PROJECT 4 WINE ANALYSIS

11.25.2024

Predicting Wine Quality

Collaborators:

Chuck Bui

Jack Jeffries

Beau Massie

Christopher Turner

Problem Statement

VinoVista, a renowned winery, is committed to producing consistently exceptional wines. However, they have observed variability in the quality of their white wine batches. To address this challenge, VinoVista seeks a predictive model that can accurately assess the quality of wine batches based on their chemical properties before bottling.

Project Goal:

The primary objective of this project is to develop a robust machine learning model capable of predicting wine quality on a scale of 0-10. This model will utilize a dataset containing various chemical properties of wine, such as acidity, pH, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulfates, and alcohol content.

Resources:

Wine Quality Dataset from UC Irvine Machine Learning Repository with 4899 samples

QUICK TAKE

- This data set was chosen because it contains information about the chemical propterties of white wines, such as acidity, sugar levels, and alcohol content and the data is modeled after the physicochemical wine tests.
- The number of samples in our dataset after filtering is 3,090. We used Quality as our target and there are 11 total features.
- We cleaned our database by checking for null values, dropping any outliers in 'total sulfur dioxide' above 150 as that is likely a data entry error. Our quality values range from 1-10, so we created binary classification to use in our machine learning models. We chose 7 and above as good; 6 and below as not good.
- We checked for multi collinearity among our features for possible features that could be dropped to increase our precision and recall values and filtered our data to only keep wines that pH values between 3 and 4, as this is the nominal level for white wines.

LOAD THE DATA AND BEGIN CLEANING

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

```python
# check for nulls ( no nulls), and data type
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         4898 non-null   float64
 1   volatile acidity      4898 non-null   float64
 2   citric acid           4898 non-null   float64
 3   residual sugar        4898 non-null   float64
 4   chlorides             4898 non-null   float64
 5   free sulfur dioxide   4898 non-null   float64
 6   total sulfur dioxide  4898 non-null   float64
 7   density               4898 non-null   float64
 8   pH                    4898 non-null   float64
 9   sulphates             4898 non-null   float64
 10  alcohol               4898 non-null   float64
 11  quality               4898 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

```python
# The scale for wine quality is 1-10, so we create a binary classification to use for our ML models
df['quality_label'] = np.where(df['quality'] >= 7, 'good', 'not good')
df.drop('quality', axis=1, inplace=True)
```

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 3090 entries, 1 to 4897
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         3090 non-null   float64
 1   volatile acidity      3090 non-null   float64
 2   citric acid           3090 non-null   float64
 3   residual sugar        3090 non-null   float64
 4   chlorides             3090 non-null   float64
 5   free sulfur dioxide   3090 non-null   float64
 6   total sulfur dioxide  3090 non-null   float64
 7   density               3090 non-null   float64
 8   pH                    3090 non-null   float64
 9   sulphates             3090 non-null   float64
 10  alcohol               3090 non-null   float64
 11  quality_label         3090 non-null   object 
dtypes: float64(11), object(1)
memory usage: 313.8+ KB
```

CODING PAGE 2

```
# convert categorical variables into numerical ones using one-hot encoding to use in our model
df = pd.get_dummies(df, columns=['quality_label'], drop_first=True)
```

```
df.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality_label_not good
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	True
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	True
5	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	True
6	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	True
8	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	True

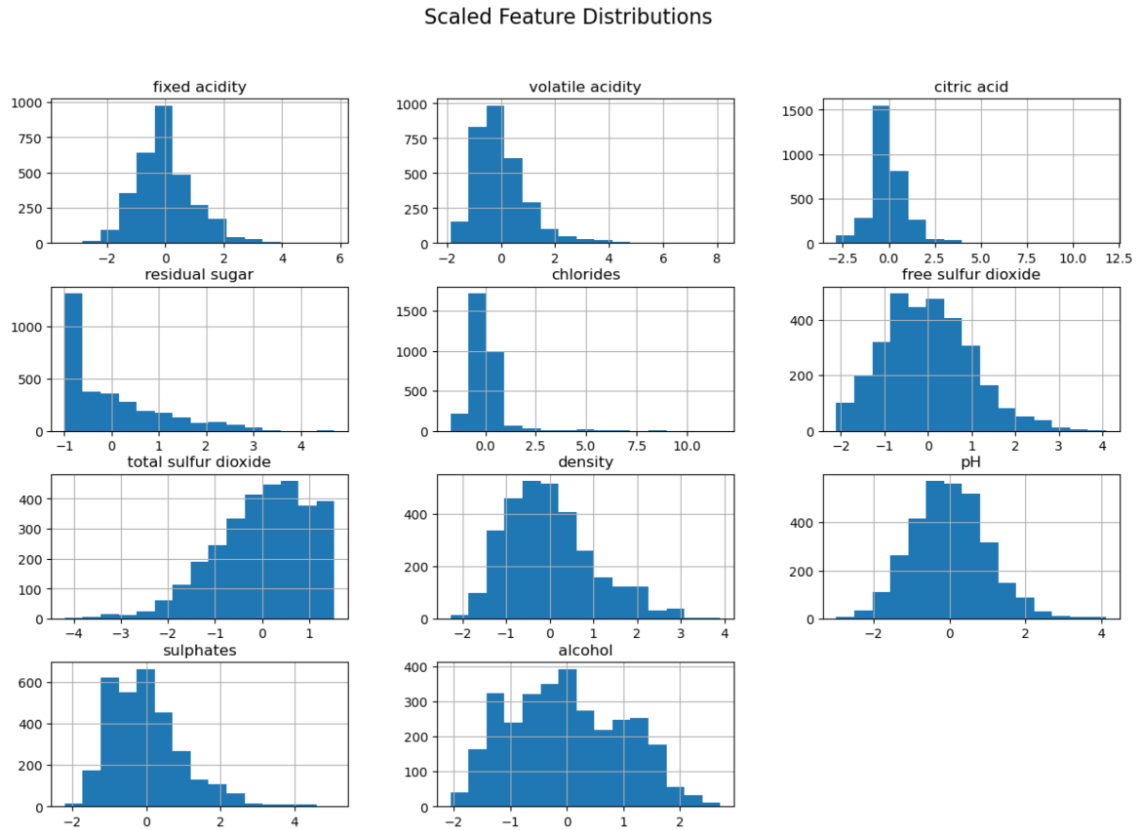
```
# check that shape
df.shape
```

```
(3090, 12)
```

```
# separate features from target
y=df['quality_label_not good']
X=df.drop(columns='quality_label_not good')
```

```
# clean it up, scale it etc
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_scaled_df = pd.DataFrame(X_scaled, columns=X.columns)
```

```
# bin our data to view distributions of our scaled data
X_scaled_df.hist(bins=15, figsize=(15, 10))
plt.suptitle('Scaled Feature Distributions', fontsize=16)
plt.show()
```



CODING PAGE 3

```
# train test module on our standard data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_scaled,
                                                    y,
                                                    random_state=1776,
                                                    stratify=y)
X_train.shape
```

```
(2317, 11)
```

```
# logistic regress
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(solver='lbfgs',
                                max_iter=200,
                                random_state=1776)
classifier.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(max_iter=200, random_state=1776)
```

```
# predictions and confusion matrix
testing_predictions = classifier.predict(X_test)
test_matrix = confusion_matrix(y_test, testing_predictions)
print(test_matrix)
```

```
[[ 78 133]
 [ 56 506]]
```

```
# classification report for our standard data
test_report = classification_report(y_test, testing_predictions)
print(test_report)
```

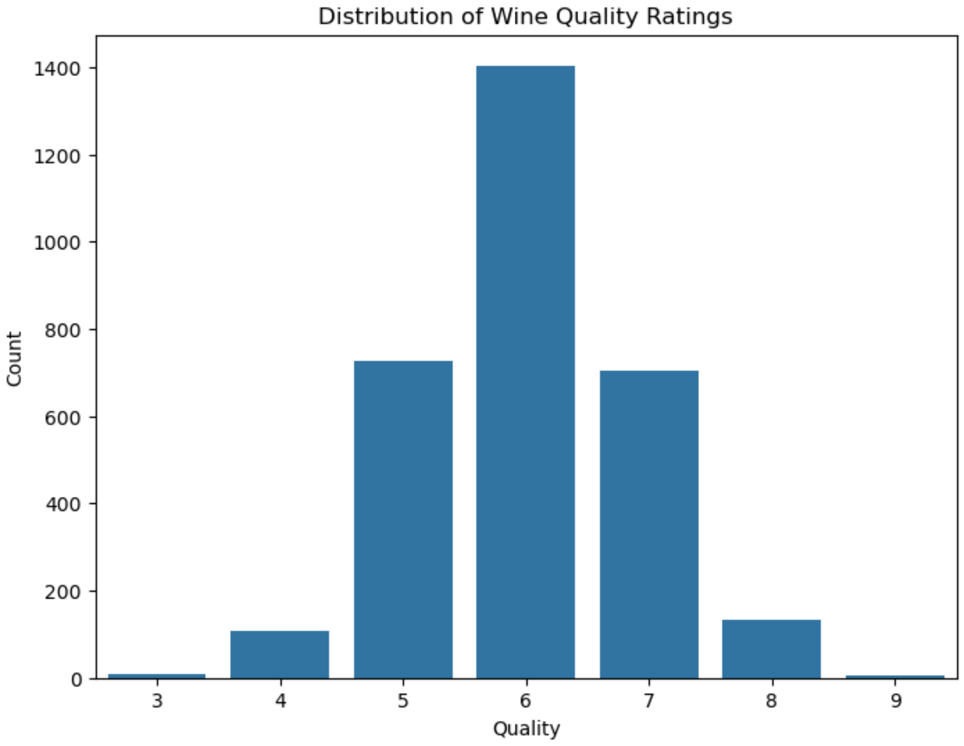
```
              precision    recall  f1-score   support

       False       0.58      0.37      0.45       211
        True       0.79      0.90      0.84       562

    accuracy                           0.76       773
   macro avg       0.69      0.64      0.65       773
weighted avg       0.73      0.76      0.74       773
```

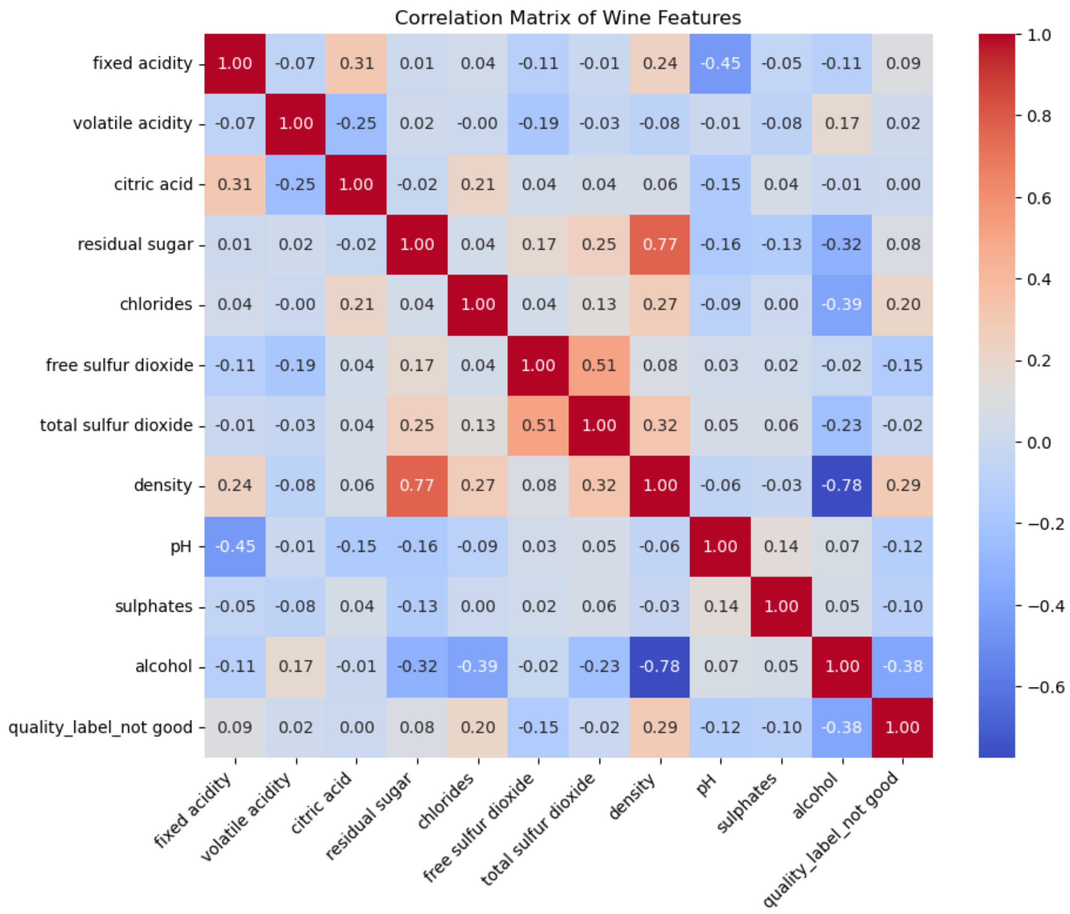
CODING PAGE 4

```
# xgboost model - our best performing model
from xgboost import XGBClassifier

xgb_clf = XGBClassifier(random_state=1776, scale_pos_weight=len(y_train[y_train == 0]) / len(y_train[y_train == 1]))
xgb_clf.fit(X_train, y_train)
predictions = xgb_clf.predict(X_test)
print(classification_report(y_test, predictions))
```

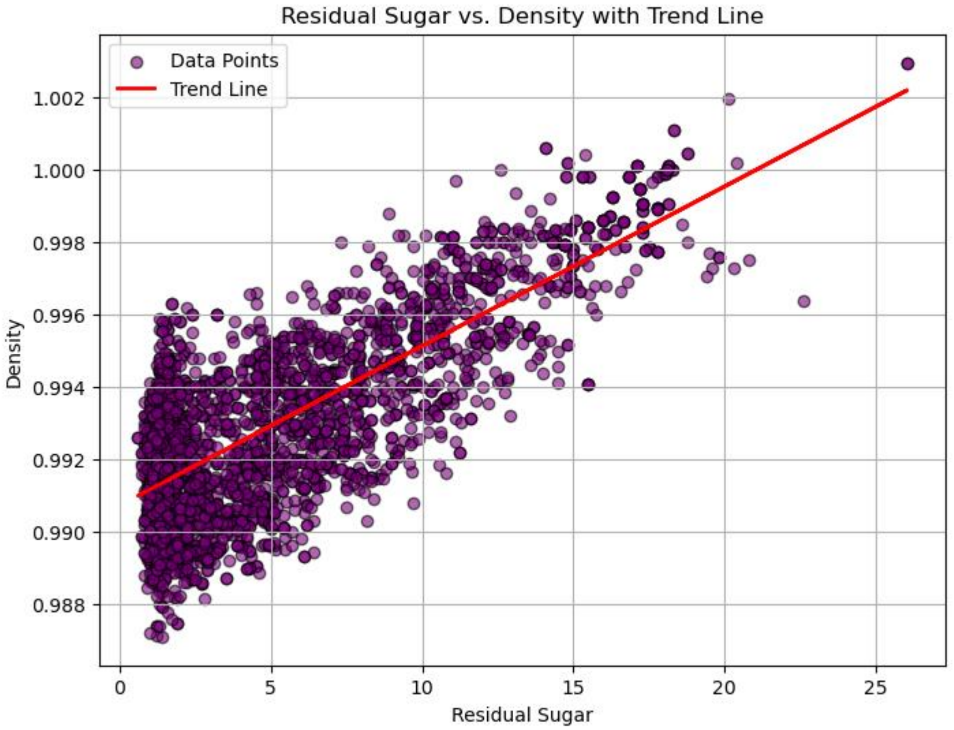
```
              precision    recall  f1-score   support

       False       0.71      0.68      0.69       211
        True       0.88      0.90      0.89       562

    accuracy                           0.84       773
   macro avg       0.80      0.79      0.79       773
weighted avg       0.83      0.84      0.84       773
```

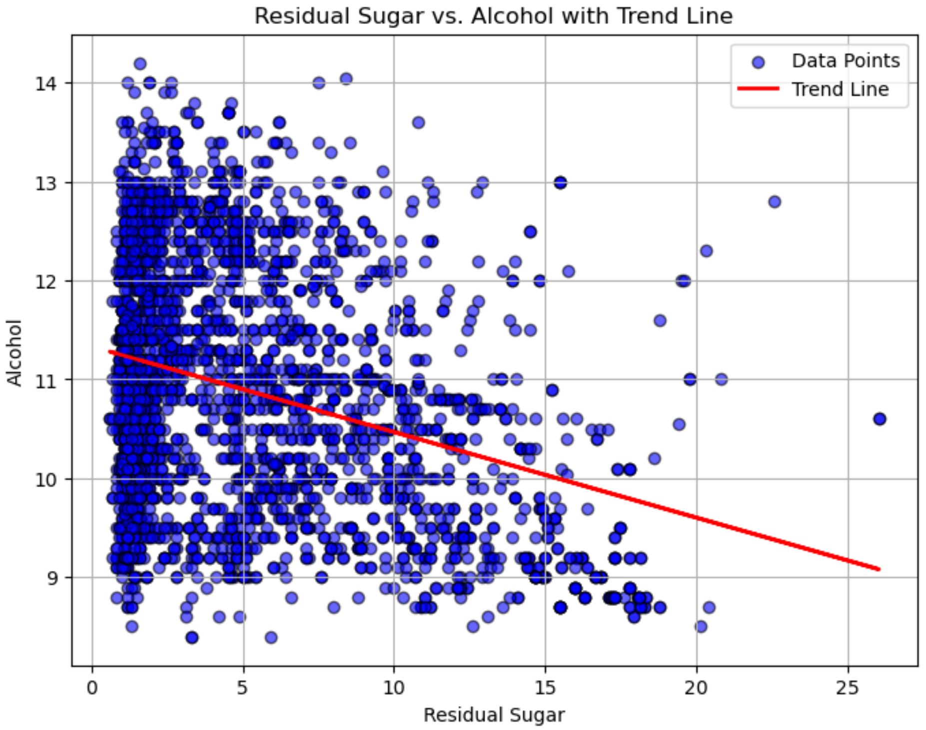
VIZ WALKTHROUGH PAGE 1



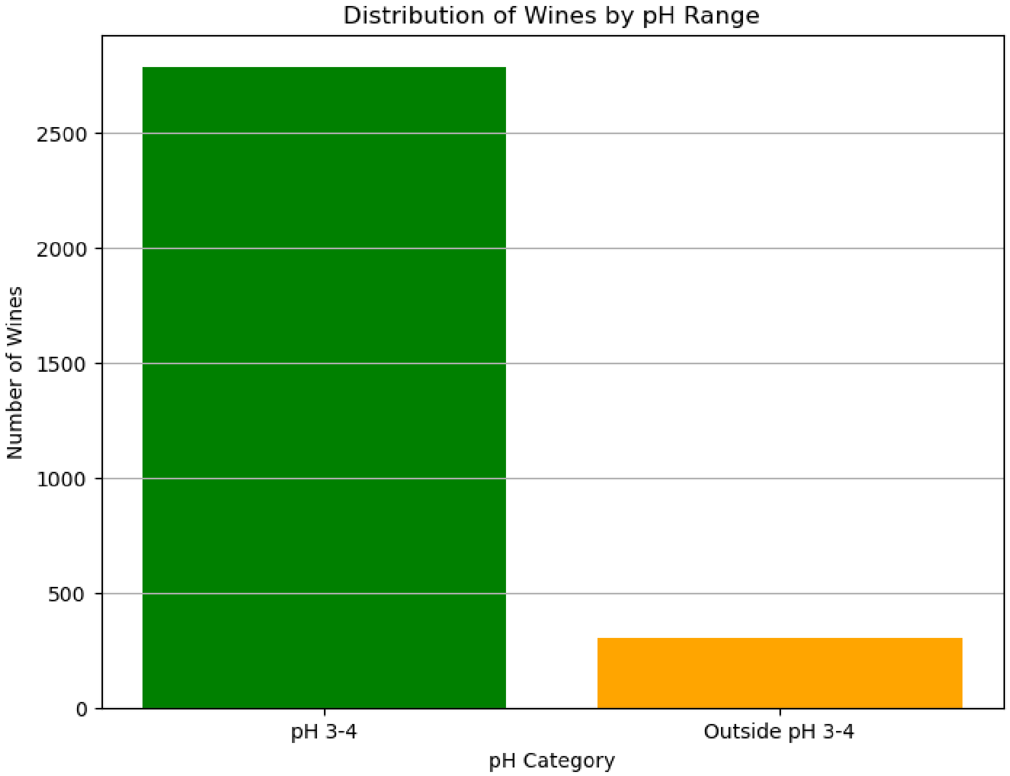


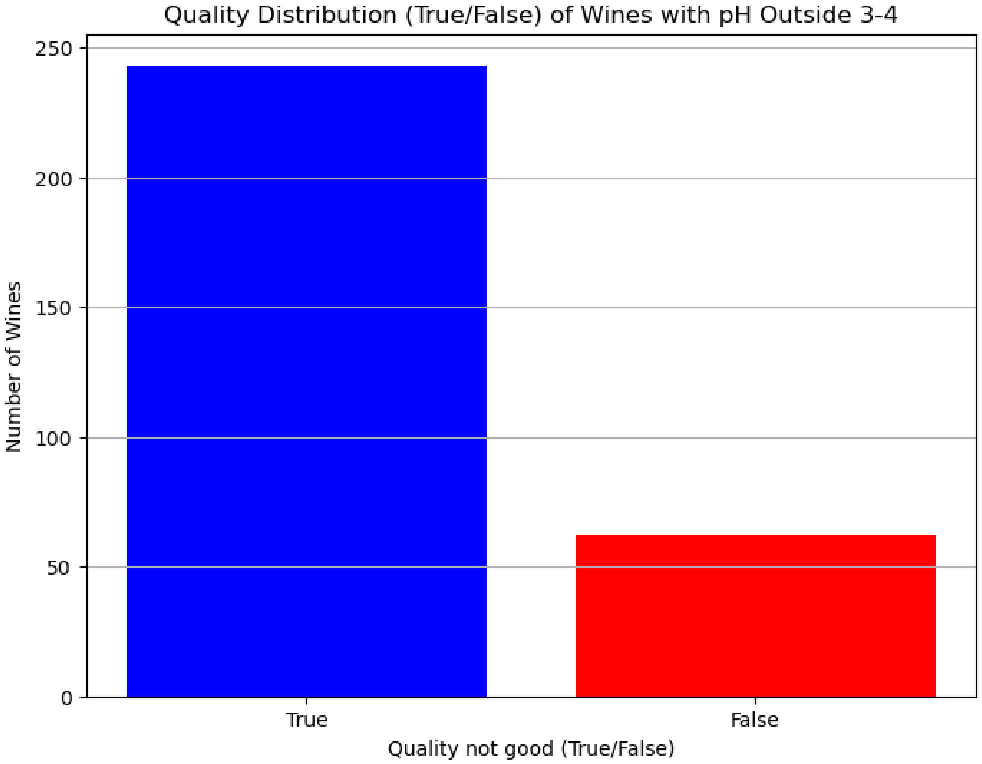
VIZ WALKTHROUGH PAGE 2





VIZ WALKTHROUGH PAGE 3





CONCLUSION

Data Insights

- Residual sugar and density are critical features for predicting wine quality.
- Chemicals like pH, alcohol content, and chlorides also play a significant role

Future Steps

- Incorporate external features like region or grape variety
- Chemicals like pH, alcohol content, and chlorides also play a significant role

WRAP AND SUGGESTIONS TO COMPANY

suggestions

- Suggestions for packaging wine or marketing directly to the consumer
- Teach the consumer with features on the labels and allow them to be part of the process, thus making them feel more loyal to the brand