Walmart Sales Forecasting Milestone Report

Problem

The aim of this project is to predict the weekly department-wise sales of 45 Walmart stores in the US using historical data

The ability to forecast sales for any business is of crucial importance. With an insight on what the expected sales would be in the near future the businesses can plan accordingly. The managers are able to schedule shifts, allocate resources, stock inventory etc based on the forecast.

The dataset is taken from Kaggle from a competition hosted by Walmart.

Client

A lot of businesses have a somewhat homogenous sales trend over time except for in the Holiday seasons. Being able to accurately predict the sales in the Holiday seasons is even more important. The challenge in predicting Holiday sales is that there isn't much data available. This problem can be applied to any industry/business whose sales are affected by the holiday seasons. They can use this information to manage resources, complete inventory, launch applicable promotions etc.

Dataset

The data that I will be using is of 45 Walmart stores. The dataset can be easily downloaded from Kaggle. It contains department wise weekly sales for each store for over 2 years, along with 12 features.

The data in available as 4 tables:

- **Train Sales** - Store wise department wise weekly sales for 2 years
- **Test Sales** - Store, department and dates
- **Features** - This contains the stores region characteristics like fuel price, temperature, Consumer affluence, unemployment percentage and active promotions over time
- **Store** - This defines the type and size of each store.

The columns descriptions are:

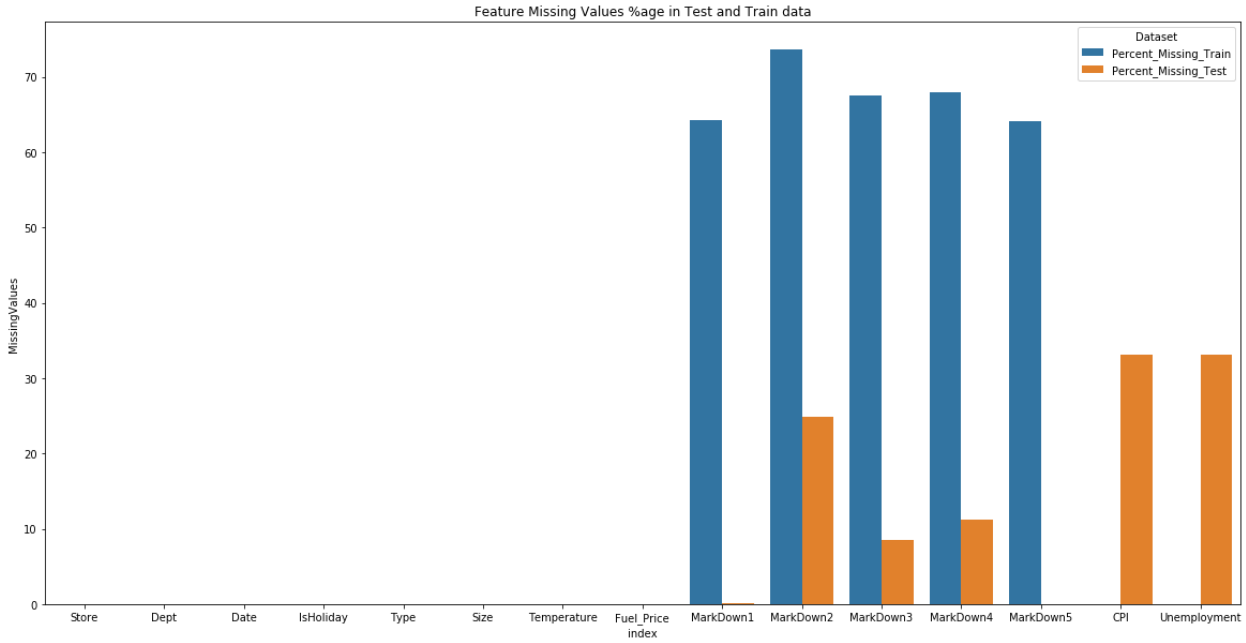
- Store - the store number
- Dept - the department number

- Date - the week
- Weekly_Sales - sales for the given department in the given store (Only present in train data)
- IsHoliday - whether the week is a special holiday week
- Type - *Description not given*
- Size - *Description not given but is self explanatory*
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time.
- CPI - the consumer price index
- Unemployment - the unemployment rate

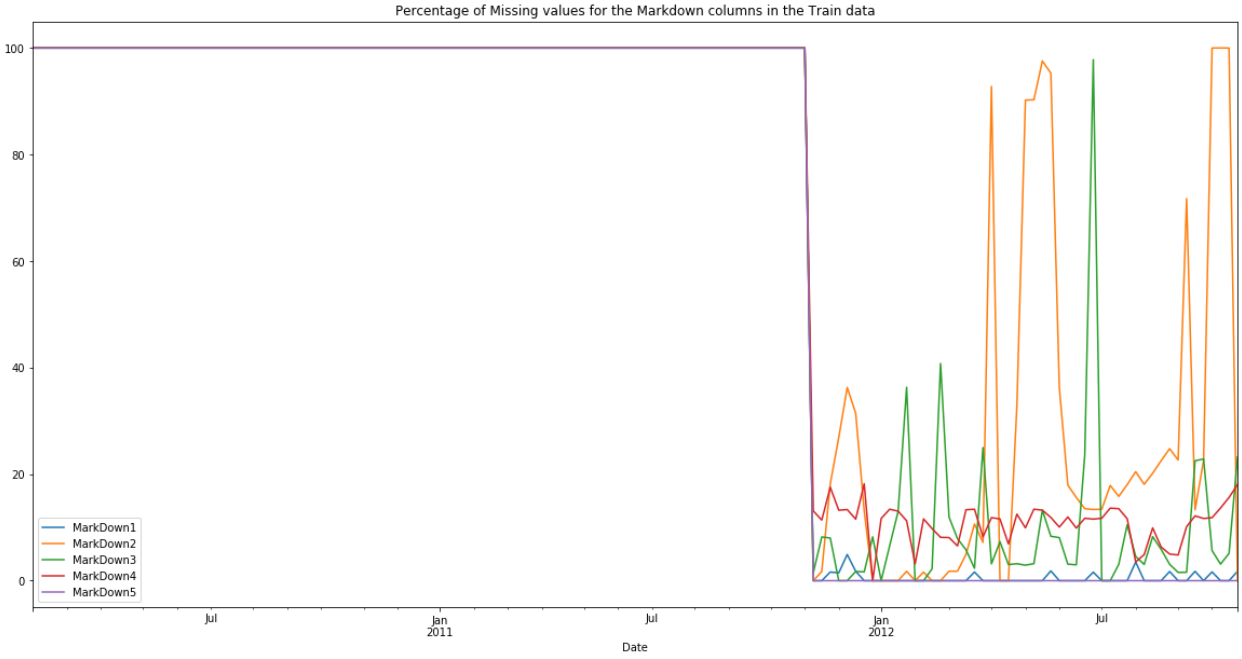
Data Wrangling

First, we input data & perform data wrangling. We conduct the following steps to clean up the data

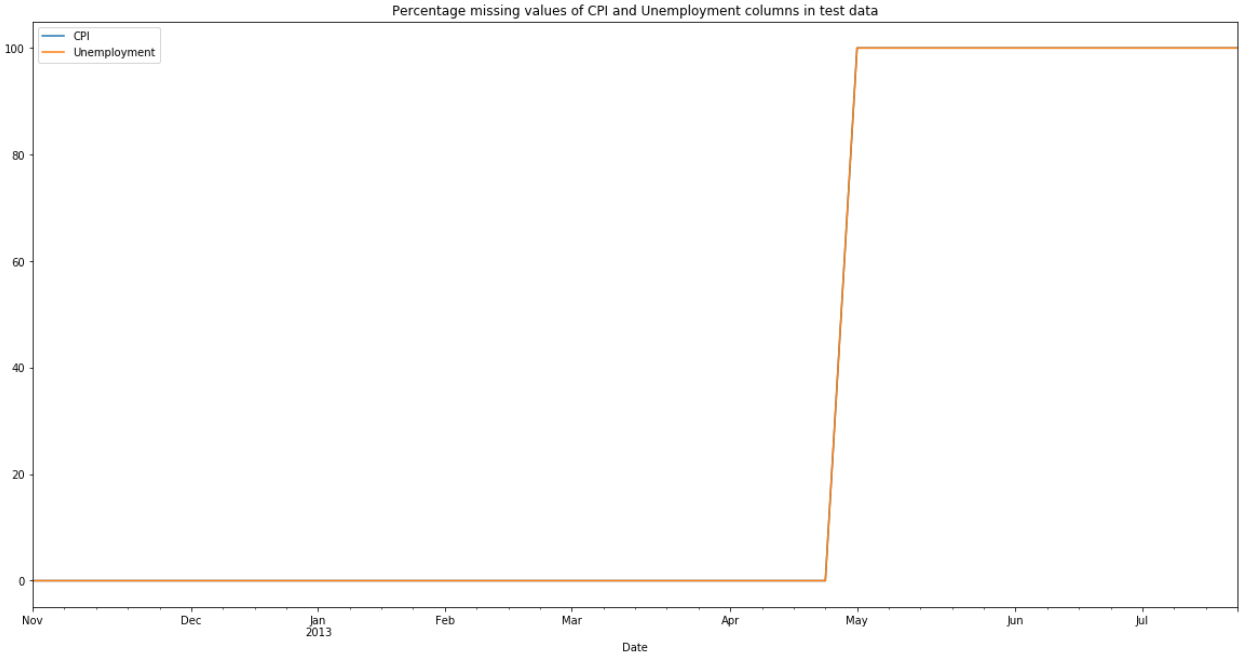
1. **Joining the tables:**
    - We join the tables together to get all the features in one table. The final table resulted in a total of 15 Features for 415,000 observations in the training data.
2. **Missing values.**
    - It was seen that there were some columns where there were missing values.We can see that most of the missing values are in the train data are in the Markdown Columns. While in the test data ⅓ of CPI and Unemployment are missing.



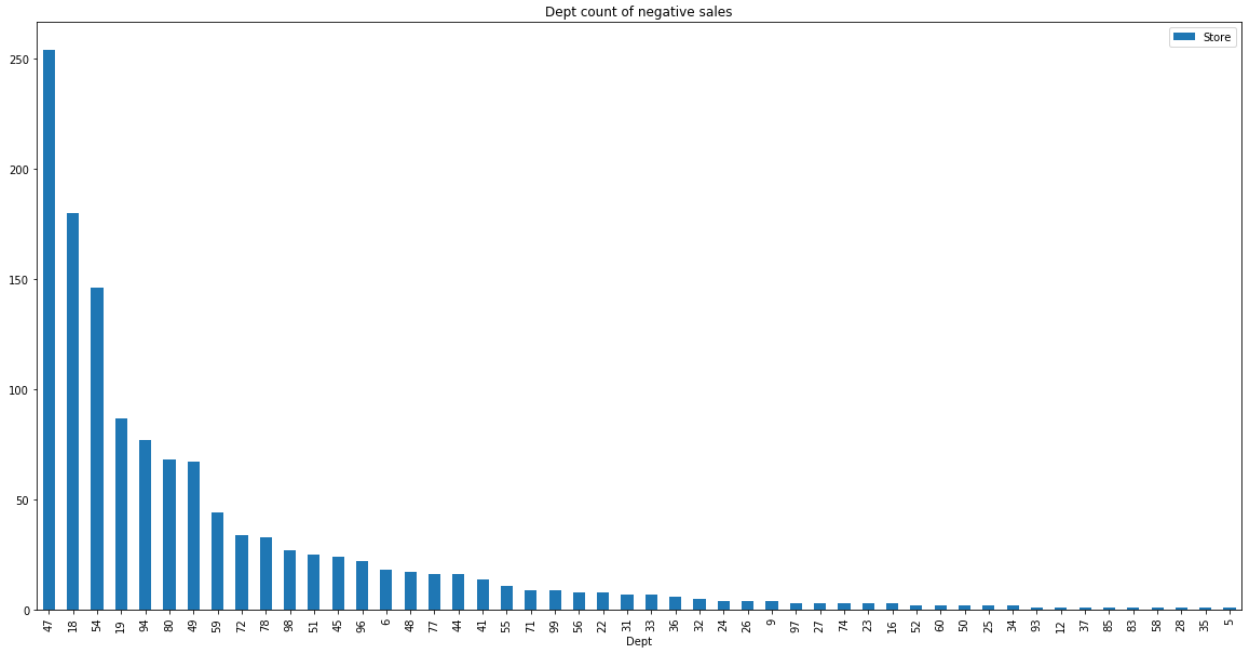
- Looking at the missing values for the Markdown Columns we see that all the markdown values are missing before Nov 2011. This implies that we only have 1 year of the Markdown columns. This issue renders implies that the markdown features would not be good predictors of the price since we won't be able to capture the yearly trend in these columns.



- Looking at the CPI and Unemployment missing features we noticed that they were missing in only the Test data. From the plot below, it can be seen that the last 3 months of the test data CPI and Unemployment columns are missing. This lead to the question as to how they are populated.



3. **Negative Sales**
    - It was observed that there were 1285 negative values which is approximately 0.305 perc of all of the data. A department wise plot is shown below.



    - The competition posters(Walmart) have not given any reason as to why that would be. A logical reasoning would be that these would be weeks where the returns were more that the sales for these weeks in these departments.

Other than the issues highlighted the data is mostly clean.

Takeaways

1. In the training data Markdowns are only available for 9 months. The test data is also for a duration of 9 months. This makes the Markdown columns not a very effective feature since the yearly trend will not be available for these features.
2. Unemployment and CPI are missing for a third of the test data. This is also a problem that will need to be dealt to make model predictions.
3. There isn't any systematic trend in the missing values in any of the columns
4. Will need to further analyse Markdown columns to see if they can be used for modeling purposes.
5. Will impute 0 in the missing values for now for all columns. Maybe a better strategy can be devised later on.

Exploratory Data Analysis

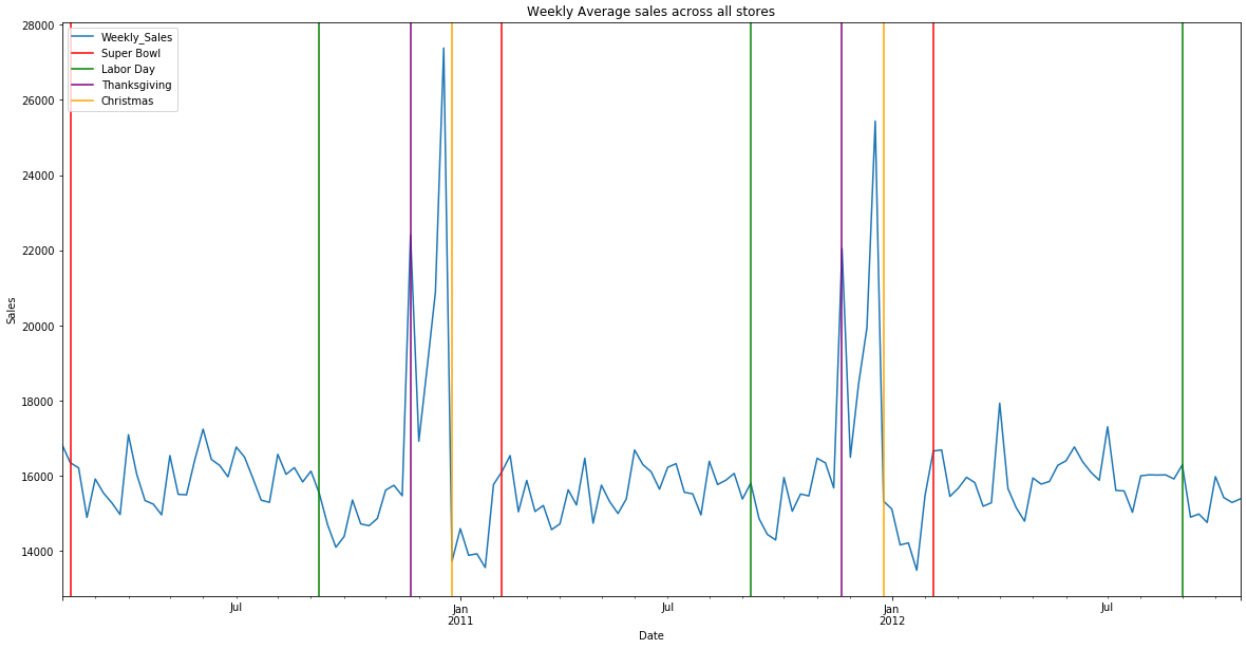
Note: The problem that we are trying to solve is fundamentally a time-series problem. To convert this into a conventional time series problem we need to find/develop/extract features that depict the Weekly_Sales trend over time.

During exploratory data analysis, we ask the following conclusions were observed:

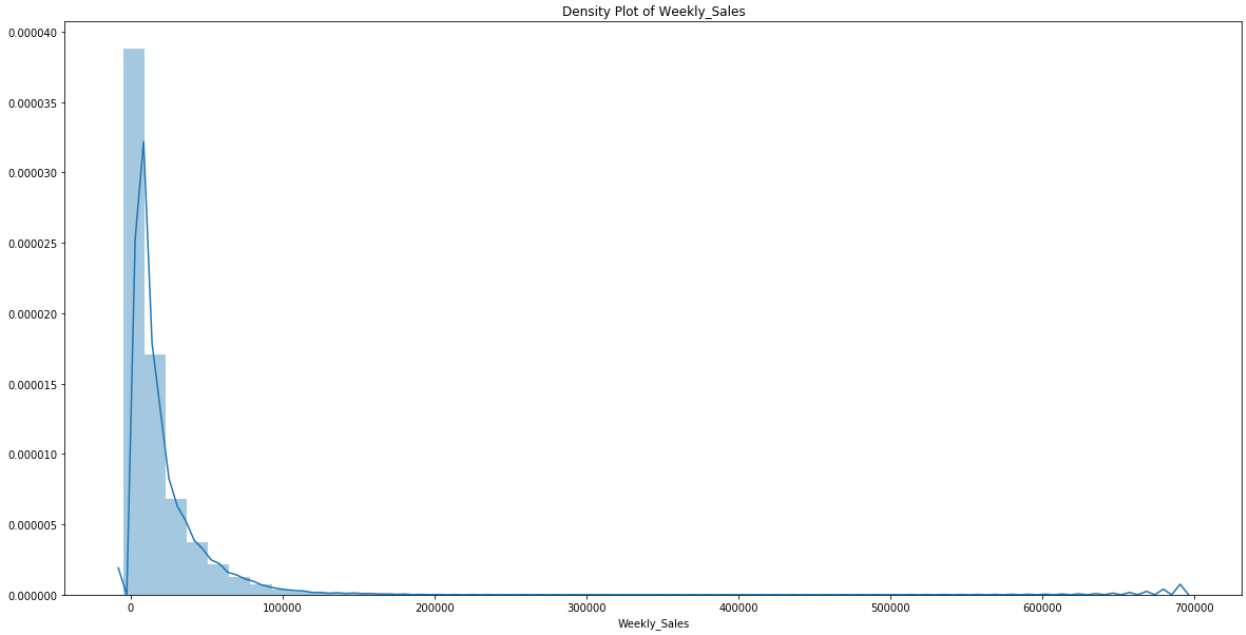
1. **Weekly_Sales trend is affected by some holidays but not all.**

   The following plot shows the trend of the average weekly sales. The vertical lines depict the holidays.The following observations can be made from the graph below

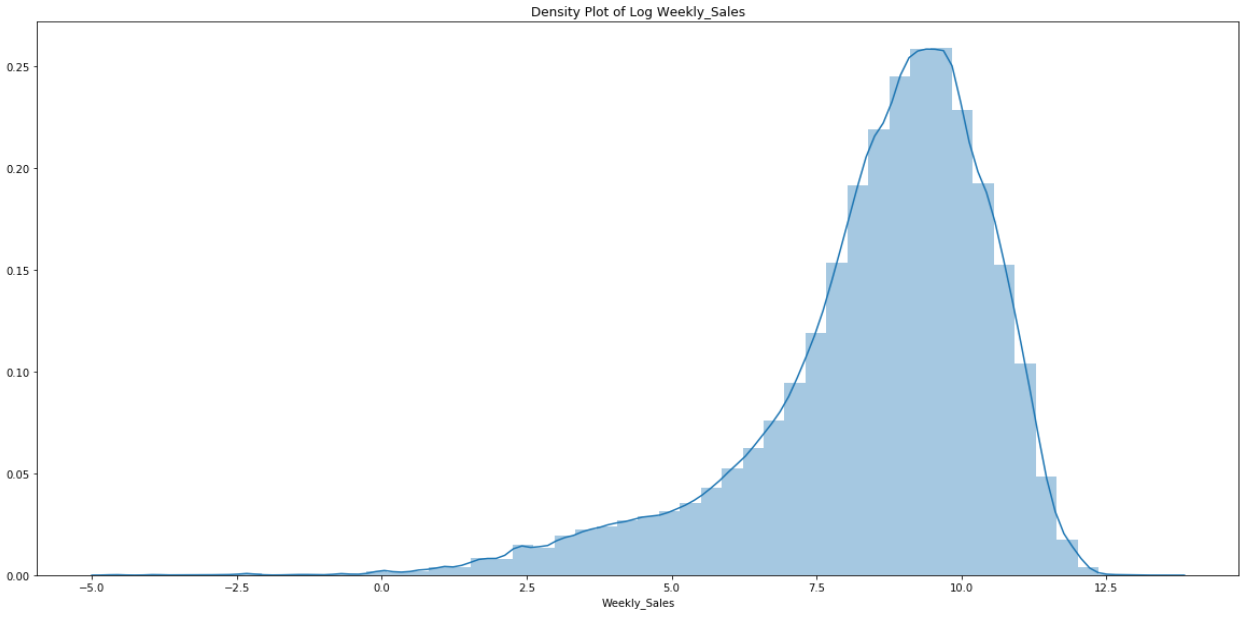
   i. Thanksgiving and Christmas holidays affect Weekly Sales more than Superbowl and Labor Day.
   ii. The sales sharply drop after all holidays except SuperBowl.
   iii. There is an offset in the weekly_sales for christmas. I can think of 2 explanations for this.
      1. The sales occur a week before the actual holiday
      2. The values of the sales are provided on a weekly basis. But the holiday is on a day. The offset maybe due to the placement of that day within the week



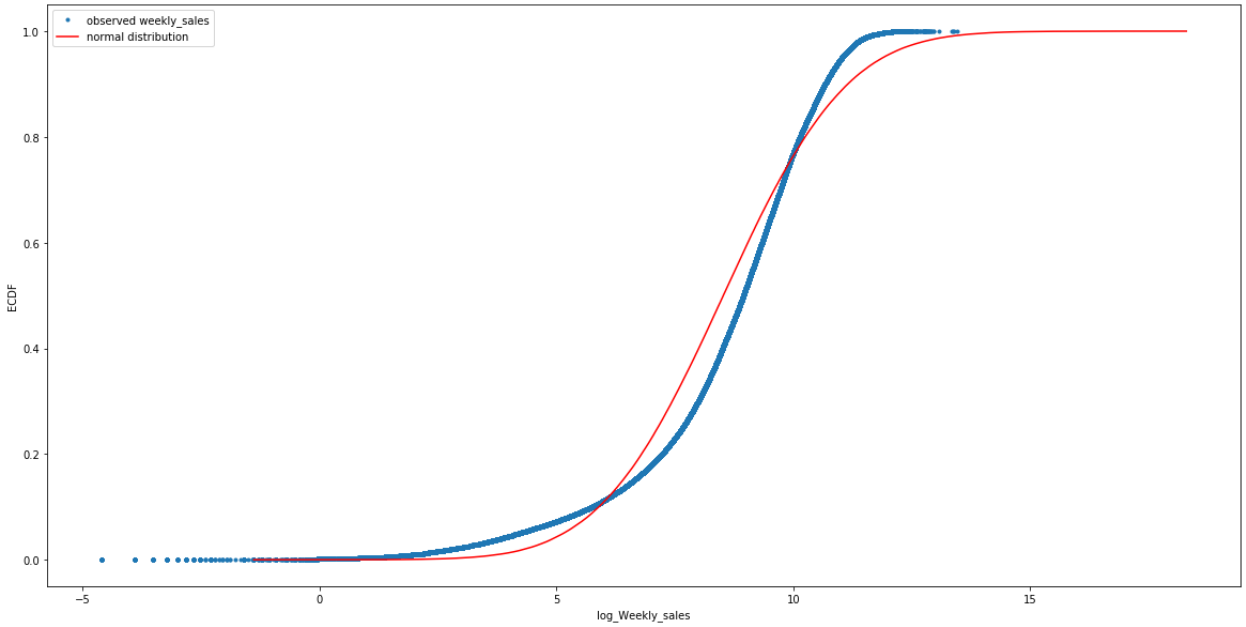
2. **The target variable (Weekly_Sales) is NOT normally distributed.**
    a. The distribution of weekly_Sales is very skewed. This limits the effectiveness of most of the parametric machine learning models which rely on the assumption that the distribution is normal.



    b. Let's look at the distribution of log scaled weekly_sales.This looks somewhat normally distributed but there is still a skew to the left



c. Performing a statistical test revealed that even the Log scaled weekly_sales aren't normally distributed



3. Most important features in predicting Weekly_Sales

a. Upon performing some analysis on the data set the following features emerged as being the most highly correlated with Weekly_Sales. Most of these features were either extracted or engineered.

i. **CPI_Cat and Size_Cat**: CPI (Customer Propensity Index) and Size (of Store) were available in the training data as continuous values. These were discretized following along the rational that came from the Exploratory Data analysis.

ii. **Which_holiday:** The training data had a column for 'IsHoliday'. It was realized during EDA that each holiday had a different trend on the weekly sales. Some holidays had virtually no effect on the sales trends e.g. Labor Day as shown below. As a result feature was created for each holiday

iii. **TillNext{Holiday}, SinceNext{Holiday}:** These features were created to depict the number of weeks since a particular holiday has passed and the number of weeks until the next holiday comes. These were created for 4 holidays thus a total of 8 features

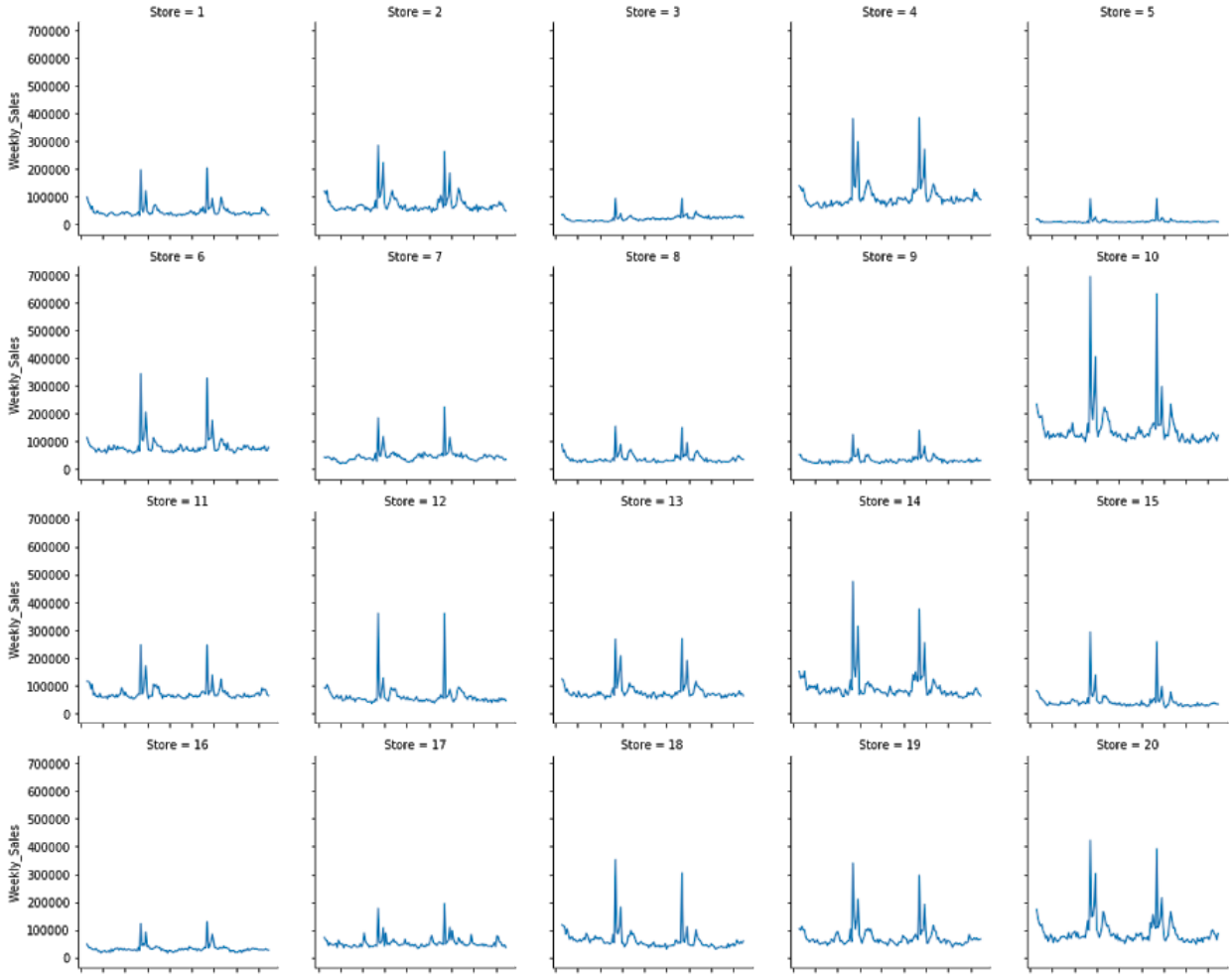
iv. **DateLagFeatures**: This feature depicts the Weekly Sales for each Store-Dept pair '**x**' weeks from the current week.

v. **DateFeatures:** This features were derived from the date of the observation. These include, Quarter,Month,Year,WeekOfMonth,WeekOfYear

vi. **Department_Contribution:** This feature is the ratio of each Department's average monthly sales with the Store's average monthly sales. The rational here is that the ratio would be a measure of how much the department's sales contribute to the sales of the store for each month.
vii. **Store_dept_month_avg**: This is the average sale for that store,dept pair for that month

4. Consistency of each department's sales trends

a. Upon visual inspection of the sales trend of a department across different Stores it was observed that most departments show similar trends if normalized against the stores total sales volume. A plot of one store is below.



5. **Redundant features**

   It was observed that all the features other than the ones mentioned above showed very weak correlation with the Weekly_Sales and thus were discarded.

Takeaways

From the analysis we can conclude the following:

i. *Weekly Sales should be transformed to the logarithmic scale so that the distribution isn't as skewed*
ii. *Weekly sales Outliers problem is somewhat sorted by transforming to the logarithmic scale, however, there might still be need to look into outliers that appear the scale change. This can be done in the modeling process*
iii. *Some more features can also be engineered however their effectiveness can only be judged when making predictions.*
iv. *Its best to use non parametric algorithms to solve the problem since they are less affected by the features or the target variable being not normal*

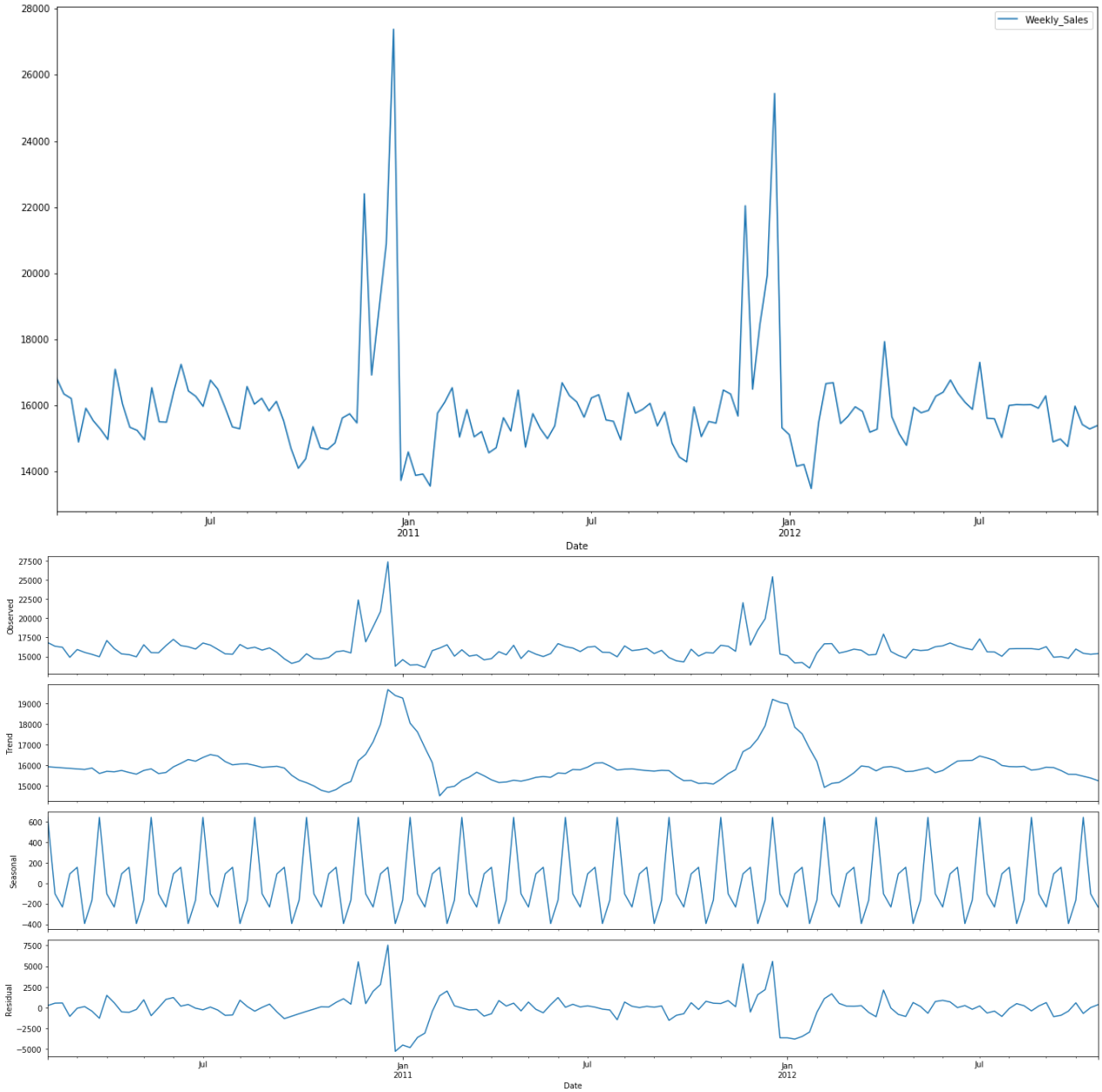
Statistical Analysis

1. **Testing for Stationarity**

   The sales predicting problem can also be approached by treating it like a time series. Most of the time series models, require that the data be non-stationary. We test for the assumption of stationarity in the series

   A statistical test called the ADFuller test was conducted. The result showed that the series is not stationary, thus if conventional Time series approaches are to be used then the necessary adjustments would need to be made to transform it to being a stationary series.

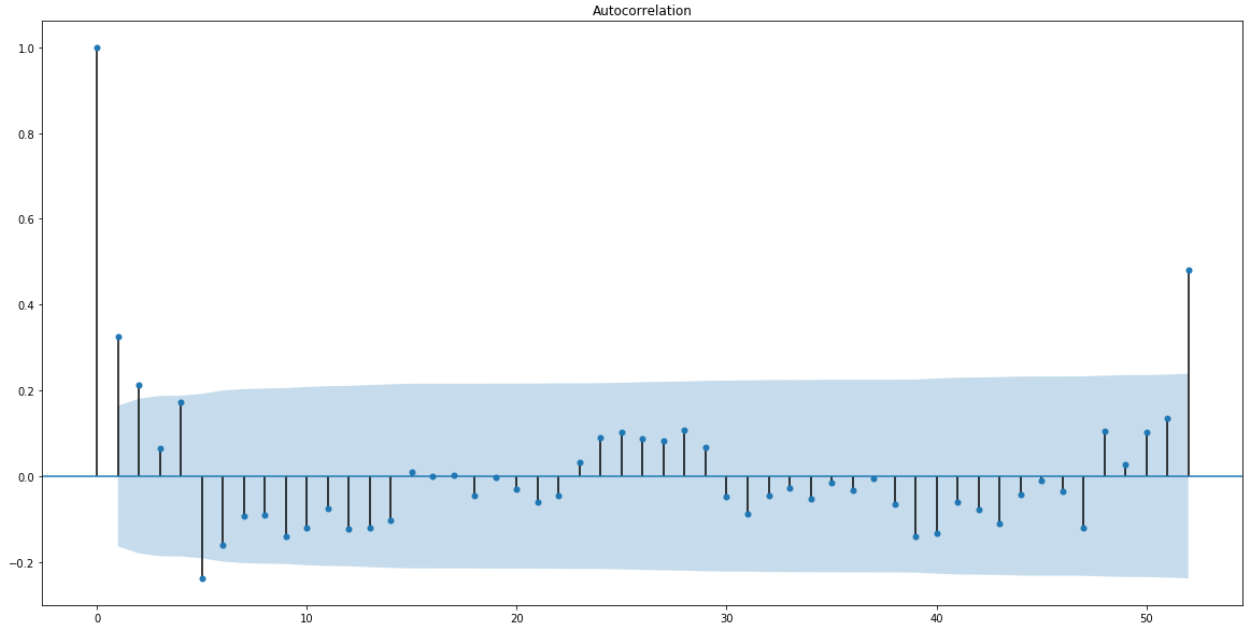
2. Series Decomposition



Decomposed series shows that the observed series is governed majorly by the trend which spikes during the months of Dec - Jan. We also see that the highest residual values are during that period as well. Other than that we notice an intra monthly seasonal trend.

3. AutoCorrelation

The weekly_sales are highly correlated with the previous 4 week's sales. The highest correlation is with last years week.



Takeaways

1. *Additional transformations will need to be made if using Time-Series based techniques*
2. *The previous weeks sales can be used as a feature*