```
# import libraries
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
FILE_PATH = r"/content/ai_job_dataset.csv"
```

```
# 1. Load the AI job dataset
df = pd.read_csv(FILE_PATH)
```

```
print("Loaded:", FILE_PATH)
```

```
Loaded: /content/ai_job_dataset.csv
Shape: (15000, 19)
    job_id              job_title  salary_usd salary_currency  \
0  AI00001  AI Research Scientist       90376             USD
1  AI00002   AI Software Engineer       61895             USD
2  AI00003          AI Specialist      152626             USD
3  AI00004           NLP Engineer       80215             USD
4  AI00005          AI Consultant       54624             EUR

  experience_level employment_type company_location company_size  \
0               SE              CT            China            M
1               EN              CT           Canada            M
2               MI              FL      Switzerland            L
3               SE              FL            India            M
4               EN              PT           France            S

  employee_residence  remote_ratio  \
0              China            50
1            Ireland           100
2        South Korea             0
3              India            50
4          Singapore           100

                                     required_skills education_required  \
0         Tableau, PyTorch, Kubernetes, Linux, NLP           Bachelor
1  Deep Learning, AWS, Mathematics, Python, Docker             Master
2     Kubernetes, Deep Learning, Java, Hadoop, NLP          Associate
3                           Scala, SQL, Linux, Python                PhD
4                        MLOps, Java, Tableau, Python             Master

   years_experience    industry posting_date application_deadline  \
0                 9  Automotive   2024-10-18           2024-11-07
1                 1       Media   2024-11-20           2025-01-11
2                 2   Education   2025-03-18           2025-04-07
3                 7  Consulting   2024-12-23           2025-02-24
4                 0       Media   2025-04-15           2025-06-23

   job_description_length  benefits_score       company_name
0                    1076             5.9    Smart Analytics
1                    1268             5.2       TechCorp Inc
2                    1974             9.4    Autonomous Tech
3                    1345             8.6     Future Systems
4                    1989             6.6  Advanced Robotics
```

```
print("Shape:", df.shape)
```

```
Shape: (15000, 19)
```

```
print(df.head(5))
```

```
    job_id              job_title  salary_usd salary_currency  \
0  AI00001  AI Research Scientist       90376             USD
1  AI00002   AI Software Engineer       61895             USD
2  AI00003          AI Specialist      152626             USD
3  AI00004           NLP Engineer       80215             USD
4  AI00005          AI Consultant       54624             EUR

  experience_level employment_type company_location company_size  \
0               SE              CT            China            M
1               EN              CT           Canada            M
2               MI              FL      Switzerland            L
3               SE              FL            India            M
```

```
     4               EN              PT           France            S

       employee_residence  remote_ratio  \
     0              China            50
     1            Ireland           100
     2        South Korea             0
     3              India            50
     4          Singapore           100

                                        required_skills education_required  \
     0         Tableau, PyTorch, Kubernetes, Linux, NLP           Bachelor
     1  Deep Learning, AWS, Mathematics, Python, Docker             Master
     2     Kubernetes, Deep Learning, Java, Hadoop, NLP          Associate
     3                        Scala, SQL, Linux, Python                PhD
     4                     MLOps, Java, Tableau, Python             Master

        years_experience    industry posting_date application_deadline  \
     0                 9  Automotive   2024-10-18           2024-11-07
     1                 1       Media   2024-11-20           2025-01-11
     2                 2   Education   2025-03-18           2025-04-07
     3                 7  Consulting   2024-12-23           2025-02-24
     4                 0       Media   2025-04-15           2025-06-23

        job_description_length  benefits_score       company_name
     0                    1076             5.9    Smart Analytics
     1                    1268             5.2       TechCorp Inc
     2                    1974             9.4    Autonomous Tech
     3                    1345             8.6     Future Systems
     4                    1989             6.6  Advanced Robotics
```

```
print(df.tail(5))
```

```
             job_id                    job_title  salary_usd salary_currency  \
     14995  AI14996            Robotics Engineer       38604             USD
     14996  AI14997  Machine Learning Researcher       57811             GBP
     14997  AI14998                 NLP Engineer      189490             USD
     14998  AI14999                   Head of AI       79461             EUR
     14999  AI15000     Computer Vision Engineer       56481             USD

           experience_level employment_type company_location company_size  \
     14995               EN              FL          Finland            S
     14996               EN              CT   United Kingdom            M
     14997               EX              CT      South Korea            L
     14998               EN              FT      Netherlands            M
     14999               MI              PT          Austria            S

           employee_residence  remote_ratio  \
     14995            Finland            50
     14996     United Kingdom             0
     14997        South Korea            50
     14998        Netherlands             0
     14999            Austria            50

                                          required_skills education_required  \
     14995                        Java, Kubernetes, Azure           Bachelor
     14996        Mathematics, Docker, SQL, Deep Learning             Master
     14997                              Scala, Spark, NLP          Associate
     14998      Java, Computer Vision, Python, TensorFlow                PhD
     14999  Scala, Azure, Deep Learning, GCP, Mathematics                PhD

            years_experience       industry posting_date application_deadline  \
     14995                 1         Energy   2025-02-06           2025-03-25
     14996                 0     Government   2024-10-16           2024-10-30
     14997                17  Manufacturing   2024-03-19           2024-05-02
     14998                 1    Real Estate   2024-03-22           2024-04-23
     14999                 2     Technology   2024-07-18           2024-08-10

            job_description_length  benefits_score       company_name
     14995                    1635             7.9  Advanced Robotics
     14996                    1624             8.2    Smart Analytics
     14997                    1336             7.4     AI Innovations
     14998                    1935             5.6    Smart Analytics
     14999                    2492             7.6     AI Innovations
```

```
# 2. Drop any column with more than 50% missing values
threshold = len(df) * 0.5
df = df.loc[:, df.isnull().sum() < threshold]
```

```
# 3. Fill missing values
null_counts = df.isnull().sum()
```

```
print("Null values per column:")
print(null_counts[null_counts > 0])
```

```
Null values per column:
Series([], dtype: int64)
```

```
#Find total number of rows that are completely duplicated
dup_count = df.duplicated().sum()
print(f"\nTotal duplicate rows: {dup_count}")
```

```

Total duplicate rows: 0
```

```
# Save the cleaned data (optional)
df.to_csv('ai_job_dataset_cleaned.csv', index=False)
print("Cleaned data saved to ai_job_dataset_cleaned.csv\n")
```

```
Cleaned data saved to ai_job_dataset_cleaned.csv
```

```
# Visualizations (matplotlib only)
import matplotlib.pyplot as plt

# Using a valid matplotlib style, for example 'ggplot'
plt.style.use('ggplot')
```

```
# Group by job title and compute average salary
avg_salary_by_job = df.groupby('job_title')['salary_usd'].mean().reset_index()
print(avg_salary_by_job.head())
```

```
               job_title     salary_usd
0           AI Architect  117436.513619
1          AI Consultant  113671.870739
2     AI Product Manager  114680.909825
3  AI Research Scientist  117897.925926
4   AI Software Engineer  114273.201531
```

```
# Sample job category table
job_category = pd.DataFrame({
    'job_title': df['job_title'].unique()[:5],  # simulate a lookup table
    'category': ['Data', 'ML', 'Analytics', 'Engineering', 'DevOps']
})

# Merge based on job_title
merged_df = pd.merge(df, job_category, on='job_title', how='left')
print(merged_df[['job_title', 'category']].head())
```

```
               job_title     category
0  AI Research Scientist         Data
1   AI Software Engineer           ML
2          AI Specialist    Analytics
3           NLP Engineer  Engineering
4          AI Consultant       DevOps
```

```
# Create a dummy stats DataFrame with index = experience_level
stats = df.groupby('experience_level')['salary_usd'].mean().to_frame('avg_exp_salary')

# Join using index
joined_df = df.set_index('experience_level').join(stats)
print(joined_df[['salary_usd', 'avg_exp_salary']].head())
```

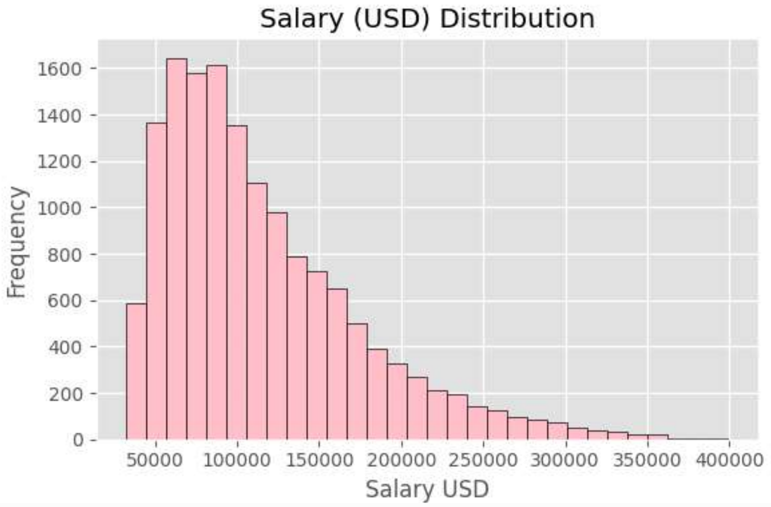
```
                  salary_usd  avg_exp_salary
experience_level
SE                     90376   122187.657845
EN                     61895    63133.377084
MI                    152626    87955.471833
SE                     80215   122187.657845
EN                     54624    63133.377084
```

```
# Take two pieces of df and stack them
top_rows = df.head(3)
bottom_rows = df.tail(3)
vertical_concat = pd.concat([top_rows, bottom_rows], axis=0)
```

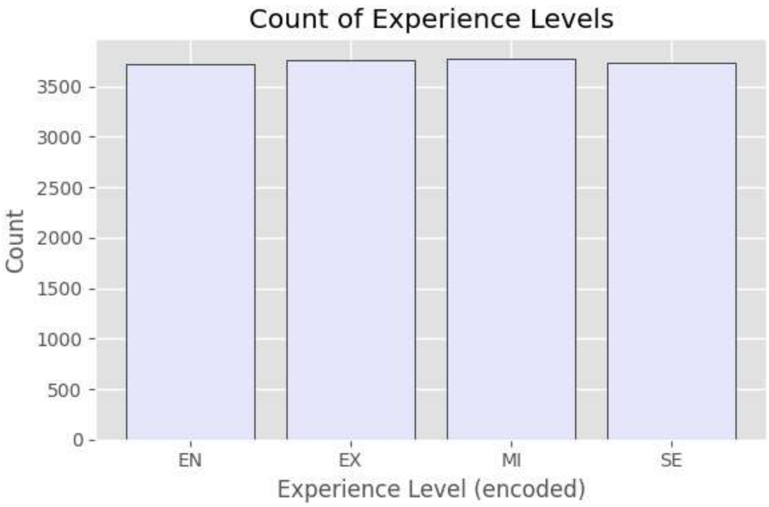
```
print(vertical_concat)
```

```
        job_id                 job_title  salary_usd salary_currency  \
0      AI00001     AI Research Scientist       90376             USD
1      AI00002      AI Software Engineer       61895             USD
2      AI00003             AI Specialist      152626             USD
14997  AI14998              NLP Engineer      189490             USD
14998  AI14999                Head of AI       79461             EUR
14999  AI15000  Computer Vision Engineer       56481             USD

      experience_level employment_type company_location company_size  \
0                   SE              CT            China            M
1                   EN              CT           Canada            M
2                   MI              FL      Switzerland            L
14997               EX              CT      South Korea            L
14998               EN              FT      Netherlands            M
14999               MI              PT          Austria            S

      employee_residence  remote_ratio  \
0                  China            50
1                Ireland           100
2            South Korea             0
14997        South Korea            50
14998        Netherlands             0
14999            Austria            50

                                       required_skills education_required  \
0             Tableau, PyTorch, Kubernetes, Linux, NLP           Bachelor
1      Deep Learning, AWS, Mathematics, Python, Docker             Master
2         Kubernetes, Deep Learning, Java, Hadoop, NLP          Associate
14997                                Scala, Spark, NLP          Associate
14998        Java, Computer Vision, Python, TensorFlow                PhD
14999    Scala, Azure, Deep Learning, GCP, Mathematics                PhD

       years_experience       industry posting_date application_deadline  \
0                     9     Automotive   2024-10-18           2024-11-07
1                     1          Media   2024-11-20           2025-01-11
2                     2      Education   2025-03-18           2025-04-07
14997                17  Manufacturing   2024-03-19           2024-05-02
14998                 1    Real Estate   2024-03-22           2024-04-23
14999                 2     Technology   2024-07-18           2024-08-10

       job_description_length  benefits_score     company_name
0                        1076             5.9  Smart Analytics
1                        1268             5.2      TechCorp Inc
2                        1974             9.4  Autonomous Tech
14997                    1336             7.4    AI Innovations
14998                    1935             5.6  Smart Analytics
14999                    2492             7.6    AI Innovations
```

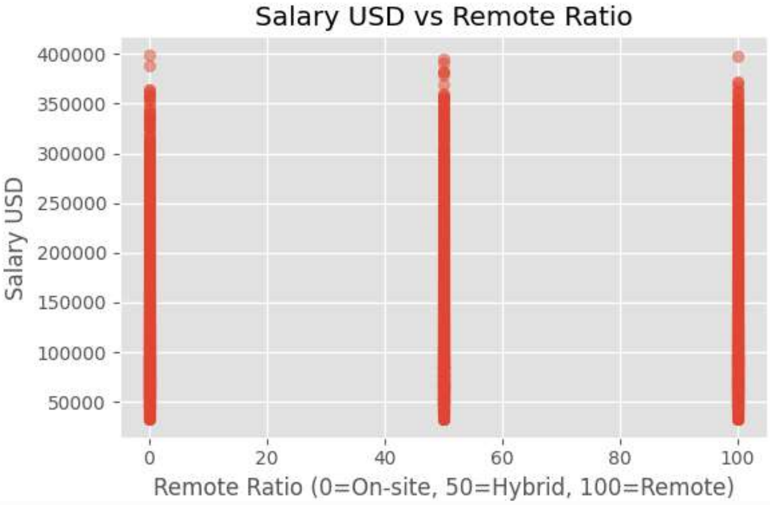
```
# Plot 1: Histogram of salary_usd
plt.figure(figsize=(6,4))
plt.hist(df['salary_usd'], bins=30, color='pink', edgecolor='black')
plt.title('Salary (USD) Distribution')
plt.xlabel('Salary USD')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



```
# Plot 2: Bar chart of experience_level counts
exp_counts = df['experience_level'].value_counts().sort_index()
plt.figure(figsize=(6,4))
plt.bar(exp_counts.index.astype(str), exp_counts.values,color = "lavender", edgecolor='black')
plt.title('Count of Experience Levels')
plt.xlabel('Experience Level (encoded)')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```

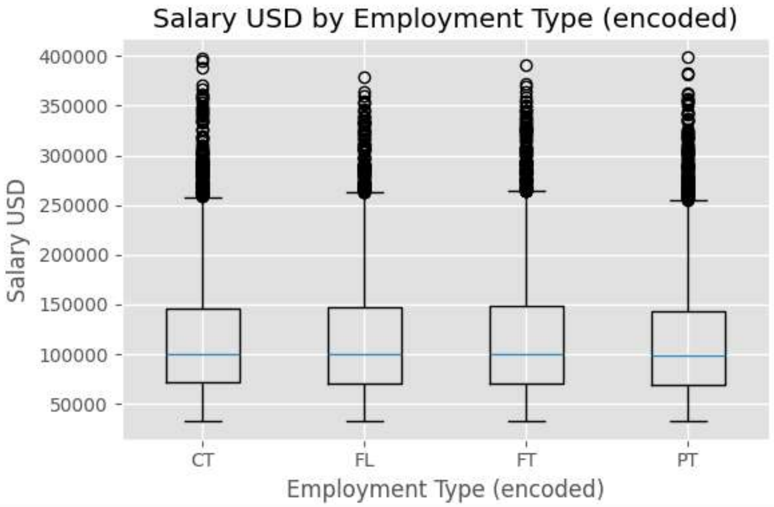


```
# Plot 3: Scatter plot salary_usd vs remote_ratio
plt.figure(figsize=(6,4))
plt.scatter(df['remote_ratio'], df['salary_usd'], alpha=0.5)
plt.title('Salary USD vs Remote Ratio')
plt.xlabel('Remote Ratio (0=On-site, 50=Hybrid, 100=Remote)')
plt.ylabel('Salary USD')
plt.tight_layout()
plt.show()
```

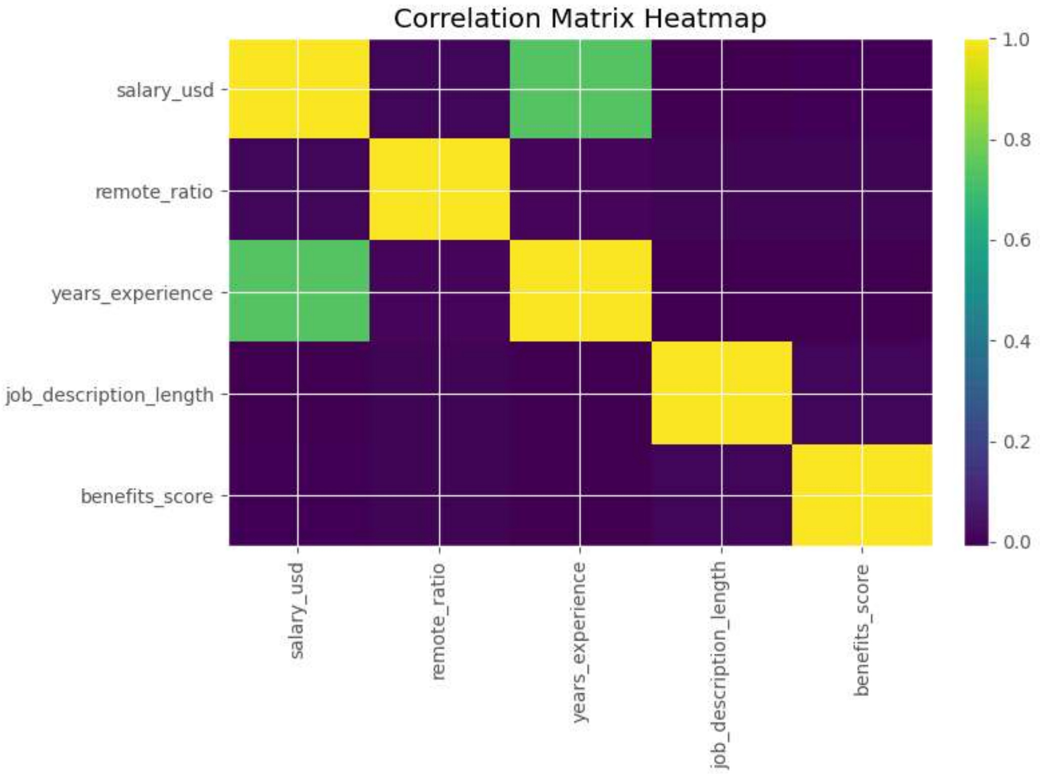


```
# Plot 4: Boxplot of salary_usd by employment_type
empl_codes = sorted(df['employment_type'].unique())
data_for_box = [df[df['employment_type']==code]['salary_usd'] for code in empl_codes]
plt.figure(figsize=(6,4))
plt.boxplot(data_for_box, labels=[str(code) for code in empl_codes])
plt.title('Salary USD by Employment Type (encoded)')
plt.xlabel('Employment Type (encoded)')
plt.ylabel('Salary USD')
plt.tight_layout()
plt.show()
```

```
<ipython-input-48-505777296>:5: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick_labels' since M
  plt.boxplot(data_for_box, labels=[str(code) for code in empl_codes])
```



```
# 7. Correlation Heatmap: numeric features
plt.figure(figsize=(8,6))
num_cols = df.select_dtypes(include=['number']).columns.tolist()
corr = df[num_cols].corr()
im = plt.imshow(corr, cmap='viridis', aspect='auto')
plt.colorbar(im, fraction=0.046, pad=0.04)
plt.xticks(range(len(num_cols)), num_cols, rotation=90)
plt.yticks(range(len(num_cols)), num_cols)
plt.title('Correlation Matrix Heatmap')
plt.tight_layout()
plt.show()
```


Correlation Matrix Heatmap
salary_usd
remote_ratio
years_experience
job_description_length
benefits_score
salary_usd
remote_ratio
years_experience
job_description_length
benefits_score