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

# Load the dataset
updated_file_path = "updated_dataset_salary_cyber.csv"
df = pd.read_csv(updated_file_path)

# Display column names to verify structure
print(" Column Names in Dataset:")
print(df.columns)

# Display first few rows to understand the structure
print("\n Dataset Preview:")
print(df.head())
```

```
 Column Names in Dataset:
Index(['job_id', 'work_year', 'experience_level', 'employment_type',
       'job_title', 'salary', 'salary_currency', 'salary_in_usd',
       'employee_residence', 'remote_ratio', 'company_location',
       'company_size', 'required_skills', 'certifications',
'posting_date',
       'current_job_openings', 'required_qualifications',
'industry_demand',
       'hiring_trends', 'candidate_profiles',
'recruitment_timelines'],
      dtype='object')

 Dataset Preview:
   job_id  work_year experience_level employment_type
job_title  \
0       1       2022               EN              FT   Cyber Program
Manager
1       2       2022               MI              FT        Security
Analyst
2       3       2022               MI              FT        Security
Analyst
3       4       2022               MI              FT     IT Security
Analyst
4       5       2022               EN              CT  Cyber Security
Analyst

   salary salary_currency  salary_in_usd employee_residence
remote_ratio  \
0   63000             USD          63000                 US
50
1   95000             USD          95000                 US
0
2   70000             USD          70000                 US
0
3  250000             BRL          48853                 BR
```

```
50
4  120000              USD         120000                  BW
100

   ... company_size   required_skills certifications posting_date  \
0  ...            S  Network Security            CEH   2022-01-01
1  ...            M               NaN            CEH   2022-01-02
2  ...            M       Azure, SOC            NaN   2022-01-03
3  ...            L               NaN          CISSP   2022-01-04
4  ...            S               NaN            CEH   2022-01-05

  current_job_openings  required_qualifications industry_demand
hiring_trends  \
0                   98         Experience-Based          Medium
Declining
1                   98          Master's Degree          Medium
Growing
2                   80       Certification Only            High
Declining
3                   14        Bachelor's Degree             Low
Stable
4                   48       Certification Only            High
Stable

  candidate_profiles recruitment_timelines
0        Experienced                 Fast
1              Mixed                 Slow
2        Entry-Level                 Slow
3        Experienced             Moderate
4        Experienced                 Slow

[5 rows x 21 columns]
```

CHECK FOR MISSING VALUES

```
missing_values = df.isnull().sum()
missing_values_table = pd.DataFrame(missing_values, columns=['Missing
Count'])

print("\n Missing Values Summary:")
print(missing_values_table)


 Missing Values Summary:
                        Missing Count
job_id                              0
work_year                           0
experience_level                    0
employment_type                     0
job_title                           0
```

```
salary                            0
salary_currency                   0
salary_in_usd                     0
employee_residence                0
remote_ratio                      0
company_location                  0
company_size                      0
required_skills                 538
certifications                  647
posting_date                      0
current_job_openings              0
required_qualifications           0
industry_demand                   0
hiring_trends                     0
candidate_profiles                0
recruitment_timelines             0
```

CHECK FOR INVALID SALARY VALUES

```
if "salary_in_usd" in df.columns:
    invalid_salaries = df[(df['salary_in_usd'] < 0) |
(df['salary_in_usd'] > 500000)]
    print("\n Checking for Invalid Salary Values:")
    print(invalid_salaries)
else:
    print("\n Skipping salary check: 'salary_in_usd' column not
found.")
```

```
 Checking for Invalid Salary Values:
      job_id  work_year experience_level employment_type  \
512      513       2022               SE              FT
647      648       2022               SE              FT
886      887       2020               EX              FT
1861    1862       2022               SE              FT
1996    1997       2022               SE              FT
2235    2236       2020               EX              FT

                          job_title  salary salary_currency
salary_in_usd  \
512              Detection Engineer  710000             GBP
910991
647      Incident Response Manager  600000             USD
600000
886   Security Engineering Manager  600000             USD
600000
1861             Detection Engineer  710000             GBP
899920
1996     Incident Response Manager  600000             USD
```

```
600000
2235  Security Engineering Manager  600000             USD
600000

     employee_residence  remote_ratio  ... company_size
required_skills  \
512                  GB           100  ...            L    AWS,
Kubernetes
647                  US           100  ...            L
NaN
886                  US            50  ...            L  Python,
Security+
1861                 GB           100  ...            L   Network
Security
1996                 US           100  ...            L
NaN
2235                 US            50  ...            L  Python,
Security+

     certifications posting_date current_job_openings  \
512    AWS Security   2023-05-28                   10
647             CEH   2023-10-10                   65
886             CEH   2024-06-05                   85
1861            CEH   2027-02-05                   87
1996          CISSP   2027-06-20                    6
2235          CISSP   2028-02-14                   33

      required_qualifications industry_demand hiring_trends  \
512           Master's Degree          Medium       Growing
647          Experience-Based            High        Stable
886          Experience-Based          Medium       Growing
1861       Certification Only            High       Growing
1996          Master's Degree             Low     Declining
2235         Experience-Based             Low       Growing

     candidate_profiles recruitment_timelines
512         Experienced                  Fast
647               Mixed              Moderate
886         Entry-Level                  Fast
1861        Experienced                  Slow
1996        Experienced              Moderate
2235        Entry-Level                  Slow

[6 rows x 21 columns]
```

CHECK FOR INVALID EXPERIENCE LEVELS

```
valid_experience_levels = ["Entry-level", "Mid-level", "Senior"]
if "experience_level" in df.columns:
```

```
    invalid_experience =
df[~df["experience_level"].isin(valid_experience_levels)]
    print("\n Checking for Invalid Experience Levels:")
    print(invalid_experience)
else:
    print("\n Skipping experience level check: 'experience_level'
column not found.")


 Checking for Invalid Experience Levels:
      job_id  work_year experience_level employment_type  \
0          1       2022               EN              FT
1          2       2022               MI              FT
2          3       2022               MI              FT
3          4       2022               MI              FT
4          5       2022               EN              CT
...      ...        ...              ...             ...
2591    2592       2020               MI              FT
2592    2593       2021               SE              FT
2593    2594       2021               SE              FT
2594    2595       2021               MI              FT
2595    2596       2021               MI              FT

                         job_title  salary salary_currency
salary_in_usd  \
0            Cyber Program Manager   63000             USD
63000
1                Security Analyst   95000             USD
95000
2                Security Analyst   70000             USD
70000
3              IT Security Analyst  250000             BRL
48853
4           Cyber Security Analyst  120000             USD
120000
...                            ...     ...             ...
...
2591        Cyber Security Analyst  140000             AUD
96422
2592  Information Security Manager   60000             GBP
82528
2593  Penetration Testing Engineer  126000             USD
126000
2594  Information Security Analyst   42000             GBP
57769
2595   Threat Intelligence Analyst   66310             USD
66310

     employee_residence  remote_ratio  ... company_size
required_skills  \
```

```
0                    US            50  ...            S   Network
Security
1                    US             0  ...            M
NaN
2                    US             0  ...            M         Azure,
SOC
3                    BR            50  ...            L
NaN
4                    BW           100  ...            S
NaN
...                 ...           ...  ...          ...
...
2591                 AU            50  ...            M  Python,
Security+
2592                 GB            50  ...            L    AWS,
Kubernetes
2593                 US           100  ...            L  Python,
Security+
2594                 GB           100  ...            L
NaN
2595                 US             0  ...            L
NaN

     certifications posting_date current_job_openings  \
0               CEH   2022-01-01                   98
1               CEH   2022-01-02                   98
2               NaN   2022-01-03                   80
3             CISSP   2022-01-04                   14
4               CEH   2022-01-05                   48
...             ...          ...                  ...
2591          CISSP   2029-02-04                   29
2592            NaN   2029-02-05                   79
2593          CISSP   2029-02-06                   27
2594   AWS Security   2029-02-07                   61
2595            CEH   2029-02-08                   56

      required_qualifications industry_demand hiring_trends  \
0            Experience-Based          Medium     Declining
1             Master's Degree          Medium       Growing
2          Certification Only            High     Declining
3           Bachelor's Degree             Low        Stable
4          Certification Only            High        Stable
...                       ...             ...           ...
2591        Bachelor's Degree            High        Stable
2592       Certification Only             Low       Growing
2593        Bachelor's Degree          Medium        Stable
2594       Certification Only            High        Stable
2595          Master's Degree          Medium     Declining

     candidate_profiles recruitment_timelines
```

```
0           Experienced                  Fast
1                 Mixed                  Slow
2           Entry-Level                  Slow
3           Experienced              Moderate
4           Experienced                  Slow
...                 ...                   ...
2591              Mixed              Moderate
2592        Entry-Level                  Fast
2593              Mixed                  Slow
2594        Experienced                  Fast
2595              Mixed                  Fast

[2596 rows x 21 columns]
```

CHECK FOR REFERENTIAL INTEGRITY (Job ID as Primary Key)

```
print("\n Checking for Unique Job ID (Primary Key):")
if df["job_id"].is_unique:
    print("Job ID column is unique and can serve as a primary key.")
else:
    print("Duplicate Job IDs found!")


 Checking for Unique Job ID (Primary Key):
Job ID column is unique and can serve as a primary key.
```

CHECK FOR DUPLICATE RECORDS

```
duplicate_rows = df[df.duplicated()]
print("\n Checking for Duplicate Entries:")
print(duplicate_rows)


 Checking for Duplicate Entries:
Empty DataFrame
Columns: [job_id, work_year, experience_level, employment_type,
job_title, salary, salary_currency, salary_in_usd, employee_residence,
remote_ratio, company_location, company_size, required_skills,
certifications, posting_date, current_job_openings,
required_qualifications, industry_demand, hiring_trends,
candidate_profiles, recruitment_timelines]
Index: []

[0 rows x 21 columns]
```

CHECK FOR OUTLIERS IN REMOTE WORK RATIO

```
if "remote_ratio" in df.columns:
    invalid_remote_ratio = df[(df['remote_ratio'] < 0) |
```

```
(df['remote_ratio'] > 100)]
    print("\n Checking for Remote Work Ratio Outliers:")
    print(invalid_remote_ratio)
else:
    print("\n Skipping remote ratio check: 'remote_ratio' column not
found.")
```

```
 Checking for Remote Work Ratio Outliers:
Empty DataFrame
Columns: [job_id, work_year, experience_level, employment_type,
job_title, salary, salary_currency, salary_in_usd, employee_residence,
remote_ratio, company_location, company_size, required_skills,
certifications, posting_date, current_job_openings,
required_qualifications, industry_demand, hiring_trends,
candidate_profiles, recruitment_timelines]
Index: []

[0 rows x 21 columns]
```

CHECK FOR INDUSTRY DEMAND & HIRING TREND CONSISTENCY

```
if "industry_demand" in df.columns and "hiring_trends" in df.columns:
    inconsistent_demand = df[(df["industry_demand"] == "Low") &
(df["hiring_trends"] == "Growing")]
    print("\n Checking for Industry Demand and Hiring Trend
Inconsistencies:")
    print(inconsistent_demand)
else:
    print("\n Skipping industry demand consistency check: Required
columns not found.")
```

```
 Checking for Industry Demand and Hiring Trend Inconsistencies:
      job_id  work_year experience_level employment_type  \
11        12       2022               SE              FT
15        16       2021               EX              FT
28        29       2022               MI              FT
35        36       2021               EN              FT
54        55       2022               EX              FT
...      ...        ...              ...             ...
2565    2566       2020               SE              FT
2572    2573       2020               SE              FT
2577    2578       2021               MI              FT
2580    2581       2020               MI              FT
2592    2593       2021               SE              FT

                            job_title  salary salary_currency
salary_in_usd  \
11     Application Security Specialist   85000             USD
```

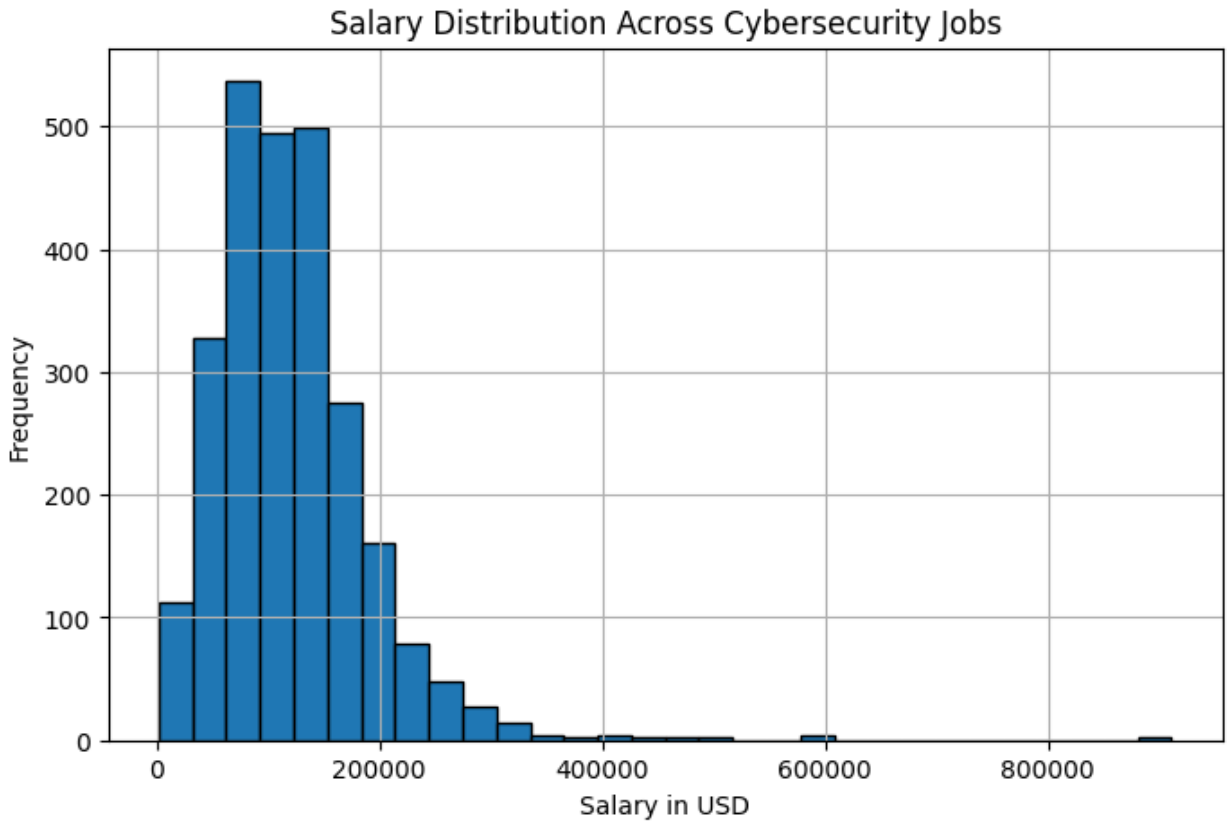
```
85000
15        Head of Information Security  232000             USD
232000
28                Vulnerability Analyst  115000             USD
115000
35                          SOC Analyst   75000             USD
75000
54    Director of Information Security  148000             USD
148000
...                                  ...     ...             ...
...
2565    Information Security Architect  198000             USD
198000
2572                  Security Engineer  149000             USD
149000
2577     Information Security Engineer  128000             USD
128000
2580      Information Security Officer  122750             USD
122750
2592      Information Security Manager   60000             GBP
82528

     employee_residence  remote_ratio  ... company_size
required_skills  \
11                   US           100  ...            L         Azure,
SOC
15                   US           100  ...            L  Python,
Security+
28                   US           100  ...            L   Network
Security
35                   US             0  ...            M
NaN
54                   US             0  ...            L         Azure,
SOC
...                 ...           ...  ...          ...
...
2565                 US           100  ...            L         Azure,
SOC
2572                 US            50  ...            L    AWS,
Kubernetes
2577                 US           100  ...            L  Python,
Security+
2580                 US           100  ...            M    AWS,
Kubernetes
2592                 GB            50  ...            L    AWS,
Kubernetes

     certifications posting_date current_job_openings  \
11     AWS Security   2022-01-12                   54
```

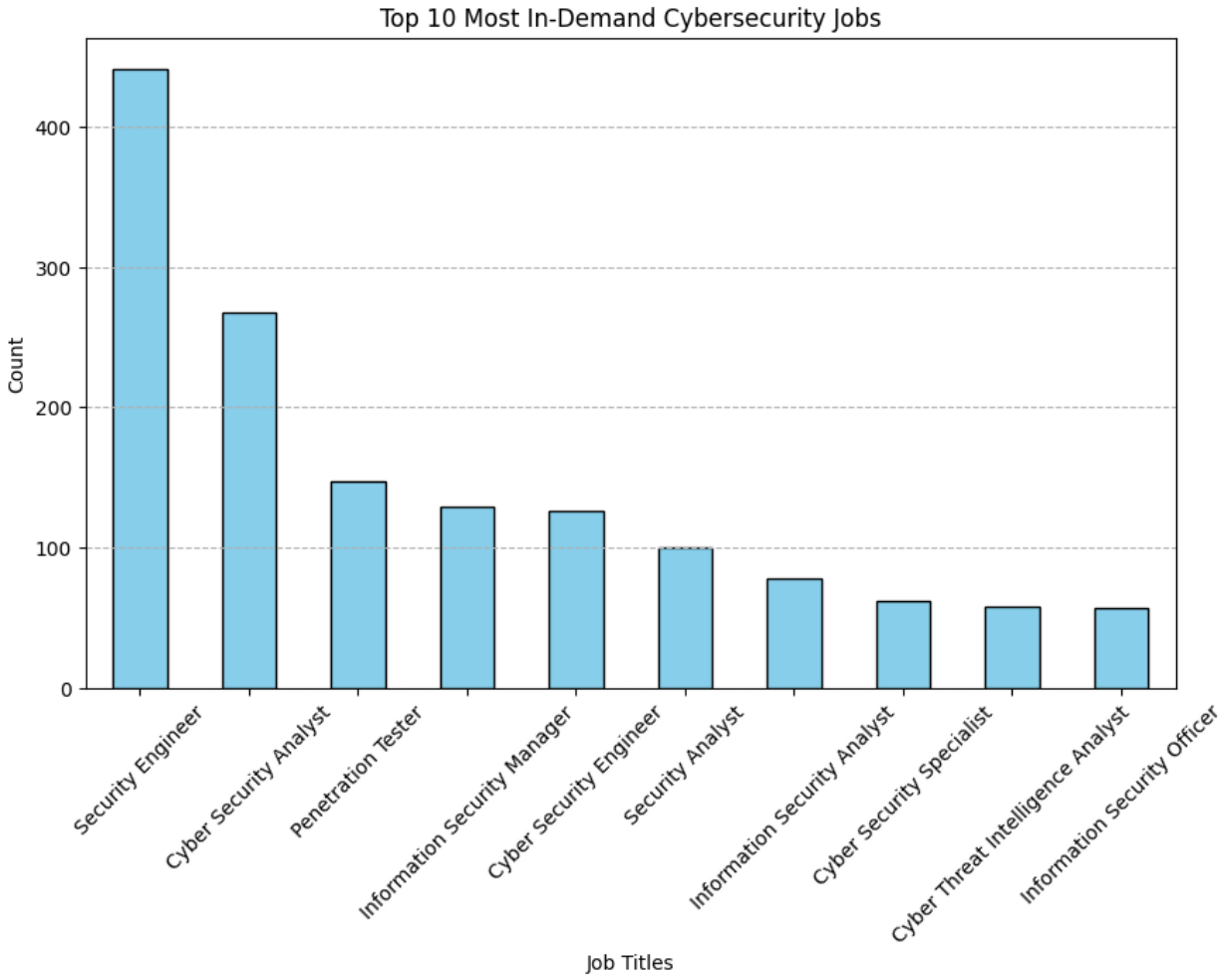
```
15     AWS Security   2022-01-16                  22
28     AWS Security   2022-01-29                  25
35              NaN   2022-02-05                  83
54     AWS Security   2022-02-24                  83
...             ...          ...                 ...
2565            CEH   2029-01-09                  42
2572            NaN   2029-01-16                  89
2577          CISSP   2029-01-21                  63
2580            CEH   2029-01-24                  96
2592            NaN   2029-02-05                  79

      required_qualifications industry_demand hiring_trends  \
11             Master's Degree             Low       Growing
15             Master's Degree             Low       Growing
28             Master's Degree             Low       Growing
35         Certification Only              Low       Growing
54         Certification Only              Low       Growing
...                       ...              ...           ...
2565       Certification Only              Low       Growing
2572        Bachelor's Degree              Low       Growing
2577        Bachelor's Degree              Low       Growing
2580       Certification Only              Low       Growing
2592       Certification Only              Low       Growing

     candidate_profiles recruitment_timelines
11                Mixed                  Slow
15          Experienced                  Slow
28                Mixed                  Slow
35          Experienced                  Slow
54          Entry-Level              Moderate
...                 ...                   ...
2565        Entry-Level                  Slow
2572        Entry-Level              Moderate
2577        Experienced              Moderate
2580        Experienced                  Fast
2592        Entry-Level                  Fast

[284 rows x 21 columns]
```

VISUALIZATION: SALARY DISTRIBUTION AND JOB DEMAND BY TITLE

```
plt.figure(figsize=(8, 5))
df['salary_in_usd'].hist(bins=30, edgecolor='black')
plt.xlabel("Salary in USD")
plt.ylabel("Frequency")
plt.title("Salary Distribution Across Cybersecurity Jobs")
plt.grid(True)
plt.show()
```

```
plt.figure(figsize=(10, 6))
top_jobs = df["job_title"].value_counts().nlargest(10)
top_jobs.plot(kind="bar", color="skyblue", edgecolor="black")
plt.xlabel("Job Titles")
plt.ylabel("Count")
plt.title("Top 10 Most In-Demand Cybersecurity Jobs")
plt.xticks(rotation=45)
plt.grid(axis="y", linestyle="--")
plt.show()
```





FINAL SUMMARY REPORT

```
print("\nFinal Validation Summary: \n")
print(f"Total Missing Values: {missing_values.sum()}")
print(f"Total Invalid Salaries: {invalid_salaries.shape[0] if
'salary_in_usd' in df.columns else 'Skipped'}")
print(f"Total Invalid Experience Levels: {invalid_experience.shape[0]
if 'experience_level' in df.columns else 'Skipped'}")
print(f"Total Duplicate Records: {duplicate_rows.shape[0]}")
print(f"Total Remote Ratio Outliers: {invalid_remote_ratio.shape[0] if
'remote_ratio' in df.columns else 'Skipped'}")
print(f"Total Industry Demand & Hiring Trend Issues:
{inconsistent_demand.shape[0] if 'industry_demand' in df.columns and
'hiring_trends' in df.columns else 'Skipped'}")
```

```
Final Validation Summary:

Total Missing Values: 1185
Total Invalid Salaries: 6
```

```
Total Invalid Experience Levels: 2596
Total Duplicate Records: 0
Total Remote Ratio Outliers: 0
Total Industry Demand & Hiring Trend Issues: 284
```