Name Rahima Siddiqui(2303.KHI.DEG.030))
Peer Name: M Humza Moeen(2303.KHI.DEG.019)
Peer Name: Osama Abdul Razzak(2303.KHI.DEG.029)

Assignment 5.3

Read data from source to DataFrame in local Spark setup and display DataFrame schema.

`tasks/4_data_pipelines/day_3_spark/data_assignment`

For numerical columns, calculate minimum, maximum and average values.

For categorical columns, create and apply UDF that will change the last letter of every word to “1”.

[33]:
```
# Import the required libraries
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col,udf
from pyspark.sql.functions import min, max, avg
from pyspark.sql.functions import udf, col
from pyspark.sql.types import StringType
```

[34]:
```
# Create the app name 'PySpark Assignment'
spark = SparkSession.builder.appName("PySpark Assignment").getOrCreate()
```

[39]:
```
# checking the given csv and found there is unnamed columns
titanic = spark.read.option("header", "false").option("inferSchema", "true").csv("titanic.csv")
titanic.show(5)
titanic.printSchema()
```

```
+---+---+---+--------------------+------+---+---+---+----------------+-------+----+----+-------------------+
|_c0|_c1|_c2|                 _c3|   _c4|_c5|_c6|_c7|             _c8|    _c9|_c10|_c11|               _c12|
+---+---+---+--------------------+------+---+---+---+----------------+-------+----+----+-------------------+
|  1|  0|  3|Braund, Mr. Owen ...|  male| 22|  1|  0|       A/5 21171|   7.25|null|   S|2020-01-01 13:45:25|
|  2|  1|  1|Cumings, Mrs. Joh...|female| 38|  1|  0|        PC 17599|71.2833| C85|   C|2020-01-01 13:44:48|
|  3|  1|  3|Heikkinen, Miss. ...|female| 26|  0|  0|STON/O2. 3101282|  7.925|null|   S|2020-01-01 13:38:11|
|  4|  1|  1|Futrelle, Mrs. Ja...|female| 35|  1|  0|          113803|   53.1|C123|   S|2020-01-01 13:32:00|
|  5|  0|  3|Allen, Mr. Willia...|  male| 35|  0|  0|          373450|   8.05|null|   S|2020-01-01 13:36:30|
+---+---+---+--------------------+------+---+---+---+----------------+-------+----+----+-------------------+
only showing top 5 rows

root
 |-- _c0: integer (nullable = true)
 |-- _c1: integer (nullable = true)
 |-- _c2: integer (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: integer (nullable = true)
 |-- _c6: integer (nullable = true)
 |-- _c7: integer (nullable = true)
 |-- _c8: string (nullable = true)
 |-- _c9: double (nullable = true)
 |-- _c10: string (nullable = true)
 |-- _c11: string (nullable = true)
 |-- _c12: timestamp (nullable = true)
```

Name Rahima Siddiqui(2303.KHI.DEG.030))
Peer Name: M Humza Moeen(2303.KHI.DEG.019)
Peer Name: Osama Abdul Razzak(2303.KHI.DEG.029)

[40]:
```
#then naming the columns according to the titanic datasets, which we download from kaggle
columns = ["PassengerId","Survived","Pclass","Name","Sex","Age","SibSp","Parch","Ticket","Fare","Cabin","Embarked","Timestamp"]
titanic = titanic.toDF(*columns)
titanic.printSchema()
```

```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
 |-- Timestamp: timestamp (nullable = true)
```

[5]:
```
titanic.show(5)
```

```
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+-------------------+
|PassengerId|Survived|Pclass|                Name|   Sex|Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Timestamp|
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+-------------------+
|          1|       0|     3|Braund, Mr. Owen ...|  male| 22|    1|    0|       A/5 21171|   7.25| null|       S|2020-01-01 13:45:25|
|          2|       1|     1|Cumings, Mrs. Joh...|female| 38|    1|    0|        PC 17599|71.2833|  C85|       C|2020-01-01 13:44:48|
|          3|       1|     3|Heikkinen, Miss. ...|female| 26|    0|    0|STON/O2. 3101282|  7.925| null|       S|2020-01-01 13:38:11|
|          4|       1|     1|Futrelle, Mrs. Ja...|female| 35|    1|    0|          113803|   53.1| C123|       S|2020-01-01 13:32:00|
|          5|       0|     3|Allen, Mr. Willia...|  male| 35|    0|    0|          373450|   8.05| null|       S|2020-01-01 13:36:30|
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+-------------------+
only showing top 5 rows
```

[41]:
```
# After that change the survived result from boolean type 0 or 1 into yes or no string type
titanic = titanic.withColumn("Survived", when(titanic["Survived"] == 0, "No").otherwise("Yes"))
titanic.show(5)
```

```
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+-------------------+
|PassengerId|Survived|Pclass|                Name|   Sex|Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Timestamp|
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+-------------------+
|          1|      No|     3|Braund, Mr. Owen ...|  male| 22|    1|    0|       A/5 21171|   7.25| null|       S|2020-01-01 13:45:25|
|          2|     Yes|     1|Cumings, Mrs. Joh...|female| 38|    1|    0|        PC 17599|71.2833|  C85|       C|2020-01-01 13:44:48|
|          3|     Yes|     3|Heikkinen, Miss. ...|female| 26|    0|    0|STON/O2. 3101282|  7.925| null|       S|2020-01-01 13:38:11|
|          4|     Yes|     1|Futrelle, Mrs. Ja...|female| 35|    1|    0|          113803|   53.1| C123|       S|2020-01-01 13:32:00|
|          5|      No|     3|Allen, Mr. Willia...|  male| 35|    0|    0|          373450|   8.05| null|       S|2020-01-01 13:36:30|
+-----------+--------+------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+-------------------+
only showing top 5 rows
```

Name Rahima Siddiqui(2303.KHI.DEG.030))
Peer Name: M Humza Moeen(2303.KHI.DEG.019)
Peer Name: Osama Abdul Razzak(2303.KHI.DEG.029)

[12]:
```
#store the numerical col in numerical df
numerical_cols = [col_name for col_name, col_type in titanic.dtypes if any(col_type.startswith(t) for t in ['int', 'bigint', 'float', 'double']]
numerical_df = titanic.select(*[col(col_name) for col_name in numerical_cols])
numerical_df.show(5)
```

```
+-----------+------+---+-----+-----+-------+
|PassengerId|Pclass|Age|SibSp|Parch|   Fare|
+-----------+------+---+-----+-----+-------+
|          1|     3| 22|    1|    0|   7.25|
|          2|     1| 38|    1|    0|71.2833|
|          3|     3| 26|    0|    0|  7.925|
|          4|     1| 35|    1|    0|   53.1|
|          5|     3| 35|    0|    0|   8.05|
+-----------+------+---+-----+-----+-------+
only showing top 5 rows
```

[42]:
```
# Compute minimum, maximum, and mean values for numerical columns
min_max_mean = numerical_df.describe()
statistics = min_max_mean.select(numerical_cols).summary("min", "max", "mean")
statistics.show()
```

```
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|summary|       PassengerId|            Pclass|               Age|             SibSp|             Parch|              Fare|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|    min|                 1|0.8360712409770491|                 0|                 0|                 0|               0.0|
|    max|               891|               891|                80|               891|               891|               891|
|   mean|497.27076840304596|179.62894264325715|167.64315089562422|180.12515025772694|179.6375301872114|297.04536731315113|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
```

[43]:
```
# Store the categorical data into cat_df
cat_cols = [col_name for col_name, col_type in titanic.dtypes if any(col_type.startswith(t) for t in ['str'])]
cat_df = titanic.select(*[col(col_name) for col_name in cat_cols])
cat_df.show(5)
```

```
+--------+--------------------+------+----------------+-----+--------+
|Survived|                Name|   Sex|          Ticket|Cabin|Embarked|
+--------+--------------------+------+----------------+-----+--------+
|      No|Braund, Mr. Owen ...|  male|       A/5 21171| null|       S|
|     Yes|Cumings, Mrs. Joh...|female|        PC 17599|  C85|       C|
|     Yes|Heikkinen, Miss. ...|female|STON/O2. 3101282| null|       S|
|     Yes|Futrelle, Mrs. Ja...|female|          113803| C123|       S|
|      No|Allen, Mr. Willia...|  male|          373450| null|       S|
+--------+--------------------+------+----------------+-----+--------+
only showing top 5 rows
```

[31]:
```
# Define UDF to change last letter of each word to "1"
def change_last_letter(word):
    if word is not None:
        words = word.split()
        for i in range(len(words)):
            words[i] = words[i][:-1] + "1"
        return " ".join(words)
    return word

change_last_letter_udf = udf(change_last_letter, StringType())
for column in cat_df.columns:
    cat_df = cat_df.withColumn(column, change_last_letter_udf(col(column)))
cat_df.show(5)
```

```
+--------+--------------------+------+----------------+-----+--------+
|Survived|                Name|   Sex|          Ticket|Cabin|Embarked|
+--------+--------------------+------+----------------+-----+--------+
|      N1|Braund1 Mr1 Owe1 ...|  mal1|       A/1 21171| null|       1|
|     Ye1|Cumings1 Mrs1 Joh...|fema11|        P1 17591|  C81|       1|
|     Ye1|Heikkinen1 Miss1 ...|fema11|STON/O21 3101281| null|       1|
|     Ye1|Futrelle1 Mrs1 Ja...|fema11|          113801| C121|       1|
|      N1|Allen1 Mr1 Willia...|  mal1|          373451| null|       1|
+--------+--------------------+------+----------------+-----+--------+
only showing top 5 rows
```

Name Rahima Siddiqui(2303.KHI.DEG.030))
Peer Name: M Humza Moeen(2303.KHI.DEG.019)
Peer Name: Osama Abdul Razzak(2303.KHI.DEG.029)

In the end, sort the data overwrite it to existing data and save it into .parquet form

```
sorted_data= titanic.orderBy(titanic.columns[0])
sorted_data.show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+----------
----+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|          Times
tamp|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+----------
----+
|          2|     Yes|     1|Cumings, Mrs. Joh...|female|  38|    1|    0|        PC 17599|71.2833|  C85|       C|2020-01-01 13:4
4:48|
|          3|     Yes|     3|Heikkinen, Miss. ...|female|  26|    0|    0|STON/O2. 3101282|  7.925| null|       S|2020-01-01 13:3
8:11|
|          4|     Yes|     1|Futrelle, Mrs. Ja...|female|  35|    1|    0|          113803|   53.1| C123|       S|2020-01-01 13:3
2:00|
|          5|      No|     3|Allen, Mr. Willia...|  male|  35|    0|    0|          373450|   8.05| null|       S|2020-01-01 13:3
6:30|
|          6|      No|     3|    Moran, Mr. James|  male|null|    0|    0|          330877| 8.4583| null|       Q|2020-01-01 13:3
1:39|
|          7|      No|     1|McCarthy, Mr. Tim...|  male|  54|    0|    0|           17463|51.8625|  E46|       S|2020-01-01 13:3
7:31|
```

```
]: try:
       sorted_data.write.mode('overwrite').parquet("titanic_results.parquet")
   except:
       print('Expection caught')
```