

Project Sales



Prepared By,

Tom Thondukandathil kuriakose-100840749

Contents

[1 Introduction 2](#_Toc90345409)

[2 Dataset used 2](#_Toc90345410)

[2.1 Definition of the columns in the dataset 2](#_Toc90345411)

[3 Research questions for insights 2](#_Toc90345412)

[4 EDA using Jupyter Notebook 3](#_Toc90345413)

[4.1 Importing the libraries 3](#_Toc90345414)

[4.2 Loading dataset in Jupyter Notebook 3](#_Toc90345415)

[4.3 Dropping unwanted columns from the dataset 3](#_Toc90345416)

[4.4 Data Audit 3](#_Toc90345417)

[4.5 Shape of the dataset 3](#_Toc90345418)

[4.6 Columns in the dataset 4](#_Toc90345419)

[4.7 Checking for missing values 4](#_Toc90345420)

[4.8 Checking for Duplicate values 4](#_Toc90345421)

[4.9 Information on the dataset 4](#_Toc90345422)

[4.10 Basic statistics of the dataset 5](#_Toc90345423)

[4.11 Converting datatype of date from object to date 5](#_Toc90345424)

[4.12 Adding columns Years , Qtr\_Year and Month for analysis 6](#_Toc90345425)

[4.13 Calculations to get Q-o-Q growth rate 6](#_Toc90345426)

[4.14 Downloading dataset into csv 7](#_Toc90345427)

[4.15 Final sales dataset 7](#_Toc90345428)

[5 Insights using Jupyter Notebook 7](#_Toc90345429)

[5.1 Insight 1 - Correction between the parameters 7](#_Toc90345430)

[5.2 Insight 2 -Store wise top 3 sales over the years 8](#_Toc90345431)

[5.3 Insight 3 - Depicting the Q-o-Q growth rate 9](#_Toc90345432)

[5.4 Insight 4 - Top 10 department by sales 9](#_Toc90345433)

[6 Insights using Tableau 10](#_Toc90345434)

[6.1 Breakup of Weekly sales and employment 10](#_Toc90345435)

[6.2 Department wise Analysis with respective to Sales, Unemployment and Holiday 11](#_Toc90345436)

[6.3 Sales analysis based on Store and Department 12](#_Toc90345437)

[7 Conclusion 14](#_Toc90345438)

# Introduction

As part of the assignment, we were required to choose one of 2 datasets provided

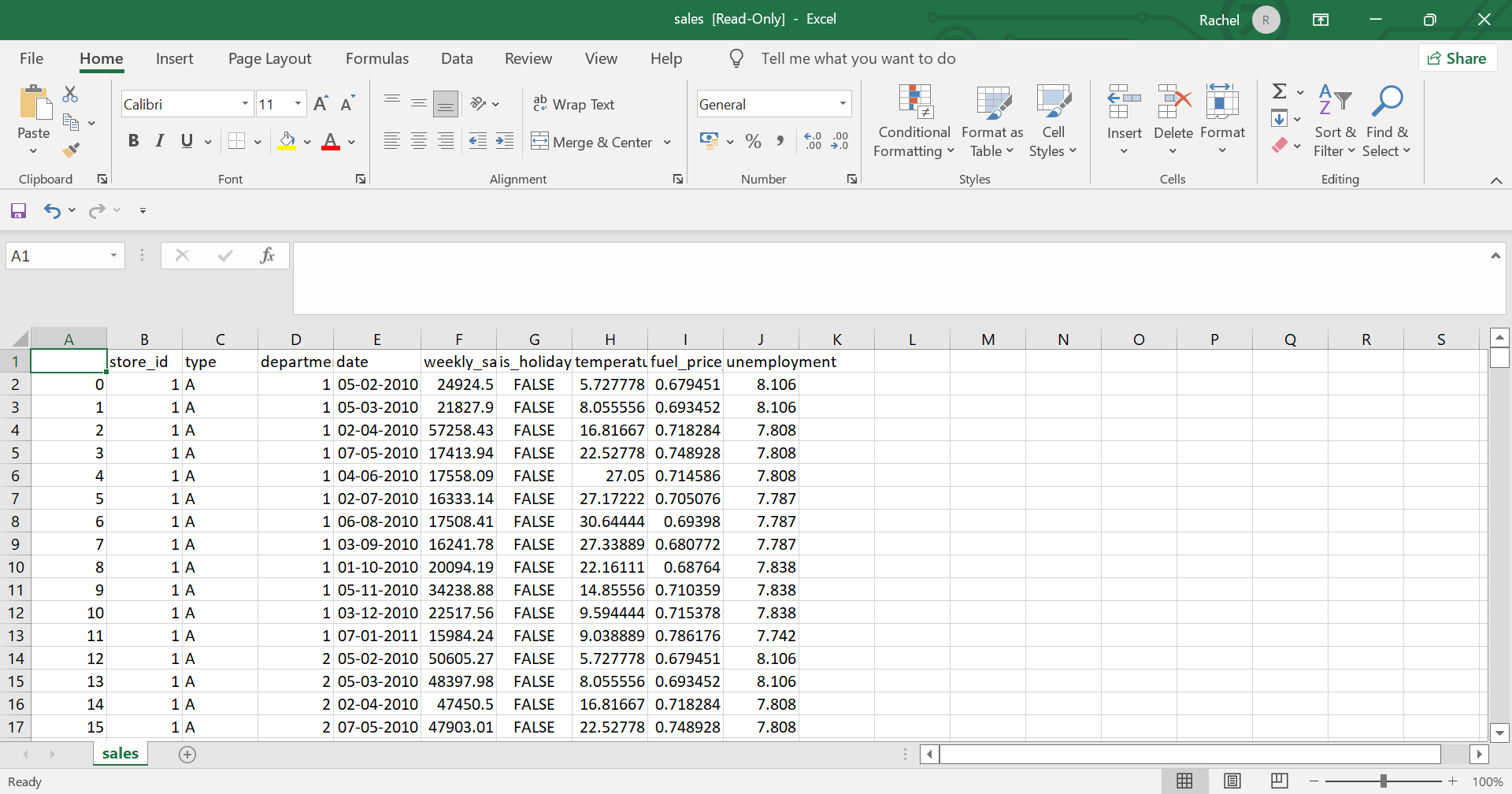
* Homeless dataset
* Sales dataset

For the purpose of this assignment Sales dataset is selected to perform EDA and to derive insights from the dataset.

# Dataset used

For this assignment Sales dataset is selected to perform EDA and get insights based on the data provided.

Below is the sample screenshot of the dataset.



## Definition of the columns in the dataset

This dataset contains weekly sales of a grocery store chain in currency CAD.

* Each store has a store ID and a specific store type.
* Sales are also separated by department\_ID.
* There is information about whether it is a holiday week or not.
* The average temperature during the week in that location is also provided.
* The average fuel price in CAD/litre the respective week is given.
* The final attribute is the unemployment rate the respective week.

# Research questions for insights

Based on the dataset, looking forward to answer the following questions

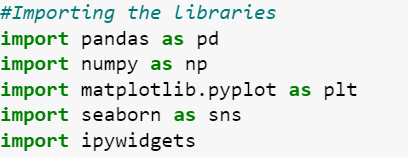
1. Correction between the parameters
2. Store-wise top 3 sales over the years
3. Depicting the Q-o-Q growth rate
4. Top 10 departments by sales
5. Breakup of Weekly sales and employment
6. Department-wise Analysis with respective to Sales, Unemployment and Holiday
7. Sales analysis based on Store and Department

# EDA using Jupyter Notebook

As part of Exploratory Data Analysis in jupyter notebook following steps were taken.

## Importing the libraries

In this step importing the libraries that are required for the analysis and computations.



## Loading dataset in Jupyter Notebook

In this step reading and loading the dataset in jupyter notebook.



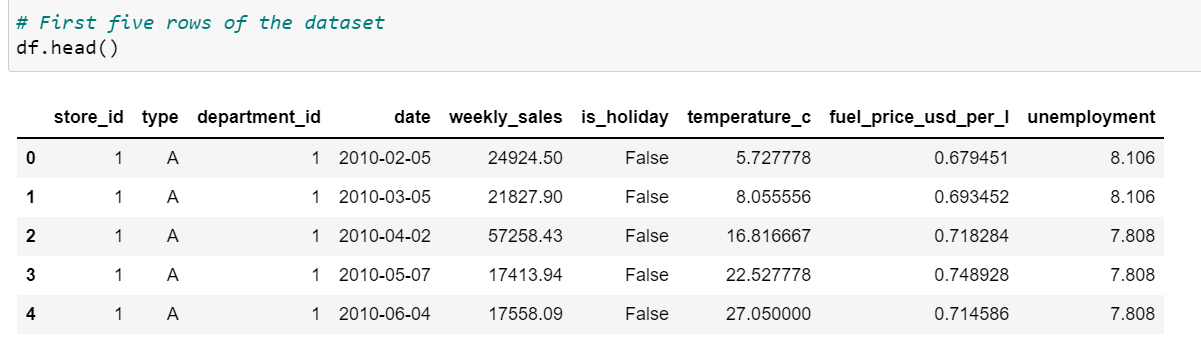
## Dropping unwanted columns from the dataset

As part of EDA process in the dataset there was an unnamed column which wasn’t adding any value, hence, it was dropped.



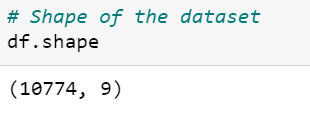
## Data Audit

Getting the initial glimpse of the dataset by displaying the top 5 rows.



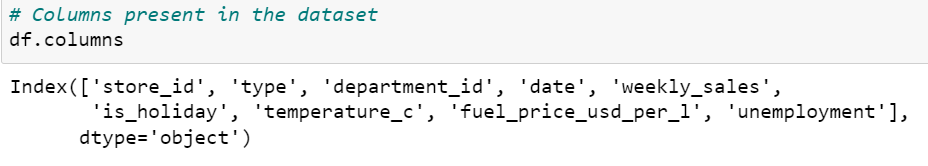
## Shape of the dataset

This provides information, the number of rows and columns respectively in the dataset.



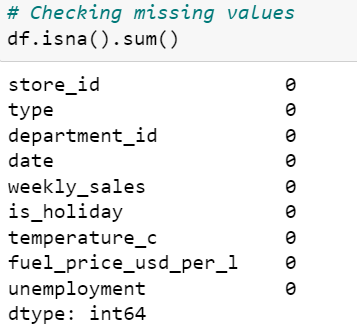
## Columns in the dataset

This provides information on the column names of the dataset.



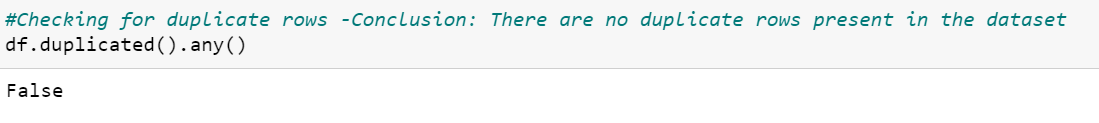
## Checking for missing values

In this step, we are checking if there are any missing values present in the dataset. As the result, respective value for each column is 0, which implies that there are no missing values.



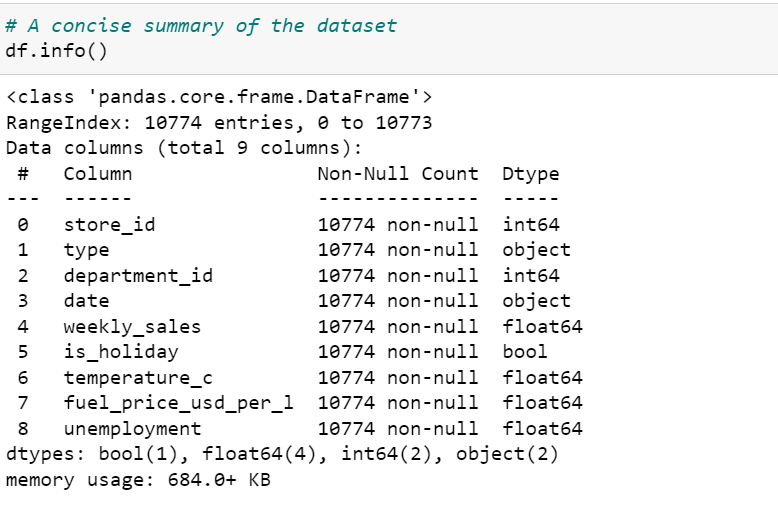
## Checking for Duplicate values

In this step, we are checking if there are any duplicate values in the dataset. The resultant value is False that means there are no duplicate values.



## Information on the dataset

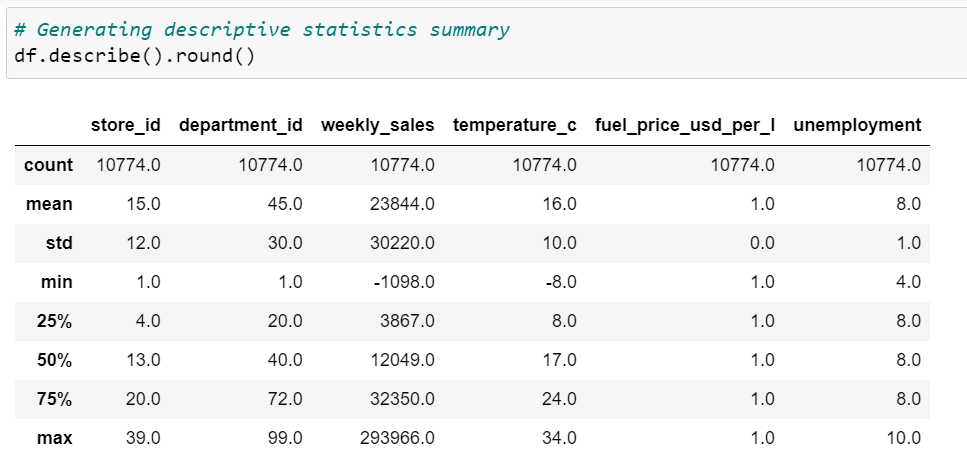
This function provides information of the dataset in terms of datatype of each column, total count of dataset, non-null items in each column.



## Basic statistics of the dataset

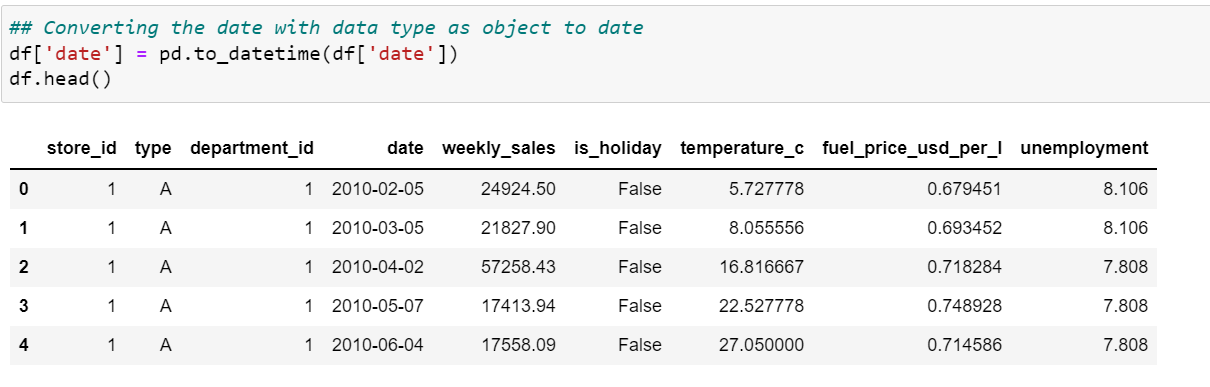
Describe is the basic function to obtain the descriptive statistics which includes, count (number of observations in the respective column), mean, standard deviation, min, 25th, 50th (median), 75th quartile and max of the respective feature in the dataset.

For example, in weekly sales the max is 293966 and minimum value is -1098. The mean is 23844 and the standard deviation from the mean is 30220.



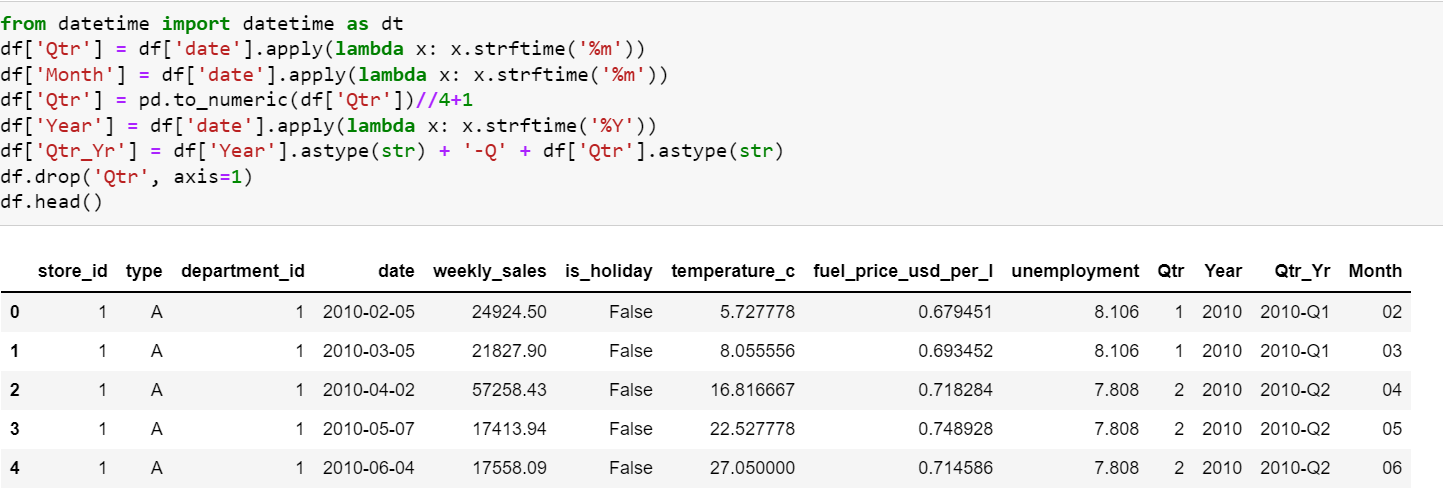
## Converting datatype of date from object to date

For analysis purposes, we need to convert the existing datatype of the date column from object to date.



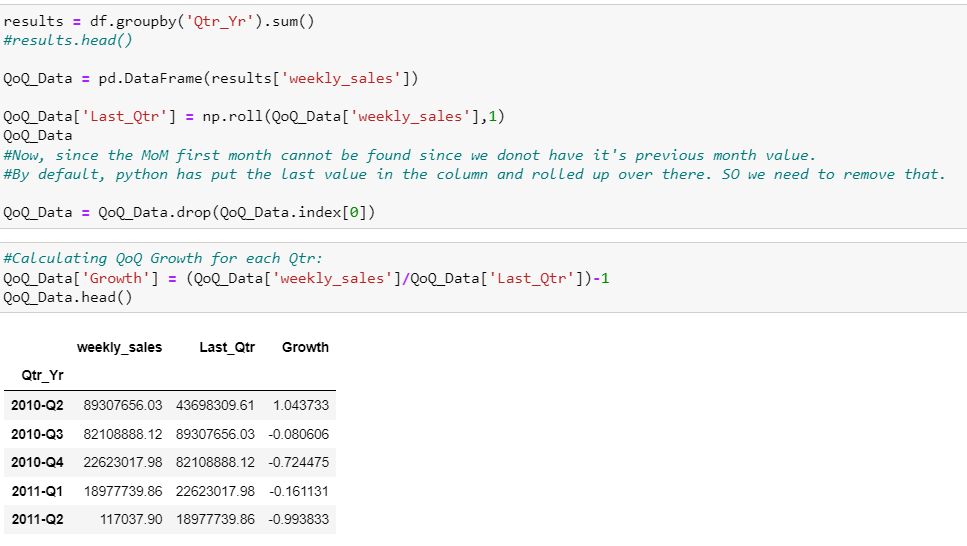
## Adding columns ‘Years’, ‘Qtr\_Year’ and ‘Month’ for analysis

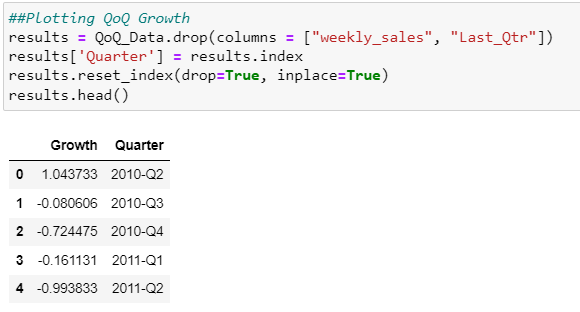
To better analyse the data, we have created new derived columns based on date. These columns are quarter, year and month.



## Calculations to get Q-o-Q growth rate

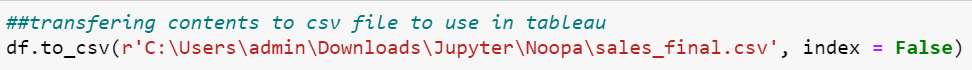
To calculate the Q-o-Q growth rate, the following steps were carried out.





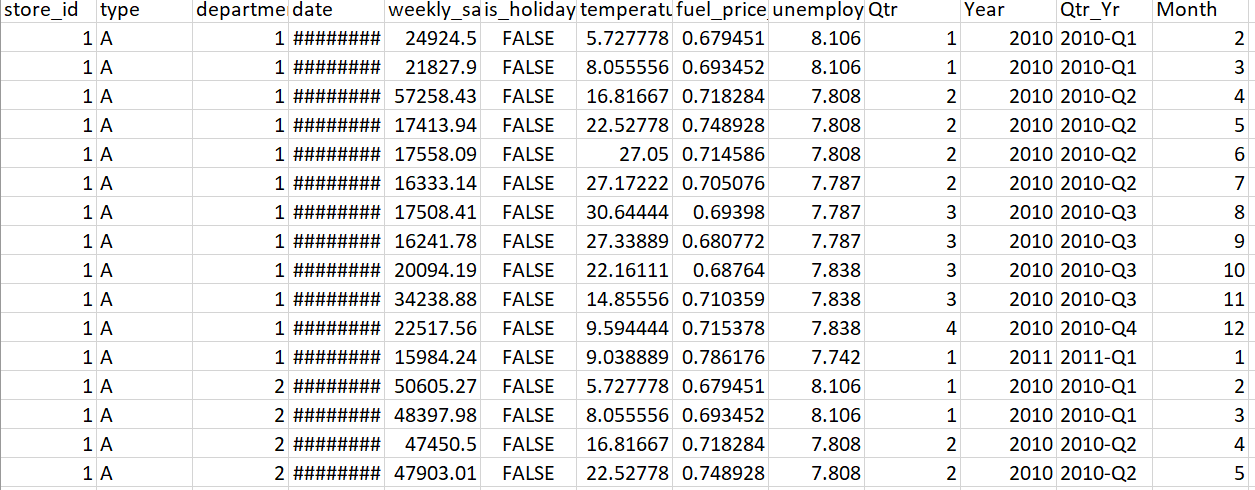
## Downloading dataset into csv

To derive insights using tableau, the dataset with additional columns; year, quarter and month is loaded into a file named ‘sales\_final’.



## Final sales dataset

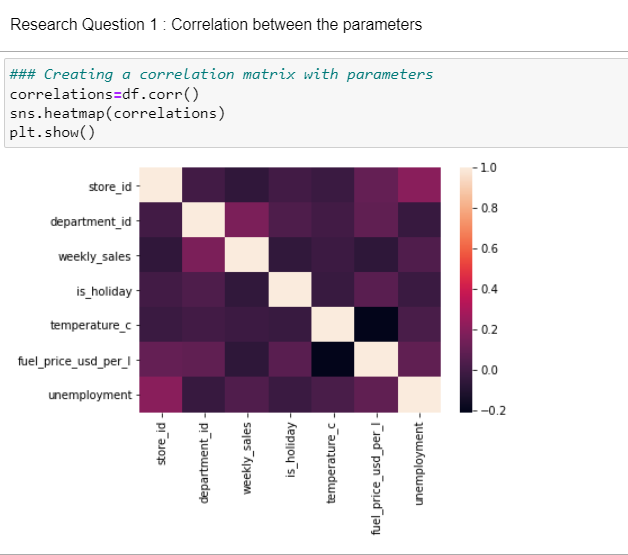
Given below is the sample dataset after saving as a new csv file.



# Insights using Jupyter Notebook

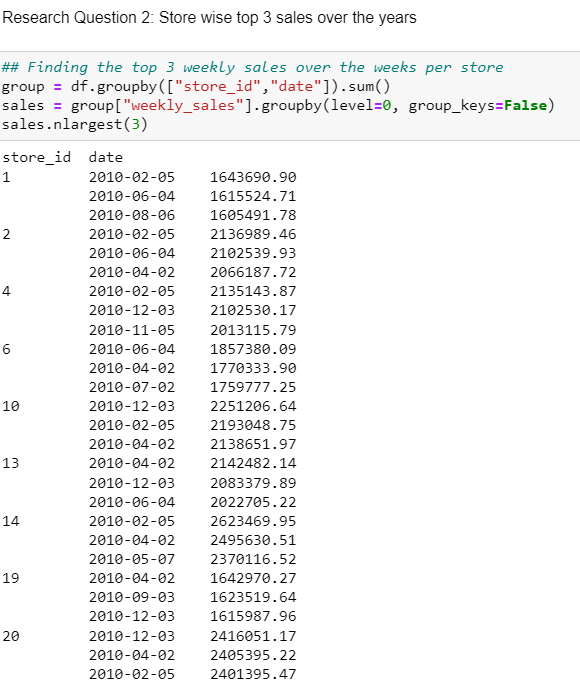
## Insight 1 - Correction between the parameters

The correlation matrix for all the parameters shows less than 0.5, except in case of correlation of the parameter with itself. Store id and unemployment seem to be highly correlated with each other as compared to others. Temperature and fuel price have negative correlation with each other.



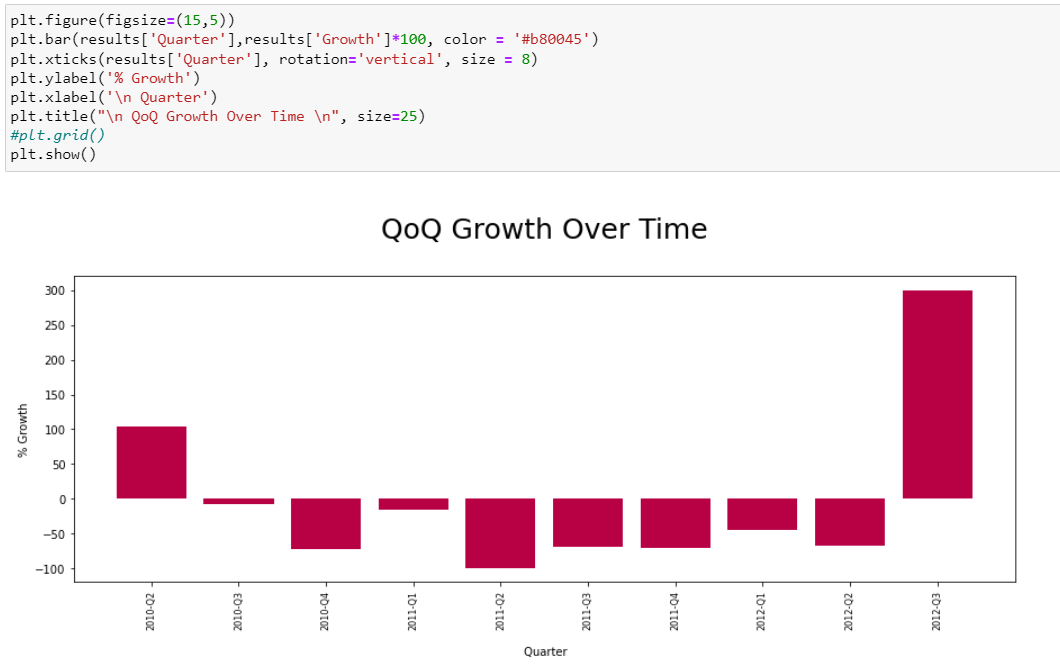
## Insight 2 -Store wise top 3 sales over the years

Using the code below in jupyter notebook, we have found out the top 3 weekly sales for all the stores for all, from all the possible dates available in the dataset.



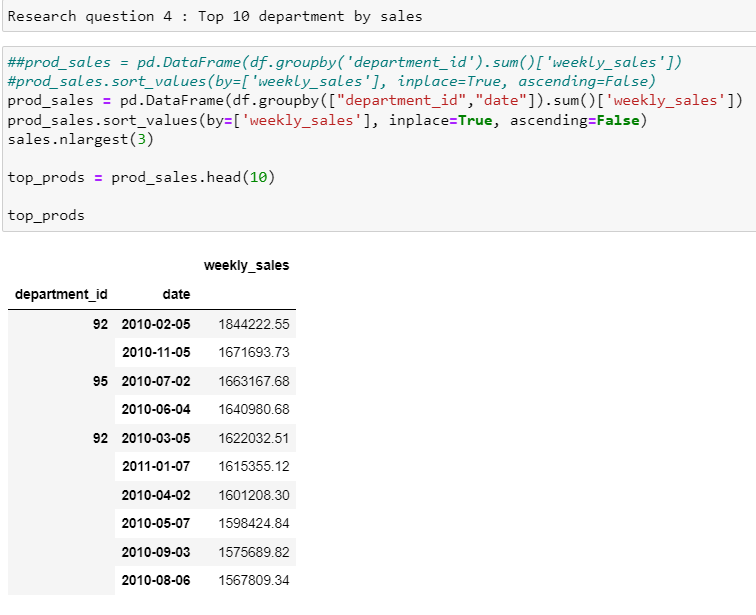
## Insight 3 - Depicting the Q-o-Q growth rate

Here, we have shown a diagrammatic representation of Q-o-Q growth rate. 2010-Q2 shows the higher growth rate at 104 %, however since then, there has been a decline in all of the quarters. There is a major increase in the growth in third quarter of the year 2012, at around 300%, considered as the best growth rate. The worst rate is at 2011-Q1 at around -99% (decline rate).



## Insight 4 - Top 10 departments by sales

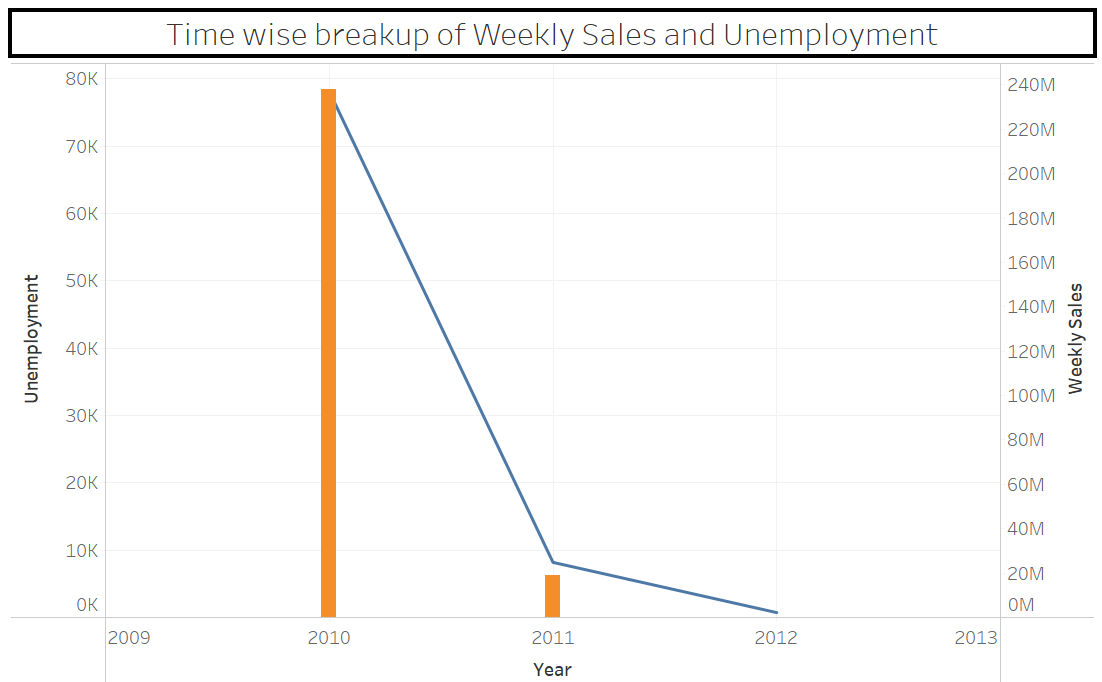
In this insight, we are showing the top 10 departments with the maximum sales over the time period given in the dataset. For example, department 92 has been consistently performing well, as 8 out of the top 10 spots belong to department 92. This seems to have been in year 2010 thereby depicting the decrement in sales over the next two years. Exceptionally, for 1 week of 2011 (2011-01-07) all other sales for the department 92 have the highest sales in 2010.



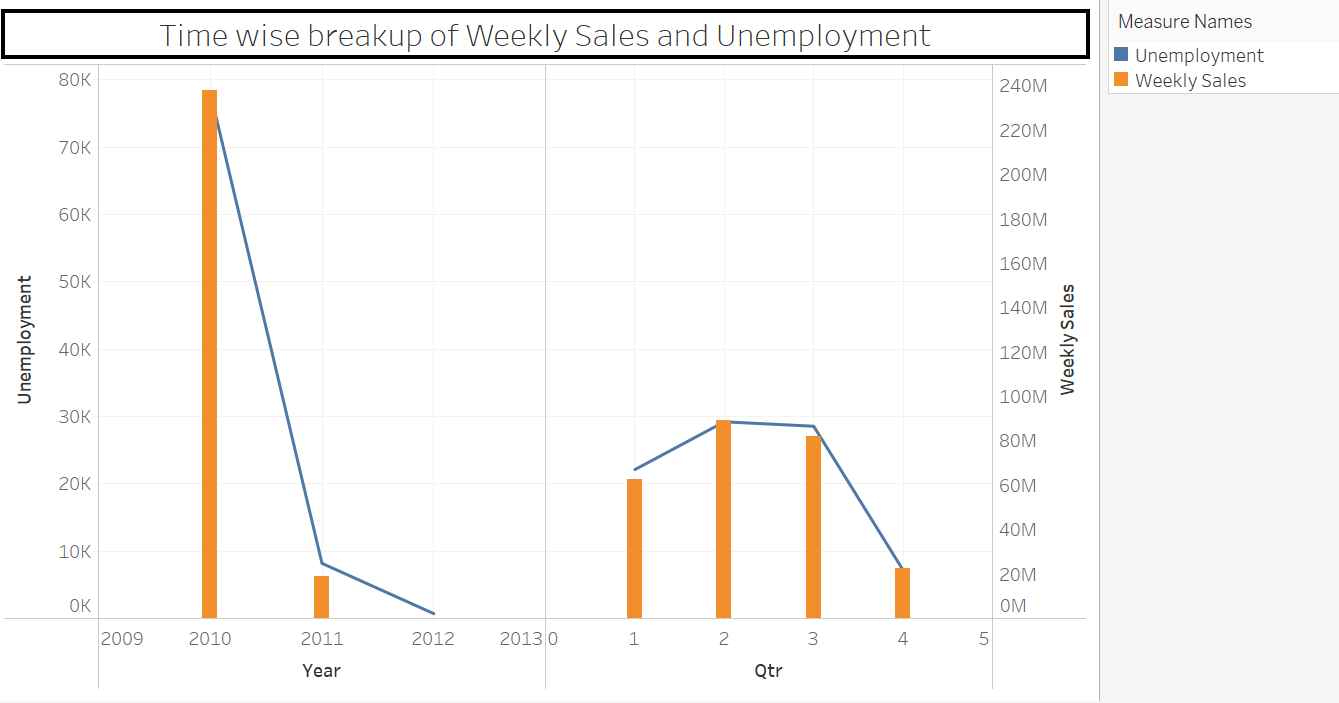
# Insights using Tableau

## Breakup of Weekly sales and employment

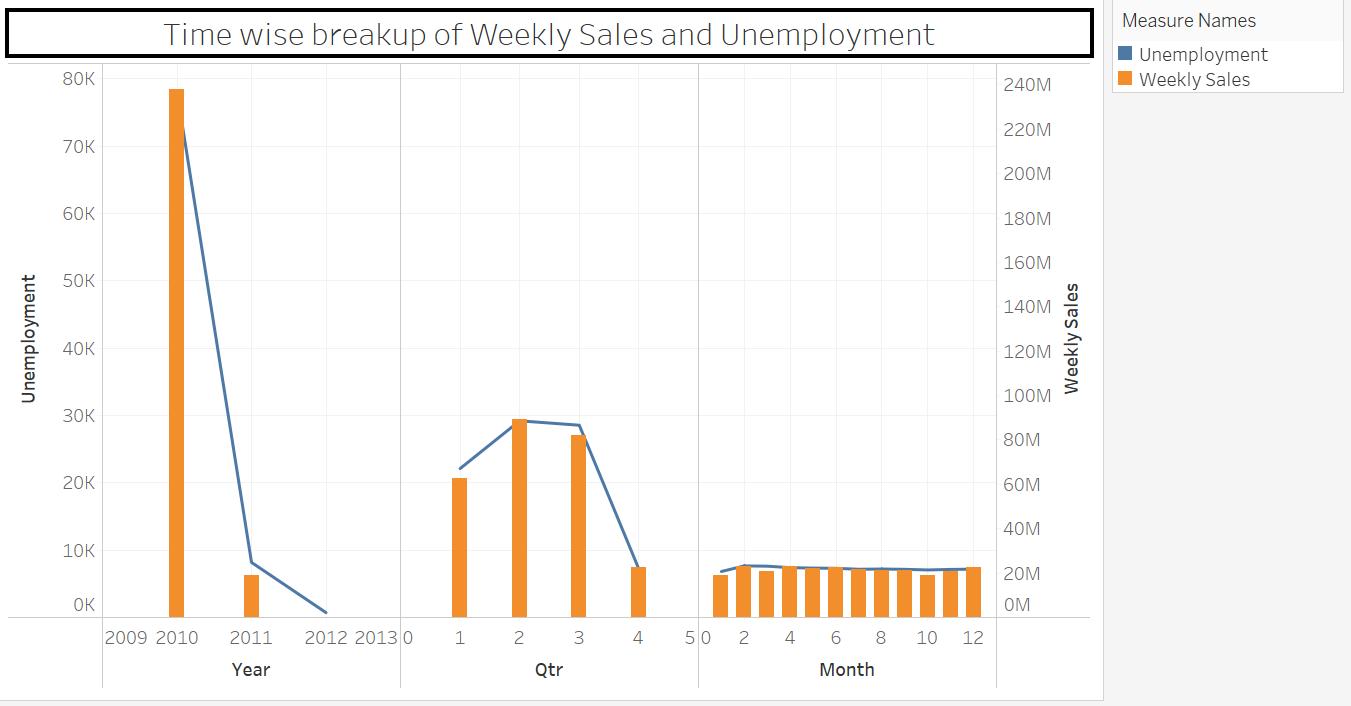
This chart shows the break down of weekly sales and unemployment. It is observed there is a direct correlation between weekly sales and the unemployment. Increase in sales leads to an increase in unemployment.



This chart can further be broken down to capture q-wise details for weekly sales and unemployment.

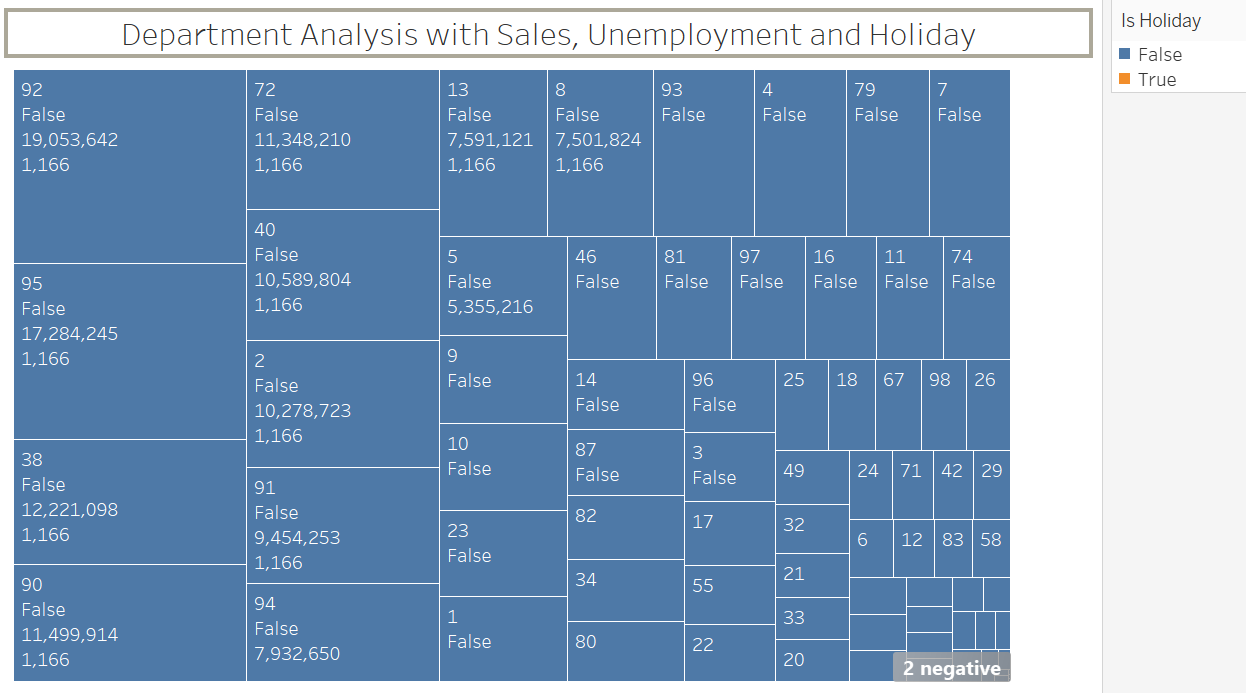


Similar break down can be done for monthly basis.



## Department wise Analysis with respective to Sales, Unemployment and Holiday

The heat map weekly sales, unemployment value and holiday in relation with department. From the visualization created, it looks like the sales are maximum when it is not a holiday.

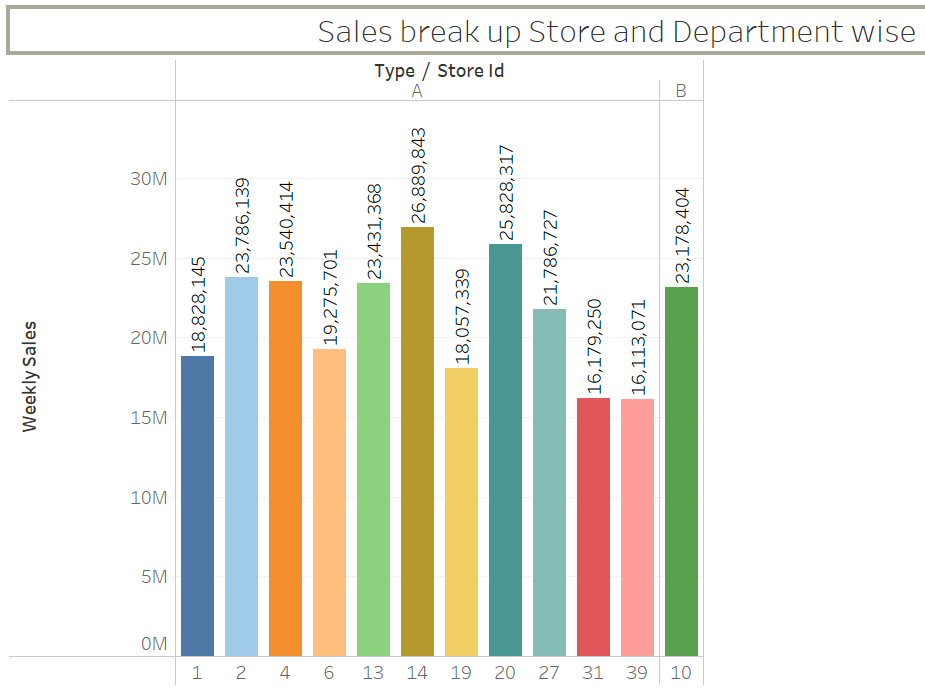


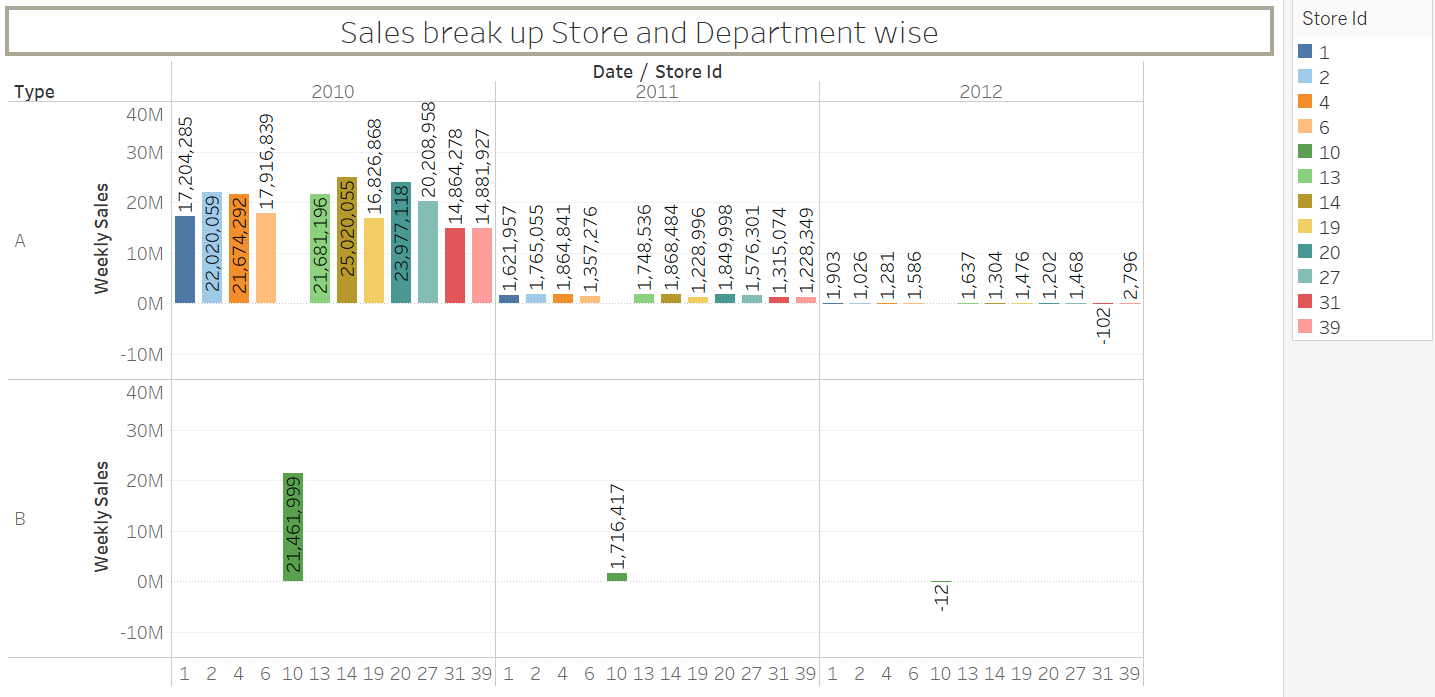
When there is holiday only a couple of stores seem to be open and the effect is visible in the sales number. When there is a holiday, Department 99 has sales of 12,150 and unemployment at 29.76. The maximum sales for a non-holiday for department 92 is 19053642 which is 1000 times more than the store 99 which operates on a holiday. Also, similar observation can be made for unemployment, when on a non-holiday, it is almost 100 times that of a holiday store.



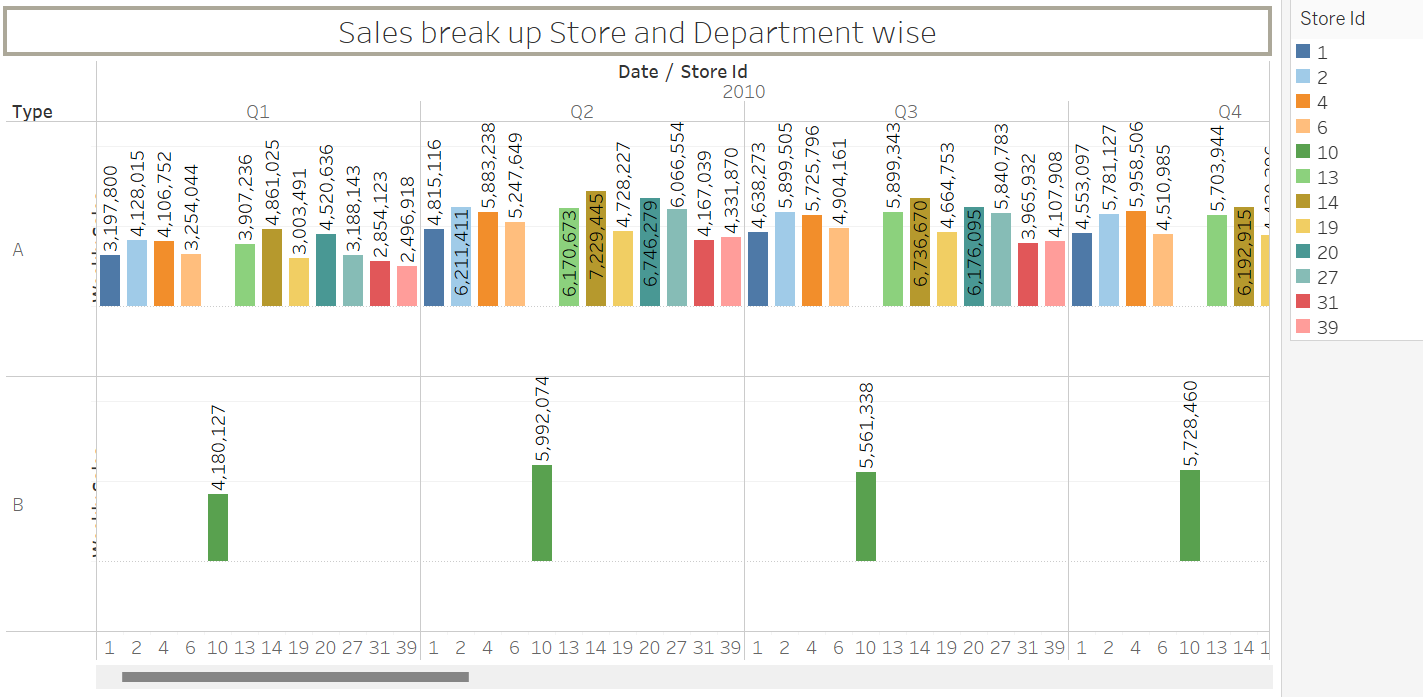
## Sales analysis based on Store and Department

This sales analysis shows only a single store operates in type B and all other stores operate under type A. This shows 90% of business is done by store A as compared to store B. Comparison between the 2 stores type and with the store id can be done based on yearly, quarterly and monthly basis.





Similarly, the break up can be shown for Quarter and month. This gives a snapshot of data within the time period to understand how the stores are behaving over the period of time.



# Conclusion

Based on the analysis of the dataset we have answered the question we wanted from the dataset, like store wise top 3 sales data over the time period. The Q-o-Q growth rate over the period wherein the maximum growth was observed in the last quarter at 300%. Top 10 department by sales shows that department 92 has been performing consistently well as compared to the other department. Sales and unemployment over the time period have a direct relation with each other, as the sales increases, unemployment also increases. Additionally, as part of analysis we have checked if holiday has significant impact on the sales over time period for a department. The maximum sales for a non-holiday department 92 is 19053642 where it is 1000 times more than store 99 which operates on a holiday. Similar observation can be made for unemployment when on a non-holiday, sales is almost 100 times that of a holiday store. Lastly, Store type A does 90% of business as compared to type B.