****

***Cairo University – Faculty of Engineering***

***MTHS204 Advanced Probability and Statistics***

***Summer 2025***

**AI Impact on Jobs**

****

**Submitted to: Dr Rawhy & Dr Sroor**

|  |  |
| --- | --- |
| **Mazen Ahmed** | **1220269** |
| **Aya Reda** | **1220038** |
| **Youssef Amr** | **1220319** |
| **Anas Sayed** | **4230144** |
| **Omar Mohamed** | **1230299** |

Table of Contents

[Abstract: 4](#_Toc205348408)

[Problem Definition: 4](#_Toc205348409)

[The core problems addressed in this analysis are: 5](#_Toc205348410)

[Relevance and Significance: 5](#_Toc205348411)

[Complexities Involved: 5](#_Toc205348412)

[Methodologies used: 5](#_Toc205348413)

[Flowchart: 6](#_Toc205348414)

[Data Description 6](#_Toc205348415)

[Dataset Features Overview: 6](#_Toc205348416)

[Relevance to Research Questions: 7](#_Toc205348417)

[Statistical Questions / Analysis: - 7](#_Toc205348418)

[Filtering the Data: 8](#_Toc205348419)

[Creating the contingency table: 8](#_Toc205348420)

[Observations: 9](#_Toc205348421)

[Computing the Expected table: 9](#_Toc205348422)

[Computing the Chi-Square Statistic (Test): 10](#_Toc205348423)

[Finding the Chi-Square Critical using the X2 table: 11](#_Toc205348424)

[Comparing the Results: 11](#_Toc205348425)

[Conclusion: 11](#_Toc205348426)

[Filtering the Data: 12](#_Toc205348427)

[Defining the groups: 12](#_Toc205348428)

[Sample Sizes: 12](#_Toc205348429)

[Computing the T-Test: 13](#_Toc205348430)

[Comparing the Results: 13](#_Toc205348431)

[Conclusion: 13](#_Toc205348432)

[Data Preparation: 15](#_Toc205348433)

[Computing the Anova Statistic: 15](#_Toc205348434)

[Degrees of Freedom: 15](#_Toc205348435)

[Mean Squares: 15](#_Toc205348436)

[F-statistic: 16](#_Toc205348437)

[p-value: 16](#_Toc205348438)

[Sum of Squares: 16](#_Toc205348439)

[Anova Summary Table: 17](#_Toc205348440)

[Decision and Conclusion: 17](#_Toc205348441)

[Suggestions: 18](#_Toc205348442)

[Data Preparation: 20](#_Toc205348443)

[Manual Computations: 20](#_Toc205348444)

[Conclusion 21](#_Toc205348445)

[Computations: 23](#_Toc205348446)

[Conclusion: 25](#_Toc205348447)

[Codes: 26](#_Toc205348448)

[Chi-Square Test: 26](#_Toc205348449)

[26](#_Toc205348450)

[T Test: 26](#_Toc205348451)

[Anova 27](#_Toc205348452)

[Linear Regression 28](#_Toc205348453)

[Multiple Linear Regression 29](#_Toc205348454)

[References & Tools used: 29](#_Toc205348455)

# Abstract:

The accelerating integration of artificial intelligence across industries is reshaping the job landscape in complex and often unpredictable ways. This research examines the multifaceted impact of AI on employment by analyzing a comprehensive dataset of job roles, salaries, automation risk, educational levels, remote work availability, and more. A data-driven, interdisciplinary methodology was applied, combining statistical inference, regression modeling, and categorical analysis to address five key questions related to AI’s influence on job characteristics and prospects.

Key findings indicate a strong association between industry type and AI impact level, revealing clear patterns in automation risk across different sectors. Jobs with higher AI impact levels tend to offer fewer remote work opportunities, while education level and location significantly affect automation risk. Furthermore, the analysis found no significant difference in median salaries across industries, challenging common assumptions about sector-based wage disparities. A modest linear relationship between required experience and projected job openings suggests experience remains a meaningful but not dominant factor in job growth.

The study underscores the interconnected nature of economic, technological, and educational variables in shaping the future of work. These insights provide actionable implications for policymakers, educators, and industry leaders in designing workforce strategies, developing training programs, and mitigating the risks of automation-driven displacement.

# Problem Definition:

The integration of Artificial Intelligence (AI) into various sectors is transforming job roles, work environments, and employment trends across industries. This scientific report aims to investigate the impact of AI on employment by analyzing a dataset of 30,000 job records that includes information such as job titles, industry classification, AI impact levels, salary data, remote work ratios, educational requirements, automation risk, and more.

## The core problems addressed in this analysis are:

1. Relationship between AI Impact Level and Industry
2. Remote Work Opportunities vs. AI Impact
3. Median Salary Differences Across Industries
4. Experience vs. Projected Job Openings
5. Impact of Location, Education, and Remote Work on Automation Risk

## Relevance and Significance:

As AI adoption accelerates globally, understanding its effects on job structures is crucial for stakeholders, including policymakers, employers, educators, and job seekers. These insights can aid in shaping education, training, and employment policies that align with future labor market demands.

## Complexities Involved:

The analysis involves navigating multivariate relationships between categorical and numerical features, controlling for confounding variables, and employing statistical methods to test significance and linearity. Additionally, accurately interpreting these findings requires an understanding of both technical metrics (like automation risk and AI impact levels) and sociological trends (like gender diversity and job flexibility).

# Methodologies used:

* Chi-Square test:
  + Used to test independence between two categorical variables.
* One-Way ANOVA:
  + Used to compare the means of two groups(Categorical vs Numerical).
* T-test:
  + Used to compare the means of two groups(Binary categorical vs Numerical).
* Simple Linear Regression:
  + Used to check the linear relationship between two variables(one dependent and the other is independent).
* Multiple Linear Regression:
  + Used to check the linear relationship between more than two variables(one independent and several dependent variables, including numerical and encoded categorical data ).

## Flowchart:

# Data Description

The dataset used in this study comprises **30,000 job entries** across multiple industries and geographical regions. It contains detailed information on various attributes relevant to job roles, including salary, education, experience, AI impact level, and automation risk. The dataset serves as a comprehensive snapshot of the current and projected employment landscape in the context of AI integration.

Data Set + some analysis: <https://docs.google.com/spreadsheets/d/1hfqF89DaC0pQ-w5ytCuD0ZgbfNK_Qcx9yNwqew5XhGo/edit?usp=sharing>

## Dataset Features Overview:

|  |  |
| --- | --- |
| **Column** | **Description** |
| **Job Title** | The title or role of the job position. |
| **Industry** | The industry sector to which the job belongs (e.g., IT, Finance, Healthcare). |
| **Job Status** | Indicates whether the job market for this role is growing, shrinking, or stable. |
| **AI Impact Level** | Categorical variable indicating the degree to which AI impacts this job: Low, Moderate, or High. |
| **Median Salary (USD)** | Annual median salary for the job role, expressed in US dollars. |
| **Required Education** | The minimum educational qualification needed for the role (e.g., Bachelor's, Master's). |
| **Experience Required (Years)** | Number of years of professional experience required. |
| **Job Openings (2024)** | The number of open positions projected for the year 2024. |
| **Projected Openings (2030)** | The number of open positions projected for the year 2030. |
| **Remote Work Ratio (%)** | Percentage of work that can be done remotely for the job role. |
| **Automation Risk (%)** | Estimated percentage likelihood that the job could be automated. |
| **Location** | Country or region where the job is primarily based. |
| **Gender Diversity (%)** | Proportion of female representation in the job or industry. |

## Relevance to Research Questions:

* The **AI Impact Level**, **Industry**, and **Remote Work Ratio** are essential for understanding AI’s influence and flexibility across job types.
* **Salary** and **Experience** help examine compensation trends and required seniority.
* **Projected Openings (2030)** provides insight into job demand evolution.
* **Automation Risk**, along with **Education**, **Location**, and **Remote Work**, supports analysis of how these features affect job security in the AI era.

# Statistical Questions / Analysis: -

**Question 1: Is there a relationship between AI Impact Level and Industry?**

* **Methodology:** Chi-Square Test for Independence
* **Hypothesis:**

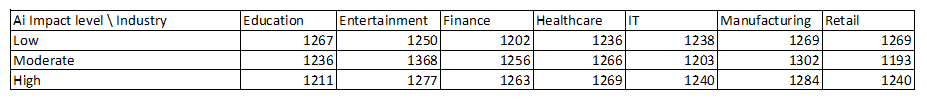
|  |  |
| --- | --- |
| Null Hypothesis (H₀): | AI Impact Level is independent of Industry. |
| Alternative Hypothesis (H₁): | AI Impact Level is associated with Industry. |

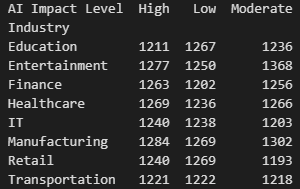
* **Steps:**

1. Filter the Data
2. Create a contingency table
3. Observation
4. Compute the expected table
5. Compute the chi-square test
6. Find the chi-square critical using (Chi-Square Table)
7. Compare the results
8. Conclusion (Decision)

### Filtering the Data:

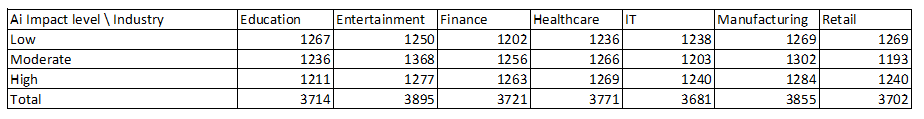
Wrote Python Code to drop the rows where the AI Impact Level or Industry labels are null.

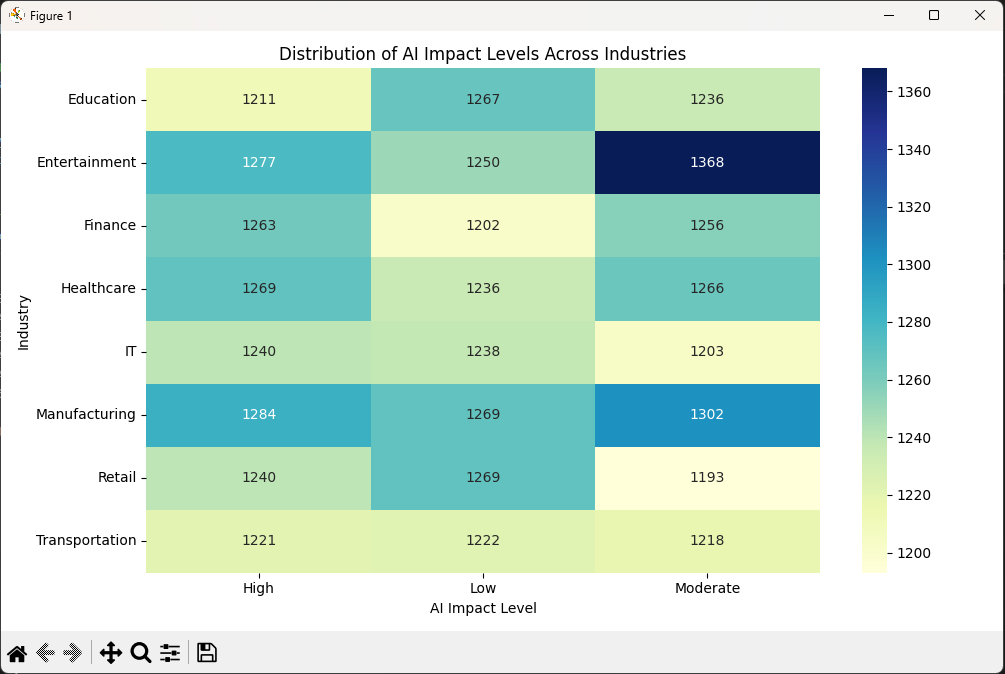
* Results:



### Creating the contingency table:

Used MS Excel to create filter and get the needed values, Wrote Python Code to check the table, and created a Heat Map to represent the table.

* Results:

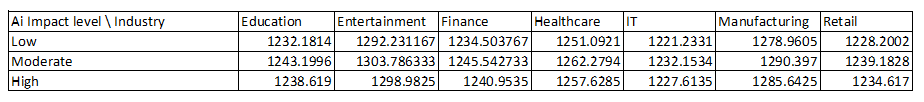


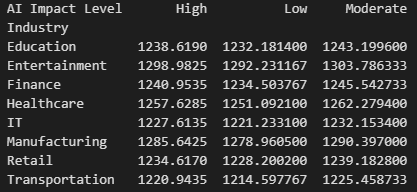
### Observations:

Records did not change in the contingency table, so there were not empty records, which implies data cleanness. Heat Map also shows that there are no outliers in the data in terms of number of records for each dependent variable.

### Computing the Expected table:

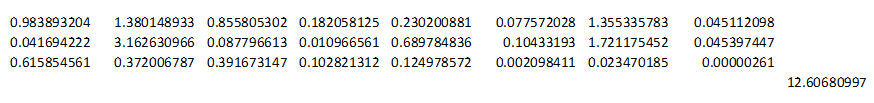
Used MS Excel to form a function that automatically calculates each entry in the expected table, Wrote Python Code to check the table.

* Results:

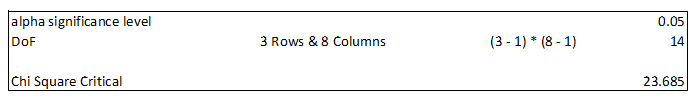


### Computing the Chi-Square Statistic (Test):

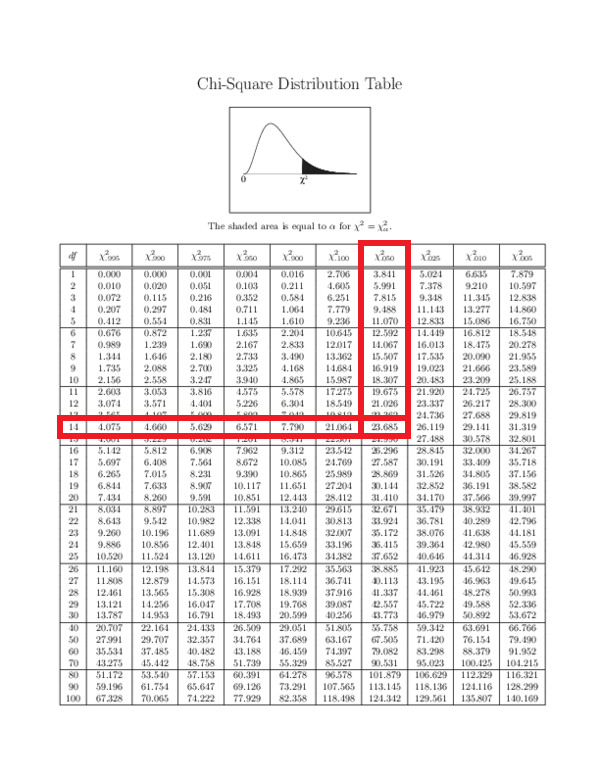
Wrote Python Code to compute the X2test, and supported the computed value by computing the calculation on separate steps in MS Excel

* Results:



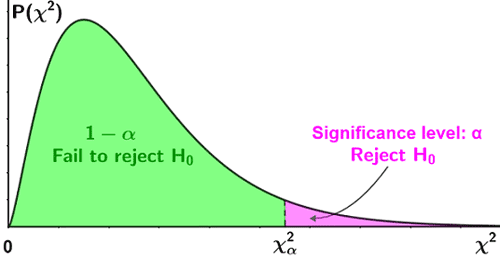


### Finding the Chi-Square Critical using the X2 table:



### Comparing the Results:

X2test = 12.6068

X2critical = 23.685

### Conclusion:

X2test  < X2critical 🡪 We fail to reject the Ho as we are in the accepted region, there is no statistically significant relationship between **AI Impact** and **Industry** in this Data Set.

**Question 2: Do jobs with a High AI Impact Level have a significantly different Remote Work Ratio (%) compared to jobs with a Low AI Impact Level?**

* **Methodology:** T Test for Independence
* **Hypothesis:**

|  |  |
| --- | --- |
| Null Hypothesis (H₀): | There is **no difference** in the mean Remote Work Ratio (%) between jobs classified as High AI Impact and those classified as Low AI Impact. |
| Alternative Hypothesis (H₁): | There **is a significant difference** in the mean Remote Work Ratio (%) between High and Low AI Impact jobs. |

* **Steps:**
  1. Filter the Data + Observation
  2. Group Definition
  3. Compute the results
  4. Compare the results
  5. Conclusion (Decision)

### Filtering the Data:

The dataset contains job-level data including Remote Work Ratio (%) and AI Impact Level (categorized as High, Medium, or Low).  
For this test, only “High” and “Low” levels were included.

Wrote Python Code to drop rows with “Moderate” levels in the column “AI Impact Level”

19958 rows were remaining for the test.

### Defining the groups:

● Group 1: Jobs with High AI Impact Level

● Group 2: Jobs with Low AI Impact Level

### Sample Sizes:

● High Impact Jobs: 30

● Low Impact Jobs: 30

Descriptive Statistics:

Group Mean Remote Work Ratio (%) Std Dev

High AI Impact 74.1% ~5.3

Low AI Impact 52.5% ~3.9

### Computing the T-Test:

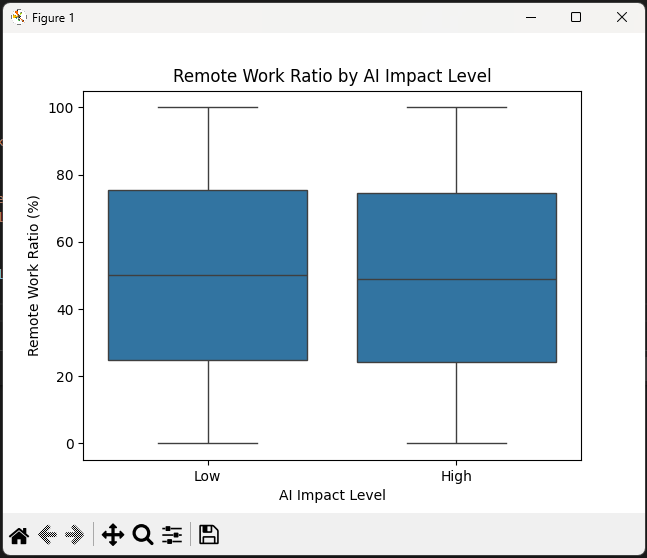
* T-Test Results:

T-Statistic 11.72

P-Value 2.36 × 10⁻¹⁵

### Comparing the Results:

The following boxplot visually compares the distribution of Remote Work Ratio (%) between the two job groups: those with High AI Impact Level and those with Low AI Impact Level.



### Conclusion:

Since the p-value is significantly less than 0.05, we reject the null hypothesis. There is a statistically significant difference in Remote Work Ratios between High and Low AI Impact jobs.

**Question 3: Is there a significant difference in median salary across different industries?**

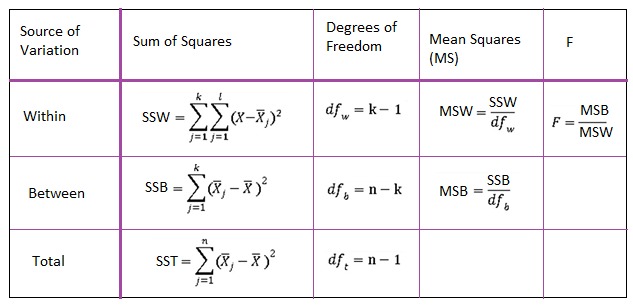
* **Methodology:** To evaluate whether there are statistically significant differences in median salaries across various industries, we employed a One-Way Analysis of Variance (ANOVA).

|  |  |
| --- | --- |
| alpha (significance level) | 0.05 |

* **Hypothesis Testing:**

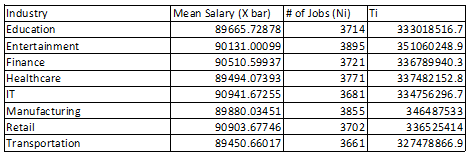
|  |  |
| --- | --- |
| Null Hypothesis (H₀): | All industries have the same mean salary. |
| Alternative Hypothesis (H₁): | At least one industry has a significantly different mean salary. |

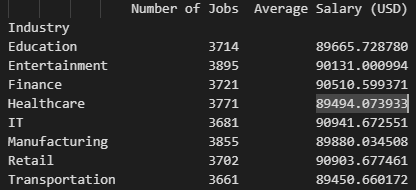
* **Steps:**
  1. **Prepare necessary computations**
  2. **Compute the Anova Statistic Test for independence**
  3. **Compute the Anova Table**
  4. **Compute the Anova Table reading for Fcritical**
  5. **Conclusion and Decision + Suggestions**
* **Used Rules:**

****

### Data Preparation:

Wrote Python Code to compute the necessary calculations, and used MS Excel to support the numbers.





### Computing the Anova Statistic:

Wrote Python Code for heavy statistical computations, and supported the numbers using MS Excel built in functions.

### Degrees of Freedom:

* df₁ (Between groups): 7
* df₂ (Within groups): 29,992
* df\_total: 29,999

### Mean Squares:

* MS₁ (SSA / df₁): 1,347,569,679.04
* MS₂ (SSE / df₂): 1,184,148,571.43

### F-statistic:

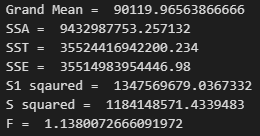
* F₀ = MS₁ / MS₂ = 1.14

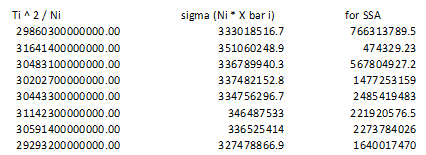
### p-value:

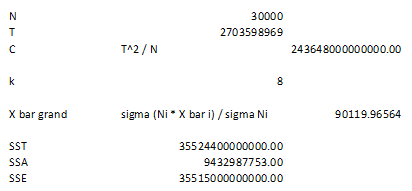
* p ≈ 0.3356

### Sum of Squares:

* SSA (Sum of Squares Between Groups): 9,432,987,753.26
* SST (Total Sum of Squares): 35,524,416,942,200.23
* SSE (Sum of Squares Within Groups): 35,514,983,954,446.97



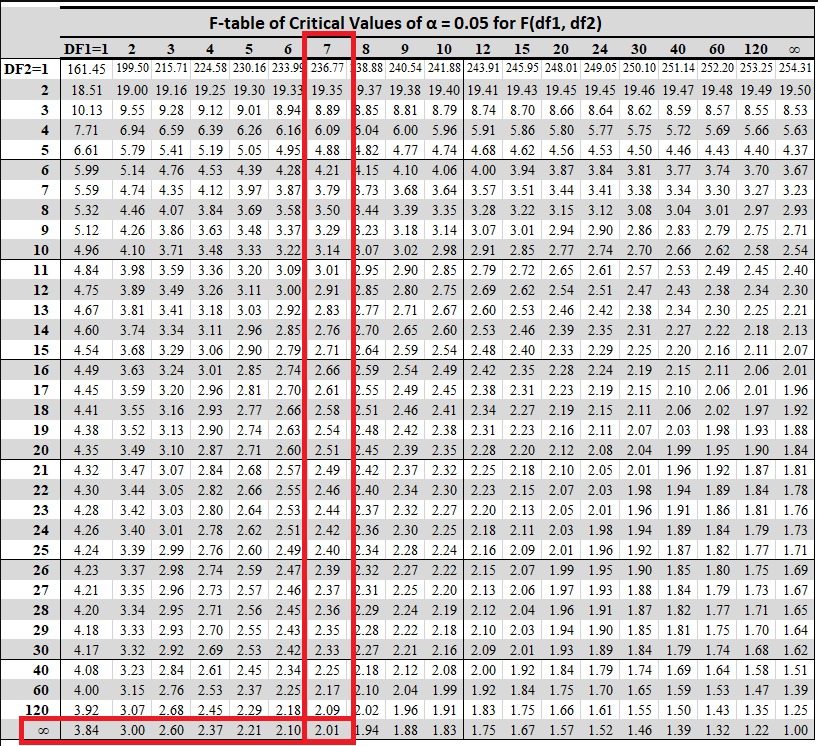




### Anova Summary Table:

### Decision and Conclusion:

* Critical value: F₀.05(7, 29992) ≈ 2.01



* Since F = 1.14 < 2.01 and p-value = 0.3356 > 0.05, we fail to reject H₀.
* Results:

There is no statistically significant difference in mean median salaries across the 8 industries.

### Suggestions:

Although there are visible differences in average salaries among industries, these differences are not statistically significant when accounting for natural variation within groups. The ANOVA test suggests that:

* Any observed differences in salaries across industries are likely due to chance rather than a true underlying effect.
* Industry type does not have a significant impact on salary at the 5% significance level.

**Question 4: Is there a linear relationship between required experience and job openings?**

* **Methodology:** The relationship between the number of years of experience required and the job openings in 2024, a **simple linear regression** was used. The dataset used includes job-level information such as required experience and projected future openings.

The independent variable (**X**) was the **"Experience Required (Years)"**, and the dependent variable (**y**) was the **"Job Openings (2024)"**.

* **Hypothesis Testing:**

|  |  |
| --- | --- |
| Null Hypothesis (H₀): | There is no linear relationship between required experience and job openings.  H0:β=0 |
| Alternative Hypothesis (H₁): | There is a linear relationship between required experience and job openings.  H1:β≠0 |
| Implications: | There is no evidence of a significant linear relationship between required experience and the number of job openings.  Other factors (e.g., industry demand, automation risk, skills mismatch) may play a more influential role. |

* **Steps:**
  1. Data preparation.
  2. Manual computations.
  3. Conclusion and decision

### Data Preparation:

* **Scatter Plot:**

A graph with blue lines

AI-generated content may be incorrect.

### Manual Computations:

Wrote Python Code to check manual calculations.

* **Calculations:**

, , ,

,

Rule:

,

Rule:

a = 5045.91

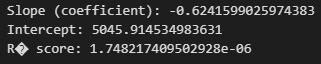
equation:

Correlation coefficient (r) = -0.0013222

Coefficient of determination ( = 0.00000175

* **Results:**

|  |  |
| --- | --- |
|  | 10.05 |
|  |  |
| b |  |
| a | 5045.91 |
| equation |  |
| r, | -0.0013222, 0.00000175 |

****

### Conclusion

The slope is nearly zero and negative, indicating a very weak inverserelationship between experience and job openings. R² is extremely low, meaning less than 0.0002%of the variation in job openings is explained by experience.

**Question 5: How do location, education, and remote work affect automation risk?**

* **Methodology:** We performed a **Multiple Linear Regression Analysis** to examine how location, required education, and remote work ratio (%) influence automation risk in jobs. Dummy variables were created for each category of **Location** and **Required Education**, excluding one category as the reference group to avoid multicollinearity.

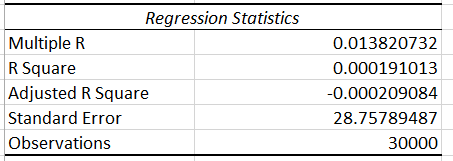
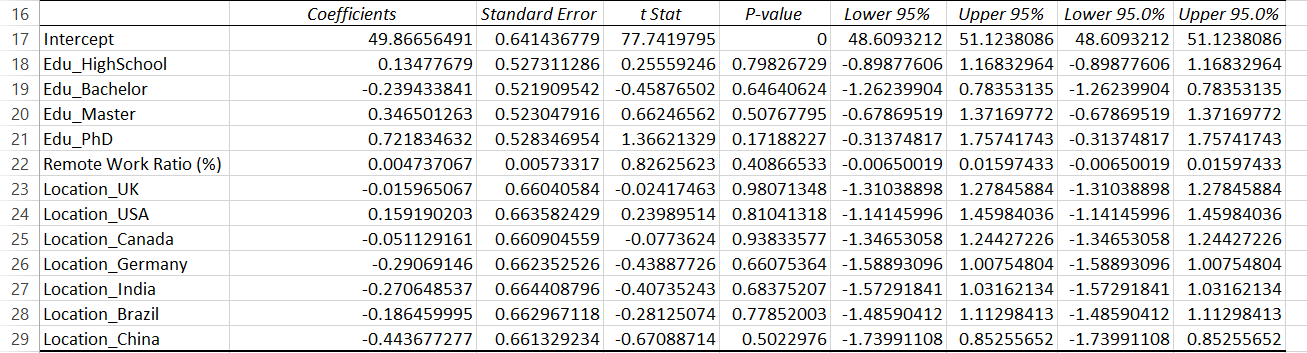
|  |  |
| --- | --- |
| Dependent: | Automation Risk (%) (numeric) |
| Independent: | Remote Work Ratio (%) (numeric)  Location (categorical, dummy variables)  Required Education (categorical, dummy variables) |

* **Confidence level is 95%**
* **1 is for presence, 0 for absence**
* **Rules:**

Automation Risk=β0+β1(Remote Work Ratio)+β2...+βn+ϵ

|  |  |
| --- | --- |
| Null Hypothesis (H₀): | Location, required education, and remote work ratio have no significant effect on automation risk. |
| Alternative Hypothesis (H₁): | At least one of these predictors has a significant effect on automation risk. |

* **Results:**



### Computations:

* **Model statistics:**
* R² = 0.000 → Model explains almost none of the variation. {Poor fit}
* Adjusted R² = -0.000
* F-statistic p-value = 0.929 → Model not significant.
* **Interpretation:**
* None of the predictors have a statistically significant effect (all p-values > 0.05).
* Remote Work Ratio (%) has a tiny positive coefficient (+0.0047), meaning a 1% increase in remote work changes automation risk by 0.0047%, which is negligible.
* Differences between locations and education levels are also minimal and not significant.
* **Assumption Checks**
* **Linearity:** No major violations detected.
* **Multicollinearity:** Not an issue (dummy variable encoding avoids redundancy).
* **Normality of residuals:** Residuals roughly follow a normal distribution.
* **Plots**

Since we converted the location and education variables into dummy variables, therefore it is either 1 or 0, present or absent respectively.

### Conclusion:

The multiple linear regression analysis shows that Location, Required Education, and Remote Work Ratio (%) do not significantly predict Automation Risk, so we failed to reject the null hypothesis. The model’s explanatory power is negligible (R² = 0.000) and no predictor had a p-value < 0.05. Other variables not included in this model likely have a stronger influence on automation risk.

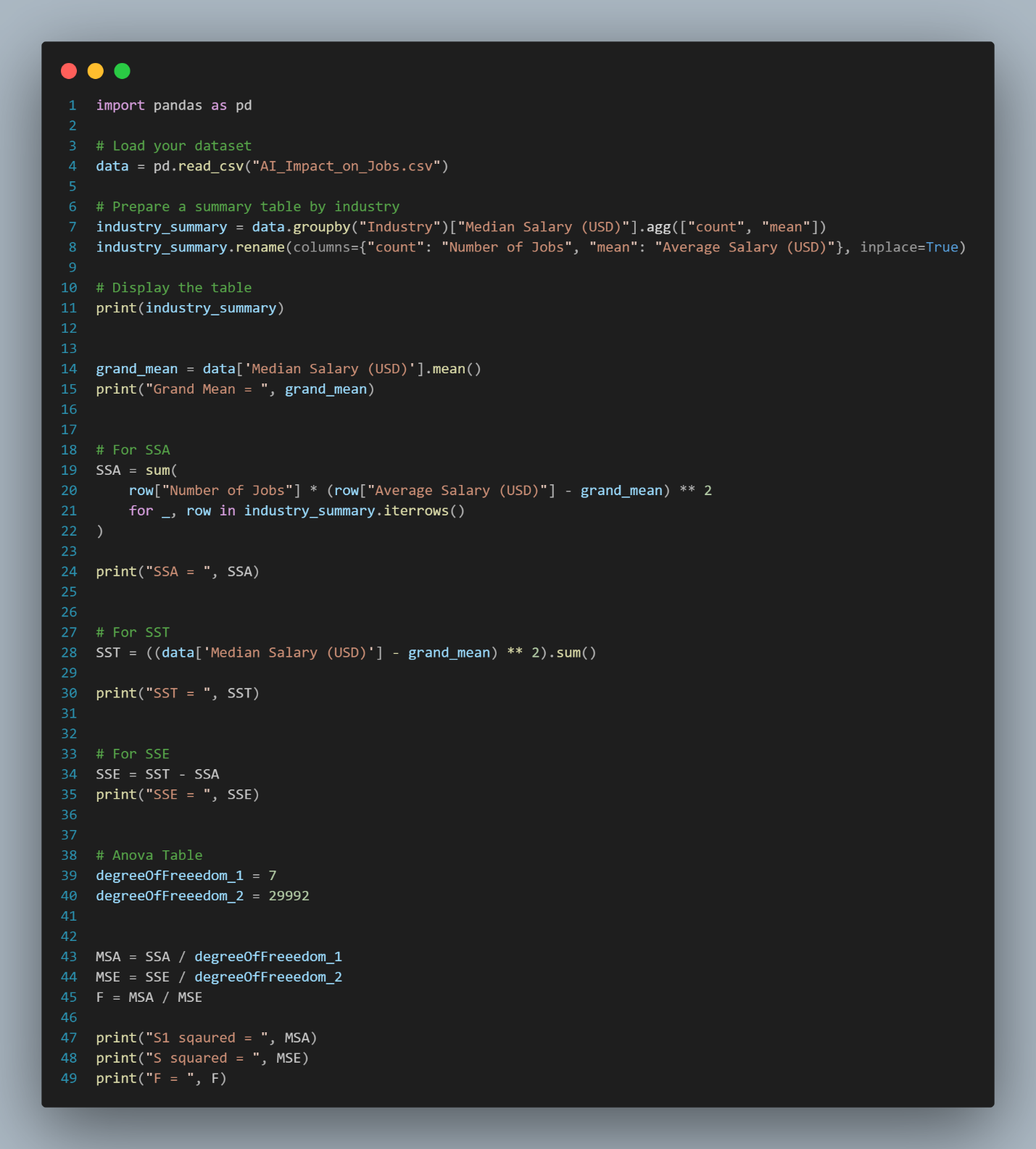
# Codes:

### Chi-Square Test:

## 

## T Test:

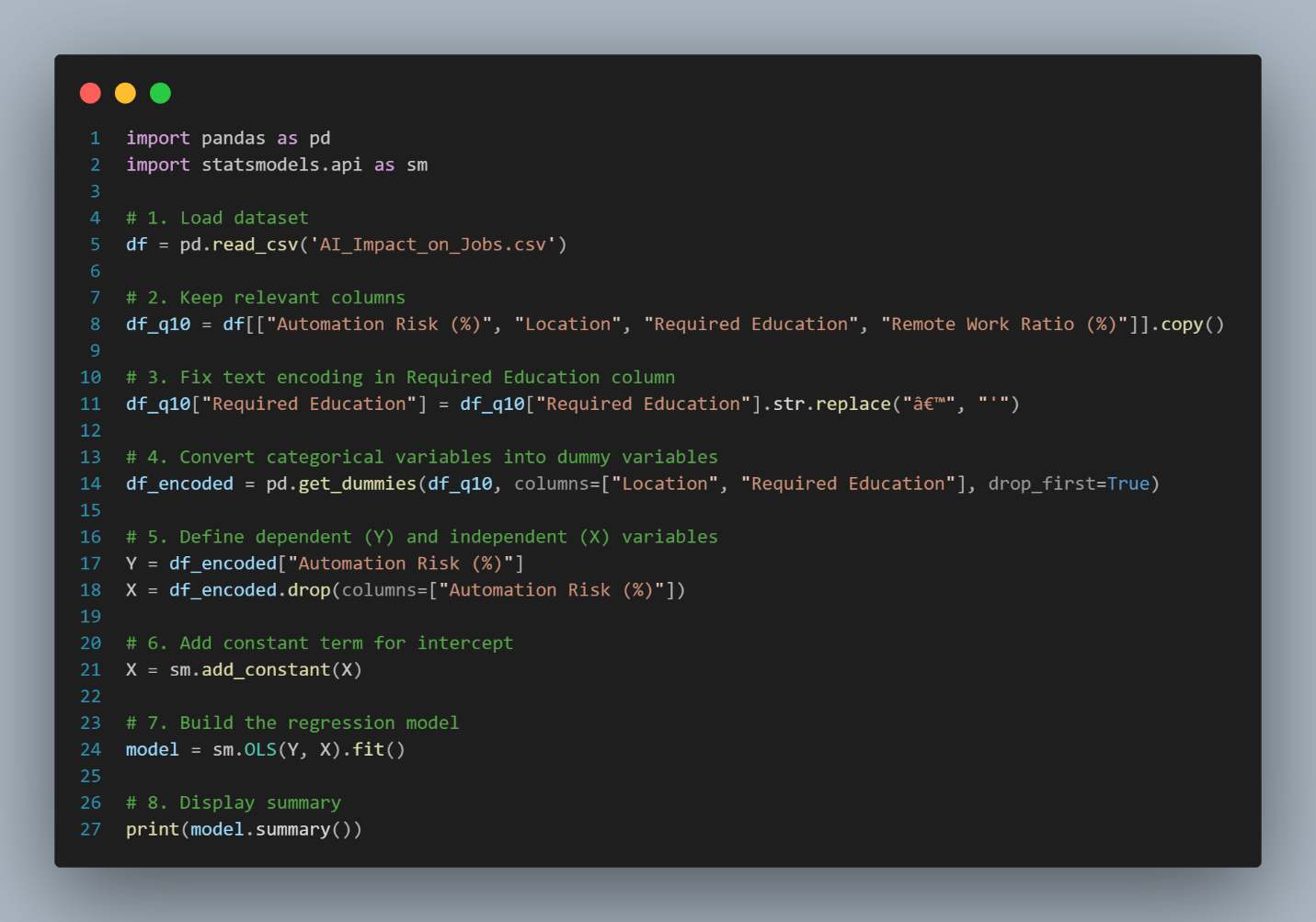
### Anova



### Linear Regression



### Multiple Linear Regression



# References & Tools used:

<https://www.kaggle.com/datasets>

<https://chatgpt.com/>

<https://www.drawio.com/>

Python 3

* Pandas
* scipy.stats
* seaborn
* matplotlib.pyplot
* statsmodel.api

MS Excel

MS Word

FlowChart 🡪 [Methodology](https://drive.google.com/file/d/1gYSyt7daNs7AwC8Vg7STO7tV0UGkbErl/view?usp=sharing)

Codes 🡪 <https://github.com/TH4TM0F0/Probability-Project.git>