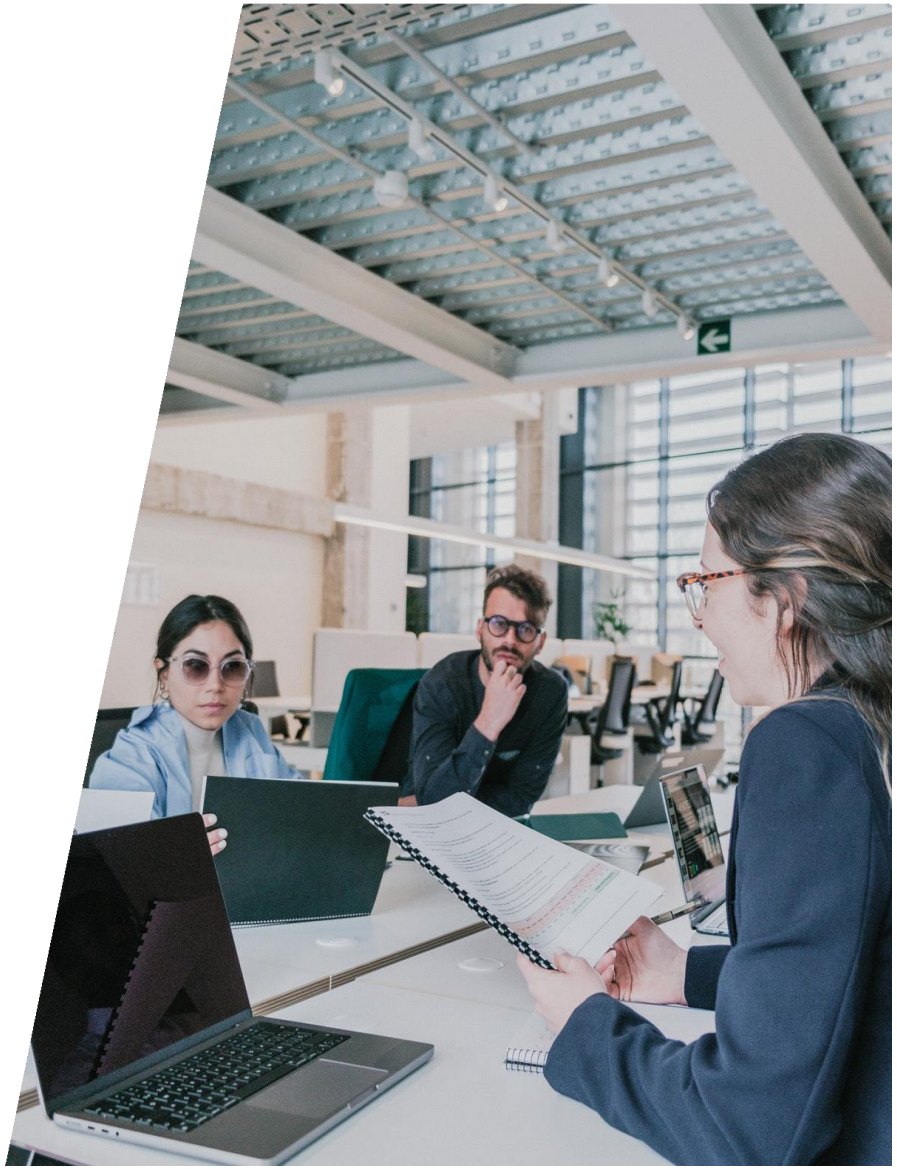
DIGITAL MARKETING CAMPAIGN CONVERSION PREDICTION



Presented By: Mohammad Amil Khan

PROBLEM STATEMENT

Develop a robust machine learning model to accurately predict customer conversions based on various demographic and engagement factors. By utilizing this model, the company aims to improve campaign targeting.

OBJECTIVE

This project aims to enhance the effectiveness of digital marketing campaigns by accurately predicting customer conversions. By leveraging machine learning, the project seeks to identify potential converters and optimize marketing strategies. The objective is to develop a robust model that predicts customer conversions based on demographic and engagement data, enabling:

- **Improve Campaign Targeting:** Identify potential converters, allowing for more precise and efficient marketing efforts.

- **Increase Conversion Rates:** Enhance the effectiveness of campaigns by focusing on the most promising leads.

WORK FLOW

Step 1 | Import Libraries	Importing Required Libraries
Step 2 | Read Dataset	Gathering and organizing data to train machine learning models.
Step 3 | Dataset Overview	Basic and Descriptive Data Overview
Step 4 | EDA	Analyzing and visualizing data patterns to understand its characteristics
Step 5 | Data Preprocessing	Preparing and cleaning data to enhance its quality and suitability
Step 6 | Split Train And Test Data	Dividing the dataset into training and testing sets to evaluate
Step 7 | Model Training	Choosing a suitable machine learning algorithm and optimizing its parameters
Step 8 | Model Evaluation	Assessing the performance of a machine learning model using metrics
Step 9 | Conclusion	Conclusion of Models and EDA
Step 10 | Power BI Dashboard	Power BI Dashboard and it's Conclusion

DATA COLLECTION & REFINEMENT

01 **Demographics**: Age, Gender, Income

0 2 **Engagement Metrics**: Adspend, socialshare, so on.

Target Variable =Conversion
0=N0 & 1 =Yes

Data types: float64(5), int64(10), object(5)

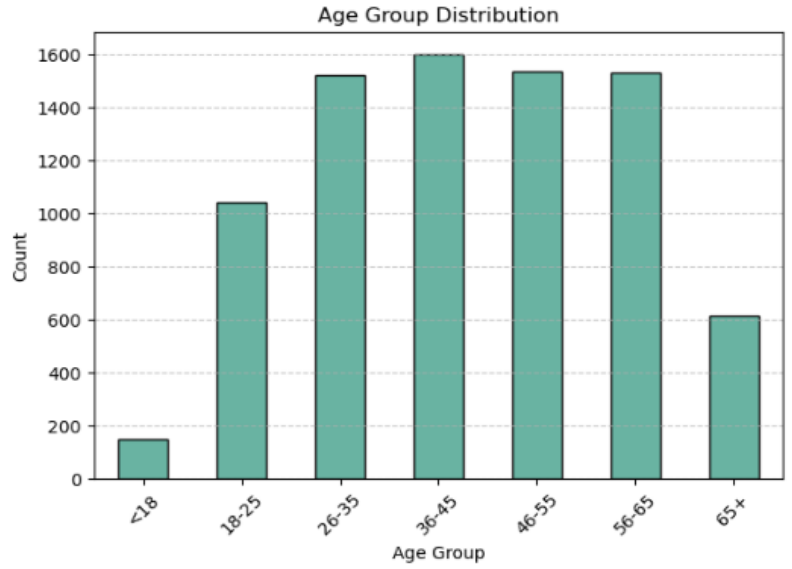
ROWS=8000

Columns=20

Variable	Description
CustomerID	Unique identifier for each customer
Age	Age of the customer
Gender	Gender of the customer (Male/Female)
Income	Annual income of the customer
CampaignChannel	Marketing channel used for the campaign (e.g., Social Media, Email, PPC, SEO, Referral)
CampaignType	Stage of marketing campaign (Awareness, Retention, Conversion, Consideration)
AdSpend	Advertising spend on the customer (in monetary units)
ClickThroughRate	Ratio of users who clicked on a marketing advertisement
ConversionRate	Ratio of users who completed the desired action (purchase, signup, etc.)
WebsiteVisits	Number of times the customer visited the website
PagesPerVisit	Average number of pages viewed per website visit
TimeOnSite	Average time (in minutes) spent on the website per visit
SocialShares	Number of times the customer shared content on social media
EmailOpens	Number of marketing emails opened by the customer
EmailClicks	Number of times the customer clicked inside marketing emails
PreviousPurchases	Number of previous purchases made by the customer
LoyaltyPoints	Loyalty points accumulated by the customer
AdvertisingPlatform	Platform used for advertising (e.g., IsConfid)
AdvertisingTool	Tool used within the advertising platform (e.g., ToolConfid)
Conversion	Whether the customer converted (0 = No, 1 = Yes)

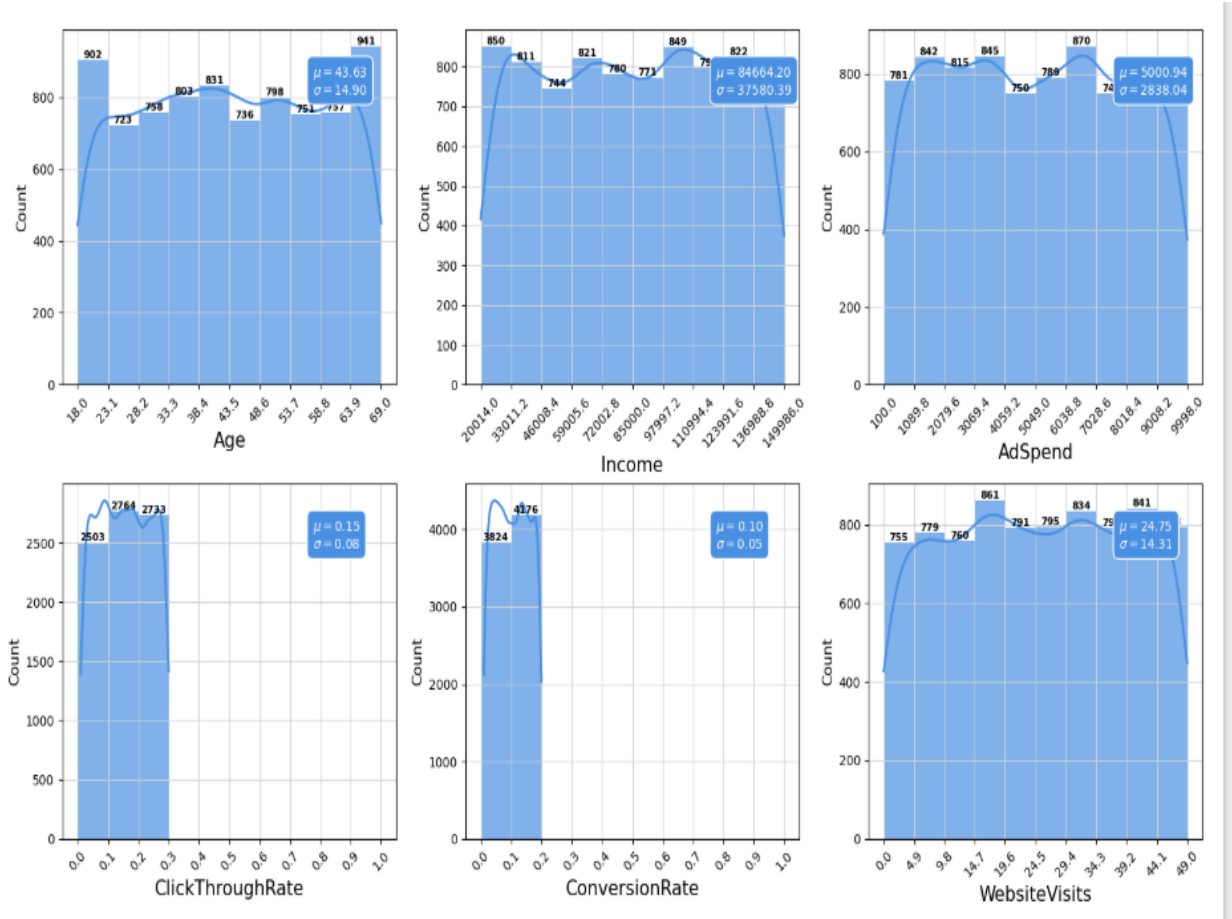
EXPLORATORY DATA ANALYSIS

- Exploratory Data Analysis (EDA) helped us understand the data structure, find patterns, identify trends, and gain valuable insights from the dataset.
- From EDA we analyze, the distribution of each features, checking the correlation between the features .
- The datasets is clean, as there is no **NULL** no **DUPLICATE** VALUES and no **OUTLIERS.**
- We used the univariate and bivariate analysis approach to gain insights into individual characteristics of the data and likewise how each feature relates to main goal: **predicting the target variable.**



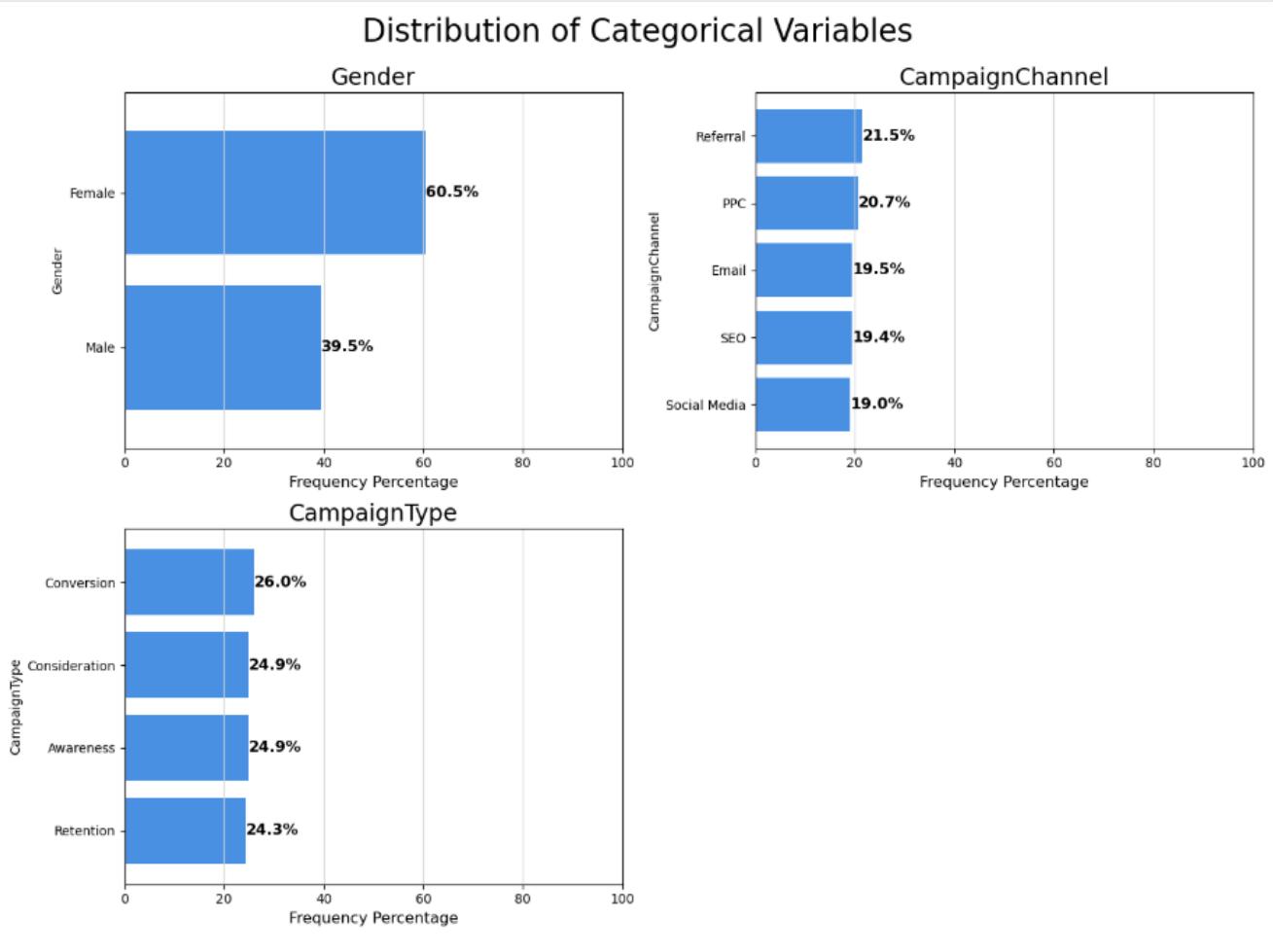
DISTRIBUTION OF CONTINUOUS VARIABLE

Age: Fairly even distribution, mean ~43.6 years.
Income: Broad spread, high variability, mean ~84,664.
Ad Spend: Evenly distributed, mean ~5,000.
Click Through Rate: Mostly between 0.1–0.3, mean ~0.15.
Conversion Rate: Concentrated between 0.05–0.2, mean ~0.10.
Website Visits: Slight peak at 20–30, mean ~24.75.



After examining the histograms of the continuous features and comparing them with the provided descriptions, the data seems consistent and falls within expected limits. **No significant noise or unrealistic values were observed among the continuous variables.

DISTRIBUTION OF CATEGORICAL VARIABLE



- **Gender**: The majority of users are female (60.5%), while males make up 39.5%.
- **Campaign Channel**: Referral is the most used channel (21.5%), followed by PPC (20.7%), Email (19.5%), SEO (19.4%), and Social Media (19.0%).
- **Campaign Type**: Campaign types are evenly distributed, with Conversion campaigns slightly leading (26.0%), followed by Consideration and Awareness (both 24.9%), and Retention (24.3%).

Income Distribution by Gender

It presents the income distribution for males and females:

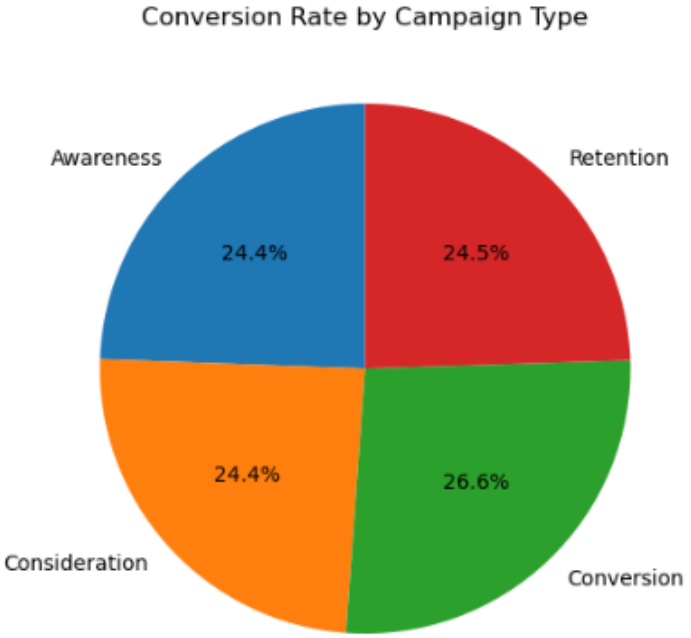
- **Females:** The median income is slightly higher than that of males. The interquartile range (IQR) is also larger, indicating more variability in female incomes. The whiskers suggest that the income range is broader for females.
- **Males:** The median income is lower compared to females, with a smaller IQR, indicating less variability. The whiskers are shorter, suggesting a narrower income range.

Conversion Rate by Campaign Type

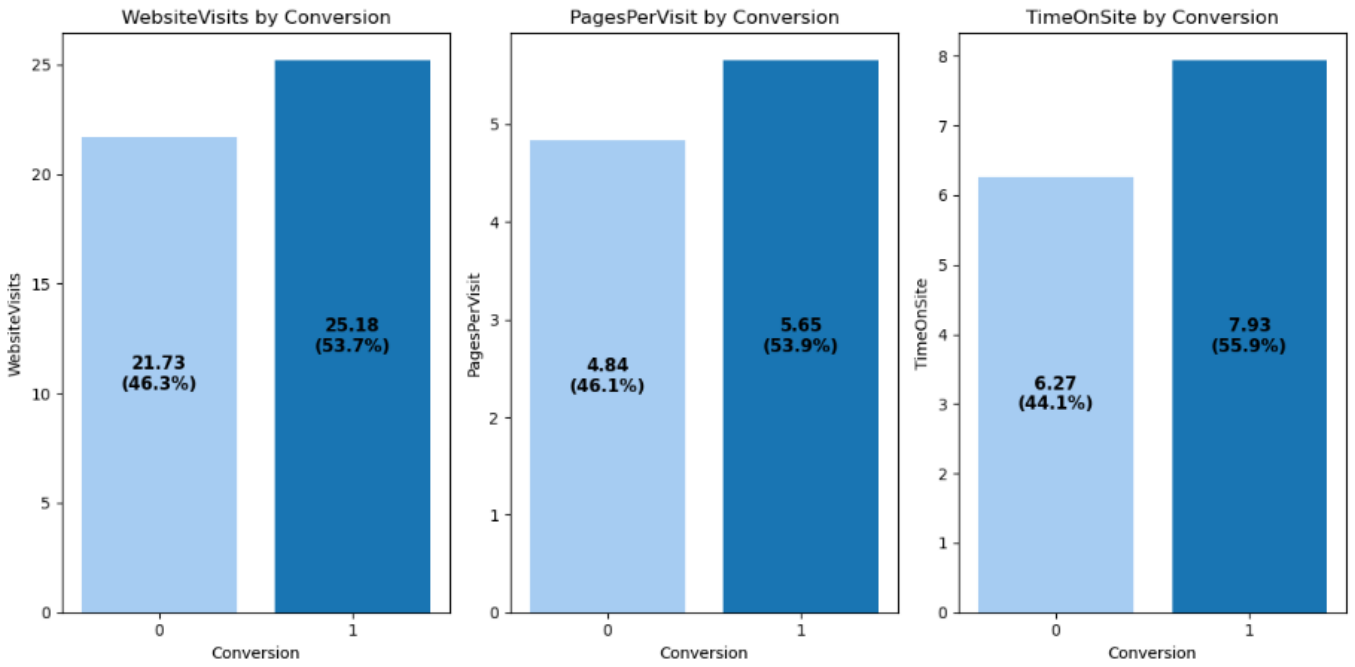
It shows the distribution of conversion rates across four types of marketing campaigns:

- Conversion: **26.6%**
- Consideration: 24.4%
- Awareness: 24.4%
- Retention: 24.5%

This chart indicates that the "Conversion" campaign type has the highest conversion rate, slightly more than the other three types, which are nearly equal.

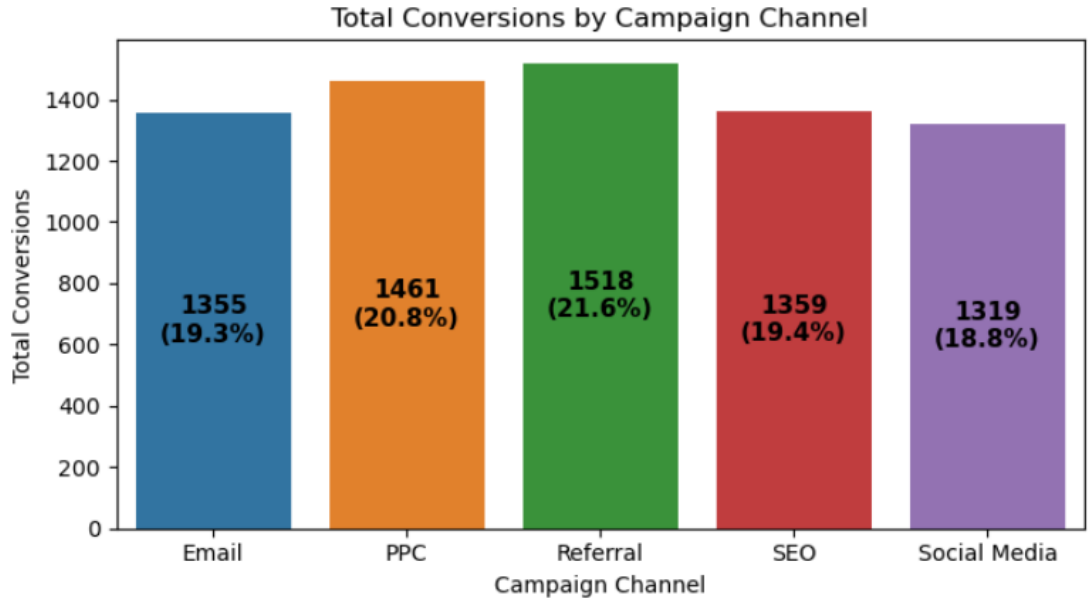


Website Engagement Metrics



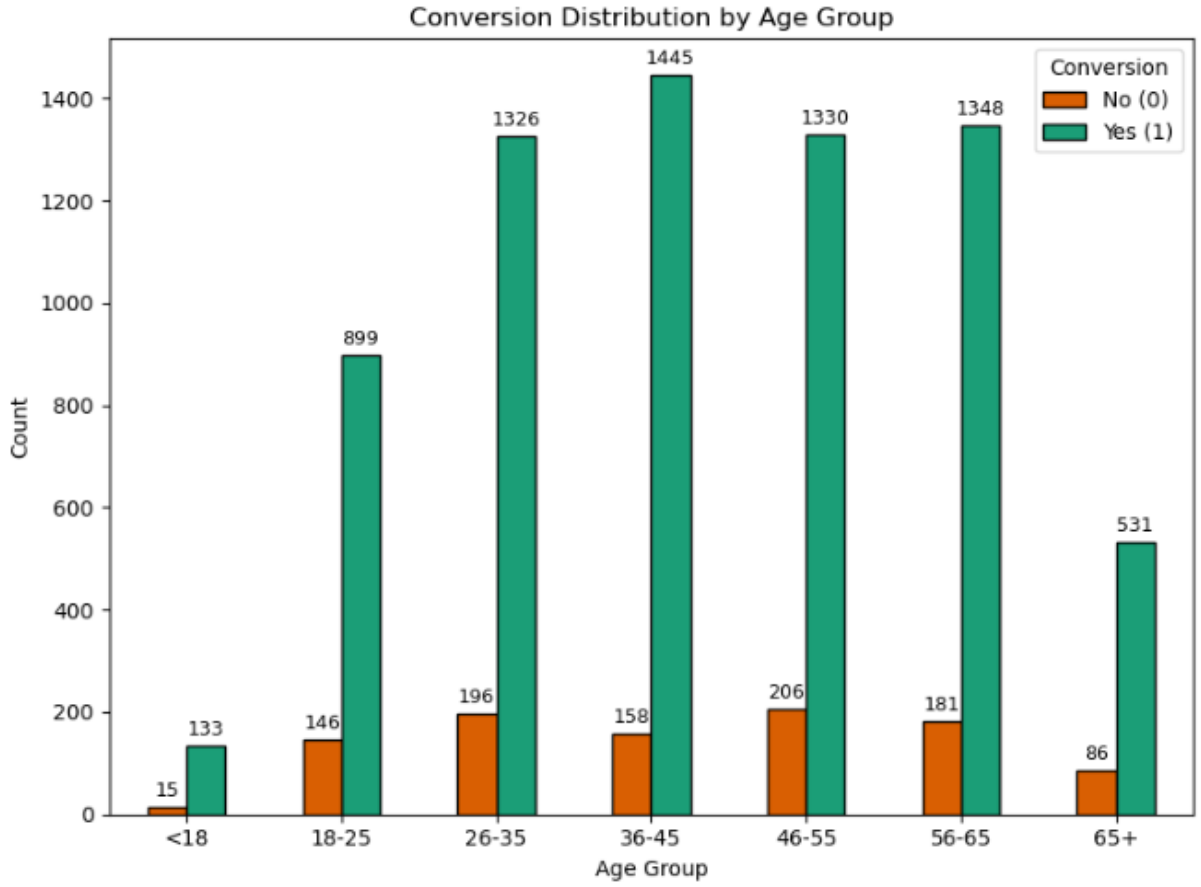
- **WebsiteVisits by Conversion:** Users who converted visited the website **more frequently** on average, indicating that higher engagement correlates with a greater likelihood of conversion.
- **PagesPerVisit by ConversionConverting:** users tend to explore **more content per visit,** suggesting that deeper interest or better navigation leads to conversion.
- **TimeOnSite by Conversion:** Converting users spend **more time** on the site, reinforcing the idea that more engaged users are more likely to convert.

Conversion by Campaign Channel



- Referral campaigns resulted in the highest number of conversions.
- PPC,SEO and Email campaigns followed closely in conversion performance.
- Social Media had the lowest conversion counts among the channels.

Conversion Distribution by Age Group



- The 36–45 age group had the highest number of conversions.
- Age groups 26–65 all showed strong conversion counts, especially 26–35, 46–55, and 56–65.
- Users under 18 had the lowest number of conversions.
- Conversion counts drop significantly in the 65+ age group compared to middle-aged groups.

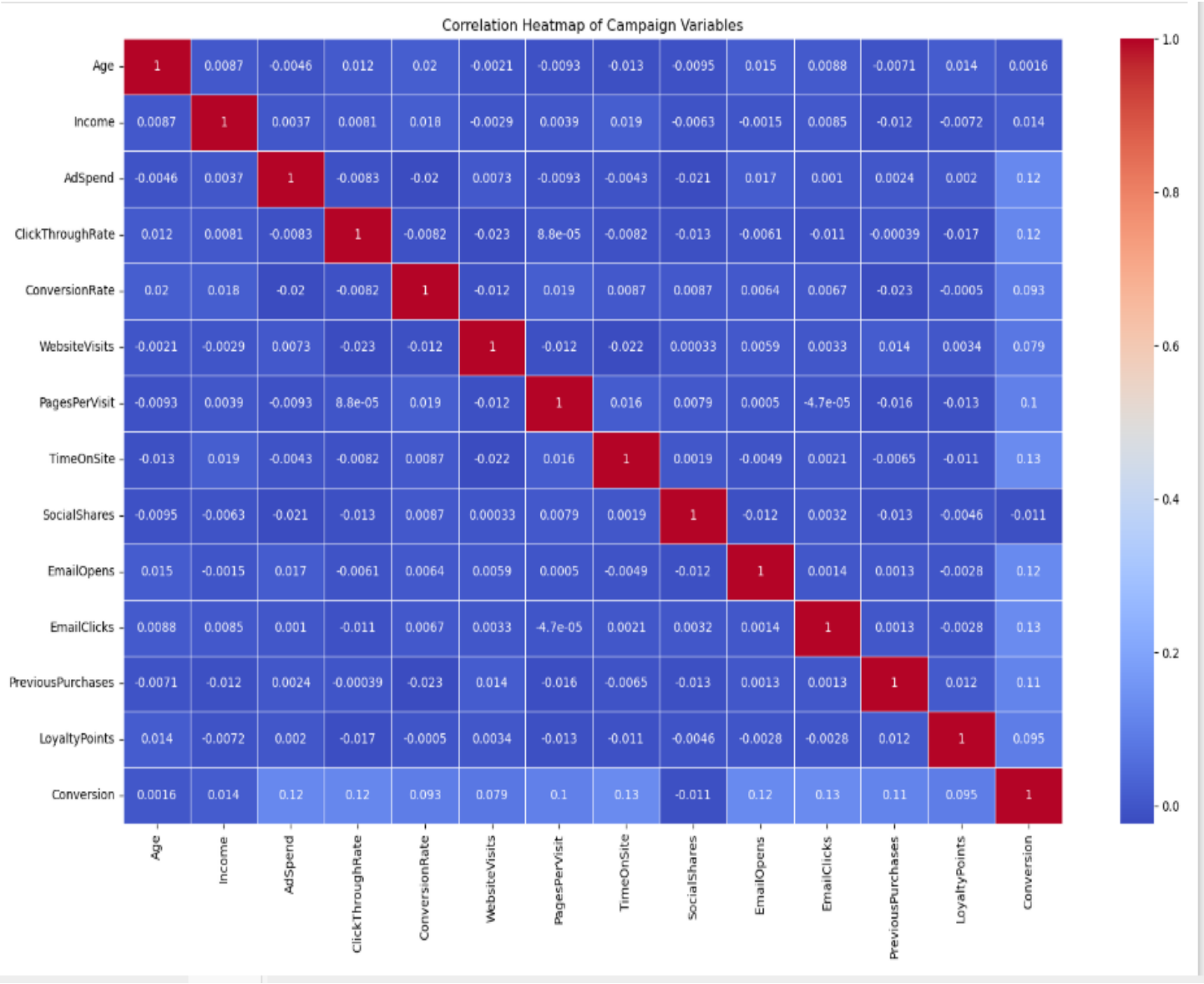
INSIGHTS FROM THE CORRELATION HEATMAP

- **Click Through Rate:** Shows a moderate positive correlation with conversion, suggesting that higher click-through rates are associated with more conversions.
- **Pages Per Visit and Time On Site:** Both have moderate positive correlations with conversion, indicating that more page views and longer site visits are linked to higher conversion rates.
- **Ad Spend:** Also positively correlated,

  suggesting that increased spending can lead

  to more conversions.

- **Email Opens and Email Clicks:** Show positive correlations with conversion, highlighting the importance of effective email campaigns.
- **Social Shares:** Has a slight positive correlation, indicating some impact on conversion.



PREPROCESSING

- **Irrelevant Feature Removal:** `CustomerID`, `AdvertisingPlatform`, and `AdvertisingTool` are removed.

```
# Drop the unnecessary columns
df = df.drop(['CustomerID', 'AdvertisingPlatform', 'AdvertisingTool'], axis=1)
```

- **Duplicates Value:** No Duplicates found in Data.

```
Number of duplicate rows: 0
```

- **Missing Value Treatment:** **No missing value** found in the dataset.

```
# Check for missing values in the dataset
df.isnull().sum().sum()
```

```
0
```

- **Outliers Treatment:** Checked outliers using **IQR** method for the continuous features and upon identifying outliers, nature of algorithm, and given small dataset size direct removal of outliers might not be best approach. No outlier found in our data.

- **Categorical Feature Encoding:** Applied **one hot encoding** to the columns like "Gender", "Campaign Type" and "Campaign Channel" since these variables are nominal variables.

```
# Perform one-hot encoding and convert boolean to integer (0/1)
df = pd.get_dummies(df, columns=['Gender','CampaignChannel', 'CampaignType'], drop_first=False)
df[df.select_dtypes(include=['bool']).columns] = df.select_dtypes(include=['bool']).astype(int)
```

- **Feature Scaling:** Standard scaling done after train and test split on numerical columns.

```
# Initialize scaler
scaler = StandardScaler()

# Fit on training data numerical columns
X_train[numerical_cols] = scaler.fit_transform(X_train[numerical_cols])

# Transform test data numerical columns
X_test[numerical_cols] = scaler.transform(X_test[numerical_cols])
```

SPLITING THE DATA INTO X & Y

- We divided the dataset into two parts: X and y.
- "**X**" typically represents the **independent** Variables, and "**y**" represents the **Dependent** (**target variable**) that we want to predict or understand.

Before Over Sampling

```
: Conversion
1    7012
0     988
```

Over sampling Using SMOTE

- We use SMOTE to oversample because the dataset is imbalanced, with significantly fewer conversion=0 instances than conversion=1.

After Over Sampling

```
from imblearn.over_sampling import SMOTE
```

```
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)

# Check the new class distribution
new_class_distribution = y_resampled.value_counts()
print(new_class_distribution)
```

```
Conversion
1    7012
0    7012
Name: count, dtype: int64
```

TRAIN TEST SPLIT

- We divided the data into training (**80%**) and testing (**20%**) sets.

MODEL SELECTION

❖ Models used:

- **Logistic Regression:** logistic Regression is commonly used for **binary classification problems**. it's preferred because it provides a **simple** an **efficient** way to **model** the **relationship** between the **independent variables** and the **probability** of a certain **outcome**.

- **Decision Tree:** Decision Tree algorithms are used for **classification** because they are **simple**, **computationally efficient**, and **effective** in handling **high-dimensional data**.
  Works best for categorical independent columns.

- **Random Forest Algorithm:** Random Forest: Random Forest is a robust supervised algorithm suitable for both regression and classification tasks.

- **Support Vector Machine:** SVM is a **powerful supervised algorithm** that works best on **smaller datasets** but on **complex ones**. Support Vector Machine(**SVM**)can be used for both **regression** and **classification** tasks, but generally, they **work best** in **classification problems**.

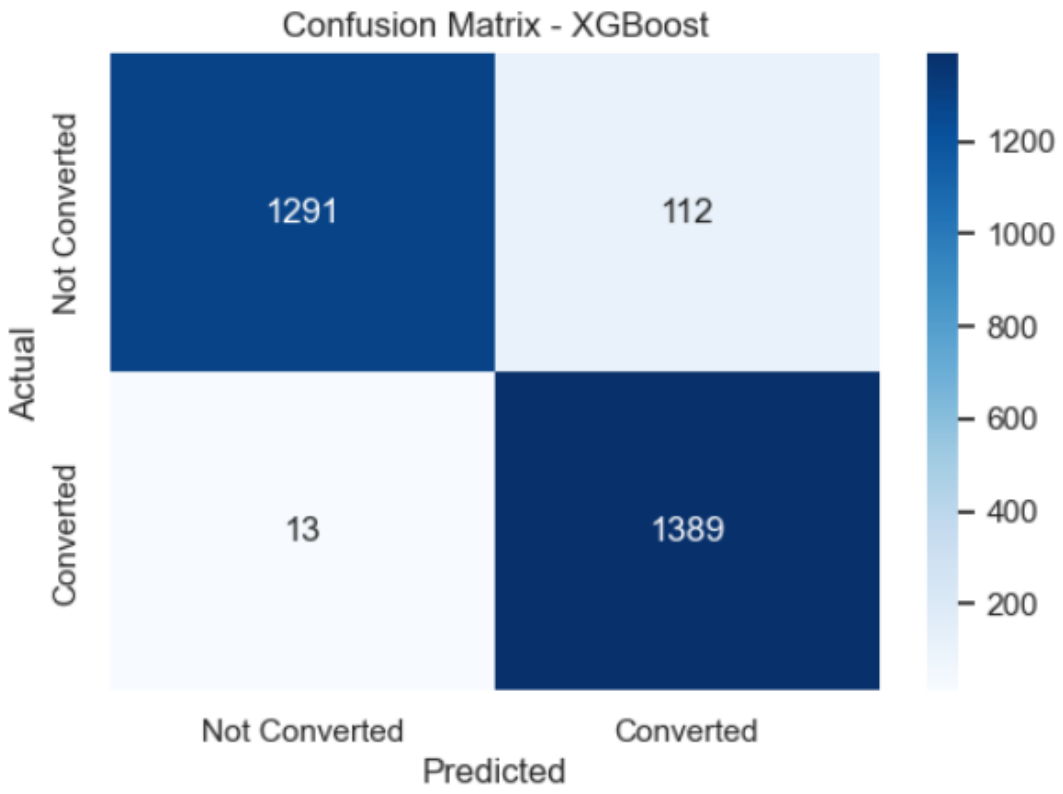
- **XGBoost:** XGBoost (Extreme Gradient Boosting) is a powerful and efficient supervised learning algorithm based on gradient boosting, designed for both classification and regression tasks.

XGBoost Report

```
Model Evaluation Metrics:
Accuracy : 0.9554
Precision: 0.9254
Recall   : 0.9907
F1 Score : 0.9569
```

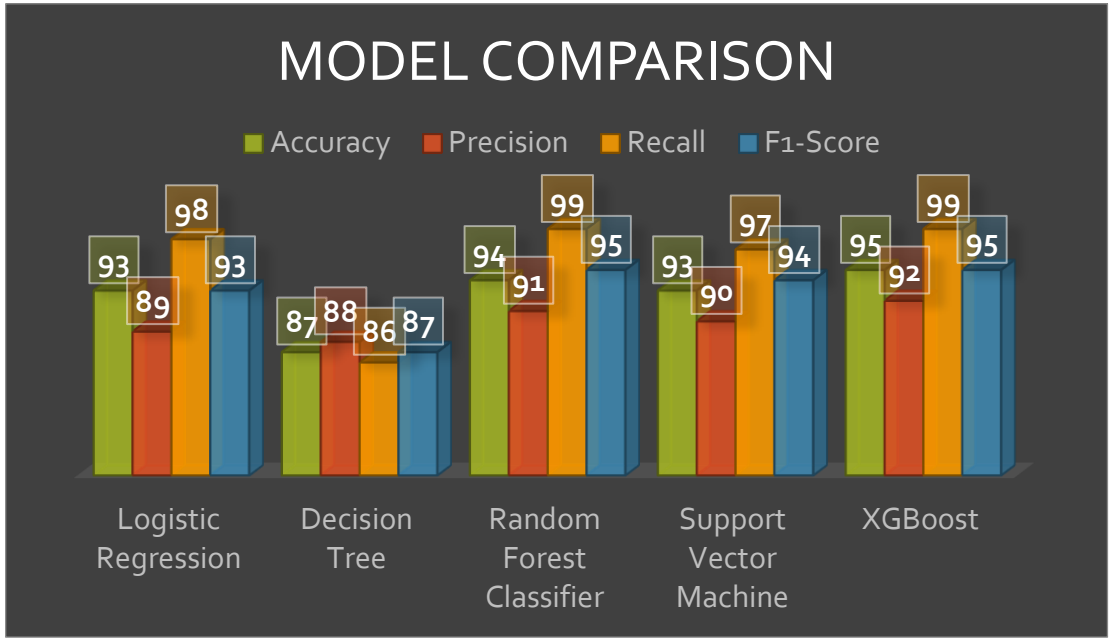
Classification Report:

	precision	recall	f1-score	support
0	0.99	0.92	0.95	1403
1	0.93	0.99	0.96	1402
accuracy			0.96	2805
macro avg	0.96	0.96	0.96	2805
weighted avg	0.96	0.96	0.96	2805



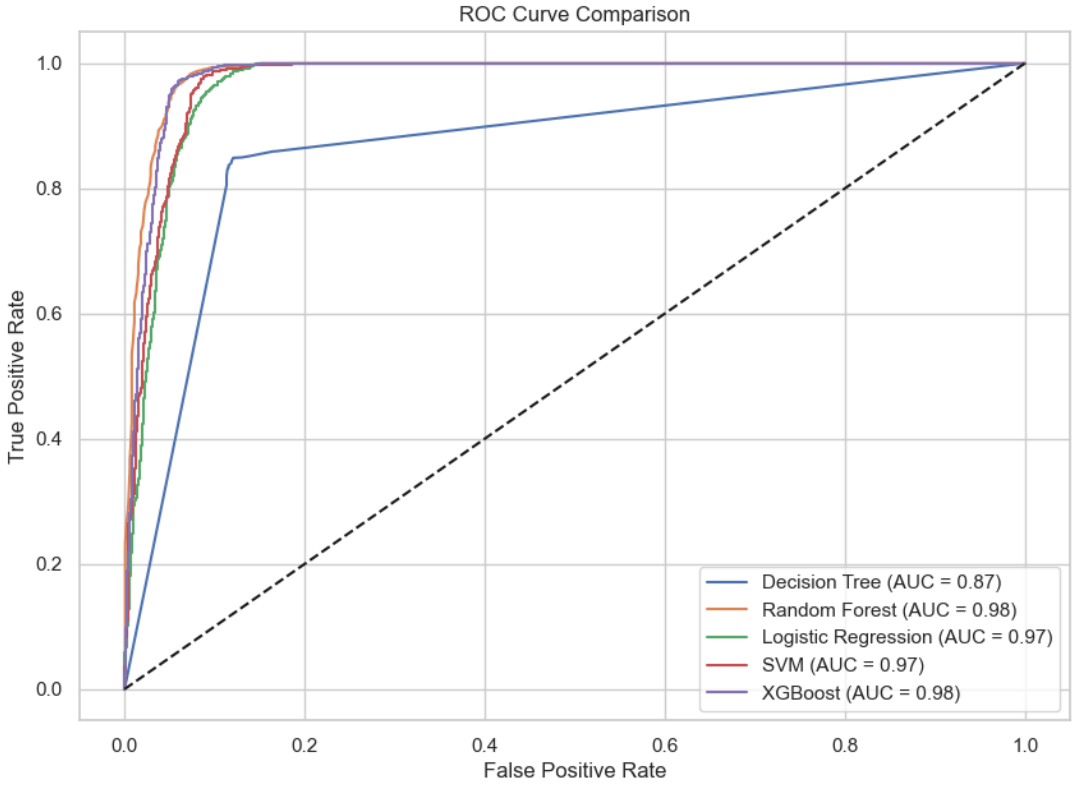
MODEL COMPARISON

MODEL	Accuracy	Precision	Recall	F1-Score
Logistic Regression	93	89	98	93
Decision Tree	87	88	86	87
Random Forest Classifier	94	91	99	95
SVM	93	90	97	94
XGBoost	95	92	99	95



- **XGBoost Leads**: With 95% accuracy, 92% precision, and 99% recall, XGBoost delivers the strongest overall performance, combining excellent predictive power and minimal false negatives.
- **Random Forest Classifier**: Matches XGBoost's 99% recall and scores a 95% F1-score, indicating a powerful balance between precision and recall.
- **Support Vector Machine (SVM)**: Achieves 93% accuracy, 90% precision, and 97% recall, maintaining solid, balanced classification performance.
- **Logistic Regression**: Performs well with 93% accuracy, high recall (98%), but slightly lower precision (89%), making it reliable for identifying positives.
- **Decision Tree**: Trails behind others with 87% accuracy, reflecting lower predictive consistency despite decent precision (88%) and recall (86%).

AUC-ROC CURVE FOR ALL MODELS



XGBoost and Random Forest lead with the highest AUC of 0.98, indicating superior classification capability and excellent distinction between classes.

SVM and Logistic Regression both show strong performance with an AUC of 0.97, closely trailing the top models.

Decision Tree lags with an AUC of 0.87, reflecting comparatively weaker model performance and lower discriminative power.

Conclusion: XGBoost and Random Forest are the most reliable choices in terms of ROC-AUC, ideal for maximizing true positives while minimizing false positives.

Final Conclusion

•**Engaged users** (more visits, time, pages) are more likely to convert.
•**Email interactions** (opens, clicks) are strong conversion drivers.
•Best-performing age group: **26–65**, especially **36–45**.
•**High-income users** convert more frequently.
•**Referral and PPC channels** are slightly more effective.
•Surprisingly, **"Conversion" campaigns** perform the worst.

Model Results
•**XGBoost** is the best model (95% accuracy, 0.98 AUC).
•**Random Forest** and **SVM** are also strong.
•**Decision Tree** performed the weakest (87% accuracy, 0.87 AUC).

Recommendations

•Focus on **high-income, aged 26–65** users.
•Invest in **email and referral campaigns**.
•Use **XGBoost** for future predictions.
•Reevaluate **"Conversion" campaign strategy**.

DIGITAL MARKETING DASHBOARD

CampaignChannel: All

Gender: All

CampaignType: All

Total Ad Spent	Total Website Visits	Avg. Time on Site	Avg. Click Rate	Avg. Conversion Rate
4,00,07,559	198013	7.728	0.1548	0.1044

Gender	Average of AdSpend	Average of Age	Average of Income
Female	5,005.08	43.75	85922.60
Male	4,994.61	43.43	82737.77

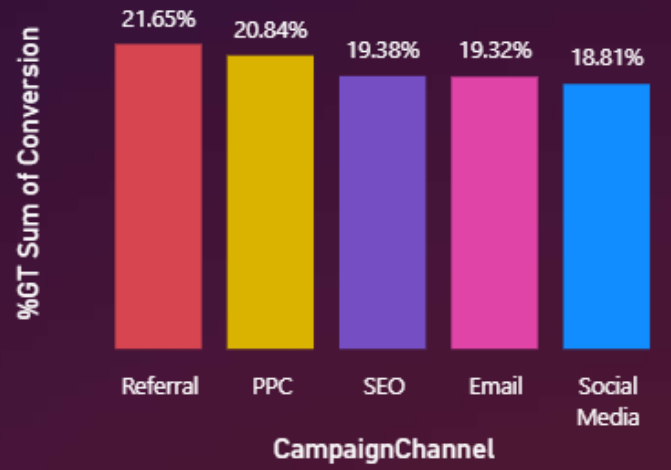
Sum of TimeOnSite	Sum of WebsiteVisits
61,821.75	198013

Average of SocialShares	Average of PagesPerVisit
49.80	5.55

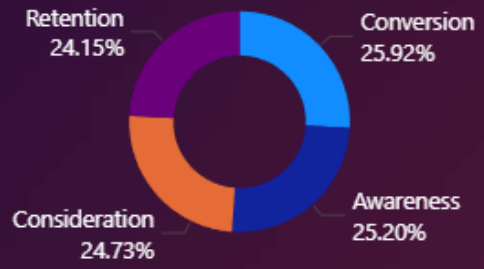
Income (bins)	10	20	30	40	50	60
20000	2733.80	2514.16	2526.27	2515.47	2391.42	2522.14
40000	2080.05	2404.15	2638.32	2403.89	2619.31	2613.30
60000	2786.27	2552.93	2459.79	2370.05	2487.69	2524.23
80000	2377.85	2331.96	2339.17	2428.62	2549.27	2528.99
100000	2524.92	2571.14	2462.63	2434.06	2526.05	2673.47
120000	2547.20	2426.94	2500.90	2544.95	2436.55	2394.01
140000	2547.71	2425.66	2424.77	2486.35	2390.72	2595.96

AdSpend By Campaign Type

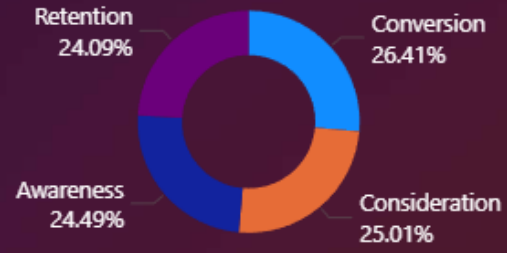
% of Conversion by CampaignChannel



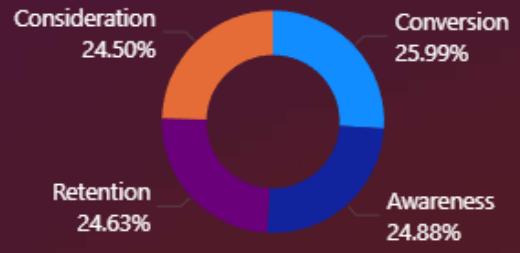
Social Engagement



Email Clicks



Email Opens



INSIGHTS

Overall Performance Metrics

- **Total Ad Spend**: ₹4,00,07,559
- **Total Website Visits**: 198,013
- **Average Time on Site**: 7.728 minutes
- **Average Click Rate**: 15.48%
- **Average Conversion Rate**: 10.44%

Campaign Effectiveness

- **Conversions by Channel**:
  - **Referral** (21.65%) leads, followed by **PPC** (20.84%), then **Email**, **SEO**, and **Social Media**.
- **Ad Spend by Campaign Type**:
  - **Referral** has the highest spend (~₹8.7M), followed closely by PPC and Email.
  - **Social Media** has the lowest ad spend (~₹7.5M).

Demographic Performance

- **Gender-wise Spend & Engagement**:
  - **Females** spend slightly more on average (₹5,005) than males (₹4,995).
  - Female users have slightly higher average income (₹85,922) vs. males (₹82,737).
- **Engagement Metrics**:
  - **Average Pages per Visit**: 5.55
  - **Average Social Shares**: 49.80

Engagement by Campaign Type

- **Social Engagement**:
  - Balanced distribution, with **Conversion (25.92%)** slightly leading.
- **Email Clicks & Opens**:
  - Email campaigns show high engagement:
    - Clicks are evenly distributed (~24-26% for Conversion, Retention, Consideration).
    - Opens are highest for **Conversion (25.99%)** and **Awareness (24.88%)**.

Conclusion

Referral and PPC channels are the most effective for conversions and receive the highest ad spend, indicating good ROI.

Conversion campaigns, despite having the highest email opens and engagement, may benefit from further optimization to improve conversion rate.

Female users show slightly better engagement and spending patterns, suggesting a potential segment for focused targeting.

Email marketing performs well and remains a strong conversion channel.

Investments in Social Media show lower conversions and may require reevaluation or retargeting.

Recommendation: Continue prioritizing Referral and PPC strategies, optimize Conversion campaigns, and leverage email insights for segmented targeting.

THANK YOU

GitHub Link