Name: Rushikesh Jyoti

Divison: A

Roll no: 27

SRN: 201901139

Question

- **Read the dataset "protein.csv" that is provided to you.**
- **Build a suitable clustering model using R/Python based on k-means clustering approach.**
- **Plot the clusters and show how the model varies with different values of k.**
- **Develop some metrics to determine the accuracy of your clustering model**

In [202]:

```
import pandas as pd
import seaborn as sb
from matplotlib import pyplot as plot
import random

from sklearn import metrics
from sklearn.preprocessing import LabelEncoder
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
```

In [203]:

```
df = pd.read_csv("./protein.csv")

df.head()
```

Out[203]:

	Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg
0	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
1	Austria	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
2	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
3	Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
4	Czechoslovakia	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0

Dataset Splitting

In [204]:

```
xtrain, xtest, ytrain, ytest = train_test_split(df.drop(["Country"], axis=1), df["Countr
y"], train_size=0.8)

print(xtrain.shape, xtest.shape, ytrain.shape, ytest.shape)
```

```
(20, 9) (5, 9) (20,) (5,)
```

Clustering Model (k=3)

Clustering Model (k=3)

In [205]:

```
model = KMeans(3)
model.fit(df.drop(["Country"], axis=1), df["Country"])

model
```

Out[205]:

```
KMeans(n_clusters=3)
```

In [206]:

```
colors = list(sb.colors.crayons.values())
colors = random.sample(colors, k=model.cluster_centers_.shape[0])
colors
```

Out[206]:

```
['#FD7C6E', '#CD9575', '#1A4876']
```

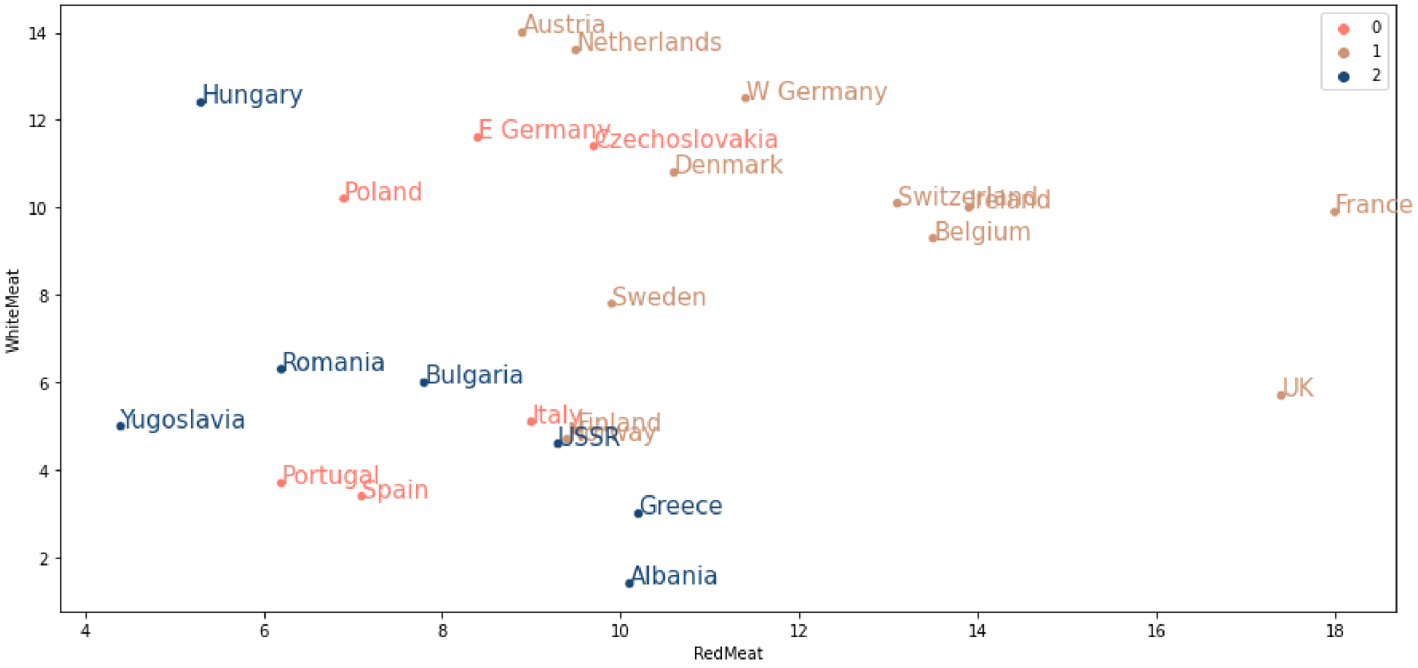
Plotting the scatter plot

In [207]:

```
plot.figure(1, (15, 7))
sb.scatterplot(x='RedMeat', y='WhiteMeat', hue=model.labels_, data=df, legend="full", pa
lette=colors)

for i in range(df.shape[0]):
    plot.text(x=df['RedMeat'][i], y=df['WhiteMeat'][i], s=df['Country'][i], fontdict={'s
ize': 15,
    'color': colors[model.predict(
        df.drop(['Country'], axis=1))[i]
    ]})
    pass

plot.show()
```



Clustering Model (k=7)

In [208]:

```
model = KMeans(7)
```

```
model.fit(df.drop(["Country"], axis=1), df["Country"])

model
```

Out[208]:

```
KMeans(n_clusters=7)
```

In [209]:

```
colors = list(sb.colors.crayons.values())
colors = random.sample(colors, k=model.cluster_centers_.shape[0])
colors
```

Out[209]:

```
['#E6A8D7', '#FAA76C', '#F0E891', '#DEAA88', '#6DAE81', '#FCE883', '#FD7C6E']
```

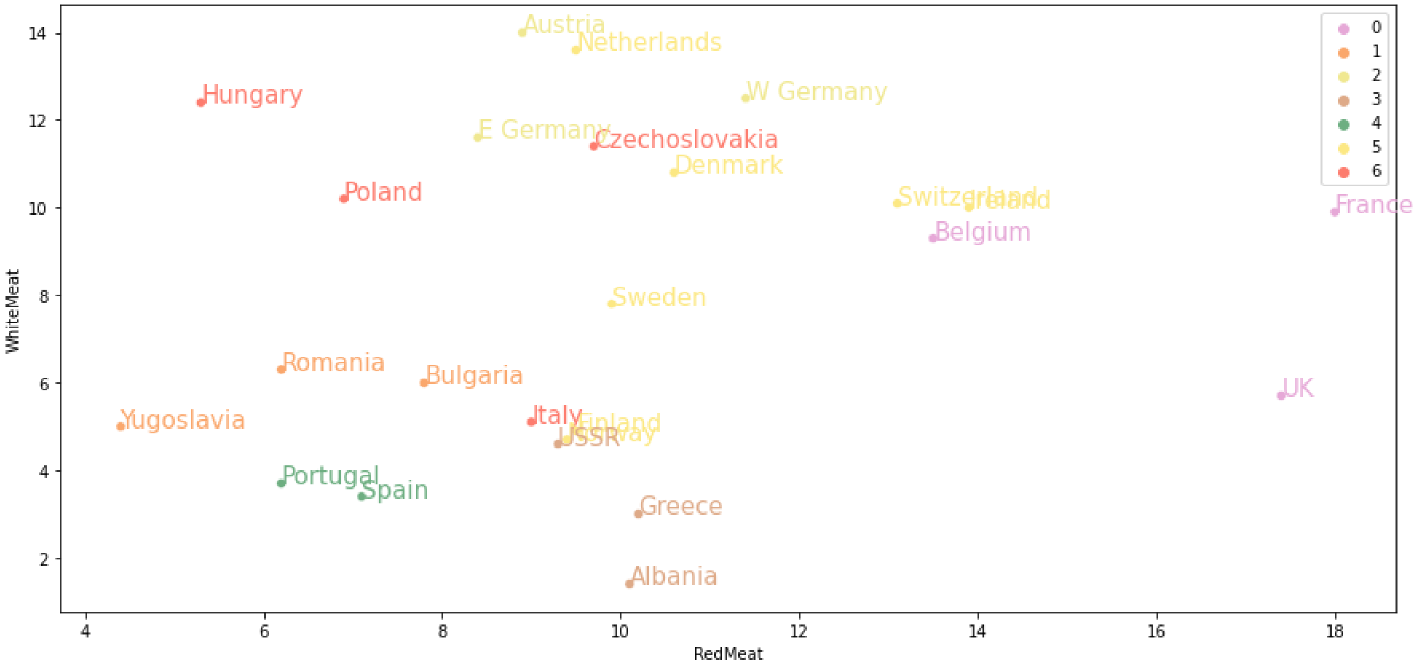
Plotting

In [210]:

```
plot.figure(1, (15, 7))
sb.scatterplot(x='RedMeat', y='WhiteMeat', hue=model.labels_, data=df, legend="full", pa
lette=colors)

for i in range(df.shape[0]):
    plot.text(x=df['RedMeat'][i], y=df['WhiteMeat'][i], s=df['Country'][i], fontdict={'s
ize': 15,
    'color': colors[model.predict(
        df.drop(['Country'], axis=1))[i]
    ]})
    pass

plot.show()
```


Austria
Netherlands
Hungary
W Germany
E Germany
Czechoslovakia
Denmark
Poland
Switzerland
Ireland
France
Belgium
Sweden
Romania
Bulgaria
UK
Yugoslavia
Italy
Finland
USSR
Norway
Portugal
Spain
Greece
Albania
● 0
● 1
● 2
● 3
● 4
● 5
● 6
WhiteMeat
RedMeat