**Data Science with Python**

**PROJECT REPORT**

**(Project Semester January-April 2025)**

***An Exploratory Analysis of Global Developer Trends and Influencing Demographics Using Stack Overflow Survey (2023)***

**Submitted by:**

**Shyam Krishna**

**Registration No: 12307238**

**Programme and Section: B. Tech Cse K23-GS**

**Course Code: -INT375**

**Under the Guidance of**

**Gargi Sharma UID: 29439**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

**I, Shyam Krishna , student of B. Tech Cse (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.**

**Date: 11-04-2025**

**Signature:**

****

**Registration No: 12307238**

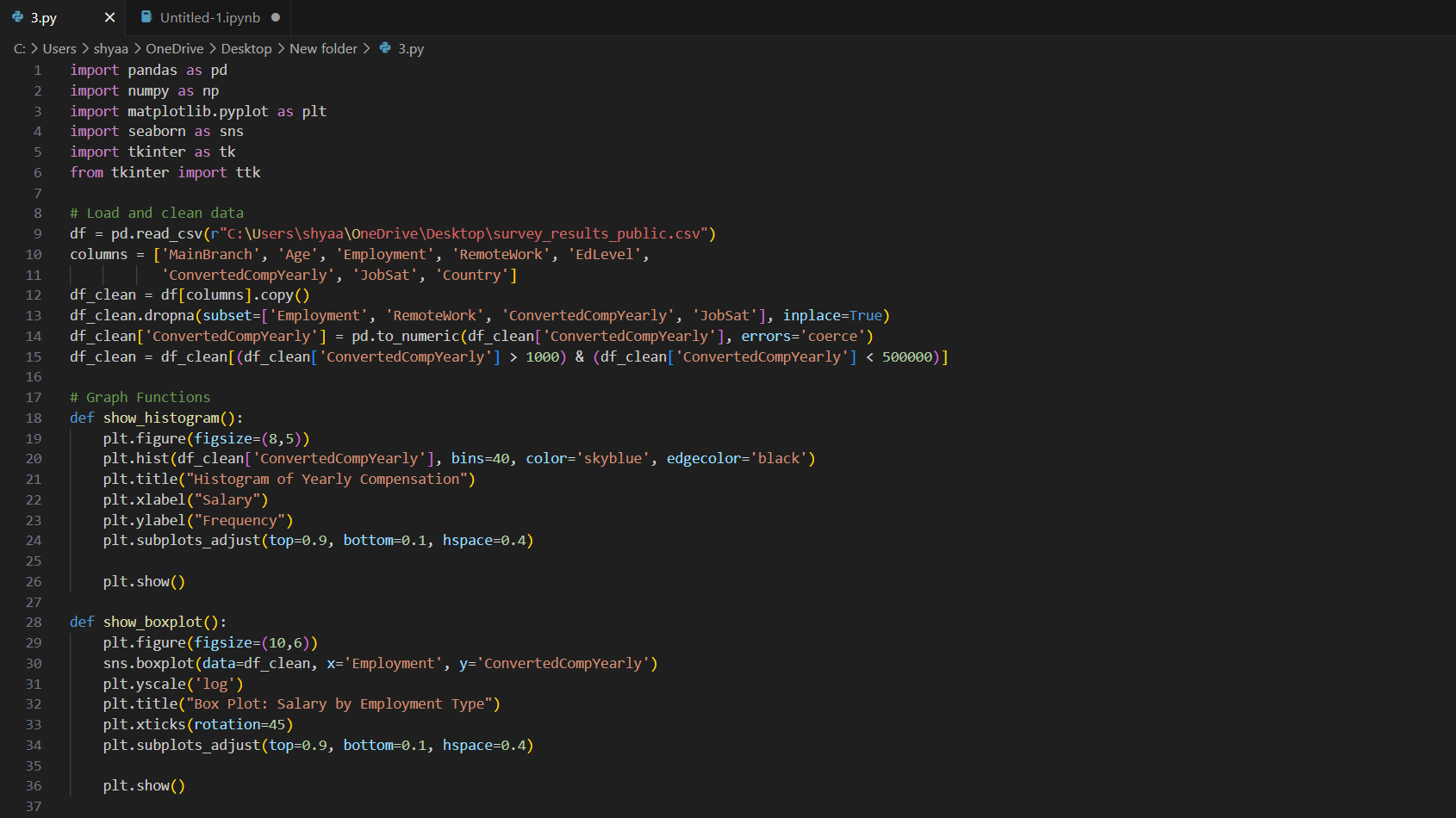
**Name of the student: Shyam Krishna**

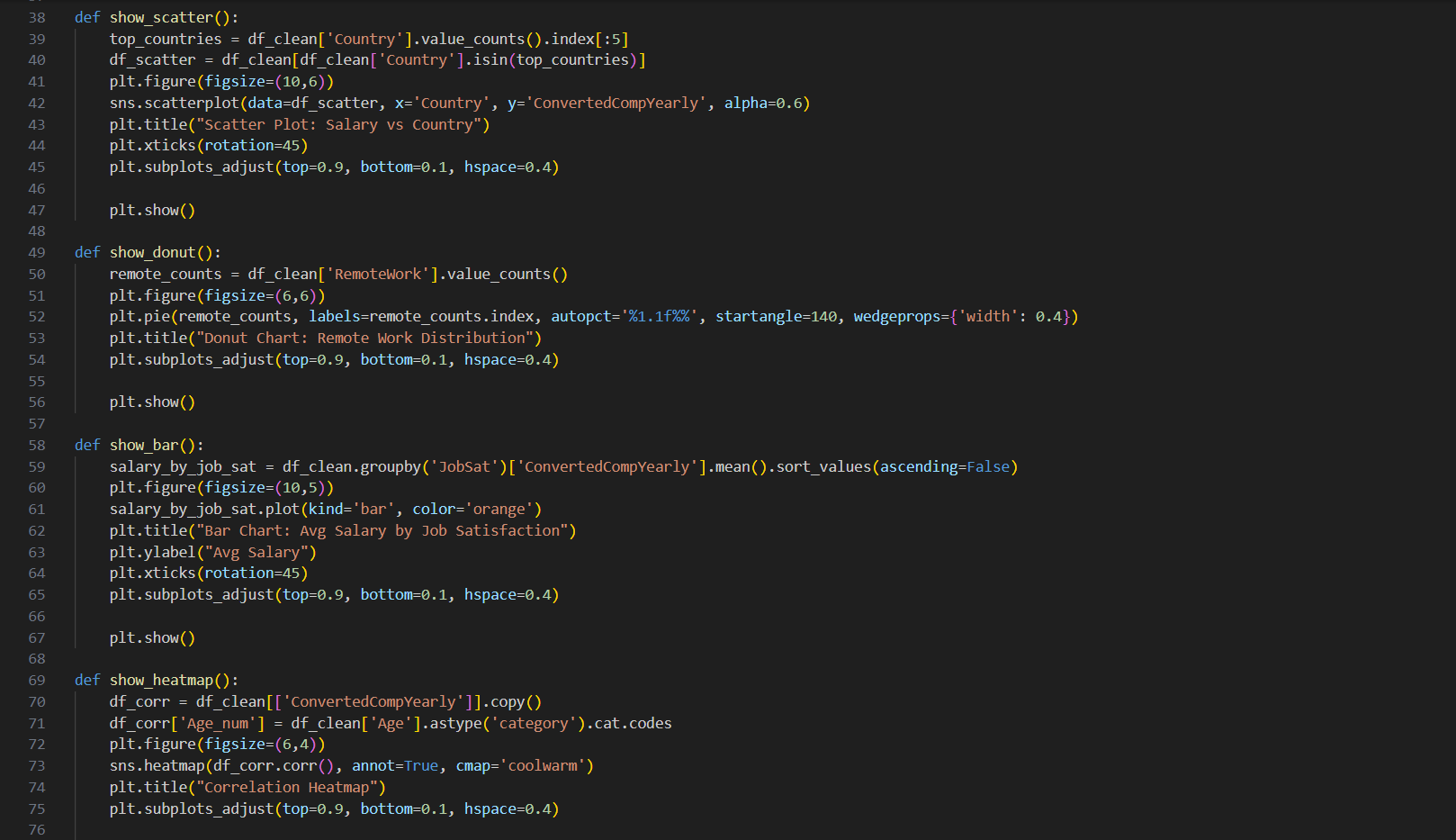
**Introduction**The *Stack Overflow Developer Survey* serves as a critical annual benchmark for understanding the global developer ecosystem. By capturing data on skills, tools, education, and workplace conditions, it offers unparalleled insights into the evolving preferences and challenges of software professionals. This report presents a structured exploratory analysis of the 2023 survey dataset, focusing on key trends in technology adoption, compensation patterns, and demographic influences. The findings aim to illuminate industry-wide shifts and empower stakeholders—from developers to tech leaders—to make data-driven decisions in a rapidly changing digital landscape.

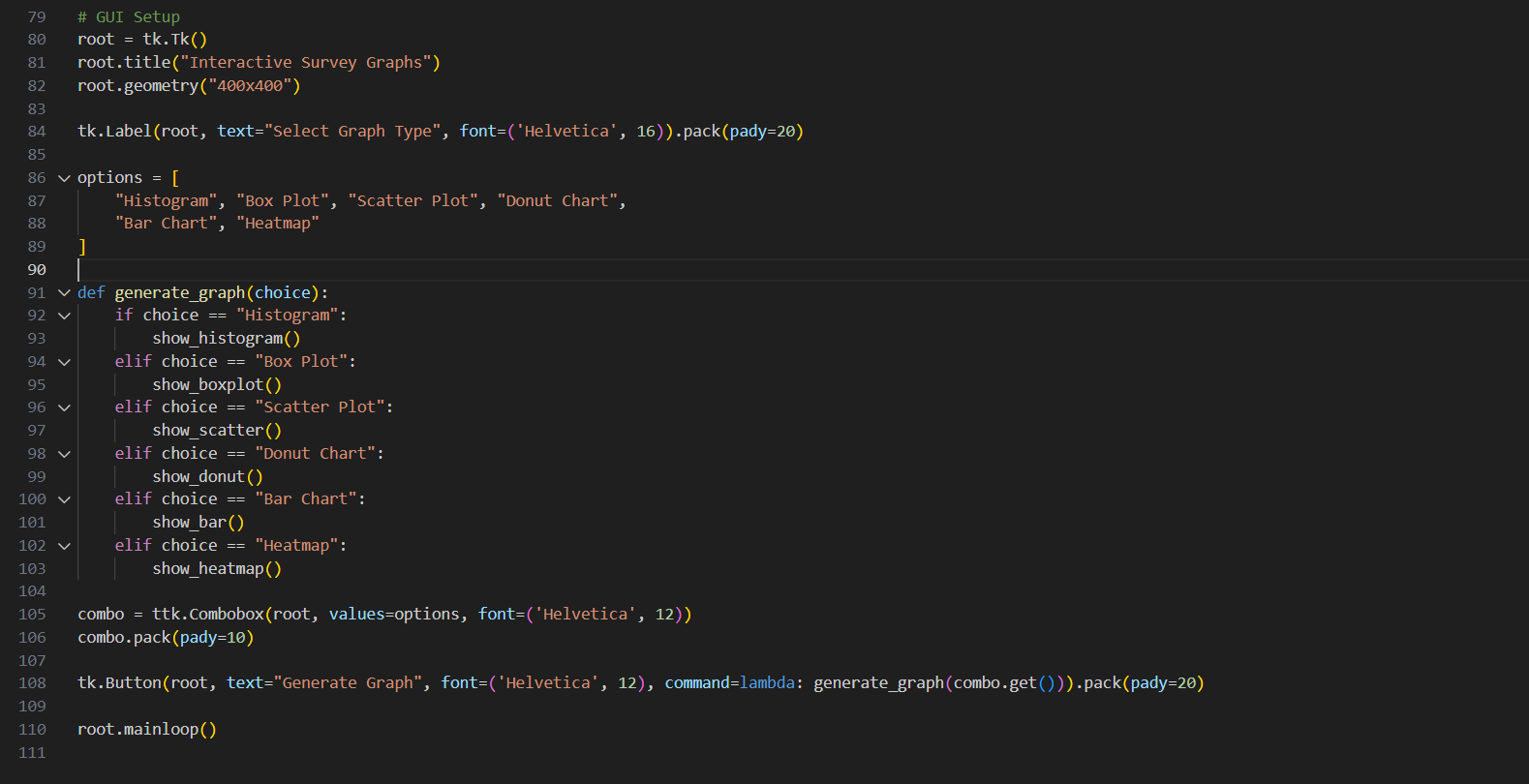
**Source of Dataset**

Link: https://survey.stackoverflow.co/

**💻EDA PROCESS**

****

****

****

**ii. General Objectives of the Project**

This report analyses six core research questions:

1. Analyse yearly compensation distribution.
2. Compare yearly compensation by employment type.
3. Compare yearly compensation by country.
4. Compare remote work distribution.
5. Compare average salary by job satisfaction.
6. Analyse Correlation Between Yearly Compensation and Age.

Each section includes statistical analysis, visualizations, and interpretive insights.

iii. Libraries & Tools Used  
• Pandas, NumPy: Data cleaning, transformation, and aggregation

• Matplotlib, Seaborn: Visualizations to uncover trends in compensation, job satisfaction, and technology adoption.

• Tkinter: Interactive GUI development for user-friendly data exploration.

**iv. Analysis & Results**

**🔹 Objective 1: Analyse Yearly Compensation Distribution**

I plotted a histogram to visualize the distribution of yearly compensation among respondents.

**Key Metrics**

* X-axis: Salary (Converted Compensation Yearly)
* Y-axis: Frequency (number of respondents in each salary range)

**Key Insight**

* The majority of respondents earn between $0 to $100,000, with frequency peaking around the lower salary bands.
* There's a sharp decline in frequency as compensation increases beyond $100,000, with very few outliers above $300,000.
* The distribution is right-skewed, indicating that high salaries are rare and most professionals earn on the lower end of the scale.

**Visualization**

* Histogram with 40 bins displaying the frequency distribution of yearly compensation.
* Custom styling: Sky blue bars with black edges for clear contrast and readability.

**i. Introduction**

Understanding salary distribution is critical for evaluating income patterns within the tech workforce. In this section, I analyze the yearly compensation figures to identify the most common salary ranges, detect outliers, and observe the overall distribution trend.

**ii. General Description**

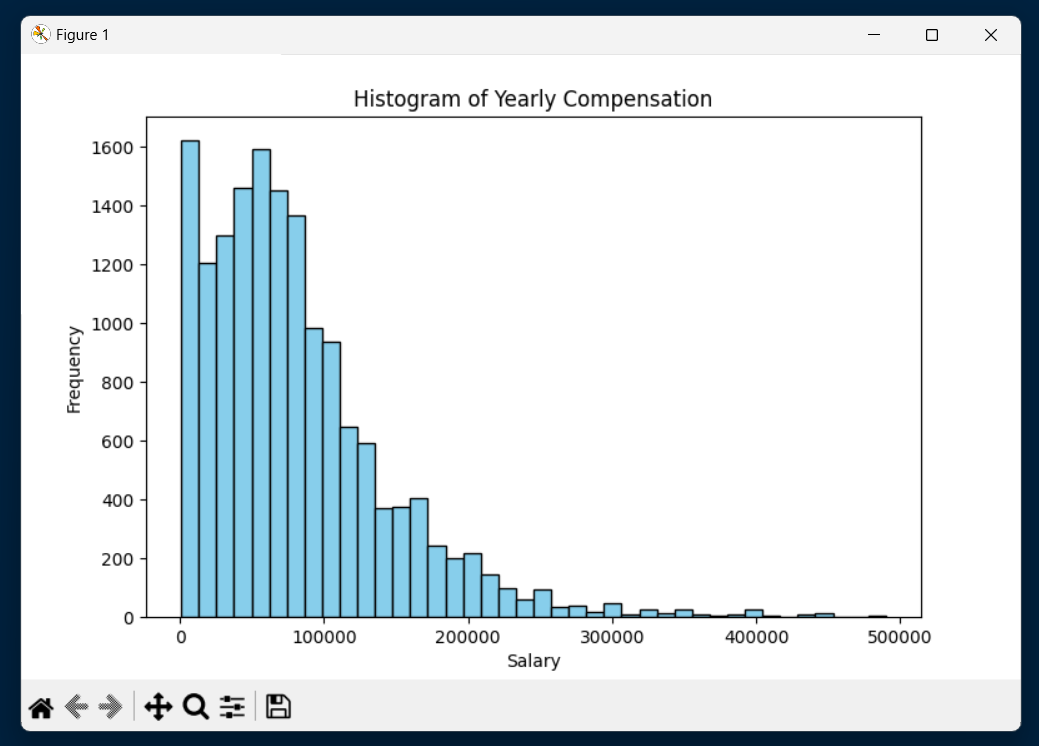
I’ll:

1. Use the ConvertedCompYearly column from the cleaned dataset.
2. Plot a histogram with 40 bins to capture a detailed view of compensation ranges.
3. Examine how compensation is distributed — identifying where most salaries cluster and how quickly frequencies taper off as compensation rises.

**iii. Functions & Formulas**

1. Histogram Plotting:
   * Used plt.hist() with parameters:
     + bins=40
     + color='skyblue'
     + edgecolor='black'
2. Defined axis labels and title for clear communication.
3. Applied plt.subplots\_adjust() to optimize spacing and layout.





**Objective 2: Compare Yearly Compensation by Employment Type**

I created a box plot to analyze how yearly compensation varies across different employment types.

**Key Metrics**

* **X-axis:** Employment Type
* **Y-axis:** Yearly Compensation (on a logarithmic scale)

**Key Insight**

* **Full-time employed professionals** show a wide compensation range with a high median salary and numerous high-value outliers.
* **Students, part-time workers, and unemployed respondents** generally have lower compensation figures, with median salaries significantly below those of full-time professionals.
* The log scale highlights the presence of substantial salary outliers in most employment categories, especially among full-time workers and self-employed professionals.
* Certain categories like **retired individuals and those unable to work** cluster tightly at lower compensation levels.

**Visualization**

* **Box plot of salary distribution** for each employment category.
* Applied a **logarithmic scale on the Y-axis** to handle skewed salary data and better visualize compensation spread and outliers.
* X-axis labels rotated for improved readability.

**i. Introduction**

Employment type can significantly influence an individual’s compensation. In this section, I analyze how yearly salaries differ across various employment categories, from full-time professionals to students and unemployed individuals, to identify income disparities and compensation trends within the workforce.

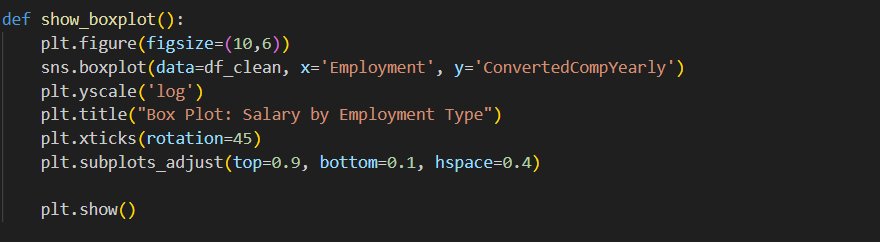
**ii. General Description**

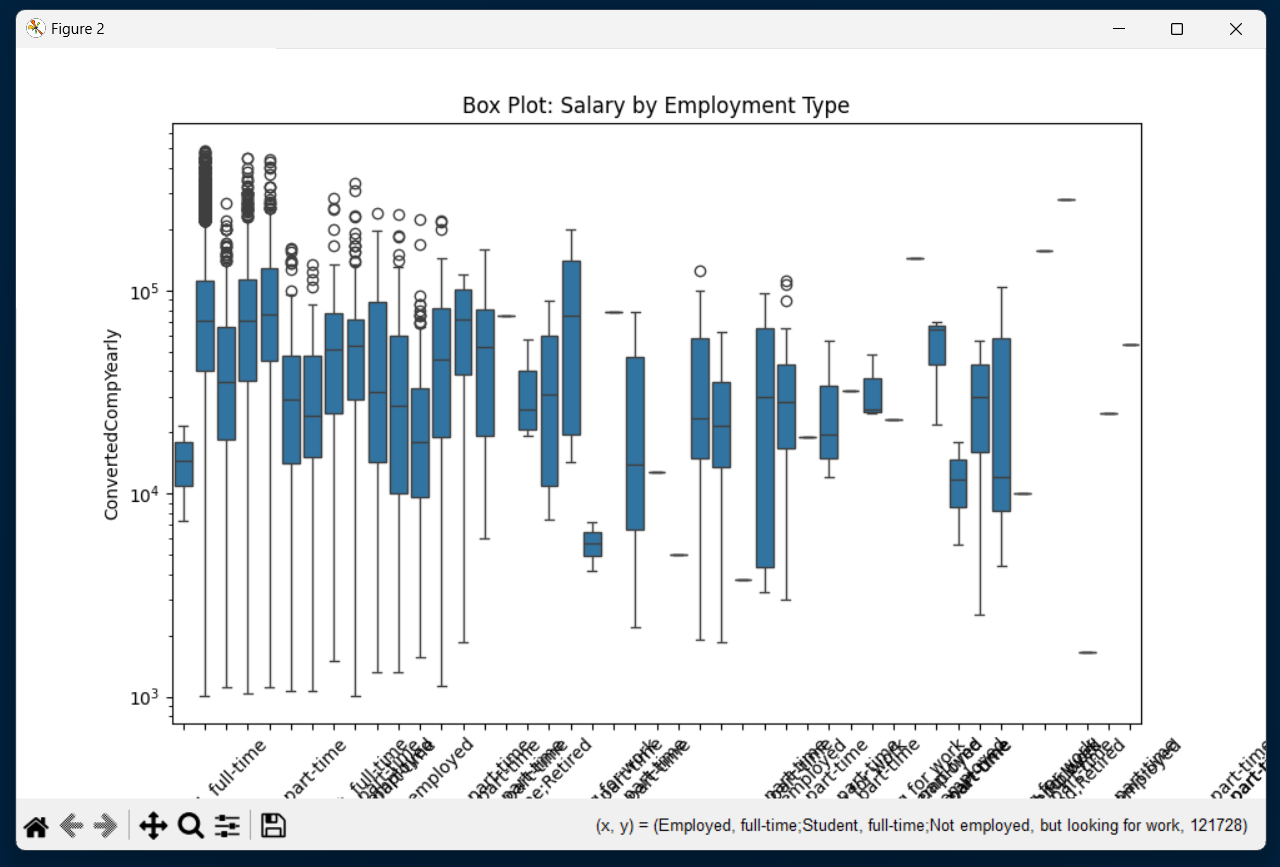
I’ll:

1. Use the **Employment** and **ConvertedCompYearly** columns from the cleaned dataset.
2. Plot a **box plot** to display the distribution, median, interquartile range (IQR), and outliers for each employment type.
3. Use a **logarithmic scale** on the Y-axis to normalize salary data and emphasize differences across groups.
4. Observe which employment categories show higher salary medians and greater variability.

**iii. Functions & Formulas**

1. **Box Plot Creation:**
   * Used sns.boxplot() with parameters:
     + x='Employment'
     + y='ConvertedCompYearly'
     + data=df\_clean
2. Applied plt.yscale('log') to handle large disparities in compensation data.
3. Added axis labels, title, and rotated X-axis labels for clarity.
4. Adjusted figure layout using plt.subplots\_adjust() for optimal spacing.





**🔹 Objective 3: Compare Yearly Compensation by Country**

I created a scatter plot to analyze how yearly compensation varies across respondents from the top five countries by respondent count.

**Key Metrics**

* X-axis: Country
* Y-axis: Yearly Compensation

**Key Insight**

* North America shows a higher overall concentration of high salaries, with several respondents earning over $400,000.
* Germany and Northern Ireland also feature high earners but in fewer numbers compared to North America.
* India and Ukraine cluster heavily at the lower end of the compensation spectrum, with fewer high-value outliers.
* The scatter plot confirms a visible income disparity between countries, highlighting stronger salary distributions in certain economies.

**Visualization**

* Scatter plot of yearly salary distribution for the top five respondent countries.
* Applied alpha=0.6 for point transparency to manage overlapping data points.
* X-axis labels rotated for readability.
* Cleaned and filtered data for only the top five countries with the highest respondent count.

**i. Introduction**

Geographical location can play a pivotal role in determining salary levels due to differences in economies, living standards, and market demands. In this section, I analyze how compensation levels compare across respondents from the five most represented countries in the dataset.

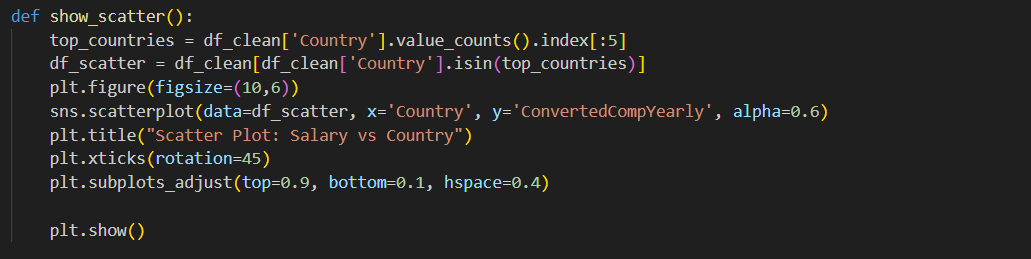
**ii. General Description**

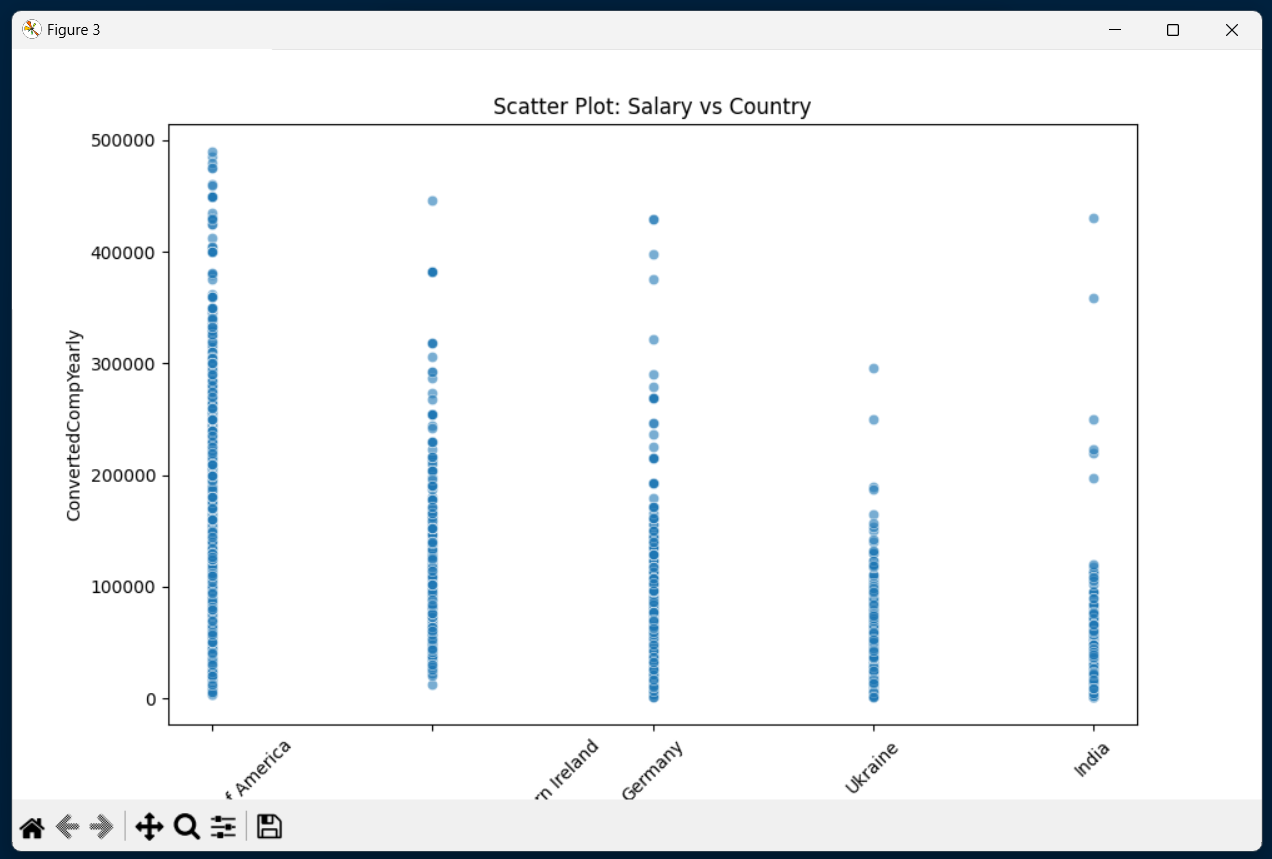
I’ll:

1. Use the Country and ConvertedCompYearly columns from the cleaned dataset.
2. Identify the top five countries with the most respondents.
3. Plot a scatter plot of yearly salaries against these countries to visualize income spread and outliers.
4. Compare concentration patterns and high-value earners between countries.

**iii. Functions & Formulas**

1. Data Filtering:
   * Extracted the top five countries using value\_counts().index[:5].
   * Filtered the dataset to only include respondents from these countries.
2. Scatter Plot Creation:
   * Used sns.scatterplot() with parameters:
     + x='Country'
     + y='ConvertedCompYearly'
     + alpha=0.6 for better visual layering.
3. Added axis labels, title, and rotated X-axis labels for improved readability.
4. Adjusted figure layout using plt.subplots\_adjust() for optimal spacing.





**🔹 Objective 4: Compare Remote Work Distribution**

I created a donut chart to analyze how remote, in-person, and hybrid work arrangements are distributed among respondents in the dataset.

**Key Metrics**  
• Category: RemoteWork  
• Values: Count and percentage distribution across categories

**Key Insight**  
• Hybrid work is the most common arrangement, accounting for 43.2% of respondents.  
• Fully remote roles closely follow at 40.4%.  
• In-person positions are the least common at 16.4%.  
• The chart highlights the continued dominance of flexible and remote-friendly work models in the current workforce.

**Visualization**  
• Donut chart representing the distribution of work modes.  
• Displayed percentage labels for each category.  
• Used a hollow-centered pie chart (donut chart) for clear, modern visual presentation.

**i. Introduction**

The COVID-19 pandemic permanently altered professional work structures, accelerating the adoption of remote and hybrid models. In this analysis, I assess the current distribution of work preferences among respondents, providing insights into post-pandemic work culture and its lasting shifts.

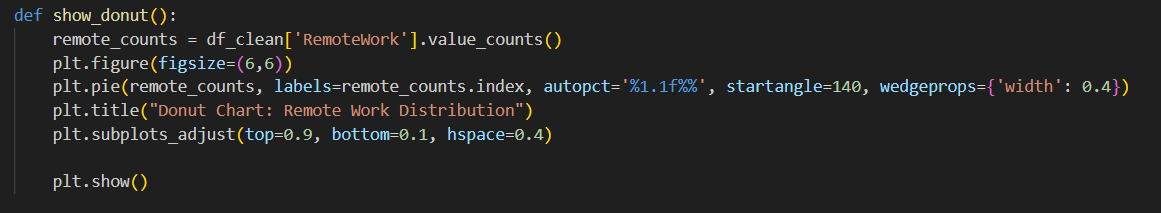
**ii. General Description**

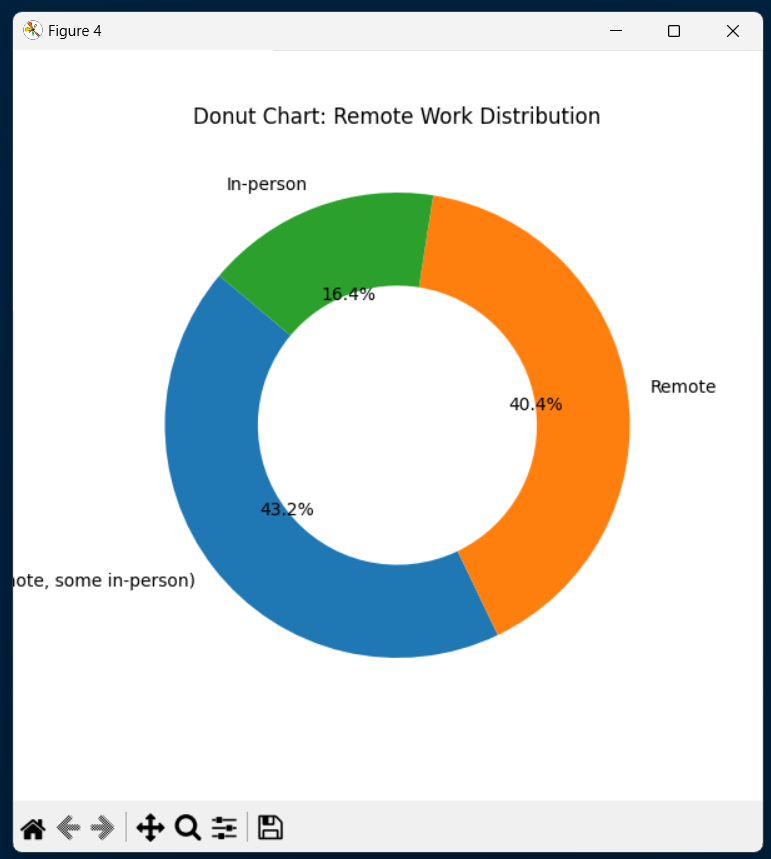
I’ll:

1. Use the RemoteWork category from the cleaned dataset.
2. Calculate the number of respondents for each work arrangement:  
   • Remote  
   • In-person  
   • Hybrid (some remote, some in-person)
3. Convert these counts into percentages to understand the overall distribution.
4. Visualize the results using a donut chart for clear comparative insight.

**iii. Functions & Formulas**

1. Data Aggregation  
   • Extracted value counts of the RemoteWork category from the cleaned dataset.
2. Donut Chart Creation  
   • Plotted a pie chart with a hollow center to visually represent the percentage distribution of each work mode.
3. Chart Formatting  
   • Applied percentage labels for each category, customized figure size, and layout for clarity and readability.





**🔹 Objective 5: Compare Average Salary by Job Satisfaction**

I created a bar chart to analyze how average yearly compensation varies across different levels of job satisfaction among respondents.

**Key Metrics**  
• X-axis: Job Satisfaction Levels  
• Y-axis: Average Yearly Salary

**Key Insight**  
• Higher job satisfaction correlates with higher average salary levels.  
• Respondents reporting 'Very satisfied' or 'Satisfied' tend to earn more than those less content with their jobs.  
• The chart emphasizes the potential financial benefit of aligning job roles with employee satisfaction.

**Visualization**  
• Bar chart representing the average salary for each job satisfaction category.  
• Applied color customization for visual emphasis.  
• X-axis labels rotated for readability.  
• Adjusted chart layout for balanced spacing.

**i. Introduction**

Employee satisfaction isn’t just about morale — it often reflects in financial outcomes. In this analysis, I examine how job satisfaction levels are associated with average annual compensation, helping identify whether happier professionals tend to earn more.

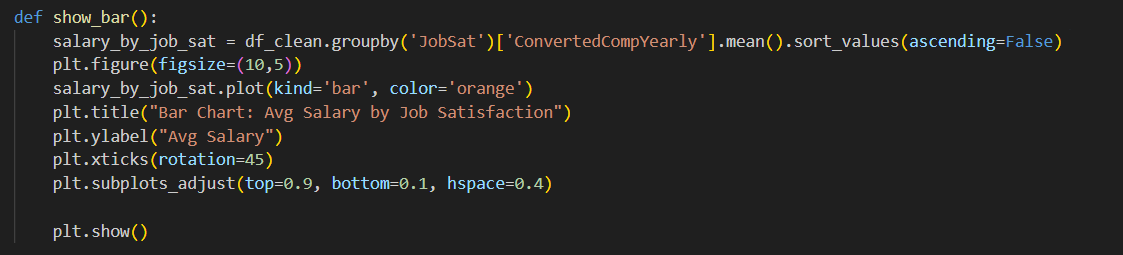
**ii. General Description**

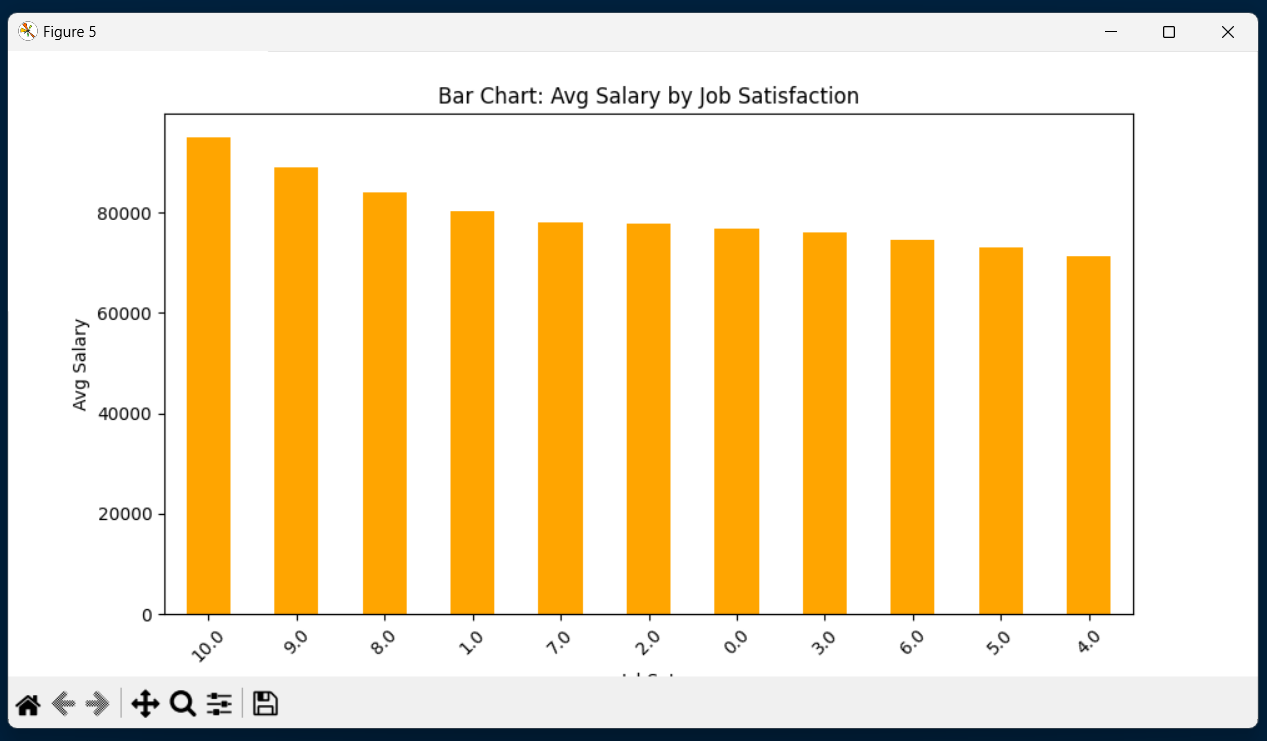
I’ll:

1. Use the **JobSat** (Job Satisfaction) and **ConvertedCompYearly** (Yearly Salary) columns from the cleaned dataset.
2. Group respondents by job satisfaction levels.
3. Calculate the mean salary for each group.
4. Plot these averages in a bar chart to visualize income distribution across satisfaction levels.

**iii. Functions & Formulas**

1. **Data Aggregation**  
   • Grouped data by **JobSat** using the groupby() function.  
   • Calculated mean salary values for each satisfaction category.
2. **Bar Chart Creation**  
   • Plotted a bar chart to display average salaries by job satisfaction level.
3. **Chart Formatting**  
   • Set figure size for clarity.  
   • Applied custom colors for distinction.  
   • Rotated X-axis labels to avoid overlap.  
   • Adjusted layout spacing for clean visual presentation.





**🔹 Objective 6: Analyze Correlation Between Yearly Compensation and Age**

I created a correlation heatmap to examine the relationship between yearly compensation (ConvertedCompYearly) and age (Age\_num) among survey respondents.

**Key Metrics**

* Variables Analyzed:
  + ConvertedCompYearly: Yearly salary (continuous numeric values).
  + Age\_num: Age of respondents (converted from categorical to numeric codes).
* Correlation Coefficient Range: -1 (strong negative) to +1 (strong positive).

**Key Insight**

* The heatmap reveals a moderate positive correlation (0.33) between age and yearly compensation.
* This suggests that, on average, older respondents tend to report higher salaries, though other factors may influence this trend.
* The diagonal values (1.0) confirm perfect self-correlation for each variable (as expected).

**Visualization**

* Heatmap with color gradients (coolwarm palette) to highlight correlation strength.
* Annotated coefficients for clarity.
* Adjusted layout (figure size, spacing) to ensure readability.

**i. Introduction**

Understanding how demographic factors like age relate to income is critical for workforce analysis. This heatmap explores whether years of experience (proxied by age) correlate with higher compensation among tech professionals.

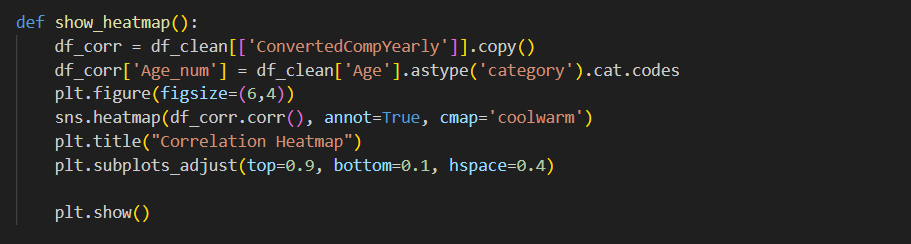
**ii. General Description**

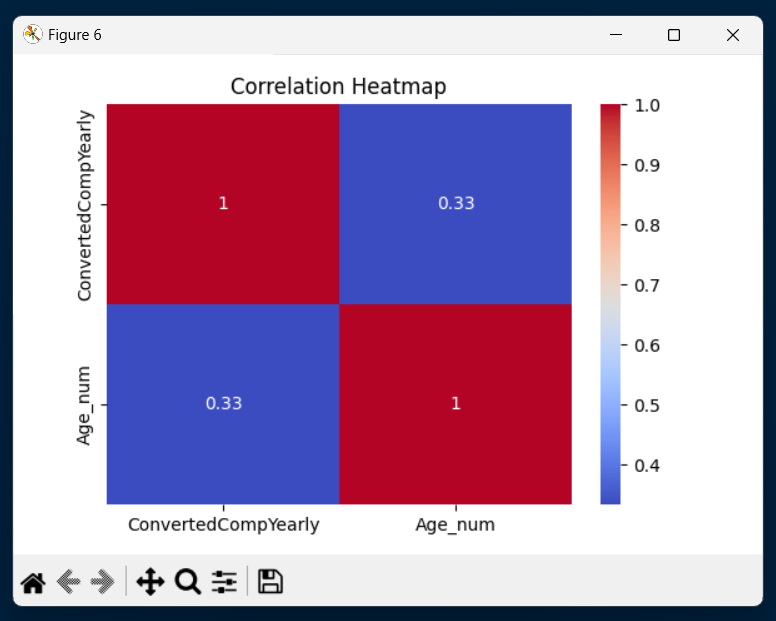
To achieve this, I:

1. Extracted ConvertedCompYearly (salary) and converted Age to numeric codes (Age\_num).
2. Computed pairwise correlations between the variables.
3. Visualized the correlation matrix using a heatmap, with annotations for precision.

**iii. Functions & Formulas**

1. Data Preparation:
   * Created a subset dataframe with df\_clean[['ConvertedCompYearly']].
   * Mapped categorical Age to numeric values using .astype('category').cat.codes.
2. Correlation Calculation:
   * Used df\_corr.corr() to generate the correlation matrix.
3. Heatmap Customization:
   * Applied sns.heatmap() with annot=True for coefficient labels.
   * Adjusted figure size (figsize=(6,4)) and layout spacing (subplots\_adjust).

****

****

**v. Summary & Key Insights**  
• Developers with higher education and experience tend to earn higher salaries, though job satisfaction plays a significant role in compensation trends.  
• Efforts to support developer growth should focus on:

* Promoting continuous learning for emerging technologies.
* Addressing disparities in compensation based on gender, region, or education.
* Encouraging workplace policies that align with developer preferences (e.g., remote work flexibility).  
  • Positive trends in technology adoption (e.g., Python, JavaScript dominance) suggest a stable ecosystem, tracking shifts in tools/languages remains critical.

**vi. Limitations**  
• Self-reported data may introduce biases (e.g., salary over/under-reporting).  
• Regional granularity is limited for smaller or underrepresented countries.  
• Temporal scope reflects only 2023 trends , longitudinal analysis would strengthen insights.

**vii. Conclusion**  
This project leverages the *2023 Stack Overflow Developer Survey* to uncover key trends in demographics, technology use, and compensation. By identifying correlations , the analysis provides actionable insights for developers, employers, and educators. Future work could explore year-over-year trends or drill deeper into underrepresented groups.