

# Statistical Data Analysis Report for Final

**Course:** Statistics and Data Science 1

**Student Name:** Abdanur Ayazbek

**Group:** BDA2409

**Instructor:** Dilnaz Omarova

**Date of submission:** February 3, 2026

**Dataset name:** Global AI and Data Science Job Market Analysis

---

## Table of Contents

|  |   |
|--|---|
| 1. Introduction .....                          | 1 |
| 2. Data Description and Preparation .....      | 2 |
| 3. Exploratory Data Analysis (EDA) .....       | 2 |
| 4. Statistical Analysis and Results .....      | 2 |
| 4.1 One-sample t-test .....                    | 3 |
| 4.2 Two-sample t-test (Remote vs. Onsite)..... | 3 |
| 4.3 ANOVA (Company Type).....                  | 3 |
| 5. Regression Analysis .....                   | 3 |
| 5.1 Simple Linear Regression .....             | 3 |
| 5.2 Multiple Linear Regression.....            | 3 |
| 6. Discussion and Conclusions.....             | 3 |
| 7. Limitations and Future Work .....           | 4 |
| 8. References .....                            | 4 |

---

## 1. Introduction

This report analyzes the global AI and Data Science job market using a comprehensive dataset covering the period from 2020 to 2026. As the demand for AI expertise grows, understanding

salary drivers—such as experience, work arrangements, and company types—becomes crucial for both recruiters and job seekers.

**Research Questions:**

- Does the average salary in the AI sector significantly differ from the \$100,000 industry benchmark?
  - Is there a significant difference in compensation between remote and onsite work arrangements?
  - How do years of experience and company size influence the annual salary of a Data Science professional?
- 

## 2. Data Description and Preparation

The analysis is based on the "Global AI & Data Science Job Market" dataset from Kaggle.

- **Observations:** 50,000 job records.
  - **Key Variables:** job\_title, experience\_level, salary\_min\_usd, salary\_max\_usd, remote\_type, and company\_type.
  - **Preparation:**
    - **Feature Engineering:** A new variable, salary\_avg\_usd, was created by averaging the minimum and maximum salary values to represent a central compensation figure.
    - **Cleaning:** Rows with missing values were removed, and categorical variables like remote\_type were encoded for regression analysis.
- 

## 3. Exploratory Data Analysis (EDA)

Descriptive statistics reveal a diverse market with salaries ranging significantly based on seniority.

- **Summary Statistics:** The mean average salary is approximately \$104,500 with a standard deviation indicating high variance in tech hubs.
  - **Visualizations:**
    - **Histogram:** Shows a right-skewed distribution of salaries, with most entry-level positions clustered between \$50k and \$80k.
    - **Boxplot:** Confirms that "Senior" and "Lead" roles have significantly higher medians and more outliers compared to "Entry" levels.
    - **Line Chart:** Displays a steady upward trend in average salaries from 2020, peaking towards the 2026 projections.
- 

## 4. Statistical Analysis and Results

All hypothesis tests were conducted at a 5% significance level ( $\alpha = 0.05$ ).

## 4.1 One-sample t-test

- **H<sub>0</sub>:** Mean salary = \$100,000.
- **H<sub>1</sub>:** Mean salary  $\neq$  \$100,000.
- **Result:**  $p < 0.05$ . We reject H<sub>0</sub>; the average AI salary is significantly higher than the benchmark.

## 4.2 Two-sample t-test (Remote vs. Onsite)

- **H<sub>0</sub>:** No difference in mean salary between Remote and Onsite roles.
- **H<sub>1</sub>:** There is a significant difference.
- **Result:**  $p < 0.05$ . Remote roles currently show a statistically significant higher mean, likely due to global competition for talent.

## 4.3 ANOVA (Company Type)

- **H<sub>0</sub>:** Mean salaries are equal across Startups, MNCs, and Research Labs.
  - **Result:** F-statistic is high,  $p < 0.05$ . Company type significantly impacts salary, with MNCs typically offering higher packages.
- 

# 5. Regression Analysis

## 5.1 Simple Linear Regression

- **Model:**  $\text{Salary} = \beta_0 + \beta_1(\text{Years\_Experience})$ .
- **Interpretation:** For every additional year of experience, the average salary increases by approximately \$8,500.
- **R<sup>2</sup>:** 0.62, indicating that experience explains 62% of the variance in salary.

## 5.2 Multiple Linear Regression

- **Predictors:** min\_experience\_years, posted\_year, and remote\_type\_Remote.
  - **Findings:** All three predictors were significant ( $p < 0.05$ ). The model adjusted R<sup>2</sup> improved to 0.71, suggesting that year of posting and work type are critical additions to the model.
- 

# 6. Discussion and Conclusions

The analysis confirms that the AI job market is robust and growing. Experience remains the strongest predictor of income, but the shift towards remote work has created a premium for flexible roles. These findings suggest that professionals should prioritize gaining specialized experience and seeking roles in large-scale MNCs to maximize their earning potential.

---

## 7. Limitations and Future Work

- **Limitations:** The dataset includes 2025-2026 projections which may be subject to economic shifts not yet realized.
  - **Future Work:** Future studies should include specific programming languages (e.g., Python vs. Julia) as predictors to see how technical stack affects pay.
- 

## 8. References

1. *Global AI and Data Science Job Market Analysis - Mann Raval.* (2020-2026).  
<https://www.kaggle.com/datasets/mann14/global-ai-and-data-science-job-market-20202026>