**Introduction**

This report explores the data wrangling process using the real-world dataset "Supermarket Sales Forecast." The goal is to prepare and transform the unprocessed data into a format that will work for modelling and analysis. The first section of the report introduces data wrangling as a critical phase of data analysis. The "Supermarket Sales Forecast" dataset is then described in general, along with its goal and intended predictions.

The report will then move on to describe the various data wrangling methods that will be used. These methods include investigating the dataset to understand its structure and spot any problems with the data's quality, handling missing data with imputation techniques, addressing errors and inconsistencies with data cleaning, and carrying out data transformation procedures like encoding and scaling.

Subsequently, the essay discusses the step of developing a straightforward linear regression model to forecast product sales at supermarket locations after the data wrangling phase is complete. The dataset will be divided into training and testing sets, and measures like mean squared error or R-squared will be used to assess the model's performance.

The report focuses on how crucial documentation is for the data wrangling process because it enables others to verify the findings. The report ends by underlining the importance of data wrangling in preparing data for analysis and modelling as well as its potential to produce insightful recommendations for improving grocery sales performance.

**Explore the Data**

Just like a typical supermarket dataset, it includes variables such as outlet identifier, establishment year, item weight, item fat content, item visibility, item type, item MRP, outlet size, outlet location type, outlet type, and item outlet sales. In this section, we will explore the relationship among these variables to better understand the factors that affect sales.

Now, we will explore the thought process behind the relationship of each variable and item outlet sales. Although it acts as a unique store ID and helps in differentiating various outlets, the outlet identifier is not directly related to other features. To determine whether the establishment year influences sales performance, the store's age is compared with item outlet sales. Factors like packing, transportation costs, and customer preferences, item weight may have an indirect impact on sales. To ascertain whether low-fat items are preferred by consumers and how this affects purchasing decisions, the item's fat content is analysed. It is anticipated that item visibility, which is the percentage of display space given to a product, will have a favourable effect on sales. To understand the significance of product placement, the relationship between item visibility and item outlet sales is investigated. The item type categorization helps identify popular product categories that contribute significantly to sales. To ascertain the impact of various product categories, the relationship between item type and item outlet sales is examined. Product prices also known as item MRP is examined to determine whether certain price points attract buyers or whether more expensive products create more sales. To find out whether larger stores have a competitive advantage in terms of sales performance, outlet size—which reflects the ground surface covered by a store—is analysed. To understand the impact of the outlet site type on customer behaviour and purchasing habits, it is necessary to know the type of city in which a store is located. Just like item type, outlet type is analysed to better understand how store types—grocery stores, supermarkets—affect sales and consumer preferences.

After gaining a clear understanding of the factors that may or may not impact item outlet sales, I proceeded to clean the data (which will be explained in the following section) before exploring its relationship with every variable. This was done to ensure unbiased relationships and conclusions. To facilitate easy reference, I categorized the variables into numerical and categorical variables. Generally, most variables do have a distinct relationship with item outlet sales excluding the following. For the numerical variables, I observed that item weight does not have a direct correlation with item outlet sales, and outlet establishment year is a discrete variable. Therefore, I decided to drop the item weight column. Instead, I created a new column called "No\_of\_Years," which represents the outlet establishment year subtracted from the current year. Regarding the categorical variables, I found that item fat content did not exhibit a significant relationship with item outlet sales. Consequently, I also dropped this variable from further analysis.

**Cleanse the Data**

Not all data given is clean and perfect for analysis. It often contains errors or inconsistencies. Thus, in this section, we will explore how data cleaning is performed, an essential step to ensure accurate and reliable data, enabling a more accurate and meaningful relationship.

Firstly, I loop through the values under each column to identify missing values, duplicate records, and inconsistent formatting. Through this step, I found out that under column item fat content exist inconsistencies regarding the naming of Low Fat and Regular. As such, I proceeded to replace the data in the column from ‘LF’ to ‘Low Fat’, ‘reg’ to ‘Regular’ and ‘low fat’ to ‘Low Fat’. With that, the dataset becomes cleaner for further analysis or modelling, ensuring uniformity and consistency in the item fat content column.

Next, the number of missing values is calculated for each column, and I found out that only the item weight column and the outlet size column contain missing values. As a result, I narrowed down to the null values of each column to look for potential patterns or relationships associated with the missing values. Through this step, I found out that the missing values for the item weight column are all under outlets which were established in the year 1985. Hence, I handled missing values in the item weight column by creating a dictionary that maps item identifiers to their respective weights based on non-null values. It then replaces the missing values in the item weight column using dictionary mapping. Regarding the outlet size column, there isn't an obvious pattern or relationship associated with the missing values. With that being said, I proceeded to do data exploration (explained in the previous section) before performing missing value imputation. I narrowed down to frequent imputation and adding a bespoke category. Ultimately, I used frequent imputation as I felt that adding a bespoke category may increases the cardinality of the variable, potentially affecting the model's performance.

Lastly, dealing with outliers. To begin with, I created diagnostic plots for all the numerical variables to look at their distribution. Through this step, I found out that item visibility data is right skewed and contains anomalies. With that in mind, thought of windsorization and capping techniques. However, what made me choose the capping technique ultimately was due to the skewness of the data. Executing this step helps mitigate the impact of outliers by capping extreme values for the item visibility column. With that, the dataset becomes more consistent, allowing for more accurate analysis and modelling.

**Data Transformation**

Data transformation is essential for getting the data ready for analysis since it enhances data quality, meets specific needs, and makes modelling and analysis processes more precise and efficient. It aids in revealing patterns, connections, and insights in the data that might not be immediately obvious in their raw form. In this section, we will explore categorical encoding and discretization /binning are carried out.

Categorical variables are often encoded in strings. However, the Python library for machine learning does not support strings as values. Thus, we need to perform categorical encoding, to transform strings into numbers. Firstly, I looked through the columns of each categorical variable to identify what type of encoding method is most suited for each type of variable. With that, I categorised them into different encoding methods. Under one hot encoding, we have variables item type, outlet location type, outlet type and outlet identifier. Categorising them so is because they are nominal categorical variables that do not show any intrinsic order. Under ordered ordinal encoding, we have outlet size variable as it has an inherent order, for this case, "Small," "Medium," and "Large,". Therefore, I used ordered ordinal encoding to assign numerical values according to their order. During the process, I also considered using ordinal encoding for variable item types if there is an order or hierarchy among the categories. Ultimately, I used one hot encoding as it produced a better linear regression model.

The variables are assumed to be regularly distributed in linear regression. If not, we can perform discretization /binning for numerical variables and might reveal linear correlations between variables and the target (outlet item sales). This indicates that changing variables may help linear machine learning models perform better. With that in mind, I looked at the distribution of each of the variables and decided to use Logarithm transformation for item mrp and Square Root transformation for item visibility variables. This is because item mrp has a wide range of values. Using the logarithm of the variable can make the distribution more symmetric and reduce the impact of extreme values. However, after performing the transformation, I felt that the data changed too much and decided to use Box-Cox transformation for item mrp instead. Regarding the item visibility variable, what made me decide to use Square Root transformation rather than Logarithm transformation, was that the data contain the value 0, preventing me from using Logarithm transformation.

**Feature Engineer**

A key component of data pre-processing that is also important for machine learning and data analysis is feature engineering. It entails converting unprocessed data into a format better suited to modelling and analysis. The feature engineering technique and its justification for improving the performance of machine learning models are explored in this section.

The extent and magnitude of the features affect several machine learning algorithms in different ways. Particularly, the feature scale affects the coefficients of the linear models, meaning that altering the feature scale will alter the coefficients' value. As a result, when features are on a comparable scale, we can compare how important they are. With that, I went on to try Min-Max scaling and Standard scaling. After trying both scaling methods and comparing the variable distributions, I felt that Min-Max scaling would be a better option for my model. This is not only because it shows a better normal distribution, but also is more useful when preserving the range of values. With that, it prevents certain variables from dominating the model due to their larger values, allowing for more accurate modelling.

Often creating new features require one to have domain knowledge of the industry. It is done by combining 2 or more variables using mathematical functions. As such I went on to analyse the variables to see if there are any possible equations I can create. Unfortunately, I did not see any correlation in the limited number of numerical variables in this dataset. Therefore, I decided to use polynomial expansion to capture the complexity of the relationships in the data, ultimately improving model performance.

**Linear Regression Model**

A basic machine learning model called linear regression is used to find a correlation between independent factors and dependent variables. The goal of the model is to identify the straight line with the best fit, which minimizes the discrepancies between the predicted and actual values of the dependent variable. The goal of this study is to use the independent factors, or X variables, to predict the target Y variable, which in this case is item outlet sales.

Finding the ideal values for the intercept and coefficients of the linear equation is a necessary step in the linear regression procedure. The coefficients show the change in the item outlet sales variable corresponding to the change in each X variable, assuming other variables are maintained constant, whereas the intercept shows the value of the item outlet sales when all X variables are 0. The goal is to identify the optimal values that result in minimal errors, achieving a best-fit line.

We may calculate the link between the X factors and the item outlet sales using the linear regression model. The regression coefficients, which offer important details on the strength and direction of the link between each independent variable and the target variable, can be obtained by fitting the model to the dataset. Positive coefficients signify a positive association, which means that rising predictor variable levels are correlated with rising item outlet sales. Negative coefficients, on the other hand, imply an inverse relationship.

R-square is a measure of how much of the variation in the response variable can be accounted for by a linear model. R-square values that are closer to 1.0 are more indicative of improved model performance. The finished model has a value of 0.618 and 0.599, showing that the data wrangling procedures were successful, and the model is reasonably dependable for use. The fact that the value is not very near 1.0 further guarantees that the measures done are appropriate and won't overfit.

In conclusion, by using linear regression analysis, we may determine how independent factors and item outlet sales are related. We can obtain an understanding of the variables influencing sales success by constructing a best-fit line and optimizing the intercept and coefficients. Businesses can use predictive modelling to make well-informed judgments that will optimize their strategy and increase overall sales.

**Summary and Further Improvements**

In this report, the data wrangling process for the "Supermarket Sales Forecast" dataset is explored. Beginning with an introduction of the dataset and its objectives, the report discusses the value of data wrangling. Examining the dataset, dealing with missing data, correcting mistakes and inconsistencies, and carrying out data transformation operations are just a few of the different data-wrangling techniques that are covered.

Understanding the connection between variables and item outlet sales is the main goal of the data exploration. Understanding the connection between variables and item outlet sales is the main goal of the data exploration section. Examined are the connections between factors including outlet identifier, establishment year, item weight, fat content, visibility, kind, and MRP, as well as outlet size, location type, and type of outlet. The results of the additional study are excluded for variables like item weight and item. Item weight and item fat content are variables that do not show a distinct relationship with item outlet sales. Thus, both variables are being dropped.

The procedure used to clean the dataset is described in the section on data cleaning. Errors, missing values, and inconsistent data are all found and fixed. For instance, by substituting various naming traditions with standardized values, inconsistent values in the item fat content column are eliminated. Using imputation approaches based on non-null values, missing values in the item weight column are addressed. To achieve more accurate analysis and modelling, outliers in the item visibility column are addressed by using capping approaches.

The data is subsequently transformed using data transformation techniques so that it is ready for analysis. String values are converted into numerical representations via categorical encoding. Nominal categorical variables are encoded using the one-hot encoding, whereas variables having an intrinsic order are encoded using ordered ordinal encoding. It is possible to find linear correlations with the target variable by discretizing or binning numerical variables. Although square root and logarithm transformations are taken into consideration because of the data distribution, the item MRP variable eventually uses the Box-Cox transformation.

The data is subsequently transformed using data transformation techniques so that it is ready for analysis. String values are converted into numerical representations via categorical encoding. Nominal categorical variables are encoded using the one-hot method, whereas variables having an intrinsic order are encoded using the ordered ordinal method. It is possible to find linear correlations with the target variable by discretizing or binning numerical variables. Although square root and logarithm transformations are taken into consideration because of the data distribution, the item MRP variable eventually uses the Box-Cox transformation.

Overall, most of the techniques used did play a role in improving the final regression model value. This is apart from numerical variable transformations and feature scaling. Ultimately, I felt that this would be the best regression model for my existing ability. Perhaps in future, what I can work on is to strengthen my domain knowledge so that I can make the necessary comparisons and a more accurate evaluation of the dataset.