Bans and Bubbles: Analysis of Moderation in VTuber Streams

A Project Report

Presented to

DATA-228-12

Fall, 2023

By

Abdul Sohail Ahmed (016769610)

Poojan Gagrani (016795285)

Kashish Thakur (016919383)

Somna Sattoor (014640769)

Swetha Neha Kutty Sivakumar (016780712)

December 10, 2023

ABSTRACT

Bans and Bubbles: Analysis of Moderation in VTuber Streams

Abdul Sohail Ahmed, Poojan Gagrani, Somna Sattoor, Swetha Neha Kutty Sivakumar, & Kashish Thakur

The rise of Virtual YouTubers (VTubers) has become a prominent trend in the expanding digital and virtual entertainment industry. This research investigates the complex dynamics of VTuber events, with a specific focus on measures related to audience involvement, including live chat conversations, Super Chat donations, and moderating activities. Using a comprehensive dataset obtained from Kaggle, the study thoroughly examines these variables to identify the characteristics that contribute to successful VTuber engagements. By utilizing Azure's powerful array of data tools, the project extensively investigates the data using Azure Databricks and Azure Synapse, resulting in a comprehensive study. The acquired insights seek to aid VTubers in enhancing their content development and engagement strategies, while also enabling viewers to navigate the VTuber ecosystem more efficiently. Expected results encompass the cultivation of a sophisticated comprehension of audience preferences and behaviors, resulting in more tailored and influential viewer experiences. The project aims to improve the digital interface between VTubers and their audiences, thereby promoting the long-term viability and expansion of this pioneering type of entertainment.

Acknowledgments

The creation of this study was greatly aided by Professor Andrew H. Bond and TA Venkatasai Rao Dheekonda, for whom the authors are extremely grateful.

Table of Contents

List of Tables

List of Figures

Chapter 1

1.1 Project Goals and Objectives

The primary aims and objectives of this study are to explore the dynamic realm of VTubers, virtual performers who have captivated a worldwide audience through their online personalities. The project will utilize sophisticated data analysis methods to analyze viewer involvement on several platforms. This analysis will include studying chat interactions, donation patterns, and moderator actions to assess the effectiveness of the material and understand audience attitude. The objective is to create a comprehensive depiction of the factors that motivate viewer engagement and loyalty within the VTuber community, highlighting material that strongly resonates with the audience.

Moreover, the initiative is carefully intended to function as a navigational tool for affiliation agencies, identifying VTubers who continuously create material that resonates positively with the public, thereby revealing opportunities for brand collaborations. The project offers full insights by utilizing Azure's extensive data capabilities, including data input to Azure Data Lake, processing with Azure Databricks, and rich exploratory data analysis in Synapse. The outcome of this analysis will be a website that is easy for users to navigate, displaying condensed insights using PowerBi visuals. This platform aims to not only direct viewers to top-notch VTuber material but also offer rising VTubers valuable data to improve their skills and increase their presence in the competitive realm of digital entertainment.

1.2 Problem and Motivation

The emergence of Virtual YouTubers (VTubers) has introduced a novel era of digital entertainment, captivating global audiences with their distinctive fusion of animation and interactivity. Nevertheless, as this virtual phenomenon expands, the intricacy of comprehending

audience involvement and content impact also increases. This project aims to analyze large volumes of live stream data to reveal trends in user engagement and the effectiveness of content. This project employs Azure's data orchestration technologies to systematically monitor chat participation, Super Chat donations, and moderation events. The goal is to support VTubers in providing impactful content and assist viewers in making informed choices. The objective of this study is to offer practical insights for VTubers and affiliation agencies, to promote constructive content production and establish long-lasting brand collaborations.

The endeavor is motivated by the potential to greatly improve the VTuber ecosystem, making it more captivating for viewers and also more profitable and manageable for content creators. The project aims to provide a strong analytical framework that utilizes cloud-based services to handle and visualize data. This will enable a thorough understanding of the VTuber landscape. This involves creating a flexible and protected framework for handling and displaying data, resulting in a user-friendly website that highlights important discoveries. The motive goes beyond simple analysis; it involves fostering a dynamic online community that flourishes through well-informed decision-making and strategic development, benefiting both viewers and VTubers.

1.3 Project Results and Expected Deliverables

The primary result of this project is the creation of a customized interactive analytics dashboard for VTuber event data, offering comprehensive information on VTuber engagement. The dashboard is an important output that provides a detailed perspective on the dynamics of VTuber streams, encompassing chat interactions and moderating occurrences. Another notable result is the acquired knowledge of content moderation, which unveils the real-time management of VTuber streams to promote constructive interactions.

Expected deliverables from this project are manifold:

1. A Power BI dashboard that visually represents VTuber event data dynamically and interactively.
2. Comprehensive analytical reports providing detailed insights about VTuber engagement and trends of content moderation.
3. An efficient and reliable implementation of Google OAuth Single Sign-On for secure and uncomplicated access to the dashboard.
4. A website designed using the .NET Framework that is easy to use, allowing users to easily explore and understand the presented information.

This extensive collection of deliverables will function as a significant asset for comprehending and promoting constructive content generation and audience involvement within the VTuber community.

Chapter 2 Background and Related Work

2.1 Background and Used Technologies

Background

The project is focused on the concept of Vtubers, which is a virtual YouTube community, where online entertainers use a virtual avatar in place of their actual identity to interact with the users. During online streaming, Vtubers connect with their fans, using Chats and Superchats. Chats are the general unpaid medium of interaction used by the fans, whereas Super chats are the medium of interaction used by fans with paid memberships. The dataset used in the project consists of Vtuber content consisting of a large quantity of data about Chats, Super chats, and moderation events like Deletion and Ban events.

Technologies Used

All the project tasks are accomplished using the data orchestration tools of the Azure ecosystem, which are explained below:

1. **Azure Data Lake:** The data is ingested into the Data Lake, which enables it to perform Big Data analytics. It allows the unlimited storage of structured, semi-structured, and unstructured data of any size. This property of Azure Data Lake makes it suitable for big data projects.
2. **Azure Databricks:** It is an Apache Spark-based analytics platform, which has a lake house architecture that along with Databricks SQL helps in performing data warehousing tasks. It is used to perform Data Cleaning and Preprocessing, with the help of Python data frames. All the null values are removed and missing values are either imputed or removed based on the context of the information.
3. **Azure Synapse Analytics:** It is an integrated analytics service, which is used for performing exploratory data analysis. This analysis is used for initial data understanding and for identifying the outliers in the data. The insights obtained can be used to perform further data analytics while creating visualization dashboards.
4. **Power BI:** It is an interactive visualization dashboard service offered by Microsoft which is integrated with other Azure services. It enables interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their reports and dashboards.
5. **DotNet Framework:** It is a software development framework that is used to build the web application.

6. **GoogleOAuth:** It is a security feature that allows a user to successfully sign in to several linked but independent software systems with a single ID. It is used to sign in to the website via Gmail ID.

2.2 Literature Survey

Kim and Yoo (2021) have done a comparative study between the user experience of existing general and virtual YouTubers. The data is collected using a survey, which is prepared by the AttrakDiff evaluation method. This questionnaire is used for quantitative analysis of four dimensions of usability which are pragmatic quality, hedonic quality-identity, hedonic quality-simulation, and attractiveness with 28 items (pairs of opposite adjectives positioned at both ends) rated on a 7-point scale (-3: negative adjective ~ 3: positive adjective). The survey was collected from a total of 50 people (25 men and 25 women) in their 20s to 30s. The study aims to further identify the preferences and features of the user experience using AttrakDiff evaluation techniques from the simple rating comparison for existing virtual YouTubers (Vtubers).

The research done by Advantia and NoHikari (2023) aims to explore the factors that influence the growth of subscribers to Virtual YouTubers (VTubers) in Vietnam. The dataset includes an intake of 97 VTubers chosen at random, and the study uses regression analysis to examine the link between the number of videos, views, and subscribers. According to the study's findings, increasing the number of video uploads does improve the number of subscribers, however the number of views has a direct impact on the number of subscribers Hence the view count is identified to be a high impact factor in increasing the subscriber count. The main goal of the study is to understand the different growth strategies for Vtubers in the context of Digital media, majorly focusing on the Vietnamese market.

Mamat et al. (2022) performed a study to identify Malaysian youth's interest in anime culture and its subcultures such as VTube. Since the 1990s, anime, manga, and drama series have been among the most popular and well-received forms of Japanese popular culture, and their translation into multiple languages has made them globally recognized. An online survey utilizing Google Forms was completed by 104 people. Respondents were Japanese language students from two Malaysian public universities. Some of the responders are members of a Japanese Cultural Club at one of the universities. The survey was divided into three sections: respondents' background, respondents' interest in anime culture, and their interest in VTubers. The results showed that the most popular culture is anime, followed by manga, songs, and seiyuu (voice actor/actress). The study identifies fantasy as the most popular anime genre, followed by humor, love stories (romance), and mystery. The majority of respondents had heard the term "VTuber" from anime, friends, or internet outlets such as YouTube. The statistics show that VTubers are an important and well-known product of Japanese culture among Japanese culture lovers and that they will certainly gain widespread momentum in the future.

Chapter 3 Data Overview

3.1 Data Description

The VTubers dataset is obtained from Kaggle, where the data is uploaded by a reliable source. The data is in the form of Parquet files, which is an efficient column storage format. The size of the data is about 13GB consisting of information related to Vtuber Channel, Vtuber Chat data, Vtuber Superchat Data, Vtuber Chat Statistics, Vtuber Super Chat Statistics, and information about Ban and Deletion Events. The dataset collected covers three months in the year 2020. The data is ingested into the Azure Data Lake, where all the collected Parquet files

are stored and fetched for further Data Processing. As described earlier the complete dataset consists of seven files. The following tables describe the columns in each file.

The Table below shows the Column names, data type, and Description of the corresponding columns of the Channel table.

Table 1: Channel table

Column Name	Type	Description
channelId	string	A unique identifier for each channel
name	string	Name of the Channel
englishName	nullable string	Name of the Channel in English
affiliation	string	Channel Affiliation
group	nullable string	Group
subscriptionCount	number	The count of the subscriptions
videoCount	number	Total number of videos uploaded by the channel
photo	string	Channel Icon

The Table below shows the Column names, data type, and Description of the corresponding columns of the Chat Data table.

Table 2: Chat Data table

Column Name	Type	Description
timestamp	string	ISO 8601 UTC timestamp
authorChannelId	string	author channel id
bodyLength	number	Chat message length

isMember	nullable boolean	is a member
videoId	string	source video id
channelId	string	source channel id

The Table below shows the Column names, data type, and Description of the corresponding columns of the Superchat Data table.

Table 3: Superchat Data table

Column Name	Type	Description
timestamp	string	ISO 8601 UTC timestamp
authorChannelId	string	author channel id
amount	number	purchased amount
currency	string	three-letter currency symbol
significance	number	significance
channelId	string	source channel id

The Table below shows the Column names, data type, and Description of the corresponding columns of the Deletion Events table.

Table 4: Deletion Events table

Column Name	Type	Description
timestamp	string	UTC timestamp
id	string	chat id

retracted	boolean	is deleted by author oneself
videoId	string	source video id
channelId	string	source channel id

The Table below shows the Column names, data type, and Description of the corresponding columns of the Bans Events table.

Table 5: Ban Events table

Column Name	Type	Description
timestamp	string	UTC timestamp
authorChannelId	string	channel id
videoId	string	source video id
channelId	string	source channel id

The Table below shows the Column names, data type, and Description of the corresponding columns of the Chat Statistics table.

Table 6: Chat Statistics table

Column Name	Type	Description
channelId	string	A unique identifier for each channel
period	string	Interested period (%Y-%M)
chats	number	Number of chats
memberChats	number	Number of chats with membership status attached

uniqueChatters	number	Number of unique chatters
uniqueMembers	number	Number of unique members appeared on the live chat
bannedChatters	number	Number of unique chatters marked as banned by mods
deletedChats	number	Number of chats deleted by mods

The Table below shows the Column names, data type, and Description of the corresponding columns of the Super Chat Statistics table.

Table 7: Super Chat Statistics table

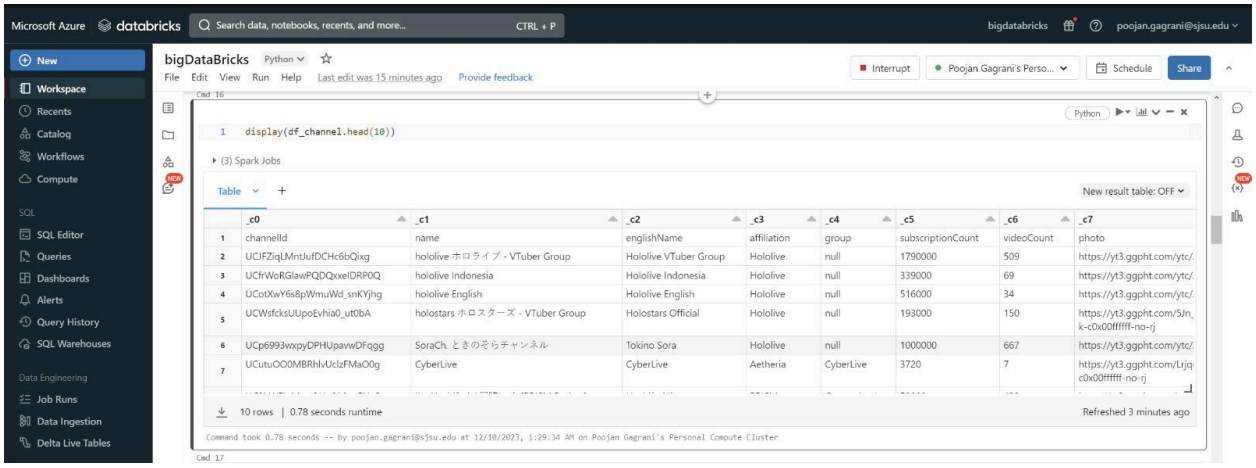
Column Name	Type	Description
channelId	string	A unique identifier for each channel
period	string	Interested period (%Y-%M)
superChats	number	Number of super chats
uniqueSuperChatters	number	Number of unique super chatters
totalSC	number	Total amount of super chats
averageSC	number	The average amount of super chat
totalMessageLength	number	Total message length
averageMessageLength	number	Average message length
mostFrequentCurrency	string	Most frequent currency
mostFrequentColor	string	Most frequent color

The Collected files are further merged into 4 separate files based on the context and relevance of the information available in them:

- Merged Data Files for Channel, Chats, Ban, and Deletion files
- Merged Data File for Channel and Chat Statistics File
- Merged Data File for Channel and Super Chat Statistics File
- Merged Data File for Channel and Super Chat File

3.2 Dataset Samples

The following figures show the sample VTubers raw data collected from Kaggle.



	_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7
1	channelId	name	englishName	affiliation	group	subscriptionCount	videoCount	photo
2	UCJFZiqLMntJufDCHc6bQixg	hololive ホロライブ - VTuber Group	Hololive VTuber Group	Hololive	null	1790000	509	https://yt3.ggpht.com/ytc/
3	UCfrWoRGlawPQDQxxeIDRP0Q	hololive Indonesia	Hololive Indonesia	Hololive	null	339000	69	https://yt3.ggpht.com/ytc/
4	UCotXwY6s8pWmuWd_snKYjhg	hololive English	Hololive English	Hololive	null	516000	34	https://yt3.ggpht.com/ytc/
5	UCWsfcksUUpoEvhia0_ut0bA	holostars ホロスターズ - VTuber Group	Holostars Official	Hololive	null	193000	150	https://yt3.ggpht.com/5Jn_k-c0x00ffffff-no-rj
6	UCp6993wxpyDPHUpavwDFqgg	SoraCh. ときのそらチャンネル	Tokino Sora	Hololive	null	1000000	667	https://yt3.ggpht.com/ytc/
7	UCutuOO0MBRhlvUclzFMaO0g	CyberLive	CyberLive	Aetheria	CyberLive	3720	7	https://yt3.ggpht.com/Ltjq c0x00ffffff-no-rj

Fig. 1. Sample from the unprocessed Channel dataset

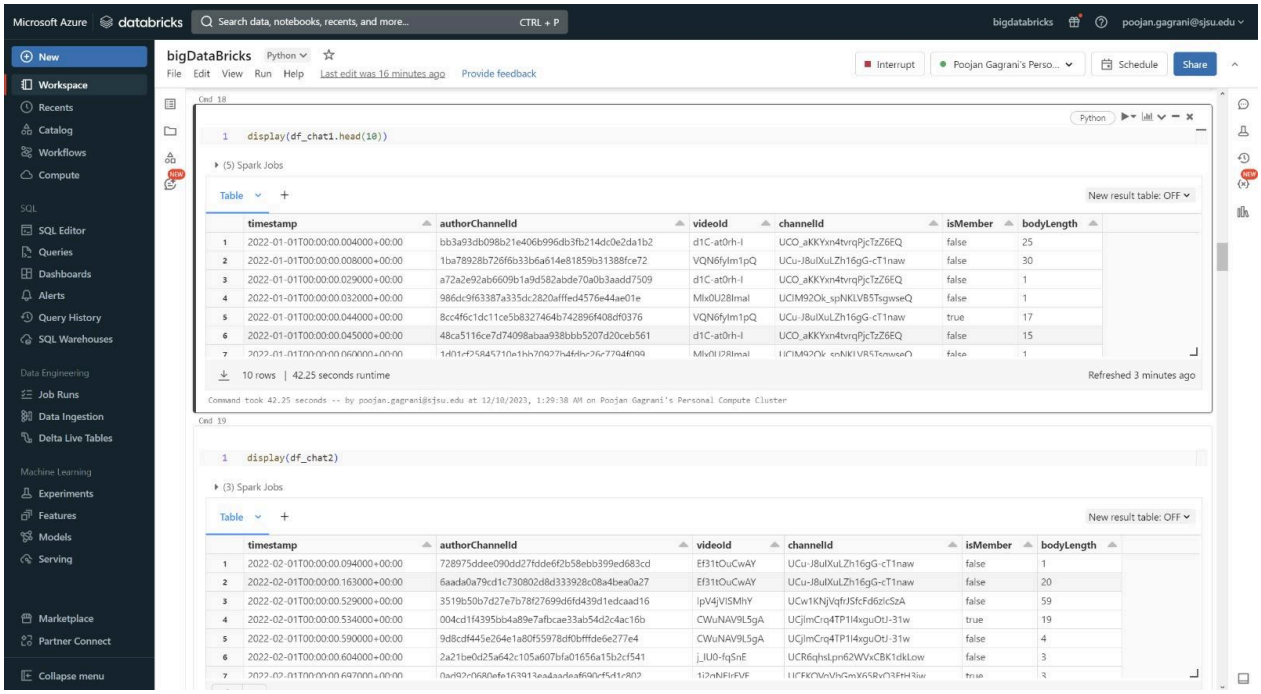


Fig. 2. Sample from the unprocessed Chat dataset

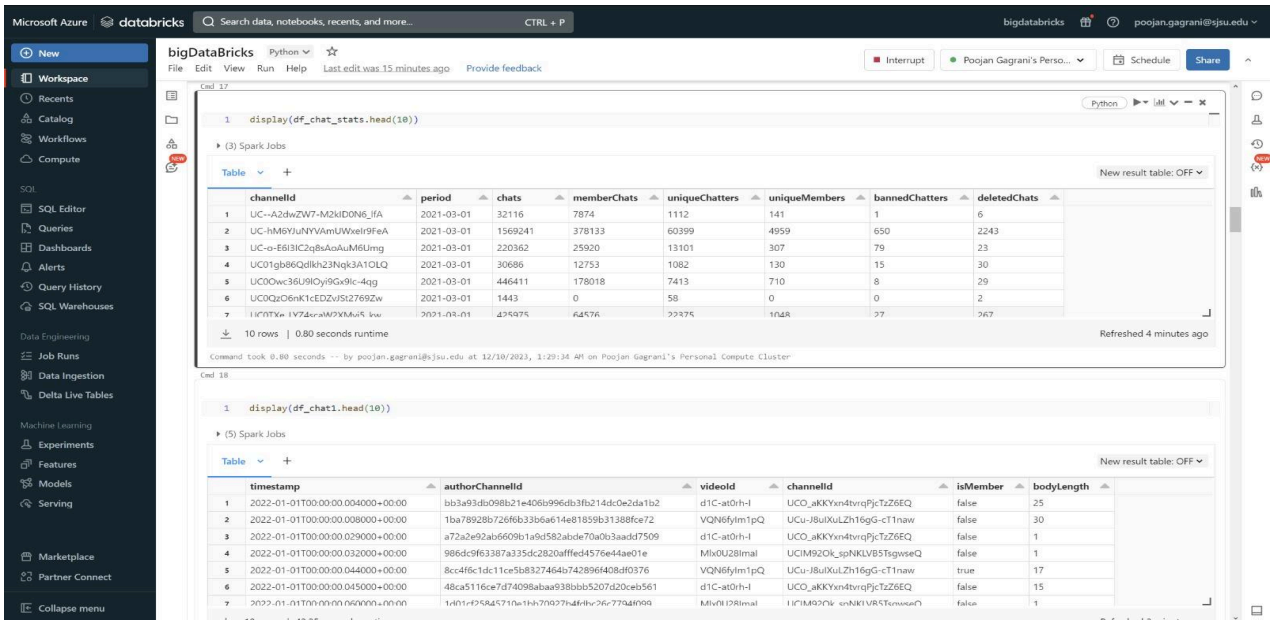


Fig. 3. Sample from the unprocessed Chat Statistics dataset

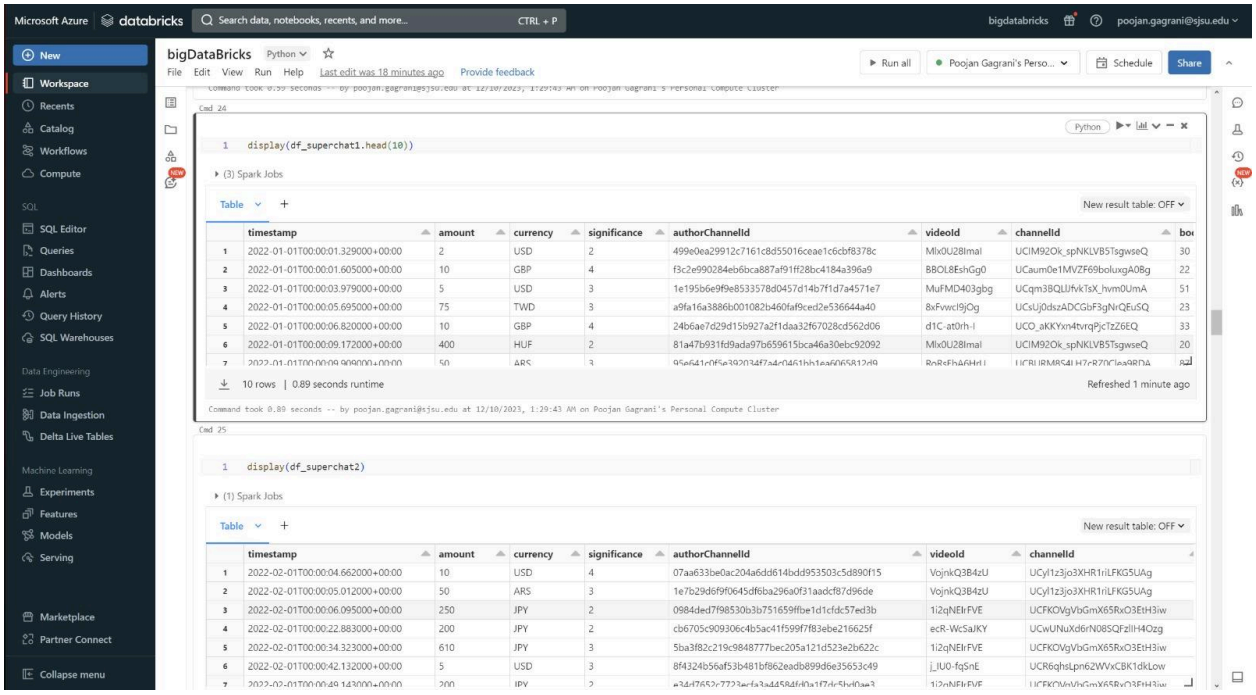


Fig. 4. Sample from the unprocessed SuperChat dataset

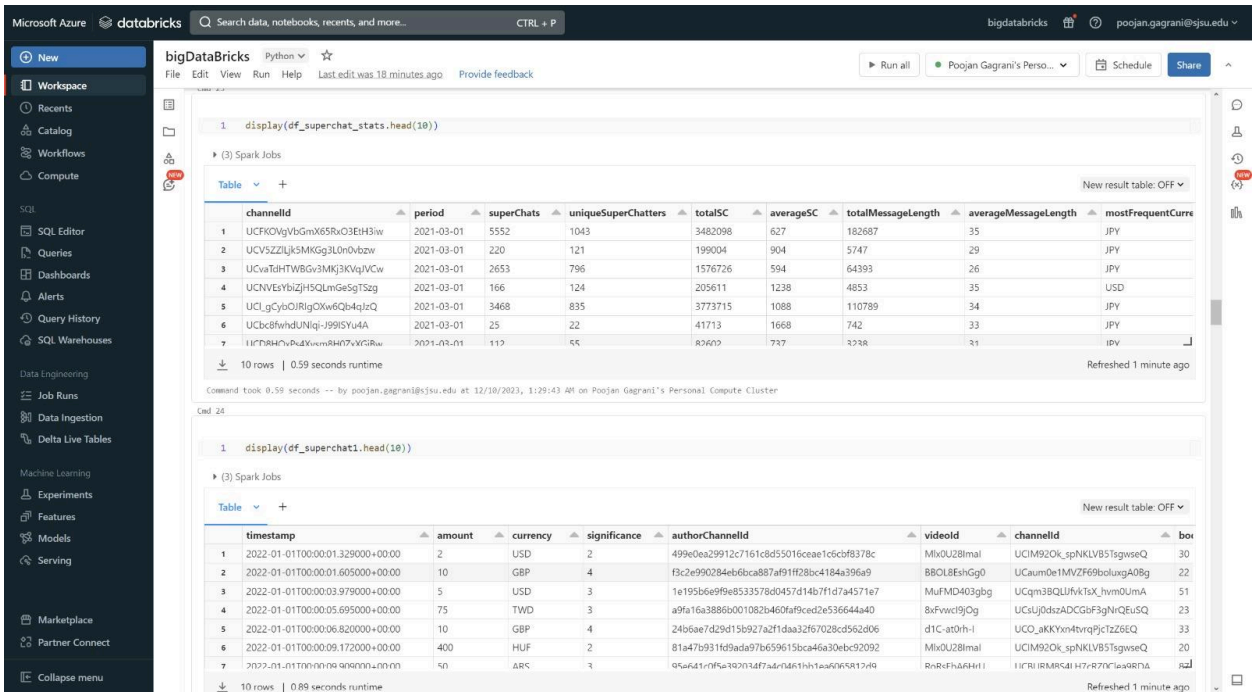


Fig. 5. Sample from the unprocessed SuperChat Statistics dataset

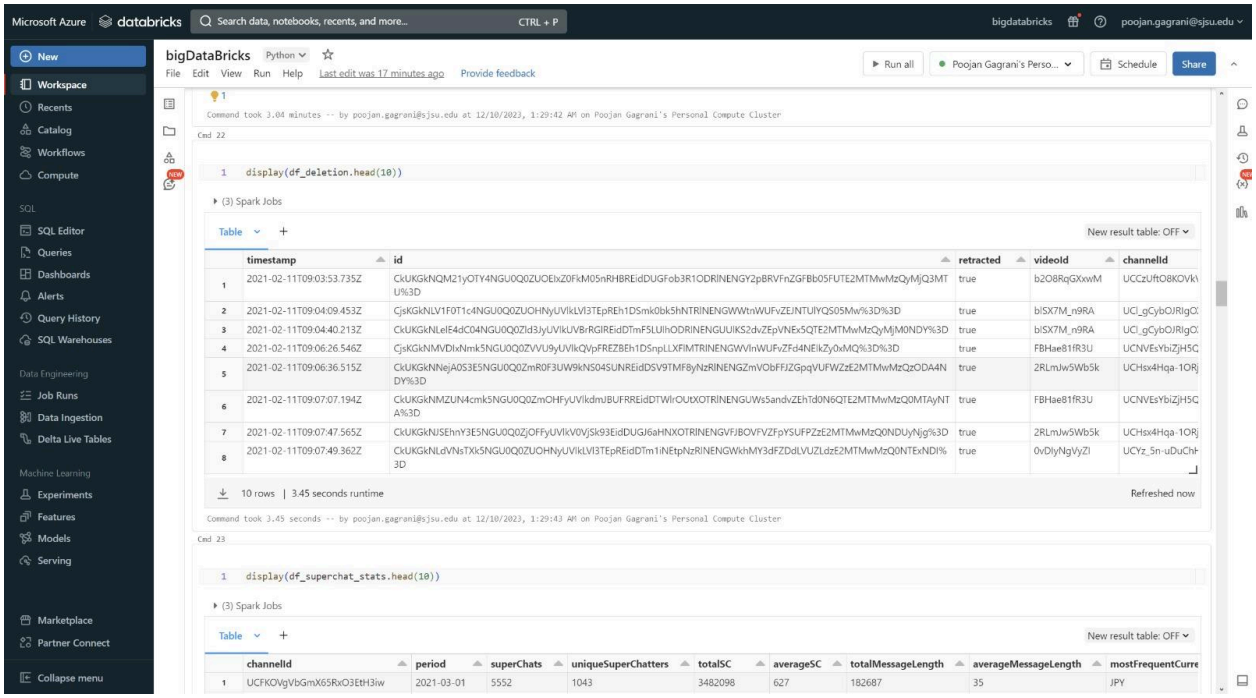


Fig. 6. Sample from the unprocessed Deletion Events dataset

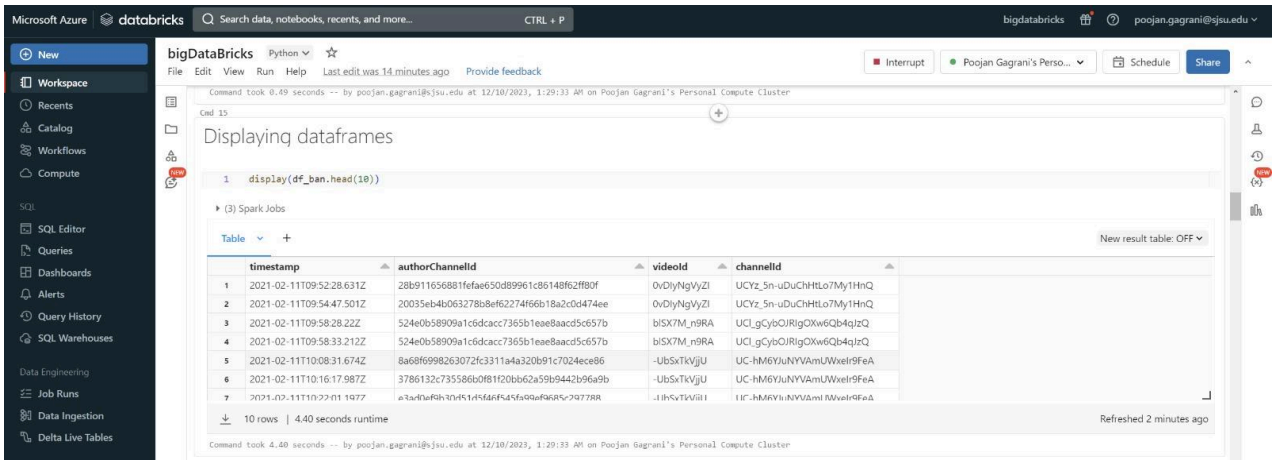


Fig. 7. Sample from the unprocessed Ban Events dataset

Chapter 4 Architecture and Project Flow

4.1 Architecture

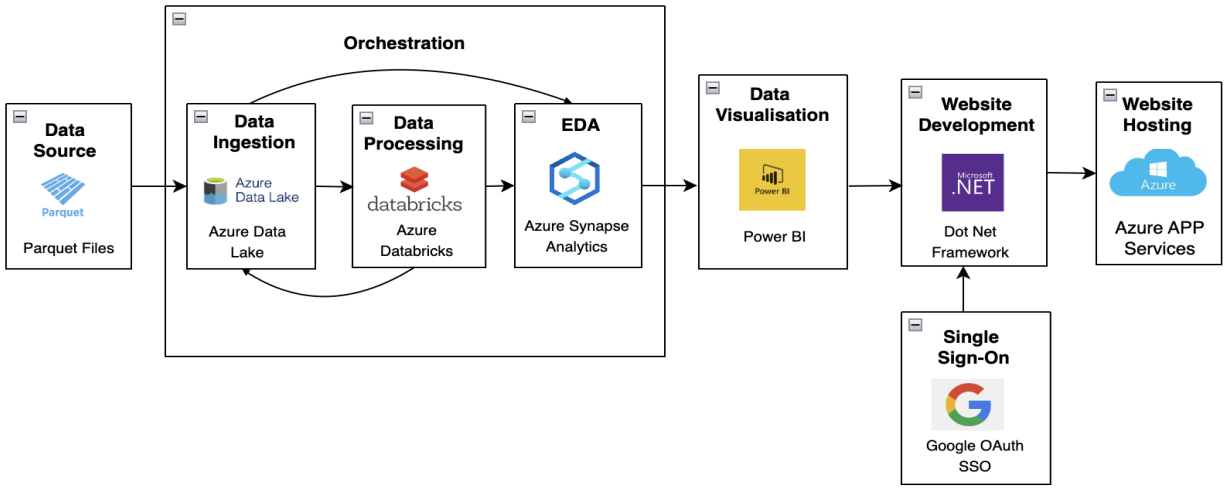


Fig. 8. Diagram of the Project Architecture

4.2 Data Modeling and ER Diagram

The merged files were created in the Azure Data Lake following which a series of data cleaning operations were performed in Azure Databricks. The data model designed for Vtubers data consists of four-dimensional tables and one fact table. The dimensional table includes the SuperChats, Chats, SuperChat statistics, and Chat Statistics data. Apart from the Super Chat Statistics table, all three tables have a mandatory one-to-many relationship with the Channel fact table on the one side and the dimensional table on the many sides. The channel fact table consists of quantitative data such as the total number of videos and subscription counts. Each dimensional table provides context to the channel as it contains descriptive attributes. The dimensional tables maintain a unique identifier column to determine every instance of dimension and are connected to the channel fact table by the ChannelID column. The below figure represents the data modeling designs of this study.

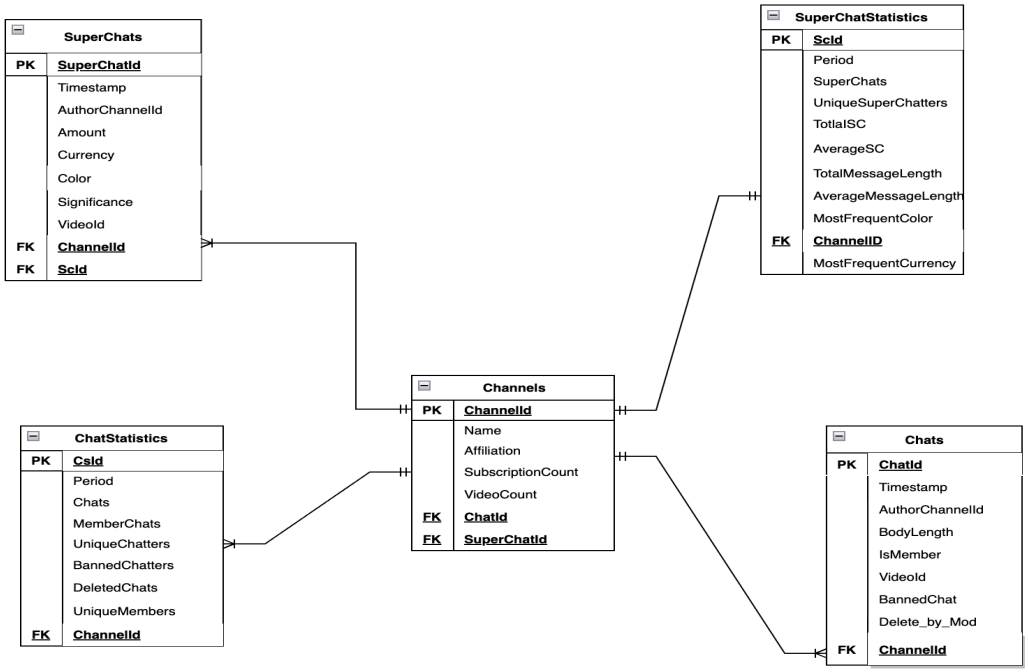


Fig. 9. ER Diagram

4.3 Project Flow

The project is implemented by utilizing a comprehensive set of tools within the Azure environment for effective data flow. The Azure Data Lake serves as the starting point for ingesting a humongous volume of data extracted from Kaggle. After importing the data to Data Lake, rigorous steps of pre-processing the data are performed using Azure Databricks.

The preprocessed data is then used to understand the underlying pattern, find outliers, and examine the characteristics of the data. The analytical results are then visualized using PowerBI dashboards. The dashboard shows the in-depth analysis and is interactive to further filter the visualization as required. The visualizations are then published to the website using the Dot Net framework. The project incorporates Google OAuth Single Sign-On (SSO) to ensure safe and efficient user authentication. This way users can seamlessly access the website. Finally, the built

web applications are deployed in a dependable and scalable manner through the use of Azure Web Services for website hosting. The overall data flow of the project reflects an integrated approach for data management, analysis, visualization, and website development in the Azure ecosystem. The table below shows the list of tools used in this project along with the corresponding purpose for each tool.

Table 8: Tools used along with the purpose

Tool	Purpose
Azure Data Lake	For Scalable Data Ingestion and allows storage of vast amounts of data
Azure Databricks	Used for large-scale data processing
Azure Synapse Analytics	Exploratory Data Analysis
Power BI	Analysis and Visualizations
Dot Net Framework	Website Development
Google OAuth SSO	Secure and Streamlined User Authentication.
Azure Web Services	Website Hosting

Chapter 5 Data Engineering

5.1 Importing Data

The raw parquet files are first loaded into the Azure data lake. The data files loaded contained Vtubers ban events, channel details, chat statistics details, deletion events, super chat statistics, and super chat details.

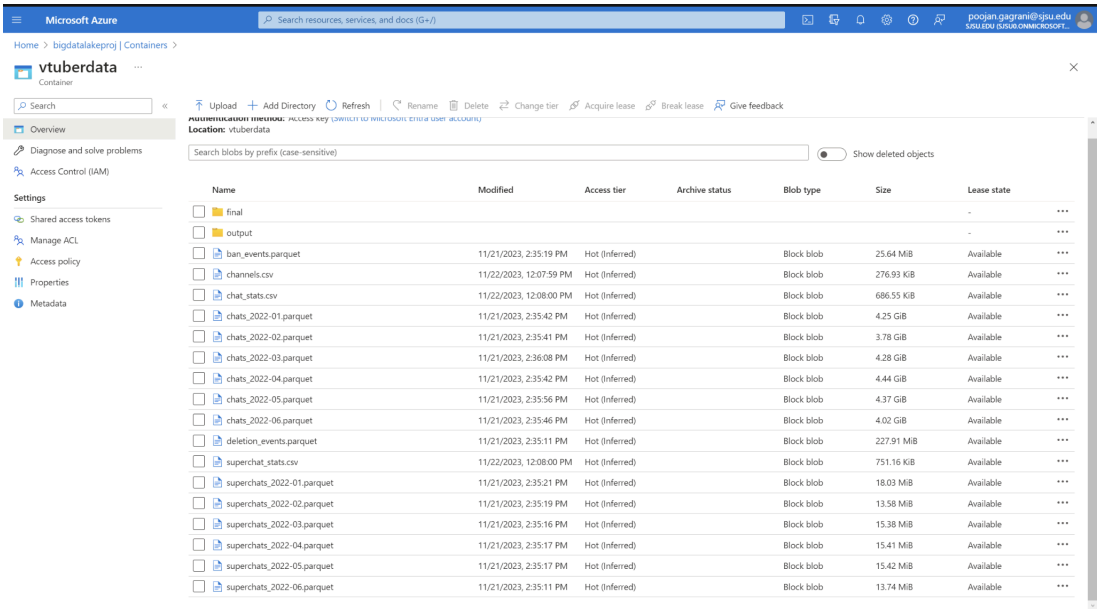


Fig. 10. Raw data files loaded into Azure data lake container

5.2 Data Cleaning

It is a crucial phase in any data analytics project and involves various methods that help ensure data consistency and integrity. All the raw Vtuber data was loaded into Azure Databricks and then the data was read using spark data frames and following data cleaning operations were performed. The data cleaning process involves the following steps.

1. Setting up Azure Databricks spark notebook and reading files
2. Filtering and reducing ban data from Jan to Feb 2022
3. Filtering and reducing chat stats from Jan to Feb 2022
4. Filtering and reducing deletion from Jan to Feb 2022
5. Filtering and reducing superchat stats from Jan to Feb 2022
6. Filtering and reducing superchat stats from Jan to Feb 2022
7. Dropping Name, Group, and Photo columns from channel
8. Mapping the color for the significance field in Superchat

9. Generating chatId for chat
10. Generating superChatId for a chat

The data cleaning code snippets are attached below.

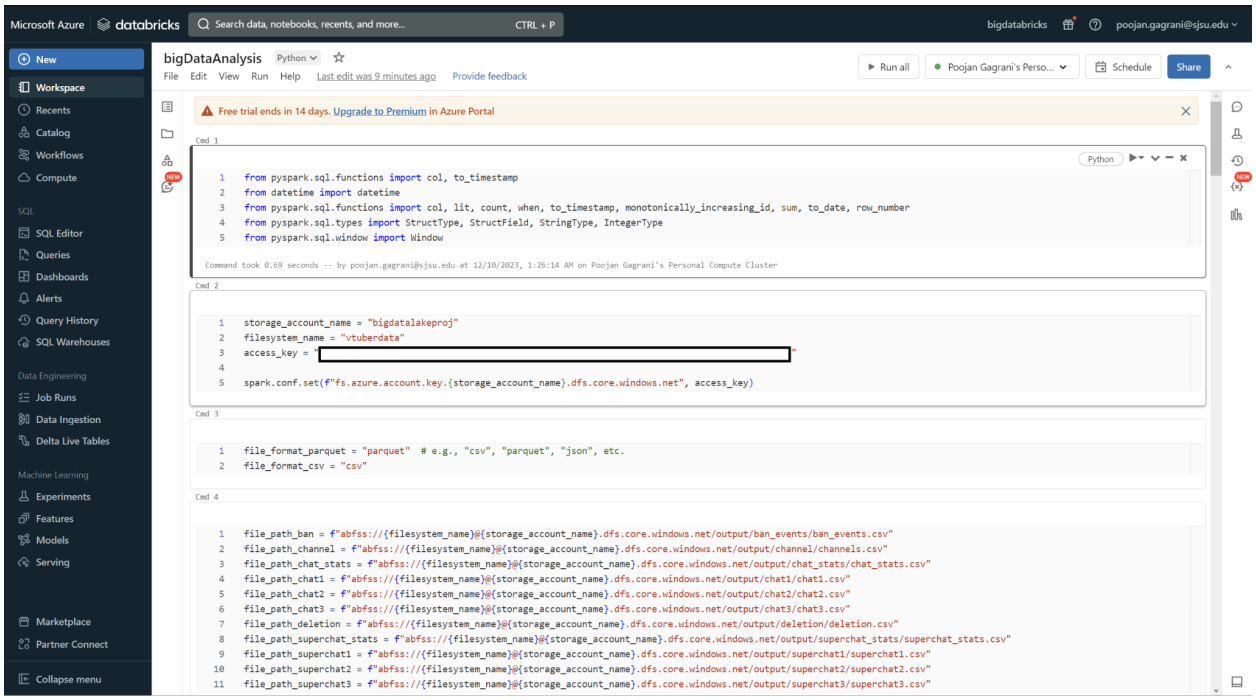


Fig. 11. Setting up Azure Databricks spark notebook and reading files

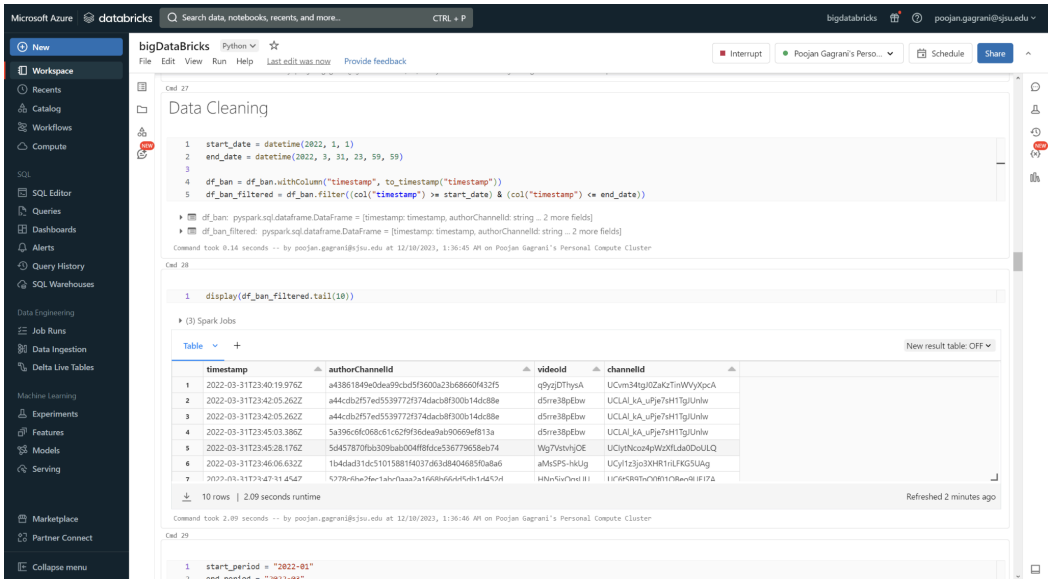


Fig. 12. Filtering and reducing ban data from Jan to Feb 2022

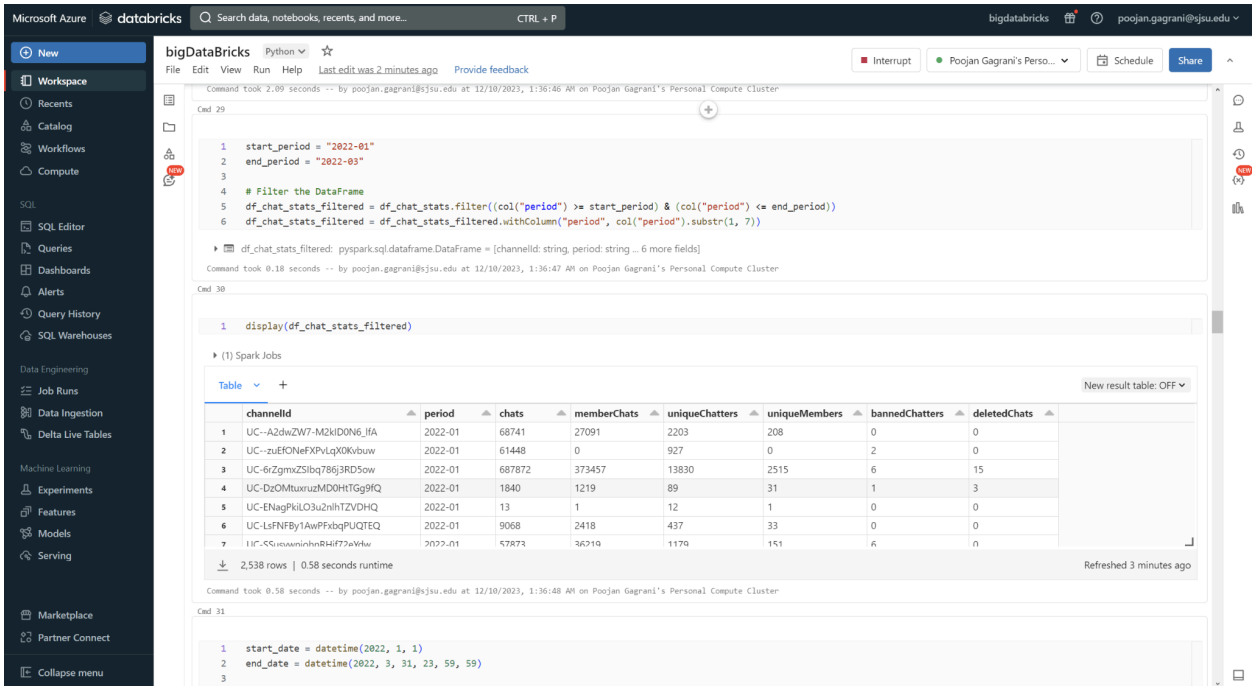


Fig 13. Filtering and reducing chat stats from Jan to Feb 2022

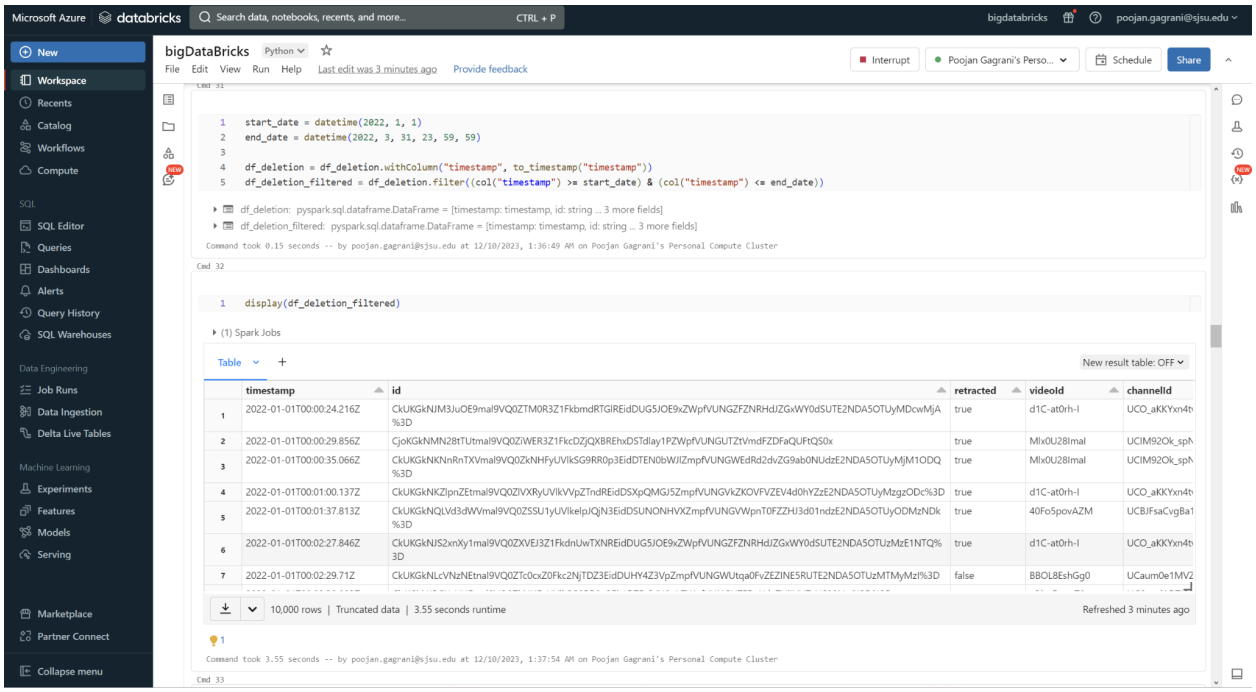


Fig. 14. Filtering and reducing deletion from Jan to Feb 2022

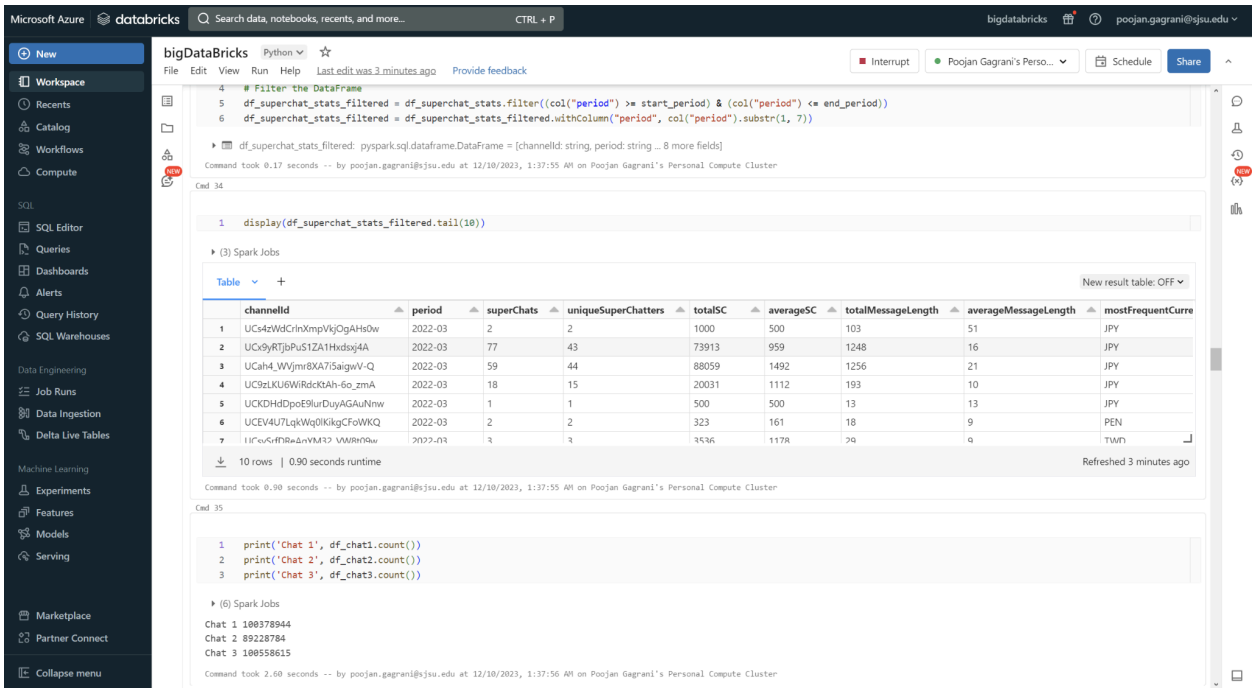


Fig. 15. Filtering and reducing superchat stats from Jan to Feb 2022

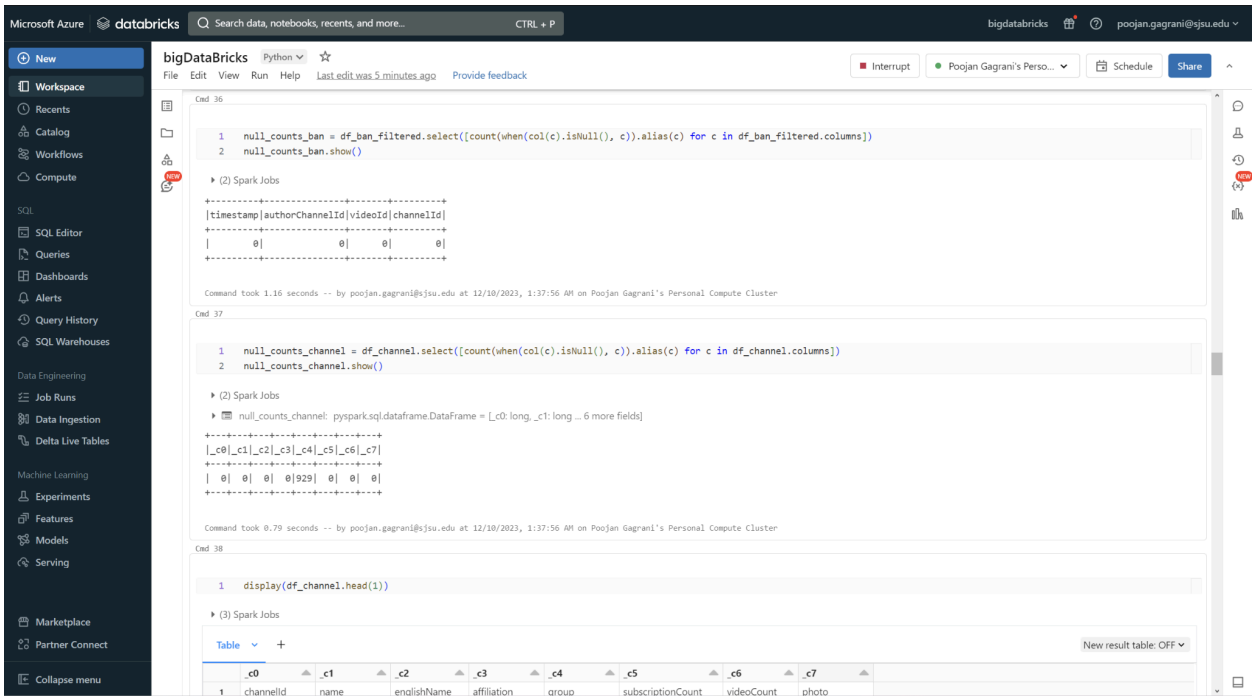


Fig. 16. Filtering and reducing superchat stats from Jan to Feb 2022

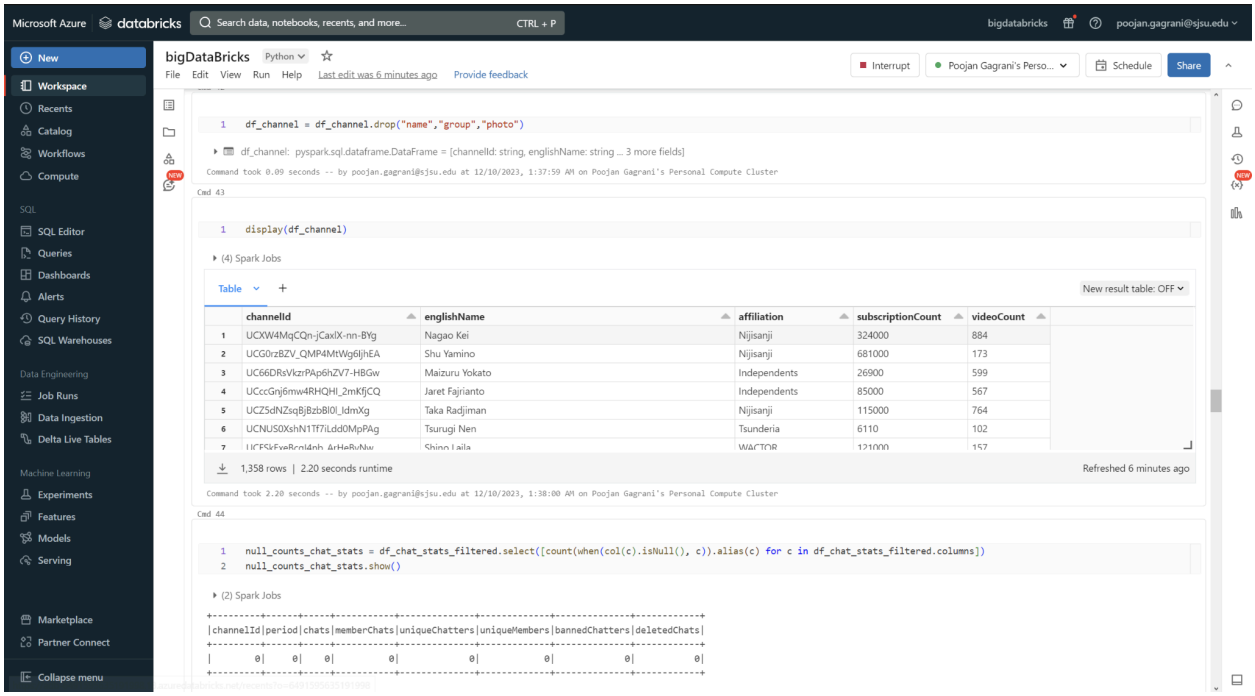


Fig. 17. Dropping Name, Group, and Photo columns from channel

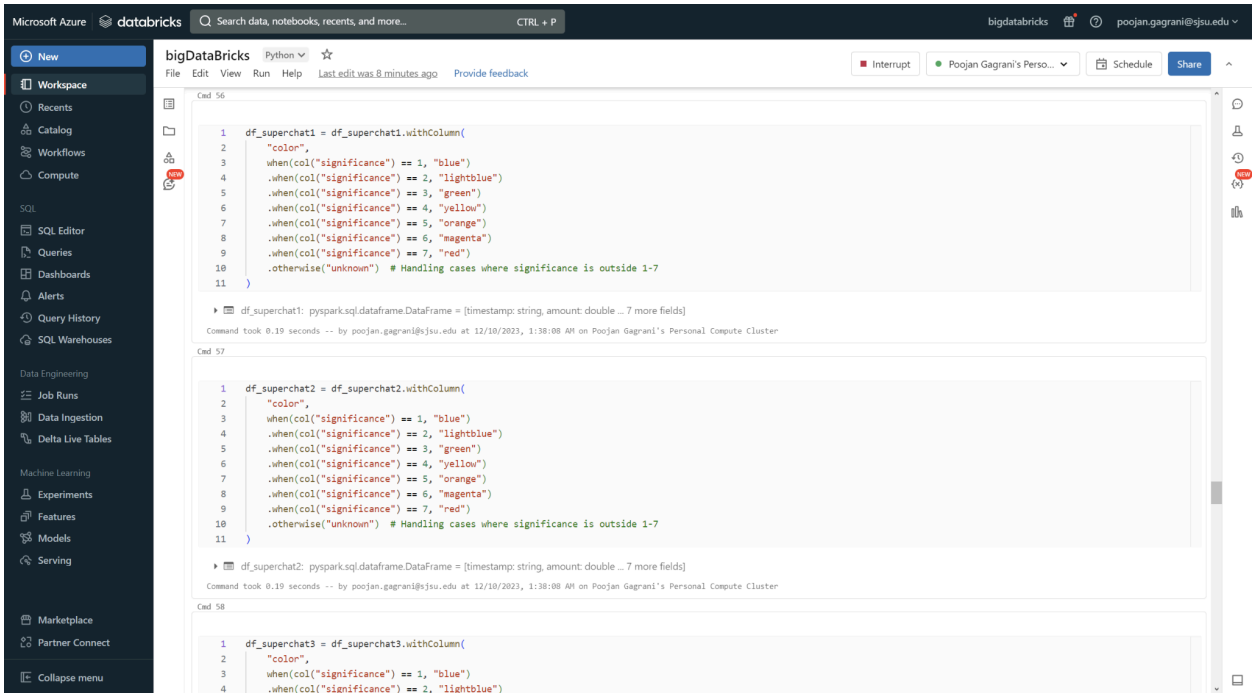


Fig. 18. Mapping the color for the significance field in superchat

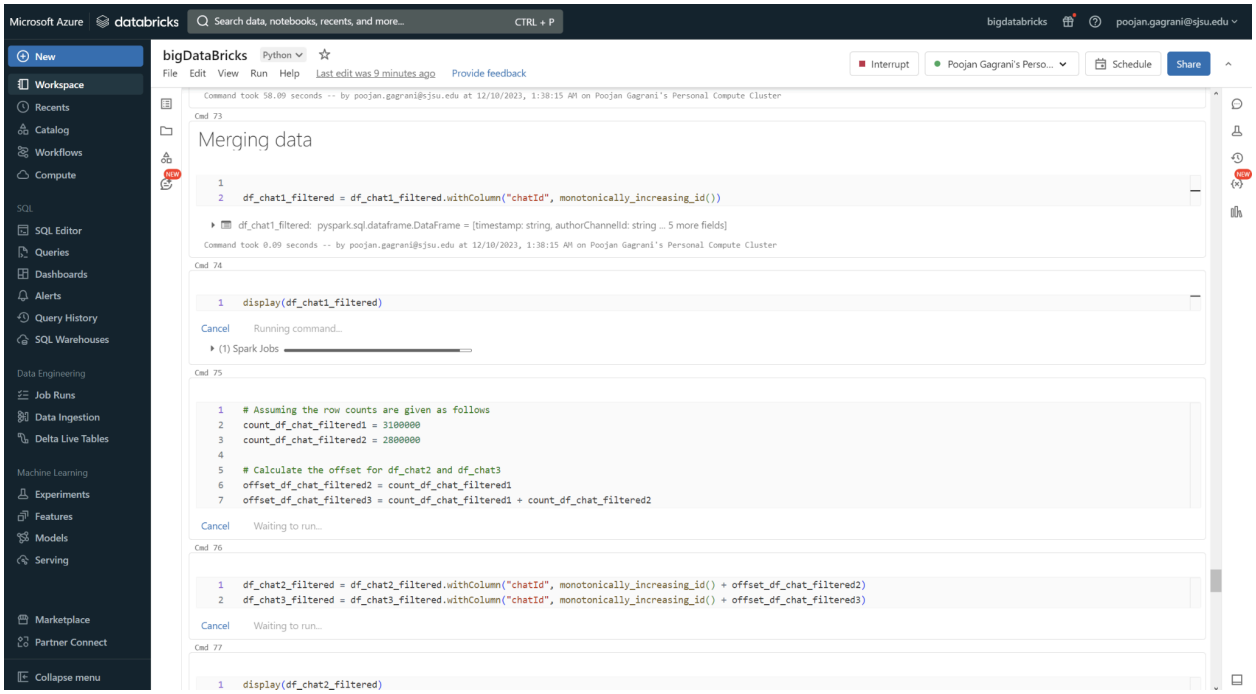


Fig. 19. Generating chatId for chat

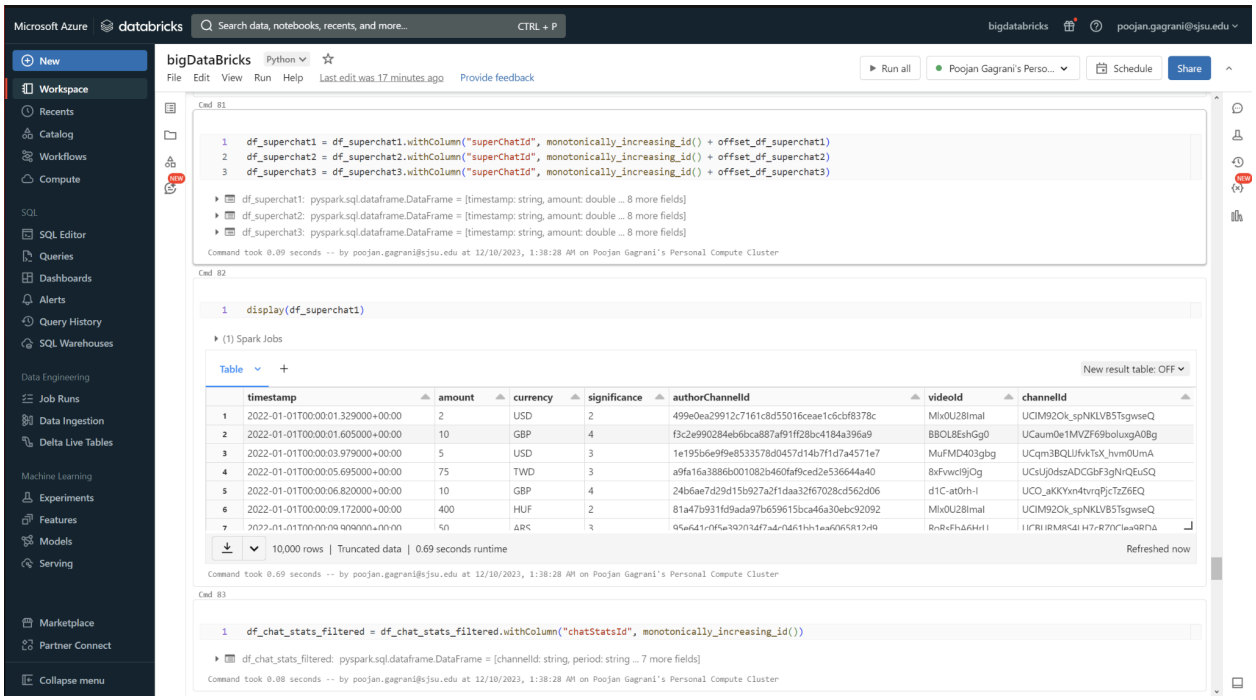


Fig. 20. Generating superChatId for chat

5.3 ETL process

The ETL pipeline was built in with the source being Datalake the data was loaded into Databricks where transform operations were performed, and a schema was achieved having four tables Channels, Chats, Chat Statistics, SuperChat, Superchat Statistics, and the destination was Azure Synapse Analytics where the Load operation was done.

On successful completion of the initial data cleaning the data was then extracted from the data lake using spark dataframes on Azure Databricks and the data was merged hence the transformations were applied where the final output tables were Channels, Chats, Chat Statistics, SuperChat, Superchat Statistics.

a. Extracting Data

The data is extracted from the data lake using the Azure data bricks. The data is read into spark data frames and further transformations are applied.

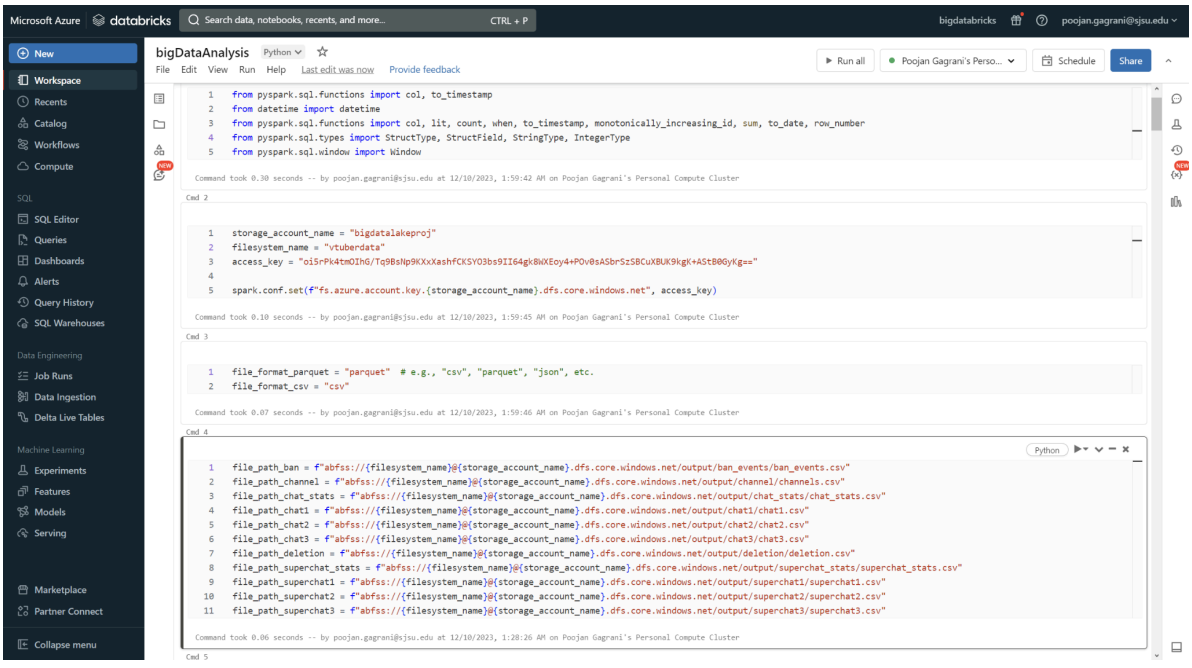


Fig. 21. Extracting cleaned files from the pipeline

b. Transforming Data

The data transformations are applied to where the Chat and Superchat data are merged first, then after merging the chat and deletion events data frames are combined, and later chat and ban events are merged. Finally, the Chat and Superchats are merged with Channel, and the final four resultant tables are created

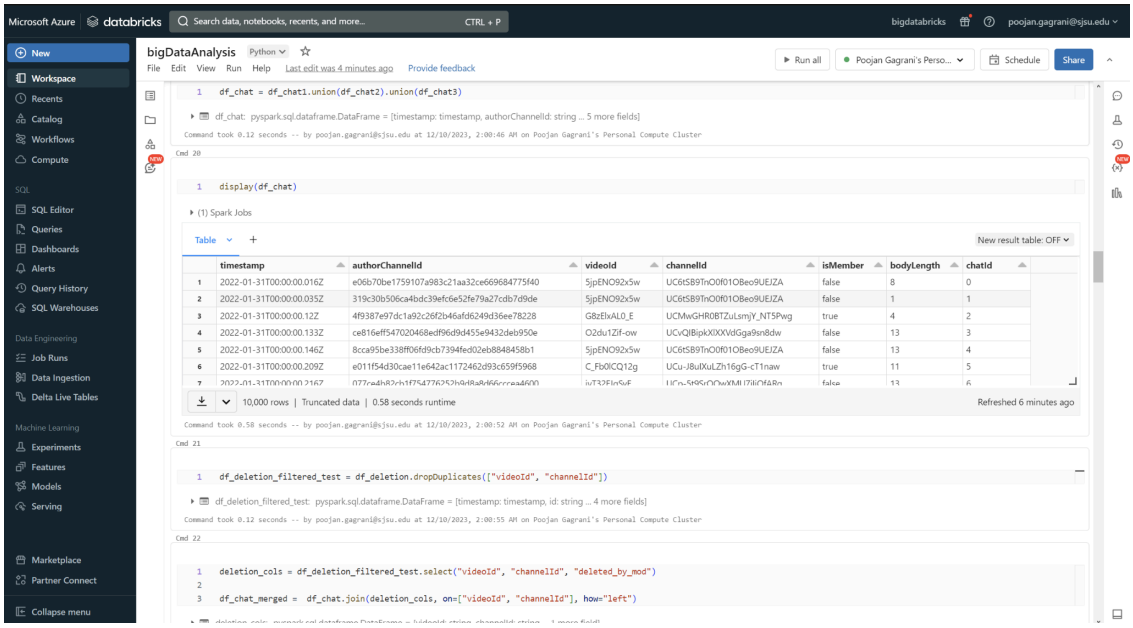


Fig. 22. Merging Chat data

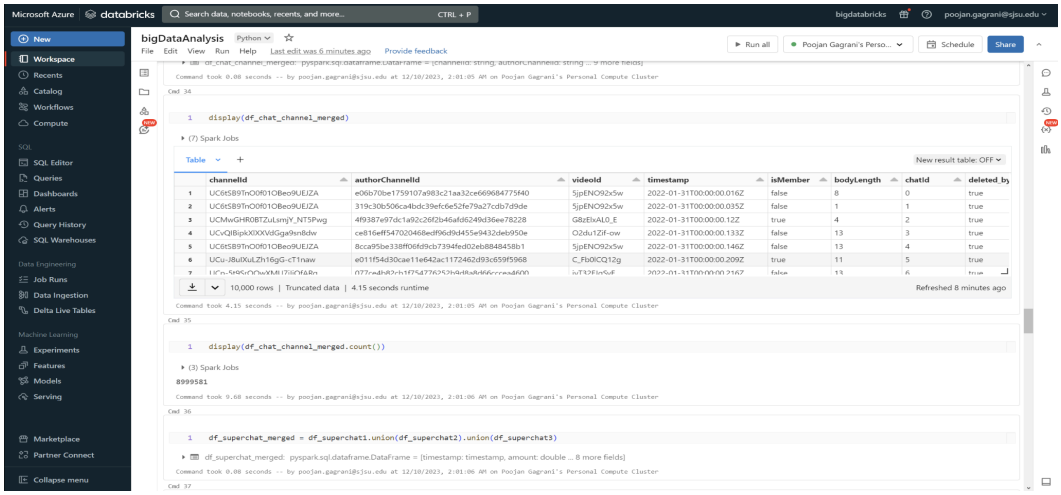


Fig. 23. Merging Chat and Channel data

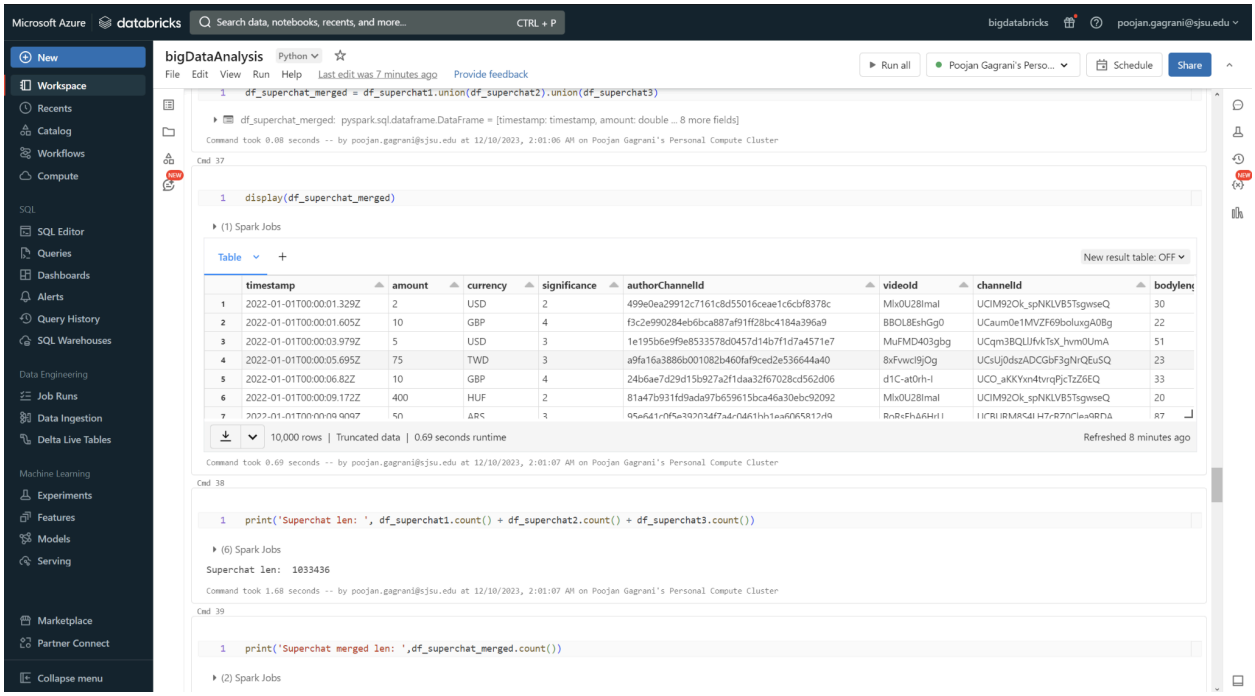


Fig. 24. Merging Superchat data

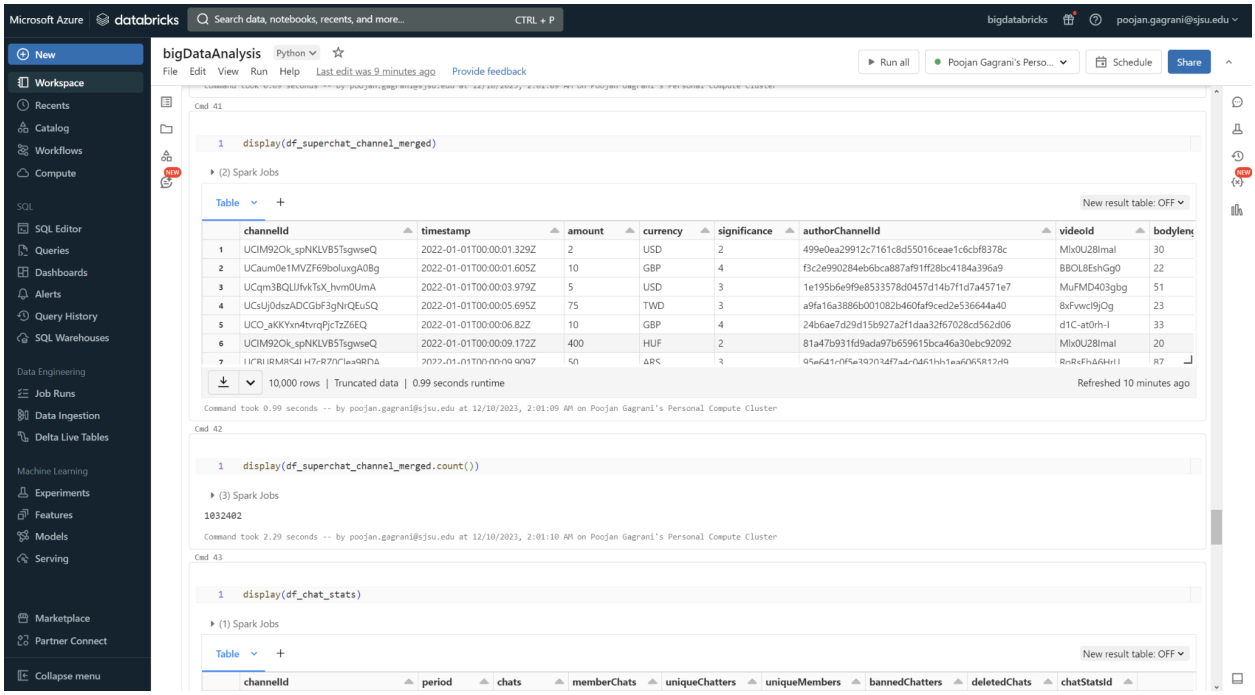


Fig. 25. Merging Superchat and Channel data

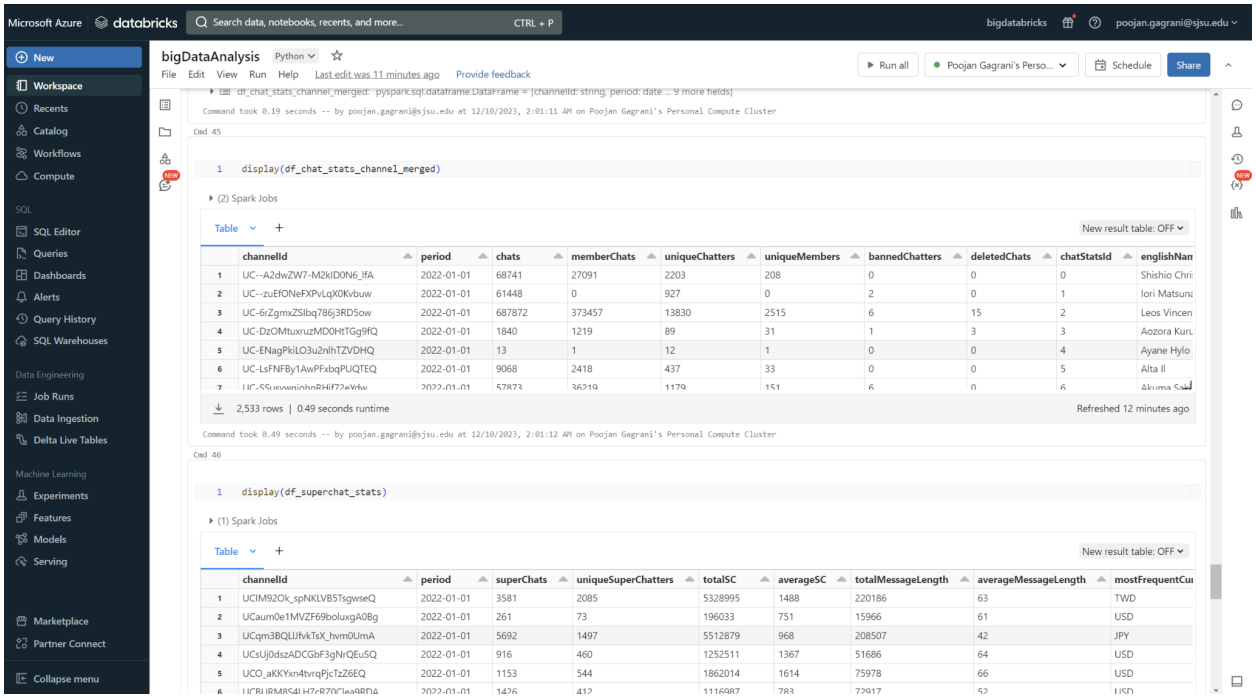


Fig. 26. Merging Chat Stats and Channel data

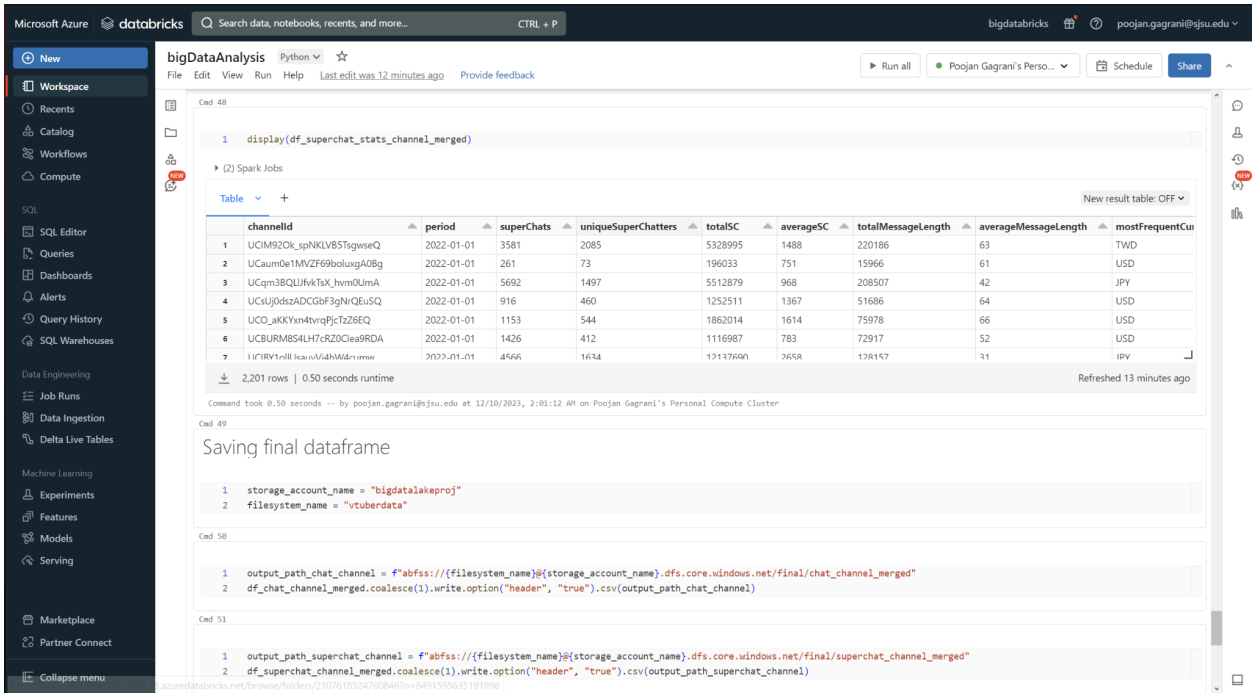


Fig. 27. Merging Superchat Stats and Channel data

c. Loading data

The merged data is moved from the source Azure data lake to the destination Azure Synapse Analytics where the EDA is performed using built-in SQL serverless to understand better the data and Visualization is done using the built-in Power-BI to extract data in the BI tool where the dashboarding is done.

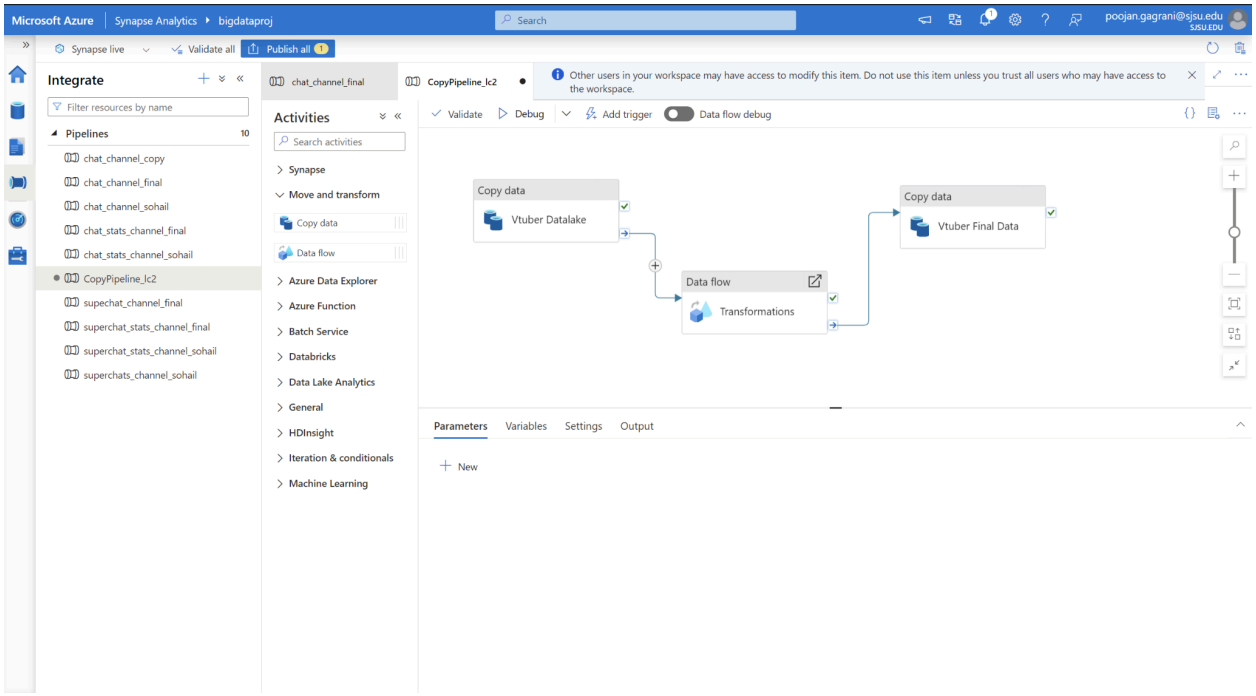


Fig. 28. Pipeline for moving final data from source to destination

Chapter 6 Data Analytics

6.1 Data Analytics using SQL Script

Once the data was cleansed, transformed, and merged in Azure Databricks it was loaded back into the Azure Synapse Analytics to perform analysis using SQL script capability. The analysis is performed on these datasets to get meaningful insights about the affiliations and channels using chat and super chat statistics as well as performing trend analysis like observing the number of banned chatters over the period to understand the behavior of the audience. Below

are some of the queries that were used to perform analysis along with the insights gained from it.

Query 1: Total number of Super Chats and Unique Super Chatters in each Period

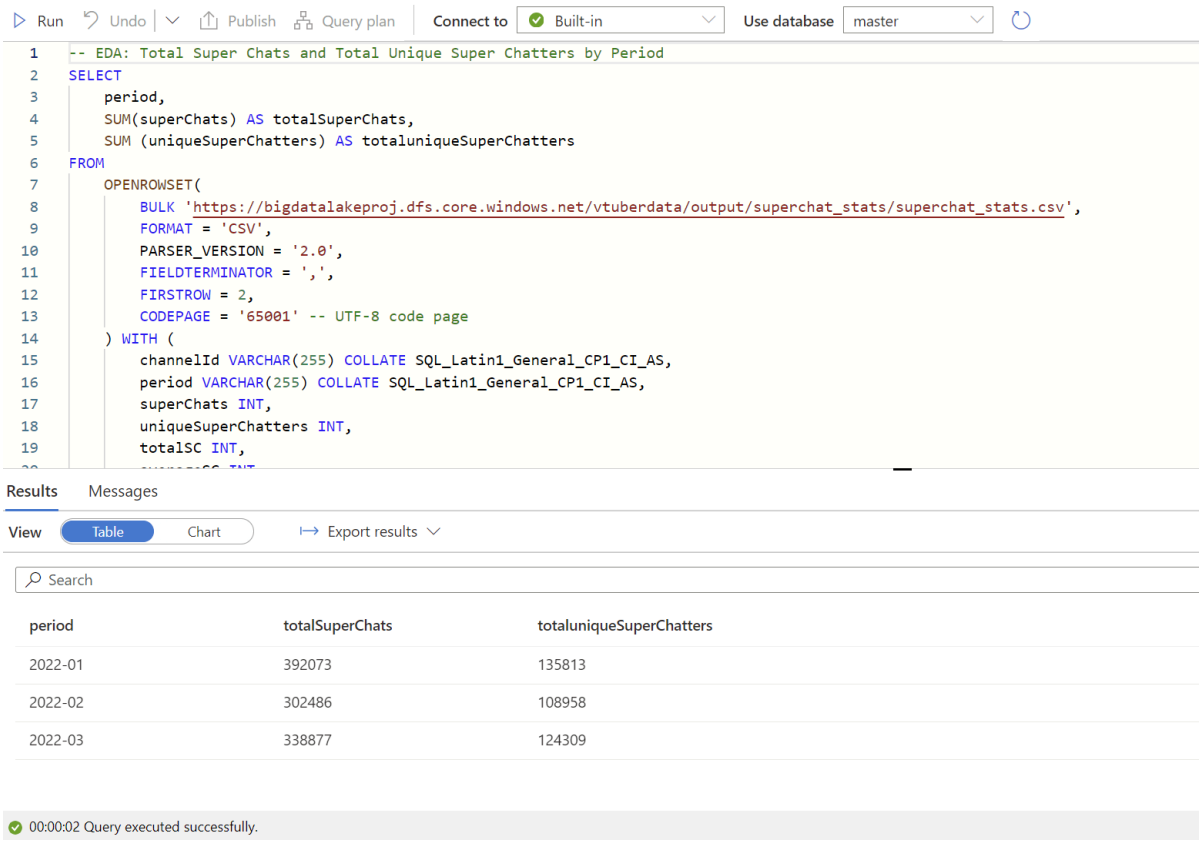


Fig. 29. SQL query with output to show the total number of Super Chats and Unique Super Chatters in each Period

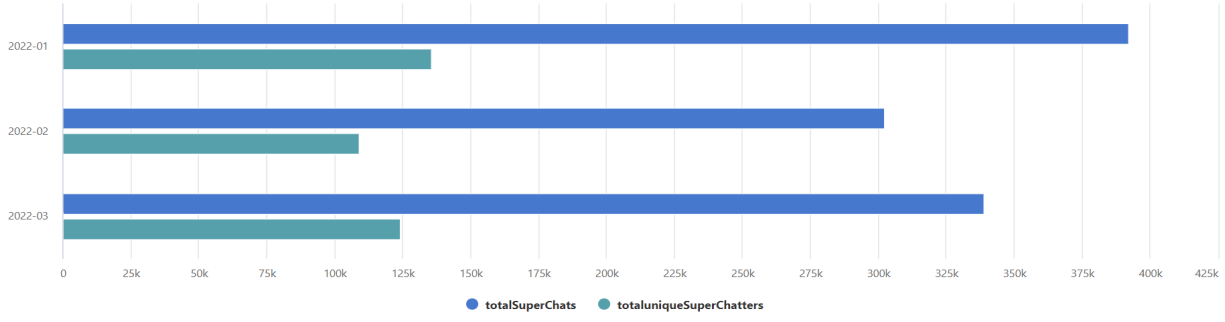


Fig. 30. Horizontal Bar plot of the total number of Super Chats and Unique Super Chatters over the Period

This query result and horizontal bar plot shows the total number of unique super chatters that are involved and the total number of super chats that are done in all the channels over 3 months. It is quite visible that the number of super chats and the number of unique super chats were highest in January then it decreased in February and it increased again in March. A probable reason to see this trend could be that the engagement between the streamer and the viewer got reduced during this time or the content delivered by most of the channels was not satisfactory for the viewers. This is important for streamers to determine the strategy that could be used to optimize their approach towards viewers, it could also help to provide information about the periods where maximum chances of super chats are there maybe due to some holidays or festivals this could be helpful for streamers to create more attractive content and gain more viewers.

Query 2: Total number of Banned Chatters by Period

```
-- Exploratory Data Analysis (EDA) for Total Banned Chatters by Period

SELECT
    period AS [Period],
    SUM(bannedChatters) AS [Total Banned Chatters]
FROM
    OPENROWSET(
        BULK 'https://bigdatalakeproj.dfs.core.windows.net/vtuberdata/output/chat_stats/chat_stats.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        FIELDTERMINATOR = ',',
        FIRSTROW = 2,
        CODEPAGE = '65001' -- UTF-8 code page
    ) WITH (
        channelId VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        period VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        chats INT,
        memberChats INT,
```

Results Messages

View Table Chart Export results

Search

Period	Total Banned Chatters
2022-01	30124
2022-02	30305
2022-03	38694

Fig. 31. SQL query with output to show the total number of Banned Chatters by Period

Fig. 32. Line plot to show the trend of total Banned Chatters over the Period

This query result and line plot shows the total number of chatters that have been banned over 3 months. It is easy to see that over months the number of chatters that are getting banned is increasing. A steep increase can be seen in March. A probable reason could be that there were a large number of chatters who were violating the channel guidelines or were using abusive language in the charts resulting in an instant ban of the id. It could be a useful indicator for the streamers as well as the streaming platform to incorporate stricter guidelines so that a safe and peaceful environment is maintained. It could also be helpful for new streamers to determine which kind of content they should be away from so that it would result in less number of guideline violations by the chatters.

Query 3: Top 10 Currencies by total Super Chat Amount

```sql
-- EDA: Total Super Chat Amount by Currency
SELECT TOP 10
    mostFrequentCurrency AS currency,
    SUM(totalSC) AS totalSuperChatAmount
FROM
    OPENROWSET(
        BULK 'https://bigdatalakeproj.dfs.core.windows.net/vtuberdata/output/superchat_stats/superchat_stats.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        FIELDTERMINATOR = ',',
        FIRSTROW = 2,
        CODEPAGE = '65001' -- UTF-8 code page
    ) WITH (
        channelId VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        period VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        superChats INT,
        uniqueSuperChatters INT,
        totalSC INT,
        averageSC INT,
        totalMessageLength INT,
        averageMessageLength INT,
        mostFrequentCurrency VARCHAR(10) COLLATE SQL_Latin1_General_CP1_CI_AS,
        mostFrequentColor VARCHAR(10) COLLATE SQL_Latin1_General_CP1_CI_AS,
        superChatStatsId INT
    ) AS [result]
GROUP BY
    mostFrequentCurrency
ORDER BY
    totalSuperChatAmount DESC;
```

Results Messages

View Table Chart Export results

Search

currency	totalSuperChatAmount
JPY	954849472
USD	180244882
TWD	96875715
HKD	7288098
KRW	4708020
MXN	3669933
ARS	1535878
PHP	1034491
CAD	948038
MYR	387312

Fig. 33. SQL query with output to show top 10 Currencies by total Super Chat Amount

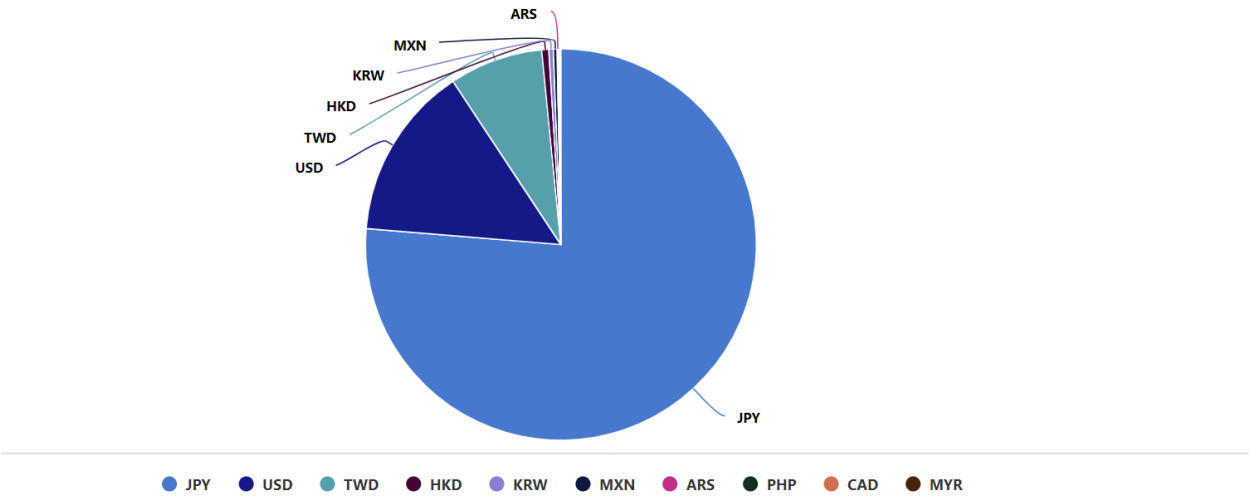


Fig. 34. Pie chart to show total Super Chat Amount from top 10 Currency

The above query result and the pie chart shows the top 10 currencies by total super chat amount. It is quite evident that the maximum amount of super chats was in Japanese Yen followed by USD. One probable reason could be that the maximum number of viewers and the streamers are from Japan and America which would inevitably lead to a large amount of donation from the currency local to that country. Another reason could be that some of the top VTubers are from Japan and America and they accept super chats only in specific currency would could also lead to a contribution towards the total amount donated through that currency. This information could be really helpful for the new viewers as it can help them understand the payment preferences of the community. It could also be helpful for new streamers to create content related to that country so that more viewers from that country can be attracted which could lead to the growth of that streamer.

Query 4: Top 10 Channels with the highest Banned Chatters percentage by Period

```sql
-- EDA: Top Channels with Highest Banned Chatters Percentage with Period
SELECT TOP 10
    channelId,
    englishName,
    affiliation,
    LEFT(period, 7) AS Period,
    (CAST(SUM(bannedChatters) AS FLOAT) / NULLIF(SUM(chats), 0)) * 100 AS bannedChattersPercentage
FROM
    OPENROWSET(
        BULK 'https://bigdatalakeproj.dfs.core.windows.net/vtuberdata/final/chat_stats_channel/chat_stats_channel.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        FIELDTERMINATOR = ',',
        FIRSTROW = 2,
        CODEPAGE = '65001' -- UTF-8 code page
    ) WITH (
        channelId VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        period VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        chats INT,
        memberChats INT,
        uniqueChatters INT,
        uniqueMembers INT,
        bannedChatters INT,
        deletedChats INT,
        chatStatsId INT,
        englishName VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
        affiliation VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS
    ) AS [result]
GROUP BY
    channelId, englishName, affiliation, LEFT(period, 7)
```

Results Messages

View Table Chart Export results

Search

channelId	englishName	affiliation	Period	bannedChattersPercentage
UCdCczCYPOePShTfbx6tzz_A	Hoshino Char	Independents	2022-03	6.25
UCu5nZcaVhqjskh--rhQv0rA	Minerva Rosaline	Independents	2022-01	2.92533098844482
UCJFZiqLMntJufDCHc6bQixg	Hololive VTuber Group	Hololive	2022-02	2.29628401766822
UCu5nZcaVhqjskh--rhQv0rA	Minerva Rosaline	Independents	2022-02	2.05879631057778
UCCK072ns58t1nYqX1CoaZ-g	Mythia Batford	Independents	2022-03	1.93268381541179
UCu5nZcaVhqjskh--rhQv0rA	Minerva Rosaline	Independents	2022-03	1.91258318538769
UCCK072ns58t1nYqX1CoaZ-g	Mythia Batford	Independents	2022-01	1.90815727105151
UCCK072ns58t1nYqX1CoaZ-g	Mythia Batford	Independents	2022-02	1.66282664776196
UCge_6FJHyeOCxRtWCmaVTAQ	Lumi Celestia	MAHAS	2022-03	1.44587991164067
UC7fk0CB07ly8oSl0aqKkqFg	Nakiri Ayame	Hololive	2022-03	1.36549840691853

Fig. 35. SQL query and output to show the top 10 Channels with the highest Banned Chatters percentage by Period

Figure 35 shows the query result of the top 10 channels having the highest percentage of banned chatters grouped by period. It is quite evident that the streamer named “Hoshino Char” without any affiliation has the highest percentage of banned chatters in March. The probable reason could be the shift in the content of the stream which some percentage of viewers didn’t like or might have found offensive which could have resulted in the chats by the viewers that needed to be moderated. This information could be really helpful for the streamer to determine the factors leading to the banned chats and modify the content of the stream accordingly so that it could result in healthy engagement by the community and a positive environment maintained on the streaming platform. It could be helpful for new viewers to understand which channels have strict moderation policies and can determine based on that whether they want to watch that particular streamer or not.

Query 5: Affiliation details of top 10 Channels by Average Subscribers

```sql
-- Query to get details for the top 10 affiliations
WITH TopAffiliations AS (
    SELECT TOP 10
        affiliation,
        COUNT(*) AS total_channels,
        AVG(CONVERT(INT, subscriptionCount)) AS avg_subscribers,
        MAX(CONVERT(INT, subscriptionCount)) AS max_subscribers,
        MIN(CONVERT(INT, subscriptionCount)) AS min_subscribers,
        SUM(CONVERT(INT, subscriptionCount)) AS total_subscribers
    FROM
        OPENROWSET(
            BULK 'https://bigdatalakeproj.dfs.core.windows.net/vtuberdata/output/channel/channels.csv',
            FORMAT = 'CSV',
            PARSER_VERSION = '2.0',
            FIELDTERMINATOR = ',',
            FIRSTROW = 2,
            CODEPAGE = '65001' -- UTF-8 code page
        ) WITH (
            channelId VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
            englishName VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
            affiliation VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
            subscriptionCount VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
            videoCount VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS
        ) AS [result]
    GROUP BY
        affiliation
    ORDER BY
        total_subscribers DESC
)
SELECT
```

Results Messages

View Table Chart Export results

Search

affiliation	total_channels	avg_subscribers	max_subscribers	min_subscribers	total_subscribers
Kizuna Ai Inc.	3	1551633	3080000	84900	4654900
Hololive	92	785527	4100000	9320	72268520
VShojo	15	402273	1430000	52600	6034100
Nijisanji	207	293853	1540000	383	60827723
VSpo	16	256093	453000	84500	4097500
Twitch Independents	60	207092	2060000	7950	12425550
Nori Pro	13	147830	883000	38200	1921800
KAMITSUBAKI	18	116690	706000	5850	2100420
Independents	470	105500	2510000	0	49585391
774inc	31	94726	602000	9920	2936520

Fig. 36. SQL query and output containing Affiliation details about top 10 Affiliations by Average Subscribers

The above query result shows the details of the top 10 affiliations ordered based on average subscribers which include details like the total number of channels that are part of the affiliation, average number of subscribers, maximum number of subscribers that is equal to the subscriber count of a particular channel of that affiliation, minimum number of subscribers that is equal to the subscriber count of a particular channel of that affiliation, and total number of subscribers. It can be easily seen from the output that the affiliation named “Kizuna Ai inc.” has the highest number of average subscribers but only three channels are part of it which shows that each channel is performing well and have decent viewership and followers. Even though the number of channels involved in Hololive and Nijisanji is high the average subscriber count is still good which shows that affiliation is famous among the audience and people prefer to watch channels related to them. This is useful for the new incoming streamer to determine which affiliation to associate with to reach a larger audience and also it could be helpful for the people who are new to this streaming platform to determine whose content to watch.

Chapter 7 Data Analytics

7.1 Data Visualization

The data visualization process was performed by creating visualizations using merged datasets to determine the underlying pattern and gain valuable insights about affiliations and channels using channel statistics, super chat statistics, and chat statistics. Power BI software was used to create various visualizations and they were incorporated into three dashboards named chat stats, super chat stats, and banned and deleted.

Figure 37 illustrates the top 10 affiliations based on the total number of channels associated with them. It is quite evident from the pie chart that Nijisanji has the highest number

of channels. Independents refer to the channels that are not associated with any affiliations. Hololive follows the Nijisanji. This information is useful for new streamers to determine the affiliation to choose from based on their popularity among streamers. It could also be helpful for a new viewer to understand which channel to explore based on the popularity of the affiliations.

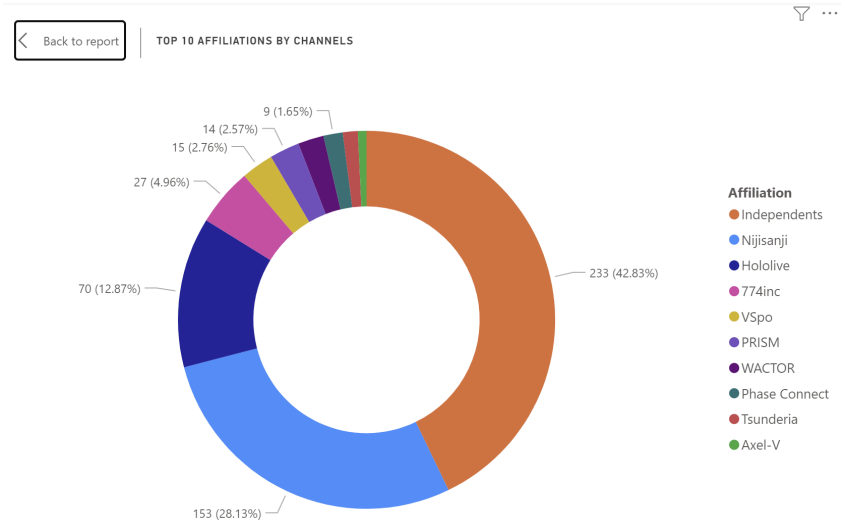


Fig. 37. Top 10 Affiliations by Channels

Figure 38 shows the total number of videos uploaded by the top 10 affiliations. It can be seen that Nijisanji and Hololive have a very large number of video uploads. The probable reason could be due to a large number of channels present in them which on a cumulative level results in a large number of uploads. These many videos could attract new viewers as they allow the ground to explore varied content uploaded by the channels associated with these affiliations which could ultimately lead to an increase in popularity and the subscriber count. It could be useful for existing streamers to determine whether to stay with their current affiliation or move on to the popular ones where content is uploaded more frequently and has a larger viewer base.

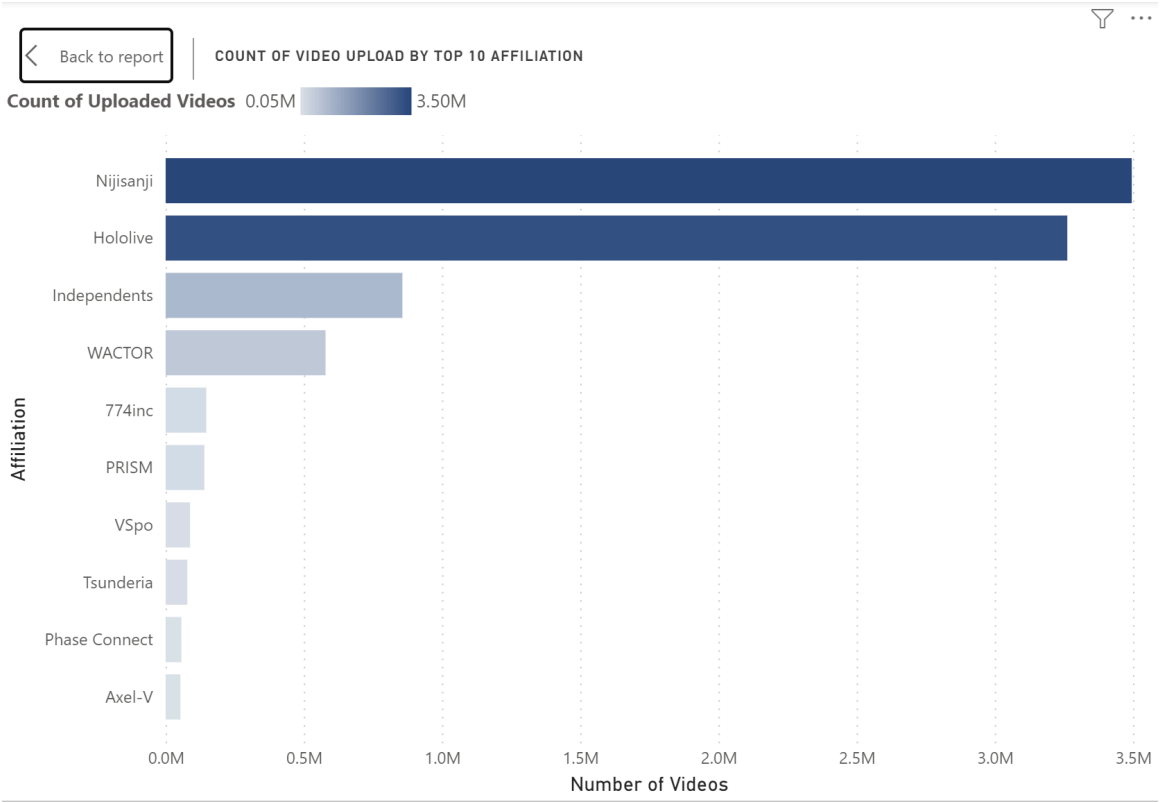


Fig. 38. Top 10 Affiliations by count of Video uploads

Figure 39 shows the top 15 channels based on the total number of member chats. The channel named Gawr Gura has the highest number of member chats followed by Yukihana Lamy and Ceres Fauna. This is useful for viewers to determine which channel has a more engaged community associated with it. This could allow the viewer to get a better understanding of the content by interacting with the members. It could also be a nice starting point for new streamers where they can understand channels like Gwar Gura by performing in-depth research on their content, strategies followed for streaming, and how the streamer keeps the community together

to get a brief idea about how to gain more popularity and get higher viewership.

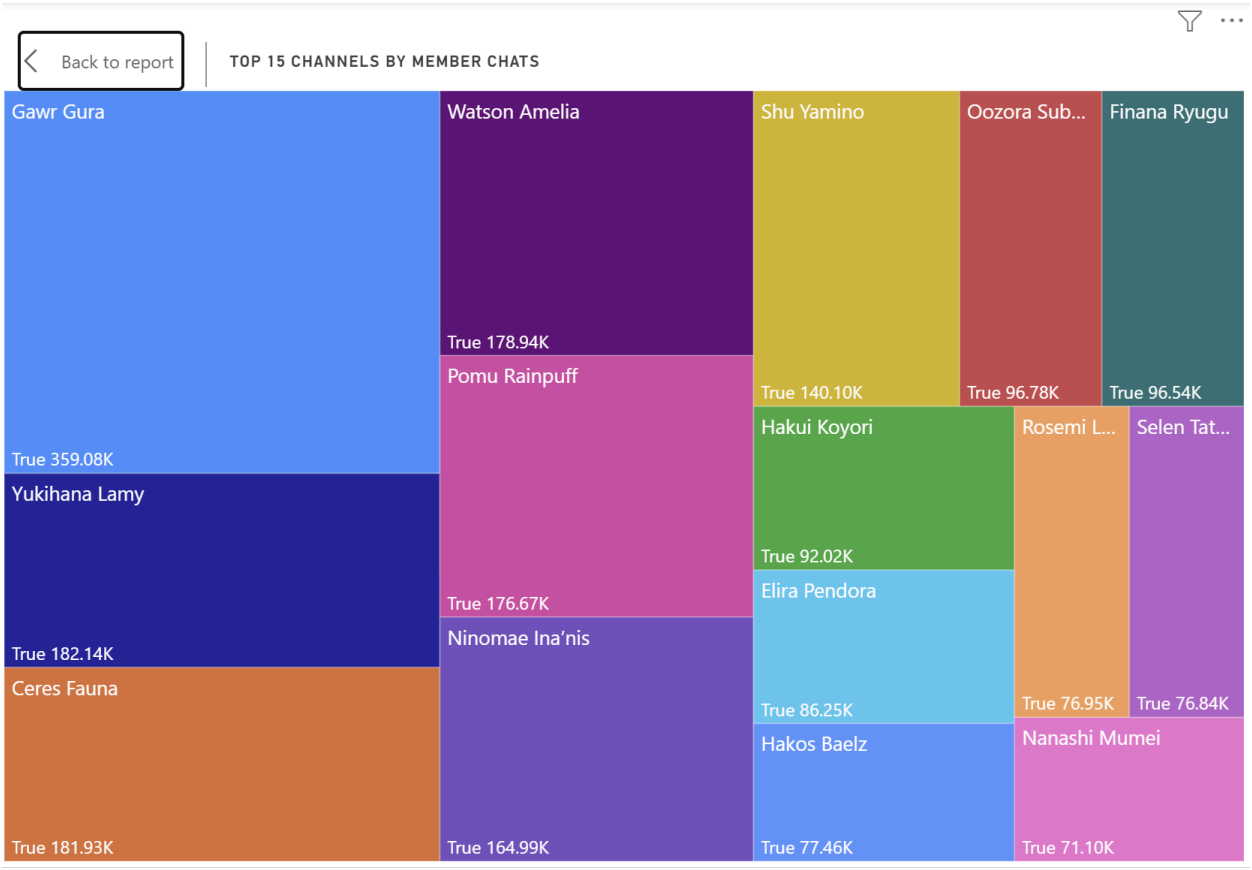


Fig. 39. Top 15 Channels by Member Chats

Figure 40 depicts the trend of the number of members in the top 5 affiliations over 3 months. It is quite evident from the plot that Hololive has the highest number of members in all three periods followed by Nijisanji and it is quite understandable because the number of channels involved with it is also very large. In Hololive the number of new members that are getting added is decreasing every month and in the case of Nijisanji it increased for February but then it decreased again in March. It could help new viewers gauge the popularity of the affiliations. Hololive is consistently able to maintain a strong membership and fanbase that shows the trust of the audience. This information is useful for new streamers to determine the affiliation they could join based on community support.

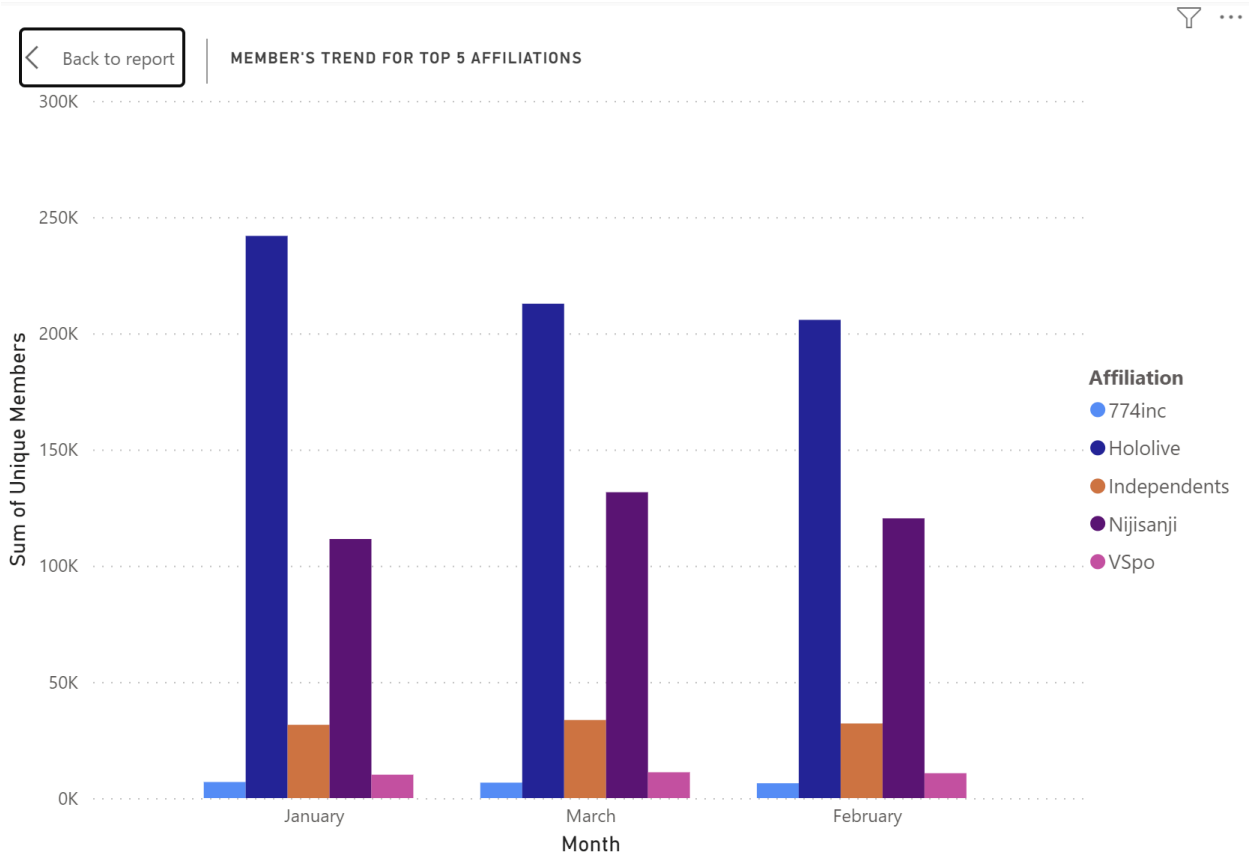


Fig. 40. Member's trend for top 5 Affiliations

Figure 41 shows the top 10 channels based on the total number of super chats. In the plot, the channel named Kson has the highest number of super chats followed by the channel named Hakui Koyori. The probable reasons could be the interesting content created by these streamers that attract more viewers to donate or the streamer maintains a healthy engagement with the viewers. This information could be helpful for new incoming streamers to determine the type of content or the strategies of community engagement that attract more audience and inspire people to contribute financially to the channel and can incorporate those insights to improve their streams. For new viewers, it provides information about the popularity of the channel and also the creation of a supportive and interactive environment where viewers contribute willingly.

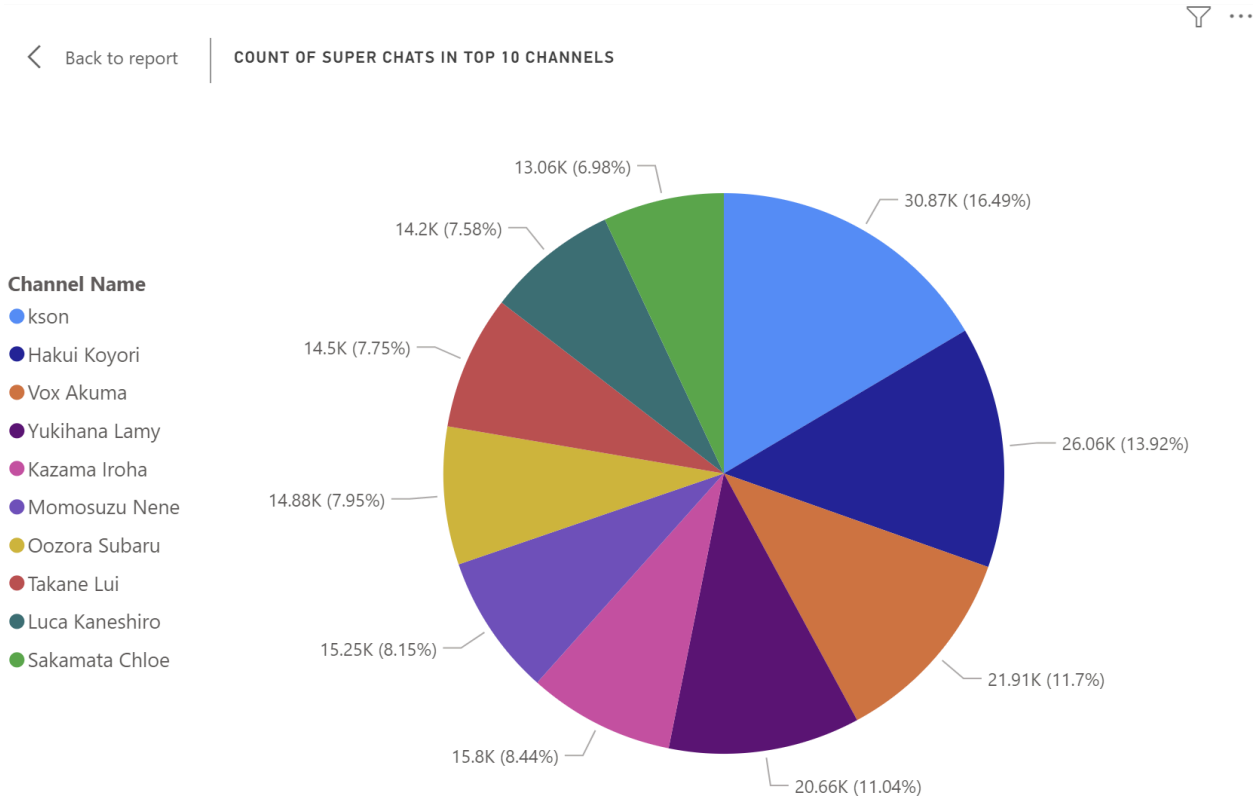


Fig. 41. Count of Super Chats in top 10 Channels

Figure 42 shows the top 10 channels that gained the maximum amount. The channel named Hakos Balez gained the highest amount through the streams followed by the channel named Sakamata Chloe. A probable reason could be the content created by these streamers was good and engaging which resulted in attracting a large amount of viewers and encouraged them to contribute to the streams financially. Another reason could be that these channels might have a decent fanbase but they are constantly supporting the channel financially which results in a large amount of money received by the streamer as a whole. It could be helpful for streamers to create more engaging content and follow the monetization techniques of these channels to get a large amount of money and become more popular. For viewers, it could portray the ability of the

streamer to maintain a loyal fanbase by highly appreciating the viewer who contributes financially and also maintaining a healthy relationship with the viewer community.

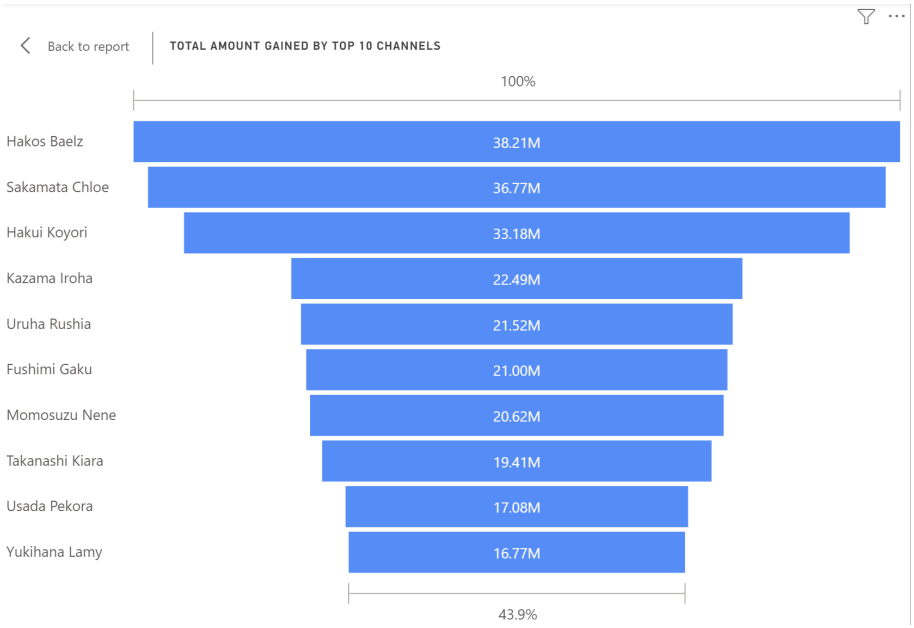


Fig. 42. Total Amount gained by the top 10 Channels

Figure 43 depicts the top 10 affiliations by the count of banned chats. It is quite visible that the affiliation named WACTOR has the highest number of banned chats followed by Nijisanji. Here, independence is not an affiliation rather it is a collection of streamers with no affiliation associated with them. This indicates that an affiliation named WACTOR might have stricter community guidelines in place which could have resulted in frequent bans of the chats and might show the enforcement of a healthy environment. Another reason could be the content created by the streamer might have been offensive or revolting to some viewers which resulted in the chats that were not according to the guidelines. This shows the information about the community guideline policy of the affiliation and it also portrays the information that may be the streamer is showing negative behavior towards the audience. This could be helpful for the streaming platform and also the affiliation to take strict action against that channel.

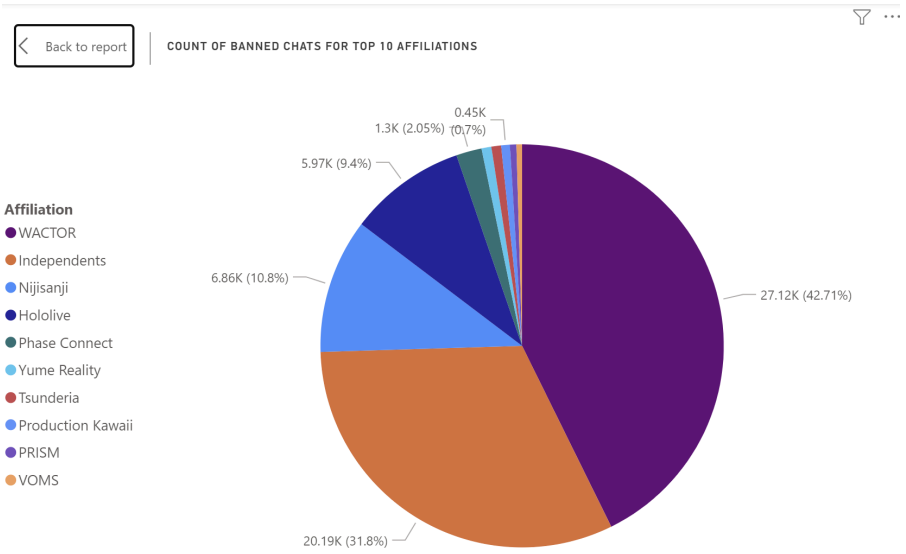


Fig. 43. Count of Banned Chats for top 10 Affiliations

Figure 44 depicts the top 10 channels by total number of deleted chats. It is quite evident that Gawr Gura has the highest number of deleted chats followed by Shu Yamino. A probable reason could be the moderators are trying to maintain a healthy environment by deleting chats that are offensive or contain abusive language. It could also be possible that the streamer is having a fight over the stream or might be saying something offensive that would have resulted in chats that were automatically deleted by the moderation policy. There is a possibility of both positive and negative take on this. If it is positive then it means that the streamer is trying to maintain a healthy environment and trying to ensure that new viewers will feel welcome and will be able to enjoy the streams without any issues. It would also look positive in the eyes of the viewer as it portrays that the streamer is trying to maintain proper community guidelines. This could prove useful for new streamers to determine the strategies needed to be followed for moderation such that there is proper communication between streamer and audience but at the same also ensured that no viewer will feel disheartened or lose interest due to negative

comments. It could also help new streamers understand what kind of things to look out for so that it does not result in any offensive content or arguments with some viewers.

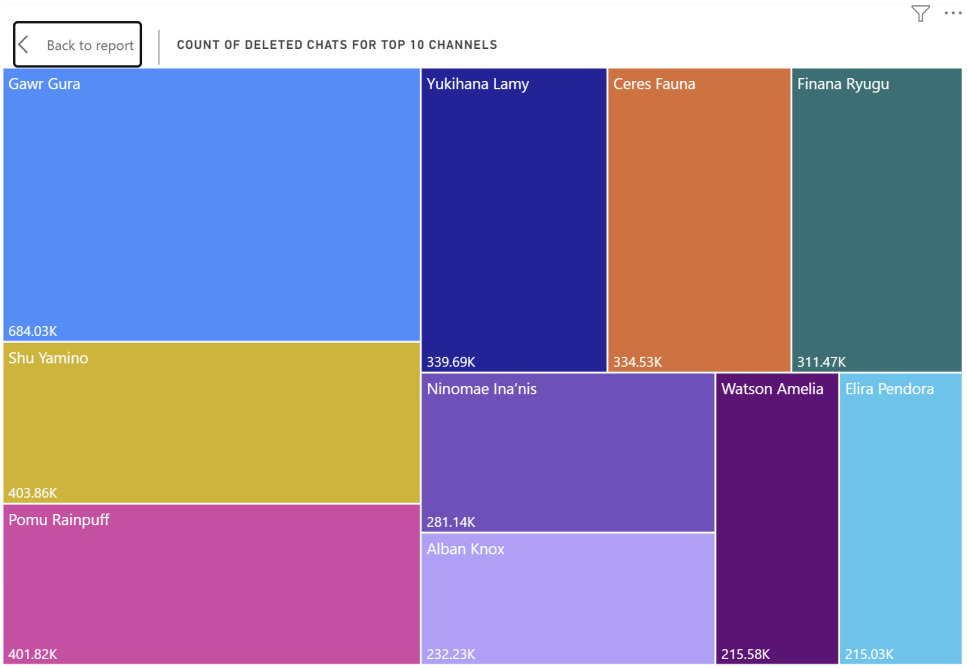


Fig. 44. Count of Deleted Chats for top 10 Channels

7.2 Web Application to Host Dashboard

The web application was created using the ASP.NET core framework to host the Power BI dashboard that contains various plots that provide useful insights about the VTuber's data. ASP.NET is the most commonly used web development framework that is used to create more dynamic websites. In this study, ASP.NET is used to create a landing page that contains some brief details about the objective of the study and the benefits that this study will have for new viewers and aspiring VTubers. Figure 45 shows the landing page of the website.

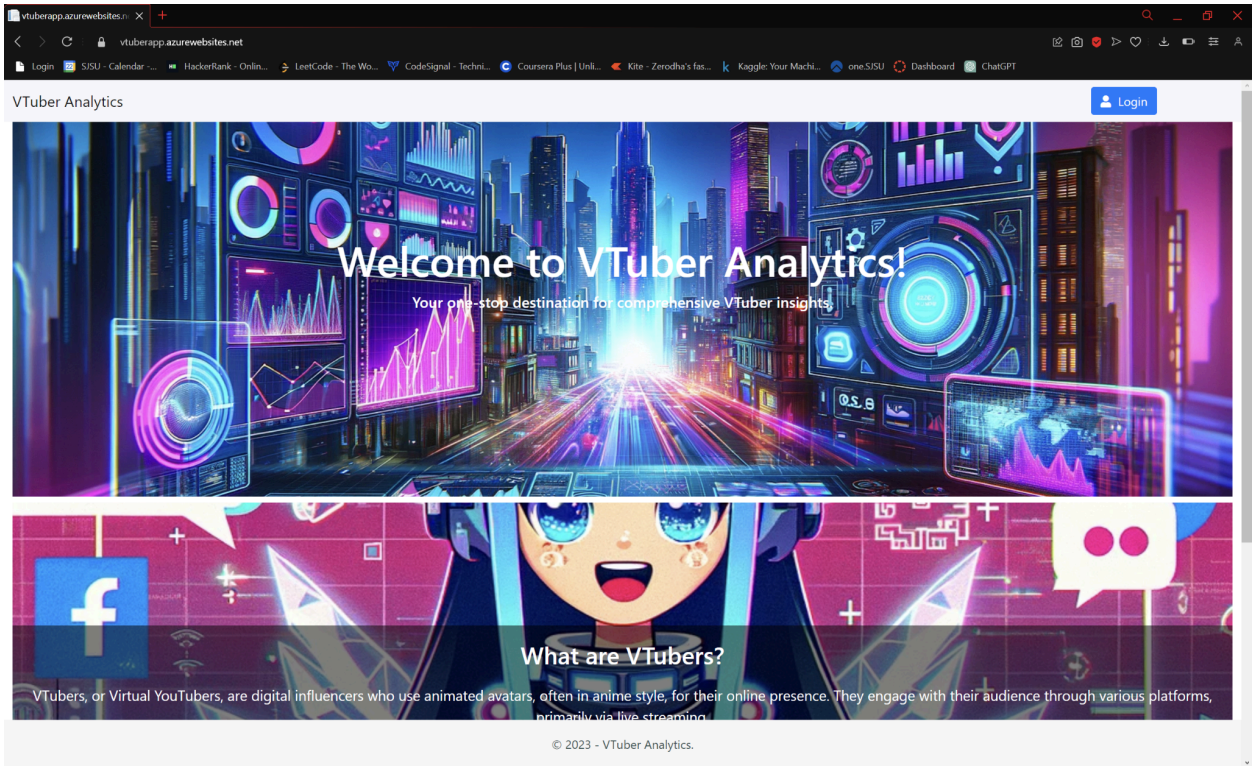


Fig. 45. Website Landing page

The Single Sign On is also implemented on this landing page using Google OAuth SSO. In this using the single login credentials of Google the user can access multiple software systems and websites. Users can log into the website by clicking on the login button at the top right-hand side of the homepage. When it is pressed, the user will be directed to the sign-in with Google option where the user will be asked to use either the existing Gmail ID that was already used to log in or sign in using a new ID. Figure 46 shows the sign-in with the Google page.

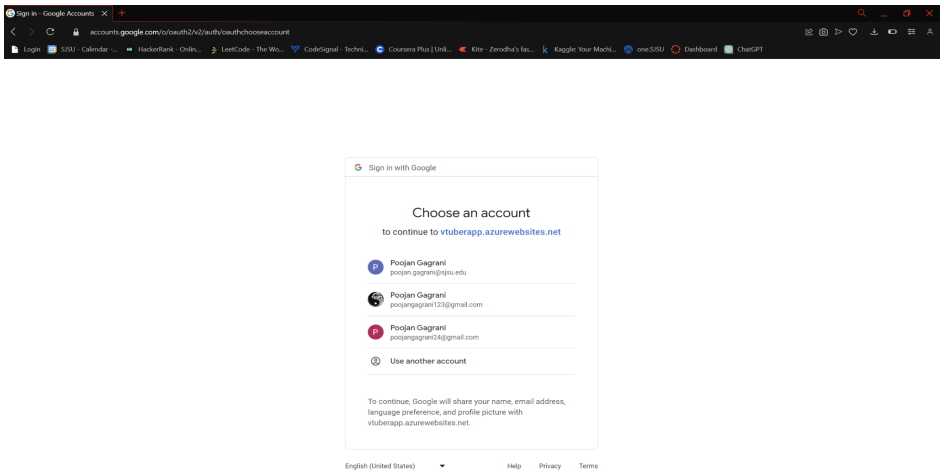


Fig. 46. Login with Google OAuth (SSO)

After successful login, the user can access multiple tabs or sheets of the Power BI dashboard that is hosted on this website. Since the dashboard is interactive users that are either streamers or viewers can play around with the plots to get useful insights. After using the website the user can log out by clicking on the logout button at the top right-hand side of the website and when it is pressed the user will be redirected to the home page or landing page. Figure 47 shows the website with the hosted dashboard.

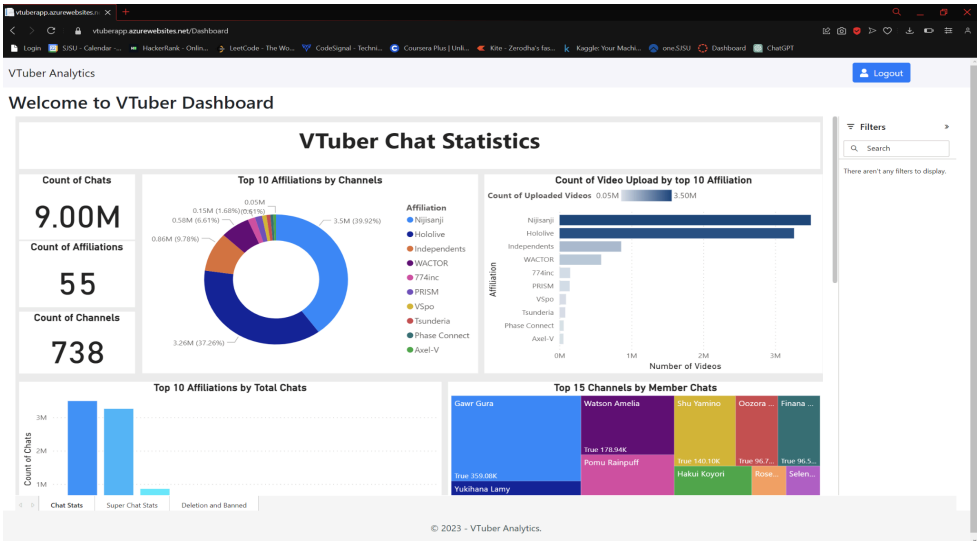


Fig.47. VTuber Dashboard hosted on the website

Chapter 8

8.1 Github

It is essential to store the final code used for the project as part of the disaster management and recovery plan. For the part of saving data the data is stored in the online repository platform GitHub where the up-to-date code's backup is stored at all times, this tool not only facilitates code backup but also facilitates a collaborative environment where multiple contributors can work together ensuring that they have up-to-date code available and can also build on top of it and can save their changes and update the code base at all times.

For this project, a GitHub repository was created and the code files were committed. It comprises data cleaning, transformation, and the web application code which was implemented for the project.

GitHub repository URL: https://github.com/poojan243/VTuber-BigData-Analytics

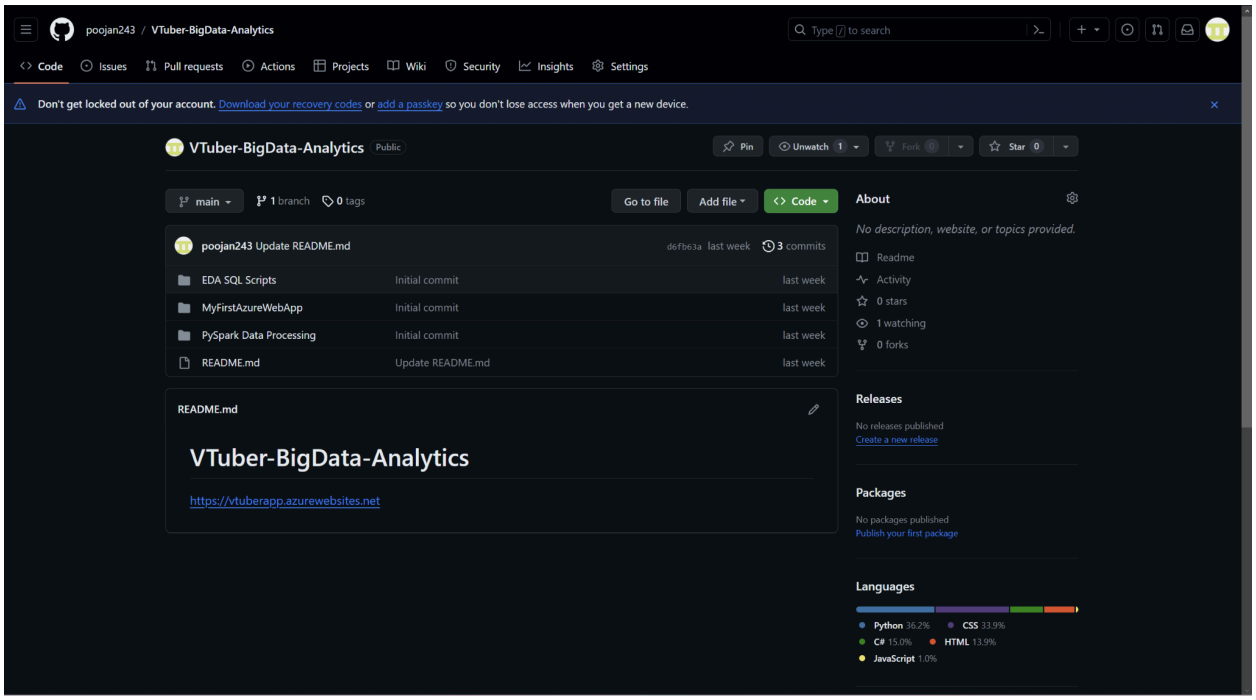


Fig. 48. Github repository of the project

8.2 Team Members and IDs

Name	Student ID
Abdul Sohail Ahmed	016769610
Poojan Gagrani	016795285
Kashish Thakur	016919383
Swetha Neha Kutty Sivakumar	016780712
Somna Sattoor	014640769

Chapter 9

9.1 Conclusion

9.2 Future Work

Creating an improved pipeline capable of handling even bigger amounts of data (TBs) at scale. Better data preparation can be done by using more capable spark clusters that can manage and preprocess larger datasets. When a greater set of data is available, including additional statistics. Enhancing the website by incorporating numerous login choices such as AzureAD, Facebook, and Twitter logins. Enabling user-specific dashboards on a website, for example, a chat metrics dashboard for a streamer that focuses on stats related to that broadcaster and allows them to easily access insights.

References

Ahmad, N., & Mamat, R. (2022). The Consumption of Popular Culture Products among Japanese Language Learners at Universiti Teknologi Mara. *International Journal of Academic Research in Business & Social Sciences*, *12*(1). https://doi.org/10.6007/ijarbss/v12-i1/12183

Amalia, S. I., & Mohammad, W. (2023). Analysis of the Effect of the Number of Views and Number of Videos on the Number of Virtual YouTuber Subscribers in Vietnam. *Himeka: Journal of Interdisciplinary Social Sciences*, *1*(1), 108–114. Retrieved from https://journal.chishikinh.my.id/index.php/himeka/article/view/23

Kim, D., & Yoo, H. (2021). A comparative study of user experience according to one-person media Virtual YouTuber (VTuber) and general YouTuber. *Asia-Pacific Journal of Convergent Research Interchange*, *7*(5), 1–10. https://doi.org/10.47116/apjcri.2021.05.01