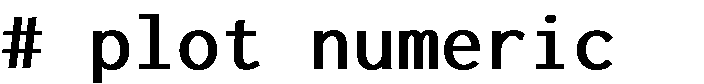


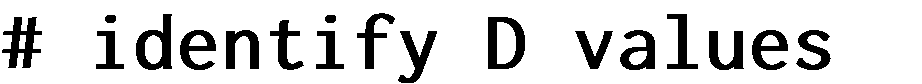


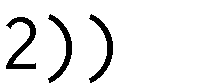
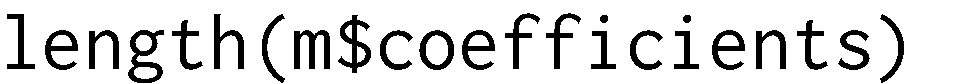
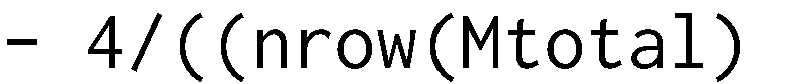
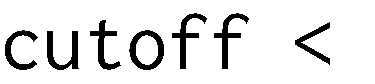


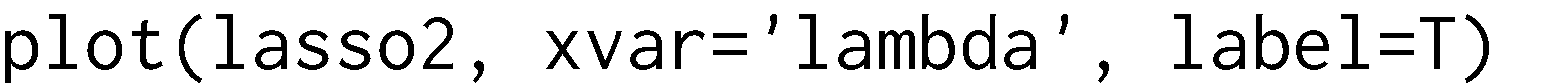


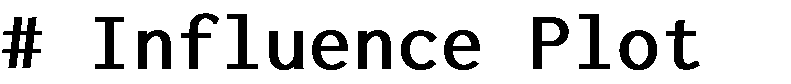


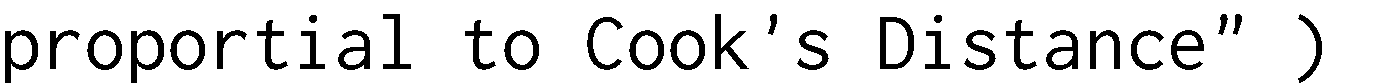
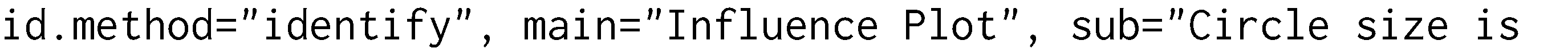
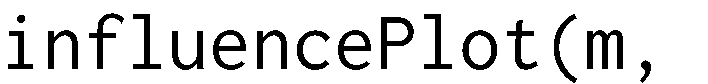


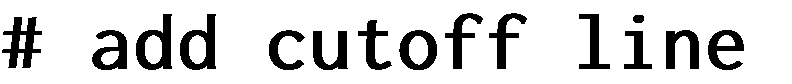


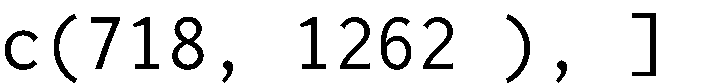
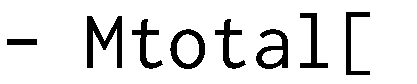
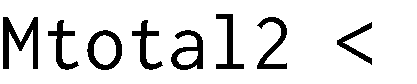




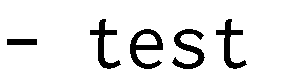
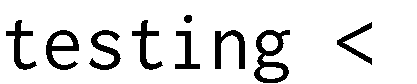
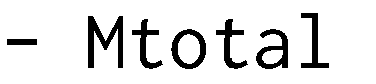
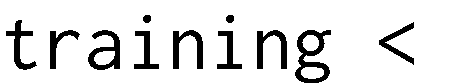


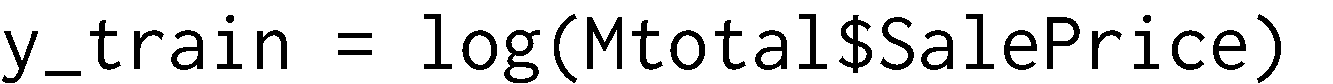


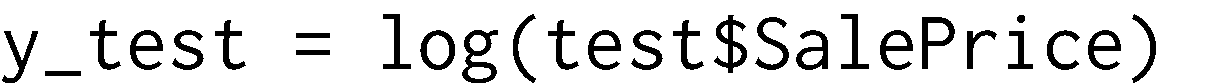


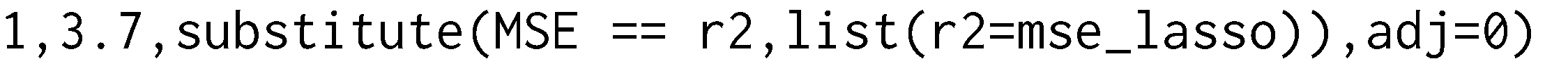
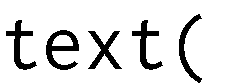
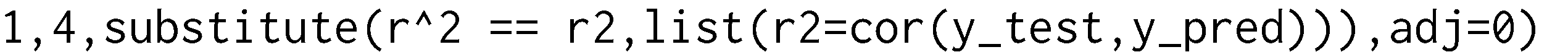
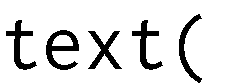
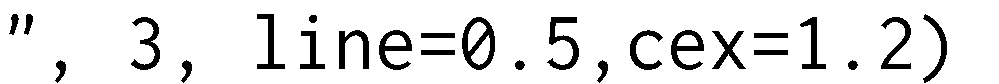
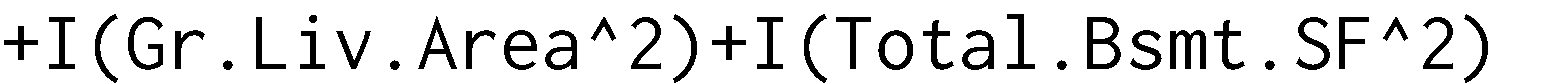
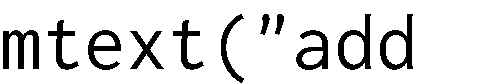
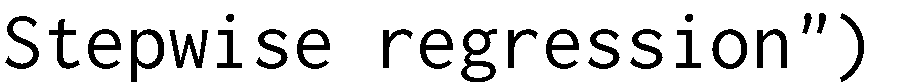
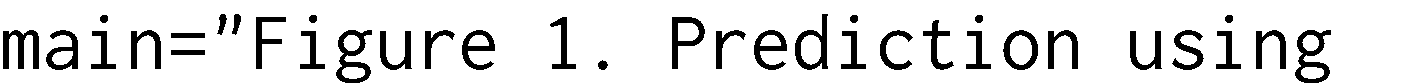
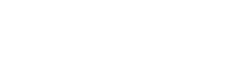
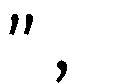
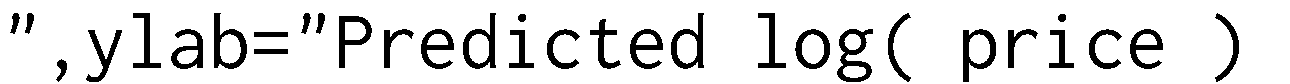
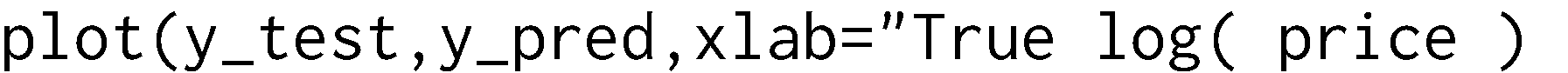
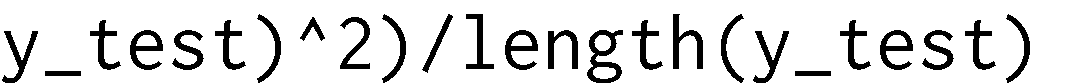
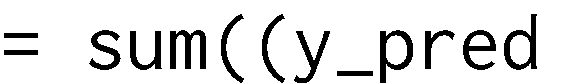
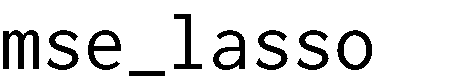


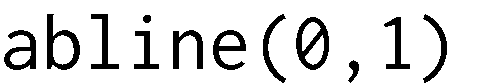
## Plot predicted vs actual & residuals

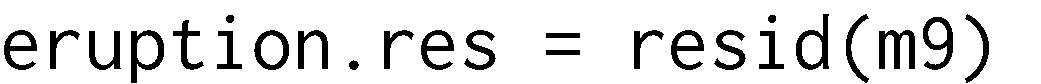


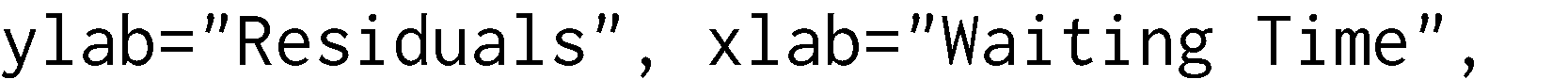
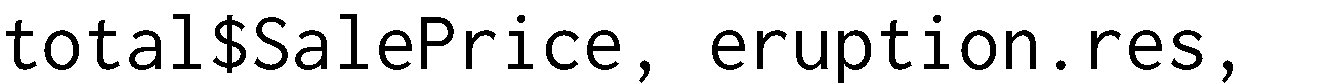
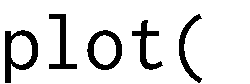


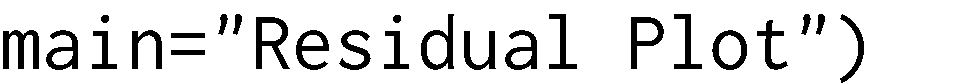


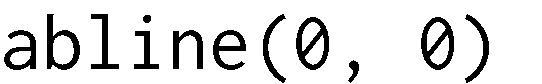


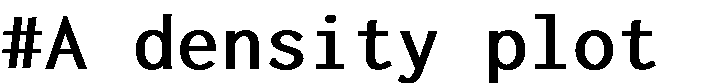


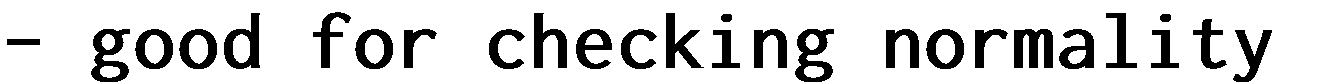
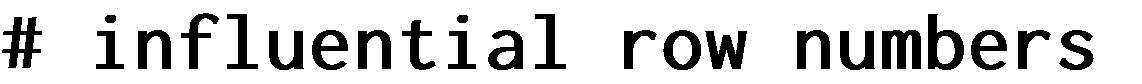
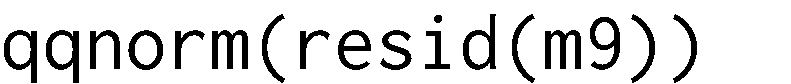
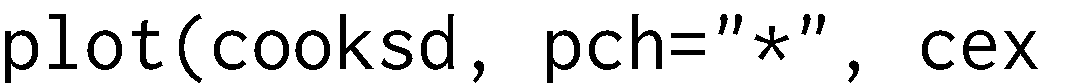


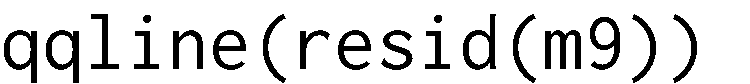


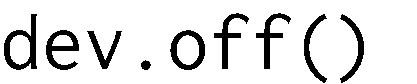


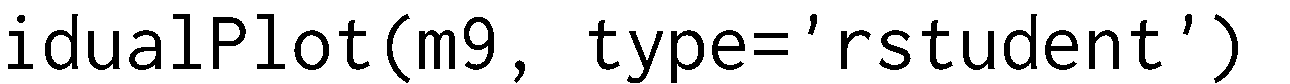
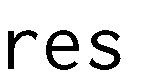


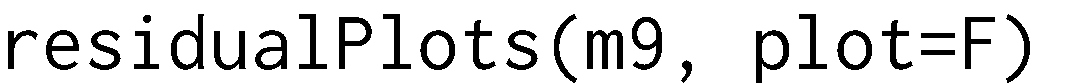


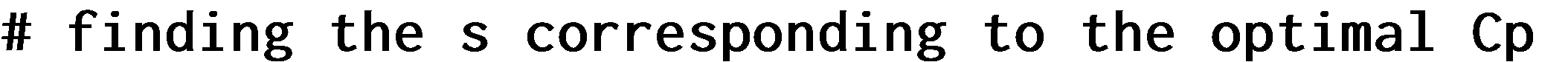


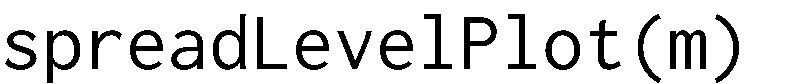




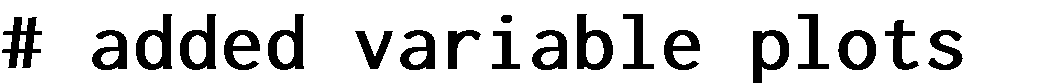


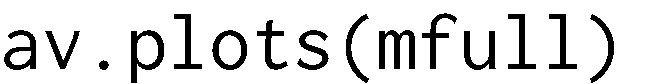


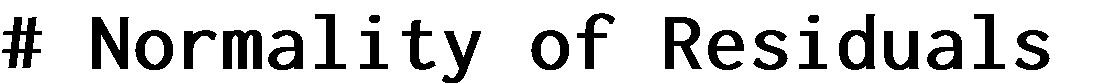


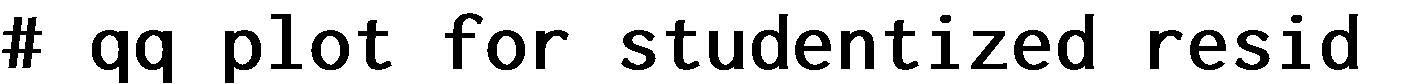


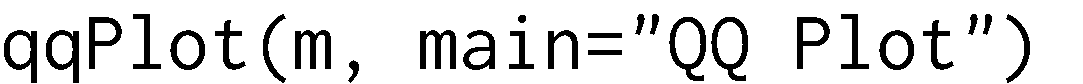
## More plots

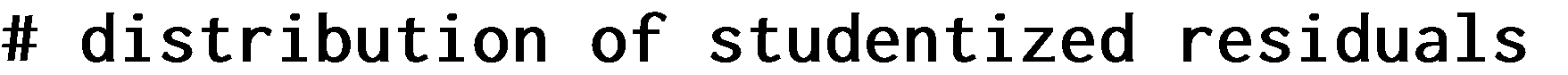


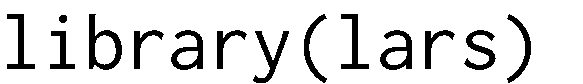


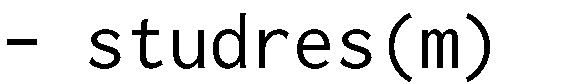
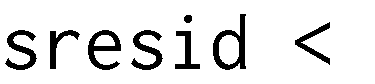


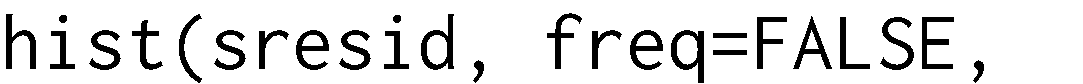


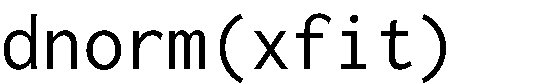
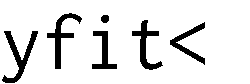
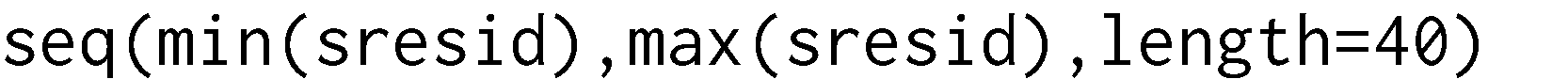
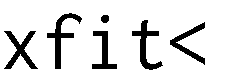
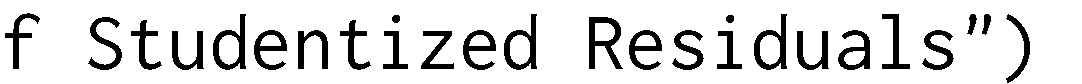
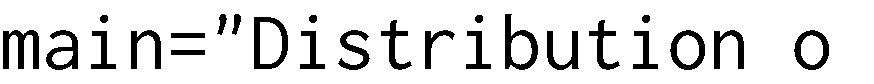
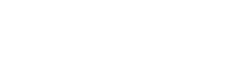


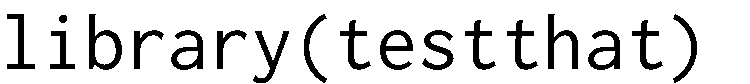


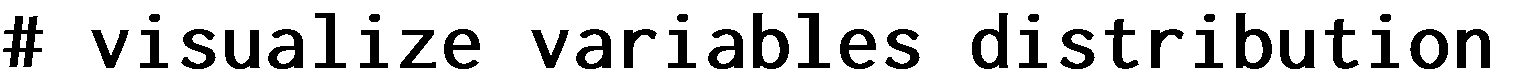


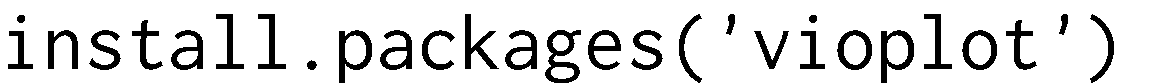


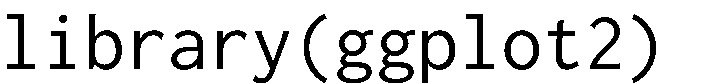


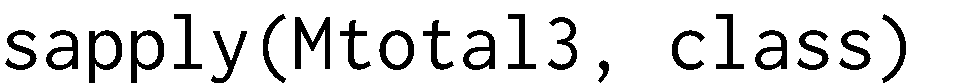


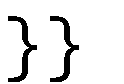
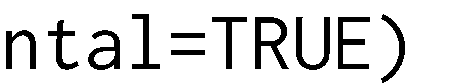
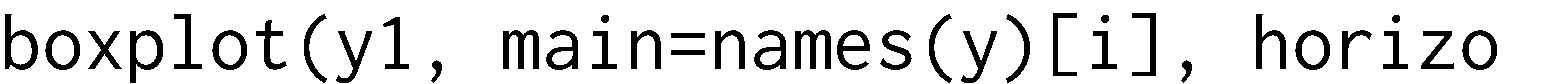
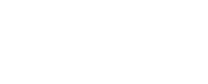
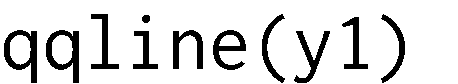
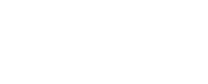
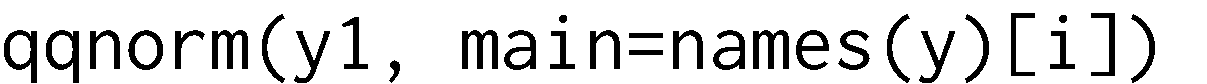
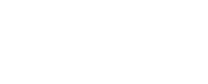
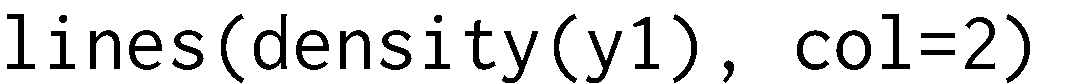
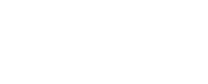
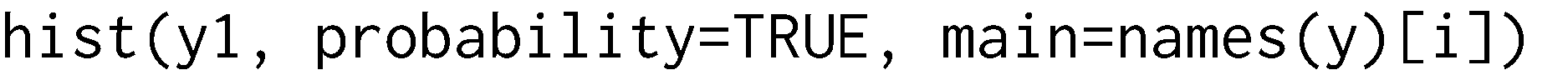
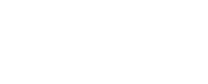
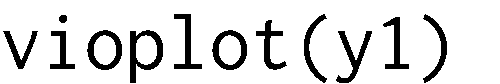
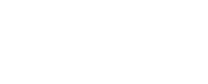
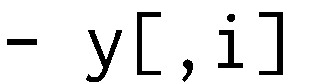
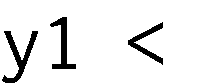
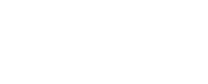
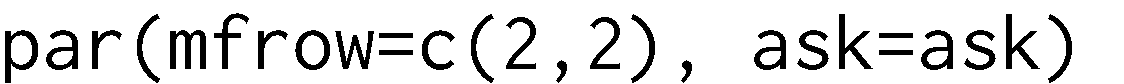
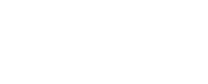
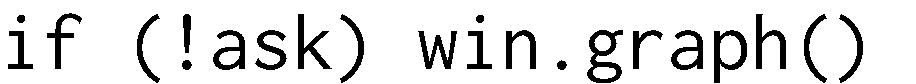
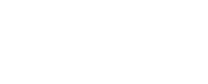
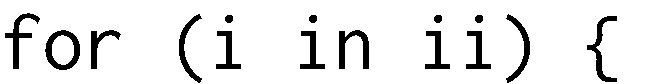
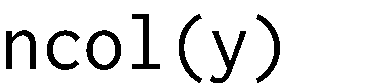
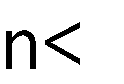
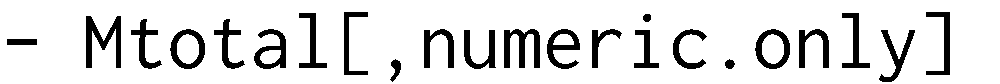
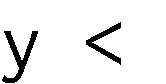
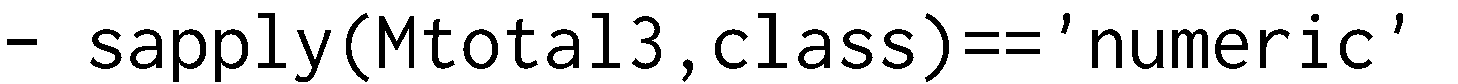
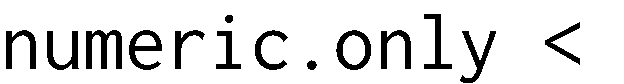
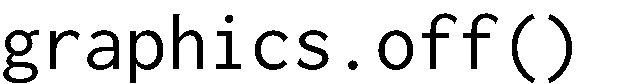
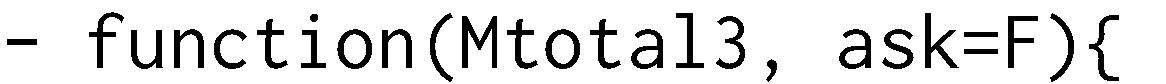
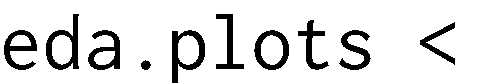
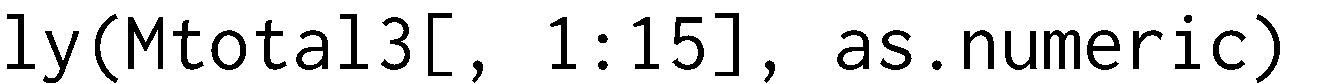
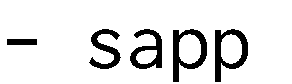
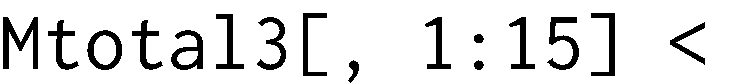


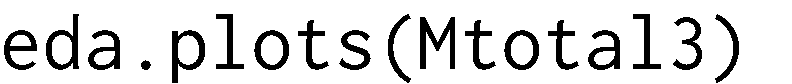


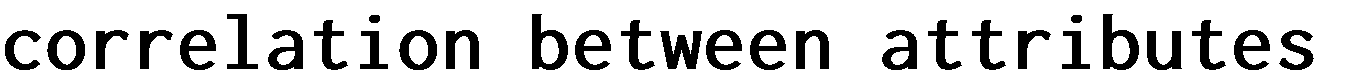
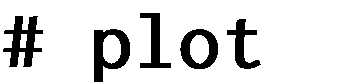


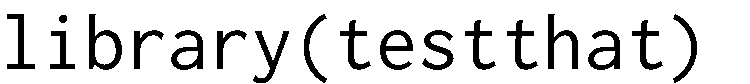


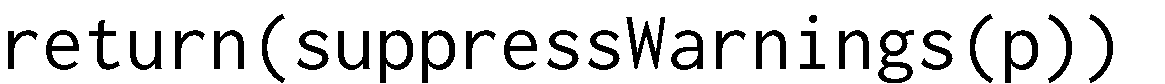
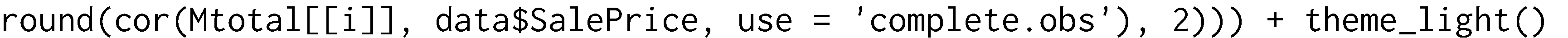
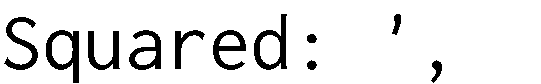
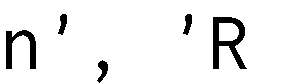
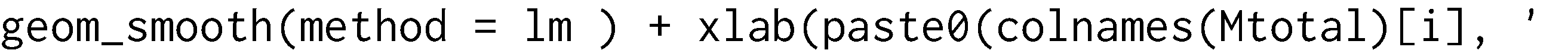
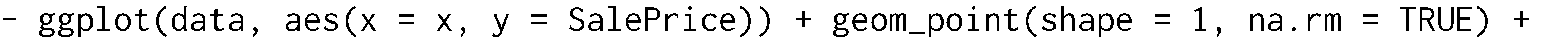
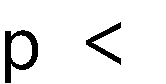
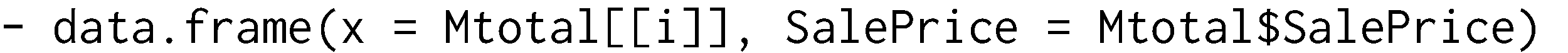
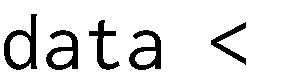
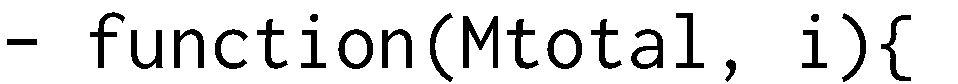
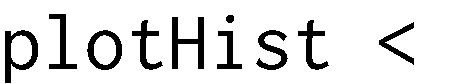
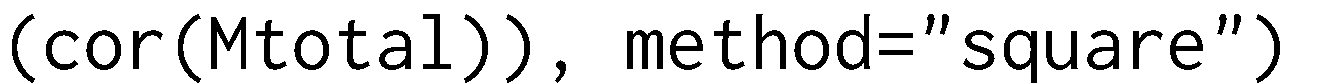
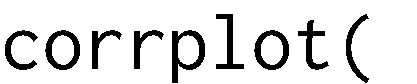




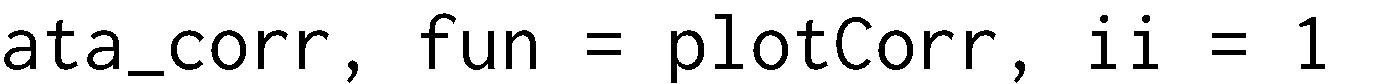
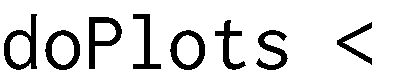




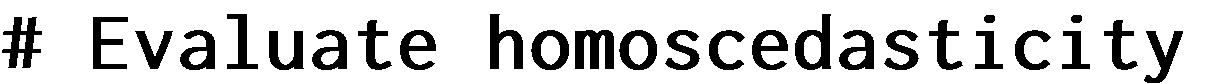


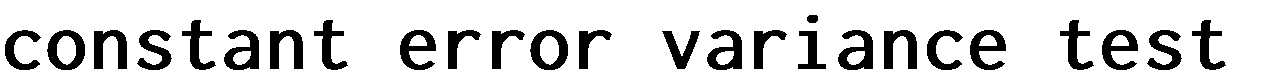
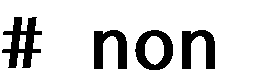


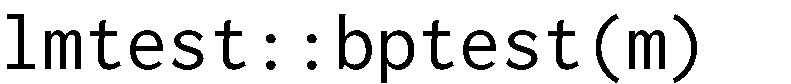


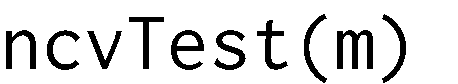


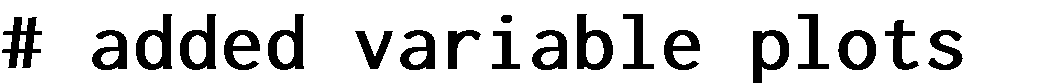
## Evaluate regression model assumptions

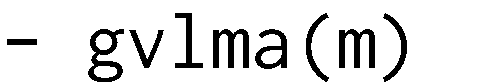
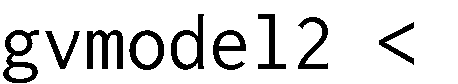
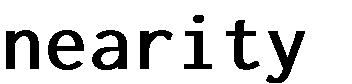
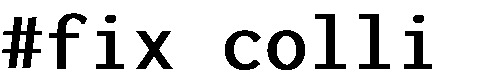
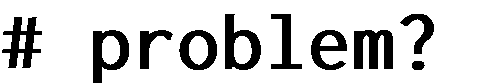
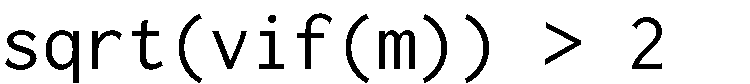
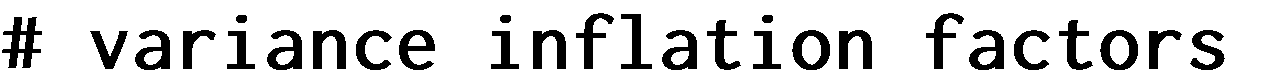
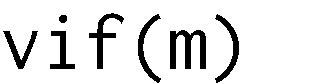


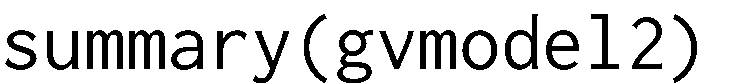


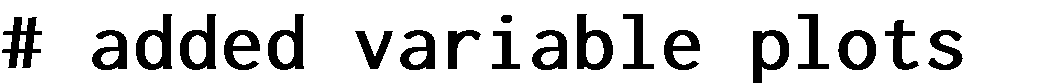


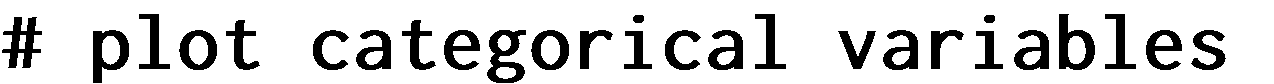


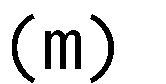
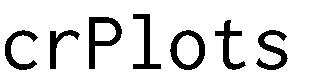


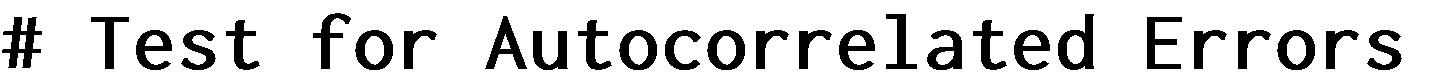


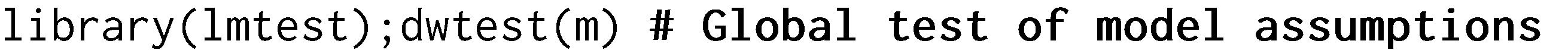


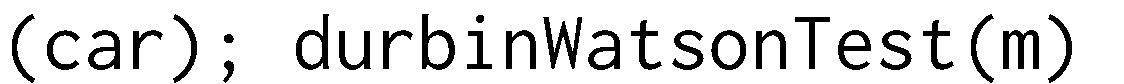
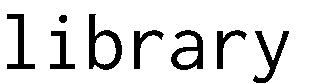


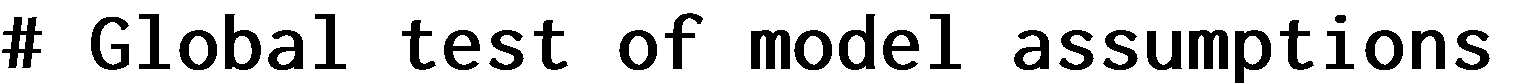


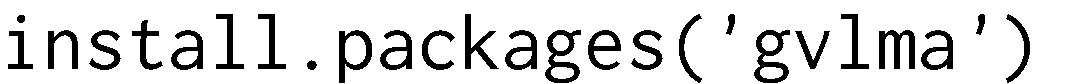


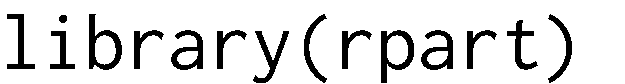


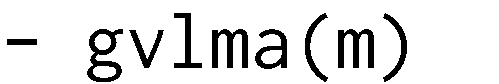
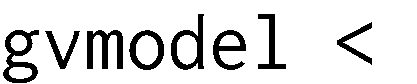


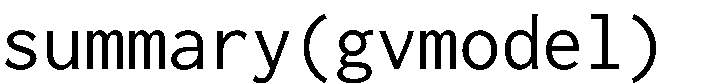


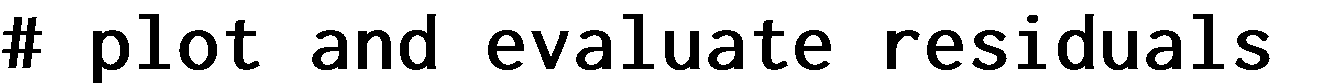


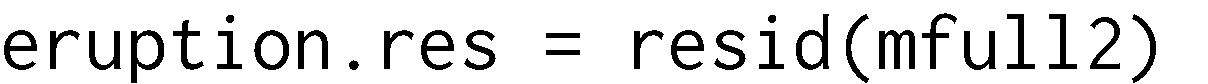


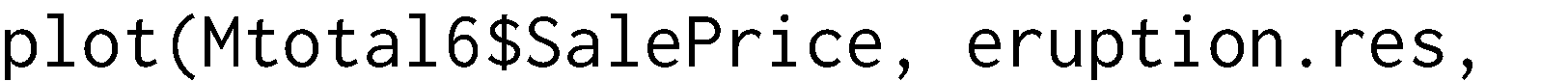


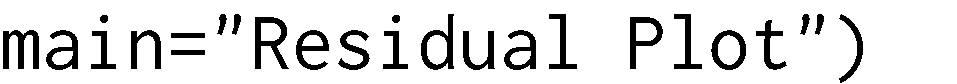
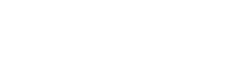
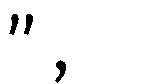
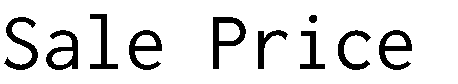
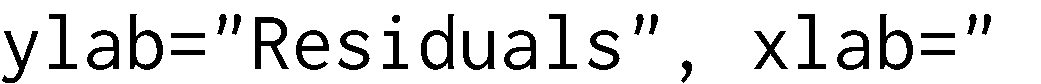
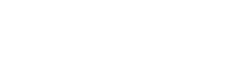


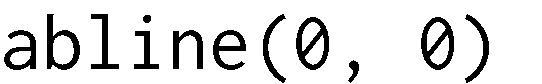


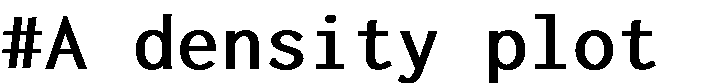


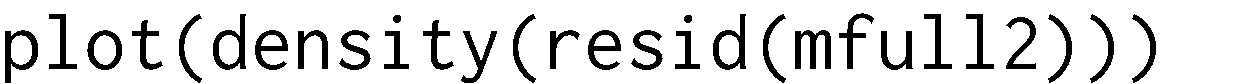


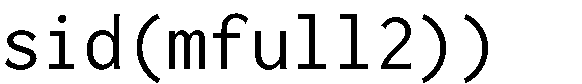
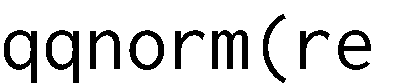
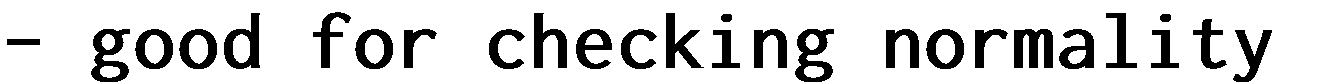
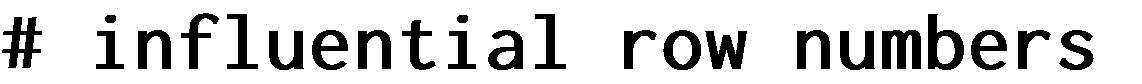


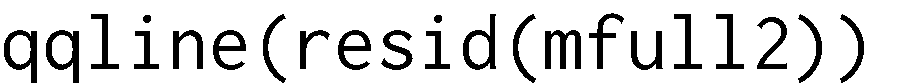


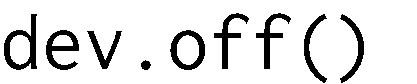


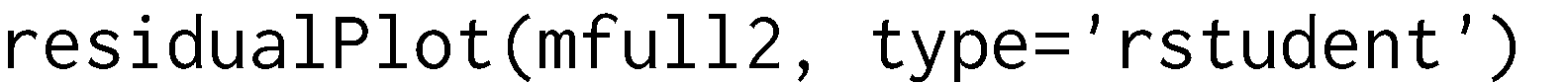


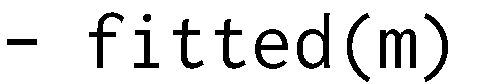
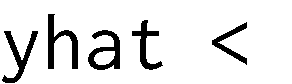
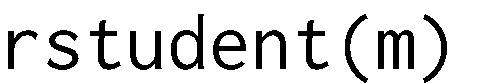
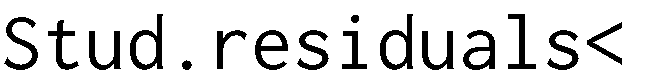
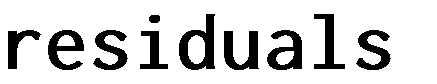
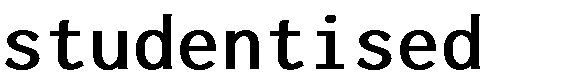
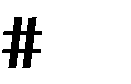


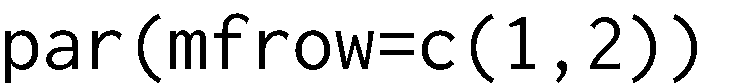


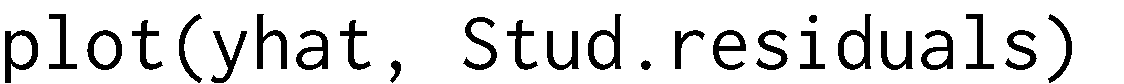


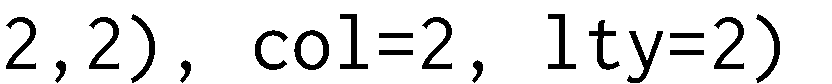
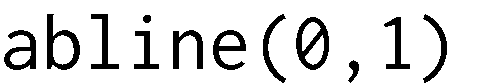


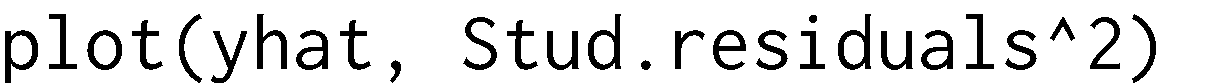


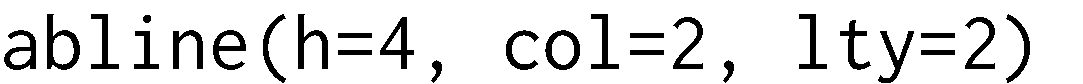


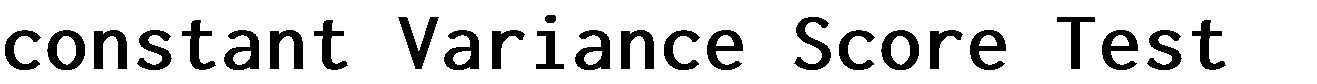
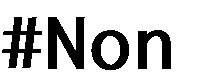


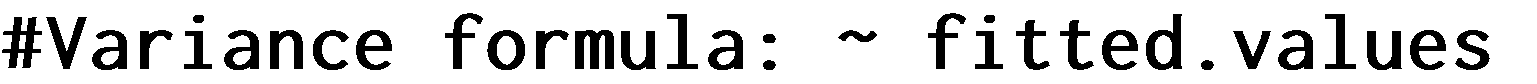


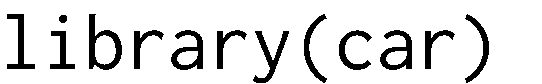


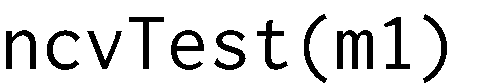


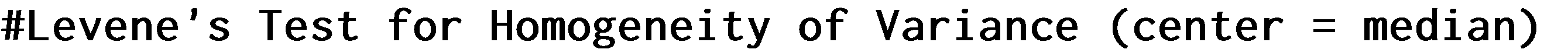


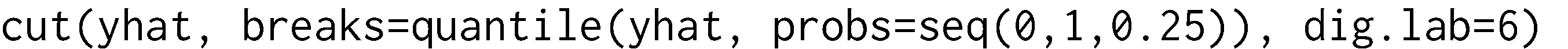
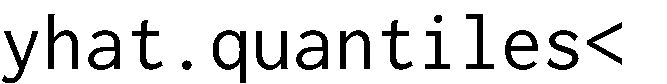


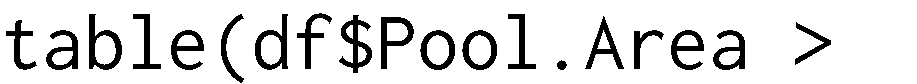


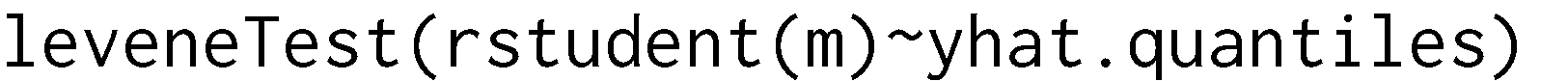


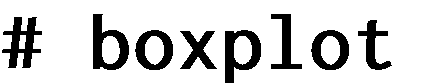


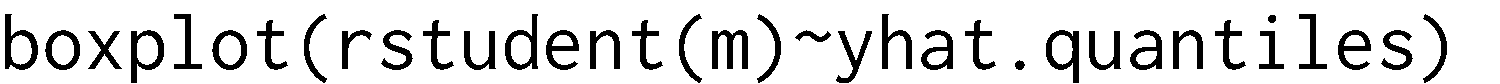


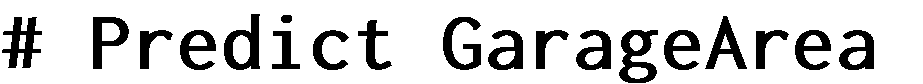


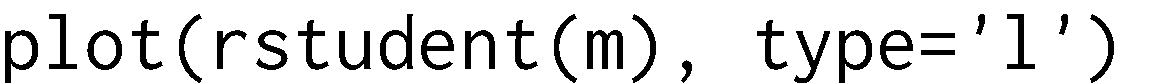


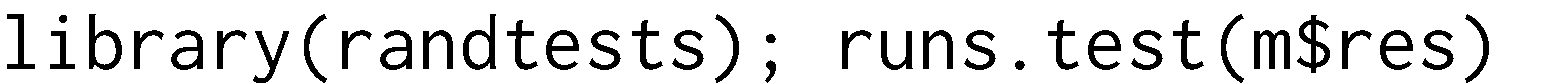




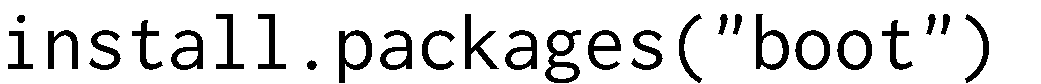


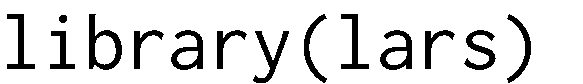


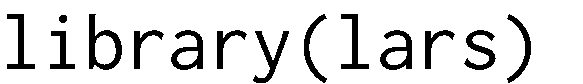


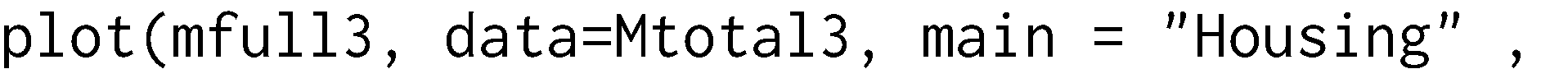


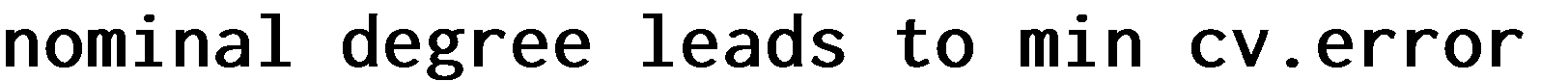
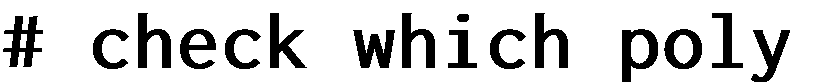
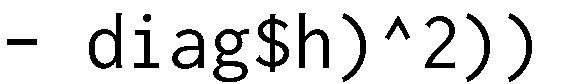
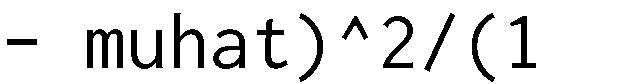
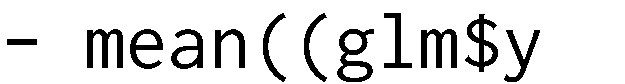
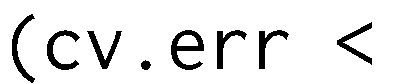
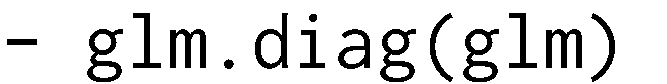
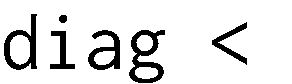
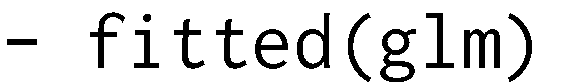
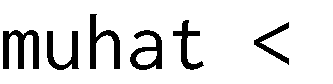
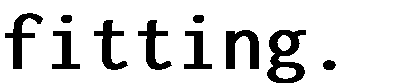
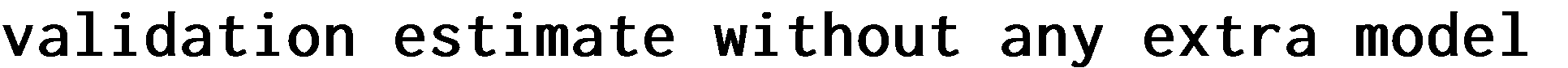
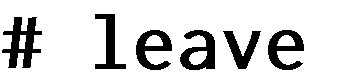
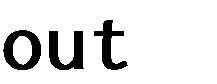
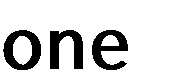
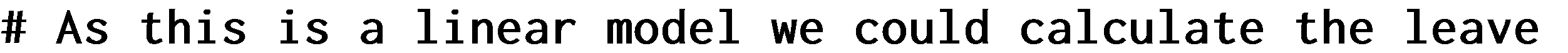
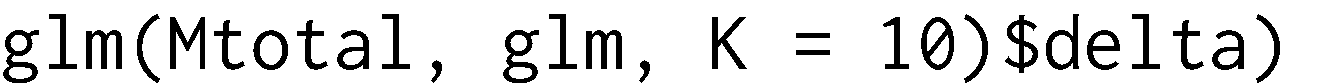
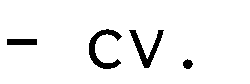
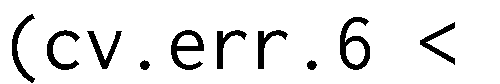
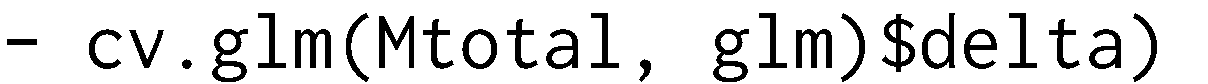
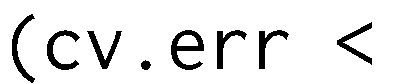
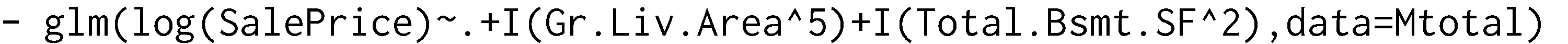
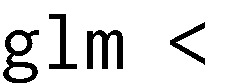
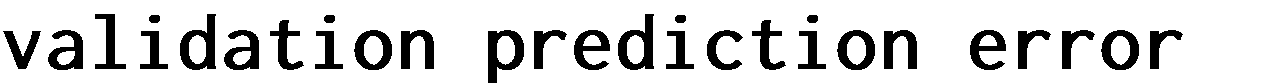
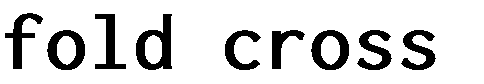
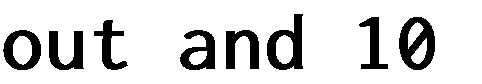
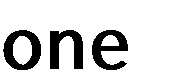
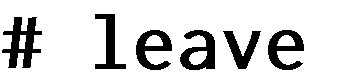
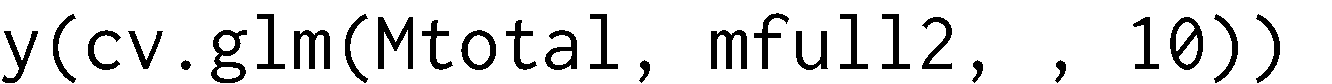
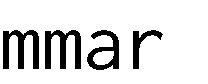
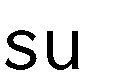
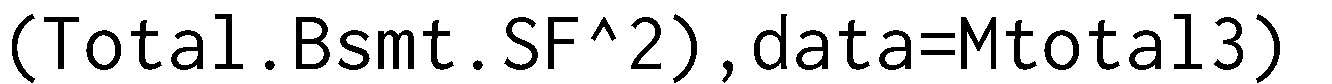
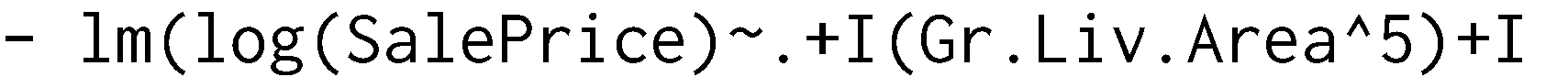
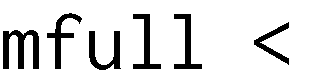
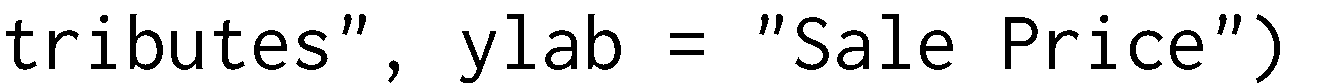
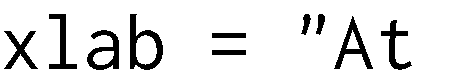
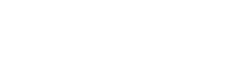
## Improve model

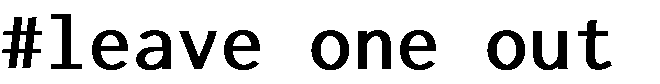


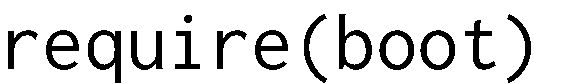


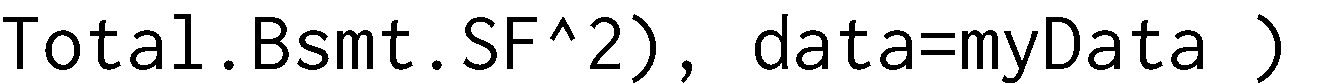
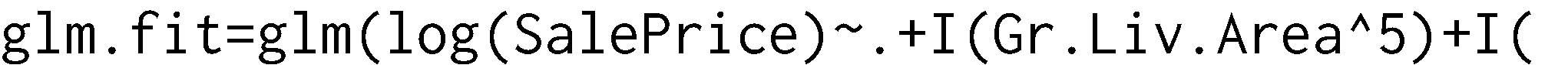
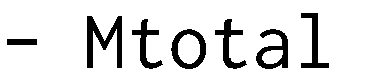
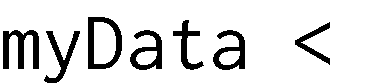


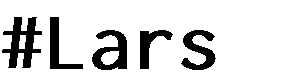


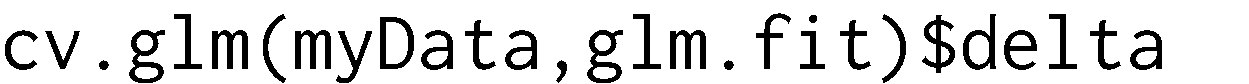


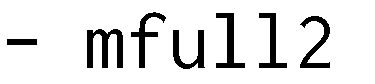
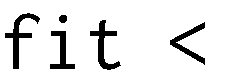


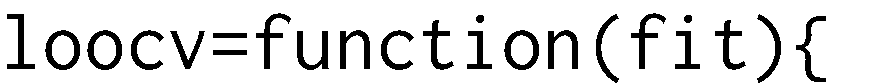


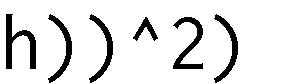
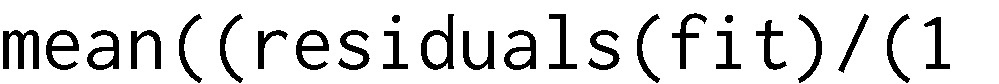
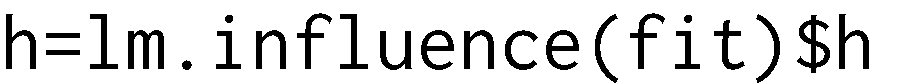




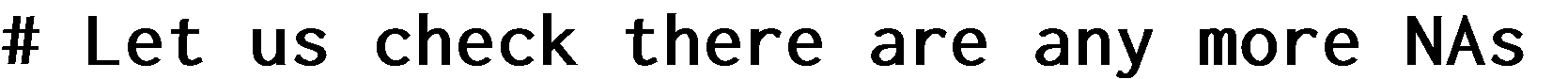


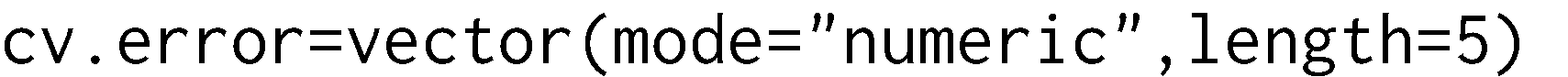


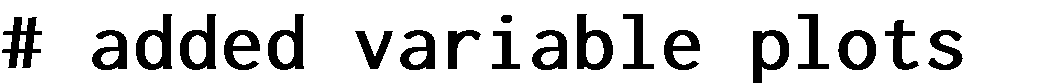


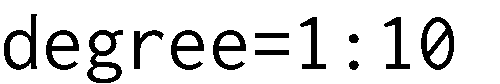


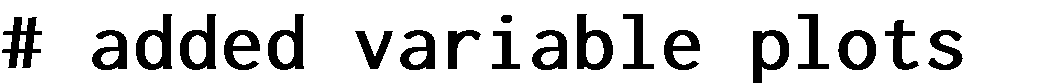


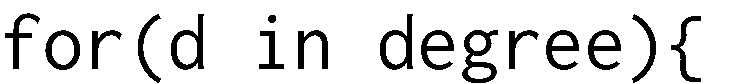


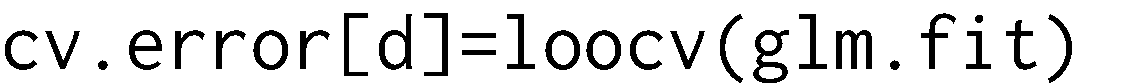
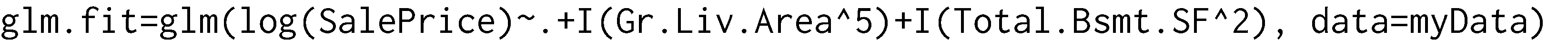




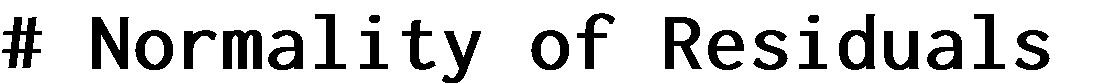


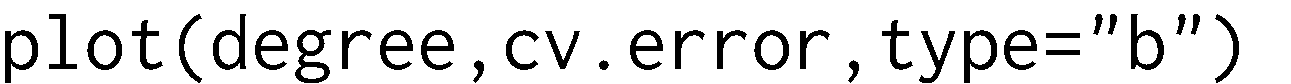


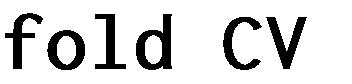
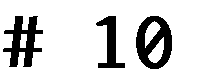


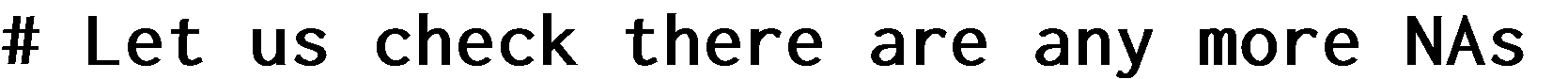


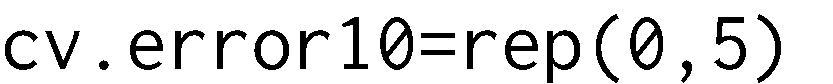


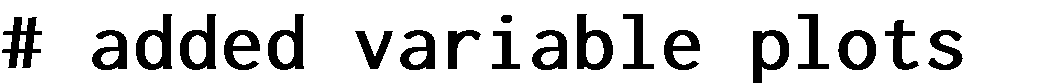


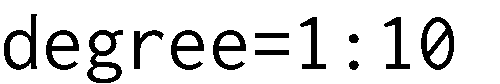


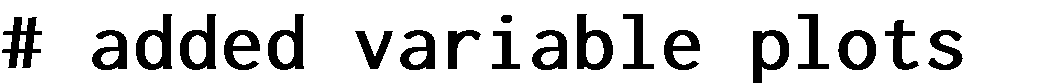


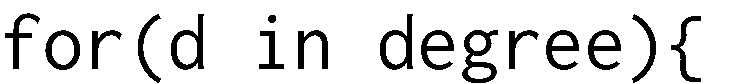


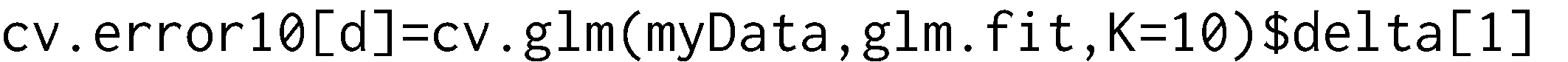
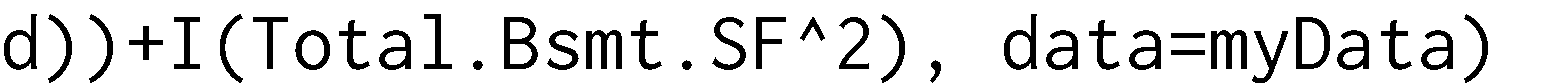
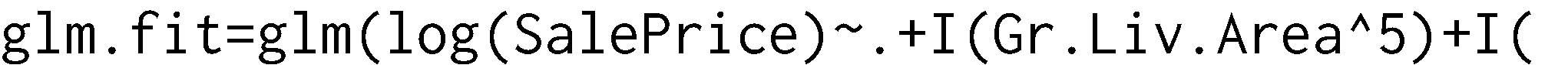




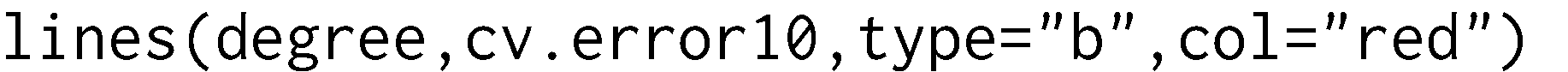




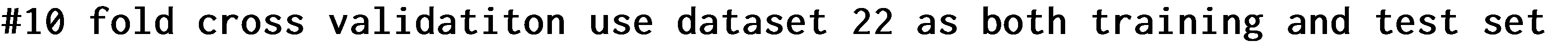


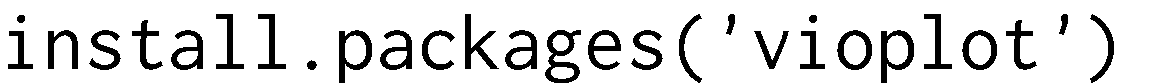


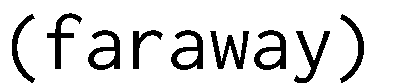
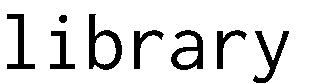


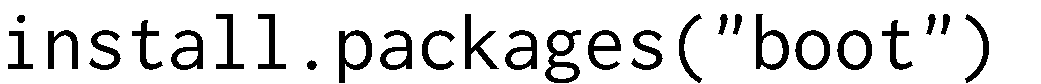


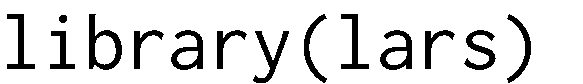
## Test model

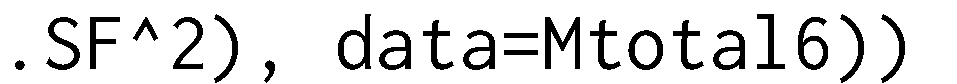
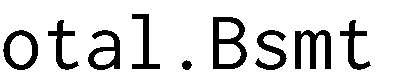
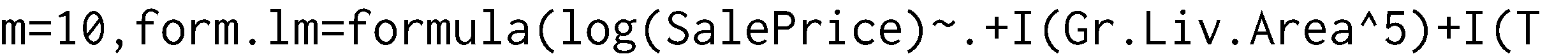
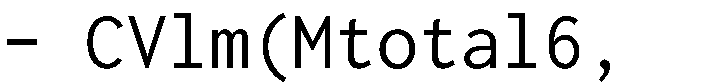
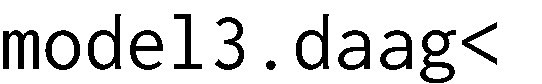


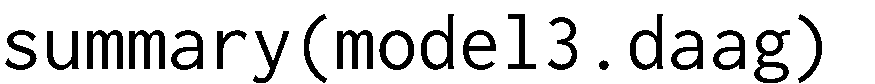


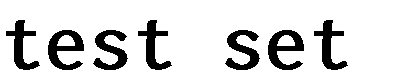
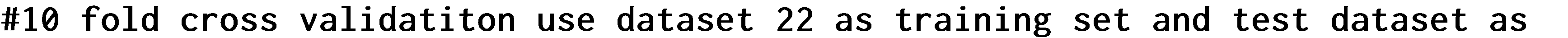


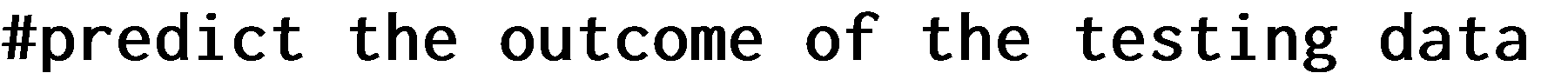


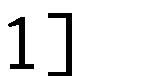
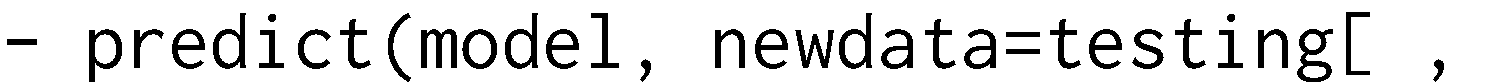
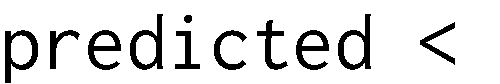


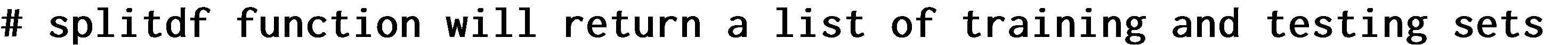


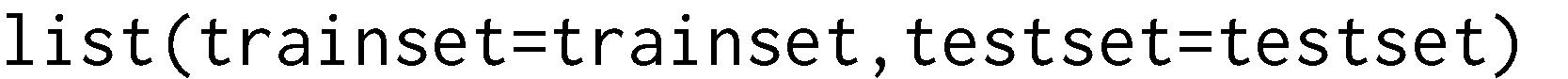
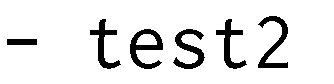
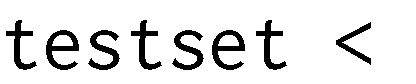
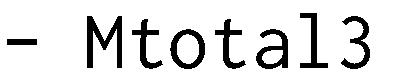
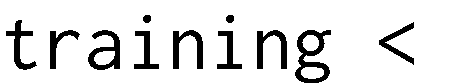
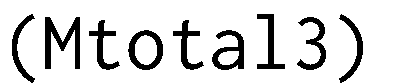
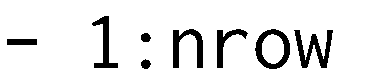
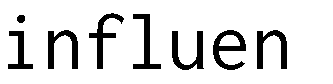
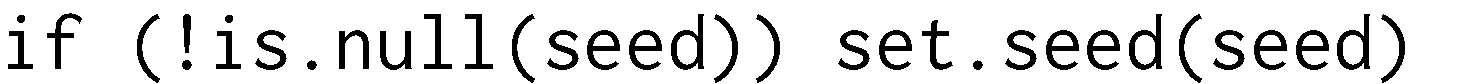
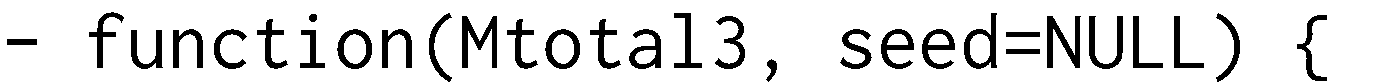
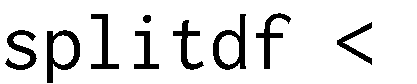


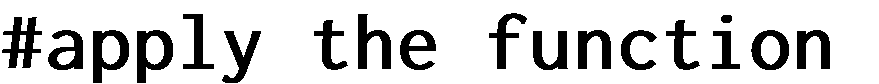


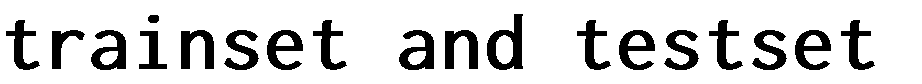
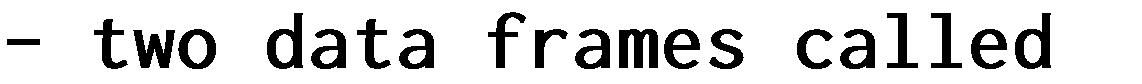
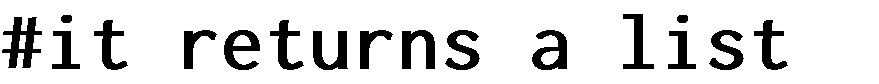
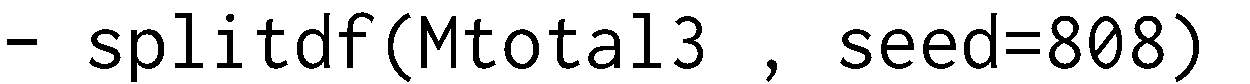
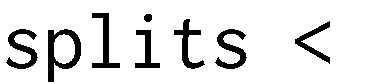


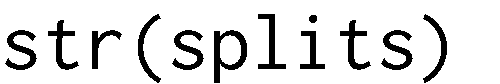


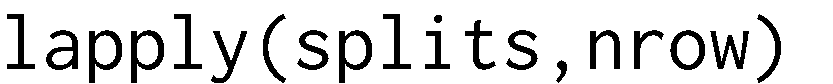


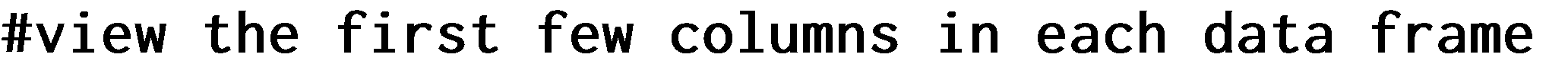


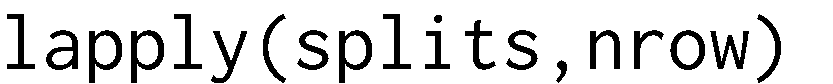


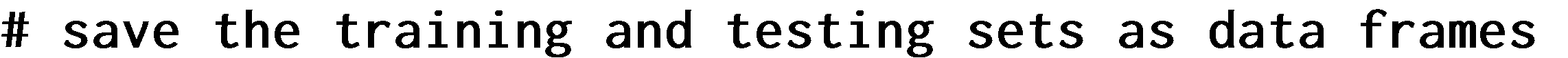


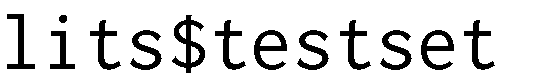
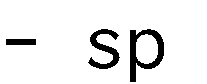
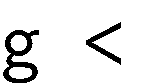
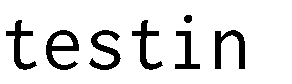
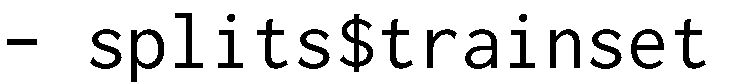
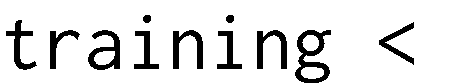


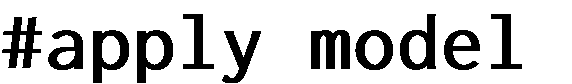


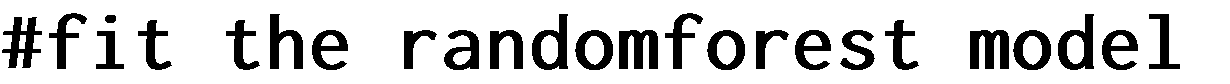


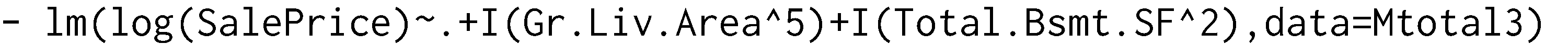
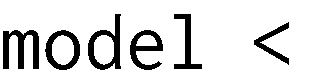


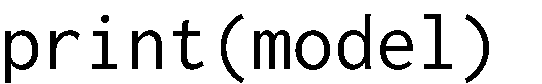


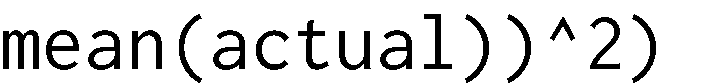
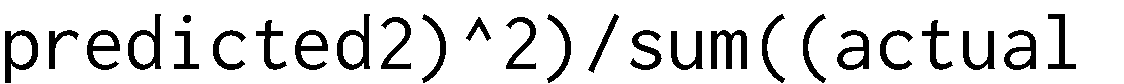
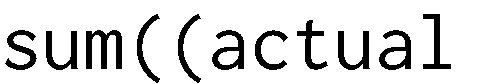
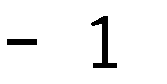
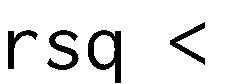
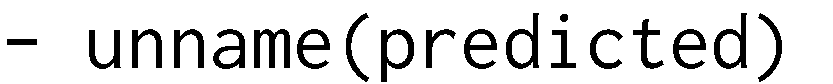
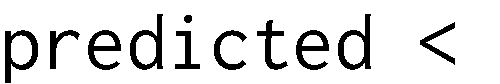
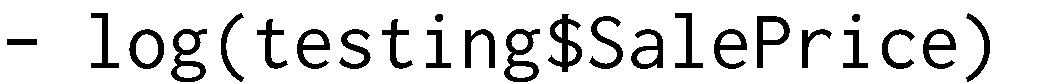
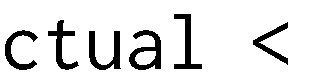
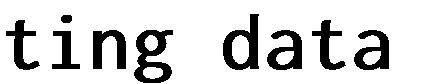
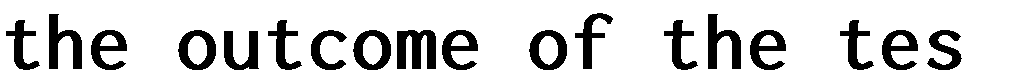
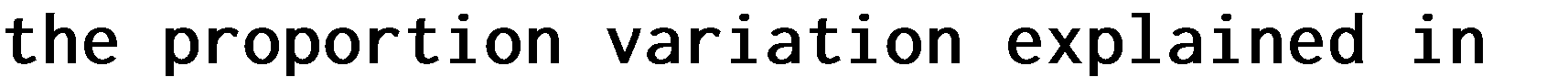
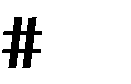


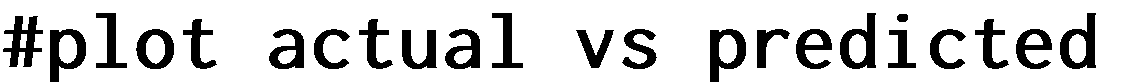


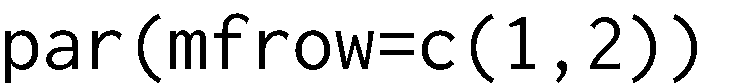


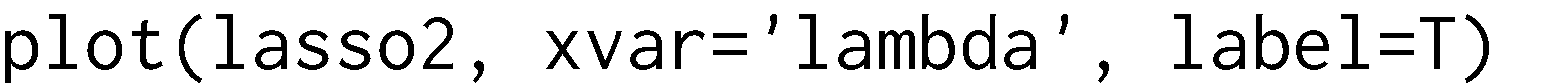


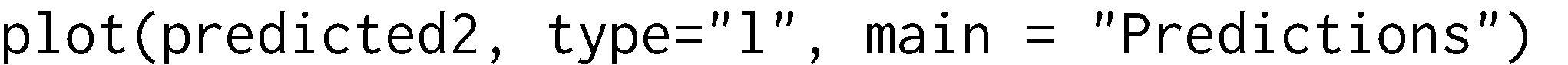


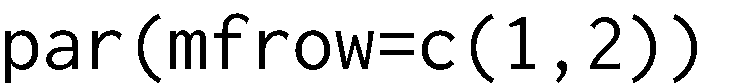


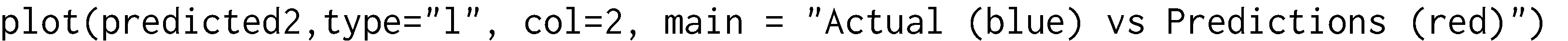


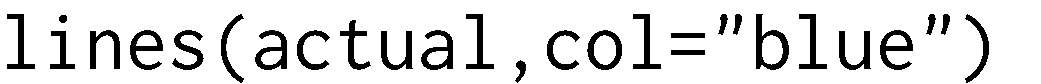


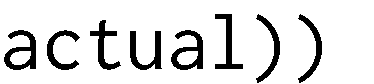
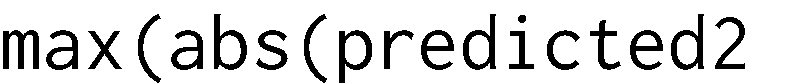












# Bibliography

#### Introduction

Journal of Statistics Education Volume 19, Number 3(2011), https://ww2.amstat.org/publications/jse/v19n3/Decock/DataDocumentation.txt

Journal of Statistics Education Volume 19, Number 3(2011),<http://ww2.amstat.org/publications/jse/v19n3/decock.pdf>

Goldberger, Arthur S. (1991). "Multicollinearity". A Course in Econometrics. Cambridge: Harvard University Press. pp. 245–53.

#### Data cleaning

https:[//www.kaggle.com/bisaria/house-prices-advanced-regression-techniques/handling-missing-data](http://www.kaggle.com/bisaria/house-prices-advanced-regression-techniques/handling-missing-data)

#### Exploratory analysis

https:[//www.kaggle.com/notaapple/house-prices-advanced-regression-techniques/detailed-exploratory-](http://www.kaggle.com/notaapple/house-prices-advanced-regression-techniques/detailed-exploratory-) data-analysis-using-r.

#### Lasso

[http://statweb.stanford.edu/~tibs/lasso/simple.html](http://statweb.stanford.edu/%7Etibs/lasso/simple.html) <http://stats.stackexchange.com/questions/68431/interpretting-lasso-variable-trace-plots> **Regression model interpretation**

<http://www.statisticssolutions.com/assumptions-of-linear-regression/> https://en.wikipedia.org/wiki/Multivariate\_normal\_distribution [http://people.duke.edu/~rnau/rsquared.htm](http://people.duke.edu/%7Ernau/rsquared.htm) <http://strata.uga.edu/8370/rtips/regressionPlots.html>

<https://www.r-bloggers.com/checking-glm-model-assumptions-in-r/>

<http://www.r-tutor.com/elementary-statistics/simple-linear-regression/residual-plot>

**Test linear assumptions** [http://people.duke.edu/~rnau/testing.htm](http://people.duke.edu/%7Ernau/testing.htm) <http://svitsrv25.epfl.ch/R-doc/library/car/html/cr.plots.html>

<http://www2.kenyon.edu/Depts/Math/hartlaub/Math305%20Fall2011/R.htm> <http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot> <http://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/> <http://www.statisticssolutions.com/autocorrelation/>

[Goldberger, Arthur S.](https://en.wikipedia.org/wiki/Arthur_Goldberger) (1991). ["Multicollinearity".](https://books.google.com/books?id=mHmxNGKRlQsC&amp;pg=PA245) A Course in Econometrics. Cambridge: Harvard University Press. pp. 245–53.

#### K-fold cross validation

https:[//www.r-bloggers.com/evaluating-logistic-regression-models/](http://www.r-bloggers.com/evaluating-logistic-regression-models/)

https:[//www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-](http://www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-) r/

#### Model improvement

https:[//www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-](http://www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-) r/

<http://r-statistics.co/Outlier-Treatment-With-R.html> [http://lojze.lugos.si/~darja/software/r/library/car/html/outlier.test.html](http://lojze.lugos.si/%7Edarja/software/r/library/car/html/outlier.test.html) https://[www.r-bloggers.com/cross-validation-for-predictive-analytics-using-r/](http://www.r-bloggers.com/cross-validation-for-predictive-analytics-using-r/) <http://gerardnico.com/wiki/r/cross_validation>