B2.342 PEC2 TFM Corporacion Favorita

December 4, 2017

0.0.1 Trabajo Final de Master

1 Predicción de Ventas de Comestibles Corporación Favorita

2 Análisis Exploratorio de Datos

2.0.1 Autor Gabriel Kreplak

2.0.2 Diciembre 2017

2.0.3 PEC 2

Como actividad prevista en la planificación del TFM, este análisis exploratorio de datos, en adelante EDA, se centrará en la carga e inspección de los distintos ficheros incluídos en la competición. Este notebook se convertirá en un capítulo del trabajo final.
La estructura de este análisis será la siguiente:
1. Carga de datos e Inspección - Conversión de datos en formato csv a dataframes python y chequeo preliminar de integridad.
2. Exploración de datos auxiliares - Exploración de todos los archivos excepto finchero train.csv
3. Exploración de datos de entreno de la predicción

In [3]:
```
# Importing the relevant libraries
import IPython.display
import json
import pandas as pd
import sys
sys.path.append('/Users/gabrielkreplak/anaconda3/lib/python3.6/site-packages')
import seaborn as sns
import squarify
%matplotlib inline
import missingno as msno
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import numpy as np
from scipy.fftpack import fft
from matplotlib import pyplot as plt
```

```
# D3 modules
from IPython.core.display import display, HTML, Javascript
from string import Template
```

3 1. Carga de Datos e Inspección

En este apartado se cargará todos los ficheros de datos excepto train.csv, que tendrá su tratamiento personalizado.
Esto se debe a que el fichero consta de 125,5 millones de líneas con transacciones de las tiendas, lo que requeriría una capacidad de proceso muy excepcional. Popr ese motivo trabajaremos incialmente con un 5% (6 millones de líneas).

```
In [4]: items = pd.read_csv("../Favorita/input/items.csv")
        holiday_events = pd.read_csv("../Favorita/input/holidays_events.csv")
        stores = pd.read_csv("../Favorita/input/stores.csv")
        oil = pd.read_csv("../Favorita/input/oil.csv")
        transactions = pd.read_csv("../Favorita/input/transactions.csv",parse_dates=['date'])
        # I read in the full training data just to get prior information and here is the output:
        # Output: "125,497,040 rows | 6 columns"
        train = pd.read_csv("../Favorita/input/train.csv", nrows=6000000, parse_dates=['date'])
```

```
In [26]: print("There are {0} rows and {1} columns in the items data".
               format(items.shape[0], items.shape[1]))

         print("There are {0} rows and {1} columns in the holiday_events data".
               format(holiday_events.shape[0], holiday_events.shape[1]))

         print("There are {0} rows and {1} columns in the stores data".
               format(stores.shape[0], stores.shape[1]))

         print("There are {0} rows and {1} columns in the oil data".
               format(oil.shape[0], oil.shape[1]))

         print("There are {0} rows and {1} columns in the transactions data".
               format(transactions.shape[0], transactions.shape[1]))

         print("There are {0} rows and {1} columns in the train data".
               format(train.shape[0], train.shape[1]))
```

```
There are 4100 rows and 4 columns in the items data
There are 350 rows and 6 columns in the holiday_events data
There are 54 rows and 5 columns in the stores data
There are 1218 rows and 2 columns in the oil data
There are 83488 rows and 3 columns in the transactions data
There are 6000000 rows and 6 columns in the train data
```

In [6]:
```
## Archivo train
## Cada línea = 1 transacción. 6 columnas: Fecha, Item, numero, y tienda
train.head()
```

Out[6]:

	id	date	store_nbr	item_nbr	unit_sales	onpromotion
0	0	2013-01-01	25	103665	7.0	NaN
1	1	2013-01-01	25	105574	1.0	NaN
2	2	2013-01-01	25	105575	2.0	NaN
3	3	2013-01-01	25	108079	1.0	NaN
4	4	2013-01-01	25	108701	1.0	NaN

In [7]:
```
## Archivo Transaccions
## Numero de transacciones por día y tienda desde 2013
transactions
```

Out[7]:

	date	store_nbr	transactions
0	2013-01-01	25	770
1	2013-01-02	1	2111
2	2013-01-02	2	2358
3	2013-01-02	3	3487
4	2013-01-02	4	1922
5	2013-01-02	5	1903
6	2013-01-02	6	2143
7	2013-01-02	7	1874
8	2013-01-02	8	3250
9	2013-01-02	9	2940
10	2013-01-02	10	1293
11	2013-01-02	11	3547
12	2013-01-02	12	1362
13	2013-01-02	13	1102
14	2013-01-02	14	2002
15	2013-01-02	15	1622
16	2013-01-02	16	1167
17	2013-01-02	17	1580
18	2013-01-02	18	1635
19	2013-01-02	19	1369
20	2013-01-02	23	1381
21	2013-01-02	24	2605
22	2013-01-02	25	1038
23	2013-01-02	26	1008
24	2013-01-02	27	1386
25	2013-01-02	28	950
26	2013-01-02	30	708
27	2013-01-02	31	1401
28	2013-01-02	32	776
29	2013-01-02	33	1163
...	...	...	...
83458	2017-08-15	25	849

```
83459 2017-08-15         26          534
83460 2017-08-15         27         1543
83461 2017-08-15         28         1343
83462 2017-08-15         29         1302
83463 2017-08-15         30          825
83464 2017-08-15         31         1360
83465 2017-08-15         32          615
83466 2017-08-15         33          919
83467 2017-08-15         34         2007
83468 2017-08-15         35          612
83469 2017-08-15         36         1192
83470 2017-08-15         37         1373
83471 2017-08-15         38         1445
83472 2017-08-15         39         1425
83473 2017-08-15         40         1392
83474 2017-08-15         41         1003
83475 2017-08-15         42          995
83476 2017-08-15         43         1482
83477 2017-08-15         44         3815
83478 2017-08-15         45         3685
83479 2017-08-15         46         3197
83480 2017-08-15         47         3581
83481 2017-08-15         48         2722
83482 2017-08-15         49         2814
83483 2017-08-15         50         2804
83484 2017-08-15         51         1573
83485 2017-08-15         52         2255
83486 2017-08-15         53          932
83487 2017-08-15         54          802

[83488 rows x 3 columns]
```

In [8]:
```
# Días de fiesta locales y nacionales
holiday_events
```

Out[8]:

	date	type	locale	locale_name
0	2012-03-02	Holiday	Local	Manta
1	2012-04-01	Holiday	Regional	Cotopaxi
2	2012-04-12	Holiday	Local	Cuenca
3	2012-04-14	Holiday	Local	Libertad
4	2012-04-21	Holiday	Local	Riobamba
5	2012-05-12	Holiday	Local	Puyo
6	2012-06-23	Holiday	Local	Guaranda
7	2012-06-25	Holiday	Regional	Imbabura
8	2012-06-25	Holiday	Local	Latacunga
9	2012-06-25	Holiday	Local	Machala
10	2012-07-03	Holiday	Local	Santo Domingo
11	2012-07-03	Holiday	Local	El Carmen

12	2012-07-23	Holiday	Local	Cayambe
13	2012-08-05	Holiday	Local	Esmeraldas
14	2012-08-10	Holiday	National	Ecuador
15	2012-08-15	Holiday	Local	Riobamba
16	2012-08-24	Holiday	Local	Ambato
17	2012-09-28	Holiday	Local	Ibarra
18	2012-10-07	Holiday	Local	Quevedo
19	2012-10-09	Holiday	National	Ecuador
20	2012-10-12	Transfer	National	Ecuador
21	2012-11-02	Holiday	National	Ecuador
22	2012-11-03	Holiday	National	Ecuador
23	2012-11-06	Holiday	Regional	Santo Domingo de los Tsachilas
24	2012-11-07	Holiday	Regional	Santa Elena
25	2012-11-10	Holiday	Local	Guaranda
26	2012-11-11	Holiday	Local	Latacunga
27	2012-11-12	Holiday	Local	Ambato
28	2012-12-05	Additional	Local	Quito
29	2012-12-06	Holiday	Local	Quito
..	...	...	...	...
320	2017-07-23	Holiday	Local	Cayambe
321	2017-07-24	Additional	Local	Guayaquil
322	2017-07-25	Additional	Local	Guayaquil
323	2017-08-05	Holiday	Local	Esmeraldas
324	2017-08-10	Holiday	National	Ecuador
325	2017-08-11	Transfer	National	Ecuador
326	2017-08-15	Holiday	Local	Riobamba
327	2017-08-24	Holiday	Local	Ambato
328	2017-09-28	Holiday	Local	Ibarra
329	2017-09-29	Transfer	Local	Ibarra
330	2017-10-07	Holiday	Local	Quevedo
331	2017-10-09	Holiday	National	Ecuador
332	2017-11-02	Holiday	National	Ecuador
333	2017-11-03	Holiday	National	Ecuador
334	2017-11-06	Holiday	Regional	Santo Domingo de los Tsachilas
335	2017-11-07	Holiday	Regional	Santa Elena
336	2017-11-10	Holiday	Local	Guaranda
337	2017-11-11	Holiday	Local	Latacunga
338	2017-11-12	Holiday	Local	Ambato
339	2017-12-05	Additional	Local	Quito
340	2017-12-06	Holiday	Local	Quito
341	2017-12-08	Holiday	Local	Loja
342	2017-12-08	Transfer	Local	Quito
343	2017-12-21	Additional	National	Ecuador
344	2017-12-22	Holiday	Local	Salinas
345	2017-12-22	Additional	National	Ecuador
346	2017-12-23	Additional	National	Ecuador
347	2017-12-24	Additional	National	Ecuador
348	2017-12-25	Holiday	National	Ecuador

```
349  2017-12-26  Additional  National                         Ecuador

                                description  transferred
0                         Fundacion de Manta        False
1               Provincializacion de Cotopaxi        False
2                         Fundacion de Cuenca        False
3                   Cantonizacion de Libertad        False
4                   Cantonizacion de Riobamba        False
5                      Cantonizacion del Puyo        False
6                   Cantonizacion de Guaranda        False
7               Provincializacion de Imbabura        False
8                  Cantonizacion de Latacunga        False
9                        Fundacion de Machala        False
10                 Fundacion de Santo Domingo        False
11                 Cantonizacion de El Carmen        False
12                   Cantonizacion de Cayambe        False
13                    Fundacion de Esmeraldas        False
14              Primer Grito de Independencia        False
15                      Fundacion de Riobamba        False
16                        Fundacion de Ambato        False
17                        Fundacion de Ibarra        False
18                   Cantonizacion de Quevedo        False
19                Independencia de Guayaquil         True
20      Traslado Independencia de Guayaquil        False
21                             Dia de Difuntos        False
22                   Independencia de Cuenca        False
23       Provincializacion de Santo Domingo        False
24              Provincializacion Santa Elena        False
25                  Independencia de Guaranda        False
26                Independencia de Latacunga        False
27                   Independencia de Ambato        False
28                      Fundacion de Quito-1        False
29                        Fundacion de Quito        False
..                                       ...          ...
320                  Cantonizacion de Cayambe        False
321                  Fundacion de Guayaquil-1        False
322                    Fundacion de Guayaquil        False
323                  Fundacion de Esmeraldas        False
324             Primer Grito de Independencia         True
325  Traslado Primer Grito de Independencia        False
326                     Fundacion de Riobamba        False
327                       Fundacion de Ambato        False
328                       Fundacion de Ibarra         True
329                       Fundacion de Ibarra        False
330                  Cantonizacion de Quevedo        False
331                Independencia de Guayaquil        False
332                            Dia de Difuntos        False
333                  Independencia de Cuenca        False
```

```
334      Provincializacion de Santo Domingo        False
335           Provincializacion Santa Elena        False
336               Independencia de Guaranda        False
337              Independencia de Latacunga        False
338                 Independencia de Ambato        False
339                    Fundacion de Quito-1        False
340                      Fundacion de Quito         True
341                       Fundacion de Loja        False
342             Traslado Fundacion de Quito        False
343                                Navidad-4        False
344                Cantonizacion de Salinas        False
345                                Navidad-3        False
346                                Navidad-2        False
347                                Navidad-1        False
348                                  Navidad        False
349                                Navidad+1        False

[350 rows x 6 columns]
```

In [9]:
```
#cotización del PEtróleo por día desde 2013
oil
```

Out[9]:

	date	dcoilwtico
0	2013-01-01	NaN
1	2013-01-02	93.14
2	2013-01-03	92.97
3	2013-01-04	93.12
4	2013-01-07	93.20
5	2013-01-08	93.21
6	2013-01-09	93.08
7	2013-01-10	93.81
8	2013-01-11	93.60
9	2013-01-14	94.27
10	2013-01-15	93.26
11	2013-01-16	94.28
12	2013-01-17	95.49
13	2013-01-18	95.61
14	2013-01-21	NaN
15	2013-01-22	96.09
16	2013-01-23	95.06
17	2013-01-24	95.35
18	2013-01-25	95.15
19	2013-01-28	95.95
20	2013-01-29	97.62
21	2013-01-30	97.98
22	2013-01-31	97.65
23	2013-02-01	97.46
24	2013-02-04	96.21

```
25    2013-02-05       96.68
26    2013-02-06       96.44
27    2013-02-07       95.84
28    2013-02-08       95.71
29    2013-02-11       97.01
...          ...         ...
1188  2017-07-21       45.78
1189  2017-07-24       46.21
1190  2017-07-25       47.77
1191  2017-07-26       48.58
1192  2017-07-27       49.05
1193  2017-07-28       49.72
1194  2017-07-31       50.21
1195  2017-08-01       49.19
1196  2017-08-02       49.60
1197  2017-08-03       49.03
1198  2017-08-04       49.57
1199  2017-08-07       49.37
1200  2017-08-08       49.07
1201  2017-08-09       49.59
1202  2017-08-10       48.54
1203  2017-08-11       48.81
1204  2017-08-14       47.59
1205  2017-08-15       47.57
1206  2017-08-16       46.80
1207  2017-08-17       47.07
1208  2017-08-18       48.59
1209  2017-08-21       47.39
1210  2017-08-22       47.65
1211  2017-08-23       48.45
1212  2017-08-24       47.24
1213  2017-08-25       47.65
1214  2017-08-28       46.40
1215  2017-08-29       46.46
1216  2017-08-30       45.96
1217  2017-08-31       47.26

[1218 rows x 2 columns]
```

In [10]:
```
# Items: Catálogo y agrupación de productos a la venta. Casa con train.
items
```

Out[10]:

	item_nbr	family	class	perishable
0	96995	GROCERY I	1093	0
1	99197	GROCERY I	1067	0
2	103501	CLEANING	3008	0
3	103520	GROCERY I	1028	0
4	103665	BREAD/BAKERY	2712	1

5	105574	GROCERY I	1045	0
6	105575	GROCERY I	1045	0
7	105576	GROCERY I	1045	0
8	105577	GROCERY I	1045	0
9	105693	GROCERY I	1034	0
10	105737	GROCERY I	1044	0
11	105857	GROCERY I	1092	0
12	106716	GROCERY I	1032	0
13	108079	GROCERY I	1030	0
14	108634	GROCERY I	1075	0
15	108696	DELI	2636	1
16	108698	DELI	2644	1
17	108701	DELI	2644	1
18	108786	CLEANING	3044	0
19	108797	GROCERY I	1004	0
20	108831	POULTRY	2416	1
21	108833	EGGS	2502	1
22	108862	GROCERY I	1062	0
23	108952	CLEANING	3024	0
24	111223	GROCERY I	1034	0
25	111397	GROCERY I	1072	0
26	112830	GROCERY I	1044	0
27	114778	GROCERY I	1016	0
28	114790	GROCERY I	1004	0
29	114799	PERSONAL CARE	4126	0
...	...	...	...	...
4070	2127992	GROCERY I	1028	0
4071	2128628	BEVERAGES	1122	0
4072	2128799	BEVERAGES	1148	0
4073	2129334	GROCERY I	1086	0
4074	2129350	GROCERY I	1086	0
4075	2129387	GROCERY I	1068	0
4076	2129515	GROCERY I	1042	0
4077	2129616	BEVERAGES	1124	0
4078	2129678	GROCERY I	1030	0
4079	2129786	GROCERY I	1016	0
4080	2129790	GROCERY I	1094	0
4081	2129892	GROCERY I	1092	0
4082	2129994	GROCERY I	1092	0
4083	2130131	GROCERY I	1092	0
4084	2130219	GROCERY I	1094	0
4085	2130265	GROCERY I	1094	0
4086	2130352	GROCERY I	1094	0
4087	2130474	GROCERY I	1064	0
4088	2130521	GROCERY I	1040	0
4089	2130526	GROCERY I	1030	0
4090	2130553	LIQUOR,WINE,BEER	1318	0
4091	2131010	LIQUOR,WINE,BEER	1328	0

```
4092   2131572         GROCERY I   1002          0
4093   2131699         GROCERY I   1002          0
4094   2132163         GROCERY I   1040          0
4095   2132318         GROCERY I   1002          0
4096   2132945         GROCERY I   1026          0
4097   2132957         GROCERY I   1068          0
4098   2134058         BEVERAGES   1124          0
4099   2134244  LIQUOR,WINE,BEER   1364          0

[4100 rows x 4 columns]
```

```
In [11]: # Usando Librería missingno visualizo valores nulos
         # Vemos que sólo hay valores nulos en oil.date
         print("Nulls in Oil columns: {0} => {1}".
               format(oil.columns.values,oil.isnull().any().values))
         print("="*70)
         print("Nulls in holiday_events columns: {0} => {1}".
               format(holiday_events.columns.values,holiday_events.isnull().any().values))
         print("="*70)
         print("Nulls in stores columns: {0} => {1}".
               format(stores.columns.values,stores.isnull().any().values))
         print("="*70)
         print("Nulls in transactions columns: {0} => {1}".
               format(transactions.columns.values,transactions.isnull().any().values))
```

```
Nulls in Oil columns: ['date' 'dcoilwtico'] => [False  True]
======================================================================
Nulls in holiday_events columns: ['date' 'type' 'locale' 'locale_name' 'description' 'transferre
======================================================================
Nulls in stores columns: ['store_nbr' 'city' 'state' 'type' 'cluster'] => [False False False Fal
======================================================================
Nulls in transactions columns: ['date' 'store_nbr' 'transactions'] => [False False False]
```

4 2. Exploración de Datos Auxiliares

4.1 oil.csv

Este fichero contiene los precios diarios del petroleo desde 2013 hasta la acrtualidad.
Este es un dato importante porque la economía ecuatoriana es fuertemente dependiente del petróleo y cuando éste baja la economía se para.
El siguiente es un gráfico interactivo con los precios del petróleo, habiendo eliminado valores erróneos.
Muestra la caída pronunciada del segundo semestre del 2014 desde los U$S 100 hasta los 40 actuales.

```
In [12]: trace = go.Scatter(
                name='Oil prices',
```

```
        x=oil['date'],
        y=oil['dcoilwtico'].dropna(),
        mode='lines',
        line=dict(color='rgb(20, 15, 200, 0.8)'),
        fillcolor='rgba(68, 68, 68, 0.3)',
        #fillcolor='rgba(0, 0, 216, 0.3)',
        fill='tonexty' )

data = [trace]

layout = go.Layout(
        yaxis=dict(title='Daily Oil price'),
        title='Daily oil prices from Jan 2013 till July 2017',
        showlegend = False)
fig = go.Figure(data=data, layout=layout)
py.iplot(fig, filename='pandas-time-series-error-bars')
```

4.2 Datos de Tiendas

Este archivo incluye metadatos de las 54 tiendas referidos a ciudad, estado y "cluster" que es una agrupación de tiendas similares que se distribuyen en 17 grupos diferentes.
A continuación se incluyen sendos treemaps que dan idea de la distribución de tiendas por ciudad y por estado respectivamente

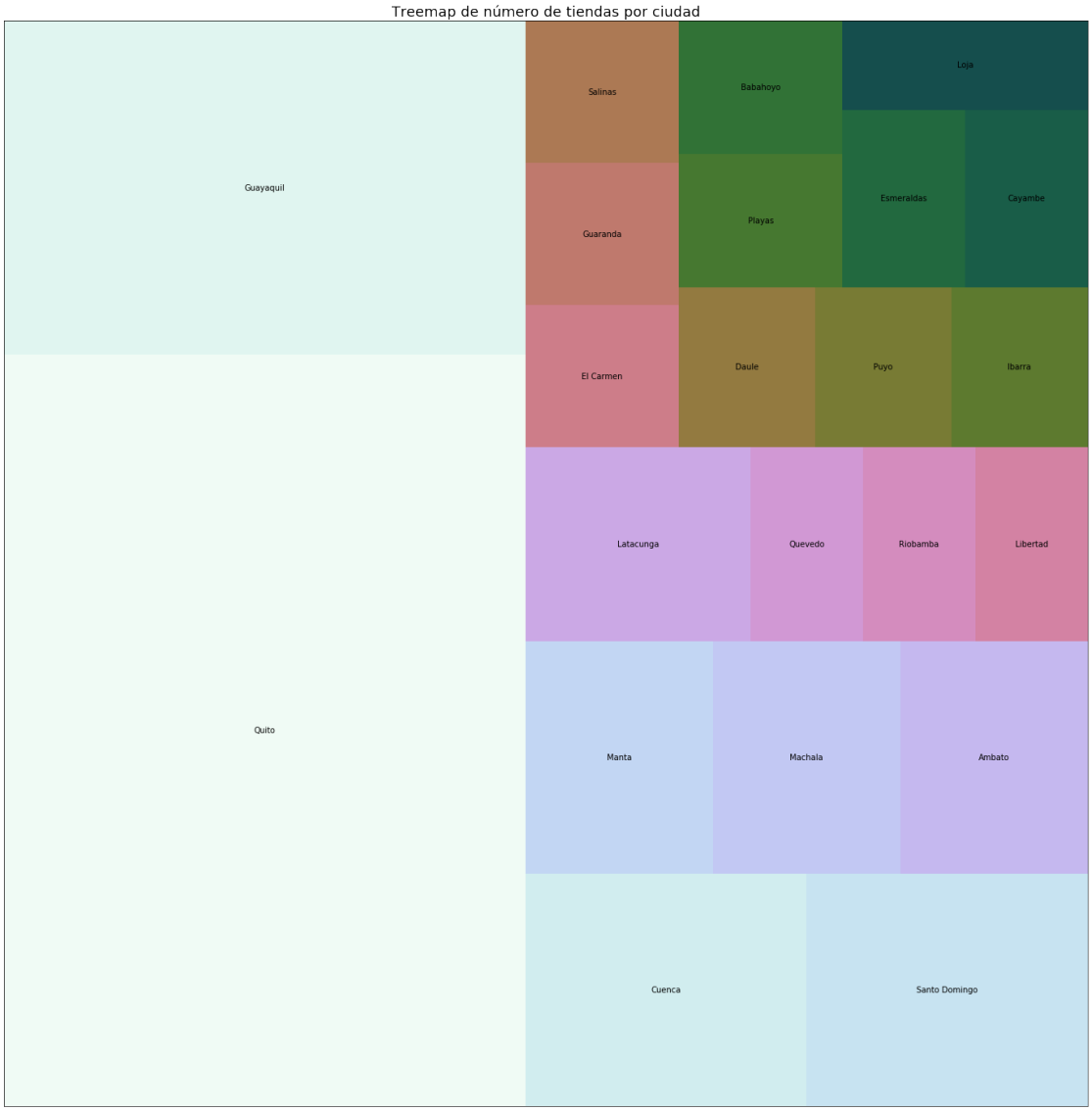
```
In [14]: stores
```

Out[14]:

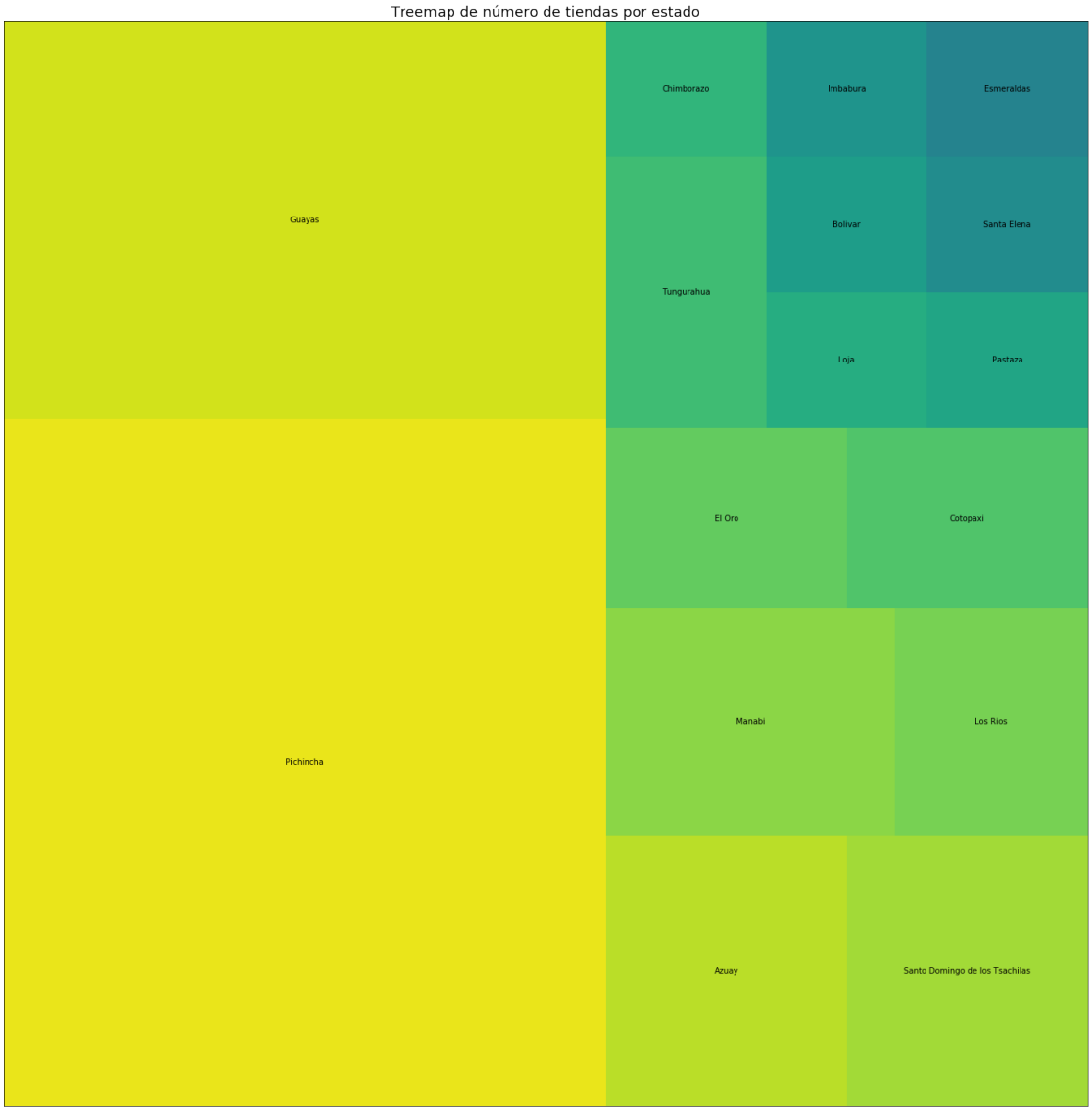
	store_nbr	city	state	type	cluster
0	1	Quito	Pichincha	D	13
1	2	Quito	Pichincha	D	13
2	3	Quito	Pichincha	D	8
3	4	Quito	Pichincha	D	9
4	5	Santo Domingo	Santo Domingo de los Tsachilas	D	4
5	6	Quito	Pichincha	D	13
6	7	Quito	Pichincha	D	8
7	8	Quito	Pichincha	D	8
8	9	Quito	Pichincha	B	6
9	10	Quito	Pichincha	C	15
10	11	Cayambe	Pichincha	B	6
11	12	Latacunga	Cotopaxi	C	15
12	13	Latacunga	Cotopaxi	C	15
13	14	Riobamba	Chimborazo	C	7
14	15	Ibarra	Imbabura	C	15
15	16	Santo Domingo	Santo Domingo de los Tsachilas	C	3
16	17	Quito	Pichincha	C	12
17	18	Quito	Pichincha	B	16
18	19	Guaranda	Bolivar	C	15
19	20	Quito	Pichincha	B	6
20	21	Santo Domingo	Santo Domingo de los Tsachilas	B	6

21	22	Puyo	Pastaza	C	7
22	23	Ambato	Tungurahua	D	9
23	24	Guayaquil	Guayas	D	1
24	25	Salinas	Santa Elena	D	1
25	26	Guayaquil	Guayas	D	10
26	27	Daule	Guayas	D	1
27	28	Guayaquil	Guayas	E	10
28	29	Guayaquil	Guayas	E	10
29	30	Guayaquil	Guayas	C	3
30	31	Babahoyo	Los Rios	B	10
31	32	Guayaquil	Guayas	C	3
32	33	Quevedo	Los Rios	C	3
33	34	Guayaquil	Guayas	B	6
34	35	Playas	Guayas	C	3
35	36	Libertad	Guayas	E	10
36	37	Cuenca	Azuay	D	2
37	38	Loja	Loja	D	4
38	39	Cuenca	Azuay	B	6
39	40	Machala	El Oro	C	3
40	41	Machala	El Oro	D	4
41	42	Cuenca	Azuay	D	2
42	43	Esmeraldas	Esmeraldas	E	10
43	44	Quito	Pichincha	A	5
44	45	Quito	Pichincha	A	11
45	46	Quito	Pichincha	A	14
46	47	Quito	Pichincha	A	14
47	48	Quito	Pichincha	A	14
48	49	Quito	Pichincha	A	11
49	50	Ambato	Tungurahua	A	14
50	51	Guayaquil	Guayas	A	17
51	52	Manta	Manabi	A	11
52	53	Manta	Manabi	D	13
53	54	El Carmen	Manabi	C	3

In [16]:
```
fig = plt.figure(figsize=(16, 11))
marrimeko=stores.city.value_counts().to_frame()
ax = fig.add_subplot(111, aspect="equal")
ax = squarify.plot(sizes=marrimeko['city'].values,label=marrimeko.index,
                   color=sns.color_palette('cubehelix_r', 28), alpha=1)
ax.set_xticks([])
ax.set_yticks([])
fig=plt.gcf()
fig.set_size_inches(40,25)
plt.title("Treemap de número de tiendas por ciudad", fontsize=18)
plt.show();
```



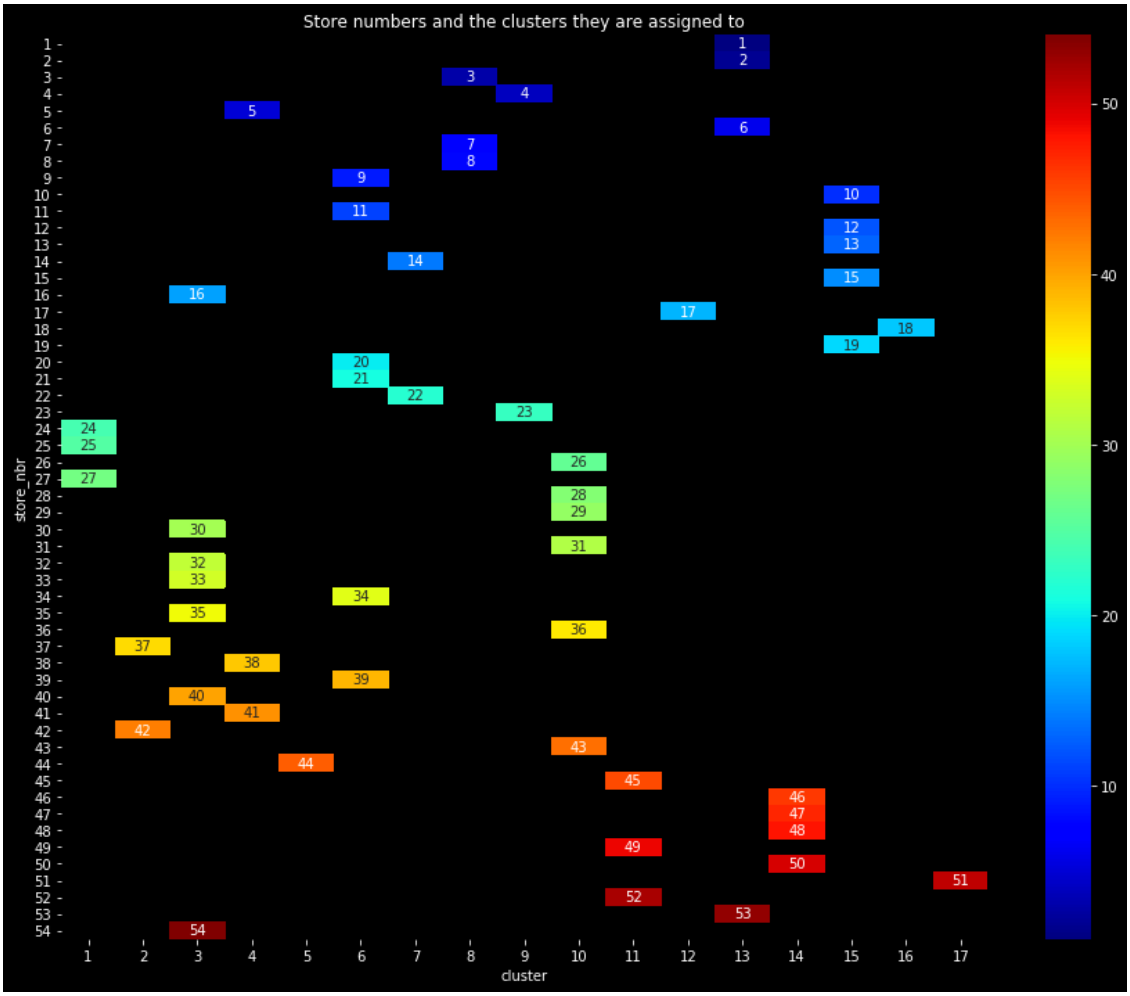

```
In [17]: fig = plt.figure(figsize=(14, 11))
         marrimeko=stores.state.value_counts().to_frame()
         ax = fig.add_subplot(111, aspect="equal")
         ax = squarify.plot(sizes=marrimeko['state'].values,label=marrimeko.index,
                            color=sns.color_palette('viridis_r', 28), alpha=1)
         ax.set_xticks([])
         ax.set_yticks([])
         fig=plt.gcf()
         fig.set_size_inches(40,25)
         plt.title("Treemap de número de tiendas por estado", fontsize=18)
         plt.show()
```



El siguiente gráfico explora si existe un patrón que asocia número de tienda con cluster. En una fase posterior asociaré cluster con otros atributos de la tienda para buscar mas posibles patrones.

```
In [18]: # Finally plot the seaborn heatmap
         plt.style.use('dark_background')
         plt.figure(figsize=(15,12))
         store_pivot = stores.dropna().pivot("store_nbr","cluster", "store_nbr")
         ax = sns.heatmap(store_pivot, cmap='jet', annot=True, linewidths=0, linecolor='white')
         plt.title('Store numbers and the clusters they are assigned to')

Out[18]: <matplotlib.text.Text at 0x110fc2518>
```

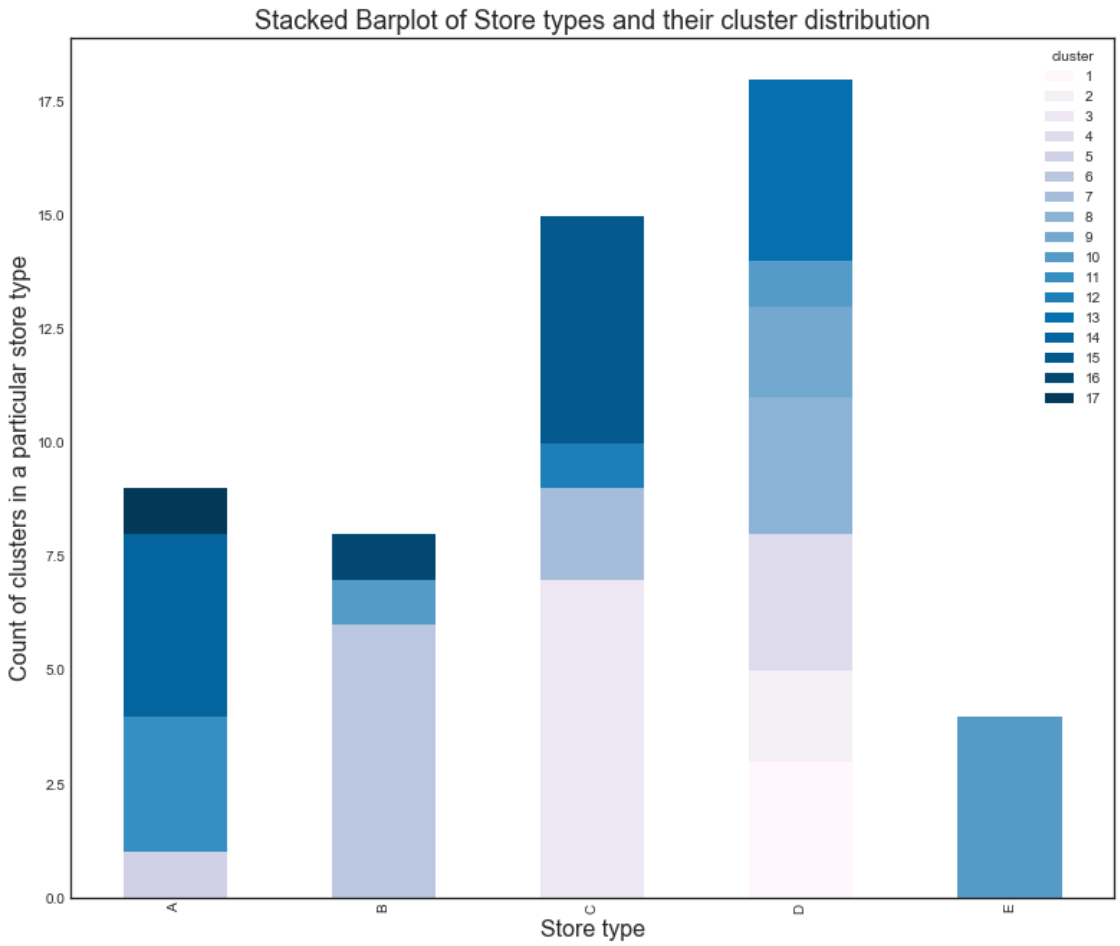



```
In [20]: neworder = [23, 24, 26, 36, 41, 15, 29, 31, 32, 34, 39,
                     53, 4, 37, 40, 43, 8, 10, 19, 20, 33, 38, 13,
                     21, 2, 6, 7, 3, 22, 25, 27, 28, 30, 35, 42, 44,
                     48, 51, 16, 0, 1, 5, 52, 45, 46, 47, 49, 9, 11, 12, 14, 18, 17, 50]
```

```
In [21]: plt.style.use('dark_background')
         nbr_cluster = stores.groupby(['store_nbr','cluster']).size()
         nbr_cluster.unstack().iloc[neworder].plot(kind='bar',stacked=True, colormap= 'tab20', f
         plt.title('Store numbers and the clusters they are assigned to', fontsize=14)
         plt.ylabel('')
         plt.xlabel('Store number')
         plt.show()
```

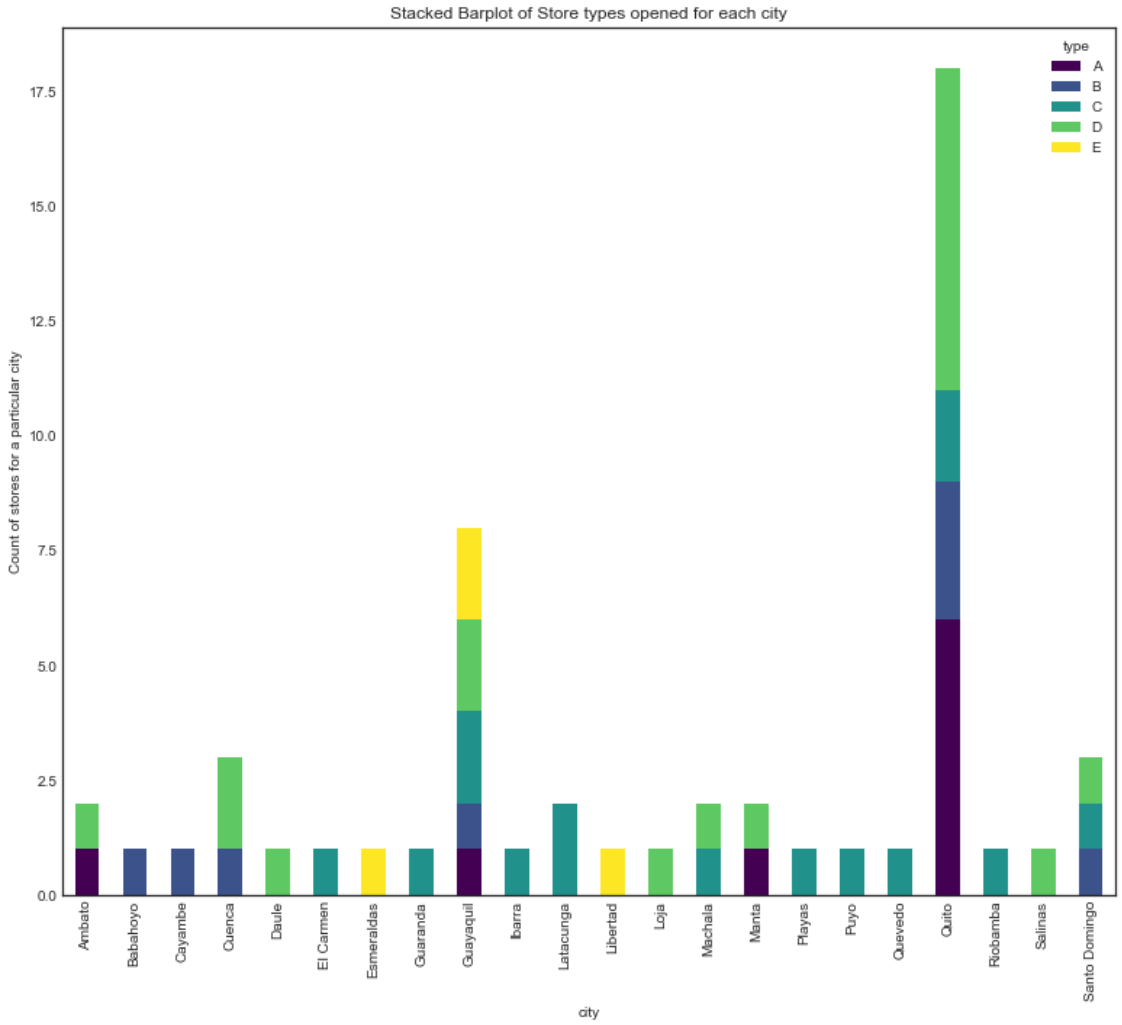
A continuación un gráfico de barras acumuladas correspondientes a tipos de tiendas para visualizar la distribución de los clusters dentro de cada tipo.
La conclusión primera y mas visible es que las tiendas tipo E sólo pertenecen al cluster 10.

```
In [22]: plt.style.use('seaborn-white')
         #plt.style.use('dark_background')
         type_cluster = stores.groupby(['type','cluster']).size()
         type_cluster.unstack().plot(kind='bar',stacked=True, colormap= 'PuBu', figsize=(13,11),
         plt.title('Stacked Barplot of Store types and their cluster distribution', fontsize=18)
         plt.ylabel('Count of clusters in a particular store type', fontsize=16)
         plt.xlabel('Store type', fontsize=16)
         plt.show()
```



A continuación se ve la distribución de tipos de tienda en cada una de las ciudades. Destacable el numero de tiendas en la capital Quito y en menor medida, Guayaquil, la ciudad mas poblada, con un número de tiendas muy superior al resto de las ciudades del país.

```
In [23]: # plt.style.use('dark_background')
         plt.style.use('seaborn-white')
         city_cluster = stores.groupby(['city','type']).store_nbr.size()
         city_cluster.unstack().plot(kind='bar',stacked=True, colormap= 'viridis', figsize=(13,1
         plt.title('Stacked Barplot of Store types opened for each city')
         plt.ylabel('Count of stores for a particular city')
         plt.show()
```



4.3 Datos de Festivos

Existen fiestas locales, regionales y nacionales. Además un día laboral se puede convertir en fiesta por decreto (puente) y aparece como tipo Additional. La columna transferred indica que esta fiesta ha sido desplazada, precisamente para mantener el calendario laboral sin añadir vacaciones adicionales.

```
In [24]: holiday_events
```

```
Out[24]:          date        type    locale                  locale_name  \
         0  2012-03-02     Holiday     Local                        Manta
         1  2012-04-01     Holiday  Regional                     Cotopaxi
         2  2012-04-12     Holiday     Local                       Cuenca
         3  2012-04-14     Holiday     Local                     Libertad
         4  2012-04-21     Holiday     Local                     Riobamba
         5  2012-05-12     Holiday     Local                         Puyo
```

6	2012-06-23	Holiday	Local	Guaranda
7	2012-06-25	Holiday	Regional	Imbabura
8	2012-06-25	Holiday	Local	Latacunga
9	2012-06-25	Holiday	Local	Machala
10	2012-07-03	Holiday	Local	Santo Domingo
11	2012-07-03	Holiday	Local	El Carmen
12	2012-07-23	Holiday	Local	Cayambe
13	2012-08-05	Holiday	Local	Esmeraldas
14	2012-08-10	Holiday	National	Ecuador
15	2012-08-15	Holiday	Local	Riobamba
16	2012-08-24	Holiday	Local	Ambato
17	2012-09-28	Holiday	Local	Ibarra
18	2012-10-07	Holiday	Local	Quevedo
19	2012-10-09	Holiday	National	Ecuador
20	2012-10-12	Transfer	National	Ecuador
21	2012-11-02	Holiday	National	Ecuador
22	2012-11-03	Holiday	National	Ecuador
23	2012-11-06	Holiday	Regional	Santo Domingo de los Tsachilas
24	2012-11-07	Holiday	Regional	Santa Elena
25	2012-11-10	Holiday	Local	Guaranda
26	2012-11-11	Holiday	Local	Latacunga
27	2012-11-12	Holiday	Local	Ambato
28	2012-12-05	Additional	Local	Quito
29	2012-12-06	Holiday	Local	Quito
..	...	...	...	...
320	2017-07-23	Holiday	Local	Cayambe
321	2017-07-24	Additional	Local	Guayaquil
322	2017-07-25	Additional	Local	Guayaquil
323	2017-08-05	Holiday	Local	Esmeraldas
324	2017-08-10	Holiday	National	Ecuador
325	2017-08-11	Transfer	National	Ecuador
326	2017-08-15	Holiday	Local	Riobamba
327	2017-08-24	Holiday	Local	Ambato
328	2017-09-28	Holiday	Local	Ibarra
329	2017-09-29	Transfer	Local	Ibarra
330	2017-10-07	Holiday	Local	Quevedo
331	2017-10-09	Holiday	National	Ecuador
332	2017-11-02	Holiday	National	Ecuador
333	2017-11-03	Holiday	National	Ecuador
334	2017-11-06	Holiday	Regional	Santo Domingo de los Tsachilas
335	2017-11-07	Holiday	Regional	Santa Elena
336	2017-11-10	Holiday	Local	Guaranda
337	2017-11-11	Holiday	Local	Latacunga
338	2017-11-12	Holiday	Local	Ambato
339	2017-12-05	Additional	Local	Quito
340	2017-12-06	Holiday	Local	Quito
341	2017-12-08	Holiday	Local	Loja
342	2017-12-08	Transfer	Local	Quito

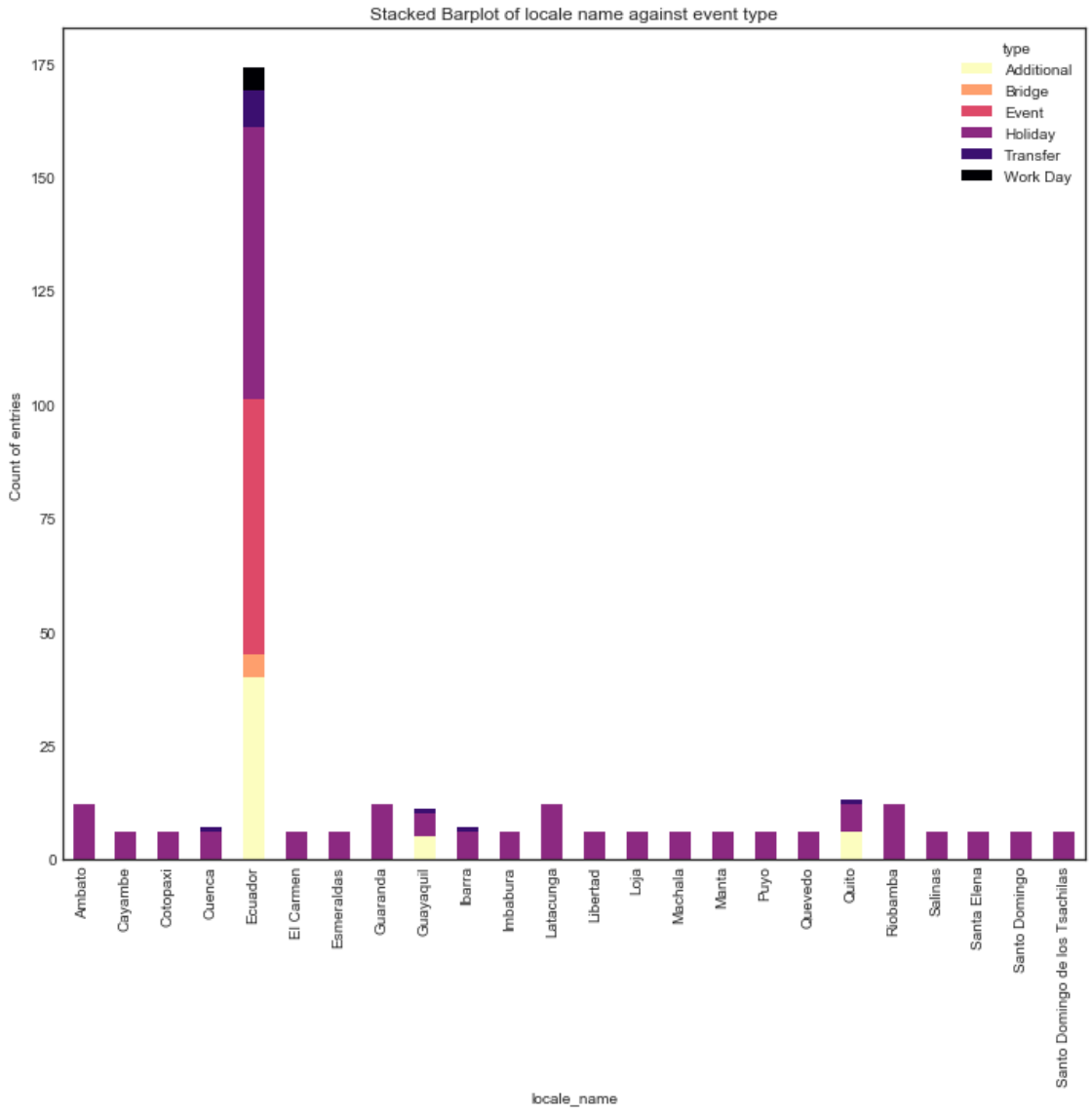
```
343  2017-12-21  Additional  National                      Ecuador
344  2017-12-22     Holiday     Local                      Salinas
345  2017-12-22  Additional  National                      Ecuador
346  2017-12-23  Additional  National                      Ecuador
347  2017-12-24  Additional  National                      Ecuador
348  2017-12-25     Holiday  National                      Ecuador
349  2017-12-26  Additional  National                      Ecuador

                              description  transferred
0                        Fundacion de Manta        False
1             Provincializacion de Cotopaxi        False
2                      Fundacion de Cuenca        False
3                 Cantonizacion de Libertad        False
4                 Cantonizacion de Riobamba        False
5                    Cantonizacion del Puyo        False
6                 Cantonizacion de Guaranda        False
7             Provincializacion de Imbabura        False
8                 Cantonizacion de Latacunga        False
9                       Fundacion de Machala        False
10                Fundacion de Santo Domingo        False
11                Cantonizacion de El Carmen        False
12                  Cantonizacion de Cayambe        False
13                   Fundacion de Esmeraldas        False
14            Primer Grito de Independencia        False
15                    Fundacion de Riobamba        False
16                      Fundacion de Ambato        False
17                      Fundacion de Ibarra        False
18                  Cantonizacion de Quevedo        False
19                Independencia de Guayaquil         True
20     Traslado Independencia de Guayaquil        False
21                           Dia de Difuntos        False
22                  Independencia de Cuenca        False
23       Provincializacion de Santo Domingo        False
24            Provincializacion Santa Elena        False
25                Independencia de Guaranda        False
26                Independencia de Latacunga        False
27                   Independencia de Ambato        False
28                      Fundacion de Quito-1        False
29                        Fundacion de Quito        False
..                                       ...          ...
320                 Cantonizacion de Cayambe        False
321                 Fundacion de Guayaquil-1        False
322                   Fundacion de Guayaquil        False
323                  Fundacion de Esmeraldas        False
324           Primer Grito de Independencia         True
325  Traslado Primer Grito de Independencia        False
326                   Fundacion de Riobamba        False
327                     Fundacion de Ambato        False
```

```
328                    Fundacion de Ibarra         True
329                    Fundacion de Ibarra        False
330                Cantonizacion de Quevedo        False
331              Independencia de Guayaquil        False
332                         Dia de Difuntos        False
333                 Independencia de Cuenca        False
334      Provincializacion de Santo Domingo        False
335           Provincializacion Santa Elena        False
336               Independencia de Guaranda        False
337              Independencia de Latacunga        False
338                 Independencia de Ambato        False
339                    Fundacion de Quito-1        False
340                      Fundacion de Quito         True
341                       Fundacion de Loja        False
342             Traslado Fundacion de Quito        False
343                                Navidad-4        False
344                Cantonizacion de Salinas        False
345                                Navidad-3        False
346                                Navidad-2        False
347                                Navidad-1        False
348                                  Navidad        False
349                                Navidad+1        False

[350 rows x 6 columns]
```

In [25]:
```
plt.style.use('seaborn-white')
# plt.style.use('dark_background')
holiday_local_type = holiday_events.groupby(['locale_name', 'type']).size()
holiday_local_type.unstack().plot(kind='bar',stacked=True, colormap= 'magma_r', figsize
plt.title('Stacked Barplot of locale name against event type')
plt.ylabel('Count of entries')
plt.show()
```



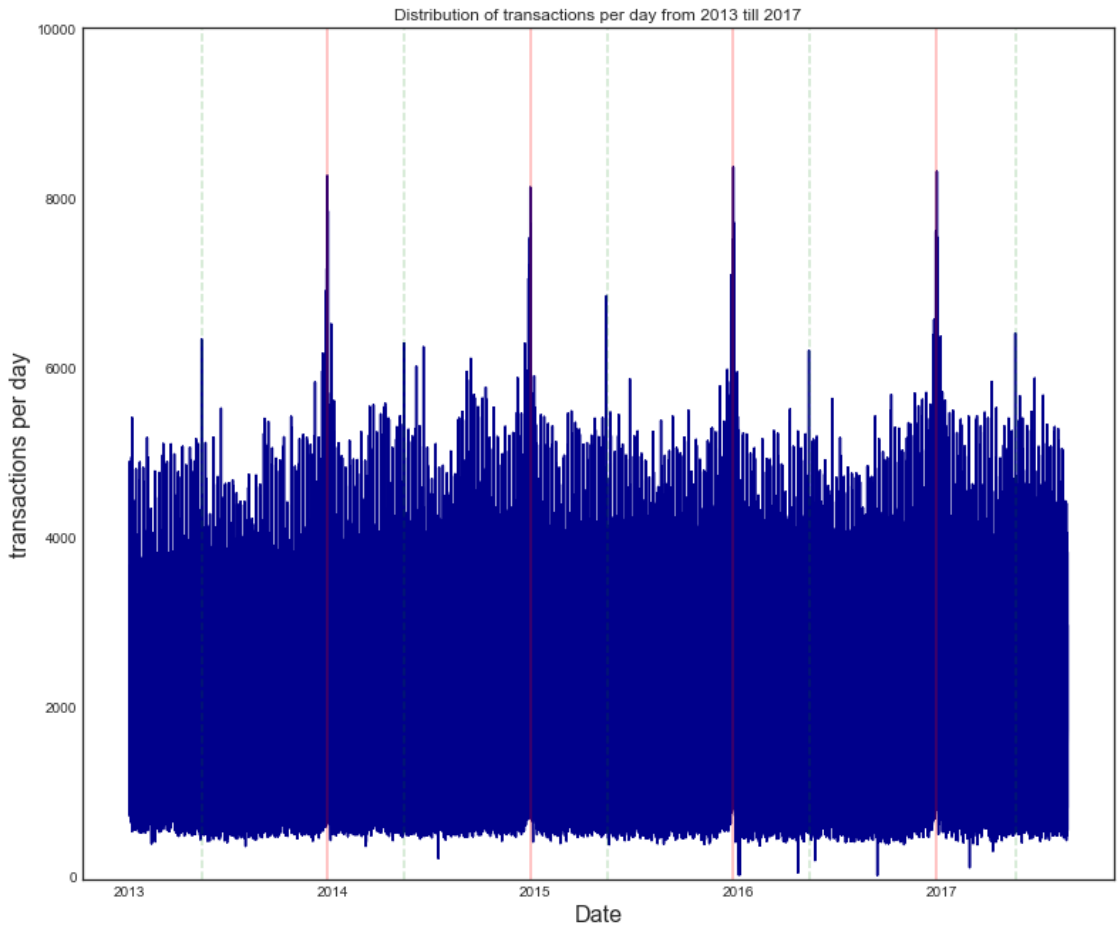
Porporción de fiestas nacionales (barra Ecuador), respecto a los demas tipos.
Datos de transacciones
Un dato importante a considerar es que la ventana de tiempo de las transacciones incluidas en este apartado corresponde al tiempo de los datos de training.

```
In [27]: transactions.head()
```

Out[27]:

	date	store_nbr	transactions
0	2013-01-01	25	770
1	2013-01-02	1	2111
2	2013-01-02	2	2358
3	2013-01-02	3	3487
4	2013-01-02	4	1922

```
In [28]: plt.style.use('seaborn-white')
         plt.figure(figsize=(13,11))
         plt.plot(transactions.date.values, transactions.transactions.values, color='darkblue')
         plt.axvline(x='2015-12-23',color='red',alpha=0.3)
         plt.axvline(x='2016-12-23',color='red',alpha=0.3)
         plt.axvline(x='2014-12-23',color='red',alpha=0.3)
         plt.axvline(x='2013-12-23',color='red',alpha=0.3)
         plt.axvline(x='2013-05-12',color='green',alpha=0.2, linestyle= '--')
         plt.axvline(x='2015-05-10',color='green',alpha=0.2, linestyle= '--')
         plt.axvline(x='2016-05-08',color='green',alpha=0.2, linestyle= '--')
         plt.axvline(x='2014-05-11',color='green',alpha=0.2, linestyle= '--')
         plt.axvline(x='2017-05-14',color='green',alpha=0.2, linestyle= '--')
         plt.ylim(-50, 10000)
         plt.title("Distribution of transactions per day from 2013 till 2017")
         plt.ylabel('transactions per day', fontsize= 16)
         plt.xlabel('Date', fontsize= 16)
         plt.show()
```



En este gráfico de transacciones por día se aprecia el pico de ventas que ocurre coincidiendo con el final de año, (linea vertical roja) y en menor medida el aumento de ventas a mediados de

mayo (linea vertical punteada verde).
Este patrón probablemente coincide con campañas de ofertas y descuentos promovidos por la Corporación Favorita.

4.4 Datos de productos

La información de los productos se limita a asignar familia a cada producto y como se verá abajo, la inmensa mayoría on comestibles. Indica también si son precederos

```
In [32]: items.head()
```

Out[32]:

	item_nbr	family	class	perishable
0	96995	GROCERY I	1093	0
1	99197	GROCERY I	1067	0
2	103501	CLEANING	3008	0
3	103520	GROCERY I	1028	0
4	103665	BREAD/BAKERY	2712	1

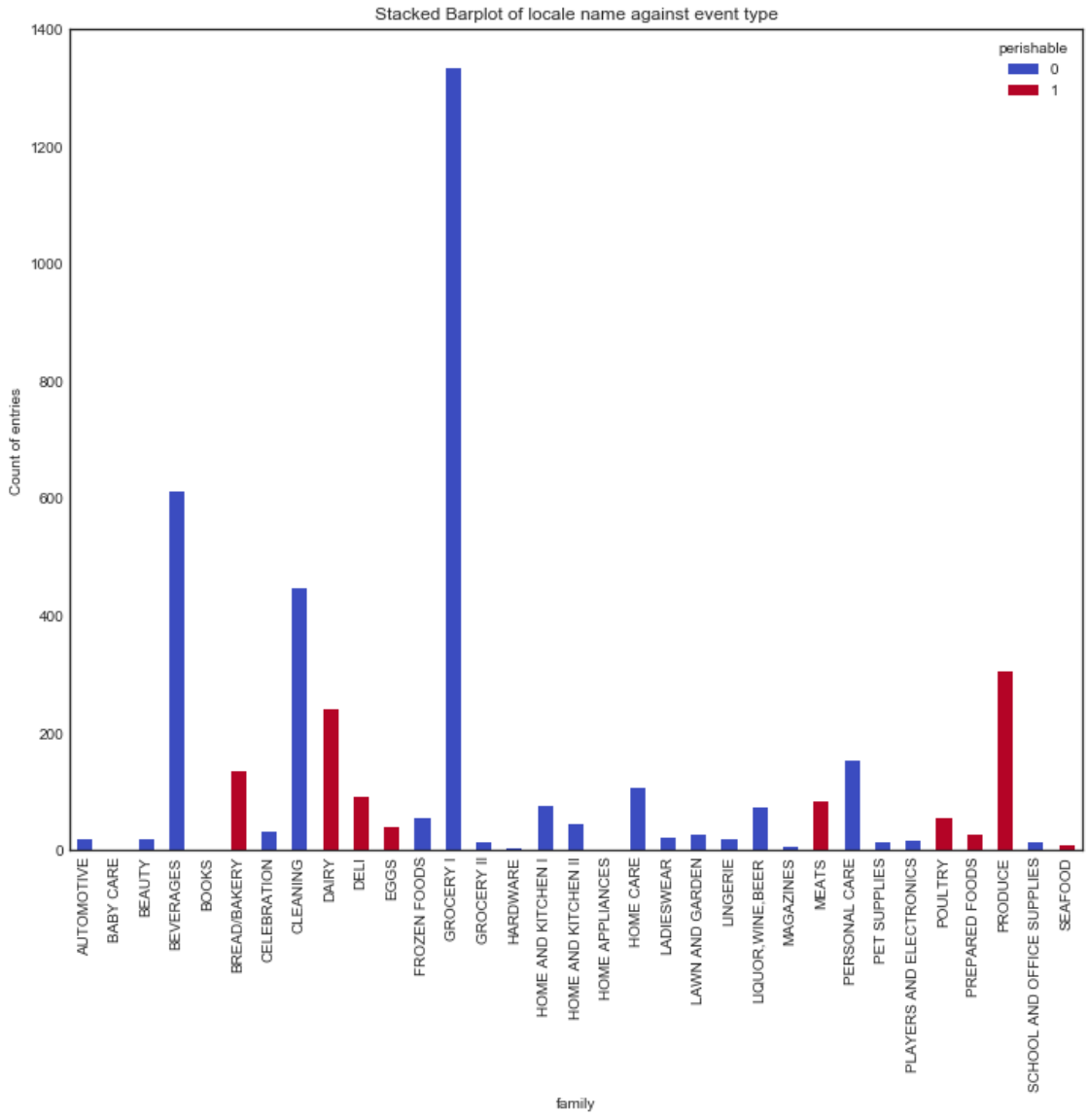
```
In [29]: x, y = (list(x) for x in zip(*sorted(zip(items.family.value_counts().index,
                                                  items.family.value_counts().values),
                                                 reverse = False)))
         trace2 = go.Bar(
             y=items.family.value_counts().values,
             x=items.family.value_counts().index,
             marker=dict(
                 color=items.family.value_counts().values,
                 colorscale = 'Portland',
                 reversescale = False
             ),
             orientation='v',
         )

         layout = dict(
             title='Counts of items per family category',
              width = 800, height = 800,
             yaxis=dict(
                 showgrid=False,
                 showline=False,
                 showticklabels=True,
         #         domain=[0, 0.85],
             ))

         fig1 = go.Figure(data=[trace2])
         fig1['layout'].update(layout)
         py.iplot(fig1, filename='plots')
```

Este gráfico interactivo indica la proporción logarítmica de numero de productos, siendo los mas numerosos los comestibles, bebidas, limpieza y lacteos.

```
In [33]: plt.style.use('seaborn-white')
         fam_perishable = items.groupby(['family', 'perishable']).size()
         fam_perishable.unstack().plot(kind='bar',stacked=True, colormap= 'coolwarm', figsize=(1
         plt.title('Stacked Barplot of locale name against event type')
         plt.ylabel('Count of entries')
         plt.show()
```



El gráfico anterior revela las familias de producto que son perecederas.

4.4.1 Referencias:

- Dataframe with all Date-Store-Item Combinations

- Comprehensive Python and D3.js Favorita analytics. https://www.kaggle.com/arthurtok/comprehensive-python-and-d3-js-favorita-analytics

- Memory Optimization and EDA on entire dataset. https://www.kaggle.com/jagangupta/memory-optimization-and-eda-on-entire-dataset