**DSA210 REPORT– MURAT GÜRBÜZ 29429**

**Data Analysis Report**

1. **Introduction**

This report provides a comprehensive explanation of the data analysis process and visualization techniques applied using Python libraries such as Pandas, Matplotlib, Seaborn, and Statsmodels. The objective is to analyze a dataset encompassing various metrics related to fitness and lifestyle, including calories intake, exercise duration, step count, sleep duration, and weight over time.

The analysis aims to identify patterns, correlations, and trends within the data that can offer meaningful insights into personal health and fitness behaviors. By leveraging statistical summaries, graphical representations, and time series forecasting, the report seeks to provide a well-rounded understanding of how different lifestyle factors interact and influence each other.

Specifically, the data exploration process begins with loading and cleaning the dataset, followed by generating descriptive statistics for numerical columns to understand the basic properties of the data. Visualization techniques are extensively used to illustrate distributions, relationships, and time-dependent changes in key variables. Additionally, ARIMA models are employed for time series forecasting, enabling the prediction of future values for selected metrics such as steps and weight.

The insights derived from this analysis can be useful for individuals aiming to monitor and improve their health by understanding the impact of various factors on their fitness outcomes. The findings may also serve as a foundation for more advanced analyses or applications, such as developing predictive models or personalized health dashboards.

**2. Data Loading and Preliminary Analysis**

The data is loaded from an Excel file using Pandas. Initially, only numeric columns are selected for statistical analysis, ensuring that operations are performed on relevant data types. A basic statistical summary of the numeric columns is generated, providing measures such as mean, standard deviation, minimum, and maximum values.

**3. Visualization Techniques**

A variety of plots and graphs are employed to visually summarize the data and uncover patterns.

**3.1 Histograms**

Histograms are created for each numeric column to display the frequency distribution of the values. The histograms help in understanding the spread and skewness of the data.

**3.2 Correlation Matrix and Heatmap**

A correlation matrix is computed to understand the linear relationships between different numeric variables. The correlation values are visualized using a heatmap, where color intensity indicates the strength of the correlation. This helps in identifying highly correlated pairs of variables.

**3.3 Scatter Plots**

Scatter plots are used to explore pairwise relationships between specific variables. For example, one plot examines the relationship between calorie intake and exercise duration, while another explores the connection between sleep duration and water intake. These plots help in visually identifying potential correlations or patterns.

**3.4 Boxplots**

Boxplots are used to display the distribution of numeric variables across different categories. One example involves plotting exercise duration by different activity levels, helping to compare how activity levels influence exercise duration.

**3.5 Time Series Visualization**

To analyze trends over time, the date column is converted to a datetime format and set as the index. A line plot is created to visualize the weight changes over time in relation to exercise duration. This provides insights into how weight fluctuates over the recorded period.

**4. Time Series Forecasting**

ARIMA models are employed to forecast future values of certain variables. The steps involved in time series forecasting include defining the ARIMA model with specific parameters, fitting the model to the data, and predicting future values.

**4.1 Steps Forecast**

An ARIMA model is built to forecast the number of steps for the next 10 days. After fitting the model, a forecast is generated, and the predicted values are visualized alongside actual values for comparison.

**4.2 Weight Forecast**

A similar ARIMA model is used to forecast the weight trend for the next 30 days. The predicted weight values are plotted against the actual recorded values, providing a visual representation of the expected trend.

**5. Conclusion**  
This report provides a detailed and structured approach to data analysis by integrating several key steps, including data exploration, visualization, and time series forecasting. Through the use of Python libraries and statistical techniques, a comprehensive understanding of the dataset was achieved, enabling the identification of significant patterns and trends related to fitness and lifestyle behaviors.

The exploratory data analysis highlighted crucial relationships among variables such as calorie intake, exercise duration, step count, sleep duration, and weight over time. Visual representations, including histograms, scatter plots, boxplots, and correlation heatmaps, were instrumental in uncovering insights into the distribution and interaction of these variables. For example, the correlation analysis revealed potential dependencies that could inform better personal fitness management strategies.

Moreover, time series forecasting using ARIMA models offered predictive insights into future trends, allowing for proactive decision-making regarding lifestyle adjustments. The accuracy of these forecasts can serve as a foundation for further development of predictive tools aimed at personalized health monitoring.

Overall, the findings of this analysis underscore the value of data-driven approaches in understanding complex behaviors and improving health outcomes. By applying systematic data analysis techniques, this report not only provides actionable insights but also sets the stage for more advanced predictive modeling efforts.

Future efforts could benefit from integrating additional data sources, employing more sophisticated machine learning algorithms, and developing interactive dashboards for real-time monitoring. These enhancements would further enhance the utility and applicability of the analysis in both personal and professional health management contexts.