CPSC 392 Final Project

Spanish Red Wine Data

Anne Marie Santich and Lauren Szlosek

Our Data Set

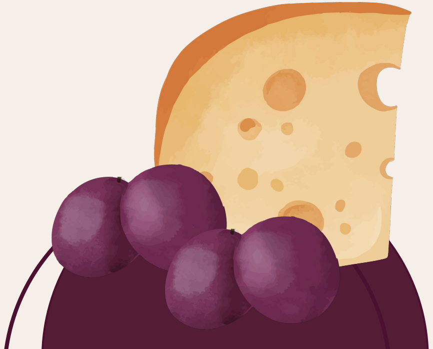
- Spanish Wine Quality Data
- 7500 rows and 11 columns
- Variables:
  - winery: Winery name
  - wine: Name of the wine
  - year: Year in which the grapes were harvested
  - rating: Average rating given to the wine by the users [from 1-5]
  - num_reviews: Number of users that reviewed the wine
  - country: Country of origin [Spain]
  - region: Region of the wine
  - price: Price in euros [€]
  - type: Wine variety
  - body: Body score, defined as the richness and weight of the wine in your mouth [from 1-5]
  - acidity: Acidity score, defined as wine's "pucker" or tartness; it's what makes a wine refreshing and your tongue salivate and want another sip [from 1-5]

fedesoriano. (April 2022). Spanish Wine Quality Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/datasets/fedesoriano/spanish-wine-quality-dataset

Data Cleaning

Missing values and "N.V." were dropped

Dummy variables created for the top 8 wine types



01

Which predictor (year, rating, num_reviews, type, body, acidity) has the strongest coefficient when predicting the price of wine?

supervised model

Linear Regression Set Up

01 Predictors

year, rating, num_reviews, body, acidity, Albarino, Mencia, Priorat Red, Red, Ribera Del Duero Red, Rioja Red, Tempranillo, Toro Red, type_other

02 Model Validation

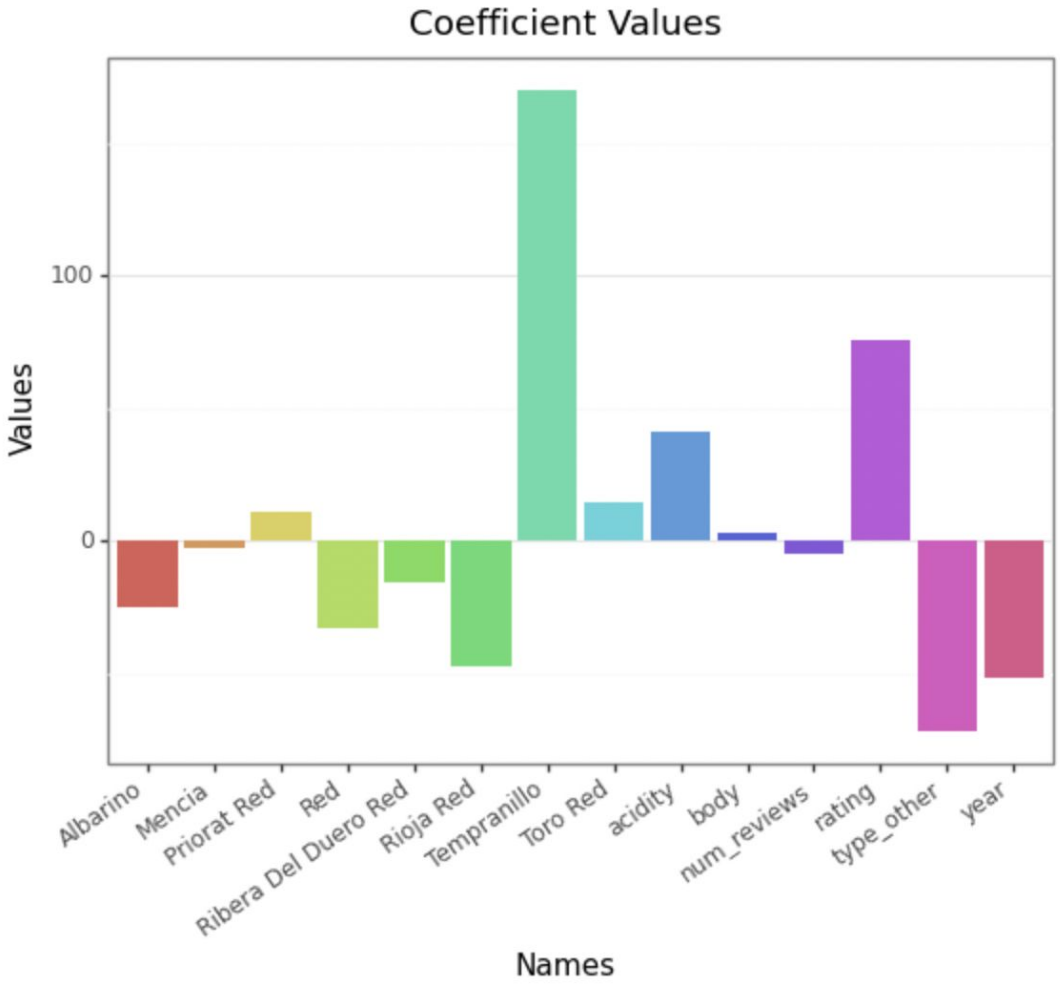
Train/Test split of 80/20

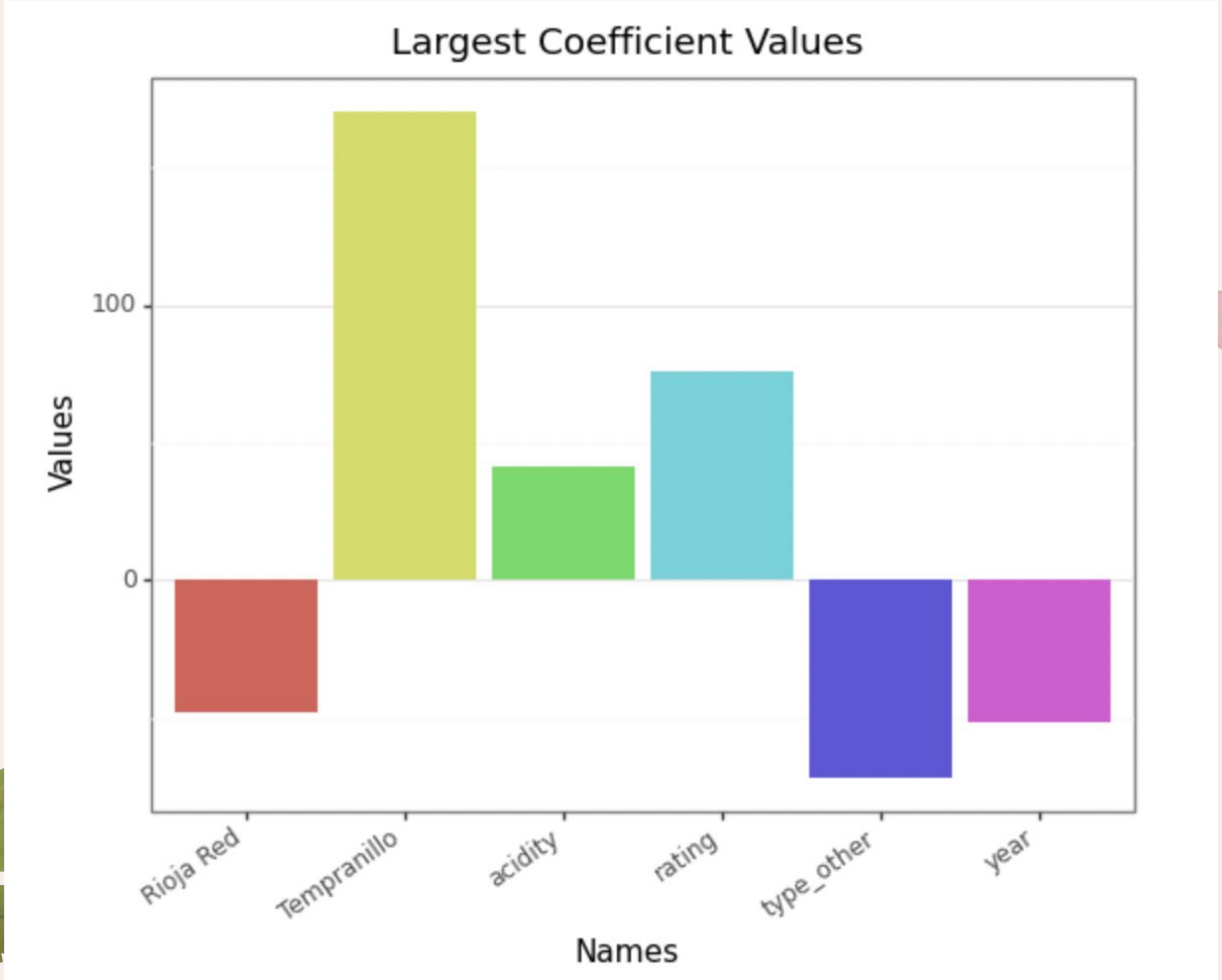
03 Z-Score

Continuous and interval variables were z-scored

04 Create + Fit

Linear Regression model was fit on the training set


Coefficient Values
Values
100
0
Albarino
Mencia
Priorat Red
Red
Ribera Del Duero Red
Rioja Red
Tempranillo
Toro Red
acidity
body
num_reviews
rating
type_other
year
Names

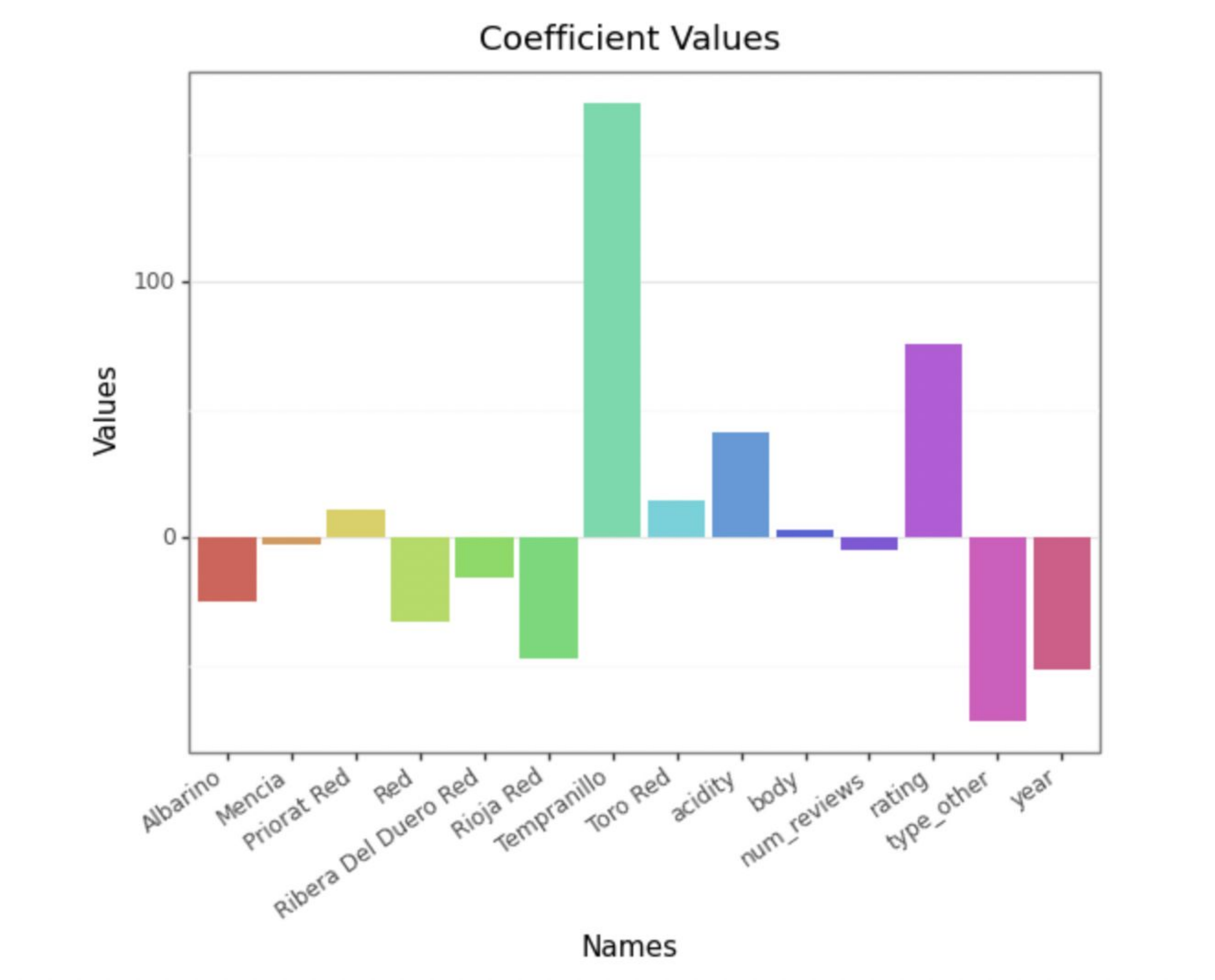

Largest Coefficient Values
Values
100
0
Rioja Red
Tempranillo
acidity
rating
type_other
year
Names

02

After finding the predictor with the strongest coefficient, how would the accuracy (r2) of the predictions change, if removed?

supervised model

Anne Marie's analysis



Linear Regression Set Up pt. 2

01 Accuracy original LR

Calculated the R2 based on the linear regression created by Anne Marie

02 New Predictors

year, rating, num_reviews, body, acidity, Albarino, Mencia, Priorat Red, Red, Ribera Del Duero Red, Rioja Red, ~~Tempranillo~~, Toro Red, type_other

03 Model Validation + Z-score

Train/Test split of 80/20

Continuous and interval variables were z-scored

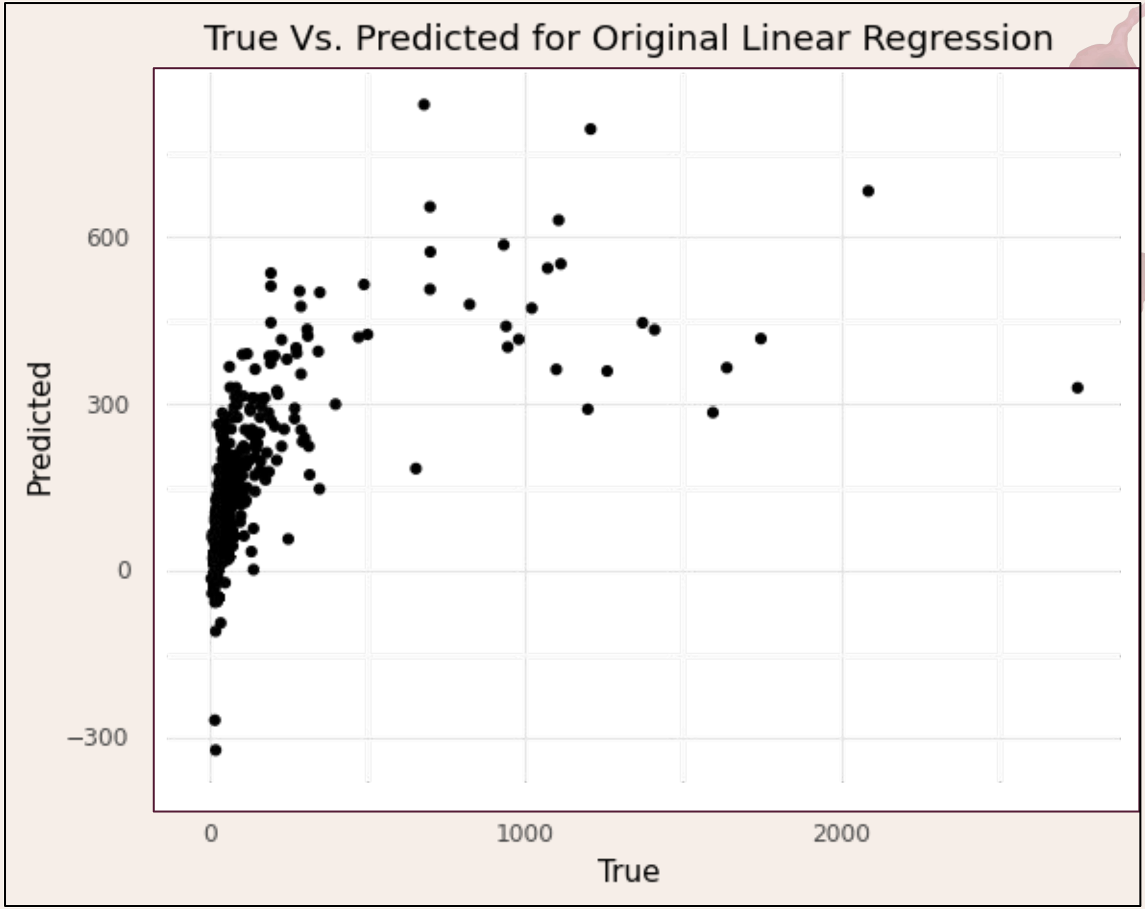
04 Create + Fit

Linear Regression model was fit on the new training set

Original R2:

train: 0.39322857330960703

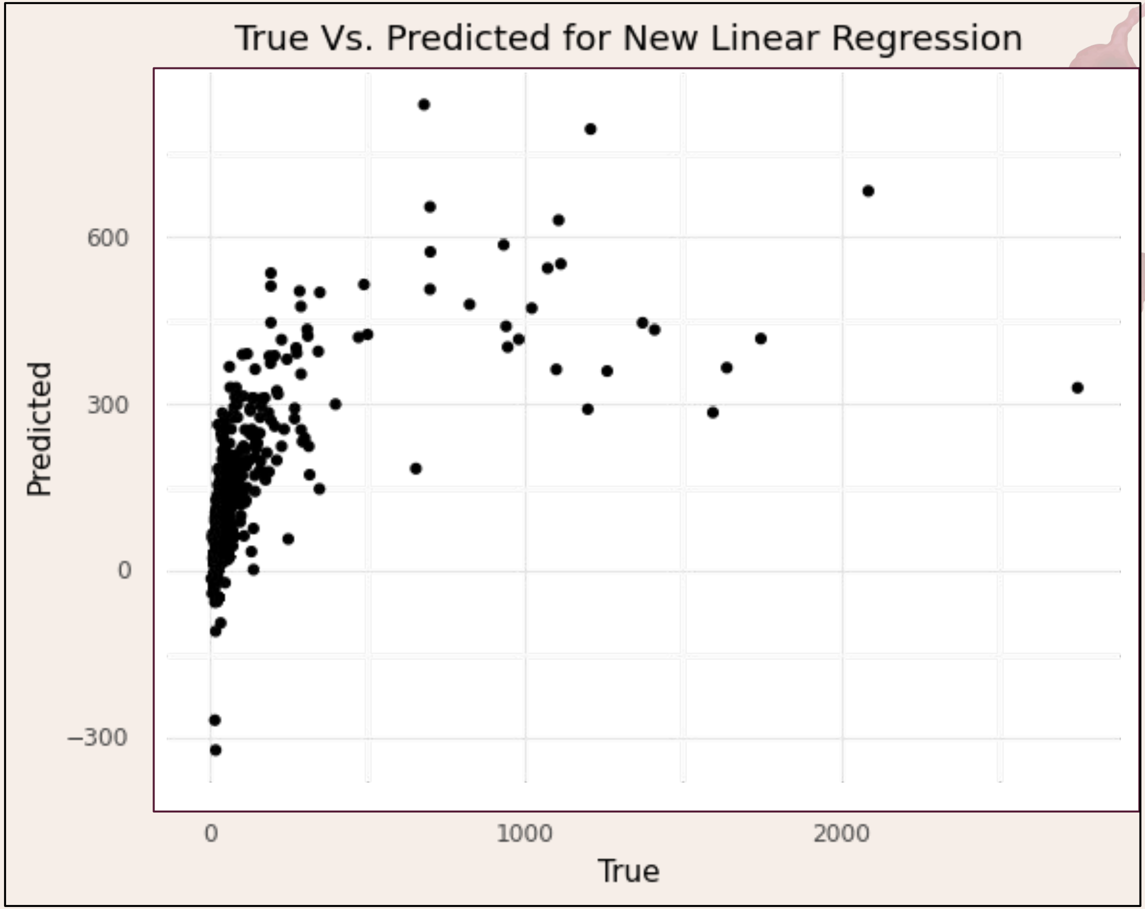
test: 0.4096839221169387



New R2:

train: 0.39322857330960703

test: 0.4096839221169385



03

How much does the MAE change when using Lasso regularization on the linear regression model?

dimensionality reduction

Lasso Set Up

01

Predictors + TTS

The same predictors and Train/Test split was used

02

Pipeline

Z-score object and empty Lasso model

03

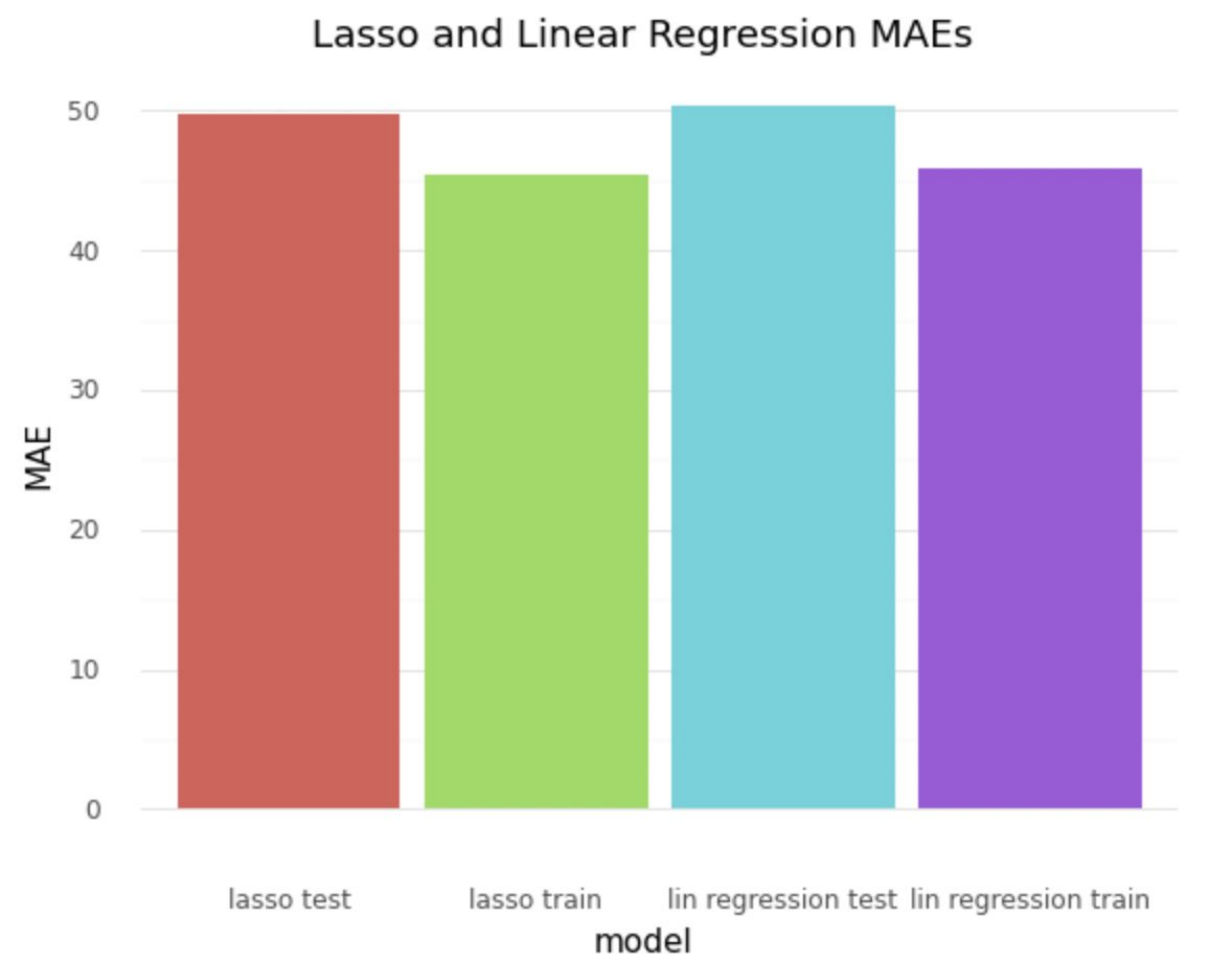
Grid Search

Finding the best value of lambda (penalty)

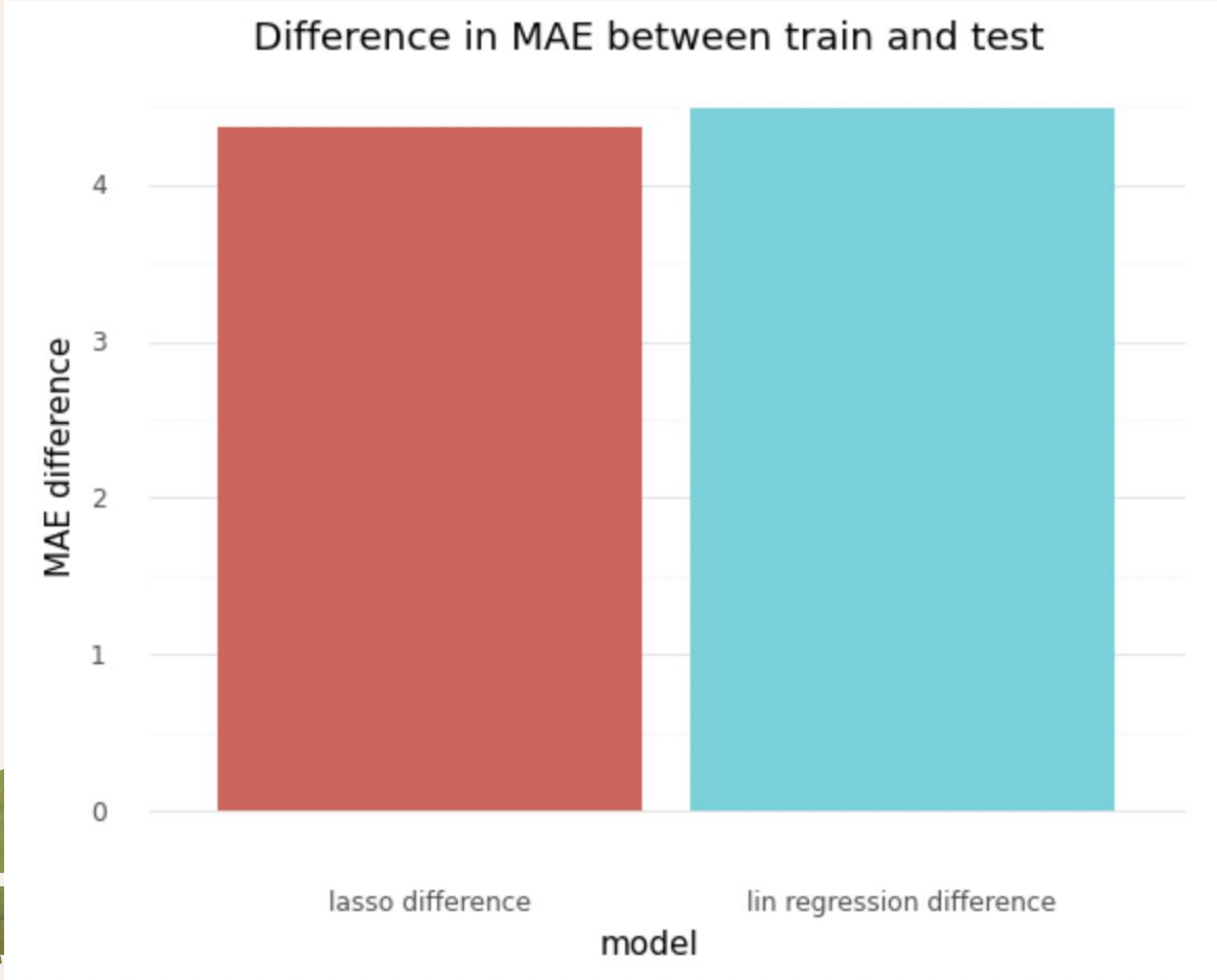
04

Fit

Lasso model fit on the training data



Lasso Train MAE: 45.43
Lasso Test MAE: 49.80
LR Train MAE: 45.82
LR Test MAE : 50.30


Difference in MAE between train and test
MAE difference
0
1
2
3
4
lasso difference
lin regression difference
model

04

Based on the features price, rating, and number of reviews, what clusters form? What types of wines can be inferred from these clusters?

clustering

Gaussian Mixture Model Set Up

01

Scatter plots

Calculated scatter plots for each pairing of the three features

02

Number of components

Created a line graph, displaying the silhouette scores of different numbers of components, decided on 6

03

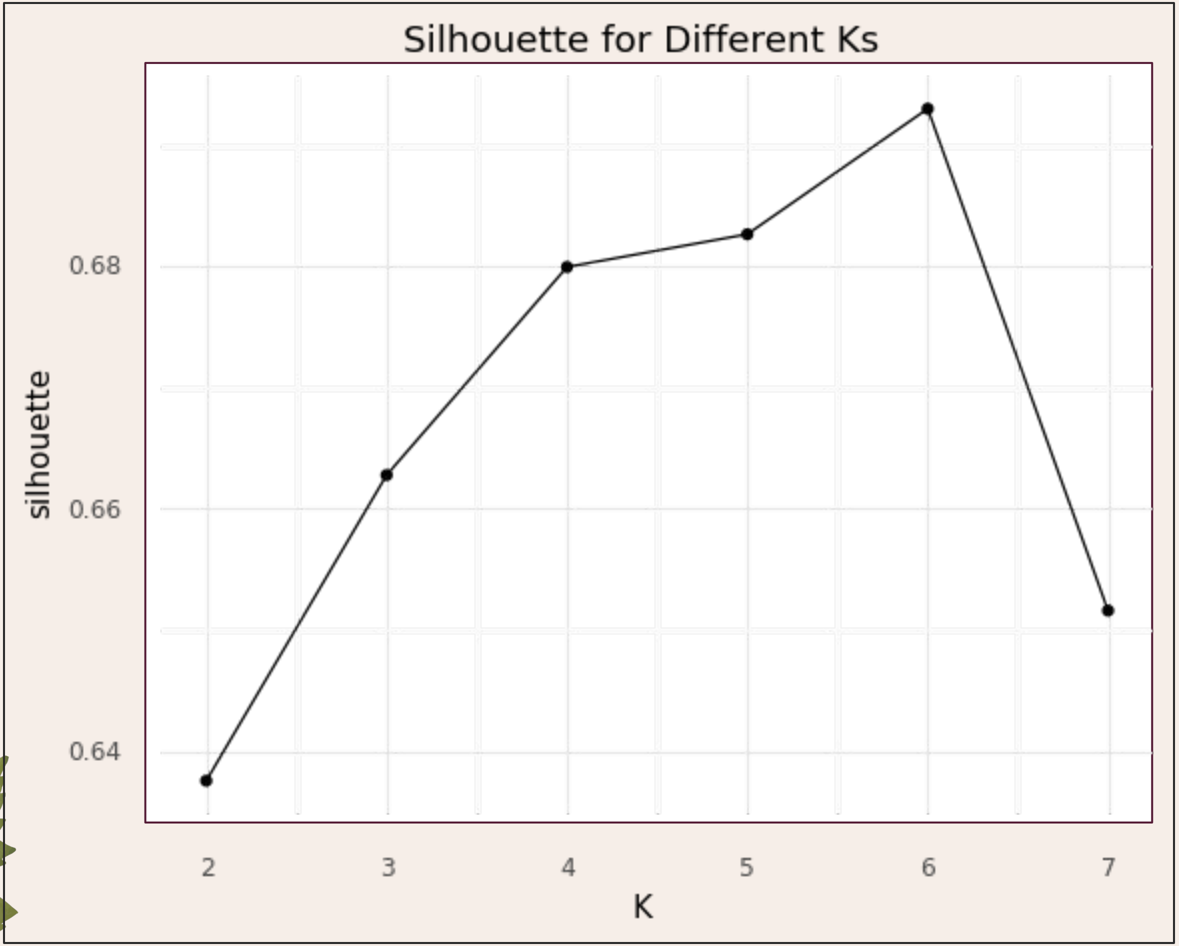
Z-score

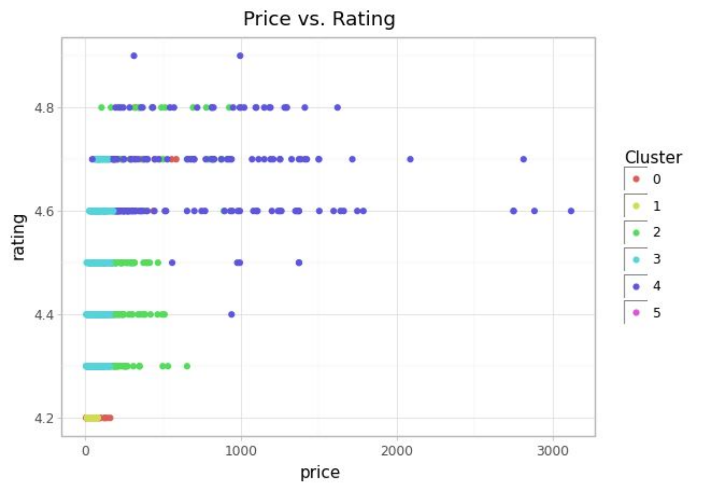
Continuous and interval variables were z-scored

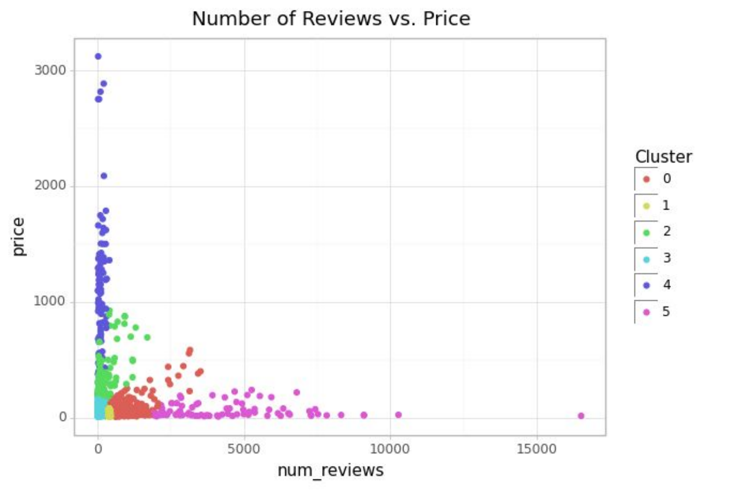
04

Create + Fit

Gaussian Mixture Model model was fit on each pair of features


Silhouette for Different Ks
silhouette
0.68
0.66
0.64
2
3
4
5
6
7
K





0 - low priced, medium number of reviews, variety of ratings

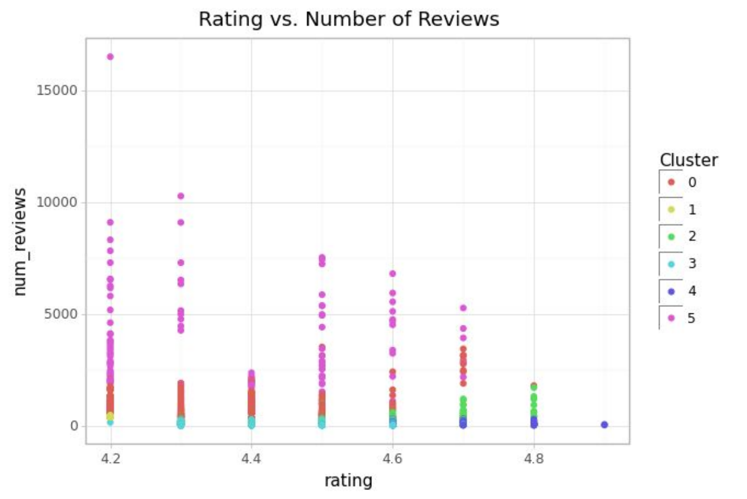
1 - lowest priced, not many reviews, lowest rated

2 - lower price, lower amount of reviews, higher rated

3 - lower priced, lowest amount of reviews, variety of ratings

4 - variety of prices, not many reviews, high rated

5 - lower priced, most reviews, variety of ratings



Silhouette score:

0.687758089