Introduction to Data Science — Assignment 0

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh** TA(s): Kianoush Arshi

Deadline: Tuesday, Esfand 15th, 11:59 PM

Introduction

In this assignment, we are going to delve into web scraping and perform an introductory data analysis. This homework will be a hands-on exercise that will help you become familiar with the process of extracting data from websites and conducting basic statistical analysis.
An incomplete notebook is provided for you, which guides you throughout the assignment.

Tasks

1. **Environment Setup:** Install the required libraries such as Beautiful Soup, Selenium, pandas, numpy, matplotlib, and seaborn.
2. **Web Scraping:** Write a script to scrape transaction data from Etherscan.io. Use Selenium to interact with the website and Beautiful Soup to parse the HTML content.
3. **Data Sampling and Analysis:** Once the data is collected, create a sample from the dataset. Compare the sample statistics (mean and standard deviation) with the population statistics.

Data Collection (Etherscan)

In this section, we will use web scraping to gather transaction data from the **Ethereum blockchain** using the Etherscan block explorer. Our objective is to collect transactions from the last 10 blocks on Ethereum. To accomplish this task, we will employ web scraping techniques to extract the transaction data from the Etherscan website. The URL we will be targeting for our data collection is: https://etherscan.io/txs.
Follow the steps in the notebook, and also check the considerations!

Data Analysis

Now that we have collected the transaction data from Etherscan, the next step is to conduct an initial analysis. This task will involve the following steps:

- **Load the Data:** Import the collected transaction data into a pandas DataFrame.
- **Data Cleaning:** Clean the data by converting data types, removing any irrelevant information, and handling duplicate values.
- **Statistical Analysis:** Calculate the mean and standard deviation of the population. Evaluate these statistics to understand the distribution of transaction values. The analysis and plotting will be on Txn Fee and Value.
- **Visualization:** This phase involves the creation of visual representations to aid in the analysis of transaction values. The visualizations include:
    - A histogram for each data column, which provides a visual representation of the data distribution. The selection of bin size is **crucial** and should be based on the data's characteristics to ensure accurate representation. Provide an explanation on the bin size selection!
    - A normal distribution plot fitted alongside the histogram to compare the empirical distribution of the data with the theoretical normal distribution.
    - A box plot and a violin plot to identify outliers and provide a comprehensive view of the data's distribution.

Data Sampling and Analysis

In this section, we will delve into the process of data sampling and perform an initial analysis on the transaction data we have collected. Our objective is to understand the distribution of transaction values by sampling the data and comparing the sample statistics with the population statistics. This task will involve the following steps:

- **Load the Data:** Import the collected transaction data into a pandas DataFrame.
- **Data Cleaning:** Clean the data by handling missing values, converting data types, and removing any irrelevant information.
- **Simple Random Sampling (SRS):** Create a sample from the dataset using a simple random sampling method. This involves randomly selecting a subset of the data without regard to any specific characteristics of the data.
- **Stratified Sampling:** Create another sample from the dataset using a stratified sampling method. This involves dividing the data into strata based on a specific characteristic (e.g., transaction value) and then randomly selecting samples from each stratum. Explain what you have stratified the data by and why you chose this column.
- **Statistical Analysis:** Calculate the mean and standard deviation of the samples and the population. Compare these statistics to understand the distribution of transaction values.
- **Visualization:** Plot the distribution of transaction values and fees for both the samples and the population to visually compare their distributions.

Questions

1. What are some potential limitations when using web scraping for data collection? Specifically, what problems did you face while fetching data from Etherscan? What problems can these limitations cause in your analysis?
2. What can make your analysis untrustworthy? What are your solutions?
3. How did the visualization help you in understanding the data? What could you interpret from the plots?
4. How do the two sampling methods differ in their output? Compare these and explain which one is a better fit to the population.

Notes

- Upload your work as a zip file in this format on the website: DS_CA0_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.