CMDA-3654

Final Report

Michael La Vance

8/13/21

Introduction

Everyone needs housing and different buyers value certain aspects of the house more than others. Using price as a quantification of desirability, it is possible to see correlations between features and price to determine whether or not the land will be desirable or not. The discussion will primarily focus on California and Virginia.

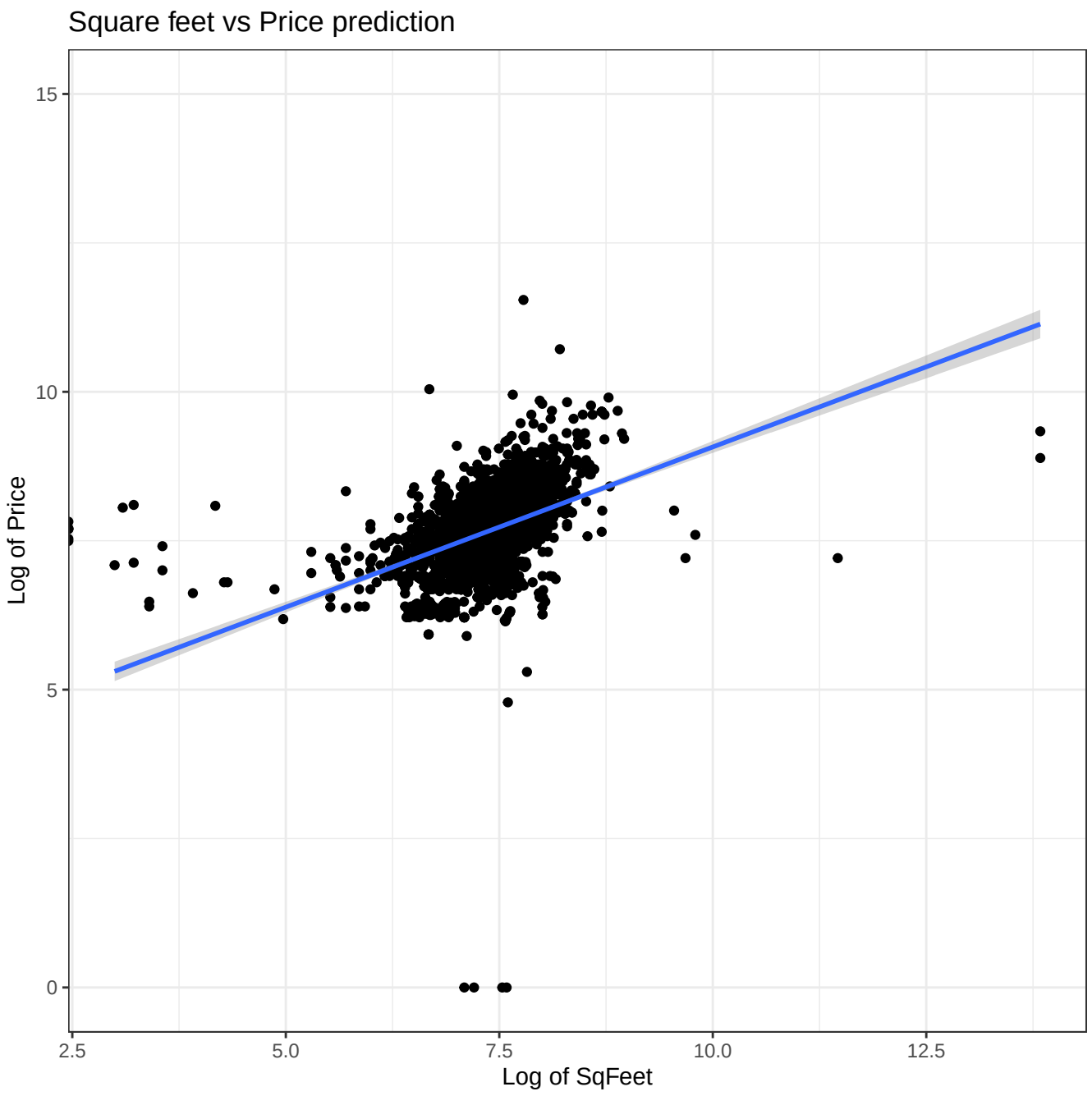
Description of the Data

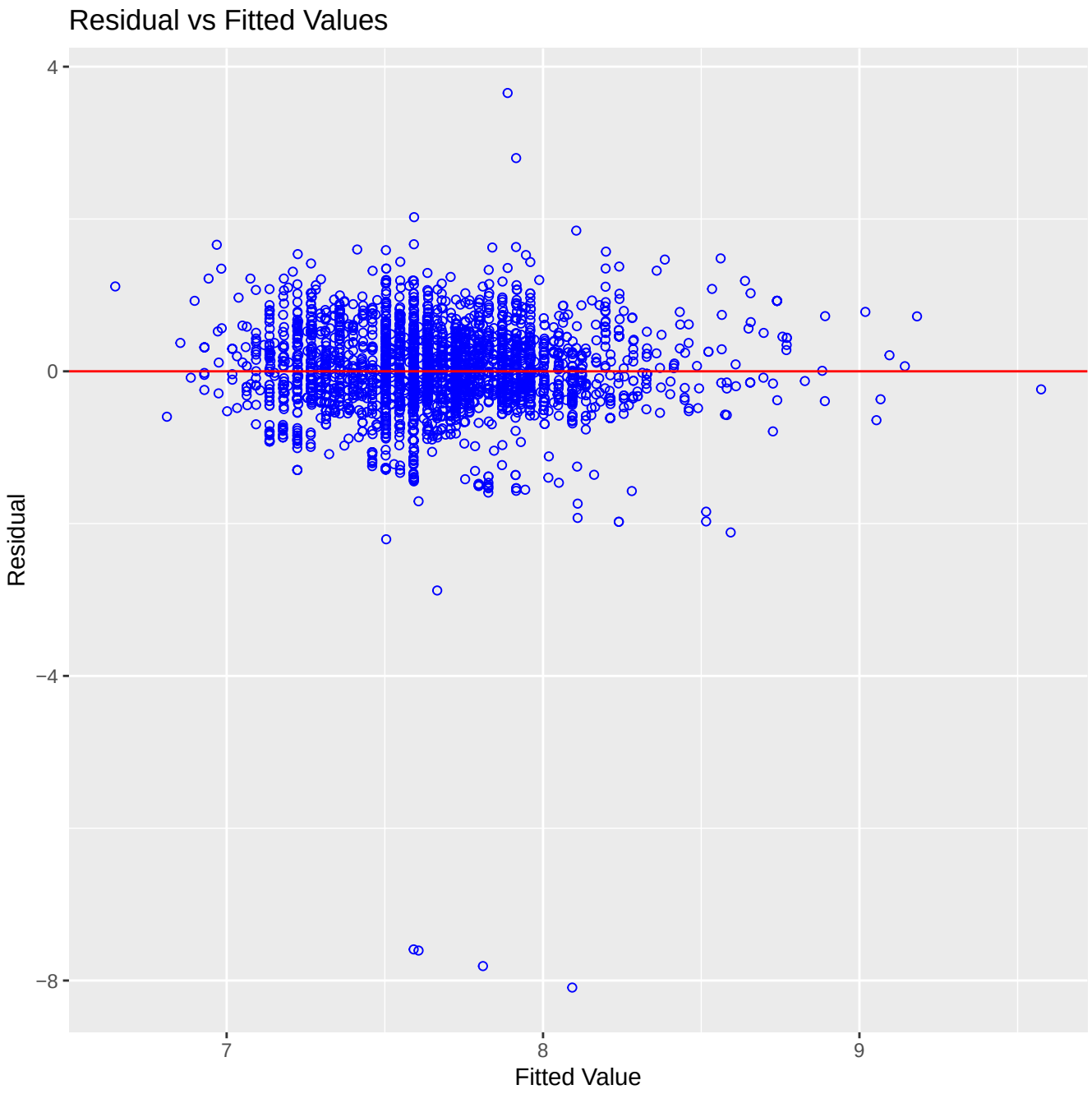
The data for this report was taken from Kaggle, but the data itself is a compilation of of craigslist listings through the United States. It includes the price and other qualitative variables about the listing. The price will be the response throughout the investigation and will be compared to other the other factors. The variables that will be used for this report are: price, type of housing, sqFeet , number of beds, number of baths, if cats are allowed, if dogs are allowed, if smoking is allowed, if it is wheel chair accessible, if you can charge electric vehicles, if it is furnished, options for laundry, options for parking, and state.

Methods

I used regression to create a model that allows for easy comparison between the price increase due to categorical variables. I had also planned to use clustering to see if states tended to have the same priorities, but the variables created started to require >8GB to store in memory. The hardware used to create this report has a system memory of 16GB, so this approach was unreasonable. As the predictions are multivariate, showing a graph of only the price vs one variable will be unable to show the predictive power of the model, residual graphs, the difference between the predicted value and real value, will be shown during the analysis.

Results and Discussion

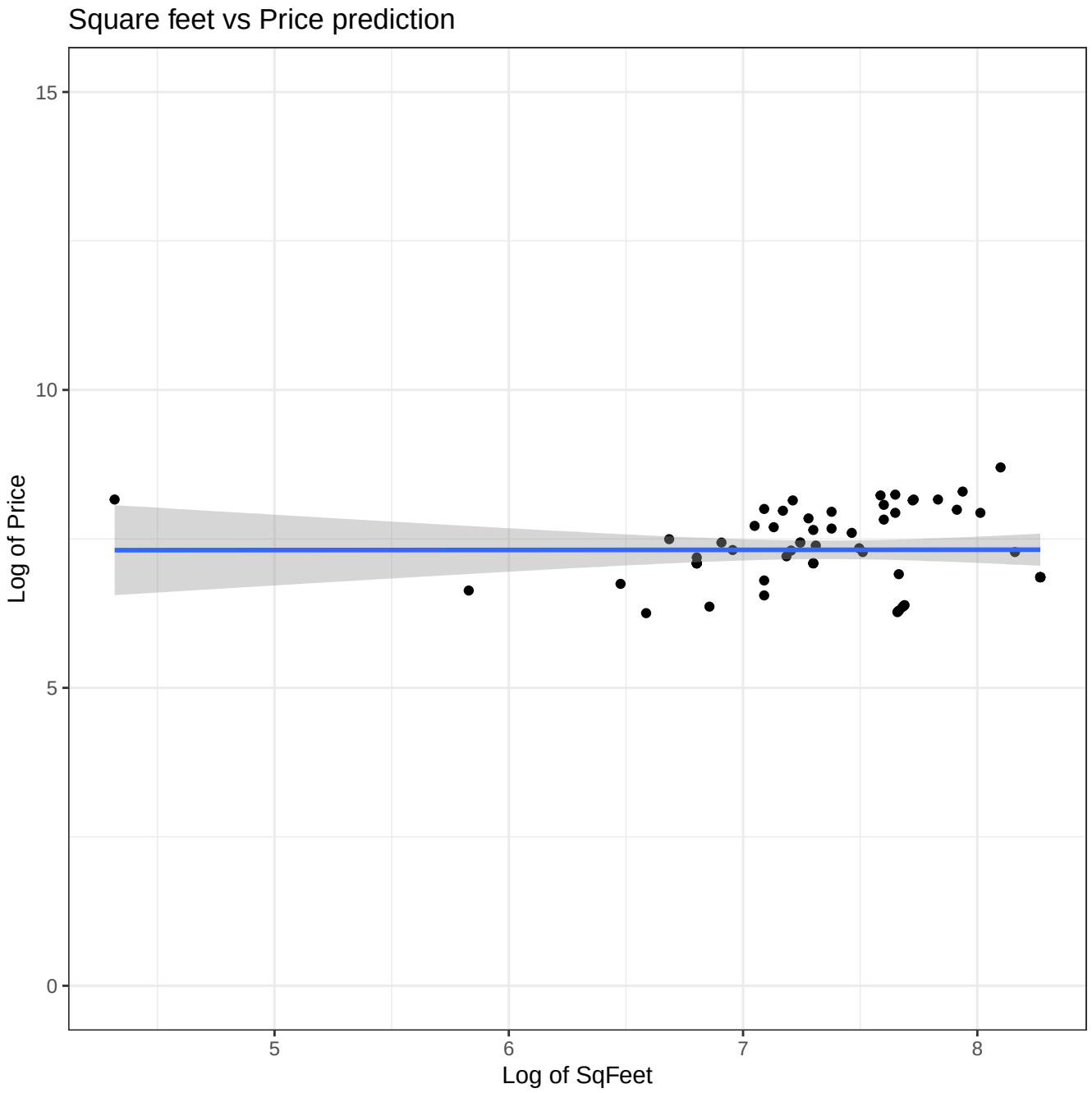


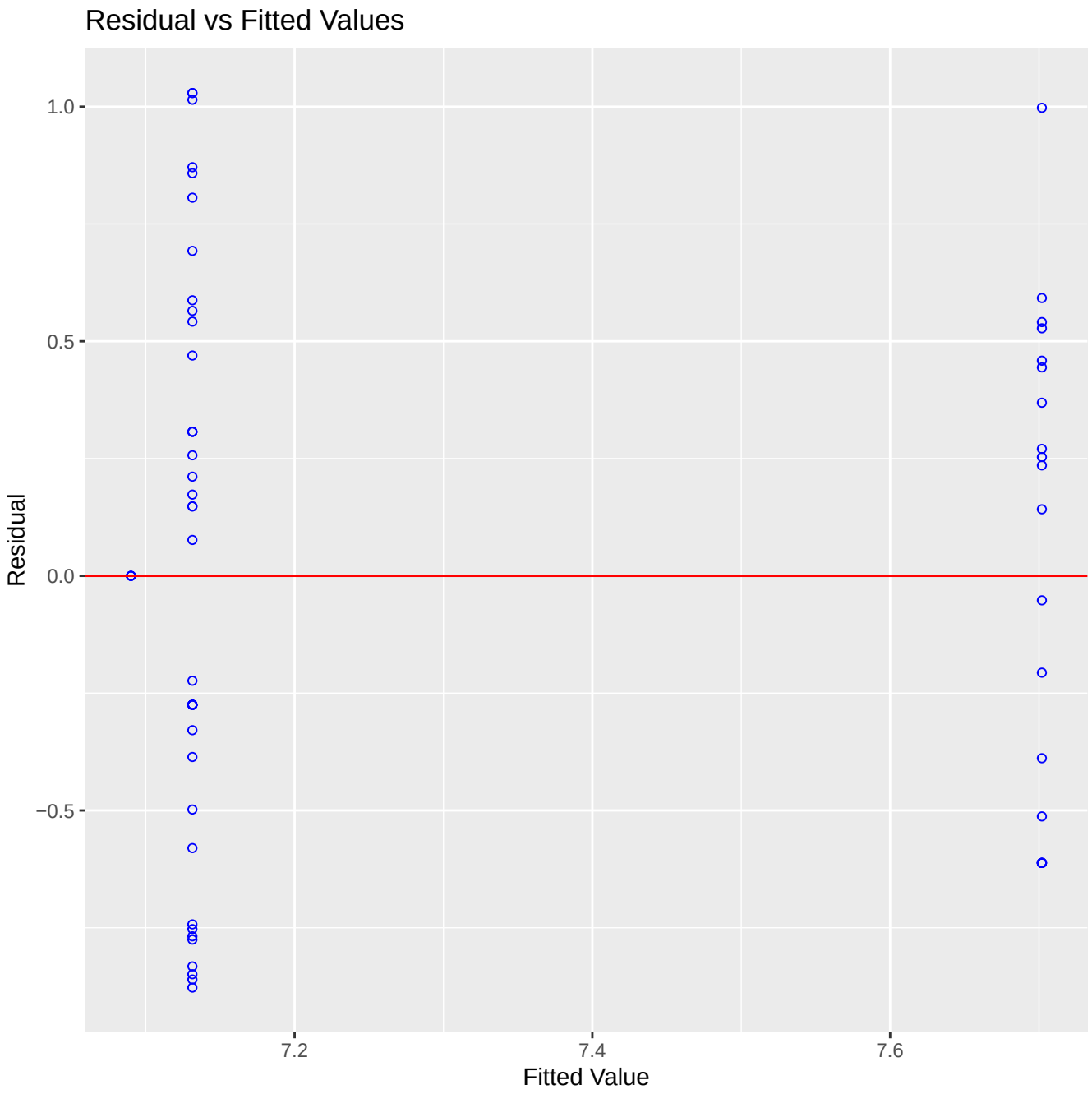


California Housing Prediction

The storage of the regression became rather messy when automating the operation, so the table will only include certain states. The regression prediction plot looks like a bad fit, but that is only because the model includes more predictive variables than only square feet. This model suggests that the number of baths has the strongest impact on pricing, predicting an increase of $1,181 per bath, wheelchair access is the strongest decrease in price at a predicted $-501 if the house is wheel chair accessible. The p-value for this particular model is <2e-16. The p-value for the number of bath changing the price is also <2e-16.

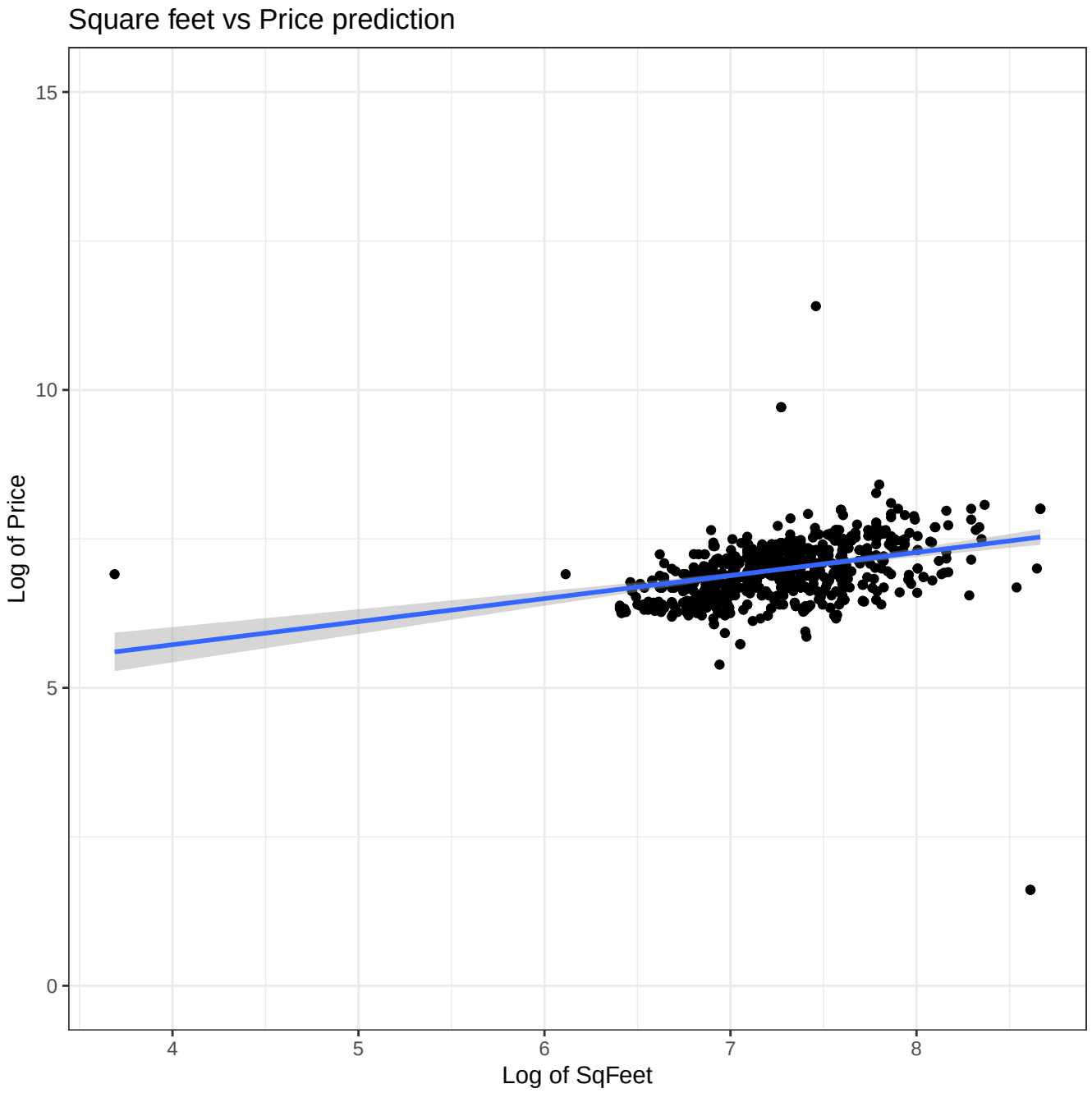
Washington DC housing Prediction

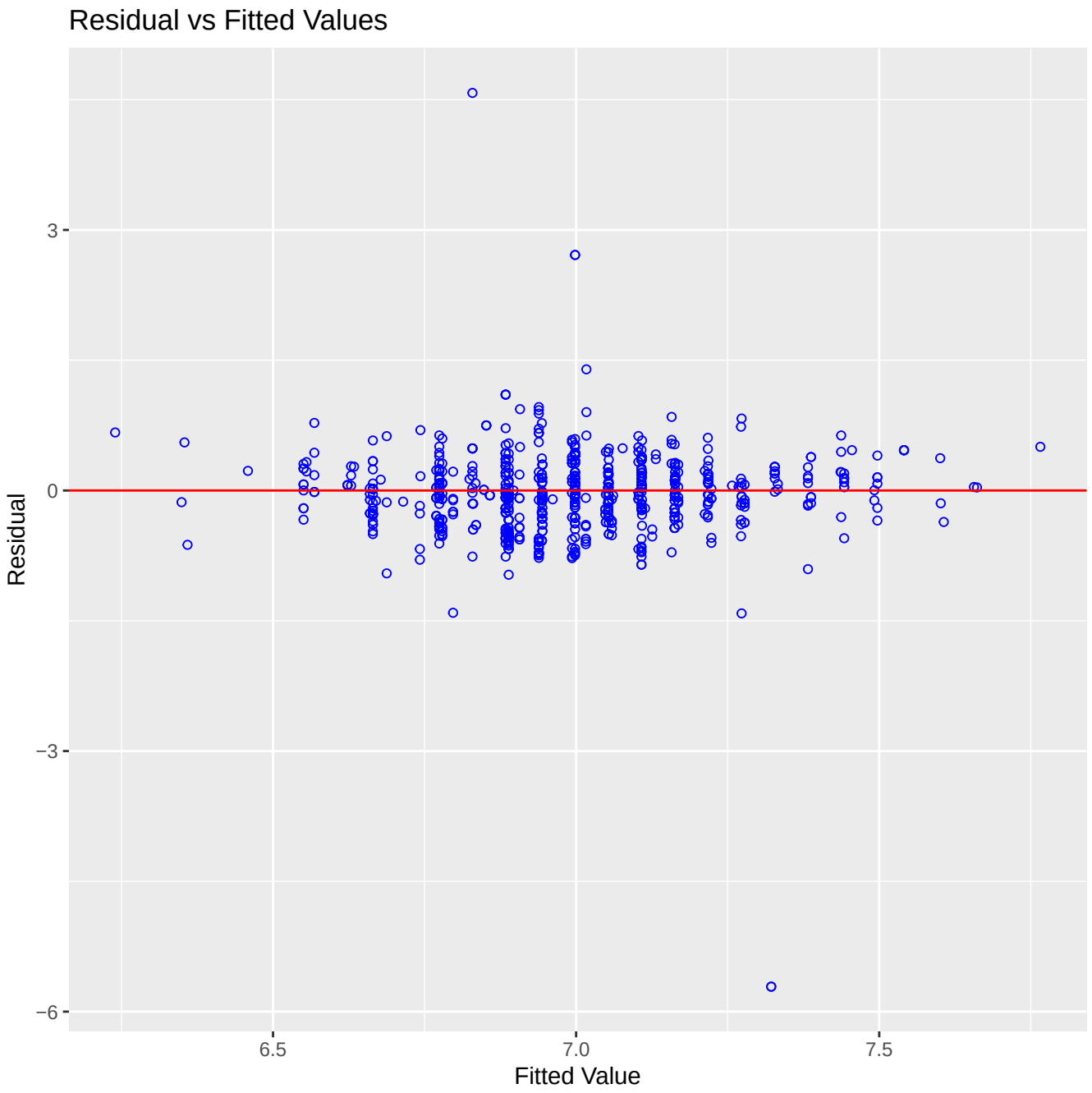




The model created by this data for DC only looks at 3 Boolean variables, if smoking is allowed, if it comes furnished, and if it can charge electric vehicles. Each success reduces the predicted price by $1270, $665, and $962 respectively. While this model is still unable to be used reliably in a real world scenario, the p-value for this model is %.09, so it is better than nothing for predicting within our data set.

Virginia Housing Prediction





This model only looks at total square feet of the house to predict the price, with an increase of \$.15 every foot^2. the p-value for this model is 3%. There were quite a few models which did not make sense and some states did not have any model which could be used to predict the house price. For example one particular model was only using the number of beds as a way of reducing the total price. The residual plot looks to have a wave pattern involved inside of it, this could suggest that there is another variable that would help improve the model, or the issue could lie in the data.

Conclusions

Some of the states models lacked observations or necessary variables to create strong predictive models, but some were able to provide models that showed the given states priorities in housing. The model showed that California housing listings tended to increase the most with the number of baths available, and the model for Virginia sowed us that the only reliable predictor for this state is the feet^2 of the house.

Citations

Data: https://www.kaggle.com/austinreese/usa-housing-listings

Appendix

```
#reading Data
housing <- read.csv("housing.csv")

#removing extras
house <- housing[,c(5:17,22)]

#removing everything for 0 price
house$price[house$price == 0] <- NA
house<-house[complete.cases(house),]

#applying log of price to make graphs look better
house$price <- log(house$price)

#modifying levels

levels(house$type) <- unique(house$type)
levels(house$laundry_options) <- unique(house$laundry_options)
levels(house$parking_options) <- unique(house$parking_options)
levels(house$state) <- unique(house$state)

#creating housing subsets
temp <- 1:length(unique(house$state))
modfit_best <- temp

#getting only houses
house.house <- subset(house, house$type == "house")

statenames <- unique(house$state)
#finding best regression model by state houses
for(i in temp){
  modfit0 <-lm(price ~ 1, data = house.house[house.house$state==statenames[i],
                                             c(-2,-12:-14)])
  modfit_full <- lm(price ~ ., house.house[house.house$state==statenames[i],
                                           c(-2,-12:-14)])
  modfit_best[i] <-
  stepAIC(modfit0,  scope = list(lower = modfit0, upper = modfit_full),
          direction = "both",trace = 0)
}
#getting specifically California
#these last sections were used again in the DC and VA sections
#with slight variable changes
modfit0.ca <-lm(price ~ 1, data = house.house[house.house$state=="ca",
                                              c(-2,-12:-14)])
modfit_full.ca <- lm(price ~ ., house.house[house.house$state=="ca",
                                            c(-2,-12:-14)])
modfit_best.ca <-
  stepAIC(modfit0.ca,  scope = list(lower = modfit0.ca, upper = modfit_full.ca),
          direction = "both",trace = 0)


#creating dataframe
best.bystate <- cbind(modfit_best,statenames)
best.bystate <- as.data.frame(best.bystate)
```

```r
#creating graphs
house_with_pred <- data.frame(
  house.house[house.house$state=="ca",c(-2,-12:-14)],
  predict(modfit_best.ca, interval = "predict"))

ggplot(data = house_with_pred, aes(x = log(sqfeet), y = price)) + geom_point() +
  geom_smooth(method = "lm") + theme_bw() + ylim(0, 15) +
  ggtitle("Square feet vs Price prediction") + xlab("Log of SqFeet") +
  ylab("Log of Price")

#residual plot of California model
ols_plot_resid_fit(modfit_best.ca)
```