DATA ACQUISITION & PRELIMINARY ANALYSIS

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 Artificial Intelligence (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.

```
[3]  df.head()
```

Python

	Job_Title	Average_Salary	Years_Experience	Education_Level	AI_Exposure_Index	Tech_Growth_Factor	Automation_Probability_2030	Risk_Category	Skill_1	Skill_2	Sk
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	

Figure 1. Columns present in the dataset

DATA STORAGE & DATA 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 also duplicated values. Any errors in the data were subsequently removed.

```
[7]  df.duplicated().sum()
```

Python

```
...  np.int64(0)
```

Figure 2. Checking for duplicated values

```
[6]  df.isna().sum()

...  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 3. Checking for null values

Z-Score in Data Analysis

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.

```
[9]
numeric_cols = df.select_dtypes(include=['float64', 'int64'])
df = df[(np.abs(stats.zscore(numeric_cols)) < 3).all(axis=1)]

[10]
df.shape

(3000, 18)
```

Figure 4. Transforming the dataset

```
[4]
df.info()

<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 5. Dataset information

PRELIMINARY ANALYSIS

The dataset was explored by visualising the data in the form of bar charts. The relationship between the years of experience required and the average salary was studied across all professions.

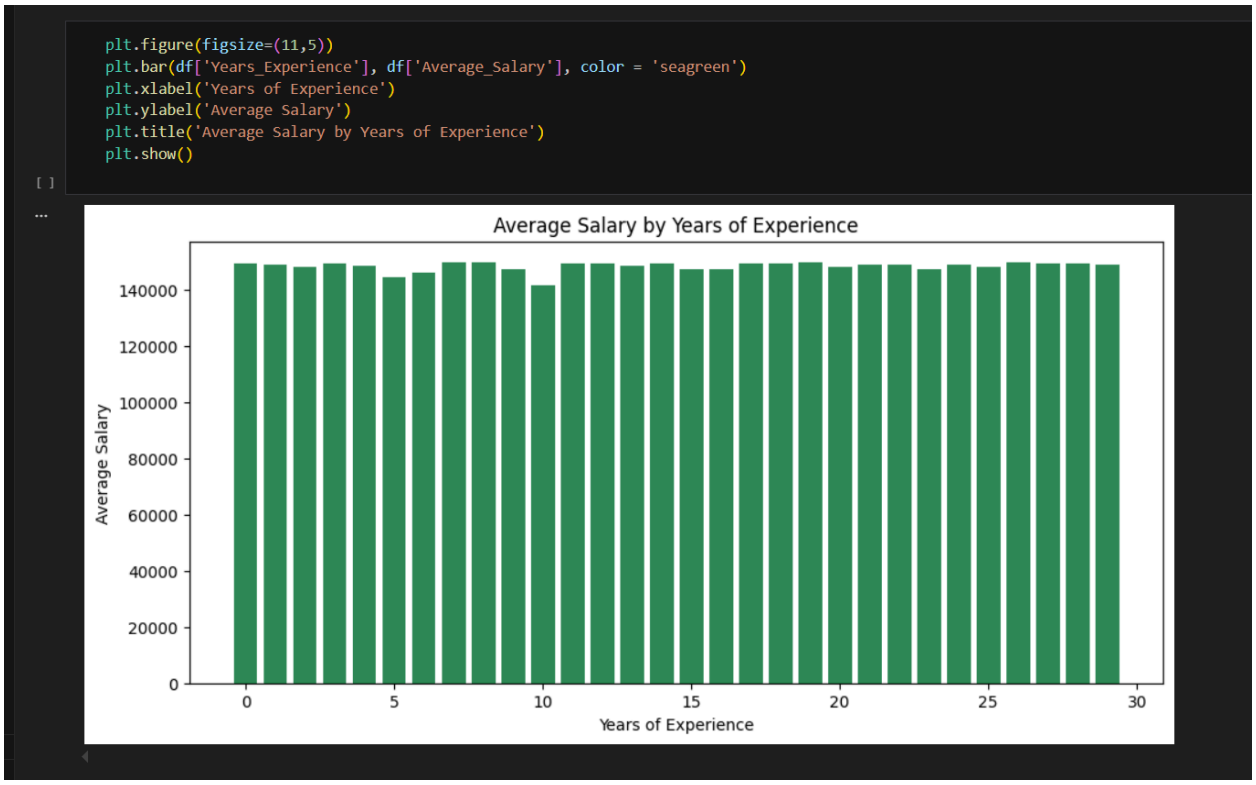

Figure 6. Bar Graph representing average salary by years of experience

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 bar graph below shows the relationship between salary and work experience for the profession of Financial Analyst.

```
[16]  filtered_finance = df.query("Job_Title == 'Financial Analyst'")
      ✓ 0.0s

      avg_by_exp_finance = filtered_finance.groupby('Years_Experience')['Average_Salary'].max().to_frame()
      avg_by_exp_finance.reset_index(inplace=True)
      #not used
[26]  ✓ 0.0s
```

Figure 7. Code snippet of filtering the job title 'Financial Analyst'

```
[25]  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()
      ✓ 0.2s
```

Figure 8. Bar 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.

The relationship to be studied next is that of educational qualifications and salary. The graph below represents this relationship.

```
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 Education for Financial Analyst')
plt.show()
```

[40] ✓ 0.2s

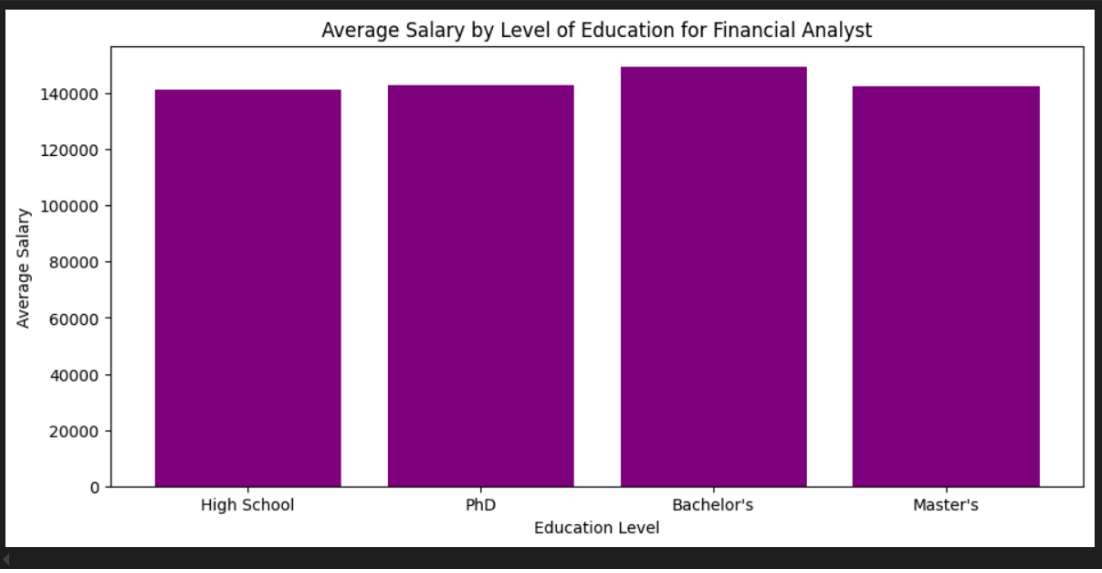

Figure 9. Graph representing Average Salary vs. Education Level for a Financial Analyst

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,

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

[43] ✓ 0.0s Python

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

Figure 10. Code snippet showing the frequency of people with different educational qualifications

Thus, the initial trends were studied for the role of a Financial Analyst. The next steps for analysis of the dataset would be to find patterns and trends in average salary and education level at a more granular level.