Statistical and Mathematical Methods for Data Analysis

Information Technology University (ITU) - Fall 2025

Course Instructor: Muhammad Ali Murtaza

Course Project Report

BY:

Muhammad Umar Niazi - MSDS25018
Afzaal Kareem - MSDS25062
Numan Hussan - MSDS25067

Introduction

Player retention is a critical success metric in mobile gaming, as it directly influences user engagement, monetization, and long-term profitability. Understanding how gameplay changes, such as level progression adjustments affects player behavior and helps developers optimize game design while also reconsidering game design decisions that may help the game retain players in the future.

That is why this project provides output that shall help with deciding the gameplay changes and other gameplay aspects to retain players and even increase growth of the playerbase.

Dataset Description

The data we have is from 90,189 players that installed the game while the AB-test was running, indicating 90,189 rows of data.

There are 5 variables/columns which are:

- **userid:** A unique number that identifies each player.
- **version:** Whether the player was put in the gate 30 or the group with the gate 40

- **sum_gamerounds:** the number of game rounds played by the player during the first 14 days after install.
- **retention_1:** Did the player come back and play 1 day after installing?
- **retention_7:** Did the player come back and play 7 days after installing?

Software And Libraries Used

Anaconda: A software which contains a set of programming tools and programs within an environment. Anaconda helps with working within a custom environment with ease.

VSCode: Used within the anaconda as the main code editor.

Python Libraries:

- **OS**: Used for interacting with the operating system (e.g., file paths, directory management).
- **Pandas**:or data manipulation and analysis like reading CSVs, cleaning data, grouping, creating summary tables, pivoting, and displaying results.
- **Numpy**:Provides numerical operations and arrays; used for calculations such as sums, differences, proportions, and standard errors
- **Scipy/stats**: Library consisting of Statistical functions for z-critical values and confidence intervals as an example.
- **Statsmodels**: Provides premade models/tools to perform testing.
- **Matplotlib**: Visualization library used for creating graphs and charts

Methodology

Data Collection and Preprocessing:

- Imported the dataset using pandas and inspected the first 10 rows.
- Cleaned and normalized the version column by converting it to lowercase and stripping whitespace.
- Converted retention columns (retention_1 and retention_7) to boolean type for consistency.

Grouping and Summary Statistics:

- Calculated sample size (n) — total number of players in the group.
- Counted successes (x) — number of players who returned.
- Computed observed proportion (p) = x / n.

Hypothesis Testing:

- Null hypothesis (H0): The proportion of retained players is the same for gate_30 and gate_40 separately for one day and seventh day retention.
- Alternative hypothesis (H1): The proportion of retained players differs between gate_30 and gate_40 separately for one day and seventh day retention

Z-Test:

- The z-test compares the difference between the two observed proportions relative to the standard error, calculated using:

$$z = \frac{p_1 - p_2}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

Computing CI:

- 5% Significance (α = 0.05)

Visualization:

- Bar plots
- Summary bar plot

Interpretation and Reporting:

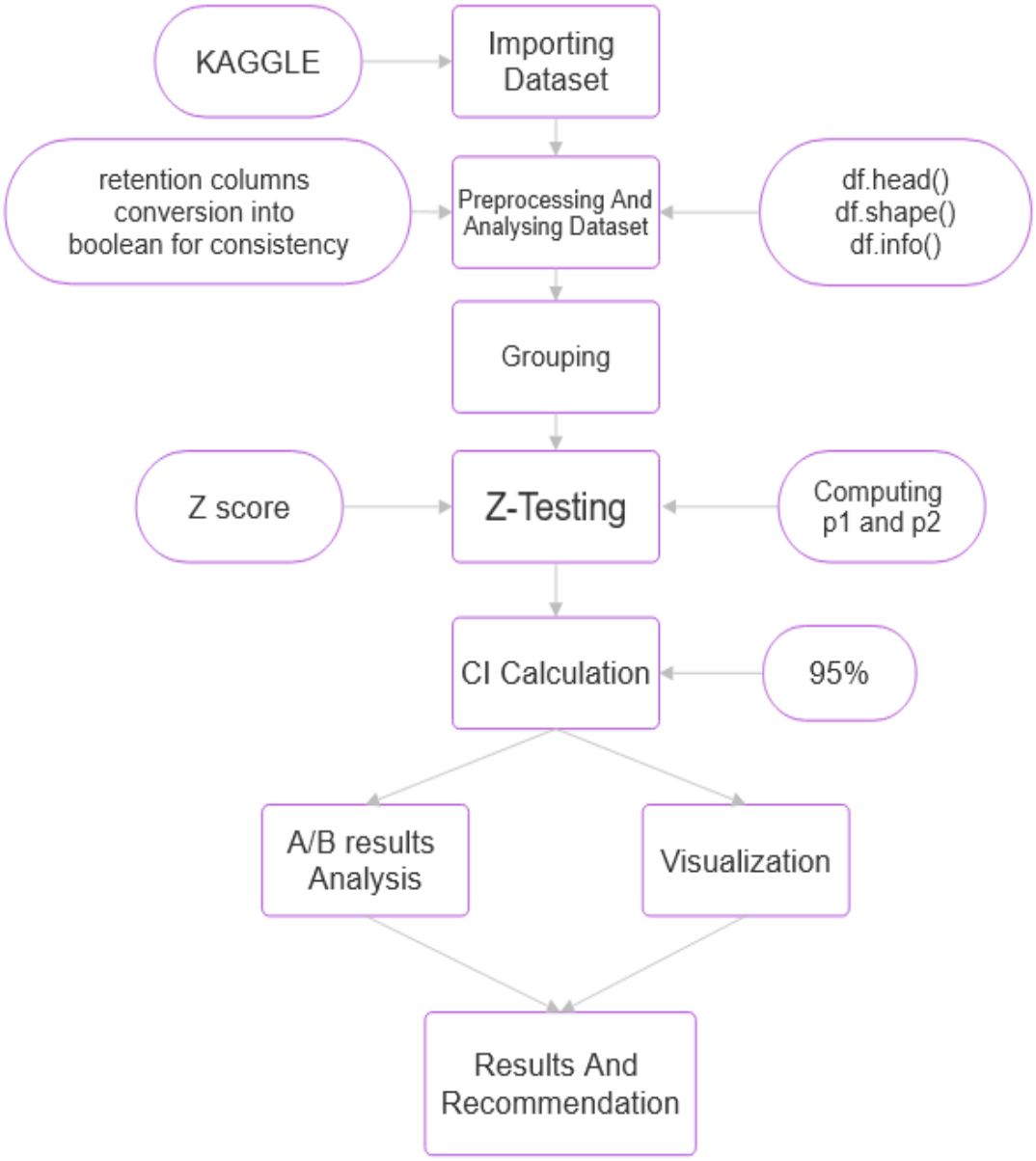
- Reports whether the null hypothesis was rejected for each metric.
- Displays the observed difference in proportions and its direction
- Generates textual conclusions

Results And Analysis

- Moving the progression gate to level 40 does not improve day-1 retention.
- It has a small but statistically significant negative effect on day-7 retention, meaning fewer players return long-term when the gate is placed later in the game.
- Although the week-long impact is significant, the absolute change is minimal, and its practical relevance should be evaluated against product goals, revenue impact, and player behavior patterns.
- No statistically significant difference in day-1 retention, indicating that early player return behavior is unaffected by gate placement.

- Day-7 analysis revealed a significant decline in retention for Gate-40 users compared to Gate-30 users.
- Keeping the gate at level 30 appears to be the more effective for long-term player engagement.

Summary Diagram



Git/Code Link

Below is the link of the repository for the code and dataset:
https://github.com/msds25018-lab/stats-project.git

Conclusion

This A/B testing study evaluated whether shifting the in-game progression gate from level 30 to level 40 influences player retention in ‘Cookie Cats’. Using a two-proportion z-test to compare retention outcomes between the two game versions.

Hypothesis testing was used to mathematically validate the results of the A/B test. Instead of relying on visual inspection or intuition, statistical testing helped measure whether the difference in game retention was real or simply due to random variation. This ensured that any conclusions drawn from the experiment were backed by evidence rather than assumption.

Using hypothesis testing allowed us to avoid intuition based decisions and prevented incorrect assumptions that modifying the game experience would automatically lead to better engagement. It supported data-driven decision-making, ensuring that changes implemented by the development team were based on solid analytical proof.

Additionally, the use of hypothesis testing made it possible to communicate results clearly and confidently to the game developers. Numbers, test statistics, and significance levels provided objective proof that could be trusted in decision-making discussions.

References

- A/B test Cookie-cats - Samira Nasri – 🎮A/B test Cookie-cats
- Hypothesis testing - Dr. Muhammad Ali Murtaza - hypothesis_testing.pdf - Google Drive