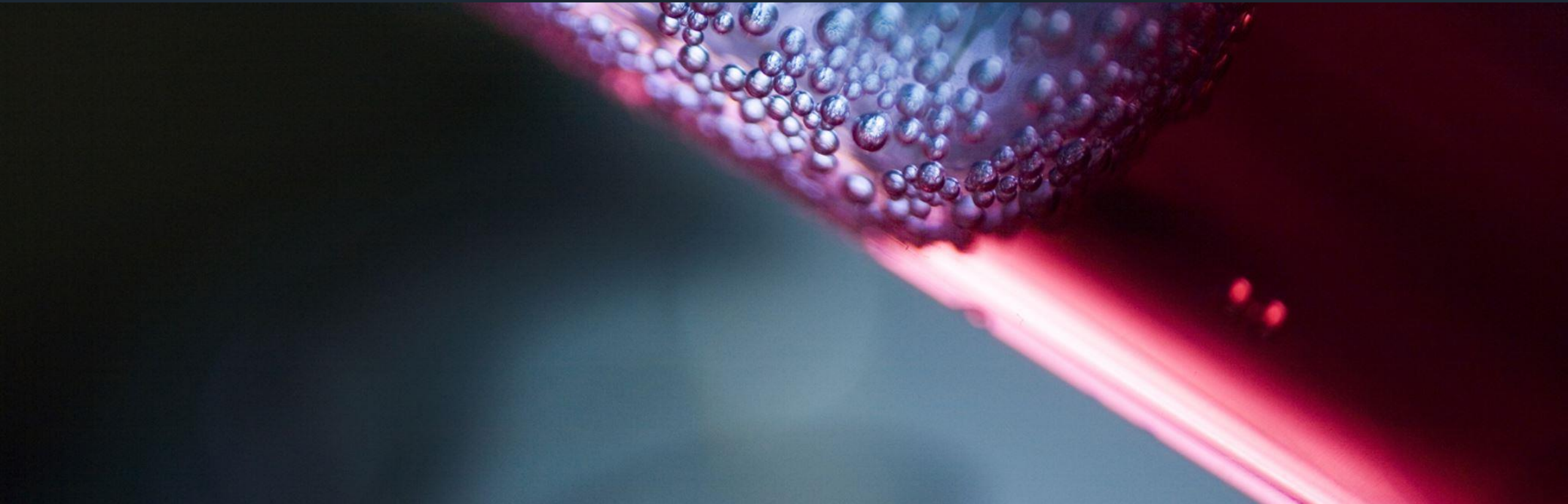# *Improving Wine Quality through Data Science*

Name: Gabriel Pantoja

Data Overview

	winery	wine	year	rating	num_reviews	country	region	price	type	body	acidity
0	Teso La Monja	Tinto	2013	4.9	58	Espana	Toro	995.00	Toro Red	5.0	3.0
1	Artadi	Vina El Pison	2018	4.9	31	Espana	Vino de Espana	313.50	Tempranillo	4.0	2.0
2	Vega Sicilia	Unico	2009	4.8	1793	Espana	Ribera del Duero	324.95	Ribera Del Duero Red	5.0	3.0
3	Vega Sicilia	Unico	1999	4.8	1705	Espana	Ribera del Duero	692.96	Ribera Del Duero Red	5.0	3.0
4	Vega Sicilia	Unico	1996	4.8	1309	Espana	Ribera del Duero	778.06	Ribera Del Duero Red	5.0	3.0

→ Data source: Public dataset from Kaggle

→ Contains physicochemical properties of red and white wines from Spain

→ Includes measurements like:

  → Acidity

  → Sugar

  → Chlorides

  → Sulphates

→ Also includes wine quality rating by experts on scale of 0 (worst) to 10 (best)

→ 6497 observations of red and white wines

→ 11 variables per wine

→ Goal is to analyze chemical properties that drive quality

1st Visual

→ This bar chart displays the mean price for each wine type in the dataset. The height of each bar represents the average price of wines of that type. There are noticeable differences in average price between various wine types.
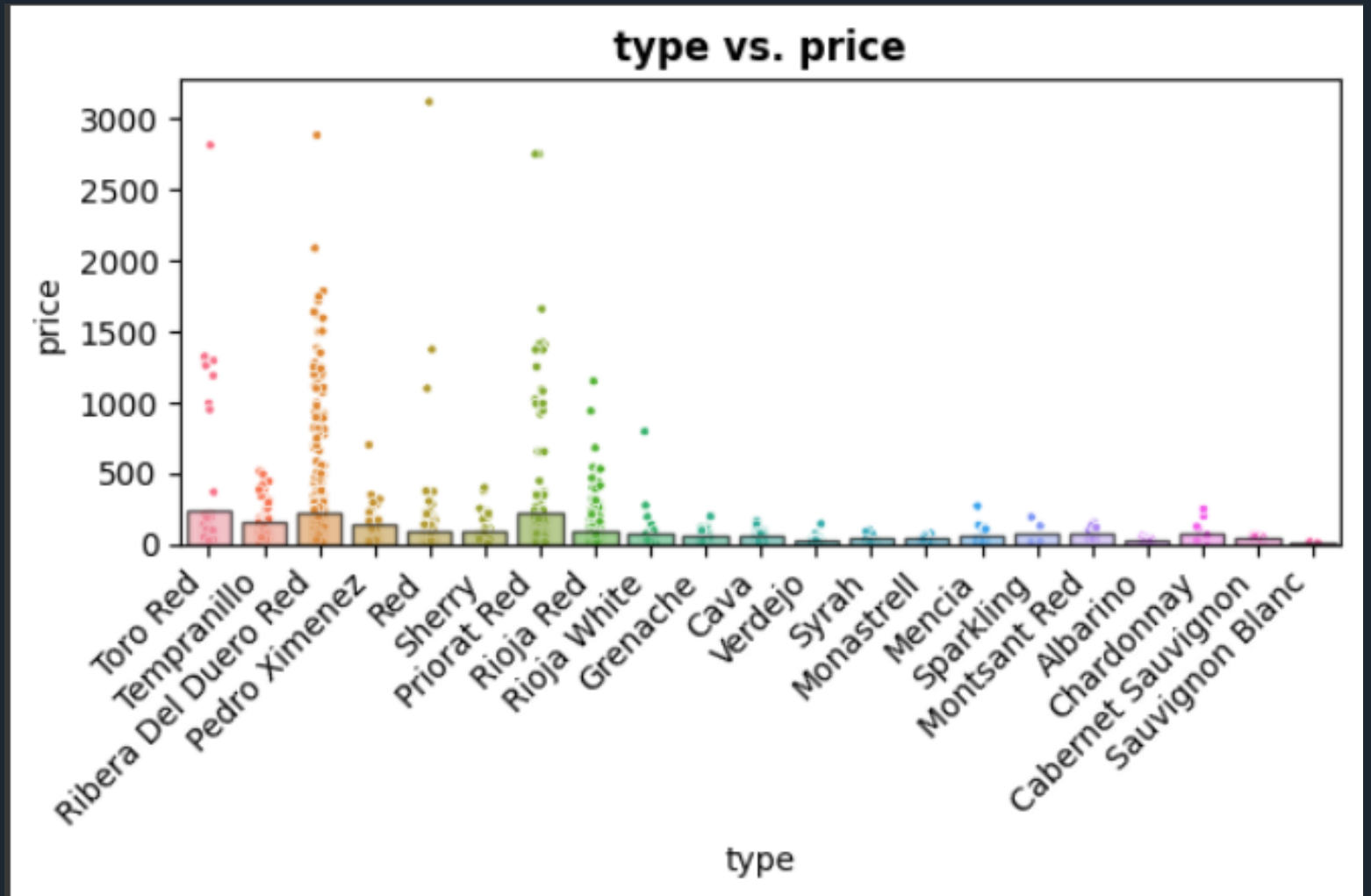
2nd Visual

→ This scatterplot shows the relationship between acidity levels and price for the wines in the dataset. Each point represents a single wine sample. There is a clear positive correlation between acidity and price.

Relationship Between Acidity and Price
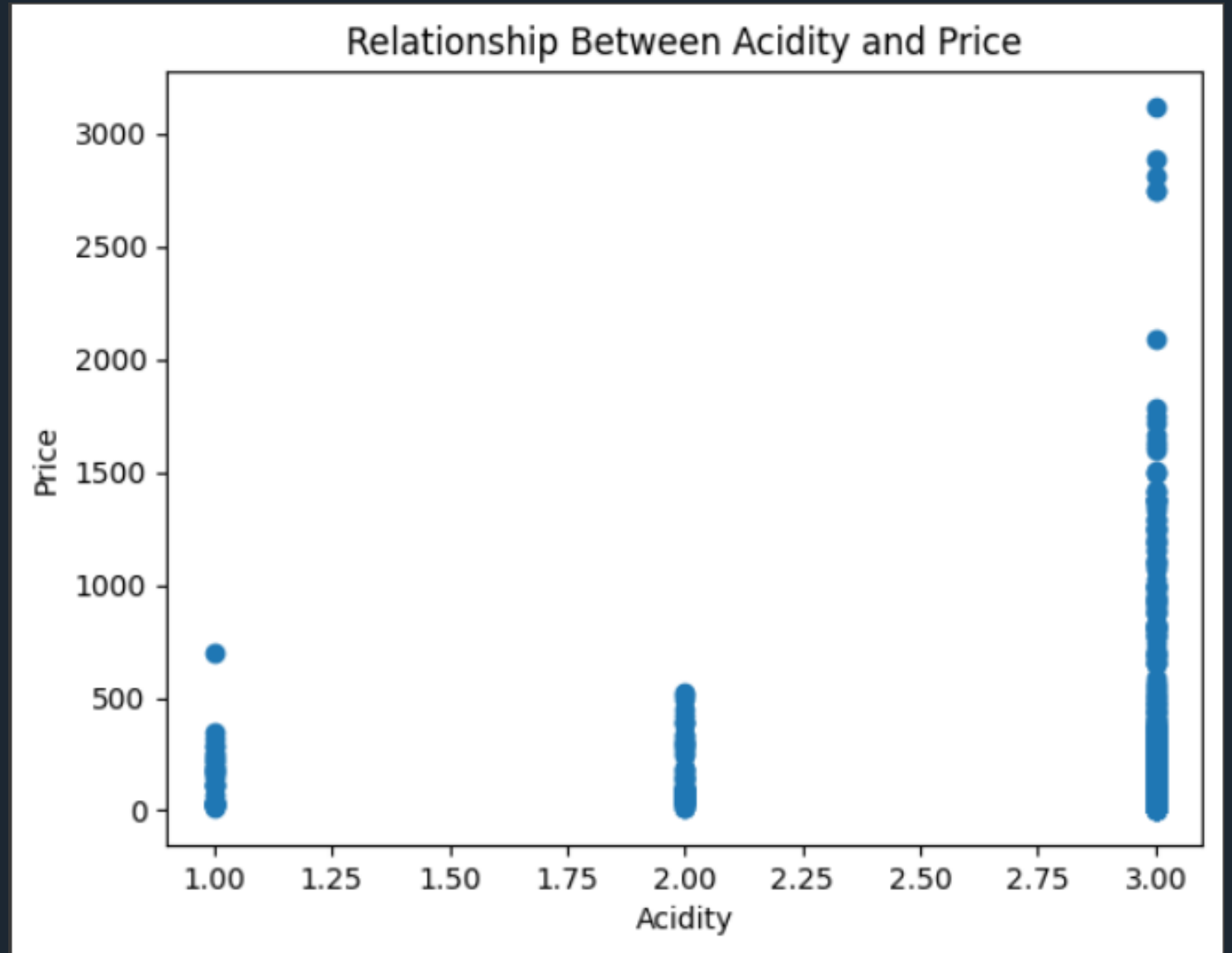
strengths and limitation

→ Used Random Forest and KNN regression models

→ Random Forest R-squared:

  → 0.67 on training data

  → 0.43 on test data

→ Indicates moderate fit, room for improvement

→ KNN with PCA pre-processing achieved MSE of 71,031 on test data

→ Limitations:

  → Moderate R-squared shows imperfect fit

  → Overfitting risk based on train/test difference

  → Does not account for all factors affecting quality

  → For winemakers:

→ Models capture some drivers of quality

→ But unpredictability means need to refine models

→ Use models to guide production changes and gauge impact on quality

Final Recommendation

→ Leverage ratings and reviews data to guide winemaking decisions

→ Increase production of highly rated wine varieties and regions

→ Adjust wine body and acidity profiles based on reviewer taste feedback