**Logo

Description automatically generated**

**ALY 6040 – Data Mining Applications**

**Instructor: Justin Grosz**

**Northeastern University**

Parveen Abbu

Neha Wakade

Meghana Musunoori

Surya Akhil Tadiparthi

**Introduction**

We have chosen one of the leading retail stores in the US, Walmart retail sales data as our group project. Using this data set we would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. We will try to predict demand accurately using factors like Temperature, Fuel price CPI, Unemployment Index, etc. The data consist of 6435 rows and 8 columns.

The following fields are included in the given dataset:

1. Store — the store number
2. Date — the week of sales
3. Weekly Sales — sales for the given store in USD ($)
4. Holiday Flag — whether the week is a special holiday week 1 — Holiday week 0 — Non-holiday week
5. Temperature — Temperature on the day of sale
6. Fuel Price — Cost of fuel in the region
7. CPI — Prevailing consumer price index
8. Unemployment — Prevailing unemployment rate

**Objective:**

We plan to do EDA analysis on this dataset and come up with three interesting facts:

1. Find out the store with maximum fluctuations in sales.
2. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together
3. Plot monthly and quarterly view of sales in millions
4. Plot correlation matrix of each independent variable

**Analysis 1: - Store with maximum fluctuations in sales**

In order to find the store with maximum fluctuations in sales, we should look at the coefficient of the variation in weekly sales figures. The coefficient of variation (CoV) is a measure of relative variability. It is the ratio of the standard deviation to the mean (average).  Pandas in Python gives us the ability to group all the stores with their store Ids and aggregate their weekly sales figures to get standard deviation(std) and mean.

Table

Description automatically generated

Figure : Store with maximum variation in sales

We found that the store 14 has the highest CoV 15.71 among all stores. This tells us that the sales at this store fluctuates a lot!

**Analysis 2: - Holiday-week with highest sales**

We are focusing only on four major holiday weeks namely **Super Bowl, Labour Day, Thanksgiving** and **Christmas.** We will compare the mean sales on these holidays with the mean sales on non-holiday weekends.

Holiday Events

1. Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12
2. Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12
3. Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12
4. Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12

We filtered rows from the dataset and calculated mean of weekly sales for these holidays.

**A picture containing text

Description automatically generated**

Figure : Comparison of holiday week sales with non-holiday week sales

We observed that Thanksgiving week has the highest mean sales $ ~1.5M and Christmas week sales has the lowest mean sales $~1M. Christmas week mean sale is even lower than the non-holiday week mean sale.

**Analysis 3: - Visualize monthly and quarterly view of sales**

To visualize monthly and quarterly sales, we will use weekly sales figures and aggregate them to calculate monthly and quarterly sales figures.

## Chart, line chart, histogram Description automatically generated

Figure :Monthly sales at all stores

Chart, line chart

Description automatically generated

Figure : Quarterly sales at all stores

From the above figures, we observed that highest sum of sales is recorded in January month of 2011 and 2012.   
  
Q2–2011(Apr 2011 to Jun 2011) was the worst quarter for sales and Q3 – 2012(Jul 2012 to Sep 2012) the best quarter for the sales.

**Reference:**

[**https://www.kaggle.com/vik2012kvs/walmart-dataretail-analysis**](https://www.kaggle.com/vik2012kvs/walmart-dataretail-analysis)

[**https://pandas.pydata.org/**](https://pandas.pydata.org/)

[**https://medium.com/analytics-vidhya/a-data-science-project-for-beginners-exploratory-data-analysis-eda-d334f58f94ee**](https://medium.com/analytics-vidhya/a-data-science-project-for-beginners-exploratory-data-analysis-eda-d334f58f94ee)