COMPETENCIES

4030.5.2 : **Multiple Regression**
The graduate employs multiple regression algorithms with categorical and numerical predictors in describing phenomena.

4030.5.3 : **Regression Implications**
The graduate makes assertions based on regression modeling.

INTRODUCTION

As a data analyst, you will assess data sources for their relevance to specific research questions throughout your career. In your previous coursework, you have performed data cleaning and exploratory data analysis on your data. You have seen basic trends and patterns and can now start building more sophisticated statistical models. In this course, you will use regression models. You will explore both linear regression and logistic regression models and their assumptions.

For this task, you will select **one** of the provided data files from the "Data Sets and Associated Data Dictionaries" found in the Web Links section.

You will then review the data dictionary related to the raw data file you have chosen and prepare the data set file for linear regression modeling. The organizations connected with the given data sets for this task seek to analyze their operations and have collected variables of possible use to support the decision-making processes. You will analyze your chosen data set using linear regression modeling, create visualizations, and deliver the results of your analysis.

REQUIREMENTS

Your submission must be your original work. No more than a combined total of 30% of the submission and no more than a 10% match to any one individual source can be directly quoted or closely paraphrased from sources, even if cited correctly. The similarity report that is provided when you submit your task can be used as a guide.

You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.

*Tasks may **not** be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc., unless specified in the task requirements. All other submissions must be file types that are uploaded and submitted as attachments (e.g., .docx, .pdf, .ppt).*

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:
   1. Summarize **one** research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple linear regression in the initial model.

2. Define the goals of the data analysis.

   Note: Ensure that your goals are within the scope of your research question and are represented in the available data.

Part II: Method Justification

B. Describe multiple linear regression methods by doing the following:
   1. Summarize **four** assumptions of a multiple linear regression model.
   2. Describe **two** benefits of using Python or R in support of various phases of the analysis.
   3. Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I.

Part III: Data Preparation

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:
   1. Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code.
   2. Describe the dependent variable and *all* independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.
   3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.
   4. Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.
   5. Provide the prepared data set as a CSV file.

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced linear regression model by doing the following:
   1. Construct an initial multiple linear regression model from *all* independent variables that were identified in part C2.
   2. Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.
   3. Provide a reduced linear regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

E. Analyze the data set using your reduced linear regression model by doing the following:
   1. Explain your data analysis process by comparing the initial multiple linear regression model and reduced linear regression model, including the following element:
      - a model evaluation metric
   2. Provide the output and *all* calculations of the analysis you performed, including the following elements for your reduced linear regression model:
      - a residual plot
      - the model's residual standard error
   3. Provide an executable error-free copy of the code used to support the implementation of the linear regression models using a Python or R file.

Part V: Data Summary and Implications

F. Summarize your findings and assumptions by doing the following:
   1. Discuss the results of your data analysis, including the following elements:
      - a regression equation for the reduced model
      - an interpretation of the coefficients of the reduced model

- the statistical and practical significance of the reduced model
- the limitations of the data analysis

2. Recommend a course of action based on your results.

Part VI: Demonstration

G. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements:

- an identification of the version of the programming environment
- a comparison of the initial multiple linear regression model you used and the reduced linear regression model you used in your analysis
- an interpretation of the coefficients of the reduced model

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the "WGU" option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto's website.

To submit your recording, upload it to the Panopto drop box titled "Regression Modeling – NBM3 | D208." Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

H. List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

J. Demonstrate professional communication in the content and presentation of your submission.