120A3051
Shreya Idate
Batch: E3

EXPERIMENT NO. 3

Aim: Data Visualization / Exploratory Data Analysis for the selected data set using Matplotlib.

a) Create a bar graph, contingency table using any 2 variables.

b) Create normalized histogram.

c) Describe these graphs and table indicates.

Theory:

Matplotlib is one of the most widely used data visualization libraries in Python. From simple to complex visualizations, it's the go-to library for most. It allows you to create every type of chart with a great level of customization. This page provides some general tips that can be applied on any kind of chart made with matplotlib like customizing titles or colors. Once installed, matplotlib must be imported, usually using import matplotlib.pyplot as plt. Then one can use the functions available in the plt object.

Data Visualization:

Data visualization is the graphical representation of information and data. The data visualization is one of the most important fundamental toolkits of a data scientist. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Visualizations are the easiest way to analyze and absorb information. Visuals help to easily understand the complex problem. They help in identifying patterns, relationships, and outliers in data. It helps in understanding business problems better and quickly. It helps to build a compelling story based on visuals. Insights gathered from the visuals help in building strategies for businesses. It is also a precursor to many high-level data analysis for Exploratory Data Analysis(EDA) and Machine Learning(ML).

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected.

EDA is an important first step in any data analysis. Understanding where outliers occur and how variables are related can help one design statistical analyses that yield meaningful results. In biological monitoring data, sites are likely to be affected by multiple stressors.

Thus, initial explorations of stressor correlations are critical before one attempt to relate stressor variables to biological response variables. EDA can provide insights into candidate causes that should be included in a causal assessment.

Bar Graph

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axes of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

Contingency Table

Contingency Table is one of the techniques for exploring two or more variables. It is basically a tally of counts between two or more categorical variables. Contingency Tables give clear correlation values between those variables, thus making it much more useful to understand the data for further information extraction.

Normalized Histogram

Normalized Histogram: A histogram is a frequency distribution that depicts the frequencies of different elements in a dataset. This graph is generally used to study frequencies and determine how the values are distributed in a dataset. Normalization of histogram refers to mapping the frequencies of a dataset between the range [0, 1] both inclusive.

Programs & Outputs:

Creating table

In [3]:
```
import pandas as pd
```

In [4]:
```
df = pd.DataFrame({'Order': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
                             11, 12, 13, 14, 15, 16, 17, 18, 19, 20],
                   'Product': ['TV', 'TV', 'Comp', 'TV', 'TV', 'Comp',
                               'Comp', 'Comp', 'TV', 'Radio', 'TV', 'Radio', 'Radio',
                               'Radio', 'Comp', 'Comp', 'TV', 'TV', 'Radio', 'TV'],
                   'Country': ['A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B', 'B',
                               'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C']})

df
```

Out[4]:

	Country	Order	Product
0	A	1	TV
1	A	2	TV
2	A	3	Comp
3	A	4	TV
4	B	5	TV
5	B	6	Comp
6	B	7	Comp
7	B	8	Comp
8	B	9	TV
9	B	10	Radio
10	B	11	TV
11	B	12	Radio
12	C	13	Radio
13	C	14	Radio
14	C	15	Comp
15	C	16	Comp
16	C	17	TV
17	C	18	TV
18	C	19	Radio
19	C	20	TV

Creating contingency table:

Without margins:

In [6]:
```
# creating contingency table

ct = pd.crosstab(index=df['Country'], columns=df['Product'])
ct
```

Out[6]:

Product	Comp	Radio	TV
Country			
A	1	0	3
B	3	2	3
C	2	3	3

With margins:

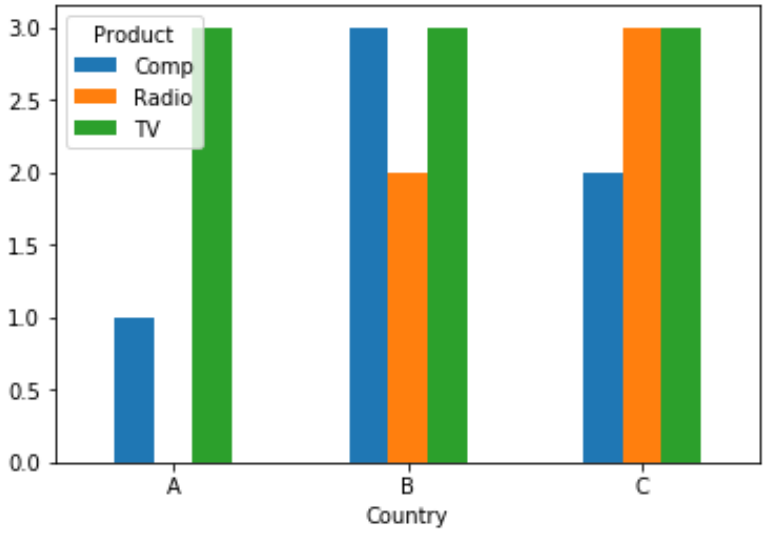
```
In [7]: # contingency table with margins

ct = pd.crosstab(index=df['Country'], columns=df['Product'],margins=True)
ct
```

Out[7]:

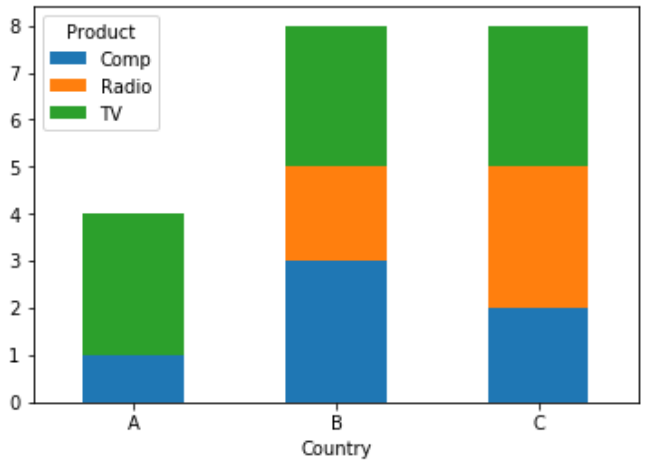
Country	Comp	Radio	TV	All
A	1	0	3	4
B	3	2	3	8
C	2	3	3	8
All	6	5	9	20

Plotting graph of the contingency table:

```
In [18]: # plotting bar graph
%matplotlib inline

bargraph = ct.plot.bar(rot=0)
```



In [19]:
```
# plotting stacked bar graph

stacked_plot = ct.plot(kind="bar",stacked=True, rot=0)
```

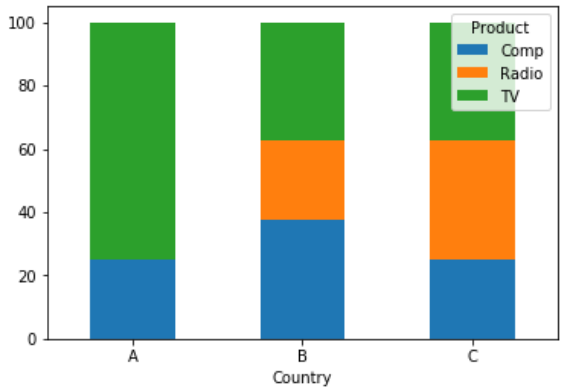


In [20]:
```
# Normalized table - by index

normalized_ct= pd.crosstab(df['Country'],
            df['Product'],
            dropna = False,
            normalize = 'index' # convert absolute to row proportions
           ).round(3)*100

normalized_ct
```

Out[20]:

Product	Comp	Radio	TV
Country			
A	25.0	0.0	75.0
B	37.5	25.0	37.5
C	25.0	37.5	37.5

In [22]:
```
normalized_graph = normalized_ct.plot(kind="bar", stacked=True, rot=0)
```

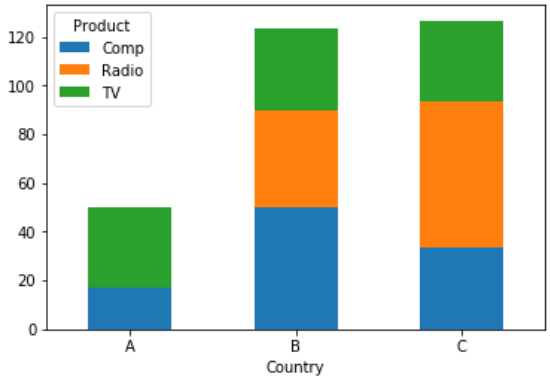


In [23]:
```
# Normalized table - by column

normalized_ct2 = pd.crosstab(df['Country'],
            df['Product'],
            dropna = False,
            normalize = 'columns' # convert absolute to columns proportions
           ).round(3)*100

normalized_ct2
```

Out[23]:

Country	Comp	Radio	TV
A	16.7	0.0	33.3
B	50.0	40.0	33.3
C	33.3	60.0	33.3

In [24]:
```
normalized_graph2 = normalized_ct2.plot(kind="bar", stacked=True, rot=0)
```



Conclusion:

Successfully performed Data Visualization / Exploratory Data Analysis for the selected data set using Matplotlib.