WALMART RETAIL ANALYSIS AND SALES PREDICTION

Chapter 1

Data Analytics:

- Data analytics is the science of analyzing raw data to make conclusions about that information.
- The techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
- Data analytics help a business optimize its performance.

Applications:

- **Security:** Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas. In a few major cities like Los Angeles and Chicago, historical and geographical data has been used to isolate specific areas where crime rates could surge. On that basis, while arrests could not be made on a whim, police patrols could be increased. Thus, using applications of data analytics, crime rates dropped in these areas.
- **Transportation:** Data analytics can be used to revolutionize transportation. It can be used especially in areas where you need to transport a large number of people to a specific area and require seamless transportation. This data analytical technique was applied in the London Olympics a few years ago. For this event, around 18 million journeys had to be made. So, the train operators and TFL were able to use data from similar events, predict the number of people who would travel, and then ensure that the transportation was kept smooth.
- **Risk detection:** One of the first data analytics applications may have been in the discovery of fraud. Many organizations were struggling under debt, and they wanted a solution to this problem. They already had enough customer data in their hands, and so, they applied data analytics. They used ‘divide and conquer’ policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting. Eventually, it led to lower risks and fraud.
- **Risk Management:** Risk management is an essential aspect in the world of insurance. While a person is being insured, there is a lot of data analytics that goes on during the process. The risk involved while insuring the person is based on several data like actuarial data and claims data, and the analysis of them helps insurance companies to realize the risk.
- **Delivery:** Several top logistic companies like DHL and FedEx are using data analysis to examine collected data and improve their overall efficiency. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well

as the most cost-efficient transport means. Using GPS and accumulating data from the GPS gives them a huge advantage in data analytics.

- **Fast internet allocation:** While it might seem that allocating fast internet in every area makes a city 'Smart', in reality, it is more important to engage in smart allocation. This smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause. It is also important to shift the data allocation based on timing and priority. It is assumed that financial and commercial areas require the most bandwidth during weekdays, while residential areas require it during the weekends. But the situation is much more complex. Data analytics can solve it. For example, using applications of data analysis, a community can draw the attention of high-tech industries and in such cases, higher bandwidth will be required in such areas.
- **Reasonable Expenditure**: When one is building Smart cities, it becomes difficult to plan it out in the right way. Remodelling of the landmark or making any change would incur large amounts of expenditure, which might eventually turn out to be a waste. Data analytics can be used in such cases. With data analytics, it will become easier to direct the tax money in a cost-efficient way to build the right infrastructure and reduce expenditure.
- **Interaction with customers:** In insurance, there should be a healthy relationship between the claims handlers and customers. Hence, to improve their services, many insurance companies often use customer surveys to collect data. Since insurance companies target a diverse group of people, each demographic has their own preference when it comes to communication.
- **Planning of cities:** One of the untapped disciplines where data analysis can really grow is city planning. While many city planners might be hesitant towards using data analysis in their favour, it only results in faulty cities riddled congestion. Using data analysis would help in bettering accessibility and minimizing overloading in the city.
- **Healthcare:** While medicine has come a long way since ancient times and is ever-improving, it remains a costly affair. Many hospitals are struggling with the cost pressures that modern healthcare has come with, which includes the use of sophisticated machinery, medicines, etc. But now, with the help of data analytics applications, healthcare facilities can track the treatment of patients and patient flow as well as how equipment are being used in hospitals. It has been estimated that there can be a 1% efficiency gain achieved if data analytics became an integral part of healthcare, which will translate to more than $63 billion in healthcare services.
- **For Travelling:** If you ever thought travelling is a hassle, then data analytics is here to save you. Data analysis can use data that shows the desires and preferences of different customers from social media and helps in optimizing the buying experience of travellers. It will also help companies customize their own packages and offer and hence boost more personalized travel recommendations with the help data collected from social media.

- **Managing Energy:** Many firms engaging with energy management are making use of applications of data analytics to help them in areas like smart-grid management, optimization of energy, energy distribution, and automation building for other utility-based companies. How does data analytics help here? Well, it helps by focusing on controlling and monitoring of a dispatch crew, network devices, and management of service outages. Since utilities integrate about millions of data points within the network performance, engineers can use data analytics to help them monitor the entire network.
- **Internet searching:** When you use Google, you are using one of their many data analytics applications employed by the company. Most search engines like Google, Bing, Yahoo, AOL, Duckduckgo, etc. use data analytics. These search engines use different algorithms to deliver the best result for a search query, and they do so within a few milliseconds. Google is said to process about 20 petabytes of data every day.
- **Digital advertisement:** Data analytics has revolutionized digital advertising, as well. Digital billboards in cities as well as banners on websites, that is, most of the advertisement sources nowadays use data analytics using data algorithms. It is one of the reasons why digital advertisements are getting more CTRs than traditional advertising techniques. The target of digital advertising nowadays is focused on the analysis of the past behaviour of the user.

Challenges of Implementation:

- **Need For Synchronization Across Disparate Data Sources:** As data sets are becoming bigger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is overlooked, it will create gaps and lead to wrong messages and insights.
- **Acute Shortage Of Professionals Who Understand Big Data Analysis:** The analysis of data is important to make this voluminous amount of data being produced in every minute, useful. With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been created in the market. It is important for business organizations to hire a data scientist having skills that are varied as the job of a data scientist is multidisciplinary. Another major challenge faced by businesses is the shortage of professionals who understand Big Data analysis. There is a sharp shortage of data scientists in comparison to the massive amount of data being produced.
- **Getting Meaningful Insights Through The Use Of Big Data Analytics:** It is imperative for business organizations to gain important insights from Big Data analytics, and also it is important that only the relevant department has access to this information. A big challenge faced by the companies in the Big Data analytics is mending this wide gap in an effective manner.

- **Getting Voluminous Data Into The Big Data Platform:** It is hardly surprising that data is growing with every passing day. This simply indicates that business organizations need to handle a large amount of data on daily basis. The amount and variety of data available these days can overwhelm any data engineer and that is why it is considered vital to make data accessibility easy and convenient for brand owners and managers.
- **Uncertainty Of Data Management Landscape:** With the rise of Big Data, new technologies and companies are being developed every day. However, a big challenge faced by the companies in the Big Data analytics is to find out which technology will be best suited to them without the introduction of new problems and potential risks.
- **Data Storage And Quality:** Business organizations are growing at a rapid pace. With the tremendous growth of the companies and large business organizations, increases the amount of data produced. The storage of this massive amount of data is becoming a real challenge for everyone. Popular data storage options like data lakes/ warehouses are commonly used to gather and store large quantities of unstructured and structured data in its native format. The real problem arises when a data lakes/ warehouse try to combine unstructured and inconsistent data from diverse sources, it encounters errors. Missing data, inconsistent data, logic conflicts, and duplicates data all result in data quality challenges.
- **Security And Privacy Of Data:** Once business enterprises discover how to use Big Data, it brings them a wide range of possibilities and opportunities. However, it also involves the potential risks associated with big data when it comes to the privacy and the security of the data. The Big Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, the rise of voluminous amount of data increases privacy and security concerns.

Chapter 2

1. Linear Regression:

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

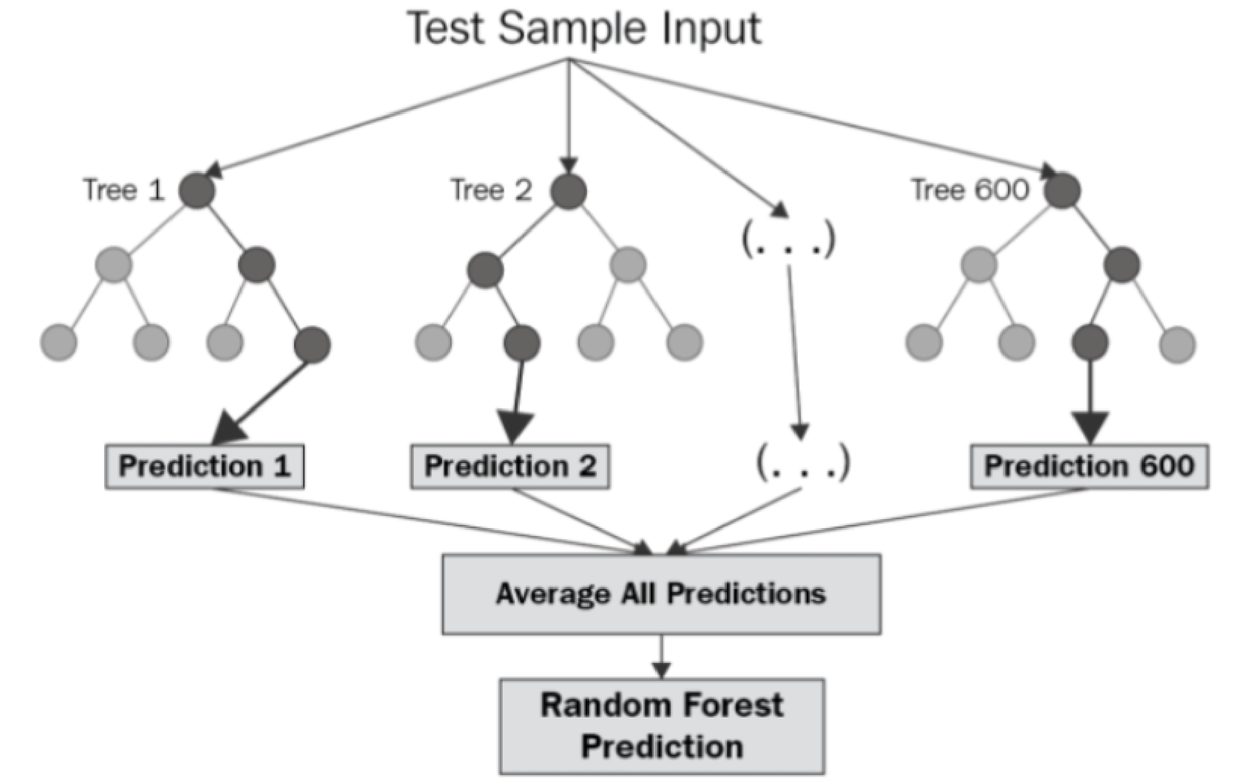
The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = B0 + B1*x$$

2. Random Forest Regressor:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



The diagram above shows the structure of a Random Forest. You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

Chapter 3

Data set Description:

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 some times, 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, Labour 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 **

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the file WalmartStoresales. 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

Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13

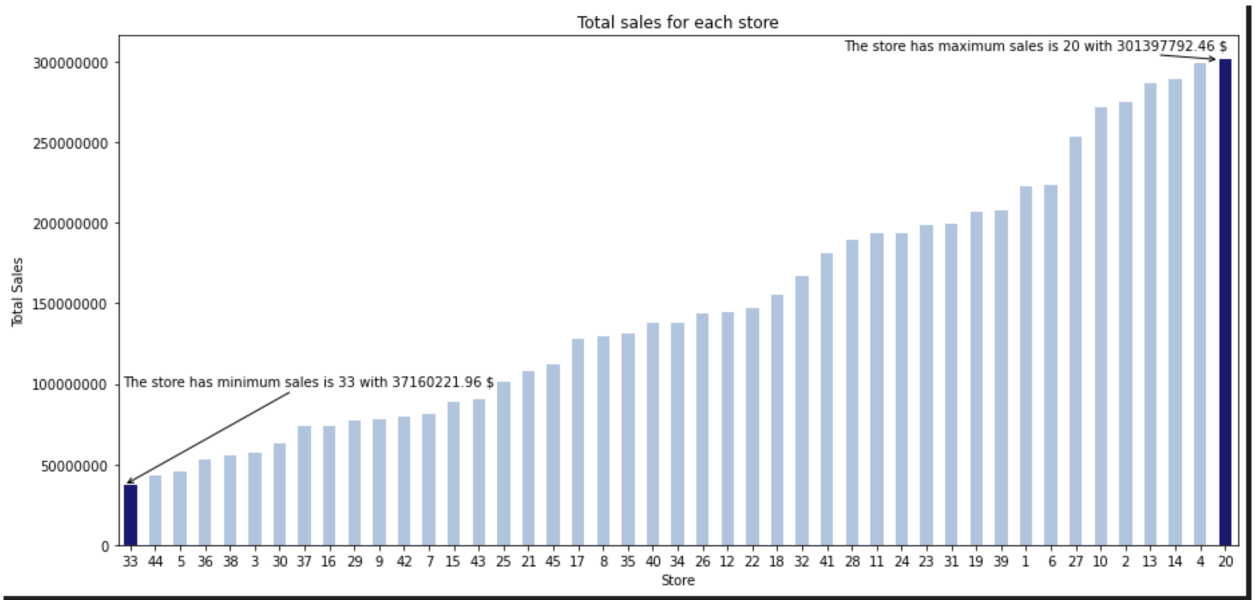
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13

Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

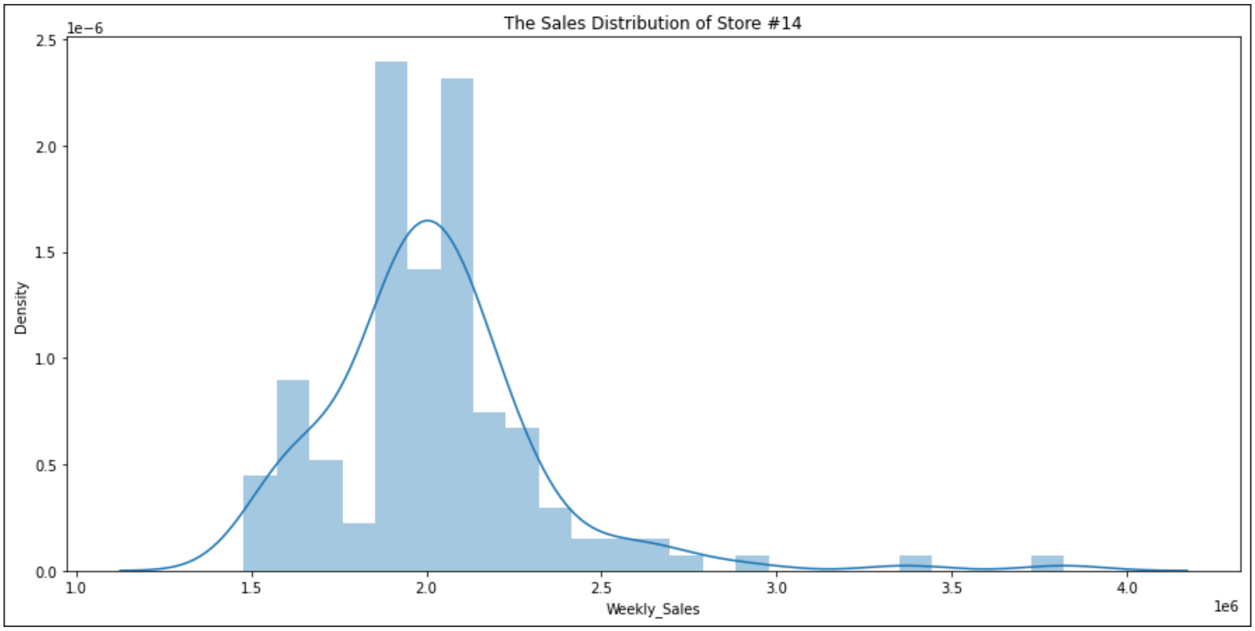
Chapter 4

Data Visualization

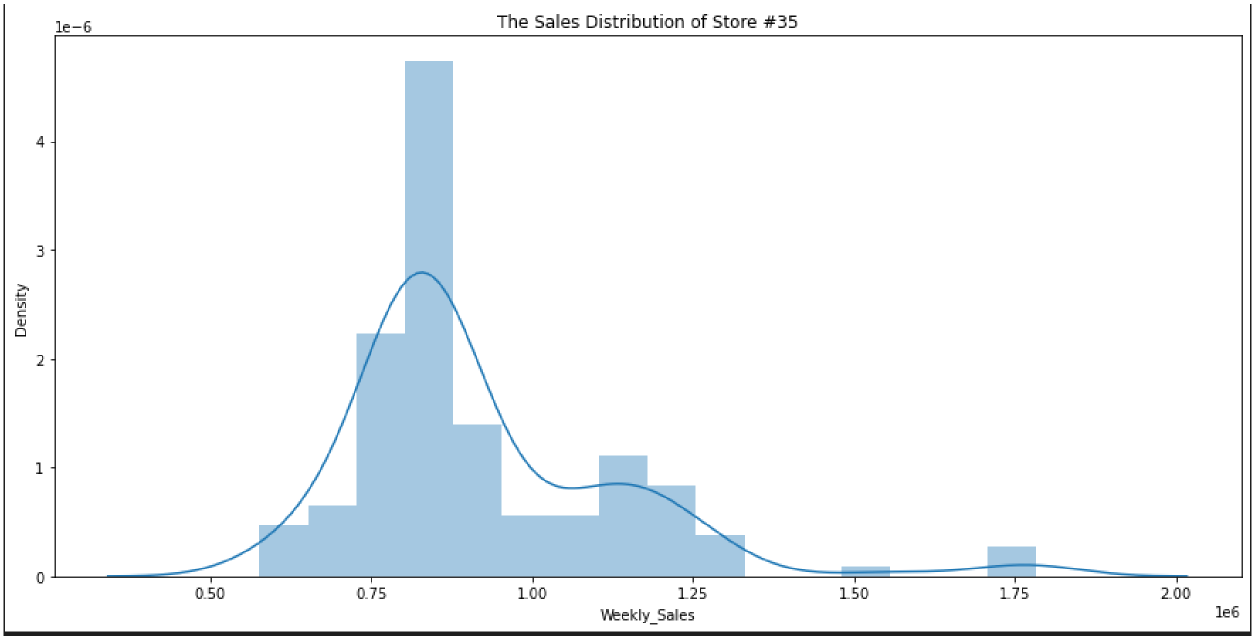
1. Which store has minimum and maximum sales?



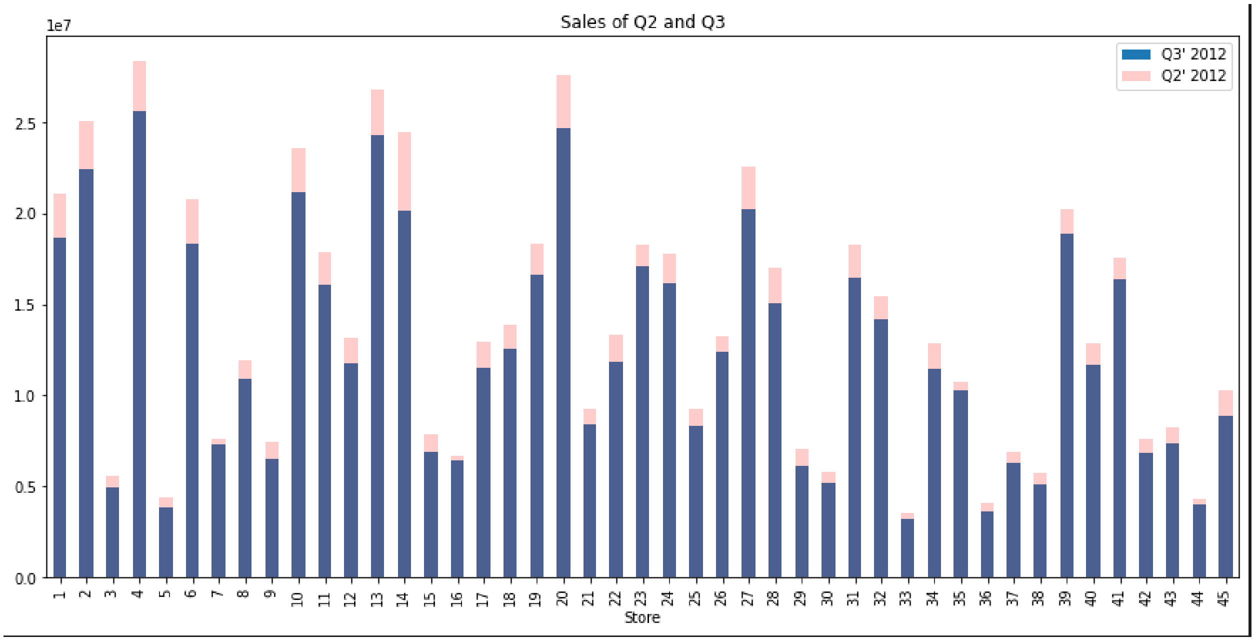
2. Which store has maximum standard deviation i.e., the sales vary a lot. Also, what is the coefficient of mean to standard deviation?



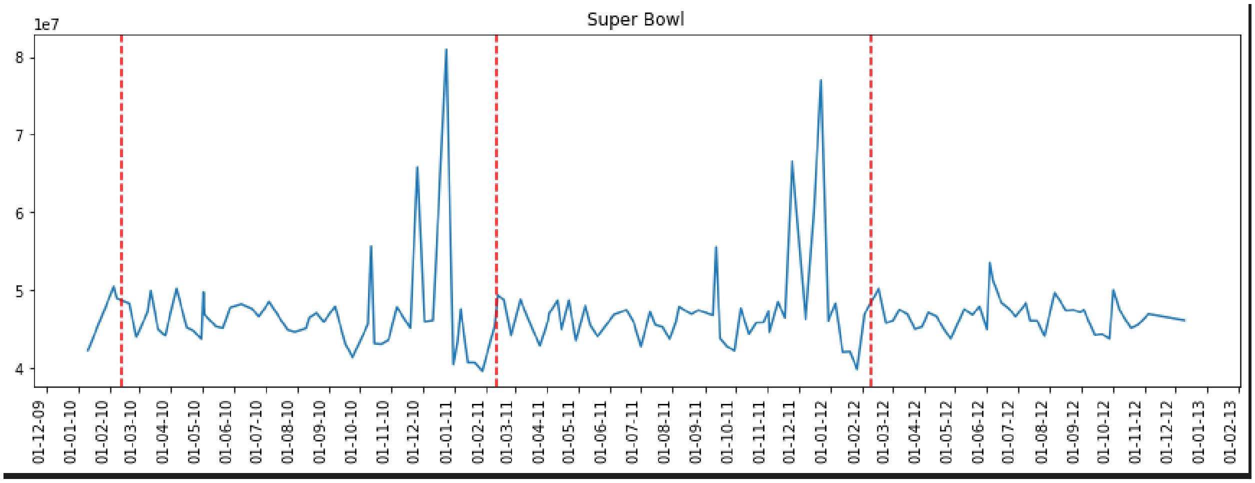
3. Distribution of store that has maximum coefficient of mean to standard deviation.

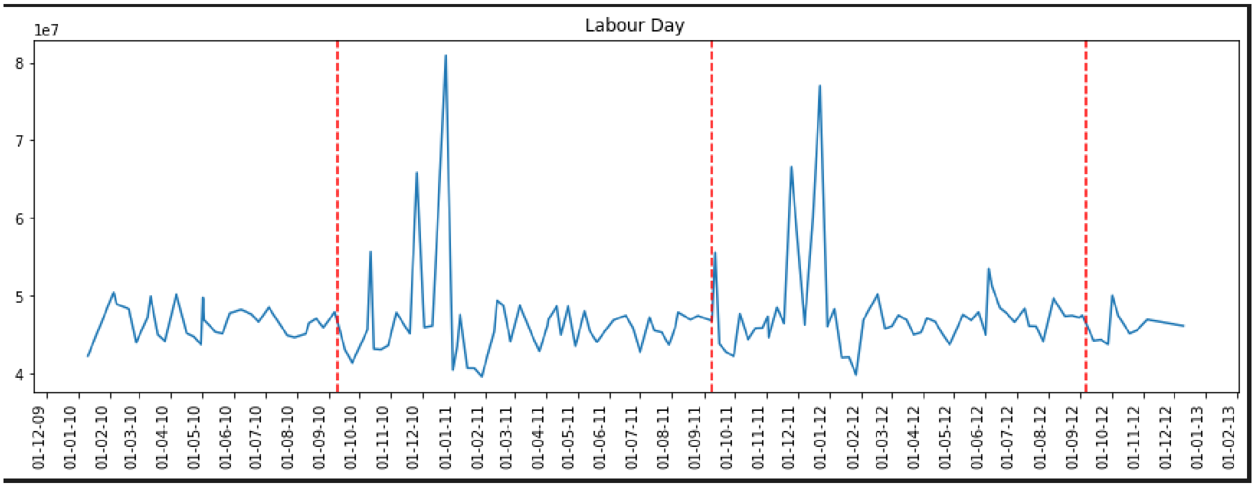


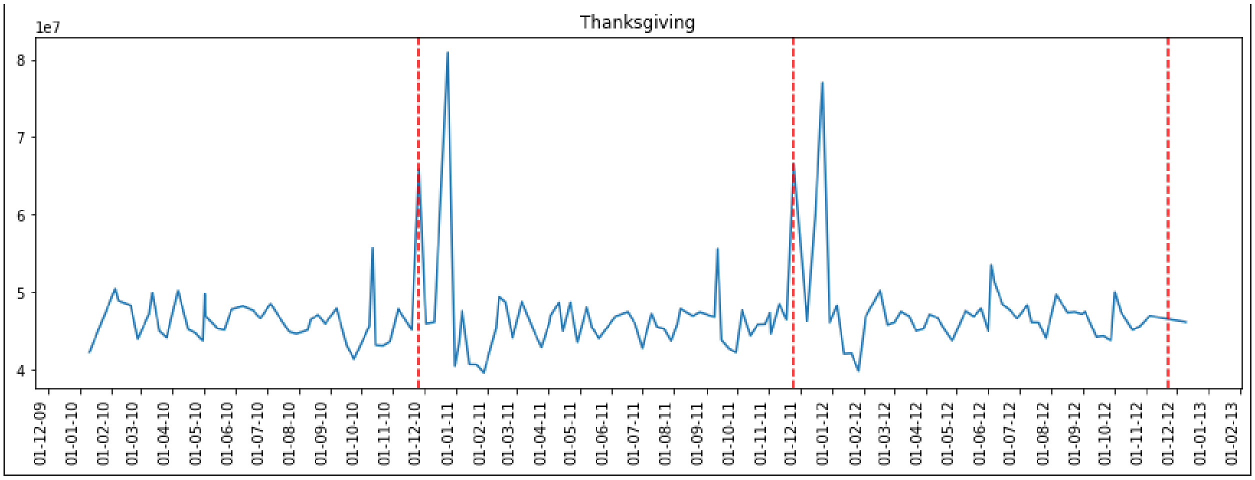
4. Which store has good quarterly growth rate in 2012?

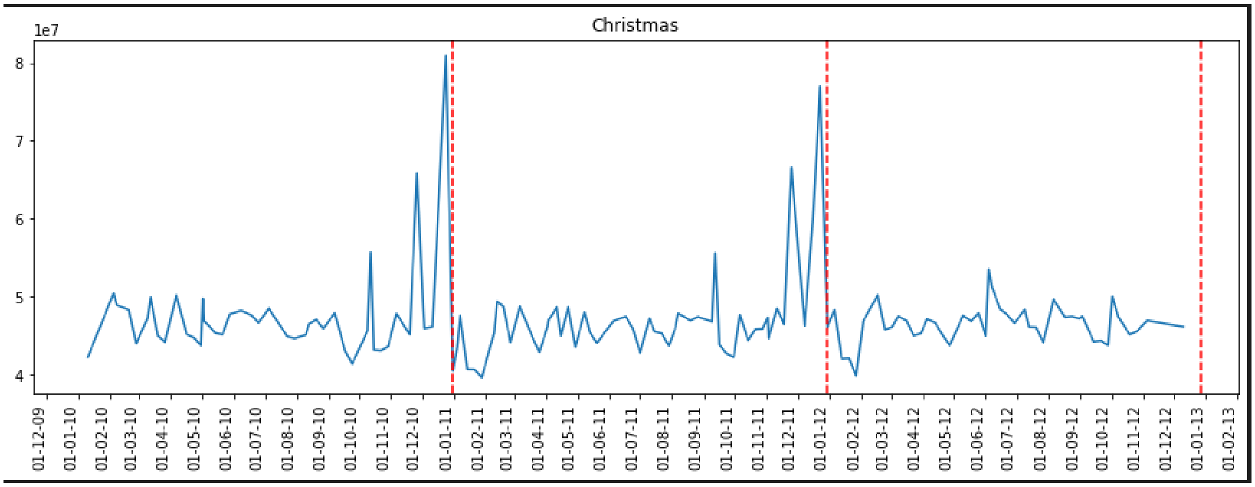


5. Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together

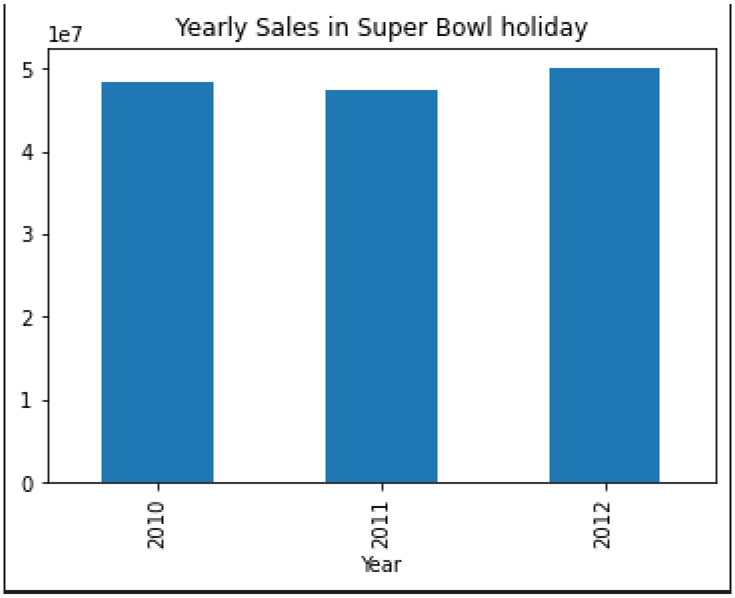


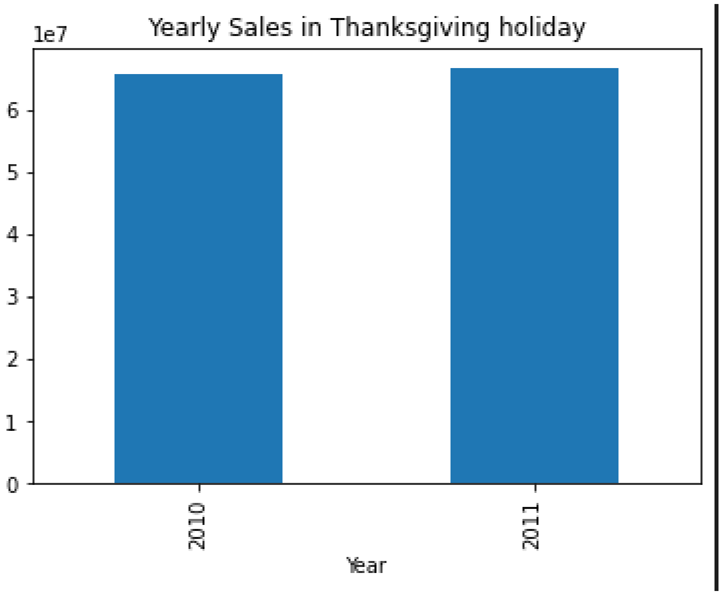


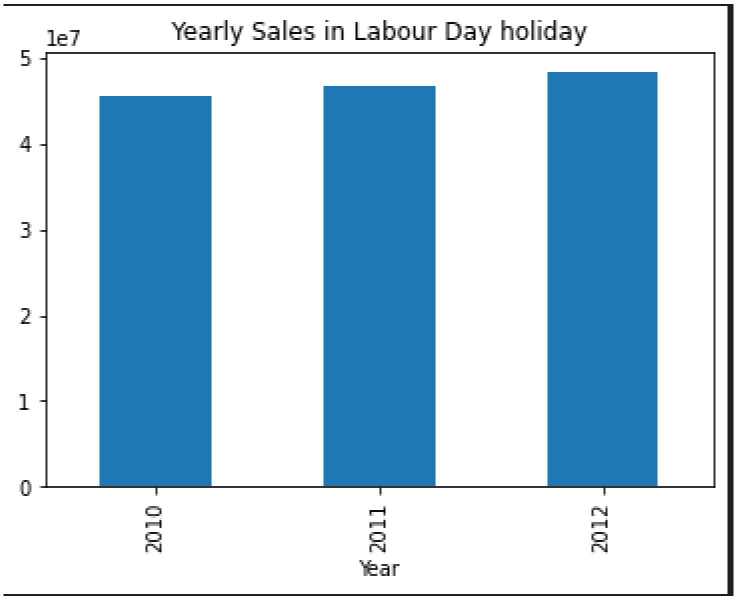


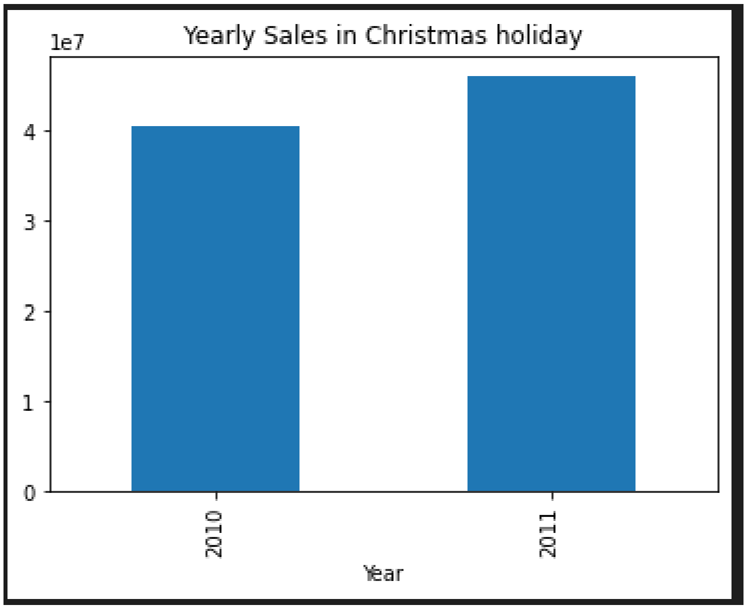


6. The sales increased during thanksgiving. And the sales decreased during christmas.

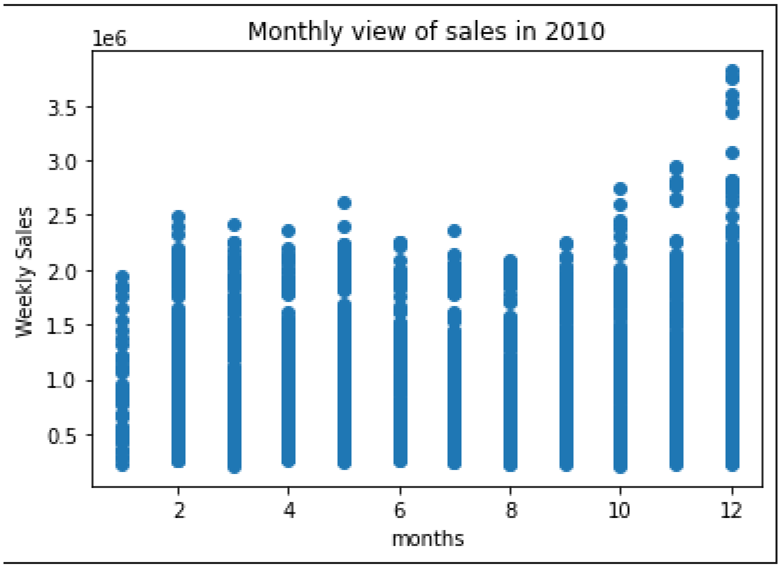

Yearly Sales in Super Bowl holiday
1e7
5
4
3
2
1
0
2010
2011
2012
Year

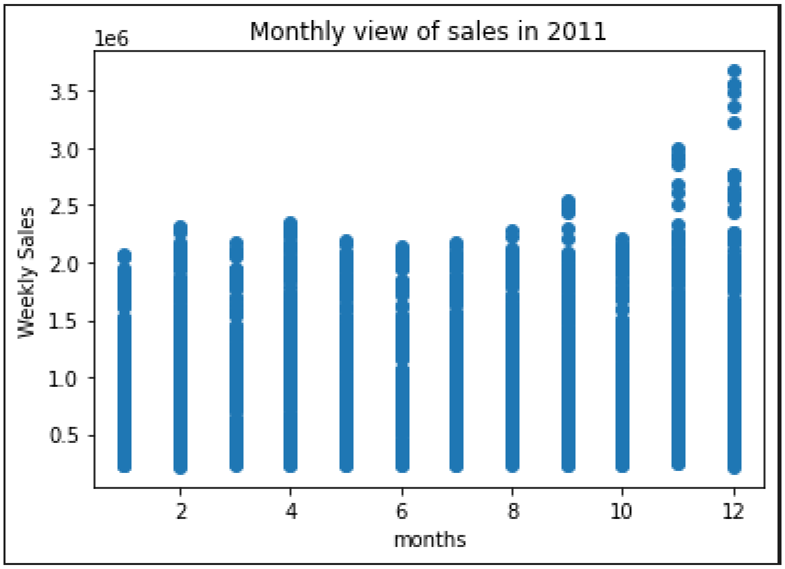

Yearly Sales in Thanksgiving holiday
1e7
6
5
4
3
2
1
0
2010
2011
Year

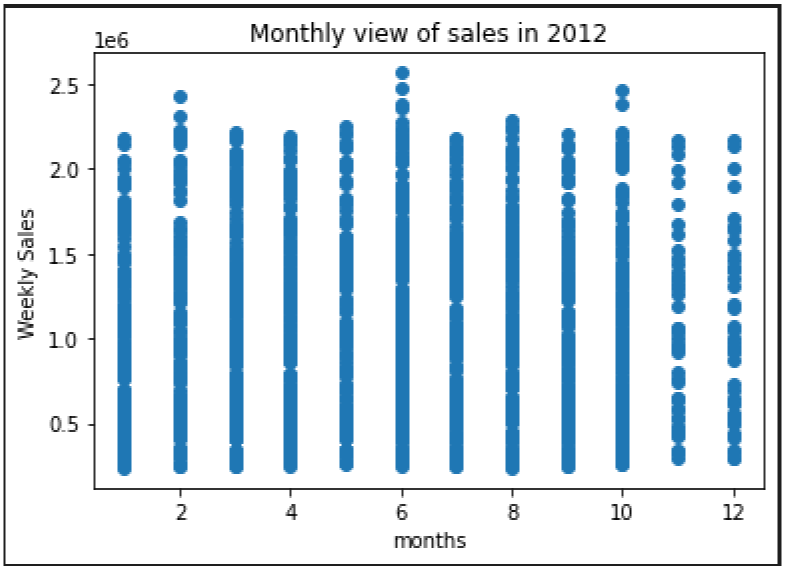




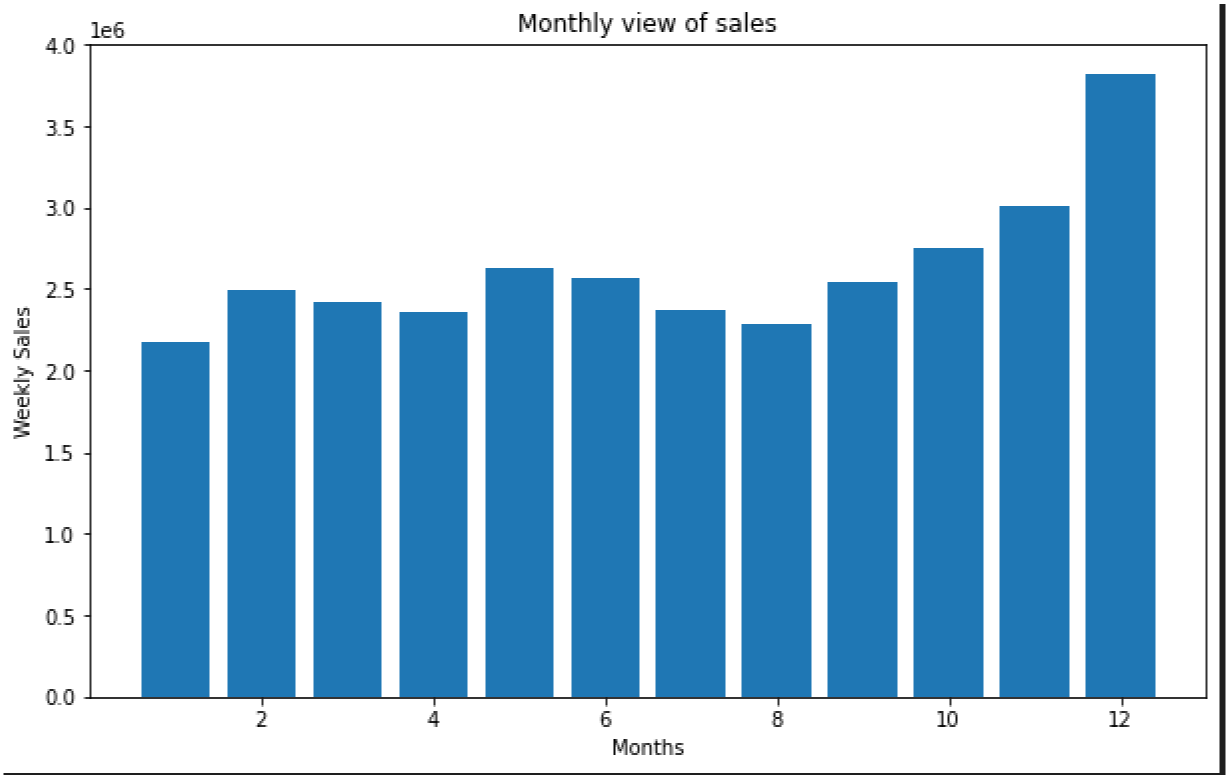
7. Provide a monthly and semester view of sales in units and give insights


Monthly view of sales in 2010
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Weekly Sales
2
4
6
8
10
12
months

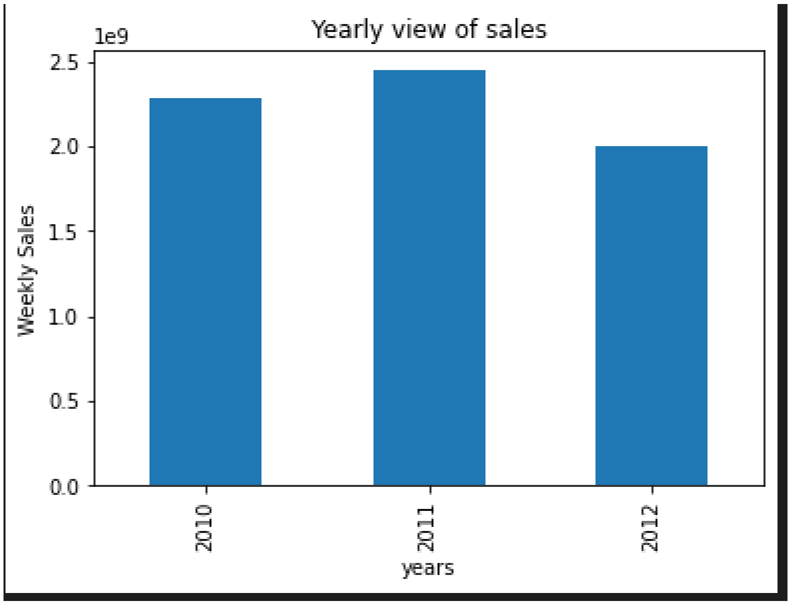

Monthly view of sales in 2011
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Weekly Sales
2
4
6
8
10
12
months



8. Monthly view of sales



9. Yearly view of sales:



Chapter 5

Future Scope:

1. We will test different algorithms such as XGBRegressor and ExtraTreesRegressor and compare accuracy with the highest accuracy we could achieve.
2. As here available data is less, so the loss difference is not extraordinary . But in large datasets of sizes in Gigabytes and Terabytes this trick of simple averaging may reduce the loss to a great extent.

Conclusion:

1. The conclusion is that Random Forest Regressor and Linear regression have the accuracy of 12.6 and 94.2 respectively

Chapter 6

References:

1. https://www.upgrad.com/blog/data-analytics-applications/
2. https://www.investopedia.com/terms/d/data-analytics.asp

3. https://www.kaggle.com/zarahshibli/retail-analysis-with-walmart-data/comments