Importing necessary libraries

In [37]:
```
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_ind
from scipy.stats import f_oneway
```

Reading Data

In [2]:
```
df = pd.read_csv("/kaggle/input/salary-by-job-title-and-country/Salary.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	Age	Gender	Education Level	Job Title	Years of Experience	Salary	Country	Race	Senior
0	32.0	Male	1	Software Engineer	5.0	90000.0	UK	White	0
1	28.0	Female	2	Data Analyst	3.0	65000.0	USA	Hispanic	0
2	45.0	Male	3	Manager	15.0	150000.0	Canada	White	1
3	36.0	Female	1	Sales Associate	7.0	60000.0	USA	Hispanic	0
4	52.0	Male	2	Director	20.0	200000.0	USA	Asian	0

In [4]:
```
df.shape
```

Out[4]:
```
(6684, 9)
```

In [5]:
```
df.isna().sum()
```

Out[5]:
```
Age                   0
Gender                0
Education Level       0
Job Title             0
Years of Experience   0
Salary                0
Country               0
Race                  0
Senior                0
dtype: int64
```

In [7]:
```
df.describe()
```

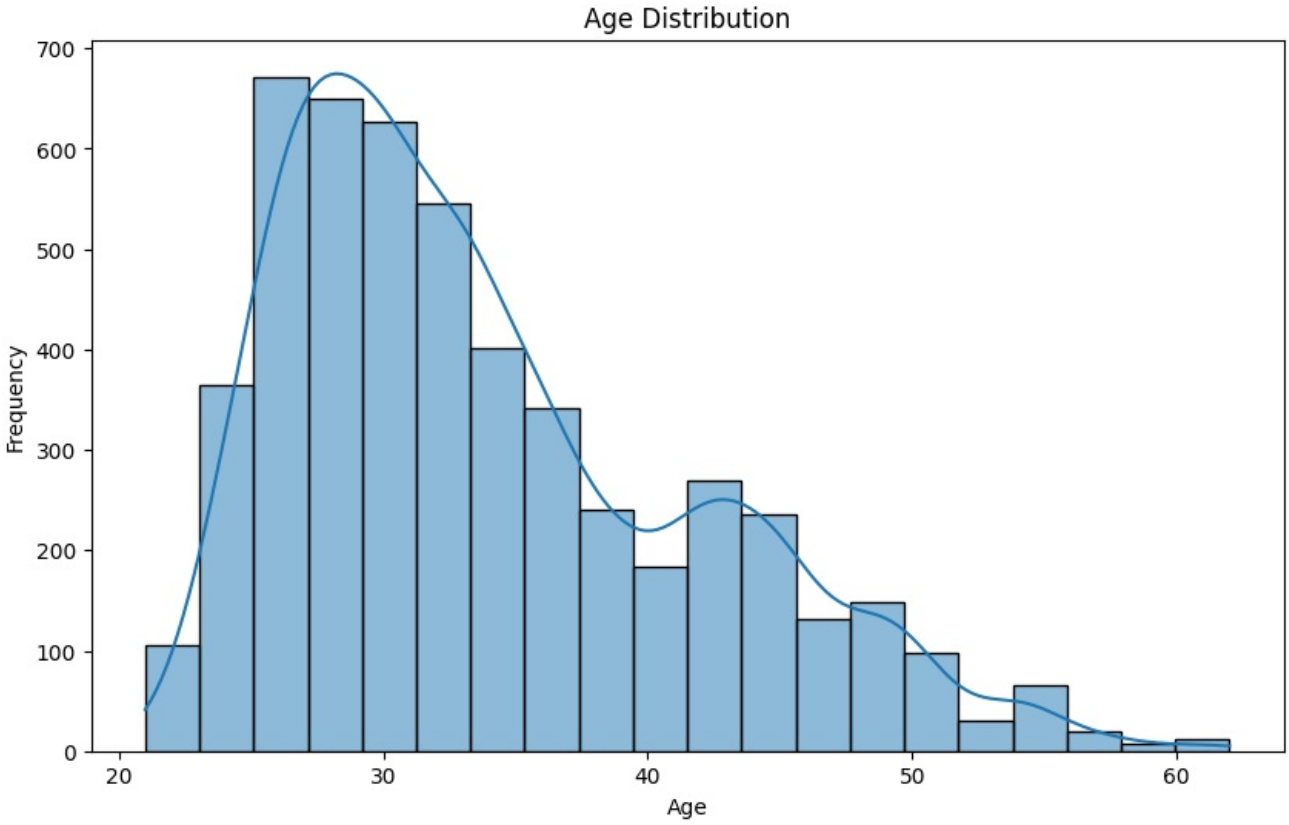
Out[7]:

	Age	Education Level	Years of Experience	Salary	Senior
count	6684.000000	6684.000000	6684.000000	6684.000000	6684.000000
mean	33.610563	1.622382	8.077723	115307.175194	0.143477
std	7.595994	0.880474	6.029305	52806.810881	0.350585
min	21.000000	0.000000	0.000000	350.000000	0.000000
25%	28.000000	1.000000	3.000000	70000.000000	0.000000
50%	32.000000	1.000000	7.000000	115000.000000	0.000000
75%	38.000000	2.000000	12.000000	160000.000000	0.000000
max	62.000000	3.000000	34.000000	250000.000000	1.000000

Visualizations

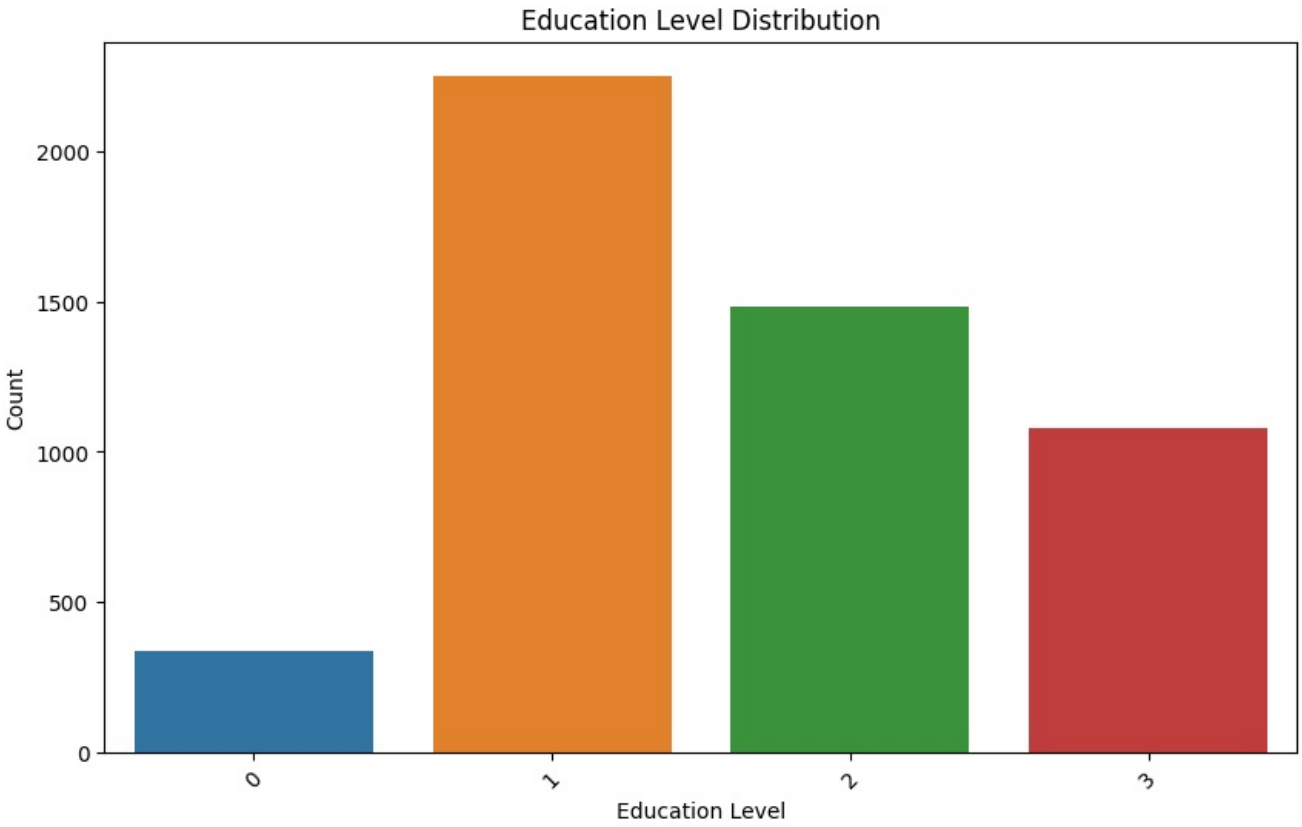
Age Distribution

In [11]:
```
plt.figure(figsize=(10, 6))
sns.histplot(df['Age'], bins=20, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



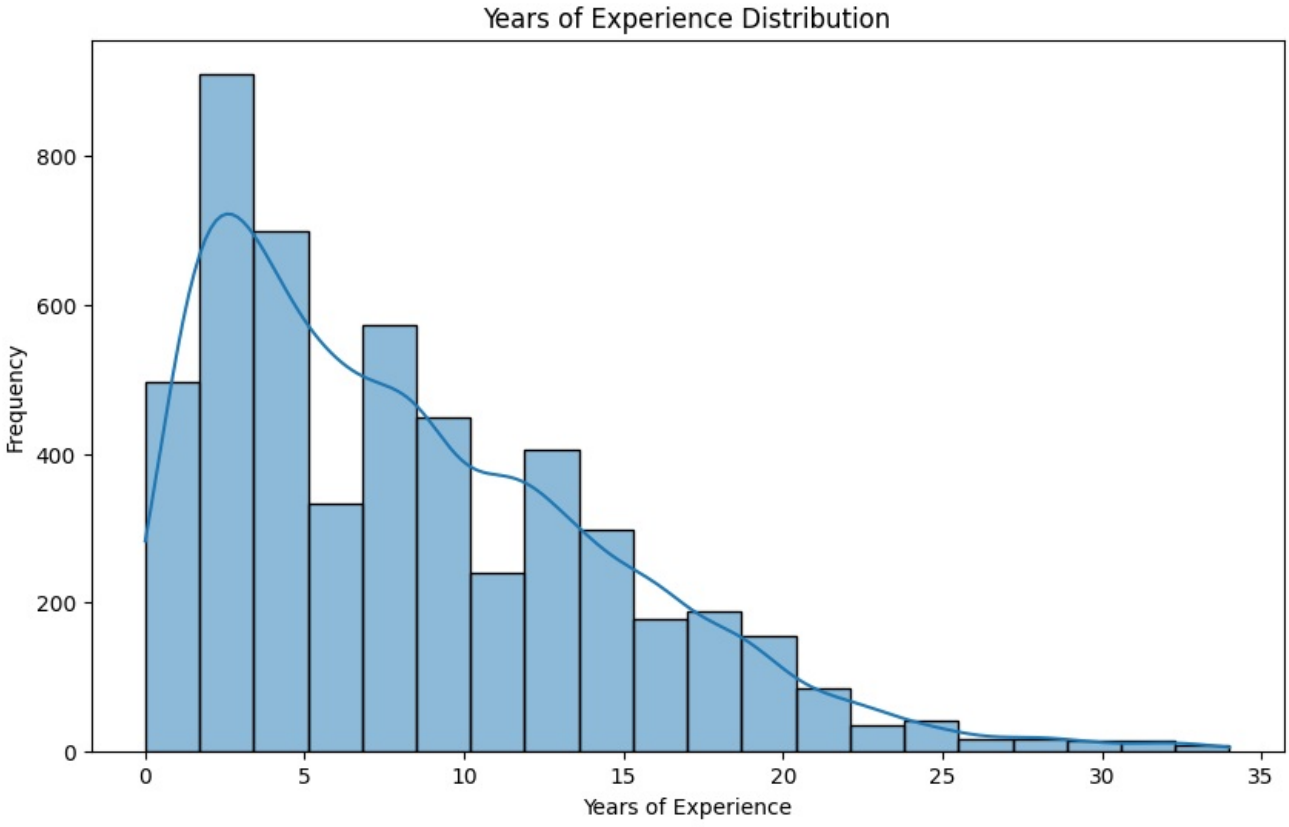
Education Level Distribution

In [13]:
```
plt.figure(figsize=(10, 6))
sns.countplot(x='Education Level', data=df)
plt.title('Education Level Distribution')
plt.xlabel('Education Level')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```



Years of Experience Distribution

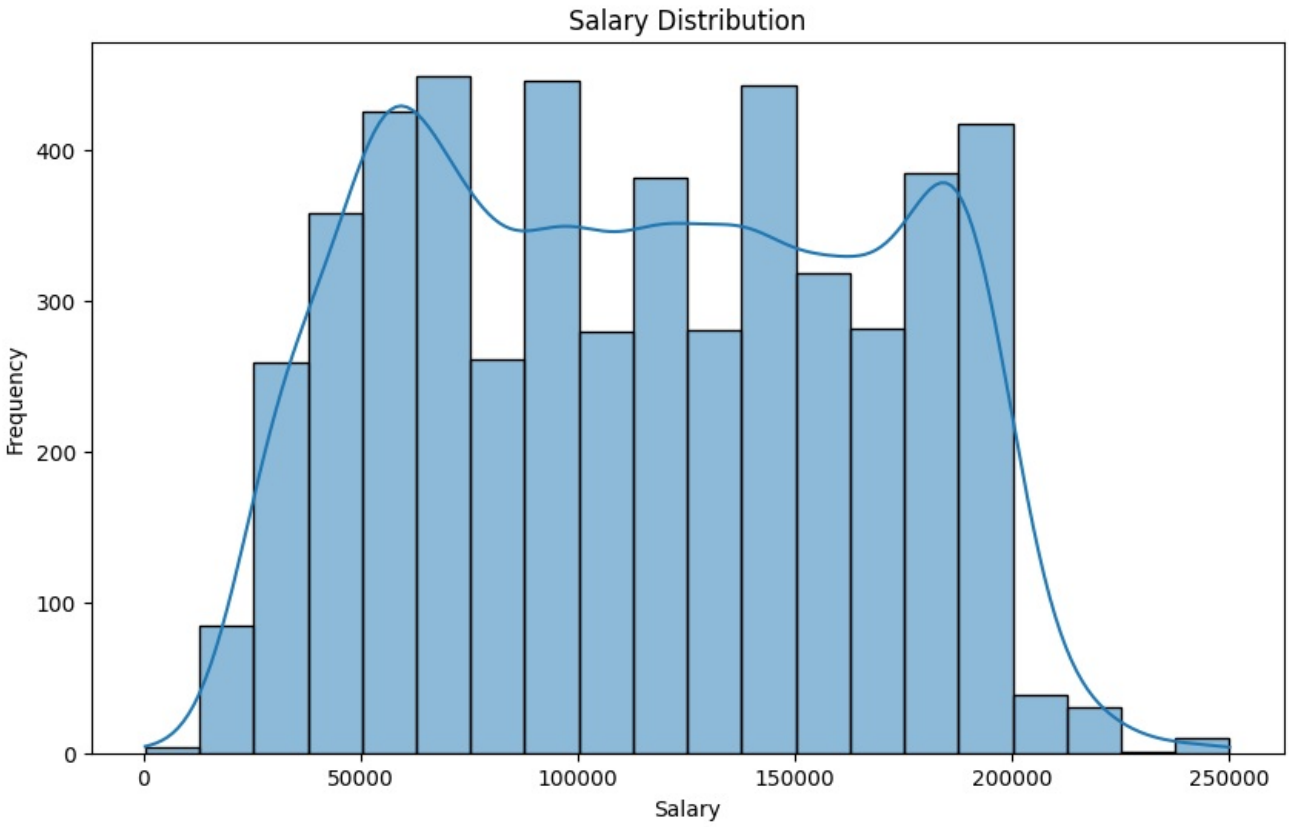
In [16]:
```
plt.figure(figsize=(10, 6))
sns.histplot(df['Years of Experience'], bins=20, kde=True)
plt.title('Years of Experience Distribution')
plt.xlabel('Years of Experience')
plt.ylabel('Frequency')
plt.show()
```



Salary Distribution

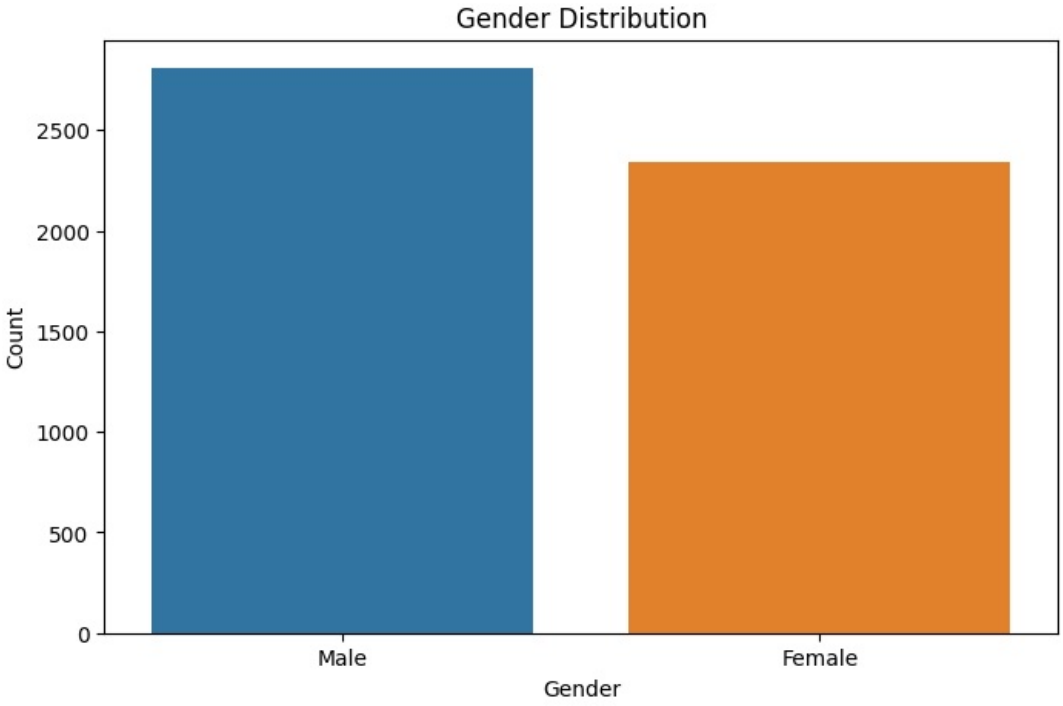
In [17]:
```
plt.figure(figsize=(10, 6))
sns.histplot(df['Salary'], bins=20, kde=True)
plt.title('Salary Distribution')
```

```
plt.xlabel('Salary')
plt.ylabel('Frequency')
plt.show()
```



Gender Distribution

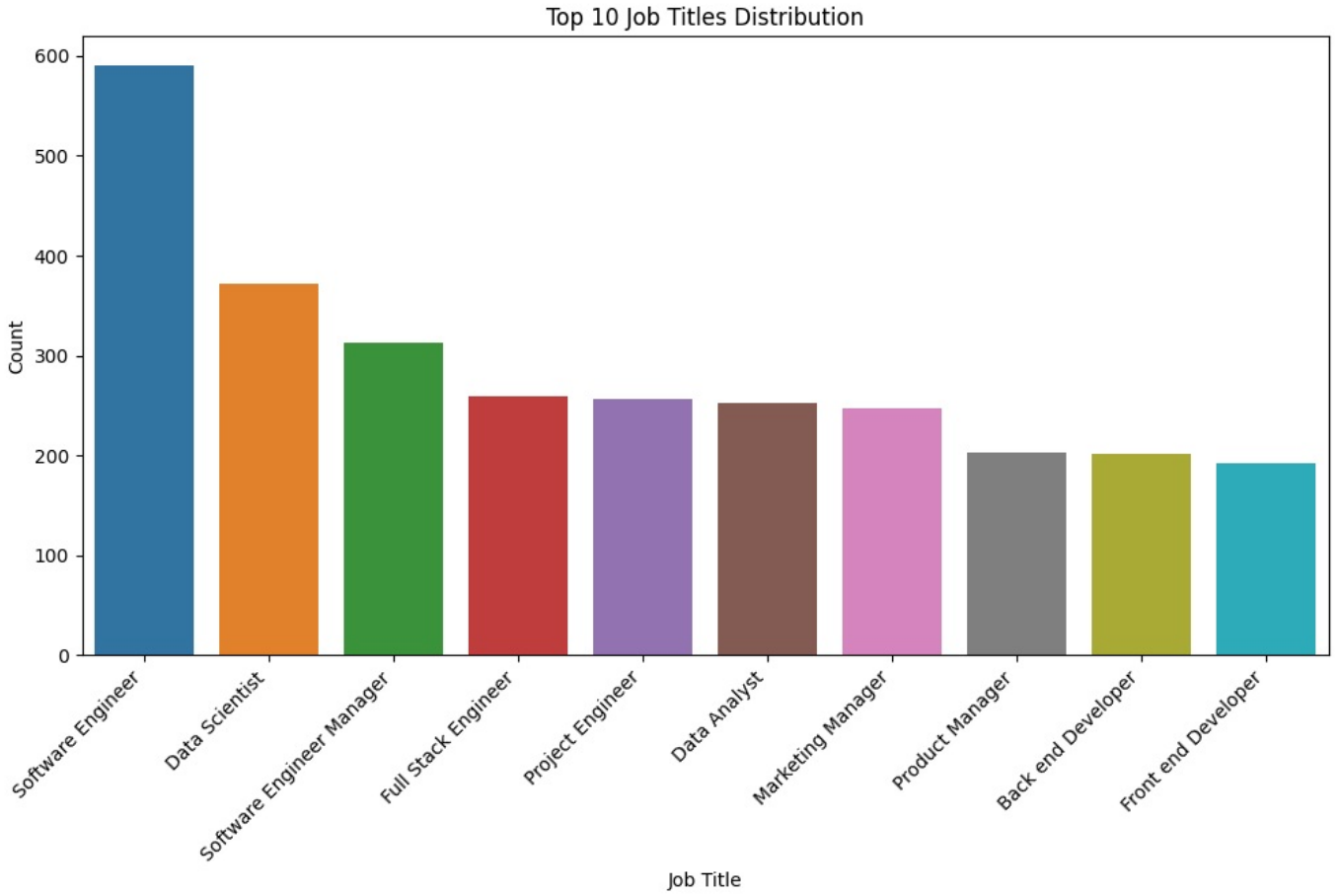
In [18]:
```
plt.figure(figsize=(8, 5))
sns.countplot(x='Gender', data=df)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



Job Title Distribution

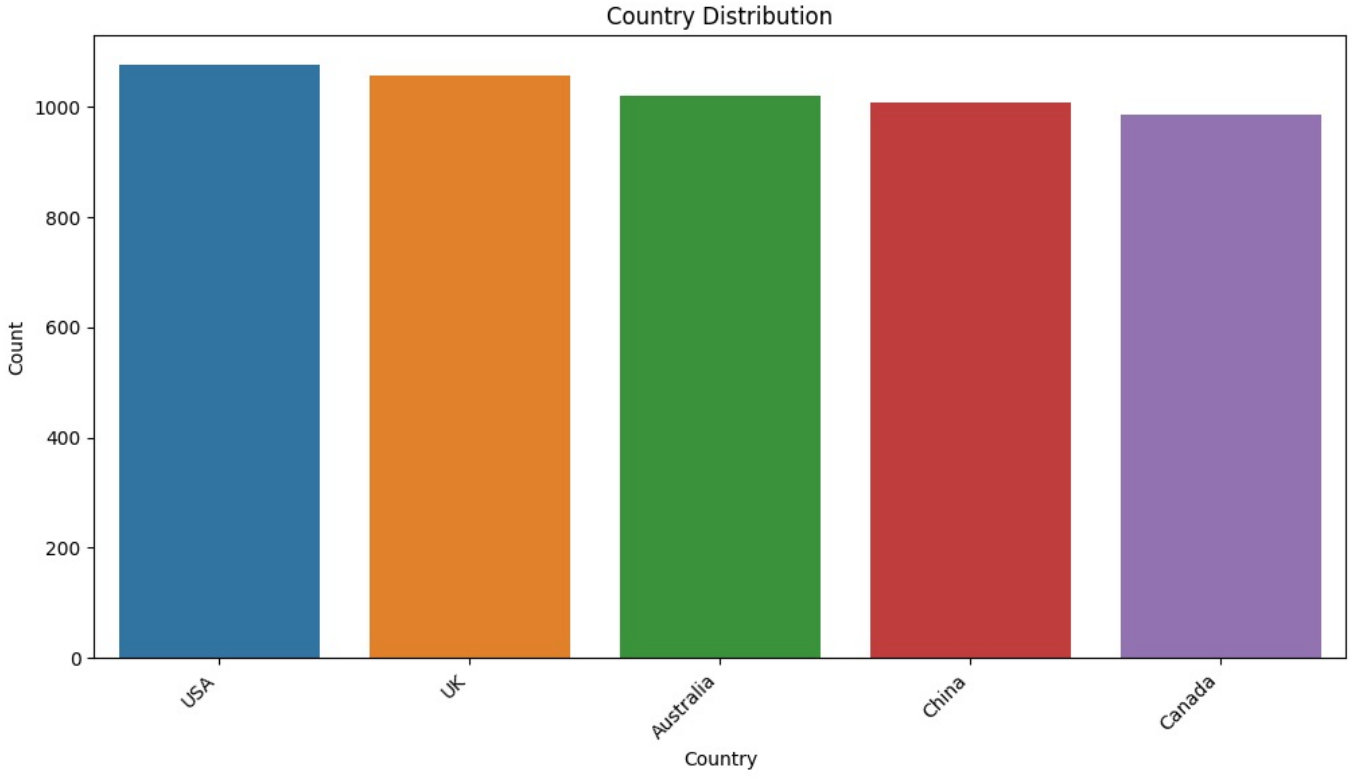
In [24]:
```
plt.figure(figsize=(12, 6))
top_job_titles = df['Job Title'].value_counts().nlargest(10).index
sns.countplot(x='Job Title', data=df, order=top_job_titles)
plt.title('Top 10 Job Titles Distribution')
```

```
plt.xlabel('Job Title')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.show()
```



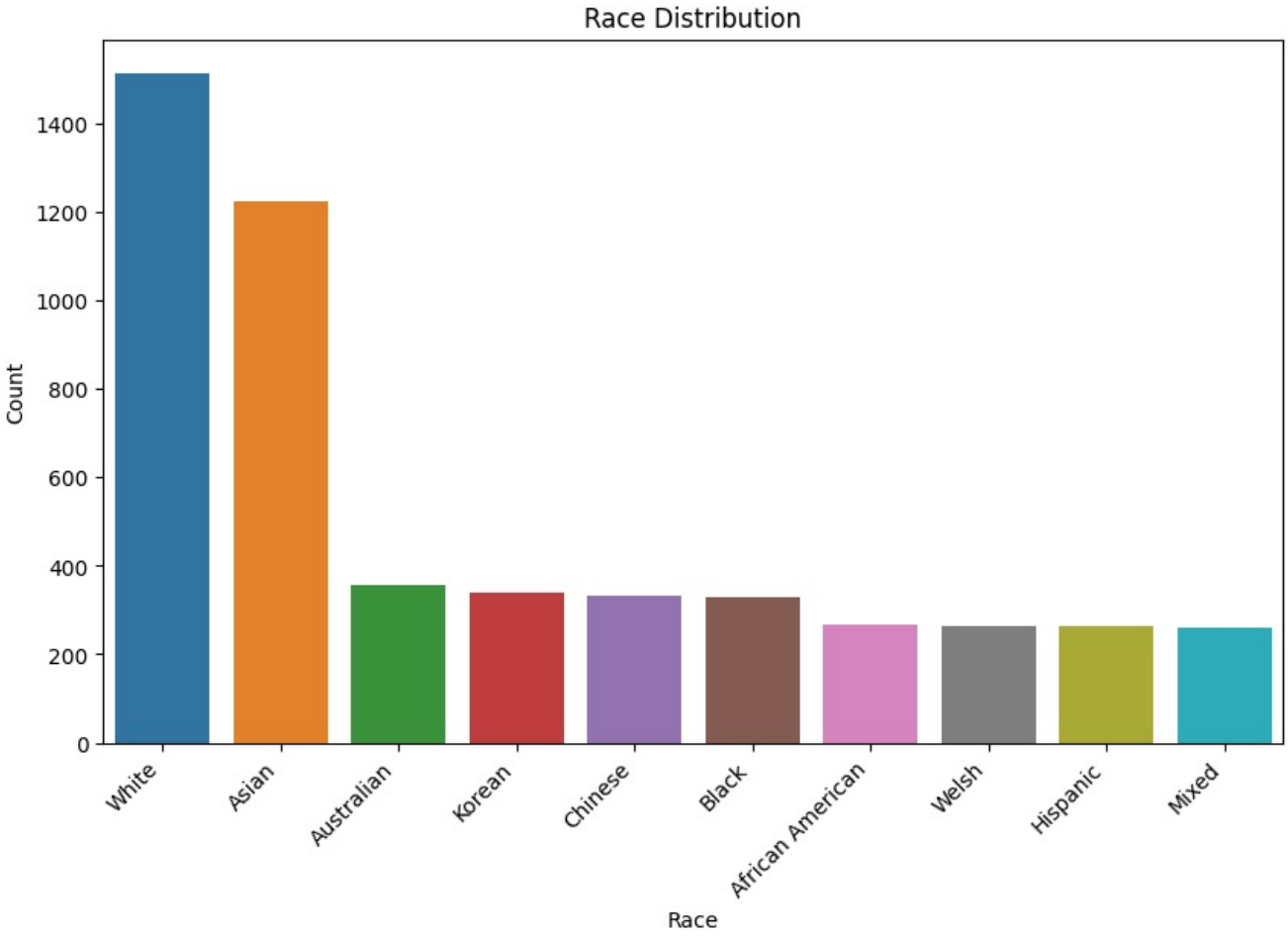
Country Distribution

In [25]:
```
plt.figure(figsize=(12, 6))
sns.countplot(x='Country', data=df, order=df['Country'].value_counts().index)
plt.title('Country Distribution')
plt.xlabel('Country')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.show()
```



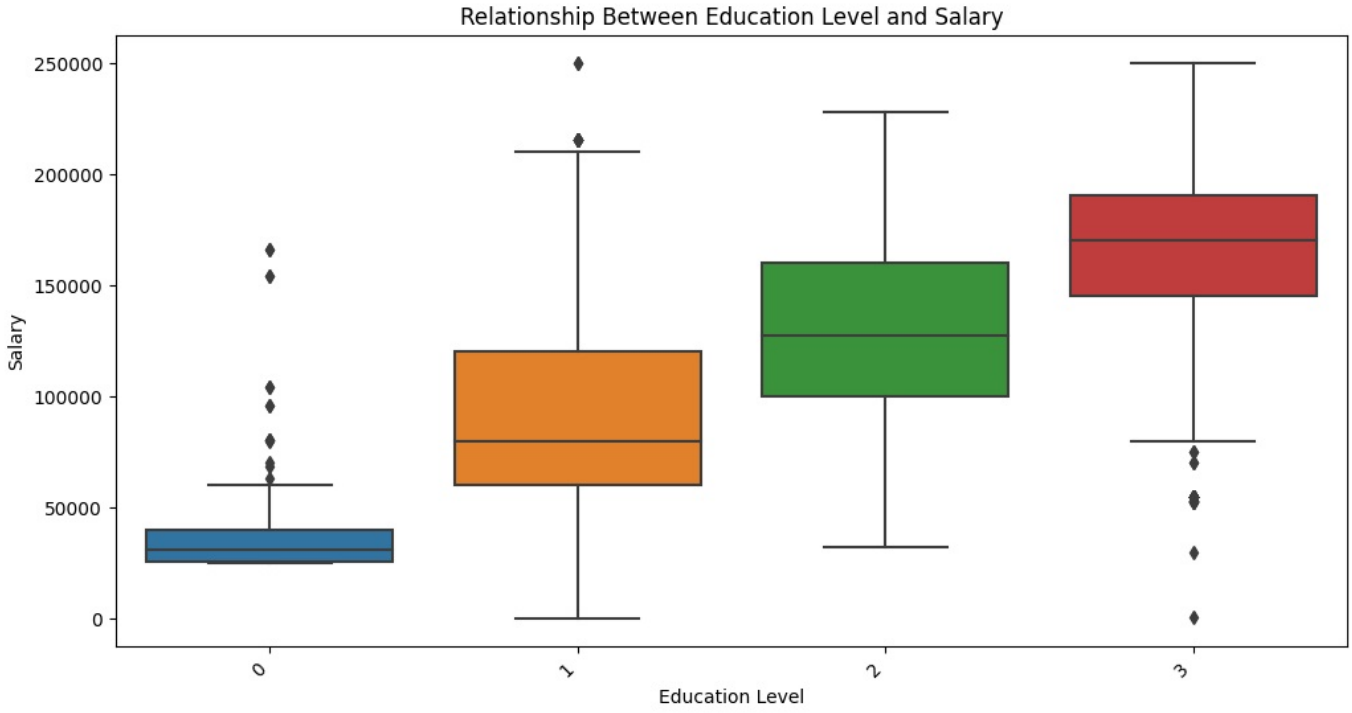
Race Distribution

In [26]:
```
plt.figure(figsize=(10, 6))
sns.countplot(x='Race', data=df, order=df['Race'].value_counts().index)
plt.title('Race Distribution')
plt.xlabel('Race')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right')
plt.show()
```



Analyzing the relationship between education level and salary.

In [28]:
```
# Visualization - Box plot
plt.figure(figsize=(12, 6))
sns.boxplot(x='Education Level', y='Salary', data=df)
plt.title('Relationship Between Education Level and Salary')
plt.xlabel('Education Level')
plt.ylabel('Salary')
plt.xticks(rotation=45, ha='right')
plt.show()
```



In [31]:
```
# Correlation Analysis
correlation_coefficient = df['Education Level'].corr(df['Salary'])
print(f'Correlation Coefficient: {correlation_coefficient}')
```

```
Correlation Coefficient: 0.6442066338489312
```

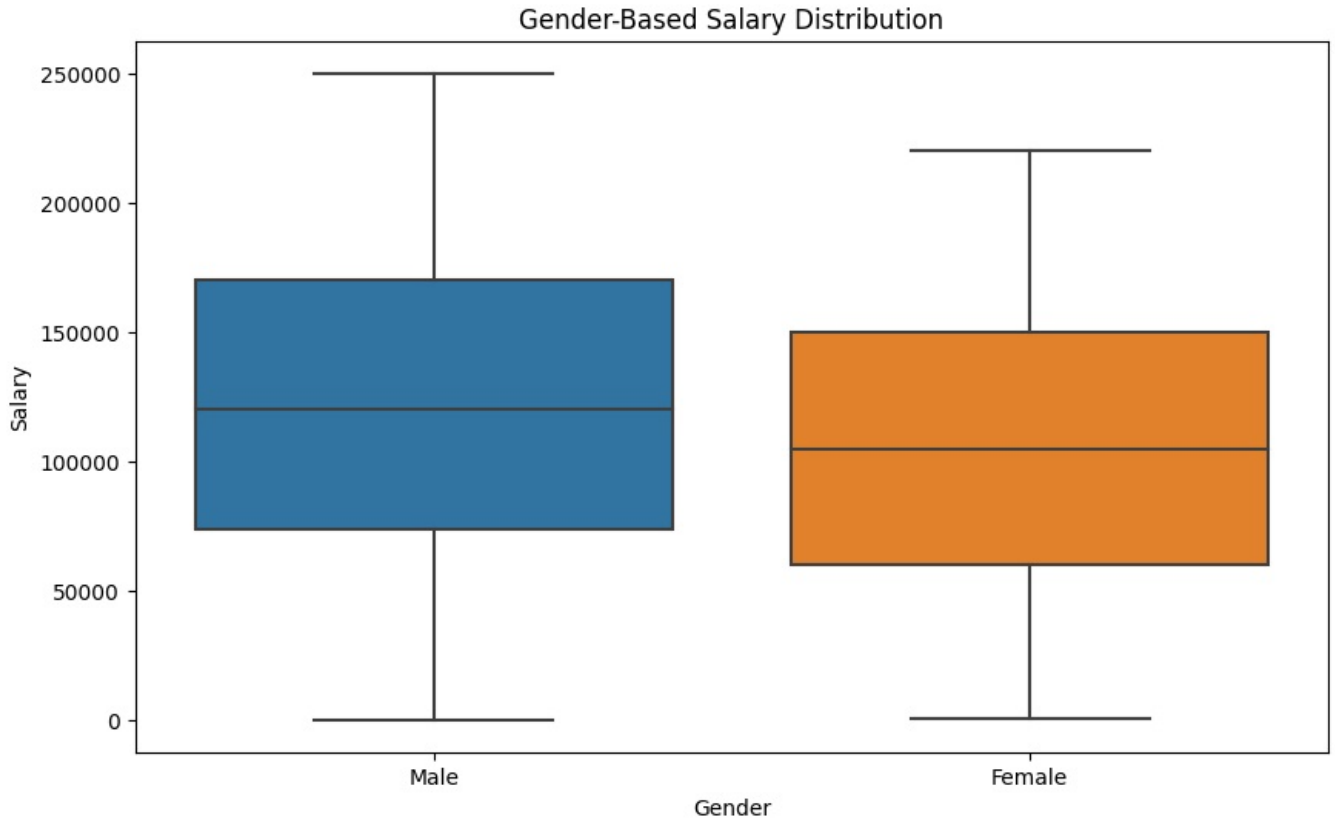
Examining how years of experience correlate with salary.

In [32]:
```
# Visualization - Scatter plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Years of Experience', y='Salary', data=df)
plt.title('Relationship Between Years of Experience and Salary')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()

# Correlation Analysis
correlation_coefficient = df['Years of Experience'].corr(df['Salary'])
print(f'Correlation Coefficient: {correlation_coefficient}')
```


```
Correlation Coefficient: 0.8160302465535905
```

Investigating potential gender based salary gaps.

In [33]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Gender', y='Salary', data=df)
plt.title('Gender-Based Salary Distribution')
plt.xlabel('Gender')
plt.ylabel('Salary')
plt.show()
```



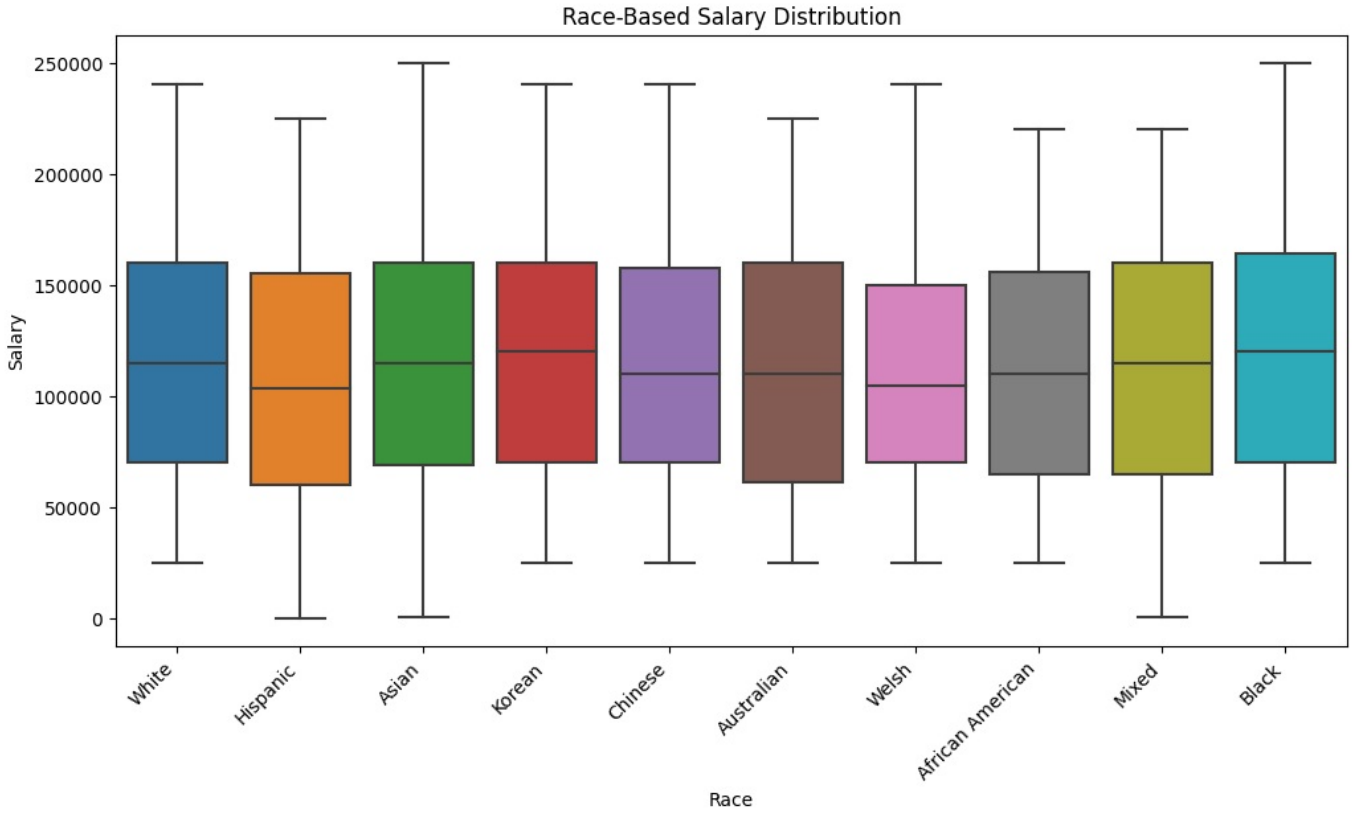
In [35]:
```
male_salaries = df[df['Gender'] == 'Male']['Salary']
female_salaries = df[df['Gender'] == 'Female']['Salary']

t_stat, p_value = ttest_ind(male_salaries, female_salaries)
print(f'T-statistic: {t_stat}\nP-value: {p_value}')
```

```
T-statistic: 9.385135283398887
P-value: 9.198486428013757e-21
```

With a t-statistic of 9.39 and an extremely low p-value (9.20e-21), there is strong evidence to reject the null hypothesis that there is no difference in salaries between different gender groups.

Investigating potential race-based salary gaps.

In [36]:
```
plt.figure(figsize=(12, 6))
sns.boxplot(x='Race', y='Salary', data=df)
plt.title('Race-Based Salary Distribution')
plt.xlabel('Race')
plt.ylabel('Salary')
plt.xticks(rotation=45, ha='right')
plt.show()
```



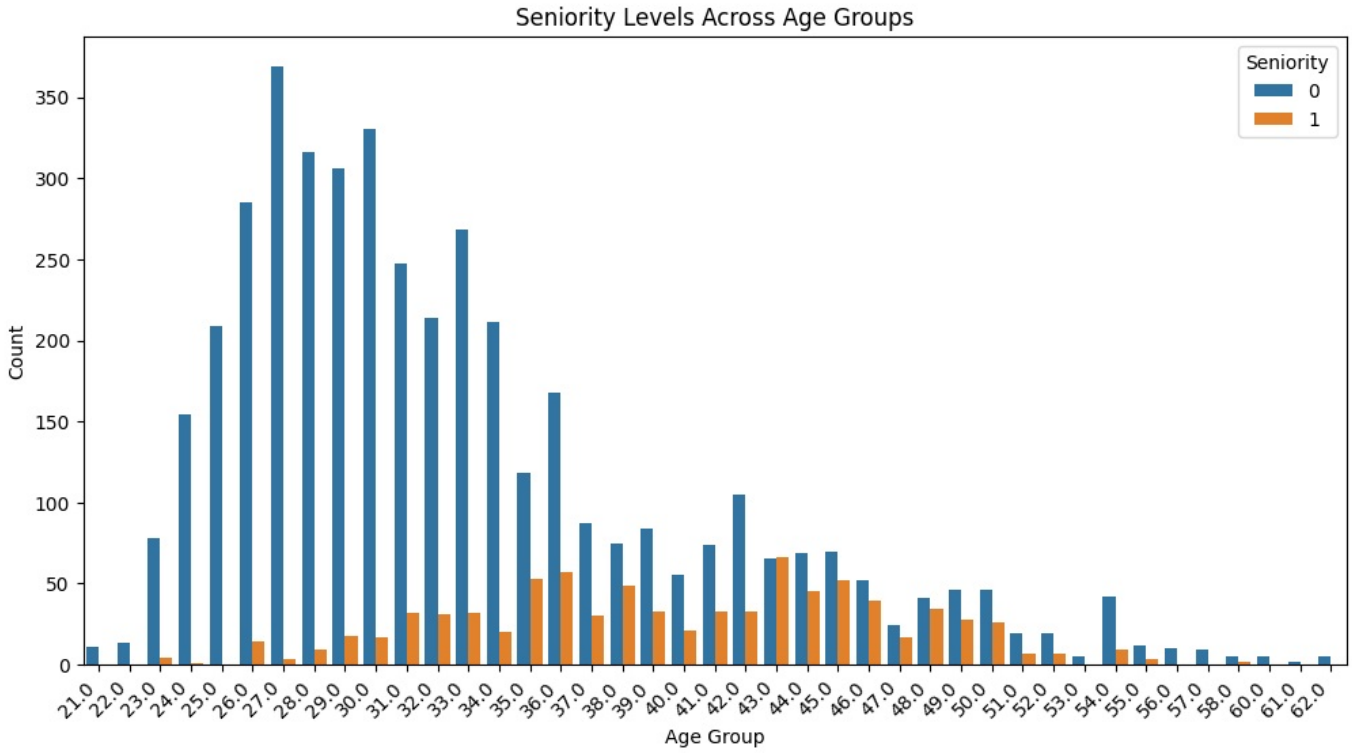
In [38]:
```
race_groups = [df[df['Race'] == race]['Salary'] for race in df['Race'].unique()]
f_stat, p_value = f_oneway(*race_groups)
print(f'F-statistic: {f_stat}\nP-value: {p_value}')
```

```
F-statistic: 0.8969678545953949
P-value: 0.5269201554713177
```

With an F-statistic of 0.897 and a p-value of 0.527, the p-value is higher than the typical significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis that there are no statistically significant differences in salaries among different race groups.

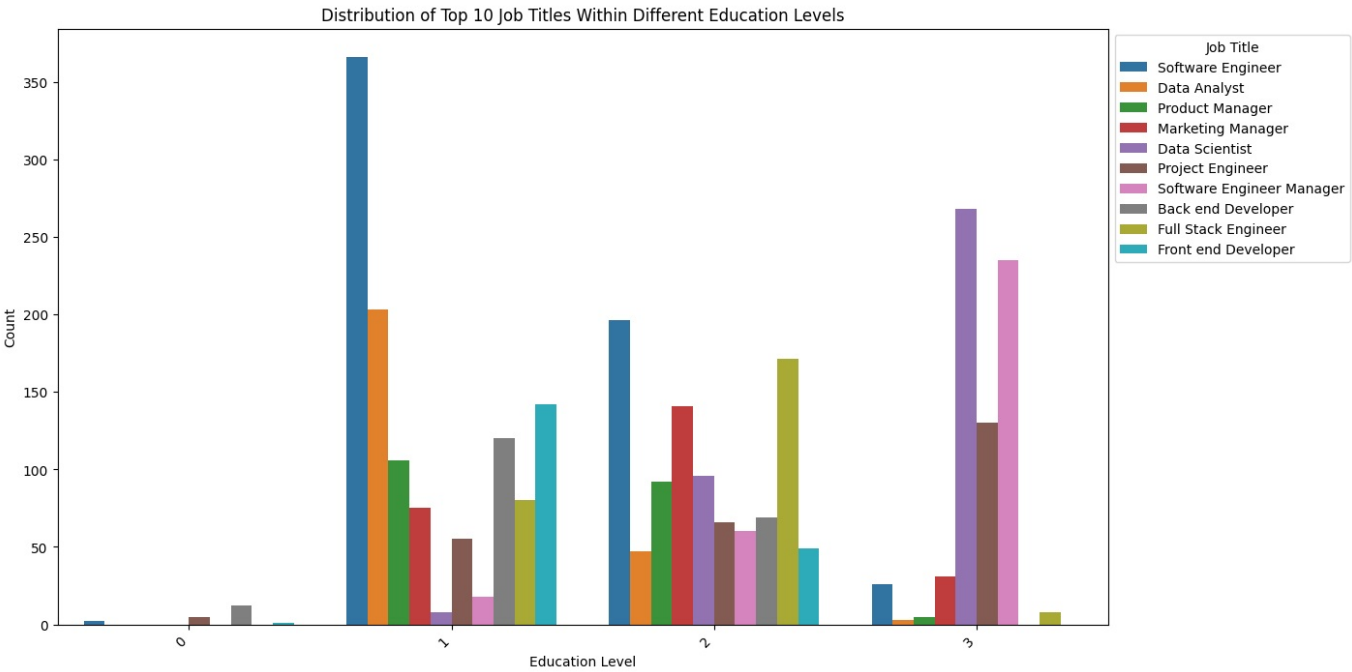
Exploring how seniority levels vary across different age groups.

In [47]:
```
plt.figure(figsize=(12, 6))
sns.countplot(x='Age', hue='Senior', data=df)
plt.title('Seniority Levels Across Age Groups')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.legend(title='Seniority')
plt.xticks(rotation=45, ha='right')
plt.show()
```



Analyze the distribution of top 10 job titles within different education levels.

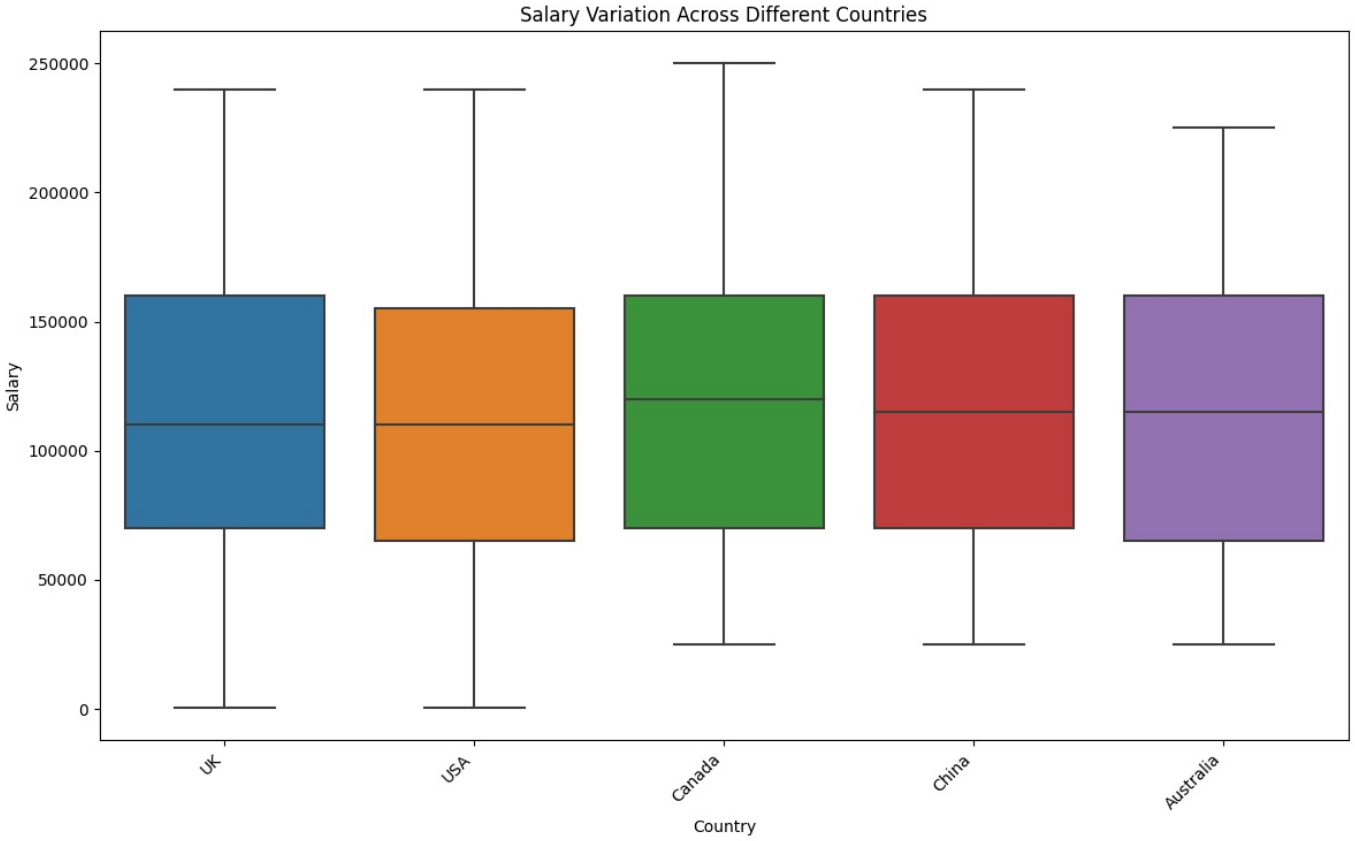
In [51]:
```
# Get the top 10 job titles
top_job_titles = df['Job Title'].value_counts().nlargest(10).index

# Filter the DataFrame for only the top 10 job titles
df_top_job_titles = df[df['Job Title'].isin(top_job_titles)]

# Visualization - Stacked bar chart
plt.figure(figsize=(14, 8))
sns.countplot(x='Education Level', hue='Job Title', data=df_top_job_titles)
plt.title('Distribution of Top 10 Job Titles Within Different Education Levels')
plt.xlabel('Education Level')
plt.ylabel('Count')
plt.legend(title='Job Title', bbox_to_anchor=(1, 1))
plt.xticks(rotation=45, ha='right')
plt.show()
```



Explore how salaries vary across different countries.

In [52]:
```
# Visualization - Box plot
plt.figure(figsize=(14, 8))
sns.boxplot(x='Country', y='Salary', data=df)
```

```
plt.title('Salary Variation Across Different Countries')
plt.xlabel('Country')
plt.ylabel('Salary')
plt.xticks(rotation=45, ha='right')
plt.show()
```



Conclusion

Thank you for accompanying me through this enlightening journey of exploratory data analysis (EDA). Even though I have only scratched the surface of the dataset's potential, I have gained the insights I sought. This exploration has bestowed upon me valuable knowledge and answered the questions that plagued my mind. While there are numerous other possible conclusions and analyses, I have decided to end my EDA here.

I trust that you found my notebook engaging and, more importantly, beneficial. I welcome any feedback you may have, and I assure you that I read and respond to each one with utmost sincerity.

I wish you the best of luck in your endeavors!

For more Follow me on Linkedin

https://www.linkedin.com/in/paras-dahiya03

In [ ]: