Week 6 **Verifying data integrity**

Group - 17

Rama Sai Durga Machara Manikumar Gunnam Pavan Venkata  
Sai Krishna Teja Kagitha  
Sujan Budha  
Suneel Kumar Buri  
Vikas Reddy Gurrala

School for Professional Studies, Saint Louis University  
IS-5960-03: Masters Research Project  
Maria Weber  
February 20, 2025

**1. Revised Problem Statement**

The primary challenge in cybersecurity job recruitment is aligning industry demand with available talent. HR professionals and recruitment specialists struggle with understanding salary benchmarks, skill demand, and workforce gaps. The **Employability Analytics Dashboard** provides real-time insights into job market trends, salaries, and skill requirements, helping HR professionals make informed hiring decisions.

**2. Mapping of Action Components to Data Fields**

|  |  |
| --- | --- |
| **Action Component** | **Relevant Data Fields** |
| Identify high-demand skills | job\_title, required\_skills, certifications |
| Benchmark salaries for roles | job\_title, salary\_in\_usd, company\_location, experience\_level |
| Analyze job market trends | job\_title, employment\_type, work\_year, company\_location |
| Recognize workforce gaps | current\_job\_openings, required\_qualifications, industry\_demand |
| Improve recruitment strategies | hiring\_trends, candidate\_profiles, recruitment\_timelines |

**3. Data Validation Checks**

**3.1 Referential Integrity Validation**

* **Primary Key Integrity:** job\_id is unique and serves as the primary key. No duplicate job entries found.
* **Referential Integrity:** Ensured all foreign keys reference valid parent records.

**3.2 Field-Level Integrity Checks**

* **Salary Validation:** Flagged **6 outliers** with salaries above $500,000.
* **Experience Level Validation:** Found **2596 incorrect values** (EN, MI, SE → Needs standardization to Entry-level, Mid-level, Senior).
* **Remote Work Ratio Validation:** No outliers; all values between 0-100.
* **Industry Demand vs. Hiring Trend Conflicts:** **284 inconsistencies** where demand is low but hiring is growing.

**3.3 Missing Values Check**

No missing values in critical fields. Previously missing required\_skills and certifications were addressed.

**4. Code Implementation for Data Integrity Checks**

The validation was performed using **Python**, utilizing **pandas** and **matplotlib** for analysis and visualization.

import pandas as pd

import matplotlib.pyplot as plt

# Load dataset

df = pd.read\_csv("updated\_dataset\_salary\_cyber.csv")

# Check for missing values

missing\_values = df.isnull().sum()

print(missing\_values)

# Salary validation

invalid\_salaries = df[(df['salary\_in\_usd'] < 0) | (df['salary\_in\_usd'] > 500000)]

print(invalid\_salaries)

# Experience level validation

valid\_experience\_levels = ["Entry-level", "Mid-level", "Senior"]

invalid\_experience = df[~df["experience\_level"].isin(valid\_experience\_levels)]

print(invalid\_experience)

# Check primary key integrity

if df["job\_id"].is\_unique:

print("Job ID column is unique and serves as a primary key.")

# Salary distribution visualization

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()

**5. Manual Data Adjustments**

* **Mapping Incorrect Experience Levels** (EN, MI, SE → Entry-level, Mid-level, Senior)
* **Handling Extreme Salaries** (Limiting salaries to a reasonable max threshold)
* **Industry Demand Conflicts** (Manual review of positions where demand is marked low but hiring is growing)

**6. AI and External Resources Used**

* Used **ChatGPT** for Python code generation and validation strategy.
* Dataset sourced from Kaggle for external reference and analysis.

**6.1 Prompts Used with Generative AI**

1. Generate Python code to validate data integrity, including primary key uniqueness, salary outliers, and missing values.
2. How to detect inconsistencies in industry demand vs. hiring trends using Python?
3. Best practices for cleaning invalid experience levels in a dataset?
4. Create visualizations for salary distribution and job demand trends in cybersecurity data.

**7. References**

* Revaldo, D. (2023). Salary Cyber Security Jobs Dataset. Kaggle. Retrieved from <https://www.kaggle.com/datasets/dannyrevaldo/salary-cyber-security-jobs>
* **Pandas Library Documentation**: <https://pandas.pydata.org/docs/>
* **Matplotlib Visualization Guide**: <https://matplotlib.org/stable/contents.html>