< Big data Final Project >

Soomin, Nicholas, Yehonatan

1. Obtain the Dataset

(minimum 1,000 rows and 10 features recommended).

Netflix Userbase Dataset

2. Dataset Description

The dataset provides a snapshot of a sample Netflix user base, showcasing various aspects of user subscriptions, revenue, account details, and activity. Each row represents a unique user, identified by their User ID. The dataset includes information such as the user's subscription type (Basic, Standard, or Premium), the monthly revenue generated from their subscription, the date they joined Netflix (Join Date), the date of their last payment (Last Payment Date), and the country in which they are located.

Additional columns have been included to provide insights into user behavior and preferences. These columns include Device Type (e.g., Smart TV, Mobile, Desktop, Tablet) and Account Status (whether the account is active or not).

The dataset contains 10 columns and 2500 rows. Here's an overview of the key columns:

Feature	**Example Values**	**Type**	**Size**
User ID	1,2,3,4......2000	Integer	4
Subscription Type	Basic,Standard,Premium	String	10
Monthly Revenue	10,12,15	Integer	2
Join Date	01-05-23, 18-03-22...	String	8
Last Payment Date	01-05-23, 18-03-22...	String	8
Country	Germany, France, Mexico...	String	16
Age	28,35,42....	Integer	2
Gender	Male,Female	String	6
Device	Smartphone,Tablet,laptop....	String	16
Plan Duration(Days)	600, 425....	Integer	4

After data transformation

User_ID	INTEGER
Subscription_Type	INTEGER
Monthly_Revenue	INTEGER
Join_Date	DATE
Last_Payment_Date	DATE
Country	INTEGER
Age	INTEGER
Gender	INTEGER
Device	INTEGER
Plan_Duration_Days	INTEGER

3. Data Cleaning and Transformation

<Convert categorical data into numerical values >
Subscription Type: ['Basic-0 ', 'Premium-1', 'Standard-2']
Country: ['Australia-0', 'Brazil-1', 'Canada-2', 'France-3', 'Germany-4', 'Italy-5', 'Mexico-6', 'Spain-7', 'United Kingdom-8', 'United States-9']
Gender: ['Female-0', 'Male-1']
Device: ['Laptop-0', 'Smart TV-1', 'Smartphone-2', 'Tablet-3']

< Delete meaningless column >
The plan_duration column contained ‘1 month’ in each row. We believed this misrepresented what the column should mean, which is the length of time that a customer has had their subscription. We dropped the original column, replacing it with a column which represents the difference between the time the customer joined and their last payment date. Also we transfer column names space to _ to make using sql easier.

< Adding column >
We add Plan_Duration_days which describes Last Payment Date - Join Date.

4. Define Questions

1. Revenue Analysis

1. How has total monthly revenue changed over time, and what growth patterns can be identified?

2. How does revenue distribution vary by country, and what can be inferred about the market value of each region?

2. Identifying Key User Groups Based on Age and Gender

3. What devices are most preferred by the key user groups?

4. Which subscription plans are most popular among the key user groups?

5. Create the ETL Pipeline

- Extract:
  We began by extracting the Netflix customer dataset from a CSV file containing various attributes such as `User_ID`, `Subscription_Type`, `Monthly_Revenue`, `Join_Date`, `Last_Payment_Date`, `Country`, and more.

- Transform
  - Column cleaning
    Removed unnecessary columns such as `Plan_Duration` and standardized column names by replacing spaces with underscores.
  - Handling categorical variable
    Encoded `Subscription_Type`, `Country`, `Gender`, and `Device` into numeric values for analysis.
  - Handling numeric fields
    Converted fields like `Monthly_Revenue` and `Age` to their correct data types, ensuring invalid values were addressed appropriately.
  - Feature engineering
    Created Lifetime Value (LTV) using the formula:
    `LTV = Monthly_Revenue * Plan_Duration_Days / 30`.

- Load :

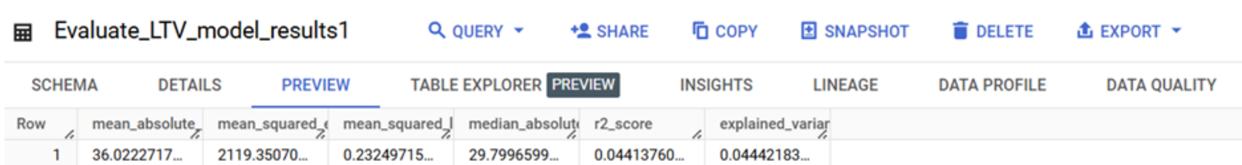
- After transforming the data, we utilized Google Cloud Dataflow for efficient data loading and integration into BigQuery:
  - The cleaned CSV file was uploaded to Google Cloud Storage.
  - A Dataflow pipeline was created to read, process, and load the data into a BigQuery table named `cleaned_subscription_data_with_ltv`.

6. Analysis and Prediction

<Revenue Analysis>

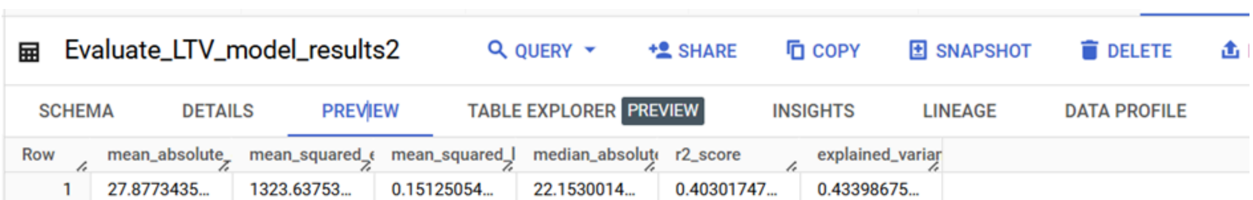
Calculating LTV

Model 1: Linear_reg



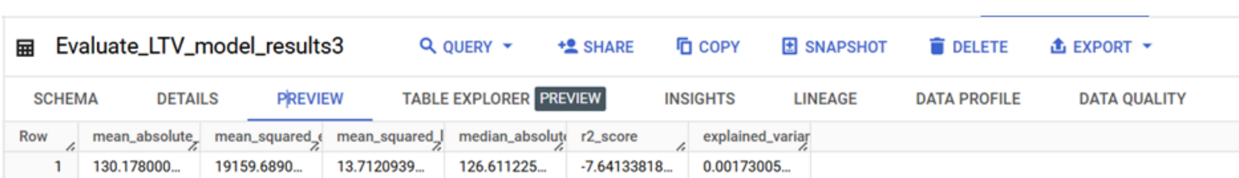
Row	mean_absolute_	mean_squared_	mean_squared_l	median_absolut	r2_score	explained_variar
1	36.0222717...	2119.35070...	0.23249715...	29.7996599...	0.04413760...	0.04442183...

Model 2: Boosted Tree reg



Row	mean_absolute_	mean_squared_	mean_squared_l	median_absolut	r2_score	explained_variar
1	27.8773435...	1323.63753...	0.15125054...	22.1530014...	0.40301747...	0.43398675...

Model 3: DNN reg



Row	mean_absolute_	mean_squared_	mean_squared_l	median_absolut	r2_score	explained_variar
1	130.178000...	19159.6890...	13.7120939...	126.611225...	-7.64133818...	0.00173005...

To predict customers' Lifetime Value (LTV), which is a continuous value, we decided to use regression models. Initially, we started with a simple and fast Linear Regression model due to its ease of implementation. However, the results were poor, indicating that the relationships in the data were not linear. To address this, we applied a Boosted Tree Regression model, which is capable of capturing non-linear relationships. This approach showed improved performance compared to Linear Regression. To further enhance the results, we used a Deep Neural Network (DNN) Regression model, as it is designed to learn complex patterns and interactions within the data. Unfortunately, the DNN model did not perform well, likely due to the limited size of the dataset, which caused overfitting. In the future, we plan to scale the data, expand the dataset, and optimize the models to improve LTV prediction accuracy.

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	mean_absolute_error	mean_squared_error	mean_squared_log_e	median_absolute_err	r2_score	explained_variance
1	129.1963986625...	18903.88255044...	11.70600293103...	125.3546788851...	-7.52596520870...	0.002158797744...

Ultimately, we attempted to upgrade the TreeBoost model by adjusting variables such as max_iterations, learn_rate, and subsample, but the results worsened**.**

Predicting revenue

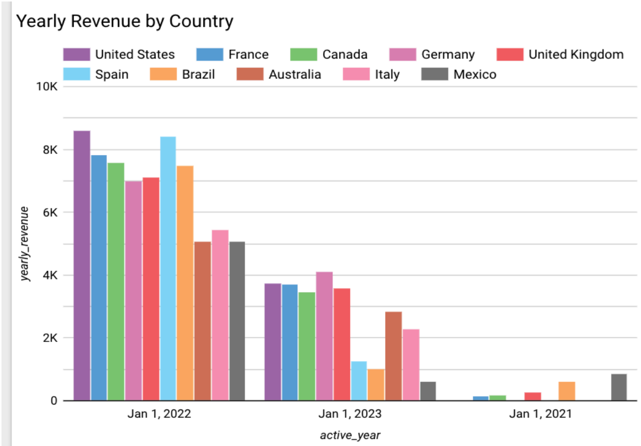
We chose the ARIMA model because it is well-suited for time series analysis, effectively capturing trends and patterns in historical data to make accurate future predictions.

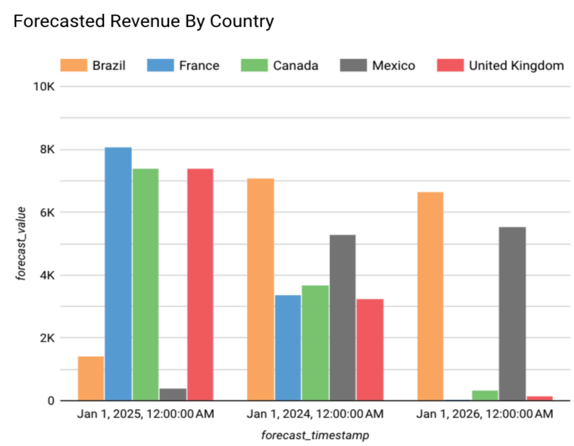
Actual values by year:

Row	active_year	Country	yearly_revenue
1	2022-01-01	Australia	5088.0
2	2023-01-01	Australia	2844.0
3	2022-01-01	Brazil	7500.0
4	2023-01-01	Brazil	1020.0
5	2021-01-01	Brazil	624.0
6	2022-01-01	Canada	7584.0
7	2023-01-01	Canada	3468.0
8	2021-01-01	Canada	180.0
9	2022-01-01	France	7836.0
10	2023-01-01	France	3708.0
11	2021-01-01	France	156.0
12	2022-01-01	Germany	6996.0
13	2023-01-01	Germany	4116.0
14	2022-01-01	Italy	5460.0
15	2023-01-01	Italy	2304.0
16	2022-01-01	Mexico	5076.0
17	2023-01-01	Mexico	624.0
18	2021-01-01	Mexico	864.0
19	2022-01-01	Spain	8424.0
20	2023-01-01	Spain	1284.0
21	2022-01-01	United Kingdom	7116.0
22	2023-01-01	United Kingdom	3600.0
23	2021-01-01	United Kingdom	288.0

Forecasted values by year :

Row	forecast_timestamp	Country	forecast_value
1	2024-01-01 00:00:00 UTC	Brazil	7086.327690439...
2	2025-01-01 00:00:00 UTC	Brazil	1437.312800635...
3	2026-01-01 00:00:00 UTC	Brazil	6666.105533950...
4	2024-01-01 00:00:00 UTC	Canada	3693.319769002...
5	2025-01-01 00:00:00 UTC	Canada	7387.254475561...
6	2026-01-01 00:00:00 UTC	Canada	340.7772320413...
7	2024-01-01 00:00:00 UTC	France	3381.867820694...
8	2025-01-01 00:00:00 UTC	France	8088.327297221...
9	2026-01-01 00:00:00 UTC	France	37.49981475726...
10	2024-01-01 00:00:00 UTC	Mexico	5307.268535707...
11	2025-01-01 00:00:00 UTC	Mexico	397.9816506546...
12	2026-01-01 00:00:00 UTC	Mexico	5527.961717428...
13	2024-01-01 00:00:00 UTC	United Kingdom	3260.229742040...
14	2025-01-01 00:00:00 UTC	United Kingdom	7393.269992295...
15	2026-01-01 00:00:00 UTC	United Kingdom	163.4772492804...





Since the data only spanned three years, we created a model to predict revenue on a monthly basis instead. I will use Mexico as an example to explain.

Actual values by month:

104	2021-09	Mexico	24
105	2021-10	Mexico	48
106	2021-11	Mexico	72
107	2021-12	Mexico	72
108	2022-01	Mexico	72
109	2022-02	Mexico	72
110	2022-03	Mexico	72
111	2022-04	Mexico	72
112	2022-05	Mexico	84
113	2022-06	Mexico	110
114	2022-07	Mexico	160
115	2022-08	Mexico	211
116	2022-09	Mexico	286
117	2022-10	Mexico	387
118	2022-11	Mexico	423
119	2022-12	Mexico	423
120	2023-01	Mexico	438
121	2023-02	Mexico	438
122	2023-03	Mexico	438
123	2023-04	Mexico	448
124	2023-05	Mexico	448

Forecasted values for 5 month :

31	2023-06-01 00:00:00 UTC	Mexico	456.899997...
32	2023-07-01 00:00:00 UTC	Mexico	476.981808...
33	2023-08-01 00:00:00 UTC	Mexico	497.063619...
34	2023-09-01 00:00:00 UTC	Mexico	517.145430...
35	2023-10-01 00:00:00 UTC	Mexico	537.227241...

Answering questions

1) How has total monthly revenue changed over time, and what growth patterns can be identified?

Row	Subscription_Mc	Total_Revenue	Avg_Revenue
1	2022-02-07	12.625	12.625
2	2022-03-05	78.8109071...	15.7621814...
3	2022-08-08	87.2013675...	17.4402735...
4	2022-09-05	253.972392...	21.1643660...
5	2022-09-30	259.261977...	21.6051647...
6	2022-08-17	134.439871...	14.9377635...
7	2022-09-17	134.439871...	14.9377635...
8	2022-05-11	24.2147008...	12.1073504...
9	2022-07-01	86.66	21.665
10	2022-07-07	175.903731...	19.5448590...
11	2022-08-20	100.135317...	16.6892195...
12	2023-01-21	186.509041...	16.9553674...
13	2023-02-11	167.286796...	18.5874218...
14	2022-06-18	89.0883333...	17.8176666...
15	2022-09-06	178.148282...	17.8148282...

Revenue saw accelerated growth from Q2 to Q3 2022, likely driven by seasonal campaigns or strategic initiatives, followed by stabilization in 2023, sustaining the higher revenue levels achieved.

2) How does revenue distribution vary by country, and what can be inferred about the market value of each region?

Row	Country_Name	Total_Revenue	Avg_Revenue	Total_Customers
1	France	8067.800000000...	13.46878130217...	599
2	United States	7994.800000000...	13.23642384105...	604
3	Canada	7376.033333333...	12.63019406392...	584
4	United Kingdom	7341.666666666...	13.15710872162...	558
5	Germany	6932.366666666...	12.90943513345...	537
6	Brazil	6895.966666666...	13.01125786163...	530
7	Spain	6461.533333333...	13.37791580400...	483
8	Australia	5094.966666666...	13.03060528559...	391
9	Mexico	5030.233333333...	12.99801894918...	387
10	Italy	4956.633333333...	13.54271402550...	366

France and the United States lead in total revenue and customer counts, making them the most valuable markets. Emerging markets like Brazil and Spain show growth potential, while Italy, despite lower customer numbers, stands out with the highest average revenue per customer, indicating opportunities for targeted premium strategies.

< **Identifying Key User Groups Based on Age and Gender** >

We tried to cluster the users to identify key user groups. Since the K-means model groups users based on similarity, we believed it would better capture and represent the characteristics of each group.

Cluster Result:

Row	centroid_id	Feature	Feature_Value
1	1	Age	38.27
2	1	Gender	0.24
3	1	Subscription_Type	0.36
4	1	Device	0.5
5	1	Lifetime_Value	122.62
6	2	Age	34.5
7	2	Gender	0.94
8	2	Subscription_Type	1.55
9	2	Device	1.63
10	2	Lifetime_Value	163.63
11	3	Age	43.36

Evaluation:

Query results

JOB INFORMATION RESULTS CHART JSON

Row	Inertia
1	3.265590622083...

Finding Key user group: (based on LTV)

Query results

JOB INFORMATION RESULTS CHART JSON E

Row	Cluster_ID	Total_Users	Avg_Lifetime_Value
1	2	105	162.2628571428...

Feature of key group users

Query results

SAVE RESULTS

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Cluster_ID	Predominant_Gender	Subscription_Type	Device	Avg_Age
1	2	Male	Standard	Smartphone	35.47826086956...

Answering questions

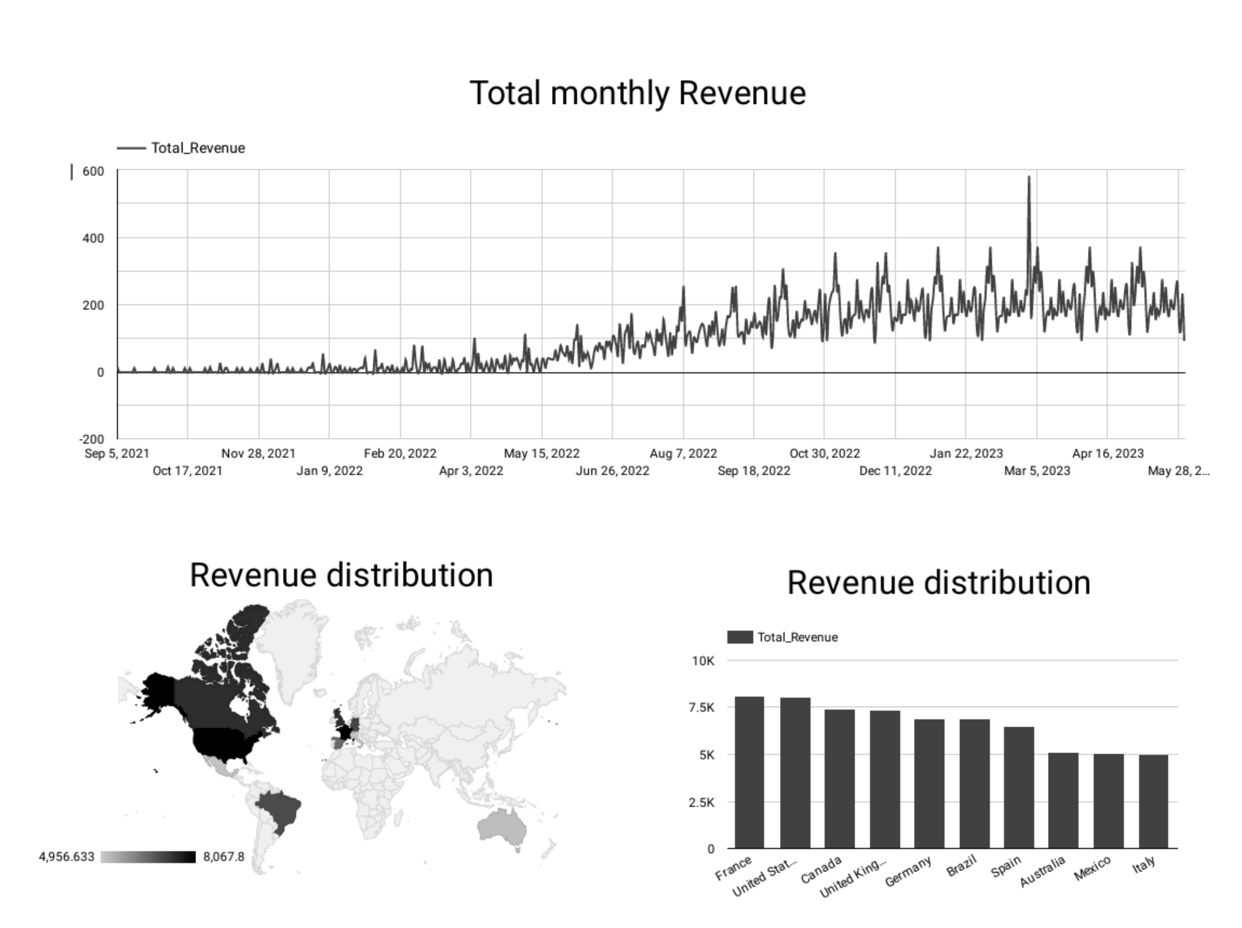
3) What devices are most preferred by the key user groups?

The data indicates that **smartphones** are the most preferred devices among key user groups. This suggests that the majority of users in the group rely on mobile access for the service, emphasizing the importance of optimizing the platform for mobile use

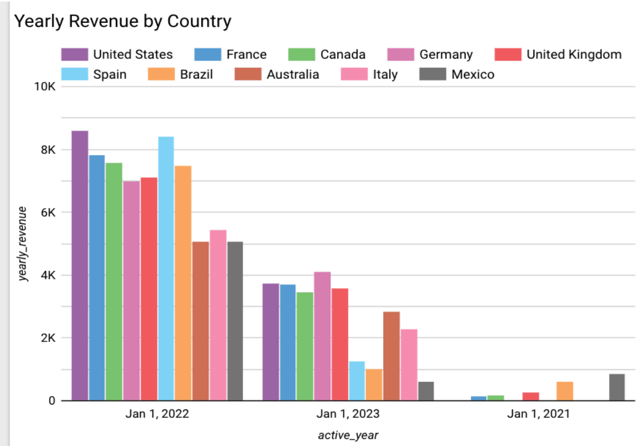
4) Which subscription plans are most popular among the key user groups?

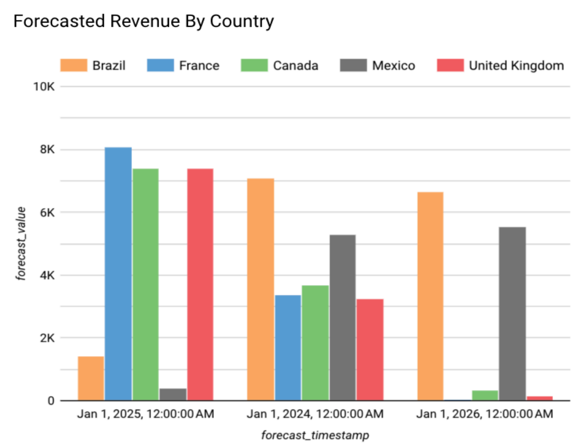
The most popular subscription plan among the key user groups is the **Standard plan**. This likely reflects a preference for mid-tier features or pricing, balancing affordability and functionality.

7. Visualizations

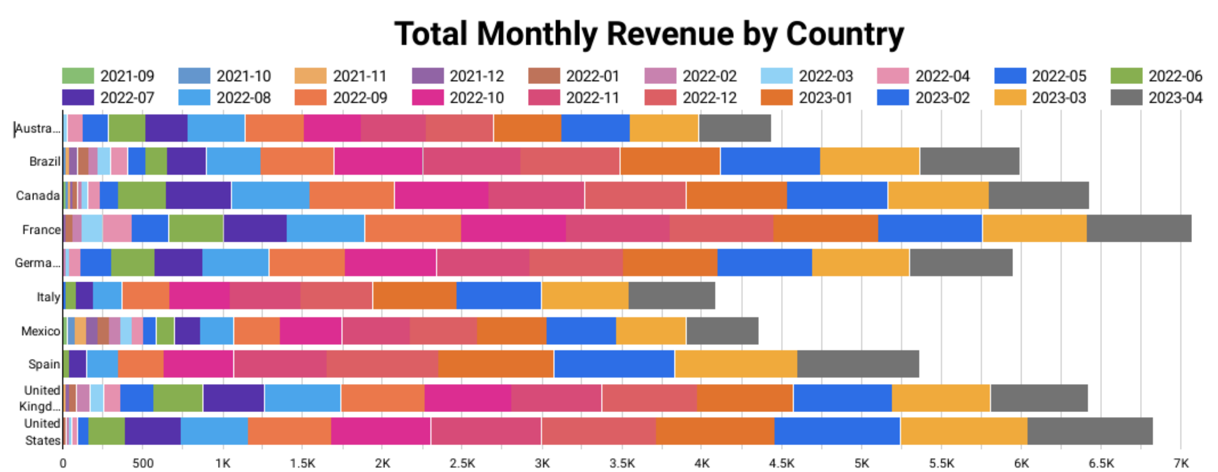


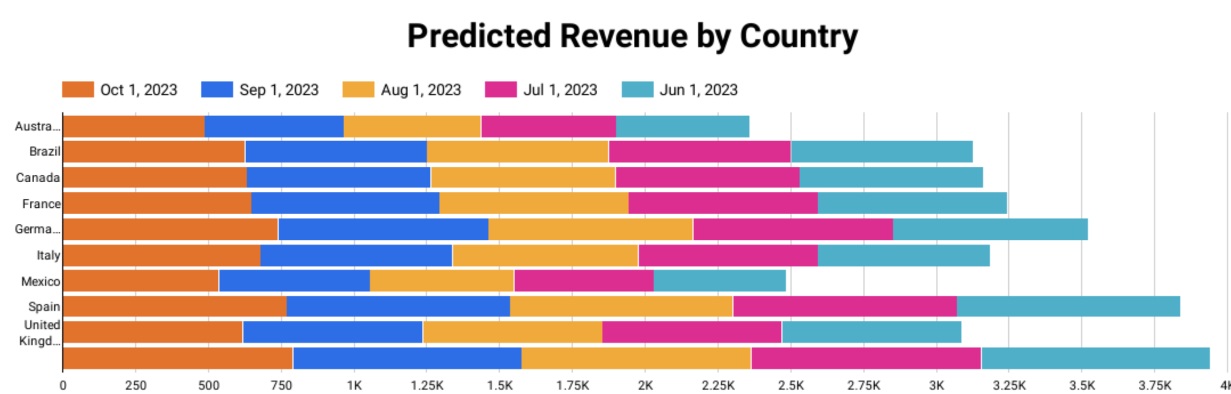
We visualized total revenue over time using Looker Studio to illustrate trends. This visualization can be utilized to analyze the company's revenue performance or identify patterns in revenue flow based on monthly characteristics, providing insights for business applications. We also analyzed revenue by region and represented it using bar charts and map charts.





This model was limited by the amount of years which were tracked in the dataset. Because of this, it was unable to have a meaningful forecast for future years.





A better representation of this visualization would be for each country to get its own chart. This would make it much simpler to see patterns of growth or decline in revenue, with the ability to compare forecasted data with the given data.

Another way to clarify changes in revenue would be to segment the time-slots to be season based(winter, spring summer, fall), rather than month to month. This could help find more meaningful patterns, like finding which season people are more likely to purchase a subscription.