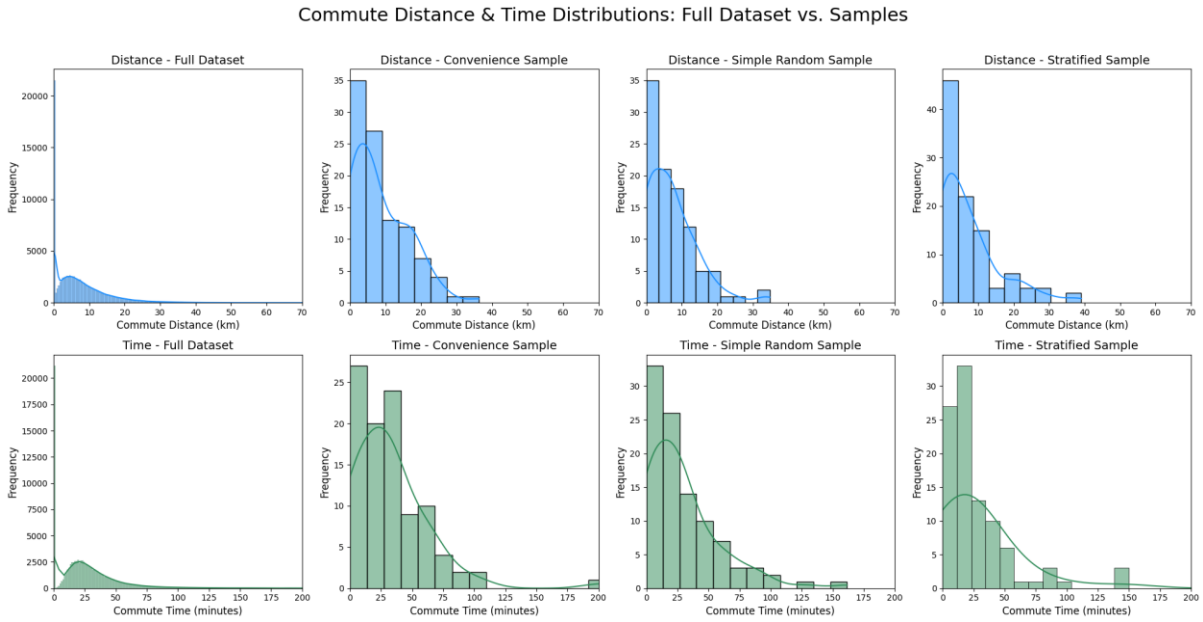
Problem 1

a. Analysis and Comparison of Samples vs Population

To evaluate the quality of different sampling strategies, we compared the population (100,000 commuters) with three types of samples (each of size 100): convenience, simple random, and stratified. The comparison focused on commute distance, remote work days, car ownership, commute method, and commute time.

1. **Commute Distance (km) & Time Distributions**



Commute Distance (km)

- **Full Dataset:** The average commute distance for the entire population is approximately **7.44 km**.
- **Convenience Sample:** The average distance is **8.62 km**, which is notably higher than the population mean. This suggests a potential bias, as the first 100 records may not be representative of the overall population's commuting habits.
- **Simple Random Sample:** With an average of **7.25 km**, this sample's mean is the closest to the population average, making it the most accurate for estimating this variable.
- **Stratified Sample:** The average distance is **7.48 km**, also very close to the population mean.

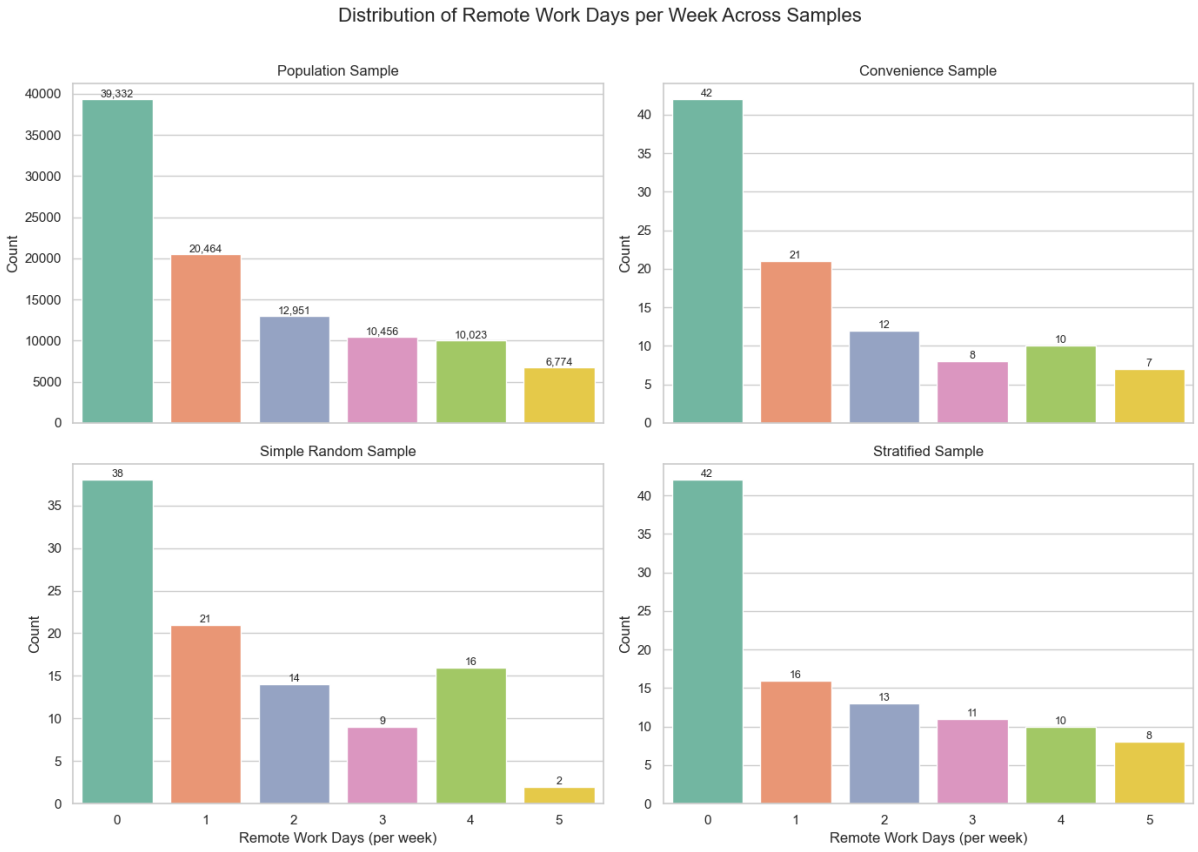
The histograms for commute distance show that the distributions of the simple random and stratified samples are the most similar to the full dataset, whereas the convenience sample's distribution is shifted slightly to the right, indicating a higher average.

Commute Time (minutes)

- **Full Dataset:** The average commute time for the population is **28.75 minutes**.
- **Convenience Sample:** This sample has an average commute time of **33.79 minutes**, which is a significant overestimation compared to the true population mean.
- **Simple Random Sample:** With an average of **28.98 minutes**, this sample is a very close estimate of the population average.
- **Stratified Sample:** The average is **34.48 minutes**, which, similar to the convenience sample, is a considerable overestimation.

For commute time, the simple random sample again proves to be the most representative, with a mean that is nearly identical to the population's.

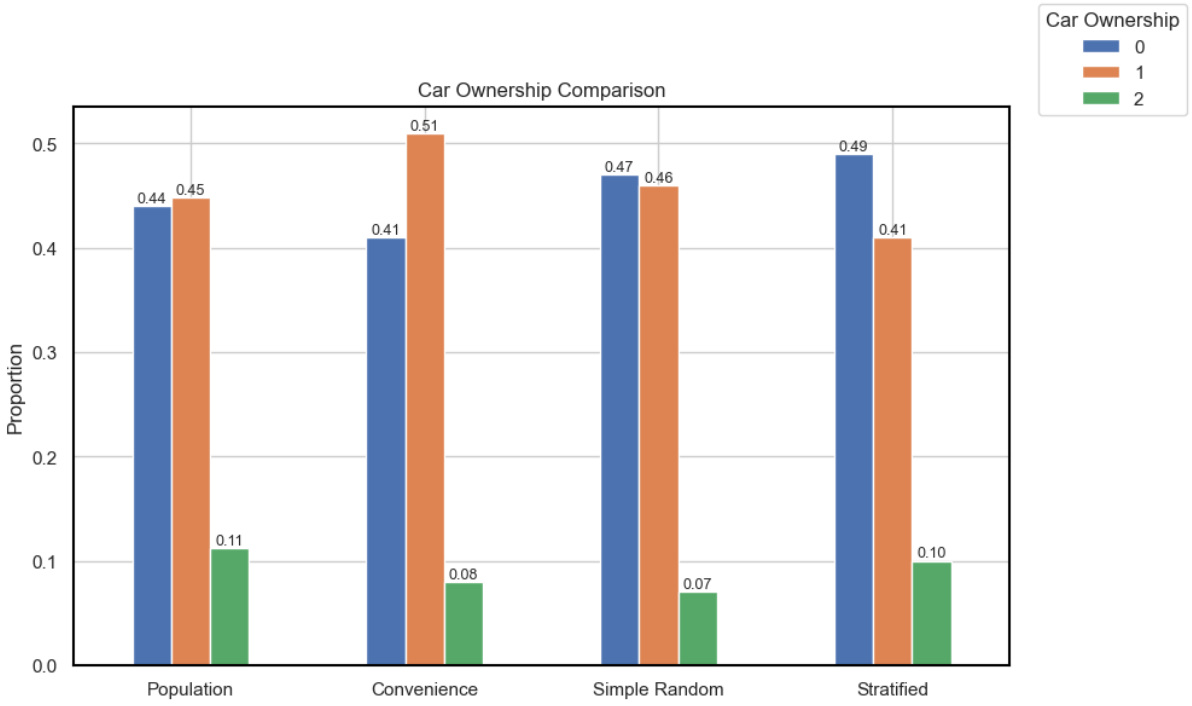
2. **Remote Work Days (per week)**



- **Full Dataset:** The average number of remote work days is 1.52 per week.
- **Convenience Sample:** The average is 1.44, a slight underestimation.
- **Simple Random Sample:** The average is 1.50, which is an excellent estimate.
- **Stratified Sample:** The average is 1.55, also a very good estimate.

All three samples provide a reasonably accurate estimate for remote work days, but the simple random sample is the most accurate of the three

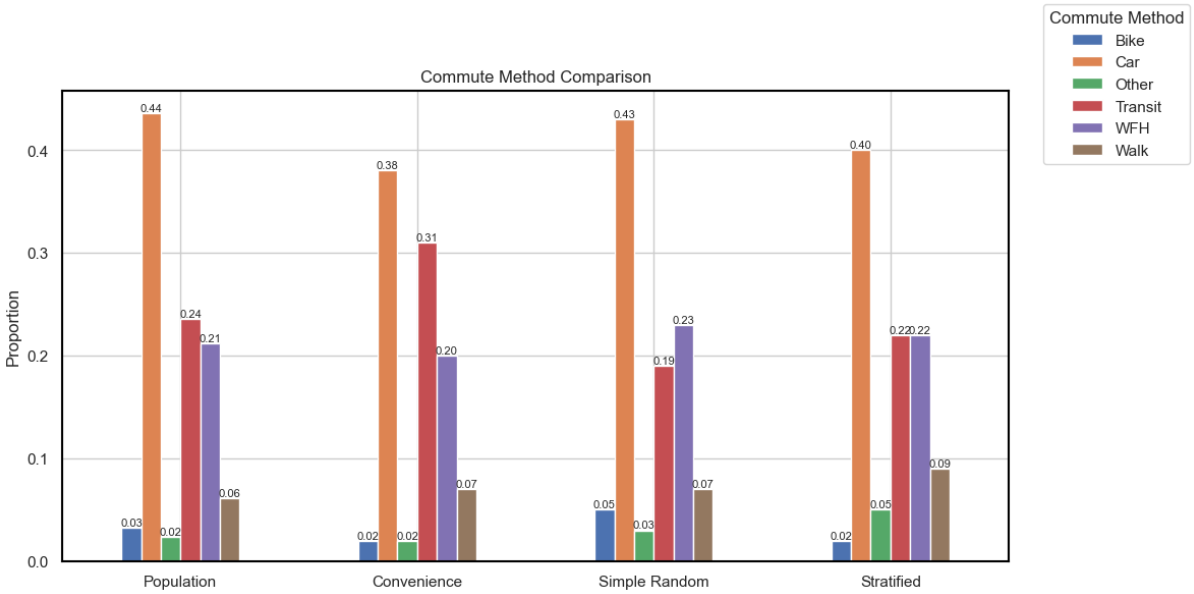
3. Car Ownership (0–2 cars)



The distribution of car ownership provides insights into how well each sample reflects the full population.

- **Full Dataset: 44.0%** have 0 cars, **44.8%** have 1, and **11.2%** have 2.
- **Simple Random Sample:** The distribution (**47%**, **46%**, **7%**) is the closest to the true population distribution, accurately representing the proportions of different car ownership levels.
- **Convenience Sample:** This sample significantly overestimates the proportion of people with 1 car (**51%**) and underestimates those with 0 cars (**41%**) and 2 cars (**8%**).
- **Stratified Sample:** This sample has the highest proportion of people with 0 cars (**49%**) and the lowest with 1 car (**41%**), showing a different distribution than the population.

4. **Commute Method (Walk, Bike, Transit, Car, Other, WFH)**



The distribution of commute methods reveals further disparities.

- **Full Dataset:** The most common method is **Car (43.6%)**, followed by **Transit (23.6%)** and **WFH (21.2%)**.
- **Simple Random Sample:** The proportions of this sample are the most similar to the full dataset's, especially for the top three methods (**Car: 43%**, **Transit: 19%**, **WFH: 23%**).
- **Convenience Sample:** This sample overrepresents those who use **Transit (31%)** and underrepresents those who use a **Car (38%)**.
- **Stratified Sample:** This sample overestimates the use of Other methods and underestimates car commutes.

The simple random sample's bar chart closely matches the distribution of the full dataset, confirming its accuracy in representing the population's commuting habits.

b. Which sampling method is most accurate in estimating average commute distance, commute time, and average remote workdays?

Based on the analysis, the **simple random sample** is the most accurate method for estimating the average commute distance, commute time, and remote work days.

Explanation:

A simple random sample gives every record in the dataset an equal chance of being selected. This method, when applied to a large and diverse population, tends to produce a sample that is highly representative of the population as a whole. As seen in the analysis,

the means, medians, and distributions of the key variables in the simple random sample were consistently the closest to the full dataset's values.

In contrast, the **convenience sample** is the least accurate because it is inherently biased, only including the first 100 records. This approach doesn't account for the diversity of the population and can lead to misleading conclusions, as demonstrated by its overestimation of average commute distance and time.

The **stratified sample**, while also an improvement over the convenience sample, did not perform as well as the simple random sample for these specific variables. Stratification is highly effective when you want to ensure the sample accurately represents a specific characteristic, such as income bracket. However, for variables like commute distance and time, which may not be directly tied to the stratification variable, a simple random sample can sometimes provide a more accurate overall picture