Assignment

A set of numbers satisfies Bernford's law if the leading digit d occurs with probability

$$\begin{aligned}\pi_d &= \log_{10}(d+1) - \log_{10}(d) \\ &= \log_{10}\left(\frac{d+1}{d}\right) \\ &= \log_{10}\left(1+\frac{1}{d}\right)\end{aligned}$$

Therefore, the frequency distribution of the leading digit d in such a set is the following: $\pi_1 = 30.1\%$, $\pi_2 = 17.6\%$, $\pi_3 = 12.5\%$, $\pi_4 = 9.7\%$, $\pi_5 = 7.9\%$, $\pi_6 = 6.7\%$, $\pi_7 = 5.8\%$, $\pi_8 = 5.1\%$ and $\pi_9 = 4.6\%$.

In this assignment, you will examine this law when applied to a fashion dataset "FashionDataset.csv" from Kaggle.
`https://www.kaggle.com/datasets/mukuldeshantri/ecommerce-fashion-dataset`

This dataset is a collection of about 31,000 women fashion products. Categories covered in this dataset are western wear, Indian wear, perfumes and fragrances, watches and nightwear.

The dataset has the following columns:

1. BrandName: Mentions the brand of the product
2. Details: Details about the product
3. Size: Sizes available
4. MRP: This is max retail price
5. SellPrice: This is the price after discount
6. Category: Category of the product

We will focus on the "SellPrice" column. You compute the real distribution $F = (f_1, f_2, \ldots, f_9)$ of frequencies of the leading digit in these prices and compare F to two models:

1. Model 1: equal-weight distribution: each leading digit has the same frequency $1/9 = 11.11\%$ In other words, your predicted model of frequencies in this model is a 9-digit vector $P = (1/9, 1/9, \ldots, 1/9)$
2. Model 2: leading digit follows the Bernford's law. In this model, the prediction is a 9-digit vector $\pi = (\pi_1, \pi_2, \ldots, \pi_9)$

Questions:

1. plot 3 histograms for the frequencies for real distribution, equal-weight and Bernford (for each digit)
2. plot 2 histograms for the relative errors for Models 1 and 2 (for each digit)
3. compute RMSE (root mean squared error) for model 1 and 2. Which model is closer to the real distribution?
4. take 3 categories of your choice For each of these categories do the following: (a) compute F, P and π. (b) using RMSE as a "distance" metric, for which of these chosen three countries is the distribution "closest" to equal weight P?
5. discuss your findings