Analysis of Netflix Userbase Dataset

Baraa Fatima Zohra LALAGUI

Limitless Learning

Abstract. This paper presents a comprehensive analysis of the Netflix Userbase dataset, aiming to uncover insights and relationships between various features. The analysis includes data cleaning, exploratory data analysis, and visualization techniques to provide valuable insights for decision-making processes.

1 Introduction

For streaming services like Netflix to optimize revenue generation, improve user experience, and make better content recommendations, they must have a thorough understanding of user behavior and preferences. In order to obtain knowledge about user demographics, subscription preferences, and usage patterns, we examine the Netflix Userbase dataset in this study.

2 Background

The Netflix Userbase dataset includes data on device usage, subscription type, demographics, and other pertinent characteristics of its consumers. Trends, inclinations, and other elements impacting user engagement and platform retention can be found by analyzing this dataset.

3 Dataset Description

The dataset from Netflix Userbase has 10 columns and 2500 rows. The dataset offers a thorough understanding of user interactions with the platform through a combination of datetime, category, and numerical variables. User ID, monthly revenue, age, kind of subscription, nation, gender, device, join date, last payment date, and plan duration are the columns to be used for the analysis.

4 Analytical Findings

4.1 Data Cleaning

- Deleting redundant data (User ID: unique value in each row, Plan Duration: same value in all rows)
- Renaming the columns to more convenient and useful names
- Standardizing the format of the columns values (e.g., gender (m, f))
- Dropping incorrect data: rows with

$$lastPaymentDate < joinDate$$

Creating Columns

– Created the column 'join time days' to know how many days the user was subscribed using

$$lastPaymentdate - joinDate$$

– Created a column for each device type (binary column)

4.2 Exploratory Data Analysis

– Use bar plots to visualize the number of users in each value in each categorical feature:

- Visualizing "device" as shown in the figure below:

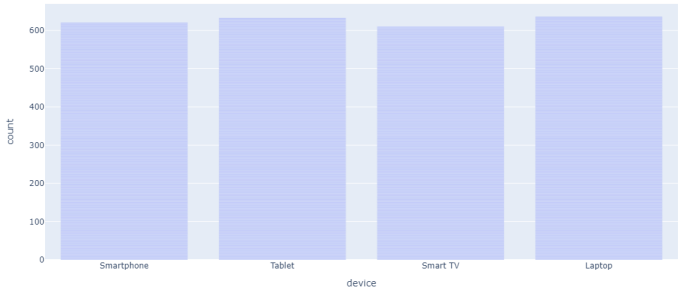

Fig. 1: Visualization of "device"

- Visualizing "country" as shown in the figure below:

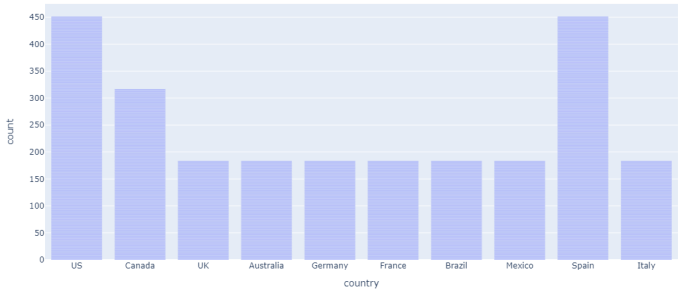

Fig. 2: Visualization of "country"

- Visualizing "subscription type" as shown in the figure below:

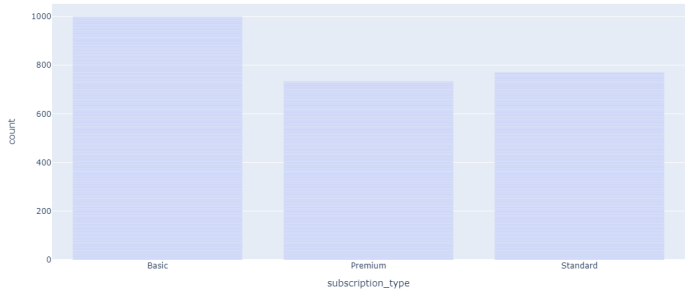

Fig. 3: Visualization of "subscription type"

– Checking the percentage of each gender using pie chart

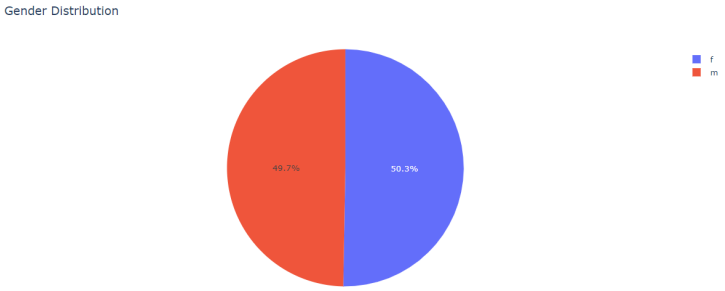

Fig. 4

The dataset was basically unbiased towards gender, where the sample of males was approximately equal to the females sample, so we can say that our testing sampling is representative of the population.

– Checking the distribution of the 'join time' of users

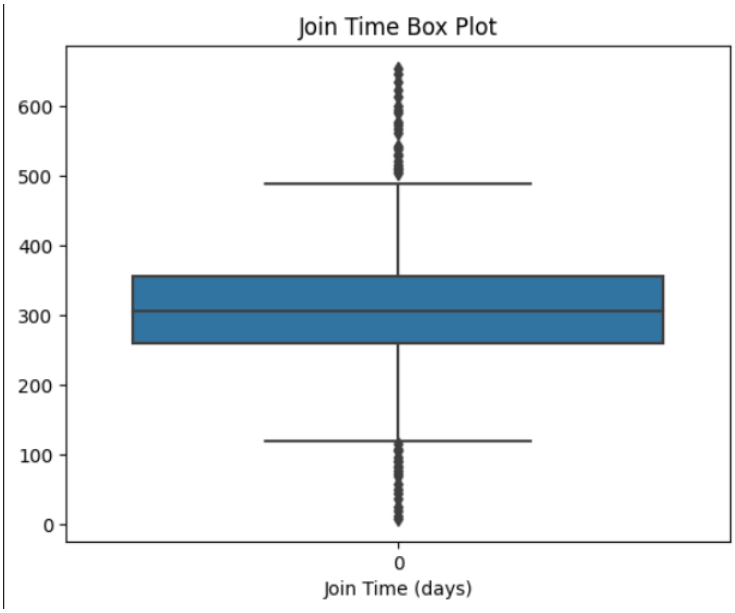

Fig. 5

The mean of the joining time is 300 days which is approximately a year, for further details, we plotted the relationship between subscription type and the join time

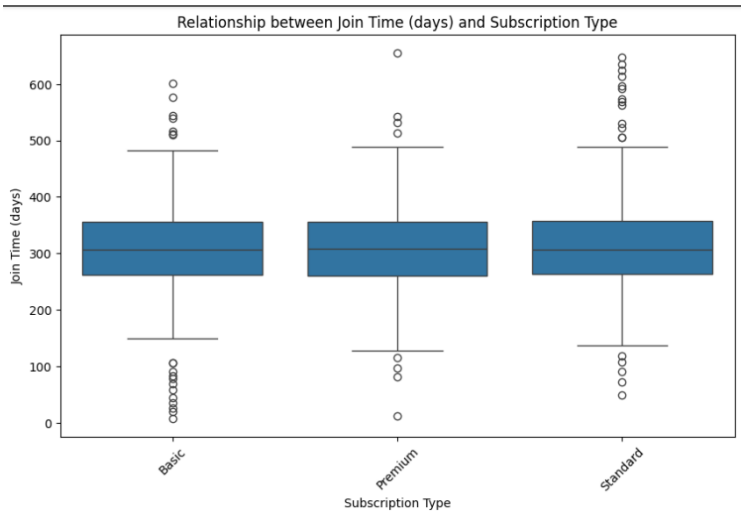

Fig. 6

Notice that they all have the same mean but the main difference is in the values beyond the IQR, in the basic subscription more people drop the subscription after a shorter period, whereas in the standard subscription we

have more people continuing to use the subscription, could be because of its quality

– Checking the distribution of each subscription type under each device type

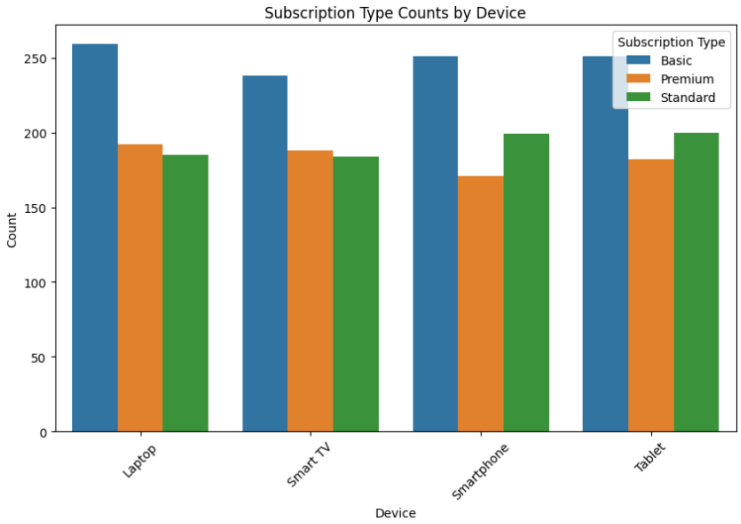

Fig. 7

We notice that bigger devices have more premium subscriptions contradictory to smaller devices that have more standard subscriptions, the basic subscription type is the most used

– Plotting the relationship between the monthly revenue and the subscription type

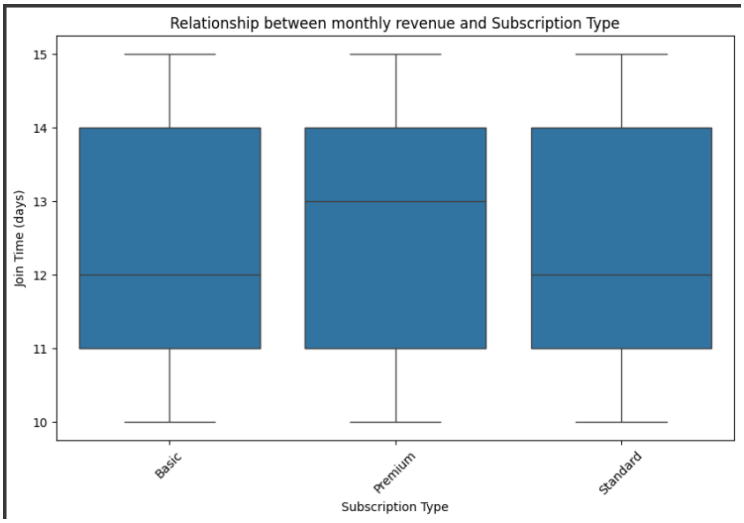

Fig. 8

The premium version leads to higher revenue compared to the standard and the basic

5 Conclusion

Important information about user behavior and preferences was gained from the examination of the Netflix Userbase dataset. Streaming systems can boost user contentment, optimize marketing strategies, and improve content recommendation algorithms by utilizing these insights. Predictive modeling to estimate user attrition and maximize monthly subscription income may be the main focus of future study.