Applied ML: Data Analytics in Python

Sanjana Sahayaraj

Install PySpark

Fetch Java and Spark

In [1]:
```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://dlcdn.apache.org/spark/spark-3.2.2/spark-3.2.2-bin-hadoop3.2.tgz
```

Uncompress Spark

In [2]:
```
!tar xf spark-3.2.2-bin-hadoop3.2.tgz
```

Install PySpark and set environment variable

In [3]:
```
!pip install -q findspark
```

In [4]:
```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.2.2-bin-hadoop3.2"
```

In [5]:
```
import findspark
findspark.init()
```

Create Spark Session

In [6]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("sample").enableHiveSupport().getOrCreate()
```

Import Pandas

In [7]:
```
import pandas as pd
```

Reading data from my google drive

Original data link: https://www.kaggle.com/datasets/parisrohan/credit-score-classification

In [8]:
```
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials
```

In [9]:
```
# Authenticate notebook to read your Google Drive
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

In [14]:
```
# Alternately data can also be ready from github using pandas as follows:
# url = 'https://raw.githubusercontent.com/SanjanaSahayaraj/AppliedML/main/Analytics/Data/credit-data/train.csv
# pandas_df = pd.read_csv(url, low_memory=False)
```

In [10]:
```
downloaded = drive.CreateFile({'id':'16ca5d_xOB90hfG03AGeroMnqs5QvNqBW'}) #change this to your GDrive location
downloaded.GetContentFile('train.csv')
```

The data will be loaded in a datastructure called Dataframe which is a 2D structure to store and operate on tabular data with rows and columns.

Differences between pandas and spark dataframe: https://www.geeksforgeeks.org/difference-between-spark-dataframe-and-pandas-dataframe/

In [11]:
```
pandas_df = pd.read_csv('train.csv', low_memory=False)
```

In [12]:
```
spark_df = spark.read.option("header",True).csv('train.csv')
```

Display sample data

In [14]:
```
pandas_df.head(10) #display 10 rows in pandas dataframe
```

Out[14]:

	ID	Customer_ID	Month	Name	Age	SSN	Occupation	Annual_Income	Monthly_Inhand_Salary	Num_Bank_Acc
0	0x1602	CUS_0xd40	January	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	
1	0x1603	CUS_0xd40	February	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	
2	0x1604	CUS_0xd40	March	Aaron Maashoh	-500	821-00-0265	Scientist	19114.12	NaN	
3	0x1605	CUS_0xd40	April	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	
4	0x1606	CUS_0xd40	May	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	
5	0x1607	CUS_0xd40	June	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	NaN	
6	0x1608	CUS_0xd40	July	Aaron Maashoh	23	821-00-0265	Scientist	19114.12	1824.843333	
7	0x1609	CUS_0xd40	August	NaN	23	#F%$D@*&8	Scientist	19114.12	1824.843333	
8	0x160e	CUS_0x21b1	January	Rick Rothackerj	28_	004-07-5839	_______	34847.84	3037.986667	
9	0x160f	CUS_0x21b1	February	Rick Rothackerj	28	004-07-5839	Teacher	34847.84	3037.986667	

10 rows × 28 columns

In [16]:
```
spark_df.show(7) # display 7 rows in spark dataframe
```

```
+------+-----------+--------+-------------+----+-----------+----------+-------------+--------------------+----
-------------+---------------+-------------+-----------+--------------------+------------------+-------------
--------+--------------------+--------------------+----------+----------------+------------------------+-------
-------------+--------------------+--------------------+------------------------+-------------------+---------
--------+------------+
|    ID|Customer_ID|   Month|         Name| Age|        SSN|Occupation|Annual_Income|Monthly_Inhand_Salary|Num_
Bank_Accounts|Num_Credit_Card|Interest_Rate|Num_of_Loan|        Type_of_Loan|Delay_from_due_date|Num_of_Delayed
_Payment|Changed_Credit_Limit|Num_Credit_Inquiries|Credit_Mix|Outstanding_Debt|Credit_Utilization_Ratio|  Credi
t_History_Age|Payment_of_Min_Amount|Total_EMI_per_month|Amount_invested_monthly|   Payment_Behaviour|   Monthly
_Balance|Credit_Score|
+------+-----------+--------+-------------+----+-----------+----------+-------------+--------------------+----
-------------+---------------+-------------+-----------+--------------------+------------------+-------------
--------+--------------------+--------------------+----------+----------------+------------------------+-------
-------------+--------------------+--------------------+------------------------+-------------------+---------
--------+------------+
|0x1602|  CUS_0xd40| January|Aaron Maashoh|  23|821-00-0265| Scientist|     19114.12|   1824.8433333333328|
            3|              4|            3|          4|Auto Loan, Credit...|                 3|
       7|               11.27|                 4.0|         _|          809.98|      26.822619623699016|22 Year
s and 1 Mo...|                   No|  49.57494921489417|      80.41529543900253|High_spent_Small_...|312.494088
67943663|        Good|
|0x1603|  CUS_0xd40|February|Aaron Maashoh|  23|821-00-0265| Scientist|     19114.12|                 null|
            3|              4|            3|          4|Auto Loan, Credit...|                -1|
    null|               11.27|                 4.0|      Good|          809.98|       31.94496005538421|
           NA|                   No|  49.57494921489417|     118.28022162236736|Low_spent_Large_v...|284.629162
49607184|        Good|
|0x1604|  CUS_0xd40|   March|Aaron Maashoh|-500|821-00-0265| Scientist|     19114.12|                 null|
            3|              4|            3|          4|Auto Loan, Credit...|                 3|
       7|                   _|                 4.0|      Good|          809.98|       28.60935202206993|22 Year
s and 3 Mo...|                   No|  49.57494921489417|       81.699521264648|Low_spent_Medium_...| 331.20986
28537912|        Good|
|0x1605|  CUS_0xd40|   April|Aaron Maashoh|  23|821-00-0265| Scientist|     19114.12|                 null|
            3|              4|            3|          4|Auto Loan, Credit...|                 5|
       4|                6.27|                 4.0|      Good|          809.98|      31.377861869582354|22 Year
s and 4 Mo...|                   No|  49.57494921489417|      199.4580743910713|Low_spent_Small_v...|223.451309
72736786|        Good|
|0x1606|  CUS_0xd40|     May|Aaron Maashoh|  23|821-00-0265| Scientist|     19114.12|   1824.8433333333328|
            3|              4|            3|          4|Auto Loan, Credit...|                 6|
    null|               11.27|                 4.0|      Good|          809.98|      24.797346908844986|22 Year
s and 5 Mo...|                   No|  49.57494921489417|     41.420153086217326|High_spent_Medium...|341.489231
03222177|        Good|
|0x1607|  CUS_0xd40|    June|Aaron Maashoh|  23|821-00-0265| Scientist|     19114.12|                 null|
            3|              4|            3|          4|Auto Loan, Credit...|                 8|
       4|                9.27|                 4.0|      Good|          809.98|       27.26225871052017|22 Year
s and 6 Mo...|                   No|  49.57494921489417|      62.430172331195294|            !@9#%8| 340.47921
17872438|        Good|
|0x1608|  CUS_0xd40|    July|Aaron Maashoh|  23|821-00-0265| Scientist|     19114.12|   1824.8433333333328|
            3|              4|            3|          4|Auto Loan, Credit...|                 3|
      8_|               11.27|                 4.0|      Good|          809.98|       22.53759303178384|22 Year
s and 7 Mo...|                   No|  49.57494921489417|      178.3440674122349|Low_spent_Small_v...| 244.56531
67062043|        Good|
+------+-----------+--------+-------------+----+-----------+----------+-------------+--------------------+----
-------------+---------------+-------------+-----------+--------------------+------------------+-------------
--------+--------------------+--------------------+----------+----------------+------------------------+-------
-------------+--------------------+--------------------+------------------------+-------------------+---------
--------+------------+
only showing top 7 rows
```

Get shape and column names

In [17]:
```
pandas_df.columns #columns displayed in pandas
```

Out[17]:
```
Index(['ID', 'Customer_ID', 'Month', 'Name', 'Age', 'SSN', 'Occupation',
       'Annual_Income', 'Monthly_Inhand_Salary', 'Num_Bank_Accounts',
       'Num_Credit_Card', 'Interest_Rate', 'Num_of_Loan', 'Type_of_Loan',
       'Delay_from_due_date', 'Num_of_Delayed_Payment', 'Changed_Credit_Limit',
       'Num_Credit_Inquiries', 'Credit_Mix', 'Outstanding_Debt',
       'Credit_Utilization_Ratio', 'Credit_History_Age',
       'Payment_of_Min_Amount', 'Total_EMI_per_month',
       'Amount_invested_monthly', 'Payment_Behaviour', 'Monthly_Balance',
       'Credit_Score'],
      dtype='object')
```

In [18]:
```
pandas_df.shape #shape displayed in pandas
```

Out[18]:
```
(100000, 28)
```

In [19]:
```
spark_df.printSchema() #schema displayed in spark including data types
```

```
root
 |-- ID: string (nullable = true)
 |-- Customer_ID: string (nullable = true)
 |-- Month: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- SSN: string (nullable = true)
 |-- Occupation: string (nullable = true)
 |-- Annual_Income: string (nullable = true)
 |-- Monthly_Inhand_Salary: string (nullable = true)
 |-- Num_Bank_Accounts: string (nullable = true)
 |-- Num_Credit_Card: string (nullable = true)
 |-- Interest_Rate: string (nullable = true)
 |-- Num_of_Loan: string (nullable = true)
 |-- Type_of_Loan: string (nullable = true)
 |-- Delay_from_due_date: string (nullable = true)
 |-- Num_of_Delayed_Payment: string (nullable = true)
 |-- Changed_Credit_Limit: string (nullable = true)
 |-- Num_Credit_Inquiries: string (nullable = true)
 |-- Credit_Mix: string (nullable = true)
 |-- Outstanding_Debt: string (nullable = true)
 |-- Credit_Utilization_Ratio: string (nullable = true)
 |-- Credit_History_Age: string (nullable = true)
 |-- Payment_of_Min_Amount: string (nullable = true)
 |-- Total_EMI_per_month: string (nullable = true)
 |-- Amount_invested_monthly: string (nullable = true)
 |-- Payment_Behaviour: string (nullable = true)
 |-- Monthly_Balance: string (nullable = true)
 |-- Credit_Score: string (nullable = true)
```

In [20]:
```
pandas_df.dtypes #datatypes of columns displayed in pandas
```

Out[20]:
```
ID                           object
Customer_ID                  object
Month                        object
Name                         object
Age                          object
SSN                          object
Occupation                   object
Annual_Income                object
Monthly_Inhand_Salary       float64
Num_Bank_Accounts             int64
Num_Credit_Card               int64
Interest_Rate                 int64
Num_of_Loan                  object
Type_of_Loan                 object
Delay_from_due_date           int64
Num_of_Delayed_Payment       object
Changed_Credit_Limit         object
Num_Credit_Inquiries        float64
Credit_Mix                   object
Outstanding_Debt             object
Credit_Utilization_Ratio    float64
Credit_History_Age           object
Payment_of_Min_Amount        object
Total_EMI_per_month         float64
Amount_invested_monthly      object
Payment_Behaviour            object
Monthly_Balance              object
Credit_Score                 object
dtype: object
```

In [21]:
```
print((spark_df.count(), len(spark_df.columns))) #shape displayed in spark
```

```
(100000, 28)
```