Nayan Kadhre - Milestone 2 – Data, Methods, Analysis Reporting

1. Project Title

Predicting Follower Growth on Twitch: The Impact of Hours Streamed

2. Research Question / Problem Statement

- **Main Question:** Does the number of hours a streamer spends streaming predict the total number of followers gained on Twitch?
- **Hypothesis:** Streamers who spend more hours streaming tend to gain more followers due to increased visibility and audience engagement.

3. Dataset Information

Dataset Source: Kaggle – Top Streamers on Twitch (Aayush Mishra, 2021)
About Data:

Columns	Description	Type
Channel	Name of the Twitch channel	Categorical
Watch time(Minutes)	Total minutes watched by the viewers	Numeric (int)
Stream time(minutes)	Total minutes streamed by the streamer	Numeric (int)
Peak viewers	Highest viewers the channel had.	Numeric (int)
Average viewers	Average viewers the channel has.	Numeric (int)
Followers	Number of followers of the channel	Numeric (int)
Followers gained	Follower gained in a period of one year	Numeric (int)
Views gained	Views gained in a period of one year	Numeric (int)
Partnered	Is the channel a twitch partner	Boolean
Mature	Is the content intended for adults	Boolean
Language	The language the streamer streams in	Categorical

4. Data Issues Identified and Handling

Issues Identified	Description	Handling Strategy
Inconsistent column names	Columns Watch time(Minutes), Stream time(minutes) not using consistent naming convention	Renamed columns to Watch time (mins), Stream time (mins) for better readability
Unclear Time Representation	Columns Watch time(Minutes), Stream time(minutes) had time period in mins which really large values, making it harder to understand watch and stream times	Created new columns Watch time (days) and Stream time (days) Converted to days
Inconsistent Stream times	Some stream times exceed the one year collection period, implying data inaccuracies or unit discrepancies.	Reached out to data author to confirm data collection time period, If not response will remove those rows
Outliers	Extreme stream times and followers gained are skewing the averages	Removed outliers using IQR method
High right skewed data.	Important columns “Stream time” and “Followers gained” are highly right skewed.	Applied log transform to reduce skewness

Note:

- As stated in the dataset description, the data was collected over the span of one year.
- No missing values are identified for the dataset.
- For the purpose of our project objective most important columns are “Stream time” and “Followers gained”. (But exploration of other columns can also provide important contextual information).
- This data is biased and does not capture the entire population correctly as these are only the "TOP" 1000 streamers, and streamers with lower stream times may not be captured in this data completely here. So, if our model were to be trained on this data, it may not be able to predict accurately for all the values.