Wine Quality Analysis

PRASANTH BATTULA

Abstract

In this project, I am analyzing the wine quality dataset by using different variable i.e. fixed acidity, volatile acidity, citric acid, free sulfur dioxide, chlorides, sulfates, density, residual sugar, alcohol, total sulfur dioxide, and pH.

Motivation

Nowadays many numbers of people are taking wine. But did they know they are drinking quality wine or not? The wine consists of lots of ingredients and every ingredient will affect the quality of the wine. I want to predict how this wine quality is changing with respect to the amount of each ingredient.

Dataset

Wine Quality dataset.

Source: Kaggle

Data Preparation and Cleaning

- Checked for null values

Check if any null values exists

```
In [4]: data.isnull().any()
```

```
Out[4]: fixed acidity           False
        volatile acidity        False
        citric acid             False
        residual sugar          False
        chlorides               False
        free sulfur dioxide     False
        total sulfur dioxide    False
        density                 False
        pH                      False
        sulphates               False
        alcohol                 False
        quality                 False
        dtype: bool
```

No null values!!!

Data Preparation and Cleaning

- Explored the data and observed the range of values of each attribute

	mean	std	min	25%	50%	75%	max
fixed acidity	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000
volatile acidity	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000
citric acid	0.270976	0.194801	0.00000	0.0900	0.26000	0.420000	1.00000
residual sugar	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000
chlorides	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100
free sulfur dioxide	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000
total sulfur dioxide	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000
density	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369
pH	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000
sulphates	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000
alcohol	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000
quality	5.636023	0.807569	3.00000	5.0000	6.00000	6.000000	8.00000

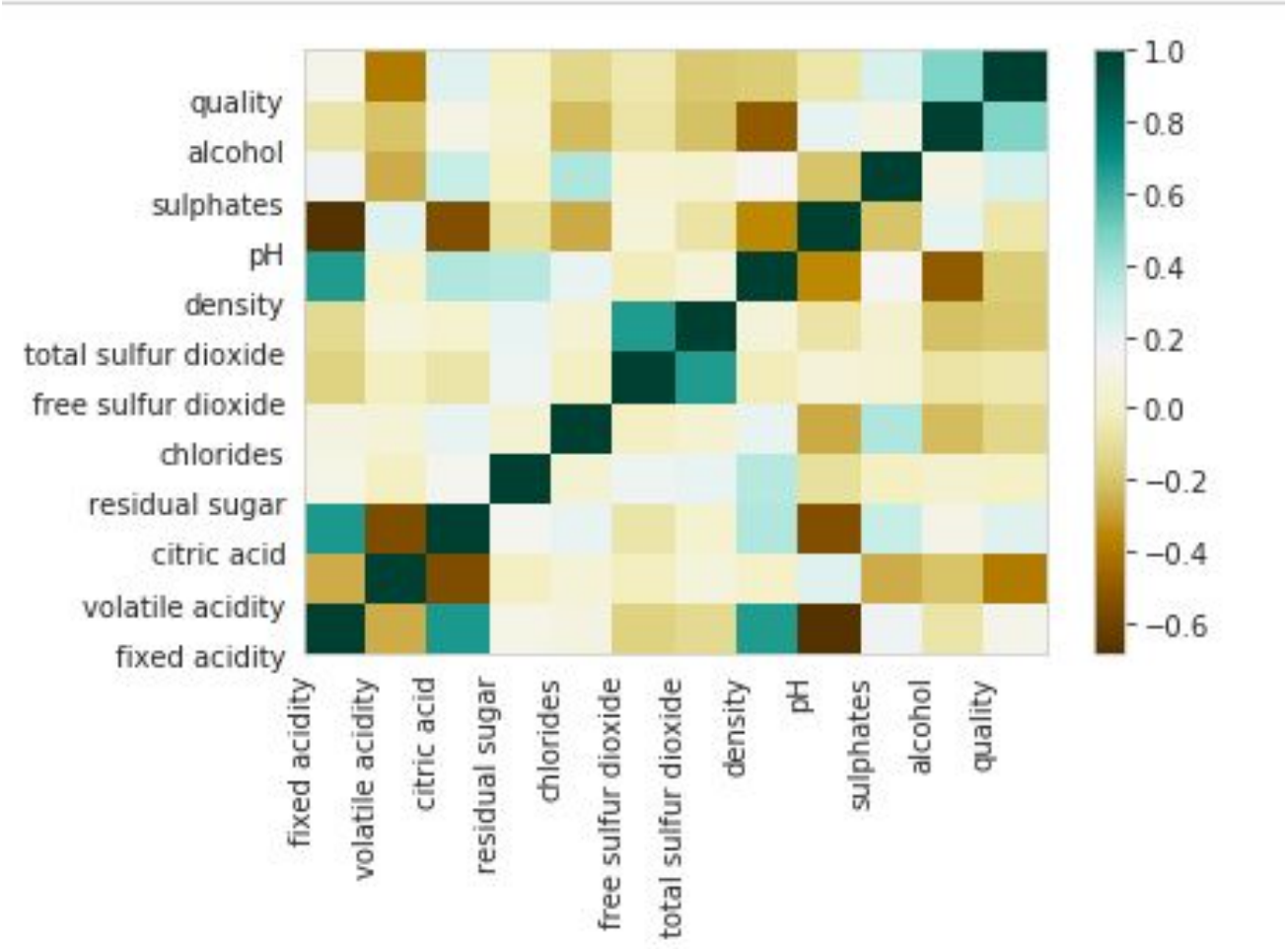
Research Questions

1. Classification of Wine into three different categories based on the quality of wine
2. How residual sugar, density, and alcohol affect the quality of wine
3. How the values of variables are distributed
4. Effect of other variables on wine quality

Methods

For this analysis, I have used pandas for exploring the data, matplotlib and seaborn for data visualization.

Findings

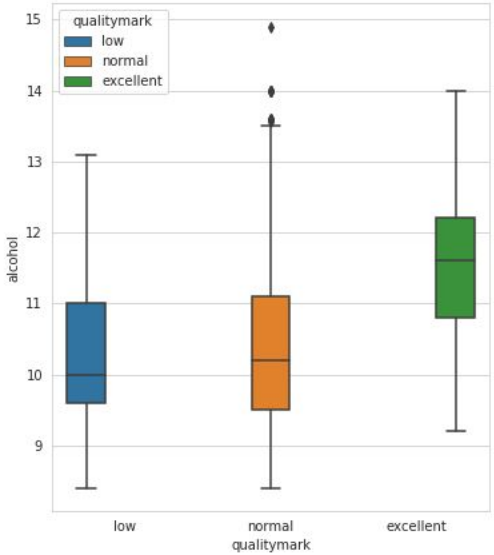

quality
alcohol
sulphates
pH
density
total sulfur dioxide
free sulfur dioxide
chlorides
residual sugar
citric acid
volatile acidity
fixed acidity
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

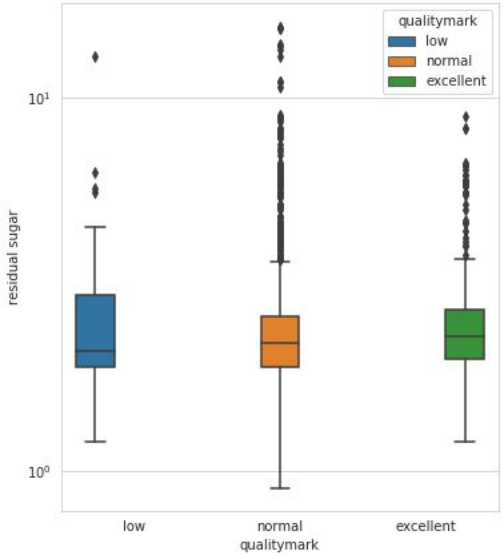
Findings

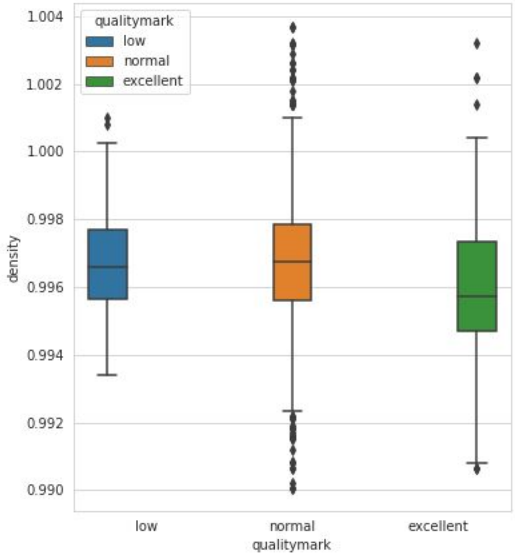
In the above slide, we can observe how every variable correlated with others.

Findings

Wine quality vs residual sugar, density and alcohol







**Please see bottom pages for high quality plots

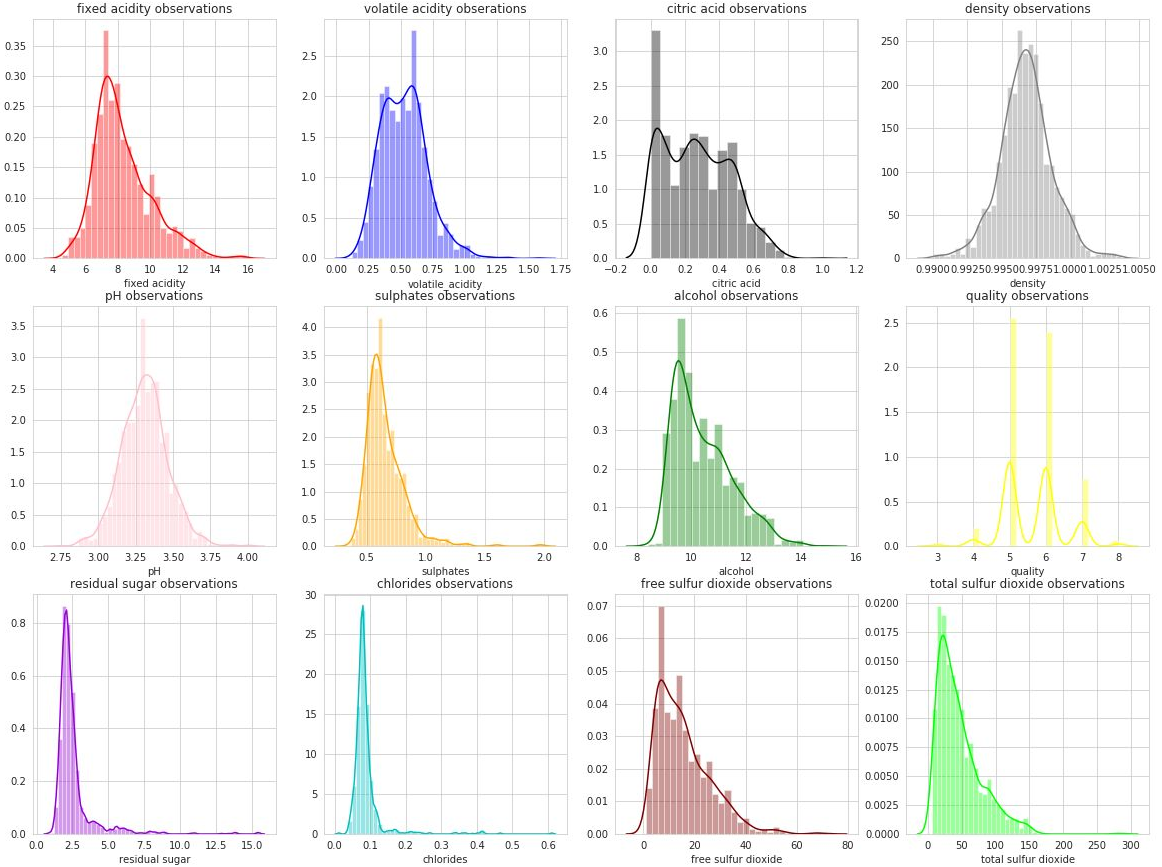
Findings

Conclusions from above boxplots:

- Wine quality is having a positive relationship with alcohol
- There is no correlation between quality and residual sugar
- Density vs Quality gives a negative correlation

Findings

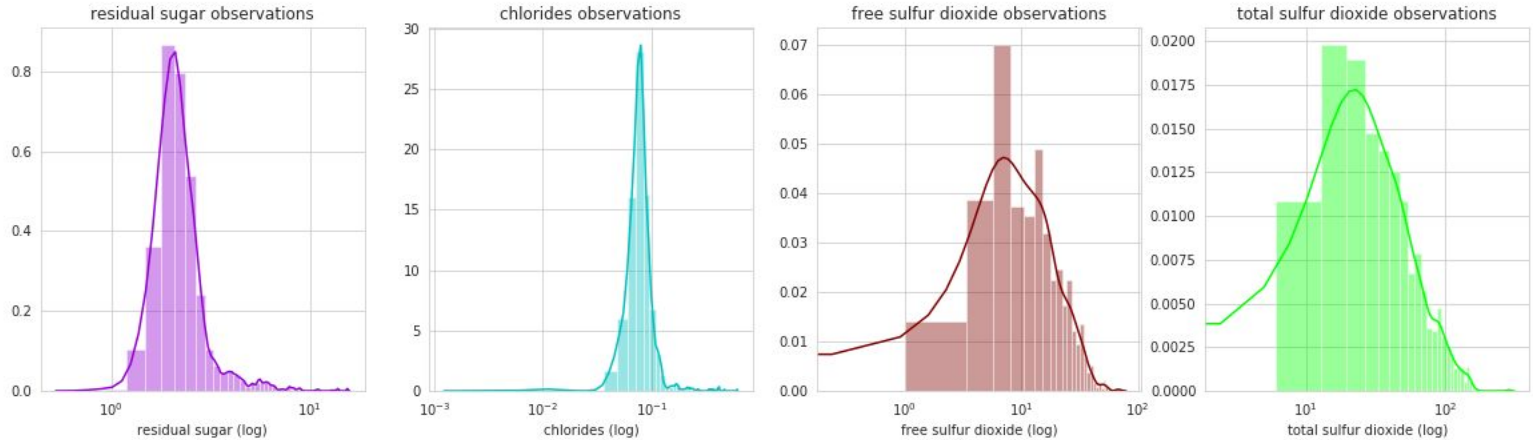
Univariate analysis



**Please see bottom pages for high quality plots

Findings

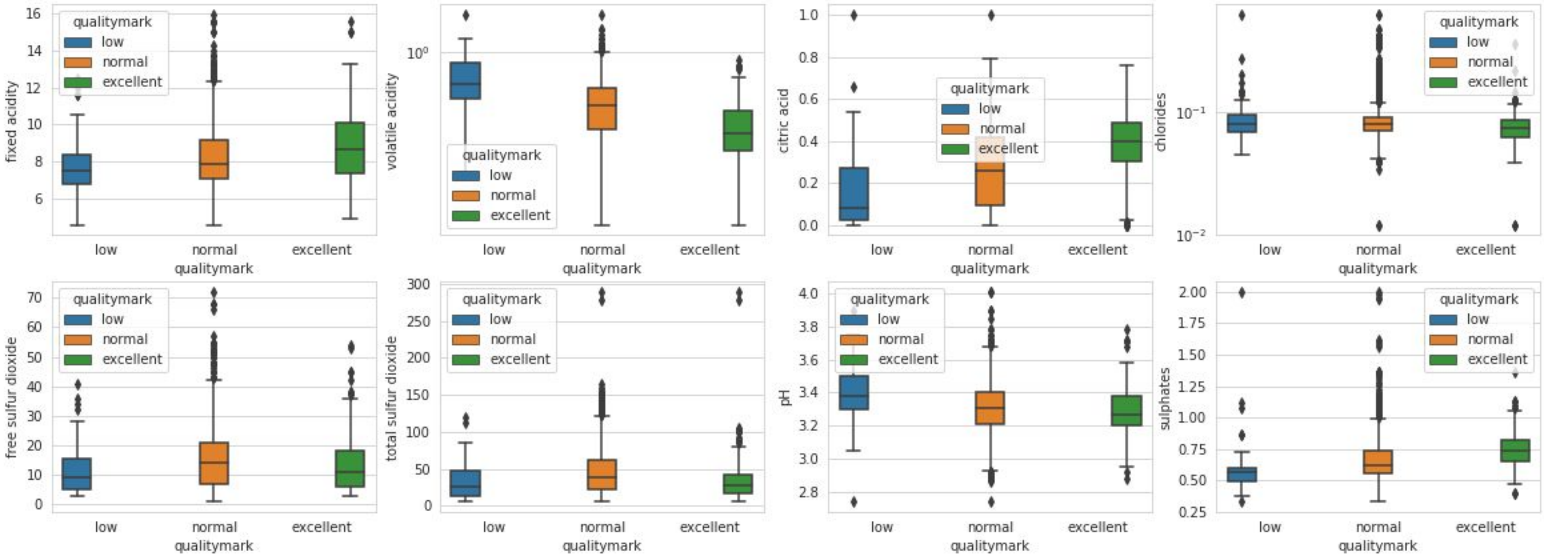
Univariate analysis



**Please see bottom pages for high quality plots

Findings

Wine quality with respect to other variables



**Please see bottom pages for high quality plots

Limitations

- We can observe in box plots, there are lot of outliers in almost all variables for medium quality wine.

Conclusions

- Fixed acidity, Citric acid, Sulphates are having a positive relationship with the quality of wine
- Low volatile acidity, low pH and low chloride values are resulting in good quality of the wine.
- Wine quality is having a positive relationship with alcohol
- There is no correlation between quality and residual sugar
- Density vs Quality gives a negative correlation

Acknowledgements

Data Source: Kaggle.com

References

1. Concepts learned in Python for Data Science.
2. Wine quality dataset from Kaggle

Wine quality analysis

Objectives

- Classify the data into poor, normal and excellent quality wines based on the qulaity
- Visualize how residual sugar, density and alcohol effect the quality of wine
- Univariate analysis
- How others variables effect the quality of wine

In [1]:
```
# import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Read the wind quality dataset using pandas library and display first five rows

In [2]:
```
data = pd.read_csv('./data/winequality-red.csv')
data.head()
```

Out[2]:

	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

Display range of values for all columns using the descirbe method

In [3]:
```
des = data.describe()
des.drop(['count'],axis=0,inplace=True)
des.T
```

Out[3]:

	mean	std	min	25%	50%	75%	max
fixed acidity	8.319637	1.741096	4.60000	7.1000	7.90000	9.200000	15.90000
volatile acidity	0.527821	0.179060	0.12000	0.3900	0.52000	0.640000	1.58000
citric acid	0.270976	0.194801	0.00000	0.0900	0.26000	0.420000	1.00000
residual sugar	2.538806	1.409928	0.90000	1.9000	2.20000	2.600000	15.50000
chlorides	0.087467	0.047065	0.01200	0.0700	0.07900	0.090000	0.61100
free sulfur dioxide	15.874922	10.460157	1.00000	7.0000	14.00000	21.000000	72.00000
total sulfur dioxide	46.467792	32.895324	6.00000	22.0000	38.00000	62.000000	289.00000
density	0.996747	0.001887	0.99007	0.9956	0.99675	0.997835	1.00369
pH	3.311113	0.154386	2.74000	3.2100	3.31000	3.400000	4.01000
sulphates	0.658149	0.169507	0.33000	0.5500	0.62000	0.730000	2.00000
alcohol	10.422983	1.065668	8.40000	9.5000	10.20000	11.100000	14.90000
quality	5.636023	0.807569	3.00000	5.0000	6.00000	6.000000	8.00000

Check if any null values exists

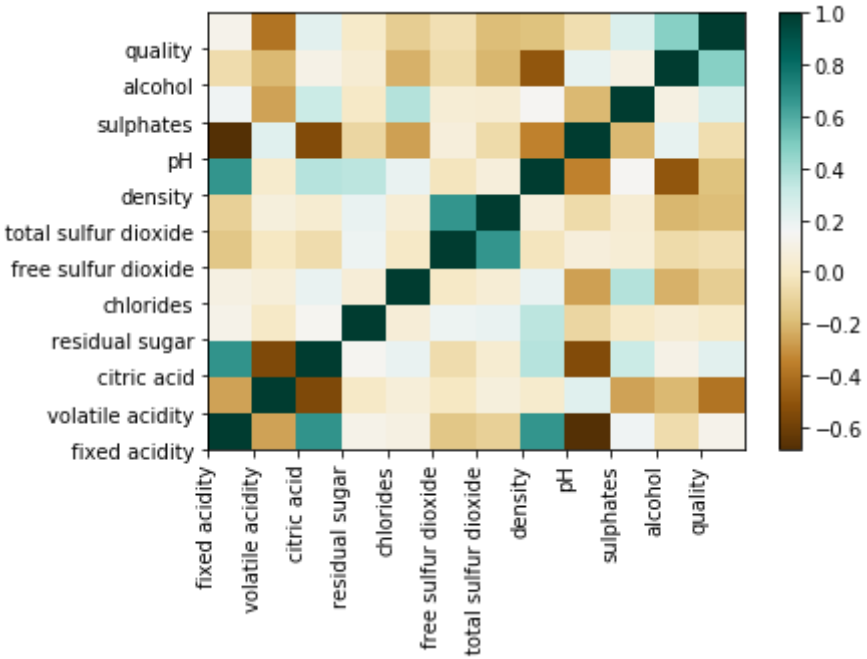
In [4]:
```
data.isnull().any()
```

Out[4]:
```
fixed acidity           False
volatile acidity        False
citric acid             False
residual sugar          False
chlorides               False
free sulfur dioxide     False
total sulfur dioxide    False
density                 False
pH                      False
sulphates               False
alcohol                 False
quality                 False
dtype: bool
```

No null values!!!

Visualize how one attribue correlated with other attributes

In [5]:
```
corr = data.corr(method='pearson')
fig = plt.figure().add_subplot(111)
plt.pcolor(corr, cmap='BrBG')
plt.colorbar()
labels = data.columns
fig.set_xticks(np.arange(len(labels)))
fig.set_yticks(np.arange(len(labels)))
fig.set_xticklabels(labels, rotation=90)
fig.set_yticklabels(labels, rotation=0)
plt.show()
fig.figure.savefig('corrrelation.png')
```



Add Quality mark column

In [6]:
```
Poor = data[data['quality']<5]
Normal = data[(data['quality']>=5) | (data['quality']<7)]
Excellent = data[data['quality']>=7]
Poor['qualitymark'] = 'low'
Normal['qualitymark'] = 'normal'
Excellent['qualitymark'] = 'excellent'

frames = [Poor, Normal, Excellent]
data_with_mark = pd.concat(frames)
```

```
/home/prasanth/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html
#returning-a-view-versus-a-copy
  after removing the cwd from sys.path.
/home/prasanth/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html
#returning-a-view-versus-a-copy
```

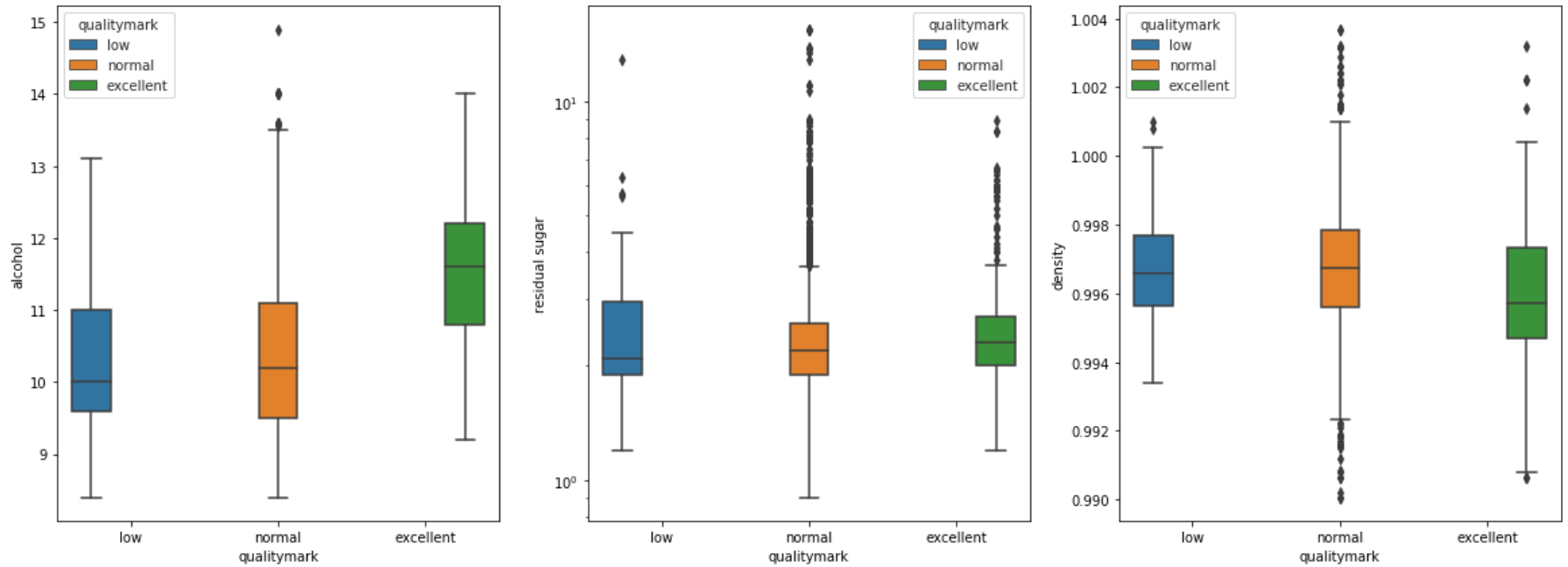
In [7]:
```
data_with_mark.head()
```

Out[7]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	qualitymark
18	7.4	0.590	0.08	4.4	0.086	6.0	29.0	0.9974	3.38	0.50	9.0	4	low
38	5.7	1.130	0.09	1.5	0.172	7.0	19.0	0.9940	3.50	0.48	9.8	4	low
41	8.8	0.610	0.30	2.8	0.088	17.0	46.0	0.9976	3.26	0.51	9.3	4	low
45	4.6	0.520	0.15	2.1	0.054	8.0	65.0	0.9934	3.90	0.56	13.1	4	low
73	8.3	0.675	0.26	2.1	0.084	11.0	43.0	0.9976	3.31	0.53	9.2	4	low

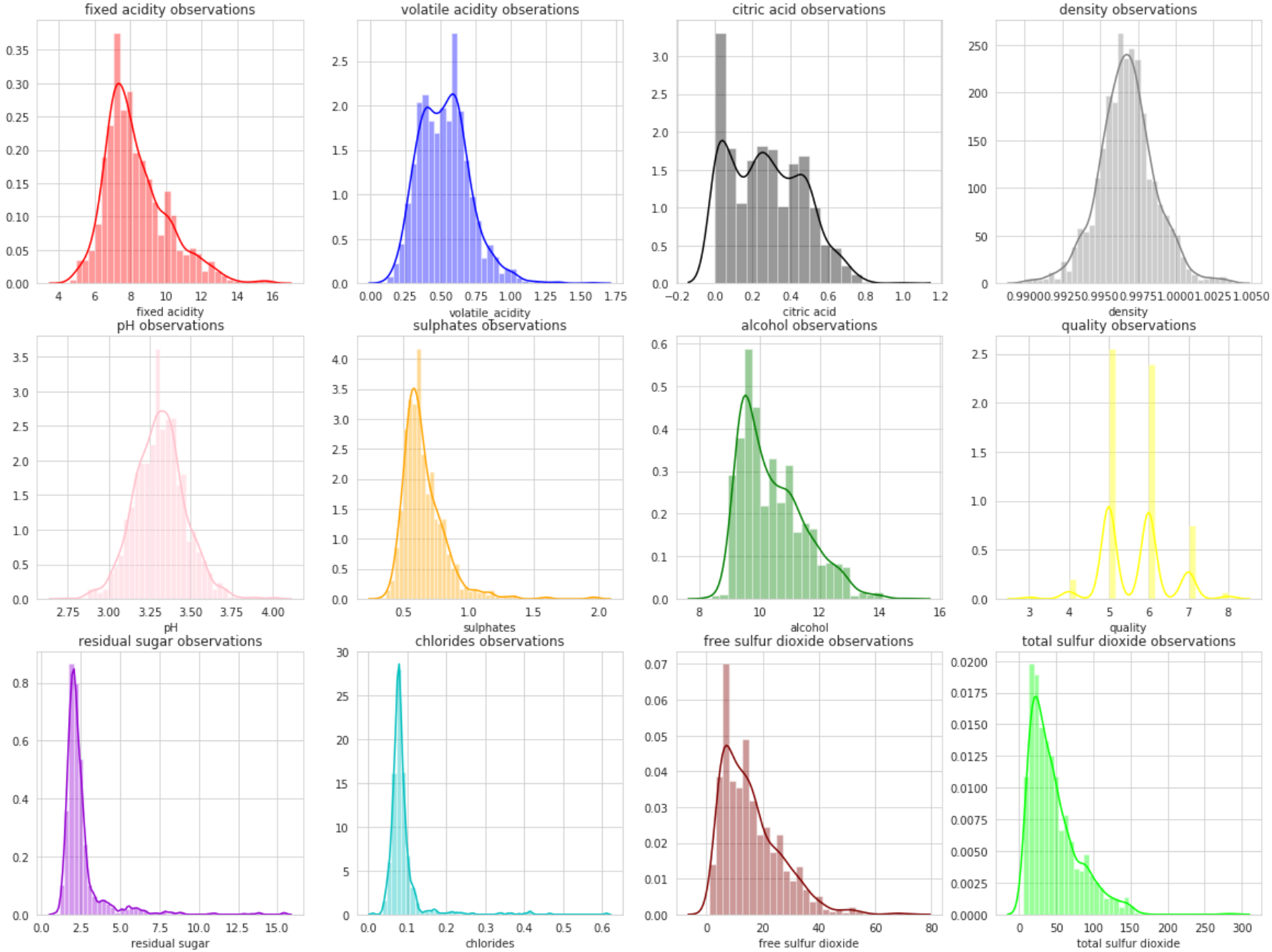
Create visualizations to find how residual sugar, density and alcohol effects the quality of wine

In [8]:
```
fig, axis = plt.subplots(ncols=3, figsize=(20,7))
sns.set_style("whitegrid")
sns.boxplot(x='qualitymark', y = 'alcohol', hue='qualitymark', data=data_with_mark, ax=axis[0])
sns.boxplot(x='qualitymark', y = 'residual sugar', hue='qualitymark', data=data_with_mark, ax=axis[1]).se
t_yscale('log')
sns.boxplot(x='qualitymark', y = 'density',hue='qualitymark', data=data_with_mark, ax=axis[2])
fig.savefig('binary_1.png')
```



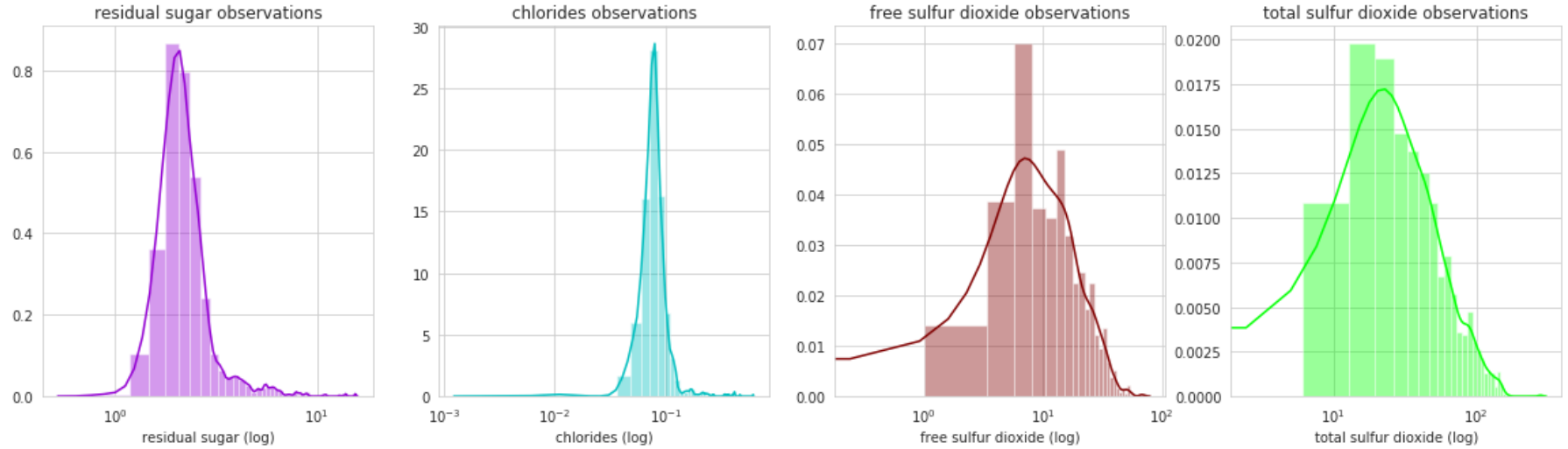
- Wine quality is having a positive relationship with alcohol
- There is no correlation between quality and residual sugar
- Density vs Quality gives a negative correlation

In [9]:
```
fig, axis = plt.subplots(3, 4, figsize=(20, 15))
sns.distplot(data['fixed acidity'].values, color='Red', ax=axis[0,0], axlabel='fixed acidity').set_title(
'fixed acidity observations')
sns.distplot(data['volatile acidity'].values, color='Blue', ax=axis[0,1], axlabel='volatile_acidity').set
_title('volatile acidity obserations')
sns.distplot(data['citric acid'].values, color='Black', ax=axis[0,2], axlabel='citric acid').set_title('c
itric acid observations')
sns.distplot(data['density'].values, color='Gray', ax=axis[0,3], axlabel='density').set_title('density ob
servations')
sns.distplot(data['pH'].values, color='Pink', ax=axis[1,0], axlabel='pH').set_title('pH observations')
sns.distplot(data['sulphates'].values, color='Orange', ax=axis[1,1], axlabel='sulphates').set_title('sulp
hates observations')
sns.distplot(data['alcohol'].values, color='Green', ax=axis[1,2], axlabel='alcohol').set_title('alcohol o
bservations')
sns.distplot(data['quality'].values, color='Yellow', ax=axis[1,3], axlabel='quality').set_title('quality
 observations')
sns.distplot(data['residual sugar'].values, color='darkviolet', ax=axis[2,0], axlabel='residual sugar').s
et_title('residual sugar observations')
sns.distplot(data['chlorides'].values, color='c', ax=axis[2,1], axlabel='chlorides').set_title('chlorides
observations')
sns.distplot(data['free sulfur dioxide'].values, color='maroon', ax=axis[2,2], axlabel='free sulfur dioxi
de').set_title('free sulfur dioxide observations')
sns.distplot(data['total sulfur dioxide'].values, color='lime', ax=axis[2,3], axlabel='total sulfur dioxi
de').set_title('total sulfur dioxide observations')
fig.savefig('univar_1.png')
```


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

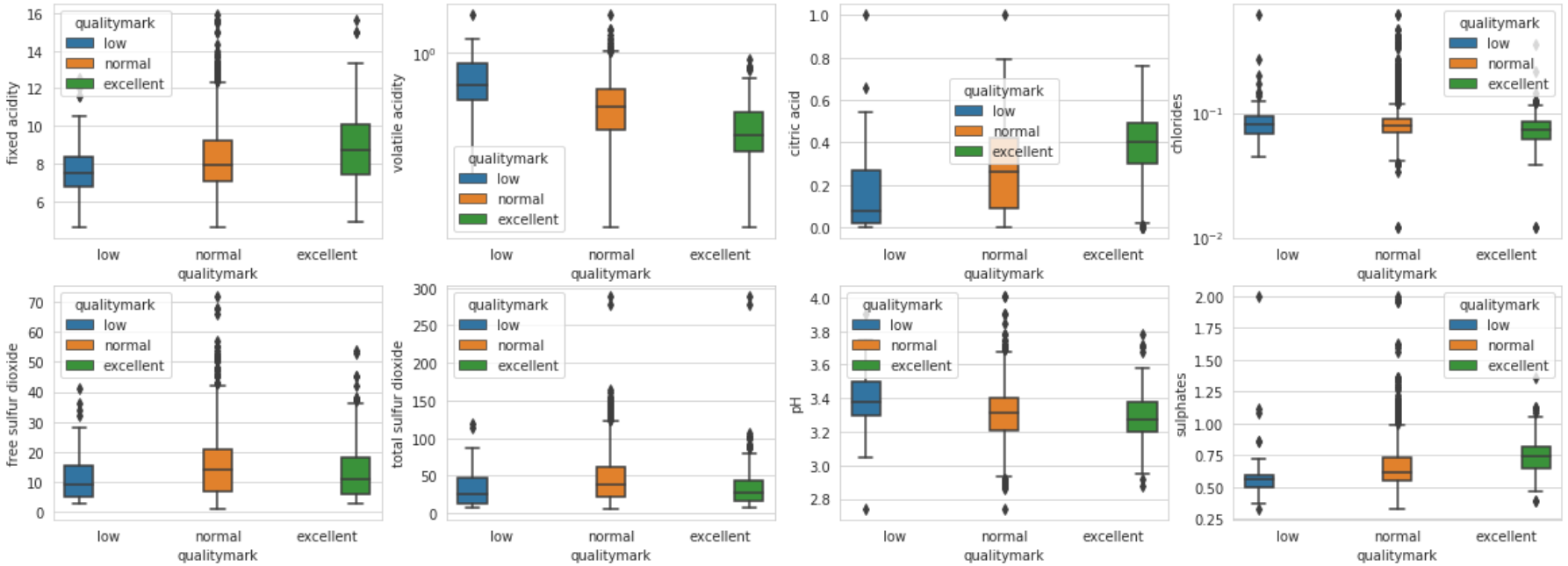
From above visualizations we can observe how the values distributed for every attribute. But residual sugar, chlorides, free sulfur dioxide and total sulfur dioxide plots are not clear so we will normailze by plottting on a log scale.

In [10]:
```
fig, axis = plt.subplots(1, 4, figsize=(20, 5))
ax1 = sns.distplot(data['residual sugar'].values, color='darkviolet', ax=axis[0], axlabel='residual sugar
(log)')
ax1.set_title('residual sugar observations')
ax1.set_xscale('log')
ax1 = sns.distplot(data['chlorides'].values, color='c', ax=axis[1], axlabel='chlorides (log)')
ax1.set_title('chlorides observations')
ax1.set_xscale('log')
ax1 = sns.distplot(data['free sulfur dioxide'].values, color='maroon', ax=axis[2], axlabel='free sulfur d
ioxide (log)')
ax1.set_title('free sulfur dioxide observations')
ax1.set_xscale('log')
ax1 = sns.distplot(data['total sulfur dioxide'].values, color='lime', ax=axis[3], axlabel='total sulfur d
ioxide (log)')
ax1.set_title('total sulfur dioxide observations')
ax1.set_xscale('log')
fig.savefig('univar_2.png')
```



Now we see how remaining variables correalted with the quality of wine

In [11]:
```
fig, axis = plt.subplots(2,4, figsize=(20,7))
sns.set_style("whitegrid")
sns.boxplot(x='qualitymark', y = 'fixed acidity', hue='qualitymark', data=data_with_mark, ax=axis[0,0])
sns.boxplot(x='qualitymark', y = 'volatile acidity', hue='qualitymark', data=data_with_mark, ax=axis[0,1
]).set_yscale('log')
sns.boxplot(x='qualitymark', y = 'citric acid',hue='qualitymark', data=data_with_mark, ax=axis[0,2])
sns.boxplot(x='qualitymark', y = 'chlorides',hue='qualitymark', data=data_with_mark, ax=axis[0,3]).set_ys
cale('log')
sns.boxplot(x='qualitymark', y = 'free sulfur dioxide',hue='qualitymark', data=data_with_mark, ax=axis[1,
0])
sns.boxplot(x='qualitymark', y = 'total sulfur dioxide',hue='qualitymark', data=data_with_mark, ax=axis[1
,1])
sns.boxplot(x='qualitymark', y = 'pH',hue='qualitymark', data=data_with_mark, ax=axis[1,2])
sns.boxplot(x='qualitymark', y = 'sulphates',hue='qualitymark', data=data_with_mark, ax=axis[1,3])
fig.savefig('binvar_2.png')
```



From above visulaizations

- Fixed acidity, Citric acid, Sulphates are having a positive relationship with the quality of wine
- Low volatile acidity, low pH and low chloride values are resulting in good quality of the wine.
- Lot of outliers for chlorides value for normal wine quality