SYBO

05 - 23 - 2024

SUBWAY SURFER

Patterns in Speedrunning: a data mining approach

Written by

Matteo Del Prato

m.matteodelprato@gmail.com

+39 334 84 75 543

SYBO

Table of Contents

1 Introduction

In the contemporary digital landscape, mobile gaming has emerged as a predominant form of entertainment, engaging millions of users worldwide. Among the myriad of mobile games, "Subway Surfers," developed by SYBO Games, stands out as a remarkably popular title, captivating players with its dynamic gameplay and vibrant graphics. This report delves into an in-depth data analysis of "Subway Surfers," leveraging a comprehensive dataset sourced from Speedrun.com, which encompasses game-wide and level-specific category leaderboards. The objective is to uncover patterns, trends, and insights within the Subway Surfers speedrunning community, offering valuable perspectives for game developers, marketers, and the broader gaming community.

The dataset provides a rich compilation of speedrun records, detailing the fastest completion times across various categories, the prevalence of verified runs, and the diverse timing methods employed. Additionally, it includes information on the platforms used for speedruns, player demographics, and other pertinent details. By analyzing these elements, the report aims to shed light on the dynamics of the Subway Surfers speedrunning ecosystem, highlighting key contributors, popular strategies, and emerging trends.

The analysis is structured into four main sections: descriptive analysis, diagnostic analysis, prescriptive analysis, and predictive analysis. Each section addresses specific business goals, providing a comprehensive understanding of the dataset and its implications for SYBO Games.

1.1 Company's overview

SYBO's financial performance from 2020 to 2022 reflects a period of significant growth and strategic shifts.

2020

In 2020, SYBO reported a gross profit of approximately € 14.6 million, marking an improvement from the previous year, which was primarily driven by increased digital entertainment consumption during the pandemic. However, the company faced a net financial loss of about € 1.6 million, resulting in a net loss of around € 627.000 for the year. Despite this, SYBO managed to maintain a stable total asset base of approximately € 29.2 million and an equity ratio of 50.61%, reflecting a solid financial foundation.

2021

In 2021, SYBO's financial performance improved markedly. The gross profit increased to about € 23.6 million, driven by a 28% rise in topline revenue from Subway Surfers. The company achieved an operating profit of approximately € 7 million, a significant turnaround from the previous year's loss. Net financials improved to a positive €1.3 million, resulting in a net profit of around € 6.9 million. Total assets increased to approximately € 33.1 million, with an equity ratio of 60.39%, indicating enhanced financial stability and growth prospects.

2022

The year 2022 marked a transformative period for SYBO with the acquisition by Miniclip. This strategic move led to extraordinary financial outcomes. Revenue for the year stood at about €51.5 million, while gross profit was approximately €21.7 million. Despite reporting an operating loss of around €7.3 million, the net profit soared to about €387.7 million, primarily due to the financial gains from the sale of SYBO ApS and Sybo Peopleco ApS. The equity ratio surged to 416.78%, reflecting the substantial capital inflow from the sale. These financial metrics underscore SYBO's robust financial health and strategic success in navigating the competitive mobile gaming landscape.

1.2 Subway surfers

Subway Surfers, SYBO's flagship game, has demonstrated exceptional performance and sustained popularity over the years. In 2020, Subway Surfers continued to attract over 1 million daily new installs, achieving a milestone of 3 billion downloads globally. The game secured a strong active user base with approximately 100 million monthly players, highlighting its enduring appeal and engagement with users worldwide.

The success of Subway Surfers significantly bolstered SYBO's financial performance in 2021. The game was the most downloaded mobile game globally, according to Forbes, contributing to a 28% increase in topline revenue compared to 2020. Subway Surfers' consistent performance, even nearly a decade after its launch, underscores its status as a mobile gaming evergreen with significant revenue-generating potential. The game's success also facilitated SYBO's ability to attract top talent and expand its workforce to over 115 employees, representing diverse nationalities and enhancing its global reach.

In 2022, the game's performance continued to be a cornerstone of SYBO's financial success. Despite the company's strategic sale to Miniclip, Subway Surfers' robust user engagement and revenue streams remained integral to SYBO's financial stability and growth. The acquisition by Miniclip, the largest mobile gaming deal in Europe for the year, further validated Subway Surfers' market value and strategic importance in SYBO's portfolio. This strategic shift not only provided a substantial financial windfall for SYBO but also positioned Subway Surfers for continued growth and innovation under Miniclip's stewardship.

YEAR	REVENUE (€ M)
2018	92.21
2020	110.65
2022	142.93

Table 1: total revenues per year

YEAR	USERS (M)
2014	140
2019	100
2022	150

Table 2: monthly active users per year

COUNTRY	DOWNLOADS (M)
India	56
Brazil	33
United States	21
Indonesia	17
Mexico	12
Vietnam	11.7

Table 3:Downloads by country

YEAR	DOWNLOADS (BN)
2015	1
2018	2
2019	2.5
2020	3

Table 4: worlwide downloads per year

2 Dataset description

For this project, the dataset “Subway Surfers” was chosen to work with. The dataset has been published on Kaggle on the 01/12/2024 by Willian Oliveira Gibin, and offers a comprehensive compilation of speedrun leaderboards for the popular game Subway Surfers, sourced from the renowned speedrunning platform Speedrun.com and clustered according to rules.

The dataset is structured with 2.827 entries and 24 columns and encompasses both gamewide category leaderboards and level-specific category leaderboards, providing a rich resource for delving into the intricate details of the Subway Surfers speedrunning community.

The dataset's structure is outlined, highlighting the distinction between gamewide leaderboards and per-level leaderboards, the latter being further segmented into different categories for each level. The 'place' column serves as a key metric, denoting the player's rank in the respective leaderboard.

The methodology to collect data and build it has been the following;

1. Trend Analysis of Subway Surfers Speedrun Times
2. Correlation Between Platform Choice and Speedrun Performance
3. Identification of Key Contributors in the Subway Surfers Speedrunning Community

The project was done with python and can be further inspected at my GitHub account.

FEATURE	DESCRIPTION	TYPE
run_id	Unique identifier for each speedrun record	Numerical
player_id	Unique identifier for each player	Numerical
place	Rank or position of the speedrun in its respective leaderboard category	Numerical
speedrun_time	Time taken to complete the speedrun, recorded in seconds	Numerical
is_verified	Boolean indicator denoting whether the speedrun has been verified	Categorical
is_verified	Boolean indicator denoting whether the speedrun has been verified	Categorical
verify_date	Date on which the speedrun was verified	Datetime
submitted_date	Date on which the speedrun was submitted	Datetime
examiner_id	Unique identifier of the moderator who verified the speedrun	Numerical
timing_type	Method used to time the speedrun (e.g., real-time, in-game time)	Categorical
platform_id	Unique identifier for the platform on which the speedrun was performed	Numerical
platform_name	Name of the platform (e.g., Android, iOS)	Categorical

FEATURE	DESCRIPTION	TYPE
platform_released_year	Year the platform was released	Numerical
is_level_cat	Boolean indicator indicating if the speedrun is for a specific level category	Categorical
level_id	Unique identifier for the game level	Numerical
level_name	Name of the game level	Categorical
level_rules	Specific rules for the level category	Categorical
game_id	Unique identifier for the game	Numerical
category_id	Unique identifier for the speedrun category	Numerical
cat_name	Name of the speedrun category (e.g., Any%, Coin Collection)	Categorical
cat_rules	Rules for the speedrun category	Categorical
player_name	Username of the player	Categorical
player_country	Country of the player	Categorical
player_pronouns	Preferred pronouns of the player	Categorical
player_signup_date	Date the player signed up on Speedrun.com	Datetime

Table 5: dataset description

3 Business understanding

3.1 Business goals

The overall business goal of this project is to to uncover patterns, trends, and insights within the Subway Surfers speedrunning community, offering valuable perspectives for game developers, marketers, and the broader gaming community.

Key insights will be helpful for all stakeholders involved at SYBO Games to enhance the product performance (refine game mechanics, introduce features to align with player interests,improve overall player engagement...) Furthermore, marketers can leverage these insights to tailor promotional campaigns and create targeted content that resonates with the community, ultimately boosting player retention and acquisition.

Aligning game development strategies with these insights allows SYBO Games to make data-driven decisions that enhance the gaming experience. This includes refining game levels, adjusting difficulty settings, and introducing new challenges that cater to the preferences of top-performing speedrunners. Moreover, by recognizing the characteristics of top players and understanding their preferences, SYBO Games can foster a stronger community through personalized interactions and rewards.

3.2 Business questions

The following questions were used to guide the analysis. Such questions fall under Descriptive, Diagnostic, Predictive and Prescriptive analyses categories.

3.2.1 Descriptive analysis

1. What is the distribution and range of speedrun times across different categories?
2. How many players participate in each category, and what is the geographical distribution of these players?
3. What are the most popular platforms used for playing Subway Surfers in speedrun attempts?
4. What is the trend in the number of speedruns submitted per year?

3.2.2 Diagnostic analysis

1. Why do certain categories have faster or slower average completion times?
2. What are the common characteristics of top - performing speedrunners compared to mid - performers ?
3. What's the distribution of top - performers?
4. Which variables are correlated?

3.2.3 Predictive analysis

1. Which model best predicts speedrun time?

3.2.3 Prescriptive analysis

1. Platform Optimization
2. Feature Implementation
3. Data Analytics and Reporting
4. Game Design and Balancing
5. UI/UX
6. Community and Social Features
7. Retention and Acquisition Strategies

4 Data understanding: descriptive and diagnostic analysis

4.1 Project setup

Once the dataset has been uploaded on Phyton, I have been working using Jupyter Notebook and as first thing handled null values.

FEATURE	NULL VALUES	ACTION	REASON
player_id	6	Dropped	The number of null values is very low compared to the total data
level_id	2827	Dropped	Feature was displaying only null values
level_name	2827	Dropped	Feature was displaying only null values
level_rules	2827	Dropped	Feature was displaying only null values
player_name	6	Dropped	The number of null values is very low compared to the total data

FEATURE	NULL VALUES	ACTION	REASON
player_country	1077	Dropped	Dropping Nan values would consist in a 38% decrease of the dataset, therefore missing data was replaced with a substituted 'Unknown' value
player_pronouns	1996	Dropped	Dropping Nan values would consist in a 71% decrease of the dataset, therefore missing data was replaced with a substituted 'Unknown' value

Table 6: Features with corresponding NaN values, actions taken and reasons

'Platform_id' and 'platform_name' display related values, therefore only 'platform_name' feature has been kept.

'Category_id 'and 'cat_name' have the same behaviour, therefore and the first feature was dropped.

FEATURE	UNIQUE VALUES	ACTION	REASON
timing_type	1 (realtime)	Filled (Unknown)	Feature only displays a unique value
game_id	1 (subsurf)	Filled (Unknown)	Feature only displays a unique value

Table 7: Features with corresponding unique values, actions taken and reasons

4.2 Outliers analysis

After pre-processing the data, I analysed the outliers in order to decide which actions to take with them. I considered all features, and took tailored made decisions to aim for a better overall result . To identify those outliers, I used box plot diagrams so that I could get a visual summary of the variables' distribution and easier detect those values that fall far outside the expected range.

‘place’

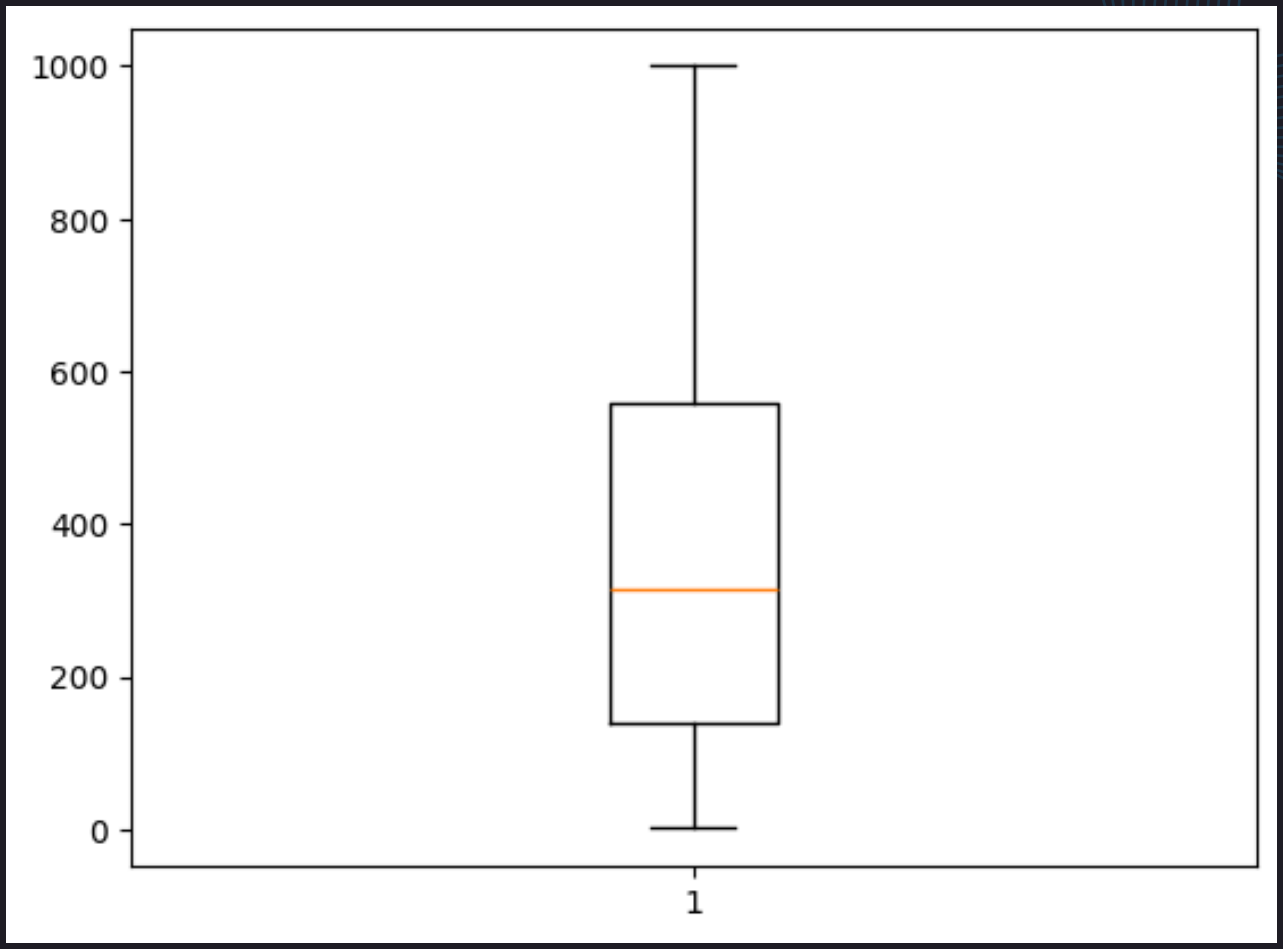


Figure 1: Box plot diagram showing ‘place’ feature

The interquartile range (IQR) of 500 indicates moderate variability in the rankings, with the middle 50% of players having ranks between approximately 250 and 750.

Since higher rank positions are represented by lower values, the lower quartile (Q1 ≈ 250) suggests that the top 25% of players have ranks of 250 or better. This group represents the most skilled or dedicated players in the speedrunning community. The upper quartile (Q3 ≈ 750) indicates that the bottom 25% of players have ranks of 750 or worse, which includes those who may be less experienced or less competitive in speedrunning.

The lack of outliers in the box plot suggests consistency in the rankings data. There are no extreme ranks that deviate significantly from the overall distribution, indicating a fairly uniform competition level among players.

QUARTILE	%	RANKINGS
Top Quartile	25	Players ranked 1 to 250
Second Quartile	25 - 50	Players ranked 251 to 500
Third Quartile	50 - 75	Players ranked 501 to 750
Fourth Quartile	75 - 100	Players ranked 751 to 1000

Table 8: players segmentation based on 'place' feature

'speedrun_time'

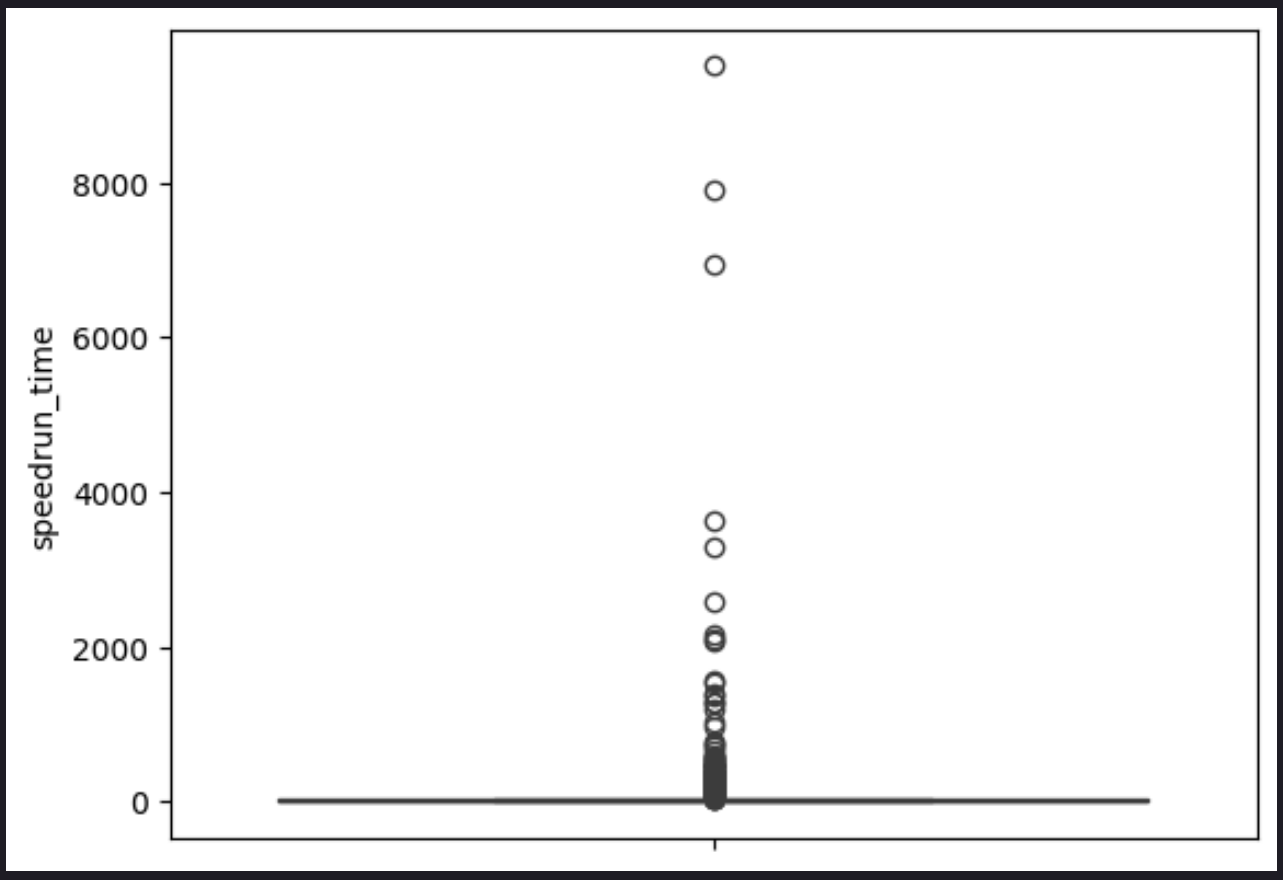


Figure 2: Box plot diagram showing 'speedrun_time' feature

The distribution shows a significant number of extreme outliers extending up to over 9000 seconds (approximately 150 minutes), while the central box (IQR) is very compressed near the bottom, indicating that the majority of 'speedrun_time' values are much lower than the extreme outliers.

The presence of very high outliers suggests that some players have exceptionally long performances, which could be due to various factors such as extended play sessions, different game modes, or incorrect data entries. The compressed IQR, on the pther hand, highlights that most players have relatively shorter performances, with only a few extreme cases pulling the upper range significantly higher.

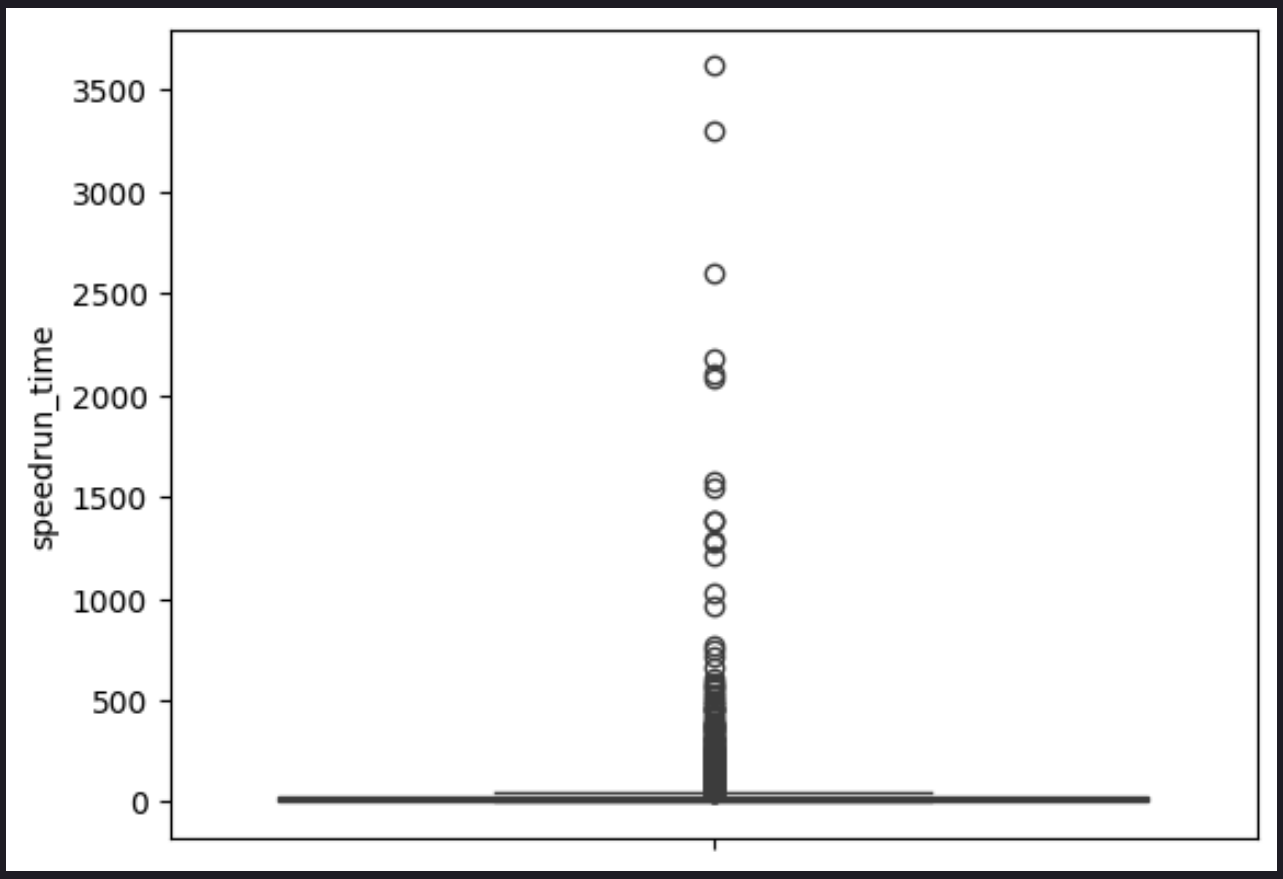


Figure 3: Box plot diagram showing 'speedrun_time' feature after outliers removal

After removing extreme outliers, the maximum 'speedrun_time' has been reduced to around 3500 seconds (approximately 58 minutes).

While the IQR and median remain relatively unchanged, the spread of the data is more visible, showing a clearer distribution of 'speedrun_time' values.

The presence of outliers up to 3500 seconds suggests that while the majority of players have shorter 'speedrun_time' values, there are still several significant outliers that indicate varied gameplay durations or strategies.

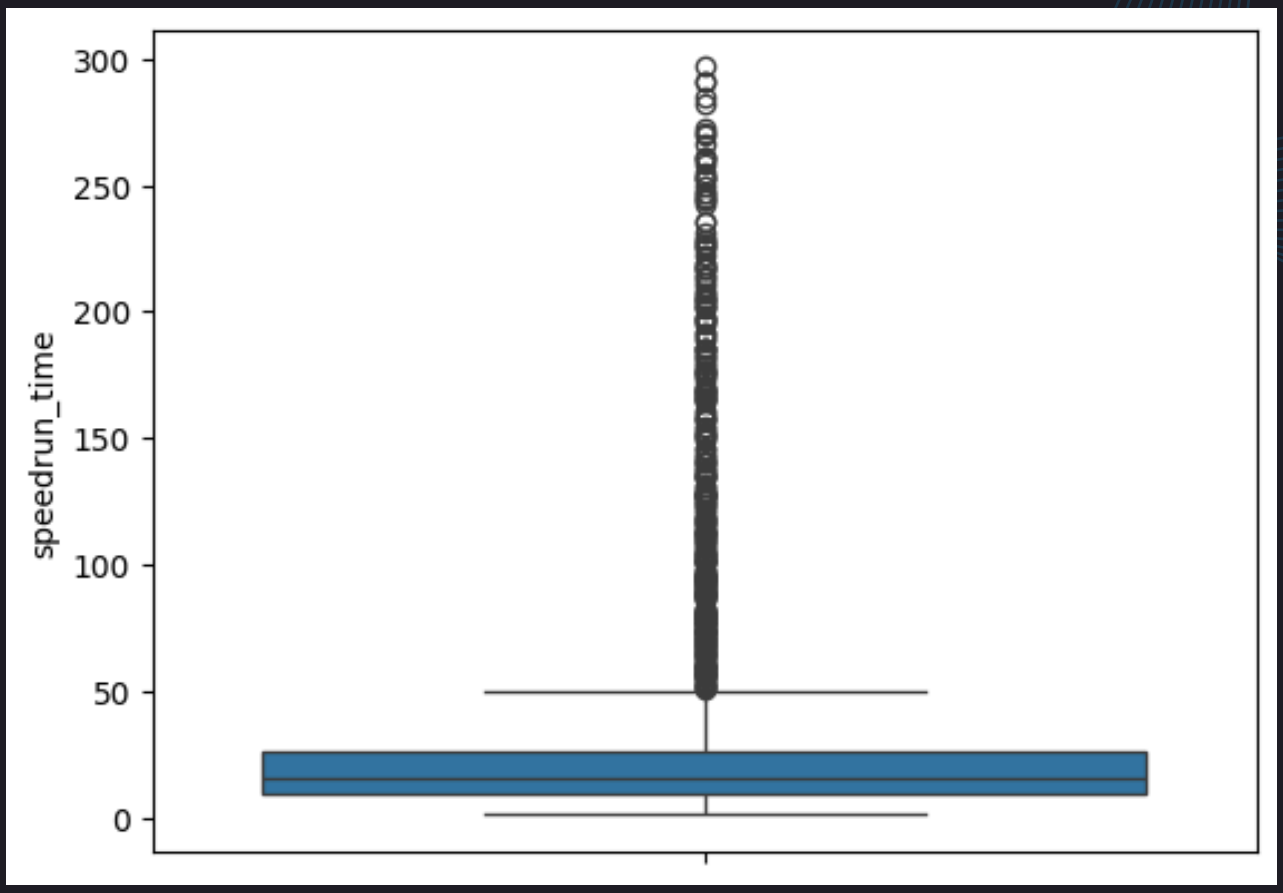


Figure 4: Box plot diagram showing 'speedrun_time' feature after final outliers removal

After the final removal of higher outliers, the final plot shows a maximum 'speedrun_time' of 300 seconds (5 minutes). The box plot now has a more balanced appearance, with a well-defined IQR and whiskers extending to more moderate values.

This cleaned dataset provides a clearer view of the 'speedrun_time' distribution without the influence of extreme outliers. The maximum 'speedrun_time' of 300 seconds indicates that most players complete their runs within a 5-minute timeframe, making it easier to focus on the core player base and understand general trends in speedrun performance.

4.3 Answering descriptive analysis questions

After cleaning the datset from null values and outliers, I dug deeper into the analysis by taking on from the descriptive stage.

4.3.1 What is the distribution and range of speedrun times across different categories?

First of all, I computed a descriptive statistic for 'speedrun_time' grouped by 'cat_name' to better understand the relationship between the two features along with each category.

- **COUNT** represents the number of observations (speedrun times) recorded for each category
- **MEAN** represents the average 'speedrun_time' for each category (sum of all speedrun times in a category divided by observations count)
- **STD** shows 'speedrun_time' standard deviation (amount of variation / dispersion from the mean) of for each category
- **MIN** indicates the shortest recorded speedrun time for each category
- **25 %** represents the first quartile of 'speedrun_time' for each category, which is the median of the lower half of the data (indicating that 25% of the speedrun times are below this value)
- **50 %** represents the second quartile (Q2) or the median speedrun time for each category
- **75 %** represents the third quartile of 'speedrun_time' for each category, which is the median of the upper half of the data
- **MAX** shows the longest recorded speedrun time for each category

CAT_NAME	COUNT	MEAN	STD	MIN	25 %	50 %	75 %	MAX
All Powerups	64.0	99.2	60.7	22.8	58.3	77.6	142.6	270.7
Coins	1001.0	21.7	4.8	13.6	17.1	21.4	26.1	30.3
Mystery Hurdles	36.0	203.7	52.8	88.0	156.3	220.0	246.1	291.2
Mystery box	34.0	18.4	11.6	1.6	10.8	19.2	21.8	61.7
Obtain Item	476.0	28.8	54.1	3.5	4.2	5.5	15.4	282.4
Score (x38-)	792.0	32.5	44.7	9.0	10.2	12.5	26.8	296.8
Score (x39+)	356.0	14.9	22.7	6.6	7.7	8.8	9.6	197.2

Table 9: descriptive statistic for 'speedrun_time' grouped by 'cat_name'

The 'Mystery Hurdles' category has the highest mean and median ‘speedrun time’, indicating it generally takes longer to complete runs in this category: on the other hand, the 'Coins' category has one of the shortest mean and median times. While the 'Obtain Item' category shows high variability in speedrun times (std) - suggesting different strategies or difficulties in completing this category - the 'Coins' category has a lower std value, indicating more consistency in ‘speedrun time’.

Overall ‘speedrun_time’ distribution

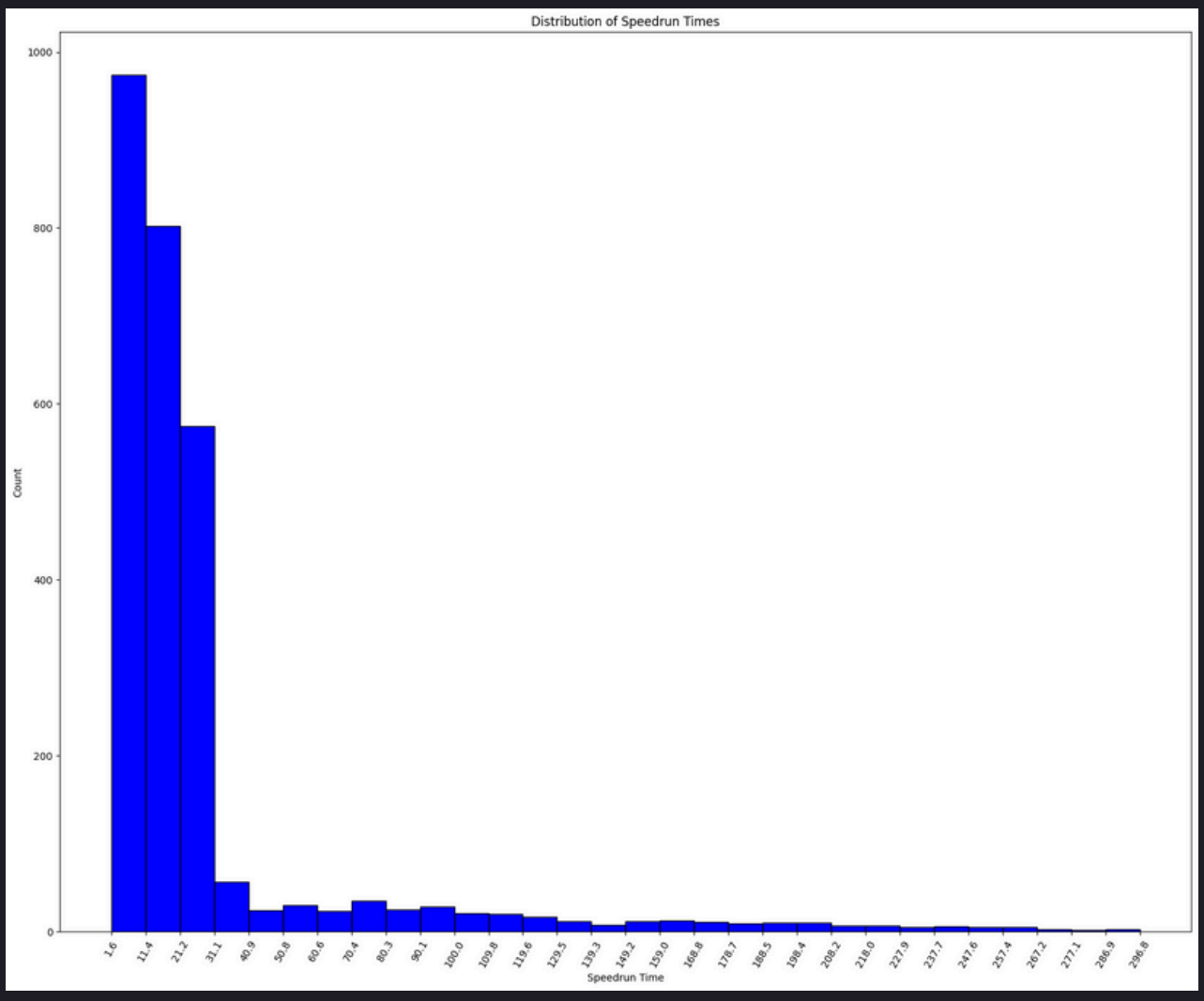


Figure 5: Distribution of ‘speedrun time’

The overall distribution of ‘speedrun time’ is heavily right-skewed, with a significant majority of runs concentrated at the lower end of the speedrun time spectrum, indicating that most speedruns are completed quickly, with only a few taking a longer time.

The highest bar is at the 1.6-second mark, along with those at 11.4 and 21.2 seconds, suggesting these are common completion times and that's the primary cluster where most speedruns fall.

As speedrun times increase, the frequency of runs decreases sharply: there are very few runs with speedrun times exceeding 60 seconds, and significant low counts beyond 180 seconds.

Myster Hurdles

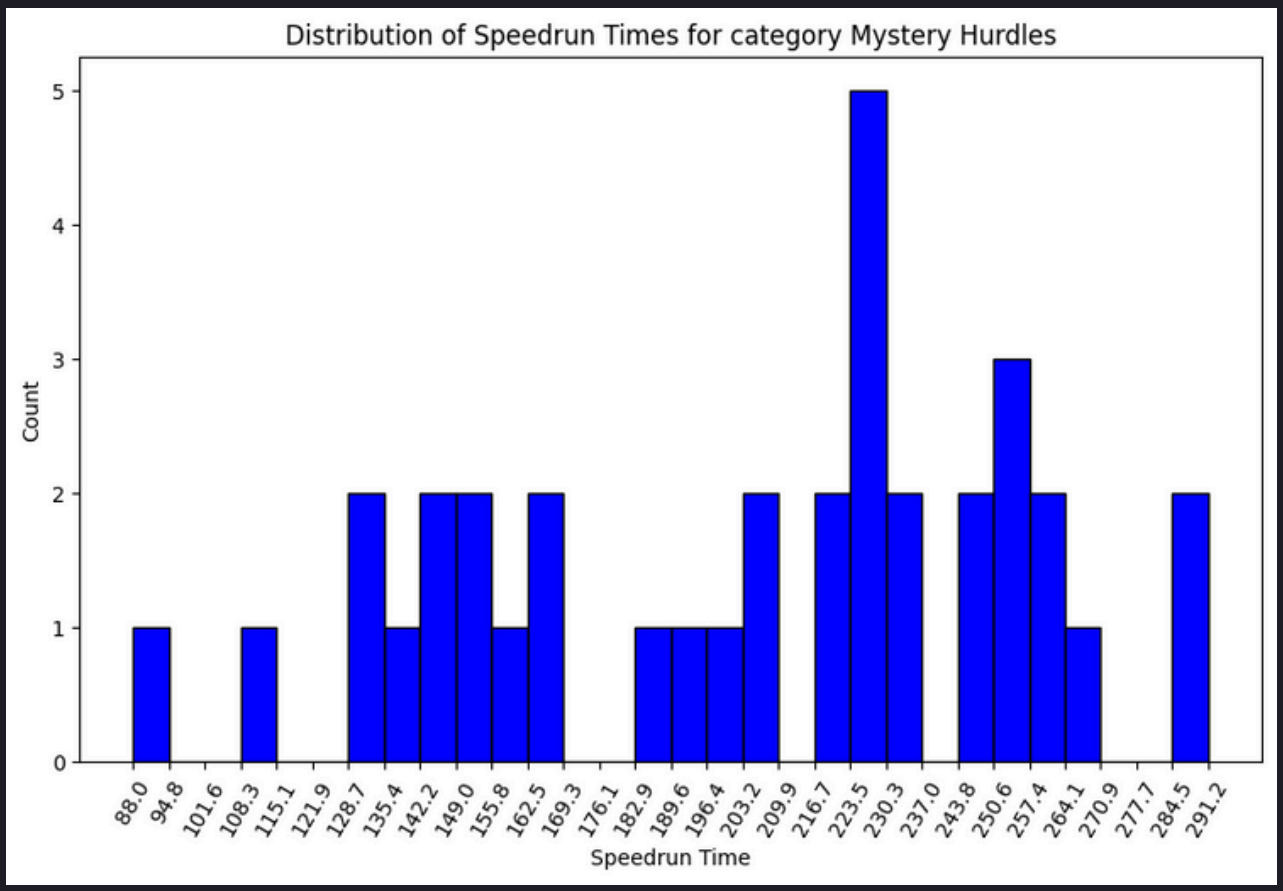


Figure 6: Distribution of 'speedrun times' for 'Mystery Hurdles' category

If we consider the distribution of 'speedrun times' for 'Mystery Hurdles' category, we can see a clear peak at around 220 seconds (3.6 min) indicating a popular completion time. Overall the graph shows no clear central tendency, expressing a varied range of speedrun times. Mystery Hurdles is, overall, one of the two categories which shows more spread and higher times, indicating more complex or varied tasks.

Coins

If we consider the distribution of 'speedrun times' for 'Coins' category, we can clearly see how the graph shows a significant peak in the range of 15 seconds, indicating a

common completion time for many speedrunners. A broad spread of speedrun times is displayed, but the majority are clustered between 14 and 29 seconds, suggesting that most players completed the coin collection within this time frame.

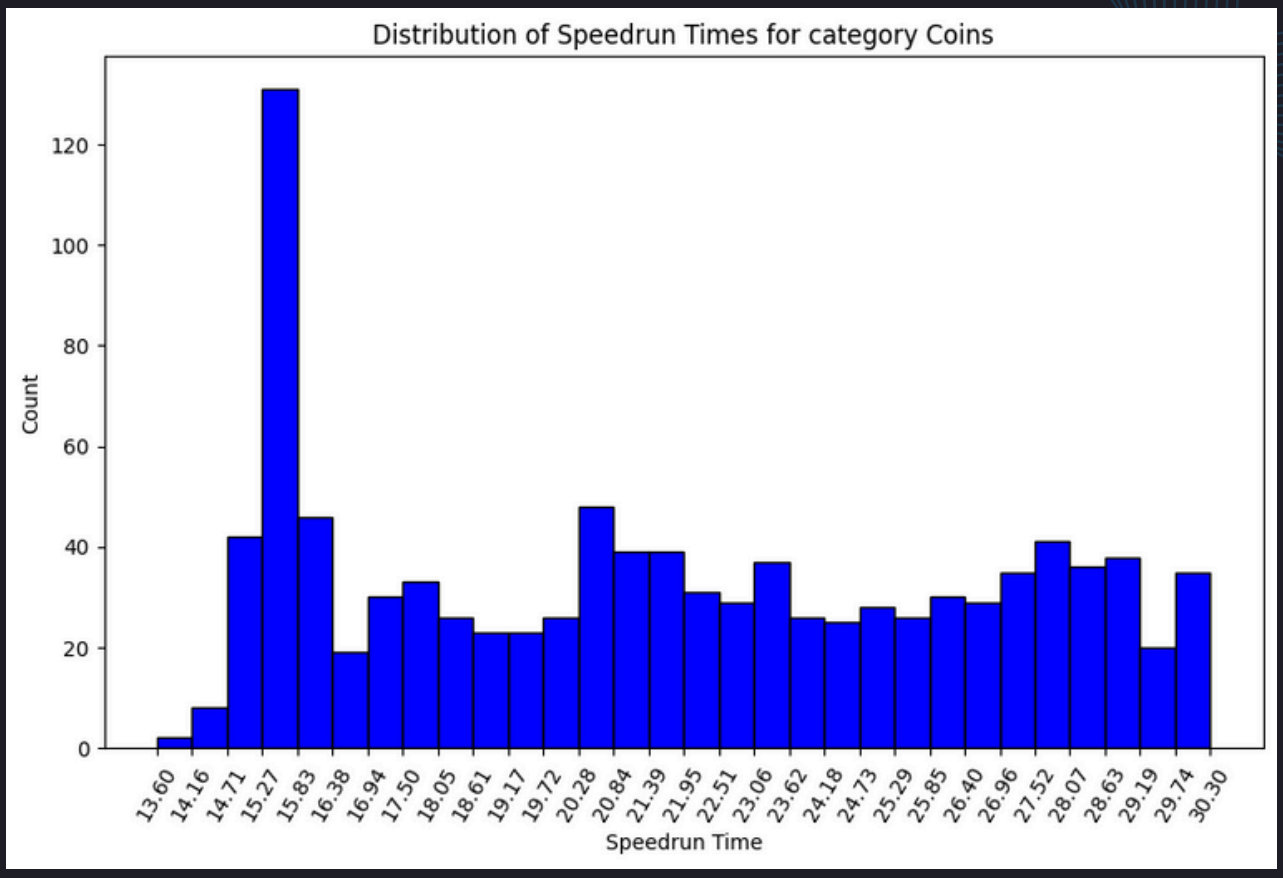
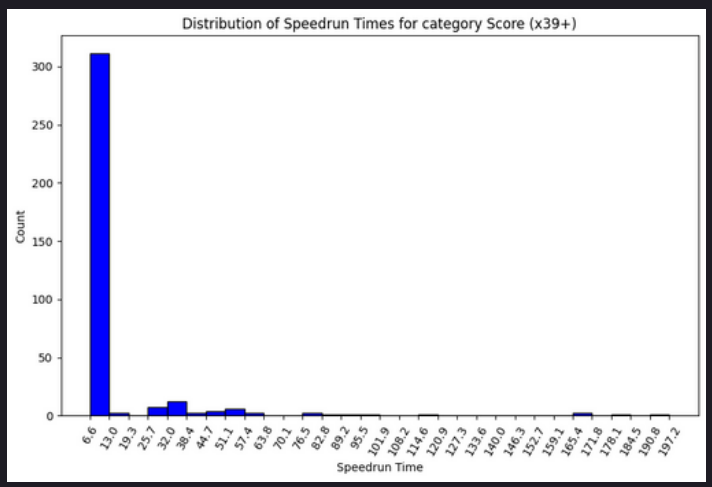


Figure 7: Distribution of 'speedrun times' for 'Coins' category

'Coin' is one of the two categories which have quicker and most concentrated completion times, indicating straightforward tasks.

Score (x39+), (x38-)



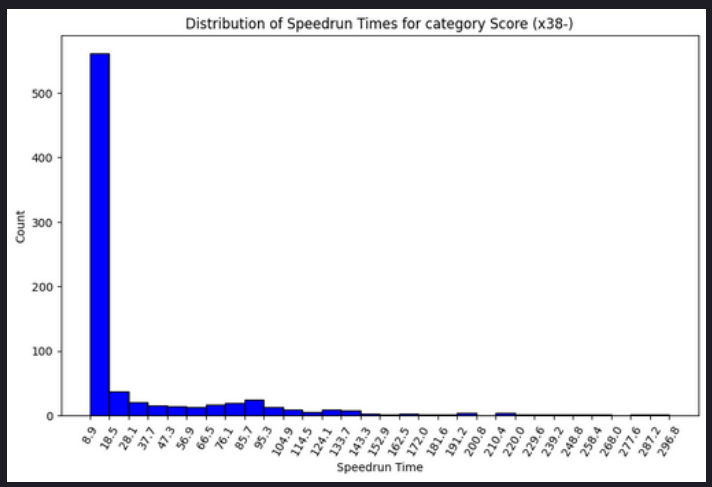


Figure 8,9 Distribution of 'speedrun times' for 'Scores' categories (x39+,x38-)

For lower scores (x38-), the distribution is heavily skewed towards lower times with a significant peak at around 8-9 seconds, showing how the majority of players can quickly achieve the score goal within this time although there's a small minority which achieves longer completition time.

Considering higher scores (x39+), the graph also shows a heavy skew towards lower times, with a peak around 6-7 seconds indicating an even faster completion time for many players. The distribution quickly drops off, showing that achieving the score is very quick for most.

In general, score categories (x38- and x39+) are highly skewed towards very fast completion times, suggesting high proficiency or ease in achieving the goals.

All Powerups

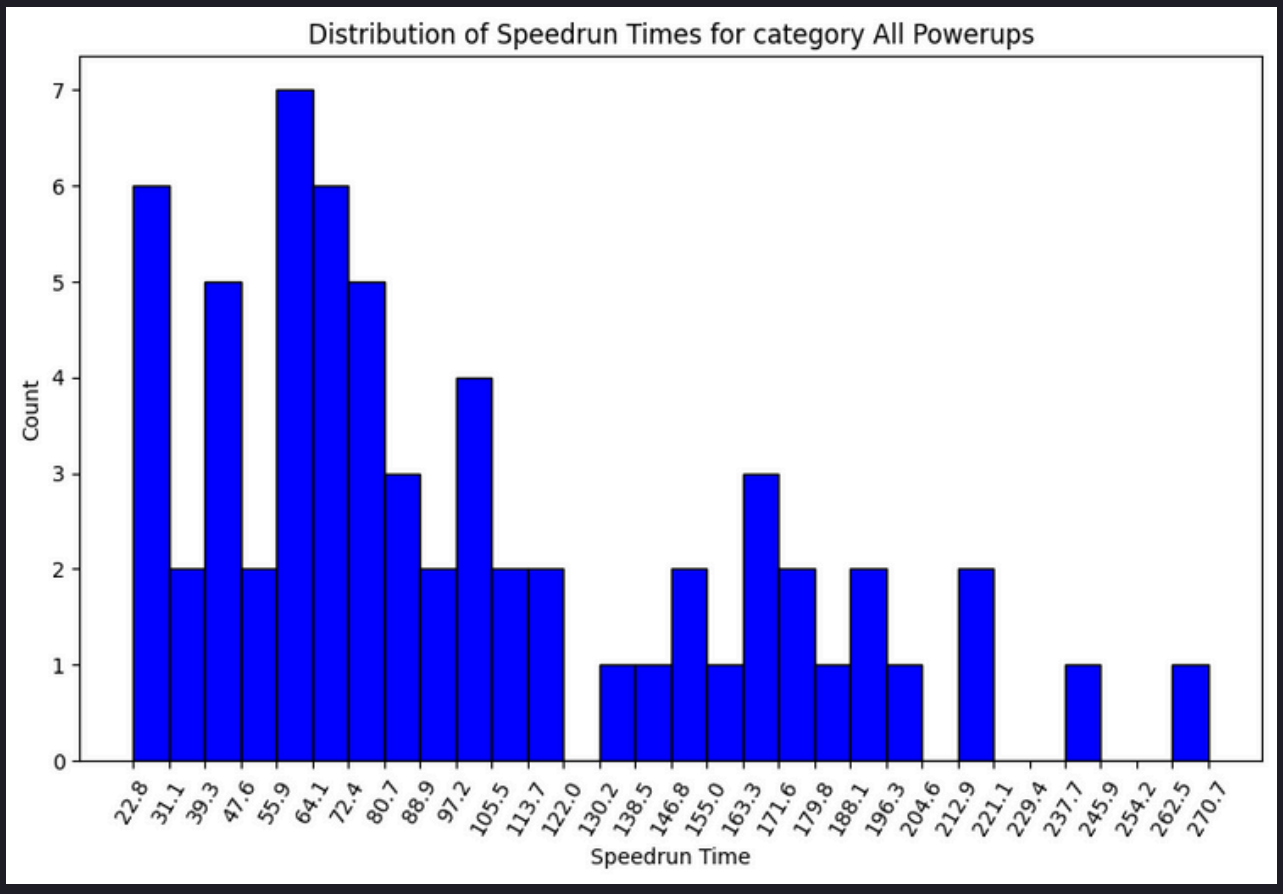


Figure 10 Distribution of 'speedrun times' for 'All Powerups' category

The distribution is more spread out with a higher variance compared to the score categories. Peaks around 55-73 seconds indicate a common completion time range. The spread suggests that collecting all power-ups takes a variable amount of time for different players.

All Powerups, along with Mystery Hurdles, show more spread and higher times indicating more complex or varied tasks

Obtain item

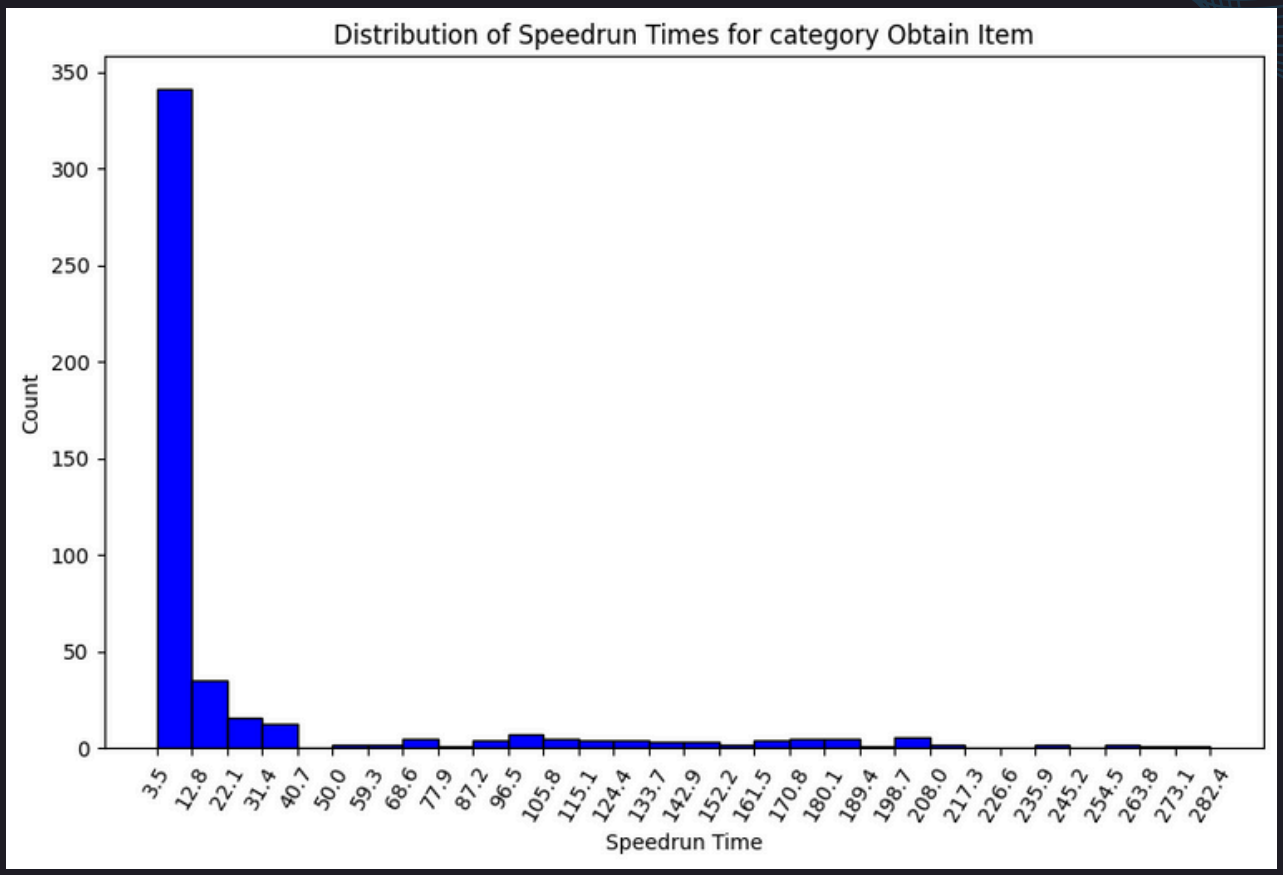


Figure 11: Distribution of 'speedrun times' for 'Obtain Item' category

The graph shows a significant peak at around 3-12 seconds, indicating a very quick task for most players. Only few players take longer, indicating the task's ease and quick completion nature.

Obtain Item, along with Coins, has faster and most condensed completion time, hinting at straightforward tasks.

Mystery box

We can see how the distribution has a notable peak around 15-16 seconds, with a wider spread compared to some other categories, having time ranging from 1.5 to over 30 seconds. The peak suggests that most players complete the task within this range, but with significant variation. Along with All Powerups, this category show more spread and higher times, indicating more complex or varied tasks.

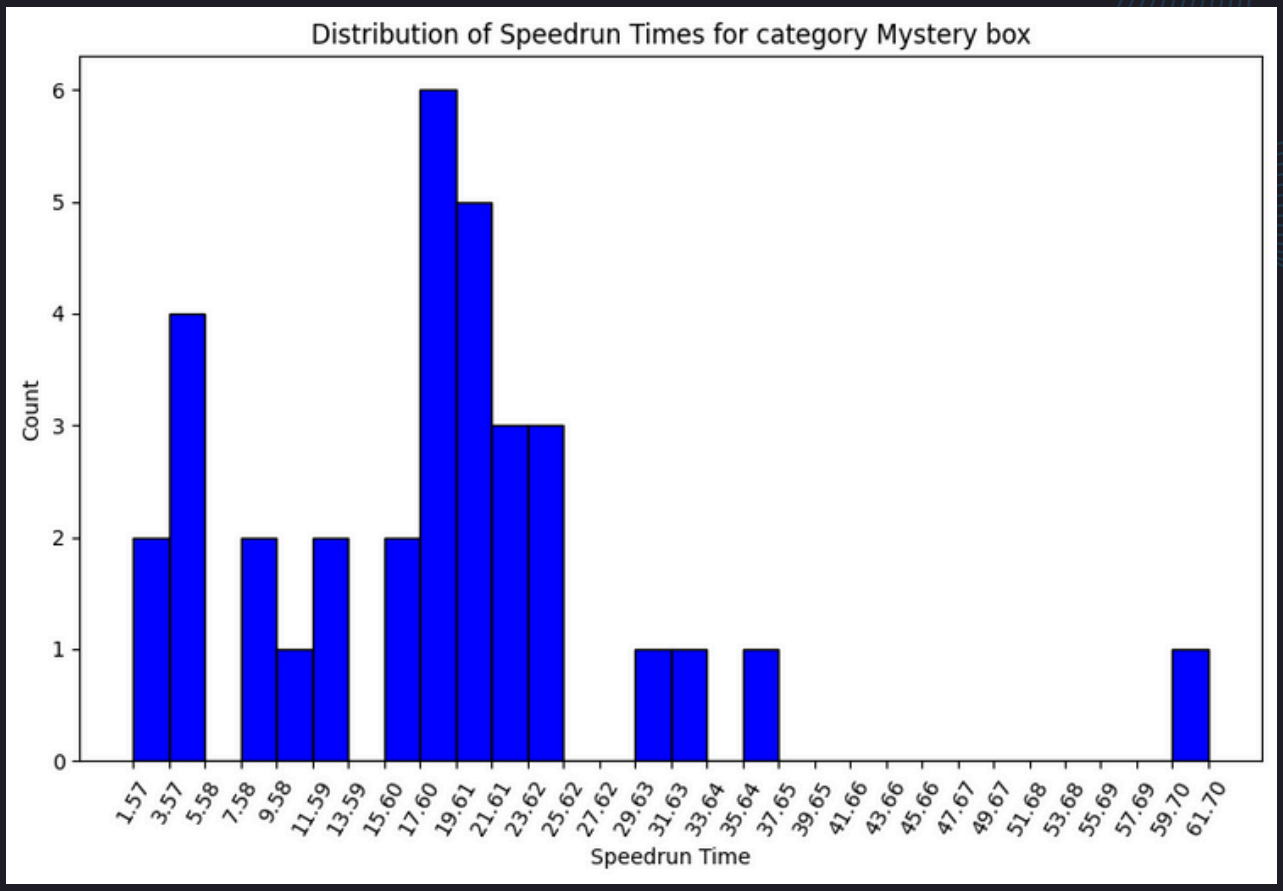


Figure 12: Distribution of 'speedrun times' for 'Mystery Box' category

4.3.2 How many players participate in each category, and what is the geographical distribution of these players?

First of all, I analysed players' presence in each category.

The graph shows how the 'Coins' category is the most popular, with approximately 1000 players participating making collecting coins as quickly as possible one of the most common goal among speedrunners.

Score categories 'Score (x38-)' and 'Score (x39+)' also show significant player participation, with around 800 and 500 players, respectively.

The 'Obtain Item' category has a moderate level of participation, with approximately 400 players. On the other hand, 'All Powerups','Mystery Hurdles' and 'Mystery Box' categories have the least participation, with numbers significantly lower than the other categories.

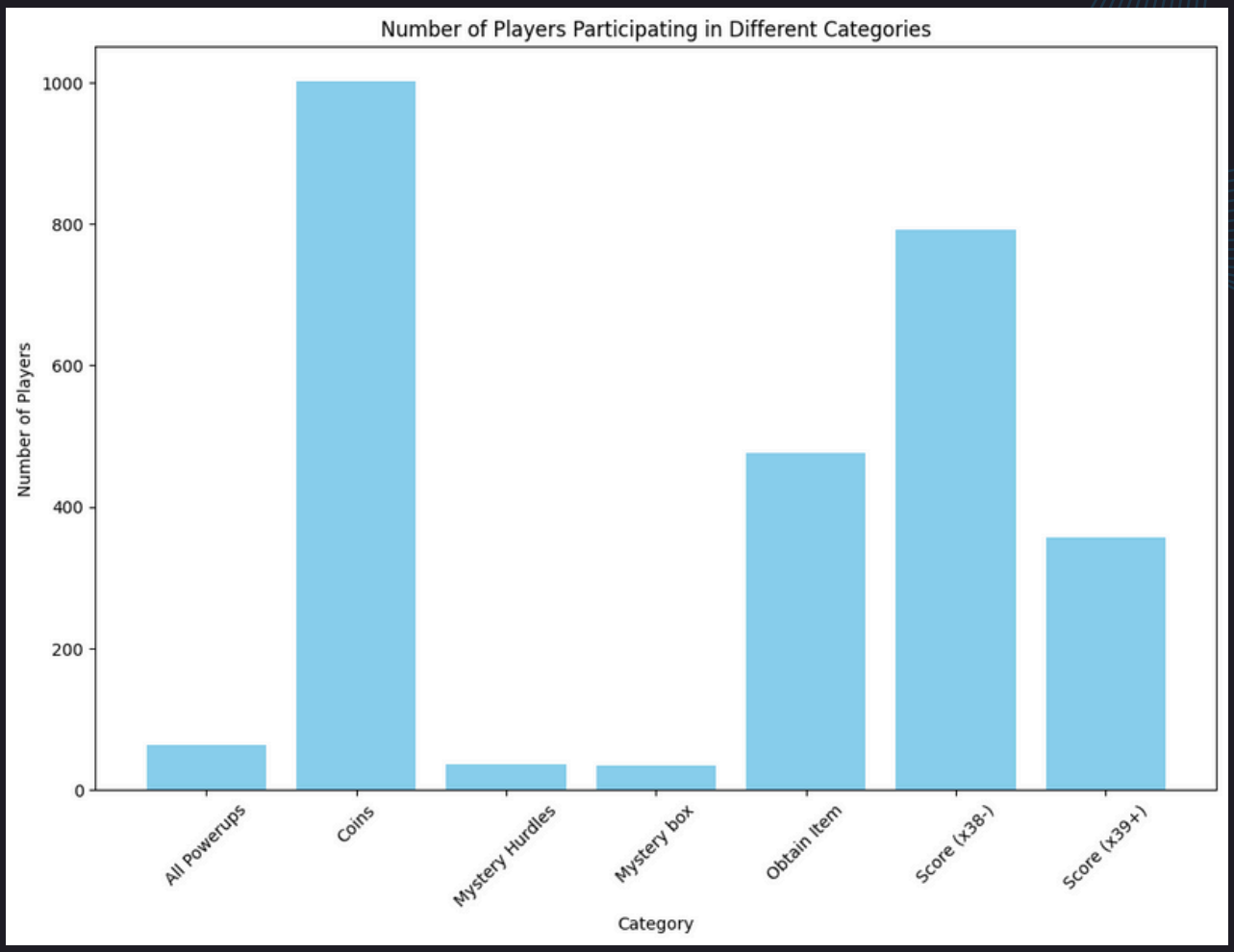


Figure 13: Number of players participating in different categories

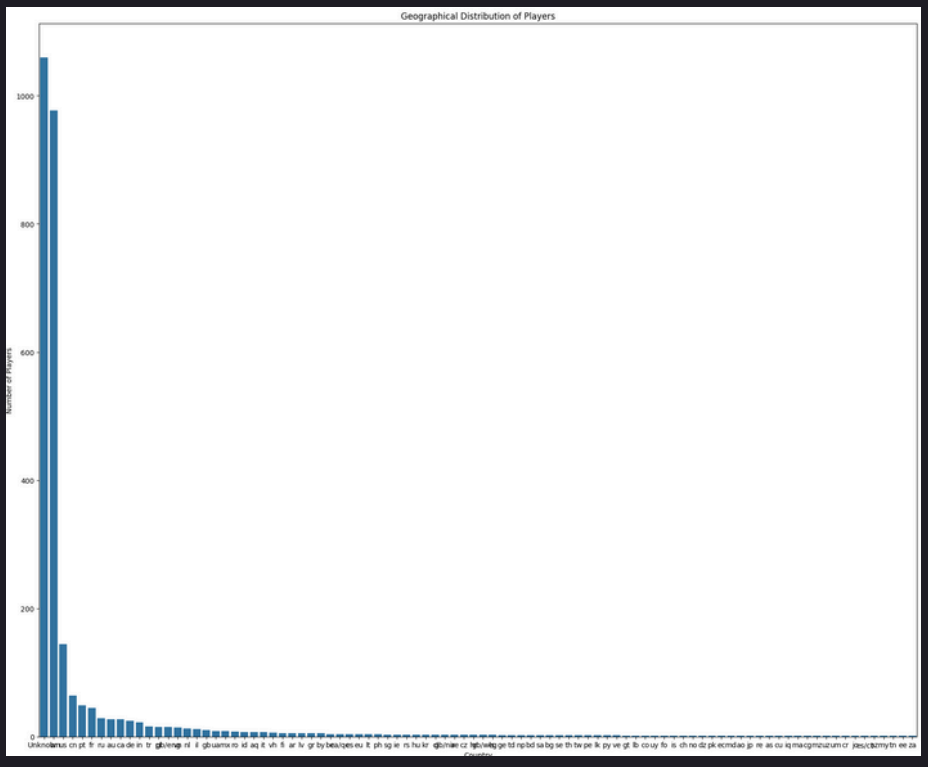


Figure 14: Geographical distribution of player (including 'Unknown' values)

The second step consisted in inspecting the geographical distribution of the players. I started by including 'Unknown' countries, but ended with dropping the value to aim for a clearer insight.

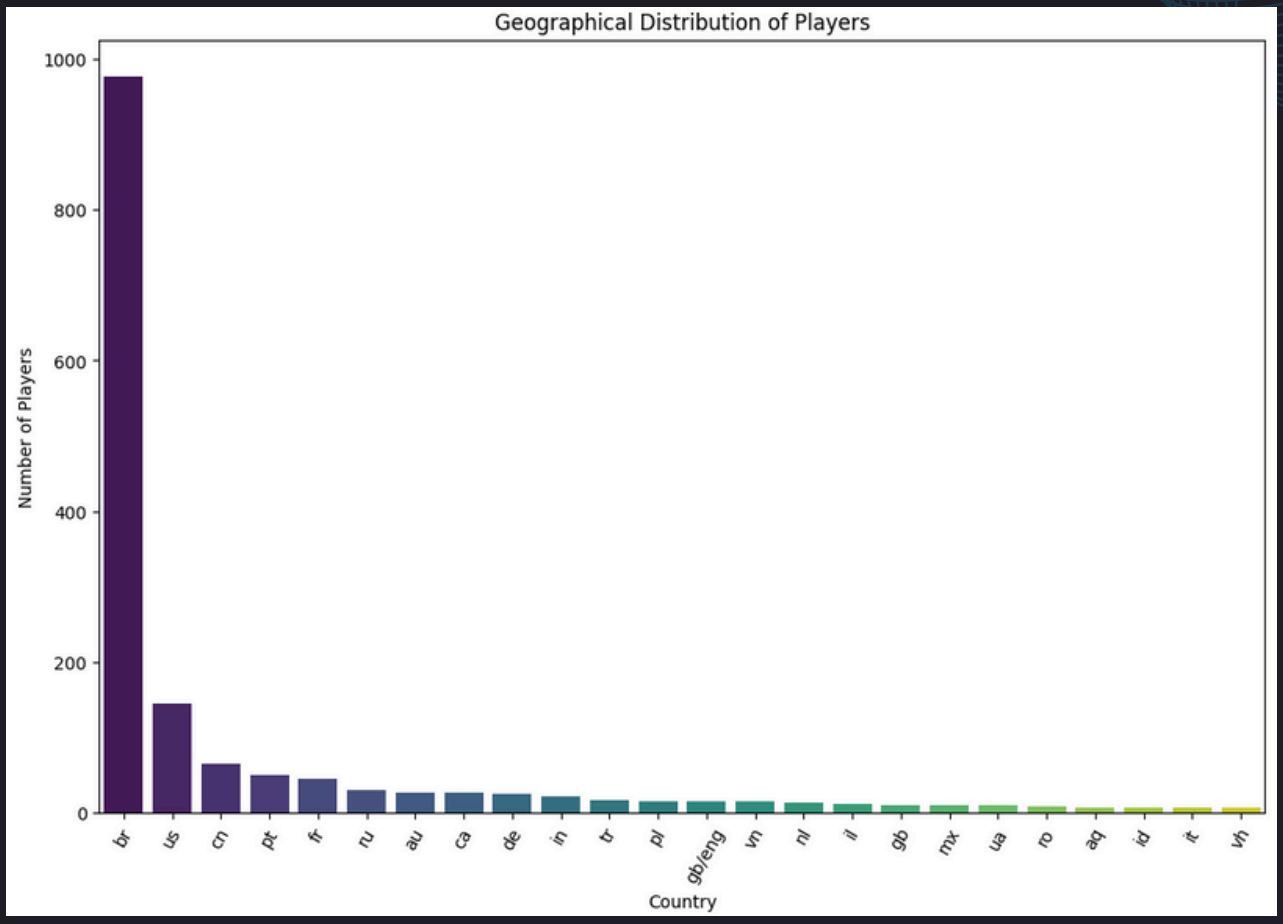


Figure 15: Geographical distribution of player (excluding 'Unknown' values)

The majority of players (close to 1000) are brasilian, followed by american ones (although with a significant lower presence); this is mainly due to the nationality of the dataset.

Other countries with visible - but less lower - participation include China, Portugal, France, Russia, Australia, Canada, Germany and India.

Numerous other countries, including Turkey, Germany , Australia, Mexico and others, show lower participation rates.

4.3.3 What are the most popular platforms used for playing Subway Surfers in speedrun attempts?

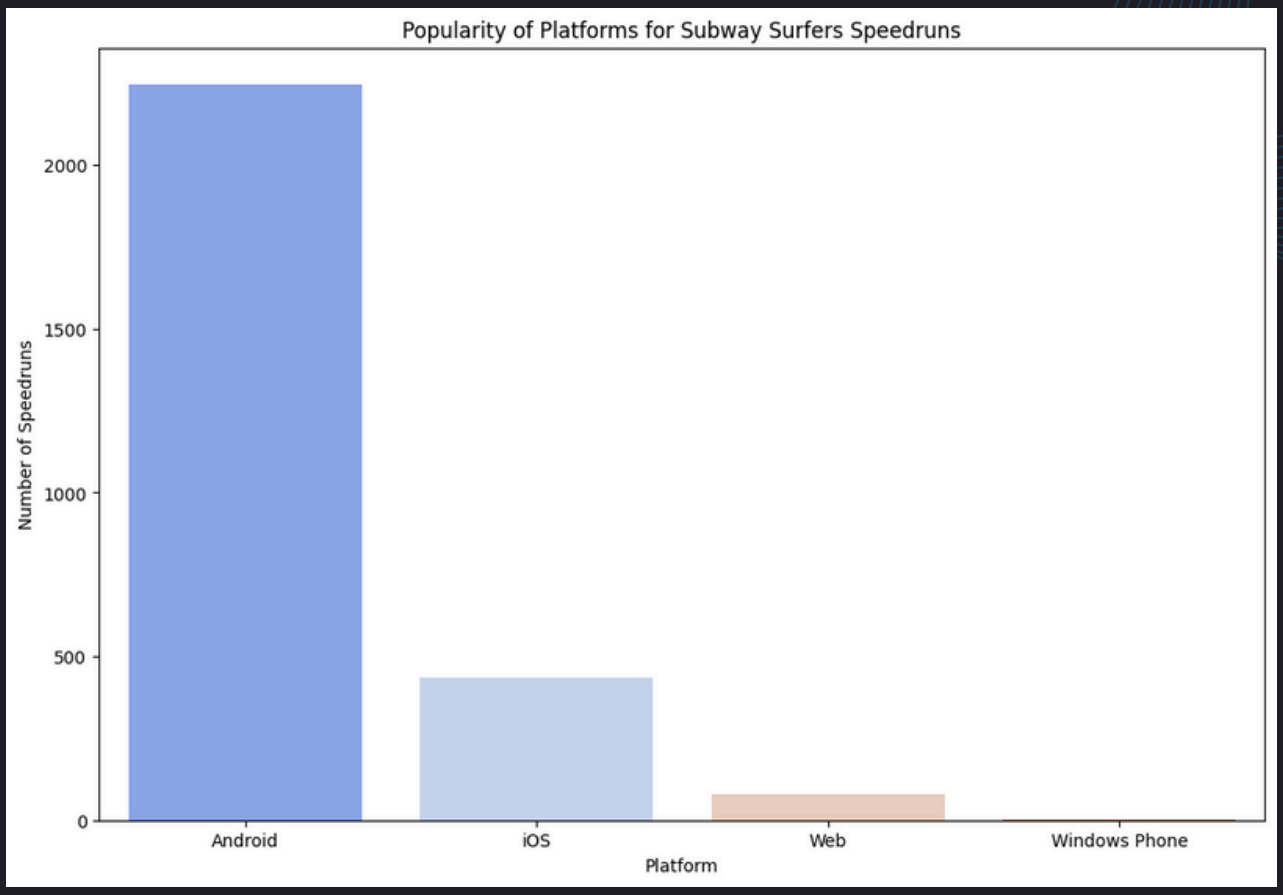


Figure 16: Popularity of platforms for speedruns

The bar chart illustrates the popularity of different platforms for Subway Surfers speedrun attempts among Android, iOS, Web, and Windows Phone.

Android is the most popular platform by a significant margin, with over 2000 speedrun submissions. iOS follows with a much lower count, slightly above 500. The Web platform has minimal submissions, highlighting its lack of popularity for speedruns. Windows Phone shows an almost negligible number of submissions, indicating it is the least used platform for this purpose.

The fact that Android platform outscores iOS one is mainly due to the fact that, according to Statista, nearly 82 % of mobile devices in Brazil (major nationality in the dataset) operate on Android, whereas over 18 % operate on Apple. Consistent with global trends, hardly any other operating systems are used for mobile devices in Brazil.

The low numbers for Web and Windows Phone suggest these platforms are not favored for optimal speedrun performance, possibly due to performance or user base limitations.

4.3.4 What is the trend in the number of speedruns submitted per year?

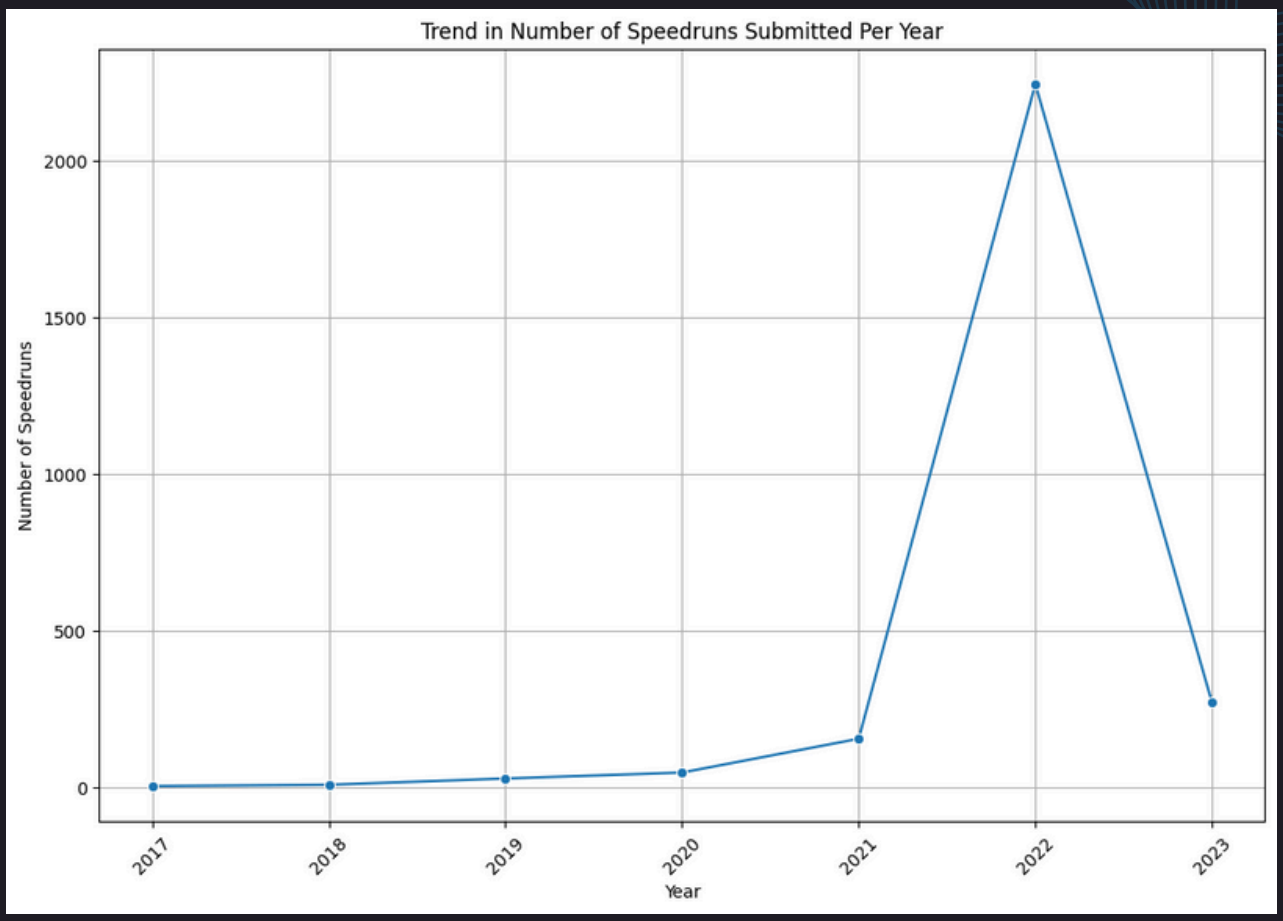


Figure 17: Trend in speedruns submitted per year

The line chart illustrates the trend in the number of speedruns submitted per year from 2017 to 2023.

There is a gradual increase in submissions from 2017 to 2020, evolving with a significant rise in 2021 and peaking in 2022 with over 2000 submissions, indicating a surge in interest or activity in that year. A sharp decline in 2023 suggests either a decrease in interest or external factors affecting submissions. The peak in 2022 may correspond to specific events, updates, or marketing strategies that spiked player engagement.

The overall trend shows growing popularity until 2022, followed by a notable drop in the subsequent year.

4.4 Answering diagnostic analysis questions

Once finished the descriptive analysis, I took on the diagnostic one to deeper analyse data, eventually creating clusters and new variables.

4.4.1 Why do certain categories have faster or slower average completion times?

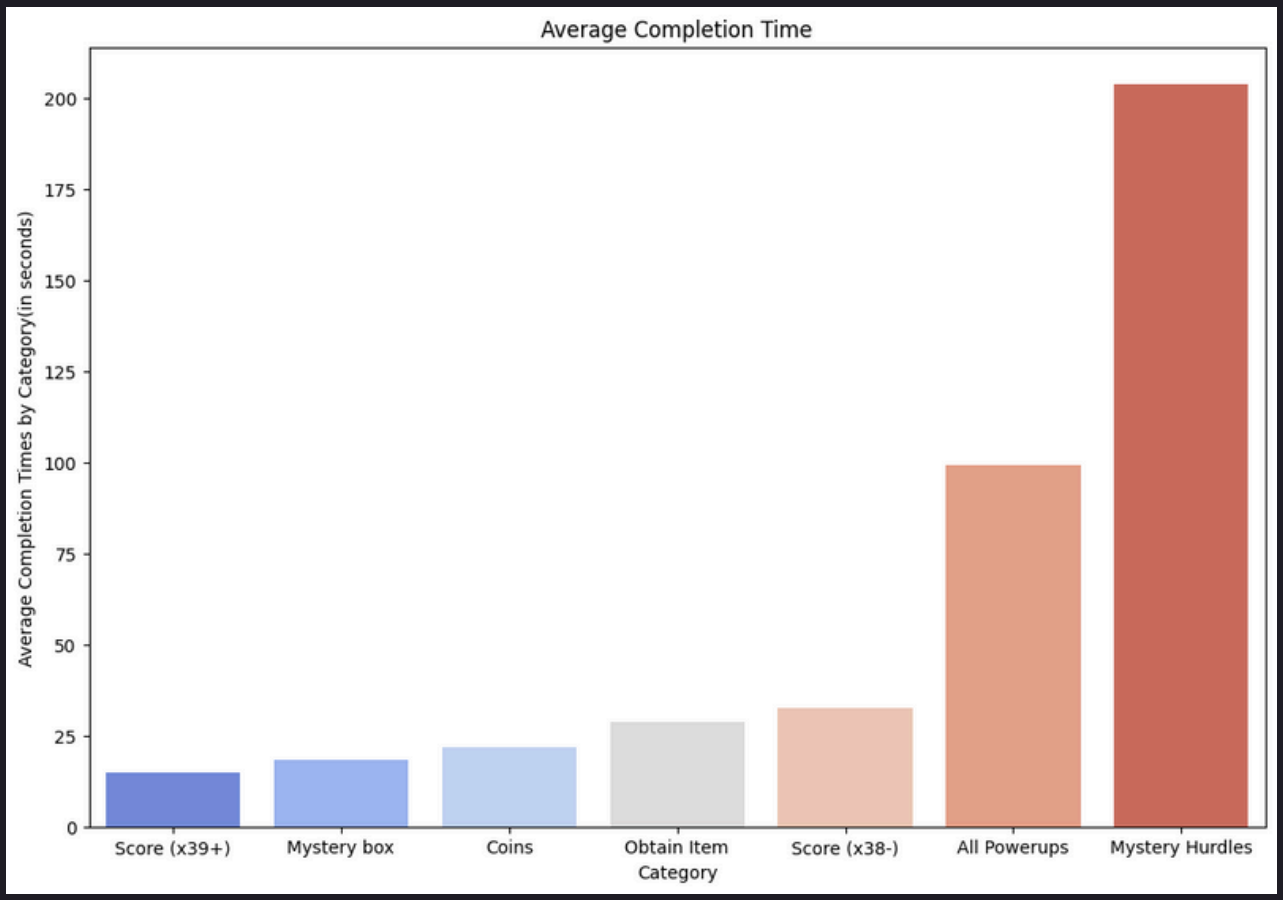


Figure 18: Average completition time per category

The bar chart displays the average completion times for different categories in Subway Surfers speedruns.

"Mystery Hurdles" has the highest average completion time, around 200 seconds, indicating it is the most time-consuming category. Looking at the community's feedback on Reddit and YouTube can be helpful to get a sense of the level of difficulty players face.

"All Powerups" also has a relatively high completion time. Categories like "Score (x38-)" and "Obtain Item" have moderate average times, while "Coins," "Mystery Box," and

"Score (x39+)" have the lowest average times, around 25 seconds or less.

4.4.2 What are the common characteristics of top - performing speedrunners compared to mid - performers ?

To be able to proceed with the diagnostic analysis, i had to create two new variables identifiying top perfomers (10% highest speedrun time values) and mid performers (speedrun time values in the range of the highest 10 % - 90 %)

Country distribution

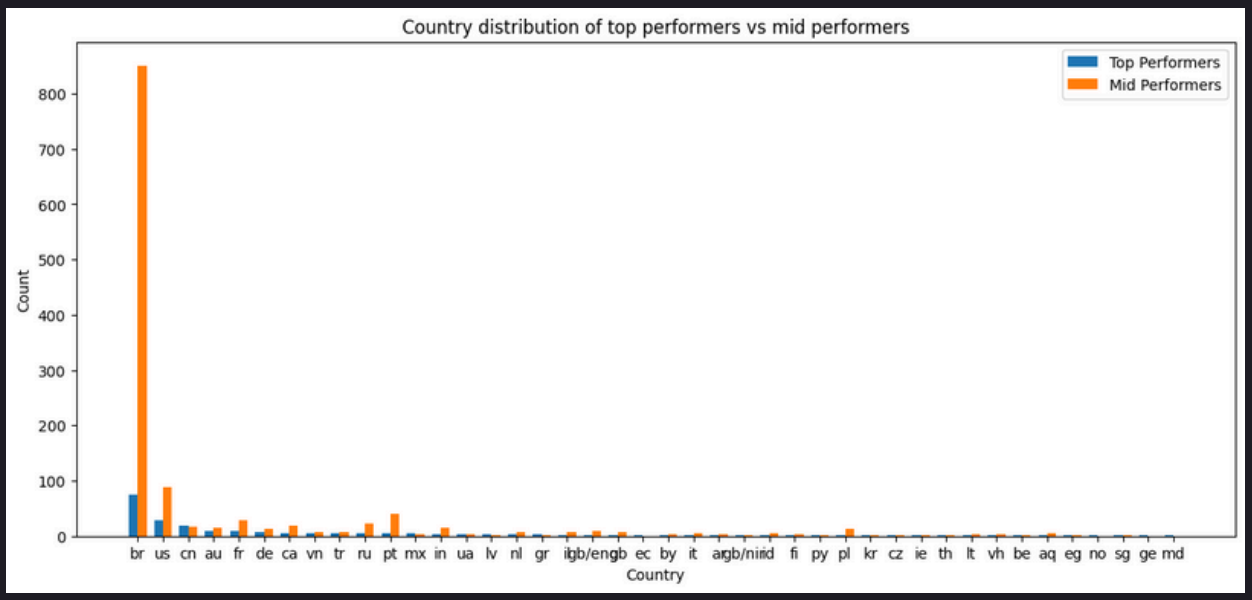


Figure 19: Country distribution of top performers vs mid perrformers

The first graph shows the country distribution of top and mid performer. While Brasil, USA and China host the most player (as seen in chapter 4.3.2), China is the only country in which the number of top performers outscores the number of mid perfomers.

Other countries to host a significant number of mid performance players are France, Canada, Russia and Portugal. Apart from Brasil, USA and China, a notable number of top perfomers also comes from Austria, France, Canada and Germany.

Preferred categories

The following graph compares metric values by performer type across different categories. Top performers performer way better in the "Obtain Item" category, having 50 % higher metric values if compared to mid performers. This could mean

they either play mostly in this category, or achieve higher scores.

On the other hands, metric values are inverted in the "Mystery box" category.

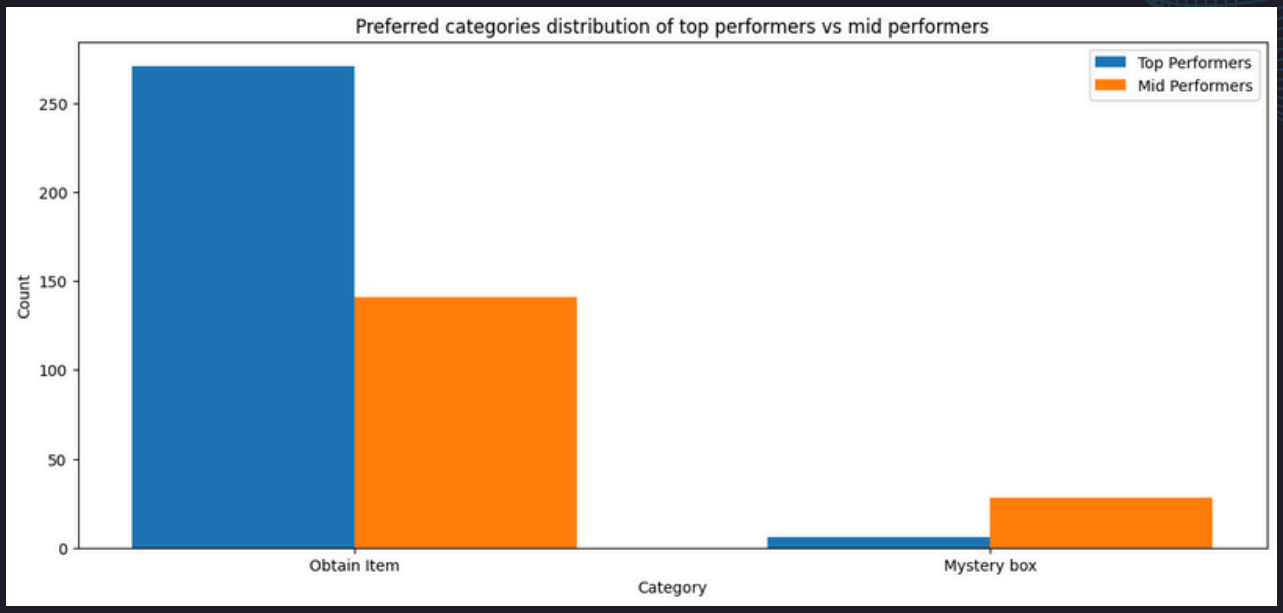


Figure 20: Preferred categories distribution of top performers vs mid perrformers

Preferred platform

This third graph inspects top and mid performers' preferred platforms. Android users have the highest number of mid performers, while top performers are fewer in number but still higher comapred to other platforms. iOS and web platforms have significantly fewer users, with mid performers still outnumbering top performers.

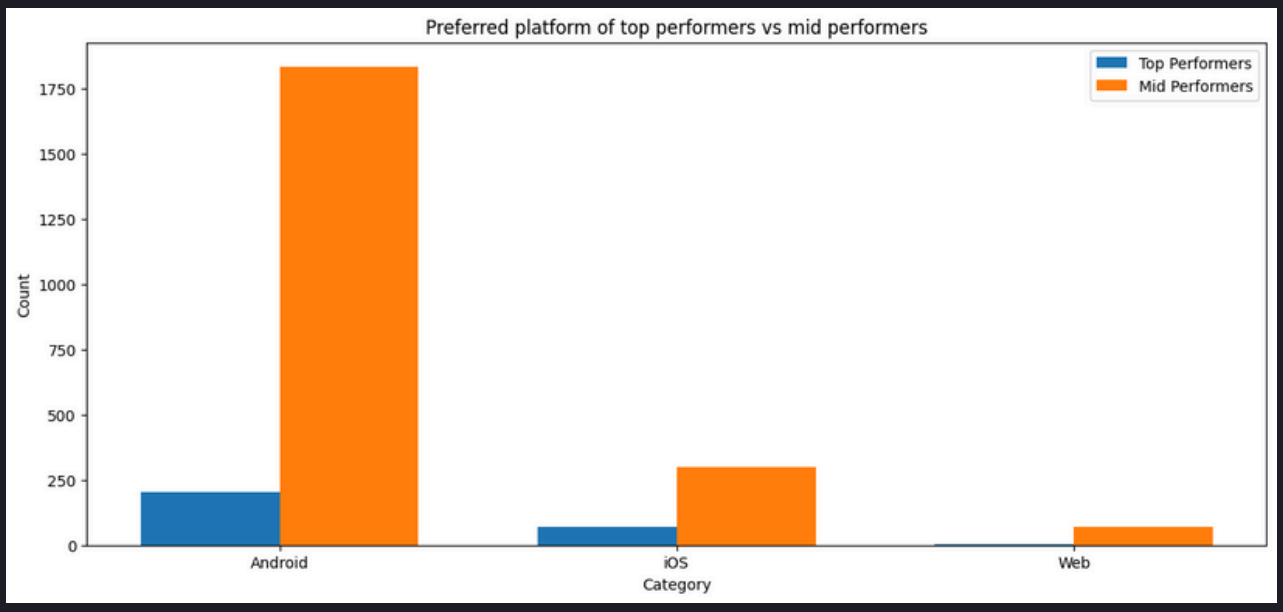


Figure 21: Preferred platform of top performers vs mid perrformers

Players' pronouns

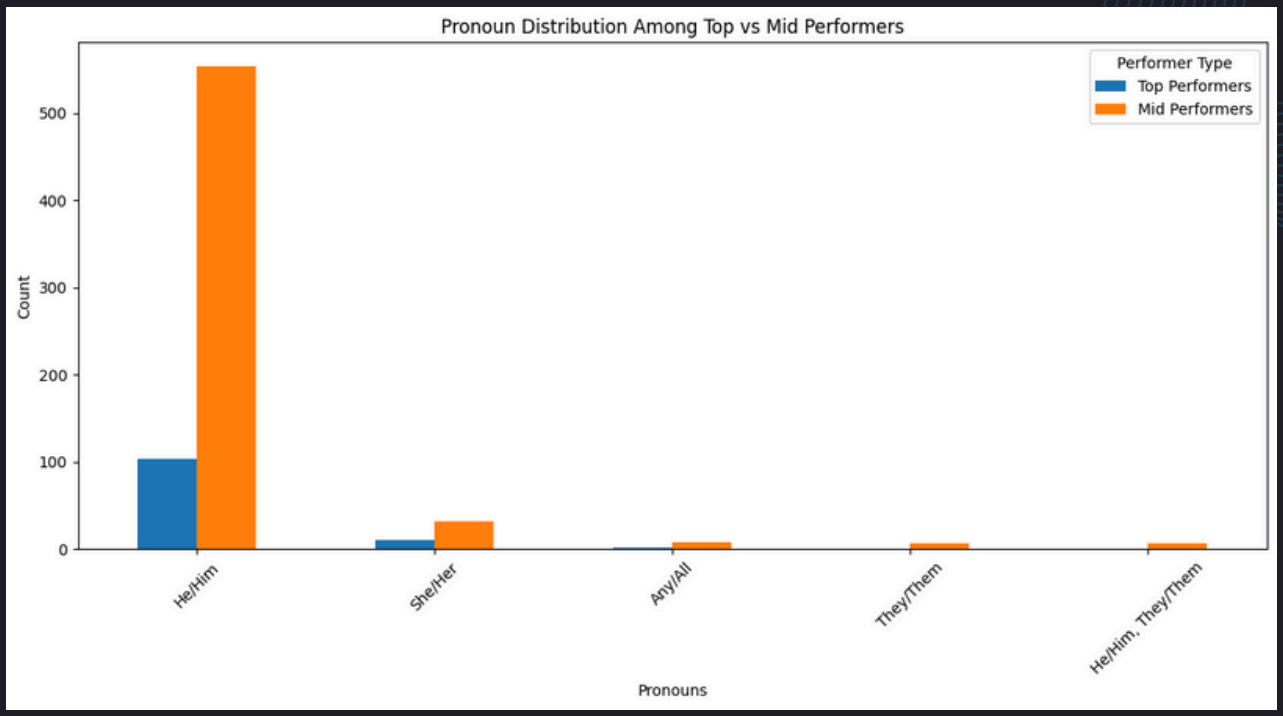


Figure 22: Pronouns distribution among top performers vs mid perrformers

The fourth graph shows pronoun distribution among top and mid performers. The majority of both top and mid performers use "He/Him" pronouns, with mid performers being more prevalent. Other pronouns like "She/Her" and "They/Them" are minimally represented among both top and mid performers.

4.4.3 What's the distribution of top - performers?

I computed a Principal Component Analysis (PCA) of the dataset with KMeans clustering to highlight top performers in the community. The method consists in reducing the dimensional view of a dataset in order to capture the first major variance in the data. Each point in the plot represents an individual data entry, where the color represents top performer status (here gold stands for top performers while blue identifies mid performers).

Three clusters up to the groups of points were detected. The gold points of top performers tend to clump together: they are located predominantly around the upper right and central regions of the plot, meaning that the characteristics of top performers are somewhat dominant. The first principal Component 1 (X-Axis) and second principal Component 2 (Y-Axis) captured the most significant variance in the dataset, thus it kind of simplified the data in a way. Spread along the X- and Y-axes indicates variation in the data: hence, a large spread in more than one direction would show that more than one feature is responsible for performance outcomes.

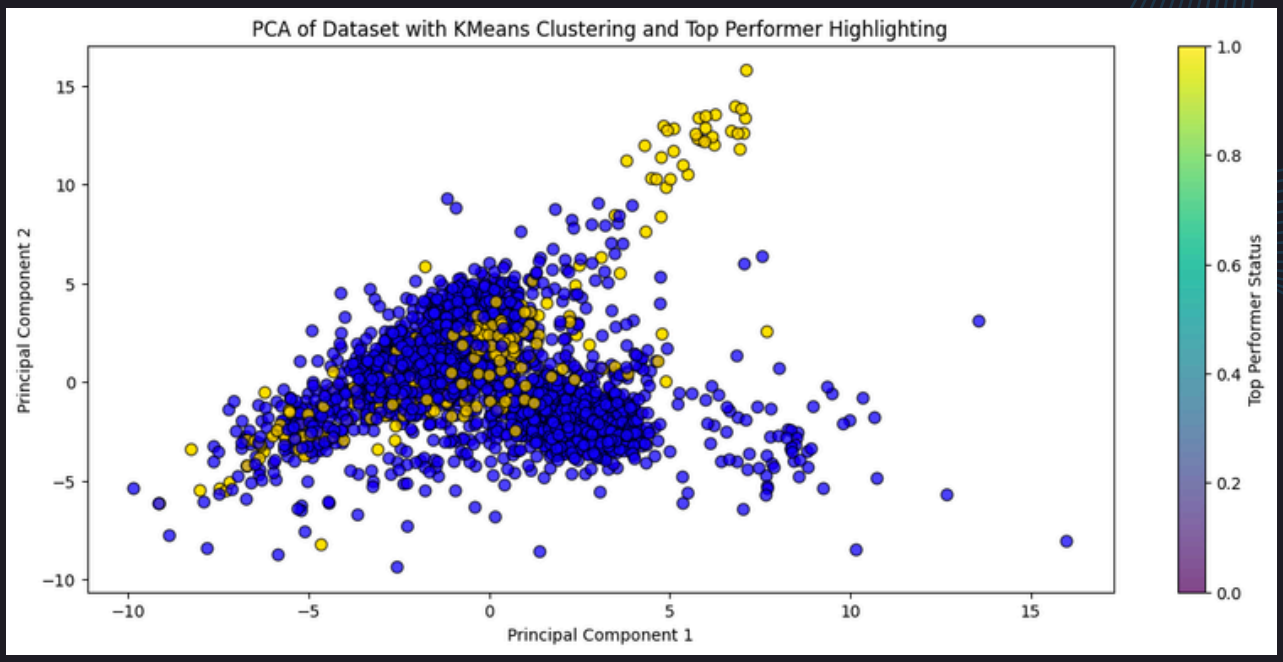


Figure 23: Preferred platform of top performers vs mid perrformers

On the right upper quadrant, there are more high performers. This is likely to correspond to a cluster of attributes associated with high performance. Although diffusion of the very top performers throughout the mid performers implies that top performance might also occur under very different conditions, then most certainly some feature combinations do enhance the probability significantly.

There are some overlappings between the two, which would suggest that while best performers clearly are a group in their own right, there are some product features in which best performers overlap with mid ones. Preponderant evidence of this overlap is that the top performers do not so much differ with the mid performers and some of the mid performers come so close to the top but fail to hit the peak.

It is important to note the clustering and distribution of top performers with respect to establishing the features toward high performance. Such features can help establish an understanding of the components associated with top performance and possibly guide strategies for improvement. The top clusters with high performers can be examined for their characteristics and this will point out the areas pinpointing for improvements and optimizations implementation in the game to game developers and analysts. The understanding derived from the PCA and clustering can further be used to come up with predictive models against feature profiles for potential high performers, which works well in the personalization of recommendations and enriched gaming experiences. Knowing where the top performers are in terms of distributions across the speedrunning community may help in creating different, tailored content or challenges to have a varied environment which attracts new users.

4.4.4 Which variables are correlated?

The correlation matrix visualizes the relationships between different numeric variables in the dataset, quantifying the strength and direction of their linear associations. The correlation coefficient values range from -1 to 1, where:

- 1 indicates a perfect positive correlation
- -1 indicates a perfect negative correlation
- 0 indicates no correlation

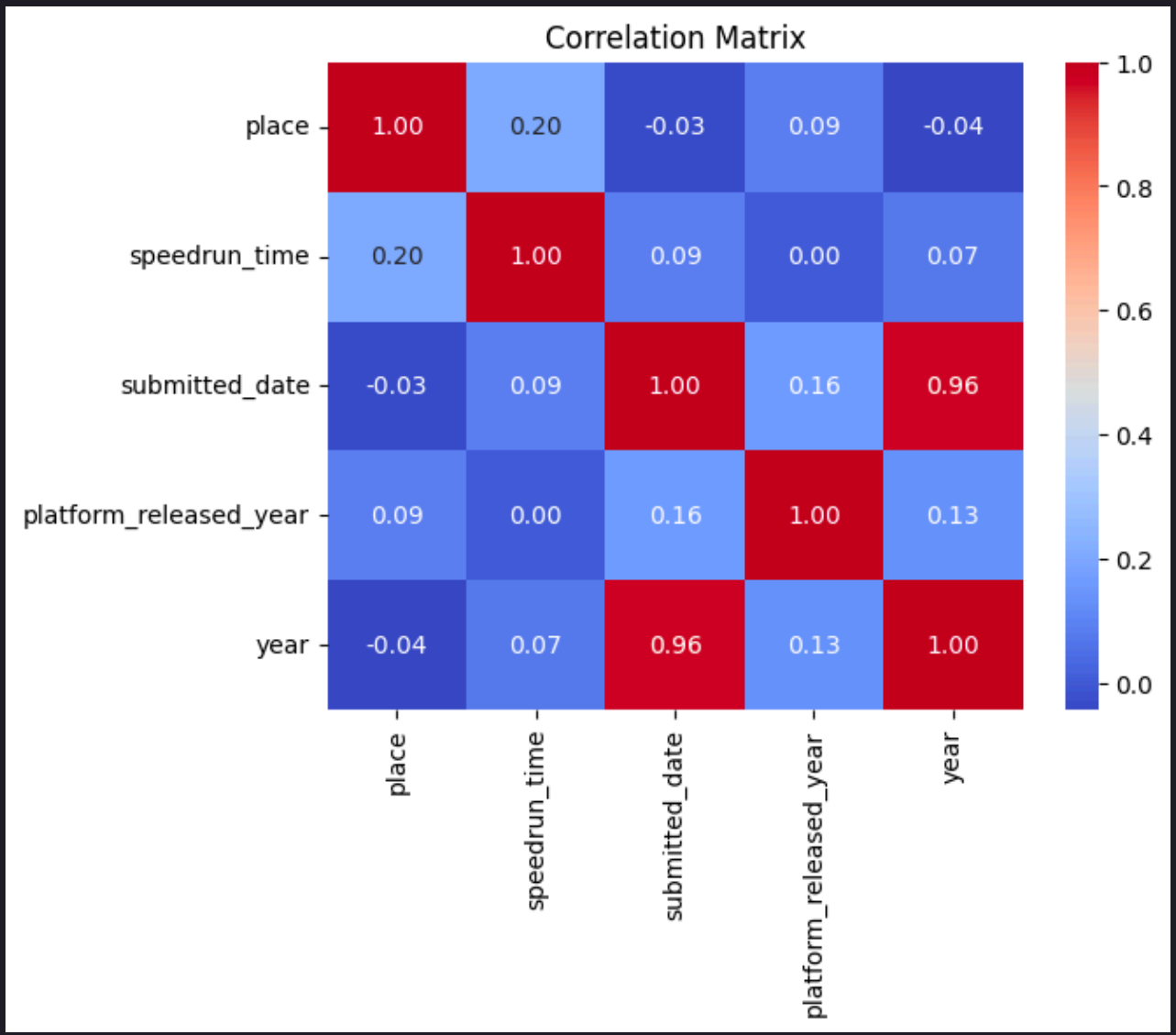


Figure 23: Preferred platform of top performers vs mid perrformers

Firstly, there is a weak positive correlation between place and speedrun time (0.20), indicating that higher ranks (lower place values) are slightly associated with better performance (lower speedrun times). This suggests that faster speedrun times tend to achieve better rankings.

Additionally, the correlation between speedrun time and submitted date is also weakly positive (0.09), indicating no significant relationship between the timing of a speedrun submission and its completion time.

A very strong positive correlation (0.96) exists between submitted date and year, which is expected as both variables are related to time, showing that submissions naturally align with their respective years.

Furthermore, there is a weak positive correlation between submitted date and platform released year (0.16), suggesting that newer platforms might slightly influence the timing of submissions.

Interestingly, the correlation between place and submitted date (-0.03) and between place and platform released year (0.09) are both weak, implying no meaningful relationship between the rank achieved and the submission date or the platform release year.

Overall, the matrix highlights that while time-related variables exhibit strong correlations, other variables like place, speedrun time, and platform released year do not strongly influence each other.

Using the output obtained from the PCA, I then proceeded to compute a variable factor map.

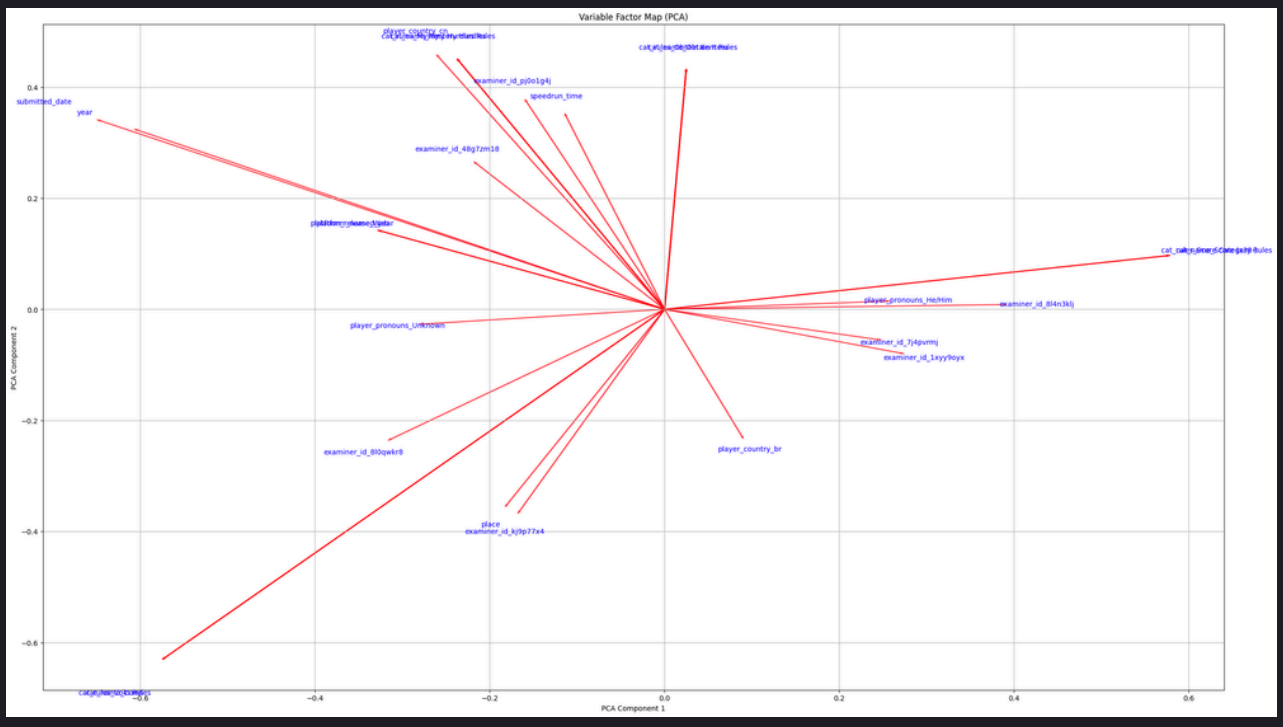


Figure 24: Variable factor map

The map showcases the relationship and influence of various features on the principal components. Each arrow represents a feature, with the direction and length of the arrow indicating its contribution and correlation with the principal components.

The map reveals that features like 'year' and 'submitted_date' have strong positive correlations with the first principal component, as indicated by their long arrows pointing in the same direction.

Conversely, features such as 'cat_name_Score (x38-)' and 'cat_name_Score (x39+)' are positively correlated with the second principal component.

The map also highlights that some features, like 'examiner_id_8loqwkr8' and 'place,' have strong negative correlations with the first principal component, as shown by their arrows pointing in the opposite direction.

Additionally, features such as 'player_country_br' and 'player_pronouns_Unknown' are located close to the origin, indicating a lower influence on the principal components.

The visualization helps identify which features contribute the most to the variance captured by the principal components, allowing for better interpretation and understanding of the underlying data structure.

5 Modelling: predictive analysis

5.1 Building the model / Logistic regression

For the predictive analysis, I tried to find the best model available among the ones I could use. As I needed to best predict speedrun_time values (useful to cluster players as top or mid performer), I used both Linear Rgeression and Neural Network.

5.5.1 Which model best predicts speedrun time?

Linear Regression

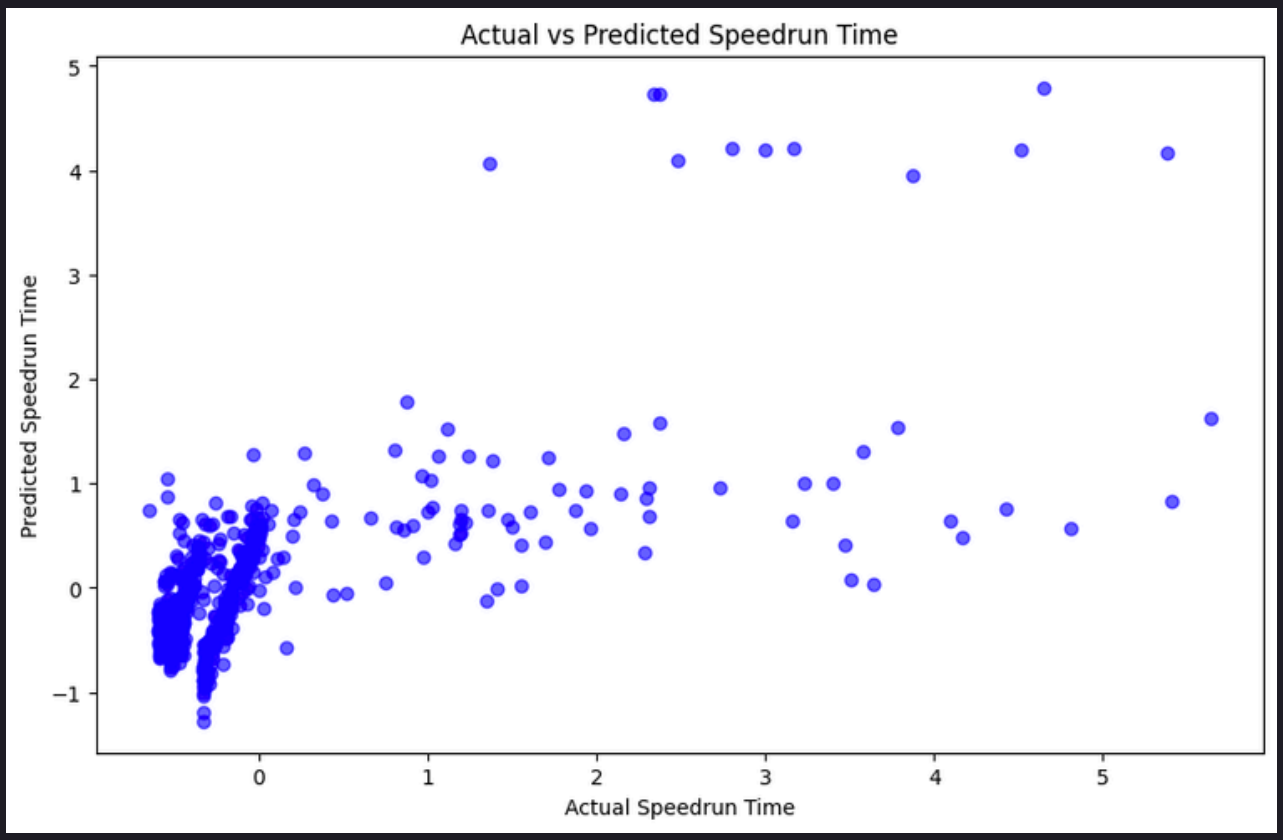


Figure 25: Scatter plot of actual vs predicted speedrun time (

MSE (MEAN SQUARE ERROR)	R² SCORE
0.5519410463752866	0.43491158064074165

table 10: MSE and R² score values

The MSE value indicates the average squared difference between the actual and predicted values, showing that the model's predictions are moderately accurate but still have room for improvement. The R² score, which measures the proportion of variance in the dependent variable that is predictable from the independent variables, is 0.4349, suggesting that approximately 43.5% of the variance in speedrun_time is explained by the model.

This indicates a moderate level of predictive power, but it also highlights that over half of the variance is unexplained, implying that additional features or more complex modelling techniques might be needed to improve accuracy.

The scatter plot of actual vs. predicted speedrun times reveals that while many predictions are close to the actual values, there are several instances where the model underestimates or overestimates significantly, particularly for higher speedrun times. This discrepancy suggests potential non-linearity or outliers in the data that the linear model cannot capture effectively.

Neural Network

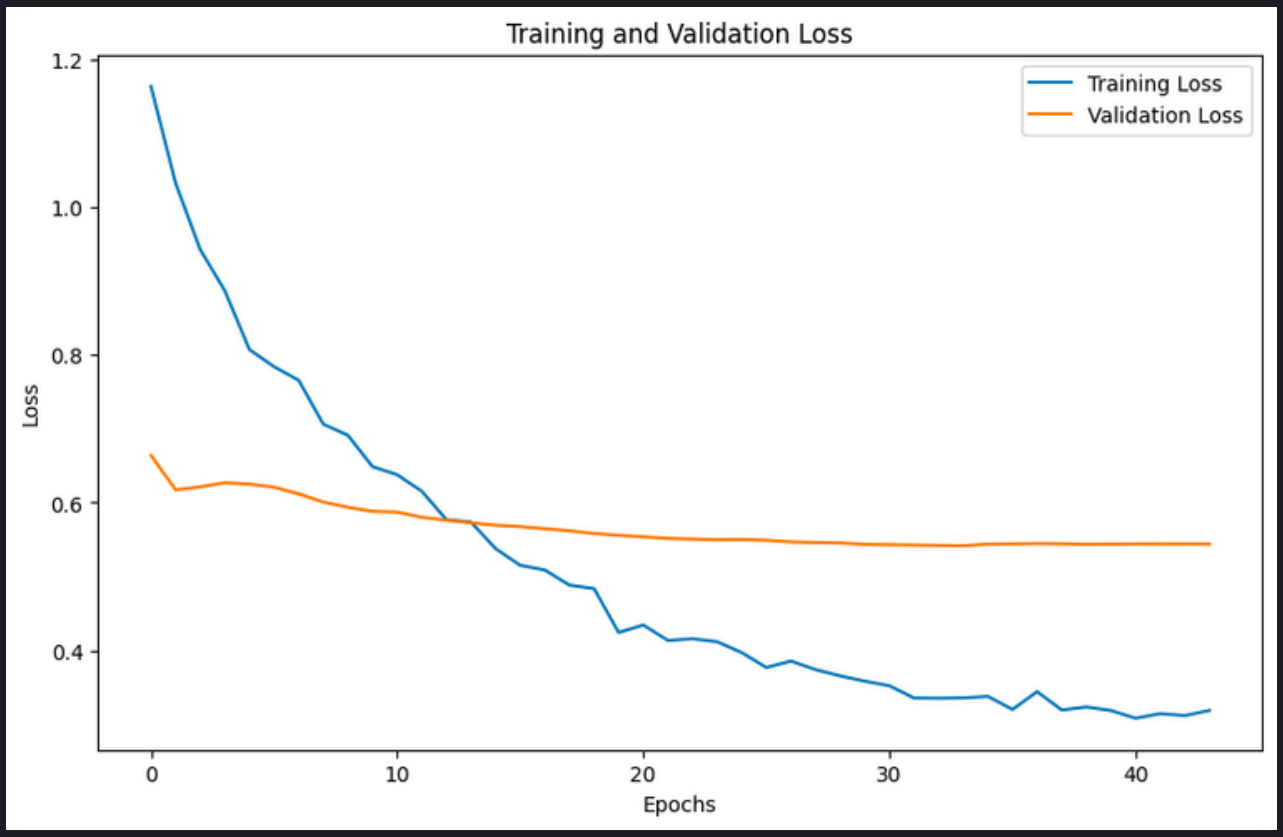


Figure 26: Scatter plot of actual vs predicted speedrun time

The neural network model designed to predict speedrun_time consists of several layers. The training and validation loss curves show a steady decrease in training loss, which indicates that the model is learning from the data. The validation loss, however, flattens out after an initial decrease and does not improve significantly beyond a certain point. This suggests that while the model is effective in learning from the training data, its performance on unseen data (validation set) stabilizes, indicating potential overfitting where the model learns the training data well but does not generalize as effectively.

The neural network's architecture and the training and validation loss trends indicate a

well-constructed model that minimizes training error.

However, the validation loss curve plateau suggests that further tuning or more complex architectures might be required to enhance the model's perfromance.

6 Evaluation: prescriptive analysis

For the product management team at SYBO the insights derived from the data analysis and predictive modeling offer valuable guidance to be addressed to different teams within the company. To translate these insights into concrete actions, I will outline specific recommendations for programmers, designers, and marketers to enhance the overall player experience and engagement.

Platform Optimization
Focus on optimizing game performance for Android devices given their dominance among the player base: ensure smooth gameplay, faster loading times, and reduced bugs on Android platforms. Address the needs of iOS and Web users as well by improving performance and stability on these platforms.

Feature Implementation
Introduce features that cater to top performers, such as advanced performance tracking, leaderboards specific to high achievers and exclusive in-game rewards. This can motivate competitive players to continue engaging with the game.

Data Analytics and Reporting
Implement robust data collection and analytics frameworks to continuously monitor player behavior, platform usage, and performance metrics. This ongoing analysis can

help identify emerging trends and areas for further optimization. Make sure the infrastructure supports seamless integration of new data points and features that may arise from future analyses or player feedback.

Game Design and Balancing

Utilize insights to create game levels and challenges that appeal to both top and mid performers. For instance, design specific levels that offer higher rewards for top performers while providing accessible challenges for mid performers.
Balancing the difficulty levels across different categories is key to ensure a rewarding experience for all players: categories such as "Mystery Hurdles" - which have higher completion times - could be adjusted based on players' performances.

UI/UX

Enhance the UI/UX to highlight achievements and progress for top performers. This can include dynamic leaderboards, personalized dashboards, and visual indicators of performance improvement.

Community and Social Features

Develop social features that encourage community building and interaction among players (community chat, forums, social media integrations...). Share success stories, tips and strategies from top performers to inspire the broader player base. Organise online events, tournaments and live streams to showcase top performers and provide opportunities for everyone to participate and win rewards. Introduce collaborative challenges or team-based events that leverage the competitive nature of top performers while encouraging mid performers to improve and participate.

Retention and Acquisition Strategies

Implement retention strategies that focus on rewarding consistent play and improvement (loyalty programs, daily challenges, exclusive in-game items for long-term players...). Develop acquisition strategies that leverage the popularity of specific platforms and regions: for example, targeted ads and promotions in Brazil can capitalize on the large Android user base.

These recommendations - if implemented - will help SYBO Games to increase overall player experience, strenghten their community and uphold retention and acquisition.

SYBO

SUBWAY SURFER

Patterns in Speedrunning: a data mining approach

Written by

Matteo Del Prato

m.matteodelprato@gmail.com

+39 334 84 75 543