# Pokémon Data Analysis Project Report

Pokémon Data Analysis Project Report

**Introduction**

For this project, I chose to analyse a Pokémon dataset because of my interest in both data analysis and Pokémon. Pokémon is a globally recognised franchise with a huge variety of characters, attributes, and forms, making it an ideal case for learning data analysis techniques. Analysing real-world datasets that align with personal interests can make the learning process more enjoyable and relevant. This dataset, sourced from Kaggle, provides comprehensive data on Pokémon from Generations 1 to 6, including information on their names, categories, types, and various base stats (Banik, 2017). To improve the accuracy and reliability of the analysis, I also referenced Serebii.net, a trusted Pokémon database, for cross-checking Pokémon names, forms, and abilities (Serebii.net, 2024).

This project aimed to develop and demonstrate skills in core areas of data analysis, such as data import, cleaning, modelling, and summary statistics, all within Excel. By working with Pokémon data, I was able to explore analytical methods and gain insights into the structure and diversity of Pokémon across multiple generations.

**Data Import**

The original dataset was downloaded from Kaggle as a CSV file and imported directly into Excel. For transparency and data integrity, the imported data was placed in a dedicated "RawData" sheet and left completely untouched during the rest of the analysis. This approach allowed for comparisons between the original dataset and the final cleaned data, making it easier to spot errors or unintended changes throughout the project.

Additional research was performed using Serebii.net to verify details such as Pokémon names, alternative forms, and specific abilities. When inconsistencies were discovered in the dataset, Serebii.net was used as the authoritative reference to ensure that the project relied on the most accurate and up-to-date information available.

To further improve workflow and avoid data loss, I made a duplicate of the original data. Keeping a backup of the initial import is a key best practice in data analysis projects, as it provides a safe point to revert to if cleaning steps cause unexpected issues.

Data Cleaning and Preparation

Cleaning and preparing the data took place primarily in the "Final Cleaned Data" sheet. This stage included several essential tasks:  
- Removing duplicates: Ensured that every Pokémon appeared only once, which is critical for **accurate analysis.**- Handling missing values: Where data was missing (for example, a type or category), information was filled using trusted sources such as Serebii.net. If the information was unavailable, affected entries were flagged and excluded from relevant summaries.  
- Standardising category labels: Labels for Pokémon categories such as “Regular,” “Legendary,” and “Mega” were reviewed and unified, preventing analysis errors due to inconsistencies in naming or spelling.  
- Trimming and correcting names: The TRIM function in Excel was applied to remove unnecessary leading or trailing spaces from text fields, which is especially important for accurate lookups and joins.

A secondary dataset containing Pokémon abilities was also cleaned and imported. This dataset, sourced from Kaggle and cross-referenced with Serebii.net, was loaded as a CSV using Excel’s import tool. The TRIM function was used here as well to ensure all names matched cleanly. The XLOOKUP formula was then used to merge Pokémon abilities into the main dataset by matching Pokémon names.

Helper columns were created in the Final Cleaned Data sheet to support flexible analysis:  
- IsMega: Flags whether a Pokémon is a Mega form based on its name.  
- IsLegendary: Marks Pokémon as Legendary or not, using the category label.  
- FormType: Classifies each Pokémon’s form as Mega, Galar, Alola, or Regular using keyword searches within the name field.

These columns made it easier to group, filter, and summarise Pokémon data based on form, rarity, or other attributes.

**Data Modelling**

Once the data was cleaned and helper columns were added, the dataset was converted into an Excel Table for dynamic referencing and easier formula application. Bringing in ability data through XLOOKUP added depth to the analysis, enabling new questions about how abilities are distributed across different Pokémon forms and types. The dynamic nature of Excel Tables also allowed summary formulas and pivot tables to update automatically as the data changed.

These enhancements allowed for new forms of insight, such as exploring how many Pokémon of a certain type have unique abilities, or comparing Mega and Legendary Pokémon to Regular forms. The flexible structure made it possible to quickly answer questions and adjust the analysis as needed throughout the project.

**Data Analysis**

I analysed the cleaned and modelled data using a combination of formula-driven summary tables, pivot tables, and supporting charts. Key aspects of this analysis included:  
- Counting Pokémon by Type 1: The COUNTIF formula was used to tally the number of Pokémon for each Type 1 value. This highlighted the most and least common Pokémon types, such as Water being the most prevalent with 112 entries.  
- Average HP by Category: The AVERAGEIF formula calculated the mean HP for each major Pokémon category (Regular, Legendary, Mega), revealing that Legendary Pokémon have higher average HP than Regular Pokémon.  
- Total Attack by Type 1: The SUMIF formula was applied to calculate total Attack stats for each Type 1, helping to identify which types tend to have the highest total offensive power.

All summary tables and calculations were linked to the Excel Table structure to ensure that results updated automatically as changes were made during cleaning or further analysis.

**Data Visualisation**

Visualisations were created in a dedicated “Charts and Visuals” sheet using Excel’s built-in bar and pie chart tools. These visual representations focused on:  
- The distribution of Pokémon by Type 1 (bar chart).  
- Comparison of average stats by category and type (bar and pie charts).  
- Category comparisons, such as the relative frequency of Mega, Legendary, and Regular Pokémon.

Conditional formatting was applied to the summary tables, using colour scales to highlight unusually high or low values. This visual approach helped make trends and outliers more immediately visible and accessible.

**Insights and Findings**

Several important findings were identified during analysis:  
- Type Distribution: Water type is the most common in the dataset, while types such as Dragon, Ghost, and Fairy are comparatively rare.  
- Legendary vs Regular Pokémon: Legendary Pokémon generally possess higher average HP and Attack stats compared to Regular Pokémon, supporting the idea that they are intended to be more powerful and unique within the game.  
- Mega Forms: Mega Pokémon consistently display significant increases in attack strength relative to their base forms, illustrating the impact of form changes on gameplay.

The dataset covers Pokémon from Generations 1 through 6, ensuring a broad historical perspective for the analysis. These results provide valuable insights not only into Pokémon themselves but also into how game designers have structured different categories and roles.

Conclusion and Reflection

Completing this project allowed me to develop essential data analysis skills, especially in cleaning, merging, and modelling data using Excel’s advanced features. Working with a familiar and engaging dataset made it easier to maintain interest and ensured the project remained relevant to both my academic and personal goals. The process also reinforced the importance of keeping raw data untouched, maintaining backups, and using authoritative references when correcting errors. Screenshots could not be added due to each time one was to be taken the entire worksheet errored and removed whatever was on sheet.

This project showed how powerful well-structured data can be for uncovering patterns, testing hypotheses, and drawing conclusions. Future work could include expanding the analysis to cover later generations, deeper exploration of abilities, or even integrating advanced statistical testing or machine learning. Overall, this project gave me hands-on experience with the full data analysis workflow, preparing me for more complex analytical challenges in the future.

## Appendix: Formula Explanations

Appendix: Formula Explanations

COUNTIF  
Counts cells matching a condition — used to count Pokémon by Type 1.  
Example: =COUNTIF(tbl\_Pokemon[Type 1], "Fire")

AVERAGEIF  
Calculates the average of cells meeting a condition — used to find average HP by category.  
Example: =AVERAGEIF(tbl\_Pokemon[Category], "Legendary", tbl\_Pokemon[HP])

SUMIF  
Sums values where condition matches — used to total Attack by Type 1.  
Example: =SUMIF(tbl\_Pokemon[Type 1], "Water", tbl\_Pokemon[Attack])

IF and SEARCH (Helper Columns)  
Checks text for substrings to classify Pokémon forms dynamically.  
Example: =IF(ISNUMBER(SEARCH("Mega",[@Name])), "Mega", "Regular")

XLOOKUP  
Looks up a value in a range and returns the corresponding value from another range. Used to bring in abilities based on Pokémon names.  
Example: =XLOOKUP([@Name], Abilities!$A$2:$A$1000, Abilities!$B$2:$B$1000, "No Ability")

TRIM  
Removes all leading and trailing spaces from a text string, ensuring clean comparisons and lookups.  
Example: =TRIM(A2)

## References

Banik, R. (2017). Pokémon with stats [Data set]. Kaggle. https://www.kaggle.com/rounakbanik/pokemon  
Serebii.net. (2024). Pokémon Database. https://www.serebii.net