# Lab 1

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

(Note: In all lab submissions including Lab 1, all graphs generated by SAS EM should have your first and last name in the footer whenever possible, otherwise your screen shot should include your project name ‘Homework-your first and last name’ to show the lab is your own original work.)

## 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 SAS® Enterprise MinerTM:

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.
2. I have used Walmart dataset and using DMDB tool I have resulted summary statistics for each column in the dataset.
3. When I ran the diagram, we can see various statistical parameters in the result window.
4. There are no missing values in any of the columns.
5. The total number of entries in each column are 6435.
6. We can see mean measures of each column. CPI: 171.58, Fuel\_Price: 3.36, Temperature: 18.44 etc.

A screenshot of a computer

Description automatically generated

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.
2. We can see below three screenshots for each question mentioned.
3. We can see that highest weekly sales recorded in store 20. While second highest weekly sales recorded in store 4.

A graph showing a store

Description automatically generated with medium confidence

1. In the below screenshot we can see the average weekly sales value is 2107676 at store 20.

A screenshot of a graph

Description automatically generated

1. In the below screenshot we can see the median weekly sales value is 2073951 at store 4.

A screenshot of a graph

Description automatically generated

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.
2. Using Graph Explore node, I have plotted bar graph between Week and Week sales with statistical parameters.
3. I found that week 47 have highest sales with value of 78226817

A screenshot of a graph

Description automatically generated

1. To find out highest average weekly sales in the week 47
2. Interestingly, week 47 is the thanksgiving week hence the sales are higher in terms of mean and highest sales.

A screenshot of a computer

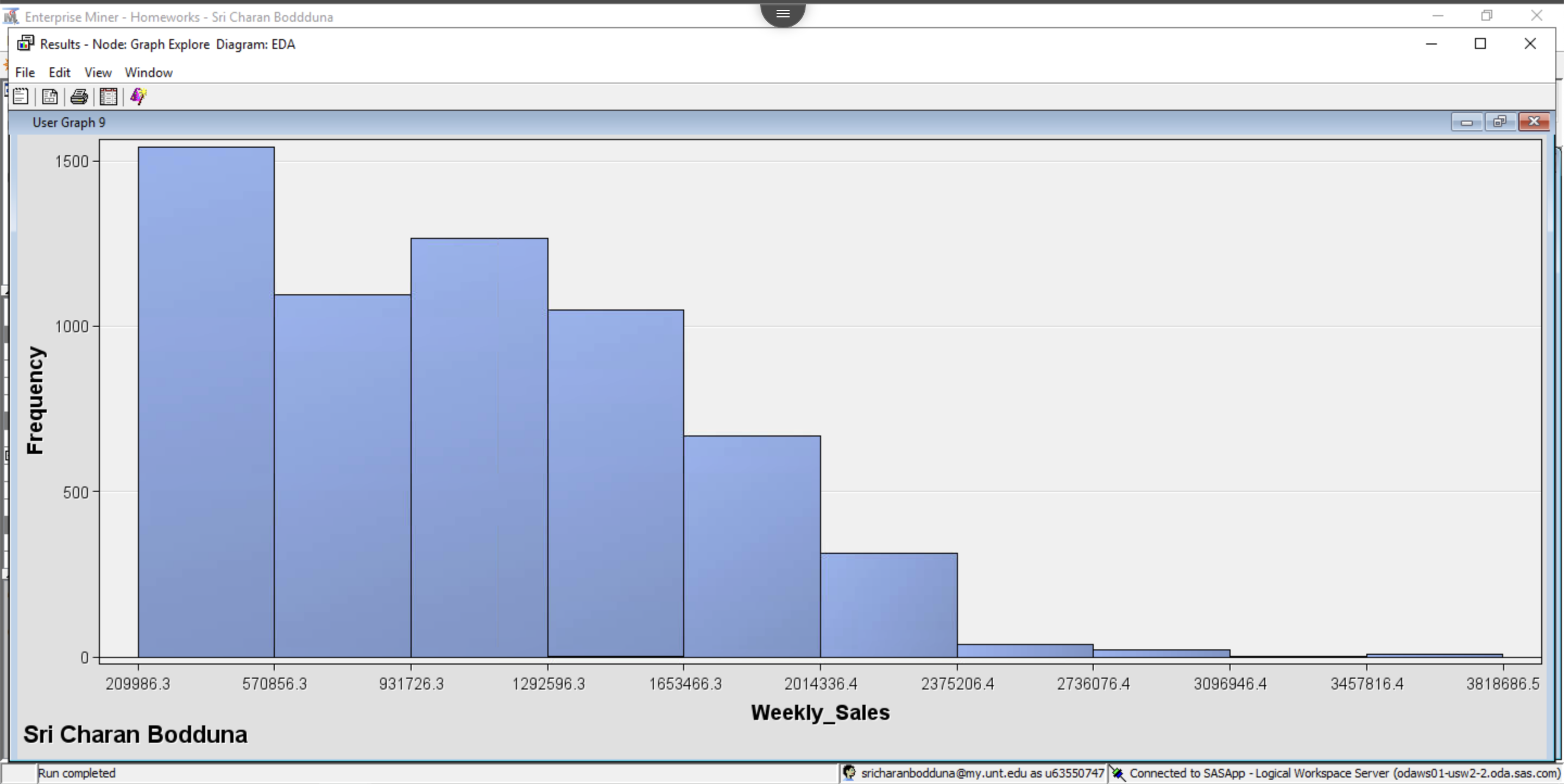
Description automatically generated

1. We can also see median sales of the Walmart are highest in the week 47 with value of 1773743.

A screenshot of a computer

Description automatically generated

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.
2. To understand the data distribution of weekly sales, I have plotted histogram chart using graph explore node.
3. When I look at the chart, I can observe that chart is right skewed.
4. We can see the weekly sales with lower values are highest in frequency. Which means products of lower value have highest sales in the Walmart store.
5. As the values of the sales increases, the frequency or purchasing power is reduced.



1. Look at the ‘The Graph Explore Node’ section of 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.
2. I have plotted pairwise scatter plot using Graph Explore node.
3. We can see various relationships between the features in the dataset.
4. There is a low positive relationship between fuel\_price and week.
5. There is a half highly positive relationship and half negative relationship between temperature and week of year. That means half winter season and half summer season.
6. There is a no relationship between week and temperature, week and weekly sales.
7. Highly positive relationship between week and week of the year.

