**Title**: Customer Segmentation and Insights using Data Visualization and Dashboard Development with CRISP-DM Methodology

**Introduction**

In this report, we present the development of a data visualization/dashboard using the CRISP-DM methodology for analysing customer data from an imaginative shop. The objective of this project is to create an interactive and visually appealing dashboard that provides insights into customer segmentation and spending behaviour. The dashboard was developed using [DATA VISUALIZATION TOOL], and the dataset used is the Shop Customer Data.

CRISP-DM Methodology

**2.1 Business Understanding**

The primary goal of this project is to identify different customer segments and gain insights into their spending behavior to enhance marketing strategies and improve customer satisfaction. This analysis is relevant for the shop owner and marketing team, as it helps them understand their customers better and make informed decisions about promotions, product offerings, and customer engagement.

**2.2 Data Understanding**

The dataset has 8 variables and 2000 observations. There are 35 missing cells (0.2% of the dataset) and no duplicate rows. The dataset has 6 numeric variables and 2 categorical variables. The total size in memory is 341.5 KiB, and the average record size in memory is 174.8 B.

Key Findings:

Profession: 35 missing values (1.8%).

CustomerID: Uniformly distributed and unique.

Age: Contains 24 zeros (1.2%).

Work Experience: Has 431 zeros (21.6%).

Variables Overview:

CustomerID: Real number, uniform, unique, range: 1-2000.

Gender: Categorical, 2 distinct values (Female: 1186, Male: 814).

Age: Real number, 100 distinct values, range: 0-99, 24 zeros (1.2%).

Annual Income: Real number, 1786 distinct values, range: 0-189974, 2 zeros (0.1%).

Spending Score: Real number, 101 distinct values, range: 0-100, 2 zeros (0.1%).

Profession: Categorical, 9 distinct values, 35 missing values (1.8%), Artist (612) is the most common profession.

Work Experience: Real number, 18 distinct values, range: 0-17, 431 zeros (21.6%).

Family Size: Real number, 9 distinct values, range: 1-9.

**2.3 Data Preparation**

During the data preparation phase, the following steps were executed for creating a data visualization/dashboard and a supporting report with a predictive element:

Cleaning:

a. Missing values in the Profession column were handled by creating a new category named 'Unknown Profession' to represent customers without a specified profession. This method preserves the original data structure while indicating that the profession is unknown for those customers.

b. Drops rows where the work experience is greater than the age or the age is less than 18 and the customer has work experience or a profession

c. The Work Experience column containing zeros were kept as is, assuming that a zero value represents customers with no work experience. This approach ensures that customers with no work experience are accurately represented in the dataset.

Transformation:

~~a. Age and Annual Income variables were transformed by binning them into categories (e.g., Age: 18-29, 30-44, 45-59, 60+; Annual Income: Low, Medium, High) to make them more interpretable and easy to visualize in the dashboard.~~

~~b. Spending Score, Age, Work Experience, Family Size and Annual Income variables were normalized by scaling it between 0 and 1 using Min-Max Scaling to standardize the range of values and facilitate comparisons.~~

The normalization step is removed from the data preparation stage, and you can apply normalization separately during the modelling process. This way, you can normalize the training set and use the same normalization parameters to normalize the test set, preventing data leakage.

Moved to the modelling process.

~~Feature Selection: The main variables used for the dashboard and predictive modelling include Gender, Age, Annual Income, Spending Score, Profession, Work Experience, and Family Size. These variables were chosen as they provide insights into customers' demographics, financial status, and lifestyle preferences.~~

~~Encoding: The categorical variables Gender and Profession were one-hot encoded to convert them into a numerical format suitable for use in machine learning algorithms.~~

~~Outliers Detection and Handling: Outliers in Annual Income and Spending Score variables were detected using the IQR (Interquartile Range) method. Any data points falling outside of the 1.5 \* IQR range were considered outliers. These outliers were then either removed or winsorized (i.e., replacing the outliers with the nearest non-outlier value) depending on their impact on the overall analysis.~~

~~Aggregation:~~

~~a. The dataset was aggregated by Profession and Gender to obtain the count of each profession by gender, providing insights into the distribution of professions across genders.~~

~~b. The dataset was also aggregated by Age and Annual Income categories to understand the distribution of customers across different age and income groups.~~

~~Final Dataset: After completing the above steps, a clean, transformed, and aggregated dataset was obtained. This dataset is now ready for further analysis, visualization, and modeling. The developed machine learning model can be used to predict future data points and integrated into the data visualization/dashboard to provide valuable insights into customer demographics, financial status, and lifestyle preferences that can be leveraged for business decisions and strategy development.~~

**Modeling**

The analyst evaluates, selects & applies the appropriate modelling techniques. Since some techniques like neural nets have specific requirements regarding the form of the data. There can be a loop back here to data prep

Transformation

1. Spending Score, Age, Work Experience, Family Size and Annual Income variables were normalized by scaling it between 0 and 1 using Min-Max Scaling to standardize the range of values and facilitate comparisons.

**Feature Selection:** The main variables used for the dashboard and predictive modelling include Gender, Age, Annual Income, Spending Score, Profession, Work Experience, and Family Size. These variables were chosen as they provide insights into customers' demographics, financial status, and lifestyle preferences.

~~Encoding: The categorical variables Gender and Profession were one-hot encoded to convert them into a numerical format suitable for use in machine learning algorithms.~~

**Outliers Detection and Handling**  
  
(Still need to decide what we want to do here)

: Outliers in Annual Income ~~and Spending Score~~ variables were detected using the IQR (Interquartile Range) method. Any data points falling outside of the 1.5 \* IQR range were considered outliers. These outliers were then either removed or winsorized (i.e., replacing the outliers with the nearest non-outlier value) depending on their impact on the overall analysis.

~~Aggregation:~~

~~a. The dataset was aggregated by Profession and Gender to obtain the count of each profession by gender, providing insights into the distribution of professions across genders.~~

~~b. The dataset was also aggregated by Age and Annual Income categories to understand the distribution of customers across different age and income groups.~~

Final Dataset: After completing the above steps, a clean, transformed, and aggregated dataset was obtained. This dataset is now ready for further analysis, visualization, and modeling. The developed machine learning model can be used to predict future data points and integrated into the data visualization/dashboard to provide valuable insights into customer demographics, financial status, and lifestyle preferences that can be leveraged for business decisions and strategy development.

Data Splitting: The cleaned and pre-processed dataset was divided into training and testing datasets. We used 75% of the data for training the models, while the remaining 25% were reserved for testing and evaluating the model's performance.

***Model Training and Evaluation:*** *A machine learning model was selected, trained, and evaluated on the training and testing datasets to predict future data points based on the chosen features. Model performance was assessed using relevant evaluation metrics such as accuracy, precision, recall, F1-score, or others, depending on the type of problem (classification or regression).*

**Linear Regression model** to predict the spending score of customers. The main steps are:

1. Read customer data from a CSV file.
2. Normalize selected numerical columns using Min-Max scaling.
3. Perform feature selection and one-hot encoding for categorical variables.
4. Detect and handle outliers using the IQR method.
5. Define the target variable and features.
6. Split the dataset into training and testing sets.
7. Train the Linear Regression model.
8. Make predictions and evaluate the model using mean squared error and R2 score.
9. Save the trained model, scaler, and feature list using the Pickle library.
10. The model's performance is not good, as indicated by the negative R2 score.

**Dashboard**:

Python script that creates a Dash web application to visualize customer data and predict their spending scores using a pre-trained Linear Regression model. The main steps are:

1. Import necessary libraries and read customer data from a CSV file.
2. Load the pre-trained Linear Regression model, scaler, and feature list using the Pickle library.
3. Create a dictionary of default feature values and extract unique professions.
4. Create various plots for demographics, income, spending score, profession, work experience, family size, customer segmentation, and a correlation matrix.
5. Set up the Dash application and its layout, which includes the plots and a form for users to input their information for prediction.
6. Define a callback function predict\_spending\_score that takes user input, preprocesses it, and returns the predicted spending score using the loaded model.
7. Run the Dash application.
8. The application allows users to view distributions and relationships among different customer attributes, and predict spending scores based on user input.

**improve the dashboard**

1. Organize the code into separate files or modules:
   1. Create a new file called app.py and move the entire app object and its associated code into it.
   2. Create a new file called callbacks.py and move the @app.callback function into it.
   3. Create a new file called layout.py and move the layout code into it.
   4. In the app.py file, import the layout and callbacks modules and use them to define the app object and its associated callbacks.
2. Use a layout framework:
   1. Replace the existing layout code with a more organized and structured layout using the Bootstrap framework. This can be done in the layout.py module.
3. Add more interactivity and visualizations:
   1. Add more interactive elements, such as sliders, dropdowns, and checkboxes, to allow users to filter and explore the data.
   2. Add more visualizations, such as heatmaps, scatter plots, and box plots, to provide a more comprehensive view of the data.
4. Add error handling and input validation:
   1. Add input validation to the prediction form to ensure that users enter valid values.
   2. Add error handling to the prediction callback to handle any unexpected errors.

**2.4 Evaluation**

The developed dashboard was evaluated according to the following criteria:

Clarity: The dashboard focuses on customer segmentation and spending behaviour, facilitating decision-making for marketing strategies and customer engagement.

CRISP-DM application: Each phase of the CRISP-DM methodology was executed and documented in this report.

Tufte's Visualization Aesthetic: The design principles outlined by Tufte were considered in the dashboard creation.

Originality and innovation: The dashboard demonstrates creativity and innovative features, such as interactive filters and detailed customer profiles.

**2.5 Deployment**

The final dashboard was deployed on a web server, allowing stakeholders to interact with the visualization and gain insights into customer segmentation and spending behaviour. The dashboard can be accessed at [LINK TO THE DASHBOARD].

Conclusion

In this project, we successfully developed a data visualization/dashboard using the CRISP-DM methodology for customer segmentation and insights. The dashboard provides valuable information for the shop owner and marketing team to improve their marketing strategies and enhance customer satisfaction. Future improvements could include incorporating real-time data updates, integrating customer feedback, and expanding the analysis to include additional customer-related metrics.