Data Analytics Capstone Topic Approval Form

Student Name: Sean Simmons

Student ID: 009752842

Capstone Project Name: Multiple Linear Regression Analysis on Video Game Sales

Project Topic:

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

Research Question: Can we use video game sales information to create a predictive multiple linear regression model to predict sales?

> **Hypothesis**: **Null hypothesis**-. We can not create a model to statistically significantly predict video game sales.
> **Alternate Hypothesis**- We can create a model to statistically significantly predict video game sales.

Context: The reason we are performing this analysis is to create a model that can be used to predict video game sales for a perceived business problem of predicting sales based on past trends. By doing this, we can inform the business of what videogame to create, what genre to market it as, what system to release it on, and other information to predict sales. At the end of this analysis, we hope to be able to predict video game sales based on the information about the video game and inform the business of their projected sales or the characteristics of games that would produce high sales.

Data: There is an existing dataset compiled on Kaggle with open source rights. The dataset describes Video game sale information with sales, name, system, genre, year, publisher, and sales by region information. Please see the following link or the link in the sources section of this document: Video Game Sales | Kaggle**.**

Data Gathering: I will introduce an existing dataset into my coding environment (jupyter labs v3.44, python v3).

Data Analytics Tools and Techniques: I will use python to create a predictive multiple linear regression model to see if we can predict video game sales based on the variables in the data set (genre, system, year, etc.).

> **Justification of Tools/Techniques:** Multiple Linear Regression is the tool to use here because it can identify a linear relationship between our continuous target variable (sales) and several other independent variables the other variables in the dataset). Our dataset consists of a target continuous variable and several categorical variables that will need to be encoded as numeric variables. Once encoded, these other variables can be used to create a model to predict sales, our key metric (Yadav 2021).

Project Outcomes: The key outcome is a regression equation that can be used to predict sales. More specifically, I also expect to retrieve a condition number indicating multicollinearity, p values of the variables in the data set, R value for prediction strength of the model, and a standard residual error value that will show us the overall strength of our predictive multiple linear regression model for sales. At the end I expect to be able to recommend a course of action to a business based on the model.

Projected Project End Date: 03/15/2023

Sources:

Smith, G. (2016, October 26). *Video game sales*. Kaggle. Retrieved March 6, 2023, from https://www.kaggle.com/datasets/gregorut/videogamesales

Yadav, H. (2021, May 8). *Multiple Linear Regression Implementation in Python*. Machine Learning with Python. https://medium.com/machine-learning-with-python/multiple-linear-regression-implementation-in-python-2de9b303fc0c

Course Instructor Signature/Date:

X☐ 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: 3/6/23

Reviewed by: Daniel J. Smith, PhD, MBA

Comments: N/A