Understanding Statistical Relationships Between Features in a Kaggle Real Estate Dataset

Mark Torres
May 8, 2019
S&DS 363: Multivariate Statistics for the Social Sciences
Professor Jonathan Reuning-Scherer

Introduction

I am a statistics and data science major, so I've been trying to find ways to practice what I've been learning in class. As a part of that, I've tried my hand at Kaggle competitions, where I can practice the techniques I've been learning in class in a more applied context. To that end, I'm performing my analysis on the Kaggle Ames Housing dataset, which is a part of a beginner-friendly competition on the Kaggle platform. Because of the many features included in the dataset and the (relatively) high number of observations, I thought this dataset would be appropriate for multivariate analysis.

Design and Primary Questions

In my analysis, I will primarily focus on the relationships across the different features of a home and see how they relate to each other. Although I can simply fit a regression to this dataset (and I already have, for the purposes of the competition), I believe that taking the time to analyze the features of the data will lead to a much richer understanding of the nature of the dataset. I am interested in using three multivariate techniques in particular to answer the following questions:

- MANOVA: how are a home's sale price and total surface area affected by the features of the home? Do certain combinations of features relate to differences in the mean sale price and/or the mean surface area?

- Factor Analysis: are there certain combinations of features of a home that indicate a "latent variable" or underlying unifying factor?

- Discriminant Analysis: do homes that sell for "low", "medium", or "high" prices have fundamentally different features, or are they similar in some respects but different in others?

Description of the Data

The data is sourced from an ongoing Kaggle competition (see https://www.kaggle.com/c/house-prices-advanced-regression-techniques) where the goal is to predict the sale price of homes. The competition is beginner-friendly and is designed to introduce people to the basics of machine learning and statistical methods. The data was originally compiled by Dean de Cock as a beginner dataset for data science education. For some features, there were NAs (for example, some homes didn't have pools, and so these were encoded as NAs, not zeroes) that I imputed with zeroes. As per **Table 1**, the majority of my variables were continuous. Many of them, however, were on integer 1-10 scales, which I treated as continuous rather than as discrete 10-level categorical variables.

Variables	Description	Type
OverallQual	Rating of the quality of the home, from 1-10	Continuous
TotalSF	The total surface area of the home, in square ft.	Continuous
logSalePrice	The (log) sale price of the home, in dollars	Continuous
BldgType	The type of building	Categorical
LotArea	The area of the lot of the home	Continuous
GrLivArea	The area of the living room/living area	Continuous
TotRmsAbvGrd	The number of total rooms in the home rated 'above average'	Continuous
BedroomAbvGr	The number of bedrooms in the home rated 'above average	Continuous
salePriceCategory	The category of the home, based off sale price. Either 'low', 'medium', or 'high	Categorical
logTotalSF	The (log) total surface area of the home, in square ft.	Continuous

Table 1. A description of the variables in the dataset, a brief description of the variables, and whether they were considered categorical or continuous

Descriptive Plots and Summary Statistics

Testing for Multivariate Normality

To test for multivariate normality, I used a chi-squared quantile plot (see **Figure 1**). As is evident in the plot, the data veers from multivariate normality in the upper ranges of the plot, which I attribute to some notably excessive homes whose features (i.e. size, price, number of rooms) was much larger than other homes in the dataset. After filtering those homes from the dataset, I see that the data better fits multivariate normality (see **Figure 2**). Because of this, I filter out the outliers, so I reduce the dataset from 1460 observations to 1316 observations. The deviation from multivariate normality is still an issue in the higher quantiles, but I believe that the dataset conforms well to the assumption for practical applications.

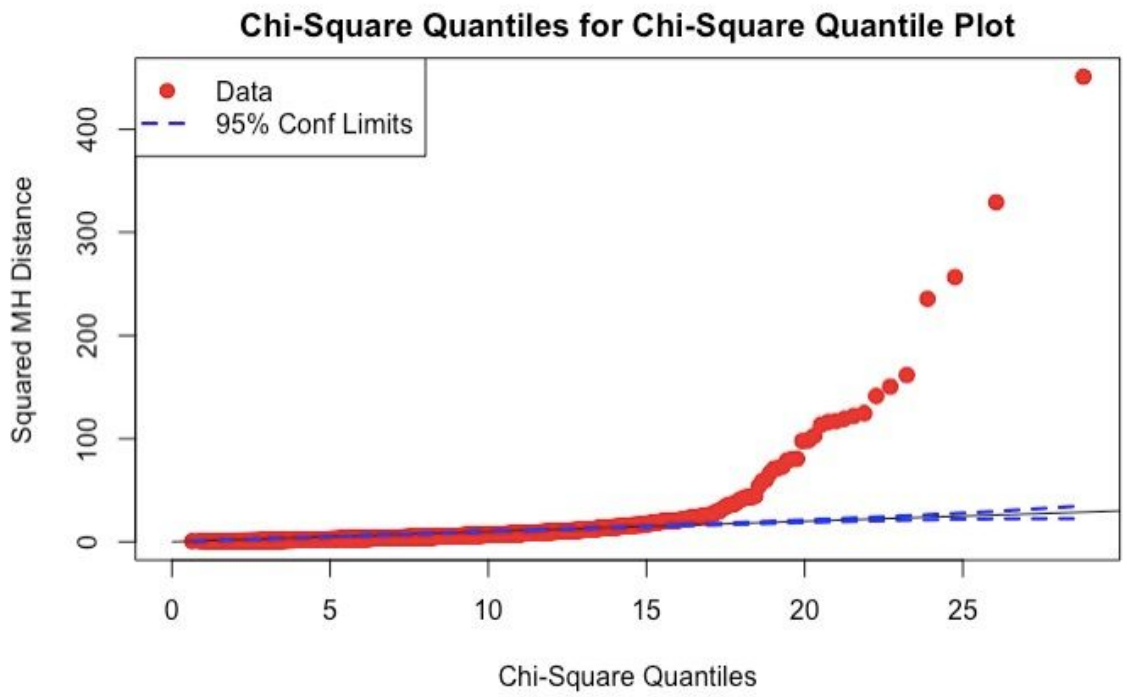


Figure 1. A chi-squared quantile plot of the complete dataset. The observations veer from multivariate normality in the upper quantiles of the graph, which can be attributed to a small number of excessive homes, whose qualities (e.g. size, price, number of rooms) is extravagant compared to other homes.

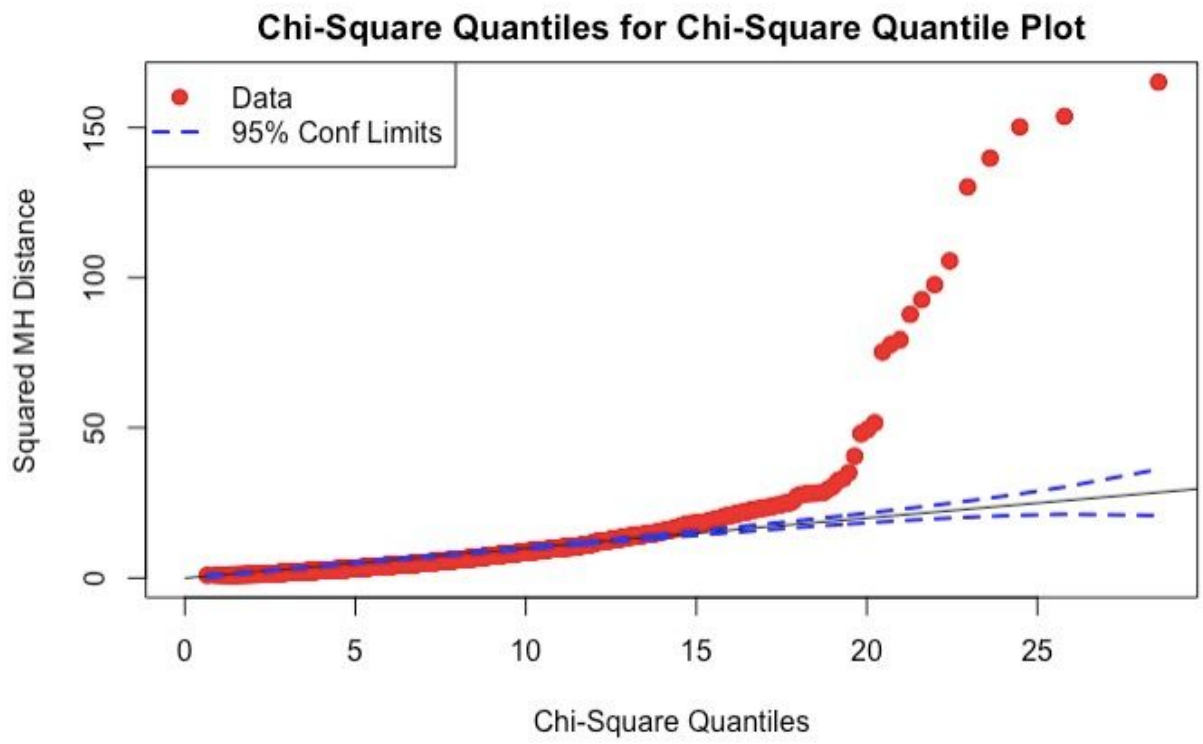


Figure 2. A chi-squared quantile plot of the dataset after removing outliers in the dataset. The observations much better fit the assumptions of multivariate normality, so I'll use this reduced dataset.

Descriptive Statistics of the Numerical Variables

For the numeric variables in my dataset, Table 2 represents a table of summary statistics (mean, variance, minimum, and maximum). I see that some of the features are on a clear 1-10 (or similar range) integer scale, while others are on a truly continuous scale. In this analysis, I treat both these types as numeric continuous variables.

Categories	Mean	Variance	Minimum	Maximum
OverallQual	6.099	1.913	1.000	10.000
TotalSF	1509.619	271611.418	334.000	5642.000
logSalePrice	12.024	0.160	10.460	13.534
LotArea	10516.828	99625649.650	1300.000	215245.000
GrLivArea	1515.464	276129.633	334.000	5642.000
TotRmsAbvGrd	6.518	2.642	2.000	14.000
BedroomAbvGr	2.866	0.665	0.000	8.000
logTotalSF	7.264	0.111	5.811	8.638

Table 2. A table of the summary statistics for the numerical variables in the dataset.

Correlation Matrix of the Numerical Variables

For the numerical variables in my dataset, I see from **Figure 2** that the values aren't on the same scale (for example, OverallQual varies from 1 to 10 while LotArea varies from 1300 to 215,000). Because of this, I felt that doing a covariance matrix wouldn't illustrate a proper comparison of the spread across variables. Because of that, I plotted the correlation matrix (see **Figure 3**) of the different features in the dataset. From the correlation matrix, I see that the features in the dataset have high positive correlations. The high magnitude of the correlations is promising for factor analysis since it suggests that there could be a feature that underlies these variables.

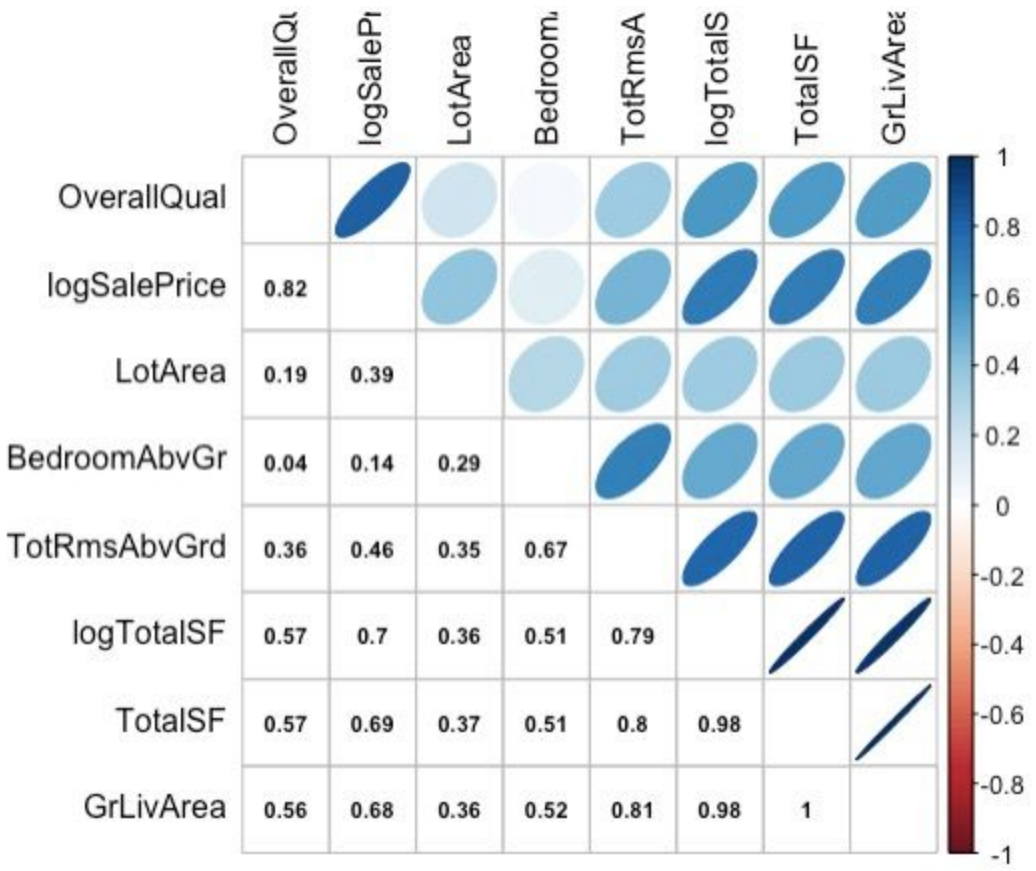
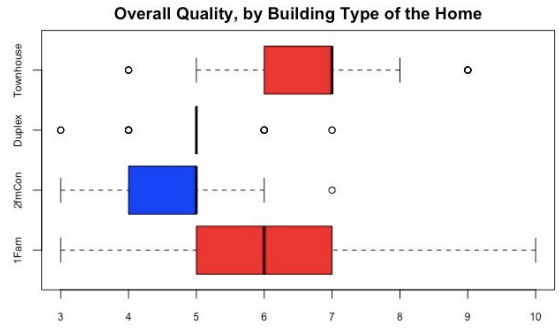
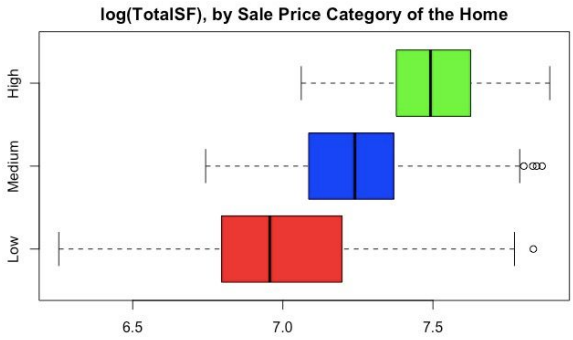


Figure 3. A correlation matrix of the numerical variables in the dataset. I see relatively strong positive correlations across the board, which suggests that these variables are strong candidates for factor analysis. I see what is likely a case of multicollinearity between TotalSF and GrLivArea. Because of this, I don't use both features jointly in any given analysis, but I keep them both because I used them separately for different purposes.

Box Plots of the Categorical Variables

I used a series of boxplots to analyze the relationship between categorical variables and a few of the numerical variables. The boxplots (see **Figure 4**) illustrate that there are some clear differences in the values of the numeric variables (e.g. price and size) based on the levels of the categorical variables. This is promising for both MANOVA and discriminant analysis, since it suggests that the homes in the dataset can be segregated into distinct groups, and that the traits of these homes vary depending on which group they fall under.





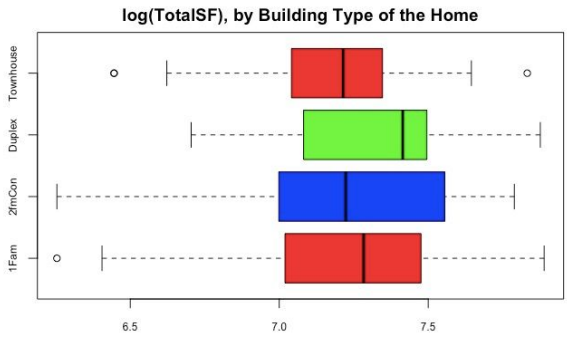


Figure 4. (*left*) Comparing SalePriceCategory (the category of the sale price) to the overall quality of the home (*top left*) and to the (log) total surface area of the home (*bottom left*). I also compared (*right*) BldgType (the type of building that the home is in) to the overall quality of the home (*top right*) and to the (log) total surface area of the home (*bottom right*). I see that the values of OverallQual and logTotalSF clearly vary across the different levels of the homes. This suggests both that the MANOVA should reveal differences in group means and that the discriminant analysis should reveal a distinct separation of groups.

Multivariate Analysis

Method 1: MANOVA

Main Question to Answer: *How do the multivariate means of the dataset vary by the groups I divide the observations into?*

Translation: *Do different prices of homes (whether the home price was low, medium, or high) or the type of building the home is in (e.g. whether it was a duplex or apartment) affect the multivariate means.*

For my first statistical analysis, I would like to use MANOVA to examine the relationships between variables. The variables that I will be using are the following:

- For my two categorical predictors, I will use the category of the home's price (low, medium, high) and the type of building the home is in (1-family home, 2-family home, duplex, or townhouse).
- For my continuous predictor, I will use the overall quality of the home (on a scale from 1 to 10).

- For my two continuous response variables, I will use the total surface area of a home and the size of the lot of the home.

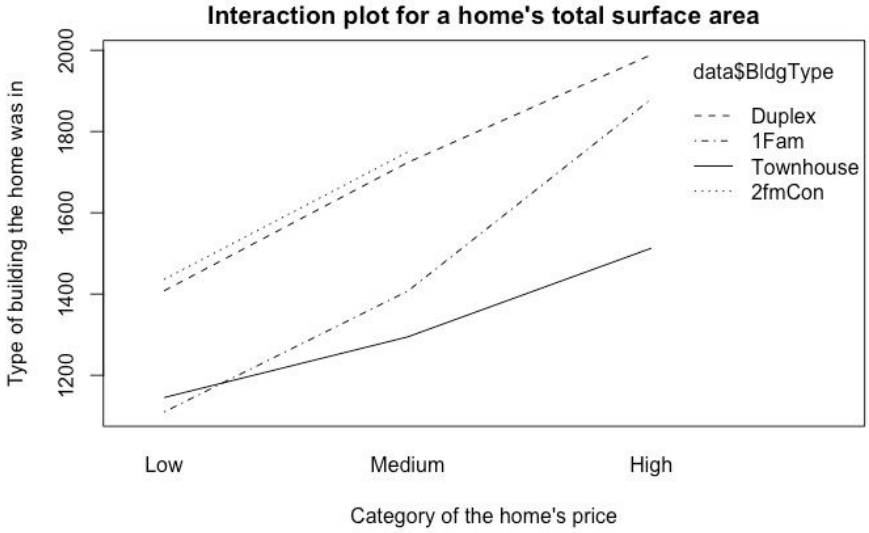
Before running any analysis, I would expect that the continuous response variables would vary based on the category of the home's price and the building type. For example, an expensive home is probably larger, so I predict that there would be a relationship between a home's price category and the size of the home.

Interaction Plots

I begin the analysis by running an interaction plot for each of the response variables (see **Figure 5**). The interaction plots are designed to illustrate how the values of the response variable (in these cases, a home's total surface area and a home's lot area) vary depending on the groups the home is classified under. In a case where there is no interaction between the levels of both the categorical predictors, I'd expect no intersection of the lines. For example, if there were no interaction between building type and home price, then I'd expect that if a home typically has a larger surface area if it's more expensive, the degree to which this occurs shouldn't be affected by the type of building the home is in.

From the analysis, I see that there is a slight interaction between a home's price category and the type of building when analyzing a home's total surface area. For 1-family homes and 2-family condos, I see a slight interaction effect. This tells us that typically, 1-family homes are larger than 2-family homes, unless the home was relatively cheap, in which case 2-family homes are larger than 1-family homes.

I see a stronger interaction effect between a home's price category and the type of building when analyzing a home's lot area. I observe that 1-family homes and 2-family condos are typically larger than duplexes, unless the home is relatively cheap, in which case duplexes are larger than both 1-family homes and 2-family condos.



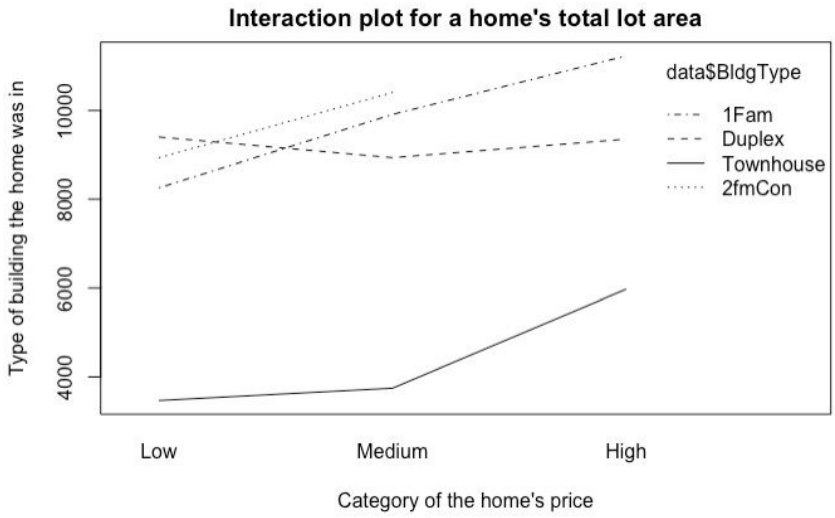


Figure 5. Interaction plots for the features. I observe that in the case of a home's total surface area, there is a slight interaction effect (*top*). From this, I see that 1-family homes are typically larger than 2-family homes, unless the home is relatively cheap, in which case 2-family homes are typically larger than 1-family homes. I observe a stronger interaction effect in the case of examining a home's total lot area. I see this most strongly in the case of 1-family, 2-family condo, and duplex homes. The plot tells us that 1-family homes and 2-family condos are typically larger than duplexes, unless the home is relatively cheap, in which case duplexes are larger than both 1-family homes and 2-family condos.

Running MANOVA

To continue the analysis, I can run a MANOVA. The results are illustrated below in **Figure 6**. In the univariate case, salePriceCategory and BldgType were significant, and the interaction between the two was significant as well, for both TotalSF and LotArea. This tells us that both the total surface area of a home and the lot area of a home were different depending on the levels of the home's price and of the home's building type, and that there was an interaction effect between a home's price and its building type (which I observed from the interaction plots). In the multivariate case, using both Wilks lambda (the typical test statistic) and the Pillai statistic (which is used in cases where the residuals are non-normal), I see a significant effect of the two categorical predictors and the interaction between them. I used the Pillai statistic in addition to Wilks lambda to verify that the test would still yield significance even if the residuals weren't assumed to be normal. The findings of significance observed in the MANOVA verify the trends from the interaction plots.

```
 Response TotalSF :
                                   Df    Sum Sq  Mean Sq  F value    Pr(>F)
data$salePriceCategory              2 103800272 51900136 535.3938 < 2.2e-16 ***
data$BldgType                       3  10184481  3394827  35.0205 < 2.2e-16 ***
data$salePriceCategory:data$BldgType    5   2929187   585837   6.0434 1.539e-05 ***
Residuals                        1305 126504400    96938
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response LotArea :
                                   Df     Sum Sq    Mean Sq  F value    Pr(>F)
data$salePriceCategory              2 1462534110  731267055 149.3602 < 2.2e-16 ***
data$BldgType                       3 3866369054 1288789685 263.2334 < 2.2e-16 ***
data$salePriceCategory:data$BldgType    5   91293626   18258725   3.7293  0.002345 **
Residuals                        1305 6389274434    4895996
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                                   Df   Wilks approx F num Df den Df    Pr(>F)
data$salePriceCategory              2 0.50881  262.054      4   2608 < 2.2e-16 ***
data$BldgType                       3 0.59248  130.037      6   2608 < 2.2e-16 ***
data$salePriceCategory:data$BldgType    5 0.96315    4.943     10   2608 4.008e-07 ***
Residuals                        1305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

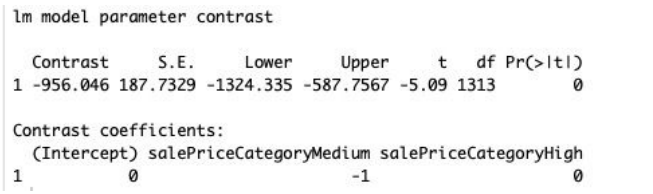
```
                                   Df  Pillai approx F num Df den Df    Pr(>F)
data$salePriceCategory              2 0.49120  212.426      4   2610 < 2.2e-16 ***
data$BldgType                       3 0.42350  116.856      6   2610 < 2.2e-16 ***
data$salePriceCategory:data$BldgType    5 0.03717    4.943     10   2610 3.997e-07 ***
Residuals                        1305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. Results of univariate (*top*) MANOVA and multivariate MANOVA, using the Wilks (*middle*) and Pillai (*bottom*) statistics. In both the univariate and multivariate cases, the predictors were significant and the interaction between them was significant. This tells us that group means were statistically significantly different from each other among the levels of the groups. I used both the Wilks and Pillai statistics to show that the test would still be robust even if normality of residuals wasn't assumed.

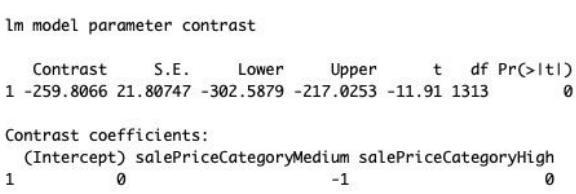
Fitting univariate contrasts

To better understand the differences in means when comparing levels of factors and combinations of factors, I'll use univariate contrasts (multivariate contrasts aren't possible in R). The MANOVA allowed us to see that the group means were different, but univariate contrasts allow us to see which specific group means are

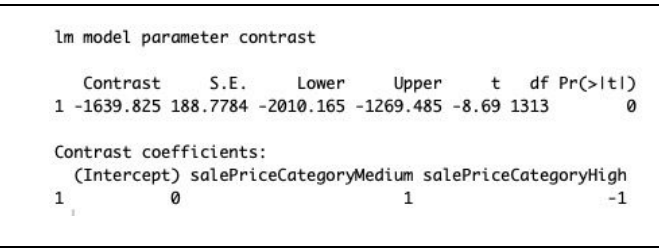
different. In the analysis, I focus on looking at the groups of home price ("low", "medium", "high") and see how a home's total surface area and lot area vary across the levels. The results are outlined in **Figure 7** (*below*). What I observe here is that for each pairwise comparison, the contrasts reveal a statistically significant difference. In particular, the high *t*- statistics observed from the univariate contrasts suggest that the group means vary greatly. This is verified by the table of group means, which suggests a noticeable difference in the total surface area and lot area across group levels.



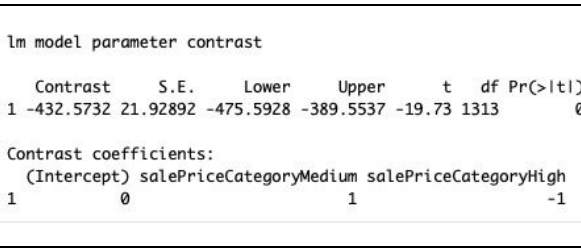
```
lm model parameter contrast

  Contrast     S.E.     Lower     Upper     t   df Pr(>|t|)
1 -956.046 187.7329 -1324.335 -587.7567 -5.09 1313        0

Contrast coefficients:
  (Intercept) salePriceCategoryMedium salePriceCategoryHigh
1           0                      -1                     0
```



```
lm model parameter contrast

   Contrast     S.E.     Lower     Upper      t   df Pr(>|t|)
1 -259.8066 21.80747 -302.5879 -217.0253 -11.91 1313        0

Contrast coefficients:
  (Intercept) salePriceCategoryMedium salePriceCategoryHigh
1           0                      -1                     0
```



```
lm model parameter contrast

   Contrast     S.E.     Lower     Upper     t   df Pr(>|t|)
1 -1639.825 188.7784 -2010.165 -1269.485 -8.69 1313        0

Contrast coefficients:
  (Intercept) salePriceCategoryMedium salePriceCategoryHigh
1           0                       1                    -1
```



```
lm model parameter contrast

   Contrast     S.E.     Lower     Upper      t   df Pr(>|t|)
1 -432.5732 21.92892 -475.5928 -389.5537 -19.73 1313        0

Contrast coefficients:
  (Intercept) salePriceCategoryMedium salePriceCategoryHigh
1           0                       1                    -1
```

salePriceCategory <fctr>	LotArea <dbl>
Low	8052.751
Medium	9008.797
High	10648.621

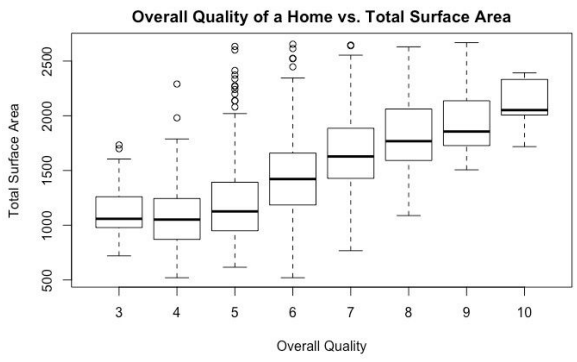
salePriceCategory <fctr>	TotalSF <dbl>
Low	1148.061
Medium	1407.867
High	1840.440

Figure 7. A summary of the univariate contrasts for salePriceCategory (the category of the home price, with levels = "low", "medium", "high") with respect to a home's total surface area (TotalSF) and a home's lot area (LotArea). I found significant differences between low-price homes and medium-price homes with respect to a home's lot area (*top left*) and a home's total surface area(*top right*). I also found significant differences between medium-price and high-price homes with respect to a home's total lot area (*middle left*) and a home's total surface area (*middle right*). The high t-statistics suggests that the difference between groups is strong, which is verified by the table of group means with respect to total lot area (*bottom left*) and surface area (*bottom right*).

Checking for relationships between predictors and response variables

Before fitting a generalized linear model, I'll fit some plots between the categorical predictors and the response variables to see if there's a linear relationship between them. Because I already established that there are group differences in the response variables based on the price of the home and the building type of the home (see above), I'll examine the relationship between my continuous predictor, the overall quality of a home, and my response variables. The analysis (see **Figure 8**) seems to suggest that there is some mild linear relationships,

where homes deemed to be of higher quality have more total surface area and larger lots, but a linear model would be better able to tease out the strength of the relationship.



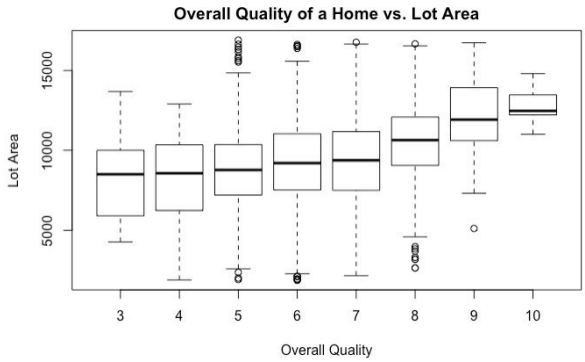


Figure 8. Plotting the relationship between the overall quality of a home vs. total surface area (*left*) and lot area (*right*). I use a boxplot instead of a scatter plot because the overall quality rating of a home could be considered either a continuous variable or a categorical variable (since it is on a 1-10 scale), and the scatter plot wasn't very illustrative compared to the boxplot.

Fitting a generalized linear model

Now, I can fit a generalized linear model. The results (see **Figure 9** below) suggest that my covariate, a home's overall quality rating, is significantly related with a home's total surface area, but isn't significantly related with a home's lot area, after accounting for a home's price and building type. This seems to suggest that, for two homes that have the same price and same building type, knowing that a home has a high overall quality rating would suggest that it has a high total surface area (so, larger homes seem more appealing), but doesn't tell you anything about the lot area (so, you don't know how big the property is, but you know the interior of the home is large and spacious).

```
Coefficients: (1 not defined because of singularities)
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                            721.159     52.455  13.748  < 2e-16 ***
data$salePriceCategoryMedium                           222.938     24.529   9.089  < 2e-16 ***
data$salePriceCategoryHigh                             581.015     33.240  17.479  < 2e-16 ***
data$BldgType2fmCon                                    333.586     69.999   4.766 2.09e-06 ***
data$BldgTypeDuplex                                    317.265     57.965   5.473 5.29e-08 ***
data$BldgTypeTownhouse                                  -8.502     60.056  -0.142   0.8874
data$OverallQual                                        78.640     10.096   7.790 1.36e-14 ***
data$salePriceCategoryMedium:data$BldgType2fmCon        52.402    143.586   0.365   0.7152
data$salePriceCategoryHigh:data$BldgType2fmCon             NA         NA      NA       NA
data$salePriceCategoryMedium:data$BldgTypeDuplex        47.425     93.616   0.507   0.6125
data$salePriceCategoryHigh:data$BldgTypeDuplex        -102.280    223.782  -0.457   0.6477
data$salePriceCategoryMedium:data$BldgTypeTownhouse   -155.756     72.292  -2.155   0.0314 *
data$salePriceCategoryHigh:data$BldgTypeTownhouse     -380.775     76.466  -4.980 7.22e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 304.5 on 1304 degrees of freedom
Multiple R-squared:  0.5034,    Adjusted R-squared:  0.4992
F-statistic: 120.2 on 11 and 1304 DF,  p-value: < 2.2e-16
```

```
Coefficients: (1 not defined because of singularities)
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           8225.343    381.361  21.568  < 2e-16 ***
data$salePriceCategoryMedium                          1653.111    178.331   9.270  < 2e-16 ***
data$salePriceCategoryHigh                            2966.784    241.665  12.276  < 2e-16 ***
data$BldgType2fmCon                                    678.420    508.910   1.333  0.18274
data$BldgTypeDuplex                                   1149.401    421.419   2.727  0.00647 **
data$BldgTypeTownhouse                               -4789.993    436.621 -10.971  < 2e-16 ***
data$OverallQual                                        5.657     73.398   0.077  0.93858
data$salePriceCategoryMedium:data$BldgType2fmCon     -179.712   1043.901  -0.172  0.86334
data$salePriceCategoryHigh:data$BldgType2fmCon           NA         NA      NA       NA
data$salePriceCategoryMedium:data$BldgTypeDuplex    -2123.824    680.608  -3.120  0.00185 **
data$salePriceCategoryHigh:data$BldgTypeDuplex      -3023.972   1626.944  -1.859  0.06330 .
data$salePriceCategoryMedium:data$BldgTypeTownhouse -1381.277    525.581  -2.628  0.00869 **
data$salePriceCategoryHigh:data$BldgTypeTownhouse    -471.822    555.928  -0.849  0.39620
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2214 on 1304 degrees of freedom
Multiple R-squared:  0.459,     Adjusted R-squared:  0.4544
F-statistic: 100.6 on 11 and 1304 DF,  p-value: < 2.2e-16
```

Figure 9. The results of the generalized linear model, with respect to total surface area (*left*) and lot area (*right)*. The continuous covariate, OverallQual (the quality of the home) is significantly related to total surface area, but not to lot area, after controlling for the categorical predictors. I don't analyze the effects of the categorical predictors since I already analyzed those in previous tests.

Checking chi-squared quantile plot of residuals

To check model assumptions, I also created a chi-squared quantile plot of the residuals. Although there is some deviation away from normality in the upper quantiles, for the most part the chi-squared quantile plot shows that the residuals conform to the assumption of multivariate normality (see **Figure 10**). Because of this, there is no need to modify the model.

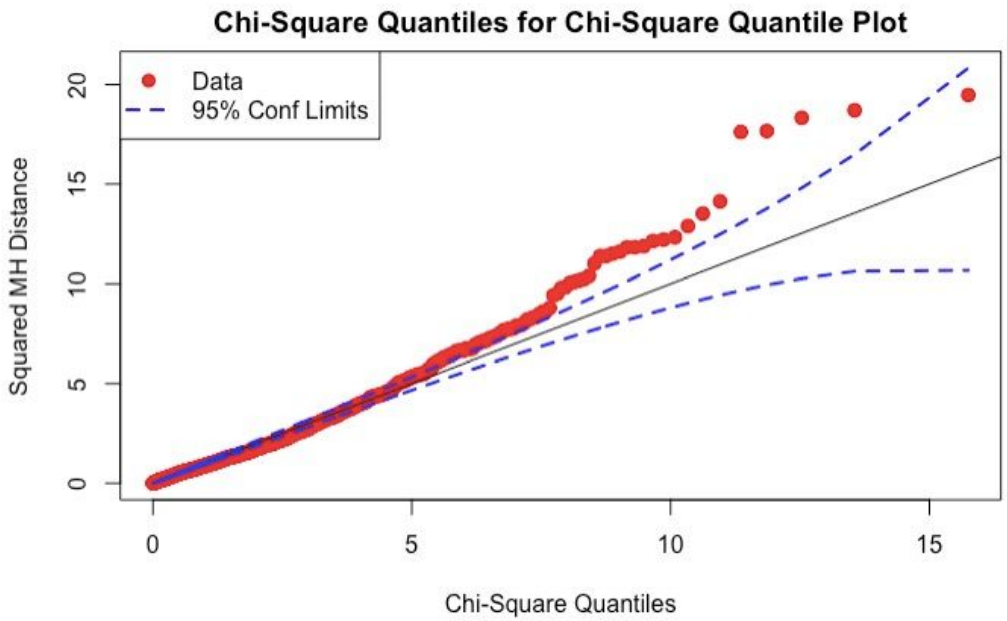


Figure 10. A chi-square quantile plot of the residuals, which suggests that the model fits the assumptions.

Method 2: Factor Analysis

Main Question to Answer: *are there certain combinations of features of a home that indicate a "latent variable" or underlying unifying factor?*

For my second statistical analysis, I would like to use factor analysis to examine the relationships between variables and see if there is an underlying explanatory variable that relates these features. In this section, I will use the following features and see if there could be an underlying "latent variable" that connects them:

- logSalePrice: the (log) sale price of the home
- TotRmsAbvGrd: the number of rooms rated to be above average (all rooms, such as bedrooms, bathrooms, study, etc.)
- BedroomAbvGrd: the number of bedrooms rated to be above average (there is a strong correlation between this feature and TotRmsAbvGrd of r = 0.67, but there likely isn't multicollinearity between these features, meaning that I can use them both in this analysis).
- TotalSF: the total surface area of the home

Relationships across the variables, calculating KMO, and determining suitability for factor analysis

I created a correlation matrix to analyze the relationships between variables (see **Figure 11**). The correlation matrix shows that the variables are strongly correlated to each other, which suggests that they could be strong candidates for factor analysis. To verify this, I calculate a KMO score as well (see **Figure 11**). The KMO score of 0.66 is ok, and I'll take this score as an indication that the data is appropriate for factor analysis.

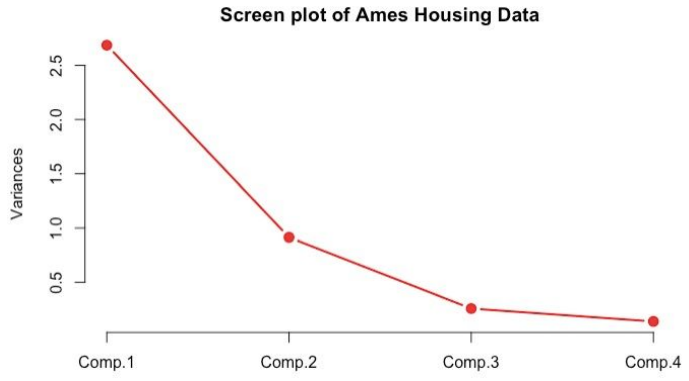
	TotalSF	logSalePrice	TotRmsAbvGrd	BedroomAbvGr
TotalSF	1.00000	0.69414	0.80140	0.51205
logSalePrice	0.69414	1.00000	0.46359	0.13861
TotRmsAbvGrd	0.80140	0.46359	1.00000	0.67118
BedroomAbvGr	0.51205	0.13861	0.67118	1.00000

```
$KMO
[1] 0.66188

$MSA
                    MSA
TotalSF       0.64888
logSalePrice  0.60676
TotRmsAbvGrd  0.69333
BedroomAbvGr  0.69364
```

Figure 11. Determining suitability for factor analysis, by using a correlation matrix (*left*) and calculating a KMO score (*right*). Both of these methods show that the data is appropriate for factor analysis.

Determining the number of latent factors

I determined the number of latent factors using both the scree plot and the" eigenvalue > 1"methods (see **Figure 12**). Based off the scree plot, I could use either 1 or 2 latent factors, so I use the "eigenvalue > 1" method as a "tiebreaker" and thus I'll use only 1 latent factor. From this, it seems like the features that I chose are connected through 1 underlying "latent" variable. This could make sense, since it seems like homes which have certain qualities often have other distinct qualities as well (i.e. a large home could also have a lot of bedrooms), so it seems almost as if the qualities of a home could be determined by the "type" of home it is.




```
Comp.1 Comp.2 Comp.3 Comp.4
  2.69   0.91   0.26   0.14
```

Figure 12. Determining the number of latent factors using the scree plot (*left*) and "eigenvalue > 1" (*right*) methods. Based on the scree plot, either 1 or 2 latent factors is appropriate, but based on the "eigenvalue > 1" method, only 1 latent factor is appropriate. Based on this logic, I'll use only 1 latent factor.

Perform factor analysis and determine an extraction method

I tried the principal axis factoring (PAF) and iterative PCA extraction methods, and compare the two using RMSR (root-mean square residual). I would like to reduce the value of the RMSR, so I would choose the method with the lower value. Based off the RMSR values (see **Figure 13**), it seems like both PAF and iterative PCA do equally well in describing my data (I also checked the number of residuals greater than 0.05, and it was the same for both methods). I had this result as well in my problem set and I verified that the PAF and iterative PCA methods are indeed different. Based on the results, I will use the PAF method for further analysis, knowing that both methods would give similar results.

Method <fctr>	RMSR <dbl>
Principal Axis Factoring	0.122
Iterative PCA	0.122

Figure 13. Comparing the principal axis factoring (PAF) and iterative PCA methods using RMSR. The identical scores indicate that both extraction methods work equally well.

Using a varimax rotation and interpreting the final factors

I used principal axis factoring with varimax rotation so that each indicator has a high loading on only one factor (in this scenario, since I only have one factor, it doesn't make a difference, but in the case of multiple factors a varimax rotation would be appropriate for maximizing loading in one direction). From the loading plot after applying the varimax rotation (see **Figure 14**), I see particularly high loadings for TotalSF and TotRmsAbvGrd, which suggests that the indicators are related by an underlying factor about the size of the home: larger homes have larger surface areas, more rooms, more bedrooms, and are more expensive.

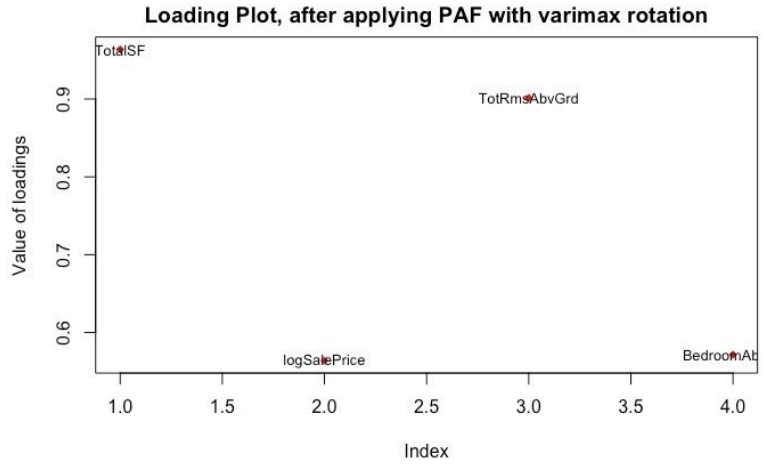


Figure 14. Loading plots after applying principal axis factoring with a varimax rotation. The plot is 2-dimensional, but since there is only 1 latent factor, the index of the variables (an arbitrary measure) is used as the x-axis, while the value of the loading is plotted on the y-axis. TotalSF and TotRmsAbvGrd have particularly high loadings, while logSalePrice and BedroomsAbvGr have very low loadings, which suggests that the variables are connected by some latent factor regarding the size of the home.

Method 3: Discriminant Analysis

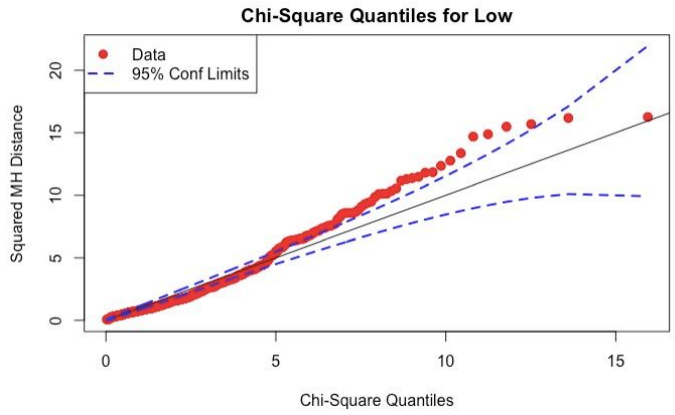
Main Question to Answer: *do homes that sell for "low", "medium", or "high" prices have fundamentally different features, or are they similar in some respects but different in others?*

For my third statistical analysis, I would like to see whether I can discriminate between homes that sell for either "low", "medium", or "high" prices using the following features:

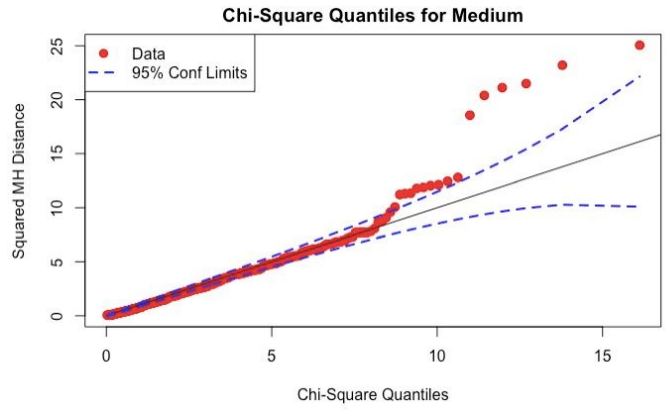
- TotalSF: the total surface area of the home
- LotArea: the area of the lot that the home is found in
- TotRmsAbvGrd: the number of rooms rated to be above average (all rooms, such as bedrooms, bathrooms, study, etc.)

Evaluating assumptions for discriminant analysis

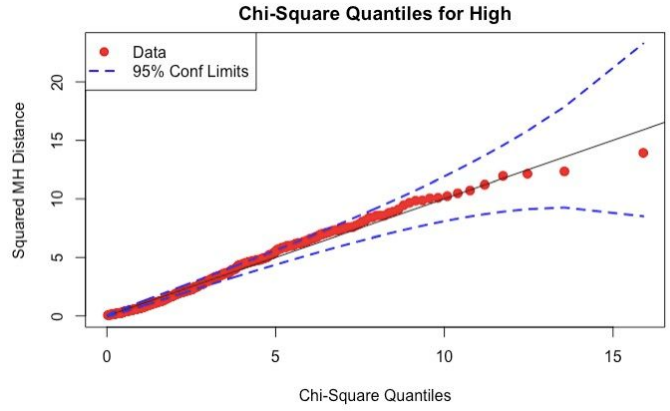
To evaluate the assumptions for discriminant analysis, I created chi-squared quantile plots and examined the covariance matrices for each price level (see **Figure 15**) . From this, I observe that the assumptions of multivariate normality and similarity are both met.



	TotalSF	TotRmsAbvGrd	LotArea
TotalSF	118636.10	393.5583	88240.65
TotRmsAbvGrd	393.56	1.9583	576.58
LotArea	88240.65	576.5802	6194407.75



	TotalSF	TotRmsAbvGrd	LotArea
TotalSF	88475.12	284.3875	77831.89
TotRmsAbvGrd	284.39	1.5923	889.16
LotArea	77831.89	889.1620	8989018.51



	TotalSF	TotRmsAbvGrd	LotArea
TotalSF	113633.44	319.5205	379239.8
TotRmsAbvGrd	319.52	1.7261	1238.7
LotArea	379239.84	1238.6745	8369565.6

Figure 15. Chi-squared quantile plots for "low" (*top left*), "medium" (*middle left*), and "high" (*bottom left)* homes. Although there is some deviation from normality, the data seems to fit closely enough for practical applications, and so the assumption of multivariate normality is met. The covariance matrices for "low" (*top right)*, "medium" (*middle right*), and "high" (*bottom right*) follow the rule of thumb established in class, which is that since there are no entries in one covariance matrix that is more than four times that of the corresponding entry in another covariance matrix, the similarity assumption is met.

Performing stepwise discriminant analysis

Now, I performed stepwise discriminant analysis in order to determine the best model. From the stepwise regression, I only have one significant variable in the model (TotalSF, the total surface area of the home). Because of this, I'll force in LotArea (the home's lot area) as a non-significant variable. I chose to force in LotArea rather than TotRmsAbvGrd because LotArea was less correlated to TotalSF than TotRmsAbvGrd was to TotalSF, so I thought LotArea would be able to capture differences between the levels of home prices that weren't captured by TotalSF.

Demonstrating whether group means are different

From the MANOVA analysis above, I've already established that the group means are different, so I won't redo that analysis in this section.

How many discriminating functions are significant? What is their relative discriminating power?

From my analysis (see **Figure 16**), I found that only one discriminating function was significant. Based off the trace, the first discriminating function explained (just about) 100% of the between-class variance.

```
Prior probabilities of groups:
    Low  Medium    High
0.32599 0.35486 0.31915

Group means:
        TotalSF TotRmsAbvGrd   LotArea
Low    -0.72789    -0.468754 -0.389665
Medium -0.12403    -0.079438 -0.070639
High    0.88139     0.567126  0.476560

Coefficients of linear discriminants:
                  LD1      LD2
TotalSF       1.69739 -0.34834
TotRmsAbvGrd -0.62600 -0.33420
LotArea       0.30197  1.04555

Proportion of trace:
LD1 LD2
  1   0
```

Figure 16. Analyzing the discriminating functions. From the output, it's clear that only one discriminating function is significant and it accounts for (rounding up) 100% of the between-class variance. I hesitate to say 100% because I would be surprised the discriminating function accounts for *all* the variance, but it seems to account for almost all the variance.

What is the discriminating ability of each function?

I evaluate the discriminating ability of the discriminating function using both regular and LOOCV classification, as can be seen in **Figure 17**. From this analysis, I see that the model is relatively accurate, with a 64% accuracy rate for both regular classification and LOOCV classification.

Method <fctr>	Accuracy <dbl>
Regular	0.64
LOOCV	0.64

Figure 17. Evaluating the discriminating ability of the discriminating function, using both regular and LOOCV classification.

Which of the original variables are the "best" discriminators?

From looking at the coefficients of the first linear discriminant in **Figure 16**, I see that the magnitude of the TotalSF coefficient is larger than the other two features. Since I observed that TotalSF was the only significant feature, it makes sense that TotalSF is the "best" discriminator out of my original variables.

Plotting and interpreting the discriminant analysis (DA) function scores

I now analyze the plot of the DA scores. From this plot (see **Figure 17**), I see that there is some clear differentiation between low, medium, and high priced homes, but the difference in DA scores isn't as clear-cut as, say, the Iris dataset. From this, it seems like I was able to successfully segregate the low, medium, and high priced homes based on several qualities of the homes.

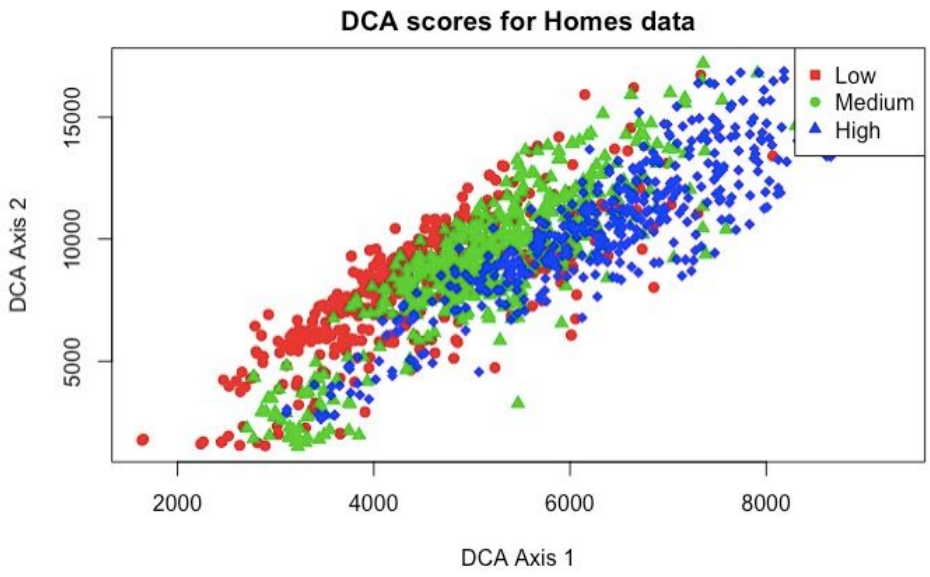


Figure 17. Plot of DA scores for the dataset. From this plot, the low, medium, and high priced homes are clearly separated, even though the segregation isn't perfect.

Conclusion and Discussion

I was interested in better understanding the Kaggle Ames Housing dataset through the use of multivariate statistics. I included the discussion and interpretation of my results throughout my analysis, so I will focus this section on thoughts and insights, while also briefly touching on some points for further analysis.

Each of the methods provided different insights into the dataset:

MANOVA:

The MANOVA analysis demonstrated that the different types of homes, whether they were separated based on their price ("low", "medium", "high") or building type ("1-family home", "2-family condo", "townhouse", "duplex") had distinctly different qualities

Factor Analysis

The factor analysis showed that there was an underlying latent factor (namely, something along the lines of the size of the home) that related home size, number of rooms, home price, and lot area. It was interesting to see, through factor analysis, that there seemed to be an underlying explanation that connected these seemingly correlated, but distinctly separated, variables.

Discriminant Analysis

Discriminant Analysis revealed that when the homes are grouped based on price ("low", "medium", and "high"), they can be reliably separated into distinct groups using a combination of features present in the dataset. Namely, understanding the total surface area of a home was very important in discriminating the price category of the home. Logically, this makes sense, since larger homes are likely more expensive, so by knowing the size of the home, one can likely estimate how much it costs.

It was interesting to take the time to analyze the Kaggle Ames Housing dataset, rather than just fitting a machine learning model. I can better explain why certain trends appear in the data and come up with future questions to explore based on relationships I've observed. I was able to notice connections in the dataset that have real-life applications (for example, knowing the size of the home is often enough to determine the price of the home) that wouldn't have necessarily appeared had I only fitted a predictive model. I felt that I got a much richer understanding of the intricacies of the data and how the different features relate to each other after having analyzed the data using multivariate statistics.