**Data Analytics Capstone Topic Approval Form**

**Student Name:** Kelseyann Wright

**Student ID:** 00565814

**Capstone Project Name:** Predicting Jeopardy Question Dollar Values with Logistic Regression

**Project Topic**: Natural Language Processing of Jeopardy Questions

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

**Research Question:** Can jeopardy question and answer keywords be used to predict the dollar value of jeopardy questions?

**Hypothesis**:   
**Null hypothesis**- Jeopardy question and answer keywords cannot be used to predict the dollar value of jeopardy questions with greater than 75% test accuracy.   
**Alternate Hypothesis**-. Jeopardy question and answer keywords can be used to predict the dollar value of jeopardy questions with greater than 75% test accuracy.

**Context:** The key to winning Jeopardy is being able to answer the high value questions. Contestants preparing for the game would benefit from being able to identify the dollar value of potential questions.

**Data:**The data needed to address the research question should include Jeopardy questions, answers, and their dollar value. An existing dataset exists on Kaggle contains data for over 200,000 Jeopardy questions that can be used for this analysis. The data is public domain and was obtained originally by the owner scraping https://www.j-archive.com/, a fan-created and run website. The dataset on Kaggle is owned by Aravind Ram Nathanm, although there are many datasets on Kaggle that use the same source data and are ‘owned’ by others. Because the data is publicly available and does not violate any privacy laws, I am able to use it for this analysis.

**Data Gathering:** The data is available on Kaggle as a JSON file. The file will be downloaded to my personal computer for use in this analysis. The JSON file will be loaded and converted into a Pandas dataframe in Python using Visual Code Studio.

**Data Analytics Tools and Techniques**: A Logistic Regression will be used for text classification of the dataset. Natural Language Processing techniques, such as removal of stop words and tokenization, will be used to pre-process the data. Analyses will be completed in Python.

**Justification of Tools/Techniques:** Logistic Regression was chosen for this analysis due to its simplicity and detailed model summary. Logistic Regression can be generalized to a dependent variable with multiple classes and has the added benefit of providing coefficients for features, allowing more important features to be identified. Logistic regression has been shown to be successful for multi-class text classification (Li, 2018).

**Project Outcomes**: The project outcome will be a trained and tested Logistic Regression model that predicts jeopardy question dollar values based on keywords in the question and answer. In addition, the most important keywords (based on coefficient values) will be identified.

**Projected Project End Date**: 2/25/2024

**Sources**:

*Data source:* <https://www.kaggle.com/datasets/aravindram11/jeopardy-dataset-updated>, *scraped from* <https://www.j-archive.com/>

*Other sources:*

Li, S. 2018. Multi-Class Text Classification Model Comparison and Selection. Towards Data Science. <https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568>

**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: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.