

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

The screenshot shows a Kaggle notebook titled 'notebook24f3ff955' with a 'Failed to save draft' warning. The notebook is in a 'Draft Session (39m)' and contains two code cells. The first cell, labeled [1]:, contains the code `df.info()`. The output of this code is displayed below the code cell, showing the structure of the Iris dataset. The second cell, labeled [22]:, contains the code `import matplotlib.pyplot as plt` and `matplotlib inline`. The right sidebar of the notebook shows the 'Data' section with 'irisdataset' and 'iris.data' listed. The 'Output' section shows '44.1MB / 19.6GB' and the file path '/kaggle/working'. The 'Settings' section is expanded, showing 'Schedule a notebook run' and 'Code Help'. The bottom of the screen shows a Windows taskbar with various application icons and a search bar.

notebook24f3ff955 Failed to save draft.

File Edit View Run Add-ons Help

Share Save Version 0

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

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

[1]:

```
df.info()
```

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

[22]:

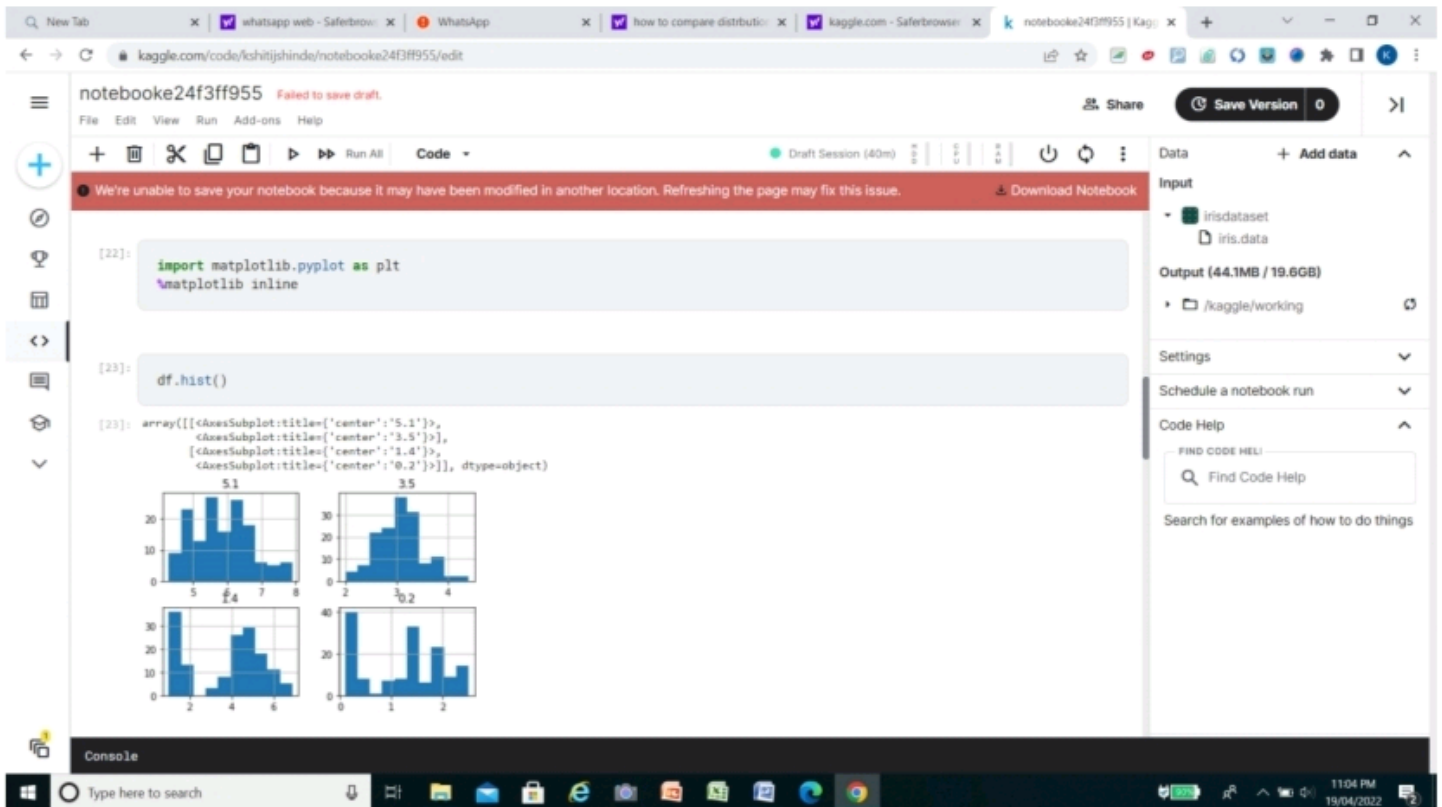
```
import matplotlib.pyplot as plt
matplotlib inline
```

Console

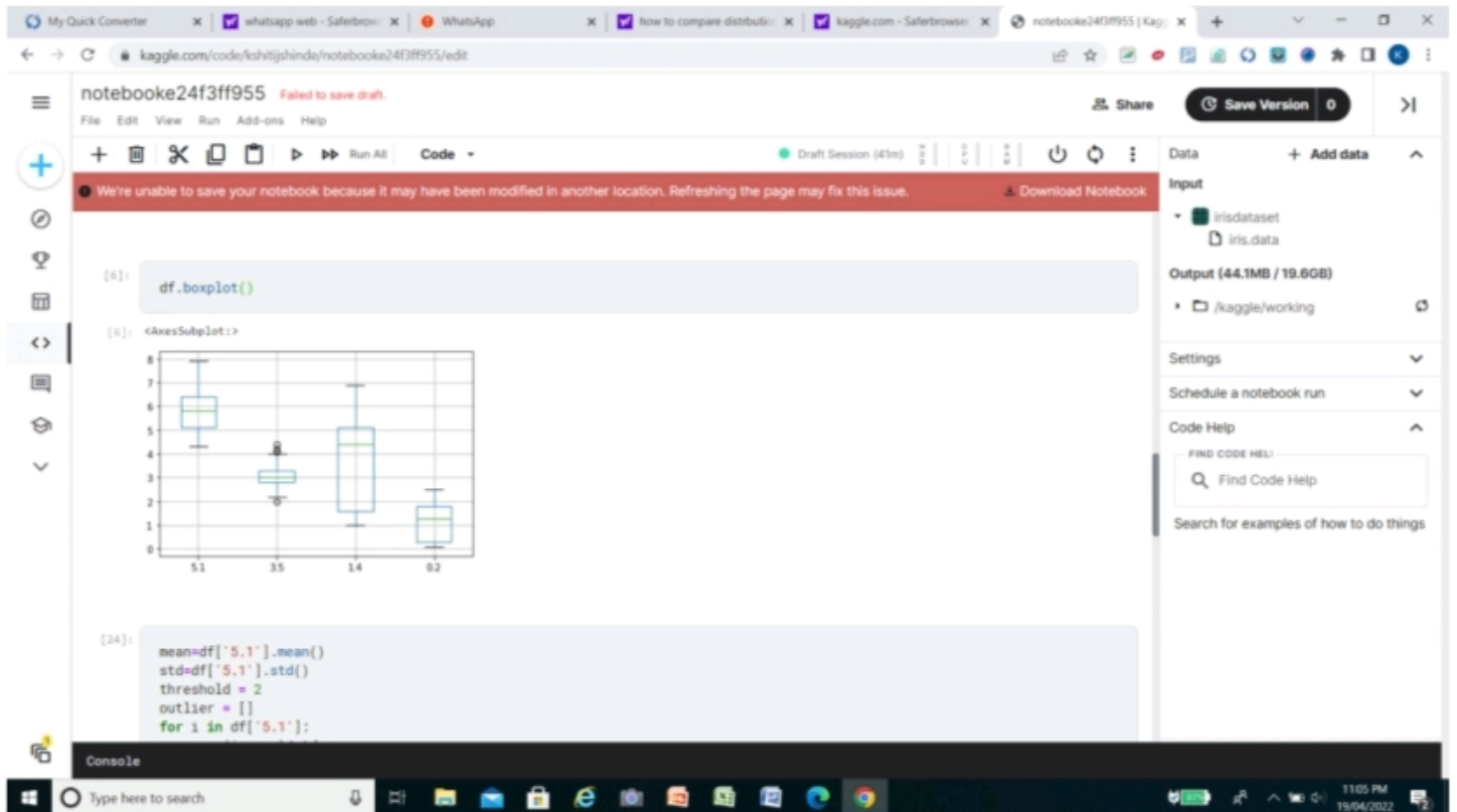
Type here to search

11:03 PM 19/04/2022

## 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.

Identify outliers :

The screenshot shows a Kaggle notebook titled 'notebook24f3ff955' with a 'Failed to save draft' warning. The notebook is in a 'Draft Session (41m)' and contains two code cells. The first cell identifies outliers in the '5.1' dataset, and the second cell identifies outliers in the '3.5' dataset. The output of the first cell is 'outlier in dataset is [7.6, 7.7, 7.7, 7.7, 7.9, 7.7]' and the output of the second cell is 'outlier in dataset is [4.0, 4.4, 4.1, 4.2]'. The right sidebar shows the 'Data' section with 'irisdataset' and 'iris.data' files, and the 'Output (44.1MB / 19.6GB)' section. The bottom of the screen shows a Windows taskbar with the time '11:05 PM' and date '19/04/2022'.

```
[24]: mean=df['5.1'].mean()
std=df['5.1'].std()
threshold = 2
outlier = []
for i in df['5.1']:
    z = (i-mean)/std
    if z > threshold:
        outlier.append(i)
print('outlier in dataset is', outlier)

outlier in dataset is [7.6, 7.7, 7.7, 7.7, 7.9, 7.7]
```

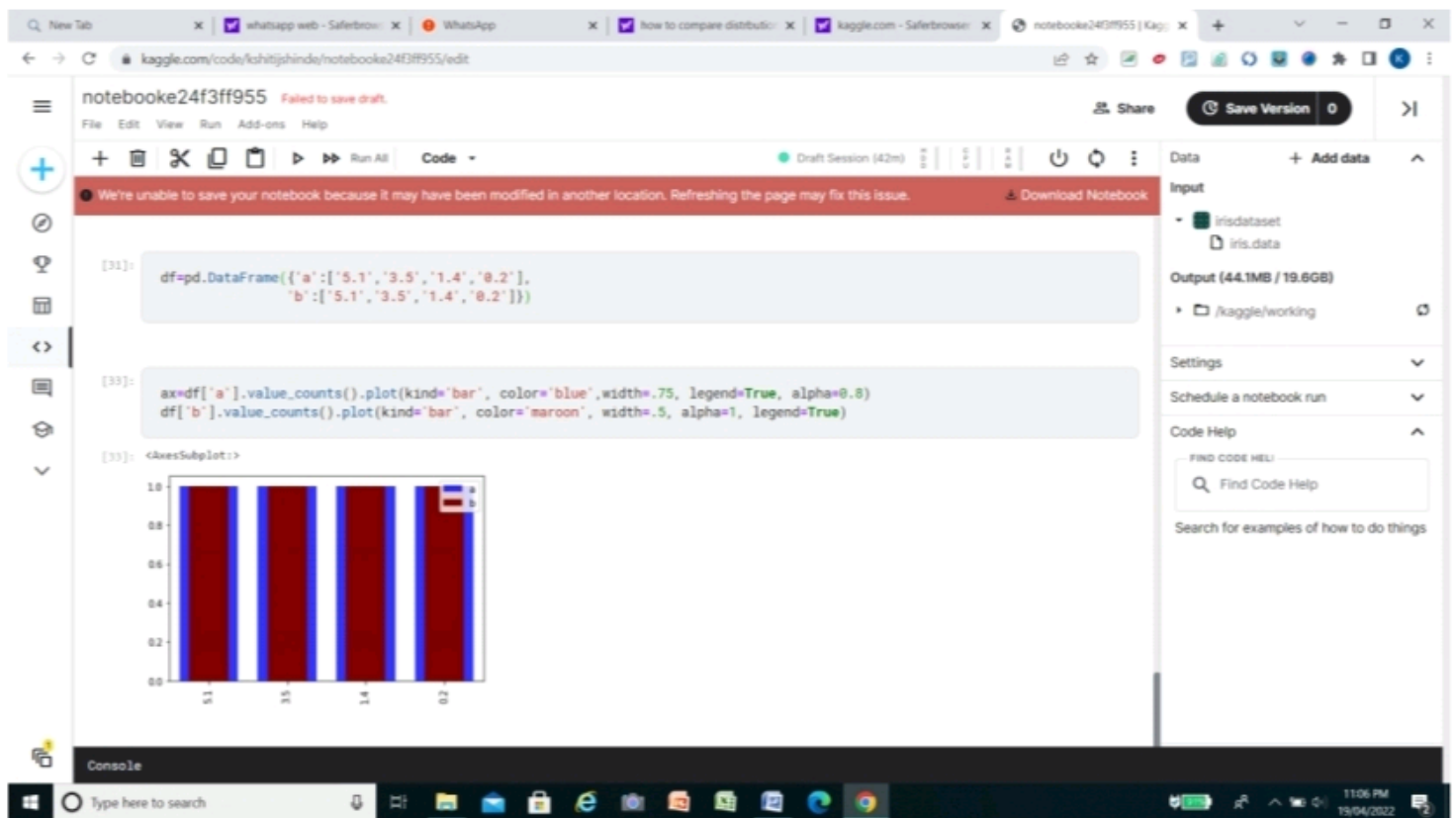
```
[25]: mean=df['3.5'].mean()
std=df['3.5'].std()
threshold = 2
outlier = []
for i in df['3.5']:
    z = (i-mean)/std
    if z > threshold:
        outlier.append(i)
print('outlier in dataset is', outlier)

outlier in dataset is [4.0, 4.4, 4.1, 4.2]
```

Console



## Comparing distributions :



# OUTPUT IN RSTUDIO :

## 1. Importing dataset :

The screenshot displays the RStudio interface with the following components:

- Environment Pane:** Shows two data objects: 'a' (5139 obs. of 14 variables) and 'acs' (7811 obs. of 14 variables).
- Console:** Contains the following R code and output:

```
R 3.3.3 - 64-bit  
> acs <- read.csv(url("http://stat511.cwick.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  
[38] 67 51 73 60 30 40 77 69 52 62 50 70 54 57 46 88 51 46 52 58 44 64 45 31 54 45 43 57 57 43 27 59 47 62 80 30 48 47  
[77] 78 55 37 29 59 48 29 34 29 41 38 62 74 43 65 24 34 58 48 60 86 59 28 65 47 35 56 58 42 43 50 67 86 43 51 37 33 36  
[115] 31 80 65 61 67 65 65 70 55 24 90 80 49 35 73 55 33 61 45 33 88 74 73 60 61 47 63 43 53 53 39 41 44 70 57 76 70 40  
[153] 62 49 55 23 24 23 38 35 63 41 72 54 71 56 30 71 49 29 55 33 46 45 68 34 62 59 71 64 59 51 42 74 74 64 61 63 57 46  
[191] 38 73 42 60 57 67 34 42 71 42 69 74 62 75 49 59 66 52 53 61 87 50 41 47 59 50 51 29 43 30 56 58 34 60 59 28 45 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 51 32 47 67 57 55 30 65 47 34 58  
[267] 74 66 41 40 21 48 56 49 52 51 53 42 38 75 95 35 63 46 66 50 72 47 64 71 56 71 39 62 73 48 39 39 72 51 60 70 44 59  
[305] 69 43 68 42 42 68 52 69 42 76 54 34 32 68 64 44 49 75 61 39 52 95 51 74 27 59 40 36 25 26 35 59 28 54 54 48 59  
[343] 49 41 64 68 29 71 38 58 43 71 58 24 57 46 54 33 45 42 31 80 81 84 44 60 66 40 77 56 49 79 50 61 37 49 55 69 65 41  
[381] 55 39 51 53 47 79 51 32 42 41 30 66 54 53 75 49 57 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 51 66 54 29 77 38 34 31 27 40 70 48 51 68 40 41 31 44 32 62 44 47 39 49 58 34 56 74 48 72 50 55  
[495] 69 75 69 63 37 46 43 58 47 66 88 51 62 51 48 47 36 53 60 84 57 45 74 62 66 38 35 63 57 34 57 52 70 64 40 58 50 43  
[533] 28 34 86 31 37 27 78 87 73 50 43 43 29 65 34 30 62 72 54 62 36 59 56 75 41 70 45 35 72 32 62 75 69 37 55 38 36 85  
[571] 64 53 51 68 78 55 62 29 27 65 30 51 54 58 24 62 58 47 72 47 46 63 86 34 51 34 52 62 55 58 74 56 62 71 52 73 29 25  
[609] 71 59 57 60 89 49 66 75 54 72 59 69 62 87 58 46 52 22 55 52 58 50 40 49 80 64 55 31 47 76 65 43 21 66 29 60 65 71  
[647] 56 63 59 55 53 26 44 63 76 36 33 51 38 42 70 75 32 62 39 64 63 55 59 40 43 32 37 42 66 66 77 85 49 84 84 28 49 78  
[685] 57 64 65 37 49 28 71 31 49 42 45 77 39 34 41 43 39 62 46 48 53 68 69 62 47 73 62 61 46 57 66 73 37 43 33 30 63 63  
[723] 64 52 65 51 52 62 42 58 63 59 51 72 45 51 72 36 38 24 45 55 66 34 75 26 35 57 57 56 76 55 27 51 48 60 57 34 66 53
```

## 2. Transforming Data :

The screenshot displays the RStudio interface with the following components:

- Environment Panel:** Shows two datasets: `DATA` (5139 obs. of 14 variables) and `ACS` (7811 obs. of 14 variables).
- Console:** Contains the following R code and its output:

```
R> RAI3 ~-~  
[131] 28 34 66 35 37 27 78 87 73 50 43 43 29 65 34 30 62 72 54 62 36 39 56 75 41 70 61 35 72 32 62 75 69 37 55 38 36 85  
[171] 64 33 33 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 22 55 52 58 50 40 49 80 64 35 32 47 78 65 43 21 66 29 60 65 71  
[647] 56 63 59 55 13 26 44 63 76 36 33 51 38 42 70 75 32 63 39 64 63 55 59 40 43 32 37 42 66 66 77 85 49 84 64 28 49 78  
[685] 57 64 63 37 49 29 71 32 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 62 62 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 48 60 57 34 66 53  
[761] 37 36 34 68 43 44 76 40 67 30 60 44 33 71 56 54 82 62 54 49 63 40 62 41 65 56 70 95 34 35 58 57 85 60 63 71 61 28  
[799] 75 95 25 74 39 33 31 37 51 46 44 41 77 51 75 47 43 30 44 66 57 56 50 47 65 46 50 72 57 52 89 38 42 73 30 25 46 78  
[837] 52 52 47 33 26 75 49 47 38 67 35 35 44 52 47 72 49 42 64 68 68 62 34 35 61 70 66 74 60 62 66 58 56 31 30 68 60 50  
[875] 57 71 58 37 69 57 84 88 48 60 62 43 67 50 86 62 61 59 63 54 48 31 71 50 59 70 53 50 84 71 81 43 54 50 56 76 56 58  
[913] 53 68 48 60 38 36 74 65 51 74 56 43 27 72 39 42 69 63 78 49 57 45 77 54 34 37 38 39 73 57 62 39 65 75 55 50 71 55  
[951] 71 58 70 40 25 35 30 54 29 41 50 32 84 55 62 67 43 55 29 23 53 75 62 54 59 62 44 45 86 55 57 53 43 76 57 49 43 54  
[989] 40 35 72 53 74 29 46 58 67 36 49 43  
[reached getOption("max.print") -- omitted 6811 entries ]  
> a <- subset(acs, age_husband > age_wife)  
> acs[1,2]  
[1] 62  
> a <- subset(acs, age_husband > age_wife)  
> mean(acs$age_husband)  
[1] 34.32776  
> median(acs$age_husband)  
[1] 35  
> quantile(acs$age_wife)
```



### 3. Getting Statistical Averages from data :

The screenshot displays the RStudio interface with the following components:

- Environment Pane:** Shows two data objects: 'a' (5139 obs. of 14 variables) and 'acs' (7811 obs. of 14 variables).
- Console:** Contains the following R commands and their outputs:

```
> quantile(acs$age_wife)
[1] 0% 25% 50% 75% 100%
[1] 19 40 51 63 95

> var(acs$age_wife)
[1] 220.527

> sd(acs$age_wife)
[1] 14.85015

> summary(acs)
household      age_husband    age_wife    income_husband    income_wife    bedrooms    electricity    gas    number_children    internet    mode    own
min. : 48      min. : 17.00    min. : 19.00    min. : -63000    min. : -63000    min. : 0.000    min. : 0.000    min. : 4.0
1st Qu.: 389236  1st Qu.: 42.00    1st Qu.: 40.00    1st Qu.: 24000    1st Qu.: 6225    1st Qu.: 3.000    1st Qu.: 80.0
Median : 764131  Median : 55.00    Median : 53.00    Median : 43000    Median : 18110  Median : 3.000    Median : 110.0
Mean : 759565   Mean : 54.32     Mean : 52.08     Mean : 59831    Mean : 28985    Mean : 3.117    Mean : 131.9
3rd Qu.: 1137444 3rd Qu.: 66.00    3rd Qu.: 63.00    3rd Qu.: 70500    3rd Qu.: 39600    3rd Qu.: 4.000    3rd Qu.: 1460.0
Max. : 1492278   Max. : 95.00     Max. : 95.00     Max. : 1756000    Max. : 1421000  Max. : 10.000    Max. : 1500.0

gas    number_children    internet    mode    own    language
min. : 3.00    min. : 0.0000    Length:7811    Length:7811    Length:7811    Length:7811
1st Qu.: 3.00    1st Qu.: 0.0000    Class :character    Class :character    Class :character    Class :character
Median : 20.00    Median : 0.0000    Mode :character    Mode :character    Mode :character    Mode :character
Mean : 45.09     Mean : 0.6855
3rd Qu.: 70.00    3rd Qu.: 1.0000
Max. : 190.00    Max. : 12.0000
```

#### 4. Plotting Data :

