Basic data analytics on Iris

Tushar B. Kute,
http://tusharkute.com

Problem Statement

- Download the Iris flower dataset or any other dataset into data frame. Use python or R to perform following:
  - How many feathers are there and their types (nominal or numeric)?
  - Compute and display summary statistics for each feature available in the dataset (e.g. minimum value, maximum value, range, standard deviation, variance and percentiles.
  - Data visualization- creating histogram for each feature in the dataset to illustrate the feature distribution. Plot each histogram.
  - Create boxplot for each feature in the dataset. All of the boxplots should be combined in the single plot. Compare distributions and identify the outliers.

The extra modules needed

- Pandas
- Matplotlib
- Sklearn
- How to install?
  - `sudo apt-get install python3-pandas`
  - `sudo apt-get install python3-matplotlib`
  - `sudo apt-get install python3-sklearn`

The Dataset

https://archive.ics.uci.edu/ml/datasets/iris

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1953764

The Dataset

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Import packages

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
```

Get information

```
iris = load_iris()

data_set = pd.DataFrame(data= np.c_[iris['data'], iris['target']],
                        columns= iris['feature_names'] + ['target'])

# Complete information of all attributes of Iris
print('\n', 'DATA SET INFORMATION'.center(45, '_'))
print(data_set.info())

# Description with Statistics of Iris Dataset
print('\n', 'STATISTICAL INFORMATION'.center(45, '_'))
print(data_set.describe())
```

Get information

```
# Locate a specific type of Data type in Data set
print('\n', 'COLUMNS DTYPE (IF NOMINAL)'.center(45, '_'))
print(data_set.select_dtypes(include=['category']))

# Data types of Iris column wise, to locate ordinal & nominal
print('\n', 'COLUMNS DTYPE (ALL)'.center(45, '_'))
print(data_set.dtypes)

# Memory occupancy done by Dataset
print('\n', 'DATA SET MEMORY USAGE'.center(45, '_'))
print(data_set.memory_usage())
```

Missing Values

```
# Counting any missing data (if any else zero)
def num_missing(x):
  return sum(x.isnull())

#Applying per column:
print('\n', 'MISSING VALUE CHECK'.center(45, '_'))
print("Missing values per column:")
print(data_set.apply(num_missing, axis=0))
#axis=0 defines that function is to be applied on each column
```

Histogram

```
# Plotting Histogram
# #1 All Features
data_set[['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)','petal width
(cm)']].plot.hist(bins=10,

title='All features')
plt.show()

# #2 Only 2 features at a time
data_set[['sepal length (cm)', 'sepal width (cm)']].plot.hist(bins=10, title='Sepal
Features')
plt.show()

data_set[['petal length (cm)','petal width (cm)']].plot.hist(bins=10, title='Petal
Features')
plt.show()
```

Boxplot

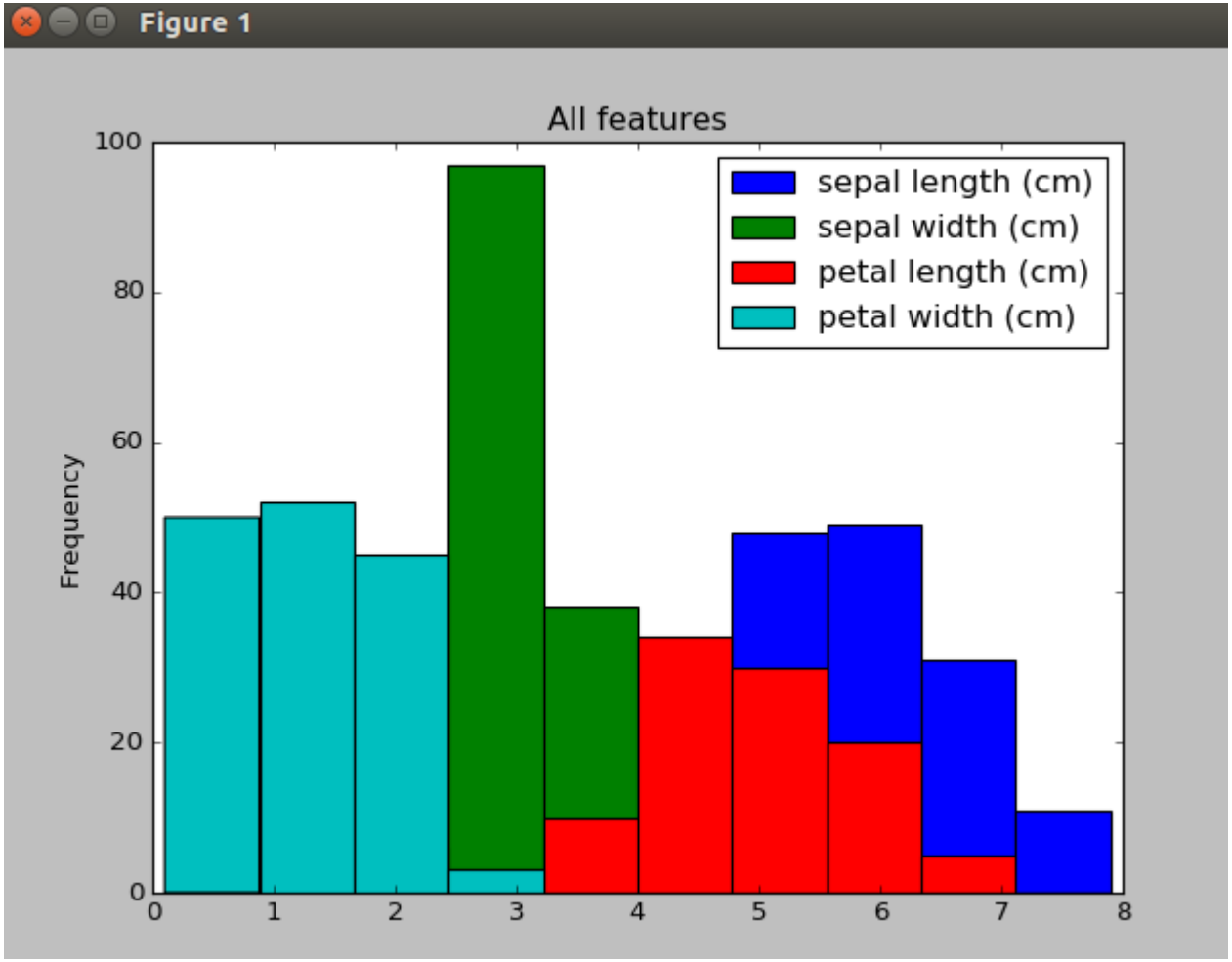
```
# Plotting Boxplot
data_set.plot.box(title="All Features with outliers")
plt.show()
# Try noticing the 'o', those are outliers. The ones who
(IQR) are Outliers
```

Sample Output

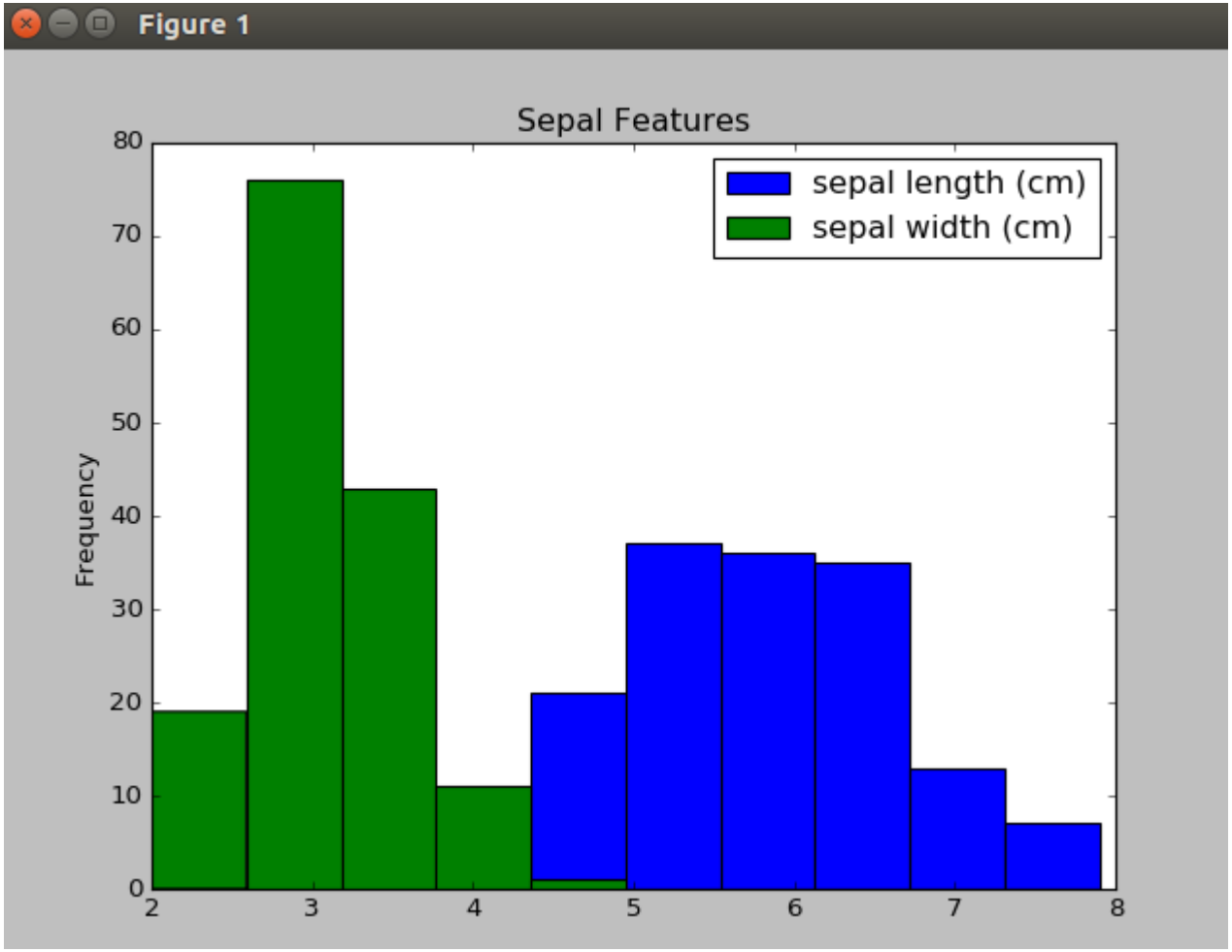
```
_____________DATA SET INFORMATION_____________
<class 'pandas.core.frame.DataFrame'>
Int64Index: 150 entries, 0 to 149
Data columns (total 5 columns):
sepal length (cm)    150 non-null float64
sepal width (cm)     150 non-null float64
petal length (cm)    150 non-null float64
petal width (cm)     150 non-null float64
target               150 non-null float64
dtypes: float64(5)
memory usage: 7.0 KB
None

____________STATISTICAL INFORMATION____________
       sepal length (cm)  sepal width (cm)  petal length (cm)  \
count         150.000000        150.000000         150.000000
mean            5.843333          3.054000           3.758667
std             0.828066          0.433594           1.764420
min             4.300000          2.000000           1.000000
25%             5.100000          2.800000           1.600000
50%             5.800000          3.000000           4.350000
75%             6.400000          3.300000           5.100000
max             7.900000          4.400000           6.900000
```

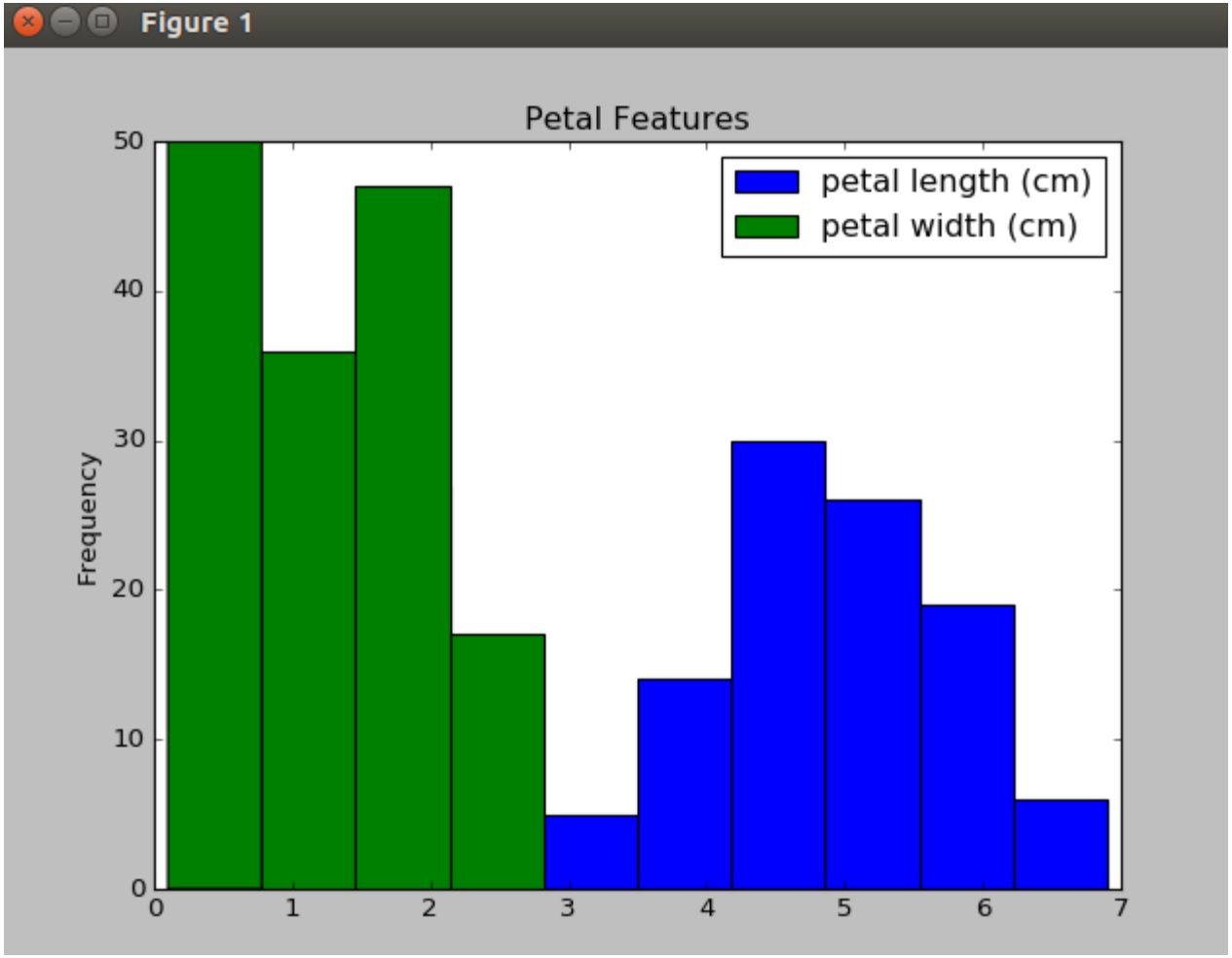
Sample Output



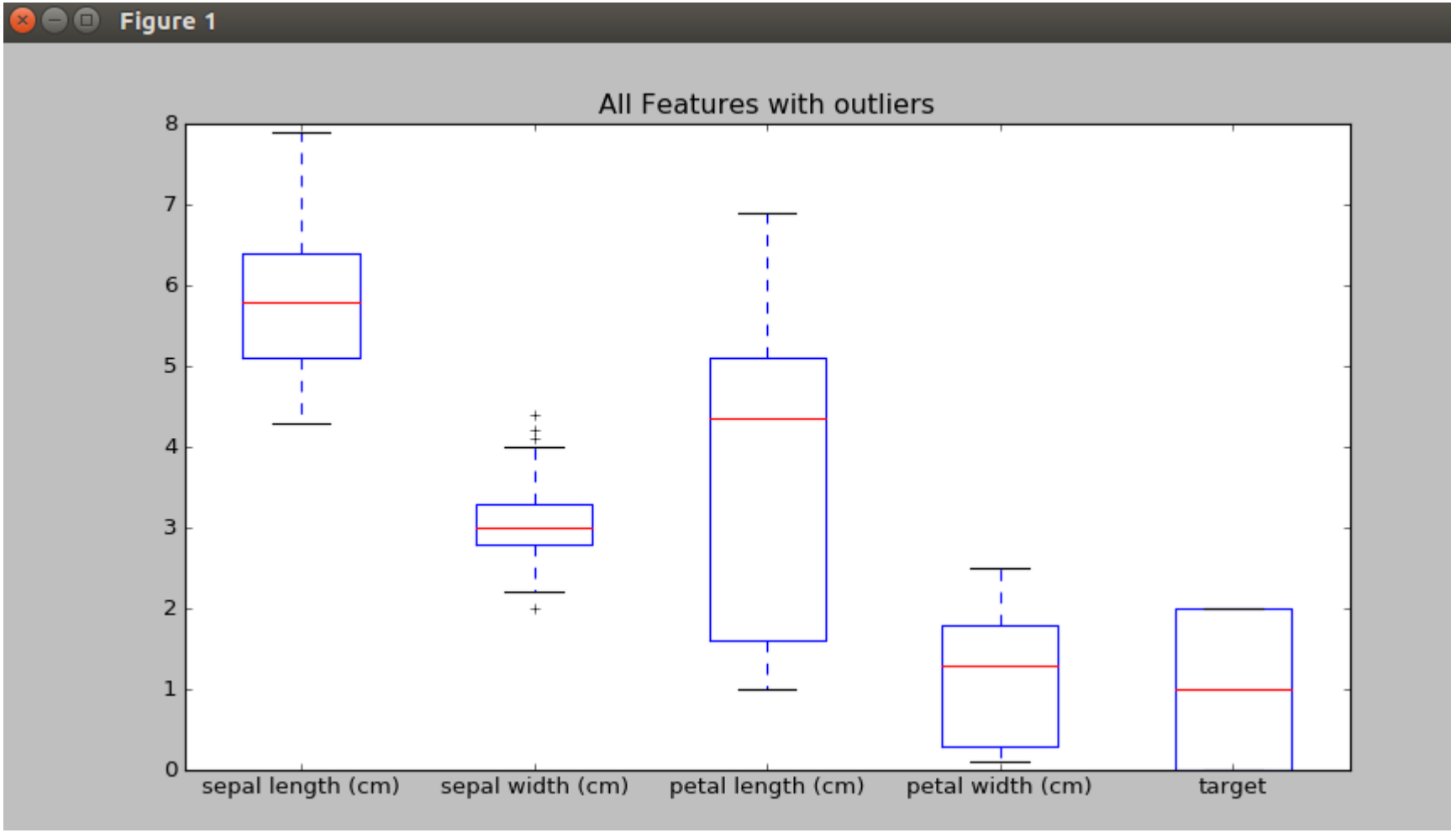
Sample Outputs



Sample outputs



Sample output



Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

Web Resources
http://mitu.co.in
http://tusharkute.com

tushar@tusharkute.com

contact@mitu.co.in