Loading dataset into FTP location and then to HDFS Directory:

```
[therohitsaha08gmail@sl-cdp-prod-en10 ~]$ cd Capstone/ECommerce
[therohitsaha08gmail@sl-cdp-prod-en10 ECommerce]$ ls
olist_public_dataset.csv
[therohitsaha08gmail@sl-cdp-prod-en10 ECommerce]$ hdfs dfs -ls
Found 6 items
drwxrwxr-x   - therohitsaha08gmail hadoop          0 2024-01-29 11:42 .sparkStaging
drwxrwxr-x   - therohitsaha08gmail hadoop          0 2024-01-25 08:41 ECommerce
drwxrwxr-x   - therohitsaha08gmail hadoop          0 2024-01-24 11:31 Weather_Analysis_Project
drwxrwxr-x   - therohitsaha08gmail hadoop          0 2024-01-18 12:12 YoutubeDataProject
drwxrwxrwx   - therohitsaha08gmail hadoop          0 2023-08-02 09:59 hive_demo
drwxrwxr-x   - therohitsaha08gmail hadoop          0 2023-04-22 13:18 simplilearn
[therohitsaha08gmail@sl-cdp-prod-en10 ECommerce]$ hdfs dfs -ls ECommerce
Found 1 items
-rw-rw-r--   2 therohitsaha08gmail hadoop   10304725 2024-01-25 08:41 ECommerce/olist_public_dataset.csv
[therohitsaha08gmail@sl-cdp-prod-en10 ECommerce]$
```

Load dataset into Spark DataFrame from HDFS:

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.1.7.2.15.1-1
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_332)
Type in expressions to have them evaluated.
Type :help for more information.

scala> var df = spark.read.format("csv").option("delimiter", ",").option("header", "false").option("inferSchema", "true").load("ECommerce/olist_public_dataset.csv").toDF("id" ,"order_status" ,"order_products_value" ,"order_freight
_value" ,"order_items_qty" ,"customer_city" ,"customer_state" ,"customer_zip_code_prefix" ,"product_name_length" ,"product_description_length" ,"product_photos_qty" ,"review_score" ,"order_purchase_timestamp" ,"order_aproved_at" ,
"order_delivered_customer_date")
df: org.apache.spark.sql.DataFrame = [id: int, order_status: string ... 13 more fields]

scala> df.show()
+---+------------+--------------------+------------------+---------------+------------------+--------------+------------------------+-------------------+--------------------------+------------------+------------+----------------
--------+----------------+-----------------------------+
| id|order_status|order_products_value|order_freight_value|order_items_qty|     customer_city|customer_state|customer_zip_code_prefix|product_name_length|product_description_length|product_photos_qty|review_score|order_purchase_t
imestamp|order_aproved_at|order_delivered_customer_date|
+---+------------+--------------------+------------------+---------------+------------------+--------------+------------------------+-------------------+--------------------------+------------------+------------+----------------
--------+----------------+-----------------------------+
|  1|   delivered|                79.0|              17.8|              1|          Luziania|            GO|                     728|                 50|                       201|                 2|           5|          02/10/
17 10:56|  02/10/17 11:07|               10/10/17 21:25|
|  2|   delivered|               119.9|             27.16|              1|         Joinville|            SC|                     892|                 50|                       511|                 3|           5|          24/07/
18 20:41|   26/07/18 3:24|               07/08/18 15:27|
|  3|   delivered|              519.99|             41.69|              1|             Serra|            ES|                     291|                 48|                      1156|                 2|           1|           08/08
/18 8:38|   08/08/18 8:55|               17/08/18 18:06|
|  4|   delivered|                29.5|             17.92|              1|    RIO DE JANEIRO|            RJ|                     222|                 21|                       207|                 2|           4|          18/11/
17 19:28|  18/11/17 19:45|                02/12/17 0:28|
|  5|   delivered|               26.77|             23.11|              1|         Sao Paulo|            SP|                      40|                 41|                       451|                 1|           5|          13/02/
18 21:18|  13/02/18 22:20|               16/02/18 18:17|
|  6|   delivered|               419.9|             23.02|              1|      Santa Adelia|            SP|                     159|                 46|                       692|                 3|           5|          09/07/
17 21:57|  09/07/17 22:10|               26/07/17 10:57|
|  7|   delivered|                65.0|             16.21|              1|          Varginha|            MG|                     370|                 47|                       893|                 3|           5|          11/04/
17 12:22|  13/04/17 13:25|               13/04/17 13:25|
|  8|   delivered|               29.99|             19.82|              1|         Sao Paulo|            SP|                      48|                 55|                       613|                 1|           4|          16/05/
17 13:10|  16/05/17 13:22|               26/05/17 12:55|
|  9|   delivered|               59.99|             51.14|              1|           Carajas|            PA|                     685|                 28|                        96|                 4|           5|          23/01/
```

Extract only ‘date’ from all 3 timestamp columns respectively and store in a new dataframe:

```
scala> var date_df = df.withColumn("PurchaseDate",to_date(unix_timestamp(col("order_purchase_timestamp"), "dd/MM/yy H:mm").cast("timestamp"))).withColumn("ApproveDate",to_date(unix_timestamp(col("order_aproved_at"), "dd/MM/yy H:mm
").cast("timestamp"))).withColumn("DeliveryDate",to_date(unix_timestamp(col("order_delivered_customer_date"), "dd/MM/yy H:mm").cast("timestamp")))
date_df: org.apache.spark.sql.DataFrame = [id: int, order_status: string ... 16 more fields]

scala> date_df.show()
+---+------------+--------------------+------------------+---------------+------------------+--------------+------------------------+-------------------+--------------------------+------------------+------------+----------------
--------+----------------+-----------------------------+------------+-----------+------------+
| id|order_status|order_products_value|order_freight_value|order_items_qty|     customer_city|customer_state|customer_zip_code_prefix|product_name_length|product_description_length|product_photos_qty|review_score|order_purchase_t
imestamp|order_aproved_at|order_delivered_customer_date|PurchaseDate|ApproveDate|DeliveryDate|
+---+------------+--------------------+------------------+---------------+------------------+--------------+------------------------+-------------------+--------------------------+------------------+------------+----------------
--------+----------------+-----------------------------+------------+-----------+------------+
|  1|   delivered|                79.0|              17.8|              1|          Luziania|            GO|                     728|                 50|                       201|                 2|           5|          02/10/
17 10:56|  02/10/17 11:07|               10/10/17 21:25|  2017-10-02| 2017-10-02|  2017-10-10|
|  2|   delivered|               119.9|             27.16|              1|         Joinville|            SC|                     892|                 50|                       511|                 3|           5|          24/07/
18 20:41|   26/07/18 3:24|               07/08/18 15:27|  2018-07-24| 2018-07-26|  2018-08-07|
|  3|   delivered|              519.99|             41.69|              1|             Serra|            ES|                     291|                 48|                      1156|                 2|           1|           08/08
/18 8:38|   08/08/18 8:55|               17/08/18 18:06|  2018-08-08| 2018-08-08|  2018-08-17|
|  4|   delivered|                29.5|             17.92|              1|    RIO DE JANEIRO|            RJ|                     222|                 21|                       207|                 2|           4|          18/11/
17 19:28|  18/11/17 19:45|                02/12/17 0:28|  2017-11-18| 2017-11-18|  2017-12-02|
|  5|   delivered|               26.77|             23.11|              1|         Sao Paulo|            SP|                      40|                 41|                       451|                 1|           5|          13/02/
18 21:18|  13/02/18 22:20|               16/02/18 18:17|  2018-02-13| 2018-02-13|  2018-02-16|
|  6|   delivered|               419.9|             23.02|              1|      Santa Adelia|            SP|                     159|                 46|                       692|                 3|           5|          09/07/
17 21:57|  09/07/17 22:10|               26/07/17 10:57|  2017-07-09| 2017-07-09|  2017-07-26|
|  7|   delivered|                65.0|             16.21|              1|          Varginha|            MG|                     370|                 47|                       893|                 3|           5|          11/04/
17 12:22|  13/04/17 13:25|               13/04/17 13:25|  2017-04-11| 2017-04-13|  2017-04-13|
|  8|   delivered|               29.99|             19.82|              1|         Sao Paulo|            SP|                      48|                 55|                       613|                 1|           4|          16/05/
17 13:10|  16/05/17 13:22|               26/05/17 12:55|  2017-05-16| 2017-05-16|  2017-05-26|
|  9|   delivered|               59.99|             51.14|              1|           Carajas|            PA|                     685|                 28|                        96|                 4|           5|          23/01/
17 18:29|   25/01/17 2:50|               02/02/17 14:08|  2017-01-23| 2017-01-25|  2017-02-02|
| 10|   delivered|               56.99|             16.13|              1|           Resende|            RJ|                     275|                 59|                       641|                 3|           5|          29/07/
17 11:55|  29/07/17 12:05|               16/08/17 17:14|  2017-07-29| 2017-07-29|  2017-08-16|
| 11|   delivered|               599.0|             15.69|              1|      Guaratingueta|            SP|                     125|                 43|                      1686|                 2|           4|          16/05/
```

1. DAILY INSIGHTS

a. Sales

Total sales :

```
scala> var sales_df = date_df.groupBy(col("PurchaseDate")).agg(round(sum(col("order_products_value")),2).as("Total_Sales")).orderBy(asc("PurchaseDate"))
sales_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, Total_Sales: double]

scala> sales_df.show()
+------------+-----------+
|PurchaseDate|Total_Sales|
+------------+-----------+
|  2016-09-04|      200.0|
|  2016-09-05|       59.0|
|  2016-09-13|       99.0|
|  2016-09-15|       90.0|
|  2016-10-02|      56.99|
|  2016-10-03|      962.3|
|  2016-10-04|     8827.6|
|  2016-10-05|    6264.19|
|  2016-10-06|    8214.07|
|  2016-10-07|    6282.83|
|  2016-10-08|    4575.48|
|  2016-10-09|    2736.38|
|  2016-10-10|    4130.77|
|  2016-10-22|       49.0|
|  2016-12-23|       82.5|
|  2017-01-05|    3891.87|
|  2017-01-06|     391.98|
|  2017-01-07|     357.36|
|  2017-01-08|     552.58|
|  2017-01-09|     559.88|
+------------+-----------+
only showing top 20 rows
```

Total sales in each city :

```
scala> var sales_city_df = date_df.groupBy(col("PurchaseDate"),col("customer_city")).agg(round(sum(col("order_products_value")),2).as("Sales_per_City")).orderBy(asc("PurchaseDate"))
sales_city_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, customer_city: string ... 1 more field]

scala> sales_city_df.show()
+------------+--------------------+-------------+
|PurchaseDate|       customer_city|Sales_per_City|
+------------+--------------------+-------------+
|  2016-09-04|       Foz do Iguacu|        200.0|
|  2016-09-05|              Betim|         59.0|
|  2016-09-13|    Feira de Santana|         99.0|
|  2016-09-15|           Fortaleza|         90.0|
|  2016-10-02|       Montes Claros|        56.99|
|  2016-10-03|           Eunapolis|        139.9|
|  2016-10-03|            Jundiai|         45.8|
|  2016-10-03|      Belo Horizonte|         81.9|
|  2016-10-03|             Taubate|         58.9|
|  2016-10-03|              Bauru|        455.0|
|  2016-10-03|          Sao Paulo|         42.0|
|  2016-10-03|      Estrela do Sul|         89.9|
|  2016-10-03|             Guacui|         48.9|
|  2016-10-04|             Recife|        267.4|
|  2016-10-04|Vitoria da Conquista|        330.0|
|  2016-10-04|        Paulo Afonso|         45.9|
|  2016-10-04|           Sao Borja|         18.9|
|  2016-10-04|             Canoas|        46.99|
|  2016-10-04|         Uberlandia|         30.9|
|  2016-10-04|            Campinas|         34.9|
+------------+--------------------+-------------+
only showing top 20 rows
```

Total sales in each state :

```
scala> var sales_state_df = date_df.groupBy(col("PurchaseDate"),col("customer_state")).agg(round(sum(col("order_products_value")),2).as("Sales_per_State")).orderBy(asc("PurchaseDate"))
sales_state_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, customer_state: string ... 1 more field]

scala> sales_state_df.show()
+------------+--------------+--------------+
|PurchaseDate|customer_state|Sales_per_State|
+------------+--------------+--------------+
|  2016-09-04|            PR|         200.0|
|  2016-09-05|            MG|          59.0|
|  2016-09-13|            BA|          99.0|
|  2016-09-15|            CE|          90.0|
|  2016-10-02|            MG|         56.99|
|  2016-10-03|            ES|          48.9|
|  2016-10-03|            BA|         139.9|
|  2016-10-03|            MG|         171.8|
|  2016-10-03|            SP|         601.7|
|  2016-10-04|            SP|        3100.61|
|  2016-10-04|            PE|        1066.4|
|  2016-10-04|            RJ|        1287.52|
|  2016-10-04|            MG|        1215.49|
|  2016-10-04|            RS|         984.39|
|  2016-10-04|            GO|         270.69|
|  2016-10-04|            SC|         109.9|
|  2016-10-04|            DF|          67.9|
|  2016-10-04|            PR|          99.0|
|  2016-10-04|            ES|          69.9|
|  2016-10-04|            BA|         555.8|
+------------+--------------+--------------+
only showing top 20 rows
```

b. Orders

Total number of orders :

```
scala> var total_order = date_df.groupBy("PurchaseDate").agg(count("*").as("Total_Orders")).orderBy(asc("PurchaseDate"))
total_order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, Total_Orders: bigint]

scala> total_order.show()
+------------+------------+
|PurchaseDate|Total_Orders|
+------------+------------+
|  2016-09-04|           1|
|  2016-09-05|           1|
|  2016-09-13|           1|
|  2016-09-15|           1|
|  2016-10-02|           1|
|  2016-10-03|           8|
|  2016-10-04|          63|
|  2016-10-05|          48|
|  2016-10-06|          51|
|  2016-10-07|          46|
|  2016-10-08|          43|
|  2016-10-09|          26|
|  2016-10-10|          39|
|  2016-10-22|           1|
|  2016-12-23|           1|
|  2017-01-05|          32|
|  2017-01-06|           4|
|  2017-01-07|           4|
|  2017-01-08|           6|
|  2017-01-09|           5|
+------------+------------+
only showing top 20 rows
```

City-wise order distribution :

```
scala> var city_order = date_df.groupBy(col("PurchaseDate"),col("customer_city")).agg(count("*").as("Total_Orders")).orderBy(asc("PurchaseDate"))
city_order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, customer_city: string ... 1 more field]

scala> city_order.show()
+------------+------------------+------------+
|PurchaseDate|     customer_city|Total_Orders|
+------------+------------------+------------+
|  2016-09-04|     Foz do Iguacu|           1|
|  2016-09-05|            Betim|           1|
|  2016-09-13|  Feira de Santana|           1|
|  2016-09-15|         Fortaleza|           1|
|  2016-10-02|    Montes Claros|           1|
|  2016-10-03|    Belo Horizonte|           1|
|  2016-10-03|            Guacui|           1|
|  2016-10-03|         Eunapolis|           1|
|  2016-10-03|         Sao Paulo|           1|
|  2016-10-03|    Estrela do Sul|           1|
|  2016-10-03|          Taubate|           1|
|  2016-10-03|            Bauru|           1|
|  2016-10-03|          Jundiai|           1|
|  2016-10-04|            Serra|           1|
|  2016-10-04|            Garca|           1|
|  2016-10-04|          Campinas|           1|
|  2016-10-04|          Anapolis|           1|
|  2016-10-04|           Buritis|           1|
|  2016-10-04|Santana de Parnaiba|           1|
|  2016-10-04|         Sao Paulo|          11|
+------------+------------------+------------+
only showing top 20 rows
```

State-wise order distribution :

```
scala> var state_order = date_df.groupBy(col("PurchaseDate"),col("customer_state")).agg(count("*").as("Total_Orders")).orderBy(asc("PurchaseDate"))
state_order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, customer_state: string ... 1 more field]

scala> state_order.show()
+------------+--------------+------------+
|PurchaseDate|customer_state|Total_Orders|
+------------+--------------+------------+
|  2016-09-04|            PR|           1|
|  2016-09-05|            MG|           1|
|  2016-09-13|            BA|           1|
|  2016-09-15|            CE|           1|
|  2016-10-02|            MG|           1|
|  2016-10-03|            ES|           1|
|  2016-10-03|            BA|           1|
|  2016-10-03|            MG|           2|
|  2016-10-03|            SP|           4|
|  2016-10-04|            SP|          24|
|  2016-10-04|            PE|           4|
|  2016-10-04|            RJ|           9|
|  2016-10-04|            MG|          11|
|  2016-10-04|            RS|           4|
|  2016-10-04|            GO|           4|
|  2016-10-04|            SC|           1|
|  2016-10-04|            DF|           1|
|  2016-10-04|            PR|           1|
|  2016-10-04|            ES|           1|
|  2016-10-04|            BA|           3|
+------------+--------------+------------+
only showing top 20 rows
```

Average review score per order :

```
scala> var avg_rev = date_df.groupBy(col("PurchaseDate")).agg(round(sum(col("review_score"))/count("*"),3).as("Avg_Rev_per_Order")).orderBy(asc("PurchaseDate"))
avg_rev: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, Avg_Rev_per_Order: double]

scala> avg_rev.show()
+------------+-----------------+
|PurchaseDate|Avg_Rev_per_Order|
+------------+-----------------+
|  2016-09-04|              3.0|
|  2016-09-05|              5.0|
|  2016-09-13|              4.0|
|  2016-09-15|              5.0|
|  2016-10-02|              5.0|
|  2016-10-03|            3.625|
|  2016-10-04|            3.857|
|  2016-10-05|            4.229|
|  2016-10-06|              4.0|
|  2016-10-07|            3.739|
|  2016-10-08|            4.233|
|  2016-10-09|            3.923|
|  2016-10-10|            4.308|
|  2016-10-22|              1.0|
|  2016-12-23|              5.0|
|  2017-01-05|            4.125|
|  2017-01-06|             3.25|
|  2017-01-07|              3.0|
|  2017-01-08|            4.333|
|  2017-01-09|              4.8|
+------------+-----------------+
only showing top 20 rows
```

Average freight charges per order :

```
scala> var avg_fre = date_df.groupBy(col("PurchaseDate")).agg(round(sum(col("order_freight_value"))/count("*"),3).as("Avg_Frei_per_Order")).orderBy(asc("PurchaseDate"))
avg_fre: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [PurchaseDate: date, Avg_Frei_per_Order: double]

scala> avg_fre.show()
+------------+-----------------+
|PurchaseDate|Avg_Frei_per_Order|
+------------+-----------------+
|  2016-09-04|            39.88|
|  2016-09-05|            14.16|
|  2016-09-13|            19.93|
|  2016-09-15|            26.89|
|  2016-10-02|            15.15|
|  2016-10-03|           15.705|
|  2016-10-04|           20.357|
|  2016-10-05|           23.849|
|  2016-10-06|           21.093|
|  2016-10-07|           19.256|
|  2016-10-08|           24.621|
|  2016-10-09|           23.502|
|  2016-10-10|           21.252|
|  2016-10-22|            37.97|
|  2016-12-23|            44.16|
|  2017-01-05|           20.032|
|  2017-01-06|           15.883|
|  2017-01-07|           17.373|
|  2017-01-08|           47.932|
|  2017-01-09|           16.624|
+------------+-----------------+
only showing top 20 rows
```

Average time taken to approve the orders (order approved – order purchased) :

```
scala> import org.apache.spark.sql.types.DoubleType
import org.apache.spark.sql.types.DoubleType

scala> var approval_diff = df.withColumn("order_purchase_timestamp", to_timestamp(col("order_purchase_timestamp"), "dd/MM/yy H:mm")).withColumn("order_aproved_at", to_timestamp(col("order_aproved_at"), "dd/MM/yy H:mm")).withColumn
("ApprovalDifference(InMinutes)", ((col("order_aproved_at").cast("long") - col("order_purchase_timestamp").cast("long")) / 60).cast(DoubleType)).select("id", "order_purchase_timestamp", "order_aproved_at", "ApprovalDifference(InMi
nutes)")
approval_diff: org.apache.spark.sql.DataFrame = [id: int, order_purchase_timestamp: timestamp ... 2 more fields]

scala> approval_diff.show()
+---+------------------------+------------------+----------------------------+
| id|order_purchase_timestamp|  order_aproved_at|ApprovalDifference(InMinutes)|
+---+------------------------+------------------+----------------------------+
|  1|     2017-10-02 10:56:00|2017-10-02 11:07:00|                        11.0|
|  2|     2018-07-24 20:41:00|2018-07-26 03:24:00|                      1843.0|
|  3|     2018-08-08 08:38:00|2018-08-08 08:55:00|                        17.0|
|  4|     2017-11-18 19:28:00|2017-11-18 19:45:00|                        17.0|
|  5|     2018-02-13 21:18:00|2018-02-13 22:20:00|                        62.0|
|  6|     2017-07-09 21:57:00|2017-07-09 22:10:00|                        13.0|
|  7|     2017-04-11 12:22:00|2017-04-13 13:25:00|                      2943.0|
|  8|     2017-05-16 13:10:00|2017-05-16 13:22:00|                        12.0|
|  9|     2017-01-23 18:29:00|2017-01-25 02:50:00|                      1941.0|
| 10|     2017-07-29 11:55:00|2017-07-29 12:05:00|                        10.0|
| 11|     2017-05-16 19:41:00|2017-05-16 19:50:00|                         9.0|
| 12|     2017-07-13 19:58:00|2017-07-13 20:10:00|                        12.0|
| 13|     2018-06-07 10:06:00|2018-06-09 03:13:00|                      2467.0|
| 14|     2018-07-25 17:44:00|2018-07-25 17:55:00|                        11.0|
| 15|     2018-03-01 14:14:00|2018-03-01 15:10:00|                        56.0|
| 16|     2018-06-07 19:03:00|2018-06-12 23:31:00|                      7468.0|
| 17|     2018-01-02 19:00:00|2018-01-02 19:09:00|                         9.0|
| 18|     2017-12-26 23:41:00|2017-12-26 23:50:00|                         9.0|
| 19|     2017-11-21 00:03:00|2017-11-21 00:14:00|                        11.0|
| 20|     2017-10-26 15:54:00|2017-10-26 16:08:00|                        14.0|
+---+------------------------+------------------+----------------------------+
only showing top 20 rows


scala> var avg_approve = approval_diff.agg(round(avg("ApprovalDifference(InMinutes)"),3).as("Average Time To Approve(Minutes)")).show()
+-------------------------------+
|Average Time To Approve(Minutes)|
+-------------------------------+
|                        625.149|
+-------------------------------+

avg_approve: Unit = ()
```

Average order delivery time :

```
scala> var delivery_diff = df.withColumn("order_purchase_timestamp", to_timestamp(col("order_purchase_timestamp"), "dd/MM/yy H:mm")).withColumn("order_delivered_customer_date", to_timestamp(col("order_delivered_customer_date"), "d
d/MM/yy H:mm")).withColumn("DeliveryDifference(InHours)", ((col("order_delivered_customer_date").cast("long") - col("order_purchase_timestamp").cast("long")) / 3600).cast(DoubleType)).select("id", "order_purchase_timestamp", "orde
r_delivered_customer_date", "DeliveryDifference(InHours)")
delivery_diff: org.apache.spark.sql.DataFrame = [id: int, order_purchase_timestamp: timestamp ... 2 more fields]

scala> delivery_diff.show()
+---+------------------------+-----------------------------+---------------------------+
| id|order_purchase_timestamp|order_delivered_customer_date|DeliveryDifference(InHours)|
+---+------------------------+-----------------------------+---------------------------+
|  1|     2017-10-02 10:56:00|          2017-10-10 21:25:00|         202.48333333333332|
|  2|     2018-07-24 20:41:00|          2018-08-07 15:27:00|         330.76666666666665|
|  3|     2018-08-08 08:38:00|          2018-08-17 18:06:00|         225.46666666666667|
|  4|     2017-11-18 19:28:00|          2017-12-02 00:28:00|                      317.0|
|  5|     2018-02-13 21:18:00|          2018-02-16 18:17:00|          68.98333333333333|
|  6|     2017-07-09 21:57:00|          2017-07-26 10:57:00|                      397.0|
|  7|     2017-04-11 12:22:00|          2017-04-13 13:25:00|                      49.05|
|  8|     2017-05-16 13:10:00|          2017-05-26 12:55:00|                     239.75|
|  9|     2017-01-23 18:29:00|          2017-02-02 14:08:00|                     235.65|
| 10|     2017-07-29 11:55:00|          2017-08-16 17:14:00|         437.31666666666666|
| 11|     2017-05-16 19:41:00|          2017-05-29 11:18:00|          303.6166666666667|
| 12|     2017-07-13 19:58:00|          2017-07-19 14:04:00|                      138.1|
| 13|     2018-06-07 10:06:00|          2018-06-19 12:05:00|         289.98333333333335|
| 14|     2018-07-25 17:44:00|          2018-07-30 15:52:00|         118.13333333333334|
| 15|     2018-03-01 14:14:00|          2018-03-12 23:36:00|          273.3666666666667|
| 16|     2018-06-07 19:03:00|          2018-06-21 15:34:00|         332.51666666666665|
| 17|     2018-01-02 19:00:00|          2018-01-20 01:38:00|          414.6333333333333|
| 18|     2017-12-26 23:41:00|          2018-01-08 22:36:00|          310.9166666666667|
| 19|     2017-11-21 00:03:00|          2017-11-27 18:28:00|         162.41666666666666|
| 20|     2017-10-26 15:54:00|          2017-11-08 22:22:00|         318.46666666666664|
+---+------------------------+-----------------------------+---------------------------+
only showing top 20 rows


scala> var avg_delivery = delivery_diff.agg(round(avg("DeliveryDifference(InHours)"),3).as("Average Time To Deliver(Hours)")).show()
+-----------------------------+
|Average Time To Deliver(Hours)|
+-----------------------------+
|                       301.39|
+-----------------------------+

avg_delivery: Unit = ()
```

2. WEEKLY INSIGHTS

new dataframe having conversion of ‘order purchase date’ to ‘week number’ :

```
scala> var weekly_df = date_df.withColumn("WeekNumber", weekofyear(col("PurchaseDate")))
weekly_df: org.apache.spark.sql.DataFrame = [id: int, order_status: string ... 17 more fields]

scala> weekly_df.show()
| id|order_status|order_products_value|order_freight_value|order_items_qty|     customer_city|customer_state|customer_zip_code_prefix|product_name_length|product_description_length|product_photos_qty|review_score|order_purchase_t
imestamp|order_aproved_at|order_delivered_customer_date|PurchaseDate|ApproveDate|DeliveryDate|WeekNumber|
|  1|   delivered|                79.0|               17.8|              1|          Luziania|            GO|                     728|                 50|                       201|                 2|           5|          02/10/
17 10:56|  02/10/17 11:07|               10/10/17 21:25|  2017-10-02| 2017-10-02|  2017-10-10|        40|
|  2|   delivered|               119.9|              27.16|              1|         Joinville|            SC|                     892|                 50|                       511|                 3|           5|          24/07/
18 20:41|   26/07/18 3:24|               07/08/18 15:27|  2018-07-24| 2018-07-26|  2018-08-07|        30|
|  3|   delivered|              519.99|              41.69|              1|             Serra|            ES|                     291|                 48|                      1156|                 2|           1|           08/08
/18 8:38|   08/08/18 8:55|               17/08/18 18:06|  2018-08-08| 2018-08-08|  2018-08-17|        32|
|  4|   delivered|                29.5|              17.92|              1|    RIO DE JANEIRO|            RJ|                     222|                 21|                       207|                 2|           4|          18/11/
17 19:28|  18/11/17 19:45|                02/12/17 0:28|  2017-11-18| 2017-11-18|  2017-12-02|        46|
|  5|   delivered|               26.77|              23.11|              1|         Sao Paulo|            SP|                      40|                 41|                       451|                 1|           5|          13/02/
18 21:18|  13/02/18 22:20|               16/02/18 18:17|  2018-02-13| 2018-02-13|  2018-02-16|         7|
|  6|   delivered|               419.9|              23.02|              1|      Santa Adelia|            SP|                     159|                 46|                       692|                 3|           5|          09/07/
17 21:57|  09/07/17 22:10|               26/07/17 10:57|  2017-07-09| 2017-07-09|  2017-07-26|        27|
|  7|   delivered|                65.0|              16.21|              1|          Varginha|            MG|                     370|                 47|                       893|                 3|           5|          11/04/
17 12:22|  13/04/17 13:25|               13/04/17 13:25|  2017-04-11| 2017-04-13|  2017-04-13|        15|
|  8|   delivered|               29.99|              19.82|              1|         Sao Paulo|            SP|                      48|                 55|                       613|                 1|           4|          16/05/
17 13:10|  16/05/17 13:22|               26/05/17 12:55|  2017-05-16| 2017-05-16|  2017-05-26|        20|
|  9|   delivered|               59.99|              51.14|              1|           Carajas|            PA|                     685|                 28|                        96|                 4|           5|          23/01/
17 18:29|   25/01/17 2:50|               02/02/17 14:08|  2017-01-23| 2017-01-25|  2017-02-02|         4|
| 10|   delivered|               56.99|              16.13|              1|           Resende|            RJ|                     275|                 59|                       641|                 3|           5|          29/07/
17 11:55|  29/07/17 12:05|               16/08/17 17:14|  2017-07-29| 2017-07-29|  2017-08-16|        30|
| 11|   delivered|               599.0|              15.69|              1|     Guaratingueta|            SP|                     125|                 43|                      1686|                 2|           4|          16/05/
17 19:41|  16/05/17 19:50|               29/05/17 11:18|  2017-05-16| 2017-05-16|  2017-05-29|        20|
| 12|   delivered|                62.0|              16.19|              1|           Tijucas|            SC|                     882|                 50|                       285|                 1|           4|          13/07/
17 19:58|  13/07/17 20:10|               19/07/17 14:04|  2017-07-13| 2017-07-13|  2017-07-19|        28|
| 13|   delivered|               250.0|              35.02|              1|          Mirassol|            SP|                     151|                 57|                       771|                 1|           1|          07/06/
18 10:06|   09/06/18 3:13|               19/06/18 12:05|  2018-06-07| 2018-06-09|  2018-06-19|        23|
| 14|   delivered|                69.9|              12.93|              1|         SAO PAULO|            SP|                      82|                 57|                       427|                 1|           4|          25/07/
```

a. Sales

Total sales :

```
scala> var weeklysales_df = weekly_df.groupBy("WeekNumber").agg(round(sum(col("order_products_value")),2).as("Total_Weekly_Sales")).orderBy("WeekNumber")
weeklysales_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, Total_Weekly_Sales: double]

scala> weeklysales_df.show()
+----------+------------------+
|WeekNumber|Total_Weekly_Sales|
+----------+------------------+
|         1|         173944.85|
|         2|         236810.92|
|         3|         255674.81|
|         4|         256544.96|
|         5|         265820.53|
|         6|         268544.45|
|         7|         256302.52|
|         8|         283705.55|
|         9|         308138.64|
|        10|          293062.7|
|        11|         288693.72|
|        12|         305988.97|
|        13|         258064.74|
|        14|         280830.14|
|        15|         258832.68|
|        16|         295689.26|
|        17|         293175.73|
|        18|         308817.36|
|        19|         363995.98|
|        20|          364922.4|
+----------+------------------+
only showing top 20 rows
```

Total sales in each city :

```
scala> var weeklysales_city_df = weekly_df.groupBy(col("WeekNumber"),col("customer_city")).agg(round(sum(col("order_products_value")),2).as("Total_Weekly_Sales")).orderBy("WeekNumber")
weeklysales_city_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, customer_city: string ... 1 more field]

scala> weeklysales_city_df.show()
+----------+--------------------+-----------------+
|WeekNumber|       customer_city|Total_Weekly_Sales|
+----------+--------------------+-----------------+
|         1|           Rancharia|             19.9|
|         1|         Carapicuíba|           173.38|
|         1|  Balneario Camboriu|            101.8|
|         1|      RIO DE JANEIRO|         12889.46|
|         1|Ferraz de Vasconc...|             94.9|
|         1|              Areado|             89.9|
|         1|                 Luz|             30.0|
|         1|             Amparo|            400.69|
|         1|              Belem|           1043.29|
|         1|            Diadema|            81.56|
|         1|     Itaquaquecetuba|           227.49|
|         1|           Louveira|             35.0|
|         1|         Altinopolis|            58.99|
|         1| Sao Miguel Do Oeste|            203.0|
|         1|          Analandia|             27.9|
|         1|Capitao Leonidas ...|            53.99|
|         1|             Penedo|           149.99|
|         1|            Resende|            219.0|
|         1|        Santa Maria|           369.84|
|         1|Sao Jose do Vale ...|            98.99|
+----------+--------------------+-----------------+
only showing top 20 rows
```

Total sales in each state :

```
scala> var weeklysales_state_df = weekly_df.groupBy(col("WeekNumber"),col("customer_state")).agg(round(sum(col("order_products_value")),2).as("Total_Weekly_Sales")).orderBy("WeekNumber")
weeklysales_state_df: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, customer_state: string ... 1 more field]

scala> weeklysales_state_df.show()
+----------+--------------+-----------------+
|WeekNumber|customer_state|Total_Weekly_Sales|
+----------+--------------+-----------------+
|         1|            TO|           1074.0|
|         1|            SE|            40.89|
|         1|            RO|           979.86|
|         1|            SC|          4246.08|
|         1|            PI|            438.7|
|         1|            RN|           862.58|
|         1|            AM|           133.79|
|         1|            PE|           3388.4|
|         1|            PR|           7578.1|
|         1|            MT|          3468.43|
|         1|            BA|          6745.95|
|         1|            MS|          2744.11|
|         1|            MG|         21649.59|
|         1|            AC|            288.49|
|         1|            SP|         69261.83|
|         1|            ES|           2640.4|
|         1|            PB|          1062.59|
|         1|            CE|          2723.16|
|         1|            RJ|          22600.0|
|         1|            PA|          1501.98|
+----------+--------------+-----------------+
only showing top 20 rows
```

b. Orders

Total number of orders :

```
scala> var totalweekly_order = weekly_df.groupBy("WeekNumber").agg(count("*").as("Total_Weekly_Orders")).orderBy("WeekNumber")
totalweekly_order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, Total_Weekly_Orders: bigint]

scala> totalweekly_order.show()
+----------+-------------------+
|WeekNumber|Total_Weekly_Orders|
+----------+-------------------+
|         1|               1439|
|         2|               1875|
|         3|               1964|
|         4|               1956|
|         5|               2077|
|         6|               2148|
|         7|               2060|
|         8|               2145|
|         9|               2412|
|        10|               2246|
|        11|               2221|
|        12|               2381|
|        13|               2047|
|        14|               2190|
|        15|               2005|
|        16|               2258|
|        17|               2359|
|        18|               2570|
|        19|               2802|
|        20|               2797|
+----------+-------------------+
only showing top 20 rows
```

City-wise order distribution :

```
scala> var weekly_city_order = weekly_df.groupBy(col("WeekNumber"),col("customer_city")).agg(count("*").as("Total_Weekly_Orders")).orderBy("WeekNumber")
weekly_city_order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, customer_city: string ... 1 more field]

scala> weekly_city_order.show()
+----------+--------------------+-------------------+
|WeekNumber|       customer_city|Total_Weekly_Orders|
+----------+--------------------+-------------------+
|         1|Ferraz de Vasconc...|                  2|
|         1|           Americana|                  6|
|         1|  Balneario Camboriu|                  2|
|         1|      RIO DE JANEIRO|                113|
|         1|               Belem|                  8|
|         1|         Santa Maria|                  4|
|         1|     Itaquaquecetuba|                  3|
|         1| Sao Miguel Do Oeste|                  2|
|         1|      Cachoeira Alta|                  1|
|         1|   Sao Jose do Egito|                  1|
|         1|Vitoria da Conquista|                  2|
|         1|            Papucaia|                  1|
|         1|    Fernando Prestes|                  1|
|         1|    Pontes e Lacerda|                  1|
|         1|             Itaocara|                  1|
|         1|        Rondonopolis|                  1|
|         1|          Bataguassu|                  1|
|         1| Sao Jose dos Campos|                  5|
|         1|       Ilha Solteira|                  1|
|         1|             Cajamar|                  2|
+----------+--------------------+-------------------+
only showing top 20 rows
```

State-wise order distribution :

```
scala> var weekly_state_order = weekly_df.groupBy(col("WeekNumber"),col("customer_state")).agg(count("*").as("Total_Weekly_Orders")).orderBy("WeekNumber")
weekly_state_order: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, customer_state: string ... 1 more field]

scala> weekly_state_order.show()
+----------+--------------+-------------------+
|WeekNumber|customer_state|Total_Weekly_Orders|
+----------+--------------+-------------------+
|         1|            TO|                  5|
|         1|            SE|                  2|
|         1|            SC|                 42|
|         1|            PI|                  4|
|         1|            RO|                  6|
|         1|            RN|                  6|
|         1|            AM|                  3|
|         1|            PE|                 28|
|         1|            PR|                 78|
|         1|            MT|                 20|
|         1|            MS|                 20|
|         1|            MG|                170|
|         1|            AC|                  2|
|         1|            SP|                592|
|         1|            ES|                 27|
|         1|            PB|                  7|
|         1|            CE|                 19|
|         1|            RJ|                195|
|         1|            PA|                 13|
|         1|            BA|                 52|
+----------+--------------+-------------------+
only showing top 20 rows
```

Average review score per order :

```
scala> var weekly_avg_rev = weekly_df.groupBy(col("WeekNumber")).agg(round(sum(col("review_score"))/count("*"),3).as("Avg_Rev_per_Order")).orderBy("WeekNumber")
weekly_avg_rev: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, Avg_Rev_per_Order: double]

scala> weekly_avg_rev.show()
+----------+-----------------+
|WeekNumber|Avg_Rev_per_Order|
+----------+-----------------+
|         1|            4.054|
|         2|             4.06|
|         3|             4.03|
|         4|            4.076|
|         5|             4.09|
|         6|            4.041|
|         7|            4.057|
|         8|            4.042|
|         9|            4.053|
|        10|            4.054|
|        11|            4.059|
|        12|            4.034|
|        13|            4.071|
|        14|             4.08|
|        15|            4.039|
|        16|            4.029|
|        17|            4.029|
|        18|            4.105|
|        19|            4.043|
|        20|            4.044|
+----------+-----------------+
only showing top 20 rows
```

Average freight charges per order :

```
scala> var weekly_avg_fre = weekly_df.groupBy(col("WeekNumber")).agg(round(sum(col("order_freight_value"))/count("*"),3).as("Avg_Rev_per_Order")).orderBy("WeekNumber")
weekly_avg_fre: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, Avg_Rev_per_Order: double]

scala> weekly_avg_fre.show()
+----------+-----------------+
|WeekNumber|Avg_Rev_per_Order|
+----------+-----------------+
|         1|           21.869|
|         2|            22.24|
|         3|           21.588|
|         4|           21.808|
|         5|           21.738|
|         6|           21.306|
|         7|           21.871|
|         8|           22.863|
|         9|           21.608|
|        10|           21.881|
|        11|           21.646|
|        12|           22.257|
|        13|           21.792|
|        14|           21.503|
|        15|           21.064|
|        16|           21.736|
|        17|           21.162|
|        18|           21.233|
|        19|           21.614|
|        20|           22.053|
+----------+-----------------+
only showing top 20 rows
```

converting the three ‘timestamp’ columns to unified format:

```
scala> var df2 = date_df.withColumn("WeekNumber", weekofyear(col("PurchaseDate"))).withColumn("order_purchase_timestamp", to_timestamp(col("order_purchase_timestamp"), "dd/MM/yy H:mm")).withColumn("order_aproved_at", to_timestamp(
col("order_aproved_at"), "dd/MM/yy H:mm")).withColumn("order_delivered_customer_date", to_timestamp(col("order_delivered_customer_date"), "dd/MM/yy H:mm"))
df2: org.apache.spark.sql.DataFrame = [id: int, order_status: string ... 17 more fields]

scala> df2.show()
+---+------------+--------------------+-------------------+---------------+-------------------+--------------+-----------------------+-------------------+--------------------------+------------------+------------+---------------
--------+-------------------+-----------------------------+------------+-----------+------------+----------+
| id|order_status|order_products_value|order_freight_value|order_items_qty|      customer_city|customer_state|customer_zip_code_prefix|product_name_length|product_description_length|product_photos_qty|review_score|order_purchase_t
imestamp|   order_aproved_at|order_delivered_customer_date|PurchaseDate|ApproveDate|DeliveryDate|WeekNumber|
+---+------------+--------------------+-------------------+---------------+-------------------+--------------+-----------------------+-------------------+--------------------------+------------------+------------+---------------
--------+-------------------+-----------------------------+------------+-----------+------------+----------+
|  1|   delivered|                79.0|               17.8|              1|           Luziania|            GO|                    728|                 50|                       201|                 2|           5|     2017-10-02
10:56:00|2017-10-02 11:07:00|          2017-10-10 21:25:00|  2017-10-02| 2017-10-02|  2017-10-10|        40|
|  2|   delivered|               119.9|              27.16|              1|          Joinville|            SC|                    892|                 50|                       511|                 3|           5|     2018-07-24
20:41:00|2018-07-26 03:24:00|          2018-08-07 15:27:00|  2018-07-24| 2018-07-26|  2018-08-07|        30|
|  3|   delivered|              519.99|              41.69|              1|              Serra|            ES|                    291|                 48|                      1156|                 2|           1|     2018-08-08
08:38:00|2018-08-08 08:55:00|          2018-08-17 18:06:00|  2018-08-08| 2018-08-08|  2018-08-17|        32|
|  4|   delivered|                29.5|              17.92|              1|     RIO DE JANEIRO|            RJ|                    222|                 21|                       207|                 2|           4|     2017-11-18
19:28:00|2017-11-18 19:45:00|          2017-12-02 00:28:00|  2017-11-18| 2017-11-18|  2017-12-02|        46|
|  5|   delivered|               26.77|              23.11|              1|          Sao Paulo|            SP|                     40|                 41|                       451|                 1|           5|     2018-02-13
21:18:00|2018-02-13 22:20:00|          2018-02-16 18:17:00|  2018-02-13| 2018-02-13|  2018-02-16|         7|
|  6|   delivered|               419.9|              23.02|              1|       Santa Adelia|            SP|                    159|                 46|                       692|                 3|           5|     2017-07-09
21:57:00|2017-07-09 22:10:00|          2017-07-26 10:57:00|  2017-07-09| 2017-07-09|  2017-07-26|        27|
|  7|   delivered|                65.0|              16.21|              1|           Varginha|            MG|                    370|                 47|                       893|                 3|           5|     2017-04-11
12:22:00|2017-04-13 13:25:00|          2017-04-13 13:25:00|  2017-04-11| 2017-04-13|  2017-04-13|        15|
|  8|   delivered|               29.99|              19.82|              1|          Sao Paulo|            SP|                     48|                 55|                       613|                 1|           4|     2017-05-16
13:10:00|2017-05-16 13:22:00|          2017-05-26 12:55:00|  2017-05-16| 2017-05-16|  2017-05-26|        20|
|  9|   delivered|               59.99|              51.14|              1|            Carajas|            PA|                    685|                 28|                        96|                 4|           5|     2017-01-23
18:29:00|2017-01-25 02:50:00|          2017-02-02 14:08:00|  2017-01-23| 2017-01-25|  2017-02-02|         4|
| 10|   delivered|               56.99|              16.13|              1|            Resende|            RJ|                    275|                 59|                       641|                 3|           5|     2017-07-29
11:55:00|2017-07-29 12:05:00|          2017-08-16 17:14:00|  2017-07-29| 2017-07-29|  2017-08-16|        30|
| 11|   delivered|               599.0|              15.69|              1|      Guaratingueta|            SP|                    125|                 43|                      1686|                 2|           4|     2017-05-16
19:41:00|2017-05-16 19:50:00|          2017-05-29 11:18:00|  2017-05-16| 2017-05-16|  2017-05-29|        20|
| 12|   delivered|                62.0|              16.19|              1|            Tijucas|            SC|                    882|                 50|                       285|                 1|           4|     2017-07-13
19:58:00|2017-07-13 20:10:00|          2017-07-19 14:04:00|  2017-07-13| 2017-07-13|  2017-07-19|        28|
| 13|   delivered|               250.0|              35.02|              1|           Mirassol|            SP|                    151|                 57|                       771|                 1|           1|     2018-06-07
```

Average time taken to approve the orders (order approved – order purchased) :

```
scala> var weekly_approval_diff = df2.withColumn("TimeDifference(InMinutes)", ((col("order_aproved_at").cast("long") - col("order_purchase_timestamp").cast("long")) / 60).cast(DoubleType)).groupBy("WeekNumber").agg(avg("TimeDiffer
ence(InMinutes)").as("WeeklyAverageTimeToApprove(Minutes)")).orderBy("WeekNumber")
weekly_approval_diff: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, WeeklyAverageTimeToApprove(Minutes): double]

scala> weekly_approval_diff.show()
+----------+-----------------------------------+
|WeekNumber|WeeklyAverageTimeToApprove(Minutes)|
+----------+-----------------------------------+
|         1|                  645.6601806810285|
|         2|                  646.3057630736392|
|         3|                  762.2146863844977|
|         4|                 498.90439672801637|
|         5|                   664.701686746988|
|         6|                  600.4473438956197|
|         7|                  559.6588465298142|
|         8|                  574.5076923076923|
|         9|                  557.2153526970955|
|        10|                  401.4129175946548|
|        11|                 386.83520936515083|
|        12|                  492.9088235294118|
|        13|                  590.9213098729227|
|        14|                  546.5223744292238|
|        15|                   619.062874251497|
|        16|                 1503.6458333333333|
|        17|                  802.8473282442748|
|        18|                  716.4544037412315|
|        19|                           516.1725|
|        20|                  532.5956381837683|
+----------+-----------------------------------+
only showing top 20 rows


scala> var weekly_avg_approve = weekly_approval_diff.agg(round(avg("WeeklyAverageTimeToApprove(Minutes)"),3).as("Weekly AVG Approval Time(InMinutes)")).show()
+-----------------------------------+
|Weekly AVG Approval Time(InMinutes)|
+-----------------------------------+
|                            629.027|
+-----------------------------------+

weekly_avg_approve: Unit = ()
```

Average order delivery time :

```
scala> var weekly_delivery_diff = df2.withColumn("TimeDifference(InHours)", ((col("order_delivered_customer_date").cast("long") - col("order_purchase_timestamp").cast("long")) / 3600).cast(DoubleType)).groupBy("WeekNumber").agg(avg("TimeDifference(InHours)").as("WeeklyAverageTimeToDeliver(Hours)")).orderBy("WeekNumber")
weekly_delivery_diff: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [WeekNumber: int, WeeklyAverageTimeToDeliver(Hours): double]

scala> weekly_delivery_diff.show()
+----------+-------------------------------+
|WeekNumber|WeeklyAverageTimeToDeliver(Hours)|
+----------+-------------------------------+
|         1|              328.11180406212657|
|         2|               333.7524894746478|
|         3|               328.2062041884815|
|         4|               335.1090483757061|
|         5|               358.5847804391221|
|         6|              371.29964521851343|
|         7|               372.8246552304069|
|         8|              409.42559897089575|
|         9|              427.27816314806836|
|        10|               403.0652822642667|
|        11|               383.0805335199878|
|        12|              323.54522983521264|
|        13|              328.86308505284353|
|        14|              299.66362570356443|
|        15|               296.0002390710383|
|        16|               301.6511227431345|
|        17|               296.3451316550927|
|        18|              277.55423344017095|
|        19|               257.7608514292731|
|        20|               279.1140763711543|
+----------+-------------------------------+
only showing top 20 rows


scala> var weekly_avg_delivery = weekly_delivery_diff.agg(round(avg("WeeklyAverageTimeToDeliver(Hours)"),3).as("Weekly AVG Delivery Time(InHours)")).show()
+--------------------------------+
|Weekly AVG Delivery Time(InHours)|
+--------------------------------+
|                         303.926|
+--------------------------------+

weekly_avg_delivery: Unit = ()
```

c. Total freight charges :

```
scala> var weeklyFreight_df = weekly_df.groupBy("WeekNumber").agg(round(sum(col("order_freight_value")),2).as("Total_Freight_Charges")).orderBy("WeekNumber").show()
+----------+---------------------+
|WeekNumber|Total_Freight_Charges|
+----------+---------------------+
|         1|             31469.52|
|         2|             41699.72|
|         3|              42398.0|
|         4|             42656.18|
|         5|              45150.2|
|         6|             45765.64|
|         7|             45054.13|
|         8|             49041.98|
|         9|             52119.31|
|        10|             49145.77|
|        11|             48075.61|
|        12|             52992.88|
|        13|             44608.61|
|        14|             47091.58|
|        15|             42233.19|
|        16|              49080.8|
|        17|             49920.09|
|        18|             54567.61|
|        19|             60562.31|
|        20|             61682.45|
+----------+---------------------+
only showing top 20 rows

weeklyFreight_df: Unit = ()
```

d. Distribution of freight charges in each city :

```
scala> var weeklyFreight_city_df = weekly_df.groupBy(col("WeekNumber"),col("customer_city")).agg(round(sum(col("order_freight_value")),2).as("City_Freight_Charges")).orderBy("WeekNumber").show()
+----------+--------------------+--------------------+
|WeekNumber|       customer_city|City_Freight_Charges|
+----------+--------------------+--------------------+
|         1|Ferraz de Vasconc...|               36.81|
|         1|           Americana|               93.16|
|         1|  Balneario Camboriu|               42.56|
|         1|      RIO DE JANEIRO|             2637.07|
|         1|               Belem|              176.16|
|         1|         Santa Maria|              149.99|
|         1|     Itaquaquecetuba|               80.58|
|         1| Sao Miguel Do Oeste|               45.63|
|         1|      Cachoeira Alta|               15.66|
|         1|   Sao Jose do Egito|               24.84|
|         1|Vitoria da Conquista|               39.54|
|         1|           Papucaia|               16.21|
|         1|    Fernando Prestes|               10.96|
|         1|    Pontes e Lacerda|               20.01|
|         1|            Itaocara|               18.09|
|         1|       Rondonopolis|               16.79|
|         1|         Bataguassu|               18.23|
|         1| Sao Jose dos Campos|               73.56|
|         1|       Ilha Solteira|                49.3|
|         1|             Cajamar|               26.65|
+----------+--------------------+--------------------+
only showing top 20 rows

weeklyFreight_city_df: Unit = ()
```