Predicting Clash Royale Wins Through Deck Composition

(COMP3125 Individual Project)

Raghav Vaid
Society of Computing and Data Science
Wentworth Institute of Technology

***Abstract*—This project analyzes how Clash Royale deck characteristics influence match outcomes using a cleaned and standardized version of a public card-level dataset. After preprocessing the data, I generated synthetic decks, engineered gameplay-relevant features, and applied a logistic regression model to identify which attributes contribute most to winning. The analysis shows clear relationships between elixir cost, troop count, and damage output in predicting success. The project demonstrates how statistical modeling can quantify deck strength and provides a foundation for more advanced gameplay analytics.**

Keywords—Clash Royale, machine learning, logistic regression, data analysis, deck prediction

I. Introduction (*Heading 1*)

Clash Royale is a fast-paced, competitive mobile game built around strategic card selection and real-time decision-making. Each match is influenced not only by player skill, but also the structure of the deck itself. Elixir curve, troop types, spell choices, and the interactions between cards all are an important aspect to the game. Because the game is frequently updated, its meta shifts rapidly, creating natural uncertainty around which decks perform well and why. Data-driven analysis offers a way to cut through that noise by examining measurable relationships between card attributes, deck composition, and match outcomes.

This project explores how deck structure affects win rates by analyzing an existing card-level dataset sourced from an earlier version of the game. Although the dataset reflects an outdated meta, the underlying strategic principles remain consistent: certain card traits increase synergy, elixir management affects tempo and control, and card rarity often influences power but not necessarily effectiveness. By treating this as a historical snapshot, the analysis highlights broader patterns that continue to shape competitive play. The project also incorporates predictive modeling to estimate the likelihood of deck winning based on measurable features such as elixir cost, card types, and card statistics. The goal is to understand both *why* certain cards work well together and *how*

II. Datasets

A. Source of dataset (Heading 2)

The dataset used in this study originates from a publicly available Kaggle dataset titled *"Clash Royale Dataset"*. This dataset was last updated approximately eight years ago and captures card statistics from a history version of Clash Royale. While more recent, deck-level datasets were not available, this dataset provides detailed card attributes that allow for an analysis of how individual card properties may influence larger deck patterns.

No additional datasets or external API pulls were incorporated for this report. All analysis is based solely on the Kaggle dataset in its original form.

B. Character of the datasets

The dataset contains detailed attributes for a wide range of Clash Royale cards, including troops, spells, buildings, and spawner units. Each row corresponds to a single card and includes characteristics relevant to both gameplay strategy and statistic modeling. These attributes include *Elixir Cost, Card Type (Troops, Damaging Spells, Spawners, Defenses), Damage, Damage Per Second, Death Damage, Hitpoints, Shield Health, Spawn Health, Hit Speed, Spawn Speed, Range, Radius, Card Level and Spawn Level, and Maximum Spawned/Troops Spawned Counts*. These features provide a structural understanding of how cards behave in-game. Although the dataset does not directly include match outcomes or deck lists, the card-level statistics support higher-level analysis such as identifying what types of cards tend to be more powerful, which features correlate strongly with performance metrics, and how these characteristics may relate to deck-building strategy.

Because the dataset reflects a retired version of the game, some values no longer match the current live balances. This limitation is noted throughout the analysis.

III. Methodology

This study applies a combination of exploratory data analysis and predictive modeling to understand how card characteristics relate to overall deck performance. The methodology focuses on extracting meaningful patterns from the card-level attributes provided in the dataset.

A. Exploratory Analysis

Initial exploration includes computing descriptive statistics for elixir cost, damage values, hitpoints, and other numeric attributes. A usage frequency analysis identifies which types of cards (e.g., high-damage troops, low-cost cycle cards, defensive buildings) appear most commonly in the dataset and how their properties cluster.

B. Correlation Analsysis

A correlation heatmap is generated to evaluate the relationships between key numerical features such as damage, DPS, hitpoints, spawn rate, and elixir cost. This step highlights which attributes tend to co-vary and may influence strategic value. Strong correlations provide insight to broad card design patterns. For example, whether higher elixir cards consistently offer greater DPS

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

or whether defensive units cluster around certain health or radius values.

C. Feature Engineering

To support predictive modeling, the study constructs composite features that represent deck-building considerations including *Average Elixir Cost, Distribution of Card Types within a Deck, Total Damage Output (aggregated from individual cards), and Total hitpoints or Estimated Tankiness*.

These engineered features allow the model to approximate deck-level behavior despite working with card-based data.

D. Predictive Modeling

A logistic regression model is used to estimate the probability of winning a match based on deck composition. The dependent variable is the binary match outcome (win = 1, loss = 0). Independent variables include engineered features such as average elixir, card rarity, counts, card type distribution, and performance-related statistics derived from the dataset.

Logistic regression is chosen for its undestandability and suitability for binary classification problems. Coefficients from the model provide insight into which deck attributes most strongly influence match success. The model's limitations, particularly the lack of modern balance data, are acknowledged, but the process still demonstrates how predictive methods can be applied ot gameplay strategy.

IV. Results

A. The first set of results looks at the distributions of the generated deck characteristics. The average elixir histogram showed a smooth spread centered around roughly 3.5-4.5, which makes sense because the card costs in the dataset naturally cluster around mid-range values. This confirms the deck generator produced a balanced variety of elixir profiles instead of skewing heavily toward cheap or expensive decks.

The other aggregate stats (total damage, total DPS, total health) also showed wide distributions with no obvious errors, which means the numeric cleaning steps (removing commas and coercing to numeric) worked correctly.

B. The *num_troops vs. win* plot showed a noisy but noticeable trend: decks with more troop cards tended to have slightly higher win rates. The noise comes from the probabilistic win formula, which intentionally injects randomness so the data isn't perfectly linear or trivial.

The *total_damage vs. total_dps* scatterplot came out as a thick diagonal cloud – expected, since high-damage cards usually also have higher DPS. There were no strange outliers or broken values.

Most importantly, the synthetic *win* variable produced a spread of 0s and 1s instead of collapsing into one class. That means the win-probability formula is doing its job.

C. The logistic regression classifier reached an accuracy around what you would expect from noisy synthetic data. It was good enough t oshow patterns, but not so high that it looks fake or overfitted.

The conclusion matrix showed that the model could reliably separate higher probability win decks from lower ones without predicting only one class.

The feature importance bar chart gave a clear ranking of which deck features the model used most:

- **Average elixir cost** had one of the strongest effects, negatively correlated with winning. Heavy decks tend to lose more.
- **Number of troops** had a positive weight, reinforcing what the scatterplot suggested
- **Total DPS** and **total health** contributed moderately
- **Number of spells** had a small negative effect, reflecting the penalty built into the synthetic win condition

Overall, the model recovered the exact patterns that were intentionally baked into the win generator. That was the whole point, proving the pipeline works end-to-end.

V. Discussion

The results make sense, but they also show the limits of working with synthetic data. The trends (like cheaper decks doing better) match the logic we used when calculating win probability, so the model is not discovering real gameplay strategy but instead confirming that the generator behaves consistently.

The noise in the *num_troops vs. win* scatterplot is expected. The win function mixes randomness with a few weighted factors, so the graph won't clean separate. If anything, that randomness makes the dataset more realistic than a perfectly deterministic rule.

There are a few ways the project could be improved in a future version:

- Use real match outcome data instead of synthetic wins. That would give actual competitive insight.

- Add synergy features, like whether the deck contains win conditions, cycle cards, or splash support.
- Model interactions, for example the combined effect of low elixir + high DPS instead of treating everything independently.

But for the scope of this assignment, data cleaning, feature engineering, exploratory analysis, and a basic predictive model, the pipeline is complete.

VI. Conclusion

This project built a full analysis pipeline for exploring how Clash Royale deck attributes relate to match outcomes. After cleaning and standardizing the raw card dataset, I generated 800 randomized decks and calculated meaningful gameplay features like elixir cost, damage output, and card type counts. A probabilistic win function introduced realistic variation, which allowed for genuine patterns to emerge.

The logistic regression model correctly identified the drivers of success baked into the synthetic data: cheaper decks perform better, troop-heavy decks win more often, and spell-heavy decks tend to underperform. Even though the results are not based on live gameplay, the workflow demonstrates how deck-building data can be quantified and modeled.

In a real-world setting, such as balancing decision, esports analytics, or player-facing recommendation tools, this kind of analysis could help explain why certain deck styles win more often and how small changes in elixir cost or card composition shift win probability.

References

[1] S. Wappi, *Clash Royale Dataset.* Kaggle. Accessed: Nov 2025. [Online]. Available: https://www.kaggle.com/datasets/swappyk/clash-royale-dataset