Yasmine Aitouny
Daniel Carmona
Terisha Kolenchery
05/10/2020

Wine Not - Exploring Wine Prices and Varieties

Background:

The goal of this project was to reveal whether the wine prices are related in any way to the wine popularity and type of wine. The Wine Reviews CSV is the dataset used in this project provided by Kaggle along with the Google GMaps API. The CSV consists of approximately 150k entries and 11 fields containing information on wine points, price, variety, the location where the wine is from and the winery where the wine was made. The data contains 619 varieties of wine from 46 different countries around the world.

In this report we will discuss how our group approached the following questions:

- Is there a relationship between wine price and popularity?
- Is there a relationship between type of wine and price?

Our methodology began with cleaning the dataset mentioned above. Through our data analysis and exploration, we noticed general trends in terms of which countries produce the most wine and what types of wine are most popular. We ran two-sample t-tests to determine if there was a statistically significant difference between average prices and average points awarded for red and white wines. We then conducted a simple linear regression analysis and a multivariate regression analysis to determine the correlation between price, country of origin, points, and wine type.

Key Takeaways:

Price and Points - there is a moderate, positive correlation between price and points for both red and white wines. The r-values for the simple linear regression for red and white wine were .539 and .536, respectively.

Price and Type of Wine - there is a statistically significant difference between the price of red wine and white wine. On average, red wine is higher priced than white wine. This difference

makes sense since the process to make red wine is more labour intensive and therefore would pass through costs (Zhang and Rosentrater).

Points and Type of Wine - there is a statistically significant difference between the points assigned to red and white wines. The difference in average points is much smaller than the difference in price - this is likely due to the fact that there were lower and upper bounds on the amount of points any given wine could earn, whereas there is theoretically no upper bound on the pricing of a wine.

Description of Data Sources/ Cleaning:

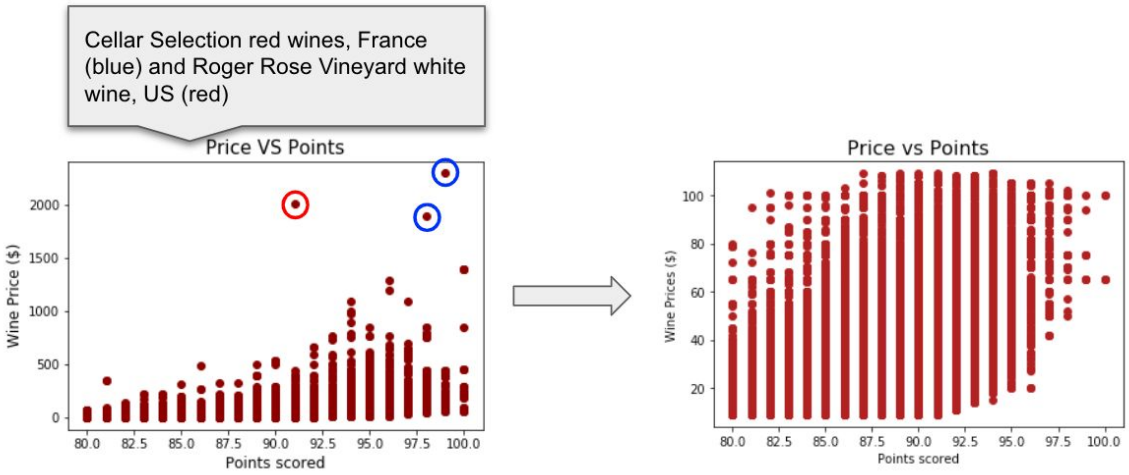
The raw data was read into Jupyter Notebook and we looked at the columns to see if there were any rows with missing fields. We decided to drop rows with NaN values for the columns relating to country and price since price was the independent variable we were trying to predict and country information was vital to our heatmap. We then looked at the unique types of wine varieties there were and exported the value counts to a separate CSV. We then manually searched and entered in whether the variety was a red or white wine for any variety that had 30 or more wines in the dataset.

```
#Outliers removed df
priceZ = wine_df[~((wine_df < (Q1 - 1.5 * IQR)) |(wine_df > (Q3 + 1.5 * IQR))).any(axis=1)]

print(f'Total number of rows pre-cleaning: {len(df)}')
print(f'Total number of rows post-cleaning: {len(priceZ)}')

priceZ.head(5)
```

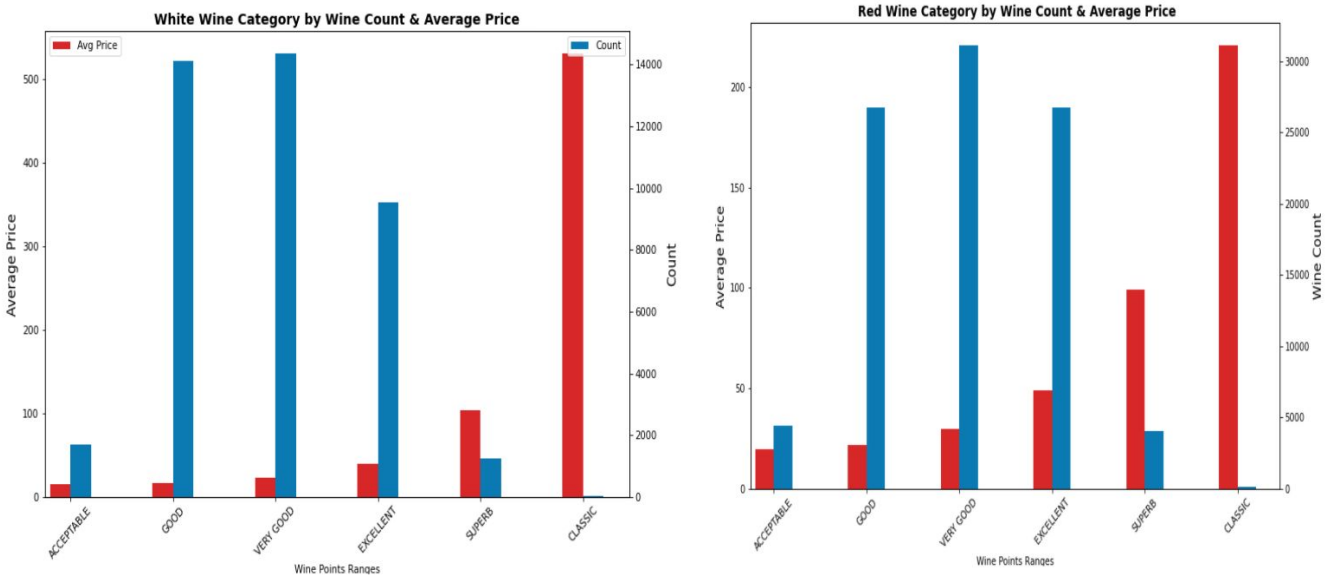
```
Total number of rows pre-cleaning: 150930
Total number of rows post-cleaning: 133762
```

	country	description	designation	points	price	province	region_1	region_2	variety	winery	Counts	Red?	Still_Red
0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz	12671.0	True	1
1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	Toro	Tinta de Toro	Bodega Carmen Rodríguez	221.0	True	1

Later on, to treat outliers, we found the top 2.5% and bottom 2.5% quantile ranges and removed them as outliers. Below is a comparison of the data before and after trimming. Trimming helped make the data more uniform in its spread.

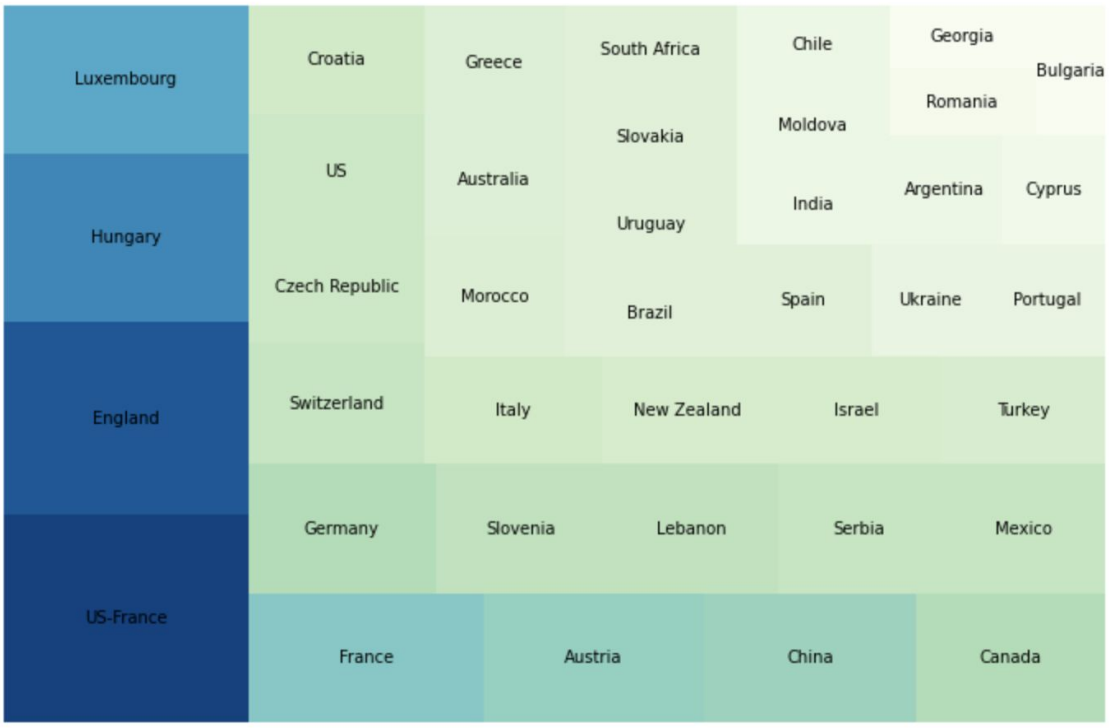


Additionally, we binned points data into categories based on Wine Enthusiast points scale (80-100) and category (Acceptable to Classic). We then separated out the data set into two smaller sets with solely red wines and solely white wines. An example visualization with binning is displayed below.

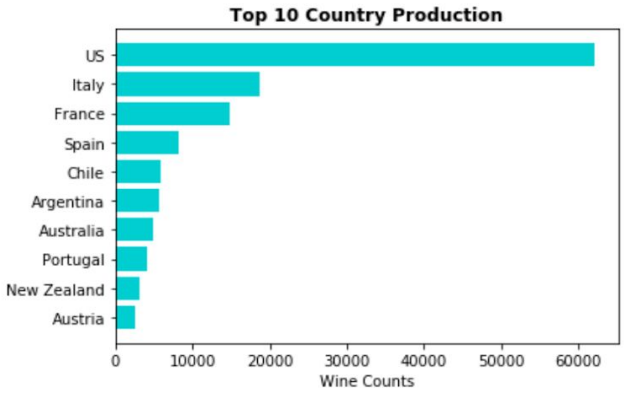


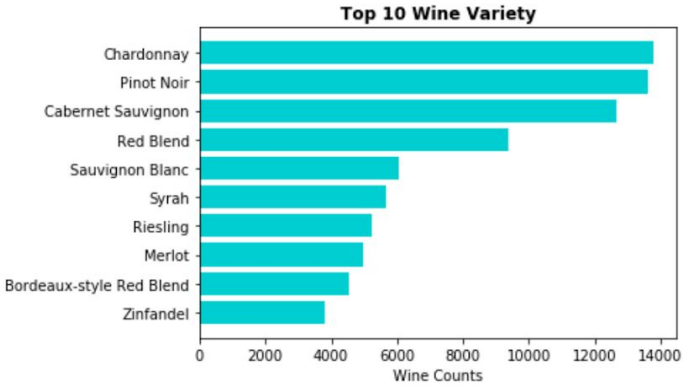
<u>Data Exploration:</u>

In our data exploration, we initially looked at what countries had the highest average prices and points. Below is one of the visualizations we created for average white wine price per country - for additional visualizations, see Appendix B.



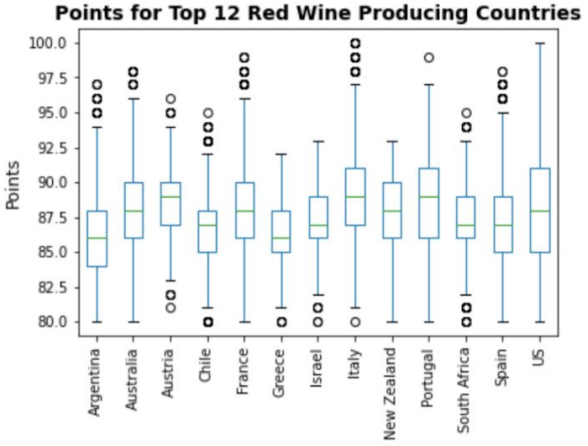
As we looked at the value counts for countries, we quickly realized that simply looking at the average price or points across all countries wasn't holistic because not all countries produced the same amount of wines. In the visualization, "US-France" (a blended wine with grapes from both countries) only had one wine that was highly priced. In order to better understand the dataset, we generated histograms to show the top wine producing countries and varieties. We decided on these cutoff points because the next highest producing country and variety was significantly lower.

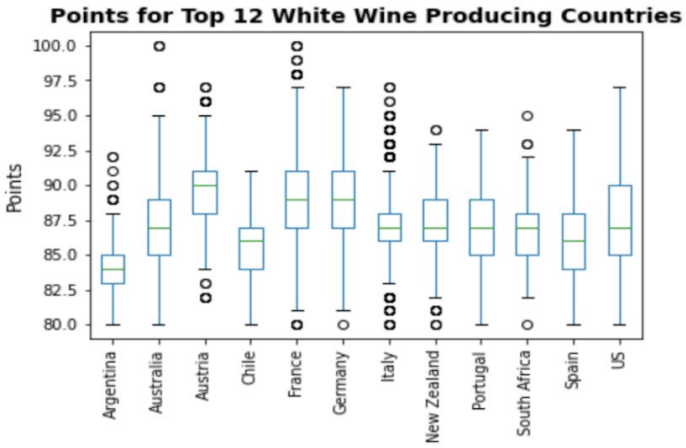




After we decided which countries to look at for our preliminary exploration, we looked at the spread of points for red wine and white wine. As seen below, Italy has the highest number of

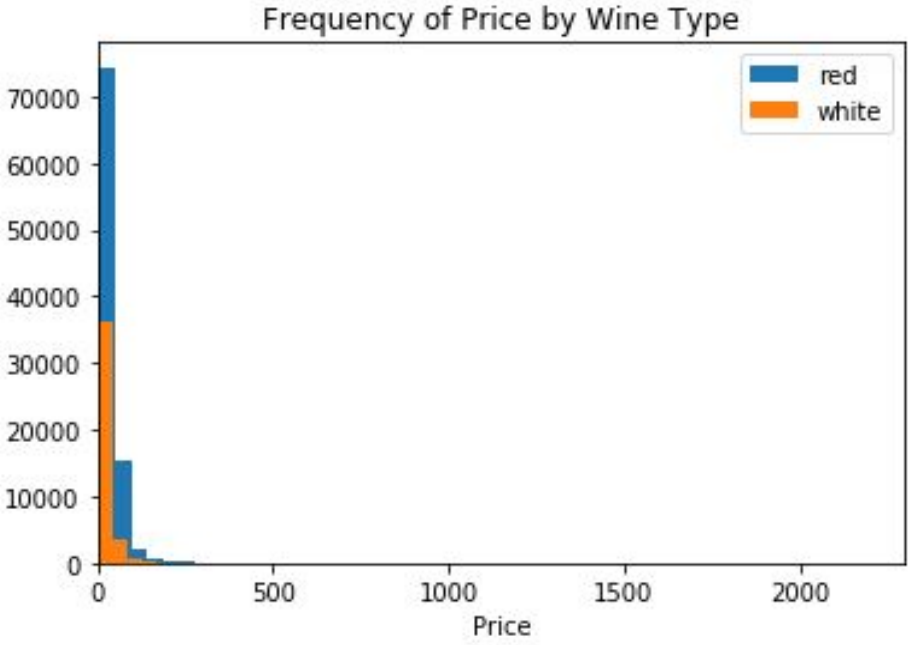
red wine points, while the United States has the largest spread of wine points for red wine. For white wine, Austria has the highest average points awarded.





We also looked at the spread of price data, and quickly realized that the spread of the data was much larger than points since Wine Enthusiast has bounds on the points system whereas there is no upper bound on the price of a wine.

For our data analysis, we looked to find correlations with a focus on price and what factors into the price of wine in general. We created various graphs and charts for both red and white wines based on: price vs points, country vs price, and country vs points. The extremity of the disparity can be much better seen when looking at all wines and not the top 12 countries.



Finally, our data exploration included heatmaps of the production density of different countries for red and white wines. The maps confirmed what we expected - most production is centered in the United States and Europe and there's significantly more red wine production than there is white wine production. Though listed, many of the provinces did not exist as geographical coordinates and were located via Google search and added manually into the dataset. Below is the map for red wine, see Appendix B for the white wine production heatmap.

<u>Data Analysis:</u>

Once our group had a more descriptive context, we then turned to answering our two main questions.

Relationship Between Price and Type of Wine

In order to answer this question, our group decided to run two-sample t-tests on the data. It is important to note that another test might have technically been more appropriate given the skew of the price data and the fact that we couldn't be sure that the sample was random. Alternatively, we could have taken random samples from the red and white wine subsets. However, we decided to stick with the t-test because that was the most appropriate tool we had for this analysis.

Our hypotheses and conclusions for the two tests were as follows:

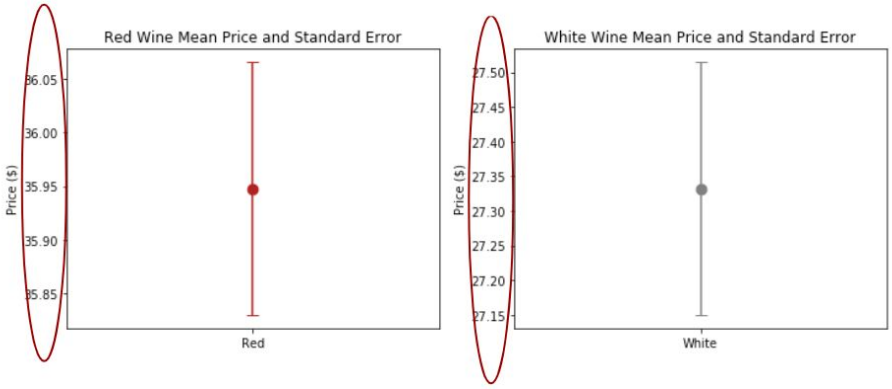
- Red vs White Wine - Price

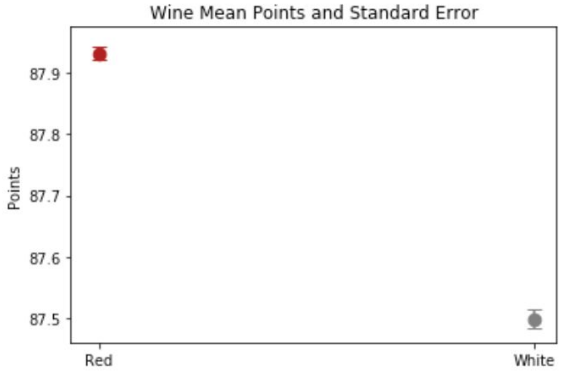
    - Null Hypothesis: There is no difference in average price between red and white wines.
    - Alternate Hypothesis: There is a difference in the average price of red and white wines.
    - Conclusion: At a 5% significance level, we reject the null hypothesis that there is no difference in average price between red and white wines since the p-value is 0.00.

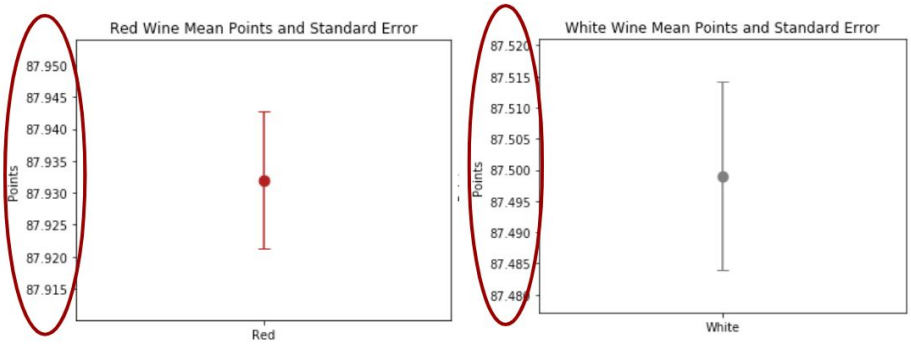
- Red vs White Wine - Points
    - Null Hypothesis: There is no difference in average points awarded between red and white wines.
    - Alternate Hypothesis: There is a difference in the average points awarded for red and white wines.
    - Conclusion: At a 5% significance level, we reject the null hypothesis that there is no difference in average points awarded between red and white wines since the p-value is 0.00.

The graphs below show the average prices and points for red and white wines along with the standard error bars. The lack of overlap between the error bars is a visual representation of the statistically significant difference in means.


Wine Mean Price and Standard Error
Price ($)
36
34
32
30
28
Red
White


Red Wine Mean Price and Standard Error
Price ($)
36.05
36.00
35.95
35.90
35.85
Red
White Wine Mean Price and Standard Error
Price ($)
27.50
27.45
27.40
27.35
27.30
27.25
27.20
27.15
White


Wine Mean Points and Standard Error
Points
87.9
87.8
87.7
87.6
87.5
Red
White


Red Wine Mean Points and Standard Error
Points
87.950
87.945
87.940
87.935
87.930
87.925
87.920
87.915
Red
White Wine Mean Points and Standard Error
Points
87.520
87.515
87.510
87.505
87.500
87.495
87.490
87.485
White

Our group looked at possible explanations for the disparity between the prices of red and white wines. We found that the demand for red wine changed drastically in the early 2010's when Chinese consumers entered the wine market as cultural norms towards foreign wines shifted (Goldstein). In particular, upper echelons of Chinese society view vintage red wines as a luxury item and therefore are willing to pay higher prices for bottles (Wang).

Additionally, red wines are more labor intensive - red wine ferments the skins and seeds, which are then extracted (Zhang and Rosentrater). This process leaves red wine less susceptible to oxidation, which then requires secondary fermentation (Zhang and Rosentrater). White wine is fermented by yeast, chilled, and then filtered, which is a much quicker process (Zhang and Rosentrater).

Relationship Between Price and Points

In order to answer this question our group first trimmed the data as described in the data-cleaning section. The trimming changed the range of the data and increased the correlation between price and points, as seen below.

Pre-Trimming Stats and Correlations

	points	price	Counts	Still_Red
count	134289.000000	134289.000000	134289.000000	134289.000000
mean	87.799931	33.319438	6340.900990	0.694942
std	3.225317	36.561578	5153.366014	0.460434
min	80.000000	4.000000	30.000000	0.000000
25%	86.000000	16.000000	1455.000000	0.000000
50%	88.000000	24.000000	4987.000000	1.000000
75%	90.000000	40.000000	12671.000000	1.000000
max	100.000000	2300.000000	13775.000000	1.000000

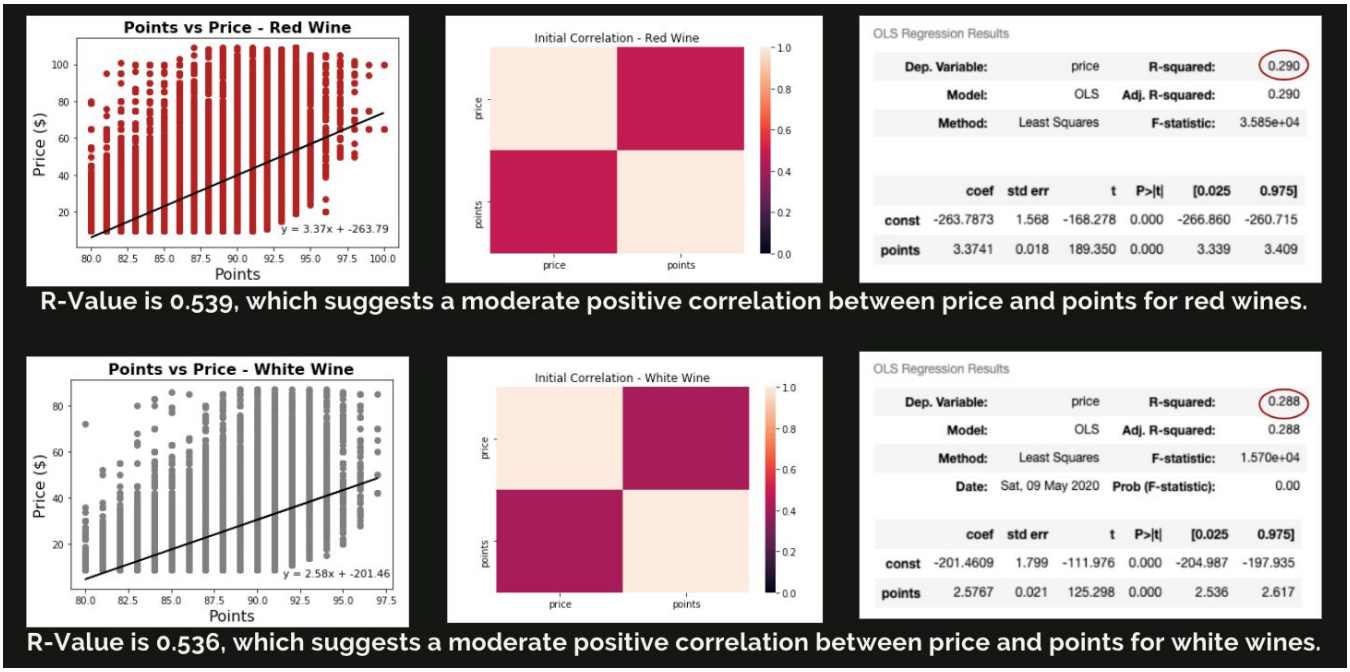
	points	price	Counts	Still_Red
points	1.000000	0.459624	0.072734	0.061808
price	0.459624	1.000000	0.107070	0.108498
Counts	0.072734	0.107070	1.000000	-0.028155
Still_Red	0.061808	0.108498	-0.028155	1.000000

Post- Trimming Stats and Correlations

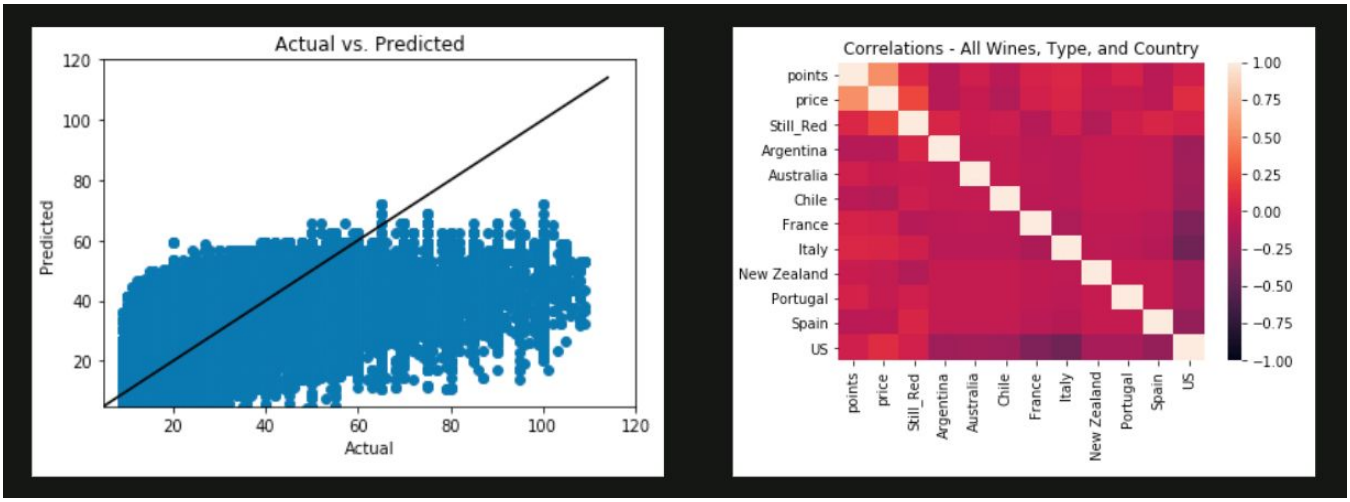
	points	price	Counts	Still_Red
count	126584.000000	126584.000000	126584.000000	126584.000000
mean	87.767245	30.078193	6363.028124	0.692979
std	3.093177	18.675083	5155.713547	0.461260
min	80.000000	9.000000	30.000000	0.000000
25%	86.000000	16.000000	1455.000000	0.000000
50%	88.000000	25.000000	4987.000000	1.000000
75%	90.000000	39.000000	12671.000000	1.000000
max	100.000000	109.000000	13775.000000	1.000000

	points	price	Counts	Still_Red
points	1.000000	0.535937	0.066897	0.070902
price	0.535937	1.000000	0.184860	0.222282
Counts	0.066897	0.184860	1.000000	-0.025279
Still_Red	0.070902	0.222282	-0.025279	1.000000

After trimming the dataset we ran a regression model to predict the price of a wine given the Wine Enthusiast rating. The graph, line of best fit, and r-squared values for red and white wine are displayed below.



Since the r-squared values for red and white wine were approximately the same and the correlations were also close we decided to run a multivariate regression analysis on the combined trimmed data set. In order to do this analysis we first dummified the top nine countries. The actual vs predicted graph and correlation heatmap is reproduced below, as well as the OLS regression results. Note that r-squared value increased over the simple model and each of the independent variables remains statistically significant.

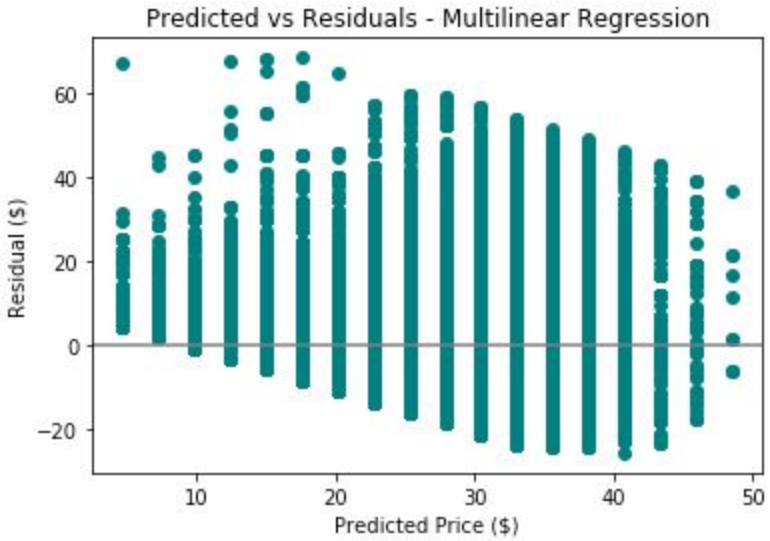


OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.350
Model:	OLS	**Adj. R-squared:**	0.350
Method:	Least Squares	**F-statistic:**	6336.
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	0.00
Time:	13:25:55	**Log-Likelihood:**	-4.8716e+05
No. Observations:	117681	**AIC:**	9.744e+05
Df Residuals:	117670	**BIC:**	9.745e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-221.2869	1.141	-193.906	0.000	-223.524	-219.050
points	3.0485	0.015	209.871	0.000	3.020	3.077
Still_Red	8.1038	0.100	81.344	0.000	7.909	8.299
Argentina	-27.9245	0.228	-122.213	0.000	-28.372	-27.477
Australia	-24.6496	0.249	-98.839	0.000	-25.138	-24.161
Chile	-28.6038	0.228	-125.435	0.000	-29.051	-28.157
France	-20.5781	0.192	-107.151	0.000	-20.955	-20.202
Italy	-20.5101	0.183	-112.063	0.000	-20.869	-20.151
New Zealand	-24.4948	0.286	-85.784	0.000	-25.054	-23.935
Portugal	-29.1171	0.280	-104.169	0.000	-29.665	-28.569
Spain	-25.5453	0.210	-121.641	0.000	-25.957	-25.134
US	-19.8635	0.154	-128.972	0.000	-20.165	-19.562

While the increase in r-squared values is encouraging, it's important to note the lack of linear trend in the actual vs predicted graph. This lack of trend is reinforced when looking at a plot of the predicted values versus the residual values. The data is not spread with uniform distance from the y = 0 line and isn't random. The skew of points above y = 0 suggests that our models often underestimate wine prices.



Limitations of the Dataset / Recommendations:

- Accuracy of wine type sorting - the sorting was done manually and in general batches, which lent to a lack of nuance in the sorting. Take for example, the category sparkling wine. Ultimately, wines under this variety were categorized as white wines, but red sparkling wines do exist and might exist within the dataset. This lack of nuance could be addressed by going through each individual row and looking up the specific wine, but given time constraints this approach didn't make sense.
- Missed variables - a couple of data points that could add accuracy to our model would be aging length of the wine and climate data about rainfall during peak growing seasons. These data points weren't easily available and so fell out of the scope for this analysis.
- Impact of auction-based price setting mechanisms - Bordeaux wines are often sold via auction mechanism in a futures market (Zylberberg and Harans). Having an auction to set prices generally drives prices up, either artificially or not. It's possible that there are more outliers in our dataset than originally thought and that might explain why we severely underestimated wine prices.
- Further statistical analysis - Given more time, it would have been interesting to complete ANOVA and T-Tests on the varieties and countries to figure out if there is a statistical difference in means.

Appendix A: Works Cited

Goldstein, Jacob. "The Bordeaux Bubble." *NPR*, NPR, 4 Aug. 2011, www.npr.org/sections/money/2011/08/04/138981124/the-bordeaux-bubble.

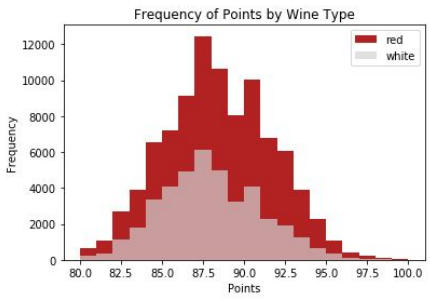
Wang, Michael. "A COMPARATIVE STUDY OF WINE CONSUMER BEHAVIOR IN CHINA AND THE UNITED STATES: DOES CULTURE AFFECT CONSUMER BEHAVIOR? ." *A COMPARATIVE STUDY OF WINE CONSUMER BEHAVIOR IN CHINA AND THE UNITED STATES: DOES CULTURE AFFECT CONSUMER BEHAVIOR?* , *California State Polytechnic University, Pomona*, California State Polytechnic University, Pomona, 2017, dspace.calstate.edu/bitstream/handle/10211.3/189331/WangMichael_Project2017.pdf?sequence=3.

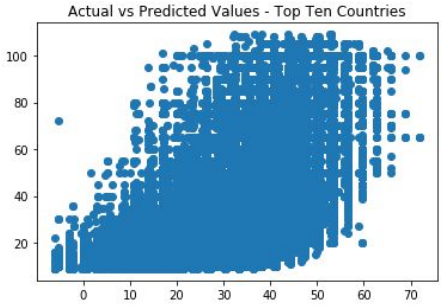
Zhang, Congmu, and Kurt A. Rosentrater. "Estimating Economic and Environmental Impacts of Red-Wine-Making Processes in the USA." *Fermentation*, vol. 5, no. 3, 2019, p. 77., doi:10.3390/fermentation5030077.

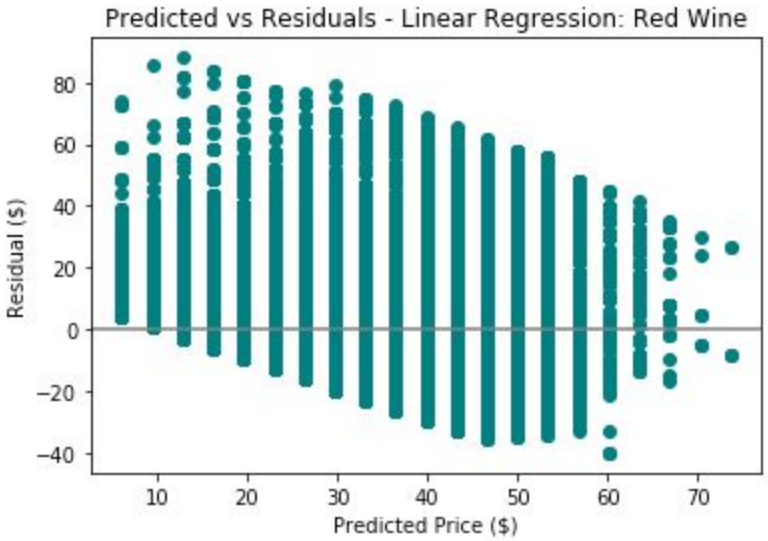
Zylberberg, Shawn, and Julie Harans. "Buying Futures." *Wine Spectator*, 27 Mar. 2019, www.winespectator.com/articles/buying-futures-3495.

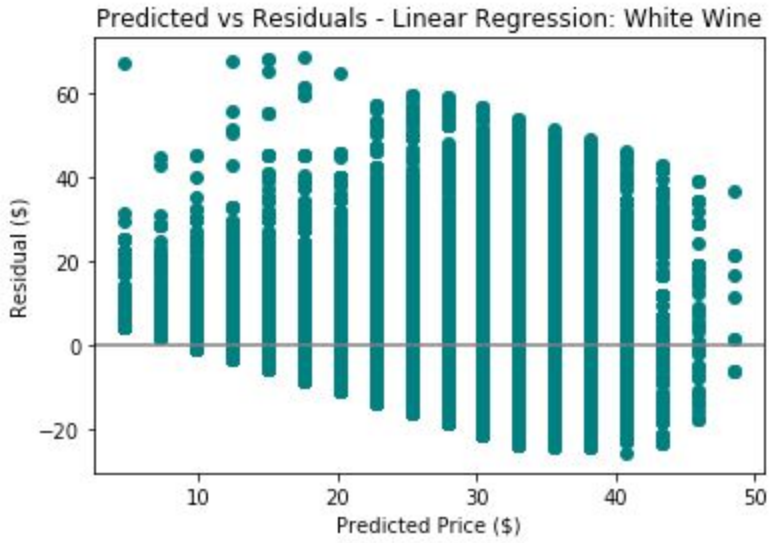
Appendix B: Misc Visualizations

Wine Points Ranges	Avg Price	Wine Count
ACCEPTABLE	19.883348	4432
GOOD	22.028680	26778
VERY GOOD	29.941696	31130
EXCELLENT	48.955550	26794
SUPERB	99.166625	4045
CLASSIC	220.888889	144



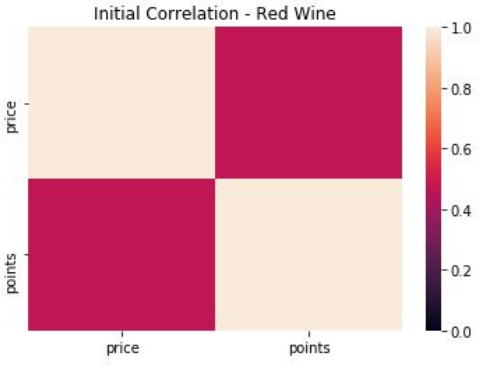


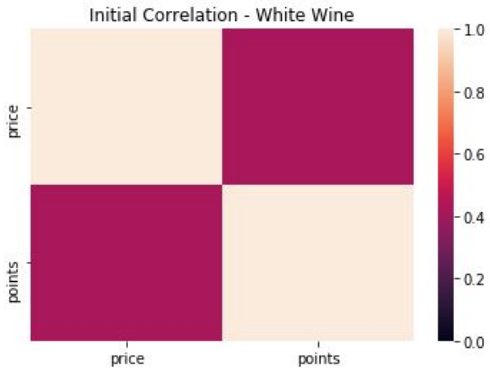




	price	points
price	1.000000	0.474773
points	0.474773	1.000000

	price	points
price	1.000000	0.414576
points	0.414576	1.000000





Red Wine Regression:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.290
Model:	OLS	**Adj. R-squared:**	0.290
Method:	Least Squares	**F-statistic:**	3.585e+04

	coef	std err	t	P>|t|	[0.025	0.975]
const	-263.7873	1.568	-168.278	0.000	-266.860	-260.715
points	3.3741	0.018	189.350	0.000	3.339	3.409

White Wine Regression:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.288
Model:	OLS	**Adj. R-squared:**	0.288
Method:	Least Squares	**F-statistic:**	1.570e+04
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	0.00

	coef	std err	t	P>|t|	[0.025	0.975]
const	-201.4609	1.799	-111.976	0.000	-204.987	-197.935
points	2.5767	0.021	125.298	0.000	2.536	2.617

Combined Wine Data

	points	price	Counts	Still_Red
count	126584.000000	126584.000000	126584.000000	126584.000000
mean	87.767245	30.078193	6363.028124	0.692979
std	3.093177	18.675083	5155.713547	0.461260
min	80.000000	9.000000	30.000000	0.000000
25%	86.000000	16.000000	1455.000000	0.000000
50%	88.000000	25.000000	4987.000000	1.000000
75%	90.000000	39.000000	12671.000000	1.000000
max	100.000000	109.000000	13775.000000	1.000000

	points	price	Counts	Still_Red
points	1.000000	0.535937	0.066897	0.070902
price	0.535937	1.000000	0.184860	0.222282
Counts	0.066897	0.184860	1.000000	-0.025279
Still_Red	0.070902	0.222282	-0.025279	1.000000

Average Price - Red Wine