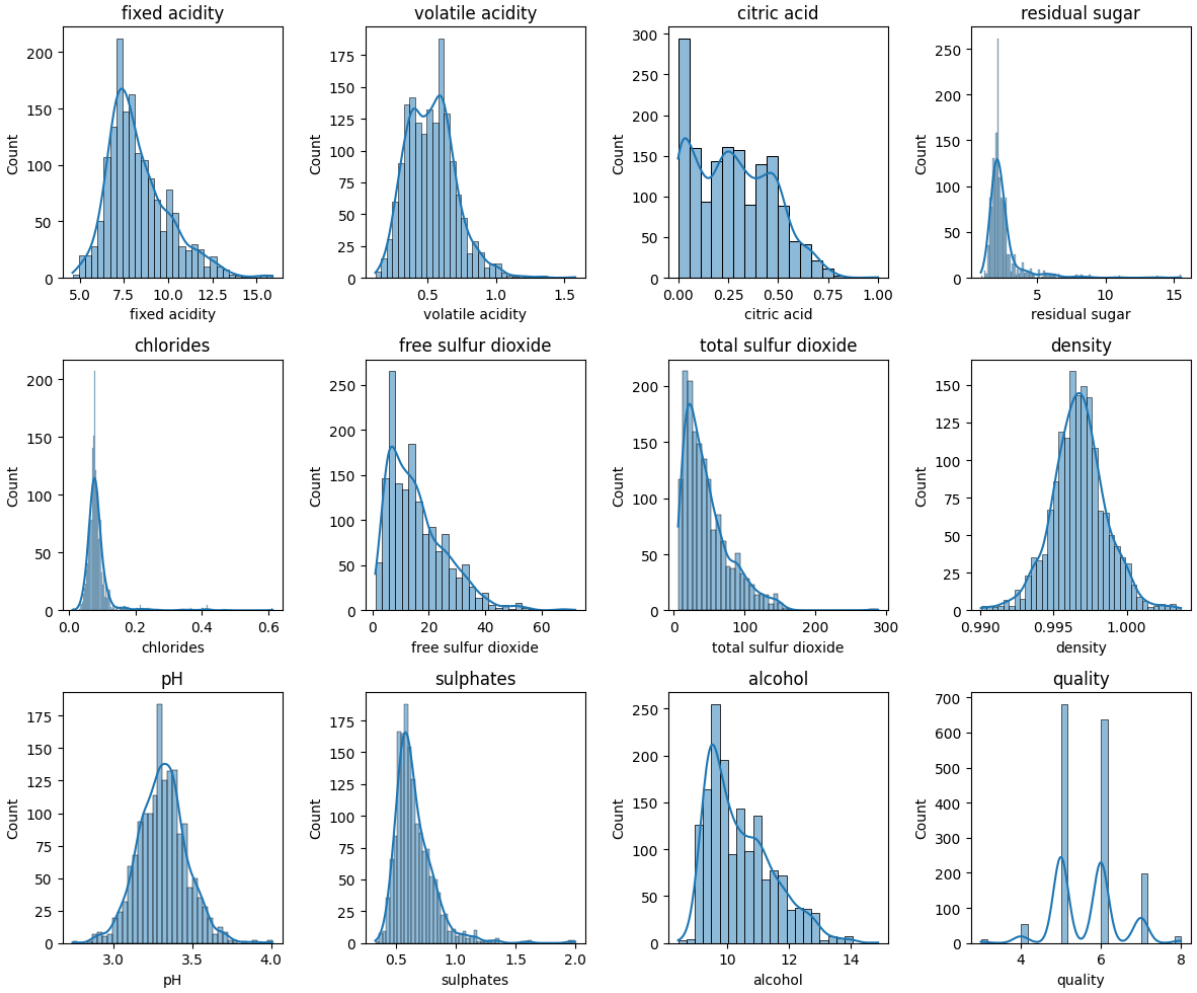
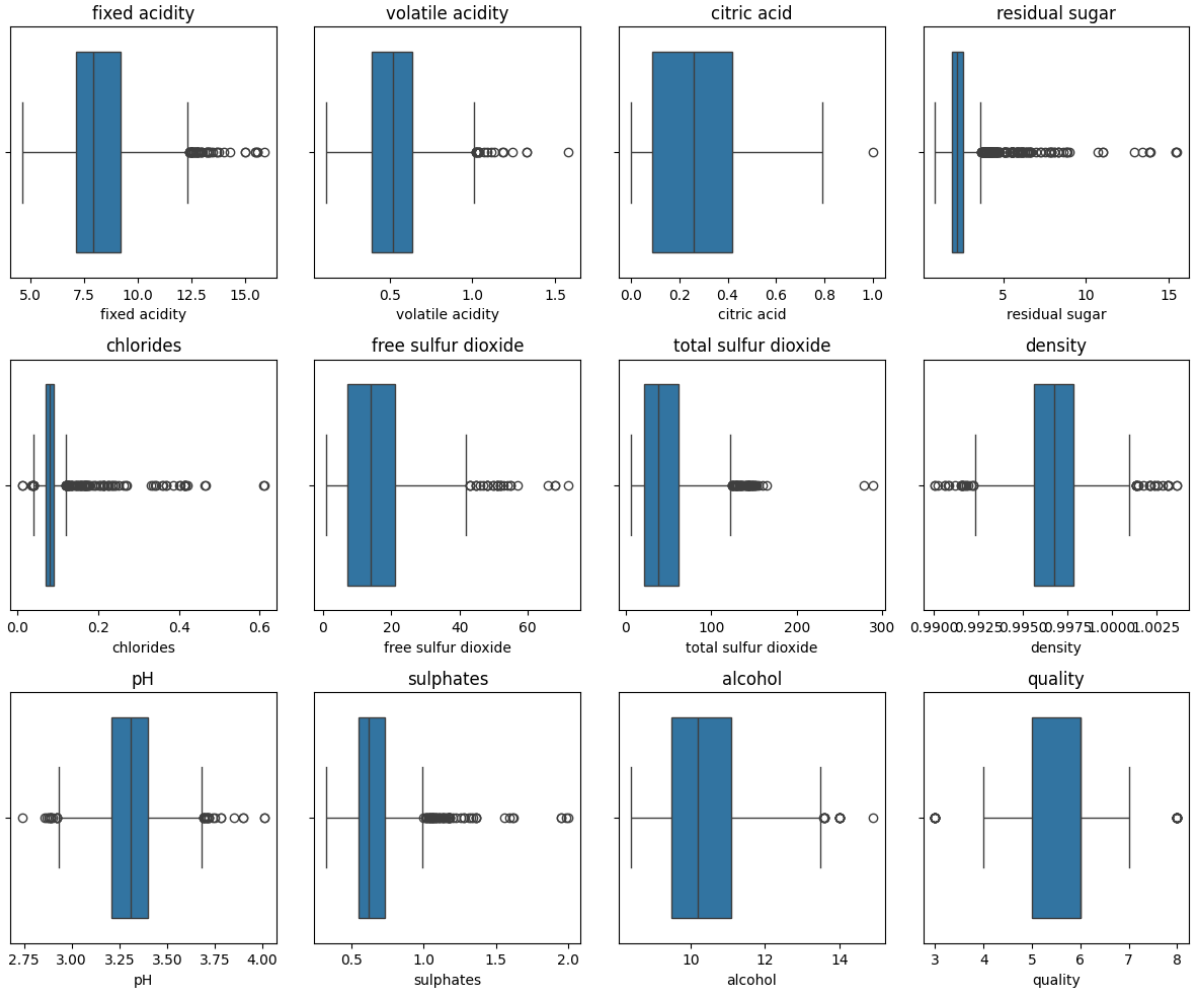
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# 데이터 불러오기
data = pd.read_csv('winequality-red.csv')

# 데이터 확인
print(data.head())

# 각 column의 데이터 타입 및 결측치 확인
print(data.info())

# 각 column의 기술통계량 확인
print(data.describe())

# 각 column의 분포 시각화
plt.figure(figsize=(12, 10))
for i, col in enumerate(data.columns):
    plt.subplot(3, 4, i + 1)
    sns.histplot(data[col], kde=True)
    plt.title(col)
plt.tight_layout()
plt.show()

# 이상치 확인 (boxplot으로 확인)
plt.figure(figsize=(12, 10))
for i, col in enumerate(data.columns):
    plt.subplot(3, 4, i + 1)
    sns.boxplot(x=data[col])
    plt.title(col)
plt.tight_layout()
plt.show()
```

```
   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
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
...
25%       3.210000     0.550000     9.500000     5.000000
50%       3.310000     0.620000    10.200000     6.000000
75%       3.400000     0.730000    11.100000     6.000000
max       4.010000     2.000000    14.900000     8.000000
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

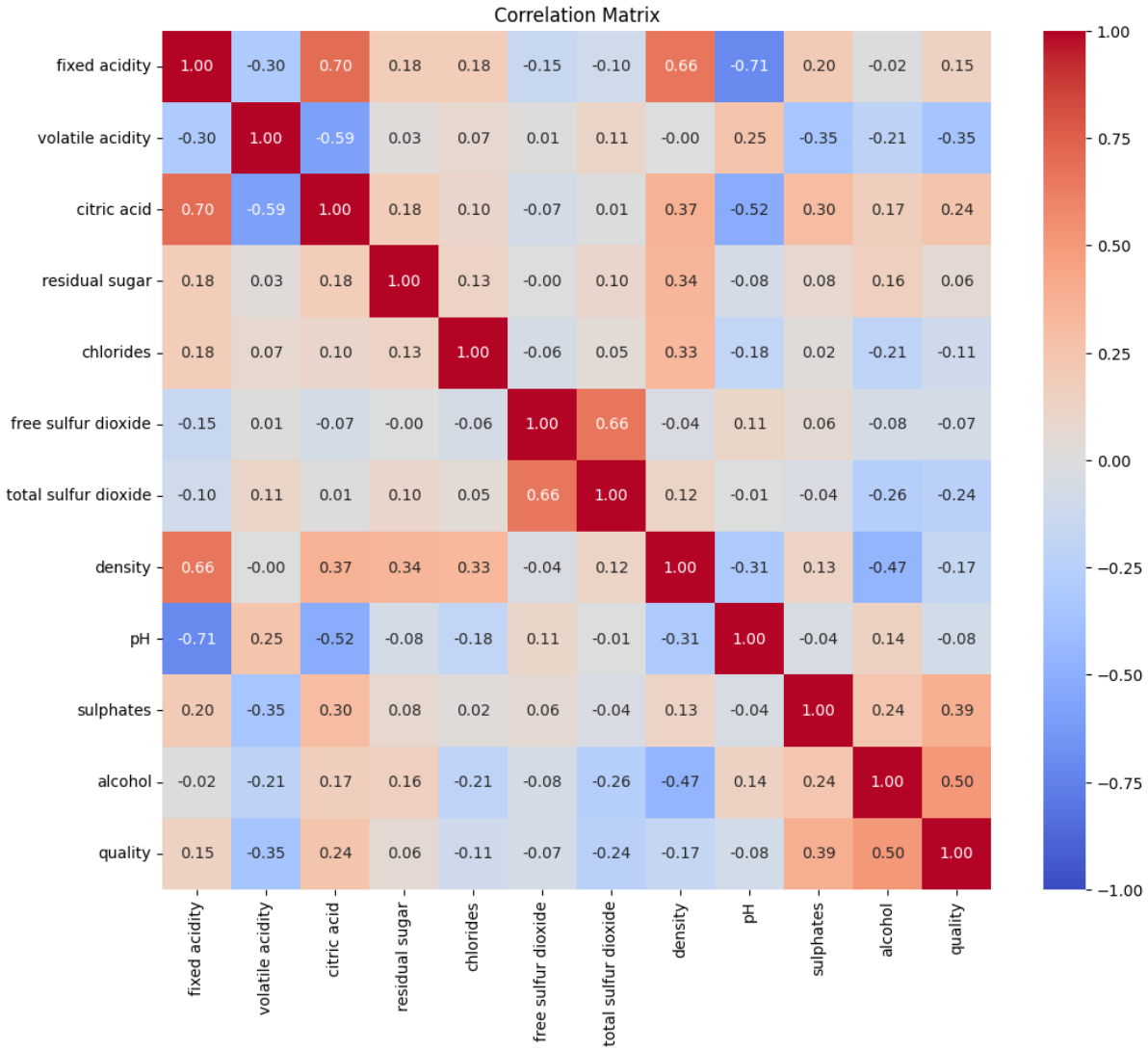

fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
Count

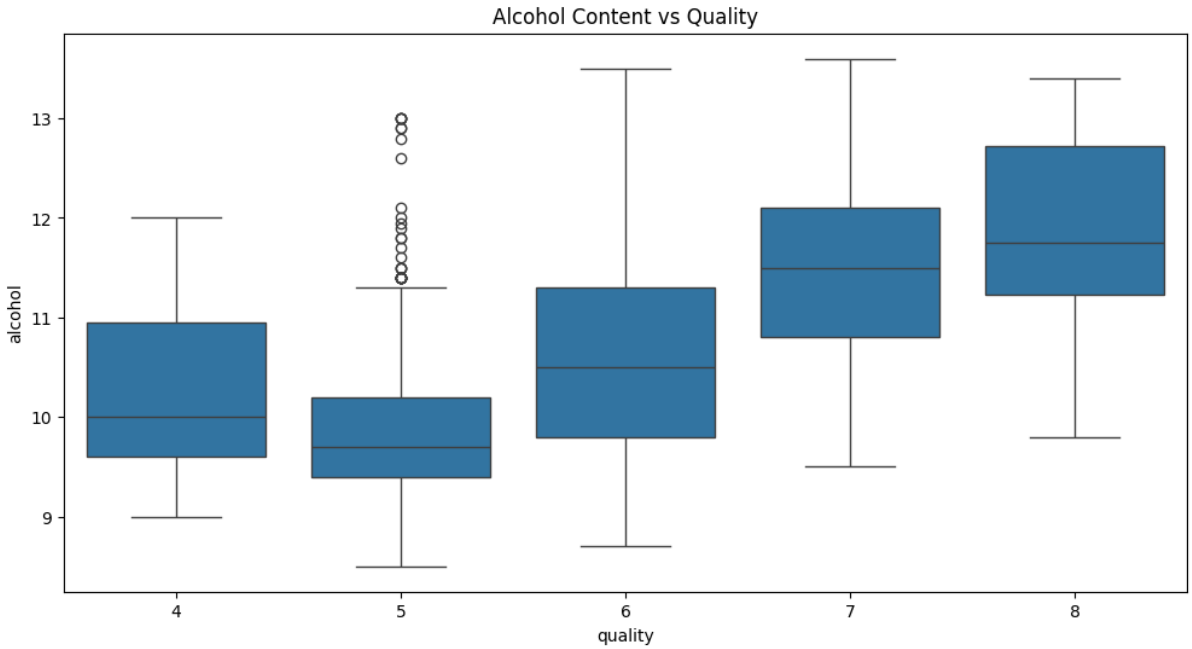

fixed acidity
5.0 7.5 10.0 12.5 15.0
fixed acidity
volatile acidity
0.5 1.0 1.5
volatile acidity
citric acid
0.0 0.2 0.4 0.6 0.8 1.0
citric acid
residual sugar
5 10 15
residual sugar
chlorides
0.0 0.2 0.4 0.6
chlorides
free sulfur dioxide
0 20 40 60
free sulfur dioxide
total sulfur dioxide
0 100 200 300
total sulfur dioxide
density
0.9900 0.9925 0.9950 0.9975 1.0000 1.0025
density
pH
2.75 3.00 3.25 3.50 3.75 4.00
pH
sulphates
0.5 1.0 1.5 2.0
sulphates
alcohol
10 12 14
alcohol
quality
3 4 5 6 7 8
quality

```python
# 결측치 확인 및 처리
print(data.isnull().sum())

# Z-score 이용한 이상치 제거
from scipy.stats import zscore

z_scores = zscore(data)
abs_z_scores = abs(z_scores)
filtered_entries = (abs_z_scores < 3).all(axis=1)
data = data[filtered_entries]

# Min-Max 데이터 스케일링
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
data_scaled = scaler.fit_transform(data)
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

```python
# 상관 분석
correlation_matrix = data.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", vmin=-1, vmax=1)
plt.title('Correlation Matrix')
plt.show()

# 품질과 다른 변수들 간의 관계 시각화 (예시)
plt.figure(figsize=(12, 6))
sns.boxplot(x='quality', y='alcohol', data=data)
plt.title('Alcohol Content vs Quality')
plt.show()
```






```python
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import mean_squared_error, r2_score

# 데이터 준비
X = data.drop('quality', axis=1)
y = data['quality']

# 데이터 분할
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
from sklearn.metrics import roc_auc_score
from itertools import combinations

# 모든 컬럼의 부분집합 생성
all_columns = X.columns
all_subsets = []

for i in range(1, len(all_columns) + 1):
    subsets_i = list(combinations(all_columns, i))
    all_subsets.extend(subsets_i)

# 모델 평가를 위한 함수 정의
def evaluate_model_auc(subset, X_train, X_test, y_train, y_test):
    # 모델 선택 (예시로 랜덤 포레스트 분류 모델 사용)
    model = RandomForestClassifier(random_state=42)

    # 선택된 변수들로 데이터셋 재구성
    X_train_subset = X_train[list(subset)]
    X_test_subset = X_test[list(subset)]

    # 모델 학습
    model.fit(X_train_subset, y_train)

    # 예측 확률 계산
    y_pred_proba = model.predict_proba(X_test_subset)

    # AUC 계산
    auc = roc_auc_score(y_test, y_pred_proba, multi_class='ovo')

    return auc

# 각 부분집합에 대해 AUC 계산
results = []
for subset in all_subsets:
    auc = evaluate_model_auc(subset, X_train, X_test, y_train, y_test)
    results.append((subset, auc))
```

```python
# 결과를 AUC 기준으로 정렬
sorted_results = sorted(results, key=lambda x: x[1], reverse=True)

# 상위 10개 부분집합에 대한 결과 출력
for subset, auc in sorted_results[:10]:
    print(f"Subset: {subset}")
    print(f"AUC: {auc:.4f}")
    print()
```

```
Subset: ('volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'sulphates', 'alcohol')
AUC: 0.8531

Subset: ('volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'density', 'sulphates', 'alcohol')
AUC: 0.8526

Subset: ('volatile acidity', 'residual sugar', 'free sulfur dioxide', 'density', 'pH', 'sulphates')
AUC: 0.8495

Subset: ('citric acid', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'density', 'sulphates', 'alcohol')
AUC: 0.8492

Subset: ('fixed acidity', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'pH', 'sulphates', 'alcohol')
AUC: 0.8479

Subset: ('volatile acidity', 'residual sugar', 'free sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol')
AUC: 0.8479

Subset: ('fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'total sulfur dioxide', 'density', 'sulphates', 'alcohol')
AUC: 0.8472

Subset: ('citric acid', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'pH', 'sulphates', 'alcohol')
AUC: 0.8463

Subset: ('volatile acidity', 'residual sugar', 'density', 'sulphates', 'alcohol')
...

Subset: ('fixed acidity', 'volatile acidity', 'residual sugar', 'free sulfur dioxide', 'density', 'sulphates', 'alcohol')
AUC: 0.8461
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...