# Data Mining LAB : Experiment 2

## Submitted By:

Name: Debatreya Das

Roll No. 12212070

CS A4

Data Mining LAB

### Problem Statement

Download any two datasets of your choice comprising different types of attributes. Compute and analyze central tendency (mean, median and mode), dispersion (range, quartiles, interquartile range, variance and standard deviation) of different attributes and covariance and correlation matrix for the given datasets. Discuss your observations regarding which operation is logically apt for given attribute type and regarding characteristics of datasets which can be observed based on covariance and correlation matrix.

The objective is to understand the central tendencies and dispersion of the data attributes.

### Data Set

* **WINE**: [Wine Dataset](http://../wine/wine.data) (Link won’t work here)
* **IRIS**: [Iris Dataset](http://../iris/iris.data) (Link wont work here)

### Code and Output

* [**WINE Notebook**](#_wgvx4f7fijz0)
* [**IRIS Notebook**](#_3dnhewwdml0i)

### Observations

No such observations (The objective is to understand the central tendencies and dispersions of the data attributes.)

## Wine Notebook

# Statistical Analysis: Wine Dataset

We will calculate the following statistics for each attribute

* [] Central Tendancy
  + Mean
  + Median
  + Mode
* [] Dispersion
  + Range
  + Quartiles
  + InterQuartile Range
  + Variance
  + Santard Deviation
* [] Covariance Matrix
* [] Correlation Matrix

## Importing necessary Libraries

**import** pandas **as** pd  
**import** numpy **as** np  
**import** seaborn **as** sns  
**import** matplotlib.pyplot **as** plt

## Load datasets

wine\_df = pd.read\_csv('../wine/wine.data', header=None)  
wine\_df.columns = [  
 "Class", "Alcohol", "Malic\_acid", "Ash", "Alcalinity\_of\_ash",   
 "Magnesium", "Total\_phenols", "Flavanoids", "Nonflavanoid\_phenols",  
 "Proanthocyanins", "Color\_intensity", "Hue", "OD280/OD315\_of\_diluted\_wines", "Proline"  
]  
wine\_df

Class 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/OD315\_of\_diluted\_wines 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]

## Compute central tendency (mean, median, mode)

mean\_values = wine\_df.mean()  
median\_values = wine\_df.median()  
mode\_values = wine\_df.mode().iloc[0] *# mode() returns a dataframe, so we take the first row*  
  
*# Combine them into a single DataFrame*  
central\_tendency\_table = pd.DataFrame({  
 'Mean': mean\_values,  
 'Median': median\_values,  
 'Mode': mode\_values  
})  
  
*# Display the table*  
central\_tendency\_table

Mean Median Mode  
Class 1.938202 2.000 2.00  
Alcohol 13.000618 13.050 12.37  
Malic\_acid 2.336348 1.865 1.73  
Ash 2.366517 2.360 2.28  
Alcalinity\_of\_ash 19.494944 19.500 20.00  
Magnesium 99.741573 98.000 88.00  
Total\_phenols 2.295112 2.355 2.20  
Flavanoids 2.029270 2.135 2.65  
Nonflavanoid\_phenols 0.361854 0.340 0.26  
Proanthocyanins 1.590899 1.555 1.35  
Color\_intensity 5.058090 4.690 2.60  
Hue 0.957449 0.965 1.04  
OD280/OD315\_of\_diluted\_wines 2.611685 2.780 2.87  
Proline 746.893258 673.500 520.00

## Measures of Dispersion (range, quartiles, InterQuartile Ranges, Variance, Standard Deviations)

wine\_dispersion = wine\_df.describe().T  
wine\_dispersion['IQR'] = wine\_dispersion['75%'] - wine\_dispersion['25%']  
wine\_dispersion['Variance'] = wine\_dispersion['std'] \*\* 2  
  
wine\_dispersion

count mean std min 25% \  
Class 178.0 1.938202 0.775035 1.00 1.0000   
Alcohol 178.0 13.000618 0.811827 11.03 12.3625   
Malic\_acid 178.0 2.336348 1.117146 0.74 1.6025   
Ash 178.0 2.366517 0.274344 1.36 2.2100   
Alcalinity\_of\_ash 178.0 19.494944 3.339564 10.60 17.2000   
Magnesium 178.0 99.741573 14.282484 70.00 88.0000   
Total\_phenols 178.0 2.295112 0.625851 0.98 1.7425   
Flavanoids 178.0 2.029270 0.998859 0.34 1.2050   
Nonflavanoid\_phenols 178.0 0.361854 0.124453 0.13 0.2700   
Proanthocyanins 178.0 1.590899 0.572359 0.41 1.2500   
Color\_intensity 178.0 5.058090 2.318286 1.28 3.2200   
Hue 178.0 0.957449 0.228572 0.48 0.7825   
OD280/OD315\_of\_diluted\_wines 178.0 2.611685 0.709990 1.27 1.9375   
Proline 178.0 746.893258 314.907474 278.00 500.5000   
  
 50% 75% max IQR \  
Class 2.000 3.0000 3.00 2.0000   
Alcohol 13.050 13.6775 14.83 1.3150   
Malic\_acid 1.865 3.0825 5.80 1.4800   
Ash 2.360 2.5575 3.23 0.3475   
Alcalinity\_of\_ash 19.500 21.5000 30.00 4.3000   
Magnesium 98.000 107.0000 162.00 19.0000   
Total\_phenols 2.355 2.8000 3.88 1.0575   
Flavanoids 2.135 2.8750 5.08 1.6700   
Nonflavanoid\_phenols 0.340 0.4375 0.66 0.1675   
Proanthocyanins 1.555 1.9500 3.58 0.7000   
Color\_intensity 4.690 6.2000 13.00 2.9800   
Hue 0.965 1.1200 1.71 0.3375   
OD280/OD315\_of\_diluted\_wines 2.780 3.1700 4.00 1.2325   
Proline 673.500 985.0000 1680.00 484.5000   
  
 Variance   
Class 0.600679   
Alcohol 0.659062   
Malic\_acid 1.248015   
Ash 0.075265   
Alcalinity\_of\_ash 11.152686   
Magnesium 203.989335   
Total\_phenols 0.391690   
Flavanoids 0.997719   
Nonflavanoid\_phenols 0.015489   
Proanthocyanins 0.327595   
Color\_intensity 5.374449   
Hue 0.052245   
OD280/OD315\_of\_diluted\_wines 0.504086   
Proline 99166.717355

# Compute covariance and correlation matrices

wine\_covariance = wine\_df.cov()  
wine\_correlation = wine\_df.corr()

### WINE Covariance Matrix

wine\_covariance

Class Alcohol Malic\_acid Ash \  
Class 0.600679 -0.206515 0.379039 -0.010555   
Alcohol -0.206515 0.659062 0.085611 0.047115   
Malic\_acid 0.379039 0.085611 1.248015 0.050277   
Ash -0.010555 0.047115 0.050277 0.075265   
Alcalinity\_of\_ash 1.340364 -0.841093 1.076332 0.406208   
Magnesium -2.315495 3.139878 -0.870780 1.122937   
Total\_phenols -0.348835 0.146887 -0.234338 0.022146   
Flavanoids -0.656091 0.192033 -0.458630 0.031535   
Nonflavanoid\_phenols 0.047177 -0.015754 0.040733 0.006358   
Proanthocyanins -0.221413 0.063518 -0.141147 0.001516   
Color\_intensity 0.477339 1.028283 0.644838 0.164654   
Hue -0.109368 -0.013313 -0.143326 -0.004682   
OD280/OD315\_of\_diluted\_wines -0.433737 0.041698 -0.292447 0.000762   
Proline -154.667651 164.567185 -67.548867 19.319739   
  
 Alcalinity\_of\_ash Magnesium Total\_phenols \  
Class 1.340364 -2.315495 -0.348835   
Alcohol -0.841093 3.139878 0.146887   
Malic\_acid 1.076332 -0.870780 -0.234338   
Ash 0.406208 1.122937 0.022146   
Alcalinity\_of\_ash 11.152686 -3.974760 -0.671149   
Magnesium -3.974760 203.989335 1.916470   
Total\_phenols -0.671149 1.916470 0.391690   
Flavanoids -1.172083 2.793087 0.540470   
Nonflavanoid\_phenols 0.150422 -0.455563 -0.035045   
Proanthocyanins -0.377176 1.932832 0.219373   
Color\_intensity 0.145024 6.620521 -0.079998   
Hue -0.209118 0.180851 0.062039   
OD280/OD315\_of\_diluted\_wines -0.656234 0.669308 0.311021   
Proline -463.355345 1769.158700 98.171057   
  
 Flavanoids Nonflavanoid\_phenols \  
Class -0.656091 0.047177   
Alcohol 0.192033 -0.015754   
Malic\_acid -0.458630 0.040733   
Ash 0.031535 0.006358   
Alcalinity\_of\_ash -1.172083 0.150422   
Magnesium 2.793087 -0.455563   
Total\_phenols 0.540470 -0.035045   
Flavanoids 0.997719 -0.066867   
Nonflavanoid\_phenols -0.066867 0.015489   
Proanthocyanins 0.373148 -0.026060   
Color\_intensity -0.399169 0.040121   
Hue 0.124082 -0.007471   
OD280/OD315\_of\_diluted\_wines 0.558262 -0.044469   
Proline 155.447492 -12.203586   
  
 Proanthocyanins Color\_intensity Hue \  
Class -0.221413 0.477339 -0.109368   
Alcohol 0.063518 1.028283 -0.013313   
Malic\_acid -0.141147 0.644838 -0.143326   
Ash 0.001516 0.164654 -0.004682   
Alcalinity\_of\_ash -0.377176 0.145024 -0.209118   
Magnesium 1.932832 6.620521 0.180851   
Total\_phenols 0.219373 -0.079998 0.062039   
Flavanoids 0.373148 -0.399169 0.124082   
Nonflavanoid\_phenols -0.026060 0.040121 -0.007471   
Proanthocyanins 0.327595 -0.033504 0.038665   
Color\_intensity -0.033504 5.374449 -0.276506   
Hue 0.038665 -0.276506 0.052245   
OD280/OD315\_of\_diluted\_wines 0.210933 -0.705813 0.091766   
Proline 59.554334 230.767480 17.000223   
  
 OD280/OD315\_of\_diluted\_wines Proline   
Class -0.433737 -154.667651   
Alcohol 0.041698 164.567185   
Malic\_acid -0.292447 -67.548867   
Ash 0.000762 19.319739   
Alcalinity\_of\_ash -0.656234 -463.355345   
Magnesium 0.669308 1769.158700   
Total\_phenols 0.311021 98.171057   
Flavanoids 0.558262 155.447492   
Nonflavanoid\_phenols -0.044469 -12.203586   
Proanthocyanins 0.210933 59.554334   
Color\_intensity -0.705813 230.767480   
Hue 0.091766 17.000223   
OD280/OD315\_of\_diluted\_wines 0.504086 69.927526   
Proline 69.927526 99166.717355

### WINE Correlation Matrix

wine\_correlation

Class Alcohol Malic\_acid Ash \  
Class 1.000000 -0.328222 0.437776 -0.049643   
Alcohol -0.328222 1.000000 0.094397 0.211545   
Malic\_acid 0.437776 0.094397 1.000000 0.164045   
Ash -0.049643 0.211545 0.164045 1.000000   
Alcalinity\_of\_ash 0.517859 -0.310235 0.288500 0.443367   
Magnesium -0.209179 0.270798 -0.054575 0.286587   
Total\_phenols -0.719163 0.289101 -0.335167 0.128980   
Flavanoids -0.847498 0.236815 -0.411007 0.115077   
Nonflavanoid\_phenols 0.489109 -0.155929 0.292977 0.186230   
Proanthocyanins -0.499130 0.136698 -0.220746 0.009652   
Color\_intensity 0.265668 0.546364 0.248985 0.258887   
Hue -0.617369 -0.071747 -0.561296 -0.074667   
OD280/OD315\_of\_diluted\_wines -0.788230 0.072343 -0.368710 0.003911   
Proline -0.633717 0.643720 -0.192011 0.223626   
  
 Alcalinity\_of\_ash Magnesium Total\_phenols \  
Class 0.517859 -0.209179 -0.719163   
Alcohol -0.310235 0.270798 0.289101   
Malic\_acid 0.288500 -0.054575 -0.335167   
Ash 0.443367 0.286587 0.128980   
Alcalinity\_of\_ash 1.000000 -0.083333 -0.321113   
Magnesium -0.083333 1.000000 0.214401   
Total\_phenols -0.321113 0.214401 1.000000   
Flavanoids -0.351370 0.195784 0.864564   
Nonflavanoid\_phenols 0.361922 -0.256294 -0.449935   
Proanthocyanins -0.197327 0.236441 0.612413   
Color\_intensity 0.018732 0.199950 -0.055136   
Hue -0.273955 0.055398 0.433681   
OD280/OD315\_of\_diluted\_wines -0.276769 0.066004 0.699949   
Proline -0.440597 0.393351 0.498115   
  
 Flavanoids Nonflavanoid\_phenols \  
Class -0.847498 0.489109   
Alcohol 0.236815 -0.155929   
Malic\_acid -0.411007 0.292977   
Ash 0.115077 0.186230   
Alcalinity\_of\_ash -0.351370 0.361922   
Magnesium 0.195784 -0.256294   
Total\_phenols 0.864564 -0.449935   
Flavanoids 1.000000 -0.537900   
Nonflavanoid\_phenols -0.537900 1.000000   
Proanthocyanins 0.652692 -0.365845   
Color\_intensity -0.172379 0.139057   
Hue 0.543479 -0.262640   
OD280/OD315\_of\_diluted\_wines 0.787194 -0.503270   
Proline 0.494193 -0.311385   
  
 Proanthocyanins Color\_intensity Hue \  
Class -0.499130 0.265668 -0.617369   
Alcohol 0.136698 0.546364 -0.071747   
Malic\_acid -0.220746 0.248985 -0.561296   
Ash 0.009652 0.258887 -0.074667   
Alcalinity\_of\_ash -0.197327 0.018732 -0.273955   
Magnesium 0.236441 0.199950 0.055398   
Total\_phenols 0.612413 -0.055136 0.433681   
Flavanoids 0.652692 -0.172379 0.543479   
Nonflavanoid\_phenols -0.365845 0.139057 -0.262640   
Proanthocyanins 1.000000 -0.025250 0.295544   
Color\_intensity -0.025250 1.000000 -0.521813   
Hue 0.295544 -0.521813 1.000000   
OD280/OD315\_of\_diluted\_wines 0.519067 -0.428815 0.565468   
Proline 0.330417 0.316100 0.236183   
  
 OD280/OD315\_of\_diluted\_wines Proline   
Class -0.788230 -0.633717   
Alcohol 0.072343 0.643720   
Malic\_acid -0.368710 -0.192011   
Ash 0.003911 0.223626   
Alcalinity\_of\_ash -0.276769 -0.440597   
Magnesium 0.066004 0.393351   
Total\_phenols 0.699949 0.498115   
Flavanoids 0.787194 0.494193   
Nonflavanoid\_phenols -0.503270 -0.311385   
Proanthocyanins 0.519067 0.330417   
Color\_intensity -0.428815 0.316100   
Hue 0.565468 0.236183   
OD280/OD315\_of\_diluted\_wines 1.000000 0.312761   
Proline 0.312761 1.000000

## Iris Notebook

# Statistical Analysis: Iris Dataset

We will calculate the following statistics for each attribute

* [] Central Tendancy
  + Mean
  + Median
  + Mode
* [] Dispersion
  + Range
  + Quartiles
  + InterQuartile Range
  + Variance
  + Santard Deviation
* [] Covariance Matrix
* [] Correlation Matrix

## Importing necessary Libraries

**import** pandas **as** pd

## Load datasets

iris\_df = pd.read\_csv('../iris/iris.data', header=None)  
iris\_df.columns = [  
 "sepal\_length", "sepal\_width", "petal\_length", "petal\_width", "class"   
]  
iris\_df

sepal\_length sepal\_width petal\_length petal\_width class  
0 5.1 3.5 1.4 0.2 Iris-setosa  
1 4.9 3.0 1.4 0.2 Iris-setosa  
2 4.7 3.2 1.3 0.2 Iris-setosa  
3 4.6 3.1 1.5 0.2 Iris-setosa  
4 5.0 3.6 1.4 0.2 Iris-setosa  
.. ... ... ... ... ...  
145 6.7 3.0 5.2 2.3 Iris-virginica  
146 6.3 2.5 5.0 1.9 Iris-virginica  
147 6.5 3.0 5.2 2.0 Iris-virginica  
148 6.2 3.4 5.4 2.3 Iris-virginica  
149 5.9 3.0 5.1 1.8 Iris-virginica  
  
[150 rows x 5 columns]

## Handle 'class' attribute which is a categorial data

*# Handling categorical data*  
  
*# Iris-setosa -> 0*  
iris\_df.loc[iris\_df['class'] == 'Iris-setosa', 'class'] = 0  
*# Iris-versicolor -> 1*  
iris\_df.loc[iris\_df['class'] == 'Iris-versicolor', 'class'] = 1  
*# Iris-virginica -> 2*  
iris\_df.loc[iris\_df['class'] == 'Iris-virginica', 'class'] = 2  
  
iris\_df

sepal\_length sepal\_width petal\_length petal\_width class  
0 5.1 3.5 1.4 0.2 0  
1 4.9 3.0 1.4 0.2 0  
2 4.7 3.2 1.3 0.2 0  
3 4.6 3.1 1.5 0.2 0  
4 5.0 3.6 1.4 0.2 0  
.. ... ... ... ... ...  
145 6.7 3.0 5.2 2.3 2  
146 6.3 2.5 5.0 1.9 2  
147 6.5 3.0 5.2 2.0 2  
148 6.2 3.4 5.4 2.3 2  
149 5.9 3.0 5.1 1.8 2  
  
[150 rows x 5 columns]

## Compute central tendency (mean, median, mode)

mean\_values = iris\_df.mean()  
median\_values = iris\_df.median()  
mode\_values = iris\_df.mode().iloc[0] *# mode() returns a dataframe, so we take the first row*  
  
*# Combine them into a single DataFrame*  
central\_tendency\_table = pd.DataFrame({  
 'Mean': mean\_values,  
 'Median': median\_values,  
 'Mode': mode\_values  
})  
  
*# Display the table*  
central\_tendency\_table

Mean Median Mode  
sepal\_length 5.843333 5.8 5.0  
sepal\_width 3.054 3.0 3.0  
petal\_length 3.758667 4.35 1.5  
petal\_width 1.198667 1.3 0.2  
class 1.0 1.0 0

## Measures of Dispersion (range, quartiles, InterQuartile Ranges, Variance, Standard Deviations)

iris\_dispersion = iris\_df.describe().T  
iris\_dispersion['IQR'] = iris\_dispersion['75%'] - iris\_dispersion['25%']  
iris\_dispersion['Variance'] = iris\_dispersion['std'] \*\* 2  
  
iris\_dispersion

count mean std min 25% 50% 75% max IQR \  
sepal\_length 150.0 5.843333 0.828066 4.3 5.1 5.80 6.4 7.9 1.3   
sepal\_width 150.0 3.054000 0.433594 2.0 2.8 3.00 3.3 4.4 0.5   
petal\_length 150.0 3.758667 1.764420 1.0 1.6 4.35 5.1 6.9 3.5   
petal\_width 150.0 1.198667 0.763161 0.1 0.3 1.30 1.8 2.5 1.5   
  
 Variance   
sepal\_length 0.685694   
sepal\_width 0.188004   
petal\_length 3.113179   
petal\_width 0.582414

# Compute covariance and correlation matrices

iris\_covariance = iris\_df.cov()  
iris\_correlation = iris\_df.corr()

### IRIS Covariance Matrix

iris\_covariance

sepal\_length sepal\_width petal\_length petal\_width class  
sepal\_length 0.685694 -0.039268 1.273682 0.516904 0.530872  
sepal\_width -0.039268 0.188004 -0.321713 -0.117981 -0.148993  
petal\_length 1.273682 -0.321713 3.113179 1.296387 1.371812  
petal\_width 0.516904 -0.117981 1.296387 0.582414 0.597987  
class 0.530872 -0.148993 1.371812 0.597987 0.671141

### IRIS Correlation Matrix

iris\_correlation

sepal\_length sepal\_width petal\_length petal\_width class  
sepal\_length 1.000000 -0.109369 0.871754 0.817954 0.782561  
sepal\_width -0.109369 1.000000 -0.420516 -0.356544 -0.419446  
petal\_length 0.871754 -0.420516 1.000000 0.962757 0.949043  
petal\_width 0.817954 -0.356544 0.962757 1.000000 0.956464  
class 0.782561 -0.419446 0.949043 0.956464 1.000000