In [5]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from bokeh.plotting import figure, output_file, show
from bokeh.layouts import row
from bokeh.io import output_notebook
import statsmodels.api as sm
import statsmodels.formula.api as smf
from patsy import dmatrices
import sklearn
import sklearn.metrics
from sklearn import ensemble
from sklearn import linear_model
import warnings
warnings.filterwarnings('ignore')
output_notebook()
%matplotlib inline
```

(https://bokeh.pydata.org) BokehJS 1.3.4 successfully loaded.

In [19]:

```
import os
os.chdir("/Users/anantkataria/Downloads")
```

In [22]:

```
url = "winequality-white.csv"
wine = pd.read_csv(url)
```

In [23]:

```
wine.head(n=5)
```

Out[23]:

	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
0	7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6
1	6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9...
2	8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;1...
3	7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4...
4	7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4...

In [24]:

```
wine = pd.read_csv(url, sep=";")
wine.head(n=5)
```

Out[24]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	al
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	

In [26]:

```
wine.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

In [27]:
```
wine.describe()
```

Out[27]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	tot
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440

In [28]:
```
wine.isnull().sum()
```

Out[28]:
```
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
```

In [29]:
```
wine.rename(columns={'fixed acidity': 'fixed_acidity','citric acid'
:'citric_acid','volatile acidity':'volatile_acidity','residual suga
r':'residual_sugar','free sulfur dioxide':'free_sulfur_dioxide','to
tal sulfur dioxide':'total_sulfur_dioxide'}, inplace=True)
wine.head(n=5)
```

Out[29]:

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	to
0	7.0	0.27	0.36	20.7	0.045	45.0	
1	6.3	0.30	0.34	1.6	0.049	14.0	
2	8.1	0.28	0.40	6.9	0.050	30.0	
3	7.2	0.23	0.32	8.5	0.058	47.0	
4	7.2	0.23	0.32	8.5	0.058	47.0	

In [30]:
```
wine['quality'].unique()
```

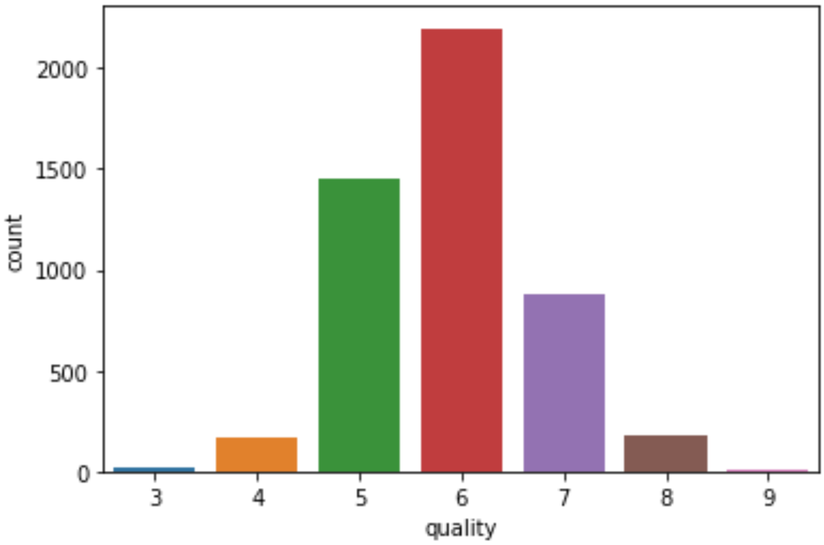
Out[30]:
```
array([6, 5, 7, 8, 4, 3, 9])
```

In [31]:
```
wine.quality.value_counts().sort_index()
```

Out[31]:
```
3      20
4     163
5    1457
6    2198
7     880
8     175
9       5
Name: quality, dtype: int64
```

In [32]:
```
sns.countplot(x='quality', data=wine)
```

Out[32]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x10fff42d0>
```



In [33]:
```
conditions = [
    (wine['quality'] >= 7),
    (wine['quality'] <= 4)
]
rating = ['good', 'bad']
wine['rating'] = np.select(conditions, rating, default='average')
wine.rating.value_counts()
```

Out[33]:
```
average    3655
good       1060
bad         183
Name: rating, dtype: int64
```

In [34]:

```
wine.groupby('rating').mean()
```

Out[34]:

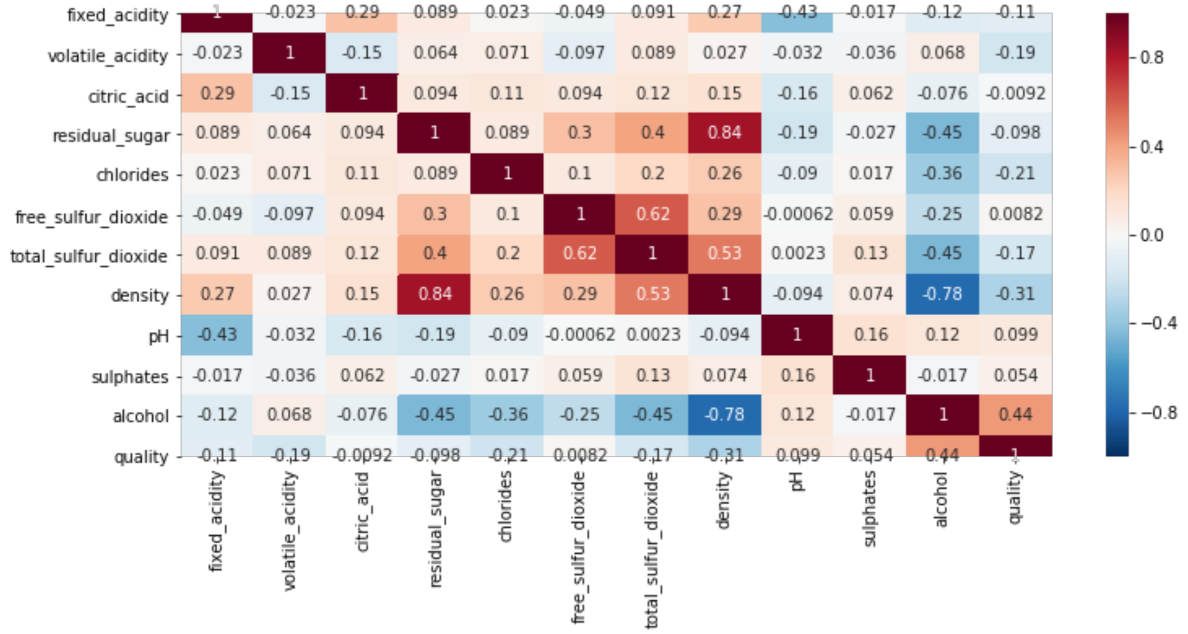
rating	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_diox
average	6.876060	0.277086	0.337877	6.797729	0.047740	35.962
bad	7.180874	0.375984	0.307705	4.821038	0.050557	26.633
good	6.725142	0.265349	0.326057	5.261509	0.038160	34.550

In [35]:

```
correlation = wine.corr()
plt.figure(figsize=(12, 5))
sns.heatmap(correlation, annot=True, linewidths=0, vmin=-1, cmap="R
dBu_r")
```

Out[35]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x126e3a2d0>
```



	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1	-0.023	0.29	0.089	0.023	-0.049	0.091	0.27	-0.43	-0.017	-0.12	-0.11
volatile_acidity	-0.023	1	-0.15	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068	-0.19
citric_acid	0.29	-0.15	1	0.094	0.11	0.094	0.12	0.15	-0.16	0.062	-0.076	-0.0092
residual_sugar	0.089	0.064	0.094	1	0.089	0.3	0.4	0.84	-0.19	-0.027	-0.45	-0.098
chlorides	0.023	0.071	0.11	0.089	1	0.1	0.2	0.26	-0.09	0.017	-0.36	-0.21
free_sulfur_dioxide	-0.049	-0.097	0.094	0.3	0.1	1	0.62	0.29	-0.00062	0.059	-0.25	0.0082
total_sulfur_dioxide	0.091	0.089	0.12	0.4	0.2	0.62	1	0.53	0.0023	0.13	-0.45	-0.17
density	0.27	0.027	0.15	0.84	0.26	0.29	0.53	1	-0.094	0.074	-0.78	-0.31
pH	-0.43	-0.032	-0.16	-0.19	-0.09	-0.00062	0.0023	-0.094	1	0.16	0.12	0.099
sulphates	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.13	0.074	0.16	1	-0.017	0.054
alcohol	-0.12	0.068	-0.076	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.017	1	0.44
quality	-0.11	-0.19	-0.0092	-0.098	-0.21	0.0082	-0.17	-0.31	0.099	0.054	0.44	1

In [36]:

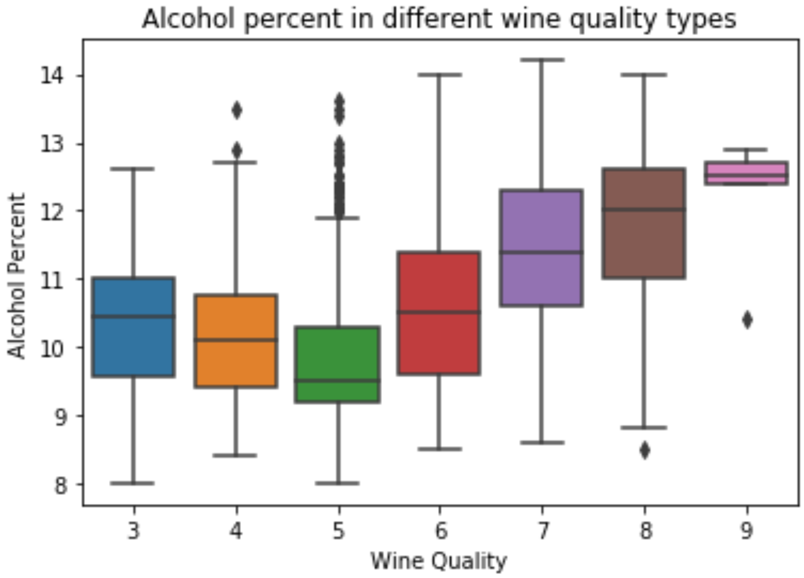
```
correlation['quality'].sort_values(ascending=False)
```

Out[36]:

```
quality                 1.000000
alcohol                 0.435575
pH                      0.099427
sulphates               0.053678
free_sulfur_dioxide     0.008158
citric_acid            -0.009209
residual_sugar         -0.097577
fixed_acidity          -0.113663
total_sulfur_dioxide   -0.174737
volatile_acidity       -0.194723
chlorides              -0.209934
density                -0.307123
Name: quality, dtype: float64
```

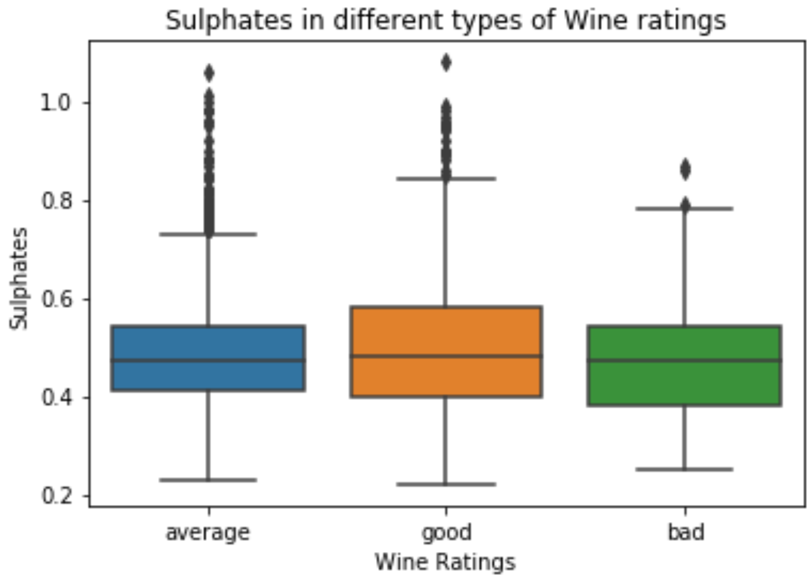
```
In [37]: bx = sns.boxplot(x="quality", y='alcohol', data = wine)
         bx.set(xlabel='Wine Quality', ylabel='Alcohol Percent', title='Alcohol percent in different wine quality types')
```

```
Out[37]: [Text(0, 0.5, 'Alcohol Percent'),
          Text(0.5, 0, 'Wine Quality'),
          Text(0.5, 1.0, 'Alcohol percent in different wine quality types')
         ]
```



```
In [38]: bx = sns.boxplot(x="rating", y='sulphates', data = wine)
         bx.set(xlabel='Wine Ratings', ylabel='Sulphates', title='Sulphates in different types of Wine ratings')
```

```
Out[38]: [Text(0, 0.5, 'Sulphates'),
          Text(0.5, 0, 'Wine Ratings'),
          Text(0.5, 1.0, 'Sulphates in different types of Wine ratings')]
```

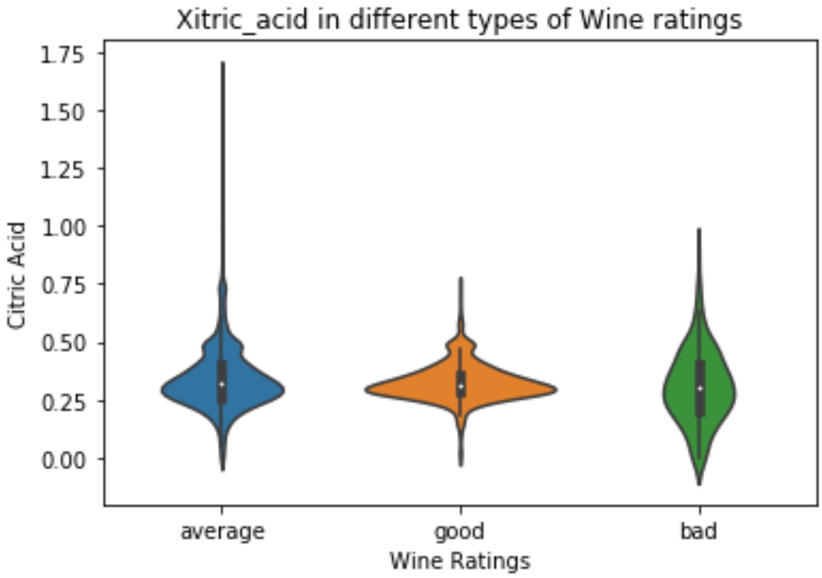


In [39]:

```
bx = sns.violinplot(x="rating", y='citric_acid', data = wine)
bx.set(xlabel='Wine Ratings', ylabel='Citric Acid', title='Xitric_a
cid in different types of Wine ratings')
```

Out[39]:

```
[Text(0, 0.5, 'Citric Acid'),
 Text(0.5, 0, 'Wine Ratings'),
 Text(0.5, 1.0, 'Xitric_acid in different types of Wine ratings')]
```

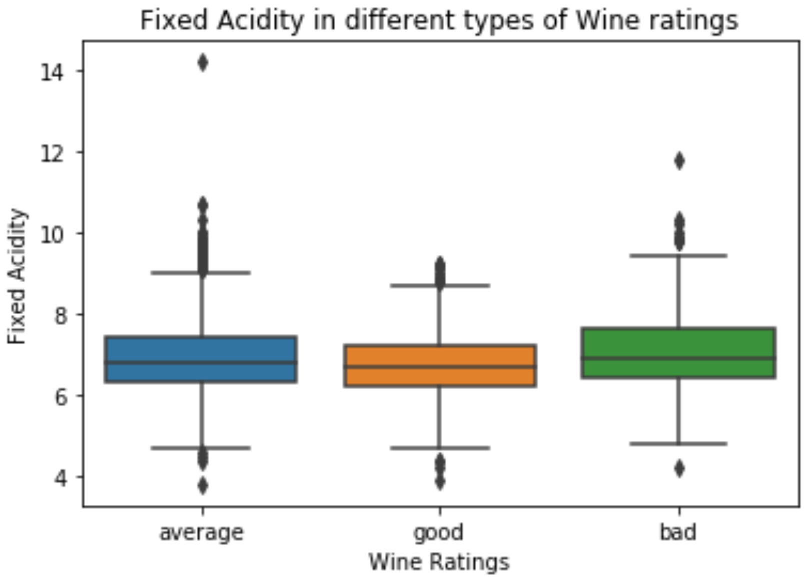


In [40]:

```
bx = sns.boxplot(x="rating", y='fixed_acidity', data = wine)
bx.set(xlabel='Wine Ratings', ylabel='Fixed Acidity', title='Fixed
Acidity in different types of Wine ratings')
```

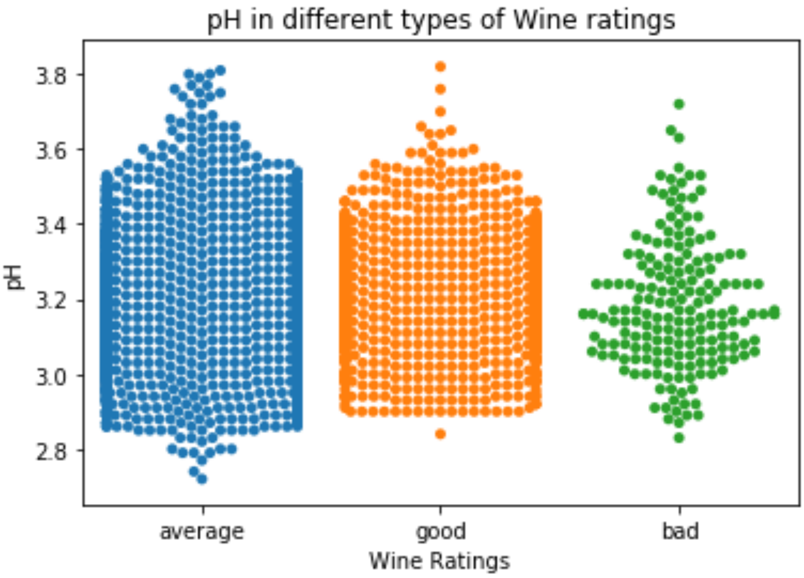
Out[40]:

```
[Text(0, 0.5, 'Fixed Acidity'),
 Text(0.5, 0, 'Wine Ratings'),
 Text(0.5, 1.0, 'Fixed Acidity in different types of Wine ratings'
)]
```



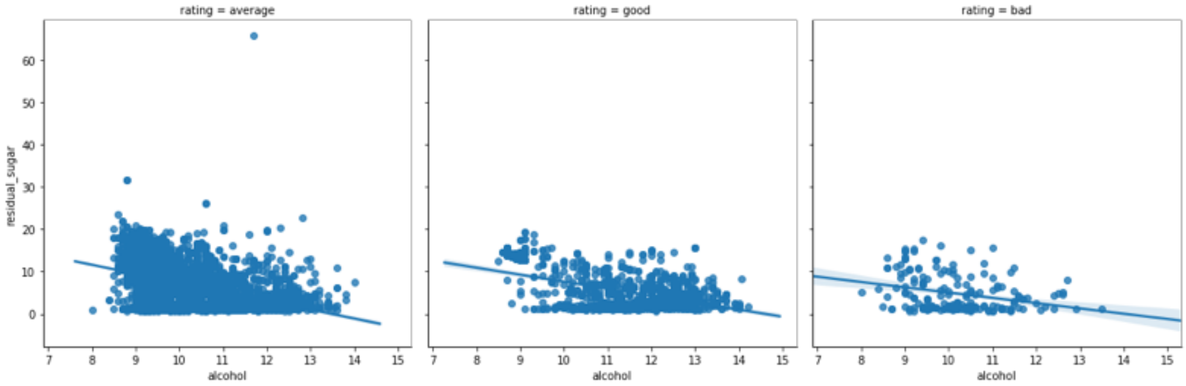
In [42]:
```
bx = sns.swarmplot(x="rating", y="pH", data = wine);
bx.set(xlabel='Wine Ratings', ylabel='pH', title='pH in different t
ypes of Wine ratings')
```

Out[42]:
```
[Text(0, 0.5, 'pH'),
 Text(0.5, 0, 'Wine Ratings'),
 Text(0.5, 1.0, 'pH in different types of Wine ratings')]
```



In [44]:
```
sns.lmplot(x = "alcohol", y = "residual_sugar", col = "rating", dat
a = wine)
```

Out[44]:
```
<seaborn.axisgrid.FacetGrid at 0x12c090510>
```



In [ ]:

In [ ]: