DSCI510 Final Report

Name and Author: Yu-Chieh Chen

1. **Motivation/Rationale for the project: Describe the question you wanted to answer and why is interesting:**

The desire to learn more about the complex Pokémon universe is the driving force behind my project, "Pokémon Type Analysis and Predictive Modeling," which focuses on statistical analysis of Pokémon types, Mega Evolution predictive modeling, and Pokemon popularity based on their attributes.

The questions I would like to explore:
1. Statistical Analysis of Pokemon Types: Pokémon types are a key component of the game mechanics, influencing player choices, strategies, and battles. This is the subject of this statistical analysis. I want to find any notable patterns or variances between the average total base stats of the various Pokémon types. This approach is interesting because it can highlight difficulties with game design preferences, underlying balancing concerns, and possible player strategies. It considers how player experience and strategy are affected by game design, adding to the conversation on design philosophy and game balance.

2. Analyzing Pokemon popularity: My goal is to examine the relationship between the popularity of Pokémon, as measured by mentions on the Pokémon Fandom forum, and their in-game statistics, including Attack, HP, and Total numbers. This study is fascinating since it investigates if gameplay utility of Pokémon or other elements like lore and design influence fan preferences. Gaining an understanding of these dynamics can help marketers and game developers create more engaging characters and increase fan engagement in this wildly popular series by illuminating the factors that contribute to a Pokémon's attractiveness.

3. Predicting Mega Evolution: Mega Evolution is a transformative ability that only affects a subset of Pokémon, but it can drastically alter a player's power and strategy in competitive games. Determining which Pokémon can undergo Mega Evolution based on their attributes is an attempt to comprehend the fundamentals of what qualifies particular Pokémon for this kind of evolution. This feature is interesting because it combines data analytics with game mechanics to offer insights that may help guide fan theories and game strategy.

2. **Description of data sources: What data did you collect? How did you collect it? How many data samples did you collect?**

a. Specify exact data sources (e.g., URLs) and your approach to extract the data.
1. Pokémon API: I made use of the Pokémon API. It offers up-to-date, complete data on Pokémon attributes. The Pokémon species data, including properties like "name," "base_happiness," "capture_rate," "forms_switchable," "gender_rate," "has_gender_differences," "hatch_counter," "is_baby," "is_legendary," "is_mythical," and "order," may be easily extracted thanks to this API.

How I collect it: I created several functions to extract the data

- Fetching Specific Pokémon Species Data (get_pokemon_species_data):This function returns comprehensive data for a particular Pokémon species that has been given a name or ID. It makes sure we only collect the most pertinent data by choosing a predetermined set of attributes from the API response.

- Listing All Pokémon Species (get_all_pokemon_species_names): This obtains the names of every Pokémon species.

- Fetching and Consolidating Data (fetch_and_consolidate_pokemon_data):This function retrieves detailed information about each species using the list of species names. After that, it compiles the data into a pandas DataFrame for simpler analysis and archiving.

- Saving Data to CSV (save_data): Given a pandas DataFrame and a file path, this method saves the dataframe to a CSV file at the specified location.

- Command-Line Interface (main):It uses argparse to process command-line arguments, allowing the script to function in multiple modes based on user input. Additionally, it lets you scrape a certain number of entries (--scrape) or save the data you scrape to a file (--save).

I collected data on all 1,025 Pokémon species with attributes: "name," "base_happiness," "capture_rate," "forms_switchable," "gender_rate," "has_gender_differences," "hatch_counter," "is_baby," "is_legendary," "is_mythical," and "order."

2. Web Scraping of Pokemon fandom forum:The main source of information was the Pokémon Fandom forum. Fans debate anything related to Pokémon, including games, TV series, and general trivia, on this lively community forum.

How I collect it:
- I used a technique known as web scraping to get information from the forum. To traverse the forum, mimic page scrolls and clicks to load further content, and retrieve the page source, the specific function scrape_pokemon_fandom was utilized. In order to collect as much information as possible without overloading the website's server, the script was made to scrape the site for a maximum of seven minutes.

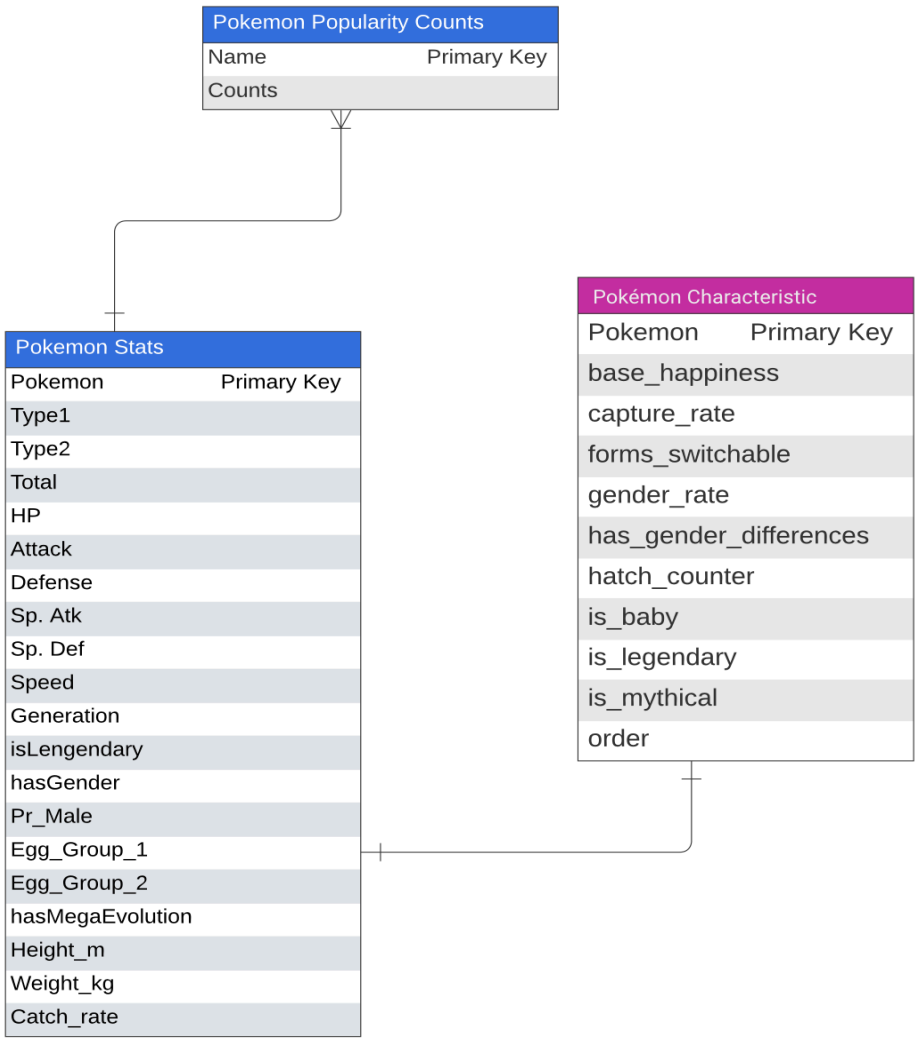
In particular, the script counts how many times each Pokémon name appeared in the scraped text from the list. The 451 distinct Pokémon names that were cited in the forum postings made up the entire dataset. Every name that was brought up was counted and documented, giving a thorough picture of Pokémon's popularity among the forum's users.

This approach made sure that a solid body of data was gathered, which reflected the conversations and interests of Pokémon enthusiasts today. This provided insightful information about the popularity trends of different Pokémon among the fan base.

3. Pokemon Statistical Dataset: This dataset contains information on 721 Pokémon, including their number, name, first and second type, and basic stats: HP, Attack, Defense, Special Attack, Special Defense, and Speed,etc. I directly downloaded it from Kaggle.

b. Describe what has been changed from your original plan, what challenges you encountered or resolved.
Originally, one of my web scraping sources was to gather data simply by scraping a Reddit post that listed the top 30 Pokémon. However, this approach rapidly became limited owing to its specific focus and lack of community participation. I asked a teaching assistant for guidance after running into problems with a narrow breadth of data and inadequate insights, and she suggested a more comprehensive strategy. I moved our data gathering to the Pokémon Fandom forum as a result. I use the Pokémon Fandom forum to gather and count the instances of Pokémon names in order to study Pokemon Popularity. In order to obtain a more comprehensive and precise estimate of popularity, I utilized Selenium WebDriver to dynamically scrape mentions of Pokémon names from the forum. Despite being technically challenging, this strategy helped me effectively manage dynamic material and enforce site usage guidelines. In the end, I generated a dataset with 451 distinct Pokémon names, providing a thorough understanding of their popularity to the community.

3. **Integrated Data Model: describe the data model and provide an informal entity-relationship diagram. You can reuse the one you provided in submission 2, but update as necessary if it changed in the final version of the project.**



- Inner Join between Pokémon Stats and Pokémon Characteristics:

The two datasets are combined according to the Pokémon Name with the intention of keeping only those Pokémon that are included in both databases. The combined columns from the two datasets provide matching_df, which displays Pokémon that are shared by both.

- Outer Join between Matching_df and Pokémon Popularity Counts:

By adding popularity data to the detailed attributes from matching_df, all Pokémon from matching_df are retained, and NaN is used to fill in any gaps in the popularity data.
Result: The final dataset containing information on each Pokémon's popularity, characteristics, and statistics.

4. **Analyses/Visualizations**
   a. Describe what analysis techniques you used.

Statistical Analysis of Pokemon Types (Question 1):
For the statistical analysis of Pokémon types, I utilized several techniques to assess the competitiveness and effectiveness of Pokémon types in battles:

- Average Calculation: Using pandas, I calculated the average total stats, defense stats, and attack stats for each Pokémon type. This approach helped in quantifying the overall strength, defensive capabilities, and offensive capabilities of Pokémon by type. The mean() function was applied to the respective columns after grouping the data by Pokémon type, which allowed for a concise summary of each type's performance metrics.
- Data Aggregation: After computing the averages, I aggregated these statistics into a new DataFrame. This DataFrame served as a consolidated view of the type-based performance metrics, facilitating easier comparison and analysis.
- Sorting: The aggregated DataFrame was sorted by the average total stats column to prioritize and highlight the types with the highest overall strength.

Analysis of Pokemon Popularity (Question 2):
Preparation of Popularity Subdataset:
To focus the analysis on relevant attributes that might influence Pokémon popularity, I created a subdataset named pop_df from the main dataset. This subdataset includes the following columns:

- Name: Pokémon's name for identification.
- Capture Rate: Reflects the difficulty of capturing the Pokémon.
- Forms Switchable: Indicates whether the Pokémon can switch between different forms.
- Total: Sum of all statistical attributes, representing overall strength.
- Is Legendary: Denotes if the Pokémon is legendary.
- Counts: Popularity counts from mentions on the Pokémon Fandom forum.
- Body Style: Describes the physical appearance or structure of the Pokémon.
- Is Mythical: Indicates if the Pokémon is considered mythical. Mythical Pokémon are typically acquired through unique distributions or occasions. They are regarded as extremely rare in the game world and frequently have special powers and high stats.
- Has Mega Evolution: Shows whether the Pokémon has the capability to Mega Evolve.

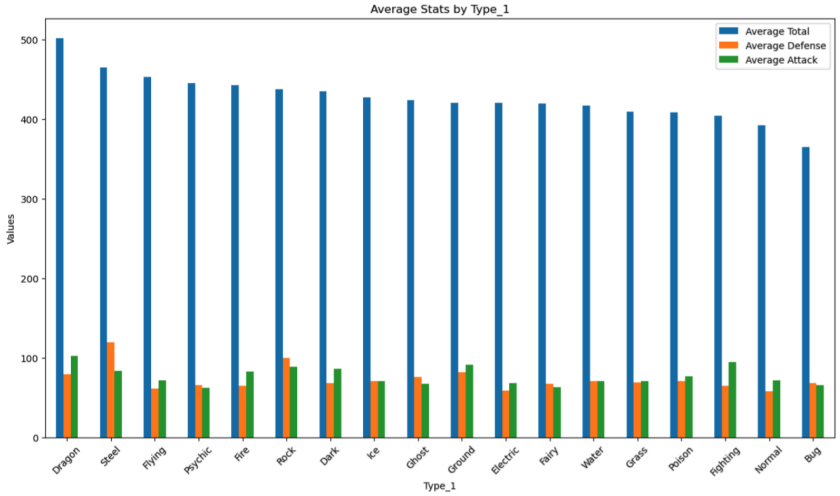
For analyzing Pokémon popularity, I explored the relationship between their mentions on the Pokémon Fandom forum and various in-game statistics and attributes:

- Correlation Analysis: I investigated how different attributes such as capture rate, total stats, and mythical status correlate with Pokémon popularity. This involved plotting these attributes against popularity counts to visually assess trends and relationships.
- Comparative Analysis: Using bar plots, I compared the average popularity across different Pokémon characteristics such as mythical status, mega evolution capability, switchable forms, legendary status, and body style. This comparison was crucial in identifying attributes that potentially influence popularity.

Predicting Mega Evolution(Question 3):
The analysis methods used for Question 3 include preprocessing the data, training the model, and utilizing a Random Forest classifier to evaluate the results. First, every categorical variable in the dataset is eliminated, and then the 'capture_rate' column is eliminated as well. Next, the data is divided into a target variable and characteristics, with the latter concentrating on a Pokémon's ability to Mega Evolve. To prepare for model training using the Random Forest, which is designed with 100 trees to provide robustness and prevent overfitting, the dataset is separated into training (80%) and testing (20%) sets. On the training set, the model is trained, and on the testing set, predictions are made.Ultimately, the precision, recall, and F1-score for the model's capacity to predict Pokémon's Mega Evolution are assessed, along with accuracy and other comprehensive metrics included in a classification report. This all-encompassing strategy successfully integrates performance assessment, machine learning, and data management to tackle Question 3's predictive modeling job.

b. Describe the figures you made, how you made them, and their elements and meaning.

Figures for Statistical Analysis of Pokemon Types (Question 1):
Bar Chart of Average Stats: Using the integrated Data Frame I made, I made a bar chart that shows the average total stats, defense, and attack for each type of Pokémon. The matplotlib and seaborn libraries were used to create this chart, which shows the three segments (attack, defense, and total) of each Pokémon type as bars. This allows for a clear visual difference of each kind's skills. To aid in reading and understanding, the chart has labeled axes, a title, and a legend.

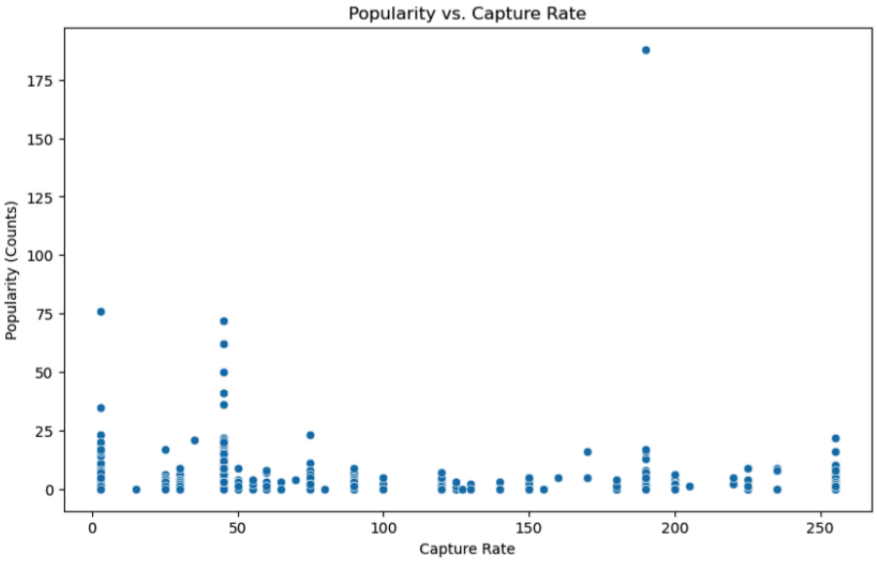


Figures for Analysis of Pokemon Popularity (Question 2):
I used Seaborn to visualize the correlations between Pokémon popularity and their qualities. It offers a high-level interface for creating visually appealing and educational statistical visuals. Seaborn's aptitude for deftly and effectively managing intricate data visualizations was the primary factor in this decision.

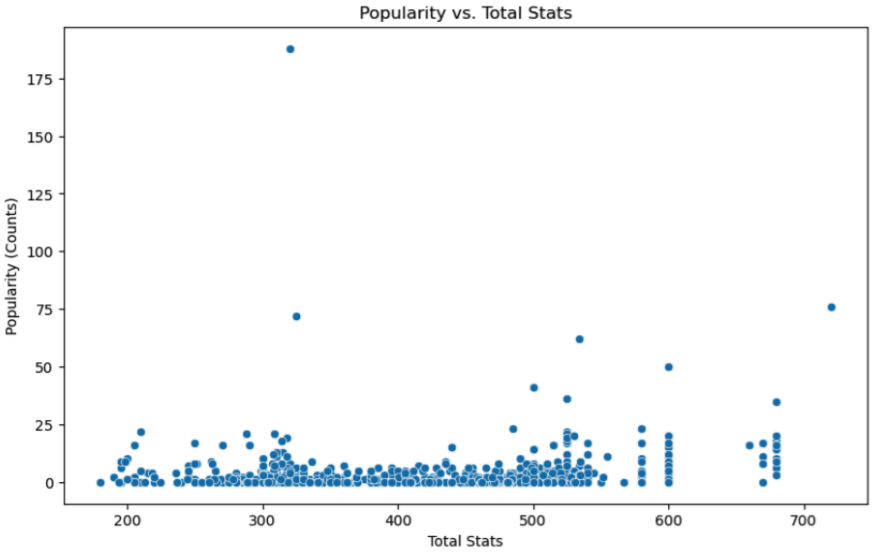
The pop_df sub dataset was used to create each visualization for question 2, and I used Seaborn to create scatter and bar plots to examine various facets of Pokémon popularity:

Scatter Plots

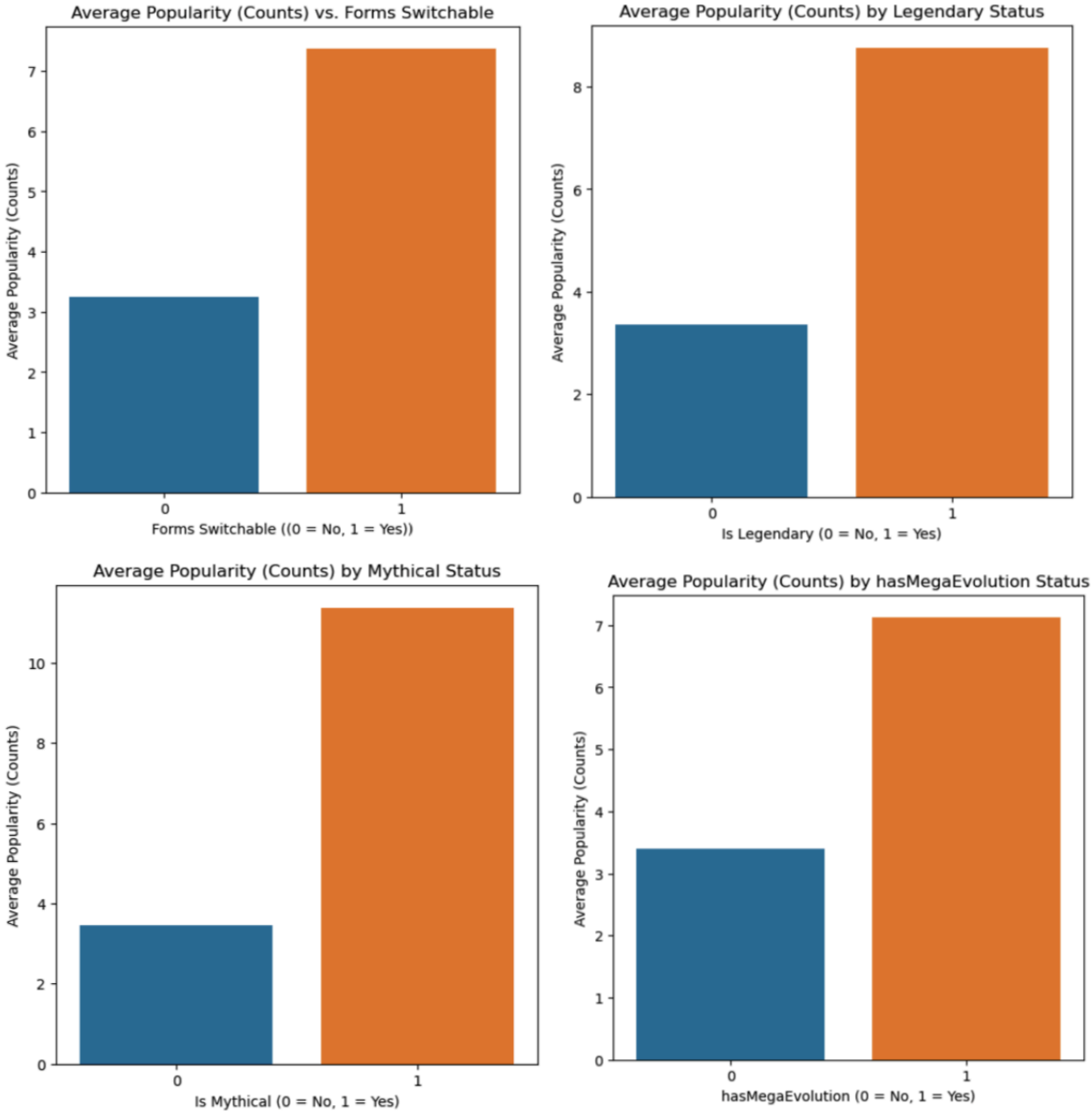
<u>Capture Rate vs. Popularity:</u>
The seaborn package with pop_df was used to create this scatter plot, which plots capture_rate on the x-axis and Counts (popularity) on the y-axis. The relationship between popularity and capture rate is not clearly correlated in the visualization.



<u>Total Stats vs. Popularity:</u>
Similarly, this scatter plot was created using seaborn by plotting Total stats against Counts on the y-axis. Similar to the figure from before, there is no obvious pattern or relationship seen here between a Pokémon's popularity and total stats.



Bar Plots

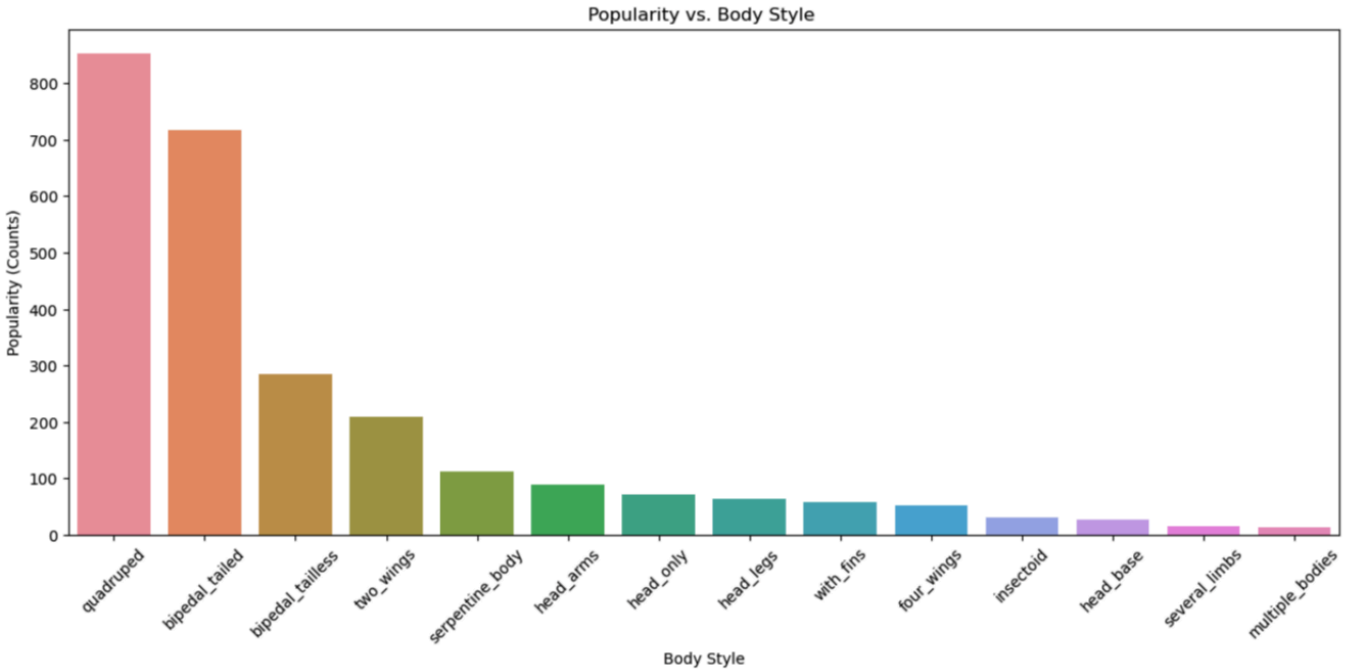


Average Popularity by Forms Switchable: Investigates whether the ability to switch forms affects a Pokémon's popularity.

Average Popularity by Legendary Status: Assesses whether legendary Pokémon are more popular than non-legendary ones.

Average Popularity by Mythical Status: Explores if mythical Pokémon are generally more popular than non-mythical ones.

Average Popularity by Mega Evolution Status: Compares the popularity of Pokémon capable of Mega Evolution with those that are not.



Popularity by Body Style: Using the Python Seaborn package, a bar chart representing Pokémon popularity by body style was created. A bar plot was made with Body_Style on the x-axis and Total_Counts on the y-axis to show the relative popularity of each body style after aggregating and sorting the total mentions for each style from the pop_df DataFrame. Targeting content and creating games can benefit from this visualization, which clearly shows which Pokémon body types are most popular. It also offers insights into the aesthetic preferences of the Pokémon community.

5. **Conclusions: Describe your findings and their impact.**

Question 1: Analysis of Pokémon Types

- Types with Balanced Stats: With moderate values across all three measures, the profiles of the Electric, Fairy, Water, and Grass types are relatively balanced. They are adaptable to several warfare tactics even though they might not be the best in any one area.
- Lower End of Spectrum: Normal and Bug kinds have the lowest average overall stats, which may limit their usefulness in extremely competitive environments unless they are combined with particular tactics or used in particular specialized roles.
- Top 3 Competitive Choices:
    - Dragon Type: They are the most powerful in both offensive and defensive positions since they have the highest overall numbers and are excellent attackers.
    - Steel Type: Having the strongest defense of all Pokémon varieties, Steel Types have strong general stats that make them formidable opponents in combat.
    - Rock Type: Rock types are well-suited for strategies that call for offensive power and durability because of their strong defensive and assault skills.

Question 2: Analysis of Pokémon Popularity

- Scatter Plots Analysis: As I explore the relationship between capture rate, overall stats(numerical variables), and Pokémon popularity as determined by forum mentions (Counts) using the scatter plots above, I find that the data points are widely distributed and devoid of any clear patterns. The absence of a clear correlation implies that rising Pokémon popularity is not systematically correlated with either improved capture rates or total statistics.
- Bar Plots Analysis: I also investigate the relationship between Pokémon popularity as determined by average counts of mentions and specific traits like legendary status, the ability to swap forms, mythical status, and the ability to mega evolve. I also looked at the variations in popularity among the various Pokémon body types. The analysis's refined findings are as follows:
    - Forms Switchable: Pokémon capable of switching forms are generally more popular, highlighting player preference for versatility in gameplay.
    - Legendary Status: Legendary Pokémon consistently show higher popularity, attributed to their unique abilities and pivotal roles in Pokémon lore.
    - Mythical Status: Pokémon that are considered mythical are also generally more popular, likely due to their rareness and mystique, enhancing their allure and engagement within the community.
    - Mega Evolution Capability: Pokémon with the ability to Mega Evolve are more popular on average, showcasing a preference for enhanced and dynamic battle capabilities.
    - Body Styles: The 'quadruped' and 'bipedal_tailed' body forms are more frequently mentioned, possibly due to their prevalence in media or appealing aesthetics, indicating a preference that could influence game design and marketing strategies.

Question 3: Predicting Mega Evolution:

0.9790209790209791

	precision	recall	f1-score	support
0	0.98	1.00	0.99	131
1	1.00	0.75	0.86	12
accuracy			0.98	143
macro avg	0.99	0.88	0.92	143
weighted avg	0.98	0.98	0.98	143

- Precision and Recall:
    - The model has a precision of 0.98 and a recall of 1.00 for Class 0 (Cannot Mega Evolve). This indicates that it virtually never misclassifies Pokémon that can Mega Evolve as those that cannot, making it extremely accurate in predicting Pokémon that cannot.
    - Class 1: Capable of Mega Evolution With a precision of 1.00, the model accurately predicts every Mega Evolution event, proving its accuracy. But with a

recall of 0.75, it appears that 25% of Pokémon who are capable of Mega Evolution are overlooked by the model, which treats them as incapable.

- F1-Score:
    - The F1-Score is a statistic that provides a balanced perspective of model performance, particularly in imbalanced datasets, by calculating the harmonic mean of precision and recall. The F1-score for Pokémon with Mega Evolution is 0.86, indicating a trade-off between flawless precision and reduced recall.
- The impact of this machine learning model:
    - Understanding Game Mechanics:The machine learning model I developed, with an astounding 98% accuracy rate, can correctly identify which Pokémon are eligible for Mega Evolution. The competitive techniques used in Pokémon games can be greatly impacted by this high degree of predictive power. Players may perhaps obtain a tactical advantage in battles by selecting Pokémon with greater knowledge when it comes to their teams by being aware of the precise traits that indicate Mega Evolution.
    - Guiding Fan Theories and Content Creation: In my opinion,the knowledge gained from this data-driven method can spark debates and fan theories regarding possible Mega Evolutions in the future or the fundamental causes of some Pokémon's special ability. With the model, content creators may delve deeper into the exploration and discussion of game mechanics in forums, tutorials, and videos, increasing community involvement and comprehension of the nuances of the game.

6. **Future Work: Given more time, what direction would you take to improve your project?**

Given more time and resources, there are several directions I would pursue to improve this project:

- including More Data Sources: To capture changing patterns and player input, including datasets from social media and new game releases could enhance analysis and offer a more complex picture of Pokémon efficacy and popularity.
- Advanced Analytical Techniques: To gain a deeper understanding of data insights and player emotions surrounding various Pokémon varieties, advanced machine learning models and sentiment analysis are implemented.
- Cross-Validation and Model Optimization: To increase predicted accuracy and guarantee dependability, use strong cross-validation techniques and optimize model parameters.
- Interactive Visualization Tools: By creating dynamic dashboards that enable data exploration, interactive dashboards can improve user engagement and help them better visualize the intricacies of Pokémon statistics and predictions.