# Exploratory Data Analysis Report

Dataset: Stack Overflow Developer Survey

## 1. Objective

The goal of this analysis is to understand the distribution of developer salaries, demographics (gender, age), and identify potential outliers in the dataset.

## 2. Data Overview

**Rows:** ~11,552

**Columns:** Multiple, including `ConvertedComp` (annual compensation in USD), `Gender`, and `Age`., `ConvertedComp` values represent annual salaries converted to USD based on February 1, 2019 exchange rates, assuming 12 months and 50 work weeks.

---

# 3. Key Analyses & Insights

- Salary Distribution (`ConvertedComp`)

The salary data is **right-skewed** with most respondents earning within the lower range.

Median salary: **~$57,000 USD**.

Histogram and KDE plots reveal a high concentration of salaries between $0 and $100,000, with some extreme high salaries (outliers)

- Gender Demographics

**Majority** of respondents identified as *Man*.

A subset identified as *Woman*, *Non-binary*, or other categories.

Median salary for women respondents: **~$54,000 USD** (slightly lower than overall median).

- Age Distribution

Ages range from teenage years to late 70s, but most respondents are between 20 and 40 years old.

Five-number summary of **`Age`**:

Min: ~10 years

Q1: ~25 years

Median: ~29 years

Q3: ~35 years

Max: ~99 years

Age histogram shows a strong concentration in the **25–35 age group**.

- Outlier Detection

**ConvertedComp** boxplot reveals extreme outliers on the higher end (>$250,000 USD).

Interquartile Range (IQR) method confirms significant outliers in the top 1–2% of salaries.

---

## 4. Observations

Salary distribution is heavily skewed, indicating that while a small portion earns very high salaries, the majority earn moderate incomes.

Gender disparity exists in both representation and median salaries.

The dataset primarily represents a **young developer population**.

---

## 5. Recommendations

**Outlier Treatment:** For statistical modeling, consider capping or transforming high salaries to reduce skewness.

**Gender Pay Analysis:** Further exploration into pay differences across genders could uncover systemic disparities.

**Age & Career Stage Study:** Cross-analyze age with experience and salary to find career progression trends.

**Regional Analysis:** Incorporate country/region data for more localized insights.

## Visualizations

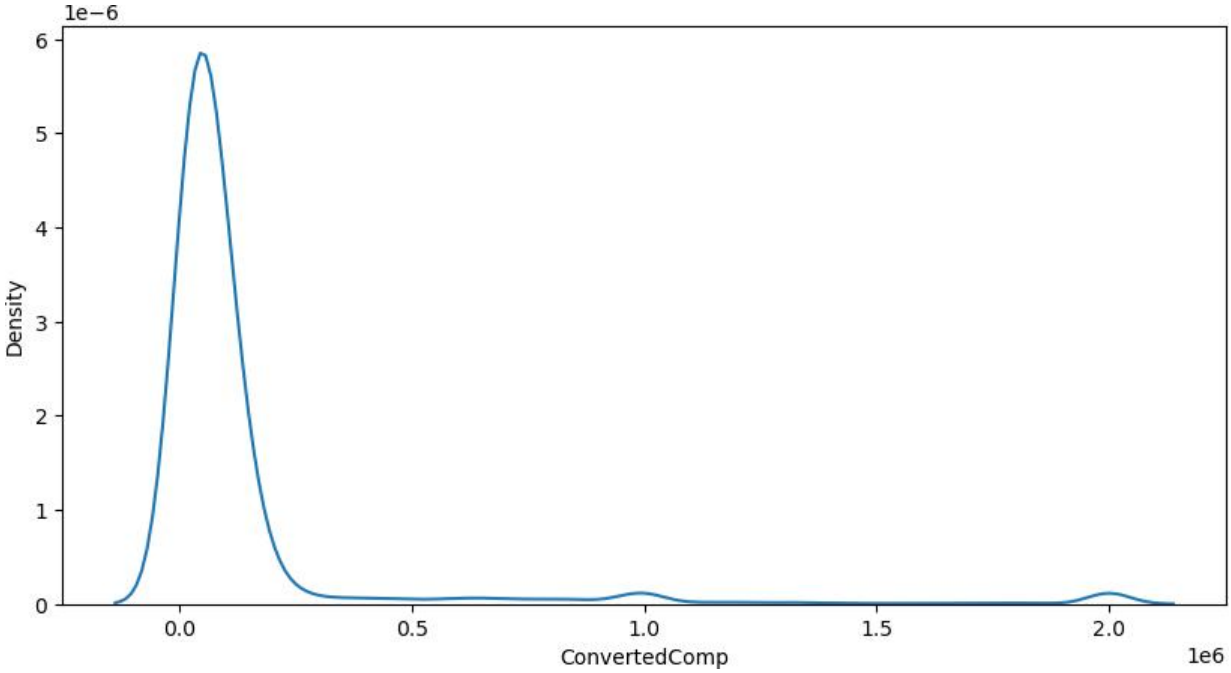


Figure 1: Understand the shape of the salary distribution.

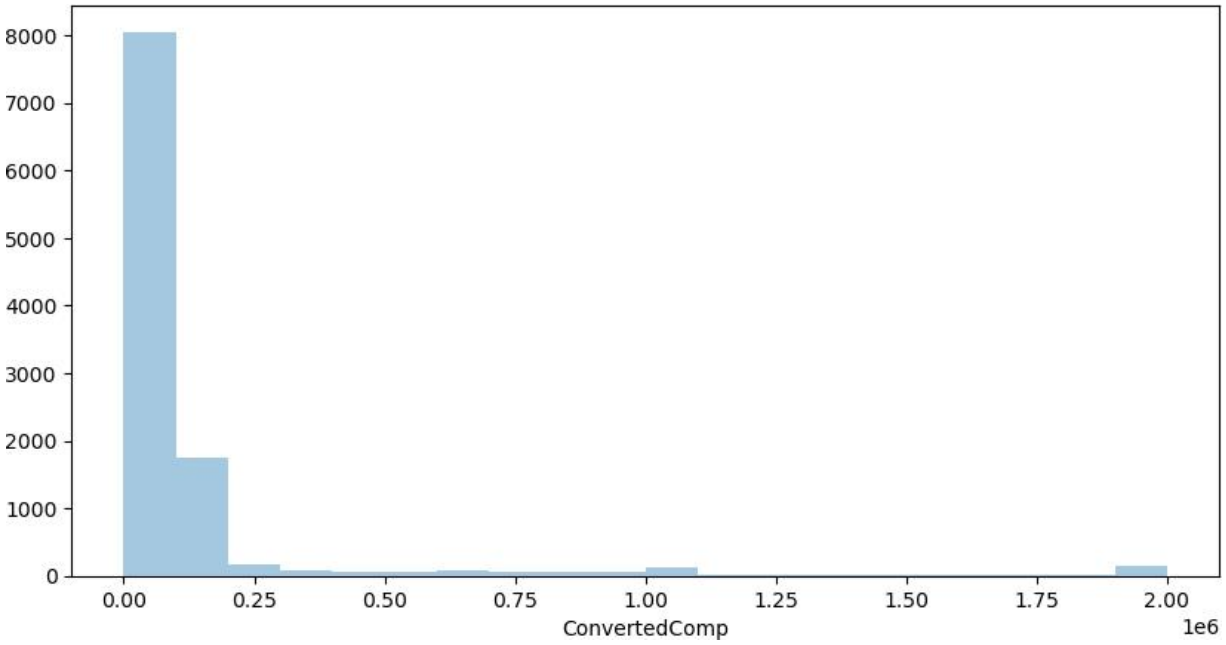


Figure 2: Most respondents earn in the lower salary bins

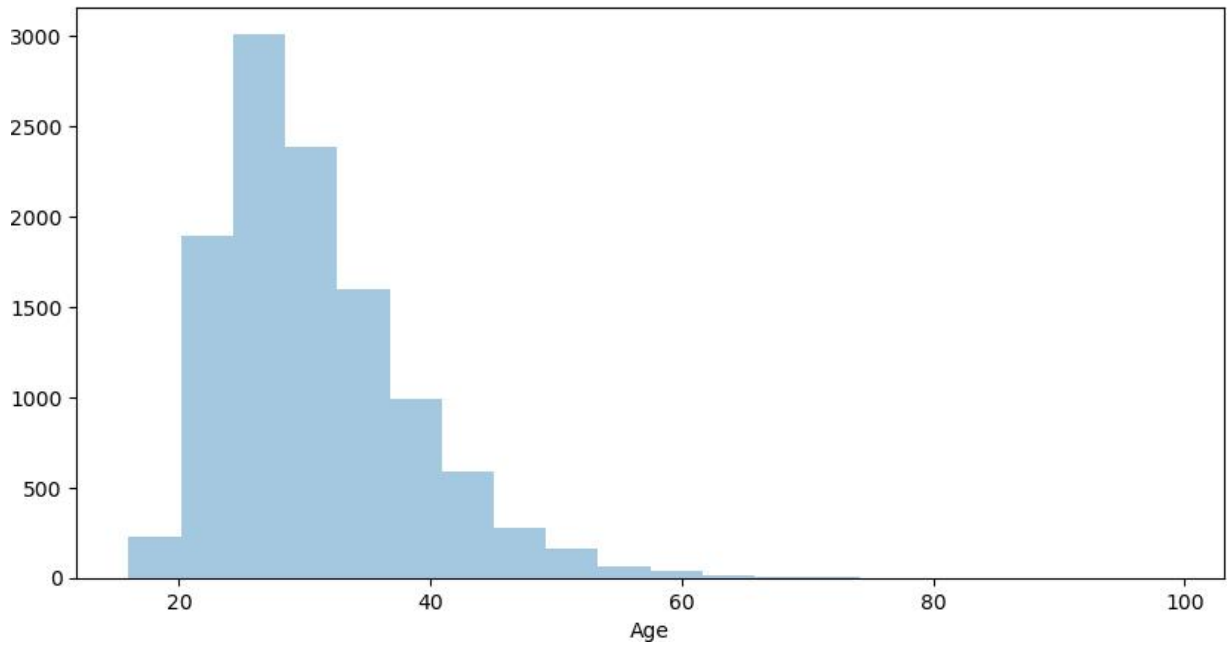


Figure 3: Age distribution of survey respondents.

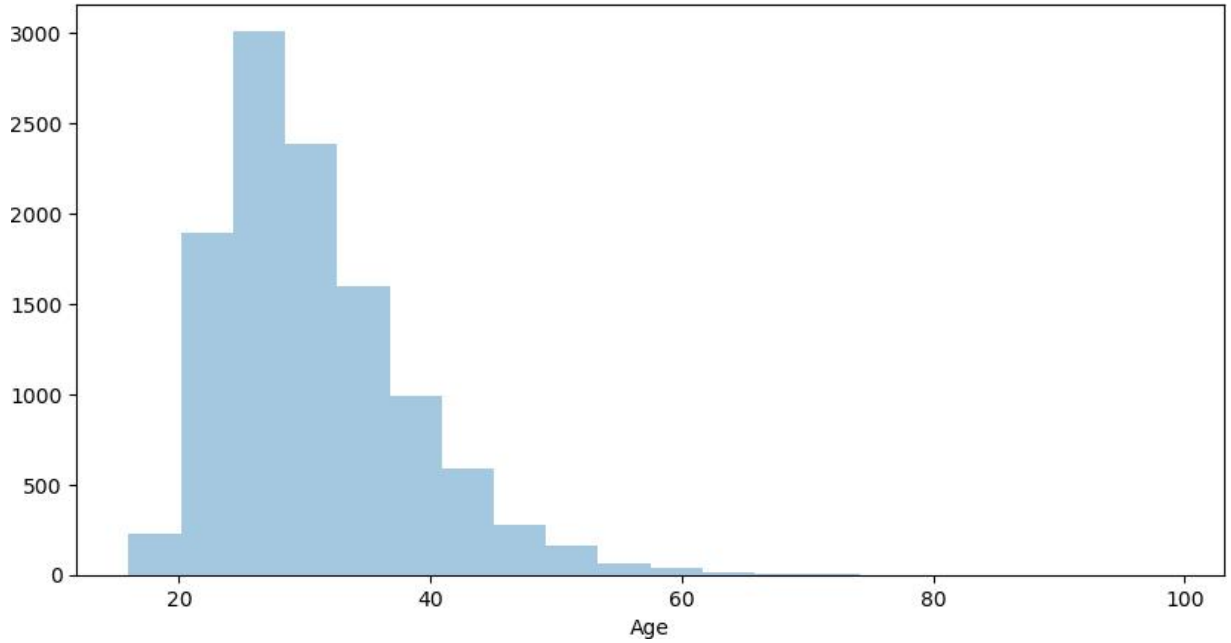


Figure 4: **Ages are distributed** among survey respondents.

Likely reveals a **concentration in younger age ranges** (20 – 40 years old), with fewer older respondents.

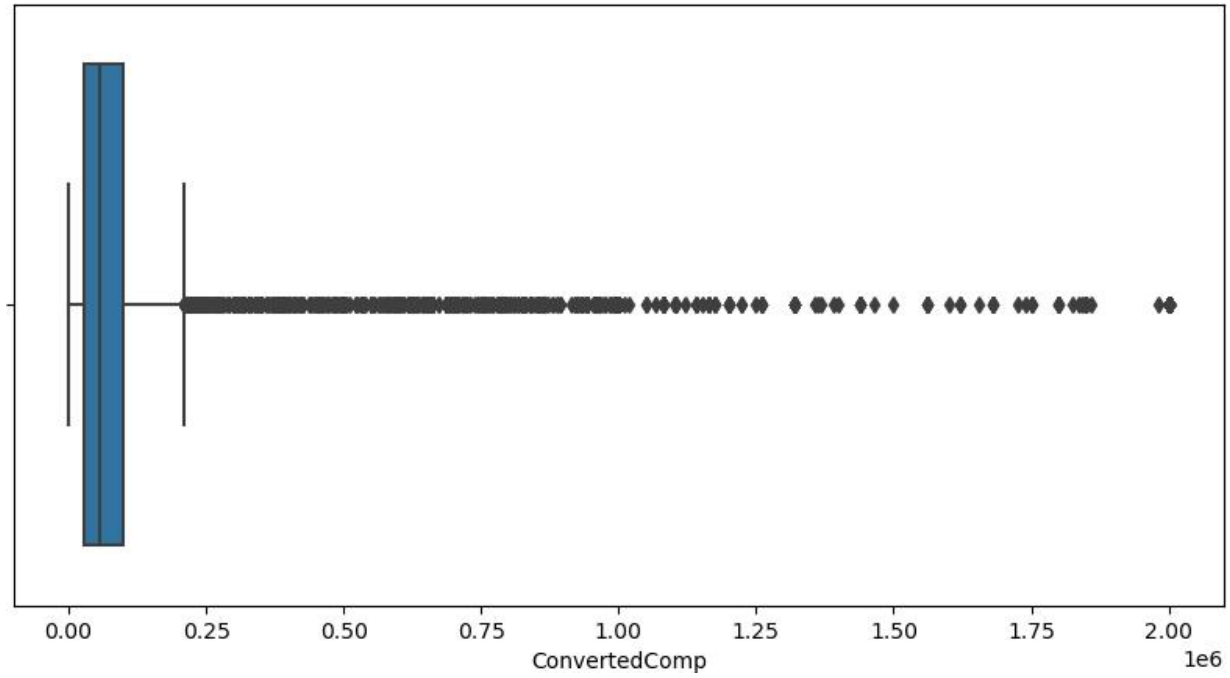


Figure 5: `outliers` in salary data.

In dataset, many high-end outliers likely appear far above the upper whisker, showing that a small number of people earn much higher salaries than the majority.Confirms that the salary distribution is heavily skewed.

## Recommendations

• Apply a log or box-cox transformation to ConvertedComp for modeling to reduce skewness.

• Perform statistical tests (e.g., t-test or Mann-Whitney) to analyze gender-based salary differences.

• Remove or cap extreme outliers when using algorithms sensitive to scale or use robust models.

• Engineer temporal and experience-based features if available to improve predictive power.

• Consider stratified sampling or resampling if you build models on subgroups (e.g., by country or role).