Mid-Term Exam

2023-10-13

Iowa Houses (What a Housing Crisis)

Here we set up the preliminaries:

```r
iowa_all <- read.csv("../data/Iowa.csv")
#Taking only the 4 variables required
iowa <- iowa_all[,c("Id","OverallQual","YearBuilt","SalePrice")]
```

Then, we count the number of houses for each value in `OverallQual`. I'm sure there's a built-in function to do the counting, but I can't for the life of me remember it right now, so we brute force it with a `for`-loop:

```r
count <- numeric(10)
for (i in 1:10) {
  count[i] <- sum(iowa$OverallQual == i)
}

q11 <- as.data.frame(cbind(c(1:10), count))
colnames(q11) <- c("OverallQual", "Count")
```

Now, we find the most expensive house built in 2007:

```r
iowa_2007 <- iowa[iowa$YearBuilt ==  2007,]
q12 <- iowa_2007[iowa_2007$SalePrice ==max(iowa_2007$SalePrice), 1]
```

Picky Connie

Here, we pick out the houses built in or after `2000`, with quality at least equal to `7`:

```r
q13 <- iowa[iowa$YearBuilt >= 2000 & iowa$OverallQual >= 7,]
```

Now, we find the cheapest house that also has an overall quality of `10`. I have chosen to subset the original data frame so that we only have the houses of best quality before finding the cheapest of these houses:

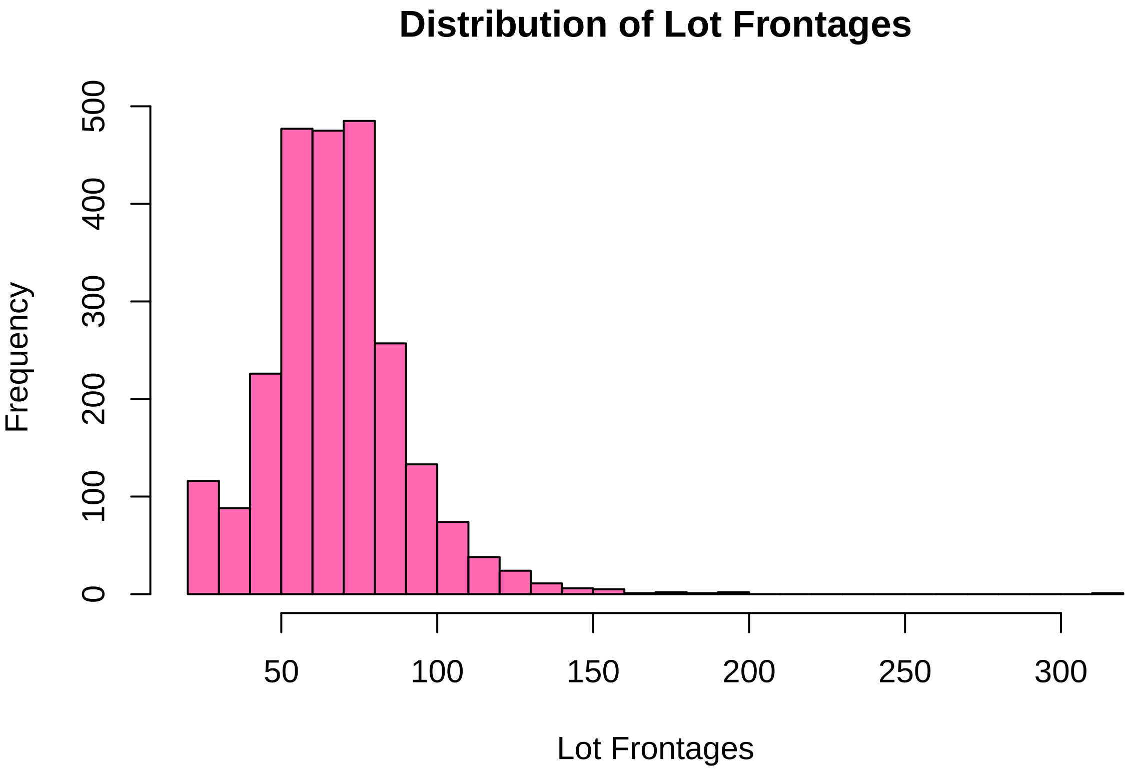
```r
best_quality <- q13[q13$OverallQual == 10,]

q14 <- best_quality[best_quality$SalePrice == min(best_quality$SalePrice), 1]
```

Lot Frontage

Here, we take only the `LotFrontage` column and remove the missing values:

```r
lot_frontage <- iowa_all$LotFrontage
#removing the missing values
lot_frontage <- lot_frontage[is.na(lot_frontage)==FALSE]
```

Here is the histogram showing the distribution of lot frontages:



As we can see, the distribution does not look symmetric. It has a bit of a thin tail to the right (if that's the right terminology) that isn't reflected on the left side.

An appropriate estimate for the 50th percentile is:

```
## 50%
##  68
```

Here, we perform the 1000 bootstrap resamples as required, and calculate the confidence interval:

```
sample_size <- length(lot_frontage)
one_resampled_frontage<- function(x){
  bootstrap <- lot_frontage[sample(sample_size, replace = TRUE)]
  return(quantile(bootstrap, 0.5))
}

many_resamples <- sapply(1:1000, one_resampled_frontage)

ci_lower_bound <- quantile(many_resamples, 0.025)
ci_upper_bound <- quantile(many_resamples, 0.975)
```

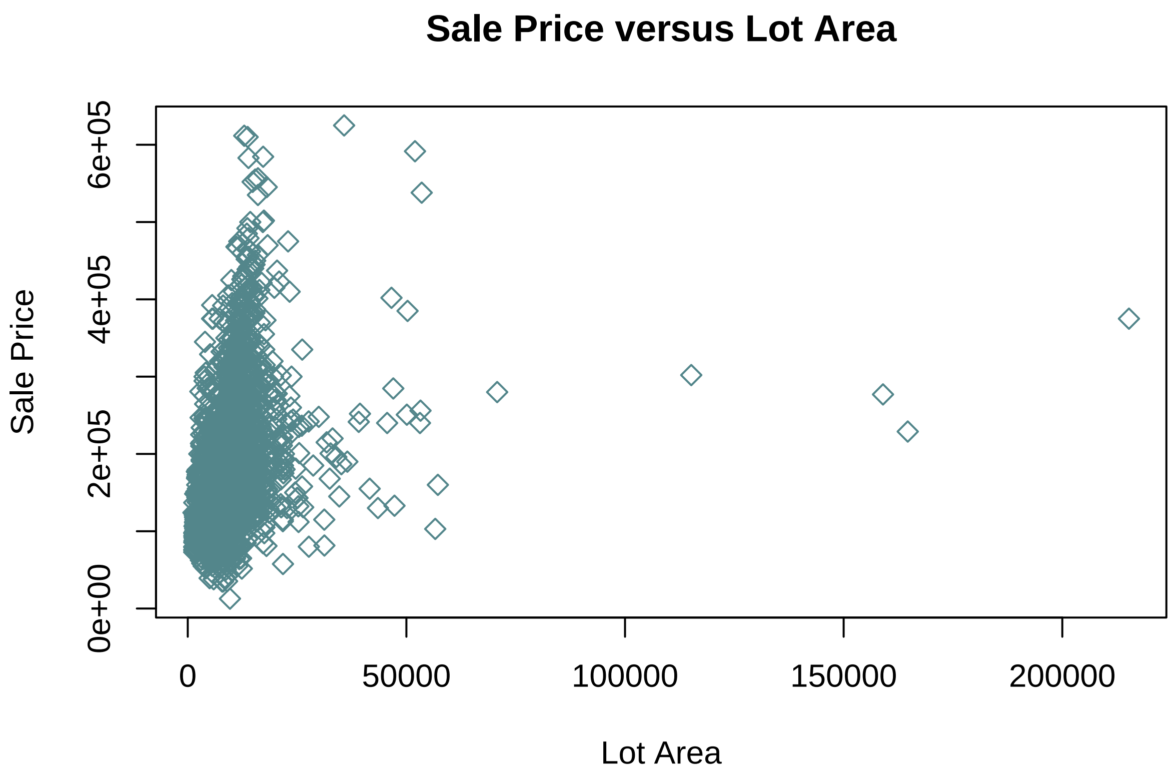
My 95%-confidence interval is (66,70).

Big Space = Big $$?

Here, we take the relevant columns and put them in a new data frame, just for convenience:

```
iowa_scatter <- iowa_all[,c("SalePrice", "LotArea")]
```

Here's the required scatterplot:



The relationship looks like it could plausibly be linear, but there are many outliers to the right of the plot, which would affect how a linear model is fit to the data. If it is linear, it would be a positive relationship.

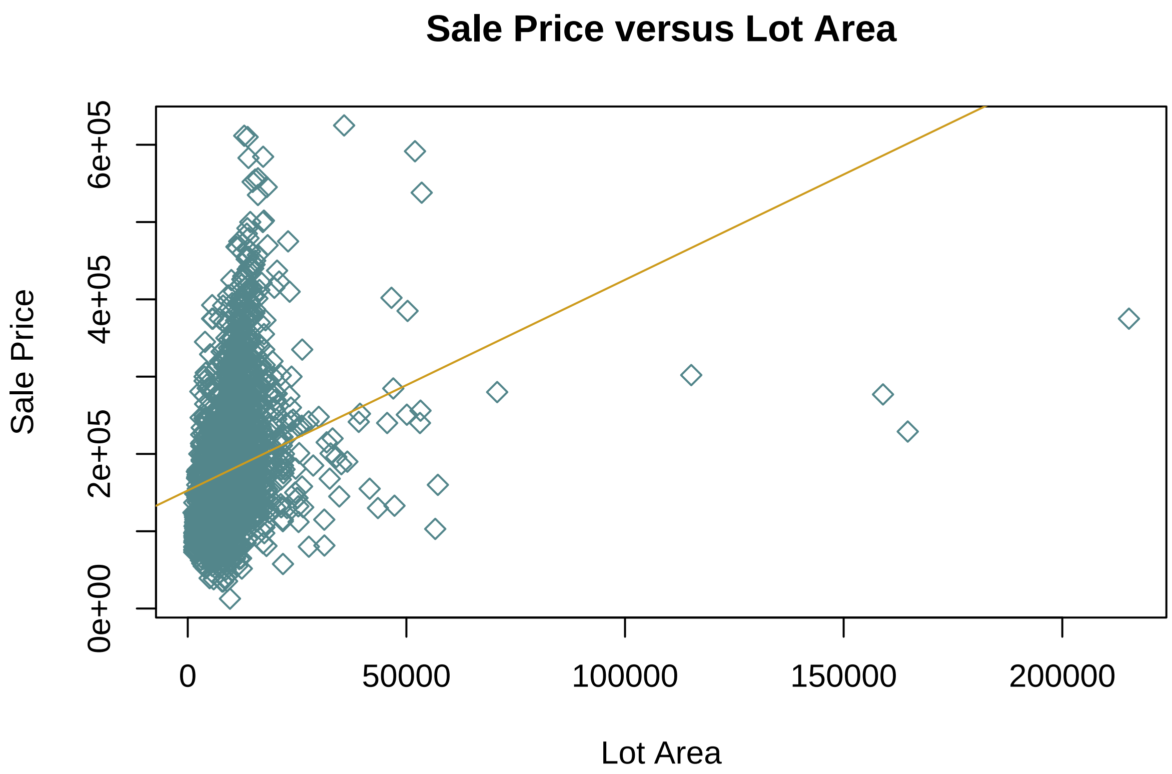
Now, performing the linear regression:

```
linear_regression <- lm(SalePrice ~ LotArea, data = iowa_scatter)
variance_prop <- summary(linear_regression)$r.squared
```

The proportion of variance explained by this model is

```
## [1] 0.07392833
```

Here, we add the best fit line to the scatter plot from before:



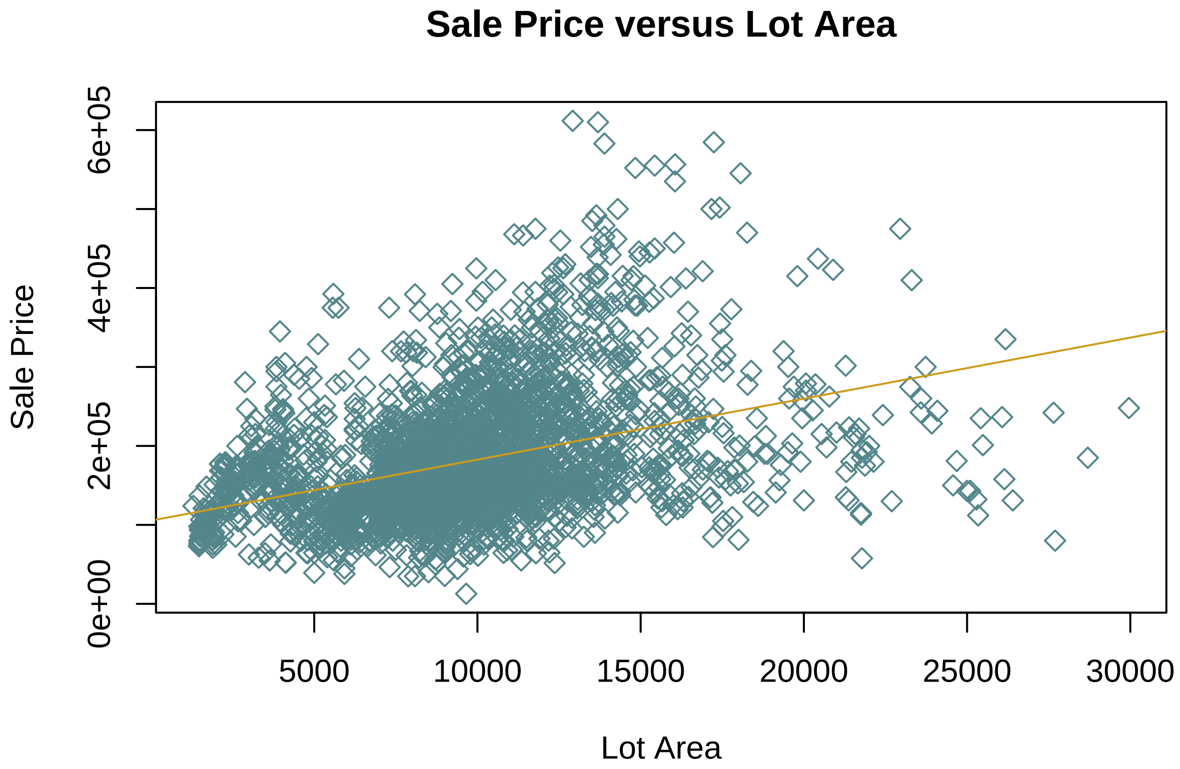
Now, we remove the extreme values from the data and repeat the regression:

```
iowa_scatter_no_extreme <- iowa_scatter[iowa_scatter$LotArea <= 30000,]
linear_regression_no_extreme <-lm(SalePrice ~ LotArea, data = iowa_scatter_no_extreme)
variance_prop_no_extreme <- summary(linear_regression_no_extreme)$r.squared
```

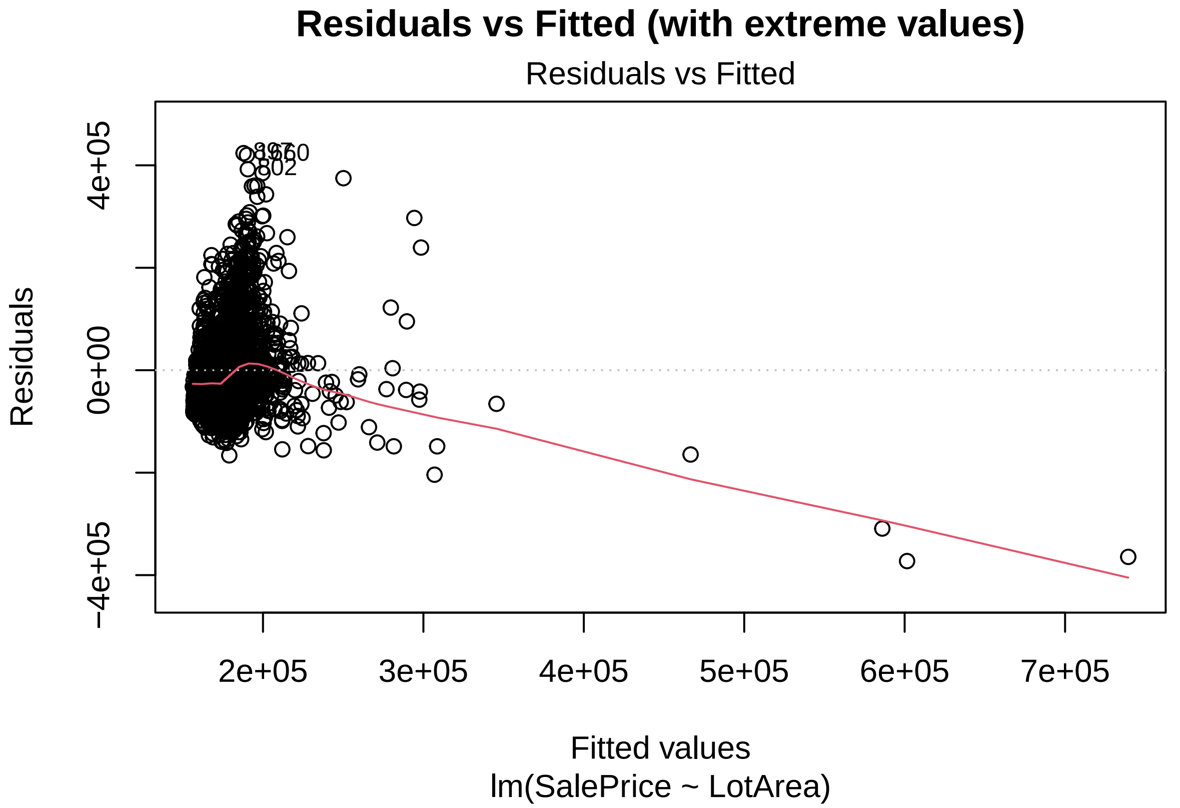
The proportion of variance explained by this model is

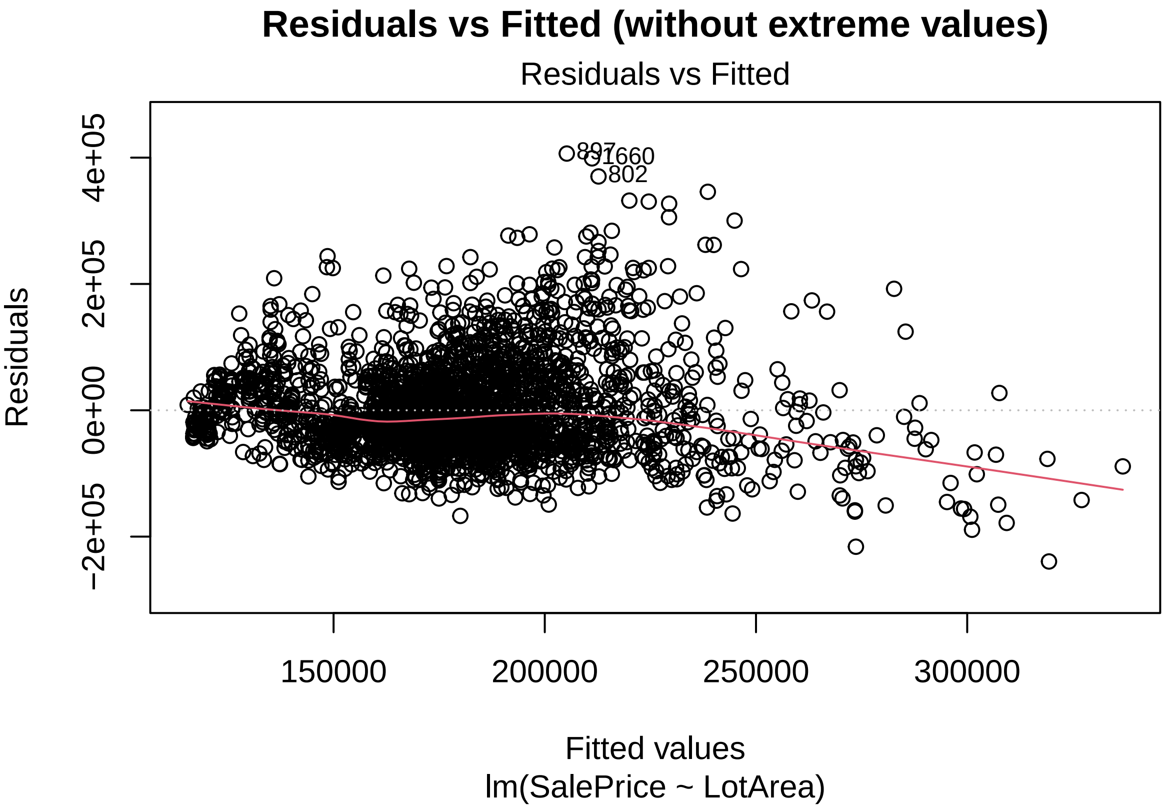
```
## [1] 0.150208
```

Here we have the scatter plot of the data without extreme values and the linear model of that data:



Below we have the two diagnostic plots for the linear regressions we performed:





The second model is preferred because the proportion of variance explained by the second model is higher than that of the first model. Looking at the diagnostic plots above, we also see that the first model has a more uneven spread of residuals around the mean, whereas the second model is more even (although there is still some curves which indicate that a linear model might not be the best). Hence, we would rather pick the second model.

Here, we see the coefficients of the estimated equation and calculate the predicted price:

```
coefficients <- linear_regression_no_extreme$coefficients
predicted_price <- coefficients[1] + coefficients[2]*15000
```

The estimated equation is `y = 7.72x + 1.05*10^5`, rounded to 3 significant figures, where `y` is the `SalePrice` and `x` is the `LotArea`.

The predicted price for a house with a lot area of 15000 is 221282.9.

Note that in this calculation, we did not use the rounded figures.