# Data Analytics Project

## Global Freelancers Dataset

From Raw Data Cleaning to Power BI Visualization

**Prepared by**
Mohamed Abbad Andaloussi

November 2025

# Contents

# 1 Introduction

The objective of this project is to perform an end-to-end Data Analytics workflow on a dataset of global freelancers. Starting from a raw CSV file containing inconsistent formatting, the data was cleaned, transformed, analyzed in Python, and then visualized using Power BI to generate business insights.

This report presents each phase of the workflow:

- Data import and inspection.
- Data cleaning and preprocessing.
- Exploratory Data Analysis (EDA).
- Power BI dashboard creation.

# 2 Dataset Description

## 2.1 Raw Dataset

The original file `global_freelancers_raw.csv` contained:

- 1000 freelancers.
- 12 columns stored incorrectly due to delimiter issues.
- Various inconsistencies: mixed delimiters, missing values, text noise, inconsistent gender notation, currency symbols, percentage signs, etc.

## 2.2 Cleaned Dataset

After cleaning, the final dataset `global_freelancers_clean.csv` contains consistent and analysis-ready fields.

Table 1: Data Dictionary (Cleaned Dataset)

| Column | Type | Description |
|---|---|---|
| freelancer_id | String | Unique ID. |
| name | String | Cleaned full name. |
| gender | Category | Male / Female / Other. |
| age | Numeric | Age (20–60). |
| country | Category | Freelancer country. |
| language | Category | Native working language. |
| primary_skill | Category | Main skill area. |
| years_experience | Numeric | Experience in years. |
| hourly_rate_usd | Numeric | Cleaned hourly rate (USD). |
| rating | Numeric | Rating (0–5). |
| is_active | Boolean | Activity status. |
| client_satisfaction_pct | Numeric | Satisfaction percentage. |

# 3 Methodology

## 3.1 Tools

The project used:

- **Python (Pandas, NumPy)** for cleaning and EDA.
- **Jupyter Notebook** (`Global_freelance_DataSet.ipynb`).
- **Power BI** for data visualization and dashboarding.

## 3.2 Step 1 — Importing Libraries

Python libraries used:

- `pandas` for data manipulation.
- `numpy` for numerical operations.

## 3.3 Step 2 — Reading the Raw CSV

The raw file initially loaded incorrectly (all data in one column). Fixing required specifying the correct delimiter:

```
df = pd.read_csv("global_freelancers_raw.csv", sep=";")
```

This extracted all 12 columns properly.

## 3.4 Step 3 — Cleaning Column Names

Operations applied:

- Remove extra spaces.
- Convert to lowercase.
- Replace spaces with underscores.

## 3.5 Step 4 — Cleaning Text Columns

For text columns:

- Remove trailing and leading spaces.
- Normalize missing values ("NA", "N/A", "none", empty strings, "-", "?" → NaN).
- Standardize text formatting (title case).

## 3.6 Step 5 — Numeric Cleaning

A custom function removed:

- Currency symbols: $, , £.
- Letters such as "USD".
- Percent signs (%).

Then values were converted to numeric using:

```
pd.to_numeric(..., errors="coerce")
```

## 3.7 Step 6 — Categorical Normalization

**Gender Mapping**

- f, female, F, woman → Female
- m, male, M, man → Male
- anything else → Other

**is_active Mapping**

- 1, yes, true, Y → True
- 0, no, false, N → False

### 3.8 Step 7 — Removing Duplicates

Using `freelancer_id` as a key ensured uniqueness, so no major duplicates were found.

### 3.9 Step 8 — Missing Value Imputation

- Numeric columns: imputed with the median.
- Categorical columns: imputed with the mode.

### 3.10 Step 9 — Outlier Treatment

- Rating clipped to the range 0–5.
- Satisfaction clipped to 0–100.
- Hourly rate winsorized using the Interquartile Range (IQR).

### 3.11 Step 10 — Saving the Clean Dataset

```
df.to_csv("global_freelancers_clean.csv", index=False)
```

## 4 Exploratory Data Analysis (EDA)

Key results after cleaning:

### 4.1 Global Statistics

- Total freelancers: **1000**.
- Countries represented: **21**.
- Languages: **16**.
- Primary skills: **10**.

### 4.2 Demographics

- Mean age: around **40.5** years.
- Mean experience: around **11.2** years.

### 4.3 Top Skills

Most frequent skills:

- DevOps (112 freelancers).
- UI/UX Design (109 freelancers).
- Blockchain Development (105 freelancers).

### 4.4 Hourly Rate Insights

- Mean hourly rate: around **51.3** USD.
- Highest average rates in countries such as India, Argentina and Indonesia.

### 4.5 Gender Difference in Rates

- Female average hourly rate $\approx$ 51.9 USD.
- Male average hourly rate $\approx$ 50.7 USD.

### 4.6 Rating vs Satisfaction

The correlation between rating and client satisfaction is approximately zero, so there is no strong linear relationship between the two variables in this dataset.

## 5 Power BI Dashboard

### 5.1 Data Loading

The cleaned dataset `global_freelancers_clean.csv` was imported into Power BI. No complex data model was required, as the data is stored in a single, clean table.

### 5.2 Measures

Examples of measures defined in Power BI:

- Total Freelancers.
- Average Hourly Rate.
- Average Client Satisfaction.
- Average Rating.

## 5.3 Visuals

The dashboard includes:

- KPI cards (total freelancers, average rate, satisfaction, rating).
- Bar charts (freelancers by skill, hourly rate by country).
- Column charts (satisfaction by skill, hourly rate by gender).
- Slicers for country, skill, gender and activity status.

## 5.4 Screenshots

Figure 5 shows the main visuals built in Power BI for this project.

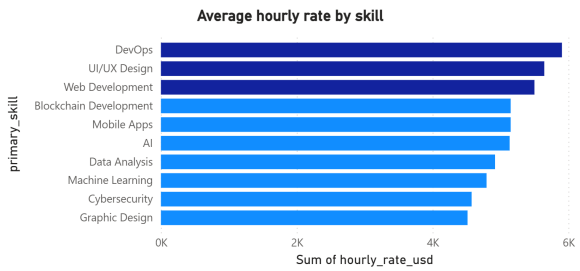


Figure 1: *
Average hourly rate by skill

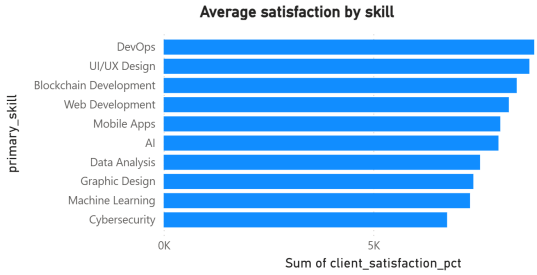


Figure 2: *
Average satisfaction by skill

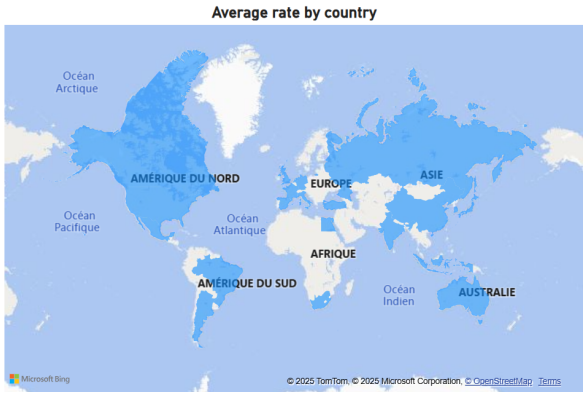


Figure 3: *
Average hourly rate by country

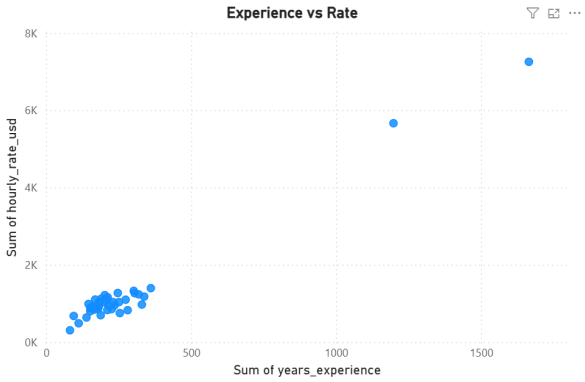


Figure 4: *
Experience vs hourly rate

Figure 5: Main Power BI visuals used in the Global Freelancers dashboard.

# 6 Conclusion

This project demonstrates a complete data analytics workflow:

- Cleaning and transforming raw data in Python.
- Performing exploratory data analysis.
- Building an interactive dashboard in Power BI.

It strengthened key skills for a junior Data Analyst:

- Data cleaning (Pandas).
- Descriptive statistics and EDA.
- Dashboard design and storytelling with data.