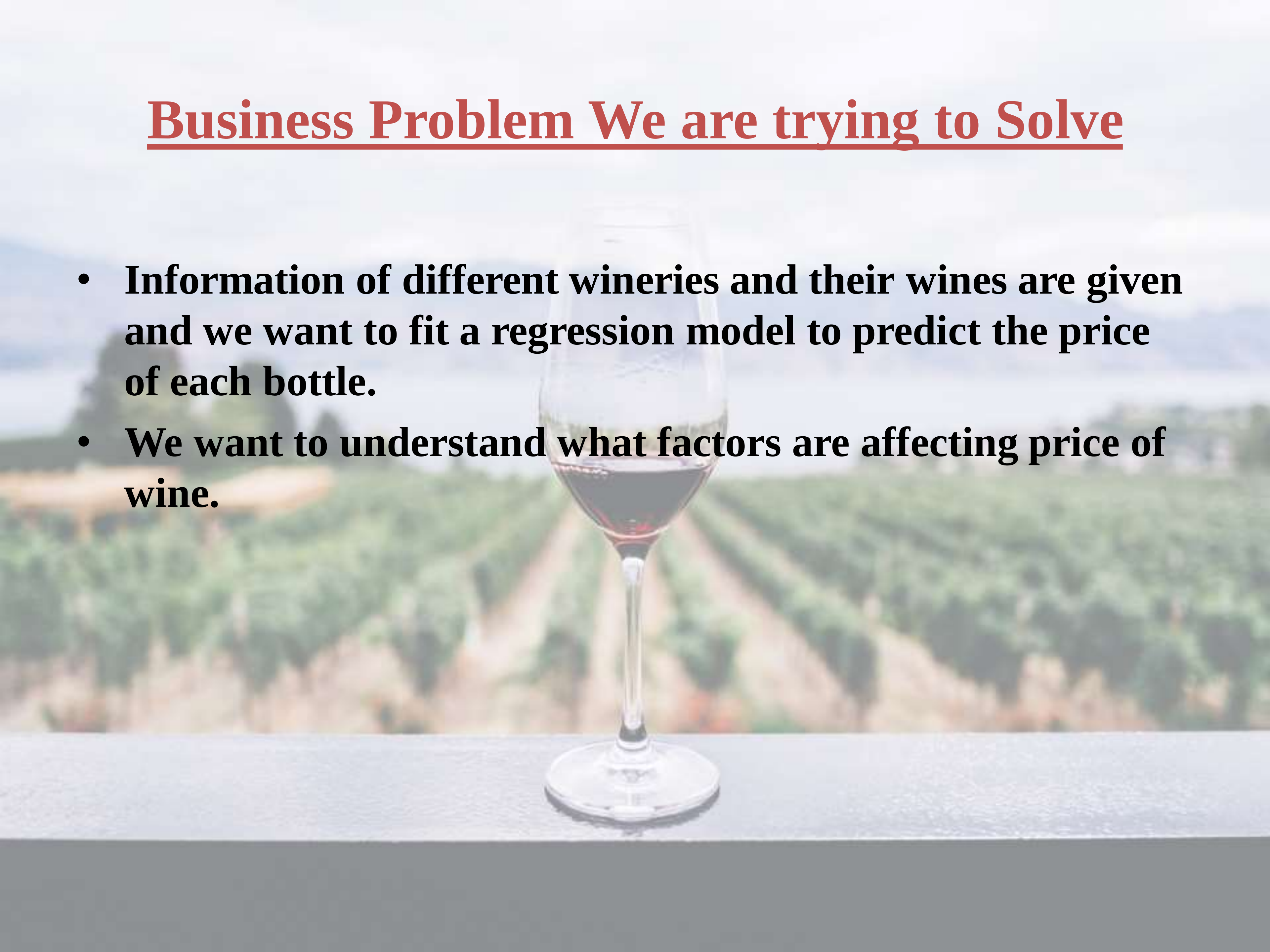
Spanish Wine Price Prediction Project

Business Problem Understanding

- **Business Problem We are trying to Solve**
- **Constraints**
- **Scope**
- **Objectives**

Business Problem We are trying to Solve

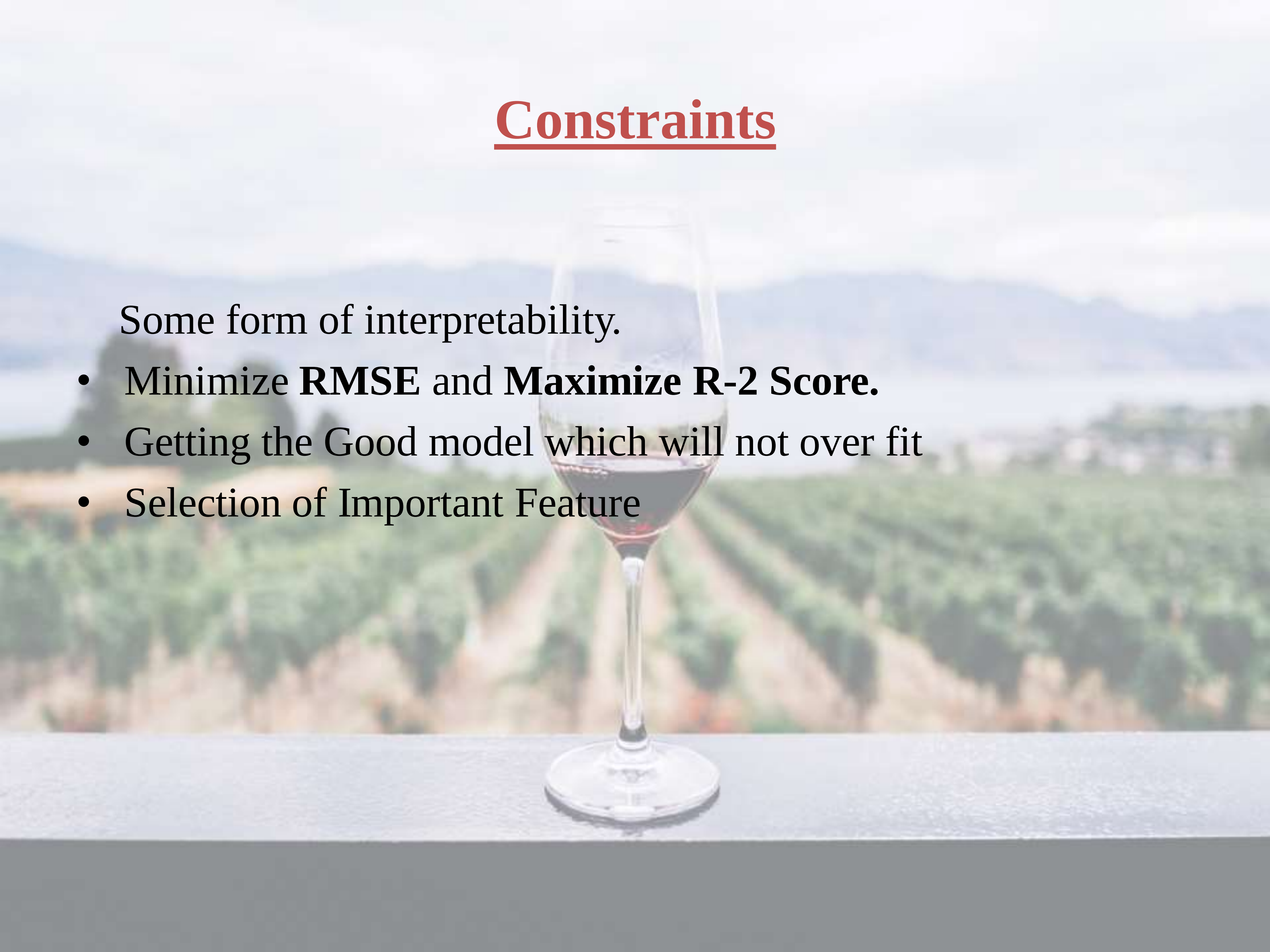
- **Information of different wineries and their wines are given and we want to fit a regression model to predict the price of each bottle.**
- **We want to understand what factors are affecting price of wine.**

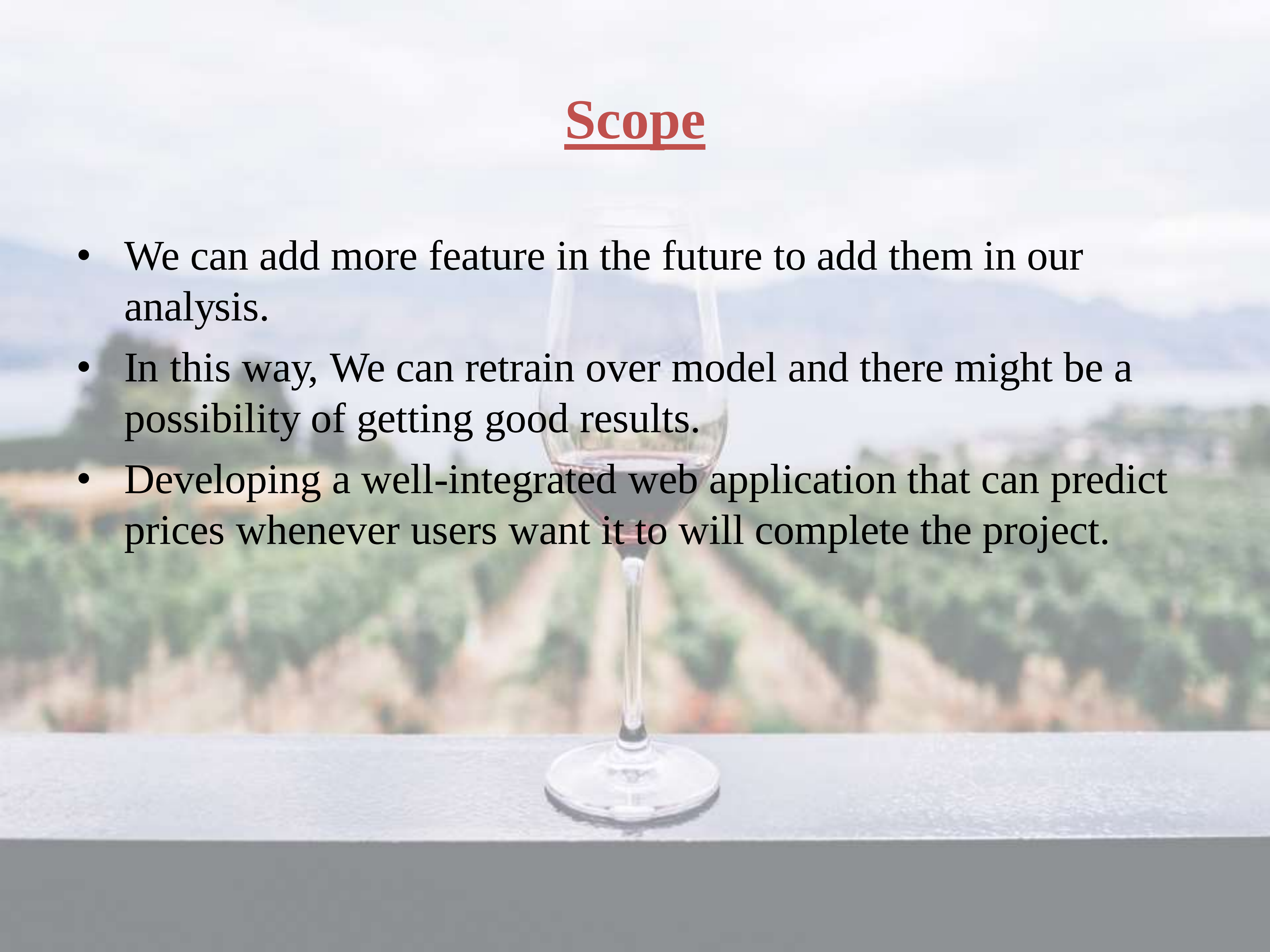


Constraints

Some form of interpretability.

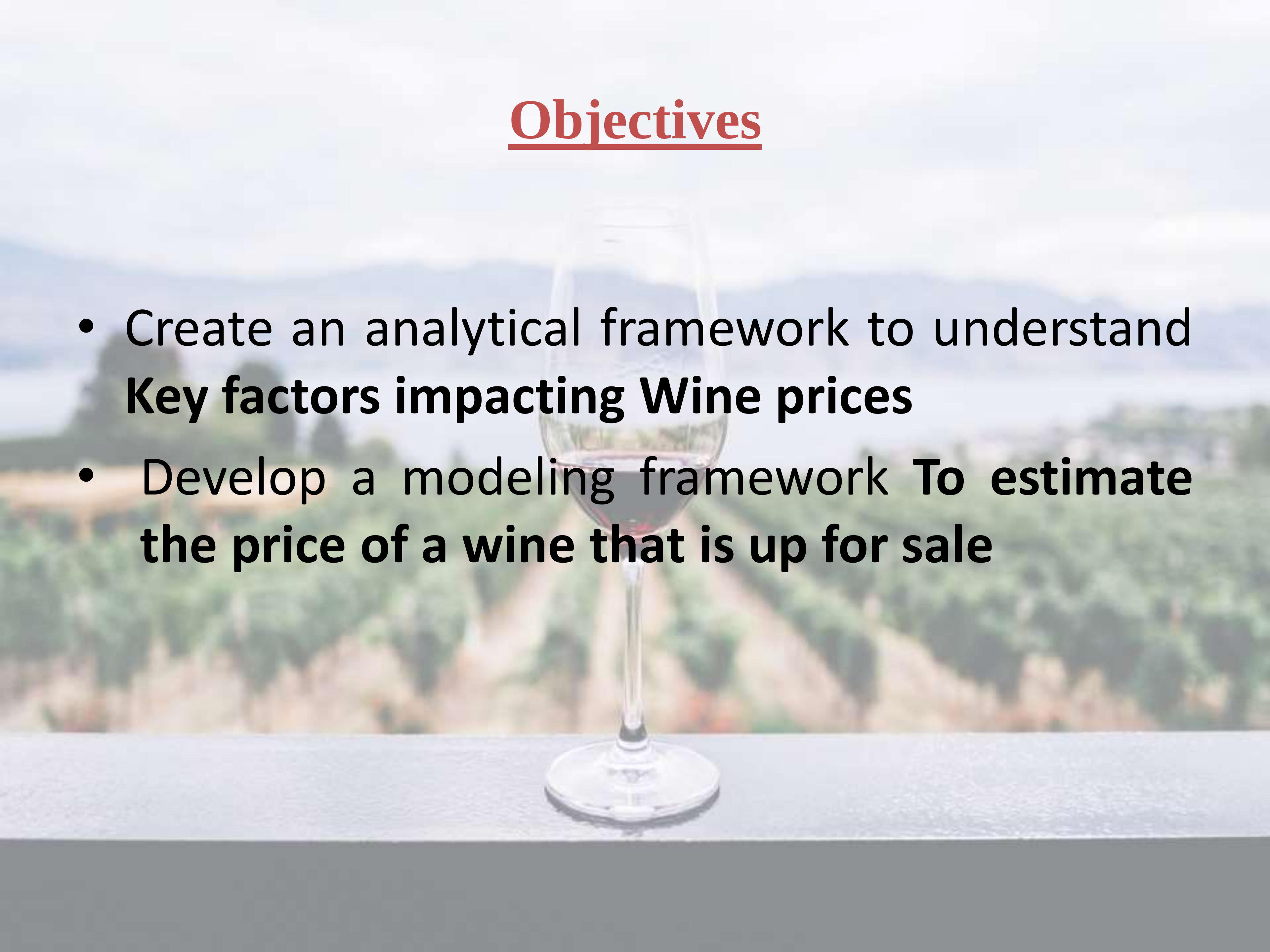
- Minimize **RMSE** and **Maximize R-2 Score.**
- Getting the Good model which will not over fit
- Selection of Important Feature





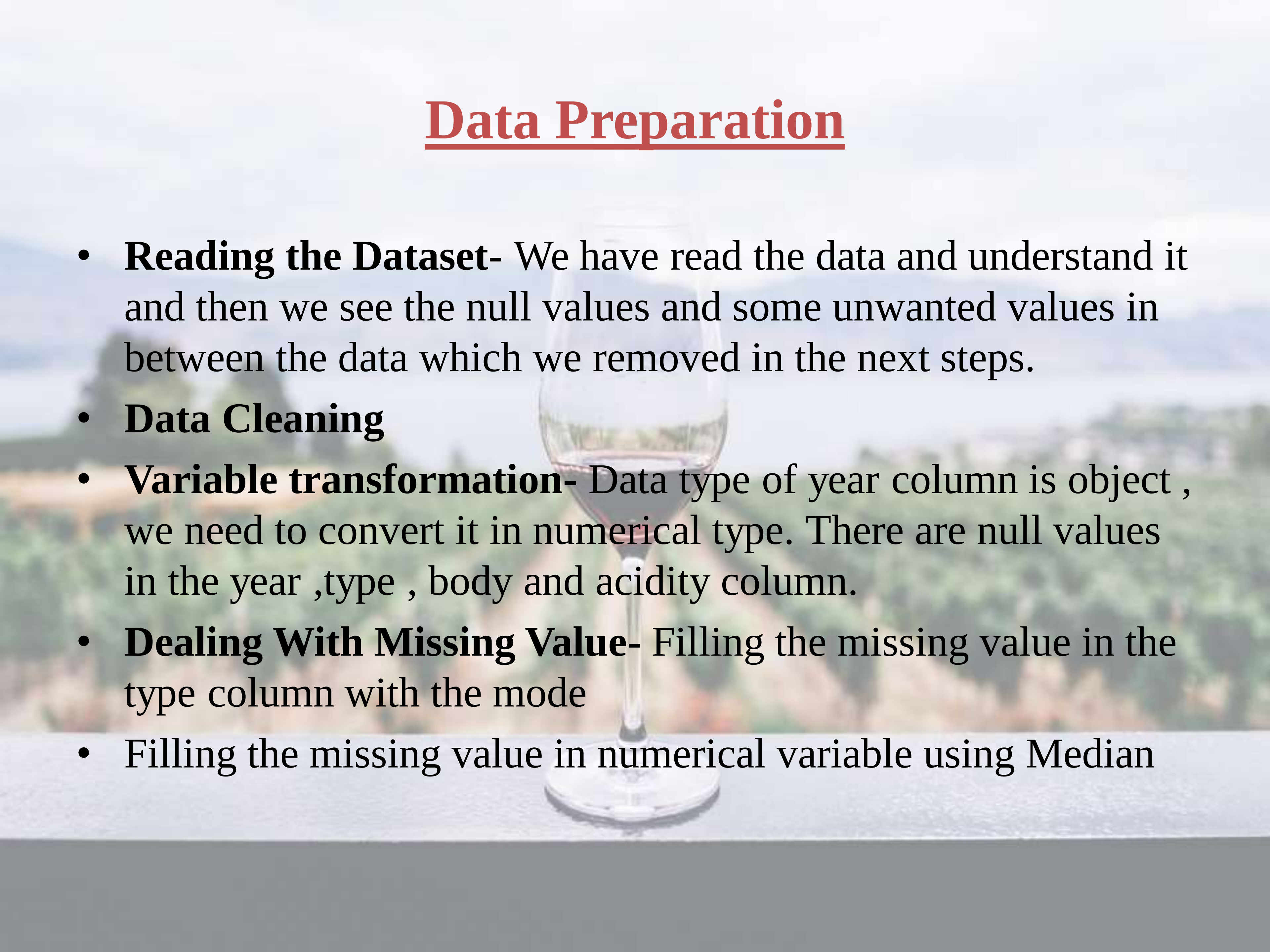
Scope

- We can add more feature in the future to add them in our analysis.
- In this way, We can retrain over model and there might be a possibility of getting good results.
- Developing a well-integrated web application that can predict prices whenever users want it to will complete the project.



Objectives

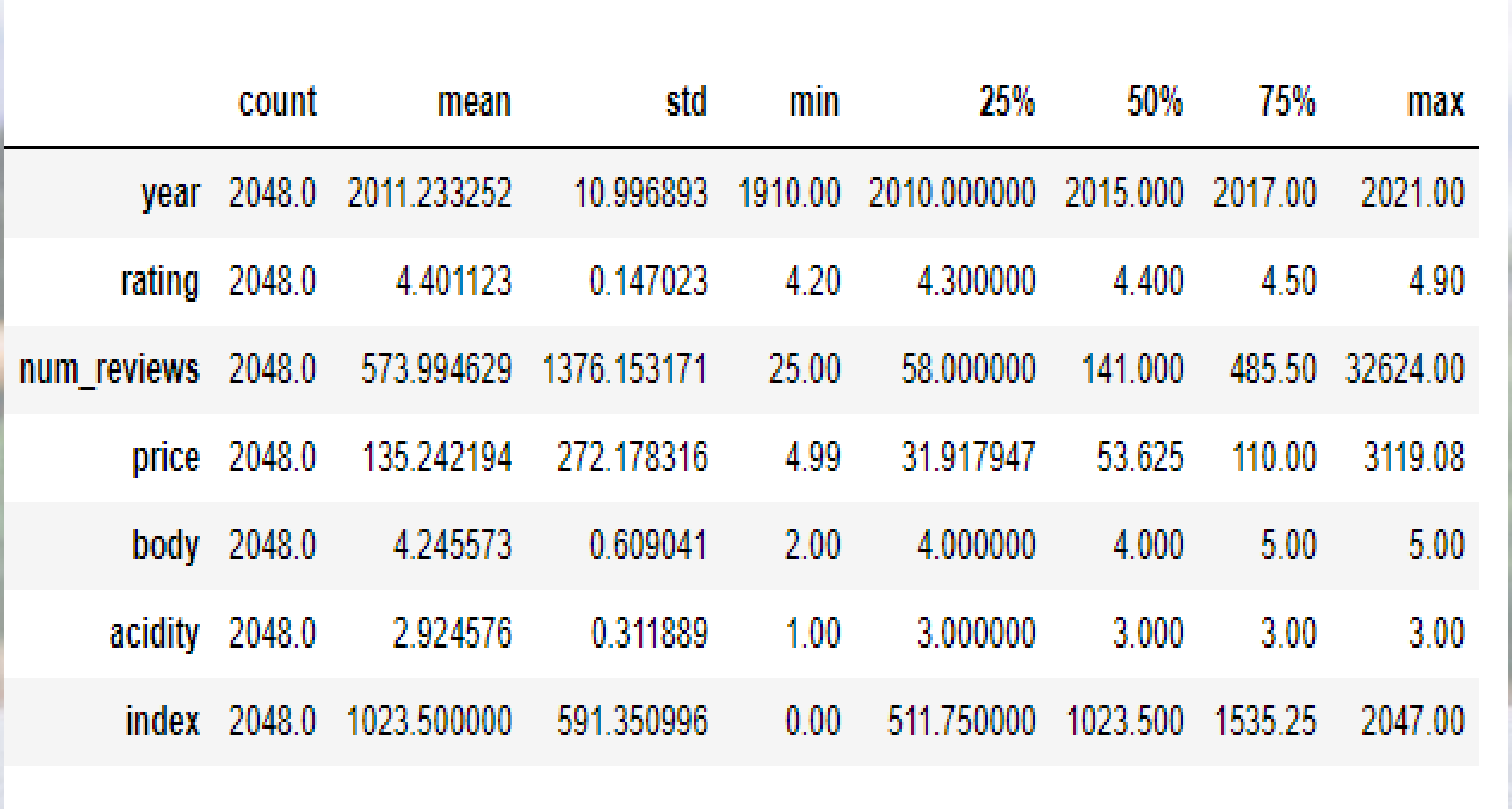
- Create an analytical framework to understand **Key factors impacting Wine prices**
- Develop a modeling framework **To estimate the price of a wine that is up for sale**



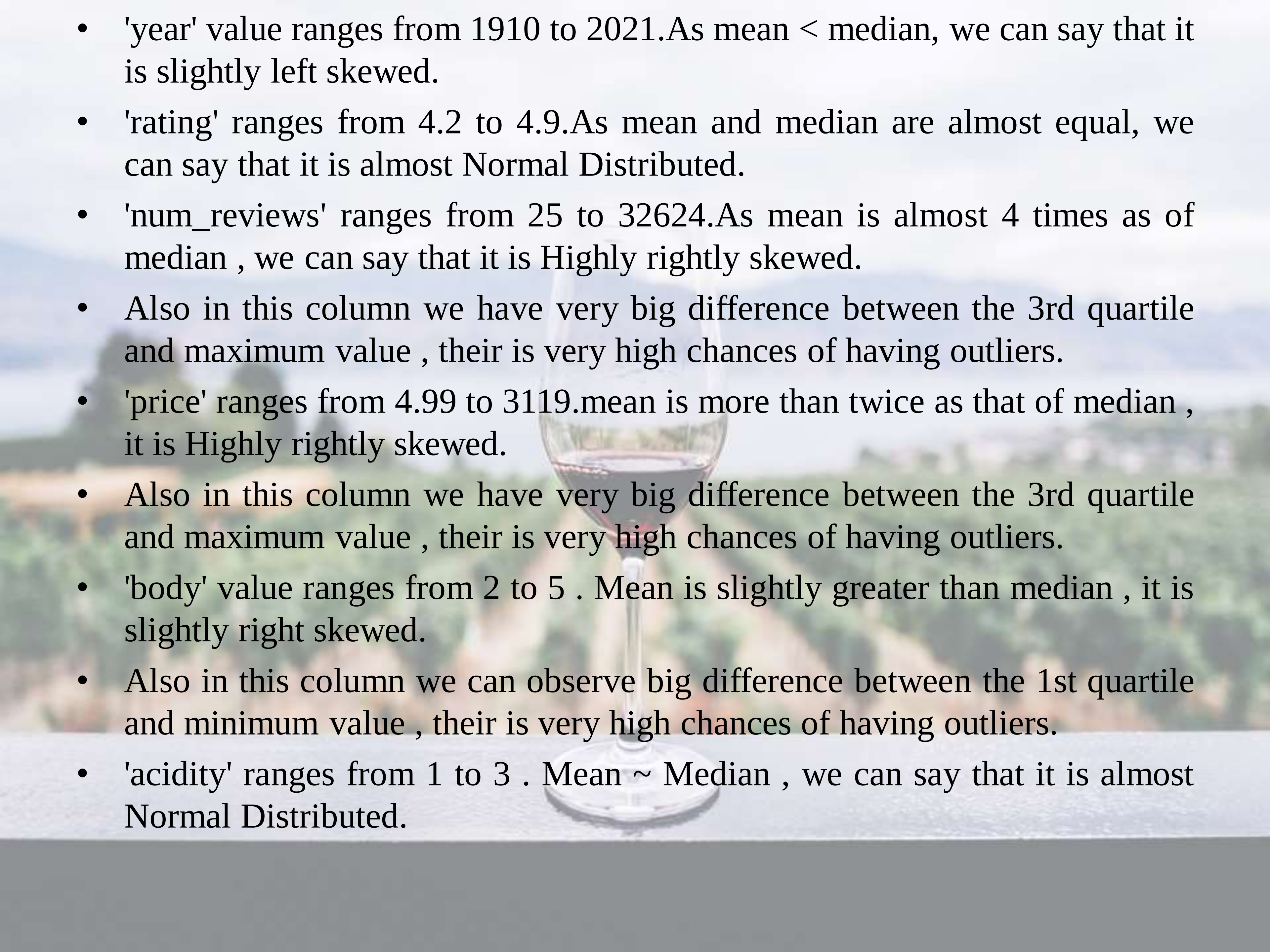
Data Preparation

- **Reading the Dataset-** We have read the data and understand it and then we see the null values and some unwanted values in between the data which we removed in the next steps.
- **Data Cleaning**
- **Variable transformation-** Data type of year column is object , we need to convert it in numerical type. There are null values in the year ,type , body and acidity column.
- **Dealing With Missing Value-** Filling the missing value in the type column with the mode
- Filling the missing value in numerical variable using Median

Analysis Of the Data

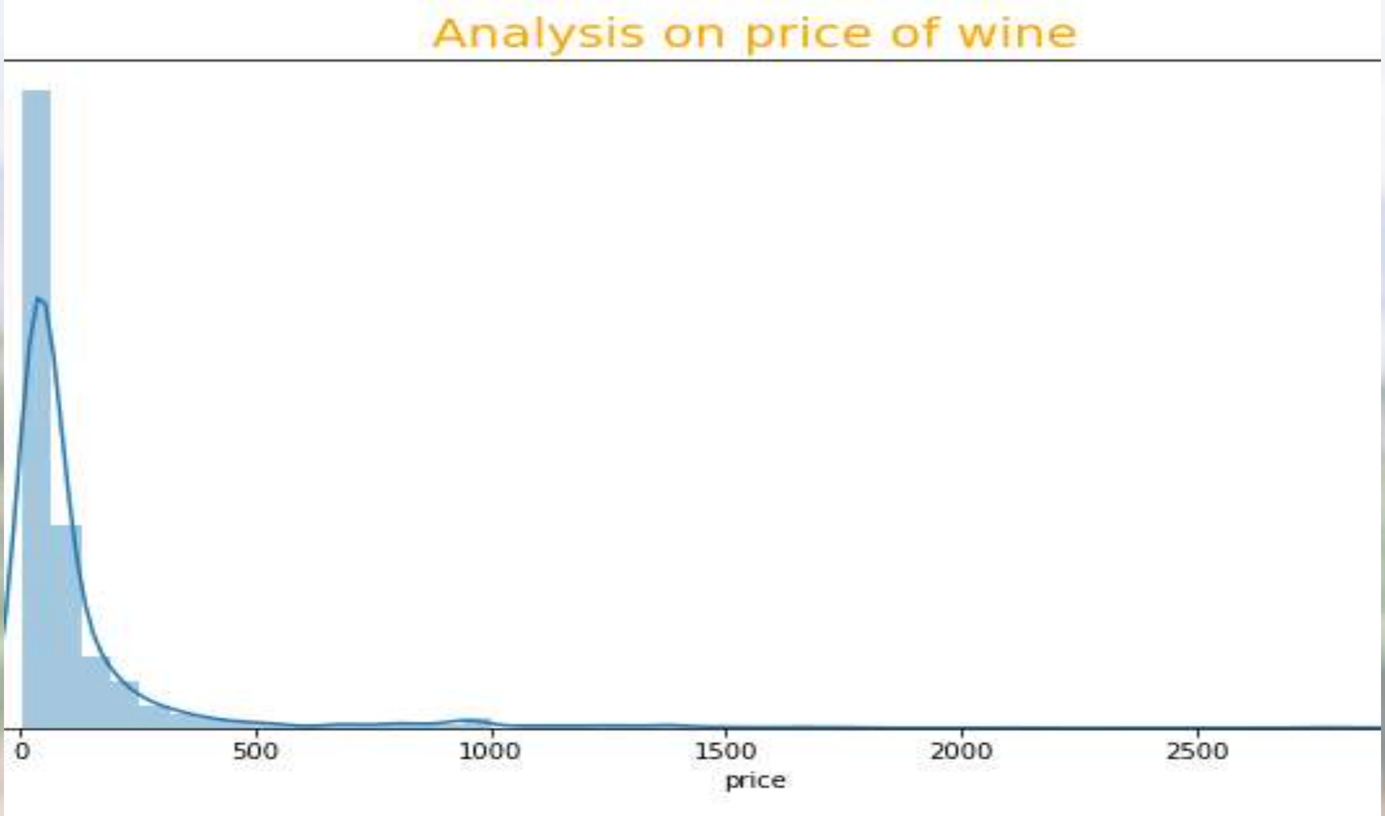


	count	mean	std	min	25%	50%	75%	max
year	2048.0	2011.233252	10.996893	1910.00	2010.000000	2015.000	2017.00	2021.00
rating	2048.0	4.401123	0.147023	4.20	4.300000	4.400	4.50	4.90
num_reviews	2048.0	573.994629	1376.153171	25.00	58.000000	141.000	485.50	32624.00
price	2048.0	135.242194	272.178316	4.99	31.917947	53.625	110.00	3119.08
body	2048.0	4.245573	0.609041	2.00	4.000000	4.000	5.00	5.00
acidity	2048.0	2.924576	0.311889	1.00	3.000000	3.000	3.00	3.00
index	2048.0	1023.500000	591.350996	0.00	511.750000	1023.500	1535.25	2047.00



- 'year' value ranges from 1910 to 2021.As mean < median, we can say that it is slightly left skewed.
- 'rating' ranges from 4.2 to 4.9.As mean and median are almost equal, we can say that it is almost Normal Distributed.
- 'num_reviews' ranges from 25 to 32624.As mean is almost 4 times as of median , we can say that it is Highly rightly skewed.
- Also in this column we have very big difference between the 3rd quartile and maximum value , their is very high chances of having outliers.
- 'price' ranges from 4.99 to 3119.mean is more than twice as that of median , it is Highly rightly skewed.
- Also in this column we have very big difference between the 3rd quartile and maximum value , their is very high chances of having outliers.
- 'body' value ranges from 2 to 5 . Mean is slightly greater than median , it is slightly right skewed.
- Also in this column we can observe big difference between the 1st quartile and minimum value , their is very high chances of having outliers.
- 'acidity' ranges from 1 to 3 . Mean ~ Median , we can say that it is almost Normal Distributed.

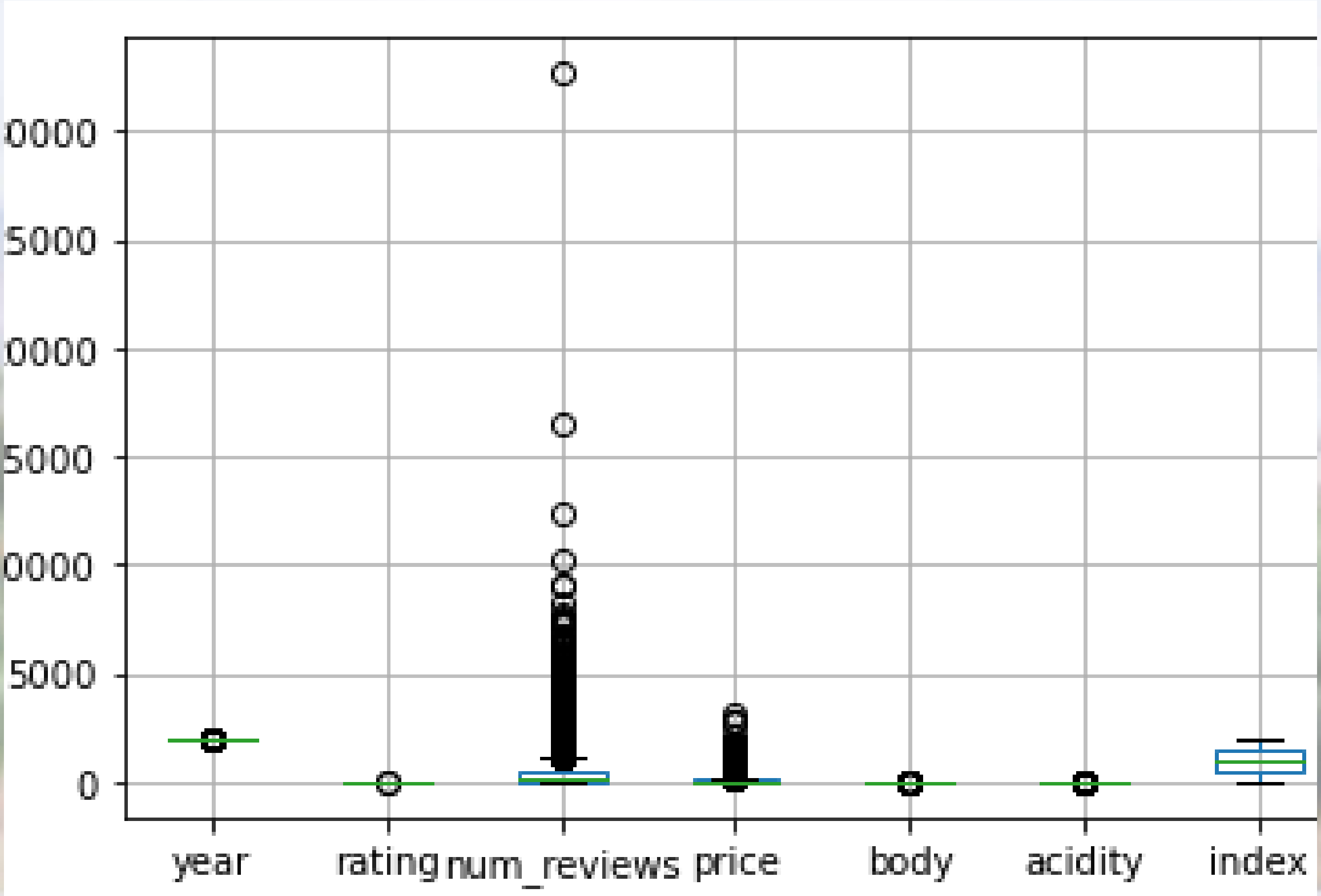
Analysis of Price (Target Variable)



Price of most of the wines less than 500.

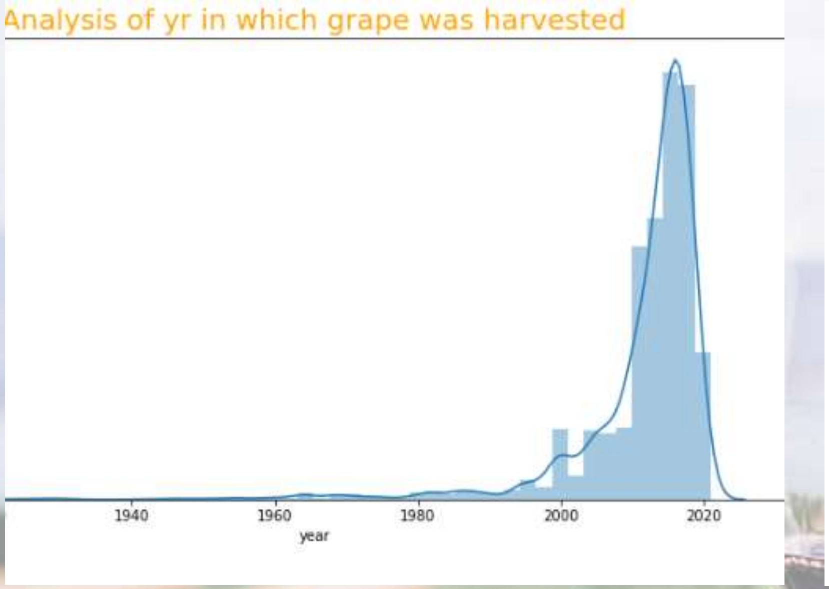
The above graph shows that price has right skewness. And we know that the assumption of linear regression tells us that the distribution of dependent variable has to be normal, so for that reason we had converted it into normal distribution.

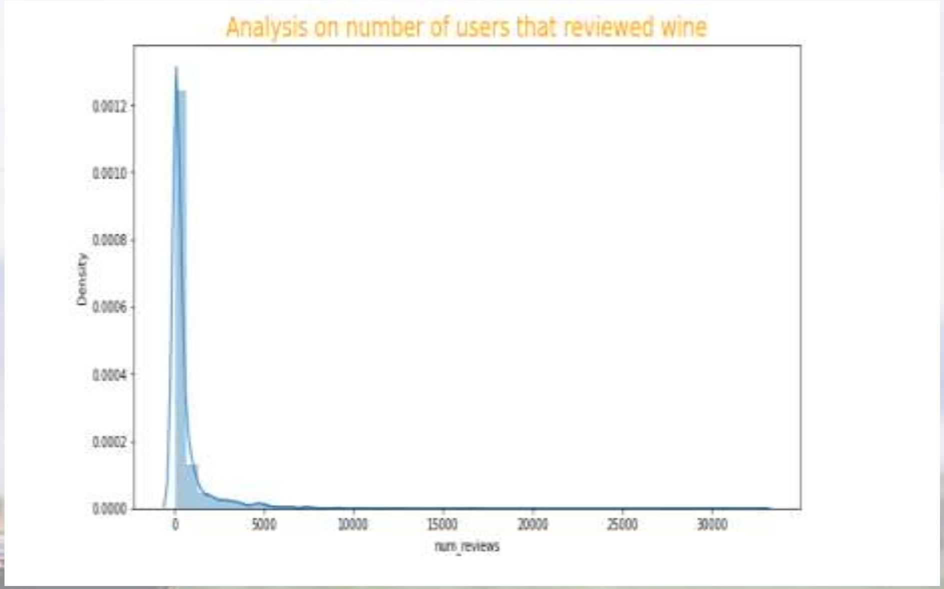
BOX-Plots

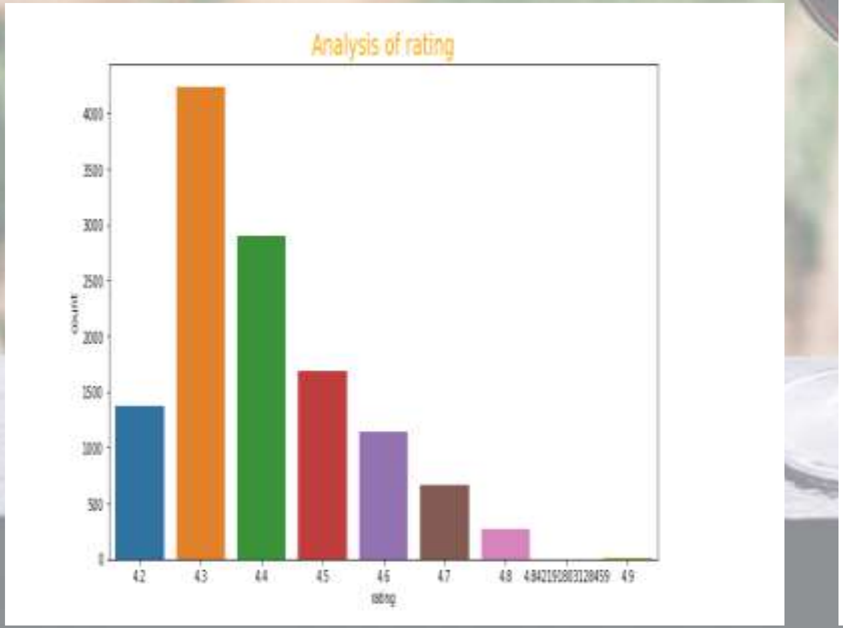


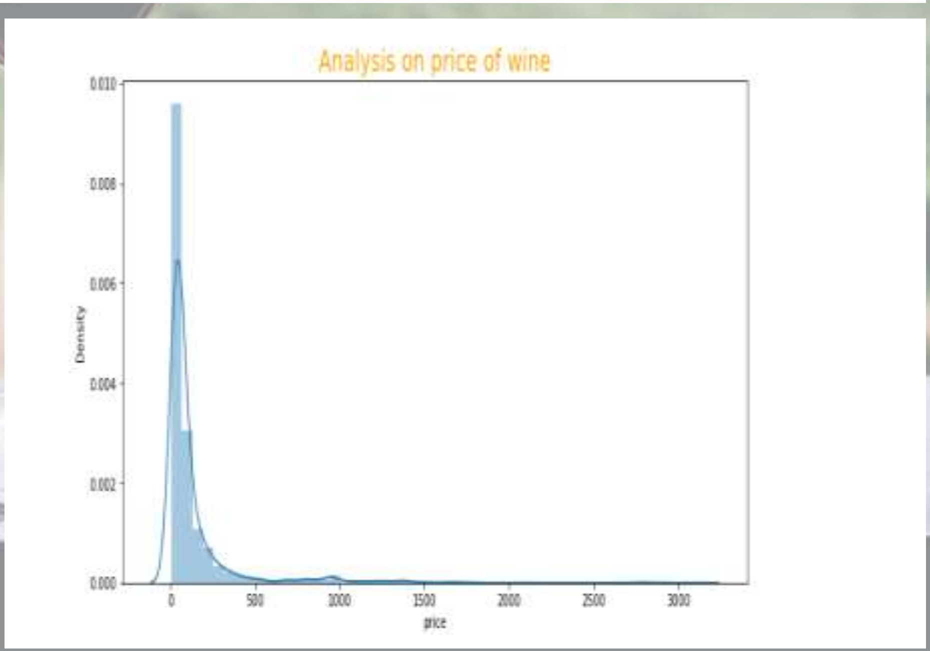
We can see, all of the columns contain outliers, but num_reviews contain most outliers from them.

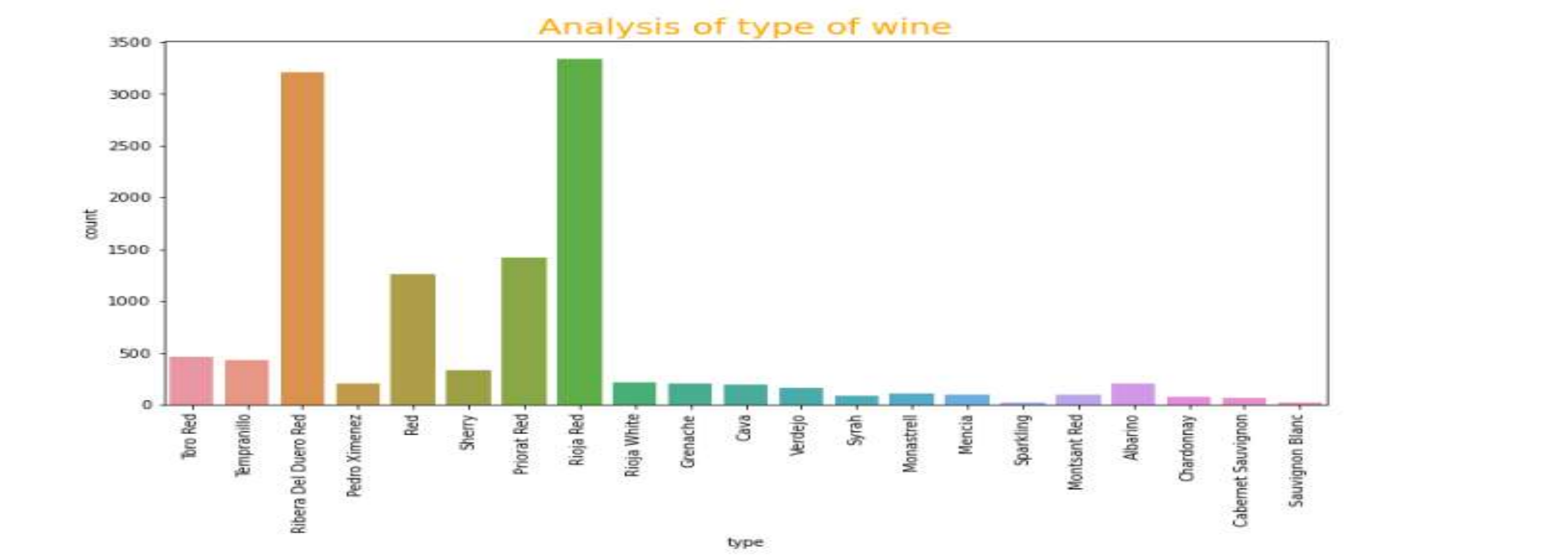
UNIVARIATE ANALYSIS

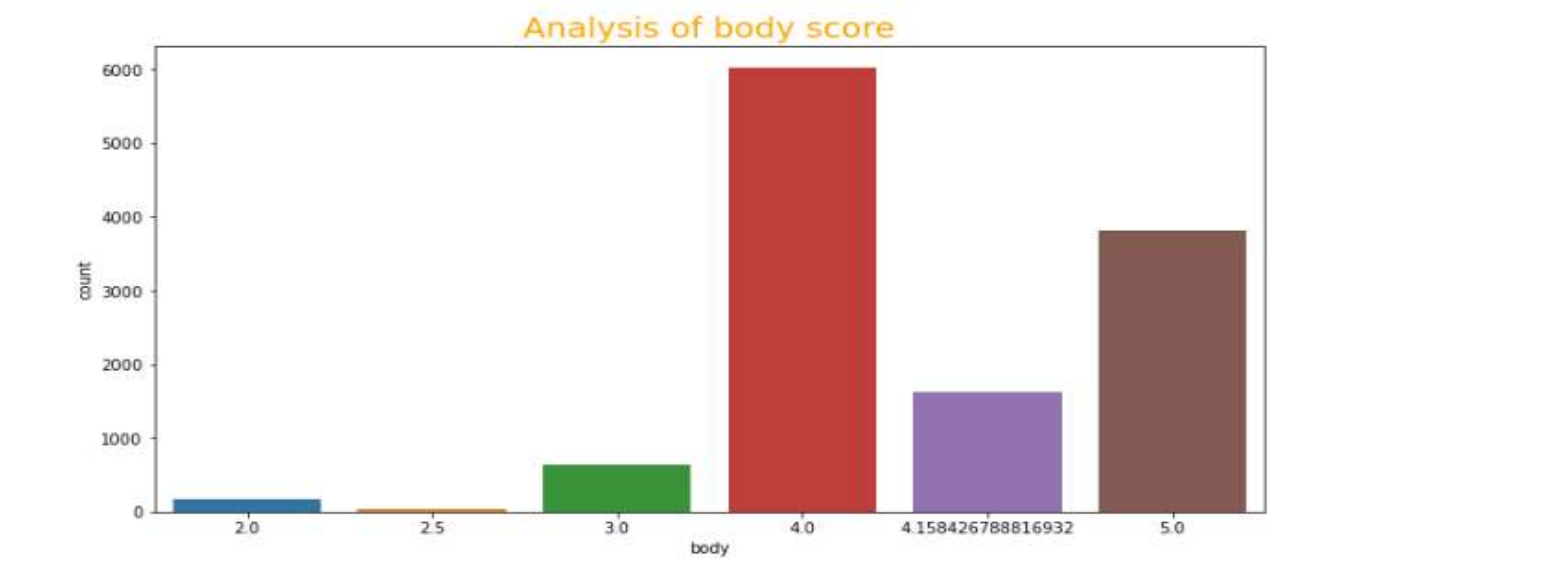

Analysis of yr in which grape was harvested


Analysis on number of users that reviewed wine

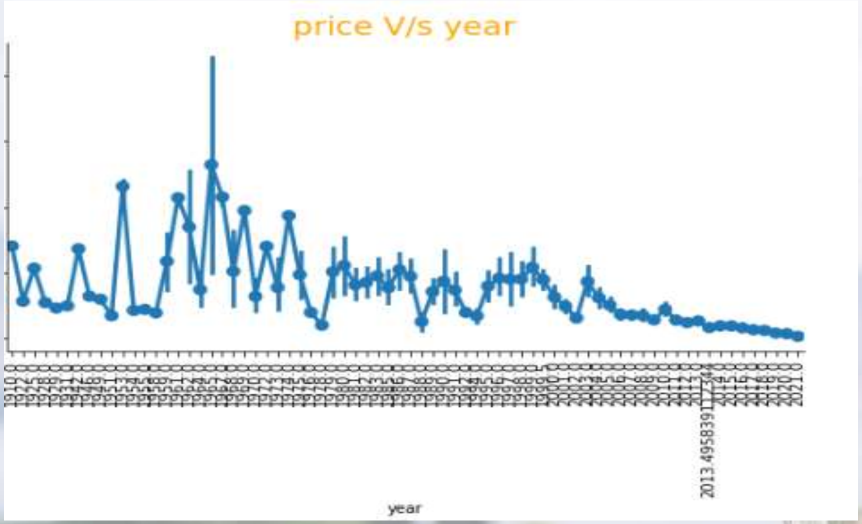

Analysis of rating

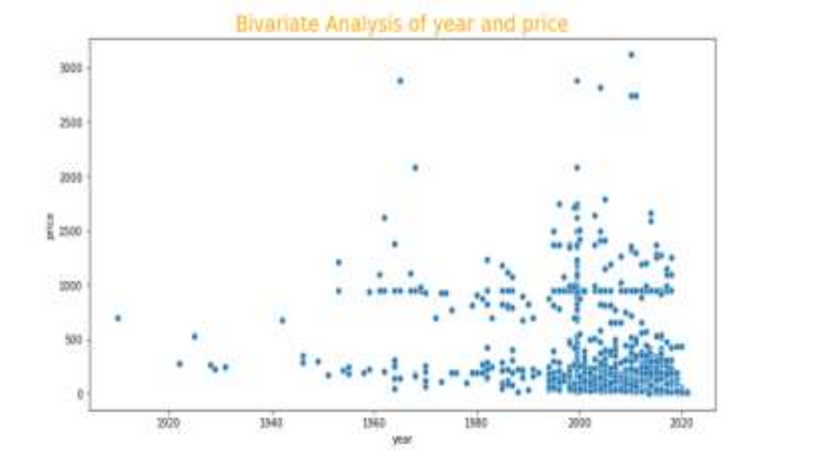

Analysis on price of wine

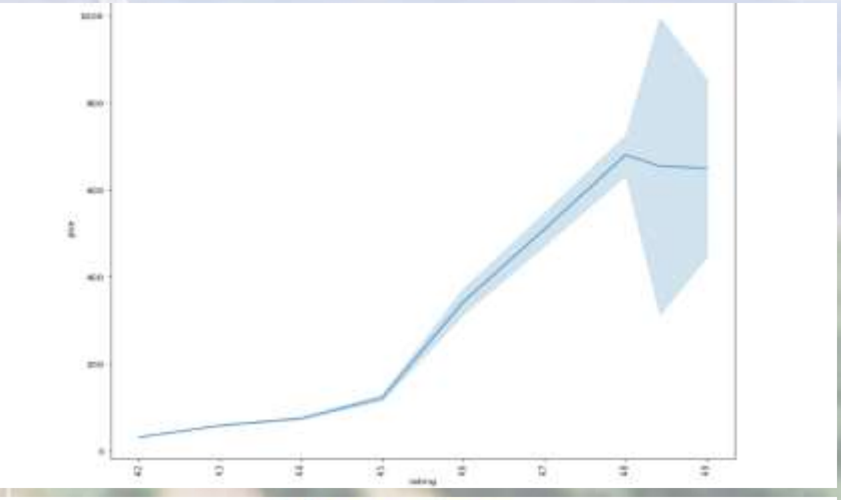

Analysis of type of wine

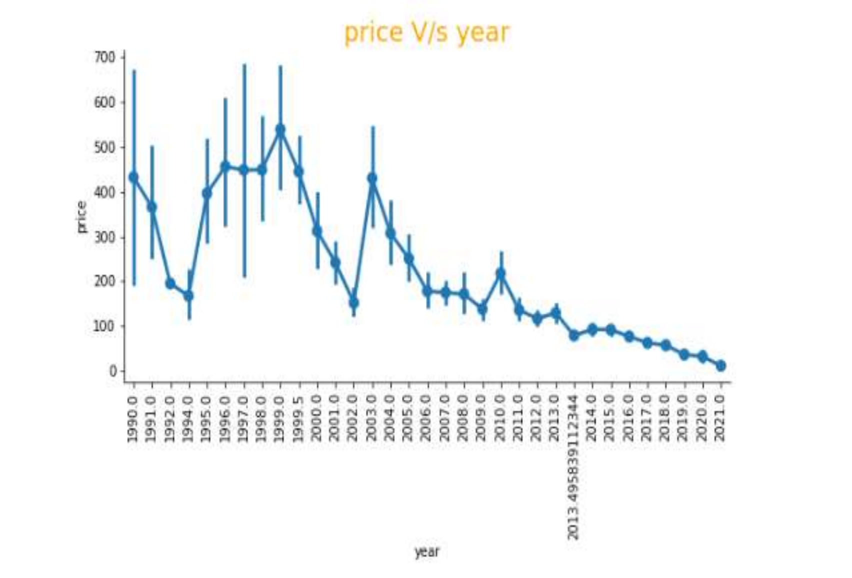

Analysis of body score

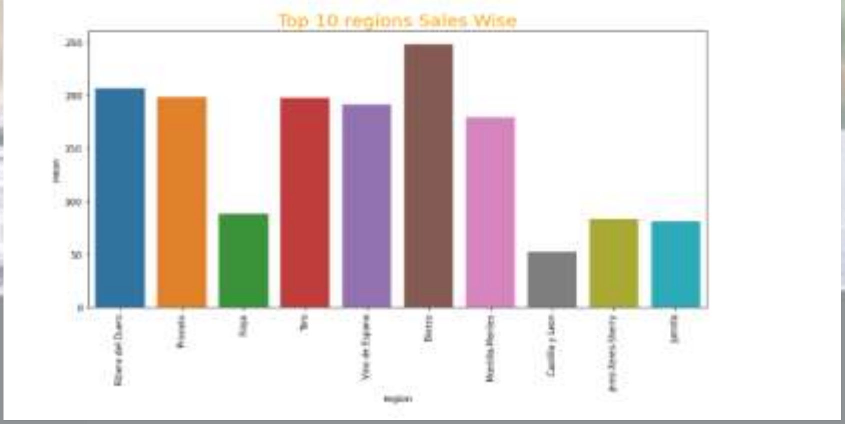
BIVARIATE ANALYSIS

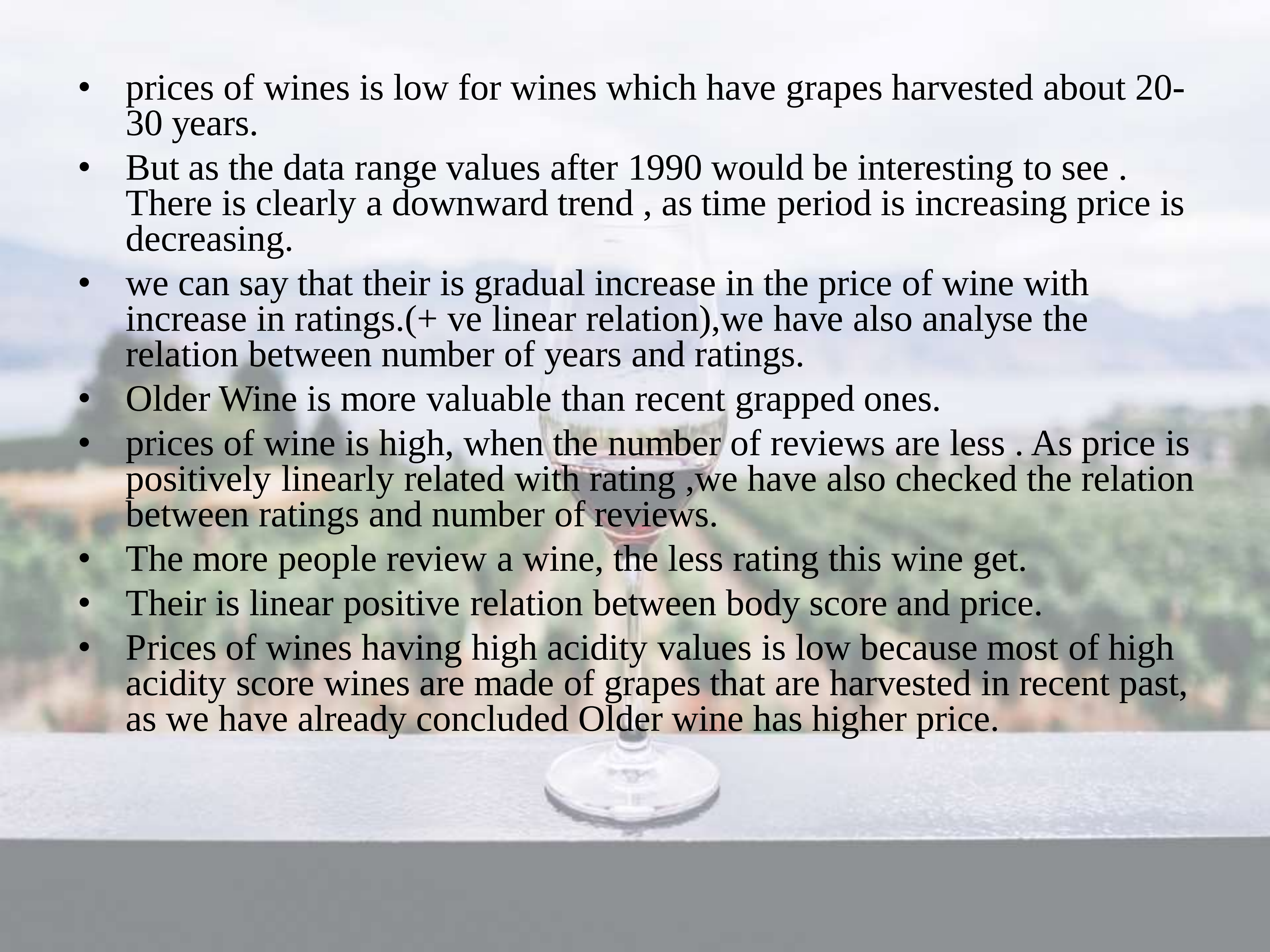






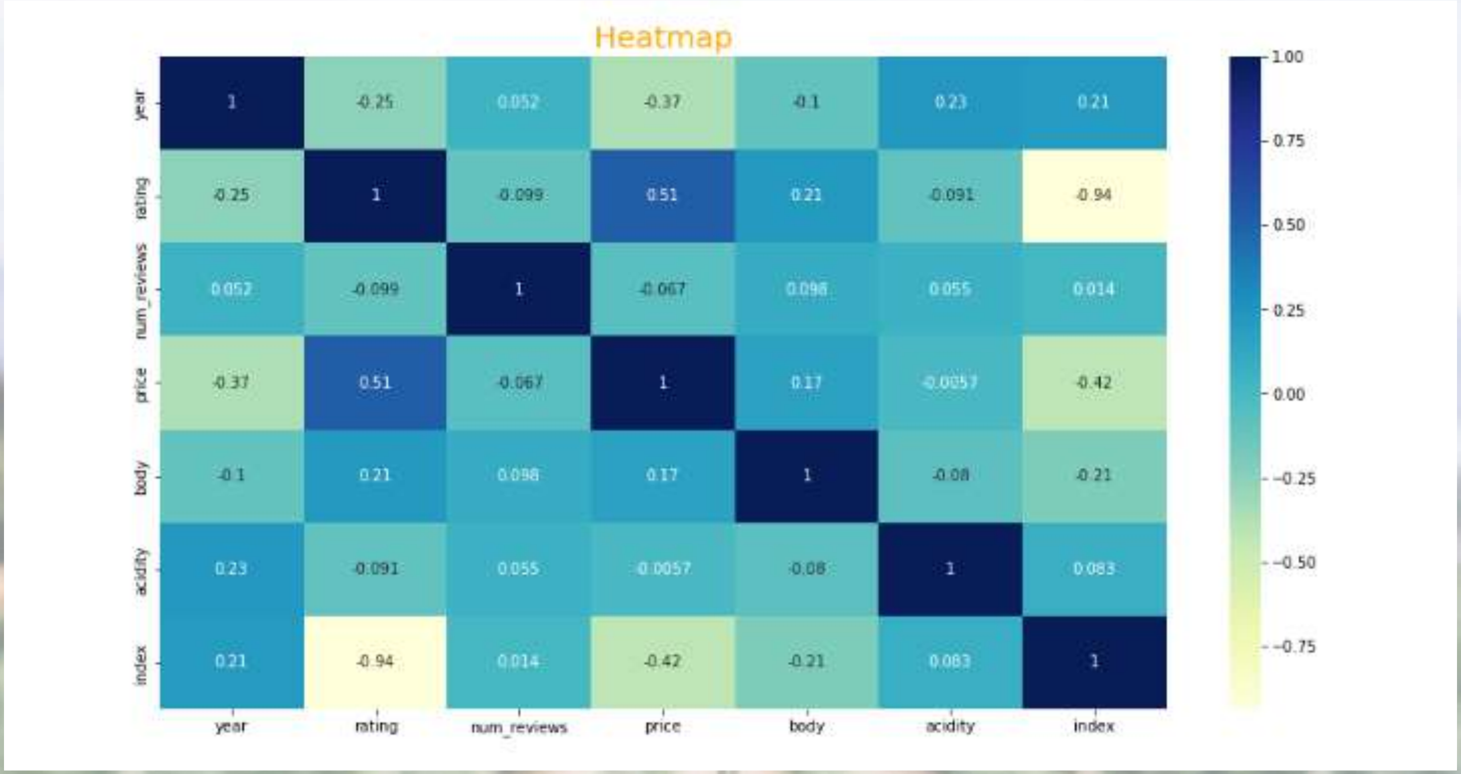






- prices of wines is low for wines which have grapes harvested about 20-30 years.
- But as the data range values after 1990 would be interesting to see . There is clearly a downward trend , as time period is increasing price is decreasing.
- we can say that their is gradual increase in the price of wine with increase in ratings.(+ ve linear relation),we have also analyse the relation between number of years and ratings.
- Older Wine is more valuable than recent grapped ones.
- prices of wine is high, when the number of reviews are less . As price is positively linearly related with rating ,we have also checked the relation between ratings and number of reviews.
- The more people review a wine, the less rating this wine get.
- Their is linear positive relation between body score and price.
- Prices of wines having high acidity values is low because most of high acidity score wines are made of grapes that are harvested in recent past, as we have already concluded Older wine has higher price.

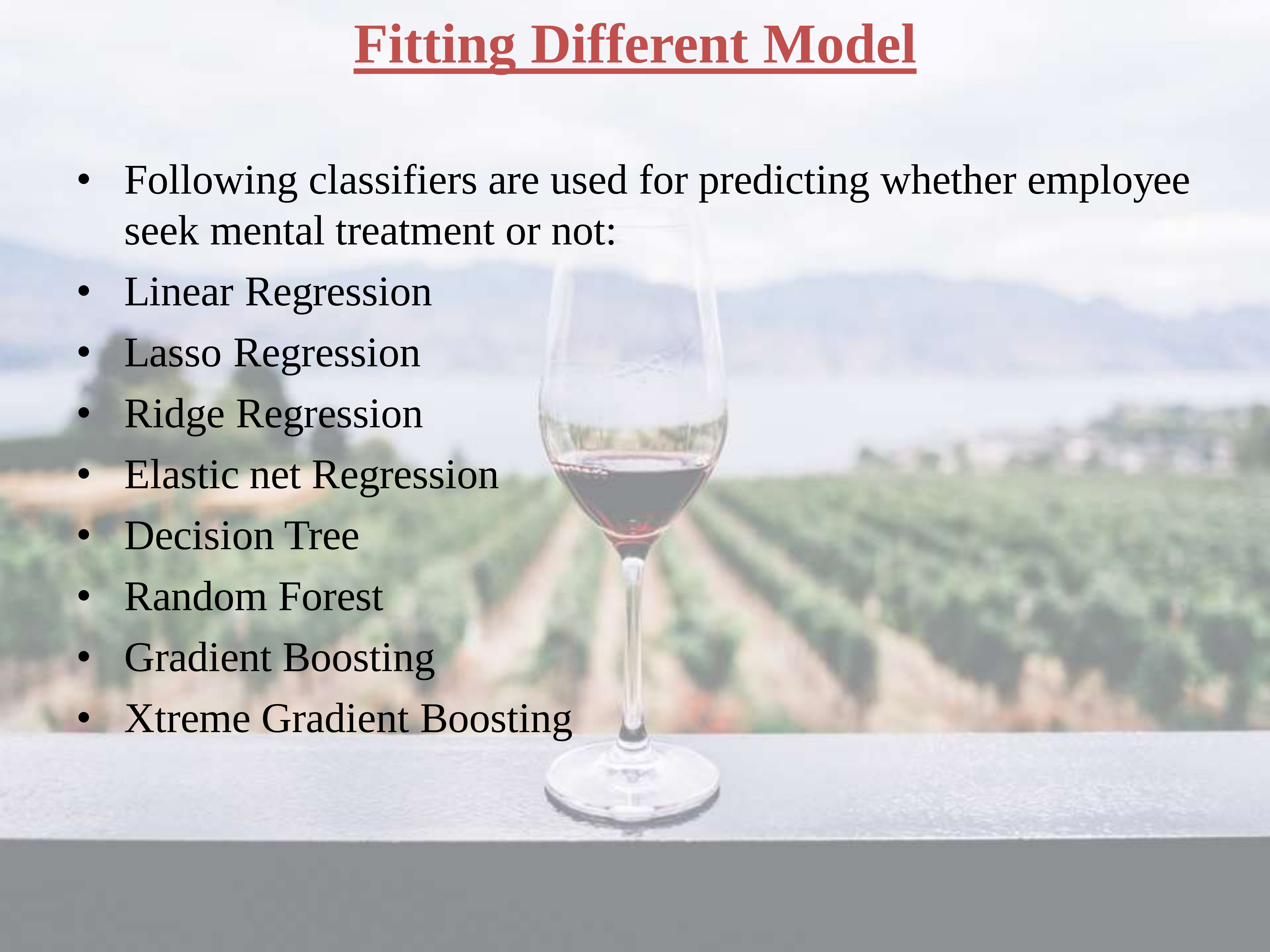
Heat-map



	year	rating	num_reviews	price	body	acidity	index
year	1	-0.25	0.052	-0.37	-0.1	0.23	0.21
rating	-0.25	1	-0.099	0.51	0.21	-0.091	-0.94
num_reviews	0.052	-0.099	1	-0.067	0.098	0.055	0.014
price	-0.37	0.51	-0.067	1	0.17	-0.0057	-0.42
body	-0.1	0.21	0.098	0.17	1	-0.08	-0.21
acidity	0.23	-0.091	0.055	-0.0057	-0.08	1	0.083
index	0.21	-0.94	0.014	-0.42	-0.21	0.083	1

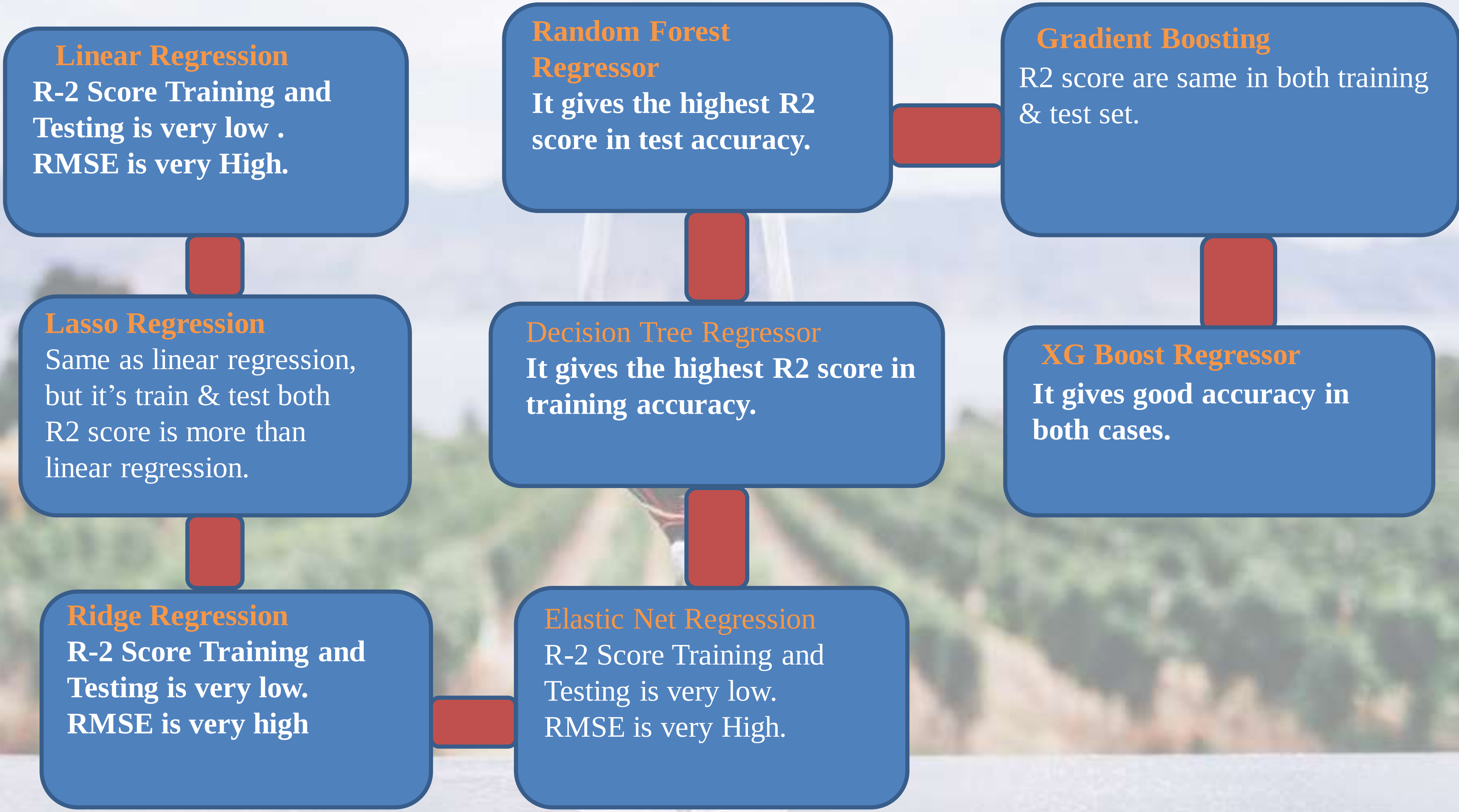
Their is high correlation between price and rating as already discussed.
We also have moderate correlation between year and acidity.
Their is also moderate relation between rating and year.
Their is high - ve correlation between price and year.

Fitting Different Model



- Following classifiers are used for predicting whether employee seek mental treatment or not:
- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic net Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Xtreme Gradient Boosting

Model Approaches Used & Why



Linear Regression

R-2 Score Training and Testing is very low . RMSE is very High.

Lasso Regression

Same as linear regression, but it’s train & test both R2 score is more than linear regression.

Ridge Regression

R-2 Score Training and Testing is very low. RMSE is very high

Elastic Net Regression

R-2 Score Training and Testing is very low. RMSE is very High.

Decision Tree Regressor

It gives the highest R2 score in training accuracy.

Random Forest Regressor

It gives the highest R2 score in test accuracy.

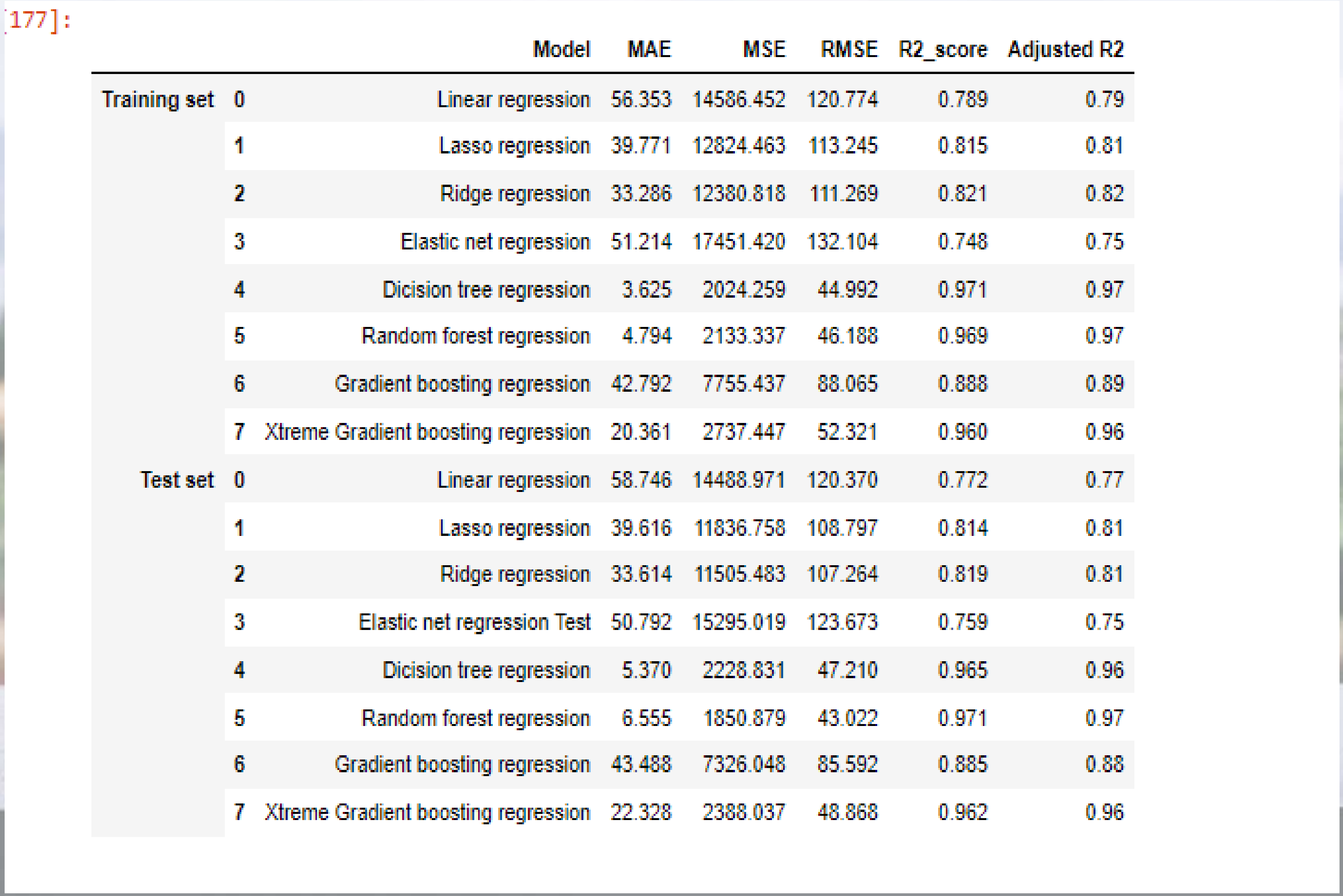
Gradient Boosting

R2 score are same in both training & test set.

XG Boost Regressor

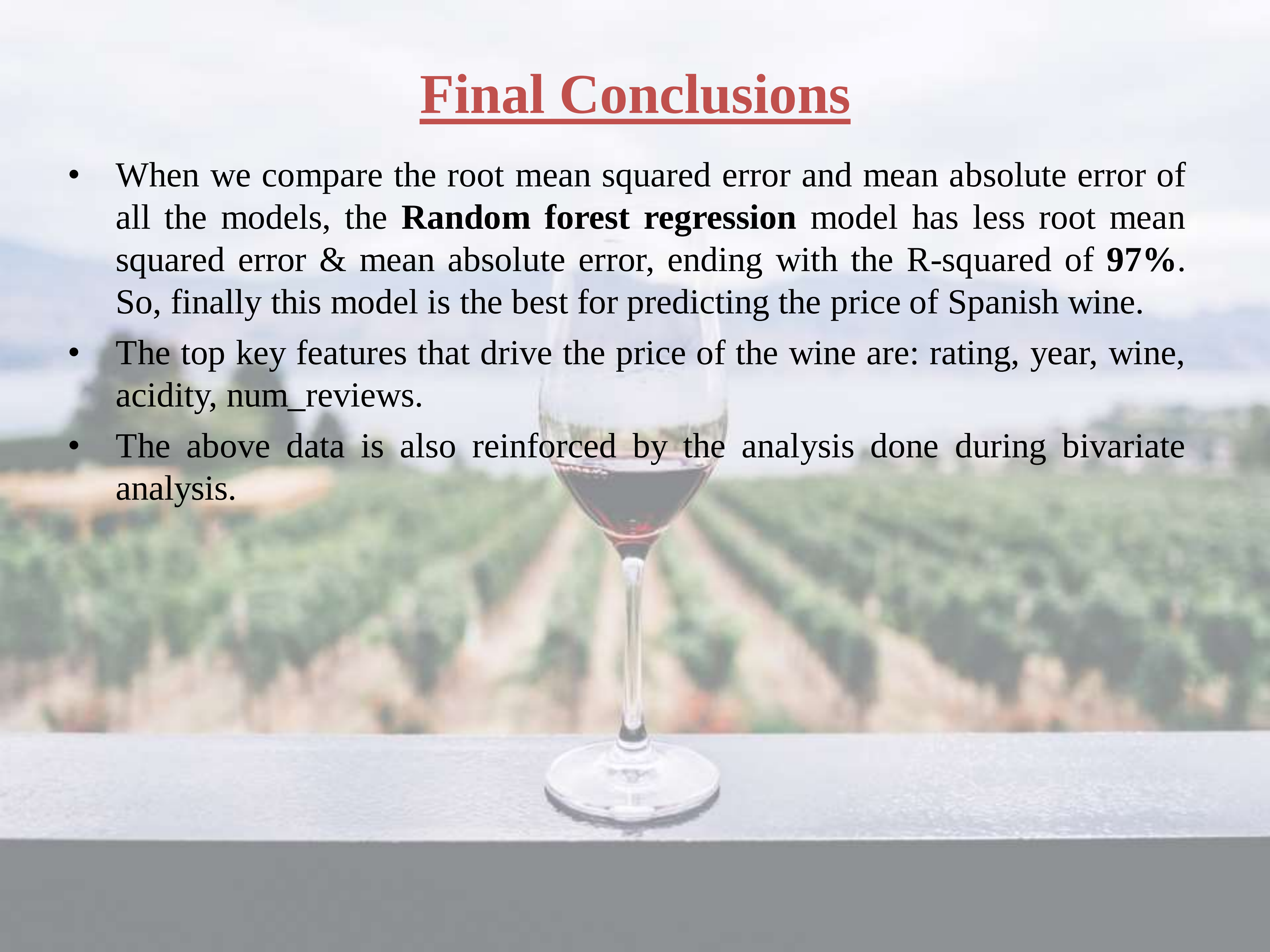
It gives good accuracy in both cases.

Performance Metrices



[177]:

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	56.353	14586.452	120.774	0.789	0.79
	1	Lasso regression	39.771	12824.463	113.245	0.815	0.81
	2	Ridge regression	33.286	12380.818	111.269	0.821	0.82
	3	Elastic net regression	51.214	17451.420	132.104	0.748	0.75
	4	Dicision tree regression	3.625	2024.259	44.992	0.971	0.97
	5	Random forest regression	4.794	2133.337	46.188	0.969	0.97
	6	Gradient boosting regression	42.792	7755.437	88.065	0.888	0.89
	7	Xtreme Gradient boosting regression	20.361	2737.447	52.321	0.960	0.96
Test set	0	Linear regression	58.746	14488.971	120.370	0.772	0.77
	1	Lasso regression	39.616	11836.758	108.797	0.814	0.81
	2	Ridge regression	33.614	11505.483	107.264	0.819	0.81
	3	Elastic net regression Test	50.792	15295.019	123.673	0.759	0.75
	4	Dicision tree regression	5.370	2228.831	47.210	0.965	0.96
	5	Random forest regression	6.555	1850.879	43.022	0.971	0.97
	6	Gradient boosting regression	43.488	7326.048	85.592	0.885	0.88
	7	Xtreme Gradient boosting regression	22.328	2388.037	48.868	0.962	0.96



Final Conclusions

- When we compare the root mean squared error and mean absolute error of all the models, the **Random forest regression** model has less root mean squared error & mean absolute error, ending with the R-squared of **97%**. So, finally this model is the best for predicting the price of Spanish wine.
- The top key features that drive the price of the wine are: rating, year, wine, acidity, num_reviews.
- The above data is also reinforced by the analysis done during bivariate analysis.

THANK YOU