Competencies

**4160.1.2** : **Performs Dimensionality Reduction**

The learner performs principal component analysis 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 using principal component analysis (PCA). 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](https://cm.wgu.edu/t5/Frequently-Asked-Questions/Understanding-Similarity-Reports/ta-p/252) 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](https://cm.wgu.edu/t5/Academic-Coaching-Center/Use-Grammarly-for-Education-Effectively/ta-p/52276) 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.](https://cm.wgu.edu/t5/WGU-Student-Policy-Handbook/Computer-System-and-Technology-Requirements/ta-p/78)  

*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 D2 through F4.

*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 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.  Explain the reasons for using PCA by doing the following:

1.  Explain how PCA can be used to prepare the selected dataset for regression analysis. Include expected outcomes.

2.  Summarize **one** assumption of PCA.

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

1.  Identify the continuous dataset variables that you will need to answer the research question proposed in part B1.

*Note: The code for the initial model from Task 1 should be used for Task 3, making sure to exclude any categorical variables.*

2.  Standardize the continuous dataset variables identified in part D1. Include a copy of the cleaned dataset.

3.  Describe the dependent variable and all independent variables from part D1 using descriptive statistics (counts, means, modes, ranges, min/max), including a screenshot of the descriptive statistics output for each of these variables.

E.  Perform PCA by doing the following:

1.  Determine the matrix of *all* the principal components.

2.  Identify the *total* number of principal components (that should be retained), using the elbow rule or the Kaiser rule. Include a screenshot of the scree plot.

3.  Identify the variance of *each* of the principal components identified in part E2.

4.  Summarize the results of your PCA.

F.  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 dataset. Provide the file(s).

*Note: The datasets should include only those principal components identified in part E2.*

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 F2 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.*

G.  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 R2 and adjusted R2 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.