Data Glacier

Your Deep Learning Partner

# Exploratory Data Analysis

## Bank Marketing (Campaign)

**Group Name:** Data Visionaries
**Name:** Abdukhakimov Asatilla
**Email:** tremendous54@gmail.com
**Country:** Uzbekistan
**Company:** Data Glacier
**Specialization:** Data Science
**Github:** https://github.com/asat94/Data-Glacier-Internship

2025.04.13

# Agenda

PROBLEM STATEMENT

DATA EXPLORATION

EDA

EDA SUMMARY

MODELLING

Data Glacier

Your Deep Learning Partner

# Problem Statement

- ABC Bank wants to sell its term deposit product to customers. Before launching the product, the bank aims to develop a model to understand whether a particular customer will buy their product or not, based on the customer's past interaction with the bank or other financial institutions.


# Data Exploration

Bank-full data

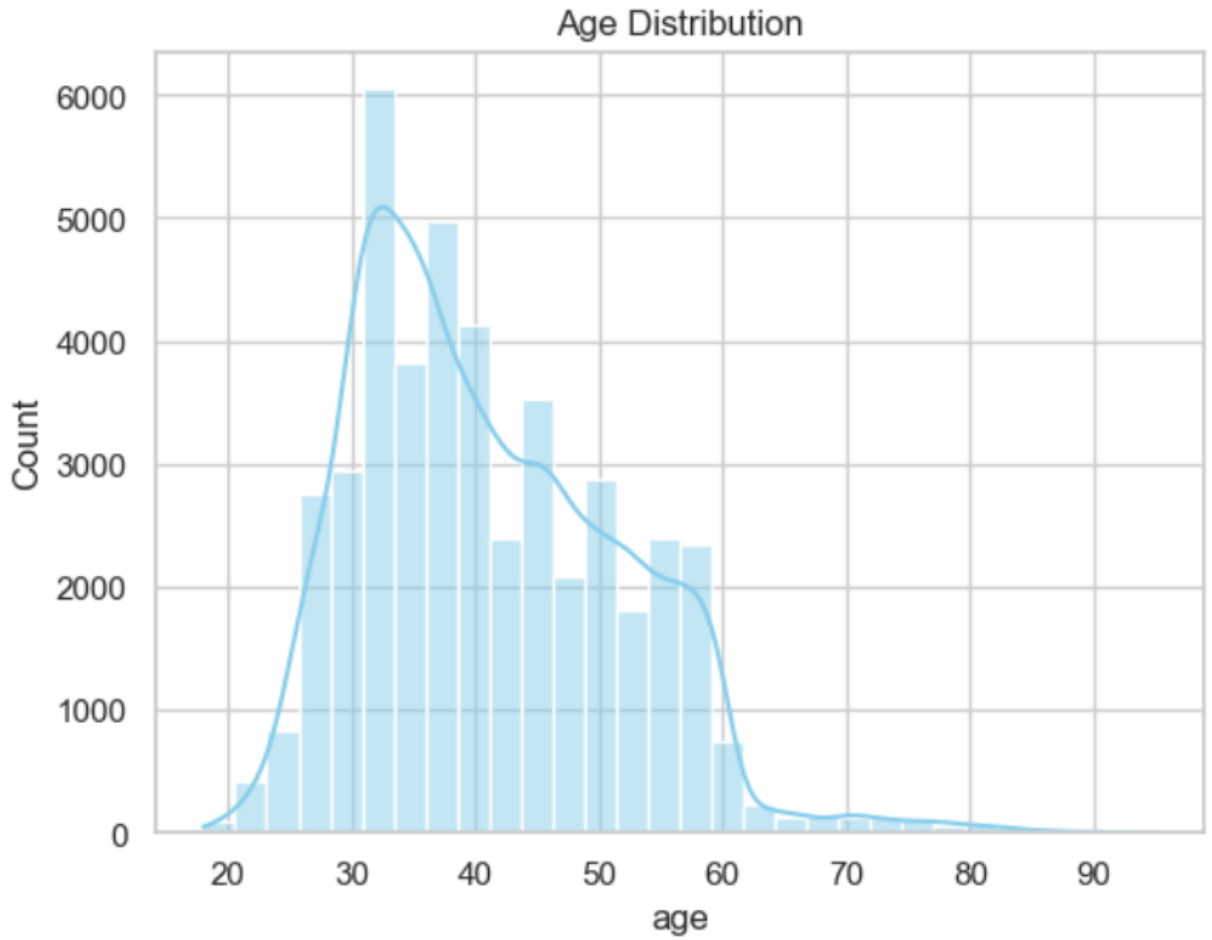
**Total Data**

- 45,211 rows
- 17 columns

```
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
```

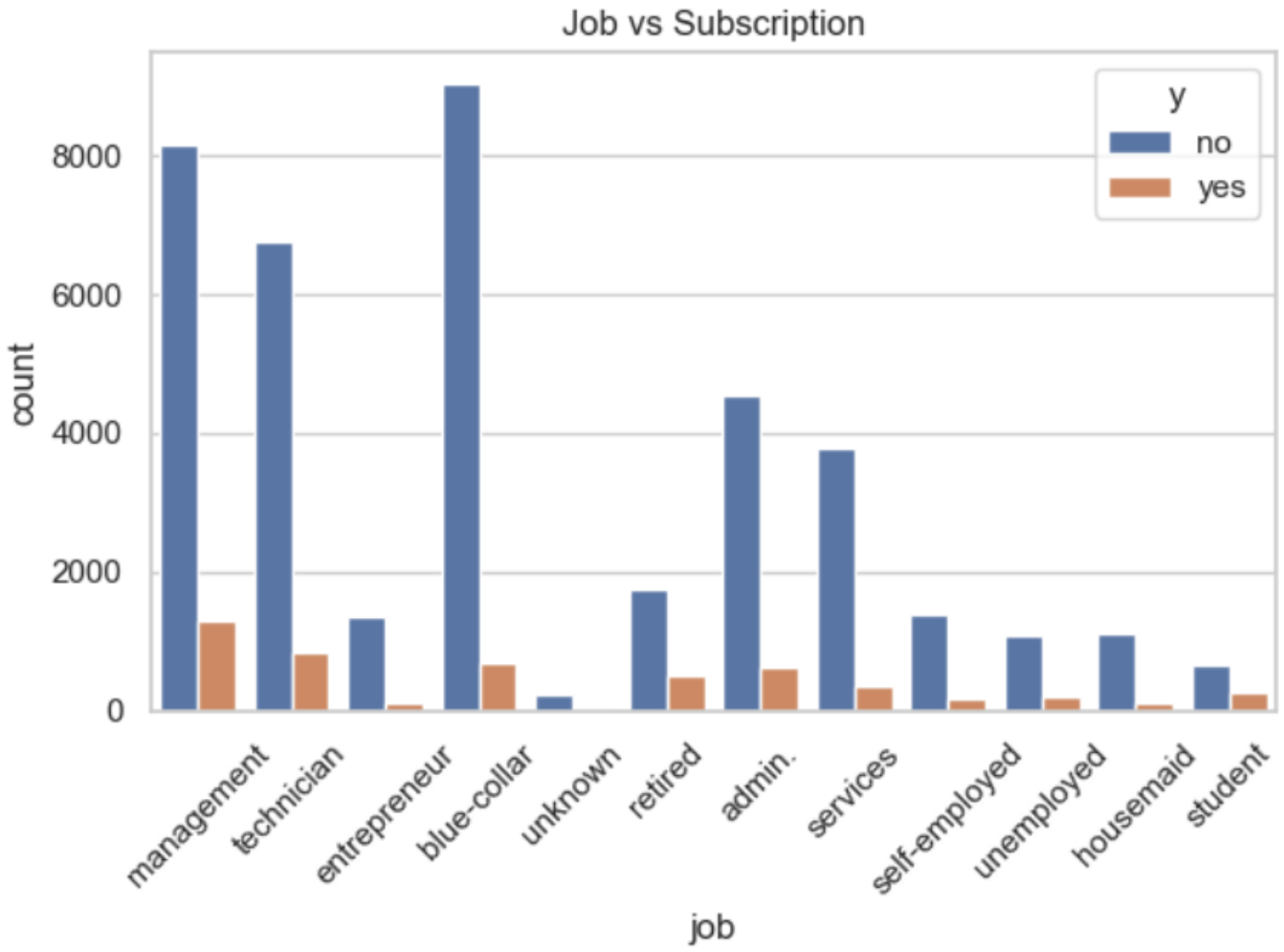
- Data Cleaned & Formatted
- Checked Missing Values & Outliers

# Age Distribution



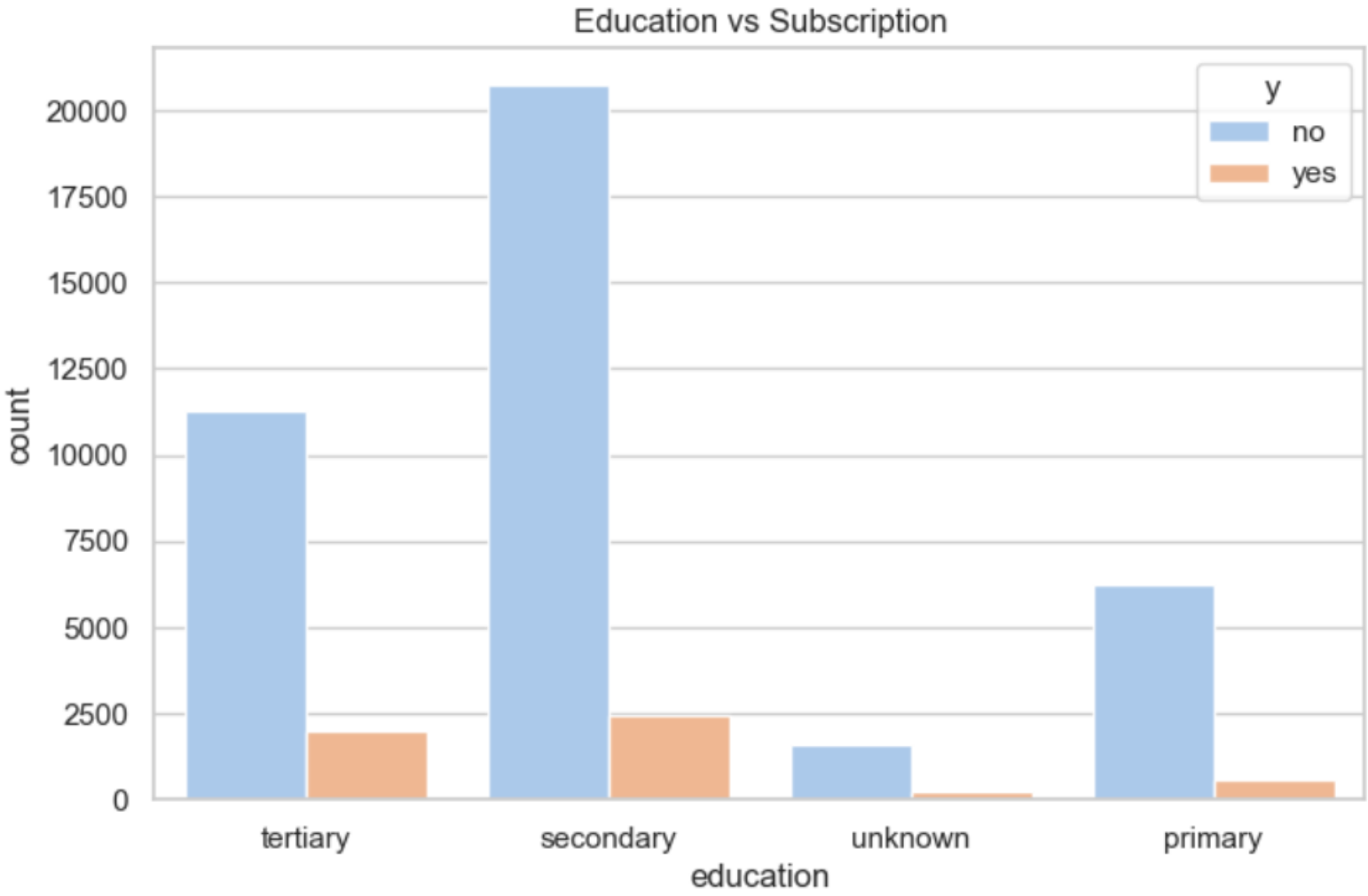

Customers are mainly between 25 to 60 years old, a key demographic for term deposits

# Job vs Subscription




Management and technician roles show higher subscription rates.

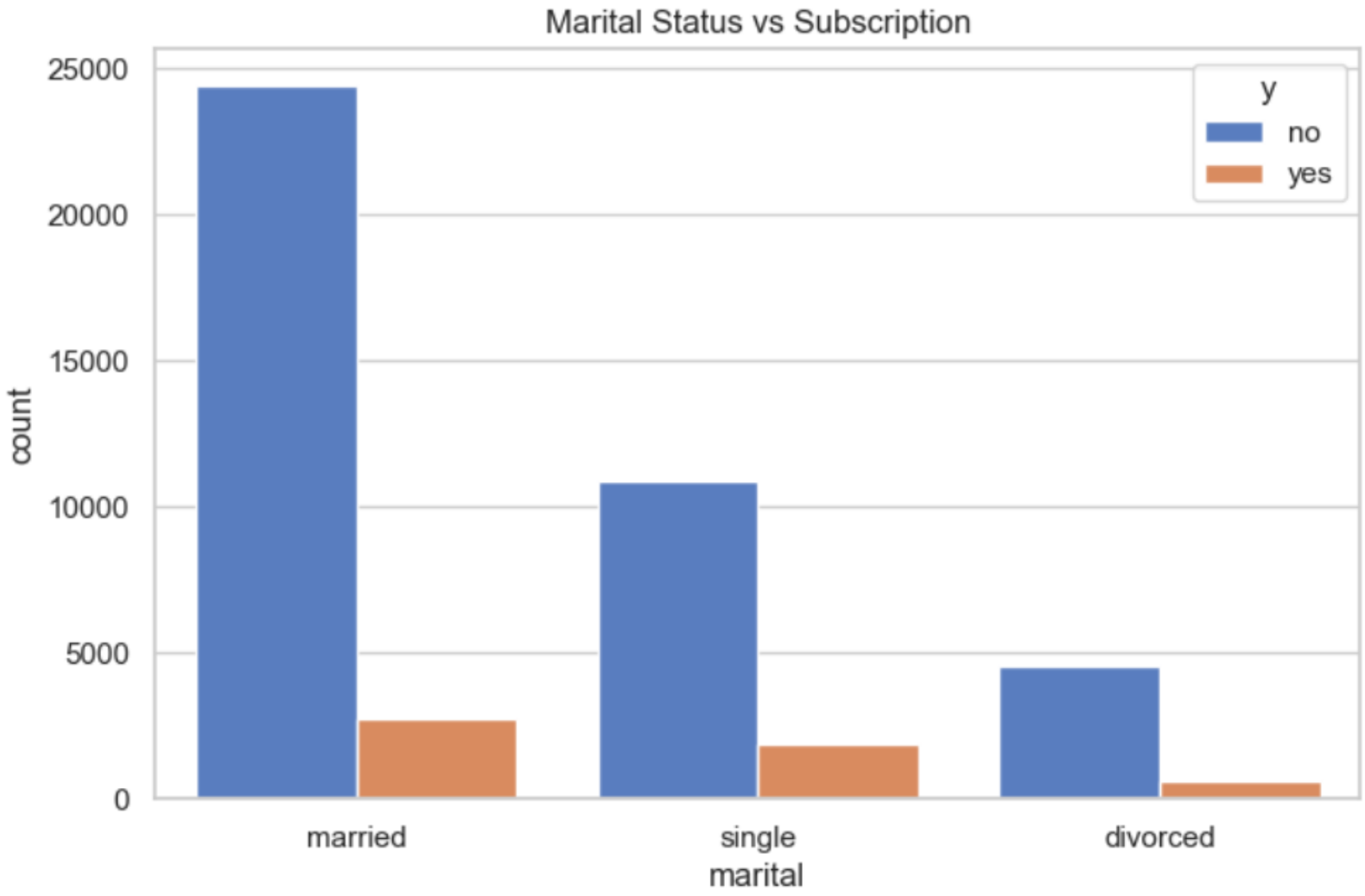
# Education vs Subscription



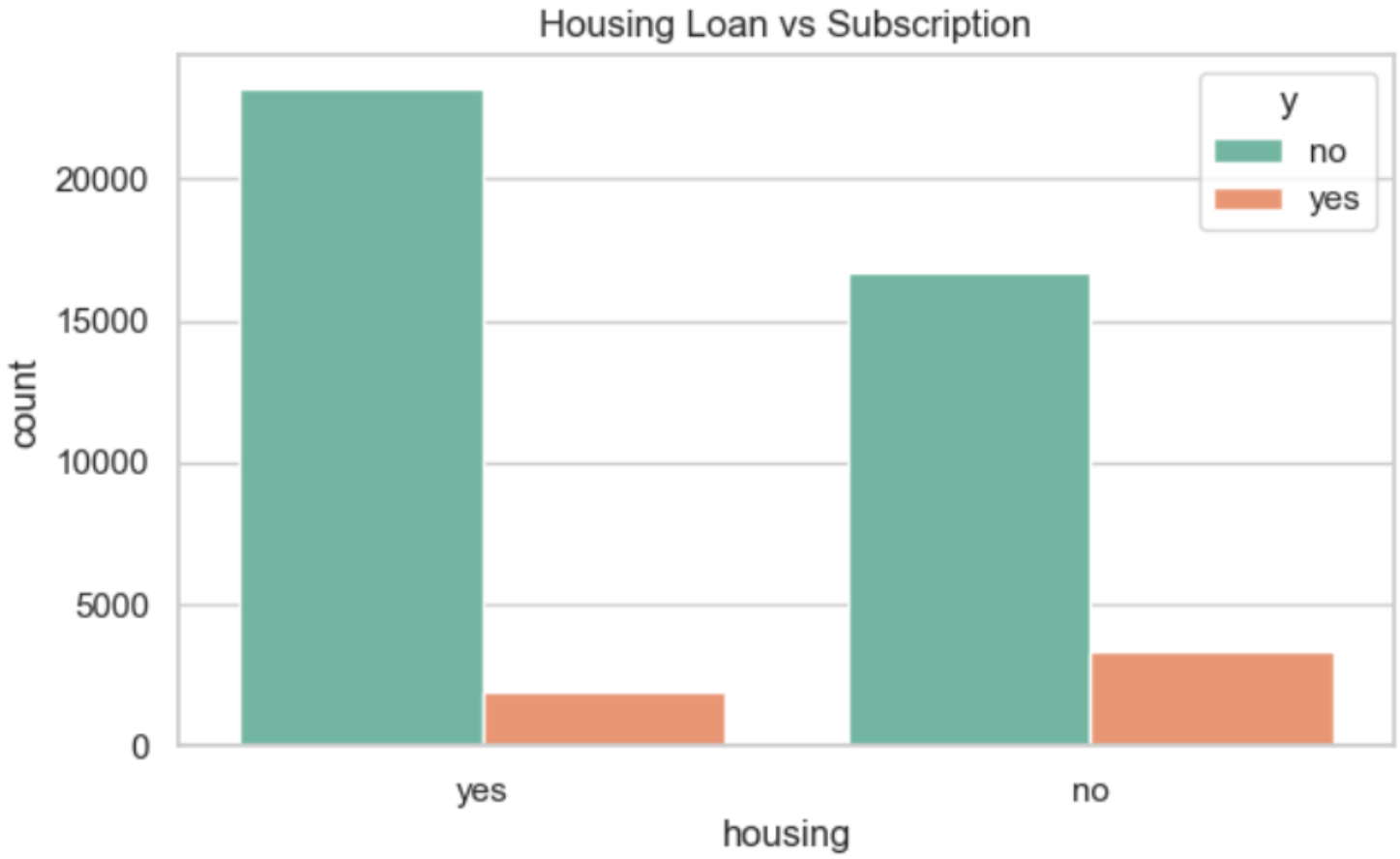

Tertiary educated customers are more likely to subscribe.

# Marital Status vs Subscription

Single customers tend to have higher subscription rates than married/divorced

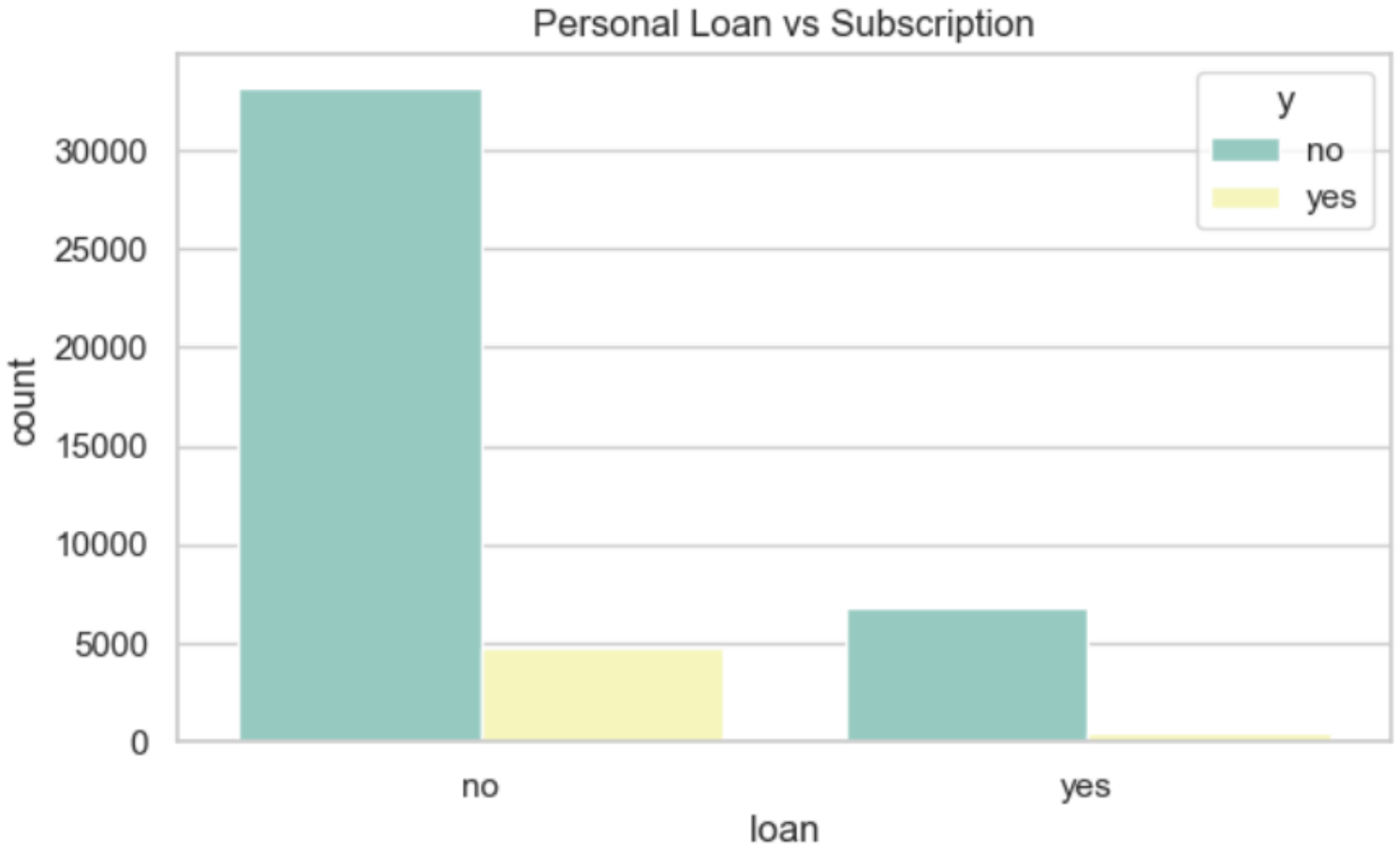


# Housing Loan Status



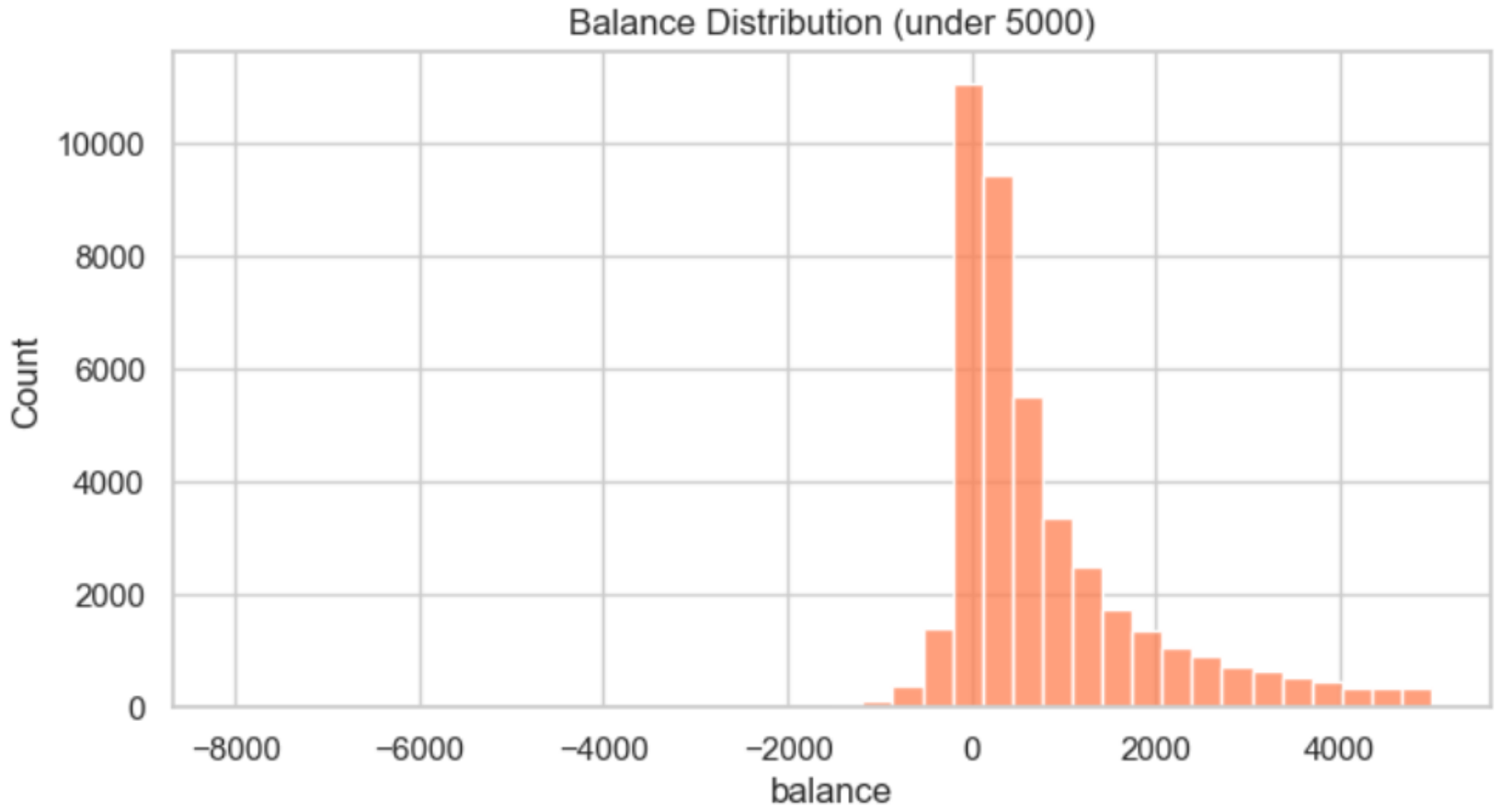

Customers without housing loans appear slightly more likely to subscribe

# Personal Loan Status



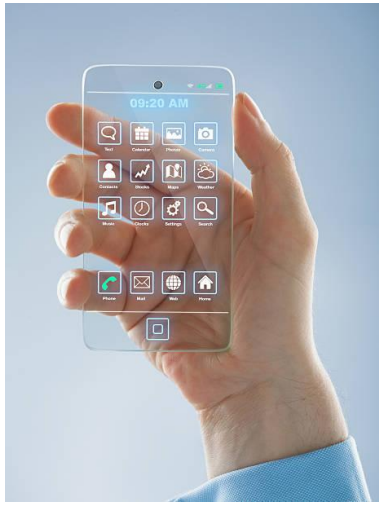

Those without personal loans show better subscription interest

# Balance Distribution (Zoomed In)

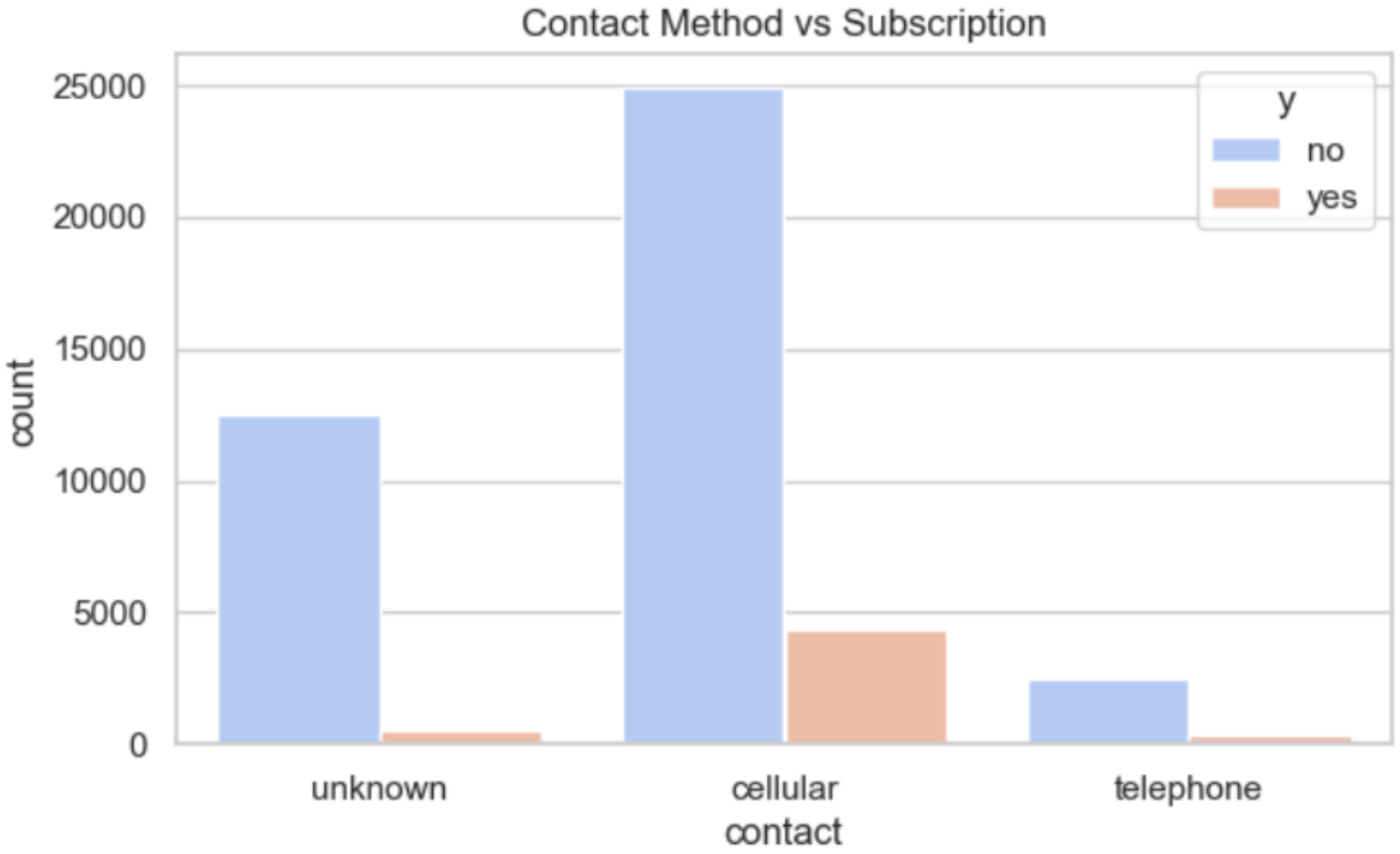



Most customers have balances under 5000, indicating a middle-income audience

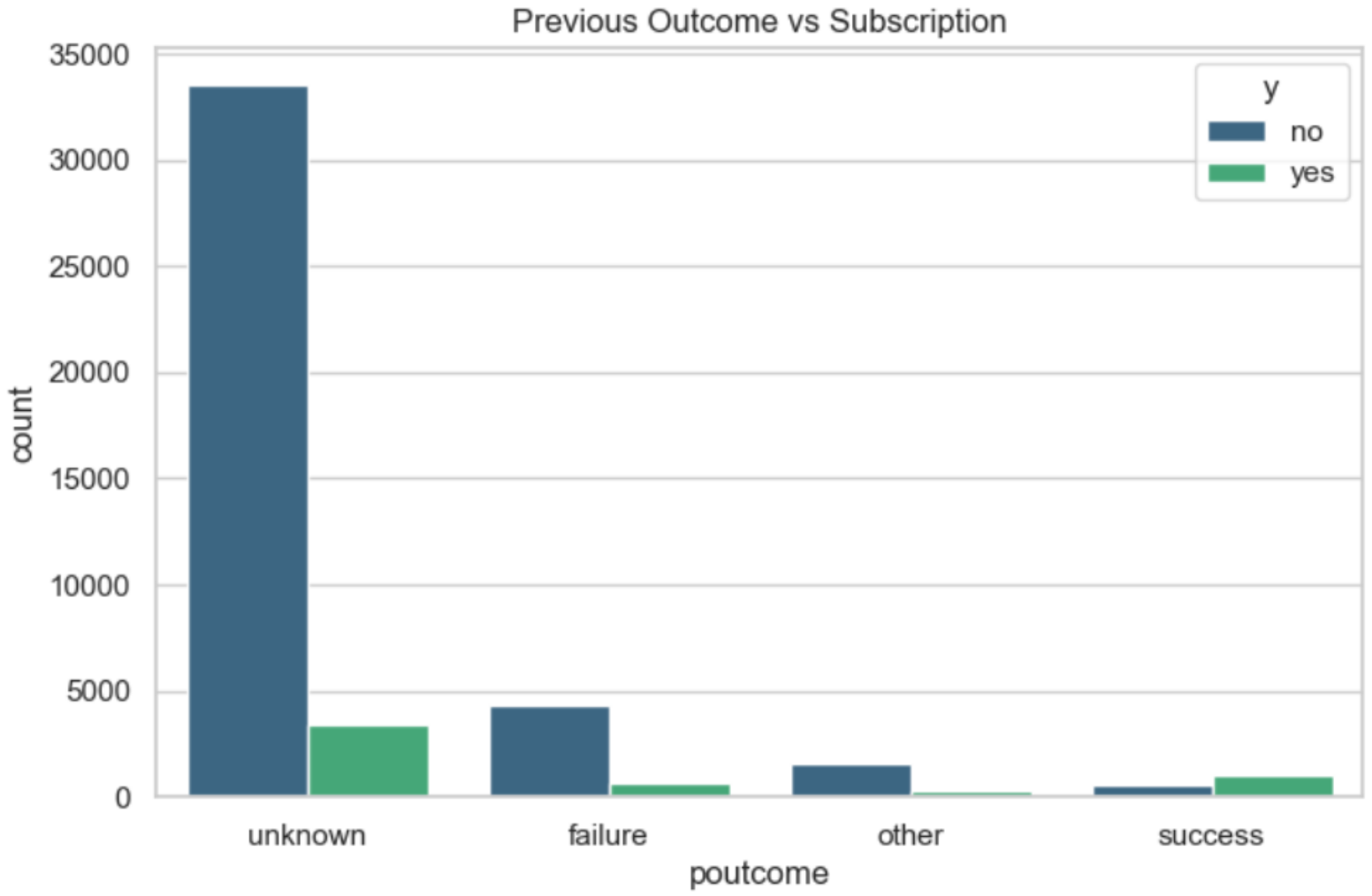
# Contact Method vs Subscription




Customers contacted via cellular show better subscription rates

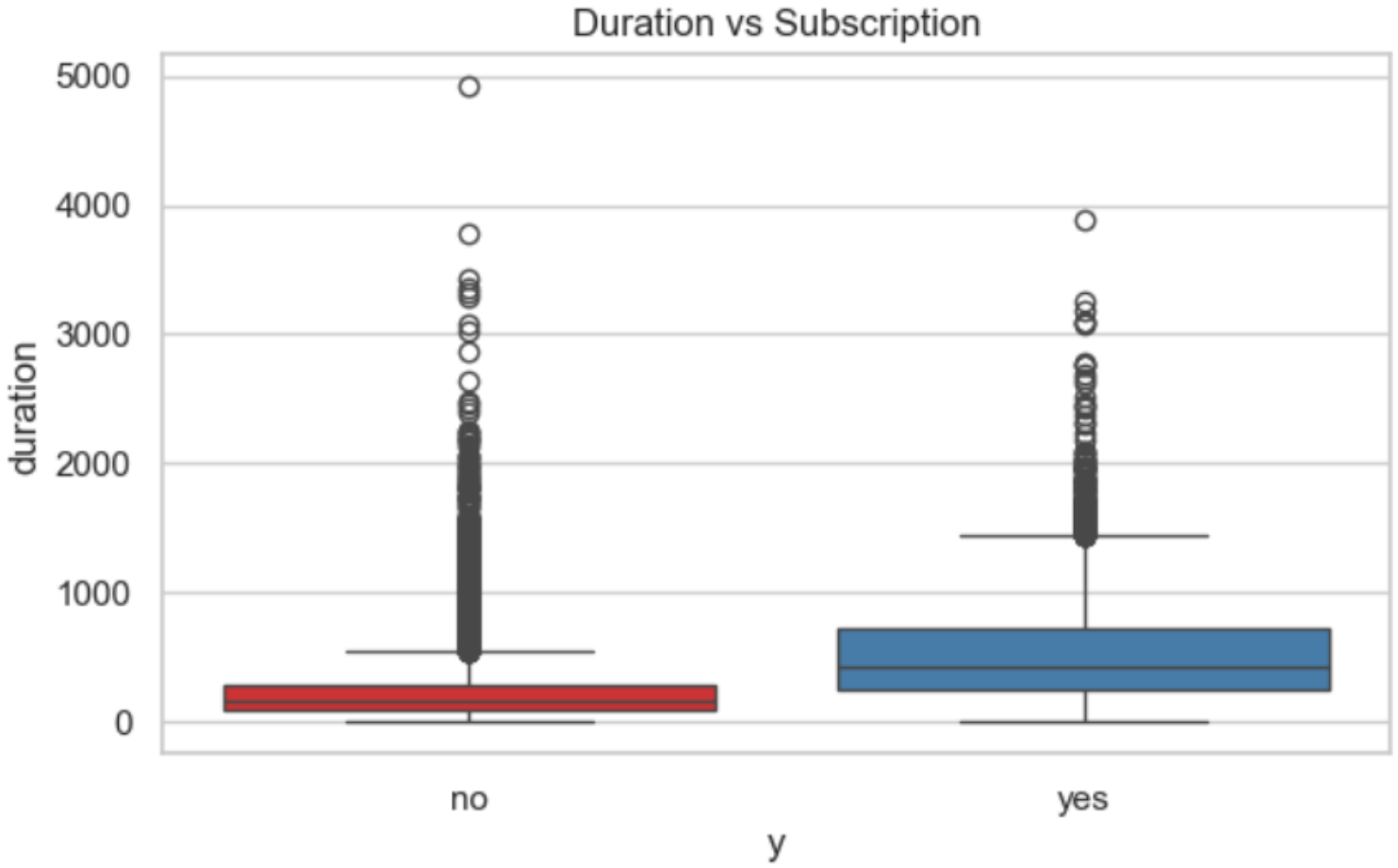


# Previous outcome vs Subscription



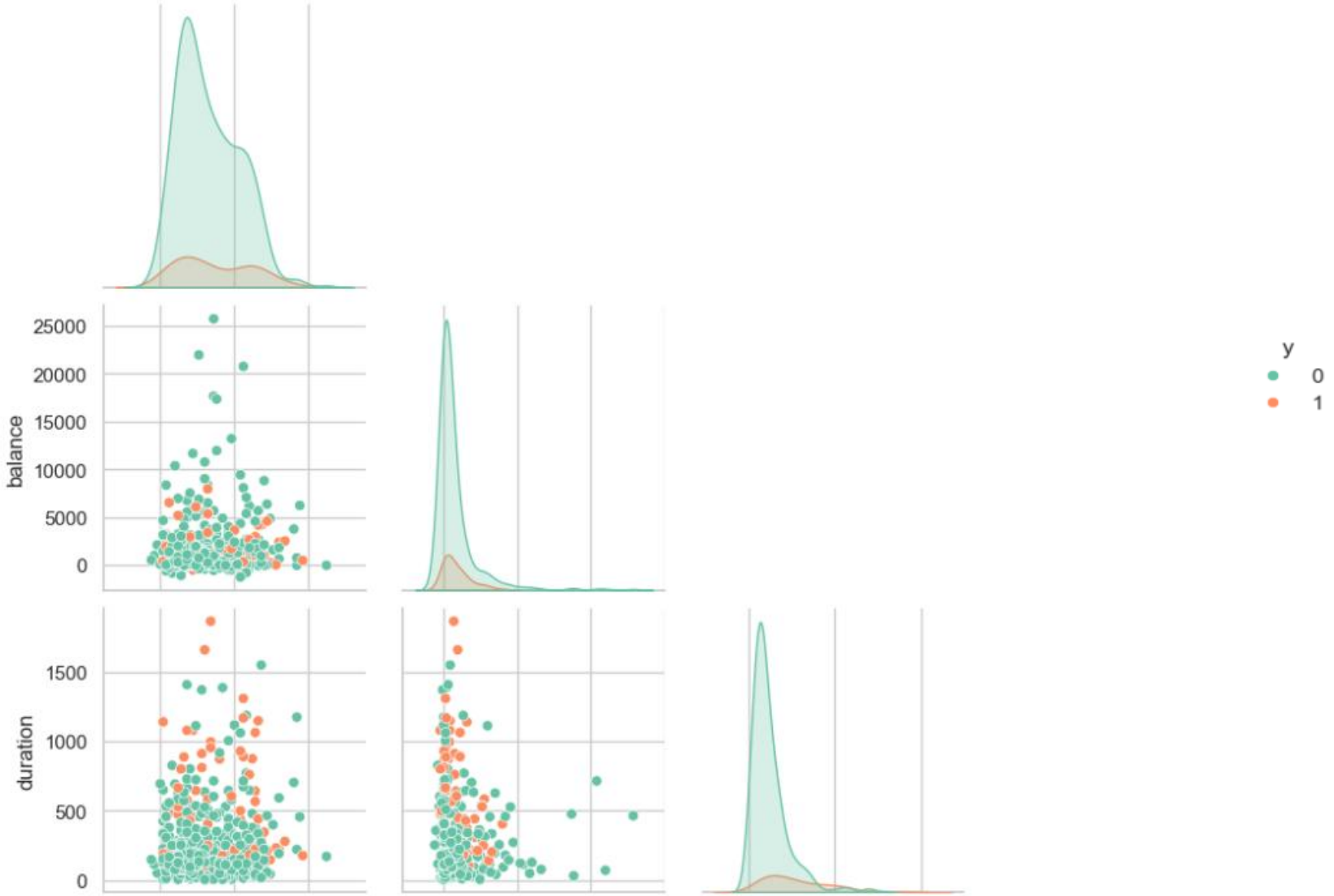

Successful outcomes in prior campaigns drastically improve subscription chances

# Duration vs Subscription

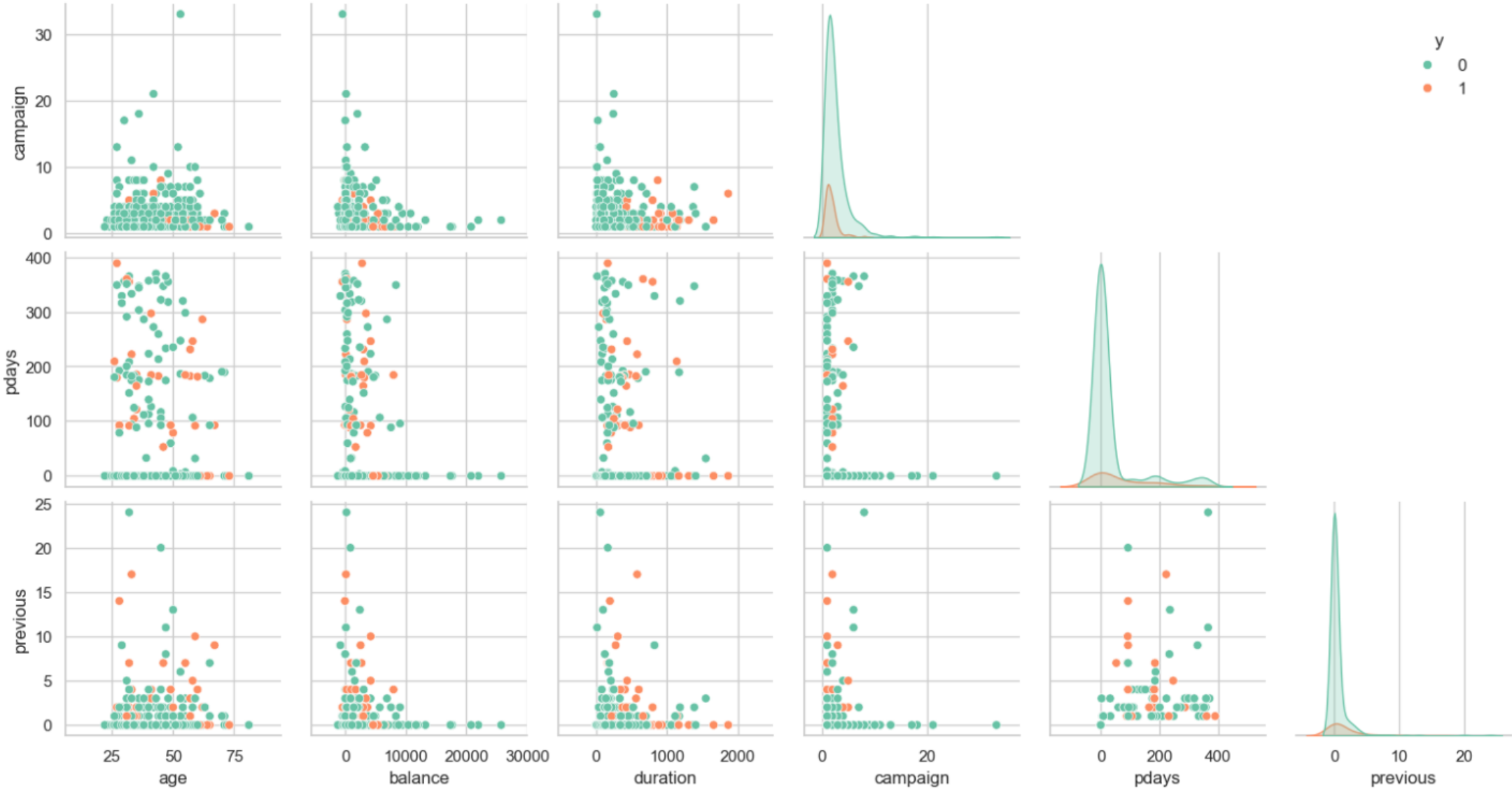



Longer call durations often lead to a 'yes' decision — indicating interest builds over time.

# Pair Plot (Scatter Matrix)



# Pair Plot (Scatter Matrix)



# EDA Summary

- Duration forms distinct clusters for subscribed (yes) vs not (no) customers.
- Subscribed customers tend to lie in higher ranges of duration and sometimes higher balances.
- No strong linear patterns between other variables and y.

# Modelling

=== LR with duration ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.98 | 0.94 | 7952 |
| 1 | 0.64 | 0.30 | 0.41 | 1091 |
| accuracy | | | 0.90 | 9043 |
| macro avg | 0.78 | 0.64 | 0.68 | 9043 |
| weighted avg | 0.88 | 0.90 | 0.88 | 9043 |

=== RF with duration ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.97 | 0.94 | 7952 |
| 1 | 0.63 | 0.41 | 0.49 | 1091 |
| accuracy | | | 0.90 | 9043 |
| macro avg | 0.78 | 0.69 | 0.72 | 9043 |
| weighted avg | 0.89 | 0.90 | 0.89 | 9043 |

=== LR w/o duration ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.99 | 0.94 | 7952 |
| 1 | 0.70 | 0.16 | 0.26 | 1091 |
| accuracy | | | 0.89 | 9043 |
| macro avg | 0.80 | 0.57 | 0.60 | 9043 |
| weighted avg | 0.87 | 0.89 | 0.86 | 9043 |

=== RF w/o duration ===

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.99 | 0.94 | 7952 |
| 1 | 0.65 | 0.20 | 0.31 | 1091 |
| accuracy | | | 0.89 | 9043 |
| macro avg | 0.78 | 0.59 | 0.63 | 9043 |
| weighted avg | 0.87 | 0.89 | 0.86 | 9043 |

- Including the **duration** variable significantly improves performance across all models.
- **Random Forest with duration** achieves the best overall performance

# THANK YOU

Data Glacier

Your Deep Learning Partner