UNDERSTANDING POKEMON GO OCCURRENCE PATTERNS

GREGORY MORGAN, ASTHA MEHTA, BERTHA SANCHEZ

University of California San Diego

PokémonGo is an augmented reality game in which players locate, capture, battle and train virtual creatures called Pokémon. These Pokémon appear in the same location as the players and can be tracked using the GPS features of their device. The appearance of a Pokémon is supposed to be linked to various factors such as time of day, type of geographic location, proximity to water, etc.

Dataset Description

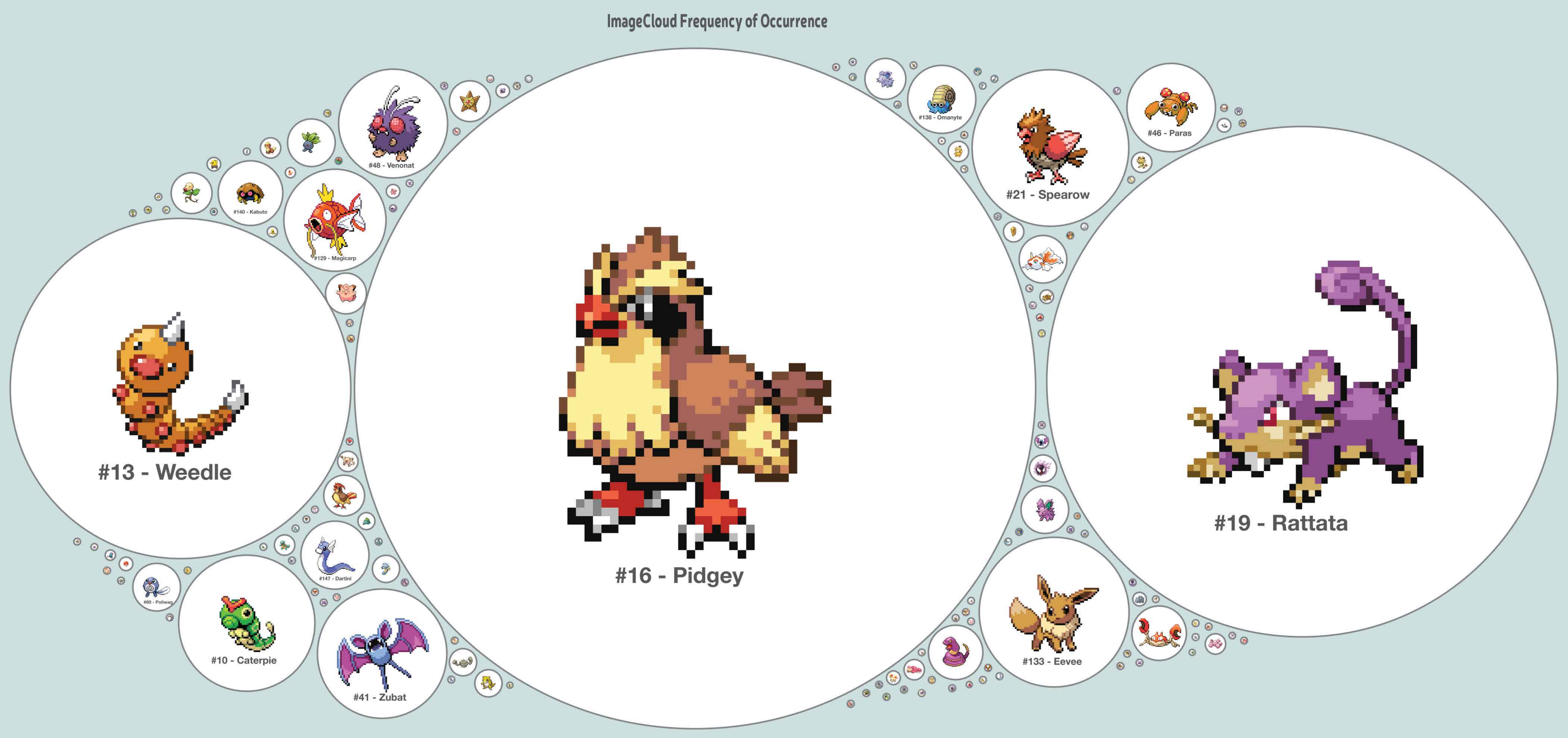
Source of main dataset: https://www.kaggle.com/semioniy/predictemall
Source of rarity dataset: http://www.pokego.org/rare-pokemon-list/

The main dataset (source: Kaggle) contains information for 151 Pokémon across 25 features making for a total of 296,022 data points collected through services like pokeRadar, NEO, darksky.net, Google's S2 geometry library, and GLFC Modis Land Cover Facility during an encounter with a Pokémon while playing PokémonGO.
Features of the dataset include information regarding the **location** the encounter took place at (longitude and latitude), the city and country, **setting** (urban, rural etc.), **terrain** (water, grassland etc.), the population of the place it was encountered, the time of day, the **environmental conditions** (**temperature, pressure, wind speed, wind bearing** etc.) at the time of the encounter, and the details regarding the proximity of a Gym or a Pokestop (determined by real world locations) to the encounter. Another feature that spans 151 columns is **co-occurrence data**; that is, which other Pokémon appear within 100m and 24 hours of the appearance of a certain Pokémon. This data is available for multiple encounters of each Pokémon.

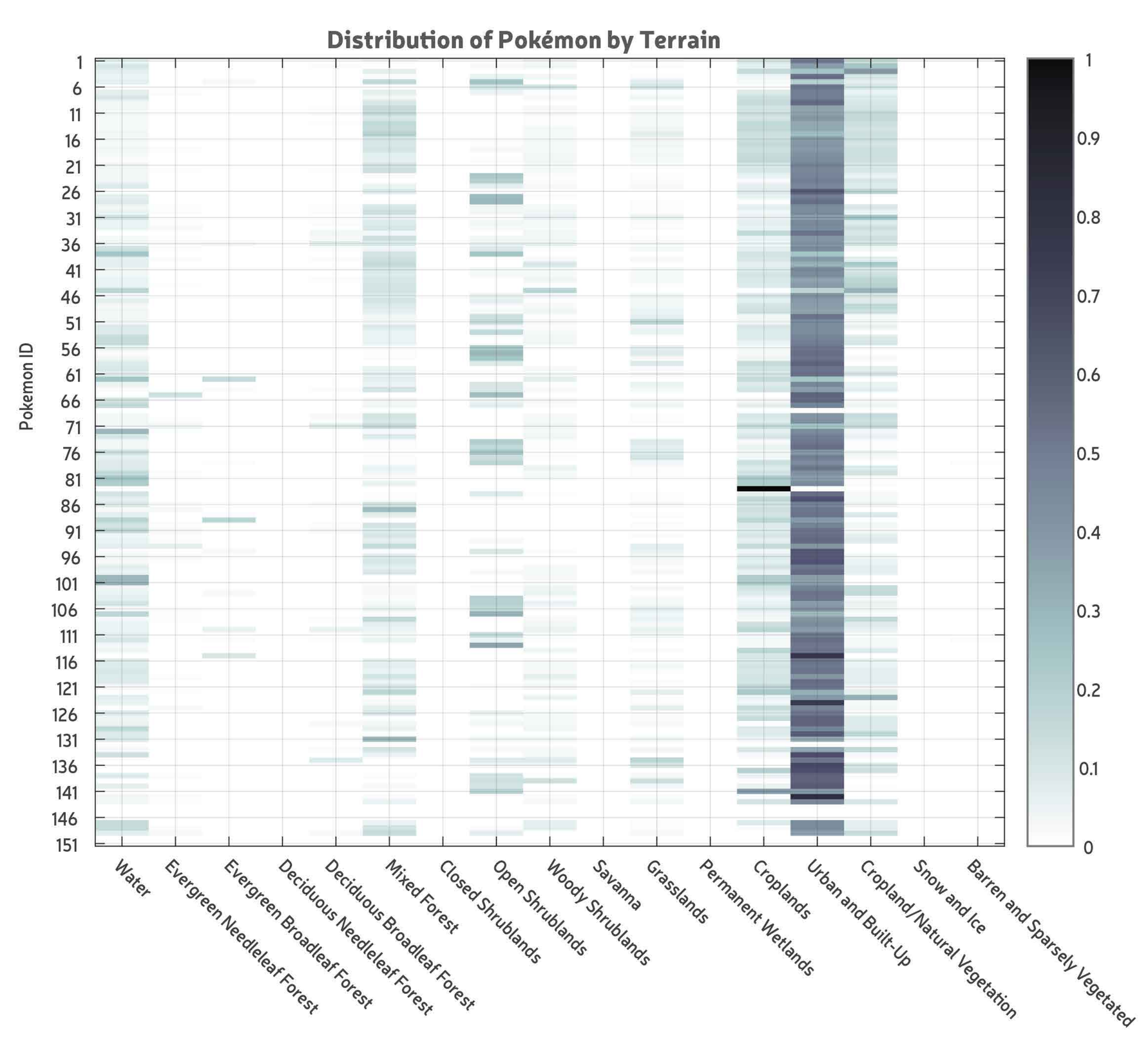
The rarity dataset (source: PokeGo) had categorical information (common, rare etc.) regarding the rarity of the Pokémon.

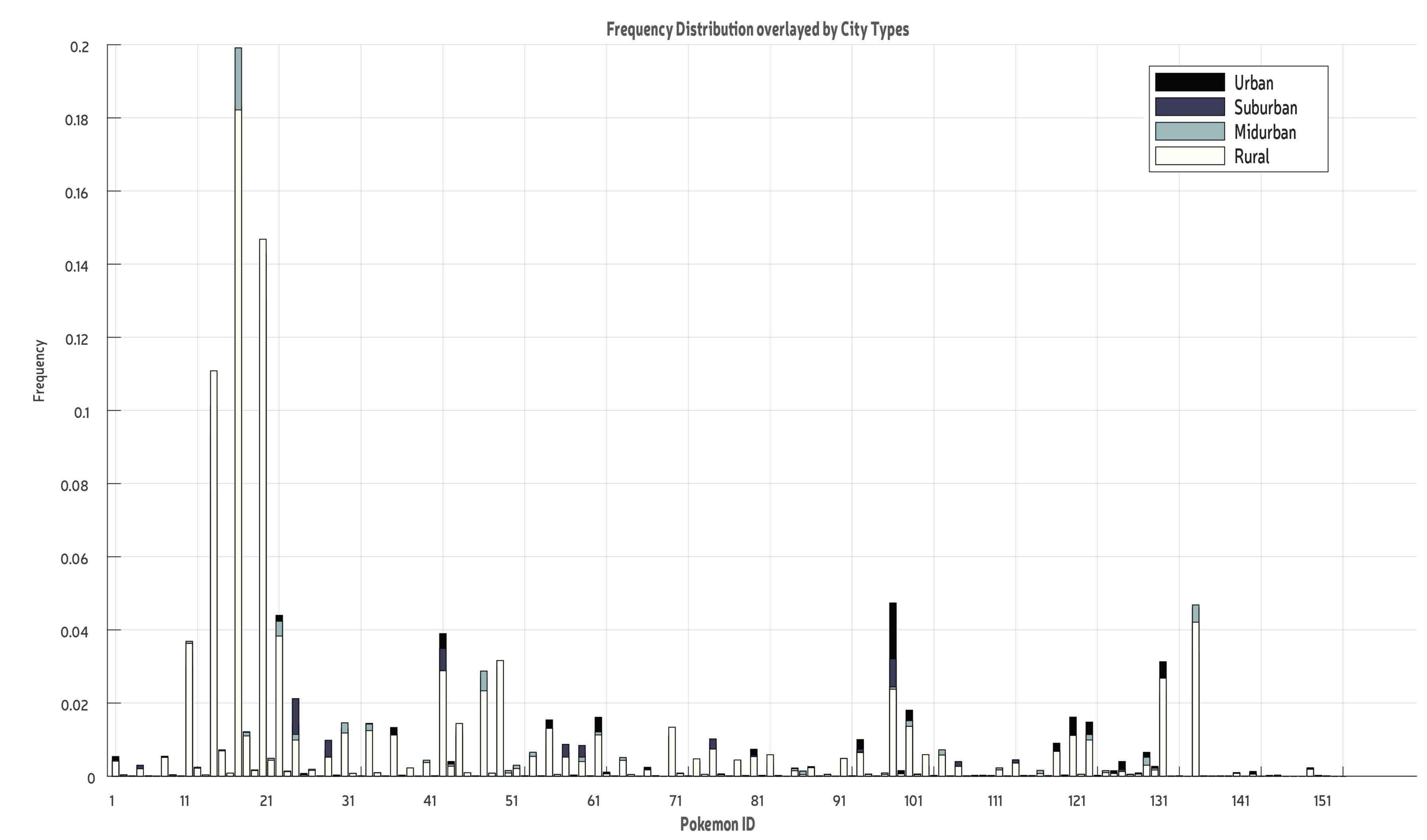
Questions

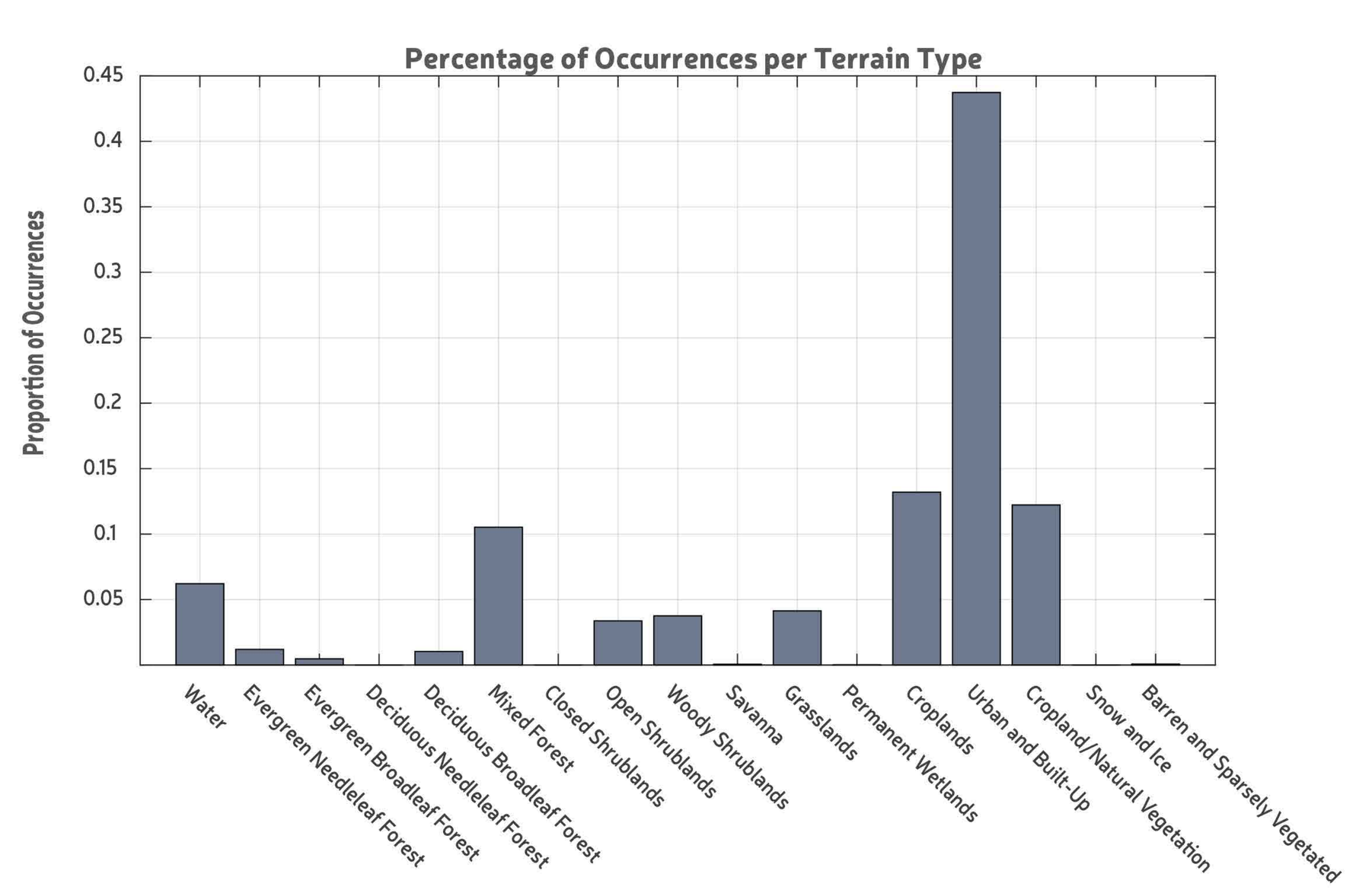
1. Does the distribution of Pokémon differ based on the setting or terrain of where they were encountered?
2. Can one classify the rarity of the Pokémon based on environmental conditions (temperature, pressure and wind speed)?
3. Which Pokémon co occur the most?

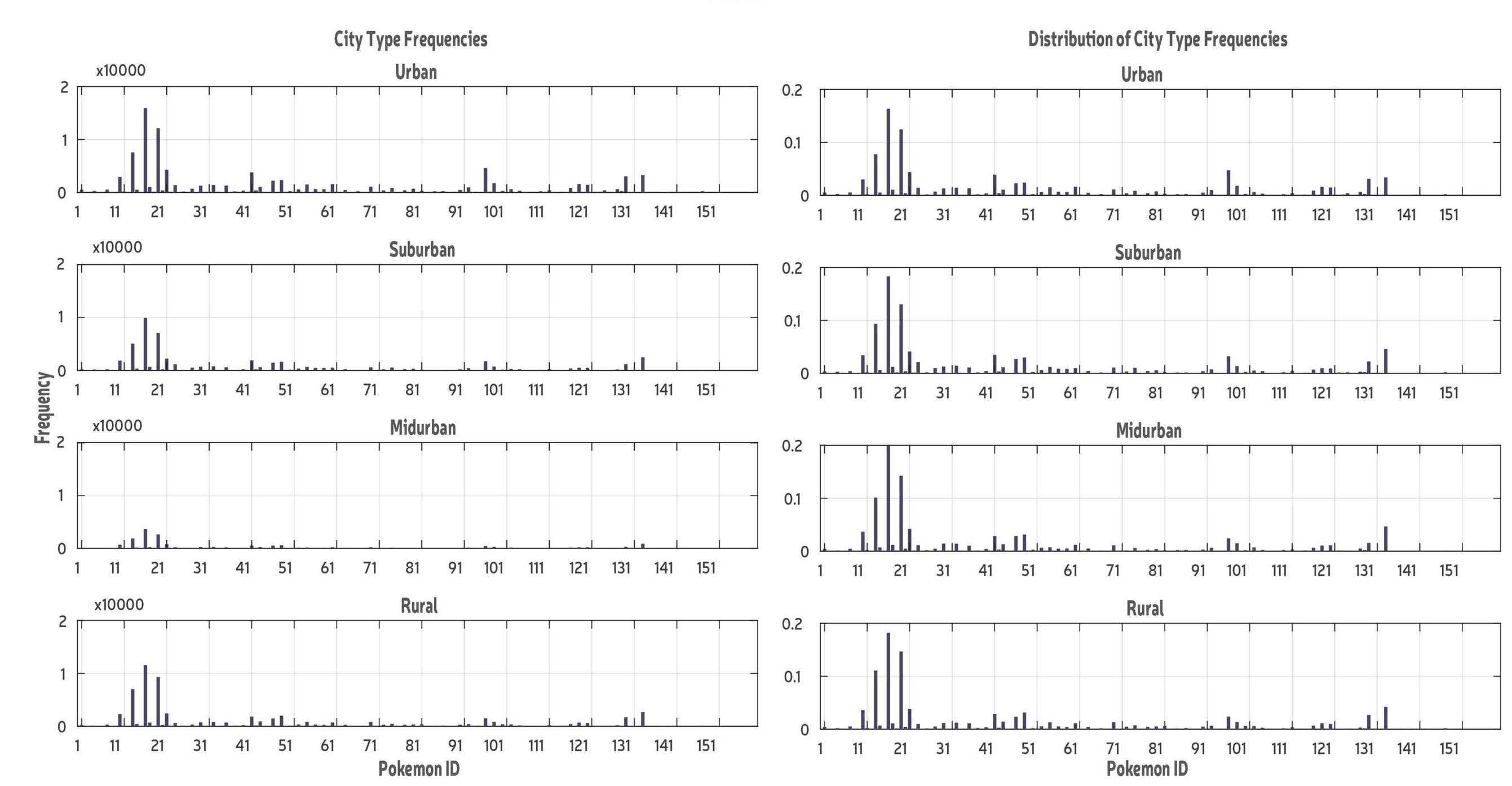


EFFECT OF SETTING ON OCCURRENCE

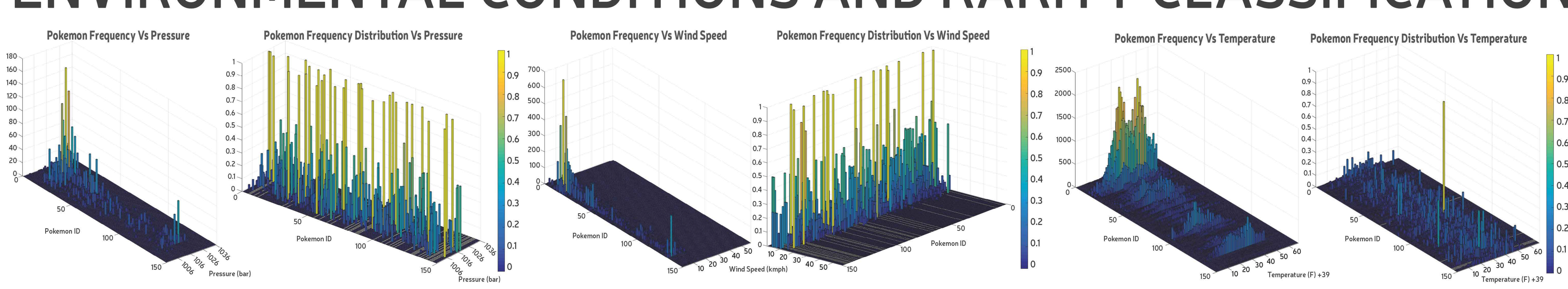


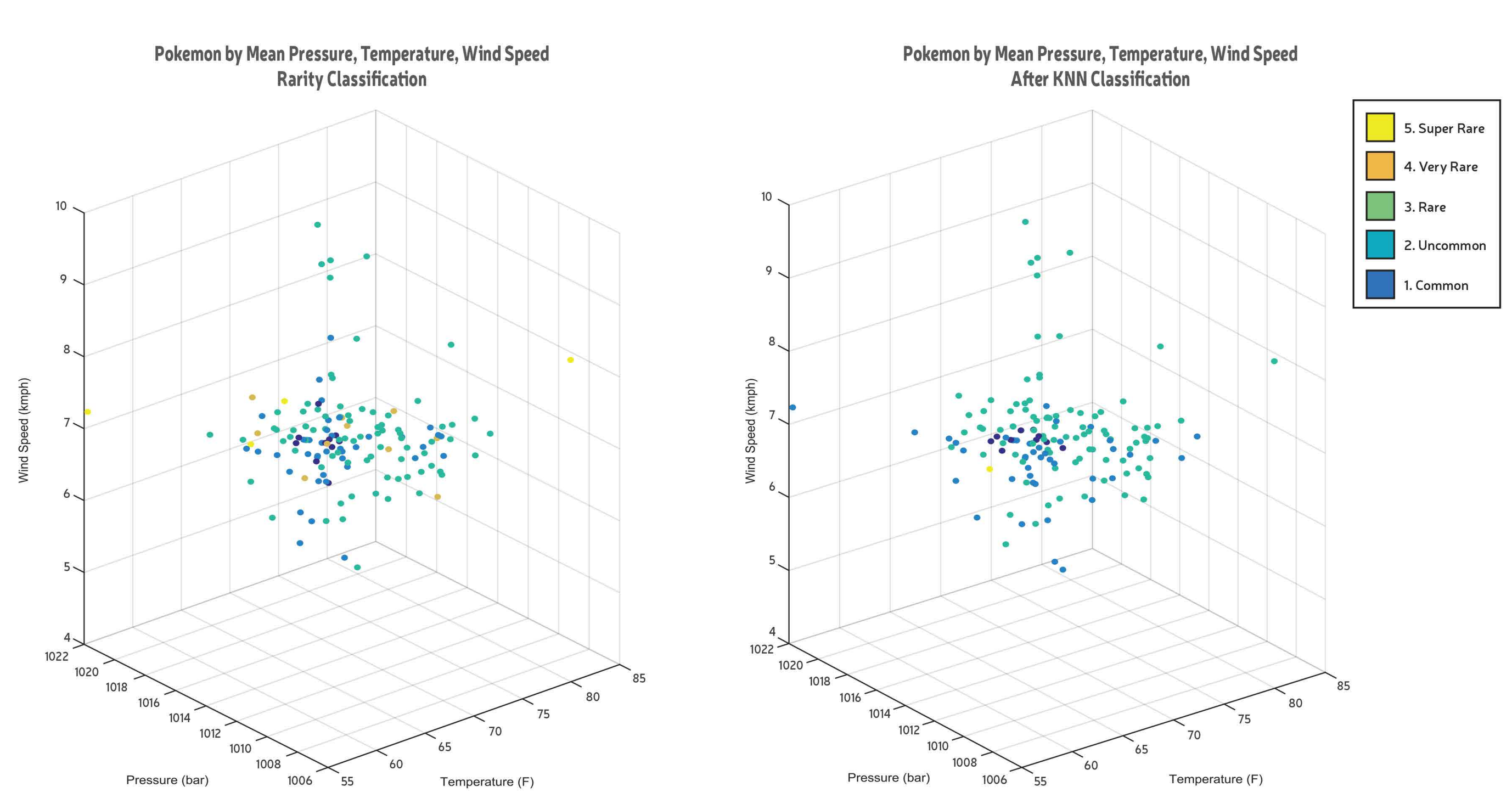


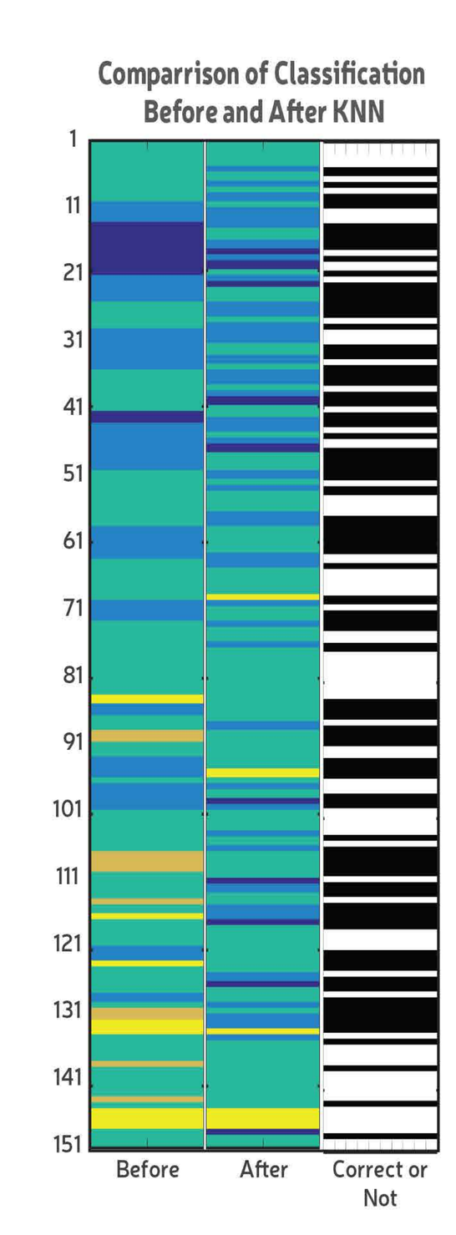


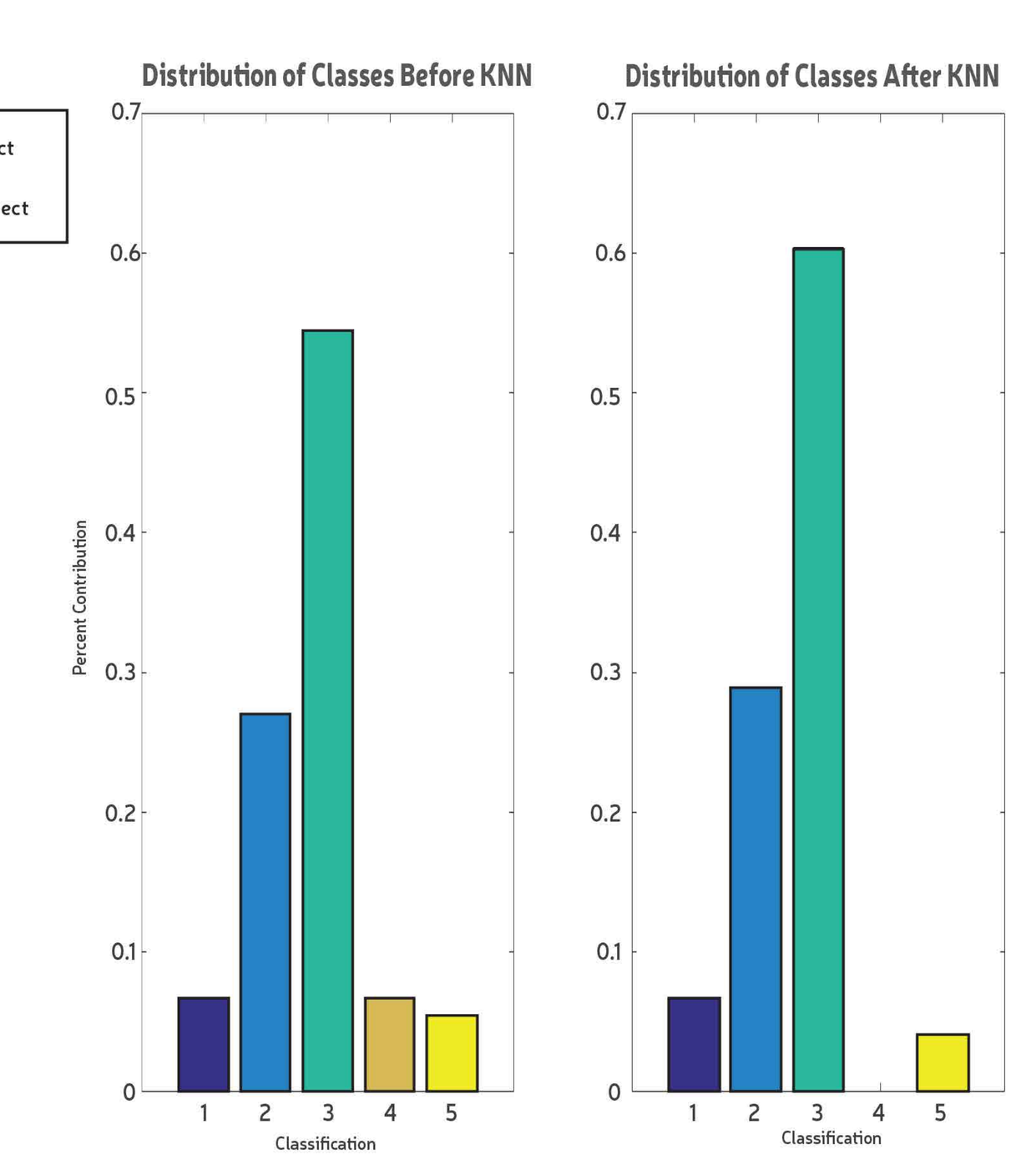


ENVIRONMENTAL CONDITIONS AND RARITY CLASSIFICATION

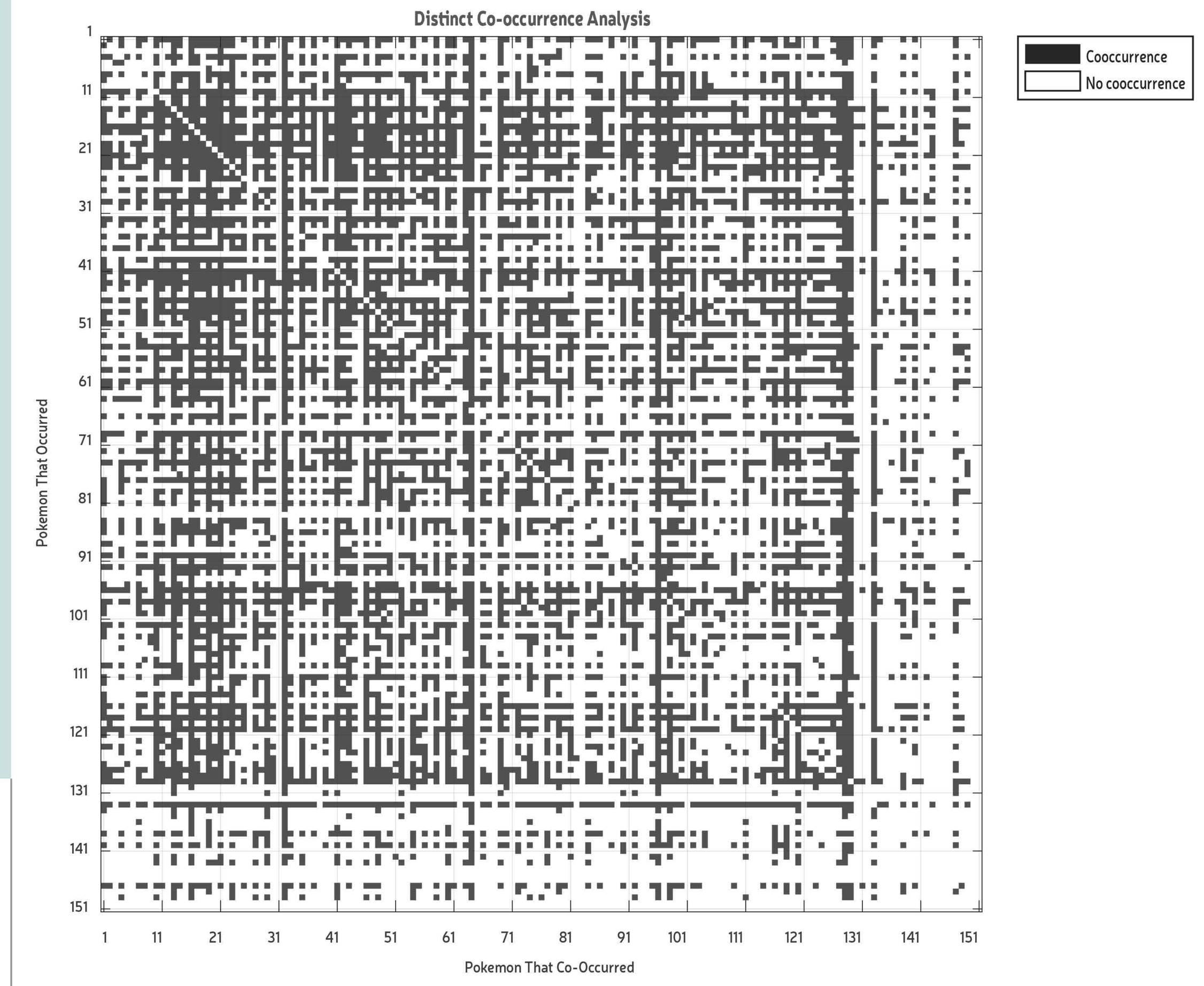


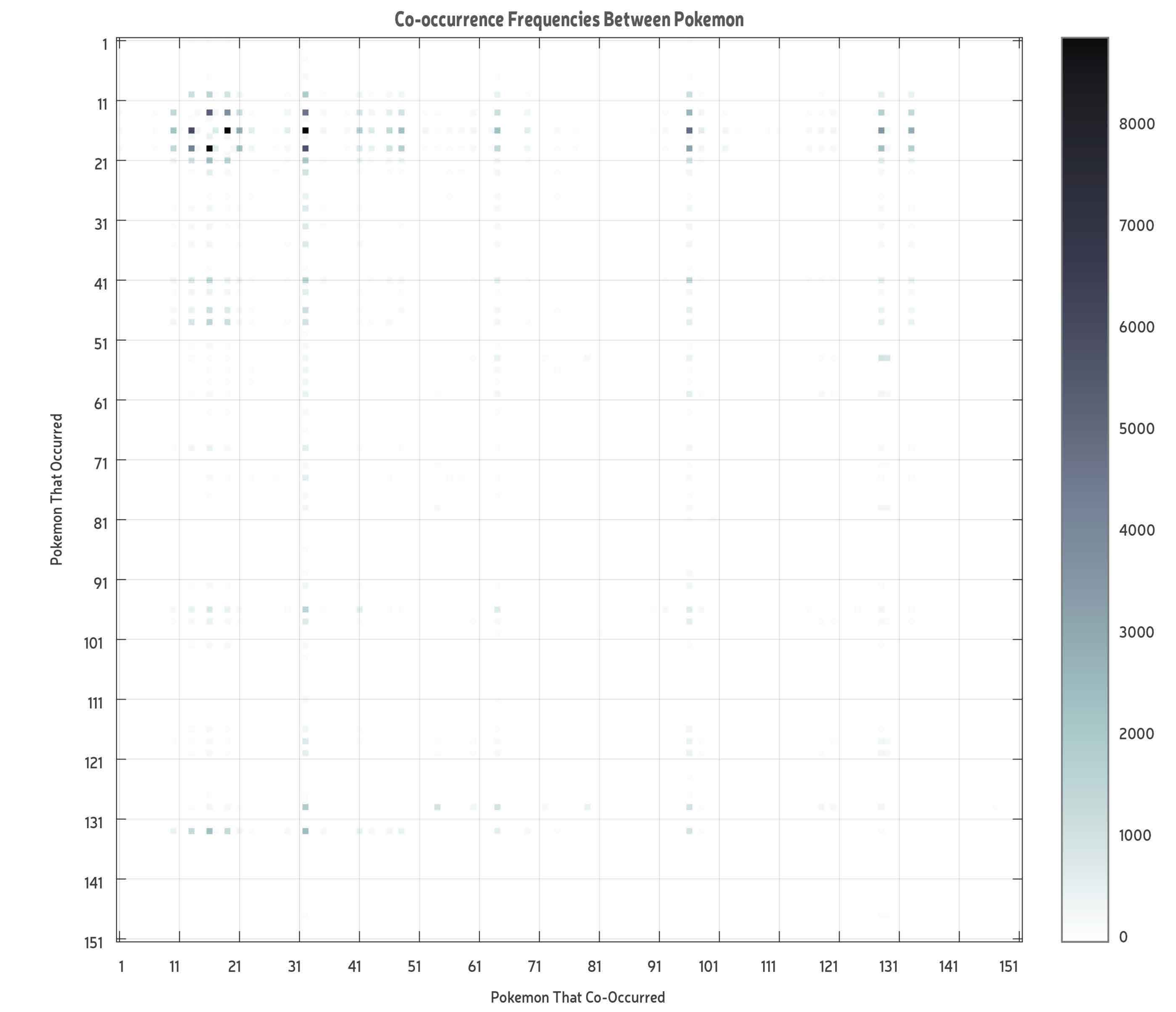


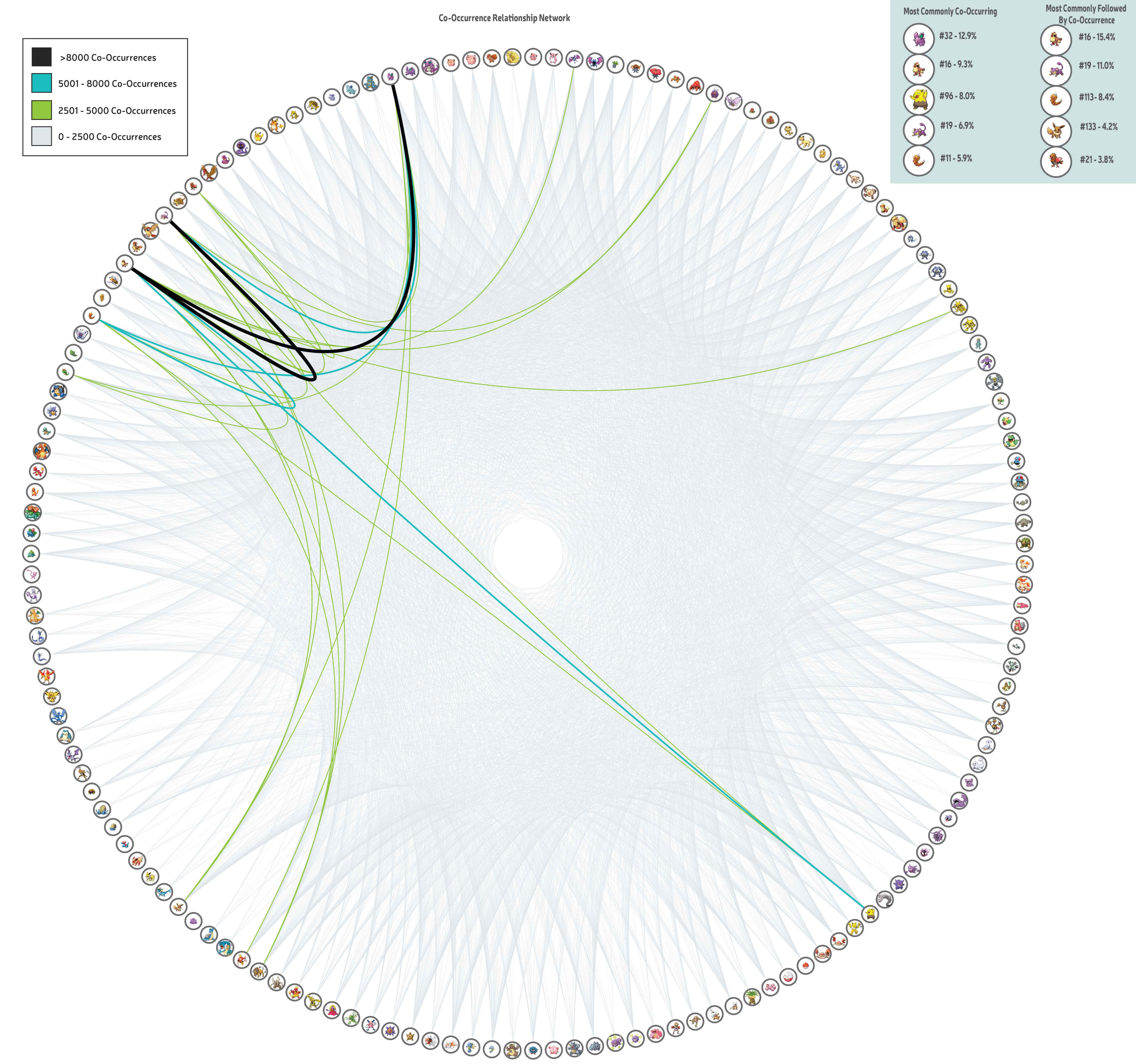




CO-OCCURRENCE BETWEEN POKEMON ENCOUNTERS







METHODOLOGY

1. Frequency Distribution
We chose to analyze using Percent Contribution to demonstrate how the values of Pokemon frequencies occurred over each feature to understand the relative occurrence patterns. Also, this allowed us to observe the frequencies on a uniform, continuous scale.
Sort the Pokémon with respect to all the settings and terrains they occurred at.
Find the frequency of the Pokémon with respect to each of the settings and terrains.
Find the percent contribution of a Pokémon using the following formula.

$$PC(x_i) = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

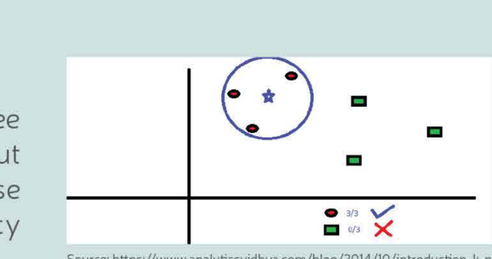
2. K Nearest Neighbors
We wanted to classify the Pokemon based on the three environmental factors. Our dataset was supervised, but our output was discreet. This method allowed us to use the conditions and our output to determine the efficacy of the combination of the features on classification.
1. Find the euclidean distance between a point to each of the other points. A point corresponds to a Pokémon and has three features (temperature (x) , pressure (y) and wind (z)).

$$\text{3D: } distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Source: http://snipd.net/distance-between-2d-and-3d-vectors-in-c

2. Sort the resulting matrix in ascending order and append a column with the rarity classification information.

3. Classify a Pokémon based on 5 of its nearest neighbors. Assign the classification that occurs the most as the classification of that Pokémon.



Source: https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/

4. Repeat the above for each Pokémon with respect to every other Pokémon to generate a classification based on environmental conditions.

5. Compare the classifications generated based on environmental conditions to the actual classifications.

3. Co-Occurrence Visualization
For this question, we were not trying to predict any feature and this question was unsupervised. The data we were working with was boolean information, we could not cluster this data for classification. Utilizing visualization techniques allowed us to represent the relationships between Pokemon that co-occurred.
Sum up every Pokémon's co occurrence with respect to the Pokémon that it co occurred with.
Find the percent contribution of each co occurrence and visualize the strength of the co occurrences.

CONCLUSION

1. While the frequencies of encounters of Pokémon in different terrains may differ slightly, the shape of the distribution remains the same. This means that the frequency of a Pokémon is more or less the same across different settings, but urban areas seem to have the most occurrences of each type of Pokémon compared to the others.
(Note that we cannot conclude whether the occurrences are biased towards urban settings or whether the data we have is predominantly generated in an urban setting.)

2. 69 out of 149 (151 - 2 as 2 Pokémon did not occur at all) were classified correctly based on environmental conditions. This means that our KNN algorithm classified the data correct 46.3% of the times. Hence, the environmental conditions of temperature, pressure and wind speed are not the best parameters to classify the rarity of a Pokémon as they classified < half correctly.

3. There were certain Pokémon that did not occur at all and hence, had 0 co occurrence. There were some Pokémon that co occurred with most other Pokémon. Of these, Pidgey (16) was responsible for the most co occurrences. Co occurrence is influenced by the frequency of occurrence of a Pokémon.

FUTURE WORK

Try KNN with other combinations of features to see which combination classifies the Pokemon rarity better.