WINE PRICE PREDICTION

THANDOLWETHU SILOYI

Business Problem

I have been provided information of different types of wines from different regions in Spain.

I want to know the factors that affect the pricing of wines in Spain

Stakeholders

- Distributors and Retailers
- Customers
- Wine producers
- Local Community
- Suppliers
- Investors
- Competitors

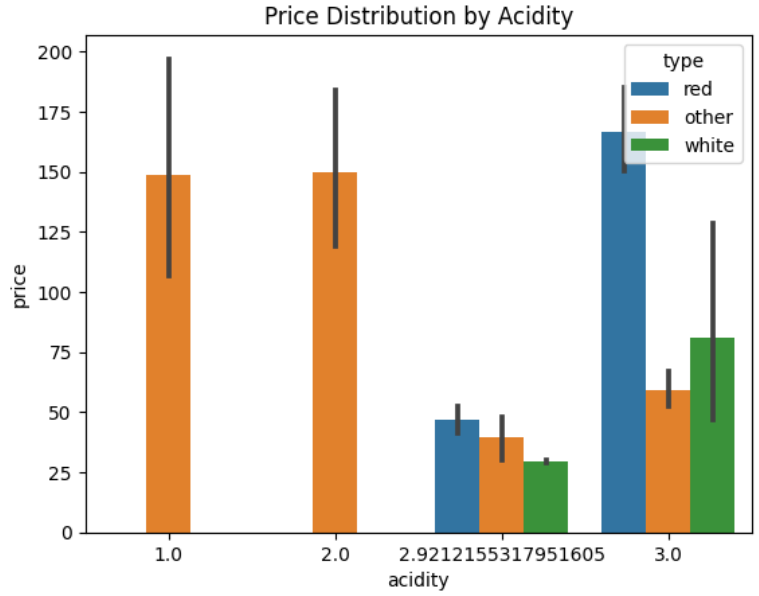
Brief Introduction of the Data

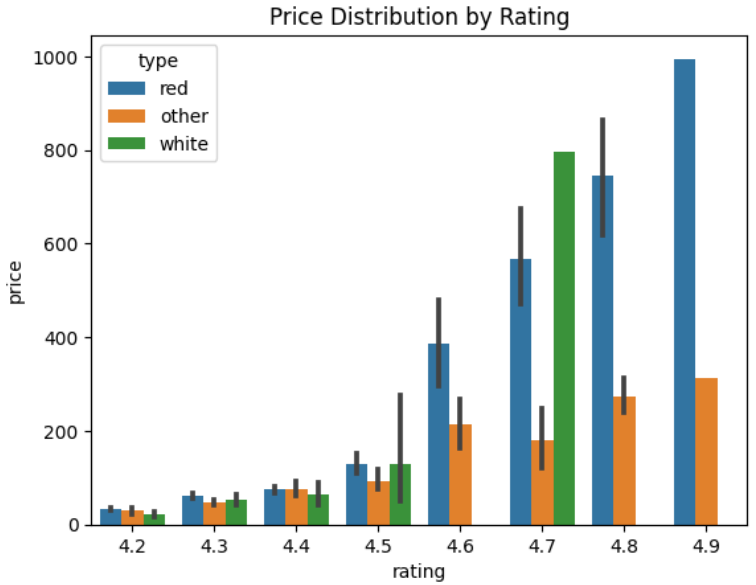
- The dataset contains 7500 different types of red wines from Spain with 11 features that describe the price, rating, and even some flavour description
- Winery name.
- Name of the wine.
- Year in which the grapes were harvested.
- Average rating given to the wine by users [from 1 – 5].
- Number of users that reviews the wine
- Country of origin
- Region of the wine
- Price in Euros
- Wine variety
- Body score, defines as the richness and weight of the wine in your mouth [from 1 – 5]
- Acidity score, defined as wine's “pucker’ or tartness; it is what makes refreshing and your tongue salivate and want another sip [form 1 -5]

Finding from the Data

- We have wines with less acidity and they are expensive.
- We also have wines with higher acidity that are affordable

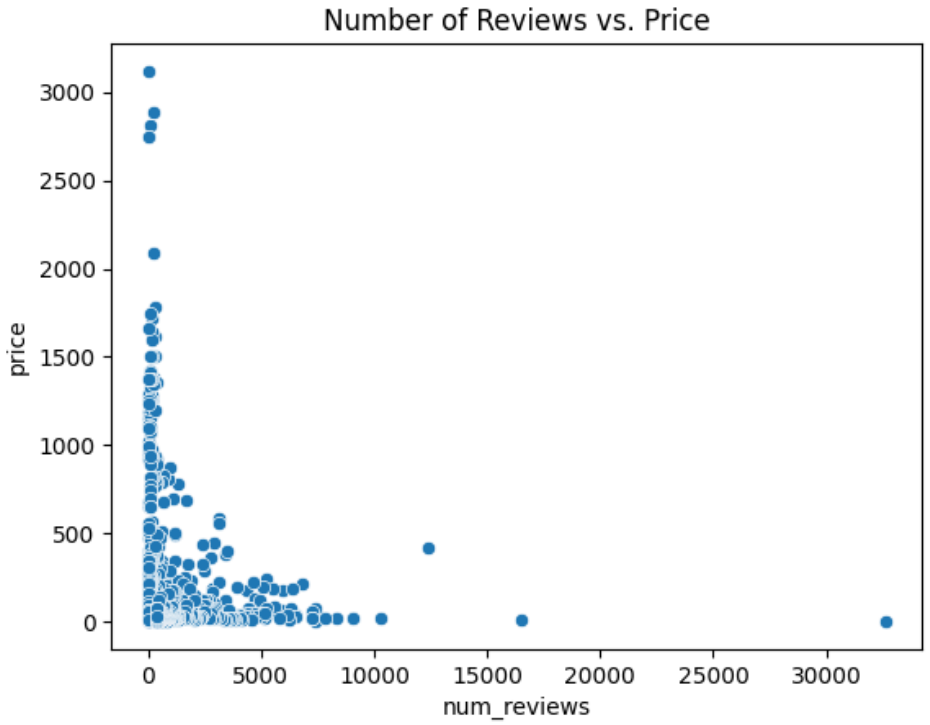
- The higher the price the higher ratings.
- The lower the price the lower the ratings.
- Red wines have more ratings followed by white and other wines



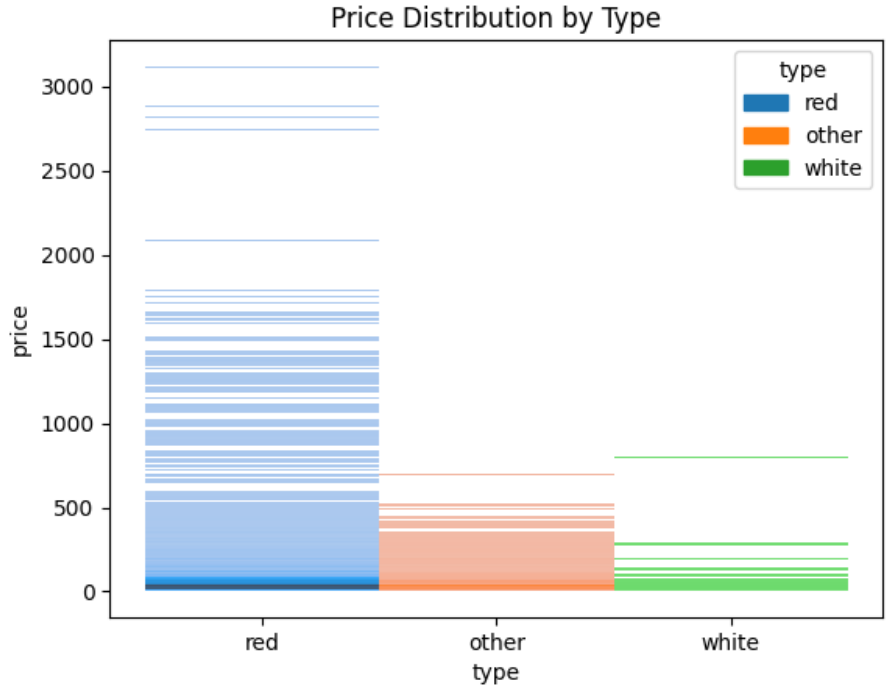


Finding from the Data…

- The less the number of reviews the higher the Price of wines.
- The higher the number of reviews the lesser the Price of wines.



- Within each wine there's a wide range of prices
- Red wines have many expensive wines.
- The most common prices are less than 500 Euros



Model used

- Random Forest Regression
- K-Nearest Neighbors Model
- PCA

Model Performances

Default	Training Data			
Random Forest Regression Model	MAE	MSE	RMSE	R^2
	24.455	5,060.281	71.136	0.932
	Test Data			
	MAE	MSE	RMSE	R^2
	69.673	27,614.264	166.175	0.616
GridSearchCV-tune				
	Training Data			
Random Forest Regression Model	52.472	14,794.261	121.632	0.801
	Test Data			
	72.935	23,390.056	152.938	0.675

Model Performances…

Default
KNN Model

Training Data

MAE	MSE	RMSE	R^2
92.813	28,129.827	167.719	0.622

Test Data

MAE	MSE	RMSE	R^2
127.842	43,361.937	208.235	0.397

GridSearchCV-tune

KNN Model

Training Data

92.813	28,129.827	167.719	0.622

Test Data

127.842	43,361.937	208.235	0.397

Random Forest

PCA
Random Forest Regression

Training Data

MAE	MSE	RMSE	R^2
24.393	4,107.402	64.089	0.945

Test Data

MAE	MSE	RMSE	R^2
71.806	31,660.225	177.933	0.560

Conclusion

- Random Forest performed well compared to KNN Model.
- Based on the comparison of Features and the price.
- We can assume that there are different factors that affect the pricing of wine such as rating, number of reviews, wine type, acidity, body, etc.