done

April 17, 2024

0.1 Red Wine Quality Analysis

Introduction

The wine industry relies heavily on consistent quality for consumer satisfaction and brand reputation. This project aims to analyze the Red Wine Quality dataset using data analysis and machine learning techniques to understand factors influencing wine quality. The analysis will focus on uncovering patterns in the physicochemical properties of the wine and their relationship to the sensory experience captured by the quality score.

Data Description

This analysis will utilize a dataset containing physicochemical (input) and sensory (output) variables for red wine. The input variables, likely containing measurements from various stages of wine production, include:

- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- pH
- Sulphates
- Alcohol

The output variable is the quality of the wine, represented by a score between 0 and 10. The source of the dataset will be acknowledged if available, and the data size (number of samples and features) will be mentioned.

Data Loading and Preprocessing

The data will be loaded using Pandas. Necessary preprocessing steps will be performed to ensure data quality and prepare it for analysis. These steps may include:

- Handling missing values: Techniques like mean/median imputation or removal will be chosen based on data characteristics.
- Scaling numerical features: To ensure all features contribute equally to the analysis, numerical features might be scaled using techniques like standardization or normalization. The specific method will be chosen based on the data distribution.

Exploratory Data Analysis (EDA)

A thorough EDA will be conducted to understand the data's characteristics and identify potential relationships between variables. This will involve creating various visualizations and conducting tests for anomalies. Specific areas of exploration include:

- **Univariate analysis:** Histograms and boxplots will be used to visualize the distributions of each variable and identify potential outliers.
- **Multivariate analysis:** Heatmaps will be used to visualize the correlation matrix, revealing relationships between all features. This can highlight potential redundancies or multicollinearity.

Statistical Inference

The target population (e.g., all types of red wines) will be defined. Based on the EDA findings, specific statistical hypotheses about the relationships between features and the target variable (quality) will be formulated.

Specifically, we will be testing the correlation between alcohol content and wine quality, and between volatile acidity and wine quality.

Confidence intervals will be constructed for relevant statistics, and significance levels will be set.

Appropriate statistical tests, such as Pearson correlation coefficient for correlation tests, will be conducted to evaluate the formulated hypotheses.

Machine Learning Models

Linear regression models will be the primary focus to predict both the quality and alcohol content of the wine using the other 11 features. Ordinary least squares regression will be the initial model, and its performance will be analyzed using metrics like:

- Feature significance: P-values of coefficients will be used to identify features that significantly contribute to the prediction. Non-significant features might be removed to avoid overfitting.
- R-squared: This metric indicates how well the model explains the variance in the target variable.
- Adjusted R-squared: This penalizes models with too many features and provides a more reliable estimate of performance.
- Information criteria: Metrics like AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) will be used to compare different model complexities and choose the one with the best balance of fit and parsimony.

Furthermore, the possibility of exploring non-linear models (e.g., decision trees, random forests) will be considered if the data suggests complex relationships. Regularization techniques like Ridge or Lasso regression might also be explored to handle potential issues with correlated features.

Model Validation and Evaluation

Cross-validation techniques will be used to validate the models and obtain a more robust estimate of their generalizability beyond the training data. The performance of the models will be evaluated using appropriate metrics like R-squared and mean squared error (MSE).

Visualization

A visually appealing and informative dashboard will be created using Looker Studio or another Business Intelligence (BI) tool. The dashboard will include at least three different chart types to effectively communicate the findings. Examples could include:

- Bar charts for feature distributions
- Scatter plots for correlations between features and target variables
- Line charts for model performance metrics

Conclusion and Suggestions for Improvement

The results of the analysis will be clearly explained, highlighting important insights gained about the relationship between physicochemical properties and wine quality.

1 Data Loading and Preprocessing

[ ]:
```
%load_ext autoreload
%autoreload 2
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly as py
import plotly.express as px
import plotly.figure_factory as ff
import plotly.subplots as sp
import plotly.graph_objs as go
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import sqlite3
from functions import *
from scipy.stats import pearsonr
from statsmodels.stats.proportion import proportion_confint
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
from statsmodels.api import OLS, add_constant
```

[ ]:
```
wine_df = pd.read_csv('winequality-red.csv')
conn = sqlite3.connect('wine_quality.db')
wine_df.to_sql('wine_quality', conn, if_exists='replace', index=False)
```

[ ]:
```
1599
```

[ ]:
```
wine_df.head()
```

[ ]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.4	0.70	0.00	1.9	0.076	
1	7.8	0.88	0.00	2.6	0.098	
2	7.8	0.76	0.04	2.3	0.092	
3	11.2	0.28	0.56	1.9	0.075	
4	7.4	0.70	0.00	1.9	0.076	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	11.0	34.0	0.9978	3.51	0.56	
1	25.0	67.0	0.9968	3.20	0.68	
2	15.0	54.0	0.9970	3.26	0.65	
3	17.0	60.0	0.9980	3.16	0.58	
4	11.0	34.0	0.9978	3.51	0.56	

	alcohol	quality
0	9.4	5
1	9.8	5
2	9.8	5
3	9.8	6
4	9.4	5

[ ]:
```
wine_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

[ ]:
```
wine_df.describe().T
```

[ ]:

	count	mean	std	min	25%	\
fixed acidity	1599.0	8.319637	1.741096	4.60000	7.1000	
volatile acidity	1599.0	0.527821	0.179060	0.12000	0.3900	

citric acid	1599.0	0.270976	0.194801	0.00000	0.0900
residual sugar	1599.0	2.538806	1.409928	0.90000	1.9000
chlorides	1599.0	0.087467	0.047065	0.01200	0.0700
free sulfur dioxide	1599.0	15.874922	10.460157	1.00000	7.0000
total sulfur dioxide	1599.0	46.467792	32.895324	6.00000	22.0000
density	1599.0	0.996747	0.001887	0.99007	0.9956
pH	1599.0	3.311113	0.154386	2.74000	3.2100
sulphates	1599.0	0.658149	0.169507	0.33000	0.5500
alcohol	1599.0	10.422983	1.065668	8.40000	9.5000
quality	1599.0	5.636023	0.807569	3.00000	5.0000

	50%	75%	max
fixed acidity	7.90000	9.200000	15.90000
volatile acidity	0.52000	0.640000	1.58000
citric acid	0.26000	0.420000	1.00000
residual sugar	2.20000	2.600000	15.50000
chlorides	0.07900	0.090000	0.61100
free sulfur dioxide	14.00000	21.000000	72.00000
total sulfur dioxide	38.00000	62.000000	289.00000
density	0.99675	0.997835	1.00369
pH	3.31000	3.400000	4.01000
sulphates	0.62000	0.730000	2.00000
alcohol	10.20000	11.100000	14.90000
quality	6.00000	6.000000	8.00000

We can see that the dataset consists of 1,599 entries, each with 12 columns that include both physicochemical properties and the quality rating of red wine. The columns are as follows:

fixed acidity

volatile acidity

citric acid

residual sugar

chlorides

free sulfur dioxide

total sulfur dioxide

density

pH

sulphates

alcohol

quality (the target variable for one of our models)**

1.0.1 Initial Observations:

Data Types: All features are numerical, which is suitable for regression models.

No Missing Values: Each column has 1,599 non-null values indicating there are no missing values.

Statistical Summary: Those 12 columns represent the following features:

- **Fixed Acidity:** The average value is 8.31, the highest value is 15.9.
- **Volatile Acidity:** The average value is 0.529, the highest value is 1.58.
- **Citric Acid:** The average value is 0.272, the highest value is 1.00.
- **Residual Sugar:** The average value is 2.523, the highest value is 15.5.
- **Chlorides:** The average value is 0.088, the highest value is 0.611.
- **Free Sulfur Dioxide:** The average value is 15.89, the highest value is 72.00.
- **Total Sulfur Dioxide:** The average value is 46.82, the highest value is 289.00.
- **Density:** The average value is 0.996, the highest value is 1.004.
- **pH:** The average value is 3.3, the highest value is 4.01.
- **Sulphates:** The average value is 0.65, the highest value is 2.00.
- **Alcohol:** The average value is 10.43, the highest value is 14.90.
- **Quality:** The average value is 5.62, the highest value is 8.00.

Afterwards, we can rename the features to be able to use them more easily.

[ ]:
```
wine_df_cleaned = remove_duplicates(wine_df)
duplicate_count = len(wine_df) - len(wine_df_cleaned)
wine_df_cleaned.info()
```

```
Removed 240 duplicate rows
<class 'pandas.core.frame.DataFrame'>
Index: 1359 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1359 non-null   float64
 1   volatile acidity      1359 non-null   float64
 2   citric acid           1359 non-null   float64
 3   residual sugar        1359 non-null   float64
 4   chlorides             1359 non-null   float64
 5   free sulfur dioxide   1359 non-null   float64
 6   total sulfur dioxide  1359 non-null   float64
 7   density               1359 non-null   float64
 8   pH                    1359 non-null   float64
 9   sulphates             1359 non-null   float64
 10  alcohol               1359 non-null   float64
 11  quality               1359 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 138.0 KB
```

We have removed 240 duplicate rows, leaving 1,359 unique entries in the dataset, we can rename the features to be able to use them more easily.

[ ]:
```
get_columns(wine_df_cleaned)
```

[ ]:
```
['fixed acidity',
 'volatile acidity',
 'citric acid',
 'residual sugar',
 'chlorides',
 'free sulfur dioxide',
 'total sulfur dioxide',
 'density',
 'pH',
 'sulphates',
 'alcohol',
 'quality']
```

[ ]:
```
wine_df_cleaned.rename(columns = {'fixed acidity': 'fixed_acidity', 'volatile␣
↪acidity': 'volatile_acidity', 'citric acid': 'citric_acid', 'residual sugar':
↪ 'residual_sugar', 'free sulfur dioxide': 'free_sulfur_dioxide', 'total␣
↪sulfur dioxide': 'total_sulfur_dioxide'}, inplace = True)
```

Next, let's identify outliers using the boxplot for each numeric column in the dataset.

This will help us determine if any extreme values need to be addressed before further analysis and modeling.

[ ]:
```
plot_box_chart(wine_df_cleaned, "Feature", "Value", "Boxplot of Features")
```

The boxplot reveals noticeable outliers in features like "free sulfur dioxide" and "total sulfur dioxide." While other features may also have outliers, the IQR (Interquartile Range) can help us further confirm these existing outliers and identify potential ones in other features.

[ ]:
```
identify_outliers(wine_df_cleaned)
```

[ ]:
```
fixed_acidity           41
volatile_acidity        19
citric_acid              1
residual_sugar         126
chlorides               87
free_sulfur_dioxide     26
total_sulfur_dioxide    45
density                 35
pH                      28
sulphates               55
alcohol                 12
quality                 27
dtype: int64
```

1.0.2 Outlier Analysis:

The dataset contains various counts of outliers across different features. Here are some notable counts:

Residual Sugar: 126 outliers

Chlorides: 87 outliers

Sulphates: 55 outliers

Total Sulfur Dioxide: 45 outliers

Fixed Acidity: 41 outliers

1.0.3 Next Steps:

Visual Exploratory Data Analysis (EDA): Before deciding on handling these outliers (e.g., capping, removing), we will visualize the data using histograms and box plots to better understand the distribution of each feature.

Statistical Summaries: Provide detailed statistics for each feature to accompany the visual analysis.#

[ ]:
```
plot_histograms(wine_df_cleaned, ["fixed_acidity", "volatile_acidity",␣
↪"citric_acid", "residual_sugar"])
```

Fixed Acidity: The distribution is right-skewed, indicating that most wines have a fixed acidity level around 7-8, with fewer wines having higher fixed acidity.

Volatile Acidity: The data is also right-skewed, showing that most wines have a volatile acidity around 0.5, with only a few wines having a volatile acidity above 1.

Citric Acid: This feature shows a bimodal distribution, indicating two groups of wines, one with low citric acid close to 0 and another with citric acid between 0.25 and 0.5.

Residual Sugar: The distribution is highly right-skewed, suggesting that most wines have low residual sugar levels, with a peak below five. Few wines have high residual sugar levels.

[ ]:
```
plot_histograms(wine_df_cleaned, ["chlorides", "free_sulfur_dioxide",␣
↪"total_sulfur_dioxide", "density"])
```

Chlorides: The distribution is right-skewed, indicating that most wines have a chloride level around 0.1, with fewer wines having higher chloride levels.

Free Sulfur Dioxide: The data is also right-skewed, showing that most wines have a free sulfur dioxide count around 10-20, with only a few wines having a count above 40.

Total Sulfur Dioxide: This feature shows a right-skewed distribution, indicating that most wines have a total sulfur dioxide count around 50, with fewer wines having a count above 100.

Density: The distribution is approximately normal, suggesting that most wines have a density around 1.0, with few wines having significantly higher or lower densities.

[ ]:
```
plot_histograms(wine_df_cleaned, ["pH", "sulphates", "alcohol", "quality"])
```

pH: The distribution is approximately normal, indicating that most wines have a pH level around 3.2, with fewer wines having significantly higher or lower pH levels.

Sulphates: The data is right-skewed, showing that most wines have a sulphate level around 0.5, with only a few wines having a level above 1.

Alcohol: This feature shows a right-skewed distribution, indicating that most wines have an alcohol level around 9-10%, with fewer wines having a level above 13%.

Quality: The distribution is approximately normal, suggesting that most wines have a quality rating around 6, with few wines having significantly higher or lower ratings.

Our initial data exploration using histograms revealed that most features (fixed acidity, volatile acidity, etc.) exhibit right-skewed distributions, indicating a concentration of wines with values clustered around a central point.

This pattern suggests that the majority of wines fall within a specific range for these features. In contrast, "density" and "pH" show distributions closer to normal, while "citric acid" has a unique bimodal distribution.

Notably, "quality" itself appears to be normally distributed.

These observations provide a foundation for further analysis, particularly investigating relationships between these features and wine quality to identify potential patterns and correlations.

[ ]:
```
corr_matrix = wine_df.corr()
print(corr_matrix)
```

```
                      fixed acidity  volatile acidity  citric acid  \
fixed acidity              1.000000         -0.256131     0.671703
volatile acidity          -0.256131          1.000000    -0.552496
citric acid                0.671703         -0.552496     1.000000
residual sugar             0.114777          0.001918     0.143577
chlorides                  0.093705          0.061298     0.203823
free sulfur dioxide       -0.153794         -0.010504    -0.060978
total sulfur dioxide      -0.113181          0.076470     0.035533
density                    0.668047          0.022026     0.364947
pH                        -0.682978          0.234937    -0.541904
sulphates                  0.183006         -0.260987     0.312770
alcohol                   -0.061668         -0.202288     0.109903
quality                    0.124052         -0.390558     0.226373

                      residual sugar  chlorides  free sulfur dioxide  \
fixed acidity               0.114777   0.093705            -0.153794
volatile acidity            0.001918   0.061298            -0.010504
citric acid                 0.143577   0.203823            -0.060978
residual sugar              1.000000   0.055610             0.187049
chlorides                   0.055610   1.000000             0.005562
free sulfur dioxide         0.187049   0.005562             1.000000
total sulfur dioxide        0.203028   0.047400             0.667666
density                     0.355283   0.200632            -0.021946
pH                         -0.085652  -0.265026             0.070377
```

```
sulphates                    0.005527   0.371260              0.051658
alcohol                      0.042075  -0.221141             -0.069408
quality                      0.013732  -0.128907             -0.050656

                      total sulfur dioxide   density        pH  sulphates  \
fixed acidity                    -0.113181  0.668047 -0.682978   0.183006
volatile acidity                  0.076470  0.022026  0.234937  -0.260987
citric acid                       0.035533  0.364947 -0.541904   0.312770
residual sugar                    0.203028  0.355283 -0.085652   0.005527
chlorides                         0.047400  0.200632 -0.265026   0.371260
free sulfur dioxide               0.667666 -0.021946  0.070377   0.051658
total sulfur dioxide              1.000000  0.071269 -0.066495   0.042947
density                           0.071269  1.000000 -0.341699   0.148506
pH                               -0.066495 -0.341699  1.000000  -0.196648
sulphates                         0.042947  0.148506 -0.196648   1.000000
alcohol                          -0.205654 -0.496180  0.205633   0.093595
quality                          -0.185100 -0.174919 -0.057731   0.251397

                       alcohol   quality
fixed acidity        -0.061668  0.124052
volatile acidity     -0.202288 -0.390558
citric acid           0.109903  0.226373
residual sugar        0.042075  0.013732
chlorides            -0.221141 -0.128907
free sulfur dioxide  -0.069408 -0.050656
total sulfur dioxide -0.205654 -0.185100
density              -0.496180 -0.174919
pH                    0.205633 -0.057731
sulphates             0.093595  0.251397
alcohol               1.000000  0.476166
quality               0.476166  1.000000
```

[ ]:
```
plot_heatmap(corr_matrix)
```

1.0.4 Correlation Analysis Insights

The correlation matrix reveals several key relationships:

Quality and Alcohol: A significant positive correlation (r=0.48) exists between alcohol and quality, suggesting that higher alcohol levels might be associated with higher quality ratings.

Quality and Volatile Acidity: A notable negative correlation ($r = -0.40$) suggests that wines with lower volatile acidity tend to have higher quality.

Density and Fixed Acidity: A strong positive correlation (r=0.67), indicating that as fixed acidity increases, so does the density of the wine.

pH and Fixed Acidity: A strong negative correlation ($r=-0.69$), which is expected as acidity and pH are inversely related.

We can move with checking the relationships between top correlated and top negatively correlated variables.

[ ]:
```
plot_top_correlations(wine_df, corr_matrix, 3)
```

When we look at the scatterplots, we see a straightforward relationship between each pair of variables:

What the Relationship Looks Like:

As one variable increases, the other one tends to increase too. This relationship can sometimes be the opposite, where one decreases as the other increases. It's predictable and consistent.

Why This Matters for our Predictive Models:

These relationships are important because they tell us that these variable pairs might help predict outcomes well in certain types of machine learning models.

For example, logistic regression models, which work well with this kind of relationship, could use these variables effectively. Next Steps:

Although the scatterplots give us good hints, we need to do more detailed statistics to be sure about these relationships. This means checking how strong and reliable these relationships are before we use them in models.

Therefore, we move to hypothesis testing:

1.0.5 Hypotheses

1. **Quality and Alcohol**:

   H0: There is no correlation between alcohol level and wine quality (= 0).

   H1: There is a correlation between alcohol level and wine quality (0).

2. **Quality and Volatile Acidity**:

   H0: There is no correlation between volatile acidity and wine quality (= 0).

   H1: There is a correlation between volatile acidity and wine quality (0).

3. **Density and Fixed Acidity**:

   H0: There is no correlation between density and fixed acidity (= 0).

   H1: There is a correlation between density and fixed acidity (0).

4. **pH and Fixed Acidity**:

   H0: There is no correlation between pH and fixed acidity (= 0).

   H1: There is a correlation between pH and fixed acidity (0).

[ ]:
```
results_alcohol_quality = test_correlation(wine_df_cleaned, 'alcohol',␣
↪'quality')
results_acidity_quality = test_correlation(wine_df_cleaned, 'volatile_acidity',␣
↪'quality')
```

```
results_density_acidity = test_correlation(wine_df_cleaned, 'density',␣
 ↪'fixed_acidity')
results_ph_acidity = test_correlation(wine_df_cleaned, 'pH', 'fixed_acidity')

results_alcohol_quality, results_acidity_quality, results_density_acidity,␣
 ↪results_ph_acidity
```

[ ]:
```
({'Correlation': 0.4803428980019927,
  'P-Value': 2.2787211325386807e-79,
  'Reject H0': True,
  '95% CI': (0.4359933265475753, 0.5246924694564101)},
 {'Correlation': -0.3952136890098409,
  'P-Value': 4.9281935807917204e-52,
  'Reject H0': True,
  '95% CI': (-0.4416621332852135, -0.3487652447344683)},
 {'Correlation': 0.6701950166538053,
  'P-Value': 6.223451737711637e-178,
  'Reject H0': True,
  '95% CI': (0.6326664454281221, 0.7077235878794885)},
 {'Correlation': -0.6866851055982757,
  'P-Value': 3.6657854879384933e-190,
  'Reject H0': True,
  '95% CI': (-0.7234435697455686, -0.6499266414509828)})
```

1. **Quality and Alcohol**
   - Correlation: 0.48
   - P-Value: 2.28×10^-79
   - Reject H0: Yes
   - 95% Confidence Interval: (0.436, 0.524)
   - Interpretation: There is a significant positive correlation between alcohol level and wine quality, suggesting that higher alcohol levels are associated with higher quality ratings.
2. **Quality and Volatile Acidity**
   - Correlation: -0.395
   - P-Value: 4.93×10^-52
   - Reject H0: Yes
   - 95% Confidence Interval: (-0.442, -0.349)
   - Interpretation: There is a significant negative correlation between volatile acidity and wine quality, indicating that lower volatile acidity is associated with higher quality ratings.
3. **Density and Fixed Acidity**
   - Correlation: 0.67
   - P-Value: 6.22×10^-178
   - Reject H0: Yes
   - 95% Confidence Interval: (0.633, 0.708)
   - Interpretation: There is a strong positive correlation between density and fixed acidity, showing that as fixed acidity increases, so does the density of the wine.
4. **pH and Fixed Acidity**
   - Correlation: -0.686

- P-Value: 3.67×10^-190
- Reject H0: Yes
- 95% Confidence Interval: (-0.723, -0.649)
- Interpretation: There is a strong negative correlation between pH and fixed acidity, which is consistent with the chemical relationship where an increase in acidity results in a lower pH.

These results affirm that the relationships observed in the scatterplots are statistically significant. This information is valuable for predictive modeling, especially when selecting features that influence wine quality.

Therefore, we move to modeling, first setting up the data for modeling.

[ ]:
```
X = wine_df_cleaned.drop(columns=['quality', 'alcohol'])
y_quality = wine_df_cleaned['quality']
y_alcohol = wine_df_cleaned['alcohol']

X_train_q, X_test_q, y_train_q, y_test_q = train_test_split(X, y_quality,␣
↪test_size=0.2, random_state=42)

X_train_a, X_test_a, y_train_a, y_test_a = train_test_split(X, y_alcohol,␣
↪test_size=0.2, random_state=42)
```

Then we build and try to predict quality

[ ]:
```
%reload_ext autoreload
```

[ ]:
```
X_train_q_const = add_constant(X_train_q)
X_test_q_const = add_constant(X_test_q)

model_quality = OLS(y_train_q, X_train_q_const).fit()

predictions_q = model_quality.predict(X_test_q_const)
r2_quality = r2_score(y_test_q, predictions_q)
aic_quality = model_quality.aic

model_quality.summary()
```

[ ]:

Dep. Variable:	quality	**R-squared:**	0.299
Model:	OLS	**Adj. R-squared:**	0.293
Method:	Least Squares	**F-statistic:**	45.96
Date:	Wed, 17 Apr 2024	**Prob (F-statistic):**	2.30e-76
Time:	23:34:09	**Log-Likelihood:**	-1130.8
No. Observations:	1087	**AIC:**	2284.
Df Residuals:	1076	**BIC:**	2339.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	171.6458	17.929	9.574	0.000	136.466	206.825
fixed_acidity	0.1396	0.029	4.897	0.000	0.084	0.196
volatile_acidity	-0.9075	0.155	-5.854	0.000	-1.212	-0.603
citric_acid	0.1613	0.186	0.867	0.386	-0.204	0.527
residual_sugar	0.0827	0.017	4.903	0.000	0.050	0.116
chlorides	-2.5900	0.502	-5.156	0.000	-3.576	-1.604
free_sulfur_dioxide	0.0043	0.003	1.531	0.126	-0.001	0.010
total_sulfur_dioxide	-0.0046	0.001	-4.930	0.000	-0.006	-0.003
density	-169.1970	18.391	-9.200	0.000	-205.284	-133.110
pH	0.3825	0.217	1.761	0.078	-0.044	0.809
sulphates	1.2305	0.142	8.692	0.000	0.953	1.508

Omnibus:	25.178	**Durbin-Watson:**	1.932
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	37.424
Skew:	-0.221	**Prob(JB):**	7.47e-09
Kurtosis:	3.794	**Cond. No.**	7.43e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.43e+04. This might indicate that there are strong multicollinearity or other numerical problems.

We have successfully built and evaluated linear regression model for predicting both quality in the dataset.

Model for Predicting Quality R-squared: 0.299 This indicates that approximately 29.9% of the variance in wine quality is explained by the model, which is moderate. AIC: 2284 The Akaike Information Criterion (AIC) helps in model selection, where lower AIC values indicate a better model. This value will be useful for comparison if we consider alternative models or subsets of features.

1.0.6 Significant Features:

fixed acidity, volatile acidity, residual sugar, chlorides, total sulfur dioxide, and density are statistically significant predictors of quality. Interestingly, citric acid and free sulfur dioxide are not significant at the 0.05 level, which may suggest these variables are less important in predicting quality in this model setup.

Model for predicting alcohol

[ ]:
```
X_train_a_const = add_constant(X_train_a)
X_test_a_const = add_constant(X_test_a)

# Fit the model
model_alcohol = OLS(y_train_a, X_train_a_const).fit()

predictions_a = model_alcohol.predict(X_test_a_const)
r2_alcohol = r2_score(y_test_a, predictions_a)
aic_alcohol = model_alcohol.aic
```

```
model_alcohol.summary()
```

[ ]:

Dep. Variable:	alcohol	**R-squared:**	0.684
Model:	OLS	**Adj. R-squared:**	0.681
Method:	Least Squares	**F-statistic:**	232.7
Date:	Wed, 17 Apr 2024	**Prob (F-statistic):**	7.35e-261
Time:	23:34:45	**Log-Likelihood:**	-1011.1
No. Observations:	1087	**AIC:**	2044.
Df Residuals:	1076	**BIC:**	2099.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	615.8133	16.060	38.345	0.000	584.301	647.325
fixed_acidity	0.5501	0.026	21.541	0.000	0.500	0.600
volatile_acidity	0.3046	0.139	2.194	0.028	0.032	0.577
citric_acid	0.8472	0.167	5.080	0.000	0.520	1.174
residual_sugar	0.2827	0.015	18.707	0.000	0.253	0.312
chlorides	-1.2387	0.450	-2.753	0.006	-2.122	-0.356
free_sulfur_dioxide	-0.0012	0.003	-0.481	0.631	-0.006	0.004
total_sulfur_dioxide	-0.0030	0.001	-3.567	0.000	-0.005	-0.001
density	-626.4235	16.474	-38.025	0.000	-658.748	-594.099
pH	3.8583	0.195	19.835	0.000	3.477	4.240
sulphates	1.2277	0.127	9.682	0.000	0.979	1.477

Omnibus:	74.150	**Durbin-Watson:**	1.914
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	140.658
Skew:	0.462	**Prob(JB):**	2.86e-31
Kurtosis:	4.500	**Cond. No.**	7.43e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.43e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

R-squared: 0.684 This model explains about 68.4% of the variance in alcohol content, which is substantially higher than the model for quality, suggesting a good fit. AIC: 2044 As with the quality model, this AIC can help compare different models for predicting alcohol

1.0.7 Significant Features:

Most features are significant with the exception of free sulfur dioxide, indicating strong predictive power across most of the chemical properties. This includes fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, and sulphates.

1.0.8 So what these results mean?

The significant predictors in each model provide insights into which chemical properties influence wine quality and alcohol content the most.

The relatively low R-squared for the quality model suggests that while the identified factors are important, there are possibly other unmeasured factors affecting quality.

The high R-squared for the alcohol model indicates a strong linear relationship between the predictors and the alcohol content, making this model very reliable for predictions within this dataset.

1.0.9 Improvements

Enhanced EDA with Visualizations: Incorporating more visual analysis would help identify outliers, data distribution properties, and more detailed inter-variable relationships.

Robust Anomaly Detection: Apply methods to detect and handle outliers, which could improve model accuracy and robustness.

Model Expansion and Comparison: Comparing multiple models, including non-linear models and ensemble methods, could provide insights into better predictive performance or feature interactions.

Feature Engineering: Investigating transformations or interactions among features could uncover more complex relationships and potentially improve model performance.

Cross-Validation: Utilizing cross-validation techniques would help ensure that the model's performance is stable across different subsets of data, enhancing its generalizability.

[ ]:
```
%reload_ext autoreload
```

[ ]:
```

```

[ ]:
```

```