JMAN GROUP

ML Training - Introduction

FT FINANCIAL TIMES

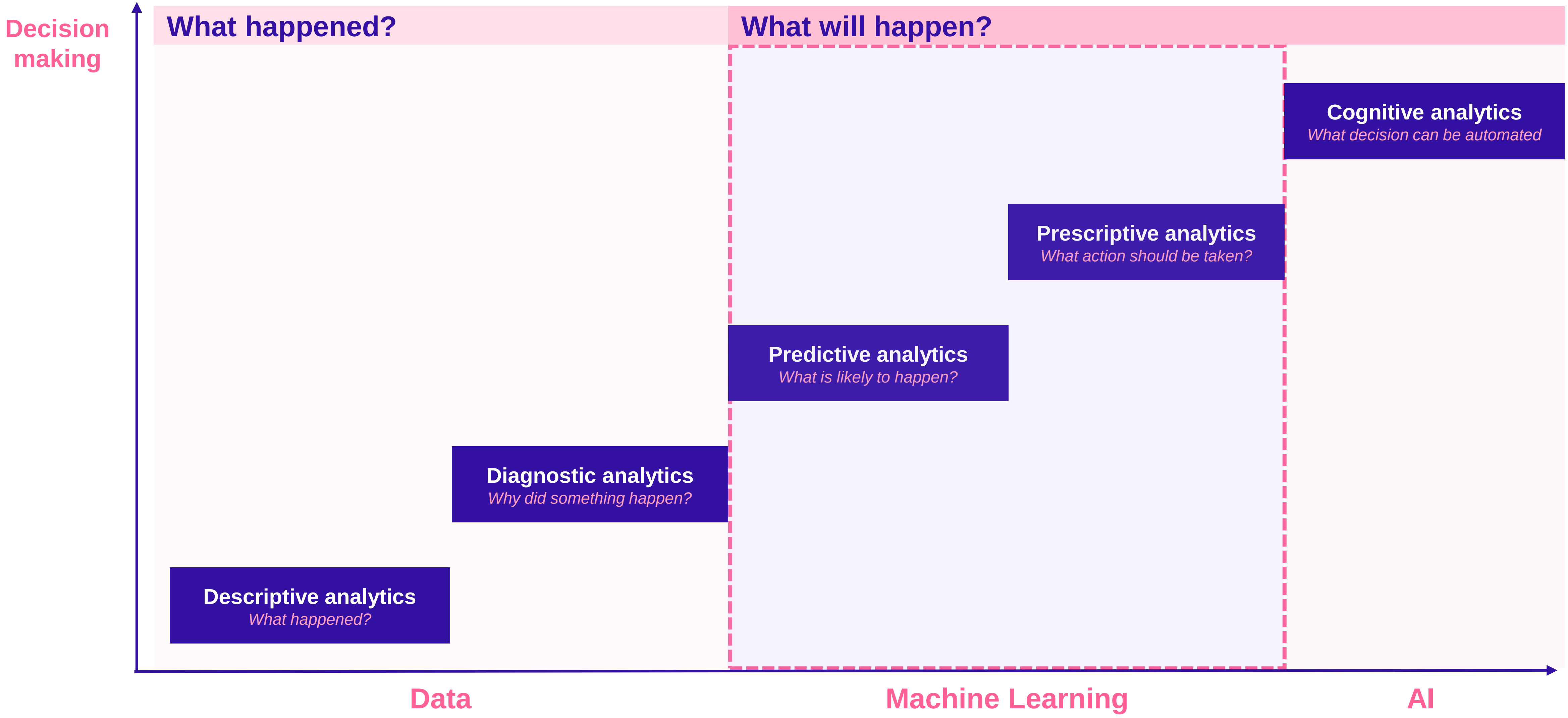
statista

UK'S LEADING MANAGEMENT CONSULTANTS 2022

Machine Learning Sessions

1. Introduction & EDA
2. Feature Engineering & Selection
3. Model Training & Evaluation
4. ML through commercial lens
5. ML Ops

Machine learning is one step further from data warehouse to automate decision making



Decision making

What happened?

What will happen?

Descriptive analytics
What happened?

Diagnostic analytics
Why did something happen?

Predictive analytics
What is likely to happen?

Prescriptive analytics
What action should be taken?

Cognitive analytics
What decision can be automated

Data

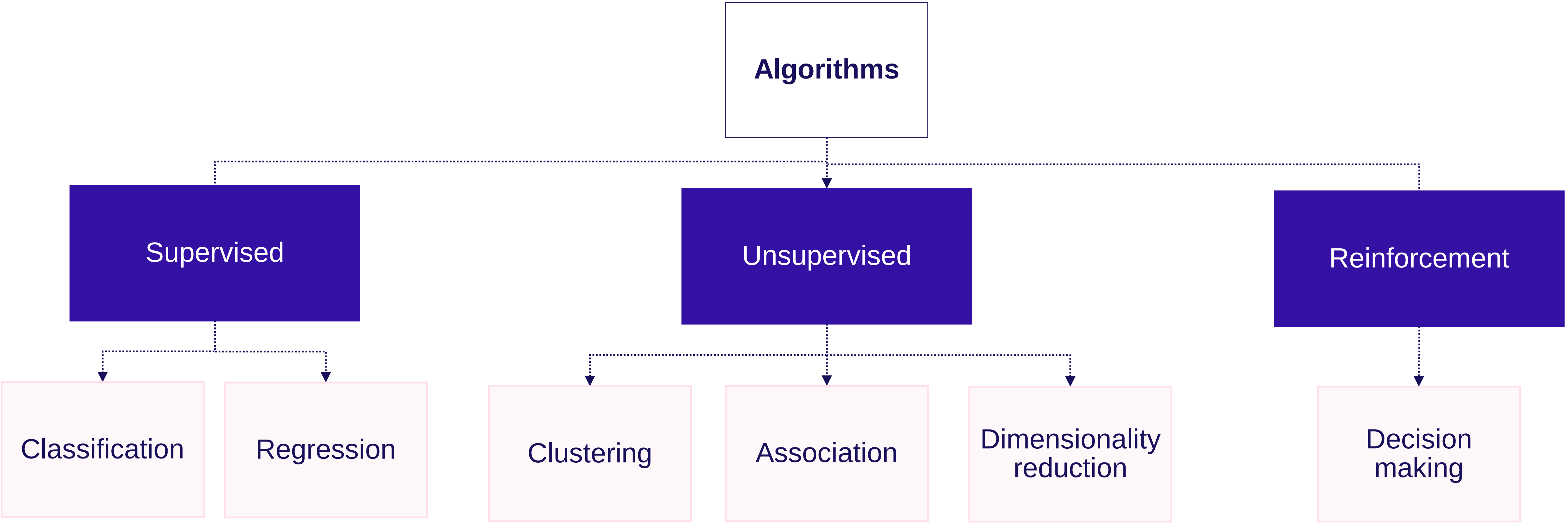
Machine Learning

AI

Why machine learning?

Machines can learn from data to **identify patterns** and **behaviours**, which can then be used to make **logical decisions** with less **human intervention**

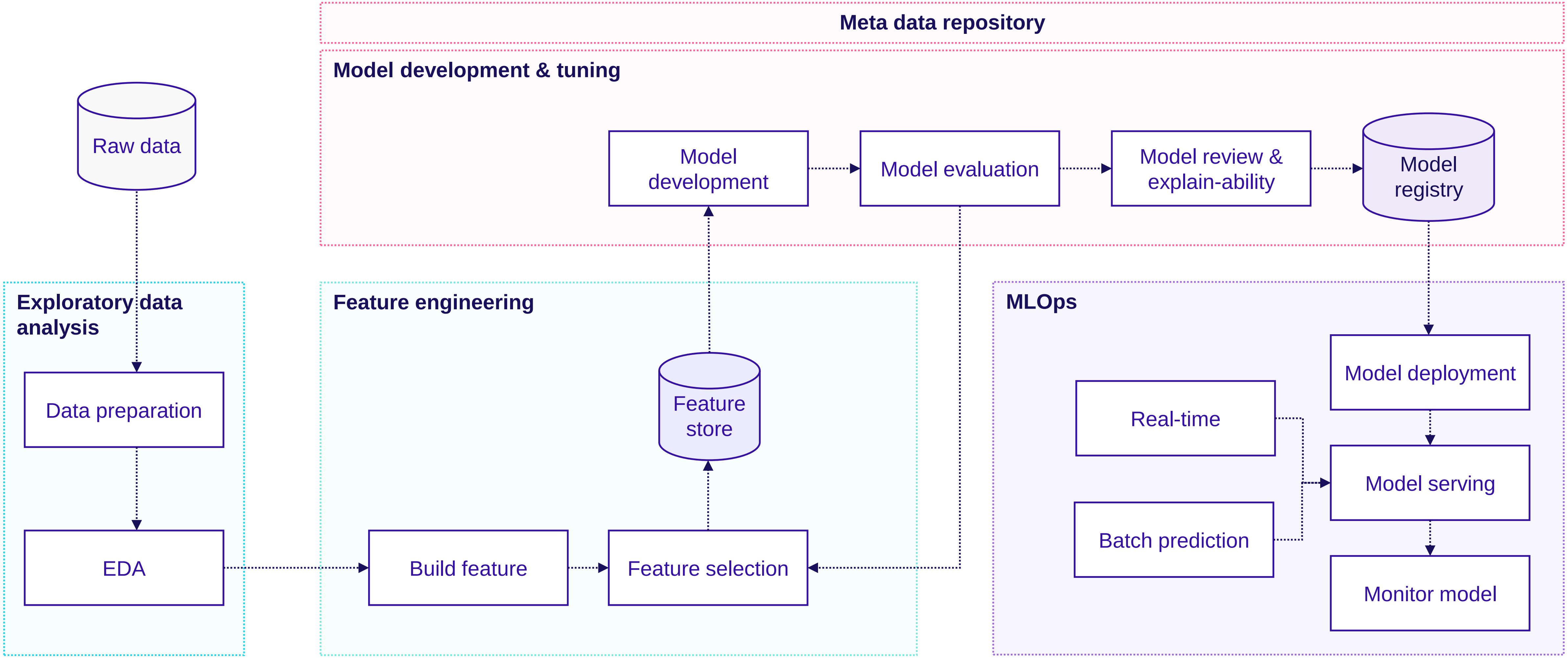
Machine learning algorithms can be broadly classified into three areas: supervised learning, unsupervised learning, and reinforcement learning



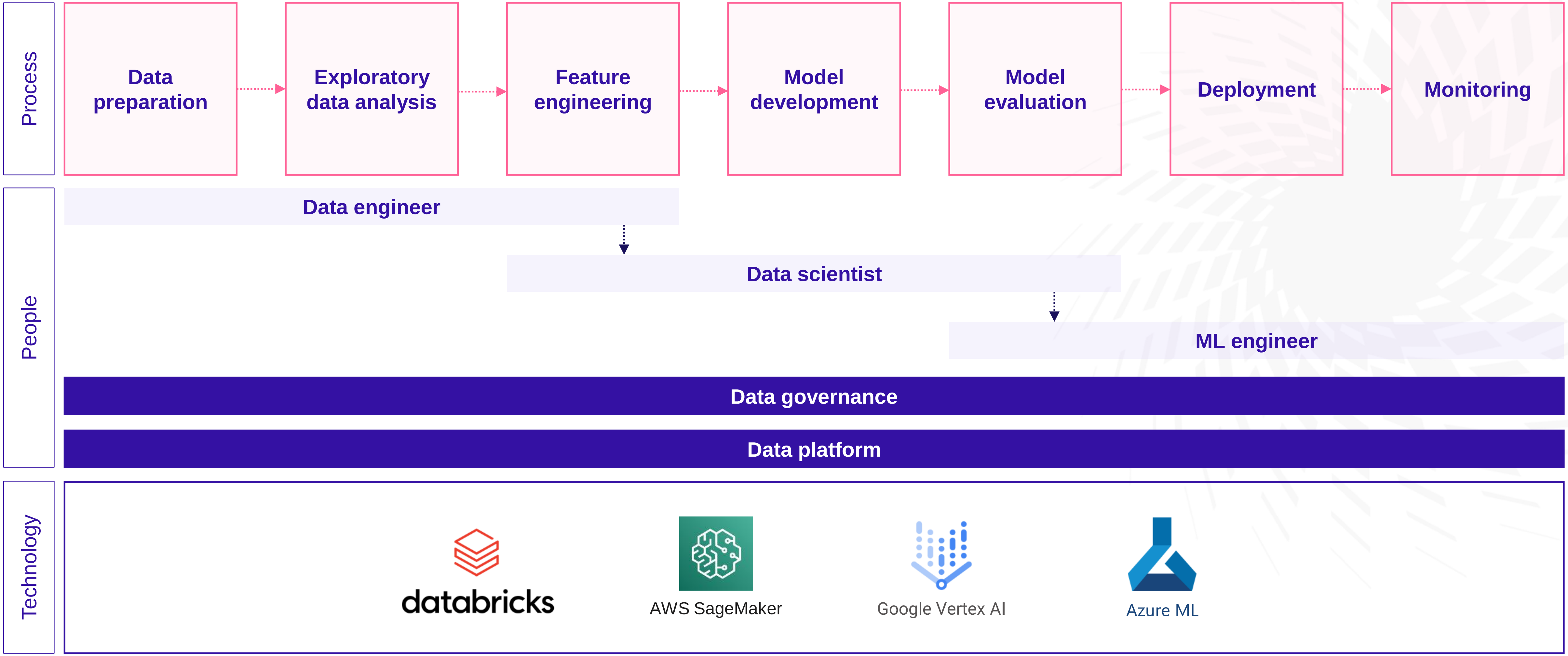
Use cases so far JMAN solved in machine learning

	Use case	*Approach*
1	**Churn prediction and compliance**	**Classification**
2	**Revenue forecasting**	**Regression**
3	**Lead conversion**	**Classification**
4	**Promotion analysis**	**Classification**
5	**Customer life-time value**	**Regression**

Machine learning end-to-end components



Process, People & Technology



Process

Data preparation

Exploratory data analysis

Feature engineering

Model development

Model evaluation

Deployment

Monitoring

People

Data engineer

Data scientist

ML engineer

Data governance

Data platform

Technology

databricks

AWS SageMaker

Google Vertex AI

Azure ML

Exploratory Data Analysis

FINANCIAL TIMES statista UK'S LEADING MANAGEMENT CONSULTANTS 2022

jmangroup.com

Context

1 Introduction

2 Data Preparation

3 Exploratory Data Analysis

4 About the data (Problem Statement)

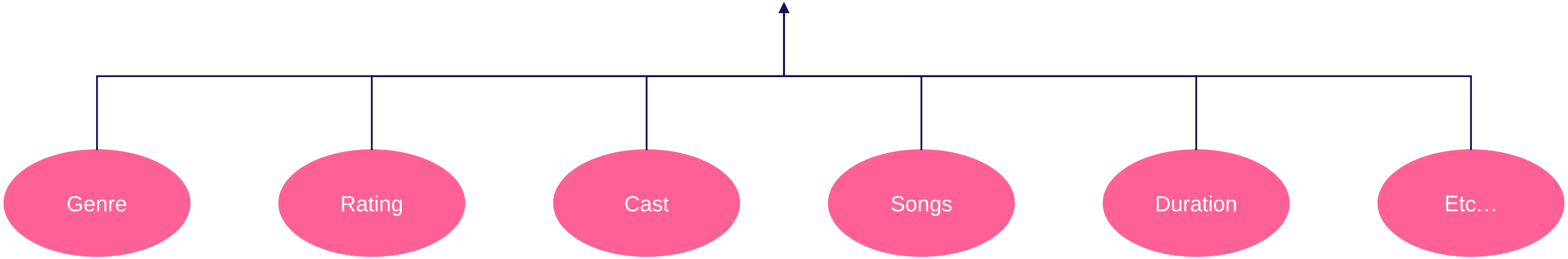
What is EDA?

Exploratory Data Analysis (EDA) in data science is the process of examining and visualizing a dataset to understand its main characteristics, patterns, and relationships. It involves using statistical methods and visualizations to uncover insights, identify outliers, and inform subsequent steps in data analysis.

Why we need EDA?

EDA helps in making helpful decisions

Scenario: You are planning a movie night and want to pick a movie everyone will enjoy. Each of them have a listed their potential movies everyone will like, and you want to make a decision based on certain criteria.



All the question mentioned above helps us in building a hypothesis on which movie will be best for the occasion and guess what this called **Exploratory Data Analysis (EDA)**

Context

1 Introduction

2 Data Preparation

3 Exploratory Data Analysis

Target and features variables

Feature

- Features are crucial for any machine learning task as they encode the information needed for the model to learn patterns and make predictions.
- For example, for describing a fruit we might use its **color**, **shape**, and **size** which will be knows as feature in this case .
- In the sense of data is nothing just **columns** in the dataset.
  - Example – all columns in our dataset except Churn, and Customer Id.

Target

- The target variable is the **variable of interest**, and it represents the outcome or result that the model is trained to predict.
- In the given dataset our target variable is **Churn.** Which represents whether a customer has left the company service or still using the service.

Types of features variables

Columns in the data are referred as feature in ML

1 Categorial Feature

Categorical features are data attributes with specific groups or labels, like names or numbers, used to group data into distinct categories.

Examples:

- Gender - Male/Female
- Whether the customer churned or not (Yes or No)

2 Numerical Feature

Numerical features represent values across a range, aiding in predicting or describing continuous data and phenomena with versatile precision.

Examples:

- The total amount charged to the customer
- The amount charged to the customer monthly

3 Date Time Feature

A date-time feature is a variable that represents a point in time, typically expressed as a combination of year, month, day, hour, minute, and second.

Examples:

Order placement date and time, e.g. "2022-02-17 14:30:00", is a date-time feature in a customer order dataset, enabling ML algorithms to identify patterns and relations in order timing.

Note: *We will be covering features in more details in upcoming sessions*

Data cleaning is essential for building correct assumptions on data

What is data cleaning?

- Cleaning and pre-processing data involves deleting null values and duplicates to assure the dataset's quality and accuracy.

Ways to do data cleaning:

- Analyse the proportion of null values and duplicates in relation to the total size of the dataset.
- Consider the possibility of data loss if a large amount of your data is missing or duplicated.
- Identify and handle outliers to prevent them from skewing analysis results.

Why is Data cleaning required?

- High-quality data leads to better model performance in machine learning and analytics tasks.
- Guarantees that the dataset is free from errors or inconsistencies, which can otherwise lead to incorrect insights.

Removing Null Values(Missing Data)

Removing Duplicate Values

Data Imputation

Technique used in Data pre-processing to fill in missing or incomplete values in a dataset

Mean, Median, Mode Imputation

- Replace missing values in the same feature/column with the mean (average), median (middle value), or mode (most frequent value).
- This method is straightforward and quick, but it may not capture complex data linkages.

Imputation using ML algorithms

- Missing values in sequential data can be interpolated using linear interpolation between nearby known values.
- Replace missing values in the feature space with values from the k-nearest neighbours.
- Using regression models trained on the remaining data, predict missing values.

Forward / Backward Fill

- To fill in missing values in time-series data, you can propagate the last observed value forward (forward fill) or the next observed value backward (backward fill).
- When missing values are likely to follow the trend of nearby data points, this function comes in handy.

Context

1 Introduction

2 Data Preparation

3 Exploratory Data Analysis

Hypothesis Building

A hypothesis is like a guess that you can check. It suggests a possible link between things, guiding your investigation during EDA by asking specific questions and looking for patterns in your data.

- **Understanding Data**
  - EDA begins with the development of hypotheses to guide your analysis.
  - Hypotheses are educated estimates regarding data linkages, patterns, or trends.
- **Diving Insights**
  - Hypotheses serve as road maps, directing your attention to key parts of the data.
  - Testing hypotheses can disclose useful information and guide further investigation.
- **Formulating Hypotheses**
  - Specific, testable hypotheses based on domain expertise should be developed.
  - They frequently involve variable relationships or group comparisons.
- **Iterative Process**
  - EDA is iterative, which means that hypotheses can change or evolve as new information is discovered.

Hypothesis Testing

• **Collecting Evidence**	• Gather data and explore visualizations during EDA to gather evidence for or against your hypothesis.
• **Visualization Tools**	• To identify patterns linked to your hypotheses, use histograms, scatter plots, box plots, and other visualizations.
• **Statistical Methods**	• To quantify relationships and establish the importance of your hypotheses, use proper statistical tests.
• **Interpretation**	• Based on the evidence collected, either accept or reject your hypotheses.
• **Refinement**	• If hypotheses are rejected, revise or reformulate them based on the analysis's new findings.

Data Visualization

Using graphical representations to communicate insights and patterns found within data

Helps in understanding complex data, identifying trends, and presenting information in a more understandable and actionable format.

1. **Bar Chart**
2. **Line Chart**
3. **Pie Chart**
4. **Histogram**
5. **Scatter Plot**
6. **Box Plot**

BAR CHART

When to Use?

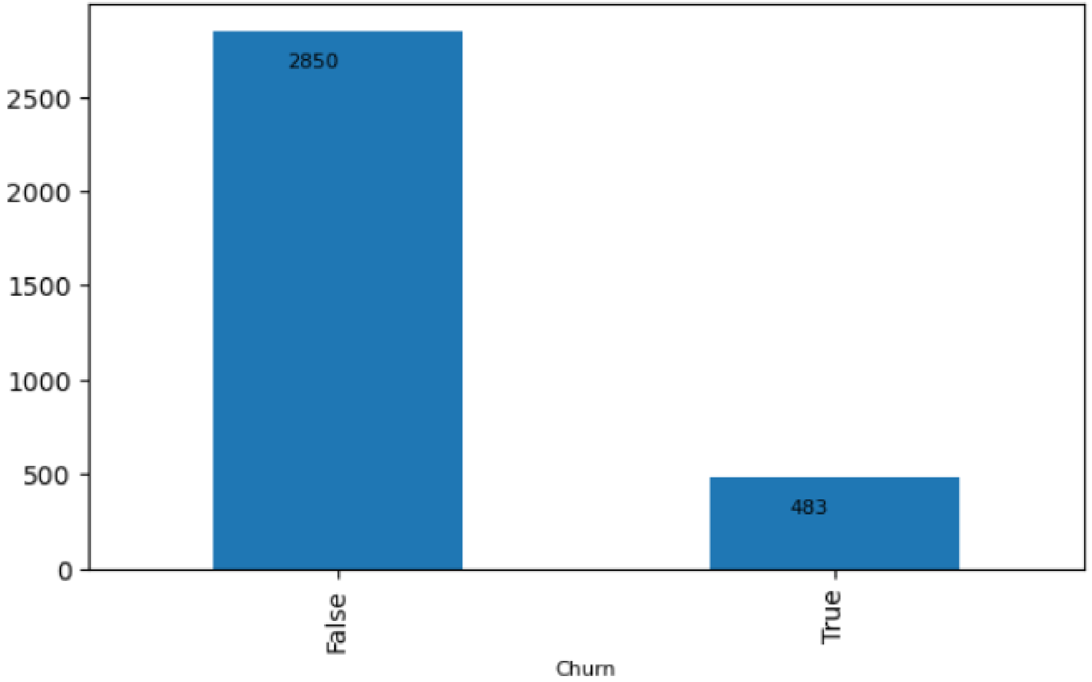
To compare the values of different categories. It is particularly useful for showing how different groups or items are distributed across categories.

Criteria for using a bar chart?

- Are you trying to compare the values of different categories?
- How many categories do you have?

Readability

- The vertical axis (y-axis) represents the values, while the horizontal axis (x-axis) displays the categories or items being compared.
- The length of each bar corresponds to the value it represents. Taller bars mean higher values.
- Categories are listed on the x-axis. Each bar is associated with a category.
- To compare values across categories, simply look at the height of the bars. Higher bars indicate larger values.



SCATTER PLOT

When to Use?

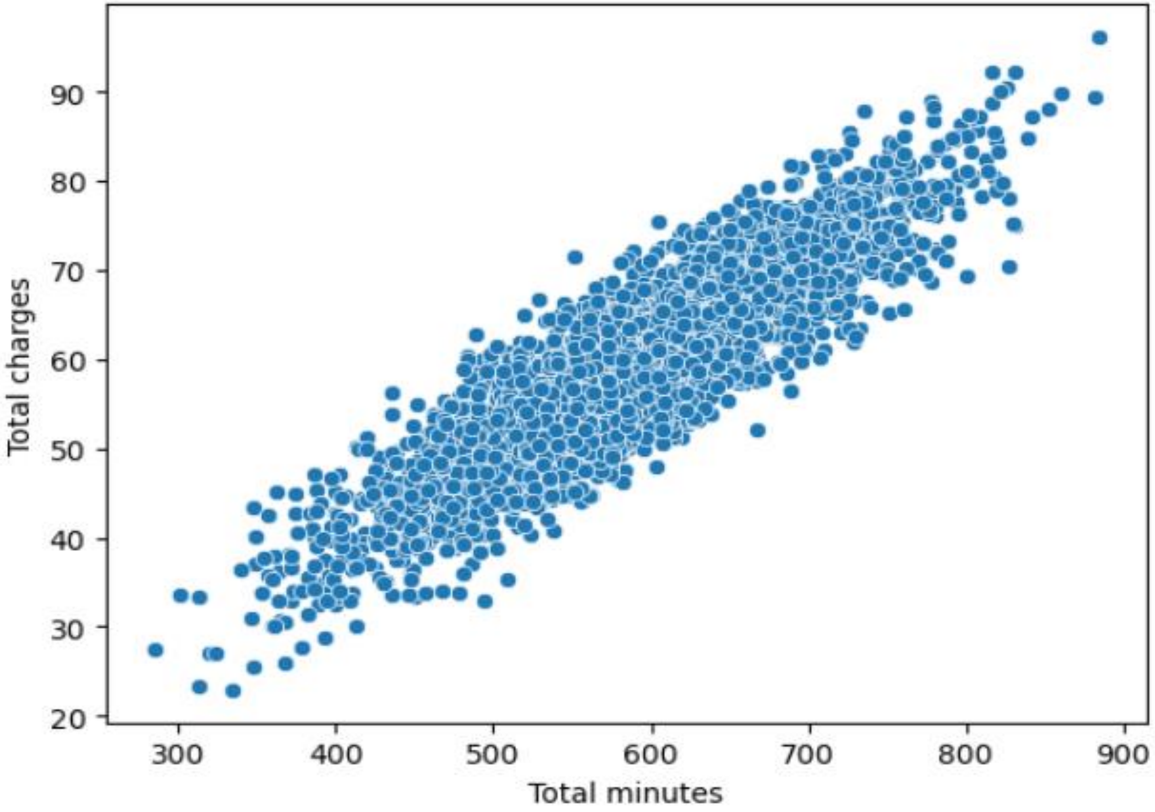
To visualize the strength and direction of a linear or nonlinear association between two variables.

Criteria for using a Scatter Plot?

- Are you trying to visualize how one variable affects another.?

Readability

- Direction: Check if points generally move upwards, indicating a positive relationship, downwards for a negative one, or are scattered with no clear trend for a lack of correlation.
- Spread: Assess how closely or widely points are distributed; a tight grouping suggests a strong correlation, while a more scattered arrangement indicates a weaker or no correlation.



BOX PLOT

When to Use?

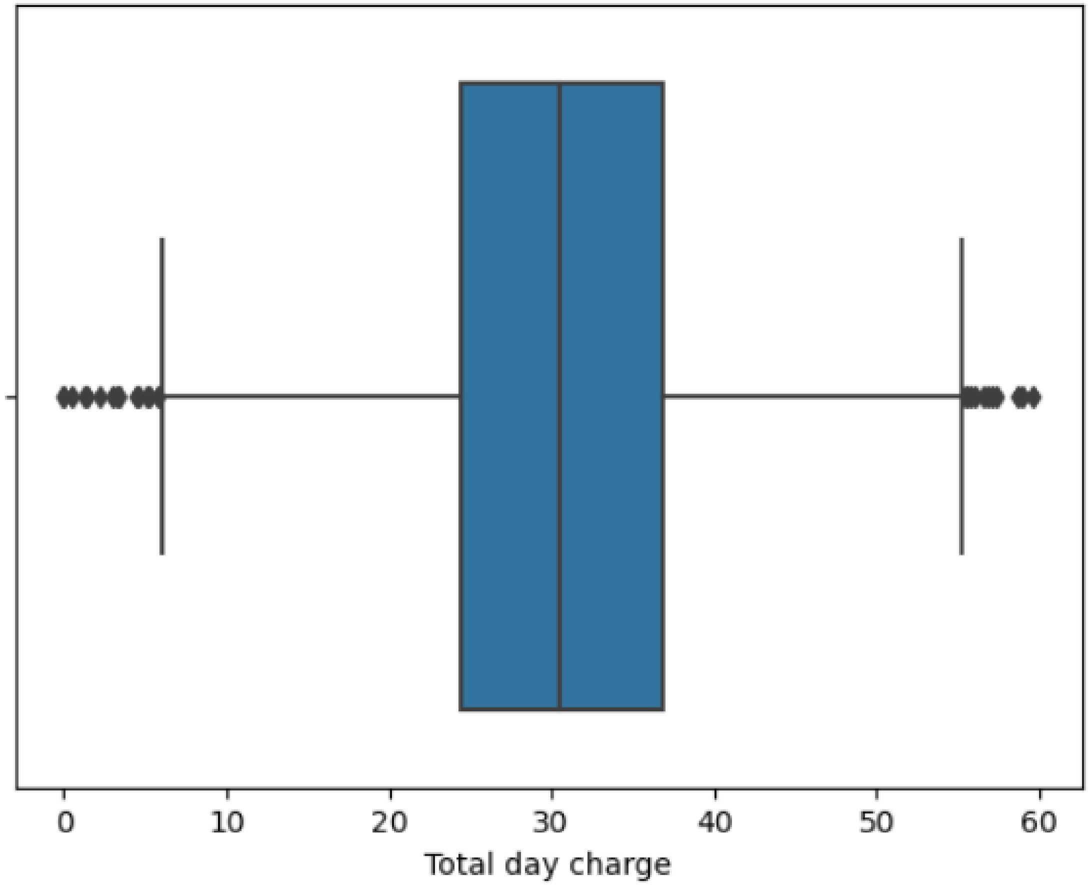
To visualize differences in the distribution of a variable between groups or to identify outliers within a group.

Criteria for using BOX PLOT?

- Are you trying to quickly identify outliers or unusual values in the data?

Readability

- Box: The box represents the interquartile range (IQR), containing the central 50% of the data points.
- Median Line: The line inside the box is the median, dividing the data into two equal halves.
- Whiskers: The lines extending from the box show the range of the data, excluding outliers.
- Outliers: Individual points outside the whiskers are potential outliers.



Problem statement

Telecom Churn analysis

Problem: The telecom industry faces a significant challenge with high customer churn rates, resulting in lost revenue and decreased market share.

The client have lots of data points which might be helpful to identify the customer behaviour. By understanding these drivers, we develop targeted interventions to retain customers and increase overall satisfaction.

Assumption

By looking at the Problem statement we can make few assumption on why customer could be churning:

- Customers who have been with the telecom company for a longer duration are less likely to churn.
- Customers who recently upgraded their plans or added new services are less likely to churn.
- Customers who have received special offers or discounts are less likely to churn.
- Churn rates vary between different service areas or regions.