PROJECT - 1 DASC 5300

OVERALL STATUS

- **About the given Dataset:-**

  A 1990 census dataset with several attributes was provided, which contained already preprocessed data where

  1) The income was biased with a threshold of 50K,
  2) "U.S." was converted to US to avoid confusion
  3) Unknown data was replaced with "?".

➔ The dataset consisted of 32560 rows having various attributes like Age, Work class, etc, without the header column names. After removing the rows having "?" in them, we had 30,161 rows in the dataset.

➔ In order to perform various functions on the dataset and visualize the output for analysis, we imported various libraries like Numpy, Pandas, Matplotlib, and Counter.

➔ Primarily, We imported libraries Pandas and Numpy for performing functions on a csv file. Pandas and Numpy are convenient to use as they provide open-source data analysis and are a manipulation tool for python.

➔ We have also imported the counter and matplotlib from the libraries for counting the occurrence of unique values and having a graphical representation respectively.

➔ In order to understand the data properly, we added a header column to the dataset. The header column had the names of the respective columns.

➔ The original dataset was named "data.csv" & the processed files were named clean_data and removed_clean_data. "Clean_data" consisted of all the rows without "?" in them and in removed_clean_data, we had all rows having "?". It is important to save the result so that we can manually check if the code and the output are right so that our analysis won't be heading in the wrong direction.

➔ To maintain the validity of the dataset, we need to first test in a small sample space. We used 175 as our sample size and 5291998 as our seed value. Seed is important as we need to make sure that each time the same sample space is used, we will get the same result, or if seed is missing, each time the result is changed even if the sample space is the same.

➔ After the validation is completed on the sample length of the data, analysis is done on the entire dataset, and then the result is checked if they are in correspondence.

➔ For the given dataset, we analyzed various columns like Education. Occupation to determine the population in each of those based on Gender and Age. We have used various threshold values in terms of Age like <21,21-40,41-60,61>. The output of these was visualized using various graphs and pie charts in matplotlib and then the analysis was concluded.

MILESTONE - PART 1

➔ The important part to run the code is to understand and import the libraries in your code. As we were using and analyzing the csv file, we imported the major libraries used to analyze the csv file, Numpy, and Pandas.

➔ In order to visualize the data, we used the most appropriate library to gain the output i.e Matplotlib.

➔ We wanted to count the occurrences of unique values in the dataset, we used Counter from Collections to execute that.

```
import pandas as pd
import numpy as np
from collections import Counter  #It is used to count the number of occurences in this case "?"
import matplotlib.pyplot as plt #It is used for plotting the data in graphical way
```

➔ To import all the libraries in python we did the following:

➔ To eliminate the rows with “?” and count them. We have used a simple loop and string match to perform this task. The code for the following is as below: -

```
count = 0
out = open('clean_data.csv','w')
out1 = open('removed_clean_data.csv','w')
for line in open('data.csv','r'):
  if("?" in line):
    count +=1
    out1.write(line)
  else:
    out.write(line)

print("The number of rows which has \'?\' mark :", count)
data1 = pd.read_csv("clean_data.csv", header = None)
```

```
The number of rows which has '?' mark : 2399
```

➔ The “removed_clean_data.csv” file contains all the rows with no “?” in them. The output as shown eliminates **2399** rows from the dataset having “?”.

➔ The “clean_data.csv” file consists of the data rows NOT having “?” in them. The number of rows in the clean data including the header column is 30,116 rows.

MILESTONE - PART 2

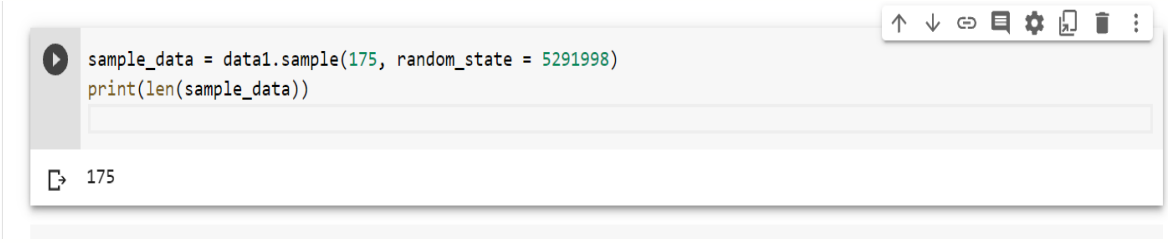
➔ In order to understand the data properly, we added a header column to the dataset. The header column had the names of the respective columns.

```
[ ]  data1.columns = ['Age','Workclass','fnlwgt','education','education-num','marital-status','Occupation',
                      'relationship','race','sex','capital-gain','capital-loss','hours-per-week',
                      'native-country','Salary']
```

➔ For analysis of various tasks, we performed tasks on a small sample size and sample seed. As we were given group 4, the sample size was 175, and the seed size, according to the instructions, was 5291998 (MM/DD/YYYY).

```
sample_data = data1.sample(175, random_state = 5291998)
```

➔ On sample size, we performed various functions which were mentioned in order to check the validity.



➔ From the above snippet, we can see the sample length was 175. We used this sample length and random seed value to the first test and analyze the data. It is always easier to check if your code and analysis are correct, if you run it with a small dataset, before approaching the entire dataset.

MILESTONE - PART 3

➔ As we took a small sample length to determine the results and analysis of the dataset, we were satisfied with the outcome and we performed the analysis of the code on the entire dataset, using the same correspondence code and logic. We can see the comparison of the output from sample size and the whole dataset in part 4 along with the analysis and graphical representation.

MILESTONE - PART 4

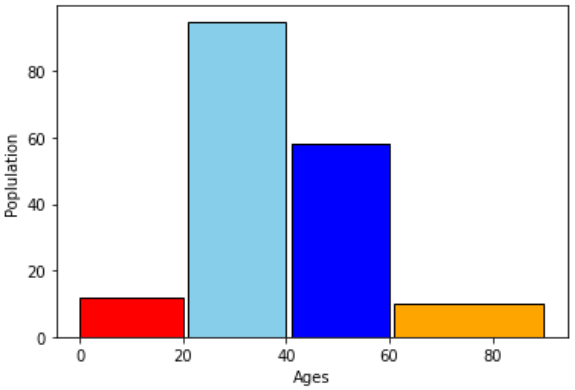
- **PART (4a)**

TASK:- HISTOGRAM OF AGE GROUPS

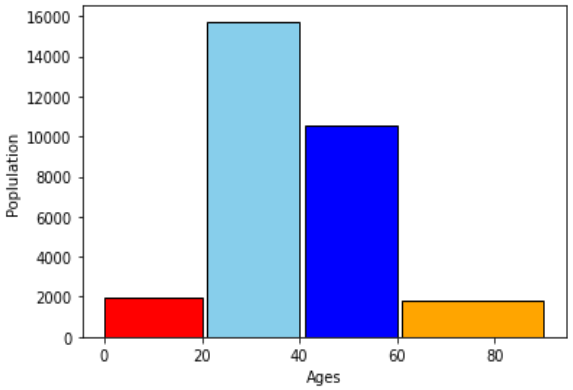
➔ We ran the code on sample data as well as the entire clean dataset.

➔ While running the data, we got correspondence between the graphs of them.

➔ We used matplotlib to display the graphical view of the datasets.



Frequency plot for sample data



Frequency plot for the entire clean data

➔ From above both graphs, we can determine that most populations are in the age group 21 - 40

➔ The highest number of population is observed in them, in both cases.

- For the sample data, we can see the following :
  1) < 20 : It's near 10 - 15
  2) 21 - 40: it's above 80
  3) 41 - 60 population is between 50-60
  4) > 60: It's less than 10.
- For the entire dataset, when the same code is used, we can observe the following numbers :
  1) < 20: It's near 1500-2000
  2) 21 - 40: population count is near 15000 - 16000
  3) 41 - 60 population is between 10000 - 11000
  4) > 60: it is less than 2000.

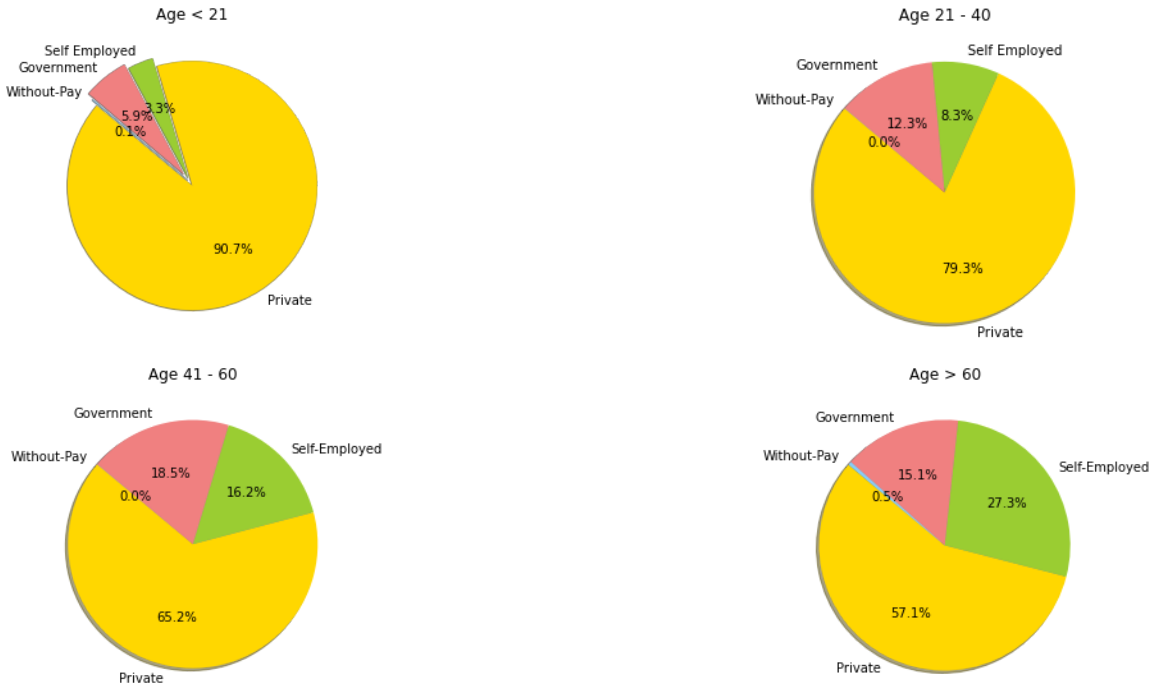
We can see a correspondence between the two graphs, indicating that the conclusion made on the sample length has matched the actual dataset.

➔ The histogram needs to be plot using matplotlib in python. The simplest code to plot any histogram is :

```
import matplotlib.pyplot as plt
x = [value1, value2, value3,....]
plt.hist(x, bins = number of bins)
plt.show().
```

- **PART (4b)**

TASK - Count the number of people in each of the following professions, Private, Self-employed, Government, and Without Pay.

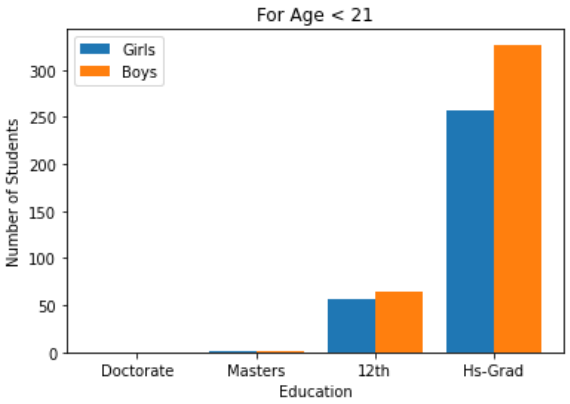


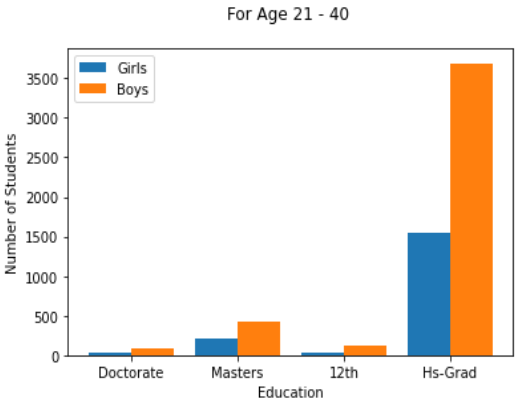
- ➔ From the above pie charts, We can state that for the age group < 21, 90.7% of people, who are of very young age, work in the Private sector, whereas 5.9% people work for various forms of government, only 3.3% of the population is self-employed & 0.1% of the population is working without pay. The majority of the minority age population is attracted towards the private corporate industry.
- ➔ For the age group 21 - 40, 79.3% population into the private industry whereas the government service population was 12.3% of the whole population serves as government workers while 8.3% of the population is self-employed
- ➔ For the age group 41 - 60, 65.2% of people are working in the private sector, whereas 16.2% of people are self-employed. 18.5% of the population is working for the government.
- ➔ For the last age group, the group above 60, the population that serves the private sector is the lowest of all age groups, 57.1%, whereas the population who is self-employed is highest of all age groups, 27.3%. 15.1% of the population serves the government and 0.5% falls in the without pay criteria.
- ➔ From the above pie chart, we can derive that, as the age group increases people tend more towards self-employment as we can see there is an increase in population as the age group also increases.
- ➔ Also, as the age group increases, we can see that the population is decreasing in the private sector. Finally, we can say that as the age group increases, people tend towards self-employment rather than a private job.

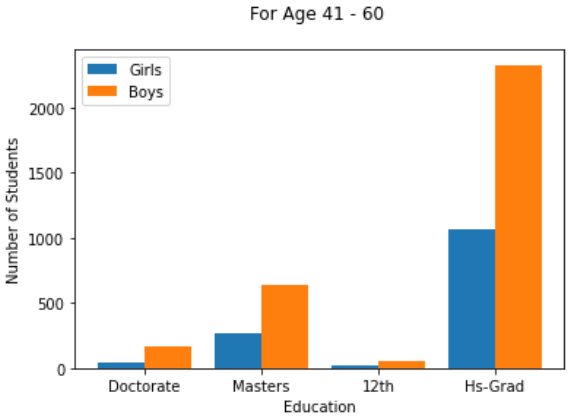
- **PART (4c)**

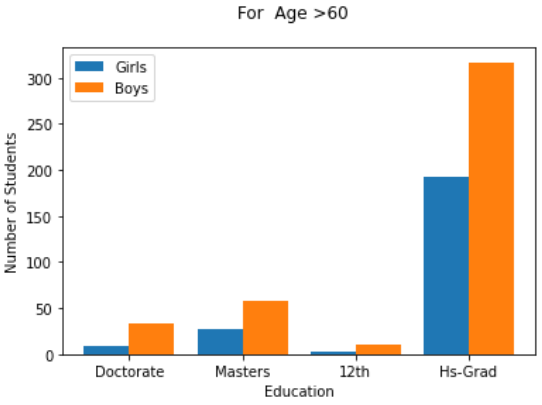
TASK 3:- Compare the education in 4 groups of males and females for the given age groups.

➔ We have grouped the 4 education streams like 12th, HS-Grad (High School Diploma), Masters, and Doctorate. We wanted to analyze the number of males and females who elect the following education.





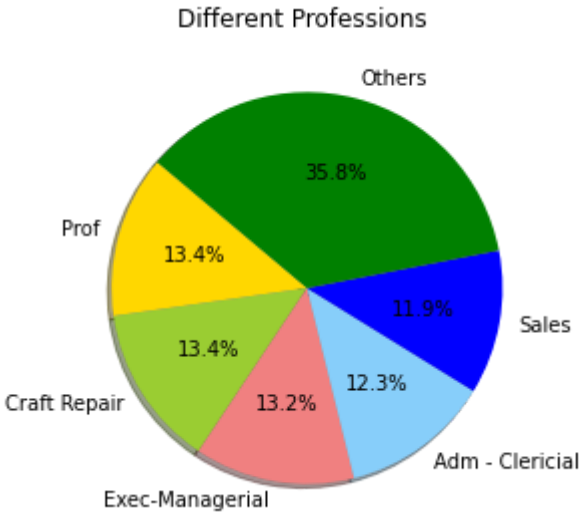


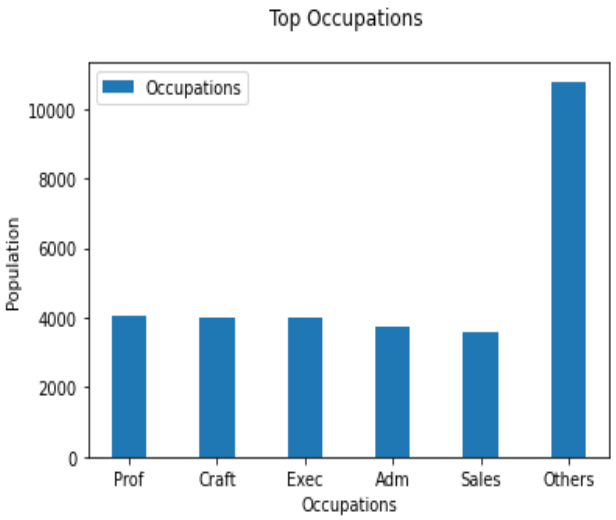


For comparing the education for age groups, we choose 4 education levels: Doctorate, Masters, 12th, and HS-Grad. The reason we choose these education levels is we wanted to analyze the difference between the count of the male and female populations participating from 12th to Doctorate. We can see that for different age groups, there are different counts of males and females for different educational levels. From the above data, we can see there is a very little amount of population in each age group who are trying to pursue master's education and doctorate education. There seems to be a very high demand for HS-Grad, as the number is significantly higher in all the age groups as compared to any other level. People are not trying to gain a full education in any particular domain as a doctorate is the highest degree in any educational field. The Number of females is less in all fields of education.

- **PART (4d)**

TASK:- To identify the Top 5 occupations for the given dataset





In order to find out the top 5 highest frequency occupations in the dataset, we need to iterate through every row in the dataset with column = "Occupation" and note down each one and count the number of times each occupation appears. We have used a function called value_counts() to count the occupation's frequency and return the one with the highest occurrence. We have created two lists named values() and counts() to store the Name and the occurrence respectively.

```
values = data1['Occupation'].value_counts().keys().tolist()
counts = data1['Occupation'].value_counts().tolist()
```

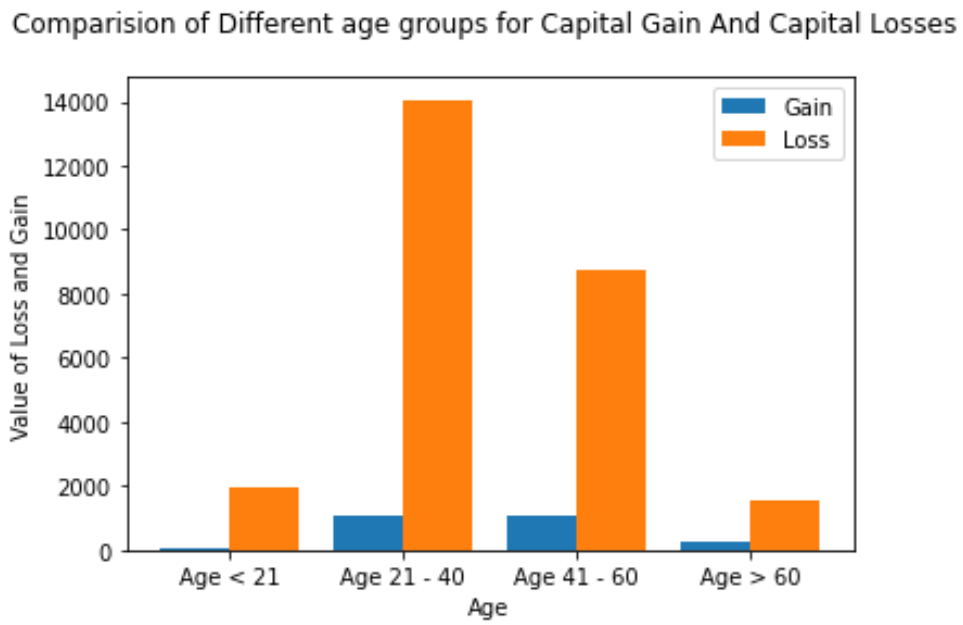
From the above code, we get that the Professions, Prof-specialty, Craft-repair, Exec-managerial, Adm-clerical, Sales have the highest number of occurrences with values 4032, 4023, 3987, 3712, 3580 respectively.

Other professions, except the ones mentioned above, cumulative have a greater count than these, but on an individual level, they have a count, less than the above ones.

The majority of the population from the overall age group prefers to be in the profession of Prof-Speciality & Craft Repair. Population in Prof-Speciality is only 10% more than that of Craft Repair.

A - PART 5

TASK 5:- Comparison of different age groups for capital gain and capital loss.



- ➔ In this case, we did an analysis of different age groups based on the criteria to analyze the number of people who have experienced the capital gain and capital loss and the summation of the loss and gain and average.
- ➔ From the graph above we can see the population of age 21-40 as an experienced loss at the highest rate as compared to any other age group. Whereas a common trait which was seen in all the age groups was that, the loss incurred to everyone was much higher than that of the profit or capital gain.
- ➔ For Age < 21 :- Average capital gain = 3013.32 & Average capital loss = 38.80
- ➔ For Age 21-40 Average capital gain = 10469.13 and Average capital loss = 80.83
- ➔ For age 41-60 Average capital gain = 15912.87 and Average capital loss = 128.02
- ➔ For age 61 and above Average capital gain = 2153.29 and Average capital loss = 127.14

B - File Description

- ➔ Basically, there are two new files created for this project.
- ➔ The first one is the clean_data.csv, it is a csv data structured file. This file contains the clean data without the special character '?'. Also, this file has a header for each column.
- ➔ The second file is removed_clean_data.csv for the removed data from the originally given dataset. This dataset contains the rows with the special character '?'.

C - Division of Labor

- ➔ We are a group of two completing this project. We divided the work equally from the start.
- ➔ From discussing the project plan to implementing it, both of us did the same amount of work.
- ➔ Initially, we started to analyze the dataset, started working on the dataset using Google Colab, which helped us to work on the same idea.
- ➔ After preprocessing the data together, we divided the data visualization task.
- ➔ The first two questions of data visualization were done by a member and the other two were done by the second member.
- ➔ We started the project on the second day, the project was allotted.
- ➔ Gradually, we completed the project 3-4 days before the submission and started to work on the project report.
- ➔ We completed our project report on time and submitted the report before the submission time.

D - Problems encountered and handling that problem in this project:-

- ➔ At first, we used highly complex functions like data frames in pandas, and later, we realized that it can be done by simple loop and string match.
- ➔ Exploring more kinds of analysis, we were unable to find a perfect and interesting analysis, but later studying the data thoroughly, we did analyze the data successfully.
- ➔ The graphical representation for the pie chart had the notations overlap, but later we overcame this with the explode function in pie and this helped us in exploring more functions for the pie-chart.
- ➔ Similarly, for the histogram, the problem we faced was to obtain the disjoint frequency graph.