Report

Stack overflow software developer survey analysis

(Big Data Analysis)

Date Analyst: Andrii Isachenko

2026

Technology stack: Python (Pandas, Matplotlib, Seaborn), Jupyter Notebook / Google Colab.

---

1. Business Problem

The goal of the project is to conduct a comprehensive audit of the state of the IT industry in 2025 based on big data.

Key analytical questions:

• What is the real income level of Python developers in different geographical regions after cleaning from anomalies?

• How have educational trends changed: is self-education replacing traditional university degrees?

• What proportion of the community works remotely and which industries offer the highest pay for Remote-format?

• How to detect data "clutter" (Outliers) in open surveys?

2. Data Acquisition

1. **Source:** Stack Overflow Annual Developer Survey 2025.

2. **Format:** Data set in CSV format

(survey_results_public.csv, survey_results_schema.csv).

3. **Size:** The sample consists of 49,191 respondents. The data includes over 70 technical and demographic indicators.

3. Data Preparation and Consolidation

At this stage, a dataset was created and data hygiene was performed:

1. **Import:** Using the Pandas library to load and initially inspect the structure.

2. **Text Normalization:** * Removing extra spaces (strip).

◦ Capitalizing categorical data (countries, employment types).

◦ Important: For technical columns (programming languages), the original case (e.g. JavaScript, HTML/CSS) was preserved to avoid damaging the terminology.

3. **Optimization:** Eliminating fragmentation of the DataFrame using the .copy() method to increase the speed of calculations.

4. Outliers Management

A two-level filtering system was implemented to ensure financial accuracy:

1. **Segmentation:** Creating a separate salary_df dataframe for respondents who reported their income.

2. **Quantile Method:** Removing the 1% of the lowest and 1% of the highest salaries. This allowed us to filter out “noise” (salaries of $1 or $1,000,000), which are often input errors or jokes.

3. **Documentation:** Creating a separate salary_outliers.csv report with suspicious data for further root cause analysis.

5. Statistical Analysis and Key Insights

5.1. Demographics and Work Experience

- **Statistics:** Average experience — 13.37 years, median and mode — 10.0 years.

- **Conclusion:** The market is represented by an experienced core of specialists. Positive asymmetry (mean > median) indicates the presence of a significant group of industry “veterans” (30+ years of experience).

5.2. Education Trends: Triumph of Online Courses

- **Result:** 21,212 people (~43%) used online courses for training.

- **Conclusion:** Traditional education is losing its monopoly. Self-education has become a full-fledged alternative, requiring employers to review their assessment criteria (skills-first approach).

5.3. Python Ecosystem and Remote Work

- **Popularity:** About 37.5% of developers use Python. This confirms its leadership in the AI era.

- **Flexibility:** 17,663 people work completely remotely. Remote work is no longer a bonus, but an industry standard.

6. Geographic Analysis of Python Compensation

Analysis of median annual salaries revealed critical gaps:

- **Leaders:** The US and Western European countries show the highest rates.

- **Emerging markets:** The huge gap (e.g. Pakistan — $8,428) explains why Remote Work is vital — it allows talent to access global capital.

7. Analysis of industries for high-paying Remote

Top 3 industries where the most Python developers work remotely with high pay:

1. Software Development (1,448 people).

2. Fintech (243 people).

3. Healthcare (233 people).

- **Conclusion:** Even conservative sectors (banking, medicine) are actively adapting to the remote hiring model.

8. Strategic conclusions and recommendations (Action Plan)

- **For recruiting:** It is recommended to focus on candidates with online course certificates, as this is the most dynamic talent pool.

- **For business:** Using the Remote model in emerging markets allows you to attract high expertise at an optimal cost.

- **For analytics:** The next step should be a deep audit of the salary_outliers.csv file to understand which specific market niches generate extremely high incomes.