*ITCS 6162*

*Knowledge Discovery in Databases*

*PROJECT DELIVERABLE 3*

ROSSMAN STORE DATA

***CRISP-DM PROCESS***

Team Members:

Urvi Jayesh Gada

Sunidhi Kabra

Darshak Mehta

Nikita Nalawade

Ravil Bikmetov

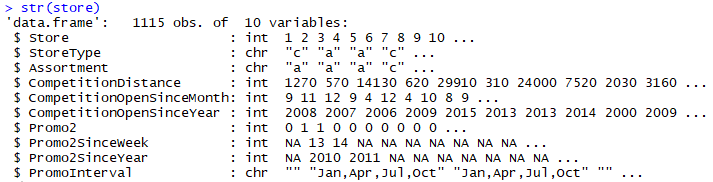
***TASK***

The Rossman Dataset provides data about the historical sales for 1,115 Rossman stores. Based on the data that has been provided, our task is to forecast the “Sales” column for the test set. They have also mentioned that some of these stored were closed temporarily for refurbishment.

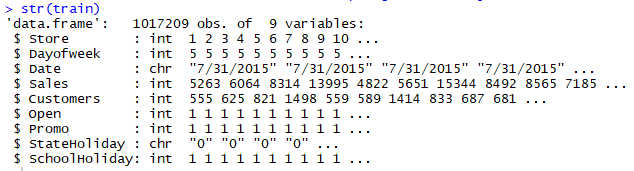
***THE DATASET***

We have been provided with three files which are the test, train and the stores dataset. There are a few columns which are common to all the three datasets while most of the fields are different.

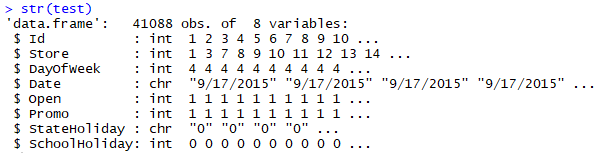
* Store Dataset



* Train Dataset



* Test Dataset



***THE CRISP-DM PROCESS***

CRISP-DM stands for **Cr**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining which follows an iterative and adaptive live cycle consisting of 6 phases.

The 6 phases of the CRISP-DM process are as follows:

1. **Business/Research Understanding Phase**

In this phase, we define the project requirements and objectives and translate them into data mining problem definitions. We also prepare the preliminary strategy to meet objectives.

1. **Data Understanding Phase**

In this phase, we collect the data and perform EDA i.e. Exploratory Data Analysis which helps in assessing the data quality. In this phase, we decide which subset of data do we wish to work with.

1. **Data Preparation Phase**

In this phase, we prepare for modelling and select the cases and variables that are appropriate for analysis. Also, we clean and prepare the data and perform transformations on certain variables so that the data is ready for the modelling tools.

1. **Modelling Phase**

In this phase, we select and apply one or more modelling techniques and calibrate the model settings to optimize the results. In some cases, additional data preparation may be required for supporting a technique.

1. **Evaluation Phase**

In this phase, we evaluate the models for effectiveness and determine whether the objectives that were defined have been achieved. We also check if some important facet of the problem has been sufficiently accounted for and make decisions regarding the data mining results before we deploy them.

1. **Deployment Phase**

In this phase, we make use of the models that have been created. A simple deployment would be to generate a report while a more complex example would be to implement parallel data mining algorithms in another department.

***DATA PREPROCESSING***

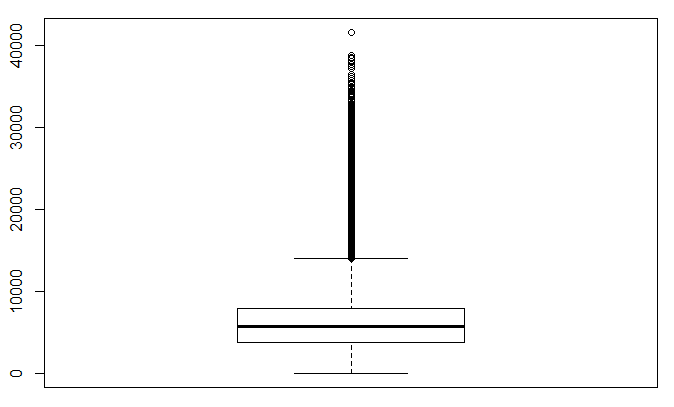
* ***What is Data Preprocessing?***

It is the technique in data mining that involves transforming the raw data into an understandable form. Most of the real-world data is incomplete, inconsistent and lacking certain behaviors or trends. To resolve these issues, data preprocessing must be carried out before going further with any data mining algorithms.

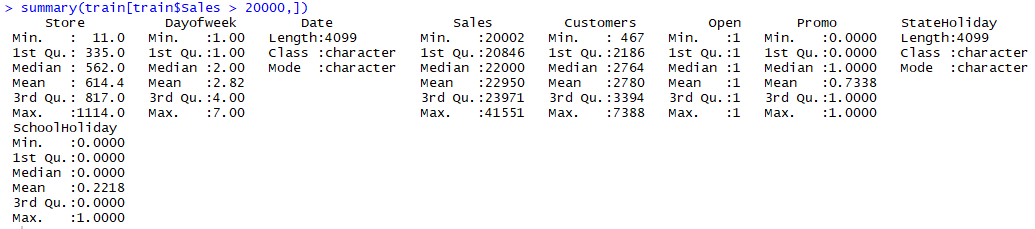
* ***Data Preprocessing on the Rossman Store Data.***

1. **Determining the Outliers in Sales**

The sales variable which is present in the train dataset is the variable that is to be predicted. Thus, the first step that we implemented was to check for the outliers in this variable. The box plot of the sales variable is as follows:



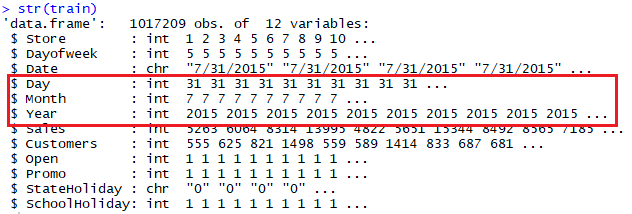
At the first instance it may appear that the values which are greater than 20000 are outliers. However, before assuming that those may be outliers, we should check for the occurrences of those values in the dataset. This is done by taking the summary of sales variable for values greater than 20000. The results obtained are as follows:



As we see that there are a lot of stores that do have extremely high sales, we conclude that sales greater than 20000 are not outliers after all.

1. **Working on the Date field in Train and Test Dataset**

The date field is expressed as a character in the form of MM/DD/YYYY. Since we know that the sales will greatly be affected by the days and months we separate this field in three variables “Day”, “Month” and “Year” and convert it to integer data type.

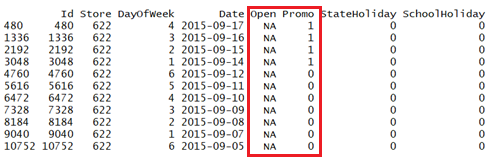


1. **Not considering the Customers field**

The customer variable doesn’t serve any purpose. Also, it isn’t used in the test set because the task is not to predict the number of customers that walk in the store but the sales of the store. So, fitting a model to predict the number of customers based on the given variables is of no use. Therefore, the customer variable isn’t used in the train dataset.

1. **Missing Values in the Open variable of the Test Dataset.**

When we check for NA values in the Open variable, we get 11 values, but they are all for the same store ‘622’. However, we also see that for the first 4 entries we have a promotional event which is carried out. Also, these NA values are only for the working days i.e. Monday to Thursday. From the dataset, for a Sunday i.e. DayofWeek = 7, the stores are closed, and the Promo field is also 0. Thus, we can assume that the store was open for the given days and thus we set this value to Open = 1 for all the records.

****

1. **Not considering the Promo2SinceWeek, Promo2SinceYear and PromoInterval**

The Promo2SinceWeek, Promo2SinceYears, PromoInterval have many values which are null. Therefore, these three fields can’t be used for the prediction of accurate results.

1. **Sales of Closed Stores**

It’s assumed that the sales of a closed shop are 0. Therefore, there is no use of predicting the sales of a closed shop. So, all the data related to closed shops is removed which reduces the size of the data by almost one-fifth of the original data.

1. **Merging fields from Store dataset to the Train dataset**

The sales of a store will always be affected by the presence of other stores in the vicinity. The Store dataset has variables like CompetitionDistance whereas the variable Sales is present in the Train dataset. We would like to use CompetitionDistance to determine the sales of the store. Therefore, we add CompetitionDistance to the Train dataset.

***EDA***

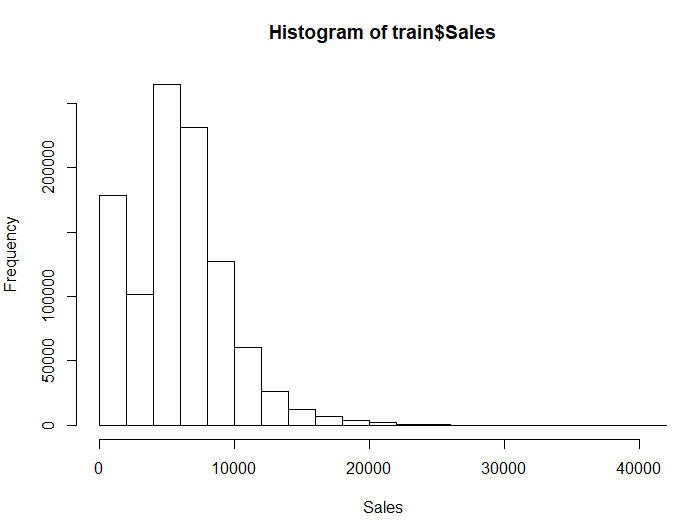
* ***What is EDA?***

EDA stands for Exploratory Data Analysis and it is an approach for data analysis to make sense from the dataset and then figure out what questions we want to ask and the best way to manipulate the data to get the needed answers.

* ***EDA on the Rossman Store Data.***

1. **Descriptive Statistics on Sales**

Since this is the variable to be predicted, we got the summary of the variable and checked the histogram. Outliers were already checked. Here, we also checked if it has NA values. The summary and the histogram are as follows:

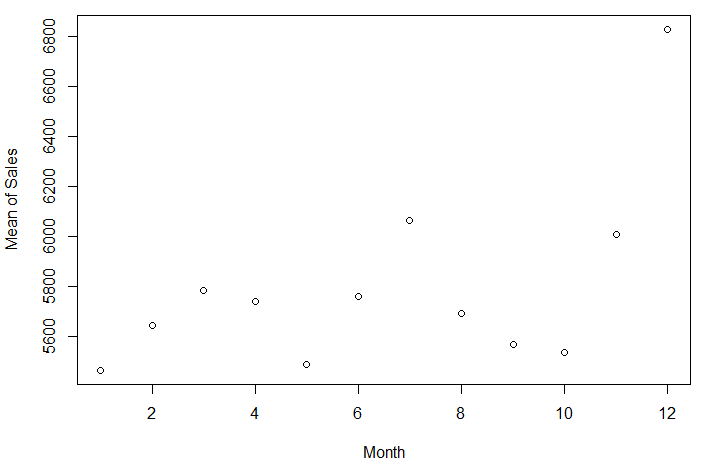




From the summary we can see that the median is slightly less than that of mean which means that the data is right skewed (positive skewness). We already know that a lot of stores have high sales indicating that there are no outliers.

1. **Variations in the Sales vs Months**

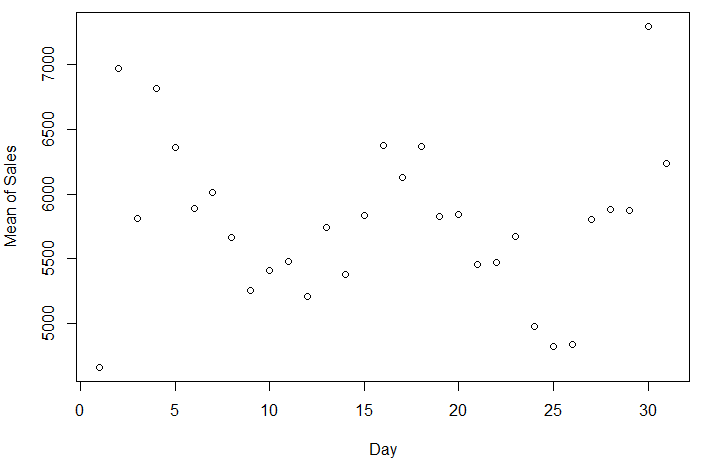
To check how the sales vales over the months of the year, we used the plot of the tapply function as got the following graph:



As we see, the sales are the maximum during the vacation period i.e. December and July. As it would be expected, the sales of the stores do tend to go higher during the holiday period as more customers visit stores due to the holidays. Thus, month will play an important role while predicting the sales of the stores.

1. **Variation in the Sales vs Days**

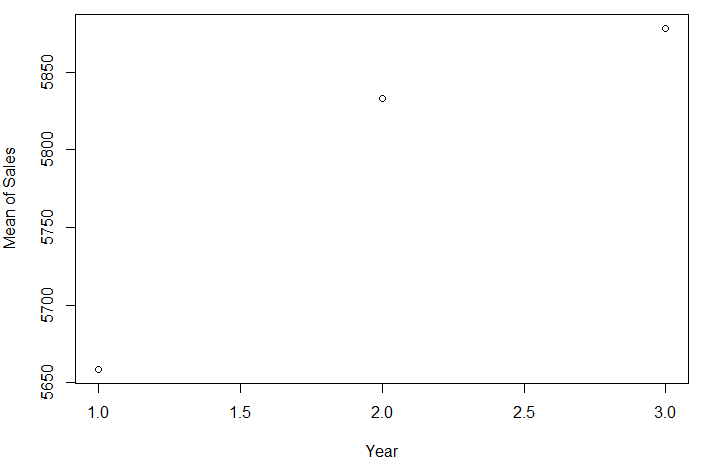
Just like for months, we also check if days of the month have any impact on the sales. The graph that we obtain is as follows:



As we know, most of the people get paid during the end of the month or during the first few days of the next month. As seen the sales are high during the start and end of the month which does match with the pay dates of the people. However, we also observe high sales during the middle of the month. Thus, further analysis may be needed to check if days does really affect the sales of the stores.

1. **Variation in the Sales vs Years**

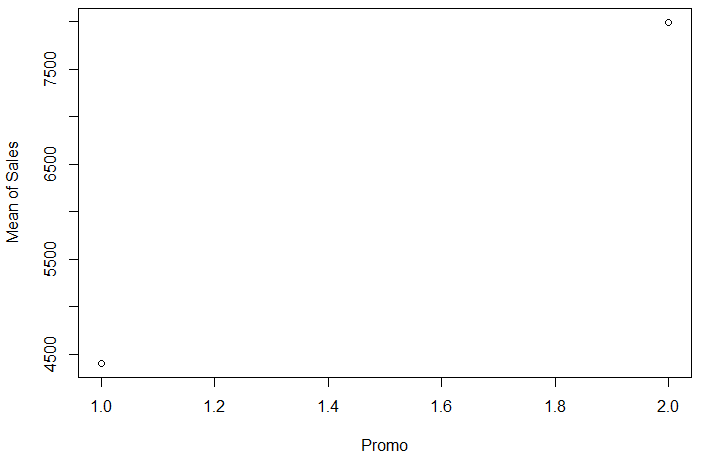
In the train datset, we have data for the years 2013, 2014 and 2015. So, we will check the trend of the sales fro these 3 months. The graph that we obtained is as follows:



As we can see that the sales goes on increasing every year. However, we are unsure as to what factor lead to the increase in the sales over the years. Thus, in order to use this for the prediction of the sales, we will have to first analyse the factors influencing the trend of increased saled over the years.

1. **Variation in the Sales vs Promo**

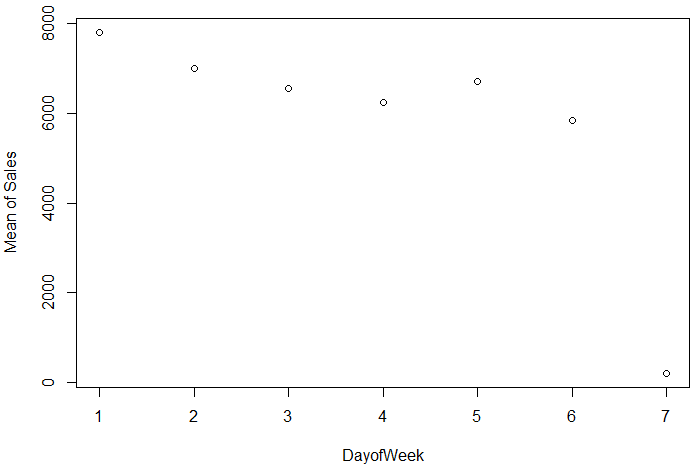
Also, promotional strategies always have a good influence over the sales. To determine that, we plot the following graph:



Thus, we see that the promotional events carried out do result in higher sales as compared to sales for stores that do not have the promotional event.

1. **Variation in the Sales vs DayofWeek**

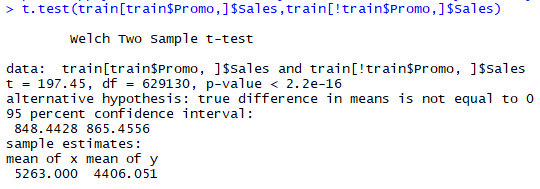
We know that the stores that are shut on Sunday have a 0 sales value while on open days the sales value is varying. Thus, to determine if the sales does get affected by the day of the week, we have the following graph:



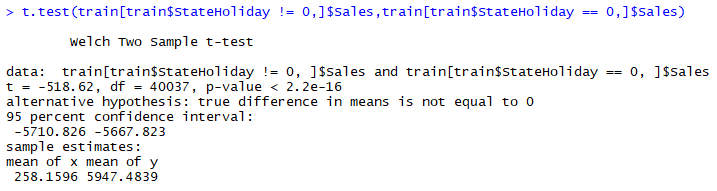
Thus, we see that people tend to go to store during the beginning and end of the working days i.e. on Monday and Friday. And even from this graph the sales on Sunday when the stores are closed is 0.

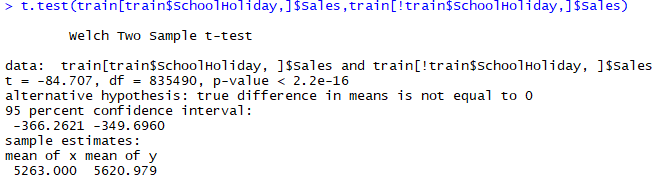
1. **T-test for Promo**

As we see, promotions have a very significant effect on sales, which are about 40 per cent higher on average during promotions.



1. **T-test for StateHoliday and SchoolHoliday**





Thus, we observe absurd variations in the variation of sales depending on StateHoliday and SchoolHoliday. Thus, we need further analysis to determine their impact of the sales of the stores.

***INTERPRETATION OF THE RESULTS***

Based on the data preprocessing and EDA done so far, we have the following interpretations:

1. Since the shops remain closed on Sundays and holidays, their sales will be 0. This is an important parameter to be considered for further analysis.
2. The sales are affected by the days of the month, specific months and the sakes increase over the years. These factors must be a part of any further analysis which may be performed.
3. Promo, StateHoliday and SchoolHoliday have an impact o the sales. These trends must be further analyzed.
4. Since the sales field does not have any outliers it means there are no erroneous values which may be present.
5. There is not much difference in the mean and median of the sales. Thus, the data is not heavily skewed.
6. Merging the data from the train and store dataset was imperative, as the train dataset did not account for the competitor’s effect on the sales.
7. Sales were affected by the day of the week. Thus, prediction algorithms must consider that factor as well.
8. There were certain fields like Promo2SinceWeek, Promo2SinceYear and PromoInterval which had a lot of null values. Predicting these values would be difficult. Thus, these fields must be avoided as predictions made on their references would produce erroneous results.

***FINAL DELIVERABLES***

* **Modelling Plans- Algorithms that can be used for Predicting the Sales**

We will use a few of the algorithms for the modelling algorithms that can predict Sales:

1. **Multiple regression** can be used for estimating store sales using a set of numerical and categorical predictors. A regression fit line will be obtained using the corresponding independent predictors. Independence of variables can be confirmed through correlation matrix. Adjusted R2 score can be used for prediction evaluation along with the lift-chart and ROC curve parameters.
   * Target variable: Sales.
   * Predictor variables: DayOfWeek, Date, Promo, CompetitionDistance, CompetitionOpenSinceYear, Promo2.
2. **Classification using Single Tree** can be used for a given limit (threshold), estimating store sales to be above or below the limit using the chosen set of predictors. During this classification, a training set is recursively divided until each following division consists of examples from one class. The whole classification process is based on recursive partition and divide and conquer algorithms. Specifically, a training set is recursively divided until each following division consists of examples from one class. The resultant tree structure is built by creating a root node and assigning all of the training data to it; selecting the best splitting attribute; adding a branch to the root node for each value of the split used to divide the data into mutually exclusive subsets; repeat previous two for each and every leaf node until the stopping criteria is reached.
3. **Classification using Random trees (forests)** can be used for estimating sales to be above or below a certain threshold. Such classification can be used as a way of averaging multiple deep decision trees trained on different parts of the same training set, with the goal of reducing the variance.
4. **K-nearest neighbor (kNN) algorithm** can be utilized for an initial clustering of stores similar in a certain set of parameters, e.g., Customer, Promo, Assortment, CompetitionDistance, etc. In kNN algorithm, an object is classified by a majority vote of its neighbors and being assigned to the class most common among its k nearest neighbors. After clustering is completed, the prediction of sales can be performed for the obtained clusters using one of the described above techniques: Single Tree or Random Trees.

* **Plan:**

|  |  |  |
| --- | --- | --- |
| Data Preparation | April 22, 2018 | Darshak, Sunidhi, Nikita |
| Modelling | April 22, 2018 | Darshak, Nikita |
| Discussions of Results | April 27, 2018 | All |
| Preparation of Final Report | April 29, 2018 | All |

***REFERENCES***

1. **Discovering Knowledge in Data:  An Introduction to Data Mining**by Larose and Larose.  (2nd edition) ISBN:  978-0-470-90874-7
2. https://www.techopedia.com/definition/14650/data-preprocessing
3. https://www.kaggle.com/anshumanyp/rossman/data
4. https://rpubs.com/gpetho/142772
5. http://www.rpubs.com/mordoobadi/Rossmann-Store-Sales