DATA C'EPT

SESSION 2 :EXPLOARARY DATA ANALYSIS

PREPARED BY :
IHEB CHAABANE

RECAP

Data

Features of a house: size, nbe of rooms,location...

Black Box

Generating a Rule

Y

Price of the house

looking for patterns,relationship between variables..

Intuition behind machine Learning

BUT HOW?

STEPS

EDA

investigate the data and summarize the key insights

FEATURE ENG

Creating features to help the model decide

MODELLING

Defining ,Adapting and Tuning a Learning algorithm

EVALUATING

Evaluate the results and conclude

OUR STRATEGY:REVERSE LEARNING

We take a quick look at all the steps and then we organize sessions to dive deeper like: advanced missing values handling,Dealing with categorical features,TimeSeries,Dealing with imbalanced datasets,Regularization and confidence,cross-validation techniques....

EDA:

EDA is a process of performing some initial **investigations** on the dataset to discover the structure and the content of the given dataset.

It involves:

- Importing a dataset
- Understanding the big picture
- Preparation
- Understanding of variables
- Study of the relationships between variables
- Brainstorming

WHY PERFORMING EDA?

- An essential part in order to understand the problem
- It helps in Feature Selection
- It helps in Feature Engineering
- It helps in choosing the appropriate learning algorithm: for example if a feature is linearly highly correlated to the target ,linear regression will probably perform well.

LET'S TAKE AN EXAMPLE:THE WINE DATASET

OBJECTIVE:

Using chemical composition of wines in order to predict the quality of it.

LOADING THE DATASET

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df =pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv',sep=';')
```

LET'S TAKE A LOOK HOW THE DATA LOOKS LIKE

```
df.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

SHAPE:

```
df.shape
```

```
(1599, 12)
```

number of columns(features+target)

number of rows(examples)

DESCRIPTIVE INFORMATION ABOUT THE DATASET:

```
df.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

LET'S DIVE DEEPER

- are there any useless or redundant variables?
- are there any duplicate columns,rows?
- does the nomenclature make sense?
- are there any new variables we want to create?

DROPPING DUPLICATES :

```
df.duplicated().sum()
```

```
240
```

```
df=df.drop_duplicates()
```

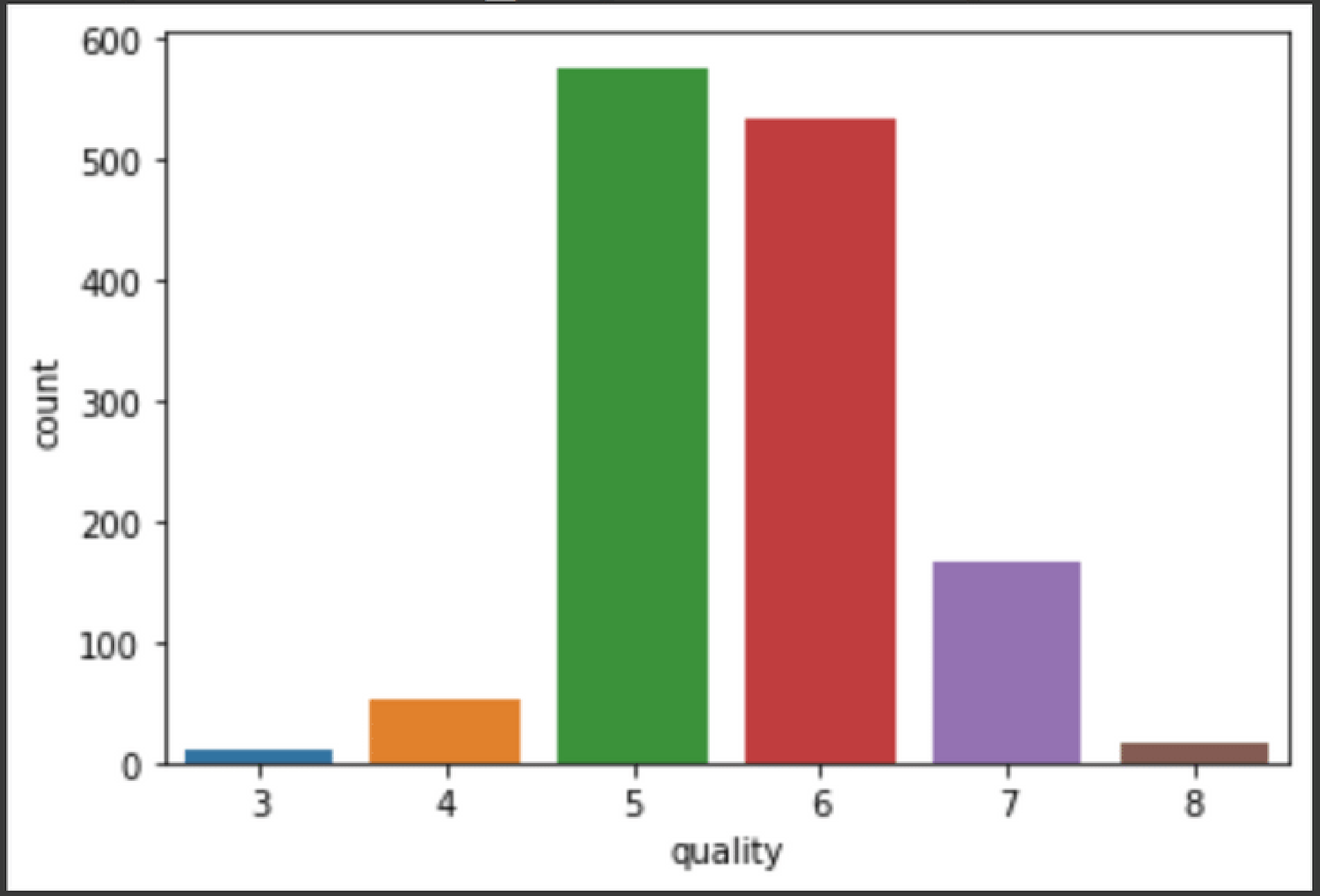
DISTRIBUTION OF THE TARGET

We're dealing with a highly imbalanced datasets,
Some classes are underrepresented .

We will learn how to deal with imbalanced Datasets
in the upcoming sessions.

```
sns.countplot(df['quality'])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_dec
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7fce188
```



MISSING VALUES :

```
df.isna().sum()/df.shape[0]
```

```
fixed acidity           0.0
volatile acidity        0.0
citric acid             0.0
residual sugar          0.0
chlorides               0.0
free sulfur dioxide     0.0
total sulfur dioxide    0.0
density                 0.0
pH                      0.0
sulphates               0.0
alcohol                 0.0
quality                 0.0
dtype: float64
```

--->No missing values

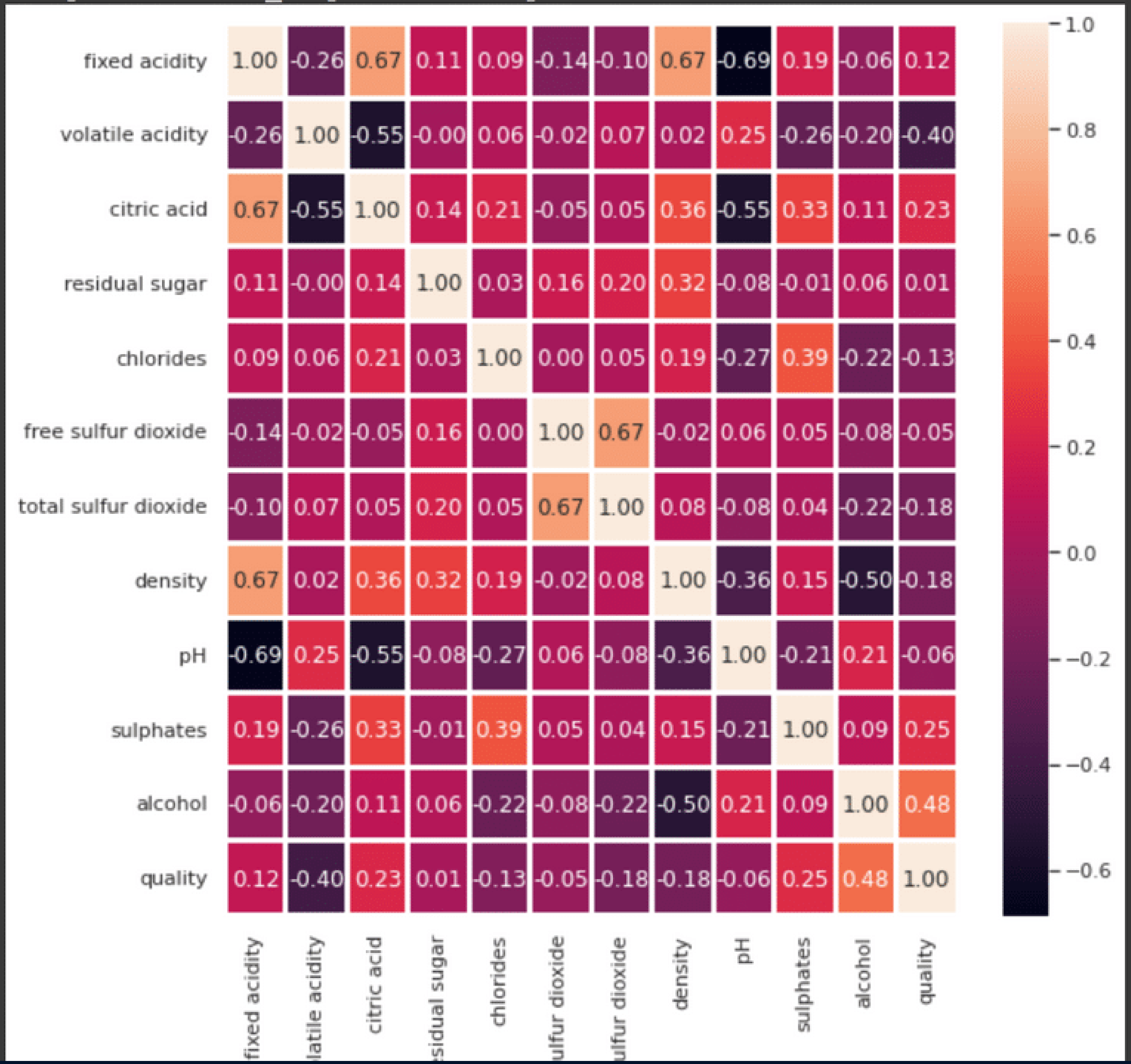
CORRELATION BETWEEN FEATURES :

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

it refers to the degree to which a pair of variables are linearly related.

```python
sns.set(rc={'figure.figsize':(9,9)})
sns.heatmap(df.corr(), annot=True, fmt='.2f', linewidth=2)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fce15e69250>
```



Quick Observation

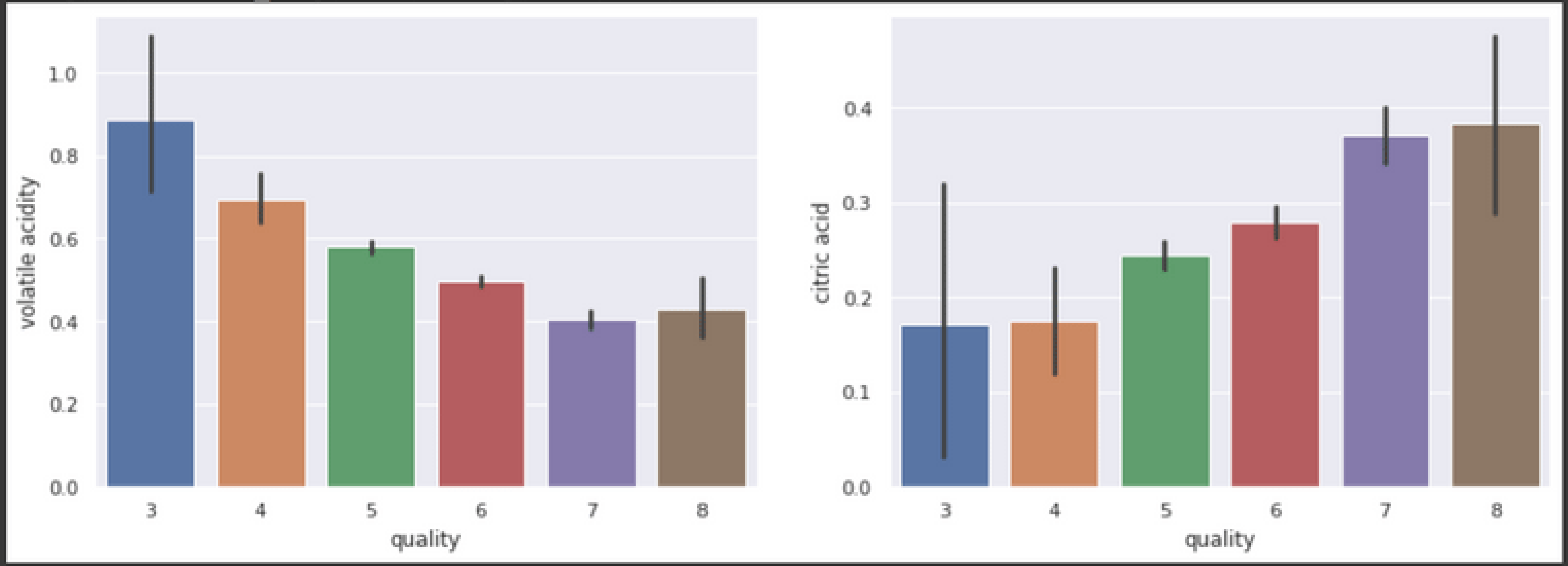
- fixed acidity has high correlation(>0.6) with density and citric acid and a negative correlation(<-0.69) with pH.
- volatile acidity seems to have a negative effect on quality (-0.41).
- free sulfur dioxide is more if total sulfur dioxide is more and vice versa.
- alcohol is having influence on quality of wine.
- alcohol and density have a negative correlation(-0.49)
- pH,free and total sulfure dioxide ,chlorides ,residual sugar do not have high correlation on quality

Ps : This only reflects linear relationship between variables

ANALYSIS:

```
fig = plt.figure(figsize=(15,5))
plt.subplot(121)
sns.barplot(data =df,x ='quality',y = 'volatile acidity')
plt.subplot(122)
sns.barplot(data =df,x ='quality',y = 'citric acid')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fce18a15d50>
```



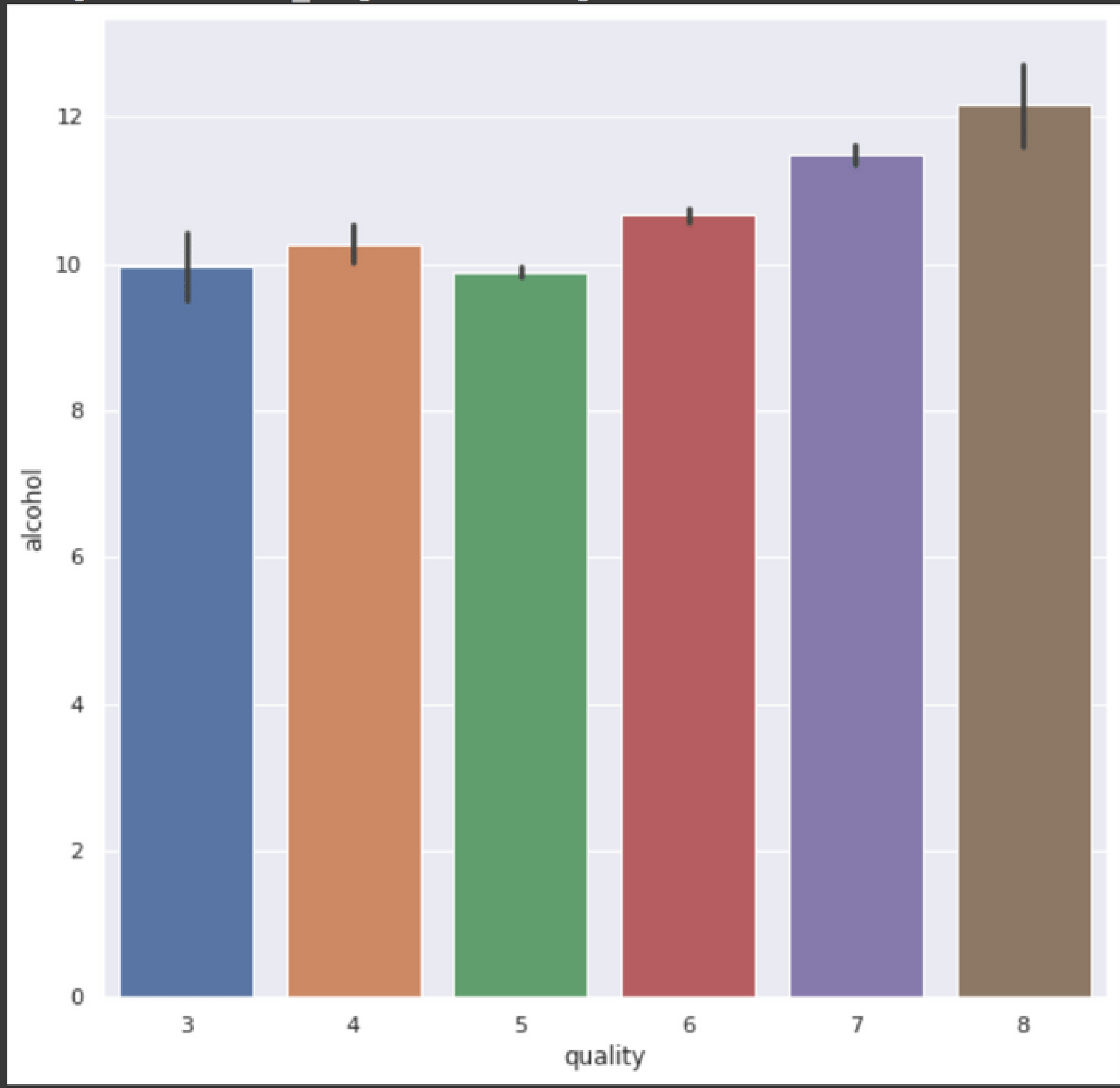
Quick Observation

- It's very distinctive from the above visualtion that a wine is of top quality if it's volatile acidity is less.
- Top quality wines usually have high citric acid.

LET'S CHECK HOW QUALITY OF THE WINE IS AFFECTED BY ALCOHOL CONTENT:

```
sns.barplot(data =df,x ='quality',y = 'alcohol')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fce15e352d0>
```



So, Quality of the wine improves when we increase it's alcohol content.

THANK YOU FOR YOUR ATTENTION!