



Ibrahima Sheik mohamed m

ID: 100781252

DCFall2020

DATA 1202 – DATA ANALYSIS TOOLS FOR ANALYTICS

DATA PERFORMANCE & ROADMAP

Final Project

December 18, 2020

Table of Contents

[Project Summary 1](#_Toc59201228)

[Sales dataset 1](#_Toc59201229)

[Reason for selection 1](#_Toc59201230)

[Tools used for analysis 2](#_Toc59201231)

[Key Statistics of the Dataset 2](#_Toc59201232)

[Descriptive statistics 2](#_Toc59201233)

[Proportion of the categorical text variables 5](#_Toc59201234)

[Correlation 6](#_Toc59201235)

[Data Quality Assessment 7](#_Toc59201236)

[Data Cleaning 11](#_Toc59201237)

[Key Insights 13](#_Toc59201238)

[Question 1 13](#_Toc59201239)

[Question 2 15](#_Toc59201240)

[Question 3 16](#_Toc59201241)

[Question 4 17](#_Toc59201242)

[Python Code 18](#_Toc59201243)

# Project Summary

The purpose of the Data Tools Final Project is to select one dataset from the given homeless and sales option, use the data roadmap and apply data analysis skills to gain insights from the dataset. This report documents the steps involved in examining the dataset, assessing for quality, cleaning the inconsistencies and finally the insights gained from the dataset.

The dataset selected for the Final Project is the “sales” dataset.

## Sales dataset

The sales dataset contains weekly sales of a grocery store chain in CAD for the years 2010, 2011 and 2012. Each store has a store ID and a specific store type. Sales are also separated by department\_ID. The dataset contains information about whether or not it is a holiday week.

Additional information like average temperature during the week in that location, average fuel price in CAD/liter that week and the unemployment rate that week are also available in the dataset.

## Reason for selection

The sales dataset is large dataset with 10774 records and 10 features (including index column). The dataset contains both numerical and categorical variables along with date values. It contains information on weekly sales of a grocery store chain in a location along with average temperature recorded, average fuel price in CAD/liter and unemployment rate during the week in that location.

This dataset seems to be a pretty good dataset for a person who would like to learn the sales trend of a store and gain insights.

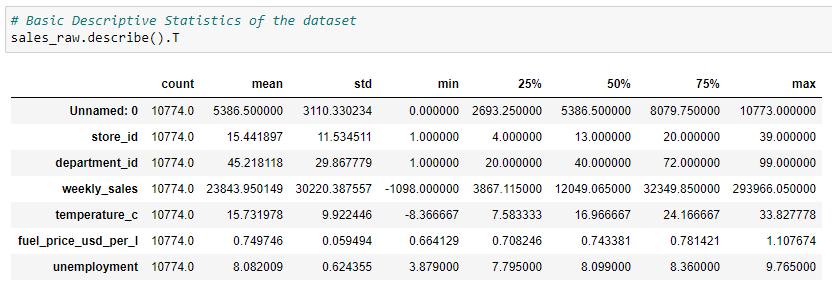
## Tools used for analysis

Python and Jupyter notebook will be majorly used for this data science project. Pandas dataframe is preferred over numpy as Pandas can handle operations with both numeric and categorical variables. Matplotlib and seaborn libraries are be used to plot graphs like boxplot to observe the correlation. Tableau is be used to observe the trend in the data and create visualizations.

# Key Statistics of the Dataset

## Descriptive statistics

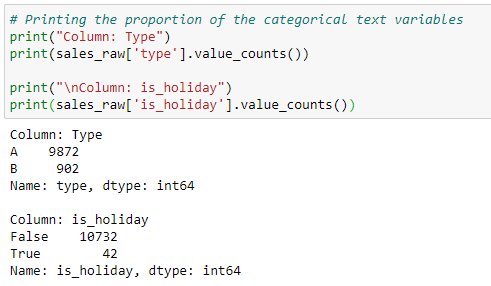
The .describe() function provides the basic statistics of the numerical columns in the dataset. The basic statistics refers to count, min, max, mean, median, standard deviation, 1st quartile (25%) and the 3rd quartile (75%) of the data within the dataset.



* Count
  1. Count represents the total number of records in the dataset which is 10774 for sales dataset.
* Unnamed: 0
  1. The Unnamed: 0 column is the first column in the dataset which holds the unique index value (0 to 10773) for the data in the dataset.
* store\_id
  1. It is the unique id of the store in a particular location.
  2. The store\_id value in the dataset ranges from 1 (min) to 39 (max) with the mean value being 15.44. The spread of the data is ±11.53 from the mean.
  3. Based on 25%, 50% and 75% value, it is clear that majority of the data in the dataset are from stores whose store\_id is <20.
* department\_id
  1. It is the unique id of the department in the grocery stores.
  2. The department\_id value in the dataset ranges from 1 (min) to 99 (max) with the mean value being 45.22. The spread of the data is ±29.87 from the mean.
  3. Based on 25%, 50% and 75% value, the majority of the data in the dataset are from departments whose department\_id is <72.
  4. There are very few records in the dataset which are from departments above 72.
* weekly\_sales
  1. It is the weekly sales of a grocery store chain in CAD.
  2. The weekly\_sales value in the dataset ranges from -1098 (min) to 293966.05 (max) with the mean value being 23843.95. The negative value in weekly sales can be due to the reason that the number of items returned to the store was more than the items sold and people used coupons instead of cash to buy products.
  3. The spread of the data is ±30220.38 from the mean, which shows that the data spread is very large. Mean (23843.95) and Median (12049.07) are not same for the weekly sales.
  4. Majority of the store have their weekly sales less than 32349.85 (75%).
  5. Max (293966.05) is away from the mean projecting that most of the data is accumulated near the min and 75% range, with a tail in the right side i.e., the weekly\_sales data is right - skewed.
* temperature\_c
  1. It is the average temperature during the week in that location.
  2. The temperature\_c value in the dataset ranges from -8.37 (min) to 33.83 (max) with the mean value being 15.73. The min and max range make sense to the data that they denote the weather during different seasons of the year i.e., negative temperature during winter season and max temperature during summer.
  3. The spread of the data is ±9.92 from the mean, which shows that the data spread is small. Mean (15.73) and Median (16.97) are close enough. The temperature\_c data seem to be normally distributed.
  4. Majority of the store sales happened when the temperature was between 25% (7.58) and 75% (24.17).
* fuel\_price\_usd\_per\_l
  1. It is the average fuel price in CAD/liter that week in the location.
  2. The fuel\_price\_usd\_per\_l value in the dataset ranges from 0.66 (min) to 1.11 (max) with the mean value being 0.75. The min and max range show that there has not been much variation in fuel price between the years 2010 and 2012.
  3. The spread of the data is ±0.06 from the mean, which shows that the data spread is very small. Mean (0.75) and Median (0.74) are almost the same.
  4. Majority of the sales happened when the fuel price per liter was below 0.78 dollars.
* unemployment
  1. It is the average unemployment rate in a week.
  2. The unemployment value in the dataset ranges from 3.88 (min) to 9.77 (max) with the mean value being 8.08. The mean and max range show that the unemployment rate has been around 8 and 9.77 for most of the weeks in the year.
  3. The spread of the data is ±0.62 from the mean. Mean (8.08) and Median (8.10) are almost the same.
  4. Max (9.77) seem to be close to mean while min (3.88) is far from mean that the unemployment data could be left-skewed.

## Proportion of the categorical text variables

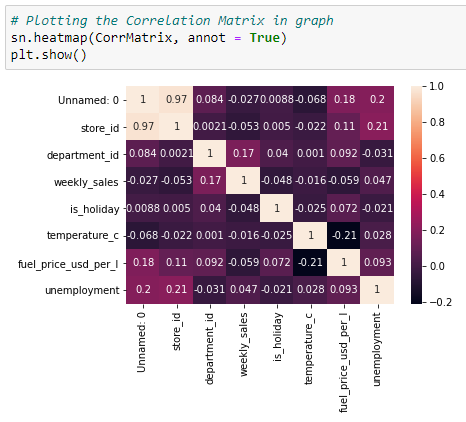
The .value\_counts() function can be used to understand the distribution of categorical text variables in the dataset.



* type
  1. It is the type of the store in a particular location. It takes the values A (9872 records) and B (902 records). Majority of the data are from the store type A and minimal data from store type B.
* is\_holiday
  1. It is a boolean field that represents if the week is a holiday week or not. It takes the values False (10732 records) and True (42 records).
  2. Majority of the data are during non-holiday weeks and very minimal data for holiday weeks which shows that most people shop during non-holiday times.

## Correlation

Correlation is the relationship between the variables present in the dataset. It can be better understood in terms of Pearson correlation coefficient and it ranges from -1 for strong negative correlation, 0 for no relation and +1 for strong positive correlation. Correlation for variables in dataset can be observed with the help correlation matrix or visually using heatmap.

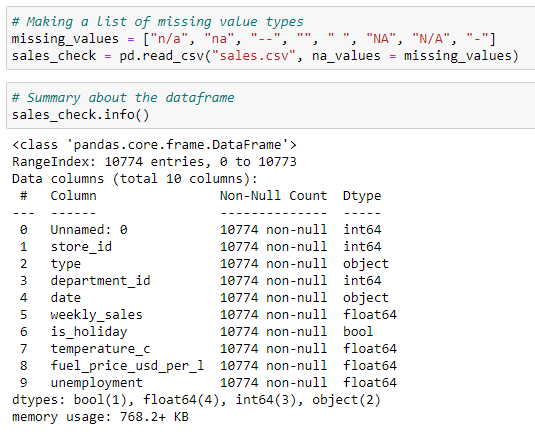


1. Weekly\_sales have weak negative correlation with is\_holiday and fuel price of the week which means that increase in fuel\_price\_usd\_per\_l had negative effect on sales but at a very minimal level.
2. is\_holiday does not seem to have any significant correlation with temperature\_c and unemployment.
3. There seems to be a moderate negative correlation (-0.21) between temperature\_c and fuel\_price\_usd\_per\_l which can be due to the increase in fuel price during winter season as many people avoid transit during snow and in winter.

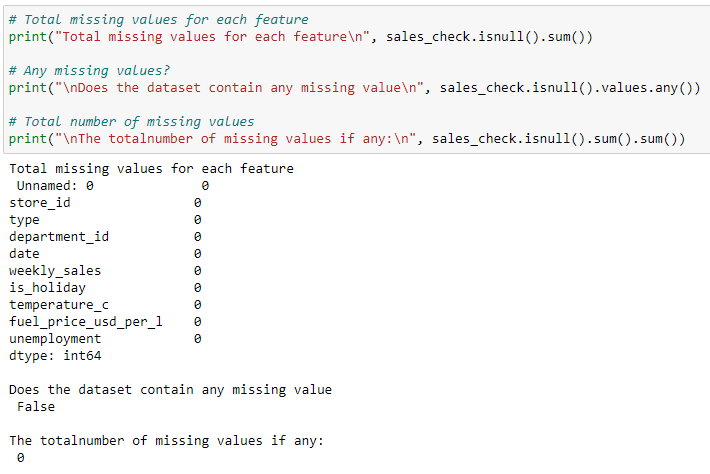
# Data Quality Assessment

This involves checking the dataset for any missing values, outlier, and duplicate records in the dataset.

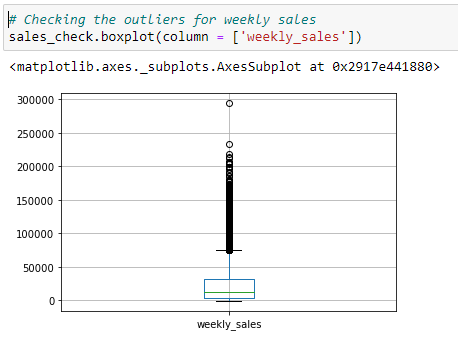
* Identifying Missing Values
  1. “Step: 3 a. Identifying NULL/ Missing values in the dataset” in the python file contains a series of steps/ code applied to the dataset to identify if there is any missing value.



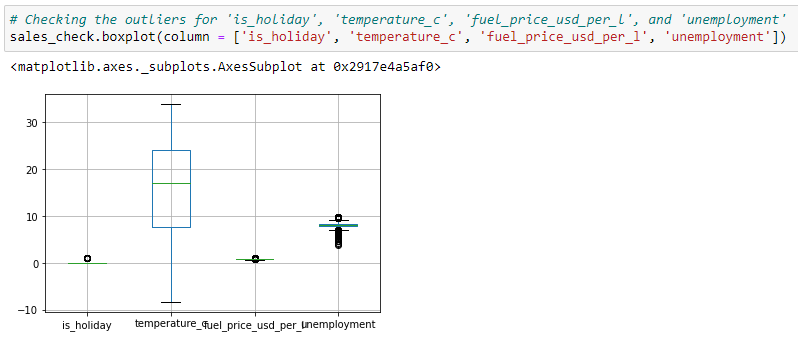
* 1. sales\_check.info() provides information on the summary of the dataframe. Non-Null Count column in the results lists the number of not null columns. Since the Non-Null Count matches the total records, there are no null/ missing values in the dataset.
  2. sales\_check.isnull().sum(), sales\_check.isnull().values.any() and sales\_check.isnull().sum().sum() functions can be used to test if the dataframe has any missing values or not.



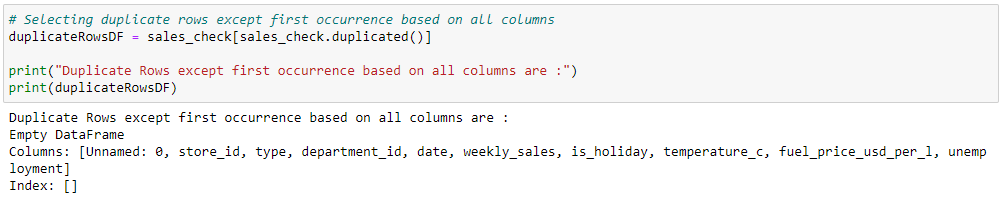
* 1. All the results from the steps showed that there is not any missing or Null value within the dataset.
  2. The first column in the dataset which holds the index value for the data does not have any header name. Pandas recognized the first column and named as unnamed: 0 column.
  3. Since Pandas dataframe creates index on its own and it matches the first column in the dataset, the unnamed: 0 can be removed from the dataframe.
* Detecting Outliers
  1. “Step: 3 b. Detecting Outliers” in the python file contains a series of steps/ code applied to the dataset to detect any outliers.
  2. Outliers, if any, can be visualized by using .boxplot() function to the dataframe. The dots in the box plot denote the outliers.
  3. The columns Unnamed: 0, store\_id, department\_id, temperature\_c does not have any outliers in it.



* 1. The weekly\_sales column in the dataset contains many outliers in it. The big number of outliers is close to max and may be due to the reason that the sales was maximum during festival season, end of year sale week and even during pandemic situation where people purchase goods in bulk prior to the lockdown.



* 1. The outlier in the is\_holiday is close to max (1) and may be due to the very limited record in weekly sales during holiday.
  2. The outlier in the fuel\_price\_usd\_per\_l is close to max and it may be due to the rare condition when the fuel price per liter increased due to the closure of international borders.
  3. The outlier in the unemployment is close to min and max. There are many outliers close to min and few near the max.
* Detecting Duplicates
  1. “Step: 3 c. Detecting Duplicates” in the python file contains steps/ code applied to the dataset to detect any duplicate records.
  2. duplicated() function can be applied to a pandas dataframe to detect if the dataframe contains any duplicate records or not.

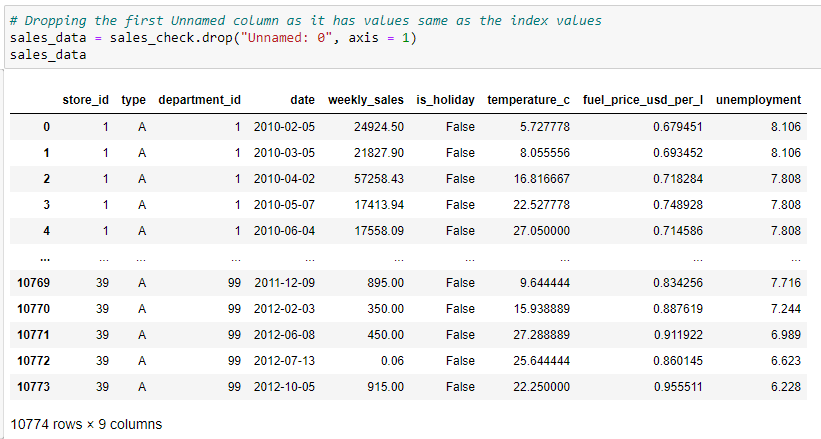


* 1. sales\_check[sales\_check.duplicated()] extracts the duplicate rows except the first occurrence based on all columns in the dataframe. From the result, it is evident that the dataset does not have any duplicate record entries in it.
  2. The first column unnamed: 0 in the dataset contains records similar to the index of the dataframe. Hence, unnamed: 0 can be removed from the dataframe to avoid duplicates.
* Data Inconsistency
  1. There is no consistency in the data for each month of the year i.e., the dataset does not contain data for some months in a year. E.g., No records for the months Jan 2010, Nov 2012, and Dec 2012.
  2. Also, not all stores have records for every month in a year.

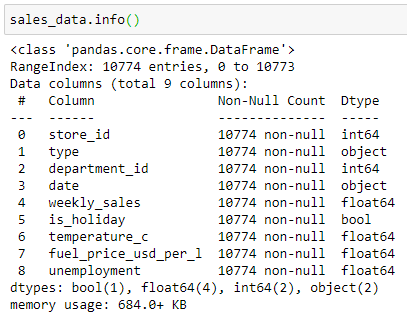
Overall, the dataset quality is medium with some gaps in information to be filled with further appropriate data collection.

# Data Cleaning

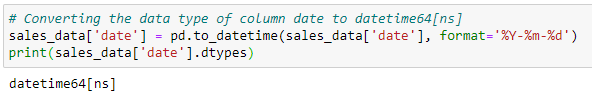
* Missing Value Treatment
  1. There are no missing or null records in the dataset that require treatment.
* Outlier Treatment
  1. The dataset contains outliers for columns weekly\_sales, is\_holiday, fuel\_price\_usd\_per\_l and unemployment. The numbers for those columns seem to be realistic that it cannot be removed from the dataset
  2. The outliers are due to the rare and extreme conditions in sales, fuel price and employment rate which is acceptable in real-world.
  3. By removing or imputing values for outliers, we might be losing some valuable information from the analysis adding bias to the dataset. Hence, the outliers reported in graph will not be imputed or removed to get meaningful insight from the dataset.
* Duplicates Treatment
  1. The first column unnamed: 0 in the dataset contains records similar to the index of the dataframe. Hence, unnamed: 0 can be removed from the dataframe to avoid duplicate columns.

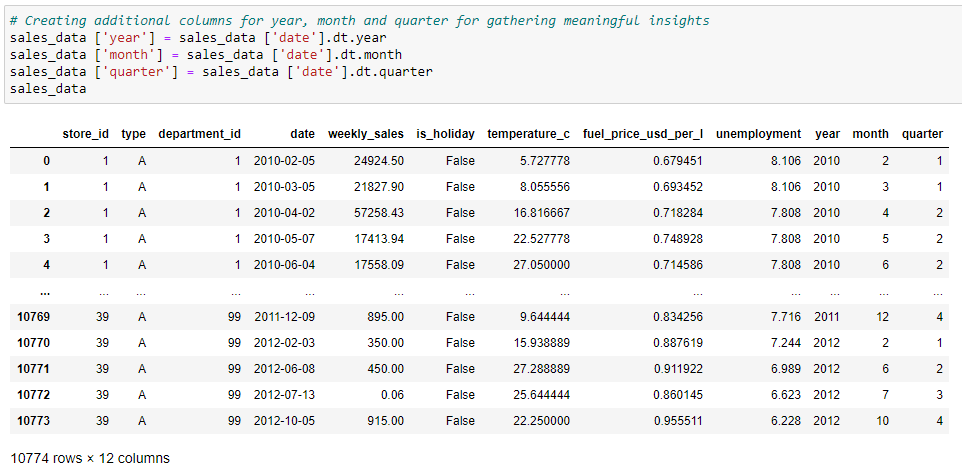


* 1. “Step: 4 a. Removing Duplicate Column” in the python file contains steps/ code applied to the dataset to exclude the first column from the dataframe.
* Converting Data Types
  1. The date column in the Pandas dataframe is stored as an object instead of datetime64[ns] datatype.



* 1. “Step: 4 b. Converting Data Types” in the python file contains steps/ code applied to the dataframe to convert the datatype of date column to datetime64[ns].





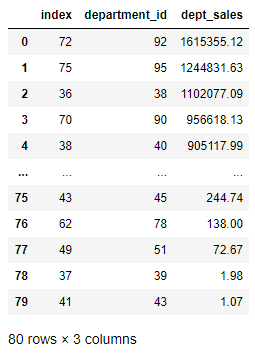
* 1. Creating additional columns for year, month, and quarter for gathering meaningful insights.
* Label Encoding
  1. The dataset contains two categorical text variables and the categories within the attributes are very less that it does not have any significant effect on the analysis.
  2. Hence, no changes made to the text variables in the dataset.

# Key Insights

## Question 1

Which department items were the most and least preferred by the customers in the year 2011?

1. Answer to this question can be obtained by:
   1. Creating a boolean variable and filtering the dataframe based on year value 2011.
   2. Using the boolean variable, creating a subset of dataframe by extracting the data for 2011 from the main dataframe sales\_data.
   3. Grouping the department\_id using groupby function to the subset and computing the sales by department using sum() function.
   4. Sorting the dataframe by dept\_sales to obtain the most and least preferred department items.
2. “Step: 5 a. Question 1” in the python file contains steps/ code applied to the dataframe to obtain the required output.

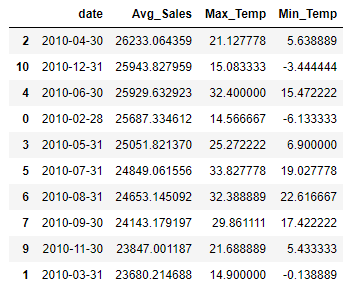


1. It is clear from the above result that the items from departments 92, 95 and 38 were most preferred by the customers and items from departments 39 and 43 were least preferred by the customers in the year 2012. Departments 92, 95 and 38 could be related to essential items for living.

## Question 2

What is the maximum and minimum temperature recorded in a month when the average monthly sales were maximum?

1. Answer to this question can be obtained by:
   1. Grouping the dataframe by month and calculating the monthly average of sales, maximum temperature in each month and minimum temperature in each month.
   2. Using merge function to merge the computed results (monthly\_avg\_sales, max\_temp, and min\_temp) to a single dataframe.
   3. Sorting the dataframe by Avg\_Sales to obtain the required result.
2. “Step: 5 b. Question 2” in the python file contains steps/ code applied to the dataframe to obtain the required output.

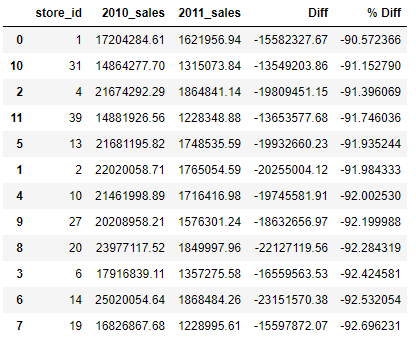


1. It is evident from the above result that the average sales were maximum during the month April 2010 and the maximum and minimum temperature recorded during that month was 21.13 and 5.64, respectively.

## Question 3

How well the sales have been in the stores in 2011 compared to the year 2010?

1. Answer to this question can be obtained by:
   1. Grouping the dataframe by store\_id and calculating the store-wise sales for the years 2010 and 2011.
   2. Merging the dataframe to compute the store-wise difference in sales.
   3. Computing the % difference in store sales between years 2010 and 2011.
   4. Sorting the dataframe by % Diff
2. “Step: 5 c. Question 3” in the python file contains steps/ code applied to the dataframe to obtain the required output.

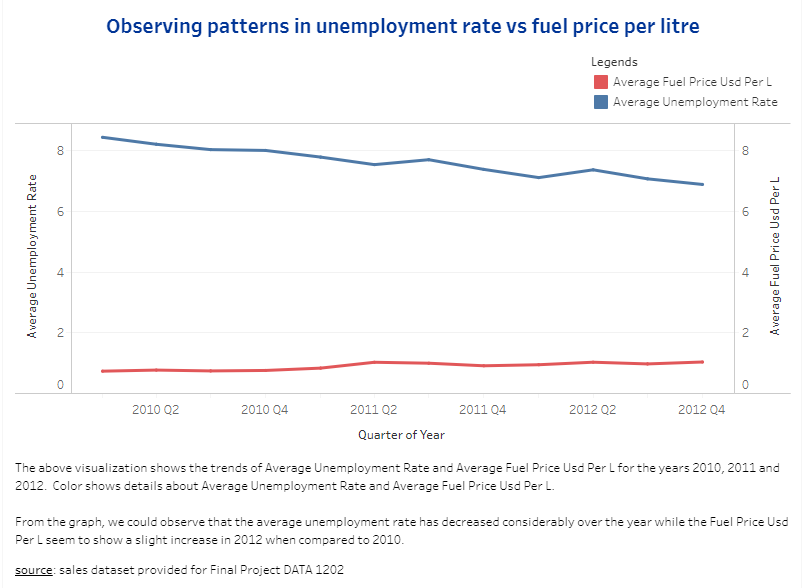


1. The above result shows that there has been a drastic decrease in store sales in 2011 when compared to the year 2010. Almost all the stores show similar trend of near 90% decrease in sales. There could be some serious reason behind this decrease which would require prompt action to be taken.

## Question 4

Is there any interesting change in trend in unemployment rate and fuel price per liter?

Tableau Public URL: <https://public.tableau.com/profile/ibrahima.sheik.mohamed#!/vizhome/Sales_16083252229010/Dashboard1?publish=yes>



1. From the graph, we could observe that the average unemployment rate has decreased considerably over the year while the Fuel Price usd Per L seem to show a slight increase in 2012 when compared to 2010.
2. This can be due to the fact that when unemployment rate decreases meaning people are getting job there comes a demand for more transit and fuel there causing the fuel price to increase.

# Python Code

The python code is available in file “Final Project – DATA 1202.ipynb”.