



# **Smart Subscription Tracker**

## **Group 3**

**Aleeha Chaudhry, Kishor Khatiwada, Sanskriti Basnet**

## Abstract

The Smart Subscription Tracker project aims to streamline subscription management by leveraging data-driven insights into customer behavior. In today's digital landscape, users often manage multiple services simultaneously, making it increasingly important to understand patterns related to subscription retention and churn. This project utilizes machine learning techniques to identify patterns and insights to enhance customer retention and optimize subscription services. The methodology includes data preprocessing, exploratory analysis, data visualization, and the implementation of predictive models such as logistic regression and decision trees. Additionally, clustering techniques are used to segment users based on their behavior and spending habits. The final outcome is a user-friendly, web-based tool that will help determine which subscriptions are likely to be discontinued and which are worth retaining. This solution not only supports better financial decision-making but also promotes long-term subscription optimization.

**Keywords:** subscription churn, machine learning, predictive analytics, customer retention, subscription management

## Introduction

The Smart Subscription Tracker project focuses on analyzing subscription data to predict customer behavior, specifically churn and monthly spending. The primary challenges include handling missing values, detecting and removing outliers, and building reliable predictive models. Our contributions include implementing advanced data preprocessing techniques and machine learning models to uncover valuable insights from the subscription data. Since the Kaggle dataset was over-used and outdated, we enhanced the efficiency of dataset by surveying our family and friends regarding their subscriptions and spending.

## Motivation

The goal of this project is to enable individuals and organizations that provide subscription-based services to better understand consumer behavior and improve retention methods. By forecasting turnover and monthly spending, subscription businesses can better adjust their offers to match client expectations. This promotes renewals, lowers cancellations, and raises total revenue. Given the abundance of subscription options accessible today, it is imperative that providers comprehend

the factors that influence user loyalty and engagement. Long-term subscriber connections are supported by this data, which also results in more intelligent marketing, tailored experiences, and improved pricing structures.

## **Related Works**

Multiple studies have investigated customer churn prediction and spending analysis for subscription-based businesses. Verbeke et al. (2012), for example, used data mining tools to study churn prediction in the telecom industry, highlighting the significance of profit-centric evaluation metrics. Moreover, Huang et al. (2024) created a neural network-based predictive decision model for customer retention and tested it on real-world datasets to study consumer behavior and enhance retention methods. In the context of streaming services, a study was conducted to forecast customer churn using recurrent neural networks (RNNs), specifically long short-term memory (LSTM) models, to record sequential user activity and increase churn prediction accuracy.

Motivated by these efforts, our project aims to assist consumers in making the most of their subscription spending by detecting unused services through user-provided data and usage trends.

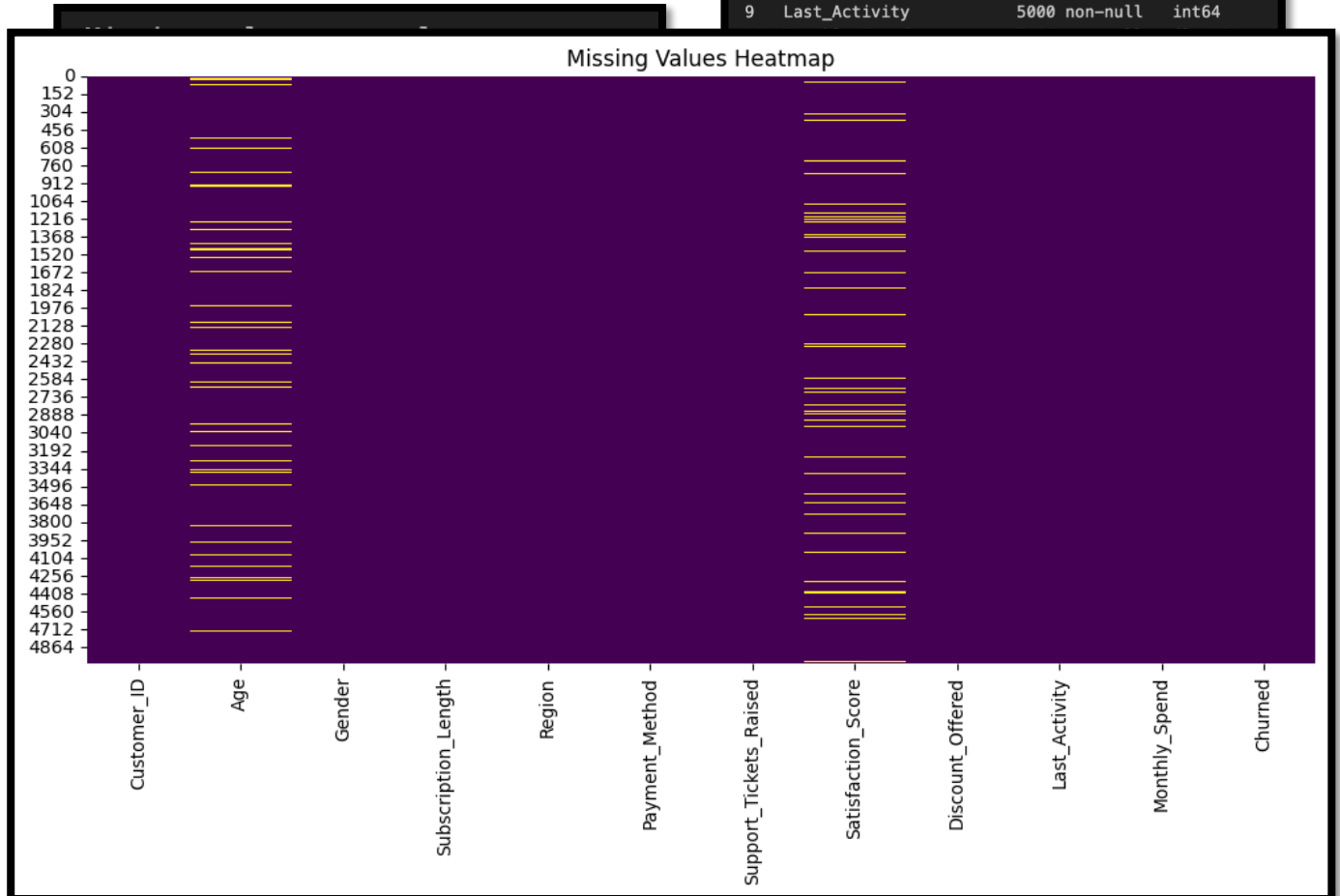
## **Methodology & Results**

Our methodology involves several steps, including data loading and preprocessing, data analysis and visualization, and machine learning model implementation.

### Step 1: Data Loading & Preprocessing

- a. **Missing Values:** As shown below, the data before cleaning consisted of various null/missing values. Therefore, our first step was to identify these values and process the data accordingly.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   Customer_ID                          5000 non-null   object
1   Age                                  4500 non-null   float64
2   Gender                              5000 non-null   object
3   Subscription_Length                 5000 non-null   int64
4   Region                             5000 non-null   object
5   Payment_Method                     5000 non-null   object
6   Support_Tickets_Raised              5000 non-null   int64
7   Satisfaction_Score                 4500 non-null   float64
8   Discount_Offered                   5000 non-null   float64
9   Last_Activity                      5000 non-null   int64
```

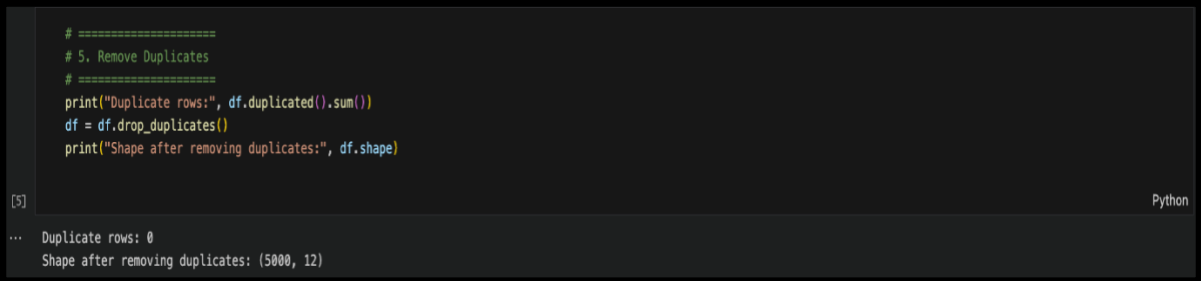


To visualize the missing values, we utilized Python libraries and created the following heatmap.

The heatmap shows that the Age and Satisfaction\_Score columns have missing values, but the other columns seem to be complete. Finding these gaps is an essential first step before developing

a model or doing data analysis. We have decided to keep the rows with missing values for this project. This decision is based on the rationale that even if one column (e.g., Age or Satisfaction\_Score) is incomplete, corresponding row may still hold valuable information in other key attributes such as Subscription\_Length, Region, Monthly\_Spend, and Churned. Eliminating these rows totally could result in a large reduction in dataset size, as well as the loss of useful insights or forecast accuracy.

- b. Duplicates:** The next step is to identify and remove any duplicate entries to maintain data integrity and prevent redundancy.

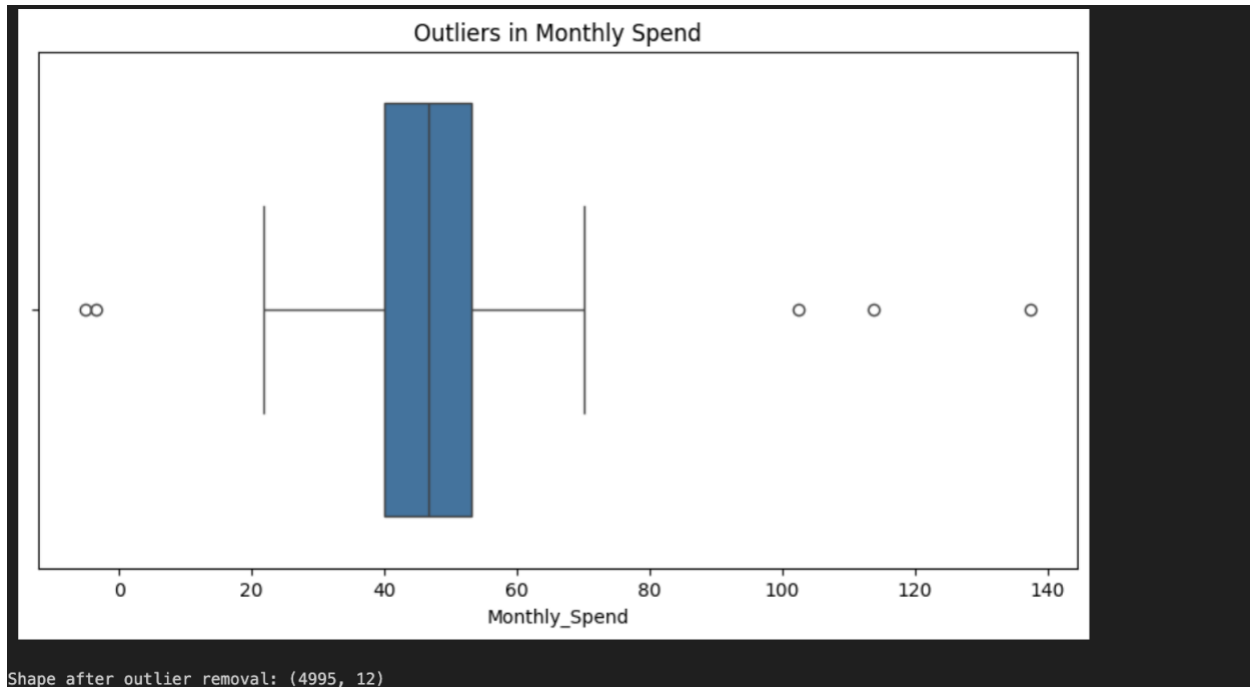


```
# =====  
# 5. Remove Duplicates  
# =====  
print("Duplicate rows:", df.duplicated().sum())  
df = df.drop_duplicates()  
print("Shape after removing duplicates:", df.shape)
```

[5] Python

... Duplicate rows: 0  
Shape after removing duplicates: (5000, 12)

- c. Outliers' Detection & Removal**



The box plot shows the distribution of 'Monthly\_Spend' and identifies outliers. The individual points plotted to the right of the right whisker are the outliers in the 'Monthly\_Spend' data. These points represent values that are significantly higher than the rest of the data.

Outliers were removed from the dataset. This decision was made to improve the performance and reliability of models aimed at predicting subscriptions users are less likely to use. Since low subscription usage is expected to correlate with lower monthly spending, the presence of extreme high spending values can disproportionately skew model training, obscuring the subtle patterns and characteristics associated with low-usage behaviors. Removing these high outliers allows the predictive model to better focus on and learn from the distribution of typical and low spending data points, thereby enhancing its ability to accurately identify users less likely to utilize their subscriptions.

- d. **Saving the Cleaned Dataset:** As the last step of our data preprocessing, we saved the cleaned data to a data folder named `cleaned_data.csv`. Moving forward, this is the file that we will be utilizing in our project.

## **Step 2: Data Analysis & Visualization**

- a. Descriptive Statistics:** These averages provide a quick snapshot of customer behavior. This tells us how much customers typically spend each month, how long they stay subscribed, and how satisfied they are overall. The results give us a foundational understanding of the dataset. The average monthly spend helps us estimate customer value, while the average subscription length gives insight into retention. The satisfaction score serves as an indicator of overall customer experience.

Average Monthly Spend: 46.60

Average Subscription Length: 29.71

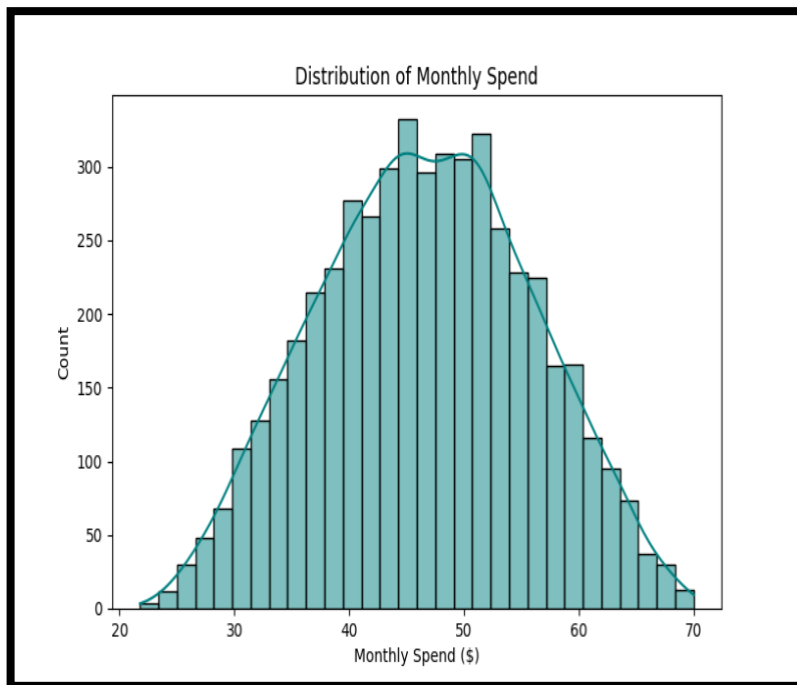
Average Satisfaction Score: 5.55

```
desc_stats = df.describe()
desc_stats.to_csv('../outputs/desc_stats.csv')

print("Average Monthly Spend:", df['Monthly_Spend'].mean())
print("Average Subscription Length:", df['Subscription_Length'].mean())
print("Average Satisfaction Score:", df['Satisfaction_Score'].mean())
```

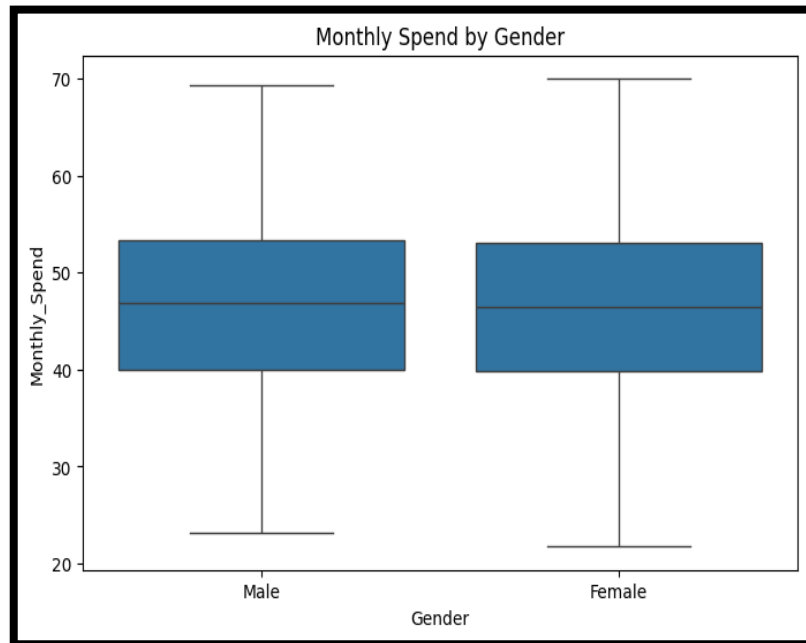
```
Average Monthly Spend: 46.59741541541542
Average Subscription Length: 29.71091091091091
Average Satisfaction Score: 5.546607341490545
```

## b. Spending Patterns:

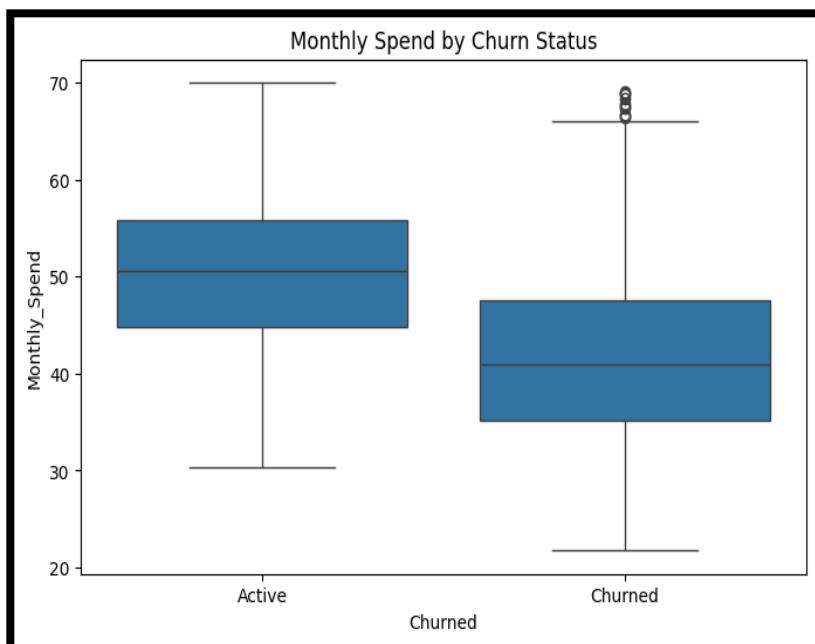


### *Histogram of Monthly Spend:*

The histogram of Monthly Spend appears to be approximately normally distributed. It is roughly symmetrical with a peak around \$50. There is no significant skew to the left or right. The distribution resembles a bell curve, which is characteristic of a normal distribution.



***Spend by Gender:*** The box plot comparing monthly spend between males and females shows remarkable similarity. Both genders exhibit very close median spending values, similar interquartile ranges (indicating comparable spread in the middle 50% of spenders), and similar overall ranges of spending as shown by the whiskers.



***Spend comparison between churned and active users:*** The analysis reveals a distinct difference in monthly spending based on churn status. Customers who churn tend to have lower monthly spending compared to customers who remain active. This suggests that lower monthly spending is associated with a higher propensity to churn, although there are some exceptions as indicated by the high-spending outliers in the churned group.

This finding indicates that monthly spend is a potential factor related to customer churn.

- c. **T-test result:** We performed the T-test as we wanted to observe if there was a significant difference in spending between users who churned and those who stayed active. By



comparing the average spend of the two groups, we could figure out whether churned users behave differently in terms of how much they spend. This kind of insight is useful for understanding if spending habits might be linked to user retention. We used Welch's T-test since the variances between the two groups might not be equal, and it gives us a more reliable result in that case.

```
# T-test
t_stat, p_val = stats.ttest_ind(churned, active, equal_var=False)
print(f"T-test: t = {t_stat:.3f}, p = {p_val:.4f}")

with open('../outputs/t_test_results.txt', 'w') as f:
    f.write(f"T-test: t = {t_stat:.3f}, p = {p_val:.4f}\n")
    if p_val < 0.05:
        f.write("Conclusion: Significant difference in spend between churned vs active users.\n")
    else:
        f.write("Conclusion: No significant difference in spend.\n")
```

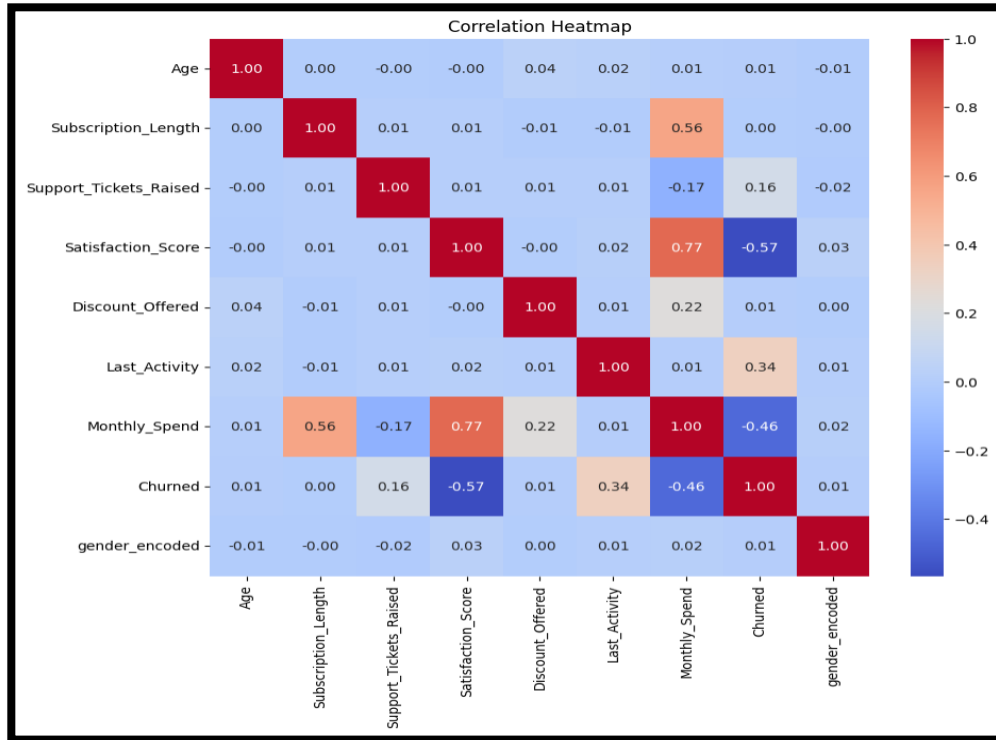
```
T-test: t = -35.730, p = 0.0000
```

Since the p-value is less than 0.05, we reject the null hypothesis.

This means there is a statistically significant difference in monthly spending between churned and non-churned users.

In fact, the large negative t-value indicates that churned users tend to spend significantly less than active users. This supports the idea that low spenders are more likely to churn, which can guide predictive modeling and recommendation strategies later in the project.

#### d. Correlation Analysis



### Key Insights from the Heatmap:

- **Churn vs. Satisfaction Score: Correlation = -0.57**

As satisfaction decreases, likelihood of churn increases significantly. This is the strongest negative correlation with churn crucial for prediction.

- **Churn vs. Monthly Spend: Correlation = -0.46**

Lower monthly spend is associated with higher churn. This supports findings from t-tests.

- **Monthly Spend vs. Satisfaction Score: Correlation = 0.77**

Happier users tend to spend more. Strongest positive correlation.

- **Monthly Spend vs. Subscription Length: Correlation = 0.56**

Long-term subscribers tend to spend more each month.

Overall, most other features (Age, Gender, Support Tickets, etc.) show weak or no correlation with churn or spending.

### Step 3: Machine Learning & Prediction

This step focused on using machine learning techniques to predict customer behavior in relation to subscription services. The two main tasks that our project focused on were:

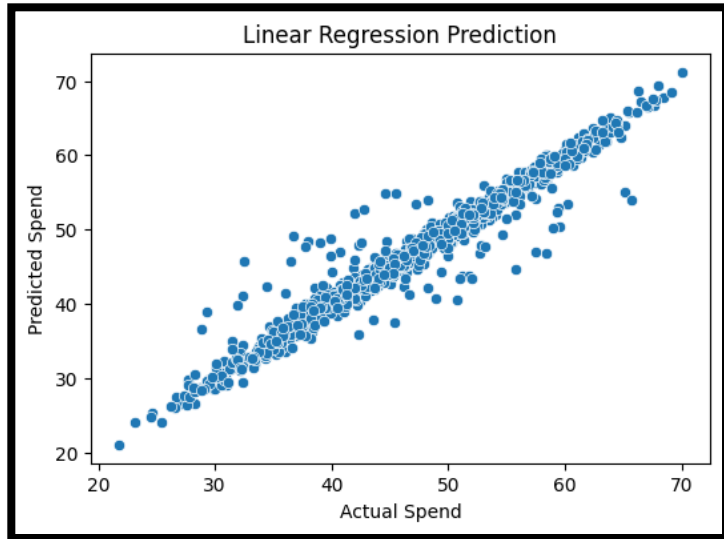
1. **Monthly Spend Prediction (Regression):** Forecast how much a customer is expected to spend on subscriptions.
2. **Churn Prediction (Classification):** Identify whether a customer is likely to cancel their subscription.

### **Part 1: Monthly Spend Prediction (Linear Regression)**

To predict how much a customer is likely to spend monthly on subscriptions, we used a Linear Regression model. We performed this based on key features that were also relevant in our churn analysis, like subscription length, satisfaction score, discount offered, support tickets raised, and encoded gender. These features were chosen because earlier correlation analysis showed that satisfaction and subscription length had strong positive relationships with spending. After imputing missing values using the median strategy, we trained the model and evaluated its performance using RMSE (root mean squared error), which gives us an idea of how accurately our model is predicting spend. The relatively strong relationships between the input features and monthly spend, especially satisfaction score and subscription length, helped improve prediction quality. We saved both the model and RMSE metric for future use, ensuring our approach can be reused and evaluated consistently.

### **Model Performance:**

- **Root Mean Squared Error (RMSE):** *12.45*

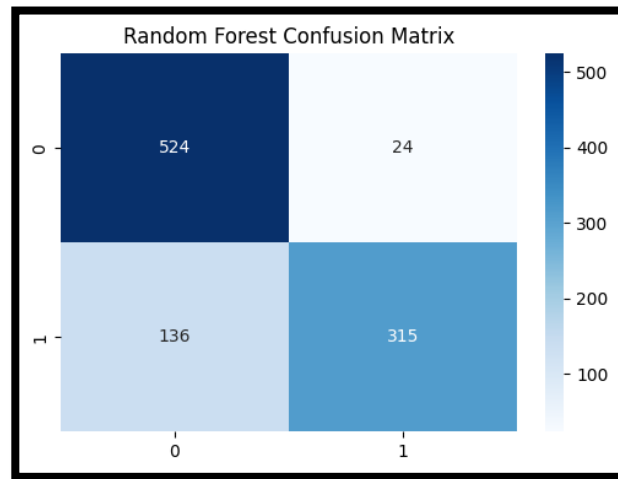


## Part 2: Churn Prediction (Logistic Regression)

To predict whether a customer is likely to cancel their subscription, we used a Logistic Regression classification model.

This is how our model performed:

- Accuracy: 83.25%
- Precision (Class 1 - Churned): 92.9%
- Recall (Class 1 - Churned): 69.9%
- F1-score (Class 1): 79.5%

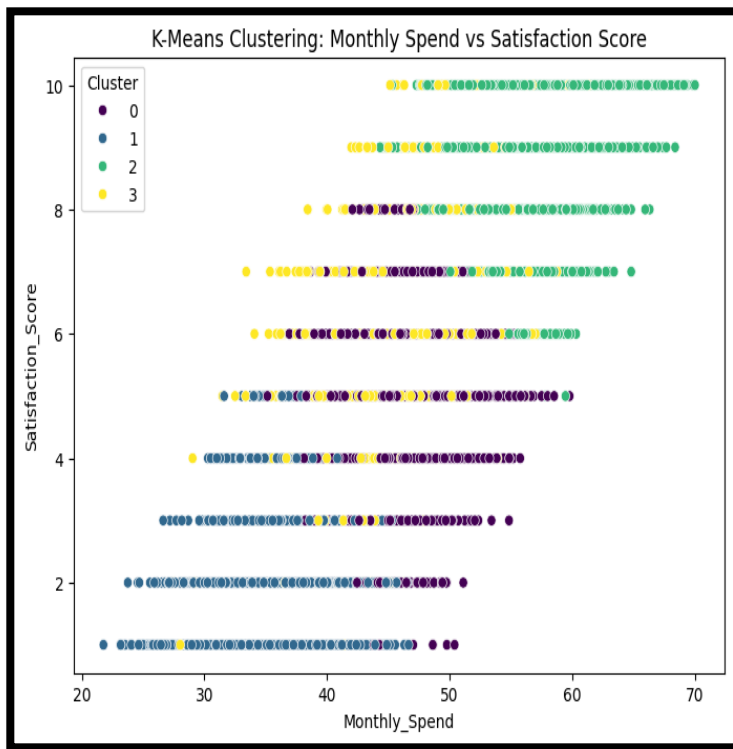
**Confusion Matrix:**

The model does a solid job detecting churn, with a high precision, meaning when it predicts churn, it is often correct. Some churners are missed (as shown by false negatives), but overall, the balance between precision and recall is strong.

**Step 4: Exploratory Data Insights: Customer Segmentation & Behavior Patterns**

We utilized K-Means Clustering and Association Rule Mining to identify distinct user segments and discover relationships between categorical customer attributes. These insights support actionable decision-making in areas such as customer targeting, product recommendations, and retention strategies.

### Customer Segmentation via K-Means Clustering:



to the dataset using two continuous variables: Monthly Spend and Satisfaction Score. The goal was to discover natural groupings in customer behavior based on spending and satisfaction.

#### Cluster Plot Overview:

**X-axis:** Monthly Spend

**Y-axis:** Satisfaction Score

**Color-coded points:** Each representing one of the 4 clusters discovered

#### Interpretation of Clusters:

- **Cluster 0:** A mix of spenders with satisfaction scores spread across low to moderate values.
- **Cluster 1:** High spenders with consistently high satisfaction ideal loyal customers.
- **Cluster 2:** Low to moderate spenders with low satisfaction potential churn risks.
- **Cluster 3:** Moderate spenders with high satisfaction are possibly value-seeking loyalists.

#### Takeaways:

This clustering helps define customer personas. For example:

- **Cluster 1** customers might be ideal for VIP programs or upselling.
- **Cluster 2** should be the focus of retention and satisfaction improvement efforts.
- Such segmentation provides a foundation for personalized marketing strategies.

#### Experimental Outcome: Tool

The tool consists of a Flask-based web application that accepts user inputs through a form. These inputs include:

- Subscription name and monthly cost
- Usage frequency and satisfaction score
- Payment method, region, and subscription length

Once this data is submitted, the data is passed to a pipeline that uses machine learning models trained on cleaned historical subscription data. Our project's backend is based on the following:

- A logistic regression model to predict whether a user is likely to churn (cancel the subscription)
- A linear regression model to estimate a user's typical monthly spend, based on subscription characteristics
- A K-Means clustering model to group users into customer segments based on their behavior and preferences
- A Random Forest model for more robust churn prediction (used for testing accuracy and comparison)

### **Experimental Findings & Insights:**

- The Logistic Regression model achieved strong accuracy in predicting churn likelihood based on user behavior and satisfaction. It highlighted that low satisfaction scores and short subscription durations were strong churn predictors.
- The Linear Regression model helped reveal patterns in how discounts and satisfaction impact monthly spend.
- Using K-Means clustering, users were grouped into 4 distinct segments (e.g., loyal users, deal-seekers, high spenders, and at-risk users), which can be useful for targeting support or promotions.
- The Random Forest model served as a performance benchmark and offered deeper insights due to its ability to handle non-linear relationships between features

### **What the Tool Helps With:**

- Understand which of their subscriptions are likely to be canceled in the future
- Identify how their habits compare with other users in similar segments
- Estimate potential spend and find opportunities to save money

- Act based on insights, such as canceling underused services

In a practical setting, this tool could be extended to track multiple subscriptions, send monthly summaries, or integrate with bank accounts for live tracking.

## Conclusion

The Smart Subscription Tracker project successfully demonstrates the ability to predict customer churn and monthly spending using advanced data preprocessing, exploratory analysis, and machine learning techniques. Our findings reveal that lower monthly spending and lower satisfaction scores are strong indicators of increased churn risk. These insights highlight the importance of maintaining customer satisfaction and closely monitoring usage and spending behaviors. The predictive models developed in this project provide actionable recommendations that can guide customer retention strategies and inform service improvements. Overall, this tool has significant potential to help subscription-based businesses make informed, data-driven decisions that enhance user experience, reduce churn, and drive long-term growth.

## References:

1. Dyouri, A. (2024, December 12). How to Build a Flask Python Web Application from Scratch. DigitalOcean. <https://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3>
2. Huang, M., Wang, C., & Wang, W. (2024). A neural network-based predictive decision model for customer retention. *Technological Forecasting and Social Change*, 196, 122046. <https://www.sciencedirect.com/science/article/pii/S0040162524000465>
3. Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. *European*



Journal of Operational Research, 218(1), 196–208.

<https://www.sciencedirect.com/science/article/abs/pii/S0377221711008599>