International Business Research; Vol. 16, No. 5; 2023
ISSN 1913-9004 E-ISSN 1913-9012
Published by Canadian Center of Science and Education

Using Statistics for Market Analysis Forecasting

Thanakit Ouanhlee[1]

[1] Independent Researcher, Thailand

Correspondence: Thanakit Ouanhlee, Phunthainorrasing, Muang, Samutsakhon, Thailand. E-mail: eric.ouanhlee@gmail.com

Received: March 20, 2023 Accepted: April 20, 2023 Online Published: April 20, 2023

doi:10.5539/ibr.v16n5p12 URL: https://doi.org/10.5539/ibr.v16n5p12

Abstract

Market analysis is a crucial aspect for any organization, business, or company because it provides a ground for decision making. Poor market analysis leads to poor decisions. On the other hand, using quality data to conduct market analysis can provide significant grounds for informed decisions. Business sectors require a clear view of future trends regarding the performance of their products, sales, stocks, employees, and customers, among others. However, defining patterns is possible only through statistical techniques of forecasting. In essence, the knowledge of market analysis forecasting using statistical tools is imperative. This article aims at providing a summary of market forecasting techniques, highlighting their interesting discoveries, and outlining some practical applications in real life. The summary covers regression analysis, handling of special events, identification of seasonality, Holt–Winters method, and forecasting for new products. Regarding regression analysis, it was found that data cleaning is an important aspect of this analysis before the actual forecasting. The data must be tested to meet the reliability and validity criteria to ensure quality data are used for forecasting. The interesting discovery with regard to handling special events was that some special events have great ripple effects, which an organization needs to plan for. Furthermore, when doing an analysis of data, it is essential to take into account the effects of seasonality. It was also ascertained that the accuracy of the Holt–Winters method is associated with its use of smoother curve, which allows a researcher to smooth time series data to make predictions. The article further illustrates that the Bass diffusion model provides more accurate forecasts than logistics and Gompertz models givens its ability to put into consideration the external and internal influence when forecasting sales of new products. One of the applications of this study is that regression models can be used in studying the effectiveness of advertisement platforms during a product marketing campaign. Sales companies can apply seasonality forecasting to understand the influence of different seasons on their products. Moreover, the data on customers' expenditure patterns can be used to forecast special events to aid in proper planning. Therefore, any business, firm, industry, or country can use forecasting to predict different components of a market.

Keywords: market analysis, market forecast, marketing seasonality forecasting, statistic, marketing statistic, statistic for marketing, analysis forecasting

1. Introduction

Market analysis has proven beneficial in industries, companies, and organizations as it helps understand the business industry at large and provides a specific view of a company's competing brands. In addition to identifying market gaps, market analysis has been hailed as a solution to finding and understanding potential target markets and creating sales forecasts. The aim of this article is to outline the key concepts/theories of market analysis forecasting, citing their basis, importance, and comparisons. The following ideas are discussed in this article: regression modeling, trends and seasonality, Holt–Winters method, and forecasting techniques. The topics discussed under regression modeling are data exploration, univariate analysis, descriptive data analysis, outlier treatment, missing value treatment, creation of dummy variables, correlation analysis, simple and multiple linear models, and assessment of the output of regression models using *R*-squared, *F*-statistic, and root-squared error. This article also covers handling of special events. In addition, the paper discusses the identification of seasonality in a time series data using special models including additive model, multiplicative model, and moving average method.

The Holt–Winters method of handling changing trends is discussed to ascertain how a time series data with trends can be modeled. Moreover, the article incorporates the use of logistic curve, Gompertz curve, and Bass

diffusion curve to show how forecasting is carried out for a new product in the market. The need to understand the market and provide insightful information to businesses necessitated the study of these market analysis techniques. Merket analysis is a key component to a business's success in the modern world. The data garnered from market analyses can direct a business to the most productive route as well as predict changes and direct volatility in the market. In other words, forecasting helps a business remain a few steps ahead and be pre-emptive with its strategies rather than reactive.

2. Regression Model

Regression model is a statistical approach that is used in a variety of fields, such as finance and investing, to assess the nature and magnitude of the connection that exists between one dependent variable (Y) and a range of independent variables (*X*) (Beers, 2022). Some of the preliminary steps needed before the actual forecasting using regression model are discussed below:

2.1 Gathering Business Knowledge

The preliminary stage of adopting a regression model is gathering knowledge about a business. Businesses have access to huge pools of raw data from which they extrapolate information about their customers' needs and their staff's skills, experience, and competencies. The way a business gathers, shares, and exploits data can provide a solid basis to its ability to develop successfully. The knowledge gathered by a business can be beneficial to everyone from a local viewpoint to a manufacturing firm. Businesses already possess raw data in the form of employees' input and feedback, records of products, documents in paper and digital forms, as well as future plans. However, the common challenge is harnessing forecasts from existing data in a coherent and productive way for modeling (Info Entrepreneurs, 2021).

Basic sources of business knowledge include (a) customer knowledge: Every business should understand their customers' needs and how they perceive the business. This could assist in developing mutually beneficial knowledge-sharing relationship by customers through engagements on future requirements. (b) Employee and supply relationship: This can be done by seeking the opinions of employees and suppliers. (c) Market knowledge: This helps in checking new developments in the respective sector regarding the performance of competitors, their selling prices, and new products launched, among others. (d) Knowledge of business environment: Any business can be affected by external factors such as politics, economic shifts, new technological progress, and societal trends. There is need to constantly assess these changes in the business world. (e) Professional associations: Publications, academic publications, government publications, and research reports could be great sources of business knowledge. (f) Trade exhibition and conferences: Every business can use this platform to find out what other players are undertaking to keep up with the latest innovations. (g) Product research and conferences: This is an important source of knowledge that can help a business create new products while retaining competitive edge.

When business knowledge is properly gathered and stored, it leads to a range of benefits such as improved goods and services offered and refined processes that are used to sell them; improved customer satisfaction and customer loyalty due to greater understanding of their requirements; increased quality of supplies because of the business's awareness of customers' needs; increased staff productivity as they feel highly appreciated in a business where their ideas are considered; improved efficiency in the business using in-house expertise; improved staffing and recruitment policies where the right staff serve different customers (Info Entrepreneurs, 2021).

2.2 Data Exploration

Data exploration is a preliminary data analysis where visual exploration is used to understand the contents of a dataset and the characteristics of the data. The characteristics of the data may include size of the data, validity of the data, reliability of the data, possible patterns among different variables in the data, or table in the data. The process of data exploration is normally conducted using automated and manual methods. Automated methods comprise data profiling, data visualization, or tabular reports presenting a clear initial view of the data and their characteristics. After the generation of automated reports, the data are filtered to identify issues. Exploration of data also requires manual scripting and queries using different languages and software such as Microsoft Excel (also commonly known as Excel), R, and Python. Visualization can be done using bar charts, histograms, line graphs, pie charts, and scatter plots, among others. The whole process of data exploration is aimed at understanding data in terms of statistics, structure, and relationships that can be used in further analyses. The data are then refined by removing unwanted parts in a process called data cleaning, thus correcting wrongly formatted elements while also redefining appropriate relationships, which is called data quality check (Rekleits, 2015).

2.3 Univariate Analysis and Exploratory Data Analysis

Exploratory data analysis is used to ascertain relationships among measures in the data and to gain insights on patterns, trends and relationships among different components of a dataset using statistics and visualization tools. Exploratory data analysis is classified into graphical and non-graphical methods, and each method is univariate, bivariate, or multivariate. In univariate analysis, only one dependent variable exists. Univariate analysis aims at deriving the data, defining and summarizing them, and analyzing the pattern present. It explores categorical and numerical variables. In univariate analysis, the following patterns can be identified: measures of central tendency (mean, mode, median), measures of dispersion (range, variance, standard deviation), and quartiles (interquartile range), among others (Shah, 2021).

Univariate analysis can be explored through:

- Frequency distribution table: This highlights the frequency of an occurrence by providing an overview of the data, making it easier to discern patterns.
- Bar chart: This is very useful when comparing different categories of data or groups of data (preferred when visualizing discrete data such as marital status).
- Histogram: This is used to display categories as bins to show the number of data points in a range (used when visualizing continuous data such as salaries).
- Pie chart: This is instrumental when assessing how a group is broken down into smaller pieces, where the slices denote the relative size of each category.
- Frequency polygon: This is useful when comparing datasets or displaying cumulative frequency distribution.

2.4 Descriptive Data Analytics in Excel

Descriptive data analysis describes, shows, or summarizes data points in a constructive way where patterns emerge to fulfill different conditions. Descriptive analysis helps in drawing conclusions regarding distributions of the data, helps in detecting outliers, and enables a researcher to identify similarities among variables for further analysis. Descriptive analysis incorporates data aggregation and data mining, and it also includes (Rawat, 2021)

- Construction of tables of means, measures of dispersion such as standard deviation and variance, and cross-tabulations.
- Measures such as discrimination and inequality, which can be studied using advanced descriptive methods.
- A table of means disintegrated into subgroups, which is normally used to indicate differences across subgroups, after which conclusions are derived.
- A cross tab or two-way tabulations, which are also used to show the proportions of variables with unique values for each of the variables available.

Descriptive analysis (Figure 1) can be categorized into four types: measures of frequency, measures of central tendency, measures of dispersion or variation, and measures of position. It is imperative to understand how a certain event is likely to occur. Measures of frequencies are used to ascertain this phenomenon by reporting counts and percentages of a variable (Rawat, 2021).

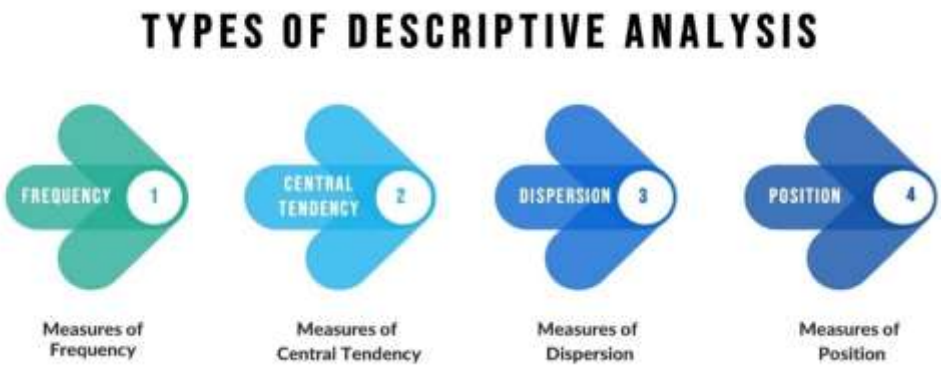

Figure 1. Types of Descriptive Analysis

Sources: Image created by the author

In Excel, descriptive analysis is conducted as follows: (a) on the *Data* tab, in the *Analysis* group, click *Data Analysis*, (b) select *Descriptive Statistics* and click *OK*, (c) select the range of the data you want to analyze, (d) select an output range, and (e) check the summary statistics box and click *OK*. The output reports mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, and count. Descriptive statistics is highly considered because of high degree and objectivity of researchers due to different characteristics of the data reported. Furthermore, descriptive analysis is more widespread than other quantitative methods and gives a broader understanding of an event. Descriptive analysis forms a basis for identifying variables and new hypotheses, which can be further assessed through inferential studies (Rawat, 2021).

2.5 Identifying and Treating Outliers in Excel

Outliers are values within a dataset that vary greatly from the others, that is, they can be either much larger or significantly smaller. They may indicate variabilities in a measurement or experimental errors (Sequitin, 2021). When entering, analyzing, and interpreting data, outliers can cause a significant change that affects the accuracy of a report. Assessing these outliers can help in identifying them and minimizing potential discrepancies they might cause.

In Excel, an outlier is a data point or set of values that is significantly different from the average expected range in a sample. Outliers influence data interpretations and cause inaccurate results through their differences from the rest of the data. Therefore, identification, calculation, and minimization of these outliers are key to ensuring accurate reports. In Excel, outliers can be calculated using the following steps (Indeed Editorial Team, 2022):

1. Reviewing the entered data: Data entered in the spreadsheet should be reviewed and verified to detect and fix errors that may create inaccuracies.
2. Sorting data values: In this case, the range of the dataset is selected, and then on the top function ribbon in Excel, *Sort & Filter* option in the *Home* tab is clicked. Then, *Custom Sort* is selected to order the dataset from smallest to largest and to implement the changes.
3. Analyzing the values: After sorting the values, large data discrepancies and outliers can be identified and eliminated.
4. Identifying data quartiles: To identify outliers in the dataset, quartiles are calculated using the quartile formula in Excel "=QUARTILE()." After adding the left bracket, the first and last cells in a data range are specified and separated by a colon, followed by a comma and then the quartile. For instance, the formula may look like this: "=QUARTILE(A5:A20, 1)" or "=QUARTILE(C3:C30, 3)."

   Where: A5 is the first cell of the data range and A20 is the last cell of the data range for the first quartile.

   C3 is the first cell of the data range and C30 is the last cell of the data range for the third quartile.
5. Defining the interquartile range: This range is calculated by subtracting the first quartile from the third quartile (e.g., "D2-D1").

Where: D2 is the third quartile and D1 is the first quartile

6. Calculating upper and lower bounds: Establishing the upper and lower bounds helps in identifying outliers. In Excel, the upper bound is found by multiplying interquartile range by 1.5 and adding it to the third quartile value: "=D2+(1.5*D4)." On the other hand, to find the lower bound, the interquartile range is multiplied by 1.5 and subtracted from the first quartile: "=D1-(1.5*D4)."

   Where: D4 is called the interquartile range.

7. Removing the outliers: After the upper and lower bounds are defined, the values that are higher than the upper bound or lower than the lower bound are removed from the dataset, or they can be adjusted to match the highest value in the average range (Indeed Editorial Team, 2022).

2.6 Identifying and Treating Missing Values in Excel

Missing data occur when there are no data stored for certain variables or participants. This could be a result of incomplete data entry, equipment malfunction, or lost files. Missing data can be categorized into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are considered MCAR if the probability of any missing value from the dataset is not related to anything else; for instance, some people started answering a survey but left out a question. Data are said to be MAR if the likelihood of missing data is related to another observed variable; for example, there are more missing values for males than that for females. MNAR data are related to the values themselves.

In Excel, missing data can be identified using the VLOOKUP function. VLOOKUP returns a #N/A error if a value is not found in the list. The VLOOKUP function is used with ISNA function to find missing values. The VLOOKUP function is defined as follows in Excel:

=IF (ISNA (VLOOKUP (value, range, 1,0)), "MISSING", "OK")

Missing data can cause problems and can sometimes lead to sampling bias, thereby hindering the generalizability of the findings due to unrepresentativeness of the sample (Bhandari, 2021). There are four ways of dealing with missing data in Excel (Garcia, 2017):

- Listwise deletion: In this scenario, all rows that have one or more column values missing are deleted. Missing values would often be deleted since they cannot contribute meaningfully to the research.
- Mean/Median/Mode imputation: In this case, all missing values in a given column are substituted with mean/median/mode calculated from all the values available in the respective column. In Excel, functions such as =AVERAGE(), =MODE(), and =MEDIAN() are used to compute mean, mode, and median, respectively.
- Last observation carried forward (LOCF): This method is used in longitudinal data where the missing value is replaced with a value available at the previous stage.
- Resurveying: In this method, data points are recollected from the respondents. This method is highly appropriate because it ensures that the missing observation is filled with an accurate rather than approximate value.

2.7 Variable Transformation in Excel

Variable transformation enables the data to work better in the model. To transform variables in Excel, multiple functions can be used to expedite the process. Variable transformation in Excel can be used in reporting, analyzing statistical data, performing mathematical operations, compiling financial data, and analyzing business analytics, among others.

Excel has a vast number of functions that can be used to transform data, change the spreadsheet appearance, and perform various mathematical operations. Some of the functions are as follows: (a) LOG10(), which allows for calculation of logarithm to base 10; (b) SQRT(), which allows for calculation of square root; (c) ABS(), which returns an absolute value of a number; (d) DATE(), which returns the sequential serial number that represents a particular date; (e) AVERAGE(), which returns the mean value of a particular variable. Further, Power Query can be used to extract, transform, and load data, and Power Pivot can be used to link imported spreadsheets, filter and sort them, and perform mathematical and logical operations.

2.8 Dummy Variable Creation in Excel

A dummy variable is a binary variable that takes a value of 0 or 1. In this regard, a numerical variable called a dummy variable can be used to represent a categorical variable in a statistical model as though it were a numerical variable that might have either of the two values. In Excel, suppose "marital status" is a categorical

variable with three levels: “single”, “married,” and “divorced.” Two dummy variables for married and divorced can be created, by making “single” a baseline value. Suppose the column representing “marital status” is C2, a dummy variable “Married” can be created as follows:

=IF(C2= “Married”, 1, 0)

The dummy variable “Divorced” is created as follows:

=IF(C2= “Divorced”, 1, 0)

2.9 Correlation Analysis

Correlation analysis is a test statistic that measures the statistical relationship, or association, between two continuous variables. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. The Pearson correlation analysis is widely used due to its ability to measure the association among variables of interest because it is based on the method of covariance. The assumptions of Pearson correlation include independence of cases, existence of linear relationship, and presence of homoscedasticity. The degree of correlation can be perfect, high, moderate, or low or show no correlation.

- Perfect correlation: Two variables are said to have a perfect correlation if the value of correlation coefficient (r) is near +/-1.
- High correlation: Two variables are said to have a high/strong correlation if r lies between +/-.5 and +/-.75.
- Moderation correlation: Variables are said to exhibit moderate correlation if r is between +/-.3 and +/-.49.
- Low correlation: Two variables have low/small correlation if r lies below +/-.29.
- No correlation: There is no correlation between two variables if $r = 0$.

In each case if r is positive, there exists a positive correlation between the two variables. If the value of r is negative, the variables are said to have a negative correlation.

In Excel, correlation analysis is conducted as follows: (a) select *Data* from the top bar menu, (b) select *Data Analysis*, (c) select *Correlation* and click *OK*, and (d) define the data range and click *OK*. This produces a correlation matrix in Excel.

3. Regression Model for Forecasting

3.1 Problem Statement

For forecasting, regression model requires a problem statement as a prerequisite citing what the model is intended to answer. Most of the problem statement entails the specification of the aim of the forecasting, that is, its objectives and specific hypothesis the model is designed to answer. A clear identification and statement of the problem provides a clear guide to regression analysis.

3.2 Simple Linear Regression Model

Simple regression model, also called ordinary least squares (OLS) model, is the most common form of regression model. The linear relationship that exists between two variables can be determined through the use of a linear regression by determining the line that provides the greatest fit. Graphically, linear regression is demonstrated using straight line with the slope showing how a change in one variable affects another variable (Beers, 2022). Simple linear regression is used to measure the relationship between one independent variable and one dependent variable. In essence, it allows a researcher to ascertain how a dependent variable changes with a change in the independent variable. The equation for linear regression model is as follows:

$$Y = B_0 + B_1X + €$$

Where,

Y is the predicted value of the dependent variable (y) at any given value of the independent variable (X),

B_0 is the intercept, that is, the predicted value of y when x is 0,

B_1 is the regression coefficient, showing the rate of change in y as x increases,

X is the independent variable, and

€ is the error term, that is, the variation in the estimate of the regression coefficient.

3.2.1 Hypotheses for Simple Linear Regression Model

Null hypothesis H_0: The independent variable (x) is not significant in predicting the dependent variable (y).

Alternative hypothesis H_1: The independent variable (x) is a significant predictor of dependent variable (y).

$$H_0: B_1 = 0$$

$$H_1: B_1 \neq 0$$

3.3 Multiple Linear Regression Model

A statistical method known as the multiple linear regression model is one that makes use of a number of different explanatory factors in order to make a prediction regarding the result of a response variable. It is an extension of OLS model because it models more than one explanatory variable (Hayes, 2022). The equation for multiple regression model is as follows:

$$yi=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta p x_{ip}+\epsilon$$

Where, for $i = n$ observations,

yi = dependent variable,

Xi = explanatory variables,

β0 = y-intercept (the predicted value of y when all the predictors are 0),

βp = slope coefficient of each explanatory variable, that is, the rate of change in y per unit change in x, and

ϵ = the error term of the model.

3.3.1 Hypotheses of Multiple Linear Regression Model

Null hypothesis H0: All the independent variables (x) are not significant in predicting the dependent variable (y).

Alternative hypothesis H1: At least one of the independent variables is significant in predicting the dependent variable.

$$H_0: B_1 = B_2 = B_3 = \ldots. = B_i = 0$$

$$H_1: B_1 \neq B_2 \neq B_3 \neq \ldots. \neq B_i \neq 0$$

3.4 Assumptions of Linear Regression Models

- There is a linear relationship between a dependent variable and an independent variable.
- There is a minimum or no presence of multicollinearity among the independent variables.
- The dependent variable should be normally distributed.
- The variance of the residual is the same for any value of x (homoscedasticity).

3.5 Assessing Accuracy of Predicted Coefficient of Regression

These two methods were used to assess the accuracy of regression coefficients: residual standard error and *R*-squared values.

3.5.1 Residual Standard Error

This is an estimate of the standard deviation, that is, the average deviation between the actual outcome and the true regression line. It is calculated as

$$\text{Residual standard error} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{df}}$$

Where,

y: The observed value,

$\hat{y}$: The predicted value, and

df: The degrees of freedom, calculated as the total number of observations-total number of model parameters.

The smaller the standard error, the more accurately a regression model fits the dataset. On the other hand, the higher the residual standard error, the worse a regression model fits the dataset. In essence, a regression

coefficient with low standard error has a high accuracy of estimation and precision.

3.5.2 R-squared (R^2)

In a regression model, R-squared is a method of statistical analysis that determines what portion of the variation in a dependent variable may be attributed to what percentage of the variance in the independent variables. It is a measure of how well the data fit the regression model. For instance, an R-squared value of .67 shows that 67% of the variation in the dependent variable is explained by the independent variable. This also indicates that 33% of the variation in the dependent variable is explained by variables not included in the model. A higher R-squared value indicates better explanation of variability by the model. Furthermore, an R-squared value can also be used to compare models as models with higher R-square values perform better than other specified models.

3.5.3 F-Statistic in Regression

In a linear regression model, F-statistic is used to test the significance of regression models. It is normally calculated as regression mean square and mean-squared error, where the former is the mean of the sum of squares and the latter is the mean of the squares of the errors (Frost, 2017). The F-statistic is based on the following hypotheses:

Null hypothesis: The null model (model with no predictors) fits the data as well as the alternative model (model with predictor variables).

Alternative hypothesis: The alternative model fits the data better than the null model.

The regression output presents the *F*-calculated value and the corresponding *p* value. The traditional method of hypothesis testing suggests that if the $f_{calculated}$ is greater than $f_{tabulated}$, the null hypothesis is rejected. In many scenarios, the use of *p* value is highly preferred and decision is made as follows: A comparison of *p* value and level of significance (alpha = .05) is used and if the *p* value is less than .05, the null hypothesis is rejected and it is concluded that the overall model is significant, that is, the alternative model performs better than the null model. If the *p* value is greater than .05, the null hypothesis cannot be rejected and it is concluded that the overall model is not significant.

3.6 Interpreting Results of Categorical Variables

Independent variables in a linear regression model can take continuous or categorical form. A categorical variable cannot be quantified and can be either nominal or ordinal. For instance, gender is a nominal variable, which also qualifies as a dummy variable. An example of an ordinal variable is the level of education. When a nominal variable has more than two categories, for example, marital status, it is normally converted to a dummy variable that can take on a value of 0 or 1. Assuming a researcher is interested in modeling the effects of gender and level of education on income, the equation would be formulated as follows:

$$Y = B_0 + B_1X_1 + B_2X_2$$

Where,

y is the income,

X_1 is gender, and

X_2 is level of education.

Assuming the output of the model is

$$Y = B_0 + 1236X_1 + 236X_2,$$

a positive coefficient of gender (0 “male”, 1 “female”), that is, $B_1 = 1236$, would mean that females have a higher income than males by $1,236. Here, males have been used as a baseline. For the level of education ($B_2 = 236$), it is interpreted that the amount of income increases by $236 as the level of education increases.

3.7 Testing for Significance of Model Coefficients

The regression output displays the coefficients, the p values, and the confidence interval. A negative coefficient shows that as the independent variable increases, the dependent variable decreases by B_i. On the other hand, a positive coefficient shows a positive correlation.

To check for significance of the independent variables, *p* value and level of significance are compared such that if the *p* value is less than the level of significance (.05), the null hypothesis is rejected and it is concluded that the respective independent variable is significant in predicting the dependent variable. Otherwise, the null hypothesis cannot be rejected if the *p* value is greater than .05. In addition, if the confidence interval does not contain zero,

the null hypothesis is rejected.

3.8 Interesting Discoveries Regarding Regression Analysis

Regression analysis is instrumental in unearthing the relationship among variables that cannot indicate causation. It is worth noting that before the actual modeling of the data in regression, it is imperative to conduct data cleaning (data transformations, outliers' removal, missing data imputation, checking for regression assumptions, etc.) and have an overview of the variables in the study through data visualization. Failure to conduct data cleaning can lead to massive poor decision making, which can mislead any organization or business. In addition, data used for regression must also pass the reliability and validity test in order to have adequate data for modeling. The majority of output problems, such as low estimation accuracies, emanate from the onset of the data collection; in other words, the issues could be rarely associated with the actual method of data analysis. This shows the importance of screening data that enter the spreadsheet of any analysis tool.

4. Handling Special Events

4.1 Special Event Forecasting

Forecasting is used to predict the outcome of an event that will aid in the planning of a particular project. Special event forecasting entails planning for events that do not follow normal patterns. This could include special sales campaigns, maintenance periods, new product launches, or holidays. Essentially, these events require extra attention during forecasting to ensure a business meets or surpasses the normal performance goals. It is also imperative to note that sometimes high sales growth do not occur during the event itself, but rather either before or after it. For instance, sales can go through the roof on the Eve of Christmas, but will drastically go down on Christmas itself and the day after. Some of the benefits of forecasting special events are

- Improved customer experience during the unavoidable ups and downs of customer demands: By using the available data, a best forecast is created, thereby maintaining positive customer experience regardless of obstacles.
- Wise expenditure of resources: A well-defined special event forecast leads to more accurate predictions on how much overtime will be needed for the event.
- Preparation for the ripple effects of the events: The forecast helps in planning for the ripple effects by looking at historical data regarding surrounding days.

Building forecast for special events requires looking at previous years' holiday patterns, comparing the holiday with the weeks leading up to it, adjusting forecasts accordingly, and conducting a post-mortem analysis.

4.2 Running Linear Regression Using Solver in Excel

Solver method is more complex than majority of the tools used for linear regression in Excel. However, it provides a background on how other methods of linear regression work in Excel. The process of linear regression using Solver method involves

1. Entering values for the slope and intercept of the equation in Excel.
2. Calculating new values of *y* variables based on those values of slope and intercept of the equation.
3. Calculating the error term between the calculated *y* values and observed *y* data.
4. Using the Solver add-in to find values of the slope and intercept that minimize the total error.

For instance, assuming column B contains *x* values, column C contains *y* values, column E contains guessed values of slope (m), and column F contains the values for the intercept (b), the following equations are obtained:

$$y\text{-calculated} = m*x + b$$

error = y-calc-y

Then, sum of the squared errors (SSE) is calculated from the Excel function as

$$SSE = SUMSQ \text{ (range of error)}$$

At this point, Solver is used to minimize the values of SSE, allowing the value in the error column to be driven to its minimum absolute value.

The Solver can be opened in Excel in the *Data* tab, and then the SSE value is entered under set objective in the *Solver* dialogue, and both the slope and intercept are changed and the new values of slope and intercept are created.

4.3 Interesting Discoveries Regarding Special Event Forecasting

Interestingly, historical data are required to understand the unique impact of special events. In case there are no perfect historical examples, data about past events that are currently applicable can be useful. It is also important to note that some holidays move dates every year; therefore, impact multiplier must be applied to the correct date. Moreover, many events can have several ripple effects and may require a clear plan for better analysis**.**

5. Identifying Seasonality

Seasonality is one of the features of a time series in which the data experience regular and predictable changes that recur every calendar year. Any pattern that repeats over a 1-year period is said to be seasonal (Kenton, 2020). Some of the models used to identify seasonality include additive model, multiplicative model, and moving average model.

5.1 Additive Model in Excel

In additive model, when the anticipated value for each data piece is the sum of the baseline component, the trend component, and the seasonality component, then seasonality is regarded to be additive. In additive model, seasonality is constant in size (Zaiontz, 2021). The recursive approach to the additive model is presented as follows:

$$u_i = \alpha(y_i - s_{i-c}) + (1-\alpha)(u_{i-1} + v_{i-1})$$

$$v_i = \beta(u_i - u_{i-1}) + (1-\beta)v_{i-1}$$

$$s_i = \gamma(y_i - u_i) + (1-\gamma)s_{i-c}$$

Where, $0 < \alpha \leq 1$, $0 \leq \beta \leq 1$, and $0 \leq \gamma \leq 1$. The predictions for the data elements y_i are given by

$$\hat{y}_i = u_{i-1} + v_{i-1} + s_{i-c}$$

For forecasts at future times, the following equation is used:

$$\hat{y}_{i+h} = u_i + hv_i + s_{i+h-ch'}$$

Where, h' = INT((h–1)/c) + 1.

5.2 Multiplicative Model

Multiplicative model is useful when seasonal variation increases over time. If c is the length of seasonal cycle, then c = 12 stands for months in a year, c = 7 for days in a week, and c = 4 for quarters in a year (Zaiontz, 2021). The recursive form of the model is as follows:

$$u_i = \alpha(y_i/s_{i-c}) + (1-\alpha)(u_{i-1} + v_{i-1})$$

$$v_i = \beta(u_i - u_{i-1}) + (1-\beta)v_{i-1}$$

$$s_i = \gamma(y_i/u_i) + (1-\gamma)s_{i-c}$$

Where, $0 < \alpha \leq 1$, $0 \leq \beta \leq 1$, and $0 \leq \gamma \leq 1$. The values of u_i variable depict the baseline, the values *of* v_i variable depict the trend (also known as the slope), and the values of s_i variable depict the seasonality component. In multiplicative model, the sum of the s_i values for any consecutive c periods of time is approximately equal to c (at least for reasonable values of α, β, γ). The prediction for the elements of the data *yi* is given by

$$\hat{y}_i = (u_{i-1} + v_{i-1})s_{i-c}$$

For forecasts at future times, the following equation is used:

$$\hat{y}_{i+h} = (u_i + hv_i)s_{i+h-ch'}$$

5.3 Moving Average Method

A moving average is a calculation that depends on a series of averages from data subsets within an entire dataset. It shows changes in averages as new data become available. Moving average is important because it helps in updating average prices, mitigating data errors associated with short-term fluctuations, and detecting changes in momentum for security (Indeed Editorial Team, 2023).

In Excel, time series data are created and arranged according to a time order. The sequence comprises of discrete time data from successive points in time that are spaced equally. After the creation of a time series data, from the Data tab, Data Analysis is selected and a dialogue box appears. In the dialogue box, the Moving Average option is selected, which is used to run the moving average analysis. Subsequently, the input and output ranges are entered. For instance, the selection values could look like "B2: L2."

The next step involves choosing an interval, that is, the set amount across which the values are to be averaged. For example, an interval of 10 shows that the average would consist of most recent data points along with the previous nine data points. The output range is selected from the first cell that relates to the first interval data point. In essence, the larger the interval, the more the peaks and valleys are smoothed out. The smaller the interval, the closer the moving averages are to the actual data points.

5.4 Interesting Discoveries About Seasonality

When seasonality is not taken into consideration, an investor may choose to buy or sell securities based on the activity at hand without accounting for seasonal change that occurs during the business cycle. However, it is imperative to consider the effects of seasonality when analyzing, for example, stocks from a fundamental point of view because they are associated with large impacts on investors' profits. In addition, it is interesting to note that economists can get a clear picture on how economy is moving when they adjust their analysis based on seasonal factors.

6. Holt–Winters Method

The Holt–Winters method is also called triple exponential smoothing, which is an incredibly popular and simple method for time series forecasting. The method is capable of forecasting both trends and seasonality. It is a combination of three other components, which are all smoothing methods (Ariton, 2021):

- Simple exponential smoothing (SES): SES assumes that the time series has no change in level and hence cannot be used with series that contains trends, seasonality, or both.
- Holts exponential smoothing (HES): This is one step beyond the SES as it allows the time series data to have a trend component. However, it cannot cope with seasonal data.
- Winters's exponential smoothing (WES): This is an extension of HES that allows for the inclusion of seasonality. Therefore, WES is what is called the Holt–Winters method.

6.1 Simple Exponential Smoothing

This method does not take into consideration trend or seasonality and only assumes that the time series data has an L level. The method is presented as follows:

$$L_t = \alpha y_t + (1 - \alpha) L_{t-1}$$

Where, y_t is the value at current step t, L_t is the level estimate for t, L_{t-1} is the previous level estimate, and α is a smoothing constant. The equation is considered to be recursive because every level estimate must be computed from previous estimates. In essence, the SES method is just a weighted average across all time periods, with the weights exponentially decaying (Ariton, 2021). Moreover, in SES forecasting, the future time step is the level of the current time step. Therefore, to improve the forecast, a method that is capable of handling trend must be incorporated.

6.2 Holt's Exponential Smoothing

This method takes into consideration the trend component; hence, it is commonly referred to as double exponential smoothing. Holt's forecast method can be presented as follows:

$$F_{t+k} = L_t + kT_t$$

Where,

L_t is the level estimate at time *t*,

k is the number of forecasts into the future, and

T_t is the trend at time *t*.

The trend estimates for a given time are computed as follows:

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$$

The equation is normally referred to as trend update equation since it updates the trend estimate of the current time step based on the difference between the previous level estimates. β is the trend equation's own smoothing constant. The values of β range from 0 to 1.

An example of a forecast is illustrated in Figure 3. The grey bar is the forecast area, while the orange line is the forecast. Blue represents the actual series, while green is the trend component (Ariton, 2021).

6.3 Winters's Exponential Smoothing

Winters's Exponential Method allows for capturing of seasonal components. It assumes that the time series has a level, trend, and seasonal component; the equation for the forecast using Winters's method is as follows:

$$F_{t+k} = L_t + kT_t + S_{t+k-M}$$

Where,

L_t is the level estimate at time *t*,

k is the number of forecasts into the future,

T_t is the trend estimate at time *t*,

S_t is the seasonal estimate at time *t*, and

M is the number of seasons.

Since the time series is assumed to have a seasonal component, the level equation must first remove the seasonality in the data to achieve proper level estimates. The level update equation can be expressed as follows (Ariton, 2021):

$$L_t = \alpha(y_t - S_{t-M}) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

For the removal of the seasonal component, there is a need to understand the type of time series seasonality being modeled, and the end of the inverse. This leads to either dividing or subtracting the seasonal component for multiplicative and additive time series, respectively.

The forecast with Winters's method captures the seasonality component.

6.4 Interesting Discoveries Regarding the Winters Method

Notably, the Winters method is widely popular as compared to other forecasting methods due to its accuracy and simplicity of use. In addition, the accuracy of the method is associated with the use of smoother curves to predict future values. Therefore, it allows one to smooth a time series and use that data to forecast areas of interest.

7. Forecasting Sales of New Products

7.1 Using Logistic Curve to Model S-Curve in Excel

A variable undergoing logistic growth initially grows exponentially, and after some time, the rate of growth decreases and the function levels off, forming a sigmoid curve or an S-shaped curve (Menezes, 2018). All logistic functions take the following form:

$$f(x) = \frac{N}{1 + Ae^{-k(x - x_0)}}$$

Where N, A, e, and k are all constants.

Excel calculates values following logistic growth and can chart them on a line graph. For instance, =A1/[1+B1exp(C1D1)] calculates values following logistic growth in Excel, where A1 is the cell containing the constant N, B1 is the cell containing the constant A, =exp() is a function in Excel that generates the exponential of a value, C1 is the cell containing constant K, and D1 is the cell containing the lowest value of x, which serves as the graph's origin.

Assuming cell D2 contains the increments of the graph, for example, growth of products after every 30 days, the formula D2 = D1 + 30 is used, and then it is extended downward to produce the x-values for the chart. The cell where the formula is inserted is also extended downward to produce y-values for the chart.

From the *Insert* tab in Excel, *Line* is selected, and from *Charts* ribbon, *2-D Line* is selected and Excel plots the function logistic growth chart. A standard logistic growth chart takes the form N = 1, K = , X_0 = 0. The sample of S-curve for a standard logistic growth chart is as illustrated in Figure 2.

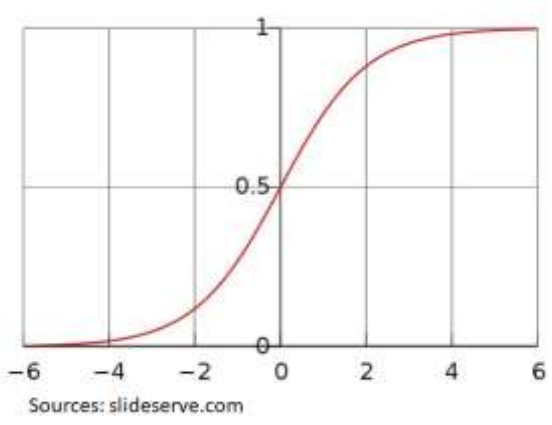

Figure 2. Standard Logistic Curve

7.2 Using Gompertz Curve to Model S-Curve in Excel

A Gompertz curve is a specific kind of mathematical model that can be used to analyze time series data. According to this theory, the rate of growth is lowest at the beginning and the end of any given period. Essentially, the right-side asymptote of the function is approached much more gradually by the curve than the lower-valued asymptote. The Gompertz equation is illustrated as follows:

$$f(t) = ae^{-be^{-ct}}$$

Where,

a is an asymptote,

b sets the displacement along the x-axis (translates the graph to the left or right),

c sets the growth rate (y scaling), and

e is the Euler's number (e = 2.71828).

As *t* grows large, f(*t*) approaches a, and the inflection points of the curve occurs for *t* = Ln b/c, and f(*t*) = a/e.

In Excel, the formula for Gompertz curve is as follows:

=a*EXP(-b*EXP(-c*C5) ….

In this case, C5 is the cell containing the time variable *t*. Drag down the values to have Gompertz forecast. A line chart can then be generated, showing sales of new products and respective Gompertz forecast across the time frame. The sample of Gompertz curve is as illustrated in Figure 3.

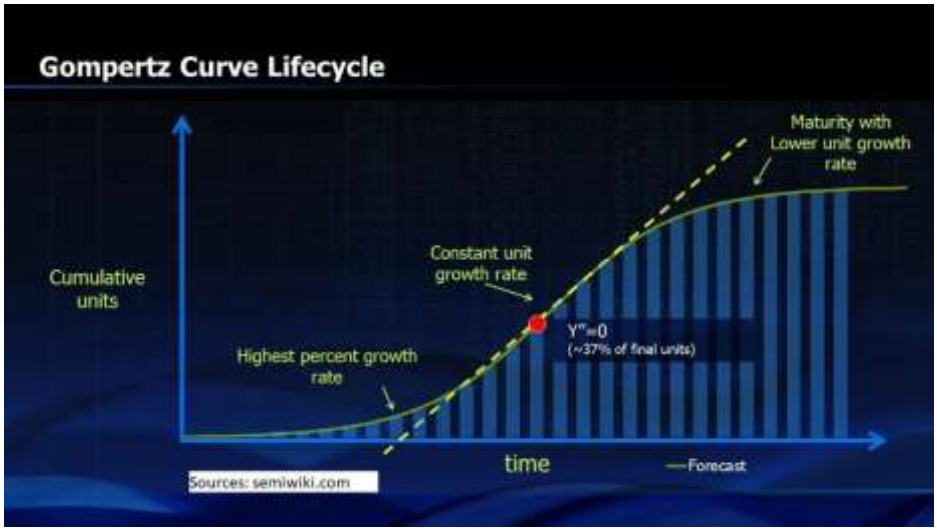

Figure 3. Gompertz Curve

7.3 Bass Diffusion Model in Excel

The Bass diffusion model is a tool for forecasting the adoption of new products and new product categories. It implements the original as well as the generalized Bass model. Unlike the Gompertz curves, the Bass model has been successfully used to forecast product sales before the product comes to the market. It also provides insights on how knowledge of new products spreads through the market.

The Bass diffusion model asserts that the diffusion of new products is determined by innovators and imitators. Innovators seek new products without caring if the product has been adopted by other people. Conversely, imitators wait to try a product until other people have successfully used it. In essence, the Bass model allows determining the relative importance of innovators and imitators in driving the spread of the product. The notations of the Bass model include (S.Y.A., 2014)

n(t) = Product sales during period *t*.

N(t) = Cumulative product sales through period *t*.

$\bar{N}$= Total number of customers in the market, supposing that all of them eventually turn to the product.

P = Coefficient of innovation or external influence.

Q = Coefficient of imitation or internal influence.

Therefore, the equation of the Bass model is as follows:

$$\text{n(t)} = P(\bar{N} - N(t-1) + \frac{(\bar{N}-N(t-1))(N(t-1)}{\bar{N}}$$

$[\bar{N} - N(t-1)]$ is an element linked to the number of people who have not used the product. This element is independent of the number of people who have already used the product [N(t-1)]. A component that is tied to the number of interactions between previous adopters [N(t-1)] and people who are yet to adopt $[\bar{N} - N(t-1)]$ represents the diffusion of the product through the market. In this case, the imitation or internal influence component shows that the previous adopters tell non-adopters about the product, hence leading to new adoptions (S.Y.A, 2014).

To fit the Bass model, the values of P, Q, and $\bar{N}$ must be found using the Solver tool in Excel. The Solver is used to determine the values of the Bass model parameters that best fit the data in the sense of minimizing the SSE. For instance, the trial values of P, Q, and $\bar{N}$ are put in cells E2:G2. Using the trial values, prediction for sales is calculated using the following formula:

$$=p*(Nbar-C5) +(q/Nbar)*C5*(Nbar-C5),$$

Where,

C5 contains the cumulative sales,

Squared errors = [n(t) – Predicted]^2,

SSE = SUM (Squared errors).

Then, the Solver is used to get estimates of P, Q, and $\bar{N}$. If the value of Q is greater than the value of P, it shows that imitation is much more important than innovation in facilitating product diffusion.

7.4 Interesting Discoveries Regarding New Product Sales Forecasting

The three models discussed are instrumental in forecasting new products in the market, as they use different parameters and techniques in fitting an S-shaped curve to the data. It is worth noting from the logistic curve that a new product in the market initially grows exponentially up to a certain time. After attaining an optimal growth, the rate of growth decreases and mostly forms an S-shaped curve. However, what is more interesting is the way Bass diffusion model carries out forecasting. By taking into consideration both external and internal influences as indicated by the Bass model, the information derived from the forecast is conclusive and has a basis. It is also quite logical to assert that the introduction of new products in the market is influenced by innovators and imitators.

8. Application of the Knowledge in Doctoral Research Projects

The knowledge of market analysis forecasting using statistics has a wide range of significant applications for researchers undertaking doctoral projects. The knowledge can be applied in comparisons for which suitable techniques are used in different types of data. Contrasting and comparing different methods by taking into account their advantages, disadvantages, and rationales behind them can aid in deducing an appropriate method to solve a problem given a specific objective or a question of a doctoral project research. Furthermore, a researcher can select one research area and apply one of the forecasting techniques to answer the hypothesis of the doctoral project.

9. Practical Applications of Market Analysis Forecasting

9.1 Application of Regression Analysis

Linear regression model is a powerful tool that can be used by several businesses, finance companies, and marketing departments to help managers understand the relationship among factors such as customer expenditures and commodity stocks. It is used in marketing to evaluate trends and work out estimates or forecasts.

For example, if in the previous years, the sales of sports products in a business have increased steadily every month, using linear regression model on the sports goods' sales data with monthly sales, the business can forecast sales for future months. Also, in a marketing campaign where a company advertises its products on different platforms such as TV or radio, linear regression model can be used to ascertain the effectiveness of each advertisement platform as well as the combined effectiveness of the platforms.

9.2 Application of Trends and Seasonality

Seasonality can be used to analyze stocks and economic trends. Moreover, organizations or companies can use seasonality to establish and make certain business decisions such as inventories and staffing. For instance, retail sales exhibit seasonality and normally experience higher expenditures during the fourth quarter of the calendar year. Therefore, seasonality can be applied in sales companies to determine the busier seasons and how they may influence the sales of products.

9.3 Application of Special Event Forecasting

Companies can use different forecasting techniques to forecast special events such as holidays or sales campaigns. For example, a company can use their customers' expenditure data patterns during previous special events to predict their expenditure patterns for the future special events, thereby properly preparing and planning for them.

9.4 Application of the Holt-Winters Method of Forecasting

Assuming a small business has information about the daily number of customers of the last few months, the Winters method can be applied to predict the number of customers for upcoming days of the month. Using the Winters model, the business will be able to identify the trend of the number of customers and the seasonality aspect. For example, the Winters smoothing method can be used to forecast future sales of electronics products in an electronics manufacturing firm.

9.5 Application of New Product Sales Forecasting

A manufacturing company can use forecasting techniques such as Bass diffusion model to forecast revenues of a new computer system based on previous data. Additionally, a company can employ the use of logistic curves to

predict future sales of different brands of phones based on the data from previous sales.

10. Conclusion

The present article is anchored on understanding the key concepts of market forecasting. It highlights the importance of the concepts studied such as regression model, trends and seasonality, effects of special events, Holt–Winters method, and forecasting techniques for new products. Regarding regression analysis, it is revealed that the use of regression model is critical to any business as it provides the platform for assessing the relationship among different variables and determining the effects of one variable on the other. Regression model also provides accurate estimates and model significance, making it highly reliable. In addition, it can predict future values of a variable (e.g., sales) from its equation, and it allows for comparison of models such that the most appropriate model in different scenarios is used. Concerning trends and seasonality, it is emphasized that every business must model seasonality and trend in its time series data to have an accurate estimate and positively impact decision making and planning for future sales. Many industries such as retail industry experience changing trends and seasonal consumer habits; therefore, understanding trends and seasonality can be instrumental in better forecast of products, stocks, and prices, among others.

Special events such as holidays and campaign periods always occur in a market cycle. Many businesses can optimize their sales before, during, or after these events. Forecasting the impacts of special events can help understand the organization's sales pattern, thereby leading to better planning prior to the events. In essence, forecasting for sales of new products can be done using different techniques, and it is upon each organization or business to choose wisely the technique. For the data exhibiting seasonality and trends, there are specific techniques that are highly suitable.

Adopting the correct forecasting procedure will give any organization a competitive edge over other market players and competitors. Therefore, this article provides a theoretical background on the concepts of market analysis forecasting, which can be used as a basis by players in the market analysis sector when considering measuring criteria to evaluate their businesses and performances.

References

Ariton, L. (2021). *A thorough introduction to holt-winters forecasting.* Medium. Retrieved from Retrhttps://medium.com/analytics-vidhya/a-thorough-introduction-to-holt-winters-forecasting-c21810b8c0e6

Beers, B. (2022). *What is regression? Definition, calculation, and example.* Investopedia. Retrieved from https://www.investopedia.com/terms/r/regression.asp

Bhandari, P. (2021). *Missing Data | Types, Explanation, & Imputation.* Scribbr. Retrieved from https://www.scribbr.com/statistics/missing-data/

Frost, J. (2017). *How to interpret the F-test of overall significance in regression analysis.* Statistics by Jim. Retrieved from https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/

Garcia, C. (2017). *Use these 4 methods to deal with missing data.* Humans of Data. Retrieved from https://humansofdata.atlan.com/2017/09/4-methods-missing-data/

Hayes, A. (2022). *Multiple linear regression (MLR) definition, formula, and example.* Investopedia. Retrieved from https://www.investopedia.com/terms/m/mlr.asp

Indeed Editorial Team. (2023). *How to calculate moving average in Excel (with example).* Indeed. Retrieved from https://www.indeed.com/career-advice/career-development/how-to-calculate-moving-average-in-excel

Indeed Editorial Team (2022). *Outliers in Excel: Definition, tips and how to calculate them.* Indeed. Retrieved from https://www.indeed.com/career-advice/career-development/how-to-calculate-outliers-in-excel

Info Entrepreneurs. (2021). *Importance of knowledge to a growing business*. Retrieved from https://www.infoentrepreneurs.org/en/guides/importance-of-knowledge-to-a-growing-business/

Kenton, W. (2020). *Seasonality: What it means in business and economics, examples.* Investopedia. Retrieved from https://www.investopedia.com/terms/s/seasonality.asp

Menezes, R. (2018). *How to plot logistic growth in Excel.* CHRON. Retrieved from https://smallbusiness.chron.com/plot-logistic-growth-excel-39372.html

Rawat, A. S. (2021). *An overview of descriptive analysis.* Analytic Steps. Retrieved from https://www.analyticssteps.com/blogs/overview-descriptive-analysis

Rekleits, I. (2015). *Exploration tutorial.* ACM SIGMOD International Conference on Management of Data, 277–281. Retrieved from https://dl.acm.org/doi/proceedings/10.1145/2723372?tocHeading=heading6

Sequitin, K. (2021). *What is an outlier. Career Foundry.* Retrieved from https://careerfoundry.com/en/blog/data-analytics/what-is-an-outlier/

Shah, K. (2021). *Exploratory analysis using univariate, bivariate, and multivariate analysis techniques.* Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/

S.Y.A. (2014). *Marketing analytics: Data-driven techniques with Microsoft Excel (2014).* Retrieved from https://apprize.best/microsoft/excel_4/28.html

Zaiontz, C. (2021). *Holt-Winters additive method. Real Statistics Using Excel.* Retrieved from https://real-statistics.com/time-series-analysis/basic-time-series-forecasting/holt-winters-additive/