Employee Data Analysis Project for ABC Company

As a culminating project, you'll be working with a dataset from ABC company, consisting of 458 rows and 9 columns. The company requires a comprehensive report detailing information about their employees across various teams. Your tasks include preprocessing the dataset, analyzing the data, and presenting your findings graphically. Here's a breakdown of what you need to do:

Preprocessing:

Correct the data in the "height" column by replacing it with random numbers between 150 and 180. Ensure data consistency and integrity before proceeding with analysis. (1 mark)

Analysis Tasks:

1. Determine the distribution of employees across each team and calculate the percentage split relative to the total number of employees. (2 marks)

2. Segregate employees based on their positions within the company. (2 marks)

3. Identify the predominant age group among employees. (2 marks)

4. Discover which team and position have the highest salary expenditure. (2 marks)

5. Investigate if there's any correlation between age and salary and represent it visually. (2 marks)

Graphical Representation:

For each of the five analysis tasks, create appropriate visualizations to present your findings effectively. (5x2 = 10 marks)

Data Story:

Provide insights gained from the analysis, highlighting key trends, patterns, and correlations within the dataset. (3 marks)

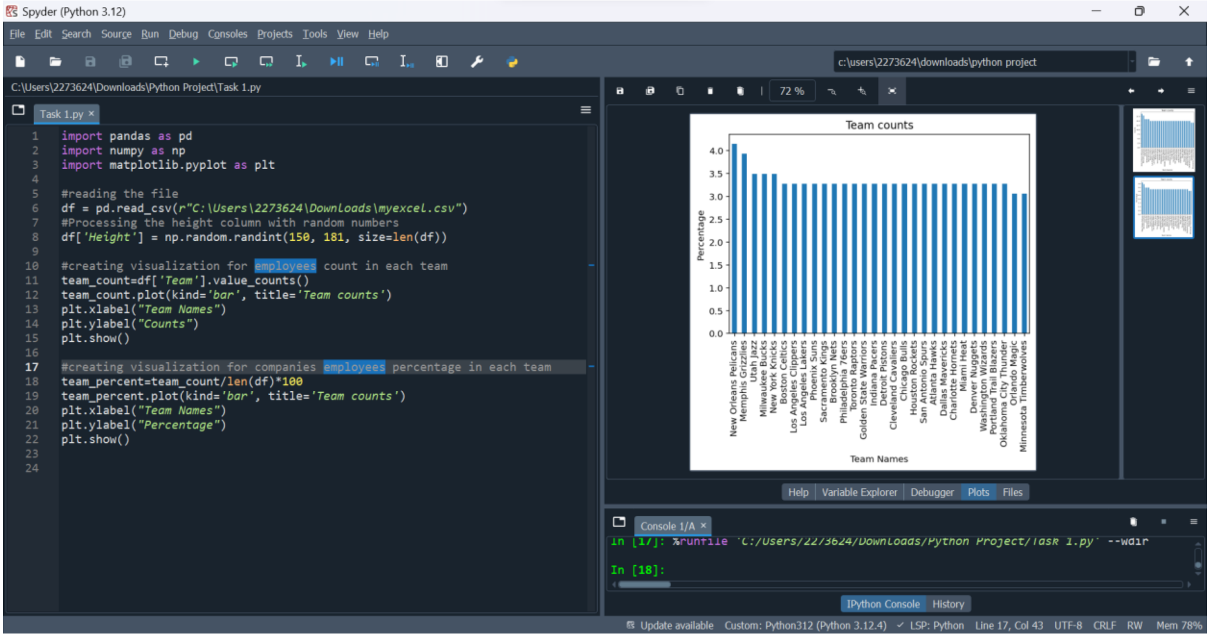
- ➔ This project focuses on analyzing ABC Company's employee dataset to derive insights into workforce distribution, salary patterns, and team dynamics. The project includes data preprocessing, statistical analysis, and visualization to present actionable findings for better organizational decision-making.

- ➔ Key Components of this project:
    - Preprocessing: Ensuring the data consistency and integrity
    - Analysis: Employee distribution, age group analysis, salary trends and correlation exploration
    - Visualization: Effective graphs and chart for analysis.
    - Data story and insights: Highlighting key trends and actionable conclusions.

Analysis Tasks:

1. Determine the distribution of employees across each team and calculate the percentage split relative to the total number of employees. (2 marks)
   a. Description: In this task we will generate the details of the employee in each team through visualization.
   b. Advantage: We can get to know which team has how many employees.
   c. Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

#reading the file
df = pd.read_csv(r"C:\Users\2273624\Downloads\myexcel.csv")
#Processing the height column with random numbers
df['Height'] = np.random.randint(150, 181, size=len(df))

#creating visualization for employees count in each team
team_count=df['Team'].value_counts()
team_count.plot(kind='bar', title='Team counts')
plt.xlabel("Team Names")
plt.ylabel("Counts")
plt.show()

#creating visualization for companies employees percentage in each
team
team_percent=team_count/len(df)*100
team_percent.plot(kind='bar', title='Team counts')
plt.xlabel("Team Names")
plt.ylabel("Percentage")
plt.show()
```

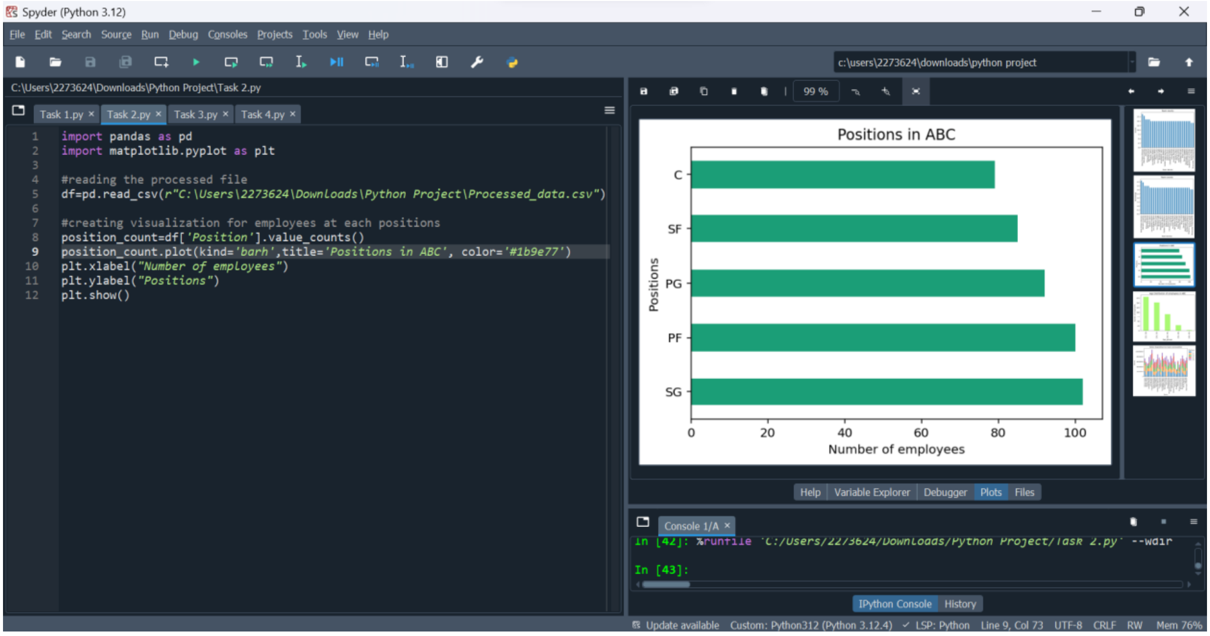


```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

#reading the file
df = pd.read_csv(r"C:\Users\2273624\Downloads\myexcel.csv")
#Processing the height column with random numbers
df['Height'] = np.random.randint(150, 181, size=len(df))

#creating visualization for employees count in each team
team_count=df['Team'].value_counts()
team_count.plot(kind='bar', title='Team counts')
plt.xlabel("Team Names")
plt.ylabel("Counts")
plt.show()

#creating visualization for companies employees percentage in each team
team_percent=team_count/len(df)*100
team_percent.plot(kind='bar', title='Team counts')
plt.xlabel("Team Names")
plt.ylabel("Percentage")
plt.show()
```

```
In [17]: %runfile 'C:/Users/2273624/Downloads/Python Project/Task 1.py' --wdir

In [18]:
```

2. Segregate employees based on their positions within the company. (2 marks)
    a. Description: In this task we are collecting the data and visualizing according to number of employees in each position in the company.
    b. Advantage: We will get the information about the number of employees working at a certain position.
    c. Code:

```
import pandas as pd
import matplotlib.pyplot as plt

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#creating visualization for employees at each positions
position_count=df['Position'].value_counts()
position_count.plot(kind='bar',title='Positions in ABC')
plt.xlabel("Positions")
plt.ylabel("Number of employees")
plt.show()
```

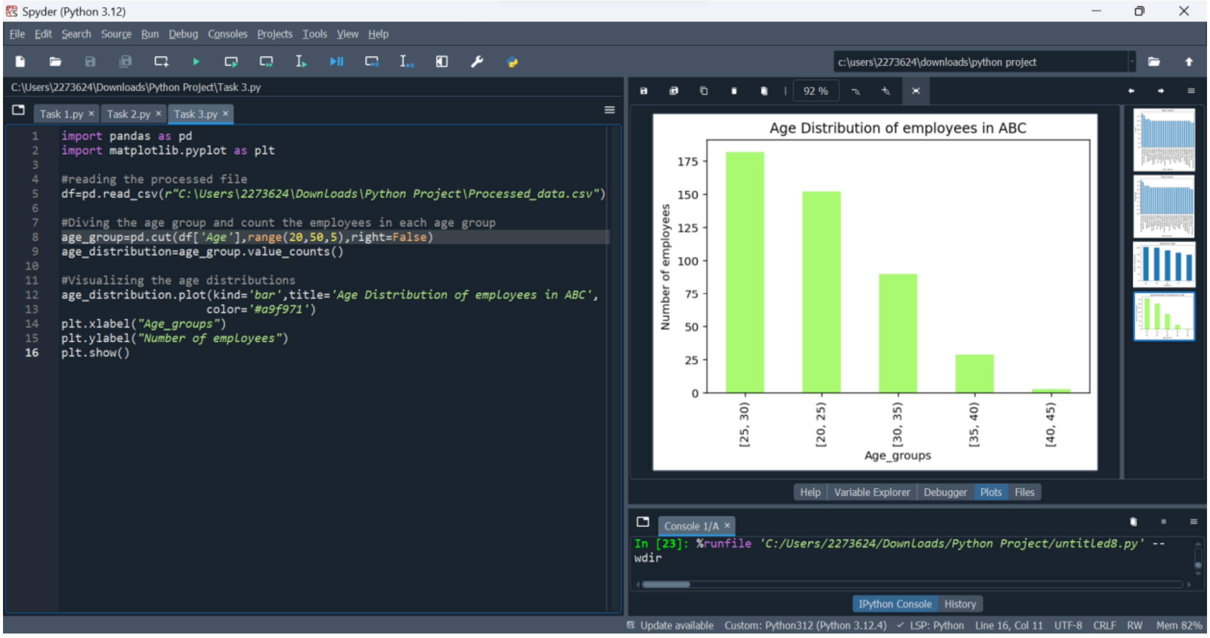


```
import pandas as pd
import matplotlib.pyplot as plt

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#creating visualization for employees at each positions
position_count=df['Position'].value_counts()
position_count.plot(kind='barh',title='Positions in ABC', color='#1b9e77')
plt.xlabel("Number of employees")
plt.ylabel("Positions")
plt.show()
```

```
In [42]: %runfile 'C:/Users/2273624/Downloads/Python Project/Task 2.py' --wdir

In [43]:
```

3. Identify the predominant age group among employees. (2 marks)
    a. Description: In this task we are dividing the employees with the specific age group. So that company can have a data about what are age groups working in the company.
    b. Advantage: Company can visualize the number of employees in a certain age group working in it and which age group is more or less.
    c. Code:

```
import pandas as pd
import matplotlib.pyplot as plt

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#Diving the age group and count the employees in each age group
age_group=pd.cut(df['Age'],range(20,50,5),right=False)
age_distribution=age_group.value_counts()

#Visualizing the age distributions
age_distribution.plot(kind='bar',title='Age Distribution of employees in ABC',
            color='#a9f971')
plt.xlabel("Age_groups")
plt.ylabel("Number of employees")
plt.show()
```



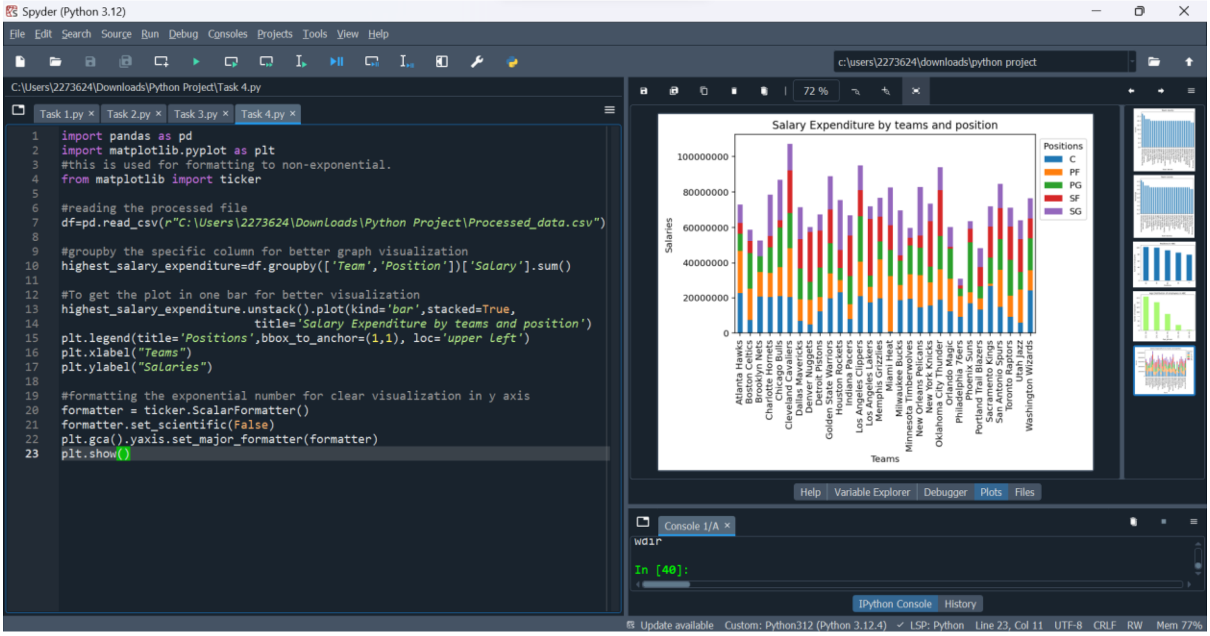
Spyder (Python 3.12)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\2273624\Downloads\Python Project\Task 3.py
Task 1.py
Task 2.py
Task 3.py
c:\users\2273624\downloads\python project
92 %

```
import pandas as pd
import matplotlib.pyplot as plt

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#Diving the age group and count the employees in each age group
age_group=pd.cut(df['Age'],range(20,50,5),right=False)
age_distribution=age_group.value_counts()

#Visualizing the age distributions
age_distribution.plot(kind='bar',title='Age Distribution of employees in ABC',
                     color='#a9f971')
plt.xlabel("Age_groups")
plt.ylabel("Number of employees")
plt.show()
```

```
In [23]: %runfile 'C:/Users/2273624/Downloads/Python Project/untitled8.py' --
wdir
```

4. Discover which team and position have the highest salary expenditure. (2 marks)
   a. Description: In this task we are visualizing those employees in a team at certain position in the company getting a specific salary.
   b. Advantage: From this data we can get the visualization of employees of a team is earning what about of money according to their position in the company.
   c. Code

```
import pandas as pd
import matplotlib.pyplot as plt
#this is used for formatting to non-exponential.
from matplotlib import ticker

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#groupby the specific column for better graph visualization
highest_salary_expenditure=df.groupby(['Team','Position'])['Salary'].sum()
#To get the plot in one bar for better visualization
highest_salary_expenditure.unstack().plot(kind='bar',stacked=True,
                title='Salary Expenditure by teams and position')
plt.legend(title='Positions',bbox_to_anchor=(1,1), loc='upper left')
plt.xlabel("Teams")
plt.ylabel("Salaries")

#formatting the exponential number for clear visualization in y axis
formatter = ticker.ScalarFormatter()
formatter.set_scientific(False)
plt.gca().yaxis.set_major_formatter(formatter)
plt.show()
```

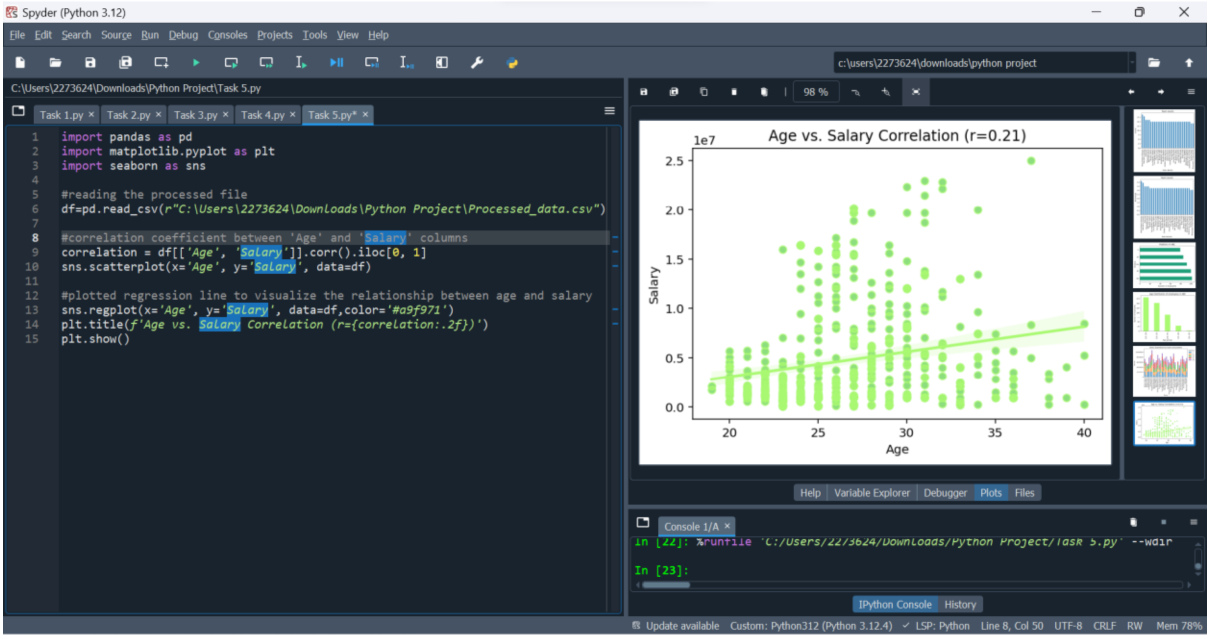


```
import pandas as pd
import matplotlib.pyplot as plt
#this is used for formatting to non-exponential.
from matplotlib import ticker

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#groupby the specific column for better graph visualization
highest_salary_expenditure=df.groupby(['Team','Position'])['Salary'].sum()

#To get the plot in one bar for better visualization
highest_salary_expenditure.unstack().plot(kind='bar',stacked=True,
                        title='Salary Expenditure by teams and position')
plt.legend(title='Positions',bbox_to_anchor=(1,1), loc='upper left')
plt.xlabel("Teams")
plt.ylabel("Salaries")

#formatting the exponential number for clear visualization in y axis
formatter = ticker.ScalarFormatter()
formatter.set_scientific(False)
plt.gca().yaxis.set_major_formatter(formatter)
plt.show()
```

```
wdir

In [40]:
```

5. Investigate if there's any correlation between age and salary and represent it visually. (2 marks)
    a. Description: In this task we are creating a correlation between age and salary column to visualize a relationship between two columns and a regression line for better visualization.
    b. Advantage: It provides a clear insight on decision making, supporting policies, salary structuring and career planning.
    c. Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#correlation coefficient between 'Age' and 'Salary' columns
correlation = df[['Age', 'Salary']].corr().iloc[0, 1]
sns.scatterplot(x='Age', y='Salary', data=df)

#plotted regression line to visualize the relationship between age and salary
sns.regplot(x='Age', y='Salary', data=df,color='#a9f971')
plt.title(f'Age vs. Salary Correlation (r={correlation:.2f})')
plt.show()
```


Spyder (Python 3.12)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\2273624\Downloads\Python Project\Task 5.py
Task 1.py Task 2.py Task 3.py Task 4.py Task 5.py*
c:\users\2273624\downloads\python project
98 %
Age vs. Salary Correlation (r=0.21)
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Salary
20
25
30
35
40
Age
Help Variable Explorer Debugger Plots Files
Console 1/A
IPython Console History
Update available Custom: Python312 (Python 3.12.4) LSP: Python Line 8, Col 50 UTF-8 CRLF RW Mem 78%

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#reading the processed file
df=pd.read_csv(r"C:\Users\2273624\Downloads\Python Project\Processed_data.csv")

#correlation coefficient between 'Age' and 'Salary' columns
correlation = df[['Age', 'Salary']].corr().iloc[0, 1]
sns.scatterplot(x='Age', y='Salary', data=df)

#plotted regression line to visualize the relationship between age and salary
sns.regplot(x='Age', y='Salary', data=df,color='#a9f971')
plt.title(f'Age vs. Salary Correlation (r={correlation:.2f})')
plt.show()
```

```
In [22]: %runfile 'C:/Users/2273624/Downloads/Python Project/Task 5.py' --wdir

In [23]:
```

Graphical Representation:

For each of the five analysis tasks, create appropriate visualizations to present your findings effectively. (5x2 = 10 marks)

1. Visualization for task 1:

Fig.1: Count in each team

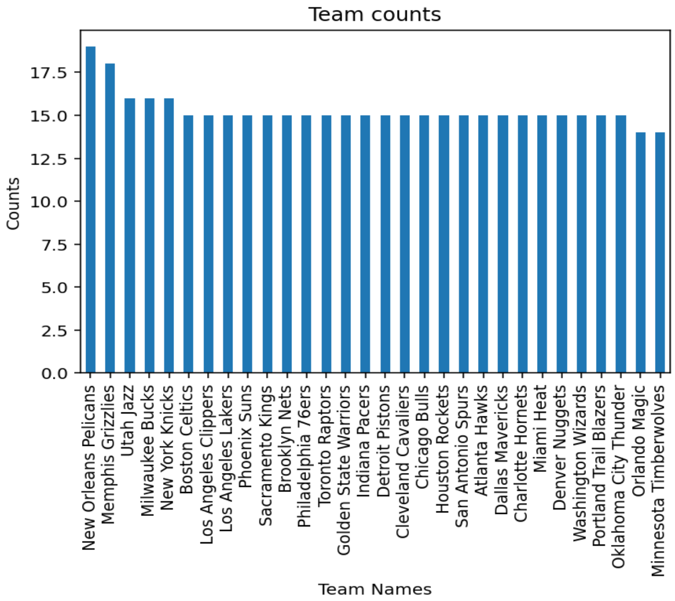
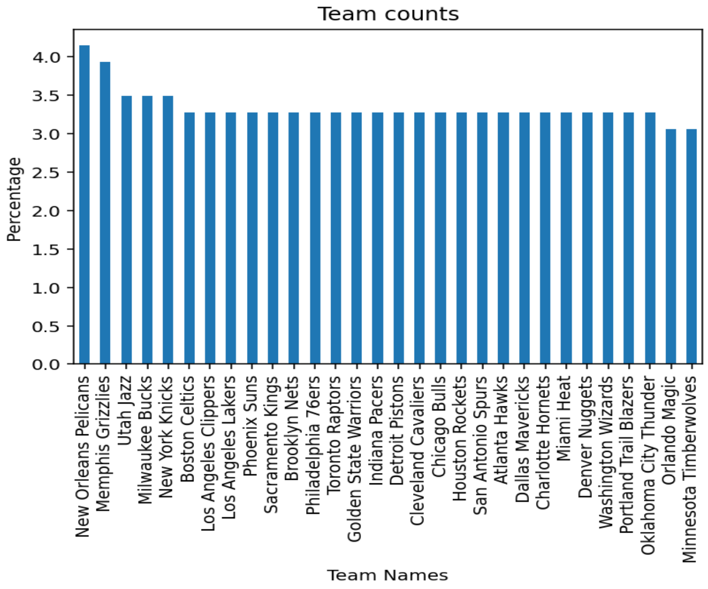
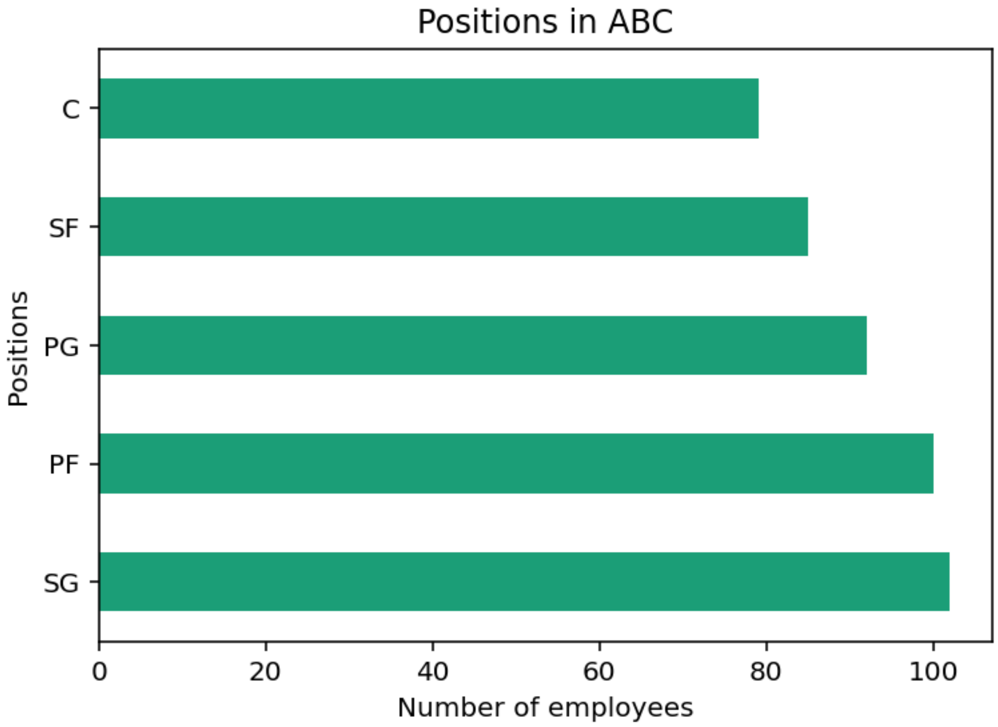


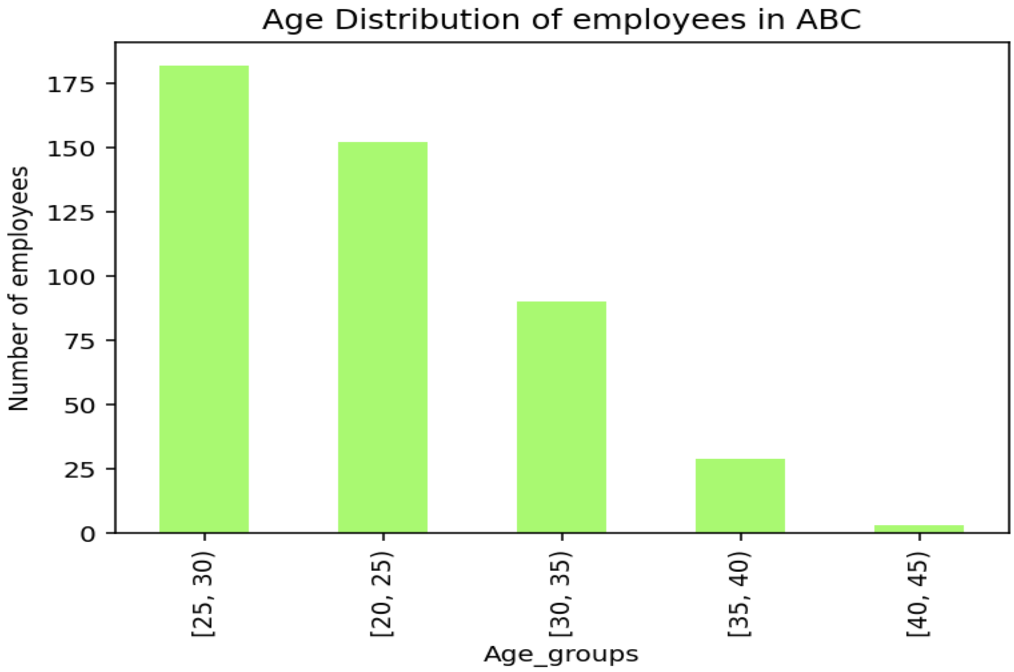
Fig.2: Percentage in each team



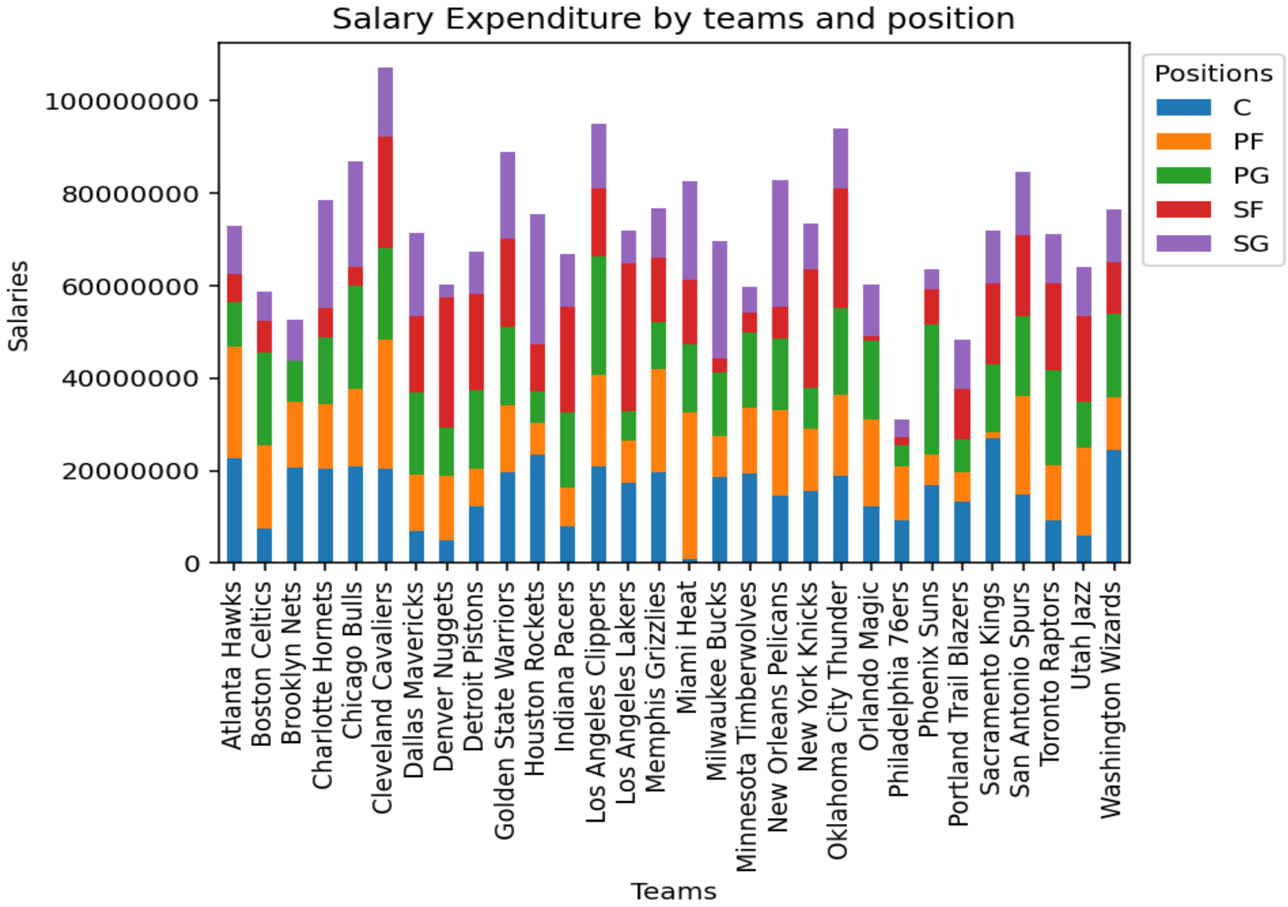
2. Visualization for task 2:



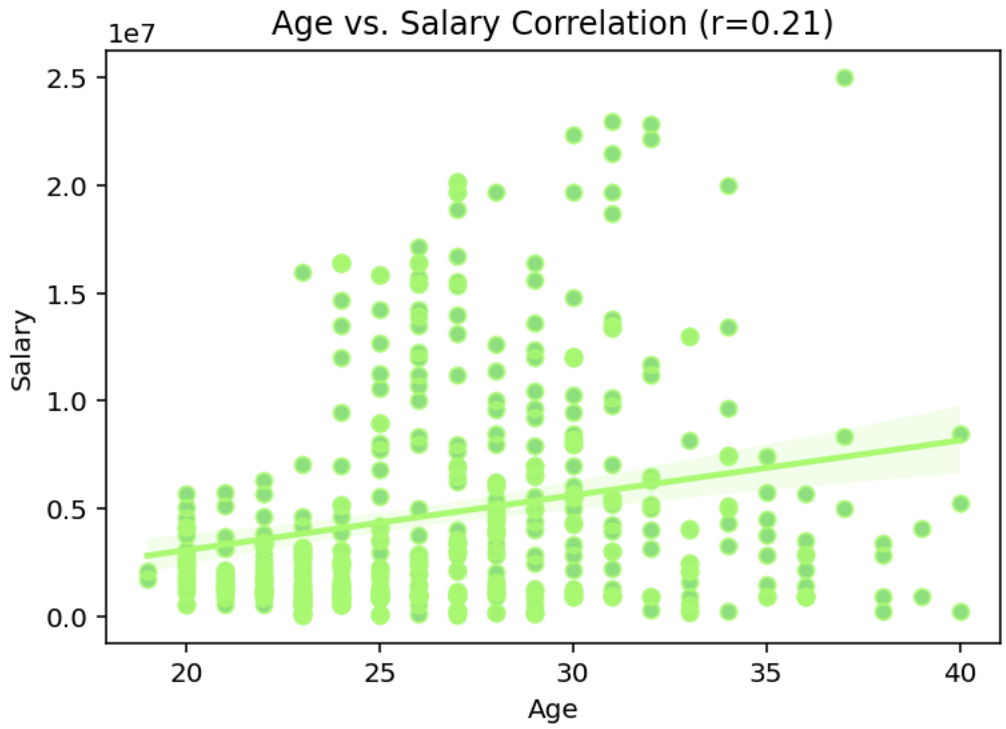
3. Visualization for task 3:



4. Visualization for task 4:



5. Visualization for task 5:



Data Story:

Provide insights gained from the analysis, highlighting key trends, patterns, and correlations within the dataset. (3 marks)

1. The largest team in the company is **New Orleans Pelicans** with **19** employees which is around **4.2%** of total workforce, indicating its significant contribution to the company's operations.
2. The position with the highest headcount is **SG** followed by **PF** and **PG**, which reflects the company's focus on specific skillset or role focus.
3. The most common age group among employees is **25–30** years followed by **20-25** years, suggesting the workforce is predominantly **experienced professionals** in their prime working years and the second common group is likely to be fresher groups helping in building the company and its core values.
4. The **Cleveland Cavaliers**, **Los Angeles Clippers**, **Oklahoma City Thunder** are the teams incurs the highest salary expenditure, possibly due to its critical role in achieving company objectives and having specialized expertise in specific skillsets.
   Among positions, **PF** followed by **SF** demands the highest salary allocation, emphasizing its value to the organization and potential of skilled professionals in this role.
5. The correlation between age and salary **positive**, with a value of **r=0.21**.