Exercise 2-2

September 10, 2023

[3]:
```
## DSC 550
## Carlos Cano
## Activity 1.2
```

[4]:
```
## Import Libraries
```

[5]:
```
import numpy as np
import pandas as pd
from scipy.stats import norm
import matplotlib.pyplot as plt
```

[6]:
```
## Step 1 - Using a data set of your choice, write an introduction explaining␣
 ↪the data set.
```

[73]:
```
## Data sourced from Kaggle
## https://www.kaggle.com/datasets/yashkmd/
 ↪credit-profile-two-wheeler-loan-dataset
```

[7]:
```
df = pd.read_csv("credit_data.csv")
df.head(10)
```

[7]:

	Age	Gender	Income	Credit Score	Credit History Length	Number of Existing Loans	Loan Amount	Loan Tenure	Existing Customer
0	31	Male	36000	604	487	5	109373	221	No
1	25	Male	50000	447	386	2	150000	89	No
2	62	Other	178000	850	503	10	69099	110	Yes
3	69	Female	46000	668	349	6	150000	148	Yes
4	52	Male	132000	601	553				
5	64	Female	127000	850	158				
6	29	Male	15000	378	89				
7	30	Other	82000	424	610				
8	52	Male	119000	753	271				
9	39	Male	101000	575	424				

```
4                        5       150000          157               No
5                       10       108702          111              Yes
6                        1        26819          108               No
7                        2       126550           92               No
8                        8       150000          251              Yes
9                        5       113257           12               No

           State        City  LTV Ratio Employment Profile  Profile Score  \
0      Karnataka      Mysuru  90.943430           Salaried             77
1      Karnataka   Bengaluru  91.135253           Salaried             43
2  Uttar Pradesh      Kanpur  40.000000           Salaried             90
3      Karnataka   Bengaluru  87.393365      Self-Employed             86
4      Karnataka      Mysuru  66.158757           Salaried             90
5     Tamil Nadu  Coimbatore  82.331250      Self-Employed             92
6  Uttar Pradesh     Lucknow  95.000000      Self-Employed             25
7    West Bengal     Kolkata  93.634577           Salaried             58
8      Rajasthan      Jaipur  75.644166         Freelancer            100
9    Maharashtra      Nagpur  68.720556           Salaried             87

          Occupation
0             Doctor
1  Software Engineer
2             Banker
3         Contractor
4            Teacher
5         Contractor
6             Farmer
7             Banker
8             Writer
9             Banker
```

[8]:
```
df.shape
```

[8]:
```
(279856, 15)
```

[9]:
```
## Contains 15 Variables, with 279,856 Rows (Individuals)
```

[10]:
```
## Step 2 - Identify a question or question(s) that you would like to explore in␣
↪your data set.
```

[11]:
```
# Question 1 - What variety of jobs are reported?
```

[12]:
```
# Question 2 - How do Credit Score differ among sexes?
```

[13]:
```
# Question 3 - What is the median income?
```

[14]:
```
## Step 3 - Create at least three graphs that help answer these questions.
## Make sure your graphs are clearly readable and are labeled appropriately and␣
↪professionally.
```

[15]:
```
## Question 1 Prepwork

JobCount = df['Occupation'].value_counts()
JobCount
```

[15]:
```
Banker                  27760
Teacher                 27356
Civil Servant           27221
Software Engineer       27146
Doctor                  26582
Shopkeeper              21405
Contractor              21090
Farmer                  20966
Business Owner          20908
Student                 18521
Graphic Designer         5723
Photographer             5706
Independent Consultant   5628
Writer                   5572
Name: Occupation, dtype: int64
```

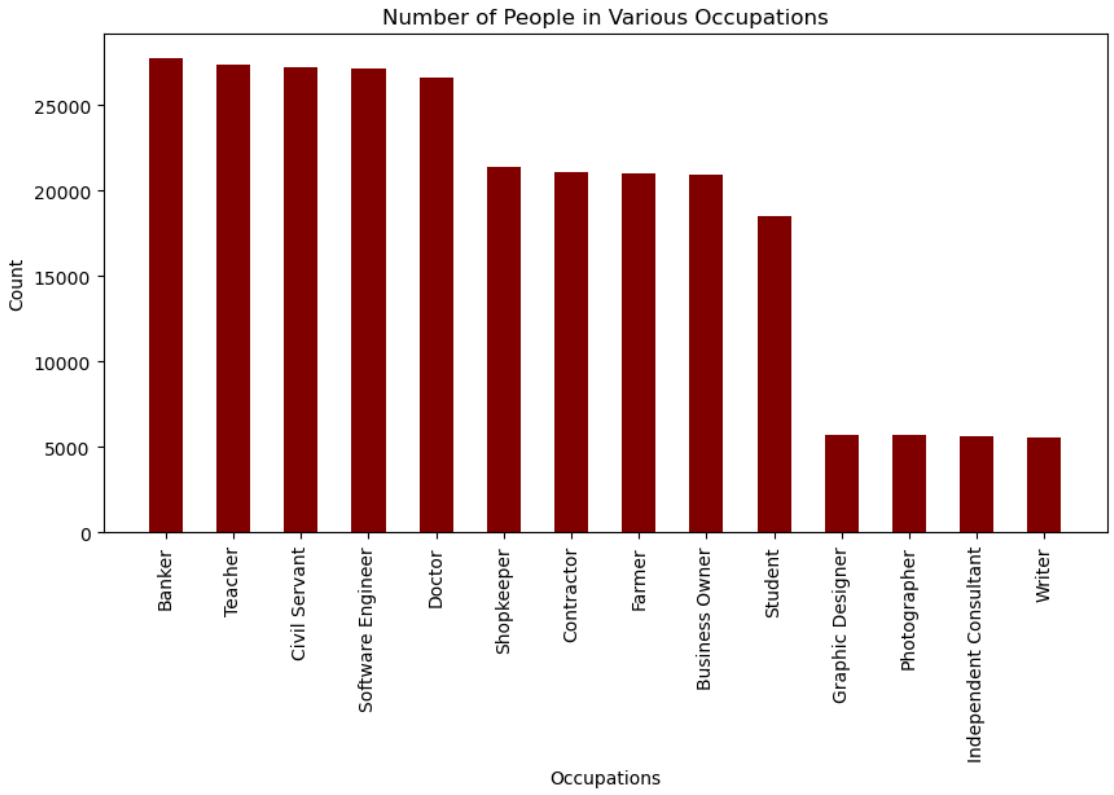
[16]:
```
## Question 1 Prepwork Continued


df1 = pd.DataFrame(JobCount)

df1 = df1.rename(columns={'Occupation':'Count'})

df1 = df1.reset_index()

df1 = df1.rename(columns={'index':'Job', 'Count':'Count'})

df1
```

[16]:

	Job	Count
0	Banker	27760
1	Teacher	27356
2	Civil Servant	27221
3	Software Engineer	27146
4	Doctor	26582
5	Shopkeeper	21405
6	Contractor	21090
7	Farmer	20966

```
8           Business Owner  20908
9                  Student  18521
10       Graphic Designer   5723
11           Photographer   5706
12  Independent Consultant   5628
13                  Writer   5572
```

[19]:
```
## Question 1 Graph

Job = list(df1["Job"])
Count = list(df1["Count"])

Count = Count
Count = np.array(Count, dtype=int)

fig = plt.figure(figsize = (10, 5))
ax = plt.subplot()

ax.ticklabel_format(useOffset=False, style='plain')

# creating the bar plot
plt.bar(Job[0:300000], Count[0:300000], color ='maroon',
        width = 0.5)

plt.xlabel("Occupations")
plt.xticks(rotation=90)
plt.ylabel("Count")
plt.title("Number of People in Various Occupations")

plt.show()
```



[20]:
```
## Question 2 Prepwork
```

[21]:
```
df2 = df[['Gender','Credit Score']]

df2.head()
```

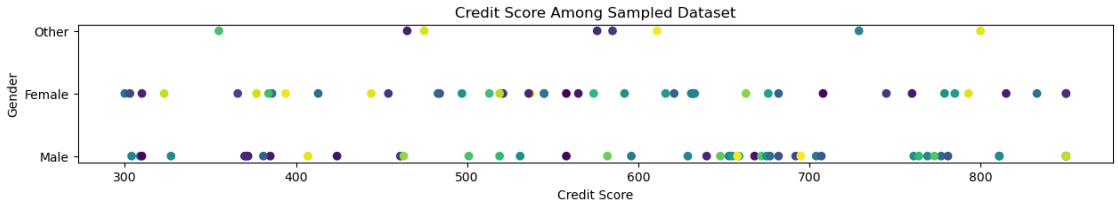
[21]:

	Gender	Credit Score
0	Male	604
1	Male	447
2	Other	850
3	Female	668
4	Male	601

[75]:
```
## Question 2 Graph

n = 100
df2_Sample = df2.sample(n)
x = df2_Sample['Credit Score']
y = df2_Sample['Gender']
colors = np.random.rand(n)
```

```
plt.figure(figsize=(15,2))

plt.scatter(x,y, c=colors)
plt.xlabel('Credit Score')
plt.ylabel('Gender')
plt.title('Credit Score Among Sampled Dataset')

plt.show
```

[75]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



[23]:
```
## Question 3 Prepwork

Income = df['Income']

df3 = pd.DataFrame(Income)

df3.head()
```

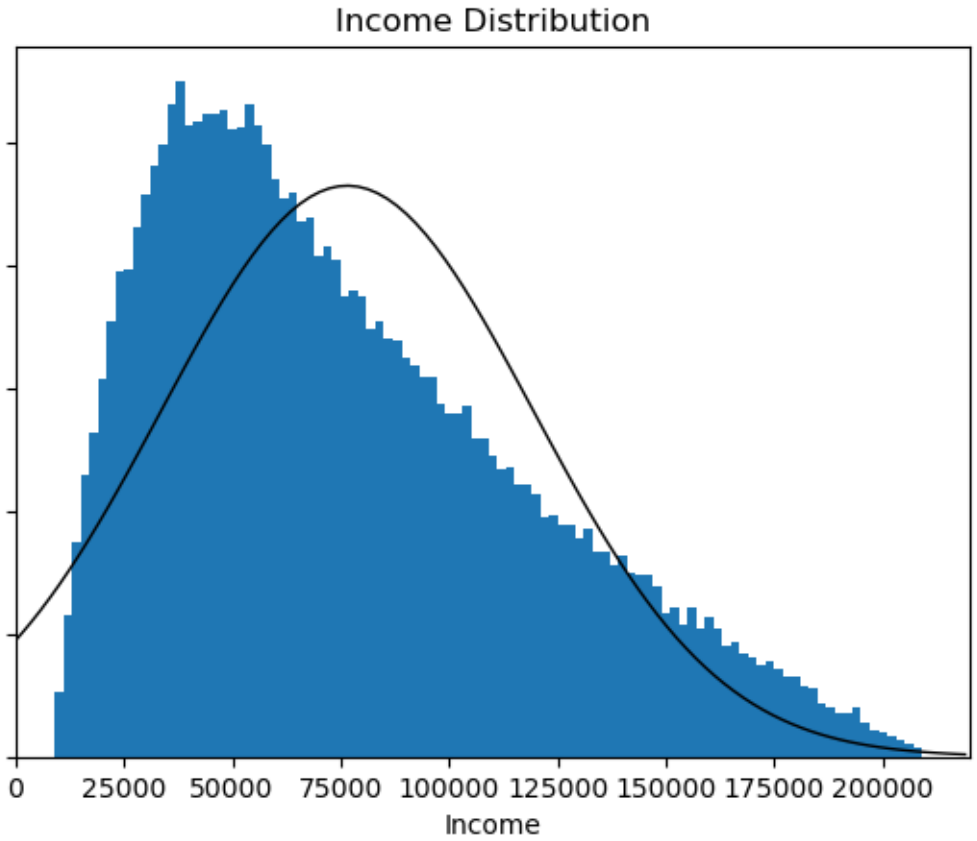
[23]:

	Income
0	36000
1	50000
2	178000
3	46000
4	132000

[56]:
```
## Question 3 Graph

mu,std = norm.fit(df3)

plt.hist(df3,bins=100,density=True)

xmin,xmax = plt.xlim()
x = np.linspace(xmin,xmax,100)
p = norm.pdf(x,mu,std)

plt.plot(x,p,'k',linewidth = 1)
```

```
plt.xlabel('Income')
plt.xlim(0,220000)
plt.ylabel('')
plt.gca().ticklabel_format(style='plain')
plt.tick_params(axis='y', which='both', labelleft=False)
plt.title('Income Distribution')
plt.show

df3["Income"].describe().apply("{0:.2f}".format)
```

[56]:
```
count    279856.00
mean      76499.16
std       42875.58
min        9000.00
25%       42000.00
50%       68000.00
75%      104000.00
max      209000.00
Name: Income, dtype: object
```



[25]:
```
## Step 4 - Explain what you have learned from each of your graphs.
```

[77]:
```
## What was learned from Question 1:

## Of the observed data from this data sample there were more bankers than any␣
↪other occupation.
```

[78]:
```
## What was learned from Question 2:

## There was higher concentration of credit scores amongst males in this sample␣
↪data.
```

[79]:
```
## What was learned from Question 3:

## In graph 3 there is a left skew in relation to income data, with a mean of␣
↪$76,499.16 based on this data sample.
```

[29]:
```
## Step 5 - Write a conclusion that summarizes your findings.
```

[30]:
```
## Conclusion:
```

[76]:
```
## Based on the finding and information gathered from this dataset, it can be␣
↪stated that there are higher rates of people in banker positions, that males␣
↪have higher concentration of better credit scores an that of those polled in␣
↪this data most are distributed within a left skew.
## Of specific note, this data is related to India and involves those submitting␣
↪applications for loans, data may be slightly skewed as a result because people␣
↪seeking a loan with deficient score would not apply. Also those in higher tax␣
↪brackets would also not be seeking loans alternatively.
```