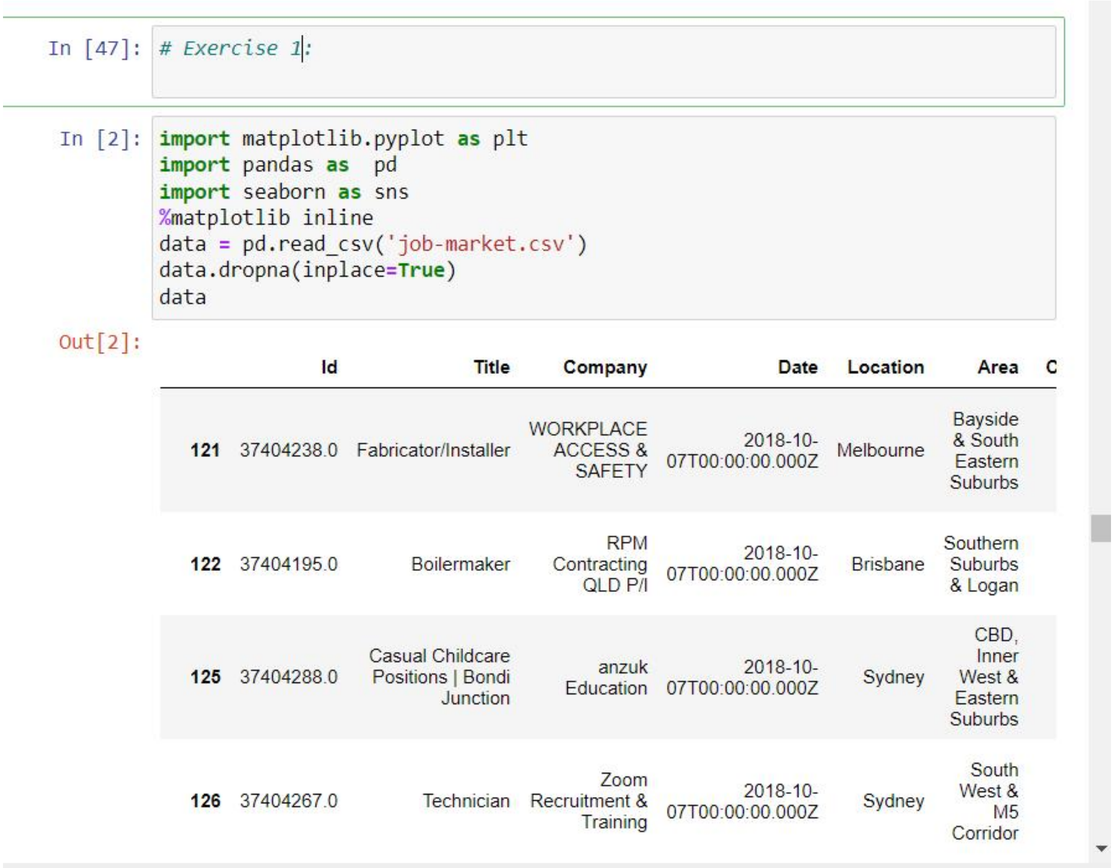
Link Github: https://github.com/paulbboone/DataMining_ThucHanh

HOMEWORK LAB 03

Exercise 1:

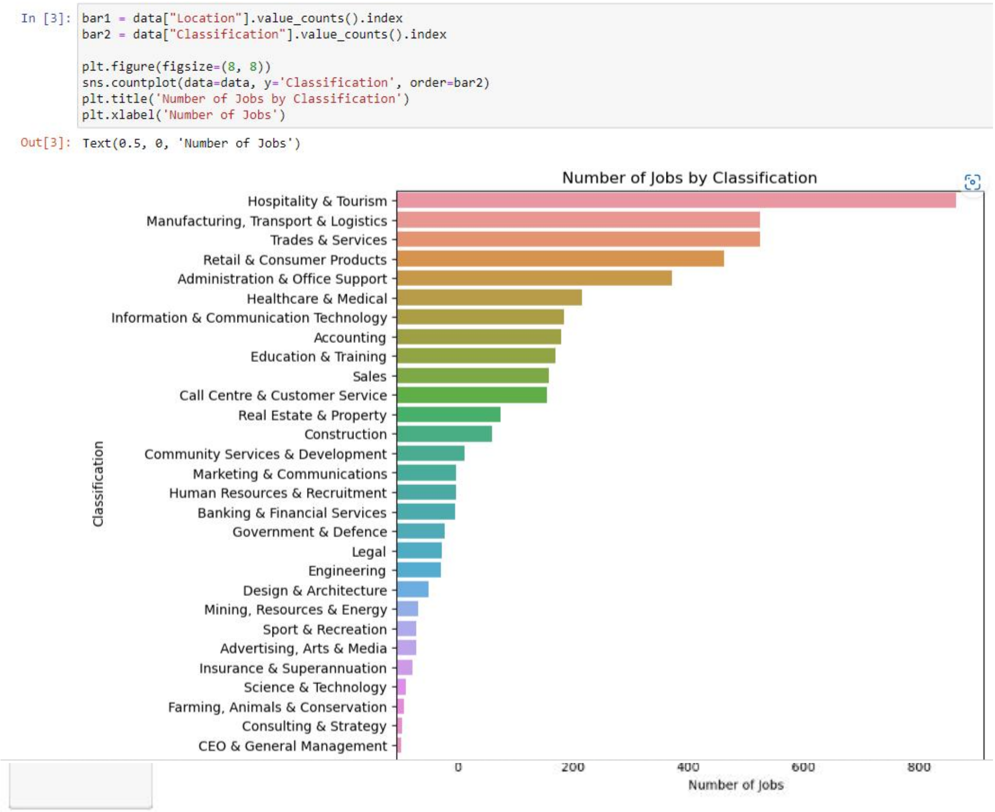


```
In [47]: # Exercise 1:
```

```
In [2]: import matplotlib.pyplot as plt
        import pandas as  pd
        import seaborn as sns
        %matplotlib inline
        data = pd.read_csv('job-market.csv')
        data.dropna(inplace=True)
        data
```

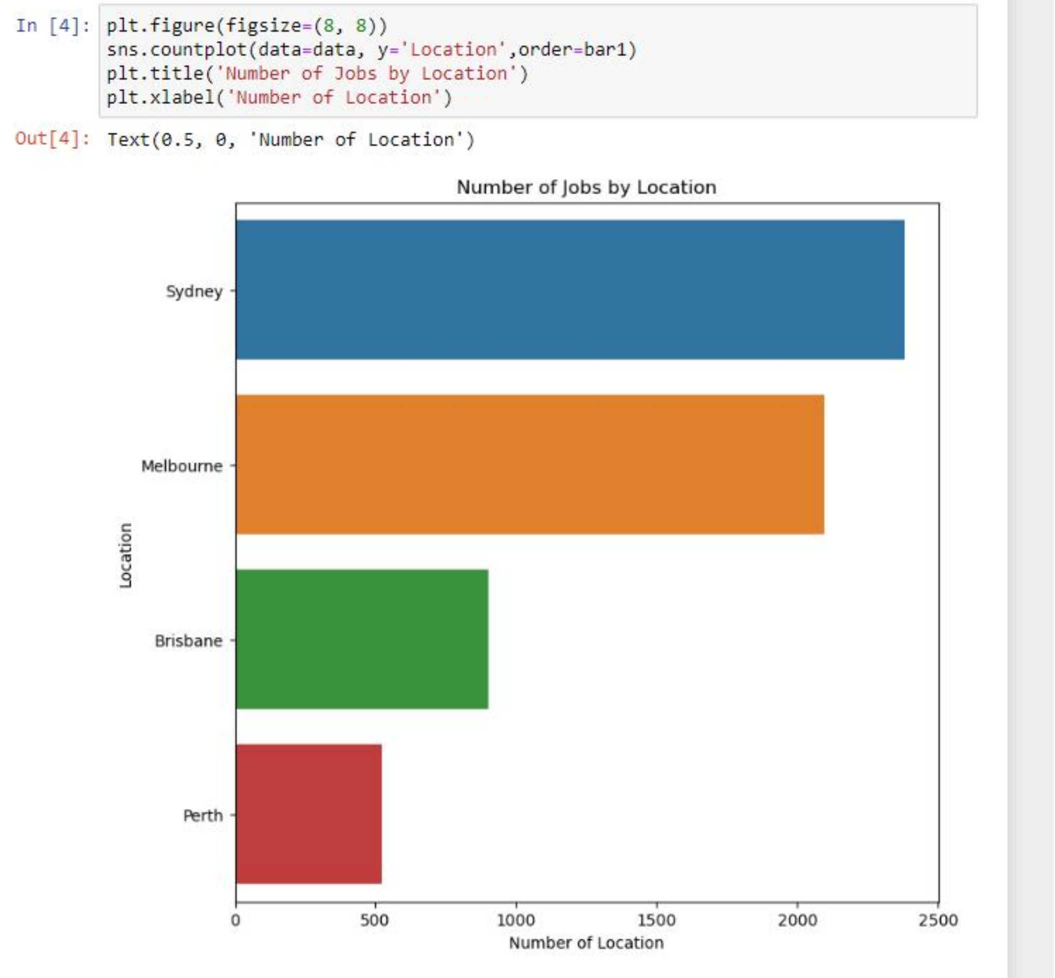
Out[2]:

	Id	Title	Company	Date	Location	Area	C
121	37404238.0	Fabricator/Installer	WORKPLACE ACCESS & SAFETY	2018-10-07T00:00:00.000Z	Melbourne	Bayside & South Eastern Suburbs	
122	37404195.0	Boilermaker	RPM Contracting QLD P/I	2018-10-07T00:00:00.000Z	Brisbane	Southern Suburbs & Logan	
125	37404288.0	Casual Childcare Positions | Bondi Junction	anzuk Education	2018-10-07T00:00:00.000Z	Sydney	CBD, Inner West & Eastern Suburbs	
126	37404267.0	Technician	Zoom Recruitment & Training	2018-10-07T00:00:00.000Z	Sydney	South West & M5 Corridor	

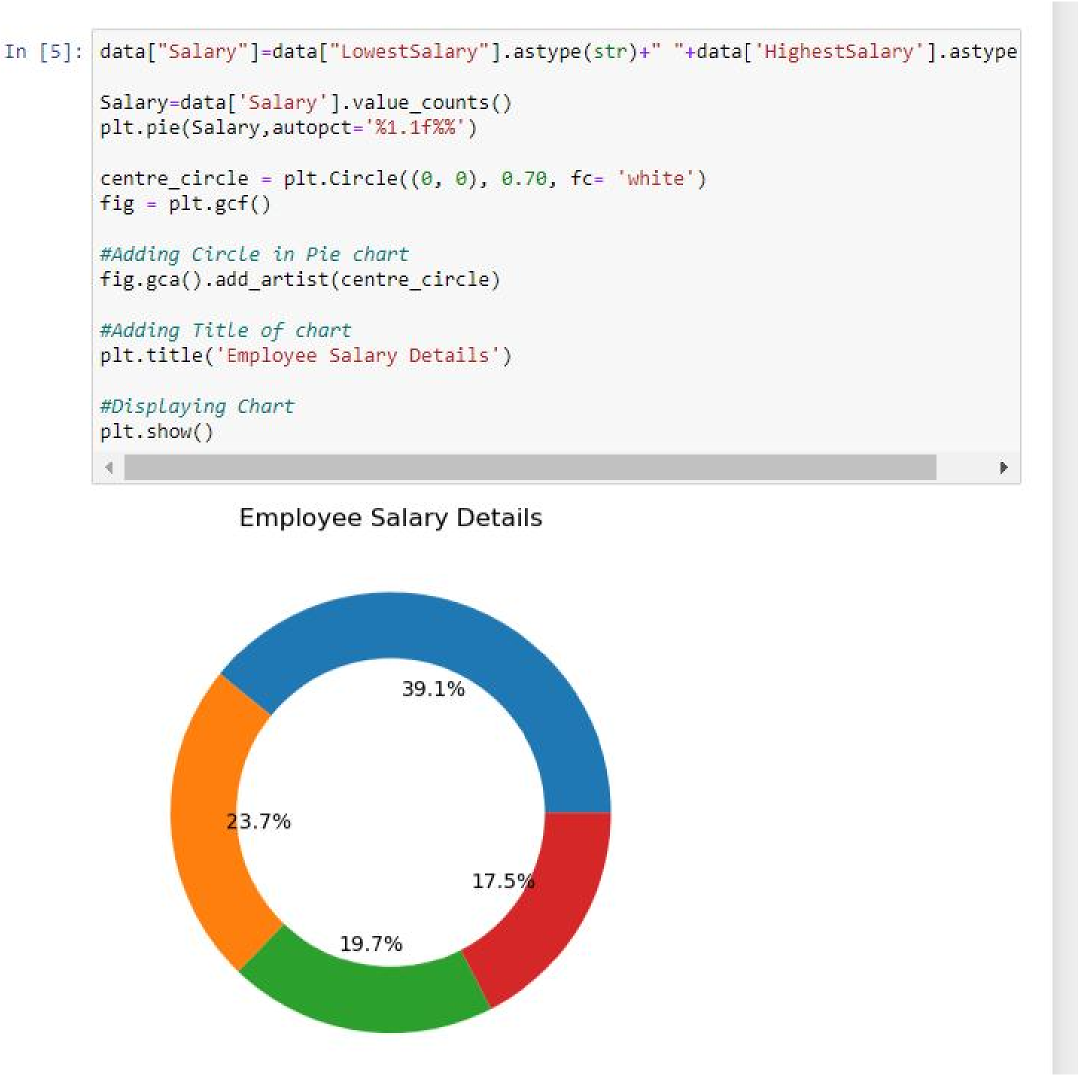


```
In [3]: bar1 = data["Location"].value_counts().index
        bar2 = data["Classification"].value_counts().index

        plt.figure(figsize=(8, 8))
        sns.countplot(data=data, y='Classification', order=bar2)
        plt.title('Number of Jobs by Classification')
        plt.xlabel('Number of Jobs')

Out[3]: Text(0.5, 0, 'Number of Jobs')
```

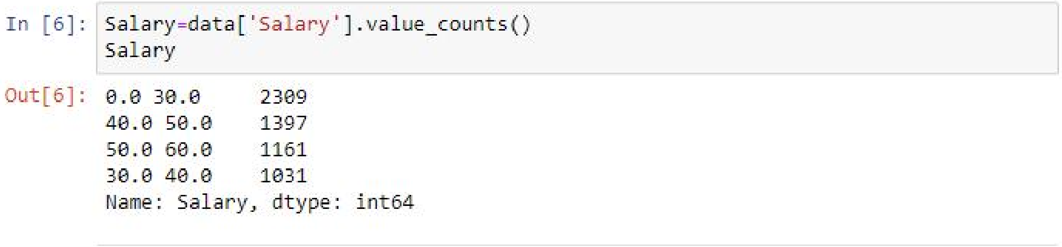
Number of Jobs by Classification
Hospitality & Tourism
Manufacturing, Transport & Logistics
Trades & Services
Retail & Consumer Products
Administration & Office Support
Healthcare & Medical
Information & Communication Technology
Accounting
Education & Training
Sales
Call Centre & Customer Service
Real Estate & Property
Construction
Community Services & Development
Marketing & Communications
Human Resources & Recruitment
Banking & Financial Services
Government & Defence
Legal
Engineering
Design & Architecture
Mining, Resources & Energy
Sport & Recreation
Advertising, Arts & Media
Insurance & Superannuation
Science & Technology
Farming, Animals & Conservation
Consulting & Strategy
CEO & General Management
Classification
0
200
400
600
800
Number of Jobs



```
In [4]: plt.figure(figsize=(8, 8))
        sns.countplot(data=data, y='Location',order=bar1)
        plt.title('Number of Jobs by Location')
        plt.xlabel('Number of Location')

Out[4]: Text(0.5, 0, 'Number of Location')
```

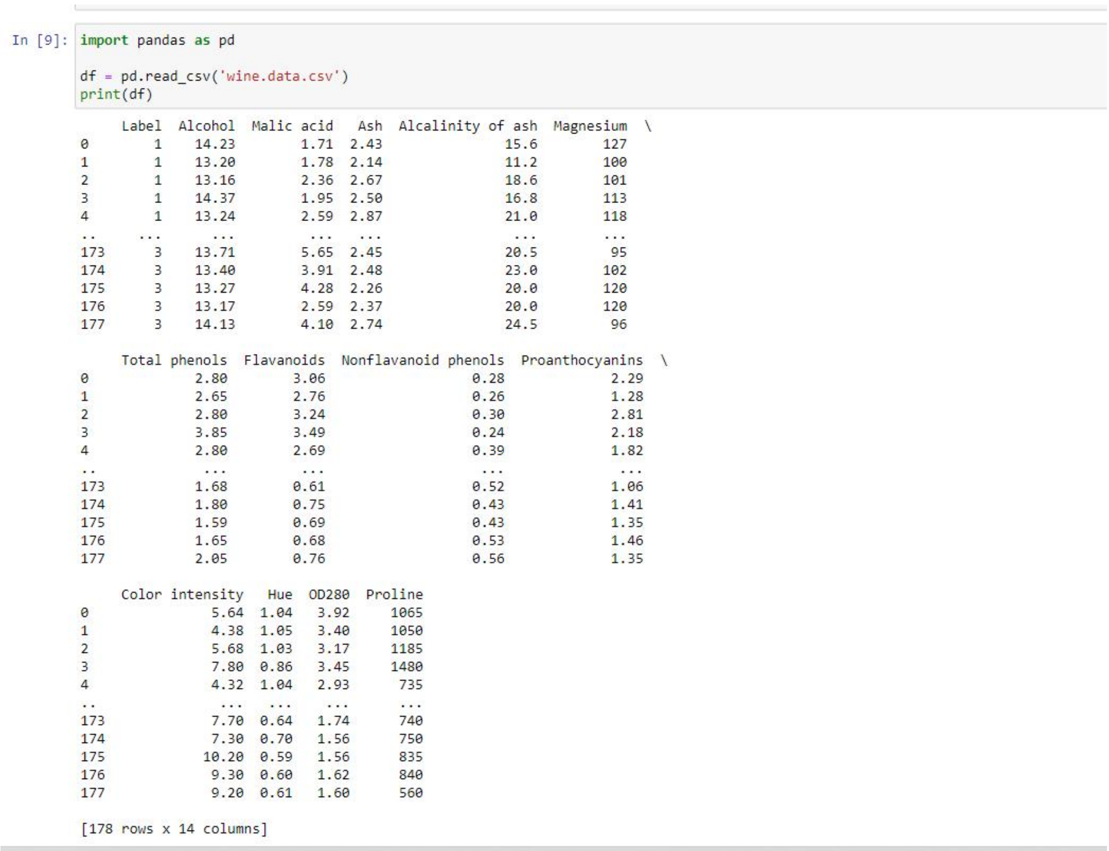


```
In [5]: data["Salary"]=data["LowestSalary"].astype(str)+" "+data['HighestSalary'].astype

        Salary=data['Salary'].value_counts()
        plt.pie(Salary,autopct='%1.1f%%')

        centre_circle = plt.Circle((0, 0), 0.70, fc= 'white')
        fig = plt.gcf()

        #Adding Circle in Pie chart
        fig.gca().add_artist(centre_circle)

        #Adding Title of chart
        plt.title('Employee Salary Details')

        #Displaying Chart
        plt.show()
```



```
In [6]: Salary=data['Salary'].value_counts()
        Salary

Out[6]: 0.0 30.0     2309
        40.0 50.0    1397
        50.0 60.0    1161
        30.0 40.0    1031
        Name: Salary, dtype: int64
```

Exercise 2:

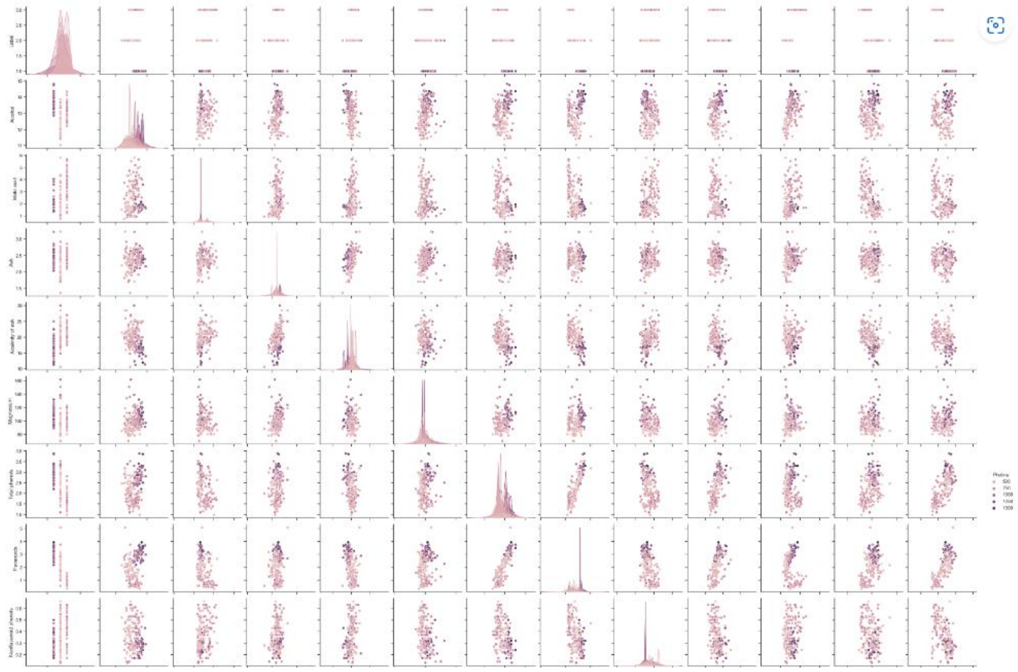


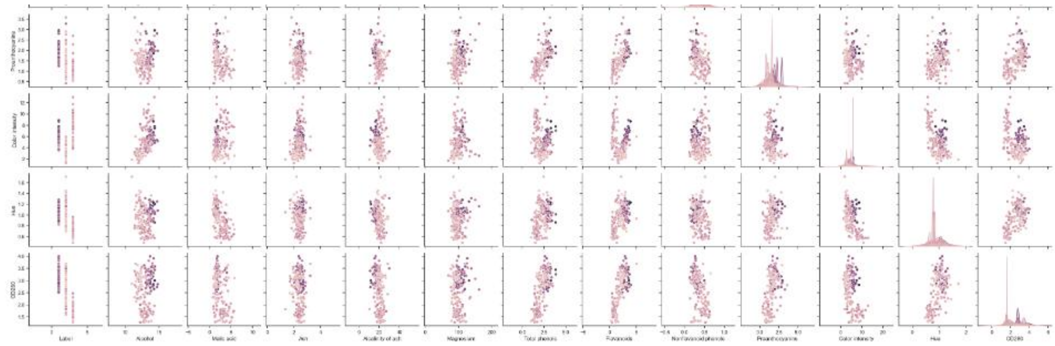
```
In [9]: import pandas as pd

        df = pd.read_csv('wine.data.csv')
        print(df)
```

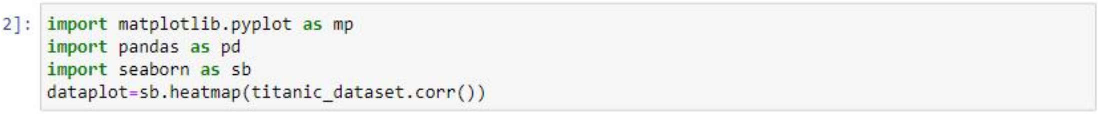
```
     Label  Alcohol  Malic acid   Ash  Alcalinity of ash  Magnesium  \
0        1    14.23        1.71  2.43               15.6        127   
1        1    13.20        1.78  2.14               11.2        100   
2        1    13.16        2.36  2.67               18.6        101   
3        1    14.37        1.95  2.50               16.8        113   
4        1    13.24        2.59  2.87               21.0        118   
..     ...      ...         ...   ...                ...        ...   
173      3    13.71        5.65  2.45               20.5         95   
174      3    13.40        3.91  2.48               23.0        102   
175      3    13.27        4.28  2.26               20.0        120   
176      3    13.17        2.59  2.37               20.0        120   
177      3    14.13        4.10  2.74               24.5         96   

     Total phenols  Flavanoids  Nonflavanoid phenols  Proanthocyanins  \
0             2.80        3.06                  0.28             2.29   
1             2.65        2.76                  0.26             1.28   
2             2.80        3.24                  0.30             2.81   
3             3.85        3.49                  0.24             2.18   
4             2.80        2.69                  0.39             1.82   
..             ...         ...                   ...              ...   
173           1.68        0.61                  0.52             1.06   
174           1.80        0.75                  0.43             1.41   
175           1.59        0.69                  0.43             1.35   
176           1.65        0.68                  0.53             1.46   
177           2.05        0.76                  0.56             1.35   

     Color intensity   Hue  OD280  Proline  
0               5.64  1.04   3.92     1065  
1               4.38  1.05   3.40     1050  
2               5.68  1.03   3.17     1185  
3               7.80  0.86   3.45     1480  
4               4.32  1.04   2.93      735  
..               ...   ...    ...      ...  
173             7.70  0.64   1.74      740  
174             7.30  0.70   1.56      750  
175            10.20  0.59   1.56      835  
176             9.30  0.60   1.62      840  
177             9.20  0.61   1.60      560  

[178 rows x 14 columns]
```

```
In [10]: df = df.drop('Label', axis=1)
```

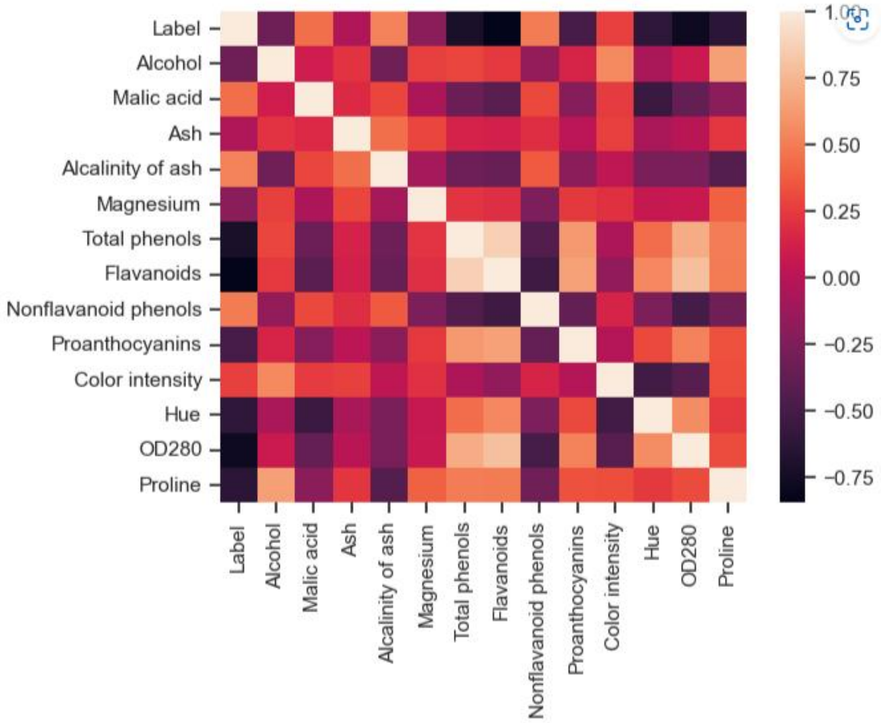
```
In [11]: titanic_dataset = pd.read_csv('wine.data.csv')
         sns.set_theme(style="ticks")
         sns.pairplot(titanic_dataset, hue='Proline')

Out[11]: <seaborn.axisgrid.PairGrid at 0x1c100c94dc0>
```







```
2]: import matplotlib.pyplot as mp
    import pandas as pd
    import seaborn as sb
    dataplot=sb.heatmap(titanic_dataset.corr())
```


Label
Alcohol
Malic acid
Ash
Alcalinity of ash
Magnesium
Total phenols
Flavanoids
Nonflavanoid phenols
Proanthocyanins
Color intensity
Hue
OD280
Proline