PDAT 610G Project Pt. II

Andrew Estes

10/10/2021

Introduction

We are utilizing the Ames, Iowa house dataset curated by Dean DeCock of Truman State University. This dataset describes 81 variables associated with 2,930 sales of individual residential property between 2006 and 2010. This project was initiated as part of a class project. Our goal for this project is to create a price predictive model. Utilizing principal component analysis (PCA) and ridge regression, we are able to create a model that accurately accounts for 70% of the houses price. See "Definitions" below for further explanation of these tests.

Definitions Principal component analysis is a tool used to reduce the number of variables while keeping preserving as much of the data's variation as possible. Basically, if there are 10 variables that account for 100% of the data, but the first 3 account for 90%, then the remaining 7 variables can be discounted.

Ridge regression is a form of linear regression that is used to analyze data models that suffer from multicollinearity.

Multicollinearity describes a relationship between two or more variables in a model that are highly correlated with each other. For example, the square foot of a garage is highly correlated to the number of vehicles the garage can handle.

Methods

The data was provided by Dean DeCock of Truman State University. It is already fairly clean although there are some missing data points within each variable. We will utilize the programming language R to do the analysis and R Markdown to transform it into a readable output. This dataset covers the sale of houses in Ames, Iowa between 2006 and 2010. This dataset has 81 variables and 2930 observations.

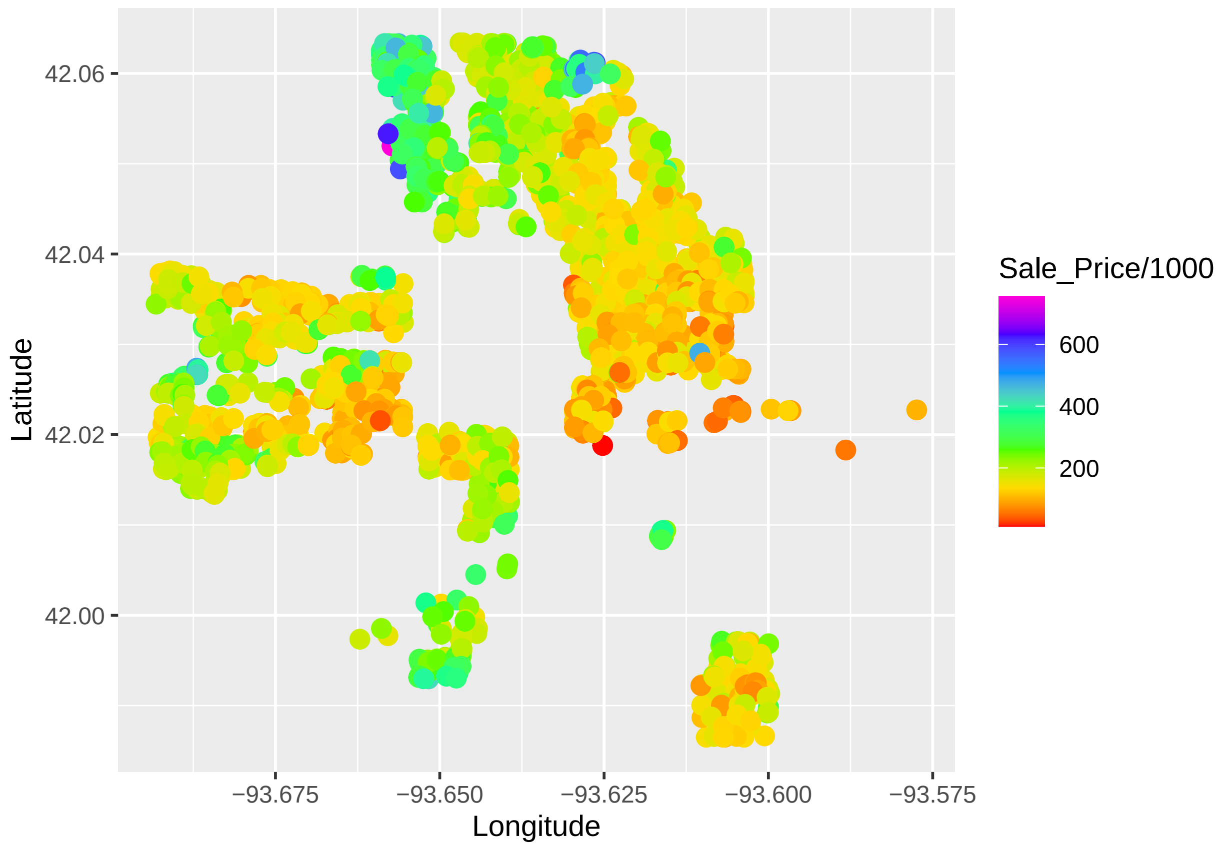
Our first step was to add a variable "Total_SF" to the dataset. Our hypothesis is total square feet is more important than the sum of square feet for each area of the house. After adding this variable, we visualize the data utilizing the R packages "tidyverse" and "mapview."

This showed us the actual location of each house sale on a geographic map. Using the ggplot function, we created a heatmap of house sales by price. This clearly showed the breakdown of house prices within Ames.

After seeing what the data said, we began formulating the data. The first step was to run the PCA on all 19 quantitative variables. The PCA said the top 3 variables were responsible for 50% of the variation. For prediction, 50% is not acceptable so we extended the model out to the top 8 variables, creating a PCA responsibility level of 75%. While this is not a great number given the quantity of variables, it is certainly something we can live with for this project.

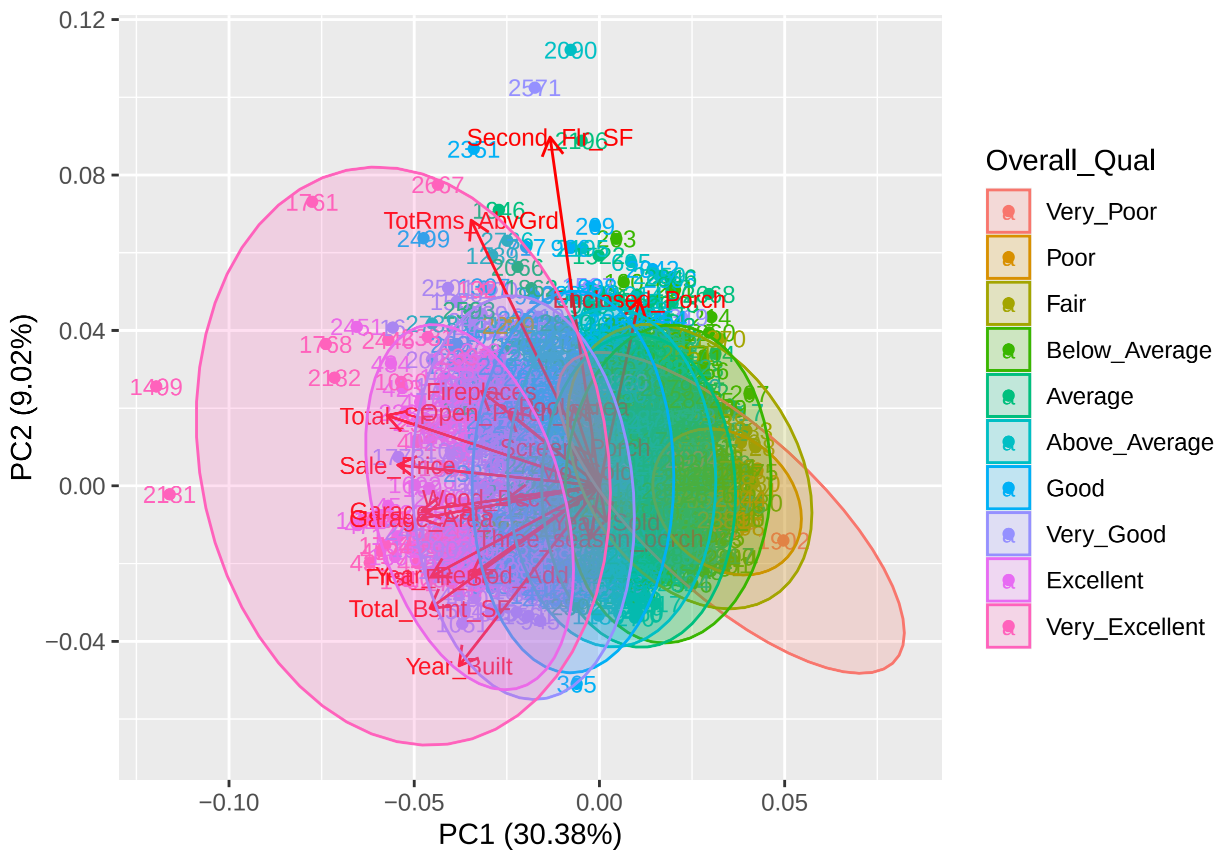
Our next step was to run a ridge regression on the models. We created a model that explained 70.26% of the price with a residual error of 43k. Using the same variables, we ran a regular linear regression model. It has near idential, albeit ever so slightly worse, predictive powers. This is due to the PCA already removing the most significant impacts of multicollinearity. If a PCA was not originally utilized, we posit the ridge regression would be significantly more accurate than the regular linear regression model.

Results

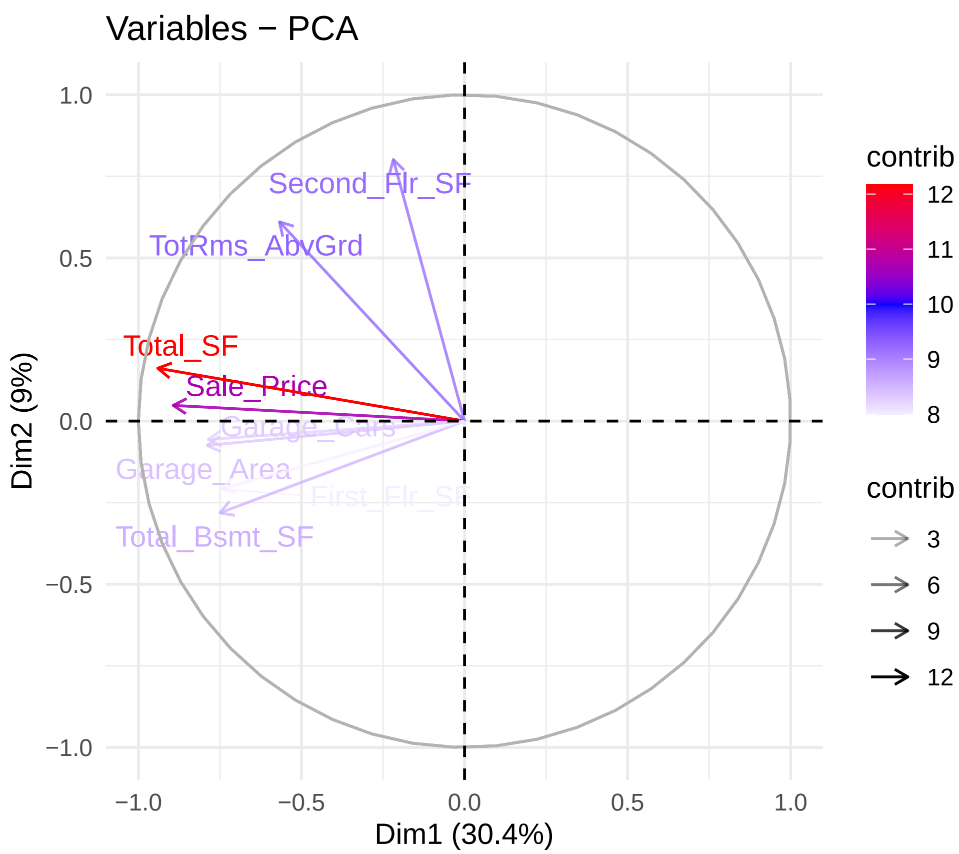


Here is a color-gradient plot seeing if there is any change in sale price based upon GPS coordinates. It shows that the north central end is the higher priced neighborhood while the central/east side is the poorer priced area of Ames.

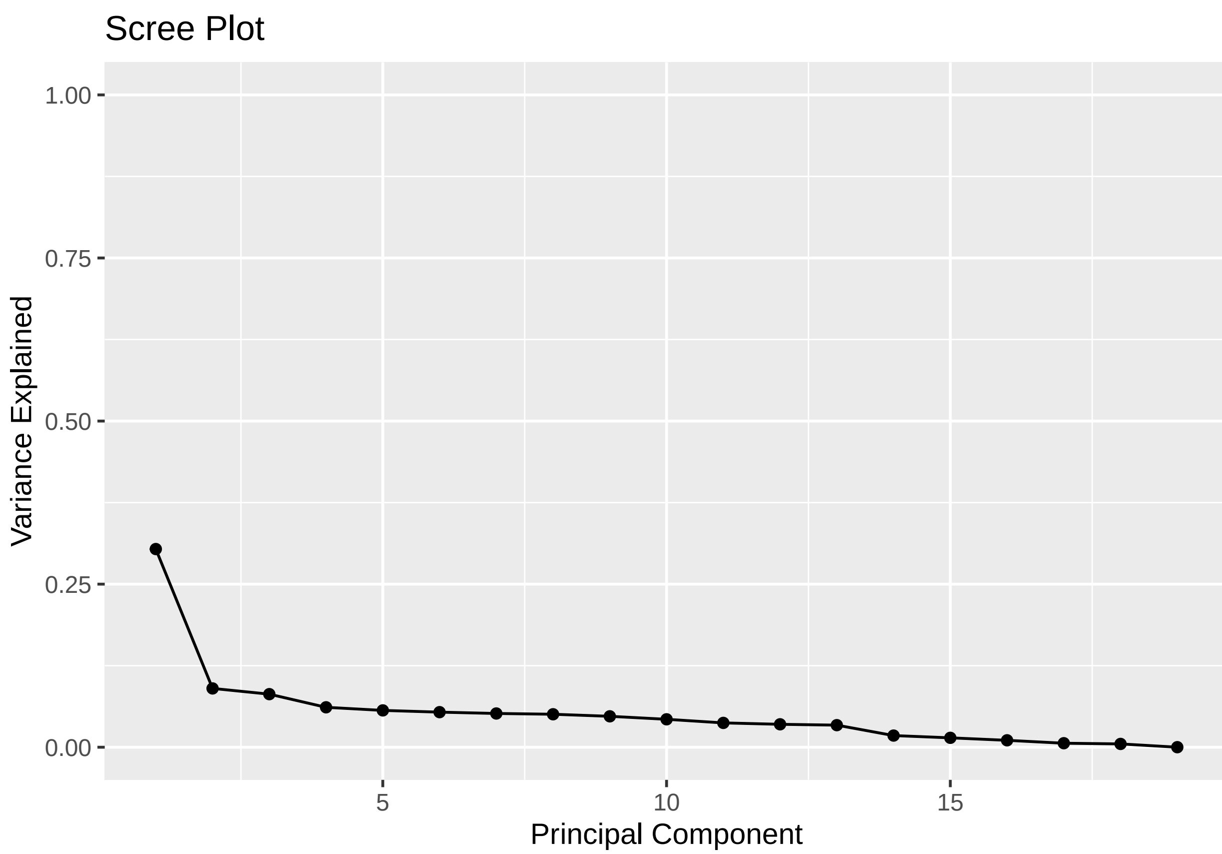
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4025 1.30887 1.24351 1.07844 1.03579 1.01091 0.99160
## Proportion of Variance 0.3038 0.09016 0.08139 0.06121 0.05647 0.05379 0.05175
## Cumulative Proportion  0.3038 0.39396 0.47535 0.53656 0.59303 0.64681 0.69857
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.98020 0.94921 0.90217 0.84198 0.81738 0.80225 0.58309
## Proportion of Variance 0.05057 0.04742 0.04284 0.03731 0.03516 0.03387 0.01789
## Cumulative Proportion  0.74913 0.79655 0.83939 0.87670 0.91187 0.94574 0.96364
##                           PC15    PC16    PC17    PC18    PC19
## Standard deviation     0.52349 0.44919 0.34233 0.31090 0.03531
## Proportion of Variance 0.01442 0.01062 0.00617 0.00509 0.00007
## Cumulative Proportion  0.97806 0.98868 0.99485 0.99993 1.00000
```



Here is a plot of overall quality and the impact of the 19 variables



Here is a clearer plot, showing the importance of the top 8 variables.



Here is the scree plot.

```
## [1] 197871.2
```

```
## [1] -5.12246
```

Here are the Residual Squared Error and R Squared amounts for the ridge regression model.

```
##
## Call:
## lm(formula = Sale_Price ~ Total_SF + Sale_Price + TotRms_AbvGrd +
##     Second_Flr_SF + Total_Bsmt_SF + Garage_Area + Garage_Cars +
##     First_Flr_SF, data = Ames2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -613141  -19610    -578   18573  284911
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -18301.257   3823.431  -4.787 1.78e-06 ***
## Total_SF          29.822      5.211   5.722 1.16e-08 ***
## TotRms_AbvGrd  -5430.471    872.708  -6.223 5.59e-10 ***
## Second_Flr_SF     49.228      6.601   7.458 1.15e-13 ***
## Total_Bsmt_SF     22.743      6.252   3.638  0.00028 ***
## Garage_Area      -12.633     10.282  -1.229  0.21929
```

```
## Garage_Cars    27185.231   2358.788  11.525  < 2e-16 ***
## First_Flr_SF      48.290      7.197   6.710 2.33e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43620 on 2922 degrees of freedom
## Multiple R-squared:  0.7026, Adjusted R-squared:  0.7018
## F-statistic:   986 on 7 and 2922 DF,  p-value: < 2.2e-16
```

Here is the summary of the regular linear regression model.

The ridge model has accounted for 70.26% of the data and has a residual error of 43613.76 dollars. The average house price is 180796.10 dollars. Our residual error with respect to the house price is 24.12%. Put another way, it is 75.88% accurate.

We also ran an ordinary linear regression using the same model. It came out very close to the ridge regression analysis. This should be expected because the PCA should have equalized most of the variance out so the ridge regression's impact was less significant than if we had began with that initially. It's R-Squared was also 70.26% but its Residual Standard Error was ever so slightly larger at 43620 dollars.

Further Discussion

Further research should be conducted by grouping the neighborhoods. That will likely explain the remaining 30% difference in predictive power. It will also reduce the residual standard error from 43k to a much smaller number.

Another consideration is the macro-economic levels. The Great Recession occurred in 2008. We treated each sale on a standard plane but time series cannot be ignored with such a large macro-event occuring in the heart of the data.

A final suggestion for further research would be to include city landmarks such as Iowa State University, Mary Greely Hospital, and river/park proximity.

References

```r
#install.packages("AmesHousing")
#install.packages('ggfortify')
#install.packages("devtools")
#install_github("kassambara/factoextra")
#install.packages("glmnet")
library(glmnet)
library(broom)
library(factoextra)
library(ggfortify)
library(tidyverse)
library(tinytex)
library(AmesHousing)
library(sf)
library(mapview)
library(devtools)

Ames <- make_ames( )
```

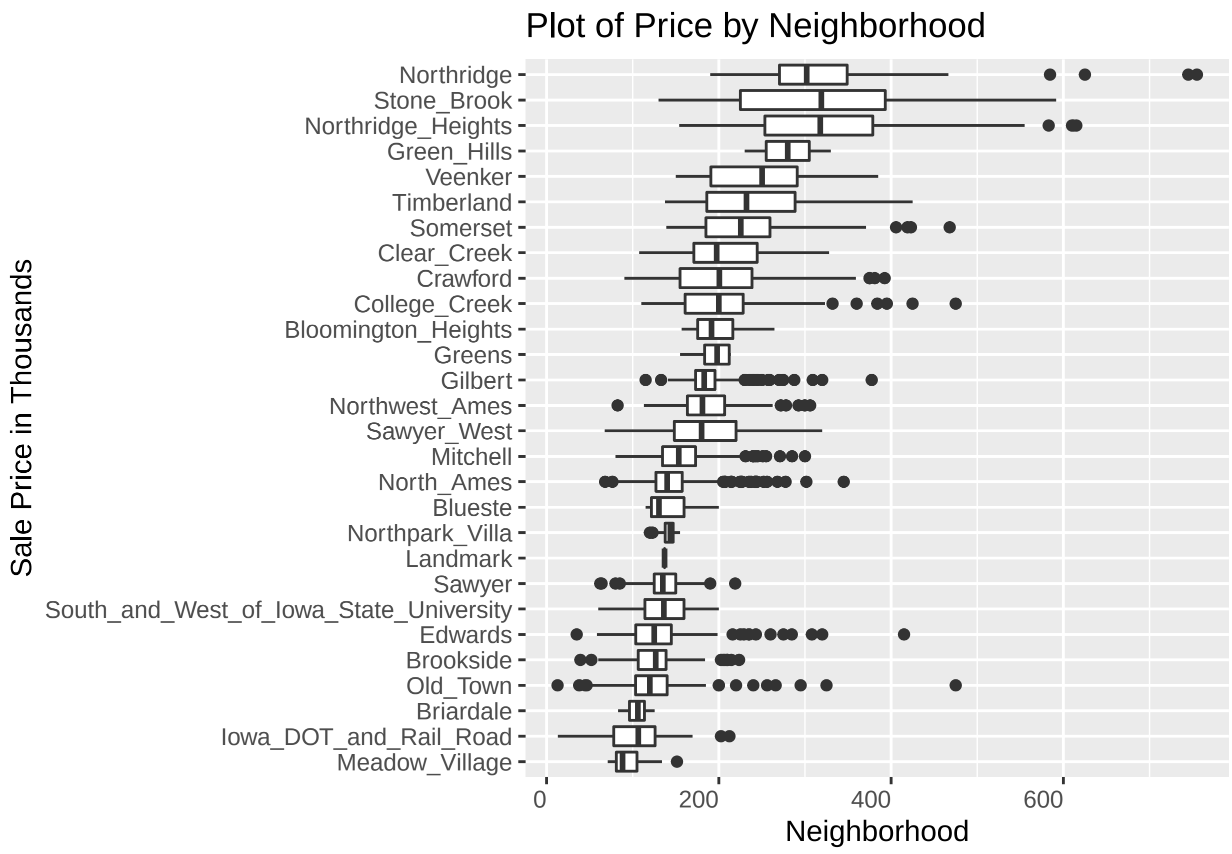
Appendix

Creating Total Square Feet variable

```
Ames2 <- Ames %>%
  mutate(Total_SF =
           Total_Bsmt_SF +
           Gr_Liv_Area +
           Garage_Area +
           Wood_Deck_SF +
           Open_Porch_SF +
           Enclosed_Porch +
           Three_season_porch +
           Screen_Porch
         )
```

Boxplot of neighborhoods broken down by price.

```
a <- ggplot(data = Ames) +
  geom_boxplot(mapping = aes(x=reorder(Neighborhood, Sale_Price/1000, na.rm = TRUE), y = Sale_Price/1000
  theme(axis.text.x = element_text(angle=0, hjust = 1)) +
  labs(title="Plot of Price by Neighborhood",x="Sale Price in Thousands", y = "Neighborhood")
a <- a + coord_flip()
a
```

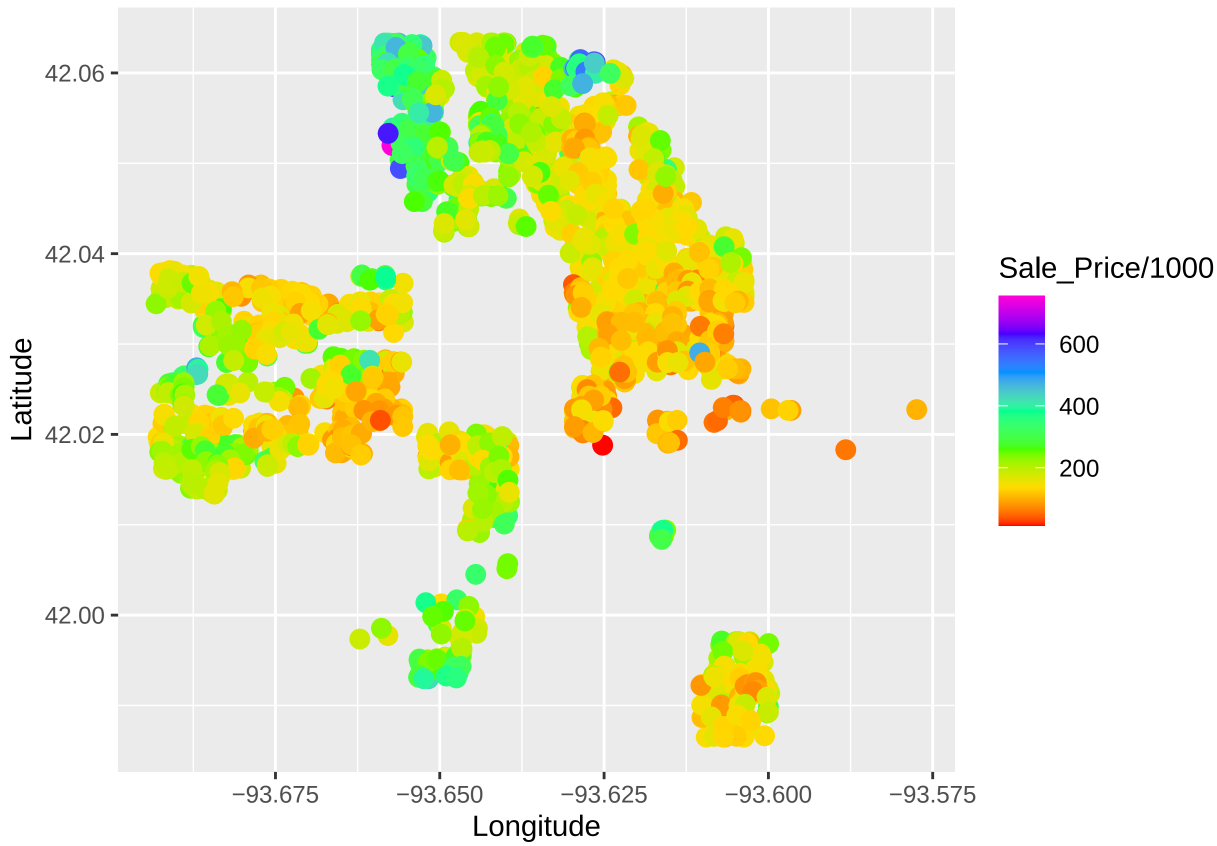


Map of houses

```
locations <- st_as_sf(Ames, coords = c("Longitude", "Latitude"), crs = 4326)
mapView(locations)
```

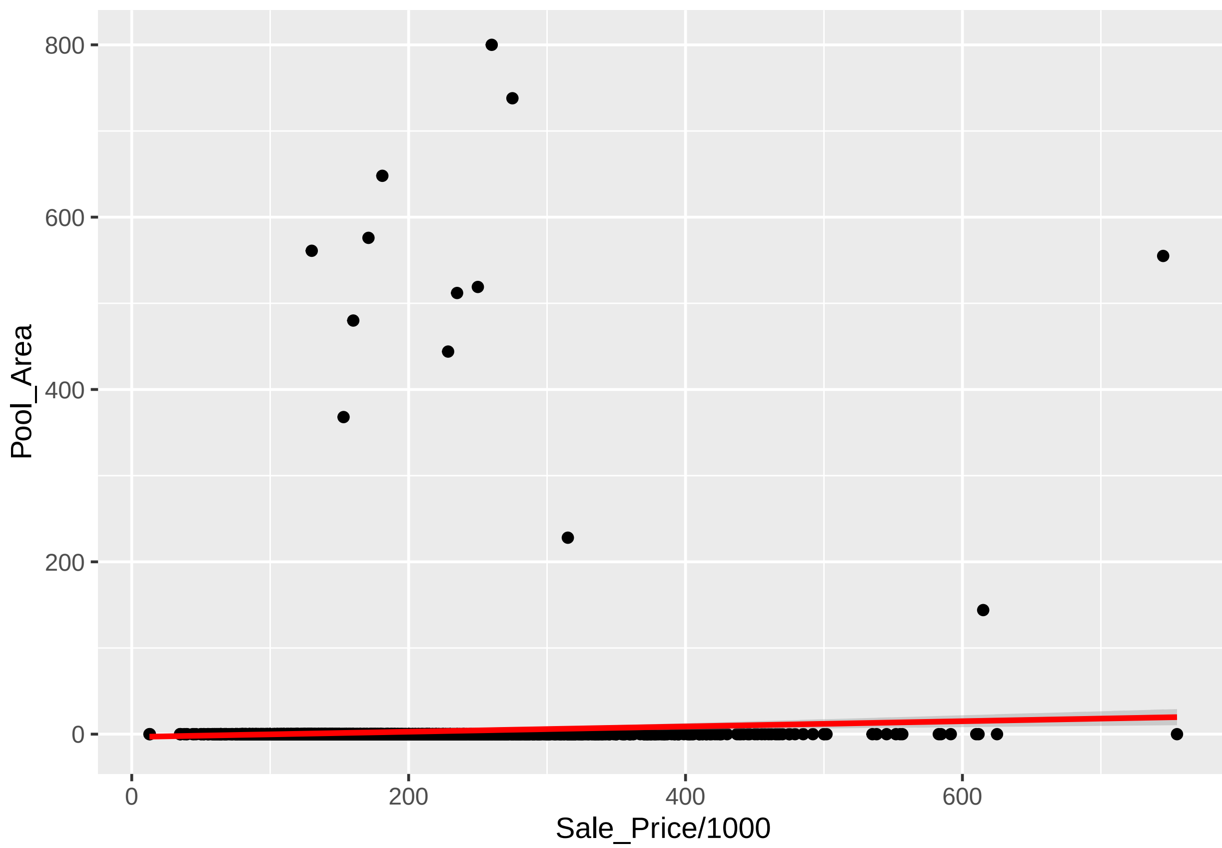
Color-gradient map of houses by sale price.

```
ggplot(Ames) +
  geom_point(data=Ames, aes(Longitude, Latitude, color = Sale_Price/1000), size = 3, lineend = "round")
  scale_color_gradientn(colours = rainbow(7))
```



Linear regression showing there is very little relation between pool size and house price.

```
ggplot(Ames, aes(x = Sale_Price/1000, y = Pool_Area)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red")
```

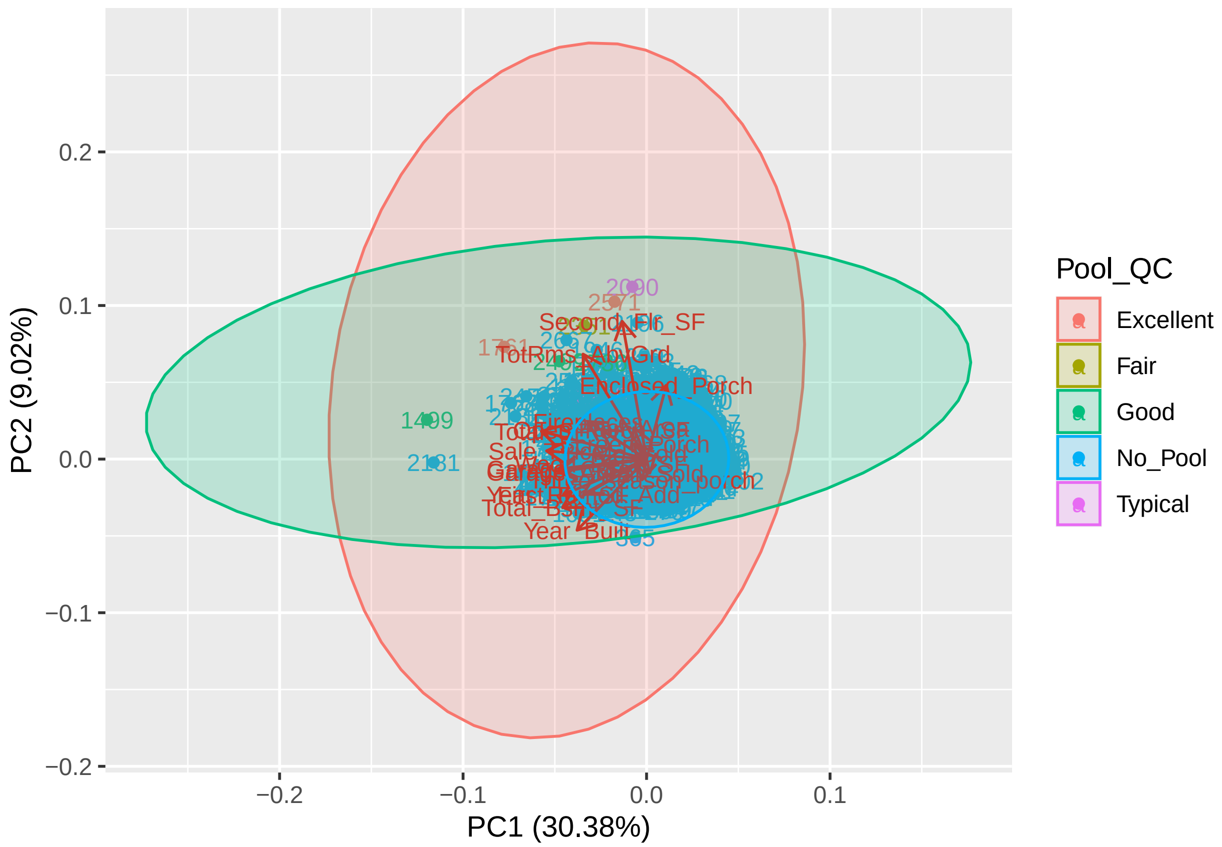


Creating dataset for PCA

```
Ames3 <- subset(Ames2, select = c(
    Sale_Price,
    Year_Built,
    Year_Remod_Add,
    Year_Sold,
    Mo_Sold,
    TotRms_AbvGrd,
    Fireplaces,
    Garage_Cars,
    Total_Bsmt_SF,
    First_Flr_SF,
    Second_Flr_SF,
    Garage_Area,
    Wood_Deck_SF,
    Open_Porch_SF,
    Pool_Area,
    Enclosed_Porch,
    Three_season_porch,
    Screen_Porch,
    Total_SF) )


ames.pca <- prcomp(Ames3, scale=TRUE )
summary(ames.pca)
```

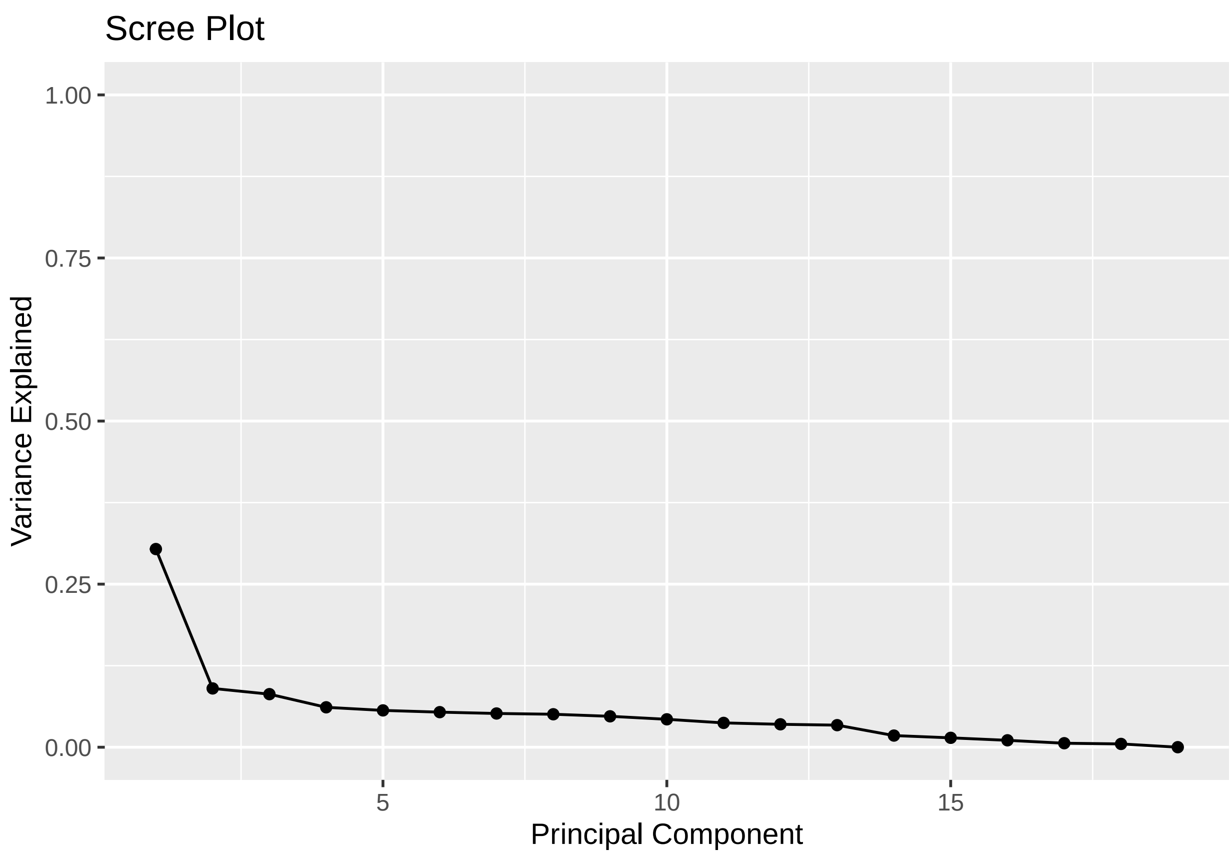
```r
str(ames.pca)
autoplot(ames.pca, data=Ames, colour = 'Pool_QC', frame=TRUE, frame.type='norm', label=TRUE, label.size
```



Scree Plot

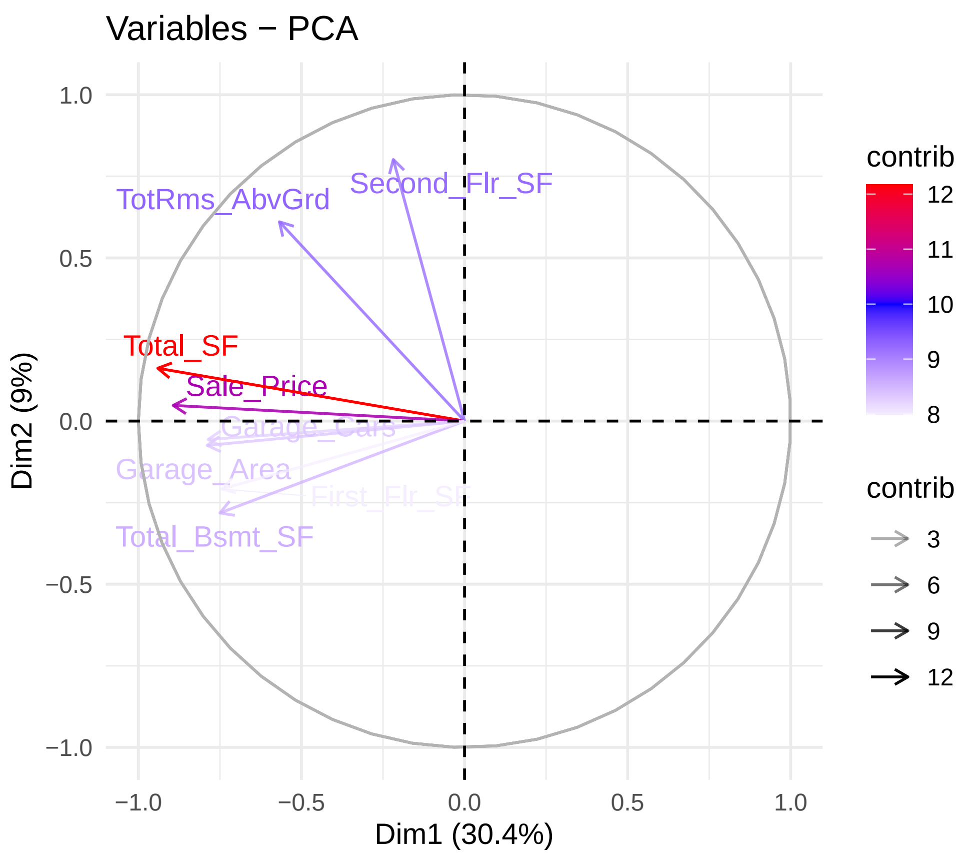
```r
var_explained <- ames.pca$sdev^2 / sum(ames.pca$sdev^2)

qplot(c(1:19), var_explained) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1)
```



PCA visualiation

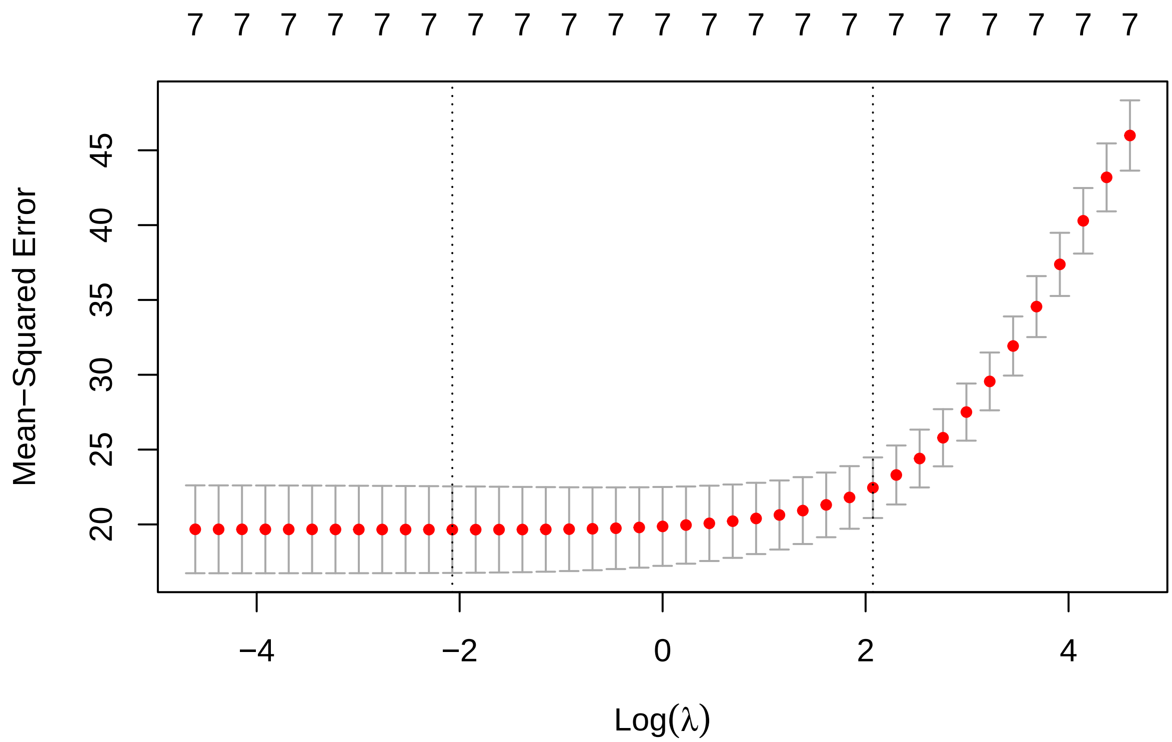
```
fviz_pca_var(ames.pca,
             alpha.var="contrib",
             col.var="contrib",
             repel = TRUE,
             select.var = list(contrib=8)
            ) +
             scale_color_gradient2(
               low="white",
               mid="blue",
               high="red",
               midpoint= 10
            ) +
  theme_minimal()
```



Ridge regression model

```
# Getting the independent variable
x_var <- data.matrix(Ames2[, c("Total_SF", "TotRms_AbvGrd", "Second_Flr_SF", "Total_Bsmt_SF", "Garage_A

# Getting the dependent variable
y_var <- (Ames2[, "Sale_Price"])
y_var <- as.numeric(unlist(y_var))
# Setting the range of lambda values
lambda_seq <- 10^seq(2, -2, by = -.1)
# Using glmnet function to build the ridge regression in r
fit <- glmnet(x_var, y_var, alpha = 0, lambda  = lambda_seq)
# Checking the model
summary(fit)
# Using cross validation glmnet
ridge_cv <- cv.glmnet(x_var, y_var/10000, alpha = 0, lambda = lambda_seq)
#Plotting MSE for Sale_Price per hundred thousand dollars
plot(ridge_cv)
```




```r
# Best lambda value
best_lambda <- ridge_cv$lambda.min
best_lambda

best_fit <- ridge_cv$glmnet.fit
summary(best_fit)
head(best_fit)
best_ridge <- glmnet(x_var, y_var, alpha = 0, lambda = 79.43000)
coef(best_ridge)


y_predict <- predict(best_fit, s = best_lambda, newx = x_var)
# Sum of Squares Total and Error
sst <- sum((y_var - mean(y_var))^2)
sse <- sum((y_predict - y_var)^2)
rse <- sqrt((sse)/(2930-7))
rse

# R squared
rsq <- 1 - sse / sst
rsq
#The optimal model has accounted for 70.26% of the data and has a residual error of $43613.76
#The average house price is $180796.10. Our residual error with respect to the house price is 24.12%.
#Put another way, it is 75.88% accurate.
```

Ordinaly linear regression model

```r
lm_fit <- lm(Sale_Price ~
               Total_SF +
               Sale_Price +
               TotRms_AbvGrd +
               Second_Flr_SF +
               Total_Bsmt_SF +
               Garage_Area +
               Garage_Cars +
               First_Flr_SF,
             data = Ames2)
summary(lm_fit)
```