Cairo University – Faculty of Engineering
MTHS204 Advanced Probability and Statistics
Summer 2025

AI Impact on Jobs

404
JOB NOT
FOUND

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:

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.

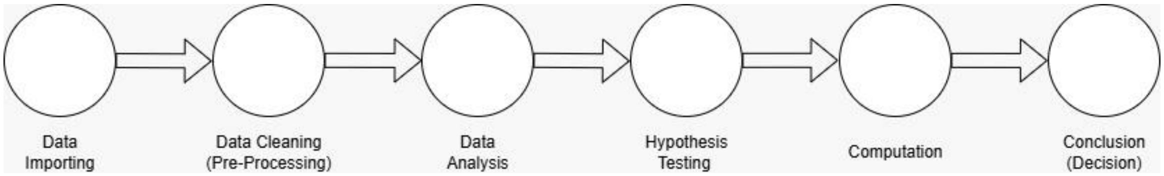
Complexities Involved:

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 (H_0):	AI Impact Level is independent of Industry.
Alternative Hypothesis (H_1):	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:

Ai Impact level \ Industry	Education	Entertainment	Finance	Healthcare	IT	Manufacturing	Retail
Low	1267	1250	1202	1236	1238	1269	1269
Moderate	1236	1368	1256	1266	1203	1302	1193
High	1211	1277	1263	1269	1240	1284	1240

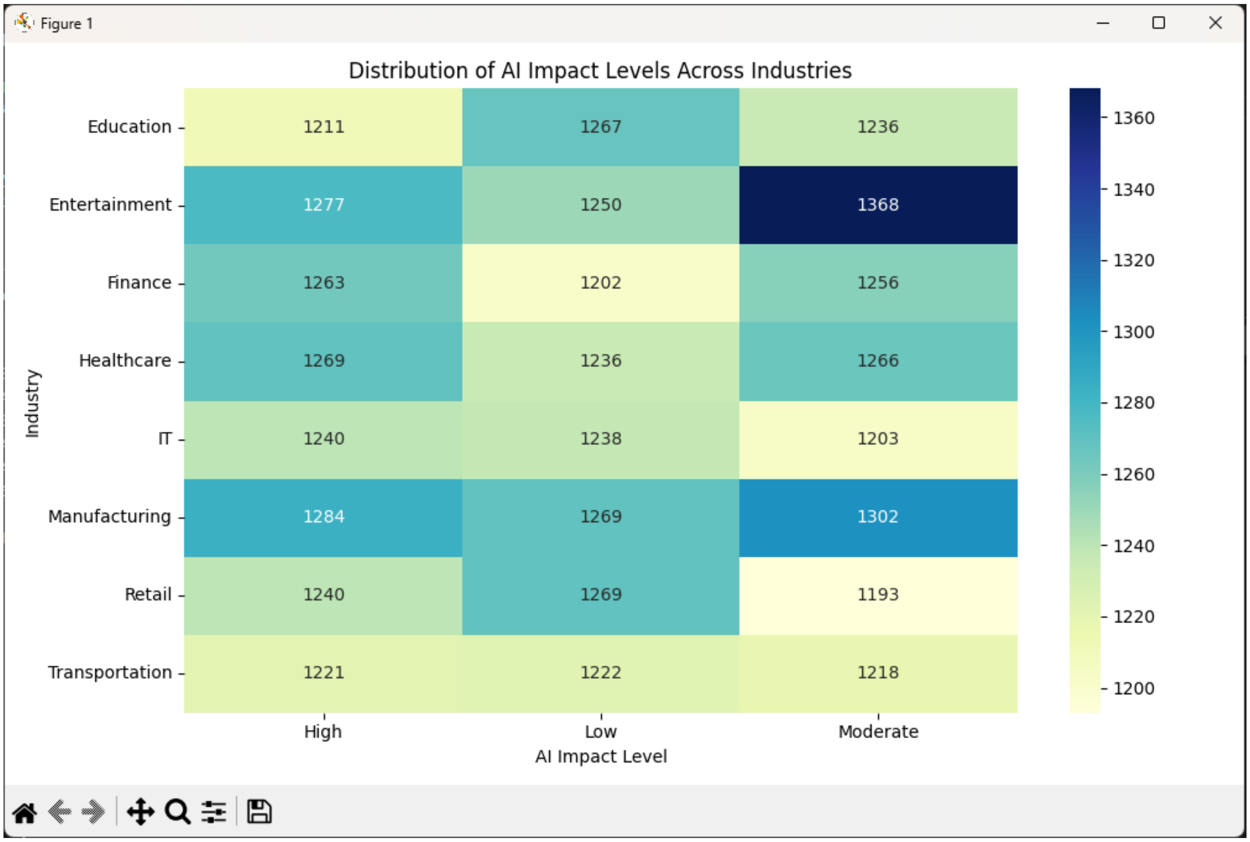
```
AI Impact Level  High   Low  Moderate
Industry
Education        1211  1267      1236
Entertainment    1277  1250      1368
Finance          1263  1202      1256
Healthcare       1269  1236      1266
IT               1240  1238      1203
Manufacturing    1284  1269      1302
Retail           1240  1269      1193
Transportation   1221  1222      1218
```

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:

Ai Impact level \ Industry	Education	Entertainment	Finance	Healthcare	IT	Manufacturing	Retail
Low	1267	1250	1202	1236	1238	1269	1269
Moderate	1236	1368	1256	1266	1203	1302	1193
High	1211	1277	1263	1269	1240	1284	1240
Total	3714	3895	3721	3771	3681	3855	3702



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:

Ai Impact level \ Industry	Education	Entertainment	Finance	Healthcare	IT	Manufacturing	Retail
Low	1232.1814	1292.231167	1234.503767	1251.0921	1221.2331	1278.9605	1228.2002
Moderate	1243.1996	1303.786333	1245.542733	1262.2794	1232.1534	1290.397	1239.1828
High	1238.619	1298.9825	1240.9535	1257.6285	1227.6135	1285.6425	1234.617

AI Impact Level	High	Low	Moderate
Industry			
Education	1238.6190	1232.181400	1243.199600
Entertainment	1298.9825	1292.231167	1303.786333
Finance	1240.9535	1234.503767	1245.542733
Healthcare	1257.6285	1251.092100	1262.279400
IT	1227.6135	1221.233100	1232.153400
Manufacturing	1285.6425	1278.960500	1290.397000
Retail	1234.6170	1228.200200	1239.182800
Transportation	1220.9435	1214.597767	1225.458733

Computing the Chi-Square Statistic (Test):

Wrote Python Code to compute the X^2_{test}, and supported the computed value by computing the calculation on separate steps in MS Excel

- Results:

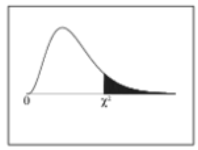
0.983893204	1.380148933	0.855805302	0.182058125	0.230200881	0.077572028	1.355335783	0.045112098	
0.041694222	3.162630966	0.087796613	0.010966561	0.689784836	0.10433193	1.721175452	0.045397447	
0.615854561	0.372006787	0.391673147	0.102821312	0.124978572	0.002098411	0.023470185	0.00000261	
								12.60680997

```
Chi-Square Statistic: 12.606809901768466
Degrees of Freedom: 14
P-value: 0.5576902129431236
```

alpha significance level			0.05
DoF	3 Rows & 8 Columns	(3 - 1) * (8 - 1)	14
Chi Square Critical			23.685

Finding the Chi-Square Critical using the X^2 table:

Chi-Square Distribution Table



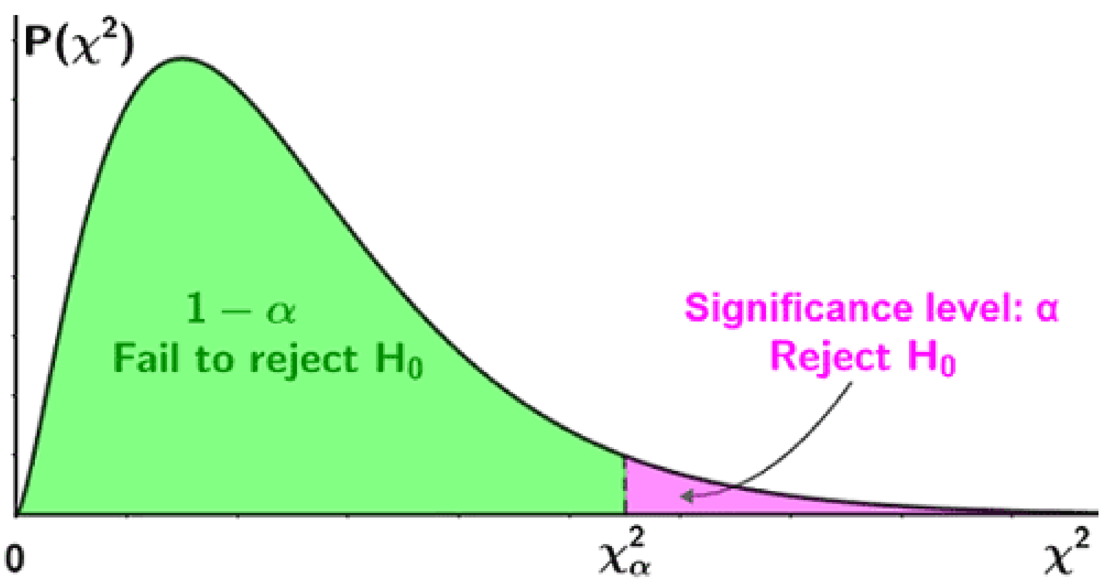
The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

Comparing the Results:

$X^2_{test} = 12.6068$

$X^2_{critical} = 23.685$



Conclusion:

$X^2_{test} < X^2_{critical}$ → We fail to reject the H_o 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 (H_0):	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 (H_1):	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

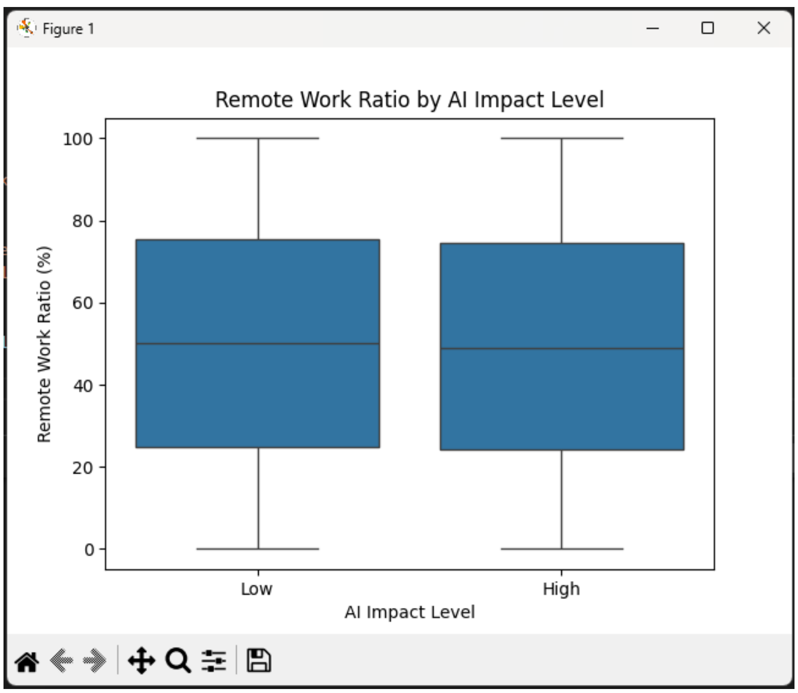
Computing the T-Test:

- 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 (H_0):	All industries have the same mean salary.
Alternative Hypothesis (H_1):	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 $F_{critical}$**
  5. **Conclusion and Decision + Suggestions**

- **Used Rules:**

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$	$df_w = k-1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2$	$df_b = n-k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^{n}(\bar{X}_j-\bar{X})^2$	$df_t = n-1$		

Data Preparation:

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

Industry	Mean Salary (X bar)	# of Jobs (Ni)	Ti
Education	89665.72878	3714	333018516.7
Entertainment	90131.00099	3895	351060248.9
Finance	90510.59937	3721	336789940.3
Healthcare	89494.07393	3771	337482152.8
IT	90941.67255	3681	334756296.7
Manufacturing	89880.03451	3855	346487533
Retail	90903.67746	3702	336525414
Transportation	89450.66017	3661	327478866.9

	Number of Jobs	Average Salary (USD)
Industry		
Education	3714	89665.728780
Entertainment	3895	90131.000994
Finance	3721	90510.599371
Healthcare	3771	89494.073933
IT	3681	90941.672551
Manufacturing	3855	89880.034508
Retail	3702	90903.677461
Transportation	3661	89450.660172

Computing the Anova Statistic:

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

Degrees of Freedom:

- df_1 (Between groups): 7
- df_2 (Within groups): 29,992
- df_total: 29,999

Mean Squares:

- MS_1 (SSA / df_1): 1,347,569,679.04
- MS_2 (SSE / df_2): 1,184,148,571.43

F-statistic:

- $F_0 = MS_1 / MS_2 = 1.14$

p-value:

- $p \approx 0.3356$

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

```
Grand Mean =  90119.96563866666
SSA =  9432987753.257132
SST =  35524416942200.234
SSE =  35514983954446.98
S1 sqaured =  1347569679.0367332
S squared =  1184148571.4339483
F =  1.1380072666091972
```

Ti ^ 2 / Ni	sigma (Ni * X bar i)	for SSA
29860300000000.00	333018516.7	766313789.5
31641400000000.00	351060248.9	474329.23
30483100000000.00	336789940.3	567804927.2
30202700000000.00	337482152.8	1477253159
30443300000000.00	334756296.7	2485419483
31142300000000.00	346487533	221920576.5
30591400000000.00	336525414	2273784026
29293200000000.00	327478866.9	1640017470

N		30000
T		2703598969
C	T^2 / N	243648000000000.00
k		8
X bar grand	sigma (Ni * X bar i) / sigma Ni	90119.96564
SST		355244000000000.00
SSA		9432987753.00
SSE		355150000000000.00

Anova Summary Table:

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares
Treatment	9432987753	7	1347569679
Error	3.55E+13	29992	1184148571
Total	3.55E+13	29999	

F critical	2.01

Decision and Conclusion:

- Critical value: $F_{0.05}(7, 29992) \approx 2.01$

F-table of Critical Values of α = 0.05 for F(df1, df2)																			
	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	38.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	9.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

- Since $F = 1.14 < 2.01$ and p-value $= 0.3356 > 0.05$, we fail to reject H_0.

- 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)"**.

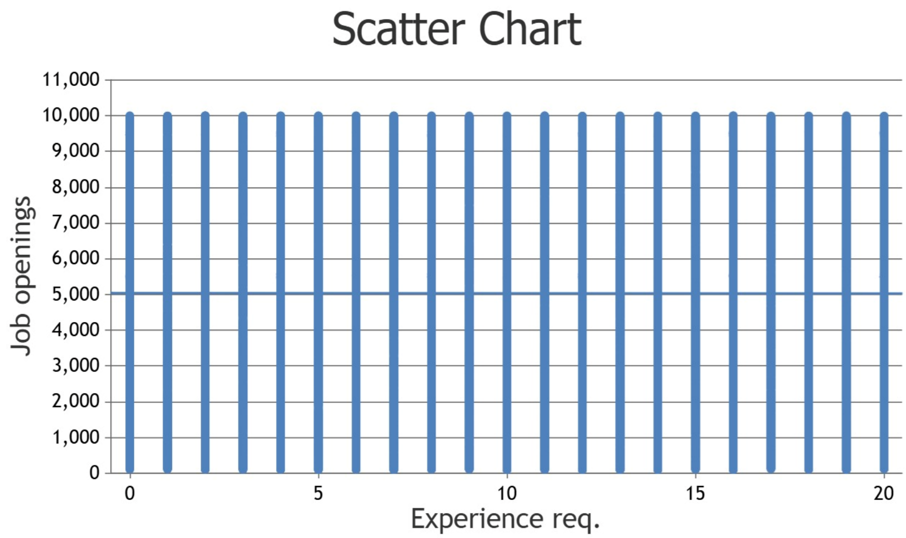
- **Hypothesis Testing:**

Null Hypothesis (H_0):	There is no linear relationship between required experience and job openings. $H0: \beta=0$
Alternative Hypothesis (H_1):	There is a linear relationship between required experience and job openings. $H1: \beta \neq 0$
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:**



Manual Computations:

Wrote Python Code to check manual calculations.

- **Calculations:**

$\sum x = 301543$, $\sum y = 151189225$, $\sum xy = 1518980643$,
$\sum x^2 = 4132857$, $\sum y^2 = 1007492493653$

Rule:

$$\bar{x} = \frac{\sum x}{n} \quad \bar{y} = \frac{\sum y}{n}$$

$$\bar{x} = \frac{301543}{30000} = 10.05, \ \bar{y} = \frac{151189225}{30000} = 5039.64$$

Rule:

$$b = \frac{n \times \sum xy - \sum x \sum y}{n \times \sum x^2 - (\sum x)^2} \qquad a = \bar{y} - b\bar{x}$$

$$b = \frac{30000 \times 1518980643 - 301543 \times 151189225}{30000 \times 4132857 - 301543^2} = -0.624 \quad a = 5045.91$$

equation: $y = 5045.91 - 0.624x$

Correlation coefficient (r) = -0.0013222

Coefficient of determination (R^2) = 0.00000175

- **Results:**

$\bar{x}$	10.05
$\bar{y}$	5039.64
b	-0.624
a	5045.91
equation	$y = 5045.91 - 0.624x$
r, R^2	-0.0013222, 0.00000175

```
Slope (coefficient): -0.6241599025974383
Intercept: 5045.914534983631
R� score: 1.748217409502928e-06
```

Conclusion

The slope is nearly zero and negative, indicating a very weak inverse relationship between experience and job openings. R^2 is extremely low, meaning less than 0.0002% of the variation in job openings is explained by experience.

<u>Question 5:</u> How do location, education, and remote work affect automation risk?

- <u>**Methodology:**</u> 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**

- <u>**Rules:**</u>

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

Null Hypothesis (H_0):	Location, required education, and remote work ratio have no significant effect on automation risk.
Alternative Hypothesis (H_1):	At least one of these predictors has a significant effect on automation risk.

➢ Results:

16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	49.86656491	0.641436779	77.7419795	0	48.6093212	51.1238086	48.6093212	51.1238086
18	Edu_HighSchool	0.13477679	0.527311286	0.25559246	0.79826729	-0.89877606	1.16832964	-0.89877606	1.16832964
19	Edu_Bachelor	-0.239433841	0.521909542	-0.45876502	0.64640624	-1.26239904	0.78353135	-1.26239904	0.78353135
20	Edu_Master	0.346501263	0.523047916	0.66246562	0.50767795	-0.67869519	1.37169772	-0.67869519	1.37169772
21	Edu_PhD	0.721834632	0.528346954	1.36621329	0.17188227	-0.31374817	1.75741743	-0.31374817	1.75741743
22	Remote Work Ratio (%)	0.004737067	0.00573317	0.82625623	0.40866533	-0.00650019	0.01597433	-0.00650019	0.01597433
23	Location_UK	-0.015965067	0.66040584	-0.02417463	0.98071348	-1.31038898	1.27845884	-1.31038898	1.27845884
24	Location_USA	0.159190203	0.663582429	0.23989514	0.81041318	-1.14145996	1.45984036	-1.14145996	1.45984036
25	Location_Canada	-0.051129161	0.660904559	-0.0773624	0.93833577	-1.34653058	1.24427226	-1.34653058	1.24427226
26	Location_Germany	-0.29069146	0.662352526	-0.43887726	0.66075364	-1.58893096	1.00754804	-1.58893096	1.00754804
27	Location_India	-0.270648537	0.664408796	-0.40735243	0.68375207	-1.57291841	1.03162134	-1.57291841	1.03162134
28	Location_Brazil	-0.186459995	0.662967118	-0.28125074	0.77852003	-1.48590412	1.11298413	-1.48590412	1.11298413
29	Location_China	-0.443677277	0.661329234	-0.67088714	0.5022976	-1.73991108	0.85255652	-1.73991108	0.85255652

Regression Statistics	
Multiple R	0.013820732
R Square	0.000191013
Adjusted R Square	-0.000209084
Standard Error	28.75789487
Observations	30000

Computations:

➢ Model statistics:

* $R^2 = 0.000$ → Model explains almost none of the variation. {Poor fit}
* Adjusted $R^2 = -0.000$
* F-statistic p-value = 0.929 → Model not significant.

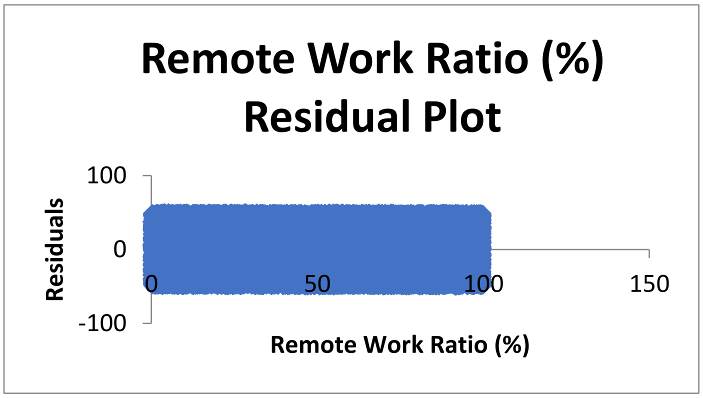
➢ Interpretation:

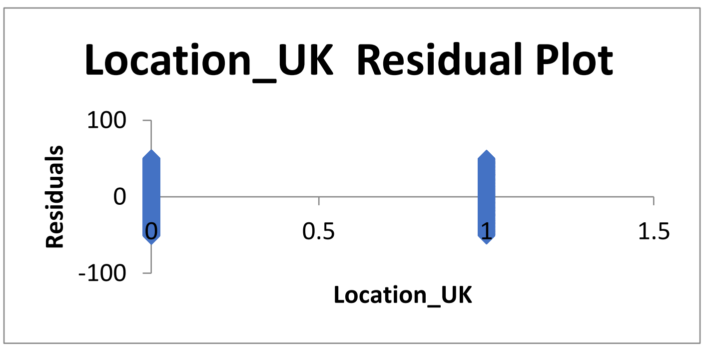
* None of the predictors have a statistically significant effect (all p-values > 0.05).
* Remote Work Ratio (%) has a tiny positive coefficient (+0.0047), meaning a 1% increase in remote work changes automation risk by 0.0047%, which is negligible.
* Differences between locations and education levels are also minimal and not significant.

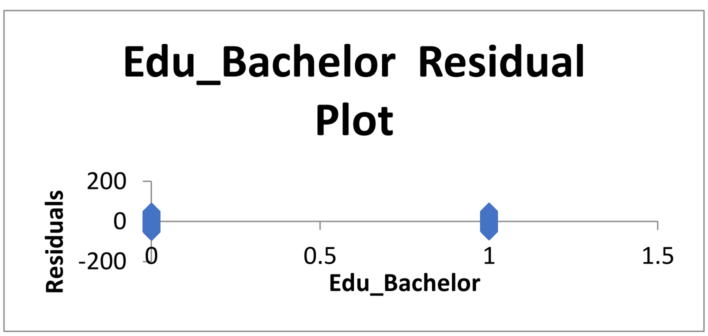
➢ Assumption Checks

- **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 ($R^2 = 0.000$) 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:

```python
import pandas as pd
from scipy.stats import chi2_contingency
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Dataset from the CSV file
dataSet = pd.read_csv("AI_Impact_on_Jobs.csv")

# Filtering the Data
cleanedData = dataSet[['AI Impact Level', 'Industry']].dropna()


#Creating the Continngency table
contingency_table = pd.crosstab(cleanedData['Industry'], cleanedData['AI Impact Level'])
print(contingency_table)


# Computing the chi-square test
chi2, p, dof, expected = chi2_contingency(contingency_table)


# Printing the Expected table]
expected_df = pd.DataFrame(expected, index=contingency_table.index, columns=contingency_table.columns)
print(expected_df)


# Print the results
print("Chi-Square Statistic:", chi2)
print("Degrees of Freedom:", dof)
print("P-value:", p)



plt.figure(figsize=(10, 6))
sns.heatmap(contingency_table, annot=True, fmt="d", cmap="YlGnBu")

plt.title("Distribution of AI Impact Levels Across Industries")
plt.xlabel("AI Impact Level")
plt.ylabel("Industry")
plt.tight_layout()
plt.show()
```

T Test:

```python
import pandas as pd
import scipy.stats as stats
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("AI_Impact_on_Jobs.csv")

# Filter for only High and Low AI Impact Levels
df_filtered = df[df['AI Impact Level'].isin(['High', 'Low'])]

# Drop rows with missing Remote Work Ratio values
df_filtered = df_filtered.dropna(subset=['Remote Work Ratio (%)'])

# Create two groups
remote_high = df_filtered[df_filtered['AI Impact Level'] == 'High']['Remote Work Ratio (%)']
remote_low = df_filtered[df_filtered['AI Impact Level'] == 'Low']['Remote Work Ratio (%)']

# Perform Welch's t-test (assumes unequal variances)
t_stat, p_val = stats.ttest_ind(remote_high, remote_low, equal_var=False)

# Print the results
print("T-statistic:", t_stat)
print("P-value:", p_val)

if p_val < 0.05:
    print("Result: Significant difference in remote work ratio between High and Low AI impact jobs.")
else:
    print("Result: No significant difference in remote work ratio.")

# Visualize the result
sns.boxplot(x='AI Impact Level', y='Remote Work Ratio (%)', data=df_filtered)
plt.title('Remote Work Ratio by AI Impact Level')
plt.ylabel('Remote Work Ratio (%)')
plt.xlabel('AI Impact Level')
plt.show()
```

Anova

```python
import pandas as pd

# Load your dataset
data = pd.read_csv("AI_Impact_on_Jobs.csv")

# Prepare a summary table by industry
industry_summary = data.groupby("Industry")["Median Salary (USD)"].agg(["count", "mean"])
industry_summary.rename(columns={"count": "Number of Jobs", "mean": "Average Salary (USD)"}, inplace=True)

# Display the table
print(industry_summary)


grand_mean = data['Median Salary (USD)'].mean()
print("Grand Mean = ", grand_mean)


# For SSA
SSA = sum(
    row["Number of Jobs"] * (row["Average Salary (USD)"] - grand_mean) ** 2
    for _, row in industry_summary.iterrows()
)

print("SSA = ", SSA)


# For SST
SST = ((data['Median Salary (USD)'] - grand_mean) ** 2).sum()

print("SST = ", SST)


# For SSE
SSE = SST - SSA
print("SSE = ", SSE)


# Anova Table
degreeOfFreeedom_1 = 7
degreeOfFreeedom_2 = 29992


MSA = SSA / degreeOfFreeedom_1
MSE = SSE / degreeOfFreeedom_2
F = MSA / MSE

print("S1 sqaured = ", MSA)
print("S squared = ", MSE)
print("F = ", F)
```

Linear Regression

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv("AI_Impact_on_Jobs.csv")

from sklearn.linear_model import LinearRegression

model = LinearRegression()
X = data[['Experience Required (Years)']]
y = data['Job Openings (2024)']

# Create and train model
model = LinearRegression()
model.fit(X, y)

# Predict values
y_pred = model.predict(X)

# Plot
plt.scatter(X, y, label='Data points')
plt.plot(X, y_pred, color='black', label='Regression line')

# Add labels
plt.xlabel("Impact Level")
plt.ylabel("Remote Work Ratio")
plt.title("Linear Regression using Scikit-Learn")
plt.legend()
plt.show()

# Print model coefficients
print("Slope (coefficient):", model.coef_[0])
print("Intercept:", model.intercept_)
print("R² score:", model.score(X, y))
```

Multiple Linear Regression

```python
import pandas as pd
import statsmodels.api as sm

# 1. Load dataset
df = pd.read_csv('AI_Impact_on_Jobs.csv')

# 2. Keep relevant columns
df_q10 = df[["Automation Risk (%)", "Location", "Required Education", "Remote Work Ratio (%)"]].copy()

# 3. Fix text encoding in Required Education column
df_q10["Required Education"] = df_q10["Required Education"].str.replace("â€™", "'")

# 4. Convert categorical variables into dummy variables
df_encoded = pd.get_dummies(df_q10, columns=["Location", "Required Education"], drop_first=True)

# 5. Define dependent (Y) and independent (X) variables
Y = df_encoded["Automation Risk (%)"]
X = df_encoded.drop(columns=["Automation Risk (%)"])

# 6. Add constant term for intercept
X = sm.add_constant(X)

# 7. Build the regression model
model = sm.OLS(Y, X).fit()

# 8. Display summary
print(model.summary())
```

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
Codes → https://github.com/TH4TM0F0/Probability-Project.git