Twitch Data Project

SQL Analysis and Matplotlib Visualization

Stephen Ellingson

Before We Begin

The Information

Twitch is the world's leading live streaming platform for gamers, with 15 million daily active users. Using data to understand its users and products is one of the main responsibilities of the Twitch Science Team.

In this project, I'll be analyzing stream viewership and chat room data.

The Data

There are two separate .csv files provided for this project, one for stream data and one for chat data.

This data is for the date January 1, 2015 and covers a variety of information (discussed in SQL analysis).

SQL + Visualization

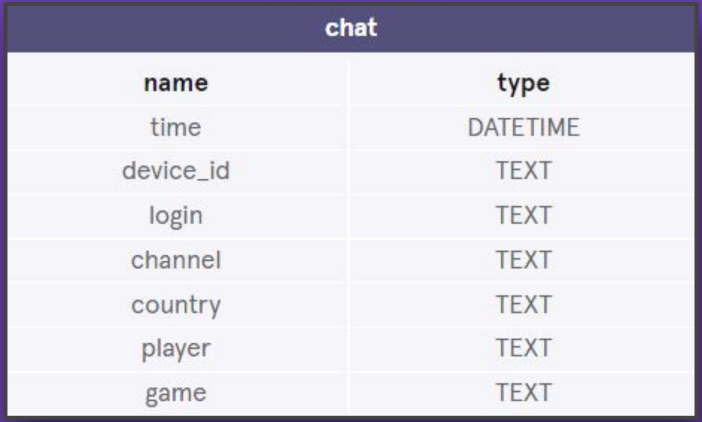
Throughout this presentation, I will be conducting EDA using SQL to understand trends in Twitch data.

I will then use Python's matplotlib library to visualize the main insights.

SQL Analysis

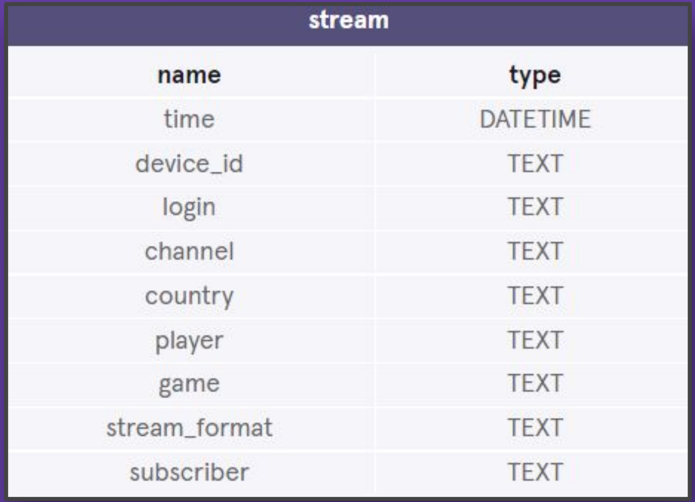
Observing Variables

- Using a few SELECT statements, we can see the columns and data types for each table.
- Game, player, and time columns will most likely be the best to work with for relevant information.



chat

name	type
time	DATETIME
device_id	TEXT
login	TEXT
channel	TEXT
country	TEXT
player	TEXT
game	TEXT

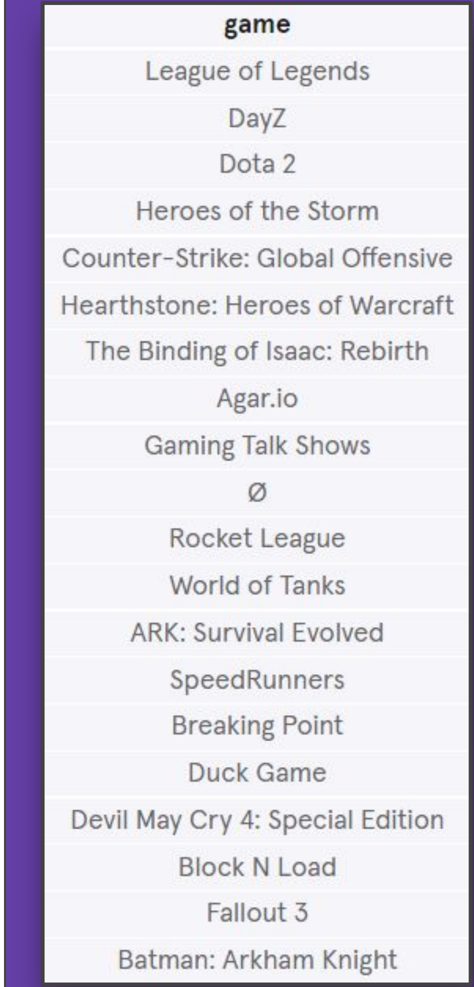


stream

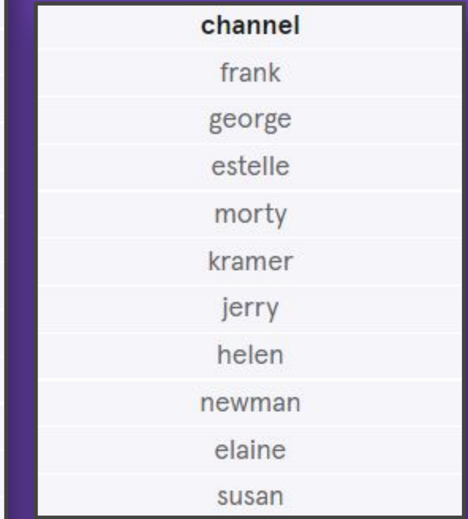
name	type
time	DATETIME
device_id	TEXT
login	TEXT
channel	TEXT
country	TEXT
player	TEXT
game	TEXT
stream_format	TEXT
subscriber	TEXT

Games and Channels

- To find what games and channels are present in the data, SELECT DISTINCT was used to identify the unique values.
- There are 20 game rows (one for non-specified streams) and 10 channels.

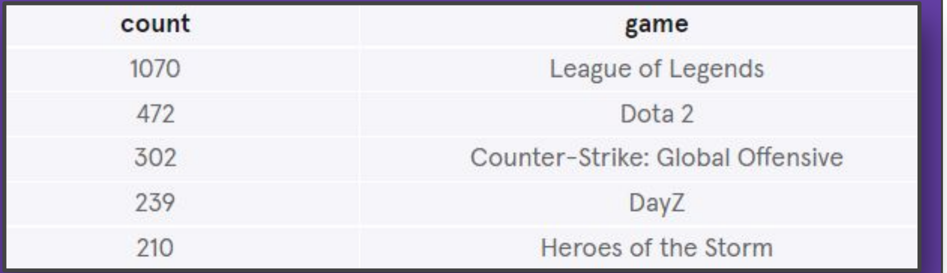


game
League of Legends
DayZ
Dota 2
Heroes of the Storm
Counter-Strike: Global Offensive
Hearthstone: Heroes of Warcraft
The Binding of Isaac: Rebirth
Agar.io
Gaming Talk Shows
Ø
Rocket League
World of Tanks
ARK: Survival Evolved
SpeedRunners
Breaking Point
Duck Game
Devil May Cry 4: Special Edition
Block N Load
Fallout 3
Batman: Arkham Knight

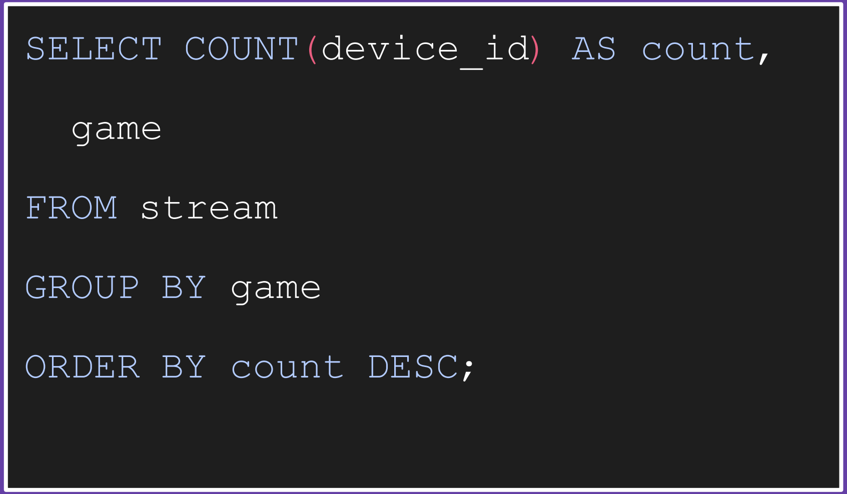


channel
frank
george
estelle
morty
kramer
jerry
helen
newman
elaine
susan

Viewer Counts by Game



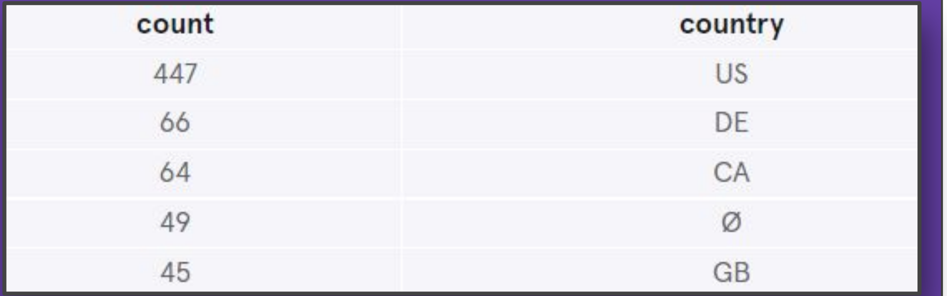
count	game
1070	League of Legends
472	Dota 2
302	Counter-Strike: Global Offensive
239	DayZ
210	Heroes of the Storm



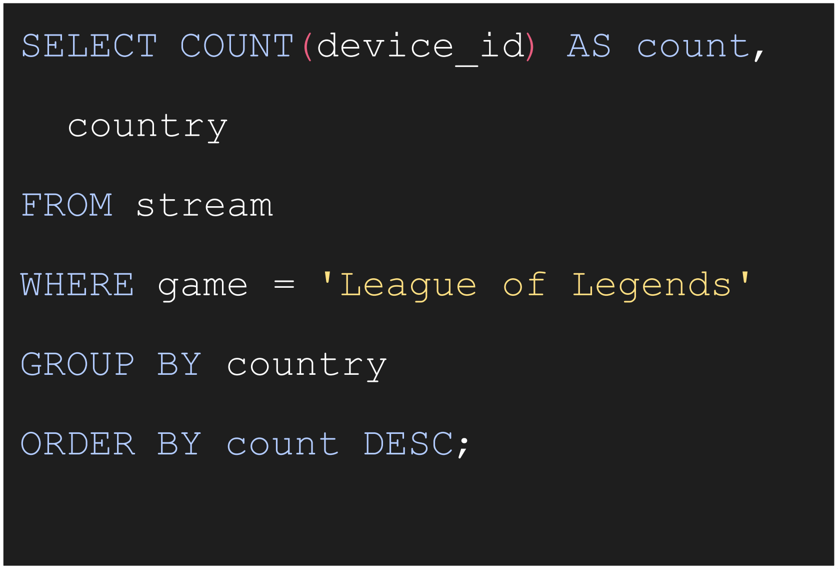
```
SELECT COUNT(device_id) AS count,

  game

FROM stream

GROUP BY game

ORDER BY count DESC;
```

- To see which games are the most popular, I queried a count of table rows and grouped by game in descending order.
- The top five games are LoL, Dota 2, CS: GO, DayZ, and Heroes of the Storm.

LoL Viewers by Country



count	country
447	US
66	DE
64	CA
49	Ø
45	GB

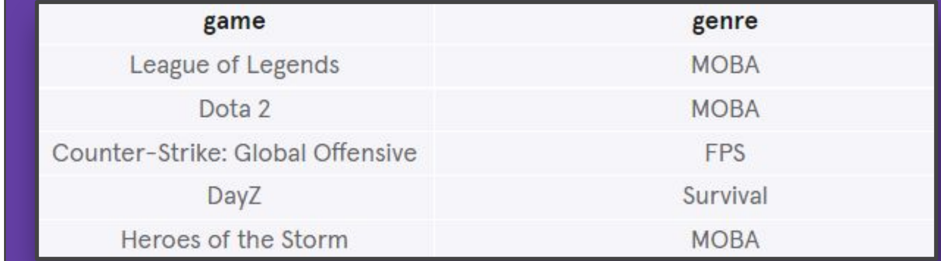


```
SELECT COUNT(device_id) AS count,
  country
FROM stream
WHERE game = 'League of Legends'
GROUP BY country
ORDER BY count DESC;
```

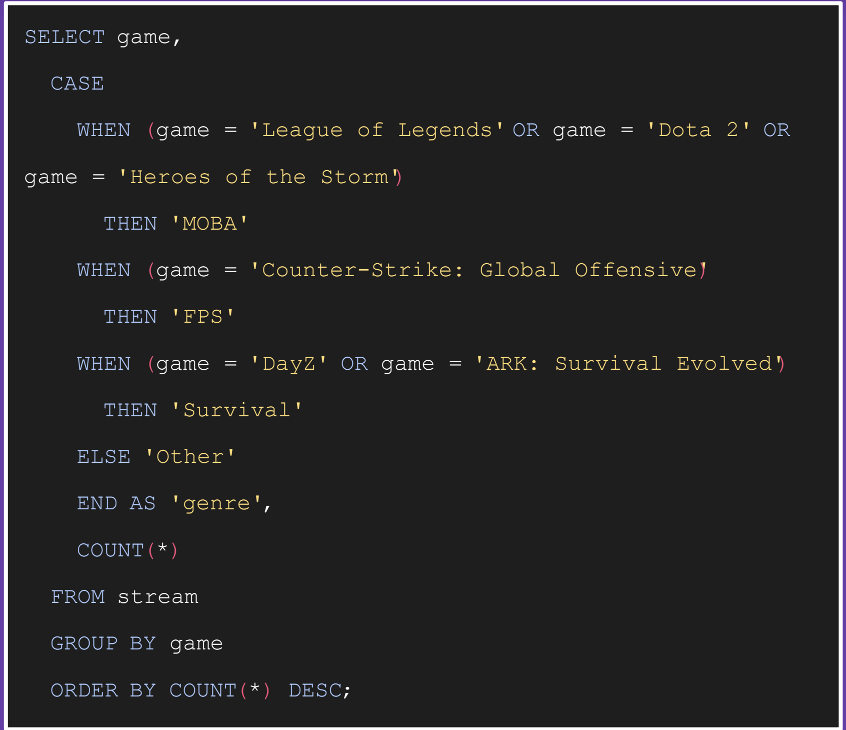
- Diving deeper into the game count analysis, I looked at League of Legends viewers by country.
- The majority of viewers are from the United States, with Germany, Canada, and Great Britain following (#4 is non-specified).

Game Genre Categories

- To see what kinds of genres are most popular, I created a genre column using a CASE statement.
- Most games with the highest viewer counts are Multiplayer Online Battle Arenas, or MOBAs, and all of the top five are multiplayer games.



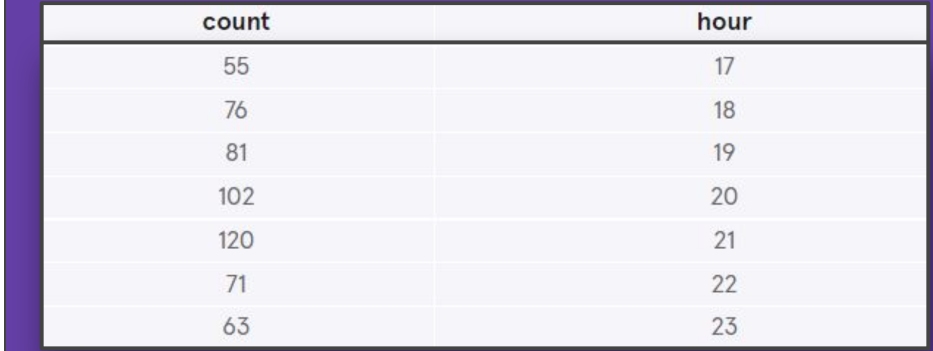
game	genre
League of Legends	MOBA
Dota 2	MOBA
Counter-Strike: Global Offensive	FPS
DayZ	Survival
Heroes of the Storm	MOBA



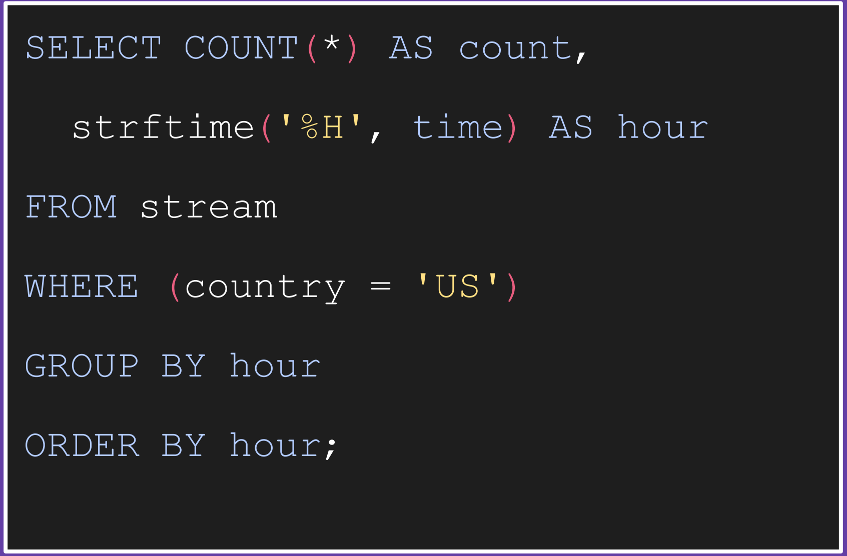
```
SELECT game,
  CASE
    WHEN (game = 'League of Legends' OR game = 'Dota 2' OR
game = 'Heroes of the Storm')
      THEN 'MOBA'
    WHEN (game = 'Counter-Strike: Global Offensive')
      THEN 'FPS'
    WHEN (game = 'DayZ' OR game = 'ARK: Survival Evolved')
      THEN 'Survival'
    ELSE 'Other'
    END AS 'genre',
    COUNT(*)
  FROM stream
  GROUP BY game
  ORDER BY COUNT(*) DESC;
```

Viewer Count per Hour

- By segmenting U.S. viewer login time by hour, hotspots in viewership can be easily seen.
- The highest spike in viewer counts can be observed at the end of the day, with the highest point at 9:00 PM.

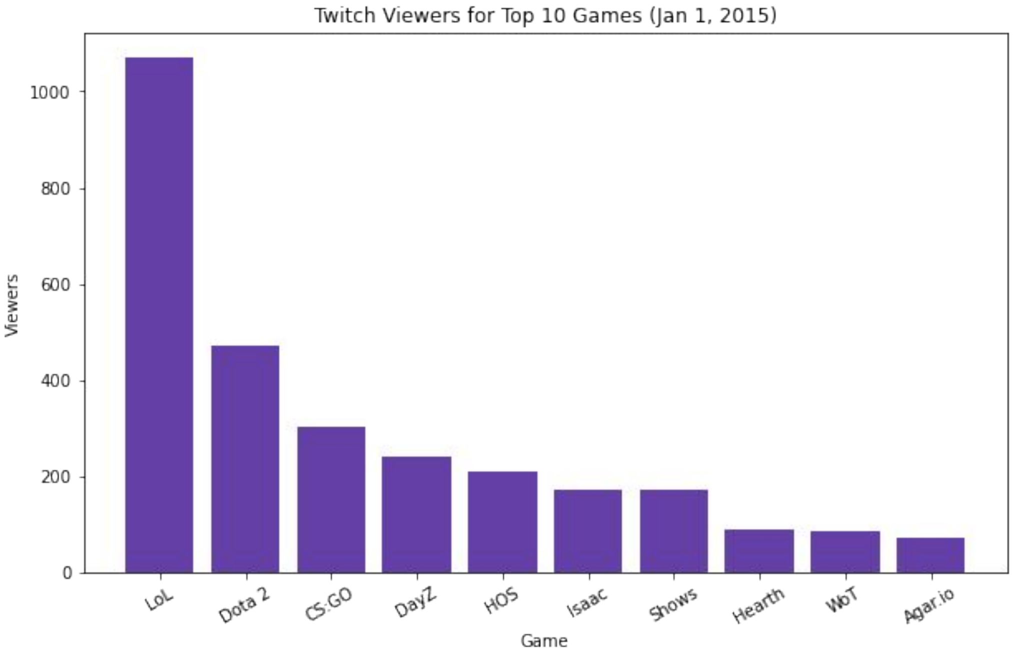


count	hour
55	17
76	18
81	19
102	20
120	21
71	22
63	23



```
SELECT COUNT(*) AS count,

  strftime('%H', time) AS hour

FROM stream

WHERE (country = 'US')

GROUP BY hour

ORDER BY hour;
```

Matplotlib Visualization

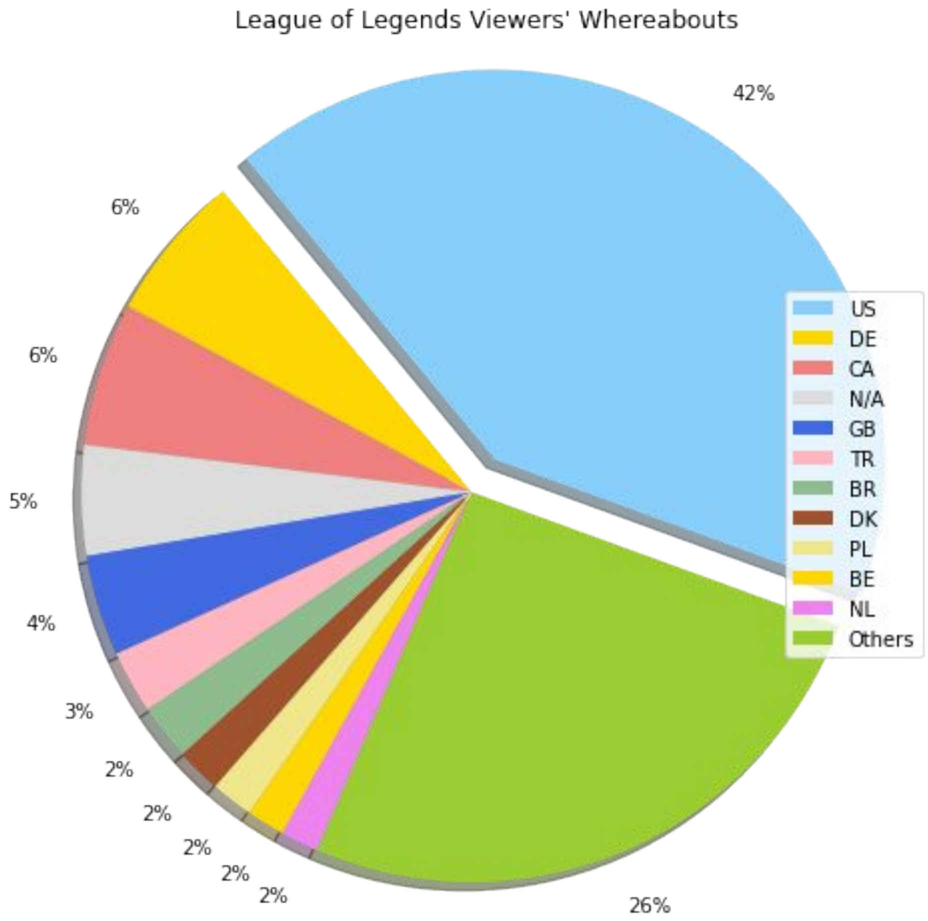


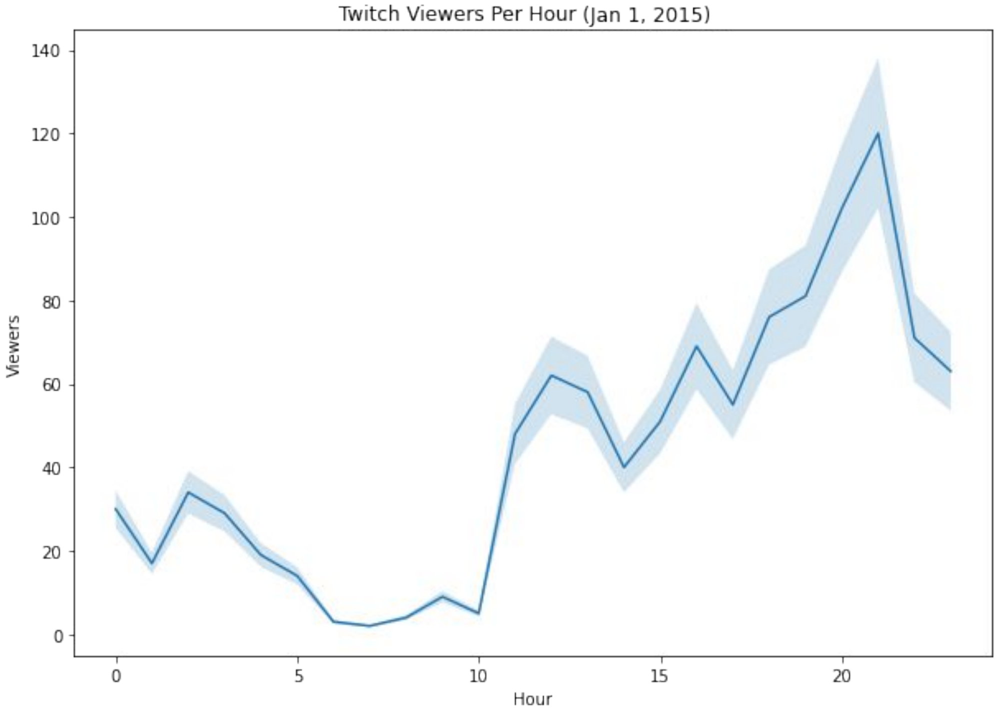
Most Popular Games Streamed

- This plot displays the top ten most popular games streamed on Twitch for January 1st, 2015.
- LoL has more than double the viewership than Dota 2 at almost 1,100 watching during this particular day.

LoL Viewers by Country

- This plot displays where League of Legends viewers tuned in on January 1st, 2015.
- The United States takes up a considerably larger percentage than any other country.





Viewer Count by Hour

- After 10 AM, there is a dramatic increase in viewers
- The highest peak occurs late at night (around 9 PM)
- A 15% error rate is displayed to compensate for some viewers potentially leaving their browsers open.

Conclusion

After analyzing the data through SQL querying, there were many insights that could be visualized. From the ones that I highlighted, we found that:

1. League of Legends was the most popular game to watch on Twitch (more than double the viewership of any other game)
2. For the top-viewed game, almost half of viewers tuned in from the United States
3. After a sudden increase in activity after 10 AM, most viewers were present on Twitch late in the night, with the most people on at 9 PM.

Feel free to provide feedback or ask questions, and thank you for viewing!