DAMG 7370 – Final Project Report
April 22, 2023

Dota 2 Statistical Analysis

Team – 2

NUID	Name	Role	Summary
002839737	Ragul Paramasivam	Data Engineer	i. Led the development of data pipelines and ETL processes. ii. Contributed to designing and implementing infrastructure for efficient data storage and processing. iii. Collaborated with team members to optimize data workflows and support project objectives.
002809300	Kanimozhi Velusamy	Data Analyst	i. Conducted exploratory data analysis to extract insights from the dataset. ii. Developed visualizations and reports to communicate findings effectively. iii. Contributed to identifying trends and patterns in the data, supporting decision-making processes.
002732528	Aishwarya Sriyapu Reddy	Clout Setup and BI Developer	i. Set up cloud infrastructure and developed BI solutions, including reports and dashboards. ii. Gathered project requirements and implemented BI artifacts to meet academic objectives. iii. Contributed to data integration and ETL processes, ensuring data accuracy and reliability.

1. Introduction

1.1. Background:

Dota 2, introduced by Valve Corporation in 2013, emerged from the vibrant modding community of Warcraft III: Reign of Chaos. The original mod, Defense of the Ancients (DotA), garnered immense popularity, prompting Valve to develop a standalone version using the Source game engine. While retaining the core gameplay mechanics of DotA, Dota 2 introduces modern graphics, features, and competitive elements. Since its release, Dota 2 has become one of the most popular and competitive multiplayer online battle arena (MOBA) games worldwide, attracting millions of players and establishing a thriving esports scene.

1.2. Objectives:

The objectives of this data engineering and business intelligence project are outlined as follows:

- Develop a comprehensive analysis of Dota 2 matches from professional tournaments.
- Utilize the Open Dota API, Apache Spark, and Tableau to gather, process, and visualize the data effectively.
- Provide detailed insights into the performance of teams and players, leveraging various performance metrics.
- Identify significant trends and patterns within the dataset to understand the dynamics of Dota 2 gameplay.
- Based on the analysis, offer actionable recommendations for teams and players to improve their performance.

1.3. Scope:

The scope of this analysis encompasses several key areas:

- Exploration of hero and team performance metrics, including win rates, pick rates, and objective completion rates, to assess effectiveness and detect trends.
- Investigation of the evolution of gameplay strategies over time, such as popular hero combinations, drafting strategies, and counterplay dynamics.

- Analysis of match outcomes and factors influencing victory or defeat, such as early-game advantages, late-game strategies, and comeback mechanisms.
- Provision of recommendations for future research directions, enhancements in data collection methods, and advanced analysis techniques to deepen the understanding of Dota 2 gameplay dynamics.

This structured approach sets clear goals and boundaries for the analysis, ensuring a focused and comprehensive examination of Dota 2 matches from professional tournaments.

2. Game Overview

2.1. Brief history of DOTA 2:

Dota 2 results from the evolution of the original Defense of the Ancients (DotA), a custom map for Warcraft III: Reign of Chaos. DotA, created by a community of passionate modders, gained immense popularity within the Warcraft III community, eventually becoming a standalone phenomenon. Valve Corporation recognized the potential of DotA and hired Ice Frog, the lead developer behind DotA, to develop a sequel in the Source game engine. Development began in 2009, with Valve's commitment to preserving the core gameplay while enhancing graphics, accessibility, and competitive features. Dota 2 was officially released in 2013, marking a new era for the franchise. Since then, Valve has continuously supported and updated the game, introducing new heroes, balance changes, and features to maintain its relevance in the gaming landscape.

2.2. Gameplay Mechanics:

Dota 2 is a highly strategic and complex game that revolves around team-based combat and objective control. Players control one of over 100 heroes, each with its own set of abilities and attributes. The game is played on a map divided into three lanes, called the top, middle, and bottom lanes, with each lane defended by defensive towers and waves of AI-controlled units called creeps. Players must navigate their hero through these lanes, battling enemy creeps and heroes and destroying enemy towers to advance toward the enemy base. Key mechanics include last hitting, where players must deliver the killing blow to enemy creeps to earn gold; denying, where players can prevent enemy heroes from gaining experience by killing their creeps; and itemization, where players purchase items to enhance their hero's abilities and stats.

2.3. Roles and Hero Selection:

Dota 2 features a diverse roster of heroes, each fulfilling specific roles within a team composition. These roles include:

- Carry: Heroes that scale well with items and deal high damage in the late game.
- Support: Heroes that assist their teammates with utility, crowd control, and healing.
- Initiator: Heroes that excel at starting team fights and disrupting the enemy team's positioning.
- Jungler: Heroes that farm neutral creeps in the jungle to gain experience and gold.
- Offlaner: Heroes that occupy the offlane, a solo lane that requires resilience and survivability.

Hero selection is a critical aspect of Dota 2, as teams must draft a balanced lineup that addresses their strengths, weaknesses, and objectives. The drafting phase consists of a series of bans and picks, where teams strategically choose heroes to counter their opponents' picks and execute their game plan effectively.

2.4. Objectives and Victory Conditions:

The primary objective in Dota 2 is to destroy the enemy team's Ancient, a large structure located within their base. To achieve this, teams must coordinate their efforts to push lanes, secure map control, and win team fights. In addition to destroying the Ancient, teams can also achieve victory through other means:

- Surrender: Teams may choose to concede defeat by voting to surrender, ending the match prematurely.
- Abandonment: If a player leaves the match for an extended period, their team may be declared the loser.

- Matches typically last between 30 to 60 minutes, although shorter or longer matches are possible depending on various factors such as team composition, strategies employed, and player skill levels.

3. Architecture

The architecture consists of Open Dota as the data source, Databricks for data orchestration and processing, PostgreSQL for data storage, and Tableau for data visualization. This architecture enables the project to collect, process, analyze, and visualize Dota 2 match data efficiently, providing valuable insights for stakeholders and decision-makers.

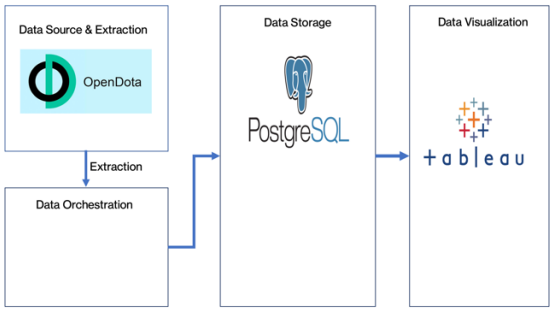

Figure 1. Architecture Diagram

i. Data Source & Extraction – Open Dota:
   a. Open Dota is the project's primary data source, providing information about Dota 2 matches, players, heroes, and tournaments.
   b. The project leverages various endpoints the Open Dota API provides to gather the necessary data for analysis.

ii. Data Orchestration – Databricks:
   a. Databricks is the data orchestration platform that facilitates data processing and analysis tasks.
   b. Databricks provides a unified analytics platform built on top of Apache Spark, enabling efficient data processing at scale.
   c. The project utilizes Databricks to execute data transformation, analysis, and machine learning tasks in a distributed and parallelized manner.

```
In [0]: draft_timing_spark.show()
```

order	pick	active_team	hero_id	player_slot	extra_time	total_time_taken	match_id
1	false	3	120	null	130	16	7699602043
2	false	2	66	null	130	19	7699602043
3	false	2	114	null	130	12	7699602043
4	false	3	91	null	130	12	7699602043
5	false	2	87	null	130	19	7699602043
6	false	2	65	null	130	45	7699602043
7	false	3	52	null	121	30	7699602043
8	true	3	78	6.0	105	18	7699602043
9	true	2	19	3.0	121	37	7699602043
10	false	3	54	null	121	25	7699602043
11	false	3	63	null	99	29	7699602043
12	false	2	86	null	99	22	7699602043
13	true	2	5	0.0	121	1	7699602043
14	true	3	51	7.0	121	28	7699602043
15	true	3	79	8.0	99	80	7699602043
16	true	2	81	1.0	78	51	7699602043
17	true	2	97	2.0	71	43	7699602043
18	true	3	48	5.0	40	60	7699602043
19	false	3	13	null	66	40	7699602043
20	false	2	22	null	58	37	7699602043

only showing top 20 rows

Figure 2. Sample data load in spark

iii. Data Storage – PostgreSQL:
   a. PostgreSQL serves as the project's data storage solution, housing processed and transformed data.

b. The project utilizes PostgreSQL to store structured data in relational tables, providing a robust and scalable storage solution for storing Dota 2 match data, player statistics, and other relevant information.

```
url = f"jdbc:postgresql://{database_host}:{database_port}/{database_name}"

remote_table = (spark.read
  .format("jdbc")
  .option("driver", driver)
  .option("url", url)
  .option("dbtable", table)
  .option("user", user)
  .option("password", password)
  .load()
)

In [0]: remote_table.show()

-------+----------+
|version|  match_id|leagueid|start_time|duration|series_id|series_type|cluster|replay_salt|radiant_win|pre_game_dur
ation|match_seq_num|tower_status_radiant|tower_status_dire|barracks_status_radiant|barracks_status_dire|first_blood
_time|lobby_type|human_players|game_mode|flags|engine|radiant_score|dire_score|radiant_team_id|    radiant_name|
radiant_logo|radiant_team_complete|dire_team_id|        dire_name|          dire_logo|dire_team_complete|radiant_c
aptain| dire_captain|metadata|patch|region| throw|   loss|          replay_url|radiant_team|dire_team|comeback|  st
omp|picks_bans|
+-------+----------+--------+----------+--------+---------+-----------+-------+-----------+-----------+-----------
-----+-------------+--------------------+-----------------+-----------------------+--------------------+-----------
-----+----------+-------------+---------+-----+------+-------------+----------+---------------+----------------+---
----------------+--------------------+------------+------------------+--------------------+------------------+-----
---------+------------+--------+-----+------+------+-------+--------------------+------------+---------+--------+
-------+----------+
|     21|7694225732|   15890|1713582234|    1687|   870159|          1|    151|  317130281|       true|
90|   6476763119|                2046|                6|                     63|                   3|           1
79|         1|           10|        2|    5|     1|           46|         12|      7915344.0|  SÖLAN E-Sports|161055
4771280220400|                    0.0|   9409423.0|Eagle Eye Esports|2.504639701999664...|                0.0|   2.515
11609E8|  1.2245299E8|    null|   54|   5.0|   0.0|30252.0|http://replay151....|        null|     null|    null|
null|      null|
```

Figure 3. Connecting to PostgreSQL

iv. Data Visualization - Tableau:

a. Tableau serves as the data visualization tool for the project, enabling the creation of interactive and visually appealing dashboards and reports.

b. The project utilizes Tableau to visualize insights and findings from data analysis.

c. Tableau's intuitive interface and powerful visualization capabilities allow stakeholders to explore and interpret the data effectively.

4. Methodologies

4.1. Data Collection:

Data will be collected using the OpenDota API, which provides access to a comprehensive range of data on Dota 2 matches from professional tournaments. This includes match details, player statistics, and team performance metrics. The API offers a reliable and efficient means of gathering relevant data for analysis.

We utilized the OpenDota API extensively to gather the data needed for our project. By leveraging the various endpoints provided by the OpenDota API, we were able to access a wealth of information on professional Dota 2 matches, players, heroes, and tournaments.

1) To collect data on professional Dota 2 matches, we made use of the `https://api.opendota.com/api/proMatches` endpoint. This allowed us to retrieve details such as match IDs, start times, durations, participating teams, and tournament information for a wide range of professional matches.
2) For detailed information on individual Dota 2 matches, we utilized the `https://api.opendota.com/api/matches/{match_id}` endpoint. By specifying the unique match ID, we were able to access comprehensive match statistics, including player performance metrics, hero picks, kills, deaths, assists, and more.
3) To gather data on professional players participating in Dota 2 tournaments, we accessed the `https://api.opendota.com/api/proPlayers` endpoint. This provided us with player-specific information such as account IDs, steam IDs, player names, and team affiliations, enabling us to analyze the performance of individual players.
4) The `https://api.opendota.com/api/players/{account_id}` endpoint allowed us to retrieve detailed statistics and performance metrics for specific Dota 2 players identified by their account IDs. This

included valuable data such as match history, win rates, hero preferences, and skill ratings, which are essential for analyzing player performance trends.

5) To obtain information about Dota 2 heroes, including hero IDs, names, and primary attributes, we accessed the `https://api.opendota.com/api/heroes` endpoint. This gave us crucial insights into hero characteristics and attributes, aiding in analyzing hero popularity, win rates, and role distributions.
6) We utilized the `https://api.opendota.com/api/heroStats` endpoint to retrieve detailed statistics and performance metrics for Dota 2 heroes. This included data such as pick rates, win rates, and performance trends across different skill brackets and game modes, enabling us to analyze hero performance comprehensively.
7) To gather information about Dota 2 tournaments and leagues, including league IDs, names, start dates, and end dates, we accessed the `https://api.opendota.com/api/leagues` endpoint. This gave us valuable insights into tournament schedules, participant information, and tournament trends.

By leveraging these OpenDota API endpoints, we collected diverse data on professional Dota 2 matches, players, heroes, and tournaments, enabling us to conduct comprehensive analysis and derive actionable insights for our project objectives.

4.2. Data Preprocessing:

Collected data will undergo preprocessing to ensure its quality and suitability for analysis. This involves cleaning the data to handle outliers and inconsistencies. Additionally, the data will be transformed and organized into a structured format conducive to analysis.

4.3. Analysis Techniques:

The analysis will use Apache Spark, leveraging its in-memory processing capabilities for efficient data analysis. The project will focus on several key areas:

- Team Performance Analysis: This involves analyzing teams' performance in professional Dota 2 tournaments. Metrics such as win rates, kill rates, and economic performance will be calculated. The analysis will identify trends and patterns in team performance and provide actionable recommendations for improvement.
- Player Performance Analysis: Individual player performance will also be analyzed using metrics such as kill rates, death rates, and gold accumulation. Trends and patterns in player performance will be identified, and recommendations for improvement will be provided.
- Match Outcome Prediction: Machine learning algorithms will be employed to predict the outcome of Dota 2 matches. Historical match data will be used to train the model and make predictions for future matches. This predictive analysis aims to provide valuable insights into match outcomes and potential factors influencing them.
- Data Visualization: Tableau will be utilized to visualize the insights gained from the data analysis. Interactive dashboards and reports will be created to present the findings clearly and concisely, enabling stakeholders to easily interpret and act upon the insights derived from the analysis.

By employing these methodologies, the project aims to comprehensively analyze Dota 2 matches from professional tournaments, providing valuable insights into team and player performance, predicting match outcomes, and presenting findings through effective data visualization techniques.

5. Exploratory Data Analysis

During the Exploratory Data Analysis (EDA) phase, we delved into the dataset using various visualization techniques to uncover insights and patterns. Here's a summary of our findings.

5.1. Basic Insights:

i. Picks vs. Non-Picks Distribution:

We examined the dataset's distribution of picks and non-picks for each hero. We visualized the frequency of hero selections in professional Dota 2 matches through stacked bar charts. This analysis revealed the popularity of certain heroes and provided insights into hero preferences among players and teams.

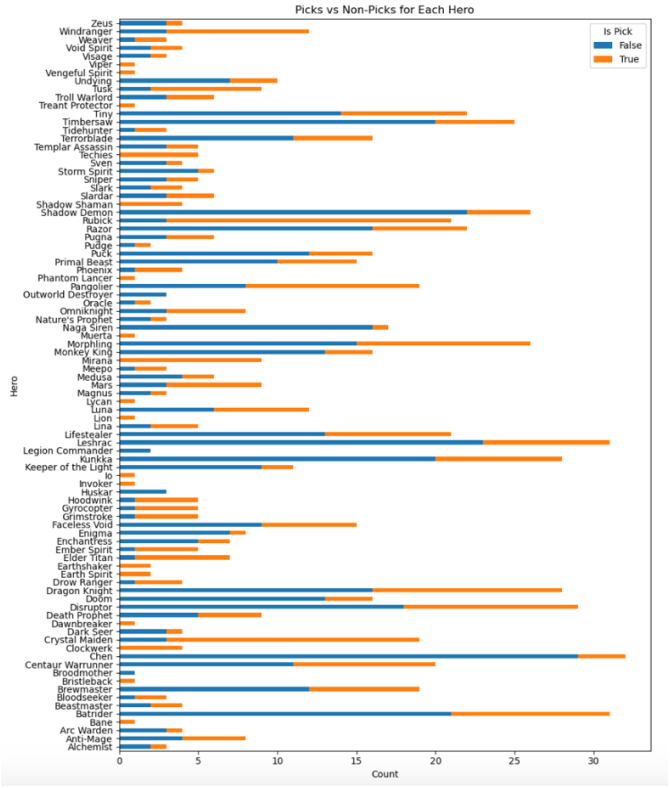

Figure 4. Picks vs. Non-Picks Distribution

ii. Distribution of Active Teams:

We visualized the distribution of active teams in the dataset by creating a pie chart. This analysis highlighted the representation of different teams in professional Dota 2 matches, shedding light on the diversity of participating teams and their respective presence in tournaments.

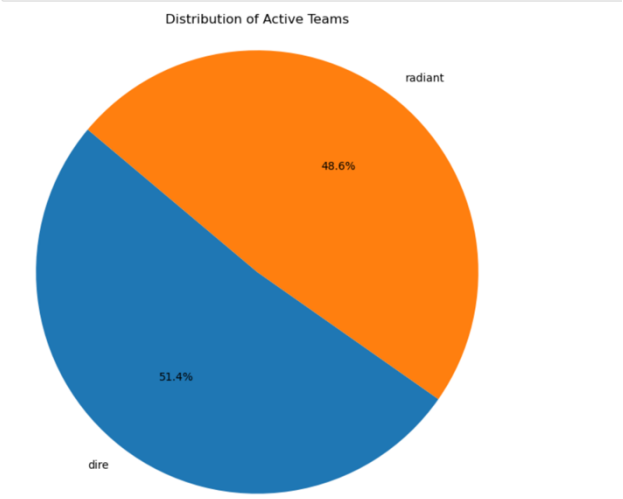

Figure 5. Distribution of Active Teams

5.2. Additional Insights:

Depending on the nature of our data and the research questions, we explored various visualization techniques, such as histograms, scatter plots, and box plots, to uncover additional insights and relationships within the data. Overall, our EDA revealed valuable insights into the distribution of picks, the presence of active teams, and the overall characteristics of the dataset. These findings serve as a foundation for further analysis and decision-making in our project.

6. Key Metrics and Insights using Tableau

6.1 Hero Performance Analysis

A.) Gold Analysis of the Hero

The graph below provides data analysis of the top 10 heroes with the highest gold accumulation. From the graph below, we can infer that:

- Hero ID 48 has the highest gold accumulation of 862442; the average gold per minute gained in the game is 732.2.

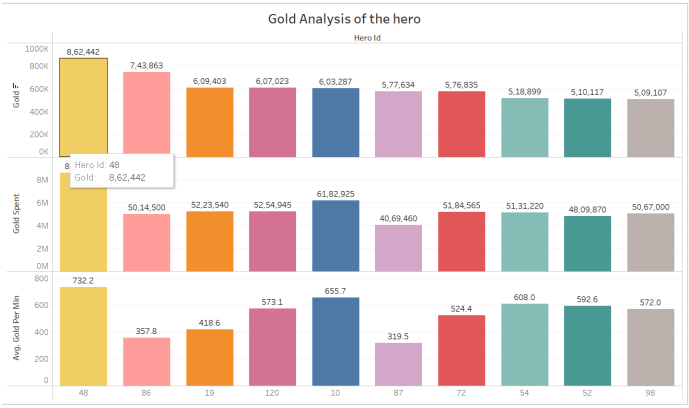

Figure 6. Gold Analysis of the Hero

B.) Hero Statistics

The graph below shows data on the Heroes' healing, damage, and kills. From the graph, we can infer:

- The graph displays the top 10 heroes with high kills and damage and corresponding healing characteristics.
- Hero 10 has the highest scores in all three characteristics: healing, Damage, and Kills, which makes him an ideal and strong player candidate for scoring points and helping the team win.

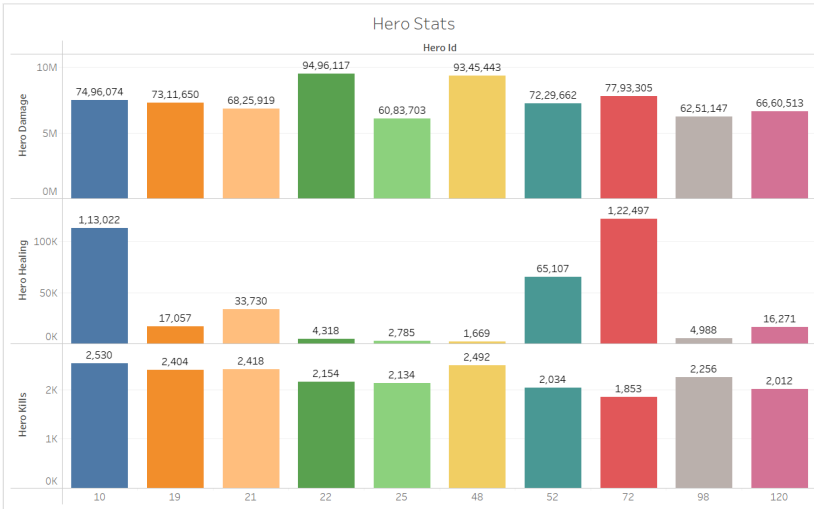

Figure 7. Hero Statistics

C.) Pick and Bans of a Hero

The graph displays the hero's number of picks and bans in all the games.

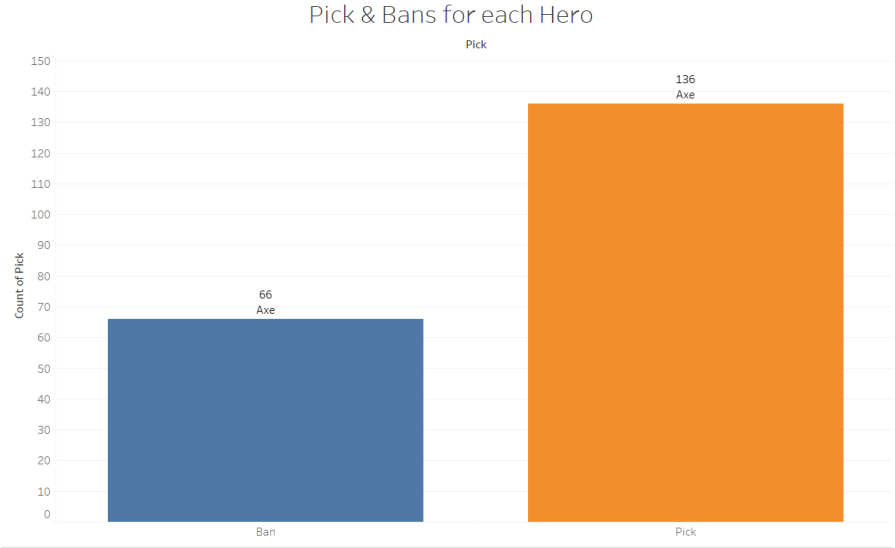

Figure 8. Pick and Bans of a Hero

D.) Top 8 Hero in-game with high Gold and XP

The graph shows heroes' top total experience (XP) and gold (earned), with hero IDs on the x-axis.

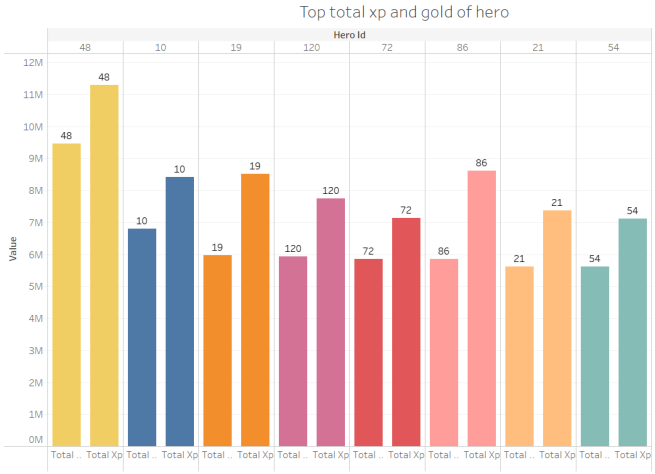

Figure 9. Top Xp and Gold of Heroes

6.2. Performance Analysis

A.) The average duration of each team's matches

The graph shows the average duration of the matches played by each team for the top 8 teams. Only three teams, 9401919, 9390116, and 5, have average match durations above the overall average of 3373.

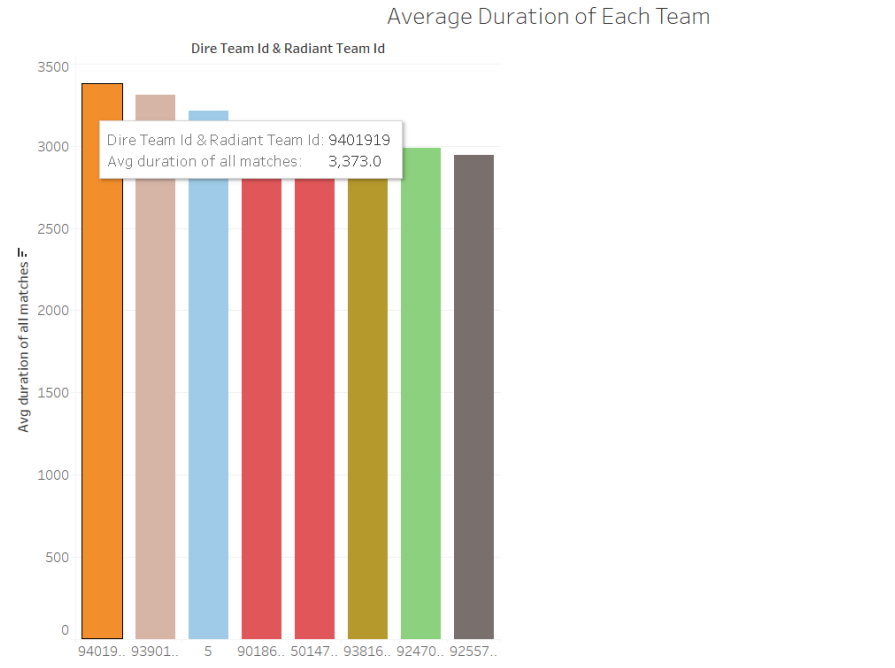

Figure 10. The average duration of each team

6.3. Match Outcomes

A.) Top Items used in the game

The graph displays the top 15 items used in this game and the top 5 items are boots, tango, braches, magic stick, tpscroll.

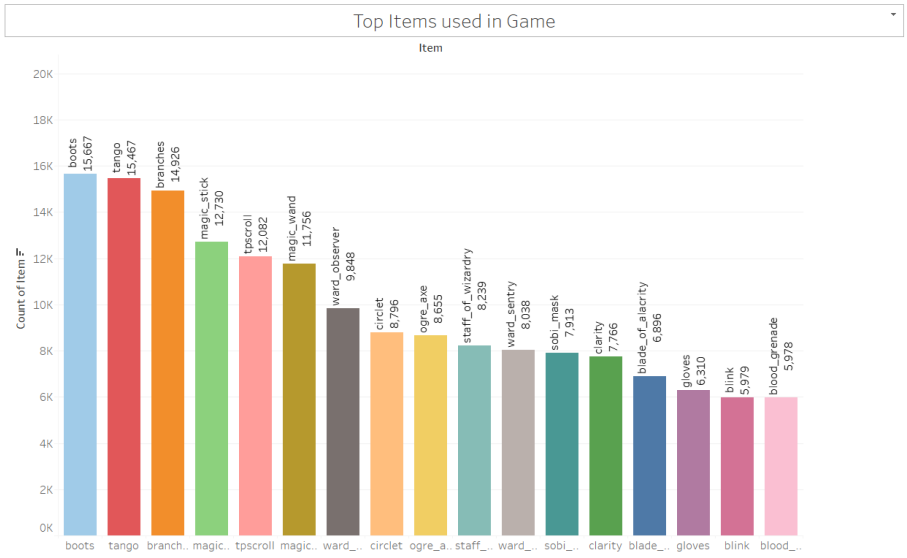

Figure 11. Top items used in-game

B.) Adv Gold and Adv Xp for each match

The graph displays the progression of Adv Gold and Adv Xp from the beginning to the end of the match between the Radiant and Dire teams.

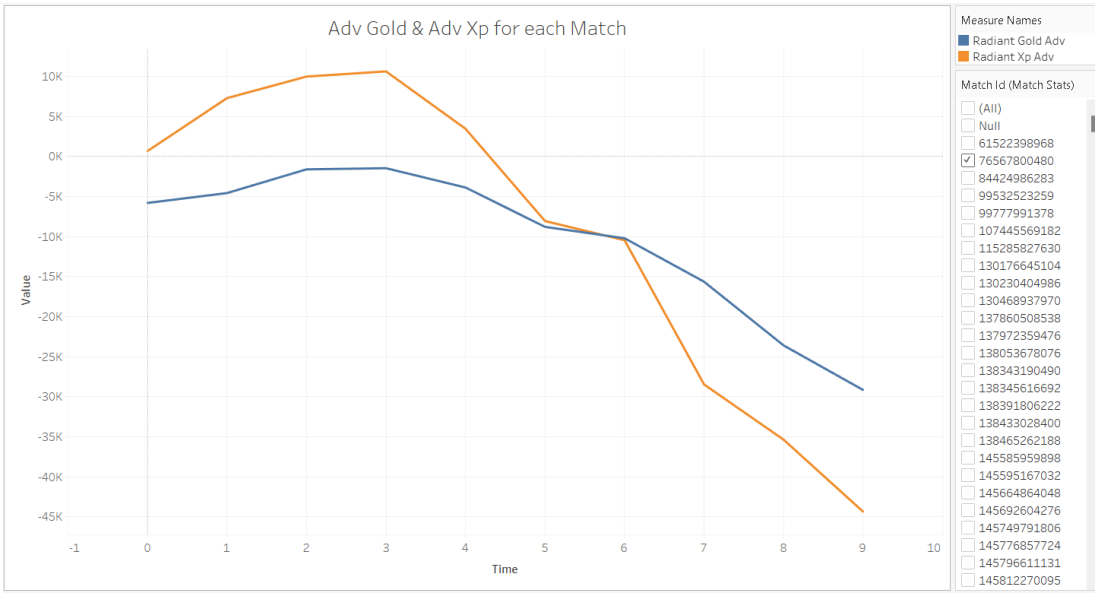

Figure 12. Adv. Gold and XP for each match

C.) Win Percentage of Teams in the Match

The table displays the Average score, Average win percentage, and time taken by each team selected based on past matches.

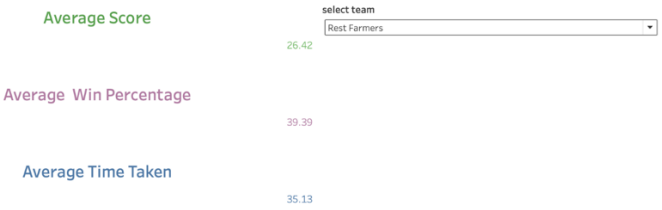

Figure 13. Win Percentage of Teams in the Match

D.) Total matches and win matches between Radiant and Dire Team

This Tree map displays,

- The total matches played between the Radiant Team and Dire Team
- Many Radiant Win matches also imply the number of Radiant lost matches.
- This graph facilitates the user's selection of Teams and Leagues, allowing them to see the match details of each league's Radiant and Dire teams.

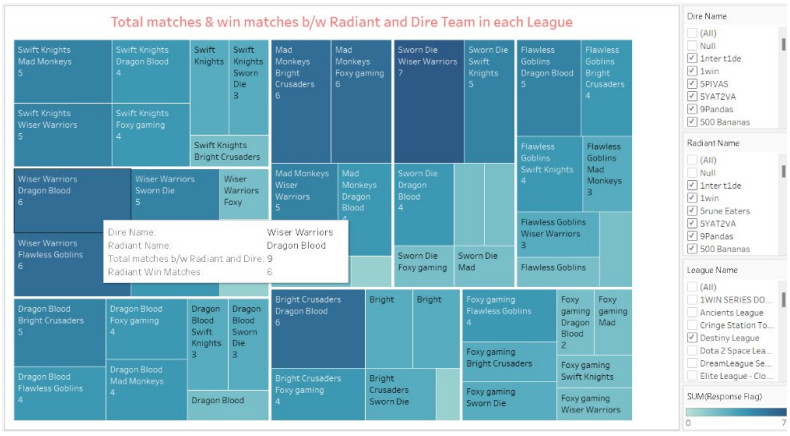

Figure 14. Total matches and win matches between Radiant and Dire Team

E.) Dashboard

- It showcases a few essential details of the match, team, and metrics.
- It also showcases the Radiant team’s total matches and wins per league.

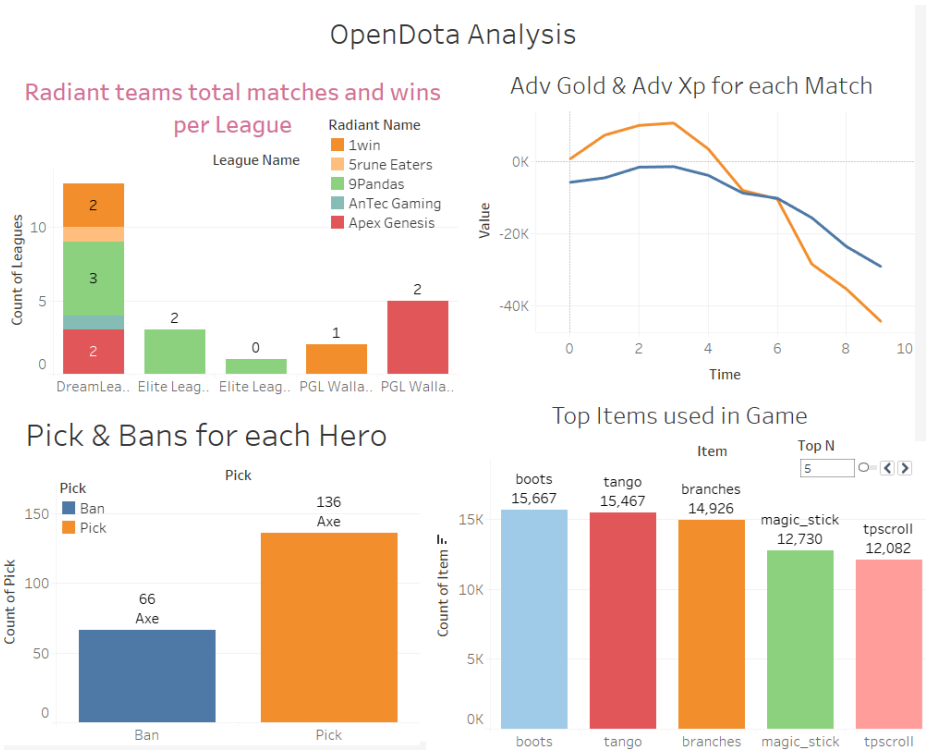

Figure 15. Open Dota dashboard

7. Future Directions

In the future, there are several avenues for further enhancing our analysis of Dota 2 matches and expanding the capabilities of our project:

7.1. Potential improvements in Data Collection:

- Explore additional data sources: We look into incorporating data from alternative sources, such as additional APIs or data scraping techniques, to enrich the dataset and capture a more comprehensive view of Dota 2 matches.
- Enhance data quality: We plan to implement measures to improve data accuracy, completeness, and reliability by addressing issues such as missing values, outliers, and data inconsistencies.
- Expand data coverage: We plan to expand the scope of data collection to include a broader range of tournaments, matches, and regions, enabling a more comprehensive analysis of Dota 2 gameplay dynamics in the future.

7.2. Advanced Analysis Techniques:

- Incorporate network analysis: We will apply network analysis techniques to model and analyze the complex interactions between heroes, players, and teams during matches, gaining insights into strategic alliances, synergy, and gameplay dynamics.
- Leverage natural language processing (NLP): We are thinking about exploring the application of NLP techniques to analyze text-based data, such as player chats and commentary transcripts, to extract sentiment, identify key themes, and understand player behavior and communication patterns.

7.3. Integration of scheduler with real-time data for live match analysis:

- Develop a scheduler: Design and implement a scheduler system to automate the retrieval and processing of real-time data from Dota 2 matches as they unfold. This will enable live match analysis and timely insights for teams, analysts, and spectators.
- Integrate real-time data feeds: Integrate with live data streams or APIs provided by platforms like Steam or tournament organizers to access real-time match information, player statistics, and event data.
- Implement real-time analytics: Leverage the capabilities of Spark Streaming or other real-time processing frameworks to perform continuous analysis of live match data, enabling instant insights, alerts, and visualizations for stakeholders.

By exploring these future directions, we aim to enhance the depth and breadth of our analysis, provide more valuable insights to the Dota 2 community, and contribute to advancing esports analytics.

8. Conclusion:

- The project delivered valuable insights into the performance dynamics of teams and individual players within professional Dota 2 tournaments. This involved thoroughly exploring various aspects of gameplay, strategy, and execution.
- The project adopted a sophisticated analytical approach, utilizing the OpenDota API, Apache Spark, and Tableau capabilities. This enabled the comprehensive exploration of multifaceted datasets and extracting meaningful patterns.
- By meticulously analyzing a substantial dataset comprising Dota 2 matches, the project uncovered actionable recommendations for enhancing performance across different dimensions. These recommendations were grounded in empirical evidence drawn from the data.
- The project's in-depth examination of team and player performance provided valuable insights into underlying trends and patterns within the Dota 2 competitive landscape. Moreover, developing a machine learning model for outcome prediction marked a significant milestone, offering the potential for informed decision-making in future matches and tournaments.

References

[1] Open Dota Documentation, the document to get the API and the information regarding it: https://docs.opendota.com

[2] Howard Chung (2024), Item data used in this project retrieved from the GitHub project https://github.com/odota/dotaconstants/blob/master/build/item_ids.json