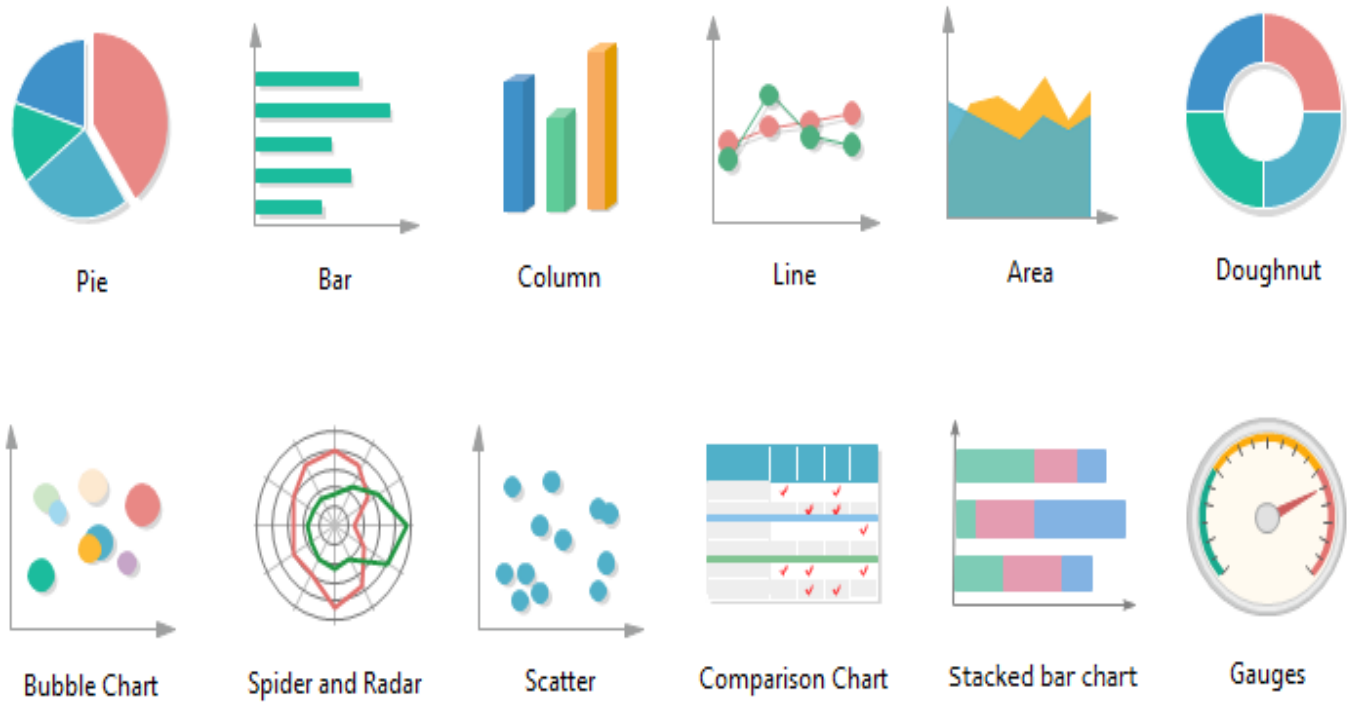
RETAIL ANALYSIS WITH WALMART DATA

PROJECT REPORT



BY

PRAJVAL RANKA
JUNIOR DATA SCIENTIST
http://linkedin.com/in/prajvalranka17

SYED TOWHEED
BUSINESS ANALYST INTERN
http://linkedin.com/in/syed-towheed-3078681a0

AKIB SHAHRIAR
DATA ANALYST INTERN
http://linkedin.com/in/akib-shahriar-b7919991

Executive Summary

The project is based on Walmart which is one of the leading retail stores in the US. 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 and needs some business solution with the help of data analytics .This projects analyzes the provided data and give insights on the pattern of weekly sales and with the help of certain KPI's helps to provide solutions. The project also aims to predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc. with the help of multilinear regression models.

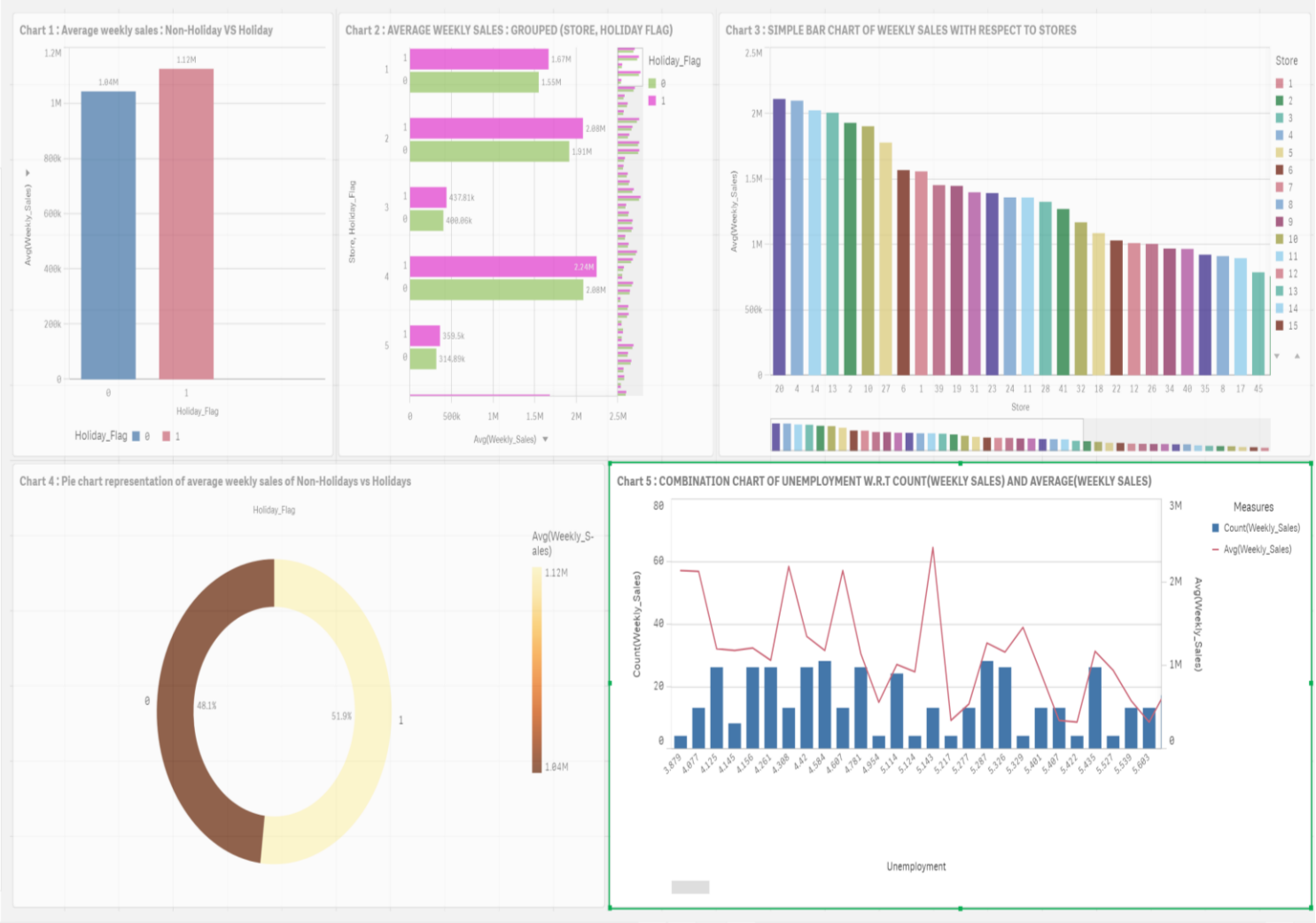
Dataset Description

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:

1. Store - the store number
2. Date - the week of sales
3. Weekly Sales - sales for the given store
4. Holiday Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
5. Temperature - Temperature on the day of sale
6. Fuel Price - Cost of fuel in the region
7. CPI – Prevailing consumer price index
8. Unemployment - Prevailing unemployment rate

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



- In chart 1, we observe that average Weekly Sales made during a holiday week is higher.
- Average Weekly Sales are $1.12 million for a holiday week as compared to $1.04 million for a non-holiday week.
- In chart 2, all the individual stores have higher average Weekly Sales for a holiday week than non-holiday week. The pink bar is for holiday and green bar for non-holiday.
- In chart 3, we observe the bar chart is displaying average Weekly Sales in the order of descending.

- Store 20 shows the highest average Weekly Sales throughout all 3 years in total and Store 33 shows the lowest average Weekly Sales. Therefore, the Weekly Sales is depending on each store as stores are performing differently.

- We can conclude similar finding through the donut chart, where out of Total Sales 51.9% of the sales were made during the holidays and 48.1% of the sales were made during the non-holidays.

- We chose average Weekly Sales as a measure and not the sum of weekly sales was because since the number of non-holiday weeks are higher than the holiday weeks. The difference between the sum of sales between holiday week & non holiday week were huge and really could not provide a good insight of the kind of comparison we were looking for.

- In the fifth chart, we tried to see if unemployment rates affected the average Weekly Sales and the number of transactions.

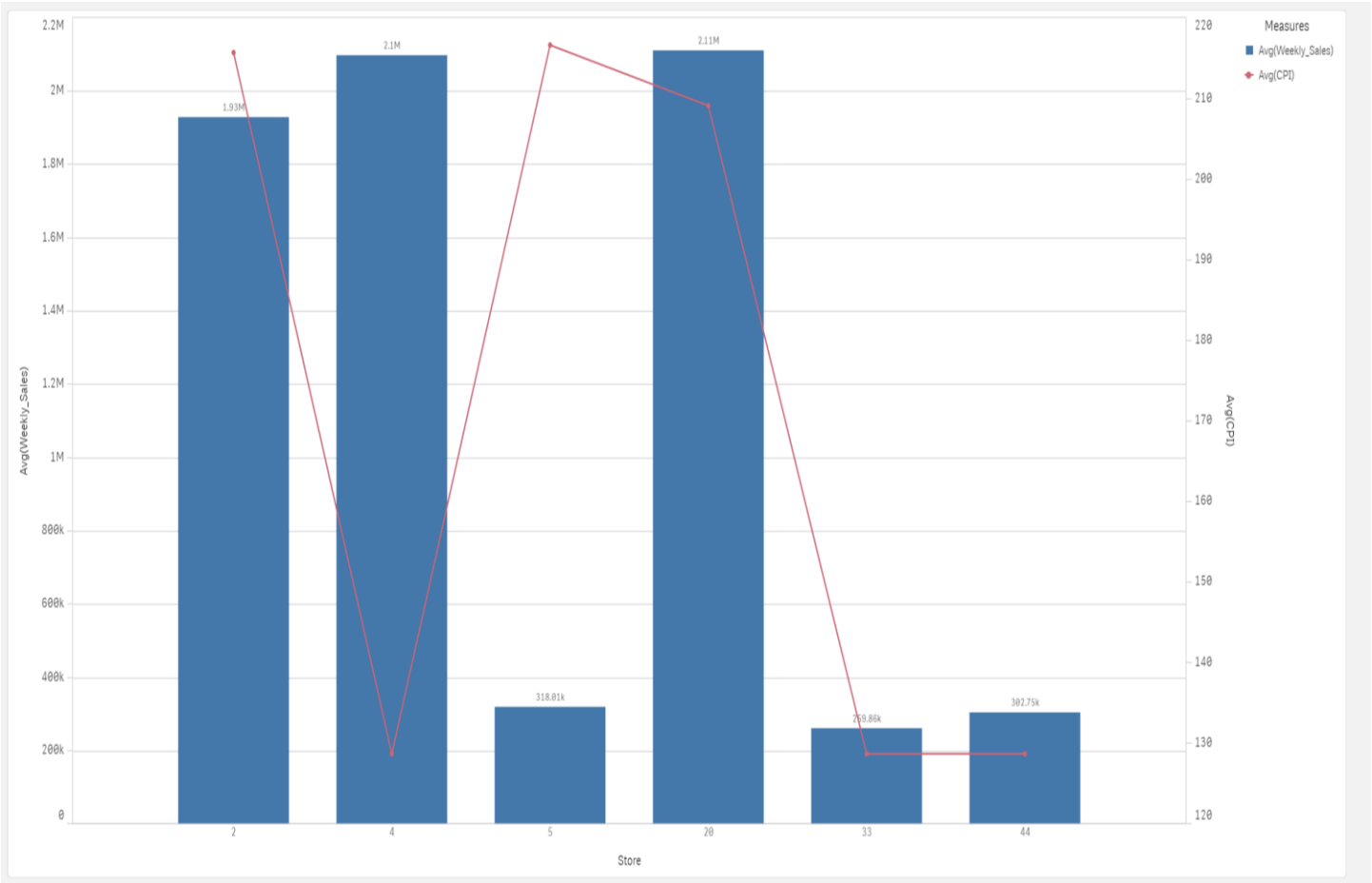
- With the change in rate of unemployment, the weekly sales and number of transactions did not respond significantly and was not too conclusive to make any final deduction.

We assume that:

1. We lack data on sales per department rather we have total store sales but since the total sales do not respond with change in unemployment rate so we assumed that bulk of the sales are from essential items department which is likely to not too responsive to the change in unemployment.

2. We assume that the reason behind more sales during a holiday is because during the holiday season people are getting bonus, and there are different store promotions offered to the consumer in terms of discounts and sales. And on a holiday people are more likely to buy more non-essential items and the foot traffic is high during a holiday season.

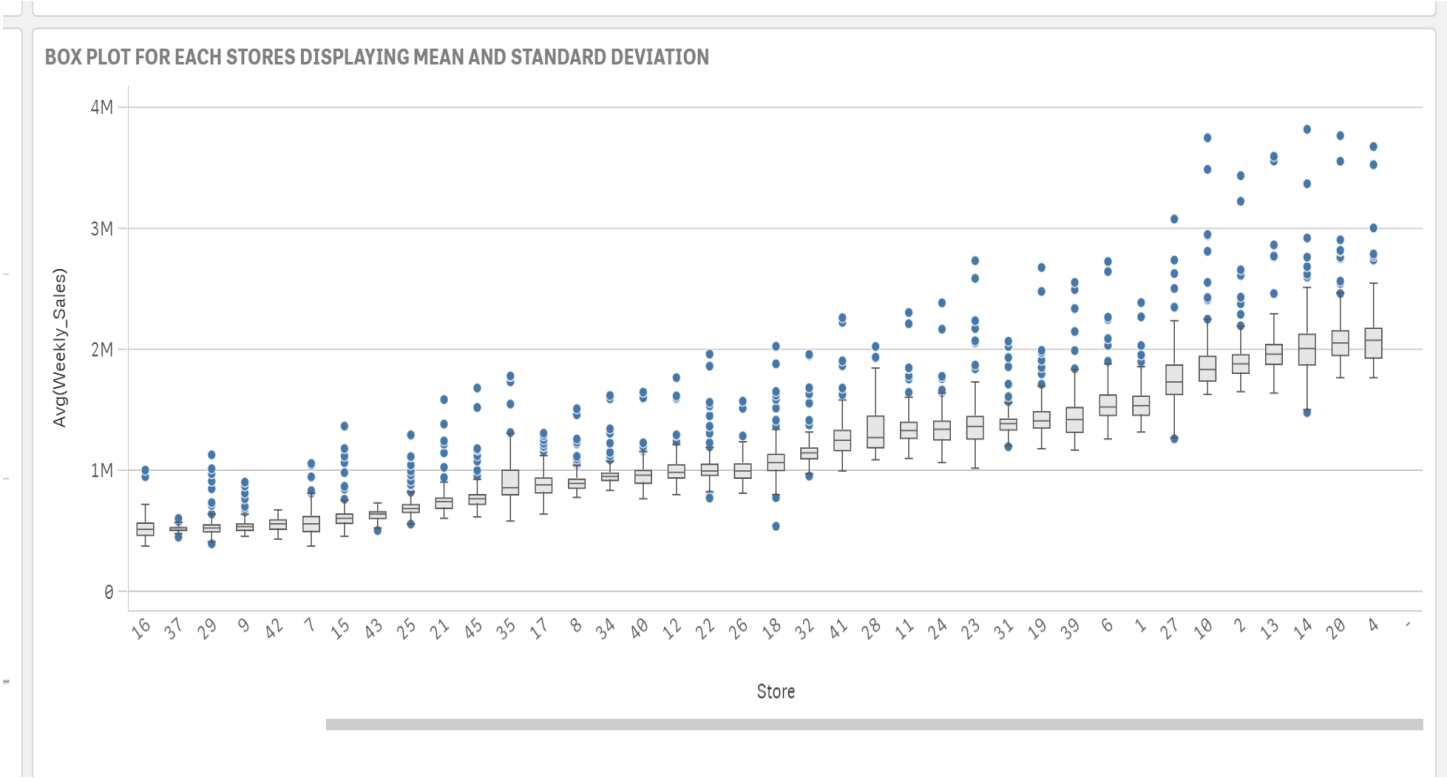
3. One of the main finding in the difference in sales in different stores might be due to difference in CPI. For instance, store 20 and store 2 with highest sales has an average CPI of 209 and higher compare to store 33 and store 44 with the lowest sales and with an average CPI of 128 significantly lower than the average CPI of 171.



- However, for Store 4 where sales are high despite of low CPI might be a result of lack of competition in the area so people are making regular visits to Walmart and making most of their essential shoppings as the options are few.

- In contrast, Store 5 has different trend where sales are low regardless of high CPI. We assume that the store is located in a town centre where despite of people with high CPI, presence of many competitor in the area are causing low sales.

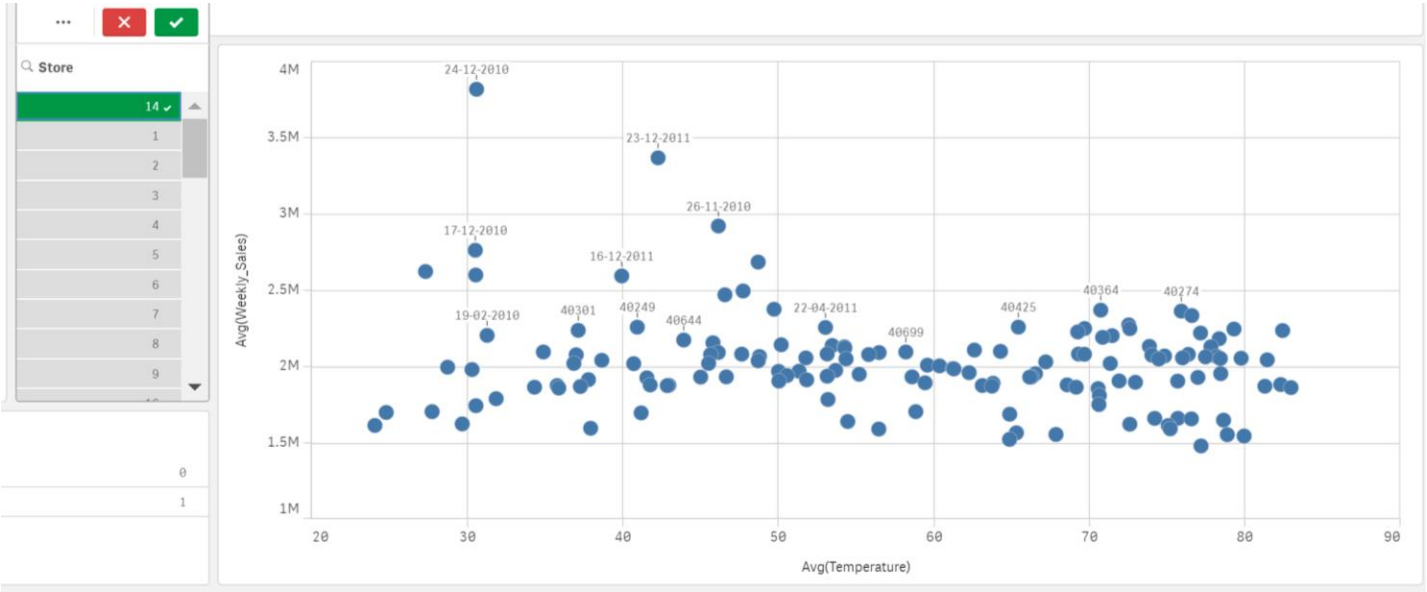
- We also assume that people with high CPI might shift to high end brands rather than Walmart, which is famous for selling value products.

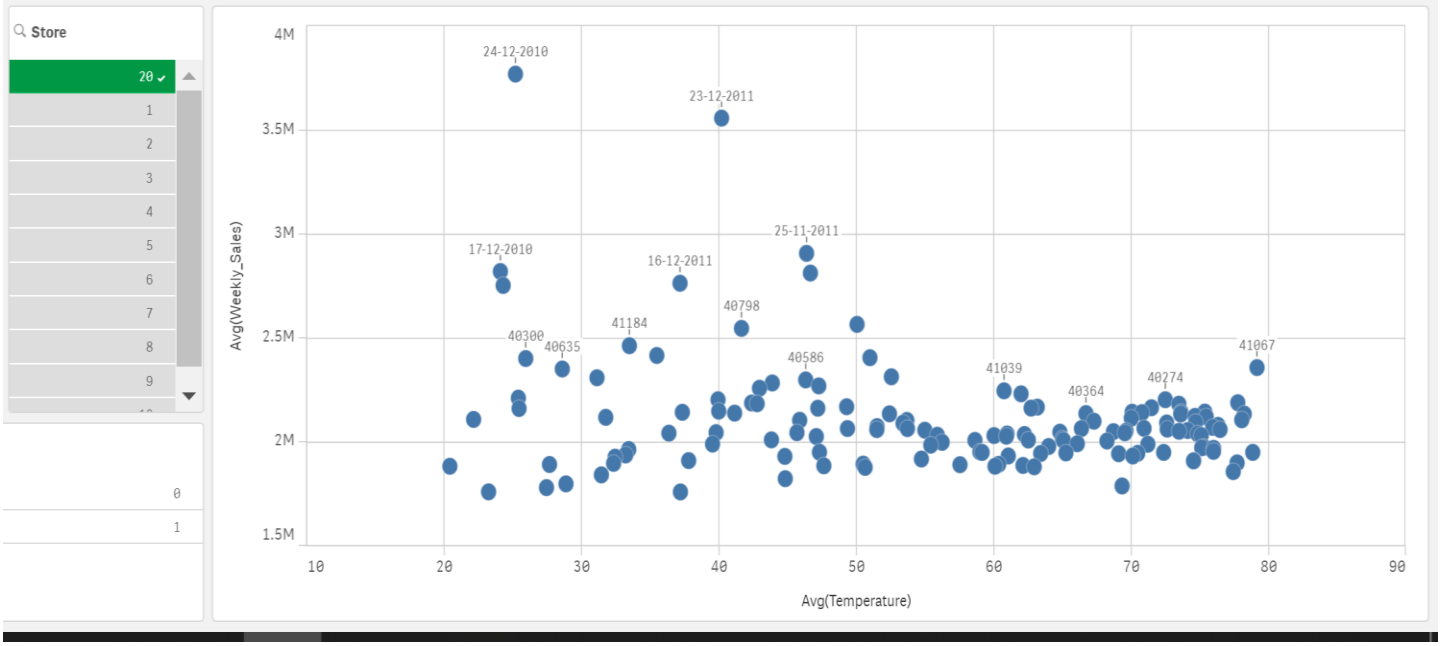


- We observed Standard Deviation and Mean for each store and plot it in box plot. The blue dots represent the outliers representing sales that stores made weekly were above or below the expected average sales.

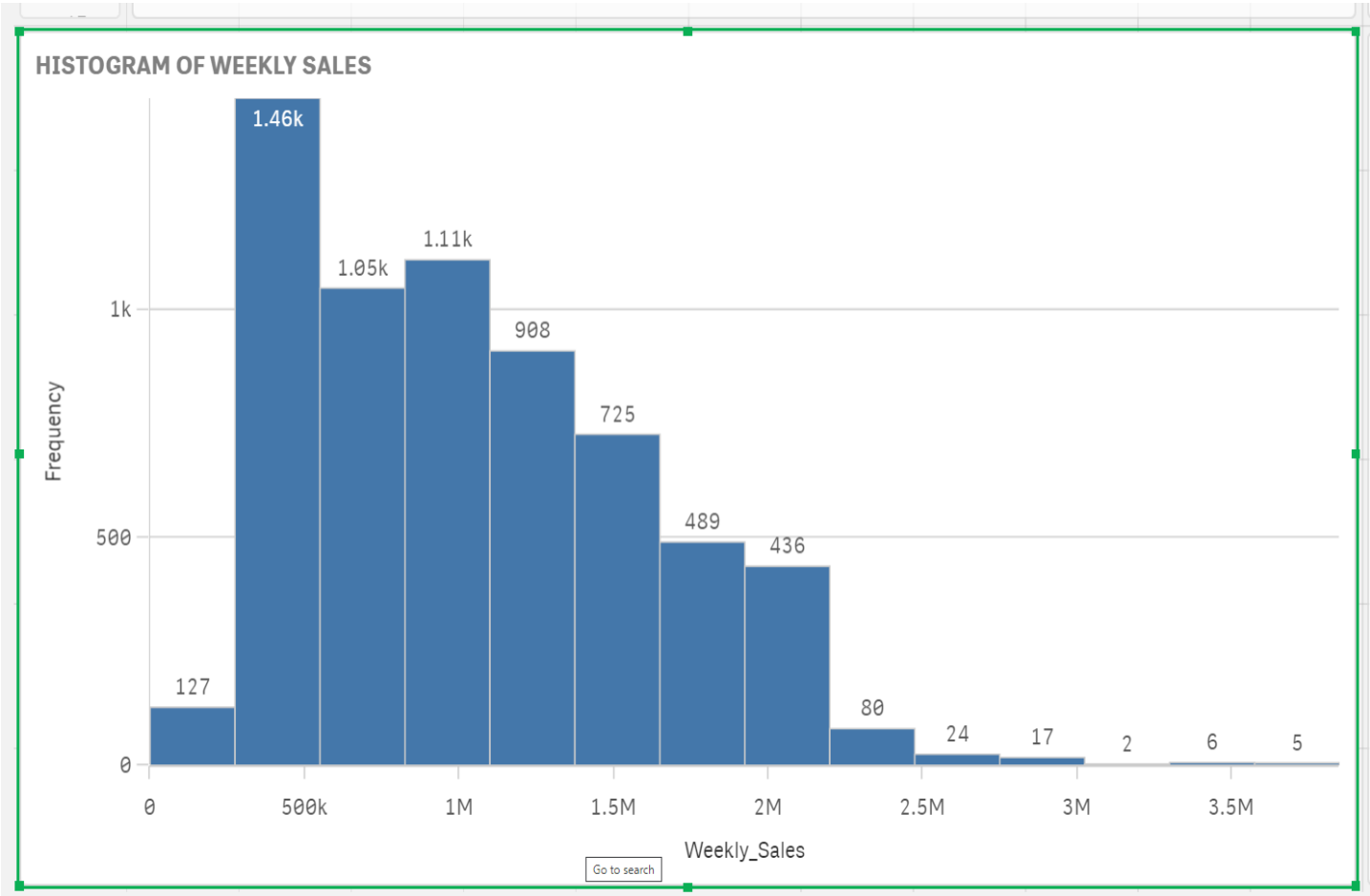
- Store number 14 as compared to other stores has a lot of variation in sales resulting in an average coefficient of standard deviation to mean as 15.

- Store number 42 has more consistent sales resulting in coefficient of variation of 9, lower than average coefficient of variation of 15.





- It is evident from the above charts that both store 20 and 14 respectively is experiencing outlier on sales during the holiday weeks of Christmas and Thanksgiving.



- Through histogram we observed right skewed distribution.

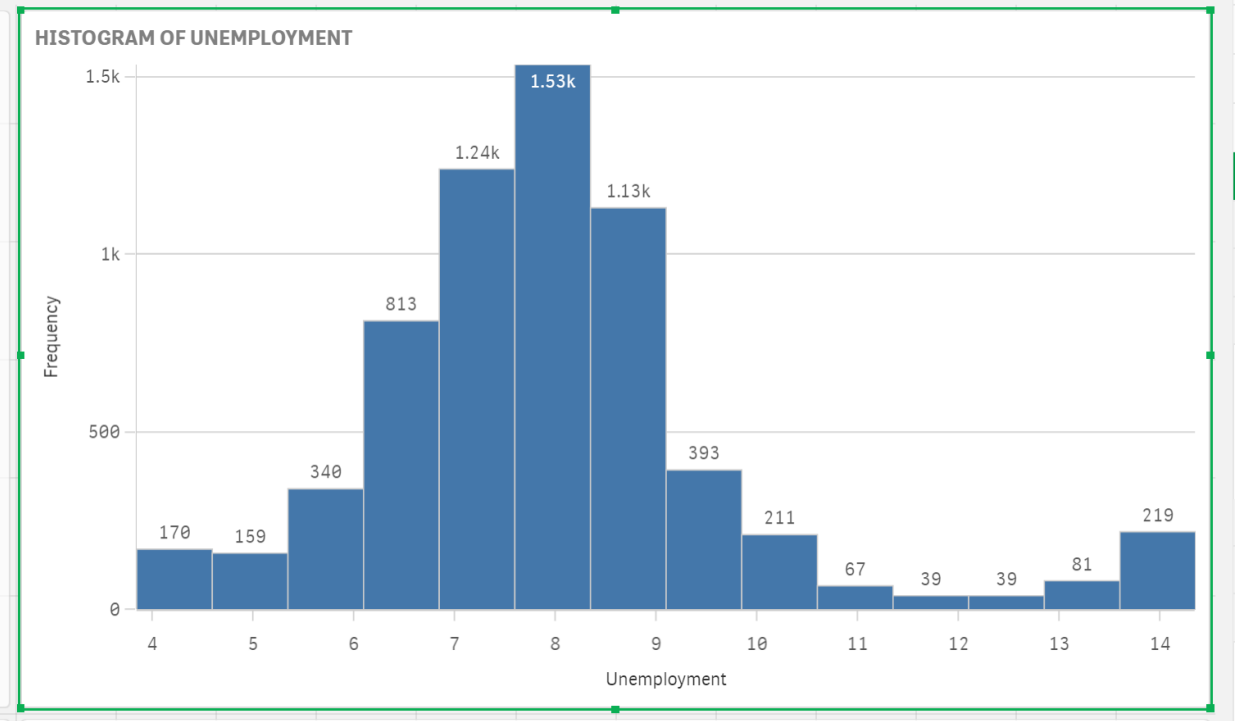
- Stores are making an average $1.05 million around 1100 times and around $500k ,1460 times.

- Therefore, in a more generalized view a store is likely to make an average between 500k to 1.05 million most of the weeks. Based on this we predicted if Walmart wants to open a new store the average weekly sales is likely to be around the range of 500k to 1.05 million in $.

  We considered $3M or more a week is “good week” and less than $500k is a “bad week”.

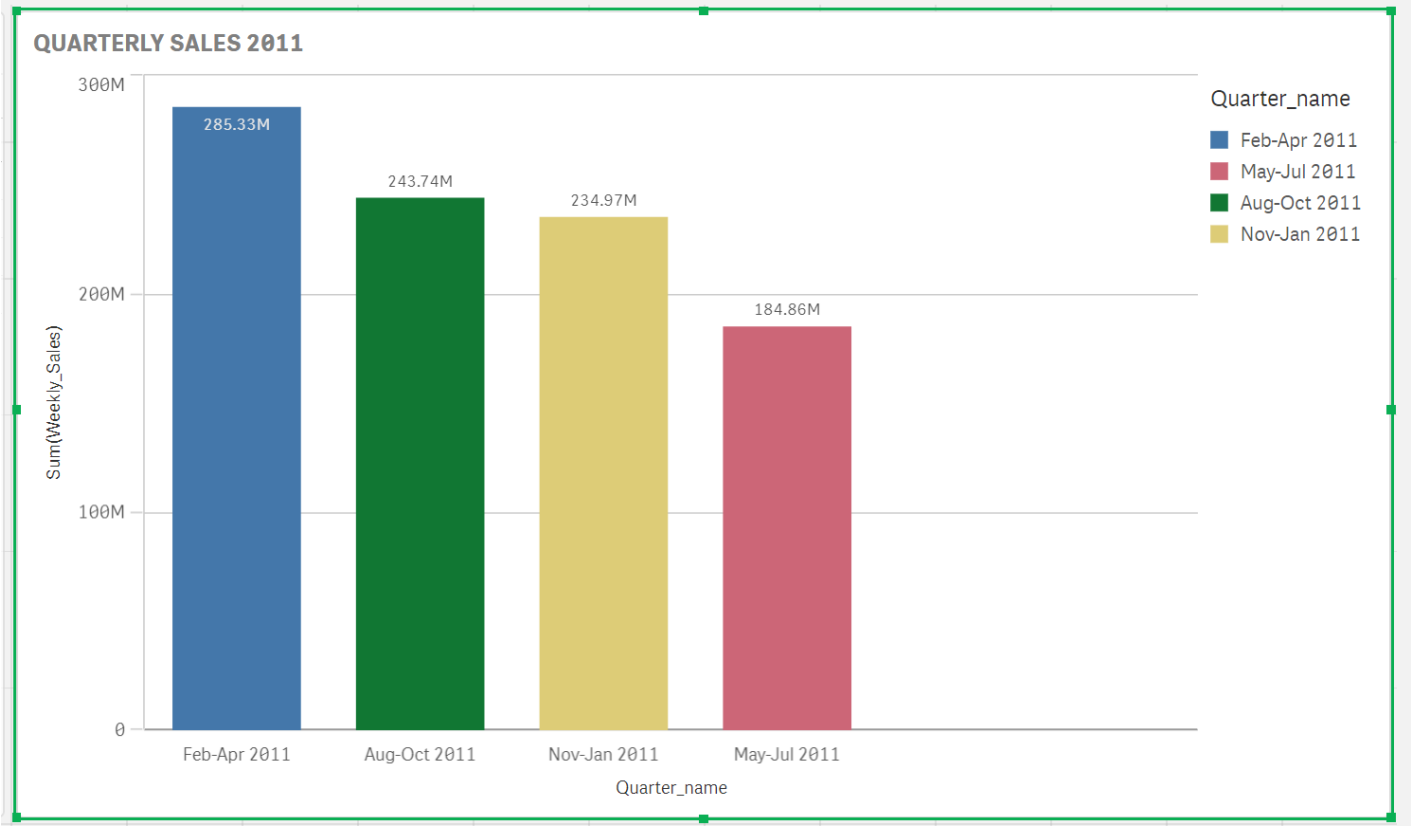
- We observed 0.20% times the stores made above 3M $ in sales (good week sales). We observe that the stores made sales less than 500k, 1.97% of the times in a week (bad sales week)

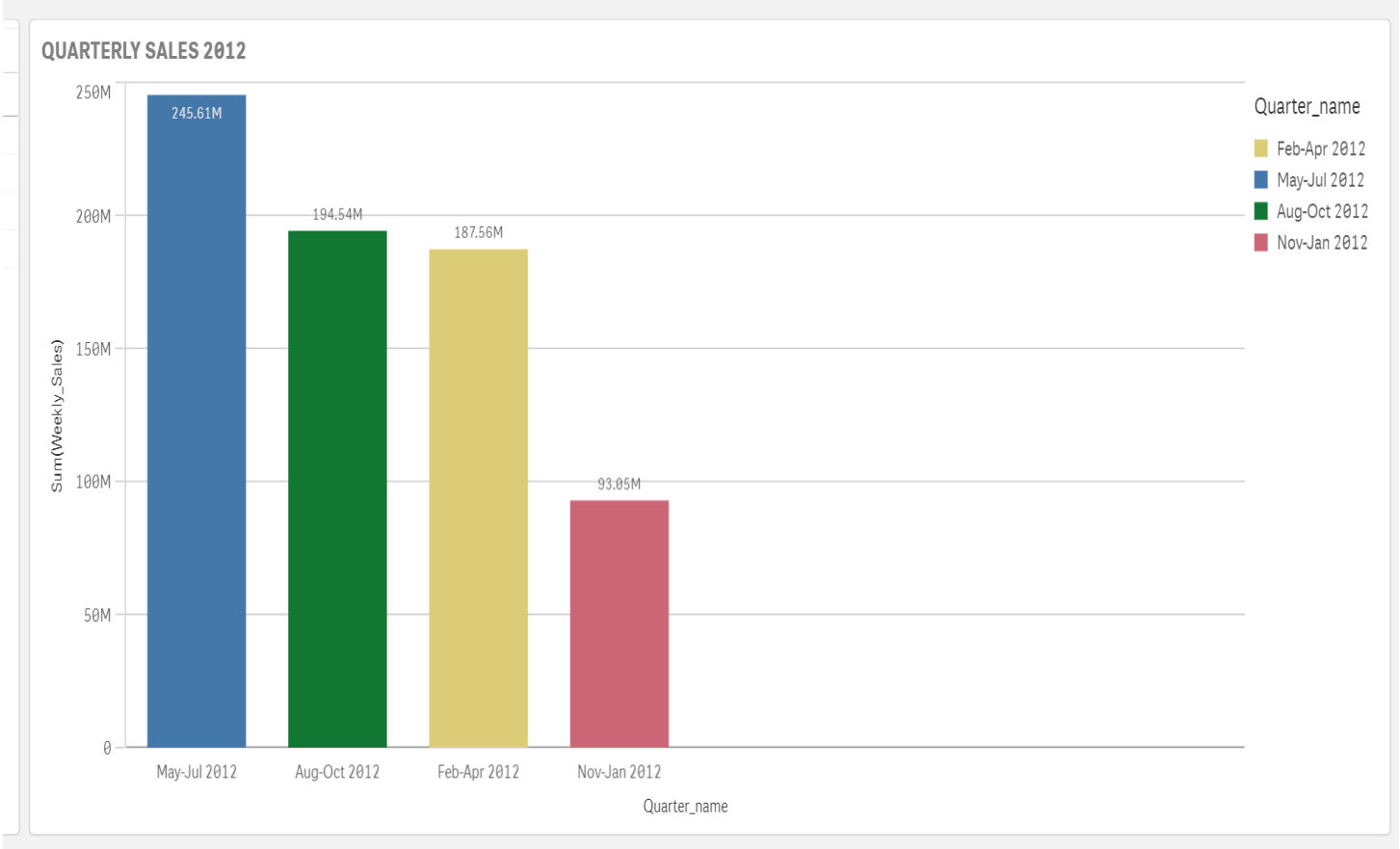
  This gives a ratio of bad week sales to good week sales of 9.85 times. Bad week sales are 9 times likelier to happen than good week sales. Efforts are needed to improve the ratio.



- We observe normal distribution curve with the unemployment rates.

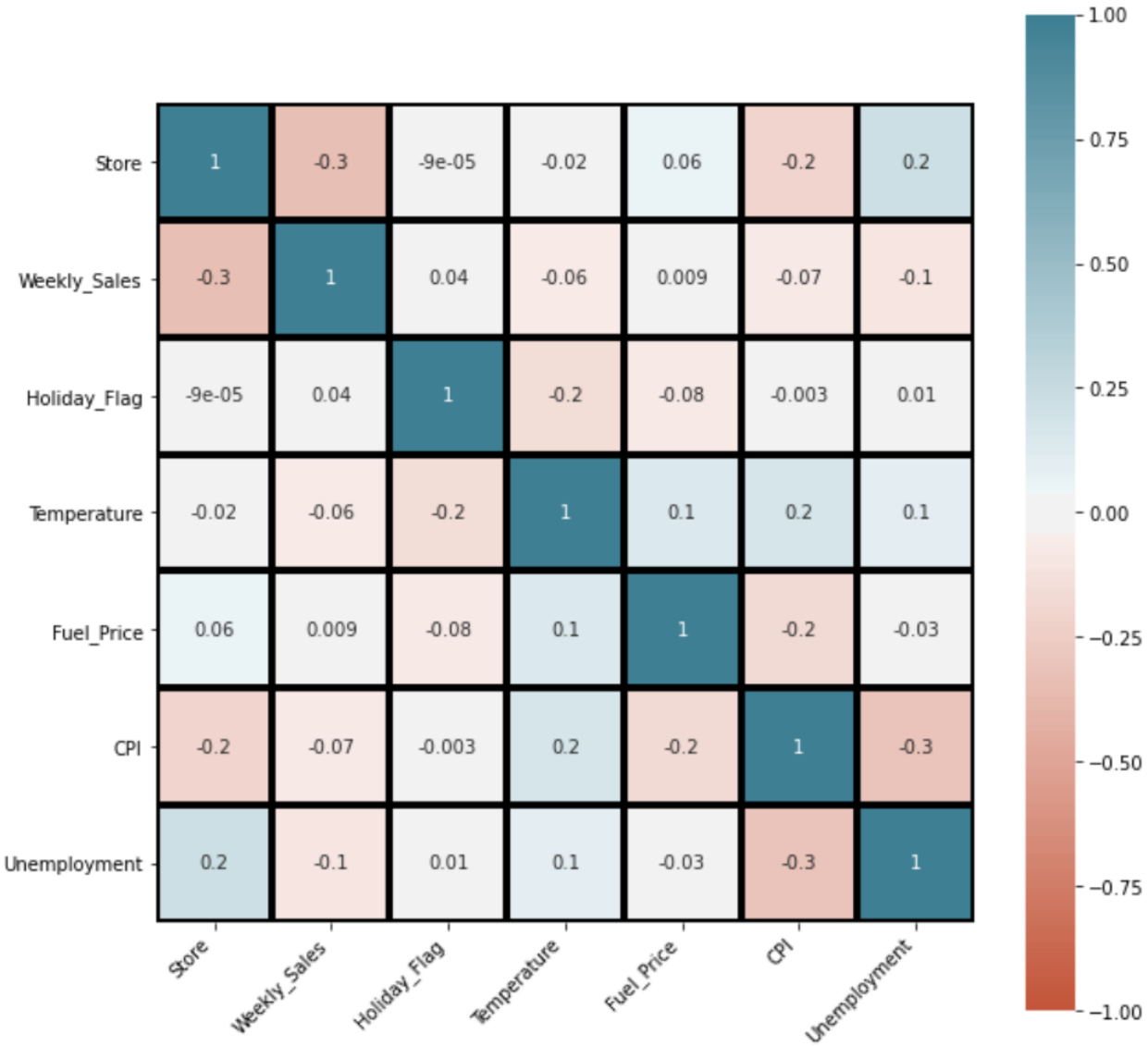
- We concluded that if Walmart wants to go with business expansion this might be a good time because an important macroeconomic factor such as unemployment rate is ranging between 6 to 9%,higher than the average of 3.7 %(national average) which gives Walmart an opportunity to hire from better pool of labour


QUARTERLY SALES 2011
300M
200M
100M
0
Sum(Weekly_Sales)
285.33M
243.74M
234.97M
184.86M
Feb-Apr 2011
Aug-Oct 2011
Nov-Jan 2011
May-Jul 2011
Quarter_name
Quarter_name
Feb-Apr 2011
May-Jul 2011
Aug-Oct 2011
Nov-Jan 2011


QUARTERLY SALES 2012
250M
200M
150M
100M
50M
0
Sum(Weekly_Sales)
245.61M
194.54M
187.56M
93.05M
May-Jul 2012
Aug-Oct 2012
Feb-Apr 2012
Nov-Jan 2012
Quarter_name
Quarter_name
Feb-Apr 2012
May-Jul 2012
Aug-Oct 2012
Nov-Jan 2012

All the three charts represent quarterly sales for the year 2010,2011 and 2012, respectively.
We can see that the highest sales per quarter is not limited to any specific quarter of the year.

For 2010 quarter of Nov-Jan made highest sales, whereas, for 2012 it was the lowest scoring quarter. We can conclude even though major holidays are during Nov-Jan quarter but only for 2010 we see a normal trend of high sales during the holiday season but not conclusive for the rest of 2 years. This abnormality is unexplainable. We are open to suggestions.



- This is a heat map showing correlation(r) between all the variable. A multicollinear heatmap is used to check how much an independent variable influences a dependent variable/target variable which in this case is Weekly sales.
  a) Correlation (r) between Weekly sales and unemployment = -0.1
  b) Correlation (r) between Weekly sales and Fuel prices = 0.009
  c) Correlation (r) between Weekly sales and Temperature =-0.06
  d) Correlation (r) between Weekly sales and CPI = -0.07

- Since the correlation values are very low for all these factors, we can conclude that weekly sales do not get influenced by the above independent variables.

Therefore, when run a linear regression model on these data we get the following summary

t[30]:

OLS Regression Results

Dep. Variable:	Weekly_Sales	**R-squared (uncentered):**	0.786
Model:	OLS	**Adj. R-squared (uncentered):**	0.786
Method:	Least Squares	**F-statistic:**	3928.
Date:	Mon, 16 Nov 2020	**Prob (F-statistic):**	0.00
Time:	21:42:13	**Log-Likelihood:**	-94179.
No. Observations:	6434	**AIC:**	1.884e+05
Df Residuals:	6428	**BIC:**	1.884e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-1.372e+04	545.520	-25.142	0.000	-1.48e+04	-1.26e+04
x2	1.28e+05	2.73e+04	4.696	0.000	7.46e+04	1.81e+05
x3	-951.1634	396.002	-2.402	0.016	-1727.459	-174.868
x4	2.999e+05	1.05e+04	28.628	0.000	2.79e+05	3.2e+05
x5	796.6228	150.042	5.309	0.000	502.490	1090.755
x6	3.156e+04	3375.270	9.350	0.000	2.49e+04	3.82e+04

Omnibus:	195.614	**Durbin-Watson:**	0.126
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	213.289
Skew:	0.445	**Prob(JB):**	4.84e-47
Kurtosis:	3.059	**Cond. No.**	744.

As we observe in the above summary, the value of **r2 = 0.786** which means that only **78.6%** of data points can be fitted and predicted by the regression line.

Therefore, we can conclude that due to lack of correlation between dependent and independent variable and lack of related factors the model is not effective enough to predict weekly sales with the help of linear regression.

DISCLAIMER: The above linear regression model is not the best outcome and is not 100% accurate . The model can be improved upon with the help of certain modification as mentioned below in the conclusion section.

Conclusion

1. While doing analysis and predictive model on data presented, we felt the absence of bulk of data and related variables. Variables such as Department weekly sales , would have given much better information and insight to understand the effect of economic factors such as CPI, unemployment and Fuel Prices .Therefore, we assumed that such data would have helped us making a better predictive model.

2. In addition to that, we relied on average concept. Since aggregate measure didn't help us much to compare result as Non-Holiday weeks were more as compared to Holiday-week. There was bias (huge difference in sales in Non-Holiday vs Holiday week) ,thus to eliminate that average weekly sales was used do a better comparison and understand the sales.