JSC370 Final Project

Cathy Pei

2024-04-26

Introduction

In this project, I aim to explore several questions: How does a wine's vintage affect its price? How are a wine's rated points correlated with its price? Additionally, how are the highest-rated and most expensive wines described by reviewers? I will also investigate the potential for predicting a wine's price using descriptive words and other characteristics. The dataset I plan to use for this inquiry comes from Kaggle and consists of wine reviews collected from WineEnthusiast during the week of June 15th, 2017. Key variables in this dataset that I will examine include country, description, points, price, variety, and vintage.

More specifically, I will extract the vintage information from the title variable and apply statistical models, such as linear models, to discern the relationships between vintage and price, and between points and price, with points representing the taste rating of the wine. Subsequently, I will construct a new categorical variable named "rating," categorizing wines into "high rating," "median rating," and "low rating" groups. In a similar vein, I will categorize wines as "expensive," "medium," or "cheap" based on their price. Following this, I will employ natural language processing (NLP) techniques to identify keywords in the descriptions of highly rated wines, as well as those that are pricey, to identify their features and flavors. Finally, I will tackle the challenge of price prediction by employing various machine learning models, such as decision trees, random forests, and gradient boosting, incorporating different feature variables.

Methods

In this section, I will clean and wrangle the dataset based on its condition, and create basic visualizations to explore the dataset. ## Data Cleaning and Wrangling

I downloaded the data in a CSV file from Kaggle and imported it as a dataframe for further data cleaning processes. I extracted the vintage year from the title variable and created a new variable for it. Then, I removed redundant variables and retained only the ones necessary for my analysis. Lastly, I created the new variables "rating" and "price_cat" based on the quantiles of the points variable and price variables.

By using the `head`, `tail`, and `summary` functions, I conducted a basic exploration of the data. I observed that there are three integer-type variables: points, price, and vintage, as well as six character-type variables: country, description, variety, title, rating, and price_cat. Furthermore, there are 8,996 missing values in price and 4,609 in vintage.

While examining outliers and potential issues in the data, I noticed that the vintage variable, extracted from the title variable, is suspicious, given its minimum of 1000 and maximum of 7200. Upon investigating specific observations, I found that this discrepancy is caused by the inclusion of the wine's name, where some wines have a vintage year in their name, but it does not necessarily represent the actual year of production. This implies the need to extract the vintage year specifically indicating when the wine was made. To address this issue, I modified the method of extracting the vintage: now, I extract the vintage from the 20s if present; otherwise, I extract the vintage as is. After this modification, the summary for the vintage variable appeared more reasonable, with a minimum of 1503 and a maximum of 2017.

Furthermore, I noticed that the maximum price is 3300, whereas its median is only 25. This implies that 3300 could be a potential outlier, and we should evaluate whether it should be removed in later analysis.

Data Exploration

There are 129,971 observations and 9 columns in the cleaned dataset. The summary statistics for the integer-type variables are shown in the table below.

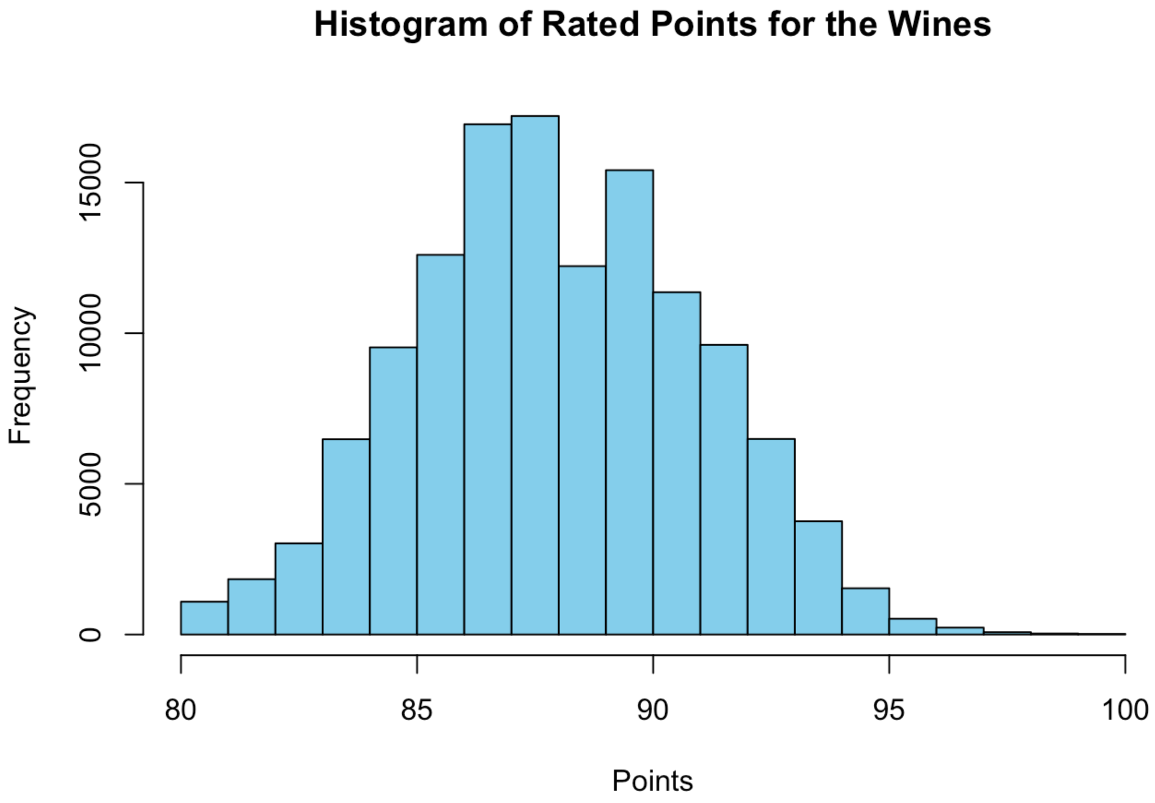
country	description	points	price	vintage	variety	title	rating	price_cat
Length:129971	Length:129971	Min. : 80.00	Min. : 4.00	Min. :1503	Length:129971	Length:129971	low rating :51493	cheap :46341
Class :character	Class :character	1st Qu.: 86.00	1st Qu.: 17.00	1st Qu.:2009	Class :character	Class :character	median rating:44843	medium :34891
Mode :character	Mode :character	Median : 88.00	Median : 25.00	Median :2011	Mode :character	Mode :character	high rating :33635	expensive:39743
NA	NA	Mean : 88.45	Mean : 35.36	Mean :2011	NA	NA	NA	NA's : 8996
NA	NA	3rd Qu.: 91.00	3rd Qu.: 42.00	3rd Qu.:2013	NA	NA	NA	NA
NA	NA	Max. :100.00	Max. :3300.00	Max. :2017	NA	NA	NA	NA
NA	NA	NA	NA's :8996	NA's :4609	NA	NA	NA	NA

The numbers of unique values for appropriate character-type variables are shown in the table below. Notice that there are 44 unique countries, 426 unique provinces, and 708 unique types of grapes.

Variable	Unique_Values
Country	44
Province	0
Variety	708

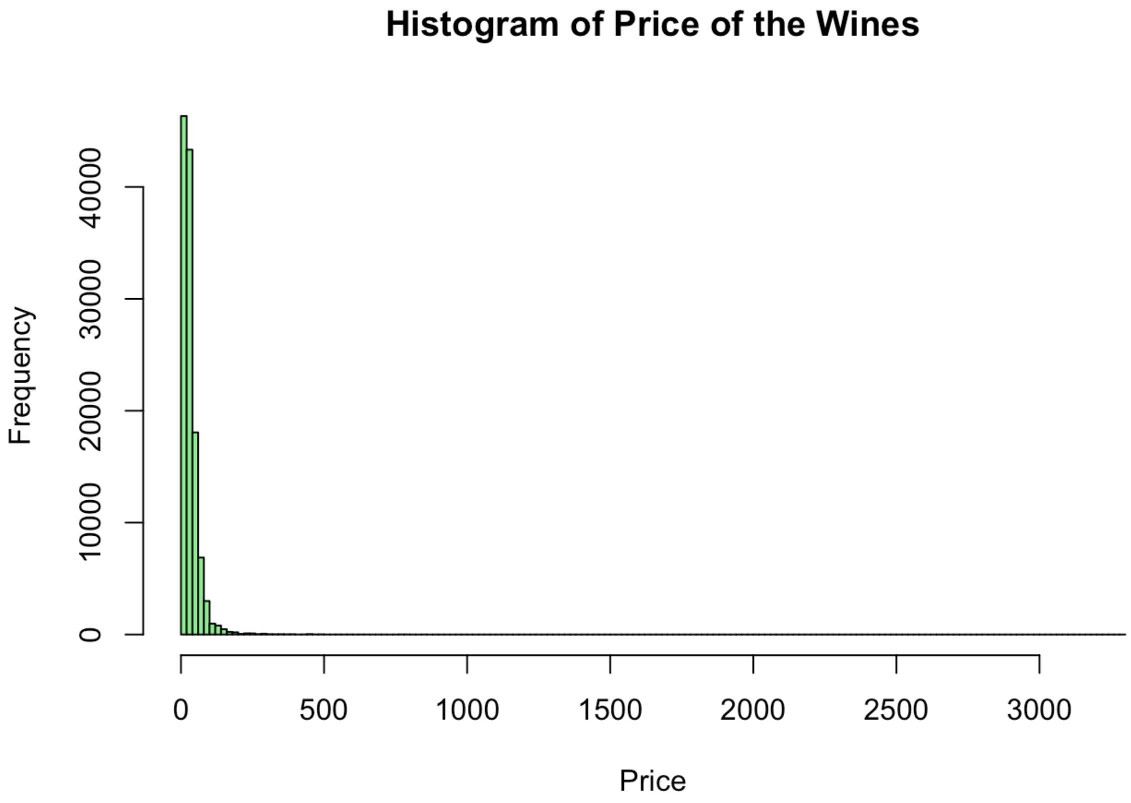
Data Visualization

Histograms for integer-type variables



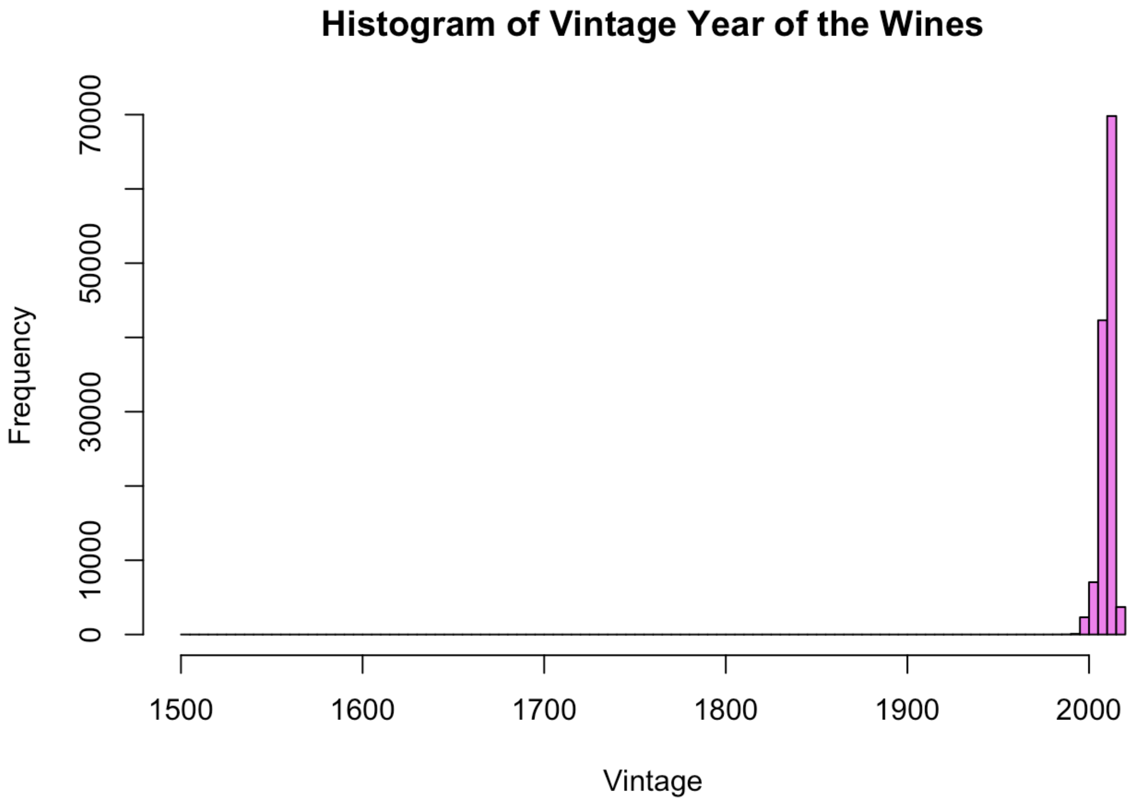
Observation

The histogram shows that the rated points for the wines are normally distributed, ranging from 80 to 100.



Observation

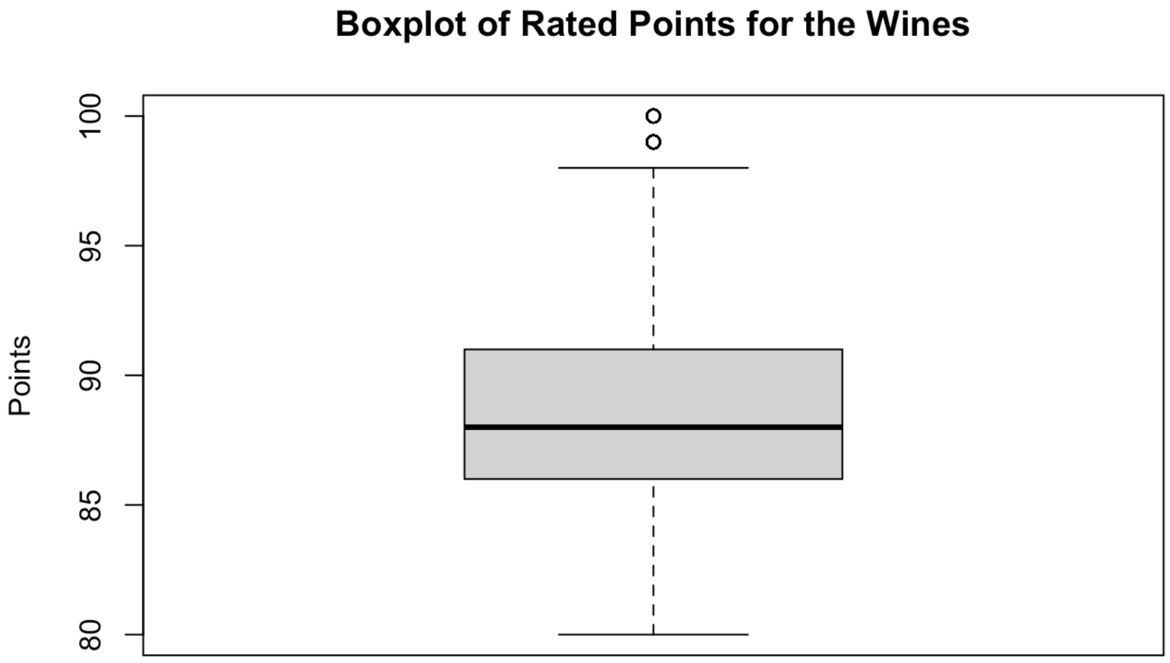
The histogram shows that the distribution of wine prices is extremely right-skewed. We can observe that most of the wines are priced within 500. As mentioned in the previous section, the long tail to the right in the histogram might be evidence of potential outliers. More details about the outliers can be shown in the boxplot.



Observation

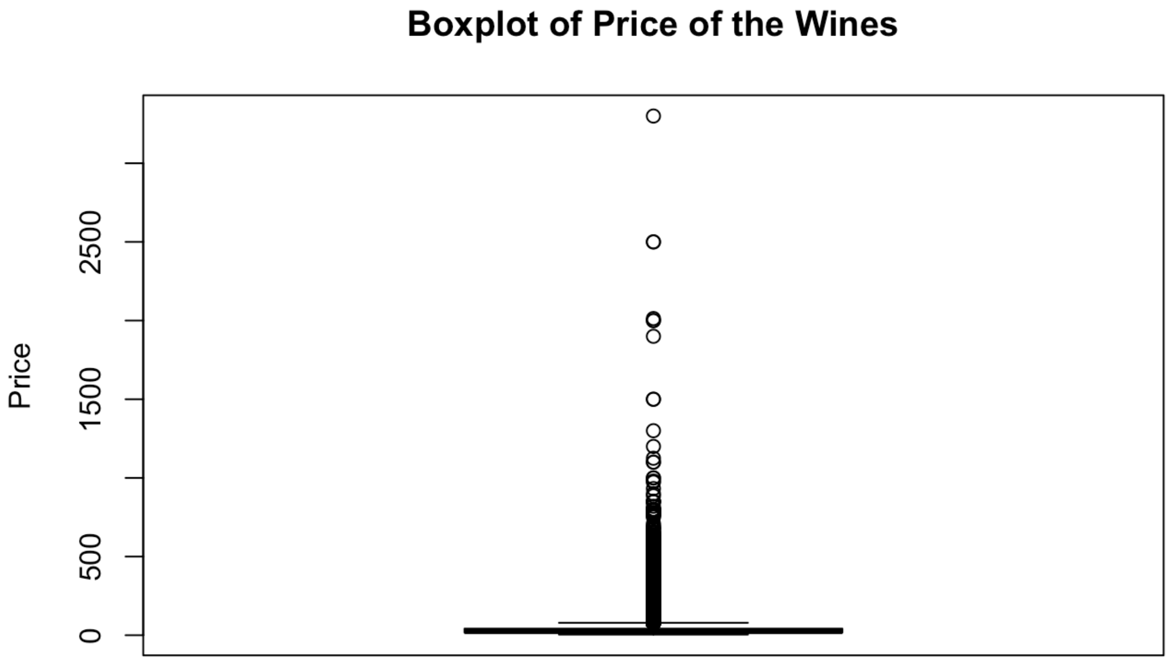
The histogram shows that the distribution of the vintage of the wine is extremely left-skewed. We can observe that most of the wines have a vintage year in the 2000s. The long tail to the left in the histogram might be evidence of potential outliers, which should be further evaluated in the boxplot.

Boxplots for integer-type variables



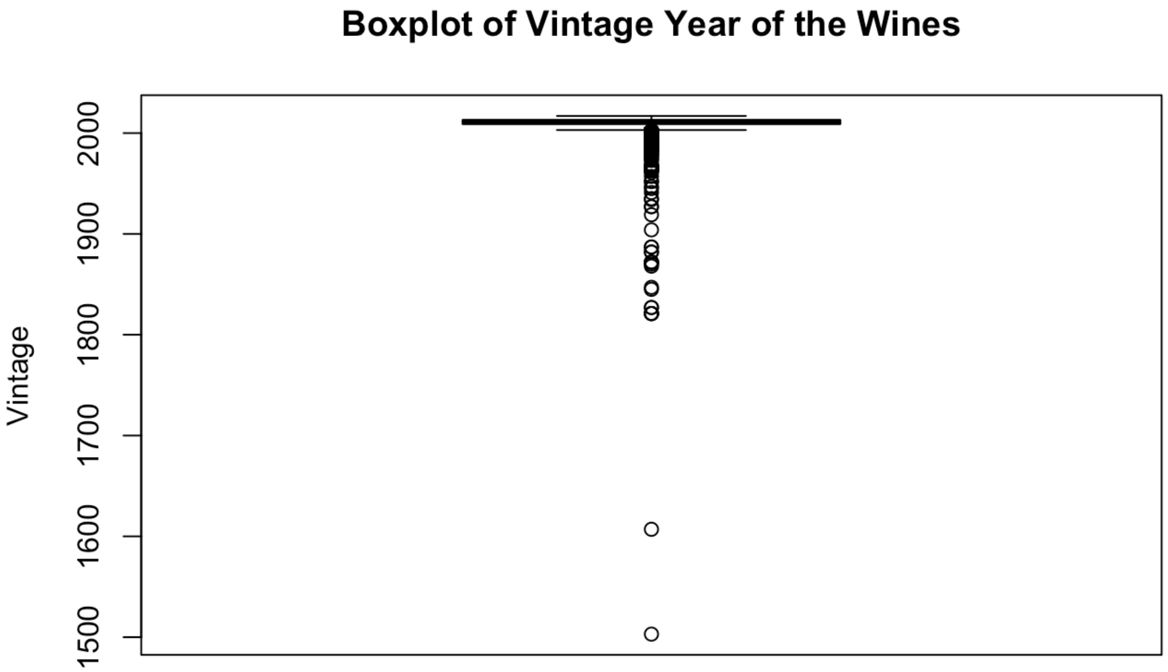
Observation

The boxplot for points looks fair; it corresponds to the conclusion we drew while investigating its histogram.



Observation

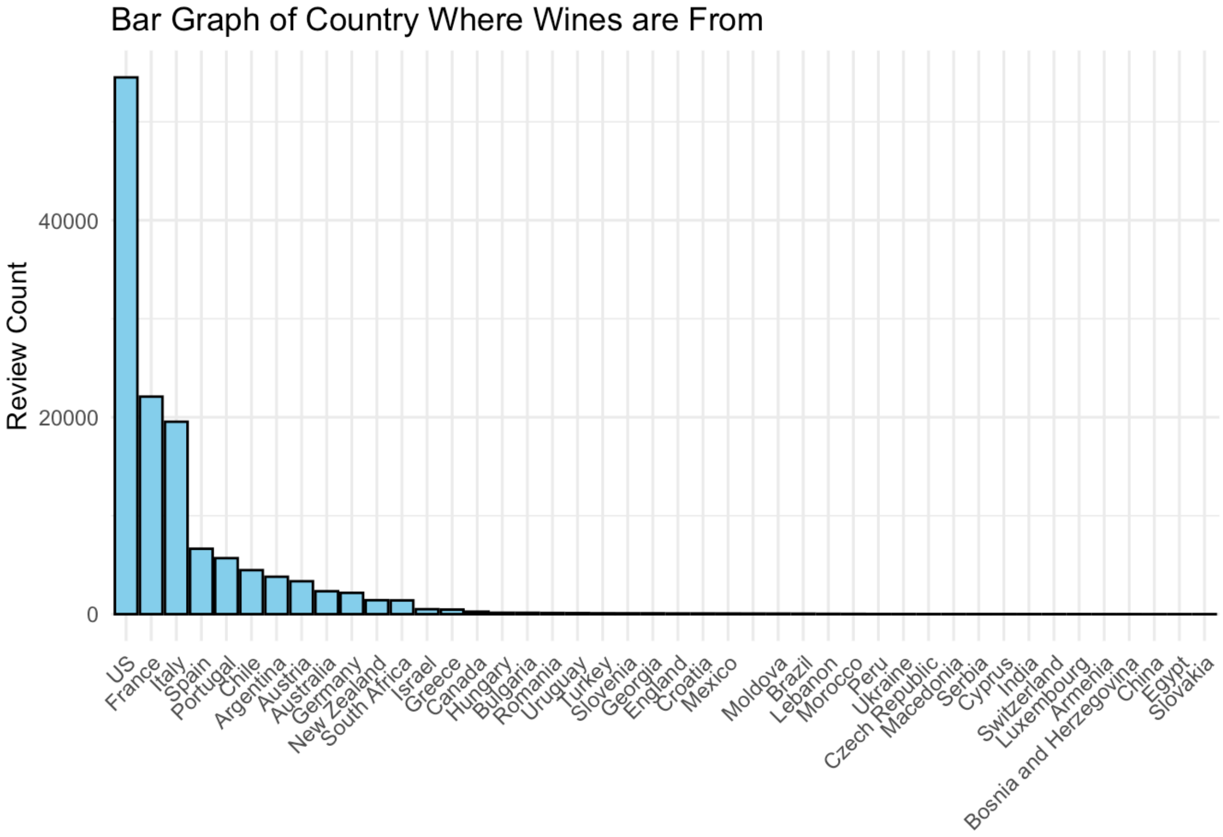
The boxplot for price suggests that there are many points considered as outliers, approximately those with prices greater than 200. Since the number of outliers is significantly large, we should carefully consider whether they should be removed to avoid biasing our statistical model or analysis in further steps.



Observation

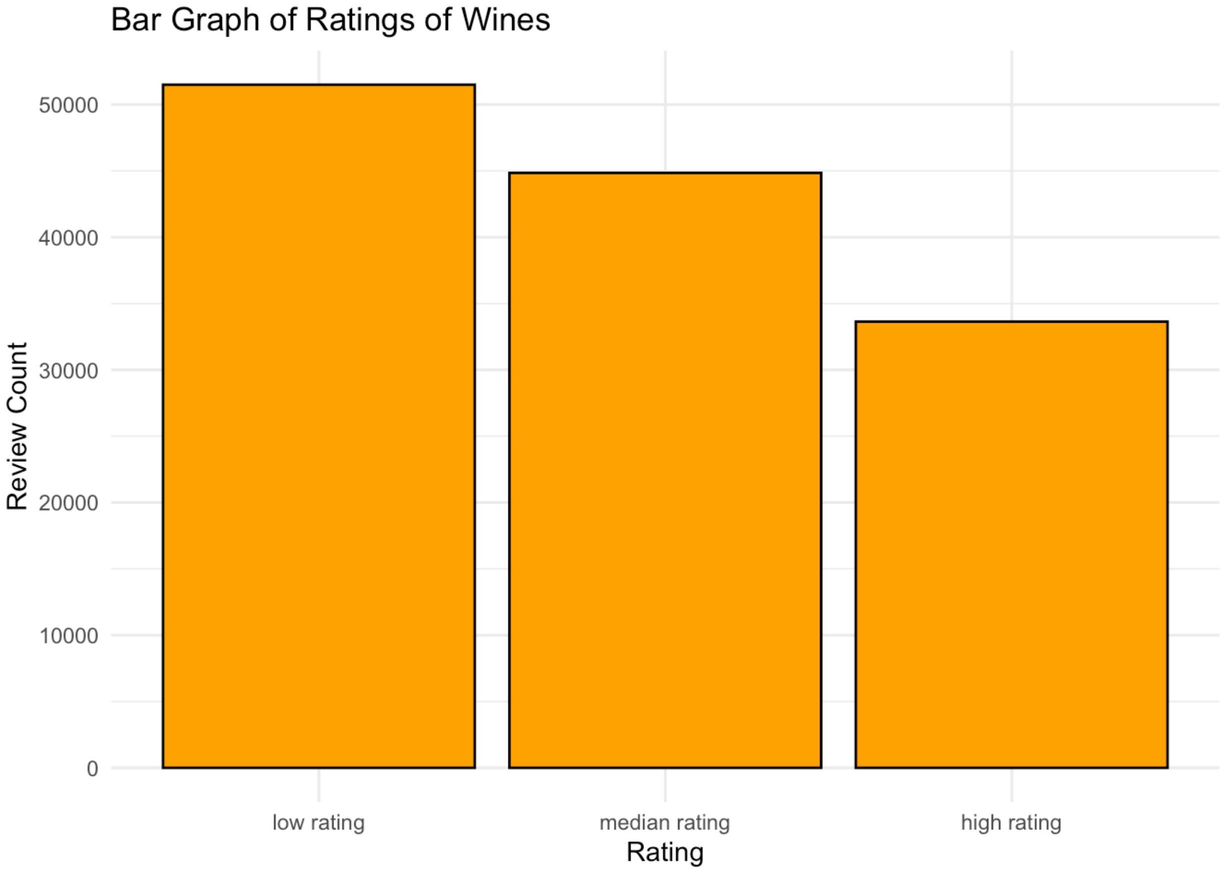
The boxplot for vintage suggests that there are many points considered as outliers, approximately those with a vintage year earlier than 2000. Similar to what we observed in price, there is a significant number of outliers to be considered during the analysis. Notice that there are two data points that are significantly far from the box, one with a vintage year in the 1600s and one in the 1500s, which should definitely be removed when conducting the analysis.

Bar plot for appropriate categorical variables



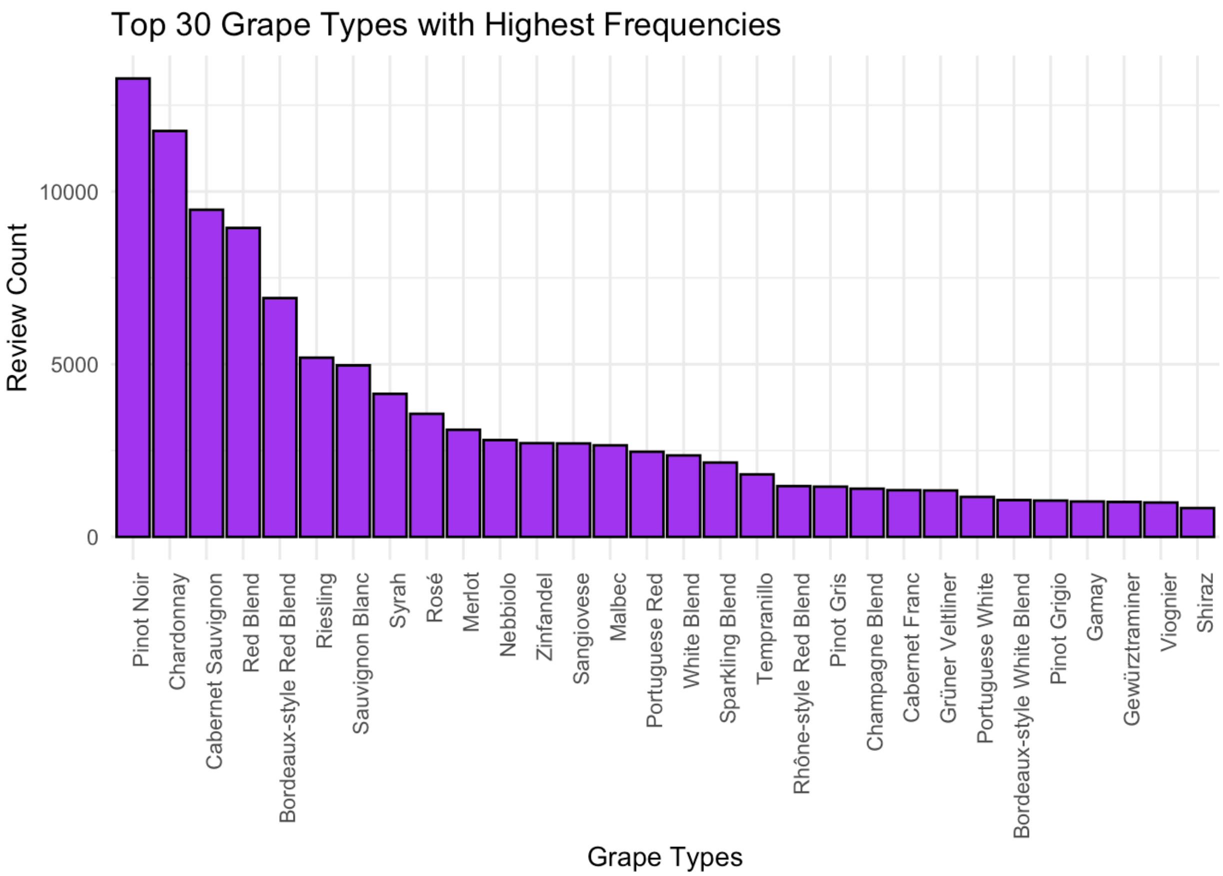
Observation

According to the bar plot, we can see that most of the reviews are based on wines made in the US. Some other countries whose wines have been reviewed extensively include France and Italy.



Observation

According to the bar graph of the ratings, we can observe that approximately 50,000 reviews give a low rating to a wine, approximately 45,000 reviews give a median rating to a wine, and approximately 35,000 reviews give a high rating to a wine.



Observation

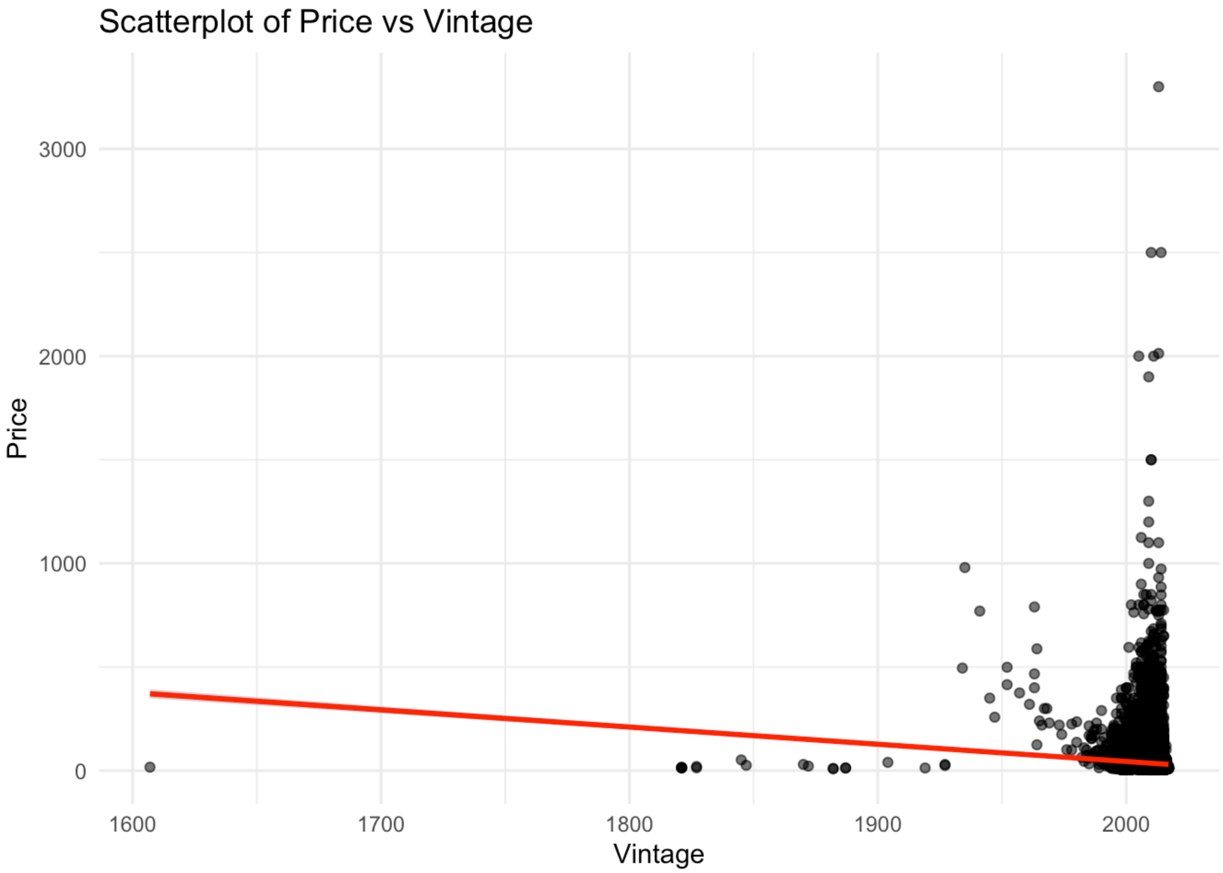
According to the bar plot of variety, we can observe that most reviews are for wines made from Pinot Noir, Chardonnay, and Cabernet Sauvignon. Note that to avoid an overcrowded graph, I have included only the top 30 grape types with the highest frequency in the data.

Modeling and Results

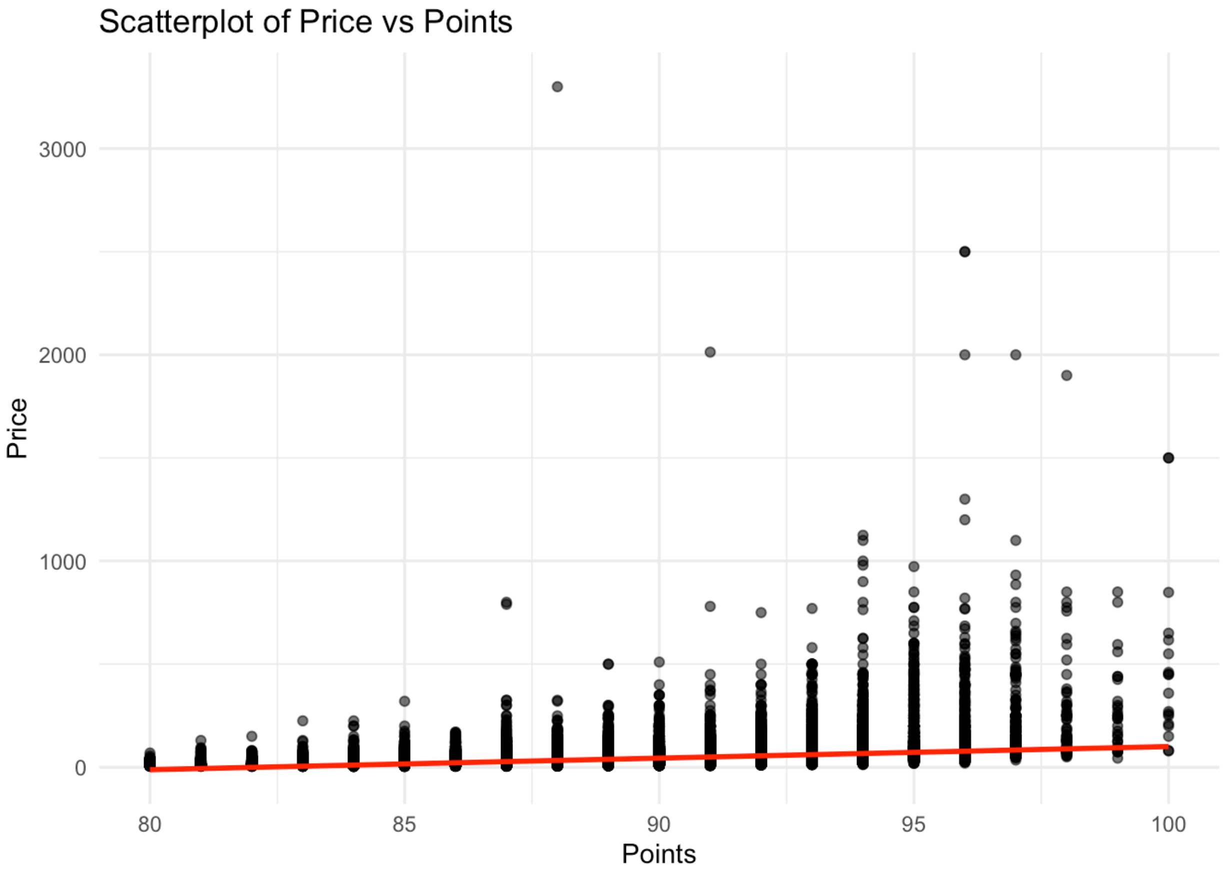
Linear model and spline model

To find the relationship between vintage and price, and points and price, I will create linear models and spline models to address this question.

```
##
## Call:
## lm(formula = price ~ vintage, data = wine_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -353.8  -17.9   -9.1    6.7 3266.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1705.28931   56.61573   30.12   <2e-16 ***
## vintage       -0.83045    0.02816  -29.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.05 on 116836 degrees of freedom
## Multiple R-squared:  0.00739,    Adjusted R-squared:  0.007382
## F-statistic: 869.9 on 1 and 116836 DF,  p-value: < 2.2e-16
```



```
##
## Call:
## lm(formula = price ~ points, data = wine_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -57.8  -14.9   -5.3    7.1 3267.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -461.17143    3.18821  -144.6   <2e-16 ***
## points         5.61474    0.03602   155.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.49 on 116836 degrees of freedom
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1721
## F-statistic: 2.43e+04 on 1 and 116836 DF,  p-value: < 2.2e-16
```

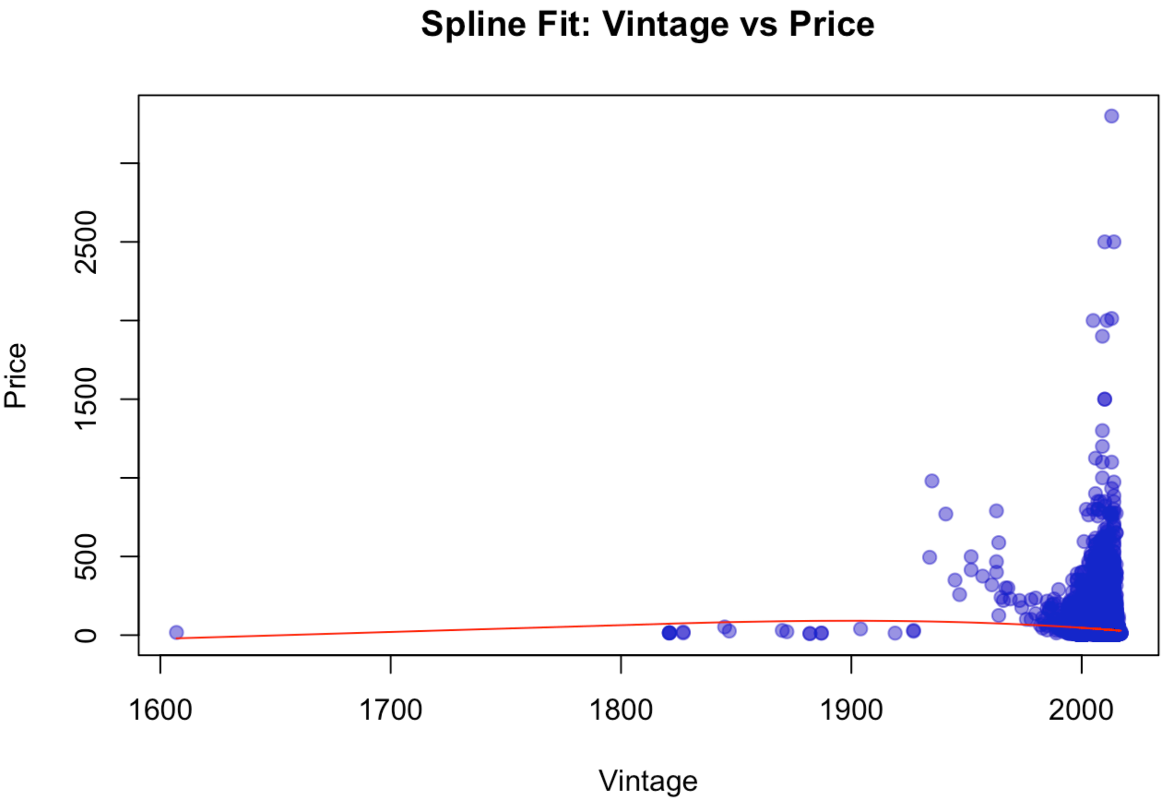


Summary on linear models

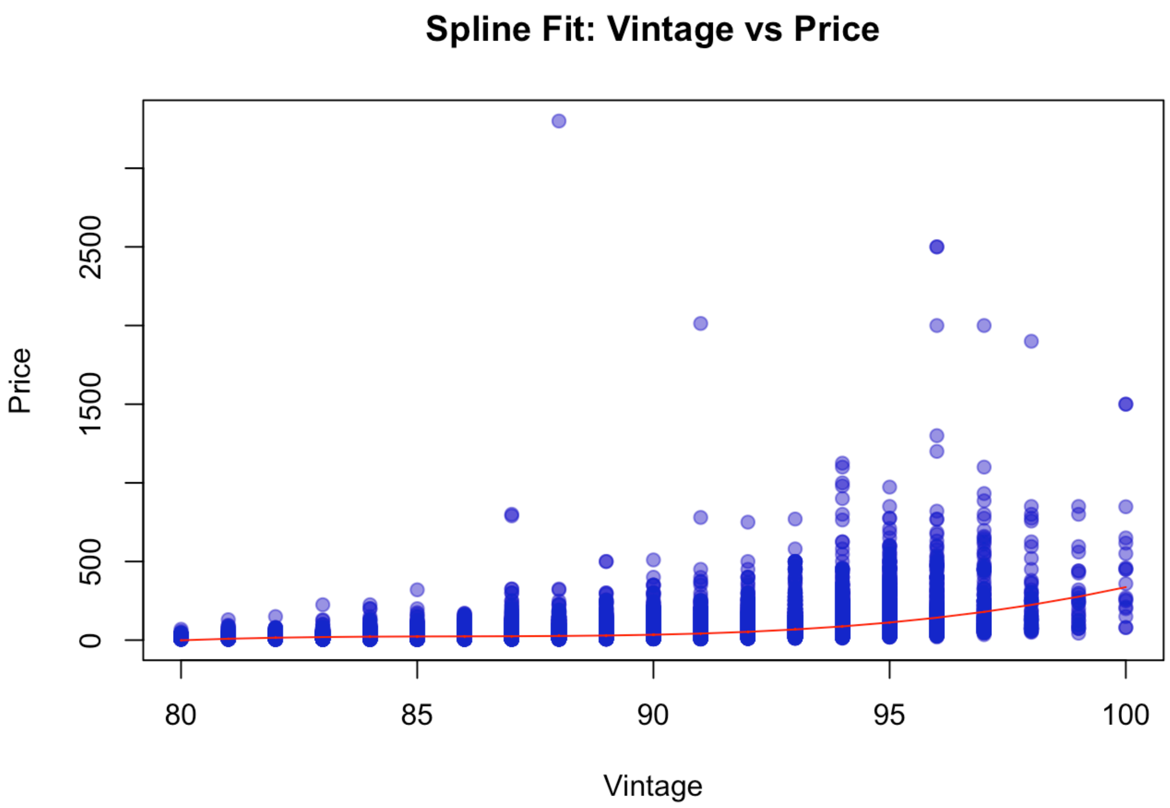
The first linear regression output indicates that vintage is a statistically significant predictor of price, with each additional year decreasing the predicted price by approximately $0.83. However, the model’s low R-squared value of 0.00739 suggests that vintage alone accounts for less than 1% of the variability in wine prices, highlighting that vintage has a minimal impact and other factors likely play a more substantial role in determining price. The large range of residuals also points to a significant amount of unexplained variability, suggesting the presence of outliers or that the relationship between vintage and price may not be linear.

The second linear regression analysis suggests a strong and positive relationship between points and price, with the points a wine receives being a significant predictor of its price. For every additional point, the price is expected to increase by about $5.61. With an R-squared of 0.1722, the model explains approximately 17.22% of the variability in wine prices, which is a substantial improvement over the model with vintage. The F-statistic is highly significant, reinforcing the significance of the model. However, the residuals indicate there is still a considerable amount of unexplained variability, and the potential influence of other factors not included in the model.

```
##
## Call:
## lm(formula = price ~ bs(vintage), data = wine_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -80.5  -17.7   -8.9    7.0 3267.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -20.09      40.88  -0.491    0.623
## bs(vintage)1    11.23      63.93   0.176    0.861
## bs(vintage)2   210.51      40.77   5.164 2.42e-07 ***
## bs(vintage)3    48.48      40.89   1.186    0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.02 on 116834 degrees of freedom
## Multiple R-squared:  0.009242,   Adjusted R-squared:  0.009217
## F-statistic: 363.3 on 3 and 116834 DF,  p-value: < 2.2e-16
```



```
##
## Call:
## lm(formula = price ~ bs(points), data = wine_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -256.8   -12.0    -5.1     5.9  3273.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.0887     0.9563  -1.138    0.255
## bs(points)1   77.2362     2.5026  30.862   <2e-16 ***
## bs(points)2  -95.2883     1.7559 -54.268   <2e-16 ***
## bs(points)3  337.9072     3.0304 111.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.48 on 116834 degrees of freedom
## Multiple R-squared:  0.2587, Adjusted R-squared:  0.2587
## F-statistic: 1.359e+04 on 3 and 116834 DF,  p-value: < 2.2e-16
```



Summary on spline models

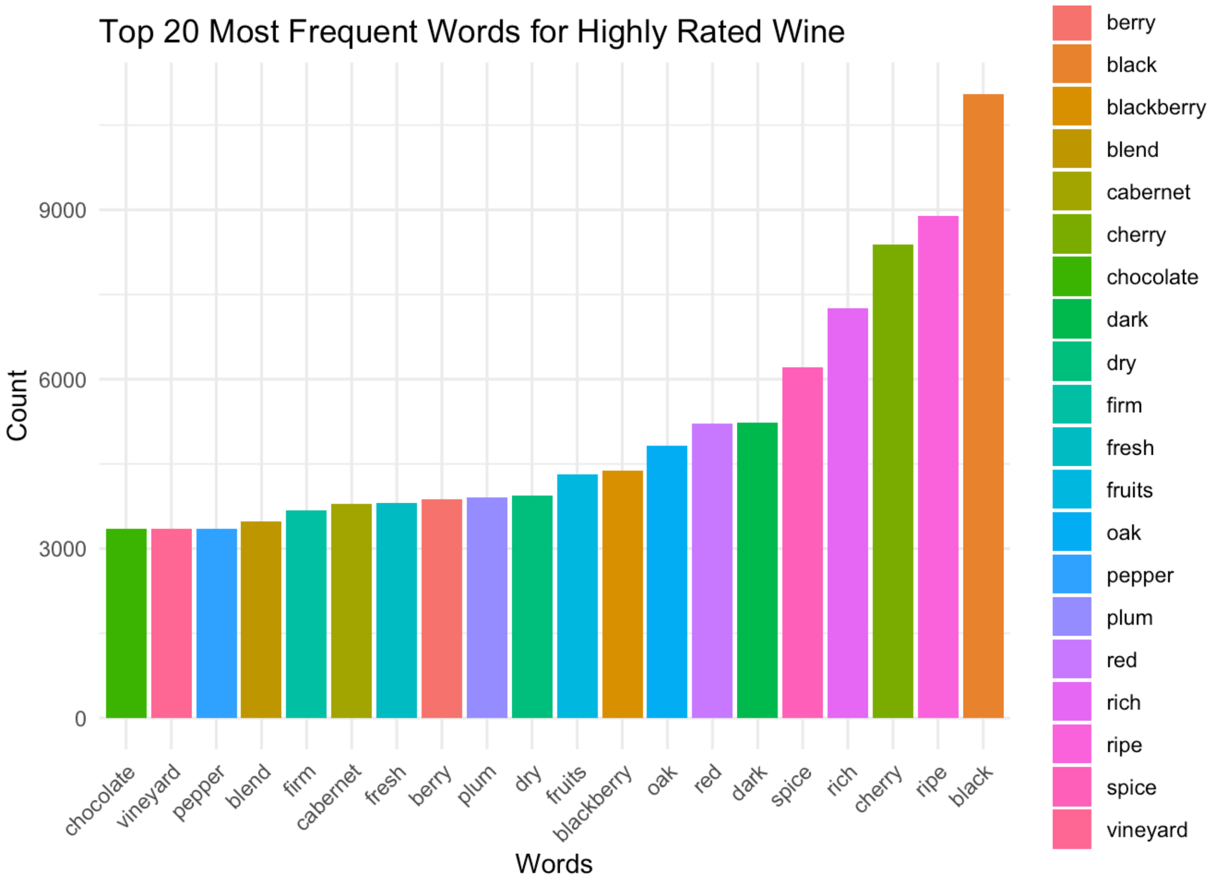
In the first spline model output for predicting wine price from vintage, using basis splines, only the second spline coefficient bs(vintage)2 is statistically significant, with a p-value well below 0.05 and a t-value of 5.164. This suggests that there is a non-linear relationship between vintage and price, but only specific parts of the vintage variable (captured by bs(vintage)2) significantly contribute to predicting price. The multiple R-squared of 0.009242 indicates that the model explains only about 0.924% of the variability in wine prices, which is very low and suggests that vintage may not be a strong predictor of price by itself. The adjusted R-squared is similarly low, reinforcing the notion that other variables may be needed to better predict price. The F-statistic is significant, indicating that the model is statistically significant overall compared to a model with no predictors. However, the low R-squared values suggest that while the spline components may be capturing some aspect of the relationship, it's likely that vintage alone is not sufficient to model price effectively.

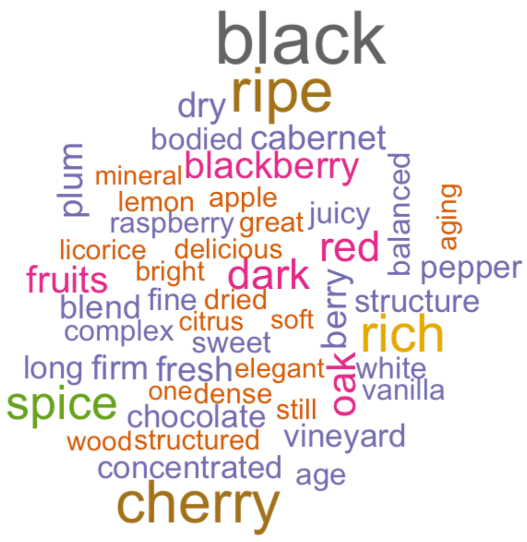
The second spline model output examines the relationship between points and price using basis splines. All spline coefficients for points are statistically significant, with p-values much less than 0.05, indicating a non-linear relationship where the impact of points on price varies at different levels of points. The model has a substantial multiple R-squared of 0.2587, meaning it explains about 25.87% of the variability in wine prices, a considerable improvement compared to the spline model of vintage against price. The adjusted R-squared is also 0.2587, which after adjusting for the number of predictors in the model, still indicates a decent fit. The significant F-statistic reaffirms that the spline model is overall a strong fit compared to a model with no predictors. The large coefficients for the spline terms suggest that as points increase, their effect on price becomes more pronounced, indicating the relationship between wine scores and their prices is complex and perhaps exponential rather than linear.

NLP

In this section, I will examine the most frequently used descriptive words that characterize highly rated and expensive wines. This analysis will involve tokenizing the words in the description variable and removing stopwords.

Highly Rated Wines

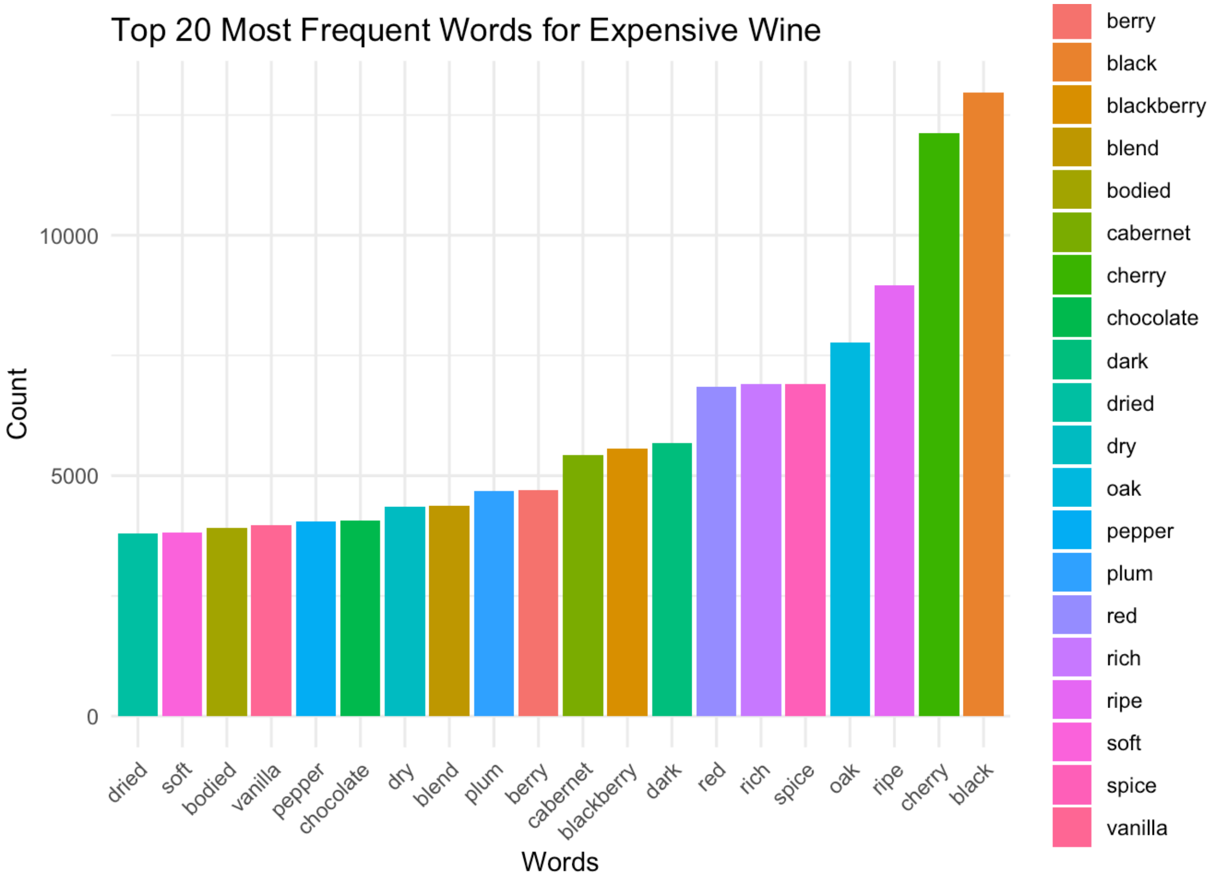


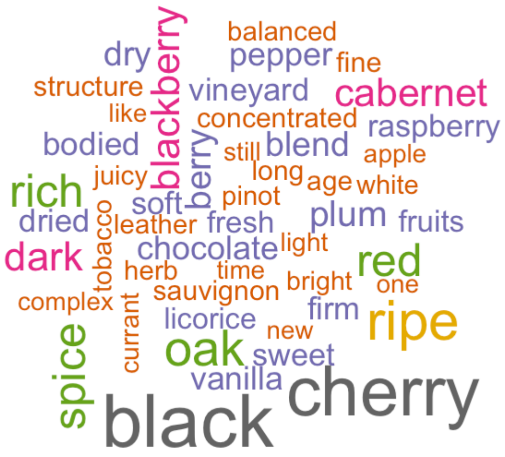


Observation and Interpretation

For highly rated wines, common descriptors include “black,” “ripe,” “cherry,” “rich,” and “spice”. These terms sketch out a sensory profile that wine enthusiasts admire. “Black” often refers to the presence of dark berries, hinting at intense and desirable flavors. “Ripe” indicates grapes harvested at their peak, offering a pronounced sweetness and robust taste. “Cherry” adds a note of both sweetness and acidity, a beloved trait in many reds. “Rich” describes a velvety, full-bodied experience, with layered flavors that enchant the palate. Lastly, “spice” points to subtle, yet intricate flavors that can arise from the grape variety, the wine’s origin, or the aging process, especially in oak which contributes additional aromatic qualities. Collectively, these terms portray wines with a profound and memorable flavor profile.

Expensive Wines


Top 20 Most Frequent Words for Expensive Wine
Count
10000
5000
0
Words
dried
soft
bodied
vanilla
pepper
chocolate
dry
blend
plum
berry
cabernet
blackberry
dark
red
rich
spice
oak
ripe
cherry
black
berry
black
blackberry
blend
bodied
cabernet
cherry
chocolate
dark
dried
dry
oak
pepper
plum
red
rich
ripe
soft
spice
vanilla


balanced
dry
blackberry
pepper
fine
structure
vineyard
cabernet
like
concentrated
raspberry
bodied
berry
still
blend
apple
long
juicy
age
white
rich
soft
pinot
tobacco
dried
leather
fresh
plum
fruits
chocolate
light
dark
red
herb
time
bright
one
sauvignon
complex
currant
firm
licorice
ripe
new
spice
oak
sweet
vanilla
cherry
black

Observation and Interpretation

For expensive wines, descriptors like “black,” “cherry,” “ripe,” “oak,” and “spice” are frequently mentioned. “Black” likely refers to robust flavors of dark fruits, similar to those noted in highly rated wines, suggesting a shared appreciation for intense fruitiness. “Cherry” brings to mind a sweetness with a touch of tartness, echoing the taste profile favored in top-rated wines. “Ripe” again implies grapes at their fullest flavor potential, a common thread that denotes quality both in highly rated and costly wines. The mention of “oak” is more specific to expensive wines, indicating that the aging process in oak barrels, which imparts vanilla and woody notes, is a valued characteristic. “Spice,” present in both categories, points to the complex flavors that give each sip depth and distinction. While both expensive and highly rated wines share several descriptors, the prominence of “oak” in expensive wines suggests that the aging process and its flavor contributions might play a more notable role in the luxury market.

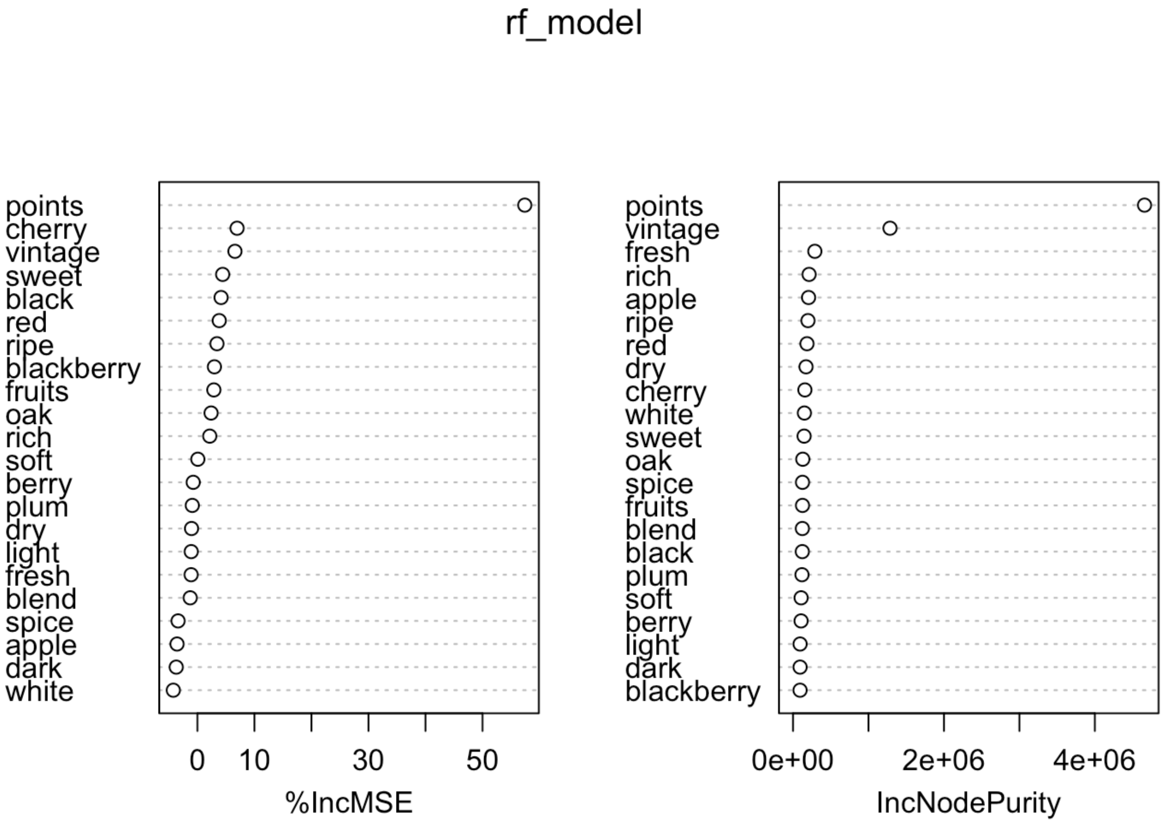
Machine learning model

In this section, I will train three different models to predict wine prices: a random forest, a boosting model, and a gradient boosting model. The independent variables used for prediction will include vintage, points, and descriptive words from the wine reviews. For instance, I’ll transform the descriptive words into numerical data using text vectorization methods. Then, I will create models to predict price using the variables I originally have and the text vector variables. Lastly, I will assess and compare each model’s performance on a testing set using Root Mean Squared Error (RMSE). Based on these evaluations, I will fine-tune the models and their hyperparameters to optimize performance.

Preparing Data

This step processes the wine dataset to prepare it for machine learning analysis. Initially, all rows containing missing values are removed, and the dataset is then halved by randomly selecting 10% of the rows to preserve vector space and avoid using up all the memory in processing large data. The descriptions of the selected wines are then tokenized into individual words, converted to character type, and filtered to remove stopwords and numeric strings, focusing only on meaningful textual content. Subsequently, the top 20 most frequent meaningful words from these descriptions are identified and retained. Next, a Document-Term Matrix (DTM) is created, recording the presence of these top words in each document as binary values (1 if present, 0 if not). This DTM is then effectively merged back into the selected_wine dataset, adding the text analysis results as new features. Finally, the prepared dataset is split into training (70%) and testing (30%) sets, using a set seed for consistent splits across different runs.

Random Forest

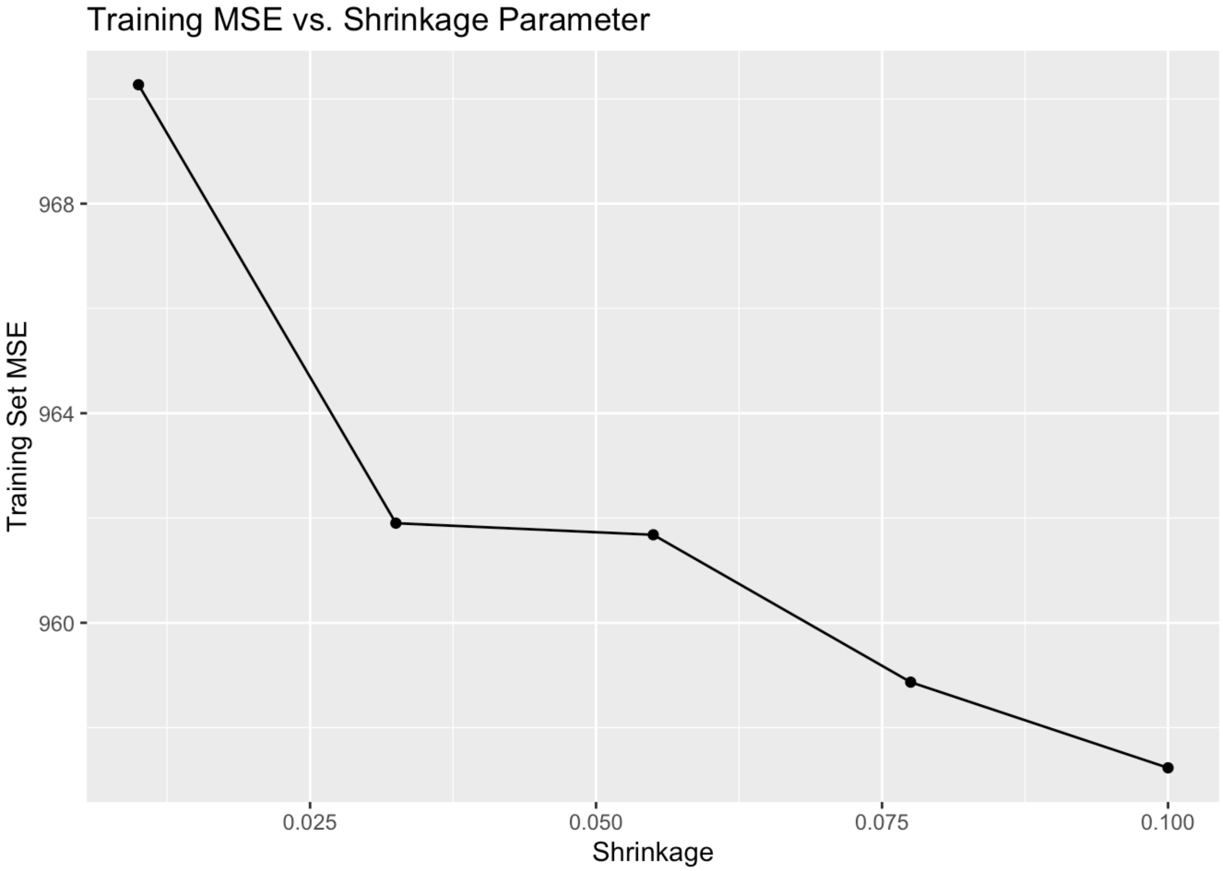


Interpretation

The variable importance plot from the Random Forest model shows that ‘points’, representing a wine’s rating, is the most significant predictor of its price, indicating a direct correlation between quality ratings and price points. ‘Vintage’, which denoting the year of production, also emerges as a critical factor, which aligns with the notion that the age and harvest conditions of the wine are important for valuation. Additionally, specific descriptors from wine reviews, like ‘cherry’, suggest that certain flavor notes are valued in the wine’s market price. These insights can be pivotal for producers and sellers in understanding which aspects of a wine—its rated quality, age, and flavor profile—are most valued in the marketplace and should be emphasized in marketing and pricing strategies.

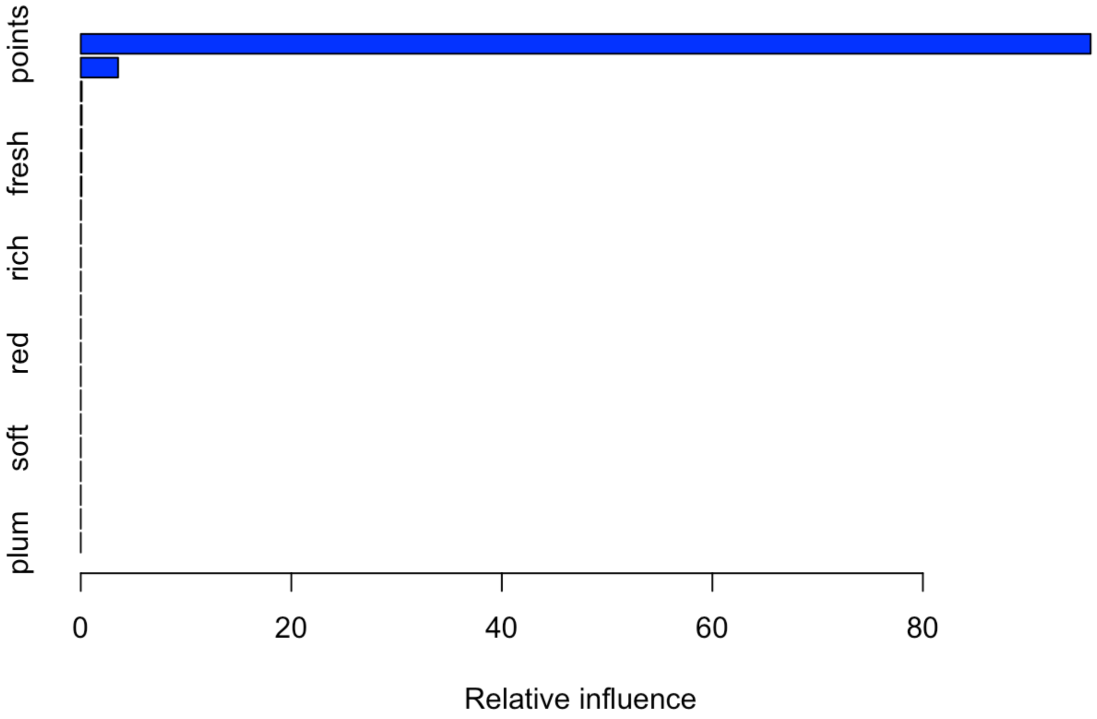
Boosting

For the boosting model, I will conduct the boosting process with 1,000 trees, considering a range of values for the shrinkage parameter λ. Then, I’ll generate a plot with various shrinkage values on the x-axis and the corresponding training set Mean Squared Error (MSE) on the y-axis to identify the optimal boosting model.



```
##   lambda      MSE
## 1 0.0100 970.2703
## 2 0.0325 961.9017
## 3 0.0550 961.6797
## 4 0.0775 958.8670
## 5 0.1000 957.2313
```

Optimal Boosting Model



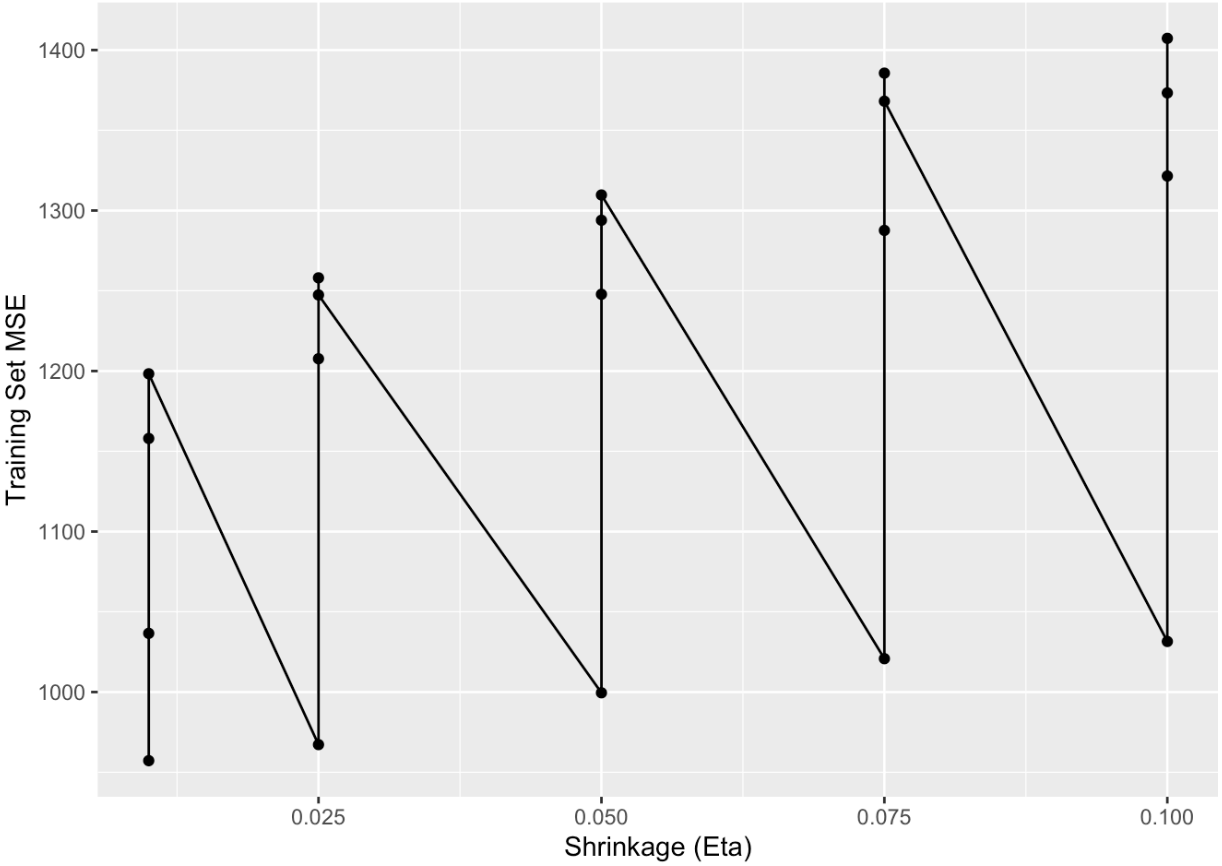
```
##                   var      rel.inf
## points         points 95.933824392
## vintage       vintage  3.543777813
## apple           apple  0.094768904
## white           white  0.091977821
## spice           spice  0.085633818
## fresh           fresh  0.074816570
## ripe             ripe  0.070792072
## blackberry blackberry  0.029702658
## cherry         cherry  0.021379794
## rich             rich  0.017829613
## light           light  0.010383333
## black           black  0.010195878
## oak               oak  0.006097930
## red               red  0.005287250
## fruits         fruits  0.003532156
## blend           blend  0.000000000
## dry               dry  0.000000000
## soft             soft  0.000000000
## sweet           sweet  0.000000000
## berry           berry  0.000000000
## dark             dark  0.000000000
## plum             plum  0.000000000
```

Interpretation

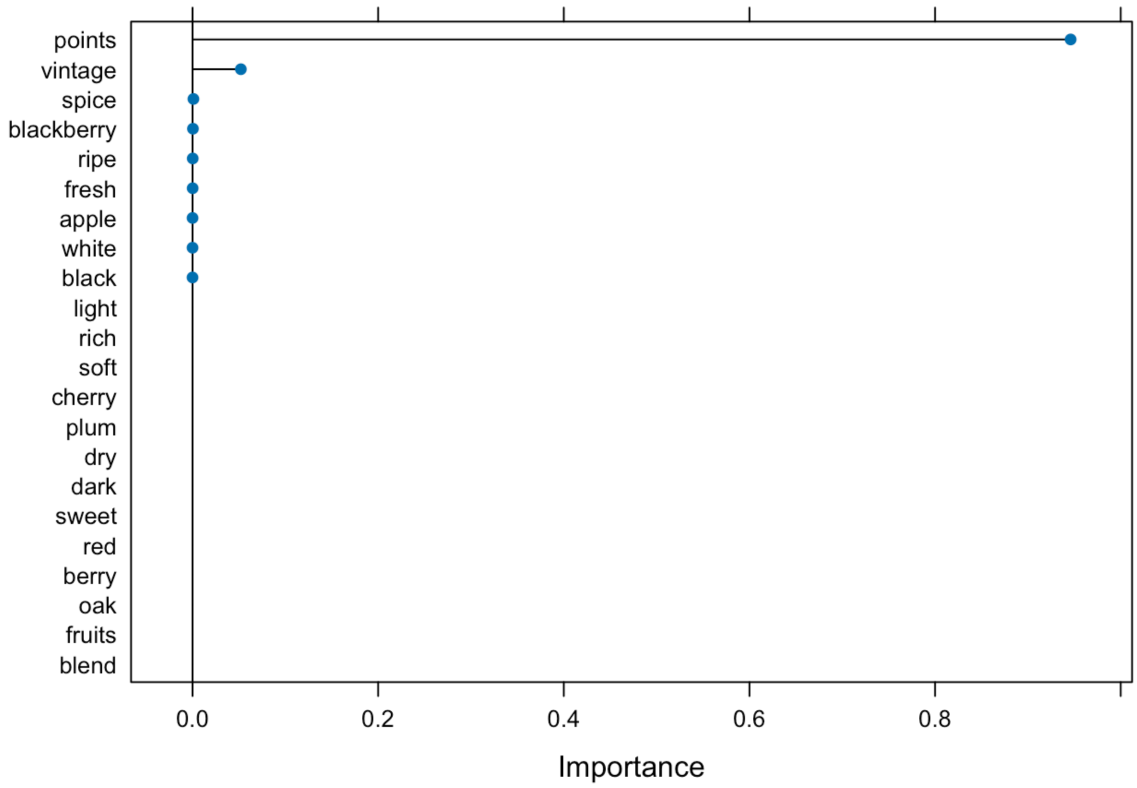
The summary of the boosting model provides a quantitative measure of the importance of different variables in predicting wine prices. Clearly, ‘points’ is the most significant predictor by a considerable margin, underscoring its strong influence on price. ‘Vintage’, also plays a notable role, albeit much less than ‘points’. Descriptive terms from reviews, such as ‘spice’, ‘apple’, and ‘fresh’, though contributing relatively minor predictive power individually, collectively offer insight into the nuanced characteristics that can affect a wine’s market value. ‘Blackberry’, ‘rich’, and ‘cherry’ follow closely, indicating specific flavors or qualities that may appeal to consumer preferences and drive pricing. These results suggest that while quality ratings vastly dominate price predictions, the nuanced sensory descriptors and vintage also carry weight in the valuation of wine.

Gradient Boosting

For the gradient boosting model, I will establish a grid search on the learning rate, denoted eta in this case. I’ll then train models across the grid and calculate the MSE for each. Finally, I will select the optimal model, which is the one with the lowest MSE.



Optimal Gradient Boosting Model



Interpretation

The gradient boosting model's summary plot illustrates the relative importance of various predictors in determining wine prices. 'Points' significantly overshadow all other variables, affirming its critical role in price prediction. 'Vintage' also appears as a key variable, suggesting that the year of production is an important factor, potentially due to the characteristics of different vintages affecting a wine's value. Descriptive terms such as 'spice' and 'blackberry' hold modest importance, indicating that specific tasting notes contribute to pricing to a lesser extent. These flavor descriptors, along with 'ripe' and 'fresh', may reflect consumer taste preferences or wine characteristics that subtly influence its market price. This plot highlights that while the subjective quality measure ('points') dominates, both the wine's age and sensory descriptors play roles in its valuation.

Model Comparison

Finally, I will compare the performance of each model using the test RMSE to determine which one has the best performance.

Test RMSE for Different Models

Model	Test_RMSE
Random Forest	31.114
Boosting	30.641
Gradient Boosting	30.366

In this comparison, the Gradient Boosting model achieves the lowest RMSE at 30.366, suggesting it is the most accurate model among the three in predicting wine prices. The Boosting model follows closely with an RMSE of 30.722, while the Random Forest model has the highest RMSE at 31.114. This indicates that, for this particular dataset, ensemble methods with boosting techniques have a slight edge in predictive accuracy over the Random Forest approach.

Conclusion

Throughout this analysis, I have addressed all the questions I posed at the outset. Using linear and spline models, I explored the relationship between a wine’s vintage and price, as well as the relationship between its rated points and price. The models indicate a negative non-linear correlation between vintage and price and a non-linear positive relationship between points and price. However, both models suggest that they explain only a small portion of the variability in price, implying that vintage and points alone are not strong predictors of price. Moreover, by applying natural language processing techniques to wine reviews, I discovered that highly-rated wines are often described with words like “black,” “ripe,” “cherry,” “rich,” and “spice,” whereas descriptors for expensive wines include “black,” “cherry,” “ripe,” “oak,” and “spice.” This offers valuable insights into the language used in wine ratings. Finally, I constructed three different models to predict wine prices, incorporating vintage, points, and vectorized descriptive variables. The gradient boosting model emerged as the top performer. The variable importance plots from each model affirm that points are the most significant predictor of price, with vintage also playing a crucial role. Descriptive variables like “cherry,” “sweet,” and “apple” have a relatively substantial impact on price predictions, offering a nuanced understanding of how specific wine characteristics can influence market value.