**Project Title: Walmart Time-Series Weekly Sales Forecast**

**Auther:**Kai Chung, Ying

**Background: Goals/ Motivation/ Importance**

The trend of e-commerce blows sky high in the recent years, a lot of mortar retail cooperation turns themselves into the competition of e-commerce (e.g. Walmart , Target etc). The concept of this project is to utilize the dataset provided by Walmart to forecast the weekly sales in 45 stores based on set of features (e.g. Store Size,Unemployment rate, Customer Price Index (CPI) etc). The first part of data analysis result is able to reflect the past or current corporation financial condition for either single store or particular department. In addition, this experiment also explores the fact that how every feature impact on the business outcome which is one of crucial operations for a corporation. In the second part of this analysis, predictive time-series model will be implemented to forecast the future sales in weekly basis. This analysis step is basically an indispensable step for a company to have a guide or estimate of the future direction of company's profit or even market stock price which helps the management team to prepare business strategy for the next round of period.

**Client / Audience**

The primary audience for this analysis is Walmart management team / executives. The findings from this research gives some hint for them to design their operation format or business plan. Furthermore, the other group of audience properly are those professional who concerns about the business condition about Walmart. The findings from the research definitly offers them some insight about the current or future sales in the mortar stores. The third group would be the people who wants to find out the sales pattern happening in Walmart store or particular department. The analysis also gives the insight how the seasonal factors changes the sales which is very valuable for professionals or merchant.

**Questions to be addressed**

- Discover the sales distribution in Store and individual department

- Discover the correlation of between features and Sales.

- Discover the features contributes differently to the predictive models

- Determine the best model(s) for forecasting the weekly sales (continuous values)?

- How’s the models perform differently in various States?

**Proposed Methodology and Techniques**

This project is entirely done by Python 3.6 version and the code is implemented in Python notebook environment.

**Dataset Exploration**

1. Dataset Summary:

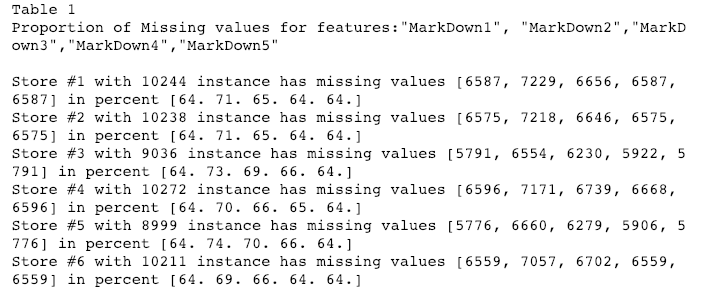
There are 421570 instances in this dataset. Among the dataset, there are total 45 retail Stores sales result in this dataset. In each store, there are about 99 Departments. The Target dependent variable which is Weekly\_sales is numeric continuous type of values. Independent features are shown as the following:

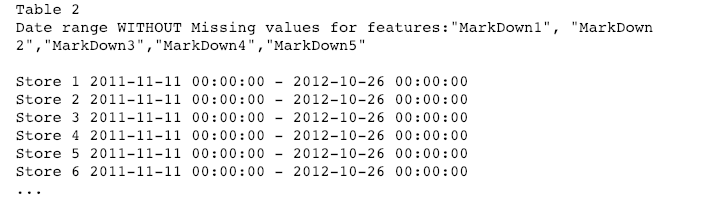
* + isHoliday(bnary)
  + Temperature(continuous)
  + Fuel Price(continuous)
  + Customer Price Index CPI(continuous)
  + Store Type(Binary)
  + Unemployment Rate(continuous)
  + MarkDown1
  + MarkDown2
  + MarkDown3
  + MarkDown4
  + MarkDown5

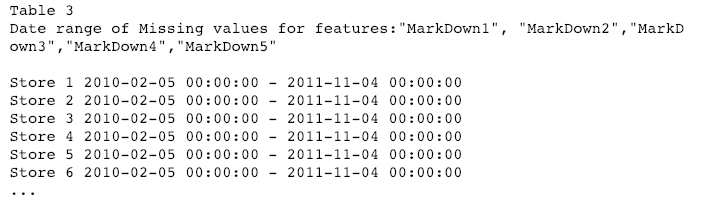
1. Missing Values:

Before we get into the analysis, we first deal with the missing values. The missing values are constantly stacking on Five features which areMarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5. Besides, missing values are not found among other features. For this reason, we are going to grub deeper into those features to check where those missing values are correlatable enough to be replaced.

The following is the basic proportion of missing values for every store. Based on the table 1, we could observe that the proportion of missing among these five features are ranging between (63%-90%). With these high volumes of missing values, we are going to grub more to check whether the values are randomly missing or not. Table 2 shows that the range of data without missing values from 2011-11-11 to 2012-10-26 and Table 3 shows that the range of data with missing values from 2010-02-05 to 2011-11-04. In this case, we could observe that the large volume of missing values are NOT randomly missed. In the consideration of the huge volumes as well as the missing patterns, I would propose to drop the five features (i.e. columns) instead of records (i.e. rows) because this is important to keep as many records as we could for the time-series analysis.

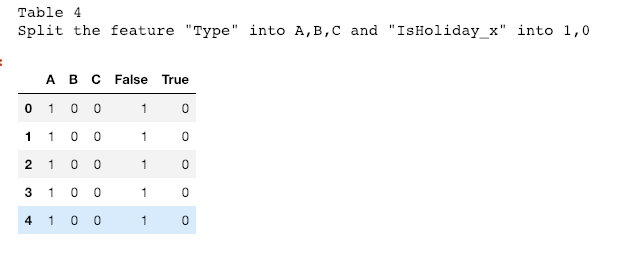






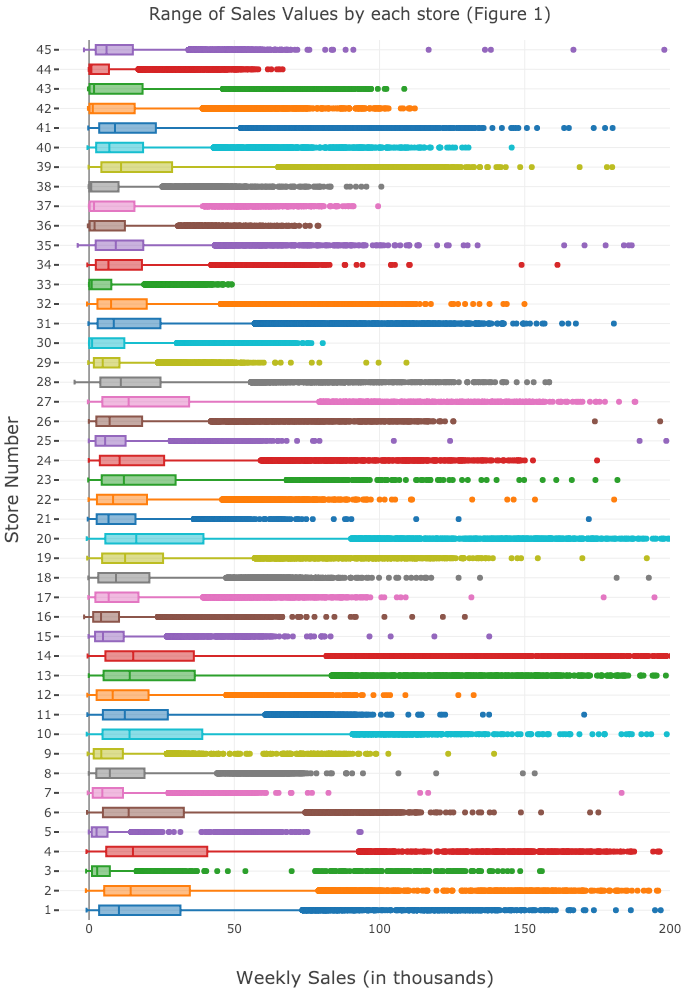
1. Categorical Features Conversion:

This dataset contains two categorical features what are  IsHoliday and Store\_Type variables. In order to proceed time-series analysis, these categorical attributes will be converted into individual variables as Table 4. As you see, the binary "IsHoliday" and "Store\_Type" features are divided into 2 columns (i.e. 1,0) and 3 columns (i.e. A,B,C)

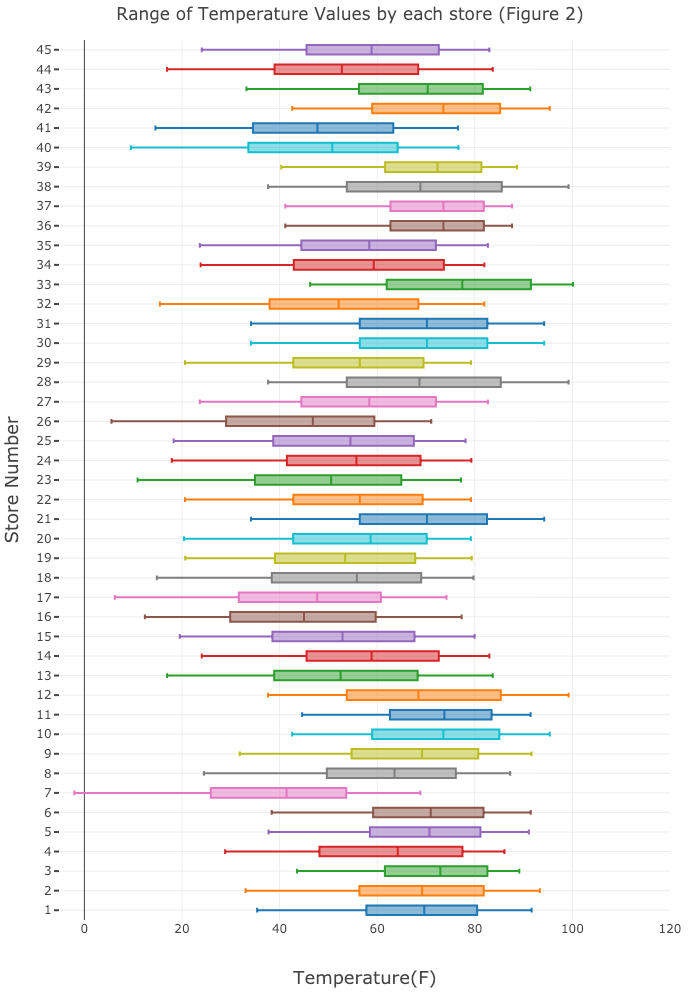


1. Outliers:

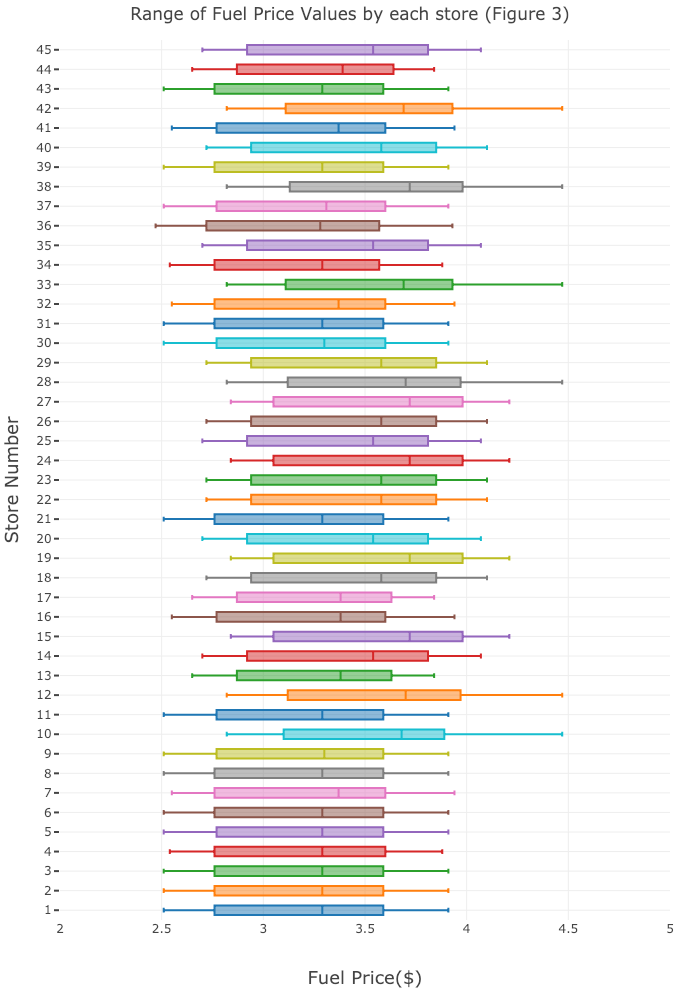
**Weekly\_Sales (Target Variables): Due to geographical or holiday features, the Sales associated with every store should varies. Refer to Figure 1, the Weekly sales mot only varies quite a bit, a considerably large amount data point are out of the range of typical range of sales. As mentioned, the variation might be caused by mixed factors. For this reason, I may just leave those "Outliers" as is for now.**

****

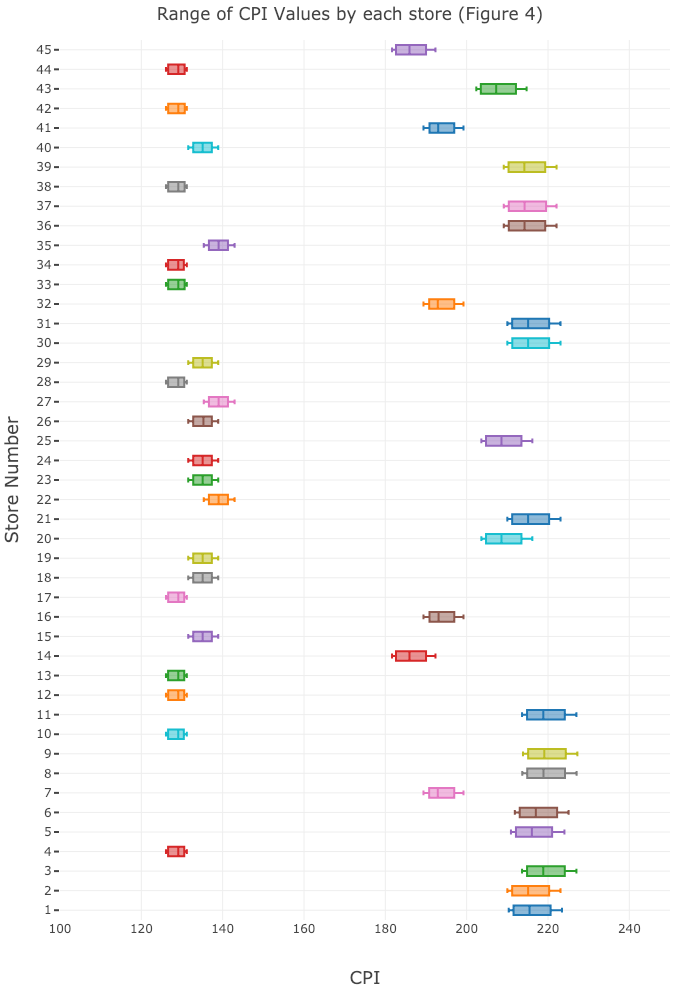
**Temperature: Refer to figure 2, the other variable "Temperature" looks very consistant across the stores. Also, there are NO outliers significantly out of their typical range among the stores**



**Fuel Price: Refer to figure 3, the feature "Fuel\_Price" which is even more consistant than the "Temperature" variable ranges between 2.5 to 4.5 among the stores. So that I would believe that there is NO outliers existing in this feature. This figure makes sense to us because Fuel Price has its market Price across the countries, so that there should be not much variation in each geographical area as expected.**



**Consumer Price Index (CPI): According to U.S. Bureau of Labor Statistics, the meaning of CPI is "a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services". Refer to Figure 4, the CPI value, ranging from 120 to 230, differs quite much among the stores. However, there is NO existing outliers shown in this dataset. This figure relatively indicates us the consuming power across stores. For hypothetical assumption, the area with higher CPI values should expect more consuming power which is supposed to generate more sales amount for the store, or vice versa. Further analysis will be conducted to measure this aspect of feature along with this analysis proceed.**



**Unemployment: The unemployment rate is another index shows the consuming power in the country. Refer to the Figure 5, the data shows that most of the values fall between 6 to 8 without outliers among the stores. Similar to the "CPI" feature, Further analysis will be conducted to measure how this feature influence to the Weekly Sales among the stores.**

