Alternate Assessment

Summary:

- Extract audience, behavioral and transactional insights using Google Analytics
- Estimate and interpret a clustering model using customers store visit data using the k-Means algorithm
- Estimate and interpret a logistic regression model to predict customer churn in the telecommunications sector

Learning Objectives:

- Create report views and interpret segment and behavior related information from Google Analytics
- Build a model to segment customers
- Build a predictive model for churn

Dear Learner,

Review the information presented in section A, section B & section C below and answer the questions mentioned in the respective sections. Please note that for answers in section A, you are required to also capture a screenshot of the Google Analytics report view to support your answers (since data from Google Analytics might vary slightly over time).

Section A – Google Analytics (10 marks)

Answer the following questions by going to the Google Analytics account of the Google Merchandise Store (https://analytics.google.com/analytics/web/demoAccount). You will need to login with your Google account

USE A DATE RANGE OF 01 January 2021 to 31 March 2021 TO ANSWER THE QUESTIONS IN THIS SECTION

YOUR ANSWERS TO EACH QUESTION SHOULD INCLUDE A SCREENSHOT OF THE REPORT VIEW EXTRACTED FROM GOOGLE ANALYTICS

Question 1 (5 marks)

Using the Acquisition > Social > Landing Pages report, report the top two landing pages for high average session duration on the 'Desktop' device category (Audience: All Users, Date Range: 01 January 2021 to 31 March 2021)?

Question 2 (5 marks)

In the Behavior > Site Content > Landing Pages report, create a System Segment for Referral Traffic. For this segment, identify the landing pages that delivered more than $100 in revenue during this period (Date Range: 01 January 2021 to 31 March 2021)?

SECTION B – Clustering & Linear Regression (25 marks)

A retail store collected data on shopping preferences of 40 visitors to their retail stores using a survey. This data is present in store_visits.xlsx. The variable descriptions to be used for analysis are presented below.

Variable Name	Description
ID	Customer ID
Income	Annual household income (in USD)
Frequency_Store_Visits	How often the respondent visits a store (on a scale of 0 – 7, with 7 being very frequent)
V1	Rating for the question "I find shopping enjoyable" on a scale of 1 – 7 (7 being highly enjoyable)
V2	Rating for the question "I am likely to compare prices while shopping" on a scale of 1 to 7 (7 being highly likely to compare)

Question 3 (5 marks)

Generate descriptive statistics for the variables Income, Frequency_Store_Visits, V1 and V2 using the KNIME Statistics node. Report your answer using the format below:

Column	Min	Max	Mean	Standard Deviation
V1				
V2				
Income				
Frequency_Store_Visits				

Question 4 (7 marks)

Using the variables Income, Frequency_Store_Visits, V1 and V2 as the inputs estimate a k-Means clustering solution for the data. **Use k = 3.**

Summarize the average values of each variable for the 3 clusters generated using KNIME. Fill in your answers in the table below.

RowID	Number of samples	V1	V2	Income	Frequency_Store_Visits
cluster_0					
cluster_1					
cluster_2					

Question 5 (3 marks)

Compare the cluster averages of all variables (i.e., V1, V2, Income and Frequency_Store_Visits) for the cluster with highest average income and the lowest average income. Present a rationale to describe the behavior of these two segments using the variables V1, V2 and Frequency_Store_Visits.

Question 6 (10 marks)

Estimate a linear regression with Frequency_Store_Visits as the outcome (dependent variable) and variables V1 & V2 as dependent variables using the Linear Regression Learner in KNIME.

Report the results of the regression in the following format using a significance level of 0.05:

RowID	Variable	Coefficient	p-value	Statistically Significant (Yes/No)?
Row1	V1			
Row2	V2			
Row3	Intercept			

SECTION C – Predicting Churn (15 marks)

Data in this section (Telco_customer_churn.xlsx) is extracted from a telecom company that served 7,043 customers in California. Apart from data on whether a customer churned or not, there is data on several other variables (variable definitions are presented in the table below).

Data Source: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

Column	Description
CustomerID	A unique ID that identifies each customer.
Count	A value used in reporting/dashboarding to sum up the number of customers in a filtered set.
Country	The country of the customer's primary residence.
State	The state of the customer's primary residence.

City	The city of the customer's primary residence.
Zip Code	The zip code of the customer's primary residence.
Lat Long	The combined latitude and longitude of the customer's primary residence.
Latitude	The latitude of the customer's primary residence.
Longitude	The longitude of the customer's primary residence.
Gender	The customer's gender Male, Female
Senior Citizen	Indicates if the customer is 65 or older Yes, No
Partner	Indicate if the customer has a partner Yes, No
Dependents	Indicates if the customer lives with any dependents Yes, No. Dependents could be children, parents, grandparents, etc.
Tenure Months	Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.
Phone Service	Indicates if the customer subscribes to home phone service with the company Yes, No
Multiple Lines	Indicates if the customer subscribes to multiple telephone lines with the company Yes, No
Internet Service	Indicates if the customer subscribes to Internet service with the company No, DSL, Fiber Optic, Cable.
Online Security	Indicates if the customer subscribes to an additional online security service provided by the company Yes, No
Online Backup	Indicates if the customer subscribes to an additional online backup service provided by the company Yes, No
Device Protection	Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company Yes, No
Tech Support	Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times Yes, No
Streaming TV	Indicates if the customer uses their Internet service to stream television programing from a third party provider Yes, No. The company does not charge an additional fee for this service.
Streaming Movies	Indicates if the customer uses their Internet service to stream movies from a third party provider Yes, No. The company does not charge an additional fee for this service.
Contract	Indicates the customer's current contract type Month-to-Month, One Year, Two Year.
Paperless Billing	Indicates if the customer has chosen paperless billing Yes, No
Payment Method	Indicates how the customer pays their bill Bank Withdrawal, Credit Card, Mailed Check
Monthly Charge	Indicates the customer's current total monthly charge for all their services from the company.

Total Charges	Indicates the customer's total charges, calculated to the end of the quarter specified above.
Churn Label	Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.
Churn Value	1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.
Churn Score	A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.
CLTV	Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.
Churn Reason	A customer's specific reason for leaving the company. Directly related to Churn Category.

Question 7 (10 marks)

Estimate a logistic regression model with the binary variable Churn Label as the outcome/dependent variables (i.e., Y) and the following variables as independent variables (i.e., X variables): Tenure Months, Multiple Lines, Streaming Movies. Report the statistical significance of all the independent variables. The output should be in the following format:

	Coefficient	**p-value**	**Statistically significant?**
Intercept			
Tenure Months			
Multiple Lines = No phone service			
Multiple Lines = Yes			
Streaming Movies = No internet service			
Streaming Movies = Yes			

Question 8 (5 marks)

Based on the output from Question 7, interpret the coefficient of (a) Multiple Lines = Yes and (b) Streaming Movies = No internet service. In your answer, you should relate these coefficients to the propensity of the customer to churn.

Please note:

- You are required to submit one file:
  - **Business Report(PDF):** In this, you need to submit all the answers to all the questions in a sequential manner. It should include a detailed explanation of the approach used, insights, inferences. You will be evaluated based on the business report

- Please go through the guidelines thoroughly before attempting the assessment.
- Any assignment found copied/ plagiarized with another person will not be graded and marked as zero.
- Please ensure timely submission as a post-deadline assignment will not be accepted.

Regards,
Program Office

Rubric

Question 1:
Correctly drilling down to the appropriate report and reporting the answer **5 marks**

Question 2:
Correctly drilling down to the appropriate report and reporting the answer **5 marks**

Question 3:
Computing and presenting the min, max, mean and standard deviation for all 4 variables **5 marks**

Question 4:
- Correctly estimating the k-Means algorithm with k = 3 and computing the number of samples per cluster **2 marks**
- Computing and presenting the average values of the variables by cluster in the specified format **5 marks**

Question 5:
Correctly inferring the relationship between the average values of V1, V2 and Frequency_Store_Visits of the two clusters to average incomes of the clusters **3 marks**

Question 6:
- Setting up the linear regression model and estimating the coefficients **8 marks**
- Interpreting p-values for statistical significance **2 marks**

Question 7:
- Setting up the logistic regression model and estimating the coefficients **8 marks**
- Interpreting p-values for statistical significance **2 marks**

Question 8:
- (a) Appropriately inferring the relationship between the regression coefficient and propensity to churn **2.5 marks**
- (b) Appropriately inferring the relationship between the regression coefficient and propensity to churn **2.5 marks**