Sale Price of Homes

By: Brian Colgrove

Introduction: Homeowners selling their houses need to have their homes appraised. This appraisal is a key part of the housing market as it shows the current cost of a home in a specific area. The appraisal is based on several factors of the home. These factors can include location, square footage, furnishings, number of rooms, number of bedrooms, etc. The seller and buyer both want the appraisal to be a good estimate of the actual price of the house. We have a dataset from Ames, Iowa, which contains information on 517 different appraisals. We have 52 missing home prices that we want to predict accurately. The dataset contains predominant features of the house which include location, living area, house style, year remodeled, central air, number of full baths, number of half baths, number of bedrooms above the garage, and the size of its garage. The location has been changed to protect privacy, but the longitude and latitude are accurate in relation to each house within the dataset.

We have explored the dataset with 8 different Figures below. In Figure 1, we can see that two-story homes have a higher median sale price than single-story or split-level homes. In Figure 2, we can see that there is a strong positive linear relationship between the categorical variable of garage size to the home's appraisal. As shown in Figure 3, there is a strong positive relationship between the above ground living area of a home and its sale price. According to Figure 4, the newer built or remodeled homes have a higher sale price than those that are older.

Its general knowledge that one's neighborhood affects one's house's sale price, so we should expect spatial correlation. To verify this assumption of spatial correlation, we have fitted a normal linear model to examine the residuals of the model. Figures 6-8 are figures of a variogram, residuals vs living area above ground, and residuals vs fitted. Figures 7 and 8 show the correlation while Figure 6 shows the model's variance structure. From Figure 6, we have plotted a variogram of the residuals from our linear model. We can now see that there is indeed spatial correlation in our dataset because of the semi-variance changes along with its distance. A variogram without spatial dependence would be a flat line, in other words, the semi-variance does not change along with the distance. Figure 5 further enforces the evidence of spatial correlation. We can see that there are higher residuals clumped together near the North-Eastern part of the map. Comparably, we have lower residuals are lower in the South-Western part of the map. This would suggest that location (i.e. Longitude and Latitude) creates a correlation between our observations. We should expect and account for spatial correlation. This means that the approximate location of a house does affect its sale price.

In addition to the correlation structure in the residuals, we should examine the variance structure of the residuals. The fitted values versus residuals plots in Figure 8 show that the variance in the residuals increases as the predicted price increases, this violates the assumption of equal variance. Further examination of the residuals against the living area of a home in Figure 7 shows that the heteroskedasticity is tied to the living area. Larger area homes appear to be more variable in the price of the home.

We need to account for spatial correlation because if we do not our model will run into some issues. For one, if we ignore the spatial correlation and heteroskedasticity in the dataset our any confidence or prediction interval from our model will not be accurate even though the price estimates are unbiased. We need a model that will quantify the uncertainty that comes from spatial correlation and heteroskedasticity. We will use a multiple linear regression model to build

a model that can predict the value of homes for the different house attributes. In our model, we will specify the correlation structure and variance structure as such to account for both the spatial correlation and heteroskedasticity in the residuals of the standard linear model. This structure will be accounted for jointly in the covariance matrix of the residuals of our specified model. This will allow us to have more accurate predictions that take into account the spatial dependence in the errors and larger variance of house price with greater living area.

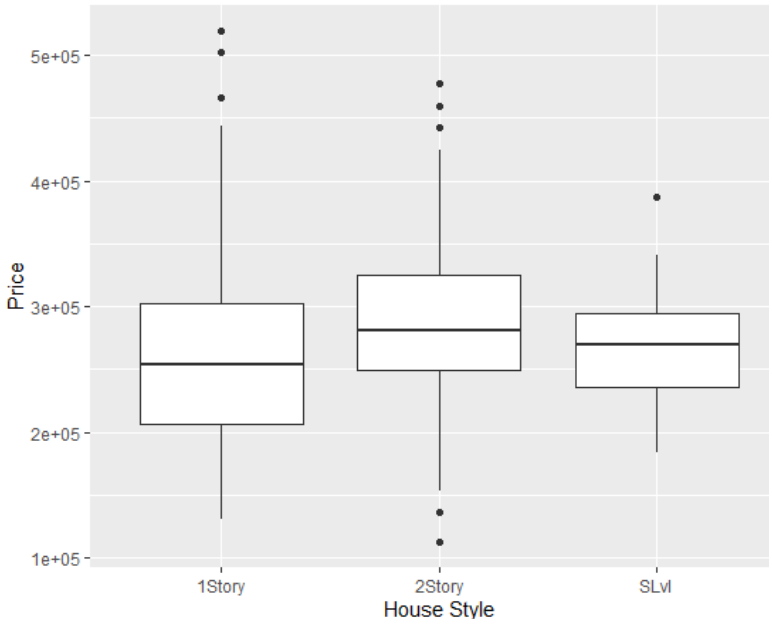

Figure 1

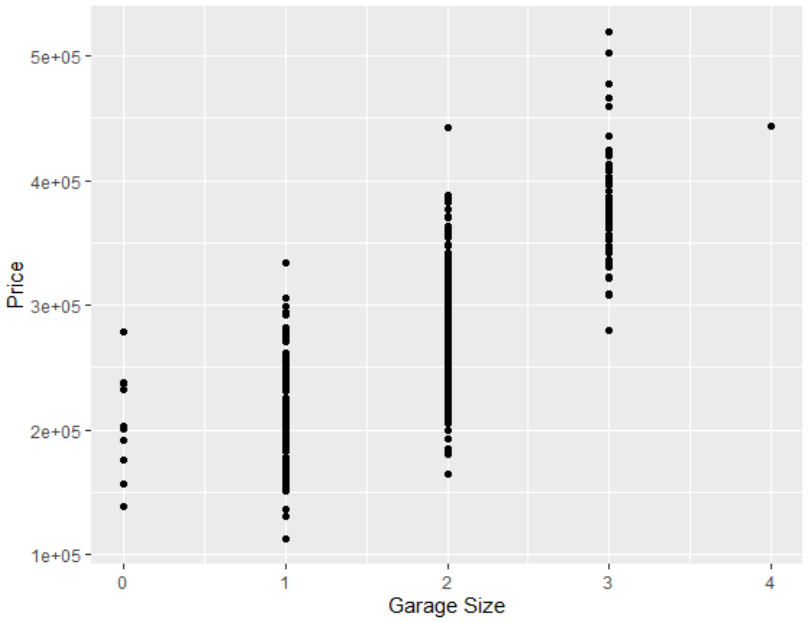

Figure 2

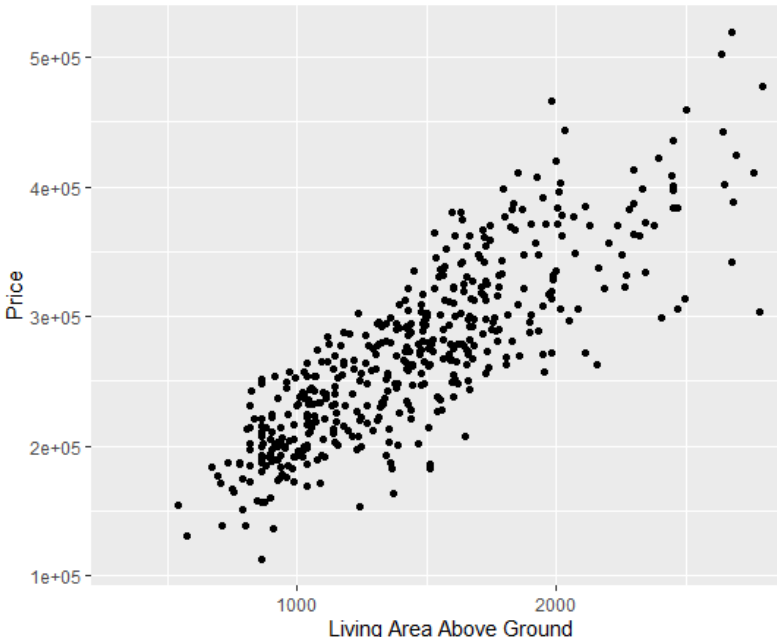

Figure 3

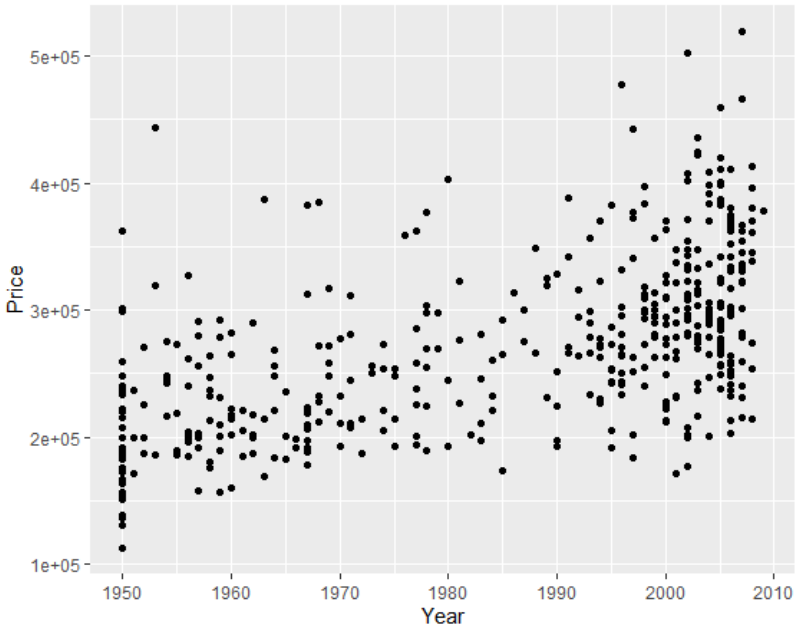

Figure 4

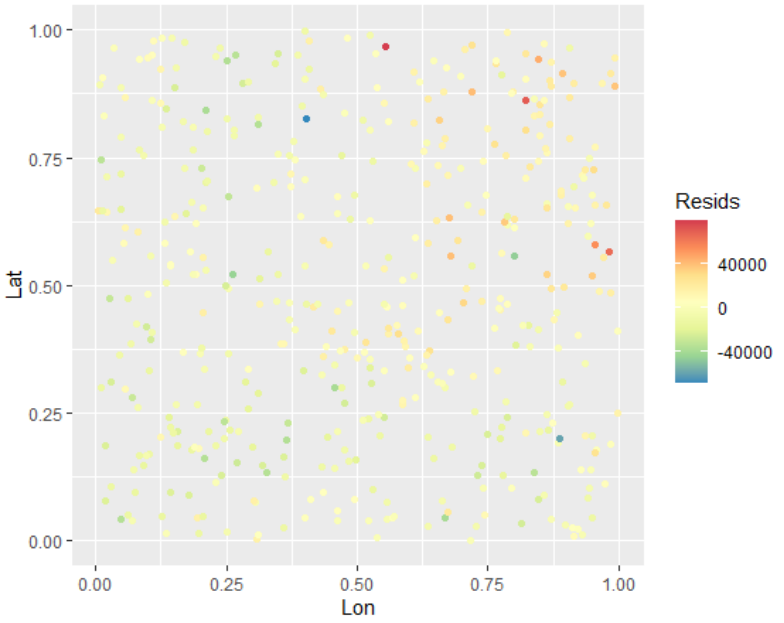

Figure 5

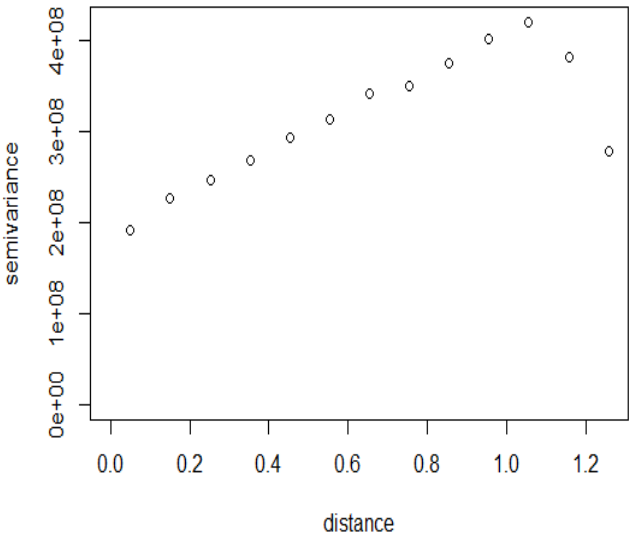

Figure 6

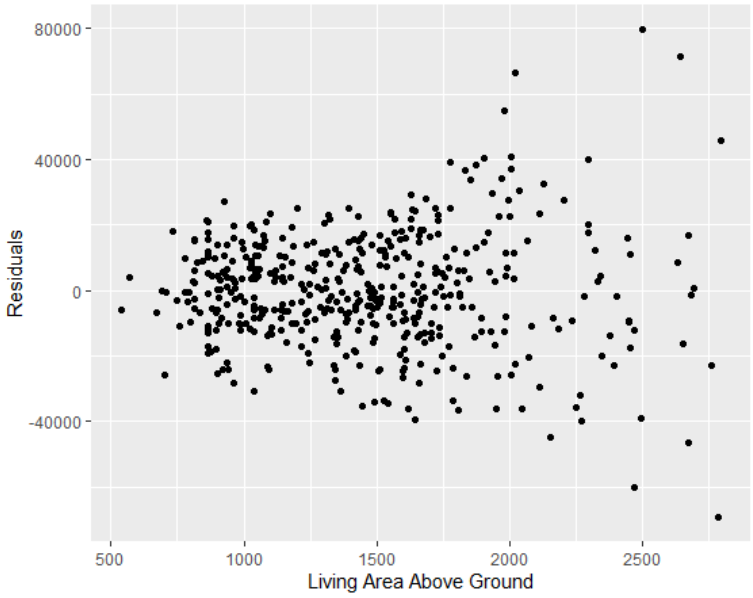

Figure 7

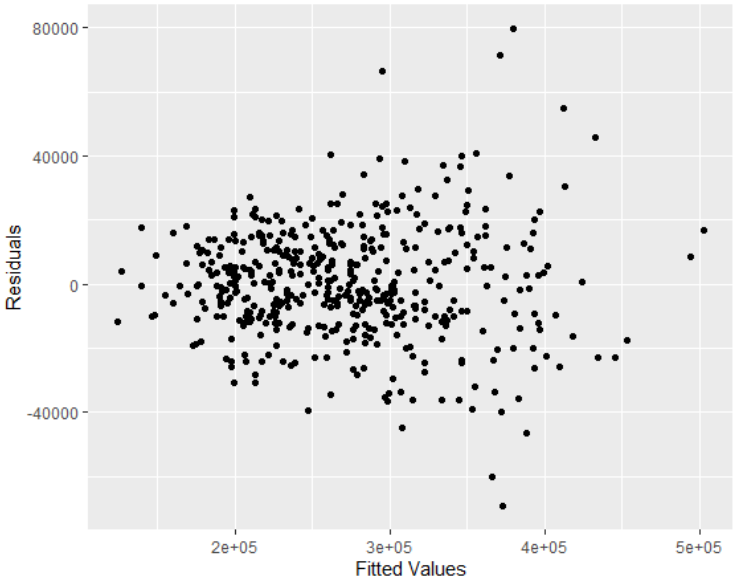

Figure 8

Statistical Model: Let y be an $N \times 1$ vector of the houses' sale prices. We assume that the response, y = Price, is normally distributed.

$$y \sim N(X\beta, \sigma^2 D^{½} R D^{½})$$

The design matrix is a 465 x 10 (removed missing rows from our original dataset) matrix of explanatory variables from our dataset. We have broken up House.style into indicator variables for each house style: Houst.Style1Story, House.Style2Story, and House.StyleSLvl.Styl.

$$X = \text{c}(1, X_{\text{Gr.Liv.Area}}, X_{\text{House.Style2Story}}, \ldots, X_{\text{Garage.Cars}})$$

β is a 10×1 vector of the coefficients for each explanatory variable of the appraisal.

$$\beta = (\beta_{0,} \beta_{\text{Gr.Liv.Area}}, \beta_{\text{House.Style2Story}}, \ldots, \beta_{\text{Garage.Cars}})$$

We can interpret the β coefficients as:

- Holding all else constant, for a 1 sq. foot increase in the living area above ground area we would expect a $\beta_{\text{Gr.Liv.Area}}$ increase in the sale price of the home.
- Holding all else constant, for a split-level home split-level we expect a $\beta_{\text{House.StyleSLvl.Styl}}$ increase in the appraisal as compared to an equivalent 1 story home.
- Holding all else constant, adding central air to a house we would expect a $\beta_{\text{Central.AirY}}$ increase to the sale price of the home.

The variance of the response is accounted for in three parts. Firstly, the spatial correlation is represented by R. We define the spatial correlation using the exponential correlation function with a nugget. If the distance $d^{(\rho)}_{ij}$ between two houses i and j are zero then the correlation is ω. If not, then spatial correlation is defined as $(1 - \omega) \exp(-d^{(\rho)}_{ij} / \varphi)$. ω is the nugget. The nugget allows for appraisal variability within the same location. The also nugget allows for the assumption that houses in the same location can vary in price even though the houses are some distance apart. φ is the range parameter. As φ increases the effect of spatial correlation diminishes less with distance.

To account for the heteroskedasticity within our data set, we use a diagonal matrix D. Along the diagonal of the matrix, the variance multiplier is $\exp(2X_{Gr.\text{Liv.Area}}\theta)$. When $\theta > 0$ the variance increase with the covariates and when $\theta < 0$ the variance decreases as the covariates increase. σ^2 is the residual standard error, which is amplified by the variance function that creates the diagonal matrix D. Now our model accounts for both spatial dependence and heteroskedasticity and can explain the housing price based on a linear combination of the explanatory variables. The next step is to check our model assumptions of linearity, equal variance, independence, and normality. Finally, we want to see how well it performs on test data.

Model Validation: Now that we have fitted our model, we need to verify it by checking our assumptions. By looking at Figure 9, we can see that the added variable plots show linearity. The added variable plots show the relationship between a single variable and the response after adjusting for the effects of all other variables. From Figure 10, we can see that our assumption of normality has been met as we have a normal distribution of the residuals. Figure 11 shows the decorrelated residuals plotted against the fitted values. The residuals seem to be distributed across the fitted values, so we have accounted for the heteroskedasticity within our model. To make sure of our assumption of normality, we performed a Kolmogorov-Smirnov test. The null hypothesis of the test is that the residuals do follow a standard normal distribution. The p-value is 0.9679, so we have failed to reject the null hypothesis.

We also want to verify that the spatial correlation has been accounted for. Figure 12 shows that the residuals across both latitude and longitude. We can see that the decorrelated residuals no longer appear to have any spatial patterns. To verify how well our model performs, we need to computer a pseud-R^2 since we used correlated data. We calculated our pseudo-R^2 by taking the correlation between our predicted values and actual values and squaring them. We found our pseudo- R^2 to be 0.93. This means that our model accounts for 93% of all variability in housing prices in Ames, Iowa. Our model is a great fit for our data, and we have accounted quite well for spatial correlation and heteroskedasticity within the data set.

Now that we have verified our model and assumptions, we want to use our model to calculate predictions for the missing housing prices. We first need to perform cross-validation to validate our predictions. We perform cross-validation by subsetting our data into two subsets. One is for training and another is for testing. We fit our model to the training data and use that model to predict the housing prices for the test data. By doing so, we can obtain estimates of our predictive bias, RPMSE, coverage, and prediction interval. We performed our cross-validation 100 times to estimate the true values bias, RPMSE, coverage, and predictions intervals. We have a bios of 351, which means we over predict by $351. We calculate our RPMSE to be $14,994 which indicates that our predictions for housing prices are 14,994 off. We have a coverage of 0.95 for our 95% prediction intervals, which is the normal coverage. Our prediction interval width is $50,000 is big. However, considering most homes are in the hundreds of thousands of dollars it’s not terrible.

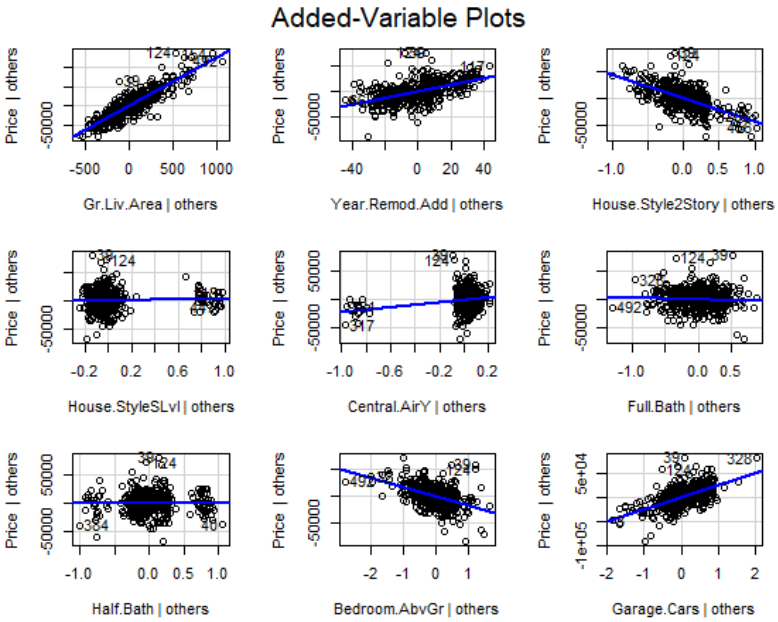

Figure 9

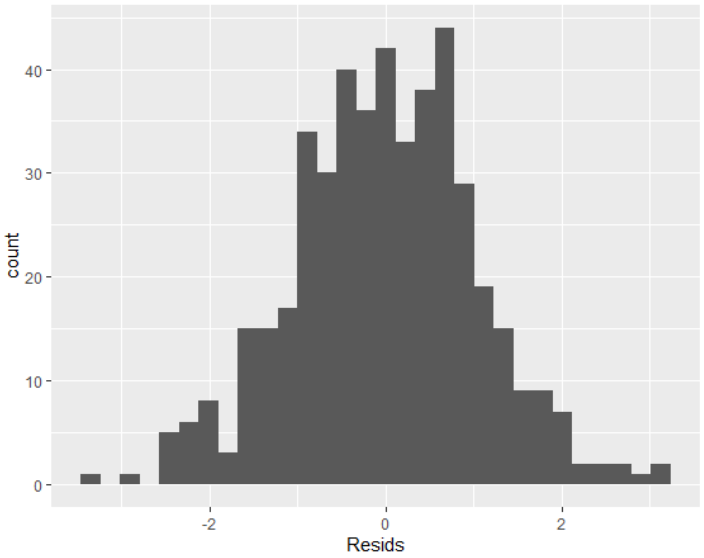

Figure 10

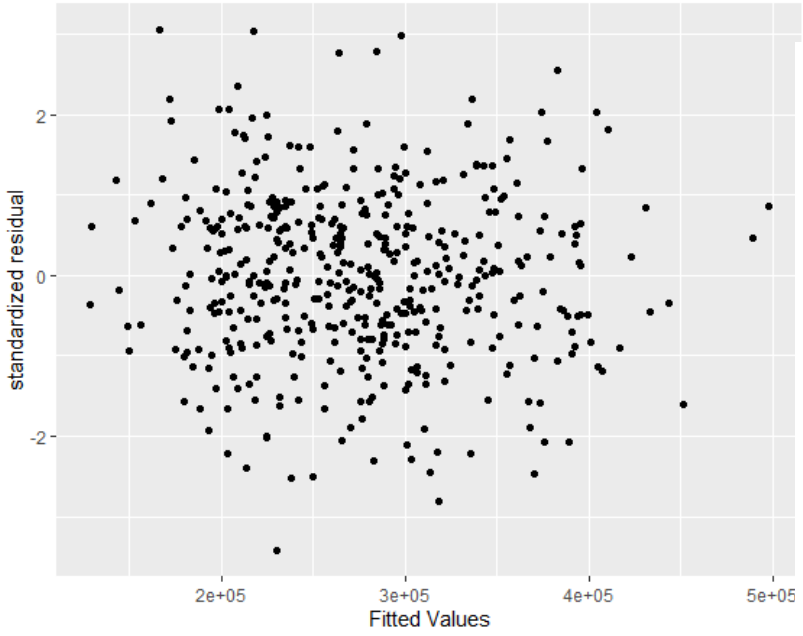

Figure 11

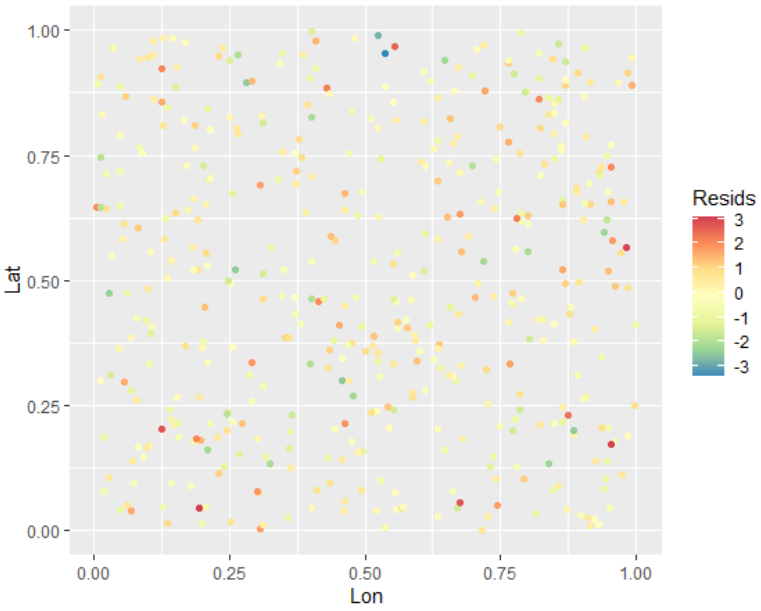

Figure 12

Analysis Results: As discussed in the last section, our pseudo-R^2 shows us that our explanatory variables explain the sale price of homes quite well. The estimates of our beta coefficients are given in Table 1 below. Holding all else constant, we estimate that for every square foot added to the house size, we should expect the sale price increase by $124.82. Holding all else constant, by having central air conditioning in a home should increase the appraisal by $21,336.08. the full bath, half bath, and split-level style all contain zero in their coefficient interval estimate. There is a negative effect on price by increasing the number of bathrooms. This seems unrealistic, so it might be that bathrooms becomes unsignificant when consider all other covariates. To test this hypothesis about bathrooms, we performed a likelihood ratio test of a model with and without the bathrooms. We got a p-value of 0.25, so we have failed to reject the null hypothesis. Therefore, the number of full baths or half baths does not affect the housing price. Unlike what we found in our exploratory analysis, we found that 2 story houses are less expensive. In summary, we can say that square footage, year remodeled, central air, and garage size all contribute to a higher appraisal. We have produced a table of the coefficients, a table of the variance, a table of our nugget, and a graph of the decorrelated residuals over their location.

Table 1

	Lwr	est	Upr
(Intercept)	-1427258.49	-1325883.61	-1224508.73
Gr.Liv.Area	116.71	124.82	132.93
Year.Remod.Add	662.75	714.95	767.15
House.Style2Story	-46484.75	-43094.53	-39704.32
House.StyleSLvl	-2960.51	767.01	4494.53
Central.AirY	17178.17	21336.08	25493.99
Full.Bath	-5743.98	-2493.70	756.58
Half.Bath	-2501.66	438.97	3379.60
Bedroom.AbvGr	-17168.60	-15386.90	-13605.19
Garage.Cars	21114.93	22888.76	24662.58

Table 2

	Lwr	est	Upr
theta	0.00062	0.00073	0.00084

Table 3

	Lwr	est	Upr
range	0.10	0.26	0.64
nugget	0.19	0.34	0.53

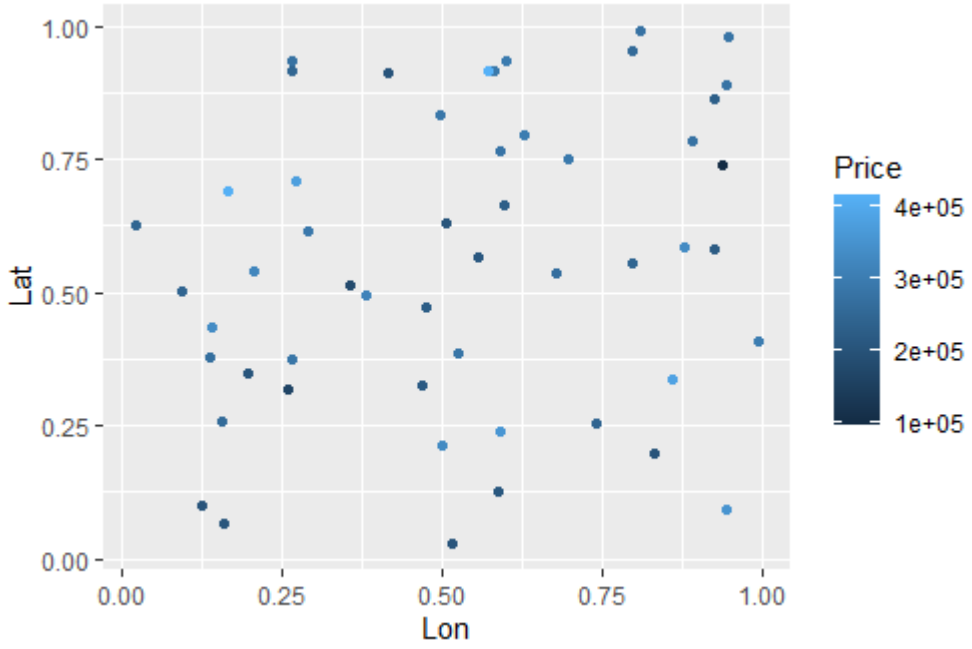

Figure 13

Conclusion: We can conclude after our analysis and validation that we can use house attributes to model the price of the home. House attributes of square footage, year built or remodeled, garage size, and central air should increase the value of a home. Looking at the style of home, we can see that 2 story homes should be less expensive. Additional bedrooms above the ground should also cause a home to be less expensive. The number of bathrooms does not significantly affect a home's appraisal when accounting for the other factors. We also know that homes with a larger area can vary more in price than smaller homes. Since home appraisers have past knowledge of appraising, a future analysis could use Bayesian models to fit the data.

Appendix:

```
#packages
library(tidyverse)
library(GGally)
library(nlme)
library(car)
library(geoR)
library(lmtest)
library(xtable)
library(DataExplorer)
source("stdres.R")
source("predictgls.R")

house <- read_csv("HousingPrices.csv")

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

plot_missing(house)
house.omit <- house[!is.na(house$Price), ]

#graphs
ggpairs(house[,-c(2:3)])

#price vs Lon
plot(house$Lon, house$Price)

#price vs Lat
plot(house$Lat, house$Price)

#price vs Gr.Liv.Area
ggplot() + geom_point(aes(y=Price, x=Gr.Liv.Area), data=house) + xlab("Living Area Above
Ground")

#Pirce vs Year.Remod.Add
ggplot() + geom_point(aes(y=Price, x=Year.Remod.Add), data=house) + xlab("Year")
```

```
#price vs House.Style
ggplot() + geom_boxplot(aes(y=Price, x=House.Style), data=house) + xlab("House Style")

#price vs Garage.Cars
ggplot() + geom_point(aes(y=Price, x=Garage.Cars), data=house) + xlab("Garage Size")

#price vs Full.Bath
plot(house$Full.Bath, house$Price)

#price vs Half.Bath
plot(house$Half.Bath, house$Price)

#change baths into one
house$Bath <- house$Full.Bath + house$Half.Bath/2

#price vs Bath
plot(house$Bath, house$Price)

#linear model
house.lm  <- lm(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air +
Full.Bath + Half.Bath + Bedroom.AbvGr + Garage.Cars , data=house)

summary(house.lm)
residuals <- house.lm$residuals

#plot of residuals across the region
ggplot(data=house[!is.na(house$Price),], mapping=aes(x=Lon, y=Lat, col=house.lm$residuals))
+
geom_point() + scale_color_distiller(palette="Spectral") +
xlab("Lon") + ylab("Lat") + labs(col="Resids")

#heteroskedasticity
ggplot() + geom_point(aes(x=house.lm$fitted.values, y=house.lm$residuals)) + xlab("Fitted
Values") + ylab("Residuals")
ggplot() + geom_point(aes(x=house.omit$Gr.Liv.Area, y=house.lm$residuals)) + xlab("Living
Area Above Ground") + ylab("Residuals")

#variogram
vargram <- variog(coords=house[!is.na(house$Price),2:3], data=residuals)
plot(vargram)

#exp model
cor.exp <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air + Full.Bath
+
            Half.Bath + Bedroom.AbvGr + Garage.Cars,
```

```
            data=house, subset=!is.na(Price),
            correlation = corExp(form=~ Lon + Lat, nugget = TRUE),
            weights = varExp(form=~Gr.Liv.Area), method="ML")

#gaus model
cor.gaus <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air + Full.Bath
+
              Half.Bath + Bedroom.AbvGr + Garage.Cars,
            data=house, subset=!is.na(Price),
            correlation = corGaus(form=~ Lon + Lat, nugget = TRUE),
            weights = varExp(form=~Gr.Liv.Area), method="ML")

#spher model
cor.spher <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air +
Full.Bath +
               Half.Bath + Bedroom.AbvGr + Garage.Cars,
             data=house, subset=!is.na(Price),
             correlation = corSpher(form=~ Lon + Lat, nugget = TRUE),
             weights = varExp(form=~Gr.Liv.Area), method="ML")

#AIC values for each model
AIC(cor.spher, cor.gaus, cor.exp)

#fit gls model
house.gls <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air +
Full.Bath + Half.Bath + Bedroom.AbvGr + Garage.Cars,
             data=house.omit,
             correlation = corExp(form=~ Lon + Lat, nugget = TRUE),
             weights = varExp(form=~Gr.Liv.Area), method="ML")

#gls without baths
house.gls2 <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air +
Bedroom.AbvGr + Garage.Cars,
              data=house.omit,
              correlation = corExp(form=~ Lon + Lat, nugget = TRUE),
              weights = varExp(form=~Gr.Liv.Area), method="ML")
#summary
summary(house.gls)

#anova between two gls models
anova(house.gls, house.gls2)

#intervals
intervals <- intervals(house.gls)

#time
```

```
system.time({
house.gls <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add + House.Style + Central.Air +
Full.Bath +
                  Half.Bath + Bedroom.AbvGr + Garage.Cars,
                data=house.omit,
                correlation = corExp(form=~ Lon + Lat, nugget = TRUE),
                weights = varExp(form=~Gr.Liv.Area), method="ML")
})

#CV
pb <- txtProgressBar(min = 0, max = 100, style = 3)
n.cv <- 100
n.test <- 200
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:nrow(house.omit), size=n.test)

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

  ## Fit a lm() using the training data
  train.lm <- house.gls <- gls(Price ~ Gr.Liv.Area + Year.Remod.Add +
                                 House.Style + Central.Air + Full.Bath +
                                 Half.Bath + Bedroom.AbvGr + Garage.Cars,
                               data=house.omit,
                               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.lm, newdframe=test.set)

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

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

  ## Update the progress bar
  setTxtProgressBar(pb, cv)
}
close(pb)

#gls cv
hist(rpmse)
hist(bias)
hist(wid)
hist(cvg)
#residuals of gls model
std.resid <- stdres.gls(house.gls)

#equal variance
ggplot() + geom_point(aes(fitted(house.gls), std.resid)) + ylab('standardized residual') +
xlab("Fitted Values")

#independence
ggplot(data=house[!is.na(house$Price),], mapping=aes(x=Lon, y=Lat, col=std.resid)) +
geom_point() + scale_color_distiller(palette="Spectral") +
xlab("Lon") + ylab("Lat") + labs(col="Resids")

#normality
ggplot() + geom_histogram(aes(std.resid)) + xlab("Resids")
ks.test(resid(house.lm), "pnorm")
ks.test(std.resid, "pnorm")

#linearity
avPlots(house.lm)

#Pseudo-R^2
cor(fitted(house.gls), house$Price[!is.na(house$Price)])^2

#predictions of na homes
pred <- predictgls(house.gls, newdframe = house[is.na(house$Price),-12])

#plot of prediction across the region
ggplot(aes(Lon, Lat, col = Prediction), data = pred) +
  geom_point() + xlab("Lon") + ylab("Lat") + labs(col="Price")
```