LAB Manual

PART A

(PART A : TO BE REFFERED BY STUDENTS)

**Experiment No.04**

**A.1 Aim:**

Apply data integration technique for the data obtained from different source.

**A.2 Prerequisite:**

**Python packages – matplotlib, pandas, numpy etc.**

**A.3 Outcome:**

**After successful completion of this experiment students will be able to integrate data from different sources and get in standard form.**

**Task**

1. Data is scattered in two CSV Files, integrate to one CSV file using Python script
2. At the time of data integration take care of invalid data, missing data etc (eg in one dataset age is mentioned in number and in other dataset it is mention in words so convert it into number or word and then integrate)
3. Check for the type of column in both the dataset it should be of similar type
4. After integrating the data check the skewness of the attribute Family Income, HSC % and SSC %
5. Give the Count of Female and Male student using Visualization
6. How many % Students are From Maharashtra and Outside Maharashtra Visualize the data
7. How many students belong to which particular religion
8. How many students have SSC and HSC percentage between 90-99, 80-89 and less than 80 Visualize the data using proper visualization technique
9. Prepare the summary of the dataset and present it

PART B

(PART B : TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case there is no Black board access available)***

|  |  |
| --- | --- |
| Roll No. C026 | Name: Anirbaan Ghatak |
| Class : B | Batch : EB1 |
| Date of Experiment: 29/08/2023 | Date of Submission: 03/08/2023 |
| Grade : |  |

**B.1 Answers of Task to be written by student:**

***(Paste your answers completed during the 2 hours of practical in the lab here)***

**Code:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

fl1 = pd.read\_excel("File1.xlsx")

fl2 = pd.read\_excel("File2.xlsx")

fl1.columns == fl2.columns

df = pd.concat([fl1, fl2], axis= 0)

df.isna().sum()

df["Home Country"].fillna(df["Home Country"].mode, inplace= True)

df["Religion"].fillna("Unkown", inplace= True)

df["Social Class"].fillna("Not Specified", inplace = True)

df.isna().sum()

df[df["Country"] == "India"]

numeric\_columns = df.select\_dtypes(include = [int,float]).columns

columns = ["Family Income", "SSC (%)", "HSC (%)"]

fig, axes = plt.subplots(nrows=7, ncols=5, figsize=(16, 16))

for i, a in enumerate(axes.flatten()):

  if i < len(columns):

    sns.boxplot(data=df, x=columns[i], ax=a)

    a.set\_title(columns[i])

  else:

    a.set\_visible(False)

fig.tight\_layout()

for i in columns:

  print(f"Skewness for {i} is: ",df[i].skew(), f"\t it's","Positively Skewed" if df[i].skew() >=0 else "Negatively Skewed" )

df["Gender"] = df["Gender"].convert\_dtypes()

len(df[df["Gender"] == "Male"])

male\_count = len(df[df["Gender"] == "Male"])

female\_count = len(df[df["Gender"] == "Female"])

plt.bar(['Male', 'Female'], [male\_count, female\_count], color=['Blue', 'Red'])

plt.xlabel('Gender')

plt.ylabel('Count')

plt.title('Gender Distribution in the Data Set')

without\_na = df.dropna()

df["Home State"].fillna(df["Home State"].mode, inplace = True)

in\_maharashtra\_withNa = len(df[df["Home State"] == "Maharashtra"])

in\_maharashtra\_without\_na = len(without\_na[without\_na["Home State"] == "Maharashtra"])

outside\_maharashtra = len(df[df["Home State"] != "Maharashtra"])

outside\_maharashtra\_without\_na = len(without\_na[without\_na["Home State"] != "Maharashtra"])

counts = [in\_maharashtra\_withNa, outside\_maharashtra]

plt.figure(figsize=(6, 4))

plt.pie(counts, labels = ["In Maharashtra", "Outside Maharashtra"], autopct="%1.1f%%" )

plt.xlabel('Category')

plt.ylabel('Count')

plt.title('Distribution of Individuals in and Outside Maharashtra with Na Values')

plt.xticks(rotation=45, ha="right")

plt.tight\_layout()

plt.show()

counts = [in\_maharashtra\_without\_na, outside\_maharashtra\_without\_na]

plt.figure(figsize=(6, 4))

plt.pie(counts, labels = ["In Maharashtra", "Outside Maharashtra"],autopct="%1.1f%%")

plt.xlabel('Category')

plt.ylabel('Count')

plt.title('Distribution of Individuals in and Outside Maharashtra without Na Values')

plt.xticks(rotation=45, ha="right")

plt.tight\_layout()

plt.show()

religions = df["Religion"].value\_counts()

plt.figure(figsize=(6, 5))

a = religions.plot(kind='bar', color='skyblue',use\_index=True)

plt.xlabel('Religion')

plt.ylabel('Count')

plt.title('Relgion of Students')

for i, v in enumerate(religions):

    a.text(i, v + 0.5, str(v), ha='center')

plt.tight\_layout()

plt.show()

marks = df[['SSC (%)', 'HSC (%)']]

X=['<80','80-89','90-99']

y=[len(marks[(marks['SSC (%)']<80) & (marks['HSC (%)']<80)]), len(marks[(marks['SSC (%)']>=80) & (marks['SSC (%)']<90) & (marks['HSC (%)']>=80) & (marks['HSC (%)']<90)]), len(marks[(marks['SSC (%)']>=90) & (marks['SSC (%)']<100) & (marks['HSC (%)']>=90) & (marks['HSC (%)']<100)])]

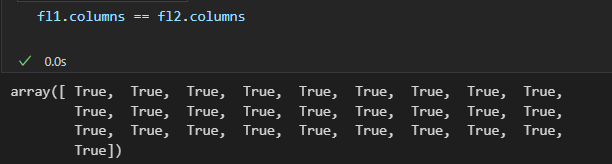
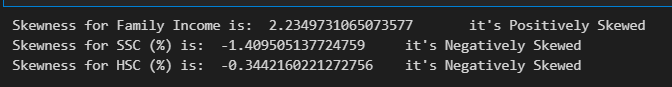
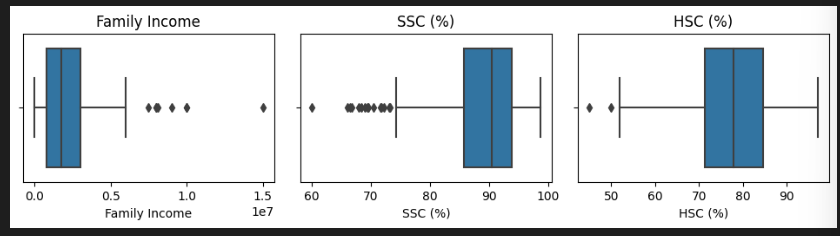
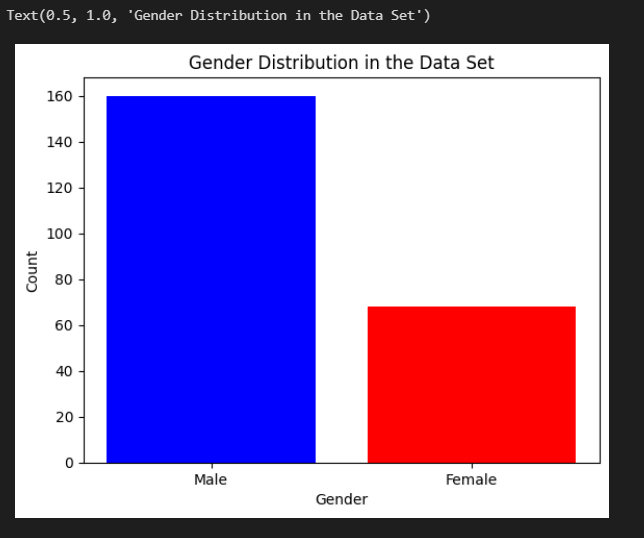
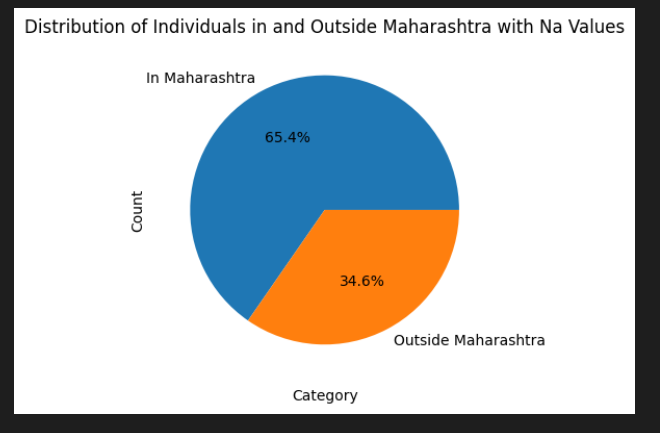
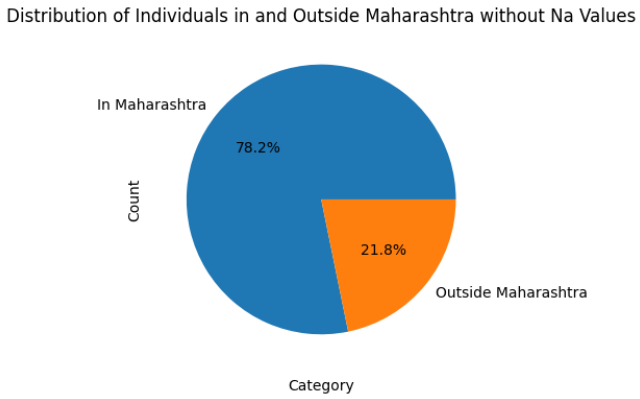
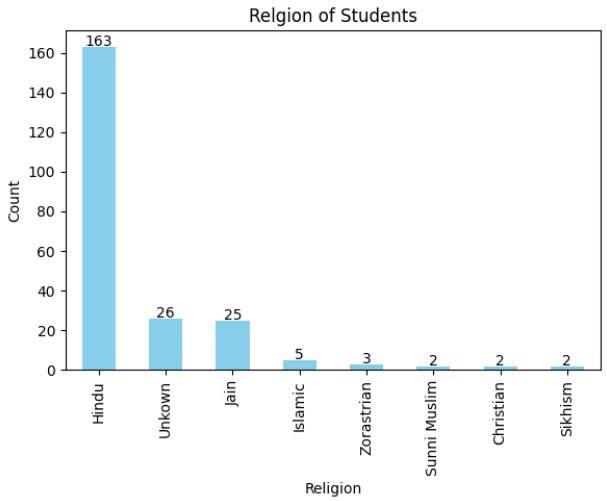
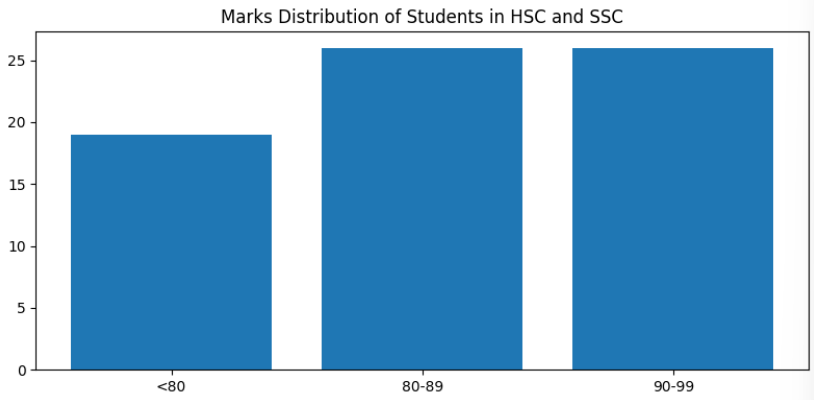
plt.figure(figsize=(8, 4))

plt.bar(X, y)

plt.title('Marks Distribution of Students in HSC and SSC')

plt.tight\_layout()

plt.show()

1. Check for the type of column in both the dataset it should be of similar type   
   
2. After integrating the data check the skewness of the attribute Family Income, HSC % and SSC %  
   
3. 
4. Give the Count of Female and Male student using Visualization  
   
5. How many % Students are From Maharashtra and Outside Maharashtra Visualize the data   
     
   
6. How many students belong to which particular religion   
   
7. How many students have SSC and HSC percentage between 90-99, 80-89 and less than 80 Visualize the data using proper visualization technique  
   

**B.2 Observations and learning:**

*Understood different types of visualization methods and merging of datasets*

**B.3 Conclusion:**

*Integrate data from different sources and get in standard form.*