# CIS 418/518: Foundation of Data Science

# Project 1

In this project, you will be performing exploratory data analysis to understand a dataset. The plots should be created using matplotlib and seaborn python libraries. Follow the instructions during class for making good and bad plots. You are required to create presentation based on the following tasks.

**What is expected to be in the presentation (Minimum Requirement):**

1. Title Slide with group member names
2. Key actions (e.g. How you dealt with missing values, how you dealt with outliers)
3. Descriptive statistics of the dataset.
4. Answers to the questions and related plots.

**Team Size:** 3

**Submission Deadline:** Please check the related announcement in the course homepage **Presentation Date:** Please check the related announcement in the course homepage

**Presentation Duration:** 10 minutes + 2 minutes

**Submission folder in d2l:** Assignments/Project\_1

**What to submit in d2l:**

1. Single Notebook with all the code for each of the tasks. Use headers and sub-headers to differentiate the tasks and the questions in a task in the notebook.
2. Presentation (ppt or pdf)

**Dataset:** The dataset for this project is available in content/dataset/project1\_ds.csv. The dataset contains information about Pokémon. There are 16 columns in the dataset. This dataset provides information about various Pokémon and their generation along with their type and different attributes. A brief description of the columns in the dataset is provided below:

***Name:*** *Name of each pokemon*

***Type.1:*** *Each Pokémon has a type, this determines weakness/resistance to attacks*

***Type.2:*** *Some Pokémon are dual type and have Type2. Note not all Pokémon will have a Type 2.*

***Total:*** *A general guide describing how strong a Pokémon is.*

***HP:*** *hit points, defines how much damage a Pokémon can withstand before fainting*

***Attack:*** *the base modifier for normal attacks*

***Defense:*** *the base damage resistance against normal attacks*

***SP..Atk:*** *special attack, the base modifier for special attacks (e.g. fire blast, bubble beam)*

***SP..Def:*** *the base damage resistance against special attacks*

***Speed:*** *determines which pokemon attacks first each round*

***Generation****: Number of Generation*

***Legendary:*** *True if Legendary Pokemon , False if not*

***Win\_rate:*** *Rate of win for the pokemon*

***Type1\_Win\_rate:*** *Rate of win for type 1 pokemon*

**Note:**

1. Tasks 5 is for CIS 518 section only.

***Task 1:*** *Reading the dataset and preparing the dataset for future tasks*

1. Import pandas, matplotlib, seaborn and any required library.
2. Read the dataset file and examine the first few records. Look at the records to get an idea about the data.
3. Check the shape of the DataFrame.
4. What do you think about the first two columns? Can you use them for analysis? If not, what can you do? Update the DataFrame based on your decision.
5. Now let’s take a more detailed look at the types of different columns in the frame. What is the data type for each of the columns.

***Task 2:*** *Handling the Missing Values*

1. Before we analyze the data, it is important to deal with any missing values in the dataset. Missing values in a dataset can be generated due to fault in data collection. It is also possible that some missing values are intended based on the data. There are several ways to deal with missing values. For example, you can either remove rows from the dataset containing missing values, replace the missing values with median or mean or mode (categorical variables) values. **Missingno** is a python library that helps us visualize the missing numbers in a dataframe.
   1. Use missingno to visualize these missing values.
   2. Check how these missing numbers are represented in the dataframe.
2. Based on the column type, what should be a good approach to handle these missing values? Do the required task to deal with the missing values.

***Task 3:*** *Handling outliers:*

1. For this assignment, we will use box plots to visualize outliers in the quantitative variables. Below is a box plot showing distribution of the different quantitative variables. Check for the presence of outliers in the dataset. How should you handle the outliers in this case?

Chart, box and whisker chart

Description automatically generated

1. Now let’s try to get a higher-level understanding about the data. Check the descriptive statistics of the data set and comment on the statistics.

***Task 4:*** *Analyzing association between two variables*

1. ? Create plots to answer the following questions:
   1. What type of Pokémon are most common and what type of Pokémon are more uncommon? Legendary and non-legendary
   2. Which generation have the maximum number of Pokémon?
   3. Which generation have the maximum number of legendary Pokémon?
   4. Which type of Pokémon’s are the most-speedy Pokémon? Here you need to determine what is a good speed for a Pokémon.
   5. Which type I Pokémon have the highest win rate?
   6. Is there any correlation between win\_rate and any other variables (e.g. Attack, speed, Defense etc.)?
   7. Does Pokémon with high attack score also possess high defense scores?

**Task 5:** *Analyzing two pokemons in a single chart*

1. Compare the HP, Attack, Defense, Sp..Atk, Sp..Def, Speed, total between any two Pokémon belonging to the class of Pokémon with highest win\_rate and lowest win\_rate. Draw these attributes in one single chart for the comparison. Do you see any major differences between the abilities?