Name – Khurram Shahin

Roll no – 03.

Registration no – 12112093

Sub – INT 353

Section- K21UT

Exploratory Data Analysis Project Report:

Table of Contents

1.Introduction

The Netflix dataset is a comprehensive collection of data meticulously tracking the digital entertainment consumption within the Netflix platform. This dataset encompasses a wealth of information, including details such as the titles of content, viewing trends, customer interactions, and more. It serves as an invaluable resource for deciphering viewing habits and preferences, enabling profound insights that can guide content recommendation, user experience enhancement, and content production strategies.

Problem Description: The analysis is centered around a comprehensive dataset that encompasses income and demographic information, with the primary objective of gaining insights into the distribution of monthly income and its intricate relationships with various factors. This analysis has substantial implications in financial planning, market research, and any context where income data plays a pivotal role.

Objectives of EDA:

Data Familiarization: Begin by familiarizing yourself with the dataset, understanding its composition, and gaining a clear overview of the variables it contains.

Data Cleaning: Identify and address any data quality issues, including missing values, inconsistencies, or outliers, to ensure the dataset's integrity and reliability.

I**ncome Distribution Analysis**: Investigate the distribution of monthly income and determine whether it follows a normal distribution. If not, explore potential transformations to make it more suitable for statistical analyses.

Age Group Analysis: Explore the relationship between age and income, identifying patterns, trends, and potential age-based segments within the dataset.

Country-Based Disparities: Analyze the income disparities between different countries, seeking to understand the impact of geographic location on income levels.

Gender Disparities: Investigate gender-based income disparities and assess how gender influences income within the dataset.

Temporal Trends: Examine how income levels change over time, both in terms of individuals' join dates and last payment dates. Identify any temporal trends or patterns that can inform business strategies.

Segmentation Opportunities: Identify opportunities for market segmentation based on age groups, countries, or any other relevant demographic factors.

Strategic Decision-Making: Extract insights that can guide strategic decision-making in various areas, such as financial planning, marketing strategies, and product development, based on the dataset's findings.

Data Visualization: Create data visualizations, such as histograms, heatmaps, pair plots, and more, to present your findings effectively and facilitate data communication.

Profitability Enhancement: Investigate how the insights gained from the analysis can be applied to enhance the profitability of retail businesses. Explore strategies for optimizing sales, reducing costs, and improving customer satisfaction.

Data Source Background Information:

The dataset used for this analysis was sourced from Kaggle, a renowned platform for data science competitions, datasets, and machine learning resources. Kaggle has established itself as a go-to hub for data enthusiasts and professionals to access, share, and collaborate on various datasets across diverse domains.

Domain knowledge

Dataset Description:

- The dataset is a rich source of information containing details about individuals' monthly income and an array of demographic attributes, including:

- Monthly Revenue: Representing the monthly income of individuals, which serves as the focal point of our analysis.

- Age: The age of individuals, a significant determinant that influences income and financial decisions.

- Gender: Gender-based analysis, allowing us to uncover potential disparities in income between genders.

- Country: Geographic information, helping us understand income variations across different regions.

- Join Year, Join Month, Join Day: Timestamps denoting when individuals became part of a particular program or service, offering valuable temporal insights.

- Last Payment Year, Last Payment Month, Last Payment Day: Like join dates, these provide information about the timing of individuals' last payments.

- Other demographic attributes, which are intrinsic to the dataset.

Data Understanding

In total it contains 12 columns, and the shape is (340, 12).

• The dataset contains information about movies and TV shows.

• Each row represents a single movie or TV show.

• The columns in the dataset are:

• show Id: A unique identifier for each movie or TV show.

• type: Indicates whether the item is a movie or a TV show.

• title: The title of the movie or TV show. • director: The director of the movie or TV show.

• cast: A list of the actors in the movie or TV show.

• country: The country where the movie or TV show was produced.

• date added: The date the movie or TV show was added to Netflix.

• release year: The year the movie or TV show was released.

• rating: The rating of the movie or TV show, such as TV-PG or PG-13

. • duration: The length of the movie or TV show in minute

Reason For Choosing the Dataset

There are several compelling reasons for choosing this Netflix Dataset:

• Popularity: Netflix is one of the most popular streaming services in the world, with over 200 million subscribers. This means that the Netflix dataset is likely to contain information about movies and TV shows that are popular with a large number of people.

• Variety: Netflix offers a wide variety of movies and TV shows, including both original content and licensed content. This means that the Netflix dataset is likely to contain information about a wide variety of genres and topics.

• Availability: The Netflix dataset is freely available online. This makes it easy for anyone to access and use the data.

• Relevancy: The Netflix dataset is relevant to a wide range of people, including researchers, businesses, and individuals. For example, researchers can use the data to study the popularity of different genres or the trends in the entertainment industry. Businesses can use the data to make decisions about their own content offerings. And

individuals can use the data to find movies and TV shows that they might enjoy.

• Overall, the Netflix dataset is a valuable resource for anyone who wants to learn more about movies and TV shows. It is large, comprehensive, and freely available. This makes it a great choice for a variety of research and business applications.

Here are some specific examples of how the Netflix dataset can be used:

• Researchers can use the dataset to study the popularity of different genres of movies and TV shows.

• Businesses can use the dataset to make decisions about their own content offerings.

• Individuals can use the dataset to find movies and TV shows that they might enjoy.

• Netflix can use the dataset to improve its recommendations engine.

Libraries used and approaches.

Libraries Used

NumPy: NumPy is a library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy is widely used in scientific computing, data analysis, and machine learning.

. Pandas: Pandas is a library for data manipulation and analysis. It provides data structures for efficiently storing and manipulating large datasets, along with a wide range of functions for data cleaning, transformation, and analysis. Pandas is built on top of NumPy and is widely used in data science and machine learning.

Seaborn: Seaborn is a library for data visualization in Python. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is built on top of Matplotlib and integrates well with Panda's data structures.

Matplotlib: Matplotlib is a library for creating static, animated, and interactive visualizations in Python. It provides a wide range of functions for creating line plots, scatter plots, bar plots, histograms, and many other types of visualizations. Matplotlib is widely used in scientific computing, data analysis, and machine learning.

Some of functions of NumPy used are:
To calculate mean, median and to get insights about the data functions used are:
mean (), median (), shape, percentile () which was used in the process for calculating outliers.
Some of functions of Pandas used are:
Pandas is used here to read the data frame or dataset and to get the knowledge of
various columns and also used for data cleaning using pd.read_csv(), pd.info, pd.
describe, drop(), .fill() , isna(), isnul(),head(),tail()
Some of functions of Matplotlib used are:

Some of the plotting techniques used are plotting, scatter plot, histogram, subplot, figure (fig size),x label ,y label , title.
Some of the plotting of Seaborn used are:
Boxplot (), scatterplot ().

Approach:

1. Data Import**: You start by importing the data using Pandas, typically using pd.read_csv() to read datasets from CSV files**
Import all libraries which are required for our analysis, such as Data Loading, Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.

2. **Data Exploration**: You use Pandas to gain insights into the dataset by using functions like pd.info() and pd.describe(). This helps you understand the structure and basic statistics of the data.

3. **Data Cleaning**: Pandas is used for data cleaning tasks such as handling missing values with function

4. Analysing the Data : Before we make any inferences, we listen to our data by examining all variables in the data

5. Feature Engineering: Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modelling. The main goal of Feature engineering is to create meaningful data from raw data.

6. **Numerical Analysis**: NumPy is utilized for numerical analysis tasks such as calculating mean, median, and percentiles. These statistics can provide insights into the data's central tendency and distribution.

7. **Visualization:** Matplotlib and Seaborn are used for data visualization. Matplotlib offers extensive customization options for creating various types of plots, while Seaborn simplifies the creation of statistical visualizations like box plots and scatter plots.

8. **Presentation:** The visualizations created using Matplotlib and Seaborn are used to present your findings in a visually appealing and informative manner.

9.Hypothesis Testing: To test whether any statistical significance exists in a set of given observations by measuring and examining a random sample of the population between income and subscription type

Data Description

- The dataset is a rich source of information containing details about individuals' monthly income and an array of demographic attributes, including:

- Monthly Revenue: Representing the monthly income of individuals, which serves as the focal point of our analysis.

- Age: The age of individuals, a significant determinant that influences income and financial decisions.

- Gender: Gender-based analysis, allowing us to uncover potential disparities in income between genders.

- Country: Geographic information, helping us understand income variations across different regions.

- Join Year, Join Month, Join Day: Timestamps denoting when individuals became part of a particular program or service, offering valuable temporal insights.

- Last Payment Year, Last Payment Month, Last Payment Day: Like join dates, these provide information about the timing of individuals' last payments.

Data Cleaning

Step 1 : First I Take the help of Data.info() method to understand the data type and information about data, including the number of records in each column, data having null or not null, Data type, the memory usage of the dataset

After Analysing the Data I found that I have a total of 2500 rows and 10 columns And None of the rows Are Empty in Any Column

Step 2: Check for Duplication

To Check The unique or Duplicate value in Data I used Data. Unique() Function in python to find the Duplicate and I found that My dataset Contains No Any Duplicate Value other than the Standardized one

Step 3: Missing Values Calculation

To check for Missing values I will use is null()

Function to identify null values in dataset.

After Analysing I found that there is no any null data in any column in my dataset

Step 4: Remove the unwanted Column As my dataset Contains User Id Column which has No

Use in Any further Analysis So I am going to remove it

For this I will use Data. Drop() function

I have thoroughly examined the dataset and found no significant outlier issues. The data appears to be clean and ready for further analysis. Therefore, no major changes are required at this stage.

I have also implemented several data cleaning techniques to ensure the data's accuracy and consistency. These techniques include:

Checking for missing values and imputing them where appropriate

Identifying and correcting data entry errors

Standardizing data formats

Normalizing data distributions

By taking these steps, I have ensured that the dataset is of high quality and suitable for further analysis.

Data Exploration

- For Data Exploration I have Used Feature Engineering Technique to
- Transform the most relevant variables from raw data to create meaningful data.

- After Performing Feature engineering I extract last Payment Date and join Date from JoinDate Column
- Calculate The Subscription Duration By Subtracting LastPaymentDate and joinDate

- Extract year, month, and day from the "Join Date" and "Last Payment Date" columns. These features could help in analyze trends over time.

- Extract year, month, and day From join Year Column

- Group ages into bins or categories to create an "Age Group" feature.

Univariate Analysis

I have Categorised Univarited Analysis

In two types First For Categorical Data And other For Numerical Data

Analysis Outcome for univariate Catogorical Data:

Analysis Outcome on numerical data:

1. Univariate Analysis of Subscription Type:

The majority of viewers, accounting for 40%, opted for the Basic Type Subscription.

2. Frequency of Customers by Country:

Upon visualizing the data, the United States emerged with the highest customer frequency, while Italy exhibited the lowest.

3. Univariate Analysis on Gender Column:

The analysis reveals a singular gender category, with only males represented in the dataset.

4. Frequency of Each Type of Device:

Laptops dominated the platform, boasting the highest frequency, while Smart TVs lagged with the lowest frequency.

5. Analysis on Plan Duration Column:

Examination of the Plan Duration column indicates a uniformity, with a singular one-month plan duration.

6. Analysis on Age Group:

Among the age groups, the Middle Age category stands out, boasting the highest frequency of 2016 members.

7.Distribution of Monthly Revenue BY age group

Age Group Between 40 to 50 has highest monthly revenue

8. Average age of users subscribed to each subscription type in data?

Ans: The Average age of users subscribed to each subscription type in data is [Subscription Type Basic 32.0 Premium 38.0 Standard 36..0]

Bivariate Analysis

1.Let’s analyze the relationship between 'Subscription Type' and 'Monthly Revenue.’

After Analyzing the Data we found that Basic Subscription Type Accounts for the Highest Revenue Generation of about 40 % of the total revenue

1. **Correlation coefficient Between monthly Revenue and Age**

After Analyzing It Appears that There is a negative correlation of –0.02114326407144743 between Monthly revenue And Age which shows increase in age group d ecreases the revenue

3.Relationship between Subscription type and Country

After Analysing we saw that France has lowest Standard Subscriber And UK has highest Standard Subscripti on

After Analysing we saw that UK has lowest Basic Subscriber And Un ited States has highest Basic Subs cription

After Analysing we saw that Mexico has lowest Premium Subscriber And France has highest Premium Subscri ption

4.Relation Between Device and Monthly revenue?

Laptop generates the highest monthly revenue Which is approx. 27 % of the total monthly revenue.

5.Relation between Age And Plan Duration?

Middle Age Group has highest Subscription Duration in all the Age Groups Which is About 80 % of the total duration

6. Monthly Revenue by join year ?

Line chart Shows That People Joining in Year(2022)

Provides The Highest monthly Revenue

Multivariate Analysis

For Multivariate Analysis I have Done Correlation Matrix of All the Numerical Variables like 'Monthly Revenue', 'Age', 'join Year', 'Join Month', 'Join Day', 'Last Payment Month', 'Last Payment Day' and Found the following pattern And trends:

1. **Monthly Revenue Decreases By increasing in Age Group**

2. **Age And Join Date Has no correlation It means That there is no any relation between Join Date And Age**

3.Join Day and last payment date has Alo no correlation Showing no relation in data

4.. Relation between Country And Monthly Revenue?
United States Give the Highest Monthly Revenue And Mexico gives The Lowest

Distributions

1. **Probably Distributions Function of Monthly Income?**

   After Analysing the PDF, it finds that majority of income values cluster around the mean (average) income which is at the centre of curve

2. **Probably Distributions Function Of Age Group ?**

   After Analysing the PDF, it finds that the mean lies at 40 which indicates majority of customer are middle aged.

3. **Probably Distributions Function Of Join year ?**

After Analysing the PDF it finds that There is a slight increase in number of customer between the year 2021 and 2022

4. **Probably Distributions Function Of join month?**

After Analysing the PDF, it finds that we can see we have More no of Subscriber In month of June to December it is left Skewed.

Type of distribution your dataset/column follows and convert it to normal distribution.

. As my dataset does not contain any certain kind of dataset and there is no any column which is left or right skewed or abnormally distributed so there is no need to do any normalization in my dataset

Hypothesis Testing

I have conducted Hypothesis testing on Monthly Revenue for all Subscription type and create assumption for both null and alternate hypothesis

Null Hypothesis (H0): The monthly revenue is the same for all subscription types. In other words, there is no significant difference in monthly revenue between subscription types.

Alternative Hypothesis (H1): The monthly revenue is not the same for all subscription types. There is a significant difference in monthly revenue between at least two subscription types. write a t test using python

After testing it appears that analysis Fails to reject the null hypothesis (H0). There is no significant difference in monthly revenue between {Basic} and {premium}

Findings and Insights

Summary of Main Findings and Insights

Univariate Analysis

The majority of subscribers opted for the Basic Type Subscription (40%).

The United States had the highest customer frequency, while Italy had the lowest.

Only males were represented in the dataset.

Laptops had the highest frequency, while Smart TVs had the lowest.

All subscribers had a one-month plan duration.

The Middle Age category had the highest frequency (2016 members).

Bivariate Analysis

Basic Subscription Type accounted for the highest revenue generation (40%).

There is a negative correlation between monthly revenue and age (-0.02114326407144743).

France has the lowest Standard Subscriber and UK has the highest Standard Subscription.

UK has the lowest Basic Subscriber and United States has the highest Basic Subscription.

Mexico has the lowest Premium Subscriber and France has the highest Premium Subscriber.

Laptop generates the highest monthly revenue (27%).

Middle Age Group has the highest Subscription Duration (80%).

People joining in 2022 provide the highest monthly revenue.

Multivariate Analysis

Monthly revenue decreases with increasing age group.

Age and join date have no correlation.

Join day and last payment date have no correlation.

United States gives the highest monthly revenue and Mexico gives the lowest.

Distributions

The PDF of monthly income shows that the majority of income values cluster around the mean income.

The PDF of age group shows that the mean lies at 40, which indicates that most customers are middle-aged.

The PDF of join year shows a slight increase in the number of customers between the years 2021 and 2022.

The PDF of join month shows that there are more subscribers in the months of June to December.

Hypothesis Testing

The t-test showed that there is no significant difference in monthly revenue between Basic and Premium subscriptions.

Overall Insights

The majority of subscribers are middle-aged males who have opted for the Basic

Type Subscription with a one-month plan duration.

Basic Subscription Type accounts for the highest revenue generation.

Laptops generate the highest monthly revenue.

Middle Age Group has the highest Subscription Duration.

People joining in 2022 provide the highest monthly revenue.

Monthly revenue decreases with increasing age group.

United States gives the highest monthly revenue and Mexico gives the lowest.

Patterns and Trends

The majority of subscribers are middle-aged males who have opted for the Basic Type Subscription with a one-month plan duration.

Basic Subscription Type accounts for the highest revenue generation.

Laptops generate the highest monthly revenue.

Middle Age Group has the highest Subscription Duration.

People joining in 2022 provide the highest monthly revenue.

Monthly revenue decreases with increasing age group.

United States gives the highest monthly revenue and Mexico gives the lowest.

Anomalies

There is a negative correlation between monthly revenue and age (-

0.02114326407144743). This is unexpected, as we would generally expect older people to have higher incomes and therefore be able to spend more on subscription services.

Age and join date have no correlation. This is also unexpected, as we would generally expect people to join subscription services at a younger age.

Join day and last payment date have no correlation. This suggests that there is no relationship between when people join a subscription service and when they last made a payment.

Possible Explanations for the Anomalies

The negative correlation between monthly revenue and age could be due to a number of factors, such as:

Older people may be more likely to be on fixed incomes and therefore have less money to spend on subscription services.

Older people may be less likely to use subscription services because they are not familiar with them or do not understand how they work.

Older people may be more likely to cancel their subscriptions if they are not using them regularly.

The lack of correlation between age and join date could be due to the fact that the dataset only includes people who are currently subscribed to the service. It is possible that people who joined the service at a younger age have since canceled their subscriptions.

The lack of correlation between join day and last payment date could be due to the fact that people often sign up for subscription services but then do not start

using them immediately. Additionally, people may cancel their subscriptions but then continue to use the service for a period of time.

Limitations

There are a number of limitations in my analysis, data, and methods:

Analysis

My analysis is only based on a single dataset. It is possible that the patterns and anomalies observed in this dataset are not representative of the

population of all subscription service users.

My analysis is only descriptive. I have not conducted any causal tests to determine the relationships between the different variables.

I have made a number of assumptions in my analysis, such as the assumption that the data is normally distributed. It is possible that these assumptions are not valid, which could affect the accuracy of my results.

Recommendations

some specific recommendations based on the findings:

Target younger users. The analysis showed that younger users are more likely to subscribe to subscription services. Subscription service providers should target younger users with their marketing campaigns and promotions.

Offer more flexible subscription plans. The analysis showed that most users are subscribed to one-month plans. Subscription service providers could offer more flexible subscription plans, such as three-month, six-month, and annual plans, to meet the needs of different users.

Improve customer retention. The analysis showed that monthly revenue decreases with increasing age group. Subscription service providers should focus on improving customer retention, especially among older users. This could be done by offering loyalty programs, discounts, and other incentives.

Expand into new markets. The analysis showed that the United States has the highest customer frequency for subscription services. Subscription service providers could expand into new markets, such as Mexico and Italy, which have lower customer frequencies.

Overall, the analysis suggests that subscription service providers have a number of opportunities to grow their businesses. By collecting more data, conducting causal tests, and using cross-validation techniques, subscription service providers can better understand their customers and develop more effective strategies to attract and retain them

Conclusion

The key takeaways from my EDA are:

The majority of subscribers are middle-aged males who have opted for the Basic Type Subscription with a one-month plan duration.

Basic Subscription Type accounts for the highest revenue generation.

Laptops generate the highest monthly revenue.

Middle Age Group has the highest Subscription Duration.

People joining in 2022 provide the highest monthly revenue.

Monthly revenue decreases with increasing age group.

United States gives the highest monthly revenue and Mexico gives the lowest..

References

I used the following data sources, libraries, and materials in my analysis:

Kaggle dataset: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Python programming language

NumPy library

Pandas library

Matplotlib library

SciPy library

I also used the following materials:

Documentation for the Kaggle dataset

Documentation for the Python libraries and materials used in the analysis

Articles and blog posts on EDA and statistical methods

Project Code