Statistics Real World Questions

Problem Statement:

Title: Data Cleaning and Transformation

You are given a dataset containing information about employees in a CSV file. The dataset contains the following columns: `EmployeeID`, `Name`, `Department`, `Salary`, and `JoinDate`. Your task is to clean and transform the data based on the following requirements:

1. Remove any duplicate rows from the dataset.
2. Fill any missing values in the `Salary` column with the average salary of the department to which the employee belongs.
3. Convert the `JoinDate` column to a `datetime` format.
4. Create a new column `Experience` that calculates the number of years each employee has been with the company, based on the `JoinDate` column and assuming the current year is 2024.

Task:

Write a function `clean_data` that takes a DataFrame as input and performs the above data cleaning and transformation tasks. The function should return the cleaned and transformed DataFrame.

Input Format:

- A CSV file with the following columns: `EmployeeID`, `Name`, `Department`, `Salary`, and `JoinDate`.

Constraints:

- The `EmployeeID` column contains unique integer identifiers.
- The `Name` column contains string values.
- The `Department` column contains string values.
- The `Salary` column contains float values, but some values might be missing (NaN).
- The `JoinDate` column contains string values in the format `YYYY-MM-DD`, but some values might be missing (NaN).

Output Format:

- A cleaned and transformed DataFrame with the same columns as the input, plus an additional `Experience` column.

Example:

Input:

EmployeeID	Name	Department	Salary	JoinDate
1	Shane Chang	Sales	70000	4/8/2024
2	Zachary Wright	HR	50000	########
3	Jennifer Moreno	Engineering	70000	########
4	Keith Vargas	Sales		########
5	Eugene Craig DDS	Sales	80000	########
6	Patricia Howell	Engineering	70000	########
7	Catherine Ramos	Engineering	80000	########
8	Steven Mcclain	HR	80000	########
9	Harry Carter	Marketing	70000	9/2/2017
10	William Turner	HR	90000	########
11	Melinda Reyes	Finance	70000	7/7/2016
12	Albert Nelson	Finance		########

Output:

EmployeeID	Name	Department	Salary	JoinDate	Experience
1	Shane Chang	Sales	70000	4/8/2024	0
2	Zachary Wright	HR	50000	#######2	8
3	Jennifer Moreno	Engineering	70000	#######2	7
4	Keith Vargas	Sales	67142.86	#######2	6
5	Eugene Craig DDS	Sales	80000	#######2	6
6	Patricia Howell	Engineering	70000	#######2	1
7	Catherine Ramos	Engineering	80000	#######2	10
8	Steven Mcclain	HR	80000	#######2	5
9	Harry Carter	Marketing	70000	9/2/2017	7
10	William Turner	HR	90000	#######2	3
11	Melinda Reyes	Finance	70000	7/7/2016	8
12	Albert Nelson	Finance	74761.9	#######2	5

Python Code Solution:

```
import pandas as pd
from datetime import datetime

def clean_data(df):
    # Remove duplicate rows
    df = df.drop_duplicates()

    # Fill missing values in the 'Salary' column with the average salary of
the department
    df['Salary'] = df.groupby('Department')['Salary'].transform(lambda x:
x.fillna(x.mean()))

    # Convert 'JoinDate' column to datetime format
    df['JoinDate'] = pd.to_datetime(df['JoinDate'])

    # Create 'Experience' column
    current_year = 2024
    df['Experience'] = current_year - df['JoinDate'].dt.year

    return df

def main():
    # Load CSV file
    csv_file = 'input_employees.csv'
    df = pd.read_csv(csv_file)

    # Clean and transform the data
    cleaned_df = clean_data(df)

    # Print results
    print(cleaned_df)

if __name__ == "__main__":
    main()
```

Test Cases :

Input : input_employees.csv

Output : if output = print(cleaned_employees.csv)

Problem Statement:

Title: Univariate Analysis of Employee Salaries

You are given a dataset containing information about employees in a CSV file. The dataset contains the following columns: `EmployeeID`, `Name`, `Department`, `Salary`, `JoinDate`, and `Experience`. Your task is to perform a basic univariate analysis on the `Salary` column.

Task:

Write a function `salary_analysis` that takes a DataFrame as input and performs the following tasks:

1. Compute the mean salary.
2. Compute the median salary.
3. Compute the standard deviation of the salary.
4. Compute the minimum and maximum salary.
5. Return these statistics as a dictionary.

Input Format:

- A CSV file with the following columns: `EmployeeID`, `Name`, `Department`, `Salary`, `JoinDate`, and `Experience`.

Constraints:

- The `EmployeeID` column contains unique integer identifiers.
- The `Name` column contains string values.
- The `Department` column contains string values.
- The `Salary` column contains float values.
- The `JoinDate` column contains datetime values.
- The `Experience` column contains integer values.

Output Format:

- A dictionary containing the mean, median, standard deviation, minimum, and maximum salary.

Example:

Input:

```
EmployeeID,Name,Department,Salary,JoinDate,Experience
1,Alice,HR,50000,2018-05-01,5
2,Bob,Engineering,75000,2017-08-15,6
```

```
3,Charlie,HR,60000,2019-03-20,4
4,David,Engineering,80000,2016-11-23,7
5,Eva,Marketing,45000,2020-02-01,3
6,Frank,Marketing,50000,2021-07-11,2
```

Output:

```
{
    'mean_salary': 60000.0,
    'median_salary': 55000.0,
    'std_salary': 14491.376746189438,
    'min_salary': 45000.0,
    'max_salary': 80000.0
}
```

Sample Code Solution:

Question: Basic Correlation Analysis

Problem Statement:

You have been given a cleaned dataset of employees in the file `cleaned_employees.csv`. This dataset contains information about employees, including their salary, department, join date, and experience.

Your task is to analyze the correlation between the employees' experience and their salary.

Input Format:

- A CSV file named `cleaned_employees.csv` with the following columns:
  - `EmployeeID`: A unique identifier for each employee.
  - `Name`: The name of the employee.
  - `Department`: The department where the employee works.
  - `Salary`: The salary of the employee.
  - `JoinDate`: The date the employee joined the company.
  - `Experience`: The number of years of experience the employee has.

Output Format:

- Compute the Pearson correlation coefficient between the `Experience` and `Salary` columns.
- Print the result.

Sample Input:

```
EmployeeID,Name,Department,Salary,JoinDate,Experience
1,Shane Chang,Sales,70000,4/8/2024,0
```

```
2,Zachary Wright,HR,50000,12/30/2016,8
3,Jennifer Moreno,Engineering,70000,2/21/2017,7
4,Keith Vargas,Sales,67142.85714,3/22/2018,6
5,Eugene Craig DDS,Sales,80000,11/29/2018,6
6,Patricia Howell,Engineering,70000,11/8/2023,1
7,Catherine Ramos,Engineering,80000,12/29/2014,10
8,Steven Mcclain,HR,80000,9/20/2019,5
9,Harry Carter,Marketing,70000,9/2/2017,7
10,William Turner,HR,90000,9/21/2021,3
```

Expected Output:

```
Pearson correlation coefficient between Experience and Salary:
```

Python Code to Solve the Problem

Question: Inferential Statistics - Confidence Interval for Mean Salary

Problem Statement:

You have been given a cleaned dataset of employees in the file `cleaned_employees.csv`. This dataset contains information about employees, including their salary, department, join date, and experience.

Your task is to compute the 95% confidence interval for the mean salary of the employees. Assume that the population standard deviation is unknown.

Input Format:

- A CSV file named `cleaned_employees.csv` with the following columns:
  - `EmployeeID`: A unique identifier for each employee.
  - `Name`: The name of the employee.
  - `Department`: The department where the employee works.
  - `Salary`: The salary of the employee.
  - `JoinDate`: The date the employee joined the company.
  - `Experience`: The number of years of experience the employee has.

Output Format:

- Print the lower and upper bounds of the 95% confidence interval for the mean salary.

Sample Input:

```
EmployeeID,Name,Department,Salary,JoinDate,Experience
1,Shane Chang,Sales,70000,4/8/2024,0
2,Zachary Wright,HR,50000,12/30/2016,8
3,Jennifer Moreno,Engineering,70000,2/21/2017,7
4,Keith Vargas,Sales,67142.85714,3/22/2018,6
```

```
5,Eugene Craig DDS,Sales,80000,11/29/2018,6
6,Patricia Howell,Engineering,70000,11/8/2023,1
7,Catherine Ramos,Engineering,80000,12/29/2014,10
8,Steven Mcclain,HR,80000,9/20/2019,5
9,Harry Carter,Marketing,70000,9/2/2017,7
10,William Turner,HR,90000,9/21/2021,3
```

Expected Output:

```
95% confidence interval for mean salary:
```

Python Code to Solve the Problem

Question: Hypothesis Testing - Salary Differences Between Departments

Problem Statement:

You have been given a cleaned dataset of employees in the file `cleaned_employees.csv`. This dataset contains information about employees, including their salary, department, join date, and experience.

You want to test whether there is a significant difference in the average salaries between two departments: `Sales` and `Engineering`.

Task: Perform a two-sample t-test to determine if there is a statistically significant difference in the mean salaries between employees in the `Sales` and `Engineering` departments.

Hypotheses:

- **Null Hypothesis (H0):** There is no significant difference in the average salaries between the `Sales` and `Engineering` departments.
- **Alternative Hypothesis (H1):** There is a significant difference in the average salaries between the `Sales` and `Engineering` departments.

Input Format:

- A CSV file named `cleaned_employees.csv` with the following columns:
  - `EmployeeID`: A unique identifier for each employee.
  - `Name`: The name of the employee.
  - `Department`: The department where the employee works.
  - `Salary`: The salary of the employee.
  - `JoinDate`: The date the employee joined the company.
  - `Experience`: The number of years of experience the employee has.

Output Format:

- Print the t-statistic and the p-value for the test.
- Based on the p-value, indicate whether you reject or fail to reject the null hypothesis at a significance level of 0.05.

Sample Input:

```
EmployeeID,Name,Department,Salary,JoinDate,Experience
1,Shane Chang,Sales,70000,4/8/2024,0
2,Zachary Wright,HR,50000,12/30/2016,8
3,Jennifer Moreno,Engineering,70000,2/21/2017,7
4,Keith Vargas,Sales,67142.85714,3/22/2018,6
5,Eugene Craig DDS,Sales,80000,11/29/2018,6
6,Patricia Howell,Engineering,70000,11/8/2023,1
7,Catherine Ramos,Engineering,80000,12/29/2014,10
8,Steven Mcclain,HR,80000,9/20/2019,5
9,Harry Carter,Marketing,70000,9/2/2017,7
10,William Turner,HR,90000,9/21/2021,3
```

Expected Output:

```
T-statistic:
P-value:
Based on the p-value, we (fail to reject)/(reject) the null
hypothesis at a significance level of 0.05.
```

Python Code to Solve the Problem