# Home Insurance

*What affects home insurance premiums?*

By: Sydney Marino

STA 9890

## Our Foundation
## The Data

Source: https://www.kaggle.com/ycanario/home-insurance#home_insurance.csv

### Raw Data

N = 256,136
P = 65 Predictors
Target Variable = Last_Ann_Prem_Gross

### Reasons for Removing

1. Predictor value would be known after the fact
2. Does not add additional info
3. Not enough data

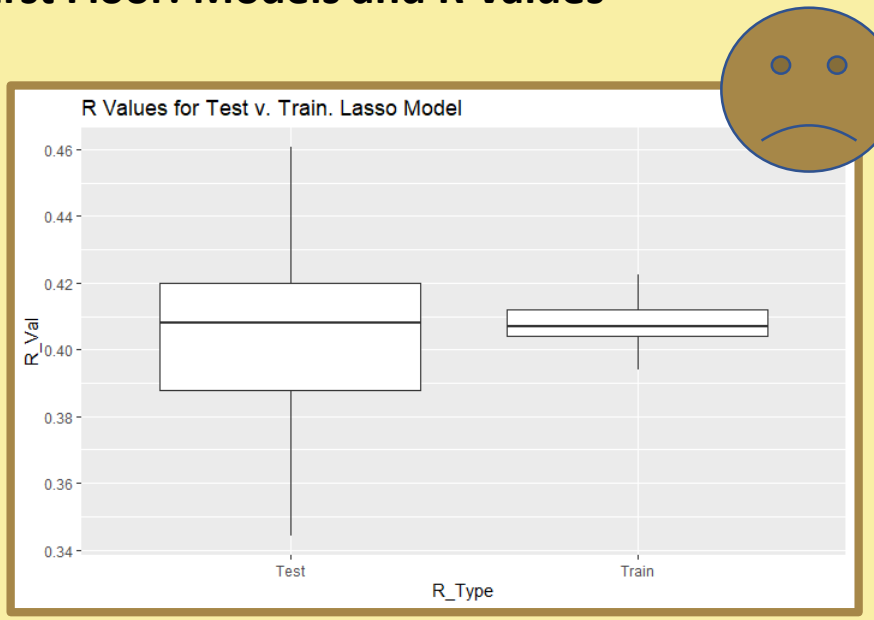
### Cleaning Data

1. Removed rows that were incomplete
2. Binarized categorical data
3. Condensing Features
4. Standardizing Numeric Features

### Final Dataset

N = 133,201
P = 44 Features (68 with Binarized)
Target Variable = Last_Ann_Prem_Gross (Numeric)
Shape: Gaussian (After standardizing)

# First Floor: Models and R Values



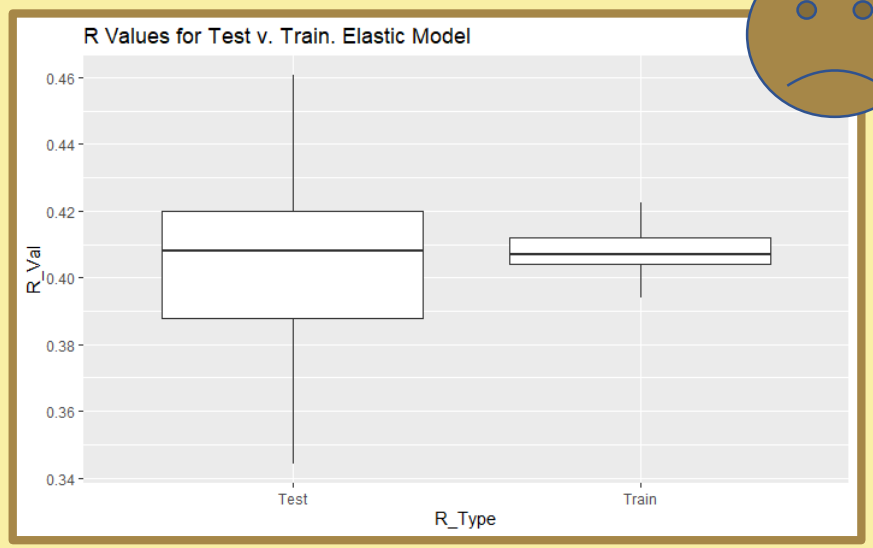

**4 Models**

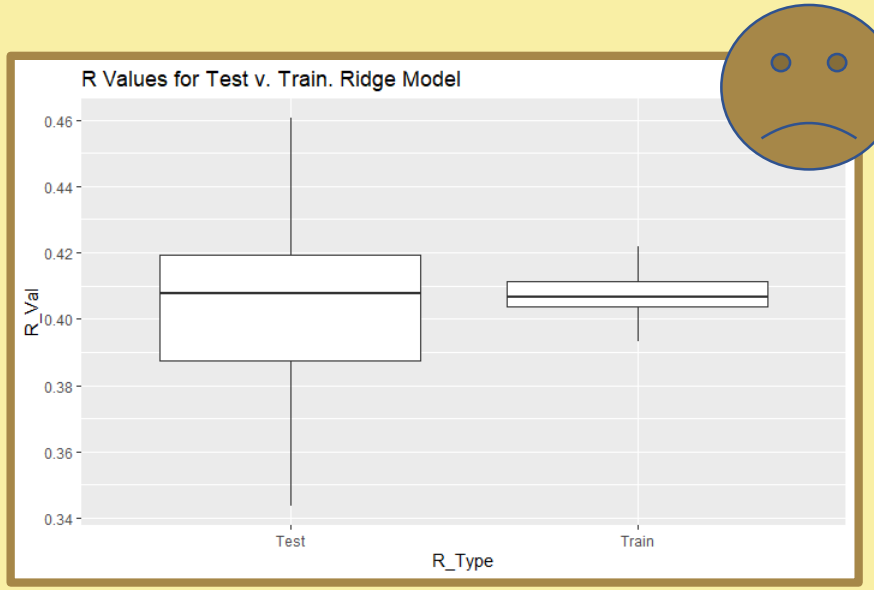
1. Lasso
α = 0

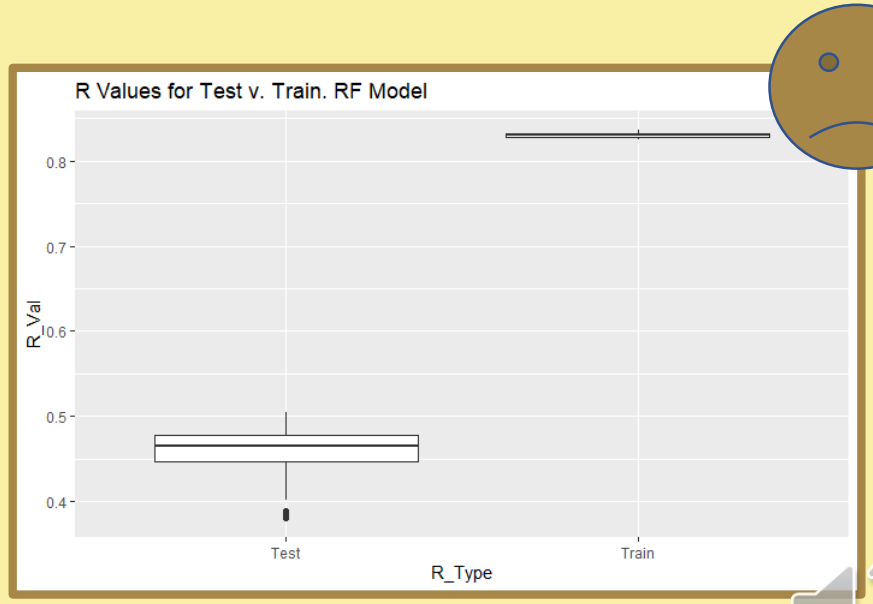
2. Ridge
α = 1

3. Elastic
α = .5

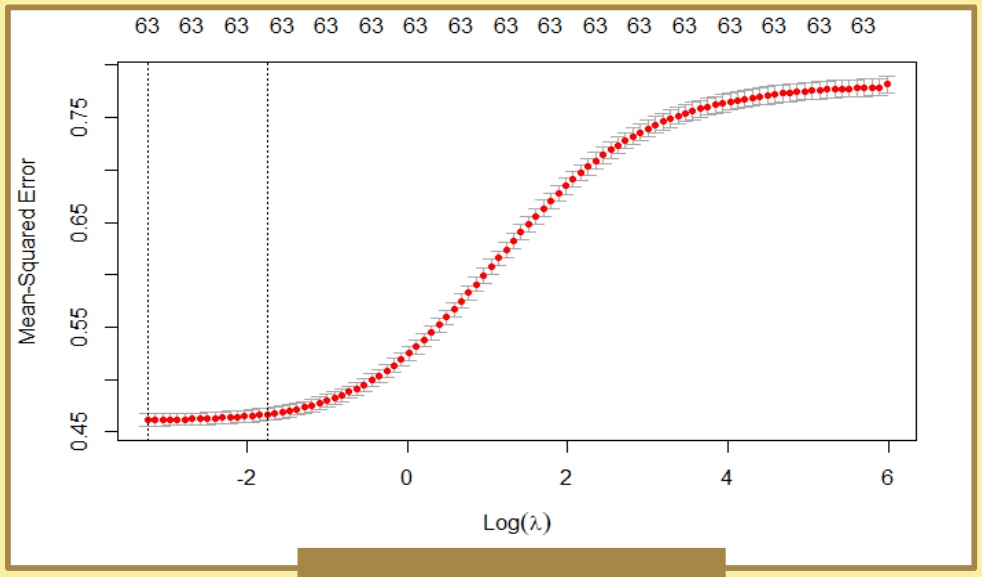
4. Random Forest
Ntree = 25
Mtry = p/3



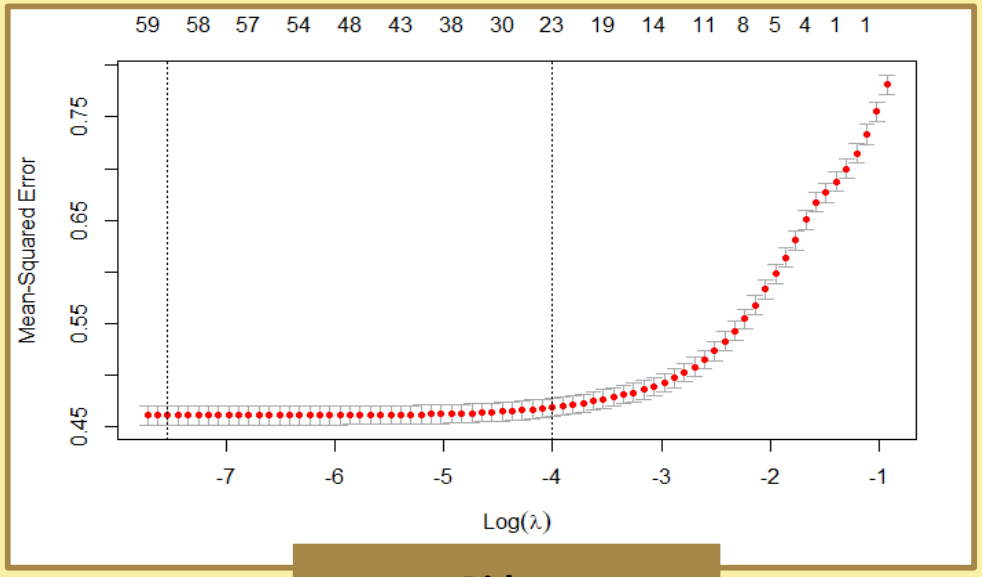



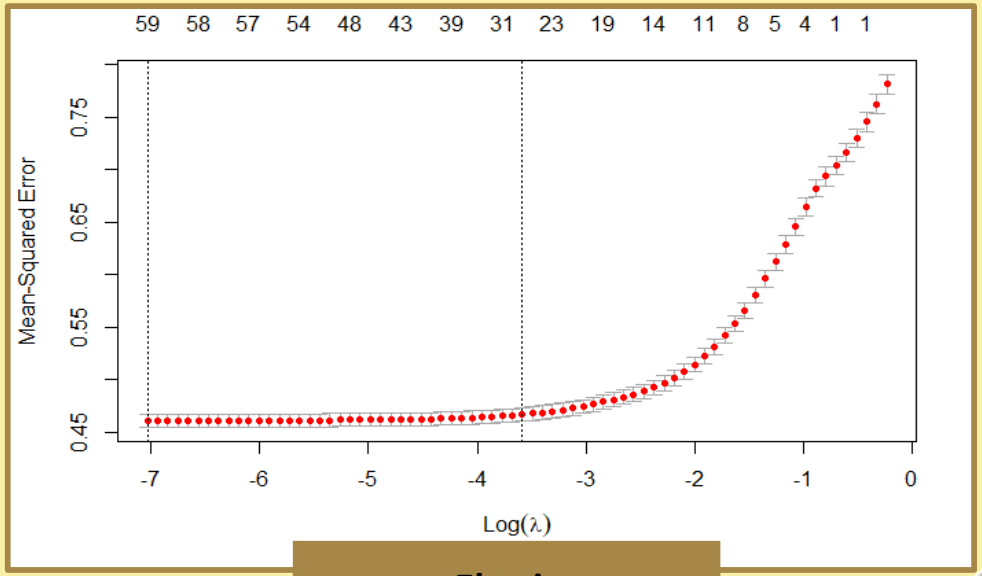
## First Floor: CV Curve




**Lasso**
λ = .039474




**Ridge**
λ =.000534102




**Elastic**
λ =.0008868418

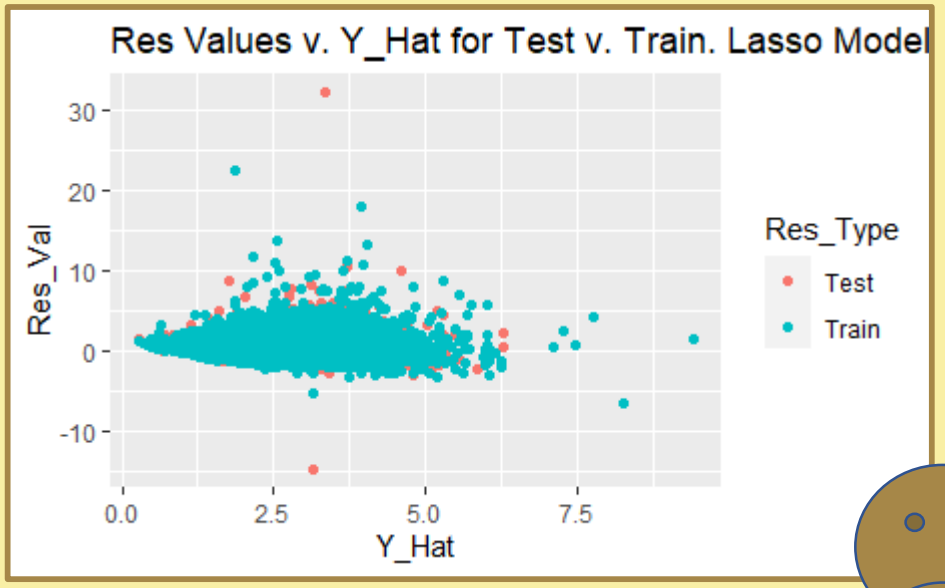
### Summary

Larger λ –> Higher Penalty

Lasso –> Highest λ Means More Variables Eliminated

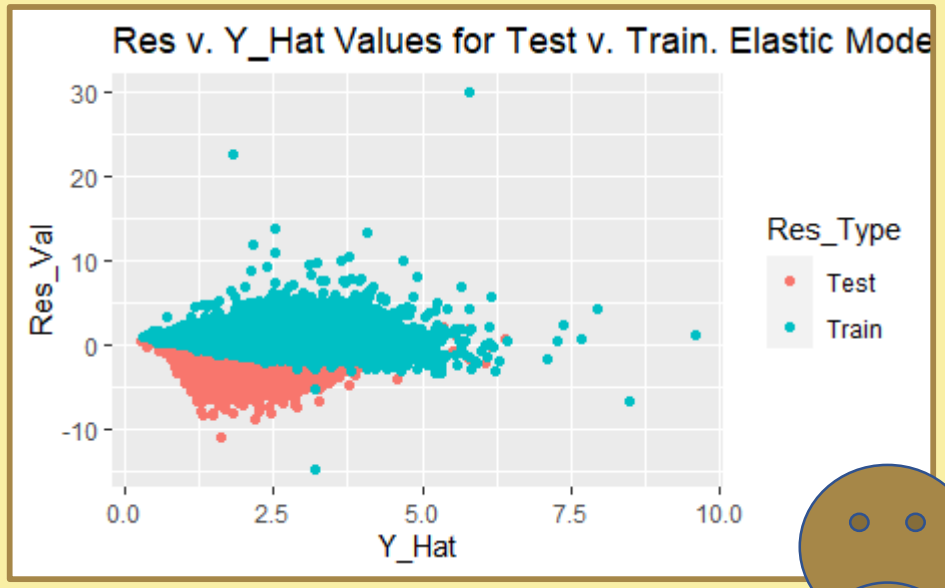
Ridge and Elastic –> Lowest λ, very close to 0, keeps most features

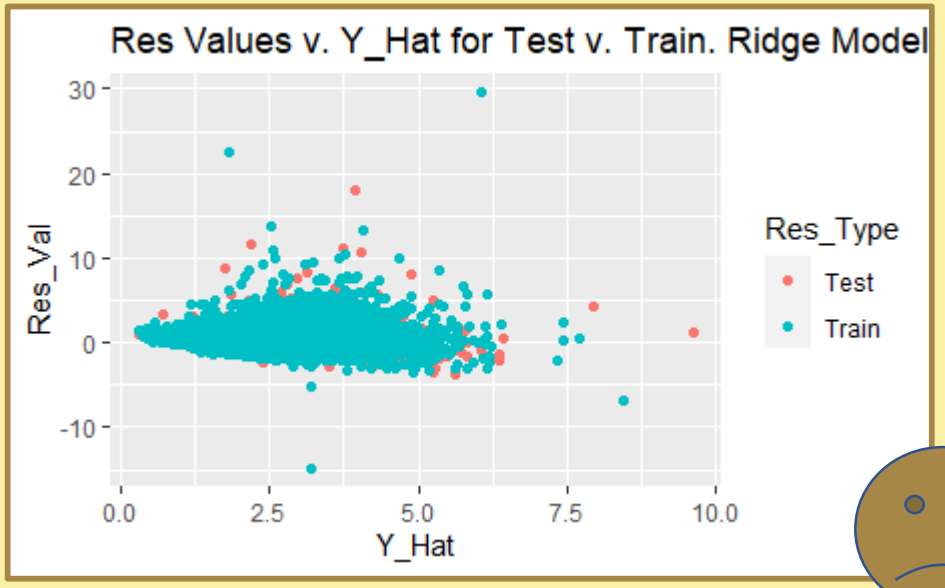
## Second Floor: Residuals




All Models are Not Good

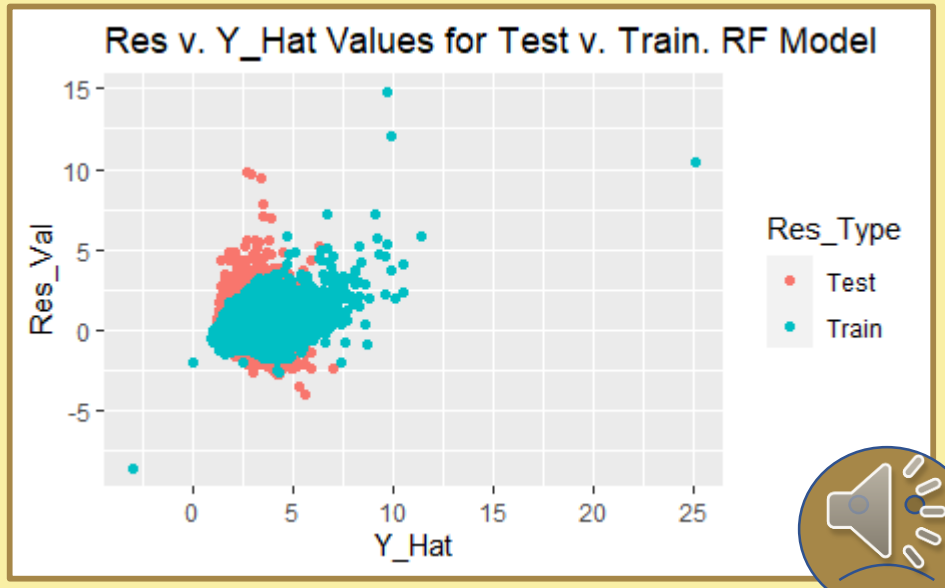
Patterns in Residuals = Bad



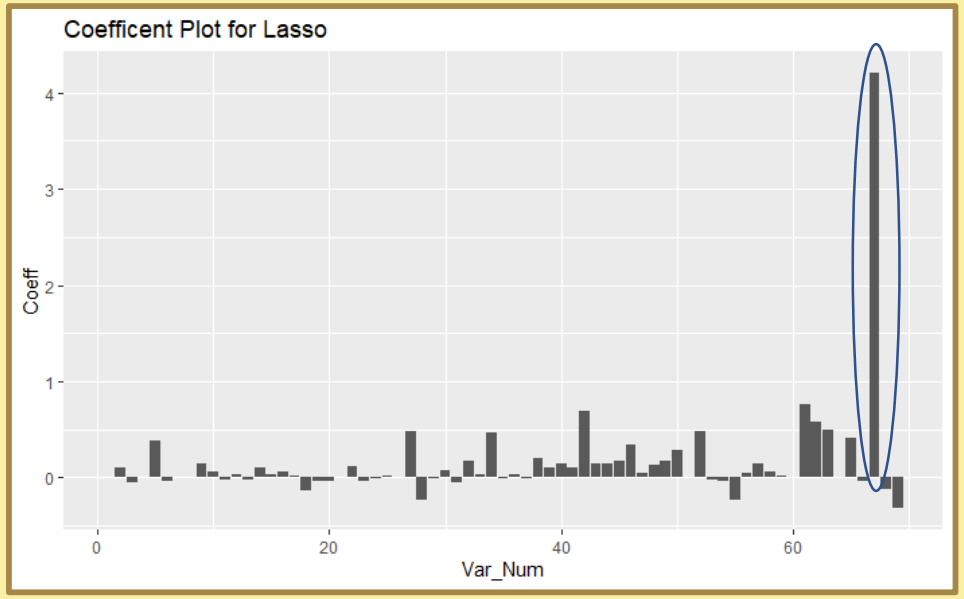




Lasso, Ridge, Elastic – Show a slightly decreasing line

RF – Shows a slightly increasing line



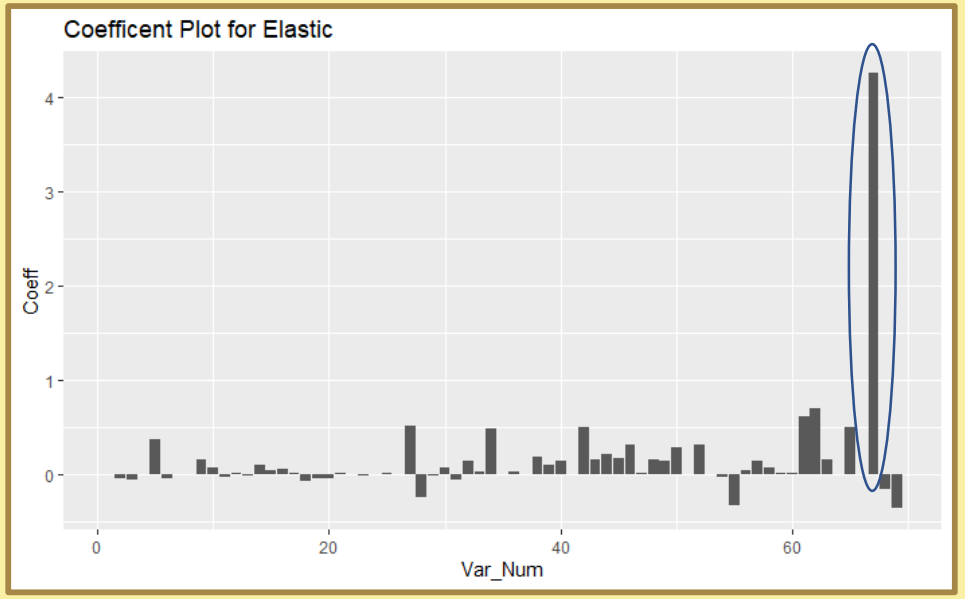
# Second Floor: Feature Importance

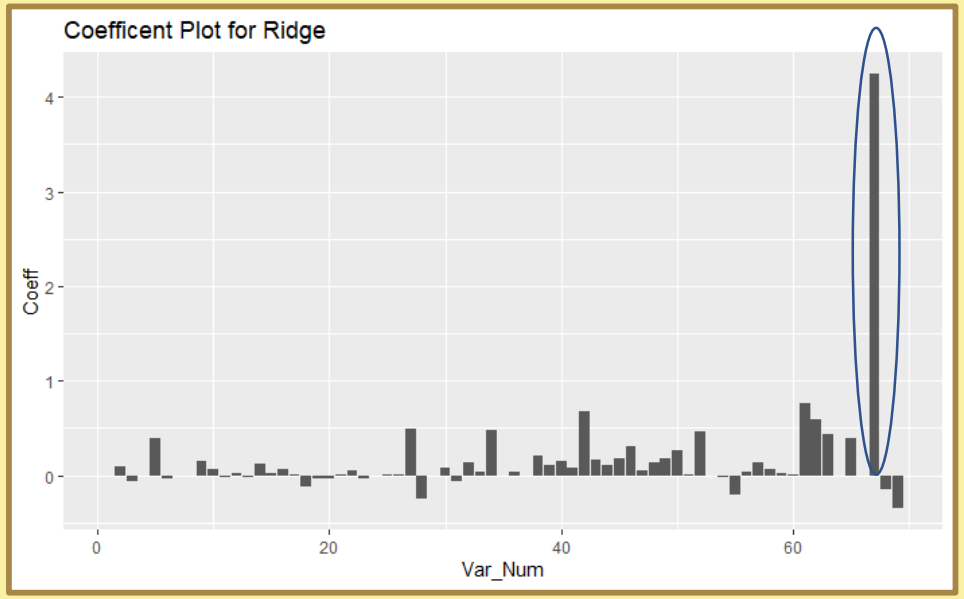



**Most Important Var**

Lasso, Ridge, Elastic:
HomeAgeCat102_to_104

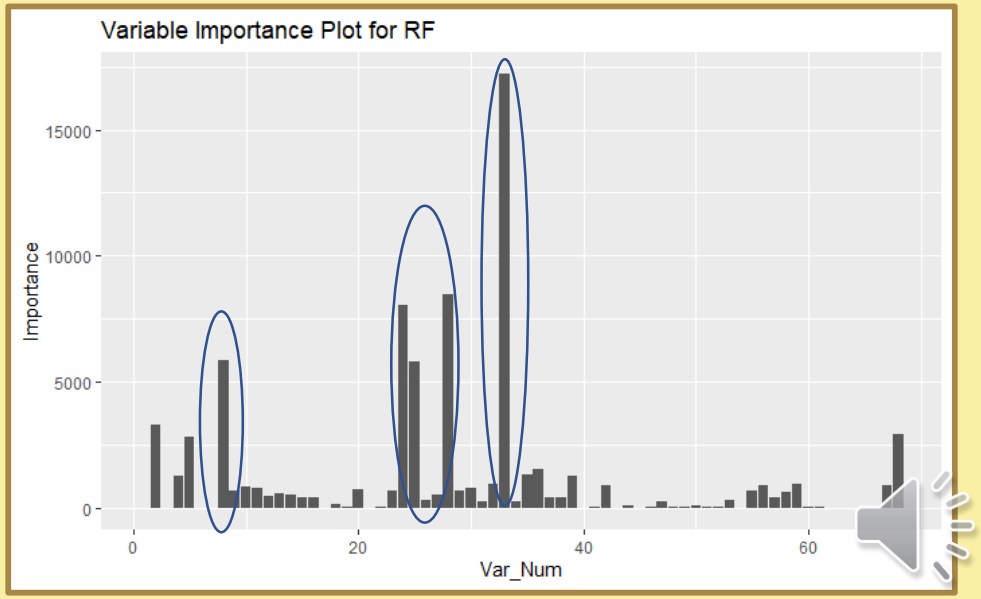
RF: Bedrooms




RF places importance on more variables than other modeling

At least 5 Variables considered important compared to others



## Roof: Summary

| Model | Perf Time | Comments |
|---|---|---|
| Lasso | 5 Mins or Less | - Best model compared to Elastic and Ridge<br>- Highest lambda compared to Elastic and Ridge – keeps more vars<br>- Residual Plot similar to Elastic and Ridge |
| Elastic | 5 Mins or Less | - Elastic and Ridge were practically the same<br>- Worst models<br>- Residual plot is decreasing linear, may want to do a first order regression instead of Lasso/Elastic/Ridge |
| Ridge | 5 Mins or Less | |
| Random Forest | 1 Day | - Very strenuous in computing power<br>- Favors more variables<br>- Most likely overfitted<br>- Better residual plot than others |

**Recommendation:**

- Research into boosted model, may work better
- Fit a first order linear regression, may perform better than others
- All of these models are not the best