PES University, Bangalore
UE15CS322 - Data Analytics

Session: Aug – Dec 2017
Week 3 – Assignment 2

Date of Submission: **Max Marks:** 20

NOTE: *In your assignment books, write the question, R commands required to get the desired output and the obtained output. For questions that require you to plot graphs, you will have to also print the obtained output (graphs/charts) and attach it in your assignment book. The solutions to the assignment must be* ***hand-written****.*

TOPIC: Sampling, Normalization and Principal Component Analysis

If you were born in the 90s, you were more than likely involved in the Pokémon phenomenon in some way or another. Whether you loved it or hated it, Pokémon took the entire world by storm when it was first introduced to the world. Created in 1995, what started as a basic video game eventually involved into its own long running TV show, a trading card game, movies and so much more. Back in its prime, Pokémon was literally everywhere you went and it continues to have a strong impact in today's youth as well. On July 6, 2016, the streets began to fill with 90s kids who grew up with this beloved franchise as well as many others. Our heads were down in our phones, and we walked all over town in pursuit of reliving our beloved childhood memories and once more become a Pokémon master.

Excerpts from 'Reliving Your Childhood with Pokémon Go', Odyssey

Pokémon initially began as a Role Play Game (RPG) and further evolved into television series and mobile games. The dataset, **pokemon.csv** includes 21 variables (associated with the RPG) per Pokémon that are:

- **Number**. Pokémon ID in the Pokédex.
- **Name**. Name of the Pokémon.
- **Type_1**. Primary type.
- **Type_2**. Second type, in case the Pokémon has it.
- **Total**. Sum of all the base stats (Health Points, Attack, Defense, Special Attack, Special Defense, and Speed).
- **HP**. Base Health Points.
- **Attack**. Base Attack.
- **Defense**. Base Defense.
- **Sp_Atk**. Base Special Attack.
- **Sp_Def**. Base Special Defense.
- **Speed**. Base Speed.
- **Generation**. Number of the generation when the Pokémon was introduced.
- **isLegendary**. Boolean that indicates whether the Pokémon is Legendary or not.
- **Color**. Color of the Pokémon according to the Pokédex.
- **hasGender**. Boolean that indicates if the Pokémon can be classified as female or male.
- **Pr_male**. In case the Pokémon has Gender, the probability of its being male. The probability of being female is, of course, 1 minus this value.
- **Egg_Group_1**. Egg Group of the Pokémon.
- **Egg_Group_2**. Second Egg Group of the Pokémon, in case it has two.
- **hasMegaEvolution**. Boolean that indicates whether the Pokémon is able to Mega-evolve or not.
- **Height_m**. Height of the Pokémon, in meters.
- **Weight_kg**. Weight of the Pokémon, in kilograms.

- **Catch_Rate**. Catch Rate.
- **Body_Style**. Body Style of the Pokémon according to the Pokédex.

The dataset was extracted from bulbapeida. More information about the attributes may be obtained here.

Question – 1 (6 points)

Building any model involves dividing the dataset into training data and test data. Choosing a training data involves sampling. For the given dataset, obtain the following samples:

(i) Simple Random Sample which includes 70% of the dataset
(ii) Systematic Sample with k = 3 (starting from the first element, i.e. 1^{st}, 4^{th}, 7^{th}, etc.)
(iii) Stratified Random Sample where the strata are 'male' and 'female'
(iv) Clustered Sample based on primary type (choose 70% of the types)

For each of these samples and for the population, plot the distribution of *Total* attribute and state which sampling technique most closely represents the entire population.

What type of data is stratified sampling most suitable for?

Question – 2 (6 points)

Calculate the distance between means of Pokémons of primary type, *Grass* and *Fire*. Consider the normalized values of *HP, Attack, Defence, Sp_Atk, Sp_Def* and *Speed*, normalized between 0 and 1 for mean. Also, plot a line graph of the mean of the normalized values of these attributes for each type of Pokémon.

Calculate the kurtosis and skew of the weights and heights of the Pokémons. What inference can you draw from these values?

Convert the values of *HP, Attack, Defence, Sp_Atk, Sp_Def, Speed, Total, Height_m* and *Weight_kg* to a new range, so that the new values have a mean 0 and standard deviation 1. Plot these transformed values on one chart (and find out which variables have a normal distribution). What is significant about distributions with mean 0 and standard deviation 1?

Question – 3 (8 points)

Ash Ketchum's Pokédex (dexter) has ceased working. He realizes that you hold a dataset on Pokémons and consults you to help him regain his database. Ash wants to store the numerical values in your dataset on his new device to refer to, in his future contests. However, he soon realizes that he would run short of memory if he stored all numerical attributes in your dataset. He can **associate only 2 numerical values** with each Pokémon on his new device.

Your dataset contains 12 numerical values – *Total, HP, Attack, Defence, Sp_Atk, Sp_Def, Speed, Generation, Pr_Male, Height_m, Weight_kg* and *Catch_Rate*. To help Ash, you decide to perform Principal Component Analysis on the dataset to reduce the entire dataset to 2 principal components.

(i) You perform preliminary analysis on the dataset and decide to drop 2 variables from consideration for the PCA procedure. Which 2 variables are **most appropriate** to be dropped? Give supporting reasons.
(ii) On the remaining 10 variables, apply the principal component analysis and plot the variance associated with each of the new variables. What is the extent of information

contained in the two most principal axes? (Note: The relative importance of the i^{th} Eigen value, λ_i, out of n eigen values is computed as $\lambda_i/(\lambda_1+\lambda_2+\ldots+ \lambda_n$.)

(iii) Print the data that needs to be stored in Ash's new device. Also provide a scatterplot of the first principal component against the second.