**A Report on Simple Linear Regression and Multiple Linear Regression**

# Introduction

The report below introduces two models: simple linear regression and multiple linear regression. The datasets used for this analysis include “Advertising.csv” and “Real\_estate.csv”. The regression models aim to explore the relationships between advertising spending and sales as well as various factors influencing house prices in the real estate market.

“Advertising.csv” contains data on advertising expenditures across different media channels namely TV, Radio and Newspapers and their sales data. “Real\_estate.csv” contains key attributes of real estate transactions including their “transaction date”, “house age” “distance to the nearest MRT station”, “number of convenient stores”, “latitude”, “longitude” and “house price of unit area”.

The objective of this analysis is to explore the accuracy of linear regression in predicting sales based on advertising spending and to make predictions on the determinants of house prices in a given area.

# Data Description

The datasets used for the regression analyses provide useful insights into different domains.

**Advertising.csv**

This dataset contains data related to advertising expenditures across different media channels and corresponding sales figures.

The variables include:

* “TV”: Advertising spending on TV commercials.
* “Radio”: Advertising spending on radio channels.
* Newspaper: Advertising spending on newspaper advertisements.
* Sales: Total sales generated as a result of advertising efforts.

**Real \_estate.csv**

This dataset provides comprehensive data on the real estate market, focusing on factors influencing house prices in a given area.

The variables include:

* “ID”: The unique identifier for each dataset.
* “Transaction date”: Date of the real estate transaction.
* “House age”: Age of the house in years.
* “Distance to the Nearest MRT station”: Proximity of the house to the nearest Mass Rapid Transit (MRT) station.
* “Number of convenience stores”: Number of convenience stores in the vicinity of the house
* “Latitude”: Geographic latitude of the house location
* “Longitude”: Geographic longitude of the house location.
* “House Price of Unit Area”: Price of the house per unit area.

# Simple Linear Regression Analysis

The results of the simple linear regression analysis conducted using the “Advertising.csv” dataset will be presented in this section. The analysis aims to explore the relationship between TV advertising spending and sales figures.

**Relationship between variables**

The independent variable in the analysis is “TV” representing the amount of money spent on TV advertising. The dependent variable is “Sales” presenting the totalsalesmade as a result of the advertising efforts.

**Q1c). Regression equation**

The regression equation derived from the simple linear regression is as follows:

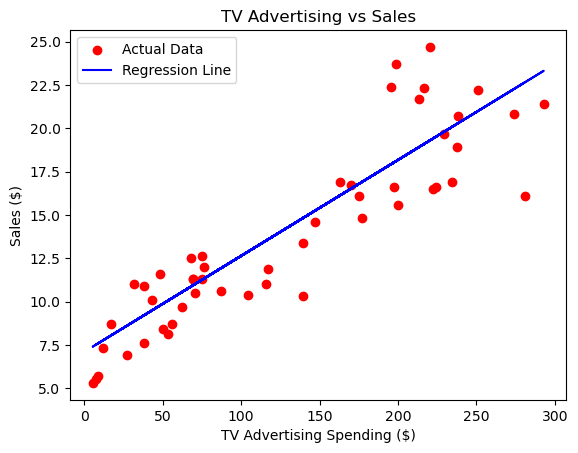
Sales = β0​ + β1​ × TV

Where:

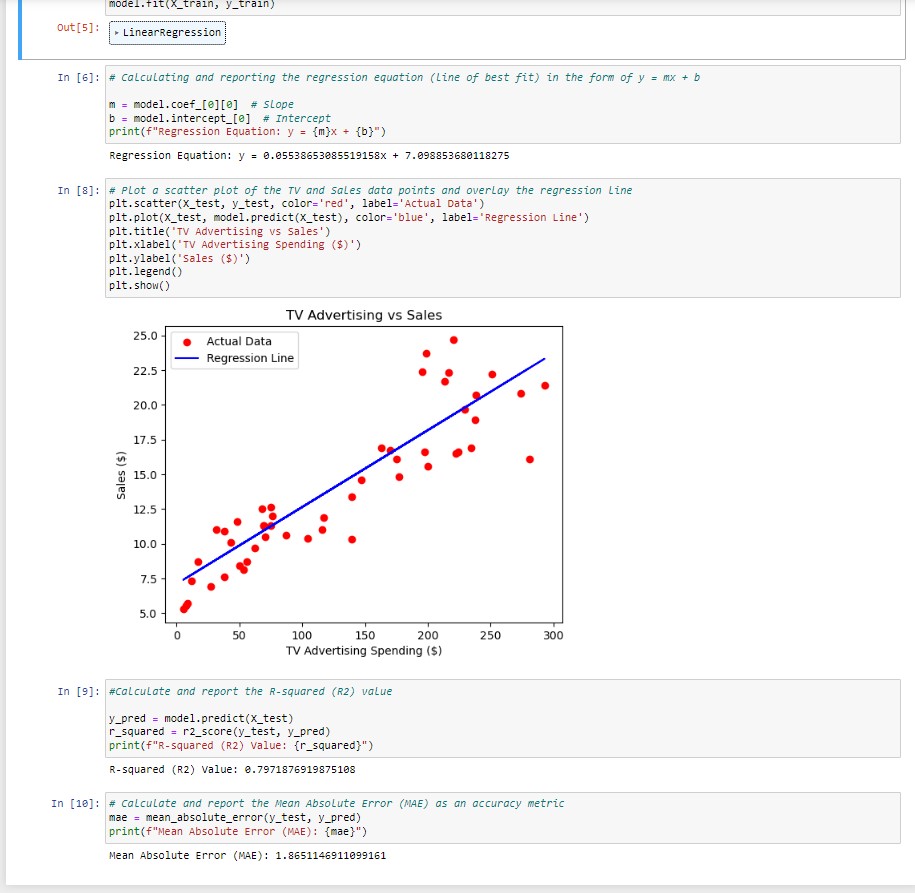
* β0:​ represents the intercept, indicating the figure for the baseline sales when TV advertising spending is zero.
* β1:​ represents the coefficient of the TV variable, and it signifies the change in sales for each unit increase in TV advertising spending. A positive coefficient suggests that higher TV advertising expenditure is associated with increased sales, while a negative coefficient implies that lower TV advertising expenditure is associated with decreased sales.

**Q1d). Visualization**

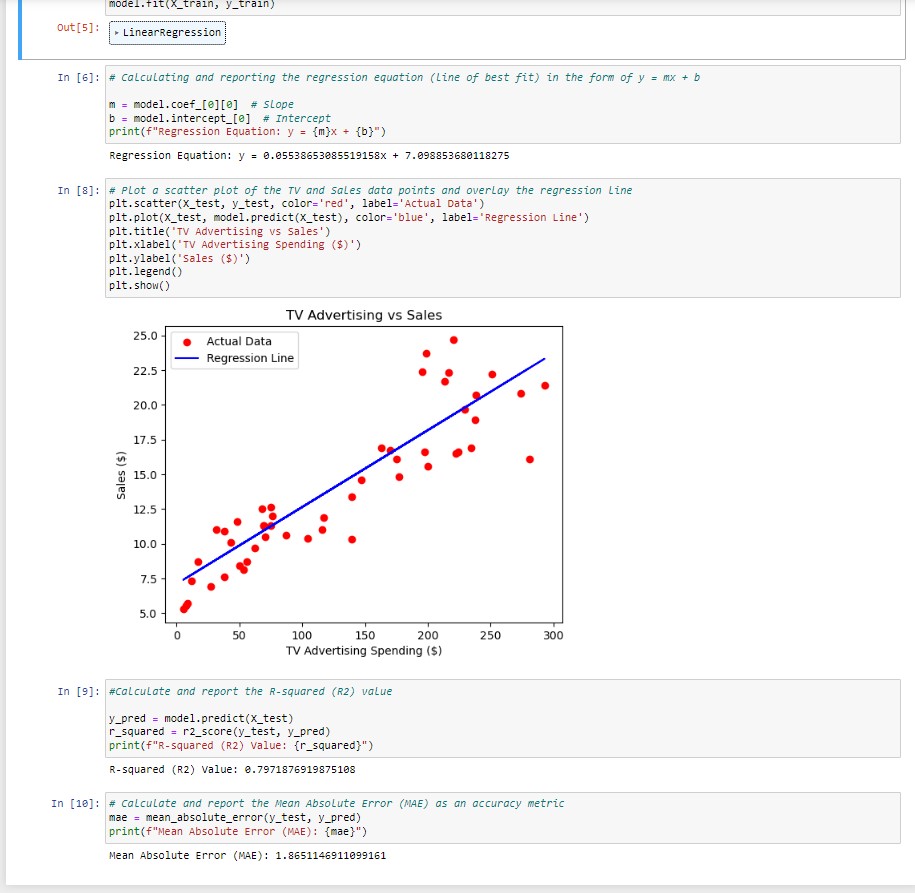
To visually represent the relationship between TV advertising spending and sales, the scatter plot below was created with TV advertising spending on the x-axis and sales figures on the y-axis. The regression line was overlaid on the scatter plot to illustrate the linear relationship between the variables.

****

**Q1e). R-squared Value to assess the goodness of the fit**

* The R-squared value is used to measure the size of the variance in the dependent variable (sales) that is explained by the independent variables in the regression model.
* It ranges from 0 to 1, where values closer to 1 indicate a better fit of the model to the data.
* An R-squared value of 1 implies that the model explains all the variability in the dependent variable, while a value of 0 indicates that the model does not explain any of the variability.
* The R-squared value for the sales variable is 0.797 which indicates that the model explains at least 80% of the variability in the dependent variable.
* 

**Q1f) Mean Absolute Error (MAE)**

* The Mean Absolute Error (MAE) measures the average absolute difference between the predicted values and the actual values of the dependent variable.
* It provides a measure of the average magnitude of errors in the predictions made by the model.
* A lower MAE indicates higher accuracy and performance of the model, as it signifies smaller discrepancies between predicted and actual values.
* The MAE for the dataset is 1.865 which indicates a high accuracy and performance of the model. It implies slight discrepancies between the predicted and actual values.
* 

# Multiple Linear Regression Analysis

This section will present the findings of the multiple linear regression analysis conducted using the "Real\_Estate.csv" dataset. The aim is to examine the relationship between various factors including the house age, distance to the nearest MRT station, and the number of convenience stores, and their collective impact on the house price per unit area. The multiple linear regression model investigates how changes in multiple independent variables collectively influence the house price per unit area. By utilising multiple predictors simultaneously, the model aims to capture the combined effects of different factors on house prices.

**Relationship between variables**

The independent variables used in this analysis include:

* “House Age”
* “Distance to the Nearest MRT Station”
* Number of Convenience Stores

The dependent variable is:

* House Price of Unit Area

**Q2c) Multiple Regression Equation**

The multiple regression equation is denoted as follows:

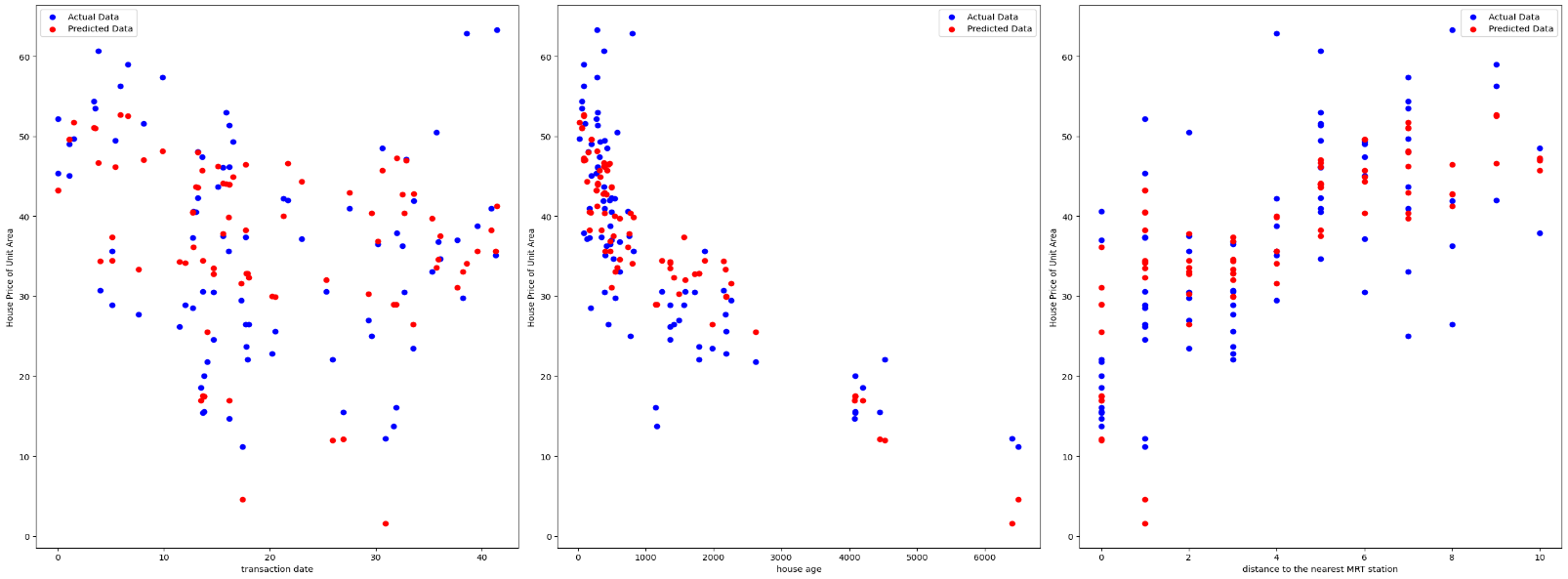
House Price of Unit Area = β0 ​+ β1 ​× House Age + β2​ × Distance to the Nearest MRT Station + β3 ​× Number of Convenience Stores + e

Where:

* β0:​ represents the intercept, indicating the baseline house price per unit area when all independent variables are zero.
* β1​, β2​, and β3:​ denote coefficients associated with each independent variable, representing the change in house price per unit area for a single-unit increase in the predictor variables.
* ϵ: denotes the error term, capturing unexplained variance in house prices not accounted for by the independent variables.

**Q2d) Visualizations**

To visualize the relationships between the independent variables and the dependent variable, the scatter plots below are created for each independent variable (House Age, Distance to the Nearest MRT Station, Number of Convenience Stores) against the house price per unit area.

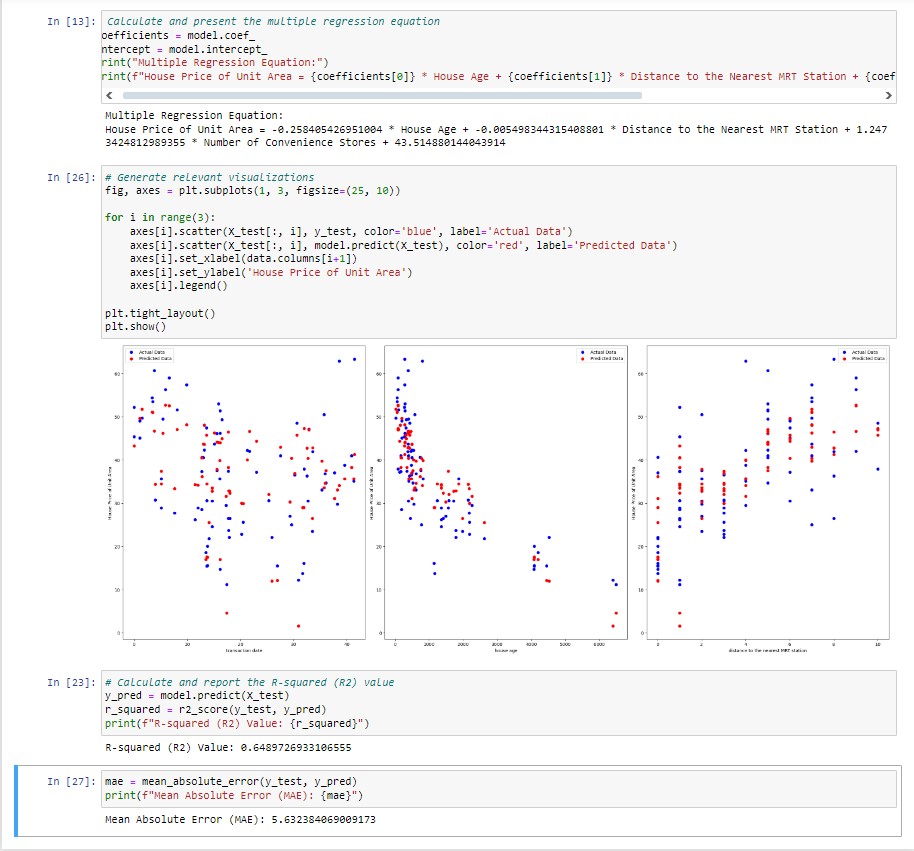
****

# ****Q2e)** R-squared (R2) Value**

* The R-squared value is used to measure the size of the variance in the dependent variable (House Price of Unit Area) that is explained by the independent variables in the regression model.
* It ranges from 0 to 1, where values closer to 1 indicate a better fit of the model to the data.
* An R-squared value of 1 implies that the model explains all the variability in the dependent variable, while a value of 0 indicates that the model does not explain any of the variability.
* The R-squared value of the above dataset was 0.649, indicating that the model explains an above-average variability of the House Price of Unit Area.

# Q2f) Mean Absolute Error

* The Mean Absolute Error (MAE) measures the average absolute difference between the predicted values and the actual values of the dependent variable.
* It provides a measure of the average magnitude of errors in the predictions made by the model.
* A lower MAE indicates higher accuracy and performance of the model, as it signifies smaller discrepancies between predicted and actual values.
* The MAE of the above dataset was 5.63, representing a slightly high margin of error between the actual and predicted values.



# Conclusion

The regression analysis conducted offers valuable insights into factors influencing sales and house prices and sheds light on the behaviors of consumers and real estate market trends.

**Summary of Findings**

* The simple linear regression analysis showed a positive relationship between TV advertising spending and number of sales. As TV advertising expenditure increased, there was a corresponding increase in sales, indicating the effectiveness of TV commercials in driving consumer demand.
* In the multiple linear regression analysis, several factors were found to significantly influence house prices per unit area. House age, distance to the nearest MRT station, and the number of convenience stores in the area emerged as key determinants, with each variable exerting a distinct impact on house prices.

**Insights from the analysis**

* The results indicate the significance of advertising expenditure, particularly on TV, in driving sales and enhancing brand recognition. Businesses can leverage this insight to optimize their marketing strategies and effectively allocate resources across different advertising mediums.
* In the real estate market, factors such as house age, proximity to transportation hubs, and access to amenities like convenience stores play important roles in determining property prices. By understanding these factors, investors and homebuyers are aided in making informed decisions and identifying rewarding investment opportunities.