Examining two datasets, which I obtained from excelanalytics, consisting of 100 records and 1000 records respectively, to analyze sales data brings forth a good understanding of the success of different predictive models' approaches. The rationale for choosing sales data is its relevance to real-world business scenarios and its complex nature, involving multiple factors that influence financial outcomes. Sales data typically includes a mix of categorical and numerical variables, making it an ideal candidate for testing different modeling approaches. Additionally, predicting sales performance is a common and crucial task in many industries, making this analysis particularly applicable to practical business challenges. While working with these datasets, I encountered some difficulties in handling the categorical variables and addressing potential overfitting in the smaller dataset. The features are consistent across both datasets, which include but are not limited to region, country, item type, channel of sale, sale order, and financial metrics such as revenue, cost, and profit. This consistency allowed for a direct comparison of model performance across different data volumes, though it also presented challenges in ensuring that the models were not overly sensitive to the specific patterns in the smaller dataset.

The analysis made on correlational relationships generally revealed high correlations between several numerical variables. Total Revenue, Total Cost, and Total Profit showed strong correlations, as did Units Sold and Unit Price. These correlational patterns indicated that linear regression could be a suitable starting point for modeling, as it can capture linear relationships between variables.

Because there were categorical variables like Item Type, one hot encoding was adopted creating a binary column for every category. Such transformation was very useful in preparing the data for both linear regression and random forest models, allowing these algorithms to effectively incorporate categorical information.

For this analysis, two algorithms were selected: linear regression and random forest. Linear regression was chosen for its simplicity, interpretability, and ability to capture linear relationships between variables. The random forest algorithm was selected as a more complex alternative, capable of modeling non-linear relationships and interactions between variables.

The reasons for using such algorithms were directly linked to the characteristics of the datasets. The relatively small number of records in both datasets (100 and 1000) meant that even complex algorithms like random forest could be applied without excessive computational burden. The presence of both numerical and categorical variables made random forest a suitable option, as it can handle mixed data types well.

A linear regression model was first built using data from all the sales records, however, Unit Cost and Total Cost were excluded due to high multicollinearity. The model initially showed signs of overfitting, with an R-squared value of 1 for the 100 sales dataset. Log

transformation was subsequently applied to the Total Revenue variable. This strategy helped reduce overfitting and capture non-linear relationships.

Interestingly, the log-transformed linear regression models and the random forest models recorded similar RMSE values for each dataset. This indicates that the log transformation effectively captured the non-linear aspects of the data, allowing the linear model to perform as well as the more complex random forest model.

From a business perspective, the linear regression model with log-transformed Total Revenue would be the preferred choice. It provides comparable performance to the random forest model while offering easier interpretability of the relationships between variables and Total Profit.

The analysis showed that the performance of the models improved with the use of more data (1000 sales vs. 100 sales), as evidenced by the lower RMSE for the larger dataset. This implies that, in this case, more data helps in building a more accurate predictive model. However, it's important to note that extremely large datasets could potentially introduce noise or irrelevant patterns, which, if not properly managed, may lead to overfitting.

Comparing the analysis between datasets, we observe that both models performed better on the larger dataset, suggesting that the additional data points helped in capturing the underlying patterns more effectively. The consistency in performance across both algorithms indicates that the relationship between the predictors and Total Profit is relatively stable and well-captured by the log-linear relationship in the regression model.

Lastly, the analysis reveals the importance of feature engineering, such as log transformation, in improving model performance. It also underscores that relatively simple models, when properly tuned, can sometimes perform as well as more complex algorithms, offering a balance between predictive power and interpretability.