ETL_task.py script is located in GitHub repository - https://github.com/azhumaba0727/spark-task

Following is documentation of some steps.

Step 1: Check restaurant data for incorrect (null) values (latitude and longitude). For incorrect values, map latitude and longitude from the OpenCage Geocoding API in a job via the REST API.

- filtering for incorrect (null) values for columns “lat” and “lng” was implemented and returned one result

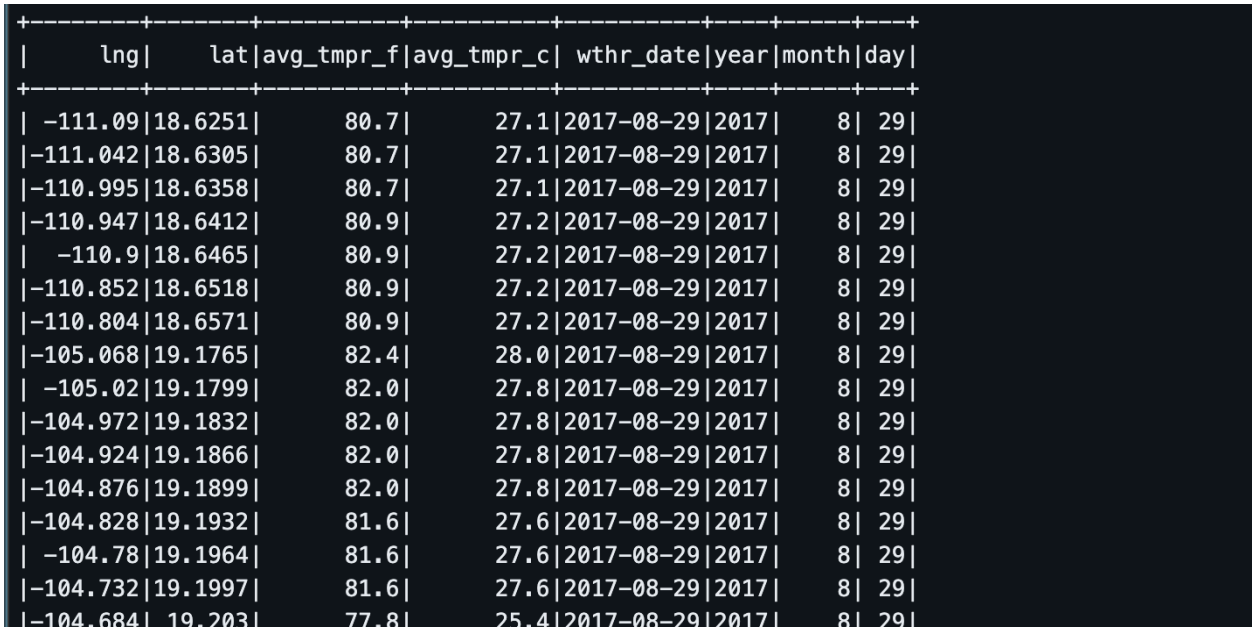
```
+-----------+------------+--------------+----------------------+-------+------+----+----+
|         id|franchise_id|franchise_name|restaurant_franchise_id|country|  city| lat| lng|
+-----------+------------+--------------+----------------------+-------+------+----+----+
|85899345920|           1|       Savoria|                 18952|     US|Dillon|null|null|
+-----------+------------+--------------+----------------------+-------+------+----+----+
```

- Yet, while reviewing documentation from OpenCage Geocoding API to implement ETL job, which will map latitude and longitude from available information seems impossible. It requires an address, and there is no address in the available data.
- For purpose of this task, information from columns “franhcise_name, “country” and “city” will be used as an address in ETL job and new column will be created.
- After “lat” and “lang” is fetched from Geocoding API, “address” column is dropped. Tables are merged.

Step 2: Implementing geohash - result

```
+-----------+-----------+------------------+----------------------+-------+-------------+------+-------+-------+
|         id|franchise_id|      franchise_name|restaurant_franchise_id|country|         city|   lat|    lng|geohash|
+-----------+-----------+------------------+----------------------+-------+-------------+------+-------+-------+
|197568495625|         10|    The Golden Spoon|                 24784|     US|      Decatur|34.578|-87.021|   dn4h|
| 17179869242|         59|         Azalea Cafe|                 10902|     FR|        Paris|48.861|  2.368|   u09t|
|214748364826|         27|     The Corner Cafe|                 92040|     US|   Rapid City| 44.08|-103.25|   9xyd|
|154618822706|         51|        The Pizzeria|                 41484|     AT|       Vienna|48.213| 16.413|   u2ed|
|163208757312|         65|       Chef's Corner|                 96638|     GB|       London|51.495| -0.191|   gcpu|
| 68719476763|         28|    The Spicy Pickle|                 77517|     US|     Grayling|44.657|-84.744|   dpgw|
|223338299419|         28|    The Spicy Pickle|                 36937|     US|       Oswego|43.452|-76.532|   dr9x|
|240518168650|         75|     Greenhouse Cafe|                 93164|     NL|    Amsterdam| 52.37|  4.897|   u173|
|128849018936|         57|The Yellow Submarine|                  5679|     FR|        Paris|48.872|  2.335|   u09w|
|197568495635|         20|       The Brasserie|                 24784|     US|Jeffersonville|39.616|-83.612|   dph9|
+-----------+-----------+------------------+----------------------+-------+-------------+------+-------+-------+
only showing top 10 rows
```

Step 3: Weather dataset - initial dataset



lng	lat	avg_tmpr_f	avg_tmpr_c	wthr_date	year	month	day
-111.09	18.6251	80.7	27.1	2017-08-29	2017	8	29
-111.042	18.6305	80.7	27.1	2017-08-29	2017	8	29
-110.995	18.6358	80.7	27.1	2017-08-29	2017	8	29
-110.947	18.6412	80.9	27.2	2017-08-29	2017	8	29
-110.9	18.6465	80.9	27.2	2017-08-29	2017	8	29
-110.852	18.6518	80.9	27.2	2017-08-29	2017	8	29
-110.804	18.6571	80.9	27.2	2017-08-29	2017	8	29
-105.068	19.1765	82.4	28.0	2017-08-29	2017	8	29
-105.02	19.1799	82.0	27.8	2017-08-29	2017	8	29
-104.972	19.1832	82.0	27.8	2017-08-29	2017	8	29
-104.924	19.1866	82.0	27.8	2017-08-29	2017	8	29
-104.876	19.1899	82.0	27.8	2017-08-29	2017	8	29
-104.828	19.1932	81.6	27.6	2017-08-29	2017	8	29
-104.78	19.1964	81.6	27.6	2017-08-29	2017	8	29
-104.732	19.1997	81.6	27.6	2017-08-29	2017	8	29
-104.684	19.203	77.8	25.4	2017-08-29	2017	8	29

After adding geohash column:

id	franchise_id	franchise_name	restaurant_franchise_id	country	city	lat	lng	geohash
197568495625	10	The Golden Spoon	24784	US	Decatur	34.578	-87.021	dn4h
17179869242	59	Azalea Cafe	10902	FR	Paris	48.861	2.368	u09t
214748364826	27	The Corner Cafe	92040	US	Rapid City	44.08	-103.25	9xyd
154618822706	51	The Pizzeria	41484	AT	Vienna	48.213	16.413	u2ed
163208757312	65	Chef's Corner	96638	GB	London	51.495	-0.191	gcpu
68719476763	28	The Spicy Pickle	77517	US	Grayling	44.657	-84.744	dpgw
223338299419	28	The Spicy Pickle	36937	US	Oswego	43.452	-76.532	dr9x
240518168650	75	Greenhouse Cafe	93164	NL	Amsterdam	52.37	4.897	u173
128849018936	57	The Yellow Submarine	5679	FR	Paris	48.872	2.335	u09w
197568495635	20	The Brasserie	24784	US	Jeffersonville	39.616	-83.612	dph9

only showing top 10 rows

Schema of final dataset:

```
# Apply renaming to the DataFrame
renamed_df = joined_df.toDF(*renamed_columns)
```

```
renamed_df: pyspark.sql.dataframe.DataFrame
  geohash: string
  id: string
  franchise_id: string
  franchise_name: string
  restaurant_franchise_id: string
  country: string
  city: string
  lat: float
  lng: float
  lng_wthr: float
  lat_wthr: float
  avg_tmpr_f: double
  avg_tmpr_c: double
  wthr_date: string
  year: integer
  month: integer
  day: integer
```