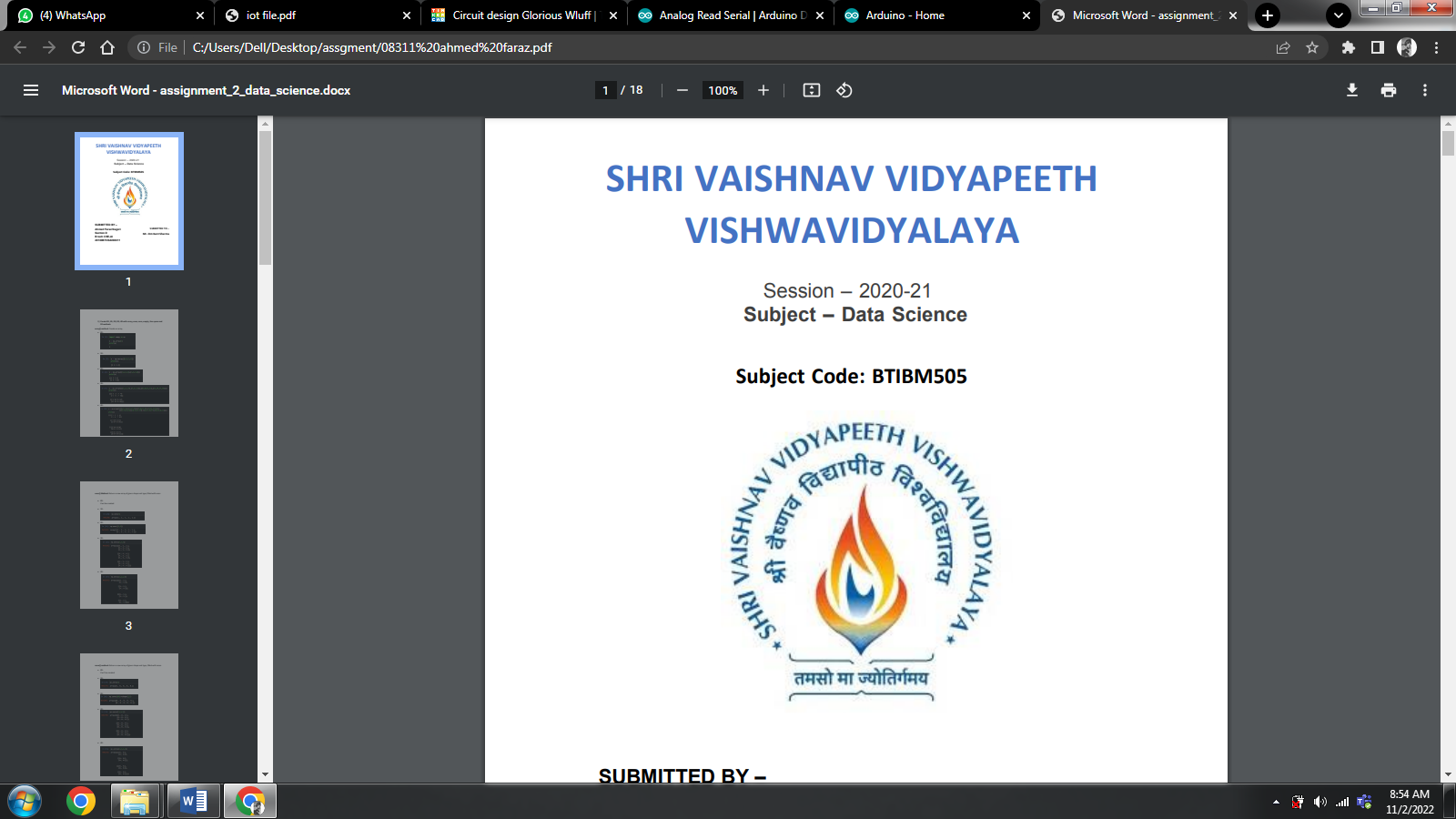
**Shri Vaishnav Vidyaapeeth Vishwavidyalaya, Indore**

**Department of Computer Science and Engineering**



Session 2021-22

3rd Year

V Semester

Subject: Introdction to Data Science

Subject Code: BTIBM505

**Major Project Code File**

**Prepared by:- Submitted to:-**

Harshita Pandit Om Kant Sharma

Ahmed Faraz Nagori

Mohammd Nadeem Khan

**CODE:-**

* **Importing the library**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

* **Importing function from cluster model in sklearn library**

from sklearn.cluster import KMeans **#for forming clusters of customer**

* **Reading the Dataset**

customerData=pd.read\_csv("C:\\Users\\Harshita\\Desktop\\customers.csv") **#from csv file "customers.csv"**

* **Visulazing the Dataset**

**#Bar Graph(Gender vs No of Customers)**

genders=customerData.Gender.value\_counts()

plt.figure(figsize=(7,7))

sns.barplot(x=genders.index, y=genders.values)

plt.show()

**#Bar Graph(Age vs No of Customer)**

df=customerData

age18\_25 = df.Age[(df.Age<=25)&(df.Age>=18)]

age26\_35 = df.Age[(df.Age<=35)&(df.Age>=26)]

age36\_45 = df.Age[(df.Age<=45)&(df.Age>=36)]

age46\_55 = df.Age[(df.Age<=55)&(df.Age>=46)]

age55above = df.Age[(df.Age>=56)]

x=["18-25","26-35","36-45","46-55","Above55"]

y=[len(age18\_25.values), len(age26\_35.values),len(age36\_45.values),len(age46\_55.values),len(age55above.values)]

plt.figure(figsize=(15,6))

plt.title=("Number of customers and ages")

plt.xlabel=("Ages")

plt.ylabel=("Number of customers")

sns.barplot(x=x,y=y)

plt.show()

**#Bar graph(Spending Score vs No of Customers)**

ss1\_20= df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=1) &(df["Spending Score (1-100)"]<=20)]

ss21\_40= df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=21) &(df["Spending Score (1-100)"]<=40)]

ss41\_60= df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=41) &(df["Spending Score (1-100)"]<=60)]

ss61\_80= df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=61) &(df["Spending Score (1-100)"]<=80)]

ss81\_100= df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=81) &(df["Spending Score (1-100)"]<=100)]

x=["1-20","21-40","41-60","61-80","81-100"]

y=[len(ss1\_20.values), len(ss21\_40.values),len(ss41\_60.values),len(ss61\_80.values),len(ss81\_100.values)]

sns.barplot(x=x , y=y)

plt.figure(figsize=(10,20))

plt.title=("Spending scores of the customers")

plt.xlabel=("Spending Scores")

plt.ylabel=("score of customers")

plt.show()

**#Bar Graph (Annual Income vs No of Customers)**

ai0\_30 = df["Annual Income (k$)"][(df["Annual Income (k$)"]>=0)&(df["Annual Income (k$)"]<=30)]

ai31\_60 = df["Annual Income (k$)"][(df["Annual Income (k$)"]>=31)&(df["Annual Income (k$)"]<=60)]

ai61\_90 = df["Annual Income (k$)"][(df["Annual Income (k$)"]>=61)&(df["Annual Income (k$)"]<=90)]

ai91\_120 = df["Annual Income (k$)"][(df["Annual Income (k$)"]>=91)&(df["Annual Income (k$)"]<=120)]

ai121\_150 = df["Annual Income (k$)"][(df["Annual Income (k$)"]>=121)&(df["Annual Income (k$)"]<=150)]

x=["0-30","31-60", "61-90","91-120","121-150"]

y=[len(ai0\_30.values), len(ai31\_60.values), len(ai61\_90.values),len(ai91\_120.values), len(ai121\_150.values)]

plt.figure(figsize=(15,6))

sns.barplot(x=x,y=y,)

plt.title=("Annual Income of customers")

plt.xlabel("Annual Income in k$ ")

plt.ylabel("Number of customers")

plt.show()

* **Analysing Dataset**

**#printing first 5 rows of teh Data Frame**

customerData.head(8)

**#to get the number of rows and columns from the data**

customerData.shape

**#Information about the dataset**

customerData.info()

**#Checking for the mission values in Dataset**

customerData.isnull().sum()

x=customerData.iloc[ : , [3,4] ].values

* **Finding no. of Clusters using ELBOW method**

**#choosing number of clusters**

L = []

for i in range(1,11):

km=KMeans(n\_clusters=i,init='k-means++',random\_state=40 )

km.fit(x)

L.append(km.inertia\_)

* **Plotting the elbow graph.....**

sns.set()

plt.plot(range(1,11),L)

plt.title('Elbow Graph')

plt.xlabel('Number of clusters')

plt.ylabel('Sum of Square')

plt.show()

**#Optimum number of cluster will be 5**

**#Training the Model(K Means Clustering Model)**

km=KMeans(n\_clusters=5,init='k-means++' ,random\_state=0)

y=km.fit\_predict(x)

y

* **Visualization**

**#Visualizing all the clusters**

​

plt.figure(figsize=(8,8))

plt.scatter(x[y==0,0], x[y==0,1], s=50, c='red', label='1')

plt.scatter(x[y==1,0], x[y==1,1], s=50, c='yellow', label='2')

plt.scatter(x[y==2,0], x[y==2,1], s=50, c='green', label='3')

plt.scatter(x[y==3,0], x[y==3,1], s=50, c='black', label='4')

plt.scatter(x[y==4,0], x[y==4,1], s=50, c='blue', label='5')

​

plt.scatter(km.cluster\_centers\_[ : ,0],km.cluster\_centers\_[ : ,1], s=100, c='cyan', label='Centroids' )

​

plt.title('Customer Groups')

plt.xlabel('Annual Income')

plt.ylabel('Spending Score')

plt.show()

​