

Report

HW1

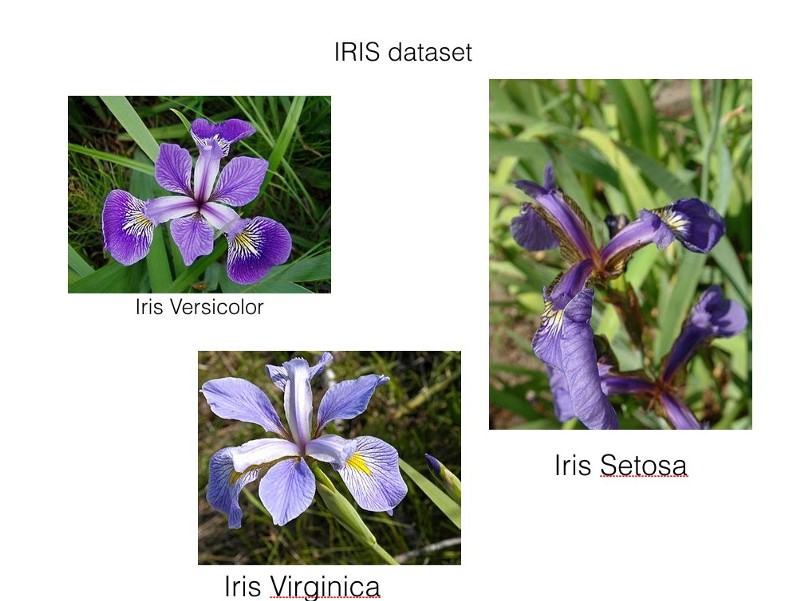
Yasaman Mirmohammad | Data Mining | Fall\_2018

# Task 1:

# Introduction

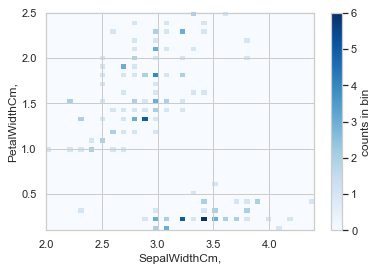
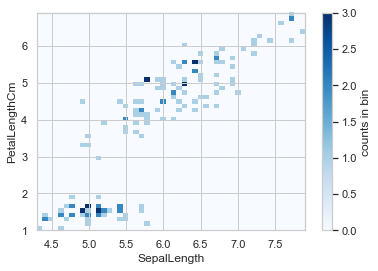
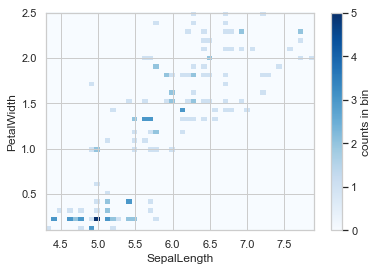
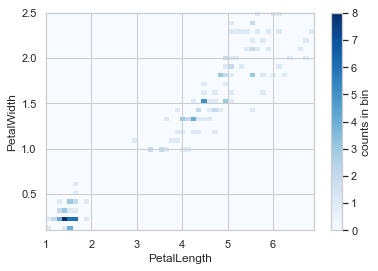
The ***Iris* flower data**is a multivariate [d](https://en.wikipedia.org/wiki/Data_set)ata set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems* as an example of linear discriminant analysis . It is sometimes called **Anderson’s *Iris* data set** because [Edgar Anderson](https://en.wikipedia.org/wiki/Edgar_Anderson) collected the data to quantify the morphologic variation of *[Iris](https://en.wikipedia.org/wiki/Iris_%28plant%29" \t "_blank" \o "Iris (plant))*flowers of three related species. Two of the three species were collected in the [Gaspé Peninsula](https://en.wikipedia.org/wiki/Gasp%C3%A9_Peninsula" \t "_blank" \o "Gaspé Peninsula) “all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus”.

* + Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  + Four (non-class) attributes
    - Sepal width and length
    - Petal width and length



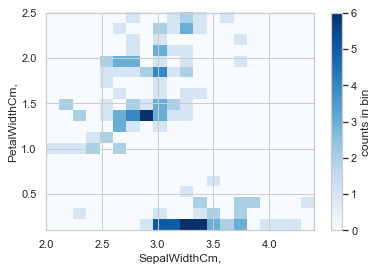
A)Plot the histograms:

2 dimenional Histogram:

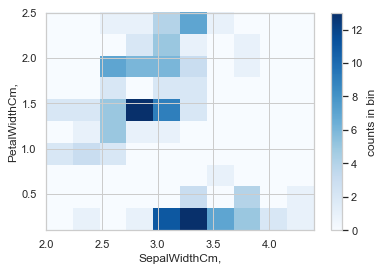


If we change number of the bins, we will get result like ths(previous results are with bin=50):

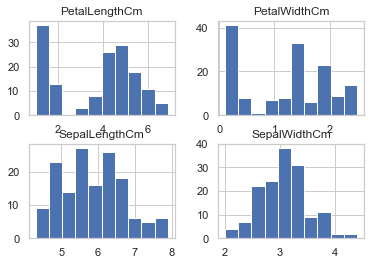
Bin=20:



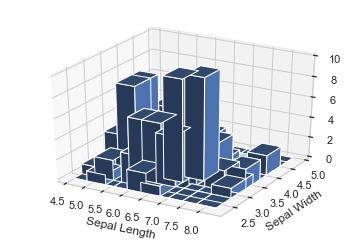
Bin=10:



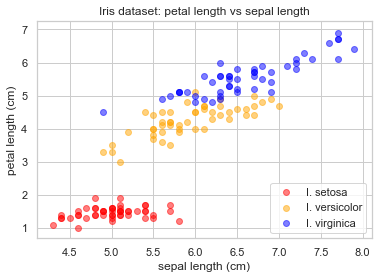
1 dimensional:

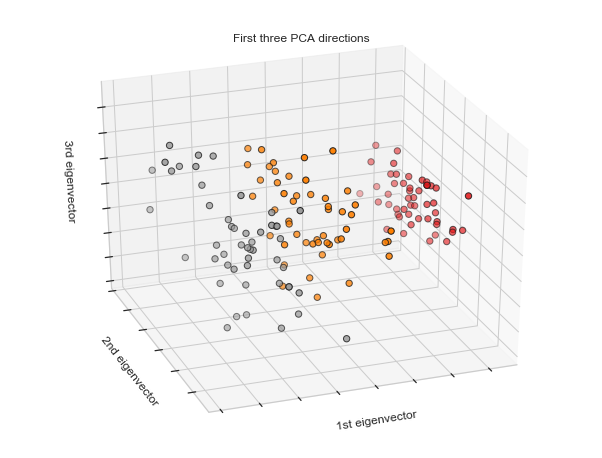


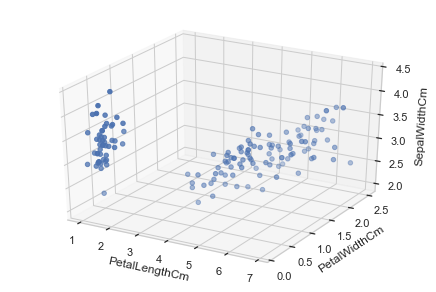
B) 3D Histogram

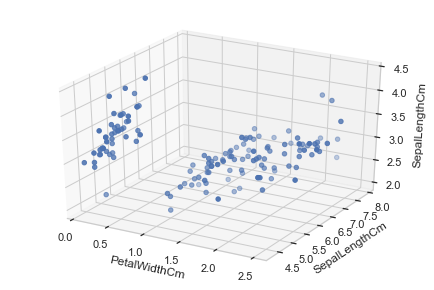
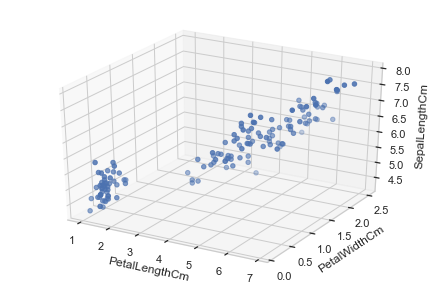


C)Scatter Data(next page):



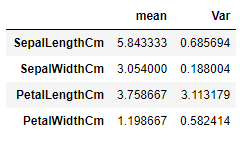


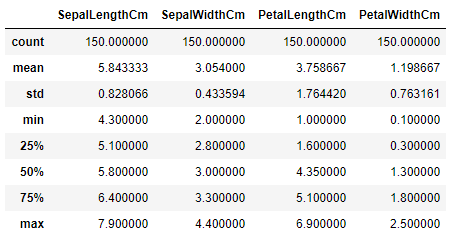




## 

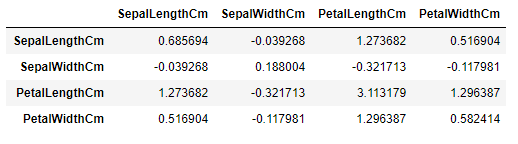
D) Mean and Variance:





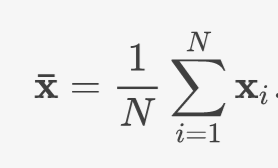
E)Covariance:

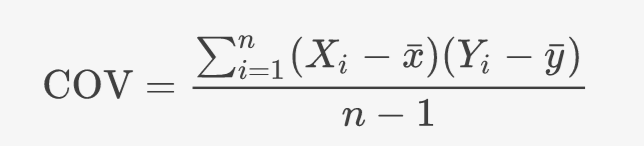
On features:



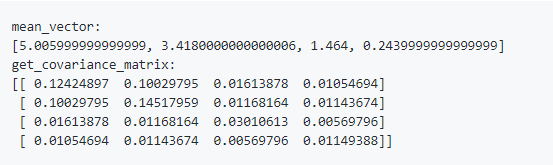
For calculating covariance on classes, we have to use Mean\_Covariance:

Mean vector:



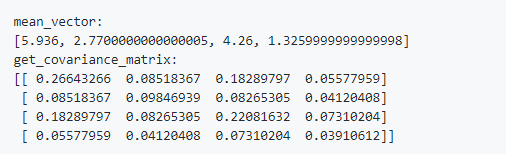


Setosa:

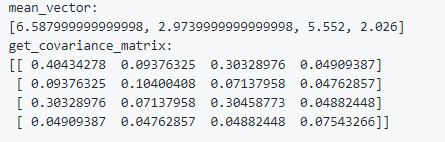


‘

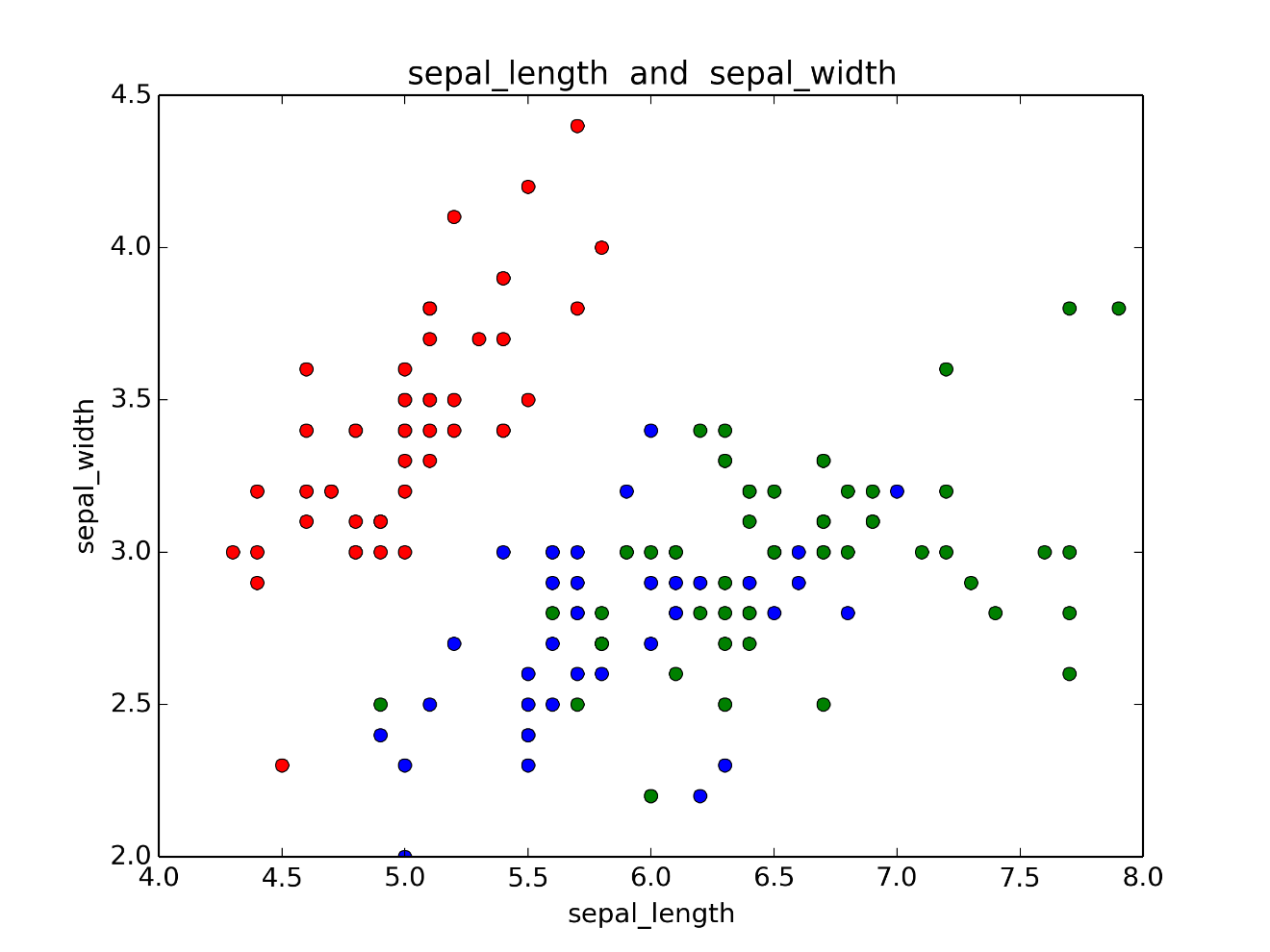
Iris\_versicolor:



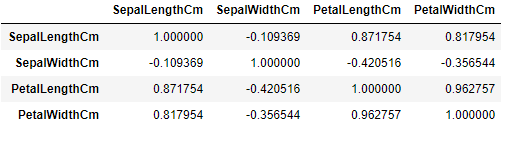
Iris\_virginica:



Considering the two features, sepal\_length and sepal\_width (mean\_vector[0] and mean\_vector[1]), we find Iris\_setosa(Red) is far from the others. By contrast, Iris\_versicolor(Blue) and Iris\_virginica(Green) are near each other.



F)Correlation:



Lets calculate correlation on each class:

Setosa-versicolor:

array([[1. , 0.7638543],

[0.7638543, 1. ]])

Setosa\_virginica:

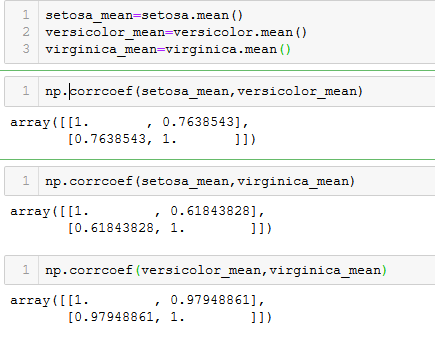
array([[1. , 0.61843828],

[0.61843828, 1. ]])

Virginica\_versicolor:

array([[1. , 0.97948861],

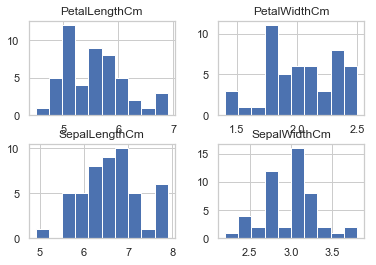
[0.97948861, 1. ]])



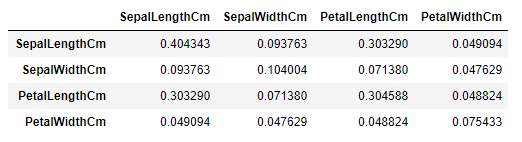
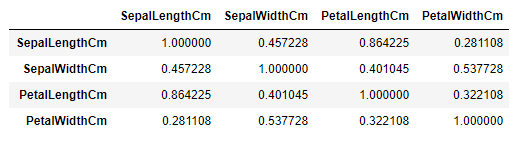
The most important difference between Correlation Matrix and Covariance .Matrix is that:

* ***Correaltion Matrix main diagonal is 1.( each feature has correlation 1 with itself)***
* Correlation is when the change in one item may result in the change in the another item. On the other hand, covariance is when two items vary together.
* Correlation is concerned as A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
* Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.
* The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between -∞ and +∞.
* Covariance is affected by the change in scale, i.e. if all the value of one variable is multiplied by a constant and all the value of another variable are multiplied, by a similar or different constant, then the covariance is changed. As against this, correlation is not influenced by the change in scale.
* Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.

G)Viriginica



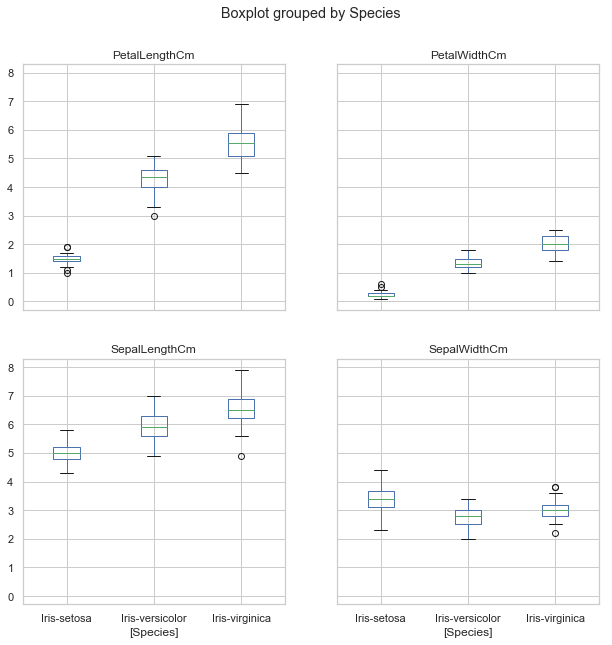
Correlation and Covariance:



Based on Correlation,”PetalLength” and “Sepal Length” are more similar.

H)Feature Selection:

I used boxplots to find out which feature seperates the classes better(I more informative)



Based On these Results, Best Features is PetalWidth

I tested this task with “SelectKBest” Function in python and the result was the same.

## Task2

### **What is gradient descent ?**

It is an optimization algorithm to find the minimum of a function. We start with a random point on the function and move in the **negative direction** of the **gradient of the function**to reach the **local/global minima**.

Cost fuction is SSE:

J(w)=12∑i(target(i)−output(i))2  ;output(i)∈R

We implement Gradient Descent with the appropriate error anf plot it.

Iteration 0 | Cost: 84604307330.630157

Iteration 1 | Cost: 82781761231.728531

Iteration 2 | Cost: 80998484776.147720

Iteration 3 | Cost: 79253631835.952148

Iteration 4 | Cost: 77546374514.400330

Iteration 5 | Cost: 75875902753.124390

Iteration 6 | Cost: 74241423947.773407

Iteration 7 | Cost: 72642162571.938477

Iteration 8 | Cost: 71077359809.180817

Iteration 9 | Cost: 69546273192.988403

Iteration 10 | Cost: 68048176254.490433

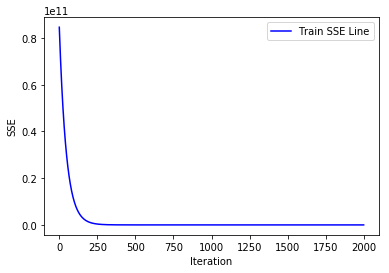
Iteration 11 | Cost: 66582358177.762253

…..

Iteration 1999 | Cost: 13541479.738889

SSE Test = 6386268.633887

Train = 13539468.826632



Bonus Part(3d plot):

