Final Housing

Drew Millane, Sam Spackman, and Naomi Zubeldia

2023-04-18

Code is in the appendix below

Section 0: Executive Summary

Housing prices are determined through an appraisal of each home's features. In this analysis, we will be investigating how well and what specific features influence pricing. Also, specifically, if house size will influence price increase. We used a model that accounted for unequal variability and spatial correlation. We found that the house size above ground does influence the variability of price increases. Other variables that influence the price as well are central air, garage size, and build/remodel dates. Also, the model that we created was able to accurately fit the data and predict new observations.

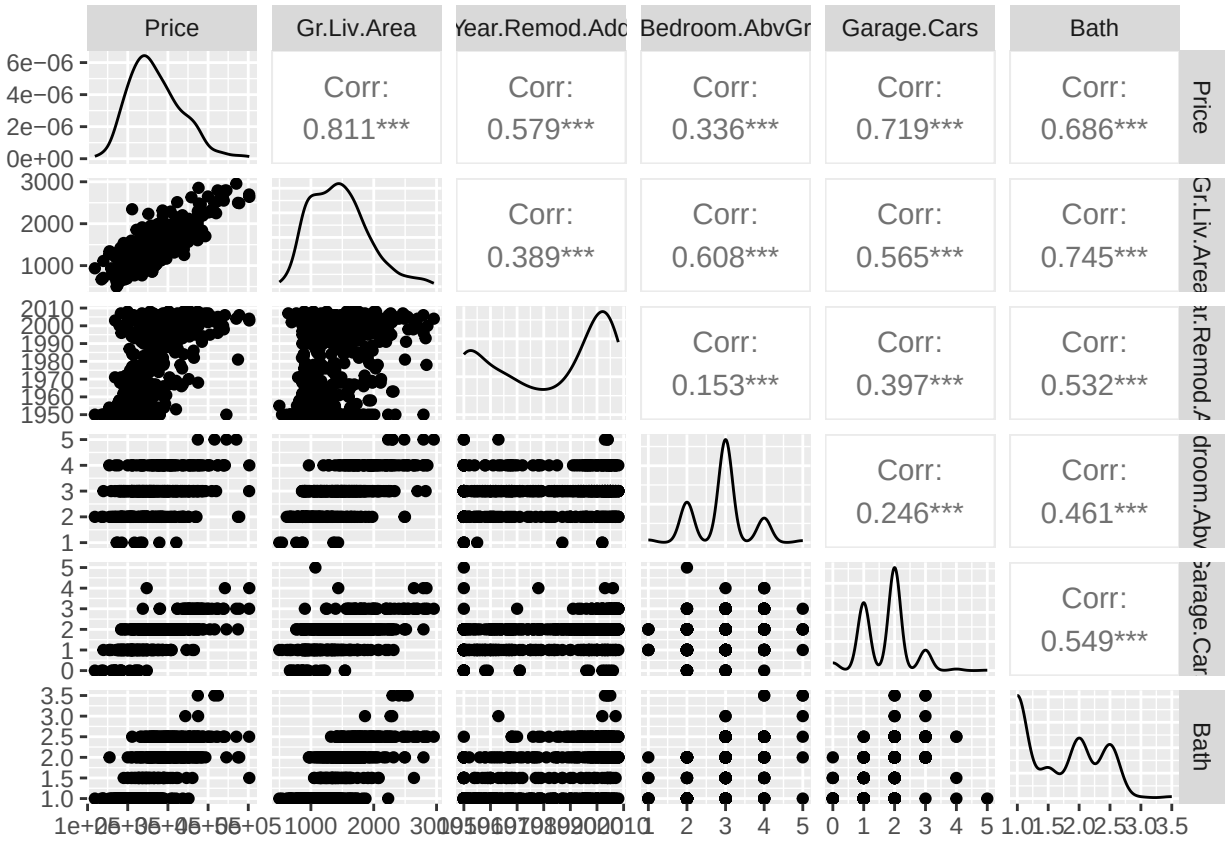
Section 1: Introduction and Problem Background

When taking out a loan to buy a house, you need to get an appraisal of the home to determine its value for collateral against your loan. The appraisal is based on many different factors of the inside and outside conditions of the house, lot size, and any improvements/additions made to it. The main goal of this analysis is to understand what factors in the appraisal are most influential on the sale price of the home. We want to know how well they explain the pricing, and if any specifically, increase the pricing as well. The variability of the sale price is also a factor of interest that we want to test if it increases with the size of the home. Lastly, we want to use the model we created to answer these questions to predict the sale prices of homes.

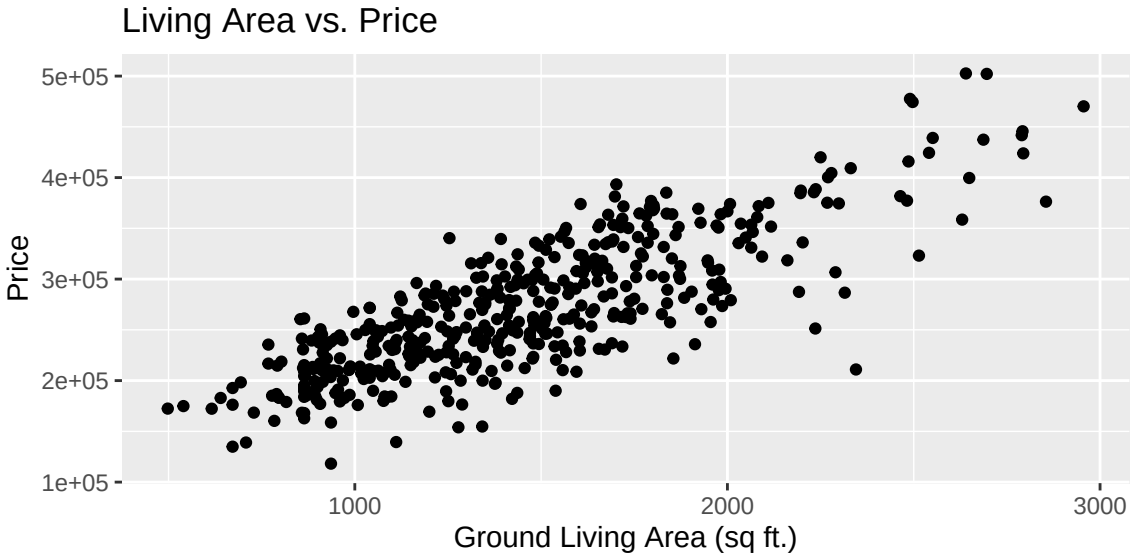
Through the analysis, we will be using a data set of homes from Ames, Iowa that has 469 observations with descriptive variables of location coordinates, ground sq. footage, style, remodeling dates, if there is central air, number of full and half bathrooms, number of bedrooms, and the size of garage.

It appears the variable of pricing has some missing values, which we can use for making predictions, and a decent range of about $400,000 of prices between the cheapest and most expensive. The above-ground square footage has a very large range of 498 ft to 2956 ft. There are 4 styles of houses we are observing with most of the remodeling having been done in the early 2000s to the later 1990s. It looks like most of the observations have a central area and an average of about 3 bedrooms and 1.5 baths. All the houses seem to have kitchens with maybe a couple of exceptions of either none or two. Lastly, the houses have an average garage size to fit about 1.5 cars with some extra room.

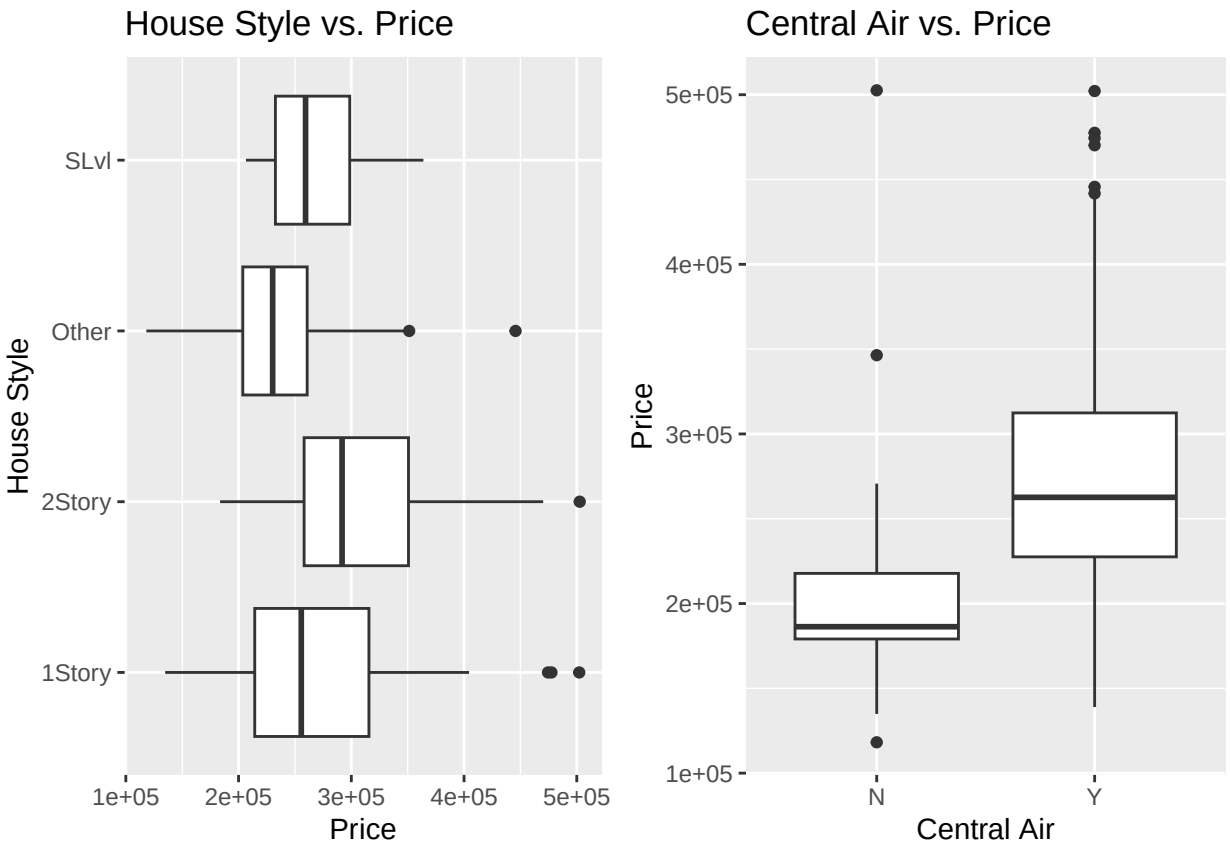
In the pairs plot below, we observe the price compared to the other numerical variables influencing the prices. It looks like Pricing has the strongest linear relationship with the above-ground square footage of the house, which brings interest to our question about if the increase of square footage increases the price variability since the correlation is very positively high too. There appear to be loose linear relationships with Price and number of bedrooms, number of bathrooms, and garage size.



Since the relationship between square footage above ground and price looked the most linear, we created the scatterplot below for the two variables. The plot appears to support the assumption of the linear relationship, but it has a cone-like spread to the points. This indicates that there may be a problem with heteroskedasticity in this data we will need to address when creating an appropriate model.



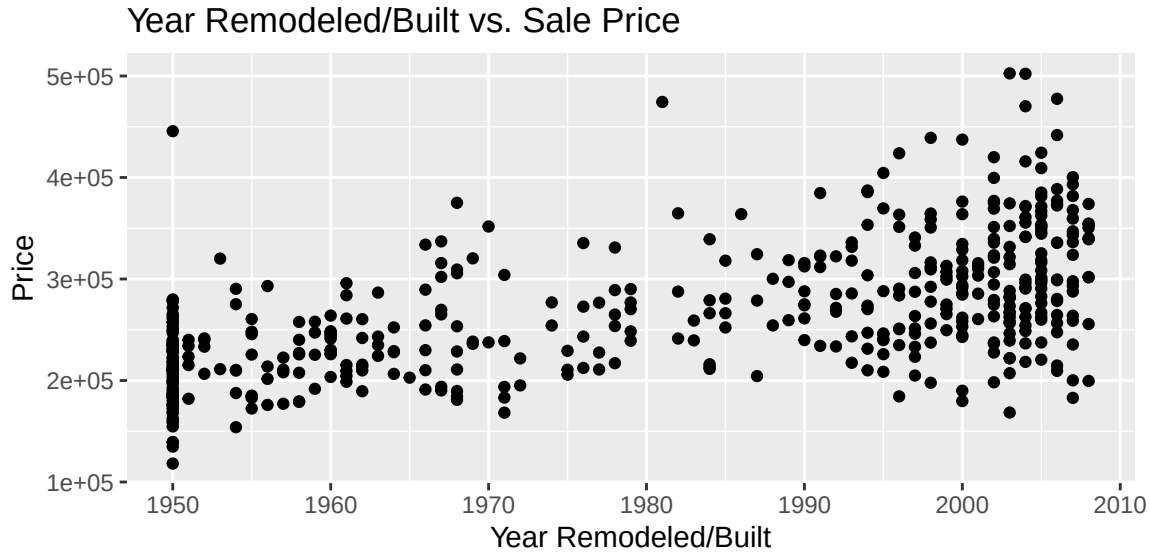
To observe the relationship between Price and categorical variables, we created the following boxplots.



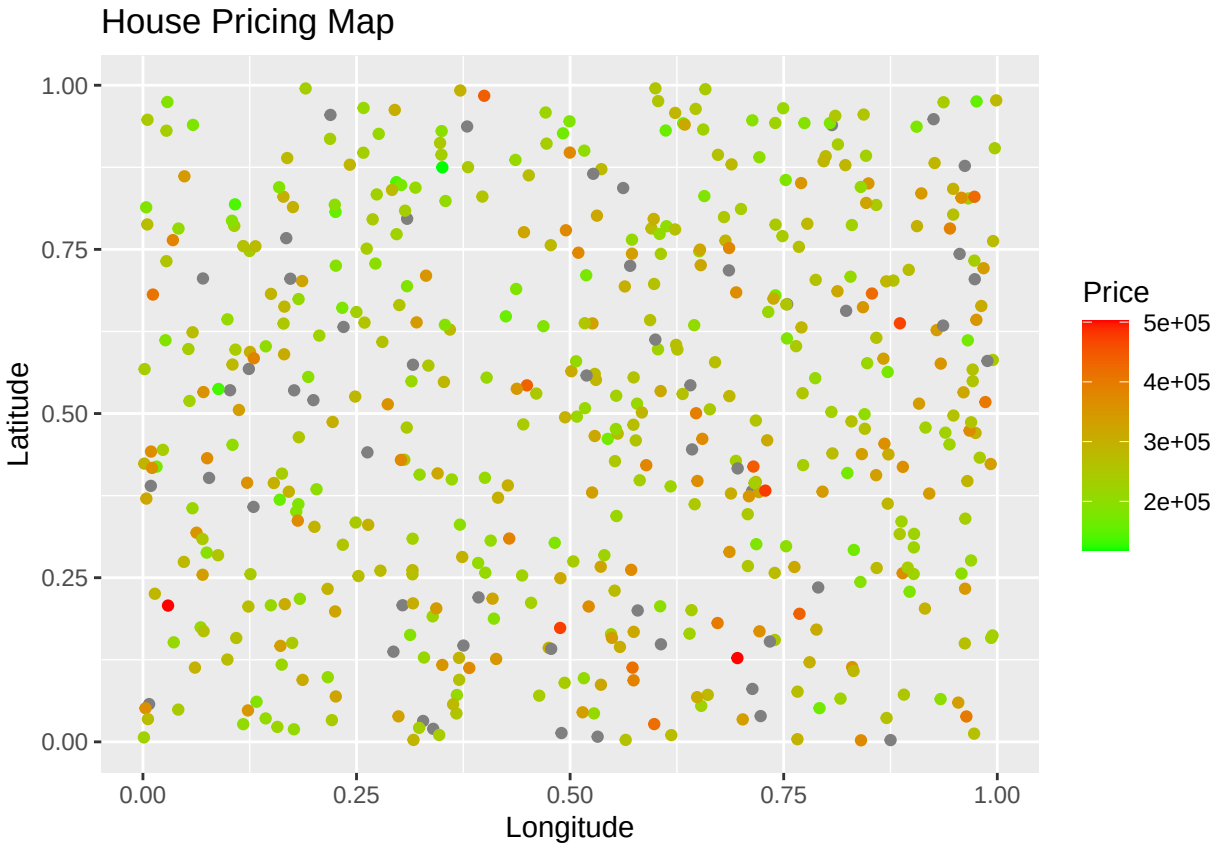
The housing styles seem to have pretty similar median prices with slightly varying spreads. Overall, it seems that the 2 story houses have greater pricing in general, which makes sense since there is "more" house. For central air in the houses, the prices of houses with no air are significantly less in price, which may be due to the lack of observations but also makes sense since it is another feature.

Below, we created another scatter plot of price versus above-ground square footage but added coloring based on housing style with lines to see the linear direction of each style. From the boxplots earlier, it looked like 2 story houses cost more usually, however, from this plot, it appears that one-story houses do because of their above-ground square footage.

Lastly in observing the data, we created the scatter plot below to see the relationship between pricing and years of renovations done to the house. Just a note, some of the observations are not renovations but rather just the build date since the house may not have had any remodeling or renovations. Overall, it appears that houses more recently build/renovated have a higher pricing. It is interesting to see a cone like shape in this graph, which again indicates that there may be heteroskedasticity in the data to address.



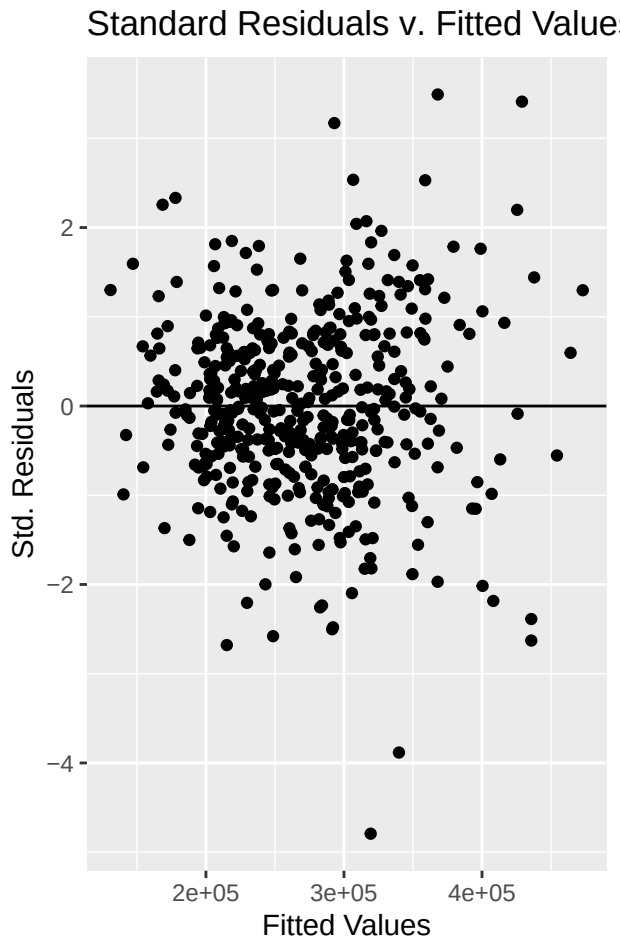
Since we have locations of the houses, we plotted the locations and the prices that we have in the plot below. There are some blank spots due to some missing values. So, it looks like there is mostly an even disbursement of the housing prices in the Ames, IA area with mostly expensive ones in the north.

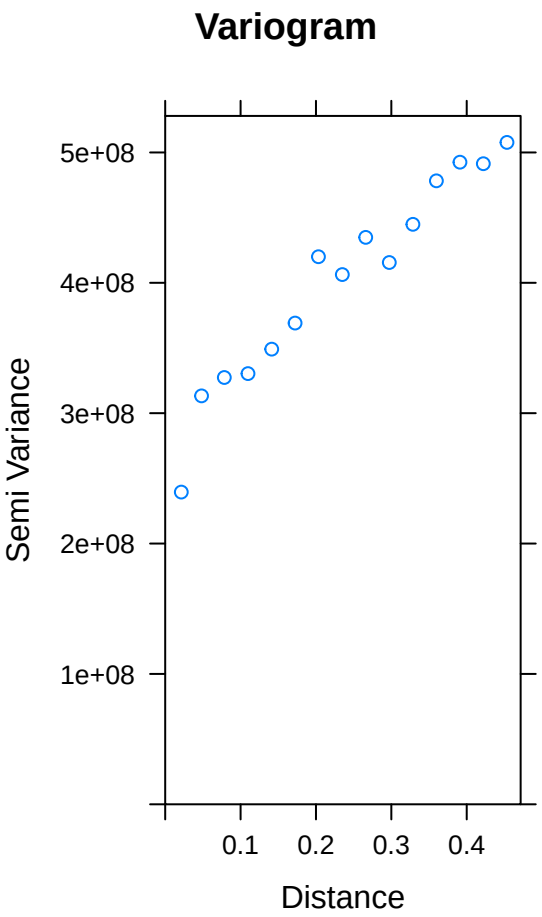


From the exploratory data analysis done on the data previously, there seems to be a problem with heteroskedasticity and also possibly some spatial correlation since we have the locations of the observations. To test if these issues are present and may affect our analysis, we created the following plots.

Table 1: AIC Score

Exp	Sphere	Gaus
10454.65	10455.25	10464.92





First, we check the equal variance assumption on a basic linear model and found that to be violated with the above graph, which proves these is heteroskedasticity in the data that needs to be accounted for. Also, we created a variogram to see if the data was independent and found that there is spatial correlation as well. If we go forward with a model that has these two issues in this analysis, we risk making incorrect conclusions about the data and not answering the questions we are trying to answer. We risk the consequences of incorrect confidence intervals and p-values for t tests as well as the results being biased. If we account for them, then we will be able to run a regression analysis on the data and be able to find what influences the housing prices and even make predictions.

Since we want to conduct this analysis correctly, we will be using a multiple linear regression model that accounts for both heteroskedasticity and spatial correlation that is a mixed model. We created a correlation matrix to account for these two issues. We will be able to observe the pricing variability with out any influence of correlation from house to house. Compared to other models, this model had the best AIC as shown by the table above.

Section 2. Spatial Heteroskedastic Model

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{M})$$

$\boldsymbol{y}$: A nx1 vector of our response variable. In this case, this is the sale price of a home.

$\boldsymbol{X}$: A model matrix of explanatory variables (covariates). In this case, we first have a column of 1's for the intercept and additional columns for each of the explanatory variables (living area in square feet, the style of dwelling, the remodel date, whether the home has central air, the number of full bathrooms and bedrooms above ground, and the size of the garage in car capacity). Each row in the matrix represents an individual home.

$\boldsymbol{\beta}$: A vector of coefficients for each explanatory variable. In our case, these are the coefficients representing the intercept (the expected sales if all the other explanatory factors were 0) and the effect for the explanatory factors on our response (sale price of a home).

σ^2: This represents the variability of y about the regression line. This will be multiplied against the d_{ii}s along the diagonal of our M correlation matrix to get the total variability.

$\boldsymbol{M}$: This represents the correlation between homes. There are d_{ii}s along the diagonal because the variability in the sale price of a home depends on the above ground living area in square feet. There are also special spatial correlation relationships between different homes along the off-diagonal represented by $\rho(s_i, s_j)$.

$\rho(s_i, s_j)$: $\rho(s_i, s_j) = e^{-\frac{||s_i - s_j||}{\phi}}$, where ϕ is the range parameter. We use iterative maximum likelihood estimation to estimate this from our data. As ϕ gets larger, the range of the spatial correlation also grows.

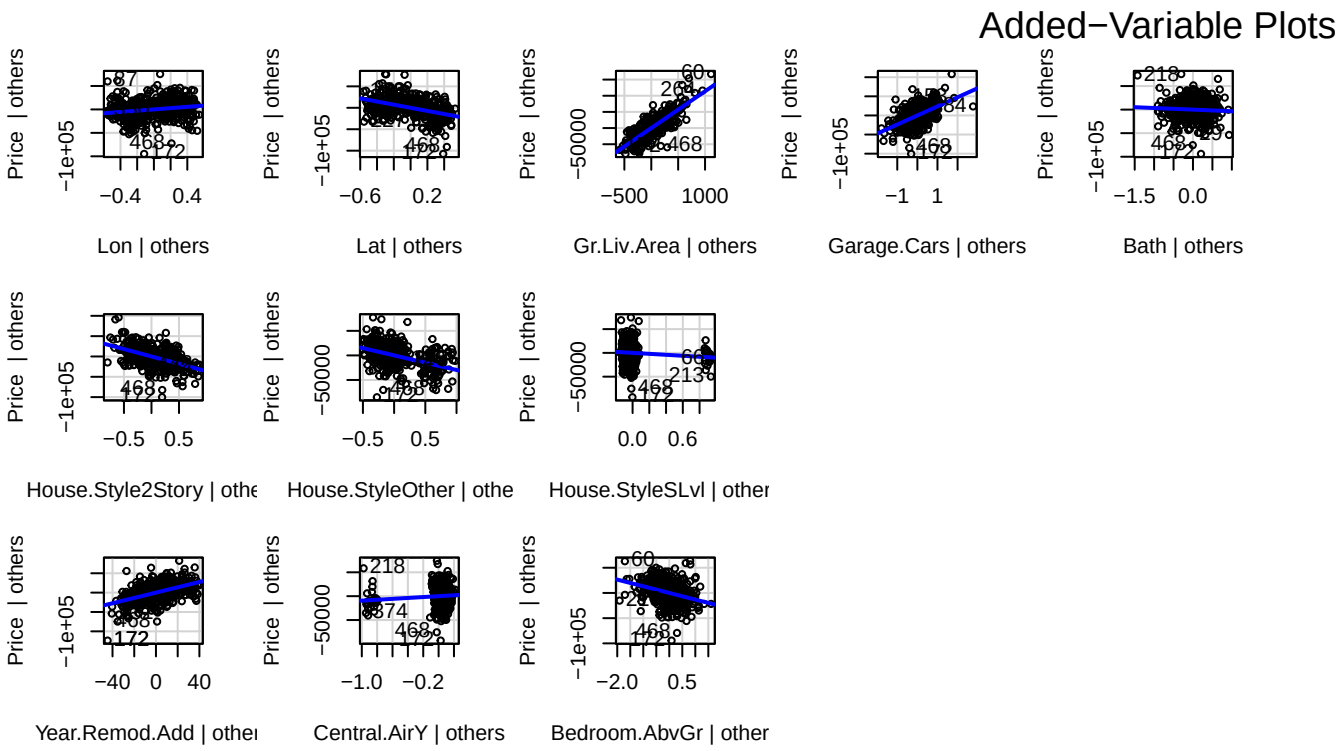
d_{ii}: $d_{ii} = e^{2\text{Gr.Liv.Area}_i\theta}$. This represents the variability of the sale price of the home in relation to the above ground living area in square feet. Our model says that the variability has an exponential relationship with the above ground living area in square feet, with this relationship depending on our θ parameter. We also calculate θ through maximum likelihood estimation. As θ increases, so does the variability of the sale price of the home.

The assumptions that we will need to asses are linearity, independence, normality, and equal variance.

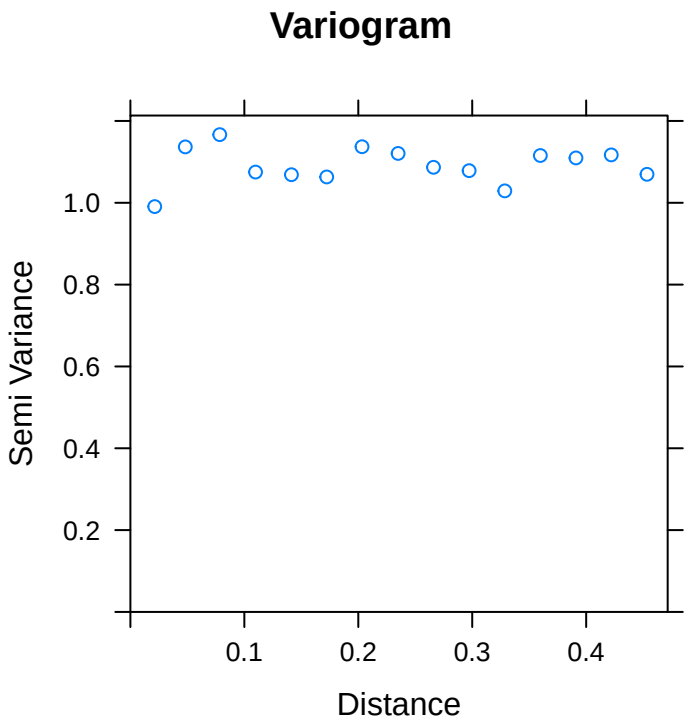
Section 3. Model Validation

To use our model, we must first verify that it meets the necessary assumptions.

We first used added-variable plots to check the linearity assumption. Our plot shows us that there are no non-linear relationships in our data, so we are good to continue.

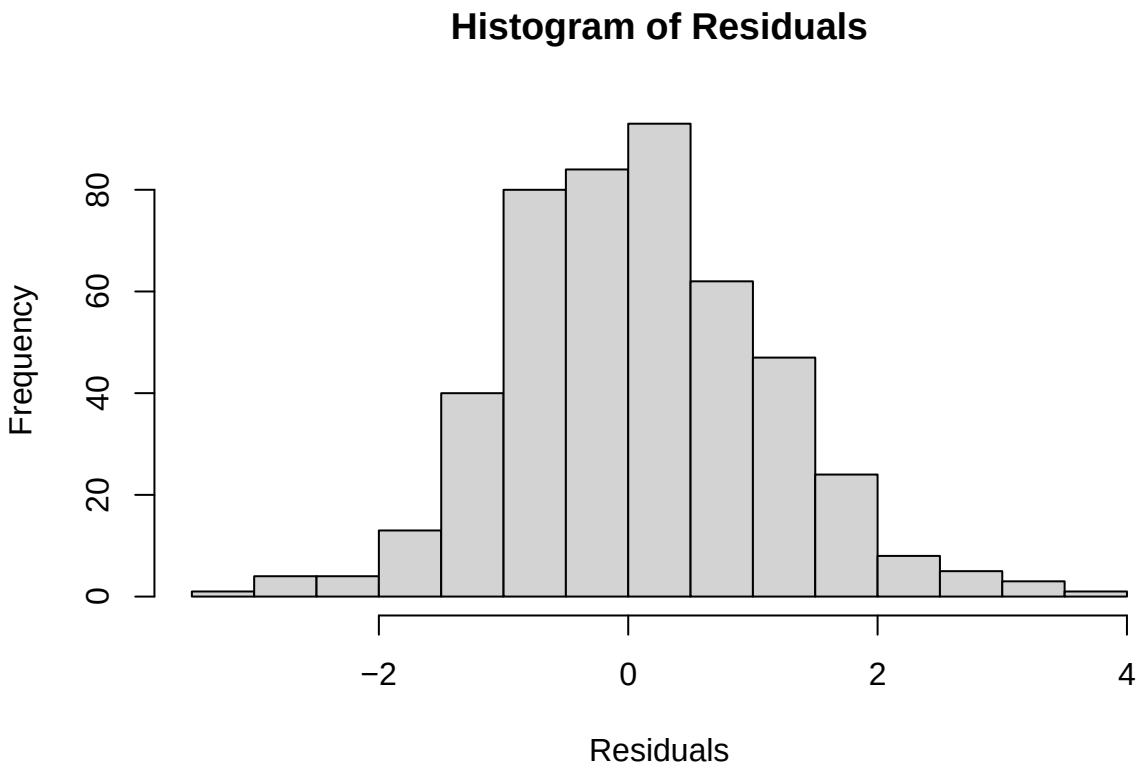


Next we plotted a variogram to show that our decorrelated residuals no longer show any correlation that our model does not account for. This variogram shows us a nice line at the top of our graph around 1, so we can safely say our model captures all of the correlation within the data.

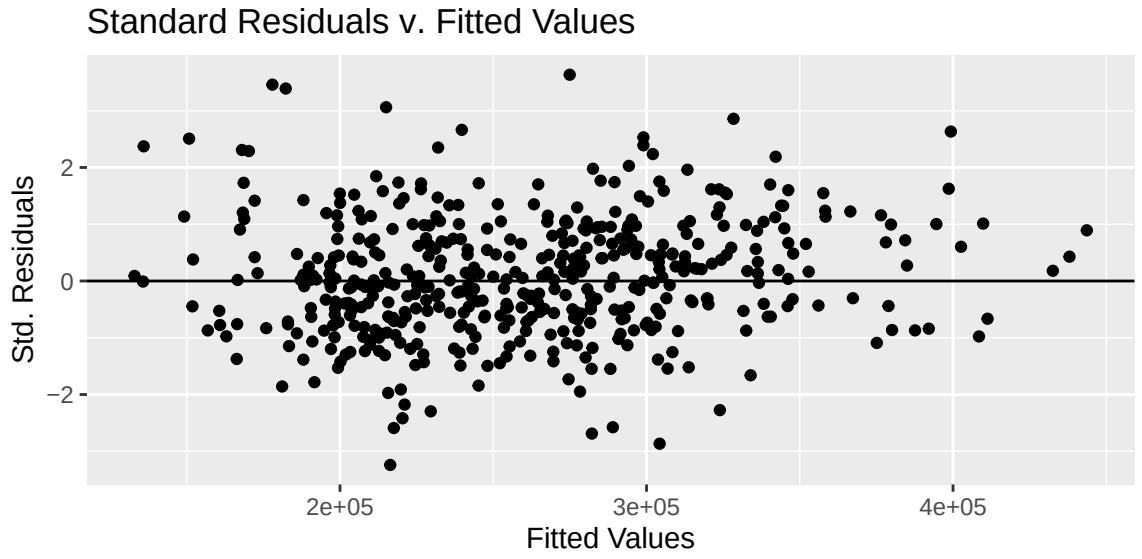


Afterwards we plotted a histogram of the residuals to test for normality. Our residuals form a nice normal

curve about our regression line, so we can say our model meets the normality assumption.



Lastly we plotted our standardized residuals against the fitted values. These showed no obvious special relationships we are not already accounting for, and the residuals varied equally above and below our fitted values.



After checking all of the necessary assumptions, our model is ready to be used for both statistical inference and prediction.

R2
0.87

We also calculated the model R-Squared below to see how well our model fits the data. This returned a value of 0.87, which means our model accounts for 0.87 percent of the data.

Cross Validation

Bias	RPMSE	Coverage	Width
23.14	17721.44	0.96	74297.23

We then performed cross-validation on our model to judge how well it performs in terms of prediction. Our prediction intervals are on average 7.4297×10^4 wide, and capture the true home sale price 0.96 percent of the time. Our model predictions are off by only 1.7721×10^4, on average, which compared to the overall range of housing prices is quite good. Compared to the standard deviation of these home prices, NA, our model is quite accurate in its predictions.

Section 4. Analysis Results

Our model and the data we used explains away 88 percent of the variability in home prices. Overall, our model explains price fairly well.

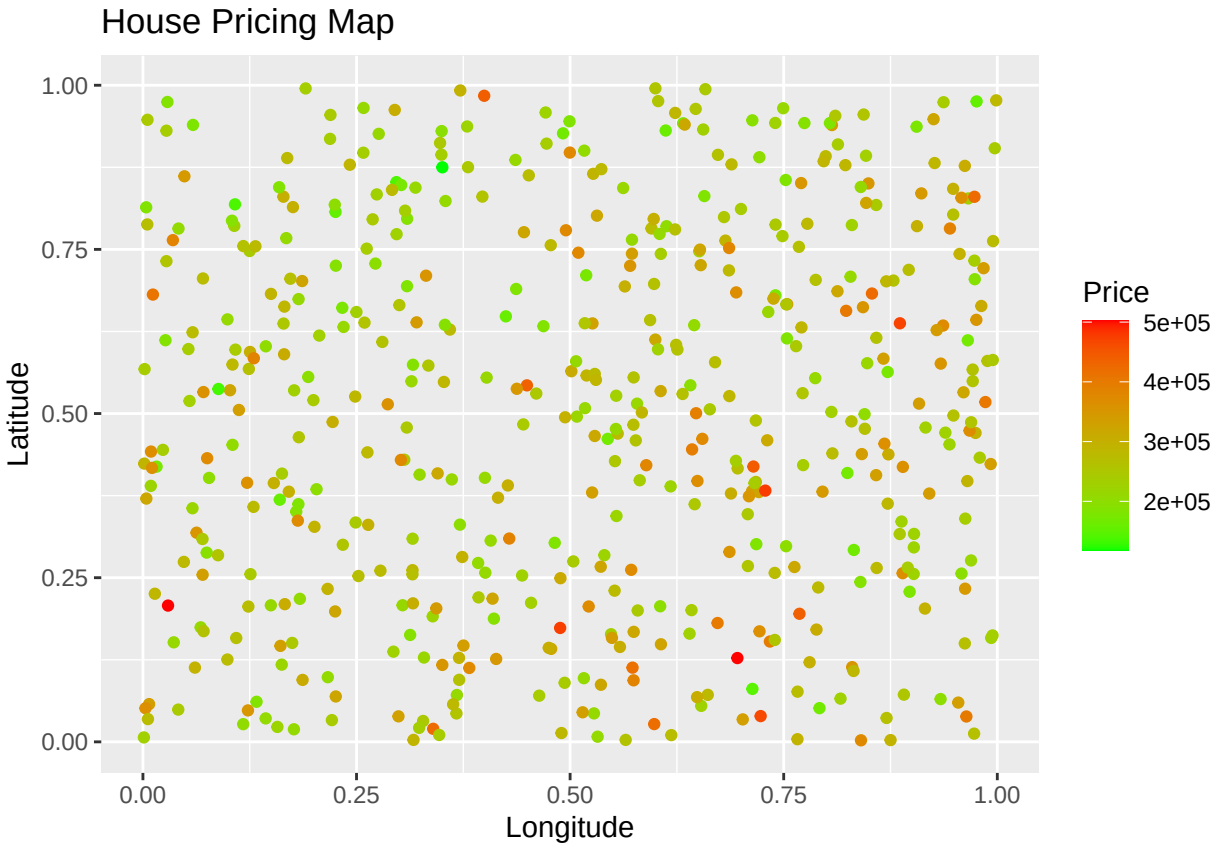
The table below tell us what factors most contributed to increasing the sale price of the home. These factors were central air, the size of the garage, the above ground living area in square feet, and the year it was remodeled.

	2.5 %	97.5 %
(Intercept)	-1.307953e+06	-1034097.6340
Gr.Liv.Area	8.850436e+01	110.2503
House.Style2Story	-3.633930e+04	-26705.0856
House.StyleOther	-2.702141e+04	-19907.2821
House.StyleSLvl	-1.555603e+04	-3462.6224
Year.Remod.Add	5.781251e+02	718.5547
Central.AirY	1.007694e+04	20553.5827
Bedroom.AbvGr	-1.655602e+04	-11587.9177
Garage.Cars	2.158252e+04	26166.4751
Bath	-2.976407e+03	5183.9166

We can also confidently say that the variability of home sale prices does grow as the size of the living area gets larger. We can assume this by looking at the variance functions positive value.

Variance
0.0006109

In the plot below, we plotted all the data points along with predicting the prices for those that were missing.



Section 5. Conclusions

From this analysis, it appears sell price is influenced by the features of the home. The model that we built for this analysis did a fine job at helping us understand the different influences on house prices and also predict prices with it as well. Specifically, we found that square footage above ground, central air, garage size, build/remodel dates influence the house price the most. Also, we were able to support our assumptions from exploring the data, that the above ground square footage does influence the variability of housing prices being more expensive, which makes sense since a larger house should cost more.

To understand more about house features influencing price, researchers should look at square footage below ground, since the above ground amount was very influential. They should also consider separating remodel and build dates as variables. The way the data is currently set up is deceiving since you cannot tell if a house was truly remodeled or not. We did find it to be a influential variable, but we feel that will change if what we suggest is done.

```
#Loading in the projects
library(GGally)
library(car)
library(MASS)
library(tidyverse)
library(lmtest)
library(multcomp)
library(nlme)
library(gstat)
library(nlme)
```

```
library(kableExtra)
#Loading in the Data
home <- read_csv('../Datasets/HousingPrices2.csv')

source("../Datasets/stdres.gls.R")
source("../Datasets/predictgls.R")

#Combining Full Bath and Half Bath
home$Half.Bath <- if_else(home$Half.Bath >= 1, home$Half.Bath/2,0)
home <- home %>%
  mutate(Bath = Half.Bath + Full.Bath) %>%
  dplyr::select(-c(Half.Bath,Full.Bath,Kitchen.AbvGr))

home$House.Style <- if_else(home$House.Style == '1.5Fin' | home$House.Style == '1.5Unf' | home$House.Sty

home$House.Style <- as.factor(home$House.Style)
home$Central.Air <- as.factor(home$Central.Air)


#Removing NA's
df <- na.omit(home)
#pairs plot of numerical variables
ggpairs(home[, c(1,4,6,8,9,10)], progress=FALSE)
#plotting living area vs price

ggplot(data=home,mapping=aes(x=Gr.Liv.Area, y=Price)) +
  geom_point() +
  ggtitle('Living Area vs. Price') +
  labs(x = 'Ground Living Area (sq ft.)')

#looking at price vs categorical variables
library(gridExtra)
p1<-ggplot(data=home,mapping=aes(y=House.Style, x=Price)) +
  geom_boxplot() +
  ggtitle('House Style vs. Price') +
  labs(y = 'House Style')

p2<-ggplot(data=home,mapping=aes(x=Central.Air, y=Price)) +
  geom_boxplot() +
  ggtitle('Central Air vs. Price') +
  labs(x = 'Central Air')

grid.arrange(p1, p2, ncol=2)
#plotting living area vs price colored by style
ggplot(data=home,mapping=aes(x=Gr.Liv.Area, y=Price, color = House.Style)) +
  geom_point() + geom_smooth(se = F) +
  ggtitle('Living Area vs. Price') +
  labs(x = 'Ground Living Area (sq ft.)')

#plot price vs remodel date
ggplot(data=home,mapping=aes(x=Year.Remod.Add, y=Price)) +
  geom_point() +
  ggtitle('Year Remodeled/Built vs. Sale Price') +
```

```
  xlab('Year Remodeled/Built')

#plot of the prices and locations
ggplot(data=home,mapping=aes(x=Lon, y=Lat, color = Price)) +
  geom_point() +
  scale_color_gradient(low = 'green', high = 'red') +  ggtitle('House Pricing Map') +
  xlab('Longitude') +
  ylab('Latitude')




#Checking variance on normal lm
basic.lm<- lm(Price~.-Lon-Lat,data=na.omit(home))

fit.vals <- predict.lm(basic.lm)
residuals <- stdres(basic.lm)
p3<-ggplot()+
  geom_point(mapping = aes(x = fit.vals, y = residuals))+
  xlab("Fitted Values") + ylab("Std. Residuals") +
  ggtitle("Standard Residuals v. Fitted Values")+
  geom_hline(mapping = aes(yintercept=0))

#create variogram of residuals
myVariogram <- variogram(object= Price~Gr.Liv.Area+House.Style+Year.Remod.Add+Central.Air+Bedroom.AbvGr


grid.arrange(p3, plot(myVariogram, main = 'Variogram', xlab = 'Distance', ylab = 'Semi Variance'), ncol=

exp_exp <- gls(model=Price~.-Lon-Lat,
           data=na.omit(home),
           weights=varExp(form=~Gr.Liv.Area),
           correlation=corExp(form=~Lon+Lat, nugget=TRUE),
           method="ML")
exp_sphere <- gls(model=Price~.-Lon-Lat,
           data=na.omit(home),
           weights=varExp(form=~Gr.Liv.Area),
           correlation=corSpher(form=~Lon+Lat, nugget=TRUE),
           method="ML")

exp_Gaus <- gls(model=Price~.-Lon-Lat,
           data=na.omit(home),
           weights=varExp(form=~Gr.Liv.Area),
           correlation=corGaus(form=~Lon+Lat, nugget=TRUE),
           method="ML")

aic <- data_frame(Exp = AIC(exp_exp), Sphere = AIC(exp_sphere), Gaus = AIC(exp_Gaus))

knitr::kable(aic, caption = 'AIC Score')
#Linear Model
lm <- lm(Price ~ ., data = na.omit(home))

#Linearity
```

```
avPlots(lm)
## Residuals ##

sres <- stdres.gls(exp_exp)

#Independence

residDF <- data.frame(Lon=na.omit(home)[['Lon']], Lat=na.omit(home)[['Lat']], decorrResid=sres)
residVariogram <- variogram(object=decorrResid~1, locations=~Lon+Lat, data=residDF)

plot(residVariogram, main = 'Variogram', xlab = 'Distance', ylab = 'Semi Variance')

#Normality
hist(sres, main = 'Histogram of Residuals', xlab = 'Residuals')
#Equal Variance
ggplot(data = na.omit(home)) +
  geom_point(mapping = aes(x = fitted(exp_exp), y = sres)) +
  xlab("Fitted Values") +
  ylab("Std. Residuals") +
  ggtitle("Standard Residuals v. Fitted Values")+
  geom_hline(mapping = aes(yintercept=0))
R2 <- cor(df$Price,predict(exp_exp))^2

r = tibble(R2 = R2)
knitr::kable(round(r,2))


source("../Datasets/predictgls.R")

n.cv <- 50 #Number of CV studies to run

n.test <- round(.1*nrow(df),2)

rpmse <- rep(x=NA, times=n.cv)

bias <- rep(x=NA, times=n.cv)

wid <- rep(x=NA, times=n.cv)

cvg <- rep(x=NA, times=n.cv)


  ## Run the CV code

for(cv in 1:n.cv){
  ## Select test observations
  test.obs <- sample(x=1:nrow(df), size=n.test)

  ## Split into test and training sets
  test.set <- df[test.obs,]
  train.set <- df[-test.obs,]

  ## Fit a gls()
```

```
  train.gls <- gls(model=Price~.-Lat-Lon,
           data=train.set,
           weights=varExp(form=~Gr.Liv.Area),
           correlation=corExp(form=~Lon+Lat, nugget=TRUE),
           method="ML")


  ## Generate predictions for the test set
  my.preds <- predictgls(glsobj=train.gls, newdframe=test.set, level=0.95)

  ## Calculate bias
  bias[cv] <- mean(my.preds[,'Prediction']-test.set[['Price']])

  ## Calculate RPMSE
  rpmse[cv] <- (test.set[['Price']]-my.preds[,'Prediction'])^2 %>% mean() %>% sqrt()


  ## Calculate Coverage
  cvg[cv] <- ((test.set[['Price']] > my.preds[,'lwr']) & (test.set[['Price']] < my.preds[,'upr'])) %>% m


  ## Calculate Width
  wid[cv] <- (my.preds[,'upr'] - my.preds[,'lwr']) %>% mean()

  ## Update the progress bar
}

results <- tibble(Bias = round(mean(bias),2), RPMSE = round(mean(rpmse),2), Coverage = round(mean(cvg),

knitr::kable(results)


## Question 2
knitr::kable(confint(exp_exp))
## Question 3
variance  <- tibble(Variance = coef(exp_exp$modelStruct, unconstrained=FALSE)[3])
knitr::kable(variance)
#Question 4

#Dataset with NA's
home_na <- home %>%
  filter(is.na(Price))

## Predictions
fit.vals <- predictgls(exp_exp,home_na)

##Fill in NA's
home_na$Price <- fit.vals$Prediction

## Combine datasets
preds <- rbind(home_na,na.omit(home))
```

```
# Graphing the temperatures
ggplot(data=preds,mapping=aes(x=Lon, y=Lat, color = Price)) + geom_point() + scale_color_gradient(low =
  ggtitle('House Pricing Map') +
  xlab('Longitude') +
  ylab('Latitude')
```