**Walmart Retail Stores Sales Analysis & Forecasting**

**Data Mining & Business Intelligence - Spring 2018**

**Team 8**

**Animesh Choudhary**

**Justin Martinek**

**Mukesh Raman**

**Dier Gao**

**Table of Contents**

1.0 Introduction

1.1 Initial Process

1.2 Research Objectives

2.0 Data Overview

2.1 General Overview

2.2 Data Breakdown

3.0 Data Pre-Process

3.1 Data Merging and Date Fix

3.2 Data Clustering

3.3 Missing Value Treatment

4.0 Compare and Contrast

4.1 Store size (Small vs Big)

4.2 Regressors

4.3 Unemployment (High vs Low)

4.4 Regressors

5.0 Residual Analysis

6.0 Event Handling

7.0 Recommendations

**1.0 Introduction**

**1.1 Initial Process**

As Team 8 was determining which type of dataset would be the best for analysis we actually reviewed several types. This initial research included dataset types such as crime stats, real estate, medical billing, credit history and retail datasets. Wanting to ensure we found a dataset that would allow us to utilize all tools and methods we have learned in OPIM 5671, we decided upon the analysis of a retail dataset. It was felt that the benefits of Time Series Forecasting would best be served with a retail dataset and would also allow for better understanding of the newly learned data analysis methods.

**1.2 Research Objectives**

After doing several initial reviews of the datasets, the team concluded that it would be best to focus on the impacts and relationships to Sales of two key variables:

1 - Store size (large and small)

2 - Unemployment (high and low)

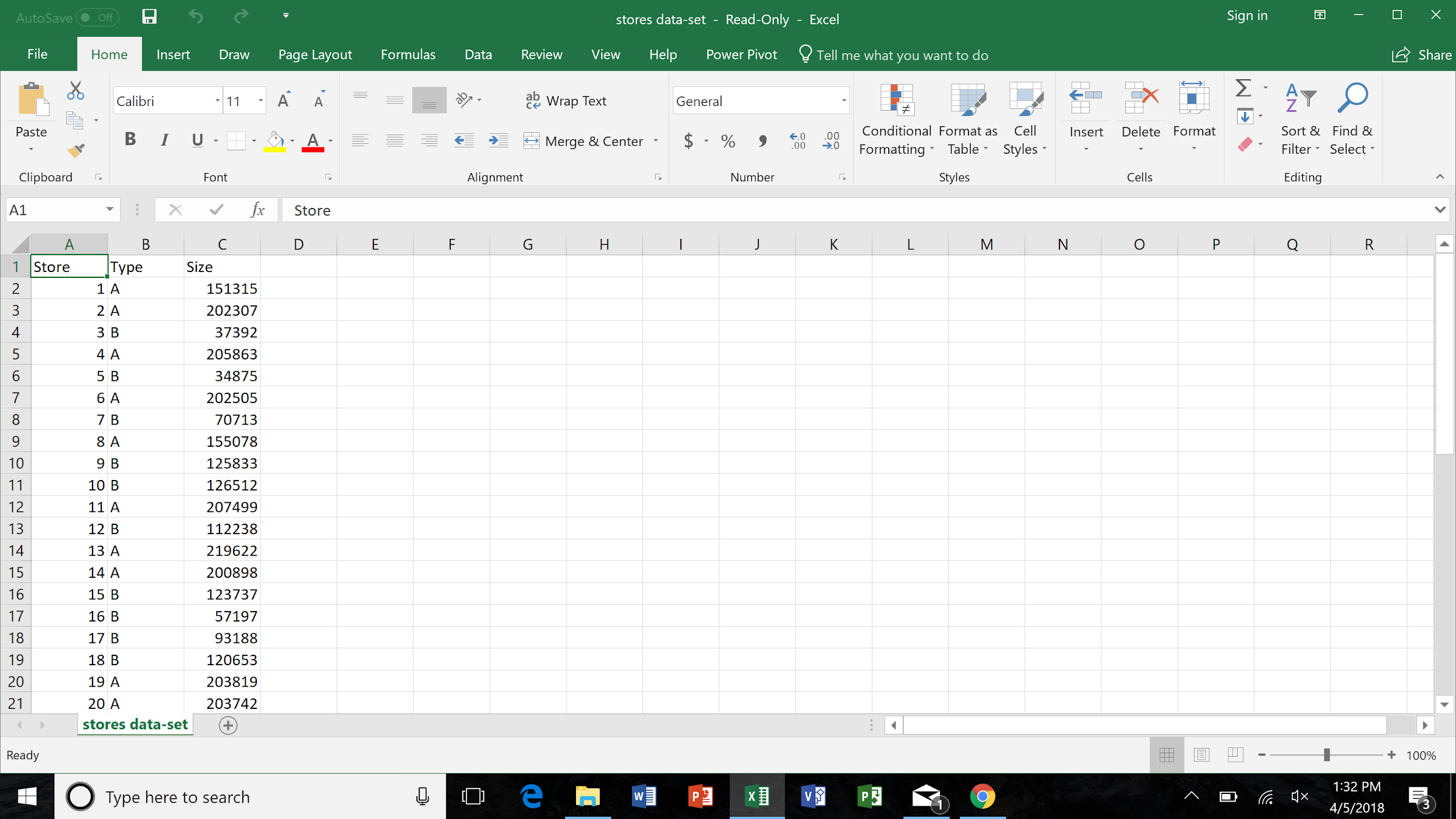
Once it was decided we would focus on these key clusters, it was then decided that the team would analyze the impacts of the remaining sub-variables to these clusters in order to determine the overall impact over time on total sales.

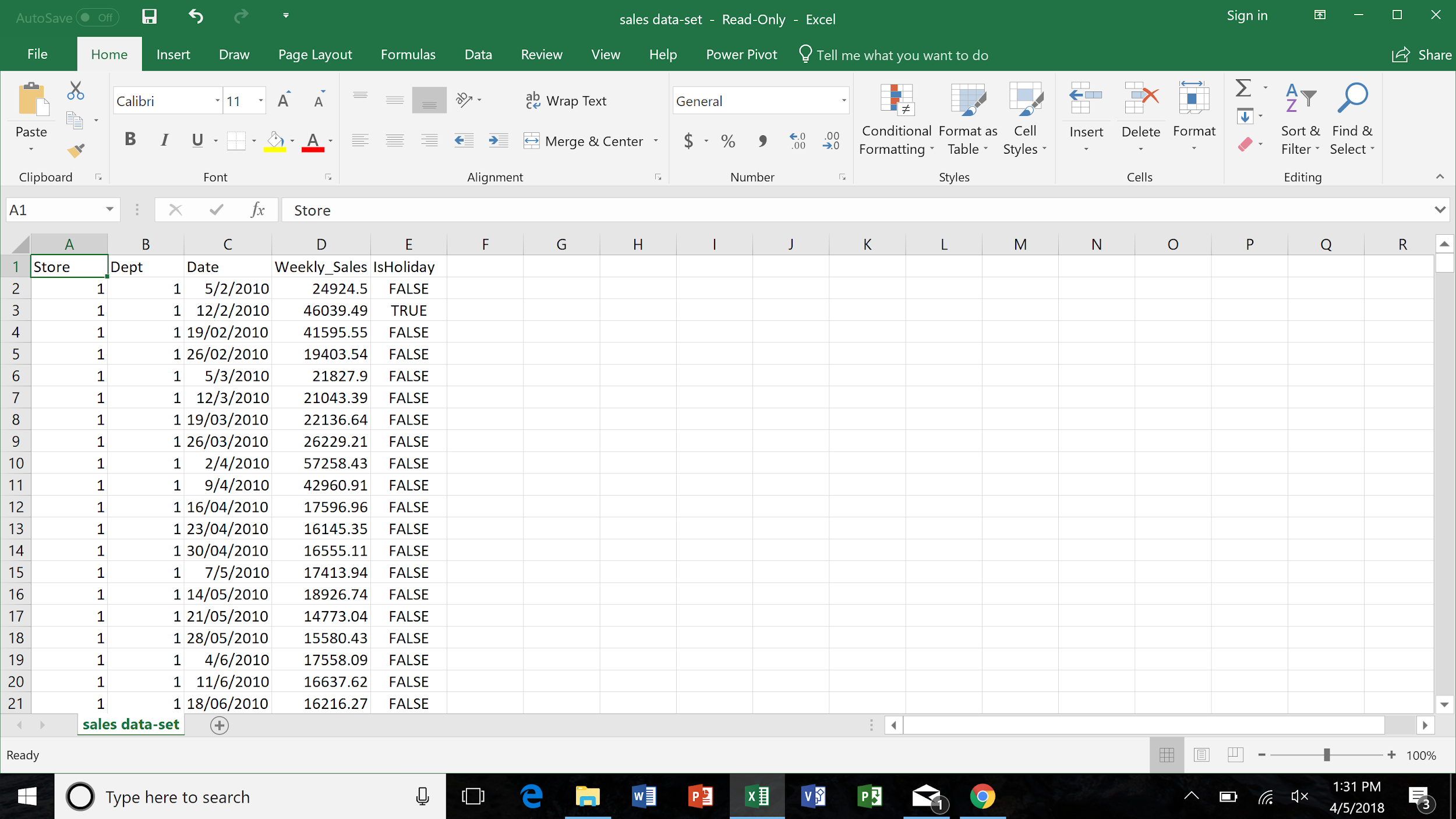
**2.0 Data Overview**

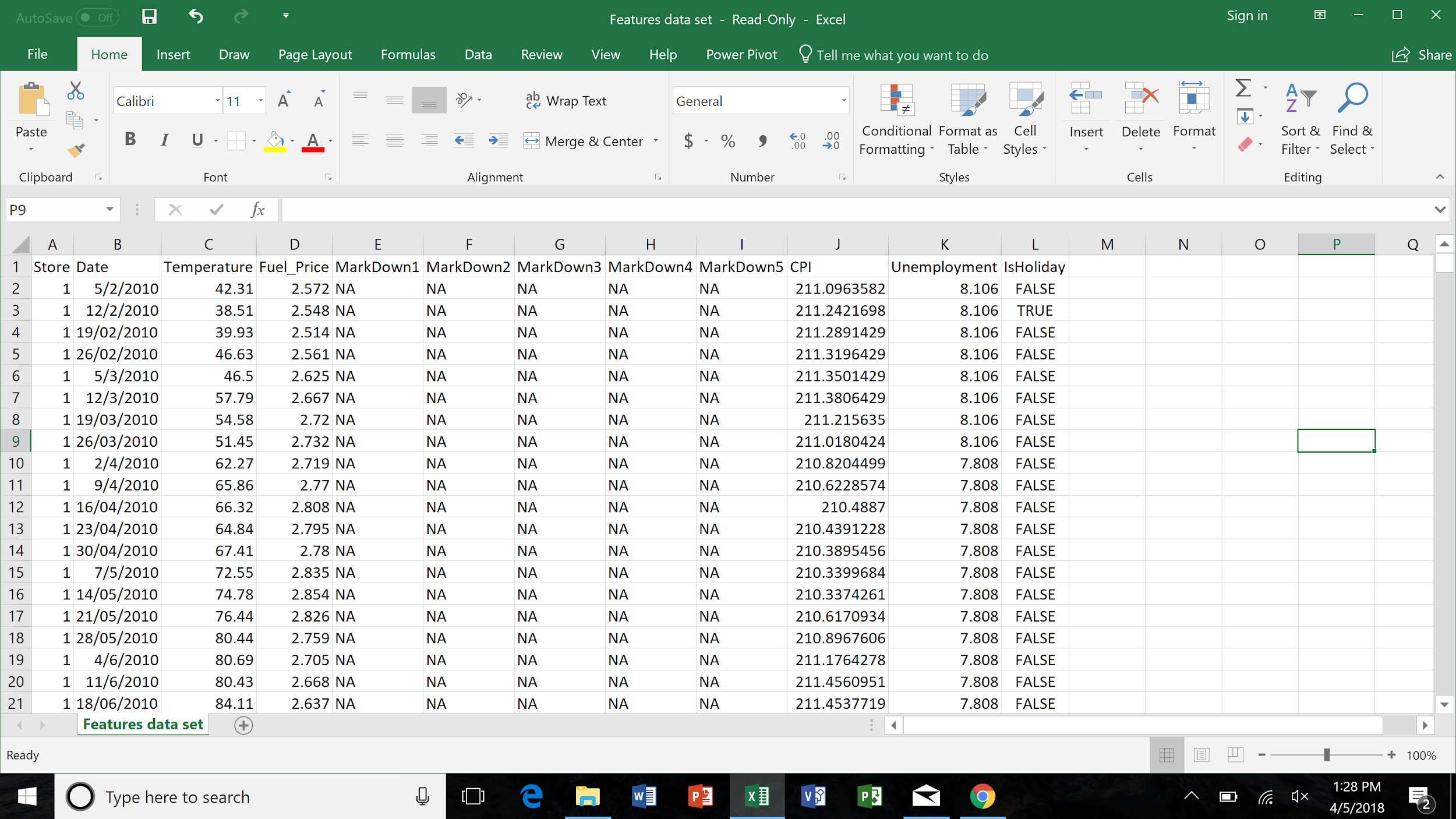
**2.1 General Overview**

The ‘Retail Sales Analytics’ dataset contains detail on 45 retail stores, located across multiple regions. Dataset was made up as follows:

* 3 raw data files - Store (Figure 1), Sales (Figure 2), and Features (Figure 3)
* Data date range - January 2010 thru December 2013

**Figure 1 **

**Figure 2 **

**Figure 3 **

**2.2 Data Breakdown**

As was stated in Section 2.1, the main dataset included 3 raw data files, titled ‘Stores’, ‘Sales’ and ‘Features’. The breakdown of the data within these raw files is as follows:

**Stores Sales Features**

**Rows/Columns**: 45/3 421,571/5 8,191/12

**Variables:**

**Stores Dataset**

|  |  |
| --- | --- |
| Variable | Description |
| Store | Individual store # in dataset (1-45) |
| Type | A or B store (superstore vs standard) |
| Size | Actual square footage |

**Sales Dataset**

|  |  |
| --- | --- |
| Variable | Description |
| Store | Individual store # in dataset (1-45) |
| Dept. | Specific department # within the store |
| Date | Ending week date for sales figures |
| Weekly\_Sales | Weekly sales figures for date. Stated in $’s |
| IsHoliday | Stated as ‘True’ or ‘False’ for if date was holiday |

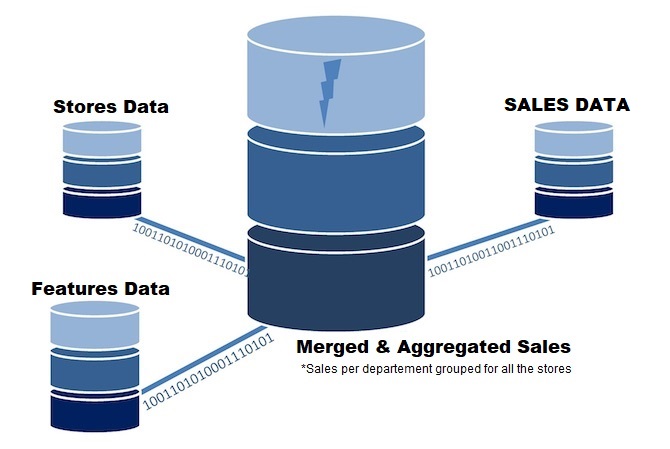
**Features Dataset**

|  |  |
| --- | --- |
| Variable | Description |
| Date | Ending week date for sales figures |
| IsHoliday | Stated as ‘True’ or ‘False’ for if date was holiday |
| Temperature | High Temperature for the date in Fahrenheit |
| Fuel\_Price | Fuel price for the date in $’s |
| CPI | Consumer Price Index stated as a whole # |
| MarkDown1 | 1st markdown or discount given |
| MarkDown2 | 2nd markdown or discount given |
| MarkDown3 | 3rd markdown or discount given |
| MarkDown4 | 4th markdown or discount given |
| MarkDown5 | 5th markdown or discount given |
| Unemployment | Unemployment rate stated as a % |
| Store | Individual store # in dataset (1-45) |

**3.0 Data Pre-Process**

**3.1 Data Merging and Date Fix**

In order to analyze the effect of various external parameter, we needed all the columns from the various files in 1 single file i.e. we needed the size information from Stores data; the temperature, CPI, Fuel Price, Unemployment, various Markdowns etc from the Features data and finally the Weekly Sales from every store through the Sales data. In order to get this, we joined the 3 files on the Store column which was common in all 3 of them.



The output of the merge had the following columns:

**Store, Sales, Date, Temperature, Fuel\_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment, IsHoliday, Type, Size**

This was done using the sqldf library in R. The code for this can be found at: <https://github.com/animvin/Walmart-Forecasting/blob/master/Merge_data_sets.R>

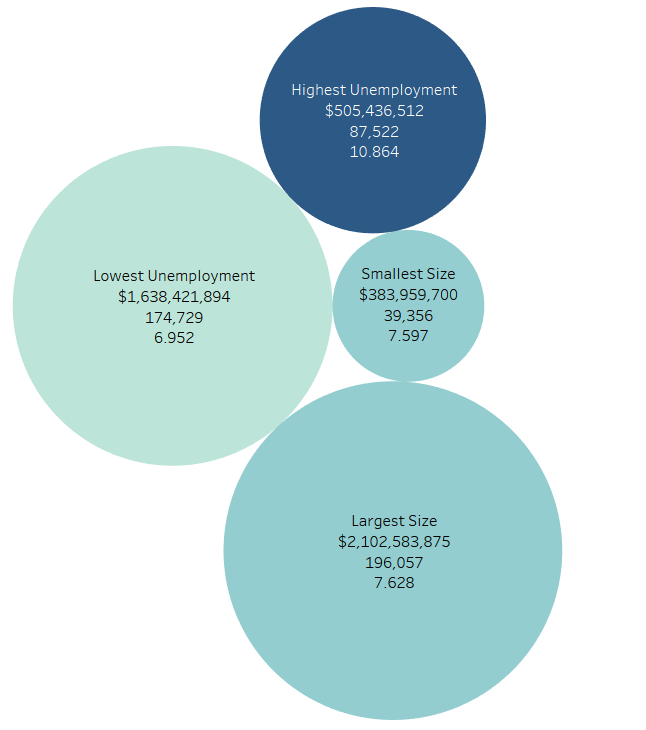
The date formats in the original file were not consistent and were in a format which could not be read by SAS TSFS. So, we changed the weekly dates in the MM-DD-YYYY format. This was done using Jupyter notebook and the code for it can be found at:

<https://github.com/animvin/Walmart-Forecasting/blob/master/DateFix.ipynb>

**3.2 Data Clustering**

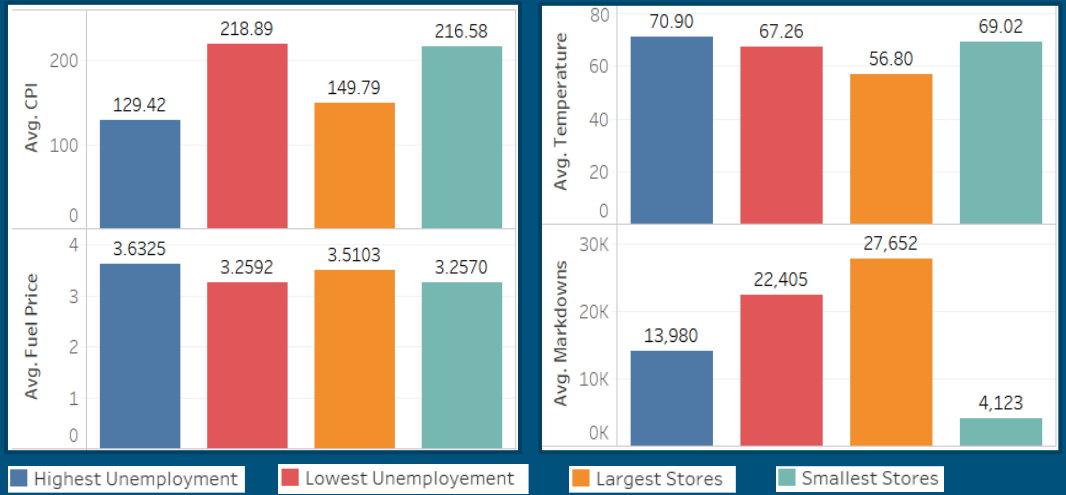
In order to answer the research questions better, we grouped the various stores into different clusters. This clustering was done using Hierarchical Clustering on the following attributes:

**CPI, Temperature, Combined Markdowns, Unemployment, Sales, Size**

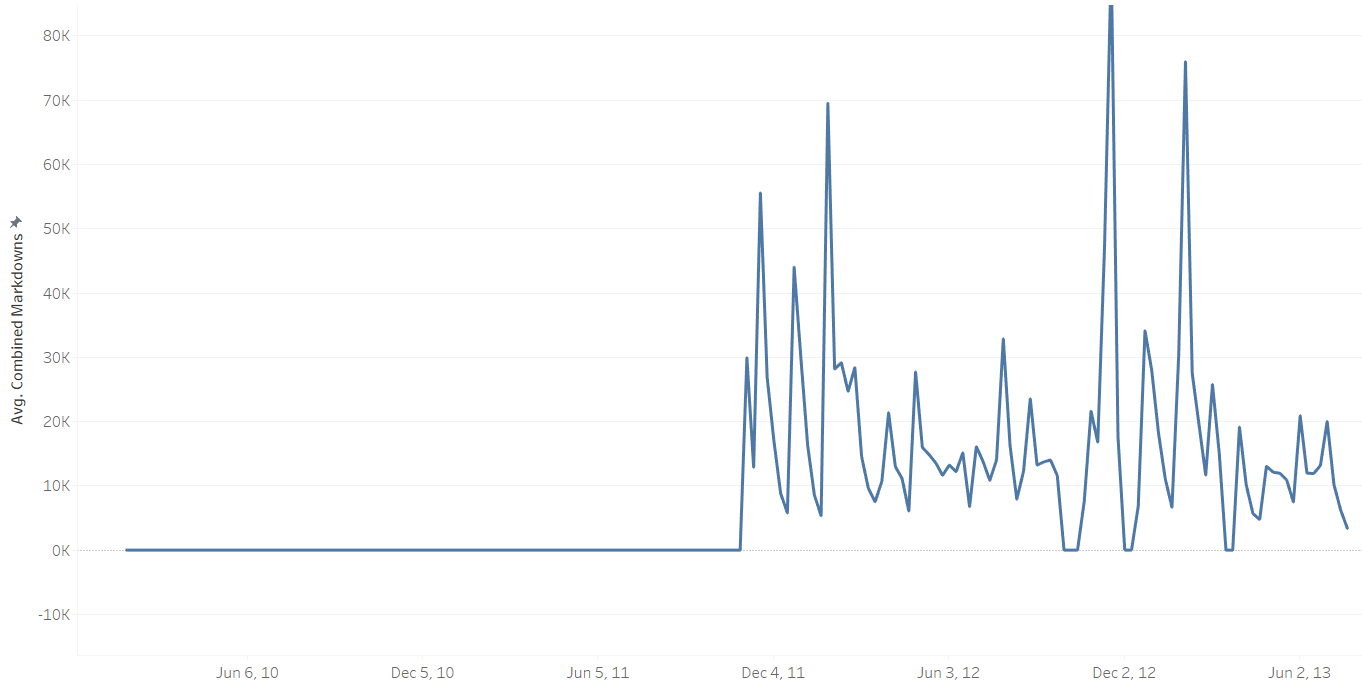


|  |  |
| --- | --- |
| Cluster | #Stores |
| Highest Unemployment | 9 |
| Lowest Unemployment | 5 |
| Smallest Size | 6 |
| Largest Size | 8 |

There were a total of 5 clusters created among which we selected 4. There was 1 cluster which had the highest unemployment rate and 9 stores, one with lowest unemployment rate and 5 stores. Similarly, there were 2 other complementary clusters in which 1 had the largest stores with 8 stores and the other had the smallest ones with 6 stores. The bubble plot above represents the 4 clusters where the size represents the store size and the intensity of color indicates the unemployment rate. These 4 clusters consisted of 28 out of the total 45 stores. The clusters had the following properties:

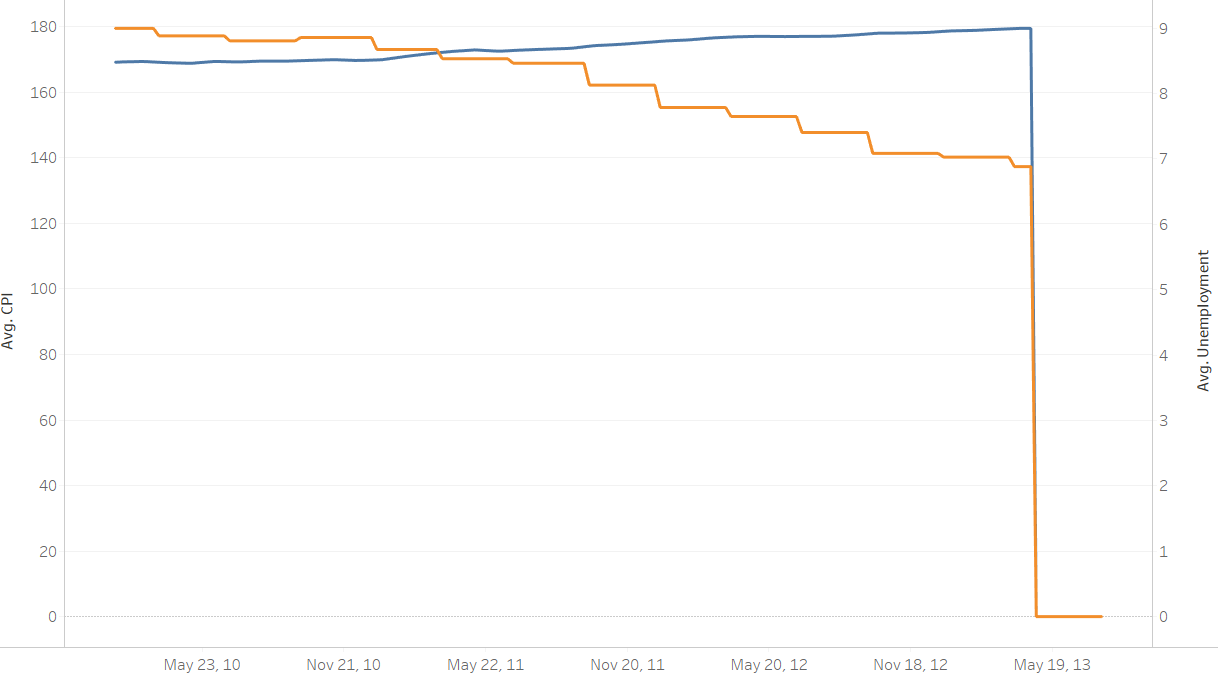
****

**3.3 Missing Value Treatment**



Markdowns:

* Missing for the initial 92 weeks
* Cannot be all assumed to be 0 for all the initial weeks
* Imputed by using the average markdowns of the same week across all the years
* Imputed independently for all clusters and markdowns

****CPI & Unemployment:

* Missing for the last 13 weeks of test set
* Cannot be assumed to be 0
* Imputed by using the average values of the various weeks in the last year
* Imputed independently for all clusters

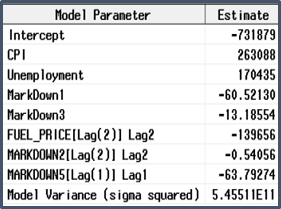
The code for cluster grouping and missing value treatment can be found at:

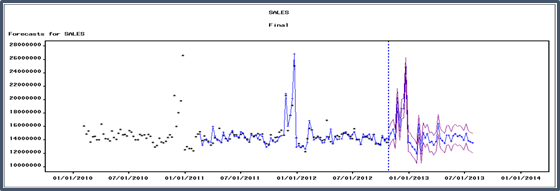
<https://github.com/animvin/Walmart-Forecasting/blob/master/ClusterGroup.ipynb>

**4.0 Compare and Contrast**

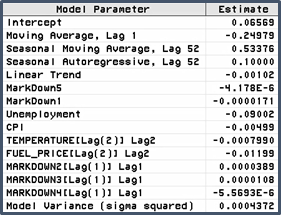
**4.1 Store size (Small vs Big)**

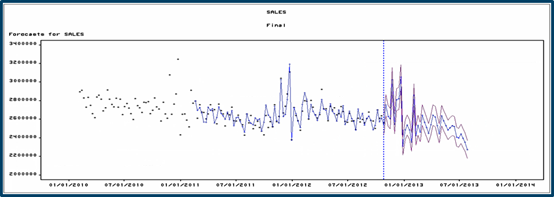
According to ACF and PACF chart, we first developed the best ARIMA model for small stores cluster and large stores cluster.





It turned out that the ARIMA(0,1,0)s is the model for large stores cluster with lowest RMSE.

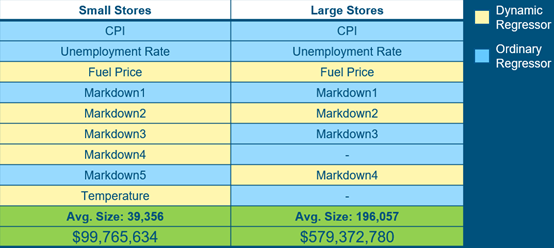




ARIMA(0,0,1)(1,1,1)s is the model for small stores with lowest RMSE. We decided to use these two models as our base model and further improvement them.

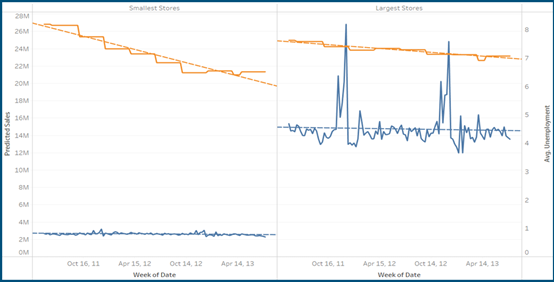
**4.2 Regressors**

Next, we started adding regressors.



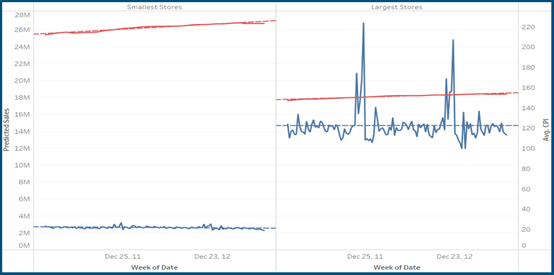
In order to compare and to show contrast, here we list all regressors we identified for two clusters. The selecting criteria is that if adding a certain regressor will decrease the RMSE without increasing BIC by too much. Now, our models are further improved.

Since we identified these factors as regressors affecting sales, we wanted to take a look at them individually in details. First, we took a look at Average Unemployment rate vs predicted Sales.



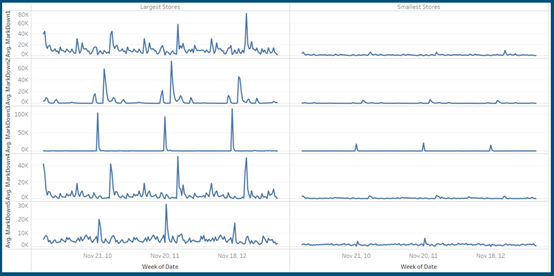
We can tell from the chart that, for small stores cluster, as unemployment rate goes down, there is only a very slight drop for sales. For large stores cluster, as unemployment goes down, we don’t see much impact on sales. The reason is that the coefficient of unemployment rate is in thousands but sales are in millions. Unemployment rate has an effect on the sales but the effect is relatively small.

Next, we visualized relationships between Average CPI and predicted Sales.



As CPI goes up, we don’t see much impact on the predicted Sales. Again, the coefficient of CPI is relatively too small compare with sales. CPI has an effect on sales, but the effect is too small to be considered.

The last common regressor is Markdowns.

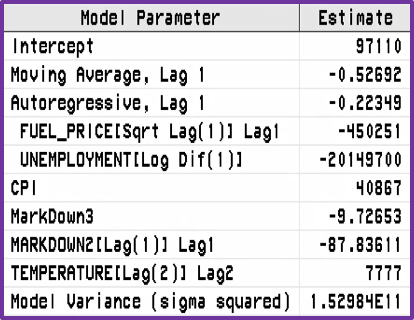


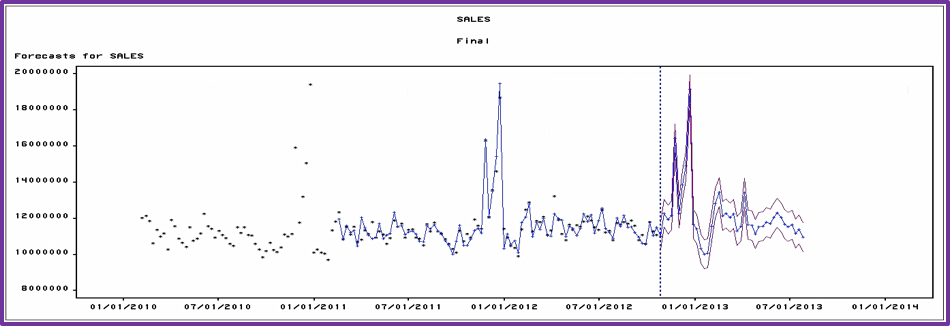
Large stores tend to run large markdowns and small stores tend to run small markdowns. However, if we look at the coefficient, it is very small. Then, we finally came into conclusion that the most important factors affecting sales is store size itself.

**4.3 Unemployment (Low vs High)**

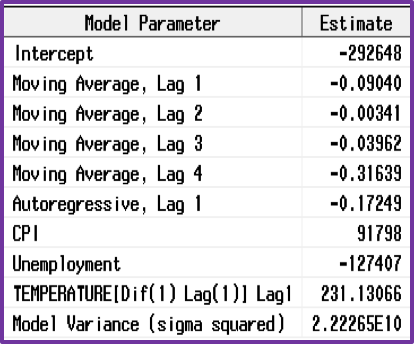
According to ACF and PACF chart, we first developed the best ARIMA model for low unemployment and high unemployment clusters.

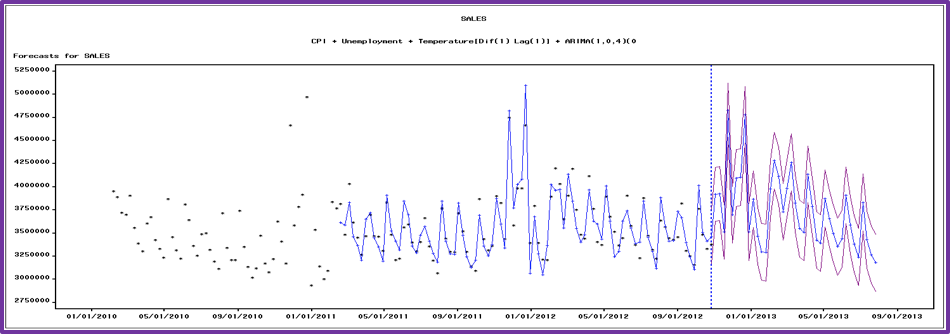
ARIMA(1,0,1)(0,1,0)s is the best model for low unemployment with the lowest RMSE value.

****

****

ARIMA(1,0,4)(0,1,0)s is the model for high unemployment cluster with lowest RMSE

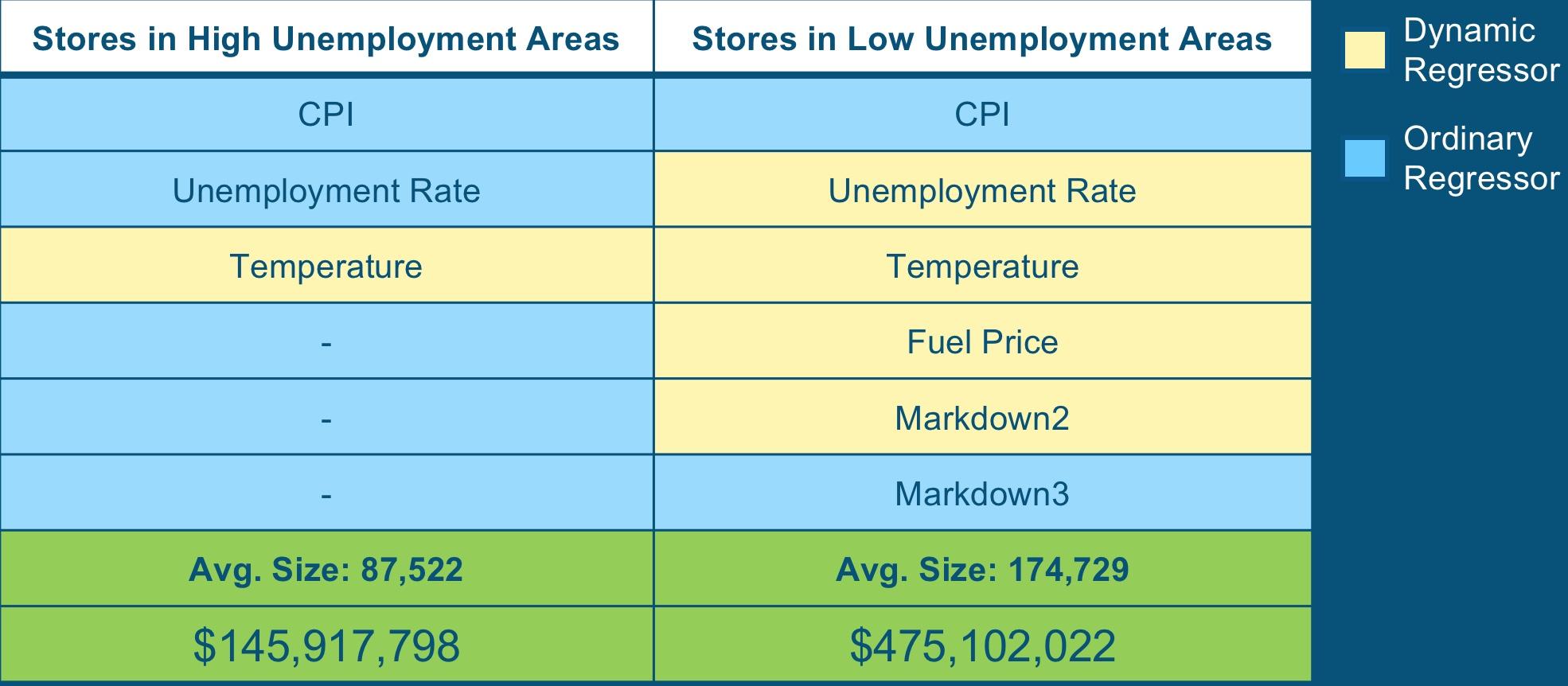
****

****

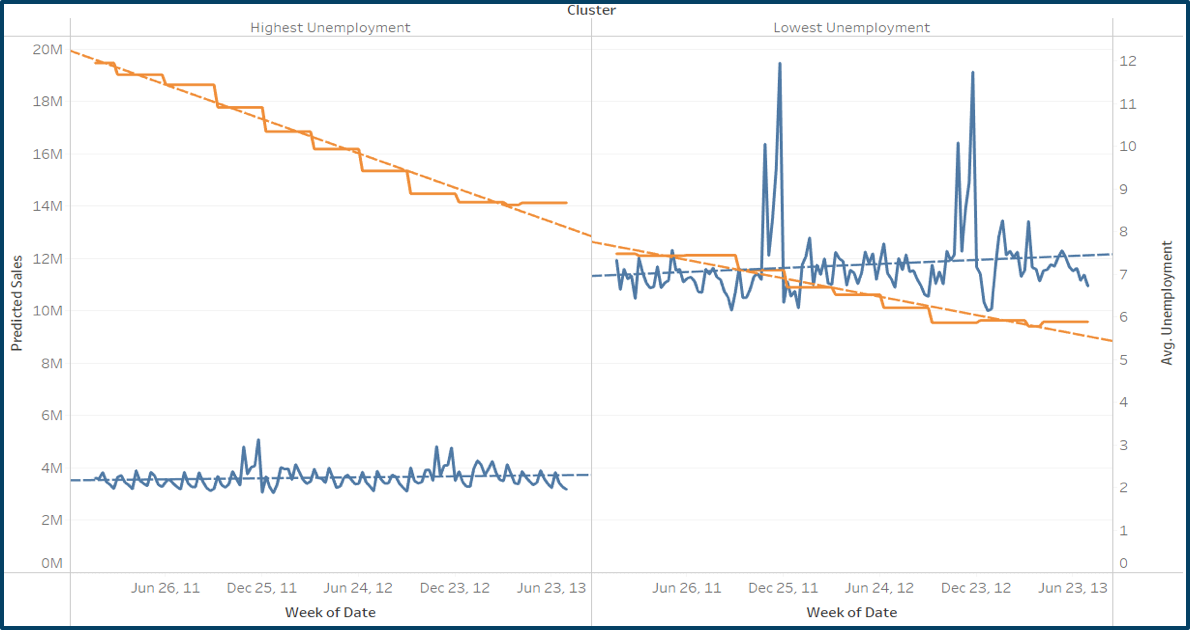
**4.4 Regressors**

Once we identified the best performing ARIMA model for both clusters, few regressors were identified for both clusters, both standard and dynamic clusters, based on trial-and-error. The list of regressors are not common between the two clusters. Some regressors suited well for low unemployment cluster as standard, but same regressor worked well as dynamic regressor for high unemployment cluster and vice-versa.

The criteria for identifying a regressor is whether it is reducing the RMSE value without significant increase in BIC value. The models derived using the regressors performed much better than the regular ARIMA models without regressors.

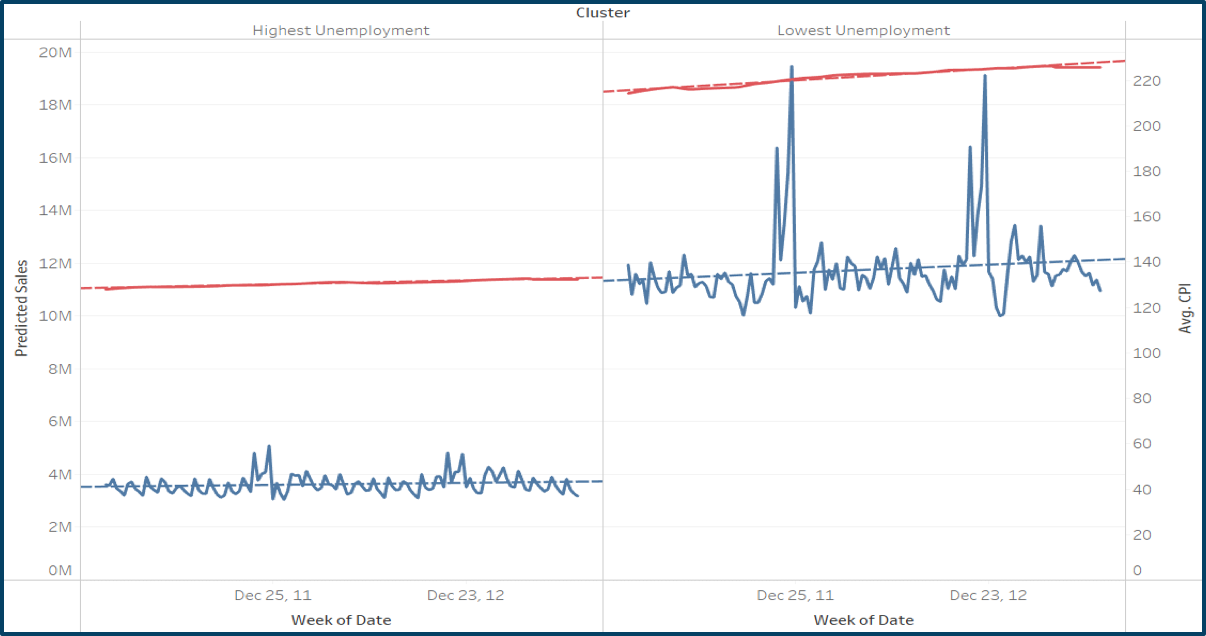


Since we identified these factors as regressors affecting sales, we wanted to take a look at them individually in details. First, we took a look at Average Unemployment rate vs predicted Sales between High and Low unemployment clusters.



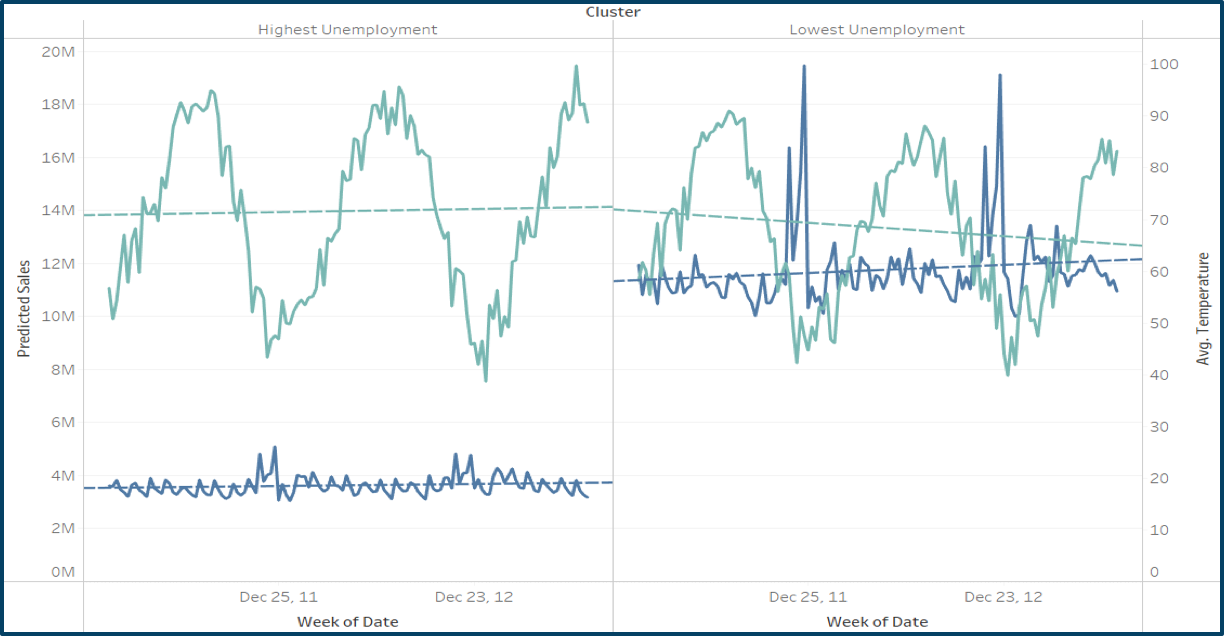
For high unemployment cluster, as unemployment rate goes down, there is no significant change in sales. For low unemployment cluster, as unemployment goes down, there is a slight upward trend in sales. The reason is that the coefficient of unemployment rate is in thousands but sales are in millions. Unemployment rate has an effect on the sales but the effect is relatively small.

Next we visualized relationships between Average CPI and predicted Sales.



As CPI goes up, we don’t see much impact on the predicted sales in high unemployment cluster. In the low unemployment cluster, as the CPI increases, sales increases but not significantly to make a big impact. Again, the coefficient of CPI is relatively too small compare with sales. CPI has an effect on sales, but the effect is too small to be considered.

The last common regressor is Markdowns.

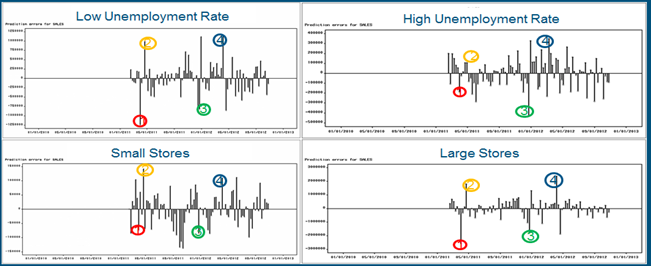


Although there is some impact of markdowns seen in the low unemployment cluster, it’s insignificant for high unemployment cluster. Although temperature plays a role in sales, it’s not very significant.

Comparing the effect on all the common factors/regressors such as CPI, unemployment and temperature on all clusters and comparing them between two groups, low/high unemployment and big/small store size clusters, we concluded that the most important factor affecting sales is store size.

**5.0 Residual Analysis**

Upon final generation of models for the 4 specific clusters, the next step was to analyze each cluster and find similarities within the residuals. As the analysis progressed, specific dates started to stand out that showed that some type of event was occurring. The team identified 4 specific events within the models (figure #), with the Good Friday holiday being a common denominator for each.

****

**6.0 Handling Events**

April 1, 2011 actually falls on Good Friday of previous year. That’s why we see a huge prediction error here. April 22, 2011 is actually the Good Friday this year. Since Good Friday falls on different week each year, we might see prediction errors on these days. Same reason applies for April 06, 2012. Dec. 23, 2011 is the Christmas eve, but it is a Friday. It may come in as different weeks, so it varies.

To tackle the errors caused due to certain occasions like Good Friday which fall in a different week each year we can use the following 2 approaches:

1. Use monthly sales predictions instead of weekly ones. In this way we would be able to forecast the sales which even when certain occasions fall on a different week every year as long as they fall within the same month.
2. Identify 2 events, one to bring down the bloated sales on the week which falls on the same number in which previous year’s occasion happened and then another even to boost the sales on the correct week on which the occasion would happen in this year. Once these events are identified on the historical data, they need to used to better forecast the actual sales on those weeks. The advantage of this method is that it would even work in cases where the occasion happens in a different month altogether. In our case due to limitations in SAS we were not able to add any interactions in the model.

**7.0 Recommendations**

* **Store Size**
  + Larger Stores lead to Higher Sales. The effect of all the external factors is massively overshadowed by the size of the store. So unless the management is restrained due to other factors such as legal compliance or excessive operating costs, they should invest in opening larger stores instead of smaller ones
  + Even in areas with High Unemployment rates, Larger stores can still perform better
* **Discounts**
  + Markdowns do not really have much impact in areas with High Unemployment rates. Spend more on Discounts in all other stores because they do lead to an increase in the sales. During winters, people tend to prefer Walmart over other stores. So, investing in more promotions during winters can lead to potentially higher sales
* **Inventory Management**
  + Inventory Management should be done over forecasts aggregated at a monthly level and not weekly as that will be more robust to distribution of sales among the weeks
* **External Factors**
  + Monitor external factors like CPI, Unemployment Rate, Fuel Price and Temperature as they do tend to have an effect on the sales in most of the stores