Kelley Denny Student ID 800563382

Intro to Data Mining ITSC 3162

Project 1

For this project, I am looking at a Genshin Impact Characters Dataset. The dataset can be obtained here. Genshin Impact is a free to play open world ARPG (action role-playing game) that is set in a fictional world where the player travels with a party of characters and can switch between whichever character they wish that is in the party. The player gains characters into the party by playing the game and obtaining characters (Wikipedia). The player can collect up to 8 characters totally for free, and the rest of the characters have to be obtained through using in-game currency that can be gotten by playing the game or through purchasing with real money (XP-Pen). I am interested in testing the balance of the game between free tier players and paying players, as the free players can only obtain up to 8 characters just through game play alone.

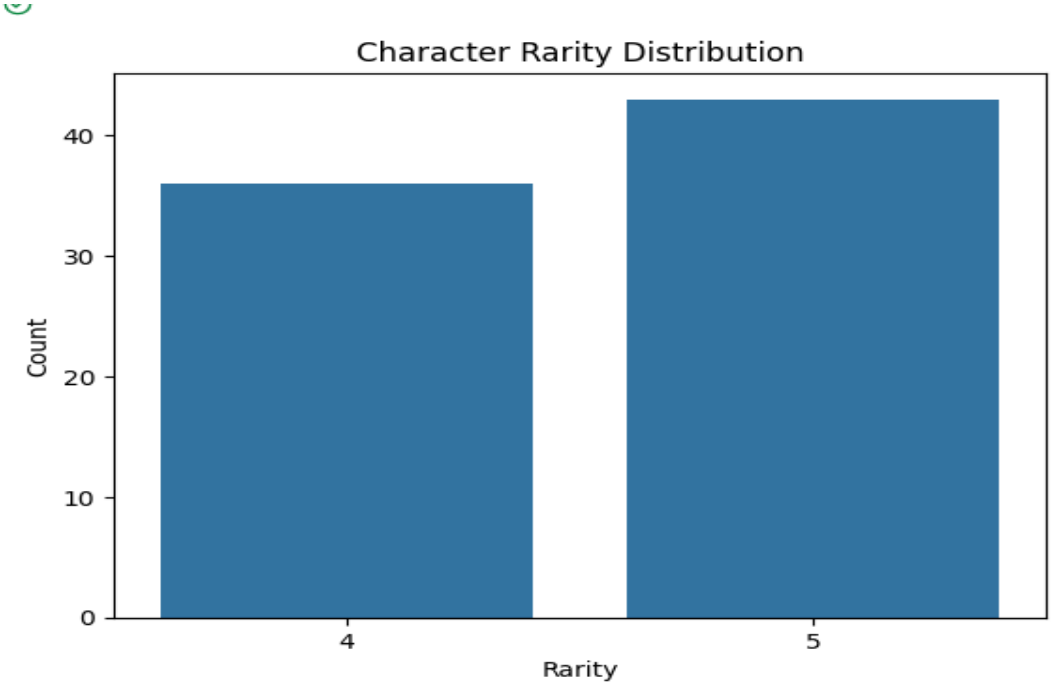
For some information about the dataset itself, it contains 81 columns and 84 unique rows (one for each character in the game). Some of the features (columns) include character rarity, weapon type, player model, region obtained, health points/attack power/defense power stats at the different level milestones for the character. The dataset appears to be fairly clean, but there are some missing values. Every row is unique, and a majority of them are integer dtypes.

Since I am interested in the balance of the game, I want to look at the relationship between the base level stats of HP, Attack, and Defense. This study hypothesizes that the character stats (HP, ATK, DEF) at base level will be higher if the character rarity is higher. To look at this relationship this study found the pertinent features of the data set to focus on to be “rarity” as the independent variable and “atk_1_20”, “def_1_20”, and “hp_1_20” as the dependent variables.

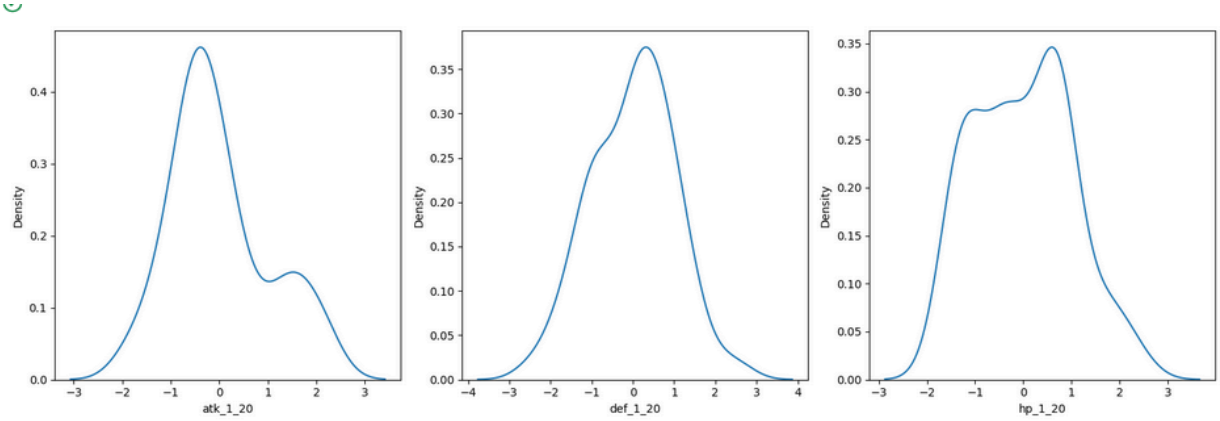
At first glance this data set appeared to be clean already but as a rule of thumb the steps for cleaning a data set include: looking for missing values (as those may skew the data or just not allow the dataset to be representative of the population in question) and deciding how to handle them if they are present, checking for duplicate values (for the same reason as checking for missing values) and handling those if present, converting data types to a consistent type (so that different features can be compared), checking for any outliers or anomalies in the dataset and normalizing those values (so that they don’t alter the results of the analysis), scaling or standardizing the observations in the dataset (so that things with a high range like health points can still be compared to things that could have a much lower range like defense points) for a streamlined analysis.

Once the data is clean, I want to first look at the distribution of the rarity levels. There are two rarity levels for the 84 different characters that can be obtained in Genshin Impact, 4-star and 5-star Rarity levels. There are more 5-star characters than there are 4-star characters as

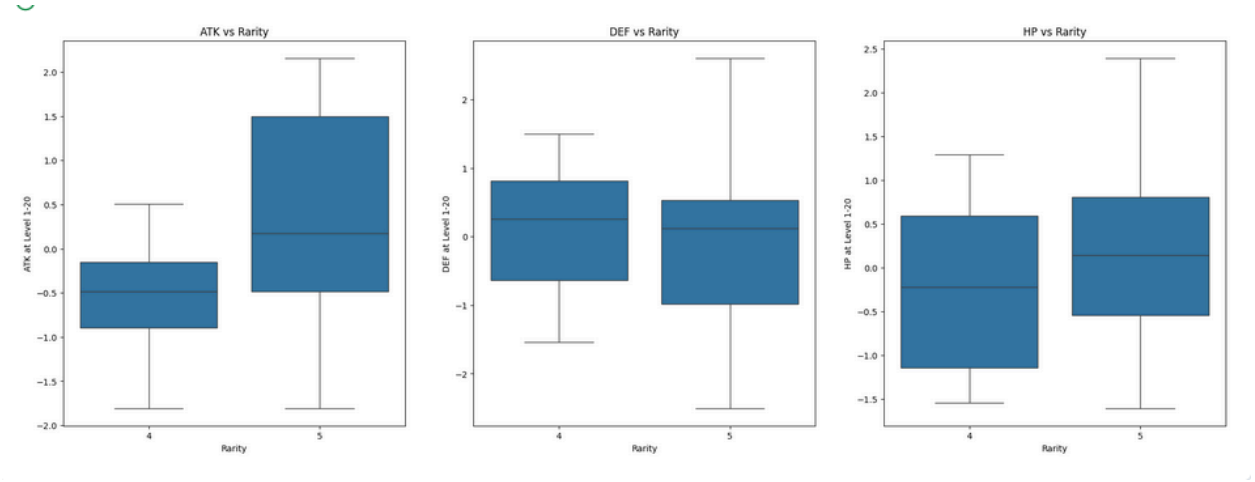
depicted in the count plot visualization. Since the player can't obtain any 5-star characters for free without a large time investment in the game ("farming"), this puts a lot more of the game behind a paywall essentially.



Next I want to see the distribution of the base level stats of the characters to see if they are pretty normally distributed. I chose to do a kernel density plot to see the distribution of the base level stats using a continuous curve instead of a histogram (Seaborn kdeplot). The stats were all pretty well distributed with a slight skew on both attack and health points. The defense stats had a larger spread than the other two stats but they were all mostly equally distributed. Here's what it looked like:



Finally I want to look at a box and whisker plot to help me visualize each of the 3 base level stats in comparison between the two rarity levels. This type of plot is a good way to see a side by side comparison of the ranges of min and max stats of each type of rarity. I noticed right away that attack stats were quite a bit higher in the 5-star rarity characters, but the defense and hp stats were closer together. My takeaway is that the most powerful fighting characters are the harder to obtain characters, enticing players to use their real money to buy a better or more powerful character to have in their party. Here is what the plot looked like:



My main takeaway from this analysis is that the higher rarity characters are more desirable for players who want to have less challenge when fighting enemies in the game, thus pushing these players to spend their money on the game despite it being a free-to-play game which could be a reason the player chose the game to begin with. The game does seem to have some balance issues and the hypothesis was correct for attack stats especially but also for health stats. Defense stats actually ended up being potentially higher in the lower rarity characters though.

The dataset is for a free to play game, but the impact of the data analysis might help to see a correlation between rarer characters and higher/better stats. This could help raise some discussions around the need for character/game balancing and fairness between the "pay-to-win" players and the non-paying players. Also, gaming communities use analysis studies to discuss the game in general, or to discuss the meta of the game. Having an analysis of the characters available in the game could help players to develop their strategies for which stats to prioritize as they level and explore the world, and perhaps even which regions to explore and farm to obtain some of the more desirable characters.

References

"How to Get Characters in Genshin Impact[a Full Guide]." *XPPen*, www.xp-pen.com/blog/how-to-get-characters-in-genshin-impact.html. Accessed 14 Sept. 2024.

"Genshin Impact." *Wikipedia*, Wikimedia Foundation, 14 Sept. 2024, en.wikipedia.org/wiki/Genshin_Impact.

"Seaborn.Kdeplot#." *Seaborn.Kdeplot - Seaborn 0.13.2 Documentation*, seaborn.pydata.org/generated/seaborn.kdeplot.html. Accessed 14 Sept. 2024.