Programming Fundamentals for Analytics
BUS4022
Individual Assignment #3 – Data Preparation and Visualization

Assignment Objective:

In this assignment you will be given a 'Raw' dataset describing several characteristics of customer loan accounts for a small bank. You will be required to import the raw data into a dataset, explore the data using a variety of techniques, identify data quality issues and make data repair/reject decisions, develop an analytical file that can support the application of advanced analysis techniques and finally, develop and apply data visualization techniques to support insights.

Deliverables:

In this assignment you will be producing a series of deliverables that will each be graded.

Deliverable 1: Data Quality Management and Audit Report

You will be required to provide a summarization of all the data quality analysis work done for each variable in the dataset. This report should provide summarization of the descriptive statistics for each variable, along with the identification of any data quality concerns and recommended actions (accept, reject and repair). Any derived or repaired data should be identified and clearly documented in terms of the underlying calculations. For example, if a data variable is being repaired through recoding, then the recoding logic should be clearly documented in the report. For any variables that are derived (such as Target Variables), they should be intuitively named with a clear explanation of the formula or calculation definition used to create the new variable.

Deliverable 2: Analytical File

You will be required to write a written response that summarizes the results and findings for each question or phase of the assignment. All responses must be supported by actual SAS output (in the form of a report or a graph as required). Please note that you are required to submit your SAS code and log files for this assignment. Please make sure code is clearly organized and documented for review. You will also be required to include supporting output from SAS procedures that you run.

Deliverable 3: Exploratory Data Analysis

You will be required to write a written response that summarizes the results and findings for each question or phase of the assignment. All responses must be supported by actual SAS output (in the form of a report or a graph as required).

Deliverable 4: Tableau Dashboard and Story

You will be required to develop a Tableau workbook that contains a combination of; underlying graphical reports, an interactive dashboard that summarizes the key results, and a presentation style

Tableau Story, that ***contains a clear narrative (i.e. story)*** around key findings and insights from your preliminary analysis of the data.

Background on the Case:

A small micro lending institution in Malawi, Africa has requested your assistance in better understanding the performance of their loan portfolio. The files provided are based on an actual case, however we will refer to the institution as the Malawi National Bank.

Micro loans are typically small loans that are provided in developing countries to assist people out of poverty. Recently the Malawi National Bank has been finding that many of the loans that have been advanced have not been paid back. To improve their loan portfolio, you have been retained to provide some analysis on their current portfolio, with the expectation of providing insights as to what may be the potential underlying factors that are causing the portfolio to deteriorate in performance.

You will use SAS programming to create a dataset for analysis that will include the information provided, plus several derived variables. The Malawi National Bank also has specific questions that they would like you to answer. In addition to answering their questions, they have requested you to explore the data using various data visualization techniques, to hopefully provide additional insight into their portfolio's performance to determine where they should focus their efforts on reducing the amount of 'bad' customers and improving on the amount of 'good' customers.

They need your preliminary analysis and data quality assessment report, along with a supporting interactive data dashboard that will start the conversation around what you're seeing and areas where you will consider doing further analysis and drill down. Your findings are due within the next 2 weeks.

Good Luck!

Instructions:

Metadata Definitions for the Raw Dataset

The LOAN_PORTFOLIO.CSV file that has been extracted for you has the following field definitions

AccountID

- A unique reference number for each account

BranchID

- A single digit number indicating the branch that has the loan. The codes are interpreted as follows:

2	Mandala
3	Kawale
4	Mzuzu
5	Blantrye
6	Salima
7	Mchinji

ProductID

- A four-digit character string indicating the type of product. The codes are interpreted as follows:

LN01	Group Business Loan
LN02	Individual Business Loan
LN03	Share Loan (used to purchase Shares)
LN04	Emergency Loan
LN05	Farming Loan
LN06	Woman's Loan
LN08	Payroll Secured Loan

DisbursedOn

- Date that the loan was advanced to the customer

Application Score

- 2-digit score ranging between 0 and 60. A higher score indicates that the branch 'predicted' that this application would be minimal risk of defaulting on the loan. A low score indicates that the branch 'predicted' that the loan was at a higher risk of defaulting.

ArrearsDays

- The number of days that the loan is in arrears (i.e. behind in its payments). The current date of this file extract was July 31, 2012.

Actual Application Grade

- Letter grade of loan application quality (A – High Quality, and D – Poor Quality)

Actual Good Bad Indicator

- Internal indicator where 1 – account is considered 'Good', and 0 – account is considered 'Bad'

Gender

- Single digit code indicating gender as follows;

M	Male
F	Female
G	Group

LiteracyLevel

- Single digit code indicating literacy level

U	Not Specified
C	College
E	Elementary
N	No formal education
O	No Formal Education
P	Primary
PT	Polytechnic
S	Secondary
T	Tertiary
U	University
X	Not Specified

Occupation

- Single digit code indicating occupation of loan applicant (no translation provided)

PR	Priest
PRO	Professor
B	Business People
C	Civil Servants
E	Employee
F	Farmer
M	Self Employed
N	Unemployed
O	Other
T	Traders
U	Not Specified

Marital Status

- Single digit code indicating marital status of loan applicant

D	Divorced
G	Group
M	Married
S	Single
U	Not Specified
W	Widow
WI	Widower

Purpose Code

- A 2-digit code indicating the purpose for the loan

1	Expanding Business
2	Land Purchase
3	House Construction
4	Buying a car
5	Business
6	School Fees
7	Buying a House
8	Paying school fees
9	Paying Medical Bills
10	Building House
11	Buying Farm Inputs
12	Purchase of Household Items
13	Other Purpose

Disbursement Amount

- The original amount of the loan that was provided to the customer. All financial figures are in local currency of Malawian Kwachas (1 US Dollar = 400 Malawian Kwachas)

Installment Amount

- The required monthly payments on the loan

ActualBalance

- The current balance of all unpaid portions of the loan

ArrearsAmount -ve number imply that the person had already paid additional amount.

- The current amount of all loan payments that have not yet been paid (are past due)

ArrearsPer - the number of times a person has not paid the loan.

Deliverable 1. Data Audit and Preliminary Assessment Report (/20 Marks)

Required components of Deliverable 1.

- Produce a PROC FREQ report for all variables that are categorical (i.e. the data is defining a category or group (i.e. branch, location, purpose code, etc.)
- Produce a PROC MEANS for all variables that are numeric (i.e. only focus on numeric variables that are measuring something, such as the balance, arrears, disbursement amount)
- Produce a PROC UNIVARIATE report for all variables that are numeric (i.e. only focus on numeric variables that are measuring something, such as the balance, arrears, disbursement amount)
- Using the Data Quality Assessment Report template AS A GUIDE, provide commentary and notes on the; accuracy, consistency and completeness of each variable. Based on your data quality assessment of each variable, determine the potential role that that variable will play in your analysis. If any repair or recoding is required of a raw variable, please summarize what the repair or recoding approach is. In situations where new derived variables are being created, please indicate them in your Data Quality/Audit Report as well.

Provide a brief write up of the interpretation of the key results in each report. Make note of most/least frequent cases, maximums and minimums, number of missing values, averages, etc.

Submit your Data Audit and Preliminary Assessment report as an attached word file called XX_Assignment_3_Part_1_Data_Audit_and_Preliminary_Assessment_Report_.doc where XX is replaced by your assigned group number (or first and last initials for non-group assignments). Please also include any supporting files (i.e. spreadsheets) as XX_Assignment_3_Part_1_Data_Audit_and_Preliminary_Assessment_Appendix.xls where XX is replaced by your assigned group number (or first and last initials for non-group assignments). Submit your supporting SAS programs as XX_Assignment_3_Audit.sas

Deliverable 2. Analytical File and Target Variable Creation (/20 Marks)

Required components of Deliverable 2.

Based on the findings of the Data Audit and Preliminary Assessment, you are required to create a 'clean' analytical file that has all the required data in a more organized manner. It is important that your analytical file is clearly laid out so that the raw (original) data elements are clearly labelled and identified, as well as all derived variables. It is also important that the variables are clearly identified or sub-grouped in terms of their role in the analysis. Any existing or derived variables that are potential candidate variables that can explain good or bad performance should be categorized as input factors, while variables that are existing or derived that describe good or bad performance, should be categorized as potential targets. (Hint: There are several variables already provided in the dataset that that can play the role of a target variable. They include; Application Score, Arrears Days, Actual Application Grade, Actual Good Bad Indicator, and Arrears Amount) For the purposes of this analysis, consider that the dataset was pulled on July 31, 2012.

Derive a calculated variable called "Days on File" as the difference between the disbursement date and the date of July 31, 2012

Derive a calculated variable called "Credit Grade" as the following;
If the Application Score is above 40 then the Credit Grade is "A``
If the Application Score is less than or equal to 40 and greater then 30 then the Credit Grade is `B`
If the Application Score is less than or equal to 30 and greater then 20 then the Credit Grade is `C`
If the Application Score is less than or equal to 20 then the Credit Grade is `D``
Using data transformation techniques covered in this course, create a MINIMUM of 5 additional derived or recoded input variables that you think may provide further explanatory ability to poor loan performance. You may also want to consider recoding variables to reduce the number of categories

Target Variable Definition:

The dataset contains several variables that in some way potentially describe what a bad loan is. Some of these variables were developed from previous models (i.e. Application Score, Actual Application Grade, and Actual Good Bad Indicator). Since you're not sure whether you can trust these derived variables, you are required to create your own target variables based on some combination of arrears amounts and arrears days. Using the techniques covered in this course, create a MINIMUM of 5 potential target variables that define a bad loan. For example, you can create a continuous target variable based on the number of days in arrears, or the dollar amount in arrears, or some combination of both arrears days and arrears dollars. Your variables can be continuous, categorical, or binary. Each target variable should represent a different 'flavour' of what a bad loan is. Try to come up with some very contrasting target variables in terms of the data type and calculation definition. In the final phase of the analysis, you will focus on the target variable(s) that you think are most appropriate in providing the best explanation of the input factors.

Submit your Analytical File as an attached Excel Workbook called;

XX_Assignment_3_Part_2_Analytical_File.xls where XX is replaced by your assigned group number (or first and last initials for non-group assignments). Submit your supporting SAS programs as XX_Assignment_3_Analytical_File.sas

Please also include the supporting explanations of all derived variables in a second excel sheet tab within the workbook.

Deliverable 3. - Exploratory Data Analysis (/10 Marks)

Required components of Deliverable 3.

Now that you have had a chance to complete your preliminary analysis and prepare your analytical file, the Malawi National Bank has a few basic questions that they would like answered to assure them that they are comfortable with your ability to handle their data.

Using the exploratory analysis method of your choice provide answers to the following. For each question provide supporting output from the SAS procedure of your choice.

(1 Marks) What Literacy Level occurs most frequently?

(1 Marks) Which Branch is largest in terms of the number of customers?

(1 Marks) Which Branch disburses the largest average loan?

(1 Marks) Which Customer AccountID has the largest number of days in Arrears?

(1 Marks) Which Gender Group (M for Males, and F for Females) has a higher proportion of "Bad" loans?

After answering their basic questions, this is your opportunity to throw all your SAS and analysis know how at this dataset and find interesting information. You can take the analysis in any direction that you want, however, you will be evaluated based on the amount of insight that you are able to derive from the information that is presented to you. At a minimum, you must demonstrate at least 1 ***relevant*** example from each of the following SAS procedures:

1. Proc freq
2. Proc means
3. Proc tabulate
4. Proc gchart
5. Proc gplot
6. Proc sql
7. Proc univariate
8. Proc sort
9. Proc contents
10. Proc summary

All output reports must be accompanied by a written summarization of the key findings of the analysis and your interpretation or recommendations.

Submit the responses to the Malawi National Bank's questions along with your analysis report as an attached word file called XX_Assignment_3_Part_3_EDA.doc where XX is replaced by your assigned group number (or first and last initials for non-group assignments). Submit your supporting SAS programs as XX_Assignment_3_EDA.sas

Deliverable 4. - Tableau Dashboard and Data Story (/50 Marks)

Required components of Deliverable 4.

The Banking Regulators have requested that the Malawi National Bank produce a key performance indicator dashboard that contains high level summaries of the portfolio's performance. Using Tableau Desktop, you have been tasked to develop a more Executive Level summarization of the key graphics, that includes a dashboard. In addition to the dashboard the Regulators have requested that the Malawi National Bank provide a brief presentation/story using Tableau that points out the key trends and observations based on the analysis. Make sure that your story has a combination of; graphics, text and interactions that will allow the regulators to get a better sense of the issues and concerns that need to be monitored on an ongoing basis.

Submit your Tableau files as a published 'read only' Tableau packaged workbook (with your dataset embedded) that can be distributed through the Tableau Reader called XX_Assignment_3_Part_4_Dashboard.twbx

(Please Note: DO NOT submit your original Tableau file (with the .twb extension) – you must do a File ….. Save As …… and select the Save as type: Tableau Packaged Workbook (*.twbx))

Authors Note:

The following case study was based on an actual field project. For further information and background on the case refer to:

https://allaninmalawi.wordpress.com/about/