Code logic for Retail case study assignment

1. The project is to design an order intelligence system which reads invoice data from kafka server and generates KPIs based on the data.
2. We need to create an EMR cluster with Spark installed along with other required services and SSH into the machine.
3. Run below commands for integration with Kafka and to read the sales data from the Kafka server
   export SPARK_KAFKA_VERSION=0.10
   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 **spark-streaming.py**
4. Storing summarised input table from various batches using spark submit command into a file
   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 **spark-streaming.py > Console-output**
5. Inside the pySpark file, preprocessing of the data is done to calculate additional derived columns such as total_cost, total_items,is_order,is_return for every invoice for each 1 minute window.
6. Calculating the time-based KPIs such as total_sales_volume, orders per minute, rate of return and average_transaction_size for a tumbling window of 1 minute.
7. Calculating time and country-based KPIs such as total_sales_volume, orders per minute, rate of return for a **tumbling window of one minute on orders on a per-country basis**
8. Storing the KPIs (both time-based and time- and country-based) in separate json files which can be used for further analysis.

PS: For a long time, facing issue with provisioning EMR cluster and downloading files from cluster. Hence I am unable to attach console output and json files. Attaching only screenshot of all files generated in master node.