Mohammad Yaser Azad | 810800003
Parham Javan | 810800008

## Summary:

This project involves conducting web scraping to gather transaction data from the Ethereum blockchain using the Etherscan block explorer. The collected data is then subjected to data sampling and analysis, including statistical analysis and visualization. The purpose is to familiarize with web scraping techniques, perform basic data analysis, and gain insights into the distribution of transaction values on the Ethereum blockchain.

## Purpose:

1. Web Scraping: The primary purpose of the project is to learn and practice web scraping techniques. By scraping transaction data from Etherscan, participants gain hands-on experience in extracting data from websites, using tools like Selenium and Beautiful Soup to parse HTML content.

2. Data Analysis: Another key objective is to perform introductory data analysis on the collected transaction data. This involves loading the data into a pandas DataFrame, cleaning it, and conducting basic statistical analysis to understand the distribution of transaction values.

3. Statistical Analysis and Visualization: Participants are tasked with calculating mean and standard deviation statistics for transaction values and fees. They are also required to visualize the data distribution using histograms, normal distribution plots, box plots, and violin plots to gain insights into the distribution and identify outliers.

4. Data Sampling: The project includes sampling the collected data using simple random sampling and stratified sampling methods. Participants analyze the sampled data and compare its statistics with the population statistics to understand the representativeness of the samples.

## Question1

**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?**

- Etherscan's rate limiting mechanisms, which may throttle or block excessive scraping activities.

- Handling dynamic content loading, especially for pages with JavaScript-based interactions.

- Ensuring the accuracy and completeness of the scraped transaction data, considering the high volume and continuous flow of transactions on the Ethereum blockchain.

These limitations can pose several challenges and potential problems for the analysis:

- Incomplete or missing data due to scraping restrictions or failures can lead to biased or inaccurate analysis results.

- Changes in website structure or data format can break existing scraping scripts, requiring constant maintenance and updates.

- Legal and ethical concerns may arise if scraping activities violate Etherscan's terms of service or applicable laws, potentially leading to legal consequences.

- Difficulty in ensuring data consistency and quality may affect the reliability and validity of the analysis findings.

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.

## Question2

**2. What can make your analysis untrustworthy? What are your solutions?**

Several factors can potentially undermine the trustworthiness of the analysis conducted in this project. Some of these factors include:

1. Data Quality Issues: If the scraped data contains inaccuracies, inconsistencies, or missing values, it can lead to biased or erroneous analysis results. Poor data quality can arise due to errors in the scraping process, website changes, or limitations in the scraping tools used.

2. Sampling Biases: If the sampling methods used are biased or not representative of the population, the analysis results may not accurately reflect the true characteristics of the data. Biases can occur if certain segments of the population are overrepresented or underrepresented in the samples.

3. Misinterpretation of Results: Incorrect interpretation of statistical analysis or visualization outputs can lead to misinterpretation of the data and erroneous conclusions. Misinterpretation may occur due to lack of domain knowledge, statistical understanding, or context.

4. Selection of Inappropriate Analysis Techniques: Using inappropriate statistical methods or visualization techniques for the data at hand can result in misleading analysis results. It's essential to choose analysis techniques that are suitable for the data characteristics and research objectives.

To address these potential issues and ensure the trustworthiness of the analysis, several solutions can be implemented:

1. Data Validation and Cleaning: Thoroughly validate and clean the scraped data to address inaccuracies, inconsistencies, and missing values. Implement data validation checks and preprocessing steps to improve data quality before conducting analysis.

2. Random Sampling and Stratification: Use random sampling techniques to ensure that samples are representative of the population. Implement stratified sampling if the data contains heterogeneous subgroups to ensure proportional representation across strata.

3. Peer Review and Validation: Engage in peer review and validation processes to verify the accuracy and reliability of the analysis. Seek feedback from colleagues, experts, or mentors to validate analysis methodologies and results.

4. Clear Documentation and Transparency: Document all steps of the analysis process, including data collection, preprocessing, analysis techniques, and interpretation of results. Ensure transparency by providing clear explanations and justifications for analysis decisions.

## Question3

**3. How did the visualization help you in understanding the data? What could you interpret from the plots?**

Visualizations play a crucial role in understanding and interpreting data by providing intuitive representations of complex information. In the context of this project, the visualizations aid in understanding the distribution of transaction values and fees, identifying outliers, and gaining insights into the underlying patterns and trends in the data. Here's how the visualizations help in understanding the data and what can be interpreted from the plots:

1. Histograms:

- Histograms provide a visual representation of the distribution of transaction values and fees.

- They help identify the frequency of occurrence of different transaction value or fee ranges.

- Interpretation: Histograms can reveal the central tendency, spread, and shape of the distribution. For example, a symmetric bell-shaped histogram suggests a normal distribution, while skewed histograms indicate asymmetry in the data.

2. Normal Distribution Plot:

- Normal distribution plots (also known as probability density plots) show how closely the data distribution matches a theoretical normal distribution.

- They help assess whether the data follows a normal distribution, which is often assumed in statistical analysis.

- Interpretation: If the data distribution closely matches the normal distribution curve, it indicates that the data is approximately normally distributed. Deviations from the curve suggest departures from normality, which may require different statistical approaches.

3. Box Plots:

- Box plots provide a summary of the distribution of transaction values or fees, including measures of central tendency (median), spread (interquartile range), and detection of outliers.

- They help visualize the variability and dispersion of the data.

- Interpretation: Box plots show the median (middle line), interquartile range (box), and outliers (points beyond the whiskers). They help identify the presence of outliers and assess the symmetry or skewness of the distribution.

4. Violin Plots:

- Violin plots combine the features of box plots and kernel density plots, providing insights into both the summary statistics and the distribution shape.

- They help compare distributions between different categories or groups.

- Interpretation: Violin plots display the distribution of transaction values or fees across different categories or groups. They offer a more detailed view of the data distribution compared to traditional box plots, making it easier to detect differences and patterns between groups.

## Question4

**4. How do the two sampling methods differ in their output? Compare these and explain which one is a better fit to the population.**

The two sampling methods used in the project, simple random sampling (SRS) and stratified sampling, differ in their approach to selecting samples from the population. Let's compare these sampling methods and evaluate which one is a better fit for the population:

1. Simple Random Sampling (SRS):

- Method: In SRS, each individual in the population has an equal chance of being selected for the sample. Samples are selected randomly without considering any specific characteristics of the data.

- Characteristics:

  - Random selection of samples without bias towards any subgroup or stratum.

  - Every individual in the population has an equal chance of being included in the sample.

- The selection of samples is independent of each other.

- Advantages:

  - Simplicity: Easy to implement and understand.

  - Unbiased: Ensures each individual in the population has an equal chance of selection.

- Disadvantages:

  - May not represent all subgroups or strata equally well if the population contains heterogeneous groups.

  - Less efficient if the population has significant variability across subgroups.

2. Stratified Sampling:

- Method: In stratified sampling, the population is divided into homogeneous subgroups or strata based on specific characteristics (e.g., transaction value ranges). Samples are then selected independently from each stratum using SRS.

- Characteristics:

  - Population divided into strata based on specific characteristics.

  - Samples are selected independently from each stratum.

  - Ensures representation of each stratum in the sample.

- Advantages:

  - Ensures representation of all subgroups or strata in the sample.

  - Provides more accurate estimates for each subgroup.

- Disadvantages:

  - Requires knowledge of population characteristics to define appropriate strata.

  - More complex to implement compared to SRS.

Comparison:

- Representation of Population: Stratified sampling ensures representation of all subgroups or strata in the sample, making it suitable for populations with heterogeneous characteristics. SRS may not adequately represent all subgroups, leading to potential bias.

- Efficiency: SRS is simpler and more straightforward to implement compared to stratified sampling. However, stratified sampling can be more efficient in terms of providing accurate estimates for each subgroup, especially when there is significant variability across subgroups.

- Accuracy: Stratified sampling generally produces more accurate estimates for each subgroup due to its targeted selection approach. SRS may yield less accurate estimates, particularly if the population contains distinct subgroups with varying characteristics.

Conclusion:

The choice between SRS and stratified sampling depends on the characteristics of the population and the research objectives. If the population is relatively homogeneous, SRS may suffice and provide efficient results. However, if the population contains distinct subgroups with different characteristics, stratified sampling is preferable as it ensures representation of all subgroups and yields more accurate estimates for each subgroup. Therefore, stratified sampling is generally considered a better fit for populations with heterogeneous characteristics.