```
#
# Purdue University Global
#
# IN402 - Modeling and Predictive Analysis
#
# Unit 7 Assignment / Module 5 Part 1 Competency Assessment
#
# Generating Descriptive Statistics
#
# Jupyter Notebook Code
#

# [1] ###############################################################################

# Library and data import.

# import all necessary initial libraries, including numpy, pandas,
# seaborn, matplotlib and math.

import sys

# Ignoring warnings

if not sys.warnoptions:
    import warnings

warnings.simplefilter("ignore")

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

import math

# [2] ###############################################################################

# Import the dataset into the development environment.

df = pd.read_csv('/home/codio/workspace/data/IN402/Churn_Modelling.csv')

# [3] ###############################################################################

# In the paper, describe the datasource and what you intend to use
# the libraries for.

# Explorative Analysis.

# Explore the contents of the dataset using .head()

df.head(10) # (NOTE: if coding in PyCharms use print(df.head(10)) instead)

# [4] ###############################################################################

# Explore the contents of the dataset using tail()

df.tail(10)

# [5] ###############################################################################

# Check if there are any missing values using isnull() functions,
# and remove them using .dropna() function (if any).

df.isna()

# [6] ###############################################################################

# Check if there are null values anywhere

df.isnull().sum()

# [7] ###############################################################################

# Check the structure and if there are any missing values using .info() function.

df.info()

# [8] ###############################################################################

# Check the descriptive statistics on numeric variables
# using the .describe() function.

df.describe()

# [9] ###############################################################################

# Check the variance of each variable

# Importing statistics module

from statistics import variance

# Set attribute values

creditScore = df['CreditScore']

age = df['Age']

tenure = df['Tenure']

balance = df['Balance']

estimatedSalary = df['EstimatedSalary']

# [10] ##############################################################################

# Display variance values

print("Variance of CreditScore is % s "% (variance(creditScore)))

print("Variance of Age is % s "% (variance(age)))

print("Variance of Tenure is % s "% (variance(tenure)))

print("Variance of Balance is % s "% (variance(balance)))

print("Variance of EstimatedSalary is % s " %(variance(estimatedSalary)))

# [11] ##############################################################################

# Build a plot to visualize customers that churned and that did not churn.

# 1st plot - between customers that churned and that did not churn

sns.countplot(x = "Exited", data = df)

# [12] ##############################################################################

# Calculate the percentage of churned customers

# Percentage of churned customers

total_customers = len(df.index)

customers_churned = df.groupby('Exited').Exited.count()[1]

perc_cust_churned = customers_churned/total_customers

print(perc_cust_churned)

# [13] ##############################################################################

# Build a histogram of credit scores for all customers

df['CreditScore'].plot.hist(bins=100, figsize=(10,5))

# Identify unique values in the Geography column

df['Geography'].unique()

# Plot the geography for all customers

sns.countplot(x='Geography', hue='Exited', data=df)

# [14] ##############################################################################

# Plot the geography for churned/non-churned customers

sns.countplot(x='Geography', hue='Exited', data=df)

# [15] ##############################################################################

# Plot the gender by the churn status

sns.countplot(x = "Exited", hue = "Gender", data = df)

# [16] ##############################################################################

# Calculate the percentage of customers by gender

churned_by_gender = df.groupby(['Gender'])['Exited'].sum()

print(churned_by_gender)

# [17] ##############################################################################

# Calculate the churn number by gender (Male):

churned_males = churned_by_gender['Male']

print('Churned males:'+ str(churned_males))

# [18] ##############################################################################

# Calculate the churn number by gender (Female):

churned_females = churned_by_gender['Female']

print('Churned females: '+ str (churned_females))

# [19] ##############################################################################

# Plot a histogram to compare the age by churn status

# Compare the age for churned

df['Age'].plot.hist()

# Plot a boxplot to identify the churned/non-churned customers by age

sns.boxplot(x="Exited", y="Age", data=df)

plt.ylim(0, 100)

# [20] ##############################################################################

# Plot the tenure for churned/non-churned customers

sns.countplot(x='Tenure', hue='Exited', data=df)

# [21] ##############################################################################

# Plot the histogram of balance for all customers.

df['Balance'].plot.hist()

# Plot a number of products by churned/non-churned status

sns.countplot(x='NumOfProducts', hue='Exited', data=df)

# [22] ##############################################################################

# Plot the Credit Card ownership by churned/non-churned status

sns.countplot(x = "HasCrCard", hue = "Exited", data = df)

# [23] ##############################################################################

# Calculate the credit card ownership by churned/non-churned status

churned_by_cc = df.groupby(['HasCrCard'])['Exited'].sum()

churned_no_cc = churned_by_cc[0]

print('Churned with credit card: '+ str(churned_no_cc))

# [24] ##############################################################################

# Calculate the credit card ownership by churned/non-churned status

churned_cc = churned_by_cc[1]

print('Churned with no credit card: ' + str (churned_cc))

# [25] ##############################################################################

# Plot the active hours for the customers by churned/non-churned status.

sns.countplot(x = 'IsActiveMember', hue = "Exited", data = df)

# [26] ##############################################################################

# Plot the estimated salary for all customers

df['EstimatedSalary'].plot.hist(bins=10000, figsize=(10,5))

plt.xlabel('EstimatedSalary')

# ###################################################################################
```