**Chargeable Income Report**

**Introduction:**

This report provides a comprehensive analysis of the financial performance of both taxable and non-taxable groups from the years 2003 to 2018. The data includes total revenue, total expenses, and net income for each group over the specified period. The objective of this report is to gain insights into the financial trends, identify patterns, and provide recommendations for optimizing financial performance.

This report aims to explore and analyse the dataset on chargeable income tax. Let's delve into the details:

1. **Data Overview:**
   * The dataset consists of information about chargeable income tax for different years of assessment.
   * It includes features such as year of assessment, tax group, number of companies assessed, total income, donations, chargeable income, group relief, loss carryback relief, chargeable income after reliefs, tax exemption, chargeable income after exemption, gross tax payable, tax deducted at source, other tax set-offs, and net tax assessed.

**2. Data Acquisition and Preprocessing**

The income tax data was loaded from a CSV file using pandas. Libraries like NumPy and warnings were imported for numerical operations and handling warnings, respectively.

Data cleaning steps included:

* Handling missing values by replacing 'na' with NaN and converting the column to float64.
* Removing negative signs from the 'chargeable\_income' column using absolute value function.

**3. Exploratory Data Analysis (EDA)**

* Dataframe information and dtypes were explored using **df.info()** and **df.dtypes.**
* Value counts were obtained for the 'chargeable\_income' column to understand income distribution.
* Head function displayed the first few rows of the data to get a glimpse of the data.
* Column names were retrieved using **df.columns.**

**4. Feature Engineering**

New features were created based on existing ones:

* **ci\_aft\_reliefs:** Chargeable income after applying group relief and loss carryback relief.
* **tax\_exemption:** Tax exemption amount calculated as a percentage of **ci\_aft\_reliefs.**
* **gross\_tax\_payable:** Gross tax payable calculated as a percentage of **ci\_aft\_exemption.**
* **tax\_deducted\_at\_source** and **other\_tax\_set\_offs:** Calculated as a percentage of **gross\_tax\_payable.**
* **net\_tax\_assessed:** Net tax assessed calculated by subtracting **tax\_deducted\_at\_source** and **other\_tax\_set\_offs** from **gross\_tax\_payable.**

**5. Data Splitting for Taxable and Non-Taxable Groups**

Data was split into two DataFrames:

* **nt:** Non-taxable group companies.
* **t:** Taxable group companies.

Further splits were made based on the year of assessment (before and after 2007) to compare pre and post trends.

**6. Percentage Calculations**

A function **percentage\_calculator** was created to calculate percentages used for feature engineering.

**7. Filling Missing Percentages for Pre-2007 Data**

Percentages for tax reliefs and tax calculations for the pre-2007 data were estimated using the average percentages obtained from the post-2007 data for the respective tax groups.

**8. Combining and Sorting Data**

The pre-processed data for all groups and years was combined using **pd.concat** and sorted by year of assessment and tax group.

**9. Data Description**

Descriptive statistics of the final data were obtained using **df.describe()**.

**10. Data Visualizations**

* Distribution of chargeable income was visualized using seaborn's **histplot**.
* Line plots with error bars (set to None) were created to show trends in chargeable income across years for taxable and non-taxable groups using **seaborn.lineplot.**
* Bar plots were used to compare donations made by the two groups over the years using **seaborn.barplot**.
* Separate histograms were created for 'group\_relief' and 'loss\_carryback\_relief' using subplots in a figure.
* Scatter plots were used to explore the relationship between **ci\_aft\_reliefs** and **tax\_exemption.**
* Distribution of **gross\_tax\_payable, tax\_deducted\_at\_source**, and **other\_tax\_set\_offs** were visualized using stripplots.
* Line plots were created to compare trends in net tax assessed for taxable and non-taxable groups.

**11. Machine Learning Model**

* The data was prepared for machine learning by creating dummy variables for the categorical feature **tax\_group** using **pd.get\_dummies.**
* Correlation between features was explored using **dt.corr()** and visualized as a heatmap using seaborn.
* Features with high correlation or deemed less relevant (e.g., number of companies assessed, donations) were dropped.
* The remaining features were separated into independent (X) and dependent variables (y).
* The data was split into training and testing sets using **train\_test\_split.**
* A Decision Tree Classifier model was chosen for prediction. The model was trained

**Conclusion:**

* + The decision tree classifier demonstrated excellent performance on the provided dataset.
  + However, achieving a perfect accuracy score may indicate potential overfitting.
  + Further analysis and validation techniques are recommended to ensure the model's robustness and generalization to unseen data.

This report provides an initial exploration of the dataset and highlights the process of building and evaluating a classification model for tax group prediction. Further analysis and refinement can enhance the model's effectiveness and reliability.