Data Validation and Dashboard Logic Documentation

Python Data Validation Script

We used the following Python code to validate our dataset. It checks for missing values, invalid entries in salary and remote ratio fields, and ensures consistent categories for experience level and employment type.

```
import pandas as pd

df = pd.read_csv("Enhanced_Cybersecurity_Job_Dataset.csv")

missing_values = df.isnull().sum()
duplicate_rows = df.duplicated().sum()
invalid_salaries = df[df["salary_in_usd"] <= 0]
invalid_remote_ratio = df[(df["remote_ratio"] < 0) |
(df["remote_ratio"] > 100)]

valid_experience_levels = ["EN", "MI", "SE", "EX"]
invalid_experience_levels =
df[~df["experience_level"].isin(valid_experience_levels)]

valid_employment_types = ["FT", "PT", "CT", "FL"]
invalid_employment_types =
df[~df["employment_type"].isin(valid_employment_types)]

invalid_salary_currency = df[~df["salary_currency"].str.match(r"^[A-
Z]{3}$", na=False)]
invalid_company_location = df[~df["company_location"].str.match(r"^[A-
Z]{2}$", na=False)]
invalid_employee_residence =
df[~df["employee_residence"].str.match(r"^[A-Z]{2}$", na=False)]
```

DAX Queries Used in Power BI

Below are the DAX formulas used to calculate insights in the dashboard, such as percentages of remote jobs, clearance-based salaries, and top-paying roles.

% Fully Remote Jobs

```
DIVIDE(
    COUNTROWS(FILTER('Dataset', 'Dataset'[Remote Work Availability] =
"Yes")),
    COUNTROWS('Dataset')
) * 100
```

% Jobs Requiring Clearance

```
DIVIDE(
    COUNTROWS(FILTER('Dataset', 'Dataset'[Security Clearance Required]
```

```
<> "No")),
    COUNTROWS('Dataset')
) * 100
```

Average Salary (Remote Jobs)

```
AVERAGEX(
    FILTER('Dataset', 'Dataset'[Remote Work Availability] = "Yes"),
    'Dataset'[salary_in_usd]
)
```

Top Paying Roles

```
TOPN(
    5,
    SUMMARIZE('Dataset', 'Dataset'[job_title], "AvgSalary",
AVERAGE('Dataset'[salary_in_usd])),
    [AvgSalary], DESC
)
```

YOY Salary Growth

```
VAR PrevYear = CALCULATE(AVERAGE('Dataset'[salary_in_usd]),
PREVIOUSYEAR('Dataset'[work_year]))
RETURN DIVIDE([Average Salary] - PrevYear, PrevYear)
```

Excel Formulas Used for Validation

If Excel was used instead of Python, the following formulas helped ensure the quality of the data before importing it to Power BI.

Check if value is missing

```
=IF(ISBLANK(A2), "Missing", "Valid")
```

Validate positive salary

```
=IF(B2>0, "Valid", "Invalid")
```

Remote ratio between 0 and 100

```
=IF(AND(C2>=0, C2<=100), "Valid", "Invalid")
```

Valid experience levels

```
=IF(OR(D2="EN", D2="MI", D2="SE", D2="EX"), "Valid", "Invalid")
```

Valid employment types

```
=IF(OR(E2="FT", E2="PT", E2="CT", E2="FL"), "Valid", "Invalid")
```