UGN1 — UGN1 TASK 1: LINEAR REGRESSION ANALYSIS

STATISTICAL DATA MINING — D600
PRFA — UGN1

COMPETENCIES

4160.1.1 : Performs Regression Analysis
The learner performs linear and logistic regressions to make recommendations based on the results.

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 task, you will build, test, and use a linear regression model to support the decision-making process.

Prepare the provided cleaned dataset file for linear regression modeling. The organizations connected with the given dataset 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 dataset using linear regression modeling, create visualizations, and deliver the results of your analysis.

You will complete this performance assessment in the provided WGU virtual lab environment.

Note: The IDE for this assessment is either Anaconda or R Studio, depending on which language you decide to use to complete the task.

REQUIREMENTS

Your submission must represent your original work and understanding of the course material. Most performance assessment submissions are automatically scanned through the WGU similarity checker. Students are strongly encouraged to wait for the similarity report to generate after uploading their work and then review it to ensure Academic Authenticity guidelines are met before submitting the file for evaluation. See Understanding Similarity Reports for more information.

Grammarly Note:
Professional Communication will be automatically assessed through Grammarly for Education in most performance assessments before a student submits work for evaluation. Students are strongly encouraged to review the Grammarly for Education feedback prior to submitting work for evaluation, as the overall submission will not pass without this aspect passing. See Use Grammarly for Education Effectively for more information.

Microsoft Files Note:
Write your paper in Microsoft Word (.doc or .docx) unless another Microsoft product, or pdf, is specified in the task directions. Tasks may not be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc. All supporting documentation, such as screenshots and proof of experience, should be collected in a pdf file and submitted separately from the main file. For more information, please see Computer System and Technology Requirements.

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.

A. Create your subgroup and project in GitLab using the provided web link by doing the following:
- Clone the project to the IDE.
- Commit with a message and push when you complete each requirement listed in parts C2 through D4.

Note: You may commit and push whenever you want to back up your changes, even if a requirement is not yet complete.

- Submit a copy of the GitLab repository URL in the "Comments to Evaluator" section when you submit this assessment.
- Submit a copy of the repository branch history retrieved from your repository, which must include the commit messages and dates.

B. Describe the purpose of this data analysis by doing the following:

1. Propose **one** research question that is relevant to a real-world organizational situation captured in the provided dataset that you will answer using multiple linear regression in the initial model.
2. Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:

1. Identify the dependent and all independent variables that are required to answer the research question and justify your selection of variables.
2. Describe the dependent variable and all independent variables from part C1 using descriptive statistics (counts, means, modes, ranges, min/max), including a screenshot of the descriptive statistics output for each of these variables.
3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables from part C1, including the dependent variable in the bivariate visualizations.

D. Perform the data analysis and report on the results by doing the following:

1. Split the data into two datasets, with a larger percentage assigned to the training dataset and a smaller percentage assigned to the test data set. Provide the files.

Note: The datasets should include only those variables identified in part C1.

2. Use the training dataset to create and perform a regression model using regression as a statistical method. Optimize the regression model using a process of your selection, including but not limited to, forward stepwise selection, backward stepwise elimination, and recursive selection. Provide a screenshot of the summary of the optimized model or the following extracted model parameters:
   - adjusted R2
   - R2
   - F statistics
   - probability F statistics
   - coefficient estimates
   - p-value of each independent variable

3. Give the mean squared error (MSE) of the optimized model used on the training set.

4. Run the prediction on the test dataset using the optimized regression model from part D2 to give the accuracy of the prediction model based on the mean squared error (MSE).

Note: The prediction run on the test dataset must use only the variables identified in the optimized regression model in part D2.

E. Summarize your data analysis by doing the following:

1. List the packages or libraries you have chosen for Python or R and justify how each item on the list supports the analysis.
2. Discuss the method used to optimize the model and justification for the approach.
3. Discuss the verification of assumptions used to create the optimized model.
4. Provide the regression equation and discuss the coefficient estimates.
5. Discuss the model metrics by addressing each of the following:
   - the R^2 and adjusted R^2 of the training set
   - the comparison of the MSE for the training set to the MSE of the test set
6. Discuss the results and implications of your prediction analysis.
7. Recommend a course of action for the real-world organizational situation from part B1 based on your results and implications discussed in part E6.

F. Provide a Panopto video recording that includes *both* a screen share of the presenter demonstrating the functionality of the code used for the analysis and a summary of the programming environment.

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 "Task 1: Linear Regression Analysis – UGN1 | D600". 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.

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

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

File Restrictions

File name may contain only letters, numbers, spaces, and these symbols: ! - _ . * ' ( )
File size limit: 200 MB
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, csv, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z

RUBRIC

A:GITLAB REPOSITORY

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A GitLab repository is not provided.	The subgroup and project are created in GitLab, but 1 or more of the given actions are not completed, or they are completed incorrectly.	The subgroup and project are created in GitLab correctly, and all of the given actions are completed correctly.

B1:PROPOSAL OF QUESTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not propose 1 question.	The submission proposes 1 question that is not relevant to a real-world organizational situation. Or the proposed question cannot reasonably be answered using linear regression analysis.	The submission proposes 1 question that is relevant to a real-world organizational situation and can be answered using linear regression analysis.

B2:DEFINED GOAL

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not define 1 goal for data analysis.	The submission defines 1 goal for data analysis, but the goal is not reasonable, is not within the scope of the scenario, or is not represented in the available data.	The submission defines 1 reasonable go for data analysis that is within the scope of the scenario and is represented in the available data.

C1:VARIABLE IDENTIFICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not correctly identify any variables.	The submission identifies fewer than three total variables, or the selection of	The submission logically identifies the dependent variable and at least 2 indepen-

	the variables is not logically justified.	dent variables that are required to answer the research question. The submission also logically justifies the selection of these variables.

C2:DESCRIPTIVE STATISTICS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
Screenshots of the descriptive statistics are not provided for the variables identified in part C1.	The submission provides screenshots of the descriptive statistics for some of the variables identified in part C1.	The submission provides screenshots of the descriptive statistics for all of the variables identified in part C1.

C3:VISUALIZATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
Visualizations of the distributions of variables from part C1 are not provided.	Univariate and bivariate visualizations of the distributions of some variables from part C1 are provided, or the dependent variable is not included in some bivariate visualizations.	Univariate and bivariate visualizations of the distributions of all variables from part C1 are provided. The dependent variable is included in all bivariate visualizations.

D1:SPLITTING THE DATA

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide the training and test dataset files.	The submission provides training and test datasets, but the split is not reasonably proportioned.	The submission provides reasonably proportioned training and test datasets.

D2:MODEL OPTIMIZATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not demonstrate model optimization.	The submission provides an incomplete screenshot of the summary of the optimized model or does not separately identify *all* of the identified model parameters.	The submission suitably demonstrates model optimization by providing a screenshot of the summary of the optimized model or separately identifies *all* of the identified model parameters.

D3:MEAN SQUARED ERROR

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The mean squared error is not provided.	The mean squared error of a model is provided prior to model optimization.	The mean squared error of the optimized model used on the training set is accurately provided.

D4:MODEL ACCURACY

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The accuracy of the prediction model is not provided.	The accuracy of the prediction model is provided, but it is not based on the mean squared error.	The accuracy of the prediction model is provided based on the mean squared error.

E1:PACKAGES OR LIBRARIES LIST

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not list the packages or libraries chosen for Python or R.	The submission lists the packages or libraries chosen for Python or R but does not justify how 1 or more items on the list support the analysis.	The submission lists the packages or libraries chosen for Python or R and justifies how each item on the list supports the analysis.

E2:METHOD JUSTIFICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss the method used to optimize the model.	The submission discusses the method used to optimize the model but does not provide justification.	The submission logically discusses the method used to optimize the model and provides justification.

E3:VERIFICATION OF ASSUMPTIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss the verification of assumptions used to create the optimized model.	The submission discusses the verification of assumptions but does not show a clear connection to model optimization.	The submission logically discusses the verification of assumptions used to create the optimized model.

E4:EQUATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The regression equation is not provided.	The regression equation is not fully provided, or the discussion of the coefficient estimates is incorrect.	The regression equation is correctly provided, and the coefficient estimates are logically discussed.

E5:MODEL METRICS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss the model metrics.	The discussion of the model metrics does not address *each* of the identified parts.	The submission cogently discusses the model metrics by addressing *each* of the identified parts.

E6:RESULTS AND IMPLICATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss both the results and implications of the prediction analysis.	The submission discusses both the results and implications of the prediction analysis, but the discussion is inadequate.	The submission adequately discusses both the results and implications of the prediction analysis.

E7:COURSE OF ACTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not recommend a course of action for the real-world organizational situation from part B1.	The submission does not recommend a reasonable course of action for the real-world organizational situation from part B1, or the course of action is not based on the results and implications discussed in part E6, or the course of action recommended is “no response.”	The submission recommends a reasonable course of action for the real-world organizational situation from part B1 based on the results and implications discussed in part E6.

F:PANOPTO RECORDING

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A Panopto video recording is not provided.	The Panopto video recording provided is missing either the screen share of the presenter demonstrating the functionality of the code used or a discussion commenting on the programming environment. Or either the demonstration or the summary is inaccurate.	The Panopto video recording provided includes *both* a screen share of the presenter demonstrating the functionality of the code used and a discussion commenting on the programming environment. *Both* the demonstration and the summary are accurate and complete.

G:SOURCES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include both in-text citations and a reference list for sources that are quoted, paraphrased, or summarized.	The submission includes in-text citations for sources that are quoted, paraphrased, or summarized and a reference list; however, the citations or reference list is incomplete or inaccurate.	The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available or the submission states no sources were used.

H:PROFESSIONAL COMMUNICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
This submission includes pervasive errors in professional communication related to grammar, sentence fluency, contextual spelling, or punctuation, negatively impacting the professional quality and clarity of the writing. Specific errors have	This submission includes substantial errors in professional communication related to grammar, sentence fluency, contextual spelling, or punctuation. Specific errors have been identified by Grammarly	This submission includes satisfactory use of grammar, sentence fluency, contextual spelling, and punctuation, which promote accurate interpretation and understanding.

been identified by Grammarly for Education under the Correctness category.

for Education under the Correctness category.

WEB LINKS

Panopto Access

Sign in using the "WGU" option. If prompted, log in with your WGU student portal credentials, which should forward you to Panopto's website. If you have any problems accessing Panopto, please contact Assessment Services at assessmentservices@wgu.edu. It may take up to two business days to receive your WGU Panopto recording permissions once you have begun the course.

Panopto FAQs

Panopto How-To Videos

WGU GitLab Environment - WGU Community

SUPPORTING DOCUMENTS

D600 Task 1 Dataset 1 Housing Information.csv