STAT 184 Course Project

Checkpoint 1 – Due Wednesday, June 25.

**You need to answer these. Generative AI should not be used to answer the following.**

## Describe the research questions you will answer.

### You need to clearly identify at least 3 questions you have about the data sets you have chosen. You also need to explain why answering this is worthwhile. Outline how you will answer these questions using the data sets you have chosen.

Does the price of a game influence the review score of a game?

Does playtime influence whether the game is recommended by the reviewer?

Do players find reviewers with more playtime more helpful?

What are the highest reviewed paid games, and their prices?

What are the highest reviewed free games?

What is the optimal price point for paid games?

How does early access effect reviews?

Answering these questions is important as it could provide a consensus of player sentiment towards price of games. This could also show how willing players are to trust reviewers that have played the game for more time.

I will answer these questions by looking at the data of the two sets and plotting the relevant columns and rows for each question.

## Describe the primary data set you have chosen.

### Where did you access the data? What are rows and columns of this data set? Will you need to clean or tidy these data? What challenges do you anticipate arising when working with these data?

The dataset is a compilation of games on the game marketplace Steam. It contains information such as game name, price, review percentage, number of review, release date, and tags.

The primary dataset comes from the r/gamedev subreddit, posted by user Sadaris. <https://www.reddit.com/r/gamedev/comments/x0qs4z/we_gathered_data_about_54000_games_in_steam_and/>

This dataset will probably need to be cleaned as there may be data that I do not need. The main challenge of working with this data will be the number of entries.

## Include below a screenshot (or a link) of the primary data that you will use for this project.

https://docs.google.com/spreadsheets/d/1D5MErWbFJ2Gsde9QxJ\_HNMltKfF6fHCYdv4OQpXdnZ4/edit?gid=352042934#gid=352042934

## Describe the secondary data sets you have chosen.

### Note, you need to have at least 1 of these.

This dataset is a is a compilation of reviews of games on Steam. It includes game name, text of the review, review score (positive or negative), and review votes. This comes from kaggle by ILJUSSA. This data will need to be cleaned as it is simplistic. The difficulty of this data will be in combining it to compare with the primary dataset.

https://www.kaggle.com/datasets/iljussa/steam-reviews