MAY 11, 2020

WINE NOT?

Yasmine Aitouny, Daniel Carmona, Terisha Kolencherry

Executive Summary

What Goes in to Pricing a Wine?

QUESTIONS

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

EXPECTATIONS

- There is a relationship between price and points, but no relationship between type of wine and price.

RESULTS

- There was a positive correlation between price and points, but not as strong as anticipated.
- Also there is a statistical difference in price depending on type of wine.

Questions and Data

WHAT TYPES OF WINE VARIETIES EXIST?

- To find the answer, the data was parsed for keyword indicators (e.g. blanc, red-blend etc.) and narrowed down by wine color: red, white.
- The data was then compiled and added to the main dataset for further analysis.
- Having the wine type as quantifiable data allowed us to narrow the scope of our research.

WHICH WINE VARIETY IS MOST POPULAR BY LOCALE?

- For the popularity heatmaps, location coordinate information was necessary.
- Latitude and longitude information for province/state was not easily accessible via API call to GMaps.
- Missing data was added manually through web search.

Questions and Data

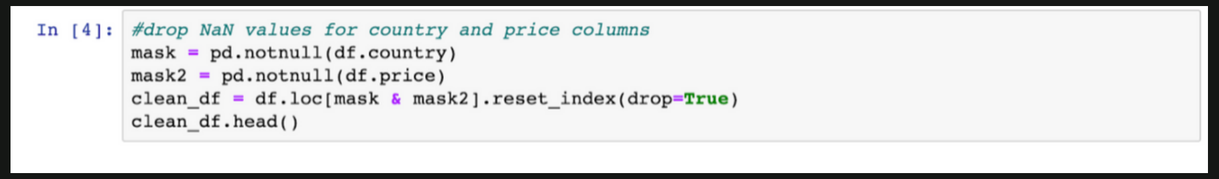
DATASET

- CSV titled Wine Reviews dated from 2017 that was derived from the wine review website, Wine Enthusiast.
- Wine Enthusiast asks reviewers to do price-blind taste tests.
- The dataset contains information on the countries of production, price, review points of red and white wine from different wineries around the world.
- Wine reviews points are categorized from 'acceptable' to 'classic'.
- Also used GMaps as described previously

Data Cleaning & Exploration

PROCESS OVERVIEW

- After importing all the necessary libraries, the first step in "sciencing the data" was to read-in the CSV file and clean the null values in the dataframe.
- Retrieved our value counts for quantifiable data, country counts, wine counts, price, region and variety.
- Filled in missing information for red and white wine labels and decimal coordinates.
- Created new CSV's with the added information and continued analysis.
- As a part of the analysis, we created bar graphs, histograms, scatter plots, correlation charts, boxplots using parsed data.
- Binned data according to Wine Enthusiast's ratings system and created visualizations.
- Created heatmaps of wine population by accessing Google API services.
- Ran statistical tests on the average prices and points for red and white wines.
- Trimmed data for regression analysis to look at a reasonable range of wine prices and then conducted regression analysis.

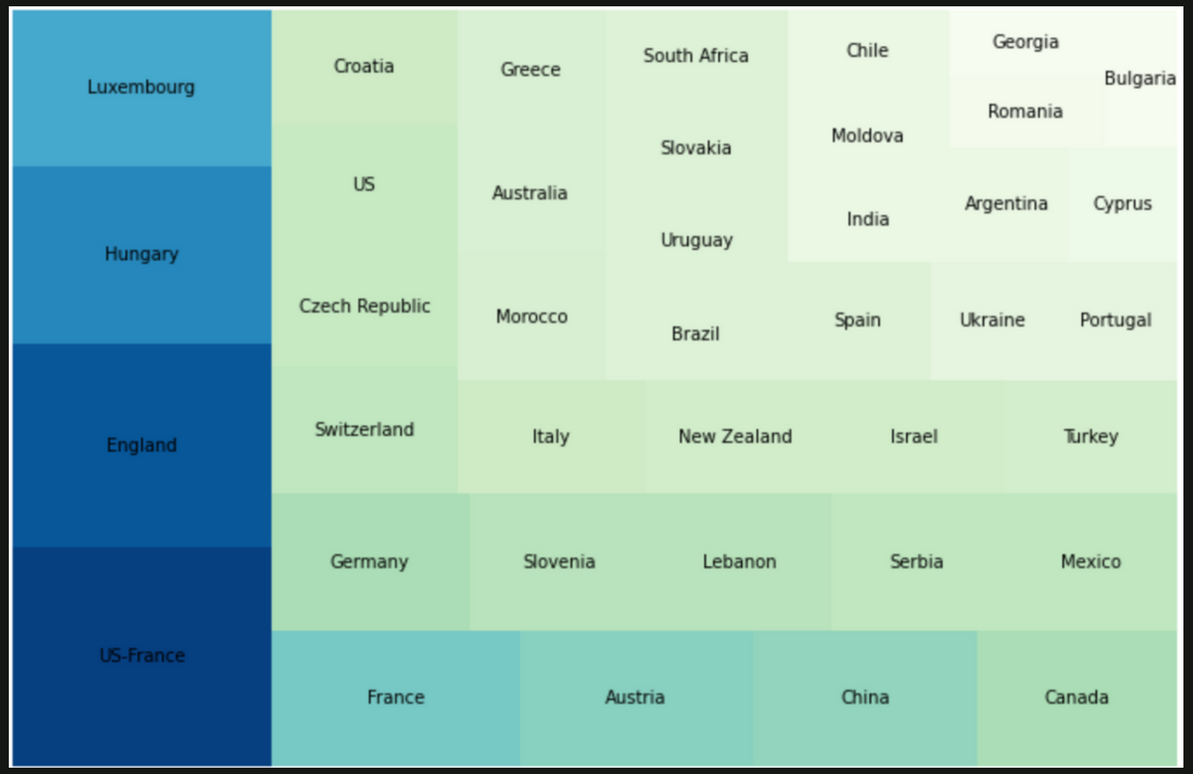

In [4]:
```
#drop NaN values for country and price columns
mask = pd.notnull(df.country)
mask2 = pd.notnull(df.price)
clean_df = df.loc[mask & mask2].reset_index(drop=True)
clean_df.head()
```

Data Cleaning & Exploration

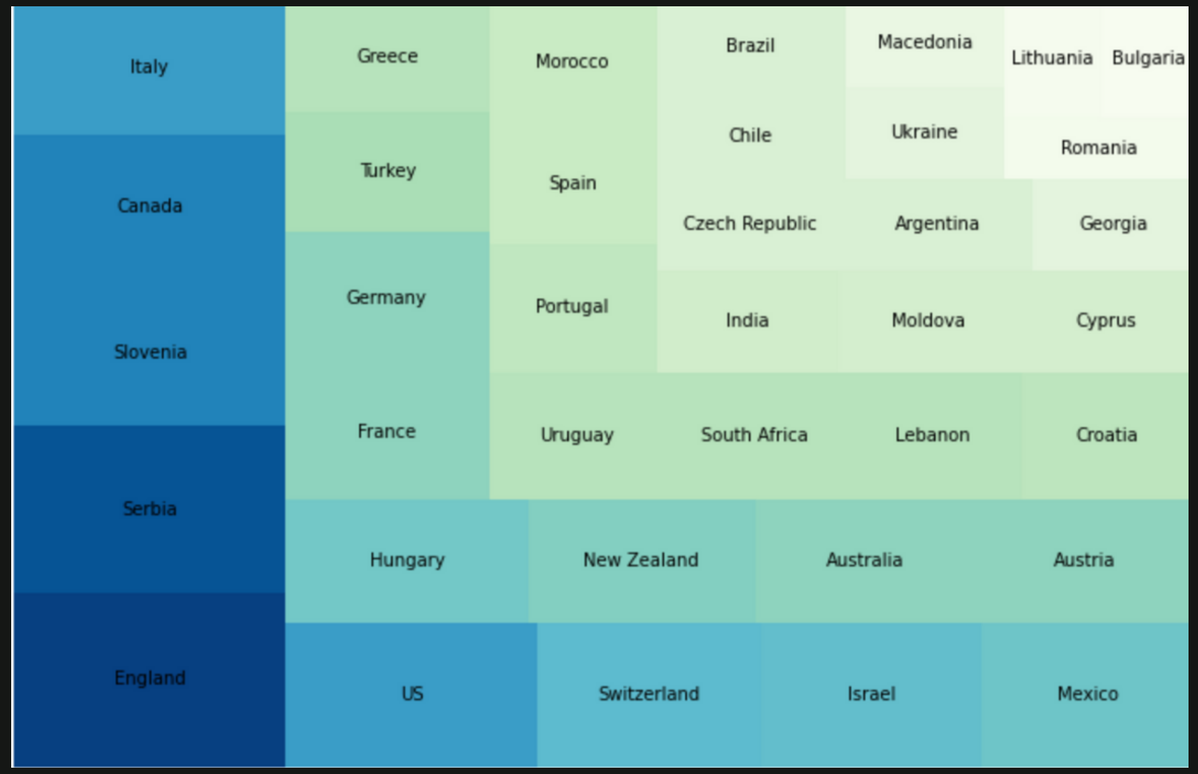
SPREAD OF DATA - PRICE

- Initially just looked at mean - but that doesn't work for this data set
- England looks like the highest priced red wine, but also has only five wines.

WHITE WINE AVERAGE PRICES



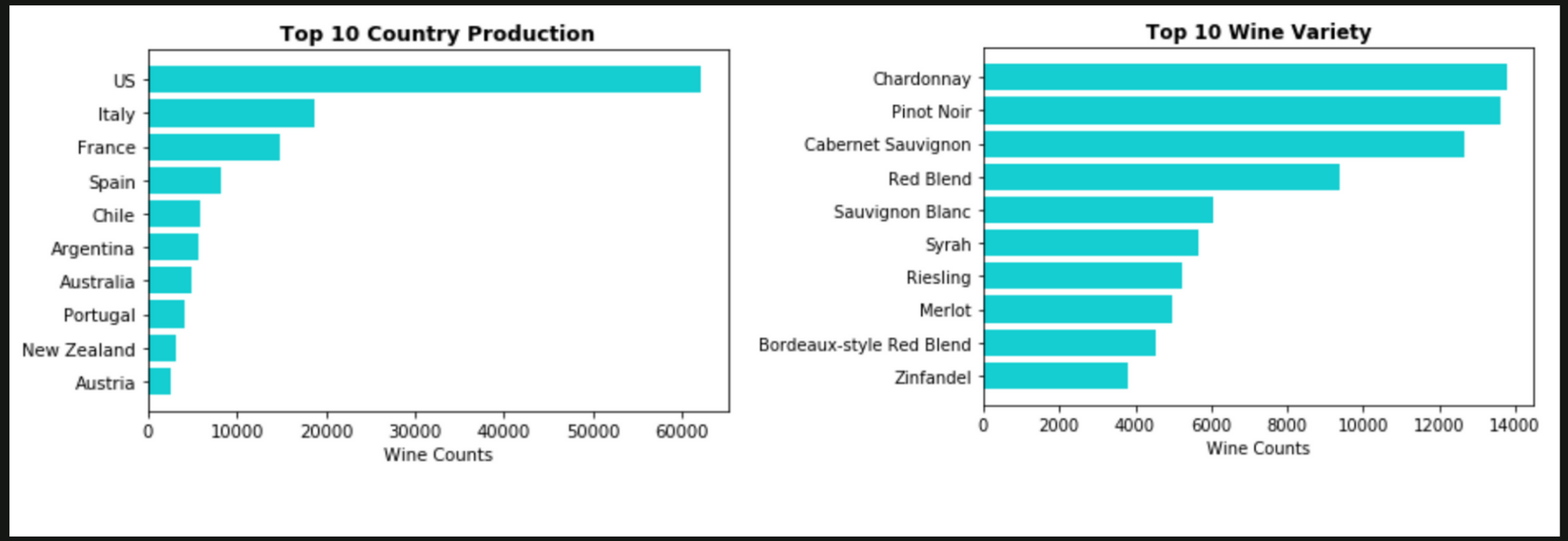
RED WINE AVERAGE PRICES



Data Cleaning & Exploration

TOP VARIETIES AND TOP COUNTRIES

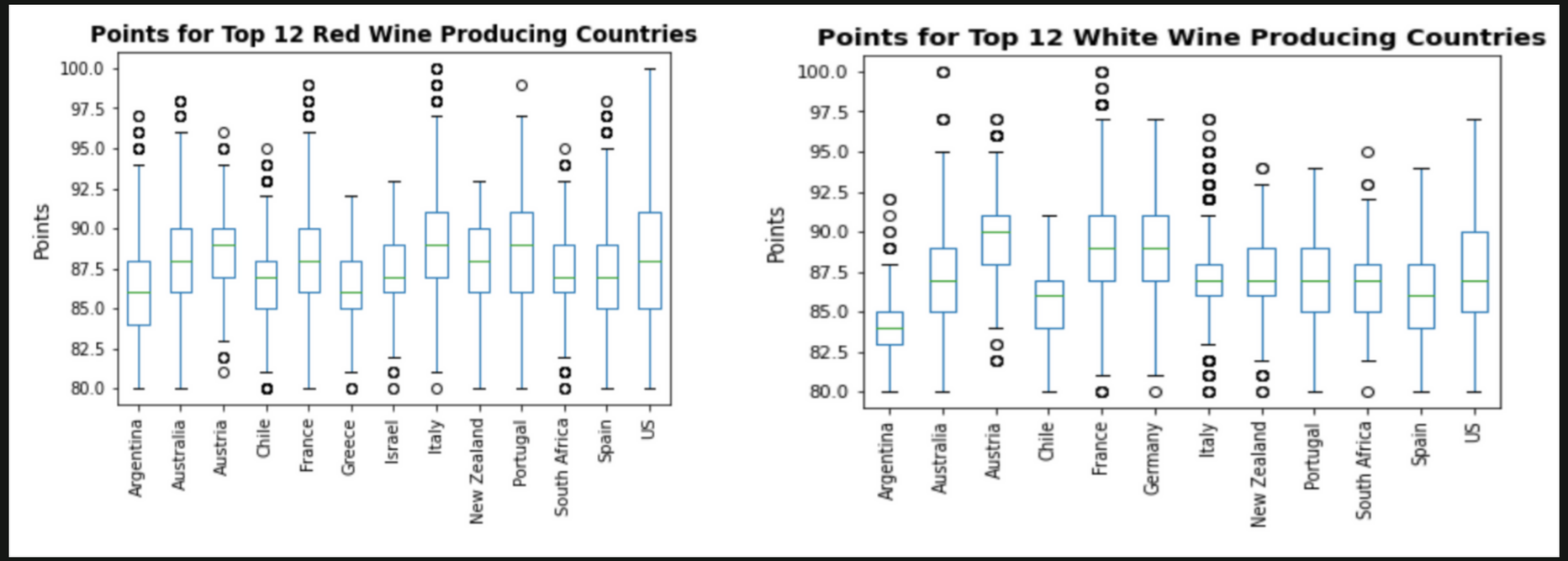
- In order to look at the spread, we calculated the ten highest wine producing by countries and varieties.
- The number one country is the US with more than 60,000 units, while the most popular wine is Chardonnay (No doubt!) with 14,000 count.



Data Cleaning & Exploration

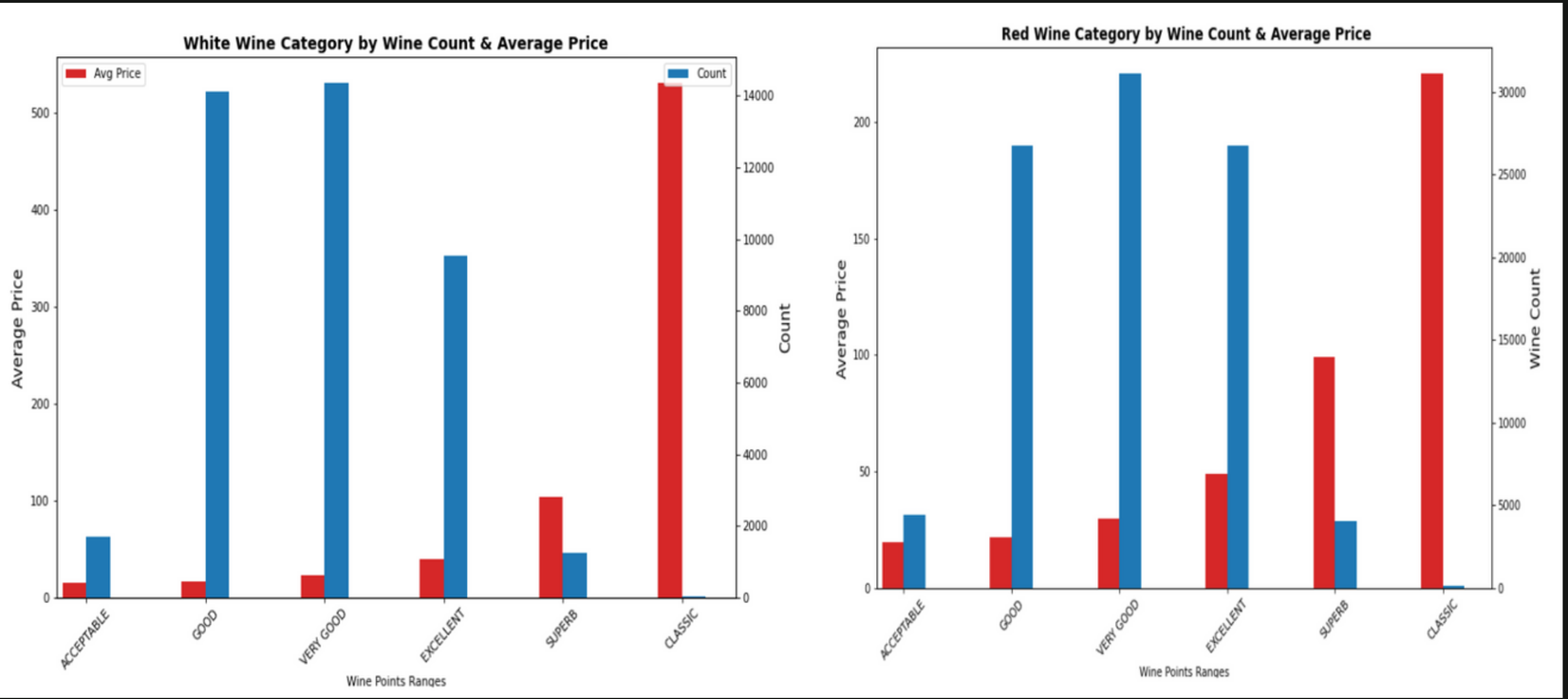
SPREAD OF DATA - POINTS

- We picked the top 12 countries because there was a large drop in counts for both white and red wines.
- For the red type Italy has the higher review points along with Austria.
- For the white type Austria has the highest average followed by France and Germany.



Data Cleaning & Exploration

BINNING



80-82 **Acceptable** Can be employed	83-86 **Good** Suitable for everyday consumption; often good value	87-89 **Very Good** Often good value; well recommended	90-93 **Excellent** Highly recommended	94-97 **Superb** A great achievement	98-100 **Classic** The pinnacle of quality

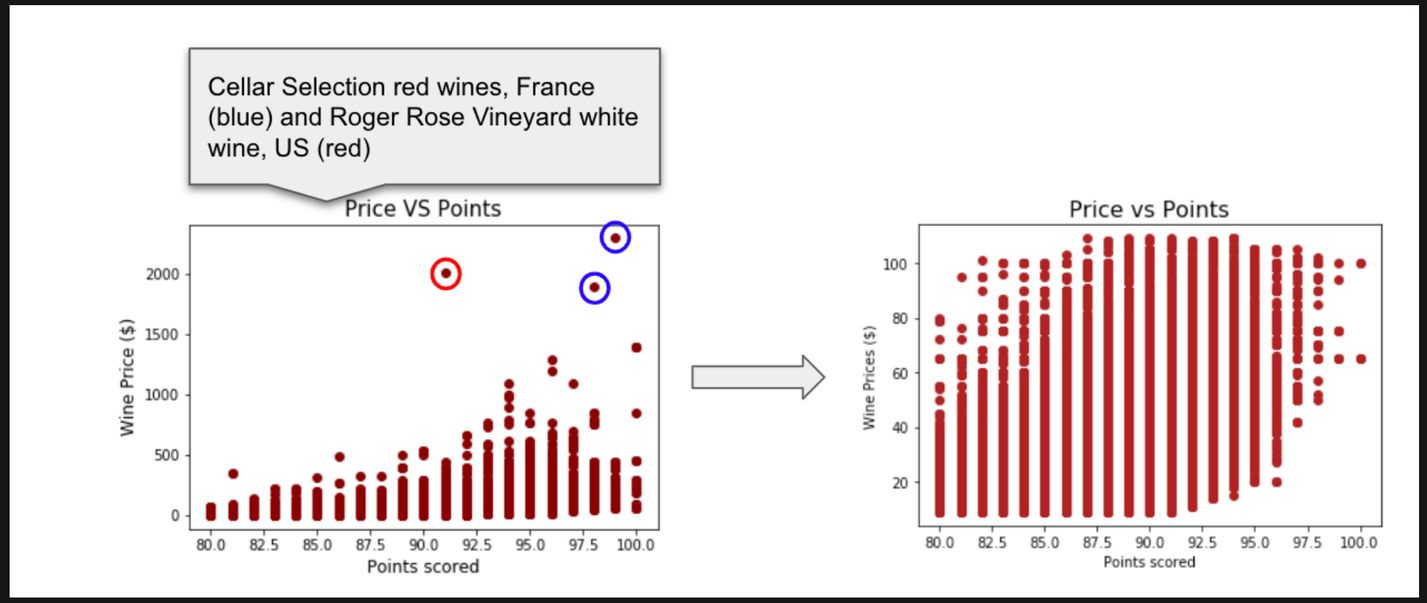
Wines are mostly symmetrically distributed across the ratings categories, with a slight skew right.

Bins were generated based on Wine Enthusiast's ratings system.

Data Cleaning & Exploration

TRIMMING

Attempted removal of top and bottom 2.5% down to 1% of data with skewed results. Ultimately, decided on removing approx. ~.00001 of the top priced wine as outliers.



Impact #1: Shifted the distribution of data points to a more compact distribution

Data Cleaning & Exploration

TRIMMING

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

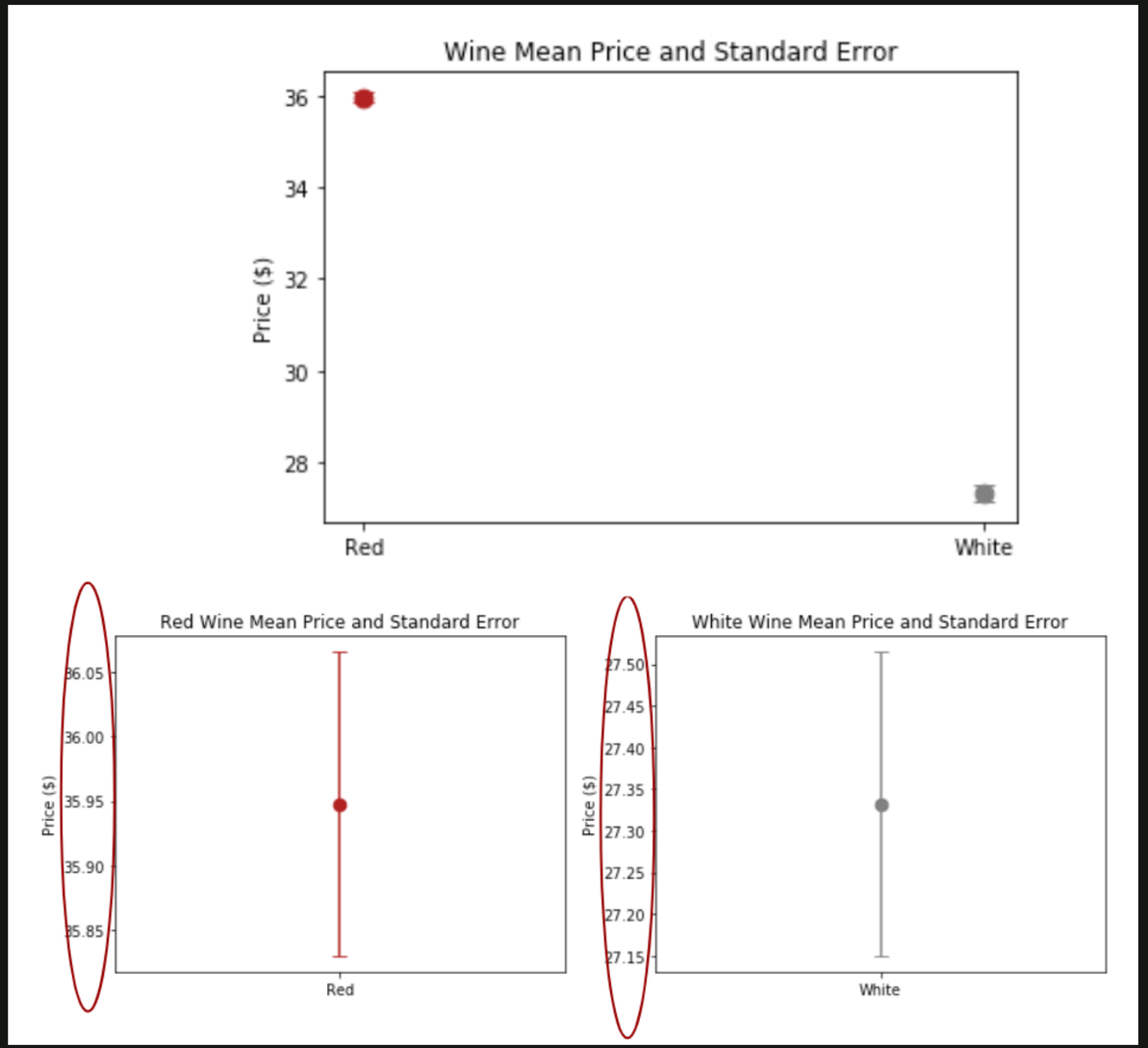
	points	price	Counts	Still_Red
points	1.000000	0.535937	0.066897	0.070902
price	0.535937	1.000000	0.184860	0.222282
Counts	0.066897	0.184860	1.000000	-0.025279
Still_Red	0.070902	0.222282	-0.025279	1.000000

Impact #2: Changed the range of data and correlation between price and points.

Statistical Analysis

RED VS WHITE WINE - PRICE

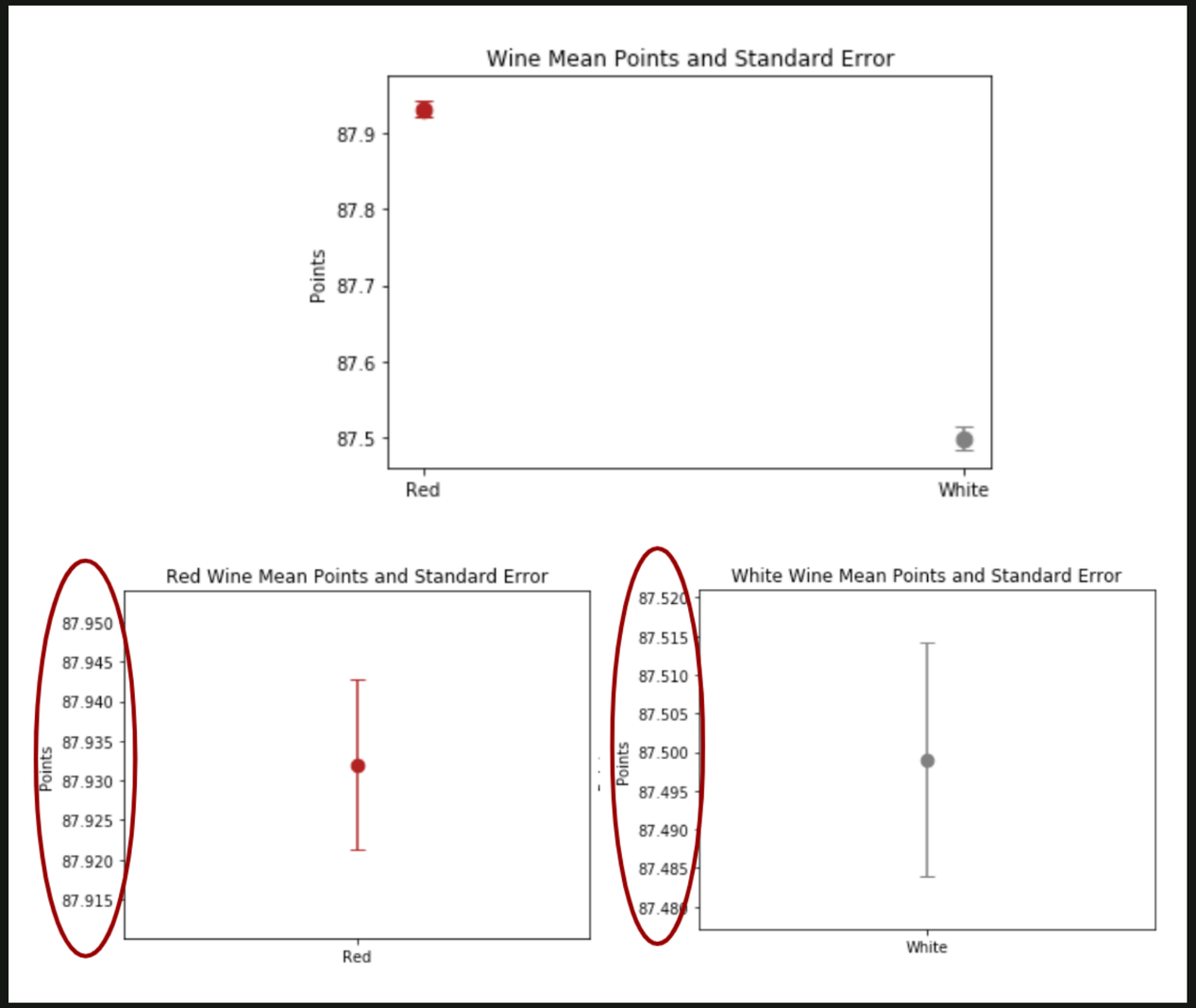
- We ran a two sample t-test comparing the average prices for red and white wines in the data set
- **Null Hypothesis:** There is no difference in average price between red and white wines.
- **Alternate Hypothesis:** There is a difference in the average price of red and white wines.
- **Conclusion:** At a 5% significance level, we reject the null hypothesis that there is no difference in the average prices of red and white wines since the p-value is 0.000.



Statistical Analysis

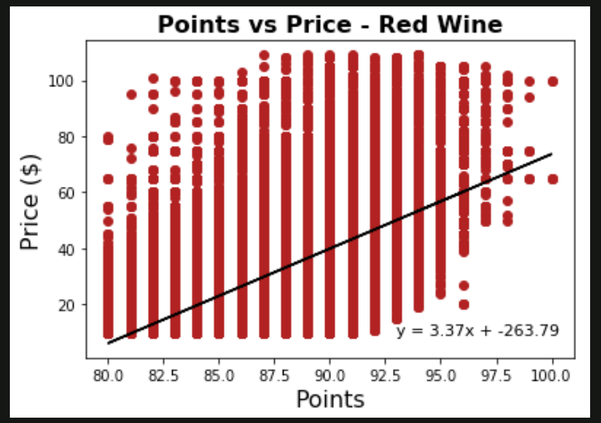
RED VS WHITE WINE - POINTS

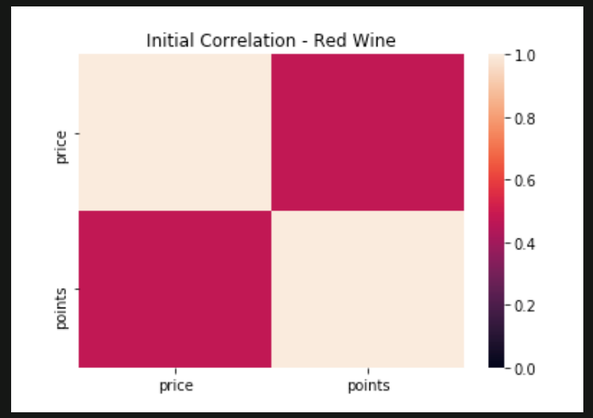
- We ran a two sample t-test comparing the average points for red and white wines in the data set
- **Null Hypothesis:** There is no difference in average point value between red and white wines.
- **Alternate Hypothesis:** There is a difference in the average point value of red and white wines.
- **Conclusion:** At a 5% significance level, we reject the null hypothesis that there is no difference in the average prices of red and white wines since the p-value is 0.000.



Regression Analysis

SIMPLE MODEL - PREDICT PRICE BASED ON POINTS



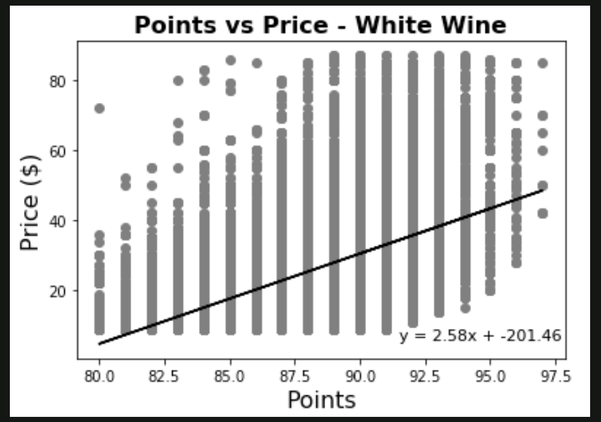


OLS Regression Results

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

R-Value is 0.539, which suggests a moderate positive correlation between price and points for red wines.



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

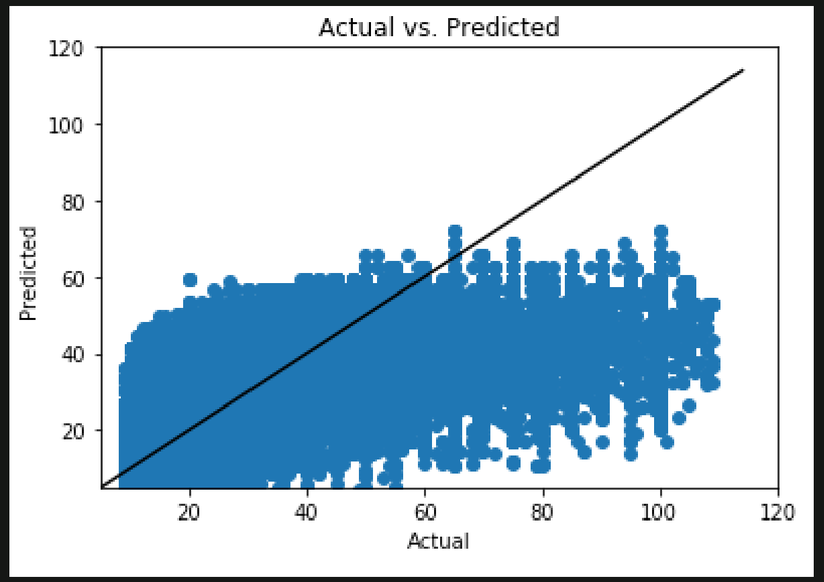
	coef	std err	t	P>|t|	[0.025	0.975]
const	-201.4609	1.799	-111.976	0.000	-204.987	-197.935
points	2.5767	0.021	125.298	0.000	2.536	2.617

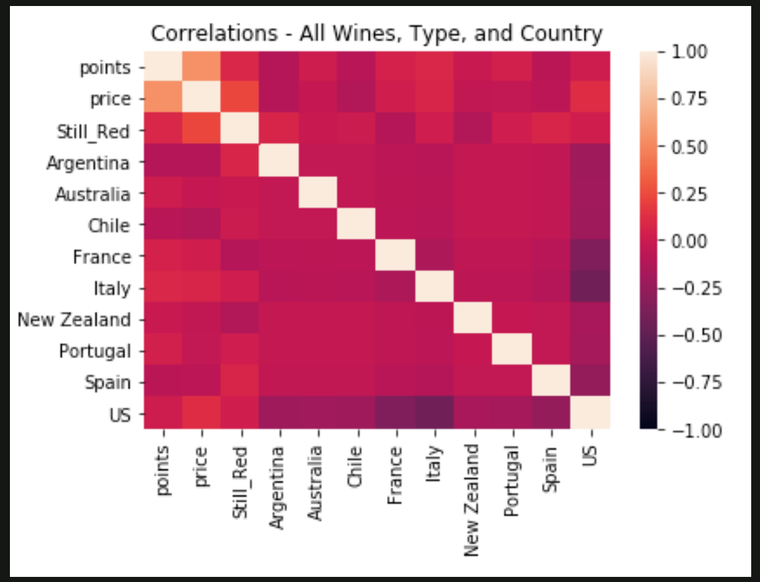
R-Value is 0.536, which suggests a moderate positive correlation between price and points for white wines.

Regression Analysis

MULTILINEAR MODEL - PREDICT PRICE BASED ON POINTS, TYPE, AND COUNTRY

Dummified country data - '1' if the wine is from X country, '0' if it isn't.





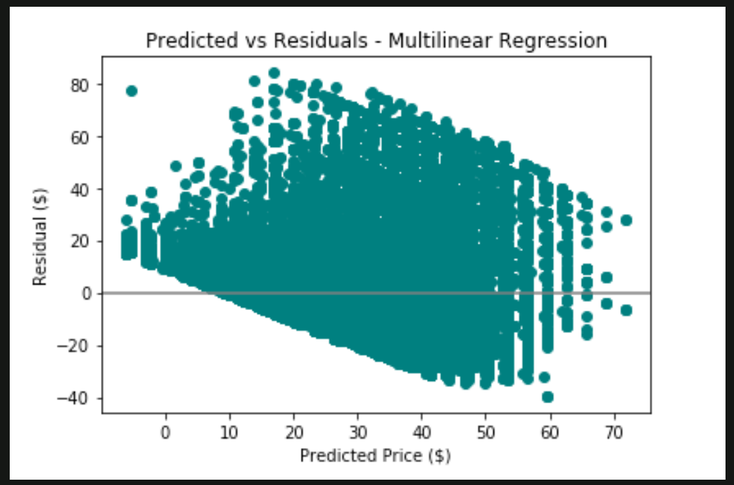
OLS Regression Results

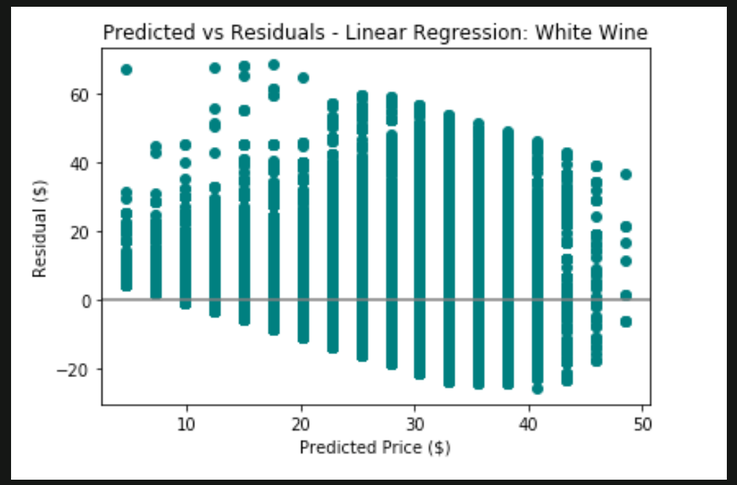
Dep. Variable:	price	**R-squared:**	0.350
Model:	OLS	**Adj. R-squared:**	0.350
Method:	Least Squares	**F-statistic:**	6336.
Date:	Sat, 09 May 2020	**Prob (F-statistic):**	0.00
Time:	13:25:55	**Log-Likelihood:**	-4.8716e+05
No. Observations:	117681	**AIC:**	9.744e+05
Df Residuals:	117670	**BIC:**	9.745e+05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

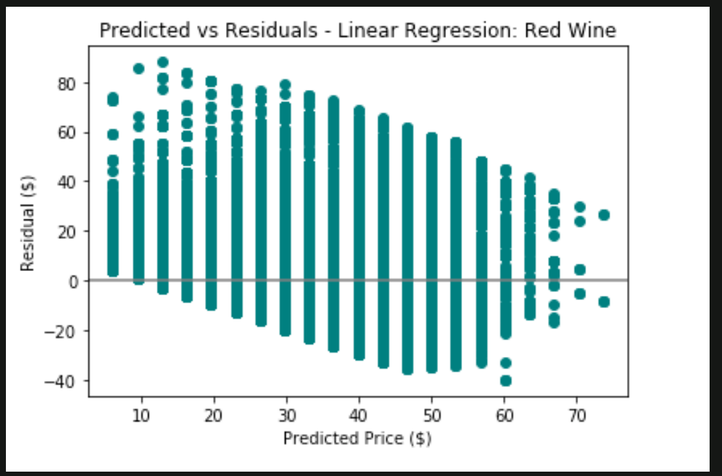
	coef	std err	t	P>|t|	[0.025	0.975]
const	-221.2869	1.141	-193.906	0.000	-223.524	-219.050
points	3.0485	0.015	209.871	0.000	3.020	3.077
Still_Red	8.1038	0.100	81.344	0.000	7.909	8.299
Argentina	-27.9245	0.228	-122.213	0.000	-28.372	-27.477
Australia	-24.6496	0.249	-98.839	0.000	-25.138	-24.161
Chile	-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

Regression Analysis

FIT OF MODELS







Discussion

RED VS WHITE WINE - PRICE FACTORS

CONSUMER DEMAND

In the early 2010's the global wine market saw a large increase in consumer demand from China, as Chinese society became more invested in foreign wines.

PRODUCTION INTENSIVITY

Red wine ferments the skins and seeds, which are then extracted. This process leaves red wine less susceptible to oxidation, which then requires secondary fermentation. White wine is fermented by yeast, chilled, and then filtered.

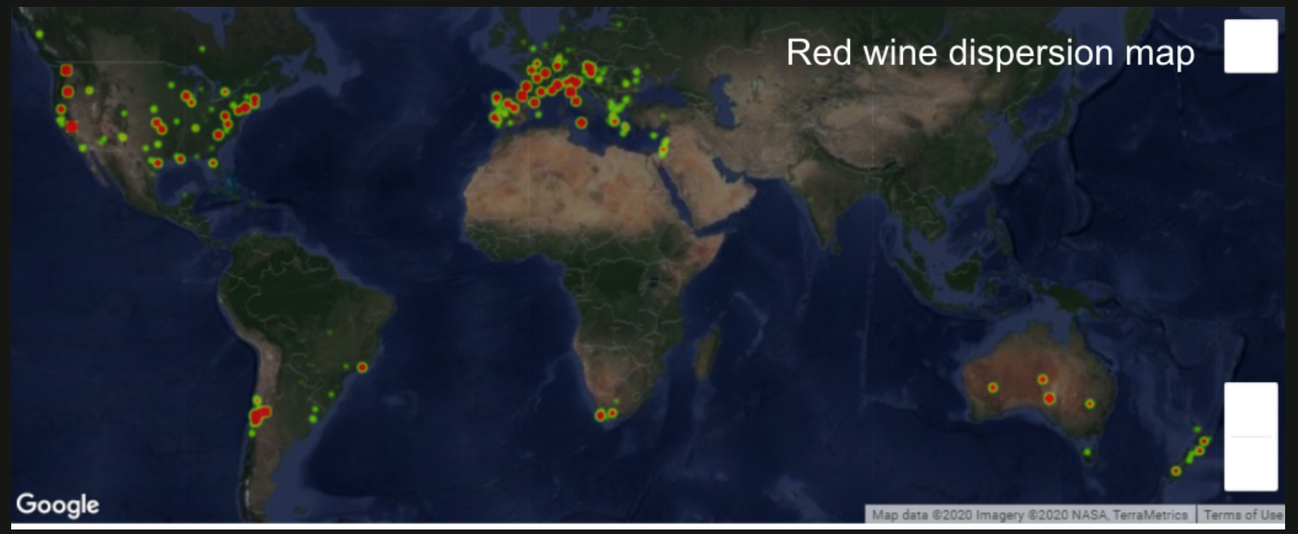
Discussion

WINE POPULARITY HEATMAPS

Red Wine Pop	
US	44550
Italy	13365
France	8021
Spain	6510
Argentina	4735
Chile	4060
Australia	3314
Portugal	2982
Austria	1620
South Africa	1277
New Zealand	1129

White Wine Pop	
US	17021
France	6467
Italy	4785
Germany	2109
New Zealand	1936
Chile	1626
Australia	1500
Spain	1217
Portugal	926
South Africa	920
Argentina	800
Austria	750
Greece	290
Israel	128
Canada	120
Hungary	116

Note: Wine population declines significantly beyond this point



Conclusion

PAIN POINTS

- Removing outliers by quantile ranges removed more data than anticipated.
  - Attempted removal of top and bottom 2.5% down to 1% of data with skewed results.
  - Ultimately, we decided on removing approx. ~.00001 of the top priced wine as outliers.
- Classifying wines as red vs. white had to be done manually, so it also cut out about 20k wines from the dataset and was more time-intensive that originally thought.
- Navigating GitHub

Conclusion

FURTHER CONSIDERATIONS/LIMITATIONS

- Location meteorological factors affecting the quality/popularity/price of wine were not taken into consideration.
- Classification of red vs white wine is not as straightforward with blends.
- There were large disparities in counts of red vs white as well as production in countries
- Regression only took into account top 12 countries - not comprehensive
- As the data was cleaned, we dropped a large amount of entries.
- Also would have been good to have aging data on wines.

QUESTIONS?