RETAIL GROCERY ANALYSIS

Documented By

Aruna Balamurugan

F07064

RETAIL GROCERY ANALYSIS

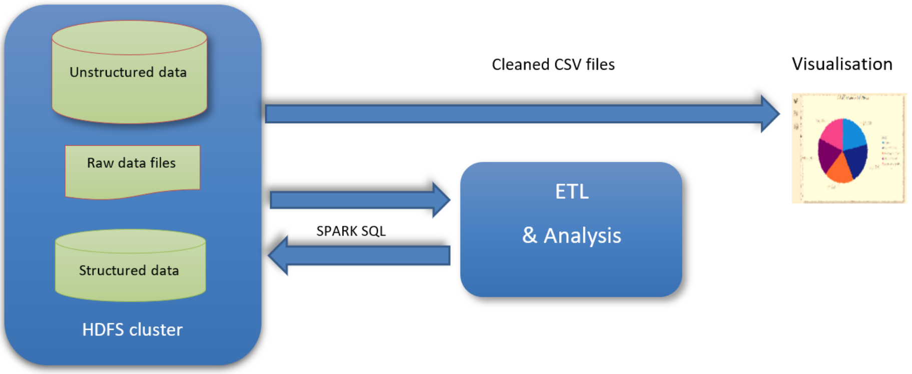
Objective:

- To perform analysis on a retail grocery company.
- To extract and build ETL pipeline and produce aggregated data into data lake.
- To analyse and provide business insights on improving sales volume and customer retention.

Solution:

- Extract unstructured raw data from HDFS located in: /user/insofe/retail/data/aws
- Use SparkSQL to read CSV file by specifying path location.
- Create schema and dataframes for each schema.
- Check for null values and redundancies.
- Merge dataframes and perform aggregations. One final dataframe for cleaned data.
- Remove duplicate columns
- Write dataframes into CSV file and load it back on HDFS. ETL pipelining is now achieved.
- Load the output CSV file into local machine and import to Power BI for visualisation.
- Remove duplicate rows and columns, null values, perform necessary transformation for analysis.
- Analyse for product-order insights by creating visuals.

Solution Architecture:

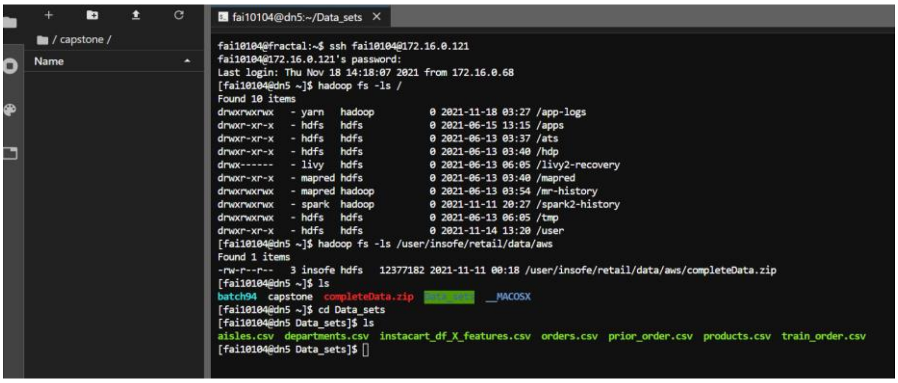


Steps:

1. Open Terminal in JupyterLab

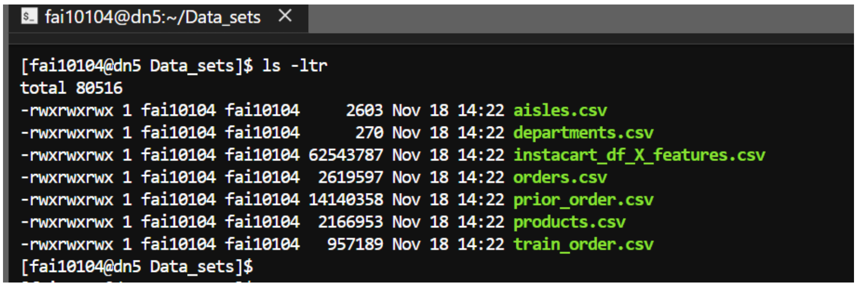
2. Connect to Hadoop by using ssh connection:
    a. Provide IP address and password.

    b. Connect to HDFS by using hadoop fs -ls /path location



```
fai10104@fractal:~$ ssh fai10104@172.16.0.121
fai10104@172.16.0.121's password:
Last login: Thu Nov 18 14:18:07 2021 from 172.16.0.68
[fai10104@dn5 ~]$ hadoop fs -ls /
Found 10 items
drwxrwxrwx   - yarn   hadoop          0 2021-11-18 03:27 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2021-06-15 13:15 /apps
drwxr-xr-x   - hdfs   hdfs            0 2021-06-13 03:37 /ats
drwxr-xr-x   - hdfs   hdfs            0 2021-06-13 03:40 /hdp
drwx------   - livy   hdfs            0 2021-06-13 06:05 /livy2-recovery
drwxr-xr-x   - mapred hdfs            0 2021-06-13 03:40 /mapred
drwxrwxrwx   - mapred hadoop          0 2021-06-13 03:54 /mr-history
drwxrwxrwx   - spark  hadoop          0 2021-11-11 20:27 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2021-06-13 06:05 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2021-11-14 13:20 /user
[fai10104@dn5 ~]$ hadoop fs -ls /user/insofe/retail/data/aws
Found 1 items
-rw-r--r--   3 insofe hdfs   12377182 2021-11-11 00:18 /user/insofe/retail/data/aws/completeData.zip
[fai10104@dn5 ~]$ ls
batch94  capstone  completeData.zip  Data_sets  __MACOSX
[fai10104@dn5 ~]$ cd Data_sets
[fai10104@dn5 Data_sets]$ ls
aisles.csv  departments.csv  instacart_df_X_features.csv  orders.csv  prior_order.csv  products.csv  train_order.csv
[fai10104@dn5 Data_sets]$ 
```

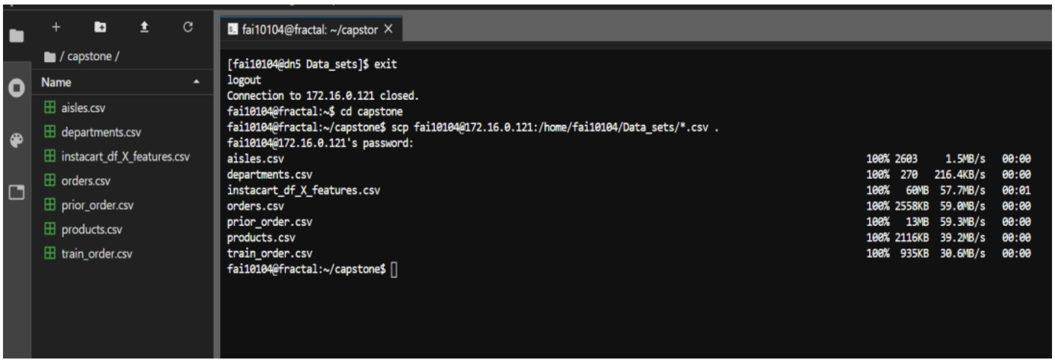
    c. Use ls -ltr for listing all files and directories in the cluser



```
[fai10104@dn5 Data_sets]$ ls -ltr
total 80516
-rwxrwxrwx 1 fai10104 fai10104     2603 Nov 18 14:22 aisles.csv
-rwxrwxrwx 1 fai10104 fai10104      270 Nov 18 14:22 departments.csv
-rwxrwxrwx 1 fai10104 fai10104 62543787 Nov 18 14:22 instacart_df_X_features.csv
-rwxrwxrwx 1 fai10104 fai10104  2619597 Nov 18 14:22 orders.csv
-rwxrwxrwx 1 fai10104 fai10104 14140358 Nov 18 14:22 prior_order.csv
-rwxrwxrwx 1 fai10104 fai10104  2166953 Nov 18 14:22 products.csv
-rwxrwxrwx 1 fai10104 fai10104   957189 Nov 18 14:22 train_order.csv
[fai10104@dn5 Data_sets]$
```

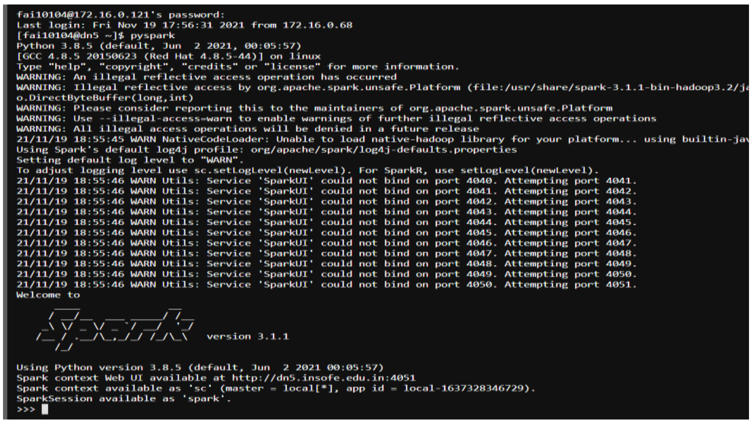
    d. Exit the connection.

    e. Use scp cluster_location local_path . for loading all CSV files into local machine (Optional)



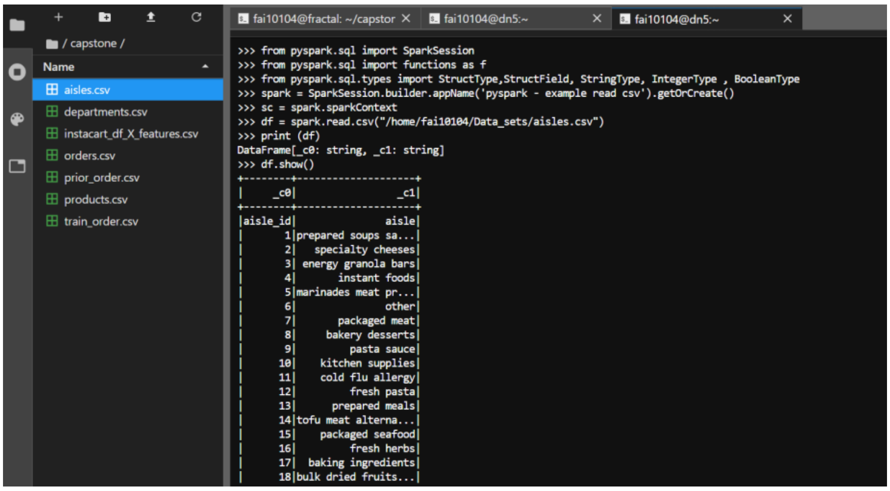
```
[fai10104@dn5 Data_sets]$ exit
logout
Connection to 172.16.0.121 closed.
fai10104@fractal:~$ cd capstone
fai10104@fractal:~/capstone$ scp fai10104@172.16.0.121:/home/fai10104/Data_sets/*.csv .
fai10104@172.16.0.121's password:
aisles.csv                                          100% 2603     1.5MB/s   00:00
departments.csv                                     100%  270   216.4KB/s   00:00
instacart_df_X_features.csv                         100%   60MB  57.7MB/s   00:01
orders.csv                                          100% 2558KB  59.0MB/s   00:00
prior_order.csv                                     100%   13MB  59.3MB/s   00:00
products.csv                                        100% 2116KB  39.2MB/s   00:00
train_order.csv                                     100%  935KB  30.6MB/s   00:00
fai10104@fractal:~/capstone$ 
```

3. Open another terminal to start connection and use Pyspark.
   (Note: PySpark is used inside the connection throughout the ETL pipeline).



```
fai10104@172.16.0.121's password:
Last login: Fri Nov 19 17:56:31 2021 from 172.16.0.68
[fai10104@dn5 ~]$ pyspark
Python 3.8.5 (default, Jun  2 2021, 00:05:57)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-44)] on linux
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/usr/share/spark-3.1.1-bin-hadoop3.2/ja
o.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/11/19 18:55:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4043. Attempting port 4044.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4044. Attempting port 4045.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4045. Attempting port 4046.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4046. Attempting port 4047.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4047. Attempting port 4048.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4048. Attempting port 4049.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4049. Attempting port 4050.
21/11/19 18:55:46 WARN Utils: Service 'SparkUI' could not bind on port 4050. Attempting port 4051.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.1
      /_/

Using Python version 3.8.5 (default, Jun  2 2021 00:05:57)
Spark context Web UI available at http://dn5.insofe.edu.in:4051
Spark context available as 'sc' (master = local[*], app id = local-1637328346729).
SparkSession available as 'spark'.
>>> 
```

4. Start a spark session by importing necessary packages.
5. Use spark.read.csv for reading the loaded CSV file:



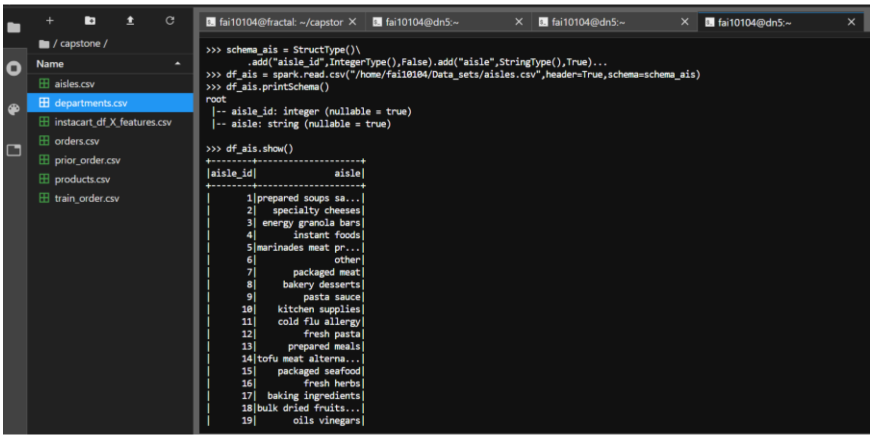
```
>>> from pyspark.sql import SparkSession
>>> from pyspark.sql import functions as f
>>> from pyspark.sql.types import StructType,StructField, StringType, IntegerType , BooleanType
>>> spark = SparkSession.builder.appName('pyspark - example read csv').getOrCreate()
>>> sc = spark.sparkContext
>>> df = spark.read.csv("/home/fai10104/Data_sets/aisles.csv")
>>> print (df)
DataFrame[_c0: string, _c1: string]
>>> df.show()
+--------+--------------------+
|     _c0|                 _c1|
+--------+--------------------+
|aisle_id|               aisle|
|       1|prepared soups sa...|
|       2|   specialty cheeses|
|       3| energy granola bars|
|       4|       instant foods|
|       5|marinades meat pr...|
|       6|               other|
|       7|       packaged meat|
|       8|     bakery desserts|
|       9|         pasta sauce|
|      10|    kitchen supplies|
|      11|    cold flu allergy|
|      12|         fresh pasta|
|      13|      prepared meals|
|      14|tofu meat alterna...|
|      15|    packaged seafood|
|      16|         fresh herbs|
|      17|  baking ingredients|
|      18|bulk dried fruits...|
```

Fig.1: df.show() is used to display data in aisle.csv

6. <u>Schema Creation:</u>

   - StructType() is used to create schema by defining fields to allow nullable(True) or not (False)
   - To convert schema to dataframe, use df=spark.read.csv("path",header=True,schema=schema)
     This dataframe df reads file located in "path", defines schema as mentioned in "schema" variable above and keeps the first row as the header.

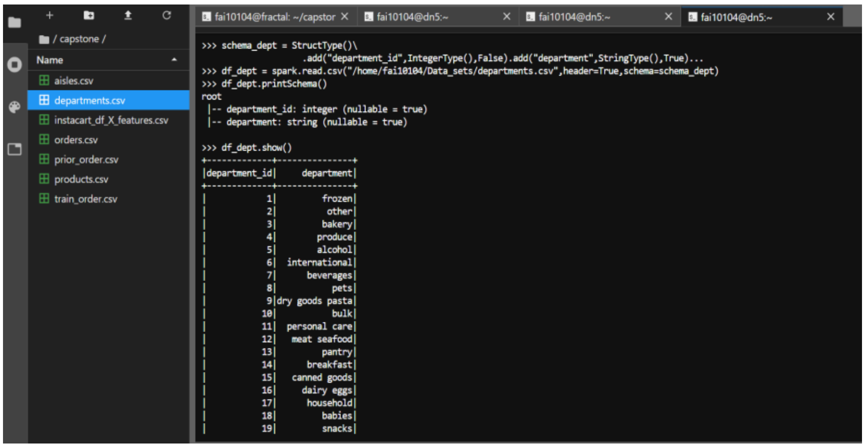
a. aisle.csv



```
>>> schema_ais = StructType()\
        .add("aisle_id",IntegerType(),False).add("aisle",StringType(),True)...
>>> df_ais = spark.read.csv("/home/fai10104/Data_sets/aisles.csv",header=True,schema=schema_ais)
>>> df_ais.printSchema()
root
 |-- aisle_id: integer (nullable = true)
 |-- aisle: string (nullable = true)

>>> df_ais.show()
+--------+--------------------+
|aisle_id|               aisle|
+--------+--------------------+
|       1|prepared soups sa...|
|       2|   specialty cheeses|
|       3| energy granola bars|
|       4|       instant foods|
|       5|marinades meat pr...|
|       6|               other|
|       7|       packaged meat|
|       8|     bakery desserts|
|       9|         pasta sauce|
|      10|    kitchen supplies|
|      11|    cold flu allergy|
|      12|         fresh pasta|
|      13|      prepared meals|
|      14|tofu meat alterna...|
|      15|    packaged seafood|
|      16|         fresh herbs|
|      17|  baking ingredients|
|      18|bulk dried fruits...|
|      19|        oils vinegars|
```

Fig.2: schema_ais = schema for aisle; df_ais = dataframe for aisle schema.

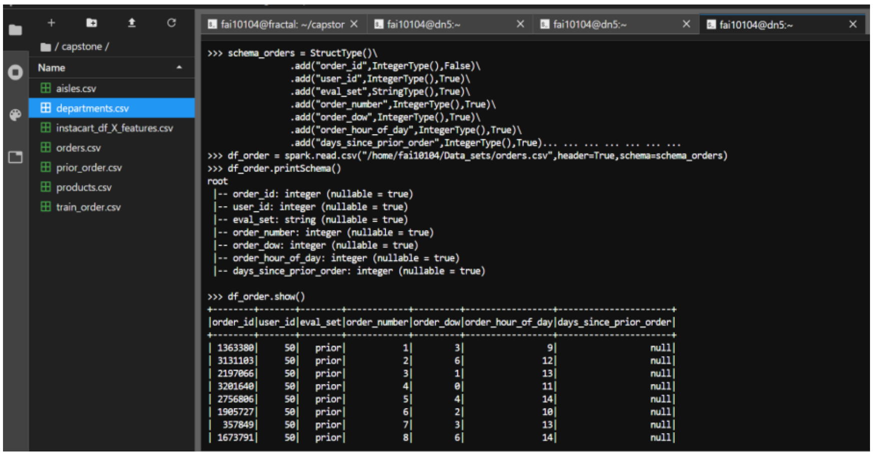
b. department.csv



```
>>> schema_dept = StructType()\
                 .add("department_id",IntegerType(),False).add("department",StringType(),True)...
>>> df_dept = spark.read.csv("/home/fai10104/Data_sets/departments.csv",header=True,schema=schema_dept)
>>> df_dept.printSchema()
root
 |-- department_id: integer (nullable = true)
 |-- department: string (nullable = true)

>>> df_dept.show()
+-------------+---------------+
|department_id|     department|
+-------------+---------------+
|            1|         frozen|
|            2|          other|
|            3|         bakery|
|            4|        produce|
|            5|        alcohol|
|            6|  international|
|            7|      beverages|
|            8|           pets|
|            9|dry goods pasta|
|           10|           bulk|
|           11|  personal care|
|           12|   meat seafood|
|           13|         pantry|
|           14|      breakfast|
|           15|   canned goods|
|           16|     dairy eggs|
|           17|      household|
|           18|         babies|
|           19|         snacks|
```

Fig.3: schema_dept = schema for dept data; df_dept = dataframe for dept schema.

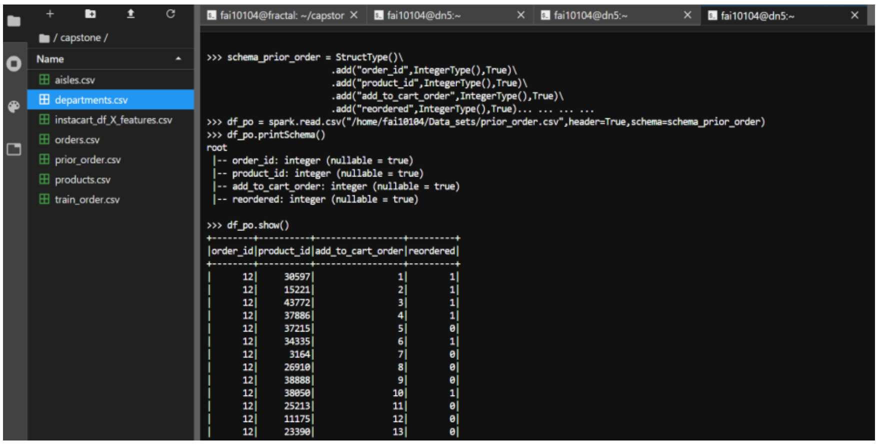
c. order.csv



```
>>> schema_orders = StructType()\
            .add("order_id",IntegerType(),False)\
            .add("user_id",IntegerType(),True)\
            .add("eval_set",StringType(),True)\
            .add("order_number",IntegerType(),True)\
            .add("order_dow",IntegerType(),True)\
            .add("order_hour_of_day",IntegerType(),True)\
            .add("days_since_prior_order",IntegerType(),True)... ... ... ... ... ... ...
>>> df_order = spark.read.csv("/home/fai10104/Data_sets/orders.csv",header=True,schema=schema_orders)
>>> df_order.printSchema()
root
 |-- order_id: integer (nullable = true)
 |-- user_id: integer (nullable = true)
 |-- eval_set: string (nullable = true)
 |-- order_number: integer (nullable = true)
 |-- order_dow: integer (nullable = true)
 |-- order_hour_of_day: integer (nullable = true)
 |-- days_since_prior_order: integer (nullable = true)

>>> df_order.show()
+--------+-------+--------+------------+---------+-----------------+----------------------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|
+--------+-------+--------+------------+---------+-----------------+----------------------+
| 1363380|     50|   prior|           1|        3|                9|                  null|
| 3131103|     50|   prior|           2|        6|               12|                  null|
| 2197066|     50|   prior|           3|        1|               13|                  null|
| 3201640|     50|   prior|           4|        0|               11|                  null|
| 2756806|     50|   prior|           5|        4|               14|                  null|
| 1905727|     50|   prior|           6|        2|               10|                  null|
|  357849|     50|   prior|           7|        3|               13|                  null|
| 1673791|     50|   prior|           8|        6|               14|                  null|
```

Fig.4: schema_orders = schema for orders data; df_order = dataframe for orders schema.

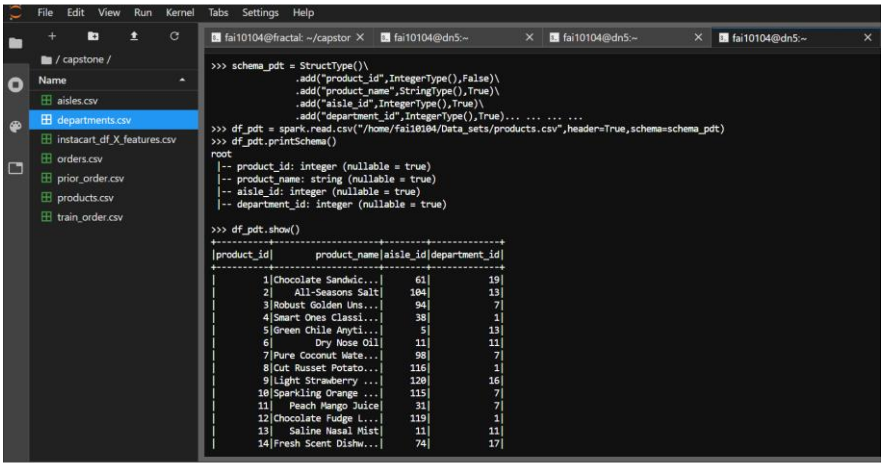
d. prior_order.csv



```
>>> schema_prior_order = StructType()\
                        .add("order_id",IntegerType(),True)\
                        .add("product_id",IntegerType(),True)\
                        .add("add_to_cart_order",IntegerType(),True)\
                        .add("reordered",IntegerType(),True)... ... ... ...
>>> df_po = spark.read.csv("/home/fai10104/Data_sets/prior_order.csv",header=True,schema=schema_prior_order)
>>> df_po.printSchema()
root
 |-- order_id: integer (nullable = true)
 |-- product_id: integer (nullable = true)
 |-- add_to_cart_order: integer (nullable = true)
 |-- reordered: integer (nullable = true)

>>> df_po.show()
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
|      12|     30597|                1|        1|
|      12|     15221|                2|        1|
|      12|     43772|                3|        1|
|      12|     37886|                4|        1|
|      12|     37215|                5|        0|
|      12|     34335|                6|        1|
|      12|      3164|                7|        0|
|      12|     26910|                8|        0|
|      12|     38888|                9|        0|
|      12|     38050|               10|        1|
|      12|     25213|               11|        0|
|      12|     11175|               12|        0|
|      12|     23390|               13|        0|
```

Fig.5: schema_prior_order = schema for prior_order data; df_po = dataframe for prior_order schema.

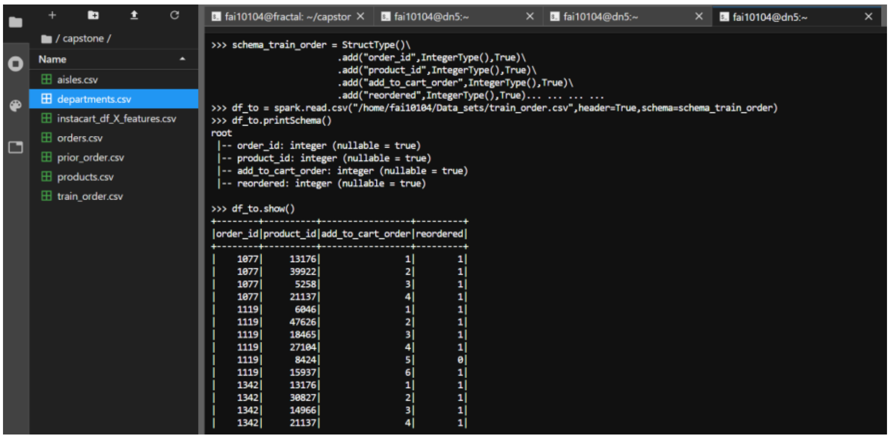
e. product.csv



```
>>> schema_pdt = StructType()\
                .add("product_id",IntegerType(),False)\
                .add("product_name",StringType(),True)\
                .add("aisle_id",IntegerType(),True)\
                .add("department_id",IntegerType(),True)... ... ... ...
>>> df_pdt = spark.read.csv("/home/fai10104/Data_sets/products.csv",header=True,schema=schema_pdt)
>>> df_pdt.printSchema()
root
 |-- product_id: integer (nullable = true)
 |-- product_name: string (nullable = true)
 |-- aisle_id: integer (nullable = true)
 |-- department_id: integer (nullable = true)

>>> df_pdt.show()
+----------+--------------------+--------+-------------+
|product_id|        product_name|aisle_id|department_id|
+----------+--------------------+--------+-------------+
|         1|Chocolate Sandwic...|      61|           19|
|         2|    All-Seasons Salt|     104|           13|
|         3|Robust Golden Uns...|      94|            7|
|         4|Smart Ones Classi...|      38|            1|
|         5|Green Chile Anyti...|       5|           13|
|         6|        Dry Nose Oil|      11|           11|
|         7|Pure Coconut Wate...|      98|            7|
|         8|Cut Russet Potato...|     116|            1|
|         9|Light Strawberry ...|     120|           16|
|        10|Sparkling Orange ...|     115|            7|
|        11|   Peach Mango Juice|      31|            7|
|        12|Chocolate Fudge L...|     119|            1|
|        13|   Saline Nasal Mist|      11|           11|
|        14|Fresh Scent Dishw...|      74|           17|
```

Fig.6: schema_pdt = schema for product data; df_pdt = dataframe for product schema.

f. train_order.csv

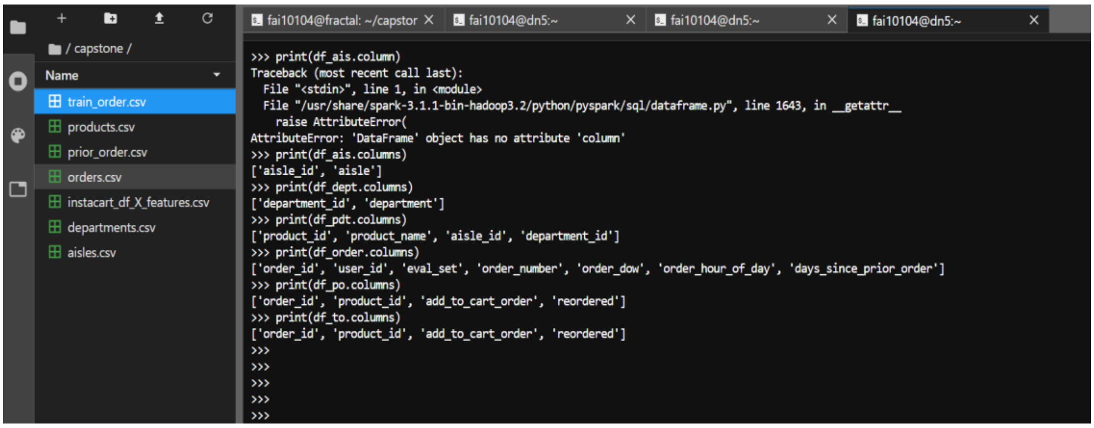


```
>>> schema_train_order = StructType()\
                        .add("order_id",IntegerType(),True)\
                        .add("product_id",IntegerType(),True)\
                        .add("add_to_cart_order",IntegerType(),True)\
                        .add("reordered",IntegerType(),True)... ... ... ...
>>> df_to = spark.read.csv("/home/fai10104/Data_sets/train_order.csv",header=True,schema=schema_train_order)
>>> df_to.printSchema()
root
 |-- order_id: integer (nullable = true)
 |-- product_id: integer (nullable = true)
 |-- add_to_cart_order: integer (nullable = true)
 |-- reordered: integer (nullable = true)

>>> df_to.show()
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
|    1077|     13176|                1|        1|
|    1077|     39922|                2|        1|
|    1077|      5258|                3|        1|
|    1077|     21137|                4|        1|
|    1119|      6046|                1|        1|
|    1119|     47626|                2|        1|
|    1119|     18465|                3|        1|
|    1119|     27104|                4|        1|
|    1119|      8424|                5|        0|
|    1119|     15937|                6|        1|
|    1342|     13176|                1|        1|
|    1342|     30827|                2|        1|
|    1342|     14966|                3|        1|
|    1342|     21137|                4|        1|
```

Fig.7: schema_train_order = schema for train_orders data; df_to = dataframe for train_order schema

7. Displaying column names:

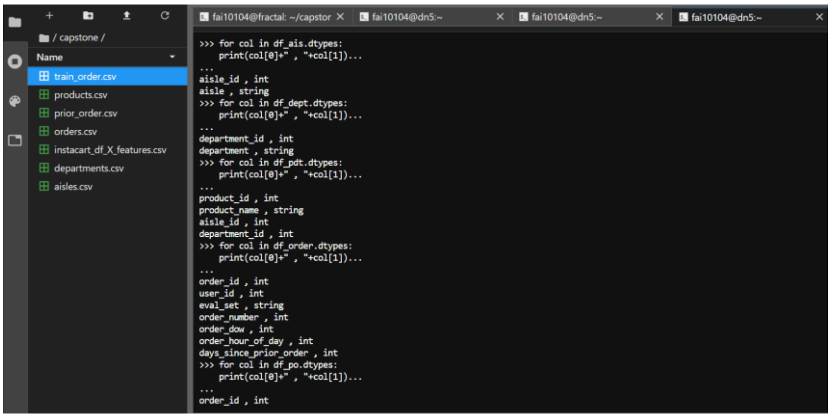
   Use print(df.columns) to print column names as a list:



```
>>> print(df_ais.column)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/share/spark-3.1.1-bin-hadoop3.2/python/pyspark/sql/dataframe.py", line 1643, in __getattr__
    raise AttributeError(
AttributeError: 'DataFrame' object has no attribute 'column'
>>> print(df_ais.columns)
['aisle_id', 'aisle']
>>> print(df_dept.columns)
['department_id', 'department']
>>> print(df_pdt.columns)
['product_id', 'product_name', 'aisle_id', 'department_id']
>>> print(df_order.columns)
['order_id', 'user_id', 'eval_set', 'order_number', 'order_dow', 'order_hour_of_day', 'days_since_prior_order']
>>> print(df_po.columns)
['order_id', 'product_id', 'add_to_cart_order', 'reordered']
>>> print(df_to.columns)
['order_id', 'product_id', 'add_to_cart_order', 'reordered']
>>>
>>>
>>>
>>>
>>>
```

8. Display the datatypes of the columns:

   Use df.dypes for diplaying data types of all columns in a dataframe.
   For loop is used to display column and its datatypes one after the other, separated by a comma.

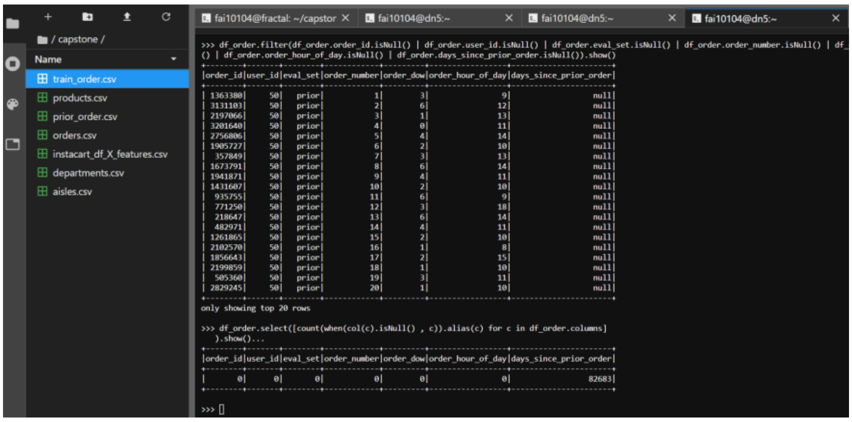


```
>>> for col in df_ais.dtypes:
    print(col[0]+" , "+col[1])...
...
aisle_id , int
aisle , string
>>> for col in df_dept.dtypes:
    print(col[0]+" , "+col[1])...
...
department_id , int
department , string
>>> for col in df_pdt.dtypes:
    print(col[0]+" , "+col[1])...
...
product_id , int
product_name , string
aisle_id , int
department_id , int
>>> for col in df_order.dtypes:
    print(col[0]+" , "+col[1])...
...
order_id , int
user_id , int
eval_set , string
order_number , int
order_dow , int
order_hour_of_day , int
days_since_prior_order , int
>>> for col in df_po.dtypes:
    print(col[0]+" , "+col[1])...
...
order_id , int
```

9. Check for Null values in the columns:

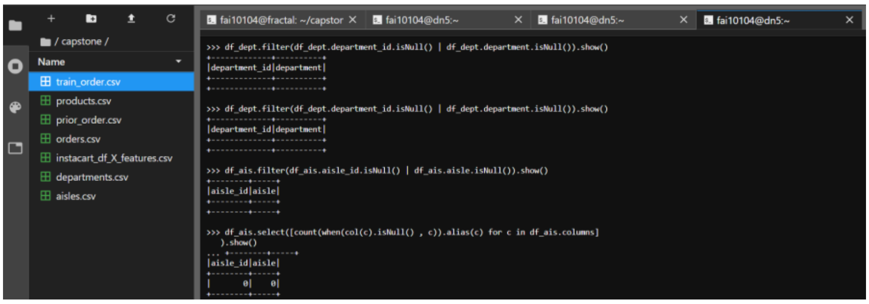
   Use df.column_name.isNull() in each column of a dataframe to display all the null values present.
   Use df.filter() to display according to the specified condition (null values, in this case)
   Use count() to display the count of null values in each column.

   a. Orders.csv (df_order)



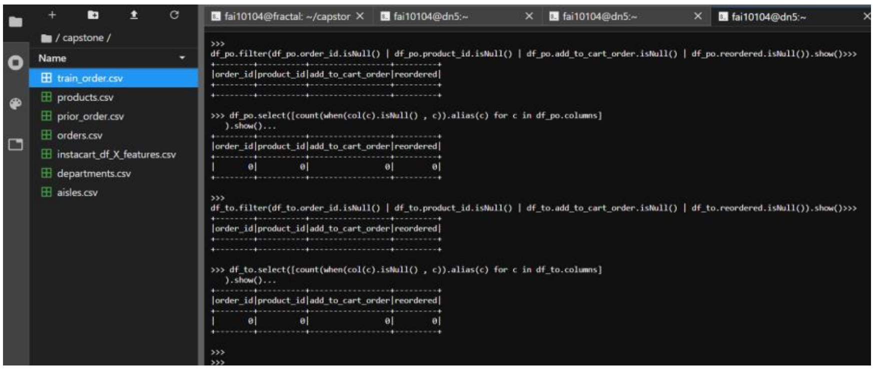
```
>>> df_order.filter(df_order.order_id.isNull() | df_order.user_id.isNull() | df_order.eval_set.isNull() | df_order.order_number.isNull() | df_
() | df_order.order_hour_of_day.isNull() | df_order.days_since_prior_order.isNull()).show()
+--------+-------+--------+------------+---------+-----------------+----------------------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|
+--------+-------+--------+------------+---------+-----------------+----------------------+
| 1363380|     50|   prior|           1|        3|                9|                  null|
| 3131103|     50|   prior|           2|        6|               12|                  null|
| 2197066|     50|   prior|           3|        1|               13|                  null|
| 3201640|     50|   prior|           4|        0|               11|                  null|
| 2756806|     50|   prior|           5|        4|               14|                  null|
| 1905727|     50|   prior|           6|        2|               10|                  null|
|  357849|     50|   prior|           7|        3|               13|                  null|
| 1673791|     50|   prior|           8|        6|               14|                  null|
| 1941871|     50|   prior|           9|        4|               11|                  null|
| 1431607|     50|   prior|          10|        2|               10|                  null|
|  935755|     50|   prior|          11|        6|                9|                  null|
|  771250|     50|   prior|          12|        3|               18|                  null|
|  218647|     50|   prior|          13|        6|               14|                  null|
|  482971|     50|   prior|          14|        4|               11|                  null|
| 1261865|     50|   prior|          15|        2|               10|                  null|
| 2102570|     50|   prior|          16|        1|                8|                  null|
| 1856643|     50|   prior|          17|        2|               15|                  null|
| 2199859|     50|   prior|          18|        1|               10|                  null|
|  505360|     50|   prior|          19|        3|               11|                  null|
| 2829245|     50|   prior|          20|        1|               10|                  null|
+--------+-------+--------+------------+---------+-----------------+----------------------+
only showing top 20 rows

>>> df_order.select([count(when(col(c).isNull() , c)).alias(c) for c in df_order.columns]
   ).show()...
+--------+-------+--------+------------+---------+-----------------+----------------------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|
+--------+-------+--------+------------+---------+-----------------+----------------------+
|       0|      0|       0|           0|        0|                0|                 82683|
+--------+-------+--------+------------+---------+-----------------+----------------------+

>>> 
```

b. Department.csv (df_dept) and aisle.csv (df_ais)



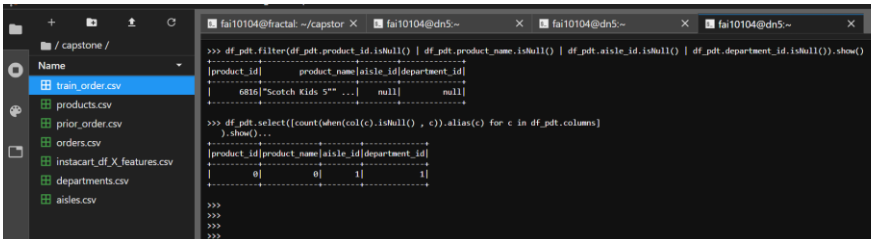
```
>>> df_dept.filter(df_dept.department_id.isNull() | df_dept.department.isNull()).show()
+-------------+----------+
|department_id|department|
+-------------+----------+
+-------------+----------+

>>> df_dept.filter(df_dept.department_id.isNull() | df_dept.department.isNull()).show()
+-------------+----------+
|department_id|department|
+-------------+----------+
+-------------+----------+

>>> df_ais.filter(df_ais.aisle_id.isNull() | df_ais.aisle.isNull()).show()
+--------+-----+
|aisle_id|aisle|
+--------+-----+
+--------+-----+

>>> df_ais.select([count(when(col(c).isNull() , c)).alias(c) for c in df_ais.columns]
   ).show()
... +--------+-----+
|aisle_id|aisle|
+--------+-----+
|       0|    0|
+--------+-----+
```

c. Prior_order.csv (df_po) and train_orders.csv (df_to)



```
>>>
df_po.filter(df_po.order_id.isNull() | df_po.product_id.isNull() | df_po.add_to_cart_order.isNull() | df_po.reordered.isNull()).show()>>>
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
+--------+----------+-----------------+---------+

>>> df_po.select([count(when(col(c).isNull() , c)).alias(c) for c in df_po.columns]
   ).show()...
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
|       0|         0|                0|        0|
+--------+----------+-----------------+---------+

>>>
df_to.filter(df_to.order_id.isNull() | df_to.product_id.isNull() | df_to.add_to_cart_order.isNull() | df_to.reordered.isNull()).show()>>>
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
+--------+----------+-----------------+---------+

>>> df_to.select([count(when(col(c).isNull() , c)).alias(c) for c in df_to.columns]
   ).show()...
+--------+----------+-----------------+---------+
|order_id|product_id|add_to_cart_order|reordered|
+--------+----------+-----------------+---------+
|       0|         0|                0|        0|
+--------+----------+-----------------+---------+

>>>
>>>
```

d. Products.csv (df_pdt)



```
>>> df_pdt.filter(df_pdt.product_id.isNull() | df_pdt.product_name.isNull() | df_pdt.aisle_id.isNull() | df_pdt.department_id.isNull()).show()
+----------+--------------------+--------+-------------+
|product_id|        product_name|aisle_id|department_id|
+----------+--------------------+--------+-------------+
|      6816|"Scotch Kids 5"" ...|    null|         null|
+----------+--------------------+--------+-------------+

>>> df_pdt.select([count(when(col(c).isNull() , c)).alias(c) for c in df_pdt.columns]
   ).show()...
+----------+------------+--------+-------------+
|product_id|product_name|aisle_id|department_id|
+----------+------------+--------+-------------+
|         0|           0|       1|            1|
+----------+------------+--------+-------------+

>>>
>>>
>>>
>>>
```

10. Merging of datasets:

Use df.join() to merge two or more datasets.

Join conditions used:

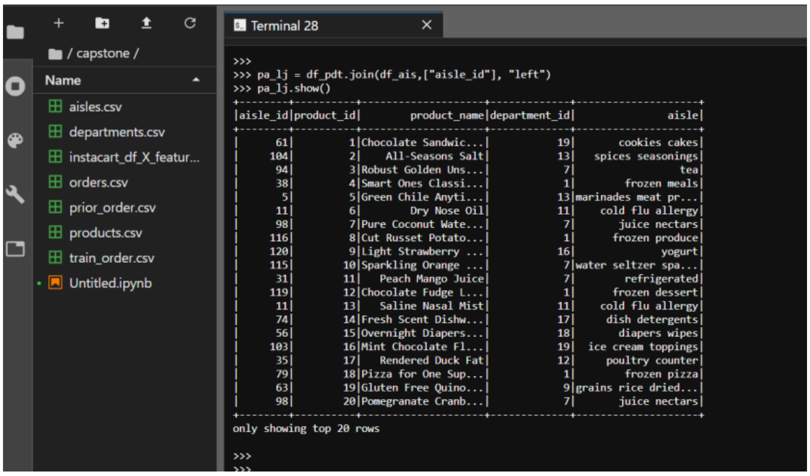
(i) Left Join Product with aisle using aisle_id.

(ii) Left join above joined product with department using department_id.

(iii) Inner join above new joined product with prior_orders and train_orders with product_id.

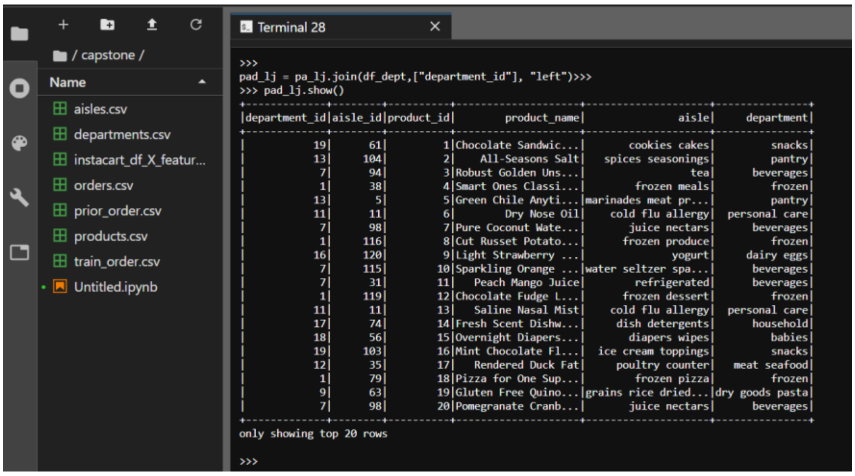
(iv) Left join Orders with prior_order and train_orders using order_id.

a. Product – Left Join - aisle:



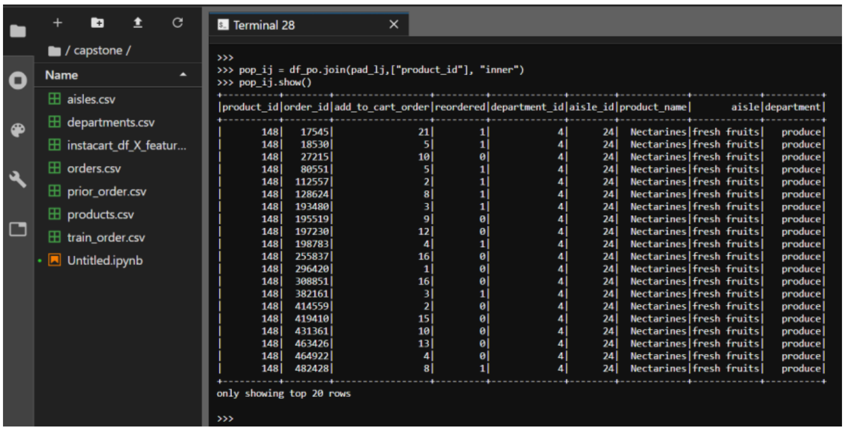
```
>>>
>>> pa_lj = df_pdt.join(df_ais,["aisle_id"], "left")
>>> pa_lj.show()
+--------+----------+--------------------+-------------+--------------------+
|aisle_id|product_id|        product_name|department_id|               aisle|
+--------+----------+--------------------+-------------+--------------------+
|      61|         1|Chocolate Sandwic...|           19|       cookies cakes|
|     104|         2|    All-Seasons Salt|           13|   spices seasonings|
|      94|         3|Robust Golden Uns...|            7|                 tea|
|      38|         4|Smart Ones Classi...|            1|        frozen meals|
|       5|         5|Green Chile Anyti...|           13|marinades meat pr...|
|      11|         6|        Dry Nose Oil|           11|    cold flu allergy|
|      98|         7|Pure Coconut Wate...|            7|       juice nectars|
|     116|         8|Cut Russet Potato...|            1|      frozen produce|
|     120|         9|Light Strawberry ...|           16|              yogurt|
|     115|        10|Sparkling Orange ...|            7|water seltzer spa...|
|      31|        11|   Peach Mango Juice|            7|        refrigerated|
|     119|        12|Chocolate Fudge L...|            1|      frozen dessert|
|      11|        13|   Saline Nasal Mist|           11|    cold flu allergy|
|      74|        14|Fresh Scent Dishw...|           17|     dish detergents|
|      56|        15|Overnight Diapers...|           18|       diapers wipes|
|     103|        16|Mint Chocolate Fl...|           19|  ice cream toppings|
|      35|        17|   Rendered Duck Fat|           12|     poultry counter|
|      79|        18|Pizza for One Sup...|            1|        frozen pizza|
|      63|        19|Gluten Free Quino...|            9|grains rice dried...|
|      98|        20|Pomegranate Cranb...|            7|       juice nectars|
+--------+----------+--------------------+-------------+--------------------+
only showing top 20 rows

>>>
>>>
```

b. Product_aisle – Left Join - department:



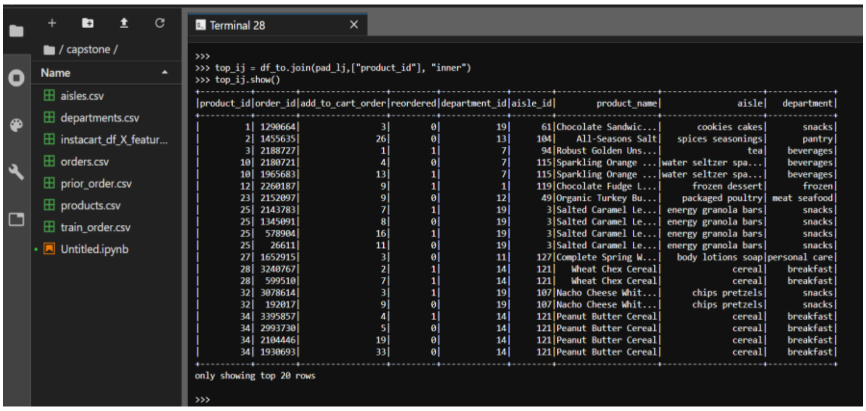
```
>>>
pad_lj = pa_lj.join(df_dept,["department_id"], "left")>>>
>>> pad_lj.show()
+-------------+--------+----------+--------------------+--------------------+---------------+
|department_id|aisle_id|product_id|        product_name|               aisle|     department|
+-------------+--------+----------+--------------------+--------------------+---------------+
|           19|      61|         1|Chocolate Sandwic...|       cookies cakes|         snacks|
|           13|     104|         2|    All-Seasons Salt|   spices seasonings|         pantry|
|            7|      94|         3|Robust Golden Uns...|                 tea|      beverages|
|            1|      38|         4|Smart Ones Classi...|        frozen meals|         frozen|
|           13|       5|         5|Green Chile Anyti...|marinades meat pr...|         pantry|
|           11|      11|         6|        Dry Nose Oil|    cold flu allergy|  personal care|
|            7|      98|         7|Pure Coconut Wate...|       juice nectars|      beverages|
|            1|     116|         8|Cut Russet Potato...|      frozen produce|         frozen|
|           16|     120|         9|Light Strawberry ...|              yogurt|     dairy eggs|
|            7|     115|        10|Sparkling Orange ...|water seltzer spa...|      beverages|
|            7|      31|        11|   Peach Mango Juice|        refrigerated|      beverages|
|            1|     119|        12|Chocolate Fudge L...|      frozen dessert|         frozen|
|           11|      11|        13|   Saline Nasal Mist|    cold flu allergy|  personal care|
|           17|      74|        14|Fresh Scent Dishw...|     dish detergents|      household|
|           18|      56|        15|Overnight Diapers...|       diapers wipes|         babies|
|           19|     103|        16|Mint Chocolate Fl...|  ice cream toppings|         snacks|
|           12|      35|        17|   Rendered Duck Fat|     poultry counter|   meat seafood|
|            1|      79|        18|Pizza for One Sup...|        frozen pizza|         frozen|
|            9|      63|        19|Gluten Free Quino...|grains rice dried...|dry goods pasta|
|            7|      98|        20|Pomegranate Cranb...|       juice nectars|      beverages|
+-------------+--------+----------+--------------------+--------------------+---------------+
only showing top 20 rows

>>>
```

c. Prior_order – Inner Join – joined_product :



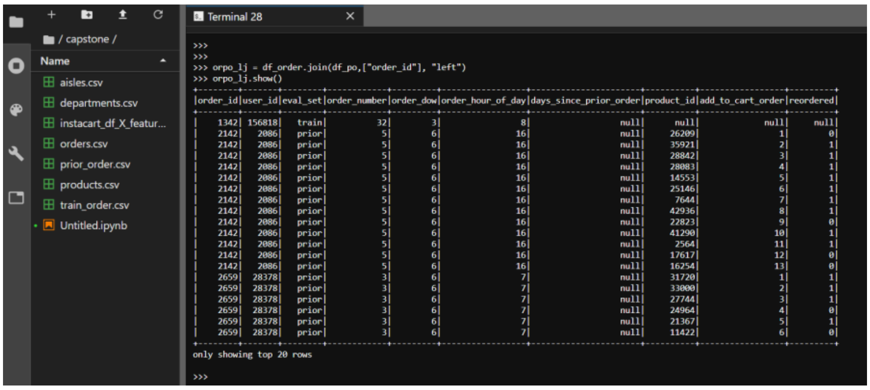
```
>>>
>>> pop_ij = df_po.join(pad_lj,["product_id"], "inner")
>>> pop_ij.show()
+----------+--------+-----------------+---------+-------------+--------+------------+------------+----------+
|product_id|order_id|add_to_cart_order|reordered|department_id|aisle_id|product_name|       aisle|department|
+----------+--------+-----------------+---------+-------------+--------+------------+------------+----------+
|       148|   17545|               21|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|   18530|                5|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|   27215|               10|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|   80551|                5|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  112557|                2|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  128624|                8|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  193480|                3|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  195519|                9|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  197230|               12|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  198783|                4|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  255837|               16|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  296420|                1|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  308851|               16|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  382161|                3|        1|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  414559|                2|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  419410|               15|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  431361|               10|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  463426|               13|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  464922|                4|        0|            4|      24|  Nectarines|fresh fruits|   produce|
|       148|  482428|                8|        1|            4|      24|  Nectarines|fresh fruits|   produce|
+----------+--------+-----------------+---------+-------------+--------+------------+------------+----------+
only showing top 20 rows

>>>
```

d. Train_order – Inner Join – joined_product:

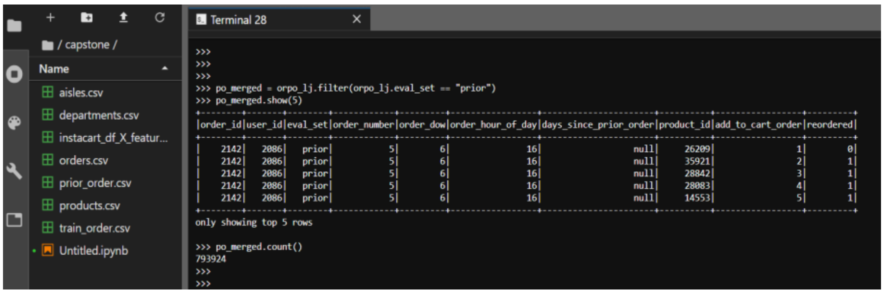


```
>>>
>>> top_ij = df_to.join(pad_lj,["product_id"], "inner")
>>> top_ij.show()
+----------+--------+-----------------+---------+-------------+--------+--------------------+--------------------+-------------+
|product_id|order_id|add_to_cart_order|reordered|department_id|aisle_id|        product_name|               aisle|   department|
+----------+--------+-----------------+---------+-------------+--------+--------------------+--------------------+-------------+
|         1| 1290664|                3|        0|           19|      61|Chocolate Sandwic...|       cookies cakes|       snacks|
|         2| 1455635|               26|        0|           13|     104|    All-Seasons Salt|   spices seasonings|       pantry|
|         3| 2188727|                1|        1|            7|      94|Robust Golden Uns...|                 tea|    beverages|
|        10| 2180721|                4|        0|            7|     115|Sparkling Orange ...|water seltzer spa...|    beverages|
|        10| 1965683|               13|        1|            7|     115|Sparkling Orange ...|water seltzer spa...|    beverages|
|        12| 2260187|                9|        1|            1|     119|Chocolate Fudge L...|      frozen dessert|       frozen|
|        23| 2152097|                9|        0|           12|      49|Organic Turkey Bu...|    packaged poultry| meat seafood|
|        25| 2143783|                7|        1|           19|       3|Salted Caramel Le...| energy granola bars|       snacks|
|        25| 1345091|                8|        0|           19|       3|Salted Caramel Le...| energy granola bars|       snacks|
|        25|  578904|               16|        1|           19|       3|Salted Caramel Le...| energy granola bars|       snacks|
|        25|   26611|               11|        0|           19|       3|Salted Caramel Le...| energy granola bars|       snacks|
|        27| 1652915|                3|        0|           11|     127|Complete Spring W...|   body lotions soap|personal care|
|        28| 3240767|                2|        1|           14|     121|   Wheat Chex Cereal|              cereal|    breakfast|
|        28|  599510|                7|        1|           14|     121|   Wheat Chex Cereal|              cereal|    breakfast|
|        32| 3078614|                3|        1|           19|     107|Nacho Cheese Whit...|      chips pretzels|       snacks|
|        32|  192017|                9|        0|           19|     107|Nacho Cheese Whit...|      chips pretzels|       snacks|
|        34| 3395857|                4|        1|           14|     121|Peanut Butter Cereal|              cereal|    breakfast|
|        34| 2993730|                5|        0|           14|     121|Peanut Butter Cereal|              cereal|    breakfast|
|        34| 2104446|               19|        0|           14|     121|Peanut Butter Cereal|              cereal|    breakfast|
|        34| 1930693|               33|        0|           14|     121|Peanut Butter Cereal|              cereal|    breakfast|
+----------+--------+-----------------+---------+-------------+--------+--------------------+--------------------+-------------+
only showing top 20 rows

>>>
```

e. Order – Left Join – prior_orders:

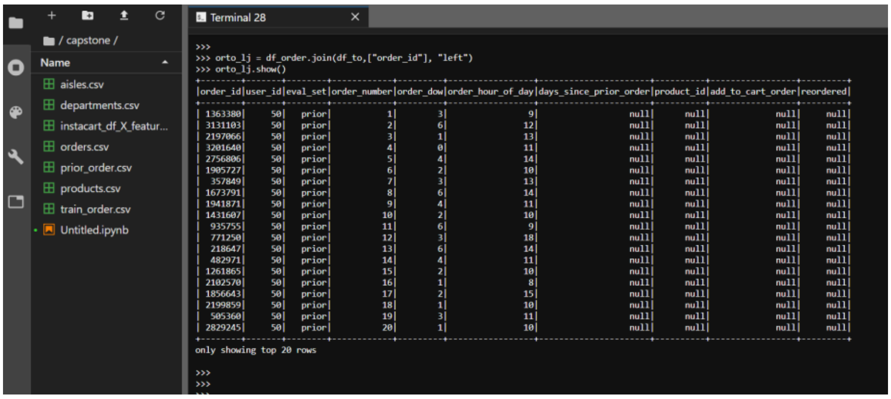


```
>>>
>>>
>>> orpo_lj = df_order.join(df_po,["order_id"], "left")
>>> orpo_lj.show()
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|product_id|add_to_cart_order|reordered|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|    1342| 156818|   train|          32|        3|                8|                  null|      null|             null|     null|
|    2142|   2086|   prior|           5|        6|               16|                  null|     26209|                1|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     35921|                2|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     28842|                3|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     28083|                4|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     14553|                5|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     25146|                6|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|      7644|                7|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     42936|                8|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     22823|                9|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     41290|               10|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|      2564|               11|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     17617|               12|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     16254|               13|        0|
|    2659|  28378|   prior|           3|        6|                7|                  null|     31720|                1|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     33000|                2|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     27744|                3|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     24964|                4|        0|
|    2659|  28378|   prior|           3|        6|                7|                  null|     21367|                5|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     11422|                6|        0|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
only showing top 20 rows

>>>
```

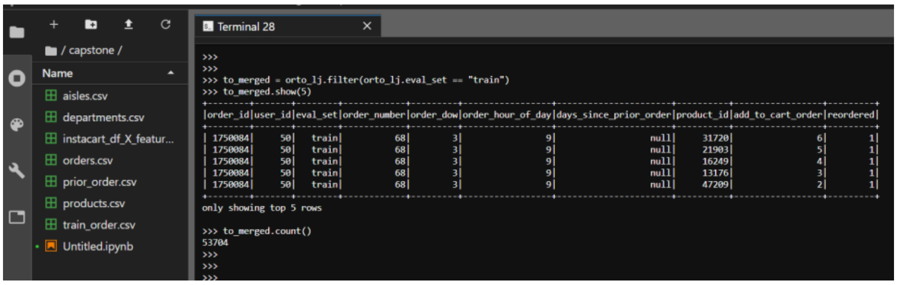


```
>>>
>>>
>>>
>>> po_merged = orpo_lj.filter(orpo_lj.eval_set == "prior")
>>> po_merged.show(5)
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|product_id|add_to_cart_order|reordered|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|    2142|   2086|   prior|           5|        6|               16|                  null|     26209|                1|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     35921|                2|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     28842|                3|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     28083|                4|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     14553|                5|        1|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
only showing top 5 rows

>>> po_merged.count()
793924
>>>
>>>
```

f. Order – Left Join – Train_orders:



```
>>>
>>> orto_lj = df_order.join(df_to,["order_id"], "left")
>>> orto_lj.show()
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|product_id|add_to_cart_order|reordered|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
| 1363380|     50|   prior|           1|        3|                9|                  null|      null|             null|     null|
| 3131103|     50|   prior|           2|        6|               12|                  null|      null|             null|     null|
| 2197066|     50|   prior|           3|        1|               13|                  null|      null|             null|     null|
| 3201640|     50|   prior|           4|        0|               11|                  null|      null|             null|     null|
| 2756806|     50|   prior|           5|        4|               14|                  null|      null|             null|     null|
| 1905727|     50|   prior|           6|        2|               10|                  null|      null|             null|     null|
|  357849|     50|   prior|           7|        3|               13|                  null|      null|             null|     null|
| 1673791|     50|   prior|           8|        6|               14|                  null|      null|             null|     null|
| 1941871|     50|   prior|           9|        4|               11|                  null|      null|             null|     null|
| 1431607|     50|   prior|          10|        2|               10|                  null|      null|             null|     null|
|  935755|     50|   prior|          11|        6|                9|                  null|      null|             null|     null|
|  771250|     50|   prior|          12|        3|               18|                  null|      null|             null|     null|
|  218647|     50|   prior|          13|        6|               14|                  null|      null|             null|     null|
|  482971|     50|   prior|          14|        4|               11|                  null|      null|             null|     null|
| 1261865|     50|   prior|          15|        2|               10|                  null|      null|             null|     null|
| 2102570|     50|   prior|          16|        1|                8|                  null|      null|             null|     null|
| 1856643|     50|   prior|          17|        2|               15|                  null|      null|             null|     null|
| 2199859|     50|   prior|          18|        1|               10|                  null|      null|             null|     null|
|  505360|     50|   prior|          19|        3|               11|                  null|      null|             null|     null|
| 2829245|     50|   prior|          20|        1|               10|                  null|      null|             null|     null|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
only showing top 20 rows

>>>
>>>
```

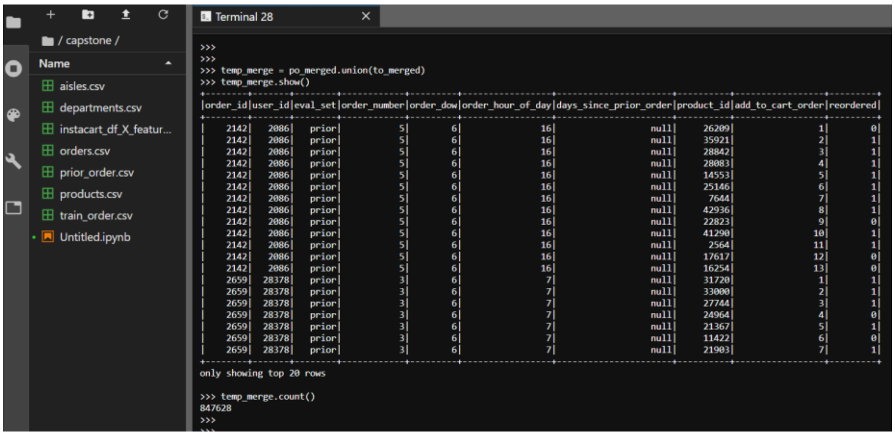


```
>>>
>>>
>>> to_merged = orto_lj.filter(orto_lj.eval_set == "train")
>>> to_merged.show(5)
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|product_id|add_to_cart_order|reordered|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
| 1750084|     50|   train|          68|        3|                9|                  null|     31720|                6|        1|
| 1750084|     50|   train|          68|        3|                9|                  null|     21903|                5|        1|
| 1750084|     50|   train|          68|        3|                9|                  null|     16249|                4|        1|
| 1750084|     50|   train|          68|        3|                9|                  null|     13176|                3|        1|
| 1750084|     50|   train|          68|        3|                9|                  null|     47209|                2|        1|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
only showing top 5 rows

>>> to_merged.count()
53704
>>>
>>>
>>>
```

11. Aggregation of datasets:

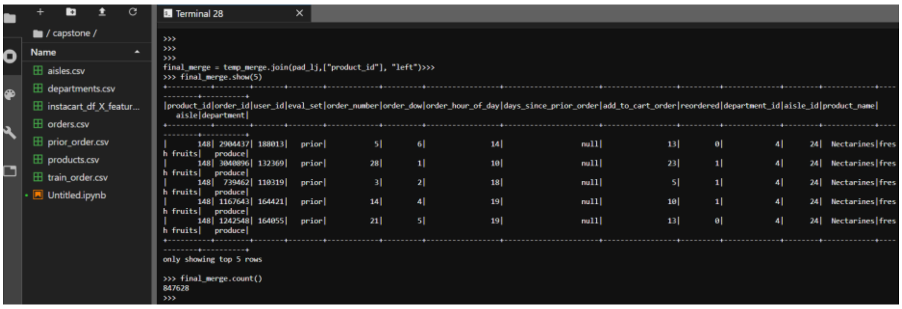
Use df1.union(df2) to combine two or more datasets.

a. Combining prior_orders and train_orders.



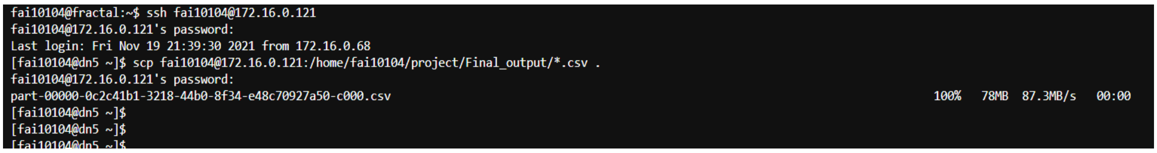
```
>>>
>>>
>>> temp_merge = po_merged.union(to_merged)
>>> temp_merge.show()
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|product_id|add_to_cart_order|reordered|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
|    2142|   2086|   prior|           5|        6|               16|                  null|     26209|                1|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     35921|                2|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     28842|                3|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     28083|                4|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     14553|                5|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     25146|                6|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|      7644|                7|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     42936|                8|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     22823|                9|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     41290|               10|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|      2564|               11|        1|
|    2142|   2086|   prior|           5|        6|               16|                  null|     17617|               12|        0|
|    2142|   2086|   prior|           5|        6|               16|                  null|     16254|               13|        0|
|    2659|  28378|   prior|           3|        6|                7|                  null|     31720|                1|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     33000|                2|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     27744|                3|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     24964|                4|        0|
|    2659|  28378|   prior|           3|        6|                7|                  null|     21367|                5|        1|
|    2659|  28378|   prior|           3|        6|                7|                  null|     11422|                6|        0|
|    2659|  28378|   prior|           3|        6|                7|                  null|     21903|                7|        1|
+--------+-------+--------+------------+---------+-----------------+----------------------+----------+-----------------+---------+
only showing top 20 rows

>>> temp_merge.count()
847628
>>>
>>>
```

b. Combine all the datasets into one dataframe.

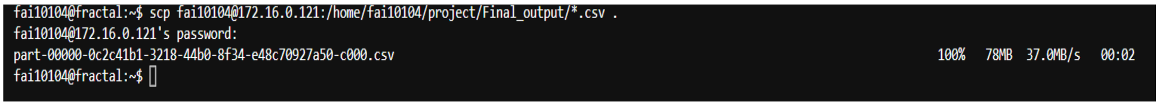


```
>>>
>>>
>>>
final_merge = temp_merge.join(pad_lj,["product_id"], "left")>>>
>>> final_merge.show(5)
+----------+--------+-------+--------+------------+---------+-----------------+----------------------+-----------------+---------+-------------+--------+------------+----
--------+----------+
|product_id|order_id|user_id|eval_set|order_number|order_dow|order_hour_of_day|days_since_prior_order|add_to_cart_order|reordered|department_id|aisle_id|product_name|
   aisle|department|
+----------+--------+-------+--------+------------+---------+-----------------+----------------------+-----------------+---------+-------------+--------+------------+----
--------+----------+
|       148| 2904437| 188013|   prior|           5|        6|               14|                  null|               13|        0|            4|      24|  Nectarines|fres
h fruits|   produce|
|       148| 3040896| 132369|   prior|          28|        1|               10|                  null|               23|        1|            4|      24|  Nectarines|fres
h fruits|   produce|
|       148|  739462| 110319|   prior|           3|        2|               18|                  null|                5|        1|            4|      24|  Nectarines|fres
h fruits|   produce|
|       148| 1167643| 164421|   prior|          14|        4|               19|                  null|               10|        1|            4|      24|  Nectarines|fres
h fruits|   produce|
|       148| 1242548| 164055|   prior|          21|        5|               19|                  null|               13|        0|            4|      24|  Nectarines|fres
h fruits|   produce|
+----------+--------+-------+--------+------------+---------+-----------------+----------------------+-----------------+---------+-------------+--------+------------+----
--------+----------+
only showing top 5 rows

>>> final_merge.count()
847628
>>>
```

12. Loading CSV file into cluster and local machine:



```
fai10104@fractal:~$ ssh fai10104@172.16.0.121
fai10104@172.16.0.121's password:
Last login: Fri Nov 19 21:39:30 2021 from 172.16.0.68
[fai10104@dn5 ~]$ scp fai10104@172.16.0.121:/home/fai10104/project/Final_output/*.csv .
fai10104@172.16.0.121's password:
part-00000-0c2c41b1-3218-44b0-8f34-e48c70927a50-c000.csv                                               100%   78MB  87.3MB/s   00:00
[fai10104@dn5 ~]$
[fai10104@dn5 ~]$
[fai10104@dn5 ~]$
```



```
fai10104@fractal:~$ scp fai10104@172.16.0.121:/home/fai10104/project/Final_output/*.csv .
fai10104@172.16.0.121's password:
part-00000-0c2c41b1-3218-44b0-8f34-e48c70927a50-c000.csv                                               100%   78MB  37.0MB/s   00:02
fai10104@fractal:~$ ▯
```

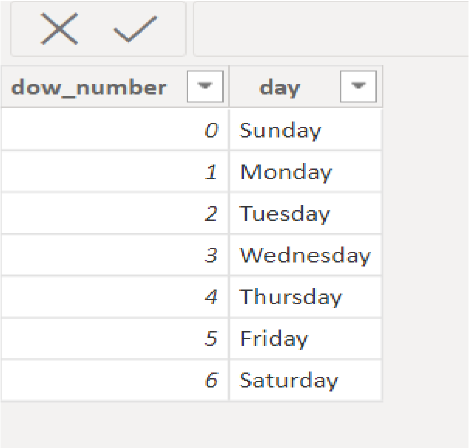
13. Visualisation:

Transformation made before loading:

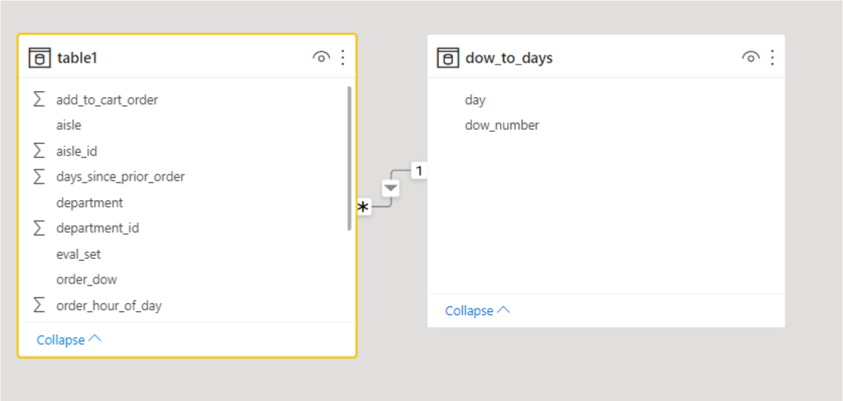
a. Removed duplicated rows and columns

b. Removed empty/null values

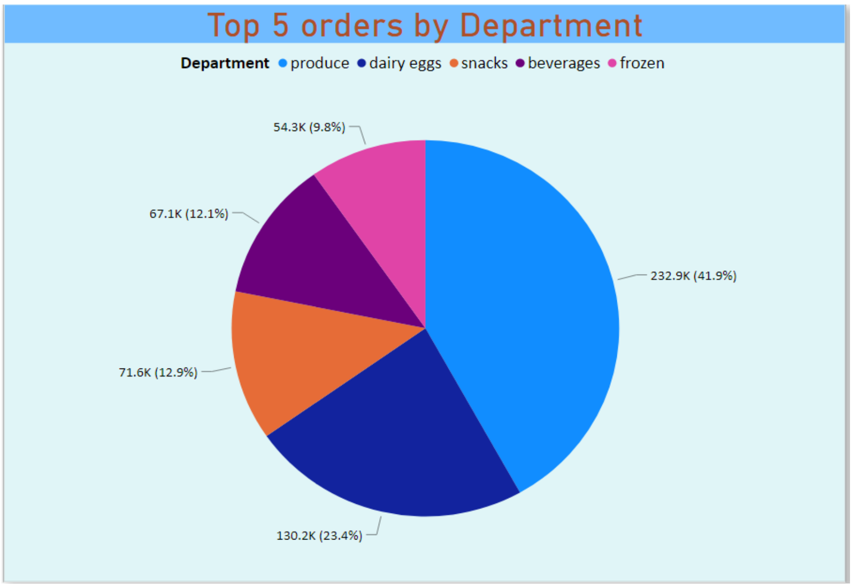
c. Created new table and relationship with the other table.



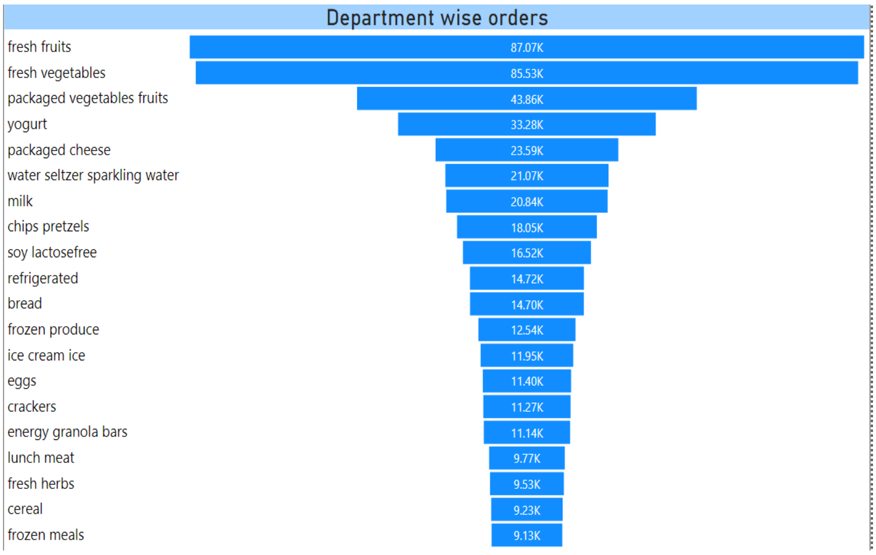
dow_number	day
0	Sunday
1	Monday
2	Tuesday
3	Wednesday
4	Thursday
5	Friday
6	Saturday



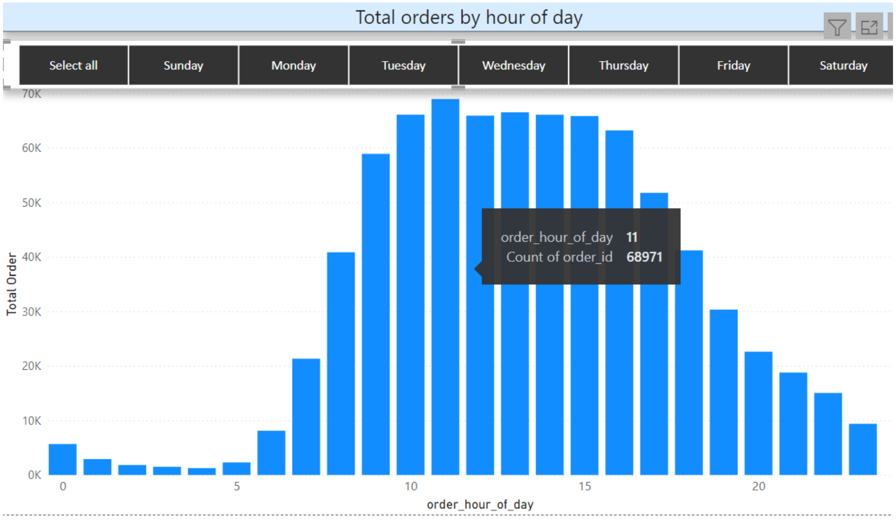
d. Created visuals:

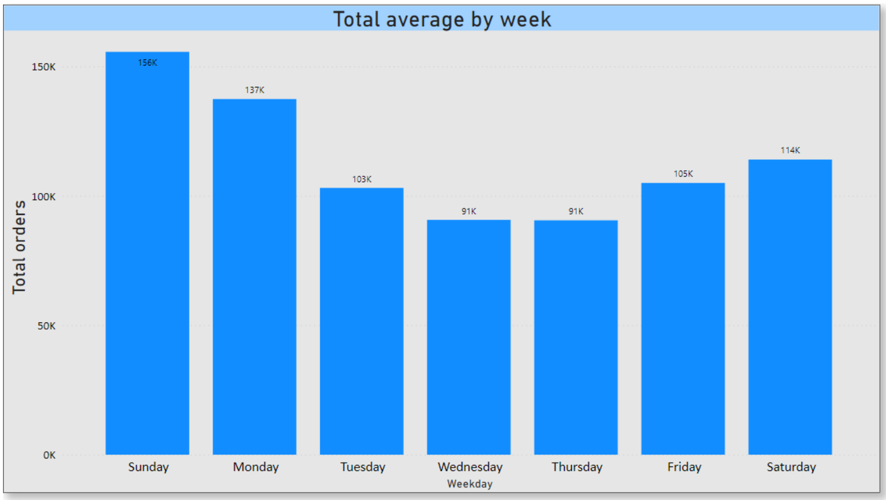


Insight: ‘Produce’ department accounted 41.9% of top 5 orders.

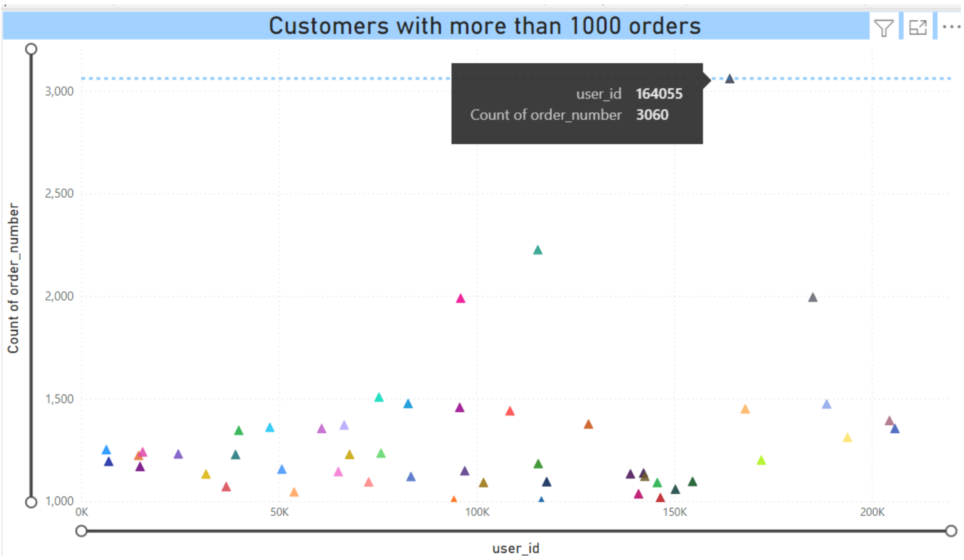


Insight: Fresh Fruits accounted around 87 thousand orders and frozen meals having lower sales. Improving cereals, frozen meals sales can drive business orders.

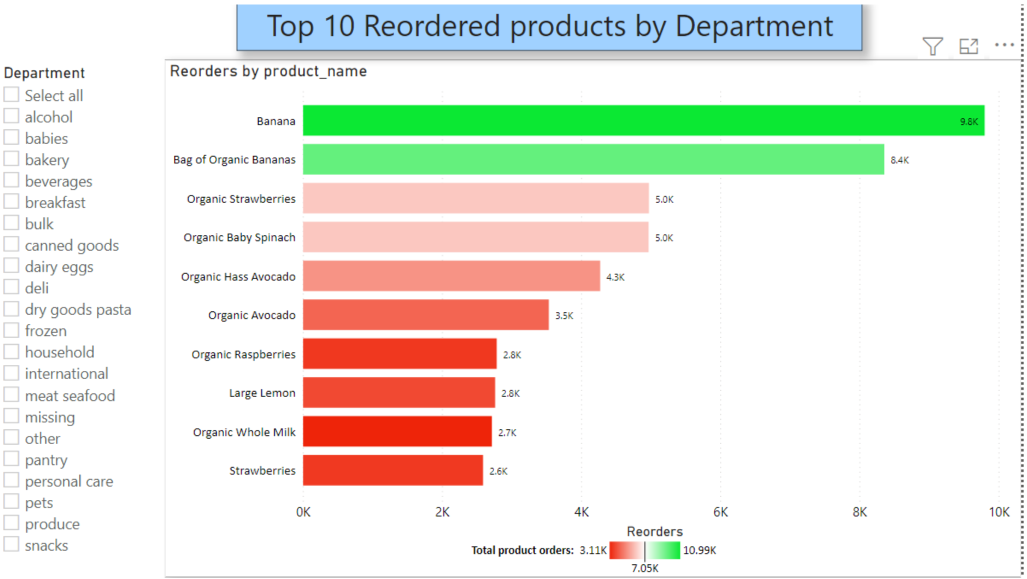




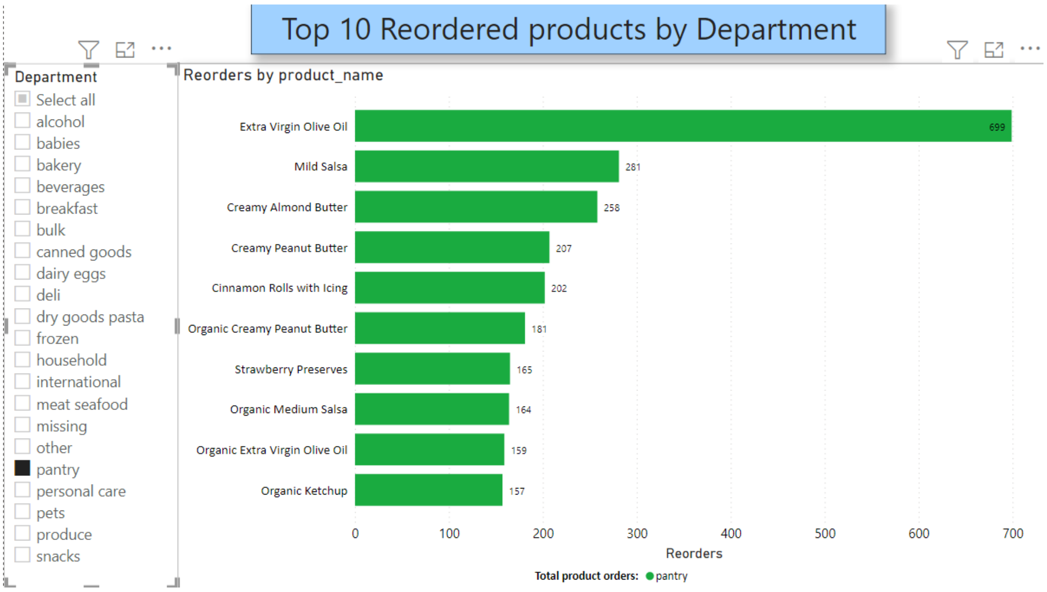
Insight: The highest orders are made around 11AM in an average week. 8AM-6PM is peak hours on average. Sunday has the highest sales while Wednesday has the lowest. Improving sales on middle of weekdays can increase business growth.



Insight: Customer_number = 164055 contributes the most accounting 3060 orders. Offering sales discounts, freebies for other customers can increase sales orders.



Insight: Banana has the highest reorders accounting 9800 orders.



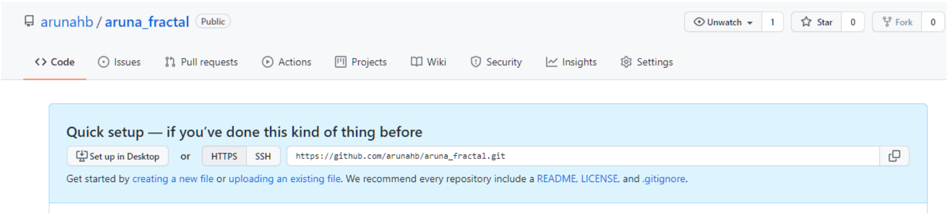
Insight: Selecting by department slicer, pantry for example provides corresponding sales orders.

Conclusion:

The sales orders can be increased by focusing on beverages and frozen department. Irregular customers can avail more discounts and offers than regular customers, on less peak hours on least sales day of the week (Wednesday). This can maintain business growth throughout.

Github repository:

a. Repository creation:

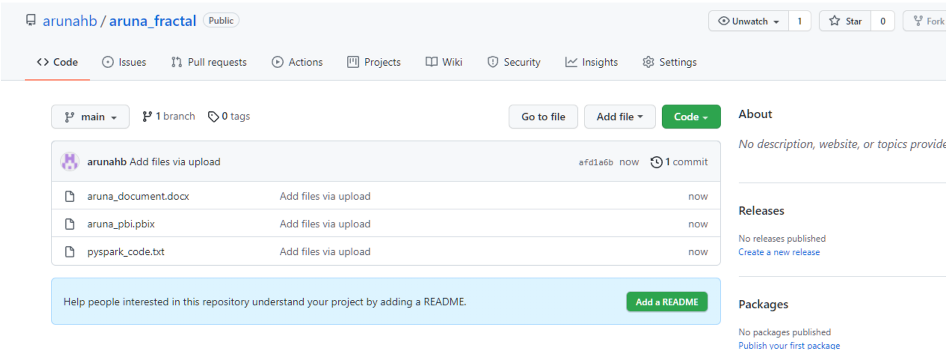


Quick setup — if you've done this kind of thing before

https://github.com/arunahb/aruna_fractal.git

Get started by creating a new file or uploading an existing file. We recommend every repository include a README, LICENSE, and .gitignore.

b. Upload supporting files:



arunahb Add files via upload afd1a6b now 1 commit

aruna_document.docx Add files via upload now

aruna_pbi.pbix Add files via upload now

pyspark_code.txt Add files via upload now

Help people interested in this repository understand your project by adding a README.

No description, website, or topics provide

No releases published
Create a new release

No packages published
Publish your first package

Github_link: https://github.com/arunahb/aruna_fractal

<u>Files submitted:</u>

aruna_document, aruna_pbi, pyspark_code, aruna_ouput.csv