# VAULT VISION

# PREDICTIVE BANK ANALYSIS

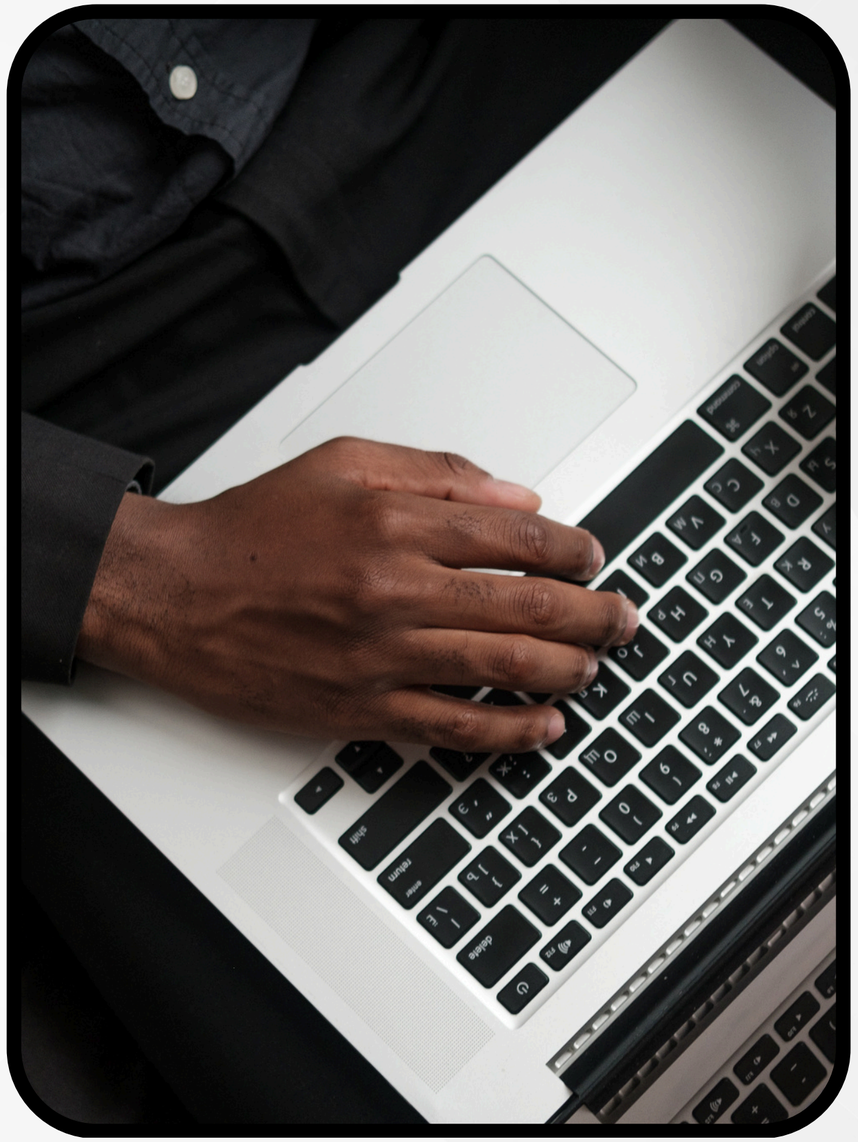
**By :**
Madhurima Dutta,
Unnati Mishra,
Samarth D S,
Sarvesh Chaudhari

# INDEX



# PROBLEM STATEMENT

To analyze Banco de Portugal data using data models such as Random Forest, Decision Tree, and Logistic Regression to identify patterns, manage risks, optimize lending decisions, and predict term deposit subscription likelihood.



# PROPOSED METHODOLOGY

1

Step-1 Data Cleaning & Feature Engineering

2

Step-2 Data Visualization
Visualizing data using different data plots

3

Step-3 Class Imbalance Problem

4

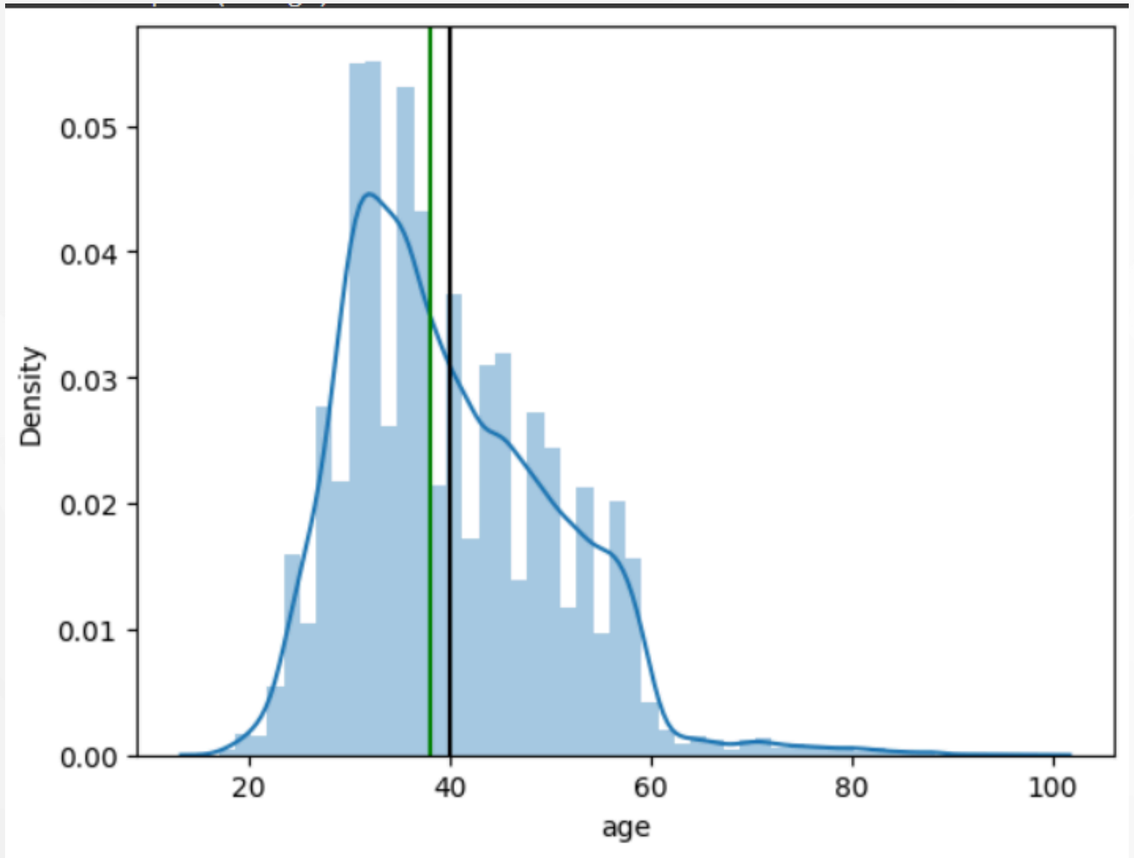
Step-4 Model Building
Random Forest, Decision Tree, Logistic Regression

# APPLIED SOLUTIONS

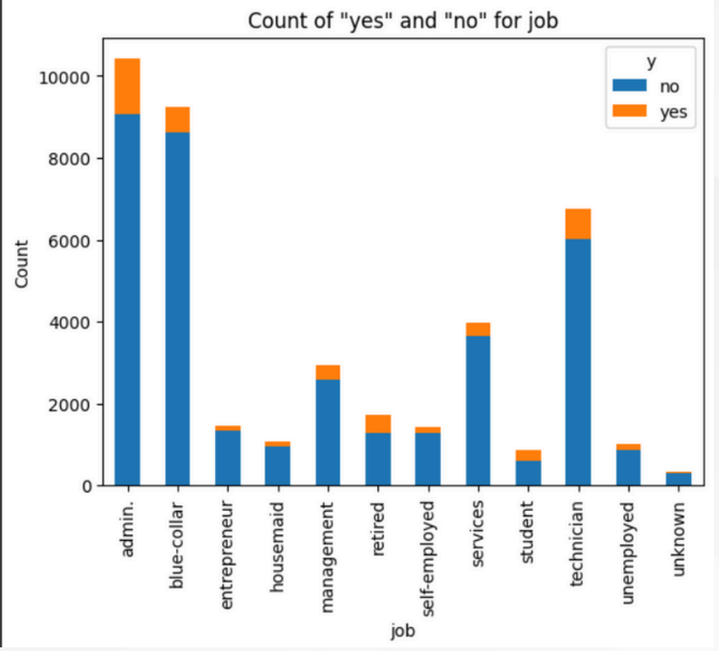
## 1. DATA CLEANING AND FEATURE ENGINEERING

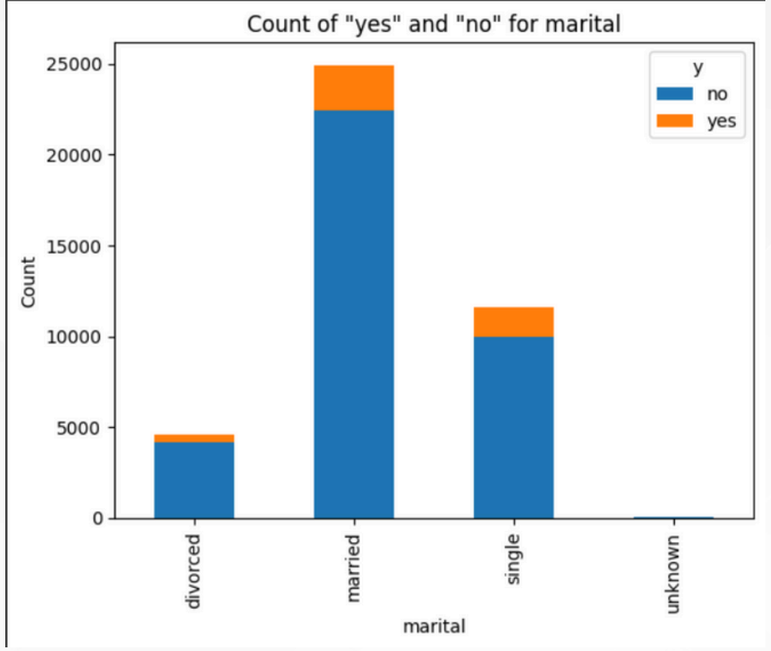
- **Missing Values Check**: Identified and ensured no missing values were present in the dataset.
- **Exploring Categorical Data**: Analyzed categorical columns (e.g., job, education, months, days etc) to understand their distribution.
- **One-Hot Encoding**: Transformed categorical variables into numerical form using one-hot encoding.
- **Correlation Analysis**: Identified highly correlated features to avoid redundancy.
- **Interaction with Target Variable**: Analyzed the relationship between categorical features and the target through visualizations.

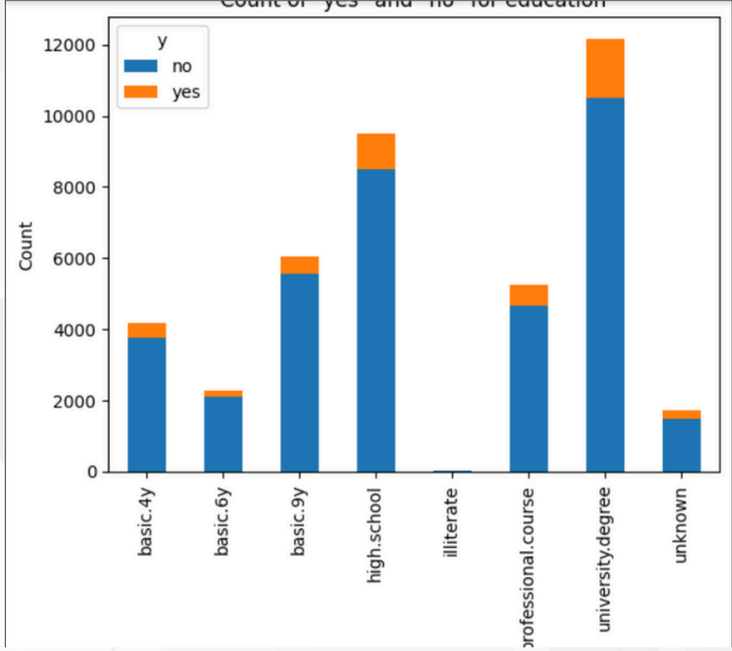
# 2. DATA VISUALIZATION

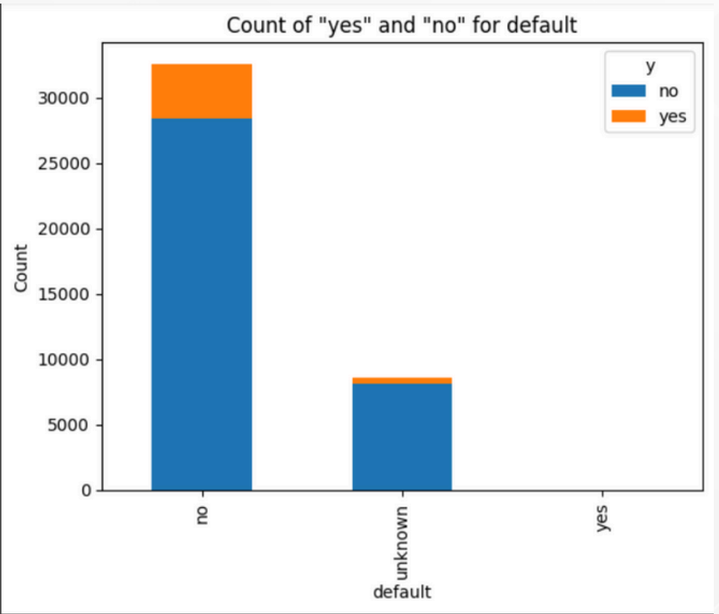


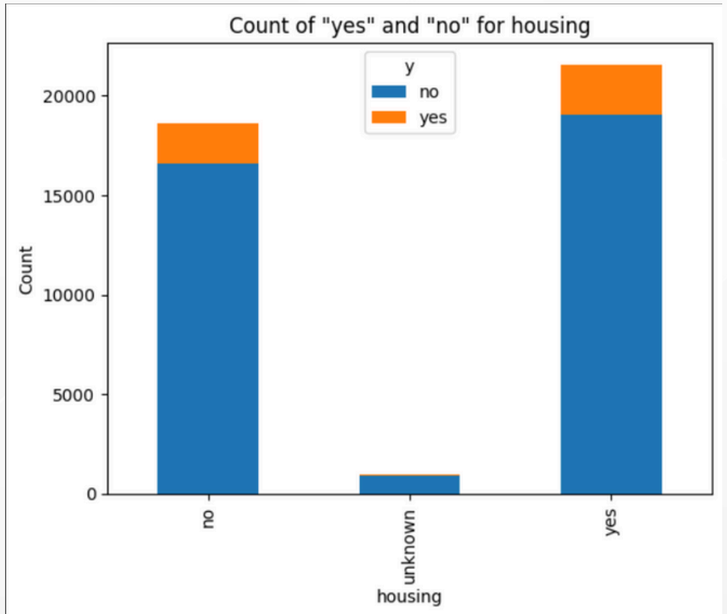

The bank calls a wide age range of clients in its telemarketing efforts, spanning from **18 to 95 years** old. Nonetheless, the bulk of callers are in their **30s and 40s** . The age distribution of the customers has a tiny standard deviation and is generally normal.

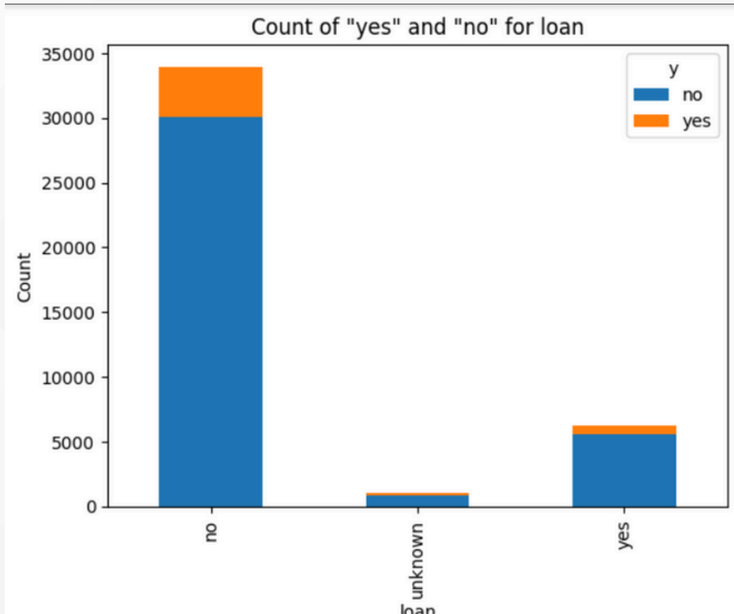

Count of "yes" and "no" for job
y
no
yes
Count
job
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown

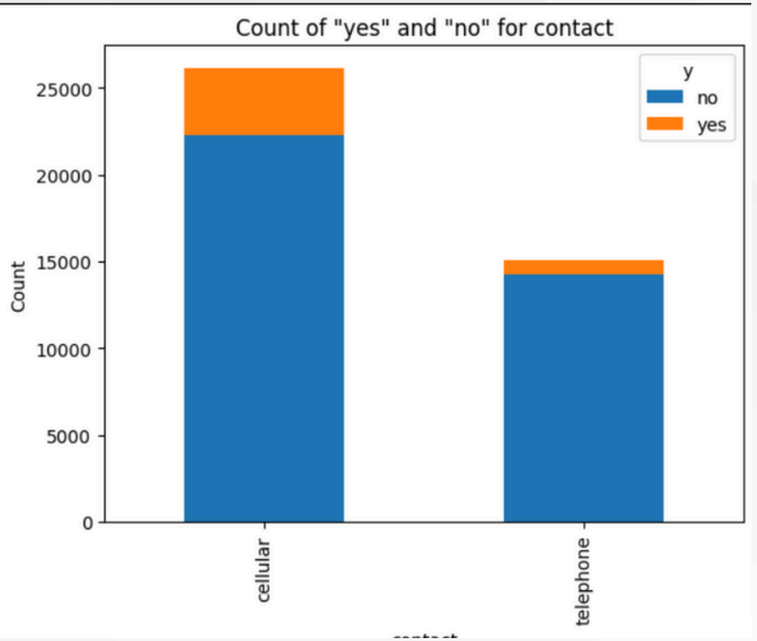

Count of "yes" and "no" for marital
y
no
yes
Count
marital
divorced
married
single
unknown

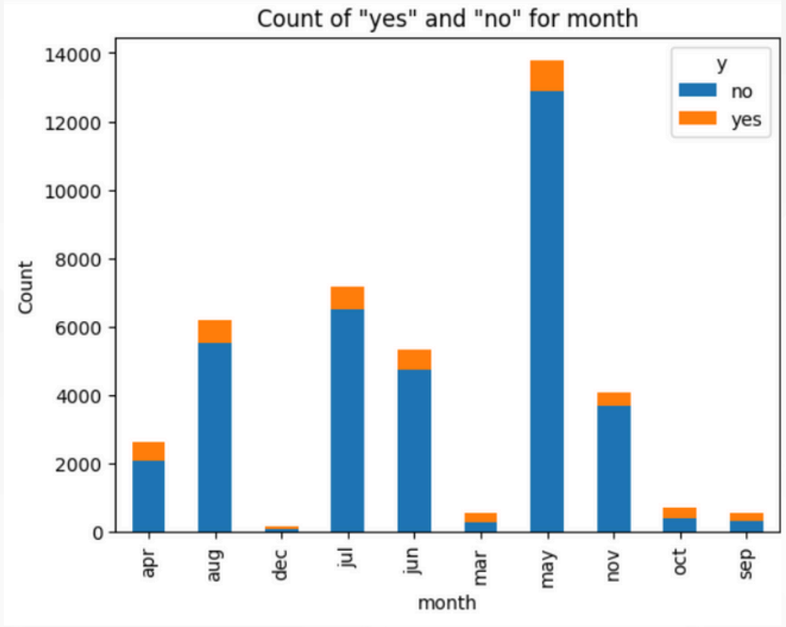

y
no
yes
Count
basic.4y
basic.6y
basic.9y
high.school
illiterate
university.degree
unknown

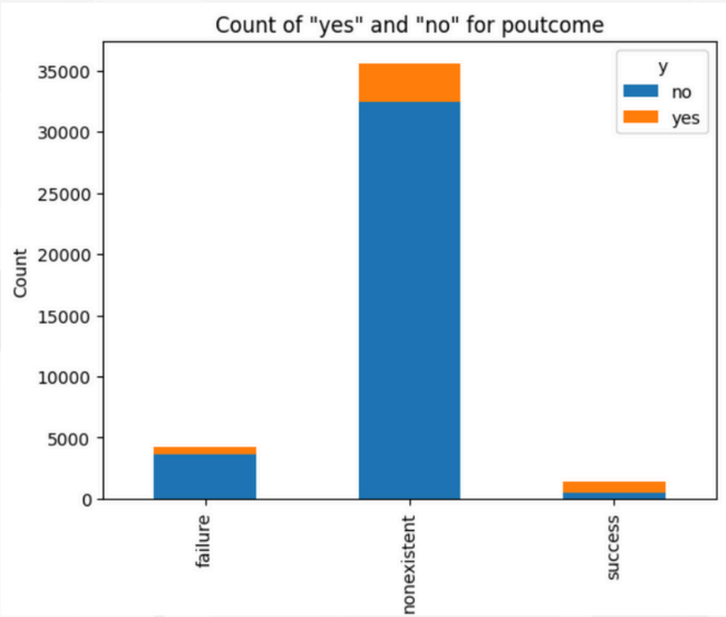

Count of "yes" and "no" for default
y
no
yes
Count
default
no
unknown
yes

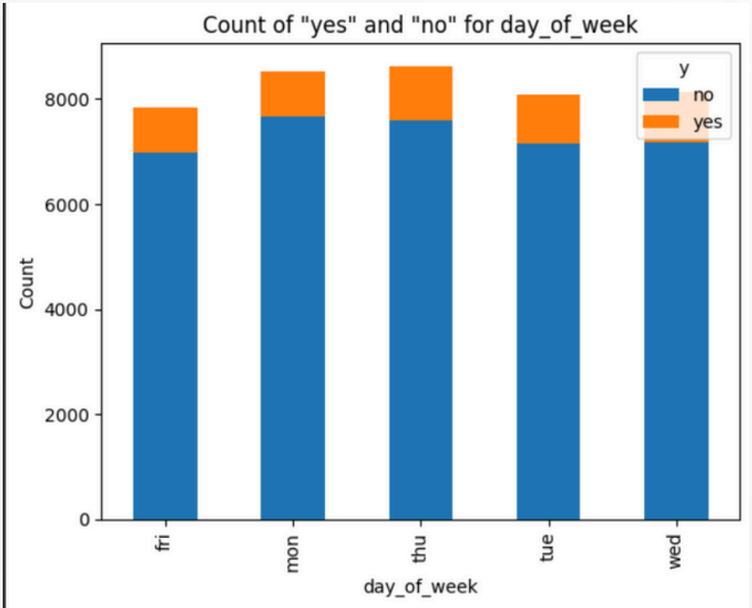

Count of "yes" and "no" for housing
y
no
yes
Count
housing
no
unknown
yes


Count of "yes" and "no" for loan
y
no
yes
Count
no
unknown
yes

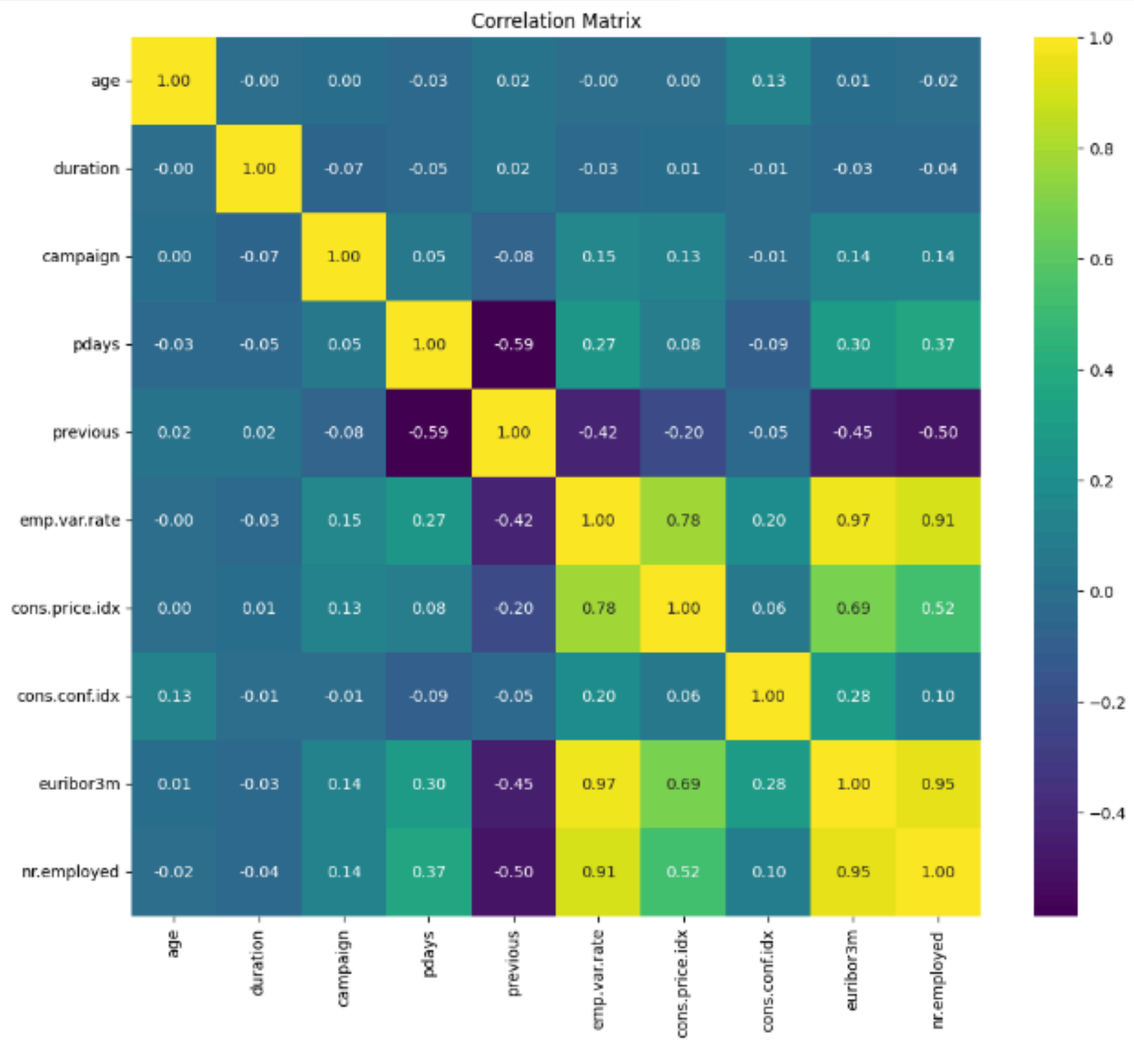


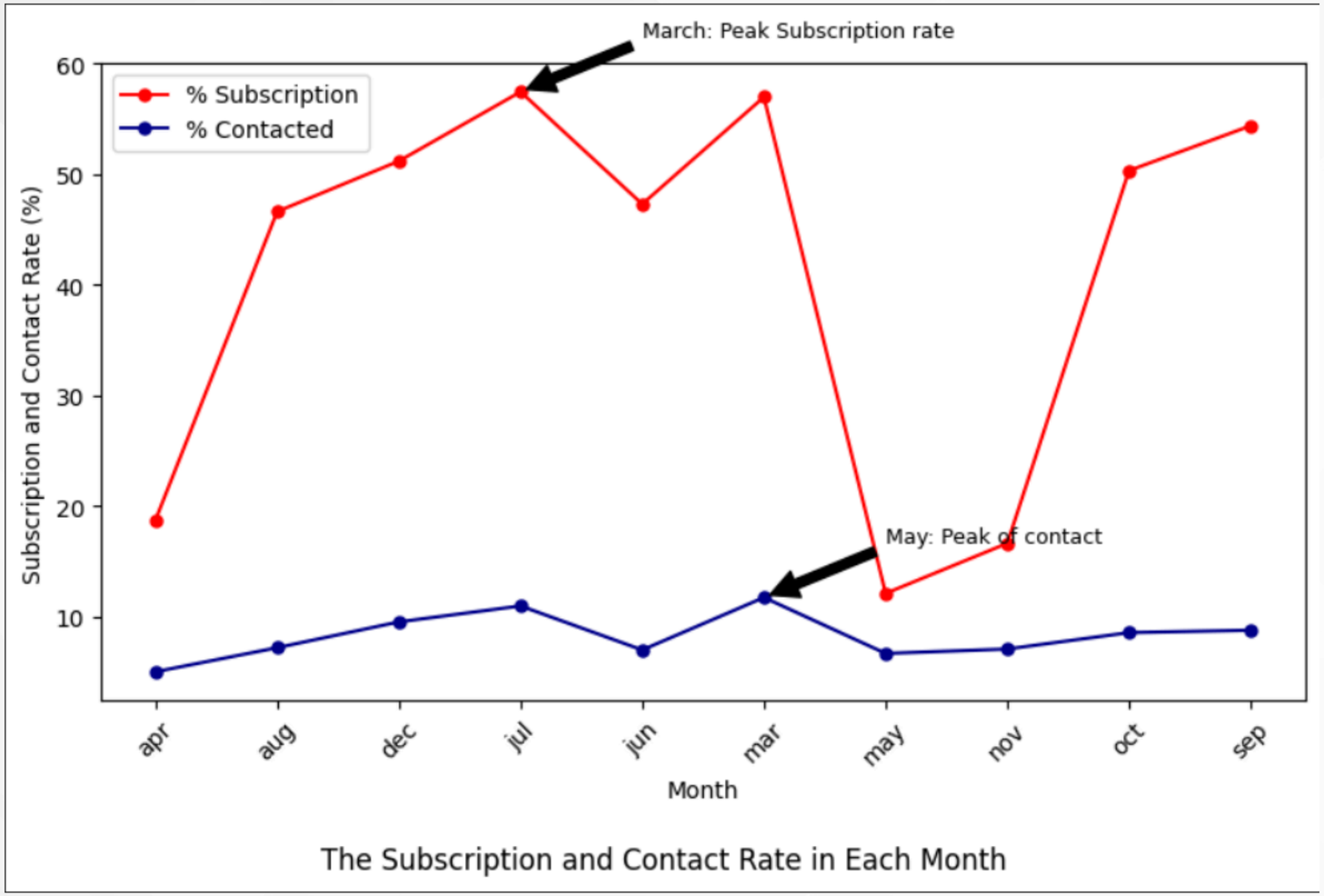




1. Plots the counts of '**yes**' and '**no**' (subscription responses) across various categorical features like job, marital status, etc.
2. Useful for seeing which groups had higher or lower subscription rates.



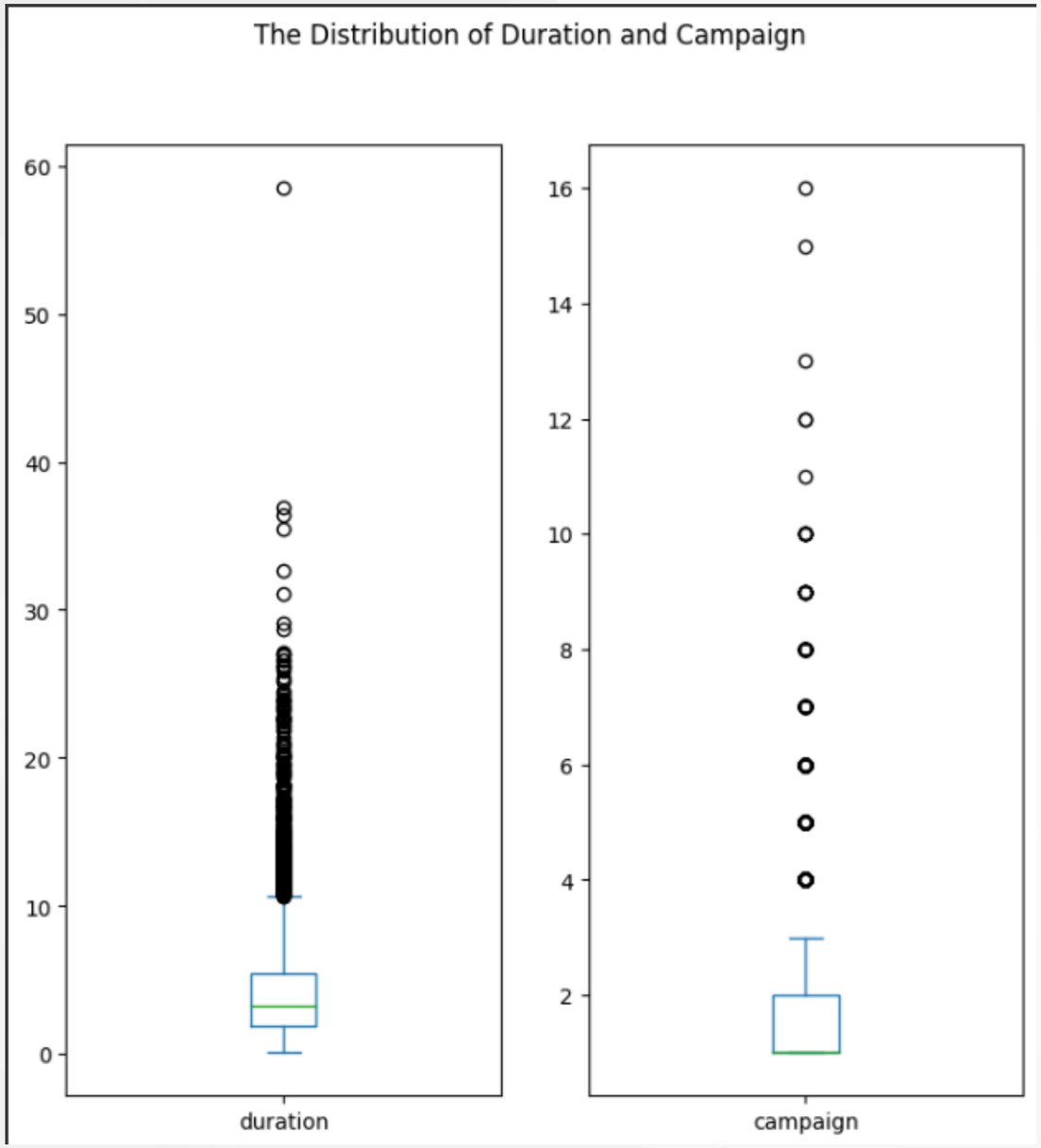

The heatmap shows the strength and direction of correlations between numerical features, ranging from **-1 (inverse) to 1 (positive)**.

The color gradient highlights strong relationships, helping identify potential multicollinearity between variables.



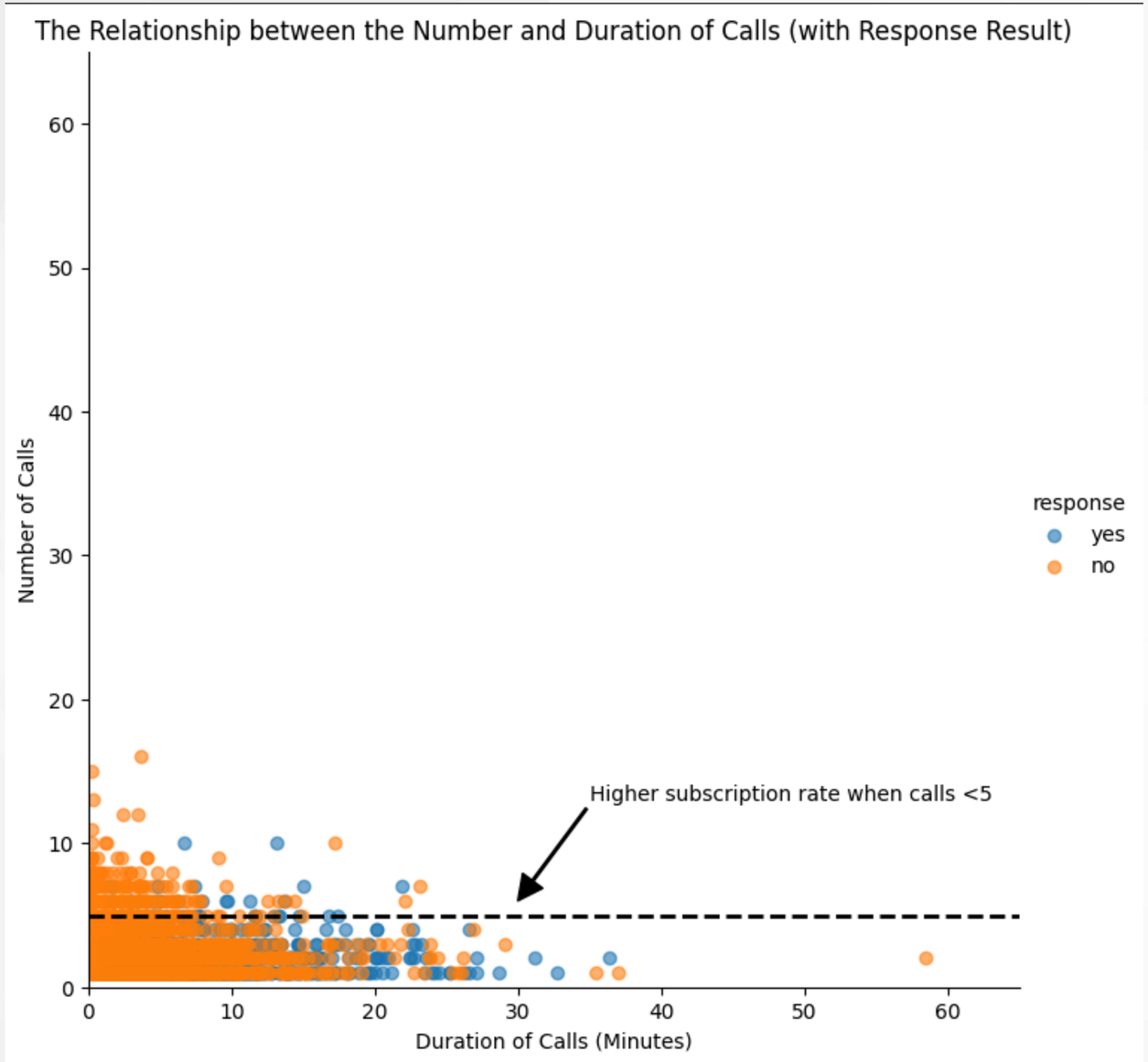

The bank contacted most clients between May and August. The highest contact rate is around **30%**, which happened in **May**, while the contact rate is closer to 0 in March, September, October, and December.

However, the subscription rate showed a different trend. The highest subscription rate occurred in **March**, which is over **50%**, and all subscription rates in September, October, and December are over 40%.




The boxplots show the distribution of call duration and the number of campaign contacts, highlighting their spread and central tendency.

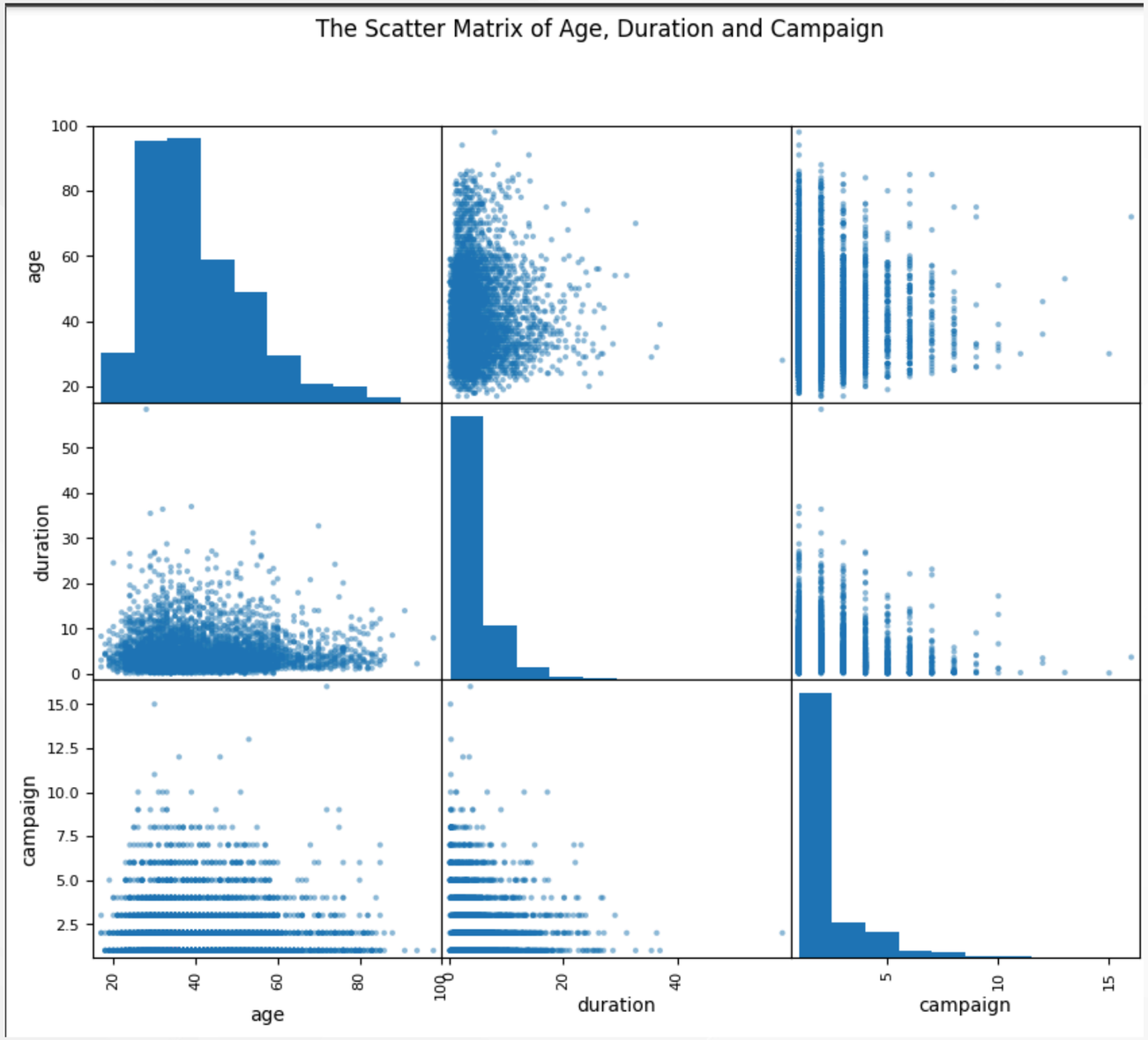
Outliers in both plots suggest cases with unusually long calls or frequent client contacts.




The scatter plot visualizes the relationship between call duration and the number of calls, categorized by client response (subscription or not).

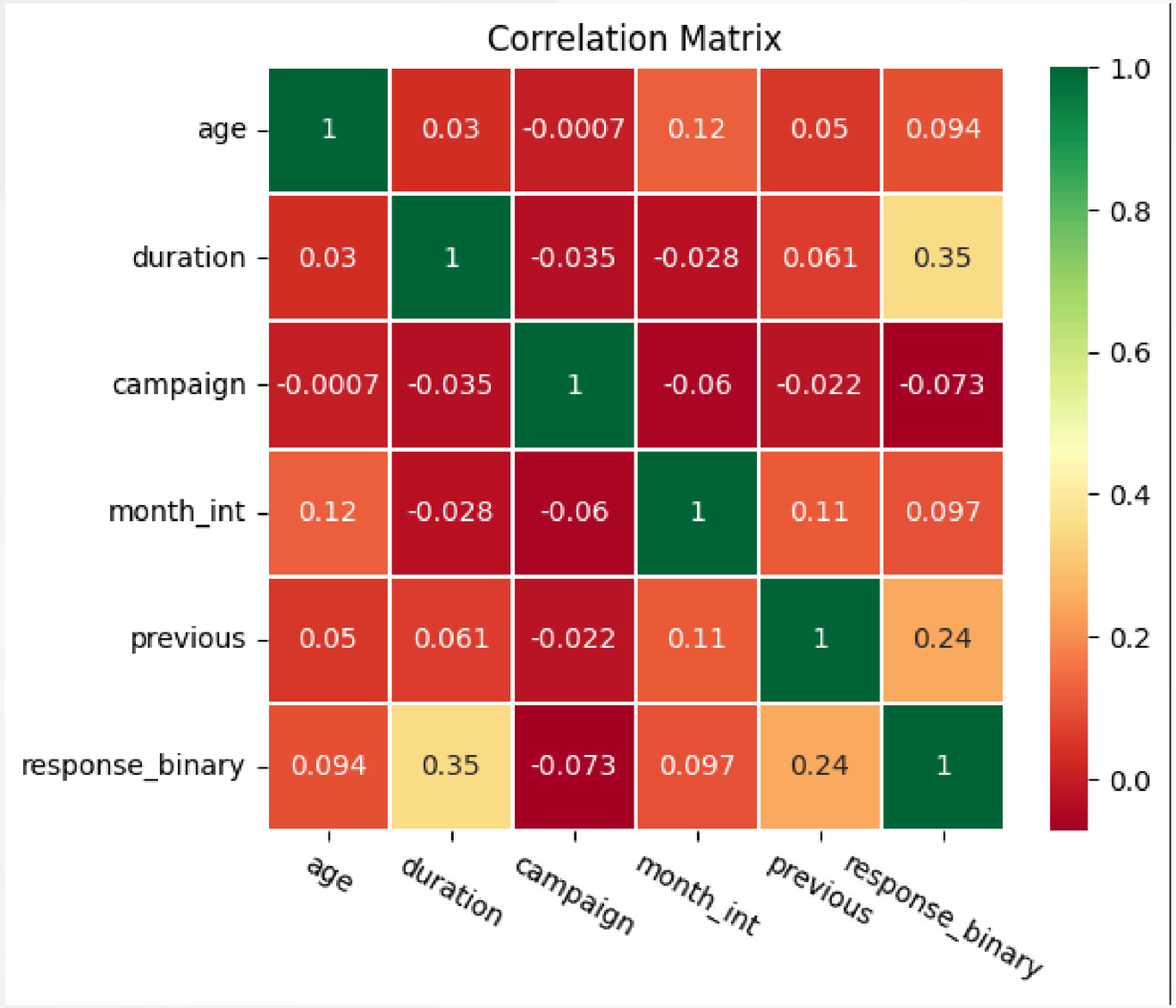
It shows that most calls are under **5 minutes**, with an annotation highlighting higher subscription rates for clients who were contacted fewer than **5 times.**

The plot provides insight into how call frequency and duration impact client decisions.



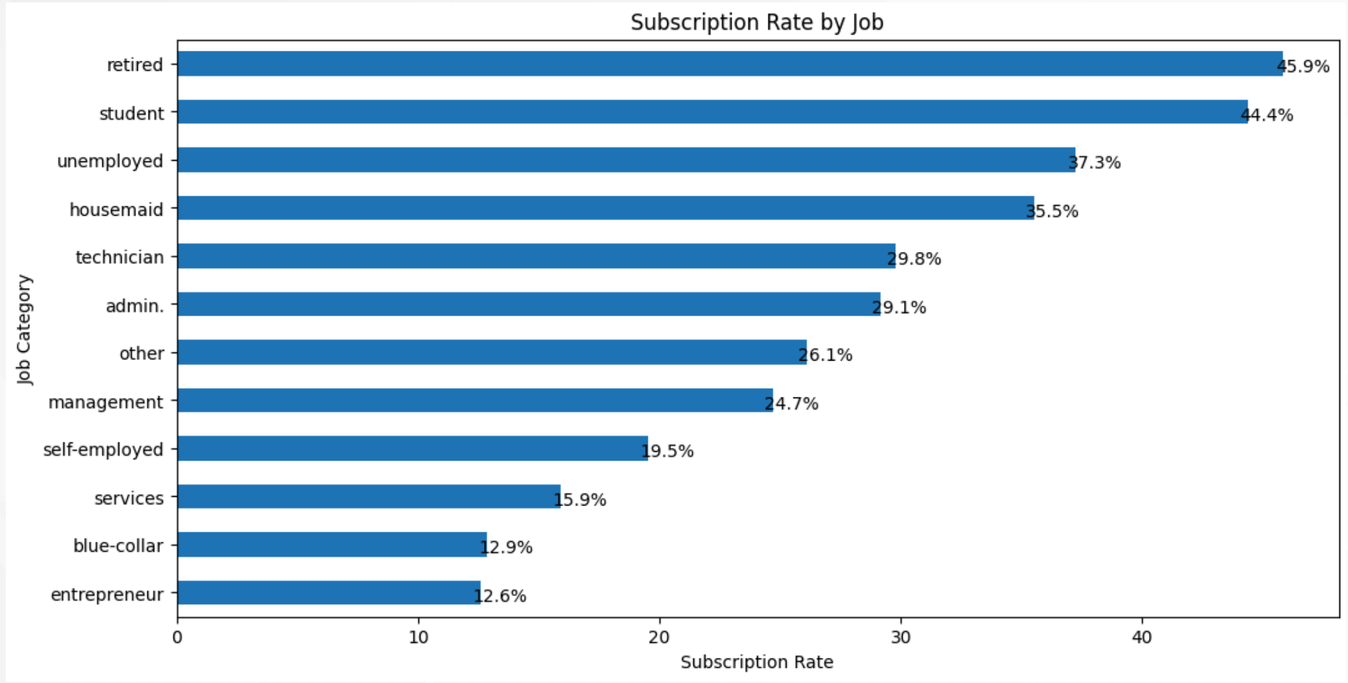

The scatter matrix illustrates the relationships among age, call duration, and the number of campaign contacts, highlighting potential correlations.

Diagonal plots show the distribution of each variable, helping to identify patterns in client behavior.



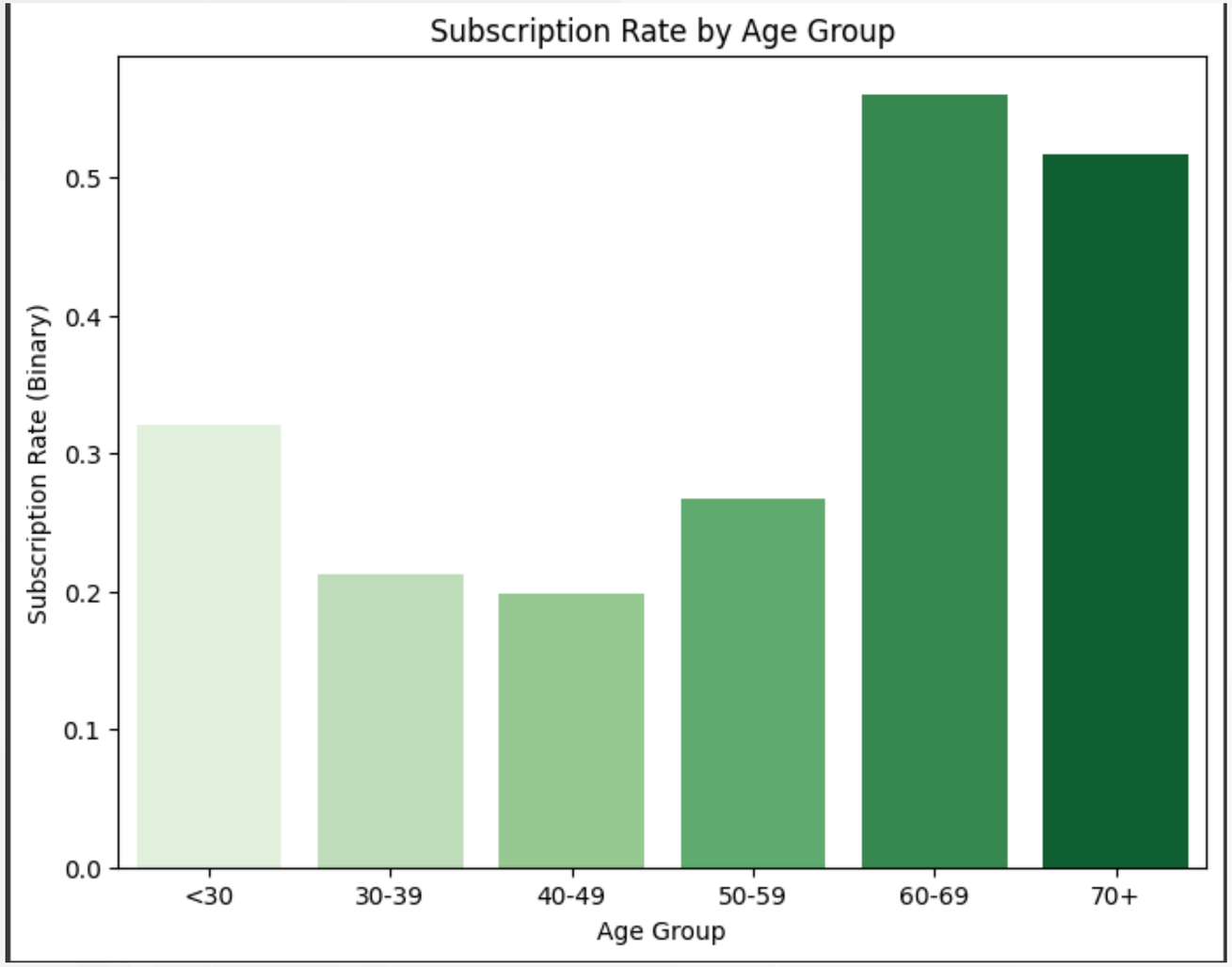

The heatmap shows the correlation matrix for variables such as age, call duration, and campaign contacts, with values indicating the strength and direction of relationships.

This visualization helps identify key factors that may influence client subscription behavior.



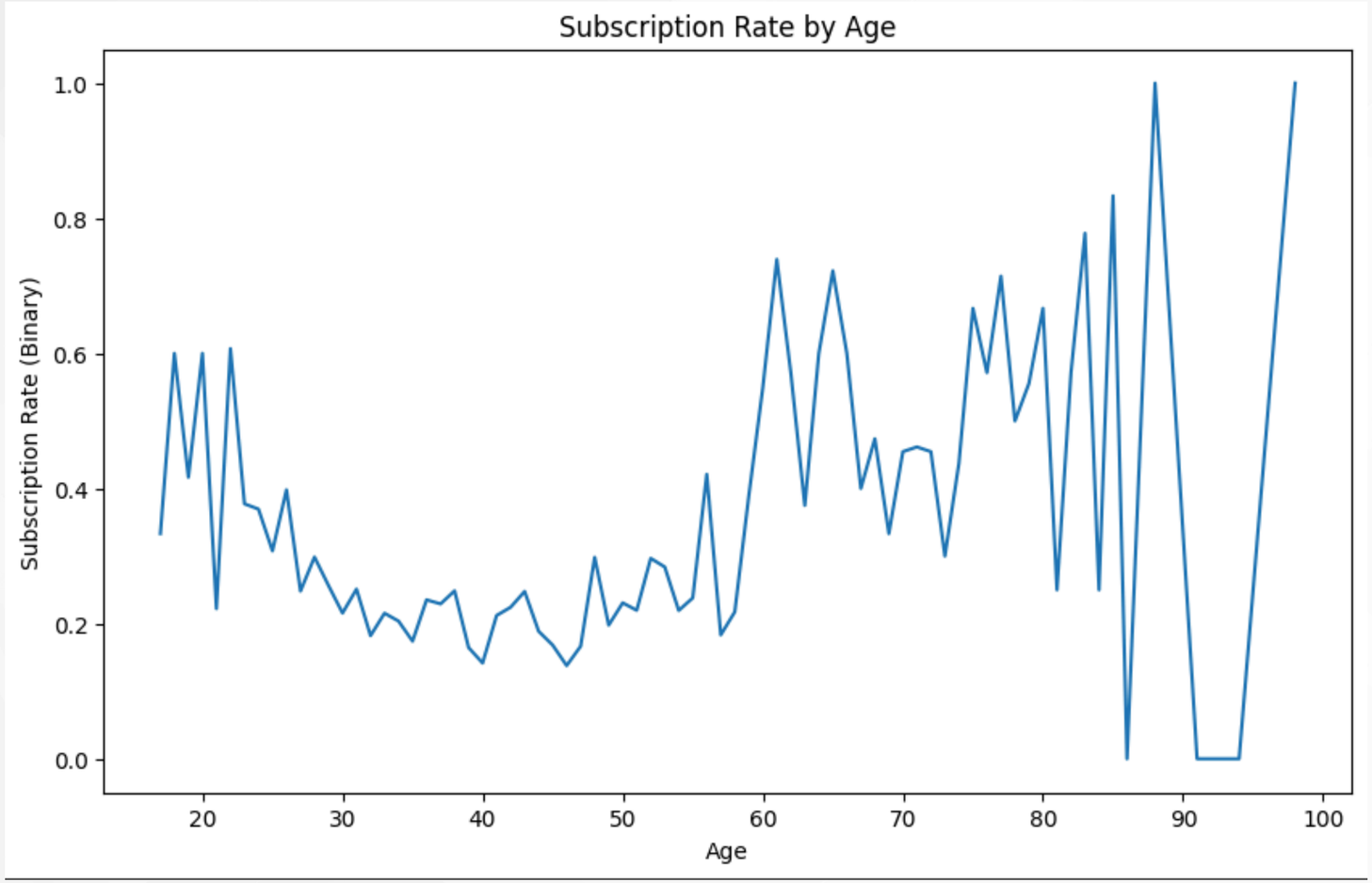

The horizontal bar chart illustrates the subscription rate by job category, showing the percentage of clients who subscribed to a term deposit.

More than **50%** of subscriptions are made up of students and retired clients, as seen by the horizontal bar chart. This conclusion is consistent with the earlier observation that younger and older clients subscribe at higher rates.




The bar chart shows the subscription rate to term deposits by age group, with rates indicating how likely clients are to subscribe.

It highlights trends in client responses, revealing which age groups are more or less likely to engage with the term deposit offers.




The line plot shows the subscription rate to term deposits by age, indicating how likelihood varies across different ages.

It helps identify age ranges that are more or less likely to subscribe, offering insights into client preferences.

# 3. CLASS IMBALANCE

1. **Class Imbalance Problem**: The dataset had a significant imbalance in the target variable y:
   - **88%** of the clients did not subscribe to a term deposit (no).
   - Only **12%** of the clients subscribed (yes).
2. **Solution**: Used SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority class. After applying SMOTE:
   - The proportion of the classes was balanced to approximately **50%** for each class, ensuring the model could better predict both subscribing and non-subscribing clients.

By oversampling the minority class, the model's performance on the minority class improved, particularly in terms of recall and F1-score.

# 4. MODEL BUILDING

Three different classification algorithms (**Random Forest, Decision Tree, Logistic regression**) were run on the dataset and the best-performing one was used to build the classification model.

All customer statistics were selected as features while the campaign outcome was set as target. **70%** of the data was used to build the classification model and **30%** was reserved for testing the model.

# 4. i. Random Forest

- **Model Application:** We applied a Random Forest model, which builds multiple decision trees and aggregates their results to make a more robust prediction.  frequency.

- **Result**:
  - Accuracy: 91.09%
  - F1-Score for ‘yes’: 0.51
  - F1-Score for ‘no’: 0.95

- **Interpretation**: The model achieved good overall accuracy, with the F1-score indicating effective handling of the minority class (clients who subscribed).

```
Accuracy: 0.9109007040543822

Classification Report:
              precision    recall  f1-score   support

          no       0.93      0.97      0.95     10968
         yes       0.67      0.41      0.51      1389

    accuracy                           0.91     12357
   macro avg       0.80      0.69      0.73     12357
weighted avg       0.90      0.91      0.90     12357


Confusion Matrix:
 [[10689   279]
 [  822   567]]
```

# 4. ii. Decision Tree

- **Model Application**: A single Decision Tree was used, where each decision splits the data based on the most informative features. The model was weighted to handle class imbalance.

- **Result**:
  - Accuracy: 89.34%
  - F1-Score for ‘yes’: 0.51
  - F1-Score for ‘no’: 0.94

- **Interpretation**: While simpler and more interpretable than Random Forest, the Decision Tree achieved slightly lower performance.

```
Accuracy: 0.8934207331876669

Classification Report:
              precision    recall  f1-score   support

          no       0.94      0.94      0.94     10968
         yes       0.53      0.50      0.51      1389

    accuracy                           0.89     12357
   macro avg       0.73      0.72      0.73     12357
weighted avg       0.89      0.89      0.89     12357


Confusion Matrix:
 [[10347   621]
 [  696   693]]
```

# 4. iii. Logistic Regression

- **Model Application**: Logistic Regression, a linear model, was applied to predict the probability of a client subscribing.

- **Result**:
  - Accuracy: 91.11%
  - F1-Score for ‘yes’: 0.43
  - F1-Score for ‘no’: 0.97

- **Interpretation**: Logistic Regression provided a solid baseline model. However, its linear nature limits its ability to capture complex patterns in the data, which explains the slightly higher performance compared to tree-based models.

```
Accuracy: 0.9112648701141054

Classification Report:
              precision    recall  f1-score   support

          no       0.93      0.97      0.95      7303
         yes       0.67      0.43      0.52       935

    accuracy                           0.91      8238
   macro avg       0.80      0.70      0.74      8238
weighted avg       0.90      0.91      0.90      8238


Confusion Matrix:
 [[7109  194]
 [ 537  398]]
```

# Conclusion

Among all algorithms, **Logistic Regression** had the highest accuracy, about **91.11%**, so it would be used to predict customers' responses.

According to previous analysis, a target customer profile can be established. The most responsive customers possess these features:

- **Feature 1**: age < 30 or age > 60
- **Feature 2**: students or retired people
- **Feature 3** : Clients with a university degree
- **Feature 4** : who had call durations longer than 200 seconds


CHART OF
**ACCURACY**

# Model Evaluation Matrices

# SUMMARY

The **Logistic Regression** algorithms were successfully used to build the estimation and classification model. The bank will be able to forecast a customer's reaction to its telemarketing campaign using these two models before making contact with them. By doing this, the bank may focus more of its marketing efforts on customers who are most likely to take term deposits and less on those who are not likely to do so.

# THANK YOU