DESCRIPTION OF CLEANED DATA

```
df.info()
```

✓ 0.3s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 11 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   CustomerID           250 non-null    object
 1   Date                 250 non-null    object
 2   SubscriptionPlan     250 non-null    object
 3   AgeGroup             250 non-null    object
 4   Login_x              250 non-null    int64 
 5   WatchTimeMinutes_x   250 non-null    int64 
 6   ContentPreference    250 non-null    object
 7   Churned              250 non-null    bool  
 8   ChurnDate            45 non-null     object
 9   TotalLogin           250 non-null    int64 
 10  TotalWatchTime       250 non-null    int64 
dtypes: bool(1), int64(4), object(6)
memory usage: 19.9+ KB
```

```
df.describe()
```

✓ 0.1s Python

	Login_x	WatchTimeMinutes_x	TotalLogin	TotalWatchTime
count	250.000000	250.000000	250.000000	250.000000
mean	0.636000	44.748000	51.708000	3702.952000
std	0.482114	47.974522	14.148337	2300.822414
min	0.000000	0.000000	10.000000	340.000000
25%	0.000000	0.000000	45.000000	1608.250000
50%	1.000000	35.000000	55.000000	3934.500000
75%	1.000000	77.500000	63.000000	5065.750000
max	1.000000	180.000000	71.000000	8716.000000

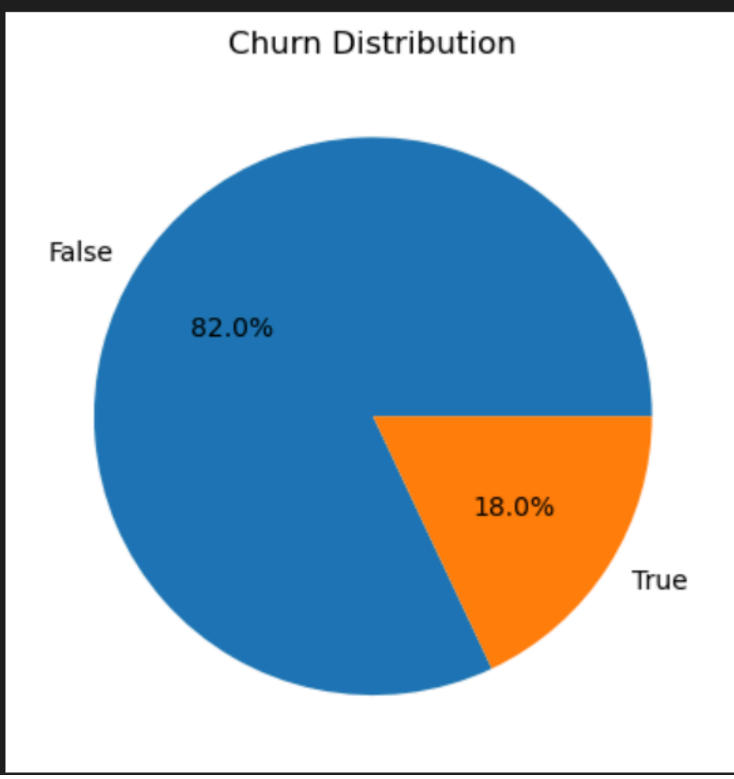
```python
df.nunique()
```

✓ 0.0s Python

```
CustomerID              250
Date                     37
SubscriptionPlan          3
AgeGroup                  4
Login_x                   2
WatchTimeMinutes_x       92
ContentPreference         4
Churned                   2
ChurnDate                36
TotalLogin               54
TotalWatchTime          242
dtype: int64
```

```python
df['Churned'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title("Churn Distribution")
plt.ylabel("")
plt.show()
```

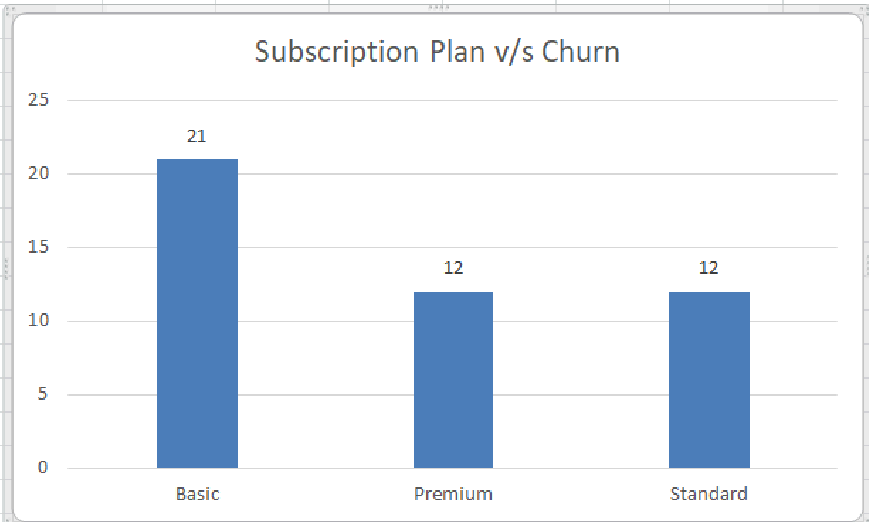
✓ 0.2s Python



UNIVARIATE ANALYSIS

Churned	TRUE

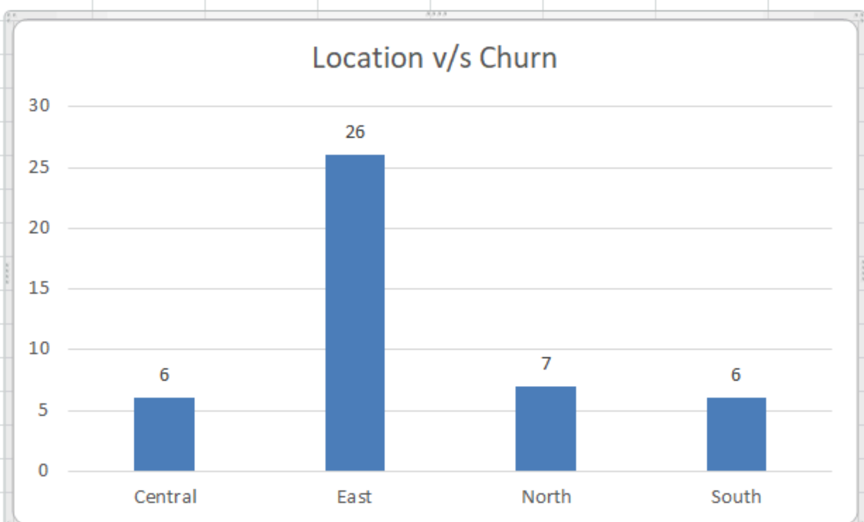
Row Labels	Count of Churned
Basic	21
Premium	12
Standard	12
Grand Total	**45**



Customers on the Basic plan show the highest churn (21) compared to Premium and Standard plans (12 each).
This indicates that users with lower-tier subscriptions are more likely to discontinue the service.

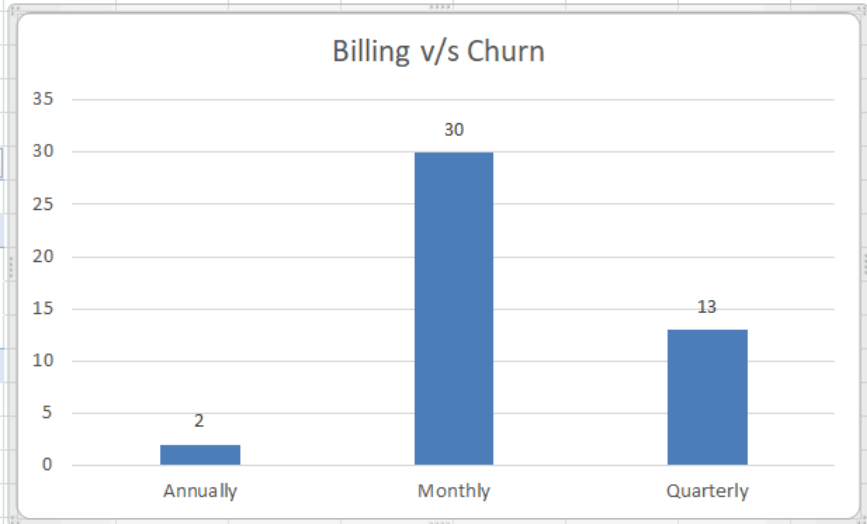
Churned	TRUE

Row Labels	Count of Churned
Central	6
East	26
North	7
South	6
Grand Total	**45**



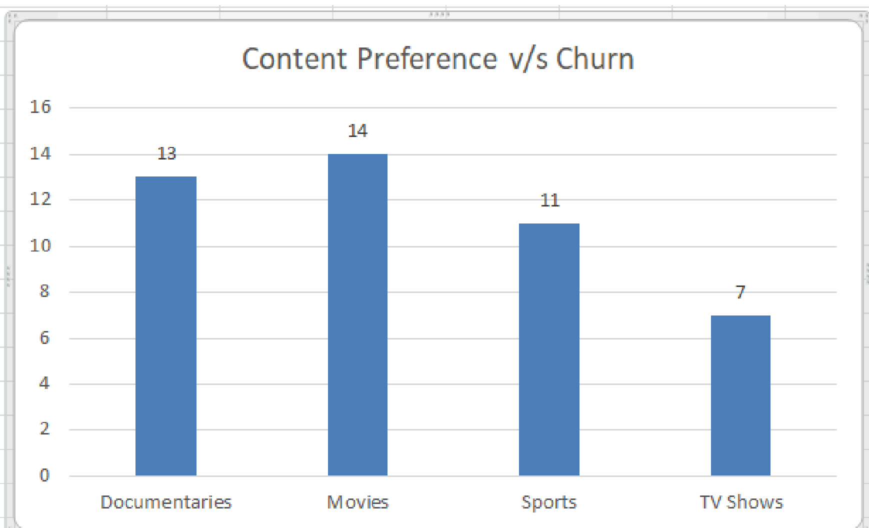
The East region records the highest churn (26 customers), which is significantly larger than other regions.
This suggests that geographic location may play a key role in customer retention, with the East requiring focused interventions.

Churned	TRUE
Row Labels	**Count of Churned**
Annually	2
Monthly	30
Quarterly	13
Grand Total	**45**



The monthly billing cycle shows the highest churn (30 customers), far exceeding quarterly (13) and annual (2).
This indicates that shorter billing cycles may increase churn risk, as customers find it easier to discontinue.

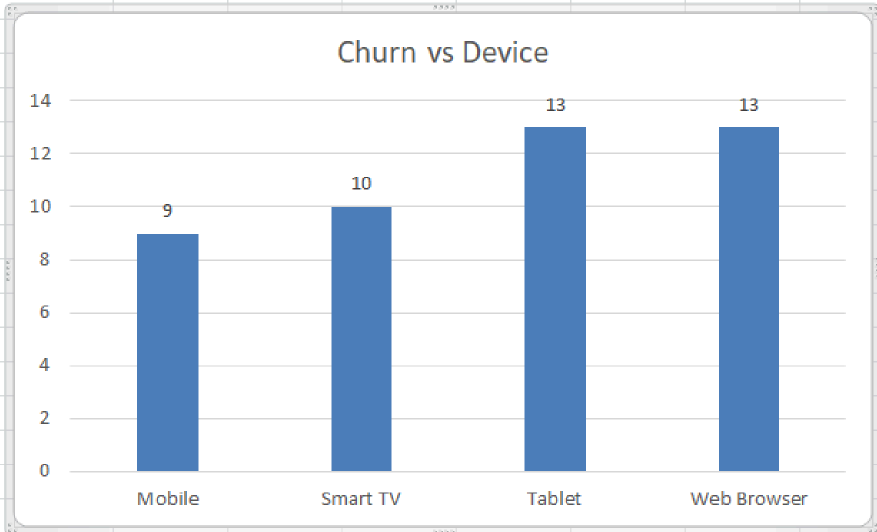
Churned	TRUE
Row Labels	**Count of Churned**
Documentaries	13
Movies	14
Sports	11
TV Shows	7
Grand Total	**45**



Customers preferring Movies (14) and Documentaries (13) have the highest churn, followed by Sports (11).
TV Shows have the lowest churn (7), indicating stronger retention among this group.

Churned	TRUE

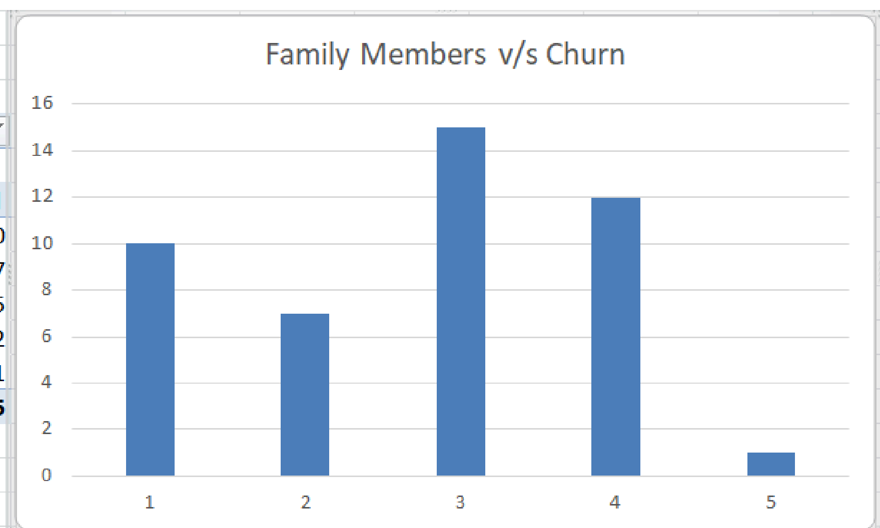
Row Labels	Count of Churned
Mobile	9
Smart TV	10
Tablet	13
Web Browser	13
Grand Total	**45**



Churn is highest among users accessing via Tablet (13) and Web Browser (13). Mobile (9) and Smart TV (10) users show relatively lower churn, suggesting stronger engagement on these devices.

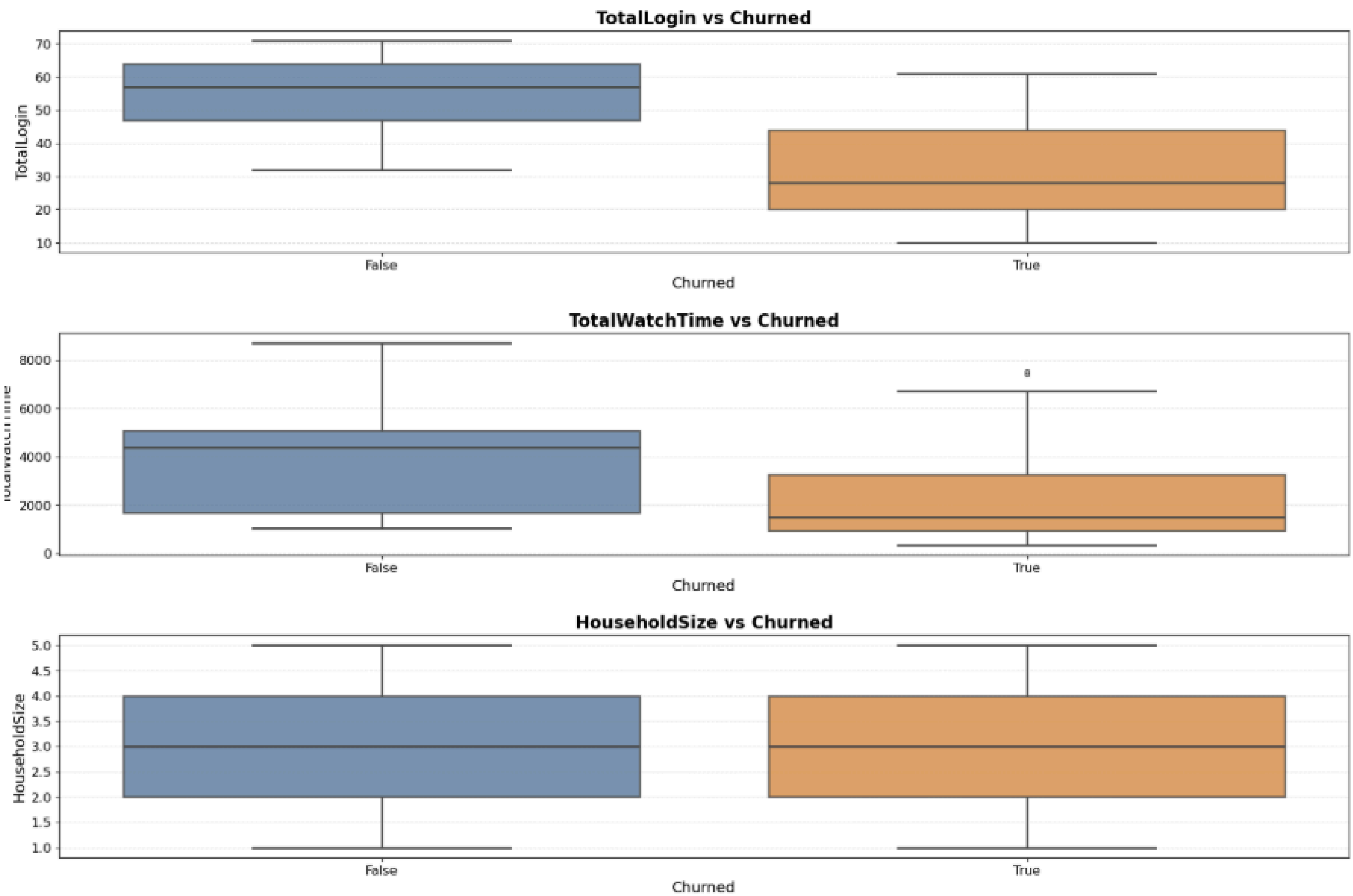
Churned	TRUE

Row Labels	Count of Churned
1	10
2	7
3	15
4	12
5	1
Grand Total	**45**

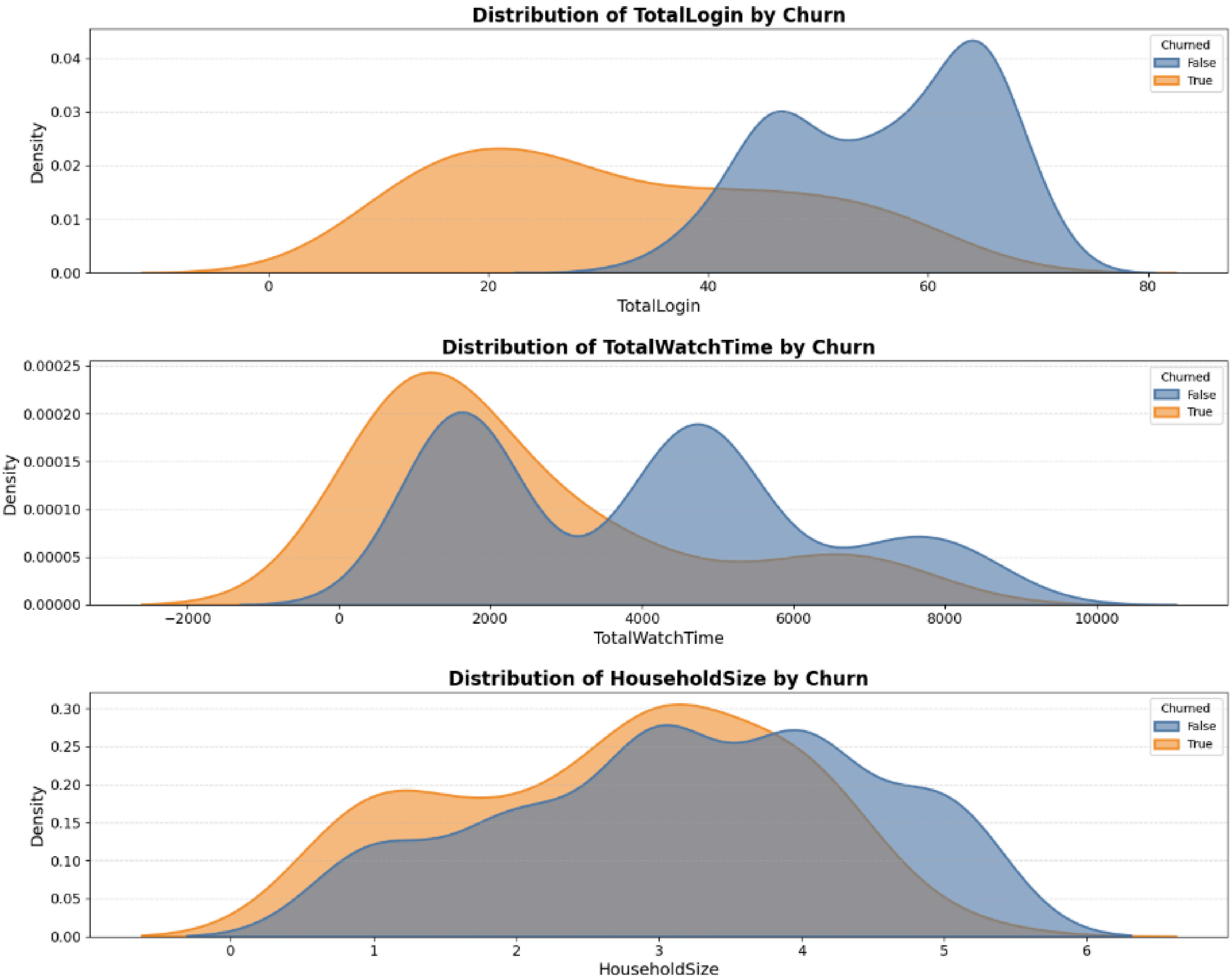


Churn is highest in families with 3 members (15), followed by 4 members (12). Very small (1–2) and very large (5) families show lower churn, suggesting mid-sized families are more prone to leaving.

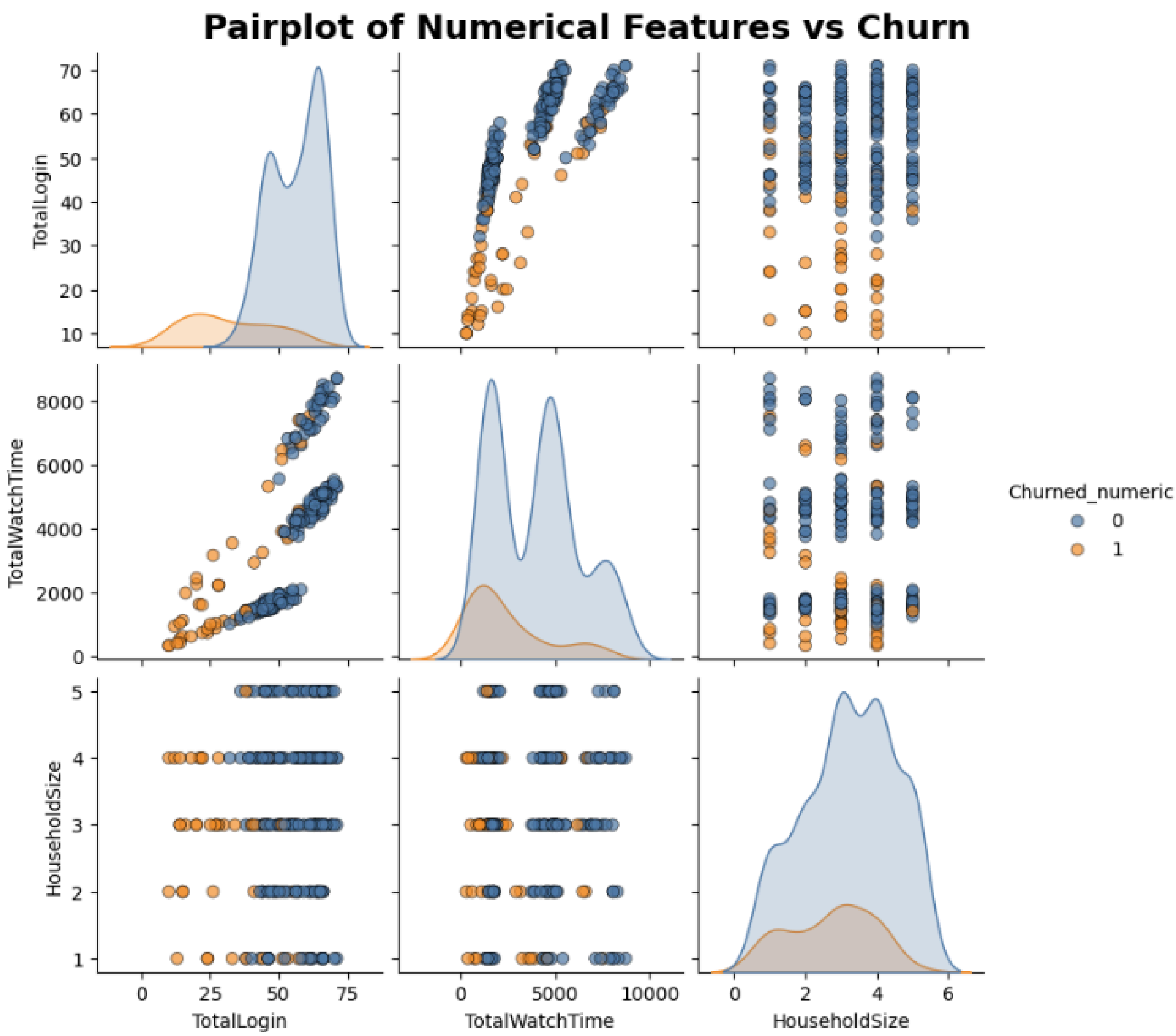
BIVARIATE ANALYSIS



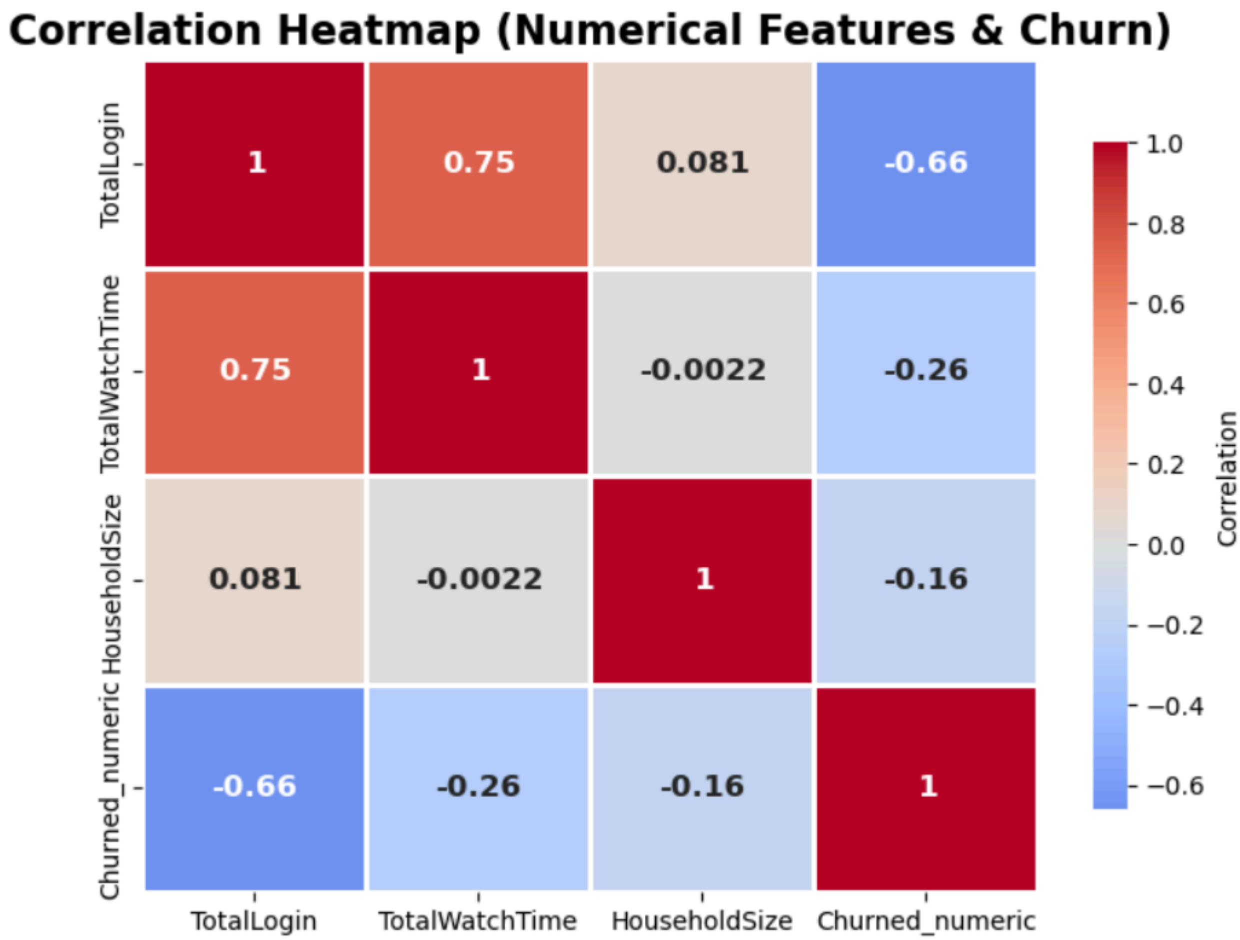
Lower engagement (logins) strongly correlates with churn.
Engagement in watch time is a key predictor of churn.
Household size is less influential than activity metrics.
This means behavioral factors (logins & watch time) are much stronger indicators of churn than demographics (like household size).


Distribution of TotalLogin by Churn
Density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
TotalLogin
Churned
False
True
Distribution of TotalWatchTime by Churn
Density
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
−2000
0
2000
4000
6000
8000
10000
TotalWatchTime
Churned
False
True
Distribution of HouseholdSize by Churn
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
2
3
4
5
6
HouseholdSize
Churned
False
True

MULTIVARIATE ANALYSIS



The correlation heatmap shows that TotalLogin and TotalWatchTime are strongly correlated (0.75), indicating similar usage patterns, while churn is negatively correlated with TotalLogin (-0.66), suggesting higher engagement reduces churn likelihood. HouseholdSize shows little correlation with churn.


Correlation Heatmap (Numerical Features & Churn)
TotalLogin
TotalWatchTime
HouseholdSize
Churned_numeric
1
0.75
0.081
-0.66
0.75
1
-0.0022
-0.26
0.081
-0.0022
1
-0.16
-0.66
-0.26
-0.16
1
TotalLogin
TotalWatchTime
HouseholdSize
Churned_numeric
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
Correlation