Steam Market Analysis

Team G8: Egor Lukjanenko, Artur Kashnikov, Arkadi Statsenko

Introduction

This project aims to analyze the Steam gaming market, focusing on understanding current market trends, genre popularity, and predicting the success of gaming projects. Steam, being a major platform for digital distribution of games, offers a wide range of data for analysis.

The Data

For this project the data has been crawled from online resources (https://steamspy.com/, https://steamcharts.com/), therefore, created and filtered two datasets, containing general statistics about games and sets of online peaks for each month since July 2012 or game release date.

Algorithms used

Our approach leverages three predictive models to gauge the market potential of games on Steam: RandomForestClassifier: Predicts the likelihood of a game achieving a user score above 70%, using genre data. RandomForestRegressor: Estimates a custom success score based on price, genres, tags, and languages. TensorFlow Regression Model: Utilizes deep learning to predict concurrent player counts (ccu), tuned for accuracy through grid search.

Statistical analysis

We thoroughly examined monthly ccu peaks, explored the popularity of different languages and genres in games, and importantly, discovered meaningful correlations among these attributes, offering new insights into gaming trend.
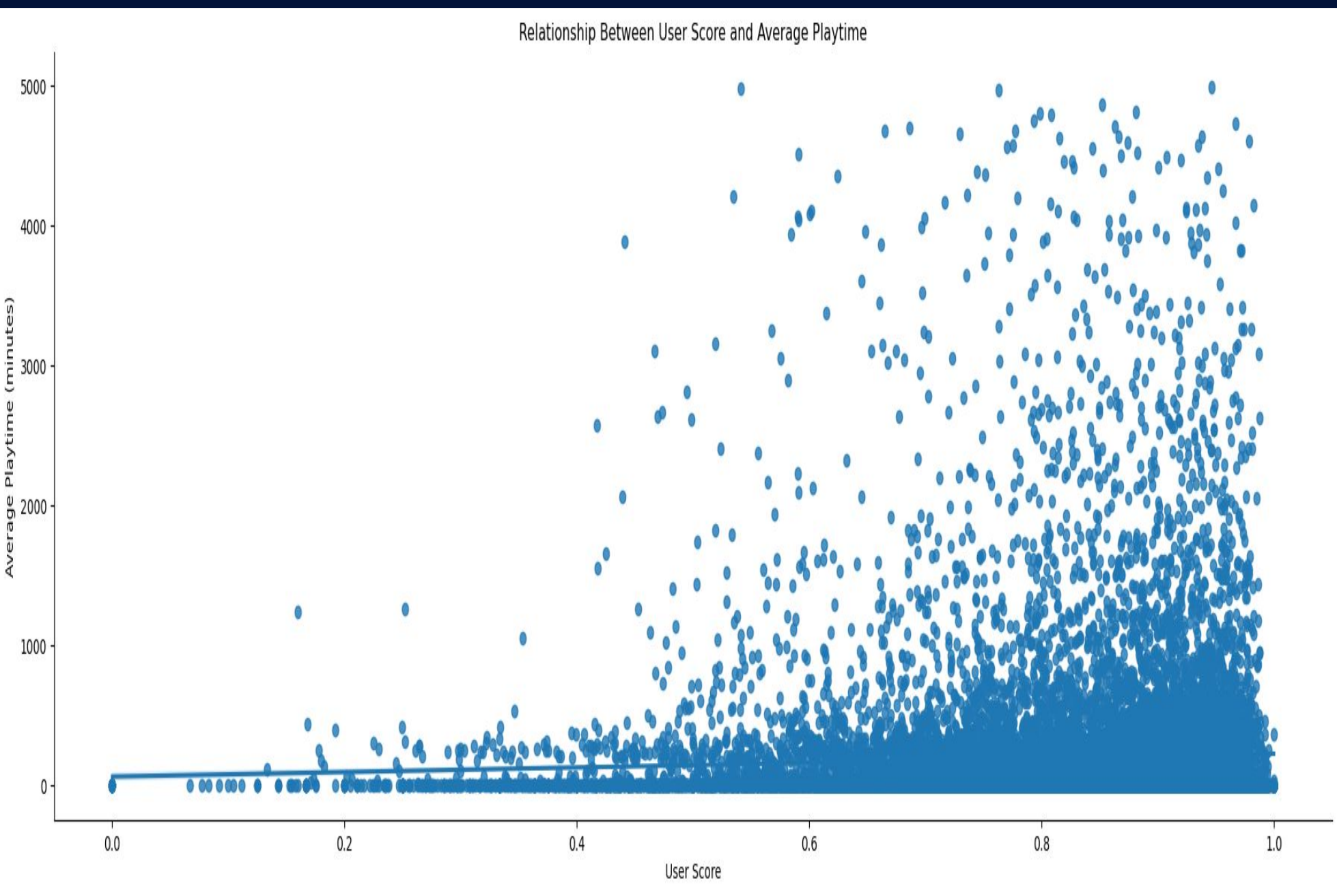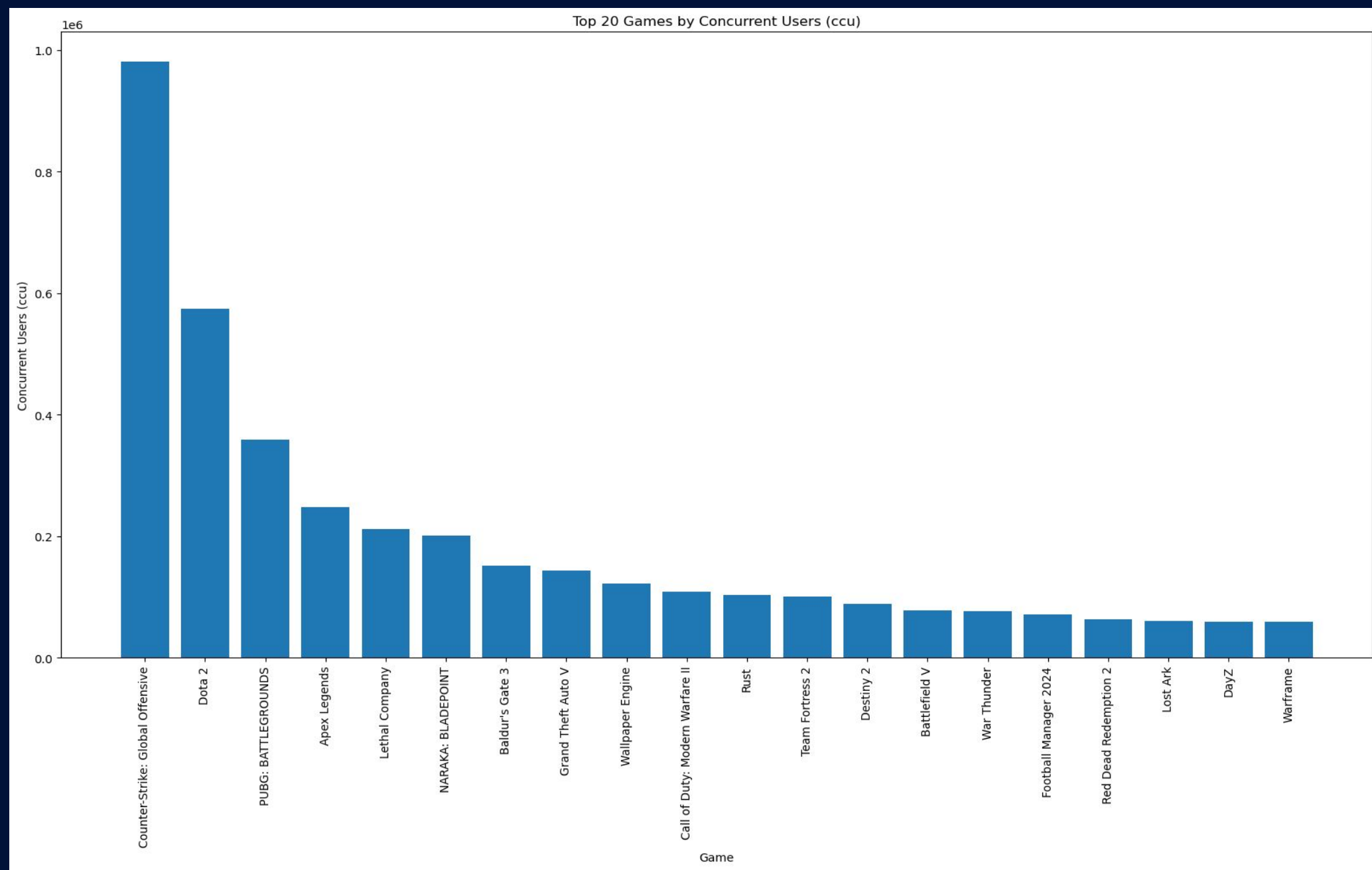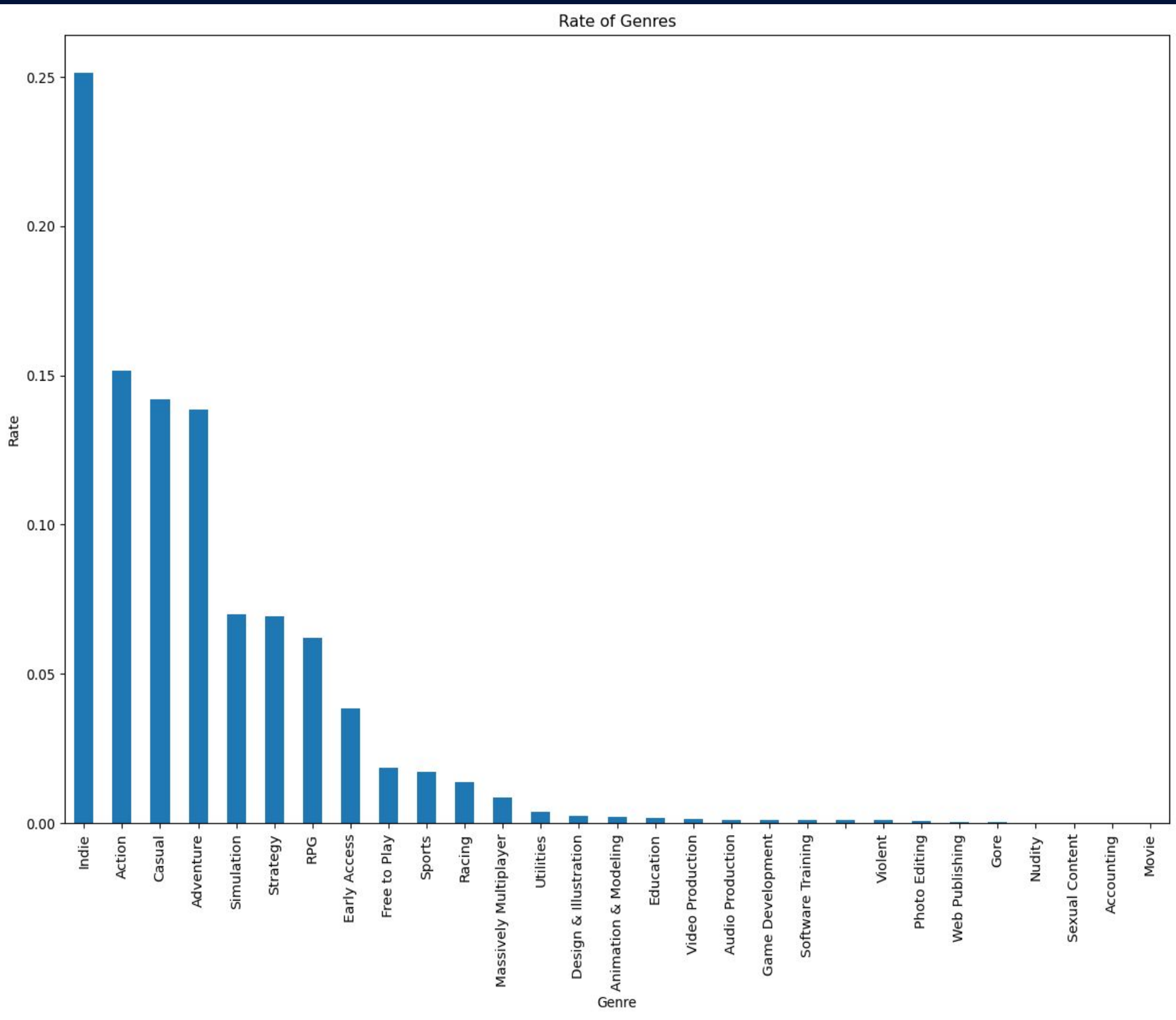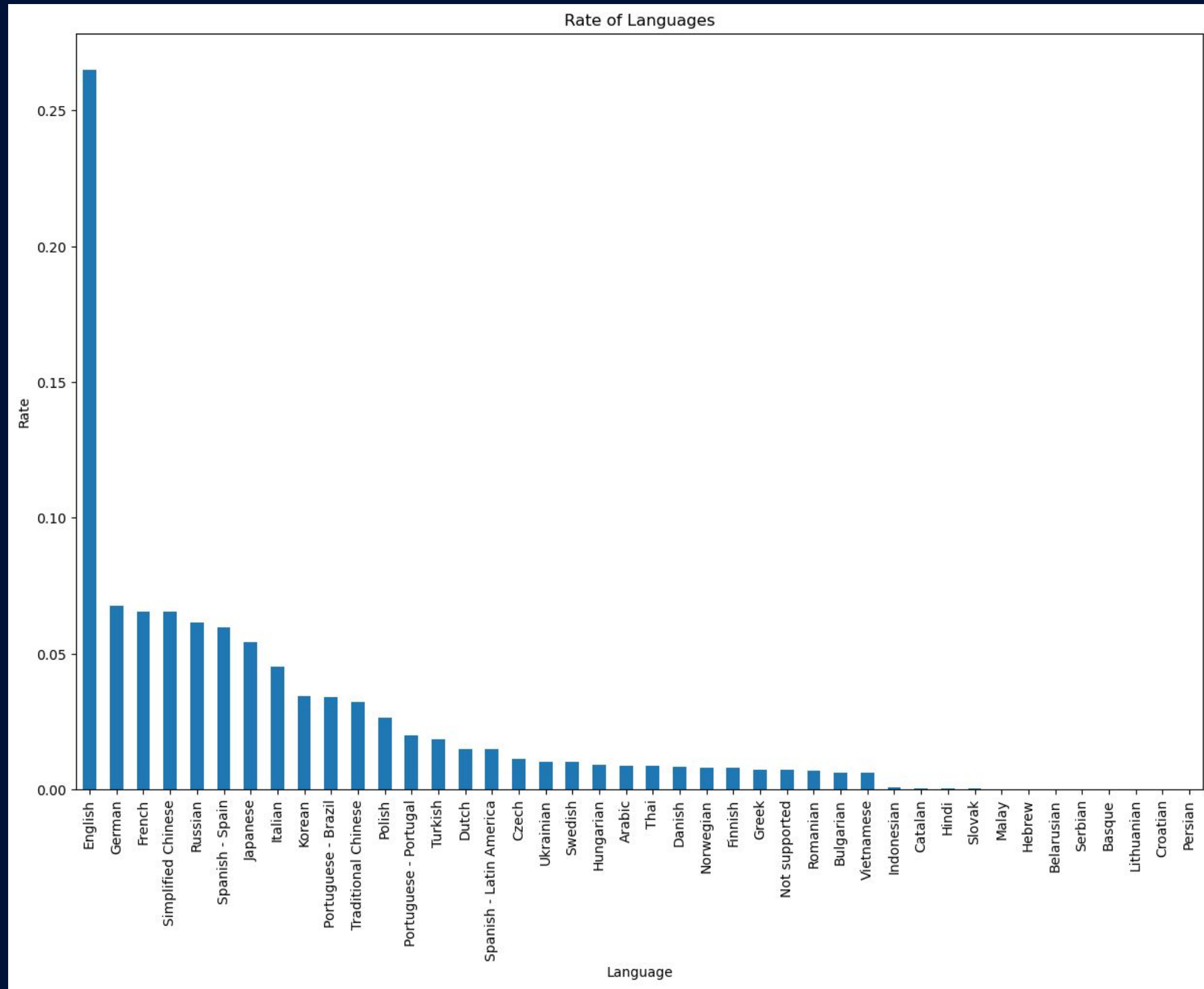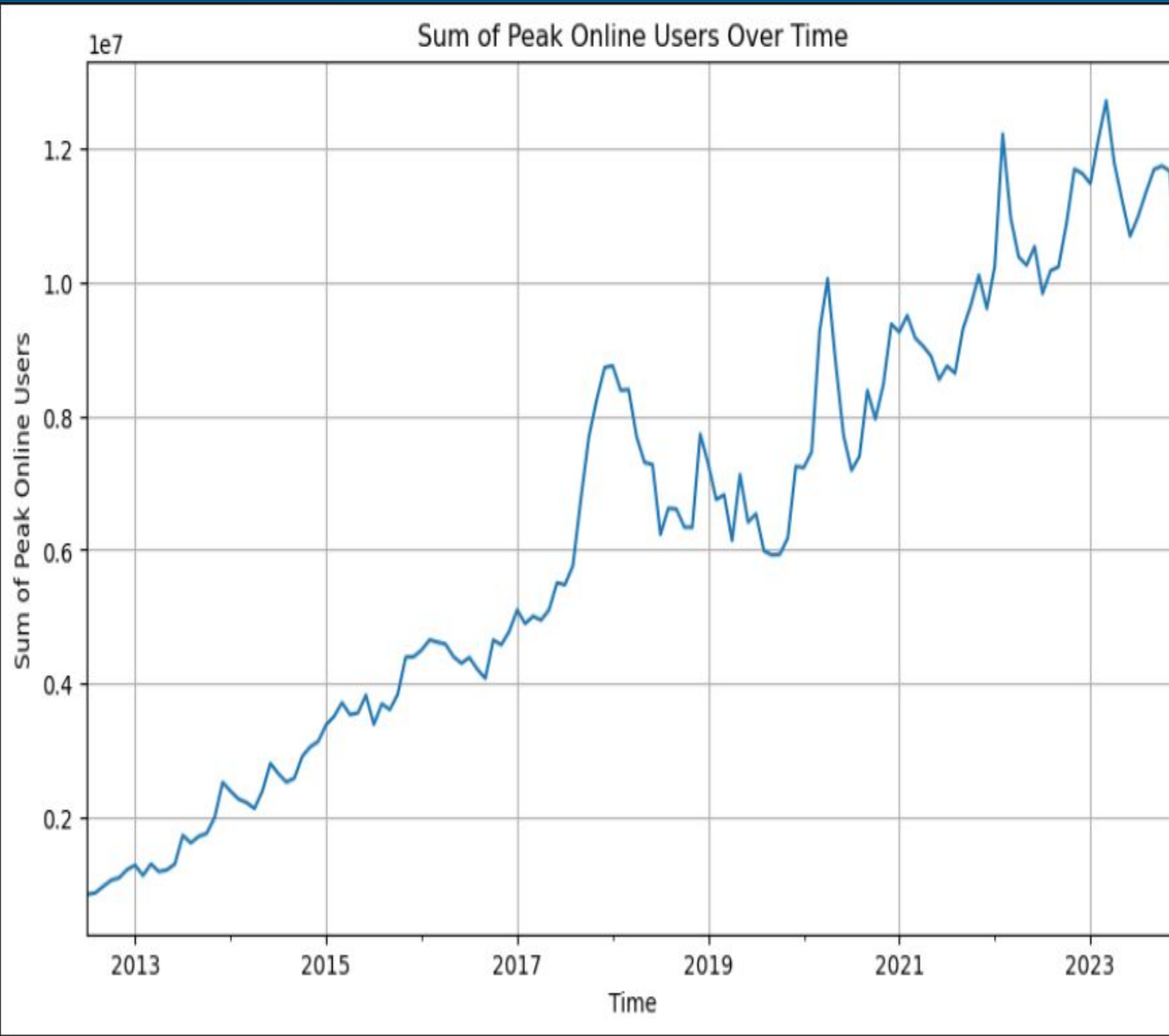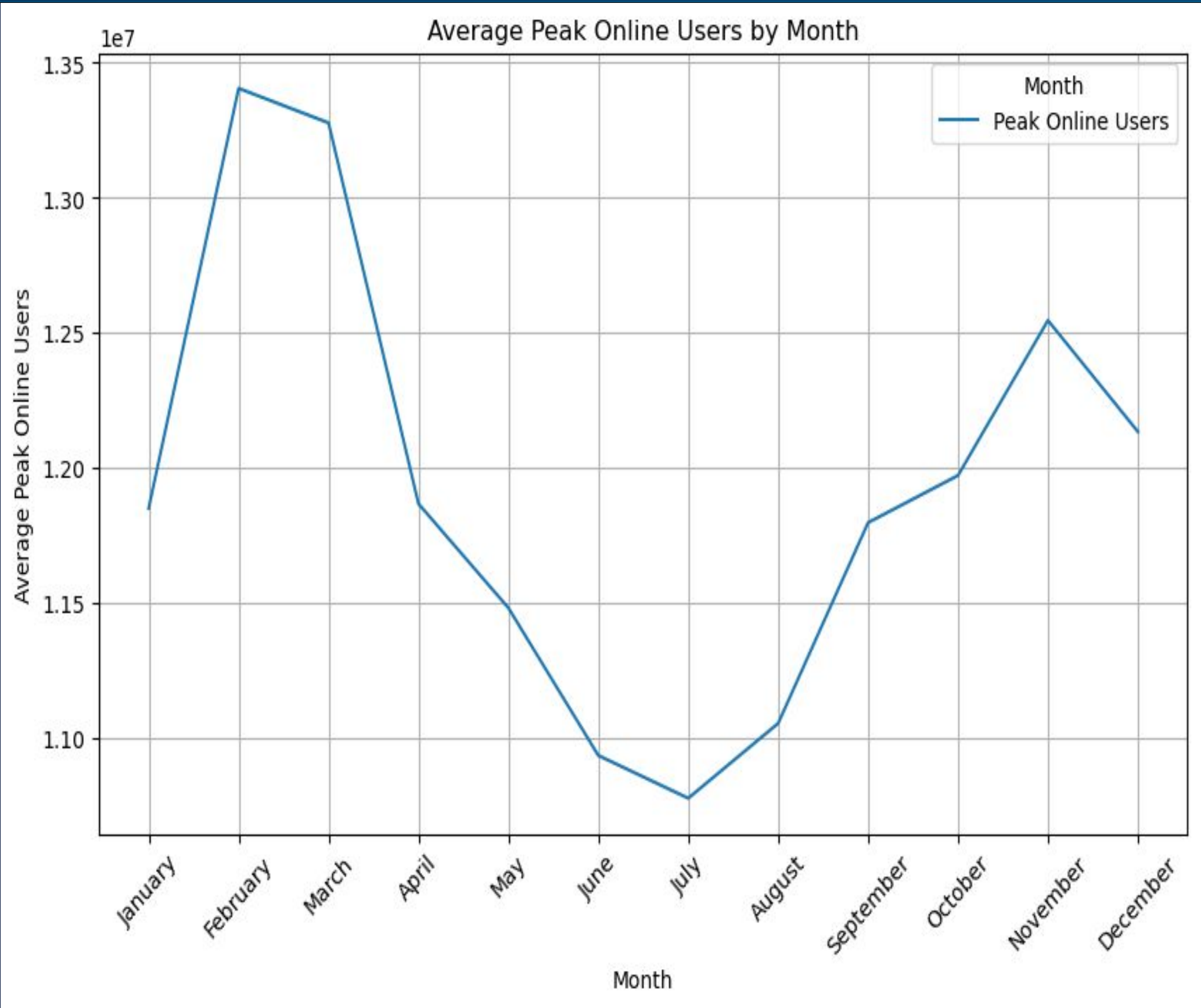
Conclusion

Despite that we rather managed to achieve our goals to make predictive models and analyze datasets, results were not as high as we expected, due to to imperfection and noisiness of gathered data. The absence of correlations in the Steam dataset can often be explained by it's size with very unpopular, low played games constituting almost all of it. The most commonly played games of Steam, even those that are probably considered as niche games by some are probably a lesser part of the Steam dataset than the tip of the iceberg is of the whole thing. The graph at the bottom left tries to show some correlation between user score and average playtime. It can be seen that even if we remove the most popular games, the overwhelming majority of it remains at the bottom with playtime being near zero with correlation also being near zero even though visually it may seem like there is some correlation. Considering all of the above it may be difficult to analyse the Steam dataset.