**David Hatchett**

**DSC540 – Data Preparation**

**Term Project Milestone 1**

**Project Subject Area:**

I wanted to do something a little more light-hearted and fun for this project. A passing side interest of mine is video game speed runs. Speed runs are a challenge players take on to compile a video in the fastest possible manner.

**Data Sources:**

* **Flat File Source:**
  + **Name:** Video Game Playtimes
  + **URL:** <https://www.kaggle.com/datasets/baraazaid/how-long-to-beat-video-games>
  + **Description:** The dataset contains a list of video games that contains how long they usually take to beat. This data is taken from pools and contains categories like 100% (do everything), Main Story (Bare minimum), and overall stats like average, minimum, and max playtimes.
* **Web Site Source:**
  + **Name:** Games Done Quick Tracker Site
  + **URL:** <https://tracker.gamesdonequick.com/tracker/>
  + **Description:** This is run by the charity group Games Done Quick, which holds several yearly events focused on speed running and raising money for charity. I plan to use one of the following datasets: Bids or Runs. The Bids dataset contains records on specific games and the money collected for specific incentives. Runs are a list of runs and when they occurred during the event. I’ve inspected the site, and the data elements are available in the raw html.
* **API Source:**
  + **Name:** SpeedRuns.com API
  + **URL:** <https://github.com/speedruncomorg/api/tree/master/version1>
  + **Description:** SpeedRuns.com is a popular site for tracking and sharing speed runs. They kindly provide an API that can be used to gather data and submit data. I’m planning to extract data on popular speed runs.

**Data Relationships:**

There is one primary and at least one secondary relationship between all the datasets. The primary relationship between all data sets is a specific video game. While I’m not sure of the exact execution, I do know that I will need to clean, transform, and potentially generate data for all the relationships to work correctly. I guess this will be cleaning up strings or creating my unique IDs.

The secondary relationship is time. The flat file I chose was created in 2022. I may wish to restrict the data from the other two sources to around the time.

**Project Pan/Challenges/Ethical Implications:**

Each dataset will have unique issues, most of which will not be apparent until loaded and reviewed. Following the suggested outline for each milestone is a good idea as it will give me time to identify the nuance of each data set.

As stated in the relationships, making the game relationship strong will need to be at the forefront of thought. At this time, I believe it would be best to leverage the ID already generated from the speedrun.com API. This means we will likely need to use the API several times during the project to search and match a game name. It also might be a good idea to extract all the games from the API and store them; however, I’m unsure which path would work best.

The other major challenge I foresee is scraping the website. I haven’t done this in a while, and I remember that cleaning out the unneeded elements can be a pain. I’m also pulling from a site with more complex tables than typical Wikipedia articles.

I am also thinking about using DuckDB for the file step in the project, as I’ve heard many good things about it. I also like how it can run regular SQL statements on top of Pandas data frames.

As of right now, there are not any ethical concerns to worry about. Some specific concerns may arise throughout the analysis, but if you find that unlikely. The results would have to provide some unfair advantage to players to be of concern.