???

...

Margarita Mondays

Tequila Tuesdays

_____ Wednesdays ?

Thirsty Thursdays

Fireball Fridays

Soju Saturdays

Sober Sundays

WINE

Predicting Wine Points
A Wine Enthusiast dataset

...

Brian Kim, Tim Tan, Ishan Supanekar, and JJ Goh

Motivation of project:

"To gain a deeper insight on which variables deem a high wine score from Wine Enthusiast"

Our Data (numerical)

Variable	Description
points	The number of points Wine Enthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >= 80)
price	The cost for a bottle of the wine
year	The vintage of wine (pulled from title)

Our Data (categorical)

Variable	Description
country	Country of origin
province	The province or state that the wine is from
region_1	The wine growing area in province or state (ie Napa)
taster_name	The taster/reviewer
variety	Grape type
winery	The winery that made the wine

Our Data (cut variables)

Variable	Description
description	Flavors and taste profile as written by reviewer
designation	The vineyard within the winery where the grapes that made the wine are from
region_2	The second wine growing area in the province or state
taster_twitter_handle	The twitter handle for the taster/ reviewer

Input Error

- Blair 2013 Roger Rose Vineyard Chardonnay (Arroyo Seco)
- Price: $2013
- Points: 91

Input Error

- Suggested Retail: $30
- Markup: 6610%
- What happened?

LIMITED RELEASE

Appellation	Arroyo Seco
Vineyard	Roger Rose
Soils	Arroyo Seco & Chular Loams
Climate	Very Cool, Region I (UCD)
Alcohol	13.8%
Oak Aging	25% new French oak, 50% neutral French oak, 25% stainless steel barrels
Production	241 cases
Sugg. Retail	$30

Data Cleaning

1. Remove the Missing Values
2. Remove Missing Values from price
3. Remove the rows with Price above $1000
4. Factorize the Categorical Variables

```r
wine_ratings <- na.omit(wine_ratings)

wine_ratings <- wine_ratings %>% filter(!is.na(wine_ratings$price)) %>%
  filter(wine_ratings$price < 1000)

#factorizing the categorical columns
for (i in c(1,4,5,6,8,9)){
  wine_ratings[,i] <- as.factor(wine_ratings[,i])
}
```

Data cleaning

```r
#provinces
wine_ratings <- wine_ratings %>%
  mutate(province_ordered = fct_infreq(province),
         province_lumped = fct_lump(province_ordered,10))
```

Lumps categorical variables

```r
n_row <- nrow(wine_ratings)
b <- clean_wine_ratings$title
x <- gregexpr("[0-9]+", b)
c <- regmatches(b,x)
df <- data.frame(matrix(c))
df <- df %>% rename(year = matrix.c.)
for (i in 1:n_row){
  df$year[i] <- ifelse(grepl("[a-z]", df$year[i]),"",df$year[i])
}
wine_ratings <- wine_ratings %>% mutate(year = df$year)
wine_ratings$year <- as.numeric(wine_ratings$year)

wine_ratings <- wine_ratings %>% filter(!is.na(wine_ratings$year))
```

Created a new column for years from title

Summary Statistics (Numeric)

	Price	Points	Year
Min	4	80	1827
Max	973	100	2017
Standard Deviation	37.07	2.95	3.55
Mean	36.59	88.7	2012

Summary Statistics (Categorical):

	Most Common	Amount
country	US	35,366
province	California	19,066
region_1	Columbia Valley (WA)	3,900
winery	Chateau Ste. Michelle	193
taster_name	Roger Voss	13,144
variety	Pinot Noir	8,355

Baseline Analysis

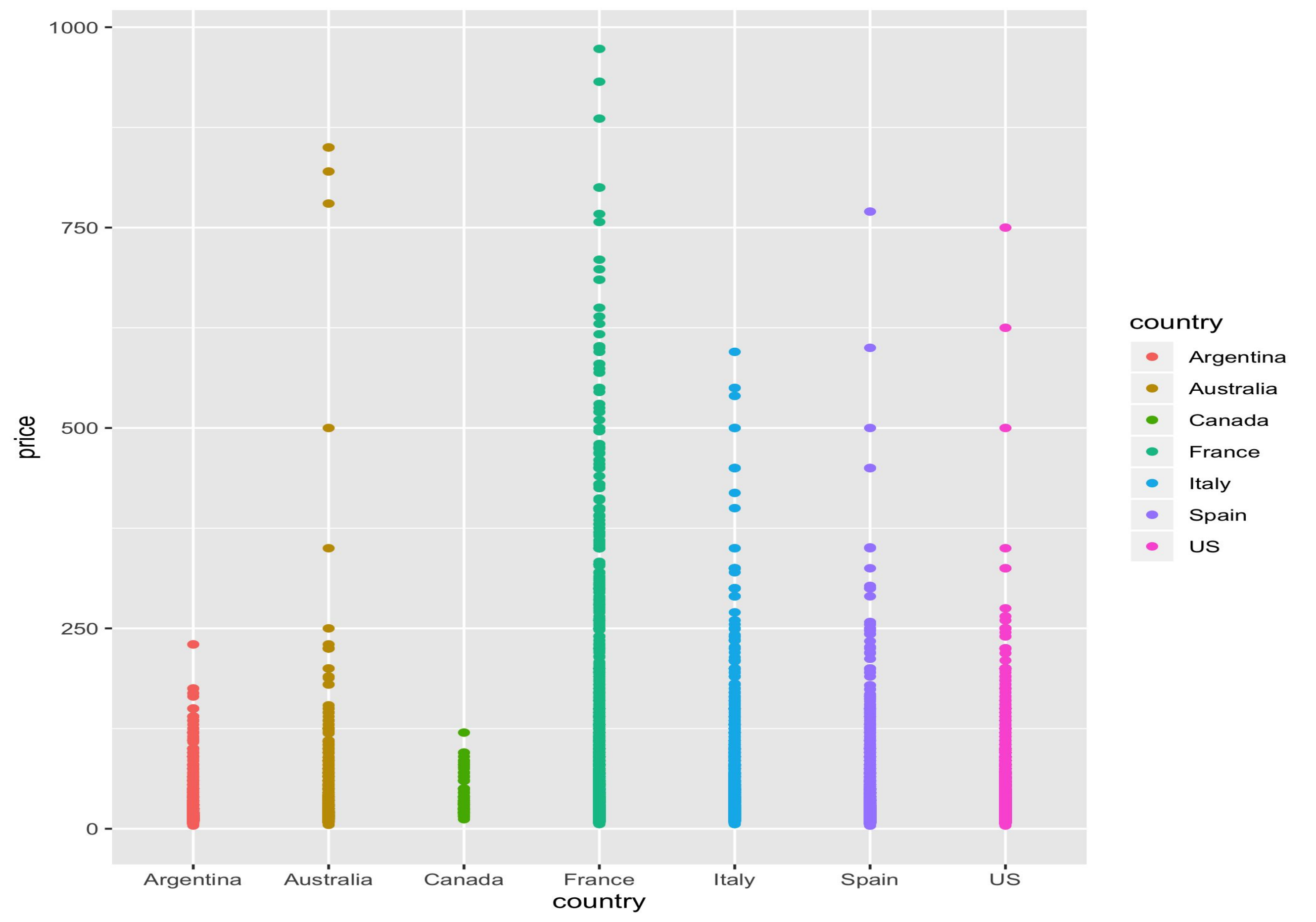
Models used

1. Backward Stepwise
2. Linear/OLS Model
3. Lasso Model
4. Random Forest Model

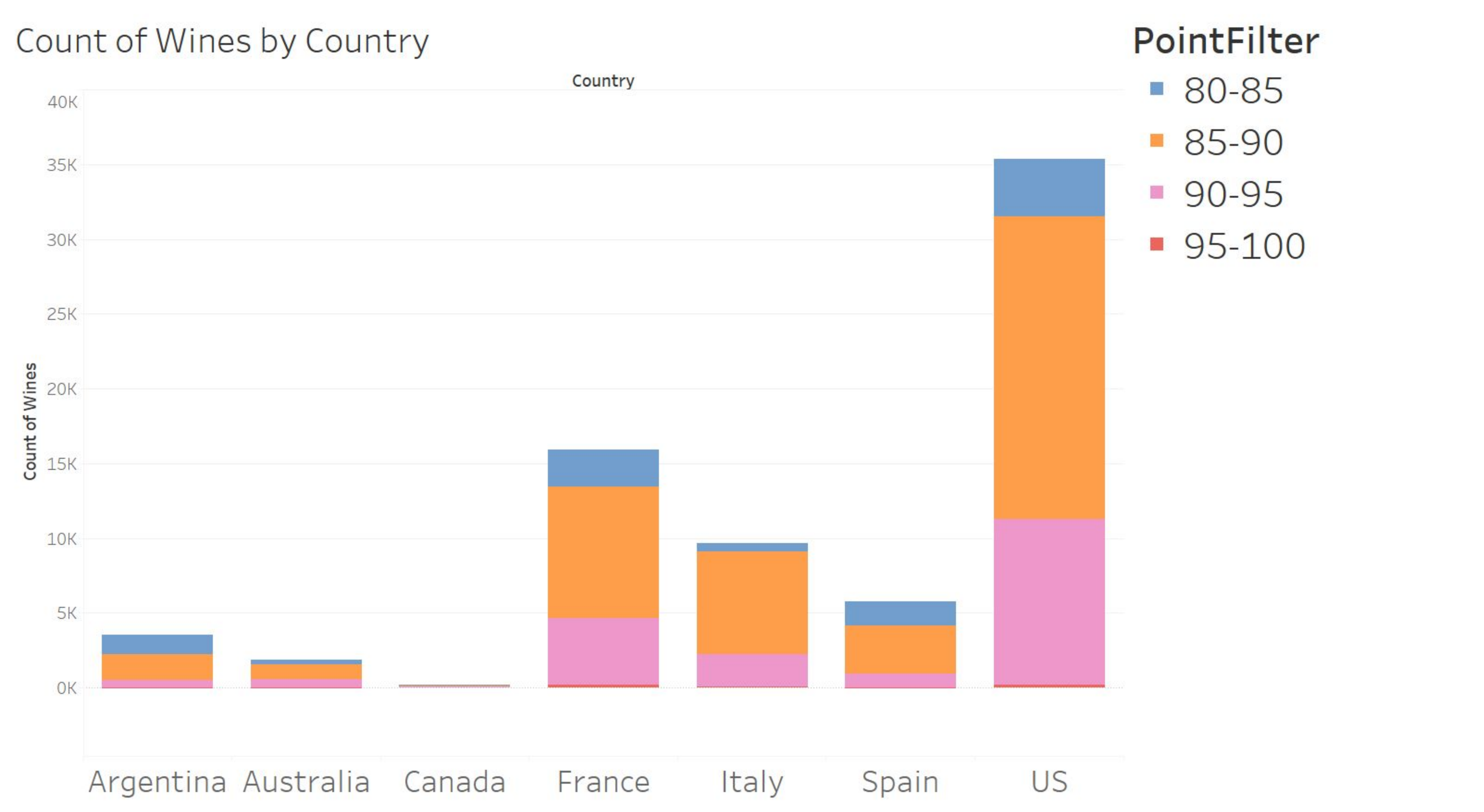
Backward stepwise to confirm findings

```r
bck_fit <-
  regsubsets(points ~.,
             data = wine_ratings_train,
             method = "backward",
             nvmax = 10)

summary(bck_fit)
```

```
price
"*"
"*"
"*"
"*"
"*"
"*"
"*"
"*"
"*"
"*"
```

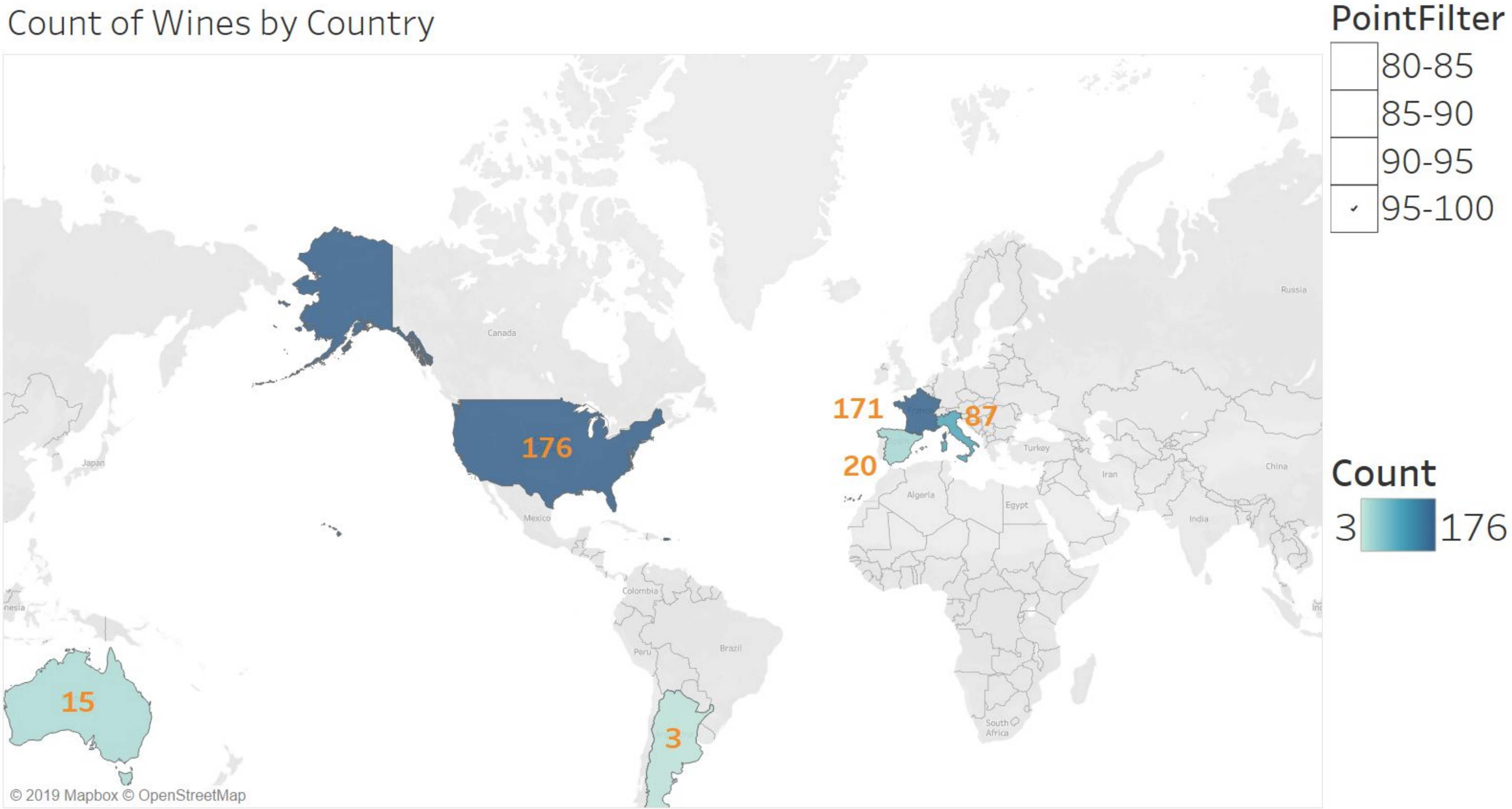


price
1000
750
500
250
0
Argentina
Australia
Canada
France
Italy
Spain
US
country
country
Argentina
Australia
Canada
France
Italy
Spain
US



Count of Wines by Country
Country
Count of Wines
40K
35K
30K
25K
20K
15K
10K
5K
0K
Argentina
Australia
Canada
France
Italy
Spain
US
PointFilter
80-85
85-90
90-95
95-100

Count of Wines by Country



Linear Model

- Predicted points against country

Coefficients:

	Estimate
(Intercept)	86.82044
countryAustralia	2.01801
countryCanada	2.84622
countryFrance	1.91106
countryItaly	2.09216
countrySpain	0.49345
countryUS	2.18616

OLS fit

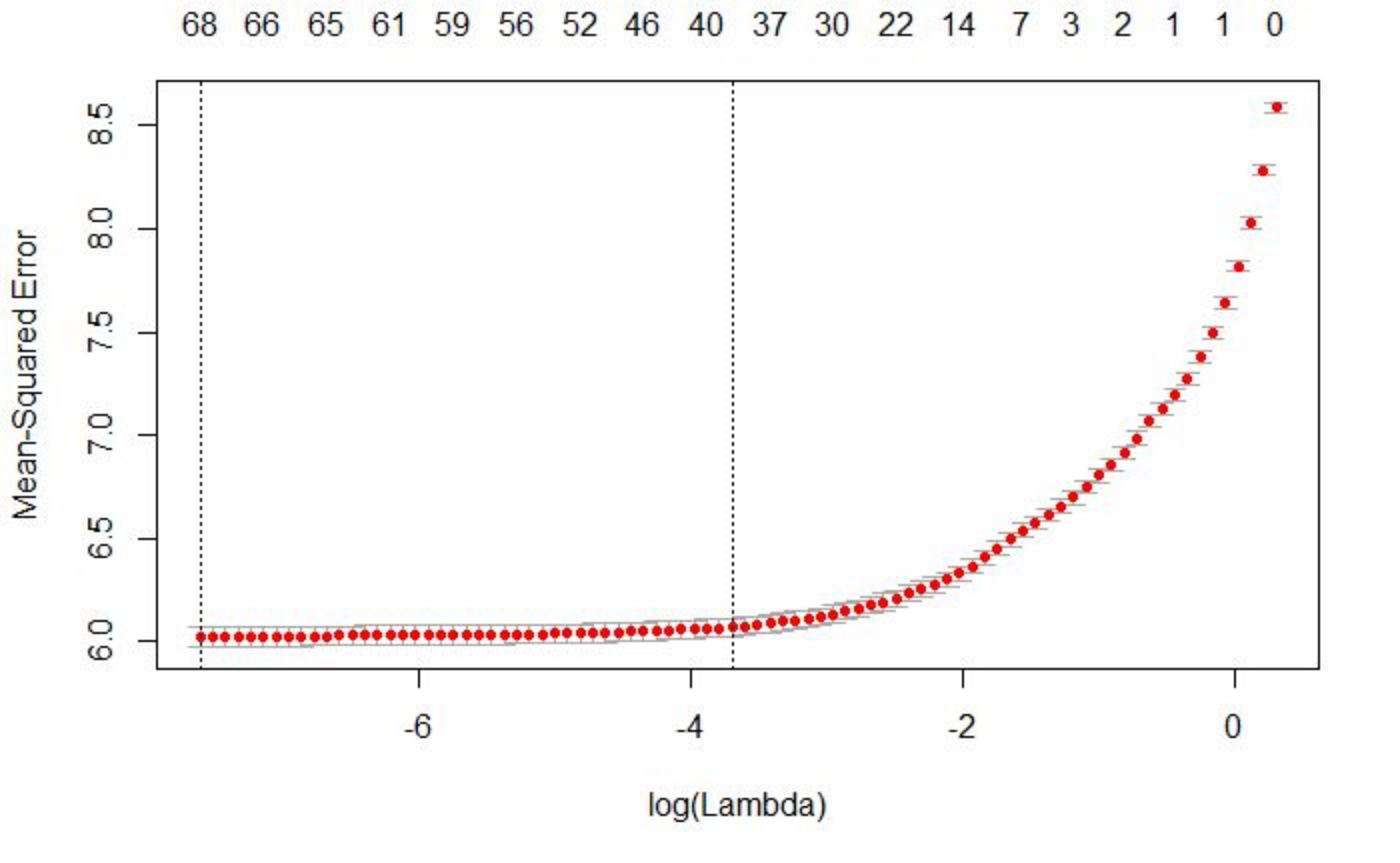
```r
ols_fit <-
  lm(points ~ .,
          data = wine_ratings_train)

preds_ols_train <- data.frame(
  preds = predict(ols_fit, newdata = wine_ratings_train,
                  type = "response"), points = wine_ratings_train$points
)
```

OLS - R-Squared / RMSE / Mean Average Error

```
R2(preds_ols_train$preds, wine_ratings_train$points)
RMSE(preds_ols_train$preds, wine_ratings_train$points)
MAE(preds_ols_train$preds, wine_ratings_train$points)
```
```

```
[1] 0.3015784
[1] 2.449904
[1] 1.957988
```
```

Lasso Model

```
Lasso_mod <-
  cv.glmnet(points ~.,
            data = wine_ratings_train, alpha = 1, nfolds = 10)
```



68 66 65 61 59 56 52 46 40 37 30 22 14 7 3 2 1 1 0
Mean-Squared Error
6.0
6.5
7.0
7.5
8.0
8.5
-6
-4
-2
0
log(Lambda)

R-Squared / RMSE / Mean Average Error

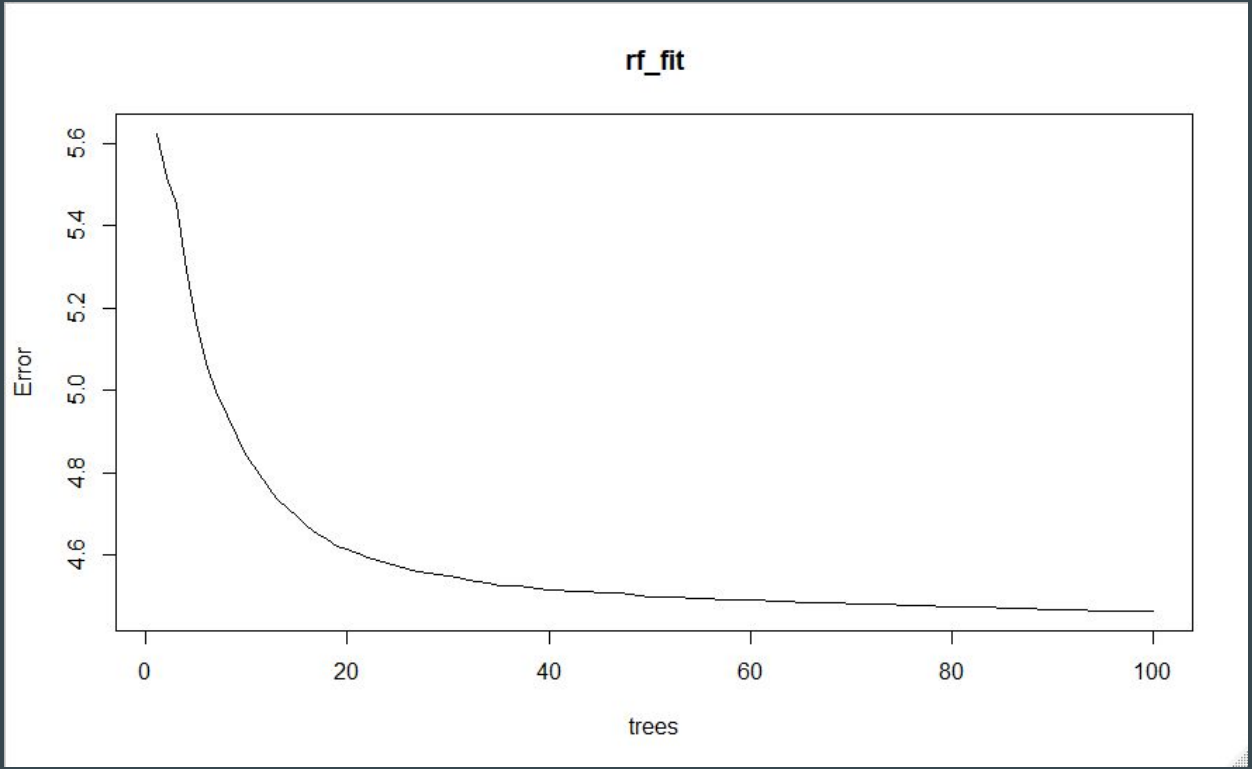
```
R2(preds_lasso_train$X1, wine_ratings_train$points)
RMSE(preds_lasso_train$X1, wine_ratings_train$points)
MAE(preds_lasso_train$X1, wine_ratings_train$points)

```
```

```
[1] 0.3013648
[1] 2.450281
[1] 1.9583
```
```

Random Forest Model

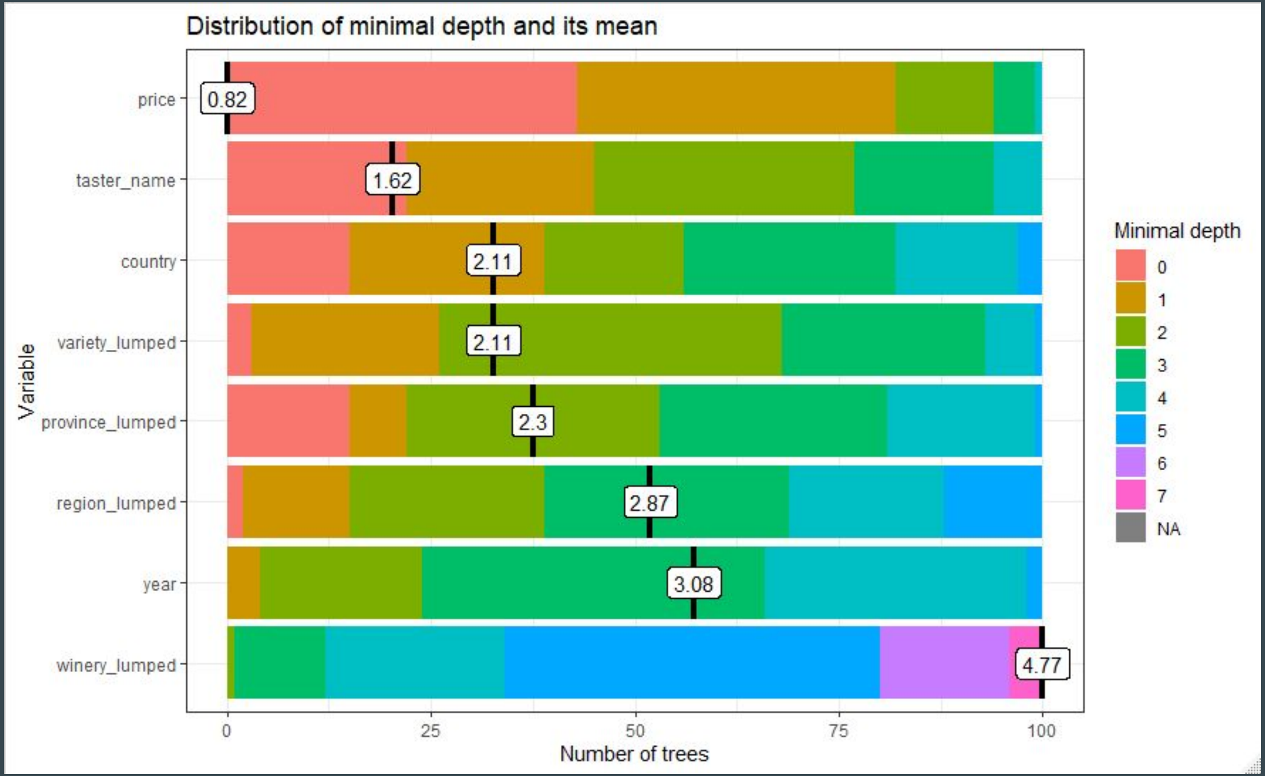
- Using 100 trees
- 3 for mtry
- Takes about 30 min to run

```
rf_fit <- randomForest(points ~ .,
                       data = wine_ratings_train,
                       mtry = 3,
                       ntree = 100)
```

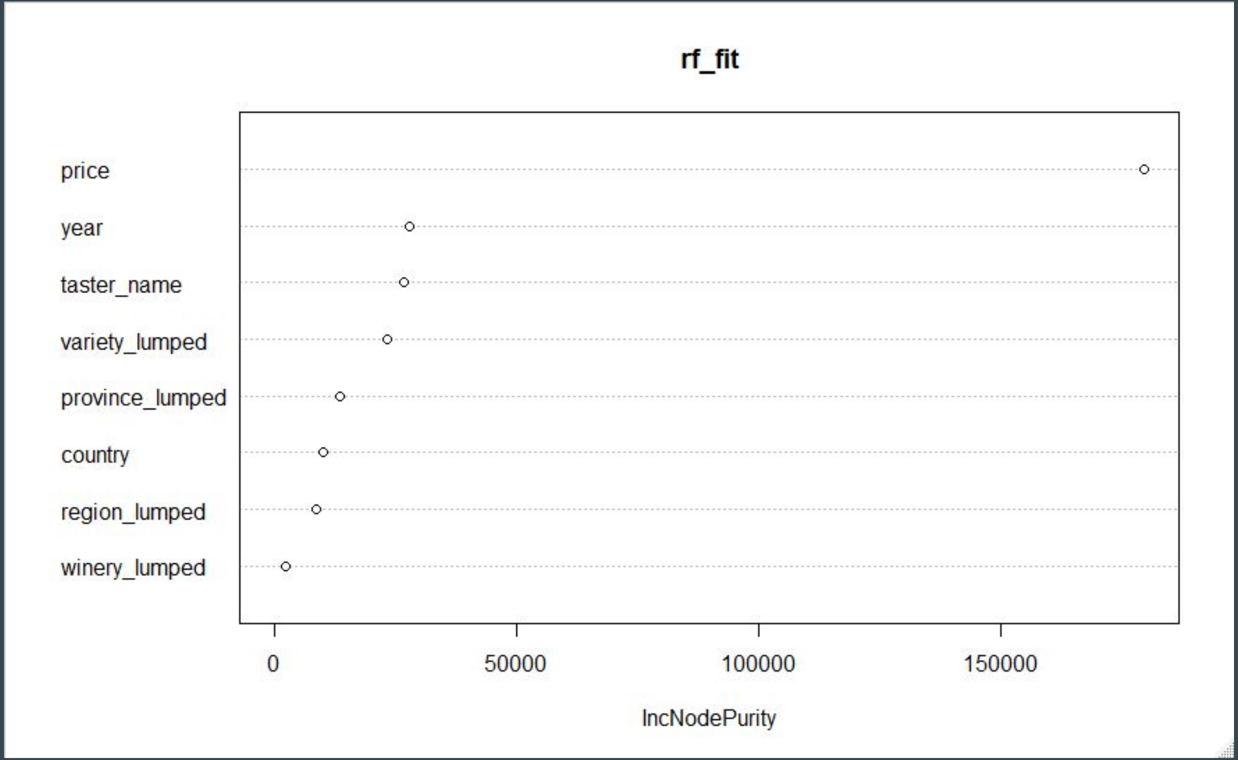
Random Forest Plot



Plot Min Depth Distribution



Variable Importance Plot



How this model do?

- R - Squared: 48.06%
- RMSE: 2.11
- MAE: 1.66

Let’s Compare Models

	OLS	Lasso	Random Forest
R- Squared	30.16%	30.13%	48.06%
RMSE	2.45	2.45	2.11
MAE	1.96	1.96	1.66

Limitations

- Wine Years
- Not taking consumers demand into consideration
  - No sales data
- Region Scarcity
- LOOCV run time (CPU power)
- Bias in scoring
- Multicollinearity between price and points

Conclusion

- Most Important Variable
  - Price (.03 point increase per dollar) , *ex: 200 dollar increase = 6 point increase*
  - Taster Name (non controllable)
  - Province - California
- Best model
  - OLS Model
    - 30.16%
    - 2.45
    - 1.96
- Focus on Pricing of Wine Bottle