# ANALYSING AUTOMATION PROBABILITY OF PROFESSIONS

INDUSTRY INTERNSHIP REPORT

*Submitted by*

VARSHINI G (241001299)

*in partial fulfilment for the award of the degree of*

BACHELOR OF TECHNOLOGY

*in*

INFORMATION TECHNOLOGY

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**RAJALAKSHMI ENGINEERING COLLEGE**

**THANDALAM**

**JANUARY 2026**

# BONAFIDE CERTIFICATE

Certified that this project report titled "**ANALYSING AUTOMATION PROBABILITY OF PROFESSIONS**" is the Bonafide work of **VARSHINI G (241001299)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**SIGNATURE** **SIGNATURE OF THE COORDINATOR**

**Dr. P. Valarmathie**

**Head of the Department**

Department of Information Technology

Rajalakshmi Engineering College

Submitted to Project Viva Voce Examination held on _________________

# ACKNOWLEDGEMENT

First, we thank the almighty God for the successful completion of the project. Our sincere thanks to our chairman **Mr. S. Meganathan, B.E., F.I.E** for his sincere endeavour in educating us in his premier institution. We would like to express our deep gratitude to our beloved Chairperson **Dr. Thangam Meganathan**, for her enthusiastic motivation which inspired us a lot in completing this project and Vice-Chairman **Mr. Abhay Shankar Meganathan B.E., M.S.,** for providing us with the requisite infrastructure.

We also express our sincere gratitude to our college principal, **Dr. S.N.Murugesan M.E., PhD.**, for his kind support and facilities to complete our work on time. We extend heartfelt gratitude to **Dr. P.Valarmathie**, **Professor and Head of the Department of Information Technology** for her guidance and encouragement throughout the work. We are very glad to thank faculty **Ms.M.Madhu Rani**, Assistant Professor and Internship coordinator of our department for their encouragement and support towards the successful completion of this project. We extend our thanks to our parents, friends, all faculty members, and supporting staff for their direct and indirect involvement in the successful completion of the project for their encouragement and support.

**VARSHINI G**

# ABOUT THE INTERNSHIP ORGANISATION – YUVAINTERN

Yuva Intern is India’s first job simulation internship platform, designed to bridge the gap between theoretical knowledge and real-world job responsibilities. By partnering with NSDC (National Skill Development Corporation) and industry experts, they have created a unique program that brings hands-on experiences and skills. "Yuva Intern" is led by Prabhjyot Kaur Juneja, Board Advisor at Henry Harvin Education.

## Mission

“Our mission at Yuva Intern is to revolutionize the internship model by making it accessible, practical, and rewarding. We aim to eliminate the guesswork and disillusionment that often come with entry-level roles. By focusing on real KRAs (Key Responsibility Areas) rather than mundane tasks, we want to ensure our interns acquire the core competences and hands-on understanding needed to excel in their future careers. Ultimately, we strive to empower the youth of India—one simulation internship at a time.”

## Vision

“We envision a future where every student and aspiring professional can explore diverse career paths without barriers—be it geographical, financial, or experiential. By seamlessly integrating technology with expert-designed simulations, we hope to create a robust ecosystem of skilled, aware, and confident professionals who contribute meaningfully to the workforce. Our vision extends beyond internships: we aim to spark a nationwide movement of experiential learning and career clarity.”

# ABSTRACT

This project uses data analysis techniques to uncover which factors affect the automation probability of professions by Artificial Intelligence (AI) in the near future. A dataset containing relevant information such as different jobs and its mean salary, educational experience, AI exposure index, automation probability, etc., was sourced from Kaggle. The data analysis processes were done by using Jupyter Notebooks, a web-based interactive environment. It supports the Python programming language, and its libraries — Pandas, SciPy, Matplotlib, etc.

The main hypothesis conjured before starting the data analysis tasks stated that mean salary and educational experience required for that particular profession was responsible for its probability of automation. The initial exploratory analysis was carried out to find relations between the data points by plotting different charts. The conclusions derived from this analysis showed that level of education did not directly influence automation probability in this dataset. Hence, another suitable factor was considered, namely the professions' AI exposure index. Further analysis done proved that mean salary and AI exposure index did influence a profession's automation probability.

# TABLE OF CONTENTS


| CHAPTER NO | TITLE | PAGE NO |
|---|---|---|
| | **ABSTRACT** | *i* |
| | **LIST OF FIGURES** | *iii* |
| I | **INTRODUCTION** | 1 |
| II | **PROJECT OBJECTIVES** | 2 |
| III | **TOOLS AND METHODOLOGY SELECTION** | 3 |
| IV | **DATA PREPARATION** | 4 |
| | 4.1 DATA ACQUISITION | |
| | 4.2 DATASET CLEANING | |
| | 4.3 DATA TRANSFORMATION | |
| V | **EXPLORATORY DATA ANALYSIS** | 7 |
| | 5.1 AVERAGE SALARY VS. YEARS OF EXPERIENCE | |
| | 5.2 AVERAGE SALARY VS. LEVEL OF EDUCATION | |
| | 5.3 AVERAGE SALARY VS. PROFESSIONS | |
| | 5.4 AI EXPOSURE VS. PROFESSIONS | |
| | 5.5 PROBABILITY OF AUTOMATION ACROSS PROFESSIONS | |
| | 5.6 FURTHER ANALYSIS | |
| VI | **CONCLUSION** | 14 |
| | **REFERENCES** | 15 |
| | **APPENDIX** | 16 |

# LIST OF FIGURES


| FIGURE NO | FIGURE NAME | PAGE NO |
|---|---|---|
| 1 | Columns present in the dataset | 4 |
| 2 | Checking for null and duplicated values | 5 |
| 3 | Code snippet of data transformation using z-score | 6 |
| 4 | Bar Graph representing average salary by years of experience | 7 |
| 5 | Graph representing average salary by years of experience for a Financial Analyst | 8 |
| 6 | Graph representing Average Salary vs. Education Level for a Financial Analyst | 9 |
| 7 | Code snippet showing the frequency of people with different educational qualifications | 9 |
| 8 | Line Graph depicting average salary vs. professions | 10 |
| 9 | Line graph depicting AI exposure vs. professions | 11 |
| 10 | Line graph depicting AI automation probability vs. professions | 12 |
| 11 | A heatmap showing the proficiency of skills in each profession | 13 |

# CHAPTER I

# INTRODUCTION

Data analysis is the process of exploring datasets to analyse patterns in the data and uncover useful information. A dataset contains raw information, which can have redundant and inconsistent values. It is essential that data must be cleaned before starting the data analysis process. This ensures the conclusions extracted from the dataset are true and faithful. Data analysis offers great value in the business sector, where optimisation of resources and labour can lead to better results with greater efficiency.

The use of Artificial Intelligence (AI) for automation of jobs presents opportunities for producing optimal output solutions. AI is currently being utilised for automating a wide variety of tasks, such as resource management and inventorying, prediction of sales by using current customer input and for creation of new tools like AI powered chatbots. Furthermore, its usage transcends from being a tool for optimising economic growth. There are innovative AI powered products used for diagnostic measures and research purposes in the pharmaceutical industry. The aim of this project is to analyse the impact of AI on professions across various sectors, and predict its risk of automation.

# CHAPTER II
# PROJECT OBJECTIVES

The objective of this project is centred around predicting which types of jobs are most likely to be AI automated in the near future. This is to be achieved by finding a pattern among professions that have a high probability of AI automation.

Most professions base their average salary based on factors such as the required educational qualifications and years of experience in the field. Entry level positions are already automated by AI products, such as handling data entry, check-in and booking systems, and other clerical operations. This leads to a hypothesis that professions with low-paying salary and repetitive tasks would be the first to be AI automated.

Professions requiring a high level of education are generally favoured as they require greater skills and offer more job security. They can be considered at less risk of AI automation compared to entry level jobs. This can also be rephrased to ask the question – “Are professions which require lesser levels of education more likely to be AI automated?”

# CHAPTER III

# TOOLS AND METHODOLOGY SELECTION

Data analysis will be done with Python 3, by using Jupyter notebooks. Jupyter notebooks is a cell-based environment, which offers output visualisation and documentation in a single document. It is best suited for data analysis, offering easy accessibility for collaboration.

The language used will be Python, as it offers a huge collection of libraries, and is the most preferred language for data analysis. The required libraries will be Pandas, SciPy and Matplotlib. These libraries serve as a foundation for data analysis using Python.

- **Pandas** - The Pandas library provides data structures to work with tabular data known as DataFrames. This makes data cleaning and filtering easier. It also provides methods for data aggregation.
- **SciPy** – It is built on top of the NumPy library. It offers an extensive collection of high-level numerical algorithms and functions used in statistics. This is useful for normalising data and other data transformations.
- **Matplotlib** - It allows data to be represented graphically by plots and graphs. It can provide data analysis to be more intuitive. Some commonly used plots are line graphs, bar graphs, histograms, among others. It also offers more complex plots and 3D graphs.

# CHAPTER IV
# DATA PREPARATION

## 4.1 DATA ACQUISITION

The dataset was sourced from Kaggle, an open-source data analysis platform. The dataset chosen for this project has the required data necessary for proceeding with the data analysis tasks. The main objective is to find common trends among jobs with high probability of automation by AI.

The dataset provides job listings along with its average salary, experience required, educational qualifications, AI exposure index, technology growth index, and its probability of AI automation in the future. The average salary of each profession can be used to uncover the pattern behind their automation probability. The same can also be done to find trends in educational experience.

| | Job_Title | Average_Salary | Years_Experience | Education_Level | AI_Exposure_Index | Tech_Growth_Factor | Automation_Probability_2030 | Risk_Category | Skill_1 | Skill_2 | S |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Security Guard | 45795 | 28 | Master's | 0.18 | 1.28 | 0.85 | High | 0.45 | 0.10 | |
| 1 | Research Scientist | 133355 | 20 | PhD | 0.62 | 1.11 | 0.05 | Low | 0.02 | 0.52 | |
| 2 | Construction Worker | 146216 | 2 | High School | 0.86 | 1.18 | 0.81 | High | 0.01 | 0.94 | |
| 3 | Software Engineer | 136530 | 13 | PhD | 0.39 | 0.68 | 0.60 | Medium | 0.43 | 0.21 | |
| 4 | Financial Analyst | 70397 | 22 | High School | 0.52 | 1.46 | 0.64 | Medium | 0.75 | 0.54 | |
| 5 | AI Engineer | 92592 | 11 | Master's | 0.29 | 0.51 | 0.10 | Low | 0.71 | 0.79 | |
| 6 | Mechanic | 107373 | 23 | PhD | 0.67 | 1.09 | 0.41 | Medium | 0.56 | 0.38 | |
| 7 | Teacher | 53419 | 12 | High School | 0.20 | 1.40 | 0.17 | Low | 0.56 | 0.70 | |
| 8 | HR Specialist | 139225 | 12 | Master's | 0.30 | 0.61 | 0.48 | Medium | 0.22 | 0.42 | |
| 9 | Customer Support | 85016 | 2 | High School | 0.01 | 1.01 | 0.80 | High | 0.22 | 0.12 | |
| 10 | UX Researcher | 82733 | 6 | High School | 0.50 | 0.80 | 0.41 | Medium | 0.04 | 0.61 | |
| 11 | Financial Analyst | 117455 | 22 | High School | 0.67 | 1.26 | 0.40 | Medium | 0.73 | 0.37 | |
| 12 | Lawyer | 79811 | 27 | High School | 0.68 | 0.52 | 0.50 | Medium | 0.23 | 0.65 | |
| 13 | Data Scientist | 115981 | 9 | High School | 0.26 | 1.16 | 0.63 | Medium | 0.56 | 0.53 | |
| 14 | Research Scientist | 96690 | 19 | Master's | 0.89 | 1.28 | 0.21 | Low | 0.08 | 0.16 | |

Figure 1. Columns present in the dataset

## 4.2 DATASET CLEANING

The dataset was made available in the format of a comma-separated values (CSV) file. It was downloaded and stored in the system, and accessed in Jupyter Notebooks by specifying the file path. The dataset was checked to find null values and duplicated values. Any errors in the data were subsequently removed.

```
df.duplicated().sum()
[4]  ✓ 0.0s                                                    Python
...  np.int64(0)

df.isna().sum()
[5]  ✓ 0.0s                                                    Python
...  Job_Title                      0
     Average_Salary                 0
     Years_Experience               0
     Education_Level                0
     AI_Exposure_Index              0
     Tech_Growth_Factor             0
     Automation_Probability_2030    0
     Risk_Category                  0
     Skill_1                        0
     Skill_2                        0
     Skill_3                        0
     Skill_4                        0
     Skill_5                        0
     Skill_6                        0
     Skill_7                        0
     Skill_8                        0
     Skill_9                        0
     Skill_10                       0
     dtype: int64
```

Figure 2. Checking for null and duplicated values

## 4.3 DATA TRANSFORMATION

The dataset did not contain any null and duplicated values. The next step in the data cleaning process is to identify the outlier values present in the dataset. The z-score is a statistical measure that measures how many standard deviations the data-point lies outside the mean range of the dataset. Data points with z-score outside the range (-3, 3) are considered outliers, and have to be investigated to ensure the validity of the dataset. The z-score is calculated by using the formula,

$$Z = (data\ point - mean) \div standard\ deviation$$

The SciPy library in Python includes the package *stats* which provides in-built functions for statistical analysis. The calculation for z-score is done by implementing the libraries in the notebook. The shape of the dataset is also checked after transforming the data.

```
numeric_cols = df.select_dtypes(include=['float64', 'int64'])
df = df[(np.abs(stats.zscore(numeric_cols)) < 3).all(axis=1)]
```
[6] ✓ 0.0s Python

```
df.info()
```
[8] ✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 18 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Job_Title                    3000 non-null   object 
 1   Average_Salary               3000 non-null   int64  
 2   Years_Experience             3000 non-null   int64  
 3   Education_Level              3000 non-null   object 
 4   AI_Exposure_Index            3000 non-null   float64
 5   Tech_Growth_Factor           3000 non-null   float64
 6   Automation_Probability_2030  3000 non-null   float64
 7   Risk_Category                3000 non-null   object 
 8   Skill_1                      3000 non-null   float64
 9   Skill_2                      3000 non-null   float64
 10  Skill_3                      3000 non-null   float64
 11  Skill_4                      3000 non-null   float64
 12  Skill_5                      3000 non-null   float64
 13  Skill_6                      3000 non-null   float64
 14  Skill_7                      3000 non-null   float64
 15  Skill_8                      3000 non-null   float64
 16  Skill_9                      3000 non-null   float64
 17  Skill_10                     3000 non-null   float64
dtypes: float64(13), int64(2), object(3)
memory usage: 422.0+ KB
```

Figure 3. Code snippet of data transformation using z-score

# CHAPTER V

# EXPLORATORY DATA ANALYSIS

## 5.1 AVERAGE SALARY VS. YEARS OF EXPERIENCE

```
plt.figure(figsize=(11,5))
plt.bar(df['Years_Experience'], df['Average_Salary'], color = 'seagreen')
plt.xlabel('Years of Experience')
plt.ylabel('Average Salary')
plt.title('Average Salary by Years of Experience')
plt.show()
```

[9] ✓ 2.3s Python


Figure 4. Bar Graph representing average salary by years of experience

The above figure shows the variations in salary as the years of experience increases. This relation takes the average salary across all professions, therefore not much variation can be observed in the graph. There will be more relevant information if one profession is used to check the above relation.

Upon further scrutinization, the dataset has many repeating job listings. The repeated listings can be treated as insights from different workplaces. The dataset has to be filtered to find the unique job titles. After filtering, the dataset was found to contain twenty job titles. From this, one profession can be selected to represent the salary-to-work experience relationship. The profession selected for analysis is that of a financial analyst.

```
plt.figure(figsize=(11,5))
plt.bar(avg_by_exp_finance['Years_Experience'], avg_by_exp_finance['Average_Salary'], color = 'turquoise')
plt.xlabel('Years of Experience')
plt.ylabel('Average Salary')
plt.title('Average Salary by Years of Experience')
plt.show()
```

[14] ✓ 0.1s Python


Figure 5. Graph representing average salary by years of experience for a Financial Analyst

The above graph shows more variation in the data. The salary is, however, not strictly increasing, which can be explained by other factors, such as the data being sourced from different companies, differing number of people, and individuals with different levels of education.

## 5.2 AVERAGE SALARY VS. LEVEL OF EDUCATION

```python
plt.figure(figsize=(11,5))
plt.bar(filtered_finance['Education_Level'], filtered_finance['Average_Salary'], color = 'purple')
plt.xlabel('Education Level')
plt.ylabel('Average Salary')
plt.title('Average Salary by Level of Educationfor Financial Analyst')
plt.show()
```

[15] ✓ 0.2s Python

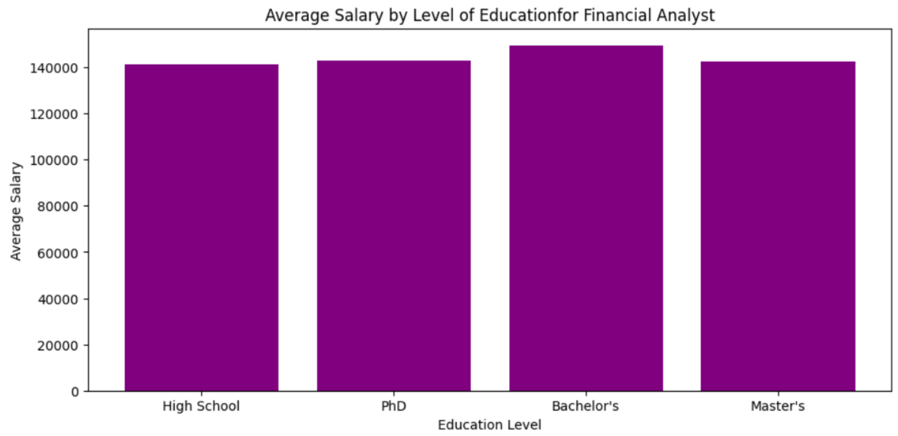


Figure 6. Graph representing Average Salary vs. Education Level for a Financial Analyst

From this graph, it can be inferred that individuals with a Bachelor degree earn higher salary than those with Master’s or a PhD. This can imply that more people find jobs after finishing their Bachelor’s degree, instead of pursuing higher education. Financial analysts with a PhD might earn more compared to others, but might be fewer in number, which can skew the results. Therefore, another relationship to explore would be to compare the population size based on their education level. The following code snippet tests this theory,

```python
filtered_finance['Education_Level'].value_counts()
```

[16] ✓ 0.0s Python

```
Education_Level
Bachelor's     56
High School    32
PhD            32
Master's       31
Name: count, dtype: int64
```

Figure 7. Code snippet showing the frequency of people with different educational qualifications

This shows that the dataset does not have an equal distribution of people having different years of experience working in the field. Furthermore, the level of education influences the mean salary of the individual, but no further relation can be made with respect to automation probability. Therefore, another parameter can be considered to measure probability of AI automation, namely the index of AI exposure in the workplace.

The dataset has already been explored to find trends in average salary and educational levels of people for a singular profession. The next step is to broaden the scope of analysis by checking for patterns in salary and the use of AI across all professions.

## 5.3 AVERAGE SALARY VS. PROFESSIONS

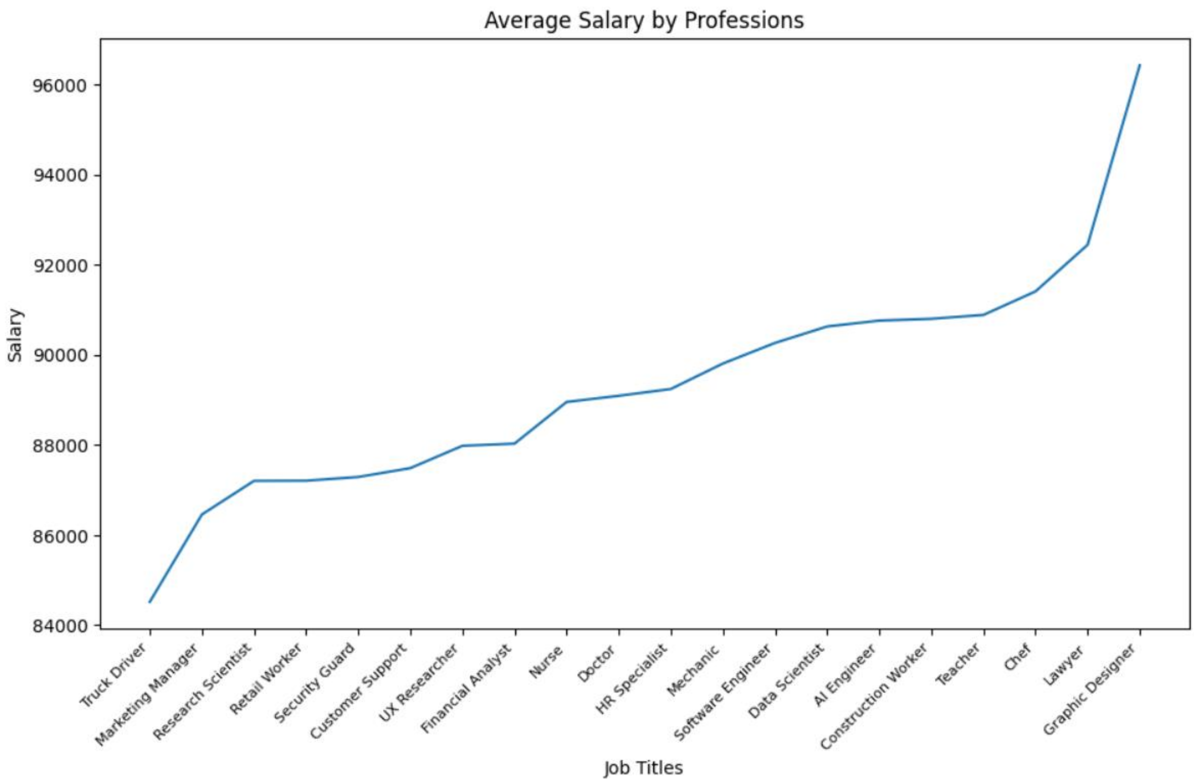


Figure 8. Line Graph depicting average salary vs. professions

The above graph shows the trends in mean salary across all professions in the dataset. The salary of twenty unique professions spans from 84,000 to 96,000 USD. According to the information available in the dataset, the salary of truck drivers is on the lower end of the spectrum, while lawyers and graphic designers have higher net worth. Low paying work might have a higher chance of being automated by AI, but this can also depend on the exposure of AI in the work field and how successfully it can be implemented. From the salary ranges across professions, a link to automation probability can be further derived.

## 5.4 AI EXPOSURE VS. PROFESSIONS

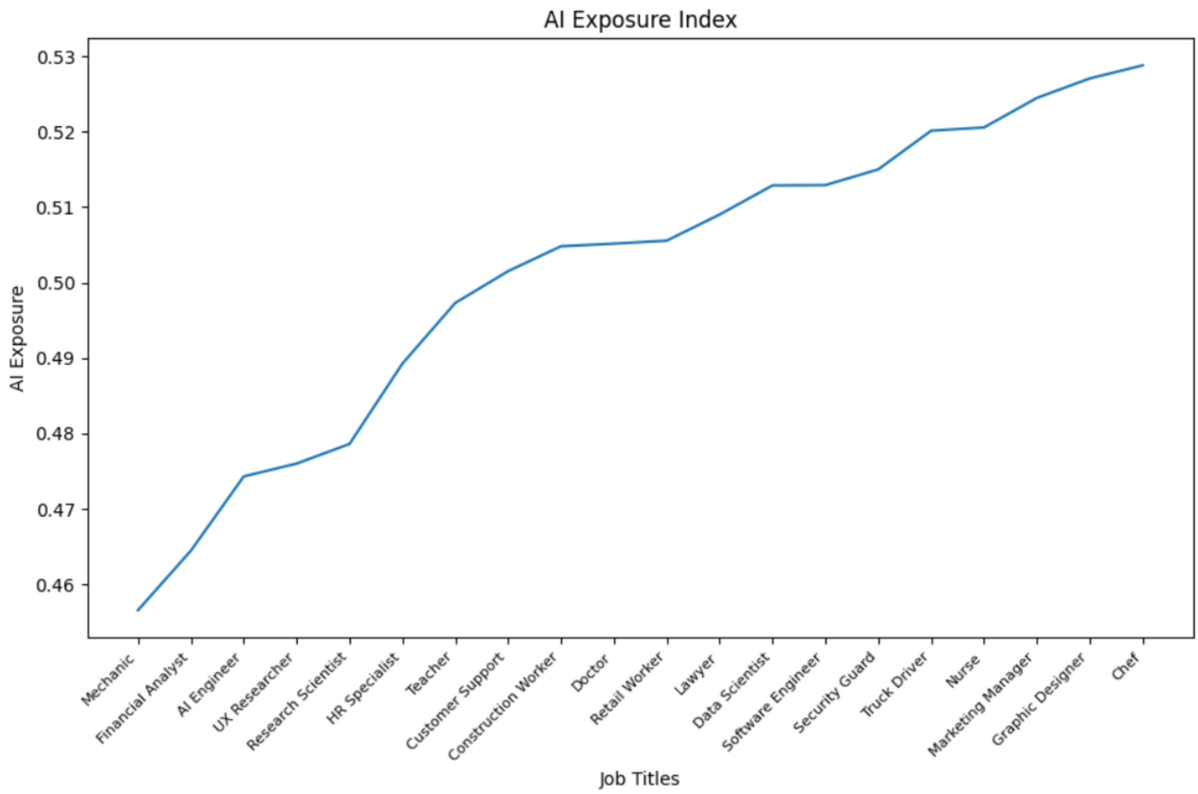


Figure 9. Line graph depicting AI exposure vs. professions

The exposure of AI tools and services can be noticed in many different fields of work, as seen in the above graph. AI exposure is measured as an index, with the values in the graph ranging from 0.46 to 0.53. It shows that mechanics and financial analysts have lesser AI exposure in their respective fields, and the professions with higher AI use in the current scenario are graphic designers and chefs.

Professions with successful implementation of AI into their field are more likely to be fully AI automated. This would significantly reduce labour costs, and the overall efficiency would be reflected in its output. This also implies that professions with higher AI exposure cannot be fully expected to be AI automated – if the implementation of AI tools is not entirely feasible for future development, and shows to be not sustainable to generate the desired results.

## 5.5 PROBABILITY OF AUTOMATION ACROSS PROFESSIONS

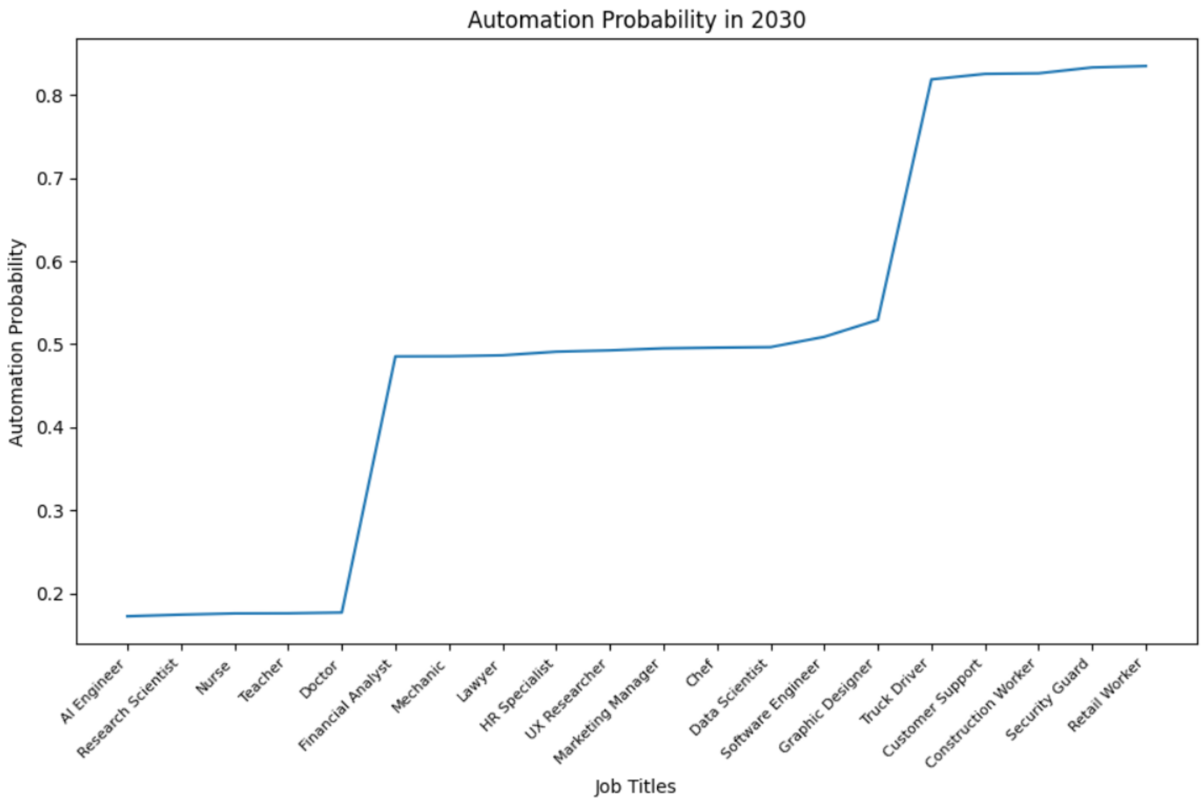


Figure 10. Line graph depicting AI automation probability vs. professions

This graph shows the visual representation of the predictive AI automation probabilities across jobs which were provided in the dataset. The probability values range from 0 to 1, with many jobs clustered around the ranges 0.1, 0.5 and 0.8 to 0.9. Many jobs have an automation probability of 0.5, with varying salaries and AI exposure. The data analysis can be made more efficiently by analysing the professions with extreme probability values.

By analysing professions with least probability of automation, it can be noted that they have substantial salaries, ranging from 88,000 to 91,000 USD. Some of these jobs have higher AI use, but this can be reasoned as unsustainable implementation of AI tools for proper automation.

The average salaries of professions with an automation probability above 0.8 are on the lower end, ranging from 84,000 to 87,000 USD. These jobs also have an AI exposure index above 0.5, indicating higher use of AI tools and services. This exposure can lead to full automation of these professions in the near future.

## 5.6 FURTHER ANALYSIS

The dataset also contains ten values indicating the proficiency of ten different skills per person in each job listing. Though these values do not influence the probability of automation of these jobs, the given data can be visualised in the form of a heatmap. The heatmap was generated with the use of seaborn, a python library that is also used to plot specialised graphs.

```
plt.figure(figsize=(8,6))
sns.heatmap(jobs_by_skill)
plt.ylabel(None)
plt.show()
```

[24] Python

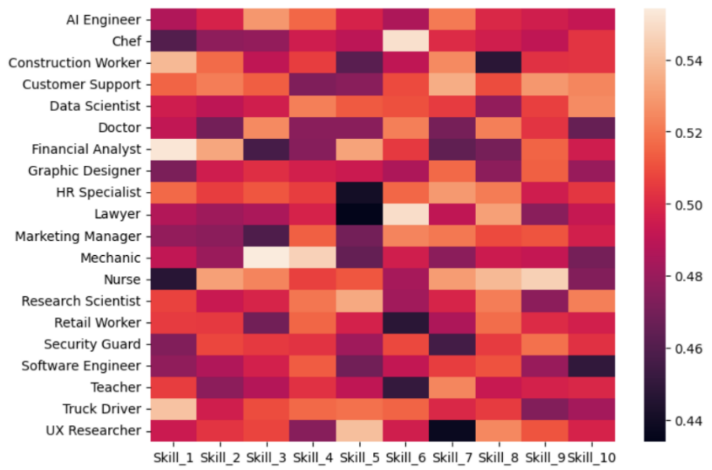


Figure 11. A heatmap showing the proficiency of skills in each profession

# CHAPTER VI
# CONCLUSION

Through rigorous data analysis, a pattern was identified by using the values of mean salary in the workplace and an index representing exposure of AI tools in that field. The initial hypothesis of this project stated a correlation between automation probability and the individual's mean salary and level of education. This was further verified by plotting several bar charts. It showed that the dataset did not have an equal distribution of people having different years of experience working in the field. Furthermore, the level of education impacted the mean salary of the individual, but no further relation could be made with respect to automation probability.

Therefore, another parameter was considered to measure probability of AI automation, namely the index of AI exposure in the workplace. By using these values, several line plots were made to find patterns in automation probability. From these graphs, some definitive conclusions were drawn; mean salary and AI exposure did affect the probability of automation by AI.

Furthermore, it was analysed that the education level of a person affected their salary per annum. Similarly, the relationship between mean salary and automation probability could be further studied. Within each profession, different people can have different work assigned to them based on their educational levels. If salary is affected by the level of education, then it is entirely possible that automation probability and mean salary are closely related.

# REFERENCES


[1] Kaggle dataset, AI impact on Jobs 2030:
https://www.kaggle.com/datasets/khushikyad001/ai-impact-on-jobs-2030/data

[2] Project Notebook by Varshini G:

https://github.com/varshini-g-07/intern-project-data-analysis/blob/main/project.ipynb

[3] Referenced kaggle notebook by Ebrahim Esmail:

https://www.kaggle.com/code/romaa200/ai-impact-on-jobs-2030

[4] Pandas documentation: https://pandas.pydata.org/docs/

[5] SciPy documentation: https://docs.scipy.org/doc/scipy/

[6] Matplotlib documentation: https://matplotlib.org/stable/index.html

[7] Seaborn documentation: https://seaborn.pydata.org/