report_py

March 9, 2023

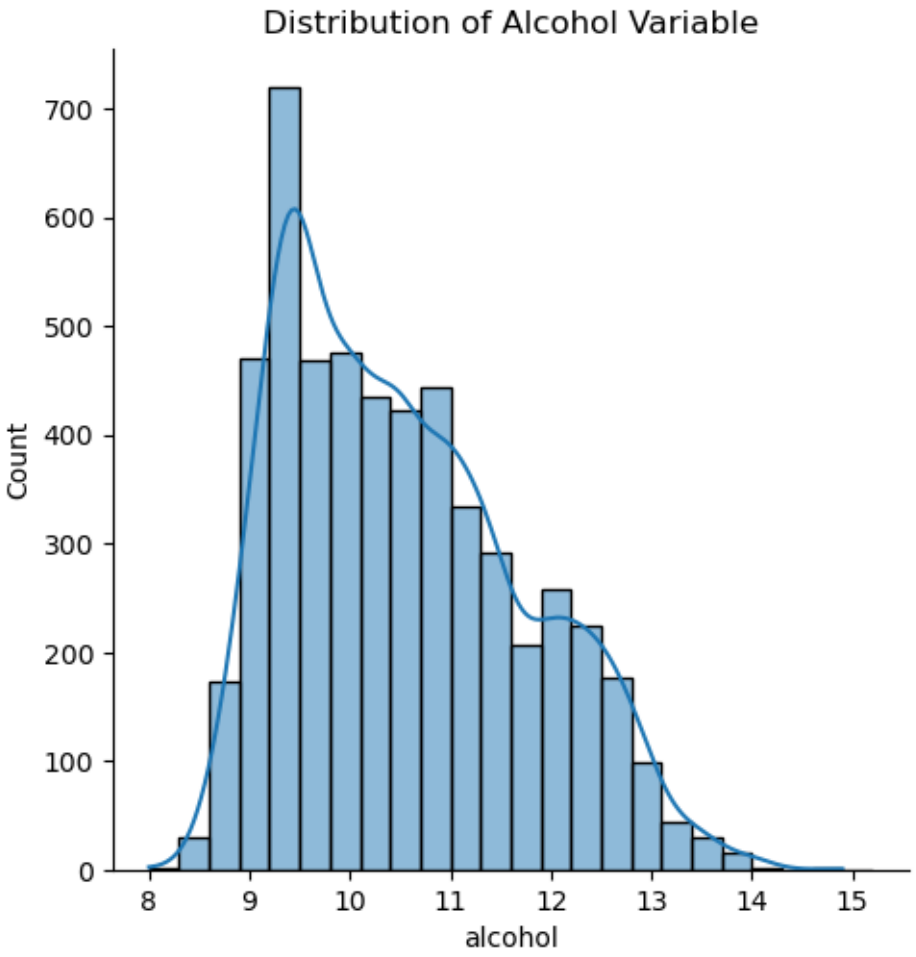
[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib
import seaborn as sns
```

[ ]:
```
wine = pd.read_csv("winequality-all.csv", comment="#")
```

0.0.1 Task 1

By exploring the distribution of alcohol levels, we can gain insights into the range and spread of the data, as well as any potential outliers or patterns. This information can be useful in identifying trends and relationships within the dataset and informing future analyses or modeling.

[ ]:
```
sns.displot(data = wine, x = "alcohol", binwidth = 0.3, kde=True)
plt.title("Distribution of Alcohol Variable")
plt.show()
```



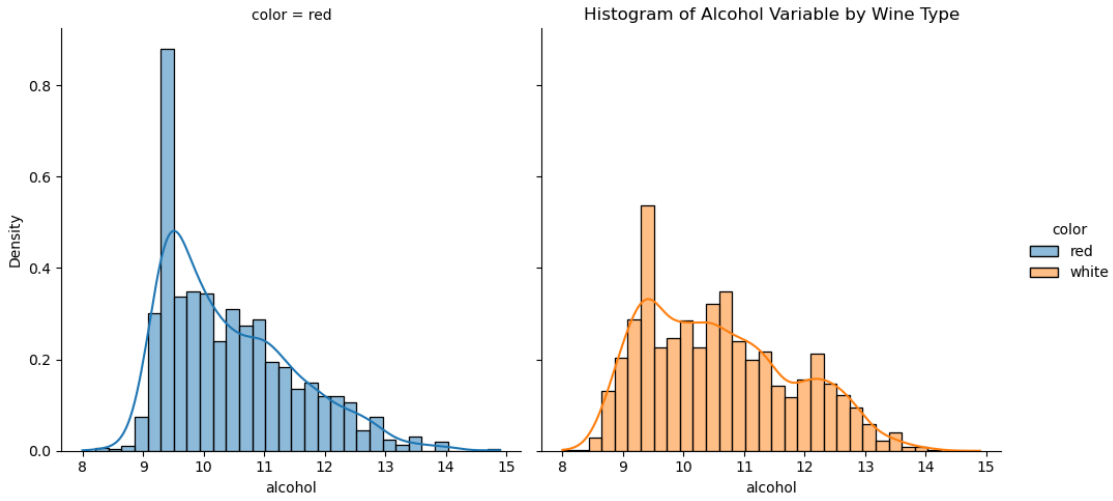
We see that much of the mass lies between 9% and 10% alcohol. After reaching an initial peak, the curve monotonically falls around zero at 14%

0.0.2 Task 2

By comparing the two distributions, we can gain insights into the differences and similarities in alcohol levels between the two types of wine, and identify any potential patterns or outliers within each dataset.

[ ]:
```
hist = sns.displot(data = wine, x = "alcohol", hue = "color", element = "bars",␣
↪stat = "density", multiple="stack", kde=True, col="color", ␣
↪common_norm=False)
plt.title("Histogram of Alcohol Variable by Wine Type")
```

[ ]:
```
Text(0.5, 1.0, 'Histogram of Alcohol Variable by Wine Type')
```

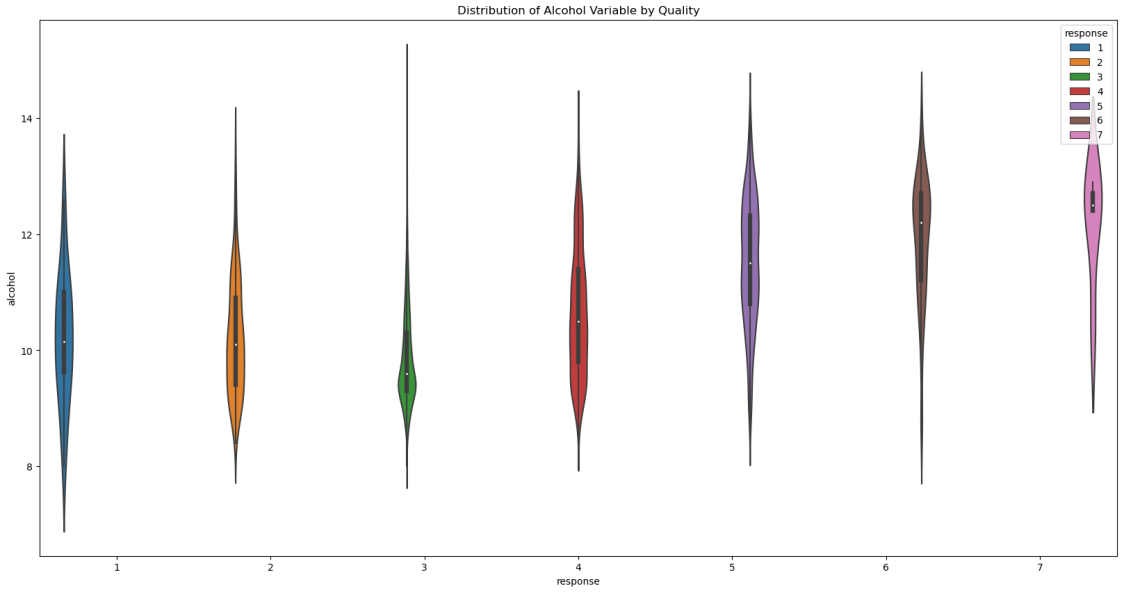


We can see that the distribution of alcohol in red wine is decidedly skewed to the left, while in white wines the alcohol is distributed somewhat more evenly.

0.0.3 Task 3

The task of comparing the distribution of alcohol variable in each possible quality group defined by a response variable involves analyzing the distribution of alcohol levels within each group and comparing them to one another.

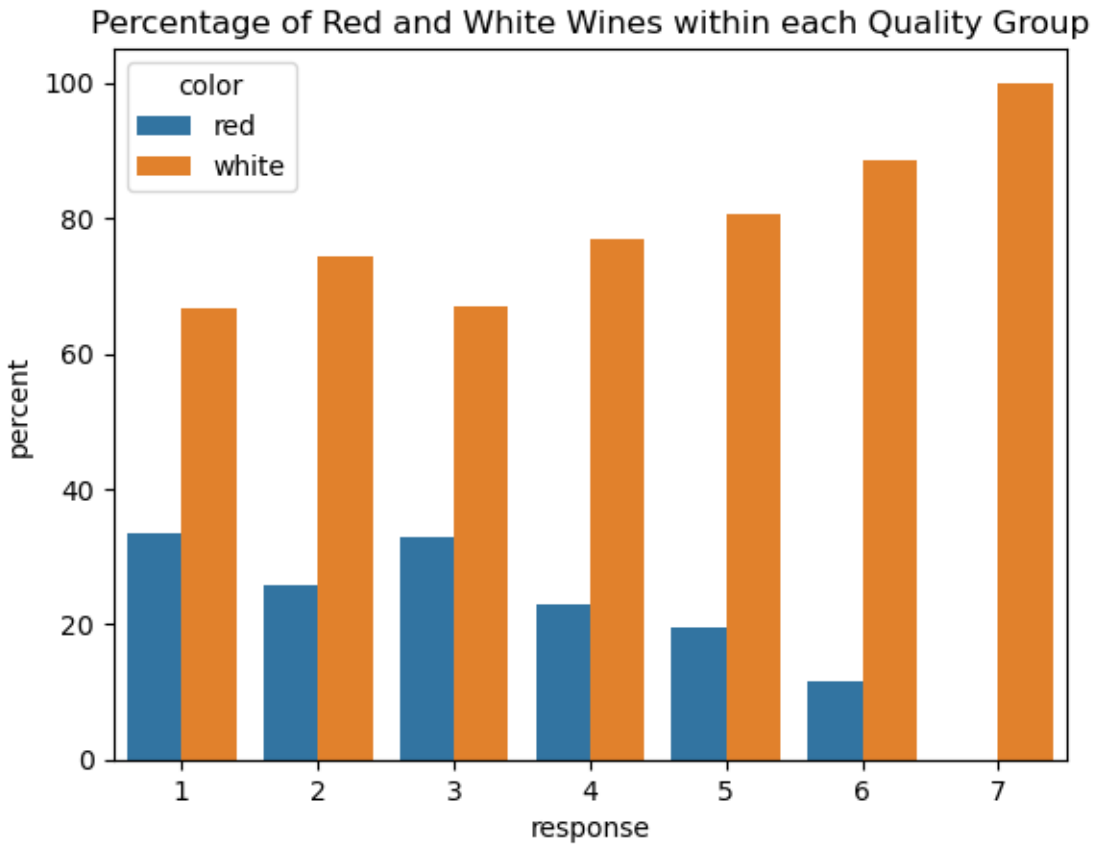
[ ]:
```
f = plt.figure()
f.set_figwidth(20)
f.set_figheight(10)
sns.violinplot(data = wine, y = "alcohol", x="response", hue = "response",␣
 ↪width=0.8)
plt.title("Distribution of Alcohol Variable by Quality")
plt.show()
```



It can be observed that from rating 3 onward, there is a definite positive linear relationship between the rating and the median amount of alcohol in the wines of the group.

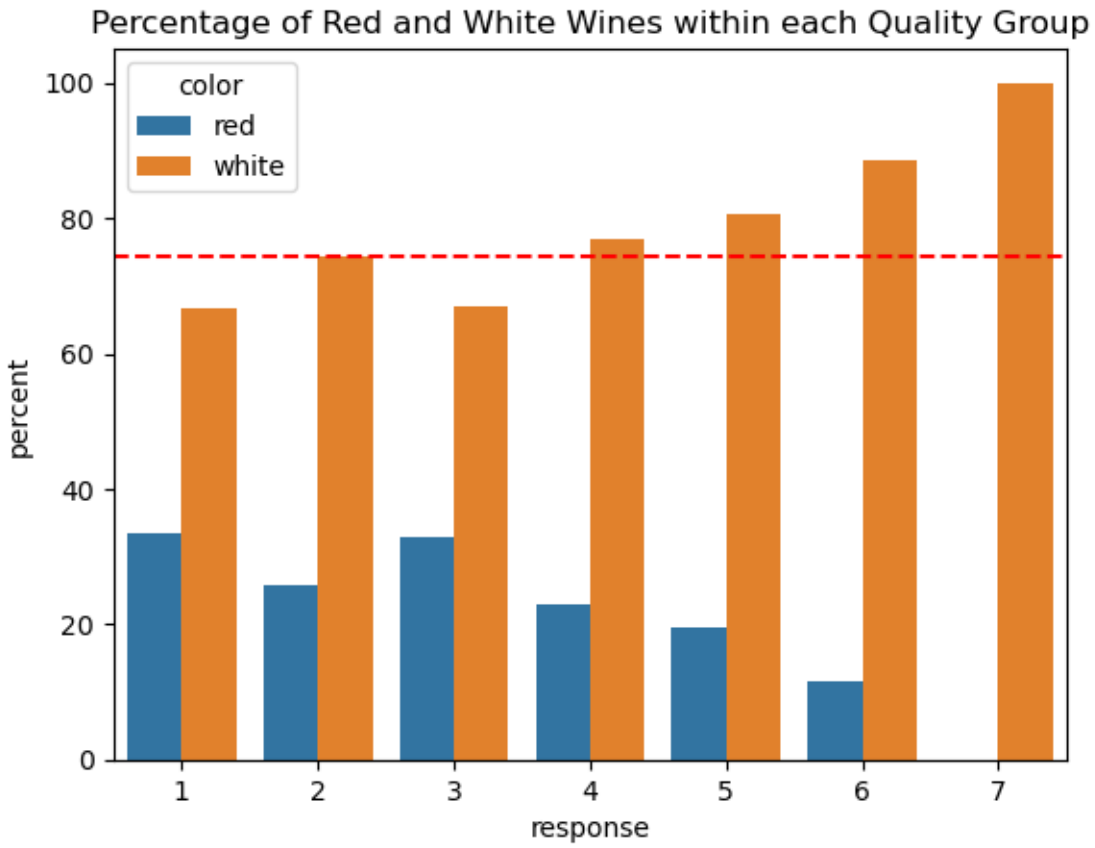
0.0.4 Task 4

[ ]:
```
quality_counts = wine.groupby(["response", "color"]).size().reset_index(name =␣
↪"count_color")
response_counts = wine.groupby(["response"]).size().reset_index(name = "count")
quality_counts = quality_counts.merge(response_counts, how="inner",␣
↪on="response")
quality_counts["percent"] = quality_counts["count_color"] /␣
↪quality_counts["count"] * 100
sns.barplot(data = quality_counts, x = "response", y = "percent", hue = "color")
plt.title("Percentage of Red and White Wines within each Quality Group")
plt.show()
```



It can be hypothesized that the better the wine rating, the less likely it is to be a red wine. However, keep in mind the predominance of white wine over red wine in the entire data set. We can set this proportion as a kind of benchmark and see in which group it is exceeded by white wine.

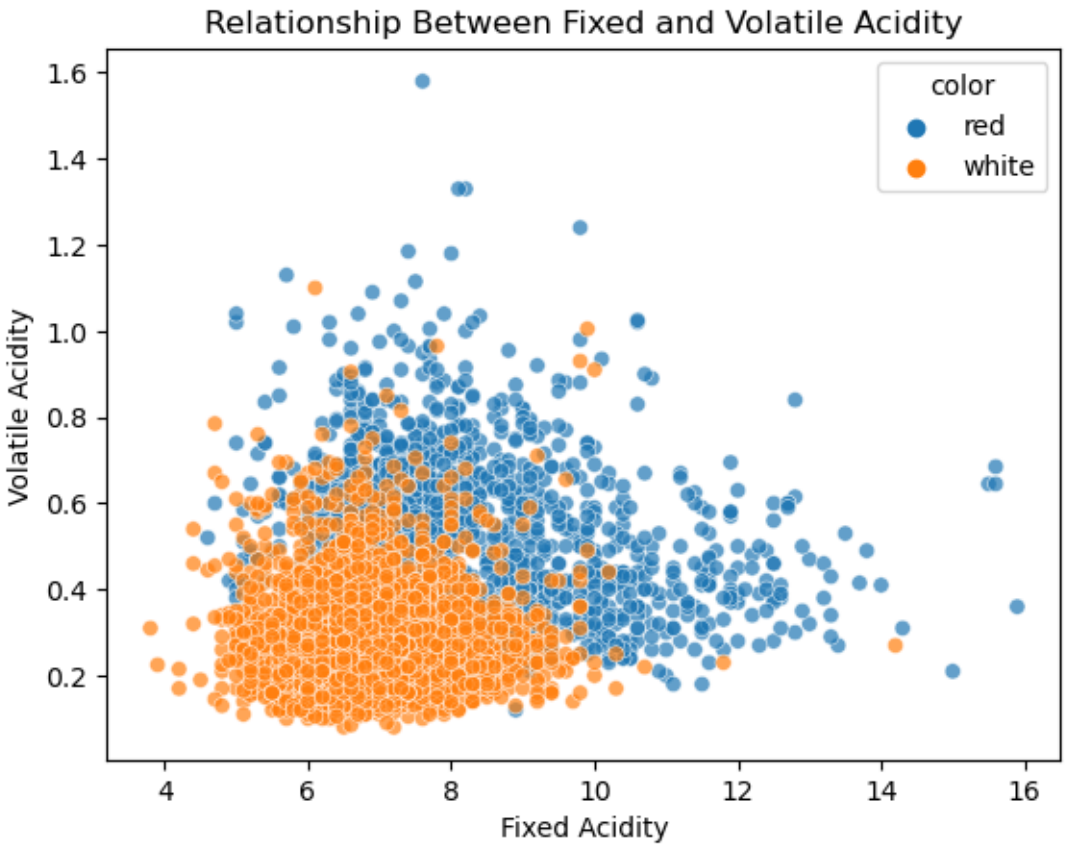
[ ]:
```
white_proportion = wine.loc[wine["color"]=='white'].shape[0]/wine.shape[0]
sns.barplot(data = quality_counts, x = "response", y = "percent", hue = "color")
plt.title("Percentage of Red and White Wines within each Quality Group")
plt.axhline(y=white_proportion*100, color='red', linestyle='--')
plt.show()
```



As we can see with the weaker ratings, there is relatively less white wine in relation to the proportion of the entire dataset. In the case of better grades, the situation is reversed.

0.0.5 Task 5

[ ]:
```
sns.scatterplot(data=wine, x="fixed.acidity", y="volatile.acidity", alpha=0.7,␣
↪hue="color")
plt.xlabel("Fixed Acidity")
plt.ylabel("Volatile Acidity")
plt.title("Relationship Between Fixed and Volatile Acidity")
plt.show()
```



We can see that despite the partial overlap between the two harvests, red wine has a much higher variance for both axes. Interestingly, in both cases, we can see that higher values of one axis do not induce higher values on the other axis.