STATISTICAL DATA MINING

ROSMANN STORE PROMOTIONS ANALYSIS

GROUP4:
TARJANI SHAH
ADRIAN TEXEIRA
DISHA NEVE
TANUJ SAWANT

DEPARTMENT/OFFICE NAME
Email | Phone | Website

TABLE OF CONTENTS

1. INTRODUCTION

Rossman is a chain of stores in 7 European countries that operates over 3000 drug stores. The sales for the stores are influenced by several factors like promotion, competition, school and state holidays, store type, seasonality and locality. The problem which we will be addressing in our project is understanding the effect of promotional data on sales and the count of customers with respect to three types of assortments.

The interesting fact about this problem is that it would help us evaluate the effect of duration of promotional offers on sales. Data available is hierarchically divided into three layers of assortment. We have data which reports daily sales for 1115 stores, spread over a duration of three years.

2. DATA ANALYSIS AND VISUALIZATION

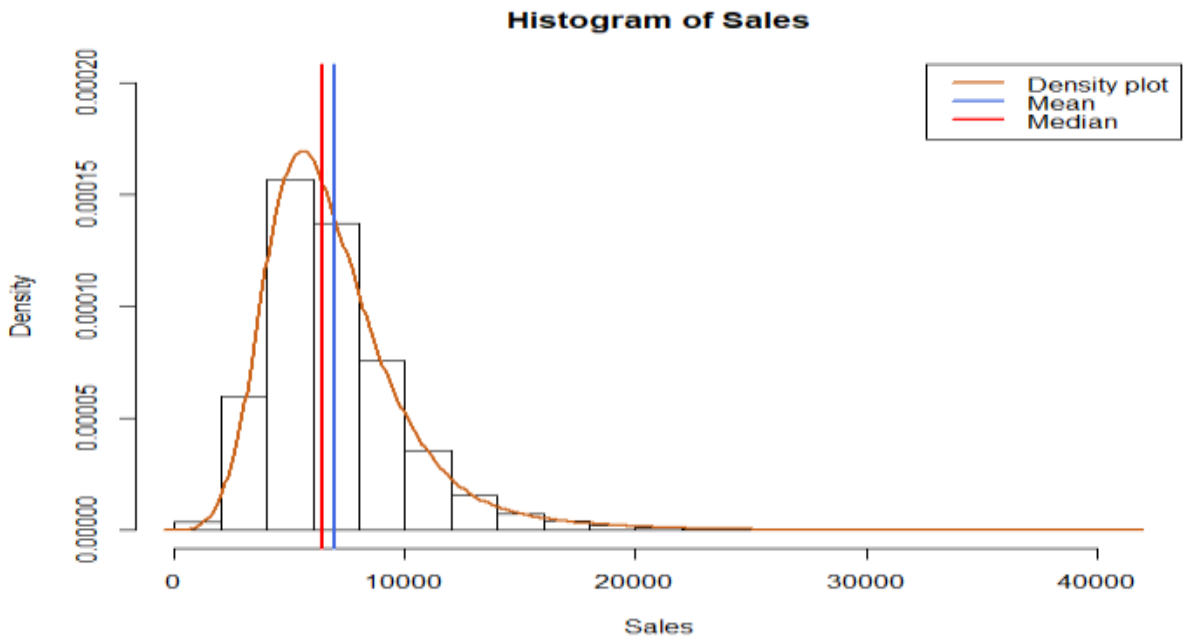
Data is sourced from Kaggle, which was provided by Rossmann for their sales prediction competition. It has over 800,000 records and 19 attributes. The attributes are as follows:

1. **Store:** Unique store ID.
2. **StoreType:** It is divided into four different categories(a,b,c,d)
3. **Assortment:** There are three levels of Assortment.
   - **Assortment a indicates Basic**
   - **Assortment b indicates Extra**
   - **Assortment c indicates Extended**
4. **DayOfWeek:** It indicates day of the week.
5. **Date:** It indicate the date of the sales.
6. **CompetitionDistance:** It indicates distance from the nearest competition store.
7. **CompetitionOpenSinceMonth:** It indicates if the competitor store has opened recently within a year's period. Stores which have been open for more than a year are indicated by value 13.
8. **Promo:** It indicates single day promotional offer.
9. **Promo2:** It indicates if the promotional offers continue over the next day (More than a day).
10. **Promo2SinceWeek:** It indicates the number of weeks since the promotion started.
11. **Promo2SinceYear:** It indicates the year since the promotion started.
12. **PromoInterval:** It indicates the months in which promotional offers are implemented(seasonal).
13. **Sales (Target Variable):** It indicates the sale for that day.
14. **Customers (Target Variable):** It indicates the number of customers for that day.
15. **Open:** It indicates whether the store was open or not.
16. **StateHoliday:** It indicates whether it was a state holiday or not.
    - **A : Public Holiday**
    - **B : Easter Holidays**
    - **C : Christmas Holiday**
17. **SchoolHoliday:** It indicates whether it was a school holiday or not.
18. **Month:** Indicates the month.
19. **Year:** Indicates the year.

Available data has 73% of promotional data and the rest is non-promotional data. The average sales for the promotional data is 6859.85 units and non-promotional data is 6949.27 units.

2.1.VARIABLE ANALYSIS

(1) Sales:



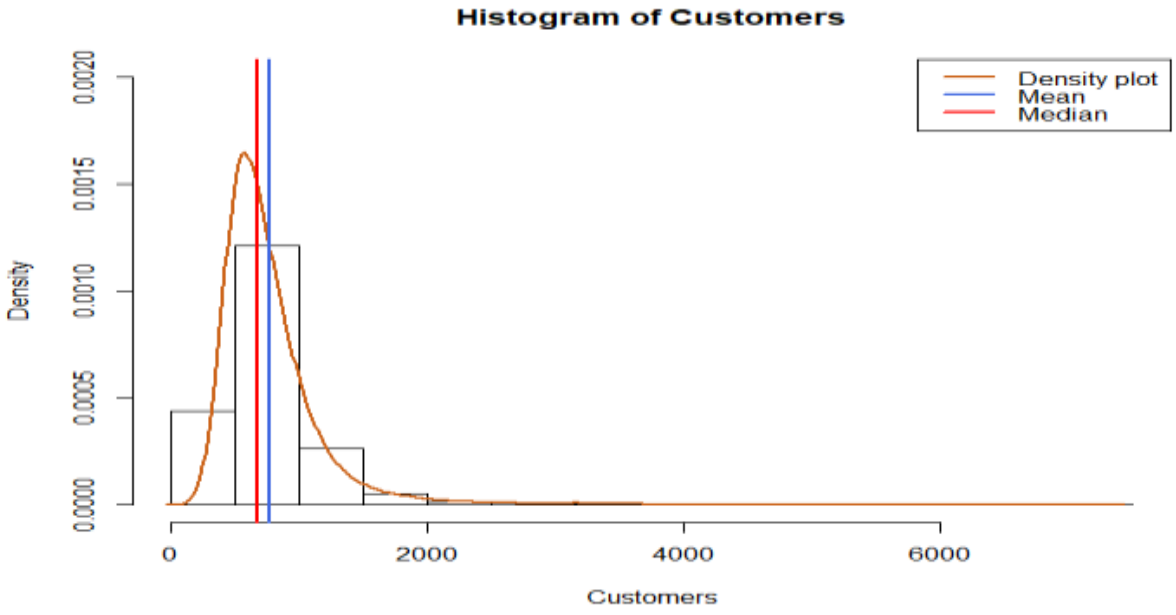
HISTOGRAM: SALES

OBSERVATIONS:

Based on the above histogram,

1. The data is right skewed.
2. Median is a better measure of central tendency and mean is greater than median.
3. The data has kurtosis.

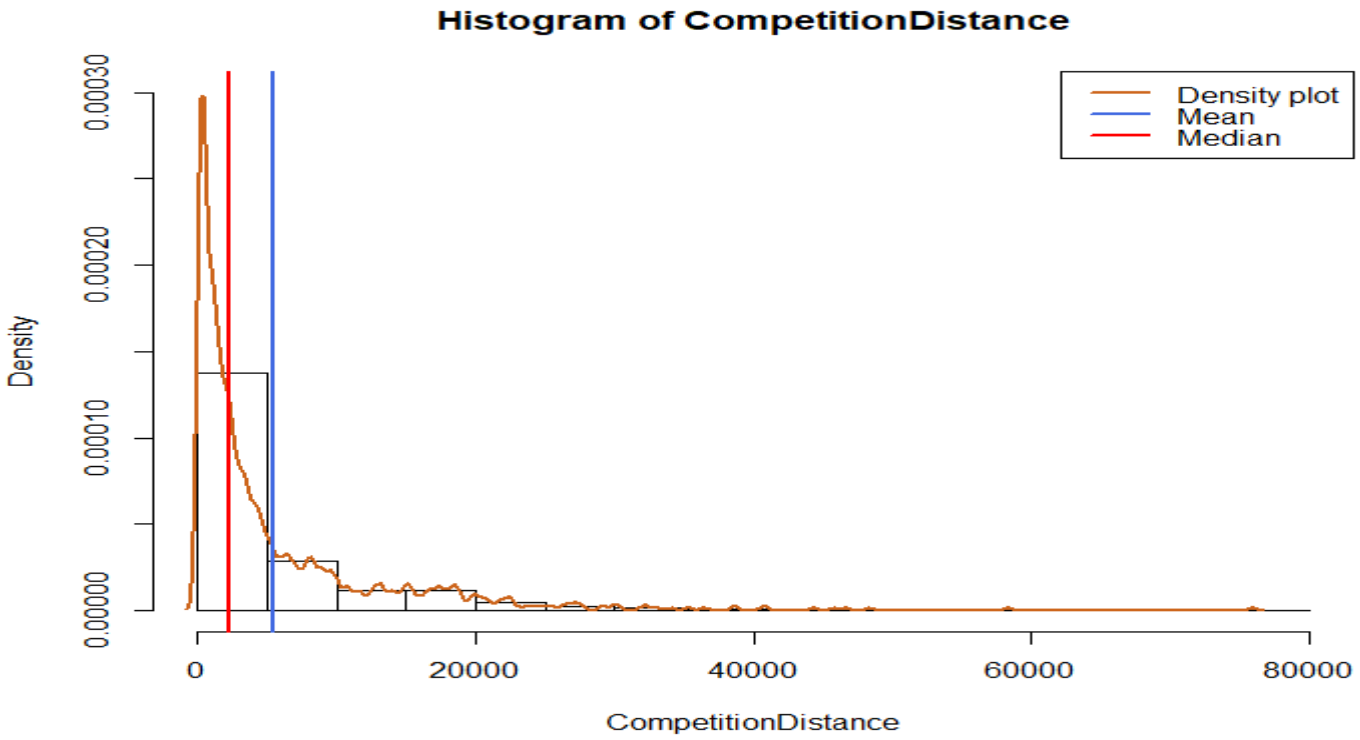
(2) Customers:



HISTOGRAM: FREQUENCY OF CUSTOMERS

OBSERVATIONS:
Based on the above histogram,
1. Histogram is slightly right skewed.
2. The data has almost normal distribution
3. The data has very high kurtosis

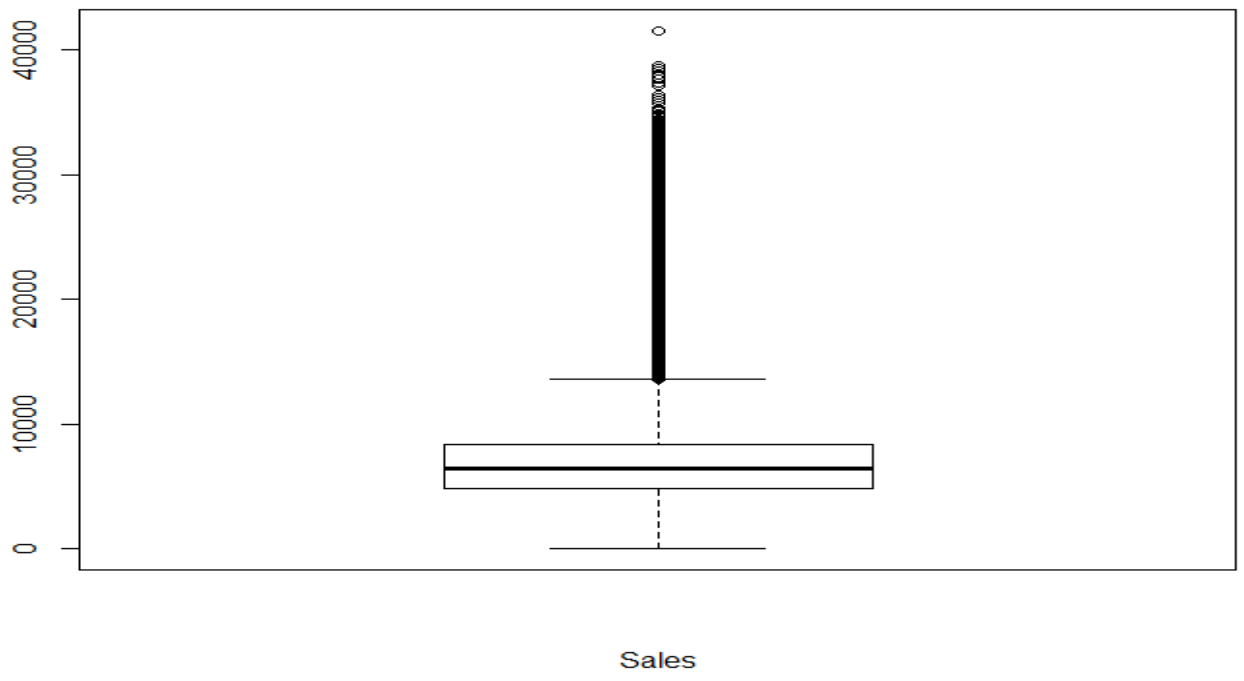
(3) Competition distance:



HISTOGRAM : COMPETITON DISTANCE

OBSERVATIONS:
Based on the above histogram,
1. The data is highly right skewed.
2. The data is not at all normally distributed
3. Median is much higher than mean

2.2.Outlier Analysis



BOXPLOT OF THE DEPENDENT VARIABLE: SALES

OBSERVATIONS

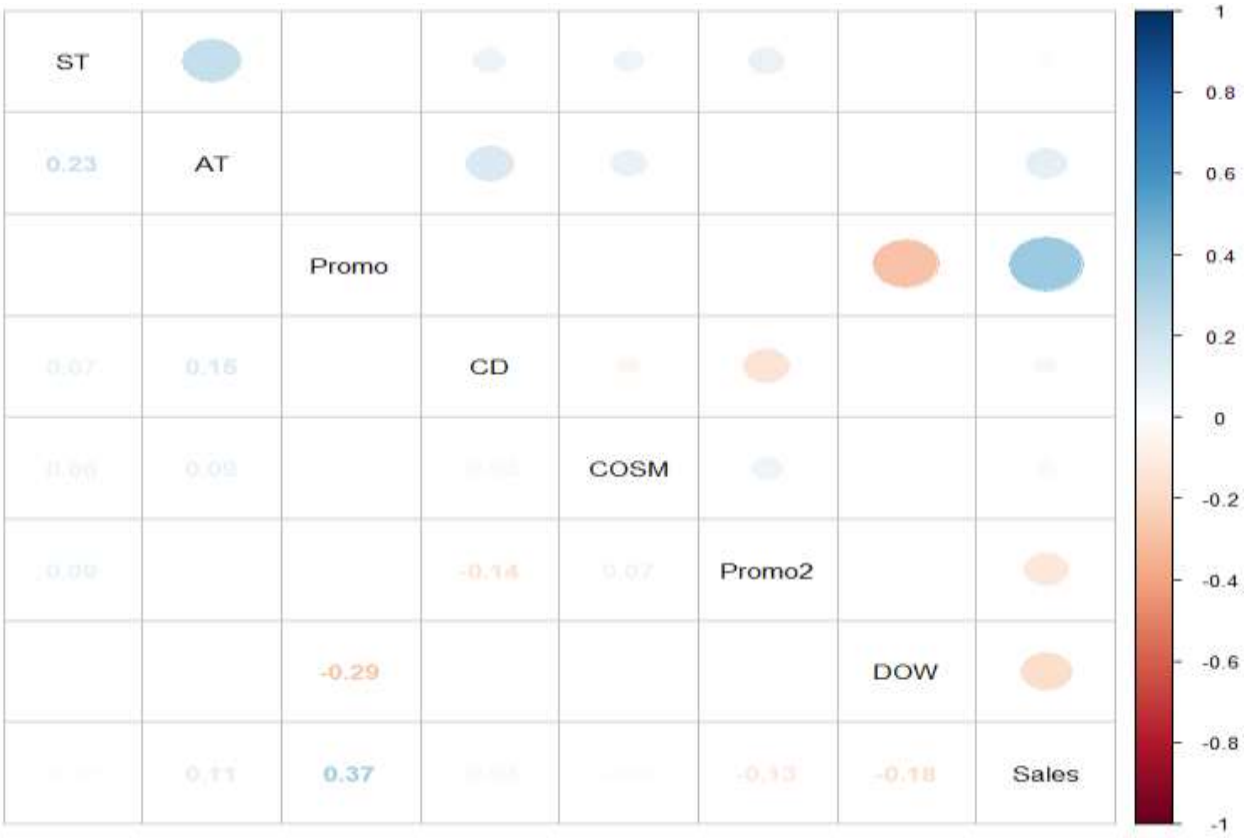
1. The boxplot indicates that there are 30744 outliers.
2. However, we cannot omit or delete any of these outliers as these are sales outliers and sales values can be extreme depending on many different factors like combination of holidays and promotional offers or considering holidays like Black Friday.
3. The outliers are less than 5% of the whole data.

3. CORE HYPOTHESIS

- If the promotional offers last more than a day, the sales will be relatively higher as compared to a single day promotion.
- If the promotional offers last more than a day, the number of customers will be relatively higher as compared to a single day promotion.
- If the assortment type is “Extended” (“type c”), with promotional offers will result in higher sales as compared to “Basic” (“type a”) and “Extra” (“type b”).
- If the assortment type is “Extended” (“type c”), the number of customers would be more when there are promotional offers, as compared to “Basic” (“type a”) and “Extra” (“type b”).

4. OLS ASSUMPTIONS

4.1.Multicollinearity:



The Correlation matrix

ST: Store type
AT: Assortment
CD: Competition Distance
COSM: Competition Open Since Month
DOW: Day of Weeks

1. Promo and Sales has positive correlation of 0.37 which indicates Promotional offers will result an increase in sales.
2. Assortment and store type has also positive correlation of 0.23, indicating that store type and assortment type have positive relationship, example store type 'a' can be further classified as either assortment type 'b' or 'c'.

   From the above correlation matrix, we can conclude that the data does not have multicollinearity.

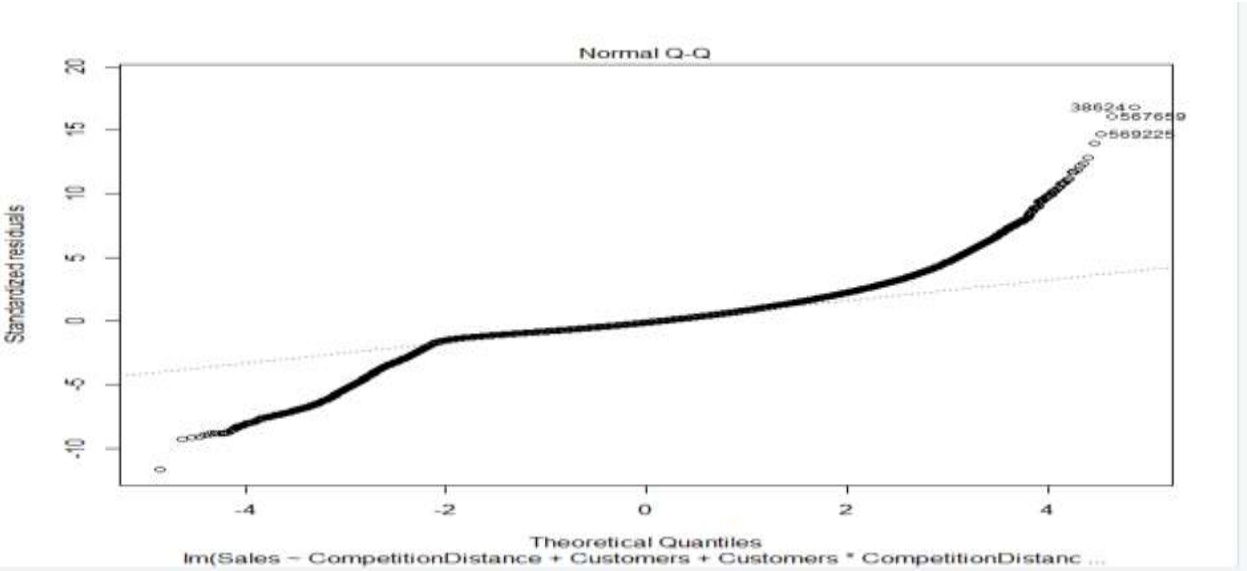
4.2.Auto correlation:

```
> m1 <- lm(Sales~ CompetitionDistance + Customers + Customers*CompetitionDistance)
> dwtest(m1)

        Durbin-Watson test

data:  m1
DW = 1.508, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

To test auto correlation, we conducted the DWTest, results are as shown in the above screenshot.

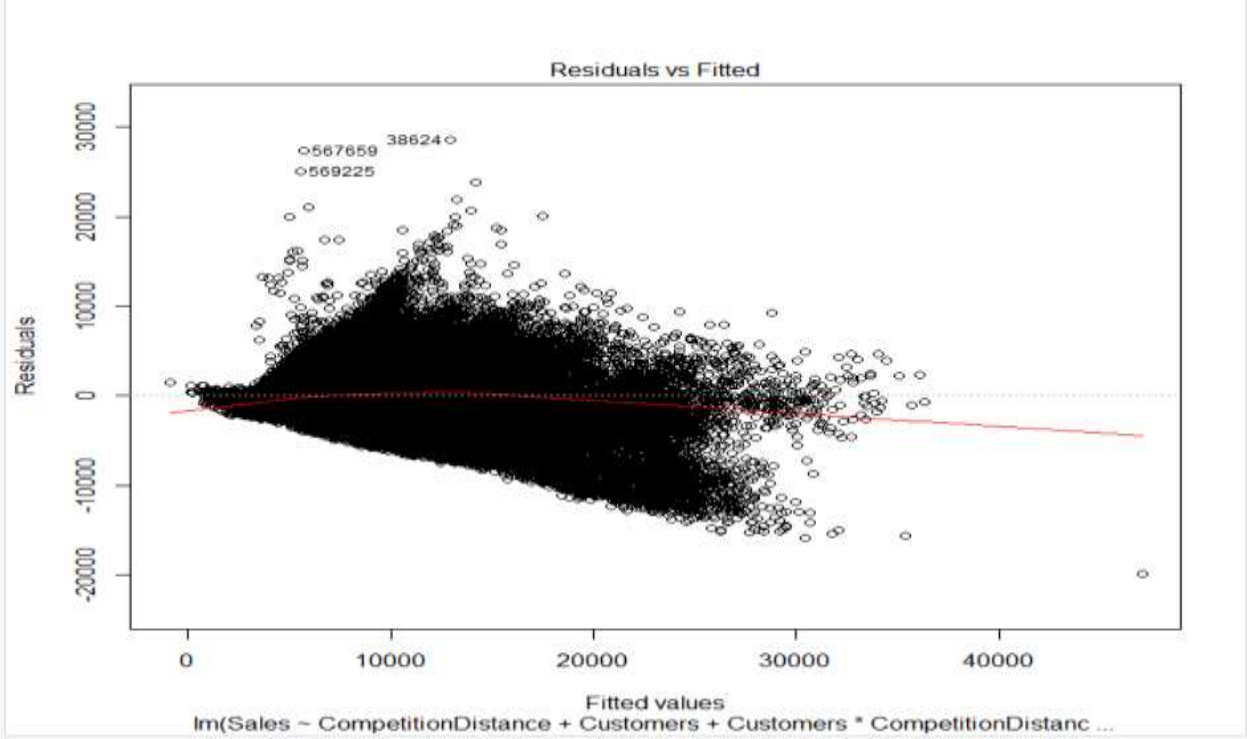
The null hypothesis states that Auto correlation does not exist. However, the P-value is less than 0.05 and hence we reject the null. It implies auto correlation exists in the data and hence violating OLS assumption.

4.3.Normality:



The above graph indicates that the data is not at all Normal with huge deviations on the extreme values. Hence violating another OLS assumption.

4.4.Homoscedasticity:



1. The above graph shows that data points are falling nearly equally on both the sides and hence it is unbiased.
2. Also, it can be seen that fanning effect exists, indicating the heteroscedastic nature and hence violating OLS assumption

Conclusion of OLS assumptions:

- It can be observed from the above tests and graphs that the OLS assumptions are being violated for the linear regression performed and hence we cannot use linear regression model for the data.
- Also, as observed the data appears to have levels, i.e multi-level, hence, we opt for Panel data modelling with Fixed effect and Random effect.

5. MODELS

Significance Indicators (p- value)

Significance		
***	Highly Significant	p < 0.001
**	More Significant	p < 0.01
*	Significant	p < 0.05
	Not Signifcant	p > 0.05

5.1POOL MODEL FOR SALES

*Pool_model<plm(Sales~as.factor(Assortment)+as.factor(Promo)+as.factor(Promo2)+as.factor(MONTH)+as.factor(YEAR)+Promo*CompetitionDistance,index = c("Store"),data = d,model = "pooling")*

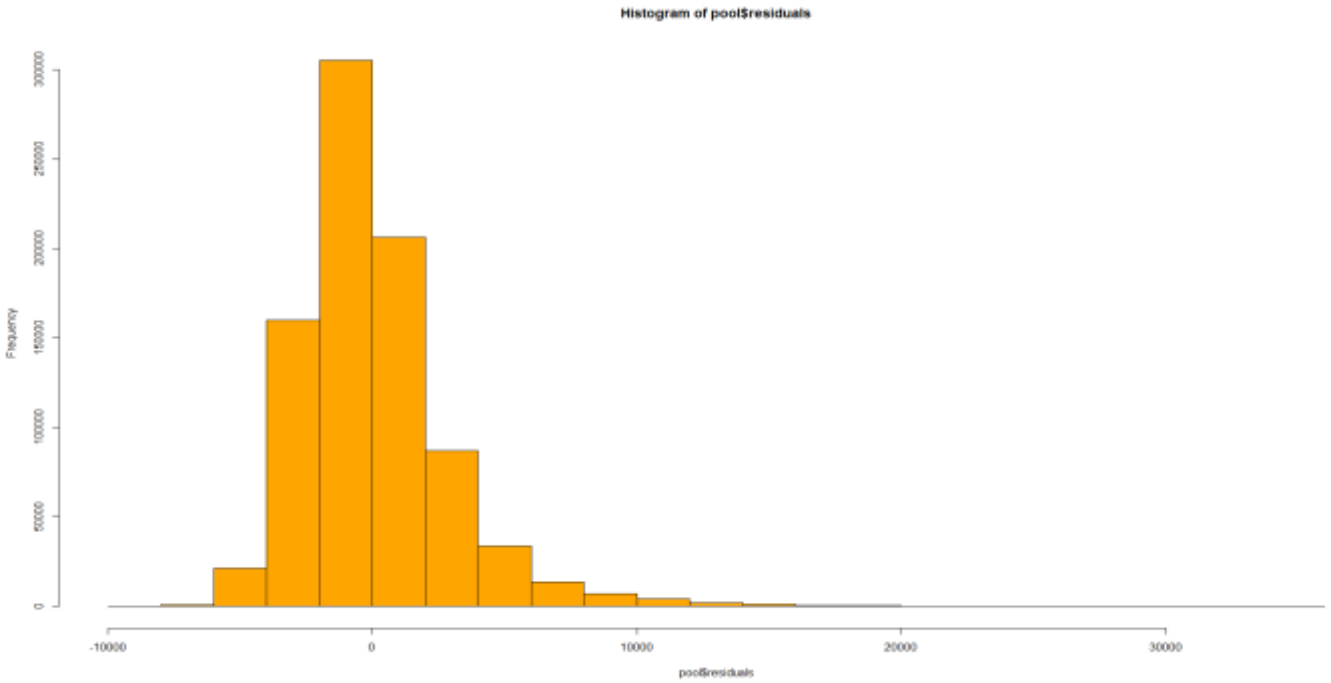
Equation:

SALES = **5626.9** + **2009.7***ASSORTMENT.B + **756.24***ASSORTMENT.C + PROMO.YES***2247.2** - PROMO2.YES***847.5** + MONTH.2***91.97** + MONTH.3***332.0** + MONTH.4***440.34** + MONTH.5***525.46** + MONTH.6***470.76** + MONTH.7***285.96** + MONTH.8***171.41** + MONTH.9***79.230** + MONTH.10***149.47** + MONTH.11***578.69** + MONTH.12***2183.4** + YEAR.2014***157.61** + YEAR.2015***328.66** - COMPETITIONDISTANCE***0.0326**+PROMO.YES: COMPETITIONDISTANCE***0.016**

```
                            Estimate  Std. Error   t-value  Pr(>|t|)
(Intercept)                5.6269e+03  1.1897e+01  472.9575 < 2.2e-16 ***
as.factor(Assortment)b     2.0097e+03  3.1017e+01   64.7920 < 2.2e-16 ***
as.factor(Assortment)c     7.5624e+02  6.1783e+00  122.4039 < 2.2e-16 ***
as.factor(Promo)YES        2.2472e+03  7.4503e+00  301.6233 < 2.2e-16 ***
as.factor(Promo2)YES      -8.4753e+02  6.1337e+00 -138.1772 < 2.2e-16 ***
as.factor(MONTH)2          9.1976e+01  1.3659e+01    6.7338 1.654e-11 ***
as.factor(MONTH)3          3.3203e+02  1.3422e+01   24.7386 < 2.2e-16 ***
as.factor(MONTH)4          4.4034e+02  1.3594e+01   32.3926 < 2.2e-16 ***
as.factor(MONTH)5          5.2546e+02  1.3665e+01   38.4539 < 2.2e-16 ***
as.factor(MONTH)6          4.7076e+02  1.3558e+01   34.7206 < 2.2e-16 ***
as.factor(MONTH)7          2.8596e+02  1.3439e+01   21.2779 < 2.2e-16 ***
as.factor(MONTH)8          1.7141e+02  1.5465e+01   11.0835 < 2.2e-16 ***
as.factor(MONTH)9          7.9230e+01  1.5646e+01    5.0639 4.109e-07 ***
as.factor(MONTH)10         1.4947e+02  1.5561e+01    9.6056 < 2.2e-16 ***
as.factor(MONTH)11         5.7869e+02  1.5736e+01   36.7746 < 2.2e-16 ***
as.factor(MONTH)12         2.1834e+03  1.5827e+01  137.9561 < 2.2e-16 ***
as.factor(YEAR)2014        1.5761e+02  6.9347e+00   22.7281 < 2.2e-16 ***
as.factor(YEAR)2015        3.2866e+02  8.4499e+00   38.8955 < 2.2e-16 ***
CompetitionDistance       -3.2678e-02  5.2484e-04  -62.2625 < 2.2e-16 ***
PromoYES:CompetitionDistance  1.0655e-02  7.8233e-04   13.6189 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    8.1143e+12
Residual Sum of Squares: 6.5154e+12
R-Squared:      0.19704
Adj. R-Squared: 0.19702
F-statistic: 10876.4 on 19 and 842132 DF, p-value: < 2.22e-16
```

Interpreting the model:

- If the assortment type is b or c the Sales increases relative to assortment type a with assortment b having the maximum increase (highest beta coefficient).
- If there is single day promotion the sales increases higher to no promotion and if the promotion is for more than one day it has negative beta coefficient hence indicating negative impact on sales.
- The sales of all other months show increase in sales as compared to January.
- Year 2014 and 2015 show an increase in sales as compared to 2013(as they have positive coefficients).
- If the competition distance is more than the sales would decrease, the coefficient is statistically significant but practically insignificant.



HISTOGRAM OF RESIDUALS(SALES_POOL)

5.2FIXED EFFECT MODEL FOR SALES

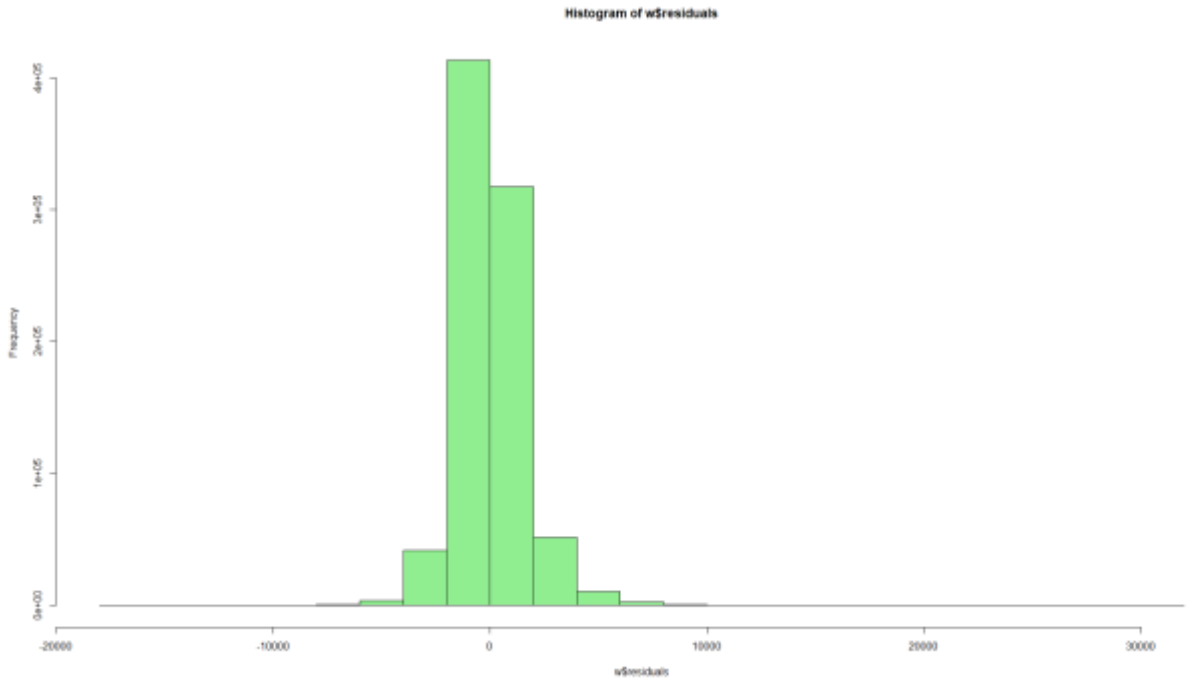
*Fixed_effect_sales<- plm(Sales~as.factor(Promo)+as.factor(MONTH)+Promo*CompetitionDistance,index = "Store",data = d,model = "within")*

Equation:

SALES = Promo.Yes***2266.6(***)** + MONTH.2***95.83 (***)**+ MONTH.3***336.1(***)** + MONTH.4***4436.57(***)** + MONTH.5***521.79(***)** + MONTH6***472.95(***)** + MONTH.7***298.43(***)** +MONTH.8***92.74(***)**+MONTH.9***2.74**+MONTH.10***70.97(***)**+MONTH.11***497.41(***)**+MONTH.12***2104.6(***)**+PromoYES:CompetitionDistance***0.002(***)**

Model Interpretation:

- We are creating a fixed effect on the store id.
- If there is a promotion, then there is increase in sales.
- As compared to January there is increase in sales as compared to all other months (positive coefficient)
- If there is a promotional offer and the competition distance is less there is increase in sales. Statistically it is significant but practically it is insignificant.



HISTOGRAM OF RESIDUALS(FIXED_EFFECT_SALES)

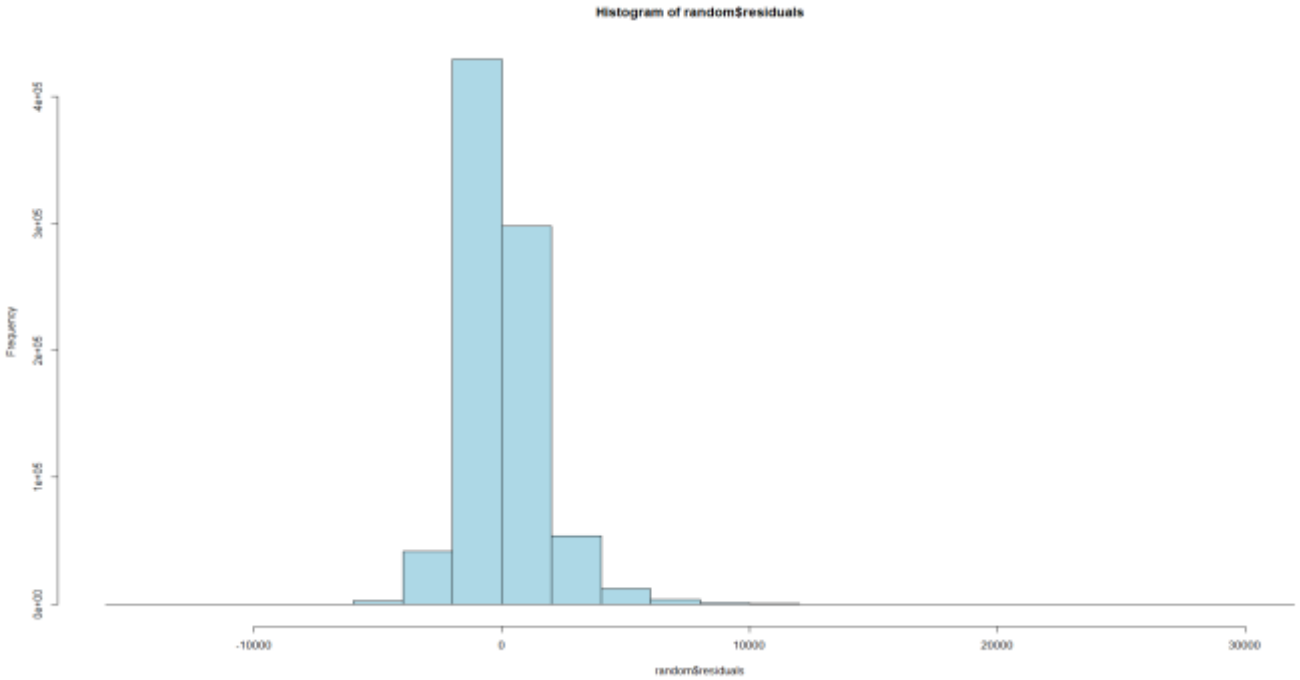
5.3RANDOM EFFECT MODEL FOR SALES

Random_model <- plm(Sales~as.factor(Promo)+as.factor(YEAR),data = d,index = c("Store"),model = "random")

Equation:
SALES = **5797.685**(***) + PROMO.YES***2309.21**(***) + YEAR.2014***161.33**(***) + YEAR.2015***202.991**(***)

Interpretation:

- This model depicts the random effect created by Store ID as this is a unique identifier.
- To preserve degrees of freedom, random effect has been created, hence instead of consuming 1115 degrees of freedom we are considering an random variable.
- The sales increase when there are single day promotional offers.
- As the year changes from 2013 to 2014-15, the sales increases across all the stores.



HISTOGRAM OF RESIDUALS(RANDOM_EFFECT_SALES)

5.4. POOLED MODEL FOR CUSTOMERS

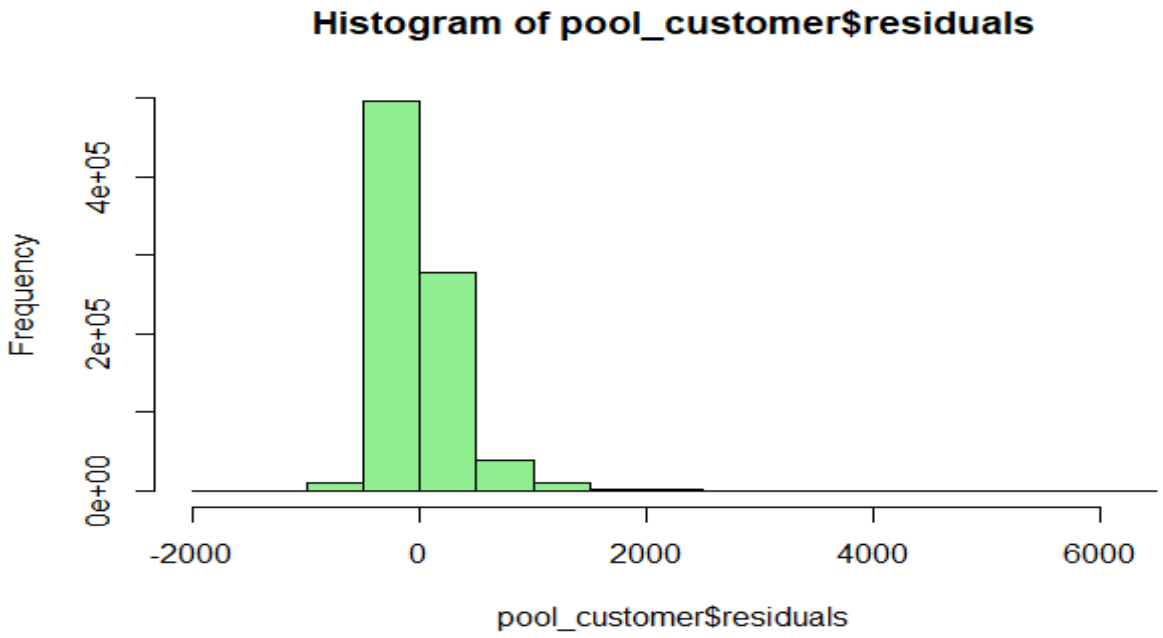
*pool_customers<plm(Customers~as.factor(Assortment)+as.factor(Promo)+as.factor(Promo2)+as.factor(MONTH)+as.factor(YEAR)+Promo*CompetitionDistance,index = "Store",data = d,model = "pooling")*

Equation:
CUSTOMERS = **765.34**(***) + ASSORTMENT.B***1289.1**(***) + ASSORTMENT.C***26.24**(***) + PROMO.YES***151.2**(***) - PROMO2.YES***177.67**(***) + MONTH.2***13.97**(***) + MONTH.3***33.20**(***) + MONTH.4***49.34**(***) + MONTH.5***54.74**(***) + MONTH.6***39.27**(***) + MONTH.7***22.96**(***) + MONTH.8***23.4**(***) + MONTH.9***19.36**(***) + MONTH.10***27.00**(***) +

MONTH.11*46.49(***) + MONTH.12*161.4(***) + YEAR.2014*9.79(***) + YEAR.2015*0.15266-COMPETITIONDISTANCE*0.008578(***) - PROMO.YES*0.0001065

Model interpretation:

- Here we are creating a pool model ignoring the fact that each store is different from other.
- Assortment type is b and c there is an increase in the number of customers relative to assortment a with assortment b having the highest increase in customer count (Highest beta coefficient)
- If there is single day promotion the number of customers increases higher to no promotion and if the promotion is for more than one day it has negative beta coefficient hence indicating negative impact on the number of customers.
- The customer count of all other months show increase in number of customers as compared to January.
- Year 2014 and 2015 show an increase in number of customers as compared to year 2013(as they have positive coefficients).



HISTOGRAM OF RESIDUALS(POOL_CUSTOMER)

5.5. FIXED EFFECT MODEL FOR CUSTOMERS

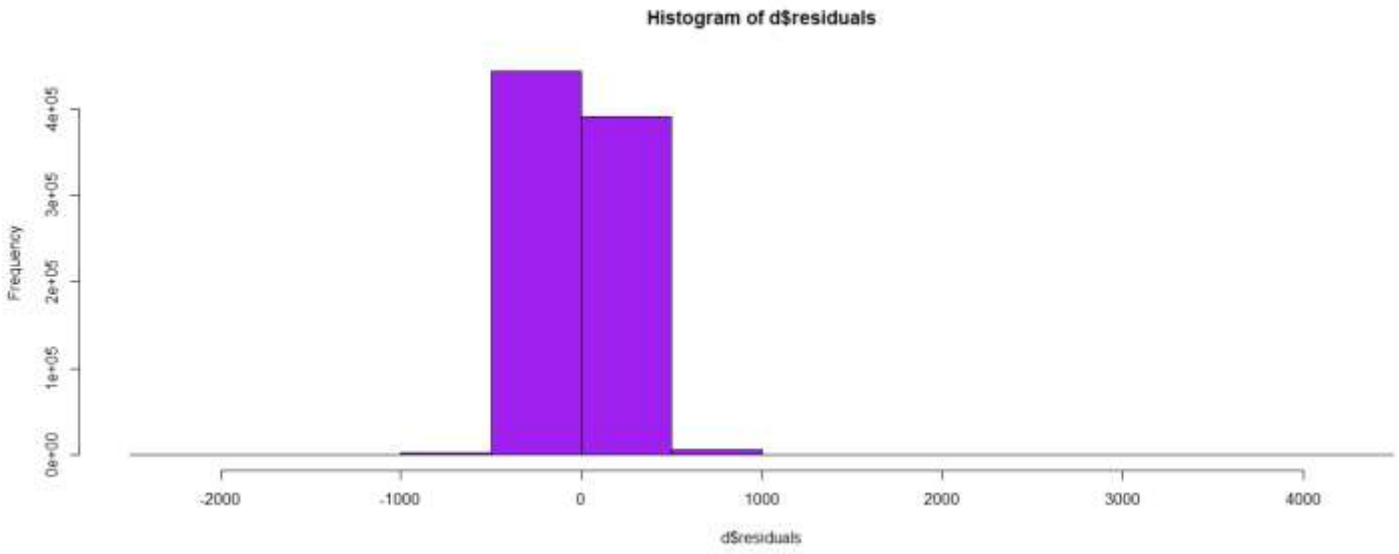
Fixed_customer <- plm(Customers~as.factor(Promo)+as.factor(YEAR)+ as.factor(StateHoliday),data = d,index = c("Store"),model = "within")

Equation:

CUSTOMERS = PROMO.YES*152.91(***) + YEAR.2014*4.2238(***) – YEAR.2015*9.201(***) + SCHOOLHOLIDAY.1*15.89(***)

Model interpretation:

- In this model we are creating a fixed effect for stores and hence we may compromise on the degrees of freedom.
- In this model if we are having promotions for a day we can observe an increase in number of customers (positive beta coefficient) by 152 customers.
- In the year 2014 the count of customers is more than year 2013 and in the year 2015 we can see a dip in the number of customers (negative beta coefficient).
- When there is a school holiday the number of customers visiting the store increases by 16 customers per store.



HISTOGRAM OF RESIDUALS(FIXED_EFFECT_MODEL)

5.6. RANDOM EFFECT MODEL FOR CUSTOMERS

random_customer <- plm(Customers~as.factor(Promo2)+as.factor(YEAR)+ StateHoliday,data = d,index = c("Store"),model = "random")

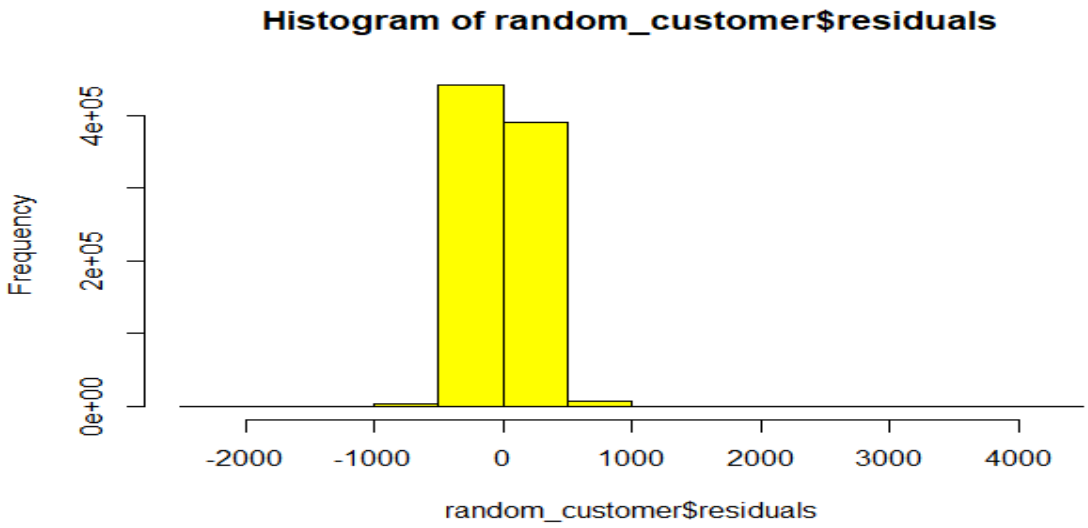
Equation:

CUSTOMER = **835.81(***)** – PROMO2.YES***160.41(***)** + YEAR.2014***5.768(***)** - YEAR.2015***5.08(***)** + STATE.HOLIDAY.A***11.61(.)** + STATE.HOLIDAY.B***67.47(***)** – STATE.HOLIDAY.C***369.75(***)**

Model interpretation:

- In this model we are creating a random effect for the stores which are our unique ids.
- As the promotion extends for more than a day we can see a decrease in the number of customers in the stores (negative beta coefficient) by 160 customers.
- In the year 2014 we can see an increase in the number customers as compared to 2013 but in the year 2015 we can see a decrease in the number of customers.

- For holidays we can see a increase in number of customers for Public Holiday(A) and Easter Holidays(B) but we can see a decrease in the number of customers for Christmas (C) due to the fact that for Christmas most people are to their relatives places or out of town.



HISTOGRAM OF RESIDUALS(RANDOM_MODEL_CUSTOMERS)

6. MODEL COMPARISONS

Dependent variable: SALES

Model_name	Test Performed	P- Value	Conclusion	R- Square	Adjusted R-Square
Sales:Pool_model	PlmTest(Pool_model)	2.2e^-16	There are significant pannel effect	0.198 or 19% variance explained by model	0.1902 or 19%
Fixed_effect_sales	pFtest(Fixed_effect_sales,Random_model)	2.2e^-16	Fixed effect model is better	0.4032 or 40% of total variance explained by model	0.402 or 40%
Random_model	pFtest(Fixed_effect_sales,Random_model)	2.2e^-16	Fixed effect model is better	0.364 or 36.4% of total variance explained by model	0.364 or 36.4%

Dependent variable: Count of Customers

Model_name	Test Performed	P- Value	Conclusion	R- Square	Adjusted R-Square
Pool_Customers	PlmTest(Pool_Customers)	2.2e^-16	There are significant pannel effect	0.21298 or 21% of total variance explained by model	0.2122 or 21%
Fixed_Customers	phtest(Fixed_customer,random_customer)	2.2e^-16	Fixed effect model is better	0.2102 or 21% of total variance explained by model	0.2110 or 21%
Random_Customers	phtest(Fixed_customer,random_customer)	2.2e^-16	Fixed effect model is better	0.001 or 0.14% of total variance explained by model	0.0010 or 0.10%

7. CONCLUSION:

After running the above Panel data models, we can conclude that

- When the promotional offers are applied there is a rise in the number of customers as well as in the Sales. That is, we can say both of our target variables (i.e Sales and Number of customers) are highly affected by the promotional offers. Hence, we can say that the promotional offers play a very crucial role to determine the sales of any given type of store.
- Our first hypothesis states that if the promotion is for more than a day the sales is higher relative to a single day promotion. But after running all the models we can observe that it is contradicting to our hypothesis. The data states that if the promotional offers are applied for a day there is an increase in sales but if the offers are for more than one day then the sales decreases.
- Our second hypothesis states that if the promotion is for more than one day then the number of customers increases. The data supports our hypothesis and we can interpret the same from the above model.
- Our third hypothesis states that if the Assortment type is “Extended” (Type c), the sales should be higher as compared to Assortment type a and b which are “Basic” and “Extra” respectively. But this hypothesis contradicts to what our data speaks after the modelling. The results show that Sales are maximum if the Assortment type is “b” which is “Extra”.
- Our fourth hypothesis states that if Assortment type is “Extended” (Type c), the number of the customers should be higher as compared to Assortment type “a” and “b” which are “Basic” and “Extra’ respectively. But this hypothesis contradicts as well. After the modelling results show that number of customers are maximum if the Assortment type is “b” which is “Extra”.
- For our second dependent variable customers we can see a decrease in the number of customers when the distance between stores increases. This can be possible because most stores are in cities and in cities there is not much distance between stores due to the paucity of space and hence increasing competition.

8. REFERENCES

- https://www.kaggle.com/c/rossmann-store-sales/data/
- https://www.princeton.edu/~otorres/Panel101R.pdf
- https://www.rossmann.de/einkaufsportal.html
- https://www.rossmann.de/einkaufsportal.html