# Lab 1

Please submit your answers in this word document, including detailed discussion for each question with supportive screenshots of your analysis in Python.

## Exploratory Data Analysis

One of the leading retail stores in the US, Walmart, 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. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. Historical sales data for 45 Walmart stores located in different regions are available.

### Dataset Description

Dataset: Walmart\_Store\_sales.xlsx

This is the historical data that covers sales from 2010–02–05 to 2012–11–01, in the file Walmart\_Store\_sales. Within this file you will find the following fields:

* Store — the store number
* Date — the week of sales
* Weekly\_Sales — sales for the given store
* Holiday\_Flag — whether the week is a special holiday week 1 — Holiday week 0 — Non-holiday week
* Temperature — Temperature on the day of sale
* Fuel\_Price — Cost of fuel in the region
* CPI — Prevailing consumer price index
* Unemployment — The prevailing unemployment rate
* Week\_of\_yr: the week of the year
* Week: the week number in the entire time period

**Holiday Events**

* Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
* Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
* Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
* Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

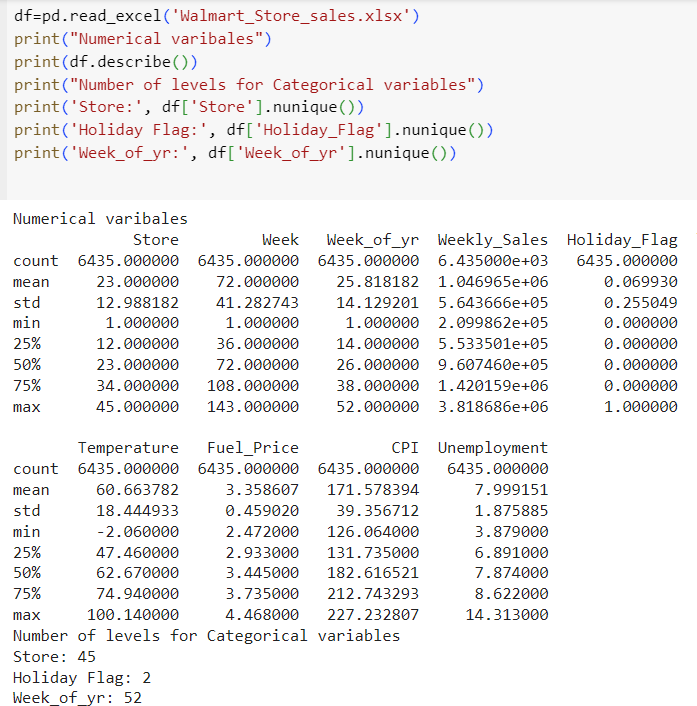
### Analysis Tasks

A new local regional manager for Walmart has contacted you to help with some data analysis of a set of 45 stores from different regions to gain some insight into the available data before jumping in to make major changes. You have been asked to answer the following based on your analysis performed in Python:

1. Show the appropriate software output of the summary statistics including missingness, min, max, mean, and standard deviation for all the numerical variables and number of levels for all the categorical variables.

**Answer:** The below screenshot displays the min, max, mean and standard deviation for the numeric variables.

Missing Values: there are no missing values in the dataset. We can confirm this by the count, in the first row which is same for all the variables/fields.

****

1. How does the weekly sales compare between these stores? Specifically, which store has the highest weekly sale value overall, which has the highest median weekly sales, and which has the highest average weekly sales? Provide a box graph and a brief discussion to explain the graph. Remember to include your first and last name in the footnote of your graph.

**Answer:**

**A screenshot of a computer screen

Description automatically generated**

**Explanation:** The box plot shows the weekly sales for each store. There are some outliers for store 4,10,14 and 20 which show much higher weekly sales. Stores 3,5, 33, 36, 38, 44 have the lowest of weekly sales. The median weekly sales vary significantly across the stores.

1. How does the weekly sales compare for the different weeks of the year? Specifically, which week has the highest weekly sale value overall, the highest median weekly sales, and the highest average weekly sales. Provide a box graph and a brief discussion to explain the graph. Remember to include your first and last name in the footnote of your graph.

**Answer:** The below screenshot displays the max, mean and median weekly sales for the different weeks of the year.

**A screenshot of a computer

Description automatically generated**

**Explanation:** The box plot clearly supports the findings. It Shows that week 51 (Christmas) had the highest sales followed by week 47 (Thanksgiving). The weekly sales do not tend to fluctuate much expect week 2,3,4 show a comparative drop in sales.

1. Make some comments on distribution of the weekly sales? What about its center and variability? Is it symmetric, right-, or left- skewed? Provide a histogram and descriptive statistics to support your claim. Remember to include your first and last name in the footnote of your graph.

**Answer:** The weekly sales data is not uniformly distributed. The median of the data is 762556.3310. It has a standard deviation of 5.643666e+05 and we can calculate the Variance as 751.2433 which is on the higher side.

We can clearly identify from the histogram that the data is skewed to the right which means there are lesser records for higher weekly sales value. There are more records for weekly sales on the left side of the histogram which means that the weekly sales were mostly lower than 1.5 millions. 500,000 weekly sales is the most frequently occurring sale. Sales higher than 3 million are very low in frequency or happen very few times.

**A screenshot of a computer

Description automatically generated**

1. Look at the pairwise scatter plot provided in the lab instruction. Generate the same pairwise scatter plot of all selected variables with different colors indicating the holidays. Then find two variables that show clear correlation with each other. There are more than one answers.

**Answer:** There is a positive linear relationship between “Fuel\_price” and “Temperature”.

Another positive linear relatonship is between “CPI” and “Temperature”.

There is a negetive linear relationship between “CPI” and “Unemployment”. Another negetive linear relationship is between “Fuel Price” and “CPI”. Hence, we can state that these variables show correlation with each other.

A collage of pink and blue dots

Description automatically generated A graph and a diagram

Description automatically generated with medium confidence