Data Analytics Capstone Topic Approval Form

Student Name: Nicole Reiswig

Student ID: 008756001

Capstone Project Name: Multiple Linear Regression Analysis on Superstore Sales Dataset

Project Topic: Predictive Model on Superstore Sales Data

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: Can a multiple linear regression model be constructed based solely on the research data?

Hypothesis: **Null hypothesis**- A predictive regression model cannot be constructed from the superstore sales data.
Alternate Hypothesis- A predictive regression model can be constructed from the superstore sales data.

Context: The contribution of this study to the field of data analytics and the MSDA program is to create a predictive model that can estimate future sales of the superstore so that the company can be adequately prepared to meet future business needs. This study will utilize a multiple linear regression model to analyze the significant predictor variables for future sales. Dong, Chen, Y., Gu, A., Chen, J., Li, L., Chen, Q., Li, S., & Xun, Q., (2020) found that multiple linear regression analysis was successful in identifying positively correlated relationships in sales data specifically with the time and review variables. In previous studies utilizing multiple linear regression for sales revenue, there has been a positive correlation between sales revenue and review ratings (Dong et al., 2020). Analyzing the superstore sales data through multiple linear regression to answer the research question is crucial because it can be determined which factors have the greatest influence on sales revenue. In knowing which factors have the greatest impact both positive and negative smart business decisions can be made to maximize sales revenue in the upcoming years.

Data: The data collected is owned by Bhanupratap Biswas, the author of the data is Bhanupratap Biswas a Kaggle expert. The dataset was collected from Kaggle.com and can be located at this web address: https://www.kaggle.com/datasets/bhanupratapbiswas/superstore-sales. This data set is updated on an annual basis. The superstore dataset has been licensed as ODC Public Domain Dedication and License which allows us to use this public data for this analysis (Open Knowledge, n.d.).

The superstore sales data consists of 9,800 rows, 18 columns, 4,922 unique Order IDs, 1,230 unique Order Dates, 1,320 unique Ship Dates, and 793 unique Customer IDs. The 18 columns in this data set are Row_ID, Order_ID, Order_Date, Ship_Date, Ship_Mode, Customer_ID, Customer_Name, Segment, Country, City, State, Postal_Code, Region, Product_ID, Category, Sub_Category, Product_Name and Sales.

Field	Data Type	Variable Type
Row_ID	Categorical	Independent
Order_ID	Categorical	Independent
Order_Date	Categorical	Independent
Ship_Date	Categorical	Independent
Ship_Mode	Categorical	Independent
Customer_ID	Categorical	Independent
Customer_Name	Categorical	Independent
Segment	Categorical	Independent
Country	Categorical	Independent
City	Categorical	Independent
State	Categorical	Independent
Postal_Code	Numeric	Independent
Region	Categorical	Independent
Product_ID	Categorical	Independent
Category	Categorical	Independent
Sub_Category	Categorical	Independent
Product_Name	Categorical	Independent

Sales	Numeric	Dependent

The study's limitations are that the dataset does not include certain factors that could have been valuable to the study such as review ratings. This data is not included because it was not collected or required in data collection. According to Chattopadhyay (2016) multiple linear regression analysis increases customer satisfaction and loyalty, predicts the number of purchases made, discovers the relationship between customer wait times and the number of complaints, and is used to estimate customers' needs. The delimitations of the study are that Order_ID, Customer_ID, Customer_Name, Country, and Product_ID will be removed from the dataset. These fields will be removed because they do not have statistically significant value in this analysis. Multiple Regression (n.d.) recommends that when using multiple linear regression the analysis should be limited to five or fewer independent variables to prevent multicollinearity.

Data Gathering: Data Treatment: The data will be downloaded from the publicly available CSV file on Kaggle.com which shows the superstore data set with data only in the United States. The data will be imported into Jupyter Lab utilizing Python and Anaconda. The data will be checked for missing, null, or duplicate values. Any inputs with missing values will be removed. The data quality is high as it has no missing, null, or duplicate values. The data contains both qualitative and quantitative values. Before manipulating the data it's important to first have visualizations to examine outliers, distribution shape and spread, relationships between variables, and correlation (Walker, 2020). Pandas has many tools that will be used to clean the data including calculating summary statistics, changing series values conditionally, evaluating and cleaning strings, working with dates, and missing value imputation (Walker, 2020). All categorical variables will be converted using encoding to binary variables. Overall data sparsity is 0%.

Data Analytics Tools and Techniques: Univariate and bivariate visualizations are created to view the distribution and spread of the data and to check for patterns. A Q-Q plot and Shapiro-Wilk were run to determine normality. Although it is not required for multiple regression analysis it will provide visualization of the data. The data analysis technique to be employed is Multiple Linear Regression Analysis. Multiple linear regression analysis is an appropriate technique because the data consists of one continuous dependent variable and several independent variables (Mishra, Pandey, Singh, Keshri, 2019). The analysis will be performed using Python in the Jupyter Lab environment. The goals and expectations of the study are to predict future sales revenue utilizing the provided superstore sales dataset. The multiple linear regression analysis will use historic sales data to find patterns between the independent and dependent variables to predict future sales. Predicting future sales revenue and the variables that have the most statistically significant impact on the sales revenue provides actionable insights for smart business decisions. The presentation layer includes univariate and bivariate visualizations and a Tableau dashboard explaining the findings.

> **Justification of Tools/Techniques:** The data analysis tools/techniques are appropriate for this analysis. Python is an open-source programming language that in terms of usability is the better choice because the syntax it uses is similar to other languages and therefore is more versatile (Ozgur, Colliau, Rogers, Hughes, & Myer-Tyson, 2017). Python is a famous data analysis language used by data scientists and is the preferred language because of its highly interactive nature and its scientific ecosystem libraries (Siddiqui, Alkadri, & Khan, 2017). According to Siddiqui (2017), SAS is an expensive solution and is between a programming language and a scripting language. R is an open-source programming language that is more difficult to learn but is great with data manipulation, statistics, and graphics functionality built-in (Siddiqui, 2017). Ozgur (2017) notes that R is geared more towards statistical roles, resembles SAS, does not rely on a computer science background or coding and is open-sourced, while Python is geared more towards the coding aspects of jobs. Utilizing Python provides a few benefits one being ease of implementation through libraries such as pandas and scikit learn. Pandas can handle data manipulation and scalability. Scikit Learn is a machine learning library that can perform multiple linear regression analyses. Multiple linear regression can consider multiple factors and produce an accurate prediction.

Project Outcomes: The anticipated outcome of this analysis is a reusable statistical model that produces actionable insights. Support for the alternative hypothesis that a predictive multiple linear regression model can be constructed utilizing the superstore sales data can be found in (Anseur, 2024).

Projected Project End Date: 05/11/2024

Sources:

Anseur, A. (2024). *Superstore Sales EDA + ML.* Kaggle.com. Retrieved April 16, 2024, from Superstore Sales | EDA + ML (kaggle.com)

Chattopadhyay, R. (2016). *Effective Business Solutions With Big Data Analytics: Key for Business Growth.* Globsyn Management Journal, 10(1/2), 87–96. Retrieved April 16, 2024, from https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=132005022&site=eds-live&scope=site&authtype=shib&custid=ns017578

Darlington, R. & Hayes, A. (2017). *Regression Analysis and Linear Models : Concepts, Applications, and Implementation.* The Guilford Press. Retrieved April 16, 2024, from https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1340168&site=eds-live&scope=site&authtype=shib&custid=ns017578

Dong, J., Chen, Y., Gu, A., Chen, J., Li, L., Chen, Q., Li, S., & Xun, Q. (2020). *Potential Trend for Online Shopping Data Based on the Linear Regression and Sentiment Analysis.* Mathematical Problems in Engineering, 1–11. Retrieved April 16, 2024, from https://doi.org/10.1155/2020/4591260

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). *Selection of appropriate statistical methods for data analysis.* Annals of Cardiac Anaesthesia, 22(3), 297–301. Retrieved April 16, 2024, from https://doi.org/10.4103/aca.ACA_248_18

Multiple Regression. (n.d.). *Multiple Regression.* Csulb.edu. Retrieved April 16, 2024, from Multiple Regression (csulb.edu)

Open Knowledge. (n.d.). *Open Data Commons Public Domain.* Opendatacommons.org. Open Data Commons Public Domain Dedication and License (PDDL) v1.0 — Open Data Commons: legal tools for open data

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, E. "Bennie." (2017). *MatLab vs. Python vs. R.* Journal of Data Science, 15(3), 355–371. Retrieved April 16, 2024, from https://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=125220011&site=eds-live&scope=site&authtype=shib&custid=ns017578

Siddiqui, T., Alkadri, M., & Khan, N. A. (2017). *Review of Programming Languages and Tools for Big Data Analytics.* International Journal of Advanced Research in Computer Science, 8(5), 1112–1118. Retrieved April 16, 2024, from https://search.ebscohost.com/login.aspx?direct=true&db=asf&AN=124636531&site=eds-live&scope=site&authtype=shib&custid=ns017578

Walker, M. (2020). *Python Data Cleaning Cookbook : Modern Techniques and Python Tools to Detect and Remove Dirty Data and Extract Key Insights.* Packt Publishing. Retrieved April 16, 2024, from https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2706043&site=eds-live&scope=site&authtype=shib&custid=ns017578

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 4/23/2024

Reviewed by: 

Comments: Click here to enter text.