House Price Analysis

Richard Yones

2023-02-23

Ames Housing Data Analysis

Context and Dataset

The Ames Housing dataset has entries of houses in the Ames housing market and their relevant information. Some of these pieces of information include classic housing information: sales price, amount of rooms, square footage, and build year.

Research Question

In this report, I will try to find a model that accurately predicts sales price (SalePrice) based upon several variables found in the data. An astute data model would be useful in determining the most important characteristics to consider when trying to manage house pricing.

Variables - Descriptive Statistics

The dependent variable of interest is sales price (SalePrice).

The predictors that I plan to use are lot area (LotArea), overall quality (OverallQual), overall condition (OverallCond), indoor square footage (a sum of GrLivArea, GarageArea, TotalBsmtSF), outdoor square footage (sum of WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea), and Kitchen Quality (KitchenQual). I have modified the original table to only include the relevant pieces of information.

```
library(tidyverse)
houses <- read_csv("~/Downloads/R1housingprices.csv")
comp_house <- houses %>%
  mutate(InSF = GrLivArea + GarageArea + TotalBsmtSF) %>%
  mutate(OutSF = WoodDeckSF + OpenPorchSF + EnclosedPorch + SsnPorch + ScreenPorch + PoolArea) %>%
  mutate(KQual = ifelse(KitchenQual %in% "Fa", 0,
                        ifelse(KitchenQual %in% "TA", 1,
                               ifelse(KitchenQual %in% "Gd", 2,
                                      ifelse(KitchenQual %in% "Ex", 3, 0))))) %>%
  select(SalePrice, LotArea, OverallQual, OverallCond, InSF, OutSF, KitchenQual, KQual)

head(comp_house)
```

```
## # A tibble: 6 × 8
##   SalePrice LotArea OverallQual OverallCond  InSF OutSF KitchenQual KQual
##       <dbl>   <dbl>       <dbl>       <dbl> <dbl> <dbl> <chr>       <dbl>
## 1    208500    8450           7           5  3114    61 Gd              2
## 2    181500    9600           6           8  2984   298 TA              1
## 3    223500   11250           7           5  3314    42 Gd              2
## 4    140000    9550           7           5  3115   307 Gd              2
## 5    250000   14260           8           5  4179   276 Gd              2
## 6    143000   14115           5           5  2638   390 TA              1
```

Note: KQual will be used in the regression in lieu of KitchenQual. I have converted KitchenQual into numeric data based on the scale. Fa, TA, Gd, Ex = 0, 1, 2, 3, respectively. Fa is the reference variable.

The following has characteristic information about each variable:

```
variables <- read_csv("~/Downloads/VariableDesc.csv")
variables
```

```
## # A tibble: 8 × 4
##   VariableName VariableType DataType DistType
##   <chr>        <chr>        <chr>    <chr>
## 1 SalePrice    IV           Quant    Continuous
## 2 LotArea      DV           Quant    Continuous
## 3 OverallQual  DV           Quant    Discrete
## 4 OverallCond  DV           Quant    Discrete
## 5 InSF         DV           Quant    Continuous
## 6 OutSF        DV           Quant    Continuous
## 7 KitchenQual  DV           Qual     <NA>
## 8 Kqual        DV           Quant    Discrete
```

Statistical data on each quantitative variable is below:

```
library(psych)
stat_house <- comp_house %>%
  select(!KitchenQual)

stat <- describe(stat_house) %>%
  select(n, median, mean, sd)
stat
```

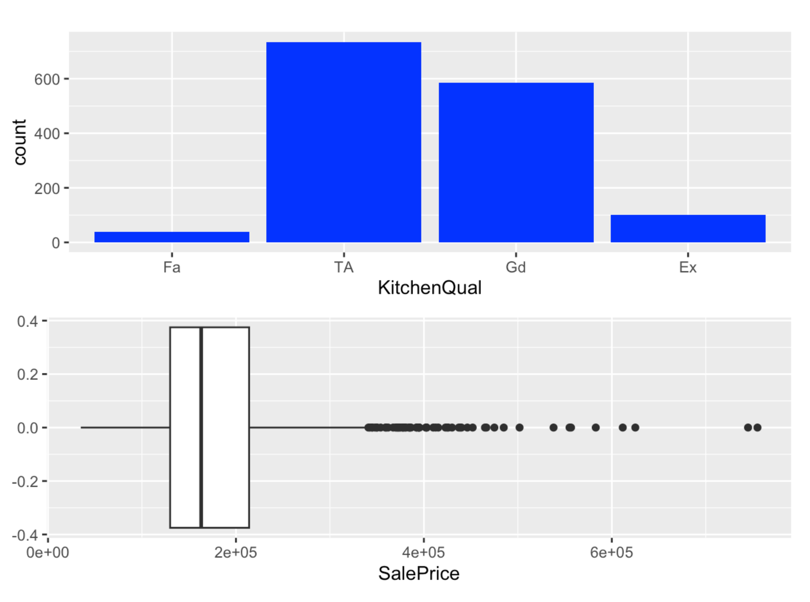
```
##                n   median      mean       sd
## SalePrice   1460 163000.0 180921.20 79442.50
## LotArea     1460   9478.5  10516.83  9981.26
## OverallQual 1460      6.0      6.10     1.38
## OverallCond 1460      5.0      5.58     1.11
## InSF        1460   2939.5   3045.87   959.53
## OutSF       1460    164.0    184.09   166.42
## KQual       1460      1.0      1.51     0.66
```

In the data set, there are 1460 total observations

Descriptive Visualizations

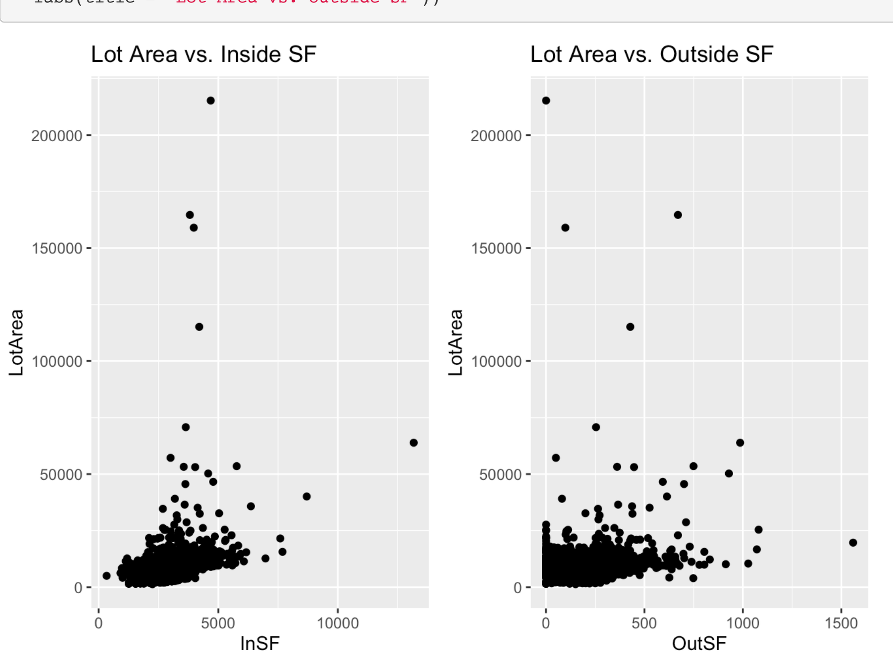
A series of plots are here to provide some visualization of the variables and how they each individually interact with SalePrice.

```
library(ggpubr)
comp_house$KitchenQual <- factor(comp_house$KitchenQual, levels = c("Fa", "TA", "Gd", "Ex"))

ggarrange(
ggplot(data = comp_house, mapping = aes(x = SalePrice)) +
  geom_boxplot(),
ggplot(data = comp_house, mapping = aes(x = KitchenQual)) +
  geom_bar(fill = "blue")) +
  coord_flip()
```



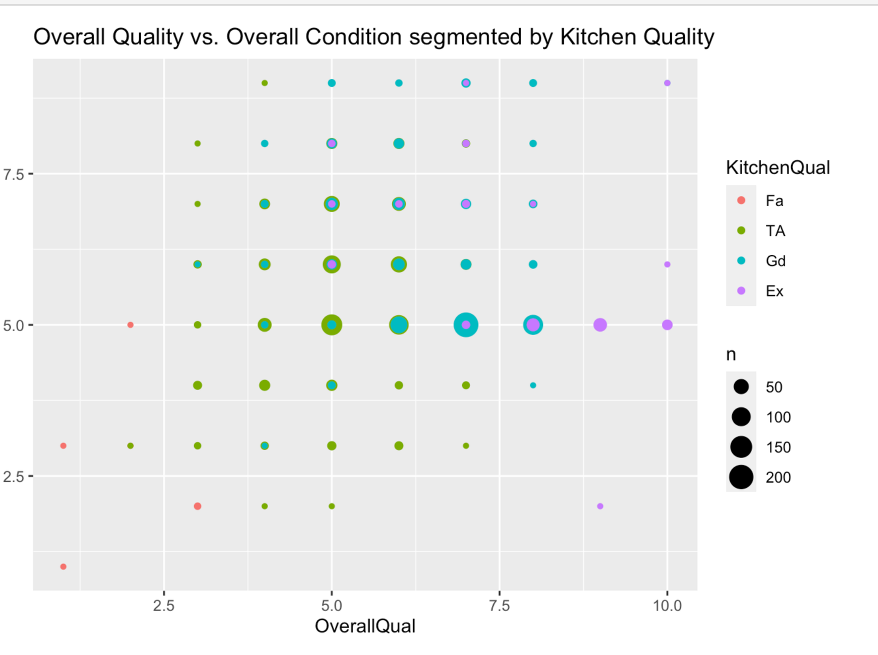
Looking at the boxplot, we see that the middle 50% of the data is consolidated into a smaller range when compared to the other quartiles. The bar graph shows that most houses have a good or average kitchen quality.

```
ggarrange(
ggplot(data = comp_house) +
  geom_point(mapping = aes(x = InSF, y = LotArea)) +
  labs(title = "Lot Area vs. Inside SF"),
ggplot(data = comp_house) +
  geom_point(mapping = aes(x = OutSF, y = LotArea)) +
  labs(title = "Lot Area vs. Outside SF"))
```



These scatterplots display InSF and OutSF against Lot Area. As you can see, there are a handful of houses that look like outliers. These outliers represent relatively smaller houses on large plots of land.

```
test <- count(comp_house, OverallCond, OverallQual)

ggarrange(ggplot(data = comp_house) +
  geom_count(mapping = aes(x = OverallQual, y = OverallCond, color = KitchenQual)) +
  labs(title = "Overall Quality vs. Overall Condition segmented by Kitchen Quality"))
```



This graph that pits OverallQual together with OverallCond is segmented by Kitchen Quality. Houses with a kitchen qual of "Ex" score in the higher reagions of both condition and quality. The inverse is true as well; lower quality kitchens are typically associated with houses of lower quality and condition. In general, the largest number of houses congregate around the middle scores.

Hypotheses

I believe that the variables chosen have a good chance of being significant predictors of SalePrice. House prices can be volatile and can depend on many conditions outside the house itself. That being said, I would expect that the very fundamental characteristics of a house (sq footage, overall perception/condition) would correlate significantly with its price.

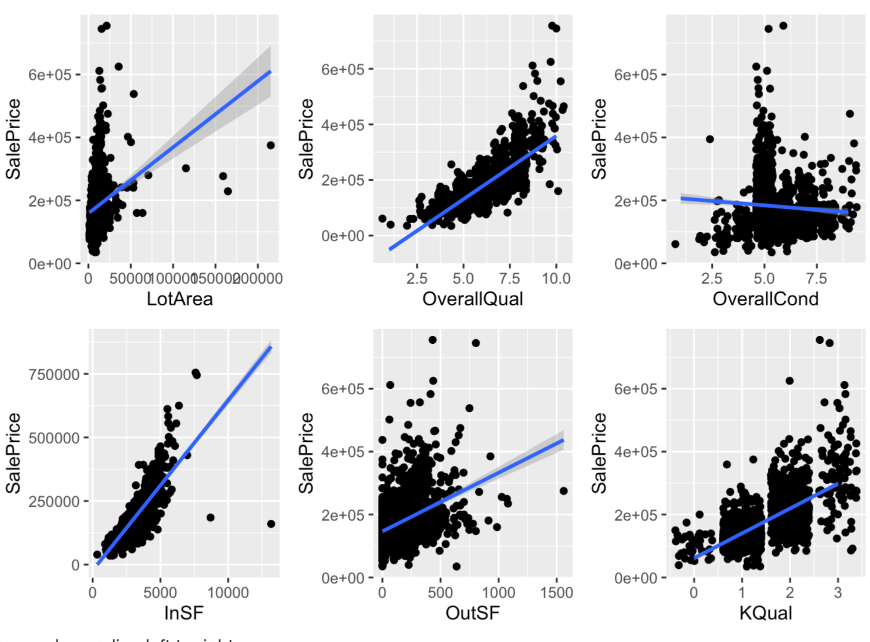
F-Test - Null and alternative hypotheses

- H0: There is **no** relationship between the independent variables (LotArea, OverallQual, OverallCond, InSF, OutSF, KitchenQual), and the dependent variable (SalePrice).
- HA: There is a **significant** relationship between the independent variables (LotArea, OverallQual, OverallCond, InSF, OutSF, KitchenQual), and the dependent variable (SalePrice).

Model Results

Individual Predictive Visualizations

```
library(ggpubr)
ggarrange(ggplot(comp_house, aes(LotArea, SalePrice)) +
  geom_point() +
  geom_smooth(method = lm),

ggplot(comp_house, aes(OverallQual, SalePrice)) +
  geom_jitter() +
  geom_smooth(method = lm),

ggplot(comp_house, aes(OverallCond, SalePrice)) +
  geom_jitter() +
  geom_smooth(method = lm),

ggplot(comp_house, aes(InSF, SalePrice)) +
  geom_point() +
  geom_smooth(method = lm),

ggplot(comp_house, aes(OutSF, SalePrice)) +
  geom_point() +
  geom_smooth(method = lm),

ggplot(comp_house, aes(KQual, SalePrice)) +
  geom_jitter() +
  geom_smooth(method = lm))
```



Correlation Coefficients Correlate to graphs reading left to right:

```
cor(comp_house$LotArea, comp_house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.2638434
```

```
cor(comp_house$OverallQual, comp_house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.7909816
```

```
cor(comp_house$OverallCond, comp_house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] -0.07785589
```

```
cor(comp_house$InSF, comp_house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.8075185
```

```
cor(comp_house$OutSF, comp_house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.390365
```

```
cor(comp_house$KQual, comp_house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.6595997
```

Note: Jitter is added to the discrete variables to more easily see the trend line relationship. **Observations**

1. The outliers in LotArea seem to drag the trend line into a more flattened position
2. The relationship that LotArea, OverallCond, and OutSF have with SalePrice seem much more randomized thans the others. (r = .264, -.078, .390)
3. Looks like OverallQual and InSF have a much better individual relationship with SalePrice. (r = .791, .808)

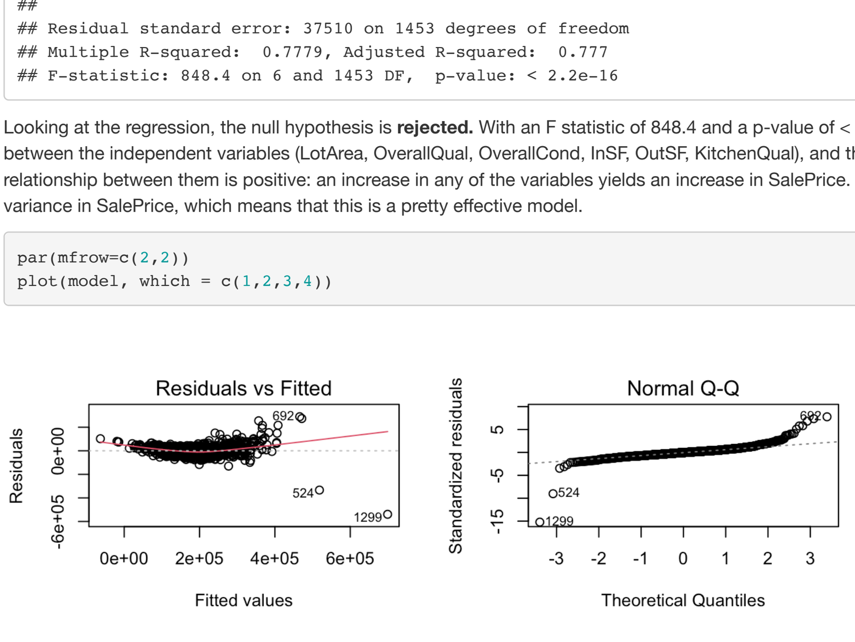
Regression Model Results

```
model <-lm(SalePrice ~ LotArea + OverallQual + OverallCond + InSF + OutSF + KQual, data = comp_house)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + OverallCond +
##     InSF + OutSF + KQual, data = comp_house)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -539116  -18202   -1246   14629  289217
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.040e+05  7.173e+03 -14.493  < 2e-16 ***
## LotArea      6.040e-01  1.051e-01   5.746 1.11e-08 ***
## OverallQual  2.052e+04  1.145e+03  17.923  < 2e-16 ***
## OverallCond  1.813e+03  9.015e+02   2.011  0.04451 *
## InSF         3.553e+01  1.593e+00  22.300  < 2e-16 ***
## OutSF        2.058e+01  6.552e+00   3.141  0.00172 **
## KQual        2.069e+04  2.027e+03  10.209  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37510 on 1453 degrees of freedom
## Multiple R-squared:  0.7779, Adjusted R-squared:  0.777
## F-statistic: 848.4 on 6 and 1453 DF,  p-value: < 2.2e-16
```

Looking at the regression, the null hypothesis is **rejected.** With an F statistic of 848.4 and a p-value of < 0.05, there is a **significant relationship** between the independent variables (LotArea, OverallQual, OverallCond, InSF, OutSF, KitchenQual), and the dependent variable (SalePrice). The relationship between them is positive: an increase in any of the variables yields an increase in SalePrice. The overall model explains **77.7%** of the variance in SalePrice, which means that this is a pretty effective model.

```
par(mfrow=c(2,2))
plot(model, which = c(1,2,3,4))
```




```
library(car)
vif(model)
```

```
##     LotArea OverallQual OverallCond        InSF       OutSF       KQual
##    1.141212    2.598381    1.043311    2.422626    1.232719    1.876713
```

Looking at the diagnostic plots, the model is very healthy. There are a few outliers in the data, and the residuals are random. Looking at the VIF scores, none of the variables reach a value > 10, which suggests no multicollinearity.

Conclusions

Results and Discussions

Since the null hypothesis was rejected, my initial thoughts were correct: fundamental characteristics of a house are related to its value. The strength also supports the part of my hypothesis that there are outside influences of a house's value that would help predict its value. The model only affects about 78% of the variation; the remainder variation could be due to these other variables.

This information could be used by large or individual realtors/investors alike. Using information from each independent variable, realtors would be able to reasonably predict what that house is worth. For "fixer-upper" projects, renovators in tandem with realtors would be able to know the main characteristics to spruce up the value of a house. Home buyers, assuming they could have access to this specific information, could make reasonable offers on houses without the fear of being completely ripped off.

Prescriptive Reccomendations

Businesses in the real estate sphere, specifically those in the Ames housing market, can use the model to optimally price houses on the market, either for buying or for selling.

Losses on investments can be reduced, and increased profits become sustainable if real estate companies can learn how to leverage this model to the best of their ability. Companies won't over pay for properties that are lackluster, increasing their margins. Furthermore, investors can renovate specific characteristics related to the independent variables to maximize the value of the house.

Key takeaways

- There is a **significant** relationship between the independent variables (LotArea, OverallQual, OverallCond, InSF, OutSF, KitchenQual), and the dependent variable (SalePrice).
- Using the model, one can reasonably predict the selling price of a house. This will help with accuracy; individuals can make educated offers on homes.

Limitations

The first limitation is that this data set only involves houses in the Ames housing market. Generalizing this data set too broadly might be innappropriate; different markets might have different significant variables. The second involves the subjective nature of KitchenQual, OverallQual, and OverallCond. There is probably not a standardized way to differentiate a house with an OverallQual of 5 vs. one with a 6.

Improving the Model When it comes to improving the model, I would have liked to have other economic/financial data about the market. This could include the amount of days that a typical house sat on sale for, average housing sale price for a certain period of time, or a house's number of owners. This would help segment the data better and help account for economic context

Looking Forward The r2 value of ~78% leaves 22% of variation that is unexplained. That is almost a quarter of the data that has a missing piece; this model is far from perfect. The question of causality is still open; causality can be found by a long series of predictive models or by a closed environment experiment. After one simple data model, it is not appropriate to say the independent variables *directly cause* SalePrice to increase.

I recommend a few options to build upon our work:

1. Use data from surrounding markets to determine if the Ames data model holds
2. Continue finding variables to optimize variation explanation while not clouding the data model.