Store Sales - Time Series Forecasting

Use machine learning to predict grocery sales

Kaggle link: Store Sales - Time Series Forecasting

Context

Forecasts aren't just for meteorologists. Governments forecast economic growth. Scientists attempt to predict the future population. And businesses forecast product demand—a common task of professional data scientists. Forecasts are especially relevant to brick-and-mortar grocery stores, which must dance delicately with how much inventory to buy. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leading to lost revenue and upset customers. More accurate forecasting, thanks to machine learning, could help ensure retailers please customers by having just enough of the right products at the right time.

Current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing.

Potential Impact

If successful, you'll have flexed some new skills in a real world example. For grocery stores, more accurate forecasting can decrease food waste related to overstocking and improve customer satisfaction. The results of this ongoing competition, over time, might even ensure your local store has exactly what you need the next time you shop.

Dataset Description

In this competition, you will predict sales for the thousands of product families sold at Favorita stores located in Ecuador. The training data includes dates, store and product information, whether that item was being promoted, as well as the sales numbers. Additional files include supplementary information that may be useful in building your models.

File Descriptions and Data Field Information

train.csv The training data, comprising time series of features store_nbr, family, and onpromotion as well as the target sales. **store_nbr** identifies the store at which the products are sold. **family** identifies the type of product sold. **sales** gives the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed

to 1 bag of chips). **onpromotion** gives the total number of items in a product family that were being promoted at a store at a given date.

test.csv The test data, having the same features as the training data. You will predict the target sales for the dates in this file. The dates in the test data are for the 15 days after the last date in the training data.

sample_submission.csv A sample submission file in the correct format.

stores.csv Store metadata, including city, state, type, and cluster. cluster is a grouping of similar stores.

oil.csv Daily oil price. Includes values during both the train and test data timeframes. (Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices.)

holidays_events.csv

- Holidays and Events, with metadata
- NOTE: Pay special attention to the transferred column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.
- Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Additional Notes

- Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

Importing all libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import *
from sklearn.linear_model import *
from math import *
from sklearn.ensemble import *
from sklearn.feature_selection import *
```

```
from sklearn.feature_extraction import *
from sklearn.naive_bayes import *
from sklearn.discriminant_analysis import *
from sklearn.preprocessing import *
from sklearn.metrics import *
from sklearn.neighbors import *
from sklearn.cluster import *
from sklearn.svm import *
import warnings
warnings.filterwarnings("ignore")
```

Importing the training datasets

In [405…
```
df_train_1 = pd.read_csv("train1.csv")
df_train_1.head()
```

Out[405]:

	id	date	store_nbr	family	sales	onpromotion
0	0	1/1/13	1	AUTOMOTIVE	0.0	0
1	1	1/1/13	1	BABY CARE	0.0	0
2	2	1/1/13	1	BEAUTY	0.0	0
3	3	1/1/13	1	BEVERAGES	0.0	0
4	4	1/1/13	1	BOOKS	0.0	0

In [406…
```
df_train_2 = pd.read_csv("train2.csv")
df_train_2.head()
```

Out[406]:

	id	date	store_nbr	family	sales	onpromotion
0	262143	5/28/13	14	MEATS	310.173000	0
1	262144	5/28/13	14	PERSONAL CARE	227.000000	0
2	262145	5/28/13	14	PET SUPPLIES	0.000000	0
3	262146	5/28/13	14	PLAYERS AND ELECTRONICS	0.000000	0
4	262147	5/28/13	14	POULTRY	47.809002	0

In [407…
```
df_train_3 = pd.read_csv("train3.csv")
df_train_3.head()
```

Out[407]:

	id	date	store_nbr	family	sales	onpromotion
0	524286	10/22/13	2	HOME AND KITCHEN I	0.0	0
1	524287	10/22/13	2	HOME AND KITCHEN II	0.0	0
2	524288	10/22/13	2	HOME APPLIANCES	0.0	0
3	524289	10/22/13	2	HOME CARE	0.0	0
4	524290	10/22/13	2	LADIESWEAR	0.0	0

In [408…
```
df_train_4 = pd.read_csv("train4.csv")
df_train_4.head()
```

	id	date	store_nbr	family	sales	onpromotion
0	786429	3/19/14	25	CELEBRATION	8.000	0
1	786430	3/19/14	25	CLEANING	485.000	0
2	786431	3/19/14	25	DAIRY	588.000	0
3	786432	3/19/14	25	DELI	116.357	0
4	786433	3/19/14	25	EGGS	119.000	0

In [409… `print("Total length of training dataframes : ",(len(df_train_1)+len(df_train`

```
Total length of training dataframes :  1048575
```

Merging all training dataframes into one single dataframe

In [410…
```
df_train = pd.concat([df_train_1,df_train_2,df_train_3, df_train_4])
df_train.head()
```

Out[410]:

	id	date	store_nbr	family	sales	onpromotion
0	0	1/1/13	1	AUTOMOTIVE	0.0	0
1	1	1/1/13	1	BABY CARE	0.0	0
2	2	1/1/13	1	BEAUTY	0.0	0
3	3	1/1/13	1	BEVERAGES	0.0	0
4	4	1/1/13	1	BOOKS	0.0	0

In [411… `print("Total length of merged dataframe : ",(len(df_train)))`

```
Total length of merged dataframe :  1048575
```

Importing testing dataset

In [412…
```
df_test = pd.read_csv("test.csv")
df_test.head()
```

Out[412]:

	id	date	store_nbr	family	onpromotion
0	3000888	2017-08-16	1	AUTOMOTIVE	0
1	3000889	2017-08-16	1	BABY CARE	0
2	3000890	2017-08-16	1	BEAUTY	2
3	3000891	2017-08-16	1	BEVERAGES	20
4	3000892	2017-08-16	1	BOOKS	0

In [413… `print("Total length of testing dataframe : ",(len(df_test)))`

```
Total length of testing dataframe :  28512
```

Checking for null values in both training and testing datasets

In [414… `df_train.isna().sum()`

```
Out[414]:   id             0
            date           0
            store_nbr      0
            family         0
            sales          0
            onpromotion    0
            dtype: int64
```

```
In [415…   df_test.isna().sum()
```

```
Out[415]:   id             0
            date           0
            store_nbr      0
            family         0
            onpromotion    0
            dtype: int64
```

Exploring other datasets

```
In [416…   df_stores = pd.read_csv("stores.csv")
           df_stores.head()
```

Out[416]:

	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

```
In [417…   df_oil = pd.read_csv("oil.csv")
           df_oil.head()
```

Out[417]:

	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

```
In [418…   df_holiday_events = pd.read_csv("holidays_events.csv")
           df_holiday_events.head()
```

Out[418]:

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

```
In [419...    df_transactions = pd.read_csv("transactions.csv")
              df_transactions.head()
```

Out[419]:

	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 [420...    print("Total length of stores dataframe          :  ",len(df_stores))
              print("Total length of oil dataframe             :  ",len(df_oil))
              print("Total length of holiday events dataframe :  ",len(df_holiday_events))
              print("Total length of transactions dataframe    :  ",len(df_transactions))
```

```
Total length of stores dataframe          :    54
Total length of oil dataframe             :    1218
Total length of holiday events dataframe :    350
Total length of transactions dataframe    :    83488
```

Checking for null values in stores, oil, holiday events, transactions dataframes

```
In [421...    df_stores.isna().sum()
```

Out[421]:
```
store_nbr    0
city         0
state        0
type         0
cluster      0
dtype: int64
```

```
In [422...    df_oil.isna().sum()
```

Out[422]:
```
date          0
dcoilwtico    43
dtype: int64
```

```
In [423...    df_holiday_events.isna().sum()
```

Out[423]:
```
date           0
type           0
locale         0
locale_name    0
description    0
transferred    0
dtype: int64
```

```
In [424...    df_transactions.isna().sum()
```

Out[424]:
```
date            0
store_nbr       0
transactions    0
dtype: int64
```

Filling in null values in oil dataset

```
In [425…  df_oil_nona = df_oil.copy()
          modal = df_oil_nona["dcoilwtico"].mode()
          na = df_oil_nona[df_oil_nona["dcoilwtico"].isna()==True].index.values
          df_oil_nona.loc[0,"dcoilwtico"] = df_oil_nona.loc[0:10,"dcoilwtico"].mean()
          for i in na[1:]:
              df_oil_nona.loc[i,"dcoilwtico"] = df_oil_nona.loc[(i-10):(i+10),"dcoilwt
          df_oil_nona.head()
```

Out[425]:

	date	dcoilwtico
0	2013-01-01	93.366
1	2013-01-02	93.140
2	2013-01-03	92.970
3	2013-01-04	93.120
4	2013-01-07	93.200

```
In [426…  df_oil_nona.isna().sum()
```

Out[426]:   date   0
             dcoilwtico   0
             dtype: int64

```
In [427…  df_holiday_events.head()
```

Out[427]:

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

Preprocessing the holiday events dataset

```
In [428…  before_transfer = df_holiday_events[df_holiday_events["transferred"]==True]
          before_transfer.head()
```

Out[428]:

	date	type	locale	locale_name	description	transferred
19	2012-10-09	Holiday	National	Ecuador	Independencia de Guayaquil	True
72	2013-10-09	Holiday	National	Ecuador	Independencia de Guayaquil	True
135	2014-10-09	Holiday	National	Ecuador	Independencia de Guayaquil	True
255	2016-05-24	Holiday	National	Ecuador	Batalla de Pichincha	True
266	2016-07-25	Holiday	Local	Guayaquil	Fundacion de Guayaquil	True

```
In [429…  after_transfer = df_holiday_events[df_holiday_events["type"]=="Transfer"]
          after_transfer.head()
```

	date	type	locale	locale_name	description	transferred
20	2012-10-12	Transfer	National	Ecuador	Traslado Independencia de Guayaquil	False
73	2013-10-11	Transfer	National	Ecuador	Traslado Independencia de Guayaquil	False
136	2014-10-10	Transfer	National	Ecuador	Traslado Independencia de Guayaquil	False
256	2016-05-27	Transfer	National	Ecuador	Traslado Batalla de Pichincha	False
265	2016-07-24	Transfer	Local	Guayaquil	Traslado Fundacion de Guayaquil	False

In [430...
```python
df_holiday_trans = df_holiday_events.drop(before_transfer.index.values,axis=
ts = after_transfer.index.values
for i in ts:
    st = df_holiday_trans.loc[i,"description"]
    ls = st.split(" ")
    if "Traslado" in ls:
        ls.remove("Traslado")
    df_holiday_trans.loc[i,"description"] = " ".join(ls)
```

In [431...
```python
df_holiday_trans[df_holiday_trans["type"]=="Bridge"]
```

Out[431]:

	date	type	locale	locale_name	description	transferred
35	2012-12-24	Bridge	National	Ecuador	Puente Navidad	False
39	2012-12-31	Bridge	National	Ecuador	Puente Primer dia del ano	False
156	2014-12-26	Bridge	National	Ecuador	Puente Navidad	False
160	2015-01-02	Bridge	National	Ecuador	Puente Primer dia del ano	False
277	2016-11-04	Bridge	National	Ecuador	Puente Dia de Difuntos	False

In [432...
```python
df_holiday_trans["type"].unique()
```

Out[432]:
```
array(['Holiday', 'Transfer', 'Additional', 'Bridge', 'Work Day', 'Event'],
      dtype=object)
```

In [433...
```python
df_holiday_trans[df_holiday_trans["type"]=="Additional"].head()
```

Out[433]:

	date	type	locale	locale_name	description	transferred
28	2012-12-05	Additional	Local	Quito	Fundacion de Quito-1	False
31	2012-12-21	Additional	National	Ecuador	Navidad-4	False
33	2012-12-22	Additional	National	Ecuador	Navidad-3	False
34	2012-12-23	Additional	National	Ecuador	Navidad-2	False
36	2012-12-24	Additional	National	Ecuador	Navidad-1	False

In [434...
```python
work = df_holiday_trans[df_holiday_trans["type"]=="Work Day"].index.values
work
```

Out[434]:
```
array([ 42,  43, 149, 161, 283])
```

```
In [435… df_holiday = df_holiday_trans.drop(work,axis=0,inplace=False)
         df_holiday.head()
```

Out[435]:

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

Displaying stores, oil, holiday events, transactions dataframes once again

```
In [436… df_stores.head()
```

Out[436]:

	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

```
In [437… df_oil_nona.head()
```

Out[437]:

	date	dcoilwtico
0	2013-01-01	93.366
1	2013-01-02	93.140
2	2013-01-03	92.970
3	2013-01-04	93.120
4	2013-01-07	93.200

```
In [438… df_holiday.head()
```

Out[438]:

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

```
In [439… df_transactions.head()
```

```
Out[439]:        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
```

Displaying training and testing datasets once again

```
In [440…  df_train.head()
```

```
Out[440]:    id   date  store_nbr       family  sales  onpromotion
         0   0  1/1/13          1   AUTOMOTIVE    0.0            0
         1   1  1/1/13          1    BABY CARE    0.0            0
         2   2  1/1/13          1       BEAUTY    0.0            0
         3   3  1/1/13          1    BEVERAGES    0.0            0
         4   4  1/1/13          1        BOOKS    0.0            0
```

```
In [441…  df_test.head()
```

```
Out[441]:        id        date  store_nbr       family  onpromotion
         0  3000888  2017-08-16          1   AUTOMOTIVE            0
         1  3000889  2017-08-16          1    BABY CARE            0
         2  3000890  2017-08-16          1       BEAUTY            2
         3  3000891  2017-08-16          1    BEVERAGES           20
         4  3000892  2017-08-16          1        BOOKS            0
```

Modifying "date" column in training dataset

```
In [442…  df_train.head()
```

```
Out[442]:    id   date  store_nbr       family  sales  onpromotion
         0   0  1/1/13          1   AUTOMOTIVE    0.0            0
         1   1  1/1/13          1    BABY CARE    0.0            0
         2   2  1/1/13          1       BEAUTY    0.0            0
         3   3  1/1/13          1    BEVERAGES    0.0            0
         4   4  1/1/13          1        BOOKS    0.0            0
```

```
In [443…  df_train_mod = df_train.copy()
          df_train_mod["date"] = pd.to_datetime(df_train_mod["date"])
          df_train_mod.head()
```

Out[443]:

	id	date	store_nbr	family	sales	onpromotion
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0
1	1	2013-01-01	1	BABY CARE	0.0	0
2	2	2013-01-01	1	BEAUTY	0.0	0
3	3	2013-01-01	1	BEVERAGES	0.0	0
4	4	2013-01-01	1	BOOKS	0.0	0

Combining training and testing datasets with stores, oil, holiday events, transactions

Combining with stores dataset

In [444...
```python
print("Length of training dataframe : ",(len(df_train_mod)))
print("Length of testing dataframe  : ",(len(df_test)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

In [445...
```python
df_stores.head()
```

Out[445]:

	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

In [446...
```python
df_train_mod["date"] = pd.to_datetime(df_train_mod["date"])

df_train_mod_1 = pd.merge(left=df_train_mod,right=df_stores,how="left",on="s
df_test_mod_1 = pd.merge(left=df_test,right=df_stores,how="left",on="store_n
```

In [447...
```python
df_train_mod_1.head()
```

Out[447]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

In [448...
```python
df_test_mod_1.head()
```

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

In [449…
```python
print("Length of training dataframe : ",(len(df_train_mod_1)))
print("Length of testing dataframe  : ",(len(df_test_mod_1)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

In [450…
```python
df_train_mod_1.isna().sum()
```

Out[450]:
```
id              0
date            0
store_nbr       0
family          0
sales           0
onpromotion     0
city            0
state           0
type            0
cluster         0
dtype: int64
```

In [451…
```python
df_test_mod_1.isna().sum()
```

Out[451]:
```
id              0
date            0
store_nbr       0
family          0
onpromotion     0
city            0
state           0
type            0
cluster         0
dtype: int64
```

Combining with oil dataset

In [452…
```python
df_oil_nona.head()
```

```
Out[452]:        date   dcoilwtico
          0   2013-01-01     93.366
          1   2013-01-02     93.140
          2   2013-01-03     92.970
          3   2013-01-04     93.120
          4   2013-01-07     93.200
```

```
In [453…  df_train_mod_1["date"] = pd.to_datetime(df_train_mod_1["date"])
          df_test_mod_1["date"] = pd.to_datetime(df_test_mod_1["date"])

          df_oil_nona["date"] = pd.to_datetime(df_oil_nona["date"])

          df_train_mod_2 = pd.merge(left=df_train_mod_1,right=df_oil_nona,how="left",o
          df_test_mod_2 = pd.merge(left=df_test_mod_1,right=df_oil_nona,how="left",on=
```

```
In [454…  df_train_mod_2.head()
```

Out[454]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

```
In [455…  df_test_mod_2.head()
```

Out[455]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

```
In [456…  print("Length of training dataframe : ",(len(df_train_mod_2)))
          print("Length of testing dataframe  : ",(len(df_test_mod_2)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

In [457… `df_train_mod_2.isna().sum()`

Out[457]:
```
id                    0
date                  0
store_nbr             0
family                0
sales                 0
onpromotion           0
city                  0
state                 0
type                  0
cluster               0
dcoilwtico       299376
dtype: int64
```

In [458… `df_test_mod_2.isna().sum()`

Out[458]:
```
id                 0
date               0
store_nbr          0
family             0
onpromotion        0
city               0
state              0
type               0
cluster            0
dcoilwtico      7128
dtype: int64
```

In [459…
```
hh = df_train_mod_2[~df_train_mod_2["dcoilwtico"].isna()]
df_train_mod_3 = df_train_mod_2.fillna(value=hh["dcoilwtico"].mean(),inplace
df_train_mod_3.head()
```

Out[459]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

In [460…
```
hh2 = df_test_mod_2[~df_test_mod_2["dcoilwtico"].isna()]
df_test_mod_3 = df_test_mod_2.fillna(value=hh2["dcoilwtico"].mean(),inplace=
df_test_mod_3.head()
```

Out[460]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

In [461…
```python
print("Length of training dataframe : ",(len(df_train_mod_3)))
print("Length of testing dataframe  : ",(len(df_test_mod_3)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

Combining with holiday events dataframe

In [462…
```python
df_holiday_mod = df_holiday.drop(["type","transferred"],axis=1,inplace=False
df_holiday_mod.head()
```

Out[462]:

	date	locale	locale_name	description
0	2012-03-02	Local	Manta	Fundacion de Manta
1	2012-04-01	Regional	Cotopaxi	Provincializacion de Cotopaxi
2	2012-04-12	Local	Cuenca	Fundacion de Cuenca
3	2012-04-14	Local	Libertad	Cantonizacion de Libertad
4	2012-04-21	Local	Riobamba	Cantonizacion de Riobamba

In [463…
```python
df_train_mod_3["date"] = pd.to_datetime(df_train_mod_3["date"])
df_test_mod_3["date"] = pd.to_datetime(df_test_mod_3["date"])

df_holiday_mod["date"] = pd.to_datetime(df_holiday_mod["date"])

df_train_mod_4 = pd.merge(left=df_train_mod_3,right=df_holiday_mod,how="left
df_test_mod_4 = pd.merge(left=df_test_mod_3,right=df_holiday_mod,how="left",
```

In [464…
```python
df_train_mod_4.head()
```

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

In [465…

```python
df_test_mod_4.head()
```

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

In [466…

```python
print("Length of training dataframe : ",(len(df_train_mod_4)))
print("Length of testing dataframe  : ",(len(df_test_mod_4)))
```

```
Length of training dataframe :  1064613
Length of testing dataframe  :  28512
```

In [467…

```python
train_na_values = df_train_mod_4.isna().sum()
train_na_values
```

```
id                 0
date               0
store_nbr          0
family             0
sales              0
onpromotion        0
city               0
state              0
type               0
cluster            0
dcoilwtico         0
locale        911361
locale_name   911361
description   911361
dtype: int64
```

```
In [468… test_na_values = df_test_mod_4.isna().sum()
         test_na_values
```

```
Out[468]:  id                0
           date              0
           store_nbr         0
           family            0
           onpromotion       0
           city              0
           state             0
           type              0
           cluster           0
           dcoilwtico        0
           locale        26730
           locale_name   26730
           description   26730
           dtype: int64
```

```
In [469… print("Number of null values in training dataset : ",(train_na_values["local
         print("Number of null values in testing dataset  : ",(test_na_values["locale
```

```
Number of null values in training dataset :  85.6
Number of null values in testing dataset  :  93.75
```

Since the resultant training and testing datasets have a very high percentage of null values, this cannot be accepted. This means that the parameter needed to be considered can be only if a day is a holiday or not.

```
In [470… df_holiday.head()
```

Out[470]:

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

```
In [471… holiday_date1 = df_holiday["date"].drop_duplicates(ignore_index=True,inplace
         holiday_date = pd.DataFrame(columns=["date","holiday?"])
         holiday_date["date"] = holiday_date1
         holiday_date["holiday?"] = 1
```

```
In [472… holiday_date.head()
```

Out[472]:

	date	holiday?
0	2012-03-02	1
1	2012-04-01	1
2	2012-04-12	1
3	2012-04-14	1
4	2012-04-21	1

```
In [473… print("Number of holidays : ",len(holiday_date))
```

```
Number of holidays :  296
```

In [474…
```python
holiday_date["date"] = pd.to_datetime(holiday_date["date"])

df_train_mod_5 = pd.merge(left=df_train_mod_3,right=holiday_date,how="left",
df_test_mod_5 = pd.merge(left=df_test_mod_3,right=holiday_date,how="left",on
```

In [475…
```python
df_train_mod_5.head()
```

Out[475]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

In [476…
```python
df_test_mod_5.head()
```

Out[476]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster	
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13	
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13	
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13	
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13	
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13	

In [477…
```python
print("Length of training dataframe : ",(len(df_train_mod_5)))
print("Length of testing dataframe  : ",(len(df_test_mod_5)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

In [478…
```python
print("Number of null values in 'holiday' column in training dataset : ",df_
print("Number of null values in 'holiday' column in testing dataset  : ",df_
```

```
Number of null values in 'holiday' column in training dataset :  911361
Number of null values in 'holiday' column in testing dataset  :  26730
```

In [479…
```python
df_train_mod_6 = df_train_mod_5.fillna(value=0,inplace=False)
df_test_mod_6 = df_test_mod_5.fillna(value=0,inplace=False)
```

In [480…
```python
df_train_mod_6.head()
```

Out[480]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

In [481… `df_test_mod_6.head()`

Out[481]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

In [482…
```
print("Length of training dataframe : ",(len(df_train_mod_6)))
print("Length of testing dataframe  : ",(len(df_test_mod_6)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

In [483…
```
print("Number of null values in 'holiday' column in training dataset : ",df_
print("Number of null values in 'holiday' column in testing dataset  : ",df_
```

```
Number of null values in 'holiday' column in training dataset :  0
Number of null values in 'holiday' column in testing dataset  :  0
```

Combining with transactions dataset

In [484… `df_transactions.head()`

Out[484]:

	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 [485…
```python
print("Length of transactions dataset : ",len(df_transactions))
```
Length of transactions dataset : 83488

In [486…
```python
df_transactions["date"] = pd.to_datetime(df_transactions["date"])

df_train_mod_7 = pd.merge(left=df_train_mod_6,right=df_transactions,how="lef
df_test_mod_7 = pd.merge(left=df_test_mod_6,right=df_transactions,how="left"
```

In [487…
```python
df_train_mod_7.head()
```

Out[487]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type	cluster
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13
1	1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13
2	2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13
3	3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13
4	4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13

In [488…
```python
df_test_mod_7.head()
```

Out[488]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

In [489…
```python
print("Length of training dataframe : ",(len(df_train_mod_7)))
print("Length of testing dataframe  : ",(len(df_test_mod_7)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

In [490… `df_train_mod_7.isna().sum()`

Out[490]:
```
id                      0
date                    0
store_nbr               0
family                  0
sales                   0
onpromotion             0
city                    0
state                   0
type                    0
cluster                 0
dcoilwtico              0
holiday?                0
transactions       145266
dtype: int64
```

In [491… `df_test_mod_7.isna().sum()`

Out[491]:
```
id                      0
date                    0
store_nbr               0
family                  0
onpromotion             0
city                    0
state                   0
type                    0
cluster                 0
dcoilwtico              0
holiday?                0
transactions        28512
dtype: int64
```

In [492…
```
train_null = df_train_mod_7["transactions"].isna().sum()
test_null = df_test_mod_7["transactions"].isna().sum()
```

In [493…
```
print("Number of null values in 'transactions' column in training dataset :
print("Number of null values in 'transactions' column in testing dataset  :
```
```
Number of null values in 'transactions' column in training dataset :  145266
Number of null values in 'transactions' column in testing dataset  :  28512
```

In [494…
```
print("Percentage of null values in 'transactions' column in training datase
print("Percentage of null values in 'transactions' column in testing dataset
```
```
Percentage of null values in 'transactions' column in training dataset :  1
3.85
Percentage of null values in 'transactions' column in testing dataset  :  10
0.0
```

In the testing dataset, there are no transactional values at all. Hence, let us calculate the perentage of all null values in both the datasets.

In [495…
```
total_per = (train_null+test_null)*100/(len(df_train_mod_7)+len(df_test_mod_
print("Total percentage : ",total_per.round(2))
```
```
Total percentage :  16.13
```

Although the null values for transactional data account for 16.13%, the testing dataset do not have any transactional values, hence we shall ignore the column.

Final datasets

```
train_col = df_train_mod_6.columns.values
loc = np.where(train_col=="sales")[0][0]
train_col = np.delete(train_col,loc)
train_col = np.append(train_col,"sales")
train_col
```

Out[496]:
```
array(['id', 'date', 'store_nbr', 'family', 'onpromotion', 'city',
       'state', 'type', 'cluster', 'dcoilwtico', 'holiday?', 'sales'],
      dtype=object)
```

```
final_train = df_train_mod_6[train_col]
final_test = df_test_mod_6.copy()
```

```
final_train.head()
```

Out[498]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster	dcoilwt
0	0	2013-01-01	1	AUTOMOTIVE	0	Quito	Pichincha	D	13	93.3
1	1	2013-01-01	1	BABY CARE	0	Quito	Pichincha	D	13	93.3
2	2	2013-01-01	1	BEAUTY	0	Quito	Pichincha	D	13	93.3
3	3	2013-01-01	1	BEVERAGES	0	Quito	Pichincha	D	13	93.3
4	4	2013-01-01	1	BOOKS	0	Quito	Pichincha	D	13	93.3

```
final_test.head()
```

Out[499]:

	id	date	store_nbr	family	onpromotion	city	state	type	cluster
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Quito	Pichincha	D	13
1	3000889	2017-08-16	1	BABY CARE	0	Quito	Pichincha	D	13
2	3000890	2017-08-16	1	BEAUTY	2	Quito	Pichincha	D	13
3	3000891	2017-08-16	1	BEVERAGES	20	Quito	Pichincha	D	13
4	3000892	2017-08-16	1	BOOKS	0	Quito	Pichincha	D	13

```
print("Length of training dataframe : ",(len(final_train)))
print("Length of testing dataframe  : ",(len(final_test)))
```

```
Length of training dataframe :  1048575
Length of testing dataframe  :  28512
```

Checking null values in final datasets

```
In [501…  final_train.isna().sum()
```

```
Out[501]:  id              0
           date            0
           store_nbr       0
           family          0
           onpromotion     0
           city            0
           state           0
           type            0
           cluster         0
           dcoilwtico      0
           holiday?        0
           sales           0
           dtype: int64
```

```
In [502…  final_test.isna().sum()
```

```
Out[502]:  id              0
           date            0
           store_nbr       0
           family          0
           onpromotion     0
           city            0
           state           0
           type            0
           cluster         0
           dcoilwtico      0
           holiday?        0
           dtype: int64
```

Exporting the datasets to respective CSV files

```
In [503…  final_train.to_csv("train_eda.csv",index=False)
          final_test.to_csv("test_eda.csv",index=False)
```

```
In [503…
```

```
In [ ]:
```