INFO 511 Final Project Report

December 16, 2025

Nayan Kadhre

Milestone 4

1 Project Title

Predicting Follower Growth on Twitch: The Impact of Hours Streamed

1.1 Team or Individual Project

- Type: Individual
- Team Members: Nayan Kadhre

1.2 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.

1.3 Project Description / Focus

This project analyzes the relationship between streaming activity and audience growth on Twitch. Using the Top Streamers dataset, it applies linear regression to determine whether hours streamed significantly predict followers gained. The goal is to identify whether effort directly influences audience expansion.

2 Dataset Description

2.1 Dataset Information

- **Dataset name:** Top Streamers on Twitch (Aayush Mishra, 2021)
- **Source:** Kaggle (Open Soruce)
- **Number of observations:** 1000 observations
- **Key variables:**
  - **Predictor Variables:** Average viewers, Followers, Followers gained, Watch time (days), Stream time (days)
  - **Response Variables:** Percentage of Followers Grained (follower_gained_perc)
- **Data Scope:** Esports/Streaming analytics

2.2 Data Access/Ethical Considerations

- **IRB Requirement:** Not required - dataset contains no personally identifiable information.
- **Data Security/Access Plan:** Data will be accessed directly from Kaggle and analyzed in a local, secure environment.

3 Data Processing and Manipulation

3.1 Exploratory Data Analysis (EDA)

The EDA started with the goal of trying to understand the data, its distribution, and to identify key variables I will be working with, as well as whether any data transformations will be required. The data description on Kaggle provided definitions of the columns. Basic information about the data shape and null values were identified. The summary statistics were calculated, which provided insight about the range, variability, and the central tendencies of the numerical variables. These values were observed specifically for the predictor variables mentioned in section 2.1. Finally, visualizations like histograms and boxplots were used to understand the distribution of the "Followers gained" and "Stream time (days)", followed by a scatter plot to see if any trend can be observed between the same two variables.

3.2 Data Processing

The data processing involved cleaning dataset and stransforming it to be ready for modeling and analysis. This Included the following data issues being handeled:

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
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
Outliers	Extreme stream times and followers gained are skewing the averages	Removed outliers using IQR method
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

Note

- For the purpose of our project objective, the most important columns are "Stream time" and "Followers gained". For these columns, outliers were removed, but the models were tested on both the data with and without outliers to check for any performance difference.
- It was identified that this data is biased and does not capture the entire population correctly, as it includes only the "TOP" 1000 streamers, and streamers with lower stream times may not be captured in this data at all. So, if our model were to be trained on this data, it may not be able to predict accurately for all values.

4 Methods and Analysis

Two types of regression models were used: simple linear and multiple linear. The simple linear model was trained on the key predictor "Stream Time (days)" with percentage of followers gained as the response variable. The variable "Followers gained" was initially supposed to be the response variable, but it was discovered that the relation between "Followers" vs "Followers gained" shows that streamers with an already large number of followers will gain fewer followers, as they already have the most followers they can get. As such, "Percentage of followers gained" was decided on as the response variable. The linear model was used to evaluate the direct relationship between these two variables.

The multiple linear model was trained with additional predictors along with "Stream Time (days)", including "Average viewers", "Followers", "Followers gained", and "Watch Time (days)". The need for using a multiple linear model arose from the insufficient performance of the simple linear model in capturing the data trend. Part of the reason for low model performance is due to data bias, as mentioned earlier, the data is not an accurate representation of the population and has significant spread and variability, making it harder for the simple linear model to fit the data.

Both models were trained on four versions of the dataset: the complete data with outliers, data without outliers, filtered data with special observations removed, and only the special observations (streamers with very few days streamed but high followers gained). This allowed us to see how outliers and unusual cases affect model performance and relationships.

Metrics like MSE, R-squared, and RMSE were compared across the various model and dataset pairs.

5 Results

The performance was compared using p-values and Mean Squared Error (MSE) across both, the simple linear models and the multiple linear models. The table below show this comparison:

Model	Data	MSE	R2 score
Simple Linear Regression	With outliers	0.07	0.02
	Without outliers	0.07	0.02
	Without Special Values	0.04	0.02
	Only special values	0.02	0.04

Model	Data	MSE	R2 score
Multiple Linear Regression	With outliers	0.05	0.38
	Without outliers	0.03	0.54
	Without Special Values	0.02	0.48
	Only special values	0.01	0.74

As per the R2 score values the Multiple Linear Model performed significantly better than the simple linear models across all the different datasets. The models which were trained on the dataset without the special values had a positive trend as compared to all the other models. Removing the outiler did not affect the model performance in any significant way.

6 Conclusion and Next Steps

Based on the analysis and findings in this project, the stream time of a streamer negatively affects follower growth, meaning that the more time a person spends streaming, the fewer followers they will gain. This conclusion is based on the available data, which is limited both in quantity and quality. A more robust dataset, representing the entire population accurately, would yield more reliable results. Hence, for any future analyses, it is recommended to collect a more comprehensive and representative dataset to ensure accurate and reliable insights.