Banking Dataset

```
#importing libraries


import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


data =
pd.read_csv(r"C:\Users\Shiuli\Downloads\DsResearch\DsResearch\Banking\banking_data.csv")
print(data.head())
print(data.info())
print(data.describe())
print(data.columns)
#Q.1
#Distribution of age: As the 'Age' column contains numerical variables, so we plot a Histogram


plt.hist(data['age'],color='blue',edgecolor='black')
plt.xlabel('Age of the client')
plt.ylabel('Number of clients')
plt.title('Distribution of age among the clients')
plt.show()


#Q.2
#To check how does the job type vary among the clients: Bar plot


job_counts = data['job'].value_counts()
job_counts.plot(kind='bar')
plt.xlabel('Job Type')
plt.ylabel('Number of Clients')
plt.title('Distribution of Job Types among the clients')
plt.show()
#Q.3
#Marital status distribution of the clients

marital_counts = data['marital'].value_counts()
marital_counts.plot(kind='bar')
plt.xlabel('Marital status')
plt.ylabel('Number of Clients')
plt.title('Distribution of Marital status')
plt.show()

#Q.4

education_counts = data['education'].value_counts()
education_counts.plot(kind='bar')
plt.xlabel('Level of education')
plt.ylabel('Number of Clients')
plt.title('Distribution of Level of education')
plt.show()

#Q.5

default_counts = data['default'].value_counts()
default_counts.plot(kind='bar')
plt.xlabel('Credit in default')
plt.ylabel('Number of Clients')
plt.title('Proportion of clients have credit in default')
plt.show()

# 815 clients among 45211 clients have credit in default
```

```
#Q.6
plt.hist(data['balance'],color='blue',edgecolor='black')
plt.xlabel('Average yearly balance')
plt.ylabel('Number of clients')
plt.title('Distribution of average yearly balance among the clients')
plt.show()

#Q.7
Housing_loan_count = data['housing'].value_counts()
print(housing_loan_count)
# 25130 clients have housing loans

#Q.8
loan_count = data['loan'].value_counts()
#print(loan_count)
# 7244 clients have personal loans

#Q.9

contact_counts = data['contact'].value_counts()
contact_counts.plot(kind='bar')
plt.xlabel('Type of communication')
plt.ylabel('Number of Clients')
plt.title('Type of communication used to contact the clients')
plt.show()

#Q.10

plt.hist(data['day'],color='blue',edgecolor='black')
plt.xlabel('Last contact day')
plt.ylabel('Number of clients')
plt.title('Distribution of the last contact day of the month')
plt.show()
#Q.11

month_counts = data['month'].value_counts()
month_counts.plot(kind='bar')
plt.xlabel('Last contact month')
plt.ylabel('Number of Clients')
plt.title('Distribution of the Last contact month of the year')
plt.show()

#Q.12

plt.hist(data['duration'],color='blue',edgecolor='black')
plt.xlabel('Duration in secs')
plt.ylabel('Number of clients')
plt.title('Distribution of the duration of the last contact')
plt.show()

#Q.13

plt.hist(data['campaign'],color='blue',edgecolor='black')
plt.xlabel('Number of contacts performed')
plt.ylabel('Number of clients')
plt.title('Number of contacts performed before the current campaign for each client')
plt.show()

#Q.14

plt.hist(data['pdays'],color='blue',edgecolor='black')
plt.xlabel('Number of Days')
plt.ylabel('Number of clients')
plt.title('Distribution of the number of days passed since the client was last contacted from a previous campaign')
plt.show()

#Q.15

plt.hist(data['previous'],color='blue',edgecolor='black')
plt.xlabel('Number of contacts')
plt.ylabel('Number of clients')
plt.title(‘Contacts performed before the current campaign for each client’)
plt.show()

#Q.16

poutcome_counts = data['poutcome'].value_counts()
poutcome_counts.plot(kind='bar')
plt.xlabel('Outcome of Previous campaign')
plt.ylabel('Number of Clients')
plt.title('Distribution of the outcome of previous campaign')
plt.show()

#Q.17

yes_no_counts = data['y'].value_counts()
yes_no_counts.plot(kind='bar')
plt.xlabel('subscribed to a term deposit')
plt.ylabel('Number of Clients')
plt.title('Distribution of subscription of clients')
plt.show()
```

```
#Q.18
data['y'] = data['y'].apply(lambda x: 1 if x == 'yes' else 0)
numeric_cols = ['age', 'balance', 'duration', 'campaign', 'pdays', 'previous', 'y']
data_numeric = data[numeric_cols]
correlation_matrix = data_numeric.corr()
print(correlation_matrix)

#To see the relation between the categorical values and y:
sns.countplot(data=data, x='job', hue='y', palette='viridis')
plt.title('Relation between job type and y')
plt.xlabel('Job Type')
plt.ylabel('Number of clients')
plt.show()
# 'management', 'technician', 'blue-collar','admin','retired' are the job type who has subscribed to a term deposite more

sns.countplot(data=data, x='housing', hue='y', palette='viridis')
plt.title('Relation between housing loan and y')
plt.xlabel('Housing loan')
plt.ylabel('Number of clients')
plt.show()
# The clients with no housing loan has subscribed to a term deposite more

#Similarly observing the countplot, clients with no personal loan, cellular communication type to contact clients, and May,June,July,August months to contact the clients are the useful situation for a term deposite
```