# Project Proposal: Global AI & Data Science Salaries

**1. Introduction & Executive Summary**

This proposal outlines a data analysis project aimed at exploring and deriving insights from a comprehensive dataset of salaries in the Data Science and AI industry. The dataset, `salaries.json`, contains over 136,000 records detailing key attributes such as job title, experience level, company location, and salary. The primary objective is to uncover and visualize trends, patterns, and key factors that influence compensation in this rapidly evolving field. The findings will provide a valuable resource for industry professionals, job seekers, and companies looking to benchmark their compensation strategies.

**2. Research Objectives**

The analysis will focus on answering the following key questions:

- **Salary Trends Over Time:** How have average salaries for various roles changed from year to year? Are salaries increasing, decreasing, or remaining stable?
- **Experience vs. Compensation:** What is the relationship between `experience_level` and salary? How significant is the jump in compensation from an entry-level position to an executive role?
- **Impact of Remote Work:** How does the `remote_ratio` and `employment_type` influence a position's salary? Is there a noticeable salary difference between on-site, hybrid, and fully remote roles?
- **Company Size as a Factor:** Is there a correlation between `company_size` (S, M, or L) and the salaries offered for similar positions?
- **Geographic and Currency-based Differences:** How do salaries vary based on the `employee_residence` and `company_location`? What are the top-paying countries for Data Science and AI roles?
- **Job Title Analysis:** Which `job_title`s consistently command the highest salaries, and which have seen the most significant growth in compensation over time?

**3. Methodology**

The project will follow a structured data analysis pipeline to ensure accurate and robust findings.

- **Data Collection & Cleaning:** The provided JSON file will be loaded into a suitable environment (e.g., Python with pandas). The data will be inspected for quality, and any missing or inconsistent values will be addressed. The `salary` field will be

cross-referenced with `salary_in_usd` to ensure accuracy, and the `salary_in_usd` field will be used for all comparative analyses to maintain a consistent currency base.

- **Exploratory Data Analysis (EDA):** Initial statistical summaries will be generated to understand the distribution of key variables. This phase will utilize visualizations such as histograms, box plots, and scatter plots to identify outliers, patterns, and potential correlations between variables (e.g., `remote_ratio` vs. `salary_in_usd`).
- **Hypothesis Testing:** Statistical tests (e.g., t-tests, ANOVA) will be conducted to formally test the hypotheses generated during the EDA phase. For instance, a test could be performed to determine if the mean salaries for different `experience_level` groups are statistically significant.
- **Predictive Modeling (Optional but Recommended):** A regression model (e.g., Linear Regression, Random Forest Regressor) could be built to predict a role's salary based on a combination of features like `experience_level`, `job_title`, `company_location`, and `company_size`. This would provide a powerful tool for salary forecasting.
- **Visualization & Reporting:** All key findings will be presented in a comprehensive report, supported by clear and insightful data visualizations (e.g., bar charts of average salary by job title, heatmaps of salary by country, etc.). The final report will summarize the project's journey, methodology, key findings, and recommendations.

**4. Deliverables**

- A documented, clean, and pre-processed dataset.
- A series of high-quality data visualizations covering the research objectives.
- A final report in a presentable format (e.g., PDF or a Jupyter Notebook), including code and commentary, that outlines the project's findings and provides actionable insights.

**5. Timeline**

- **Week 1:** Data Preparation and EDA
- **Week 2:** In-depth Analysis, Hypothesis Testing, and Model Building
- **Week 3:** Visualization, Final Report Writing, and Review

**Dataset:**
https://www.kaggle.com/datasets/adilshamim8/salaries-for-data-science-jobs?select=salaries.json