Tourism Page Engagement

Description:

- Anjali
  - What is the dataset?
  - What are the questions you want to answer?
  - Importance of problem?

EDA:

Basic stats or intuitions. Show patterns and abnormalities (related or not to goal).

- Shirley
  - Correlation map
  - Distribution of cat and numeric columns
- Kush
  - Clustering
  - Rest of EDA

Solution and Insights:

Features used, summary of results, feature importance, anything surprising?

- Haden (NN Model)
- Twinkle (Boosting)
- Ari (Knn & NB)
- (ORDER: NB, Boosting, Knn, NN)

Taken_product (Target): Whether or Not you buy a ticket in the next month

Yearly_avg_view_on_travel_page: Avg. yearly views on any travel-related page by the user

Total_likes_on_outstation_checkin_given: Total number of likes given by the user on out-of-station check-ins in the last year // Cumulative count of "likes" that a user has given on check-ins that were made outside of their usual location or station (out-of-station) over the past year

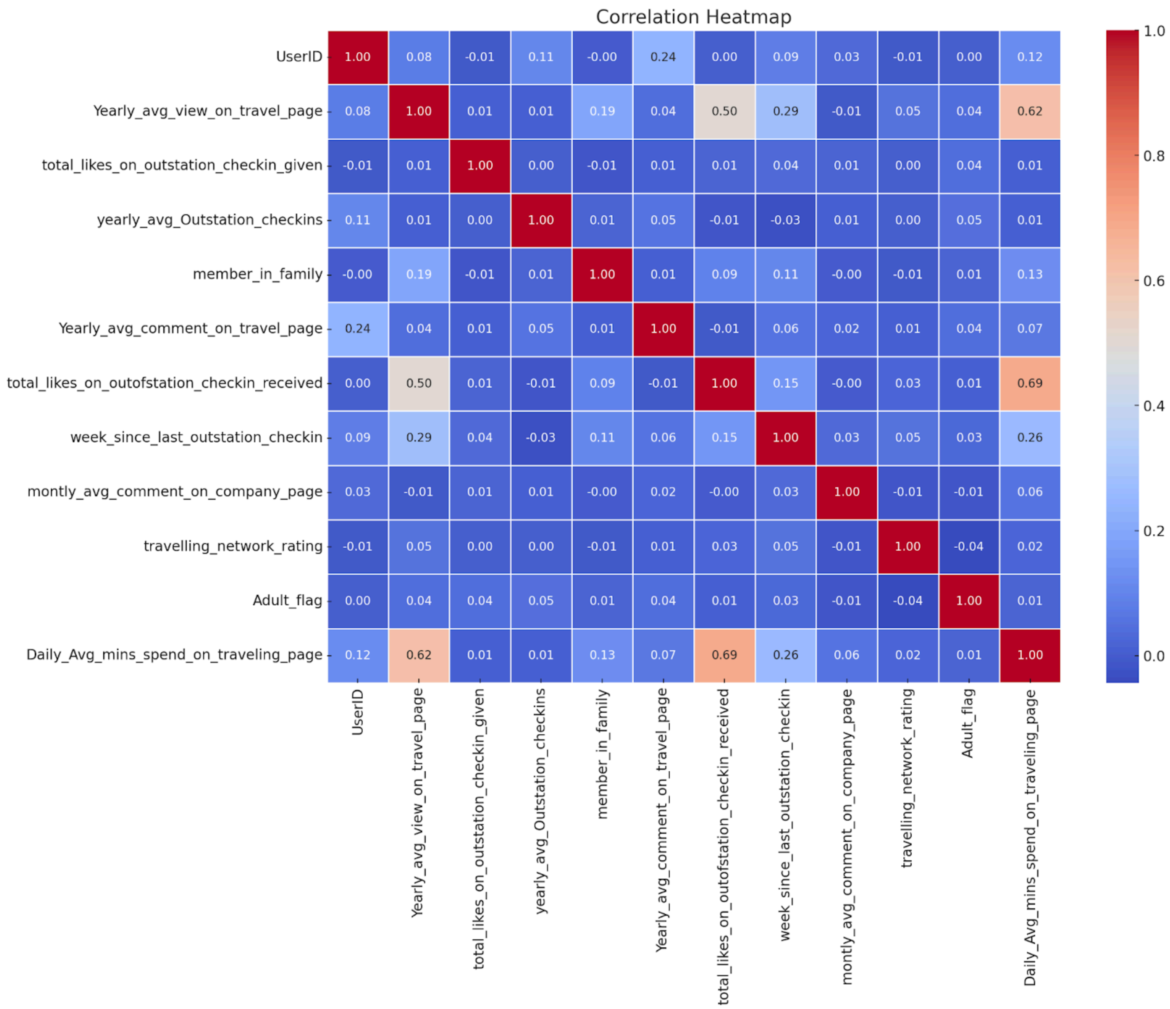
Problem Statement

Can we predict whether a customer would buy a ticket based on:

1. Psychographic Variables
2. Demographic Variables

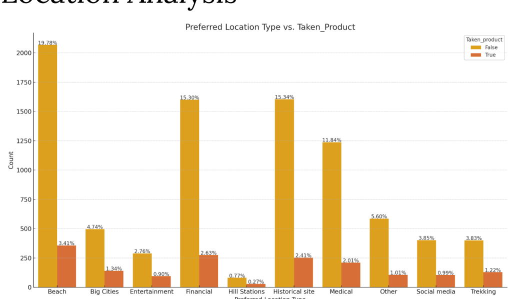
1. EDA

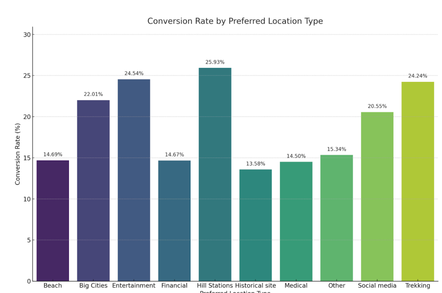
CORRELATION HEATMAP



DISTRIBUTION OF CATEGORICAL VARIABLES:

- Location Analysis

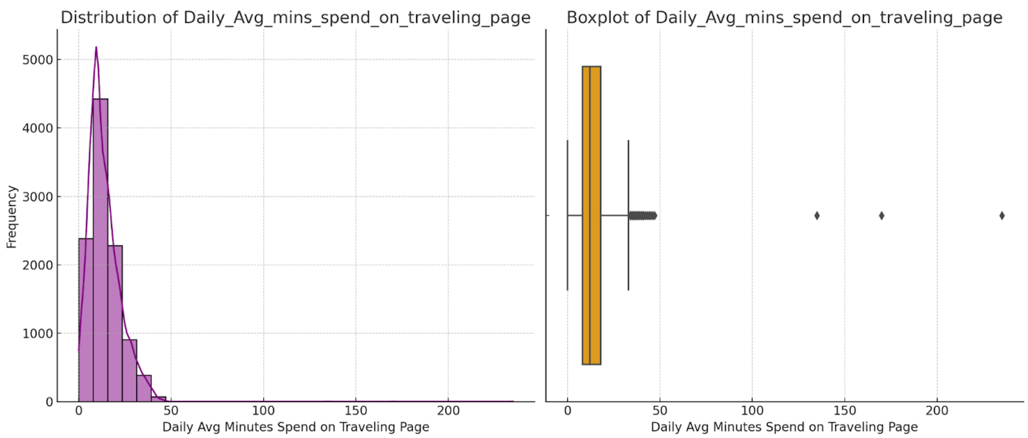




According to the analysis on preferred location, it indicates a significant discrepancy between user interest in certain locations and their likelihood of making a purchase. Despite high levels of interest in locations such as Beaches, Financial districts, and Historical Sites, the conversion rates for these categories are notably low.

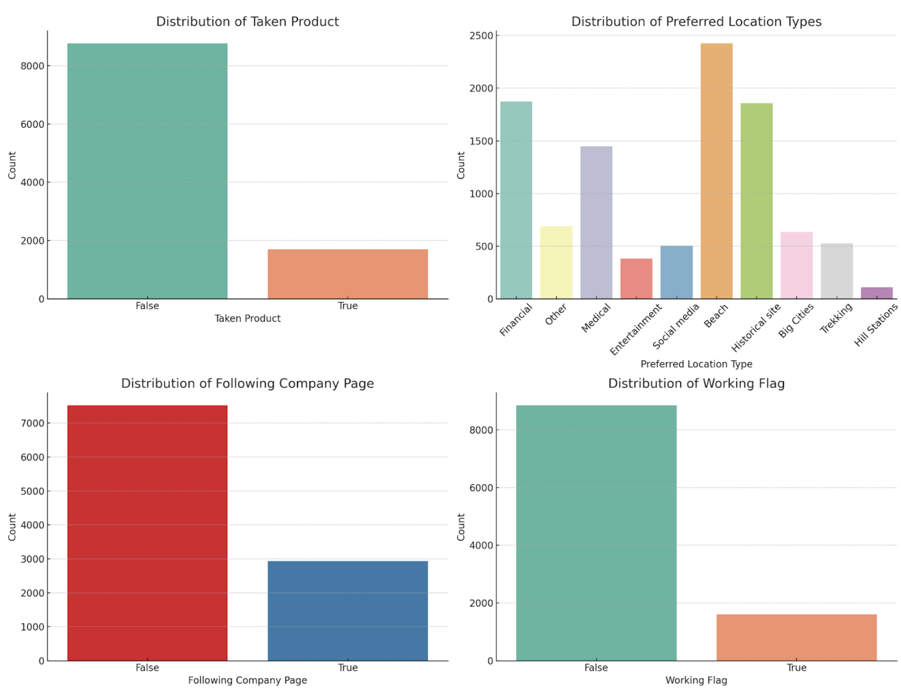
VISUALIZATION FOR THE DAILY_AVG-MINS-SPEND_ON_TRAVELLING_PAGE:

Average time spent on the company's travel page by the user



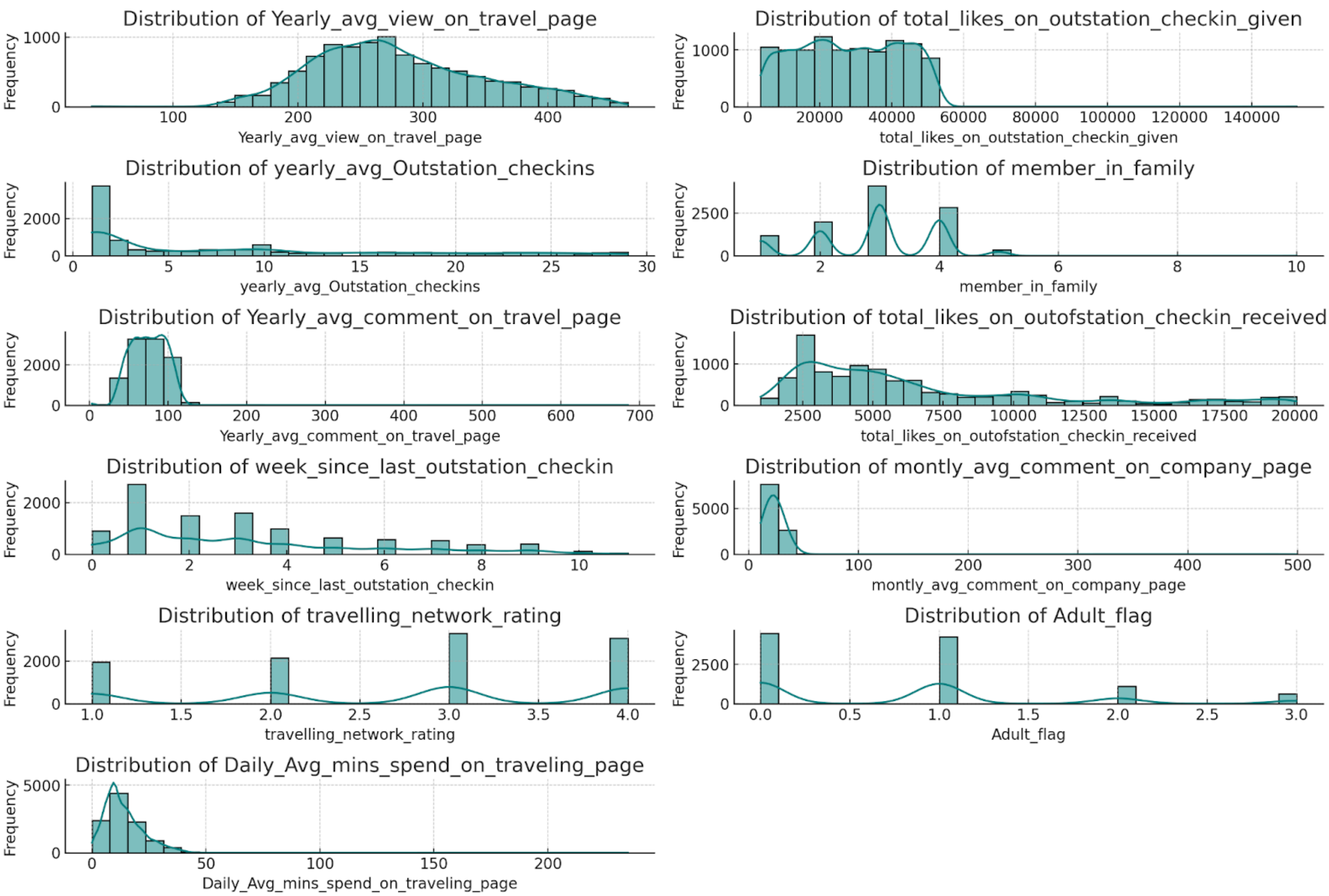
The third plot indicates whether users follow the company page.

The fourth plot shows the distribution of the working flag.



DISTRIBUTION OF NUMERICAL VARIABLES:

The histograms show the distribution of different numerical features, which can give insight into the distribution and if there's any potential skewness

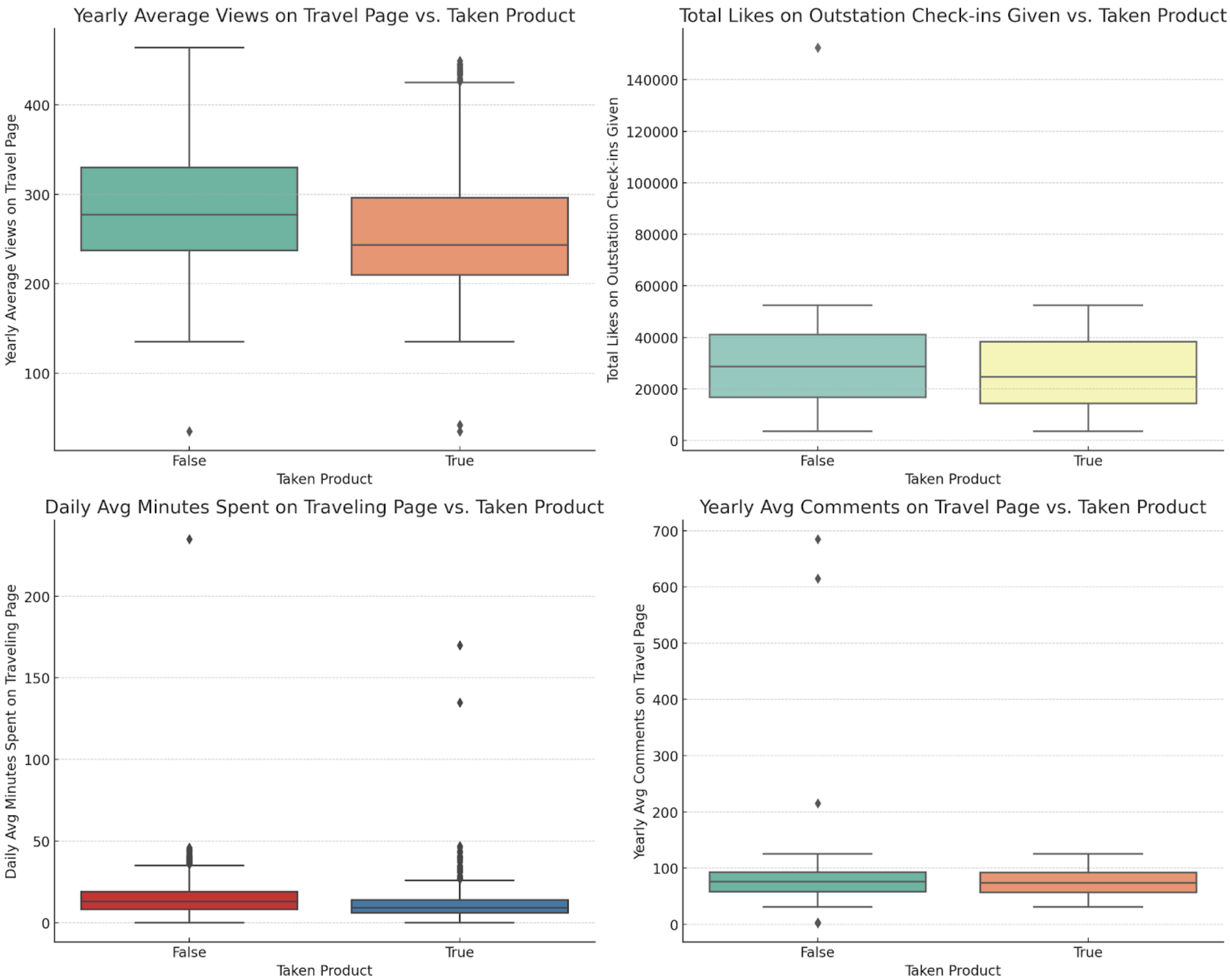


- **Distribution of yearly_avg_view on travel page**
- **Daily_avg_mis_spend on traveling page**
- **Distribution of total likes on outofstation checkin recieved**

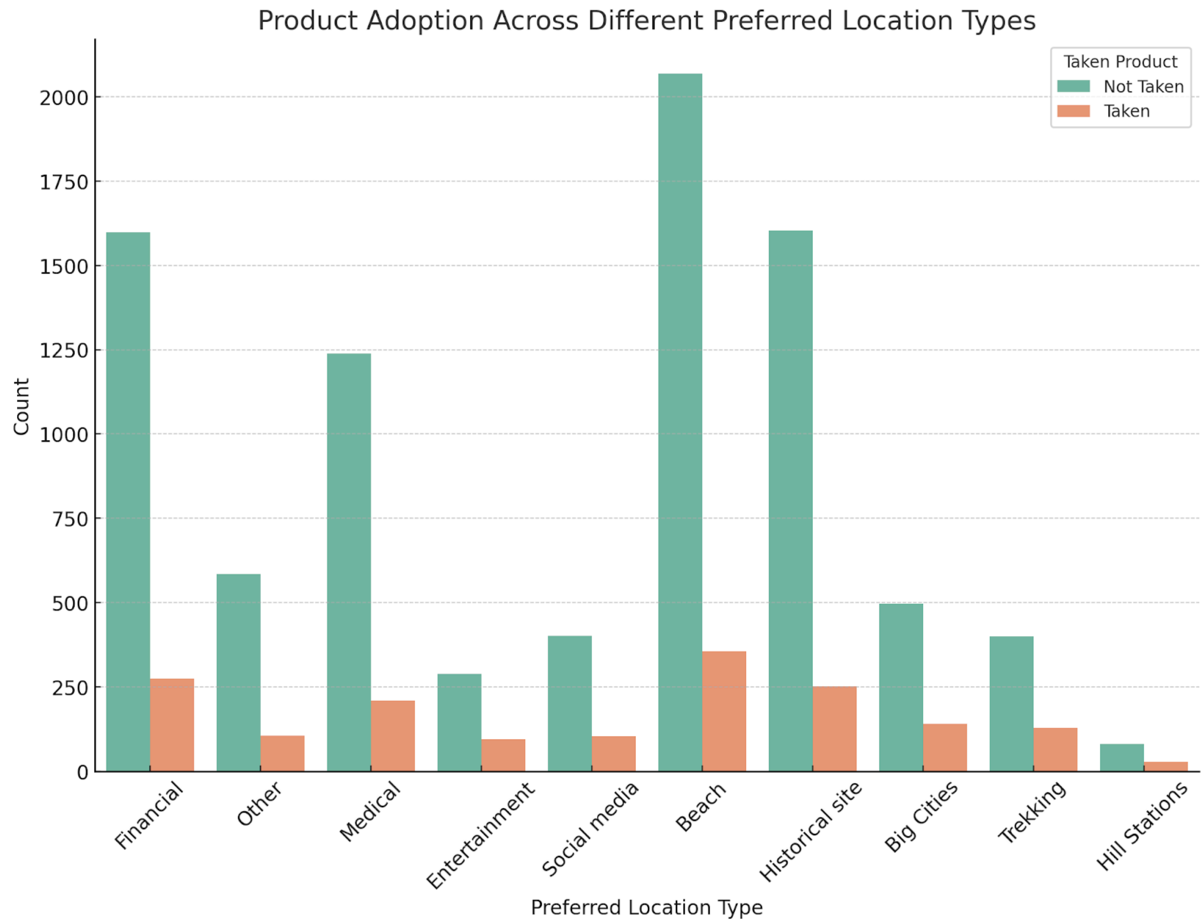
Q1: USER ENGAGEMENT VS. PRODUCT ADOPTION

Eg: Users with some higher engagement metrics (ex. Yearly_avg_views on the travel page, likes_on_outstation_check_ins) may be more likely to purchase the product

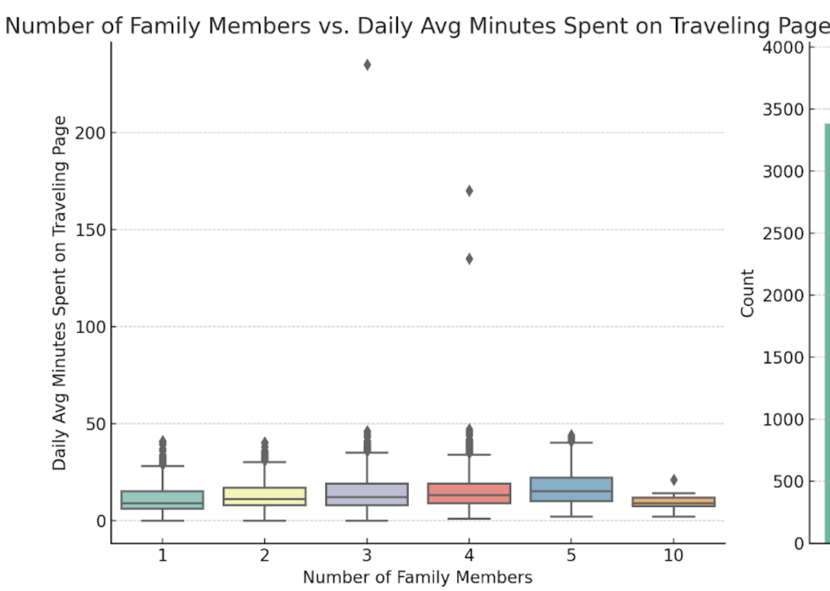
Does the level of user engagement correlate with the purchasing of the product? If yes, how?

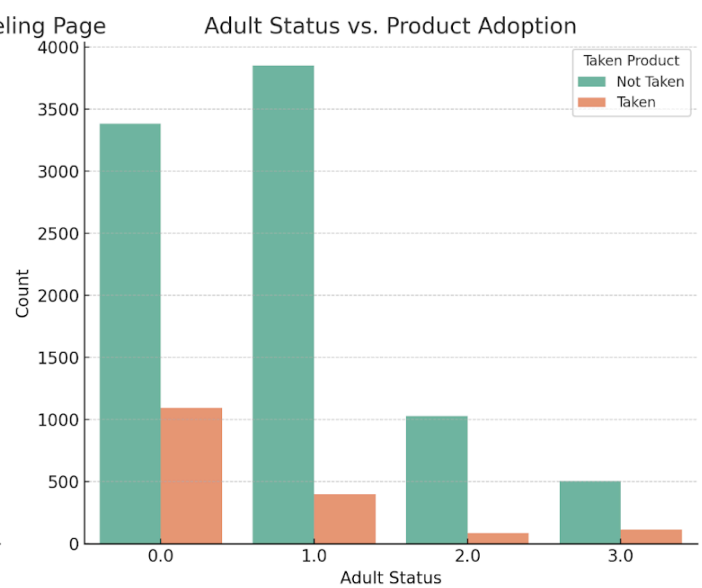


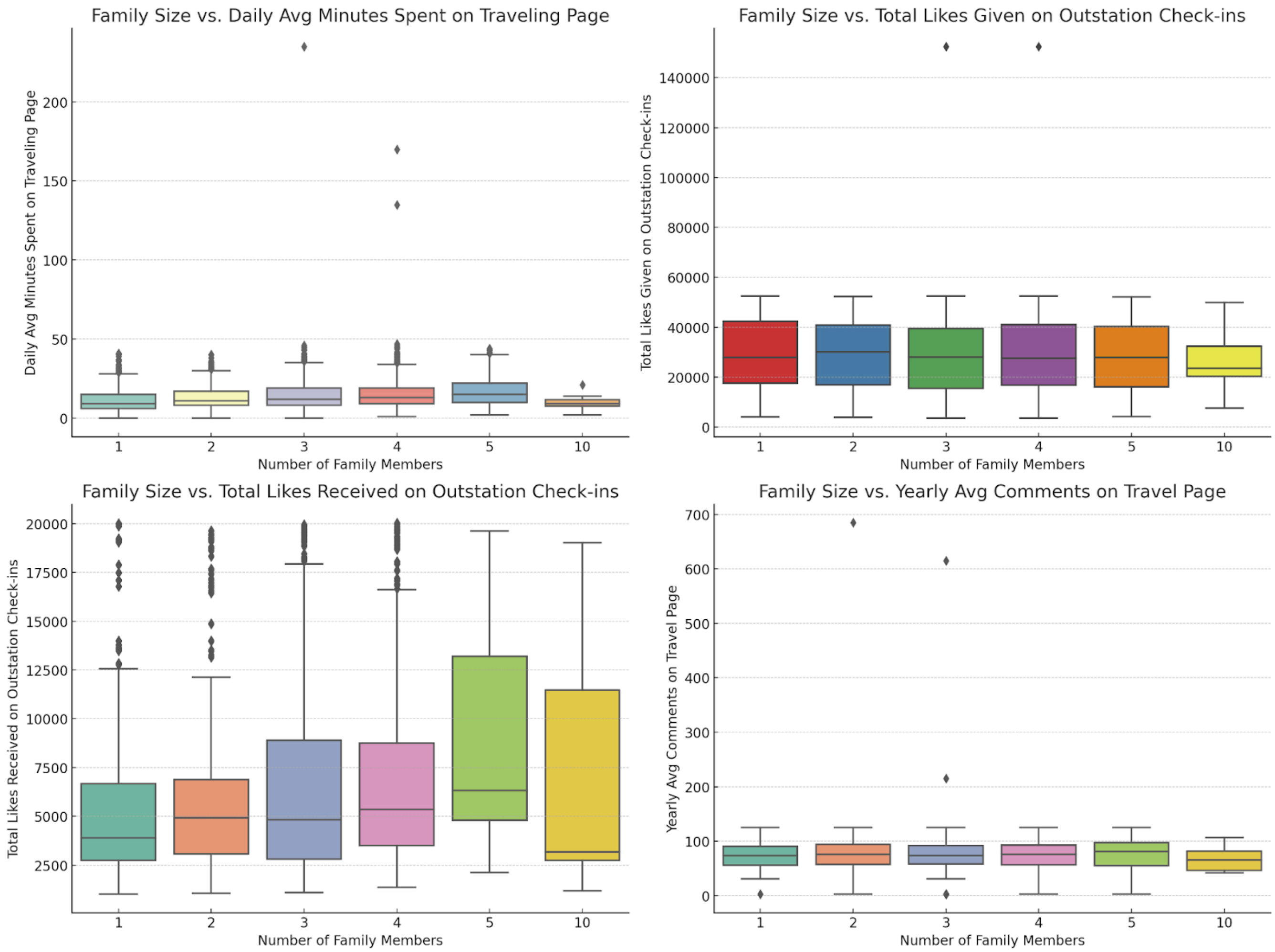
Q2: IMPACT OF THE PREFERRED LOCATION TYPE



Q3: DEMOGRAPHIC INFLUENCE



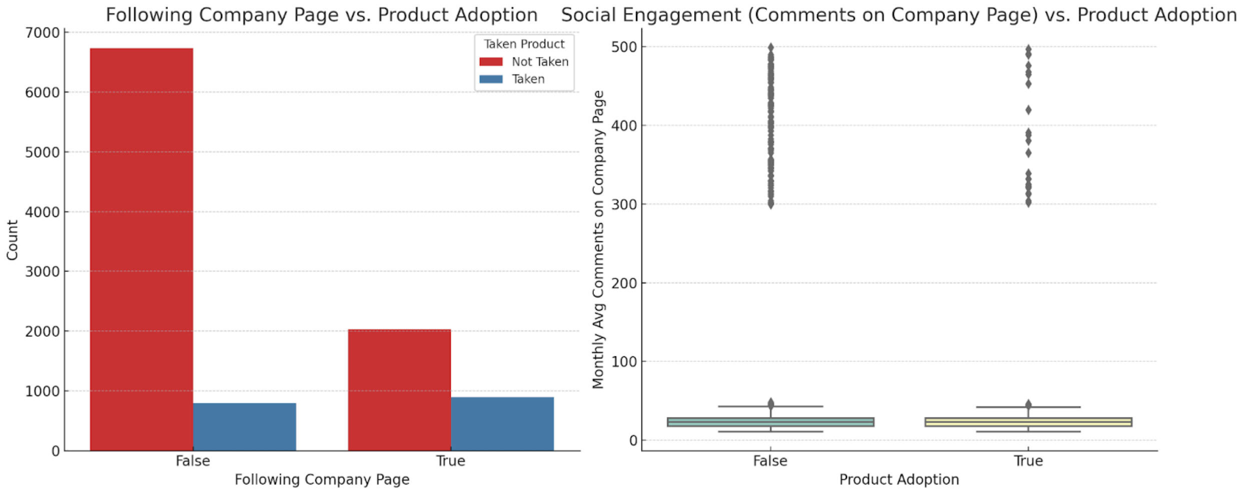




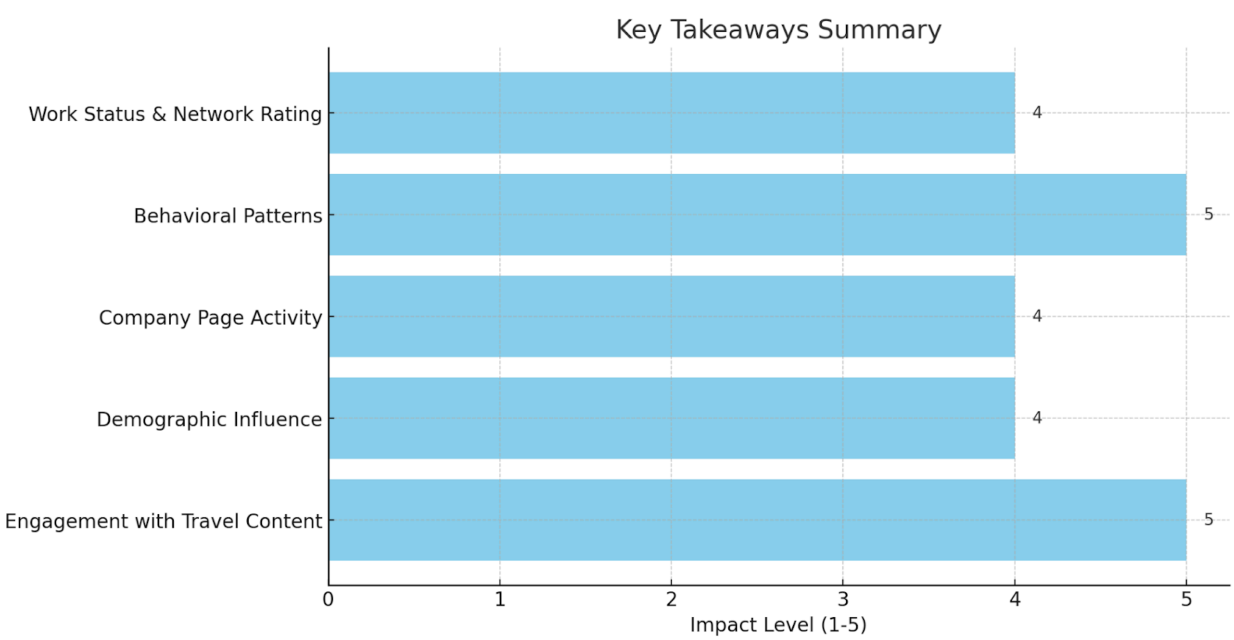
Adults Status vs Adult Flag?

For individual who made a purchase, plot their location counts by family size

Q4: EFFECT OF SOCIAL MEDIA INFLUENCE



KEY TAKEAWAYS:



2. MODELS

1. KNN

Data Split to 70 - training and 30 - testing, seed set at 42

Feature Importance:

1. total_likes_on_outstation_checkin_given
2. yearly_avg_Outstation_checkins
3. Yearly_avg_comment_on_travel_page
4. following_company_page
5. total_likes_on_outofstation_checkin_received

The KNN model achieved an accuracy of approximately 96.3% on the test set. Here's a summary of the model's performance:

Precision:

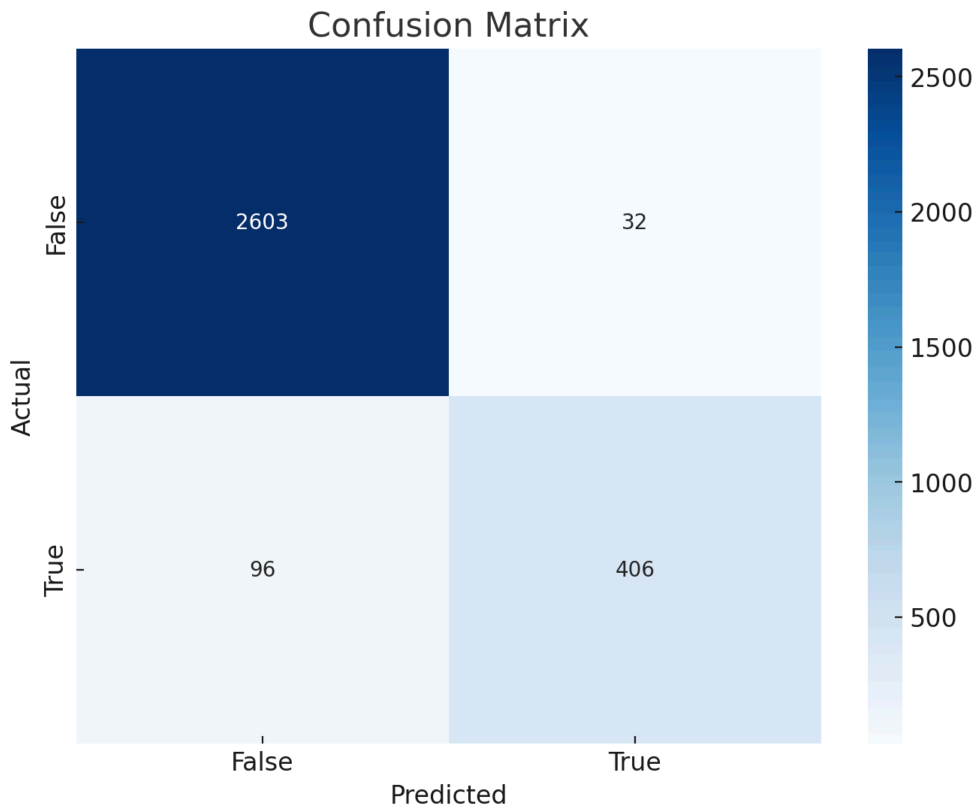
- For `False` (not taking the product): 96%
- For `True` (taking the product): 93%

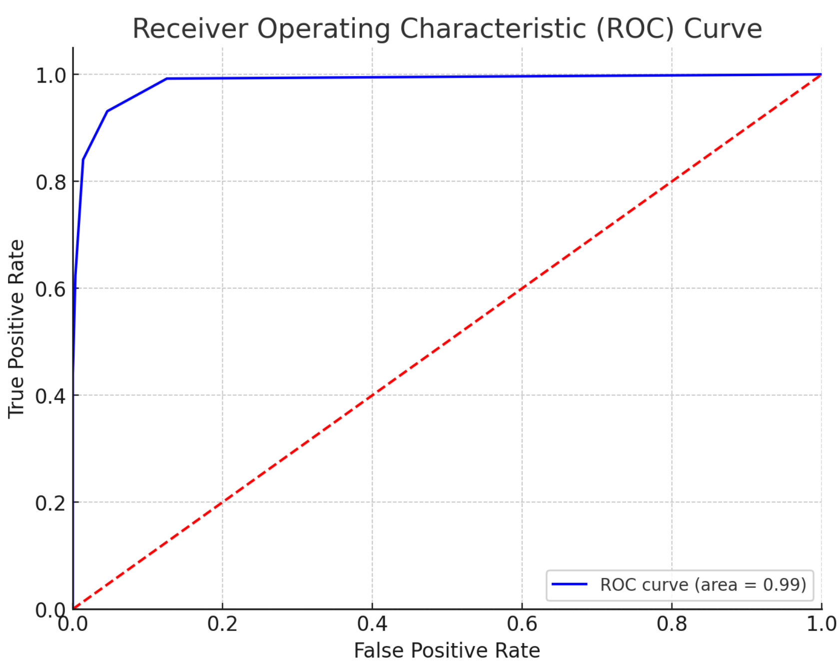
Recall:

- For `False`: 99%
- For `True`: 81%

F1-Score:

- For `False`: 98%
- For `True`: 86%


Confusion Matrix
Actual
False
True
2603
32
96
406
False
True
Predicted
2500
2000
1500
1000
500


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.99)

The curve shows the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different thresholds. The area under the curve (AUC) is approximately 0.99, indicating that the model has a high capability of distinguishing between the positive and negative classes.

2. Naive Bayes

Data Split: 80 - Training, 20 - Testing

The NB model achieved an accuracy of approximately 85.1% on the test set. Here's a summary of the model's performance:

Precision:

- **For `False` (not taking the product): 86%**
- **For `True` (taking the product): 65%**

Recall:

- **For `False`: 98%**
- **For `True`: 17%**

F1-Score:

- **For `False`: 92%**
- **For `True`: 27%**

3. Neural Network

Data Split: 70 - Training, 30 - Testing

Objective:

The goal was to develop a neural network model to accurately predict whether a product would be taken (`Taken_Product = 1`) or not (`Taken_Product = 0`), despite a significant class imbalance (8761 records as `0` and 1693 as `1`).

Model Architecture:

A deep neural network with residual connections was employed to improve learning and mitigate the vanishing gradient problem. The architecture included:

- Residual Blocks: Four blocks with linear layers, batch normalization, and ReLU activations, designed to facilitate effective gradient flow.
- Dropout Layer: A 0.4 dropout rate was used to prevent overfitting.
- Fully Connected Layer: The output was passed through a final fully connected layer with a sigmoid activation to produce the binary classification.

Handling Class Imbalance:

Class weights were applied in the loss function to penalize misclassifications of the minority class (`1`). This approach ensured the model remained focused on correctly identifying positive cases, which was critical due to the imbalance.

Optimization Techniques:

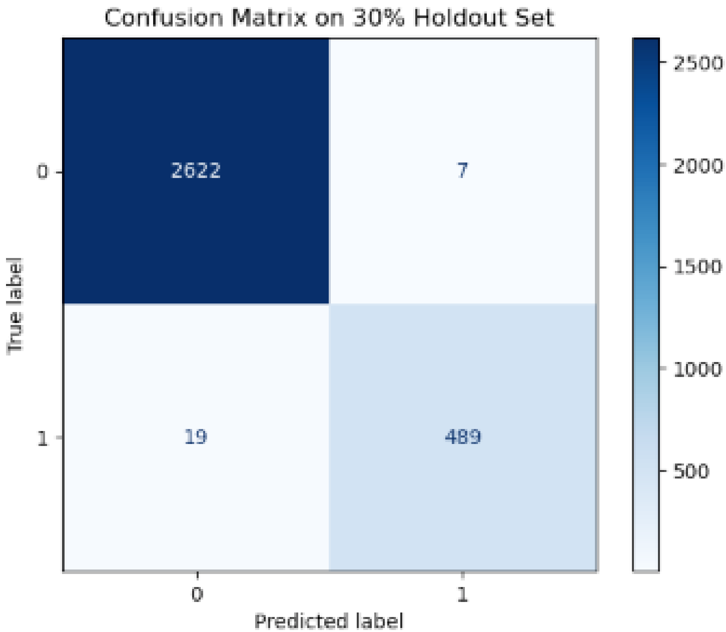
- Adam Optimizer: An adaptive learning rate optimizer was used for efficient training.
- Learning Rate Scheduler: A `ReduceLROnPlateau` scheduler adjusted the learning rate based on validation loss to ensure effective convergence.
- Gradient Clipping: To prevent exploding gradients, clipping was applied with a maximum norm of 1.0.

Results:

After training on 70% of the data and evaluating on a 30% holdout set, the model achieved:

- Accuracy: 0.9917
- Precision: 0.9859
- Recall: 0.9626
- F1 Score: 0.9741
- Specificity: 0.9973

These results indicate that the model performed exceptionally well, especially in identifying the minority class (`1`), which was the main focus of the project.


Confusion Matrix on 30% Holdout Set
2622
7
19
489
True label
Predicted label
0
1
0
1
2500
2000
1500
1000
500

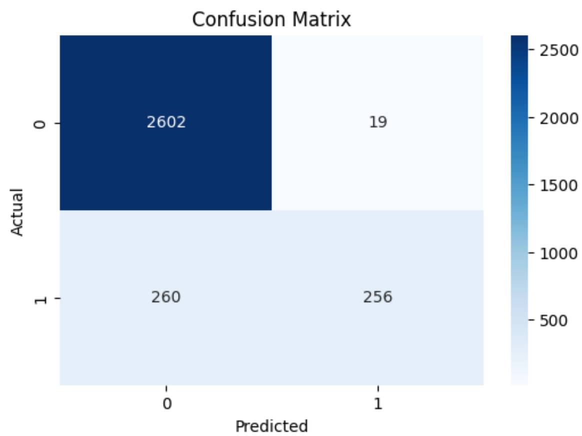
4. Boosting

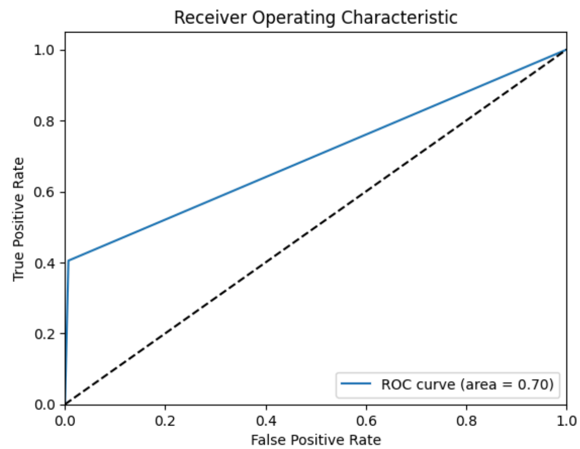
GBM

Data Split: 70 - Training 30 - Testing

Evaluation metrics:

Accuracy: 0.91 Precision: 0.93 Recall: 0.50 F1 Score: 0.65 ROC AUC Score: 0.70





XGBOOST

Data Split: 70 - Training 30 - Testing

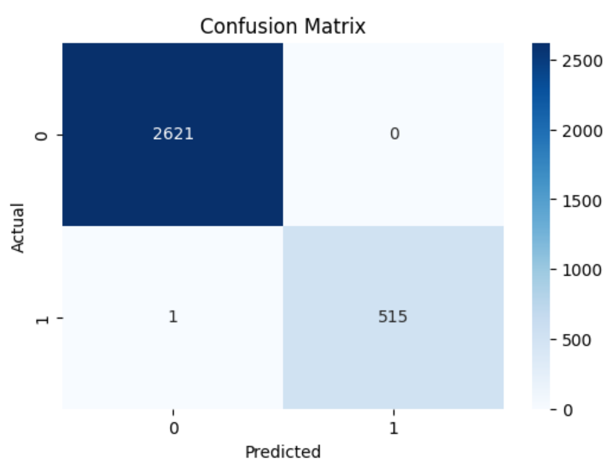
Pros:

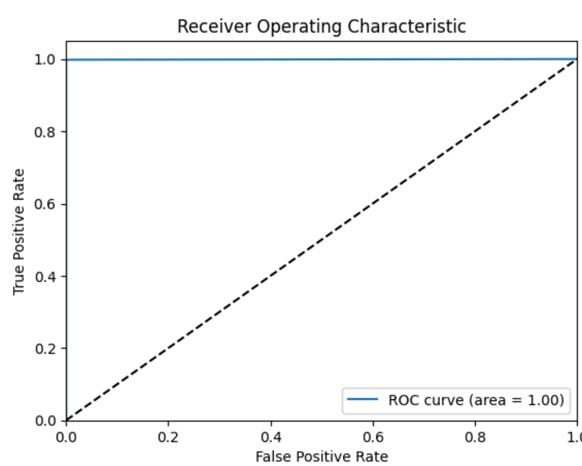
- Regularization properties reduce overfitting
- Parallel processing in XGB enabling faster training
- Produced better results than GBM

Approach: Initially a basic XGBClassifier yielded us 89% accuracy. Post this we expanded our grid to find the best parameters for the classifier. The final parameters selected using GridSearchCV are: colsample_bytree: 0.8 gamma: 0 learning_rate: 0.1 max_depth: 10 n_estimators: 1000

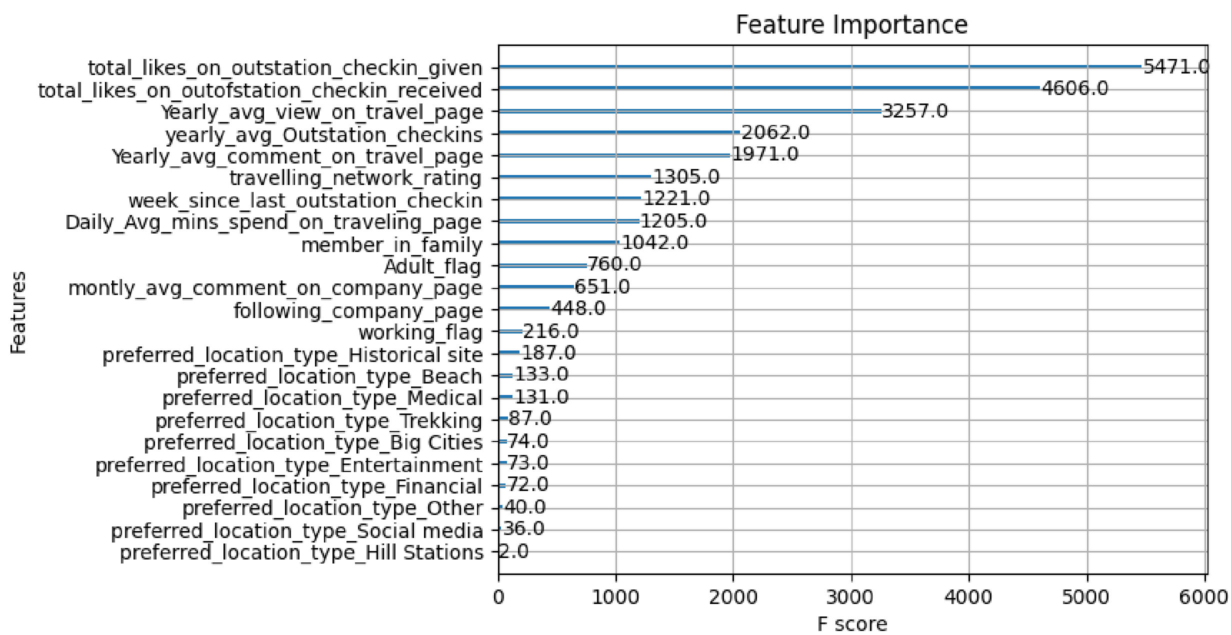
Evaluation metrics:

Accuracy: 1.00 Precision: 1.00 Recall: 1.00 F1 Score: 1.00 ROC AUC Score: 1.00





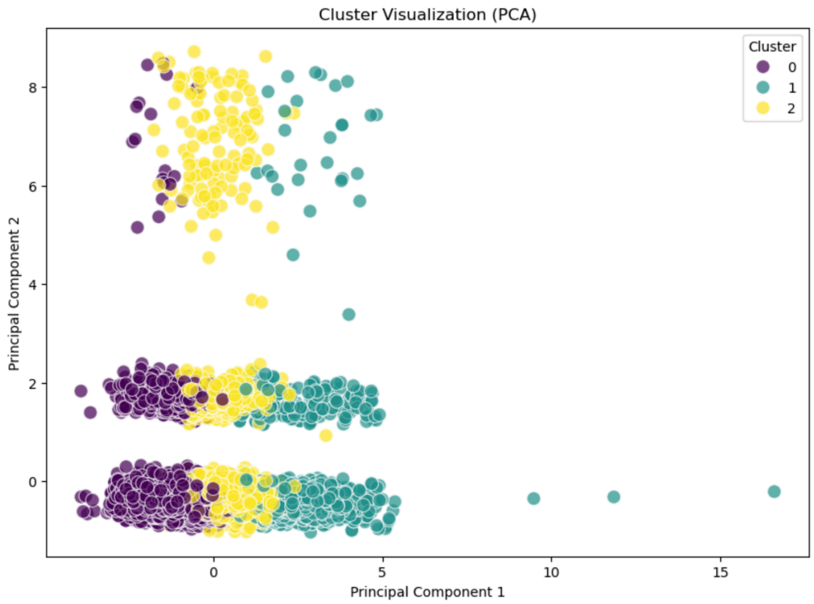
Variable Importance:

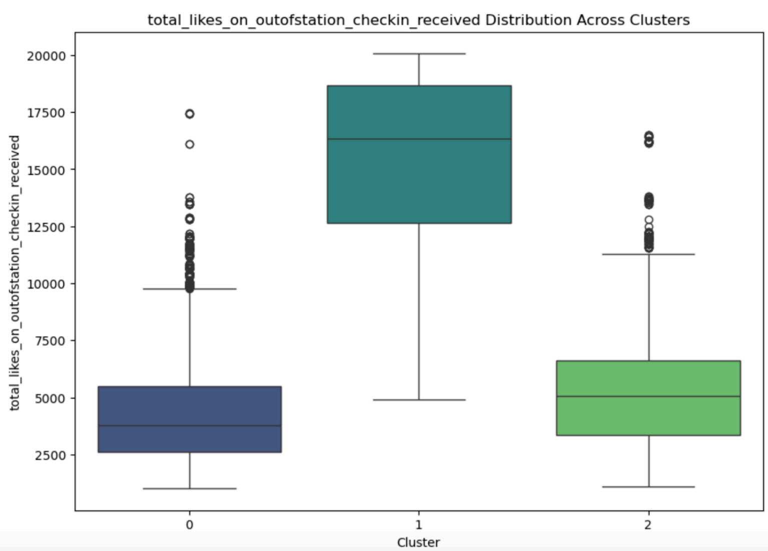


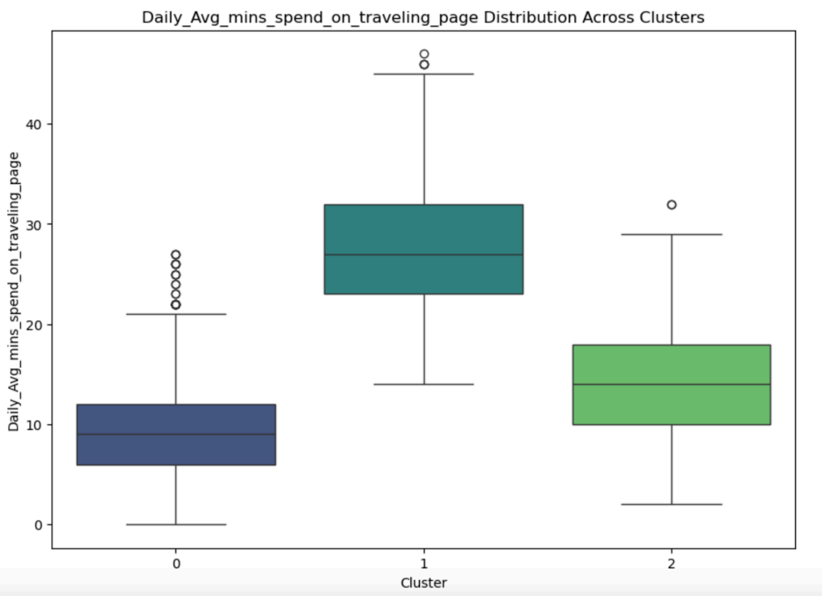
Kush Cluster:

Columns included

columns = [
    'Yearly_avg_view_on_travel_page', 'total_likes_on_outstation_checkin_given',
    'yearly_avg_Outstation_checkins', 'member_in_family',
    'Yearly_avg_comment_on_travel_page', 'total_likes_on_outofstation_checkin_received',
    'week_since_last_outstation_checkin', 'following_company_page',
'montly_avg_comment_on_company_page',
    'working_flag', 'travelling_network_rating', 'Adult_flag',
'Daily_Avg_mins_spend_on_traveling_page'
]


Cluster Visualization (PCA)
Cluster
0
1
2
Principal Component 2
Principal Component 1
0
2
4
6
8
0
5
10
15


total_likes_on_outofstation_checkin_received Distribution Across Clusters
total_likes_on_outofstation_checkin_received
Cluster
20000
17500
15000
12500
10000
7500
5000
2500
0
1
2


Daily_Avg_mins_spend_on_traveling_page Distribution Across Clusters
Daily_Avg_mins_spend_on_traveling_page
0
10
20
30
40
0
1
2
Cluster