Project Introduction:

In today's data-driven world, businesses heavily rely on data analysis to make informed decisions. This project uses a real-world dataset containing employee information. Through this project, we will:

- Apply descriptive statistics to summarize the dataset and identify key trends.
- Perform hypothesis testing to validate assumptions and explore relationships between different variables.
- Use inferential statistics to draw conclusions about the employee population from the sample data.

Complete Data Dictionary for employee.csv

Column Name	Description	Data Type
EmployeeID	Unique identifier for each employee	Integer
Name	Full name of the employee	String
Age	Age of the employee	Integer
Gender	Gender of the employee (Male, Female, Other)	String
Department	Department where the employee works	String
JobTitle	Job title of the employee	String
Salary	Annual salary of the employee	Float
HireDate	Date the employee was hired	Date
YearsAtCompany	Number of years the employee has worked at the company	Integer
EducationLevel	Highest level of education attained (e.g., Bachelor's, Master's, PhD)	String
PerformanceRating	Performance rating of the employee (scale 1-5)	Integer
MaritalStatus	Marital status of the employee	String
City	City where the employee is based	String

Column Name	Description	Data Type
State	State where the employee is based	String
Country	Country where the employee is based	String
AgeGroup	Categorized age group of the employee (e.g., 20-29, 30-39)	String
SalaryBand	Categorized salary range (e.g., 40k-60k, 60k-80k)	String
PromotionStatus	Whether the employee has been promoted (Yes/No)	String
RemoteWork	Whether the employee works remotely (Yes/No)	String
LastPromotionDate	Date of the employee's last promotion	Date

Questions for the Project

Descriptive Statistics

1. What is the mean, median, and mode of employee ages?
2. Calculate the mean, median, and range of salaries.
3. What is the standard deviation of employee salaries, and what does it indicate?
4. How many employees are there in each department? Present this data in a frequency table.
5. What is the gender distribution of employees in the dataset?

Hypothesis Testing

6. Test the hypothesis: "The average salary of male employees is significantly different from that of female employees." Use an independent t-test.
7. Test the hypothesis: "The proportion of male employees in the Sales department is greater than 50%."
8. Use ANOVA to test whether there is a significant difference in salaries among different departments.
9. Perform a chi-square test to determine if there is a significant association between department and gender.

10. Test the hypothesis: "Employees with more than 10 years at the company have a higher average salary than those with 10 years or less."
11. Test the hypothesis: "The average age of employees in the IT department is higher than in the Marketing department."
12. Test the hypothesis: "Employees with a master's degree earn more on average than those with a bachelor's degree."
13. Test the hypothesis: "The proportion of employees working remotely is different across different states."
14. Test the hypothesis: "The performance rating is independent of the number of years at the company."
15. Test the hypothesis: "Married employees have a higher average performance rating than unmarried employees."

Inferential Statistics

16. Calculate the 95% confidence interval for the mean salary.
17. Determine the 95% confidence interval for the average years at the company.
18. Calculate the correlation coefficient between age and salary and interpret its meaning.
19. Estimate the proportion of employees who have been at the company for more than 5 years and calculate a 95% confidence interval for this proportion.
20. Test the hypothesis: "The average salary differs significantly between employees in urban and rural areas."