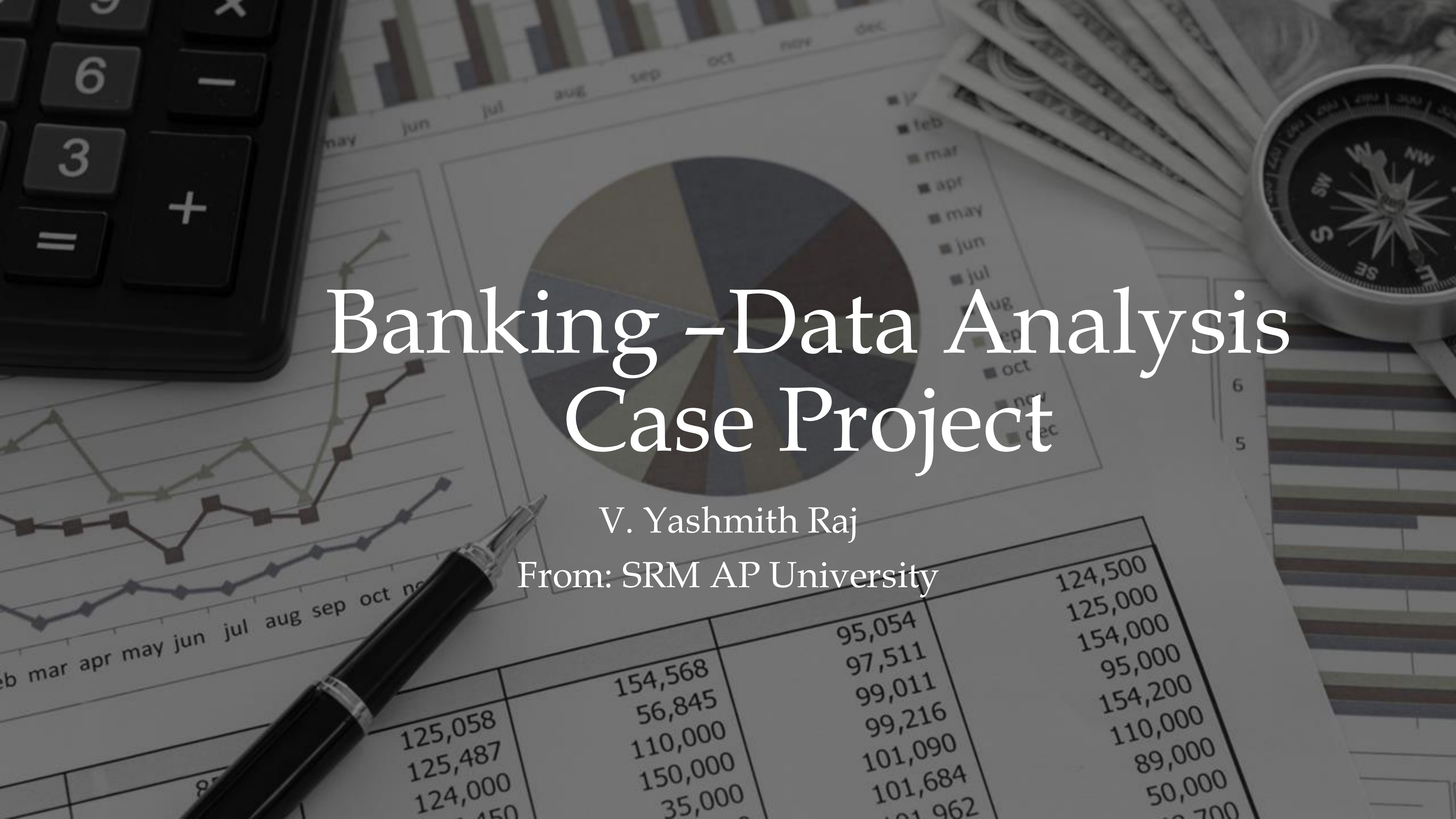


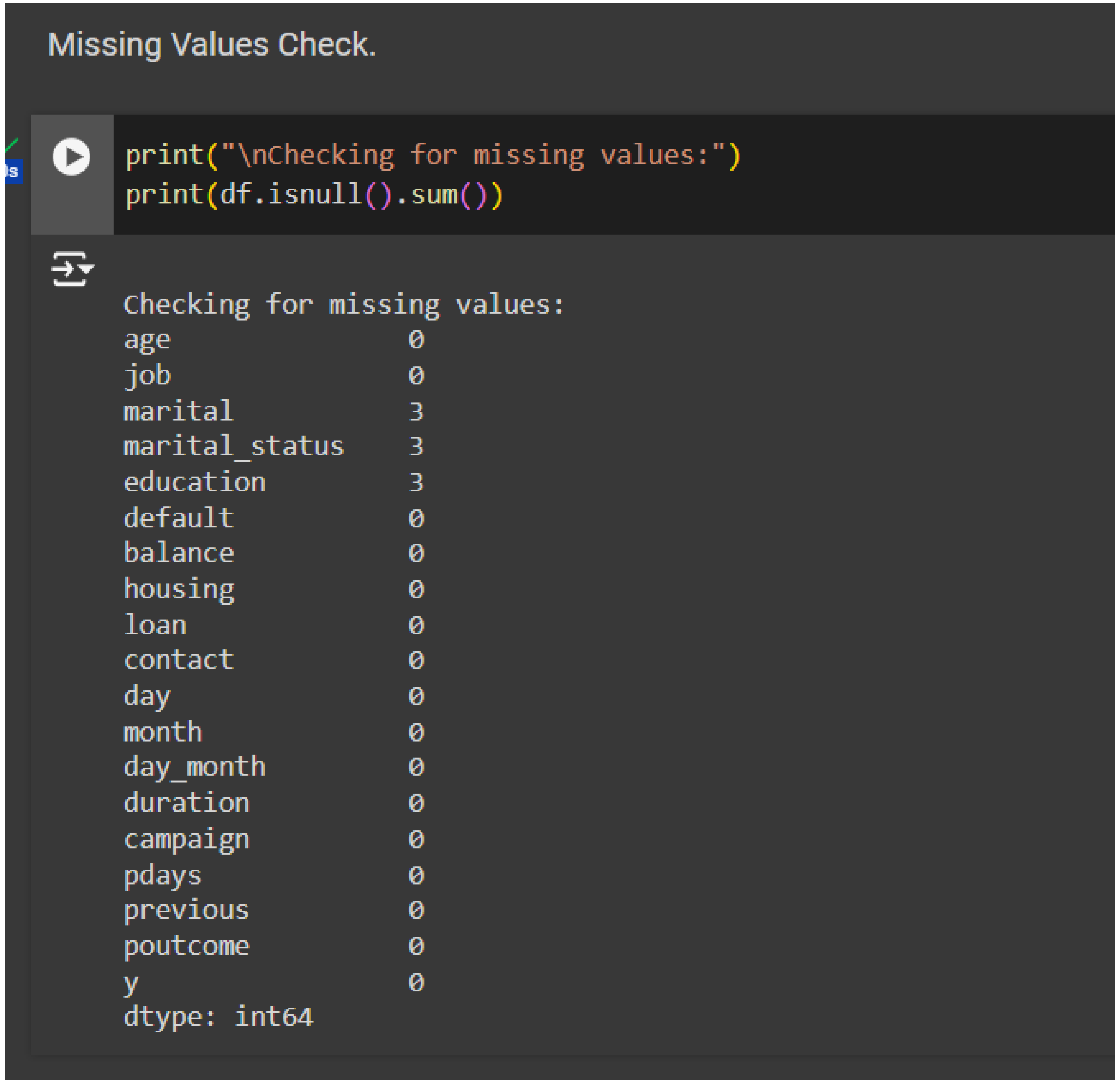
Banking –Data Analysis Case Project

V. Yashmith Raj

From: SRM AP University

Missing Values Analysis

- The code output reveals missing values in the columns **marital, marital_status,** and **education,** each with 3 missing entries.
- Given the small number of missing values relative to the overall dataset size and their random distribution without any specific pattern, we can **drop these missing values** without significantly affecting the analysis.



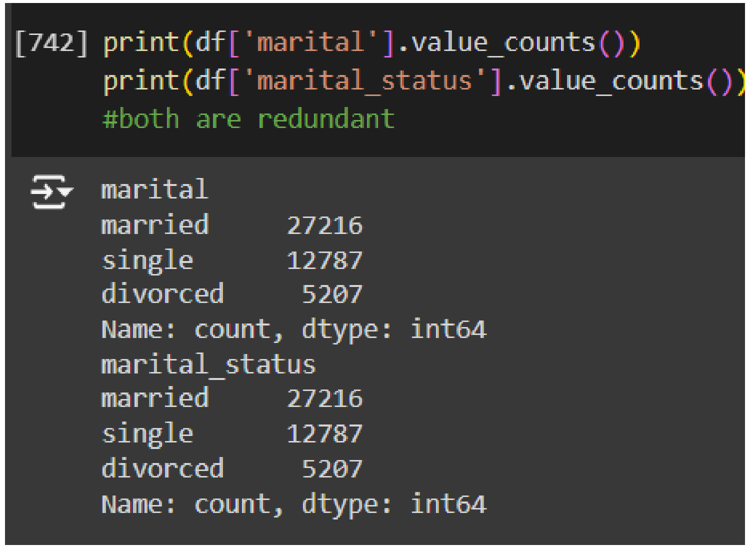
Missing Values Check.

```
print("\nChecking for missing values:")
print(df.isnull().sum())
```

```
Checking for missing values:
age               0
job               0
marital           3
marital_status    3
education         3
default           0
balance           0
housing           0
loan              0
contact           0
day               0
month             0
day_month         0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
y                 0
dtype: int64
```

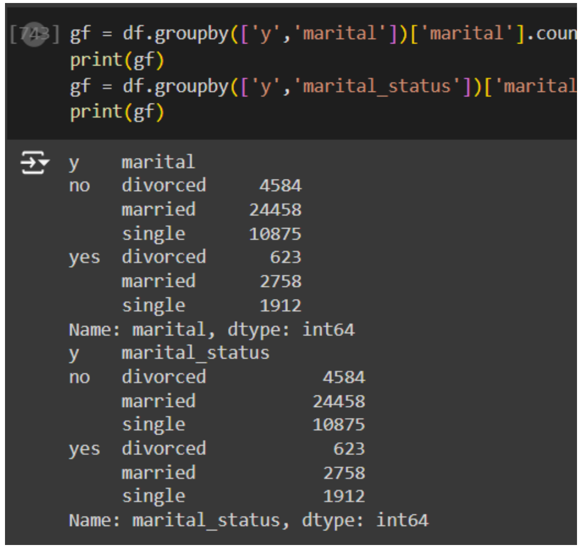
Redundant Columns

- Upon analyzing the dataset, it was found that the columns marital and marital_status contain identical information, leading to redundancy. Therefore, to maintain data integrity, we can drop one of these columns without losing any valuable information.
- Similarly, redundancy is seen between the day_month column and the separate day and month columns. To reduce data complexity, we can drop the day_month columns and retain day & month.



```
[742] print(df['marital'].value_counts())
      print(df['marital_status'].value_counts())
      #both are redundant
```

```
marital
married     27216
single      12787
divorced     5207
Name: count, dtype: int64
marital_status
married     27216
single      12787
divorced     5207
Name: count, dtype: int64
```

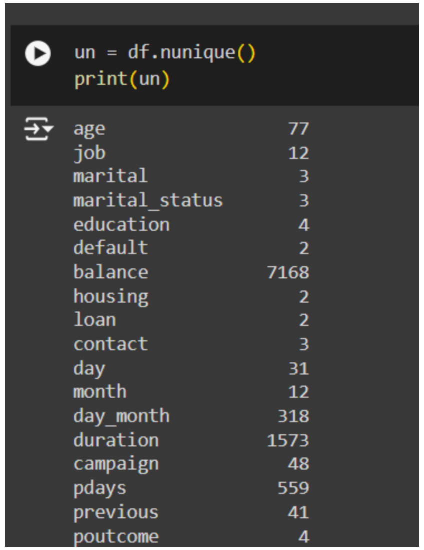


```
[743] gf = df.groupby(['y','marital'])['marital'].coun
      print(gf)
      gf = df.groupby(['y','marital_status'])['marital
      print(gf)
```

```
y    marital
no   divorced     4584
     married     24458
     single      10875
yes  divorced      623
     married      2758
     single       1912
Name: marital, dtype: int64
y    marital_status
no   divorced            4584
     married            24458
     single             10875
yes  divorced             623
     married             2758
     single              1912
Name: marital_status, dtype: int64
```

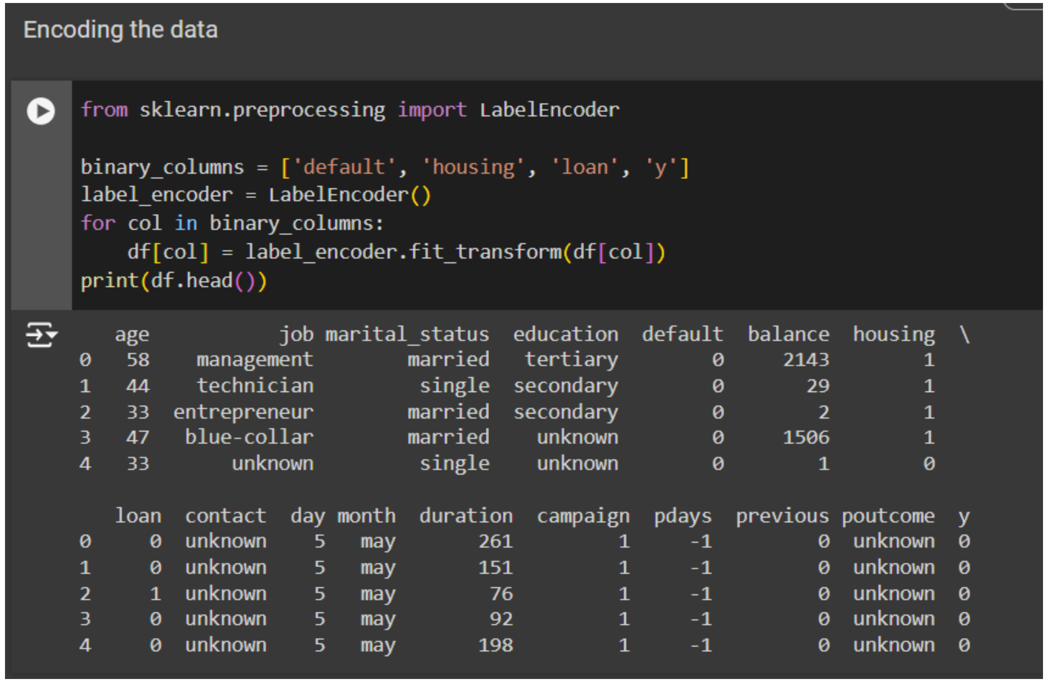
Encoding the data

- Encoding categorical variables is essential for machine learning models to process data effectively. For instance, columns like loan, housing, and y can be encoded due to their limited unique values (2 categories each). This transformation helps algorithms better understand the data and improves model performance.



```
un = df.nunique()
print(un)
```

```
age                 77
job                 12
marital              3
marital_status       3
education            4
default              2
balance           7168
housing              2
loan                 2
contact              3
day                 31
month               12
day_month          318
duration          1573
campaign            48
pdays              559
previous            41
poutcome             4
```



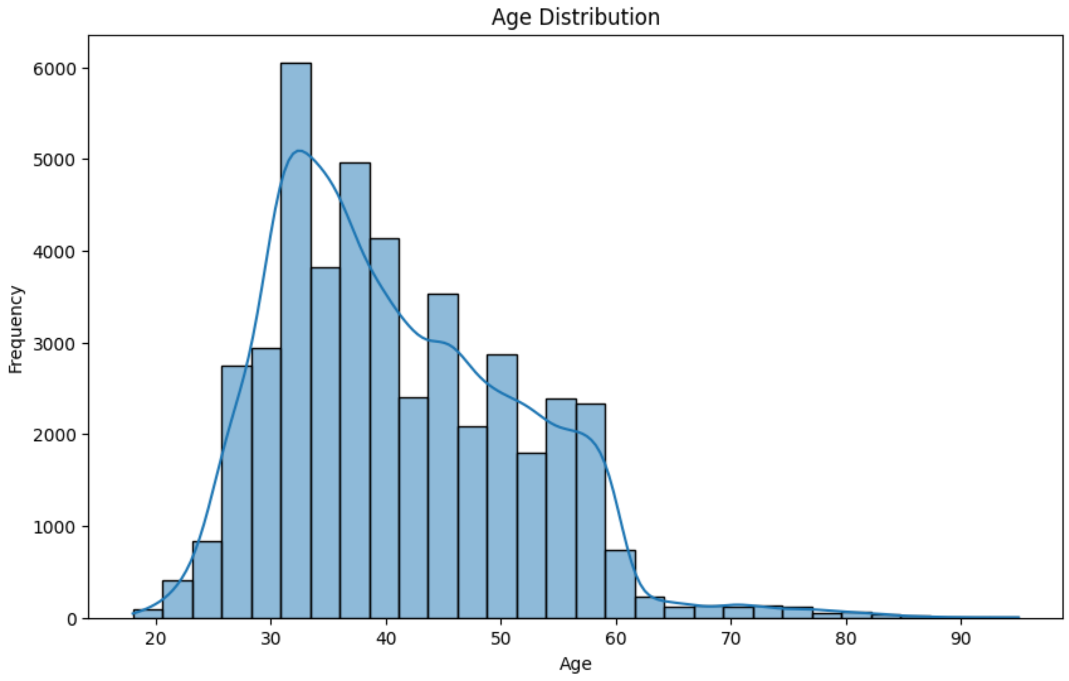
Encoding the data

```
from sklearn.preprocessing import LabelEncoder

binary_columns = ['default', 'housing', 'loan', 'y']
label_encoder = LabelEncoder()
for col in binary_columns:
    df[col] = label_encoder.fit_transform(df[col])
print(df.head())
```

```
   age           job marital_status education  default  balance  housing  \
0   58    management        married  tertiary        0     2143        1
1   44    technician         single secondary        0       29        1
2   33  entrepreneur        married secondary        0        2        1
3   47   blue-collar        married   unknown        0     1506        1
4   33       unknown         single   unknown        0        1        0

   loan  contact  day month  duration  campaign  pdays  previous poutcome  y
0     0  unknown    5   may       261         1     -1         0  unknown  0
1     0  unknown    5   may       151         1     -1         0  unknown  0
2     1  unknown    5   may        76         1     -1         0  unknown  0
3     0  unknown    5   may        92         1     -1         0  unknown  0
4     0  unknown    5   may       198         1     -1         0  unknown  0
```

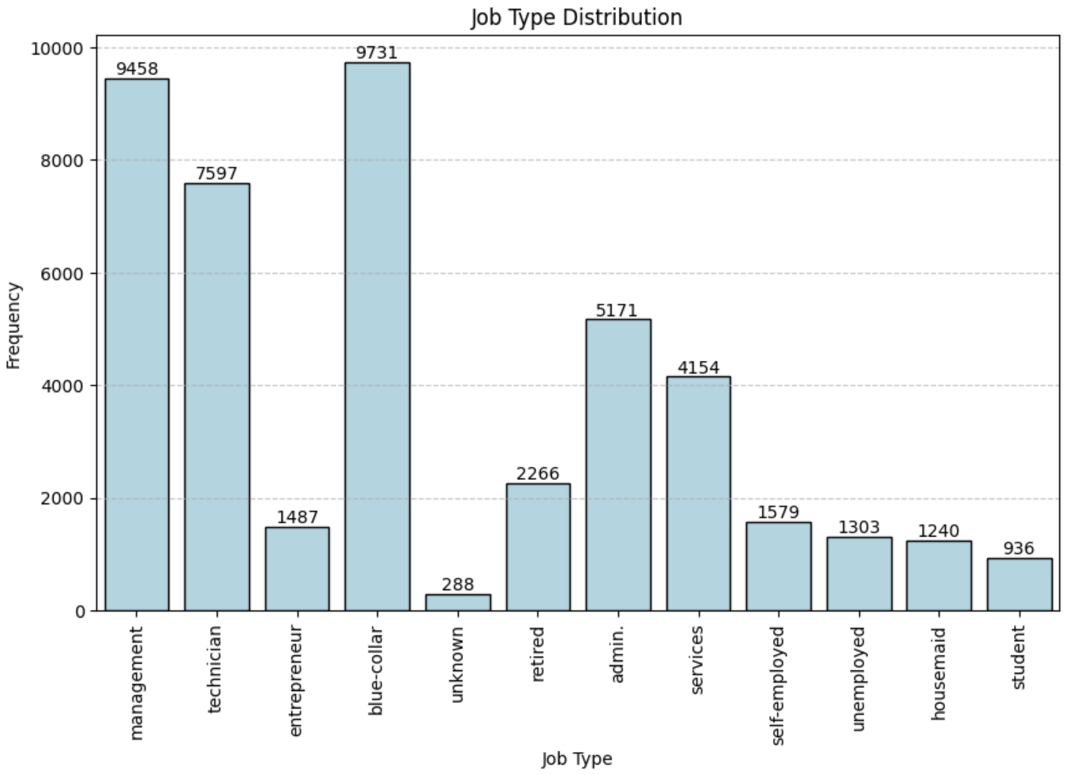
Plot 1 :
What is the distribution of age among the clients?

- The histogram displays the age distribution of the bank's clients. The x-axis represents age in years, while the y-axis indicates the number of clients in each age group.
- **Distribution Shape:** The distribution shows a right skew, characterized by a longer tail towards the older age groups. This suggests that the majority of clients are younger, with a decreasing number as the age increases.
- **Peak Age Group:** The histogram is most prominent around the 30-40 year age range. This peak indicates that the majority of clients are in their early to mid-career stages.
- **Age Range:** The age data spans from about 18 to over 90 years, offering a comprehensive view of the age diversity among clients.
- **Outliers:** There is a smaller group of clients aged 70 and above, which may be considered outliers. Examining this group's specific needs and behaviors could reveal opportunities for niche products and services.



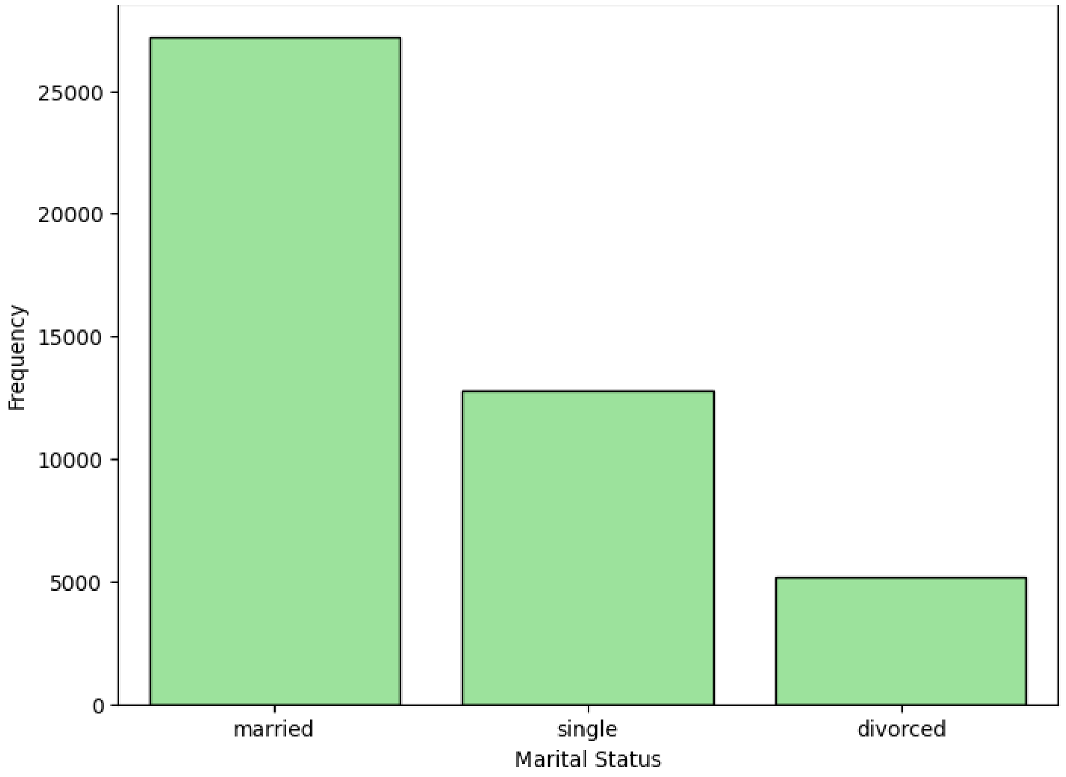
Plot 2 :
How does the job type vary among the clients?

- The bar chart shows the distribution of job type, with job categories on the x-axis and their frequencies on the y-axis.
- **Dominant Job Types:**
  - **Management and Blue-Collar Jobs:** These are the most prevalent job types among clients, with significantly higher frequencies compared to other categories. Management roles and blue-collar jobs dominate the client base, indicating a strong presence in these sectors.
  - **Technicians and Administrative Roles:** These job types also have notable representation but are less common than management and blue-collar positions.
- **Less Frequent Job Types:**
  - **Entrepreneurs, Retired, Services, Self-Employed, Unemployed, Housemaids, and Students:** These job types have much lower frequencies. Each of these categories includes fewer clients compared to the dominant job types.
- **Distribution Shape:**
  - **Skewed Distribution:** The distribution is skewed to the left, characterized by a few job types with high frequencies and a long tail with less frequent job categories. This skewness indicates that most clients belong to a few specific job types, while others are less common.
- The bank has a substantial customer base in management and blue-collar fields. Targeting marketing efforts towards these segments could be highly effective.



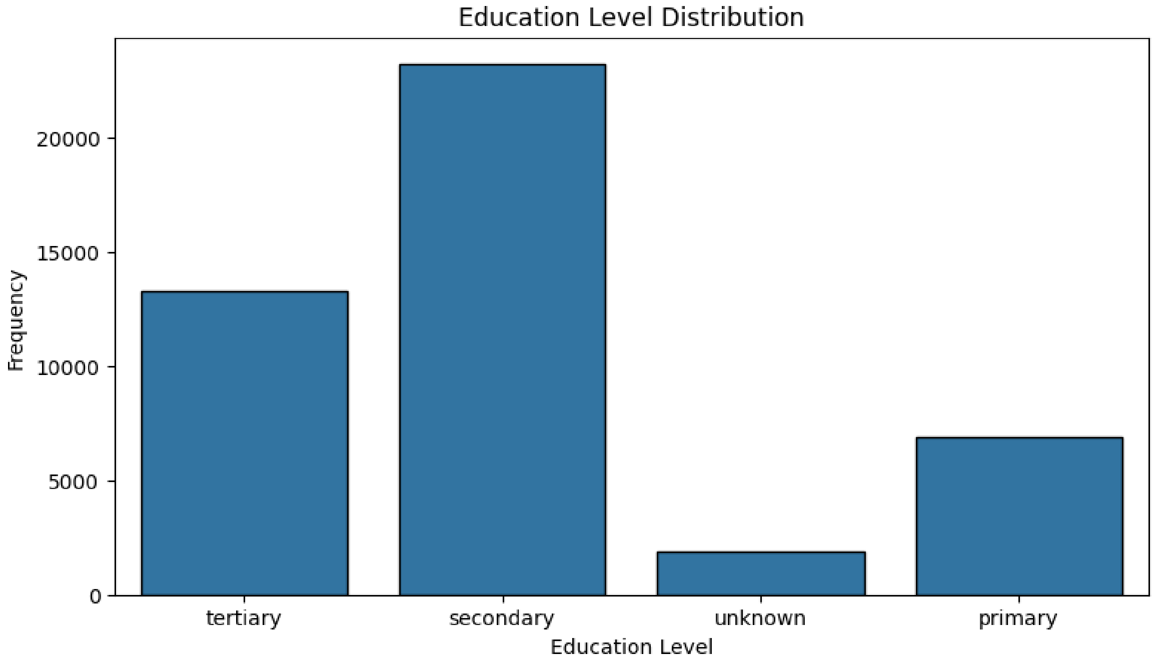
Plot 3: What is the marital status distribution of the clients?

- The bar chart displays the distribution of marital status among the bank's clients. The x-axis represents different marital statuses (married, single, divorced), and the y-axis shows the number of clients in each category.
- **Dominant Marital Status:**
  - **Married Clients:** Represent the largest segment of the client base, with significantly higher frequency compared to single and divorced clients.
- **Other Marital Statuses:**
  - **Single Clients:** The second largest group, though less frequent than married clients.
  - **Divorced Clients:** The smallest group, with the lowest frequency among the categories.
- **Distribution Shape**
  - The distribution is left-skewed, with a dominant number of married clients and fewer clients in the single and divorced categories. This indicates that most clients are married, while the other marital statuses are less common.



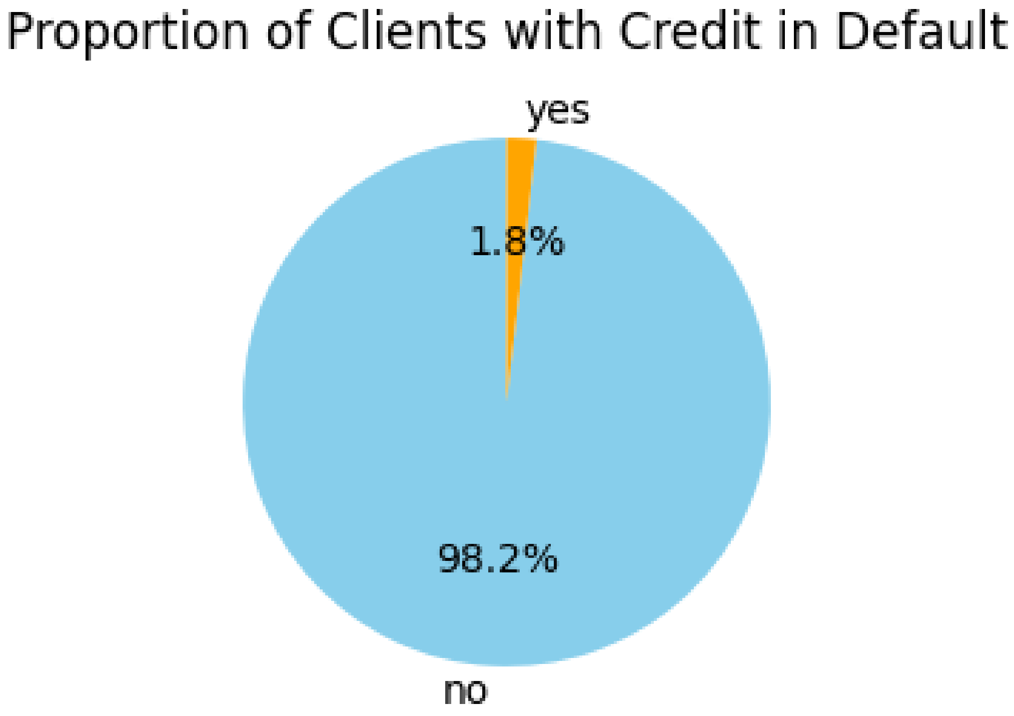
Plot 4:
What is the level of education among the clients?

- The bar chart shows the distribution of education levels among the bank's clients. The x-axis represents different education levels (tertiary, secondary, unknown, primary), while the y-axis shows the frequency or count of clients within each category.
- **Dominant Education Level:**
  - **Secondary Education:** This is the most common education level among clients, with a significantly higher count compared to other categories.
- **Other Education Levels:**
  - **Tertiary Education:** This level has a notable presence, though it is less frequent than secondary education.
  - **Unknown and Primary Education Levels:** These categories show significantly lower representation.
- **Distribution Shape:**
  - **Skewed Distribution:** The distribution is skewed to the left, with secondary education as the dominant category and tertiary, unknown, and primary education levels appearing less frequently.



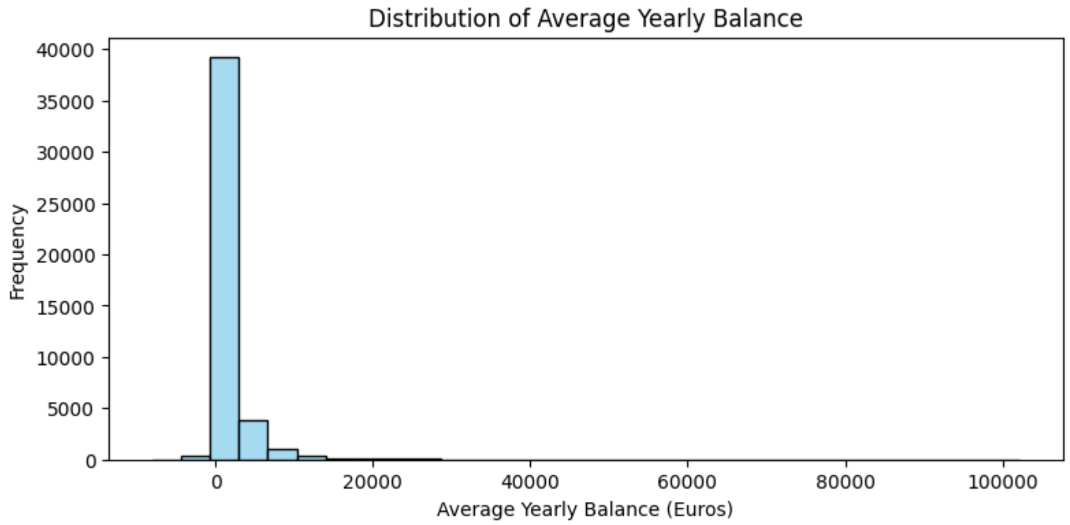
Plot 5:
What proportion of clients have credit in default?

- The pie chart displays the proportion of clients with credit in default. It is divided into two segments representing clients who have credit in default and those who do not.
- **Overwhelming Majority:** A substantial majority of clients (98.2%) do not have credit in default.
- **Small Proportion:** Only a small fraction of clients (1.8%) have credit in default.
- The bank has a relatively low rate of credit defaults among its clients, this indicates a generally healthy credit portfolio.



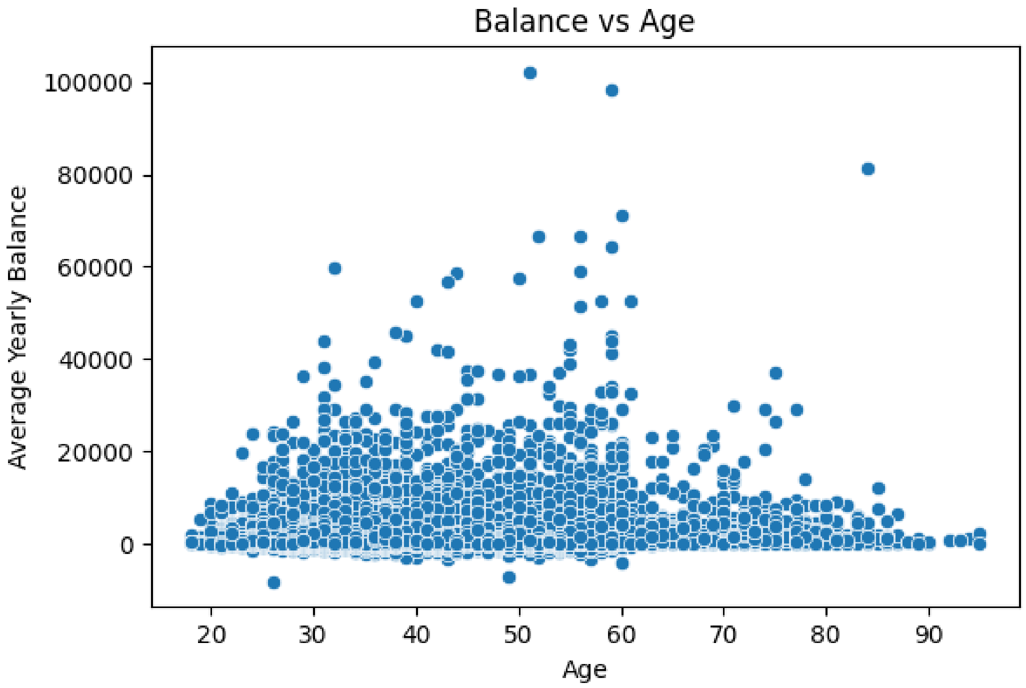
Plot 6:
What is the distribution of average yearly balance among the clients?

- The histogram illustrates the distribution of average yearly balances among clients. The x-axis represents average yearly balance in euros, while the y-axis shows the frequency of clients in each balance range.
- **Distribution Shape:** The distribution is right-skewed, indicating that most clients have lower average yearly balances, with a smaller number holding significantly higher balances.
- **Peak Frequency:** The highest frequency of clients falls within the 0 to 20,000 euros range, showing that the majority have relatively low average yearly balances.
- **Central Tendency:** The mean balance (1362.33 euros) is higher than the median balance (448.5 euros), reinforcing the right-skewed nature of the distribution.
- **Spread:** The interquartile range (IQR) of 1356 euros (from 25th percentile at 72 euros to 75th percentile at 1428 euros) indicates that the middle 50% of clients have balances within this range.
- **Outliers:** The minimum value of -8019 euros suggests the presence of outliers on the lower end, though their impact requires further analysis.



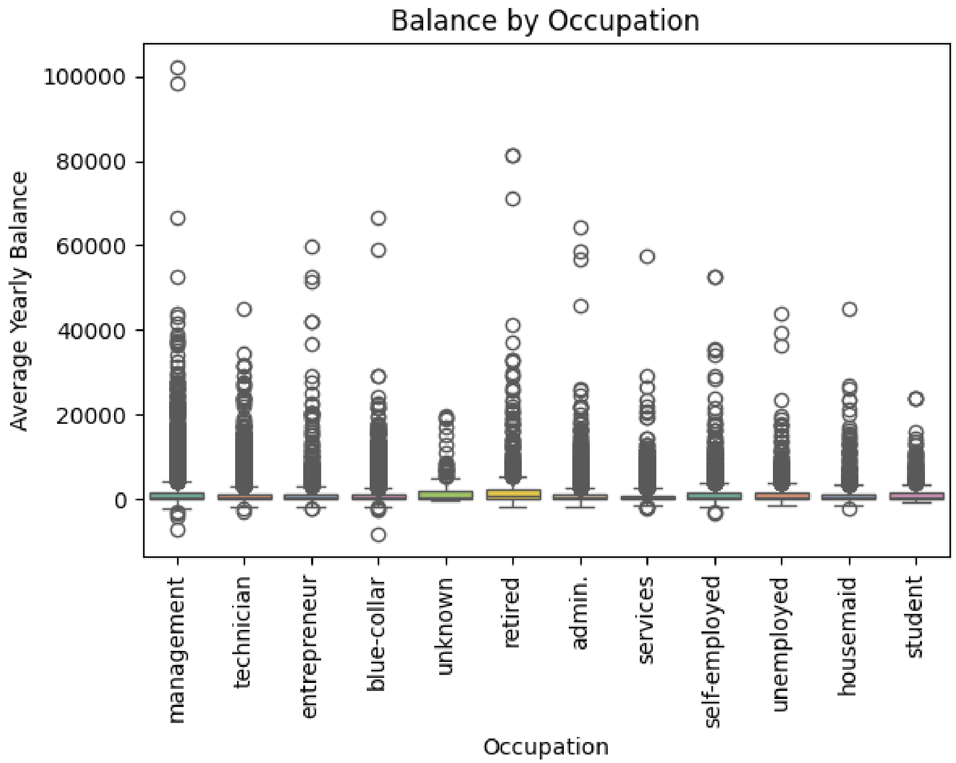
Plot 6:
What is the distribution of average yearly balance among the clients?

- Balance vs. Age
  - The scatter plot indicates a weak positive correlation between age and average yearly balance. This suggests that older clients tend to have slightly higher balances, but there's significant variability.
  - There are outliers with high balances at younger ages, indicating potential high-net-worth individuals.



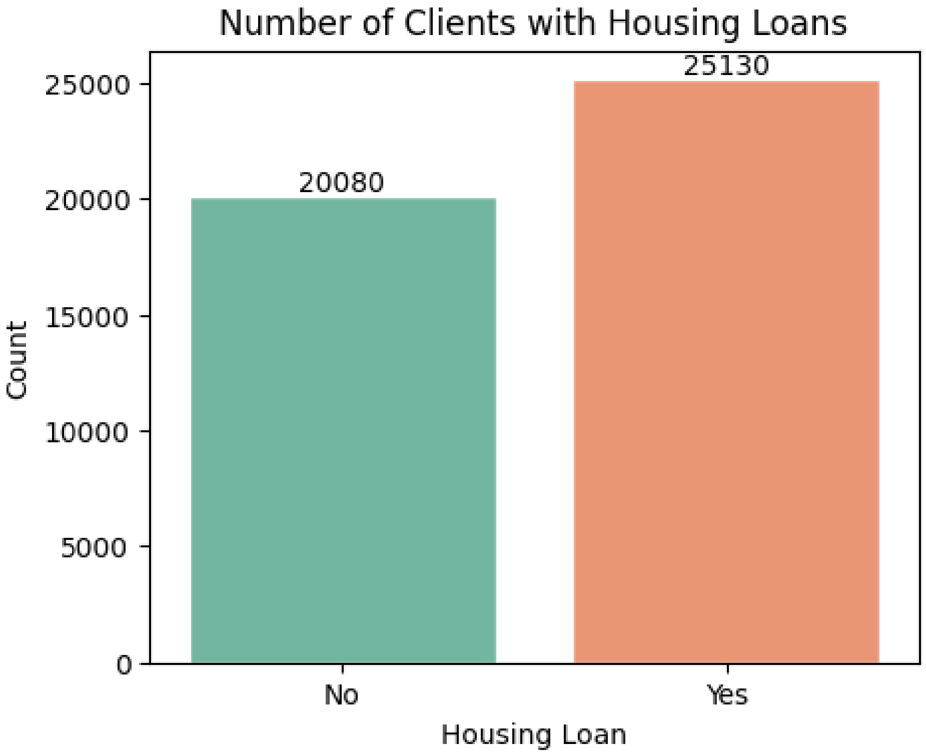
Plot 6:
What is the distribution of average yearly balance among the clients?

- Balance vs. Occupation
  - The box plots reveal substantial differences in average yearly balance across occupations.
  - Management and entrepreneur roles tend to have higher median balances and larger ranges, indicating greater financial stability or wealth.
  - Blue-collar and unemployed occupations generally exhibit lower median balances and tighter distributions.



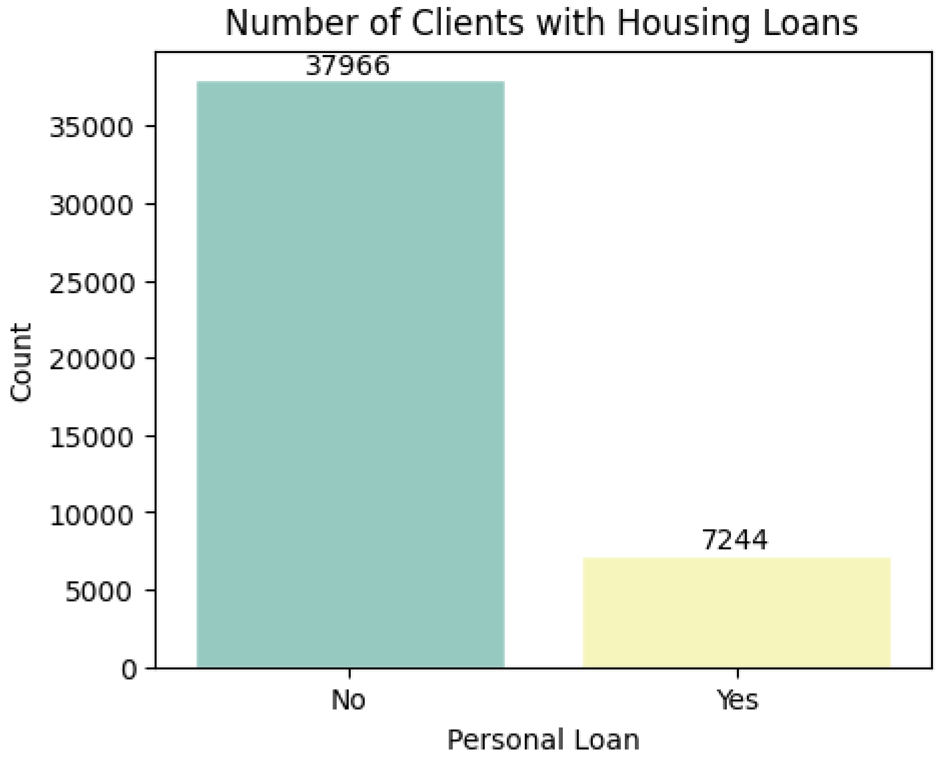
Plot 7:
How many clients have housing loans?

- **Total Clients:** Based on the chart, there are a total of 25130 + 20080 = 45210 clients.
- **Clients with Housing Loans:** A total of 25130 clients have housing loans.
- **Clients without Housing Loans:** A total of 20080 clients do not have housing loans.



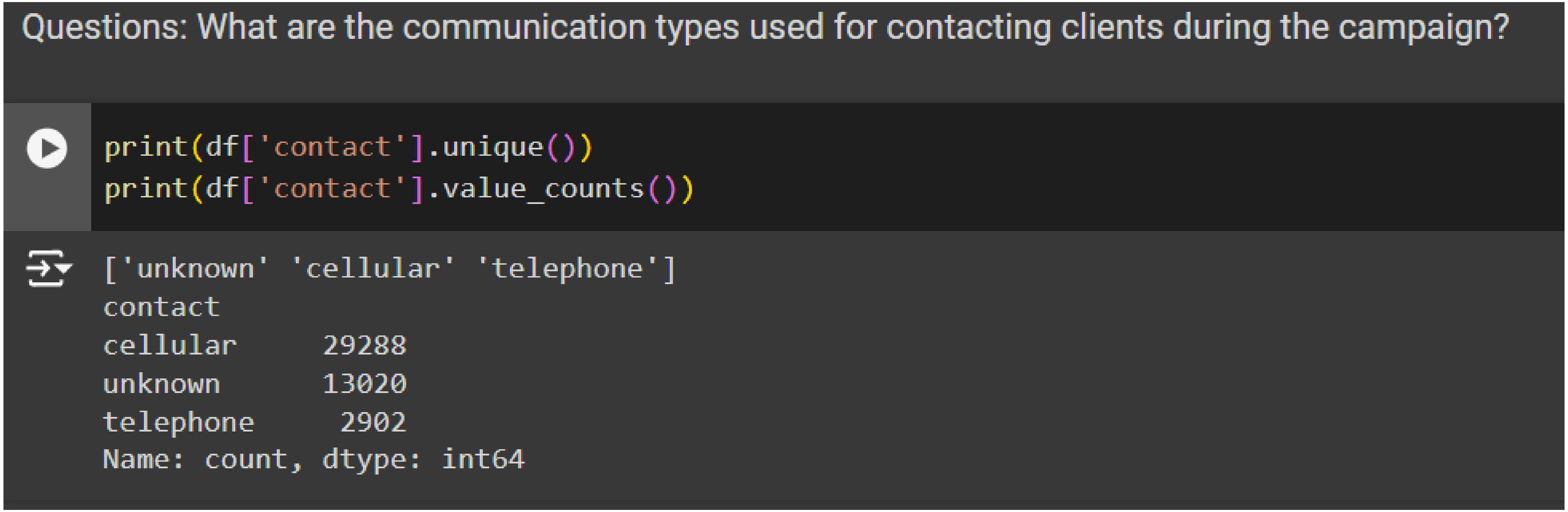
Plot 8:
How many clients have personal loans?

- **Total Clients:** Based on the chart, there are a total of 45210 clients.
- **Clients with Personal Loans:**
  - A total of 37966 clients have housing loans.
- **Clients without Housing Loans:**
  - A total of 7244 clients do not have housing loans.



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

- **Cellular:** 29288 clients were contacted via cellular phones.
- **Unknown:** 13020 clients had an unknown communication type.
- **Telephone:** 2902 clients were contacted via traditional landline phones.



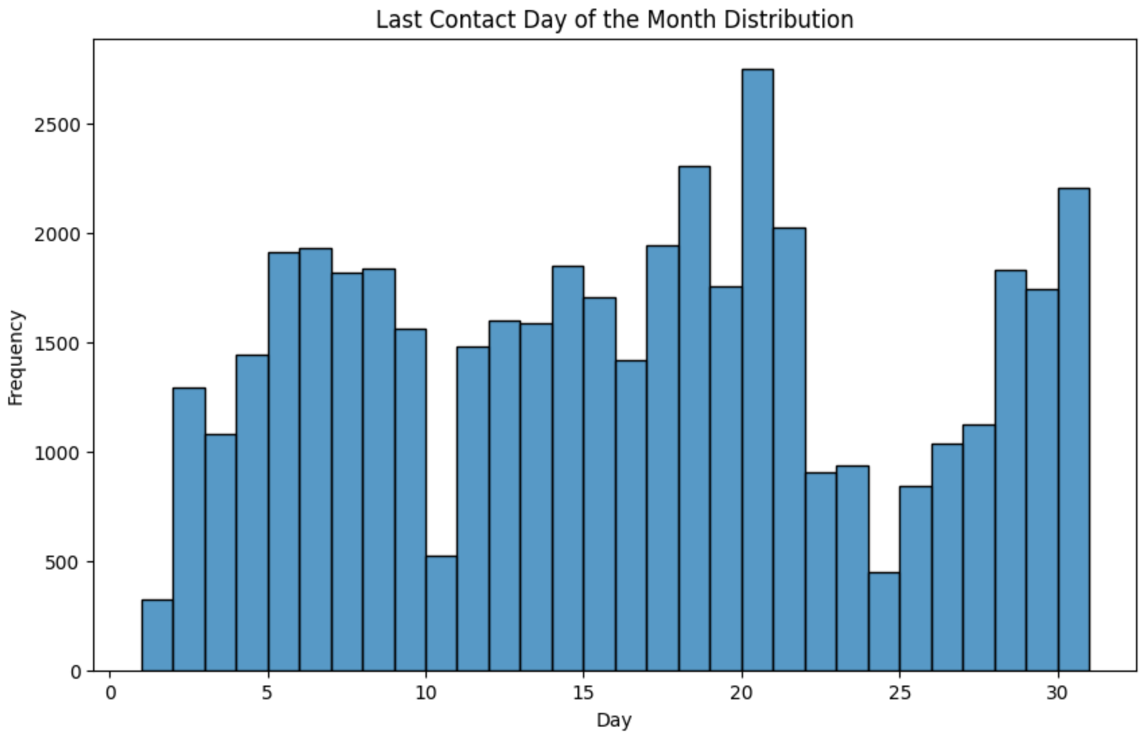
Questions: What are the communication types used for contacting clients during the campaign?

```
print(df['contact'].unique())
print(df['contact'].value_counts())
```

```
['unknown' 'cellular' 'telephone']
contact
cellular     29288
unknown      13020
telephone     2902
Name: count, dtype: int64
```

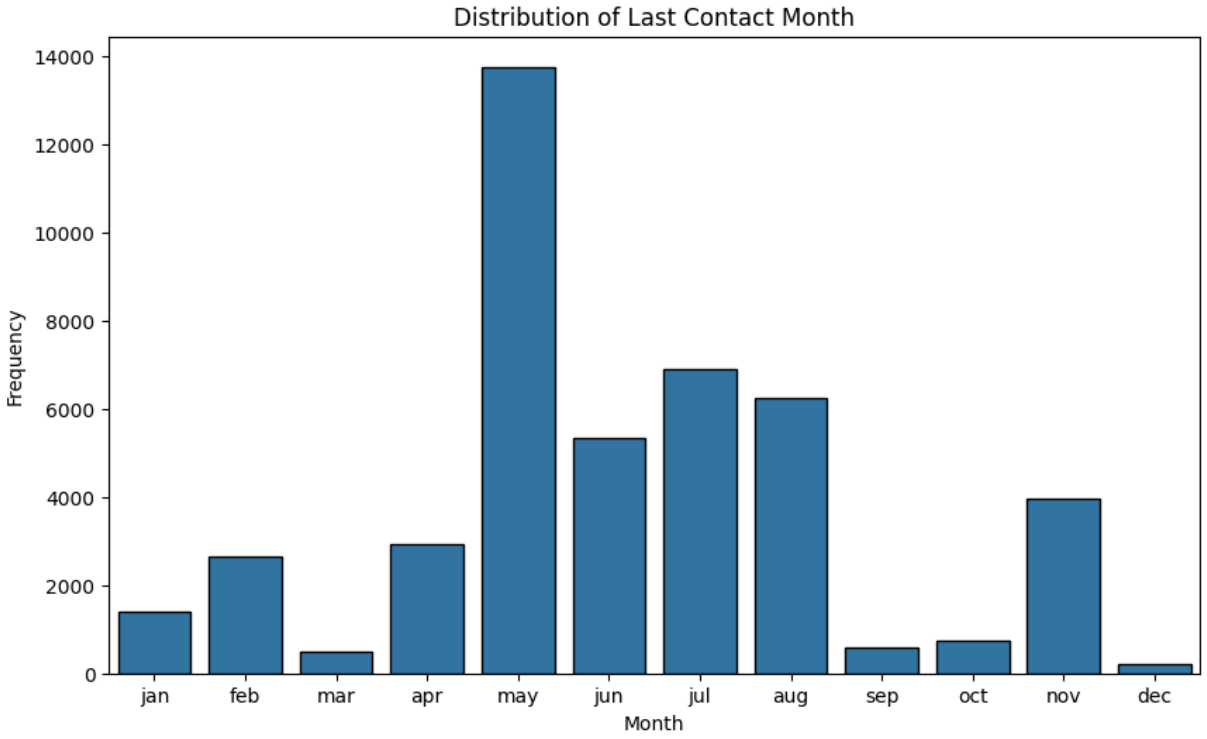
Plot 10:
What is the distribution of the last contact day of the month?

- The provided histogram visualizes the distribution of the last contact day of the month for the bank's clients. The x-axis represents the days of the month, and the y-axis shows the frequency or count of client contacts on each day.
- **Distribution Shape:** The distribution is relatively uniform across most days of the month, showing consistent contact frequencies with minor fluctuations.
- **Peak Days:** There is a noticeable increase in contact frequency towards the end of the month, especially around the 20th to 25th. This suggests that these days are particularly busy for client contacts.
- **Lower Frequency Days:** The beginning and end of the month, particularly around the 1st and the 31st, exhibit lower contact frequencies. This indicates fewer client interactions on these days.



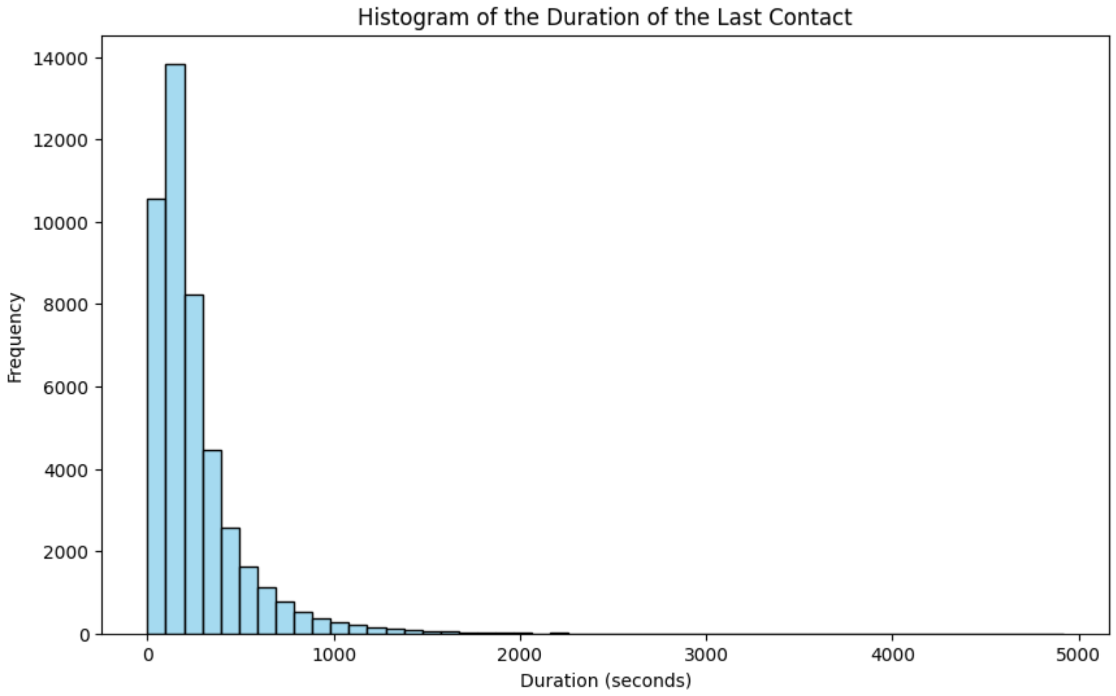
Plot 11: How does the last contact month vary among the clients?

- The bar chart illustrates the distribution of the last contact month for the bank's clients. The x-axis represents the months of the year, while the y-axis shows the frequency of client contacts in each month.
- **Peak Months:** The months of May and October show the highest frequencies of client contacts, indicating these periods were particularly active for marketing campaigns.
- **Seasonal Variation:** There is a noticeable seasonal pattern with increased contact frequencies during spring (March-May) and fall (September-November).
- **Lower Contact Months:** The months of March, September, and December have the lowest frequencies of client contacts.



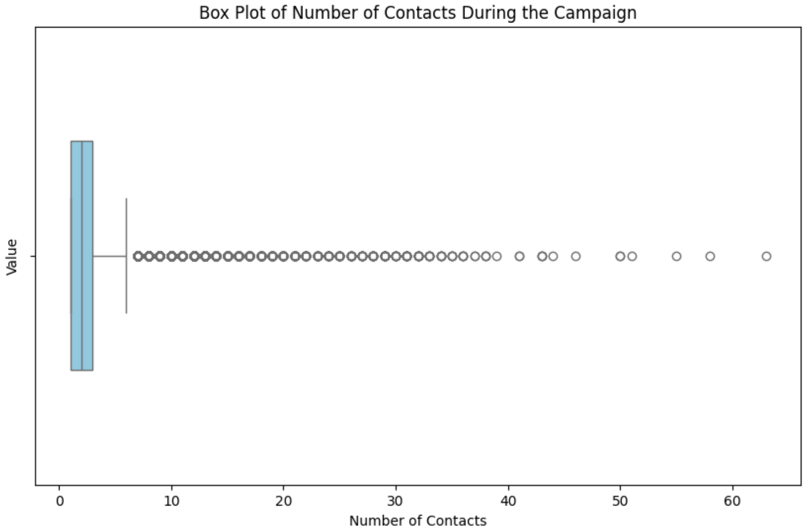
Plot 12:
What is the distribution of the duration of the last contact?

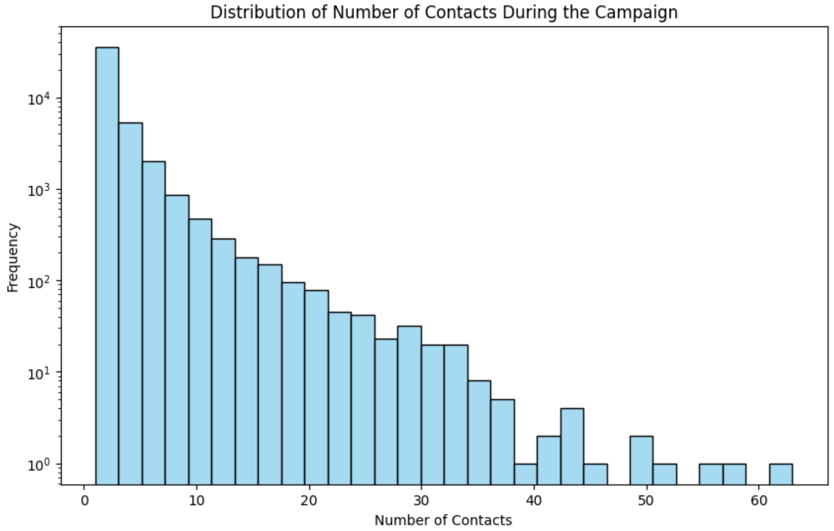
- The following histogram visualizes the distribution of the length of the last contact with clients. The x-axis represents the duration in seconds, and the y-axis shows the frequency or count of client contacts within each duration range.
- **Distribution Shape:**
  - The distribution is heavily right-skewed, indicating that a large number of contacts are short in duration, with a smaller number of longer contacts. This suggests that most interactions with clients are relatively short.
- **Peak Frequency:**
  - The highest frequency of contacts occurs within the first few hundred seconds, indicating that the vast majority of interactions are short in length.
- **Long Tail:**
  - The histogram shows a long tail to the right, suggesting that a small proportion of contacts last considerably longer. These could represent in-depth discussions or complex customer inquiries.



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

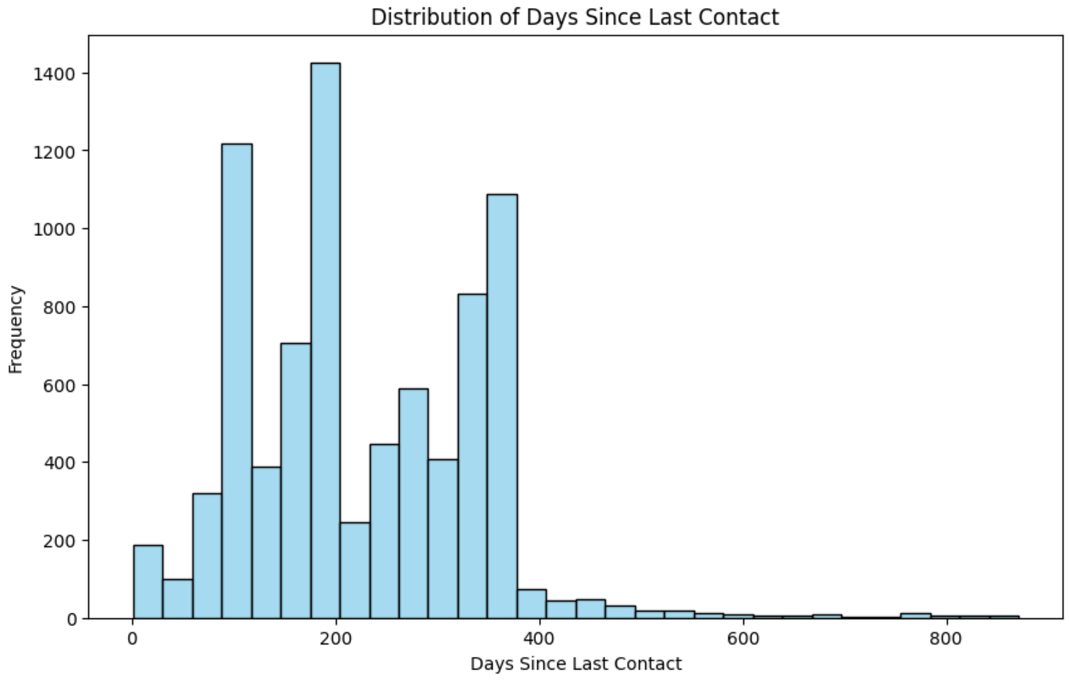
- **Right-Skewed Distribution:** The distribution is heavily right-skewed. This indicates that while most clients received a small number of contacts, there is a smaller group of clients who experienced a higher number of contacts.
- **Most Common Frequency:** The peak occurs at one contact, showing that the majority of clients were contacted only once during the campaign. This suggests that a large portion of the client base received minimal follow-ups.
- **Presence of Long Tail:** The histogram displays a long tail extending to the right, which indicates that there are clients who received multiple contacts. This long tail highlights the small subset of clients who were contacted several times.
- **Outliers:** As majority of the data is focused at the right with lesser number of call we can see that there are plenty of outliers in the box plot





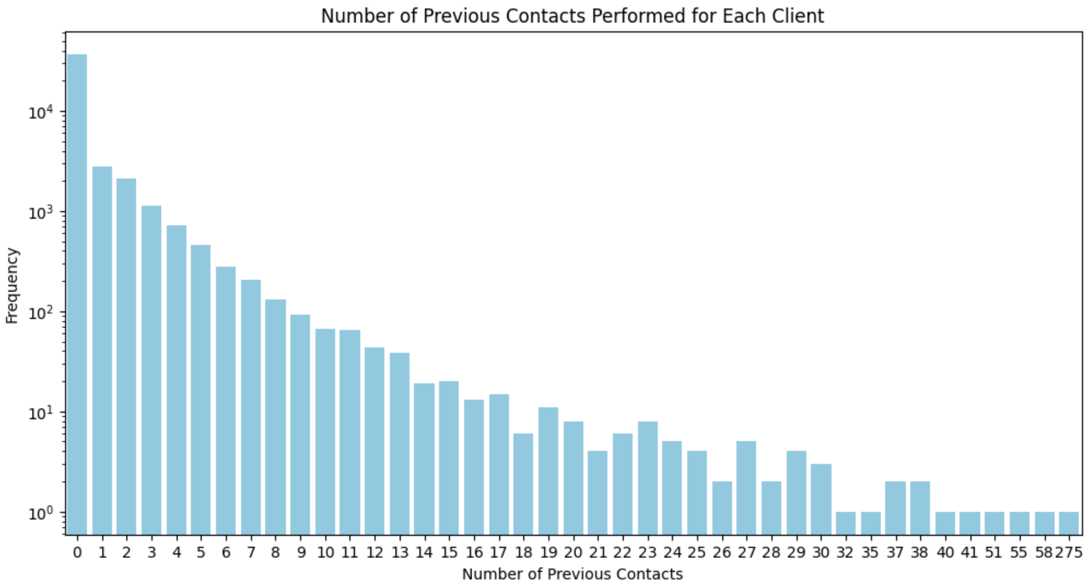
Plot 14:
What is the distribution of the number of days passed since the client was last contacted from a previous campaign?

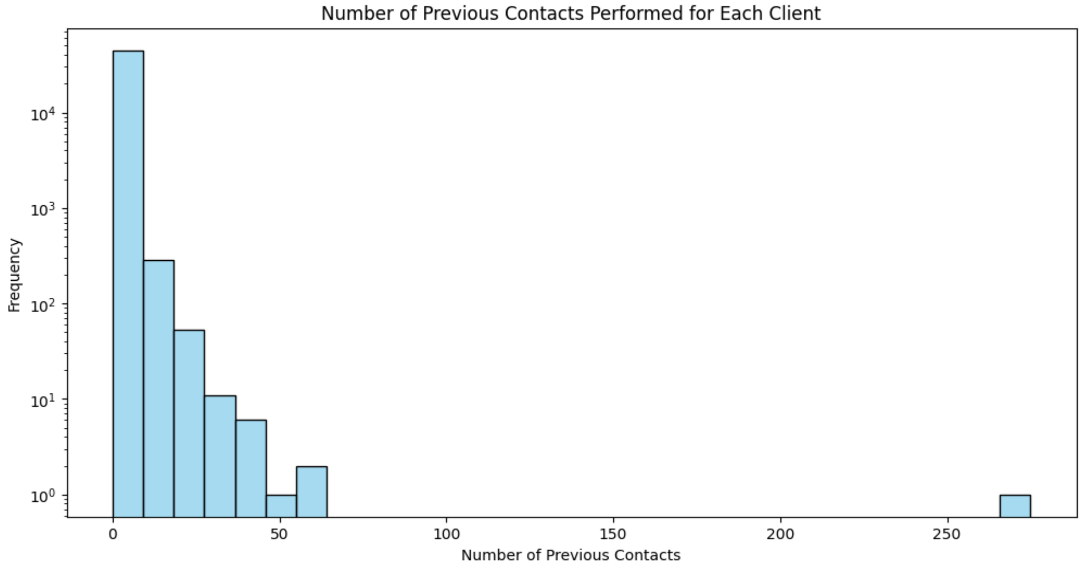
- **Most Clients Not Contacted Previously:** A large portion of clients have a pdays value of -1, indicating they were not contacted in previous campaigns, reflecting a focus on acquiring new clients.
- **Right-Skewed Distribution:** For clients who were contacted previously, the distribution is right-skewed, showing that while many clients have fewer days since their last contact, the number of clients with longer intervals decreases. The mean of 40.19 days supports this.
- **Outliers:** Clients with up to 800 days since their last contact are potential outliers, representing those who haven't been contacted for an extended period.



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

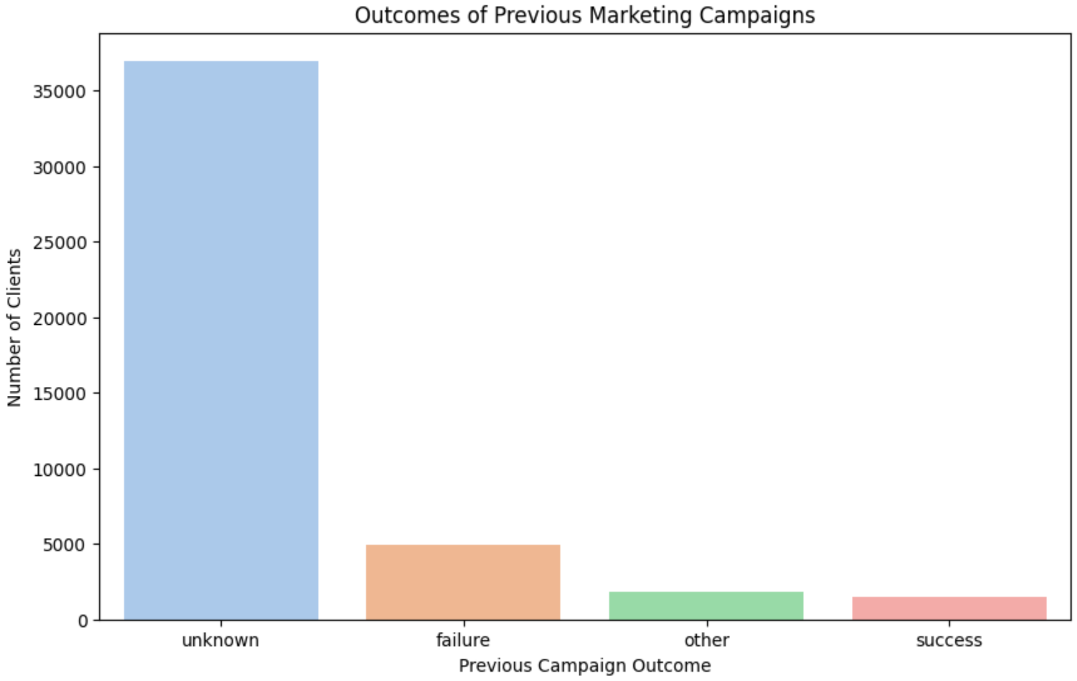
- The provided bar chart show the distribution of the number of contacts performed during the campaign for each client. The x-axis represents the number of contacts, and the y-axis represents the frequency or count of clients who experienced that number of contacts.
- **High Frequency:** The majority of clients received only one contact during the campaign, as evidenced by the tallest bar at the "0" position.
- **Decreasing Frequency:** As the number of contacts increases, the frequency of clients decreases rapidly, indicating that a smaller proportion of clients received multiple contacts.
- **Extreme Outlier:** In the histogram you can see that there seems to be a outlier in the data set which represents that 275 calls have been made to a client.

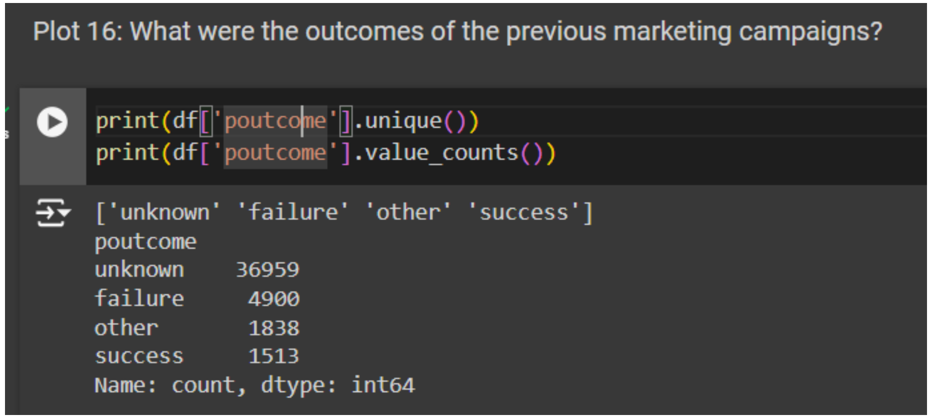




Plot 16:
What were the outcomes of the previous marketing campaigns?

- The provided bar chart visualizes the outcomes of previous marketing campaigns. The x-axis represents the different outcomes (unknown, failure, other, success), and the y-axis represents the number of clients associated with each outcome.
- **Majority Outcome:** The majority of previous campaigns have an unknown outcome, representing the largest segment of the chart. This could be due to missing data or incomplete records.
- **Failure:** The "failure" category has a significant presence, indicating that a substantial number of previous campaigns did not result in successful subscriptions.
- **Other and Success:** The "other" and "success" categories have relatively smaller representations, suggesting that these outcomes occurred less frequently.



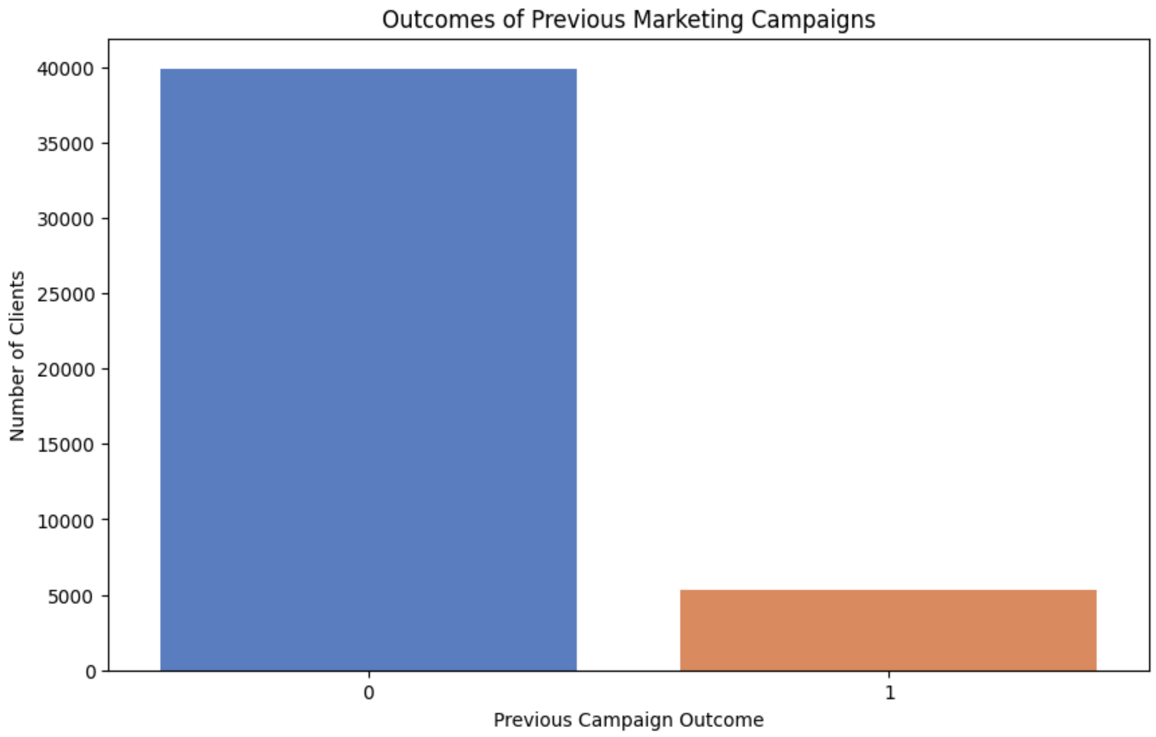

Plot 16: What were the outcomes of the previous marketing campaigns?

```
print(df['poutcome'].unique())
print(df['poutcome'].value_counts())
```

```
['unknown' 'failure' 'other' 'success']
poutcome
unknown    36959
failure     4900
other       1838
success     1513
Name: count, dtype: int64
```

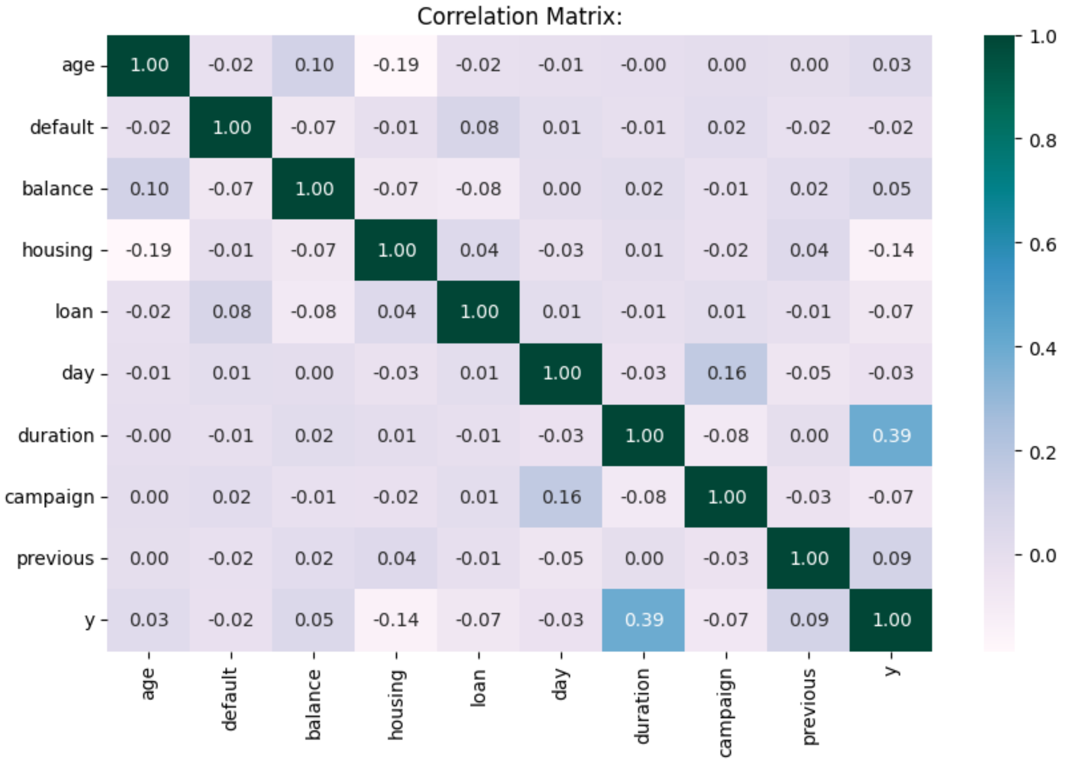
Plot 17:
What is the distribution of clients who subscribed to a term deposit vs. those who did not?

- The bar chart shows how many clients subscribed to a term deposit compared to those who didn't. On the x-axis, you see the subscription status ("yes" or "no"), and the y-axis indicates the number of clients in each category.
- **Majority Didn't Subscribe:** Most clients did not subscribe to the term deposit, with the "no" bar being significantly taller. This points to a low conversion rate for the campaign.
- **Smaller Subscriber Base:** There were 5,293 clients who subscribed to the term deposit, but a much larger number, 39,917, chose not to.
- The low conversion rate suggests that the current campaign might not be effectively reaching or convincing potential subscribers.



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

- **Strongest Positive Correlation:**
- **Duration and Subscription (y):** There's a moderate positive correlation (0.39) between the duration of the last contact and the likelihood of subscribing to a term deposit. This suggests that longer conversations might lead to higher subscription rates.
- **Weak to No Correlation:** Most variables exhibit weak or no correlation with the target variable (y - subscription). This indicates that individual factors might not be strong predictors of subscription behavior.



	age	default	balance	housing	loan	day	duration	campaign	previous	y
age	1.00	-0.02	0.10	-0.19	-0.02	-0.01	-0.00	0.00	0.00	0.03
default	-0.02	1.00	-0.07	-0.01	0.08	0.01	-0.01	0.02	-0.02	-0.02
balance	0.10	-0.07	1.00	-0.07	-0.08	0.00	0.02	-0.01	0.02	0.05
housing	-0.19	-0.01	-0.07	1.00	0.04	-0.03	0.01	-0.02	0.04	-0.14
loan	-0.02	0.08	-0.08	0.04	1.00	0.01	-0.01	0.01	-0.01	-0.07
day	-0.01	0.01	0.00	-0.03	0.01	1.00	-0.03	0.16	-0.05	-0.03
duration	-0.00	-0.01	0.02	0.01	-0.01	-0.03	1.00	-0.08	0.00	0.39
campaign	0.00	0.02	-0.01	-0.02	0.01	0.16	-0.08	1.00	-0.03	-0.07
previous	0.00	-0.02	0.02	0.04	-0.01	-0.05	0.00	-0.03	1.00	0.09
y	0.03	-0.02	0.05	-0.14	-0.07	-0.03	0.39	-0.07	0.09	1.00

Final Insights of the dataset:

- **Low Conversion Rate**: The analysis shows that a significant majority of clients did not subscribe to the term deposit. With 39,917 clients opting out compared to 5,293 who subscribed, it indicates a low conversion rate. This suggests that the current campaign might not be effectively reaching or convincing potential subscribers.
- **Contact Duration**:
  - **General Trend**: Most client interactions are relatively brief, with the highest frequency of contacts occurring within the first few hundred seconds. This trend may suggest a lack of deep engagement or interest from clients.
  - **Correlation with Subscription**: There is a moderate positive correlation (0.39) between the duration of the last contact and the likelihood of subscribing to a term deposit. This implies that longer conversations are generally associated with higher subscription rates, highlighting the importance of more extended interactions in potentially increasing subscriptions.

My Experience :

- Through this project, I gained valuable experience in data analysis, enhancing my ability to select and apply the most appropriate graphs and plots based on data types and analytical objectives.
- **Categorical Data**: I utilized count and bar plots to effectively visualize and compare the frequencies of different categories, providing clear insights into distribution and prevalence.
- **Distribution Analysis**: For numerical data, I employed histograms to analyze and visualize data distributions, allowing for a better understanding of patterns and trends.
- **Outlier Detection**: I used box plots to identify and analyze outliers, offering insights into anomalies and extreme values within the dataset.
- **Correlation Analysis**: I analyzed correlations between variables using heatmaps to identify relationships and dependencies. This included assessing the impact of contact duration on subscription rates, revealing moderate correlations which are effecting the subcription rates.