Ryan Elicker

Dr. Julie Butler

DSC 140S

6 December 2024

Digimon Final Report

Everybody has heard of the media giant that is the Pokémon franchise (*The Official Pokémon Website*), but there is another big franchise that hides in its shadow, Digimon (*Digimon Encyclopedia*). Most people wouldn’t struggle to name 5 Pokémon, but very few people could name even a single Digimon. Bringing attention to Digimon was the original reason for using the Digimon data frame for analysis. After much analysis, however, it has become clear that there are many things that Digimon has that make it unique and many systems that could use in depth analysis. These things are going to be what the analysis performed will seek to make clear.

The first thing that seems clearly important is a background on our dataset and how it works. This dataset uses the statistics for the Digimon from the video game Digimon Story: Cyber Sleuth, released in 2015 and originally for the PlayStation 4 system (Tatman). Digimon Story: Cyber Sleuth has both types and attributes, which might be confusing for somebody familiar with Pokémon. The types work like they do in Pokémon, but there are significantly less of them, only four. Virus type is effective against data type, data type is effective against vaccine type, vaccine type is effective against virus type, and free type is neutral. The attributes are more plentiful, so they will not be listed here, but they function similarly to the element system in Persona 5 (Persona 5) for the PlayStation 3, as each Digimon has its own numeric strengths and weaknesses to specific attributes. Another important system is the stage system, which shows how evolved a Digimon is. The stages, in order, are Baby, In-Training, Rookie, Champion, Ultimate, Mega, and Ultra. There are also Armor stage Digimon, but they exist outside of the normal hierarchy. Each Digimon also has memory, which is a limiter on how many Digimon can be used, generally memory cost goes up as the Digimon gets more powerful, and thus makes those Digimon more costly to use. Finally, each Digimon has values for HP(Health Points), SP(Skill Points), Attack, Defense, Intelligence, and Speed, which influence the numbers in combat. With that, anybody should have a fairly substantial knowledge of how Digimon works.

Next, we can begin with the actual analysis of the data and diving into how that could be useful to an interested party. For this task, the Python programming language was used through Jupyter Notebooks software. After importing libraries to make the program function and importing the dataset itself, the obvious first step was to obtain descriptive statistics over the numerical aspects of the data frame, things like the mean and the standard deviation of each column, the table that contains that information is contained below. A screenshot of a computer screen

Description automatically generated

Figure 1 Descriptive Statistics of Numeric Data

What seems interesting is that the mean value for the attack stat is higher than for the other basic stats besides HP, perhaps implying that attack should not be focused on too much as a player of the game. Many of these stats have relatively high standard deviations compared to their means, perhaps implying that many datapoints are widespread. The next thing that needs to be focused on are the categorical columns of data, as we have only focused so far on the columns that contain numeric values. A good way to focus on this is to print out how many datapoints fit into each category for the three major columns of categorical data, stage, attribute, and type. The amount of Digimon that fall into each category are listed in a table below. A screenshot of a computer screen

Description automatically generatedA screenshot of a computer

Description automatically generatedA white background with black text

Description automatically generated

Figure 2 Table of Distributions of Categorical Data

We’ll start with the stage section, there are more Digimon in each stage going up until you get to the strongest, as there are only 6 Ultra Digimon, there are also only 3 Armor Digimon, but they fall outside of the normal hierarchy. Next is attribute, something interesting is that there are more Digimon that possess the dark and fire attributes than possess their opposite, light and water respectively. It seems that the earth, wind, and water attributes are underrepresented in particular, so prioritizing Digimon that possess those attributes might be a wise decision for somebody trying to optimize a team in Digimon. For these attributes, a separate count plot was made to visualize the data, this is present at the bottom of the paragraph. Finally, we come to type, an interesting thing is that the free type is very underrepresented, so playing in a way that has no specific weakness might be particularly challenging for the average player because the free type is so uncommon. Because there are so many virus types and relatively few data types, a vaccine type might be the ideal Digimon to have, because it would be strong against the many virus types and weak only to the few data types. A graph of a number of bars

Description automatically generated

Figure 3 Countplot of Attributes

With a deep analysis of the patterns in the basic data out of the way, we are free to analyze the more complicated statistics such as the relationships between different statistics. For this process, we will begin by using the Pearson’s Correlation Coefficient to analyze whether or not two numeric columns are related to one another. The way this coefficient works is that it returns a decimal value between -1.0 and 1.0. If the value is strongly negative, meaning close to -1.0, that means that it is strongly negatively correlated, so as one value goes up, the other generally goes down. On the other hand, if it is strongly positive, meaning close to 1.0 the data is strongly positively correlated, so as one value goes up, the other value generally goes up with it. If the value is close to zero that implies the data isn’t related. Because it would take too long to do this one at a time for each combination of numeric columns, we will create something called a correlation matrix that calculates each of them for us. This matrix is displayed below, along with a heatmap for easier interpretation. A screenshot of a graph

Description automatically generatedA chart of a heatmap

Description automatically generated

Figure 4 Correlation Table and Heatmap of Numerical Data

As can be determined through the color bar on the right side of the heatmap, the data never went very negative, so the color bar doesn’t interpret values below about -0.2 or so. Some interesting things about the heatmap are that only SP and attack ended up having a negative correlation at all, and it's weak. Also, despite just being a number that counts each Digimon, number correlates well with everything. Additionally, SP and intelligence are actually related fairly strongly, and have the best relationship unrelated to number, though speed and memory are fairly related as well. Because SP and intelligence are so strongly correlated, it seemed like it would be a good time to use a scatter plot and line of best fit to sufficiently visualize this data, and the plot is shown below. A graph of a line graph

Description automatically generated with medium confidence

Figure 5 Scatter Plot of Intelligence vs. SP

This makes the idea of a strong correlation really make sense, as that high R-value translates to the data points sticking closely to a very positively sloped line. Another demonstration of this modeling is to do the same process for SP and Attack which is negatively correlated and only weakly, once more shown below. A graph with blue dots and a red line

Description automatically generated

Figure 6 Scatter of Attack vs. SP

This R-value is much closer to zero, so the points don’t cling nearly as closely to the line of best fit as before, and that line has a negative slope because the R-value is negative as well. I also calculated a 2-sample Kolmogorov-Smirnov test between speed and memory and received a very low p-value below 0.05, which implies they are from different distributions.

The natural next step in the analysis of Digimon data is to show the relationships for categorical columns of data. For this, we cannot use a correlation coefficient because the data doesn’t involve numbers, so we’ll use a chi-squared test, which returns a p-value that claims the data is unrelated if larger than 0.05 and related if less than 0.05. We’ll do this process for the attribute and type columns and return a contingency table and the p-value of course. A screenshot of a computer

Description automatically generated

Figure 7 Contingency Table and P-Value of Types vs. Attributes

Something interesting that the contingency table shows is that dark attributes, which have the most amount of Digimon, are almost always virus types, light has a similar pattern for vaccine types. Free types are fairly rare but it makes sense that they would have the most neutral attributes. As for the p-value from the chi-squared test it is very far below 0.05 so the data is undoubtedly significant. For more analysis and a deeper understanding, I decided to repeat the process for stage and type, results shown below. A white background with black text and numbers

Description automatically generated

Figure 8 Contingency Table and P-Value of Type vs. Stage

There are even more helpful results from this contingency table. For one, it is interesting that there are no ultra data types, ultra is the highest level of evolution in Digimon, so vaccine ultras would be powerful, because they are weak to data, and would have an advantage in the absence of data-types. Another thing is that virus types also tend to have a strong presence in most stages, which also implies that vaccine types are powerful because virus is weak to vaccine. The chi-squared test has a similar result to before, because it is so far below 0.05, the data is definitely significant.

The natural last bit of analysis for the detailed Digimon dataset is to use a machine learning program. For this step, we will be using the k-Nearest Neighbors system, which can predict categorical outputs with numerical data when properly trained. You do this by splitting the data into a portion for testing and a portion for training. Here, the training data is 95% of the data and the testing data is the remaining 5%. The output we were trying to get was stage, so the numeric inputs used were memory and speed, since those correlated well with number, which generally increases with stage. This particular test looked at the 5 nearest datapoints to determine patterns. On the particular run used for the purpose of this report, the accuracy of the test was around 86.4%. A value this high is fairly impressive for an algorithm of this type, but success can vary somewhat each run. We also used a confusion matrix to determine where the program was mis categorizing the stages and for this run the confusion matrix is displayed below. A chart of different colored squares

Description automatically generated

Figure 9 K-Nearest Neighbors Confusion Matrix

The two ultras that ended up in the testing data were the ones that ended up getting categorized incorrectly, so I believe that the test would be improved by removing the ultra-stages as outliers or adding more so that the program can properly categorize them.

After all of the trials and tribulations of analyzing the Digimon dataset, it can be determined that Digimon has many unique and complicated systems of its own that make it interesting. This report seeks to be helpful and informative for both the process of data science and the franchise of Digimon. Perhaps the data presented could be what gets a casual observer to give a Digimon game a try and truly succeed at it.

Works Cited

“Digimon Encyclopedia: Digimon Web: Digimon Official General Site.” *Digimon Web | Digimon Official General Site*, digimon.net/reference\_en/. Accessed 4 Dec. 2024.

Persona 5, Directed by Katsura Hashino, P-Studio / ATLUS, 2016, Sony PlayStation 3 game

Tatman, Rachael. “Digimon Database.” *Kaggle*, 13 July 2017, www.kaggle.com/datasets/rtatman/digidb.

“The Official Pokémon Website.” *Pokemon.Com*, www.pokemon.com/us. Accessed 4 Dec. 2024.