**Lab 5: Spark Diff**

**Pandas Code**

|  |
| --- |
| import pandas as pd  pre\_df = pd.read\_csv('pre.csv')  post\_df = pd.read\_csv('post.csv')  keys = ['InvoiceNo', 'StockCode', 'CustomerID', 'InvoiceDate']  pre\_df = pre\_df.drop\_duplicates(subset=keys)  post\_df = post\_df.drop\_duplicates(subset=keys)  merge\_df = pre\_df.merge(post\_df, on=keys, suffixes=('\_pre', '\_post'), indicator=True)  diff\_df = merge\_df[merge\_df['\_merge'] == 'both']  diff\_df['Quantity\_diff'] = diff\_df['Quantity\_post'] - diff\_df['Quantity\_pre']  diff\_df['UnitPrice\_diff'] = diff\_df['UnitPrice\_post'] - diff\_df['UnitPrice\_pre'] |

**Spark Code**

|  |
| --- |
| from pyspark.sql import SparkSession  from pyspark.sql.functions import col  # Create a Spark session  spark = SparkSession.builder.getOrCreate()  # Load the data  pre\_df = spark.read.csv('pre.csv', header=True, inferSchema=True)  post\_df = spark.read.csv('post.csv', header=True, inferSchema=True)  keys = ['InvoiceNo', 'StockCode', 'CustomerID', 'InvoiceDate']  # Drop duplicates  pre\_df = pre\_df.dropDuplicates(subset=keys)  post\_df = post\_df.dropDuplicates(subset=keys)  # Rename the columns in pre\_df  pre\_df = pre\_df.select(\*[col(c).alias(c+'\_pre') if c not in keys else c for c in pre\_df.columns])  # Rename the columns in post\_df  post\_df = post\_df.select(\*[col(c).alias(c+'\_post') if c not in keys else c for c in post\_df.columns])  # Merge pre and post dataframes  merge\_df = pre\_df.join(post\_df, on=keys, how='inner')  # Calculate differences  diff\_df = merge\_df.withColumn('Quantity\_diff', merge\_df.Quantity\_post - merge\_df.Quantity\_pre) \  .withColumn('UnitPrice\_diff', merge\_df.UnitPrice\_post - merge\_df.UnitPrice\_pre) |

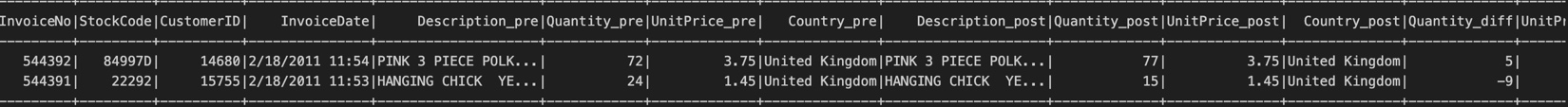
**Pandas Output 1**

diff\_df[diff\_df['Quantity\_diff'] != 0]



**Pyspark Output 1**

diff\_df.filter(diff\_df.Quantity\_diff != 0).show()



**Pandas Output 2**

diff\_df[diff\_df['UnitPrice\_diff'] != 0]



**Pyspark Output 2**

diff\_df.filter(diff\_df.UnitPrice\_diff != 0).show()

