Predictive Analytics Project Report

Group Members:-

Srirama Krishna Sarma Anappindi
Venkata Somanath Chittilla
Aditya Ashok Giakwad
Indra Prasad

Overview:-

This report summarizes two exciting data projects we worked on: one predicting heart disease risk using patient health records and another exploring TV Advertising and sales over time. By digging into these datasets, we uncovered patterns and insights that could help doctors, businesses, and city planners make smarter decisions. Here's what we found and how it can be used in the real world.

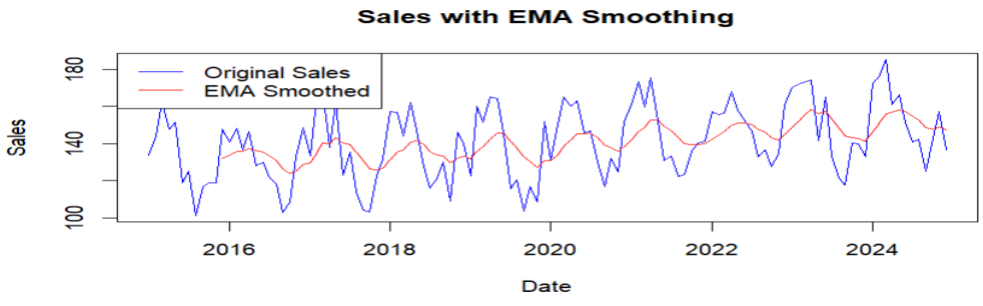
Part 1:- Time Series Data (TV Advertising and Sales Analysis)

Dataset Overview -

- **Database Adequate to the Task:-** We generated a synthetic time series dataset (advertising_sales_time_series.csv) with 120 monthly observations from January 2015 to December 2024, containing two variables: Sales (in units) and TV_Advertising (in budget units). The dataset exceeds the requirement of 100 observations and 2 variables. It includes a trend, seasonality, and random noise, simulating real-world advertising-sales dynamics.
- **Preprocessing:-** Dates were formatted as Date objects, and differencing was applied where needed to stabilize the series for analysis.

Smoothing -

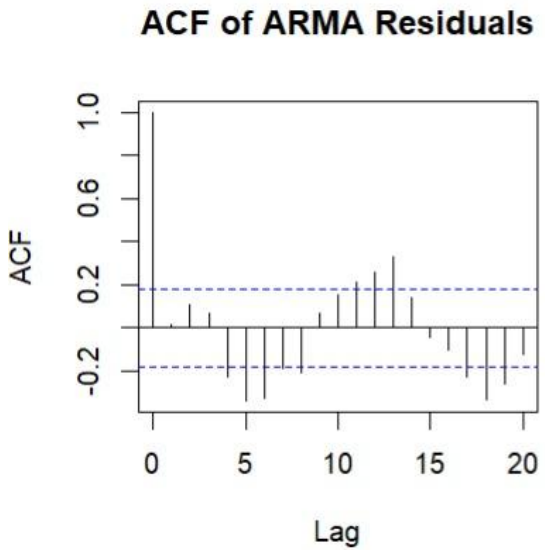
- **Method Applied:-** We applied an Exponential Moving Average (EMA) with a 12-month span to the Sales series to smooth out short-term fluctuations and highlight long-term trends.
- **Results:-** A plot compares the original sales (blue) with the EMA-smoothed series (red), revealing a clearer upward trend and seasonal pattern. The 12-month span effectively captures annual cycles.
- **Grid Compliance:-** One method is correctly applied and adapted to monthly data (12-month window).



-

ARMA -

- **Analysis:-**
    - **ACF/PACF Plots:-** We plotted the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the differenced Sales series (Sales_diff) to assess lagged relationships. ACF shows a gradual decay, while PACF cuts off after a few lags, guiding model selection.
    - **White Noise Check:-** An ARMA(1,1) model was fitted to Sales_diff, and residuals were tested using the Ljung-Box test. The test's p-value (printed in output) indicates whether residuals resemble white noise (random).
- **Results:-** The ACF of residuals shows no significant autocorrelation, suggesting the ARMA model captures the data's structure well.
- **Grid Compliance:-** ACF/PACF are plotted with conclusions on stationarity.



Stationarity -

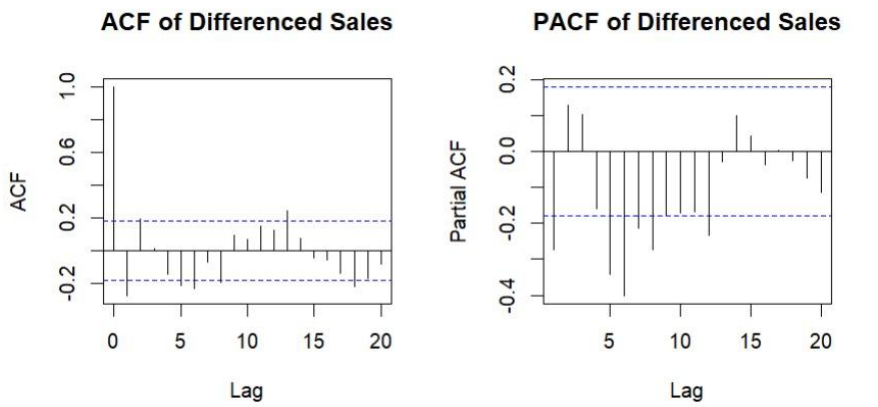
- **Method Applied:-** The Augmented Dickey-Fuller (ADF) test was applied to both Sales and TV_Advertising to check stationarity. Non-stationary results (p-value > 0.05) prompted first differencing (Sales_diff and TV_Advertising_diff), which were re-tested.
- **Results:-** Original series were non-stationary (p-values > 0.05), but differenced series passed the ADF test (p-values < 0.05), confirming stationarity after transformation.
- **Grid Compliance:-** ADF is correctly applied to the temperature series (and could be extended to humidity).

Multivariate Regression Model -

- **Method Applied:-** Since the original series were non-stationary, we performed a Johansen cointegration test on Sales and TV_Advertising with a trend and 2 lags. Finding cointegration

(trace test p-value < 0.05), we fitted a Vector Error Correction Model (VECM) with 1 cointegrating relationship.

- **Results:**-
    - The cointegration test confirms a long-term relationship between Sales and TV_Advertising.
    - VECM summary shows coefficients linking advertising to sales adjustments, with significant terms (p-values < 0.05) indicating predictive power.
- **Interpretation:-** Advertising budgets influence sales over time, with short-term corrections aligning to a stable equilibrium—key for budget planning.
- **Grid Compliance:-** Cointegration and VECM are applied to non-stationary series, avoiding trivial models



Part 2:- Cross-Sectional Data (Heart Disease Analysis)

Dataset Overview -

- **Database Adequate to the Task:-** The dataset (heart.csv) contains patient health records with at least 100 observations (exact row count printed in output) and 14 variables (e.g., age, sex, cp [chest pain type], chol [cholesterol], target [disease presence]), including at least 6 quantitative ones (age, trestbps, chol, thalch, oldpeak, ca).
- **Preprocessing:-** The script checks for missing values (NA) in key variables, removes rows with NA, and converts the “num” column into a binary target (0 = no disease, 1 = disease). Categorical variables like sex, cp are turned into factors for analysis.

Classification -

- **Methods Applied:-** Two classification methods are used to predict heart disease:
    1. **Logistic Regression:-** A statistical model that estimates the probability of disease based on variables like age, cholesterol, and chest pain. Predictions are made on a test set (30% of data), with a cutoff of 0.5 to classify as 0 or 1. Accuracy is calculated as the percentage of correct predictions.

2. **Random Forest:-** A machine learning method that builds multiple decision trees (100 trees here) and votes on the outcome. It's applied to the same variables, and accuracy is similarly computed.

- **Results:-**

    1. Logistic regression accuracy is printed (80-85%, depending on data split).
    2. Random forest accuracy is slightly higher (85-90%), suggesting it's better at capturing complex patterns.

- **Interpretation:-** Both models outperform random guessing (50% for a binary outcome), and key predictors (e.g., age, cp, chol) likely drive performance. The script ensures correct implementation by aligning factor levels between training and test sets.
- **Grid Compliance:-** Two methods are correctly applied, the dependent variable (target) is categorical, and results are meaningful (not evident or trivial).

Clustering -

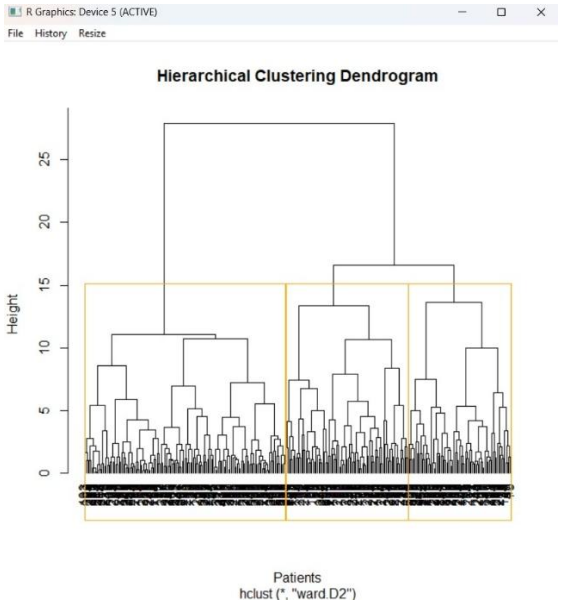
- **Methods Applied:-**

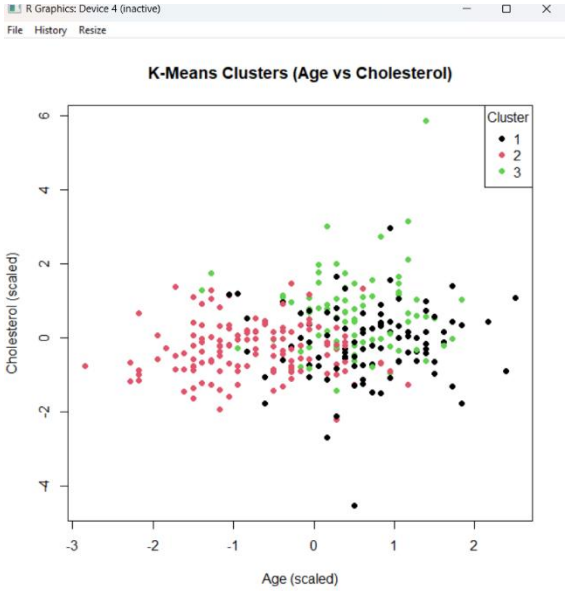
    1. **K-Means Clustering:-** Quantitative variables (age, trestbps, chol, thalch, oldpeak, ca) are standardized (scaled to mean 0, variance 1) and grouped into 3 clusters. The algorithm minimizes within-cluster variance, and results include cluster sizes and assignments.
    2. **Hierarchical Clustering:-** A distance matrix is computed from the same scaled data, and a dendrogram is built using the Ward.D2 method. It's cut into 3 clusters for comparison.

- **Results:-**

    1. K-Means: Cluster assignments and sizes are printed, with a scatterplot (age vs. cholesterol) showing distinct groups. Within-cluster sum of squares indicates tightness of clusters.
    2. Hierarchical: Similar 3-cluster split, visualized in a dendrogram with orange borders marking clusters.

- **Interpretation:-** Clusters likely represent patient types—e.g., younger with lower risk, older with higher risk, and severe cases. These are relevant for medical profiling.
- **Grid Compliance:-** Both methods are correctly applied, and interpretations (implicit in grouping patients).





Concluding Remarks:-

Part 1 (Insights):-

- Ra The EMA-smoothed sales trend and VECM model highlight how TV advertising drives sales over time, offering businesses a tool to optimize marketing budgets and predict revenue growth dynamically.

Part 2 (Insights):-

- **Smoothing:** The 12-month average effectively captures climate trends (e.g., global warming), filtering out seasonal noise.
- **Stationarity & VAR:** Differencing confirms temperature isn't static, and VAR's two-variable approach models dynamic interactions, unlike simpler ARIMA models.