DATA EXPLORATION ROADMAP

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.

PROJECT OBJECTIVES

The objective of this project is centered 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?”

TOOL & 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, numpy 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.

- **Numpy -** It provides support for multi-dimensional arrays, along with methods for numerical computation on the matrices. It is very useful in data analysis as it can support statistical analysis of the dataset.

- **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.

DATA ACQUISITION

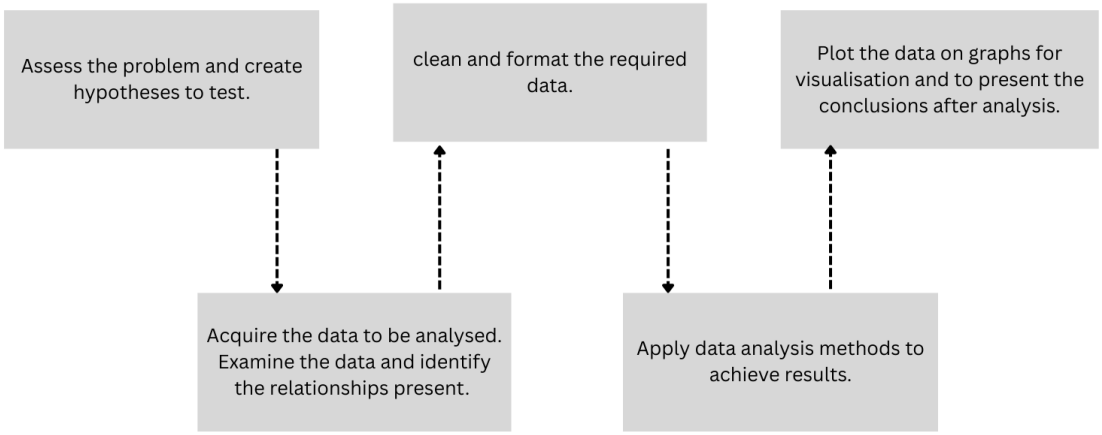
The data is to be acquired from Kaggle, an open source data science platform. The dataset to be selected must have unbiased data and also satisfy the following criteria,

- It must have various job titles across different sectors. This ensures that the data is properly distributed to realise the impact of AI across all professions.

- The dataset must also have the salary listings, required years of experience and educational qualifications.

- It should contain an index indicating the use of AI and other technologies in the profession. This can be used to identify jobs requiring greater support from AI tools.

ROAD MAP



The data analysis process across four weeks can be represented by using a road map to observe all milestones and checkpoints for the successful outcome of deliverables.

WEEK 1: Create the problem statement to be verified through data analysis. Outline the tools to be used and environment for analysis.

WEEK 2: Acquire the data to be studied through a reliable source, to ensure unbiased data. Perform data cleaning operations to remove any errors present in the dataset.

WEEK 3: Perform exploratory data visualisations by using Python and its libraries (Pandas, Numpy and Matplotlib). Document the observed findings and results.

WEEK 4: Discuss the overall strengths and weaknesses of the project and suggest optimisation techniques.

RISK ANALYSIS

Data which has not been processed to remove inconsistencies pose a challenge to the data analysis process. It can cause incorrect findings, which can lead to invalid conclusions. The dataset to be analysed must also not contain misinformation or biased data. These issues can be mitigated by cleaning the data to make it suitable for analysis.