**Programing for Analytics**  *Nandan Chetanbhai Patel*

**Walmart Sales Prediction Project**

Contents

Milestone 1: 2

Milestone 2: 3

Milestone 3: 7

Milestone 4: 12

1. **Milestone 1**
2. **Personal objective:**

The main objective for this project is to analyze the sales data of 45 Walmart stores from 2010-02-05 to 2012-11-01 and identify the factors as well as patterns and trends that have greatest impact on sales to help improve business decisions and increase revenue.

Analyzing Walmart dataset using SAS would help us to gain hands-on experience in data analysis, learn new techniques and tools, and provide valuable insights to Walmart that can inform decision-making and drive growth. The aim to identify opportunities for Walmart to enhance its customer experience and stay competitive in the retail industry.

1. **Intended outcomes:**

* Using SAS, to clean up Walmart's sales data by fixing data formats, dealing with missing values, identifying, and addressing outliers, and correcting any possible errors, to ensure the data is accurate and reliable for further analysis.
* To create a new variable in Walmart's sales data that will make it easier for us to do more in-depth analysis in the future, by adding extra information or calculations that can help us find interesting patterns or insights.
* To analyze and compare sales performance across different Walmart stores over a specified period, to identify patterns and trends, and gain insights into how different stores are performing in terms of sales.
* Sales Performance Analysis: Using SAS, will conduct analysis of Walmart's sales data to identify trends, patterns, and seasonality that can reveal why sales go up or down, including factors like Holiday, Fuel price, Unemployment, CPI, and time of year, to gain insights into potential areas for improvement. This could help Walmart plan their inventory and staffing based on different parameters.
* Data Visualization: Visualize the data using charts and graphs to communicate the findings and insights to the intended audience, such as Walmart management, investors, and stakeholders.

Overall, the intended outcomes of this project are to provide valuable insights and recommendations to Walmart's management team, leveraging SAS data analysis techniques to inform their business strategies, optimize sales performance, and enhance customer satisfaction and loyalty.

1. **Description of the needs of the intended audience:**

The Walmart dataset analysis using SAS is intended for Walmart management, marketing, and sales teams who are responsible for making data-driven decisions about sales and marketing strategies. The analysis will help them understand how different factors like holidays, weather, fuel prices, and unemployment rate can affect the sales of Walmart.

The intended audience may not be experts in statistics, so it is important to explain things in a way that is easy to understand. The analysis will show tables and charts to help explain the important information of the analysis that will help them make informed decisions about sales and marketing strategies, such as planning promotions, adjusting pricing, and staffing.

Overall, our goal is to provide actionable insights that will help Walmart improve its sales and marketing strategies.

1. **Foreseeable challenges:**

Some of the foreseeable challenges that were encountered while analyzing the Walmart Dataset are:

1. **Data quality:** The Walmart dataset may have missing or incorrect data, which can affect the accuracy of the analysis. Cleaning and transforming the data will be one of the most time consuming and necessary step required to ensure the data quality is good enough for analysis.
2. **Data exploration**: It can be challenging to determine which variables are relevant and should be used for analysis. The need to identify the most important relationships and patterns between various variables to extract valuable information.
3. **Choosing appropriate Statistical Analysis techniques**: SAS offers a variety of statistical techniques; hence we need to choose the appropriate techniques and visualizations based on the variables in the dataset to find out valuable and actionable insights.
4. **Interpretation of results**: Interpreting the results can be challenging, and we need to make sure that we understand the findings and communicate the insights effectively.
5. **MILESTONE 2:**
   * + 1. **Description of the dataset**

For this project, I will be using a data set called Walmart. This data set is in csv. format. It consists of only one worksheet. The dataset consists of **8** variables:

* Store: a character variable indicating the store number.
* Date: the date of the sales.
* Weekly\_Sales: a numerical variable indicating the weekly sales amount in dollars.
* Holiday\_Flag: a binary variable indicating whether the week contained a holiday (1) or not (0).
* Temperature: a numerical variable indicating the average temperature in the region during the week.
* Fuel\_Price: a numerical variable indicating the fuel price in the region during the week.
* CPI: a numerical variable representing the consumer price index during the week.
* Unemployment: a numerical variable indicating the unemployment rate during the week.

It consists of a total of **6,436** observations.

* + - 1. **Importing of DATA**

The plan is to use SAS to perform auditing and analysis on our dataset. Our first step is to import the data file into SAS so to view and gain a better understanding of the dataset. To accomplish this, we will utilize the PROC IMPORT procedure.

Graphical user interface, text, application

Description automatically generated

As shown in the picture below, we are importing just one dataset and displaying the first

few lines of the dataset.

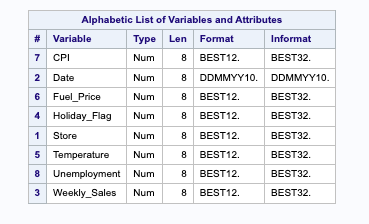
Table

Description automatically generated

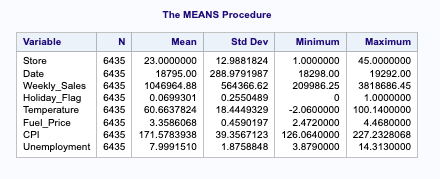
* + - 1. **Examine all variables of the dataset:**

The purpose of this step is to examine if there are any errors in the values of all variables that require steps of cleaning the dataset during data pre-processing phrase. We use proc contents and proc means to conduct the examination. The below are the codes in SAS with their results:









* + - 1. **Data Cleaning**

Upon reviewing the loaded data, and realized the importance of cleaning it up to ensure consistency and enhance its suitability for analysis. Cleaning the data involves eliminating unnecessary variables and adjusting the formatting of certain variables to optimize their compatibility with the models we intend to use in the analysis.

We are using all the variables in the dataset. We have renamed the Temperature variable to have a clear idea about which unit is used for temperature. Here, we have also changed the formatting for variables like Date, Weekly sales, Fuel price, and Unemployment to have a better understanding of the value.

Graphical user interface, text, application, email

Description automatically generated



The picture below displays the cleaned data set.

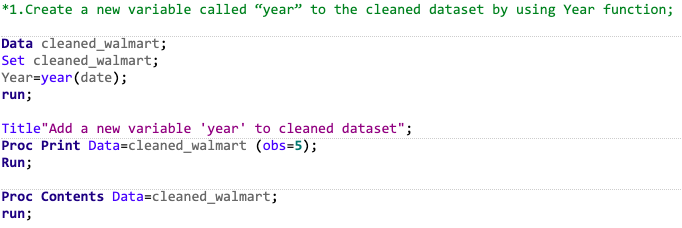
Table

Description automatically generated

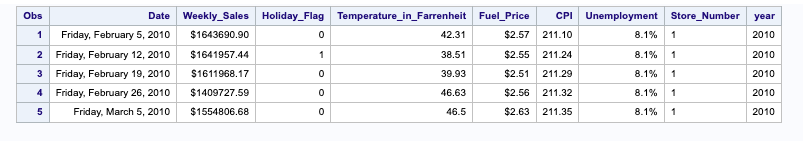
1. **MILESTONE 3:**
   * + 1. **Create a new variable called “year” to the cleaned dataset, which will allow us to conduct the exploratory analysis by years.**

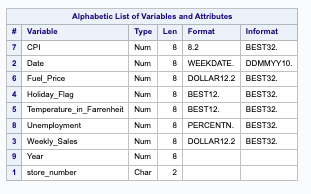
To conduct the exploratory analyses, we need to create a new variable called “year”, which can allow us to compare the annual sales performance across all 45 Walmart stores. To accomplish this task, we can use YEAR function to extract the last 4 digits of the “date” as the value of “year”. The below is the code in SAS with results attached.

**Code in SAS:**



**The Results:**

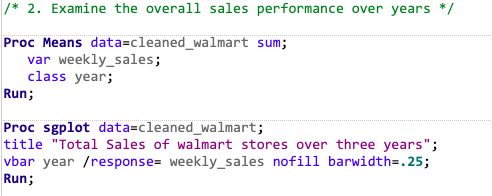




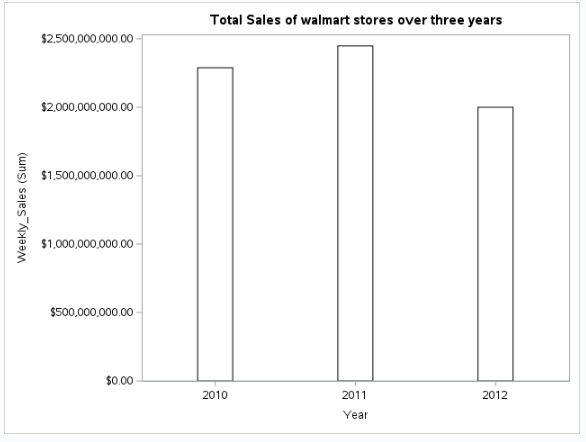
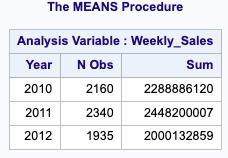
* + - 1. **How is the sales performance across stores over the three years?**

To compare the total sales over the three years, here we simply leverage the proc means to produce the overall sales grouped by year. The below are the codes in SAS and the results.

**Code in SAS:**



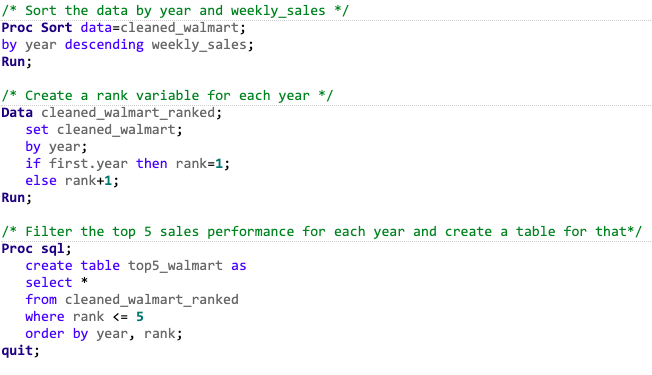
**The Results:**

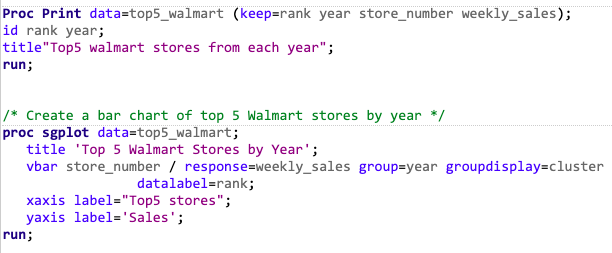


* + - 1. **Who are the top 5 stores from each year?**

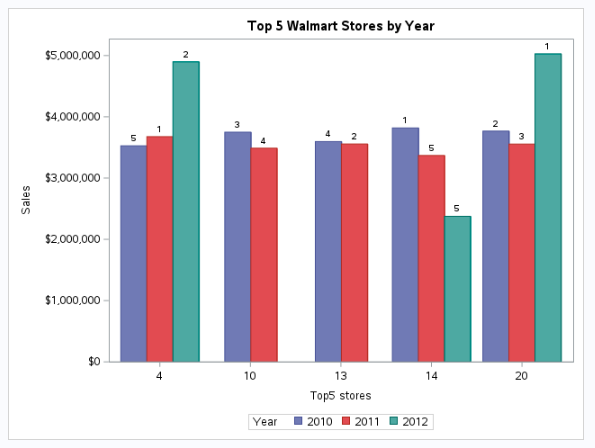
To have a quick peek of the top 5 stores for each year, and we use proc sort, if statement and pro SQL to create a table for the top 5 sales performance stores from each year (2010, 2011 and 2012). Here are the codes in SAS with the results attached as below.

**Code in SAS:**

****

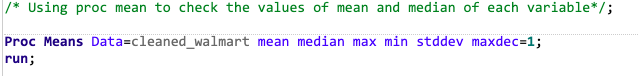
****

**The Results:**

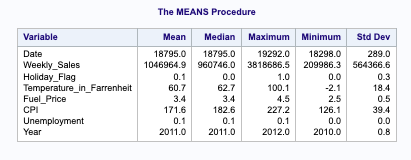


* + - 1. **The graphs or charts help us to explore the normal distribution of the dataset and identify the outliers.**

**Code in SAS:**

****

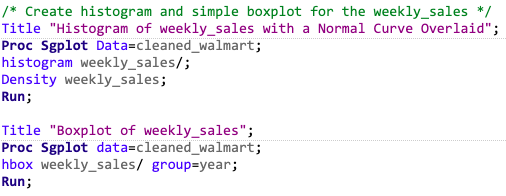
**The Results:**

****

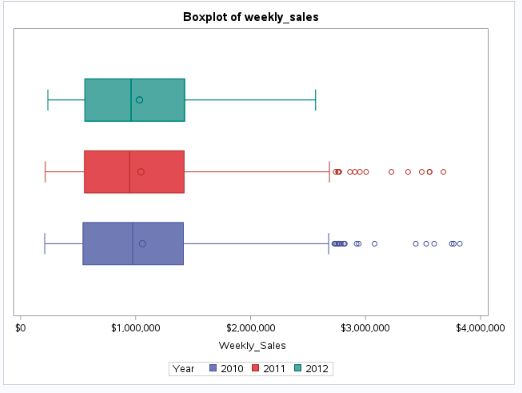
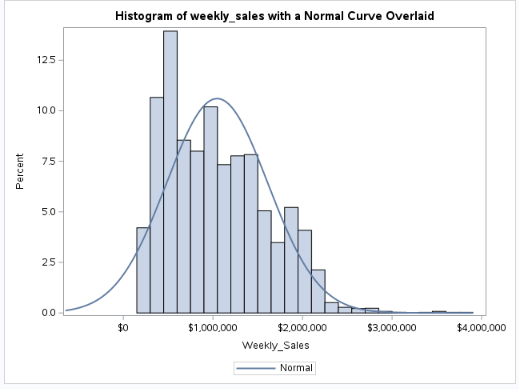
Based on the results, there is no significant difference between Mean and Median, we assume the dataset can be distributed normally, but we still want to create histogram and Boxplot to check the outliers of weekly\_sales and temperature.

We create the histogram and simple boxplot for variable - weekly\_sales , here is the code in SAS with the results attached.

**Code in SAS:**

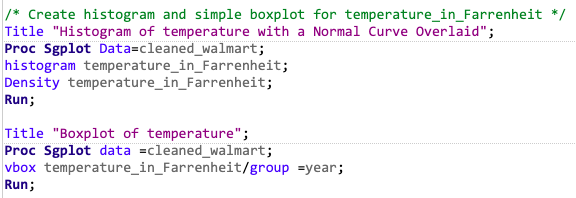
****

**The Results:**

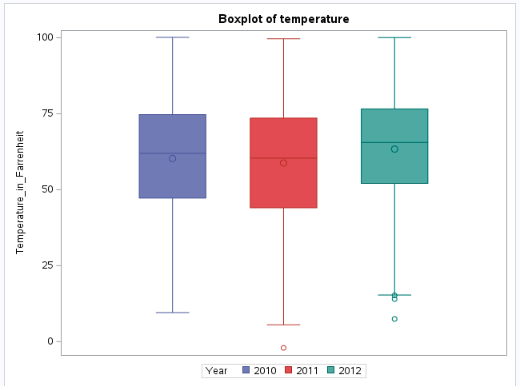
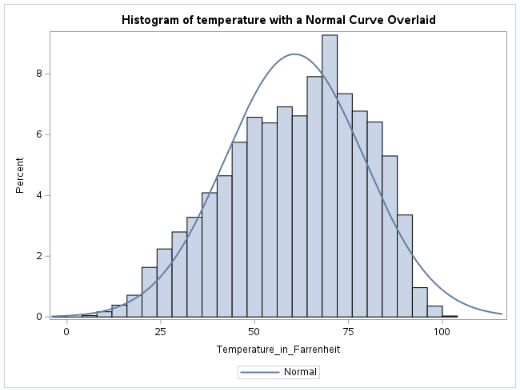


We also create the histogram and simple boxplot for variable - temperature\_in\_Farrenheit, here is the code in SAS with the results attached.

**Code in SAS:**

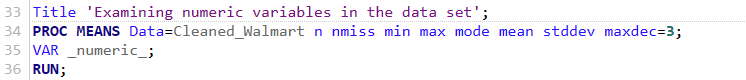


**The Results:**



Based on the histograms and boxplots for the two variables, a few outliers have been identified. Since these outliers won’t have a big impact on the distribution of the value, and decide to ignore these outliers before carrying on the further analysis.

1. **MILESTONE 4:**
   * + 1. **Examine the numeric variable of dataset.**
   * The purpose of this step is to examine if there are errors in the values of dataset that requires step of correcting errors.
   * **Code in SAS:**



* + **The results:** Table

    Description automatically generated
    - There are no missing values in the dataset.
    - Mean and Mode values have no significant differences.
    - Min and Max values are normal.
    - The dataset is ready for analysis.

1. **Identify relationship between Holiday weeks and Sales, whether Sales peak up during the holidays**
   * As variable Holiday Flag has only 02 values 0 and 1 (0 for not Holiday and 1 for Holiday), a ‘ttest’ method would be suitable to explore the relationship between sales and Holiday time.
   * **Codes in SAS**

**Text

Description automatically generated**

* + **Code to show sales in holiday week vs. sales in non holiday week**

**Graphical user interface, Word

Description automatically generated with medium confidence**

* + **The results:**
    - P value < 0.05 from ttest code indicate that Holiday has impacted on sales of Walmart storesTable

      Description automatically generated
    - The chart below indicate that Average sales in holiday week is higher than average sales in non holiday week

Graphical user interface

Description automatically generated

1. **Which holiday impact most on the sales?**
   * There are different holiday seasons during a year, and each season may have different shopping behaviours from shoppers. A graph shows total sales of Walmart stores across the time will indicate which Holiday impact most on the sales.
   * **Codes in SAS**
     + Use SQL statement to create a new table with total sales by store and date.
     + Make a chart to show the sales across the time.

**Text

Description automatically generated**

* + **The result**
    - It’s obvious that Walmart achieved highest sales during
      * 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

**Chart, histogram

Description automatically generated**

1. **What are other factors can be used to predict sales?**
   * As the dataset include different stores in different location, each location would have different values across variables, to do the analysis, a store with highest sale would be selected.
   * **Code in SAS** to create the dataset of the store with highest sales.
     + Use PROC SORT to sort the data by Stores.
     + Use PROC SUMMARY to create a new data set call store\_summary with variable weekly\_sales as total sales of a store
     + Use if condition to eliminate grand total value.
     + Use Proc Sort to identify store 20 as the store has highest sale.
     + Use Where statement to filer original data to create a new data set called store\_data with values of store 20.

Graphical user interface, text

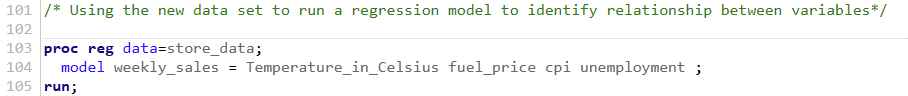
Description automatically generated

* + **The result:**

Table

Description automatically generated

* + - Using the new data set to run a regression model to identify relationship between variables.
  + **Code in SAS:**



* + **The result:**
    - R Square value of 0.069 indicates that only 6.9% of Sales can be explained by the mentioned factors. There are other factors impacted on the sales
    - P values of variables are also > 0.05 except for Temperature, which has a negative relation to sales

Table

Description automatically generated

* + Visualize the regression equation between Temperature and Sales
  + **Code in SAS:**
    - Using scatter plot with linear equation to visualize

Text

Description automatically generated

* + The results:
    - Although it’s not obvious, but we can observe a trend showing that sale is going to slow down as temperature go up.

Chart, scatter chart

Description automatically generated