ADS500B-02-FA21 {-}

Group 3

In [2]:

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

In [3]:

```
%pwd
```

Out[3]:

```
'/Users/kyledalope'
```

In [4]:

```
cd '/Users/kyledalope/downloads'
```

```
/Users/kyledalope/Downloads
```

In [5]:

```
df = pd.read_csv('bank_marketing.csv', sep=";")
```

In [6]:

```
df.head()
```

Out[6]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous
0	58.0	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0
1	44.0	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0
2	33.0	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0
3	47.0	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0
4	33.0	unknown	single	unknown	no	1	no	no	NaN	5	may	198	1	-1	0

In [7]:

```
df.count()
```

Out[7]:

```
age          43872
job          45211
marital      45211
education    45211
default      43905
balance      45211
housing      45211
loan         45211
contact      43828
day          45211
month        45211
duration     45211
campaign     45211
pdays        45211
previous     45211
poutcome     45211
deposit      45211
dtype: int64
```

In [8]:

```
df.isna()
```

Out[8]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	True	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45206	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
45207	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
45208	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
45209	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
45210	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

45211 rows × 17 columns

In [42]:

```
df.dropna()
df = df[df.loc[:]!=0].dropna()
```

In [43]:

```
df.describe() #with Nulls/Blanks removed
```

Out[43]:

	age	balance	day	duration	campaign	pdays	previous
count	7146.000000	7146.000000	7146.000000	7146.000000	7146.000000	7146.000000	7146.000000
mean	40.828296	1658.395746	14.303247	258.968654	2.056535	225.566331	3.179961
std	11.406420	3175.218171	7.897718	234.285177	1.554387	115.780516	4.738133
min	18.000000	-1884.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	32.000000	227.250000	7.000000	113.000000	1.000000	133.000000	1.000000
50%	38.000000	677.000000	14.000000	193.000000	2.000000	195.000000	2.000000
75%	47.000000	1846.250000	20.000000	323.000000	2.000000	329.000000	4.000000
max	93.000000	81204.000000	31.000000	2219.000000	16.000000	871.000000	275.000000

In [11]:

```
df.dtypes
```

Out[11]:

```
age          float64
job           object
marital       object
education     object
default       object
balance        int64
housing       object
loan          object
contact       object
day            int64
month         object
duration       int64
campaign       int64
pdays          int64
previous       int64
poutcome      object
```

```
deposit      object
dtype: object
```

In [44]:

```
df.corr() #Determine which variables can be correlated and change data type if needed
```

Out[44]:

	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.123369	0.014267	0.049438	0.002639	-0.104849	-0.000423
balance	0.123369	1.000000	0.047057	0.044030	-0.007162	-0.114372	0.001602
day	0.014267	0.047057	1.000000	-0.014402	-0.030993	-0.081245	-0.021655
duration	0.049438	0.044030	-0.014402	1.000000	-0.084832	-0.020868	0.006144
campaign	0.002639	-0.007162	-0.030993	-0.084832	1.000000	0.059477	0.121570
pdays	-0.104849	-0.114372	-0.081245	-0.020868	0.059477	1.000000	-0.020884
previous	-0.000423	0.001602	-0.021655	0.006144	0.121570	-0.020884	1.000000

In [13]:

```
client_df = df[['age', 'job', 'marital', 'education', 'default', 'housing', 'loan']]
client_df.dropna()
```

Out[13]:

	age	job	marital	education	default	housing	loan
0	58.0	management	married	tertiary	no	yes	no
1	44.0	technician	single	secondary	no	yes	no
2	33.0	entrepreneur	married	secondary	no	yes	yes
3	47.0	blue-collar	married	unknown	no	yes	no
4	33.0	unknown	single	unknown	no	no	no
...	...	...	...	...	...	...	...
45206	51.0	technician	married	tertiary	no	no	no
45207	71.0	retired	divorced	primary	no	no	no
45208	72.0	retired	married	secondary	no	no	no
45209	57.0	blue-collar	married	secondary	no	no	no
45210	37.0	entrepreneur	married	secondary	no	no	no

42605 rows × 7 columns

In [14]:

```
#Observe the average age, highest education, and the counts of the clients based on the loans
client_df.groupby('age').size().reset_index(name = 'Age Count').sort_values(by='Age Count', ascending=False)
```

Out[14]:

	age	Age Count
14	32.0	2031
13	31.0	1945
15	33.0	1921
16	34.0	1876
17	35.0	1836
...	...	...
72	90.0	2
73	92.0	2
74	93.0	2
76	95.0	2
75	94.0	1

77 rows × 2 columns

In [15]:

```
client_df['age'].mean()
```

Out[15]:

```
40.92478118161926
```

In [16]:

```
client_df.groupby('education').size().reset_index(name = 'Education Count').sort_values(by='Education Count', a
```

Out[16]:

	education	Education Count
1	secondary	23202
2	tertiary	13301

	education	Education Count
0	primary	6851
3	unknown	1857

In [17]:

```
client_df['education'].value_counts()
```

Out[17]:

```
secondary    23202
tertiary     13301
primary       6851
unknown       1857
Name: education, dtype: int64
```

In [18]:

```
client_df.groupby(['housing', 'loan']).size().reset_index(name = 'Loan Count').sort_values(by='Loan Count', asc
```

Out[18]:

	housing	loan	Loan Count
2	yes	no	20763
0	no	no	17204
3	yes	yes	4367
1	no	yes	2877

Determined the most common loan type/combination

Housing loan was the most common loan bank clients had in the data set.

In [19]:

```
Marital = client_df['marital'].value_counts()
Marital
```

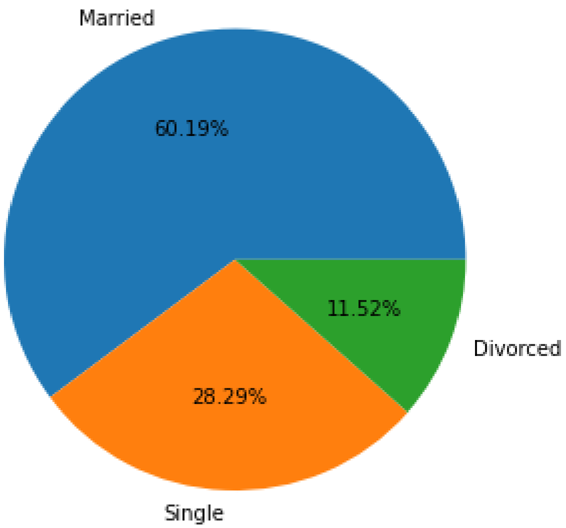
Out[19]:

```
married     27214
single      12790
divorced     5207
Name: marital, dtype: int64
```

In [20]:

```
fig = plt.figure()
ax = fig.add_axes([0, 0, 1, 1])
ax.axis('equal')
```

```
status = ['Married', 'Single', 'Divorced']
counts = [27214, 12790, 5207]
ax.pie(counts, labels = status, autopct='%1.2f%%')
plt.show()
```



In [24]:

```
contact_df = df[['month', 'duration']]
```

In [25]:

```
contact_df['month'].value_counts()
```

Out[25]:

```
may    13766
jul     6895
aug     6247
jun     5341
nov     3970
apr     2932
feb     2649
jan     1403
oct      738
sep      579
mar      477
dec      214
Name: month, dtype: int64
```

In [26]:
```
contact_df['duration'].mean()
```

Out[26]:
```
258.1630797814691
```

In [36]:
```
contact_df.groupby(['month']).sum().sort_values(by='duration', ascending=False)
```

Out[36]:

	duration
month	
may	3591856
jul	1847690
aug	1451816
jun	1298332
nov	1005000
apr	874026
feb	657742
jan	376313
oct	212767
sep	169214
mar	116579
dec	70476

In [45]:
```
deposit_df = df[['previous', 'campaign', 'deposit']]
```

In [46]:
```
deposit_df.describe()
```

Out[46]:

	previous	campaign
count	7146.000000	7146.000000
mean	3.179961	2.056535
std	4.738133	1.554387

	previous	campaign
min	1.000000	1.000000
25%	1.000000	1.000000
50%	2.000000	2.000000
75%	4.000000	2.000000
max	275.000000	16.000000

In [57]:

```
deposit_df.groupby(['previous', 'campaign']).size().reset_index(name = 'Deposit').sort_values(by='Deposit', asc
```

Out[57]:

	previous	campaign	Deposit
0	1.0	1	1320
13	2.0	1	953
1	1.0	2	645
14	2.0	2	504
24	3.0	1	435
...	...	...	...
150	19.0	2	1
152	19.0	5	1
153	19.0	7	1
155	20.0	2	1
198	275.0	2	1

199 rows × 3 columns

In [ ]: