ECE 657A : Data and Knowledge Modeling and Analysis

Assignment 1 : Basic Environment Set-up and Classification

Iris dataset

Libraries Used:

- numpy
- pandas
- seaborn
- matplotlib
- scipy
- scikit-learn

[CM1]

Question 1: Data Exploration

Importing libraries

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

Load Iris dataset

This dataset includes different features (attributes) of three Iris flower species (setosa, versicolor, virginica). The features are 'Petal Length', 'Petal Width', 'Sepal Length', 'Sepal Width'.

[2]:
```
# load dataset
df_iris= pd.read_csv('iris_dataset_missing.csv')
```

Displaying and exploring the Iris DataFrame created:

[3]:
```
df_iris.describe()
```

[3]:

	sepal_length	sepal_width	petal_length	petal_width
count	105.000000	101.000000	97.000000	105.000000
mean	5.858909	3.059083	3.812370	1.199708
std	0.861638	0.455116	1.793489	0.787193
min	4.344007	1.946010	1.033031	-0.072203
25%	5.159145	2.768688	1.545136	0.333494
50%	5.736104	3.049459	4.276817	1.331797
75%	6.435413	3.290318	5.094427	1.817211
max	7.795561	4.409565	6.768611	2.603123

[4]:
```
df_iris.head()
```

[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.045070	2.508203	3.018024	1.164924	Iris-versicolor

```
1       6.325517      2.115481      4.542052      1.413651  Iris-versicolor
2       5.257497      3.814303      1.470660      0.395348      Iris-setosa
3       6.675168      3.201700      5.785461      2.362764   Iris-virginica
4       5.595237      2.678166      4.077750      1.369266  Iris-versicolor
```

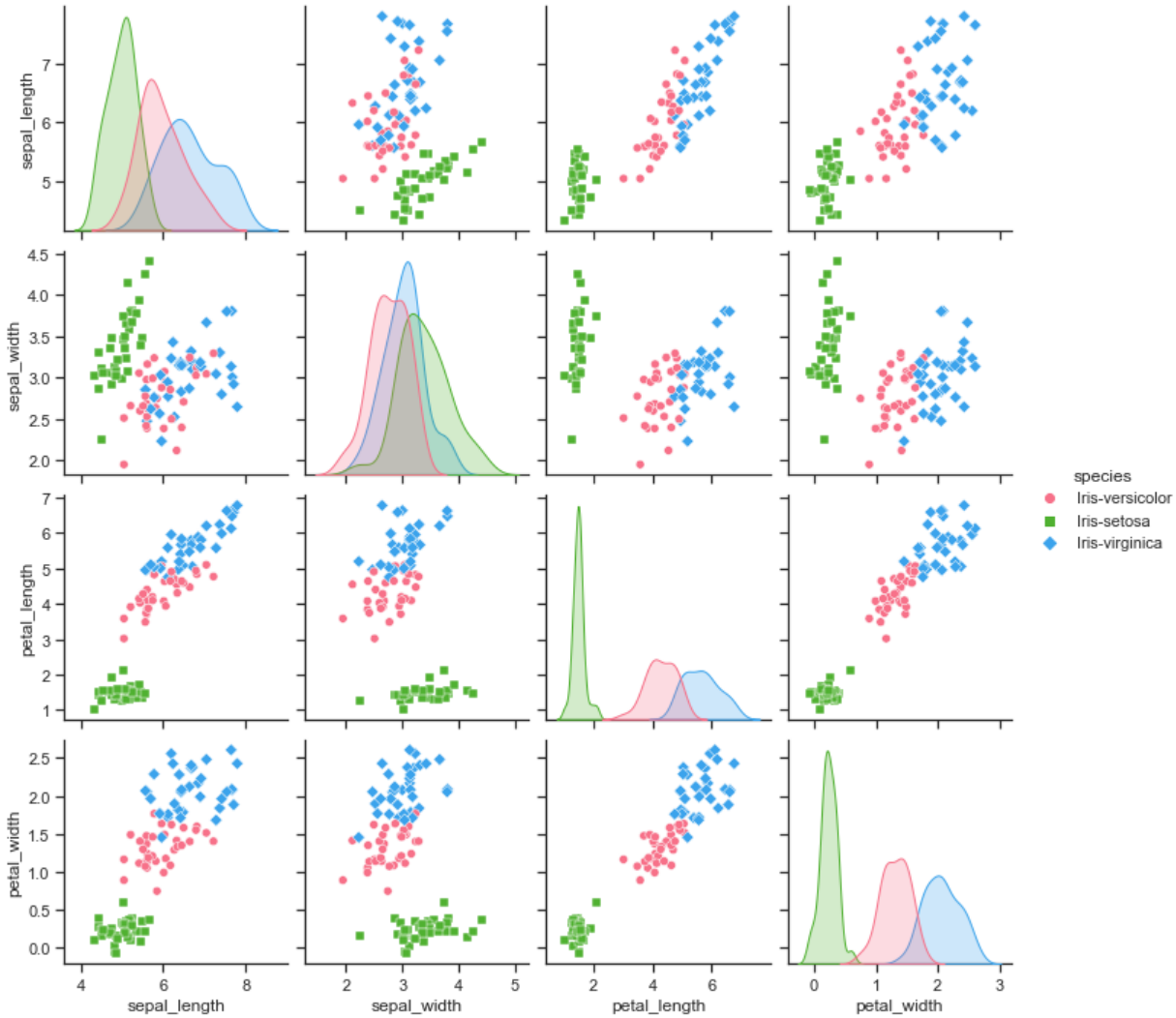
[5]:
```
df_iris.columns
```

[5]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

Visualizing the data distribution by generating "pair plots" (using pairplot method of the seaborn library)

[6]:
```
# pairplot
sns.set(style='ticks', color_codes=True)
sns.pairplot(df_iris, hue='species', palette='husl', markers=['o', 's', 'D'])
```

[6]:
```
<seaborn.axisgrid.PairGrid at 0x1e6d22ae370>
```



From the "pair plot" visualization, we observe that :

- petal length and petal width are most positively correlated as we see a linear increase between the feaures. The scatter plot aligns with a linear line function.
- we observe a similar pattern with petal length and sepal length where there is linear positive correlation.
- In all the plots, Iris-setosa is easily distinguishable and can be identified irrespective of petal or sepal features. By using petal length, we can distinctly seperate Iris-setosa.
- For Iris-versicolor and Iris-virginica, we see that the plots are mostly overlapping, but petal features provide better distinction than sepal features.

[CM2]

Correlation coefficient of each pair of features

Heat map is used to find out the correlation between different features in the dataset. High positive or negative value shows that the features have high correlation

[7]:
```
# Get correlations of each features in dataset
corrmat = df_iris.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))

# Plot heat map
iris_heat_map=sns.heatmap(df_iris[top_corr_features].corr(),annot=True,cmap='RdYlGn')
```