

| **ePROJECT DATA SCIENCE**  **(Project Title)** |
| --- |

| **Developed by**  **Members:**   | **No.** | **Student Name** | **Student ID** | | --- | --- | --- | | 1 |  |  | | 2 |  |  | | 3 |  |  | | 4 |  |  |  * **Class No.:** * **Start Date:** * **End Date:** * **Name of the Coordinator:** * **Date of Submission:** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1. GET THE DATA** |

* **Who**: Who collected the data? What perspectives or agenda might that person or group bring to the data collection? For example, data collected from partisan political groups may reflect their biases.
* **What**: What data is included? Is the data complete? Does it address the data questions?
* **When**: When was the data collected? Is it current? Is it still relevant? For example, a five-year-old dataset may be incapable of addressing a customer's current problem.
* **Where**: If the data is from an external source, where did it come from? Is the source reputable? Trustworthy? For example, user engagement data would likely be more reliable than user reported data on a survey.
* **Why**: Why was the data collected? Can data collected for one purpose be appropriately used for another purpose? For example, data collected to track student learning may not be useful for evaluating teacher and school quality.

**Example**

[**https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis**](https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis)

* **People**
* ID: Customer's unique identifier
* Year\_Birth: Customer's birth year
* Education: Customer's education level
* Marital\_Status: Customer's marital status
* Income: Customer's yearly household income
* Kidhome: Number of children in customer's household
* Teenhome: Number of teenagers in customer's household
* Dt\_Customer: Date of customer's enrollment with the company
* Recency: Number of days since customer's last purchase
* Complain: 1 if the customer complained in the last 2 years, 0 otherwise

| **2. PREPARE THE DATA** |
| --- |

* **Extracting the data from its source**: This may involve writing SQL queries to extract and transform the data from a database or downloading it from an external source.
* **Transforming/organizing the data**: This involves summarizing the data, joining data from different sources, and putting it into a format that is accessible in the tool you'll use to explore it.
* **Cleaning the data**: This involves things like removing duplicate data, correcting inaccurate or missing data, reformatting columns, and dealing with outliers.
* Finding and removing duplicate data
* Finding and removing extra spaces
* Separating data without delimiters
* Correcting inaccurate or irrelevant data
* Updating missing or incomplete data
* Standardizing the case for text
* Standardizing formats for each data type (dates, numbers, etc.)
* Working with outliers
* Visualizing to verify clean data

| **3. EXPLORE THE DATA (EXPLORATORY DATA ANALYSIS, OR EDA)** |
| --- |

* Group data
* Apply descriptive statistics to that data
* Explore and analyze the data
* Ask questions and iterate over our exploratory analysis
* Different kinds of distributions:
  + Normal distributions
  + Uniform distributions
  + Skewed distributions
* Gathering insights from those distributions without calculating descriptive statistics.
* Visualizing the spread, including the interquartile range (IQR) and outliers, for a given distribution using boxplots.
* Visualizing boxplots by categories to visually understand the relationship between numerical columns of data and those categories.

| **4. ANALYZE THE DATA** |
| --- |

**Descriptive analysis** uses past data to answer the question *What happened*? Previously collected data undergoes analysis to answer key questions about past situations or events. For example, a company would use descriptive analysis to determine which products generated the highest volume of sales in the previous year, or which employees demonstrated the greatest productivity.

**Diagnostic analysis** seeks reasons or explanations for what has happened, asking and answering the question *Why did this occur*? Because it is often difficult or impossible to prove cause and effect, diagnostic analysis tends to explore correlations and other relationships in data. For example, medical researchers have long sought to determine what causes multiple sclerosis. They have found a strong association between the incidence of the Epstein-Barr virus and multiple sclerosis, leading them to hypothesize that exposure to the Epstein Barr virus may explain subsequent multiple sclerosis diagnosis.

**Predictive analysis** uses past data to predict future occurrences or unknown values. It addresses the question: What *is likely to happen in the future*? It uses statistical methods and machine learning to identify trends, patterns, and relationships in the data. Predictive analysis is commonly used to forecast customer behavior, credit card default and fraud, market trends, labor demand, weather events, and other important business and social outcomes of interest.

**Prescriptive analysis** is the most sophisticated of the four levels, using historical data to address the question *What should we do*? It analyzes data and generates advice on possible outcomes for different decisions, helping guide decision-makers toward action. Prescriptive analysis helps to automate future decisions, such as evaluating the credit-worthiness of a borrower.

| **5. COMMUNICATE THE RESULTS** |
| --- |

1. **Design for an audience**: tailor results and visualizations to the appropriate audience. For example, a data science manager may want to hear more about the details of the analysis, while an executive may just want the conclusions. This includes using visuals and metrics that the audience will easily understand.
2. **Less is more**: the takeaway from each visual should be clear. Adding too much data or unnecessary design features can distract our audience from the main point. This is called maximizing the data-to-ink ratio. There should be very little space devoted to things that aren't communicating results.
3. **Clear labels and annotations**: The chart titles, axis titles, legends, etc. should work together so a viewer can read the chart or report and know what the data represents without guessing.
4. **Tell a story with data**: *data storytelling*, which has now become its own sub-field, involves communicating data insights via narratives and visualizations. Using elements from storytelling, such as a hook, a narrative, and a moral of the story, can make our analysis more memorable and engaging.