ALY6000: Introduction to Analytics

Student Assignment – Audits Data Analysis

Overview and Rationale

Being able to ask appropriate questions of data is an important part of the work of data analytics. It is also critical to be able to interpret the results of the analysis. This assignment is intended to familiarize you with the Audits data set and to get you thinking about key business questions you can answer from this data.

Course Outcomes

This assignment is directly linked to the following learning outcomes from the course syllabus:

- Visualize data in a compelling way to enable data driven storytelling.
- Conduct basic analytic tasks programmatically using the R language*

∗   R is not required for this assignment, but you may choose to use it if you feel confident in your R skills

Assignment Summary

There are two Excel sheets with data – one concerning "Good Working Practice" audits (GWP_Audits_Data) and one concerning "Computerized System Quality Assurance" audits (CSQA_Audits_Data). Please see the accompanying Data Dictionary to understand the fields and values.

You may use any software to perform the analyses specified below. Collaboration is encouraged, but you must not submit identical assignments.

The assignment has three parts. In the Appendix of this assignment, you are provided an example of how the questions in Part I should be answered.

Part I

Please review the Vertex Data Dictionary document as you review the Excel datasheets.

In order to understand the data we first need to run some descriptives on the data set. For both the GxP Audits and the CSQA Audits sheets, we want to look at the following variables:

- Audit Status
- In USA or OUS
- GxP Area
- Audit Type

- Audit Method
- Proposed Quarter

Start by providing the following for each variable:

1. A table that provides the frequency and percent of each value
2. A graphic representation of the count of each value
3. A graphic representation of the percent of each value
4. What business question do your descriptive analyses answer? Provide a brief discussion of the findings. If there are any unusual values, discuss them. Are these values "out of range"? If so, the data cleaning is not complete. Delete the out of range values and run the analysis again. If this is the case for any of the variables, present both the analysis with the out of range values and the analysis with the deleted out of range value.

Please first present your findings for the 2017 GxP Audits data and then the findings for the 2017 CSQA Audits data.

Note: Appendix 1 is only an example and you must complete your own analysis.

Part II

For each worksheet, compute the number of days lapsed between:

1. "Date of Intake" and "Date Q Sent". Name that variable "Days_Intake_QSent"
2. "Date Q Sent" and "Date Q Received". Name that variable "Days_QSent_QReceived. Based on the name of the variables, what do you think that variable means? Does it apply to all audits? Why?
3. "Date On Site Scheduled" and "Audit Start Date". Name that variable "Days_OnSiteScheduled_AuditStartDate. Does this variable apply to all audits? Why?
4. "Audit Start Date and "Audit End Date". Name that variable "Days_StartDate_EndDate".
5. "Audit End Date" and "Date Final Report Due". Name that variable "Days_AuditEnd_FinalReportDue"
6. "Date Final Report Due" and "Date of Completion". Name that variable "Days_FinalReportDue_CompletionDate"

Then, compute the mean and median for each of the 6 variables you have computed.

Part III

Would you recommend merging the sheets "2017 GxP Audits" and "2017 CSQA" Audits? Why or why not?

Rubrics

Category	Exceeds Standards	Meets Standards	Approaching Standards	Below Standards
Descriptives ALY6000-CO5	Accurately visualizes data in an innovative and compelling way that tells a story.	Accurately and creatively visualizes data in a useful way that tells a story.	Visualizes data in a useful way but visualizations may not be creative, or may have errors.	Data visualizations are actuate, but not visually appealing and hinder the story being told.
Data Analysis ALY6000-CO1	Correctly calculates all values based on the data and highlights each mean and median value.	Correctly calculates all values based on the data, including the mean and median values	Calculates values, but not all are correct or does not include mean and median values	Does not calculate all required values and does not include mean and median values
Recommendations ALY6000-CO4	Makes compelling recommendations supported by coherent and valid reasoning.	Makes appropriate recommendations supported by logical reasons	Makes good recommendations, but reasoning is obscure or not valid	Recommendations do not make sense given the reasons provided or recommendations are not supported by reasons.
Writing and Format	Assignment follows normal conventions of grammar and spelling and appropriate conventions of style and format that engages the reader	Assignment work follows normal conventions of grammar and spelling and has been carefully proofread. Appropriate conventions of style and format are used consistently.	Minimal errors in spelling, and grammar, and/or other writing conventions. Some transitions are choppy but not difficult to follow.	Frequent errors in spelling, grammar, and/or other writing conventions that distract the reader. Transitions are choppy and difficult to follow. Limited connection to the topic.

<p align="center">**Appendix**</p>

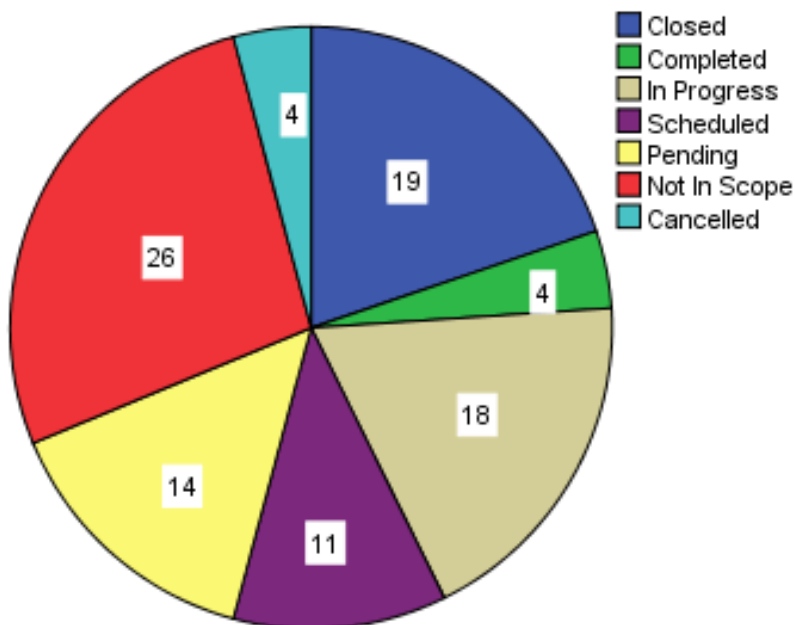<p align="center">**Assignment Part I Section a Example**</p>

Business Question:

What is the distribution of the status of the 2017 GxP Audits?
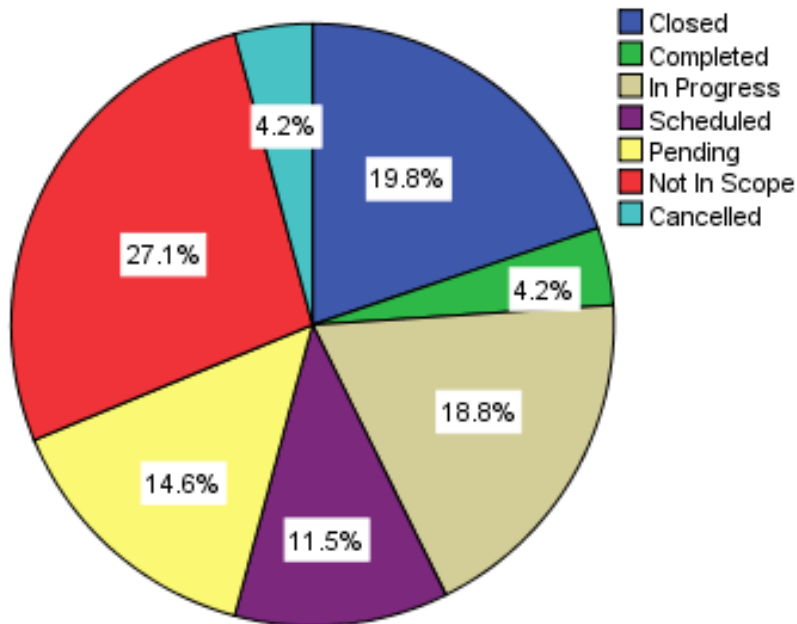
Analysis:

Descriptives Table

Audit Status		Frequency	Percent	Valid Percent
Valid	Closed	19	19.8	19.8
	Completed	4	4.2	4.2
	In Progress	18	18.8	18.8
	Scheduled	11	11.5	11.5
	Pending	14	14.6	14.6
	Not In Scope	26	27.1	27.1
	Cancelled	4	4.2	4.2
	Total	96	100.0	100.0

Audit Status Count

Audit Status Percentages

Discussion:

The data file includes information on 96 audits in 2017 for GxP areas. It is unclear if the data file includes all the known GxP audits in 2017 or if it only includes a subset.

A large percentage of all GxP Audits (27.1%) are not in scope.

19.8% of audits are closed and 4.2% of audits are completed. It is unclear what the difference between "closed" and "completed" audits is. We should perhaps ask the client. Do we really need two distinct values?

18.8% of the audits are in progress, 11.5% are scheduled and 14.6% are pending. For the pending audits, the dates of the audit process have not been established.

4.2% of the audits were canceled. It may be interesting to have a notes field where the reasons for cancelation are noted.