

**ANL252**

**Python for Data Analytics**

**End of Course Assessment**

**July 2022 Presentation**

**Ong Jia Ying - B1410552**

#### Question 1

This credit facility dataset to be analyzed comprises records of customers’ demographics, amount owed, repayment history/status etc. The data dictionary of this dataset is depicted in Appendix 1.

List the categorical and numeric variables in this dataset.

#### Categorical variables:

"ID", "GENDER", "EDUCATION", "MARITAL", "S(n)", "RATING"

#### Numeric Variables:

"LIMIT", "BALANCE", "INCOME", "AGE", "B(n)", "R(n)"

#### Question 2

Conduct four (4) data pre-processing tasks for the analysis of the data, explaining results obtained.

**Jupyter Notebook:**

import pandas as pd

import numpy as np

#read csv file

df = pd.read\_csv("ECA\_data.csv")

df.isnull().sum() # check for missing data

df.shape #check number of rows in data

#replace '-1' and '199' in column 'AGE' with NaN, -1 and 199 is impossible value in csv data

df['AGE'] = df['AGE'].replace([-1,199],None)

#drop all rows with missing value

df = df.dropna(axis = 0, how = 'any')

df.shape #check number of rows drop = 59 rows

The data set have 38 missing values under 'MARITAL', 13 missing values under 'EDUCATION' and 10 impossible value (-1, 199) under 'AGE'. Solution is to drop rows with missing and impossible data.

**Jupyter Notebook:**

#check for duplicated data

df.loc[df.duplicated(keep = 'first'),:]

#drop rows with duplicated data

df.drop\_duplicates(inplace = True)

df.shape #check if rows are dropped

There are 3 duplicated rows in the data set. Solution is to drop them.

**Jupyter Notebook:**

#check data set for formating issue.

df.info()

#strip '$' from column 'R3' data

df['R3'] = df['R3'].map(lambda x: x.lstrip('$'))

#strip comma from column 'R3' data

df['R3'] = df['R3'].str.replace((','),'')

#convert column R3 to int64

df['R3'] = pd.to\_numeric(df['R3'])

df.dtypes #show dtypes

There are non integer values in column 'R3' which might affect tabulation later. Solution is to remove the '$' and ',' from 'R3' data.

#### Question 3

Articulate five (5) relevant insights of the data, with supporting visualization for each insight.

**Jupyter Notebook:**

output = pd.pivot\_table(data=df,

index=['GENDER'],

values=['INCOME'],

aggfunc='mean')

output

output.plot(kind="bar", title = "Average Income by Gender")

0 = Male, 1 = Female. The average income per annum for Male and Female is approximately the same, between $170K - $180K.

**Jupyter Notebook:**

output1 = pd.pivot\_table(data=df,

index=['EDUCATION'],

values=['INCOME'],

aggfunc='mean')

output1

output1.plot(kind="bar", title = "Average Income by Education")

0 = Others, 1 = Postgraduate, 2 = Tertiary, 3 = High School. Average income per annum is the highest amoung postgraduate, second is others, third is tertiary graduates and the lowest is high school graduates.

**Jupyter Notebook:**

output2 = pd.pivot\_table(data=df,

index=['MARITAL'],

values=['INCOME'],

aggfunc='mean')

output2

output2.plot(kind="bar", title = "Average Income by Marital Status")

0 = Others, 1 = Single, 2 = Married. Average income earned is the highest by singles followed by married people and the lowest is amoung other marital status.

#### Question 4[¶](http://localhost:8892/notebooks/ANL252_ECA_B1410552_OngJiaYing_4Sep2022.ipynb#Question-4)

Perform linear regression modelling to predict the variable, B1, explaining the approach taken, including any further data pre-processing.

**Jupyter Notebook:**

from sklearn import linear\_model

from sklearn.model\_selection import train\_test\_split

df['BALANCE'] = df['BALANCE'].astype(int)

LR = linear\_model.LinearRegression()

LR.fit(df[['BALANCE']],df.B1)

LR.predict(1000)