ASSIGNMENT 5.3

on

Apache Spark

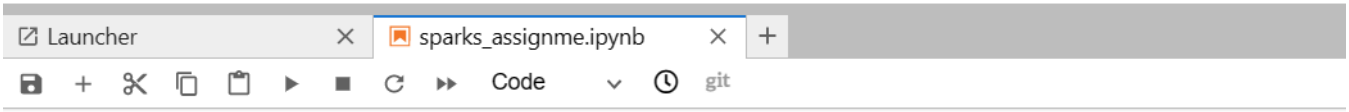
Submitted by:

Haseebullah Shaikh (2303.KHI.DEG.015)

and

Faiza Gulzar Ahmed (2303.khi.deg.001)

Solution:



Importing required libraries

[1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType
```

[2]:
```
scSpark = SparkSession.builder.appName("Spark Assignment 1").getOrCreate()
```

Defining schema and loading titanic dataset in a dataframe, also giving header

[3]:
```
file = 'titanic.csv'
header = StructType([
    StructField("PassengerId", IntegerType(), nullable=True),
    StructField("Survived", IntegerType(), nullable=True),
    StructField("Pclass", IntegerType(), nullable=True),
    StructField("Name", StringType(), nullable=True),
    StructField("Sex", StringType(), nullable=True),
    StructField("Age", DoubleType(), nullable=True),
    StructField("SibSp", IntegerType(), nullable=True),
    StructField("Parch", IntegerType(), nullable=True),
    StructField("Ticket", StringType(), nullable=True),
    StructField("Fare", DoubleType(), nullable=True),
    StructField("Cabin", StringType(), nullable=True),
    StructField("Embarked", StringType(), nullable=True),
    StructField("Timestamp", StringType(), nullable=True)
])
df_titanic = scSpark.read.csv(file, header= True, schema=header,inferSchema=True)

df_titanic
```

[3]:
```
DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Sex: string, Age: double, SibSp: int, Parch: int, Ticket: string,
Fare: double, Cabin: string, Embarked: string, Timestamp: string]
```

[4]:
```
df_titanic.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Timestamp|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-------------------+
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|  C85|       C|2020-01-01 13:44:48|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925| null|       S|2020-01-01 13:38:11|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1| C123|       S|2020-01-01 13:32:00|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05| null|       S|2020-01-01 13:36:30|
|          6|       0|     3|    Moran, Mr. James|  male|null|    0|    0|          330877| 8.4583| null|       Q|2020-01-01 13:31:39|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|  E46|       S|2020-01-01 13:37:31|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075| null|       S|2020-01-01 13:49:08|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333| null|       S|2020-01-01 13:33:42|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708| null|       C|2020-01-01 13:32:53|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|   G6|       S|2020-01-01 13:32:23|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55| C103|       S|2020-01-01 13:30:12|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05| null|       S|2020-01-01 13:33:34|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275| null|       S|2020-01-01 13:30:20|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542| null|       S|2020-01-01 13:41:17|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0| null|       S|2020-01-01 13:34:22|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125| null|       Q|2020-01-01 13:41:55|
|         18|       1|     2|Williams, Mr. Cha...|  male|null|    0|    0|          244373|   13.0| null|       S|2020-01-01 13:39:35|
```

Droping null value for avoiding nonetype error in UDF for categorical data

[5]:
```
df_titanic = df_titanic.na.drop()
```

Defining function for distributing columns according to their datatype and returning results in to a dictionary

[6]:
```
def preprocess_fun(dataframe):

    numeric_columns = []
    categorical_columns = []
    others = []

    column_types = dataframe.dtypes
    for column_name, column_type in column_types:
        if column_type == 'int' or column_type == 'float' or column_type == 'double':
            numeric_columns.append(column_name)
        elif column_type == 'string':
            categorical_columns.append(column_name)
        else:
            others.append(column_name)

    column_types_dic = {
        "numeric_columns" : numeric_columns,
        "categorical_columns" : categorical_columns,
        "others" : others
    }

    return column_types_dic
```

[7]:
```
column_types_dict = preprocess_fun(df_titanic)
```

Calculating min, max, average of numeric columns

[8]:
```
min_values = df_titanic.select([min(column).alias(column) for column in column_types_dict['numeric_columns']])
min_values.show()
```

```
+-----------+--------+------+---+-----+-----+----+
|PassengerId|Survived|Pclass|Age|SibSp|Parch|Fare|
+-----------+--------+------+---+-----+-----+----+
|          2|       0|     1|0.0|    0|    0| 0.0|
+-----------+--------+------+---+-----+-----+----+
```

[9]:
```
max_values = df_titanic.select([max(column).alias(column) for column in column_types_dict['numeric_columns']])
max_values.show()
```

```
+-----------+--------+------+----+-----+-----+--------+
|PassengerId|Survived|Pclass| Age|SibSp|Parch|    Fare|
+-----------+--------+------+----+-----+-----+--------+
|        890|       1|     3|80.0|    3|    4|512.3292|
+-----------+--------+------+----+-----+-----+--------+
```

[10]:
```
avg_values = df_titanic.select([avg(column).alias(column) for column in column_types_dict['numeric_columns']])
avg_values.show()
```

```
+-----------------+-------------------+------------------+-----------------+------------------+-------------------+-----------------+
|      PassengerId|           Survived|            Pclass|              Age|             SibSp|              Parch|             Fare|
+-----------------+-------------------+------------------+-----------------+------------------+-------------------+-----------------+
|455.3661202185792|0.6721311475409836|1.1912568306010929|35.66120218579235|0.4644808743169399|0.47540983606557374|78.68246885245901|
+-----------------+-------------------+------------------+-----------------+------------------+-------------------+-----------------+
```

defining function for categorical columns to replace last letter of word with 1

[11]:
```
def categorical_udf(words):
    text = ''
    for word in words.split():
        text = text + word[:-1] + '1' + " "
    return text.strip()

categorical_udf = udf(categorical_udf, StringType())
```

[12]:
```
for column in column_types_dict["categorical_columns"]:
    df_titanic = df_titanic.withColumn(column+"_changed", categorical_udf(col(column)))
```

[13]:
```
df_titanic.columns
```

[13]:
```
['PassengerId',
 'Survived',
 'Pclass',
 'Name',
 'Sex',
 'Age',
 'SibSp',
 'Parch',
 'Ticket',
 'Fare',
 'Cabin',
 'Embarked',
 'Timestamp',
 'Name_changed',
 'Sex_changed',
 'Ticket_changed',
 'Cabin_changed',
 'Embarked_changed',
 'Timestamp_changed']
```

[14]:
```
df_titanic.select('Name_changed', 'Sex_changed', 'Ticket_changed', 'Cabin_changed','Embarked_changed','Timestamp_changed').show()
```

```
+--------------------+-----------+--------------+-------------+---------------+-------------------+
|        Name_changed|Sex_changed|Ticket_changed|Cabin_changed|Embarked_changed|  Timestamp_changed|
+--------------------+-----------+--------------+-------------+---------------+-------------------+
|Cumings1 Mrs1 Joh...|     femal1|      P1 17591|          C81|              1|2020-01-01 13:44:41|
|Futrelle1 Mrs1 Ja...|     femal1|        113801|         C121|              1|2020-01-01 13:32:01|
|McCarthy1 Mr1 Tim...|       mal1|         17461|          E41|              1|2020-01-01 13:37:31|
|Sandstrom1 Miss1 ...|     femal1|       P1 9541|           G1|              1|2020-01-01 13:32:21|
|Bonnell1 Miss1 El...|     femal1|        113781|         C101|              1|2020-01-01 13:30:11|
|Beesley1 Mr1 Lawr...|       mal1|        248691|          D51|              1|2020-01-01 13:33:31|
|Sloper1 Mr1 Willi...|       mal1|        113781|           A1|              1|2020-01-01 13:49:11|
|Fortune1 Mr1 Char...|       mal1|         19951|  C21 C21 C21|              1|2020-01-01 13:32:41|
|Harper1 Mrs1 Henr...|     femal1|      P1 17571|          D31|              1|2020-01-01 13:48:41|
```

The End ☺