Home Appraisals Analysis

STAT 469: Analysis of Correlated Data

Nick Hass and Tyler Ward
April 20th, 2022

Executive Summary

The following report presents an in-depth statistical analysis of home prices in Ames, Iowa. We have developed a statistical model based on basic home characteristics that explains 93% of the variability in home appraisals. This model shows that certain characteristics of a home lead to higher sale price value. The home characteristics that we've identified that lead to higher sales price include: above-ground living area, split-level homes, remodeling date, central air, larger car garages, and more half bathrooms. Our statistical model predicted home sales price for several homes in Ames, and can be used to predict any home's appraisal sales price in Ames given the home's characteristics.

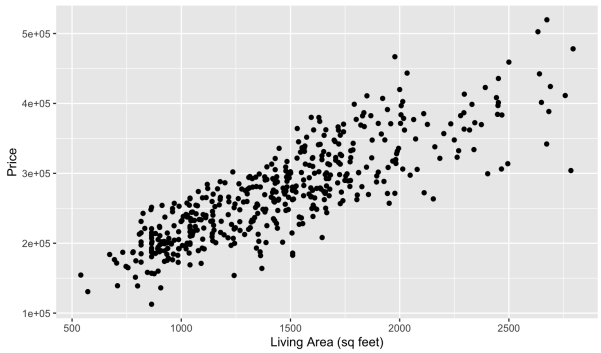
Introduction and Problem Background

With home demand and housing prices increasing every year, unbiased determination of the fair market value of a home is in high demand. This fair-market price of a home is based on a variety of factors, including amount of living area, style of dwelling, air conditioning status, construction or remodel date, number of bathrooms and bedrooms, and size of garage. All of these variables are included in a dataset of various homes in the Ames, IA area, which we will be analyzing.

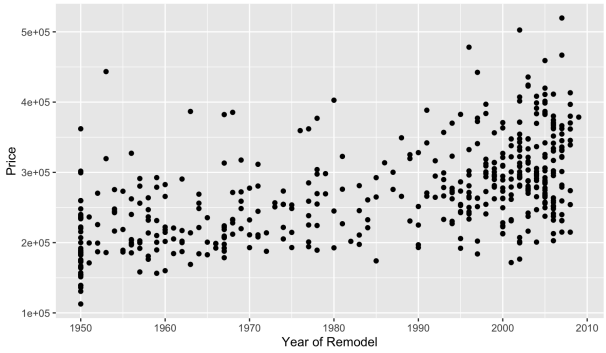
The goal of this study is to use statistical models to determine both how well home characteristics explain home sale price, and which of these characteristics significantly increase the home sale price. We will also discuss how the variability of home price changes with respect to home land area. Finally, we will use our best statistical model to appraise homes in our dataset without a sales price, showing how statistical models can be used to make home-appraisals in the world we live in today with a great degree of accuracy.

Looking at the data through a series of plots shown on the page below, we see several general relationships present in the data.

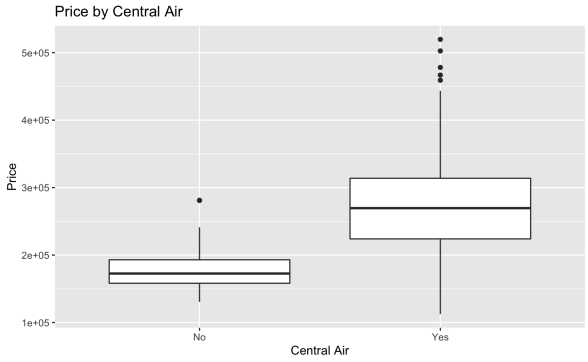
Price by Living Area



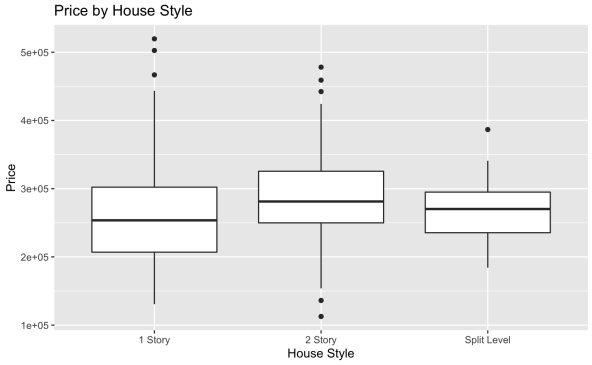
Price by Remodel Date



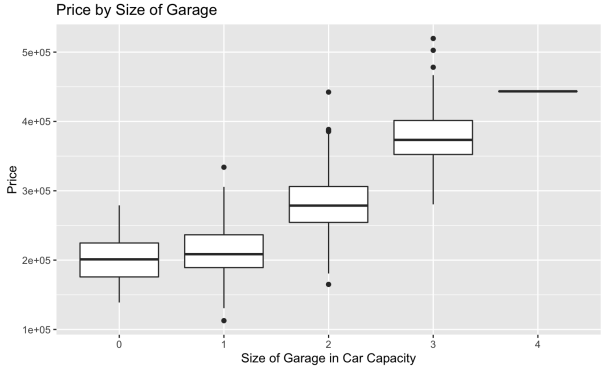
Price by Central Air



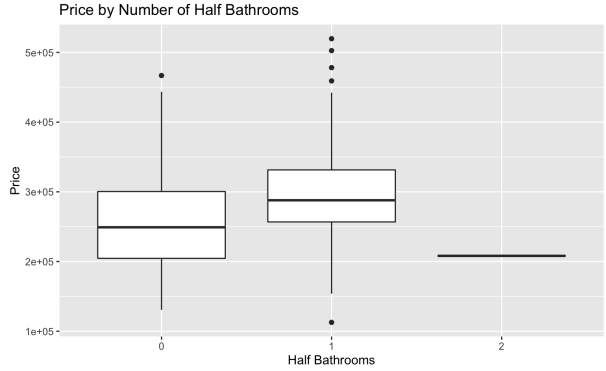
Price by House Style



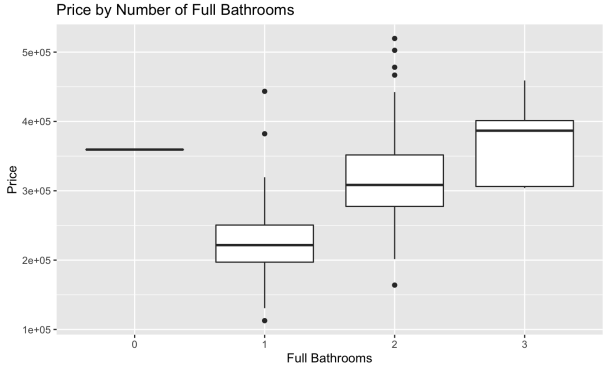
Price by Size of Garage



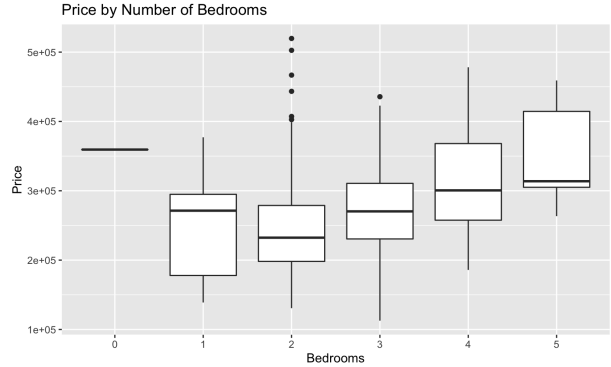
Price by Number of Half Bathrooms



Price by Number of Full Bathrooms



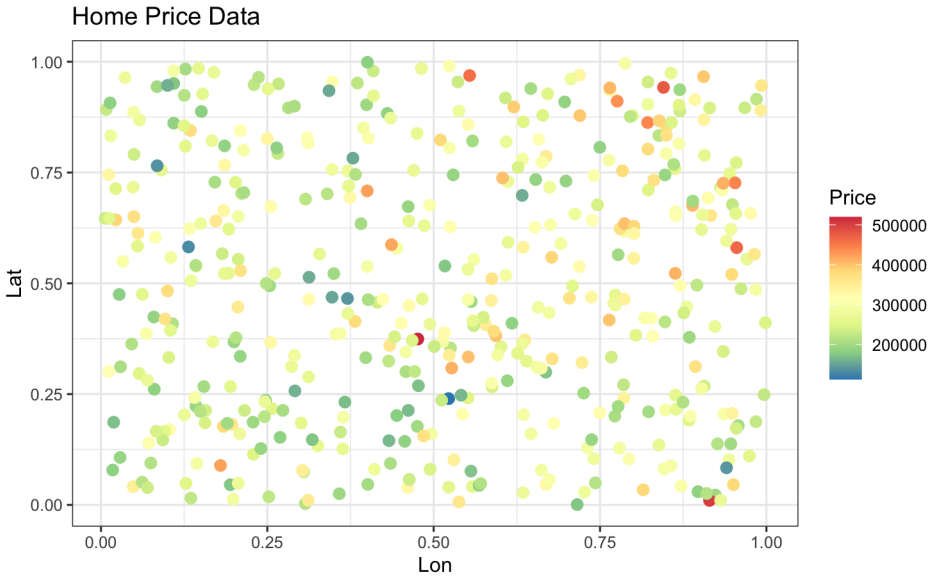
Price by Number of Bedrooms



First, The price by living area scatterplot shows a positive relationship. As the living area increases, home prices also increase. In the same plot, as shown by the funnel shape, as living area increases, it seems that the variability in home price also increases. We will examine this further as we go through our analysis. The price by remodel date scatter plot also shows a positive linear relationship. In general, as homes are newer (or have a newer remodel date), home prices also increase on average. It is also clear that most homes have either been remodeled somewhere close to 1950 or somewhere close to the year 2000, with those remodeled closer to 2000 having a higher home price, on average.

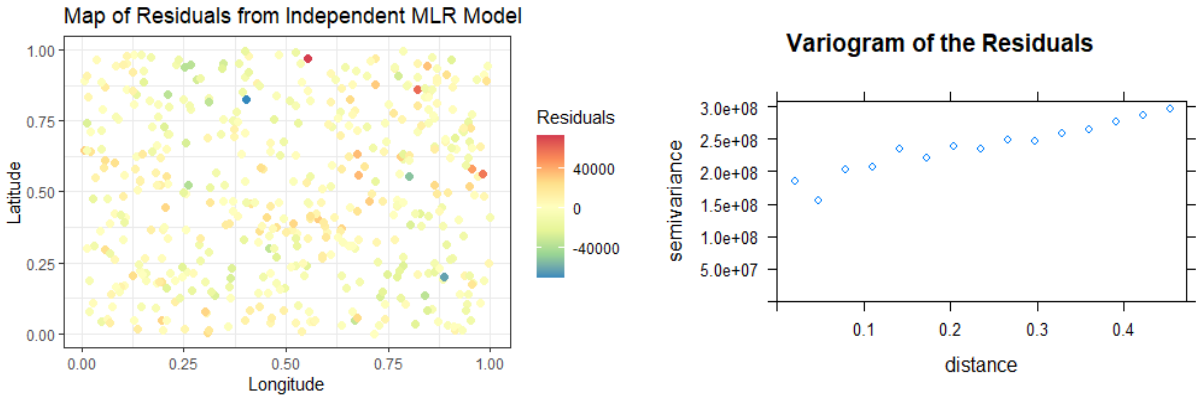
The price by central air box plot shows that having central air is associated with higher home prices. It also shows that there is a larger range of home prices for homes with central air. The price by house style box plot shows that there are different ranges of home prices associated with house style. There may be some evidence that 2-story homes are associated with higher home prices, but more will be investigated when we fit a model.

The price by size of the garage box plot shows a positive linear relationship in the distributions of each size of garage. Homes with a larger garage seem to be associated with higher home prices. The price by number of full bathrooms box plots also show a positive linear relationship in its distributions of bathrooms, with more full bathrooms in a home associated with higher home prices. On the other hand, the price by number of half bathrooms box plot shows that there is little difference in home prices on average for homes with 0 or 1 half bathrooms. The price by number of bedrooms box plot shows a positive linear relationship. More bedrooms in a home are associated with higher home prices.

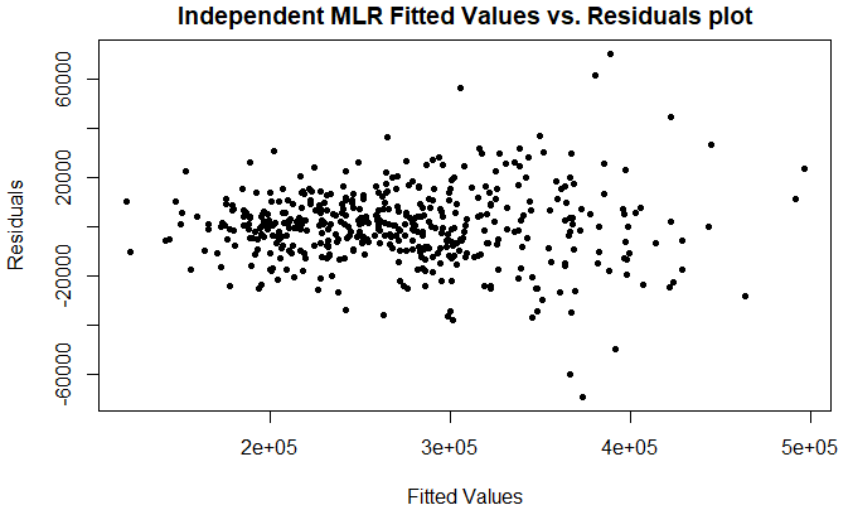


Along with these relationships, the plot of home prices at each latitude and longitude above shows that the homes in our data are close geographically and homes nearby each other seem to be similar in price. This raises the concern of potential spatial correlation between the sales prices of homes that are nearby each other. Plotting the residuals from an independent MLR model with no correlation structure confirms that houses close together seem to be highly

correlated with each other. Further evaluating this with a variogram, these residuals of a model without any correlation structure shows clear evidence of spatial correlation, with a positive linear trend in semivariance present as distance increases. We will want to have the variogram of our final model's residuals be void of this positive linear trend, so that we can confirm that this spatial correlation has been accounted for.



Additionally, as mentioned before after looking at the scatterplot of living area vs. home price, there seemed to be heteroscedasticity present with the spread of home prices increasing as land area values increase. Plotting a fitted values vs. residuals plot of our independent MLR model similarly shows signs of unequal variance in the higher fitted values, as seen in the plot below, confirming our concern of heteroscedasticity.



If we ignore accounting for the spatial correlation present in our data, our model will be unreliable, with our predictions, effects of each of the home characteristics, as well as confidence intervals corresponding to these, thrown off by this correlation. However, by accounting for this correlation, we can achieve the goals of our analysis, and use this correlation to make our predictions much more accurate than they would otherwise.

Furthermore, failure to account for the unequal variance in our model will throw off our models standard error, which will affect confidence intervals, prediction, and hypothesis testing. However, by accounting for this unequal variance using a variance function, we will be able to get a measurement of how much variability in home price varies as land area changes, and will be able to use our model to make inferences using confidence intervals and testing, along with making predictions that can be used for home appraising.

Therefore, we will analyze our data using a heteroscedastic spatial linear regression model. After fitting our model with several spatial correlation structures, a model with exponential correlation structure was deemed best in terms of AIC. We will also use an exponential variance function in relation to land area to account for unequal variance in home price as land area increases. This model is described in mathematical detail below.

Statistical Model

Heteroscedastic MLR Model with Exponential Spatial Correlation Structure:

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\Sigma} = \boldsymbol{DRD}$$

where

- $\boldsymbol{Y}$ is the vector of home price observations for a homes at given latitude-longitude locations in Aimes, Iowa
- $\boldsymbol{\beta}$ is the vector of the effects of each of the home characteristics. These effects give important interpretations in relation to home sales price, our models response variable. We interpret a few of these effects below:
    - $\beta_{FullBath}$ intuitively tells us that, holding all other variables constant, as the number of full bathrooms in a home increases, the sales price of a home will increase by $\beta_{FullBath}$, on average
    - $\beta_{CentralAir=YES}$ intuitively tells us that homes *with* Central Air will have a $\beta_{CentralAir=YES}$ ***higher*** home sales price than a home *without* central air
- $\boldsymbol{D} = diag(\sigma_1, ..., \sigma_n)$, as $\boldsymbol{D}$ is a diagonal matrix with values of an exponential variance function with respect to land area ("Gr.Liv.Area") along the diagonal; namely $dii = \exp\{2x_i\theta\}$ where x_i is values of land area and θ is a measure of how much home sales price varies as land area increases. I.e., if θ is positive, then as land area increases, the variance of sales price also increases. If θ is negative, the opposite is true.
- $\boldsymbol{R}$ is a diagonal matrix with ones along the diagonal, and values of the exponential correlation function $\rho(s_i, s_j) = \exp\{-\frac{||s_i - s_j||}{\phi}\}$ filling the rest of the matrix. This correlation structure uses a variance "nugget" in order to capture same-location

variability, which would not be captured otherwise by our exponential correlation function alone. This variance nugget is applied using the formula below:

- $Cor(s_i, s_j) = (1 - \omega)\rho(s_i, s_j)$, where ω is estimated from the data using iterative optimization

Model Assumptions

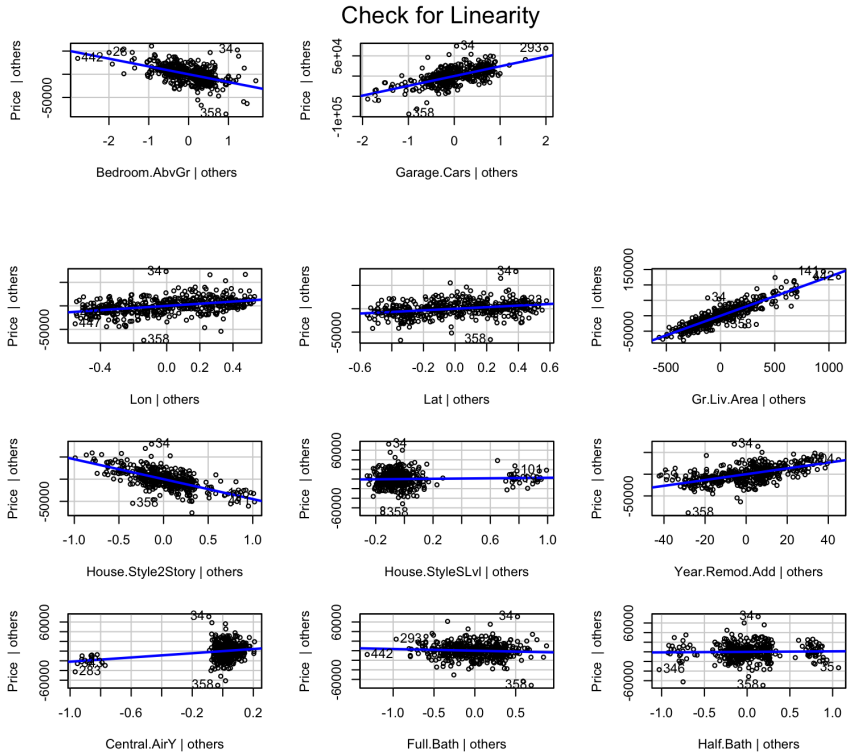
Our model operates with four main assumptions. These include assumptions of linear relationships between the variables used as effects and response, independence of variables *after accounting for spatial correlation*, normal distribution of the residuals of our model, and equal variance of the residuals of our model, after accounting for the unequal variance present from changes in home land area.

Model Validation

Here we justify our models assumptions using the plots and graphics below.

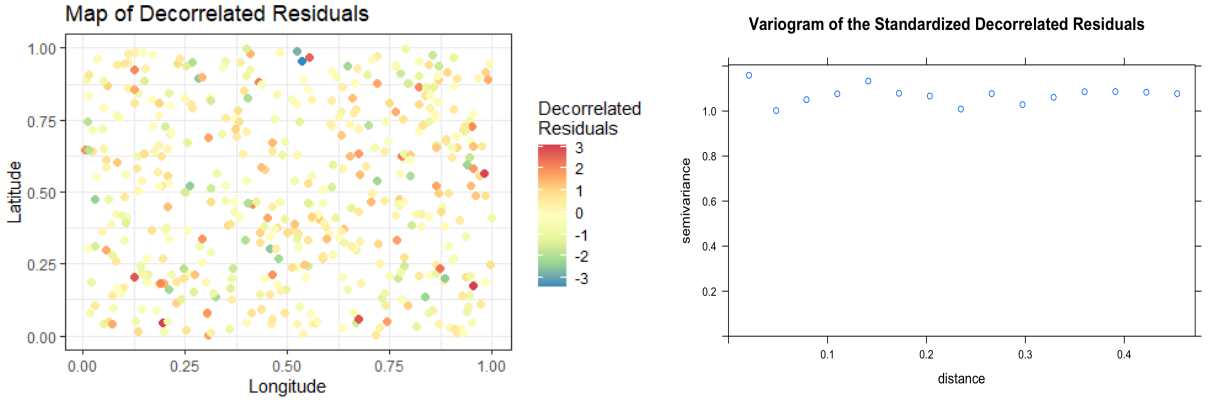
Linearity

To access our linearity assumption, we created added variable plots of the variables included in our model. These plots show linear relationships for all variables, and we can therefore reasonably assume our assumption of linearity holds true.



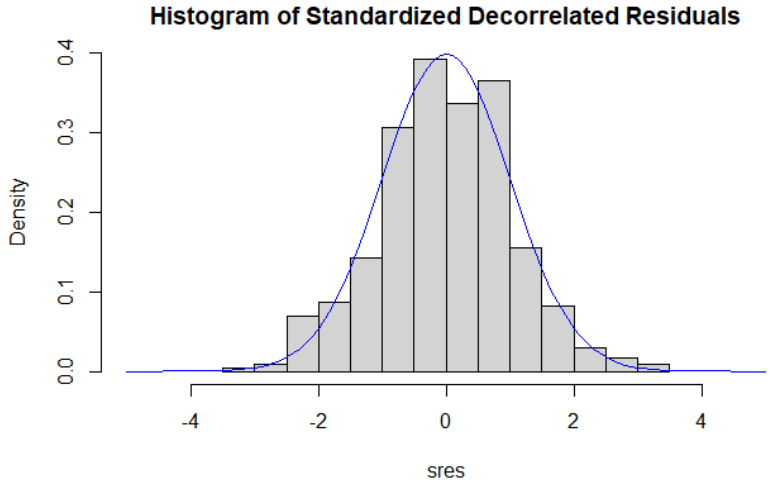
Independence

In order to see if our model fully accounted for spatial correlation, we plot a variogram of its standardized and decorrelated residuals below. This variogram shows no major trend in semivariance as distance increases (all data points, no matter the distance, show a high degree of variability), and a plot of the standardized decorrelated residuals shows much less correlation by location, so we can reasonably assume that there is independence after applying an exponential correlation function with a variance nugget.



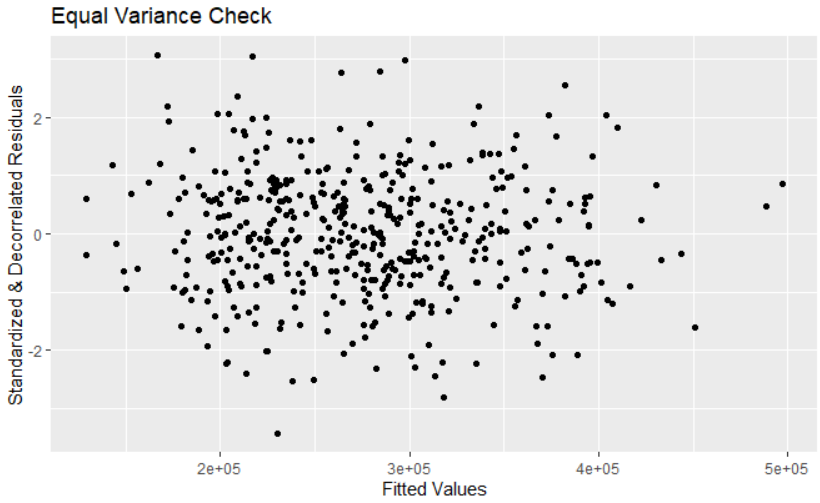
Normality of Residuals

A histogram of our model's decorrelated residuals shown below seems to be relatively normally distributed. Our assumption of normality is therefore validated.



Equal Variance of Residuals

After accounting for the homoscedasticity in home price with an exponential variance function, a scatterplot our model's fitted values vs. the decorrelated residuals shows no longer shows major signs of unequal variance, and so we can reasonably assume our equal variance assumption is now validated.



Model Fit and Predictive Accuracy

Our model fits the data well, explaining over 93% of the variation in home sales price based on the calculation of pseudo R^2. In cross validation we calculated a root predicted mean square error of 13,002.20, meaning that our predictions of housing prices are only about $13,000 dollars off, on average. This is relatively small in the context of housing prices in Ames—less than 5% of the median home price in Ames. Confirming this, our cross validation gave a bias of only $\approx$ 63.42, meaning overall our predictions are just barely higher than observed price values, on average. Additionally, our model performed with an average coverage of $\approx$ 96.48%, meaning that on average, taking many prediction intervals, approximately 96 percent of them will contain the true home price value. Therefore, we can reasonably conclude that our model does very well at predicting home prices.

Analysis Results

After verifying the predictive accuracy of our model along with validating its assumptions, we used our model to fulfill the goals of our analysis.

Effect of Home Characteristics

First, a calculation of model's pseudo R^2 tells us that home characteristics explain approximately 93.29 % of the variability in home prices. Therefore, home characteristics explain home prices very well in Aimes, Iowa.

In addition, several home characteristics lead to an increased sales price as shown below in Table 1. These factors include *living area* (in square feet above ground), a more recent *remodel year*, having *central air*, and *garage car capacity*. As these characteristics increase, the sale price of a home also increases on average. These are the factors in the table that have strictly positive 95% confidence intervals (i.e., they do not contain 0). If the confidence interval contained 0, then there would still be a reasonable chance that the home characteristic is not a significant factor in increasing the sale price of a home.

Table 1: 95% Confidence intervals for Home Characteristic Effects

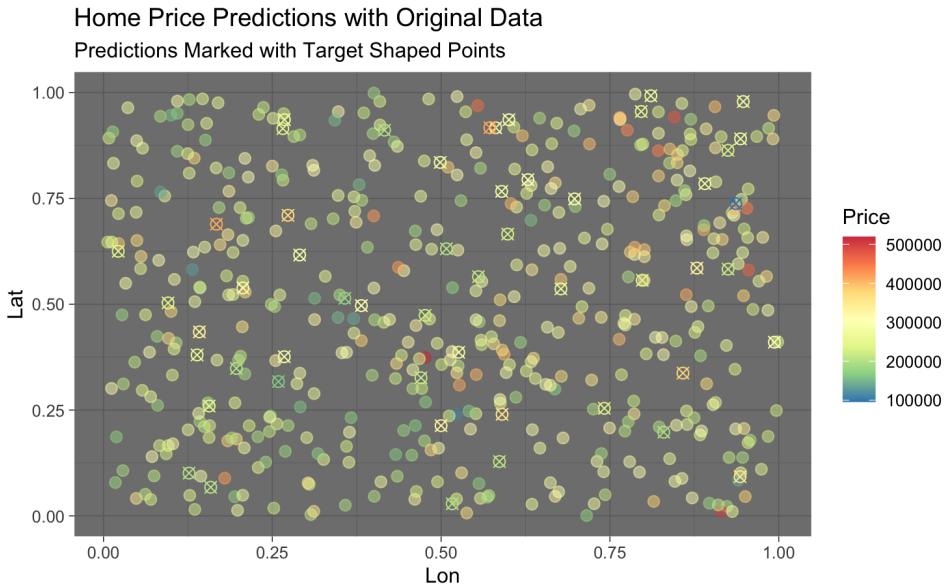
	Lower Bound	Upper Bound
Gr.Liv.Area	116.73	132.90
House.Style2Story	-46475.73	-39713.34
House.StyleSLvl	-2950.60	4484.62
Year.Remod.Add	662.89	767.01
Central.AirY	17189.23	25482.93
Full.Bath	-5735.33	747.94
Half.Bath	-2493.84	3371.78
Bedroom.AbvGr	-17163.86	-13609.93
Garage.Cars	21119.65	24657.86

Variability in Sales Price

As noted in our introductory notes, a fitted vs. residuals plot of a multiple linear regression model with no account of unequal variance shows changing variance (heteroscedasticity). Additionally, our EDA scatter plot of home price by land area seemed to have growing variance in home price as home size (land area) increased. From our model summary, we see that an estimate of the exponential variance parameter $\hat{\theta}$ is positive with a value of 0.000731. This confirms that on average, as the size of a home increases, the variance of home prices also increases. Therefore, the variance of sales price *does* increase with the size of a home as we suspected in our exploratory analysis.

Predicted Sales Prices

Finally, below is a map of the predicted sales price for homes in the data set that did not have a sales price labeled. The predictions are marked with target shaped points and the homes that already had labeled home prices are marked with a color filled circle.



Conclusions

Summary

In this analysis, we studied four questions: how well do the home characteristics explain sale price? What factors increase the sale price of a home? Does the variability of sale price increase with the size of a home? What is our predicted sales price for the homes in the data set without a sale price? Our main findings of this home appraisal analysis were in inference and prediction. We made a statistical model that explains 93% of the variability in home prices using geographic location and home characteristics. These home characteristics therefore did very well at explaining home sale prices. We found that the specific home characteristics that lead to an increased home sale price are living area, a more recent remodel year, having central air, and garage car capacity. We also found that the variability of sales price increases as the size of a home increases. Finally, we made predictions for the price of the homes in the Ames data set that did not have prices originally.

Next Steps

We are satisfied with our low root mean square error, and instead of making marginal improvements to the prediction accuracy, we would like to increase the scope of the analysis. In the next steps, we would like to add a time series component to this analysis. Home prices in the US are always fluctuating, but have skyrocketed since 2020 and have made us aware of the impact of home appreciation. This rapid appreciation in home value has impacted everyone across the country. It would be interesting to collect data on these home prices over time, to be able to predict future prices. How might the home be valued in one year? In five years? This would help those purchasing homes understand how much the home would appreciate or depreciate in value over the time. This will lead to people making better decisions about planning purchases of homes.

We are interested in combining this new time series component with similar analysis to what we have presented in this report. In order to present the future predictions, a dynamic 3-dimensional plot would be required such that we retain a spatial plot using longitude and latitude on the x and y axis, but additionally plot time on the z-axis.

Code

```
# libraries functions and options
library(tidyverse)
library(nlme) # for gls
library(intervals)
library(gstat) # for variogram
library(tinytex) # for LaTex in pdf
source("glstools-master/stdres.gls.R") # decorrelate residuals
function
source("glstools-master/predictgls.R") # predictgls function
options(scipen=999)

# data import
homes <- read_csv("HousingPrices.csv")
# Identify factor variables -- I found that using too many breaks the
gls
# make these factor variables,
# these two are factor variables
homes$House.Style <- as.factor(homes$House.Style)
homes$Central.Air <- as.factor(homes$Central.Air)

# the rest we will observe as factor variables to make explanatory
boxplots,
# i.e., use as.factor() in the ggplot call
# then they can be used as continuous variables in our model

#homes$Garage.Cars <- as.factor(homes$Garage.Cars)
# homes$Full.Bath <- as.factor(homes$Full.Bath)
# homes$Half.Bath <- as.factor(homes$Half.Bath)
# homes$Bedroom.AbvGr <- as.factor(homes$Bedroom.AbvGr)

# Separate into data set with and without missing values
homes_na <- homes %>% filter(is.na(Price))
homes_no_na <- homes %>% filter(!is.na(Price))

# First map
ggplot(data = homes, aes(x = Lon, y = Lat)) +
  geom_point(aes(col = Price), size = 2.5) +
  scale_color_distiller(palette = "Spectral", na.value = NA) +
  labs(title = "Home Price Data") +
  theme_bw()

# EDA Plots
ggplot(homes_no_na, mapping = aes(x = Gr.Liv.Area, y = Price)) +
```

```
  geom_point() +
  labs(title = "Price by Living Area",
       x = "Living Area (sq feet)")

# Newer homes tend to be more expensive
ggplot(homes_no_na, mapping = aes(x = Year.Remod.Add, y = Price)) +
  geom_point() +
  labs(title = "Price by Remodel Date",
       x = "Year of Remodel")

ggplot(homes_no_na, mapping = aes(x = Central.Air, y = Price)) +
  geom_boxplot() +
  labs(title = "Price by Central Air",
       x = "Central Air") +
  scale_x_discrete(labels = c("N" = "No", "Y" = "Yes"))

ggplot(homes_no_na, mapping = aes(x = House.Style, y = Price)) +
  geom_boxplot() +
  labs(title = "Price by House Style",
       x = "House Style") +
  scale_x_discrete(labels = c("1Story" = "1 Story", "2Story" = "2
Story",
                              "SLvl" = "Split Level"))

# Larger garages tend to be more expensive homes
ggplot(homes_no_na, mapping = aes(x = as.factor(Garage.Cars), y =
Price)) +
  geom_boxplot() +
  labs(title = "Price by Size of Garage",
       x = "Size of Garage in Car Capacity")

ggplot(homes_no_na, mapping = aes(x = as.factor(Half.Bath), y =
Price)) +
  geom_boxplot() +
  labs(title = "Price by Number of Half Bathrooms",
       x = "Half Bathrooms")

ggplot(homes_no_na, mapping = aes(x = as.factor(Full.Bath), y =
Price)) +
  geom_boxplot() +
  labs(title = "Price by Number of Full Bathrooms",
       x = "Full Bathrooms")

ggplot(homes_no_na, mapping = aes(x = as.factor(Bedroom.AbvGr), y =
Price)) +
```

```
  geom_boxplot() +
  labs(title = "Price by Number of Bedrooms",
       x = "Bedrooms")

# Simple Model
homes.lm <- lm(Price ~ ., data = homes_no_na)

# variogram to look for spatial correlation
myVariogram <- variogram(object = Price ~ Gr.Liv.Area + House.Style +
                                   Year.Remod.Add + Central.Air +
Full.Bath +
                                   Half.Bath + Bedroom.AbvGr +
Garage.Cars,
                         locations = ~Lon + Lat,
                         data = homes_no_na)
plot(myVariogram, main = "Variogram of the Residuals") # There is
spatial correlation.

# get fitted values and residuals from Independent MLR model
home.lm.fitted.vals <- fitted(homes.lm)
home.lm.resids <- resid(homes.lm)

# plot fitted values vs. standardized residuals
plot(home.lm.fitted.vals, home.lm.resids,
     main = "Independent MLR Fitted Values vs. Residuals plot",
     pch = 20, xlab = "Fitted Values", ylab = "Residuals")

# add residuals to original df and plot these residuals
homes$lmresid <- rep(NA,nrow(homes))
homes$lmresid[!is.na(homes$Price)] <- home.lm.resids

# plot a map of the residuals to look for spatial correlation
ggplot(data=homes,mapping=aes(x=Lon, y=Lat, color=lmresid)) +
geom_point(size = 2) +
scale_color_distiller(palette="Spectral",na.value=NA) +
  labs(title = "Map of Residuals from Independent MLR Model",
       color = "Residuals", x = "Longitude", y = "Latitude") +
  theme_bw()

## Model Validation
# Spatial MLR Model with heteroskedasticity: homes.lm.hetero
# fit model
homes.lm.hetero <- gls(model = Price ~ Gr.Liv.Area + House.Style +
                                 Year.Remod.Add + Central.Air +
Full.Bath +
```

```
                                Half.Bath + Bedroom.AbvGr +
Garage.Cars,
    data=homes_no_na,  # data without missing values
    weights = varExp(form=~Gr.Liv.Area), # D part (Heteroskedastic)
    correlation = corExp(form = ~Lon+Lat, nugget = TRUE), # R part
(Spatial corr)
    method="ML") ## with factor central air and house style

# Linearity: avPlots
car::avPlots(homes.lm, ask =FALSE) # added ask = FALSE so code runs
faster

# Independence by decorrelating residuals and looking at the
variogram of decorrelated residuals.
sres <- stdres.gls(homes.lm.hetero) # Decorrelate residuals
residDF <- data.frame(Lon=homes_no_na$Lon, Lat=homes_no_na$Lat,
decorrResid=sres)
residVariogram <- variogram(object=decorrResid~1, locations=~Lon+Lat,
data=residDF)
plot(residVariogram, main = "Variogram of Standardized Decorrelated
Residuals")

# add residuals to original df and plot these residuals
homes$decorresid <- rep(NA,nrow(homes))
homes$decorresid[!is.na(homes$Price)] <- sres

# plot a map of the residuals
ggplot(data=homes,mapping=aes(x=Lon, y=Lat, color=decorresid)) +
geom_point(size = 2) +
scale_color_distiller(palette="Spectral",na.value=NA) +
  labs(title = "Map of Decorrelated Residuals",
       color = "Decorrelated\nResiduals", x = "Longitude", y =
"Latitude") +
  theme_bw()

#  Normality: Histogram of standardized decorrelated residuals
hist(sres, probability = TRUE, ylim = c(0, 0.4), xlim = c(-5, 5),
     main = "Histogram of Standardized Decorrelated Residuals")
curve(dnorm(x), add = TRUE, col = "blue") # Normality holds

# Equal Variance: Fitted values vs. Standardized and decorrelated
residuals plot
fitted_values <- fitted(homes.lm.hetero)
ggplot(homes_no_na, aes(x = fitted_values, y = sres)) +
```

```
  geom_point() + ggtitle("Equal Variance Check") + # Equal variance
holds
  labs(x ="Fitted Values", y = "Standardized & Decorrelated
Residuals")

## Model Fit
cor(homes_no_na$Price, fitted_values)^2

## Model Predictive Accuracy
### Cross validation for predictive accuracy of home prices
# load Dr. Heaton's predict.gls function
source("./glstools-master/predictgls.R")

# perform monte carlo cross-validation
# get number of observations and set as n
n = nrow(homes_no_na)

# run 50 cvs
n.cv <- 50 #Number of CV studies to run
n.test <- floor(n*.10) #Number of observations in a test set
rpmse <- rep(x=NA, times=n.cv)
cvg <- rep(x=NA, times=n.cv)
bias <- rep(x=NA, times=n.cv)

pb <- txtProgressBar(min = 0, max = n.cv, style = 3) # start pb
for (cv in 1:n.cv) {
  ## Select test observations
  test.obs <- sample(x=1:n, size=n.test)

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

  ## fit model from training set
  train.gls <- gls(model = Price ~ Gr.Liv.Area + House.Style +
                               Year.Remod.Add + Central.Air +
Full.Bath +
                               Half.Bath + Bedroom.AbvGr +
Garage.Cars,
    data=train.set,  # data without missing values
    weights = varExp(form=~Gr.Liv.Area), # D part (Heteroskedastic)
    correlation = corExp(form = ~Lon+Lat, nugget = TRUE), # R part
(Spatial corr)
    method="ML")
```

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

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

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

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


  # update progress bar
  setTxtProgressBar(pb, cv)

}
close(pb)

# report rpmse and coverage
rpmse <- mean(rpmse)
cvg <- mean(cvg)
bias <- mean(bias)

# saveRDS(rpmse, file = "rpmse.Rdata")
# saveRDS(cvg, file = "cvg.Rdata")
#saveRDS(bias, file = "bias.Rdata")

# compare rpmse with sd
# rpmse # 16000
# sd(homes_no_na$Price) #86000

### Results
# 1
pr2 <- cor(homes_no_na$Price, fitted_values)^2  # Pseudo R squared =
0.9329216

# 2
homes.lm.hetero$coefficients > 0

cis<- confint(homes.lm.hetero)[-1,]
```

```
knitr::kable(cis, caption = "95% Confidence intervals for Home
Characteristic Effects")

# 3 - Variability
# This plot is also seen in the EDA
ggplot(homes_no_na, mapping = aes(x = Gr.Liv.Area, y = Price)) +
  geom_point() +
  labs(title = "Price by Living Area")

# get estimate for variance parameter theta
coef(homes.lm.hetero$modelStruct$varStruct, unconstrained=FALSE)

# 4
missingPreds <- predictgls(homes.lm.hetero, newdframe = homes_na,
level = 0.95)  # predict for missing values, save df
not_misssing_preds <- predictgls(homes.lm.hetero, newdframe =
homes_no_na)

homes_na$Price <- missingPreds$Prediction

full <- rbind(homes_no_na, homes_na)  # full <- rbind(missing,
without missing)

# Final map
ggplot(mapping = aes(x = Lon, y = Lat)) +
  geom_point(data = full[1:465,], aes(col = Price), size = 2.5, alpha
= 0.5) +
  geom_point(data = full[466:517,], aes(col = Price), size = 2.5,
shape = 13,
             alpha = 1) +
  theme_dark() +
  scale_color_distiller(palette = "Spectral", na.value = NA) +
  labs(title = "Home Price Predictions with Original Data",
       subtitle = "Predictions Marked with Target Shaped Points") +
  scale_fill_continuous(labels=scales::comma)
```