Food sales weather...

```
// Imports
import org.apache.spark.sql.functions._
import org.joda.time.format.DateTimeFormat
```

FINISHED

```
import org.apache.spark.sql.functions._
import org.joda.time.format.DateTimeFormat
```

Took 1 sec. Last updated by anonymous at March 30 2017, 7:10:21 PM.

```
%pyspark
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
sns.set_style('whitegrid')
```

FINISHED

Took 3 sec. Last updated by anonymous at March 30 2017, 8:04:11 PM.

```
%pyspark
df_train = pd.read_csv("/Users/vpandiyan/Downloads/train.csv")
```

FINISHED

```
sys:1: DtypeWarning: Columns (7) have mixed types. Specify dtype option on import or set low_m
emory=False.
```

Took 2 sec. Last updated by anonymous at March 30 2017, 8:07:27 PM.

```
%pyspark
df_train.head()
```

FINISHED

```
   Store  DayOfWeek        Date  Sales  Customers  Open  Promo StateHoliday  \
0      1          5  2015-07-31   5263        555     1      1            0
1      2          5  2015-07-31   6064        625     1      1            0
2      3          5  2015-07-31   8314        821     1      1            0
3      4          5  2015-07-31  13995       1498     1      1            0
4      5          5  2015-07-31   4822        559     1      1            0
   SchoolHoliday
0              1
1              1
2              1
3              1
4              1
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:08:01 PM.

```
%pyspark
df_train = pd.read_csv("/Users/vpandiyan/Downloads/train.csv")
df_store = pd.read_csv("/Users/vpandiyan/Downloads/store.csv")
df_test = pd.read_csv("/Users/vpandiyan/Downloads/test.csv")
```

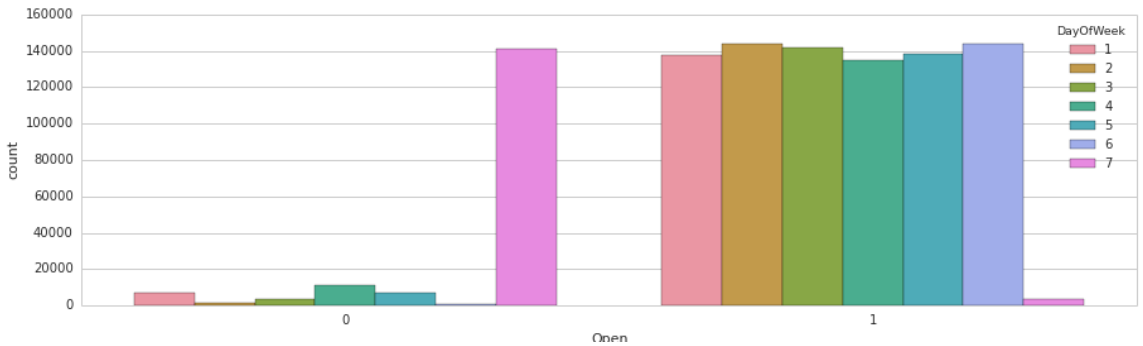
FINISHED

Took 2 sec. Last updated by anonymous at March 30 2017, 8:09:07 PM.

FINISHED

```
%pyspark
fig, (axis1) = plt.subplots(1,1,figsize=(15,4))
sns.countplot(x = 'Open', hue = 'DayOfWeek', data = df_train,)
```

```
<matplotlib.axes.AxesSubplot object at 0x1114debd0>
```



Took 1 sec. Last updated by anonymous at March 30 2017, 8:09:29 PM.

FINISHED

```
%pyspark
df_train.groupby('Date')["Sales"].mean()
```

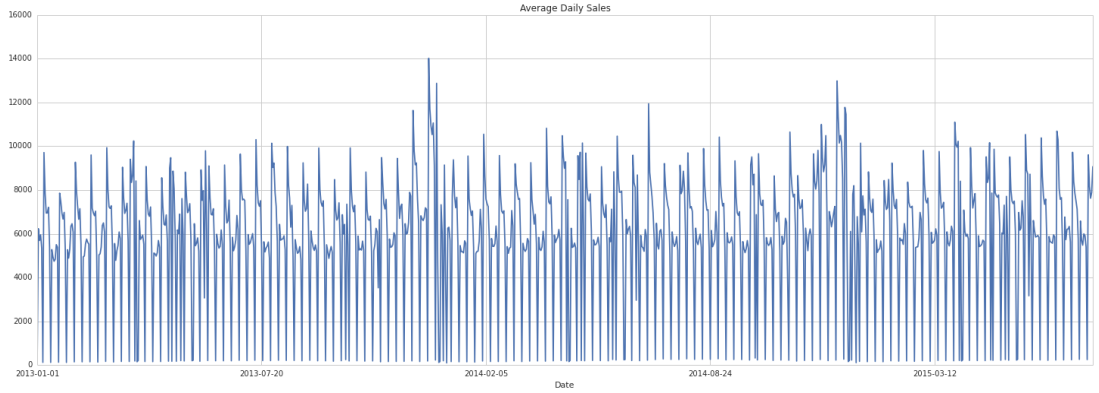
```
Date
2013-01-01      87.284560
2013-01-02    6233.030493
2013-01-03    5693.112108
2013-01-04    5954.218834
2013-01-05    5337.751570
2013-01-06     129.061883
2013-01-07    9710.177578
2013-01-08    7847.028700
2013-01-09    6947.626009
2013-01-10    6952.004484
2013-01-11    7210.139910
2013-01-12    5396.852915
2013-01-13     129.194619
2013-01-14    5279.630493
2013-01-15    4944.027803
2013-01-16    4747.103139
2013-01-17    4830.999103
```

Took 0 sec. Last updated by anonymous at March 30 2017, 8:12:42 PM.

FINISHED

```
%pyspark
average_daily_sales = df_train.groupby('Date')["Sales"].mean()
fig = plt.subplots(1,1,sharex=True,figsize=(25,8))
average_daily_sales.plot(title="Average Daily Sales")
```

```
<matplotlib.axes.AxesSubplot object at 0x115274390>
```

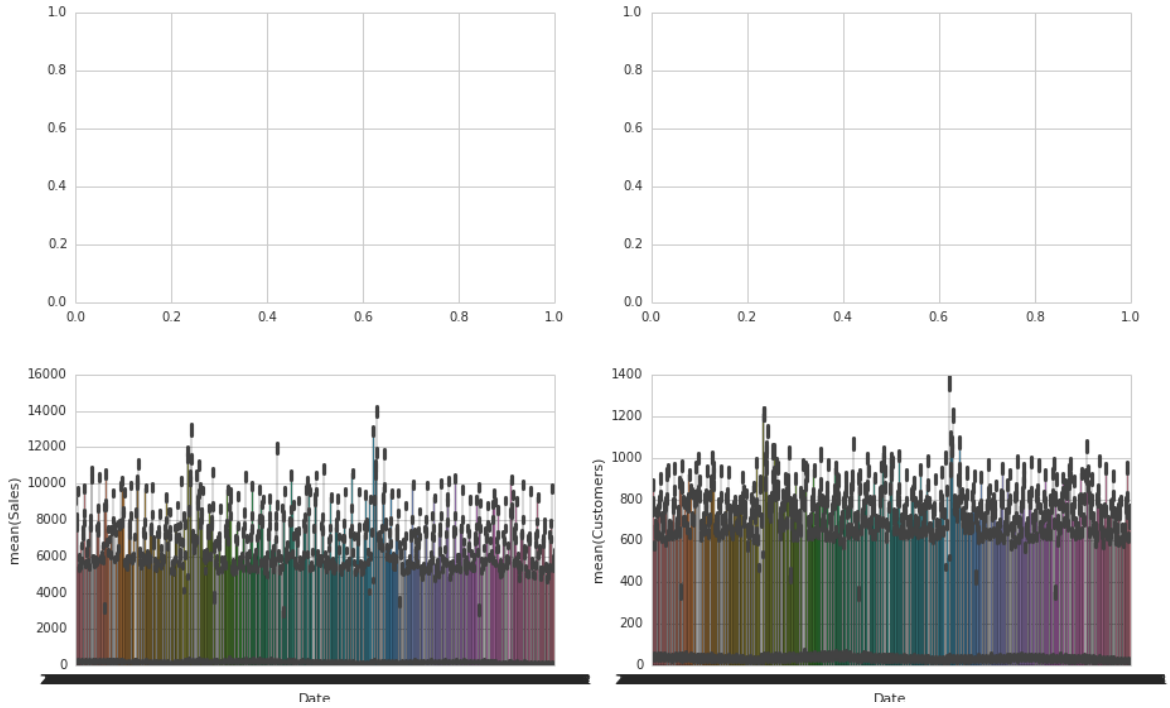


Took 1 sec. Last updated by anonymous at March 30 2017, 8:17:25 PM.

FINISHED

```
%pyspark
fig, (axis1,axis2) = plt.subplots(1,2,figsize=(15,4))

sns.barplot(x ='Date', y ='Sales', data = df_train, ax=axis1)
sns.barplot(x ='Date', y ='Customers', data = df_train, ax=axis2)
```

```
<matplotlib.axes.AxesSubplot object at 0x113b4edd0>
```

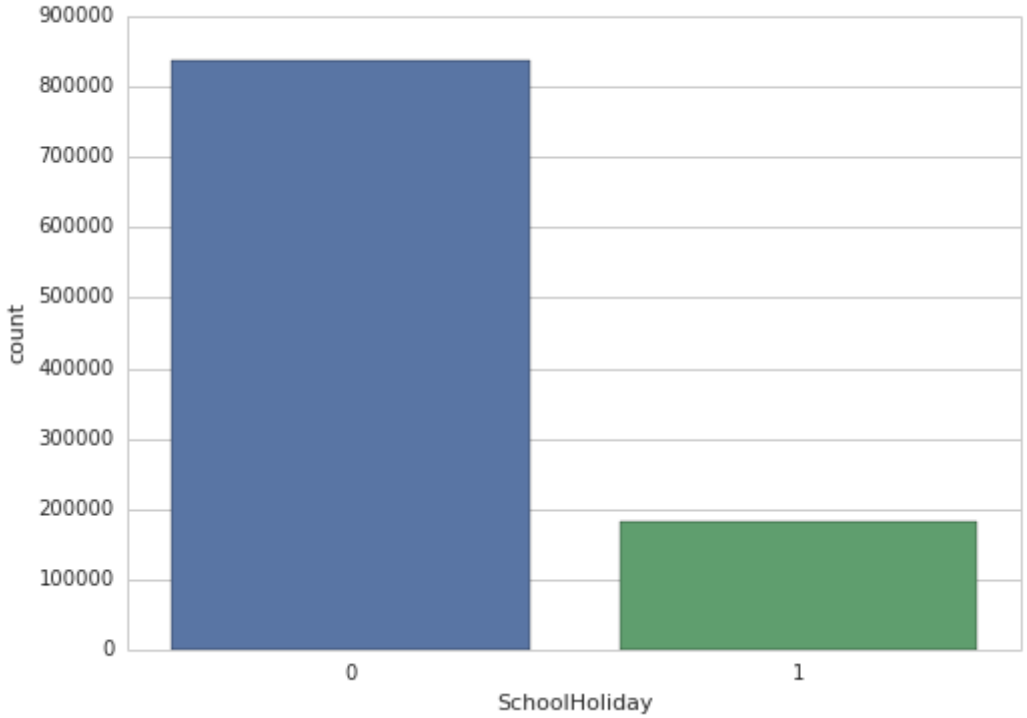


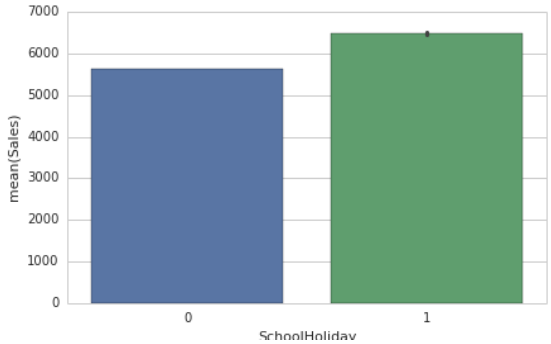
Took 3 min 22 sec. Last updated by anonymous at March 30 2017, 8:22:29 PM.

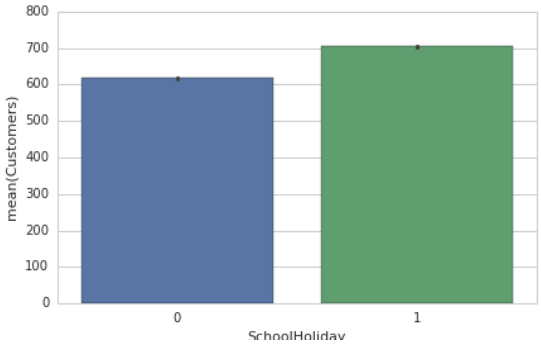
FINISHED

```
%pyspark
sns.countplot(x='SchoolHoliday', data= df_train)

fig, (axis1,axis2) = plt.subplots(1,2,figsize=(15,4))

sns.barplot(x='SchoolHoliday', y='Sales', data= df_train, ax=axis1)
sns.barplot(x='SchoolHoliday', y='Customers', data= df_train, ax=axis2)
```

```
<matplotlib.axes.AxesSubplot object at 0x116838c10>
```







Took 1 min 41 sec. Last updated by anonymous at March 30 2017, 8:33:27 PM.

```
%pyspark
```

READY