Data Mining: Project 1

Akash Shanmugam

9/9/24

Dataset link:

https://www.kaggle.com/datasets/n2cholas/competitive-pokemon-dataset/data

Data Introduction:

The dataset for this project is called the “Complete Competitive Pokemon Dataset” which in total contains 2 .csv files for analysis of the competitive Pokemon information regarding both moveset data and pokemon data for use of analyzing the connection between moves and pokemon’s attributes and stats. The questions about this project involve whether the move data of pokemon has changed in a meaningful way across the generations listed in the data as well as how moves are categorized and distributed in general. All of the data was obtained from sources around 2018 which means that the latest data the dataset contains is from the generation 7 games (Pokemom Sun, Moon, Ultra Sun, and Ultra Moon) and all further additions and retroactive changes will not be reflected in the sets that were used. For the purposes of the project, only the move dataset will be used to reflect in the visualizations shown.

Pre-Processing the Data:

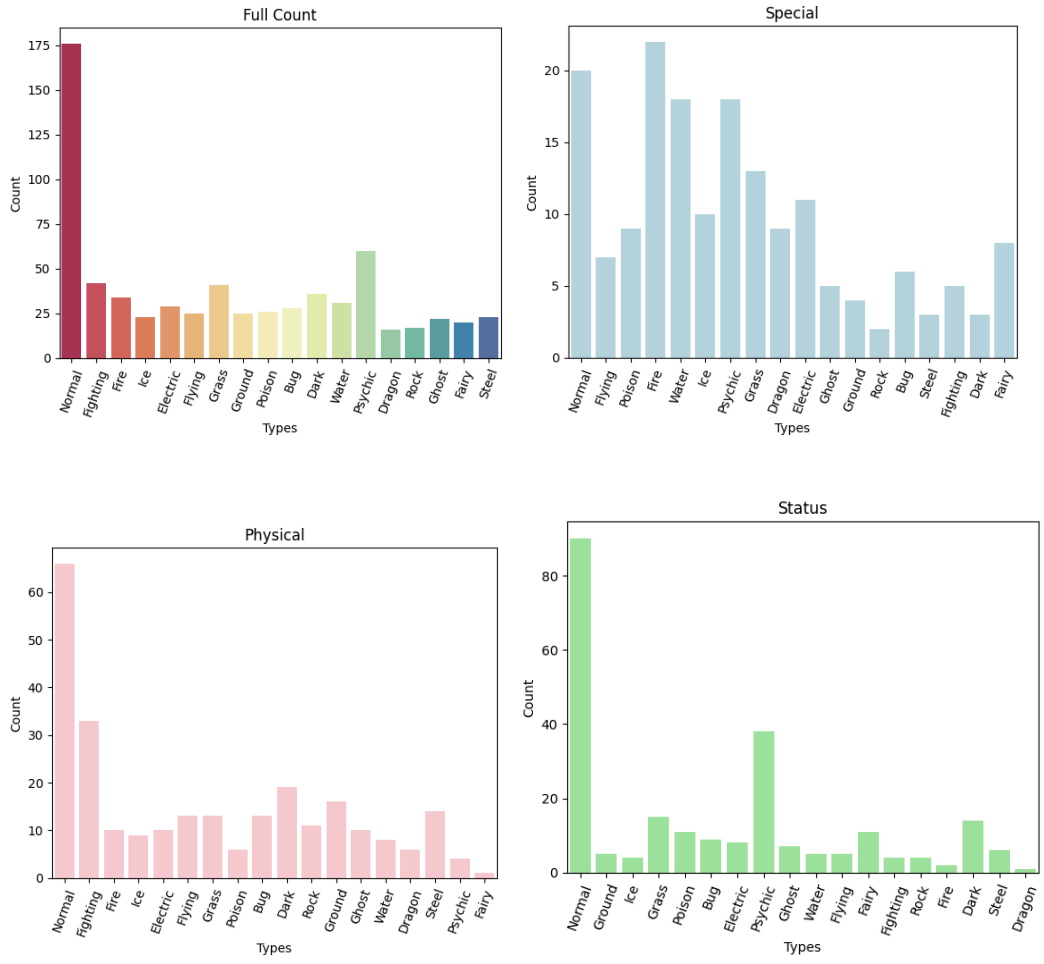
The data in the move-data.csv file was not taken from something like a survey but instead was taken using the actual move data contained in the pokemon games. This means that the amount of null values is essentially zero however I dropped possible ones anyway resulting in no change which meant I could move onto the next step which involved dropping the moves which did no damage and were listed as “None” under the power column. With this I was finished with the clean up.

```
import matplotlib.pyplot as plt
df.dropna(how='any')
df.loc[df['Power'] == "None"]
print('Number of moves: ', len(df))
df.sample()
df.columns = ['index', 'name', 'type', 'category', 'contest', 'pp', 'power', 'accuracy', 'g
df.set_index('index')
df['power'].replace('None', 0, inplace=True)
df['accuracy'].replace('None', 100, inplace=True)
df['power'] = pd.to_numeric(df['power'])
df['accuracy'] = pd.to_numeric(df['accuracy'])

df = df[(df.pp != 1) | (df.name == 'Sketch')]

df.sample()
```

Vis 1: The Count of Moves across Categories

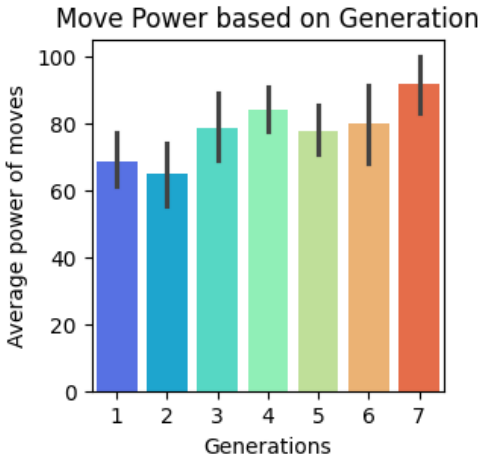
The visualizations for this specific project needed to be created in terms of breaking down how moves can be categorized but first is the number of moves themselves. This figure shows a plot of the count of the types of moves across the whole dataset and shows how the moves collected in the data have a massive tendency to be normal type across the board. The types behind normal tend to have a more consistent distribution with psychic being the highest that’s not an outlier. Dragon and rock being the ones that have the fewest amount of moves total. This fits with the fact that normal type moves are generally accessible to almost every pokemon and thus have

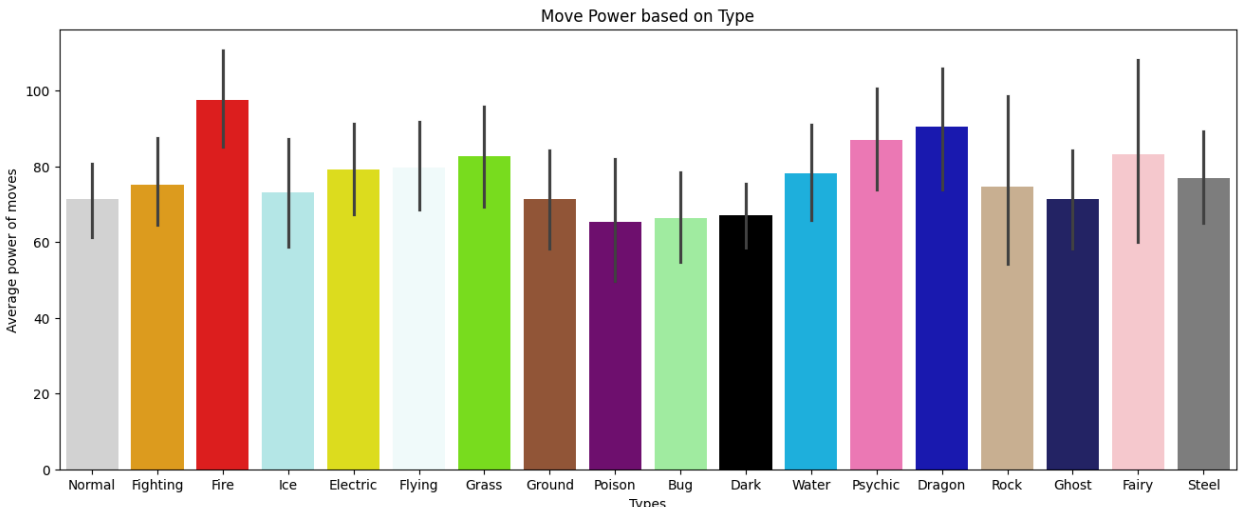
more in general. The special countplot shows that special moves have the least amount of moves available as well as that fire the number one type of move in this category with ice, water, and fairy seeing a massive increase as well. The physical plot is somewhat similar to the full list and only contains fighting being the second most abundant type as its major change as well as ground and steel being higher on placement. The status type is the least balanced of the three subplots and shows a relatively similar distribution to the original plot and emphasizses psychic being a dominating factor of all categories but physical. The main changes for the physical and special countplots reflect how the types are associated with either category and prior to generation 3 all moves of a certain type had to fall into either of them.



Vis 2: Average Move Power based on Move traits

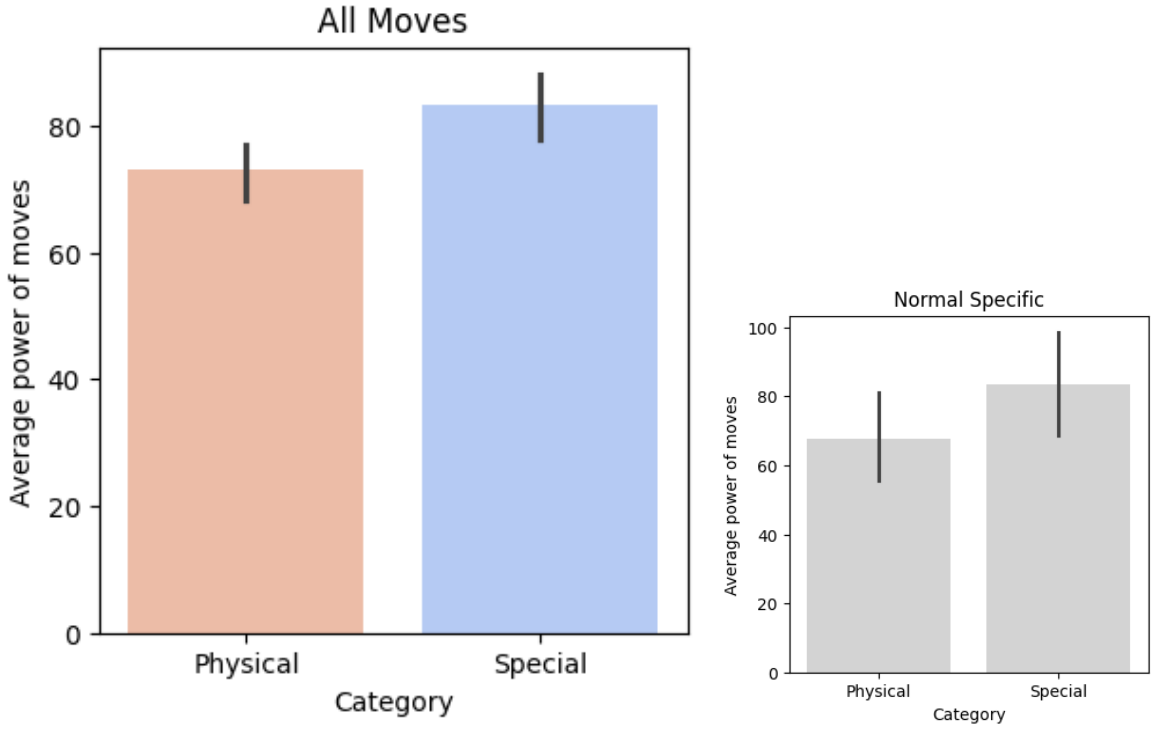
My next task was to see what the average base power of moves in each generation and type category accounted to. I used the data to plot a bar plot that showed what the average power of each move tended towards and this lead to some interesting results. First looking at the power based on generation, the base power of the generations overall seems to get slightly higher with noticeable variation each time, generation 2 being the weakest and 7 the strongest. In the type graph you can see the strongest type of move on average is fire type. Looking at the data on its own this makes sense as a lot of individual fire type moves have 150 base power which can contribute to how large the average is. The next thing I could find was that dragon type came in second place which makes sense as it is considered rarer and more powerful which is supported by the previous visualization. The fairy type was high as well but also had a larger variance which indicates that the fairy type doesn't have a majority of its moves in the area around the base power in the bar chart (around 80).





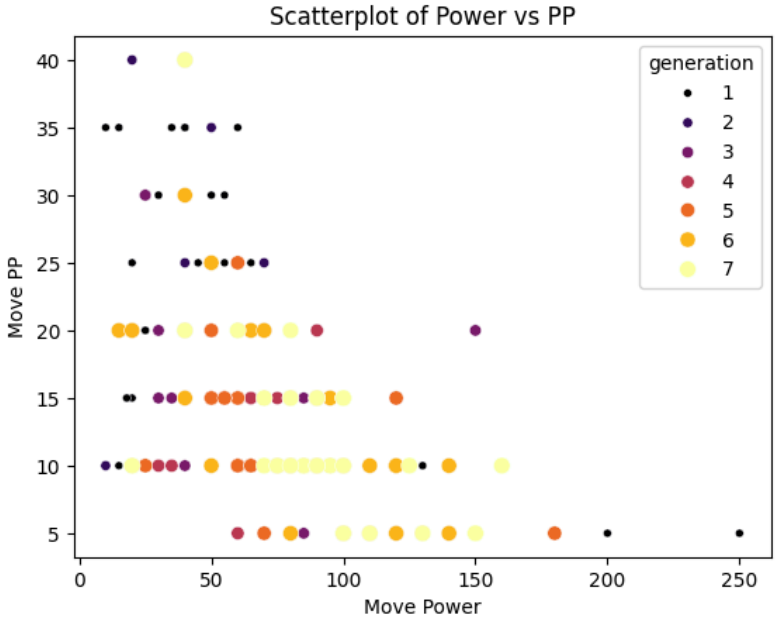
Vis 3: General Move Power by Special and Physical

The moves in Pokemon are separated into 3 categories for distinction. Those being: Physical, Special and Status. Given that status moves have no base power, they were removed from the bar plot dataset. With this visualization it is clear to see that special moves as a whole generally have more power than physical moves with a nearly identical amount of variation. This data could very much be explained by the fact that the most prevalent kind of move is normal type and that many physical normal type moves tend to be the type that are meant to be learned by pokemon early and therefore have lower base power.



Vis 4: Power vs PP across all Generations

The next figure represents how the PP value of moves correlates to the amount of base Power each move has. Looking at the visualization, there is a curve that can be seen going from the highest power to the highest PP. This makes sense given that lower power moves tend to have high PP to make sure the move can't be overused and utterly outclass any move weaker than it in base power with the same applying the other way around. With the generations being represented on the scatterplot an interesting correlation shows up where the earlier generation moves (especially 1) appear to have many low power data points with varying PP and generation 7 moves very much clustering on the 10 PP line and around 100 base power. This makes sense as a lot of the low power "basic moves" a pokemon will learn at low levels have been set in stone since the first generation. As pokemon passed in generations, a lot of the new moves created were done so to reflect the current metagame and not usurp balance which explains the later generations pooling around one area.



Impact:

The visualizations in this dataset seek to explain how the move data in the main series pokemon games have changed over the period of the series' existence as well as the distribution of the moves in the beginning and how they may trend for the future. The impact this project specifically can have is getting a better understanding of the design decisions the developers of pokemon make in order to balance the competitive metagame and keep the games interesting without ruining balance. This change in attitude while trying to maintain the essence of pokemon can be seen in how generation 1 moves had more PP or how the move power on average goes up each generation. The visualizations here can also be seen more competitively where if certain moves have more average base power, it would be favorable to both use those moves and use a pokemon that is the same type and has a better special or physical attack depending on what the type of move has more of.

References:

https://www.kaggle.com/datasets/n2cholas/competitive-pokemon-dataset/data

https://bulbapedia.bulbagarden.net/wiki/Move