fds-assignment1

September 2, 2024

```
[7]: import pandas as pd


     df = pd.read_csv('/content/assign_1 - assign_1.csv')
     df.head()
```

```
[7]:    Company   Age  Salary     Place Country  Gender
     0      TCS  20.0     NaN   Chennai   India       0
     1  Infosys  30.0     NaN    Mumbai   India       0
     2      TCS  35.0  2300.0  Calcutta   India       0
     3  Infosys  40.0  3000.0     Delhi   India       0
     4      TCS  23.0  4000.0    Mumbai   India       0
```

```
[8]: df.shape
```

```
[8]: (148, 6)
```

```
[10]: df.isnull()
```

```
[10]:      Company    Age  Salary  Place  Country  Gender
      0      False  False    True  False    False   False
      1      False  False    True  False    False   False
      2      False  False   False  False    False   False
      3      False  False   False  False    False   False
      4      False  False   False  False    False   False
      ..       …      …      …      …       …      …
      143    False  False   False  False    False   False
      144    False  False   False  False    False   False
      145    False  False   False  False    False   False
      146    False  False   False  False    False   False
      147    False  False   False  False    False   False

      [148 rows x 6 columns]
```

```
[11]: df.isnull().sum()
```

```
[11]: Company     8
      Age        18
      Salary     24
      Place      14
      Country     0
      Gender      0
      dtype: int64
```

```
[12]: df.isnull().sum().sum()
```

```
[12]: 64
```

```
[14]: ## FILLING NULL VALUES
```

```
[15]: df2 = df.fillna(value = 0)
      df2
```

```
[15]:      Company   Age  Salary     Place Country  Gender
      0        TCS  20.0     0.0   Chennai   India       0
      1    Infosys  30.0     0.0    Mumbai   India       0
      2        TCS  35.0  2300.0  Calcutta   India       0
      3    Infosys  40.0  3000.0     Delhi   India       0
      4        TCS  23.0  4000.0    Mumbai   India       0
      ..       …     …       …         …       …       …
      143      TCS  33.0  9024.0  Calcutta   India       1
      144  Infosys  22.0  8787.0  Calcutta   India       1
      145  Infosys  44.0  4034.0     Delhi   India       1
      146      TCS  33.0  5034.0    Mumbai   India       1
      147  Infosys  22.0  8202.0    Cochin   India       0

      [148 rows x 6 columns]
```

```
[17]: df2.isnull().sum().sum()
```

```
[17]: 0
```

```
[18]: df.isnull().sum().sum()
```

```
[18]: 64
```

```
[32]: # Filling the Null Values with the previous values
      df4 = df.ffill()
      df4
```

```
[32]:      Company   Age  Salary     Place Country  Gender
      0        TCS  20.0     NaN   Chennai   India       0
      1    Infosys  30.0     NaN    Mumbai   India       0
```

```
2        TCS  35.0  2300.0  Calcutta   India       0
3    Infosys  40.0  3000.0     Delhi   India       0
4        TCS  23.0  4000.0    Mumbai   India       0
..       …   …     …       …     …     …
143      TCS  33.0  9024.0  Calcutta   India       1
144  Infosys  22.0  8787.0  Calcutta   India       1
145  Infosys  44.0  4034.0     Delhi   India       1
146      TCS  33.0  5034.0    Mumbai   India       1
147  Infosys  22.0  8202.0    Cochin   India       0

[148 rows x 6 columns]
```

[34]:
```
# Filling the Null Values with the next values
df5 = df.bfill()
df5
```

[34]:
```
     Company   Age  Salary     Place Country  Gender
0        TCS  20.0  2300.0   Chennai   India       0
1    Infosys  30.0  2300.0    Mumbai   India       0
2        TCS  35.0  2300.0  Calcutta   India       0
3    Infosys  40.0  3000.0     Delhi   India       0
4        TCS  23.0  4000.0    Mumbai   India       0
..       …   …     …       …     …     …
143      TCS  33.0  9024.0  Calcutta   India       1
144  Infosys  22.0  8787.0  Calcutta   India       1
145  Infosys  44.0  4034.0     Delhi   India       1
146      TCS  33.0  5034.0    Mumbai   India       1
147  Infosys  22.0  8202.0    Cochin   India       0

[148 rows x 6 columns]
```

[37]:
```
df6 = df.ffill(axis = 1)
df6
```

[37]:
```
     Company   Age  Salary     Place Country Gender
0        TCS  20.0    20.0   Chennai   India      0
1    Infosys  30.0    30.0    Mumbai   India      0
2        TCS  35.0  2300.0  Calcutta   India      0
3    Infosys  40.0  3000.0     Delhi   India      0
4        TCS  23.0  4000.0    Mumbai   India      0
..       …   …     …       …     …    …
143      TCS  33.0  9024.0  Calcutta   India      1
144  Infosys  22.0  8787.0  Calcutta   India      1
145  Infosys  44.0  4034.0     Delhi   India      1
146      TCS  33.0  5034.0    Mumbai   India      1
147  Infosys  22.0  8202.0    Cochin   India      0
```

```
[148 rows x 6 columns]
```

[38]:

```
df7 = df.bfill(axis = 1)
df7
```

[38]:

```
    Company   Age   Salary     Place Country Gender
0       TCS  20.0  Chennai   Chennai   India      0
1   Infosys  30.0   Mumbai    Mumbai   India      0
2       TCS  35.0   2300.0  Calcutta   India      0
3   Infosys  40.0   3000.0     Delhi   India      0
4       TCS  23.0   4000.0    Mumbai   India      0
..      …     …        …        …       …      …
143     TCS  33.0   9024.0  Calcutta   India      1
144 Infosys  22.0   8787.0  Calcutta   India      1
145 Infosys  44.0   4034.0     Delhi   India      1
146     TCS  33.0   5034.0    Mumbai   India      1
147 Infosys  22.0   8202.0    Cochin   India      0

[148 rows x 6 columns]
```

[42]:

```
df8 = df.fillna({'Company':'abcd', 'Salary': 'defg'})
df8
```

[42]:

```
    Company   Age  Salary     Place Country  Gender
0       TCS  20.0    defg   Chennai   India       0
1   Infosys  30.0    defg    Mumbai   India       0
2       TCS  35.0  2300.0  Calcutta   India       0
3   Infosys  40.0  3000.0     Delhi   India       0
4       TCS  23.0  4000.0    Mumbai   India       0
..      …     …       …        …       …      …
143     TCS  33.0  9024.0  Calcutta   India       1
144 Infosys  22.0  8787.0  Calcutta   India       1
145 Infosys  44.0  4034.0     Delhi   India       1
146     TCS  33.0  5034.0    Mumbai   India       1
147 Infosys  22.0  8202.0    Cochin   India       0

[148 rows x 6 columns]
```

[48]:

```
# Filling the Null Values with the mean
df9 = df.fillna(value = df['Salary'].mean()) # Similar for median, min, max
df9
```

[48]:

```
    Company   Age       Salary     Place Country  Gender
0       TCS  20.0  5457.246575   Chennai   India       0
1   Infosys  30.0  5457.246575    Mumbai   India       0
2       TCS  35.0  2300.000000  Calcutta   India       0
3   Infosys  40.0  3000.000000     Delhi   India       0
```

```
4        TCS  23.0  4000.000000    Mumbai   India        0
..       …   …          …       …     …     …
143      TCS  33.0  9024.000000  Calcutta   India        1
144  Infosys  22.0  8787.000000  Calcutta   India        1
145  Infosys  44.0  4034.000000     Delhi   India        1
146      TCS  33.0  5034.000000    Mumbai   India        1
147  Infosys  22.0  8202.000000    Cochin   India        0

[148 rows x 6 columns]
```

[51]:

```
# Dropna() function
df10 = df.dropna() # Drops all rows having NULL Values
df10
```

[51]:

```
     Company   Age  Salary     Place Country  Gender
2        TCS  35.0  2300.0  Calcutta   India       0
3    Infosys  40.0  3000.0     Delhi   India       0
4        TCS  23.0  4000.0    Mumbai   India       0
5    Infosys  23.0  5000.0  Calcutta   India       0
6        TCS  23.0  6000.0   Chennai   India       1
..       …   …     …       …     …      …
143      TCS  33.0  9024.0  Calcutta   India       1
144  Infosys  22.0  8787.0  Calcutta   India       1
145  Infosys  44.0  4034.0     Delhi   India       1
146      TCS  33.0  5034.0    Mumbai   India       1
147  Infosys  22.0  8202.0    Cochin   India       0

[146 rows x 6 columns]
```

[54]:

```
# replace() function
import numpy as np
df12 = df.replace(to_replace = np.nan, value = 1234 )
df12
# we can replace any value using to_replace
```

[54]:

```
     Company   Age  Salary     Place Country  Gender
0        TCS  20.0  1234.0   Chennai   India       0
1    Infosys  30.0  1234.0    Mumbai   India       0
2        TCS  35.0  2300.0  Calcutta   India       0
3    Infosys  40.0  3000.0     Delhi   India       0
4        TCS  23.0  4000.0    Mumbai   India       0
..       …   …     …       …     …      …
143      TCS  33.0  9024.0  Calcutta   India       1
144  Infosys  22.0  8787.0  Calcutta   India       1
145  Infosys  44.0  4034.0     Delhi   India       1
146      TCS  33.0  5034.0    Mumbai   India       1
147  Infosys  22.0  8202.0    Cochin   India       0
```

```
[148 rows x 6 columns]
```

[57]:
```
df['Age'] = df['Age'].interpolate(method = 'linear')
df
```

[57]:
```
     Company   Age  Salary     Place Country  Gender
0        TCS  20.0     NaN   Chennai   India       0
1    Infosys  30.0     NaN    Mumbai   India       0
2        TCS  35.0  2300.0  Calcutta   India       0
3    Infosys  40.0  3000.0     Delhi   India       0
4        TCS  23.0  4000.0    Mumbai   India       0
..       …     …      …        …      …       …
143      TCS  33.0  9024.0  Calcutta   India       1
144  Infosys  22.0  8787.0  Calcutta   India       1
145  Infosys  44.0  4034.0     Delhi   India       1
146      TCS  33.0  5034.0    Mumbai   India       1
147  Infosys  22.0  8202.0    Cochin   India       0

[148 rows x 6 columns]
```

[63]:
```
df.describe()
```

[63]:
```
              Age       Salary      Gender
count  148.000000   146.000000  148.000000
mean    29.885135  5457.246575    0.222973
std     10.774449  2730.139189    0.417654
min      0.000000  1089.000000    0.000000
25%     22.000000  3030.000000    0.000000
50%     32.000000  5004.500000    0.000000
75%     36.250000  8309.250000    0.000000
max     54.000000  9876.000000    1.000000
```

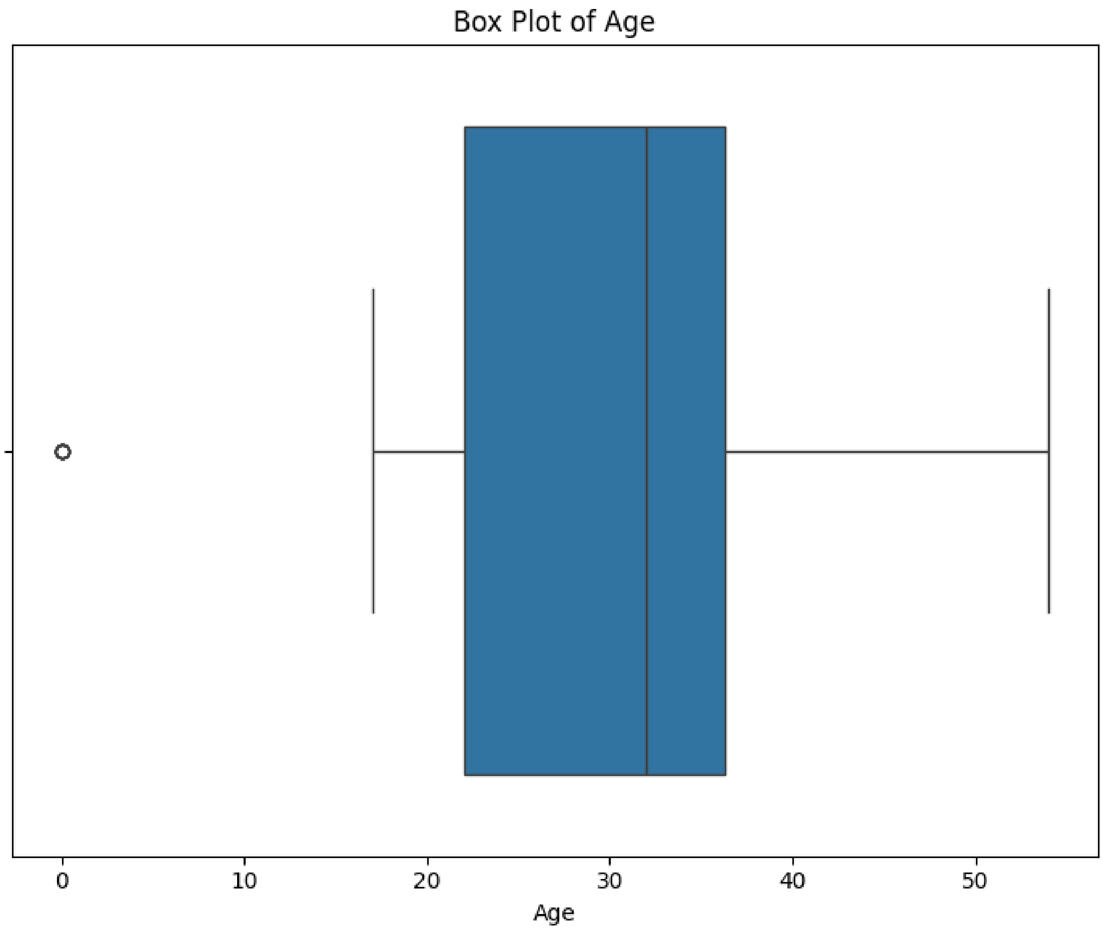
[64]:
```
df['Salary'].describe()
```

[64]:
```
count     146.000000
mean     5457.246575
std      2730.139189
min      1089.000000
25%      3030.000000
50%      5004.500000
75%      8309.250000
max      9876.000000
Name: Salary, dtype: float64
```

[85]:
```
# Outliers
import matplotlib.pyplot as plt
import seaborn as sns
```

[81]:

```
# Comparing
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
sns.boxplot(x=df['Age'])
plt.title('Box Plot of Age')



plt.tight_layout()
plt.show()
```

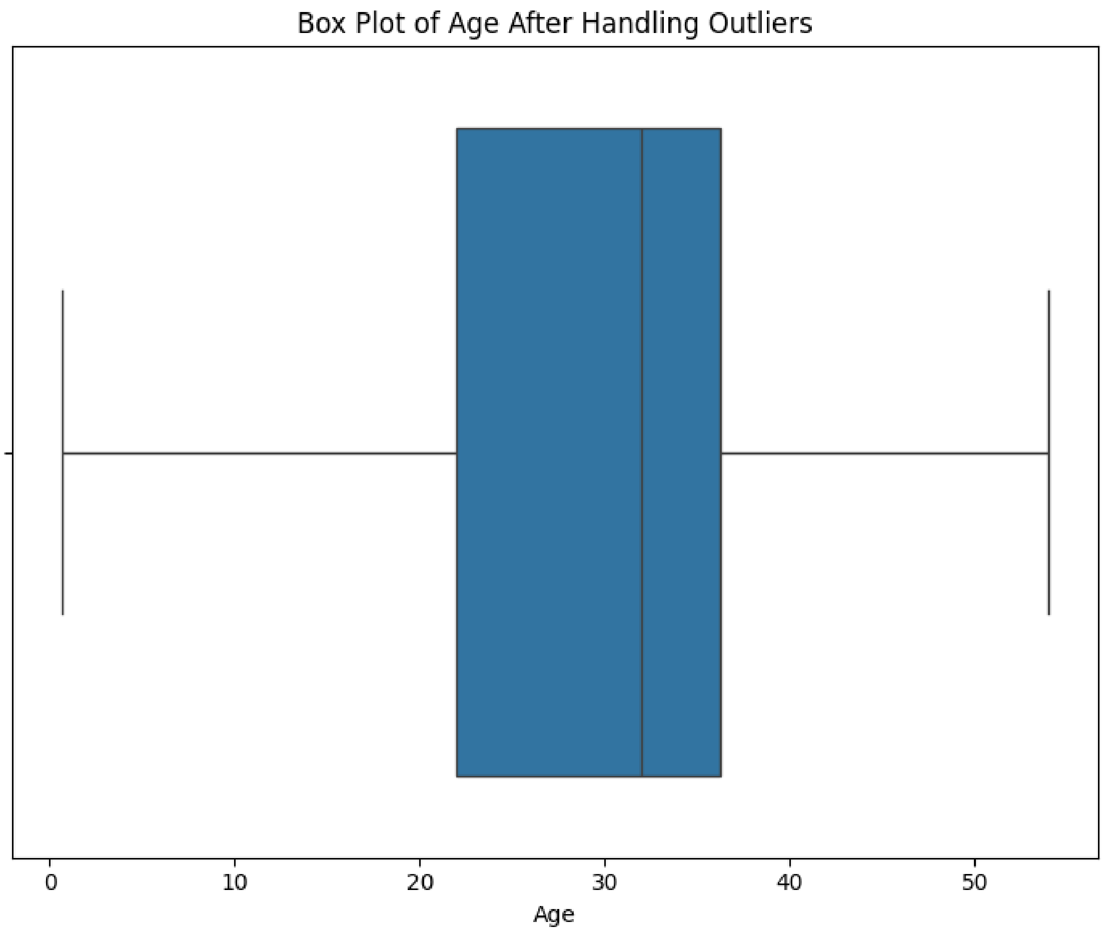


[82]:

```
# Handling OutLiers

def cap_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    df[column] = np.clip(df[column], lower_bound, upper_bound)

# Apply to Age
cap_outliers(df, 'Age')
```

[84]:

```
# Box plots after capping outliers
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
sns.boxplot(x=df['Age'])
plt.title('Box Plot of Age After Handling Outliers')


plt.tight_layout()
plt.show()
```



Box Plot of Age After Handling Outliers

[89]:

```
# Handle Inconsistent Formatting

df['Country'] = df['Country'].str.upper()

# spaces and first letter capital
df['Company'] = df['Company'].str.strip().str.title()
df['Place'] = df['Place'].str.strip().str.title()

print(df[['Age', 'Salary']].describe())

# Replace -ve values with NULL
df.loc[df['Age'] < 0, 'Age'] = pd.NA
df.loc[df['Salary'] < 0, 'Salary'] = pd.NA

# Fill the Missing Values
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Salary'].fillna(df['Salary'].median(), inplace=True)

# Remove duplicates
df.drop_duplicates(inplace=True)

# For consistency all values should be unique
print(df['Country'].unique())
print(df['Gender'].unique())
print(df['Company'].unique())
print(df['Place'].unique())
```

```
              Age       Salary
count  145.000000   145.000000
mean    29.936207  5462.531034
std     10.714007  2701.101591
min      0.625000  1089.000000
25%     22.000000  3030.000000
50%     32.000000  5004.500000
75%     36.000000  8202.000000
max     54.000000  9876.000000
['INDIA']
[0 1]
['Tcs' 'Infosys' 'Cts' 'Tata Consultancy Services' 'Congnizant'
 'Infosys Pvt Lmt']
['Chennai' 'Mumbai' 'Calcutta' 'Delhi' 'Podicherry' 'Cochin' 'Noida'
 'Hyderabad' 'Bhopal' 'Nagpur' 'Pune']
```

[90]:

```
# Handling Noise
# Removing whitespaces
df['Company'] = df['Company'].str.strip()
df['Place'] = df['Place'].str.strip()
```

[92]:
```
df.head()
```

[92]:
```
   Company   Age  Salary     Place Country  Gender
0      Tcs  20.0  5004.5   Chennai   INDIA       0
1  Infosys  30.0  5004.5    Mumbai   INDIA       0
2      Tcs  35.0  2300.0  Calcutta   INDIA       0
3  Infosys  40.0  3000.0     Delhi   INDIA       0
4      Tcs  23.0  4000.0    Mumbai   INDIA       0
```

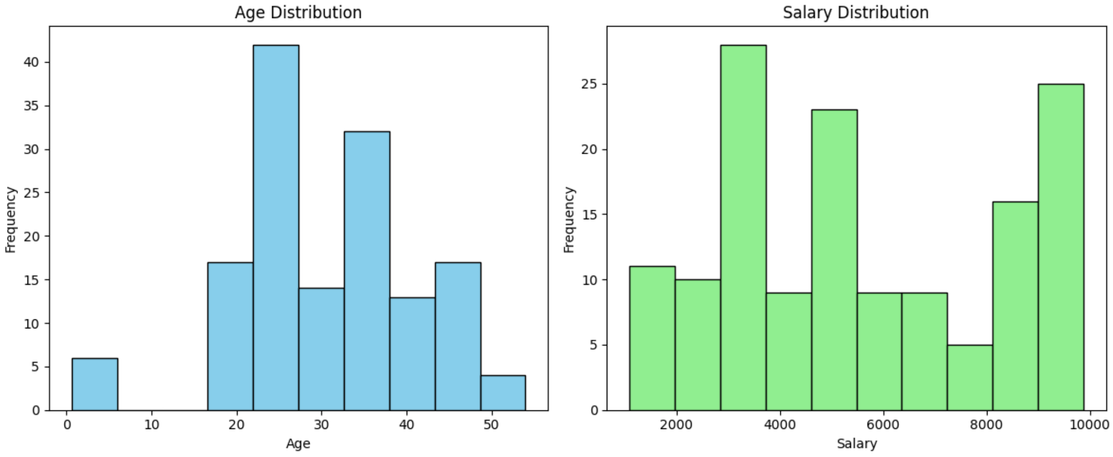
[93]:
```
# DATA EXPLORATION

print(df.describe())

print(df['Company'].value_counts())
print(df['Place'].value_counts())
print(df['Country'].value_counts())
print(df['Gender'].value_counts())
```

```
              Age       Salary      Gender
count  145.000000   145.000000  145.000000
mean    29.936207  5462.531034    0.220690
std     10.714007  2701.101591    0.416149
min      0.625000  1089.000000    0.000000
25%     22.000000  3030.000000    0.000000
50%     32.000000  5004.500000    0.000000
75%     36.000000  8202.000000    0.000000
max     54.000000  9876.000000    1.000000
Company
Tcs                          56
Infosys                      46
Cts                          37
Tata Consultancy Services     2
Congnizant                    2
Infosys Pvt Lmt               2
Name: count, dtype: int64
Place
Calcutta      46
Mumbai        35
Chennai       14
Delhi         14
Cochin        13
Noida          8
Hyderabad      8
```

```
Podicherry     3
Pune           2
Bhopal         1
Nagpur         1
Name: count, dtype: int64
Country
INDIA    145
Name: count, dtype: int64
Gender
0    113
1     32
Name: count, dtype: int64
```

[94]:

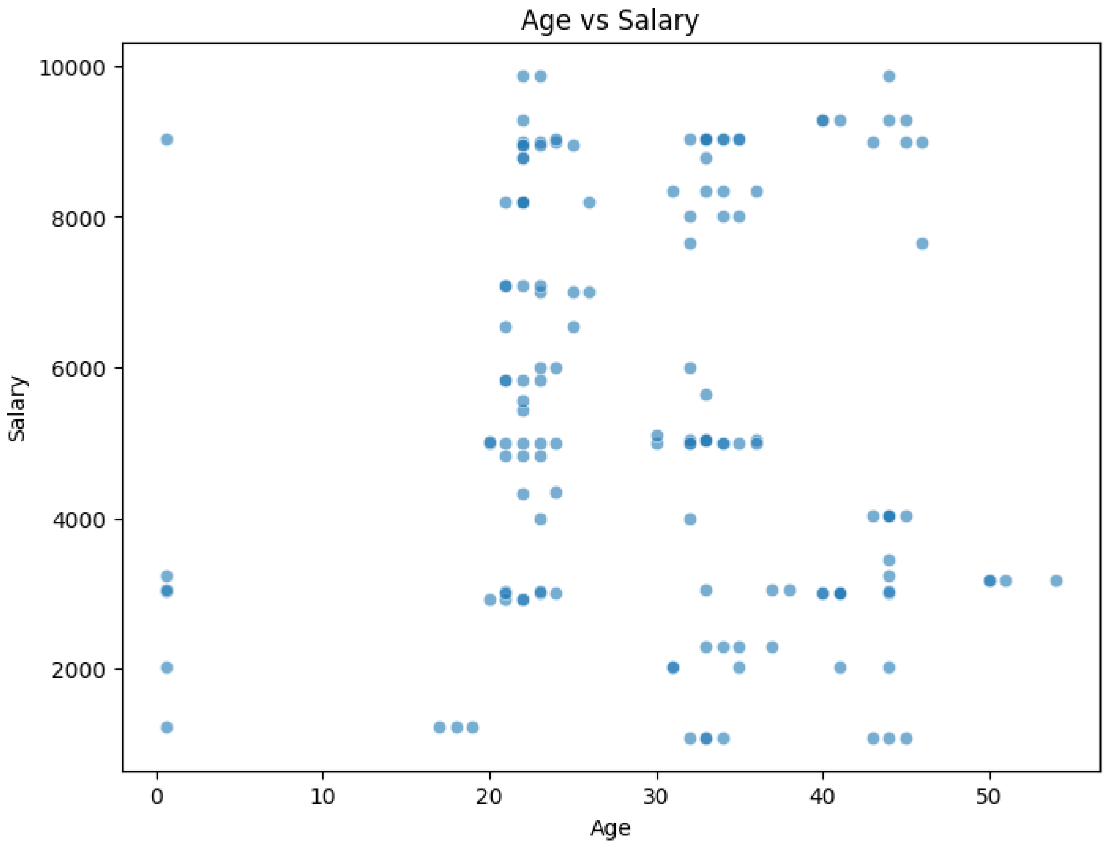
```
# Histograms
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.hist(df['Age'], bins=10, color='skyblue', edgecolor='black')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')

plt.subplot(1, 2, 2)
plt.hist(df['Salary'], bins=10, color='lightgreen', edgecolor='black')
plt.title('Salary Distribution')
plt.xlabel('Salary')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
```



[95]:

```
# Correlation matrix
correlation_matrix = df[['Age', 'Salary']].corr()
print(correlation_matrix)

# Scatter plot
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Age', y='Salary', data=df, alpha=0.6)
plt.title('Age vs Salary')
plt.xlabel('Age')
plt.ylabel('Salary')
plt.show()
```

```
             Age    Salary
Age     1.000000 -0.044161
Salary -0.044161  1.000000
```

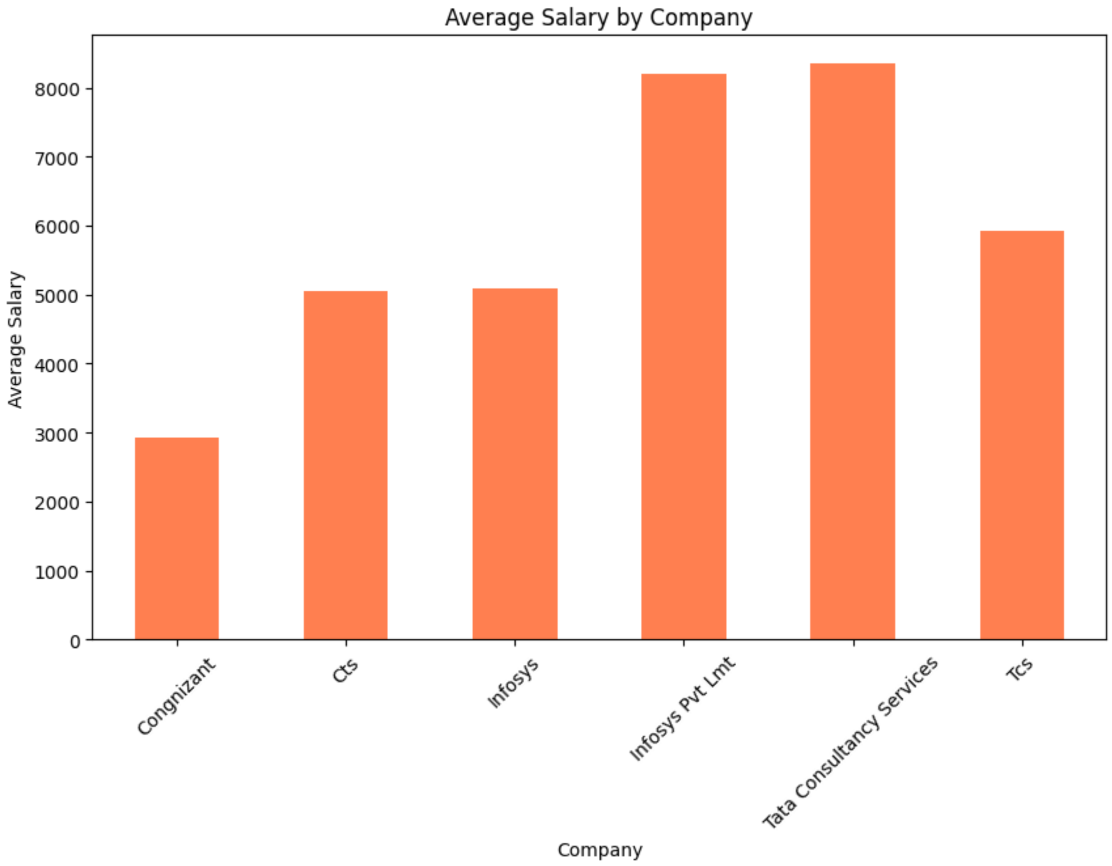


[96]:

```
# Average Salary by Company
avg_salary_by_company = df.groupby('Company')['Salary'].mean()
print(avg_salary_by_company)
```

```
avg_salary_by_country = df.groupby('Country')['Salary'].mean()
print(avg_salary_by_country)

# Bar chart
plt.figure(figsize=(10, 6))
avg_salary_by_company.plot(kind='bar', color='coral')
plt.title('Average Salary by Company')
plt.xlabel('Company')
plt.ylabel('Average Salary')
plt.xticks(rotation=45)
plt.show()
```

```
Company
Congnizant                   2934.000000
Cts                          5059.270270
Infosys                      5098.684783
Infosys Pvt Lmt              8202.000000
Tata Consultancy Services    8345.000000
Tcs                          5917.366071
Name: Salary, dtype: float64
Country
INDIA    5462.531034
Name: Salary, dtype: float64
```


Average Salary by Company
Average Salary
0
1000
2000
3000
4000
5000
6000
7000
8000
Congnizant
Cts
Infosys
Infosys Pvt Lmt
Tata Consultancy Services
Tcs
Company

```
[ ]:
```