Lab 2 Homework: Summarizing Data Using the PSID Dataset

Background

The **Panel Study of Income Dynamics (PSID)** is a longitudinal study that collects demographic, economic, and social data on individuals and families in the United States. This dataset offers insights into trends and disparities in income, employment, and demographic changes over time.

You are provided with a subset of the PSID dataset, which includes the following variables:

- **Person_ID**: A unique identifier for each individual in the dataset.
- **Year**: The year of the observation.
- **race**: The race of the individual (*1 refers to White and 0 refers to Black*).
- **age**: The age of the individual.
- **sex**: The gender of the individual (*1 refers to Male and 0 refers to Female*).
- **log_wage**: The logarithm of the individual's wage.

Your goal is to explore and summarize this data, providing descriptive statistics and visualizations to interpret the relationships within it.

Practice Questions

1. Dataset Overview

(a) Import the `PSID.csv` dataset. How many observations and variables are in the dataset?

2. Variable Descriptions

(a) For each variable in the dataset, describe what it represents in reality (categorical or numerical) and its data type in R (*e.g., integer*). For example:

- **Variable:** `Person_ID`
- **Reality:** Categorical (Unique Identifier)
- **Type in R:** Integer

3. Recode Variables

(a) The `sex` variable may be encoded as numerical values (*e.g., "1" and "0"*). Convert it into a factor (in R) if necessary. How many individuals are Male? How many are Female?

4. Age Distribution

(a) Which graph is most appropriate to visualize the distribution of `age`?

(b) Produce the graph and describe the distribution of ages in the dataset.

5. Analyzing Wage Disparities

(a) Which graph is most appropriate to compare `log_wage` across different `race`s?

(b) Are there any potential outliers in `log_wage`? If so, identify them.

6. Descriptive Statistics

(a) Calculate the mean and standard deviation for `age` and `log_wage` for the overall sample. Then, calculate the same statistics separately for each `race` and `sex`. Create a summary table like the one below:

Table 1: Descriptive Statistics by Race and Gender

	Sample Size (n)	**Mean Age**	**Std. Age**	**Mean log_wage**	**Std. log_wage**
Overall					
Black					
White					
Male					
Female					