**Module 1: Assignment**

**Mathematical Foundations for AI and Data Science**

**Business Scenario**

A consumer goods company “ABC Ltd” promotes its products by adapting marketing strategies related to active media advertisements. They are actively advertising their products through 3 different platforms – TV, Radio and Newspaper. The management has collected monthly data on Sales(in Lacs) and monthly advertising expenses over a period for analysis. The columns in the dataset are defined as below:

Date: Depicts beginning of every month in the format (mm/dd/yyyy). The figures for the other columns are last month’s summary.

TV: Monthly advertising cost spent (lacs) for advertising on TV

Newspaper: Monthly advertising cost spent (lacs) for advertising on newspaper

Radio: Monthly advertising cost spent (lacs) for advertising on radio

Sales: Number of units sold

Assume yourself to be a part of the analyst team of the company. You can complete the following tasks using **Python language**.

Use the dataset: “**Advertisement\_Sales.csv”**

**Read the file and perform the following operations:**

1. Identify the missing values/null values in each field and replace the same by the mean of that year for the respective columns.

Example: If missing values are found for TV advertisement in the year 2018, then the missing field(s) should be replaced by the mean cost of TV advertisement for that year. The same applies for other years and all other fields.

1. Find the absolute methods of dispersion (range, standard deviation, variance, interquartile range, quartile deviation and mean deviation) in each numeric field (TV, Newspaper, Radio, Sales) and compute boxplot on each field.
2. Find the yearly frequency distributions of Sales and plot graphs for visualization.
3. Compute Hypothesis Testing to validate the hypothesis whether investment in advertisements through different channels(TV, Newspaper, Radio) are affecting the Sales figure. Compute both 1-Tailed and 2-Tailed tests. Identify Type-1 and Type-2 errors(if any), justify the answer. Compute t-test, z-test and chi-square test as appropriate.
4. Identify any linear or polynomial dependency of Sales on advertisement investments through different channels independently or in any combination of the same. Plot scatter charts to demonstrate the dependencies (if any). Visualize the relationship between the following using scatter plot: i) Amount spent on advertising on TV and units sold ii) Amount spent on advertising on Newspapers and units sold iii) Amount spent on advertising on Radio and units sold

1. Apply Linear Regression algorithm to the dataset. Also find y-intercept, R^2 score and Mean Squared Error.
2. Create a time plot using the date field and the sales to visualize the sales figures for each month.
3. Apply Augmented Dickey-Fuller test on the sales data to check whether the data is stationary. Perform “differencing” as necessary to make the Sales data stationary. Confirm by reapplying the Augmented Dickey-Fuller test on the differenced data.
4. Use ARIMA time-series model to forecast the Sales appropriately. Apply ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) to determine the values of *p*, *d*, and *q*. The number of lags where ACF cuts off is *q*, and where PACF cuts off is *p*. Choose the appropriate value for d based on the differences used.
5. By analyzing the cost of advertisements for different channels (TV, Radio and Newspaper) and the Units sold over the period, apply Bayesian statistics to predict whether increase in advertisement budget would increase the sales of units. Also find the most appropriate channel for which the budget needs to be raised for an increase in sales.

**NB:** The solution of the assignment should be submitted as document doc,docx/pdf or may create zip/7z of the solutions along with screenshots and upload. The entire code should be available as plain text against each question. Paste the screenshots of output. Explain the process of working(logic) appropriately in plain text as required.