SNS_DATA_ANLAYST_PROJECT_SARAVANAN.J

1.Understanding Data Visualization

Question: Explain the importance of data visualization in data analysis. What are the key principles of effective data visualization?

Data visualization is essential in data analysis because it simplifies complex data, reveals patterns and trends, and aids decision-making. Effective visualization should be clear, accurate, and consistent to ensure the data is easily understood and accurately represented.

Key Principles:

Uniformity - Use Uniform designs to maintain coherence across the visualization

Accuracy - Represent the data truthfully, avoiding any misleading elements.

Clarity - visualization needs to be straightforward and easy to understand.

- **Simplicity:** Keep visualizations clean and uncluttered. Avoid excessive detail or unnecessary elements that can distract from the main message.
- **Clarity:** Use clear labels, titles, and legends to explain what the visualization represents. Avoid jargon or technical terms that might confuse the audience.
- **Relevance:** Ensure that the visualization is relevant to the question being asked and that it effectively conveys the desired message.
- **Accuracy:** Verify that the data is accurate and that the visualization correctly represents the data.
- **Context:** Provide context for the visualization, including information about the data source, time period, and any relevant factors.
- **Accessibility:** Consider the needs of different audiences, including those with disabilities, and ensure that the visualization is accessible to all.

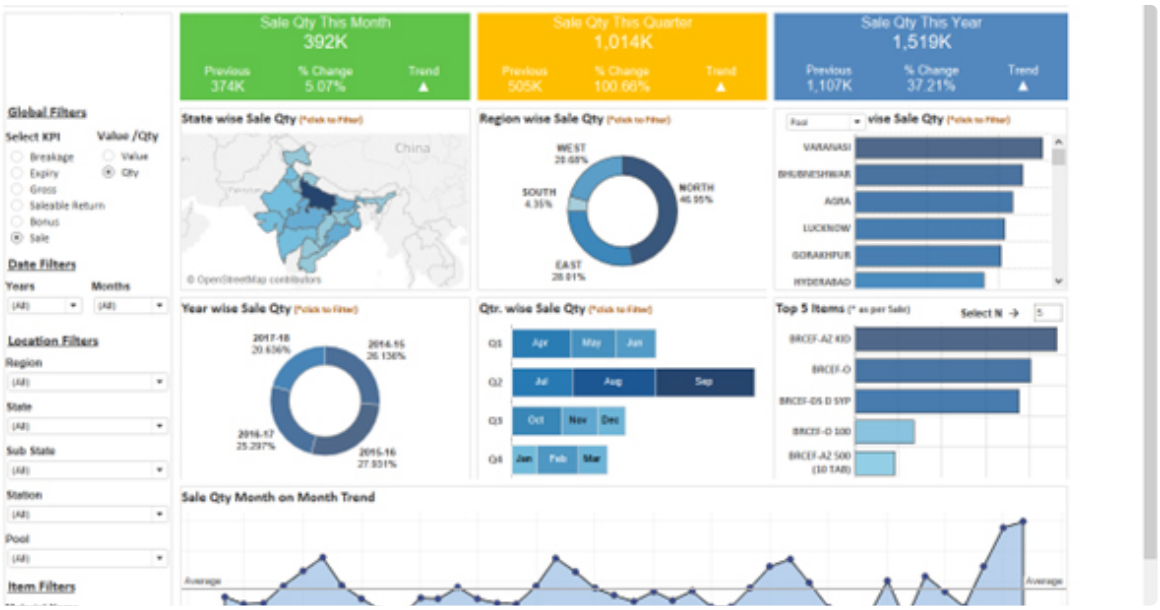
- **Interactivity:** If appropriate, use interactive elements to allow users to explore the data further and uncover additional insights.

2.Tableau Basics

Question: What are the main components of Tableau? Describe the process of creating a basic dashboard in Tableau.

Main Components of Tableau:

1. **Data Sources:** Connects to various data types like Excel, SQL, or cloud services.
2. **Sheets:** Individual workspaces where you create different visualizations.
3. **Dashboards:** A collection of sheets that provide a consolidated view of data.
4. **Stories:** A sequence of visualizations that convey a narrative.
5. **Filters:** Allow dynamic interaction with the data by focusing on specific subsets.



Creating a Basic Dashboard in Tableau:

1. **Connect to Data:** Import your data source (e.g., Excel, database).
2. **Create Sheets:** Build individual visualizations by dragging and dropping fields onto rows, columns, and marks.
3. **Design Dashboard:** Drag your created sheets onto a blank dashboard layout.
4. **Customize Layout:** Arrange the visualizations, add filters, and adjust sizes to create a cohesive view.

5. **Publish/Share:** Save and share the dashboard for others to interact with.

3.Power BI Fundamentals

Question: Discuss the main features of Power BI. How does Power BI differ from Tableau in terms of functionality and use cases?

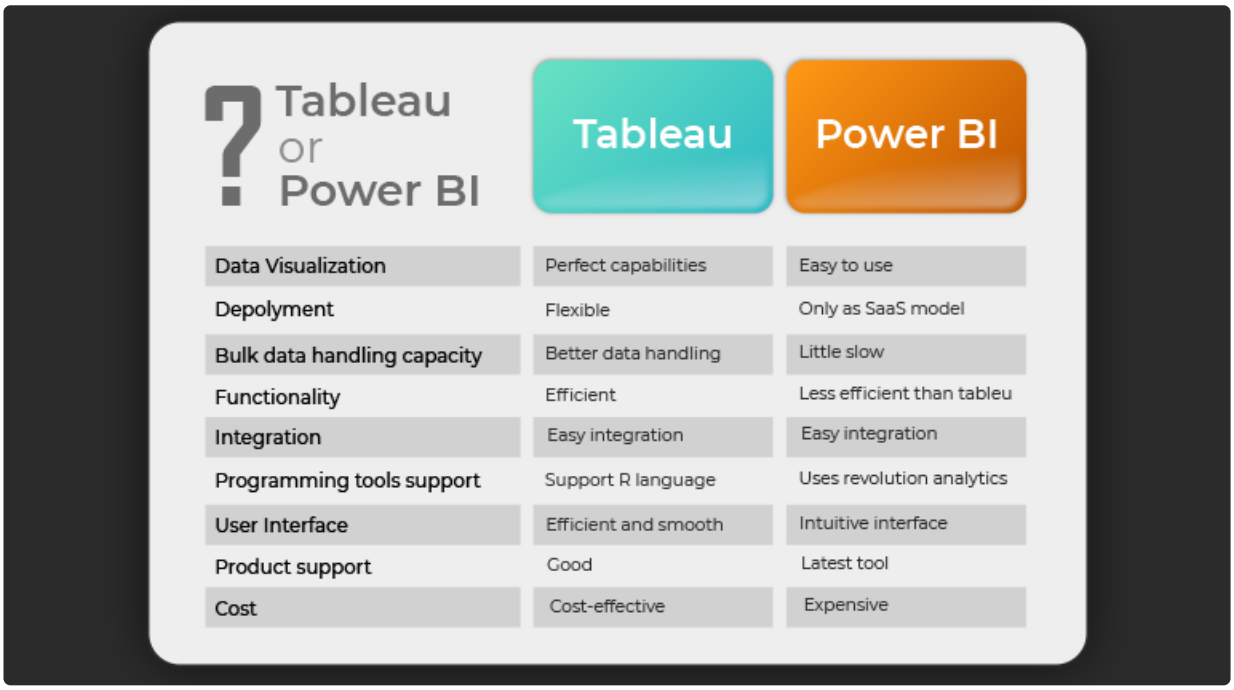


	Tableau	Power BI
Data Visualization	Perfect capabilities	Easy to use
Depolyment	Flexible	Only as SaaS model
Bulk data handling capacity	Better data handling	Little slow
Functionality	Efficient	Less efficient than tableu
Integration	Easy integration	Easy integration
Programming tools support	Support R language	Uses revolution analytics
User Interface	Efficient and smooth	Intuitive interface
Product support	Good	Latest tool
Cost	Cost-effective	Expensive

Main Features of Power BI:

1. **Data Connectivity:** Power BI offers a wide range of data connectors to various sources such as databases, online services, Excel, and more. It supports both direct query and import methods.
2. **Data Transformation:** With Power Query, users can clean, transform, and model data before creating visualizations. This includes merging, filtering, and shaping data.
3. **Data Modeling:** Users can create complex data models using relationships, calculated columns, measures, and hierarchies. DAX (Data Analysis Expressions) is used for advanced calculations.
4. **Interactive Dashboards and Reports:** Power BI enables the creation of interactive reports and dashboards with various visualization options like charts, maps, and KPIs.
5. **Publishing and Sharing:** Reports and dashboards can be published to the Power BI service, where users can share them with others, collaborate, and set up automatic data refreshes.
6. **Integration with Microsoft Ecosystem:** Seamless integration with other Microsoft tools like Excel, Azure, and Office 365 enhances its functionality and user experience.
7. **Natural Language Query:** Power BI's Q&A feature allows users to ask questions in natural language and get answers in the form of visualizations.
8. **Mobile Access:** Power BI provides mobile apps for accessing reports and dashboards on-the-go.

4.Data Cleaning and Preparation

Problem Statement: Given a dataset with missing values and inconsistencies, clean and prepare the data for analysis.

Data Set : Sales_data_sample (https://www.kaggle.com/datasets/kyanyoga/sample-sales-data?resource=download)

Steps for preprocessing:

- **Import the required libraries.**
- **Import the CSV file using `pd.read_csv` and store the dataset in a DataFrame named `df`.**
- **Analyze the dataset by checking the first few rows and examining the shape (number of rows and columns).**
- **Check the dataset's structure and data types using `df.info()`.**
- **Check for null values in each column using `df.isnull().sum()`.**
- **Drop rows with null values using `df.dropna()`.**
- **Convert the data type of the `PRICE_EACH` column from object to integer using `astype(int)`**

5.Tableau Visualization

Problem Statement: Create an interactive sales dashboard in Tableau using the provided sales dataset. The dashboard should include key metrics such as total sales, sales by region, and sales trends over time.

Link - https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls

In this project, you will be working with a dataset from the Superstore, aiming to answer 30 scenario-based questions through data visualisation and analysis. Your objective is to select the best chart for each question, explain your choice. This project will showcase your proficiency in data visualisation, critical thinking, and effective communication.

6.Power BI Report

Problem Statement: Develop a report in Power BI to analyze customer feedback data

link - https://www.kaggle.com/datasets/sudarshan24byte/online-food-dataset

Age: Age of the customer.

Gender: Gender of the customer.

Marital Status: Marital status of the customer.

Occupation: Occupation of the customer.

Monthly Income: Monthly income of the customer.

Educational Qualifications: Educational qualifications of the customer.

Family Size: Number of individuals in the customer's family.

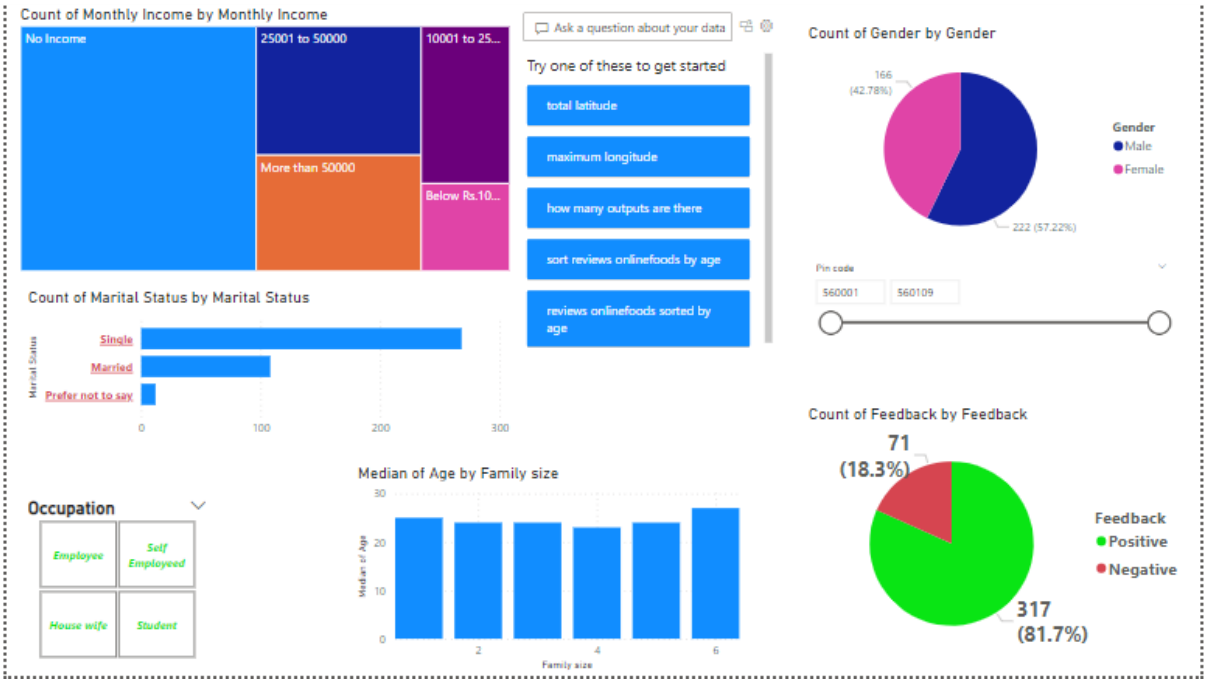
Latitude: Latitude of the customer's location.

Longitude: Longitude of the customer's location.

Pin Code: Pin code of the customer's location.

Output: Current status of the order (e.g., pending, confirmed, delivered).

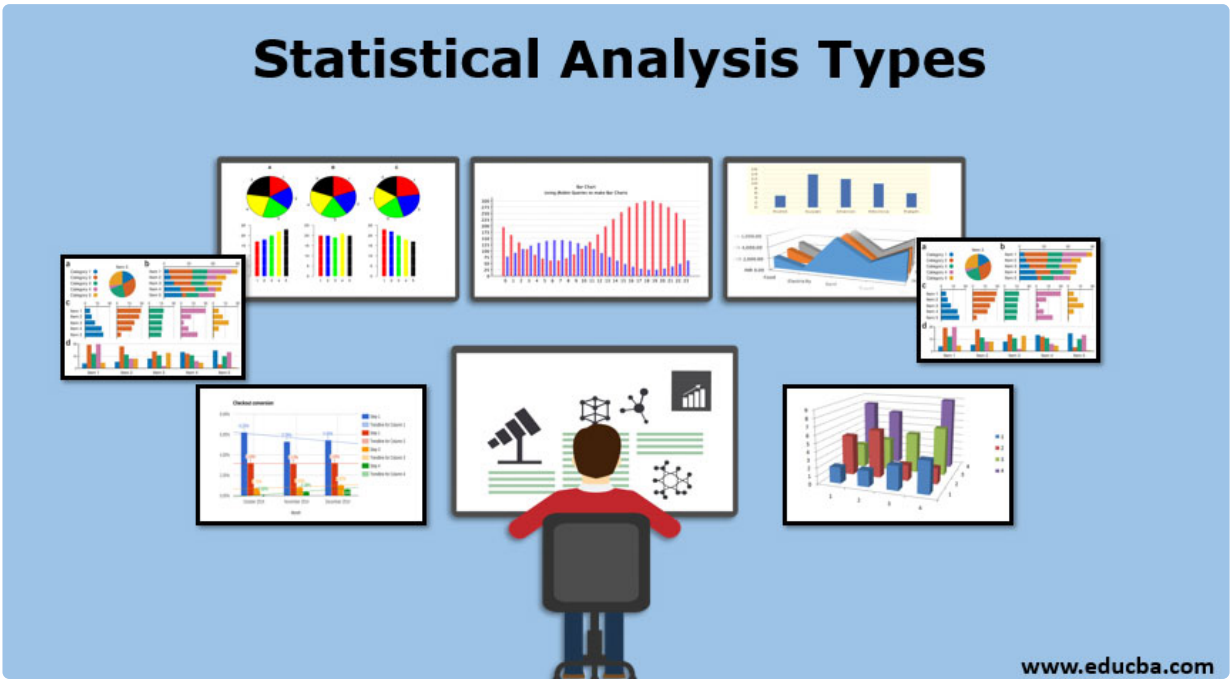
Feedback: Feedback provided by the customer after receiving the order.



Part 3: Advanced Analytics

7.Statistical Analysis

Problem Statement: Perform a statistical analysis on a given dataset to identify significant trends and correlations. Provide a summary of your findings.



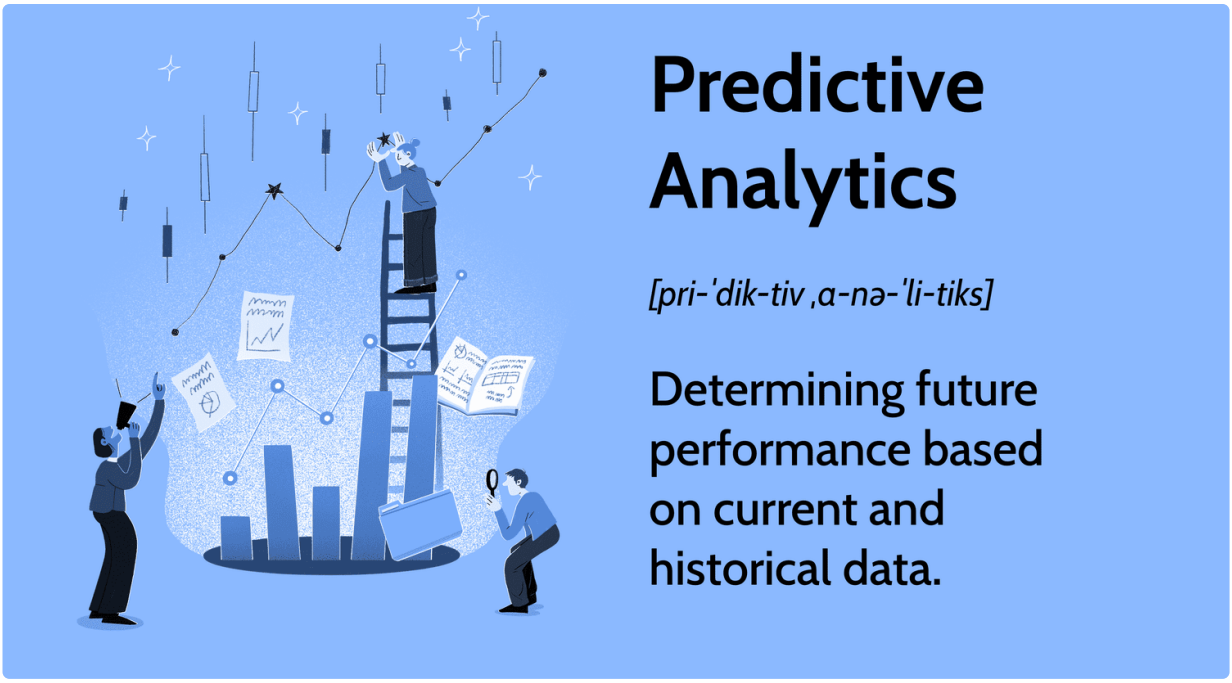
DataSet - Titanic

Key Steps Explained:

1. **Loading the Dataset**:
   - Uses `seaborn` to load the Titanic dataset.
2. **Exploratory Data Analysis**:
   - Displays the first few rows, basic info, and summary statistics.
   - Handles missing values and drops irrelevant columns.
3. **Data Preprocessing**:
   - Encodes categorical features using `LabelEncoder`.
   - Filters out extreme fare values.
4. **Correlation Analysis**:
   - Calculates and plots the correlation matrix.
5. **Hypothesis Testing**:
   - Performs Chi-Square tests to check relationships between gender/class and survival.
6. **Model Training and Evaluation**:
   - Splits the data into training and testing sets.
   - Scales features and trains an `AdaBoostClassifier`.
   - Evaluates model performance with accuracy, precision, recall, and F1 score.

8. Predictive Analytics

Problem Statement: Build a predictive model to forecast sales for the next quarter using historical sales data. Explain the steps taken and the rationale behind your model choice.



DATASET - Copper_price_Prediction

Link - https://github.com/SaravananJayavelu/Industrial_Copper_Modeling.git

- **Data Collection**: Gather historical copper price data and relevant features that might influence prices, such as economic indicators, supply data, and market trends.
- **Data Preprocessing**: Clean the data by handling missing values, removing outliers, and formatting data correctly. Transform features if necessary and split the data into training and testing sets.
- **Feature Selection**: Identify and select relevant features that are most predictive of copper prices. This may involve exploratory data analysis and feature importance techniques.
- **Model Selection**: Choose appropriate predictive models, such as linear regression, decision trees, or advanced algorithms like Random Forest or XGBoost.
- **Model Training**: Train the selected model(s) on the training data to learn the patterns and relationships in the data.
- **Model Evaluation**: Evaluate the model's performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared on the testing data.
- **Model Tuning**: Optimize model parameters and improve performance through techniques like cross-validation, hyperparameter tuning, and feature engineering.
- **Prediction and Validation**: Use the trained model to make predictions on new data and validate the predictions against actual values to assess accuracy.

- **Deployment**: Implement the model into a production environment where it can be used to make real-time predictions and inform decision-making.
- **Monitoring and Maintenance**: Continuously monitor model performance and update it as necessary to ensure it remains accurate over time.

Key Steps Explained:

1. **Loading the Dataset**:
   - Uses `seaborn` to load the Titanic dataset.
2. **Exploratory Data Analysis**:
   - Displays the first few rows, basic info, and summary statistics.
   - Handles missing values and drops irrelevant columns.
3. **Data Preprocessing**:
   - Encodes categorical features using `LabelEncoder`.
   - Filters out extreme fare values.
4. **Correlation Analysis**:
   - Calculates and plots the correlation matrix.
5. **Hypothesis Testing**:
   - Performs Chi-Square tests to check relationships between gender/class and survival.
6. **Model Training and Evaluation**:
   - Splits the data into training and testing sets.
   - Scales features and trains an `AdaBoostClassifier`.
   - Evaluates model performance with accuracy, precision, recall, and F1 score.

9.Real-World Problem Solving

Question: Imagine you are given a large dataset with customer transactions. How would you approach the task of identifying key customer segments and their behaviors? Describe the steps and tools you would use

Step 1: Data Preparation

1. **Load the Dataset**:
   - Import the dataset containing customer transactions into your data analysis tool (e.g., Python with Pandas, R with data.table).
2. **Explore the Data**:
   - Understand the structure of the dataset by examining the columns, types of data, and any apparent anomalies.
   - Summarize the basic statistics of the dataset to understand distribution and identify potential issues.

3. **Data Cleaning**:
    - Handle missing values appropriately (e.g., imputation, removal).
    - Remove duplicate records to ensure accuracy.
    - Convert data types if necessary (e.g., dates).

Step 2: Feature Engineering

1. **Generate Relevant Features**:
    - **Recency**: Calculate the number of days since the last transaction for each customer.
    - **Frequency**: Count the number of transactions each customer has made within a specified period.
    - **Monetary**: Calculate the total amount spent by each customer.
2. **Aggregate Data**:
    - Summarize customer data into a format suitable for analysis. This usually involves creating a dataset where each row represents a customer and columns represent features like recency, frequency, and monetary value.

Step 3: Segmentation Analysis

1. **Normalize the Data**:
    - Standardize or normalize the features to ensure they are on a similar scale. This step is crucial for most clustering algorithms to perform effectively.
2. **Choose a Segmentation Technique**:
    - **K-Means Clustering**: An iterative algorithm that partitions the data into K distinct clusters based on feature similarity. Determine the optimal number of clusters using methods like the Elbow Method or Silhouette Score.
    - **Hierarchical Clustering**: Builds a hierarchy of clusters using a tree-like structure (dendrogram) to determine natural groupings in the data.
3. **Apply Clustering Algorithm**:
    - Implement the chosen clustering technique to segment the customers into distinct groups.

Step 4: Interpret and Analyze Segments

1. **Analyze Segment Characteristics**:
    - Calculate and review the mean or median values of features within each segment to understand the typical customer profile for each group.
    - Identify key differences between segments, such as high-value versus low-value customers.
2. **Visualize the Segments**:
    - Use visualizations (e.g., scatter plots, pair plots) to illustrate how different customer segments are distributed across features.
    - Create charts or plots to represent the characteristics and behaviors of each segment effectively.

Step 5: Actionable Insights

1. **Develop Customer Profiles**:
    - Create detailed profiles for each segment to summarize their characteristics, behaviors, and preferences.
2. **Design Targeted Strategies**:
    - Formulate marketing, sales, and service strategies tailored to each customer segment. For example, offer personalized promotions to high-value customers or improve engagement with frequent but low-spending customers.
3. **Monitor and Refine**:
    - Continuously monitor the effectiveness of your strategies and refine the segmentation as needed based on new data or changing business objectives.

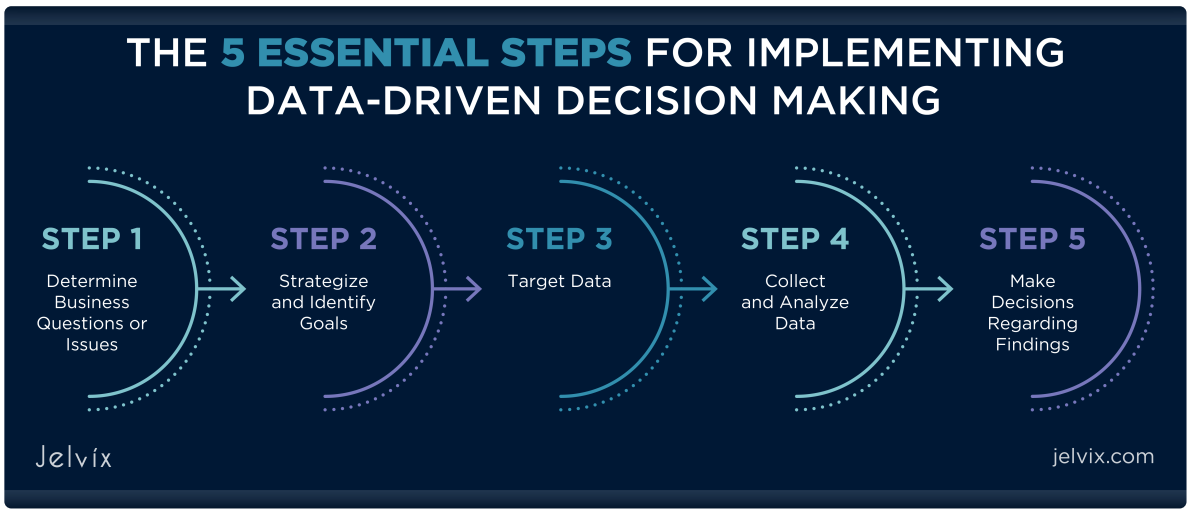
Tools and Techniques

- **Data Cleaning and Preparation**: Pandas (Python), data.table (R)
- **Feature Engineering**: Pandas (Python), dplyr (R)
- **Clustering Algorithms**: Scikit-learn (Python) for K-Means, SciPy (Python) for hierarchical clustering
- **Visualization**: Matplotlib, Seaborn (Python), ggplot2 (R)

This procedure provides a comprehensive framework for identifying and analyzing key customer segments, enabling you to leverage customer data effectively for business decision-making.

10.Data-Driven Decision Making

Question: A company wants to launch a new product and has collected survey data on customer preferences. How would you use this data to help the company make an informed decision? Outline your approach.



1. Data Collection and Preparation

1.1. Review the Survey Data

- **Understand the Dataset**: Examine the survey data to understand the types of questions asked, the response format, and the scope of the data collected.
- **Check Data Quality**: Identify any missing values, inconsistencies, or outliers that need to be addressed.

1.2. Clean the Data

- **Handle Missing Values**: Impute or remove missing values depending on the extent and importance of the missing data.
- **Standardize Responses**: Ensure that responses are standardized (e.g., converting text responses to categorical variables).

2. Data Exploration and Analysis

2.1. Descriptive Statistics

- **Summarize the Data**: Compute basic statistics (mean, median, mode, standard deviation) for numerical responses and frequency counts for categorical responses.
- **Visualize Preferences**: Use charts (e.g., bar charts, pie charts) to visualize customer preferences and trends.

2.2. Segment the Data

- **Customer Segmentation**: Identify different customer segments based on responses (e.g., demographic information, buying behavior).
- **Analyze Segments**: Compare preferences and needs across different segments to understand varying demands.

3. Hypothesis Testing

3.1. Formulate Hypotheses

- **Test Assumptions**: Develop hypotheses about customer preferences and the potential success of the new product. For example, "Customers who prefer eco-friendly products will be more likely to purchase this new product."

3.2. Perform Statistical Tests

- **Conduct Tests**: Use statistical tests (e.g., t-tests, chi-square tests) to evaluate whether differences in preferences are statistically significant.
- **Evaluate Results**: Interpret the test results to confirm or refute your hypotheses.

4. Predictive Analysis

4.1. Build Predictive Models

- **Choose Models**: Depending on the data and goals, build predictive models to estimate potential product adoption rates or customer satisfaction.
- **Train and Validate**: Train the model on historical survey data and validate its performance using techniques like cross-validation.

4.2. Analyze Predictive Results

- **Estimate Potential Success**: Use the model to forecast how different customer segments are likely to respond to the new product.
- **Interpret Predictions**: Evaluate the model's predictions and their implications for the product launch.

5. Strategic Recommendations

5.1. Identify Key Insights

- **Customer Preferences**: Summarize key insights from the survey data, such as the most desired features or price points.
- **Segment-Specific Recommendations**: Provide tailored recommendations for each customer segment based on their preferences.

5.2. Develop a Launch Strategy

- **Marketing Strategy**: Develop targeted marketing campaigns based on customer segments and preferences.
- **Product Positioning**: Position the product in a way that aligns with the identified needs and preferences of the target audience.
- **Pricing Strategy**: Set a price point that is attractive to the target segments while ensuring profitability.

6. Monitor and Adjust

6.1. Implement the Launch

- **Execute the Plan**: Roll out the product according to the developed strategy.
- **Monitor Performance**: Track key performance indicators (KPIs) such as sales, customer feedback, and market penetration.

6.2. Iterate Based on Feedback

- **Collect Feedback**: Gather feedback from customers post-launch to assess satisfaction and identify any issues.
- **Adjust Strategy**: Make necessary adjustments to the product or marketing strategy based on real-world performance and feedback.

Tools and Techniques

- **Data Cleaning and Preparation**: Pandas, R for data manipulation.
- **Descriptive and Exploratory Analysis**: Pandas, Matplotlib, Seaborn (Python) or ggplot2 (R) for visualization.

- **Hypothesis Testing**: SciPy, Statsmodels (Python) or base R functions.
- **Predictive Modeling**: Scikit-learn, XGBoost (Python) or caret, randomForest (R).
- **Visualization and Reporting**: Matplotlib, Seaborn (Python) or ggplot2, Shiny (R).

This approach provides a comprehensive framework for leveraging survey data to make informed decisions about launching a new product.