In [3]:
```
import pyarrow.parquet as pq
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import tarfile
import seaborn as sns
```

In [4]:
```
import pandas as pd
import pyarrow.parquet as pq
import pyarrow.compute as pc
import pyarrow as pa
import numpy as np


def sample_n_visitors (full, n):
    uniqueid = full["visitid"].unique().to_pandas()
    sample_nid = pa.array(uniqueid.sample(n))
    id_filter_idx = pc.is_in(full["visitid"], sample_nid)
    full_filtered = full.filter(id_filter_idx)
    return full_filtered

def read_full_week (cols, direc):
    week_dic = {}
    fileidx = np.arange(10, 16)
    for i in range(len(fileidx)):
        print('read: ' + direc + 'visitday=' + str(fileidx[i]))
        dataset = pq.ParquetDataset('Math-M148-data-january2023/hitdata7days/visitday=' + str(fileidx[i]))
        week_dic['visitday=' + str(fileidx[i])] = dataset.read(columns = cols)
    return week_dic
```

In [ ]:

In [ ]:

In [5]:
```
col = ['visitid', 'hit_time_gmt', 'pagename', 'productlist', 'ordernumber','visitdatetime','promocode']
visit_week = read_full_week(col, 'week/')
```

```
read: week/visitday=10
read: week/visitday=11
read: week/visitday=12
read: week/visitday=13
read: week/visitday=14
read: week/visitday=15
```

In [6]:
```
full_week = pd.DataFrame([],
                         columns=col)
for i in visit_week:
    full_week=pd.concat([full_week,visit_week[i].to_pandas()])
full_week
```

Out[6]:

	visitid	hit_time_gmt	pagename	productlist	ordernumber	visitdatetime
0	189763922254746751413200407250696891446	1670652028	mobile index	None	None	2022-12-10 00:00:28
1	224586517001559320985673092461365783711	1670692775	pdp:Emeril Lagasse French Door 26-Qt. Air Frye...	;NRWP5;;;;eVar1=Kitchen|eVar2=Kitchen:Kitchen ...	None	2022-12-10 11:19:35
2	813021380406890216086034508823121971521O	1670661057	checkout:1page	None	None	2022-12-10 02:30:57
3	590912592049203187912453158393064821062	1670675400	checkout:1page	None	None	2022-12-10 06:30:00
4	538021183841242432152115917402238466902	1670664191	index	None	None	2022-12-10 03:23:11
...	...	...	...	...	...	...
6004223	765252480325430649352432463936193695l6	1671170010	search:results	None	None	2022-12-15 23:53:30
6004224	395189371088209292079958646964557152O559	1671116626	mobile index	None	None	2022-12-15 09:03:46
6004225	68336520221177822661305432204360093483S	1671118574	cart	;NRTOH;1;49.99;;;;NRV1M;1;41.99;;;;NN93T;1;129...	None	2022-12-15 09:36:14
6004226	152080814150849848661185320580245477171	1671142465	search:results	None	None	2022-12-15 16:14:25
6004227	492642432931835217695149915839828693l4	1671144207	search:results	None	None	2022-12-15 16:43:27

37250107 rows × 7 columns

Purchased, and used Promo Code

In [7]:
```
transact = []
pro = []
for i,j in zip(full_week['ordernumber'],full_week['promocode']):
    if i:
        transact.append(1)

    else:
        transact.append(0)
    if j:
        pro.append(1)

    else:
        pro.append(0)

full_week['transact'] = transact
full_week['promo'] = pro
```

TimeBin

In [27]:
```
import datetime
def bin_f(x):

    for i in range(1,25):
        if i ==24:
            return str(24-1)
        elif x.time() < datetime.time(i):
            return str(i-1)


full_week["Bin"] = full_week["visitdatetime"].apply(bin_f)
```

In [43]:
```
full_week['visitdatetime']
```

Out[43]:
```
0          2022-12-10 00:00:28
1          2022-12-10 11:19:35
2          2022-12-10 02:30:57
3          2022-12-10 06:30:00
4          2022-12-10 03:23:11
                    ...
6004223    2022-12-15 23:53:30
6004224    2022-12-15 09:03:46
6004225    2022-12-15 09:36:14
6004226    2022-12-15 16:14:25
6004227    2022-12-15 16:43:27
Name: visitdatetime, Length: 37250107, dtype: datetime64[ns]
```

In [45]:
```
np.unique(full_week['date'])
```

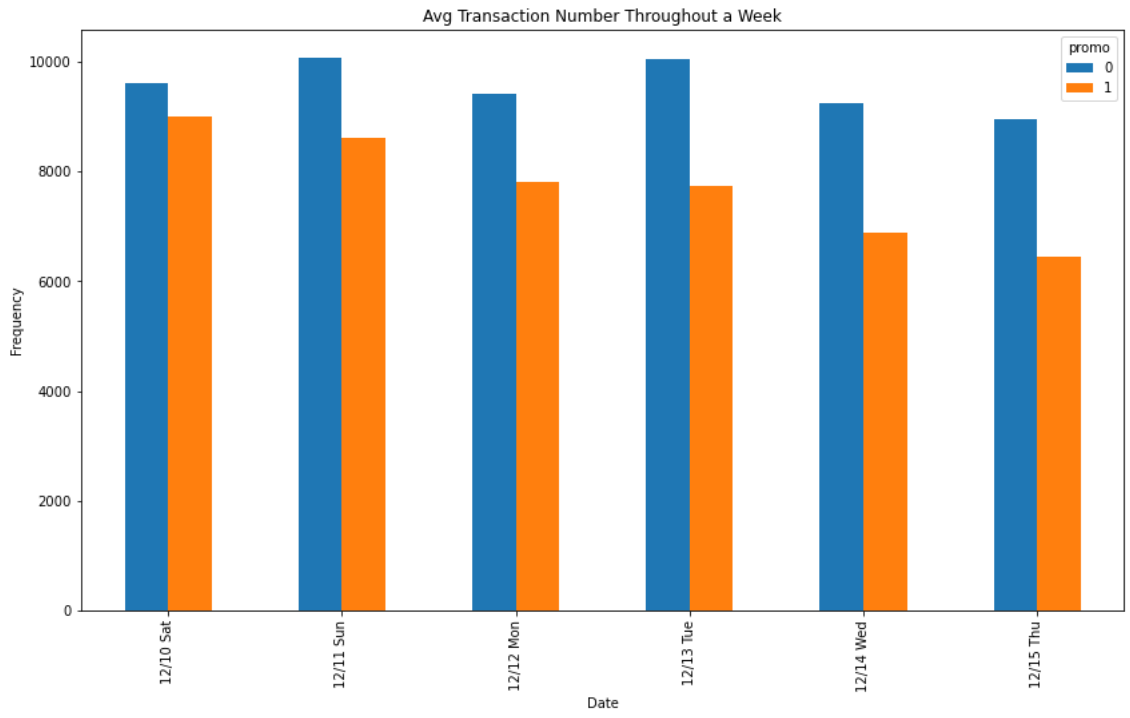
Out[45]:
```
array([datetime.date(2022, 12, 10), datetime.date(2022, 12, 11),
       datetime.date(2022, 12, 12), datetime.date(2022, 12, 13),
       datetime.date(2022, 12, 14), datetime.date(2022, 12, 15)],
      dtype=object)
```

In [44]:
```
full_week['visitdatetime'] = pd.to_datetime(full_week['visitdatetime'])
full_week['date'] = full_week['visitdatetime'].dt.date
```

In [48]:
```
full_week['date_str'] = full_week['date'].apply(lambda x: x.strftime("%m/%d %a"))
```

In [50]:
```
ax = df.plot(kind='bar', figsize=(14,8), title="Avg Transaction Number Throughout a Week")
ax.set_xlabel("Date")
ax.set_ylabel("Frequency")
```

Out[50]:
```
Text(0, 0.5, 'Frequency')
```



In [51]:
```
df = full_week.groupby(['Bin', 'promo'])['ordernumber'].count().unstack()
df.index = pd.to_numeric(df.index)
df = df.sort_index()
ax = df.plot(kind='bar', figsize=(14,8), title="Avg Transaction Number Throughout a day")
ax.set_xlabel("Time")
ax.set_ylabel("Frequency")
```

Out[51]:
```
Text(0, 0.5, 'Frequency')
```