import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns

studperf\_df=pd.read\_csv('StudentsPerformance.csv')

studperf\_df.info()

studperf\_df.info(verbose=False) # short summary

studperf\_df.describe()

print(studperf\_df.head(15))

#Remove unnecessary features (E.g. drop unwanted columns) from the dataset such as ‘lunch’ and ‘test preparation course’

stuperf\_df = studperf\_df.drop(['lunch'], axis=1,inplace = False)

print('\\n====Understanding Inplace False : The Copied Dataframe====')

print(stuperf\_df.head(3))

#Manipulate data by replacing empty column values in ‘parental level of education’ with a default value.

stuperf\_df["parental level of education"] =stuperf\_df["parental level of education"].fillna("Not applicable")

print(stuperf\_df.head(5))

print(stuperf\_df.head(5))

#Convert the attribute ‘race/ethnicity’ to have ‘groupA’ to be ‘Asian Students’, ‘groupB’ to be ‘African Students’ , ‘groupC’ to be ‘Afro-Asian #Students’, ‘groupD’ to be ‘American Students’ and ‘groupE’ to be ‘European Students’

stuperf\_df["race/ethnicity"]=stuperf\_df["race/ethnicity"].map({"group A" : "Asian students",

"group B" : "African students",

"group C" : "Afro-Asian students",

"group D" : "American students",

"group E" : "European students"})

print(stuperf\_df.head(2))

#Tally of the Number of Male & Female students who took up the ‘test preparation course’ and those who did not.

ax = sns.countplot(x="test preparation course",hue='gender',palette='Set3',data=stuperf\_df)

ax.set(title="Course completion based on gender", xlabel='Course', ylabel='Total')

plt.show()

#Total Number of Male & Female Students belonging to each student group

ax = sns.countplot(x="race/ethnicity",hue="gender",palette="Set2",data=stuperf\_df)

ax.set(title="Total number of male and female students belonging to each group", xlabel="Groups", ylabel="Total")

plt.show()

#No of students who ‘failed’(less than 40), ‘second class’(between 40 & 50).

#'first class’(between 60 & 75) and ‘distinction’(above 75) in ‘Maths’,

#‘Reading’ and ‘Writing’.

interval=(0,40,50,60,75)

categories = ["Fail", "2nd class","1st class","Distinction"]

stuperf\_df["Marks\_cats"]=pd.cut(stuperf\_df.mathscore,interval,labels=categories)

ax=sns.countplot(x="Marks\_cats",hue="gender",palette="Set1",data=stuperf\_df)

ax.set(title="Marks categorisation for math",xlabel="Categories",ylabel="Number of students")

plt.show()

stuperf\_df["Marks\_Cats"]=pd.cut(stuperf\_df.readingscore,interval,labels=categories)

ax=sns.countplot(x="Marks\_Cats",hue="gender",palette="Set1",data=stuperf\_df)

ax.set(title="Marks categorisation for reading",xlabel="Categories",ylabel="Number of students")

plt.show()

stuperf\_df["Marks\_Cats"]=pd.cut(stuperf\_df.writingscore,interval,labels=categories)

ax=sns.countplot(x="Marks\_Cats",hue="gender",palette="Set1",data=stuperf\_df)

ax.set(title="Marks categorisation for writing",xlabel="Categories",ylabel="Number of students")

plt.show()