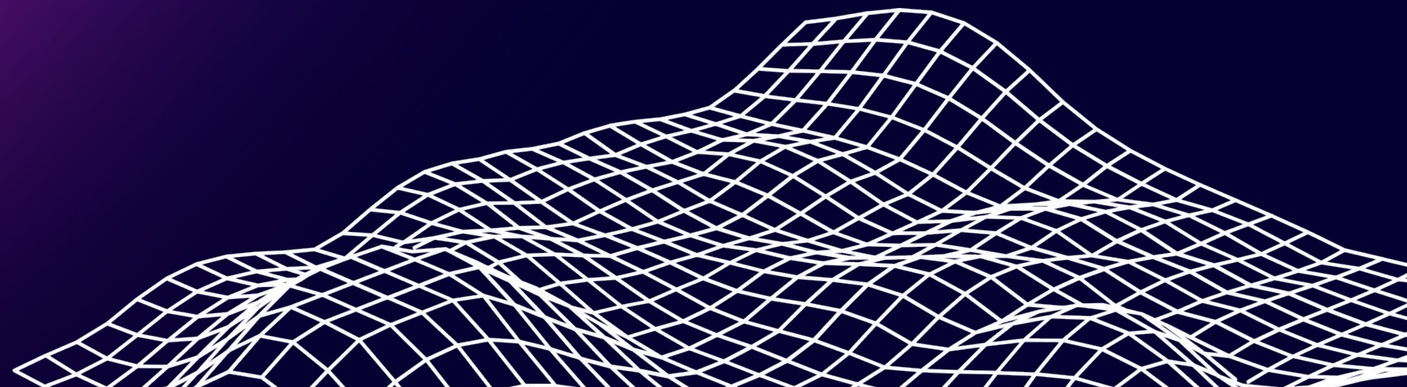
SC1015: MINI PROJECT

PREDICTING TWITCH STREAMER POPULARITY

PETER LI (N2402585L), JASMINE SUN (N2402658K)

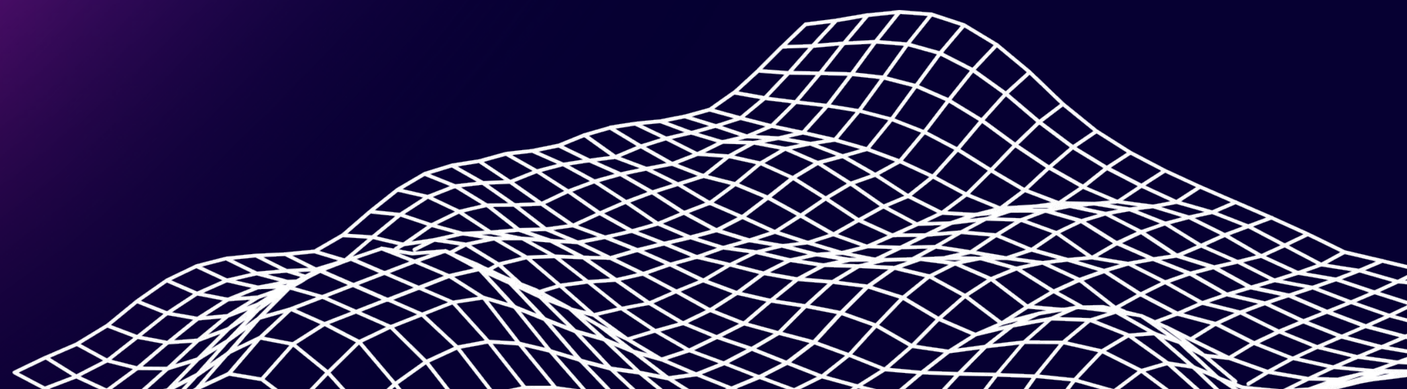
Can we predict the popularity of a Twitch streamer (measured by watch time) based on their streaming habits, and engagement metrics?



DATASET

TOP STREAMERS ON TWITCH

channel	watch_time_minutes	stream_time_minutes	peak_viewers	average_viewers	followers	followers_gained	views_gained	partnered	mature	language
xQcOW	6196161750	215250	222720	27716	3246298	1734810	93036735	True	False	English
summit1g	6091677300	211845	310998	25610	5310163	1370184	89705964	True	False	English
Gaules	5644590915	515280	387315	10976	1767635	1023779	102611607	True	True	Portuguese
ESL_CSGO	3970318140	517740	300575	7714	3944850	703986	106546942	True	False	English
Tfue	3671000070	123660	285644	29602	8938903	2068424	78998587	True	False	English



MOTIVATION

DATA-DRIVEN CONTENT STRATEGY

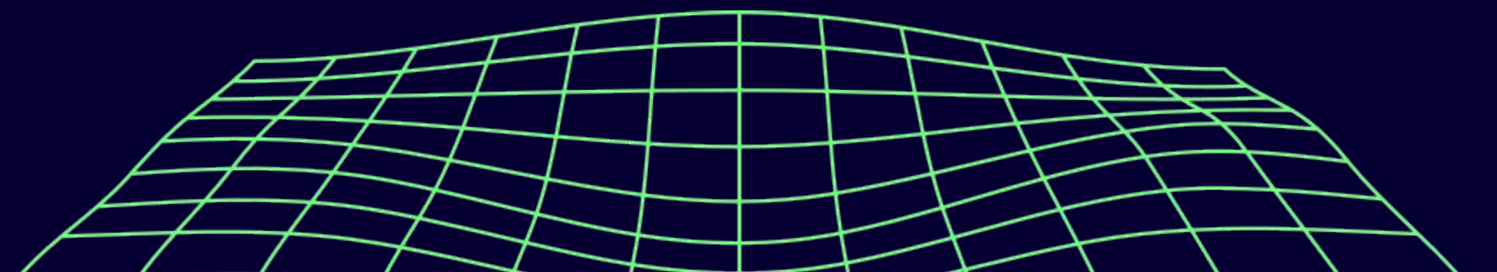
Smarter content strategies

BRAND AND SPONSORSHIP

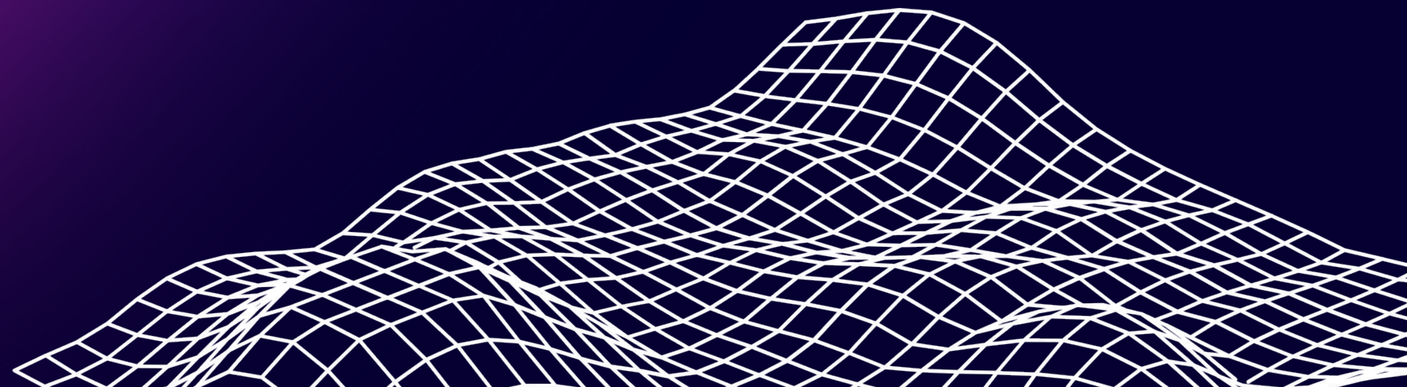
Better sponsorship decisions

PLATFORM OPTIMIZATION

Better insights to discover rising stars and improve recommendation algorithm



1 EXPLORATORY DATA ANALYSIS



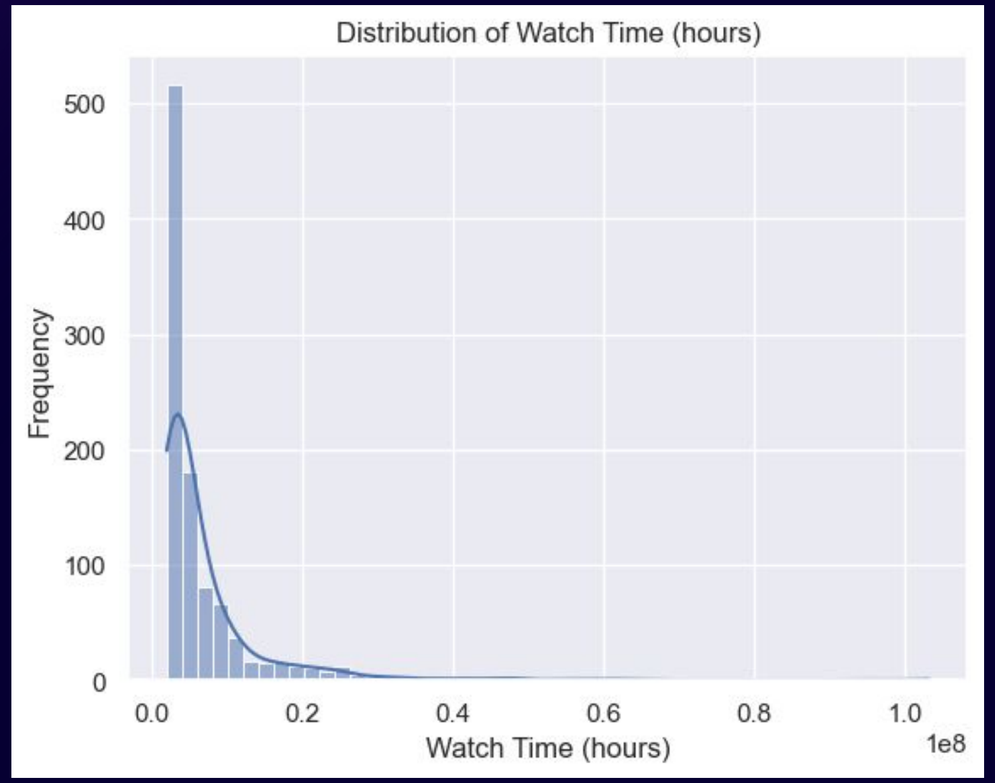
PREPARATION OF DATA

DATA COLLECTION

- Used Kaggle, an online community for data scientists

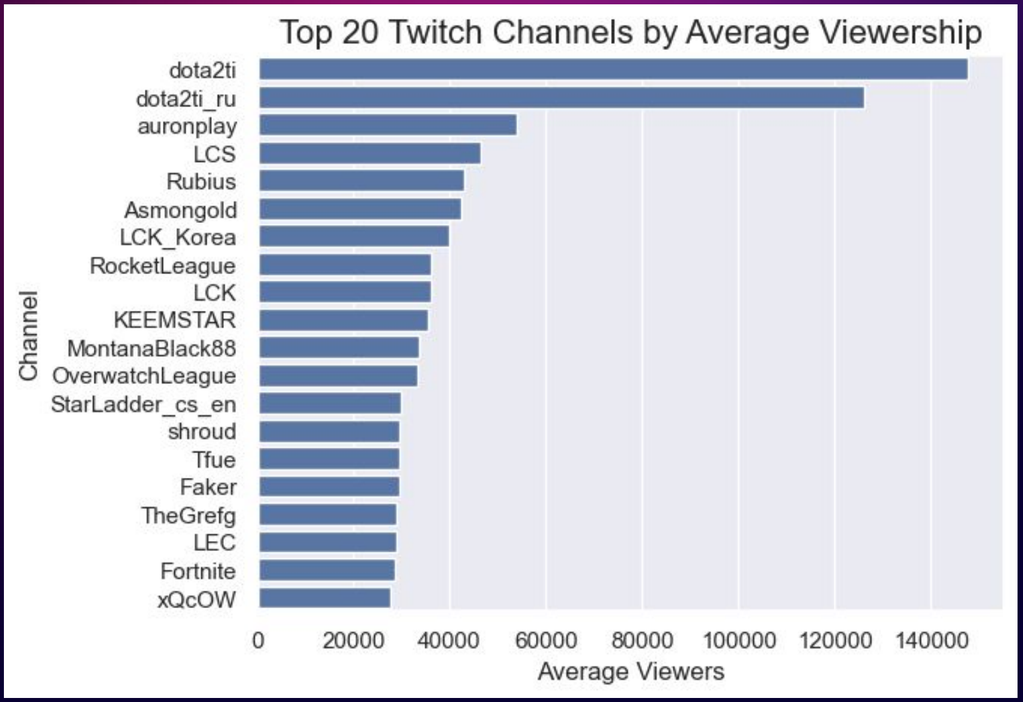
INITIAL PROCESSING

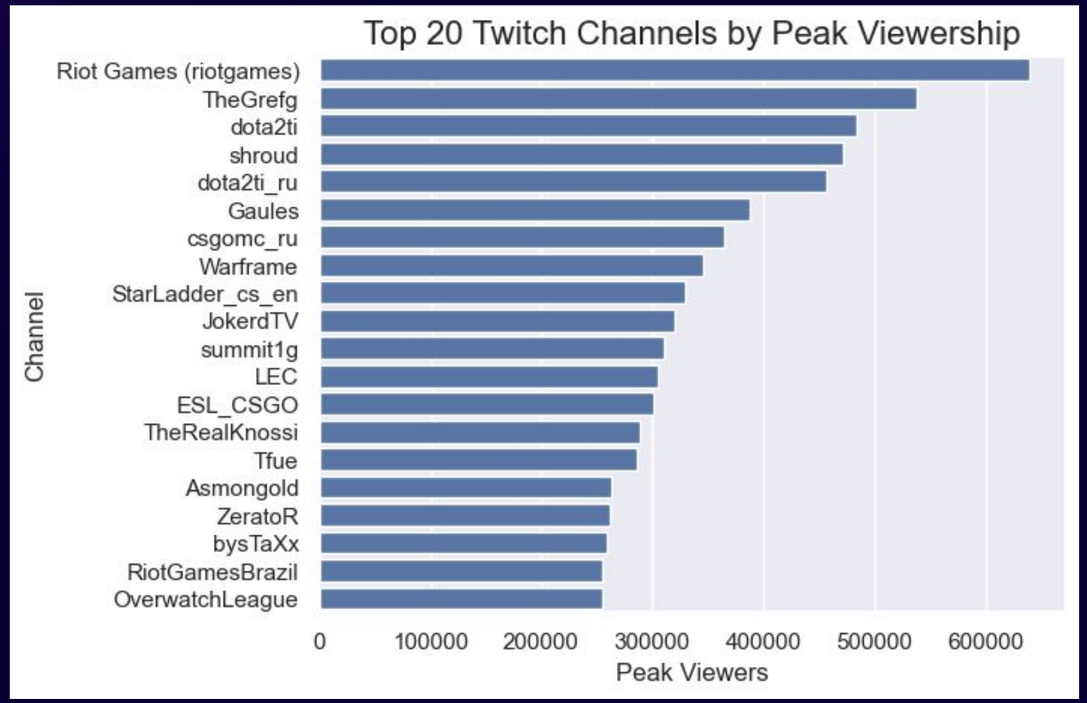
- Rename columns
- Remove null and duplicate rows
- Verify data types
- Apply log transformations to data
- Remove outliers (check correlation and outliers with box plots, bar graphs)



PREPARATION OF DATA

REMOVE OUTLIERS (i.e., TOURNAMENT STREAMING CHANNELS)



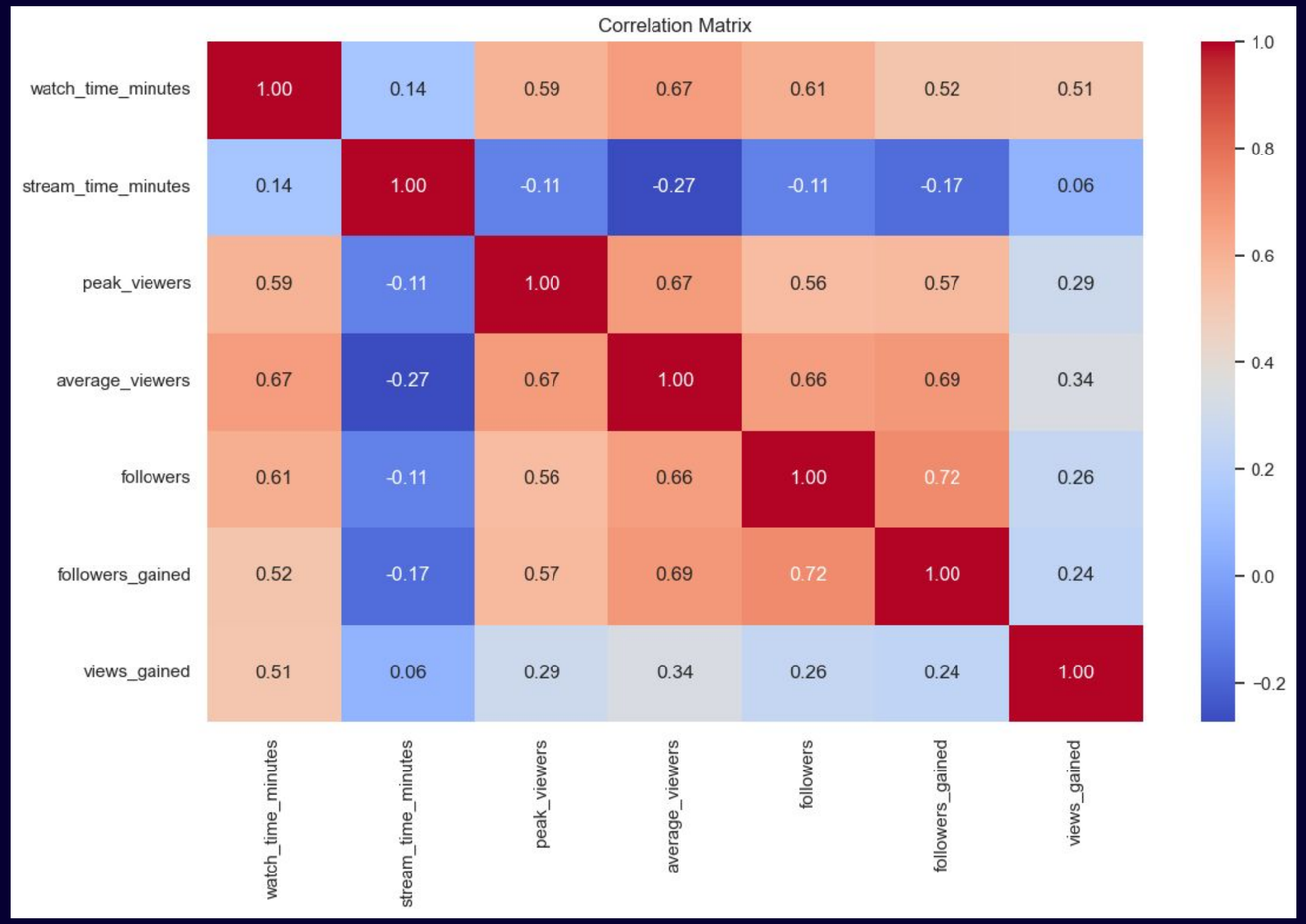


DATA VISUALIZATIONS AND OBSERVATIONS

DRAW DISTRIBUTION FOR ALL VARIABLES

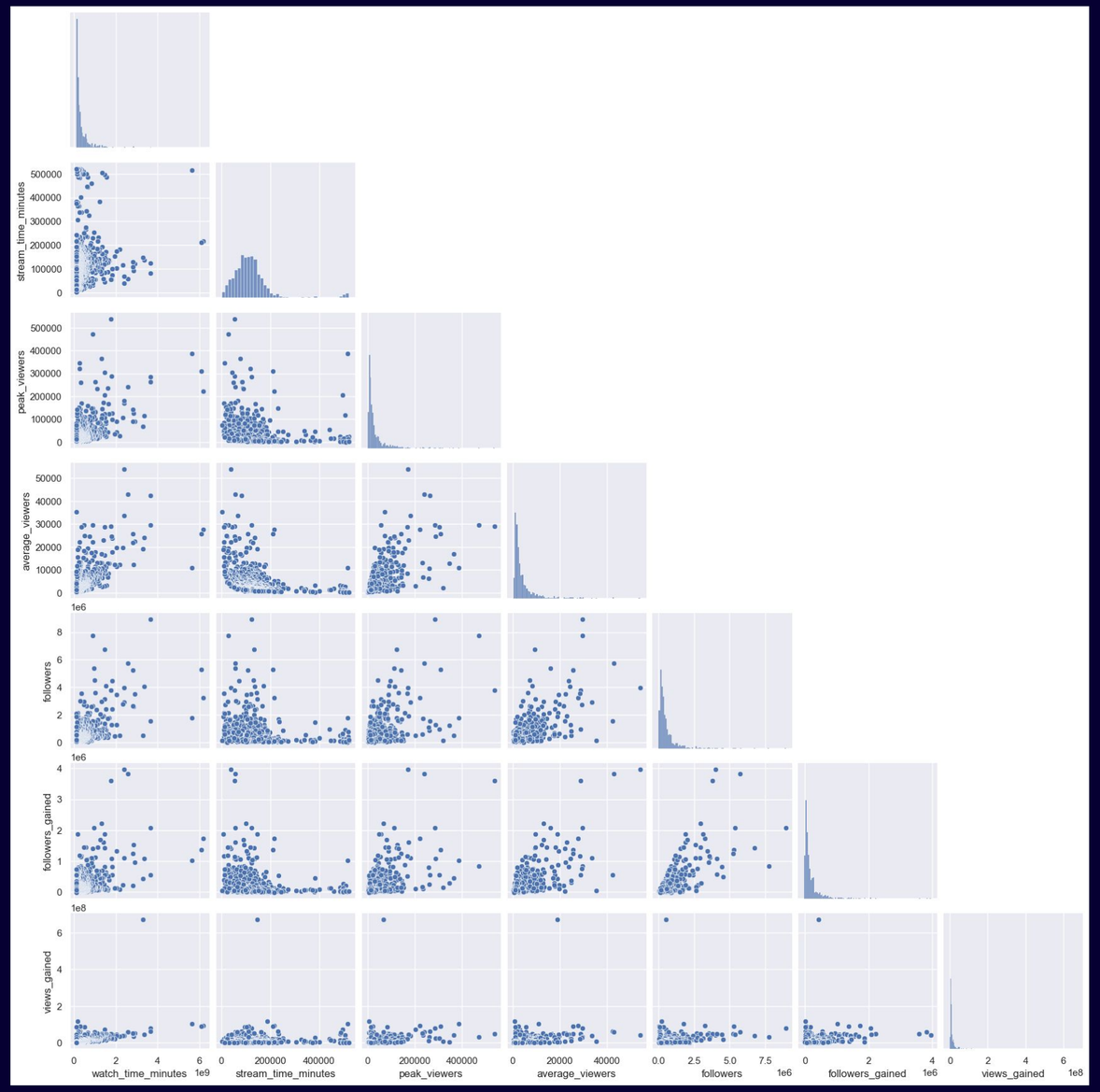
DATA VISUALIZATIONS AND OBSERVATIONS

CORRELATION MATRIX



DATA VISUALIZATIONS AND OBSERVATIONS

PAIR PLOT

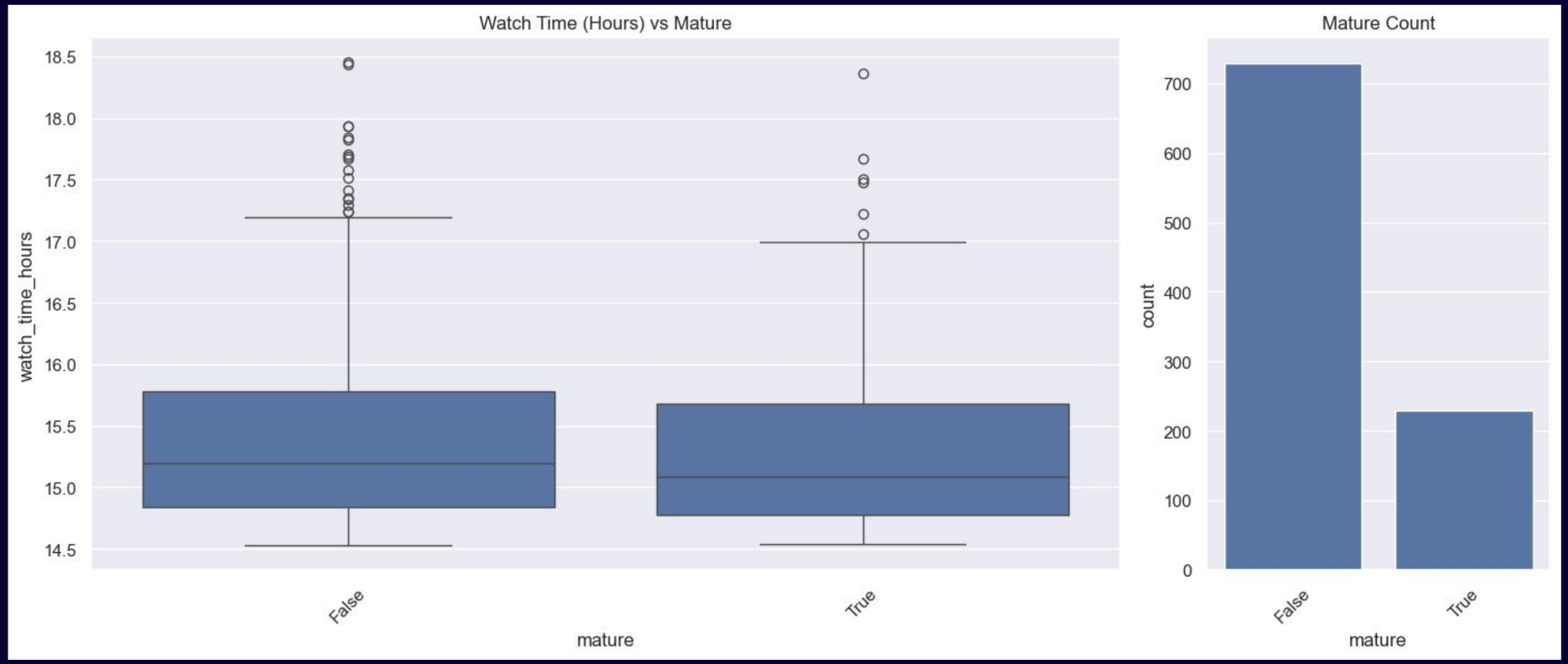


DATA VISUALIZATIONS AND OBSERVATIONS

ANALYSIS OF CATEGORICAL VARIABLES: PARTNERSHIP

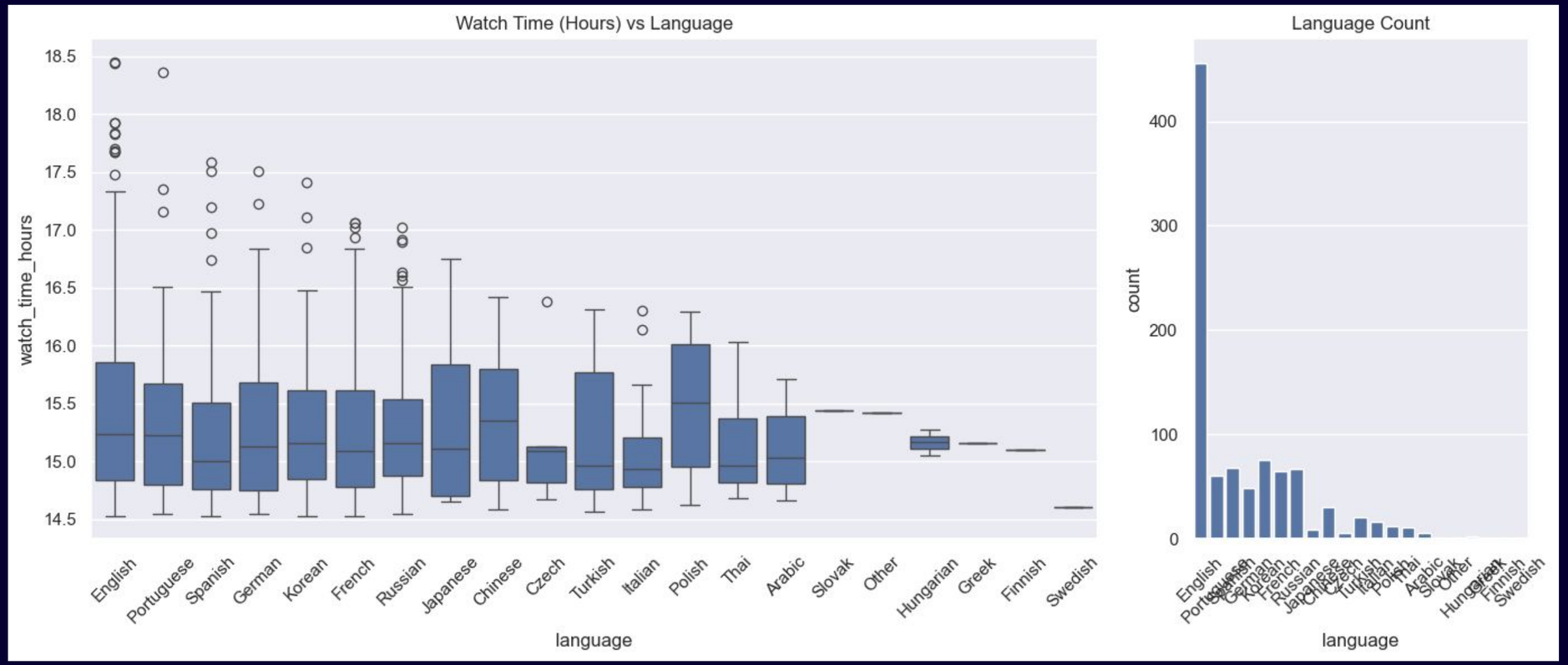
DATA VISUALIZATIONS AND OBSERVATIONS

ANALYSIS OF CATEGORICAL VARIABLES: MATURITY



DATA VISUALIZATIONS AND OBSERVATIONS

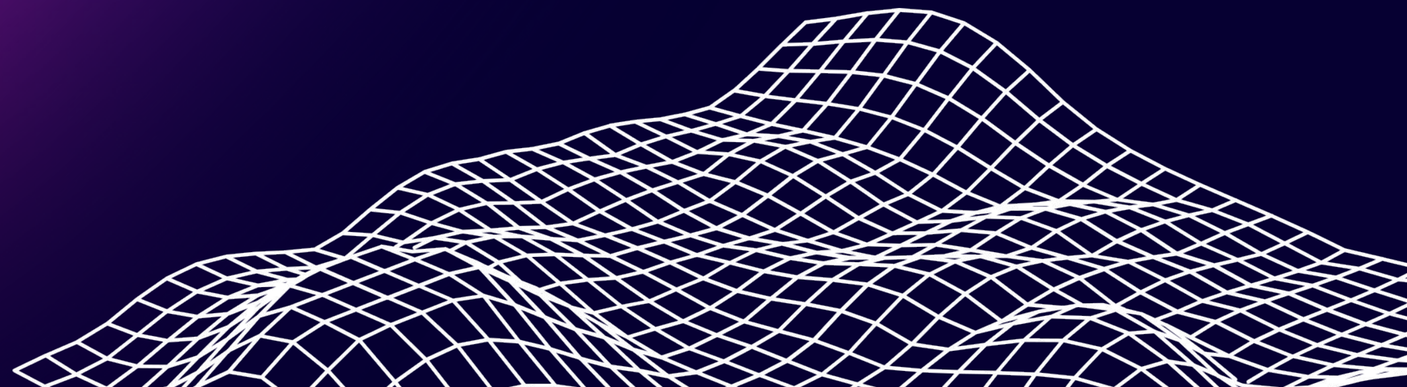
ANALYSIS OF CATEGORICAL VARIABLES: LANGUAGE



FINAL TAKEAWAYS

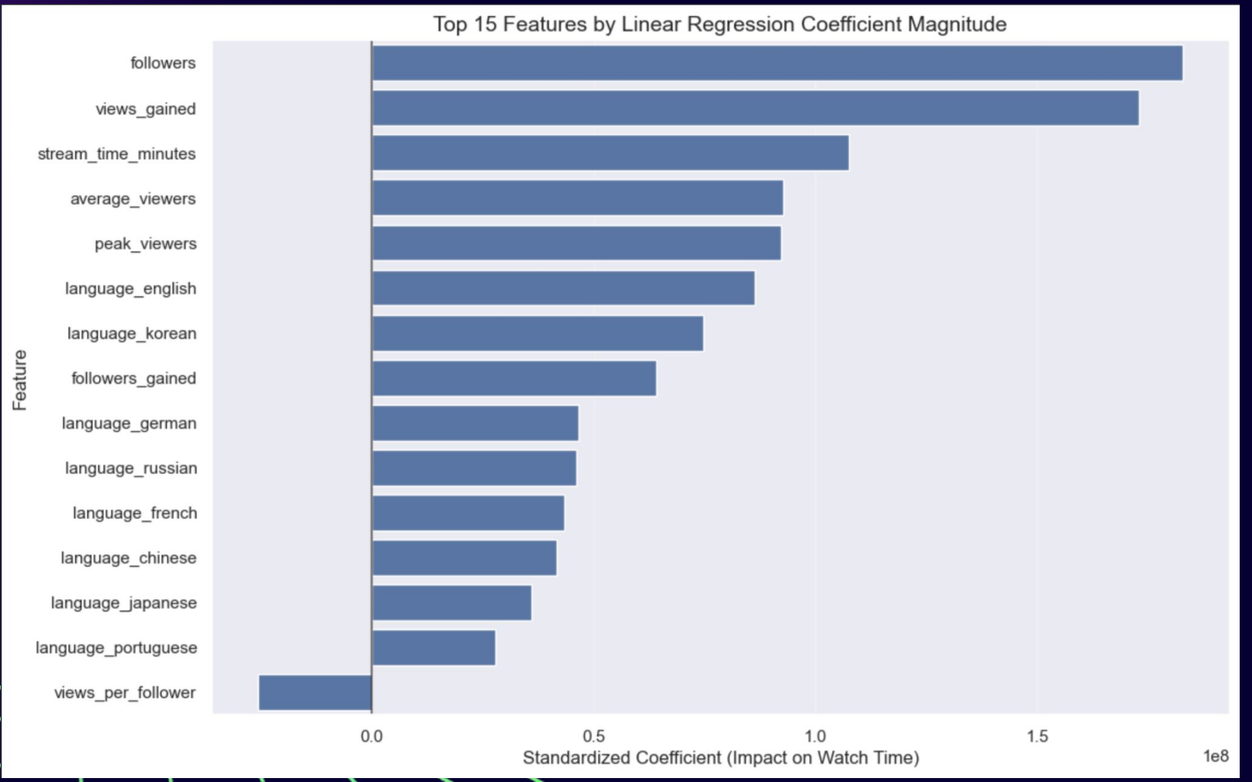
Watch time is highly influenced by viewer engagement (like average viewers), followers, and visibility (like language or partnership status).

2 MACHINE LEARNING TECHNIQUES

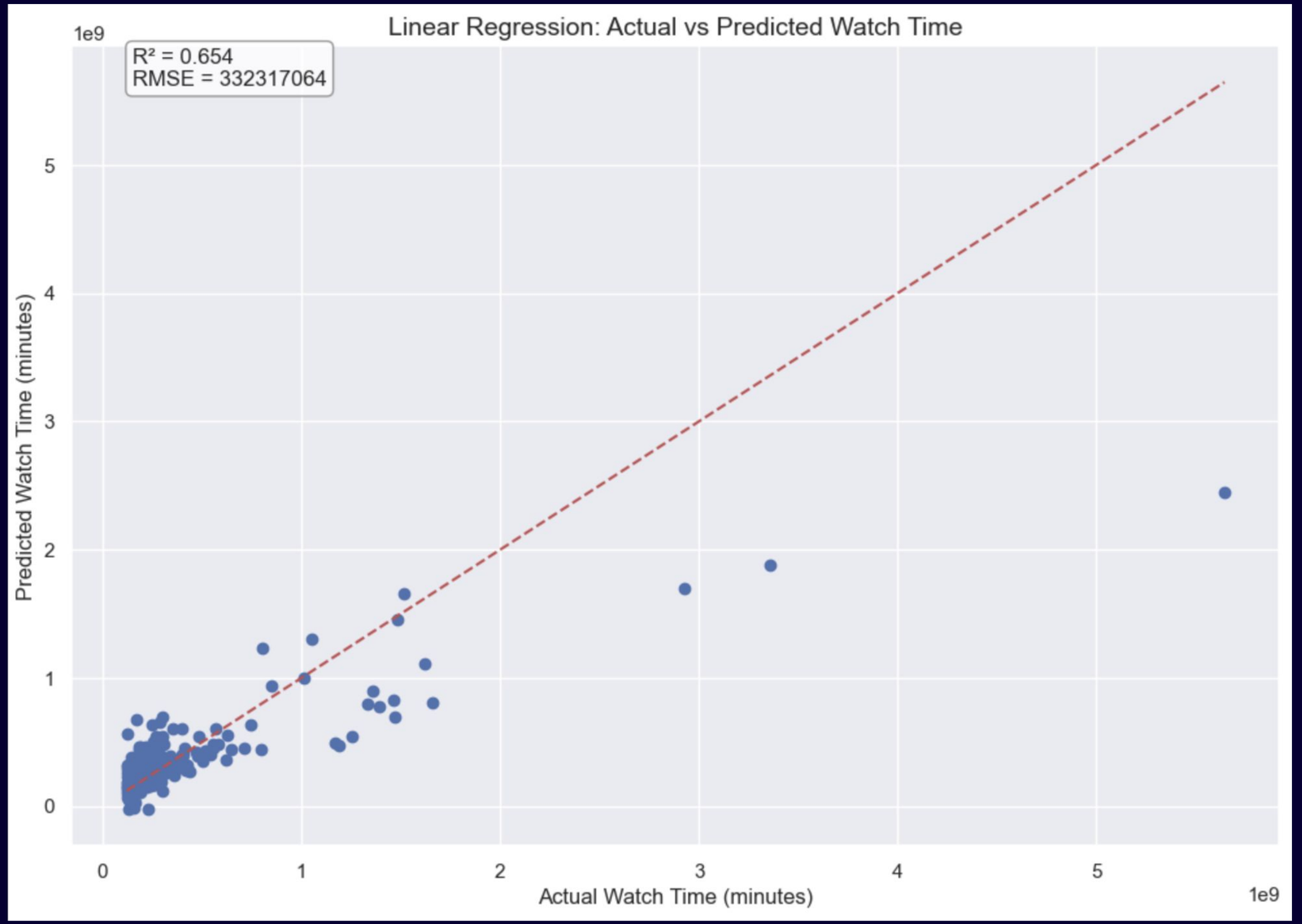


LINEAR REGRESSION

- Multiple features to predict watch time
- Pre-processing by scaling features to ensure all variables contribute proportionally to model
  - Formula: y = B_0 + B_1 + B_2
  - B_0 is intercept, B_1, B_n are feature proportionality

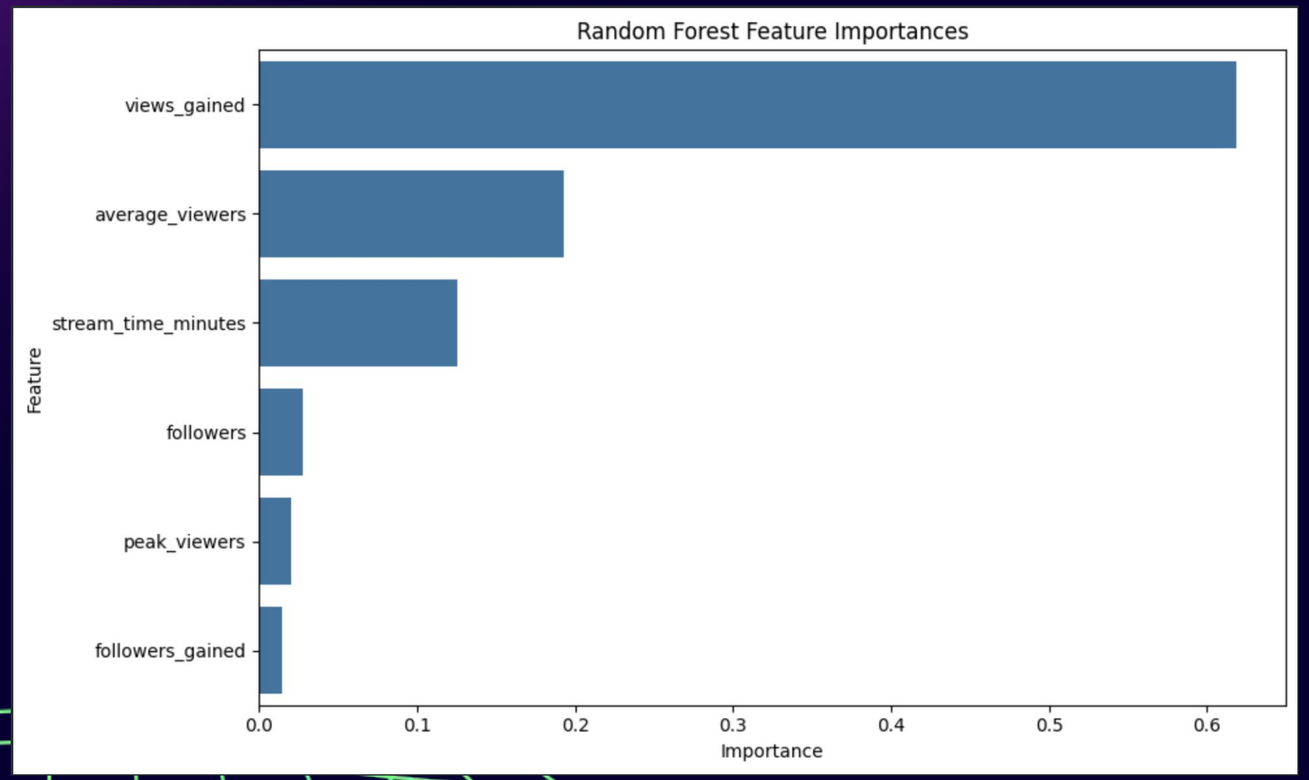


LINEAR REGRESSION PLOT

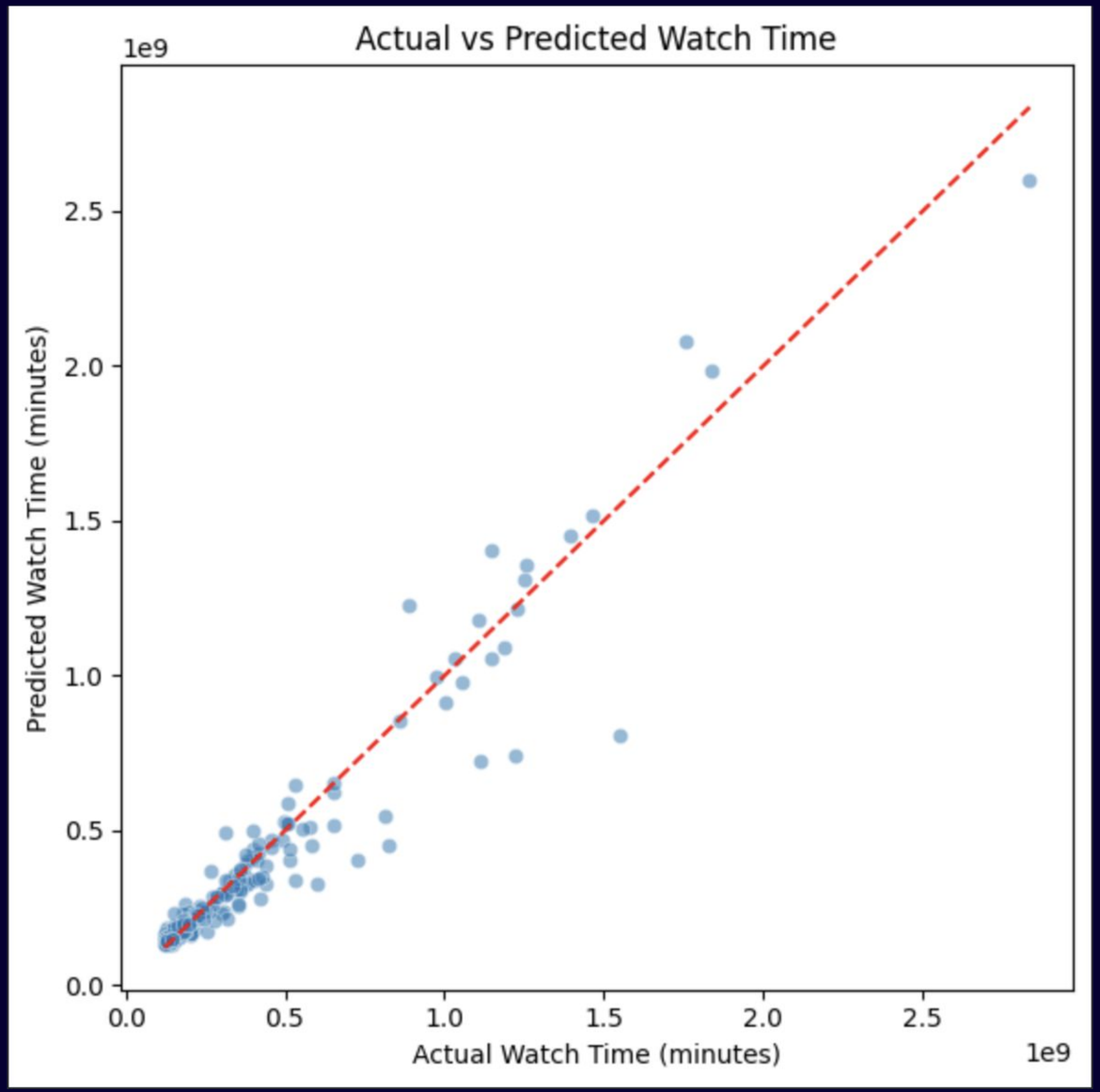


RANDOM FOREST REGRESSION

- Log-transform target for skew
- Many decision trees
  - Each tree makes decision by splitting data based on feature which gives the best improvement to the model’s prediction
- Views gained was most important feature in predicting watch time



RANDOM FOREST REGRESSION PLOT



CONCLUSION

Were we able to predict watch time based on their streaming habits and metrics?

- Yes, we identified several major factors such as views gained, followers, and stream time.

Recommendations

- It is important to focus on factors like viewer engagement and streaming consistently as a top streamer.
- Watch time and average viewers are critical metrics to take in account

Next steps?

- This dataset was pretty old (~5 years old) – would love to work with an updated dataset with a lot more metrics included which could possibly be aggregated from a site like TwitchTracker.com or other sites
- Analyze over a time-series – popular streamers often take breaks and it would be interesting to analyze whether streamers who take breaks are able to make it back to their previous peaks.

GUESS THE GOATS WHO WORKED ON THIS???

PP LI

IM THE BEST

JASMINE