Check Raw Data

The data needs to cleaned, transformed before loading into RDMS. For each raw file, checking of null values, duplicate values and datatypes are checked. All raw files are kept as it is and loaded into RDMS after data explore.

In [1]:
```
# Import libraries
import numpy as np
import pandas as pd
```

In [2]:
```
# Read Fulfilment Center Info file
data_center = pd.read_csv('../fulfilment_center_info.csv')
data_center.head()
```

Out[2]:

	center_id	city_code	region_code	center_type	op_area
0	11	679	56	TYPE_A	3.7
1	13	590	56	TYPE_B	6.7
2	124	590	56	TYPE_C	4.0
3	66	648	34	TYPE_A	4.1
4	94	632	34	TYPE_C	3.6

In [3]:
```
# Check DataType for fulfilment center
data_center.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 77 entries, 0 to 76
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   center_id    77 non-null     int64  
 1   city_code    77 non-null     int64  
 2   region_code  77 non-null     int64  
 3   center_type  77 non-null     object 
 4   op_area      77 non-null     float64
dtypes: float64(1), int64(3), object(1)
memory usage: 3.1+ KB
```

In [4]:
```
# Count the total Null values for fulfilment center
data_center.isnull().sum()
```

Out[4]:
```
center_id      0
city_code      0
region_code    0
center_type    0
op_area        0
dtype: int64
```

In [5]:
```
# Check the duplicates for fulfilment center
data_center.duplicated().sum()
```

Out[5]:
```
0
```

In [6]:
```
# Read Meal Info file
data_meal = pd.read_csv('../meal_info.csv')
data_meal.head()
```

Out[6]:

	meal_id	category	cuisine
0	1885	Beverages	Thai
1	1993	Beverages	Thai
2	2539	Beverages	Thai
3	1248	Beverages	Indian
4	2631	Beverages	Indian

In [7]:
```
# Check DataType for Meal Info
data_meal.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   meal_id   51 non-null     int64 
 1   category  51 non-null     object
 2   cuisine   51 non-null     object
dtypes: int64(1), object(2)
memory usage: 1.3+ KB
```

In [8]:
```
# Count the total Null values for Meal Info
data_meal.isnull().sum()
```

Out[8]:
```
meal_id     0
category    0
cuisine     0
dtype: int64
```

In [9]:
```
# Check the duplicates for Meal Info
data_meal.duplicated().sum()
```

Out[9]:
```
0
```

In [10]:
```
# Read Train file
data_train = pd.read_csv('../train.csv')
data_train.head()
```

Out[10]:

	id	week	center_id	meal_id	checkout_price	base_price	emailer_for_promotion	homepage_featured
0	1379560	1	55	1885	136.83	152.29	0	0
1	1466964	1	55	1993	136.83	135.83	0	0
2	1346989	1	55	2539	134.86	135.86	0	0
3	1338232	1	55	2139	339.50	437.53	0	0
4	1448490	1	55	2631	243.50	242.50	0	0

In [11]:
```
# Check DataType for Train
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 456548 entries, 0 to 456547
Data columns (total 9 columns):
 #   Column                Non-Null Count   Dtype 
---  ------                --------------   ----- 
 0   id                    456548 non-null  int64 
 1   week                  456548 non-null  int64 
 2   center_id             456548 non-null  int64 
 3   meal_id               456548 non-null  int64 
```

```
 4   checkout_price         456548 non-null  float64
 5   base_price             456548 non-null  float64
 6   emailer_for_promotion  456548 non-null  int64
 7   homepage_featured      456548 non-null  int64
 8   num_orders             456548 non-null  int64
dtypes: float64(2), int64(7)
memory usage: 31.3 MB
```

In [12]:
```
# Count the total Null values for train data
data_train.isnull().sum()
```

Out[12]:
```
id                     0
week                   0
center_id              0
meal_id                0
checkout_price         0
base_price             0
emailer_for_promotion  0
homepage_featured      0
num_orders             0
dtype: int64
```

In [13]:
```
# Check the duplicates for train data
data_train.duplicated().sum()
```

Out[13]:
```
0
```

In [14]:
```
# Read Test file
data_test = pd.read_csv('../test.csv')
data_test.head()
```

Out[14]:

	id	week	center_id	meal_id	checkout_price	base_price	emailer_for_promotion	homepage_featured
0	1028232	146	55	1885	158.11	159.11	0	0
1	1127204	146	55	1993	160.11	159.11	0	0
2	1212707	146	55	2539	157.14	159.14	0	0
3	1082698	146	55	2631	162.02	162.02	0	0
4	1400926	146	55	1248	163.93	163.93	0	0

In [15]:
```
# Check DataType for Test data
data_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32573 entries, 0 to 32572
Data columns (total 8 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id                     32573 non-null  int64
 1   week                   32573 non-null  int64
 2   center_id              32573 non-null  int64
 3   meal_id                32573 non-null  int64
 4   checkout_price         32573 non-null  float64
 5   base_price             32573 non-null  float64
 6   emailer_for_promotion  32573 non-null  int64
 7   homepage_featured      32573 non-null  int64
dtypes: float64(2), int64(6)
memory usage: 2.0 MB
```

In [16]:
```
# Count the total Null values for test data
data_test.isnull().sum()
```

Out[16]:
```
id                     0
week                   0
```

```
center_id               0
meal_id                 0
checkout_price          0
base_price              0
emailer_for_promotion   0
homepage_featured       0
dtype: int64
```