

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
```

```
In [2]: test = pd.read_csv('test.csv')
test
```

Out[2]:

	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
...
28507	3029395	2017-08-31	9	POULTRY	1
28508	3029396	2017-08-31	9	PREPARED FOODS	0
28509	3029397	2017-08-31	9	PRODUCE	1
28510	3029398	2017-08-31	9	SCHOOL AND OFFICE SUPPLIES	9
28511	3029399	2017-08-31	9	SEAFOOD	0

28512 rows × 5 columns

```
In [3]: stores = pd.read_csv("stores.csv")
stores.head()
```

Out[3]:

	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 [4]: df_test = test.merge(stores ,how = 'left',on = 'store_nbr')
df_test
```

	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
...
28507	3029395	2017-08-31	9	POULTRY	1	Quito	Pichincha	B	6
28508	3029396	2017-08-31	9	PREPARED FOODS	0	Quito	Pichincha	B	6
...

```
In [5]: oil = pd.read_csv("oil.csv")
oil.head()
```

Out[5]:

	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 [6]: df_train_1 = df_test.merge(oil, how = 'left', on = 'date')
df_train_1
```

	id	date	store_nbr	family	onpromotion	city	state	type	cluster	d
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	
...	
28507	3029395	2017-08-31	9	POULTRY	1	Quito	Pichincha	B	6	
28508	3029396	2017-08-31	9	PREPARED FOODS	0	Quito	Pichincha	B	6	

```
In [9]: events = pd.read_csv('holidays_events.csv')
events
```

Out[9]:

	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
...
345	2017-12-22	Additional	National	Ecuador	Navidad-3	False
346	2017-12-23	Additional	National	Ecuador	Navidad-2	False
347	2017-12-24	Additional	National	Ecuador	Navidad-1	False
348	2017-12-25	Holiday	National	Ecuador	Navidad	False
349	2017-12-26	Additional	National	Ecuador	Navidad+1	False

350 rows × 6 columns

```
In [10]: final_df = df_train_1.merge(events, how = 'left', on = 'date')
```

```
In [12]: final_df.shape
```

Out[12]: (28512, 15)

```
In [13]: final_df.isna().sum()
```

```
Out[13]: id                0
         date              0
         store_nbr         0
         family            0
         onpromotion       0
         city              0
         state             0
         type_x            0
         cluster           0
         dcoilwtico        7128
         type_y            26730
         locale            26730
         locale_name       26730
         description       26730
         transferred       26730
         dtype: int64
```

```
In [14]: final_df.drop(['type_y', 'locale', 'locale_name', 'description', 'transferred' ],axis=1)
```

```
In [237]: final_df_1.head(50)
```

```
Out[237]:
```

	date	store_nbr	family	onpromotion	state	type_x
0	736557	1	0	0	12	3
1	736557	1	1	0	12	3
2	736557	1	2	2	12	3
3	736557	1	3	20	12	3
4	736557	1	4	0	12	3
5	736557	1	5	12	12	3
6	736557	1	6	0	12	3
7	736557	1	7	25	12	3
8	736557	1	8	45	12	3
9	736557	1	9	18	12	3
10	736557	1	10	1	12	3
11	736557	1	11	1	12	3
12	736557	1	12	64	12	3
13	736557	1	13	0	12	3
14	736557	1	14	0	12	3
15	736557	1	15	2	12	3
16	736557	1	16	6	12	3
17	736557	1	17	0	12	3
18	736557	1	18	10	12	3
19	736557	1	19	0	12	3
20	736557	1	20	16	12	3
21	736557	1	21	5	12	3
22	736557	1	22	9	12	3
23	736557	1	23	0	12	3
24	736557	1	24	0	12	3
25	736557	1	25	18	12	3
26	736557	1	26	0	12	3
27	736557	1	27	0	12	3
28	736557	1	28	0	12	3
29	736557	1	29	0	12	3
30	736557	1	30	256	12	3
31	736557	1	31	14	12	3
32	736557	1	32	0	12	3
33	736557	10	0	0	12	2
34	736557	10	1	0	12	2

	date	store_nbr	family	onpromotion	state	type_x
35	736557	10	2	2	12	2
36	736557	10	3	43	12	2
37	736557	10	4	0	12	2
38	736557	10	5	15	12	2
39	736557	10	6	0	12	2
40	736557	10	7	30	12	2
41	736557	10	8	106	12	2
42	736557	10	9	18	12	2
43	736557	10	10	1	12	2
44	736557	10	11	1	12	2
45	736557	10	12	96	12	2
46	736557	10	13	0	12	2
47	736557	10	14	0	12	2
48	736557	10	15	2	12	2
49	736557	10	16	6	12	2

In [15]: `final_df['dcoilwtico'].isna().sum()`

Out[15]: 7128

In [16]: `final_df['dcoilwtico'] = final_df['dcoilwtico'].fillna(final_df["dcoilwtico"].me`

In [18]: `final_df.isna().sum()`

Out[18]:

id	0
date	0
store_nbr	0
family	0
onpromotion	0
city	0
state	0
type_x	0
cluster	0
dcoilwtico	0
dtype:	int64

In [19]: `final_df.drop(['city','cluster','dcoilwtico'], axis=1, inplace =True)`

In [20]: final_df

Out[20]:

	id	date	store_nbr	family	onpromotion	state	type_x
0	3000888	2017-08-16	1	AUTOMOTIVE	0	Pichincha	D
1	3000889	2017-08-16	1	BABY CARE	0	Pichincha	D
2	3000890	2017-08-16	1	BEAUTY	2	Pichincha	D
3	3000891	2017-08-16	1	BEVERAGES	20	Pichincha	D
4	3000892	2017-08-16	1	BOOKS	0	Pichincha	D
...
28507	3029395	2017-08-31	9	POULTRY	1	Pichincha	B
28508	3029396	2017-08-31	9	PREPARED FOODS	0	Pichincha	B
28509	3029397	2017-08-31	9	PRODUCE	1	Pichincha	B
28510	3029398	2017-08-31	9	SCHOOL AND OFFICE SUPPLIES	9	Pichincha	B
28511	3029399	2017-08-31	9	SEAFOOD	0	Pichincha	B

28512 rows × 7 columns

In [21]: categorical = ['family', 'state', 'type_x']

```
In [22]: from sklearn.preprocessing import LabelEncoder
label_encoding = LabelEncoder()
for column in categorical:
    final_df[column] = label_encoding.fit_transform(final_df[column])
```

In [23]: final_df

Out[23]:

	id	date	store_nbr	family	onpromotion	state	type_x
0	3000888	2017-08-16	1	0	0	12	3
1	3000889	2017-08-16	1	1	0	12	3
2	3000890	2017-08-16	1	2	2	12	3
3	3000891	2017-08-16	1	3	20	12	3
4	3000892	2017-08-16	1	4	0	12	3
...
28507	3029395	2017-08-31	9	28	1	12	1
28508	3029396	2017-08-31	9	29	0	12	1
28509	3029397	2017-08-31	9	30	1	12	1
28510	3029398	2017-08-31	9	31	9	12	1
28511	3029399	2017-08-31	9	32	0	12	1

28512 rows × 7 columns

In []:

In [24]: final_df['date'] = pd.to_datetime(final_df['date'])

In [25]: **import** datetime **as** dt
 final_df['date']=final_df['date'].map(dt.datetime.toordinal)

In [26]: final_df

Out[26]:

	id	date	store_nbr	family	onpromotion	state	type_x
0	3000888	736557	1	0	0	12	3
1	3000889	736557	1	1	0	12	3
2	3000890	736557	1	2	2	12	3
3	3000891	736557	1	3	20	12	3
4	3000892	736557	1	4	0	12	3
...
28507	3029395	736572	9	28	1	12	1
28508	3029396	736572	9	29	0	12	1
28509	3029397	736572	9	30	1	12	1
28510	3029398	736572	9	31	9	12	1
28511	3029399	736572	9	32	0	12	1

28512 rows × 7 columns

In [28]: `from sklearn.preprocessing import StandardScaler`
`x_scaled = StandardScaler()`

In [29]: `ily", "onpromotion", "state", "type_x"] = x_scaled.fit_transform(final_df[["date", "s`

In [30]: final_df

Out[30]:

	id	date	store_nbr	family	onpromotion	state	type_x
0	3000888	-1.626978	-1.700267	-1.680336	-0.336759	0.769193	0.83205
1	3000889	-1.626978	-1.700267	-1.575315	-0.336759	0.769193	0.83205
2	3000890	-1.626978	-1.700267	-1.470294	-0.240064	0.769193	0.83205
3	3000891	-1.626978	-1.700267	-1.365273	0.630191	0.769193	0.83205
4	3000892	-1.626978	-1.700267	-1.260252	-0.336759	0.769193	0.83205
...
28507	3029395	1.626978	-1.186979	1.260252	-0.288411	0.769193	-0.83205
28508	3029396	1.626978	-1.186979	1.365273	-0.336759	0.769193	-0.83205
28509	3029397	1.626978	-1.186979	1.470294	-0.288411	0.769193	-0.83205
28510	3029398	1.626978	-1.186979	1.575315	0.098369	0.769193	-0.83205
28511	3029399	1.626978	-1.186979	1.680336	-0.336759	0.769193	-0.83205

28512 rows × 7 columns

```
In [31]: final_df.drop(["id"], axis = 1, inplace = True)
```

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

Out[32]:

	date	store_nbr	family	onpromotion	state	type_x
0	-1.626978	-1.700267	-1.680336	-0.336759	0.769193	0.83205
1	-1.626978	-1.700267	-1.575315	-0.336759	0.769193	0.83205
2	-1.626978	-1.700267	-1.470294	-0.240064	0.769193	0.83205
3	-1.626978	-1.700267	-1.365273	0.630191	0.769193	0.83205
4	-1.626978	-1.700267	-1.260252	-0.336759	0.769193	0.83205

```
In [39]: train = pd.read_csv("train.csv")
train.head()
```

Out[39]:

	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

```
In [40]: stores = pd.read_csv("stores.csv")
stores.head()
```

Out[40]:

	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 [41]: df_train = train.merge(stores , how = 'left', on = 'store_nbr')
df_train.head()
```

Out[41]:

	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 [42]: oil = pd.read_csv("oil.csv")
oil.head()
```

Out[42]:

	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 [43]: df_train_1 = df_train.merge(oil, how = 'left', on = 'date')
df_train_1.head()
```

Out[43]:

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

```
In [44]: events = pd.read_csv('holidays_events.csv')
events.head()
```

Out[44]:

	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 [45]: final_train_df = df_train_1.merge(events, how = 'left', on = 'date')
final_train_df.head()
```

Out[45]:

	id	date	store_nbr	family	sales	onpromotion	city	state	type_x	cluster	dcoil
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 [46]: final_train_df.isna().sum()
```

```
Out[46]: id                0
date                0
store_nbr           0
family              0
sales               0
onpromotion         0
city                0
state               0
type_x              0
cluster             0
dcoilwtico          955152
type_y              2551824
locale              2551824
locale_name         2551824
description          2551824
transferred         2551824
dtype: int64
```

```
In [49]: oil_median = final_train_df["dcoilwtico"].median()
```

```
In [50]: final_train_df["dcoilwtico"] = final_train_df["dcoilwtico"].fillna(oil_median)
```

```
In [51]: final_train_df.isna().sum()
```

```
Out[51]: id                0
         date              0
         store_nbr         0
         family            0
         sales             0
         onpromotion       0
         city              0
         state             0
         type_x            0
         cluster          0
         dcoilwtico        0
         type_y          2551824
         locale          2551824
         locale_name      2551824
         description      2551824
         transferred      2551824
         dtype: int64
```

```
In [52]: final_train_df.drop(["type_y","locale","locale_name","description","transferred"])
```

```
In [53]: final_train_df.head()
```

```
Out[53]:
```

	id	date	store_nbr	family	sales	onpromotion	city	state	type_x	cluster	dcoilwtico
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 [54]: final_train_df.drop(["city","cluster","dcoilwtico"],axis = 1 , inplace = True)
```

```
In [55]: final_train_df.head()
```

```
Out[55]:
```

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

```
In [56]: from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
In [57]: categorical = ["family", "state", "type_x"]
```

```
In [58]: label_encoder = LabelEncoder()
for column in categorical:
    final_train_df[column] = label_encoder.fit_transform(final_train_df[column])
```

```
In [59]: final_train_df.head()
```

Out[59]:

	id	date	store_nbr	family	sales	onpromotion	state	type_x
0	0	2013-01-01	1	0	0.0	0	12	3
1	1	2013-01-01	1	1	0.0	0	12	3
2	2	2013-01-01	1	2	0.0	0	12	3
3	3	2013-01-01	1	3	0.0	0	12	3
4	4	2013-01-01	1	4	0.0	0	12	3

```
In [60]: final_train_df['date'] = pd.to_datetime(final_train_df['date'])
```

```
In [61]: import datetime as dt
final_train_df['date'] = final_train_df['date'].map(dt.datetime.toordinal)
```

```
In [62]: final_train_df.head()
```

Out[62]:

	id	date	store_nbr	family	sales	onpromotion	state	type_x
0	0	734869	1	0	0.0	0	12	3
1	1	734869	1	1	0.0	0	12	3
2	2	734869	1	2	0.0	0	12	3
3	3	734869	1	3	0.0	0	12	3
4	4	734869	1	4	0.0	0	12	3

```
In [83]: final_df.head()
```

Out[83]:

	date	store_nbr	family	onpromotion	state	type_x
0	-1.626978	-1.700267	-1.680336	-0.336759	0.769193	0.83205
1	-1.626978	-1.700267	-1.575315	-0.336759	0.769193	0.83205
2	-1.626978	-1.700267	-1.470294	-0.240064	0.769193	0.83205
3	-1.626978	-1.700267	-1.365273	0.630191	0.769193	0.83205
4	-1.626978	-1.700267	-1.260252	-0.336759	0.769193	0.83205

In []:

In [63]: `scaler = StandardScaler()`In [64]: `final_train_df[["date","store_nbr","family","onpromotion","state","type_x"]] = scaler.fit_transform(final_train_df[["date","store_nbr","family","onpromotion","state","type_x"]])`In [65]: `final_train_df.head()`

Out[65]:

	id	date	store_nbr	family	sales	onpromotion	state	type_x
0	0	-1.734609	-1.700267	-1.680336	0.0	-0.213586	0.769193	0.83205
1	1	-1.734609	-1.700267	-1.575315	0.0	-0.213586	0.769193	0.83205
2	2	-1.734609	-1.700267	-1.470294	0.0	-0.213586	0.769193	0.83205
3	3	-1.734609	-1.700267	-1.365273	0.0	-0.213586	0.769193	0.83205
4	4	-1.734609	-1.700267	-1.260252	0.0	-0.213586	0.769193	0.83205

In [67]: `final_train_df.drop(['id'], axis =1, inplace=True)`In [68]: `final_train_df`

Out[68]:

		date	store_nbr	family	sales	onpromotion	state	type_x
0		-1.734609	-1.700267	-1.680336	0.000	-0.213586	0.769193	0.83205
1		-1.734609	-1.700267	-1.575315	0.000	-0.213586	0.769193	0.83205
2		-1.734609	-1.700267	-1.470294	0.000	-0.213586	0.769193	0.83205
3		-1.734609	-1.700267	-1.365273	0.000	-0.213586	0.769193	0.83205
4		-1.734609	-1.700267	-1.260252	0.000	-0.213586	0.769193	0.83205
...	
3054343		1.726369	-1.186979	1.260252	438.133	-0.213586	0.769193	-0.83205
3054344		1.726369	-1.186979	1.365273	154.553	-0.131986	0.769193	-0.83205
3054345		1.726369	-1.186979	1.470294	2419.729	11.863180	0.769193	-0.83205
3054346		1.726369	-1.186979	1.575315	121.000	0.439212	0.769193	-0.83205
3054347		1.726369	-1.186979	1.680336	16.000	-0.213586	0.769193	-0.83205

In [72]: `x = final_train_df.drop(['sales'], axis =1)
y = final_train_df['sales']`In [73]: `from sklearn.tree import DecisionTreeRegressor
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.2, random_state = 42)
dt_model = DecisionTreeRegressor(random_state = 42)
dt_model.fit(x_train,y_train)`Out[73]: `DecisionTreeRegressor(random_state=42)`

```
In [77]: y_test_pred = dt_model.predict(x_test)
```

```
In [80]: from sklearn.metrics import mean_squared_log_error as rmse
print('RMSE :',rmse(y_test,y_test_pred))
```

RMSE : 0.4179111831844443

```
In [85]: final_df
```

Out[85]:

	date	store_nbr	family	onpromotion	state	type_x
0	-1.626978	-1.700267	-1.680336	-0.336759	0.769193	0.83205
1	-1.626978	-1.700267	-1.575315	-0.336759	0.769193	0.83205
2	-1.626978	-1.700267	-1.470294	-0.240064	0.769193	0.83205
3	-1.626978	-1.700267	-1.365273	0.630191	0.769193	0.83205
4	-1.626978	-1.700267	-1.260252	-0.336759	0.769193	0.83205
...
28507	1.626978	-1.186979	1.260252	-0.288411	0.769193	-0.83205
28508	1.626978	-1.186979	1.365273	-0.336759	0.769193	-0.83205
28509	1.626978	-1.186979	1.470294	-0.288411	0.769193	-0.83205
28510	1.626978	-1.186979	1.575315	0.098369	0.769193	-0.83205
28511	1.626978	-1.186979	1.680336	-0.336759	0.769193	-0.83205

28512 rows × 6 columns

```
In [86]: test_data = final_df.tail(28512)
```

```
In [88]: pred = dt_model.predict(test_data)
```

```
In [89]: df_sample = pd.read_csv("sample_submission.csv")
```

```
In [90]: df_sample.head()
```

Out[90]:

	id	sales
0	3000888	0.0
1	3000889	0.0
2	3000890	0.0
3	3000891	0.0
4	3000892	0.0

```
In [91]: df_sample["sales"] = pred
```


In [92]: `df_sample.head()`

Out[92]:

	id	sales
0	3000888	2.0
1	3000889	0.0
2	3000890	1.0
3	3000891	2623.0
4	3000892	0.0

In []: