Name : Samriddhi Raskar

Roll no.: 226523

Introduction to K-Means

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

We look at the data and then try to club similar observations and form different groups. Hence it is an unsupervised learning problem.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters

DATASET : MUSIC AND MENTAL HEALTH SURVEY RESULTS

Context

Music therapy, or MT, is the use of music to improve an individual's stress, mood, and overall mental health. MT is also recognized as an evidence-based practice, using music as a catalyst for "happy" hormones such as oxytocin.

However, MT employs a wide range of different genres, varying from one organization to the next.

The MxMH dataset aims to identify what, if any, correlations exist between an individual's music taste and their self-reported mental health. Ideally, these findings could contribute to a more informed application of MT or simply provide interesting sights about the mind.

Interpreting data

Background: Respondents answer generic questions focused on musical background and listening habits.

Music genres: Respondents rank how often they listen to 16 music genres, where they can select:

Never, Rarely, Sometimes, Very frequently.

Mental health: Respondents rank Anxiety, Depression, Insomnia, and OCD on a scale of 0 to 10, where:

0 - I do not experience this. 10 - I experience this regularly, constantly/or to an extreme. Additional data that does not fall in these blocks may provide useful background information. See column descriptors.

In [1]:

```
import pandas as pd
import warnings as w
w.filterwarnings('ignore')
```

In [2]:

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

In [3]:

```
df=pd.read_csv('E:/Users/Samu/mxmh_survey_results.csv')
```

In [4]:

```
df.head()
```

Out[4]:

	Timestamp	Age	Primary streaming service	Hours per day	While working	Instrumentalist	Composer	Fav genre	Exploratory	Foreign languages	...	Frequency [R&B]	Frequency [Rap]	Frequenc [Roc
0	8/27/2022 19:29:02	18.0	Spotify	3.0	Yes	Yes	Yes	Latin	Yes	Yes	...	Sometimes	Very frequently	Nev
1	8/27/2022 19:57:31	63.0	Pandora	1.5	Yes	No	No	Rock	Yes	No	...	Sometimes	Rarely	Ve frequent
2	8/27/2022 21:28:18	18.0	Spotify	4.0	No	No	No	Video game music	No	Yes	...	Never	Rarely	Rare
3	8/27/2022 21:40:40	61.0	YouTube Music	2.5	Yes	No	Yes	Jazz	Yes	Yes	...	Sometimes	Never	Nev
4	8/27/2022 21:54:47	18.0	Spotify	4.0	Yes	No	No	R&B	Yes	No	...	Very frequently	Very frequently	Nev

5 rows × 33 columns

In [5]:

```
data=df[['Age','Hours per day']]
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 736 entries, 0 to 735
Data columns (total 2 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Age            735 non-null    float64
 1   Hours per day  736 non-null    float64
dtypes: float64(2)
memory usage: 11.6 KB
```

In [6]:

```
x=df[['Age','Hours per day']]
x
```

Out[6]:

	Age	Hours per day
0	18.0	3.0
1	63.0	1.5
2	18.0	4.0
3	61.0	2.5
4	18.0	4.0
...	...	...
731	17.0	2.0
732	18.0	1.0
733	19.0	6.0
734	19.0	5.0
735	29.0	2.0

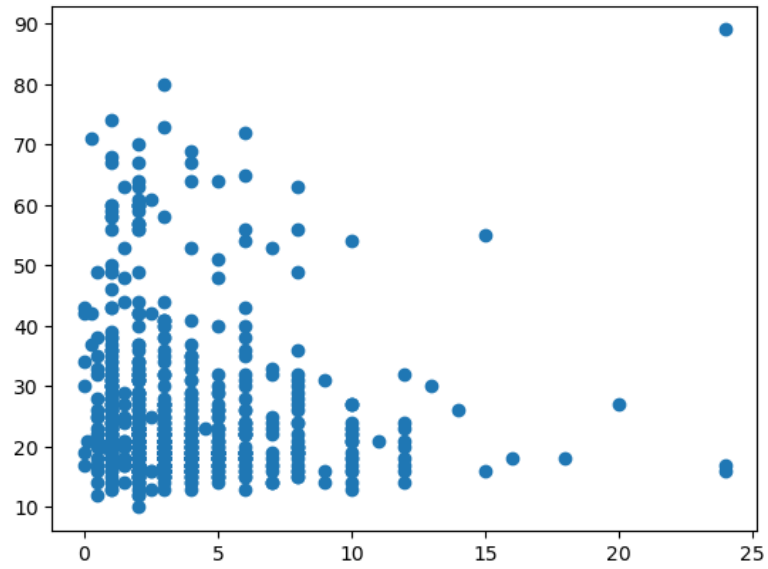
736 rows × 2 columns

In [7]:

```
plt.scatter(data['Hours per day'],data['Age'])
```

Out[7]:

```
<matplotlib.collections.PathCollection at 0x2122ee0bc10>
```



In [8]:

```
from sklearn.cluster import KMeans
```

In [9]:

```
#filtering data by removing null values and outliers
data=data.dropna(axis=0,how='any')
max(data['Hours per day'])
max(data['Age'])
data=data[data['Age']<85]
data=data[data['Hours per day']<23]
data.info()
```

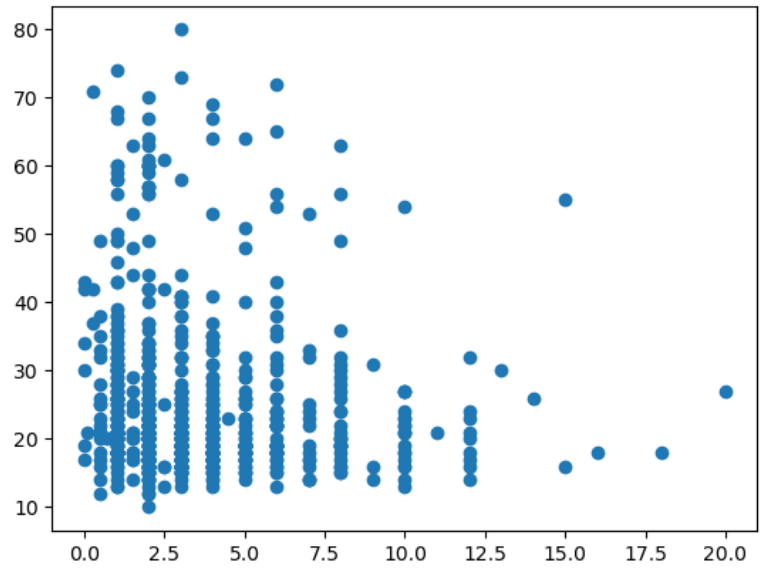
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 732 entries, 0 to 735
Data columns (total 2 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Age            732 non-null    float64
 1   Hours per day  732 non-null    float64
dtypes: float64(2)
memory usage: 17.2 KB
```

In [10]:

```
plt.scatter(data['Hours per day'],data['Age'])
```

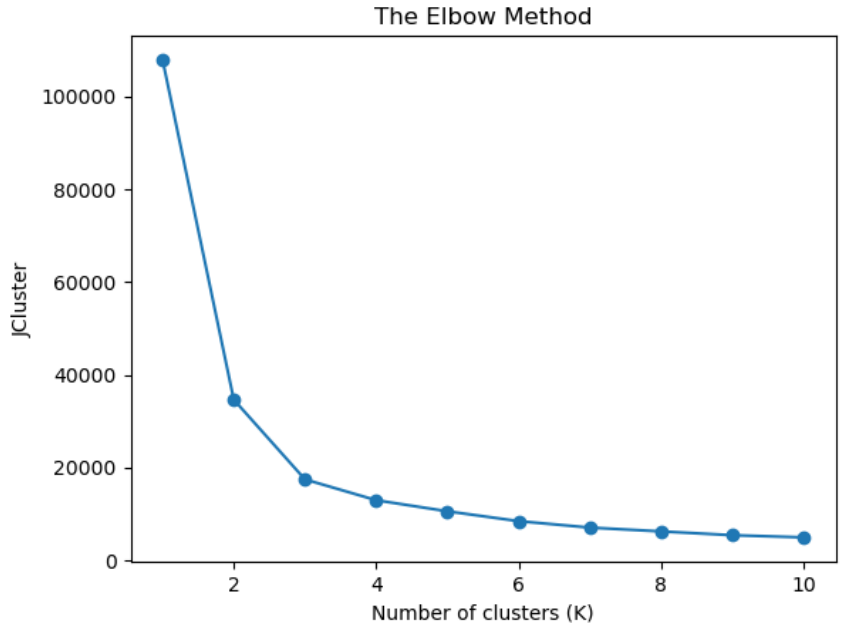
Out[10]:

```
<matplotlib.collections.PathCollection at 0x21231f0a310>
```



In [11]:

```
jcluster=[]
for i in range(1,11):
    kmeans = KMeans(n_clusters=i,init='k-means++',n_init=10)
    kmeans.fit(data)
    jcluster.append(kmeans.inertia_)
plt.plot(range(1, 11), jcluster, marker='o')
plt.title('The Elbow Method')
plt.xlabel('Number of clusters (K)')
plt.ylabel('JCluster')
plt.show()
```



In [12]:

```
model=KMeans(n_clusters=3)
model.fit(data)
```

Out[12]:

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

In [13]:

```
model.cluster_centers_
```

Out[13]:

```
array([[32.05294118,  3.22058824],
       [59.25454545,  3.14090909],
       [19.12623274,  3.61794872]])
```

In [14]:

```
cluster_number = model.predict(data)
```

In [15]:

```
len(cluster_number)
```

Out[15]:

```
732
```

In [16]:

```
len(data)
```

Out[16]:

```
732
```

In [17]:

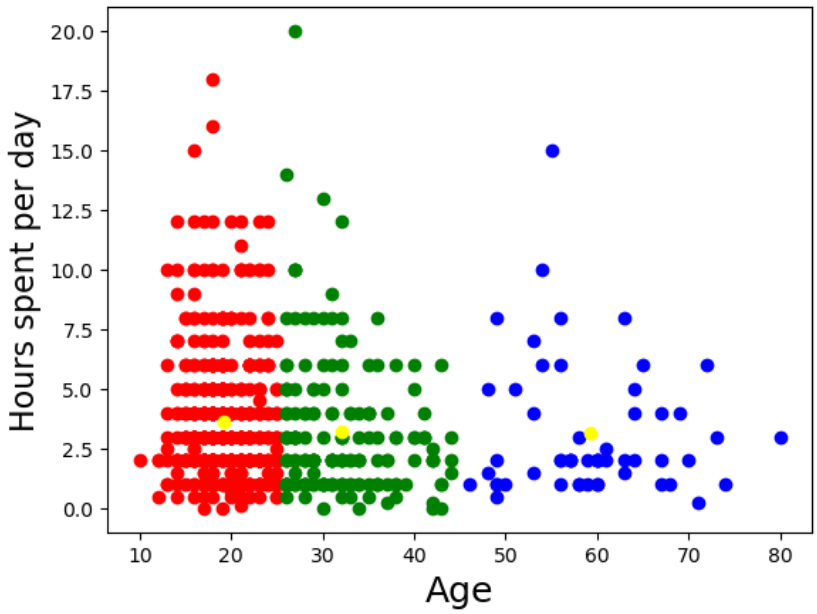
```
c0=data[cluster_number==0]
c1=data[cluster_number==1]
c2=data[cluster_number==2]
```

In [19]:

```
plt.scatter(c0['Age'],c0['Hours per day'],c='red')
plt.scatter(c1['Age'],c1['Hours per day'],c='blue')
plt.scatter(c2['Age'],c2['Hours per day'],c='green')
plt.scatter(model.cluster_centers_[:,0],model.cluster_centers_[:,1],c='yellow')
plt.xlabel('Age', fontsize=18)
plt.ylabel('Hours spent per day', fontsize=16)
```

Out[19]:

```
Text(0, 0.5, 'Hours spent per day')
```



Conclusion

By this data we understood that companies will get to know their target audience.

They can push their offers and premium plans to users who listen to the music for higher time OR they can introduce different category plans to different age groups.

From health point of view, company should introduce a prompt for listners who listen to the music for more than 5 hours per day to lower their volume below 50% As we know that our generation is at high risk of getting deaf due to the same issue.

The company should make a daily mix playlist for a person who frequently listens to a particular type of genre music.