Twitch Data Analysis Project

This project involves importing Twitch streamers data from a CSV file into a SQL Server database, performing data cleaning and manipulation using SQL queries, exporting the cleaned data back to a CSV file using SQL Server Management Studio (SSMS) Export Wizard, and presenting some of the data within charts to answer specific questions.

Introduction

The popularity of live streaming platforms like Twitch has grown significantly in recent years, with millions of users streaming and watching content daily. Analyzing Twitch data can provide valuable insights into streaming trends, popular games, and viewer behavior. This project aims to explore Twitch streamers data, clean and manipulate the data using SQL queries, and visualize key metrics to answer specific questions.

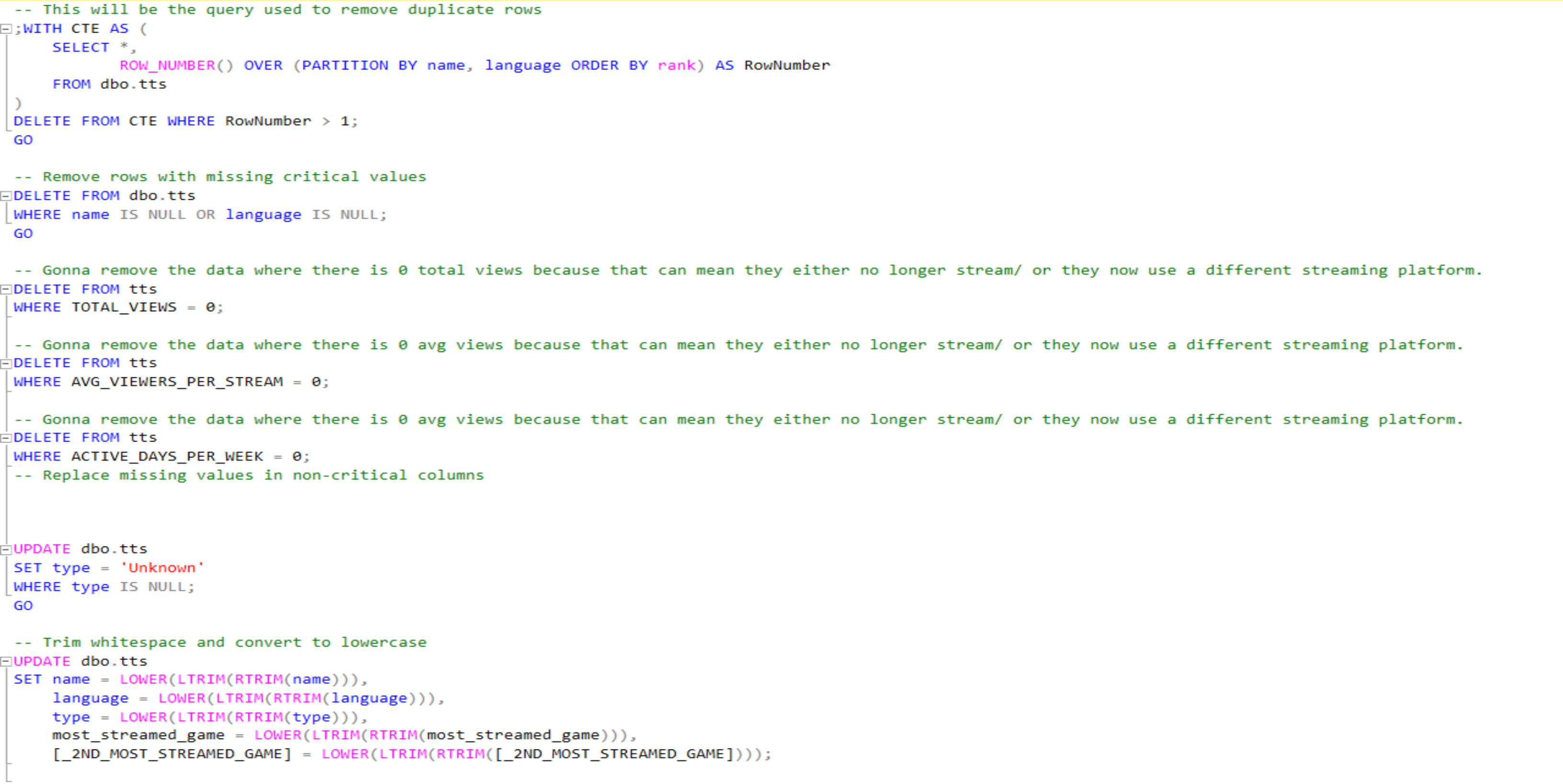
Setup Instructions

1. Database Setup

- Create a new database named `tts(Thetwitchstream)`

2. Data Cleaning

- After importing the data, it's essential to clean and preprocess it to ensure accuracy and consistency.
- Run the provided SQL queries to clean the imported data, removing duplicates, null values, and outliers.



```sql
-- This will be the query used to remove duplicate rows
;WITH CTE AS (
    SELECT *,
           ROW_NUMBER() OVER (PARTITION BY name, language ORDER BY rank) AS RowNumber
    FROM dbo.tts
)
DELETE FROM CTE WHERE RowNumber > 1;
GO

-- Remove rows with missing critical values
DELETE FROM dbo.tts
WHERE name IS NULL OR language IS NULL;
GO

-- Gonna remove the data where there is 0 total views because that can mean they either no longer stream/ or they now use a different streaming platform.
DELETE FROM tts
WHERE TOTAL_VIEWS = 0;

-- Gonna remove the data where there is 0 avg views because that can mean they either no longer stream/ or they now use a different streaming platform.
DELETE FROM tts
WHERE AVG_VIEWERS_PER_STREAM = 0;

-- Gonna remove the data where there is 0 avg views because that can mean they either no longer stream/ or they now use a different streaming platform.
DELETE FROM tts
WHERE ACTIVE_DAYS_PER_WEEK = 0;
-- Replace missing values in non-critical columns



UPDATE dbo.tts
SET type = 'Unknown'
WHERE type IS NULL;
GO

-- Trim whitespace and convert to lowercase
UPDATE dbo.tts
SET name = LOWER(LTRIM(RTRIM(name))),
    language = LOWER(LTRIM(RTRIM(language))),
    type = LOWER(LTRIM(RTRIM(type))),
    most_streamed_game = LOWER(LTRIM(RTRIM(most_streamed_game))),
    [_2ND_MOST_STREAMED_GAME] = LOWER(LTRIM(RTRIM([_2ND_MOST_STREAMED_GAME])));
```

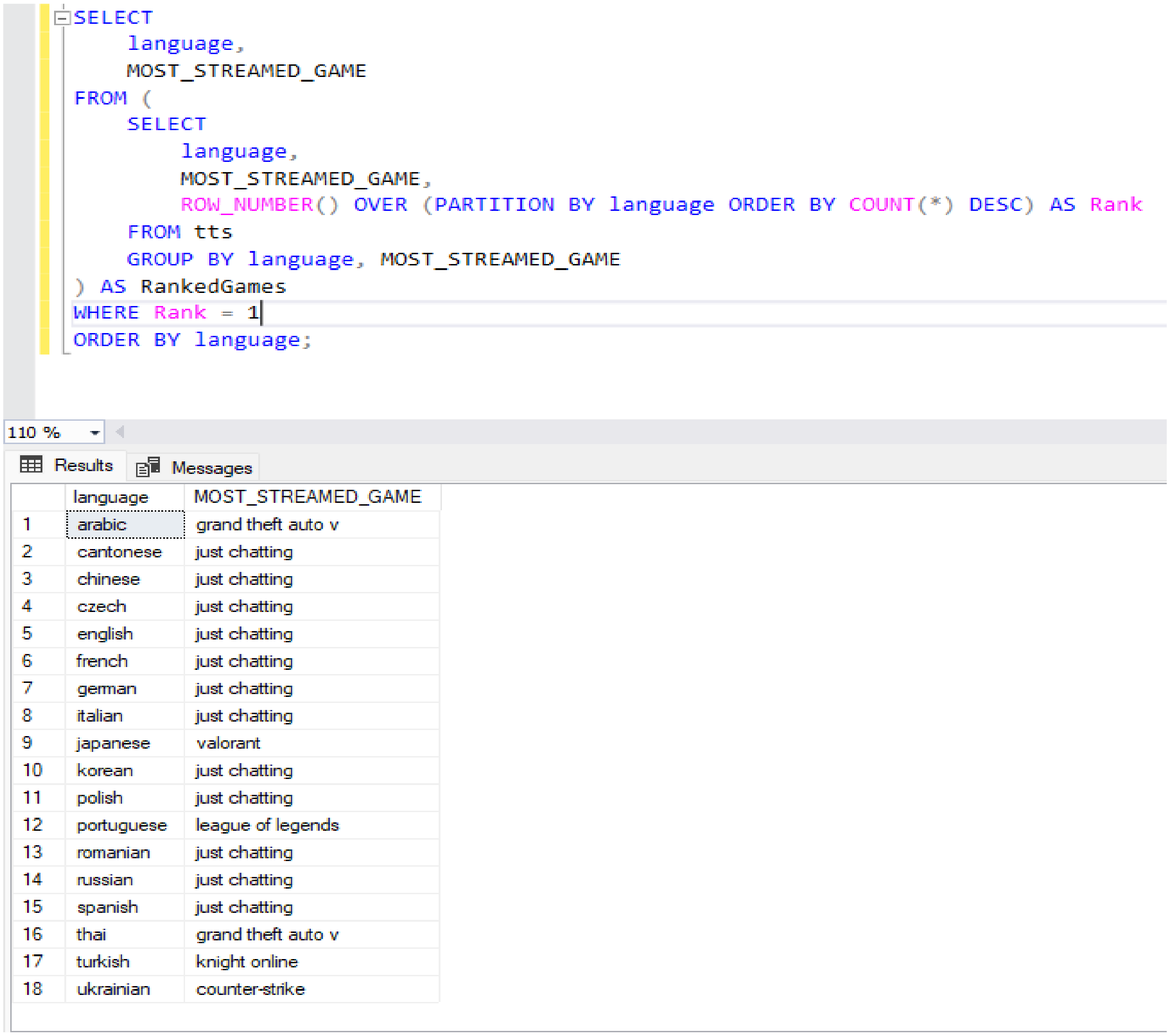
3. Export Data

- Once the data cleaning process is complete, use SQL Server Management Studio (SSMS) Export Wizard to export the cleaned data back to a CSV file.
- Follow the provided instructions in the README file to configure the Export Wizard and export the data to a CSV file.

Data Analysis and Visualization

Question 1: What game is the most streamed per language?

- Use SQL queries to aggregate the data and extract the most streamed game per language.

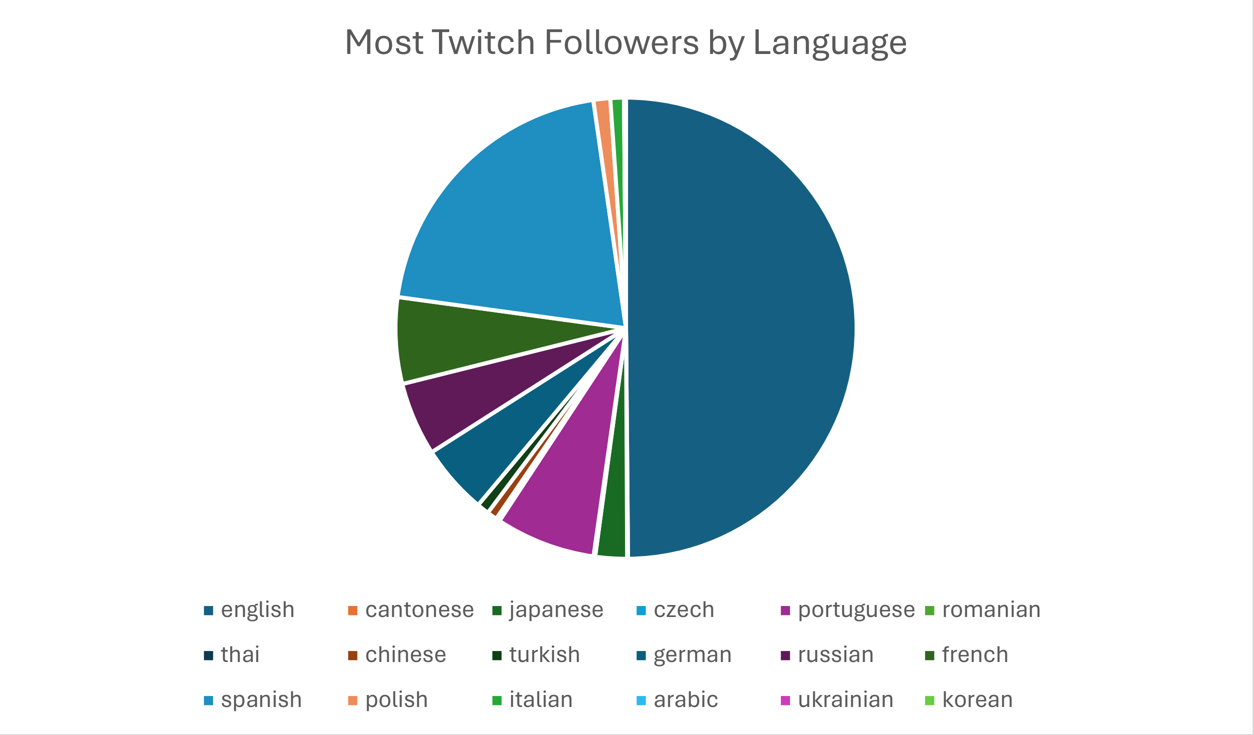


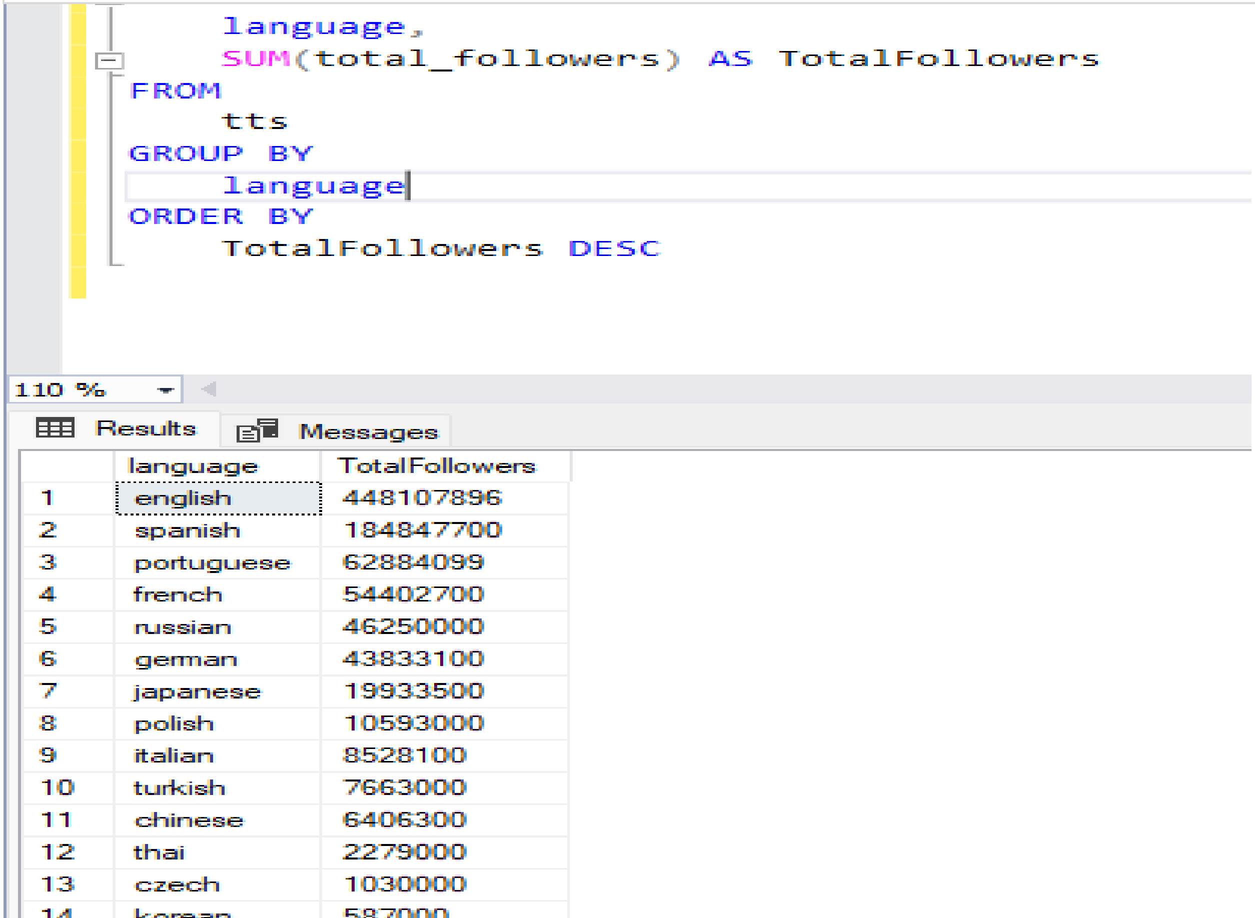
```sql
SELECT
    language,
    MOST_STREAMED_GAME
FROM (
    SELECT
        language,
        MOST_STREAMED_GAME,
        ROW_NUMBER() OVER (PARTITION BY language ORDER BY COUNT(*) DESC) AS Rank
    FROM tts
    GROUP BY language, MOST_STREAMED_GAME
) AS RankedGames
WHERE Rank = 1
ORDER BY language;
```

	language	MOST_STREAMED_GAME
1	arabic	grand theft auto v
2	cantonese	just chatting
3	chinese	just chatting
4	czech	just chatting
5	english	just chatting
6	french	just chatting
7	german	just chatting
8	italian	just chatting
9	japanese	valorant
10	korean	just chatting
11	polish	just chatting
12	portuguese	league of legends
13	romanian	just chatting
14	russian	just chatting
15	spanish	just chatting
16	thai	grand theft auto v
17	turkish	knight online
18	ukrainian	counter-strike

Question 2: What language has the most total Twitch followers?

- To determine the distribution of Twitch followers by language, we will create a pie chart.
- Aggregate the data using SQL queries to calculate the total number of Twitch followers for each language.
- Visualize the results with a pie chart to illustrate the proportion of followers for each language.



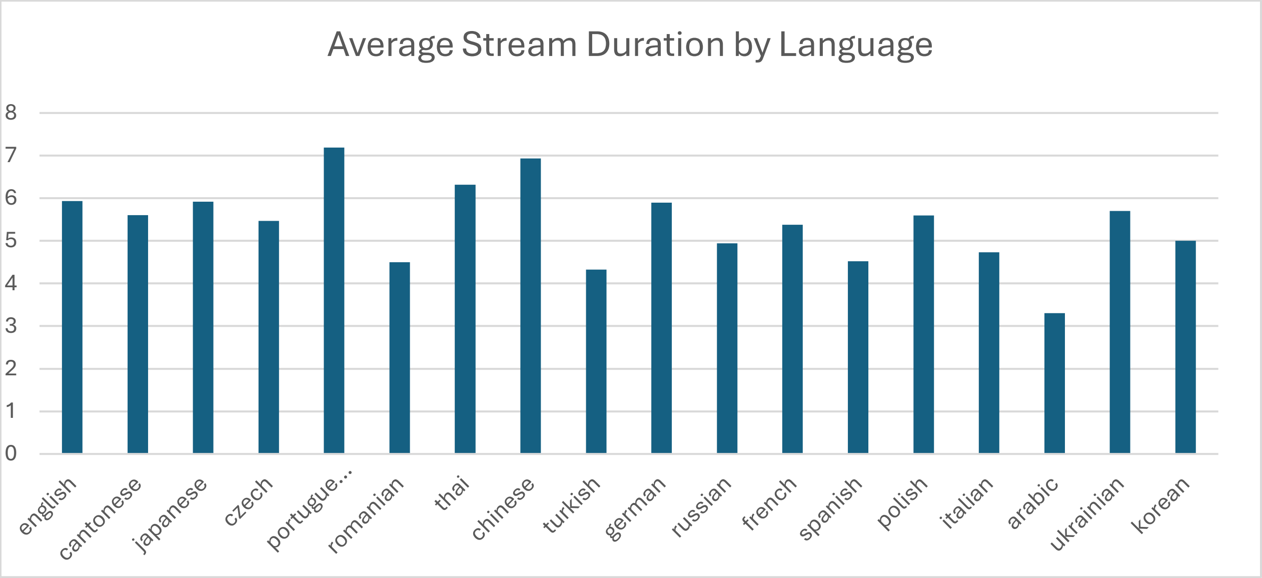


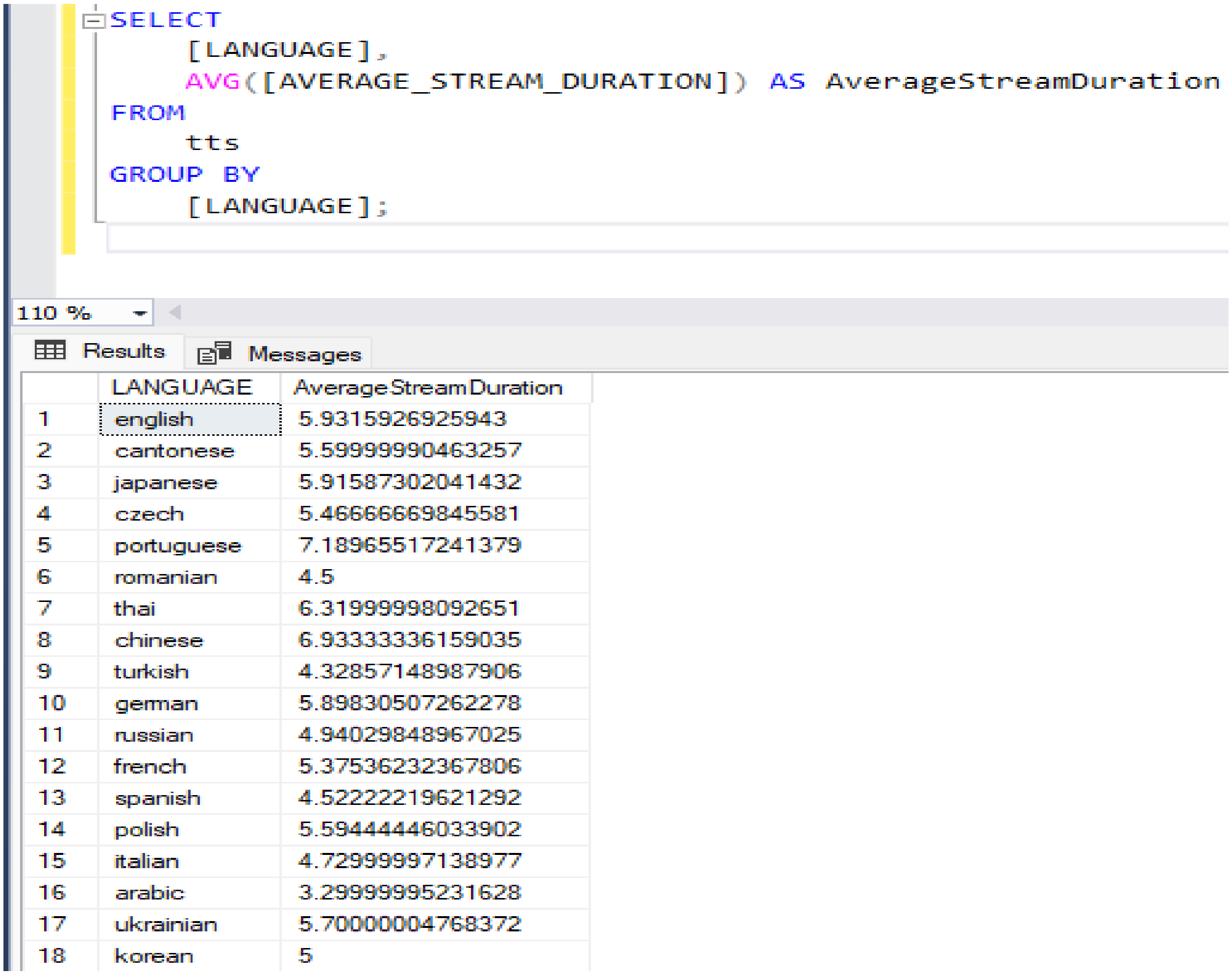
```
        language,
        SUM(total_followers) AS TotalFollowers
FROM
        tts
GROUP BY
        language
ORDER BY
        TotalFollowers DESC
```

	language	TotalFollowers
1	english	448107896
2	spanish	184847700
3	portuguese	62884099
4	french	54402700
5	russian	46250000
6	german	43833100
7	japanese	19933500
8	polish	10593000
9	italian	8528100
10	turkish	7663000
11	chinese	6406300
12	thai	2279000
13	czech	1030000
14	korean	587000

Question 3: What is the average stream duration by language?

- Analyzing the average stream duration by language will provide insights into streaming habits.
- Calculate the average stream duration for each language using SQL queries.





```
SELECT
    [LANGUAGE],
    AVG([AVERAGE_STREAM_DURATION]) AS AverageStreamDuration
FROM
    tts
GROUP BY
    [LANGUAGE];
```

	LANGUAGE	AverageStreamDuration
1	english	5.9315926925943
2	cantonese	5.59999990463257
3	japanese	5.91587302041432
4	czech	5.46666669845581
5	portuguese	7.18965517241379
6	romanian	4.5
7	thai	6.31999998092651
8	chinese	6.93333336159035
9	turkish	4.32857148987906
10	german	5.89830507262278
11	russian	4.94029848967025
12	french	5.37536232367806
13	spanish	4.52222219621292
14	polish	5.59444446033902
15	italian	4.72999997138977
16	arabic	3.29999995231628
17	ukrainian	5.70000004768372
18	korean	5

Conclusion

In conclusion, this Twitch data analysis project aims to explore and understand streaming trends, popular games, and viewer behavior on the platform. By importing, cleaning, and visualizing Twitch streamers data, we can uncover valuable insights that can inform content creators, marketers, and platform developers. This README serves as a guide to setting up the project, performing data analysis, and visualizing key metrics to answer specific questions about Twitch streaming data.

Dataset:

Top 1000 Twitch Streamers Data (kaggle.com)