I. Project Overview

The objective of this project is to conduct a comprehensive analysis of job role salaries and trends using a specific dataset. The aim is to gain insights into employment types, levels of experience, remote work percentages, and company sizes. By performing an in-depth exploratory data analysis (EDA) and leveraging various data visualization techniques, we will investigate how these variables impact salaries for different job titles. The analysis will involve identifying salary distributions, examining trends over time, and highlighting the influence of remote work and company size on compensation. This project is intended to provide a better understanding of the current job salary landscape and offer valuable insights for job seekers, employers, and policymakers.

This project incorporates several essential factors, including work experience, seniority level, employment arrangement, job designation, salary in USD, employee location, remote work percentage, company location, and company size. These factors offer a comprehensive perspective on job salaries and patterns, allowing for an examination of how different elements such as experience, employment type, and remote work impact compensation. By harnessing these factors, the project strives to reveal valuable insights for job seekers, employers, and policymakers. It will use data visualization and thorough analysis to present findings in a clear and impactful manner.

1. **Work Year**
    - The "work_year" attribute indicates the year when the job data was documented. This attribute enables us to analyze how job roles and salaries change over time, providing insights into trends in compensation and employment patterns. By studying data from multiple years, we can gain valuable insights into salary growth, emerging job roles, and the evolution of employment types over time.

2. **Experience Level**
    - The term "experience level" denotes the level of experience necessary for each job, usually classified as Junior (JR), Mid (MI), Senior (SE), and Expert (EX).

This classification helps us comprehend the relationship between experience and salary, showing how compensation rises with greater expertise. It also enables us to pinpoint which experience levels are most sought after for particular job positions, offering valuable insights for career planning and workforce development.

3. **Employment Type**
    - The employment_type attribute denotes the employment contract type for each job, including Full-Time (FT), Part-Time (PT), Contract (CT), or Freelance (). This attribute is essential for examining the impact of various employment arrangements on salary. It allows us to assess the distribution of job types within the dataset and analyze how compensation differs across full-time, part-time, contractual, and internship positions. This provides valuable insights into labor market dynamics.

4. **Job Title**
    - The "job_title" attribute specifies the particular roles within the job, such as Data Scientist, ML Engineer, or Principal Data Scientist. This attribute is crucial for conducting role-specific salary analysis, enabling us to compare compensation across different job titles. By examining job titles, we can identify the highest and lowest paying roles, determine the demand for various positions, and analyze salary trends within specific job categories, thus gaining insight into the job market landscape.

5. **Salary**
    - The salary attribute denotes the remuneration provided for the position in the local currency. This attribute is crucial for examining the diversity in salaries among various job designations, levels of experience, and types of employment. It enables us to evaluate the overall salary spread, detect any anomalies, and comprehend the spectrum of compensation within the dataset. This analysis aids in establishing salary benchmarks and comprehending pay structures across different geographic areas and sectors.

6. **Salary Currency**
    - The salary_currency field specifies the currency in which the salary is paid, such as USD, EUR, or other local currencies. This attribute is crucial for ensuring accurate comparisons and conversions of salaries across different currencies. Standardizing salaries to a common currency (USD) facilitates cross-regional and international salary comparisons, enabling a more comprehensive analysis of compensation trends across different geographic locations.
7. **Salary in USD**
    - The "salary_in_usd" attribute denotes the salary converted to USD for the purpose of standardization. This standardized measure facilitates precise comparisons of salaries across positions with varying local currencies. It is instrumental in establishing a consistent benchmark for evaluating compensation, simplifying the process of comparing salaries across diverse countries and regions. This attribute is essential for conducting global salary analysis and gaining insights into the international job market.

8. **Employee Residence**
    - employee_residence provides the country where the employee resides. This attribute is vital for analyzing how the location of the employee affects salary and understanding the distribution of remote and non-remote jobs. It helps in examining regional salary disparities, identifying high and low-paying regions, and understanding the impact of employee residence on compensation. This analysis provides insights into geographic salary trends and workforce distribution.

9. **Remote Ratio**
    - The remote_ratio attribute indicates the percentage of time the job allows for remote work, ranging from 0 (no remote work) to 100 (fully remote). This attribute is useful for examining the impact of remote work on salary and

identifying trends in remote work preferences. It helps us understand how remote work opportunities affect compensation and the demand for remote jobs, providing insights into the evolving nature of work and its influence on salary structures.

10. Company Location

- company_location indicates the country where the company is located. This attribute allows us to analyze how the location of the company affects salary and compare compensation across different regions. It helps in understanding regional salary trends, identifying high-paying job markets, and examining the impact of company location on employment patterns. This analysis provides valuable information for companies and job seekers regarding the influence of geographic location on salaries.

11. Company Size

- The company_size attribute categorizes companies as Small (S), Medium (M), or Large (L). This attribute is essential for understanding the influence of company size on salary. It helps us analyze how compensation varies across small, medium, and large companies, identifying trends in hiring practices and salary structures within different company sizes. This analysis provides insights into the relationship between company size and compensation, helping stakeholders make informed decisions regarding employment opportunities.

By analyzing these attributes, this project aims to gain a holistic understanding of job role salaries and trends. By leveraging these insights, stakeholders can make informed decisions regarding job roles, salary negotiations, and workforce strategies.

II. Libraries and Data Handling

Libraries Used

The following libraries were used in the project.

1. **Pandas**
   - Pandas is a fast, powerful, and flexible open-source data analysis and manipulation library built on top of the Python programming language. It offers data structures like DataFrame and Series that are designed for efficient data manipulation and analysis. Pandas is widely used for tasks such as data cleaning, transformation, and analysis, making it a fundamental tool in the data science toolkit.

2. **Seaborn**
   - Seaborn is a Python visualization library based on matplotlib that provides a high-level interface for creating attractive and informative statistical graphics. It is particularly useful for visualizing complex datasets and statistical models. Seaborn simplifies the process of creating visually appealing plots by providing built-in themes and color palettes, making it easy to customize plots to suit different needs.

3. **NumPy**
   - NumPy is a core library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. NumPy is widely used for tasks such as numerical computing, linear algebra, and random number generation, making it essential for scientific computing and data analysis.

4. **Matplotlib**
   - Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. The pyplot module provides a MATLAB-

like interface for creating plots and visualizations. Matplotlib is highly customizable, allowing users to create a wide variety of plots, including line plots, bar plots, histograms, scatter plots, and more. It is widely used for creating publication-quality figures and visualizations in Python.

Data Loading

Data is loaded from a CSV file into a DataFrame

- The pd.read_csv() function is used to read the CSV file into a DataFrame, which is a tabular data structure similar to a spreadsheet. Each row in the DataFrame represents a record (or observation) and each column represents a variable (or feature).

Data Cleaning and Preprocessing

In the project, preprocessing techniques were used to prepare the dataset for analysis.

- **Handling Missing Values**
  - This technique involves identifying and dealing with missing values in the dataset. Missing values can be filled in with a default value (e.g., mean, median, mode), removed entirely, or imputed using more advanced techniques.

- **Lowercasing Column Names**
  - Lowercasing column names standardizes the naming convention, making it easier to work with the dataset, especially when referencing columns in code.

- **One-Hot Encoding**
  - One-hot encoding is used to convert categorical variables into a numerical format. It creates dummy variables for each category of a categorical variable, which can then be used in machine learning models that require numerical inputs.

- **Binning**
  - Binning is the process of grouping numerical data into bins or intervals. It can be useful for reducing the impact of outliers, converting numerical data into categorical data, or simplifying complex datasets. In this case, it was used to categorize salary values into predefined bins.

In this project, several preprocessing techniques were applied to prepare the dataset for analysis. Missing values were checked and addressed, ensuring dataset completeness. Column names were standardized to lowercase for consistency. Categorical variables were converted into numerical format using one-hot encoding, enabling machine learning models' use. Salary values were binned into categories, simplifying analysis. These techniques ensured the dataset was clean, consistent, and ready for analysis.

III. Data Analysis Techniques

Exploratory Data Analysis

Data analysis is a critical step in extracting meaningful insights from datasets. In this project, several data analysis techniques were employed to explore and understand the salary trends within the dataset.

- **Grouping and Aggregation**
  - The groupby() method was used to group the data by different attributes, such as 'job_title', 'experience_level', 'employment_type', 'remote_ratio', and 'company_size'. This allows for the aggregation of salary data within each group.
  - The ['salary_in_usd'].mean() part calculated the mean salary for each group, providing insights into the average salary for different job titles, experience levels, employment types, remote work ratios, and company sizes.

- **Column Renaming**
  - The columns = ['Job_Title', 'Average_Salary_in_USD'] statement renamed the columns of the role_salary DataFrame to provide more descriptive column names, improving readability and understanding of the data.

- **Unstacking**
  - The unstack() method was used to pivot the 'experience_level', 'employment_type', 'remote_ratio', and 'company_size' columns from rows into columns. This reshaping of the data makes it easier to compare salary data across different categories and analyze trends.

- **Data Summarization**
  - The describe() method provided summary statistics for the numerical columns in the dataset. This includes count, mean, standard deviation, minimum, maximum, and quartile values. These statistics help in

understanding the distribution and characteristics of the salary data, providing valuable insights into the dataset's overall profile.

- **Groupby Aggregation**
  - Groupby aggregation involves splitting the data into groups based on a specific attribute, in this case, 'region,' and calculating the average salary in USD for each group. This technique helps understand salary variations across different regions, providing insights into regional salary trends and comparisons.

These techniques together enable a comprehensive analysis of the dataset, leading to valuable insights into salary trends and distributions across various job attributes.

Data Visualization

Data visualization is a crucial aspect of data analysis, providing insights into complex datasets through graphical representations. These techniques help in understanding the distribution of salary values, comparing average salaries across different categories, exploring relationships between variables, and identifying patterns and correlations in the data.

- **Histogram**
  - The histogram was used to visualize the distribution of salaries in USD. By dividing the data into 5 bins and including a KDE curve, the histogram provides insights into the central tendency and spread of salary values. The KDE curve helps in understanding the shape of the distribution and identifying any underlying patterns or clusters in the data.

- **Bar Plots**
  - Bar plots were employed to compare average salaries across different categories. For example, one bar plot shows the average salary by job title

and experience level, providing a visual comparison of salary levels for different roles and experience levels. Similarly, other bar plots compare average salaries based on employment type, company size, and remote work ratio, helping in understanding how these factors influence salary levels.

- **Pair Plots**
  - The pair plot created a grid of scatterplots for numerical variables ('salary_in_usd', 'experience_level', 'remote_ratio', 'company_size'). This visualization technique allows for the exploration of relationships between these variables, such as how salary relates to experience level or remote work ratio. It also provides insights into the distributions of these variables, highlighting any outliers or patterns in the data.

- **FacetGrid with Barplot**
  - This technique involved creating a grid of plots based on a categorical variable ('remote_ratio') and using bar plots to display the average salary for each job title in each grid cell. By ordering the job titles and adjusting the x-axis labels for readability, this visualization method allows for a clear comparison of average salaries across different remote work ratios and job titles.

- **Heatmap**
  - The heatmap was used to visualize the correlation between 'remote_ratio' and 'salary_in_usd'. By annotating the heatmap with correlation values and using a color scale to represent the strength of the correlation, this visualization technique helps in understanding the relationship between remote work ratio and salary levels.

- **Line Plots**
  - The visualization technique used is a line plot. It shows the trend of salaries over the years based on the dataset. Each point on the line represents the

average salary for a specific year. This visualization helps in understanding how salaries have changed over time, providing insights into the overall salary trends within the dataset. The marker 'o' is used to highlight each data point, making it easier to visualize individual salary data points for each year.

A range of data visualization techniques was utilized to analyze salary trends within a dataset. Histograms were used to visualize salary distributions, bar plots for comparing average salaries across different categories, and pair plots for exploring relationships between variables. Facet grids with bar plots were employed to compare salaries based on remote work ratios, while a heatmap was used to visualize the correlation between remote work ratios and salaries. These techniques collectively provide a comprehensive analysis of the dataset, offering insights into the factors influencing salary levels across various job attributes.

IV. Key Findings

Average Salary: Analyzing the salary based on different attributes provides valuable insights into how various factors influence salary levels.

- **Average Salary by Job Title and Experience Level**
  - The graph indicates that for almost all job titles, the highest segment (red, representing Senior - SE) contributes the most to the average salary, followed by the green segment (Mid - MI), then the orange segment (Expert - EX), and finally the blue segment (Entry - EN) contributes the least to the average salary. This suggests that, on average, employees with senior-level experience (SE) tend to earn more across various job titles compared to those with mid-level (MI), expert (EX), or entry-level (EN) experience.

- **Average Salary by Job Title and Employee Type**
  - The analysis of average salaries based on job title and employment type shows that Full-Time (FT) employees generally have the highest average salaries across most job titles. Contract (CT) employees, while lower than Full-Time, still have notable salaries, especially for roles like Principal Data Analyst. Freelance (FL) positions tend to offer lower average salaries compared to Full-Time and Contract roles. Part-Time (PT) roles consistently have the lowest average salaries among the employment types analyzed.

- **Average Salary by Job Title and Company Size**
  - The analysis of average salaries based on job title and company size reveals that larger companies (L) generally offer the highest average salaries across all job titles. Medium-sized companies (M) have average salaries that are comparable to those of large companies, indicating competitive compensation packages. Small companies (S) consistently offer the lowest average salaries among the company sizes analyzed, suggesting that salary levels may be influenced by the size and financial resources of the company.

- **Average Salary by Job Title and Remote Ratio**
  - The analysis of average salaries based on job title and remote work ratio indicates that jobs with no remote work (0) have the highest average salaries across all job titles. Surprisingly, jobs with a fully remote work arrangement (100) have average salaries comparable to those with no remote work, suggesting that remote work does not necessarily lead to lower salaries. Jobs with a partial remote work arrangement (50) consistently have the lowest average salaries among the different remote work ratios analyzed.

- **Correlation of Salary in USD and Remote Ratio**
  - The correlation heatmap reveals the relationship between different variables in the dataset. The diagonal line represents the correlation of each variable with itself, which is always 1. The off-diagonal values show the correlation between pairs of variables. In this case, the correlation between 'remote_ratio' and itself is 1, as expected. The correlation between 'remote_ratio' and 'salary_in_usd' is approximately -0.06, indicating a very weak negative correlation. This suggests that there is a slight tendency for salaries to decrease slightly as the remote work ratio increases, although the correlation is not strong.

The analysis of average salaries based on different attributes reveals several key insights. Across various job titles, employees with senior-level experience tend to earn more, followed by mid-level, expert, and entry-level experience. Full-Time positions generally offer the highest average salaries, while Contract roles also offer competitive salaries, especially for certain job titles. Larger companies tend to offer higher average salaries compared to medium and small companies, indicating a correlation between company size and salary levels. Surprisingly, jobs with no remote work and fully remote work arrangements have the highest average salaries, with partially remote jobs offering the lowest. The correlation heatmap shows a very weak negative correlation between

salary and remote work ratio, suggesting that salaries may decrease slightly as the remote work ratio increases, but the correlation is not strong. Overall, these findings provide valuable insights into the factors influencing salary levels across different job titles and employment types.

V. Advanced Analysis

Geographical Analysis: By examining how data points are distributed across different regions, this technique provides valuable insights into geographical patterns, trends, and relationships.

- **Categorization into Continents**
  - The purpose of this is to categorize company locations into broader geographical regions, facilitating regional analysis and insights. Each company location is assigned to a specific region based on its country code. This categorization allows for a more structured and meaningful analysis of the data by grouping companies into regions such as Africa, Asia, Europe, North America, South America, and Oceania.

- **Number of Employees by Region**
  - The categorization of company locations into broader geographical regions reveals a significant variation in the number of employees across these regions. North America stands out with a substantial number of 3,144 employees, indicating a high concentration of the workforce in this region. This suggests that North America is a major hub for the companies in the dataset, potentially reflecting the presence of many large and prominent companies.
  - Europe follows with 461 employees, indicating a notable but smaller concentration compared to North America. Asia, with 89 employees, shows a moderate presence, suggesting some level of activity and employment in this region. South America, Oceania, and Africa have relatively small numbers of employees, with 24, 18, and 14 respectively, indicating limited workforce representation in these regions.

- **Average Salary by Region**
  - The analysis of average salaries by region reveals significant disparities across different geographical areas. North America leads with a remarkably

high average salary of $151,000, indicating that this region offers the most lucrative compensation for employees. This suggests that companies in North America, likely including numerous high-profile firms, provide significantly higher wages compared to other regions.

- Europe and Oceania follow with average salaries of $75,400 and $74,100, respectively. These figures indicate that both regions offer competitive salaries, although they are only about half of what is offered in North America. The similar salary levels between Europe and Oceania suggest a comparable compensation landscape in these regions.
- Africa and Asia have average salaries of $43,300 and $42,000, respectively. These figures are significantly lower than those in North America, Europe, and Oceania, highlighting a considerable gap in earnings. South America has the lowest average salary at $41,300, underscoring the regional disparities in compensation.

Temporal Trends: Understanding the temporal trends in average salaries across different regions provides valuable insights into how compensation patterns have evolved over time.

- **Average Salary Trends Over the Years by Region**
  - 2020:
    - Africa: With an average salary of $10,000, Africa had the lowest average salary among all regions, reflecting a significant gap in compensation.
    - Asia: Employees earned an average of $51,000, showing a relatively higher salary compared to Africa but still lagging behind other regions.
    - Europe: The average salary was $62,400, indicating a competitive compensation level within the region.
    - North America: With an average of $130,100, North America offered the highest salaries, emphasizing its position as a lucrative region for employees.

- Oceania: The average salary was $125,000, close to North America's, suggesting a strong salary landscape.
- South America: No data was recorded for this year.

- 2021:
  - Africa: The average salary increased to $29,800, showing a substantial rise from 2020.
  - Asia: There was a decrease to $44,500, indicating a slight decline in average salaries.
  - Europe: The average salary increased to $69,400, showing a positive trend.
  - North America: With $131,800, the salaries remained high, with a slight increase from the previous year.
  - Oceania: The average salary dropped significantly to $38,800, indicating a notable decrease.
  - South America: The average salary was recorded at $23,500, providing baseline data for the region.
- 2022:
  - Africa: The average salary saw a significant increase to $83,200, marking a notable improvement.
  - Asia: The average salary slightly increased to $45,800, showing stabilization.
  - Europe: The average salary remained stable at $69,000.
  - North America: The average salary rose to $146,800, maintaining its position as the highest-paying region.
  - Oceania: The average salary increased to $84,900, showing recovery from the previous year's dip.
  - South America: The average salary increased to $46,300, showing improvement.
- 2023:

- Africa: The average salary dropped to $52,400, indicating a decline from the previous year.
- Asia: The average salary further decreased to $27,500, marking a significant decline.
- Europe: The average salary increased to $90,500, showing a strong upward trend.
- North America: The average salary continued to rise to $156,400, reinforcing its high-salary trend.
- Oceania: The average salary dropped to $60,000, showing volatility.
- South America: The average salary decreased slightly to $44,200.

VI. Visual Insights

- **Salary Distribution**
  - **Histogram:** This histogram displays the salary distribution in USD, showing the frequency of salaries across different ranges. Most salaries cluster between $50,000 and $150,000, with the highest frequency around the $100,000 mark. As the salary increases beyond $150,000, the frequency of occurrence drops significantly, tapering off sharply after $200,000. The distribution appears to follow a right-skewed pattern, indicating that higher salaries are less common, with only a few instances exceeding $300,000. This suggests that the majority of salaries fall within the lower to mid-range, with fewer high-salary outliers.

  - **Implications:** The histogram shows most salaries clustered between $50,000 and $150,000, reflecting the dataset where roles like Data Scientist and Applied Scientist fall within or exceed this range. High-salary outliers, such as a Principal Data Scientist earning $85,847 (converted to USD) and Applied Scientists earning up to $222,200, match the histogram's less frequent, higher salaries. Larger companies tend to offer higher salaries, and the dataset's senior roles command substantial pay, consistent with the histogram's pattern. Additionally, remote positions also appear within the mid-range salary peak, illustrating a diverse salary distribution influenced by company size, role seniority, and remote work options.

- **Average Salary Based on Job Title and Experience Level**
  - **Bar Plot:** A bar plot is used in this context to visually represent the average salary distribution across various job titles and experience levels. This type of plot is particularly effective because it allows for clear comparison of salaries across different job titles and experience levels, facilitating an easy understanding of complex data. The use of stacked bars helps to show the cumulative effect of different experience levels on the salary for each job title, making it easier to compare the total impact and relative contributions. Additionally, the color coding of experience levels (EN: Entry, EX:

Executive, MI: Mid, SE: Senior) provides a quick visual differentiation, aiding in the analysis of how experience impacts salary.

- **Implications:** This bar plot illustrates the average salary in USD for different job titles segmented by experience levels. Senior (SE) roles (red bars) typically command the highest salaries across most job titles, indicating that experience significantly boosts earning potential, while executive (EX) roles (orange bars) also show high salaries, often close to or sometimes surpassing senior roles. For job titles such as "Principal Data Scientist" and "Applied Scientist," there are substantial salary differences based on experience, with senior roles earning significantly more. Meanwhile, job titles like "Data Scientist" and "ML Engineer" show a steady salary increase with experience, reflecting a clear career progression.

- **Average Salary Based on Job Title and Employment Type**
  - **Bar Plot:** A bar plot is used here to visually represent the average salary distribution across different job titles and employment types. This type of plot is effective because it allows for a clear comparison of salaries across different categories, facilitating an easy understanding of complex data. The use of stacked bars helps show the cumulative effect of different employment types on the salary for each job title, making it easier to compare the total impact and relative contributions. Additionally, color coding of employment types (CT: Contract, FL: Freelance, FT: Full-time, PT: Part-time) provides a quick visual differentiation, aiding in the analysis of how employment type impacts salary.

  - **Implications:** This bar plot illustrates the average salary in USD for different job titles segmented by employment type. Full-time (FT) roles (green bars) typically command the highest salaries across most job titles, indicating that permanent, full-time positions offer more compensation. Contract (CT) and freelance (FL) roles (blue and orange bars, respectively) also show significant salaries in some job titles but generally fall below full-

time positions. Part-time (PT) roles (red bars) generally have the lowest salaries, reflecting the reduced hours and commitments associated with these roles.

- **Average Salary Based on Job Title and Company Size**
  - **Bar Plot:** A bar plot is used here to visually represent the average salary distribution across different job titles and company sizes. This type of plot is effective because it allows for a clear comparison of salaries across various categories, making complex data more understandable. The use of stacked bars helps show the cumulative effect of different company sizes on the salary for each job title, making it easier to compare the total impact and relative contributions. Additionally, color coding of company sizes (L: Large, M: Medium, S: Small) provides quick visual differentiation, aiding in the analysis of how company size impacts salary.

  - **Implications:** This bar plot illustrates the average salary in USD for different job titles segmented by company size. Larger companies (L, represented by blue bars) typically offer higher salaries across most job titles, suggesting that larger organizations have more resources to compensate their employees generously. Medium (M, orange bars) and small (S, green bars) companies generally offer lower salaries compared to large companies.

- **Average Salary Based on Job Title and Remote Ratio**
  - **Bar Plot:** A bar plot is used here to visually represent the average salary distribution across different job titles and remote work ratios. This type of plot is effective because it allows for a clear comparison of salaries across various categories, facilitating an easy understanding of complex data. The use of stacked bars helps show the cumulative effect of different remote work ratios on the salary for each job title, making it easier to compare the total impact and relative contributions. Additionally, color coding of remote ratios (0: No remote work, 50: Partial remote work, 100: Fully remote work)

provides quick visual differentiation, aiding in the analysis of how remote work impacts salary.

- **Implications:** This bar plot illustrates the average salary in USD for different job titles segmented by remote work ratios. Fully remote positions (100% remote, represented by green bars) typically offer higher salaries across many job titles, suggesting that companies may compensate more for fully remote roles to attract and retain talent. Partial remote positions (50% remote, orange bars) and non-remote positions (0% remote, blue bars) generally offer lower salaries compared to fully remote positions.

- **Average Salary Based on Job Title and Remote Ratio**
  - **Facetgrid:** The FacetGrid with bar plots is used here to visually represent the average salary distribution across different job titles, segmented by remote work ratios (0, 50, and 100%). This type of plot is effective because it allows for a detailed comparison of salaries across various categories while maintaining clarity by separating data based on remote work ratios. Using separate plots for each remote ratio category helps to isolate the effects of remote work on salaries. Additionally, ordering job titles and rotating the labels ensures readability despite the large number of categories.

  - **Implications:** These bar plots illustrate the average salary in USD for different job titles, segmented by remote work ratios (0%, 50%, and 100%). Each plot highlights how salaries vary with the extent of remote work allowed. Generally, fully remote positions (100% remote) often show higher salaries compared to partial (50%) and non-remote (0%) positions, suggesting companies might offer premium pay for fully remote roles to attract talent.

- **Correlation of Salary in USD and Remote Ratio**
  - **Correlation Heatmap:** A correlation heatmap is used here to visually represent the relationship between remote work ratio and salary in USD. This type of plot is effective because it provides a clear, color-coded visualization of the correlation coefficients, making it easy to understand the strength and direction of relationships between variables. The use of annotations adds numerical values to the visual representation, providing precise information about the degree of correlation.

  - **Implications:** The heatmap illustrates the correlation between the remote work ratio and salary in USD. The correlation coefficient between remote ratio and salary is -0.06, indicating a very weak negative correlation. This suggests that there is a slight tendency for salaries to decrease as the remote work ratio increases, but the effect is minimal. The correlation coefficient of 1.00 along the diagonal indicates perfect positive correlation, which is expected as a variable is always perfectly correlated with itself. Overall, the heatmap shows that remote work ratio has a negligible impact on salary, implying that other factors might play a more significant role in determining compensation.

- **Distribution of Employees by Residence Country**
  - **Count Plot:** A count plot is used here to visually represent the distribution of employees by their country of residence. This type of plot is effective because it provides a clear view of the frequency distribution of a categorical variable, allowing easy comparison across different categories. The order of countries is based on the frequency of occurrences, making it straightforward to identify the countries with the highest number of employees.
  - **Implications:** The count plot illustrates the distribution of employees by their country of residence. The x-axis represents different countries, and the y-axis represents the number of employees. The plot shows a highly skewed distribution, with the majority of employees concentrated in a few

countries. One country stands out with an exceptionally high number of employees, indicating it is the primary residence for the majority of the workforce.

- **Number of Employees per Region**
  - **Bar Plot:** A bar plot is used here to visually represent the number of employees by region. This type of plot is effective because it clearly shows the frequency distribution across different regions, allowing easy comparison of the number of employees in each region. The use of color coding for different regions adds an additional layer of clarity and helps in quickly distinguishing between them. Plotly Express provides interactive elements that enhance the user experience by allowing them to hover over the bars for more details.

  - **Implications:** The bar plot illustrates the distribution of employees across different regions. The x-axis represents various regions, while the y-axis represents the number of employees. The plot shows a significant concentration of employees in North America, with over 3,000 employees, making it the region with the highest number of employees by a large margin. Europe follows as the second most represented region, albeit with significantly fewer employees than North America.

- **Average Salary per Region**
  - **Bar Plot:** A bar plot is used here to visually represent the average salary distribution across different regions. This type of plot is effective because it provides a clear visual comparison of average salaries across various regions, making complex data more understandable. The use of color coding for different regions enhances clarity and helps in quickly distinguishing between them. Plotly Express provides interactive elements that enhance the user experience by allowing them to hover over the bars for more details.

- **Implications:** The bar plot illustrates the average salary in USD for different regions. The x-axis represents various regions, while the y-axis represents the average salary. The plot shows that North America has the highest average salary, significantly surpassing other regions. Europe follows with a notably high average salary as well.

- **Average Salary Trends Over the Years by Region**
  - **Line Plot:** A line plot is used here to visually represent the trends in average salary over the years across different regions. This type of plot is effective because it shows how salaries change over time, providing a clear visual representation of trends and patterns. The use of different colors for each region enhances clarity and helps in quickly distinguishing between them. Plotly Express provides interactive elements that enhance the user experience by allowing them to hover over the lines for more details and see the data points clearly.

  - **Implications:** The line plot illustrates the average salary trends in USD over the years (2020 to 2023) across different regions. The x-axis represents the years, while the y-axis represents the average salary. Each line represents a different region, allowing for easy comparison of salary trends across regions.
    - North America consistently shows the highest average salaries, with a noticeable upward trend over the years, indicating a growth in compensation in this region.
    - Europe shows a steady increase in average salaries, reflecting a positive trend in compensation.
    - Oceania experienced a significant increase in average salaries in 2022 but saw a decline in 2023.
    - Africa shows a consistent upward trend, starting from a lower base compared to other regions but catching up gradually.
    - Asia shows a slight downward trend, indicating a decrease in average salaries over the period.

- South America and Other regions have relatively stable trends, with minor fluctuations in average salaries.

VII. Conclusion

This project aims to conduct a comprehensive analysis of job role salaries and trends using a specific dataset, focusing on employment types, experience levels, remote work percentages, and company sizes. By leveraging exploratory data analysis (EDA) and various data visualization techniques, the project investigates how these variables impact salaries for different job titles, providing valuable insights for job seekers, employers, and policymakers. Key factors considered include work experience, seniority level, employment arrangement, job designation, salary in USD, employee location, remote work percentage, company location, and company size. The objective is to present a clear and impactful understanding of the current job salary landscape.

The analysis reveals several critical insights: senior-level positions generally command higher salaries across job titles, while full-time roles offer the most compensation compared to contract, freelance, and part-time roles. Larger companies tend to provide higher salaries than medium and small companies. Interestingly, jobs with no remote work or fully remote arrangements often have higher salaries compared to those with partial remote work. The correlation heatmap indicates a very weak negative correlation between remote work ratio and salary, suggesting minimal impact. Additionally, geographic analysis shows significant salary disparities, with North America leading in average salaries, followed by Europe and Oceania.

Data visualization techniques such as histograms, bar plots, pair plots, facet grids, heatmaps, and line plots were used to explore salary trends and distributions. The visualizations highlight the concentration of salaries within the $50,000 to $150,000 range and illustrate the influence of experience level, employment type, company size, and remote work ratio on compensation. Furthermore, the geographic analysis provides insights into regional salary trends and workforce distribution, with North America having the highest number of employees and the highest average salaries. These findings offer a comprehensive view of the factors influencing salary levels, aiding in informed decision-making for job seekers, employers, and policymakers.