Competencies

**4159.1.1** : **Profiles Data**

The learner interprets a data dictionary to understand the data set.

**4159.1.2** : **Interprets Statistics and Visualization**

The learner interprets probability, descriptive and inferential statistics, and visualization.

**4159.1.3** : **Wrangles Data**

The learner wrangles data to ensure accuracy, format, and integrity relevant to the task being performed.

Introduction

Throughout your career in data analytics, you will analyze data according to business and data analytic needs. You will explore basic statistics, examine correlations among variables using visualization, and perform inferential statistical analysis to provide insights relevant to business requirements.  
  
In this task, you will aggregate and analyze a large health insurance company’s data. Your goal is to uncover patterns, trends, and correlations to offer insights into business performance. You will deliver the results of your analysis to company stakeholders.

Scenario

Refer to the most recent company data provided in the "Health Insurance Dataset" and "Health Insurance Considerations and Dictionary" supporting documents to inform your work.

*Note: The IDE for this assessment is either Anaconda or RStudio, depending on which language you decide to use to complete the task. Please use the “WGU Virtual Lab Environment” web link below.*

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.*

**Part I: Univariate and Bivariate Statistical Analysis and Visualization**

*Note: You may use Python or R for implementing your coding solutions, manipulating the data, and creating visual representations. However, your responses to the following task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will****not****be accepted.*

A.  Identify the distribution of **two** continuous variables and **two** categorical variables using univariate statistics from the dataset.

1.  Represent your findings from part A visually as part of your submission.

B.  Identify the distribution of **two** continuous variables and **two** categorical variables using bivariate statistics from the dataset.

1.  Represent your findings from part B visually as part of your submission.

**Part II: Parametric Statistical Testing**

*Note: You may use Python or R for implementing your coding solutions, manipulating the data, and creating visual representations. However, your responses to the following task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will****not****be accepted.*

C.  Describe a real-world organizational situation or issue in the provided dataset by doing the following:

1.  Provide**one** research question relevant to the dataset and *any* organizational needs that can be answered through data analysis.

2.  Identify the variables in the dataset that are relevant to answering your research question from part C1.

D.  Analyze the dataset by doing the following:

1.  Identify a *parametric* statistical test that is relevant to your question from part C1.

2.  Develop null and alternative hypotheses related to your chosen parametric test from part D1.

3.  Write code (in either Python or R) to run the parametric test.

4.  Provide the output and the results of *any* calculations from the parametric statistical test you performed.

E.  Evaluate parametric test results by doing the following:

1.  Justify why you chose the statistical test identified in part D1 based on variables.

2.  Discuss the test results, including the decision to reject or fail to reject the null hypothesis from part D2.

3.  Explain how stakeholders in the organization benefit from your choice of testing method.

F.  Summarize the implications of your parametric statistical testing by doing the following:

1.  Discuss the answer to your question from part C1.

2.  Discuss the limitations of your data analysis.

3.  Recommend a course of action based on your findings.

4.  Submit your code for each test.

*Note:****One****notebook can be submitted for both statistical tests.*

**Part III: Nonparametric Statistical Testing**

*Note: You may use Python or R for implementing your coding solutions, manipulating the data, and creating visual representations. However, your responses to the following task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will****not****be accepted.*

G.  Describe a real-world organizational situation or issue in the provided dataset by doing the following:

1.  Provide **one** research question relevant to the dataset and any organizational needs that can be answered through data analysis.

2.  Identify the variables in the dataset that are relevant to answering your research question from part G1.

H.  Analyze the dataset further by doing the following:

1.  Identify a **nonparametric** statistical test that is relevant to your question from part G1.

2.  Develop null and alternative hypotheses related to your chosen nonparametric test from part H1.

3.  Write code (in either Python or R) to run the nonparametric test.

4.  Provide the output and the results of *any* calculations from the nonparametric statistical test you performed.

I.  Evaluate nonparametric test results by doing the following:

1.  Justify why you chose the statistical test identified in part G1 based on variables.

2.  Discuss test results, including the decision to reject or fail to reject the null hypothesis from part H2.

3.  Explain how stakeholders in the organization benefit from your choice of testing method.

J.  Summarize the implications of your nonparametric statistical testing by doing the following:

1.  Discuss the answer to your question from part G1.

2.  Discuss the limitations of your data analysis.

3.  Recommend a course of action based on your findings.

4.  Submit your code for *each* test.

*Note:****One****notebook can be submitted for*both*statistical tests.*