Report: R Analysis of *Grand Data Auto ViceCity*

BINF5003 Data Mining, Modeling and Biostatistics

Professor Arastonejad

Brenda Soljic, Sam Lenet, Abrar Faruque

November 13, 2025

Dataset Identification

The selected dataset for this project is titled *Grand Data Auto ViceCity*, published by *Eshum Malik (2023)* on *Kaggle*, and is available here https://www.kaggle.com/datasets/eshummalik/grand-data-auto-vicecity . The dataset originates from an open-world gaming environment, containing structured data about reviews from the *Grand Theft Auto* games. It provides a rich opportunity to analyze engagement and behavioural patterns with online gaming reviews, featuring numerous variables that are well-suited for R-based analytical tools due to the dataset's structured, data-rich contents with no missing or duplicate values.

Context of the Dataset

The *Grand Data Auto ViceCity* dataset was sourced from *Kaggle*, an online data-sharing platform that hosts community-contributed datasets. This specific dataset was created and shared by *Eshum Malik* in 2023. It compiles publicly available data about user reviews of the video game *Grand Theft Auto: Vice City*, reflecting patterns of user engagement and review sentiment, as well as voting behaviour within the games community.

The tubular dataset is provided in CSV, where each row represents an individual review, and each column contains a measurable feature. The features provided in this data set include both quantitative and qualitative metrics. Quantitative features in this data set are: the date review was made, whether the review was positive or not, how many upvotes the review received, how many comments the review received, if the reviewer purchased the game via steam, if the reviewer received the game for free as well as metrics for the reviewer themselves (such as hours played, number of games owned, etc.). Qualitative features include the unique reviewer ID and the content of the review. A notable issue with this dataset is the fact that some (298) reviews are blank (null), though this won't be an issue for this analysis; this may pose an issue for future attempts at analyzing the dataset. Due to the subjective nature of each review, determining how to score them will be the most challenging, some ways to mitigate this include using unique word count as a metric for 'effort' as well as using the reviewer metrics themselves, with the assumption that reviewers with more games and more hours invested in reviewed games will give a better review.

Due to its structure, the dataset is compatible with R and its standards in data handling libraries, such as dplyr, for example, for exploration and visualization.

Original Analysis and Anticipated Contribution

There is no documented original analysis of this dataset for the type of player base in the GTA games. This means that this is a raw, unexplored dataset without any interpretations in regards to the player-base as a whole. The absence of prior conclusions presents an opportunity to perform an independent and novel analysis using R, allowing this project to take a market-research-oriented approach.

This analysis will aim to uncover patterns in reviewer behaviour that can inform audience segmentation and engagement strategies for game developers and marketers. This could be done by classifying reviewers into distinct phenotypic groups, investigating correlations between engagement and sentiment metrics, and assessing if temporal patterns or archetypes can predict engagement. The analysis of *Grand Data Auto ViceCity* will contribute to a broader understanding of gaming communities and user behaviour by translating this data into quantifiable metrics.

Ultimately, the project aims to bridge data analytics and behavioural insights, transforming raw community feedback into structured information by applying R’s analytical and visualization tools, that support data-driven decision making for game marketers and makers.