Required

Once you have selected a data set, you will produce the deliverables listed below and submit them to one of your peers for review.Treat this exercise as an opportunity to produce analysis that are ready to highlight your analytical skills for a senior audience, for example, the Chief Data Officer, or the Head of Analytics at your company.

Sections required in your report:

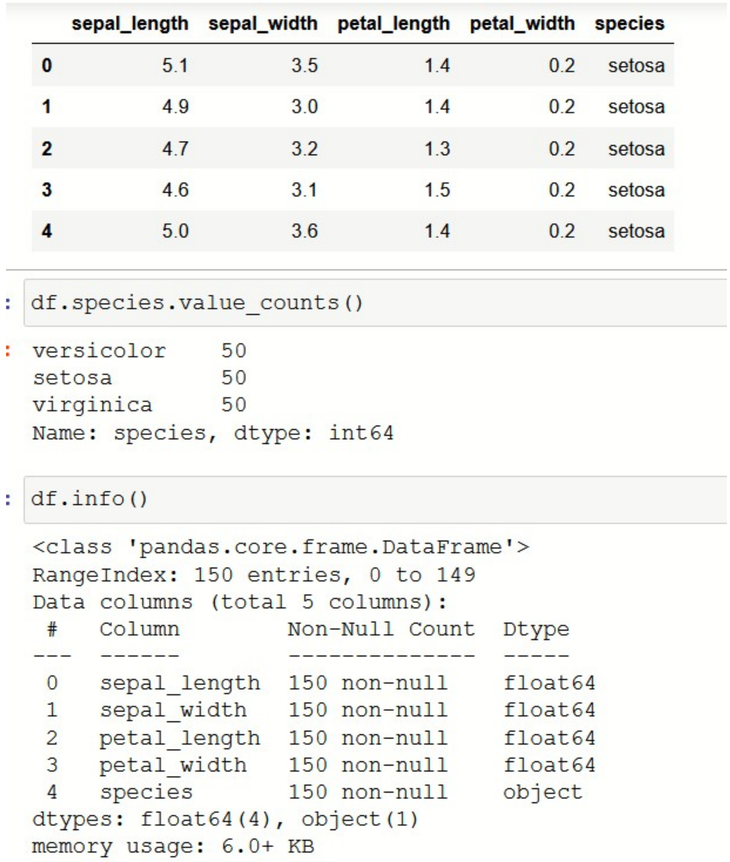
Brief description of the data set and a summary of its attributes

The Iris flower data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher - Wikipedia

The Iris data sets consists of 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 numpy.ndarray

The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width.

Chosen because this is one of the several classic data sets that have been used extensively in the statistical literature



	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

```
: df.species.value_counts()

: versicolor    50
  setosa        50
  virginica     50
  Name: species, dtype: int64
```

```
: df.info()

  <class 'pandas.core.frame.DataFrame'>
  RangeIndex: 150 entries, 0 to 149
  Data columns (total 5 columns):
   #   Column        Non-Null Count  Dtype
  ---  ------        --------------  -----
   0   sepal_length  150 non-null    float64
   1   sepal_width   150 non-null    float64
   2   petal_length  150 non-null    float64
   3   petal_width   150 non-null    float64
   4   species       150 non-null    object
  dtypes: float64(4), object(1)
  memory usage: 6.0+ KB
```

Initial plan for data exploration

- ➔ Check missing data
- ➔ Histogram of frequency and sepal length, width
- ➔ Histogram of frequency and petal length, width
- ➔ Pair plot
- ➔ Heat map to check correlation between attributes
- ➔ Swarm plot and box plot

Implementation given in notebook file

Actions taken for data cleaning and feature engineering

- ➔ Deal with the outliers individually for each species by dropping those observations
- ➔ If any observations are missing, replace with the mean value of the attribute of that species

Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

- ➔ No missing data
- ➔ Histogram give an idea of the frequency distribution of data
- ➔ Heatmap show some positive correlation between petal length and width, sepal length with petal length/width and negative correlation between sepal width with petal length and width
- ➔ Pairplot slows data is normally distributed
- ➔ From the swarm and box plot, some difference between the sepal length distribution of the 3 categories is observed.

Formulating at least 3 hypothesis about this data

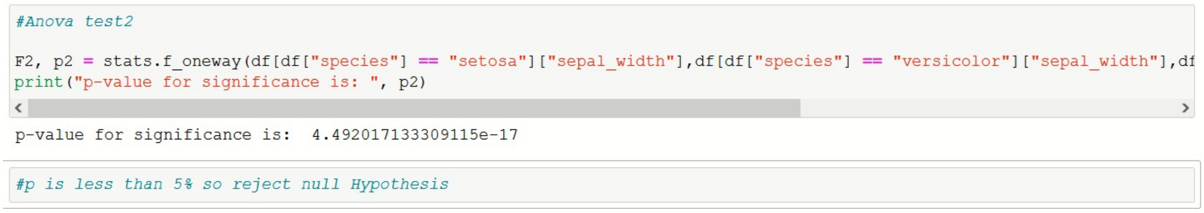
- ➔ There is no significant difference between the species in sepal width
- ➔ There is no significant difference between the species in sepal length
- ➔ There is no significant difference between sepal length of verginica, versicolor
- ➔ There is no significant difference between sepal width of verginica, versicolor

Conducting a formal significance test for one of the hypotheses and discuss the results

Anova test to test if there is a significant difference between the species in terms of sepal width.

H0 null hypothesis - there is no difference between groups

H1 alternate hypothesis - difference between groups



```
#Anova test2

F2, p2 = stats.f_oneway(df[df["species"] == "setosa"]["sepal_width"],df[df["species"] == "versicolor"]["sepal_width"],df
print("p-value for significance is: ", p2)
```

```
p-value for significance is:  4.492017133309115e-17
```

```
#p is less than 5% so reject null Hypothesis
```

p is less than 5% so reject null hypothesis. Thus there is a significant difference between the species.

Suggestions for next steps in analyzing this data

- ➔ Get more testing data to test if the applied machine learning model can predict the species given the attributes.
- ➔ Find the limits of confidence levels on previous hypotheses

A paragraph that summarizes the quality of this data set and a request for additional data if needed. - NA