# BlockStackML Bank Data Analysis

## Features of the dataset

Here's a brief breakdown of the features of the dataset:

- Age: A numerical variable that indicates the age of the client.
- Employment: The customer's employment type is indicated by this categorical variable.
- Marital: A categorical variable indicating whether or not the consumer is married.
- Education: The customer's degree of education is indicated by this categorical variable.
- Default: A binary variable that can be set to "yes" or "no" to indicate if a customer has missed credit payments.
- Balance: A numerical variable that shows the typical annual balance expressed in euros.
- Housing: A binary variable that indicates "yes" or "no" depending on whether the consumer has a home loan.
- Loan: A binary variable that indicates "yes" or "no" depending on whether the consumer has a personal loan.
- Contact: A categorical variable that indicates how the most recent marketing campaign contact was made.
- Day: A number that represents the day of the month that the client was last contacted.
- Month: The month of the year the consumer was last contacted is indicated by this categorical feature.
- Duration: A numerical variable that indicates, in seconds, how long the last contact lasted.
- Campaign: A numerical variable that shows how many contacts this customer's marketing campaign resulted in.
- Pdays: A numerical variable that shows how many days have passed since the customer was last contacted during a previous campaign (a value of -1 indicates the client has never been contacted previously).
- Previous: A numerical variable that indicates how many connections this client had prior to this campaign.
- Poutome: A categorical variable that shows the outcome of the prior marketing initiative.

- Y: The target variable, a binary class attribute denoting if the client subscribed to a term deposit ("yes" or "no").

# Statistical Analysis

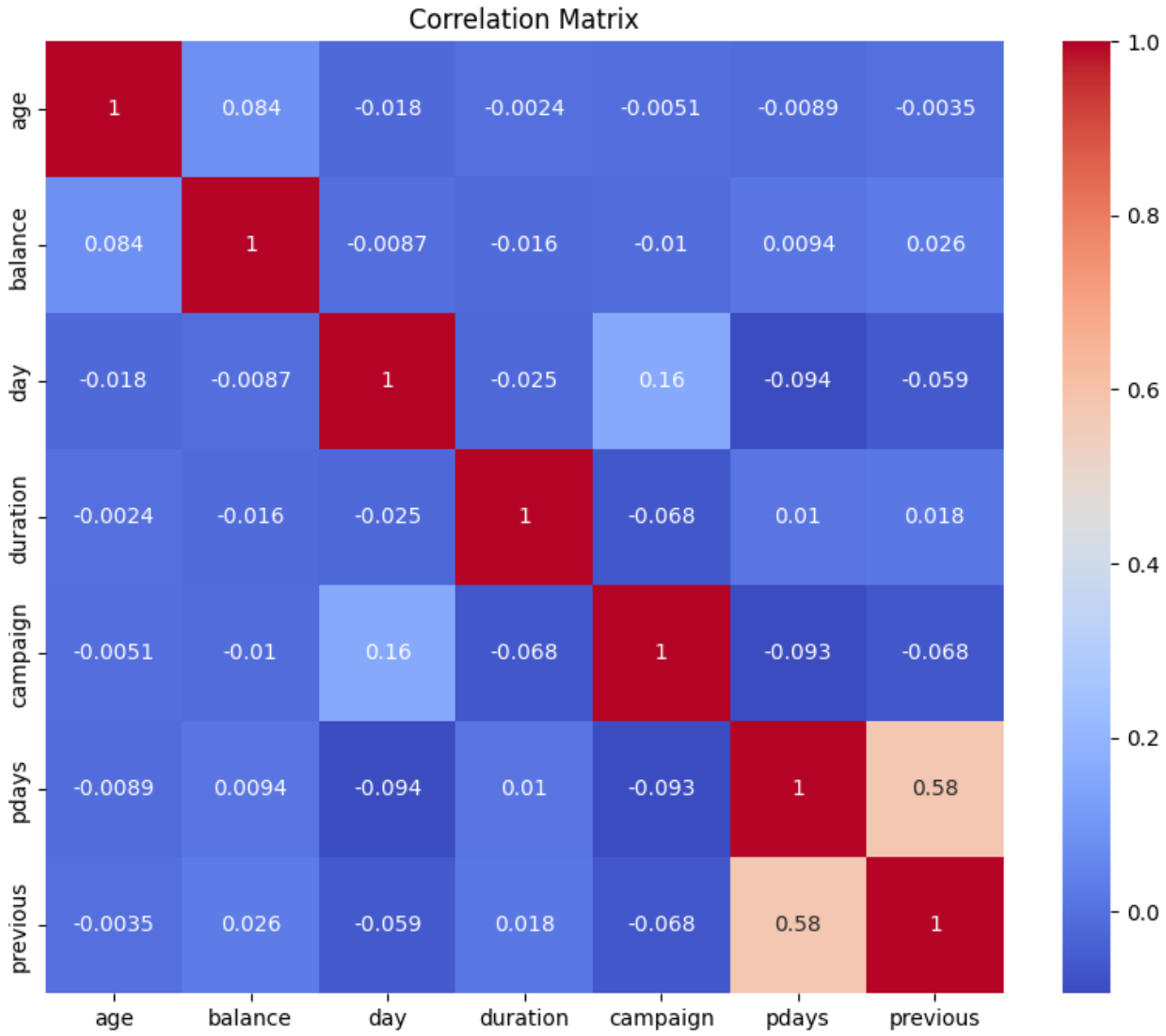
The outcomes of a thorough statistical analysis performed on a campaign-related dataset are presented in this report. This analysis aims to shed light on the variables driving customer subscription choices and offer practical suggestions for business and marketing tactics.

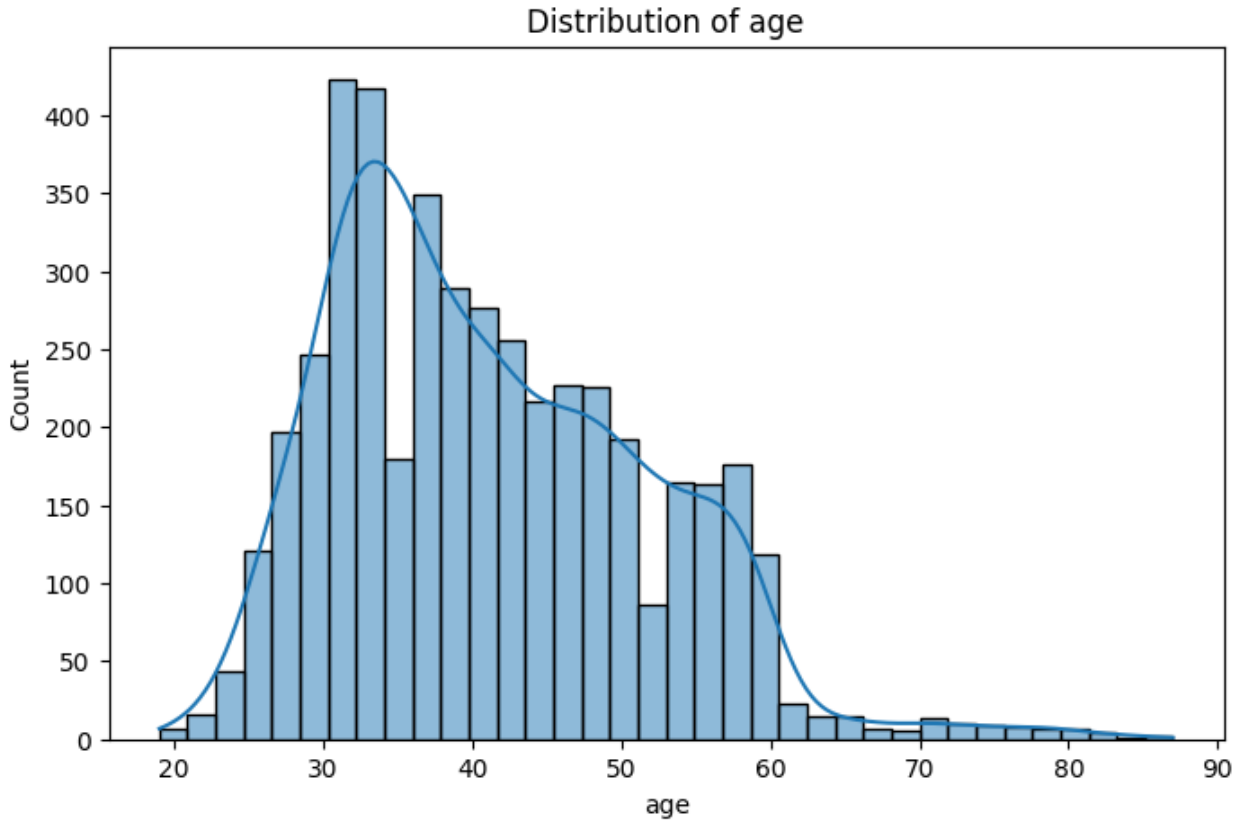
## Key Insights

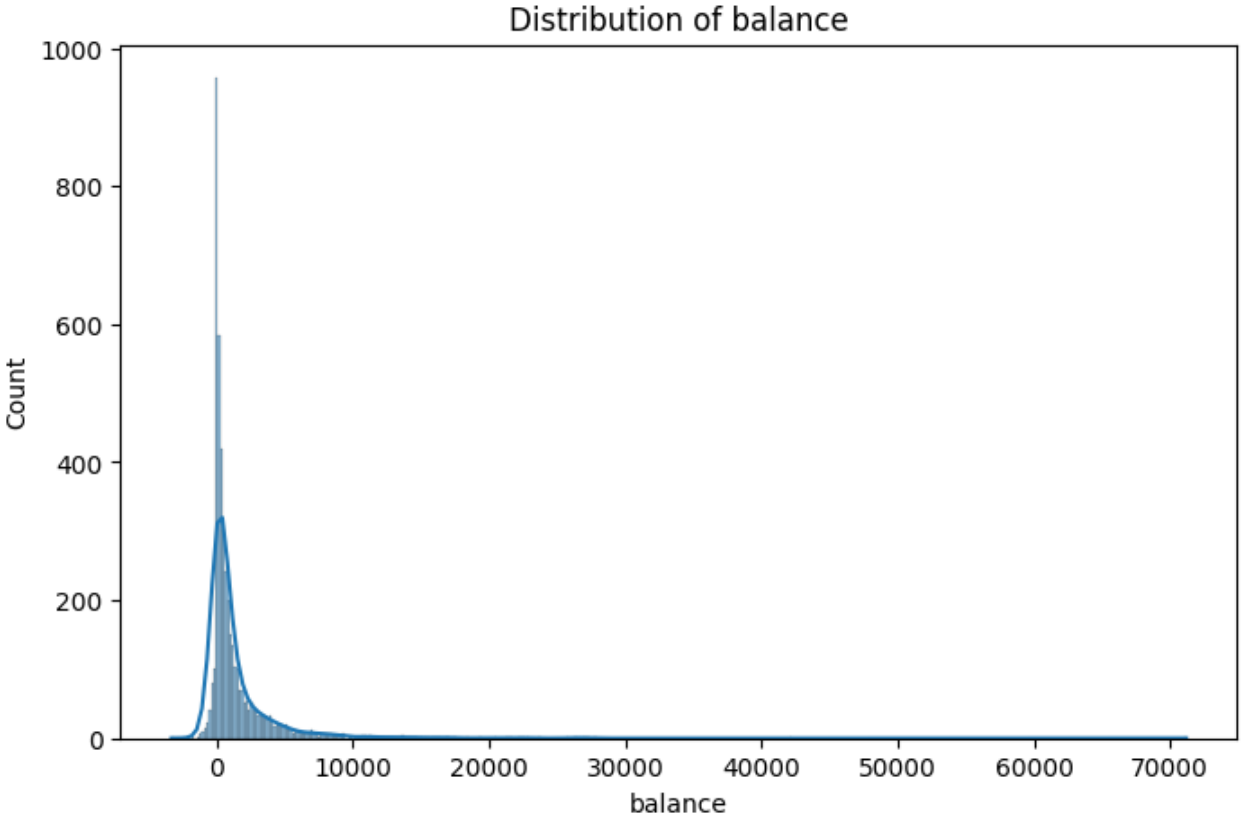
- **Subscription and Balance:** According to the analysis, there isn't much of a difference in account balance between subscribers and non-subscribers. This implies that a customer's decision to subscribe may depend on factors other than balance.
- **Work and Subscription:** It has been discovered that an individual's subscription status is correlated with the kind of work they do. This suggests that a person's choice to subscribe or not may depend on the nature of their job.
- **Marital Status and Subscription:** Since there is a strong correlation between marital status and subscription status, marital status also affects subscription decisions.
- **Education and Subscription:** There is evidence that education and subscription are correlated, making education a significant predictor of subscription decisions.
- **Default and Subscription:** Having a credit default and being a subscriber do not significantly depend on each other, in contrast to the other criteria.
- **Personal and Housing Loans:** Having a personal loan or a housing loan is linked to subscription decisions, suggesting that loan availability may influence a customer's decision to subscribe.
- **Contact Method and Subscription:** Subscription status determines which contact method is selected, suggesting that communication style might affect subscription results.
- **Month and Poutcome:** It has been discovered that there is a substantial correlation between subscription decisions and the month of the marketing campaign as well as the results of a prior campaign.

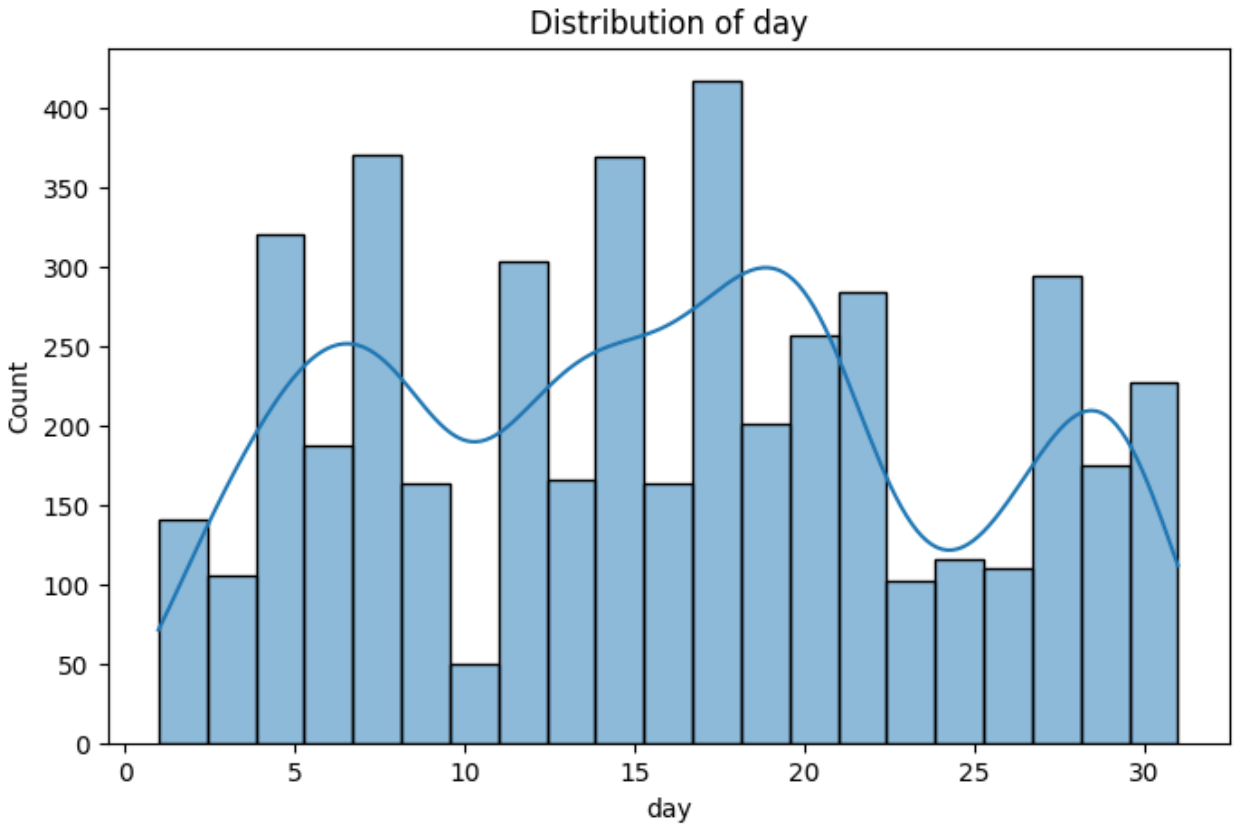
- **Balance by Education Levels**: While balance doesn't vary significantly by marital status, it does exhibit variation across different education levels. Higher education levels are associated with higher account balances.

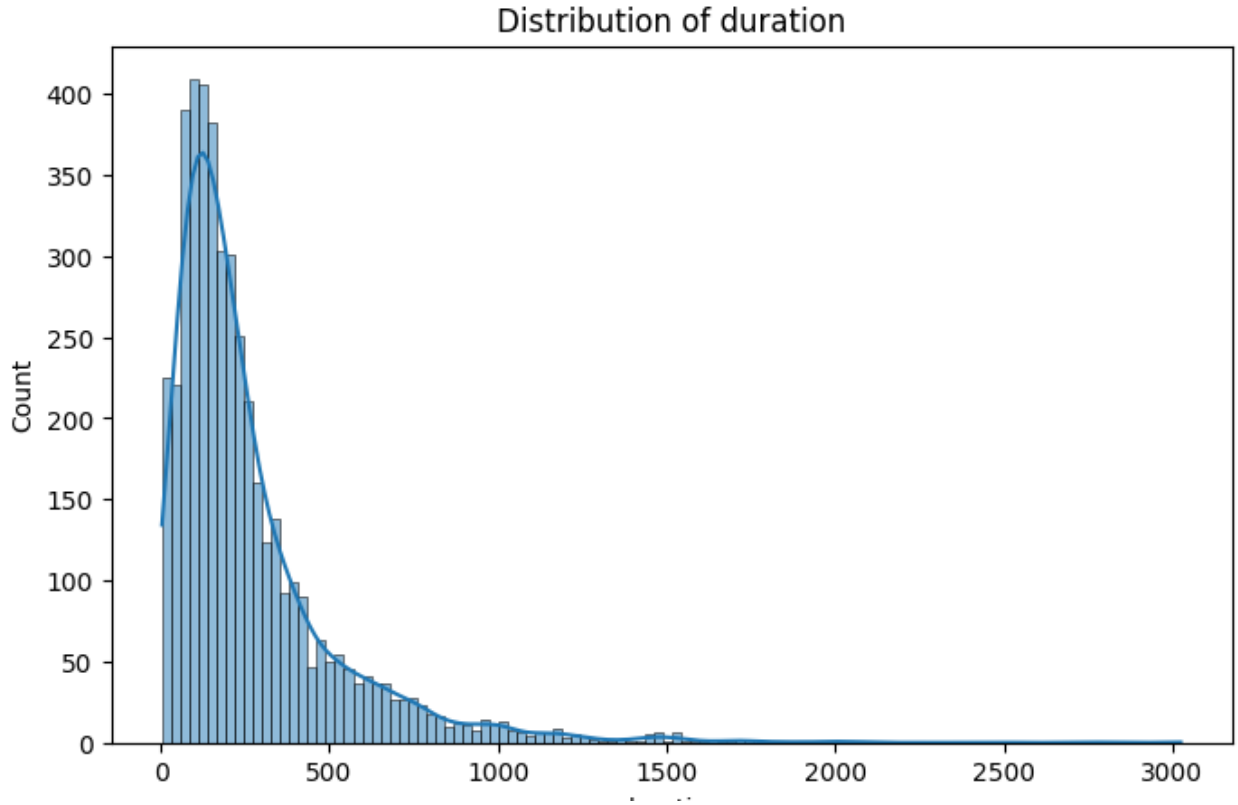
## Some Continuous Data Visualizations

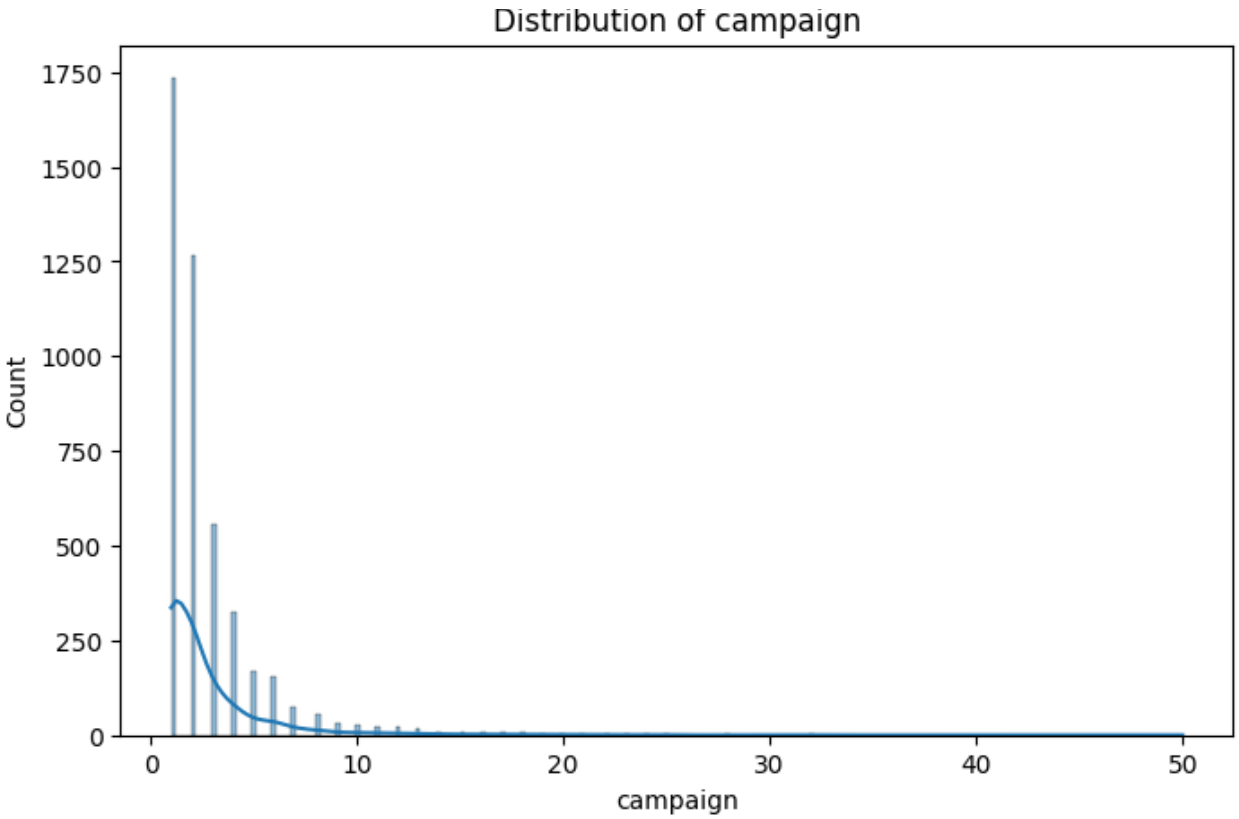



Distribution of age
Count
age
0
50
100
150
200
250
300
350
400
20
30
40
50
60
70
80
90

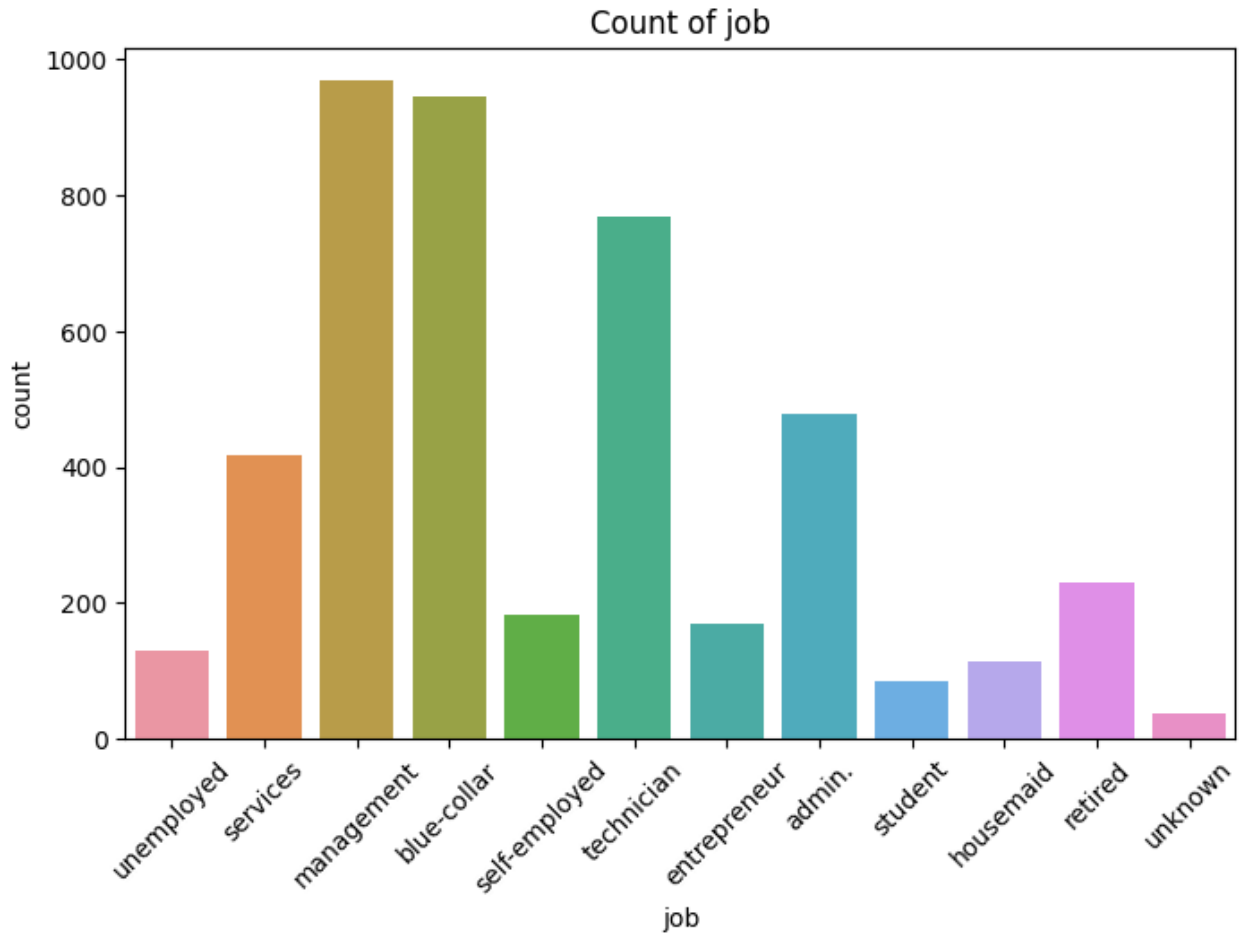

Distribution of balance
Count
balance
0
200
400
600
800
1000
0
10000
20000
30000
40000
50000
60000
70000

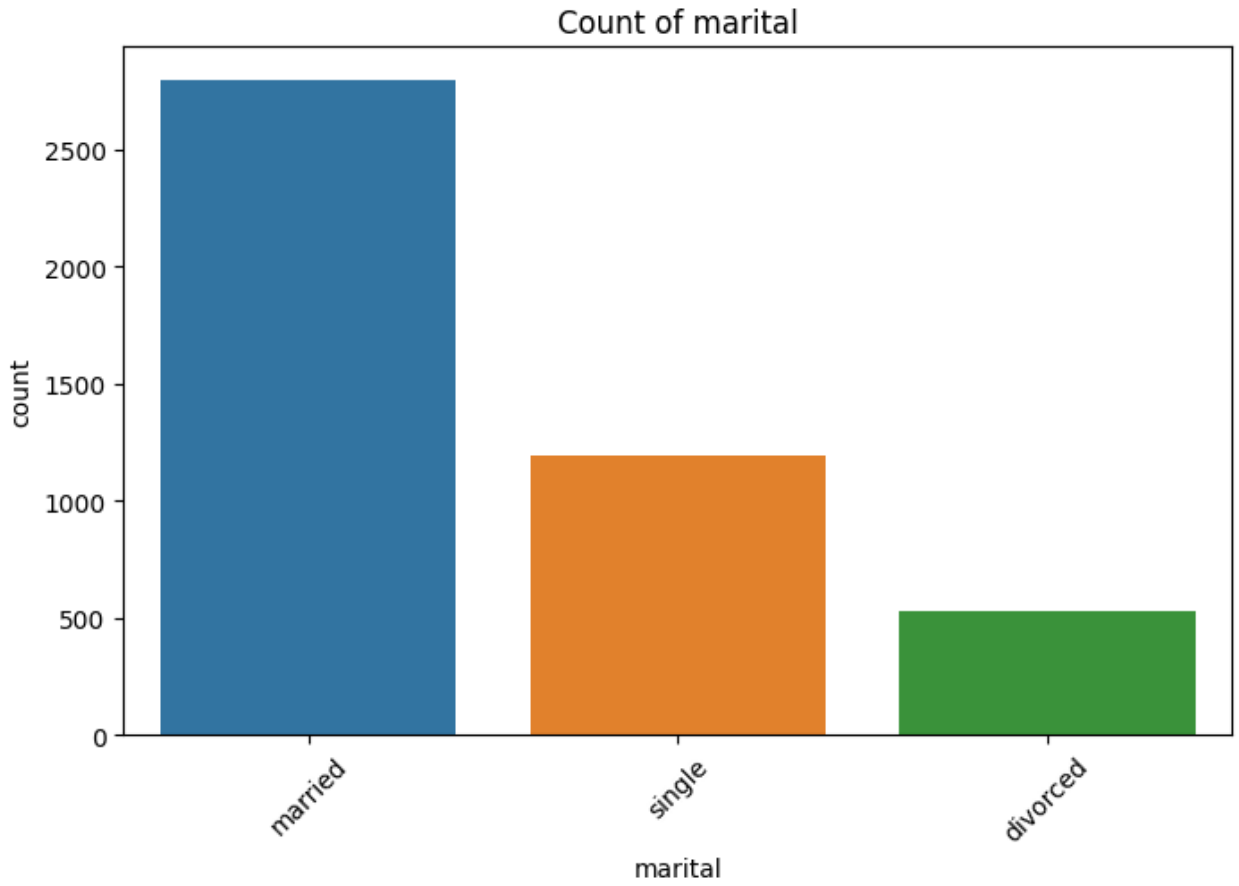

Distribution of day
Count
day
0
50
100
150
200
250
300
350
400
0
5
10
15
20
25
30

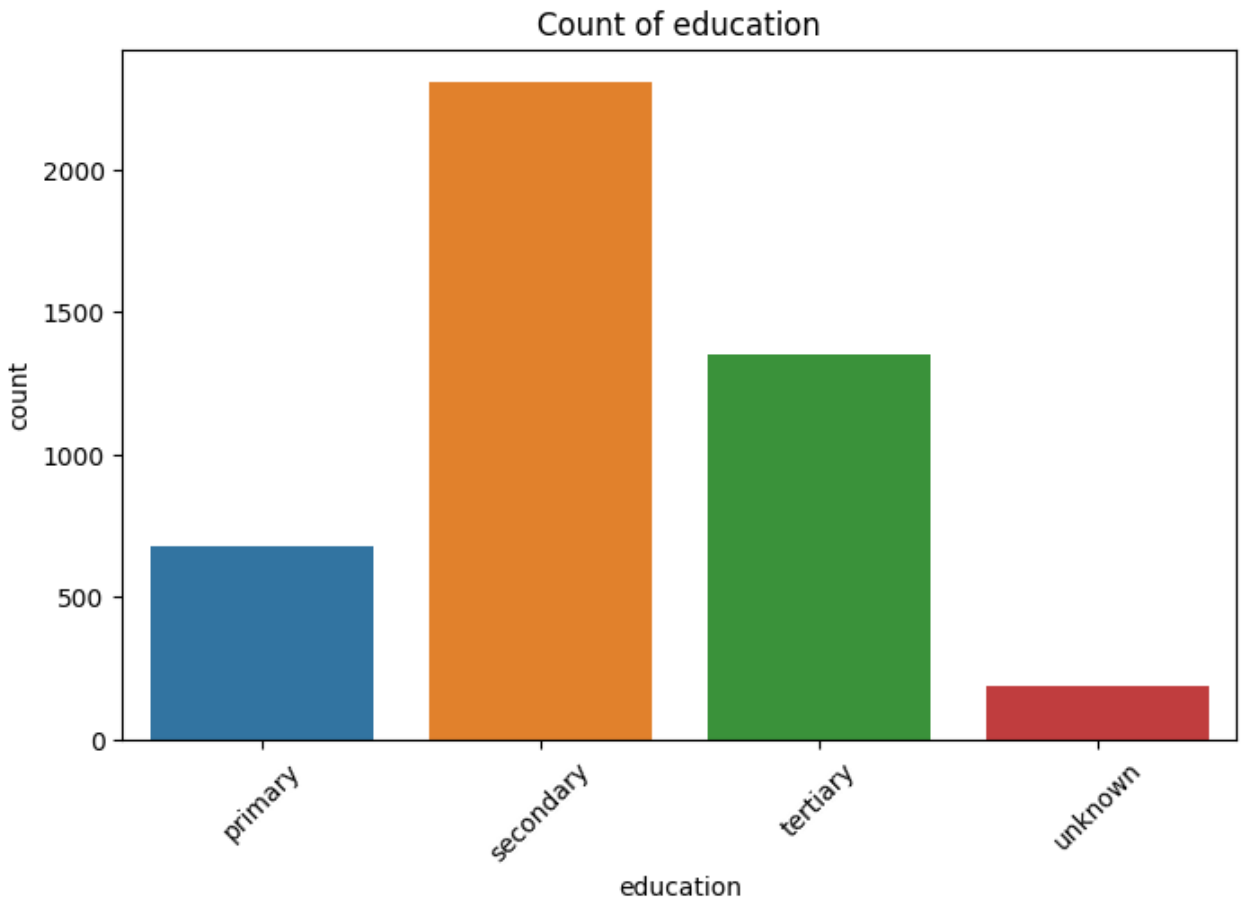

Distribution of duration
Count
0
50
100
150
200
250
300
350
400
0
500
1000
1500
2000
2500
3000

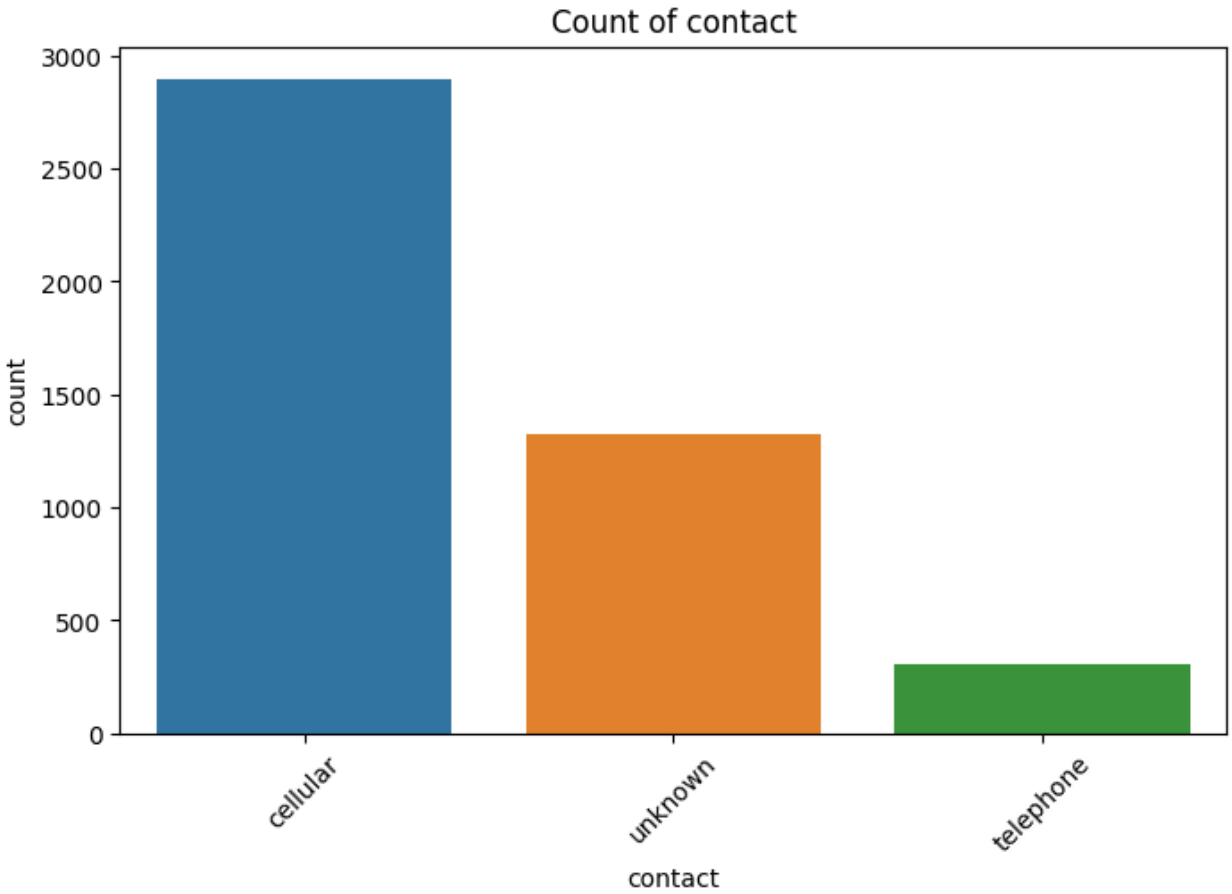

Distribution of campaign
Count
campaign
0
250
500
750
1000
1250
1500
1750
0
10
20
30
40
50

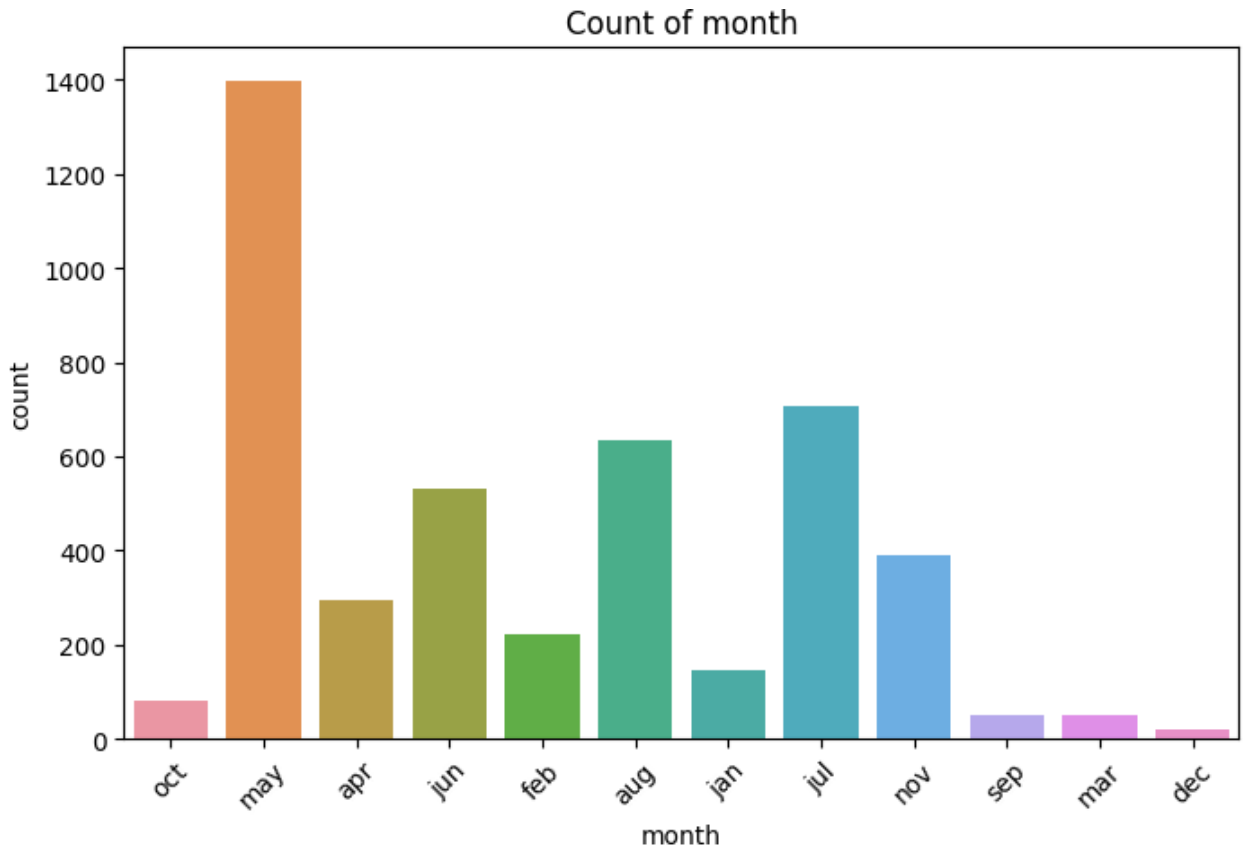
Some Continuous Data Visualizations

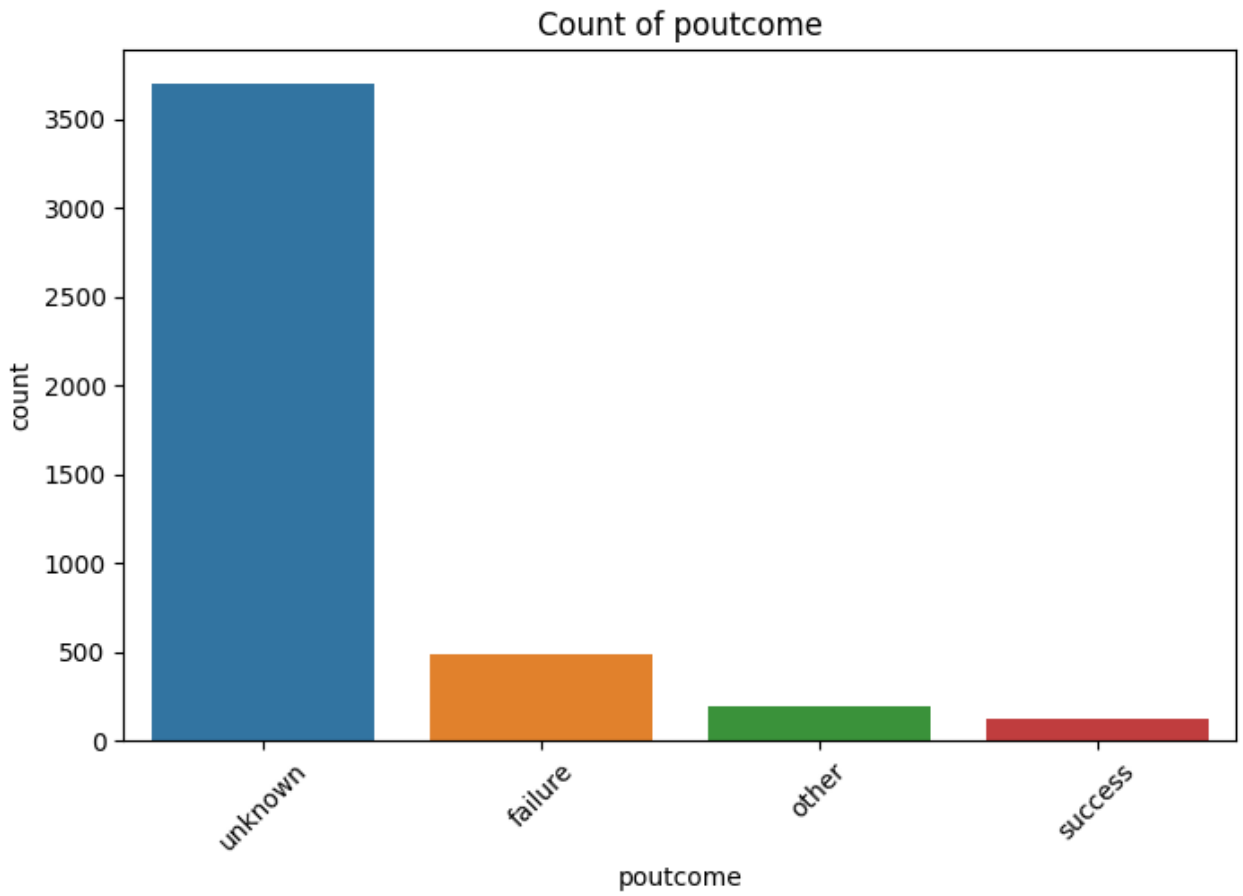


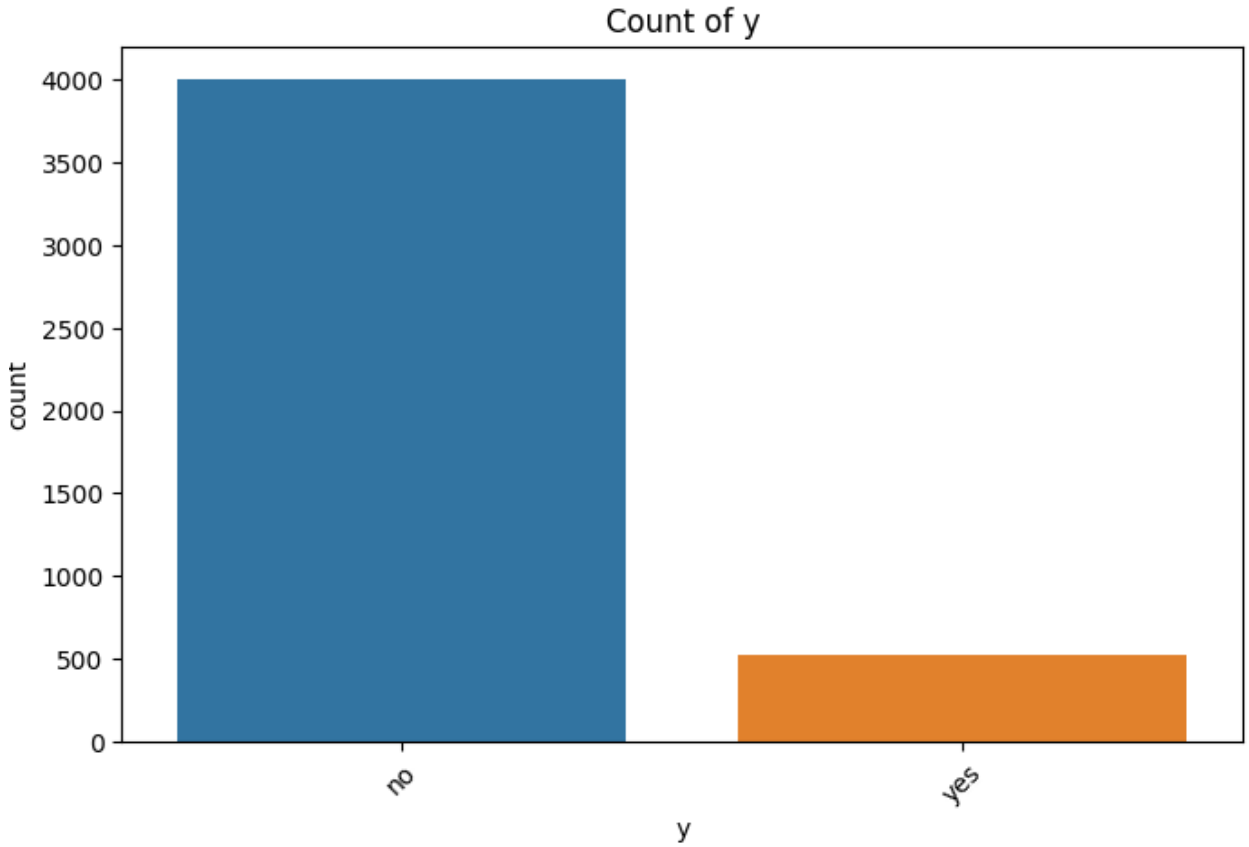

Count of marital
count
2500
2000
1500
1000
500
0
married
single
divorced
marital


Count of education
count
2000
1500
1000
500
0
primary
secondary
tertiary
unknown
education


Count of contact
3000
2500
2000
1500
1000
500
0
count
cellular
unknown
telephone
contact


Count of month
count
1400
1200
1000
800
600
400
200
0
oct
may
apr
jun
feb
aug
jan
jul
nov
sep
mar
dec
month


Count of poutcome
count
3500
3000
2500
2000
1500
1000
500
0
unknown
failure
other
success
poutcome


Count of y
4000
3500
3000
2500
2000
1500
1000
500
0
count
no
yes
y

# Recommendations

Based on the insights derived from this analysis, the following recommendations are proposed:

- **Segmentation:** To target marketing campaigns to particular groups, divide up the consumer base according to their occupation, marital status, level of education, and other demographic variables.
- **Time-Based Campaigns:** Considering seasonality and past campaign results data, think about running campaigns during particular months.
- **Loan-Specific Offers:** Create tailored offers for clients who have personal and home loans, since these things affect their choice to subscribe.
- **Contact Method Optimization:** Make the most of the "cellular" contact approach and think about customizing messages for various channels of communication.
- **Customer Education:** Look into the possibility of offering services or financial education to clients who have smaller balances but might use more financial awareness.
- **Feedback Loop:** Establish a feedback loop to monitor the results of earlier campaigns and modify them for subsequent ones.

Important insights into the variables influencing subscription decisions have been obtained from this investigation. Businesses can improve their marketing strategy and increase the success rate of obtaining new clients by utilizing this information.

# Exploratory Data Analysis

## Key Insights

**Age Distribution**

- Age Distribution: It looks like age is somewhat skewed to the right and is mostly normally distributed.
- Implication: Since most of the clientele are between the ages of 30 and 60, it makes sense that the campaign targets this demographic.

**Job Categories**

- Job Categories: Based on customer feedback, "blue-collar," "management," and "technician" are the most popular job categories.
- Implication: Based on a client's employment category, customized offers or communications can be created for them.

**Marital Status**

- Marital Status Insight: "Married" is the most common status among clients, followed by "single" and "divorced."
- Implication: Marketing plans can be created to successfully target a range of marital statuses.

**Education Levels**

- Education Levels: "Primary," "tertiary," and "secondary" are the most common educational levels among our clients.
- Implication: To improve engagement, campaigns may take into account the clients' educational backgrounds.

**Credit Default**

- Analysis of Credit Defaults: Very few customers experience a credit default.
- Implication: To foster trust, marketing materials should highlight how rare credit default is.

**Balance Distribution**

- Balance Distribution Insight: The majority of clients have relatively low balances, indicating a right-skewed balance distribution.
- Implication: Customers with smaller account balances can be eligible for exclusive financial deals or services.

**Housing and Personal Loans**

- Observation: A housing loan is held by the majority of consumers, although a small percentage of them have personal loans.
- Implication: Loan availability should be taken into account in marketing campaigns as it may affect subscribers' decisions.

**Contact Methods**

- Observation: The majority of clients were reached by phone.
- Implication: Given its effectiveness, cellular communication might be the favored means of contact.

**Contact Month**

- Information about Contact Month: "May" is the month with the most contacts.
- Implication: The campaign schedule may be influenced by seasonal patterns in the contact months.

**Duration of Last Contact**

- Observation: There is a right-skewed distribution in the length of the most recent contact.
- Implication: Campaign managers are able to evaluate the correlation between call length and results related to subscriptions.

**Number of Contacts**

- Count of Interactions Analysis: Throughout the campaign, the majority of clients received very few communications.
- Implication: These insights can be used to identify the ideal frequency of contact for the campaign.

**Previous Contacts**

- Prior Contacts: Information: Clients that have been contacted in the past show a wide range of days between contacts.
- Implication: Follow-up tactics can be created by utilizing past contact information.

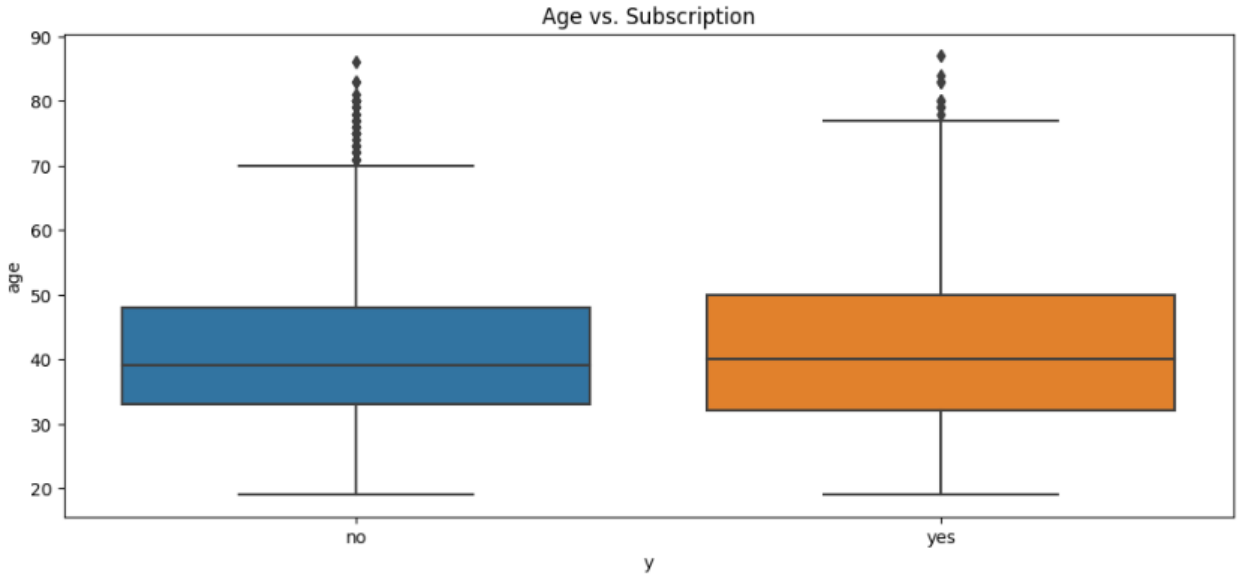
**Previous Campaign Outcomes**

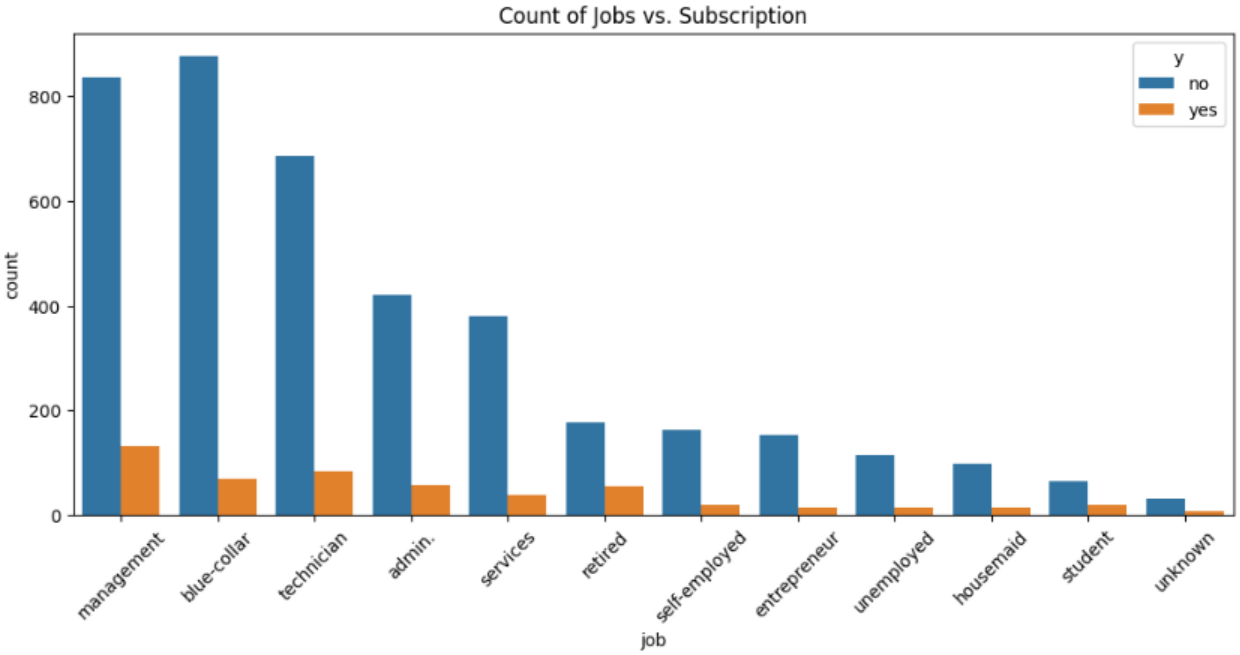
- Prior Campaign Results: "Unknown" is the most typical prior campaign result, followed by "failure."
- Implication: The large percentage of "unknown" results can point to the necessity of more thorough documentation in subsequent efforts.

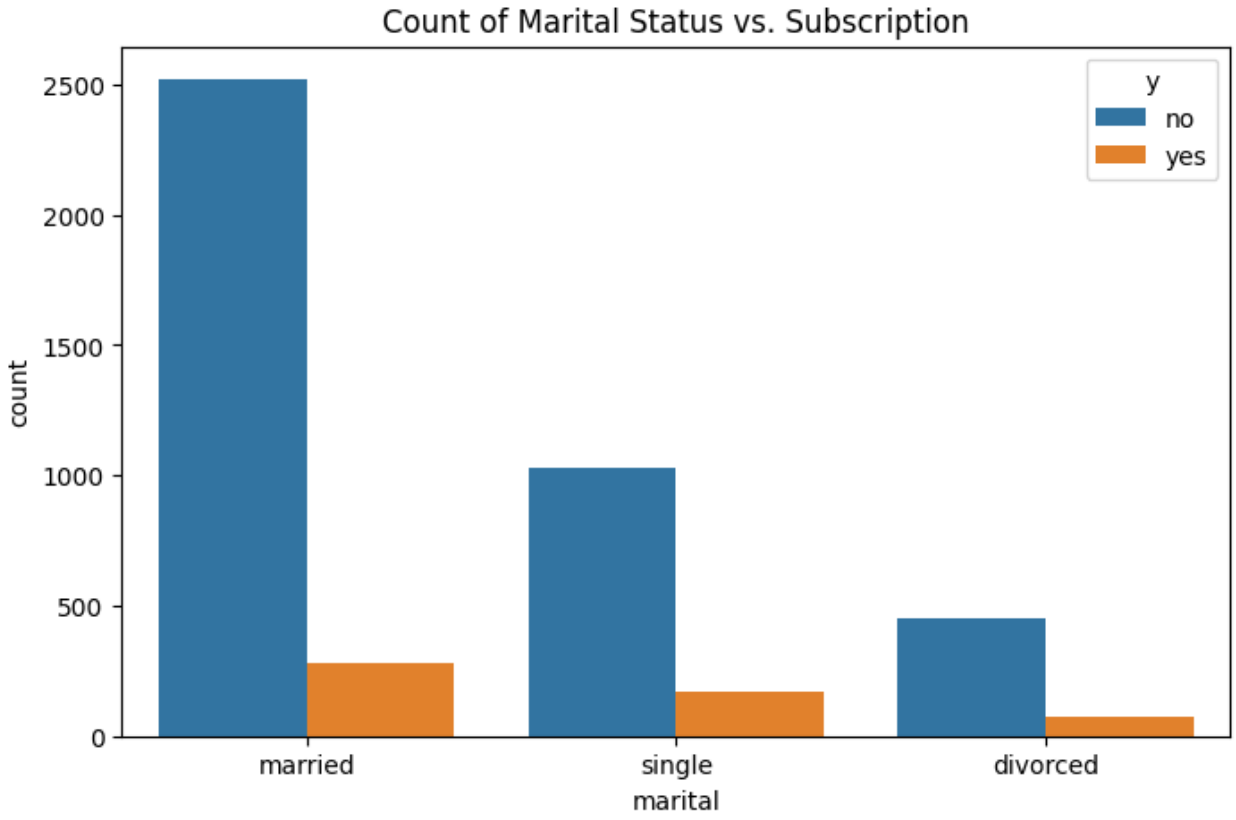
**Subscription Imbalance**

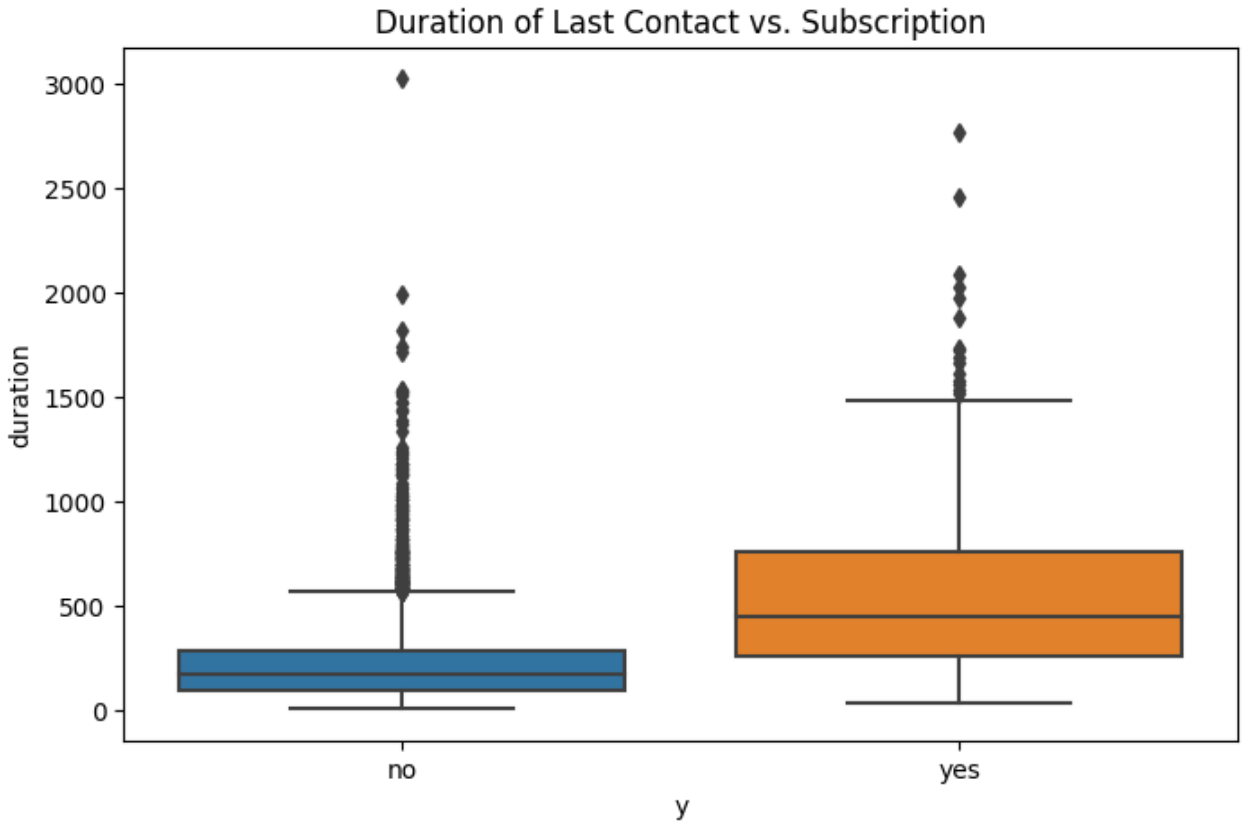
- Observation: There is an imbalance in the dataset as a large proportion of clients do not subscribe.
- Implication: When doing analysis and organizing campaigns, models, and strategies need to take this imbalance into consideration.

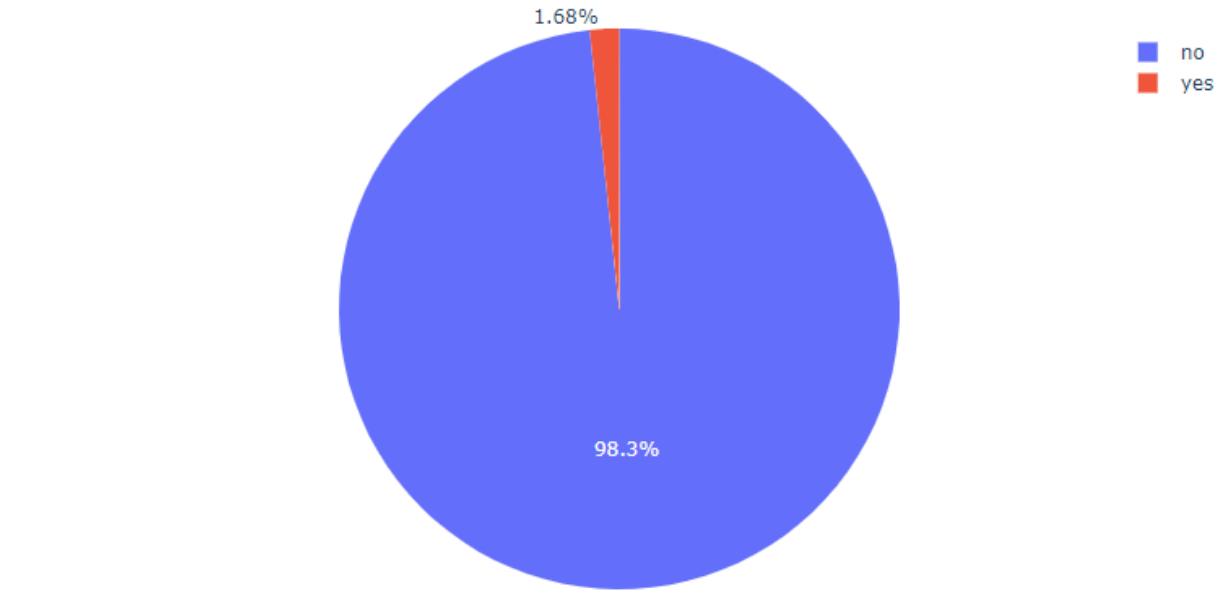
# Some EDA Visualizations




Count of Marital Status vs. Subscription
y
no
yes
count
2500
2000
1500
1000
500
0
married
single
divorced
marital


Duration of Last Contact vs. Subscription
duration
3000
2500
2000
1500
1000
500
0
no
yes
y


Distribution of Credit Default
1.68%
98.3%
no
yes

# Business Problems and Possible Solutions

By utilizing the statistical and EDA insights obtained from the dataset, we are able to tackle various business issues and suggest methods to enhance the efficacy of marketing campaigns. The following are business issues and the fixes for them:

**First Business Issue: Low Rate of Subscription**

**Observation:** There is an imbalance in the dataset as a large proportion of clients do not subscribe.

**Solution:** Take into account the following tactics to raise the subscription rate:

- Create specialized advertising campaigns: Create marketing materials that are tailored to particular customer segments and highlight the advantages of subscribing.
- Rewards and customized offers: Give subscribers exclusive deals, discounts, or tailored offerings to entice them to subscribe.
- Pay attention to past campaign results: To gain more insight into the behavior and preferences of your customers, evaluate and enhance the outcomes of prior campaigns.

**Second Business Issue: Age-Based Targeting**

**Observation:** Most of the clientele are in the age range of thirty to sixty.

**Solution:** Think about age-based targeting:

- Tailor promotional materials: Provide content that appeals to a range of age groups. For instance, concentrate on financial objectives for younger clients and retirement planning for older clients.
- When to run campaigns: Plan your campaigns for when the various age groups are most likely to be accessible and responsive.

**Third Business Issue: Strategies Dependent on Occupation**

**Observation:** Work and membership are interdependent.

**Solution:** Create marketing plans tailored to your profession:

- Task-oriented messaging: Adapt message to each job category's particular demands and interests.
- Industry collaborations: Work together with employers or related sectors to provide special services or promotions.

**Fourth Business Issue: Leveraging the Influence of Marital Status**

**Observation:** Subscription and marital status are related.

**Solution:** Make use of marital status data to boost marketing performance:

- Family-focused advertisements: Create campaigns that highlight the demands of married clients' families and more customized strategies for singles.
- Social networks: Inspire married customers to promote and recommend the services to their spouses.

**Fiveth Business Issue: Focusing on Educational Background**

**Observation:** Learning and membership are interdependent.

**Solution:** Develop marketing initiatives targeted toward education:

- Instructional materials: Provide content that is appropriate for the clients' educational background.
- Instructional workshops: Adapt classes and events to the educational background of your clientele.

**Sixth Business Issue: Optimization of Contact Methods**

**Observation:** The majority of clients were reached by phone.

**Solution:** Maximize communication channels and methods:

- Ensure that all marketing materials are optimized for mobile devices.
- Multi-channel strategy: Look into more efficient means of communication with clients.

**Seventh Business Issue: Marketing of Loan Products**

**Observation:** Few clients have personal loans, whereas the majority have house loans.

**Solution:** Effectively market loan products.

- Customized loan offers are those that are made in accordance with the client's financial circumstances and current loans.
- Cross-promotional prospects: Determine which housing loan clients would benefit from additional financial products, such as personal loans.

**Eighth Business Issue: Get in touch with Frequency Management**

**Observation:** Throughout the campaign, the majority of clients received very few communications.

**Solution:** Find the ideal frequency of contact:

Testing A/B: Try out various touch frequencies to determine the ideal ratio of interaction to irritation.

**Ninth Business Issue: Managing "Unknown" Results**

**Observation:** The most typical past campaign result was "Unknown".

**Solution:** Resolve the problem of "unknown" outcomes:

- Enhanced data gathering Put measures into place to collect more thorough data regarding customer responses.
- Client input: Invite customers to share their opinions and the rationale behind their choices.

**Tenth Business Issue: Increasing Credits to Increase Subscription Rate**

**Observation:** The majority of clients do not experience credit defaults.

**Solution:** Profit from the rarity of credit defaults:

- Credibility-driven campaigns: Point out the advantages of a subscription to customers who have good credit records.

# Predictive Analysis: Decision Tree VS Naive Bayes

Decision Tree with Hyperparameter Tuning:

Classification Report for Best Decision Tree:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| no | 0.93 | 0.95 | 0.94 | 807 |
| yes | 0.49 | 0.38 | 0.43 | 98 |
| accuracy | | | 0.89 | 905 |
| macro avg | 0.71 | 0.67 | 0.68 | 905 |
| weighted avg | 0.88 | 0.89 | 0.88 | 905 |

Decision Tree without Hyperparameter Tuning:

Classification Report for Decision Tree:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| no | 0.92 | 0.93 | 0.93 | 807 |
| yes | 0.38 | 0.36 | 0.37 | 98 |
| accuracy | | | 0.87 | 905 |
| macro avg | 0.65 | 0.64 | 0.65 | 905 |
| weighted avg | 0.86 | 0.87 | 0.87 | 905 |

An accuracy of 0.89 was attained by the Decision Tree model with hyperparameter adjustment. It demonstrated a reasonably balanced performance, demonstrating the model's ability to recognize true positives with a precision of 0.49 for the 'yes'. The recall of the 'yes' class was 0.38, indicating a limited capacity of the model to accurately detect all true positives. The 'yes' class's F1-score was 0.43, suggesting a fair balance between recall and precision.

With hyperparameter tuning off, the Decision Tree model's accuracy was 0.87. For the 'yes' class, it showed a lower precision of 0.38 and a recall of 0.36. For the 'yes' class, the F1 score was 0.37. Although this model functioned satisfactorily, the adjusted Decision Tree model outperformed it.

Naive Bayes with Hyperparameter Tuning:

```
Classification Report for Best Naive Bayes:
              precision    recall  f1-score   support

          no       0.93      0.92      0.92       807
         yes       0.38      0.41      0.40        98

    accuracy                           0.87       905
   macro avg       0.66      0.66      0.66       905
weighted avg       0.87      0.87      0.87       905
```

Naive Bayes without Hyperparameter Tuning:

```
Classification Report for Naive Bayes:
              precision    recall  f1-score   support

          no       0.93      0.92      0.92       807
         yes       0.38      0.41      0.40        98

    accuracy                           0.87       905
   macro avg       0.66      0.66      0.66       905
weighted avg       0.87      0.87      0.87       905
```

An accuracy of 0.87 was attained with the hyperparameter-tuned Naive Bayes model. For the 'yes' class, it showed balanced performance with a precision of 0.38, demonstrating the model's capacity to detect true positives. Recall for the 'yes' class was 0.41, indicating that the model could detect true positives. The 'yes' class's F1-score was 0.40, suggesting a fair balance between recall and precision.

With no hyperparameter adjustments, the accuracy of the Naive Bayes model was 0.87 as well. For the 'yes' class, it displayed an F1-score of 0.40 with a precision of 0.38 and a recall of 0.41. This model performed similarly to the modified Naive Bayes model.

# The Difference

**Accuracy Comparison**:

- Decision Tree (Hyperparameter Tuned): Accuracy of 0.89.
- Naive Bayes (Hyperparameter Tuned): Accuracy of 0.87.

**Precision and Recall for 'Yes' Class**:

- Decision Tree (Hyperparameter Tuned): Precision ('yes') of 0.49 and Recall ('yes') of 0.38.
- Naive Bayes (Hyperparameter Tuned): Precision ('yes') of 0.38 and Recall ('yes') of 0.41.

**F1-Score for 'yes' Class**:

- Decision Tree (Hyperparameter Tuned): F1-Score ('yes') of 0.43.
- Naive Bayes (Hyperparameter Tuned): F1-Score ('yes') of 0.40.

**Hyperparameter Tuning Impact**:

- The Decision Tree model demonstrated a notable improvement in accuracy (0.87 to 0.89) after hyperparameter tuning.
- The Naive Bayes model's accuracy remained the same (0.87) with and without hyperparameter tuning.

**Model Complexity**:

- Decision Trees can create more complex decision boundaries, which allows them to capture intricate relationships in the data. However, this complexity may lead to overfitting.
- Naive Bayes models are simpler and make strong independence assumptions between features. This simplicity may lead to underfitting, especially when features are correlated.

**Interpretability**:

- Decision Trees provide interpretability as they create a clear decision path based on features.
- Naive Bayes models are also interpretable, as they make assumptions about the probability distribution of features.

**Sensitivity to Feature Distribution**:

- Decision Trees are sensitive to the distribution of features and may perform poorly with imbalanced data.
- Naive Bayes models can handle imbalanced datasets relatively well.

**Applicability**:
- Decision Trees are versatile and suitable for both classification and regression tasks.
- Naive Bayes models are commonly used for classification problems, especially in text and document classification.

**Decision Boundary**:
- Decision Trees create non-linear decision boundaries, which can capture complex patterns in the data.
- Naive Bayes models assume linear decision boundaries between classes.

# Recommendation

The Decision Tree model with hyperparameter tuning is advised for deployment based on the model evaluation. It performed better for the 'yes' class, which is usually the more important class for this business challenge, in terms of precision, recall, and F1-score. However, there are still ways to make improvements, such as feature engineering and experimenting with different classification methods like Gradient Boosting or Random Forest. There is no discernible difference between the two Naive Bayes models in terms of performance measures, either with or without hyperparameter adjustment. As a result, while deciding which model to use, one should take into account aspects like processing capacity, maintenance ease, and the particular needs of the company. To increase prediction accuracy, more optimization and testing with other categorization methods can be investigated.