Grupo 1:
Cintia Mercado
Mariana Revilla
Eduardo Garza

Agosto 8, 2020

Can we predict the price of wine?

- What factors influence pricing?
- Can we determine if we are getting a good deal for our money when purchasing a bottle of wine?

Data Set

- Open Data @ Kaggle.com
- 150,930 records:
  - Wine
  - Country
  - Description
  - Grading (points)
  - Price
  - Province
  - Region_1
  - Region_2
  - Grape variety
  - Winery

Our Sample

Data Cleansing

- Dropped ID, description, region_2
- Dropped NaNs
- Re-ordered DataFrame

```
# Drop columns no needed
data.drop(columns=["Unnamed: 0", "description", "region_2"], inplace=True)

# Remove missing values
data.dropna(inplace=True)

#Order DataFrame Columns
data = data[["designation", "winery", "variety", "region_1", "province", "country", "price", "points"]]
data.head(10)
```

Sampling

- 3000 random records

```
# Sample for project purpose
sample = data.sample(n=3000, axis=0, random_state = 4)
sample.reset_index(drop = True, inplace=True)
```

Georeferencing sampling

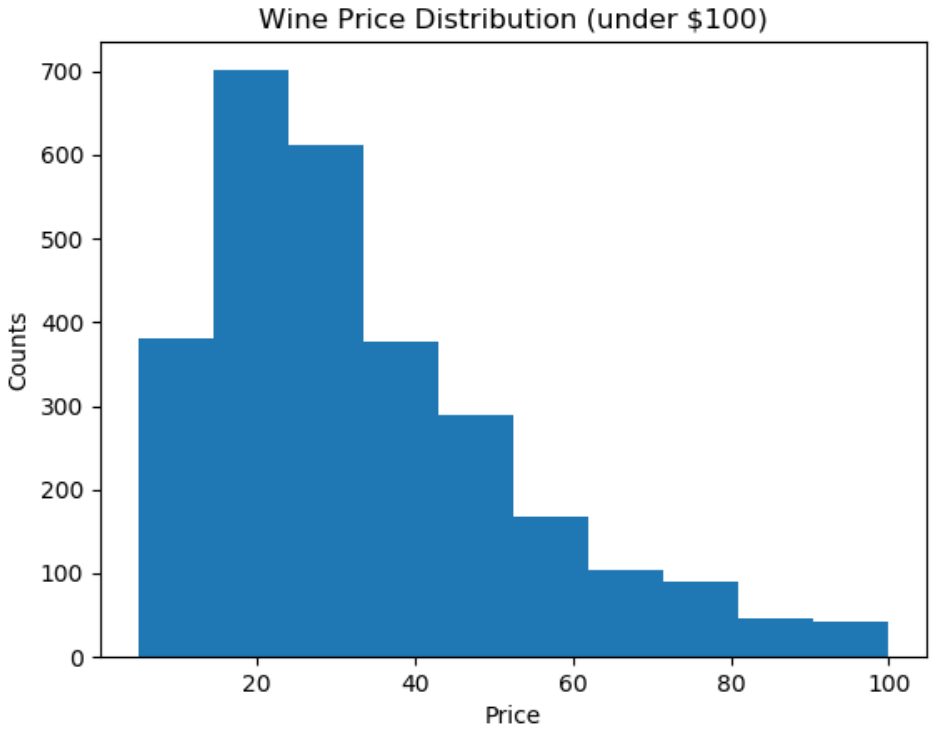
- Latitude and Longitude (geocode.api)
- With coordinates obtain altitude (elevation.api)

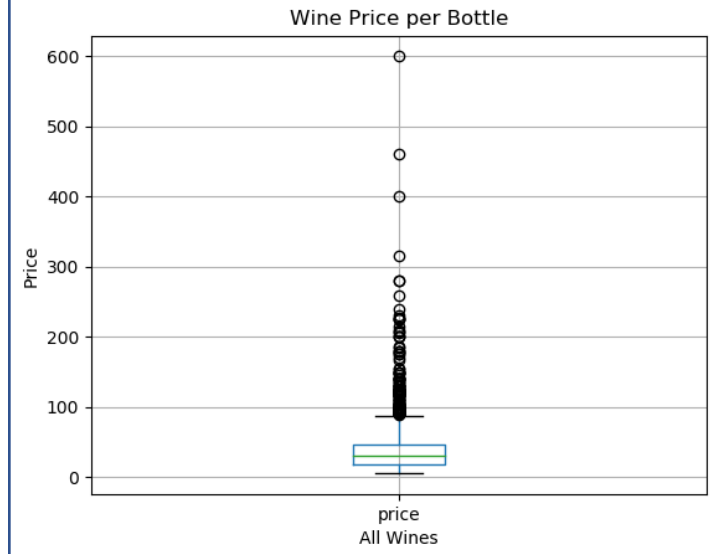
```
Processing row 9 Winery location Concannon,Livermore Valley,California 37.666199 -121.7397388 found
Processing row 10 Winery location Sandrone,Barolo,Piedmont 44.6208903 7.954625300000001 found
Processing row1 Winery elevation 323.2424926757812 found
Processing row2 Winery elevation 354.9888916015625 found
Processing row3 Winery elevation 280.32470703125 found
```

Our Sample

Sample Cleansing

- Dropped records with no Latitude, Longitude or Altitude (88 -> 2,912)
- Dropped “Canada” due to low number of observations (3 -> 2909)
- Focused on wines under $100usd (100 -> 2,809)





The lower quartile of price is: 18.0
The upper quartile of price is: 46.0
The interquartile range of price is: 28.0
The the median of price is: 30.0
Values below -24.0 could be outliers. (zero)
Values above 88.0 could be outliers.

Final Sample

- 2,809 wines
- 1,791 wineries
- 172 varieties of wine
- 6 countries
- 48 provinces

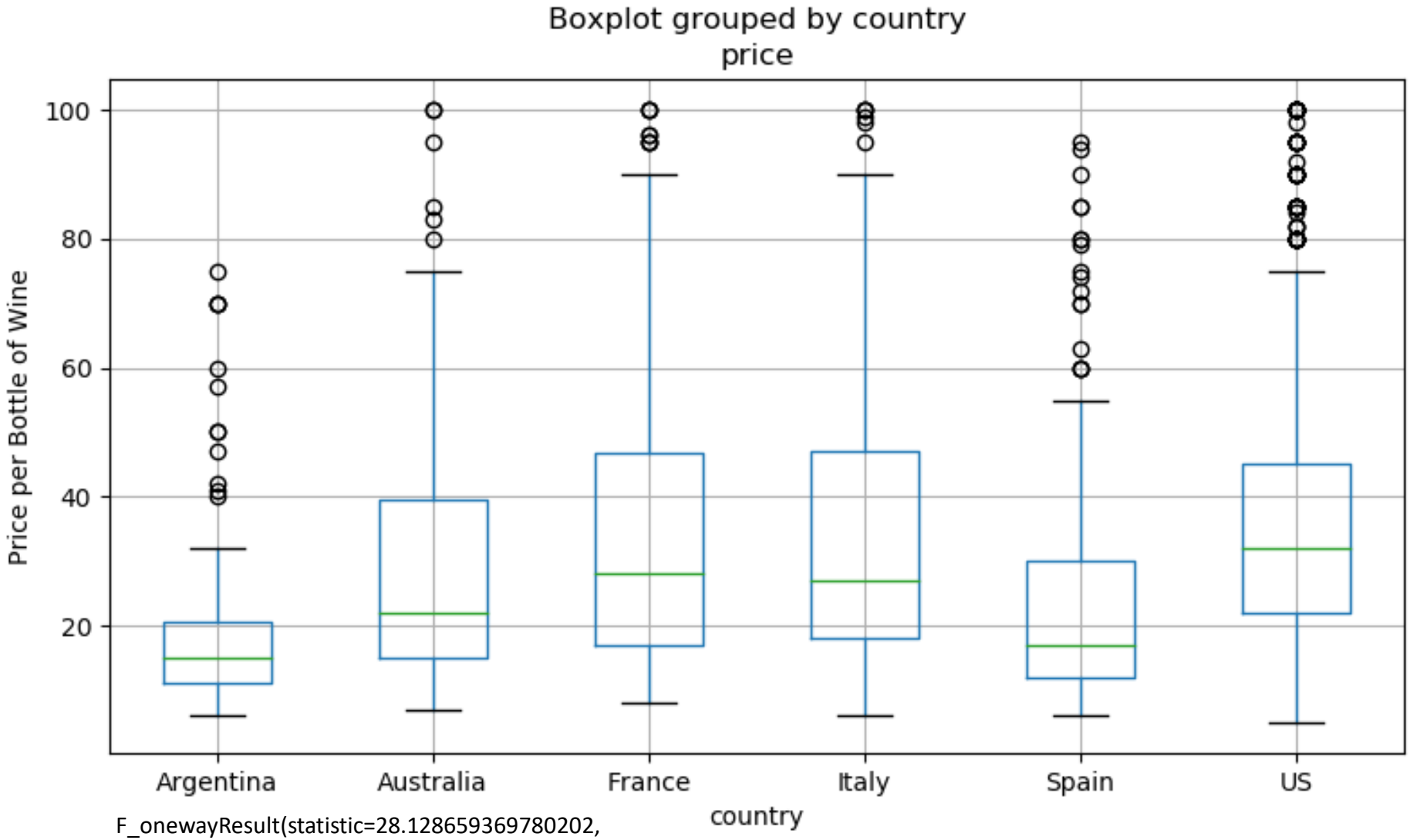
Sample Description

	price	points	latitud	longitud	elevation
count	2809.000000	2809.000000	2809.000000	2809.000000	2809.000000
mean	33.421858	88.090780	34.552359	-58.581826	270.499286
std	20.057513	3.215317	21.708023	73.716641	320.600963
min	5.000000	80.000000	-42.809838	-123.799459	0.666450
25%	18.000000	86.000000	37.599994	-122.265389	66.650963
50%	28.000000	88.000000	39.086566	-118.040206	188.161011
75%	45.000000	90.000000	44.663166	7.285526	352.602264
max	100.000000	98.000000	49.236201	153.288288	6907.093262

Price and Rating Statistics Grouped by Country

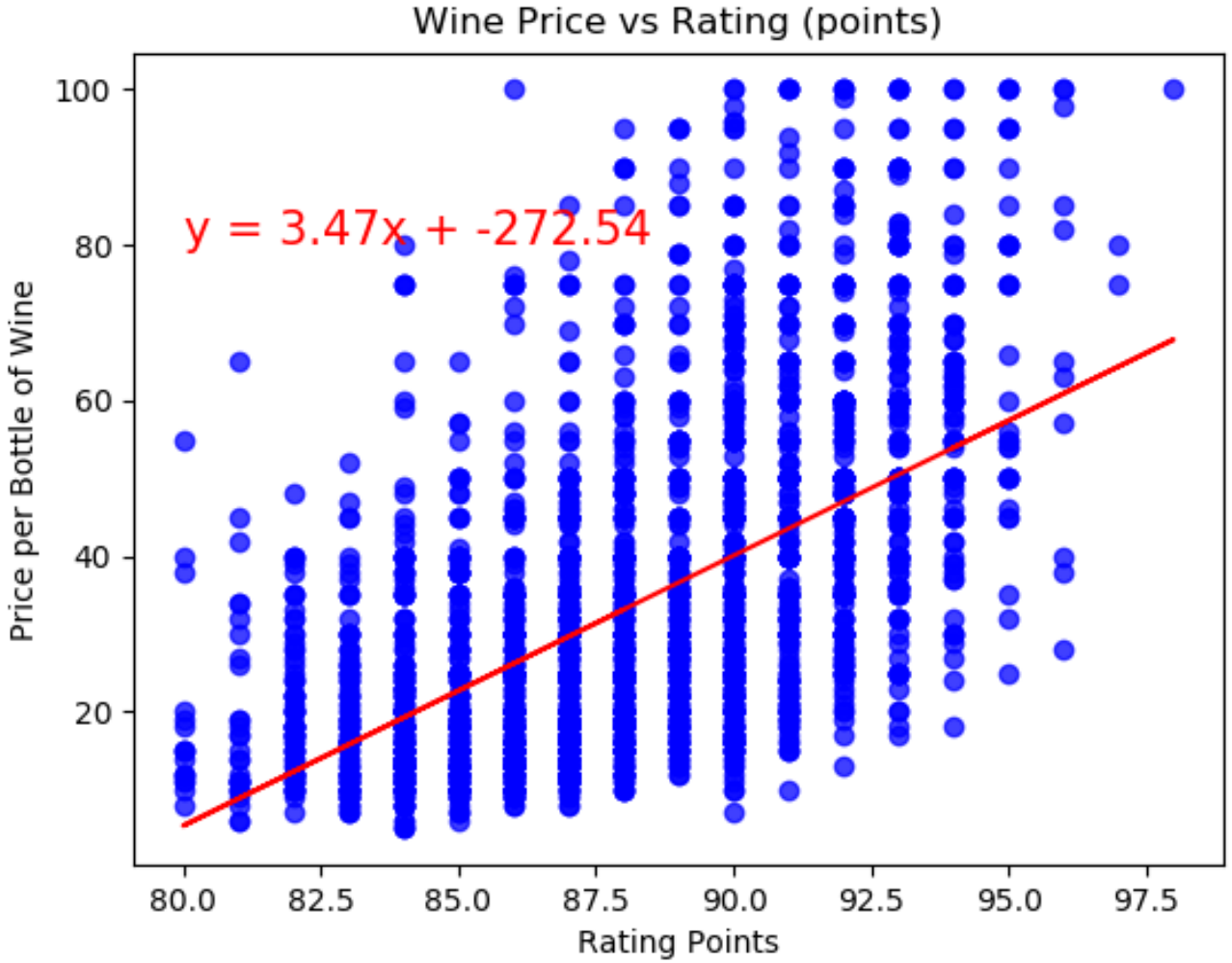
	price					points			
country	mean	median	var	std		mean	median	var	std
Argentina	19.389313	15.0	182.239577	13.499614		86.106870	86	9.850029	3.138476
Australia	30.099237	22.0	450.951615	21.235621		88.152672	88	8.222666	2.867519
France	34.234266	28.0	488.460716	22.101147		88.681818	89	8.147528	2.854387
Italy	34.035382	27.0	442.564044	21.037206		88.322160	88	6.741168	2.596376
Spain	24.451613	17.0	347.711768	18.647031		86.622120	86	8.375064	2.893970
US	35.849370	32.0	359.793364	18.968220		88.274718	88	11.800974	3.435255

Price seems to have a significant amount of variation per country



F_onewayResult(statistic=28.128659369780202, pvalue=6.700349494424998e-28)

The ANOVA demonstrates “Prices per Country” are independent groups

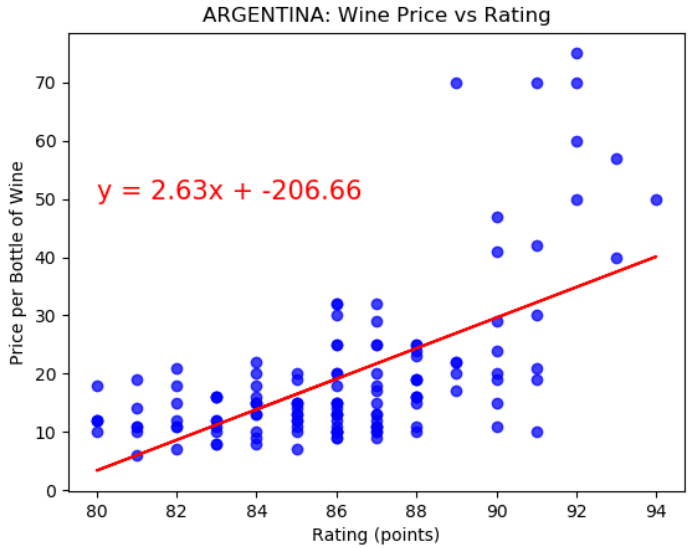


The correlation between both factors is 0.56
The r-squared is: 0.31001201962230474

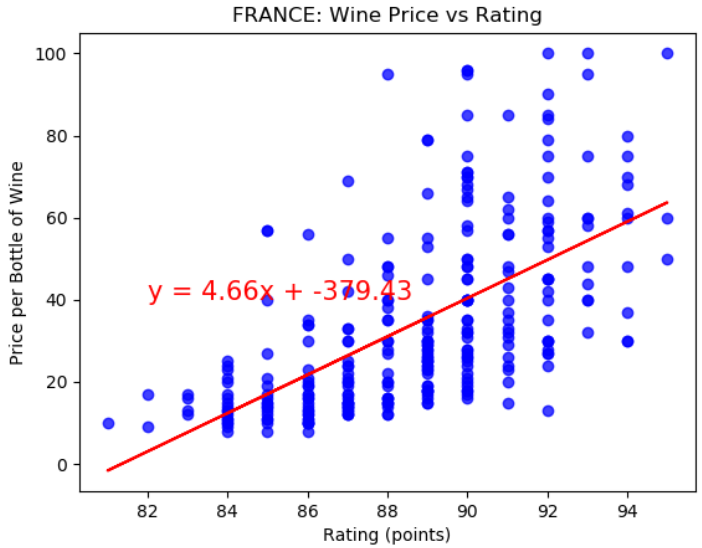
Moderate and positive correlation for prices and rating (points)

Does it mean that prices are higher for better tasting wines? or are we prone to be more easygoing when rating more expensive wines?

Same relationship of price and rating in every country

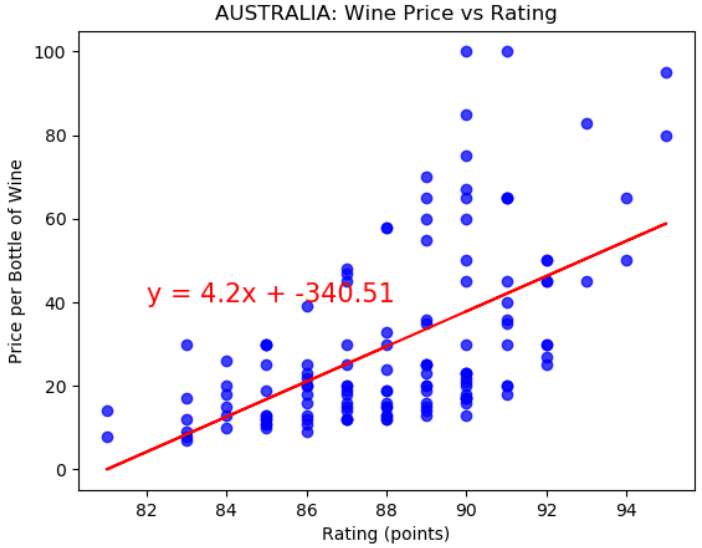


The correlation between both factors is 0.61
The r-squared is: 0.3724878603227943

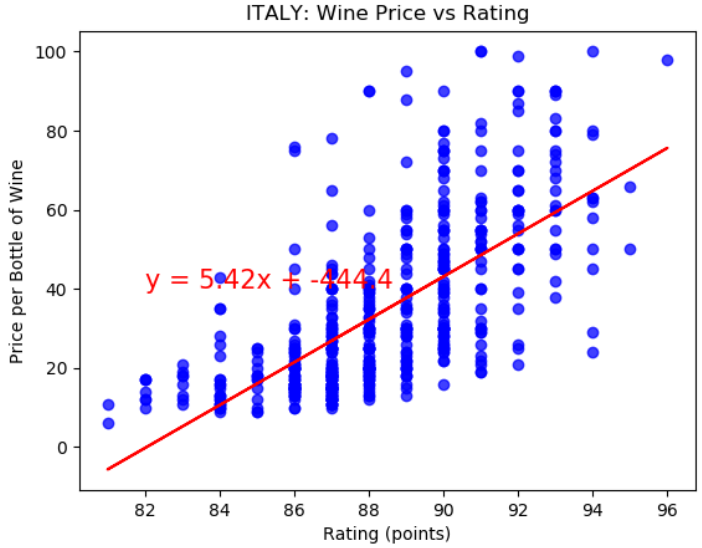


The correlation between both factors is 0.6
The r-squared is: 0.3629254372526378

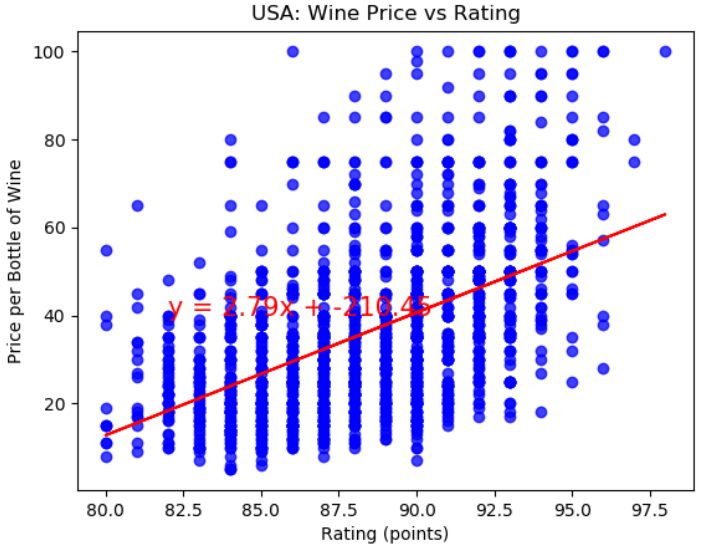
The correlation between both factors is 0.48
The r-squared is: 0.23221956621624312



The correlation between both factors is 0.57
The r-squared is: 0.3222847769928225



The correlation between both factors is 0.67
The r-squared is: 0.44695987199960985

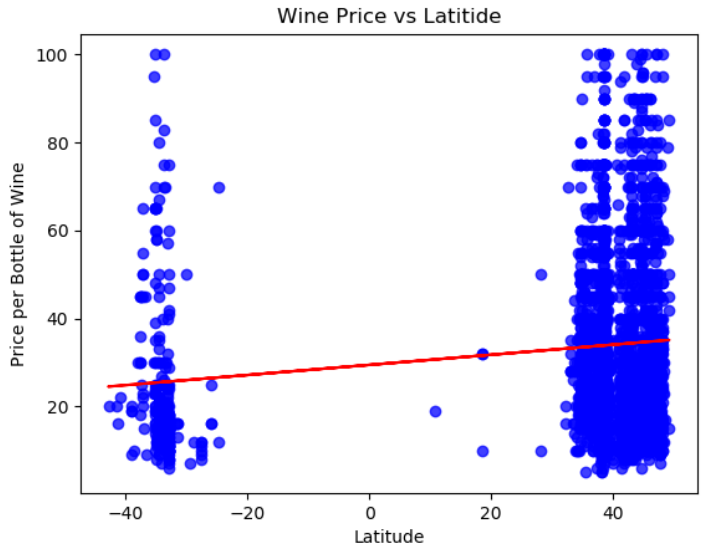


The correlation between both factors is 0.51
The r-squared is: 0.2553412894715367

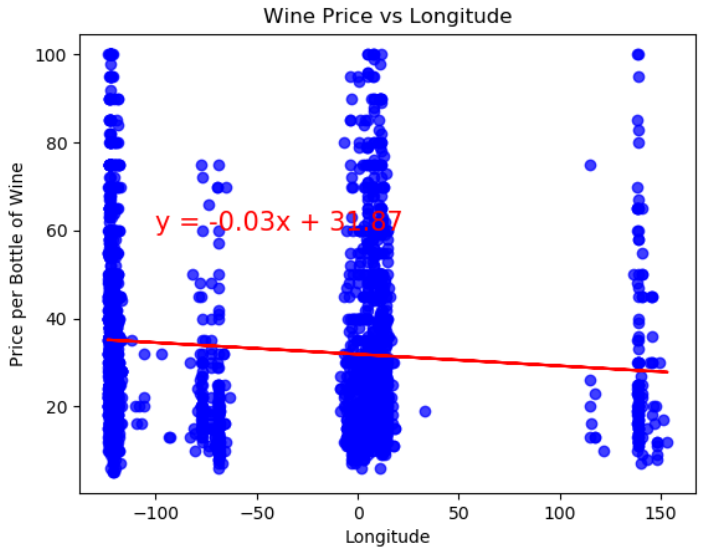
Heat Map for Wine Prices by Province, Country

Markers for Wine Prices by Province, Country

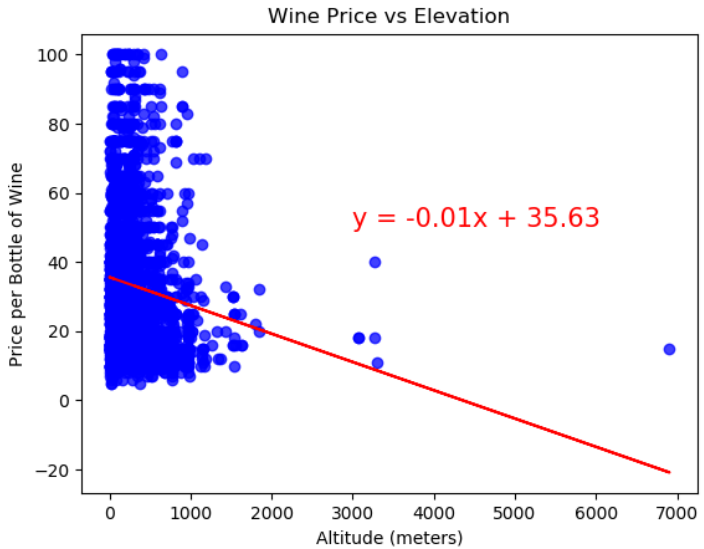
Latitude, Longitude and Elevation do not seem to be explanatory variables for wine pricing



The correlation between both factors is 0.12
The r-squared is: 0.015506105074422814



The correlation between both factors is -0.1
The r-squared is: 0.009471488415690064



The correlation between both factors is -0.13
The r-squared is: 0.01709064379223488

Regression Models

Model 1

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.763
Model:	OLS	Adj. R-squared (uncentered):	0.762
Method:	Least Squares	F-statistic:	999.8
Date:	Tue, 04 Aug 2020	Prob (F-statistic):	0.00
Time:	21:04:07	Log-Likelihood:	-12255.
No. Observations:	2809	AIC:	2.453e+04
Df Residuals:	2800	BIC:	2.458e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
points	0.5100	0.041	12.341	0.000	0.429	0.591
latitud	0.2148	0.063	3.404	0.001	0.091	0.339
longitud	0.1092	0.039	2.767	0.006	0.032	0.187
elevation	-0.0052	0.001	-3.982	0.000	-0.008	-0.003
country_Australia	-21.4376	8.712	-2.461	0.014	-38.521	-4.355
country_France	-20.1779	5.529	-3.650	0.000	-31.019	-9.337
country_Italy	-19.9438	5.505	-3.623	0.000	-30.739	-9.149
country_Spain	-25.4182	5.182	-4.905	0.000	-35.579	-15.257
country_US	-3.6810	4.937	-0.746	0.456	-13.363	6.000

Omnibus:	519.566	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	869.939
Skew:	1.221	Prob(JB):	1.24e-189
Kurtosis:	4.213	Cond. No.	1.29e+04

[1] Std errors assume covariance matrix of errors is correctly specified
[2] Strong Multicollinearity

Model 2

OLS Regression Results

Dep. Variable:	price	R-squared:	0.310
Model:	OLS	Adj. R-squared:	0.310
Method:	Least Squares	F-statistic:	1261.
Date:	Tue, 04 Aug 2020	Prob (F-statistic):	1.76e-228
Time:	21:03:59	Log-Likelihood:	-11887.
No. Observations:	2809	AIC:	2.378e+04
Df Residuals:	2807	BIC:	2.379e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-272.5440	8.621	-31.613	0.000	-289.449	-255.639
points	3.4733	0.098	35.513	0.000	3.282	3.665

Omnibus:	415.549	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	644.306
Skew:	1.031	Prob(JB):	1.23e-140
Kurtosis:	4.119	Cond. No.	2.42e+03

[1] Std errors assume covariance matrix of errors is correctly specified
[2] Strong Multicollinearity

Model 3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.320
Model:	OLS	Adj. R-squared:	0.319
Method:	Least Squares	F-statistic:	329.5
Date:	Tue, 04 Aug 2020	Prob (F-statistic):	1.16e-232
Time:	21:03:50	Log-Likelihood:	-11867.
No. Observations:	2809	AIC:	2.374e+04
Df Residuals:	2804	BIC:	2.377e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-268.0631	8.688	-30.855	0.000	-285.098	-251.028
points	3.4056	0.098	34.628	0.000	3.213	3.598
latitud	0.0326	0.016	2.069	0.039	0.002	0.064
longitud	-0.0175	0.005	-3.859	0.000	-0.026	-0.009
elevation	-0.0025	0.001	-2.422	0.016	-0.004	-0.000

Omnibus:	442.278	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	713.949
Skew:	1.062	Prob(JB):	9.29e-156
Kurtosis:	4.261	Cond. No.	1.18e+04

[1] Std errors assume covariance matrix of errors is correctly specified
[2] Strong Multicollinearity

Model 4

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.759
Model:	OLS	Adj. R-squared (uncentered):	0.759
Method:	Least Squares	F-statistic:	1473.
Date:	Tue, 04 Aug 2020	Prob (F-statistic):	0.00
Time:	21:04:12	Log-Likelihood:	-12275.
No. Observations:	2809	AIC:	2.456e+04
Df Residuals:	2803	BIC:	2.460e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
points	0.3124	0.019	16.313	0.000	0.275	0.350
country_Australia	2.5568	2.377	1.076	0.282	-2.104	7.217
country_France	6.5265	2.041	3.197	0.001	2.524	10.529
country_Italy	6.4399	1.883	3.421	0.001	2.748	10.131
country_Spain	-2.6127	2.108	-1.240	0.215	-6.745	1.520
country_US	8.2688	1.761	4.695	0.000	4.815	11.722

Omnibus:	535.860	Durbin-Watson:	1.980
Prob(Omnibus):	0.000	Jarque-Bera (JB):	908.477
Skew:	1.250	Prob(JB):	5.33e-198
Kurtosis:	4.228	Cond. No.	964.

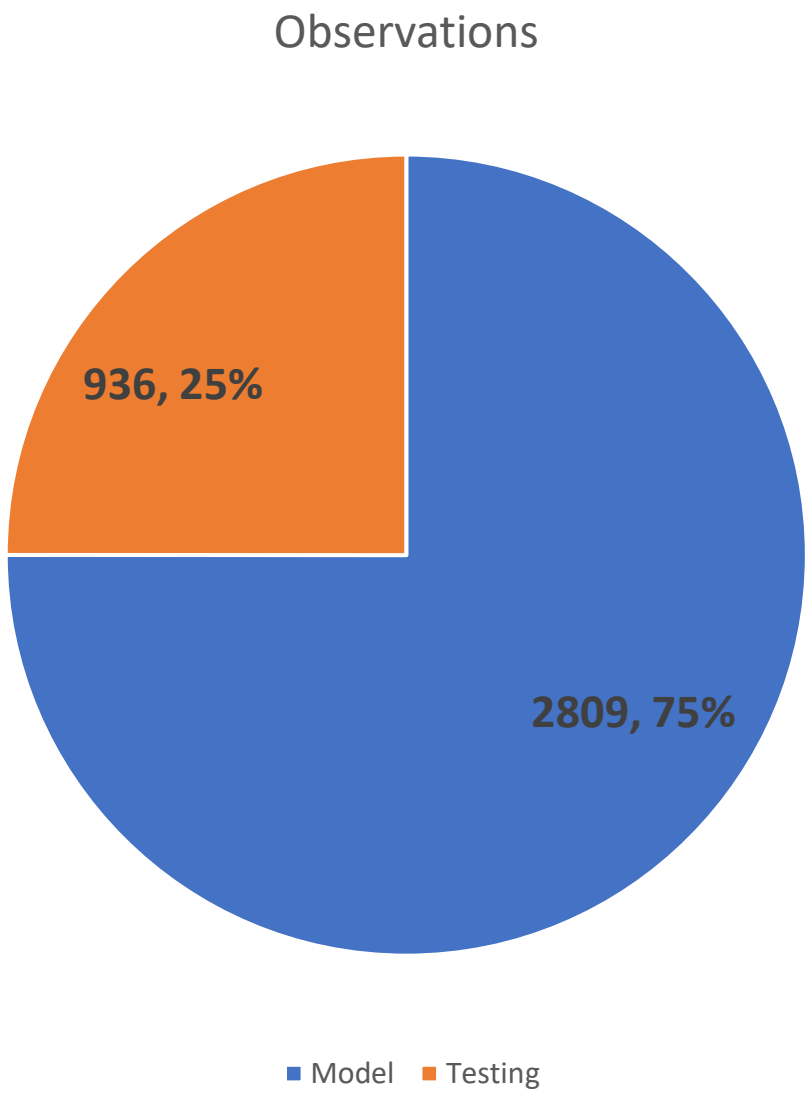
[1] Std errors assume covariance matrix of errors is correctly specified

Testing Data Set

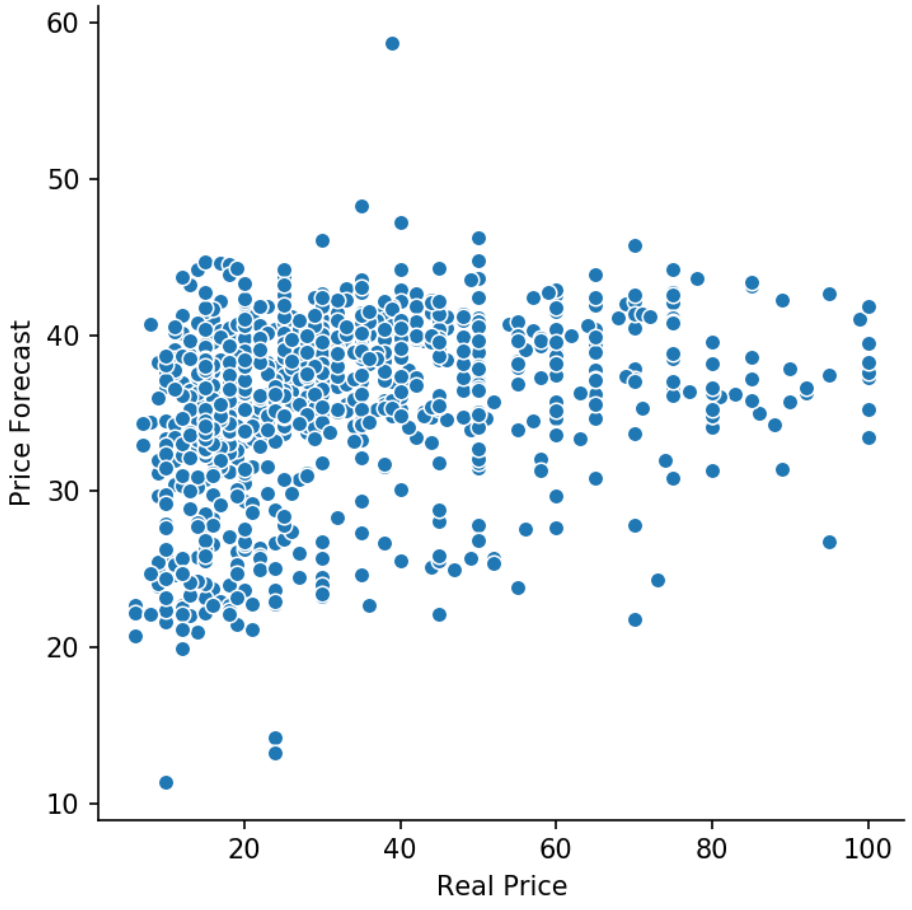
Test Sample
- 1000 random records from same dataset

Georeferencing sampling
- Latitude and Longitude (geocode.api)
- With coordinates obtain altitude (elevation.api)

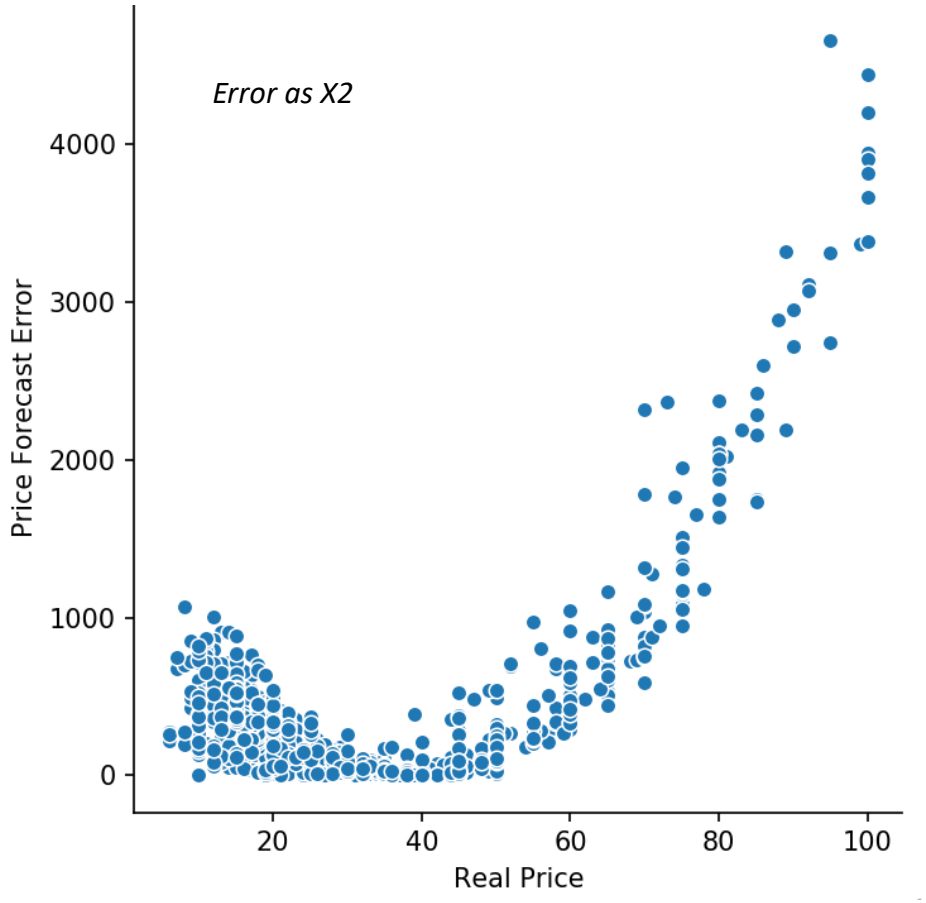
Sample Cleansing
- Dropped records with no Latitude, Longitude or Altitude (30 -> 970)
- Dropped “Canada” (5 -> 965)
- Only wines under $100usd (14 -> 936)



Model 1 Testing Results

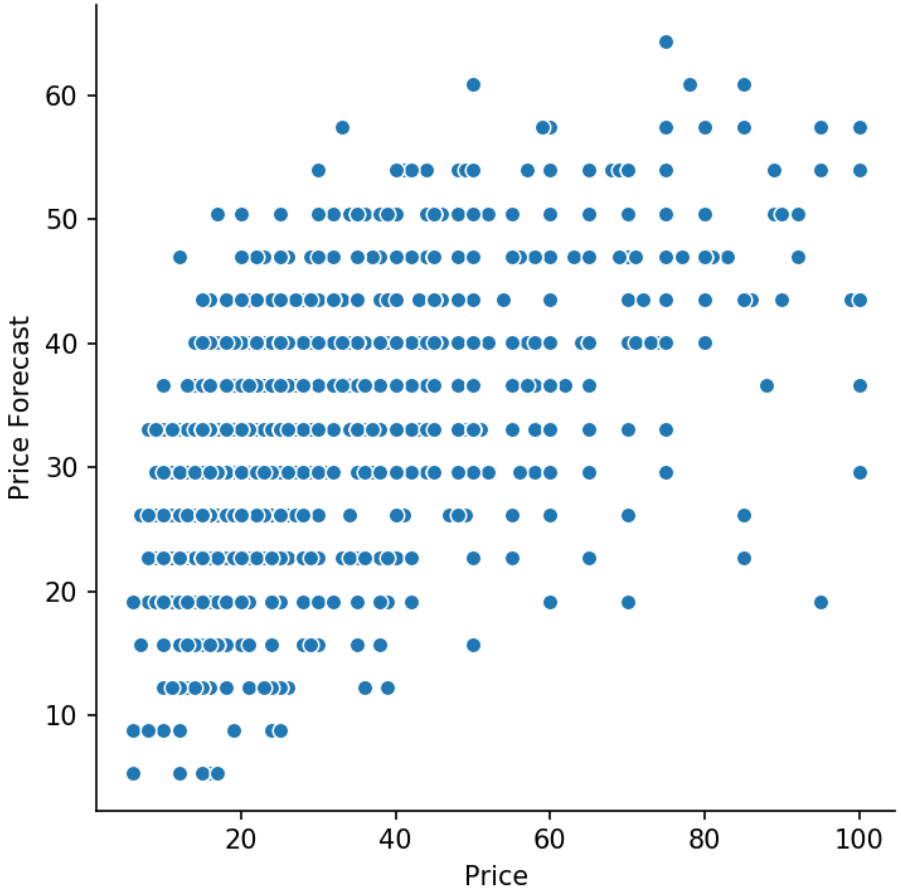


Price Forecast *= 0.51(Points) + .2148(Latitude) + .1092(Longitude) - (.0052)(Elevation) - (21.4376)(Australia) - (20.1779)(France) - (19.9438)(Italy) + (25.4182)(Spain)*

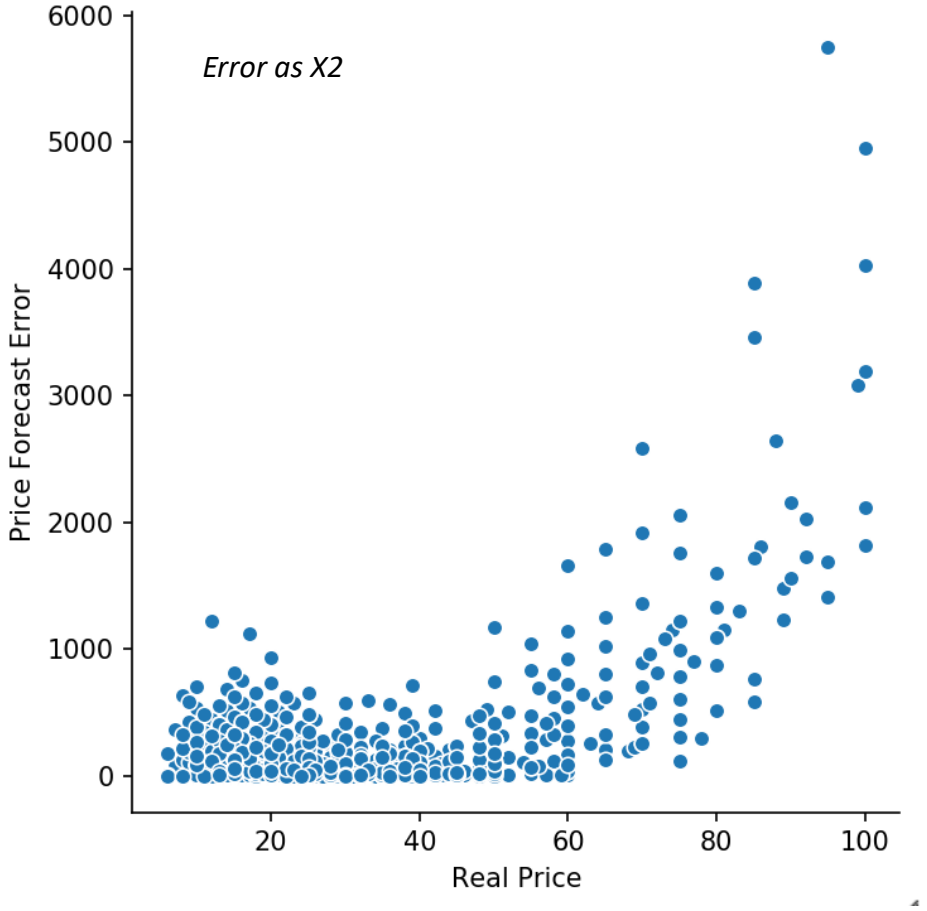


MSE = 368.09077103952404
RMSE = +- 19.1856918311413

Model 2 Testing Results

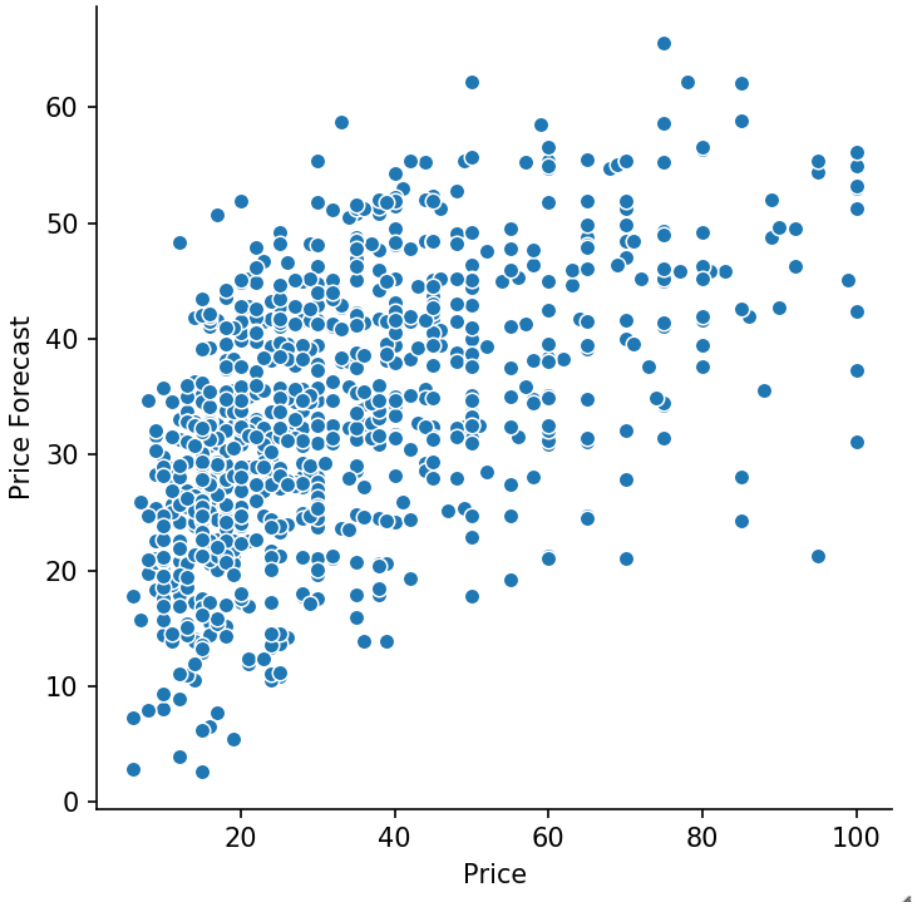


***Price Forecast** = -268.0631 + 3.4056(Points)*

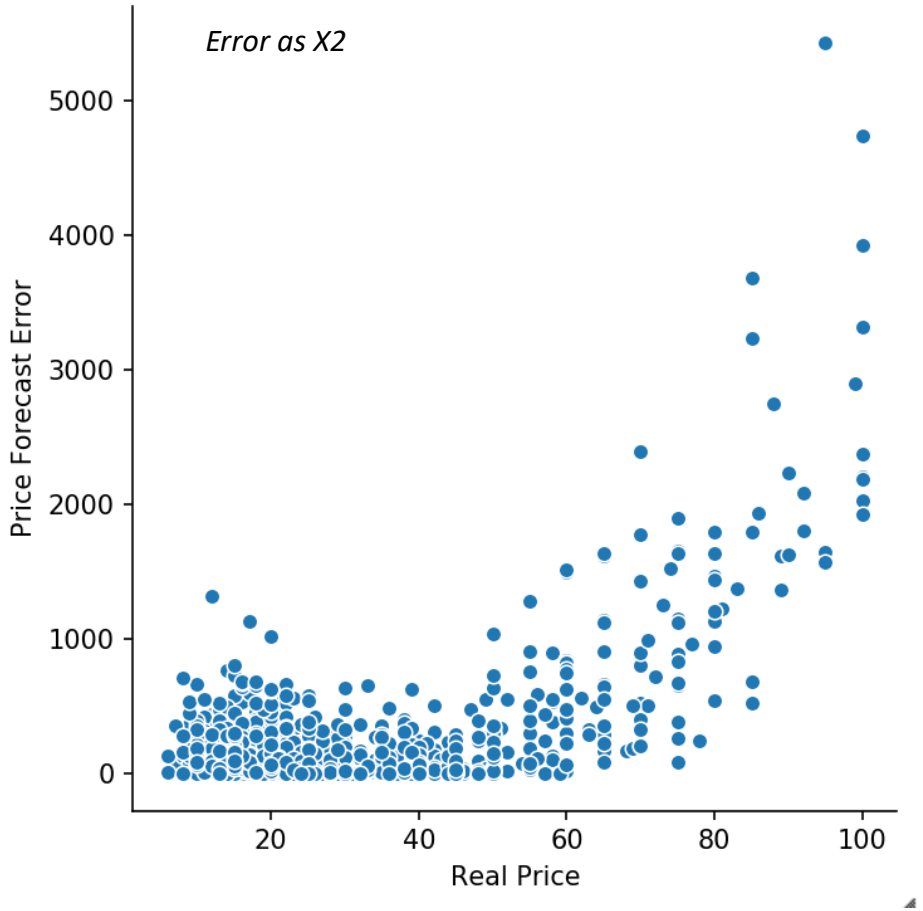


MSE = 274.20792466784167
RMSE = +- 16.559224760472382

Model 3 Testing Results



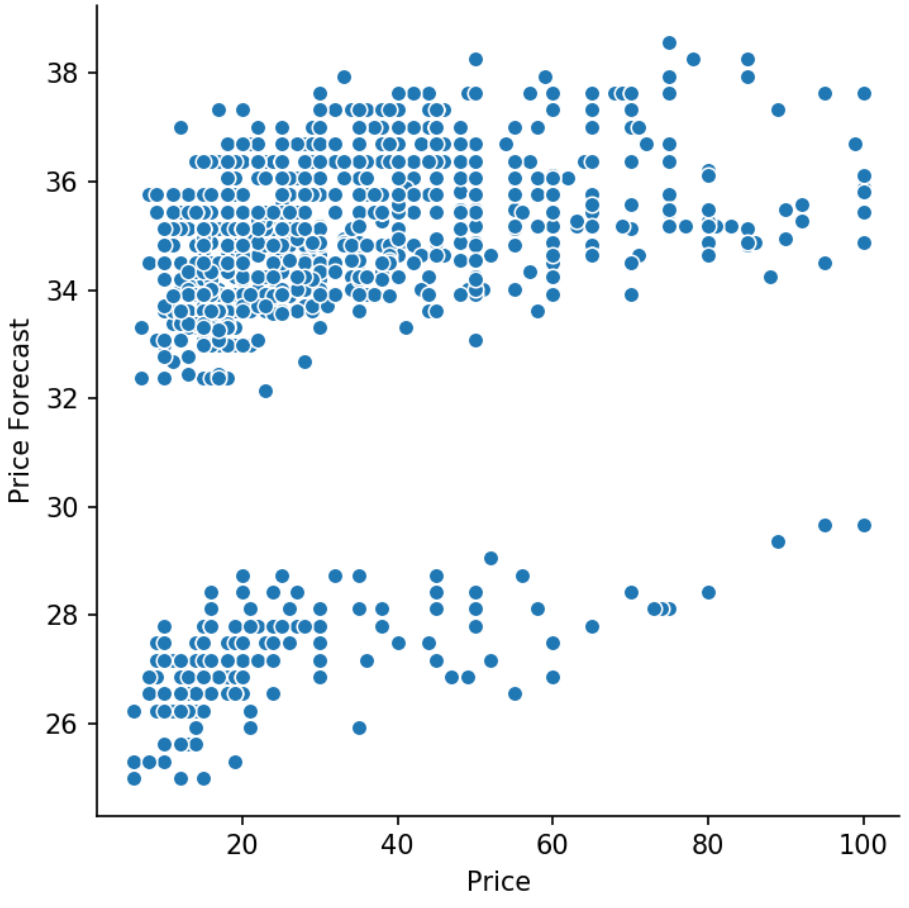
***Price Forecast** = -268.0631 + 3.4056(Points) + .0326(Latitude) - (.0175)(Longitude) - (.0025)(Elevation)*



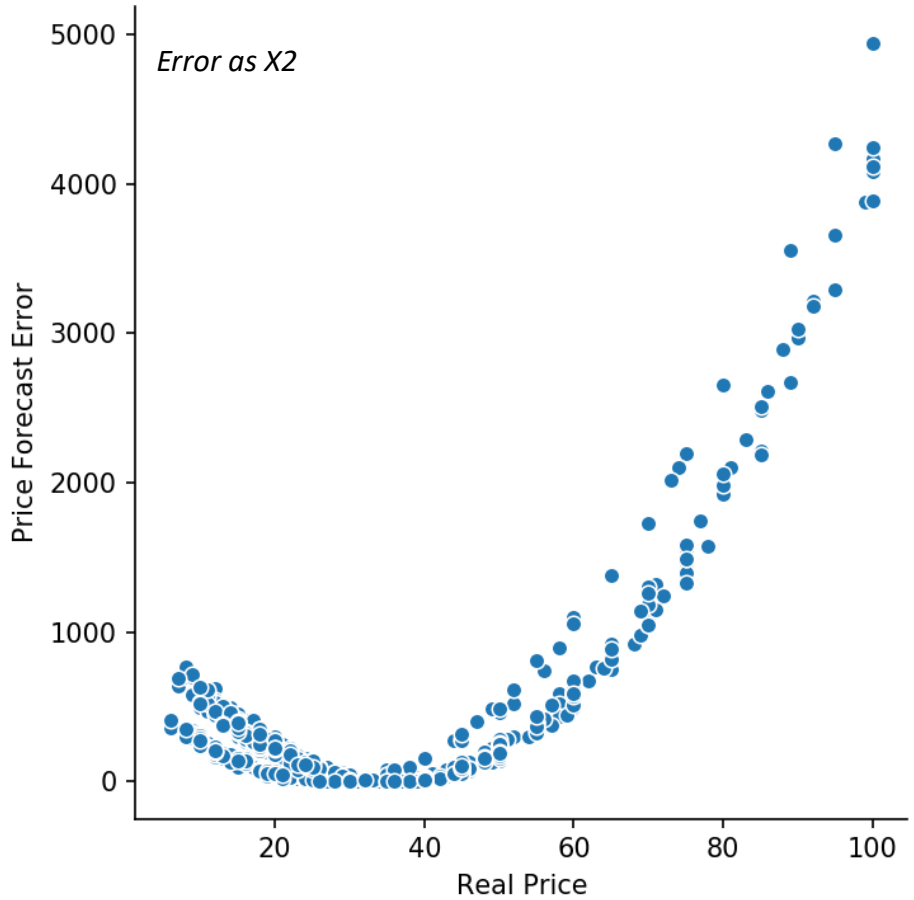
MSE = 271.8739375731748

RMSE = +- 16.48860023086177

Model 4 Testing Results



Price Forecast *= 0.3124(Points) + 6.5265(France) + 6.4399(Italy) + 8.2688(US)*



MSE = 362.7593081679807

RMSE = +- 19.04624131339254

Conclusions

- Overall, the models are more accurate for wine bottles priced between $20 and $60.
- Model 4 can be adjusted for better fit. It also depicts two price clusters: 1) High price wines: USA, Italy and France and 2) Low Price wines: Argentina, Australia and Spain.
- A larger sample data set could result in a more promising model. Our sample was limited due to budget constraints.
- Wine tasting 'ratings' seems to be an important contributor to wine prices. Future research could establish if *better ratings contribute to higher prices* or *if higher prices contribute to better ratings*. Regardless, wine producers need to pay close attention to reviews. On the other hand, wine consumers can rely on ratings when purchasing wine but keeping in mind that there are diminishing returns as the price continues to increase.
- Other factors not considered in our study should be observed to have a more solid model, as for example: grape variety, harvest year, winery years of operation as well as weather variables per region like rain per year and temperature.
- Even though the wine market is considered as a global market, our research shows *prices per country* are independent, maybe influenced by local market dynamics. Thus, pricing models per country could be done at later stage.
- Future research could be done by reframing the question as *what is the probability this is a good wine*? Using a log regression and factors such as price, ratings, country of origin, etc. could give us a probability of success.
- As there are diminishing returns for taste rating, another approach could be to use an optimization model that seeks to *maximize rating while minimizing price* to pursue *best value deals*.

Grupo 1:
Cintia Mercado
Mariana Revilla
Eduardo Garza

Agosto 8, 2020