Executive Summary

The Uncleaned Laptop Price dataset is a collection of laptop product listings from an online e-commerce website. The dataset includes information about various laptop models, such as their brand, screen size, processor, memory, storage capacity, operating system, and price. However, the dataset is unclean. It contains missing values, inconsistent formatting, and other errors that need to be addressed before the data can be used for analysis or modelling.

The dataset contains categorical and numerical variables, with most variables being categorical, including brand, model name, screen resolution, processor type, and operating system.
Some numerical variables include screen size, memory, and storage capacity.
The target variable in the dataset is the price, which is a continuous variable.

To achieve the project's objective, I employed a multiple-step approach:
1. Data Preprocessing: I tried to understand data, handle duplicate data, replace missing values, standardize values, remove outliers, and filter data.
2. Exploratory Data Analysis (EDA): I performed Univariate Analysis, Bivariate Analysis, Multi-variate Analysis and Hypothesis Testing.
Since all p-values are less than 0.05, I can reject the null hypothesis and conclude that there is a significant difference between the price and them.
3. Feature Selection and Feature Engineering: After encoding there were 82 features, then I performed RFE, VIF, removed collinear features, and backward elimination, after these, there were 27 features left.
The scores after feature selection are not much different from the scores of the original 82 features.
4. Model Selection: Before performing Linear Regression, I tested multiple Linear Regression assumptions including Linearity Assumption, No Multicollinearity among Predictors, Normality of the Error Terms, Homoscedasticity and No Autocorrelation of the Error Terms.
Then I performed Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, KNN Regression and Support Vector Regression.
After these, for Linear Regression, Lasso Regression, Ridge Regression, Random Forest and KNN Regression, the overall performance of these 5 models is very good.
Among them, KNN has the lowest MAE, MSE and RMSE.
Next, run Grid Search for hyperparameter optimization.
5. Model Training: For the KNN Regression model
R2 Score: The "Best KNN Regression" model has an R2 value of 0.79, which indicates that approximately 79% of the variance in the target variable is explained by the model. A higher R2 value suggests a better fit to the data, indicating that this model accounts for a larger portion of the variability in the dataset.
MAE (Mean Absolute Error): The MAE for the "Best KNN Regression" model is 10,843.68. MAE measures the average absolute difference between predicted and actual values, and a lower MAE indicates a model that provides more accurate predictions.

MSE (Mean Squared Error): The MSE for this model is 236,678,748.87, which is a measure of the average squared difference between predicted and actual values. Lower MSE signifies a model that has smaller prediction errors.
RMSE (Root Mean Squared Error): With an RMSE of 15,384.37, the "Best KNN Regression" model excels in providing predictions with lower magnitudes of error. RMSE is the square root of MSE and is another metric for assessing prediction accuracy.
Cross-Validation Score: The cross-validation score for this model is 0.74, which suggests that it maintains good performance across different subsets of the data during cross-validation. This indicates a robust and reliable model.
Considering these metrics collectively, KNN Regression demonstrates good performance in explanatory ability, prediction accuracy, and stability, making it the optimal choice among the given models.

6. Model Evaluation: The average R-squared score of the KNN Regression model in cross-validation is 0.74, which indicates that this model can explain the variance of the target variable well, with about 74% of the variance explained by the model.
The standard deviation is 0.036, which means there is relatively little change or volatility in the R-squared score.
These results show that this model performed well in cross-validation, has strong explanatory power for the target variable, and is highly stable, indicating that the model may have good generalization performance.

7. Interpretation and Application:

Feature importance analysis:
Feature importance reflects the contribution of each feature to the model when predicting laptop price.
According to the top 10 features ranked by feature importance, we can see that RAM, laptop type, GPU model, and screen resolution, have a high impact on laptop prices.

Model explanation:
In this scenario, I use the KNN Regression model.
The model can be interpreted as predicting laptop prices based on the presence or absence of specific features (RAM, GPU model, screen resolution, laptop type) and their numerical magnitude.

Model application:
With this model, we can predict the price of a laptop based on its characteristics.
This is very useful for e-commerce websites that can be used to develop product pricing strategies, recommend products to customers that suit their needs, etc.

Business impact analysis:
RAM, GPU model, screen resolution, and laptop type have a greater impact on price, so these factors should be considered when designing products or sales strategies.
For example, when pricing, we need to consider whether the RAM size, GPU model, etc. are high-end, which will affect the product pricing strategy.