National University of Computer and Emerging Sciences

Career Development

Job Salary Prediction

Project Deliverable 1

Team

Hammadullah Abid	21L-5241
Shaheer Ahmad Rana	21L-5193
Rameez Qureshi	21L-5301
Zahra Aslam Khan	21L-7723
M. Ibrahim Raza	21L-1795

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

March 2025

TABLE OF CONTENTS

Introduction

Problem Statement

In today's rapidly evolving job market, salary prediction plays a crucial role in career planning, recruitment strategies, and workforce management. The Career Development - Job Salary Prediction project aims to develop a predictive model that estimates salary ranges based on various job-related factors. By analyzing historical job postings and their associated salary information, this project seeks to provide valuable insights for job seekers, employers, and analysts.

The dataset used in this project consists of job postings with various attributes, including:

1. **Job ID**
2. **Experience**
3. **Qualifications**
4. **Salary Range**
5. **Location**
6. **Country**
7. **Work Type**
8. **Company Size**
9. **Job Posting Date**
10. **Preference**
11. **Job Title**
12. **Role**
13. **Job Description**
14. **Benefits**
15. **Skills**
16. **Responsibilities**
17. **Company Name**

These attributes will be processed and analyzed to determine their impact on salary predictions.

Exploratory Data Analysis

A fundamental step in this process is data wrangling, which ensures that the dataset used for analysis is clean, structured, and ready for modeling. Data wrangling includes essential steps such as data loading, cleaning, transformation, and exploratory data analysis (EDA), which help identify inconsistencies, missing values, and patterns within the data. High-quality data preparation is crucial for building an accurate and reliable salary prediction model.

Dataset Overview

This dataset contains job-related information, including various attributes that describe job postings across different locations, industries, and work types. It provides insights into job roles, required qualifications, salary ranges, and employer details.

Data Quality Considerations

Ensuring data quality is essential for accurate analysis and meaningful insights, requiring careful handling of inconsistencies, missing values, and formatting issues.

- Some features may contain missing or inconsistent values that require preprocessing.
- Salary ranges and experience levels are categorical and may need transformation for numerical analysis.
- Text-based features like Job Description and Responsibilities may require natural language processing (NLP) techniques for deeper insights.

Career Development - Job Salary Prediction

Dataset Description

The dataset consists of 1,048,576 rows and 17 features, making it a comprehensive source for job market analysis.

FEATURE DESCRIPTION TABLE		
Feature	**Description**	**Example**
Job ID	Unique identifier for each job posting.	3.98E+14
Experience	Required experience range in years.	2 to 12 Years
Qualifications	Minimum educational qualification needed for the job.	PhD
Salary Range	Expected salary range in USD for the role.	$56K-$116K
Location	City where the job is based.	Ashgabat
Country	Country where the job is located.	Turkmenistan
Work Type	Nature of employment (Full-Time, Part-Time, Intern, Temporary).	Temporary
Company Size	Number of employees in the company.	100,340
Job Posting Date	Date when the job was posted.	12/19/2022
Preference	Gender preference, if specified.	Male
Job Title	Title of the job position.	Web Developer
Role	Specific job function within the given title.	Frontend Web Developer
Job Description	Summary of responsibilities and expectations for the role.	"Design and implement UI..."
Benefits	Perks and benefits provided by the employer.	Health Insurance, PTO
Skills	Key technical or soft skills required.	HTML, CSS, JavaScript
Responsibilities	Main duties associated with the job.	Design and code UI
Company	Hiring organization name.	PNC Financial Services Group

This dataset provides a valuable resource for analyzing employment trends and developing predictive models for salary prediction.

Data Loading & Exploration

Loading Dataset

The dataset was loaded into the environment using the pandas library. Specifically, the read_csv() function was used to import the dataset from a CSV file:

1 - Data Loading

```
job_salary_df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Job_Salary_Prediction_Dataset.csv')
```

Initial Look At Data

To understand the dataset structure, the first 15 rows were displayed using:

2 - Data Overview Exploration ¶

Displaying First Fifteen Rows

```
job_salary_df.head(15)
```

	Job Id	Experience	Qualifications	Salary Range	location	Country	Work Type	Company Size	Job Posting Date	Preference	Job Title	Role	Job Description
0	1.089840e+15	5 to 15 Years	M.Tech	$59K-$ 99K	Douglas	Isle of Man	Intern	26801	4/24/2022	Female	Digital Marketing Specialist	Social Media Manager	Social Media Managers oversee an organizations...
1	3.984540e+14	2 to 12 Years	BCA	$56K-$ 116K	Ashgabat	Turkmenistan	Intern	100340	12/19/2022	Female	Web Developer	Frontend Web Developer	Frontend Web Developers design and implement u...
2	4.816400e+14	0 to 12 Years	PhD	$61K-$ 104K	Macao	Macao SAR, China	Temporary	84525	9/14/2022	Male	Operations Manager	Quality Control Manager	Quality Control Managers establish and enforce...
3	6.881930e+14	4 to 11 Years	PhD	$65K-$ 91K	Porto-Novo	Benin	Full-Time	129896	2/25/2023	Female	Network Engineer	Wireless Network Engineer	Wireless Network Engineers design,

10	2.696960e+15	3 to 10 Years	BCA	57K–104K	Manama	Bahrain	Contract	130338	7/1/2023	Female	QA Analyst	Testing Specialist	Specialists assess the per...
11	1.446190e+15	4 to 12 Years	B.Tech	64K–98K	The City of Hamilton	Bermuda	Contract	117285	10/11/2021	Male	Litigation Attorney	Family Law Attorney	Family Law Attorneys deal with legal matters r...
12	1.914120e+15	3 to 15 Years	MCA	65K–122K	Kingston	Jamaica	Part-Time	79071	1/17/2022	Both	Mechanical Engineer	Mechanical Design Engineer	Mechanical Design Engineers create and develop...
13	2.907620e+14	1 to 8 Years	B.Com	56K–86K	Banjul	Gambia	Temporary	127900	5/24/2022	Female	Network Administrator	Network Security Analyst	Protect an organizations computer networks and...
14	1.627540e+15	1 to 9 Years	MCA	57K–98K	Damascus	Syrian Arab Republic	Full-Time	92128	3/1/2022	Male	Account Manager	Sales Account Manager	A Sales Account Manager is responsible for bui...

4	1.170580e+14	1 to 12 Years	MBA	64K–87K	Santiago	Chile	Intern	53944	10/11/2022	Female	Event Manager	Conference Manager	A Conference Manager coordinates and manages c...
5	1.168310e+14	4 to 12 Years	MCA	59K–93K	Brussels	Belgium	Full-Time	23196	7/25/2023	Male	Software Tester	Quality Assurance Analyst	A Quality Assurance Analyst tests software and...
6	1.292170e+15	3 to 15 Years	PhD	63K–103K	George Town	Cayman Islands	Temporary	26119	4/10/2023	Both	Teacher	Classroom Teacher	A Classroom Teacher educates students in a spe...
7	1.498780e+15	2 to 8 Years	M.Com	65K–102K	SÃ£o TomÃ©	Sao Tome and Principe	Contract	40558	9/20/2022	Female	UX/UI Designer	User Interface Designer	User Interface Designers focus on the visual a...
8	1.680290e+15	2 to 9 Years	BBA	65K–102K	Male	Maldives	Temporary	105343	2/19/2022	Female	UX/UI Designer	Interaction Designer	Interaction Designers specialize in designing ...
9	2.556280e+14	1 to 10 Years	BBA	60K–80K	Saint John's	Antigua and Barbuda	Full-Time	102069	5/13/2022	Both	Wedding Planner	Wedding Consultant	A Wedding Consultant assists couples in planni...

This allowed for an initial assessment of the dataset, including column names, data types, and missing values.

Dataset Summary

Displaying No. of Rows and Columns

```
print(f"Rows: {job_salary_df.shape[0]}, Columns: {job_salary_df.shape[1]}")
```

```
Rows: 1048575, Columns: 17
```

To gain insights into the dataset, the following methods were used:

- **info():** Provided an overview of column names, data types, and missing values.
- **describe():** Generated summary statistics, including count, mean, standard deviation, min, max, and quartiles.

Displaying Summary of Non-Missing Values and Data Types of Columns

```
job_salary_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 17 columns):
 #   Column            Non-Null Count    Dtype  
---  ------            --------------    -----  
 0   Job Id            1048575 non-null  float64
 1   Experience        1048575 non-null  object 
 2   Qualifications    1048575 non-null  object 
 3   Salary Range      1048575 non-null  object 
 4   location          1048575 non-null  object 
 5   Country           1048575 non-null  object 
 6   Work Type         1048575 non-null  object 
 7   Company Size      1048575 non-null  int64  
 8   Job Posting Date  1048575 non-null  object 
 9   Preference        1048575 non-null  object 
 10  Job Title         1048575 non-null  object 
 11  Role              1048575 non-null  object 
 12  Job Description   1048575 non-null  object 
 13  Benefits          1048575 non-null  object 
 14  skills            1048575 non-null  object 
 15  Responsibilities  1048575 non-null  object 
 16  Company           1048575 non-null  object 
dtypes: float64(1), int64(1), object(15)
memory usage: 136.0+ MB
```

Key Observations

- The dataset consists of 1,048,575 rows and 17 columns.
- Key columns include Job Id, Experience, Qualifications, Salary Range, Location, Company Size, Job Posting Date, Role, and Company.
- The dataset contains both categorical and numerical data.

Data Cleaning

Dropping Unnecessary Columns

Certain columns were dropped to reduce noise and complexity:

1. **Job ID**
    - It Has No Predictive Value
    - It Can Mislead the Model

```
job_salary_df.drop('Job Id', axis=1, inplace=True)
```

2. **Job Description**
    - It contains unstructured textual data.
    - Each role has exactly one corresponding job description.
    - Increases noise without providing additional predictive value.

```
job_salary_df["Role"].nunique()
```

```
376
```

```
print((job_salary_df.groupby('Role')['Job Description'].nunique() == 1).sum())
```

```
376
```

```
job_salary_df.drop('Job Description', axis=1, inplace=True)
```

3. **Skills**
    - Skills listed are generic to job roles.
    - Each role has a one-to-one mapping with skills, making it redundant.
    - Unstructured data increases complexity.

```
job_salary_df["Role"].nunique()
```

```
376
```

```
print((job_salary_df.groupby('Role')['skills'].nunique() == 1).sum())
```

```
376
```

```
job_salary_df.drop('skills', axis=1, inplace=True)
```

4. **Responsibilities**
    - Responsibilities are directly tied to job roles.
    - Similar to Job Description, it does not add unique predictive value.

```
job_salary_df["Role"].nunique()
```

```
376
```

```
print((job_salary_df.groupby('Role')['Responsibilities'].nunique() == 1).sum())
```

```
376
```

```
job_salary_df.drop('Responsibilities', axis=1, inplace=True)
```

Handling Missing Data

No missing values were found in the dataset, eliminating the need for imputation or deletion.

Handling Missing Values

No Missing Value

```
job_salary_df.isnull().sum()
```

	0
Experience	0
Qualifications	0
Salary Range	0
location	0
Country	0
Work Type	0
Company Size	0
Job Posting Date	0
Preference	0
Job Title	0
Role	0
Benefits	0
Company	0

Duplicate Removal

The dataset was checked for duplicate records using:

Handling Duplicate Enteries / Records

No Duplicate Enteries / Records

```
print(job_salary_df.duplicated().sum())
```

```
0
```

- No duplicate records were found.

Data Type Conversions

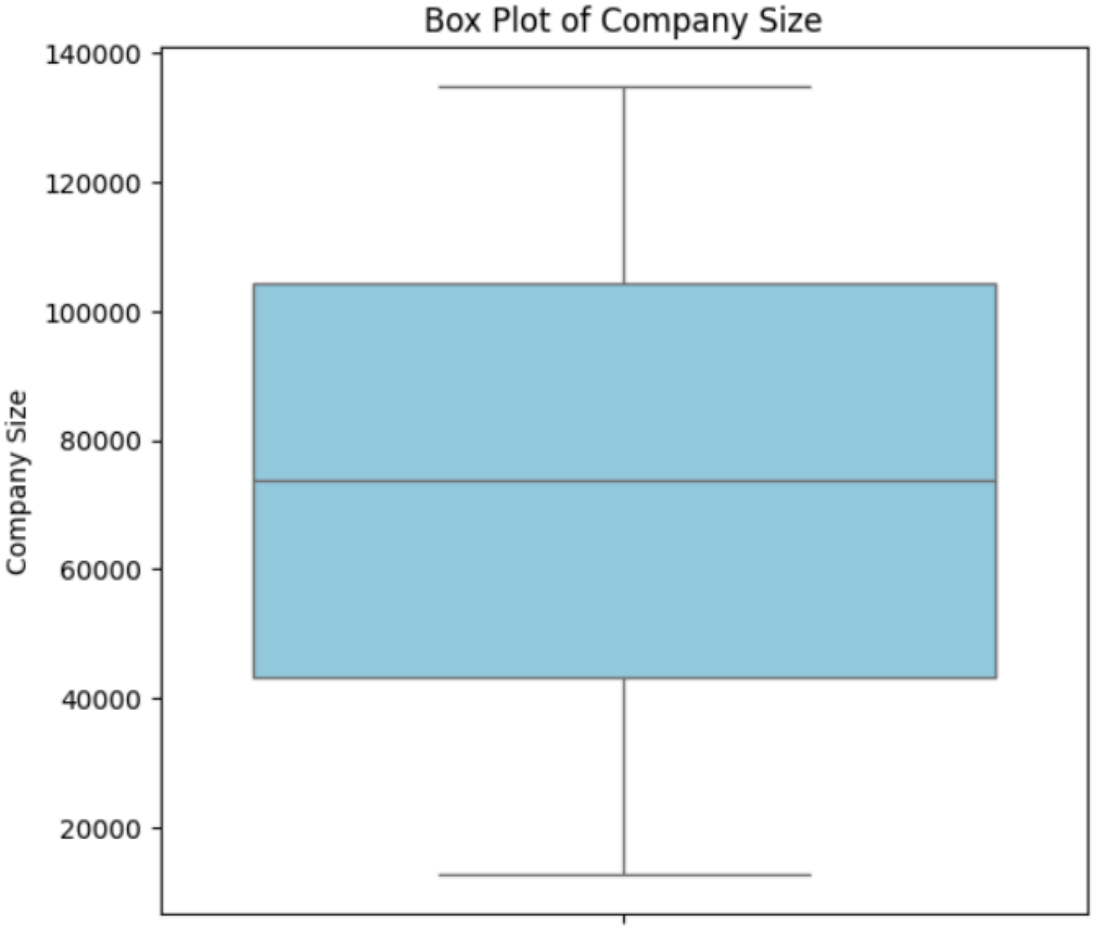
- The Job Posting Date column was converted to datetime format for better analysis.

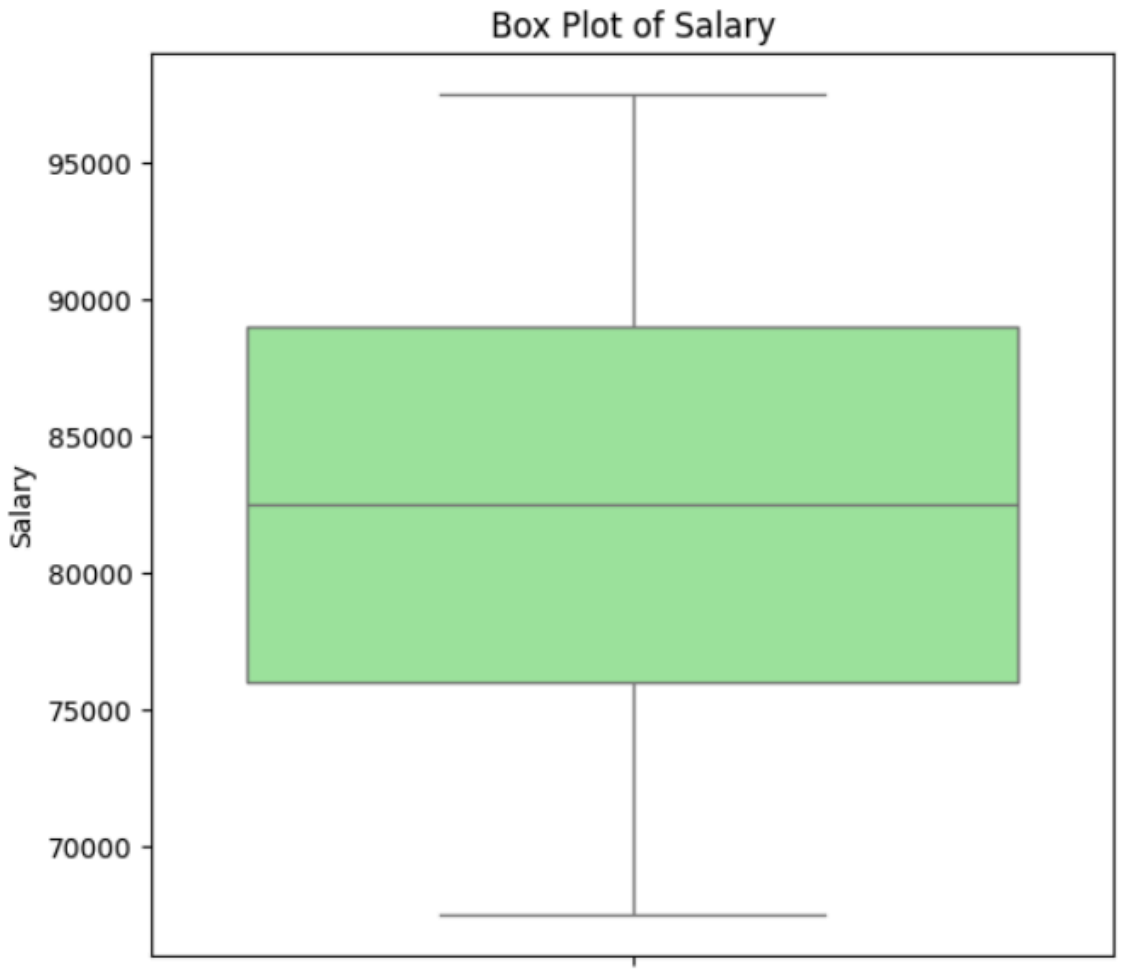
```
job_salary_df['Job Posting Date'] = pd.to_datetime(job_salary_df['Job Posting Date'])
```

Outlier Detection

- Outliers were detected using the **Interquartile Range (IQR)** method.
- No extreme outliers were found that required removal.

Career Development - Job Salary Prediction





Data Transformation

Feature Scaling

Feature scaling was applied to ensure that numerical values contributed equally to model training. Since **Company Size** varies significantly across companies, **StandardScaler** was used to standardize this feature.

For this dataset, **StandardScaler** was applied to the **Company Size** column, transforming it to have a mean of 0 and a standard deviation of 1. This ensures that the model treats company size consistently, preventing larger values from dominating the learning process.

Encoding Categorical Variables

Categorical variables were converted into numeric representations using different encoding techniques. This step ensures that models can process and learn from categorical data effectively.

One-Hot Encoding

One-hot encoding was applied to categorical variables that do not have an inherent order, such as Qualifications and Benefits.

- This method avoids imposing a false ordinal relationship between categories.
- It allows the model to learn the influence of each category independently.

Label Encoding

Label encoding was used for ordinal categorical variables like Preference and Work Type, which have an inherent ranking.

- This method maintains ordinal relationships while keeping the feature space compact.
- It avoids unnecessary complexity and improves computational efficiency.

Target Mean Encoding

Target mean encoding was applied to high-cardinality categorical columns like Company, Location, and Job Profile to capture their direct relationship with salary.

```
job_salary_df['Location'] = job_salary_df.groupby('Location')['Salary'].transform('mean')
```

- This method effectively handles categorical variables with many unique values.
- It improves predictive performance by capturing category-specific salary trends.

Additional Transformations

Several additional feature transformations were performed to enhance predictive power and ensure data consistency.

Transforming Experience Column

The Experience column was split into Min Experience and Max Experience to preserve all information while making it suitable for numerical analysis.

- This transformation allows the model to learn how both lower and upper experience bounds affect salary predictions.
- Preserves complete experience information for better pattern recognition.

```
print(job_salary_df["Experience"].unique())
```

```
['5 to 15 Years' '2 to 12 Years' '0 to 12 Years' '4 to 11 Years'
 '1 to 12 Years' '4 to 12 Years' '3 to 15 Years' '2 to 8 Years'
 '2 to 9 Years' '1 to 10 Years' '3 to 10 Years' '1 to 8 Years'
 '1 to 9 Years' '5 to 14 Years' '0 to 11 Years' '3 to 12 Years'
 '5 to 9 Years' '0 to 15 Years' '0 to 10 Years' '2 to 14 Years'
 '3 to 9 Years' '4 to 15 Years' '2 to 10 Years' '4 to 8 Years'
 '3 to 8 Years' '1 to 14 Years' '1 to 13 Years' '0 to 8 Years'
 '5 to 10 Years' '2 to 13 Years' '4 to 9 Years' '1 to 15 Years'
 '4 to 10 Years' '5 to 12 Years' '0 to 13 Years' '4 to 14 Years'
 '1 to 11 Years' '4 to 13 Years' '0 to 9 Years' '5 to 8 Years'
 '2 to 15 Years' '5 to 13 Years' '5 to 11 Years' '0 to 14 Years'
 '3 to 13 Years' '2 to 11 Years' '3 to 11 Years' '3 to 14 Years']
```

```
def transformExperience(exp):
    numbers = list(map(int, re.findall(r'\d+', exp)))
    # Return minimum and maximum experience
    return numbers[0], numbers[1]
```

```
job_salary_df[['Min Experience', 'Max Experience']] = job_salary_df['Experience'].apply(
    lambda x: pd.Series(transformExperience(x))
)
```

```
job_salary_df.drop('Experience', axis=1, inplace=True)
```

Transforming Salary Column

The Salary Range column was transformed into a single Salary column by computing the mean of the minimum and maximum values.

- This simplifies the target variable, reducing ambiguity and uncertainty.
- Ensures the model has a specific numerical value to learn from, minimizing overfitting.

```
print(job_salary_df["Salary Range"].nunique())
```

```
561
```

```
job_salary_df[['min_salary', 'max_salary']] = job_salary_df['Salary Range']\
    .str.extract(r'\$(\d+)K-\$(\d+)K').astype(float)*1000
```

```
job_salary_df['Salary'] = (job_salary_df['min_salary'] + job_salary_df['max_salary']) / 2
```

```
job_salary_df.drop(columns=['Salary Range', 'min_salary', 'max_salary'], inplace=True)
```

Combining Location and Country

The Location and Country columns were merged into a single Location column.

- This captures complete geographic information in a single feature.
- Reduces noise and redundancy in the dataset.

```
job_salary_df['Location'] = job_salary_df['Country'] + ' - ' + job_salary_df['location']
```

```
job_salary_df.drop(columns=['Country','location'], inplace=True)
```

Combining Job Title and Role

The Job Title and Role columns were combined into a new Job Profile column.

- Job title and role are closely related, and merging them reduces sparsity.
- This transformation helps in capturing complete position information in one signal.

```
job_salary_df['Job Profile'] = job_salary_df['Job Title'] + ' - ' + job_salary_df['Role']
```

```
job_salary_df.drop(columns=['Job Title','Role'], inplace=True)
```

Transforming Location Column

Target mean encoding was applied to Location to reduce high-cardinality issues.

- Directly encodes the relationship between location and salary.
- Avoids falsely imposing an order on geographic locations.

```
location_mean = job_salary_df.groupby('Location')['Salary'].mean()

job_salary_df['Location Encoded'] = job_salary_df['Location'].map(location_mean)
```

```
job_salary_df.drop(columns=['Location'], inplace=True)
```

Transforming Work Type Column

Custom label encoding was applied to the Work Type column.

- Captures the ordinal relationship between different work types.
- Reduces unnecessary complexity and improves model efficiency.

```
print(job_salary_df["Work Type"].unique())
```

```
['Intern' 'Temporary' 'Full-Time' 'Contract' 'Part-Time']
```

```
work_type_mapping = {
    'Intern': 0,
    'Part-Time': 1,
    'Temporary': 2,
    'Contract': 3,
    'Full-Time': 4
}

# Apply Custom Label Encoding
job_salary_df['Work Type'] = job_salary_df['Work Type'].map(work_type_mapping)
```

Transforming Preference Column

Label encoding was applied to the Preference column.

- The column has low cardinality with only three unique categories.
- Encoding avoids unnecessary columns while reducing model complexity.

```
print(job_salary_df["Preference"].unique())
```

```
['Female' 'Male' 'Both']
```

```
job_salary_df['Preference'] = le.fit_transform(job_salary_df['Preference'])
```

```
print(job_salary_df["Qualifications"].unique())
```

```
['M.Tech' 'BCA' 'PhD' 'MBA' 'MCA' 'M.Com' 'BBA' 'B.Tech' 'B.Com' 'BA']
```

```
job_salary_df = pd.get_dummies(job_salary_df, columns=['Qualifications'], prefix='Qual')
```

Transforming Job Profile Column

Target mean encoding was applied to Job Profile to capture its effect on salary.

- This reduces high-cardinality issues while maintaining predictive power.
- Prevents falsely imposing an order on different job profiles.

```
print(job_salary_df["Job Profile"].nunique())
```

```
376
```

```
role_salary_mean = job_salary_df.groupby('Job Profile')['Salary'].mean()

job_salary_df['Job Profile Encoded'] = job_salary_df['Job Profile'].map(role_salary_mean)
```

```
job_salary_df.drop(columns=['Job Profile'], inplace=True)
```

Transforming Benefits Column

One-hot encoding was applied to the Benefits column.

- Since benefits do not have a natural order, one-hot encoding is the best approach.
- It allows the model to learn the effect of each benefit separately, improving accuracy.

```
print(job_salary_df["Benefits"].nunique())
```

```
11
```

```
job_salary_df['Benefits List'] = job_salary_df['Benefits'].apply(lambda x: [b.strip() for b in x.strip('{}').split(',')])

# Apply Binarizer
benefits_encoded = mlb.fit_transform(job_salary_df['Benefits List'])

# Create DataFrame of encoded benefits
benefits_df = pd.DataFrame(benefits_encoded, columns=[f'Benefit_{b}' for b in mlb.classes_])

# Concatenate back to original DataFrame
job_salary_df = pd.concat([job_salary_df, benefits_df], axis=1)

#  Drop Original Benefits columns
job_salary_df.drop(['Benefits', 'Benefits List'], axis=1, inplace=True)
```

Transforming Company Column

Target mean encoding was applied to the Company column.

- Encoding was applied after saving the original DataFrame for visualization.
- It effectively handles high cardinality and improves predictive accuracy.

```
print(job_salary_df["Company"].nunique())
```

```
888
```

```
company_salary_mean = job_salary_df.groupby('Company')['Salary'].mean()

job_salary_df['Company Encoded'] = job_salary_df['Company'].map(company_salary_mean)
```

```
job_salary_df.drop('Company', axis=1, inplace=True)
```

Dataset Overview After Cleaning and Transformation

```
job_salary_df_visualization.describe()
```

	Company Size	Min Experience	Max Experience	Salary	Job Post Age (Days)	Job Posting Year
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	7.370591e+04	2.501410e+00	1.149676e+01	8.249193e+04	9.269160e+02	2.022206e+03
std	3.530437e+04	1.708398e+00	2.291408e+00	7.528770e+03	2.111549e+02	6.774980e-01
min	1.264600e+04	0.000000e+00	8.000000e+00	6.750000e+04	5.620000e+02	2.021000e+03
25%	4.312100e+04	1.000000e+00	9.000000e+00	7.600000e+04	7.440000e+02	2.022000e+03
50%	7.364100e+04	3.000000e+00	1.100000e+01	8.250000e+04	9.270000e+02	2.022000e+03
75%	1.043175e+05	4.000000e+00	1.300000e+01	8.900000e+04	1.110000e+03	2.023000e+03
max	1.348340e+05	5.000000e+00	1.500000e+01	9.750000e+04	1.292000e+03	2.023000e+03

```
job_salary_df_visualization.head()
```

	Qualifications	Work Type	Company Size	Preference	Benefits	Company	Min Experience	Max Experience	Salary	Job Post Age (Days)	Job Posting Year	Location	Job Profile
0	M.Tech	Intern	26801	Female	{'Flexible Spending Accounts (FSAs), Relocatio...	Icahn Enterprises	5	15	79000.0	1071	2022	Isle of Man - Douglas	Digital Marketing Specialist - Social Media Ma...
1	BCA	Intern	100340	Female	{'Health Insurance, Retirement Plans, Paid Tim...	PNC Financial Services Group	2	12	86000.0	832	2022	Turkmenistan - Ashgabat	Web Developer - Frontend Web Developer
2	PhD	Temporary	84525	Male	{'Legal Assistance, Bonuses and Incentive Prog...	United Services Automobile Assn.	0	12	82500.0	928	2022	Macao SAR, China - Macao	Operations Manager - Quality Control Manager
3	PhD	Full-Time	129896	Female	{'Transportation Benefits, Professional Develo...	Hess	4	11	78000.0	764	2023	Benin - Porto-Novo	Network Engineer - Wireless Network Engineer
4	MBA	Intern	53944	Female	{'Flexible Spending Accounts (FSAs), Relocatio...	Cairn Energy	1	12	75500.0	901	2022	Chile - Santiago	Event Manager - Conference Manager

Exploratory Data Analysis (EDA)

Univariate Analysis

This section examines individual features to understand their distribution, central tendency, and variability. This step is essential for identifying patterns, detecting outliers, and ensuring data consistency before exploring relationships between multiple variables.

Importance Of Univariate Analysis

- Helps understand the distribution of a single feature, allowing for better data interpretation.
- Identifies key statistical properties such as central tendency (mean, median, mode) and dispersion (variance, standard deviation, range).
- Detects outliers and anomalies that may impact downstream analyses.
- Provides fundamental insights before proceeding to more complex bivariate or multivariate analysis.

Insights from Univariate Analysis:

1. Salary Distribution:

- If a right-skewed histogram appears, most salaries are low, with some high-paying jobs.
- A box plot with outliers suggests a few exceptionally high salaries.

2. Experience Levels:

- If bimodal peaks exist, there may be two types of jobs: entry-level and senior.

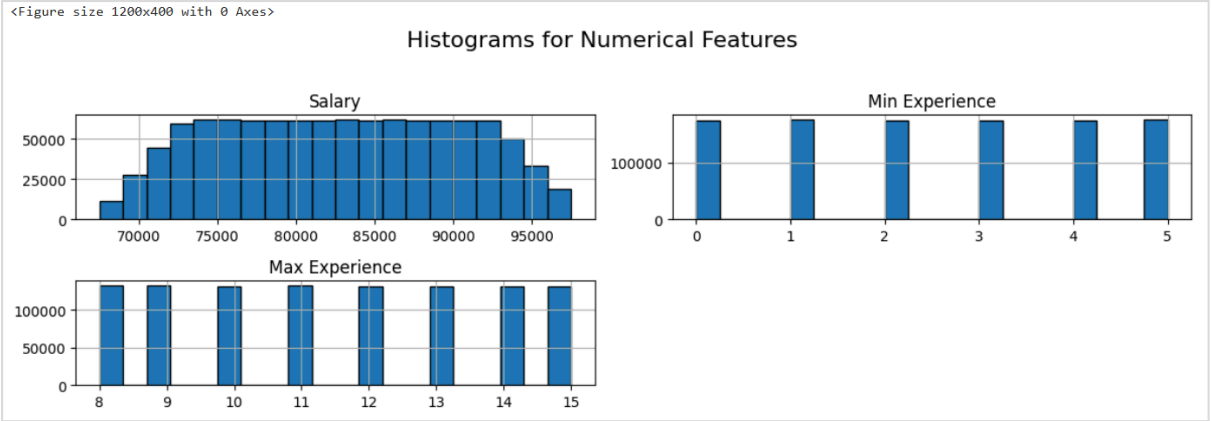
3. Categorical Variables (Job Titles, Locations):

- Bar charts can highlight the most common job roles & locations.
- Uneven distributions suggest demand varies across categories.

```
]: numerical_cols = ['Salary', 'Min Experience', 'Max Experience']

plt.figure(figsize=(12, 4))
job_salary_df_visualization[numerical_cols].hist(figsize=(12, 4), bins=20, edgecolor='black')
plt.suptitle("Histograms for Numerical Features", fontsize=16)
plt.tight_layout(rect=[0, 0, 1, 0.95])
plt.show()
```

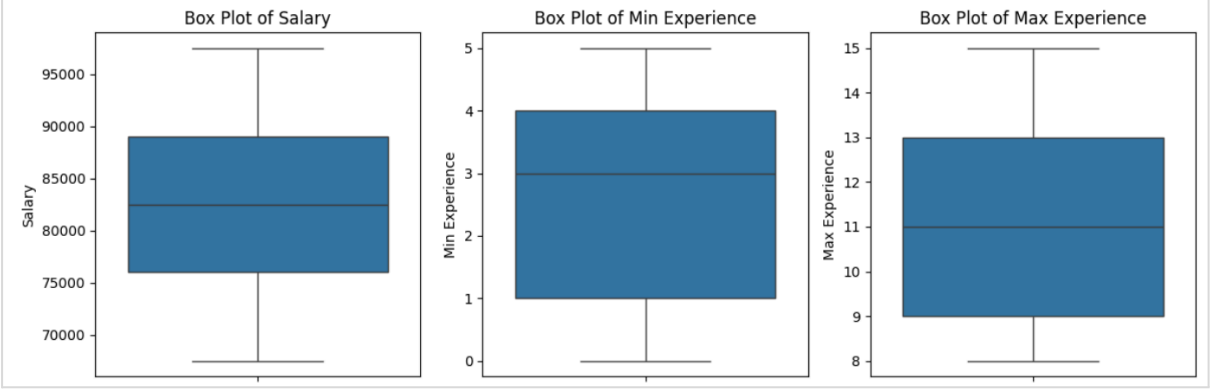
Histogram

- If a right-skewed histogram is observed, it suggests that most jobs offer lower salaries, with a few high-paying outliers.
- A bimodal distribution in experience data could indicate two distinct categories of jobs—entry-level and senior positions.
- The range and spread help in understanding the diversity of experience levels required across industries.



Box Plot

```
plt.figure(figsize=(12, 4))

for i, col in enumerate(numerical_cols):
    plt.subplot(1, len(numerical_cols), i+1)
    sns.boxplot(y=job_salary_df_visualization[col])
    plt.title(f"Box Plot of {col}")

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



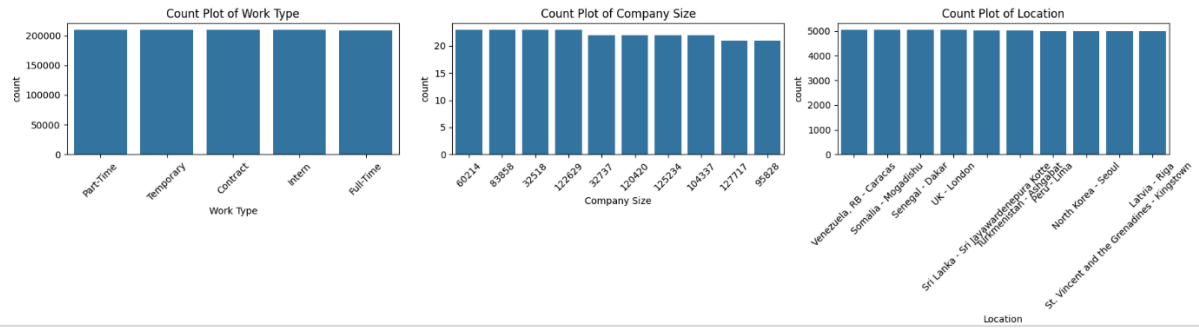
- Box plots may reveal extreme salary values, indicating roles with significantly higher pay.

Categorical Variables (Job Titles, Locations)

- Bar charts for job titles and locations can highlight the most common roles and geographic job concentrations.

- Uneven distributions suggest varying demand for specific roles or locations, which can influence job seekers' decision-making.

```
# Plot each categorical column
for i, col in enumerate(categorical_cols):
    plt.subplot(1, len(categorical_cols), i + 1)

    # Take only top 10 most frequent categories to avoid too many bars
    top_categories = job_salary_df_visualization[col].value_counts().nlargest(10).index

    sns.countplot(data=job_salary_df_visualization, x=col,
                  order=top_categories)

    plt.xticks(rotation=45)
    plt.title(f"Count Plot of {col}")

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



Bivariate Analysis

This section explores the relationships between two variables, allowing us to detect correlations, dependencies, and trends in the dataset. This analysis provides deeper insights into how different features interact, which is essential for predictive modeling and decision-making.

Insights from Bivariate Analysis:

1. Salary vs. Experience (Scatter Plot):

- A positive correlation means higher experience leads to higher salary.
- If scattered without pattern, experience does not strongly impact salary.

2. Salary vs. Job Type (Box Plot):

- If remote jobs have higher median salaries, companies pay more for remote flexibility.

3. Salary vs. Location (Bar Plot):

- If certain locations dominate, it shows regional salary variations.

4. Correlation Heatmap:

- If salary correlates strongly with experience but not company size, then experience matters more than the employer.

```
plt.figure(figsize=(8, 5))
sns.scatterplot(x=job_salary_df_visualization['Min Experience'], y=job_salary_df_visualization['Salary'], alpha=0.6)
plt.title("Scatter Plot: Salary vs. Min Experience")
plt.xlabel("Min Experience (Years)")
plt.ylabel("Salary")
plt.show()
```

Salary vs. Experience (Scatter Plot)

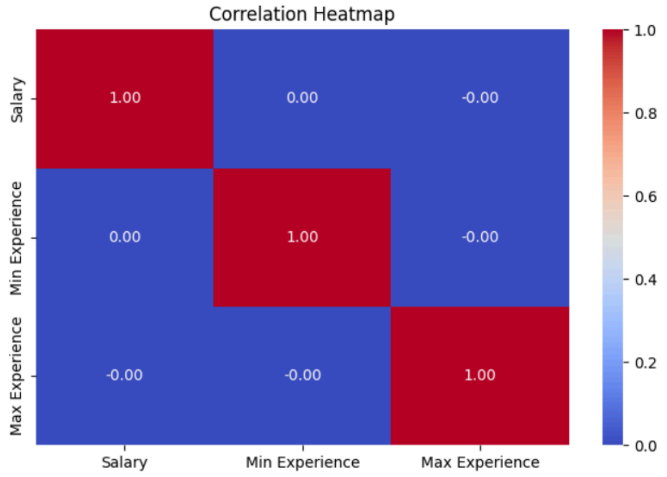
- A positive correlation suggests that higher experience generally leads to higher salaries.
- If the scatter plot appears randomly distributed, it indicates that experience does not have a strong impact on salary.

Correlation Heatmap

- If salary correlates strongly with experience but weakly with company size, it suggests that experience is a more significant factor in salary determination than the employer's size.
- This visualization helps in identifying which features have the strongest impact on salary and other key job attributes.

This figure is present on the next page.

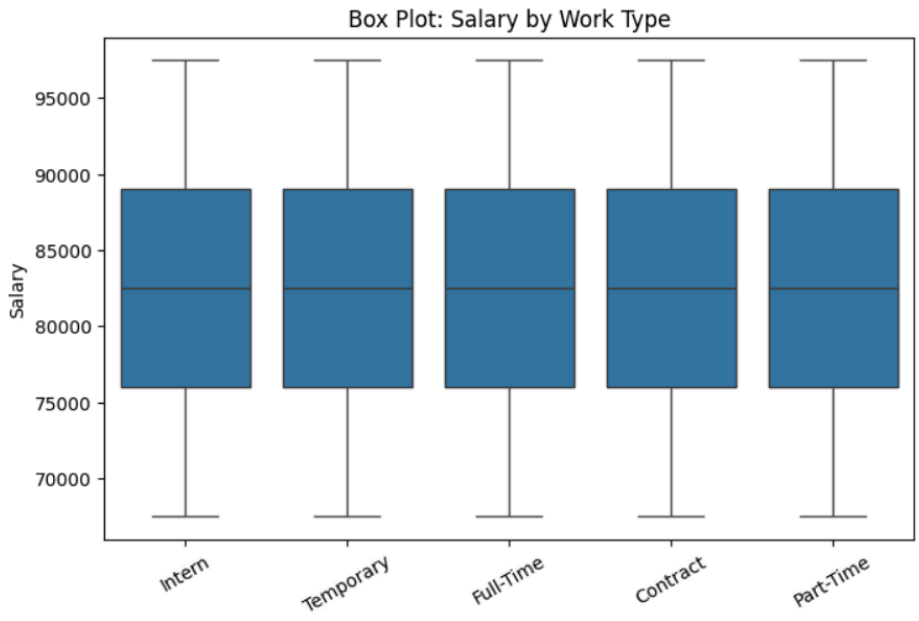
```
plt.figure(figsize=(8, 5))
sns.heatmap(job_salary_df_visualization[['Salary', 'Min Experience', 'Max Experience']].corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()
```



Salary vs. Job Type (Box Plot)

- If remote jobs show a higher median salary, it suggests that companies may offer competitive pay for remote flexibility.
- Significant variations across job types highlight potential salary disparities between full-time, part-time, contract, and internship roles.

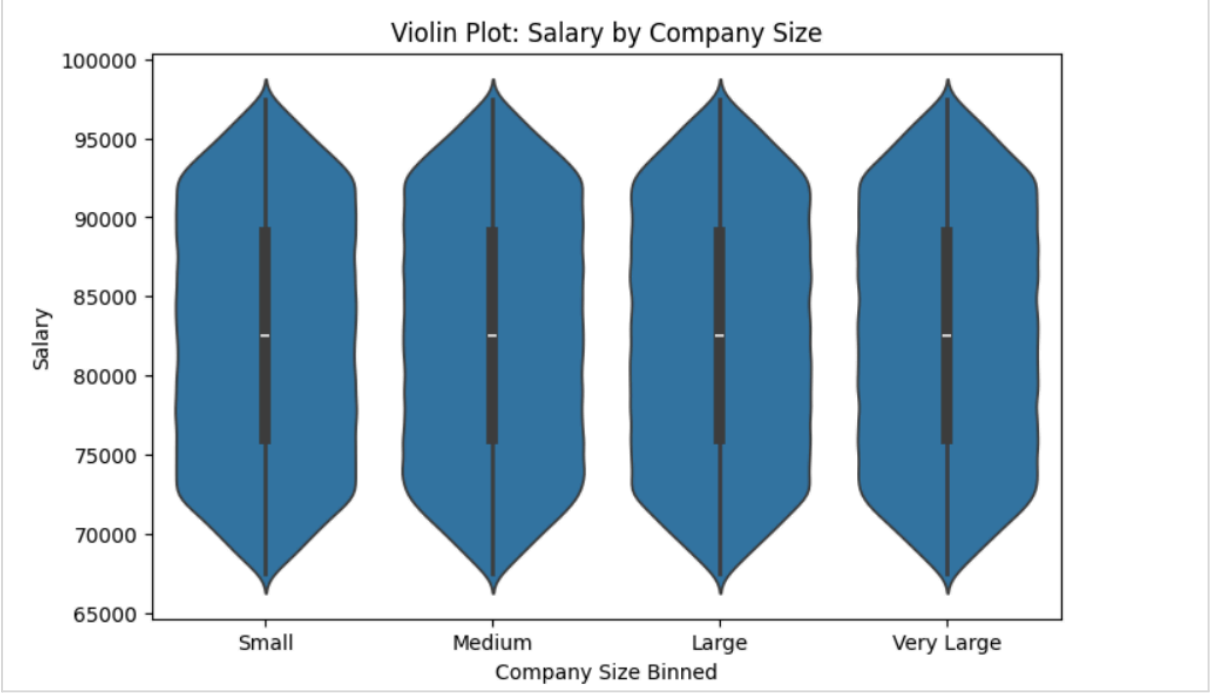
```
plt.figure(figsize=(8, 5))
sns.boxplot(data=job_salary_df_visualization, x='Work Type', y='Salary')
plt.xticks(rotation=30)
plt.title("Box Plot: Salary by Work Type")
plt.show()
```



Violin Plot

A violin plot for Salary by Company Size visualizes the distribution of salaries across different company sizes, combining aspects of a box plot and a density plot. If the violin plot shows wider sections at higher salary ranges for large companies, it indicates that larger companies offer more high-paying jobs, while narrower sections at lower salary ranges suggest smaller companies have lower salary variability.

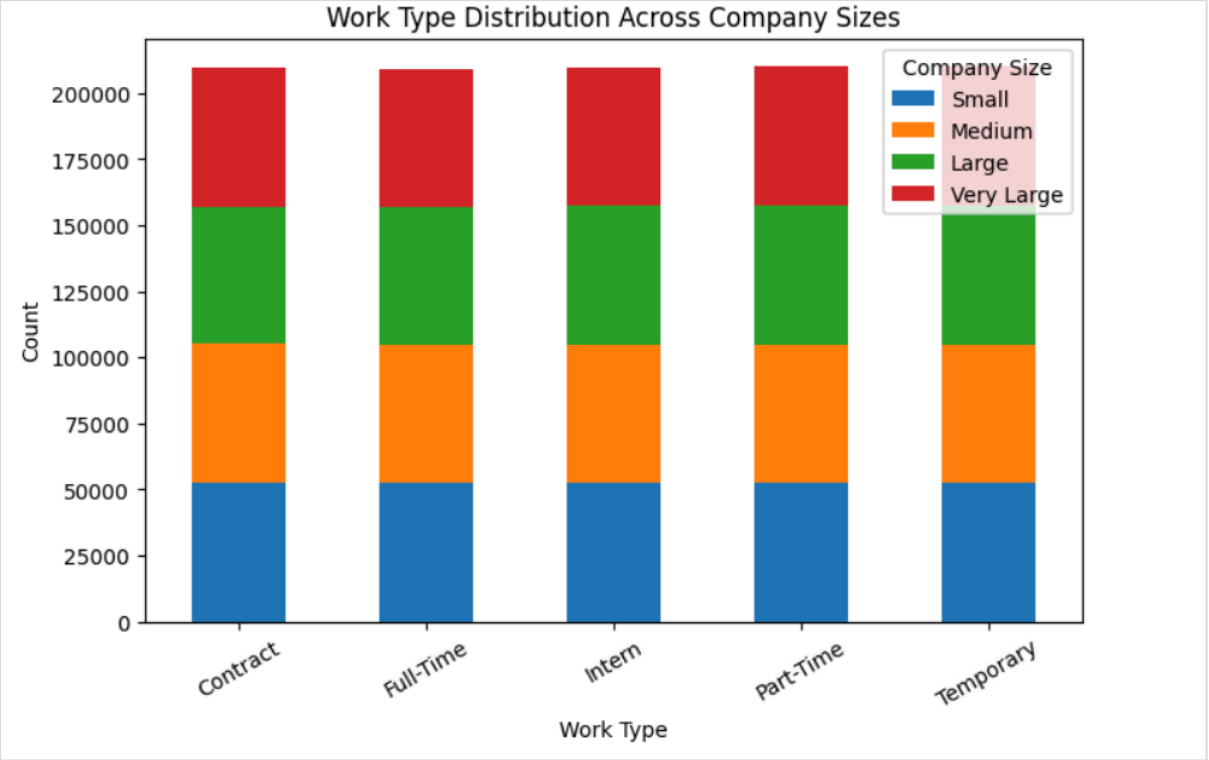
```
# Plot violin plot
plt.figure(figsize=(8, 5))
sns.violinplot(data=job_salary_df_visualization, x='Company Size Binned', y='Salary')
plt.title("Violin Plot: Salary by Company Size")
plt.show()
```




```
# Convert 'Company Size' to numeric if needed
job_salary_df_visualization['Company Size'] = pd.to_numeric(job_salary_df_visualization['Company Size'], errors='coerce')

# Create bins for Company Size (quartiles)
job_salary_df_visualization['Company Size Binned'] = pd.qcut(
    job_salary_df_visualization['Company Size'], q=4, labels=['Small', 'Medium', 'Large', 'Very Large']
)

# Crosstab and stacked bar plot
pd.crosstab(job_salary_df_visualization['Work Type'], job_salary_df_visualization['Company Size Binned']).plot(
    kind="bar", stacked=True, figsize=(8, 5)
)
plt.title("Work Type Distribution Across Company Sizes")
plt.ylabel("Count")
plt.xticks(rotation=30)
plt.legend(title="Company Size")
plt.show()
```

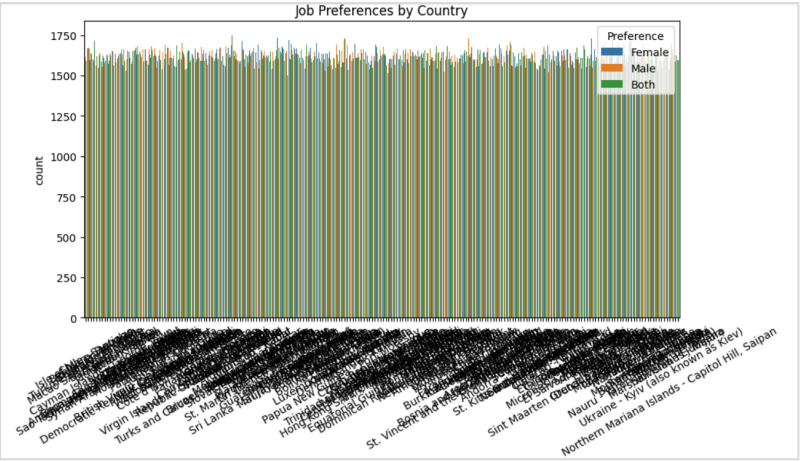
Stacked Bar Chart

This stacked bar chart visualizes the distribution of work types (e.g., Full-Time, Part-Time) across different company sizes (Small, Medium, Large, Very Large). It helps identify trends, such as whether larger companies offer more full-time roles.



Count Plot

This count plot displays the distribution of job postings across different locations, categorized by preference (e.g., Male, Female, Both). It highlights regional variations in job preferences.



Multivariate Analysis

This section analyzes multiple variables simultaneously to uncover complex interactions, providing deeper insights into the dataset's structure.

3. Multivariate Analysis

Reasons for Multivariate Analysis:

- Shows deeper insights by combining multiple factors.
- Helps in decision-making & predictive modeling.

Insights from Multivariate Analysis:

1. 3D Scatter Plot (Salary, Experience, Company Size):

- If large companies cluster at higher salaries, size affects salary significantly.

2. Grouped Box Plot (Work Type & Company Size vs. Salary):

- If remote jobs at big companies have high median salaries, it confirms that top firms pay more for flexibility.

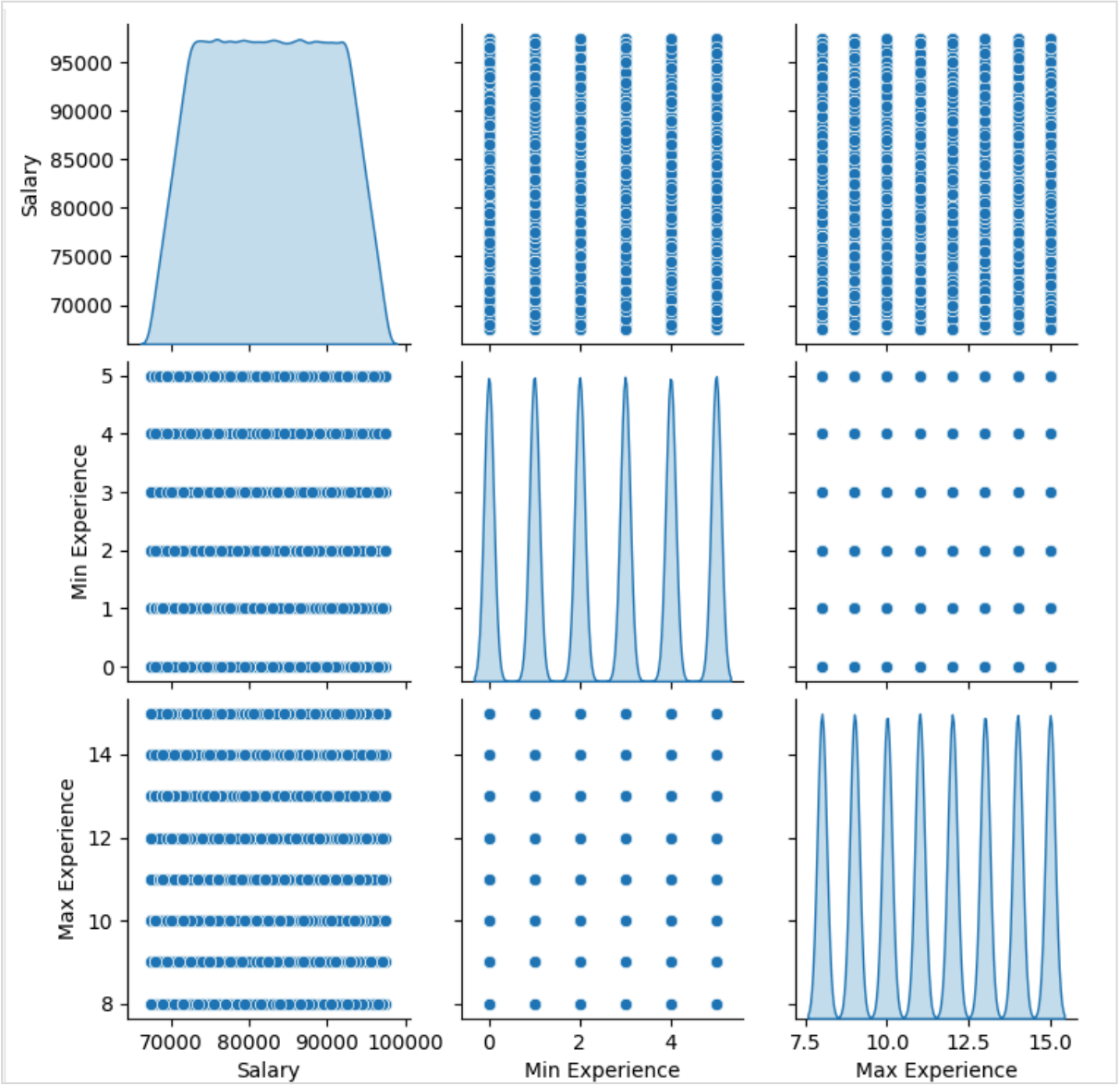
3. Heatmap (Preferences vs. Work Type):

- If most job seekers prefer remote work, companies may adjust hiring policies.

4. FacetGrid (Salary by Location & Work Type):

- If some locations show high salaries for specific work types, job seekers can target those regions.

```
sns.pairplot(job_salary_df_visualization[['Salary', 'Min Experience', 'Max Experience']], diag_kind='kde')
plt.show()
```



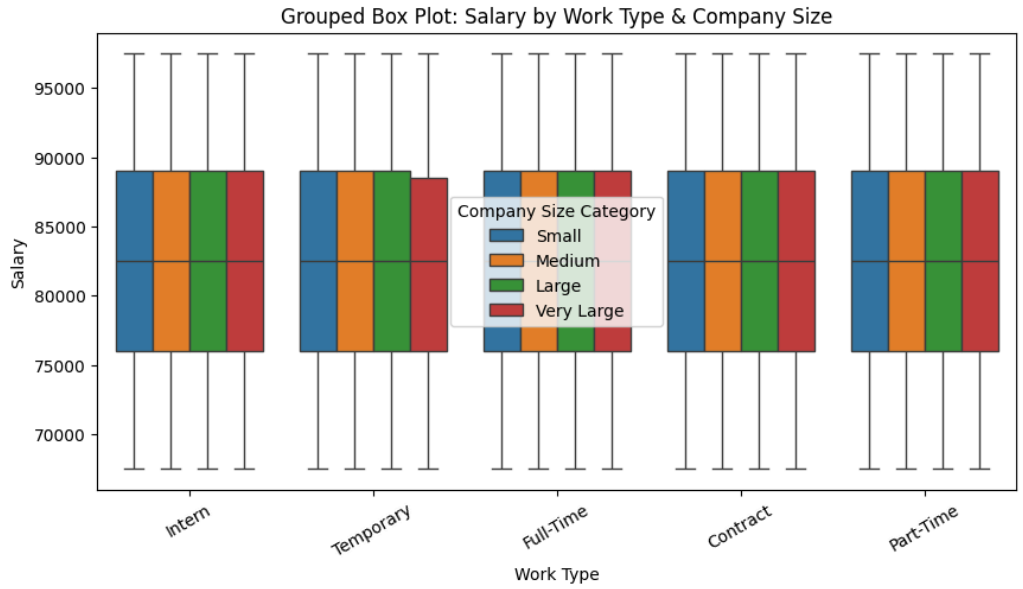
3D Scatter Plot (Salary, Experience, Company Size)

- If data points for large companies cluster at higher salary levels, it indicates that company size significantly impacts salary.
- A weak or scattered relationship suggests other factors influence salary more than company size.

Grouped Box Plot (Work Type & Company Size vs. Salary)

- If remote jobs in large companies show higher median salaries, it confirms that top firms offer better pay for flexibility.
- Significant variation across company sizes can indicate different compensation structures based on work type.

```
# Bin 'Company Size' into categories
job_salary_df_visualization['Company Size Binned'] = pd.qcut(
    job_salary_df_visualization['Company Size'], q=4, labels=['Small', 'Medium', 'Large', 'Very Large']
)

# Create boxplot
plt.figure(figsize=(10, 5))
sns.boxplot(data=job_salary_df_visualization, x='Work Type', y='Salary', hue='Company Size Binned')
plt.xticks(rotation=30)
plt.title("Grouped Box Plot: Salary by Work Type & Company Size")
plt.legend(title="Company Size Category")
plt.show()
```



FacetGrid (Salary by Location & Work Type)

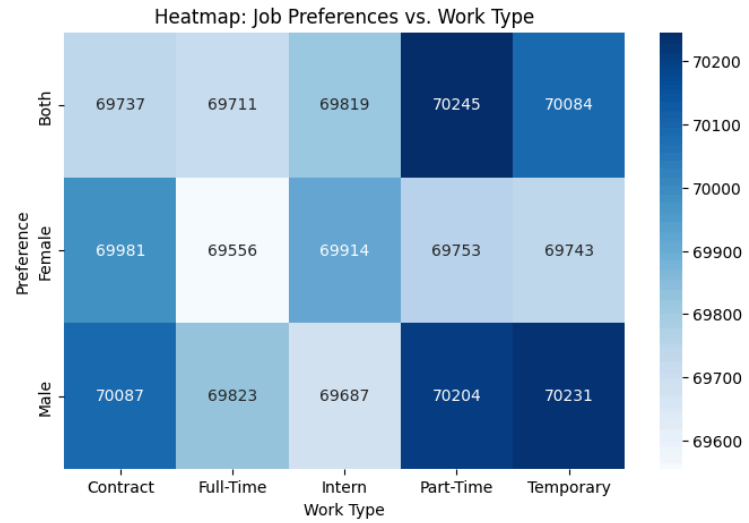
- If certain locations consistently show higher salaries for specific work types, it suggests geographical salary trends that job seekers can leverage.
- This can help professionals strategically apply for jobs in high-paying regions based on work type.

This figure is present on the next page.

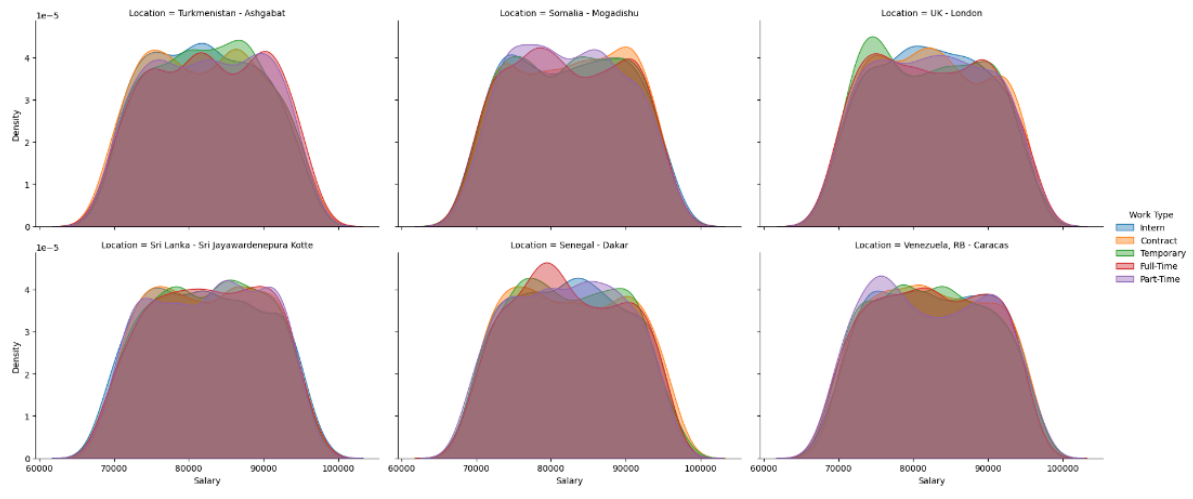
Heatmap (Preferences vs. Work Type)

- If a strong preference for remote jobs is observed, it suggests that job seekers prioritize flexibility, influencing hiring policies.
- Companies can use this insight to adjust recruitment strategies to attract talent.

```
heatmap_data = pd.crosstab(job_salary_df_visualization['Preference'], job_salary_df_visualization['Work Type'])
plt.figure(figsize=(8, 5))
sns.heatmap(heatmap_data, annot=True, cmap="Blues", fmt="d")
plt.title("Heatmap: Job Preferences vs. Work Type")
plt.show()
```




```
# Select top 6 locations based on job count
top_locations = job_salary_df_visualization["Location"].value_counts().index[:6]
filtered_df = job_salary_df_visualization[job_salary_df_visualization["Location"].isin(top_locations)]

# Create FacetGrid
g = sns.FacetGrid(filtered_df, col="Location", hue="Work Type", height=4, aspect=1.5, col_wrap=3)
g.map(sns.kdeplot, "Salary", fill=True, alpha=0.4)

# Add legend and show plot
g.add_legend()
plt.show()
```



Feature Analysis

Feature Analysis

```
job_salary_df_visualization.head()
```

	Qualifications	Work Type	Company Size	Preference	Benefits	Company	Min Experience	Max Experience	Salary	Job Post Age (Days)	Job Posting Year	Location	Job Profile	Company Size Binned
0	M.Tech	Intern	26801	Female	{'Flexible Spending Accounts (FSAs), Relocatio...	Icahn Enterprises	5	15	79000.0	1071	2022	Isle of Man - Douglas	Digital Marketing Specialist - Social Media Ma...	Small
1	BCA	Intern	100340	Female	{'Health Insurance, Retirement Plans, Paid Tim...	PNC Financial Services Group	2	12	86000.0	832	2022	Turkmenistan - Ashgabat	Web Developer - Frontend Web Developer	Large
2	PhD	Temporary	84525	Male	{'Legal Assistance, Bonuses and Incentive Prog...	United Services Automobile Assn.	0	12	82500.0	928	2022	Macao SAR, China - Macao	Operations Manager - Quality Control Manager	Large
3	PhD	Full-Time	129896	Female	{'Transportation Benefits, Professional Develo...	Hess	4	11	78000.0	764	2023	Benin - Porto-Novo	Network Engineer - Wireless Network Engineer	Very Large
4	MBA	Intern	53944	Female	{'Flexible Spending Accounts (FSAs), Relocatio...	Cairn Energy	1	12	75500.0	901	2022	Chile - Santiago	Event Manager - Conference Manager	Medium

Feature Importance Analysis

Feature importance analysis helps identify the most influential variables in predicting salary or other target outcomes. By training a **Random Forest Regressor** or **Decision Tree**, we can extract feature importance scores and rank predictors based on their contribution. A bar chart visualization provides a clear representation of which features have the highest impact on model predictions.


```
job_salary_df_visualization.head()
```

	Qualifications	Work Type	Company Size	Preference	Benefits	Company	Min Experience	Max Experience	Salary	Job Post Age (Days)	Job Posting Year	Location	Job Profile	Company Size Binned
0	M.Tech	Intern	26801	Female	{'Flexible Spending Accounts (FSAs), Relocatio...	Icahn Enterprises	5	15	79000.0	1071	2022	Isle of Man - Douglas	Digital Marketing Specialist - Social Media Ma...	Small
1	BCA	Intern	100340	Female	{'Health Insurance, Retirement Plans, Paid Tim...	PNC Financial Services Group	2	12	86000.0	832	2022	Turkmenistan - Ashgabat	Web Developer - Frontend Web Developer	Large
2	PhD	Temporary	84525	Male	{'Legal Assistance, Bonuses and Incentive Prog...	United Services Automobile Assn.	0	12	82500.0	928	2022	Macao SAR, China - Macao	Operations Manager - Quality Control Manager	Large
3	PhD	Full-Time	129896	Female	{'Transportation Benefits, Professional Develo...	Hess	4	11	78000.0	764	2023	Benin - Porto-Novo	Network Engineer - Wireless Network Engineer	Very Large
4	MBA	Intern	53944	Female	{'Flexible Spending Accounts (FSAs), Relocatio...	Cairn Energy	1	12	75500.0	901	2022	Chile - Santiago	Event Manager - Conference Manager	Medium

```
job_salary_df_visualization.info()
```

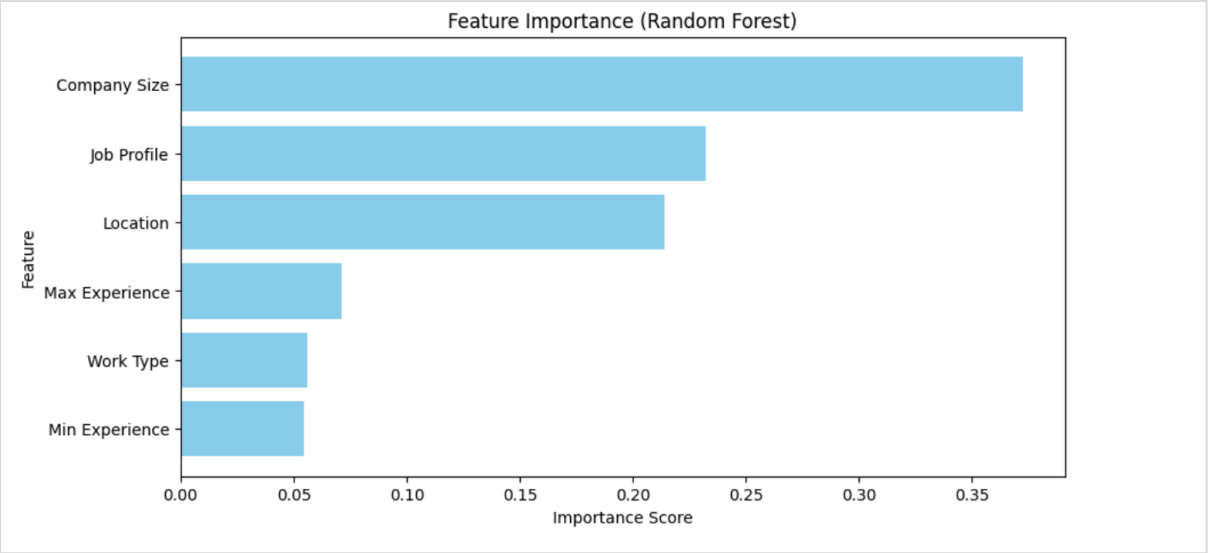
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 14 columns):
 #   Column               Non-Null Count    Dtype
---  ------               --------------    -----
 0   Qualifications       1048575 non-null  object
 1   Work Type            1048575 non-null  object
 2   Company Size         1048575 non-null  int64
 3   Preference           1048575 non-null  object
 4   Benefits             1048575 non-null  object
 5   Company              1048575 non-null  object
 6   Min Experience       1048575 non-null  int64
 7   Max Experience       1048575 non-null  int64
 8   Salary               1048575 non-null  float64
 9   Job Post Age (Days)  1048575 non-null  int64
 10  Job Posting Year     1048575 non-null  int32
 11  Location             1048575 non-null  object
 12  Job Profile          1048575 non-null  object
 13  Company Size Binned  1048575 non-null  category
dtypes: category(1), float64(1), int32(1), int64(4), object(7)
memory usage: 101.0+ MB
```

```
# Sample 20% of data to speed up training
df_sample = job_salary_df_visualization.sample(frac=0.2, random_state=42)

# Define features and target
features = ['Min Experience', 'Max Experience', 'Company Size', 'Work Type', 'Job Profile', 'Location']
df_encoded = df_sample.copy()

# Encode only categorical columns
label_encoders = {}
for col in ['Work Type', 'Job Profile', 'Location']:
    le = LabelEncoder()
    df_encoded[col] = le.fit_transform(df_encoded[col])
    label_encoders[col] = le  # Store encoders for later use (optional)

# Define X (features) and y (target)
X = df_encoded[features]
y = df_encoded['Salary']

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the optimized model
rf = RandomForestRegressor(n_estimators=25, max_depth=10, random_state=42, n_jobs=-1)
rf.fit(X_train, y_train)

# Show dataset shape
print(f"Training set: {X_train.shape}, Testing set: {X_test.shape}")
```

```
Training set: (167772, 6), Testing set: (41943, 6)
```

```
importances = rf.feature_importances_

feature_importance_df = pd.DataFrame({'Feature': X.columns, 'Importance': importances})
feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=False)
```

```
plt.figure(figsize=(10, 5))
plt.barh(feature_importance_df['Feature'], feature_importance_df['Importance'], color='skyblue')
plt.xlabel("Importance Score")
plt.ylabel("Feature")
plt.title("Feature Importance (Random Forest)")
plt.gca().invert_yaxis()
plt.show()
```



Continuous Features Correlation Analysis

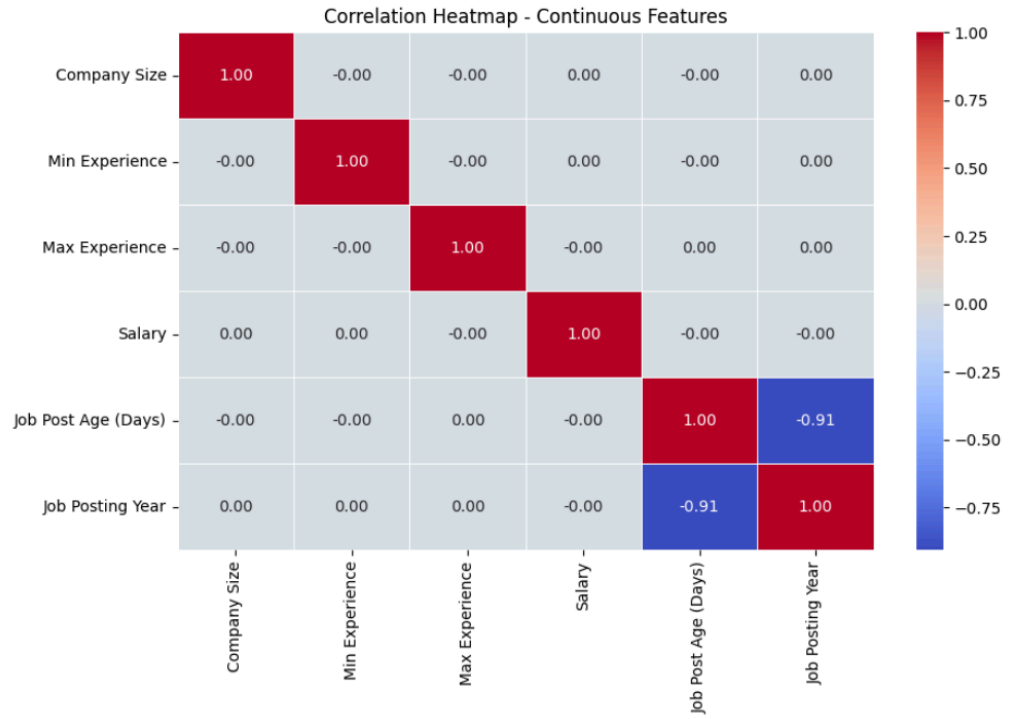
This analysis helps identify relationships between numerical features (e.g., Experience, Company Size, Salary Range) to determine their impact on salary predictions. By computing a correlation matrix and visualizing it through a heatmap, we can detect strong predictors and potential multicollinearity issues in the dataset.

```
31]: # Exclude non-numeric columns
     df_numeric = job_salary_df_visualization.select_dtypes(include=['number'])

     # Compute correlation matrix
     corr_matrix = df_numeric.corr()

     # Display correlation matrix
     print(corr_matrix)
```

```
                    Company Size  Min Experience  Max Experience    Salary  \
Company Size            1.000000       -0.000583       -0.001433  0.000469
Min Experience         -0.000583        1.000000       -0.000467  0.002010
Max Experience         -0.001433       -0.000467        1.000000 -0.000366
Salary                  0.000469        0.002010       -0.000366  1.000000
Job Post Age (Days)    -0.002082       -0.001178        0.000082 -0.000325
Job Posting Year        0.001335        0.001806        0.000498 -0.000043

                    Job Post Age (Days)  Job Posting Year
Company Size                  -0.002082          0.001335
Min Experience                -0.001178          0.001806
Max Experience                 0.000082          0.000498
Salary                        -0.000325         -0.000043
Job Post Age (Days)            1.000000         -0.906306
Job Posting Year              -0.906306          1.000000
```

Career Development - Job Salary Prediction

```
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap - Continuous Features")
plt.show()
```




```
salary_corr = corr_matrix['Salary'].drop('Salary').sort_values(ascending=False)
salary_corr
```

	Salary
Min Experience	0.002010
Company Size	0.000469
Job Posting Year	-0.000043
Job Post Age (Days)	-0.000325
Max Experience	-0.000366

dtype: float64

```
strong_corr = salary_corr[abs(salary_corr) > 0.3]
strong_corr
```

	Salary

dtype: float64

```
strong_features = strong_corr.index.tolist()

for feature in strong_features:
    plt.figure(figsize=(6, 4))
    sns.scatterplot(x=job_salary_df_visualization[feature], y=job_salary_df_visualization['Salary'], alpha=0.5)
    plt.xlabel(feature)
    plt.ylabel("Salary")
    plt.title(f"Salary vs {feature}")
    plt.show()
```

Categorical Features Analysis

This analysis examines how categorical variables such as Job Roles, Locations, Work Types, and Qualifications influence salary trends. By using groupby() to compute summary statistics (mean, median, count) and visualizing with box plots, bar charts, and violin plots, we can uncover salary disparities across different categories.

```
cat_features = ['Job Profile', 'Work Type', 'Location', 'Min Experience', 'Max Experience']

for feature in cat_features:
    grouped_data = job_salary_df_visualization.groupby(feature)['Salary'].agg(['mean', 'median', 'count'])
    print(f"\nSalary Summary for {feature}:\n")
    display(grouped_data.sort_values(by='mean', ascending=False))
```

```
Salary Summary for Job Profile:
```

Job Profile	mean	median	count
Physical Therapist - Geriatric Physical Therapist	82865.859130	83000.0	2229
Art Director - Visual Designer	82851.513802	83000.0	2246
Finance Manager - Treasury Manager	82842.177493	83000.0	2186
Network Engineer - Wireless Network Engineer	82830.108011	83000.0	2222
QA Analyst - Software QA Tester	82826.367274	83000.0	2249
...	...	...	...
Java Developer - Java Backend Developer	82156.042497	82000.0	2259
Pediatrician - Pediatric Specialist	82144.142168	82000.0	2279
Front-End Engineer - JavaScript Developer	82136.114623	82000.0	2373
Sales Manager - Key Account Manager	82132.530120	82000.0	2241
Customer Service Representative - Call Center Agent	82016.122098	81500.0	2326

376 rows × 3 columns

This code calculates and displays a salary summary for each categorical feature in cat_features. It groups the dataset by each feature and computes the mean, median, and count of salaries within each category. Sorting by mean salary helps identify which categories have the highest and lowest average earnings, providing insights into salary trends based on job profiles, work types, locations, and experience levels.

Salary Summary for Work Type:

Work Type	mean	median	count
Full-Time	82497.752164	82500.0	209090
Intern	82494.558781	82500.0	209420
Part-Time	82493.009106	82500.0	210202
Temporary	82489.536223	82500.0	210058
Contract	82484.809704	82500.0	209805

Salary Summary for Location:

Location	mean	median	count
American Samoa - Apia	82737.040816	83000.0	4900
Italy - Rome	82731.471977	83000.0	4871
Timor-Leste - Dili	82729.441306	83000.0	4779
Bosnia and Herzegovina - Sarajevo	82726.531875	82500.0	4847
Honduras - Tegucigalpa	82724.819773	83000.0	4855
...	...	...	...
Mozambique - Maputo	82292.049249	82000.0	4792
Israel - Jerusalem	82289.943313	82000.0	4763
Sweden - Stockholm	82286.745273	82000.0	4919
Guyana - Georgetown	82271.672772	82000.0	4914
Turkey - Ankara	82256.506350	82000.0	4803

216 rows × 3 columns

Salary Summary for Min Experience:

Min Experience	mean	median	count
5	82513.936637	82500.0	175401
2	82513.615174	82500.0	174401
4	82503.289983	82500.0	174469
3	82486.691972	82500.0	174669
0	82470.302565	82500.0	174409
1	82463.743965	82500.0	175226

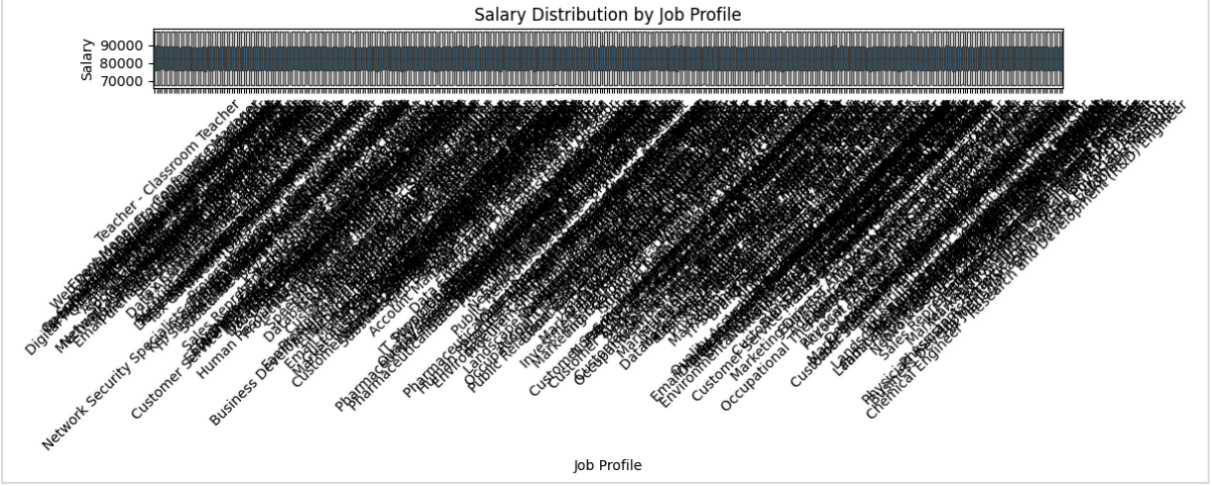
Salary Summary for Max Experience:

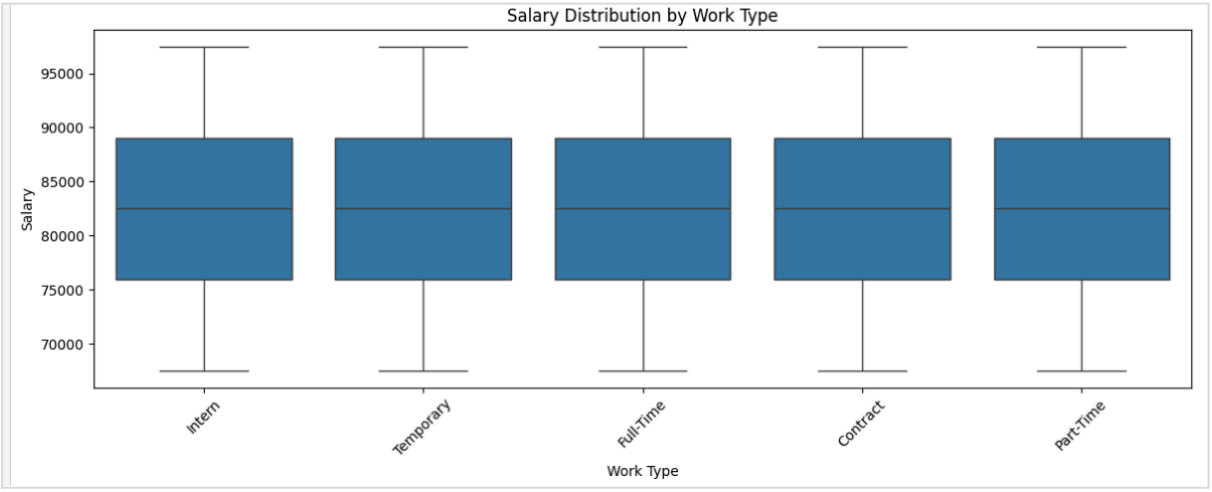
Max Experience	mean	median	count
15	82527.683922	82500.0	130816
11	82522.370361	82500.0	131558
8	82495.846961	82500.0	131470
10	82494.135077	82500.0	130607
12	82493.642195	82500.0	131256
9	82491.402877	82500.0	131323
14	82457.020451	82500.0	130946
13	82453.100713	82500.0	130599

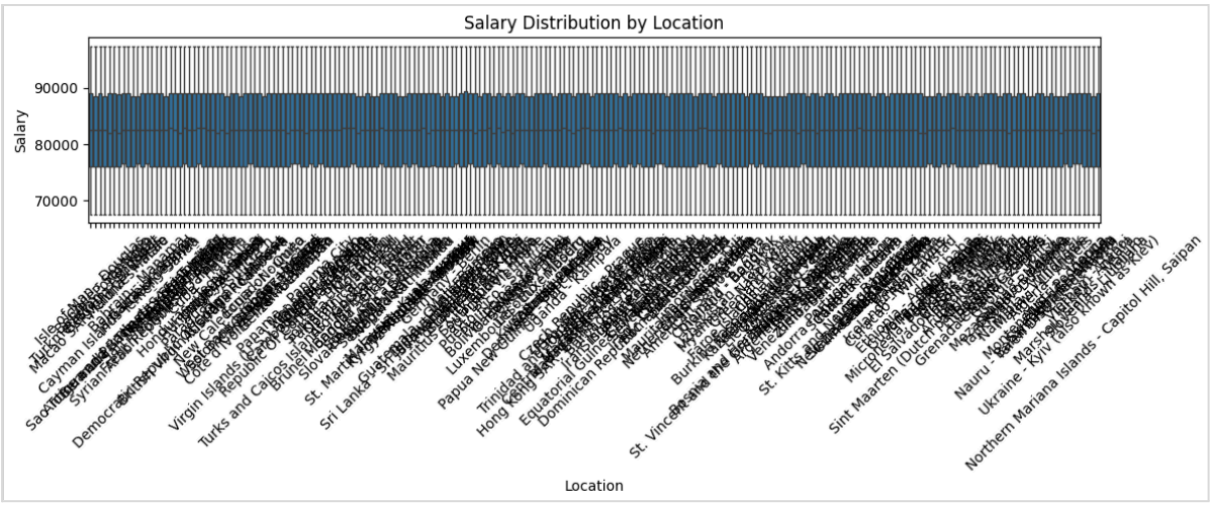
This box plot visualizes the salary distribution across different categories of a selected categorical feature (e.g., job roles, locations, work types). It highlights the median salary, interquartile range (IQR), and the presence of outliers, providing insights into salary variation and potential disparities within each category.

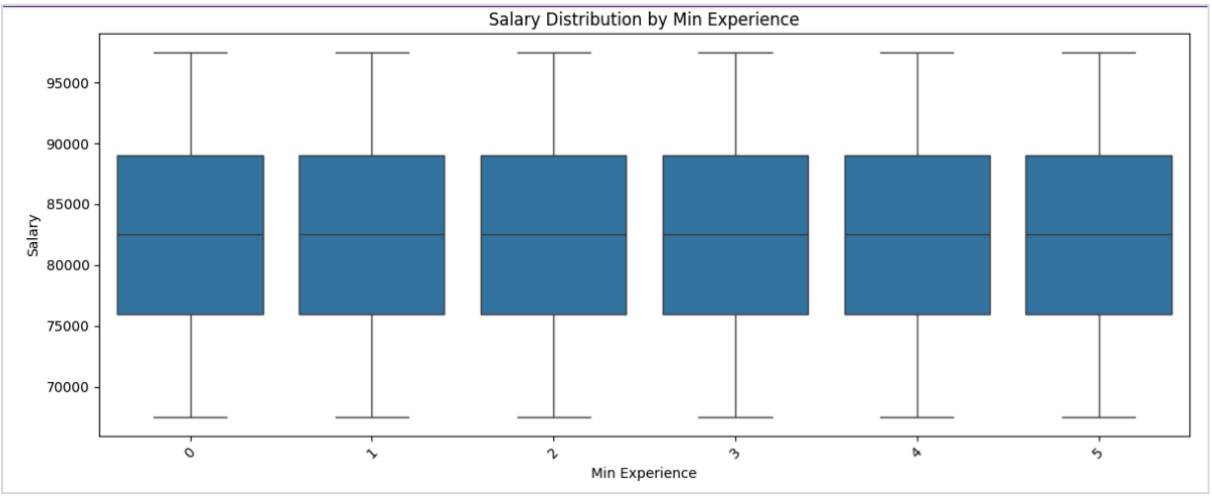
Career Development - Job Salary Prediction

```
for feature in cat_features:
    plt.figure(figsize=(12, 5))
    sns.boxplot(x=job_salary_df_visualization[feature], y=job_salary_df_visualization['Salary'])
    plt.xlabel(feature)
    plt.ylabel("Salary")
    plt.xticks(rotation=45)
    plt.title(f"Salary Distribution by {feature}")
    plt.tight_layout()
    plt.show()
```



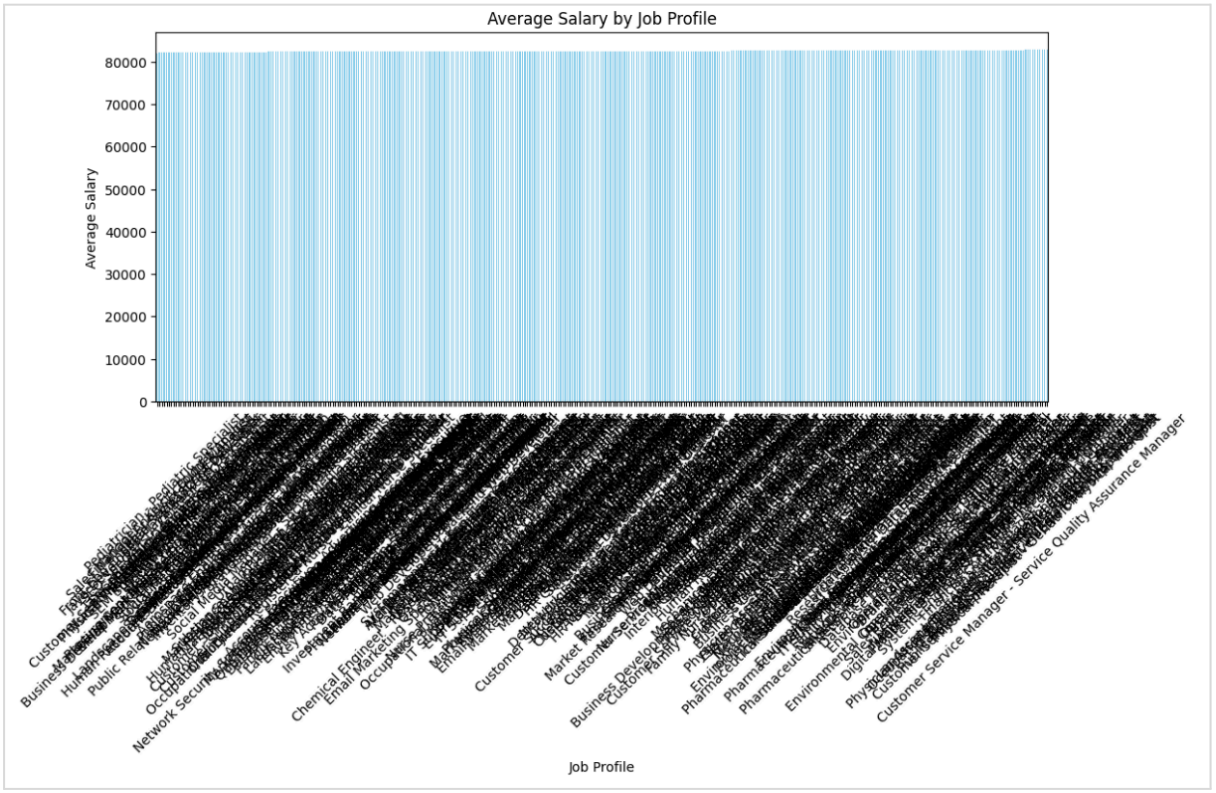


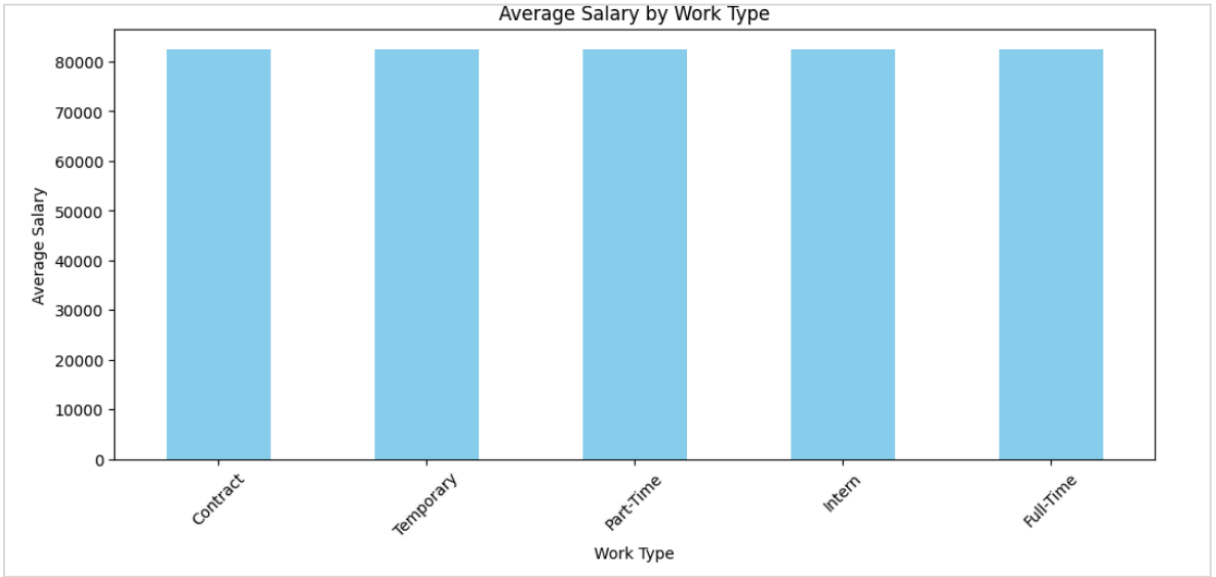


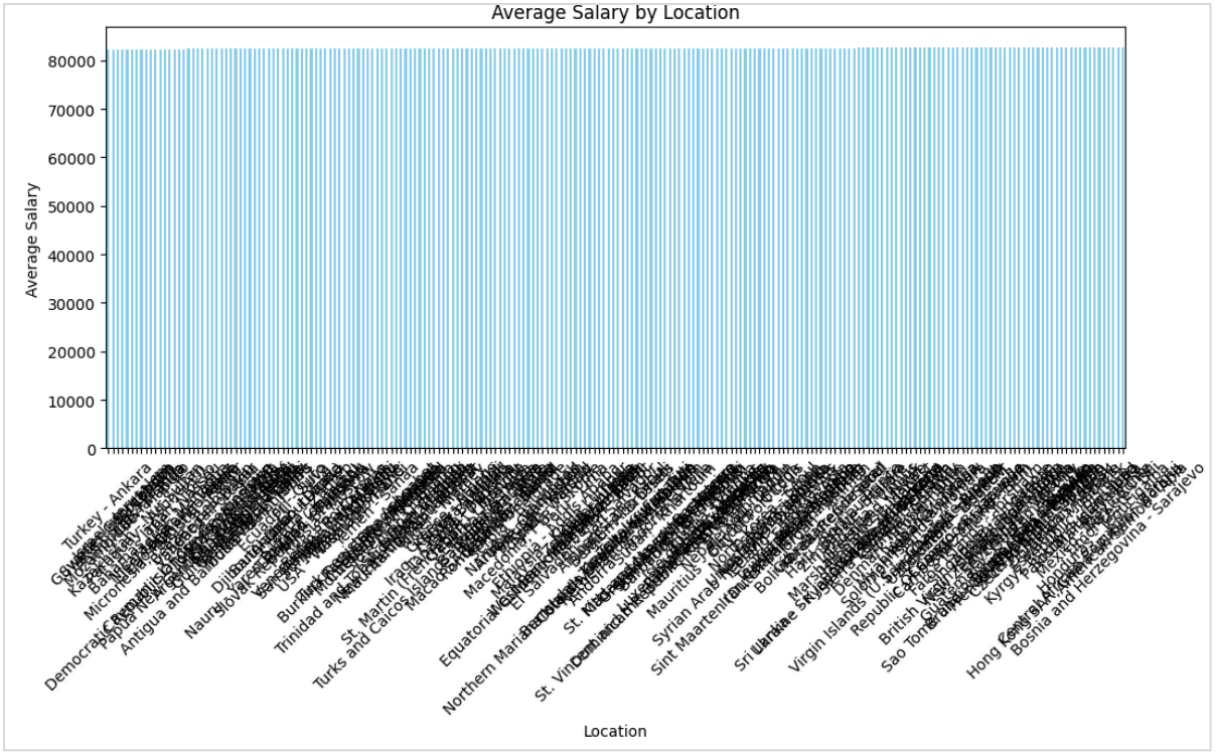


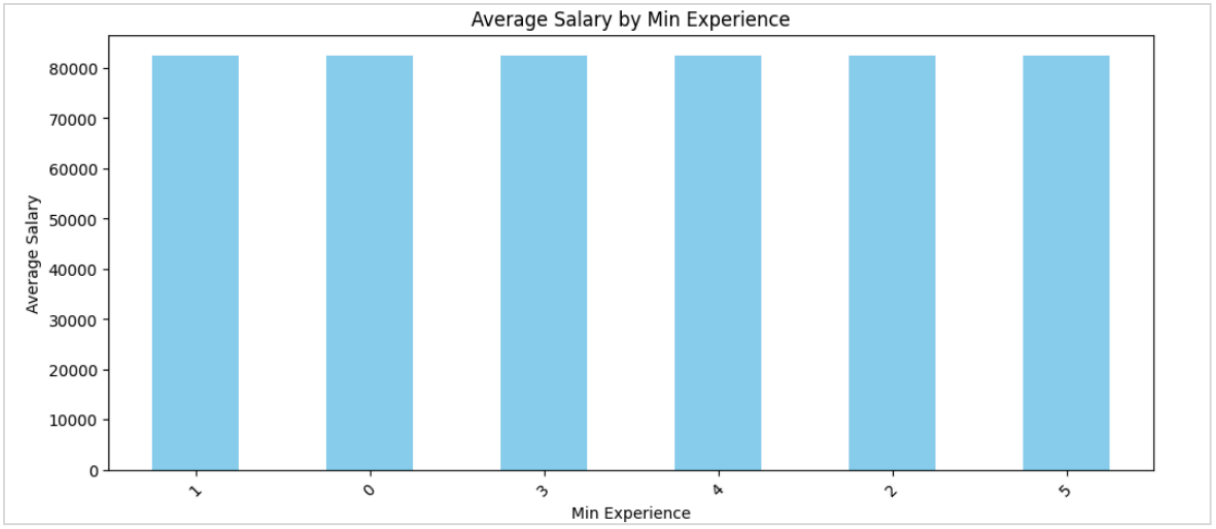
This bar chart displays the average salary for each category within a selected categorical feature (e.g., job roles, locations, work types). The bars represent the mean salary for each category, helping identify which groups tend to offer higher or lower salaries. Sorting the values makes it easier to compare salary trends across different categories

Career Development - Job Salary Prediction

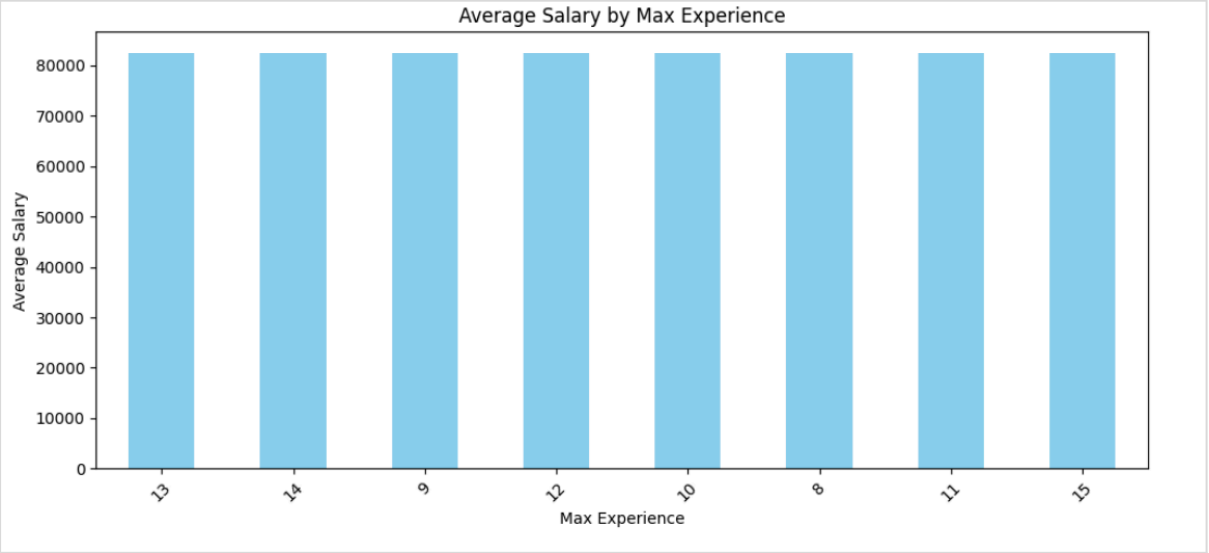




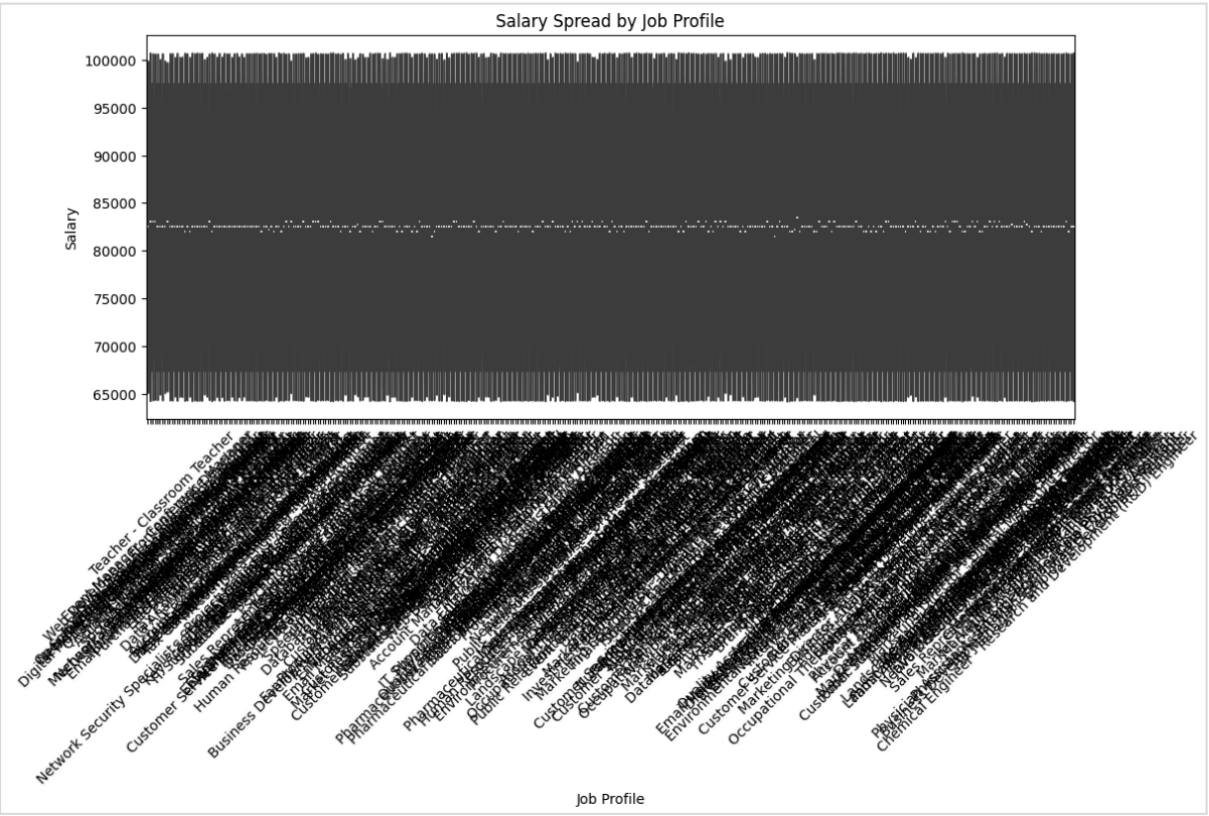

Average Salary by Location
Average Salary
0
10000
20000
30000
40000
50000
60000
70000
80000
Location

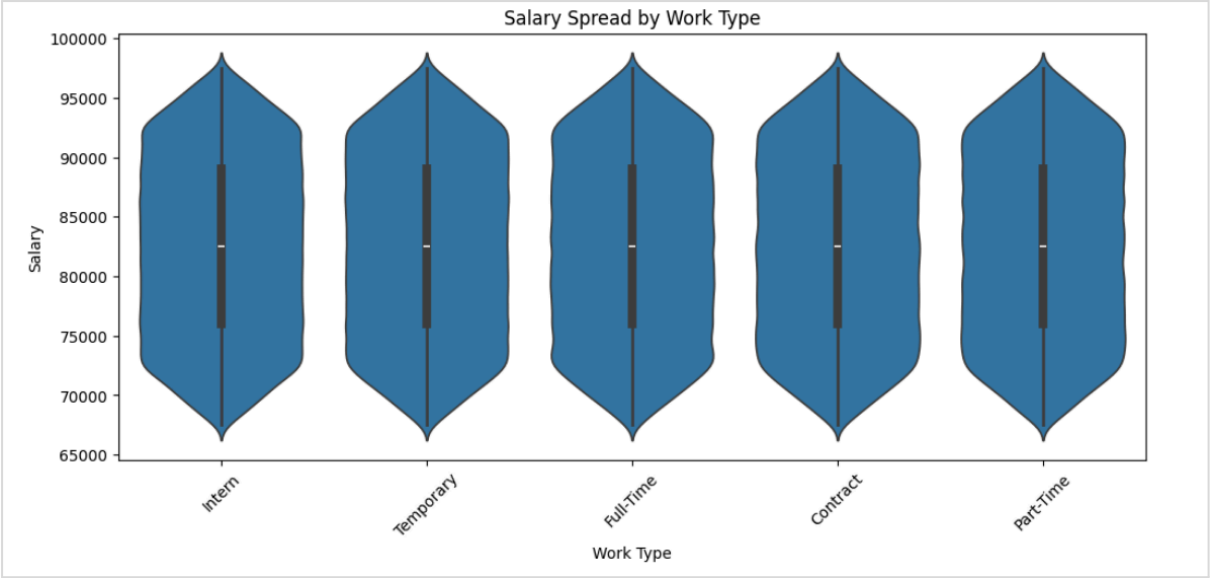

Average Salary by Min Experience
Average Salary
0
10000
20000
30000
40000
50000
60000
70000
80000
1
0
3
4
2
5
Min Experience

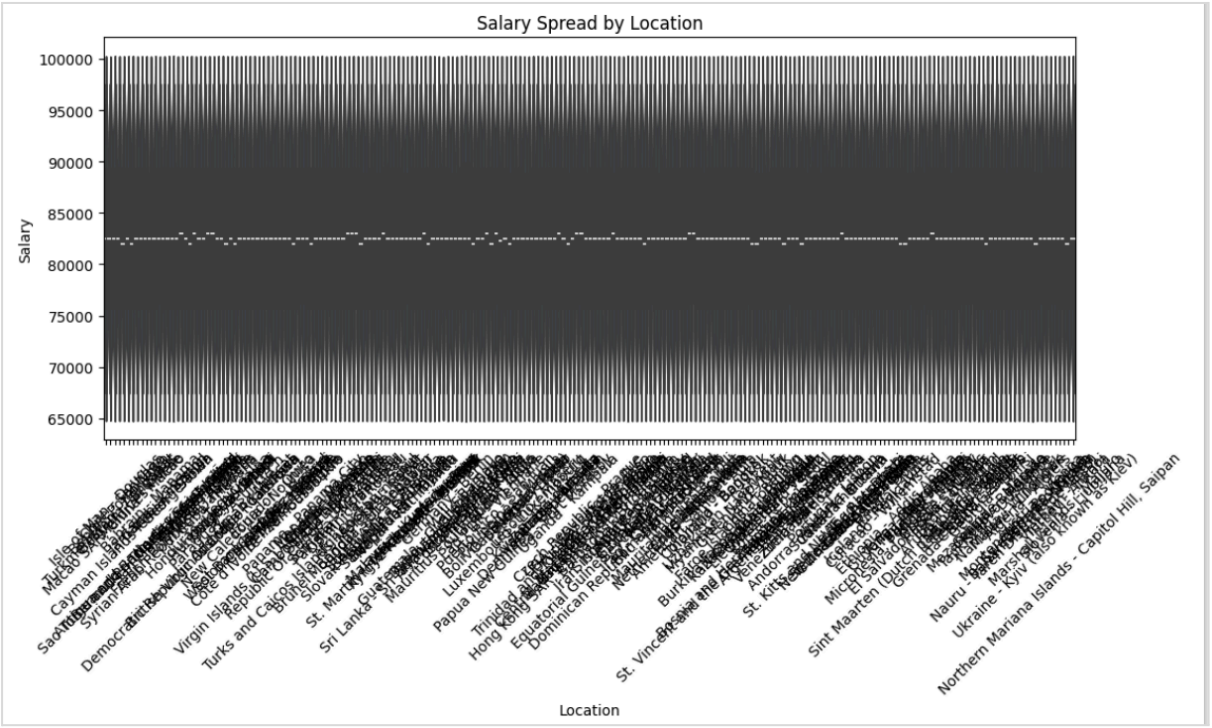
Career Development - Job Salary Prediction



This violin plot visualizes the distribution of salaries across different categorical features (e.g., job roles, locations, work types). It combines a box plot and a density plot, showing the spread, central tendency, and skewness of salaries within each category. Wider sections indicate higher data density, helping identify salary trends and variations across different groups.




Salary Spread by Work Type
Salary
100000
95000
90000
85000
80000
75000
70000
65000
Intern
Temporary
Full-Time
Contract
Part-Time
Work Type


Salary Spread by Location
Salary
100000
95000
90000
85000
80000
75000
70000
65000
Location

Career Development - Job Salary Prediction