Sale Price of Homes

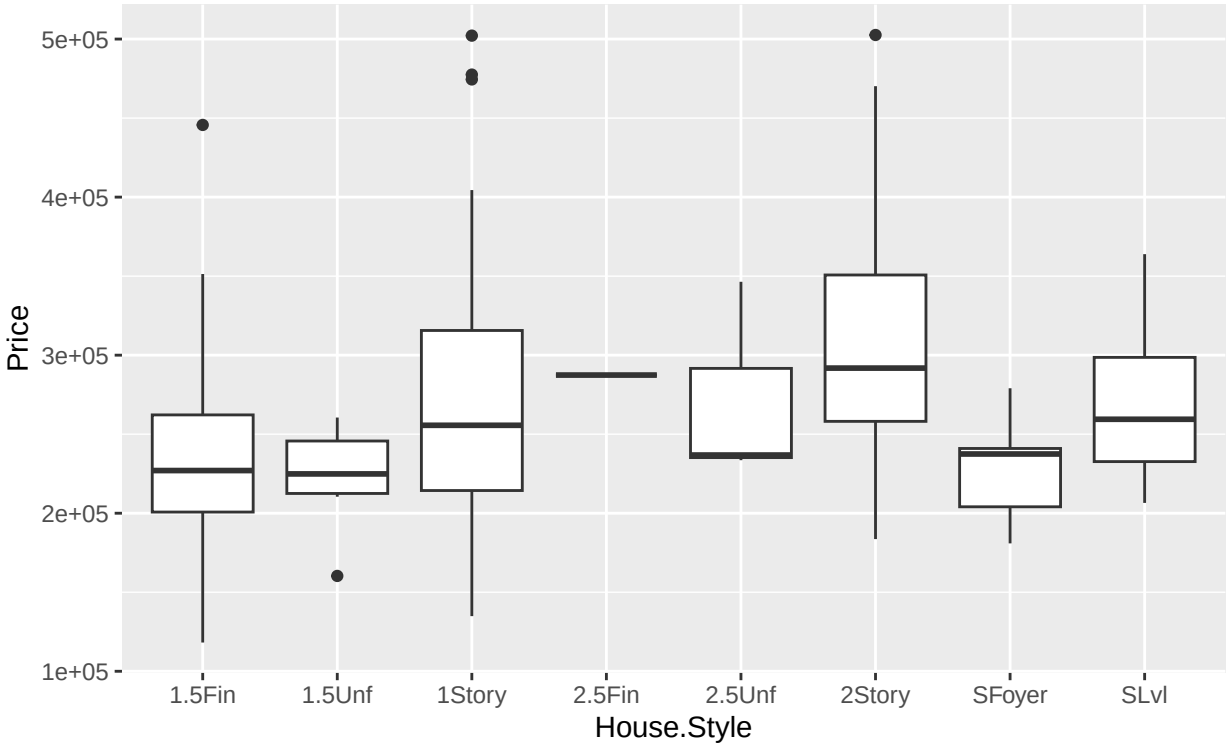
Section 0

This analysis takes a purely statistical approach, using home characteristics (longitude, latitude, living square footage, house style, remodel date, central air, bathroom count, bedroom count, and garage capacity), to appraising homes in Ames, IA. I found a few interesting things regarding the impact various characteristics have on the sale price of a home. As houses get larger in square footage, their prices begin to vary more and more. Knowing this and accounting for this led to more accurate home appraisals. I also found that certain home characteristics resulted in a larger price increase than others, specifically having more garage capacity, central air, a newer construction, and more square footage were large contributors. Location was also accounted for in my analysis, and while it played a role in the appraisal, it was limited to only longitude and latitude. Overall, I learned that home appraisals need to take many different factors into consideration, but can still be predicted with relatively high accuracy using statistics.

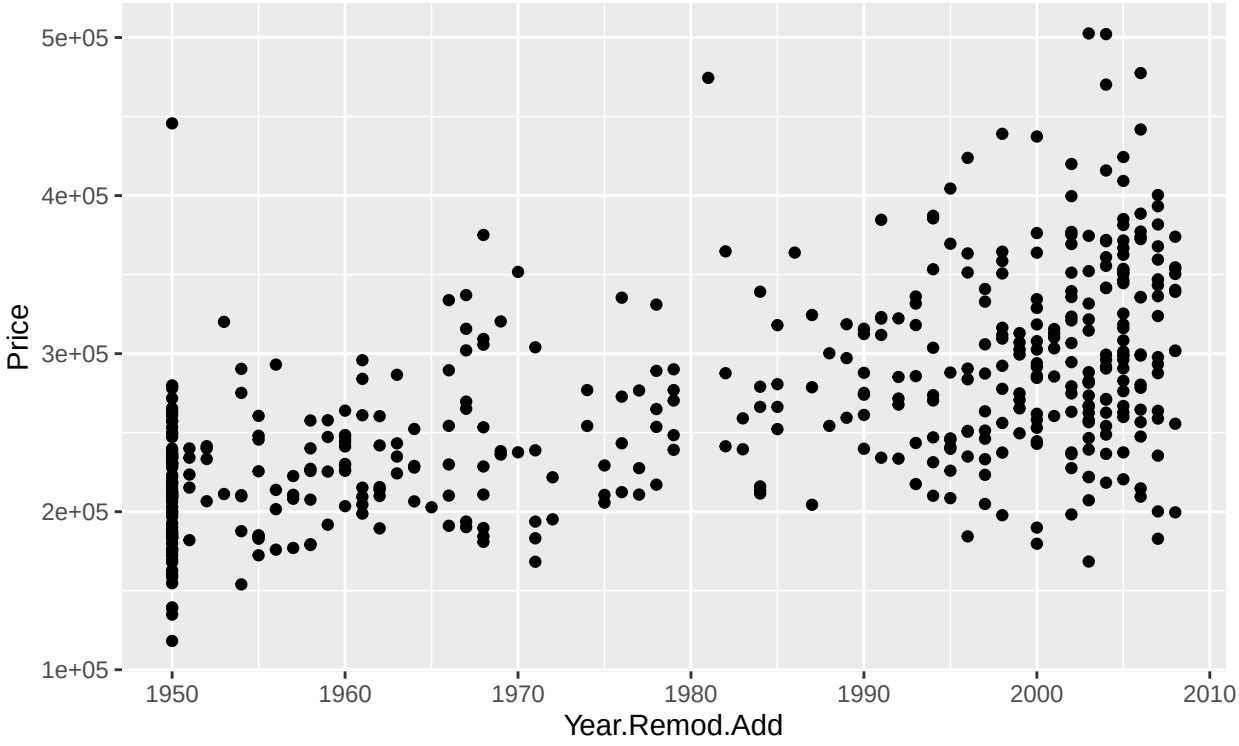
Section 1

When buying a new home with a mortgage, the home needs an appraisal of the value to be used as collateral against the loan. An appraiser would inspect the property based on factors such as exterior and interior condition, lot size, home improvements the seller made, and any renovations or additions. The aim of this analysis is to appraise various homes in the Ames, IA area using a statistics approach. The goals of the study are to learn how well home characteristics explain the price, what factors increase the price the most, do larger homes have a greater variability in price. Using the answers to those questions will then allow for the homes in the data that do not have a sale price to be appraised.

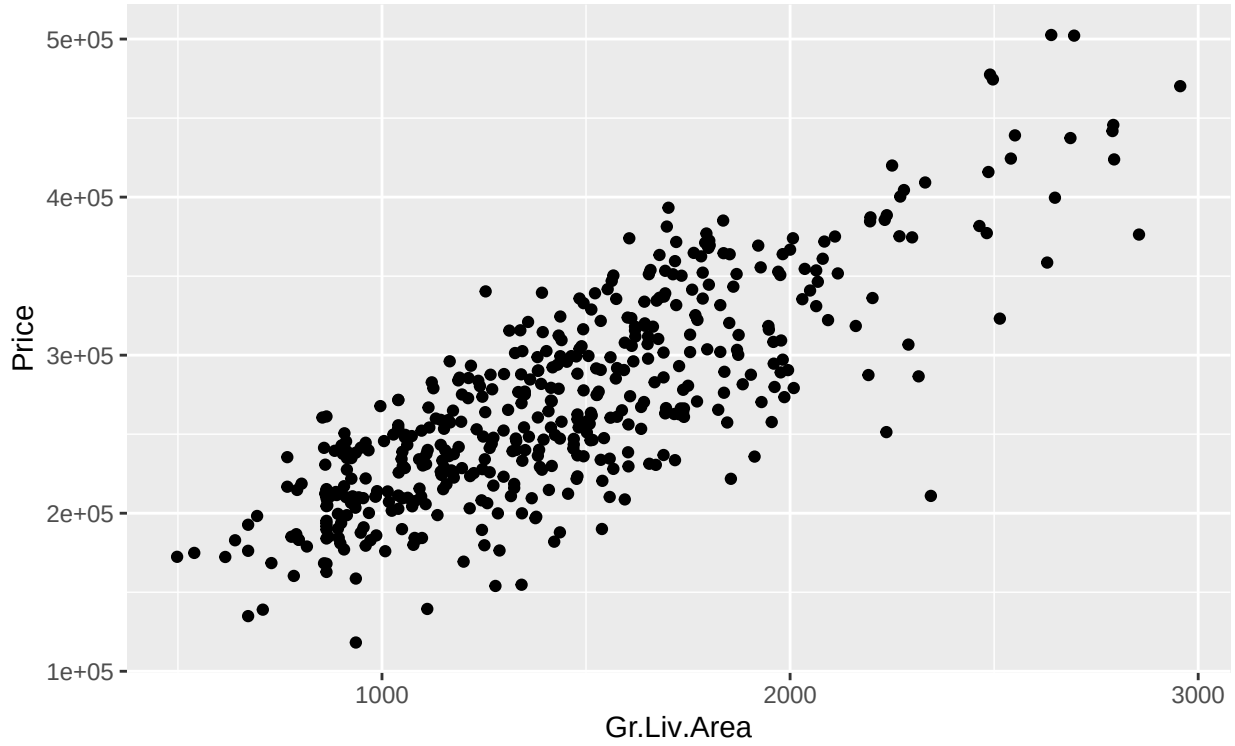
The data provides information on the sale price, transformed measures of longitude and latitude, living area square footage above ground, style of house, remodel/original construction date, if it has central air, number of full bathrooms, number of half bathrooms, number of bedrooms, and size of garage. The data comes from homes in Ames, IA from 1950 to 2008.



The average prices of all the home styles fall between $200,000-300,000, however some styles have a lot of variability while some styles have very little variability.



There is a general positive linear trend here that indicates that the price of a home increases as the remodel, or original construction, date is more recent.



This shows a fairly strong, positive linear relationship between price and above ground living area. Meaning that home price increases as living area increases.

A couple potential issues associated with the data that I want to address are the changing variability in house prices based on size of the home and the house price of houses that are close together in space affecting each other. The consequence for ignoring the change in variability in home prices is that the standard errors would be incorrect, leading to a model that fits the data poorly and doesn't predict accurately. By accounting for this issue, one of the goals of the study (learning if larger homes have a greater variability in price) will be able to be accomplished and the model will fit the data much better. The consequence for ignoring house prices being correlated based on proximity is that prediction will be inaccurate. By accounting for this issue, homes can be accurately appraised meaning that the main purpose of the study can be accomplished.

The method I will use to analyze the data started above by making a few exploratory plots to determine what type of statistical model I want to fit to the data and to learn about some of the trends of the data. From those plots and from looking at the data, I determined that I want to fit a heteroskedastic, spatial multivariate linear regression model. The next step is to use iterative optimization to get the maximum likelihood estimates of the model parameters needed for the heteroskedastic, spatial MLR. Then I will verify the model assumptions so I can make inference results to address the goals of the study.

Section 2

Statistical model: $y \sim N(X\beta, \sigma^2 DR)$

y is the response variable or target vector (the house price).

X is the design matrix containing a column for the intercept and a column to include the data from the dataset (living area, house style, year remodeled, central air, full bath, half bath, bedrooms, and garage size).

β is the vector of parameters where when all x's are zero, y is β_0 (intercept) on average and as the pth x goes up by 1, y goes up by β_p (coefficients for living area, house style, year remodeled, central air, full bath,

half bath, bedrooms, and garage size) on average.

σ^2 is the variability of y about the regression line.

D is a matrix of d_{ii} on the diagonal and the covariances everywhere else but since y_i is independent of y_j all the covariances are 0.

$\text{Var}(y_i) = \sigma^2 d_{ii}$

Using an exponential variance function $d_{ii} = exp\{2x_i\theta\}$ where x_i is a covariate (θ is a parameter estimated from the data).

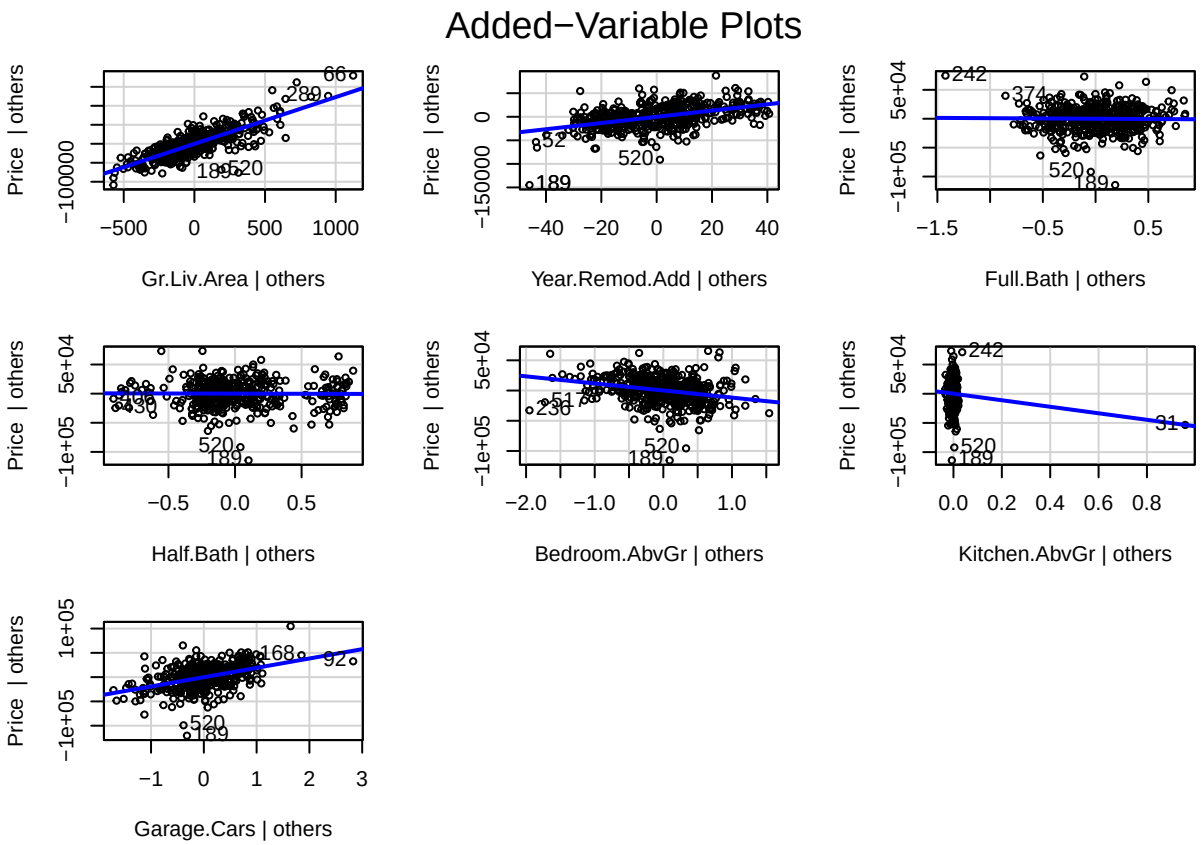
R is an n x n covariance matrix for location with 1's down the diagonal and the other points being the correlation between locations, or in this case between longitude/latitude coordinates. For example, the value in the 1st row and 2nd column is the correlation between the first longitude/latitude location (1st house) and the second longitude/latitude location (2nd house).

$\text{Cov(y)} = \sigma^2 R$

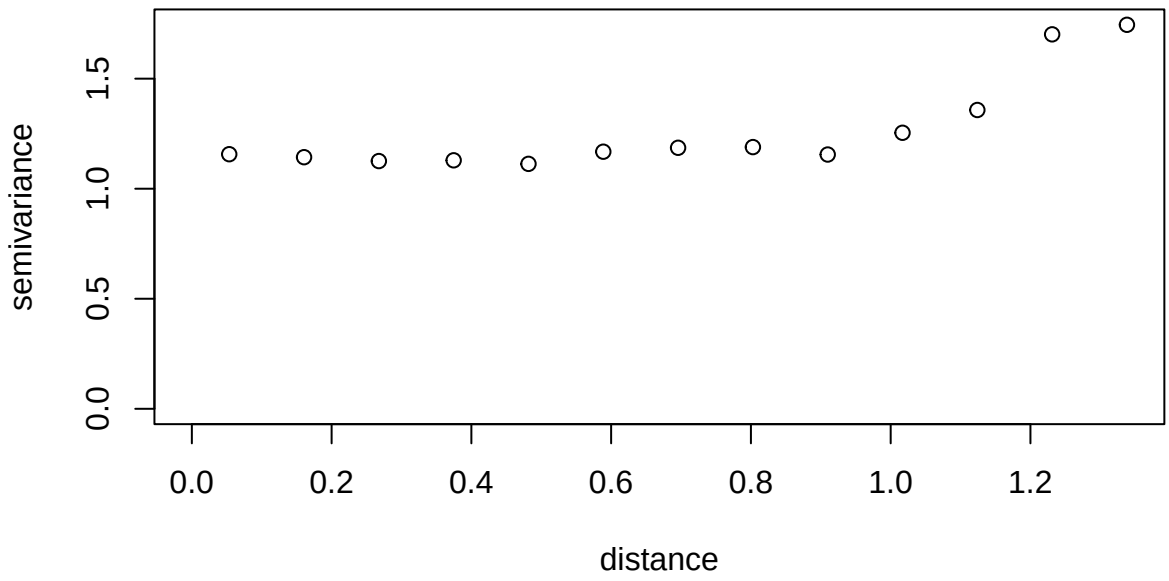
In this case the correlation function used in the model to capture spatial correlation is the exponential function.

This model uses the LINE assumptions or the linearity, independence, normality, and equal variance assumptions. For linearity there needs to be a linear relationship between the numeric explanatory variables and the response variable. For independence the residuals need to be independent of each other, or in other words there can't be correlation between the data points after decorrelating residuals. For normality the standardized residuals of the model need to be normally distributed. Lastly, for the equal variance assumption the standardized residuals need to have constant variance for all the levels of the explanatory variables, after normalizing.

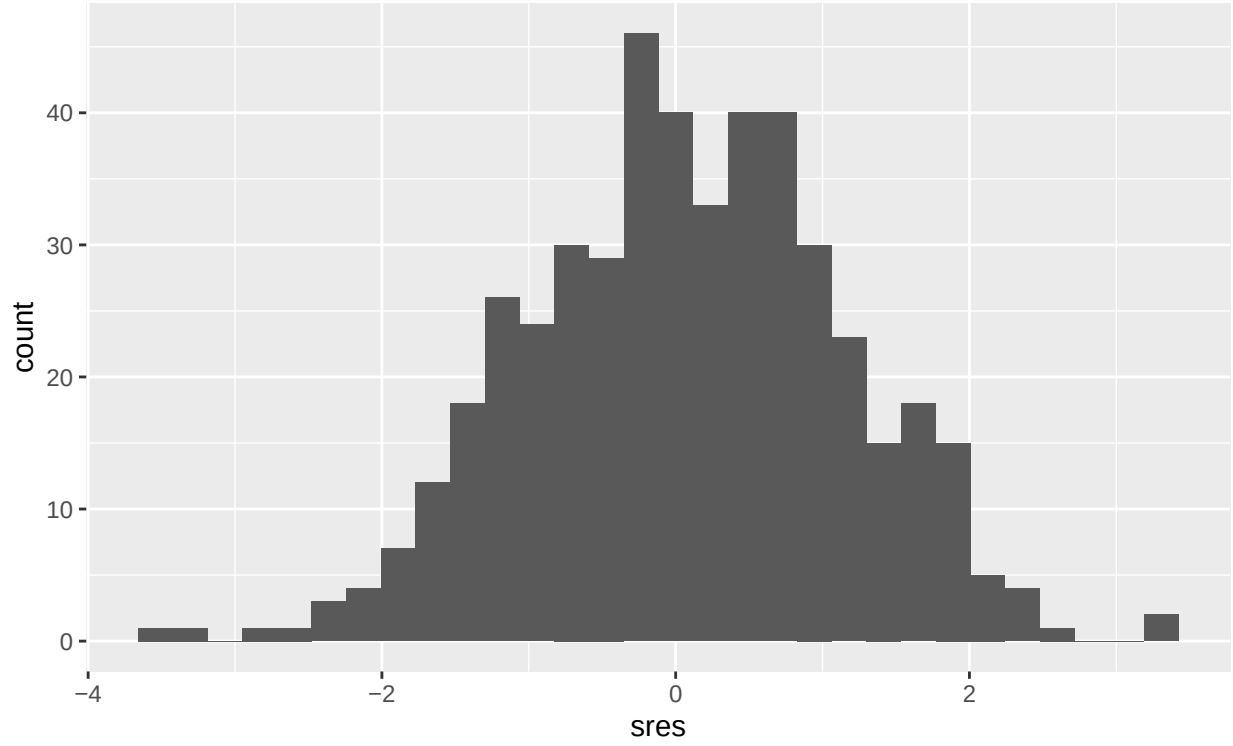
Section 3



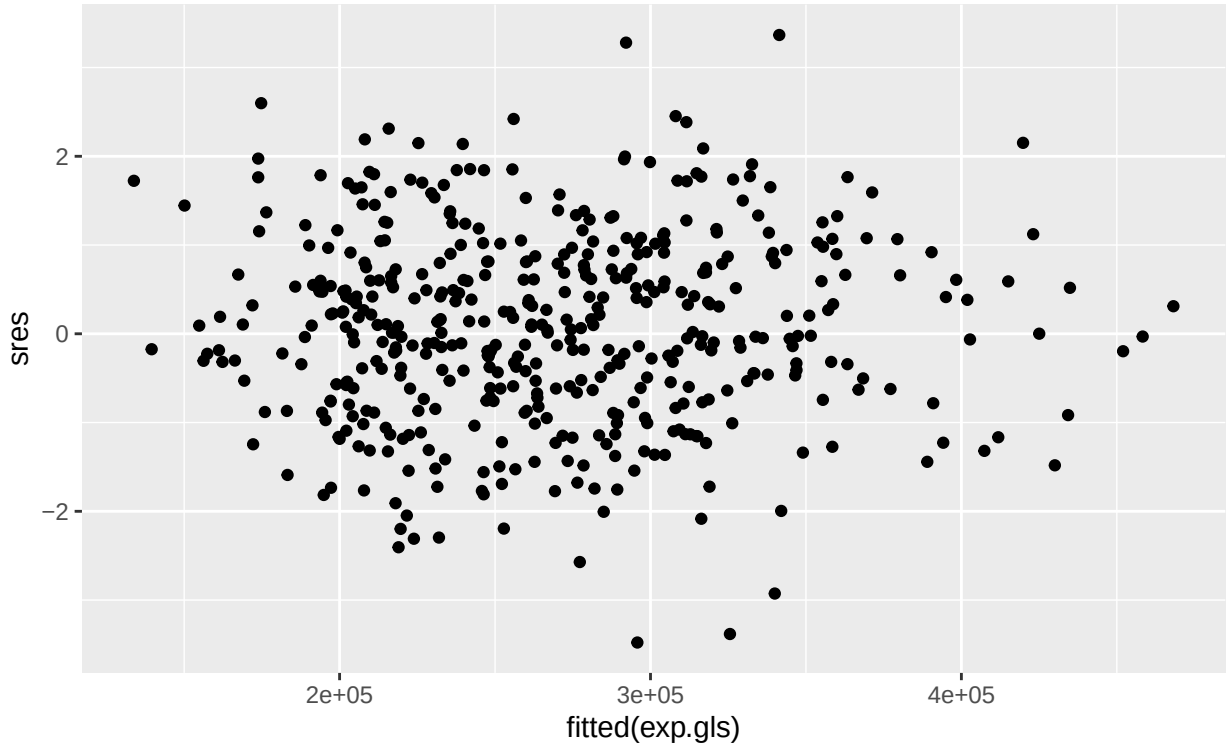
The linear assumption holds here because the added variable plots show fairly linear shapes.



The variogram has a fairly flat line, meaning that the independence assumption holds.



The histogram of the residuals looks fairly normal, so the normality assumption holds here.



The plot of the residuals vs. fitted values is centered around zero and there is no concerning shape, meaning the equal variance assumption holds here.

After calculating a couple summary statistics, I found that the model fits the data very well. The first number is the RMSE (21,845). An RMSE of 21,845 seems pretty small compared to the range (384,401) and standard deviation (65,442) of the house prices. The second number is the pseudo r-squared (0.889), which shows the percentage of the overall variance in house prices that was captured by the model.

I also ran cross-validation to see how well my model does at predicting housing prices. I did a cross-validation on the heteroskedastic, spatial MLR and an independent linear regression model to compare the results. The independent model returned an RPMSE of 1.8922097×10^4, a coverage of 0.9542857, and a width of 8.7259537×10^4. The heteroskedastic, spatial model returned an RPMSE of 1.4967307×10^4, a coverage of 0.9485714, and a width of 5.5844382×10^4. The heteroskedastic, spatial model has a lower RPMSE and Width while having a having a higher coverage. Those aspects all indicate that the model does well at predicting housing prices.

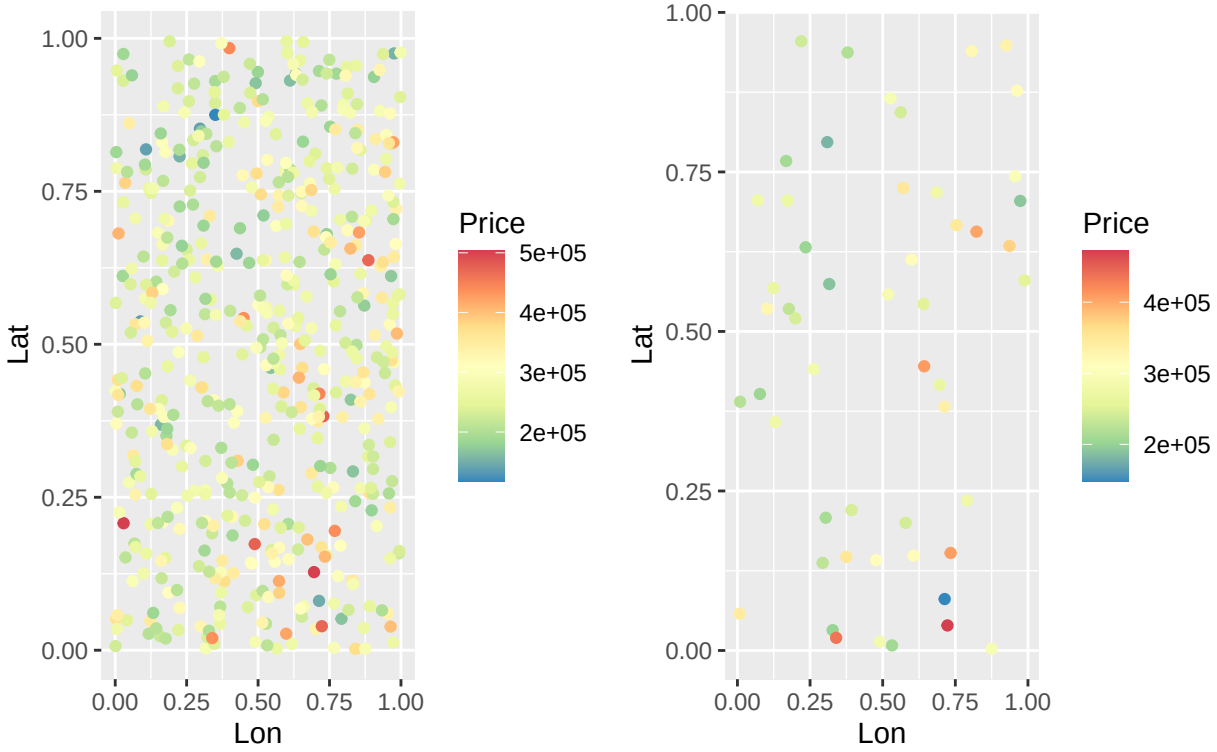
Section 4

To address the first goal of this analysis I'll refer to pseudo r-squared that was reported in the last section (0.889). This shows that 88.9% of the overall variance in house prices was captured by the model or, in other words, home characteristics explain about 89% of the variability in sale price. This means that the home characteristics do very well at explaining sale price, however they do not explain the sale price entirely.

For the second goal of this analysis to be answered I looked at the confidence intervals for the coefficients in the model. The factors that increase the sale price of home are having more living area above ground, a newer remodel or construction, central air, more garages, and a 1.5 Unf house style, a 1 story house style, an S Foyer house style, or a Split Level house style (when those house styles are compared to the 1.5 Fin house style).

To address the third goal I calculated a confidence interval for the θ parameter in the exponential variance function of the model. The interval is 0.0007 to 0.0009, and since it doesn't contain zero that means that

the variability of sale price does increase with the size of the home (as given by living area).



Above are plots to assist in accomplishing the final goal of predicting the sale price for the homes in the dataset that do not have a sale price. The plot on the left shows all of the homes in the dataset together (including the predicted prices) and the plot on the right shows only the homes with predicted prices.

Section 5

From this analysis of how different home characteristics affect the sale price of a home in Ames, IA, I found a few interesting things regarding the impact various characteristics have. As houses get larger in square footage, their prices begin to vary more and more. Knowing this and accounting for this led to more accurate home appraisals. I also found that certain home characteristics resulted in a larger price increase than others, specifically having more garage capacity, central air, a newer construction, and more square footage were large contributors. Location was also accounted for in my analysis, and while it played a role in the appraisal, it was limited to only longitude and latitude. Some of the best "next steps" to take with this would be to gather more data about location influences that could impact sale price. A few ideas that come to mind would be to get information about the neighborhood, nearby schools, and proximity to shopping/grocery. Data such as that could prove very valuable in making an even more accurate statistical approach to home appraisals.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(GGally)
library(forecast)
library(car)
```

```
library(multcomp)
library(nlme)
library(mvtnorm)
library(geoR)
library(rgdal)
library(rgeos)
library(broom)
library(spdep)
library(RSpectra)
library(sf)
library(spdep)
library(gridExtra)
source("stdres.gls.R")
source("predictgls.R")
source("moranBasis.R")
homes <- read.csv("~/STAT 469/HousingPrices2.csv", sep=",")
# Boxplot of price by house style
ggplot(data = homes, mapping = aes(x = House.Style, y = Price)) +
  geom_boxplot()
# Scatter plot of price by remodel year
ggplot(data = homes, mapping = aes(x = Year.Remod.Add, y = Price)) +
  geom_point()
# Scatter plot of price by remodel year
ggplot(data = homes, mapping = aes(x = Gr.Liv.Area, y = Price)) +
  geom_point()
# Creating a secondary data frame that has no NAs
NoNA <- na.omit(homes)
# Linear model for avplots
slm <- lm(formula = Price ~ .-Lon-Lat, data = NoNA)

# Looking at AIC and choosing model (exp spatial)
exp.gls <- gls(model = Price ~ .-Lon-Lat, data = NoNA,
               correlation = corExp(form=~Lon+Lat, nugget=TRUE),
               weights = varExp(form = ~Gr.Liv.Area), method = "ML")
aic.exp <- AIC(exp.gls)

sph.gls <- gls(model = Price ~ .-Lon-Lat, data = NoNA,
               correlation = corSpher(form=~Lon+Lat, nugget=TRUE),
               weights = varExp(form = ~Gr.Liv.Area), method = "ML")
aic.sph <- AIC(sph.gls)

gau.gls <- gls(model = Price ~ .-Lon-Lat, data = NoNA,
               correlation = corGaus(form=~Lon+Lat, nugget=TRUE),
               weights = varExp(form = ~Gr.Liv.Area), method = "ML")
aic.gau <- AIC(gau.gls)
# AVplots
avPlots(slm, terms = ~ .-House.Style-Central.Air)
# Decorrelating residuals
sres <- stdres.gls(exp.gls)
# Variogram
coormat <- as.matrix(NoNA[,2:3])
plot(variog(coords = coormat, data = as.vector(sres)))
# Histogram of the residuals
```

```
ggplot() +
  geom_histogram(mapping = aes(x = sres))
# Scatter plot of the fitted values vs. residuals
ggplot() +
  geom_point(mapping = aes(x = fitted(exp.gls), y = sres))
# RMSE
rmse <- sqrt((1/length(NoNA$Price)) * sum((NoNA$Price - fitted(exp.gls))^2))
# Pseudo R Squared
psRs <- (cor(NoNA$Price, fitted(exp.gls)))^2

# Range of log aerosol
rn <- range(NoNA$Price)
# Std dev of log aerosol
std <- sd(NoNA$Price)
# Cross validation
n.cv <- 25
n.test <- 7
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)

for(cv in 1:n.cv){
  ## Run the CV code
  ## Select test observations
  test.obs <- sample(x=1:length(NoNA[,1]), size=n.test)

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

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

  ## Generate predictions for the test set
  my.preds <- predictgls(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'])) %>% mean()

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

```r
cv.sp <- c(mean(bias), mean(rpmse), mean(cvg), mean(wid))
# Cross validation on independent model
n.cv <- 25
n.test <- 7
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)
for(cv in 1:n.cv){
  ## Select test observations
  test.obs <- sample(x=1:length(NoNA[,1]), size=n.test)

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

  ## Fit a lm() using the training data
  train.lm <- lm(formula= Price ~ .-Lon-Lat, data=train.set)

  ## Generate predictions for the test set
  my.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

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

  ## Calculate RPMSE
  rpmse[cv] <- (test.set[['Price']]-my.preds[,'fit'])^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()
}

cv.in <- c(mean(bias), mean(rpmse), mean(cvg), mean(wid))
# Looking at confidence intervals for effects
intervals <- confint(exp.gls)
# 95% confidence interval for theta
int.theta <- intervals(exp.gls, level = 0.95)
# Creating and mapping predictions of Price
sale <- homes[!complete.cases(homes),]
my.preds <- predictgls(exp.gls, newdframe=sale, level=0.95)
preds <- my.preds[,c(13, 2:12)]
names(preds)[names(preds) == 'Prediction'] <- 'Price'
houses <- rbind(NoNA, preds)

# Mapping plot of all prices
complete <- ggplot(data=houses, mapping=aes(x=Lon, y=Lat, color=Price)) +
  geom_point() +
  scale_color_distiller(palette="Spectral",na.value=NA)
# Mapping plot of only predicted prices
only.preds <- ggplot(data=preds, mapping=aes(x=Lon, y=Lat, color=Price)) +
```

```
  geom_point() +
  scale_color_distiller(palette="Spectral",na.value=NA)

grid.arrange(complete, only.preds, ncol = 2, nrow = 1)
```