DS3000 Data Analysis Project

July 15, 2024

1 Introduction

Palworld is an exciting new open-world survival crafting game developed by Pocketpair. The game features a diverse array of creatures known as Pals, which players can befriend, capture, and utilize in various ways. Players can explore vast landscapes, build structures, and engage in battles with their Pals. Each Pal has unique attributes, abilities, and elemental affinities, making the gameplay dynamic and strategic. The game also includes elements of farming, breeding, and cooperative multiplayer modes, enhancing the overall experience. For this project, students will analyze data from PALDB and PALWORLD to gain insights into the characteristics and capabilities of these Pals.

2 Project steps

2.1 Data Sources

The data for this project will be sourced from the following websites:

- PALDB - A comprehensive database for PAL world information.
- PALWORLD - The official site for PAL World game updates and information.

Both website are available for scrapping and there are a lot of information there. The first step is to discuss with your group members and decide which features you may be interested and what specific questions you want to answer in your project. This part will be in the homework 5.

You will need to scrape at least 8 features to for the project. You can use other websites if you have found better data, but please label those in your work and save all the data in local file.

2.2 Data Cleaning

The data cleaning process should include:

- Handling missing values.
- Correcting data types.
- Removing duplicates.
- Standardizing formats (e.g., units of measurement, capitalization).

and any other things you need to do before the analysis.

2.3 Data Visualizations

Students are required to create three to four different visualizations prior to the modeling. Possible options include:

- Bar charts to compare the attack power of different Pals.
- Scatter plots to visualize the relationship between speed and attack power.
- Histograms to show the distribution of health points (HP) among Pals.

2.4 Model Fitting

Two models need to be fitted to the data to answer specific research questions. Suggested models include:

- **Linear Regression** - To predict the attack power based on other variables like speed, defense, and HP.
- **Clustering (e.g., K-Means)** - To group Pals into clusters based on their attributes and identify patterns.

The goal of this project is to provide students with hands-on experience in data scraping, cleaning, visualization, and modeling. By the end of the project, students should be able to draw meaningful insights and present their findings in a coherent manner. Template will be provided when we move to the later part of the project.