Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Data Science Project: Bank Marketing Campaign

Submission Date: 15/08/2024

Batch Code: LISUM34

Group Name: Shiny Star Data Specialists

Team member's details:

Carmelo R. Casiraro, USA, Farmingdale State University, Data Analyst

Fenil Mavani, UK, University of West London, Data Science

Nazri, London, UK, University of Greenwich, Data Science

Agenda

- Problem Statement
- Dataset Information
- EDA
- EDA Summary
- Final Recommendations

Problem Statement

ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution). This is an application of the organization's marketing data

Objective

Build a Classification ML model to shortlist customers who are most likely to buy the term deposit product. This would allow the marketing team of the bank to target those customers through various marketing channels (tele marketing, SMS/email marketing etc)

Data Glacier
Your Deep Learning Partner

Dataset Information

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Dataset: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

bank-additional-full.csv: 20 inputs (including 1 target variable) and 41118 observations ordered by date (from May 2008 to November 2010)

bank-full.csv: 17 inputs (including 1 target variable) and 45211 observations (older version of the dataset with less inputs)

Dataset

Data Glacier
Your Deep Learning Partner

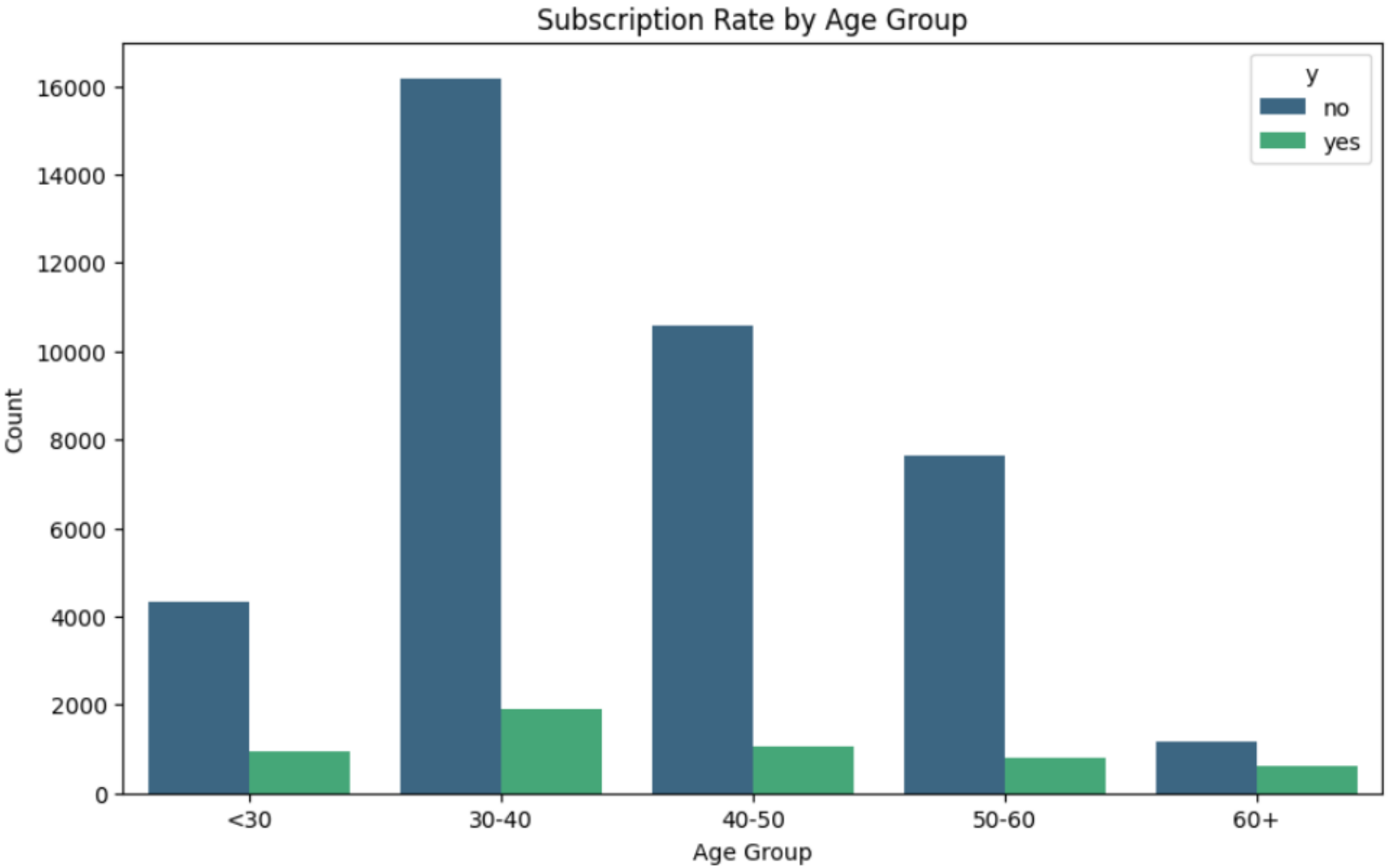
Attribute Information:

Input variables:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Exploratory Data Analysis-
Hypothesis Testing

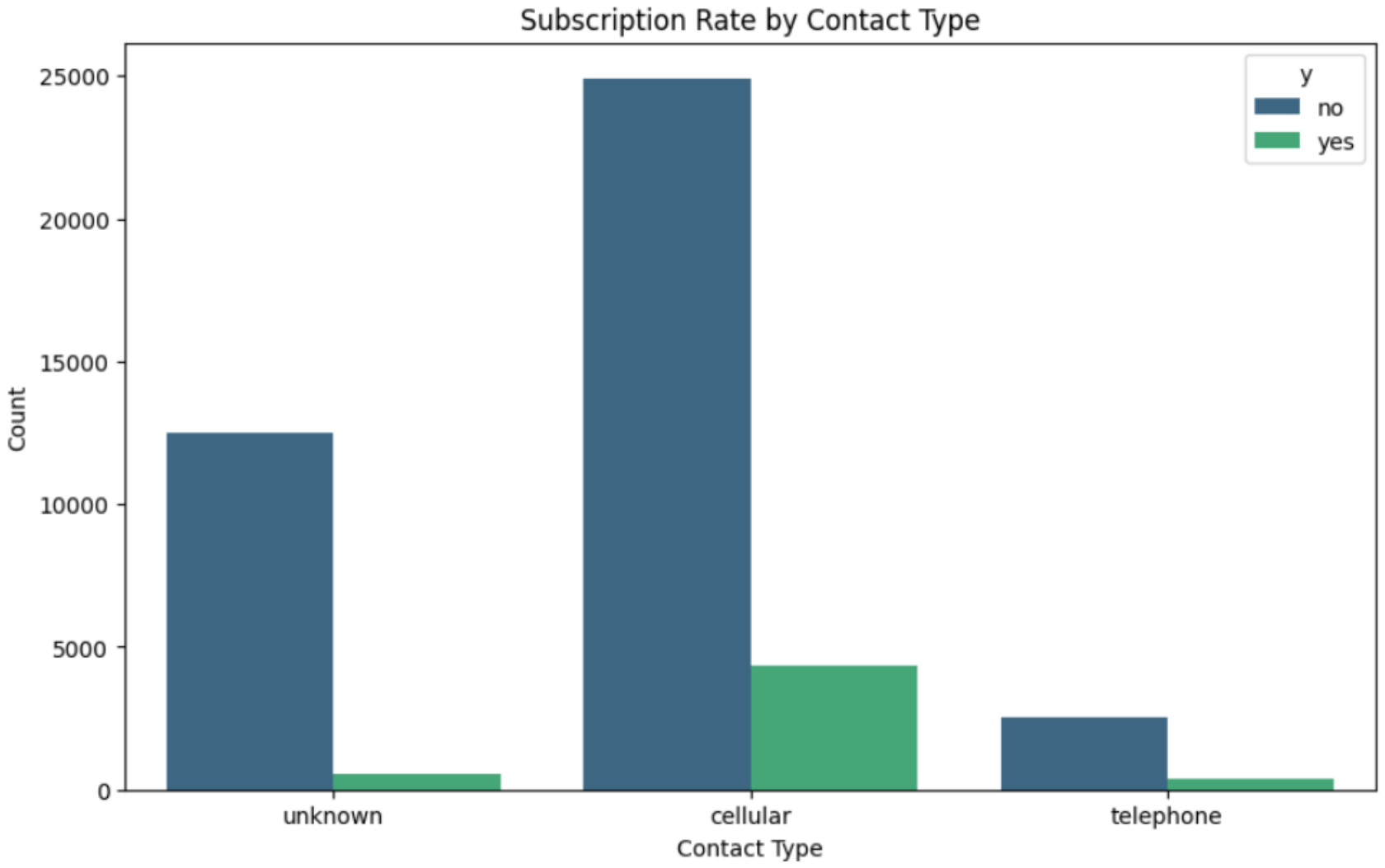
Subscription Rate by Age Group



Summary:

Subscription Rate decreases as age increases

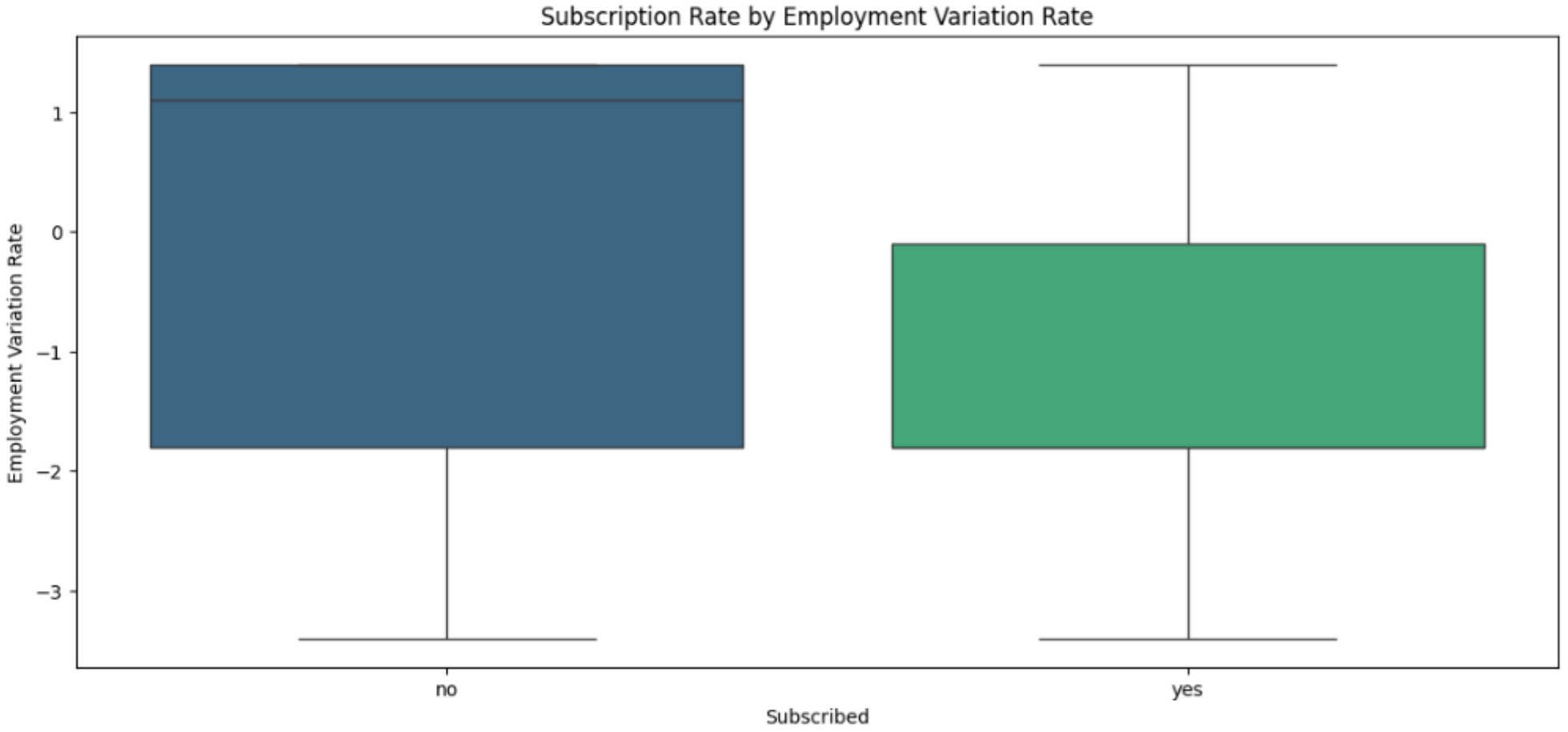
Subscription Rate by Contact Type



Summary:

Subscription rate increases with various forms of contact types.

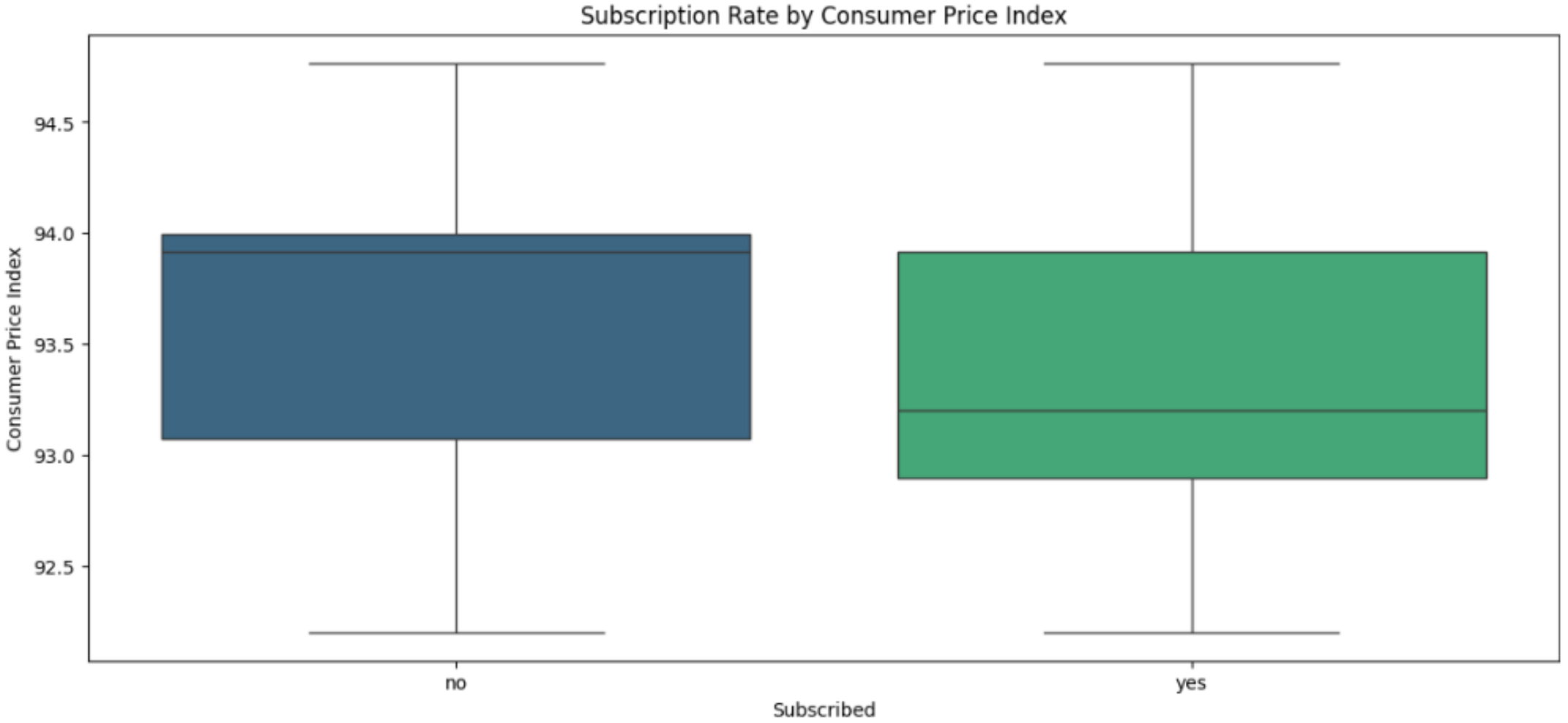
Subscription Rate by Employment Variation Rate



Summary:

Subscription rate will increase according to different economic indicators.

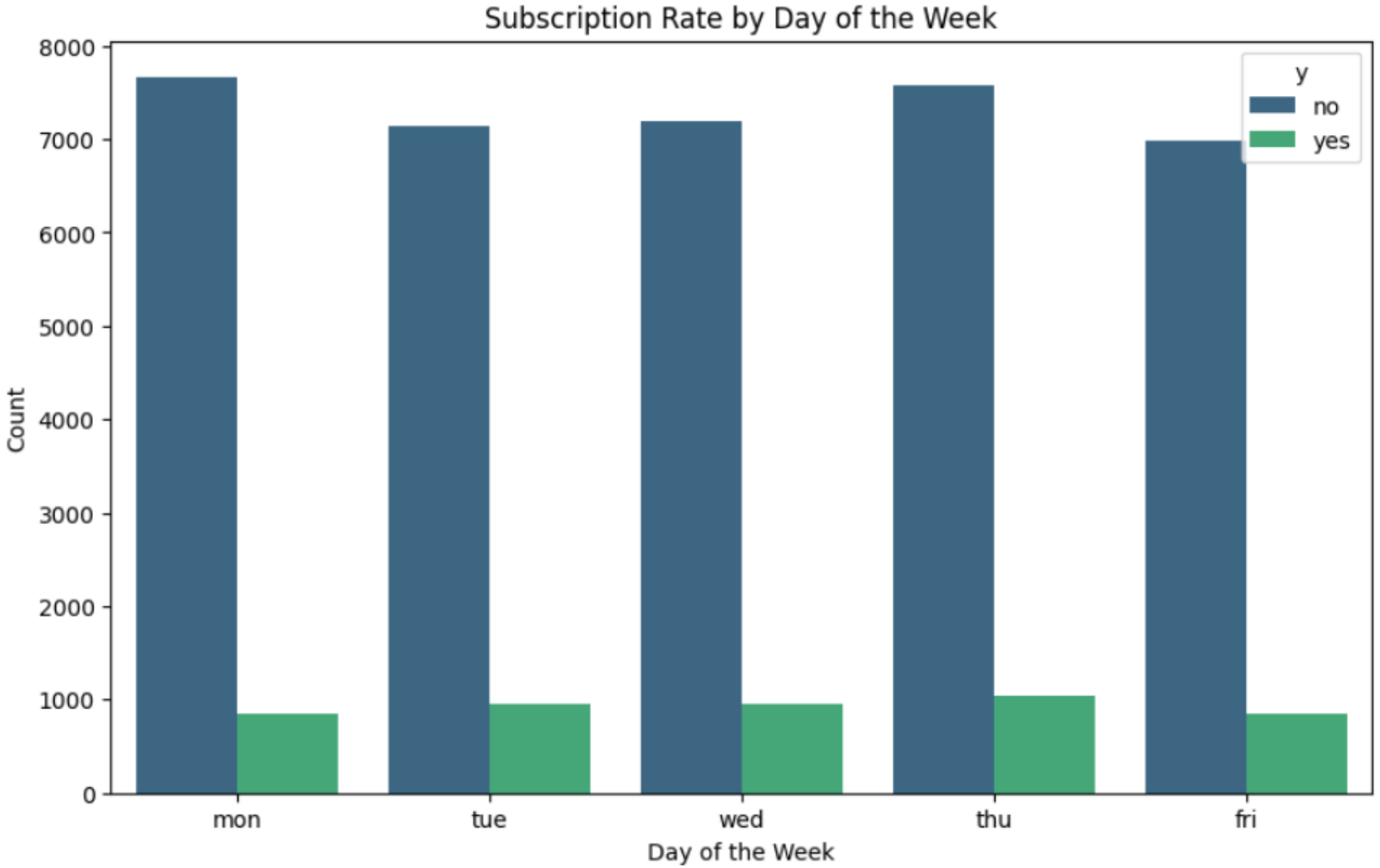
Subscription Rate by Consumer Price Index



Summary:

Subscription rate will increase according to different economic indicators.

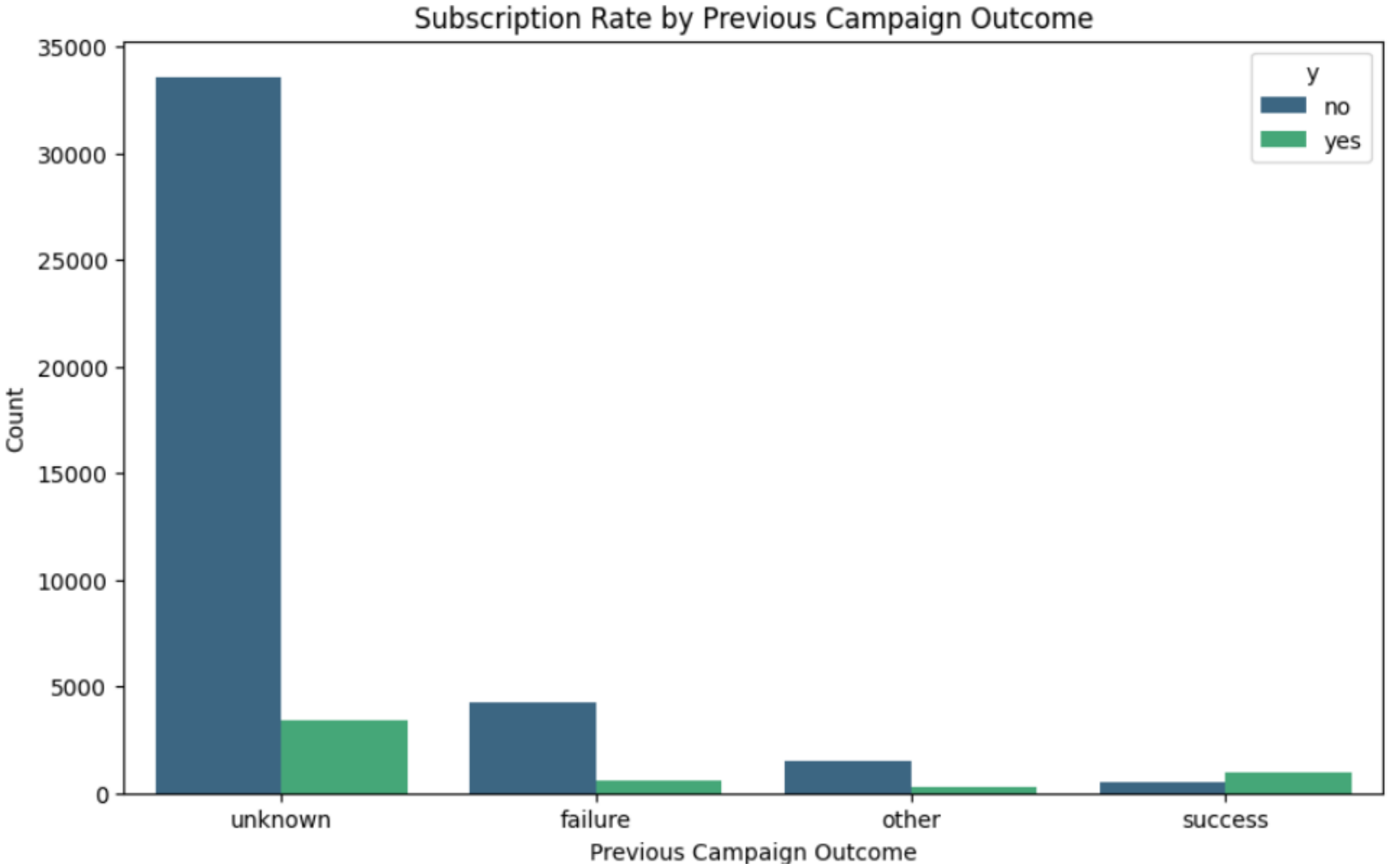
Subscription Rate by Day of the Week



Summary:

Success rate increases during different days of the week.

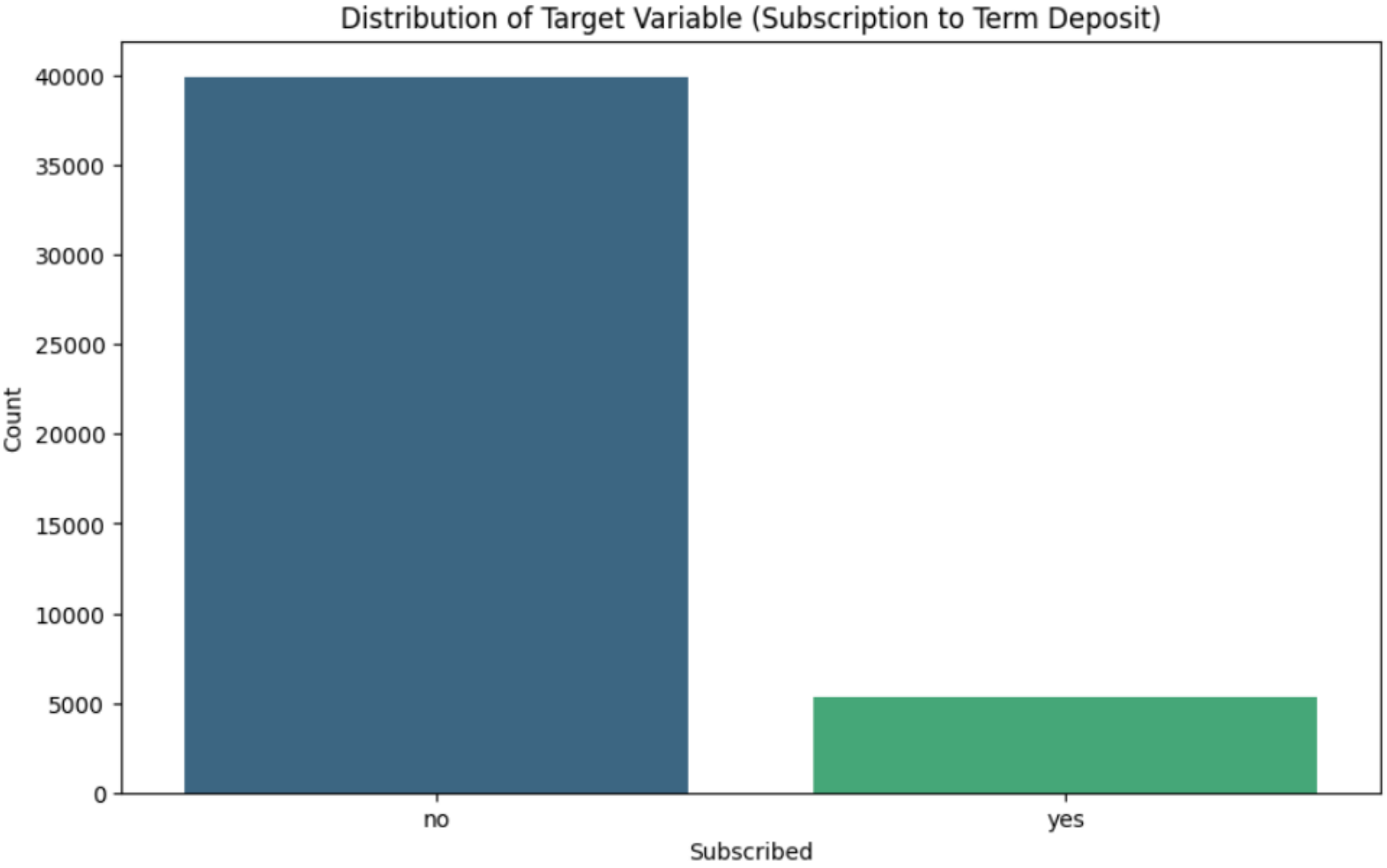
Subscription Rate By Previous Campaign Outcome



Summary:

Customers who had a positive outcome in previous campaigns are more likely to subscribe.

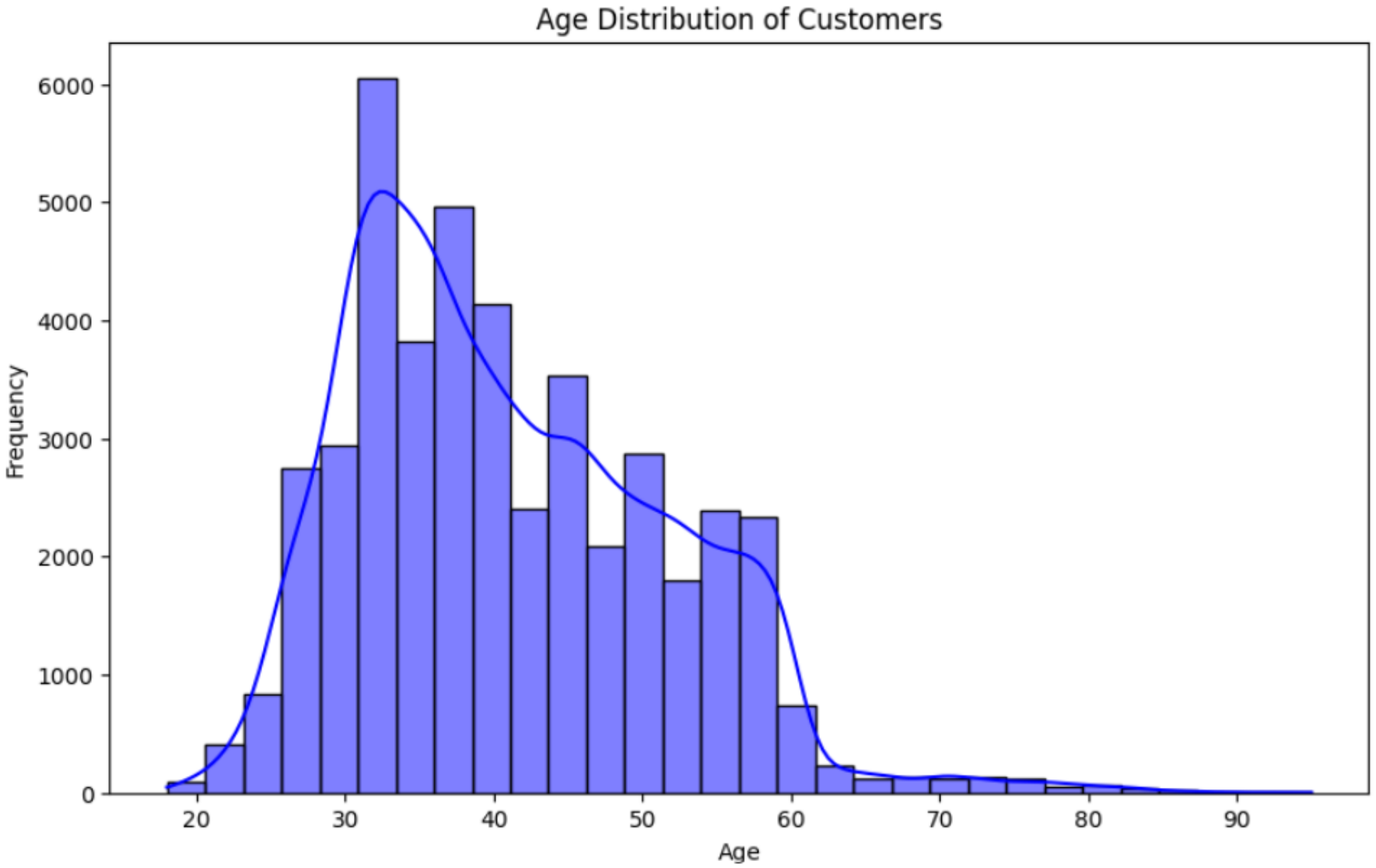
Subscription to the Term Deposit



Summary:

The distribution of subscription outcomes (target variable) follows a specific pattern influenced by key factors.

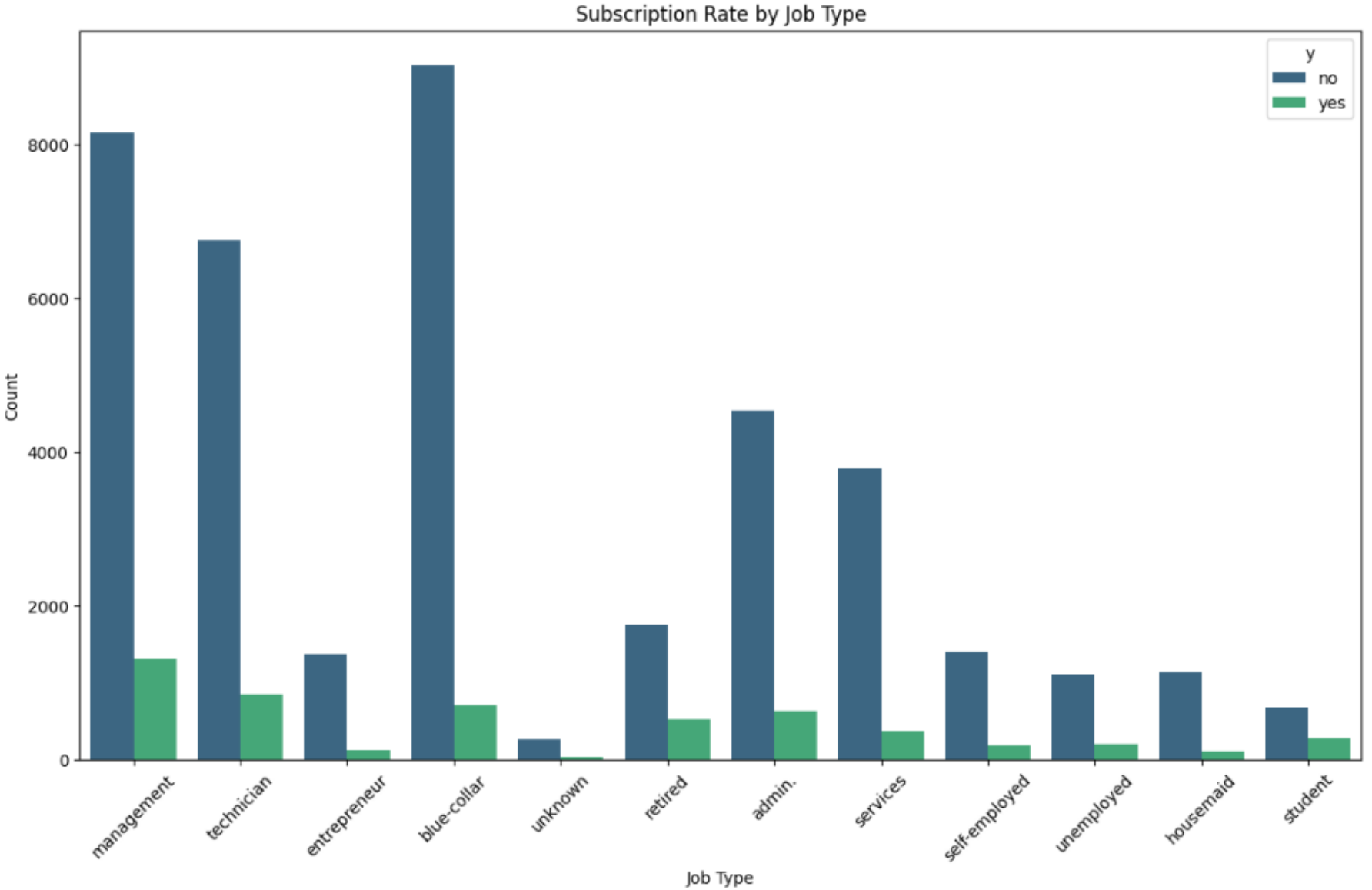
Age Distribution of Customers



Summary:

The age distribution of customers shows distinct patterns that affect their likelihood to subscribe.

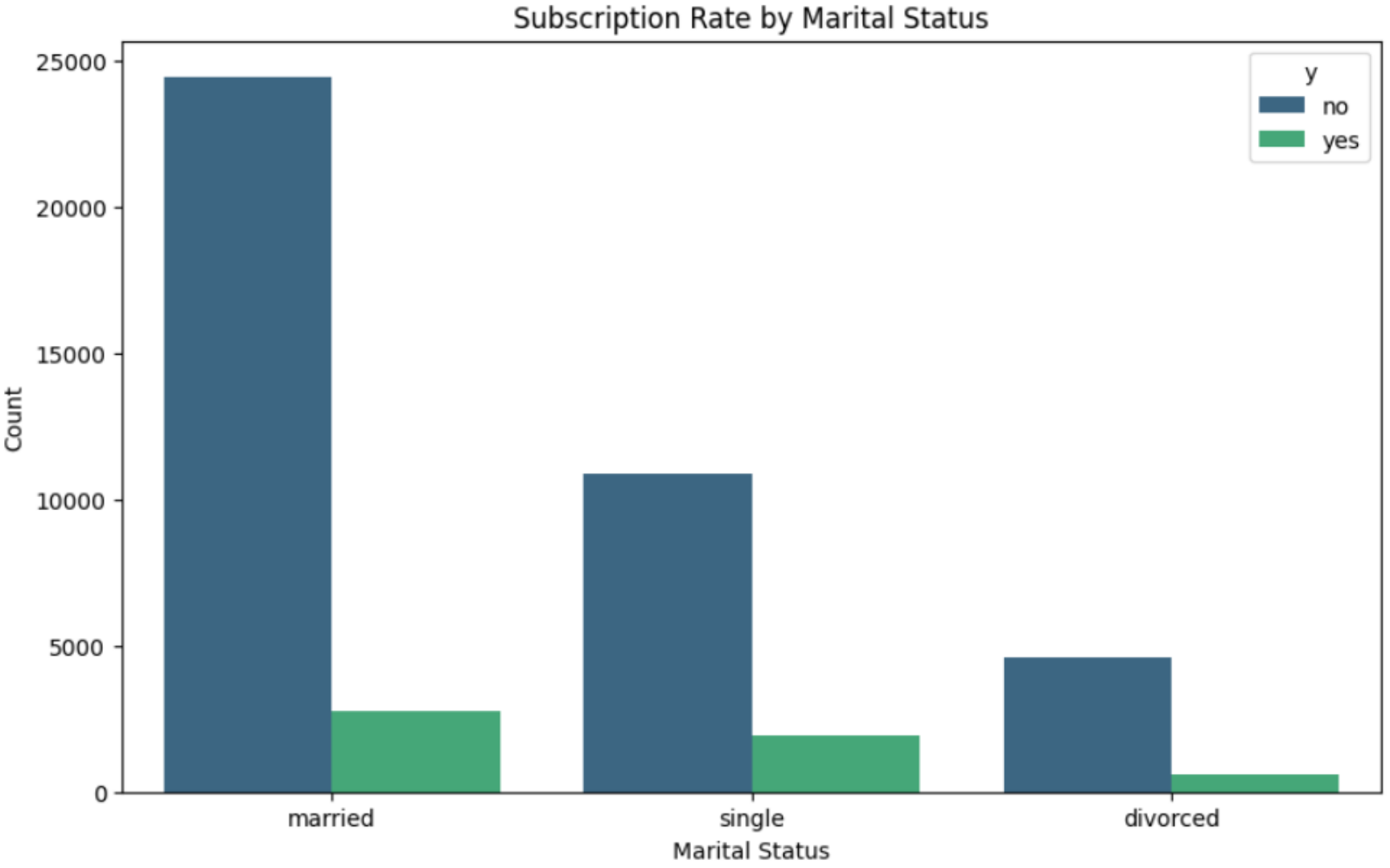
Subscription Rate by Job Type



Summary:

Certain job types are associated with higher subscription rates.

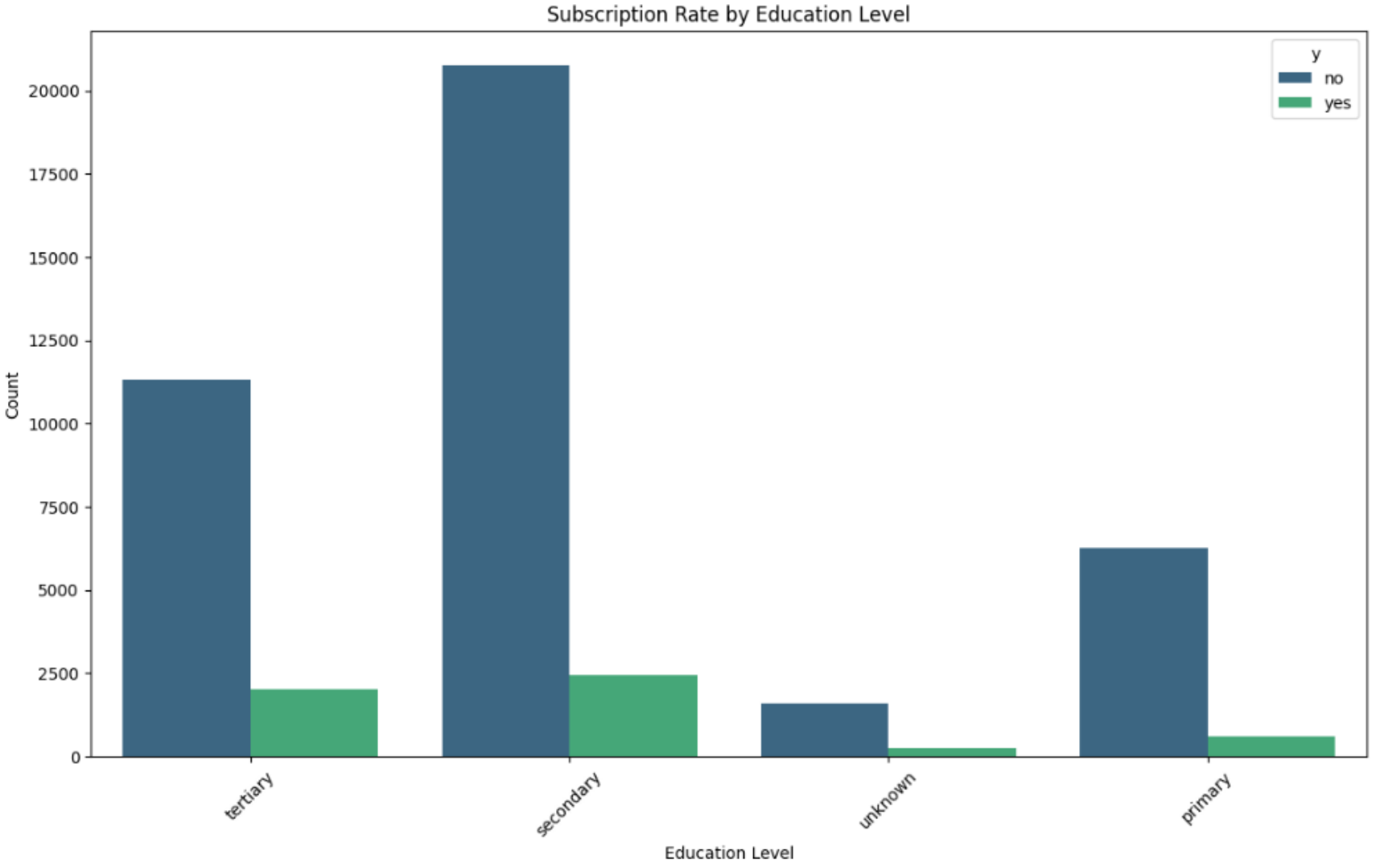
Subscription Rate by Marital Status



Summary:

Marital status influences the likelihood of subscribing, with some statuses being more inclined to subscribe than others.

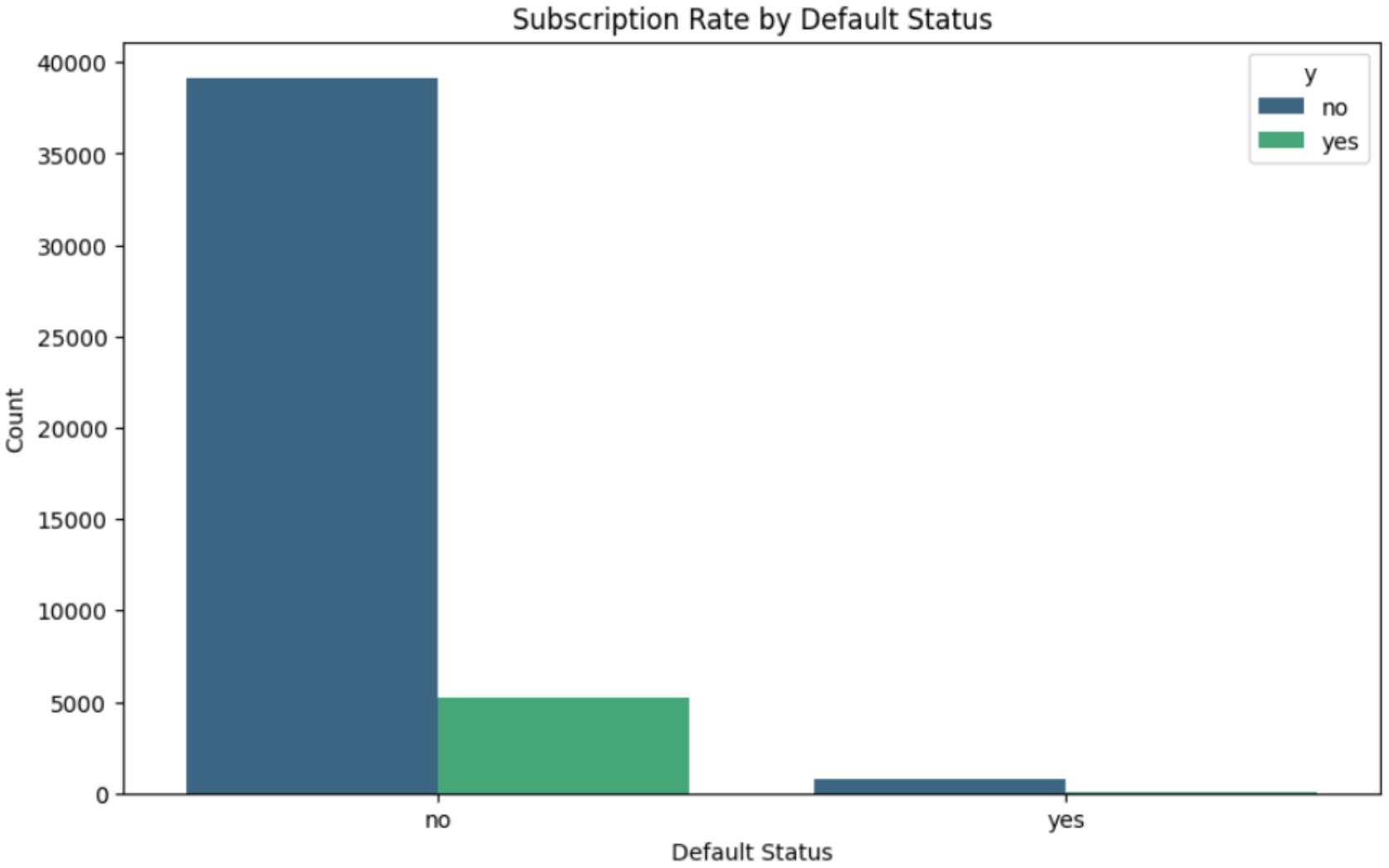
Subscription Rate by Educational Level



Summary:

Higher levels of education correlate with an increased likelihood of subscription.

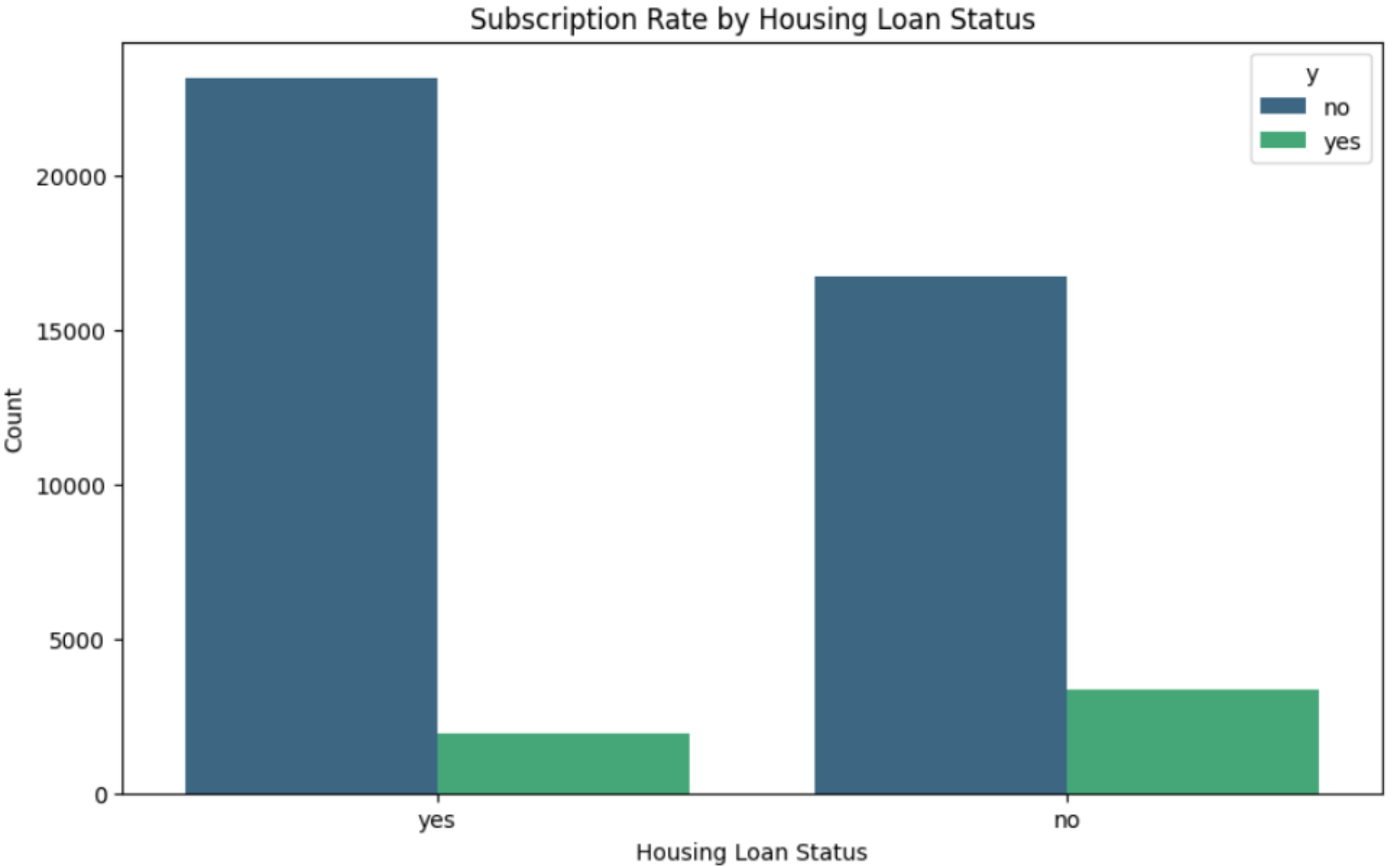
Subscription Rate by Default Status



Summary:

Customers with a history of default are less likely to subscribe.

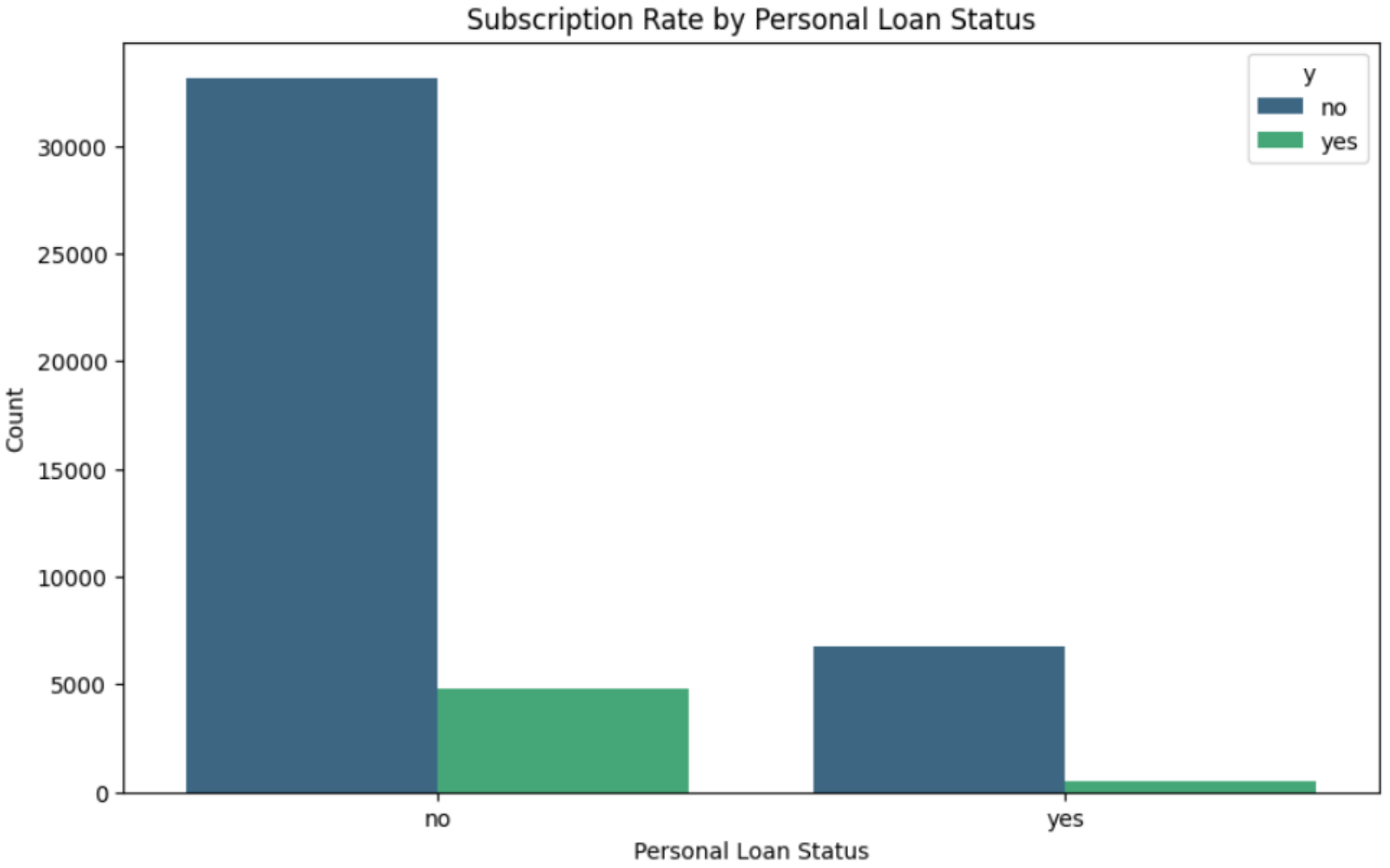
Subscription Rate by Housing Loan Status



Summary:

Customers with an existing housing loan are less likely to subscribe.

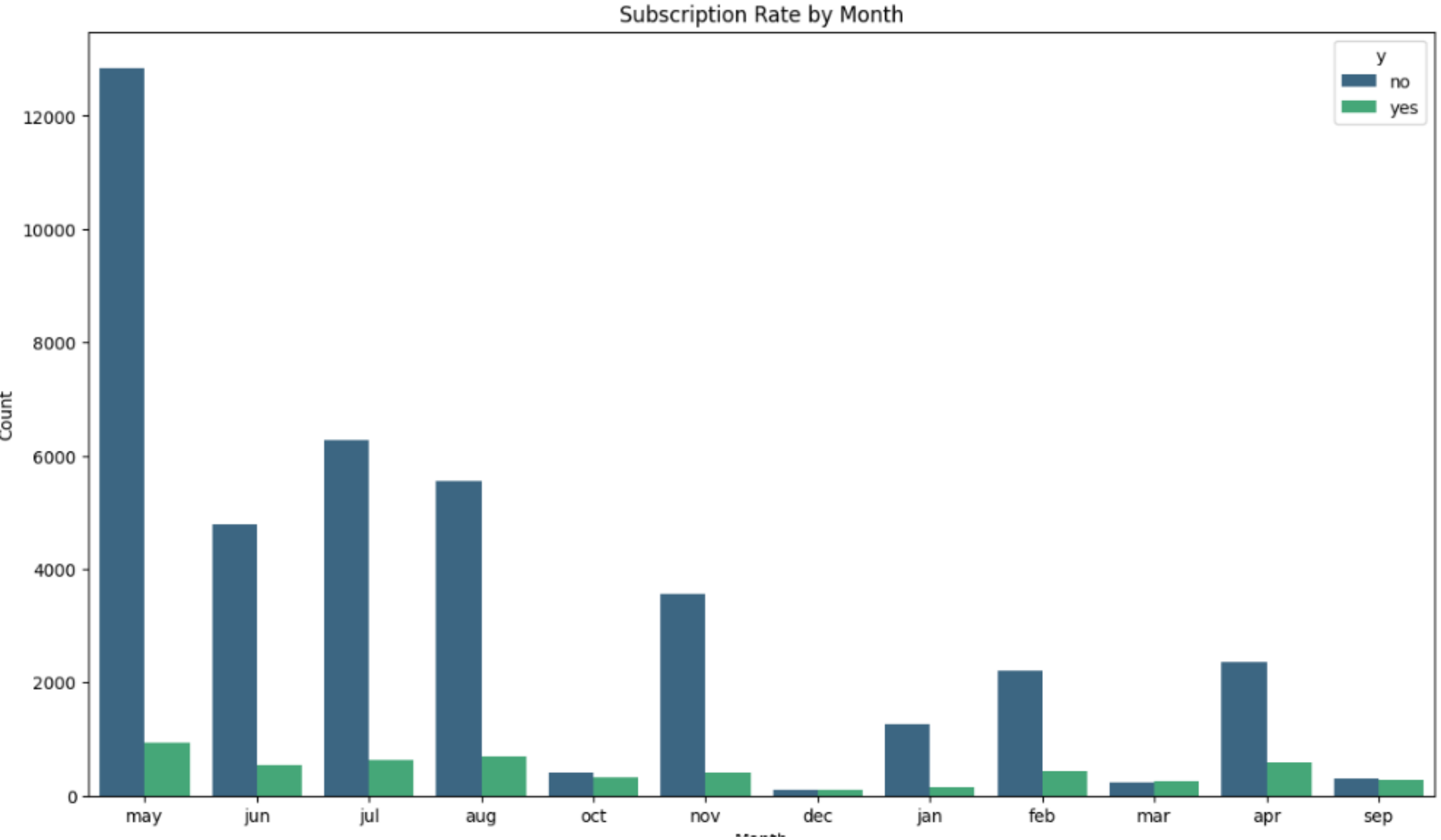
Subscription Rate by Personal Loan Status



Summary:

Holding a personal loan affects the likelihood of subscribing to new services.

Subscription Rate By Month



Summary:

Subscription rates vary across different months, indicating seasonal trends.

EDA Summary: from Hypothesis

1. Customer Demographics Matter:

Age, job type, education, and marital status significantly influence subscription rates. These demographic factors should be used to segment the customer base and tailor marketing strategies accordingly.

2. Financial Status Influences Decisions :

The presence of housing and personal loans, as well as default history, affect a customer's likelihood to subscribe. Marketing messages that address financial concerns or offer related products may improve conversion rates for these groups.

3. Communication Strategy is Key:

The method and timing of contact, including the day of the week and month, play a crucial role in subscription outcomes. Optimizing these factors based on past performance can enhance engagement and increase subscription rates.

4. Campaign Frequency and History Impact Results:

The number of contacts in a campaign and the outcomes of previous campaigns influence current subscription behaviour. It’s important to find the right balance in contact frequency and to build on the successes of past campaigns.

5. Economic Conditions Affect Behaviour:

Broader economic indicators like employment variation rates also play a role in subscription decisions. Adjusting marketing strategies to align with current economic conditions can make campaigns more effective.

6. Seasonality Should Be Considered:

Subscription rates vary by month, indicating seasonality in customer behaviour. This trend should inform the timing of campaigns, with strategic offers or promotions during low-performing months to smooth out performance across the year.

Summary:

The hypotheses reveal that both customer-specific factors (like demographics and financial status) and external factors (such as economic conditions and seasonality) significantly affect subscription behaviour. To maximize subscription rates, businesses should adopt a data-driven approach that segments customers effectively, tailor communication strategies, and adjusts campaigns based on the timing and broader economic context. This comprehensive strategy will likely lead to more successful outcomes and optimized marketing efforts.

Final Recommendation

Logistic Regression:

Reasoning: Logistic Regression is a strong candidate for binary classification problems like predicting subscription to a term deposit. It is particularly useful for interpreting the influence of different features on the target variable, aligning with the first hypothesis about demographic factors influencing subscription.

Hypotheses Fit:

Hypothesis 1: Customers between ages 30-40 are more likely to subscribe.

Hypothesis 8: Job type influences subscription rate.

Hypothesis 9: Marital status affects subscription rate.

Hypothesis 10: Education level's impact on subscription.

Random Forest Classifier:

Reasoning: Random Forest is well-suited for handling large datasets with multiple features and can model complex interactions between them. It also provides feature importance scores, which could validate hypotheses related to economic indicators or communication channels.

Hypotheses Fit:

Hypothesis 3: Impact of economic indicators.

Hypothesis 5: Previous campaign outcomes impact on subscription.

Hypothesis 12 & 13: Housing and personal loan status affect subscription.

Final Recommendation

Gradient Boosting Machines (GBM):

Reasoning: GBM is powerful for classification problems and can capture non-linear relationships. It can be particularly useful if the relationship between features like communication channels and subscription success is complex.

Hypotheses Fit:

Hypothesis 2: Success rate is higher when using cell phone communication.

Hypothesis 4: Timing of contact influences success rate.

Hypothesis 14: Month-wise subscription distribution.

Support Vector Machines (SVM):

Reasoning: SVM can be effective in high-dimensional spaces and is particularly useful when the dataset is not too large but has clear margins of separation between classes.

Hypotheses Fit:

Hypothesis 6: Distribution of the target variable 'y' (Subscription to term deposit).

Hypothesis 7: Age distribution of customers.

Neural Networks:

Reasoning: While overkill for simpler datasets, a neural network might be useful if the data is large and has non-linear dependencies that simpler models can't capture effectively. It could explore interactions between multiple features.

Hypotheses Fit:

Complex interactions not clearly delineated by individual hypotheses but potentially existing between multiple variables.

Thank You