# Task 4: Building a predictive model for Dibs organization to be able to predict future sales.

1. **Data preparation:**

To develop a robust prediction model, we recommend creating a new dataset derived from the existing one. This novel dataset should incorporate the 'Order Date' and 'Total Sales' features. The 'Total Sales' variable can be computed by multiplying the 'Quantity Ordered' and 'Price Each' values. The target variable for sales forecasting will be the 'Total Sales' field. Given the time-dependent nature of the data, it would be prudent to leverage time series forecasting methodologies within the modeling pipeline.

* Date Conversion: The Order.Date column is converted to a Date format using ymd\_hms() from the lubridate package.
* Daily Sales: The dataset is aggregated to calculate daily sales using the dplyr package.
* Sales Trends: Daily sales data is visualized to identify trends and patterns over time, likely using ggplot2.

1. **Splitting and training data**

It is recommended to divide the dataset into training and testing subsets using an 80-20% split method. The training set should encompass all data up to a specific point and will be used to construct and refine the model. Conversely, the testing set, starting from the point where the training set concludes, is utilized to assess the model's predictive accuracy.

1. **Time series analysis**

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

What sets time series data apart from other data is that the analysis can show how variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data.

First of all, lets take a look at sales overtime (daily and weekly)A graph showing a line graph

Description automatically generated A graph showing a line graph

Description automatically generated

The time series plot shows a significant drop in sales around the beginning of 2020, followed by a period of very low sales. This sharp change could be due to an outlier event or data issue that needs addressing before choosing a forecasting model.

Base on time series graph on sale over time, it’s seemed that Dib’s sale data are seasonal and on trend, so instead of using SES(simple exponential smoothing), I’ll will Holt-winter’s SES because it can cover seasonal and trend data.

The second model that will be used is Linear regression model because of it simplicity.

1. **Forecasting**

In this topic, I will use Linear regression model and Holt winter’s SES(Simple Exponential smoothing.

Linear Regression and Holt Winters’ Simple Exponential Smoothing are two different methods used for forecasting or predicting future values based on historical data. The choice between these two models depends on the characteristics of your data and the assumptions. Here are some reasons to use each model for predicting sales:

Linear Regression:

1. **Linear Trend**: If your sales data exhibits a linear trend over time, linear regression can effectively capture and model this pattern. It assumes that the relationship between the predictor variable (e.g., time) and the target variable (sales) is linear.
2. **Explanatory Variables**: Linear regression allows you to include multiple explanatory variables (e.g., advertising spending, price, economic indicators) that may influence sales. This can provide valuable insights into the factors driving sales and improve the accuracy of predictions.
3. **Interpretability**: Linear regression models are relatively simple and interpretable. The coefficients of the model can be easily understood, providing insights into the magnitude and direction of the effect of each predictor variable on sales.
4. **Assumptions**: Linear regression makes certain assumptions, such as linearity, normality of residuals, and homoscedasticity (constant variance of errors), which may be reasonable for some sales data.

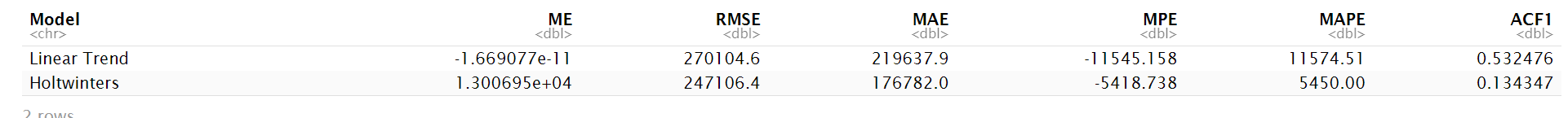
Holt-Winters' Exponential Smoothing is an extension of Simple Exponential Smoothing, designed to handle data with both trend and seasonal components. Here are some reasons why you might consider using Holt-Winters' Exponential Smoothing for predicting sales:

1. **Trend and Seasonality**: If your sales data exhibits both trend and seasonal patterns, Holt-Winters' method can effectively capture and model these components. Many businesses experience seasonal fluctuations in sales due to factors like holidays, weather, or consumer behavior patterns.
2. **Flexibility**: Holt-Winters' method offers flexibility by allowing you to choose between additive or multiplicative seasonality, depending on whether the seasonal variations are constant or proportional to the level of the series.
3. **Adaptive to Changes**: Like Simple Exponential Smoothing, Holt-Winters' method is responsive to recent changes in the data. It can adapt to shifts in the level, trend, or seasonal patterns, making it useful for sales data that may be influenced by changing market conditions or consumer preferences.
4. **Automatic Smoothing**: Similar to Simple Exponential Smoothing, Holt-Winters' method automatically smooths out random fluctuations or noise in the data, providing a cleaner representation of the underlying patterns.
5. **No Data Transformation**: Unlike some other time series methods (e.g., ARIMA models), Holt-Winters' Exponential Smoothing does not require differencing or other data transformations, making it more straightforward to apply and interpret.

A graph showing a graph of a graph

Description automatically generated with medium confidenceHolt’s Simple Exponential smoothing Model) A graph showing a graph showing a number of different types of data

Description automatically generated with medium confidence

Accuracy

From above table, Holt model is better than Linear regression model on overall performance indicating that it’s more suitable to predict sales so I would recommend using Holt’s over than Linear regression. Even though the Holt model is better and seem to be a good model but in fact, Holt model required the data to be:

* Trend: The data should exhibit a trend component, meaning that the data shows an increasing or decreasing pattern over time.
* Seasonality: The data should exhibit a seasonal component, meaning that there are repetitive patterns or cycles within the data.

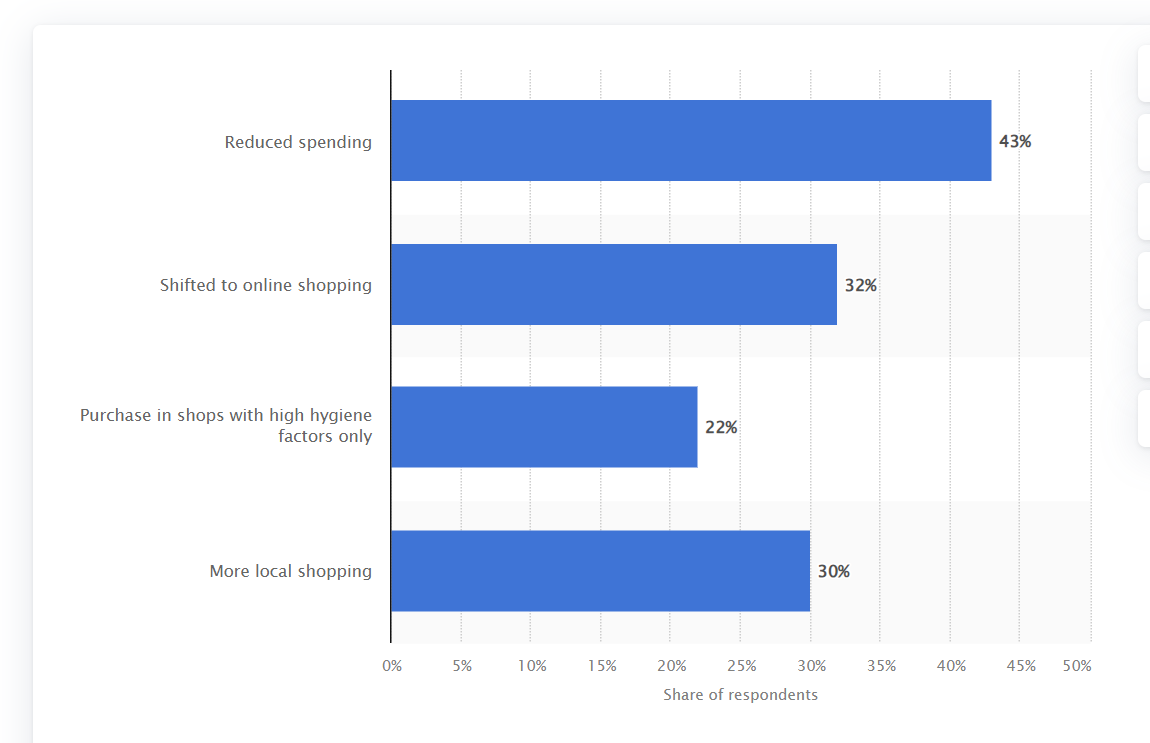
So the missing previous year’s sale data affect a lot on how this model performs thus the prediction are less reliable

Personal assessment: Since this data is incomplete without previous year’s data and in 2019-2020 period, there was black swan event happened (Covid-19) which has a huge impact on Supply chain, Customer behavior, Government’s policy…. Which would make this data very hard to be used to predict future sale.

To study further about company’s data and provide a further analysis, future sale-marketing strategies, I suggest studying about customer behavior and market-trend.

1. **Customer behaviors**

Since 2019, after lock-down happened across global. The demand for online shopping was skyrocketing because people can’t move around that easy anymore. online shopping has always been a potential market for every company in any industry.

The graph depicting that custommer tent to switch from physical to online shopping due to lock down and social distance.

The data was about 2019-2020-2021, coincidently matching the Covid period which can explain the plummeting in Dib’s sale. For most company, the change was too fast to adjust. Dib company probably cannot adapt the new market trend and customer’s service despite having store across many state in the US, Dib’s sale are so frustrating.