data engineer assignment

July 8, 2021

[1]:

```
from pyspark.sql import SparkSession
spark = SparkSession \
    .builder \
    .appName("data engineer assignment") \
    .getOrCreate()
```

[2]:

```
df = spark.read.csv('myFile0.csv',header=True)
```

[25]:

```
df.show(5)
```

```
+-------+-----------+----------+----------+--------------+---------+------------
+--------------------+----------+---------+-------+
|   Name|Customer_Id| Open_Date| Cosult_Dt|Vaccination_Id|  Dr_Name|
State|              Country|       DOB|IS_Active|Pincode|
+-------+-----------+----------+----------+--------------+---------+------------
+--------------------+----------+---------+-------+
|Aigneis|          1|1951-10-02|1920-10-03|           810|   Donnie|
Vientiane|      Czech Republic|1974-05-15|       A |     IM|
|  Elena|          2|1982-01-28|1996-12-26|           811|Christian|   Cebu
City|           Australia|1975-09-03|       A |     LU|
| Lonnie|          3|2011-07-30|1925-03-27|           812|    Lorie|
Luanda|             Tokelau|1978-09-24|       A |     NG|
|Lorenza|          4|2007-10-12|1960-04-04|           813|    Fayre|
Timbuktu|Macedonia, The Fo...|1977-09-13|       A |     AL|
|   Tani|          5|1944-04-10|1935-08-07|           814|   Dorene|Split
(city)|              Jordan|1971-01-10|       A |     KW|
+-------+-----------+----------+----------+--------------+---------+------------
+--------------------+----------+---------+-------+
only showing top 5 rows
```

[26]:

```
df = df.withColumnRenamed('Customer_name','Name')
df.show(5)
```

```
+-------+-----------+----------+----------+--------------+---------+------------
+--------------------+----------+---------+-------+
|   Name|Customer_Id| Open_Date| Cosult_Dt|Vaccination_Id|  Dr_Name|
State|             Country|       DOB|IS_Active|Pincode|
+-------+-----------+----------+----------+--------------+---------+------------
+--------------------+----------+---------+-------+
|Aigneis|          1|1951-10-02|1920-10-03|           810|   Donnie|
Vientiane|      Czech Republic|1974-05-15|       A |     IM|
|  Elena|          2|1982-01-28|1996-12-26|           811|Christian|   Cebu
City|           Australia|1975-09-03|       A |     LU|
| Lonnie|          3|2011-07-30|1925-03-27|           812|    Lorie|
Luanda|             Tokelau|1978-09-24|       A |     NG|
|Lorenza|          4|2007-10-12|1960-04-04|           813|    Fayre|
Timbuktu|Macedonia, The Fo...|1977-09-13|       A |     AL|
|   Tani|          5|1944-04-10|1935-08-07|           814|   Dorene|Split
(city)|              Jordan|1971-01-10|       A |     KW|
+-------+-----------+----------+----------+--------------+---------+------------
+--------------------+----------+---------+-------+
only showing top 5 rows
```

[27]:
```
df = df.withColumnRenamed('Last_consulted_Date','Cosult_Dt')
```

[9]:
```
df.registerTempTable('table')
```

[17]:
```
df2 = spark.sql("SELECT * from table WHERE Country=='United States'")
df2.collect()
```

[17]:
```
[Row(Name='Janis', Customer_Id='223', Open_Date='2017-08-25',
Cosult_Dt='1912-07-18', Vaccination_Id='1032', Dr_Name='Goldie',
State='Kawasaki', Country='United States', DOB='1977-12-21', IS_Active='A ',
Pincode='SL'),
 Row(Name='Cherilyn', Customer_Id='381', Open_Date='1902-03-13',
Cosult_Dt='1955-01-12', Vaccination_Id='1190', Dr_Name='Regina', State='Monaco',
Country='United States', DOB='1972-10-20', IS_Active='A ', Pincode='CC'),
 Row(Name='Tonia', Customer_Id='505', Open_Date='1980-11-30',
Cosult_Dt='1989-01-15', Vaccination_Id='1314', Dr_Name='Sallie', State='Muscat',
Country='United States', DOB='1972-06-08', IS_Active='A ', Pincode='BR'),
 Row(Name='Jorry', Customer_Id='872', Open_Date='1941-02-14',
Cosult_Dt='1951-07-16', Vaccination_Id='1681', Dr_Name='Starla', State='Dallas',
Country='United States', DOB='1985-01-04', IS_Active='A ', Pincode='TK')]
```

[28]:
```
df2.show()
```

```
+--------+-----------+----------+----------+--------------+-------+--------+----
---------+----------+---------+-------+
|    Name|Customer_Id| Open_Date| Cosult_Dt|Vaccination_Id|Dr_Name|   State|
Country|       DOB|IS_Active|Pincode|
+--------+-----------+----------+----------+--------------+-------+--------+----
---------+----------+---------+-------+
|   Janis|        223|2017-08-25|1912-07-18|          1032|
Goldie|Kawasaki|United States|1977-12-21|       A |     SL|
|Cherilyn|        381|1902-03-13|1955-01-12|          1190| Regina|
Monaco|United States|1972-10-20|       A |     CC|
|   Tonia|        505|1980-11-30|1989-01-15|          1314| Sallie|
Muscat|United States|1972-06-08|       A |     BR|
|   Jorry|        872|1941-02-14|1951-07-16|          1681| Starla|
Dallas|United States|1985-01-04|       A |     TK|
+--------+-----------+----------+----------+--------------+-------+--------+----
---------+----------+---------+-------+
```

[29]:

```
df3 = spark.sql("SELECT * from table WHERE Country=='United Kingdom'")
df4 = spark.sql("SELECT * from table WHERE Country=='Germany'")
df5 = spark.sql("SELECT * from table WHERE Country=='India'")
```

[30]:

```
df3.collect()
df4.collect()
df5.collect()
```

[30]:

```
[Row(Name='Shannah', Customer_Id='70', Open_Date='1920-08-20',
Cosult_Dt='1968-05-17', Vaccination_Id='879', Dr_Name='Rubie', State='Murmansk',
Country='India', DOB='1988-08-26', IS_Active='A ', Pincode='SC'),
 Row(Name='Gavrielle', Customer_Id='160', Open_Date='1964-03-02',
Cosult_Dt='2010-03-13', Vaccination_Id='969', Dr_Name='Mignon', State='Málaga',
Country='India', DOB='1976-03-07', IS_Active='A ', Pincode='PM'),
 Row(Name='Imojean', Customer_Id='459', Open_Date='1983-08-21',
Cosult_Dt='1967-11-25', Vaccination_Id='1268', Dr_Name='Gabi', State='Port
Vila', Country='India', DOB='1986-04-22', IS_Active='A ', Pincode='CI'),
 Row(Name='Tami', Customer_Id='964', Open_Date='1910-11-07',
Cosult_Dt='1974-11-27', Vaccination_Id='1773', Dr_Name='Kimberley',
State='Toronto', Country='India', DOB='1974-12-26', IS_Active='A ',
Pincode='DO')]
```

[31]:

```
df5.show() # India
```

```
+---------+-----------+----------+----------+--------------+---------+---------+
-------+----------+---------+-------+
|     Name|Customer_Id| Open_Date| Cosult_Dt|Vaccination_Id|  Dr_Name|
State|Country|       DOB|IS_Active|Pincode|
+---------+-----------+----------+----------+--------------+---------+---------+
-------+----------+---------+-------+
|  Shannah|         70|1920-08-20|1968-05-17|           879|    Rubie| Murmansk|
India|1988-08-26|       A |     SC|
|Gavrielle|        160|1964-03-02|2010-03-13|           969|   Mignon|   Málaga|
India|1976-03-07|       A |     PM|
|  Imojean|        459|1983-08-21|1967-11-25|          1268|     Gabi|Port Vila|
India|1986-04-22|       A |     CI|
|     Tami|        964|1910-11-07|1974-11-27|          1773|Kimberley|  Toronto|
India|1974-12-26|       A |     DO|
+---------+-----------+----------+----------+--------------+---------+---------+
-------+----------+---------+-------+
```

[32]:

```
df4.show() # Germany
```

```
+---------+-----------+----------+----------+--------------+-------+----------+-
------+----------+---------+-------+
|     Name|Customer_Id| Open_Date| Cosult_Dt|Vaccination_Id|Dr_Name|
State|Country|       DOB|IS_Active|Pincode|
+---------+-----------+----------+----------+--------------+-------+----------+-
------+----------+---------+-------+
|   Carree|        121|1949-07-04|1940-10-16|           930|   Aili|
Leeds|Germany|1974-02-07|       A |     TN|
|Kara-Lynn|        332|1951-02-04|1966-11-12|          1141| Arlena|
Kobe|Germany|1989-09-28|       A |     SG|
|      Ana|        336|1994-10-02|1955-11-22|          1145|
Xylina|Bratislava|Germany|1984-05-24|       A |     JP|
|  Delilah|        413|1974-09-02|2013-09-01|          1222|    Ida|
Apia|Germany|1978-07-29|       A |     HT|
|  Damaris|        679|1909-06-30|1930-04-27|          1488| Roxane|
Patna|Germany|1982-07-29|       A |     CA|
+---------+-----------+----------+----------+--------------+-------+----------+-
------+----------+---------+-------+
```

I have generated a data online using the following link in the csv format. https://extendsclass.com/csv-generator.html

[ ]:

[ ]: