

**ISSS602 DATA ANALYTICS LAB**

ASSIGNMENT 2

*Be Customer Wise or Otherwise*

**Prepared by:**

Sarthak Jagdish Nagapurkar

15th October 2023

Table of Contents

[1. Introduction 3](#_Toc148305928)

[2. Overview 3](#_Toc148305929)

[3. Data Used 3](#_Toc148305930)

[4. Data Preparation 3](#_Toc148305931)

[4.1 Converting ReqTimeUTC to correct datatype. 3](#_Toc148305932)

[4.2 Removing Duplicates 4](#_Toc148305933)

[4.3 Remove transactions with status ‘NA’ and derive fields for month and date of transaction. 5](#_Toc148305934)

[4.4 Pivot Table on Transaction Type and Replace missing values with 0. 6](#_Toc148305935)

[4.5 Get Total Amount transacted and Frequency of each transaction type throughout the year for each customer. 6](#_Toc148305936)

[4.6 Get data for approved and declined amounts and approved and declined frequencies for each customer. 7](#_Toc148305937)

[4.8 Deriving Clustering Variables 8](#_Toc148305938)

[5. Cluster Analysis 9](#_Toc148305939)

[5.1 Univariate Analysis of Clustering Variables 9](#_Toc148305940)

[5.2 Bivariate Analysis of Clustering Variables. 11](#_Toc148305941)

[5.3 Finding optimum number of clusters for K Means Clustering. 11](#_Toc148305942)

[5.4. K-means Clustering 14](#_Toc148305943)

[6. Cluster Summary and Interpretation 15](#_Toc148305944)

[7. Further Recommendations: 19](#_Toc148305945)

[APPENDIX A – Metadata 20](#_Toc148305946)

[Data Preparation Log 20](#_Toc148305947)

# 1. Introduction

In recent years, digital payments are becoming increasingly more prevalent across the gambling industry, bringing about questions regarding how this technology might influence individuals’ gambling behaviours and the potential of this technology to contribute to harm (Gainsbury & Blaszczynski, 2020). For example, easier access to funds could encourage overspending and less transparent payment methods could obscure the value of money. The digitization of gambling payments is partly due to a broader global shift toward cash substitutes (Khiaonarong & Humphrey, 2022). These transformations have created a research gap in terms of understanding how these digital payments might contribute to harm, and how they might provide opportunities for harm prevention.

# 2. Overview

The objective of this study is to employ Clustering Analysis to explore whether subgroups of gamblers can be distinguished by analysing their payment behaviour using data sourced from a gambling digital payments service provider. Specifically, this study aims to discern whether cluster analysis of payment transaction data (e.g., activity related to the deposit and withdrawal of funds) can reveal different types of gambling payment profiles.  While doing this, the secondary goal of this study is to identify whether certain markers of harm may be evident for further research exploration.

# 3. Data Used

For the purpose of this study, following dataset was used:

* ONLINE\_SPORTS\_DIB.csv – Data of the customer transactions

# 4. Data Preparation

Upon counterchecking, there are no missing values in all the datasets.

## 4.1 Converting ReqTimeUTC to correct datatype.

It was observed that ReqTimeUTC had a varchar datatype. It needed to be converted to datetime format. This was done using Convert Column in SAS Data Studio.

A screenshot of a computer

Description automatically generated

Figure 1: Convert ReqTimeUTC to datetime format

## 4.2 Removing Duplicates

Code snippet to check and see the duplicates:

A screenshot of a computer

Description automatically generated

Figure 2: Check and show existing duplicates (Source: Program to find duplicates.sas)

Part of the list of Duplicates found:

A screenshot of a computer

Description automatically generated

Remove Duplicates from SAS Data Studio will be used to remove all duplicates.

A screenshot of a computer

Description automatically generated

ONLINE\_SPORTS\_DIB\_1 file was created. 13 records were removed due to remove duplicates operation.

## 4.3 Remove transactions with status ‘NA’ and derive fields for month and date of transaction.

Following code snippet will help us do it.

A screenshot of a computer code

Description automatically generated

Figure 3: Remove transactions with status 'NA'. Derive transaction month and date fields.

## 4.4 Pivot Table on Transaction Type and Replace missing values with 0.

We can find transaction amount and frequency foe each transaction type in a given period by pivoting the table on Transaction type.

A screenshot of a computer

Description automatically generated

Figure 4: Pivot table on Transaction type to get Amount and Frequency for each transaction type.

## 4.5 Get Total Amount transacted and Frequency of each transaction type throughout the year for each customer.

A computer screen shot of a computer code

Description automatically generated

Figure 5: Total Amount Transacted and frequency for each transaction type for each customer.

## 4.6 Get data for approved and declined amounts and approved and declined frequencies for each customer.

Following code snippet will help us.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

Figure 6: Creating table containing approved sum and frequency and declined sum and frequency for each customer.

## 4.8 Deriving Clustering Variables

Following code snippet has been used to derive clustering variables.

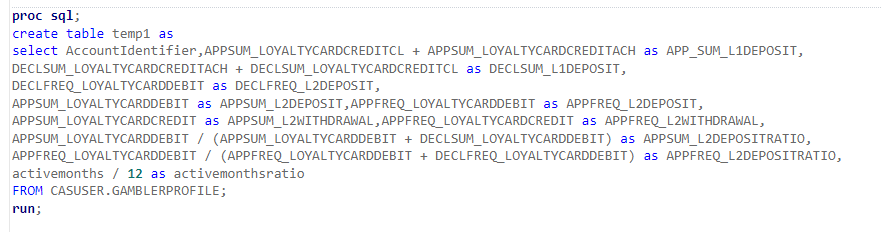


Figure 7: Deriving potential clustering variables.

Following table indicates potential clustering variables for each customer account identifier. There are 1465 total customer account identifiers (active gamblers):

|  |  |
| --- | --- |
| Variable Name | Significance |
| APPSUM\_L2DEPOSIT | Total Approved Sum for Level 2 Deposits for each customer |
| APPSUM\_L2WITHDRAWAL | Total Approved Sum for Level 2 Withdrawals for each customer |
| APPFREQ\_L2DEPOSIT | Total Approved Frequency for Level 2 Deposits for each customer |
| APPFREQ\_L2WITHDRAWAL | Total Approved Frequency for Level 2 Withdrawals for each customer |
| APPFREQ\_L2DEPOSITRATIO | Ratio of the number of Total Approved Level 2 Deposits to the Total number of attempted Level 2 Deposits for each customer. |
| activemonthsratio | Ratio of total active months (Months with Approved transactions) in the year to 12 |
| APPSUM\_L2DEPOSITRATIO | Ratio of the Total Approved Sum for Level 2 Deposits to the Total Attempted Level 2 Deposit Sum Transfer |
| DECLFREQ\_L2DEPOSIT | Total Declined Frequency for Level 2 Deposits for each customer. |
| APP\_SUM\_l1\_DEPOSIT | Total Approved Sum for Level 1 Deposits for each customer. |

# 5. Cluster Analysis

K-Means Clustering Analysis will be used to analyse the given dataset. Univariate and Correlation Analysis will be done before clustering to check for skewness and multicollinearity.

## 5.1 Univariate Analysis of Clustering Variables

A screenshot of a computer screen

Description automatically generated

Figure 8: Univariate Analysis of Clustering Variables

As shown in the SAS Viya Report above, all the clustering variables (except activemonthsratio) are skewed. Either log transform or cube transform will be used depending on direction of skewness to reduce skewness.

A computer screen shot of a code

Description automatically generated

Figure 9: Code for Log and Cube transforms for Clustering Variables

A group of blue graphs

Description automatically generated

Figure 10: Univariate Analysis of Transformed Clustering Variables

Log and Cube transformations have reduced skewness of Clustering Variables.

## 5.2 Bivariate Analysis of Clustering Variables.

Correlation Matrix is used to check if any two of clustering variables are highly correlated with each other. A blue squares with white text

Description automatically generated

Figure 11: Bivariate Analysis of Clustering Variables

Correlation Matrix indicates that APP\_SUM\_L1DEPOSIT and APP\_SUML2DEPOSIT are highly correlated with a coefficient of 0.9242. We will only consider APP\_SUML2DEPOSIT as a clustering variable and drop APP\_SUM\_L1DEPOSIT.

## 5.3 Finding optimum number of clusters for K Means Clustering.

For finding optimum number of Clusters the following code snippet has been used.

A screenshot of a computer program

Description automatically generated

Figure 12: CCC and ABC tests (Source: Determining number of clusters.sas)

Results for Cubic Clustering Criterion and Applied Box Criterion are as below.

A graph of a number of clusters

Description automatically generated

Figure 13: Plots for CCC, Pseudo F, Pseudo T-Squared

A table of data with numbers

Description automatically generated

Figure 14: Cubic Clustering Criterion Results

A screenshot of a data analysis

Description automatically generated

Figure 15: Applied Box Criterion Results

Local peak at Number of Clusters = 4 for CCC vs Number of Cluster Plot.

Applied box criterion for cluster values between 3 to 8 clearly indicates optimum number to be 4. Based on results of Cubic Clustering Criterion based on Ward's Minimum Variance Method, Pseudo F statistics, Pseudo t squared statistics, and Aligned Box Criterion the optimal number of clusters for K-Means Clustering has been derived to be 4.

## 5.4. K-means Clustering

K – means clustering has been performed using Cluster under Statistics in Explore and Visualize in SAS Viya.

A group of colorful objects

Description automatically generated with medium confidence

Figure 16: K Means Clustering

K-Means Clustering has been performed for the given transformed clustering variables.

A screenshot of a computer

Description automatically generated

Figure 17: K Means Model Summary

A screenshot of a computer

Description automatically generated

Figure 18: K Means Cluster Summary

A screenshot of a computer

Description automatically generated

Figure 19: Cluster Centroids

# 6. Cluster Summary and Interpretation

**Cluster 1: Active Gamblers with Very High Amount Wagered, High Wagering Frequency, and Lowest Amount Withdrawn from Wagering Account**

**(~32% of the customers)**

**A screenshot of a graph

Description automatically generated**

Figure 20: Parallel Coordinates plot for Cluster 1

• This cluster contains customers with

1. High amounts wagered (log\_appsuml2deposit mean = 3.802. Mean amount wagered = 6337.69 $)
2. High frequency of wagering (log\_appfreql2deposit = 1.904. Mean frequency = 78.4)
3. Low amounts withdrawn (log\_appsuml2withdrawal mean = 0.095. Mean amount withdrawn = 0.23 $)
4. High Active Months ratio (stdactivemonthsratio = 0.607. Active for ~9 months out of 12)
5. Low to medium frequency of declined wagering deposits. (Mean frequency = 1.69)
6. High frequency approval ratio for level 2 deposits (Mean frequency approval ratio = 0.96)
7. Very low frequency of level 2 withdrawals (~0)

These gamblers withdraw less from their wagering account and deposit huge amounts of money with high frequency which may indicate they are losing money. These are the gamblers which need to be analysed more to understand their demographics, behaviour patterns, financial transaction patterns. This will help us identify if they have a tendency towards compulsive gambling. From initial analysis, they seem to be compulsive gamblers.

**Cluster 2: Active Gamblers with Highest Amount Wagered, Highest Wagering Frequency, and Highest Amount Withdrawn from Wagering Account.**

**(~21% of the customers)**

**A diagram of a graph

Description automatically generated with medium confidence**

Figure 21: Parallel Coordinates plot for Cluster 2

• This cluster contains customers with

1. High amounts wagered (log\_appsuml2deposit mean = 3.987. Mean amount wagered = 9704.09 $)
2. High frequency of wagering (log\_appfreql2deposit = 2.138. Mean frequency = 133.89)
3. Very high amounts withdrawn (log\_appsuml2withdrawal mean = 3.52. Mean amount withdrawn = 3310.311 $)
4. Highest Active months ratio (stdactivemonthsratio = 0.778. Active for ~10 months out of 12)
5. Relatively High frequency of declined wagering deposits. (Mean frequency 3.67)
6. High frequency approval ratio for level 2 deposits (Mean frequency approval ratio ~ 0.96)
7. Very high frequency of level 2 withdrawals (~28)

These gamblers are the ones who are transferring huge amounts of money to their wagering account and withdrawing huge amounts of money from their wagering account. This might mean they are potentially winning huge sums of money leading to reinforcement of the online gambling tendencies.

**Cluster 3:  Slightly Inactive Gamblers with slightly low amount wagered, slightly low Wagering Frequency, and slightly low amount withdrawn from Wagering Account.**

**(~22% of customers)**

A screenshot of a computer screen

Description automatically generated

Figure 22: Parallel Coordinates Plot for Cluster 3

• This cluster contains customers with

1. Slightly low amounts wagered (log\_appsuml2deposit mean = 3.213. Mean amount wagered = 1632.05 $)
2. Slightly Low frequency of wagering (log\_appfreql2deposit = 1.4923. Mean frequency = 30.04)
3. Slightly Low amounts withdrawn (log\_appsuml2withdrawal mean = 2.63. Mean amount withdrawn = 426.58 $)
4. Slightly Inactive Months ratio (stdactivemonthsratio = -0.472. Active for ~5 months out of 12)
5. Low to Medium frequency of declined wagering deposits. (Mean frequency 1.71)
6. Slightly High frequency approval ratio for level 2 deposits (Mean frequency approval ratio ~ 0.92)
7. Slightly low frequency of level 2 withdrawals (~5.3)

These gamblers are relatively inactive compared to the first two cluster profiles. They have a lower frequency and lower amount deposited to their wagering account. This might indicate they have medium to low tendency towards online gambling and may continue in the same way or even lower their current wagering frequency and amount going further.

**Cluster 4:  Inactive Gamblers with lowest amount wagered, lowest Wagering Frequency, and very low amount withdrawn from Wagering Account.**

**(~23% of customers)**

**A diagram of a graph

Description automatically generated with medium confidence**

Figure 23: Parallel Coordinates Plot for Cluster 4

• This cluster contains customers with

1. Very Low amounts wagered (log\_appsuml2deposit mean = 2.686. Mean amount wagered = 484.28 $)
2. Very Low frequency of wagering (log\_appfreql2deposit = 0.98. Mean frequency = 8.54)
3. Slightly low amounts withdrawn (log\_appsuml2withdrawal mean = 0.12. Mean amount withdrawn = 0.31 $)
4. Very Low Active Months ratio (stdactivemonthsratio = -1.190. Active for ~3 months out of 12)
5. Low frequency of declined wagering deposits. (Mean frequency ~1.18)
6. Slightly High frequency approval ratio for level 2 deposits (Mean frequency approval ratio ~ 0.879)
7. Low frequency of Level 2 withdrawals. (Mean frequency ~ 0)

This cluster indicates very inactive gamblers. These are potentially those gamblers who didn’t find enough interest in online gaming gambling and hence have very low amount wagered and low frequency of wagering. This might indicate these gamblers might totally quit online sports gambling soon if they don’t find renewed interest to continue.

# 7. Further Recommendations:

• **Recommend collecting data regarding gamblers' demographics.**

Gamblers' demographic data like their age, income levels, education levels, and occupation could help us do in-depth analysis and help in better gambler profile segmentation.

• **Recommend collecting data for a greater number of transactions and gambler accounts.**

   The current data contains data for only 1468 customer accounts. While this might seem like a huge number it might be insufficient considering the active gambler population in the United   States. Transaction data containing > 2M records (currently 447853 records), might contain a significantly higher number of gambler accounts, will reduce the impact of outliers, and may lead to well-separated and less elongated clusters.

• **Recommend more efficient data collection.**

   a) The transaction data contained 13 duplicate values.

   b) In addition, two records had a transaction approval status 'NA'.

   c) Three customers had either an amount successfully wagered = 0 and frequency of amount successfully wagered = 0. This does not align with the definition of being an active gambler as no amount is wagered during the entire year. The data for these customers should have been excluded while preparing the dataset. An efficient and well-structured data collection process will ensure potential data quality issues like (a), (b), and (c) mentioned above do not occur.

• **Recommend collecting data about the reason for transactions being declined.**

This would help in analysing if the transaction failed due to insufficient funds in the customer accounts or due to technical issues in the digital transaction. This information can help in more efficient gambler profile segmentation.

# APPENDIX A – Metadata

# Data Preparation Log

|  |  |  |  |
| --- | --- | --- | --- |
| Item | Data Preparation Issue | Dataset | Action |
| 1 | Data type of ReqTimeUTC is considered as varchar. | Input tables:  ONLINE\_SPORTS\_DIB  Output tables:  ONLINE\_SPORTS\_DIB\_1 | Data type of ReqTimeUTC has been converted to datetime. |
| 2 | Online\_Sports\_DIB dataset contains duplicates | Input tables:  ONLINE\_SPORTS\_DIB  Output tables:  ONLINE\_SPORTS\_DIB\_1 | Remove Duplicates using SAS Data Studio. |
| 3 | Some transactions have status ‘NA’ | Input table:  ONLINE\_SPORTS\_DIB\_1  Output tables:  Participant\_1 | The data type of Interest group has been converted to character. This is not an absolute mandatory step for analysis. |
| 4 | Date and Month of Transaction is not directly usable. | Input table:  ONLINESPORTSDIBPREP  Output tables:  ONLINESPORTSDIBMONTH  Program:  Program for converting month online gambling.sas | Add fields for transactionmonth and traansactiondate and compute them using SAS Viya SQL proc. |
| 5 | Transaction Amount and frequency are not known for a group of multiple transactions. | Input table:  ONLINESPORTSDIBMONTH  Output tables:  ONLINESPORTSGROUPEDBY  Program:  Program for converting month online gambling.sas | Use sum, count in SAS Viya proc to compute transactionamount and Frequency grouped for multiple transactions. |
| 6 | Pivot table on Transaction Type | Input table:  ONLINESPORTSGROUPEDBY  Output tables:  WORK.transposed  Program:  Program for converting month online gambling.sas | Use SAS Viya sort and transpose procedures to pivot the table. |
| 7 | Missing values exist in the pivoted table. | Input table:  WORK.transposed  Output table:  missingremoved  Program:  Program for converting month online gambling.sas | Use stdize procedure in SAS Viya to replace missing values with 0 |
| 8 | Data for entire year like activedays, activemonths, amount transacted for each transaction type in the year not available. | Input table: missingremoved  Output table: gamblingbehav  Program:  Program for converting month online gambling.sas | Add and compute Fields for activedays, activemonths and compute amount transacted for each type for a customer for the entire year. Use procedure SQL in SAS Viya. Export file as GAMBLINBEHAV.csv |
| 11 | Data for total transaction amount approved for each type of transaction type for entire year not available. | Input table: gamblingbehav  Output table: approvedsum  Program: Program trying to get clustering variables.sas | Rename fields and use where and group by in SAS Viya SQL proc to get desired outcome. |
| 12 | Data for frequency of transactions approved for each type of transaction type for entire year not available. | Input table: gamblingbehav  Output table: approvedfrequency  Program: Program trying to get clustering variables.sas | Rename fields and use where and group by in SAS Viya SQL proc to get desired outcome. |
| 13 | Data for total transaction amount declined for each type of transaction type for entire year not available. | Input table: gamblingbehav  Output table: declinedsum  Program: Program trying to get clustering variables.sas | Rename fields and use where and group by in SAS Viya SQL proc to get desired outcome. |
| 14 | Data for frequency of transactions declined for each type of transaction type for entire year not available. | Input table: gamblingbehav  Output table: declinedfrequency  Program: Program trying to get clustering variables.sas | Rename fields and use where and group by in SAS Viya SQL proc to get desired outcome. |
| 15 | Data for approvedsum, approvedfrequency, declinedsum, declinedfrequency for one customer not available in a single table. | Input table: approvedsum, approvedfrequency, declinedsum, declinedfrequency  Output table: final  Program: Program trying to get clustering variables.sas | Join tables approvedsum, approvedfrequency, declinedsum, declinedfrequency on accountidentifier for active gamblers to get resultant tables. Export file as gamblerprofile.csv |
| 16 | Clustering variables need to be derived. | Input table: gamblingbehav  Output table: temp1  Program: Program for data transformations.sas | Calculate  a) approved sum for Level1 deposit, b)declined sum for Level1 deposit, c)declined frequency for Level2 deposit, d)approved sum Level 2 deposit, e)approved frequency Level 2 deposit, f)approved sum Level 2 withdrawal, g)approved frequency Level 2 withdrawal, h)approved sum level 2 deposit ratio, i) approved frequency level 2 deposit ratio, j) active months/ total months ratio |
| 17 | Transformation and standardization of required clustering variables. | Input table: temp1  Output table: ltransformed  Program: Program for data transformations.sas | Active months ratio has been standardized using mean and standard deviation obtained from SAS Visual Analytics. Rest of the clustering variables have been log/ cube transformed depending on the skewness. Export file as TRANSFORMEDDATA.csv |