Case Project: Banking

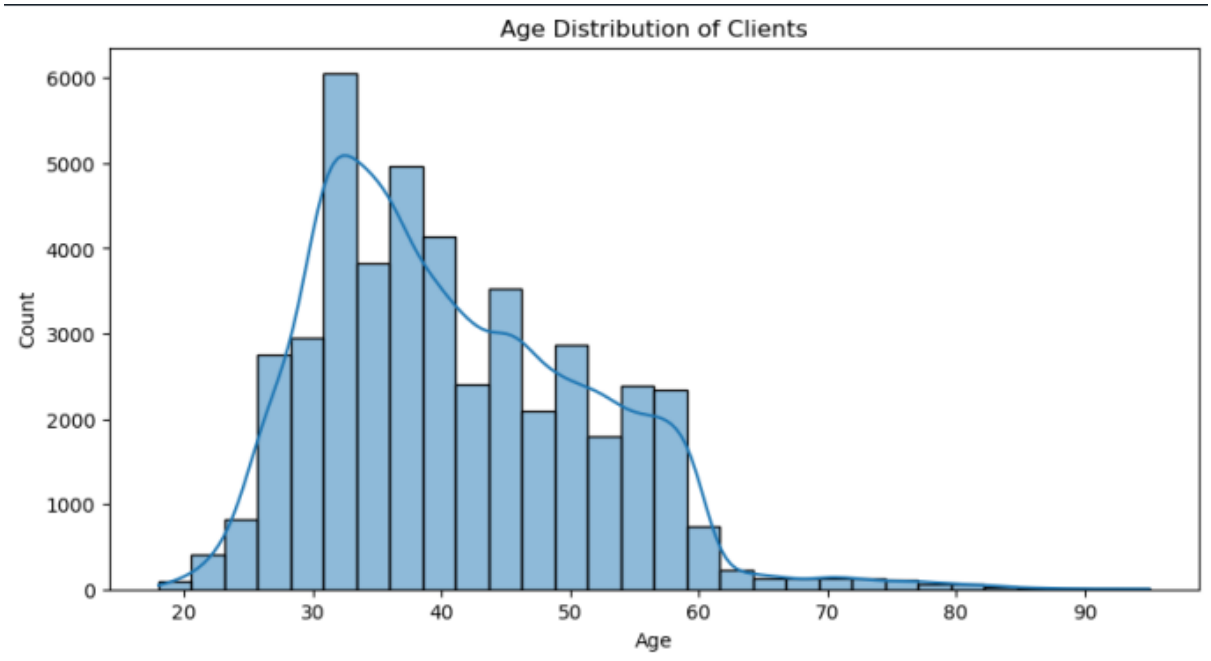
Sarthak Das

- **What is the distribution of age among the clients?**

  Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)


plt.figure(figsize=(10, 5))
sns.histplot(df['age'], bins=30, kde=True)
plt.title("Age Distribution of Clients")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```

Output:

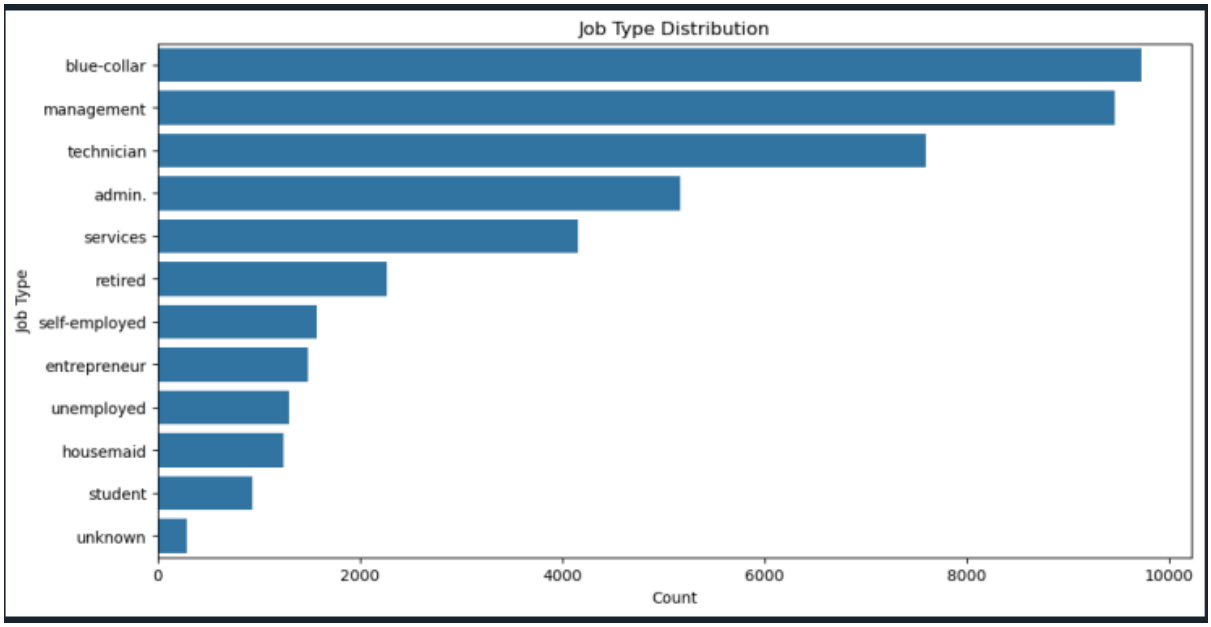


- **How does the job type vary among the clients?**

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)


plt.figure(figsize=(12, 6))
sns.countplot(y=df['job'], order=df['job'].value_counts().index)
plt.title("Job Type Distribution")
plt.xlabel("Count")
plt.ylabel("Job Type")
plt.show()
```

Output:

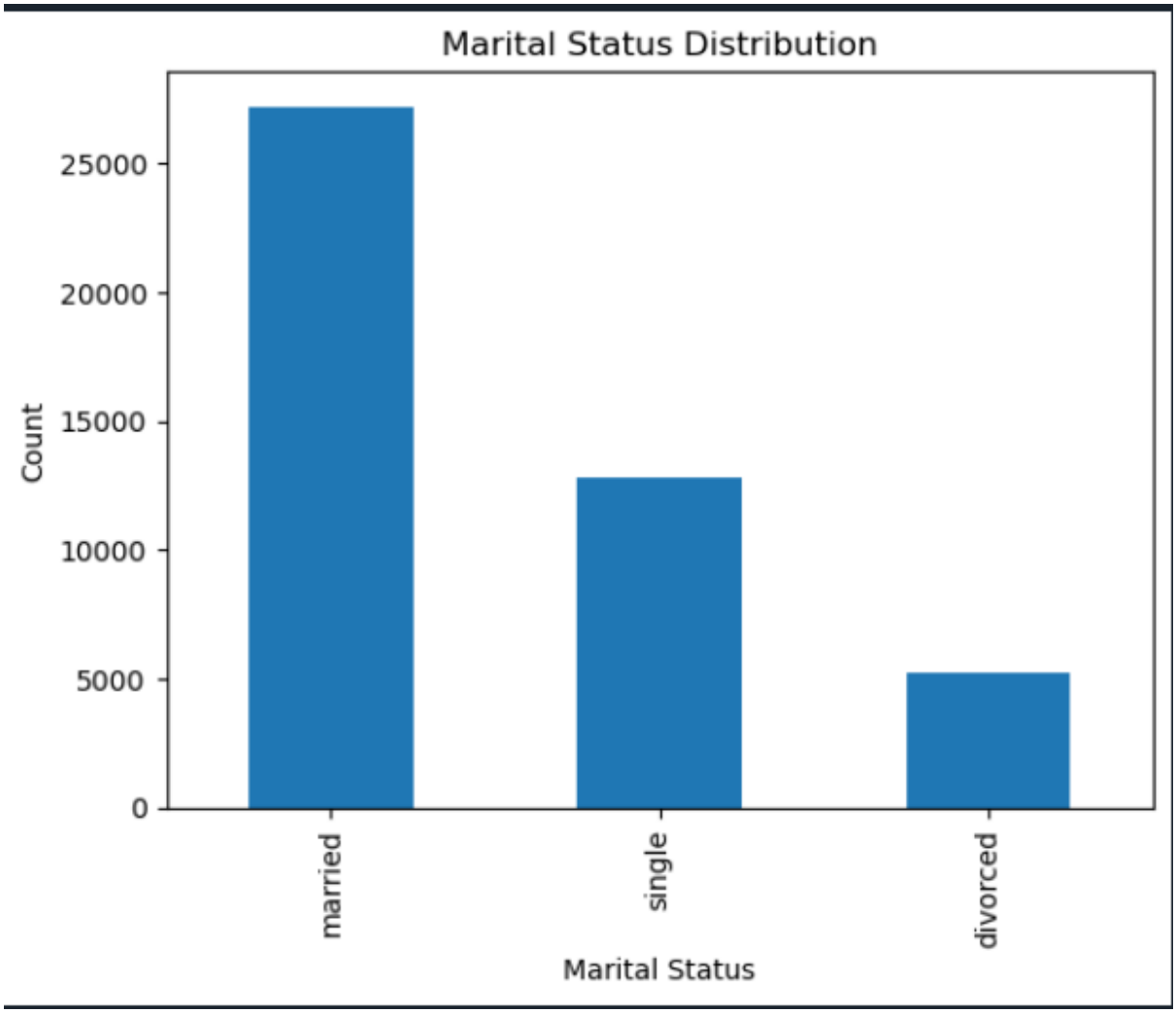


- **What is the marital status distribution of the clients?**

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)


df['marital'].value_counts().plot(kind='bar', title="Marital Status Distribution")
plt.xlabel("Marital Status")
plt.ylabel("Count")
plt.show()
```

Output:



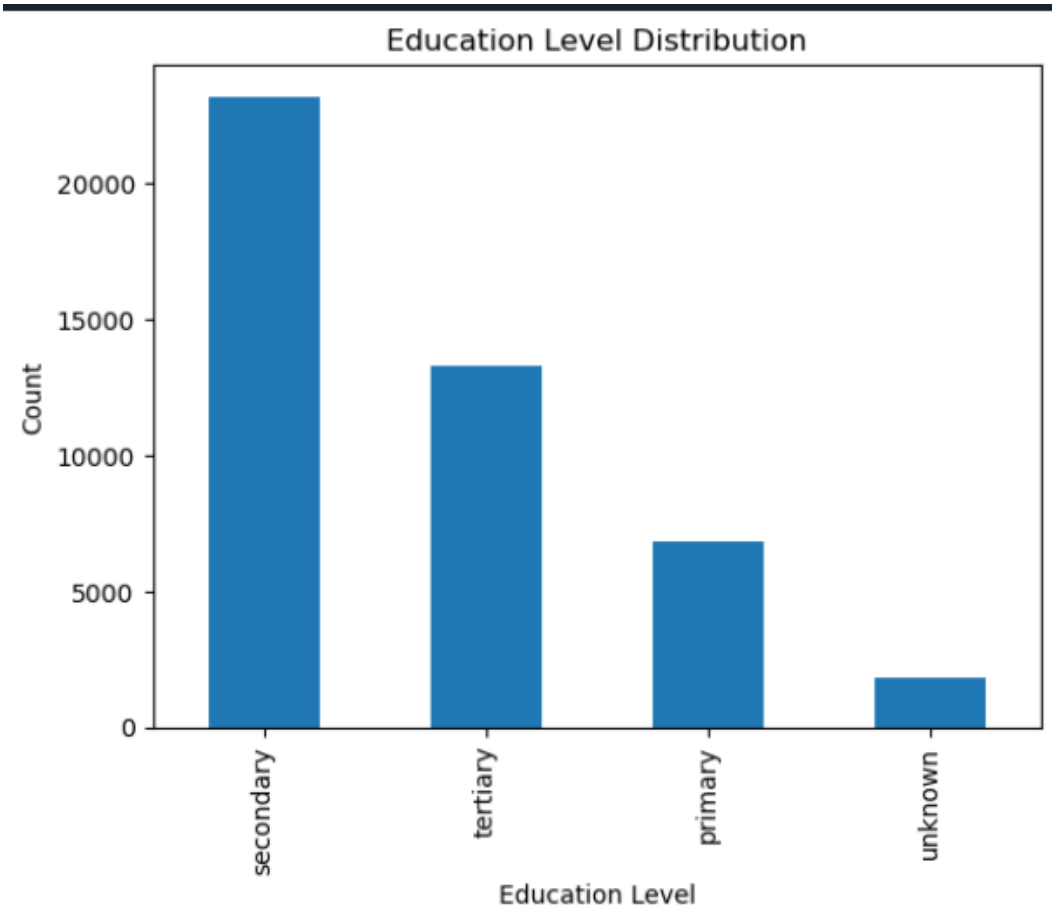
- **What is the level of education among the clients?**

  Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)

df['education'].value_counts().plot(kind='bar', title="Education Level Distribution")
plt.xlabel("Education Level")
plt.ylabel("Count")
```

plt.show()

Output:



- **What proportion of clients have credit in default?**

  Code:

  import pandas as pd

  import matplotlib.pyplot as plt

  import seaborn as sns

  file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

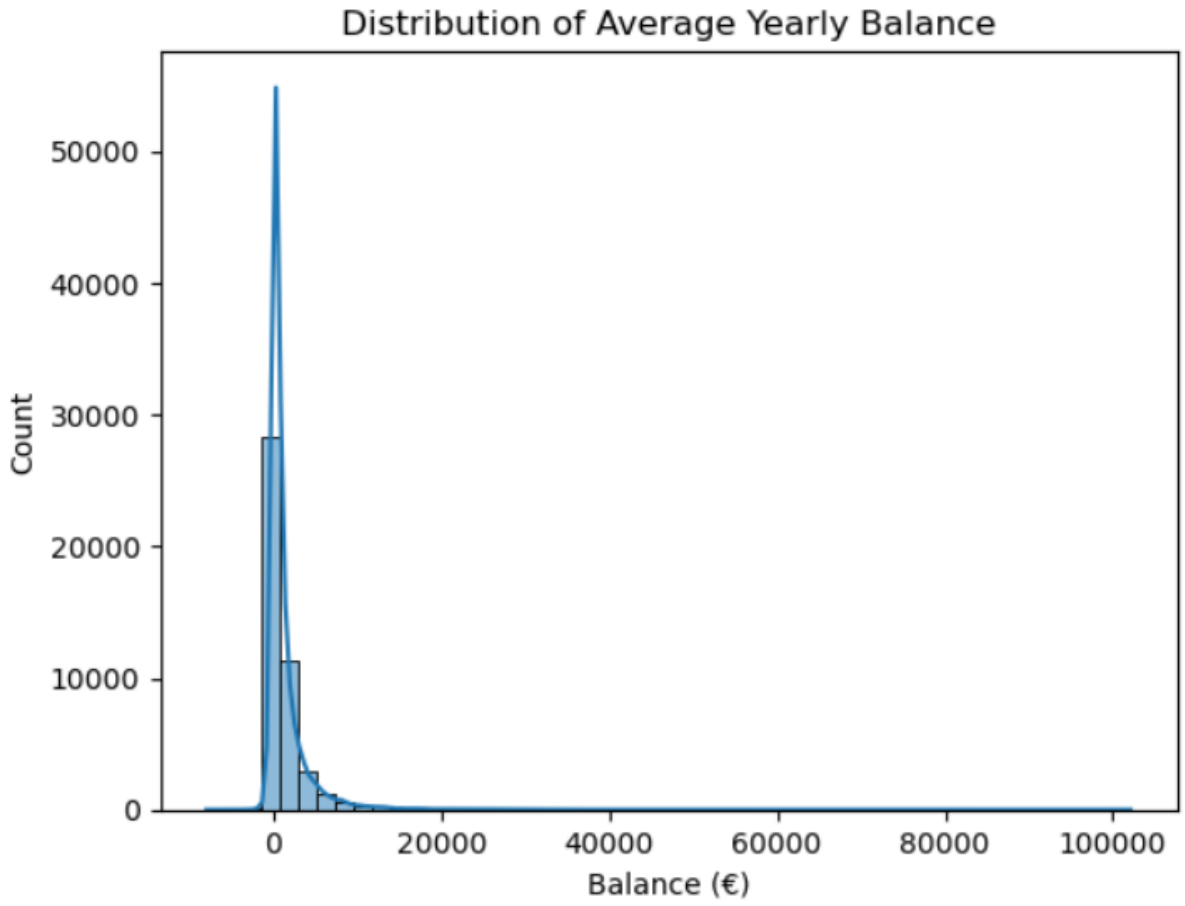
  df = pd.read_csv(file)

  print(df['default'].value_counts(normalize=True) * 100)

```
default
no     98.197541
yes     1.802459
Name: proportion, dtype: float64
```

- What is the distribution of average yearly balance among the clients?

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)



sns.histplot(df['balance'], bins=50, kde=True)
plt.title("Distribution of Average Yearly Balance")
plt.xlabel("Balance (€)")
plt.ylabel("Count")
plt.show()
```



- **How many clients have housing loans?**

  Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)


print("Housing Loan Count:\n", df['housing'].value_counts())
```

Output:

```
Housing Loan Count:
 housing
yes    25130
no     20086
Name: count, dtype: int64
```

- How many clients have personal loans?

  Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)

print("Personal Loan Count:\n", df['loan'].value_counts())
```

Output:

```
Personal Loan Count:
 loan
no     37972
yes     7244
Name: count, dtype: int64
```

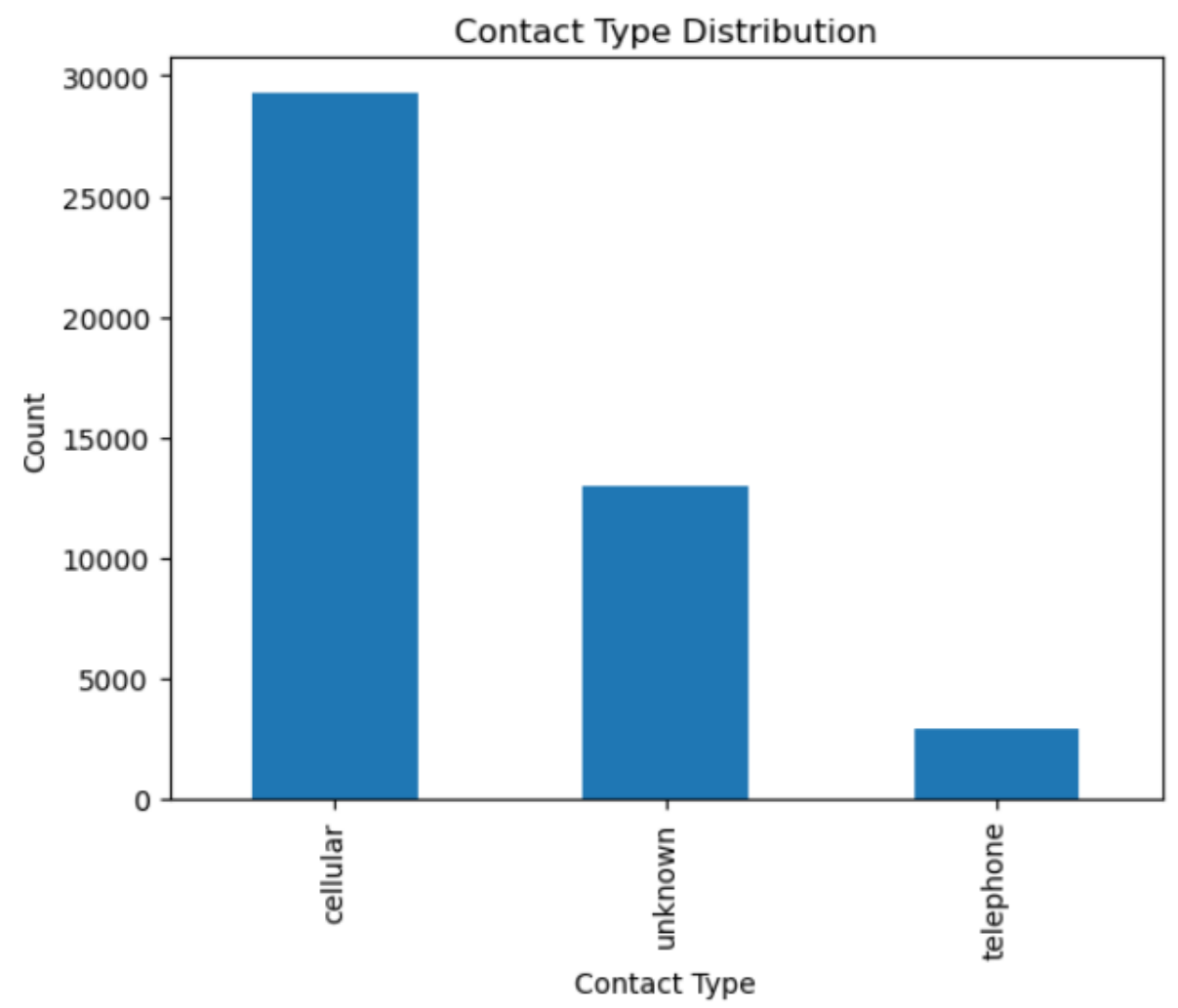
- **What are the communication types used for contacting clients during the campaign?**

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)
```

```
df['contact'].value_counts().plot(kind='bar', title="Contact Type Distribution")
plt.xlabel("Contact Type")
plt.ylabel("Count")
plt.show()
```

Output:



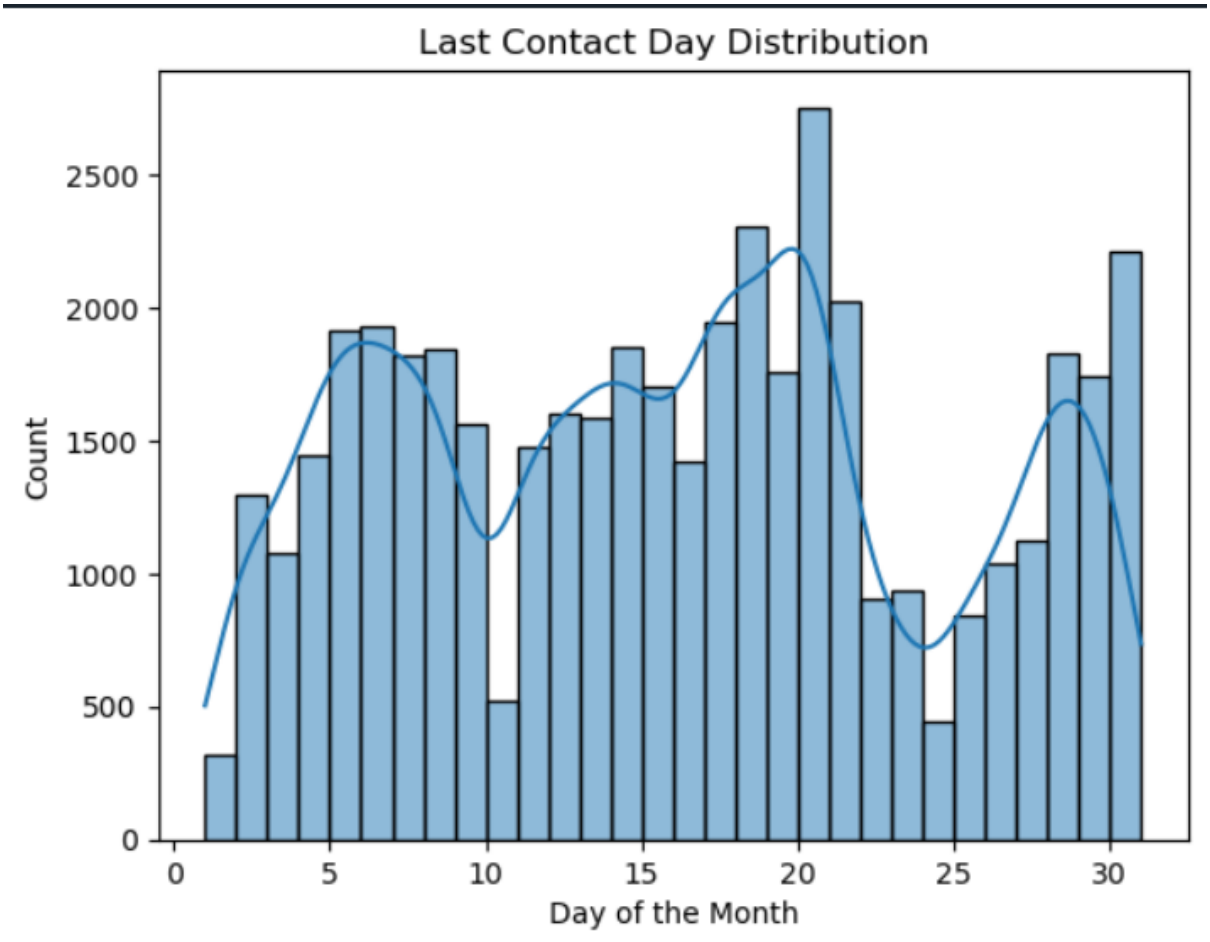
- **What is the distribution of the last contact day of the month?**

  Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)


sns.histplot(df['day'], bins=30, kde=True)
plt.title("Last Contact Day Distribution")
plt.xlabel("Day of the Month")
plt.ylabel("Count")
plt.show()
```

Output:



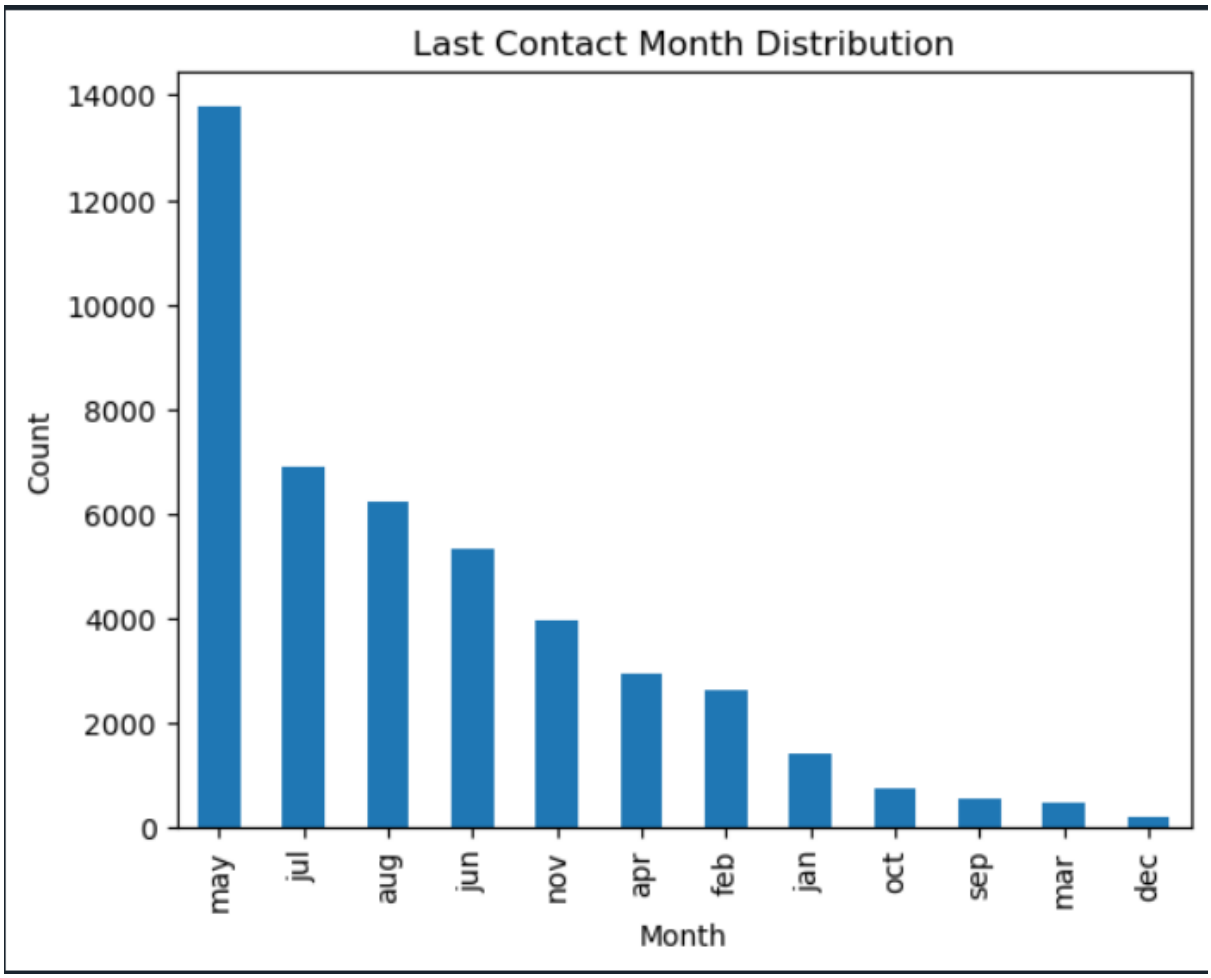
- **How does the last contact month vary among the clients?**

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
import seaborn as sns


file = r"C:\Users\Sarthak
Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)



df['month'].value_counts().plot(kind='bar', title="Last Contact Month Distribution")
plt.xlabel("Month")
plt.ylabel("Count")
plt.show()
```

Output:

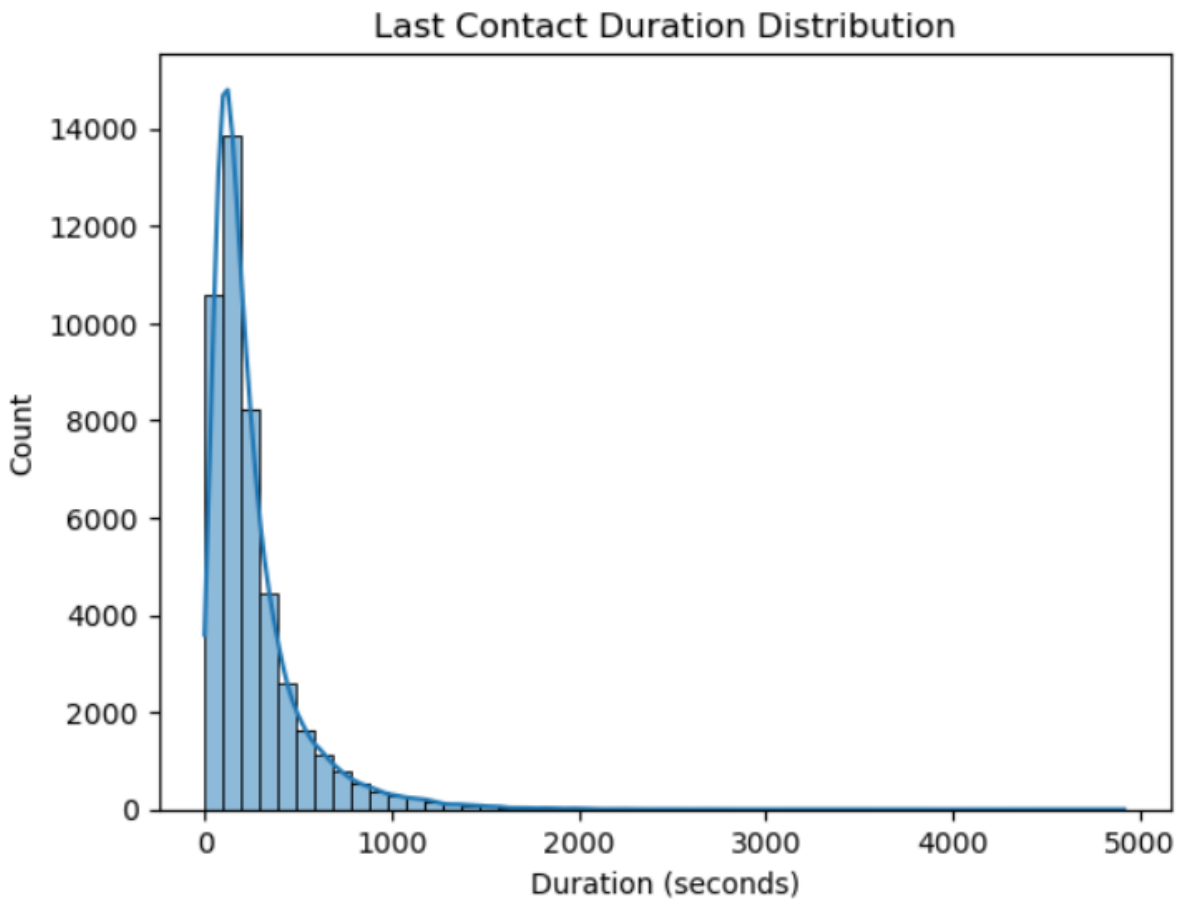


- **What is the distribution of the duration of the last contact?**

  Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file = r"C:\Users\Sarthak
Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

df = pd.read_csv(file)



sns.histplot(df['duration'], bins=50, kde=True)

plt.title("Last Contact Duration Distribution")

plt.xlabel("Duration (seconds)")

plt.ylabel("Count")

plt.show()
```

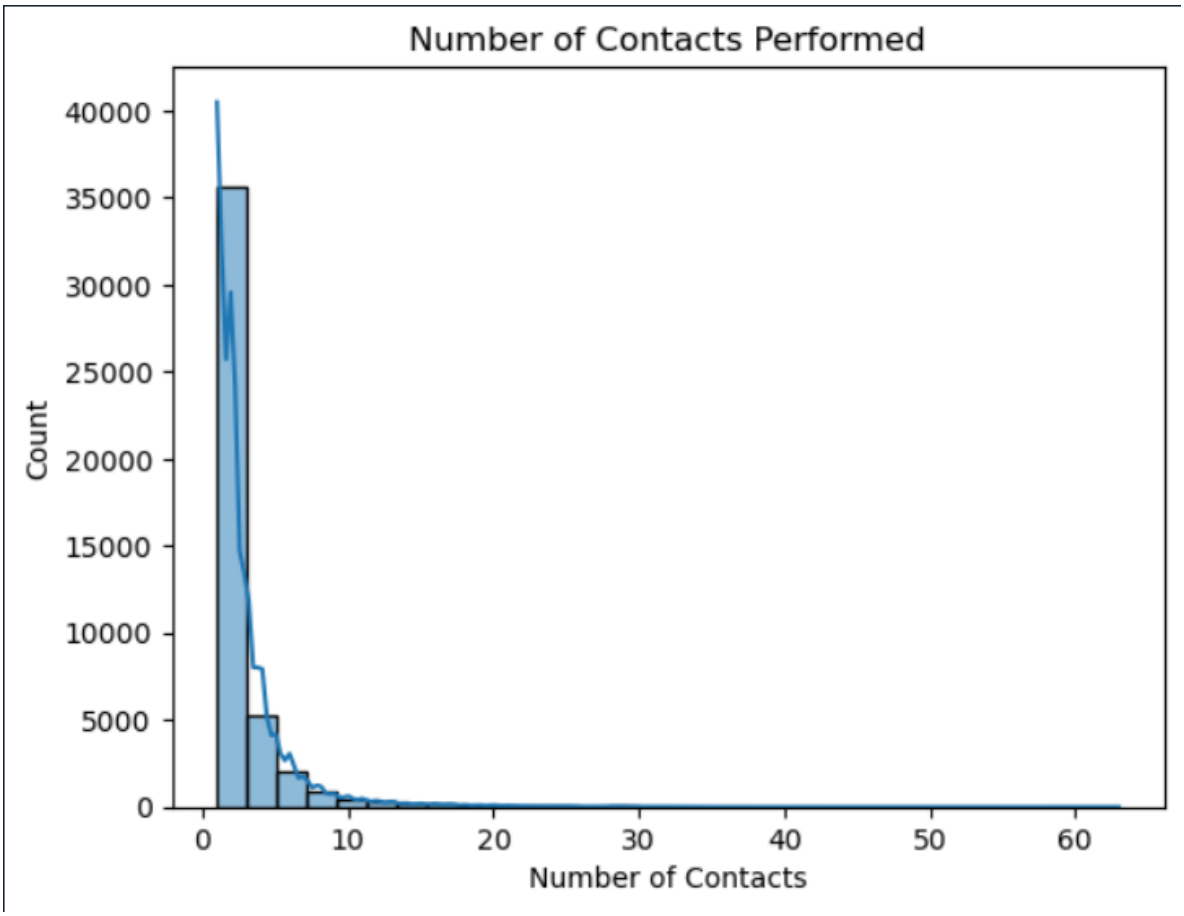
Output:



- **How many contacts were performed during the campaign for each client?**
  Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

df = pd.read_csv(file)



sns.histplot(df['campaign'], bins=30, kde=True)

plt.title("Number of Contacts Performed")

plt.xlabel("Number of Contacts")

plt.ylabel("Count")

plt.show()
```

Output:

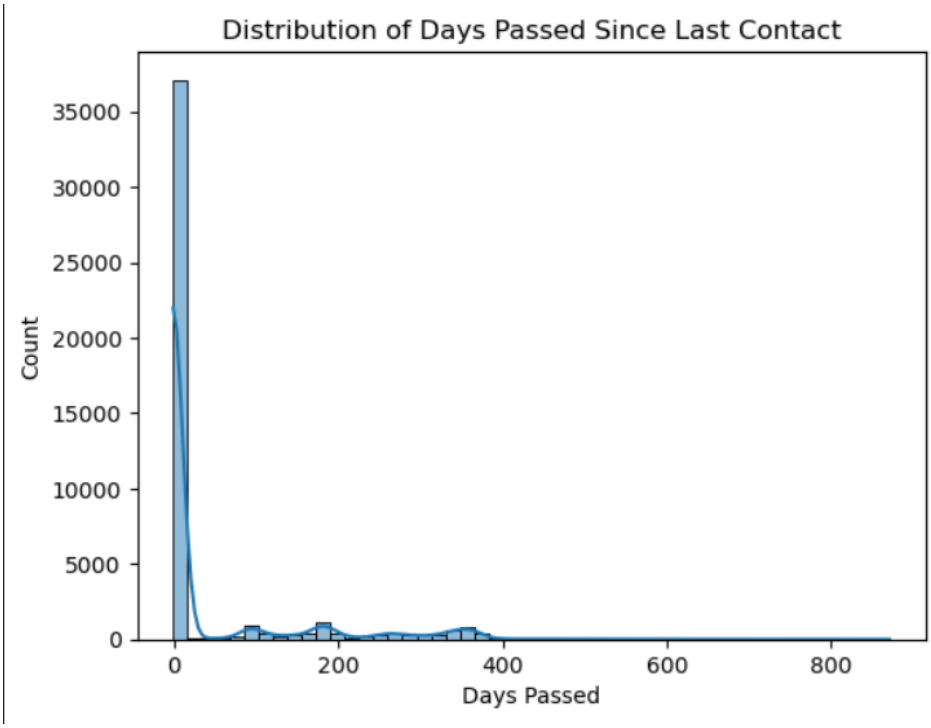


- What is the distribution of the number of days passed since the client was last contacted from a previous campaign?

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file = r"C:\Users\Sarthak
Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

df = pd.read_csv(file)



sns.histplot(df['pdays'], bins=50, kde=True)

plt.title("Distribution of Days Passed Since Last Contact")

plt.xlabel("Days Passed")

plt.ylabel("Count")

plt.show()
```

Output:

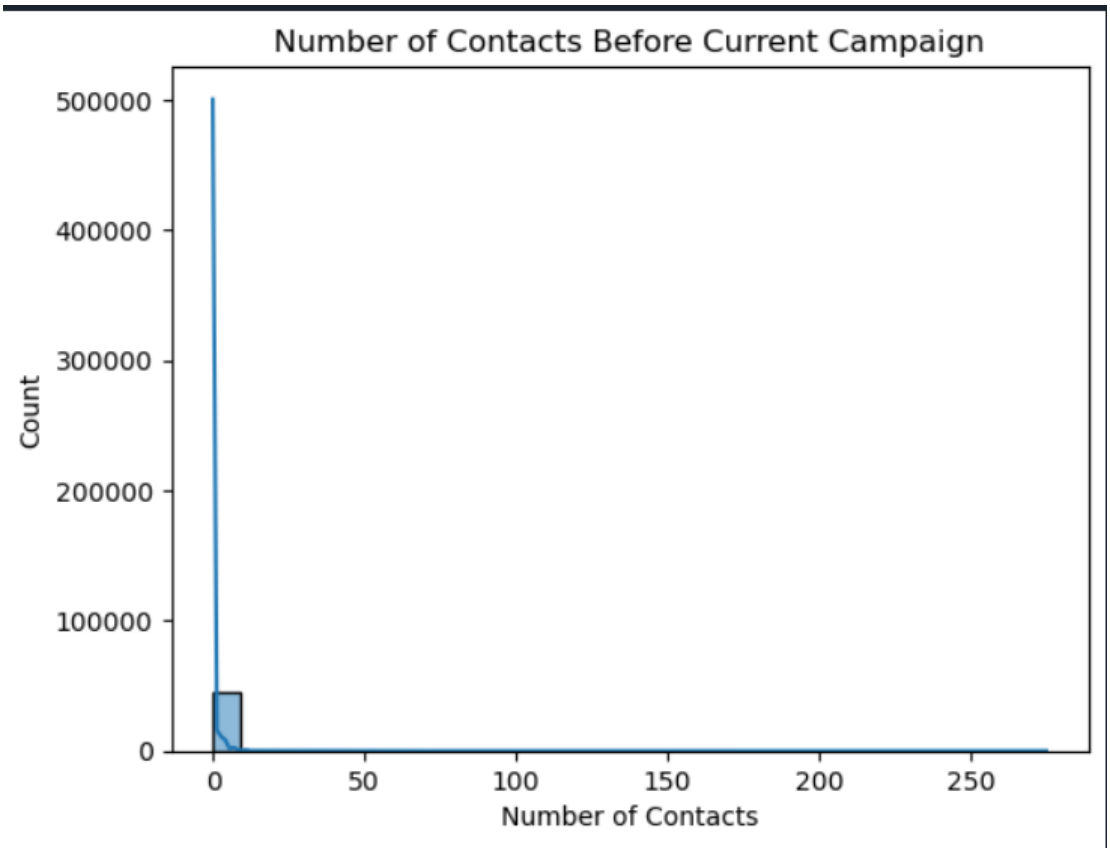


- **How many contacts were performed before the current campaign for each client?**

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

df = pd.read_csv(file)



sns.histplot(df['previous'], bins=30, kde=True)

plt.title("Number of Contacts Before Current Campaign")

plt.xlabel("Number of Contacts")

plt.ylabel("Count")

plt.show()
```

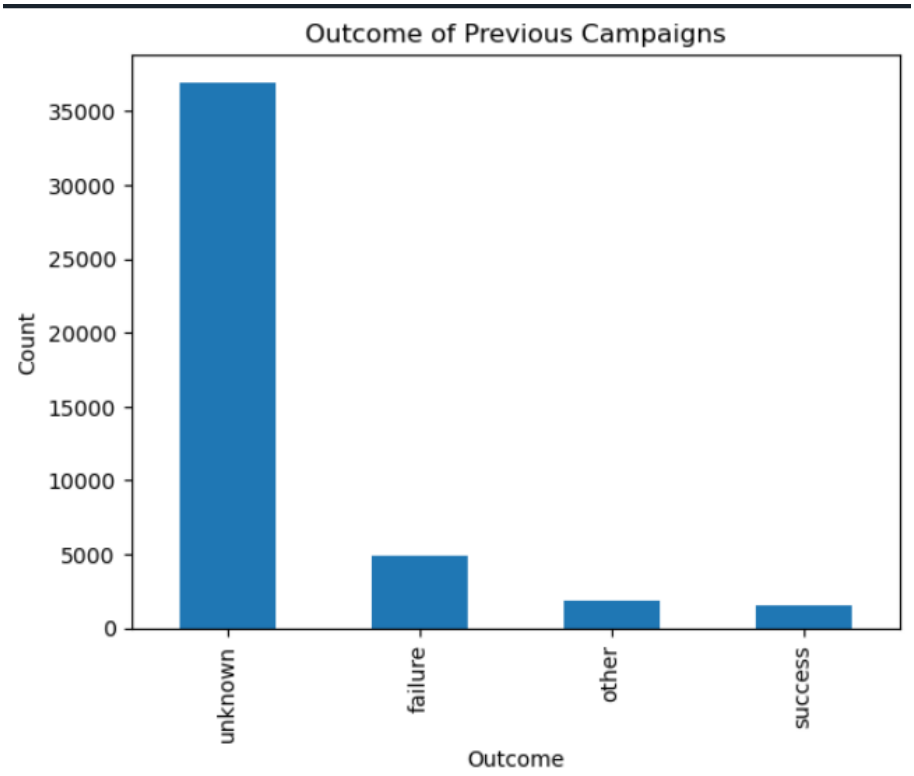
Output:



- **What were the outcomes of the previous marketing campaigns?**

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

df = pd.read_csv(file)



df['poutcome'].value_counts().plot(kind='bar', title="Outcome of Previous Campaigns")

plt.xlabel("Outcome")

plt.ylabel("Count")

plt.show()
```

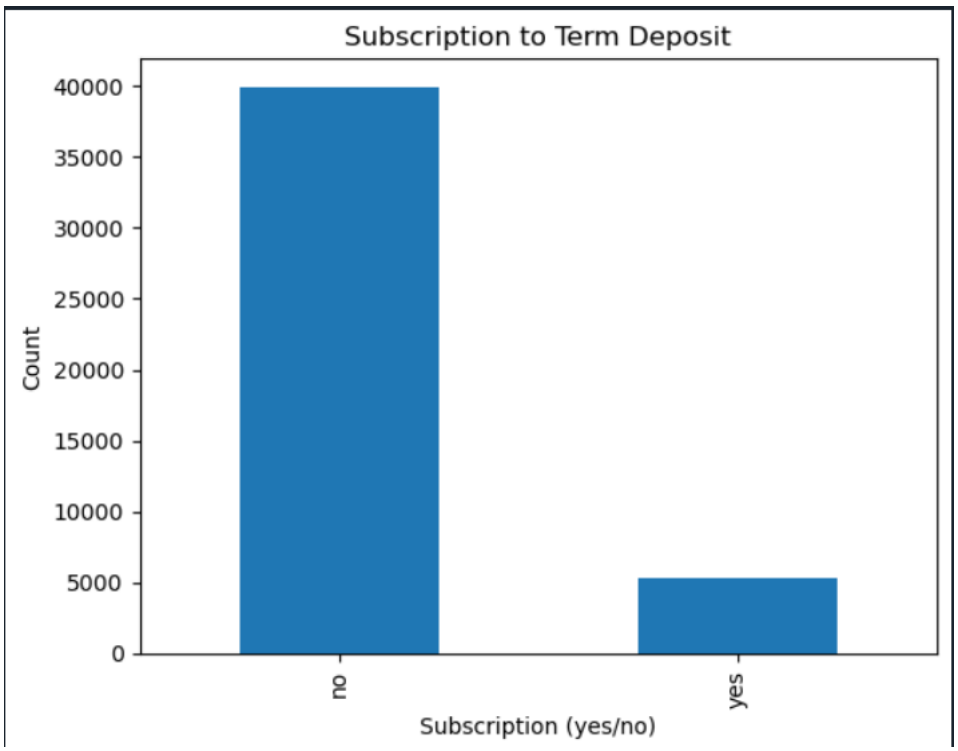


- **What is the distribution of clients who subscribed to a term deposit vs. those who did not?**

Code:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"

df = pd.read_csv(file)



df['y'].value_counts().plot(kind='bar', title="Subscription to Term Deposit")

plt.xlabel("Subscription (yes/no)")

plt.ylabel("Count")

plt.show()
```

Output:

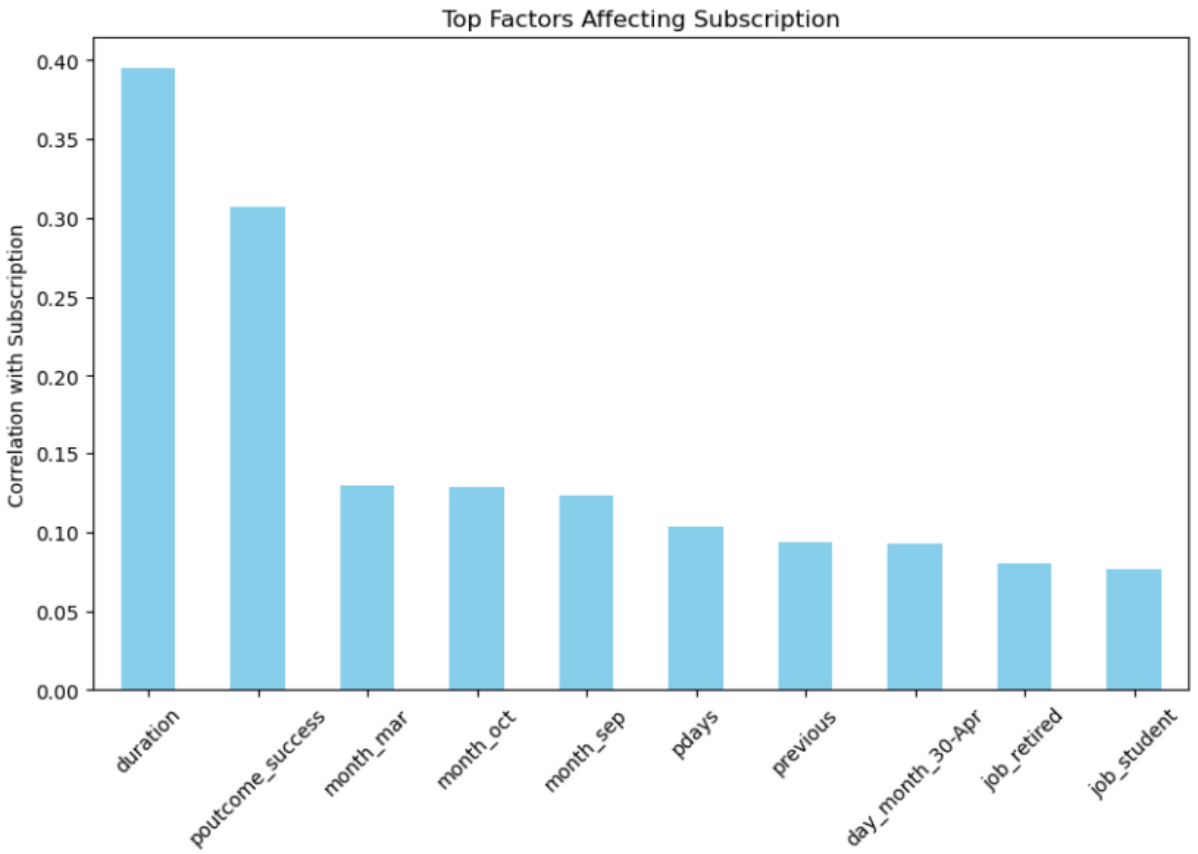


- **Are there any correlations between different attributes and the likelihood of subscribing to a term deposit?**

  Code:

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

# Load the dataset
file = r"C:\Users\Sarthak Das\Downloads\DsResearch\DsResearch\Banking\banking_data.csv"
df = pd.read_csv(file)

df['y'] = df['y'].map({'no': 0, 'yes': 1})

df_encoded = pd.get_dummies(df, drop_first=True)

correlation_matrix = df_encoded.corr()

target_correlation = correlation_matrix["y"].sort_values(ascending=False)

plt.figure(figsize=(10, 6))
target_correlation.drop("y").head(10).plot(kind='bar', color='skyblue')
plt.title("Top Factors Affecting Subscription")
plt.ylabel("Correlation with Subscription")
plt.xticks(rotation=45)
plt.show()
```

Output:


Top Factors Affecting Subscription
Correlation with Subscription
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
duration
poutcome_success
month_mar
month_oct
month_sep
pdays
previous
day_month_30-Apr
job_retired
job_student