Lab 04

Pandas

A. Multiple Choice (10 points)

1. You are analyzing a dataset named *students_data.csv* that contains information about students, including their names, ages, and grades. The dataset has columns: *Name*, *Age*, and *Grade*. You want to find out the age of the student named "John Doe".

   After loading the dataset into a pandas DataFrame named df, which of the following statements will give you the age of the student named "John Doe"?

   (a) df.loc['John Doe', 'Age']

   (b) df['Age'].loc[df['Name'] == 'John Doe']

   (c) df[df['Name'] == 'John Doe']['Age'].values[0]

   (d) df.at['John Doe', 'Age']

2. You are working with a dataset named *sales_data.csv* that contains monthly sales data for different products across various regions. The dataset has columns: *Month*, *Product*, *Region*, and *Sales*. You are tasked with analyzing which product has the highest average monthly sales across all regions.

   After loading the dataset into a pandas DataFrame named *df*, you use the *groupby* method to group the data by *Product* and calculate the average monthly sales for each product. Which of the following statements will give you the product's index with the highest average monthly sales?

   (a) df.groupby('Product')['Sales'].mean().idxmax()

   (b) df.groupby('Product').sum()['Sales'].max()

   (c) df['Product'][df['Sales'].idxmax()]

   (d) df.groupby('Product')['Sales'].max().sort_values(ascending=False).head(1)

B. Programming Exercise (30 points)

Next, we will use the Pokemon.csv dataset to practice **data filtering**. This dataset includes Pokémon attributes such as #, Name, Type 1, Type 2, Total, HP, Attack, Defense, Sp. Atk, Sp. Def, Speed, Generation, and Legendary.

Question 1 (10 points)

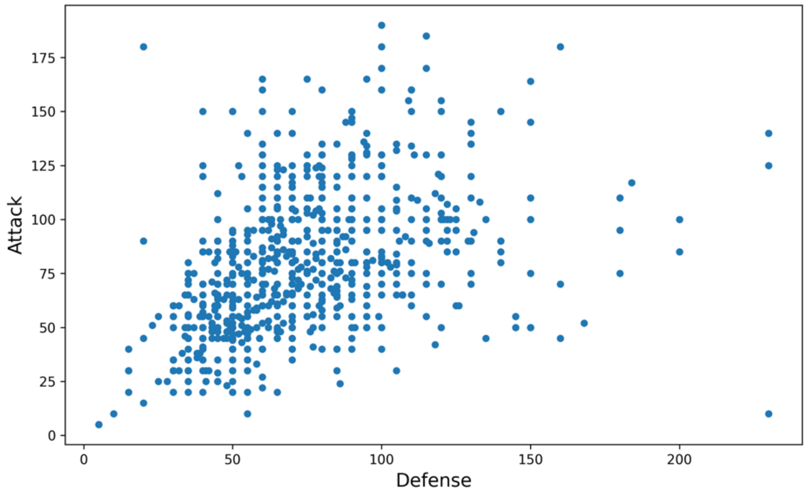
Using the dataset, determine the following:

a. What percentage of legendary Pokémon have an Attack value greater than 150? (3 pts)

b. What percentage of non-legendary Pokémon have an Attack value greater than 150? (3 pts)

c. Provide a brief description of your findings. (1 pt)

d. Given the scatter plot provided:



Identify the Pokémon that appears as an outlier in the lower right corner. (3 pts)

Question 2 (10 points)

Given a DataFrame containing Pokémon data, we are interested in understanding the distribution of Pokémon across different types.

Complete the function ***pokemon_type_count()***. This function should compute the number of Pokémon for each type1.

note:

For this question, only consider type1.

Question 3 (10 points)

The objective here is to compare the average attack value across different Pokémon types. Complete the function ***average_attack_type()***. This function should compute the average attack value for each Pokémon type.

C. Data Analysis for Climate Change (60 points)

In this part, you will work with a dataset *GlobalLandTemperaturesByState.csv* containing historical climate data for states across the world from the year 1744 to 2013. The dataset includes average temperature for various states and their respective date.

Question 1: Data Import (6 points)

Use Pandas to import the climate change dataset into a DataFrame called df_state. Then find out all the country names from the 'Country' column and print them out. (there are a total of seven unique country names.)

Question 2: Data Cleaning (12 points, 12 points)

The first step in examining any dataset involves the preparation and refinement of the data. Various forms of irregularities can occur during the data collection or curation process, and it is essential to rectify these issues before conducting any analysis.

i. Some country names include additional affiliations, such as "United States (US)". Create the function ***preprocess_data*** to simplify these names, we should discard any additional affiliations. In a broader sense, any country name in the format "name1 (name2)" should be replaced with just "name1".

ii. Create the function ***drop_missing_values*** to eliminate rows in our datasets that have missing values. Missing data can cause issues when we're analyzing the data, and the easiest way to deal with this is to delete rows that have any missing values.

Question 3: Data Analysis (12 points)

We can get an overview of our dataset by examining summary statistics. To do this, we will use Pandas to load the climate change dataset into a DataFrame and then display key statistics such as the minimum value, maximum value, average (mean), and standard deviation of the "AverageTemperature" column.

Question 4: Outlier Detection (12 points, 6 points)

We can identify outliers using the Interquartile Range (IQR) rule: a data point is considered outlier if it is at least 1.5 interquartile ranges below the first quartile (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3), i.e.,

$$\textbf{outlier} \leq Q\mathbf{1}-\mathbf{1.5}\times IQR \text{ or } \textbf{outlier} \geq Q\mathbf{3}+\mathbf{1.5}\times IQR.$$

Introduction of IQR: https://en.wikipedia.org/wiki/Interquartile_range

Create a function named ***remove_outliers***. This function will be responsible for removing rows from a DataFrame where the values in a specified column are identified as outliers based on the IQR rule.

After creating the function, apply it to the "AverageTemperature" column in our preprocessed df_country DataFrame. Then, compare the minimum value, maximum value, average (mean), and standard deviation to the original data where the ***remove_outliers*** function was not used.