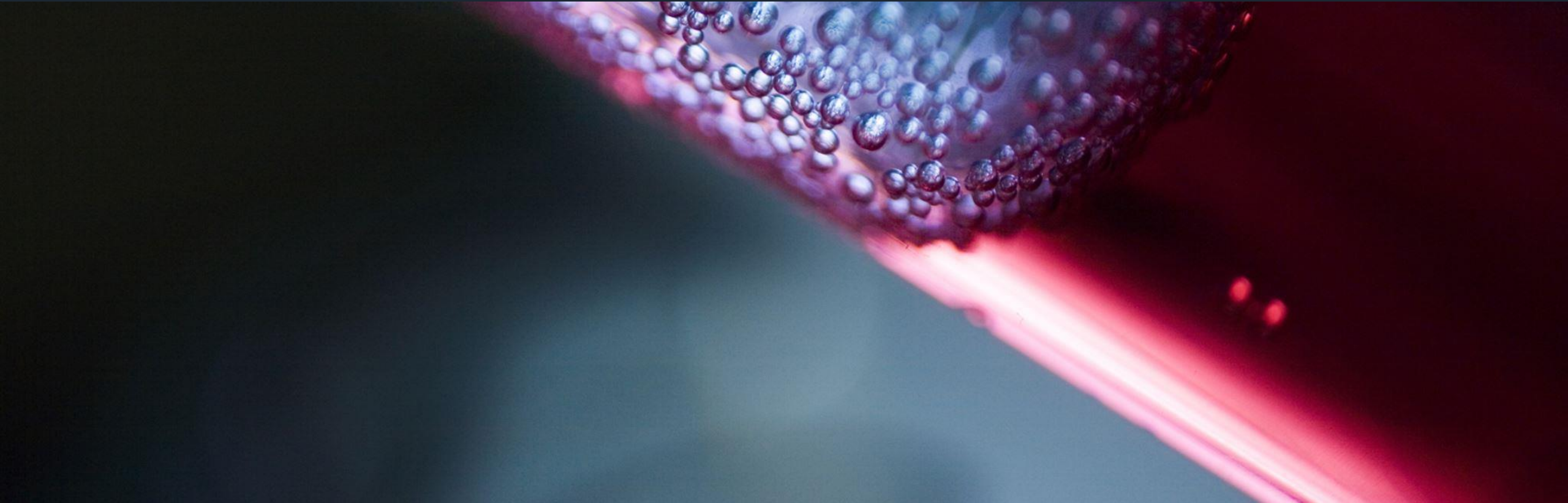
Predicting Optimal Wine Prices with Machine Learning

Name: Gabriel Pantoja



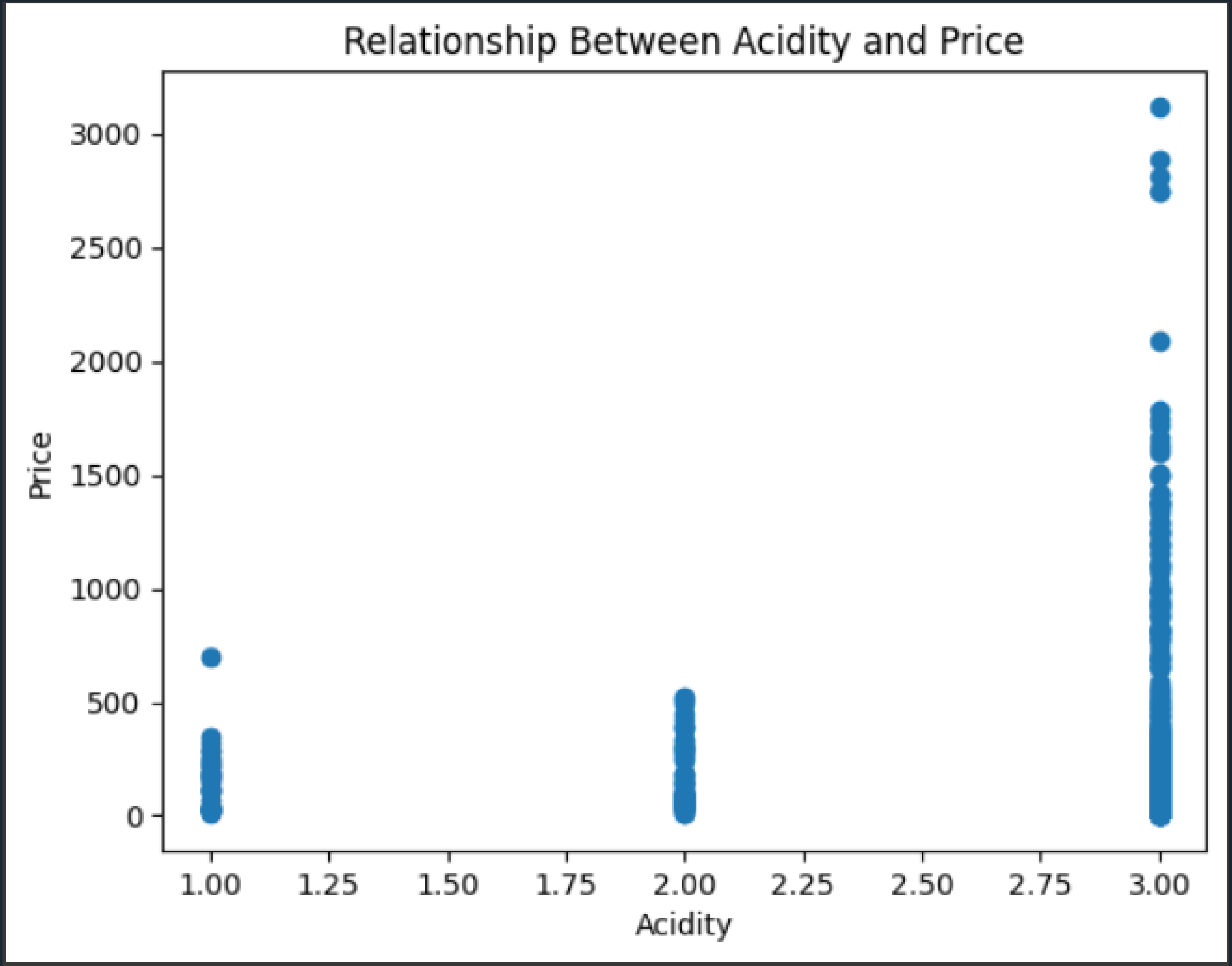
Data Overview

	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
  - → Body
  - → Type
  - → Region
  - → Country
- → Also includes wine quality rating by experts on scale of 0 (worst) to 5(best)
- → 6497 observations of red and white wines
- → 11 variables per wine

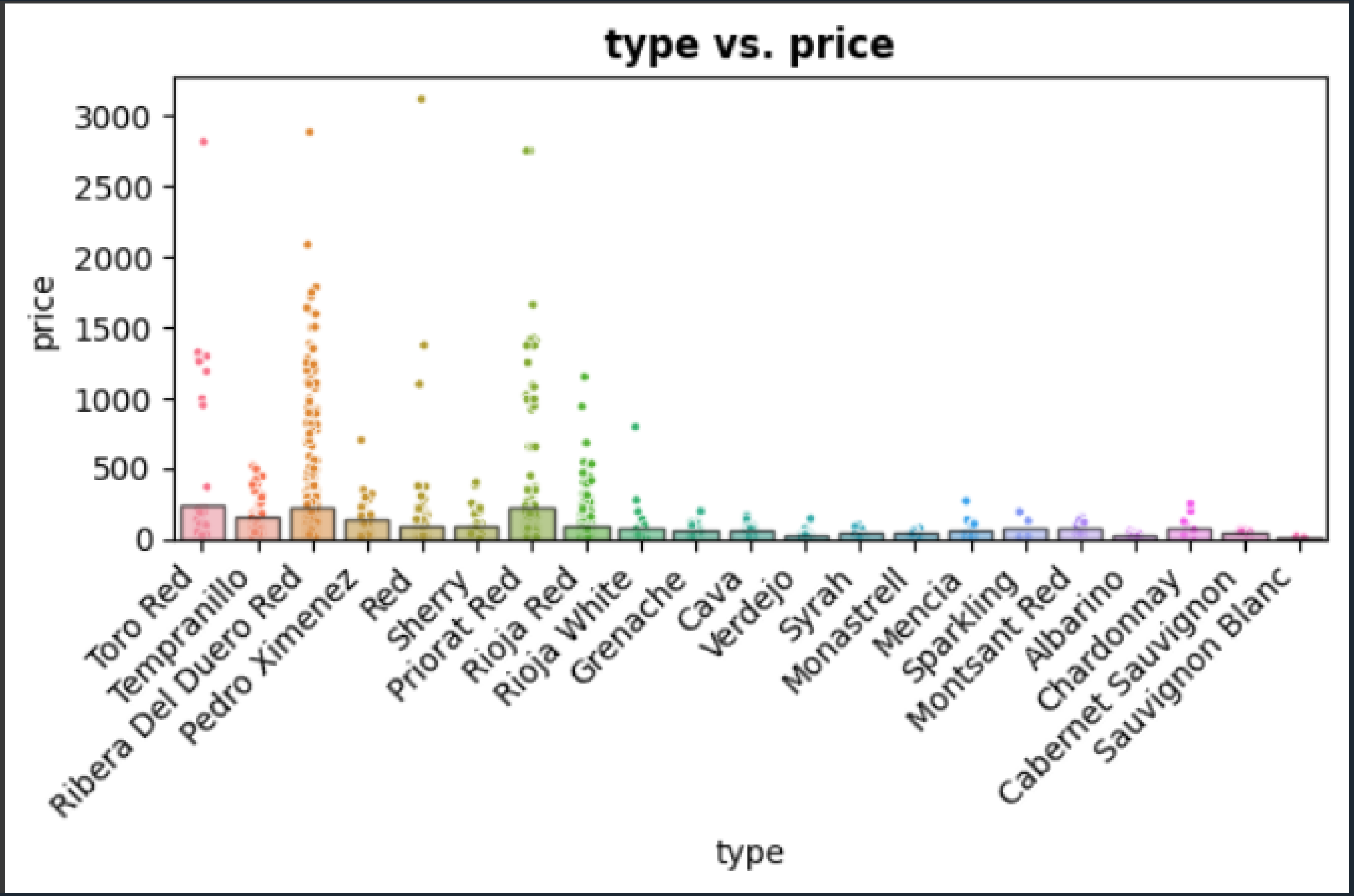
1st Visual

→ 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.



2nd 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.



strengths and limitation

- **Strengths:**
- We tested two models to predict wine quality - Random Forest and K-Nearest Neighbor (KNN).
- The Random Forest model was able to explain 67% of quality variation in the training data. But its accuracy dropped to 43% on new test data. This means it only moderately fits the data and has room for improvement.
- The KNN model achieved an error score of 71,031 on test data after using a PCA technique to simplify the inputs. Lower error is better.
- **Limitations:**
- The moderate accuracy scores mean our models don't fully capture everything that affects wine quality. There's still unpredictability we haven't accounted for.
- The difference in performance between training and test data shows overfitting. This means the models work well only on data they've seen before, not new data.
- There are still winemaking factors our models don't incorporate that impact quality. We need to keep improving them.

Final Recommendation

1

Gather more data to improve model training and avoid overfitting.

2

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