Data Science Capstone Topic Approval Form

Student Name: Shanay Murdock

Student ID: 011377935

Capstone Project Name: Multiple Linear Regression on Global Mobile Reviews Dataset

Project Topic: Predictive Model for Global Mobile Reviews 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 on the research dataset?

Hypothesis: H_0: A predictive multiple linear regression model cannot be constructed from the research dataset.

H_1: A multiple linear regression model can be constructed to predict overall customer rating at a model accuracy of > 70%.

Context:. The contribution of this study to the field of Data Analytics and the MSDA-DS program is to create a predictive model which can use product pricing and features to estimate customer survey ratings so device manufacturers or product teams can prioritize their development roadmaps. Finding what features drive the highest customer satisfaction allows companies to improve that technology to yield the highest ROI for customer satisfaction in the next model. This study will use a multiple linear regression model to analyze the significance of predictor variables and identify which variables have the strongest predictive performance on overall customer satisfaction ratings. Frantsen (2025) found that utilizing multiple linear regression is effective on survey data as a predictive method. The research found that multiple linear regression "helps to identify trends, patterns, and correlations within data for better forecasting" for purposes including customer satisfaction, quantifying impact, and estimating market demand (Frantsen, 2025). A study by J.D. Power (2025) indicates that performance, overall ease of operation, battery life, and physical design have repeatedly shown as important components of overall customer satisfaction. Frantsen further explains that understanding the relationship between a target variable and response variables allows for prediction of the response variable based on given predictor inputs (Frantsen, 2025).

Data: The data needed to be collected for the question is the publicly available information provided by Kaggle for global mobile phone review data (Kaggle, 2025). The dataset contains 50,000 rows.

The data set is made available through Kaggle. The data set includes the following variables of `brand`, `price_usd`, `battery_life_rating`, `camera_rating`, `performance_rating`, `design_rating`, `display_rating`, and the target variable `rating`. The predictor variables are broken down as follows:

https://www.kaggle.com/datasets/mohankrishnathalla/mobile-reviews-sentiment-and-specification/data

Field	Type
brand	Categorical, nominal
price_usd	Continuous
battery_life_rating	Categorical, ordinal
camera_rating	Categorical, ordinal
performance_rating	Categorical, ordinal
design_rating	Categorical, ordinal
display_rating	Categorical, ordinal

This dataset was made publicly available on Kaggle as a curated collection of 50,000 mobile phone reviews gathered through web scraping, market analysis, and content aggregation from multiple e-commerce and tech

review platforms, covering eight countries. While there is a `customer_name` field and the data was collected from publicly available review information on e-commerce platforms, the names were anonymized for research and educational use (Kaggle, 2025). Limitations: The data is restricted to the "mobile" product category, which may limit the generalizability of findings to other retail sectors. The dataset is limited to three years of data, spanning from '2022-10-22' to '2025-10-21'. The data presents historical reviews and may not reflect the current marketplace or recent model releases. The dataset also represents eight countries with different market dynamics, potentially introducing cultural bias in rating behaviors. The rating scale uses a 1-5 scale (1 being worst, 5 being best), and individual interpretation may vary across respondents. The dataset is pre-collected, so there is no controlling for widely varying data collection methods across different source platforms (e.g., Amazon vs. Flipkart may have different user bases). Delimitations: The scope is restricted from the larger dataset that includes 25 columns down to 1 target variable and 7 variables that include brand, pricing, and 5 feature-specific ratings. The model excludes the demographic data available (including country of reviewer) and relies solely on brand, pricing (USD), and feature-specific ratings using the Likert scale. There is no missing data from any row, so the full 50,000 rows are available to the model, meeting the recommendation for a sufficient number of rows for a linear regression model (Austin & Steyerberg, 2015).

Data Gathering: The Treatment of the Data: The data will be downloaded via a publicly accessible CSV from Kaggle in a dataset called "Global Mobile Reviews Dataset (2025 Edition) which shows 50,000 customer reviews of mobile phone purchases made across 8 countries. No entries have any missing data, so no data will be removed or modified for exclusion or imputation. The data quality is very high and has already been cleaned and translated so that all data is in English. The data contains both quantitative and qualitative variables. Python will be used to explicitly assign ordinal order to all rating variables and to create dummy variables for the phone brands (nominal data) as machine learning models require numeric data to process (Walker, 2024). There are no missing values in the dataset, meaning the overall data sparsity is 0%.

Data Analytics Tools and Techniques: The Design of the Study: 1. A Q-Q Plot and Shapiro-Wilk will be run to determine the normality of the data. 2. A multiple linear regression analysis will utilize stepwise analysis to identify the contribution of each independent variable in predicting the dependent variable and retain only those that are statistically significant. During the process of model fitting, the dataset will be split to allow for a training set and a test set, so the fitted model can be tested for accuracy on unseen data. The process of fitting a multiple linear model should create a model which can accurately predict the value of the dependent variable given the independent variable inputs (Géron, 2019). Univariate and bivariate graphs and correlation matrices will be in the presentation layer. Univariate graphs will check the distribution of each variable being tested, bivariate graphs will test each independent variable against the dependent variable, and a correlation heat map will be used to check the strength of the correlation of each variable against all other variables.

Justification of Tools/Techniques: Python will be used for all stages of the analysis including data loading, data preparation, visualization, and statistical analysis. McKinney (2020, p. 2) states that "[i]n the last 10 years, Python has gone from a bleeding-edge or "at your own risk" scientific computing language to one of the most important languages for data science, machine learning, and general software development in academia and industry. He goes on to further say that in many organizations, it's typical to prototype and research in a specialized language like R, but then port the work over into a language built for larger production systems; Python is a suitable language for research and prototyping as well as building (McKinney, 2020, p. 3). While SAS still remains a powerful tool, Python's graphics capabilities for data visualization has surpassed SAS and allows for more customization and flexibility (GeeksforGeeks, 2025b).

Project Outcomes: The project seeks to create a multiple linear regression model for the overall customer satisfaction rating of mobile phone purchasers based on the brand, price, and feature-specific ratings. Support for the alternative hypothesis is found in (GeeksforGeeks, 2025a) that a multiple linear regression model can be helpful in estimating the target variable by determining the strength of predictor variables with respect to mobile phone purchase data.

Projected Project End Date: 11/30/2025

Sources:

Austin, P., & ; Steyerberg, E. (2015, January 22). The number of subjects per variable required in linear regression analyses. Retrieved November 11, 2025, from https://www.sciencedirect.com/science/article/pii/S0895435615000141

Frantsen, A. (2025, May 28). *Using Linear Regression to Forecast Survey Results*. SurveyKing. Retrieved November 10, 2025, from https://www.surveyking.com/blog/linear-regression-survey-data/

GeeksforGeeks (2025a). *Linear Regression in Machine Learning*. Retrieved November 11, 2025, from Linear Regression in Machine learning

GeeksforGeeks (2025b). *SAS vs R vs Python*. Retrieved November 10, 5, from https://www.geeksforgeeks.org/blogs/sas-vs-r-vs-python/

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media, Inc.

J.D. Power (2017, March 23). *2017 Full Service Smartphone Satisfaction Study*. Retrieved November 11, 2025, from https://www.jdpower.com/business/press-releases/2017-full-service-smartphone-satisfaction-study

Kaggle. (2025, October 22). Retrieved November 1, 2025, from https://www.kaggle.com/datasets/mohankrishnathalla/mobile-reviews-sentiment-and-specification/data

McKinney, W. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed., p. 212). O'Reilly Media, Inc.

Walker, M. (2024). *Python Data Cleaning Cookbook* (2nd ed.). Packt Publishing.

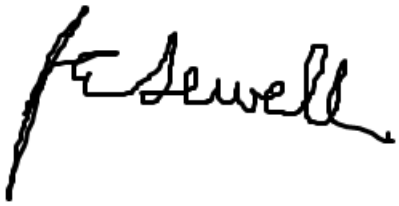
Instructor Signature/Date:

X The research is exempt from an IRB Review.

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

Instructor's Approval Status: **Approved**

Date: **November 12, 2025**



Reviewed by:

Comments: Click here to enter text.