**Table of Contents**

**An Analysis of Wine**

1. Overview……………………………………………………………………………………………………………………………………………..2

2. Methods used throughout the analysis………………………………………………………………………………………………..2

a. Data Hygiene………………………………………………………………………………………………………………………….2

b. Gaining and understanding of the data………………………………………………………………………………….2

i. Relationship of Price to Points…………………………………………………………………………………….2

ii. Variety Statistics………………………………………………………………………………………………………..5

iii. Country Statistics……………………………………………………………………………………………………….6

iv. Region Statistics………………………………………………………………………………………………………..7

c. Sampling and Cross-Validation………………………………………………………………………………………………7

d. Predictive Modeling………………………………………………………………………………………………………………8

e. Prescriptive Modeling……………………………………………………………………………………………………………8

i. Decision Analysis…………………………………………………………………………………………………………8

ii. Inventory Simulation………………………………………………………………………………………………….8

iii. Variety/Product Selection – Profit Maximization……………………………………………………….8

3. Determinants of Wine Rating…………………………………………………………………………………………………………….8

a. United States…………………………………………………..……………………………………………………………………9

b. France……………………………………………………………..…………………………………………………………………..9

c. Italy…………………………………………………………………..………………………………………………………………….9

d. Spain…………………………………………………….…………..……………………………………………………………….10

e. Portugal……………………………………………….…………………………………………………………………………….10

f. Chile…………………………………………………….………………………………………………………………………….….10

g. Argentina…………………………………………….………………………………………………………………………….….10

h. Australia…………………………………………….…………………………………………………………………………..….11

i. Austria……………………………………………….……………………………………………………………………………….11

j. New Zealand…………………………………….…………………………………………………………………………….….11

4. Conclusions………………………………………………………………………………………………………………………………………12

5. Recommendations……………………………………………………………………………………………………………………………12

a. Notable Designations…………………………………………………………………………………………………………13

**Overview**

The data analyzed was scraped from the accredited Wine Enthusiast Magazine. This data provides us with characteristics of over 280,000 wines. These characteristics include country, designation, points (rating), price, province, region, variety, winery, and description. The main objective of this analysis is to notice patterns and/or variables that drive wine rating. Wines within the dataset range from 80-100 in terms of ratings.

**Methods Used Throughout the Analysis**

*Data Hygiene*

The data scraped from Wine Enthusiast was in a very raw form, initially. There were many instances of missing data and the presence of variables that were not useful for this analysis. To carry out the necessary analytical strategies, the dataset was cleaned sufficiently while retaining relevant variables and meaningful data.

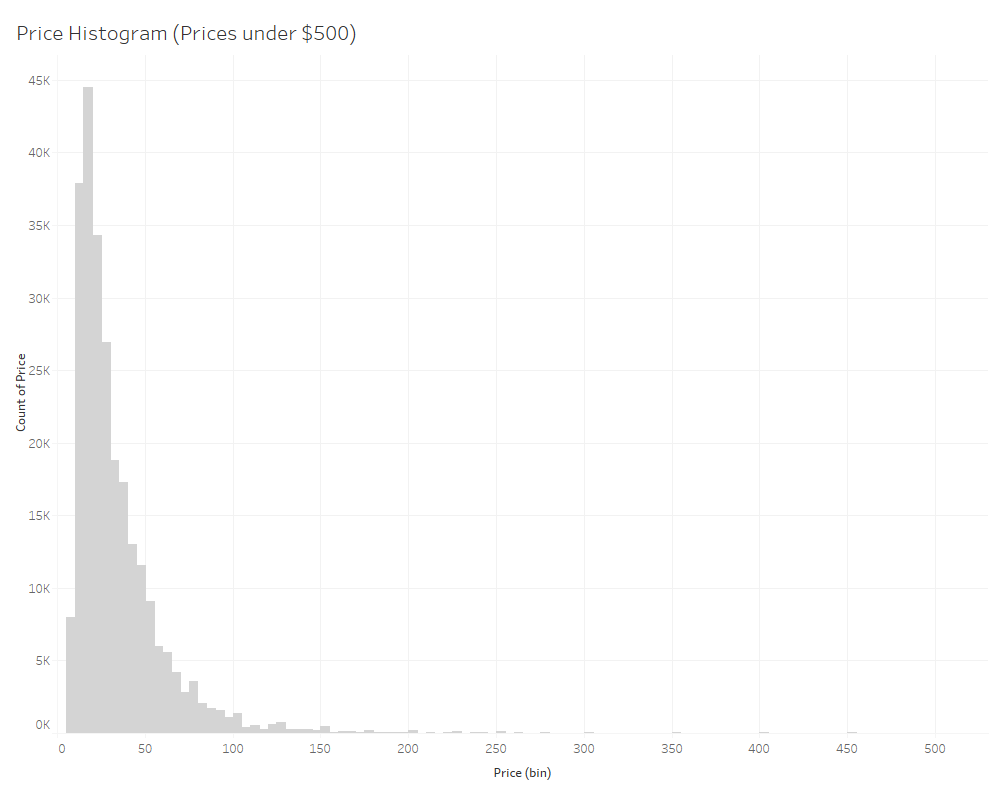
*Gaining an Understanding of the Data*

Before implementing detailed and advanced analytics strategies, it is imperative to gain an understanding of the behavior of the data. Various summary statistics were discovered in an effort to comprehend the dataset at hand.

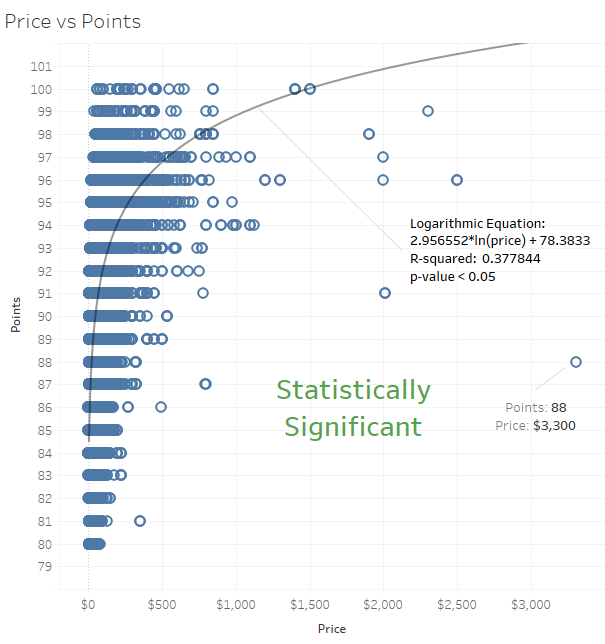
**Relationship of Price to Points**

For quantitative variables, points and price descriptive statistics are found.



Interesting findings for points are an average rating of 88, little presence of outliers, and little variability. Descriptive statistics for price show that an average price of about $34.18, presence of outliers, such as the maximum price of $3,300, and a high degree of variability. A histogram of the price data under $500 is shown below.

It is clear that the majority of prices fall between around $10 and $50.

There is a clear relationship between price and points judging by the color gradient in the average price column above.

A logarithmic trend line was used as the best fit for this relationship. There is a significant positive correlation between price and points, albeit not a very strong one with an R-squared of 0.377844.

**Variety Stats**

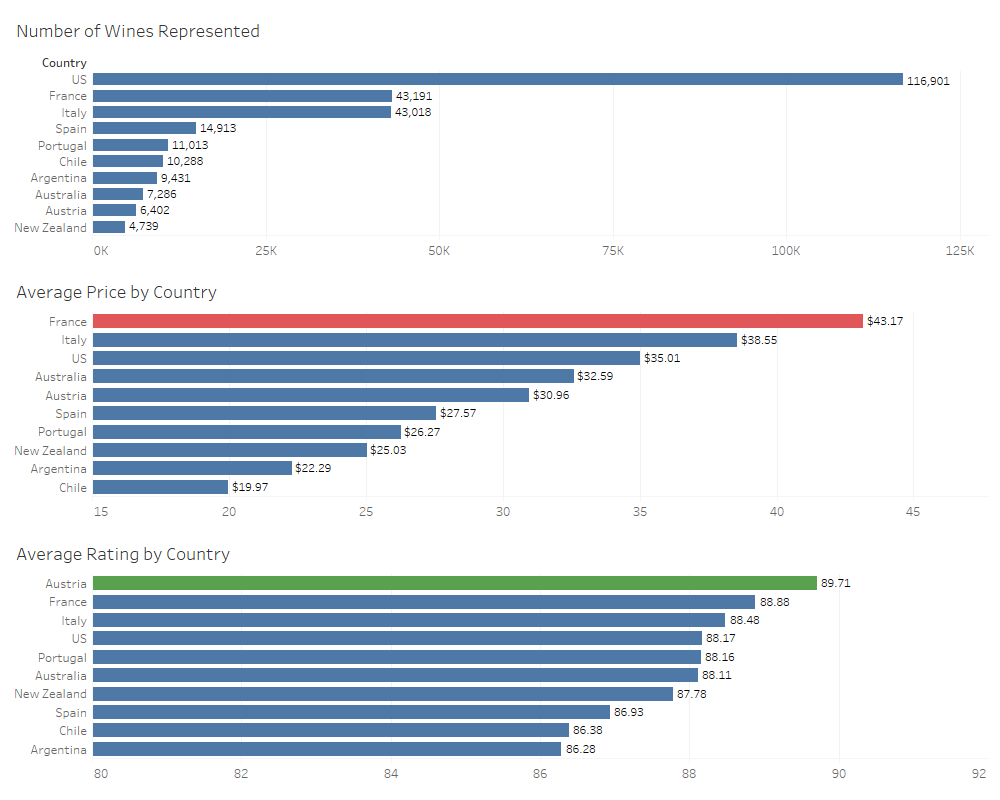
A summary of the top 25 varieties and their respective point and price information is shown below.



Pinot Noir is the most prevalent wine variety, followed closely by Chardonnay. The variety with the highest average rating, and the only average rating above 90, is Nebbiolo, but has a high average price. I contrast, Pinot Grigio has the lowest average points, and also the lowest average price.

**Country Stats**

For the purpose of this analysis, we will be looking at the top 10 countries in terms of number of wines represented. These countries’ wine selections are the most easily accessible in terms of obtaining supply for future inventories.



The dataset is made up mostly of US wines. France has a substantially higher average price than other countries, while Chile’s wines are substantially lower in price on average. Austria has the highest average wine rating, while Argentina as the lowest average rating.

**Regional Statistics**

A summary of the top 25 regions and their respective point and price information is shown below.



Napa Valley in California has the most wines represented in the dataset, followed by Columbia Valley in Washington State. The region with the highest average rating is Barolo in Italy, but has a moderately high average price. In contrast, California has the lowest average points, and also the lowest average price. This goes along with the theme presented throughout this analysis that price has a significant relationship with points.

*Sampling and Cross-Validation*

For more advanced analysis, specifically to implement predictive modeling and identify variables that determine points, the master dataset is split into 10 samples. These separate samples consist of each of the top 10 countries in terms of wines represented. The top 10 countries include United States, France, Italy, Spain, Portugal, Chile, Argentina, Australia, Austria, and New Zealand. A model validation method called cross-validation is used. These methods involve using include a train/test data split (70%/30%) and K-fold for each country’s sample. Using the training data to develop predictive models and the test data to identify overfitting and overall error is essential to verifying the fit and validity of a statistical model. K-fold cross-validation is necessary for developing a random forest predictive model.

*Predictive Modeling*

For each country’s sample, ordinary logistic regression, stepwise logistic regression, and neural network methods are used to build predictive models to predict wine points. The chosen model to be used to predict outcomes for future scenarios will have the highest accuracy measure possible, while also having balanced sensitivity and specificity. Accuracy is the proportion of total correct observations to incorrect observations. Sensitivity is the proportion of observed positives that were predicted to be positive (true positives), whereas specificity is the proportion of observed negatives that were predicted to be negative (true negatives). For the purpose of this analysis, points scores from 80-89 are classified as “No” and scores from 91-100 are classified as “Yes.” Wines with ratings from 91-100 are understood to be exceptional wines.

*Prescriptive Modeling*

**Decision Analysis**

This prescriptive model can be used for specific products and decisions. Demand levels, purchase levels, variable costs, selling price, and the coefficient of realism can be given new values to receive unique results for different decisions. This model is intended for repeated use with new and varying scenarios.

**Inventory Simulation**

Per request, a model has been developed to better predict inventory levels and costs going forward. This model is built to simulate a year's worth of inventory rotation for each week. This model was developed with the historic values provided for beginning inventory, demand, order quantity, reorder point, inventory cost, stockout cost, and order cost.

**Variety/Product Selection - Profit Maximization**

This model can be used when deciding which varieties, specific wine designations, regions, countries, etc. to include in your product portfolio to maximize profit. Once desired constraints and conditions are plugged in, the model is designed to output profit maximizing results.

**Determinants of Wine Rating**

Each country’s model has unique results. When using these predictive models to predict future scenarios, the appropriate country must be selected. Variables explored as predictors include price, province, and variety. For each of the 10 countries, the chosen model, model accuracy, model sensitivity, model specificity, and significant predictor variables are as follows:

*United States*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7268

Sensitivity: 0.7211

Specificity: 0.7301

Significant Predictor Variables: Price and provinces of Colorado, Michigan, Texas, Virginia, and Washington

*France*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7797

Sensitivity: 0.7656

Specificity: 0.7717

Significant Predictor Variables: Price and provinces of Beaujolais, Bordeaux, Burgundy, Champagne, France other, Languedoc-Roussilon, Loire Valley, Rha’ne Valley, and Southwest France

*Italy*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7938

Sensitivity: 0.7713

Specificity: 0.8057

Significant Predictor Variables: Price

Provinces include Northeastern Italy, Sicily & Sardinia, Southern Italy, Tuscany, and Veneto.

Varieties include Aleatico, Arneis, Barbera, Bordeaux-style Red Blend, Cabernet Sauvignon, Cabernet-Sauvignon-Merlot, Cannonau, Carignano, Chardonnay, Corcina, Rondinella, Molinara, Dolcetto, Falanghina, Frappato, Friulano, Gargenega, Gewarztraminer, Glera, Grillo, Insolia, Kerner, Lagrein, Lambrusco, Lambrusco Grasparossa, Maller-Thurgau, Malvasia, Marzemino, Merlot, Montepulciano, Moscato, Nebbiolo, Negroamaro, Nerello Mascalese, Nero D-Avola, Nosiola, Pallagrello, Passerina, Pecorino, Piedirosso, Pinot Bianco, Pinot Grigio, Pinot Nero, Primitivo, Prosecco, Prugnolo Gentile, Red Blend, Ribolla Gialla, Riesling, Rosa, Rosato, Sagrantino, Sangiovese, Sangiovese Grosso, Sauvignon, Sauvignon Blanc, Schiava, Sparkling Blend, Sylvaner, Syrah, Tocai Friulano, Trebbiano, Turbiana, Uva di Troia, Verdicchio, Vermentino, Vernaccia, and Viognier.

*Spain*

Model: Neural Network

Accuracy: 0.8435

Sensitivity: 0.5352

Specificity: 0.9245

Significant Predictor Variables: N/A

*Chile*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.8418

Sensitivity: 0.8055

Specificity: 0.8473

Significant Predictor Variables: Price

Provinces include Aconcagua Valley, Central Valley, Elqui Valley, and Repel Valley-Casablanca Valley

Varieties include Cabernet Sauvignon, Cabernet Sauvignon-Carmena re, Cabernet Sauvignon-Syrah, Carmena re, Chardonnay, Malbec, Merlot, Pinot Noir, Sauvignon Blanc, and Shiraz

*Portugal*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7937

Sensitivity: 0.7839

Specificity: 0.7913

Significant Predictor Variables: Price and the province, Beira Atlantico

*Argentina*

Model: Neural Network

Accuracy: 0.8858

Sensitivity: 0.5403

Specificity: 0.955

Significant Predictor Variables: N/A

*Australia*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7832

Sensitivity: 0.754

Specificity: 0.7982

Significant Predictor Variables: Price

Provinces include South Wales, Queensland, South Australia, Tasmania, Victoria, and Western Australia

Varieties include Cabernet Sauvignon, G-S-M, Grenache, Muscat, Rhane-style Red Blend, Riesling, Sacmillon, Shiraz, Grenache, and Shiraz-Viognier

*Austria*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7269

Sensitivity: 0.737

Specificity: 0.7162

Significant Predictor Variables: Price and varieties Cabernet Sauvignon, Gelber Muskateller, Merlot, Pinot Noir, Red Blend, Sauvignon Blanc, St. Laurent, Traminer, Zweigelt

*New Zealand*

Model: Forward Stepwise Logistic Regression

Accuracy: 0.7243

Sensitivity: 0.7628

Specificity: 0.7101

Significant Predictor Variables: Price

Provinces include Canterbury, Central Otago, East Coast, Hawke’s Bay, Marlborough, Martinborough, Nelson, New Zealand, Waiheke Island, and Wairapara.

Varieties include Gewarztraminer, Merlot, and White Blend

**Conclusions**

Chile has the most reliable predictive model. With that being said, every country’s model is moderately accurate. These models will be useful in predicting future scenarios involving country, price, province, and variety. Price, as a predictor of points, is a theme throughout this analysis. The summary statistics involving the entire dataset show this, and, each predictive model shows price as the most significant predictor of points.

**Recommendations**

- Utilize prescriptive models to maintain inventories, maximize profit, and moderate risk-oriented decisions.

- Implement country-specific predictive models to predict wine ratings. Understand that some models are more accurate than others.

- As a general rule, you may have to pay more for higher-rated wines.

- When looking to diversify regional and varietal wine selections, look to the data as a place to start. For instance, if looking to include more Italian wines, start with the region of Barolo.

- Consider adding the below notable wines to your selection.

**Notable Designations**

*These designations have exceptional ratings while being reasonably priced.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Designation** | **Points** | **Price** | **Country** | **Province** | **Variety** |
| En Chamberlin Vineyard | 100 | $65 | United States | Oregon | Syrah |
| Estate Vineyard | 99 | $44 | United States | California | Chardonnay |
| Maritime Vineyard | 98 | $52 | United States | California | Pinot Noir |
| Fenaughty Vineyard | 97 | $35 | United States | California | Syrah |
| Grenache Noir | 96 | $27 | United States | California | Grenache |
| Bacchus Vineyard | 95 | $20 | United States | Washington | Riesling |
| Assobio | 94 | $13 | Portugal | Douro | Portuguese Red |
| Aydie l-Origine | 93 | $12 | France | Southwest France | Tannat-Cabernet Franc |
| Follies Fonte Nossa Senhora da Vandoma | 92 | $11 | Portugal | Bairrada | Touriga Nacional-Cabernet Sauvignon |
| Toutalga | 91 | $7 | Portugal | Alentejano | Portuguese Red |