CAPSTONE SUBMISSION

An analysis of the variables which contribute to the ratings of different wines across the world

Benjamin Howard – Student ID 500901525

Ryerson University

Table of Contents

INTRODUCTION

Wine has been enjoyed by humanity for thousands of years. Yet each year, millions of Canadians purchase wines from all around the world without properly understanding how the region, price, and country of origin affect the quality and/or rating of the wine.

While wine reviews are subjective based on the tastes of professional sommeliers, there may exist common traits in wines that can help predict the quality of the wine to regular consumers. A wine's price may not necessarily indicate its quality, and there may be more important factors such as flavor, region, reviews etc.

This project seeks to determine if certain variables including the country a wine was produced in, it's price and age are reflective of the quality of the wine, and if they can be used to predict the ratings of individual wines.

LITERATURE REVIEW

To better understand the factors that contribute to the ratings, quality and price of wine, several wine articles were referenced.

The first paper reviewed was titled, "Analyzing the us retail wine market using price and consumer segmentation models". This was helpful understanding the importance of price as a variable when consumers select and purchase wines. In fact, it demonstrated that consumers prefer cheaper "jug" wine (~3$ USD) representing roughly 44% of the total volume sold. Price appears to have more of a weight than quality[1]

The second article, titled, "Reality of Wine Prices (What You Get for What You Spend)" was important as it solidified the categorization of prices that I would be using in my analysis. This will be covered in the "Approach" section of this submission. Displayed in the appendix is a helpful graphic of the categorization[2]

Lastly, the study done by researchers at **Caltech Wine Study Shows Price Influences Perception** demonstrates that consumers opinions and ratings of wines are influenced and sometimes heavily based on price. For example, "*When the subjects were told the wine cost $90 a bottle, they loved it; at $10 a bottle, not so much. In a follow-up experiment, the subjects again tasted all five wine samples, but without any price information; this time, they rated the cheapest wine as their most preferred"* [3]

DATASET

The dataset that was chosen for this analysis was scraped from Wine Enthusiast in late 2017 and has 129,941 records across 14 variables. Below is a table that displays all the different variables and their descriptions as well as data types.

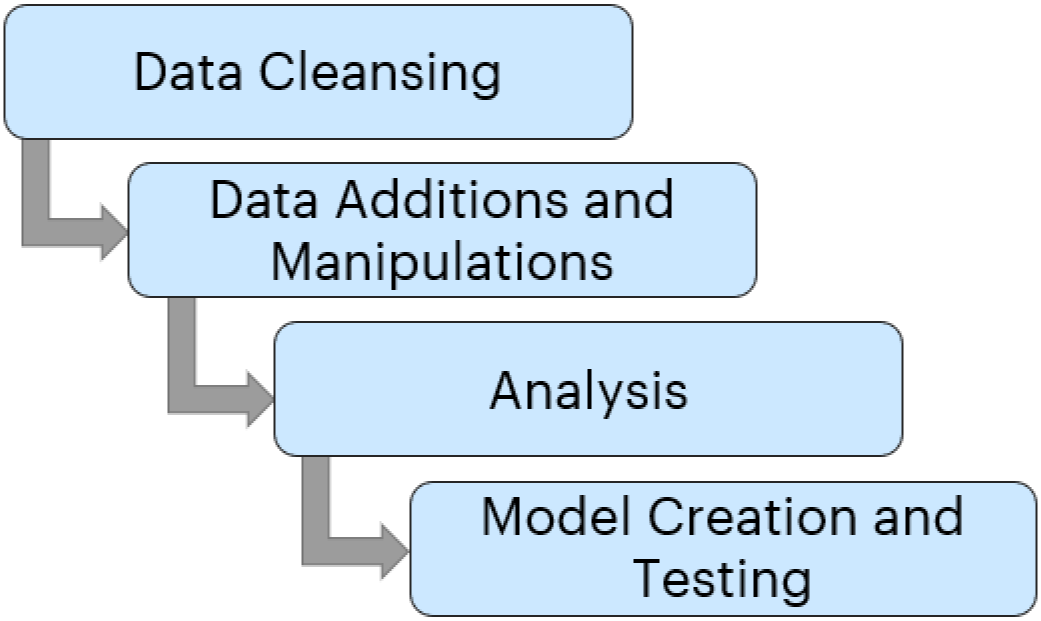
Table for Wine Enthusiast's Wine Extract

Field	Description	Data Type
Country	Country where the wine is produced	TEXT
Description	Description of the wine and flavors	VARCHAR
Designation	The vineyard within the winery where the grapes that made the wine are from	TEXT
Points	The number of points WineEnthusiast rated the wine on a scale of 1-100	INT
Price	The cost for a bottle of the wine	INT
Province	The province or state that the wine is from	TEXT
Region_1	The wine growing area in a province or state (ie Napa)	TEXT
Region_2	Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this	TEXT
Taster Name	Name of the reviewer	TEXT
Taster Twitter	Twitter handle of reviewer	VARCHAR
Title	Name of the wine	VARCHAR
Variety	Type of wine – Chardonnay for example	TEXT
Winery	Winery where it comes from	TEXT

Of course, this dataset would be transformed and cleansed thoroughly during the data cleansing portion of this analysis. More

METHODOLOGY

In this section, the methodology that was undertaken to complete this analysis will be explained and then later elaborated in subsequent parts of the submission. The following is a visual representation of the methodology.



DATA CLEANSING

This step of the methodology consists of cleaning up the dataset. Most dataset that are scraped from external sources are not cleaned up or formatted. This step includes, removing NA values, removing duplicates, eliminating outliers etc.…

Removing NA Values and Incomplete Values

The dataset contained over 8,000 NA records as well as several BLANK values. Using simple R code, these values were removed from the dataset. Any rows that did not have values for Points, Country, Price, Title, Variety and Description were removed.

Removing Variables

Several variables in the dataset were not relevant and were removed from the dataset.

- **Row Number** was removed because it came as part of the initial dataset and was not required.

- **Region_2** was removed due to the large amount of BLANK values. Additionally, this level of granularity was not necessary for the depth of this analysis.
- **Designation** was removed due to the fact that Title and Winery contained the same information, but it was not concatenated as was designation.
- **Taster Twitter** was removed due to the fact that the Twitter handle of a taster is not relevant for this analysis.
- **Removed Countries** with less than 400 records because of sample size

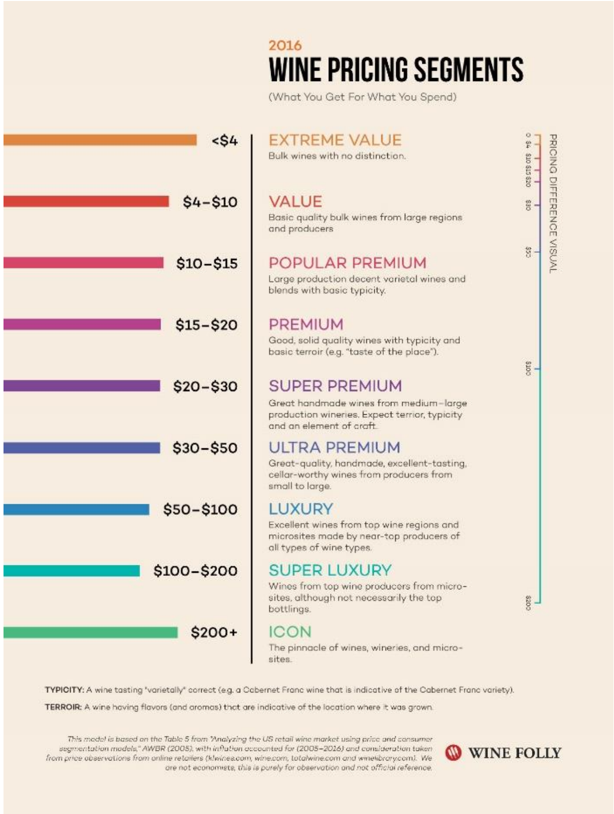
Adding new variables

This was a crucial step of the project because we needed to add variables to the dataset that were not available initially. There were two categorical variables that needed to be created so that they could be used in a categorical classification model.

- **Price Range** was created by using the price of the bottle of wine to assign it to a corresponding range category. These categories consisted of:
  - "Value" were wines that cost less than 10$ per bottle
  - "Premium" were wines that cost less than 20$ per bottle
  - "Ultra Premium" were wines that cost less than 40$ per bottle
  - "Luxury" were wines that cost less than 100$ per bottle
  - "Super Luxury" were wines that cost more than 100$ per bottle

  Important to note that there were additional categories of higher priced wines however they were removed during the outlier clean-up process which is described below

  Inspiration for these values came from a wine review site which provided this graphic of their categorizations [4]



2016

WINE PRICING SEGMENTS

(What You Get For What You Spend)

<$4	EXTREME VALUE	Bulk wines with no distinction.
$4–$10	VALUE	Basic quality bulk wines from large regions and producers
$10–$15	POPULAR PREMIUM	Large production decent varietal wines and blends with basic typicity.
$15–$20	PREMIUM	Good, solid quality wines with typicity and basic terroir (e.g. "taste of the place").
$20–$30	SUPER PREMIUM	Great handmade wines from medium–large production wineries. Expect terrior, typicity and an element of craft.
$30–$50	ULTRA PREMIUM	Great-quality, handmade, excellent-tasting, cellar-worthy wines from producers from small to large.
$50–$100	LUXURY	Excellent wines from top wine regions and microsites made by near-top producers of all types of wine types.
$100–$200	SUPER LUXURY	Wines from top wine producers from micro-sites, although not necessarily the top bottlings.
$200+	ICON	The pinnacle of wines, wineries, and micro-sites.

TYPICITY: A wine tasting "varietally" correct (e.g. a Cabernet Franc wine that is indicative of the Cabernet Franc variety).

TERROIR: A wine having flavors (and aromas) that are indicative of the location where it was grown.

This model is based on the Table 5 from "Analyzing the US retail wine market using price and consumer segmentation models," AWBR (2005), with inflation accounted for (2005–2016) and consideration taken from price observations from online retailers (klwines.com, wine.com, totalwine.com and winelibrary.com). We are not economists, this is purely for observation and not official reference.

Source located in references

- **Year** was created based on the year of the wine that was in the "Title" variable of each wine. To extract a number from the CHAR field was done through an extraction string function.

  Below is a sample of a "Title" value in the dataset.
  Rainstorm 2013 Pinot Gris (Willamette Valley)

  The extract isolated the numerical value, and created a variable called "Year" and the variable was now a NUM variable and not CHAR.

- **Age Range** was the last new variable created and this was a categorical variable derived from the newly created “Year” variable. Similar to “Price Range”, the following logic was applied to the records based on the “Year”
    - Wines between the age of 2013-2016 labeled as “Young”
    - Wines between the age of 2006-2013 labeled as “Mature”
    - Wines older than 2006 labeled as “Old”

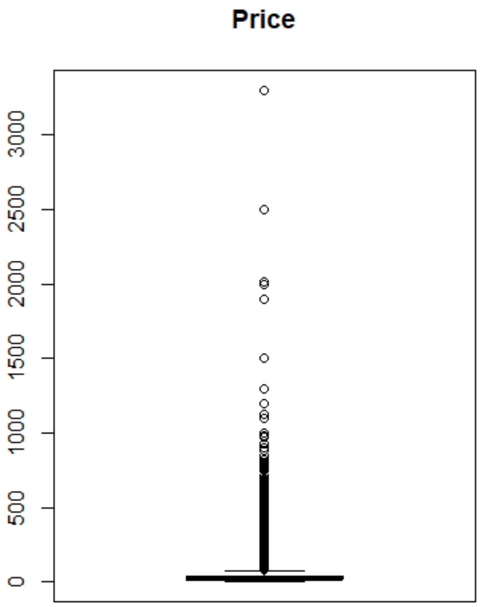
  This classification was based on numerous articles comparing the age ranges of wines for both red wines and white wines. These wines have differences in age ranges, Red wines being more acceptable to aging than White wines.

Removing Outliers

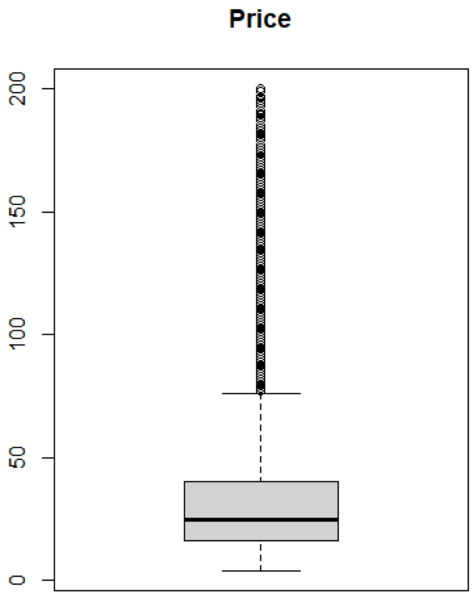
Removing outliers is a key step in Data Cleansing because the outliers may distort models and averages while not being representative of the dataset.

Below are two boxplots which show the distributions of the “Price” variable.

Boxplot of Price of Wines from initial dataset



Boxplot of Price of Wines after removing all wines above 200$



Upon initial analysis, it was discovered that 99.7% of all wines were below 200$. It was further discovered that eliminating wines over 100$ from the dataset would reduce the margin of error in the linear model. This will be described in detail in the “Modelling” section of the submission.

Additional Data Cleansing Notes

There were several other minor changes in the datasets including removing values with special characters or spelling mistakes and changing certain variable types from CHAR to NUM, and additional minor table formatting changes.

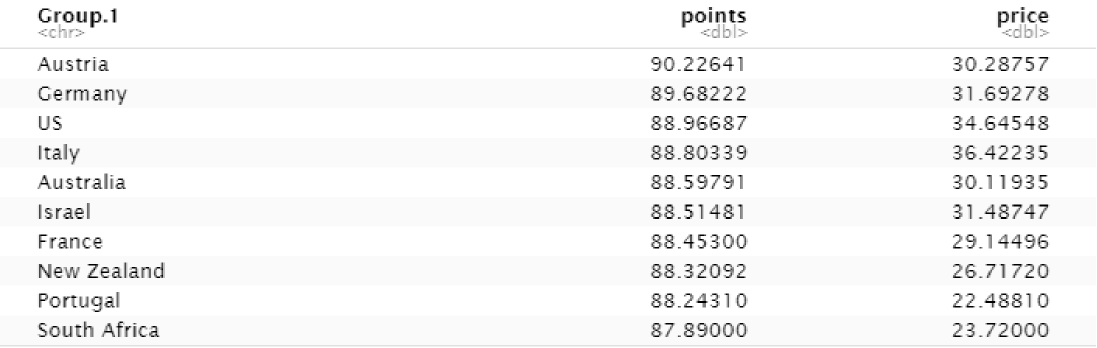
ANALYSIS

After the Data Cleansing exercise, the file consisted of 81,915 records with 14 variables, including the new created categorical variables. With the clean dataset, it was possible to do discovery analysis to assess any initial observations before running the regression and classification models.

With points and price being the primary focus of this analysis, it was key to look at the average of price per country, as well as average of points earned.

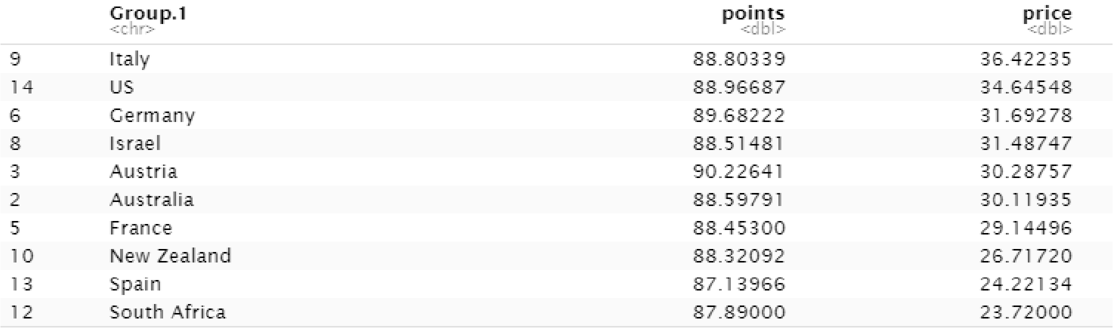
In more detail, here are top ten for each variable in descending order:

Top 10 Countries by Points (descending)



Group.1 <chr>	points <dbl>	price <dbl>
Austria	90.22641	30.28757
Germany	89.68222	31.69278
US	88.96687	34.64548
Italy	88.80339	36.42235
Australia	88.59791	30.11935
Israel	88.51481	31.48747
France	88.45300	29.14496
New Zealand	88.32092	26.71720
Portugal	88.24310	22.48810
South Africa	87.89000	23.72000

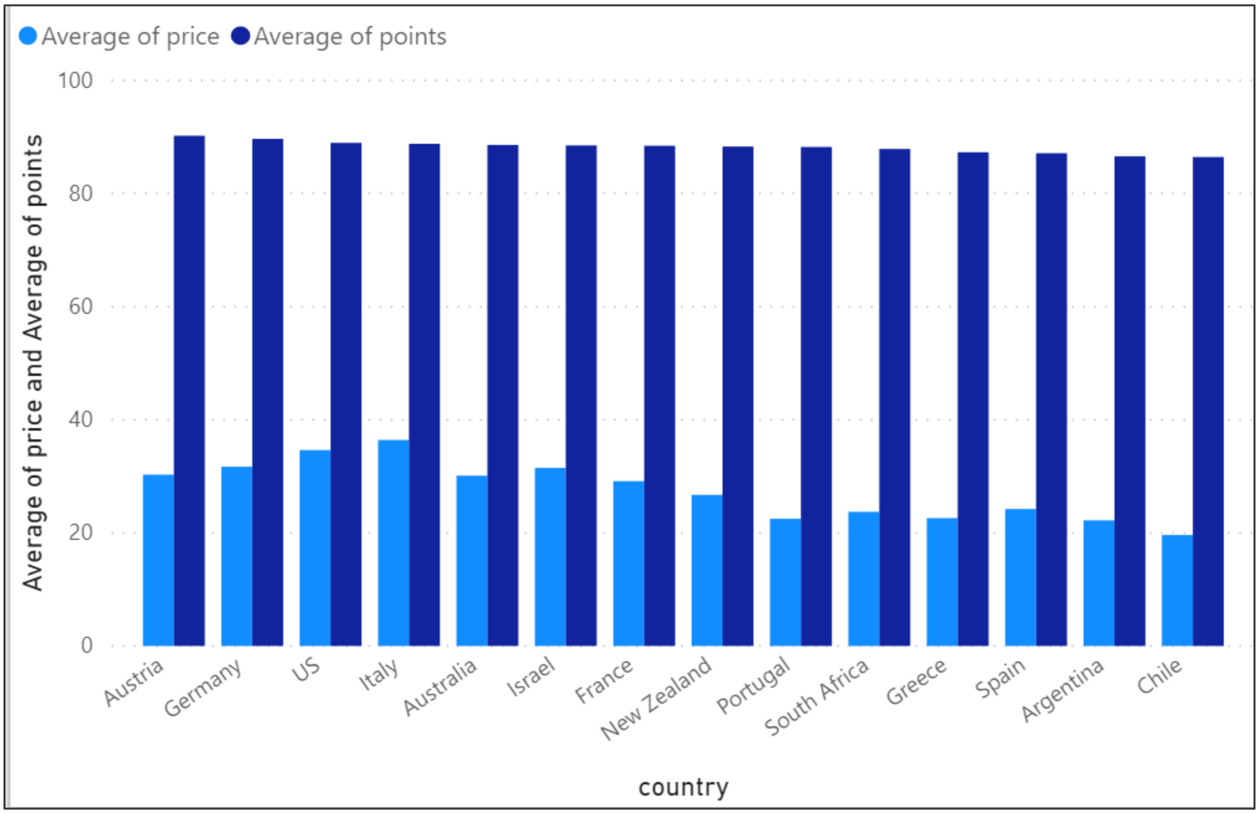
Top 10 Countries by Price (descending)



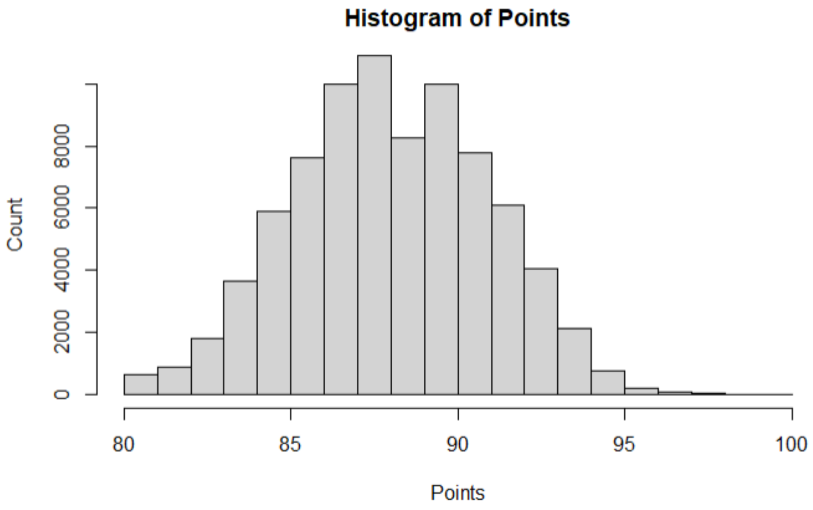
	Group.1 <chr>	points <dbl>	price <dbl>
9	Italy	88.80339	36.42235
14	US	88.96687	34.64548
6	Germany	89.68222	31.69278
8	Israel	88.51481	31.48747
3	Austria	90.22641	30.28757
2	Australia	88.59791	30.11935
5	France	88.45300	29.14496
10	New Zealand	88.32092	26.71720
13	Spain	87.13966	24.22134
12	South Africa	87.89000	23.72000

Here is a complete graphical visual of the average points and price of each country

Average Price and Points per Country



It is also important understand the distribution and potential for linear models of the dataset. Looking at the numerical values for price and years. Points was the only NUM variable that has normal distribution as evidenced below.



Lastly, before modelling it was important to determine if there was a relationship between Points and Price, and Points and Year.

The Pearson correlation values for the following are:

Points ~ Price = 0.5615601

Points ~ Year = 0.01295443

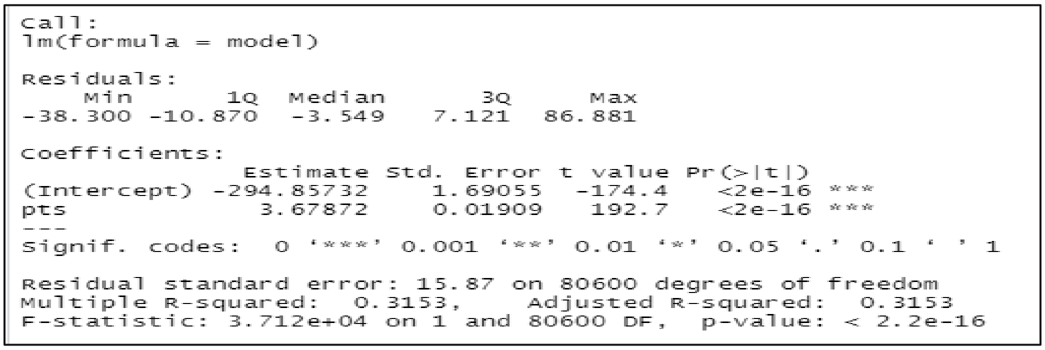
Based on these results we can determine that there is a positive correlation between Points and Price, while Points and year does not show a strong correlation.

MODELLING

Linear Regression Points ~ Price

The first model applied to the dataset is a linear regression model between Points and Price to determine if an increase in Price leads to an increase in Points.

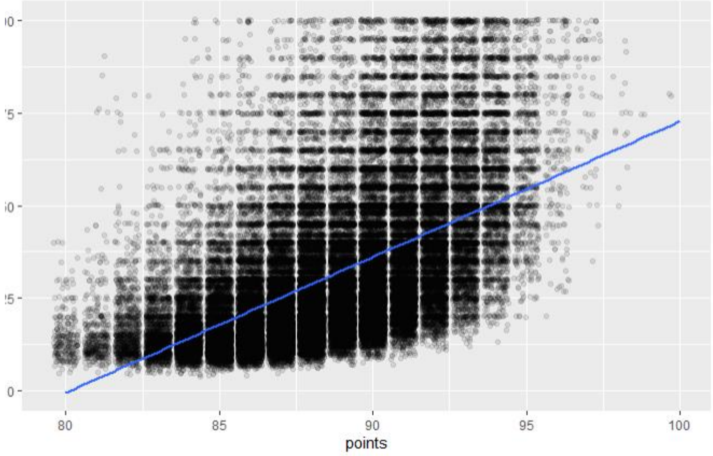
Below are the findings for the Linear Model



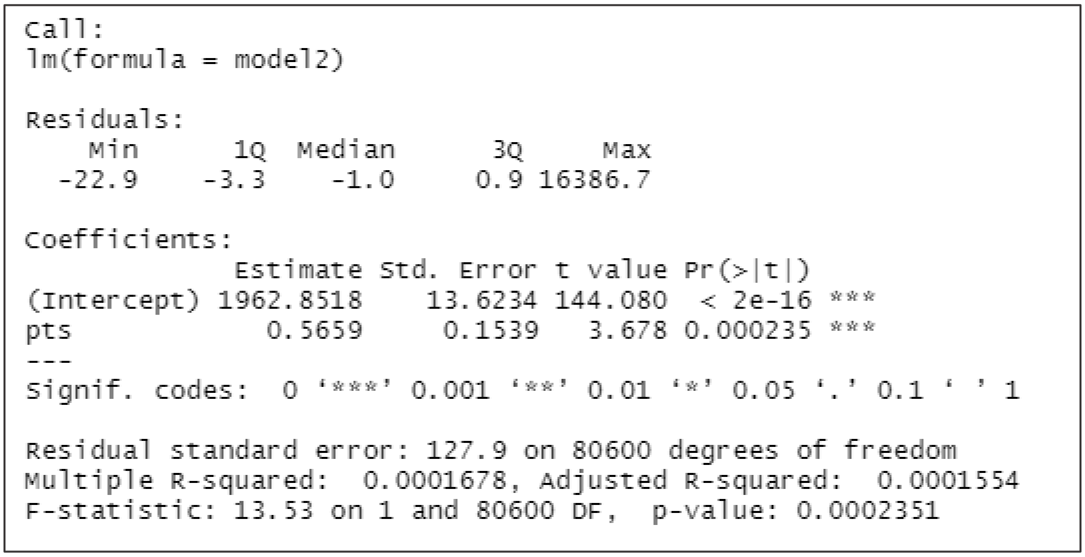
```
Call:
lm(formula = model)

Residuals:
    Min      1Q  Median      3Q     Max
-38.300 -10.870  -3.549   7.121  86.881

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -294.85732    1.69055  -174.4   <2e-16 ***
pts            3.67872    0.01909   192.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.87 on 80600 degrees of freedom
Multiple R-squared:  0.3153,	Adjusted R-squared:  0.3153
F-statistic: 3.712e+04 on 1 and 80600 DF,  p-value: < 2.2e-16
```

We can see that for every 1$ increase in price we could determine that it would be an increase of 3.67 points. However, we can also see that the margin of error is 15.87 points, which when dealing with a standard point range of 78-100 is problematic. We can see the visualization of this in the following graph. While there is a correlation and the model may point to an increase in points with an increase in price, the margin for error is too high for this to be used on a practical scale.

Points ~ Price – Linear Model



The second model that was run was based on the “Year” value.



```
Call:
lm(formula = model2)

Residuals:
    Min      1Q  Median      3Q     Max 
  -22.9    -3.3    -1.0     0.9 16386.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1962.8518    13.6234 144.080  < 2e-16 ***
pts            0.5659     0.1539   3.678 0.000235 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127.9 on 80600 degrees of freedom
Multiple R-squared:  0.0001678, Adjusted R-squared:  0.0001554 
F-statistic: 13.53 on 1 and 80600 DF,  p-value: 0.0002351
```

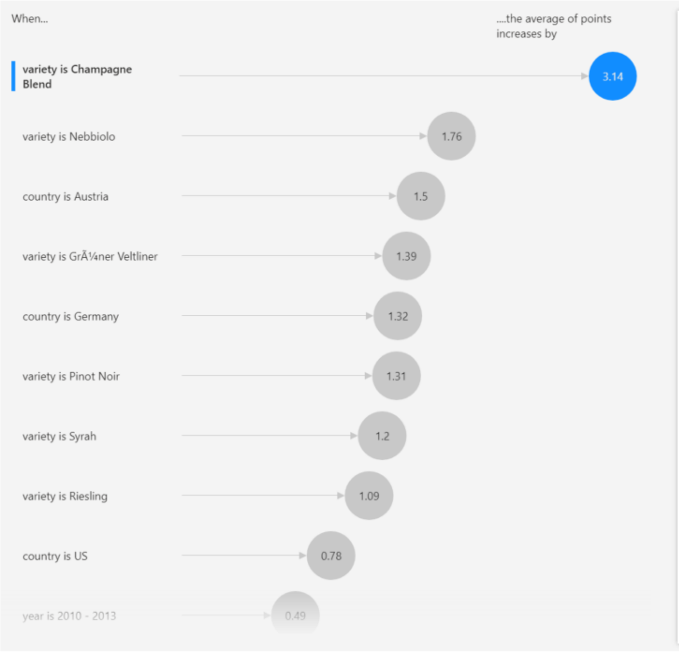
We can see that for every year decrease from 1990 we could determine that it would be an increase of 0.59 points. However, we can also see that the margin of error is 127.9 points which is an unacceptable level. The correlation value is also < 0.01 which does not show a strong correlation.

Categorical Classification for Increases

In addition to the decision tree, a key influencer algorithm was run through PowerBI and delivered some interesting outcomes for key influencers on Points. Here we can see several key variable values lead to an increasing in points. The variable with the highest influence on points is when the variety of the wine = Champagne Blend.

We also see that “Young” wines also lead to a 0.49-point increase which is contrary to the popular thought that older wines are reviewed higher.

There are also certain countries which influence the Points received for the wine including the US, Germany and Austria which can add from 0.78 points to 1.5 points.



CONCLUSION

Through our analysis we discovered that Austria has the highest average of points earned for country and that Italy had the highest average of price per bottle amongst the refined dataset. We also discovered that linear regression is difficult to run on this dataset because it is clear that we cannot express points as an equation of year and price. Classification models do display information around certain categories including variety and country.

References

[1] Cholette, S. (2006). ANALYZING THE US RETAIL WINE MARKET USING PRICE AND CONSUMER SEGMENTATION MODELS (REFEREED) . Retrieved June 16, 2020, from http://online.sfsu.edu/cholette/public_research/Sonoma-Segmentation.pdf

[2] Puckette, M. (2020, April 17). Reality of Wine Prices (What You Get For What You Spend). Retrieved June 16, 2020, from https://winefolly.com/lifestyle/reality-of-wine-prices-what-you-get-for-what-you-spend/

[3] Svitil, K. (2008, January). Wine Study Shows Price Influences Perception. Retrieved June 16, 2020, from https://www.caltech.edu/about/news/wine-study-shows-price-influences-perception-1374

[4] https://winefolly.com/lifestyle/reality-of-wine-prices-what-you-get-for-what-you-spend/