

Group A

Assignment No. 10

Aim : Data Visualization III

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., <https://archive.ics.uci.edu/ml/datasets/Iris>). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a box plot for each feature in the dataset.
4. Compare distributions and identify outliers.

Solⁿ :

OUTPUT :

The screenshot shows a Kaggle notebook interface with the following components:

- Browser Tabs:** New Tab, whatsapp web - Saferbrowser, (4) WhatsApp, how to compare distribution, kaggle.com - Saferbrowser, notebook24f3ff955 | Kaggle.
- URL:** kaggle.com/code/kshitijsinde/notebook24f3ff955/edit
- Notebook Title:** notebook24f3ff955 (Failed to save draft).
- Menu Bar:** File, Edit, View, Run, Add-ons, Help.
- Toolbar:** +, -, Run All, Code, Draft Session (39m), Run, Refresh, Help.
- Code Cell [20]:**

```
import numpy as np
import pandas as pd

df=pd.read_csv('../input/irisdataset/iris.data')
```
- Code Cell [21]:**

```
df.head()
```
- Output [21]:**

	5.1	3.5	1.4	0.2	Iris-setosa
0	4.9	3.0	1.4	0.2	Iris-setosa
1	4.7	3.2	1.3	0.2	Iris-setosa
2	4.6	3.1	1.5	0.2	Iris-setosa
3	5.0	3.6	1.4	0.2	Iris-setosa
4	5.4	3.9	1.7	0.4	Iris-setosa
- Code Cell [3]:**

```
df.info()
```
- Right Panel:** Data, + Add data, Input (irisdataset, iris.data), Output (44.1MB / 19.6GB), /kaggle/working, Settings, Schedule a notebook run, Code Help, FIND CODE HELI, Find Code Help, Search for examples of how to do things.
- Console:** Type here to search.
- Taskbar:** Windows Start button, Type here to search, File Explorer, Mail, Store, Edge, Camera, Photos, OneDrive, Google Chrome, System tray (Battery, Network, Volume, Date/Time: 11:02 PM 19/04/2022).

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.

The screenshot shows a Kaggle notebook interface. The top bar indicates the notebook is named 'notebook24f3ff955' and has failed to save a draft. The code cell [3] contains the following code:

```
2 4.6 3.1 1.5 0.2 Iris-setosa
3 5.0 3.6 1.4 0.2 Iris-setosa
4 5.4 3.9 1.7 0.4 Iris-setosa
```

The output of the code is a pandas DataFrame summary:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 149 entries, 0 to 148
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
---  -
 0   5.1         149 non-null    float64
 1   3.5         149 non-null    float64
 2   1.4         149 non-null    float64
 3   0.2         149 non-null    float64
 4   Iris-setosa 149 non-null    object  
dtypes: float64(4), object(1)
memory usage: 5.9+ KB
```

The right sidebar shows the 'Data' section with 'irisdataset' and 'iris.data' listed. The 'Output' section shows the file path '/kaggle/working'.

2. Create a histogram for each feature in the dataset to illustrate the feature distributions.

The screenshot shows the same Kaggle notebook interface. The code cell [22] contains the following code:

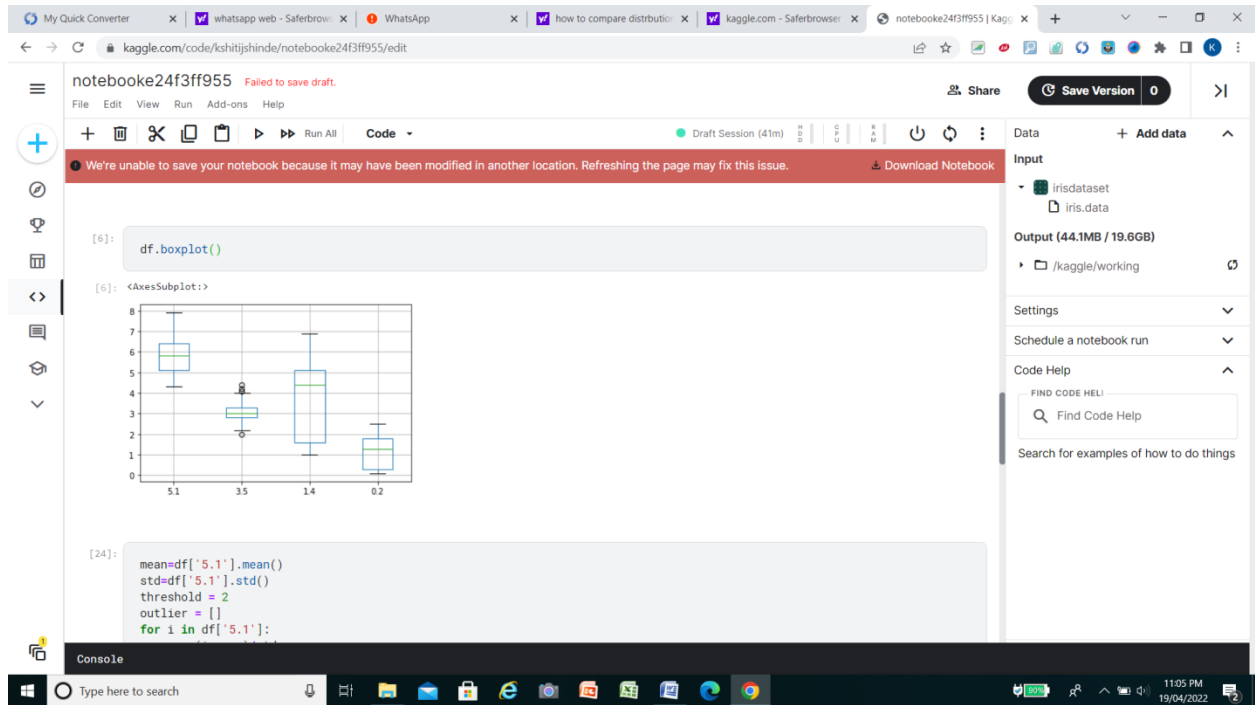
```
import matplotlib.pyplot as plt
%matplotlib inline
```

The code cell [23] contains the following code:

```
df.hist()
```

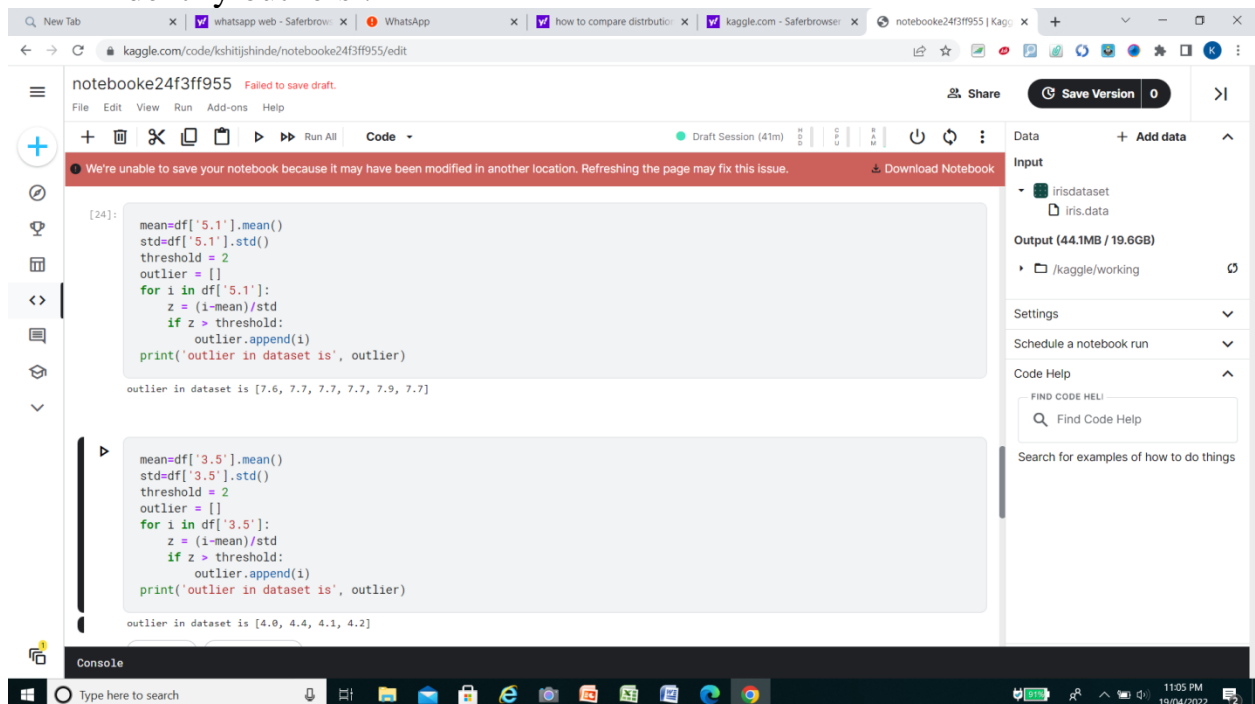
The output of the code is a 3x2 grid of histograms showing the distribution of the first three features (sepal length, sepal width, and sepal thickness) for the Iris dataset. The histograms are titled '5.1', '3.5', and '0.2' respectively.

3. Create a box plot for each feature in the dataset.



4. Compare distributions and identify outliers.

Identify outliers :



notebook24f3ff955 Failed to save draft.

We're unable to save your notebook because it may have been modified in another location. Refreshing the page may fix this issue. Download Notebook

```
[28]:
mean=df['1.4'].mean()
std=df['1.4'].std()
threshold = 1
outlier = []
for i in df['1.4']:
    z = (i-mean)/std
    if z > threshold:
        outlier.append(i)
print('outlier in dataset is', outlier)

outlier in dataset is [6.0, 5.9, 5.6, 5.8, 6.6, 6.3, 5.8, 6.1, 6.7, 6.9, 5.7, 6.7, 5.7, 6.0, 5.6, 5.8, 6.1, 6.4, 5.6, 5.6, 6.1, 5.6, 5.6, 5.9, 5.7]
```

```
[30]:
mean=df['0.2'].mean()
std=df['0.2'].std()
threshold = 1
outlier = []
for i in df['0.2']:
    z = (i-mean)/std
    if z > threshold:
        outlier.append(i)
print('outlier in dataset is', outlier)

outlier in dataset is [2.5, 2.1, 2.2, 2.1, 2.5, 2.0, 2.1, 2.0, 2.4, 2.3, 2.2, 2.3, 2.3, 2.0, 2.0, 2.1, 2.1, 2.0, 2.2, 2.3, 2.4, 2.1, 2.4, 2.3, 2.3, 2.5, 2.3, 2.0, 2.3]
```

Console

Comparing distributions :

notebook24f3ff955 Failed to save draft.

We're unable to save your notebook because it may have been modified in another location. Refreshing the page may fix this issue. Download Notebook

```
[31]:
df=pd.DataFrame({'a':['5.1','3.5','1.4','0.2'],
                'b':['5.1','3.5','1.4','0.2']})
```

```
[33]:
ax=df['a'].value_counts().plot(kind='bar', color='blue',width=.75, legend=True, alpha=0.8)
df['b'].value_counts().plot(kind='bar', color='maroon', width=.5, alpha=1, legend=True)
```

```
[33]: <AxesSubplot>
```

Console

OUTPUT IN RSTUDIO :

1. Importing dataset :

RStudio interface showing the 'acs' dataset loaded from a CSV file. The Environment pane shows 'acs' with 7811 observations and 14 variables. The console shows the R code used to load the data.

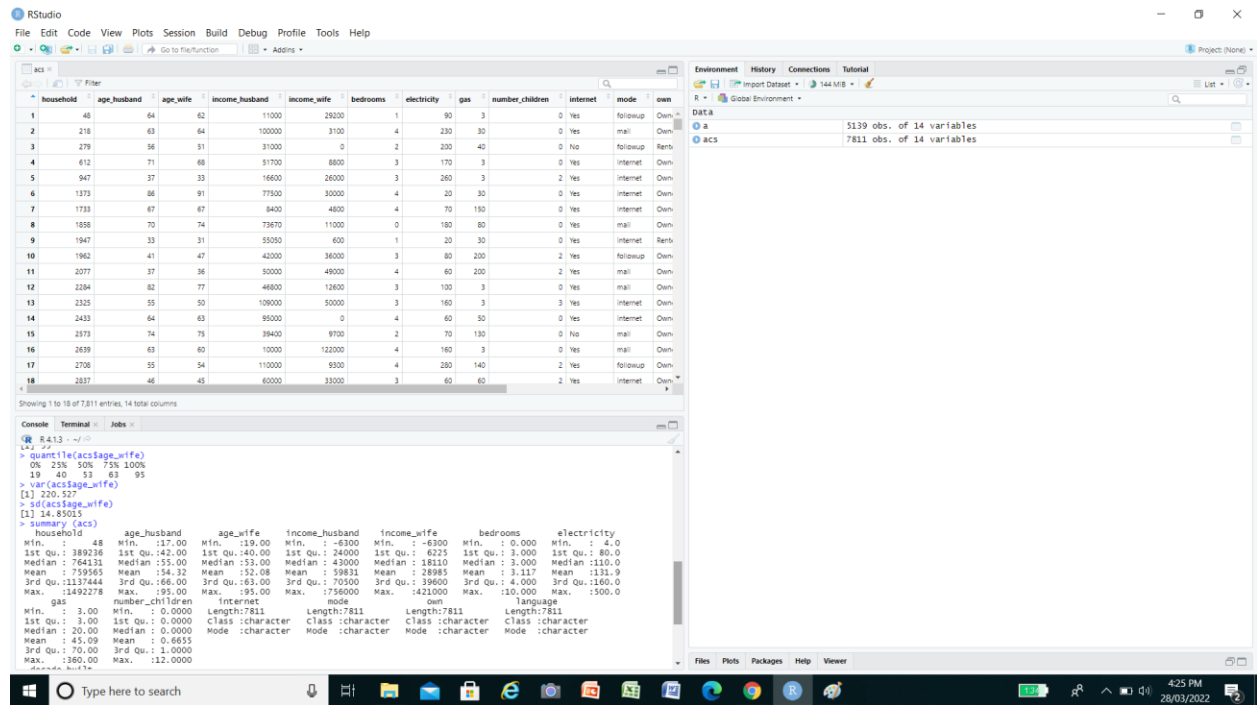
```
R413: ~
> acs <- read.csv(url("http://statlib.cwcl.co.nz/homeworks/acs_or.csv"))
> view(acs)
> acs$age_husband
[1] 64 63 56 71 37 86 67 70 33 41 37 82 55 64 74 63 55 46 63 39 63 58 45 30 84 37 50 40 49 54 77 37 58 64 69 61 33 37
[39] 67 51 73 60 30 40 77 69 32 61 50 70 54 57 46 88 31 46 52 50 44 44 45 31 54 65 43 57 57 43 27 59 47 61 90 30 48 67
[77] 78 55 37 29 59 48 29 34 29 41 18 62 74 43 65 34 58 48 60 66 59 28 65 47 33 38 58 43 50 67 66 43 51 37 33 36
[115] 31 80 65 61 67 65 65 70 55 24 90 90 49 35 73 55 33 61 45 33 68 74 73 60 61 47 63 43 53 39 41 44 70 57 76 70 40
[153] 62 55 21 24 23 38 35 63 72 54 71 56 30 71 49 29 55 33 46 45 68 34 62 59 71 64 59 51 42 74 64 81 63 57 46
[191] 38 73 42 60 57 67 34 42 71 42 69 74 62 75 49 59 66 52 33 61 87 50 41 47 59 50 51 29 43 30 56 58 34 60 59 28 65 39
[229] 61 43 72 82 58 51 71 47 56 41 84 54 47 28 66 52 32 59 42 49 44 28 37 47 63 66 55 53 32 47 67 57 55 30 65 47 34 58
[267] 74 46 43 40 21 56 49 52 53 42 18 75 95 35 43 46 66 50 72 47 44 71 58 71 39 67 73 48 39 39 72 51 60 70 44 59
[305] 69 45 68 42 42 66 32 69 42 76 54 34 24 32 68 64 44 49 73 81 39 52 95 51 74 27 59 40 36 25 26 35 59 28 54 54 46 59
[343] 49 41 64 68 29 71 38 58 43 56 54 46 54 33 45 42 31 80 81 84 44 60 66 40 77 56 49 79 50 37 49 55 69 65 41
[381] 55 39 51 53 47 79 51 32 42 41 30 66 54 53 75 49 37 63 32 42 66 54 34 35 33 36 28 80 37 60 57 64 60 57 64 55 36 77
[419] 46 40 43 26 52 33 74 35 53 56 62 81 64 61 54 47 44 72 57 79 74 74 42 63 55 40 75 52 66 34 29 65 55 69 56 60 57 39
[457] 81 57 42 45 63 68 56 54 59 77 38 51 37 40 70 46 51 68 60 41 31 44 32 62 44 70 49 58 74 74 74 72 50 55
[495] 69 71 69 63 37 46 43 58 47 46 88 31 61 51 48 47 36 53 60 84 57 45 74 61 66 38 35 63 57 34 57 52 70 64 40 58 50 43
[533] 28 34 66 35 37 27 78 87 73 50 43 43 29 65 34 30 62 72 54 42 36 39 56 75 41 70 61 35 72 32 62 75 69 37 55 38 36 85
[571] 64 53 31 68 78 33 62 29 27 65 30 31 54 58 24 62 58 47 71 47 46 61 86 34 51 34 52 62 55 58 74 56 62 71 32 73 29 35
[609] 71 59 57 60 89 49 66 75 54 72 59 69 62 87 58 46 52 25 52 58 50 40 49 80 64 33 47 76 65 43 21 66 29 60 65 71
[647] 56 63 59 55 53 26 44 63 76 36 51 58 42 70 75 32 63 39 44 63 55 59 40 43 32 37 42 66 66 77 83 49 84 28 49 78
[685] 57 64 65 37 49 28 71 31 49 42 45 77 39 34 41 43 39 62 48 48 53 68 69 62 42 73 62 61 46 57 66 73 37 43 33 30 63 63
[723] 64 52 65 51 52 42 58 63 59 51 72 65 51 72 36 38 24 45 55 66 34 75 26 35 57 37 56 76 55 27 31 46 60 57 34 66 53
```

2. Transforming Data :

RStudio interface showing data transformation code in the console. The code includes subsetting, creating new variables, and calculating statistical averages.

```
R413: ~
> a <- subset(ac, age_husband > age_wife)
> acs[4,3]
[1] 62
> a <- subset(ac, age_husband > age_wife)
> mean(ac$age_husband)
[1] 54.31776
> median(ac$age_husband)
[1] 55
> quantile(ac$age_wife)
```

3. Getting Statistical Averages from data :



4. Plotting Data :

