1)**Reading CSV /Parquet/SAS files:-**

Reading the files and storing into dataframes

Pandas: df = pd.read_csv(file_path)

df = pd.read_parquet(file_path)

df = pd.read_sas(file_path,encoding=False)

Pyspark:

df = spark.read.csv(file_path) or

df = spark.read.option('inferSchema','true').csv(file_path)

df = spark.read.parquet(file_path) or

df = spark.read.option('inferSchema','true').parquet(file_path)

2)Checking columns:-

Pandas: df.columns() or df.shape[1]

Pyspark: df.select("*").show()

3)Checking the number of records:-

Pandas : len(df.index) or df.shape[0]

Pyspark: df.count()

4)Checking the schema/datatype of the table :-

Pandas: df.info()

Pyspark: df.printSchema() or df_basket1.dtypes

5)Finding the sum of specific columns:-

Pandas : df['PRICE'].sum(axis=1)

Pyspark : df.agg({'PRICE':'sum'}).show()

6)Finding first and last 3 rows:-

Pandas : df.head(3) or df.tail(3)

Pyspark: df.head(3) or df.tail(3)

To compare:-

Read python file PY = pd.read_csv(file_path)

Read SAS file SS = pd.read_sas(file_path)

Perform the above operation on these dataframes

For eg:

PY.info()

SS.info()

It will display the schema of python and sas