Importing Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Reading the dataset

```
train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
```

Exploratory data analysis

```
train.head()
```

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

```
test.head()
```

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

```
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
```

```
train['num_orders'].describe()
```

```
count    456548.000000
mean        261.872760
std         395.922798
min          13.000000
25%          54.000000
50%         136.000000
75%         324.000000
max       24299.000000
Name: num_orders, dtype: float64
```

Checking for null values

```
train.isnull().sum()
```

```
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
```

Reading and merging.csv files

```
meal_info = pd.read_csv("meal_info.csv")
center_info = pd.read_csv("fulfilment_center_info.csv")

trainfinal = pd.merge(train, meal_info, on="meal_id", how="outer")
trainfinal = pd.merge(trainfinal,center_info, on="center_id", how="outer")
trainfinal.head()
```

	id	week	center_id	meal_id	checkout_price	base_price	emailer_for_prom
0	1379560	1	55	1885	136.83	152.29	
1	1018704	2	55	1885	135.83	152.29	
2	1196273	3	55	1885	132.92	133.92	
3	1116527	4	55	1885	135.86	134.86	
4	1343872	5	55	1885	146.50	147.50	

```
trainfinal = trainfinal.drop(['center_id', 'meal_id'], axis=1)
trainfinal.head()
```

	id	week	checkout_price	base_price	emailer_for_promotion	homepage_feat
0	1379560	1	136.83	152.29	0	
1	1018704	2	135.83	152.29	0	
2	1196273	3	132.92	133.92	0	
3	1116527	4	135.86	134.86	0	
4	1343872	5	146.50	147.50	0	

Dropping columns

```
cols = trainfinal.columns.tolist()
print(cols)
```

```
['id', 'week', 'checkout_price', 'base_price', 'emailer_for_promotion', 'homepage_featur
```

```
cols = cols[:2] + cols[9:] + cols[7:9] + cols[2:7]
print(cols)
```

```
['id', 'week', 'city_code', 'region_code', 'center_type', 'op_area', 'category', 'cuisir
```

```
trainfinal = trainfinal[cols]
trainfinal.dtypes
```

```
id                        int64
week                      int64
city_code                 int64
region_code               int64
```

```
center_type               object
op_area                  float64
category                  object
cuisine                   object
checkout_price           float64
base_price               float64
emailer_for_promotion      int64
homepage_featured          int64
num_orders                 int64
dtype: object
```

```
from sklearn.preprocessing import LabelEncoder
```

Label encoding

```
trainfinal.head()
```

	id	week	city_code	region_code	center_type	op_area	category	cuisine
0	1379560	1	647	56	TYPE_C	2.0	Beverages	Thai
1	1018704	2	647	56	TYPE_C	2.0	Beverages	Thai
2	1196273	3	647	56	TYPE_C	2.0	Beverages	Thai
3	1116527	4	647	56	TYPE_C	2.0	Beverages	Thai
4	1343872	5	647	56	TYPE_C	2.0	Beverages	Thai

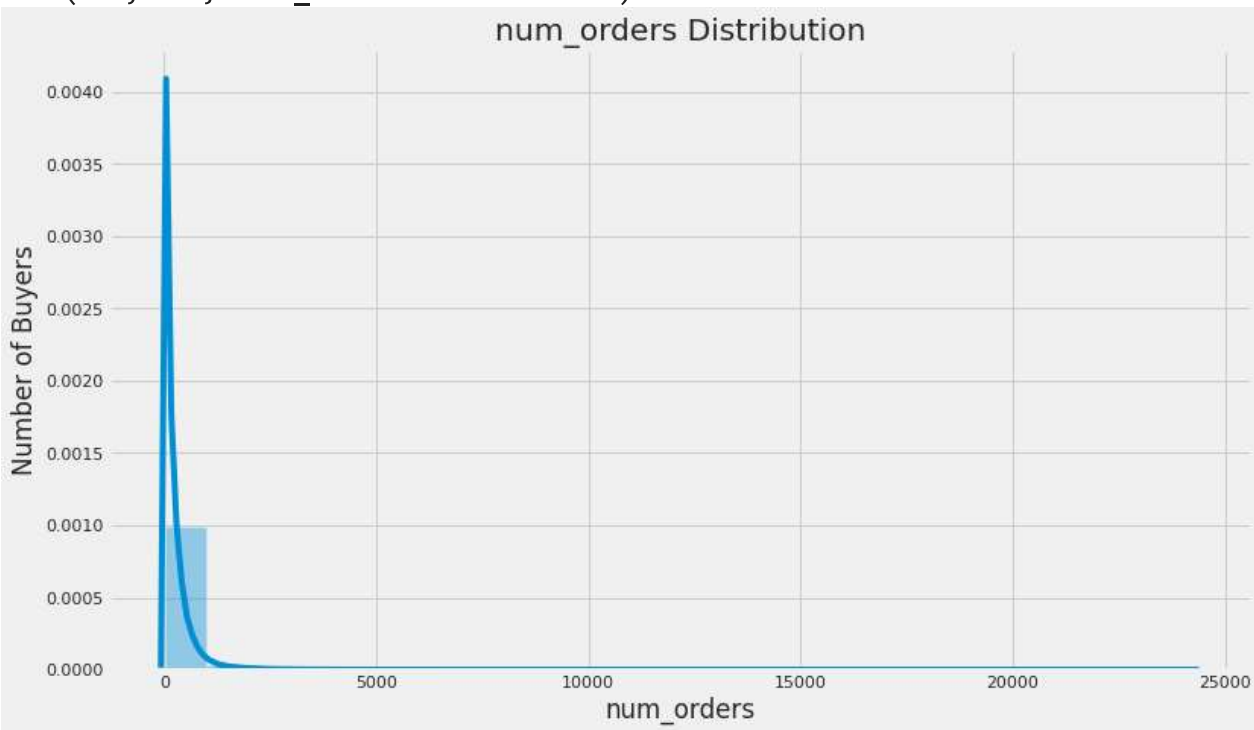
```
trainfinal.shape
```

```
(456548, 13)
```

data visualization

```
plt.style.use('fivethirtyeight')
plt.figure(figsize=(12,7))
sns.distplot(trainfinal.num_orders, bins = 25)
plt.xlabel("num_orders")
plt.ylabel("Number of Buyers")
plt.title("num_orders Distribution")
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarnin
  warnings.warn(msg, FutureWarning)
Text(0.5, 1.0, 'num_orders Distribution')
```

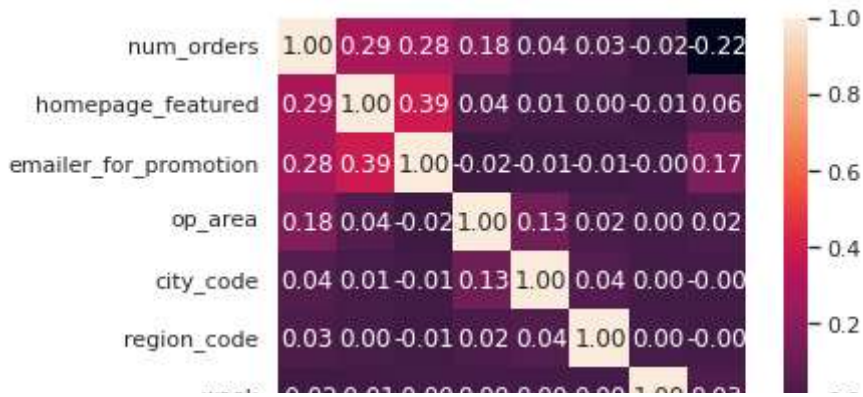



```
trainfinal2 = trainfinal.drop(['id'], axis=1)
correlation = trainfinal2.corr(method='pearson')
columns = correlation.nlargest(8, 'num_orders').index
columns
```

```
Index(['num_orders', 'homepage_featured', 'emailer_for_promotion', 'op_area',
       'city_code', 'region_code', 'week', 'base_price'],
      dtype='object')
```

```
correlation_map = np.corrcoef(trainfinal2[columns].values.T)
sns.set(font_scale=1.0)
heatmap = sns.heatmap(correlation_map, cbar=True, annot=True, square=True, fmt='.2f', ytickla

plt.show()
```



spliting the dataset into dependent and independent variable

```
features = columns.drop(['num_orders'])
trainfinal3 = trainfinal[features]
X = trainfinal3.values
Y = trainfinal['num_orders'].values
trainfinal3.head()
```

	homepage_featured	emailer_for_promotion	op_area	city_code	region_code	week
0	0	0	2.0	647	56	1
1	0	0	2.0	647	56	2
2	0	0	2.0	647	56	3
3	0	0	2.0	647	56	4
4	0	0	2.0	647	56	5

Split the dataset into train set and test set

```
from sklearn.model_selection import train_test_split
X_train, X_val, Y_train, Y_val = train_test_split(X, Y, test_size=0.25)
```