FINLATICS - Data Science Experience Program

Case Project - 1

Submitted by:
Name: Anmol Sharma
Email: anmosharma25@gmail.com

Banking Dataset

Term deposits serve as a significant revenue stream for banks, representing cash investments held within financial institutions. These investments involve committing funds for a predetermined period, during which they accrue interest at an agreed-upon rate. To promote term deposits, banks employ various outreach strategies including email marketing, advertisements, telephonic marketing, and digital marketing.

Despite the advent of digital channels, telephonic marketing campaigns persist as one of the most effective means of engaging customers. However, they necessitate substantial investment due to the requirement of large call centers to execute these campaigns. Therefore, it becomes essential to pre-identify potential customers likely to convert, enabling targeted outreach efforts via phone calls.

The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

Content:

The data is related to the direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed by the customer or not. The data folder contains two datasets:-

Banking_data.csv: 45,211 rows and 18 columns ordered by date (from May 2008 to November 2010)

Detailed Column Descriptions:

- **age:** This column represents the age of the bank client. It's a numeric variable indicating the age in years.
- **job:** This column indicates the type of job the client has. It's a categorical variable with options such as "admin.", "unknown", "unemployed", "management", etc.
- **marital:** This column represents the marital status of the client. It's a categorical variable with options such as "married", "divorced", or "single".

- **education:** This column indicates the level of education of the client. It's a categorical variable with options such as "unknown", "secondary", "primary", or "tertiary".
- **default:** This column indicates whether the client has credit in default. It's a binary variable with options "yes" or "no".
- **balance:** This column represents the average yearly balance in euros for the client. It's a numeric variable.
- **housing:** This column indicates whether the client has a housing loan. It's a binary variable with options "yes" or "no".
- **loan:** This column indicates whether the client has a personal loan. It's a binary variable with options "yes" or "no".
- **contact:** This column represents the type of communication used to contact the client. It's a categorical variable with options such as "unknown", "telephone", or "cellular".
- **day:** This column represents the last contact day of the month. It's a numeric variable.
- **month:** This column represents the last contact month of the year. It's a categorical variable with options such as "jan", "feb", "mar", etc.
- **duration:** This column represents the duration of the last contact in seconds. It's a numeric variable.
- **campaign:** This column represents the number of contacts performed during this campaign and for this client. It's a numeric variable.
- **pdays:** This column represents the number of days that passed by after the client was last contacted from a previous campaign. It's a numeric variable where -1 means the client was not previously contacted.
- **previous:** This column represents the number of contacts performed before this campaign and for this client. It's a numeric variable.
- **poutcome:** This column represents the outcome of the previous marketing campaign. It's a categorical variable with options such as "unknown", "other", "failure", or "success".
- **y:** This column is the target variable and indicates whether the client has subscribed to a term deposit. It's a binary variable with options "yes" or "no".

Questions:

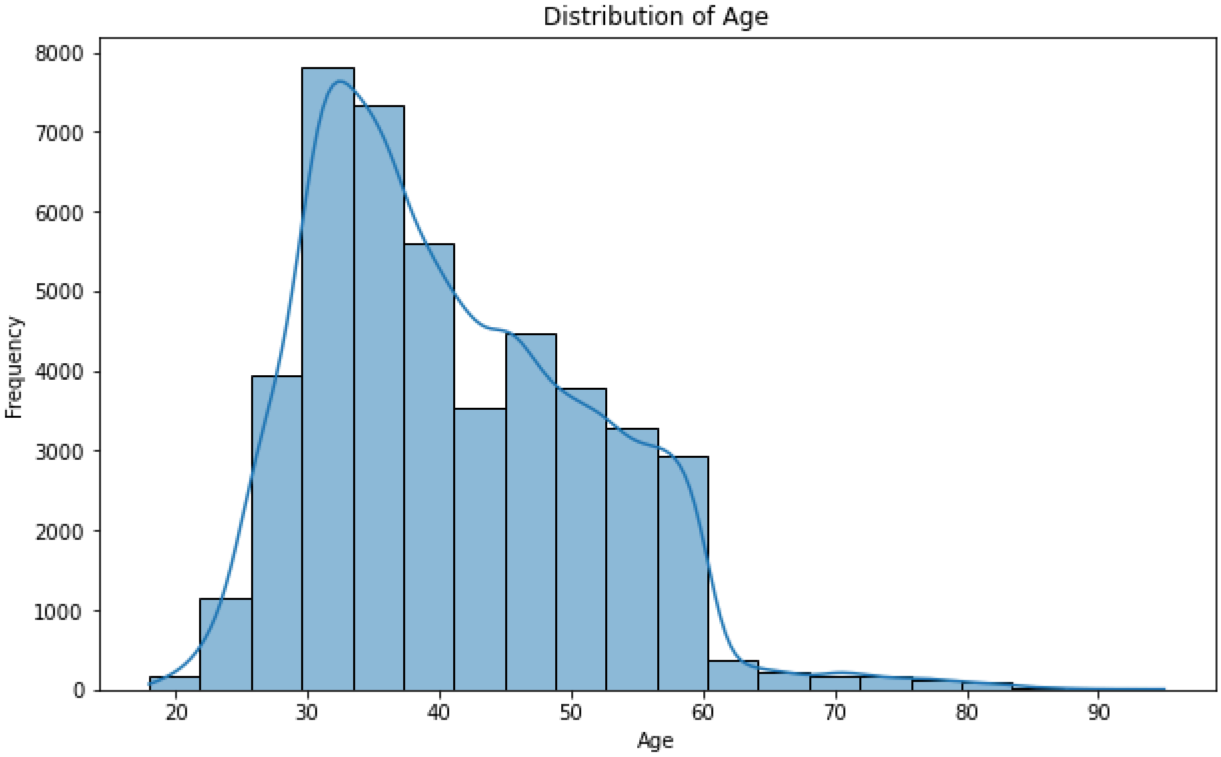
1. What is the distribution of age among the clients?
2. How does the job type vary among the clients?
3. What is the marital status distribution of the clients?
4. What is the level of education among the clients?
5. What proportion of clients have credit in default?
6. What is the distribution of average yearly balance among the clients?
7. How many clients have housing loans?
8. How many clients have personal loans?
9. What are the communication types used for contacting clients during the campaign?
10. What is the distribution of the last contact day of the month?

11. How does the last contact month vary among the clients?
12. What is the distribution of the duration of the last contact?
13. How many contacts were performed during the campaign for each client?
14. What is the distribution of the number of days passed since the client was last contacted from a previous campaign?
15. How many contacts were performed before the current campaign for each client?
16. What were the outcomes of the previous marketing campaigns?
17. What is the distribution of clients who subscribed to a term deposit vs. those who did not?
18. Are there any correlations between different attributes and the likelihood of subscribing to a term deposit?

Answers

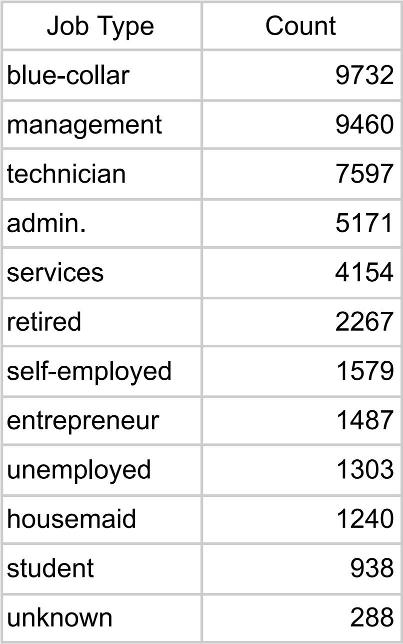
Q1) What is the distribution of age among the clients?

Answer:

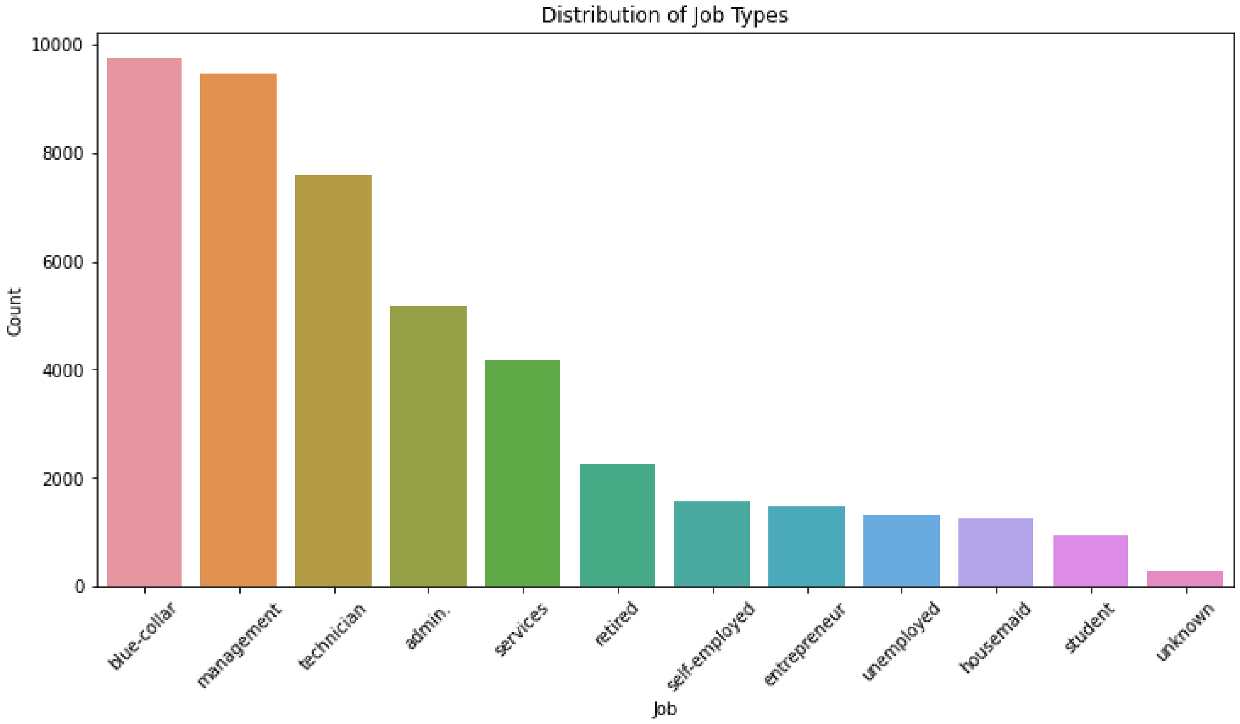


Q2) How does the job type vary among the clients?

Answer:

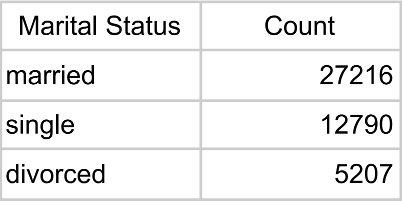


Job Type	Count
blue-collar	9732
management	9460
technician	7597
admin.	5171
services	4154
retired	2267
self-employed	1579
entrepreneur	1487
unemployed	1303
housemaid	1240
student	938
unknown	288

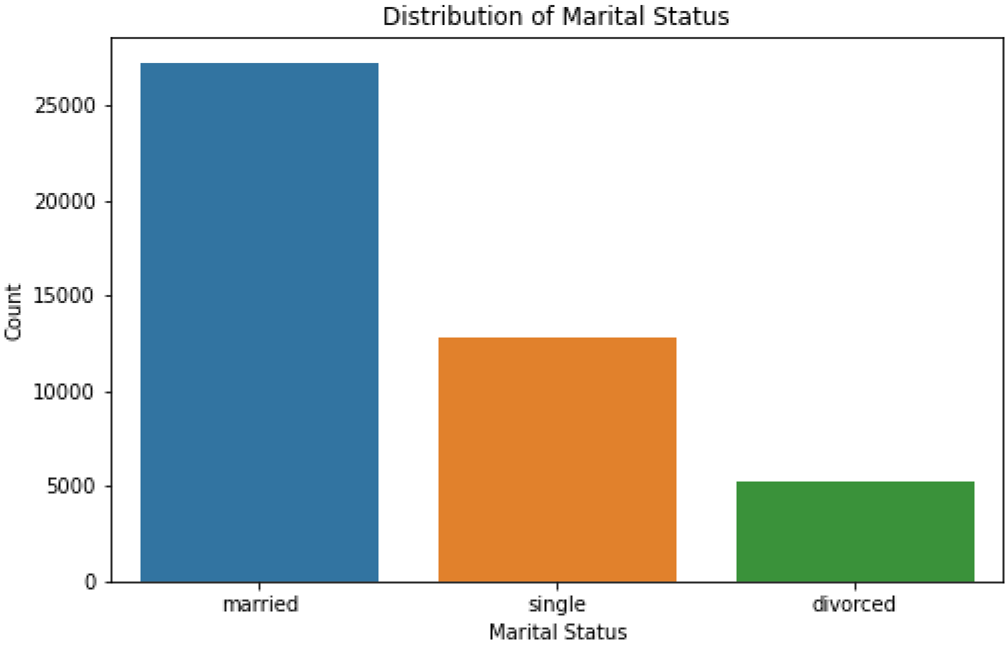


Q3) What is the marital status distribution of the clients?

Answer:



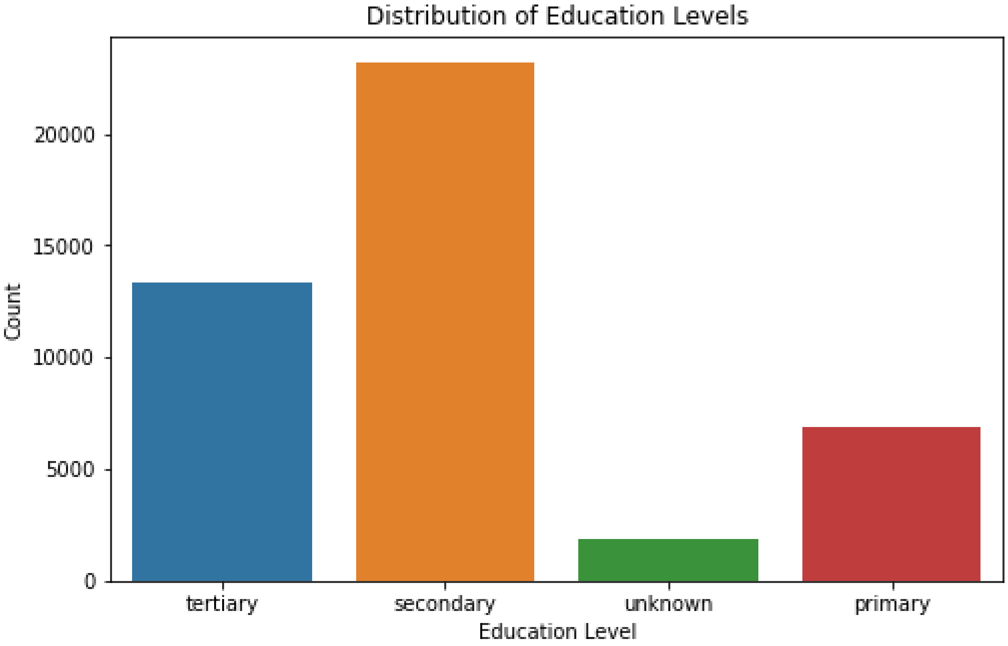
Marital Status	Count
married	27216
single	12790
divorced	5207



Q4) What is the level of education among the clients?

Answer:

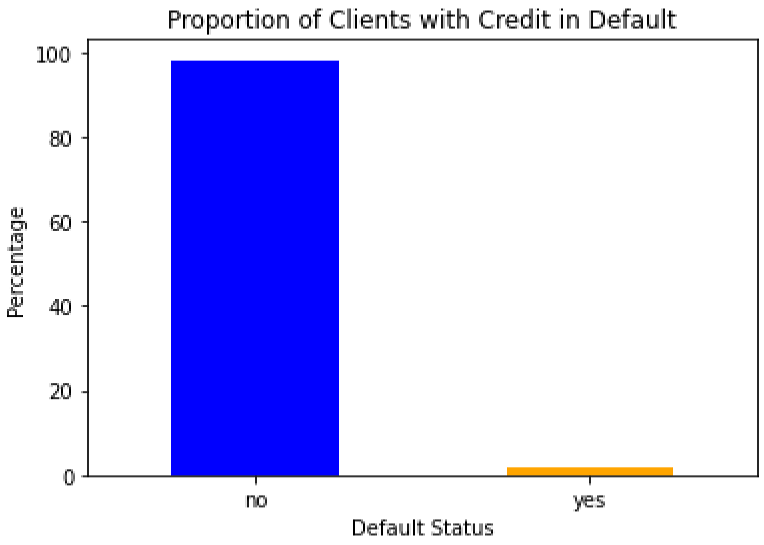
Level of Education	Count
secondary	23204
tertiary	13301
primary	6851
unknown	1857



Q5) What proportion of clients have credit in default?

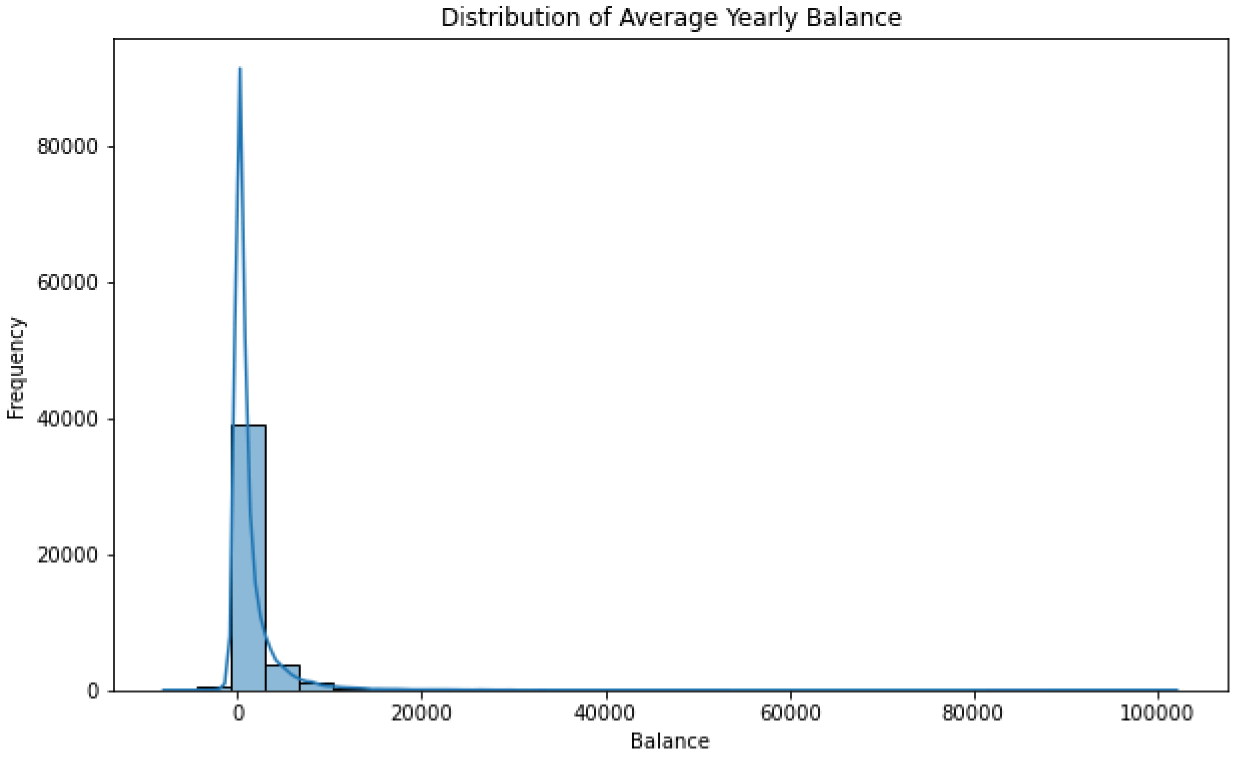
Answer:

Have credit in default?	%
no	98.197541
yes	1.802459



Q6) What is the distribution of average yearly balance among the clients?

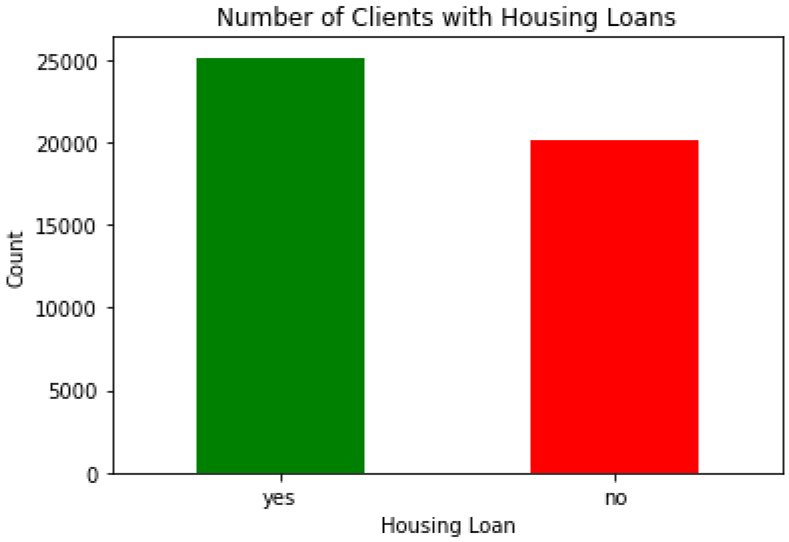
<u>Answer:</u>



Q7) How many clients have housing loans?

Answer:

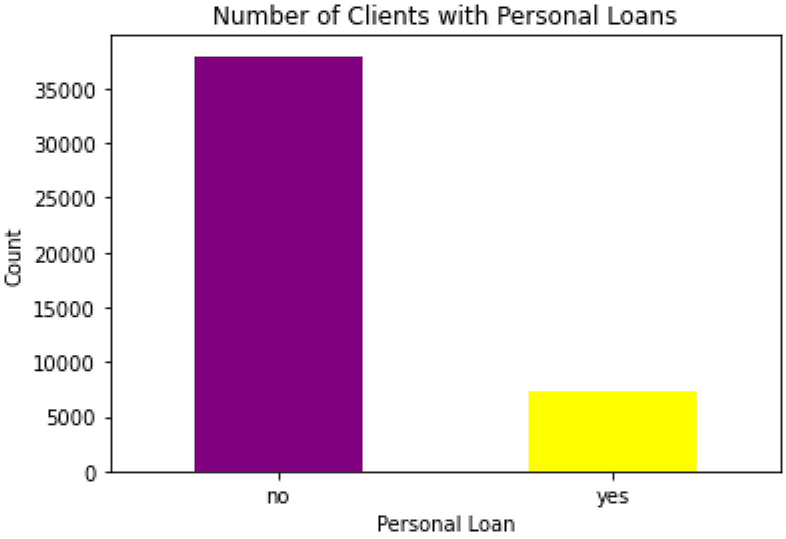
Housing loans?	Count
yes	25130
no	20086



Q8) How many clients have personal loans?

Answer:

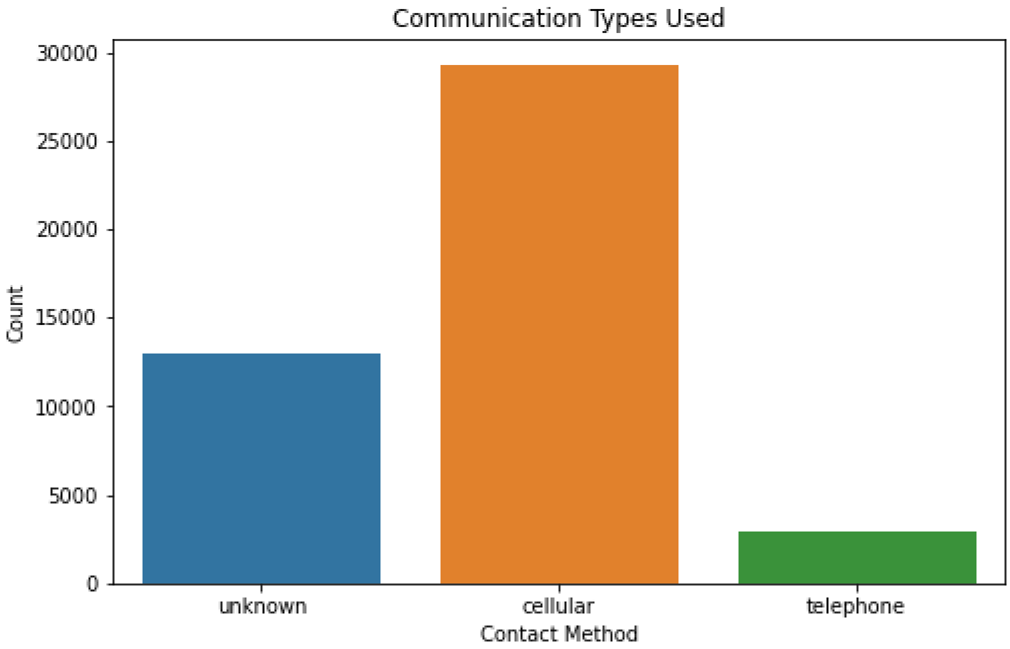
Personal loan?	Count
no	37972
yes	7244



Q9) - What are the communication types used for contacting clients during the campaign?

Answer:

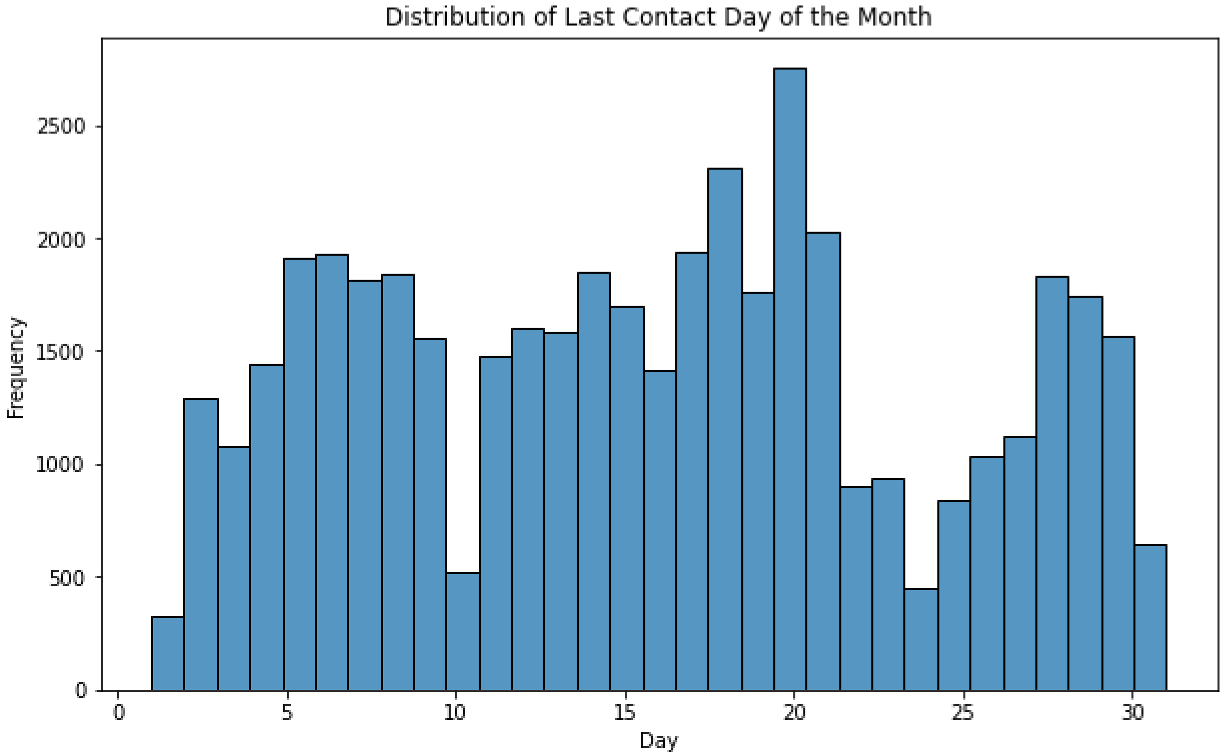
Communication types	Count
cellular	29290
unknown	13020
telephone	2906



Q10) - What is the distribution of the last contact day of the month?

Answer:

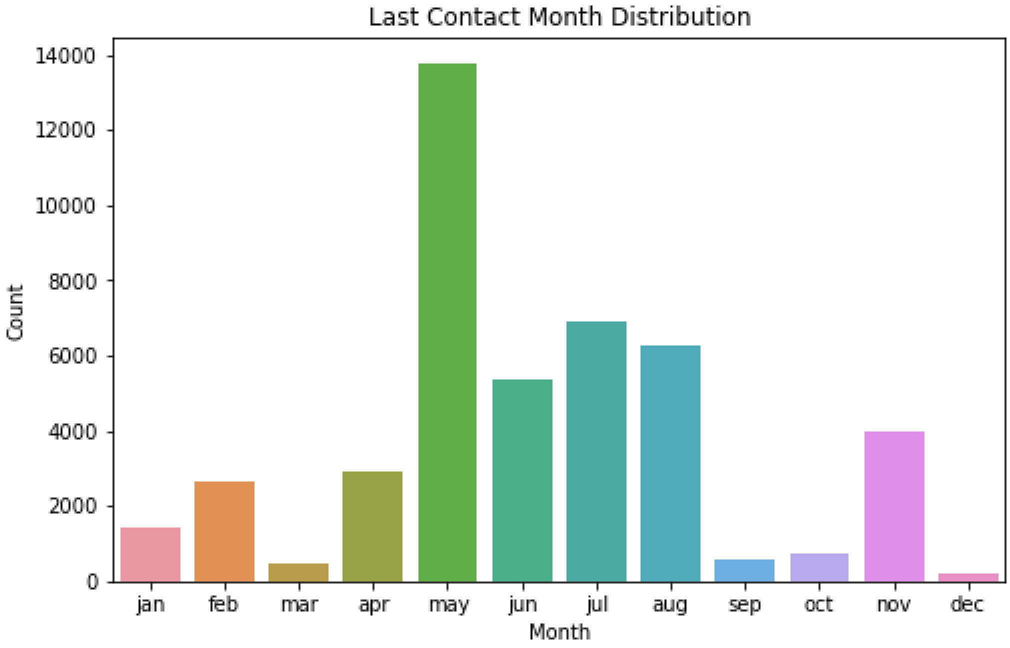
day	Frequency
1	322
2	1293
3	1079
4	1445
5	1910
6	1932
7	1817
8	1842
9	1561
10	524
11	1479
12	1603
13	1585
14	1848
15	1703
16	1417
17	1942
18	2308
19	1757
20	2752
21	2026
22	905
23	939
24	447
25	840
26	1035
27	1121
28	1830
29	1745
30	1566
31	643



Q11) How does the last contact month vary among the clients?

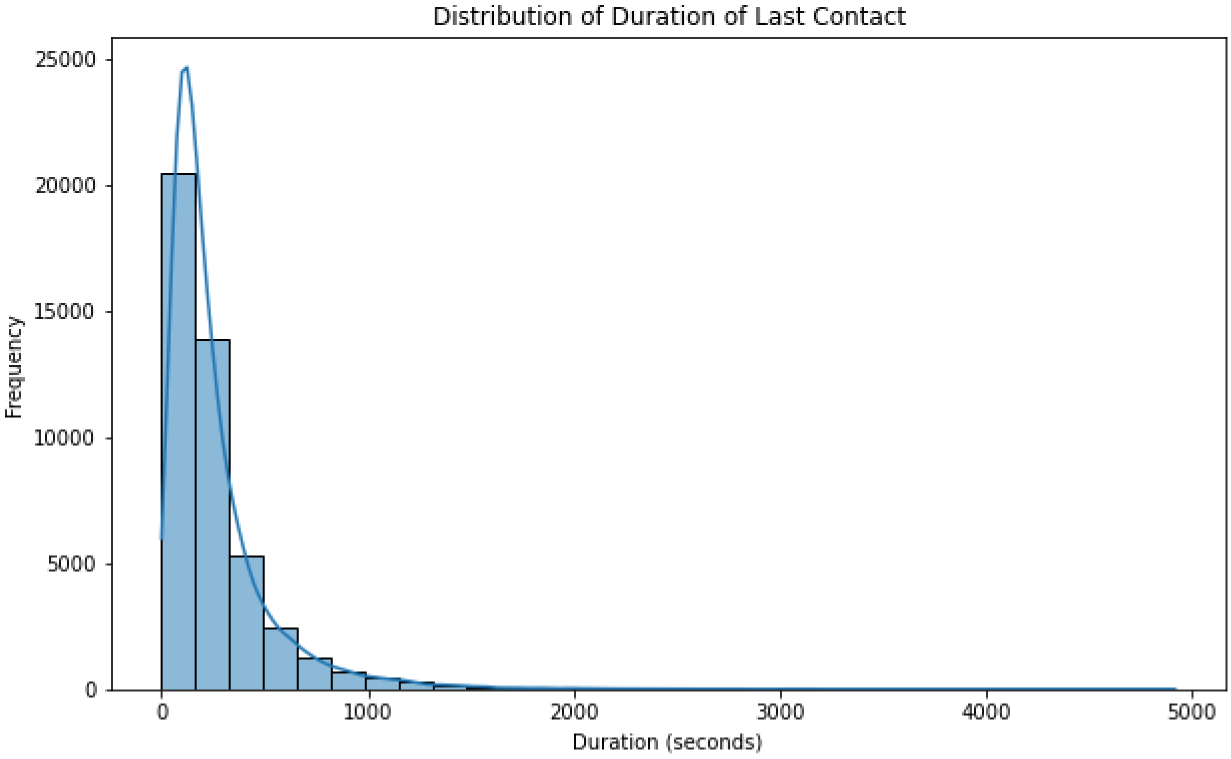
Answer:

month	count
Jan	1403
Feb	2649
mar	477
Apr	2932
may	13766
Jun	5341
Jul	6895
Aug	6247
sep	579
oct	738
Nov	3975
dec	214



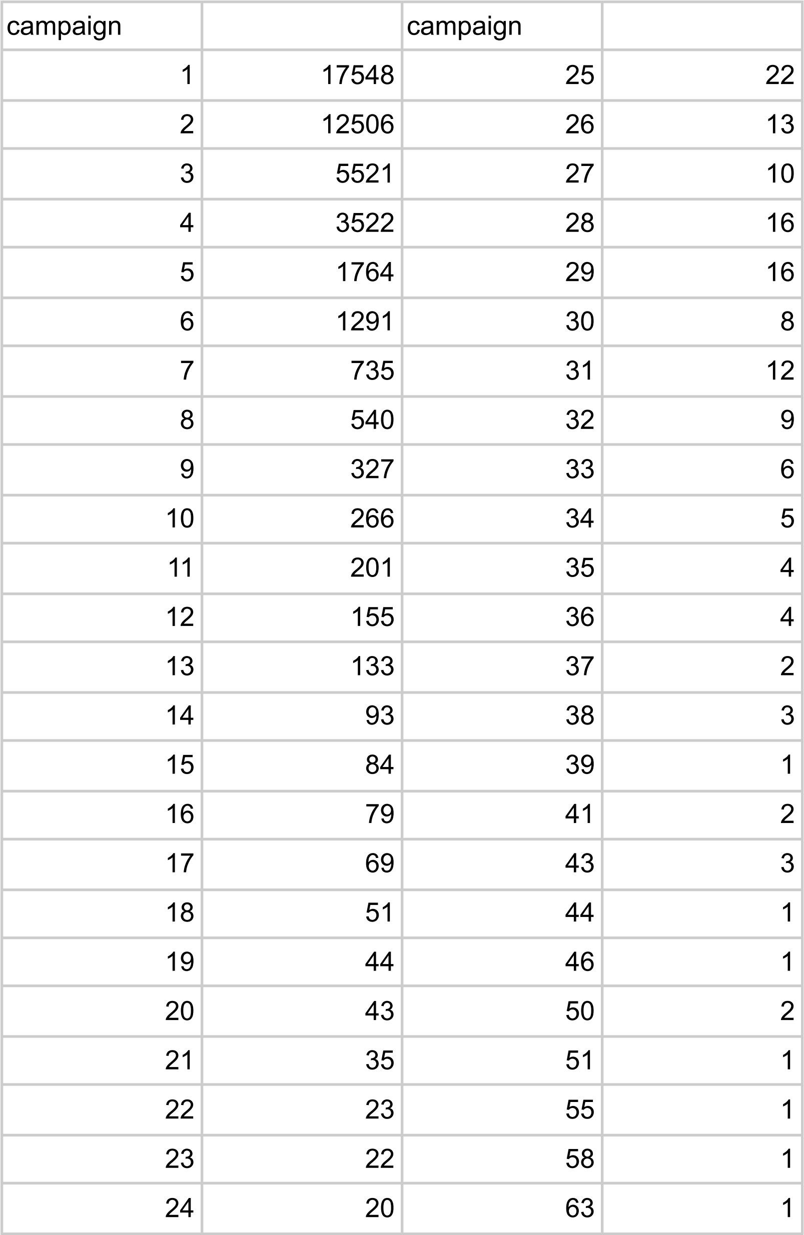
Q12) What is the distribution of the duration of the last contact?

Answer:

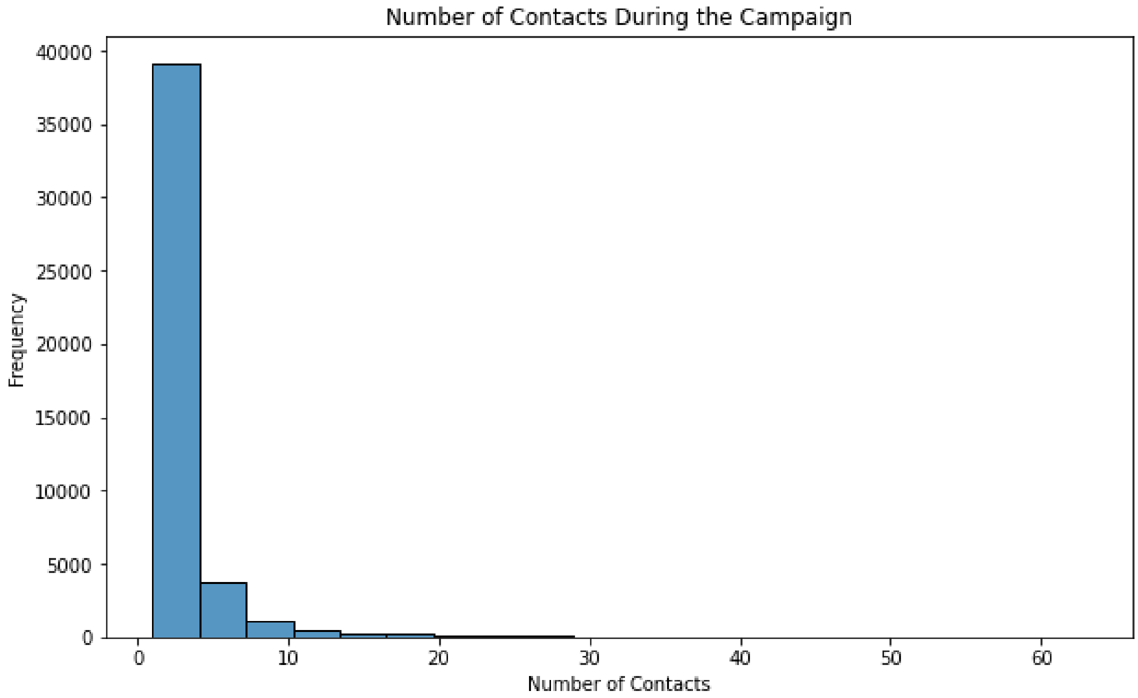


Q13) How many contacts were performed during the campaign for each client?

Answer:

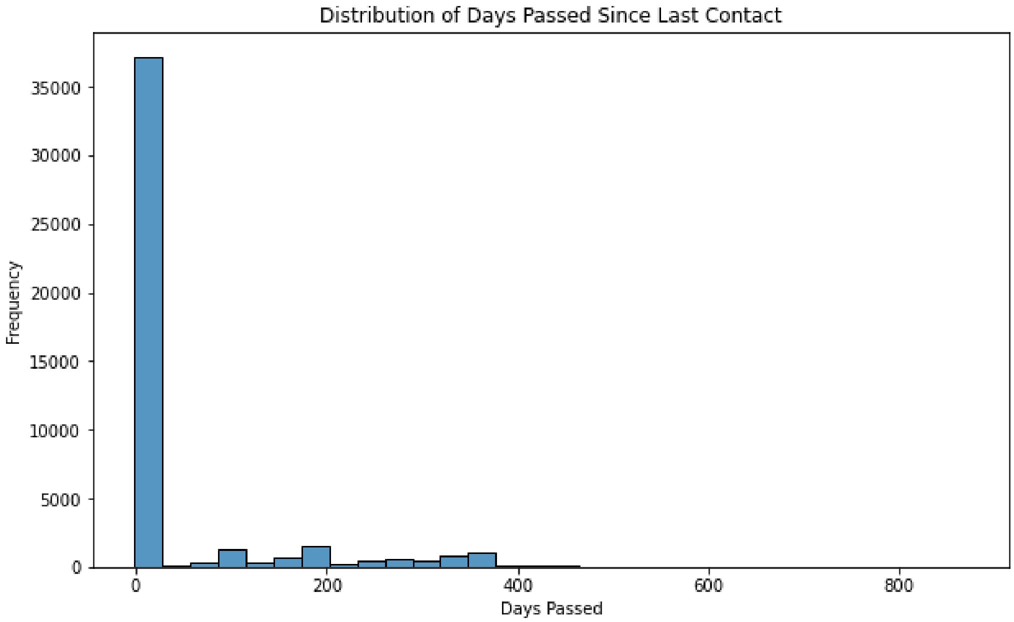


campaign		campaign	
1	17548	25	22
2	12506	26	13
3	5521	27	10
4	3522	28	16
5	1764	29	16
6	1291	30	8
7	735	31	12
8	540	32	9
9	327	33	6
10	266	34	5
11	201	35	4
12	155	36	4
13	133	37	2
14	93	38	3
15	84	39	1
16	79	41	2
17	69	43	3
18	51	44	1
19	44	46	1
20	43	50	2
21	35	51	1
22	23	55	1
23	22	58	1
24	20	63	1



Q14) What is the distribution of the number of days passed since the client was last contacted from a previous campaign?

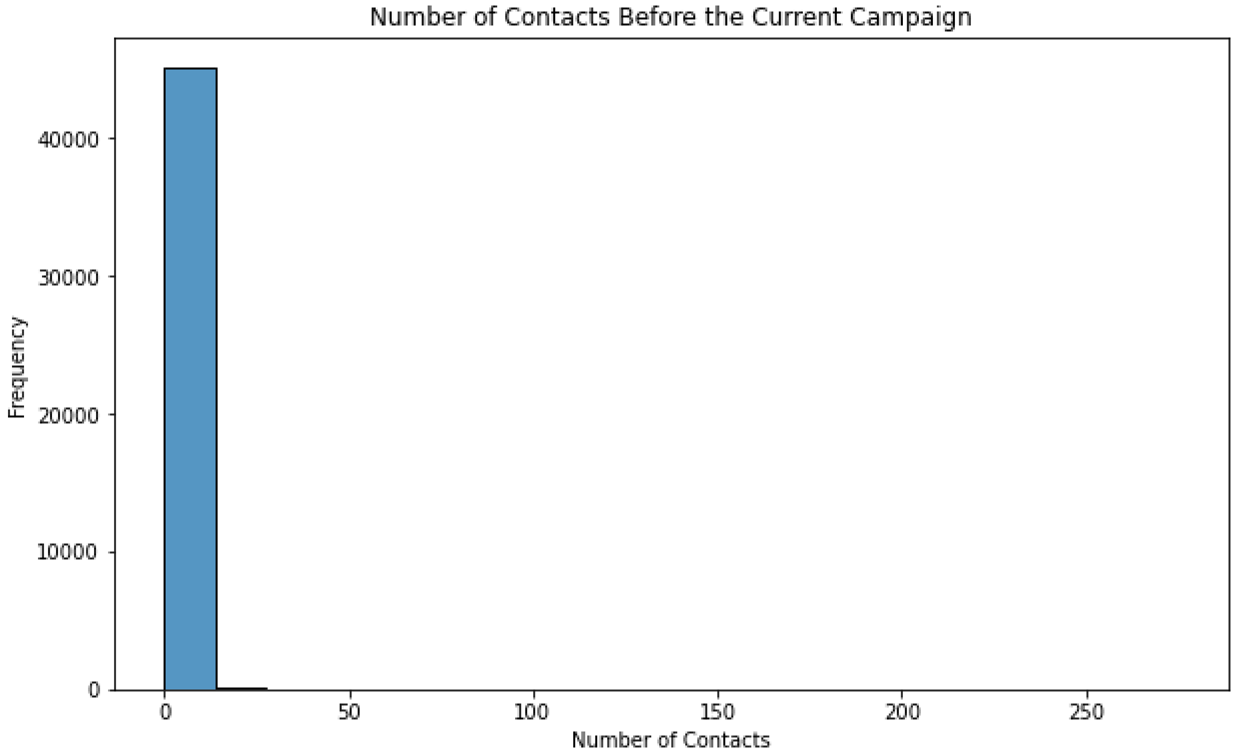
Answer:



Q15) How many contacts were performed before the current campaign for each client?

Answer:

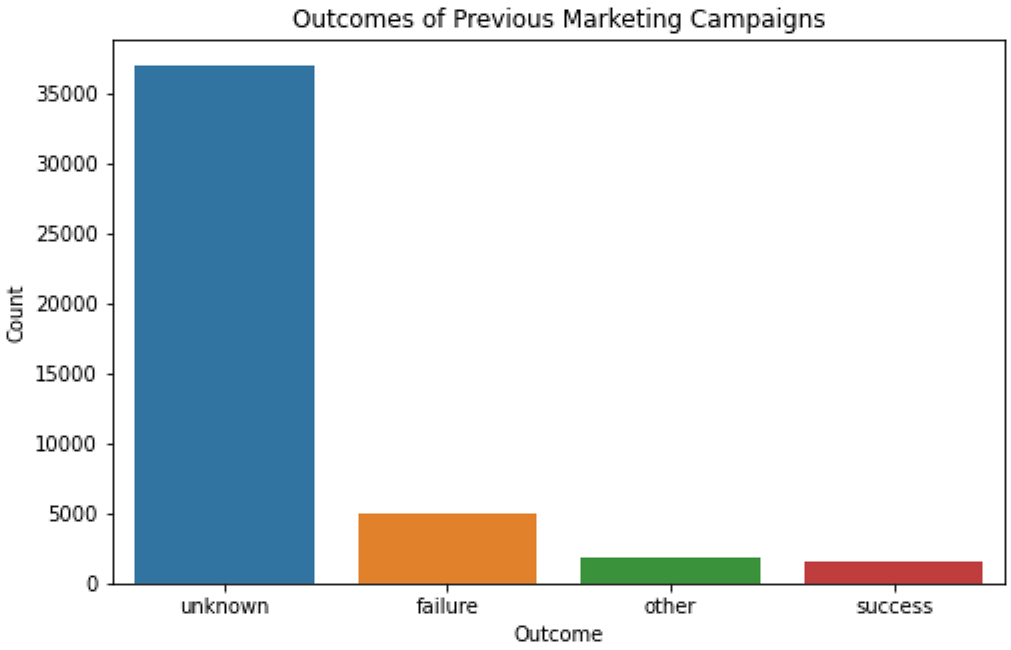
previous		previous	
0	36956	18	6
1	2772	22	6
2	2106	24	5
3	1142	27	5
4	715	21	4
5	459	29	4
6	278	25	4
7	205	30	3
8	130	38	2
9	92	37	2
10	67	26	2
11	65	28	2
12	44	51	1
13	38	275	1
15	20	58	1
14	19	32	1
17	15	40	1
16	13	55	1
19	11	35	1
20	8	41	1
23	8		



Q16) What were the outcomes of the previous marketing campaigns?

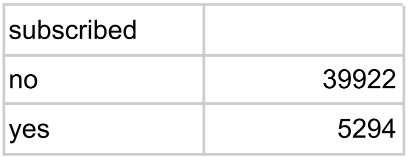
Answer:

poutcome	
unknown	36961
failure	4902
other	1840
success	1513

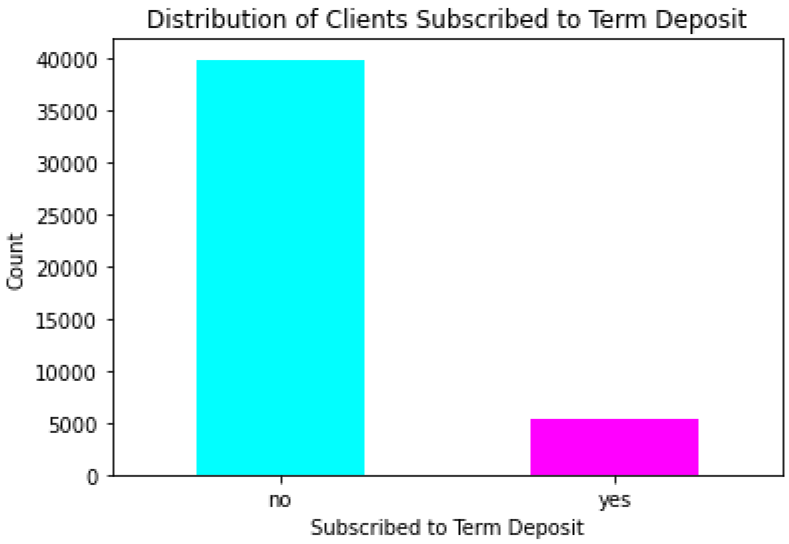


Q17) What is the distribution of clients who subscribed to a term deposit vs. those who did not?

Answer:



subscribed	
no	39922
yes	5294



Q18) Are there any correlations between different attributes and the likelihood of subscribing to a term deposit?

Answer:

attributes	coorelation
age	0.025648
balance	0.052821
day	-0.028307
duration	0.394387
campaign	-0.073294
pdays	0.103699
previous	0.093576

Python Code

Program Code:

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

df = pd.read_csv('Banking_data.csv')

# Replace infinite values with NaN
df.replace([np.inf, -np.inf], np.nan, inplace=True)

# 1. Distribution of age among the clients
print("1. What is the distribution of age among the clients?")
plt.figure(figsize=(10, 6))
sns.histplot(df['age'], bins=20, kde=True)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

# 2. Variation in job types among the clients
print("2. How does the job type vary among the clients?")
print(df['job'].value_counts())
plt.figure(figsize=(12, 6))
sns.countplot(x='job', data=df, order=df['job'].value_counts().index)
plt.title('Distribution of Job Types')
plt.xlabel('Job')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

# 3. Marital status distribution of the clients
print("3. What is the marital status distribution of the clients?")
print(df['marital'].value_counts())
plt.figure(figsize=(8, 5))
sns.countplot(x='marital', data=df)
plt.title('Distribution of Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.show()

# 4. Level of education among the clients
print("4. What is the level of education among the clients?")
print(df['education'].value_counts())
plt.figure(figsize=(8, 5))
sns.countplot(x='education', data=df)
plt.title('Distribution of Education Levels')
plt.xlabel('Education Level')
plt.ylabel('Count')
plt.show()

# 5. Proportion of clients with credit in default
print("5. What proportion of clients have credit in default?")
default_counts = df['default'].value_counts(normalize=True) * 100
print(default_counts)
plt.figure(figsize=(6, 4))
default_counts.plot(kind='bar', color=['blue', 'orange'])
plt.title('Proportion of Clients with Credit in Default')
plt.xlabel('Default Status')
plt.ylabel('Percentage')
plt.xticks(rotation=0)
plt.show()

# 6. Distribution of average yearly balance among the clients
print("6. What is the distribution of average yearly balance among the clients?")
plt.figure(figsize=(10, 6))
sns.histplot(df['balance'], bins=30, kde=True)
plt.title('Distribution of Average Yearly Balance')
plt.xlabel('Balance')
plt.ylabel('Frequency')
plt.show()

# 7. Number of clients with housing loans
print("7. How many clients have housing loans?")
print(df['housing'].value_counts())
plt.figure(figsize=(6, 4))
df['housing'].value_counts().plot(kind='bar', color=['green', 'red'])
```

```
plt.title('Number of Clients with Housing Loans')
plt.xlabel('Housing Loan')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.show()

# 8. Number of clients with personal loans
print("8. How many clients have personal loans?")
print(df['loan'].value_counts())
plt.figure(figsize=(6, 4))
df['loan'].value_counts().plot(kind='bar', color=['purple', 'yellow'])
plt.title('Number of Clients with Personal Loans')
plt.xlabel('Personal Loan')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.show()

# 9. Communication types used for contacting clients during the campaign
print("9. What are the communication types used for contacting clients during the campaign?")
print(df['contact'].value_counts())
plt.figure(figsize=(8, 5))
sns.countplot(x='contact', data=df)
plt.title('Communication Types Used')
plt.xlabel('Contact Method')
plt.ylabel('Count')
plt.show()

# 10. Distribution of the last contact day of the month
print("10. What is the distribution of the last contact day of the month?")
print(df['day'].value_counts())
plt.figure(figsize=(10, 6))
sns.histplot(df['day'], bins=31, kde=False)
plt.title('Distribution of Last Contact Day of the Month')
plt.xlabel('Day')
plt.ylabel('Frequency')
plt.show()

# 11. Variation in last contact month among the clients
print("11. How does the last contact month vary among the clients?")
print(df['month'].value_counts())
```

```
plt.figure(figsize=(8, 5))
sns.countplot(x='month', data=df, order=['jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct',
'nov', 'dec'])
plt.title('Last Contact Month Distribution')
plt.xlabel('Month')
plt.ylabel('Count')
plt.show()

# 12. Distribution of the duration of the last contact
print("12. What is the distribution of the duration of the last contact?")
plt.figure(figsize=(10, 6))
sns.histplot(df['duration'], bins=30, kde=True)
plt.title('Distribution of Duration of Last Contact')
plt.xlabel('Duration (seconds)')
plt.ylabel('Frequency')
plt.show()

# 13. Number of contacts performed during the campaign for each client
print("13. How many contacts were performed during the campaign for each client?")
print(df['campaign'].value_counts())
plt.figure(figsize=(10, 6))
sns.histplot(df['campaign'], bins=20, kde=False)
plt.title('Number of Contacts During the Campaign')
plt.xlabel('Number of Contacts')
plt.ylabel('Frequency')
plt.show()

# 14. Distribution of the number of days passed since the client was last contacted from a
previous campaign
print("14. What is the distribution of the number of days passed since the client was last
contacted from a previous campaign?")
plt.figure(figsize=(10, 6))
sns.histplot(df['pdays'], bins=30, kde=False)
plt.title('Distribution of Days Passed Since Last Contact')
plt.xlabel('Days Passed')
plt.ylabel('Frequency')
plt.show()

# 15. Number of contacts performed before the current campaign for each client
print("15. How many contacts were performed before the current campaign for each client?")
print(df['previous'].value_counts())
plt.figure(figsize=(10, 6))
sns.histplot(df['previous'], bins=20, kde=False)
plt.title('Number of Contacts Before the Current Campaign')
plt.xlabel('Number of Contacts')
plt.ylabel('Frequency')
plt.show()

# 16. Outcomes of the previous marketing campaigns
print("16. What were the outcomes of the previous marketing campaigns?")
print(df['poutcome'].value_counts())
plt.figure(figsize=(8, 5))
sns.countplot(x='poutcome', data=df)
plt.title('Outcomes of Previous Marketing Campaigns')
plt.xlabel('Outcome')
plt.ylabel('Count')
plt.show()

# 17. Distribution of clients who subscribed to a term deposit vs. those who did not
print("17. What is the distribution of clients who subscribed to a term deposit vs. those who did
not?")
print(df['y'].value_counts())
plt.figure(figsize=(6, 4))
df['y'].value_counts().plot(kind='bar', color=['cyan', 'magenta'])
plt.title('Distribution of Clients Subscribed to Term Deposit')
plt.xlabel('Subscribed to Term Deposit')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.show()

# 18. Correlations between different attributes and the likelihood of subscribing to a term deposit
print("18. Are there any correlations between different numeric attributes and the likelihood of
subscribing to a term deposit?")
numeric_columns = df.select_dtypes(include=['int64', 'float64']).columns
correlation_matrix = df[numeric_columns].corrwith(df['y'].map({'yes': 1, 'no': 0}))
print(correlation_matrix)
```

Output (Without plots):

1. What is the distribution of age among the clients?

```
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

2. How does the job type vary among the clients?

```
job
blue-collar      9732
management       9460
technician       7597
admin.           5171
services         4154
retired          2267
self-employed    1579
entrepreneur     1487
unemployed       1303
housemaid        1240
student           938
unknown           288
Name: count, dtype: int64
```

3. What is the marital status distribution of the clients?

```
marital
married     27216
single      12790
divorced     5207
Name: count, dtype: int64
```

4. What is the level of education among the clients?

```
education
secondary    23204
tertiary     13301
primary       6851
unknown       1857
Name: count, dtype: int64
```

5. What proportion of clients have credit in default?

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

6. What is the distribution of average yearly balance among the clients?

```
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

7. How many clients have housing loans?

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

￼

8. How many clients have personal loans?

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

9. What are the communication types used for contacting clients during the campaign?

```
contact
cellular     29290
unknown      13020
telephone     2906
Name: count, dtype: int64
```

￼10. What is the distribution of the last contact day of the month?

```
day
20    2752
18    2308
21    2026
17    1942
6     1932
5     1910
14    1848
8     1842
```

```
28    1830
7     1817
19    1757
29    1745
15    1703
12    1603
13    1585
30    1566
9     1561
11    1479
4     1445
16    1417
2     1293
27    1121
3     1079
26    1035
23     939
22     905
25     840
31     643
10     524
24     447
1      322
Name: count, dtype: int64
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

￼

11. How does the last contact month vary among the clients?

```
month
may    13766
jul     6895
aug     6247
jun     5341
nov     3975
apr     2932
feb     2649
jan     1403
oct      738
```

```
sep     579
mar      477
dec     214
Name: count, dtype: int64
```

12. What is the distribution of the duration of the last contact?

```
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

13. How many contacts were performed during the campaign for each client?

```
campaign
1     17548
2     12506
3      5521
4      3522
5      1764
6      1291
7       735
8       540
9       327
10      266
11      201
12      155
13      133
14       93
15       84
16       79
17       69
18       51
19       44
20       43
21       35
22       23
25       22
23       22
24       20
29       16
28       16
```

```
26      13
31      12
27      10
32       9
30       8
33       6
34       5
36       4
35       4
43       3
38       3
37       2
50       2
41       2
46       1
58       1
55       1
63       1
51       1
39       1
44       1
Name: count, dtype: int64
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

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

```
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

15. How many contacts were performed before the current campaign for each client?

```
previous
0     36956
1      2772
2      2106
3      1142
```

```
4      715
5      459
6      278
7      205
8      130
9       92
10      67
11      65
12      44
13      38
15      20
14      19
17      15
16      13
19      11
20       8
23       8
18       6
22       6
24       5
27       5
21       4
29       4
25       4
30       3
38       2
37       2
26       2
28       2
51       1
275      1
58       1
32       1
40       1
55       1
35       1
41       1
Name: count, dtype: int64
```

```
C:\Users\hp\anaconda3\envs\Finlatics\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version.
Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

16. What were the outcomes of the previous marketing campaigns?

```
poutcome
unknown    36961
failure     4902
other       1840
success     1513
Name: count, dtype: int64
```

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

```
y
no     39922
yes     5294
Name: count, dtype: int64
```

18. Are there any correlations between different numeric attributes and the likelihood of subscribing to a term deposit?

```
age         0.025648
balance     0.052821
day        -0.028307
duration    0.394387
campaign   -0.073294
pdays       0.103699
previous    0.093576
dtype: float64
```