**✅ 1. Overall Data Summary**

* **Rows:** 100,000 entries
* **Columns:** 18 columns including both numerical and categorical data
* **Data Source:** Loaded from PostgreSQL (originally CSV)

### 📌 ****2. Column-wise Observations****

| **Column** | **Key Insight** |
| --- | --- |
| job\_title | Clean — 12 unique roles; might contain typos (e.g., casing, whitespace) |
| experience\_level | 20% missing — should be imputed or encoded |
| employment\_type | 24% missing — needs imputation |
| company\_size | Good — 3 unique values, no missing data |
| company\_location | Clean — 6 countries, categorical |
| remote\_ratio | Ranges from 0 to 100 — can be used to create a binary is\_remote feature |
| salary\_currency | 5 currencies — useful for exchange normalization |
| years\_experience | Int, no missing — useful as a predictive feature |
| base\_salary | Some negative values (likely errors), must clean |
| bonus, stock\_options | Good distribution — useful for computing total\_salary |
| total\_salary | Already computed as base + bonus + stock — redundant if recomputed |
| salary\_in\_usd | Currency-normalized version — useful as a comparison target |
| currency | May be inconsistent with salary\_currency — should be cross-verified |
| conversion\_rate | Must be used to compute adjusted\_total\_usd (if re-derived) |
| adjusted\_total\_usd | Final target column (for modeling) — clean and complete |
| education, skills | **100% missing** — should be dropped unless data can be supplemented |

### ⚠️ ****3. Issues Detected****

* **Missing Values:**
  + experience\_level: 20% missing → needs imputation (mode/ML-based)
  + employment\_type: 24% missing → use mode
  + education, skills: 100% missing → drop unless filled externally
* **Outliers:**
  + base\_salary includes negative values → **must be removed**
  + salary\_in\_usd max value ≈ $2.3M — may be outlier
* **Redundancy:**
  + total\_salary is already a sum — can be dropped if recalculated
  + salary\_in\_usd vs adjusted\_total\_usd — keep the most accurate for modeling

### 🧠 ****4. Modeling Consideration****

You can use:

* **Target variable**: adjusted\_total\_usd
* **Drop or recalculate**: total\_salary, salary\_in\_usd
* **Drop**: education, skills, currency (if inconsistent)
* **Encode**: job\_title, employment\_type, company\_location, salary\_currency, company\_size
* **Create**: is\_remote from remote\_ratio

## 🔍 **Key Bivariate Analysis Insights**

### 📊 1. ****Job Title vs Adjusted Total Salary****

* **Observation:** Median salary varies significantly across job titles.
* **Insight:**
  + **Data Scientists** and **Machine Learning Engineers** show higher median salaries compared to **Data Analysts** and **Data Engineers**.
  + **Recommendation:** Use job\_title as a strong categorical predictor. Consider grouping low-frequency roles if any.

### 📊 2. ****Experience Level vs Adjusted Total Salary****

* **Observation:** Salary increases with seniority.
* **Insight:**
  + **Entry-level roles** have a clearly lower salary distribution.
  + **Mid and Senior levels** show overlapping but higher ranges.
  + **Recommendation:** Map experience\_level to an ordinal scale (e.g., Entry → 0, Mid → 1, Senior → 2, Exec → 3) for modeling.

### 📊 3. ****Company Size vs Adjusted Total Salary****

* **Observation:**
  + **Large companies** generally offer higher compensation.
  + **Small and Medium companies** have more tightly clustered lower salaries.
* **Insight:** Company size plays a moderate role in salary prediction.

### 📊 4. ****Employment Type vs Adjusted Total Salary****

* **Observation:**
  + **Full-time** employees receive much higher salaries.
  + **Part-time, Contract, and Internship roles** have substantially lower distributions.
* **Insight:**
  + Employment type is a strong categorical predictor.
  + Should be included using one-hot or label encoding.

### 📊 5. ****Remote Ratio vs Adjusted Salary****

* **Observation:**
  + There is **no consistent increase or decrease** in salary based on remote ratio alone.
  + However, **fully remote roles (100%)** show **wider variance** in salary.
* **Insight:**
  + Binary encoding (is\_remote) is still useful.
  + Explore interaction with job role and location.

### 📊 6. ****Years of Experience vs Adjusted Salary (Numerical-Numerical)****

* **Observation:**
  + Positive but **weak correlation** (~0.35–0.4).
  + Salaries generally rise with experience, but the relationship is not perfectly linear.
* **Insight:**
  + Use years of experience as a numerical feature.
  + Consider binning or using polynomial terms if needed.

## ✅ Summary Recommendations:

| **Feature** | **Type** | **Use in Model?** | **Notes** |
| --- | --- | --- | --- |
| job\_title | Categorical | ✅ Yes | Strong predictor, encode properly |
| experience\_level | Categorical | ✅ Yes | Ordinal encoding recommended |
| employment\_type | Categorical | ✅ Yes | Clear salary separation by type |
| company\_size | Categorical | ✅ Yes | Moderate influence, encode |
| years\_experience | Numerical | ✅ Yes | Keep, possible scaling or binning |
| remote\_ratio | Numerical | ✅ Yes | Convert to binary or use directly |