MarketMixModeling KPI Selection

Atchireddy chavva

Thu May 11 23:00:37 2017

Load Libraries:

Load required libraries. Will use `stepAIC` from `MASS` package for model pruning. `vif` variation inflation factor from `car` package

```
library(MASS)
library(car)
library(DataCombine)   # Pair wise correlation
library(stargazer)
library(dplyr)         # Data aggregation
source('./code/atchircUtils.R')
```

Load Data:

Load blended data from sales and marketing datasets. Data is throughly cleaned, pre-processed for model building. Refer to `DataCleaning.R` script for data preparation steps. For this capstone project we limit our model building focus to `camera_accessory`, `Home_audio` and `Gaming_accessory` product sub-categories. For simplicity will start Linear model building with numerical features, later will consider categorical features.

```
data    <- read.csv('./intrim/eleckart.csv')
```

Data Aggregation:

Weekly aggregrated data is at granularity of `product_sub_category`, For Exploratory analysis, will collapse `product_sub_category` level weekly data.

```
data_week <- data %>% group_by(week) %>%
  summarise(gmv=sum(gmv),product_mrp=mean(product_mrp),
            discount=mean(discount),sla=mean(sla),procurement_sla=mean(procurement_sla),
            n_saledays=mean(n_saledays),TV=mean(TV),Digital=mean(Digital),
            Sponsorship=mean(Sponsorship),ContentMarketing=mean(ContentMarketing),
            OnlineMarketing=mean(OnlineMarketing),Affiliates=mean(Affiliates),SEM=mean(SEM),
            Radio=mean(Radio),Other=mean(Other),TotalInvestment=mean(TotalInvestment),
            NPS=mean(NPS),list_mrp=mean(list_mrp),
            units=sum(units),COD=sum(COD),Prepaid=sum(Prepaid))
```

Units Scaling: `GMV`, `Product_mrp` are in terms of INR, while marketing spend recorded in INR Cr. Lets convert marketing spend to INR. While models won't be sensitive to these parity in units, but for easy of model explanation, will standadize units.

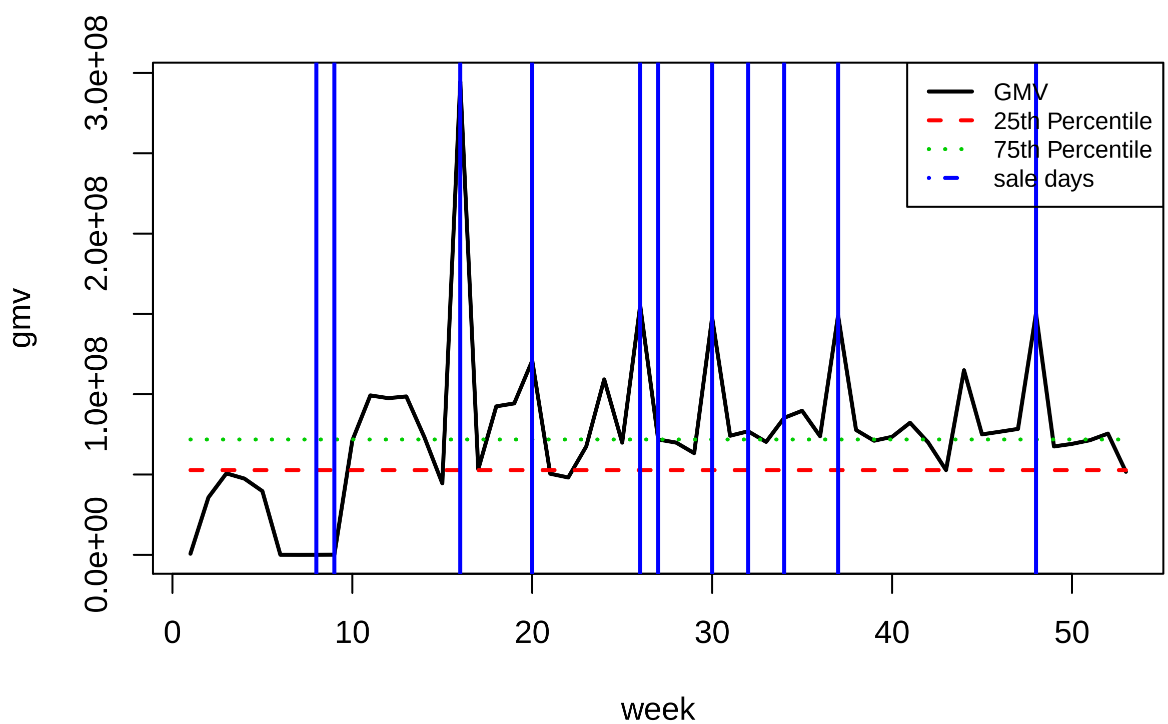
```
data_week[,c(8:17)] <- data_week[,c(8:17)]*10000000
```

Exploratory Data Analysis:

Will start analyzing the weekly `Gross Merchandize Value` the online website generating in sales. Below chart shows per week `GMV` over a period of 52 weeks. Horizontal dotted lines mark 25th 75th percentile of weekly expected `GMV`. Veritical lines represent whether any sale/promotion is carried out during the week.

```
quant <- quantile(data_week$gmv,c(0.25,0.5,0.75))
matplot(data_week$week, cbind(data_week$gmv,rep(quant[1],53),rep(quant[2],53)),
        type='l',lwd=2,xlab = 'week',ylab = 'gmv')

saledays <- data_week$week[data_week$n_saledays > 1]
abline(v=saledays,col='blue',lwd=2)
legend('topright', inset = 0, legend = c('GMV','25th Percentile',
                                         '75th Percentile','sale days'),
       lty = c(1:4), col=c(1,2,3,4), lwd = 2, cex = 0.75)
```

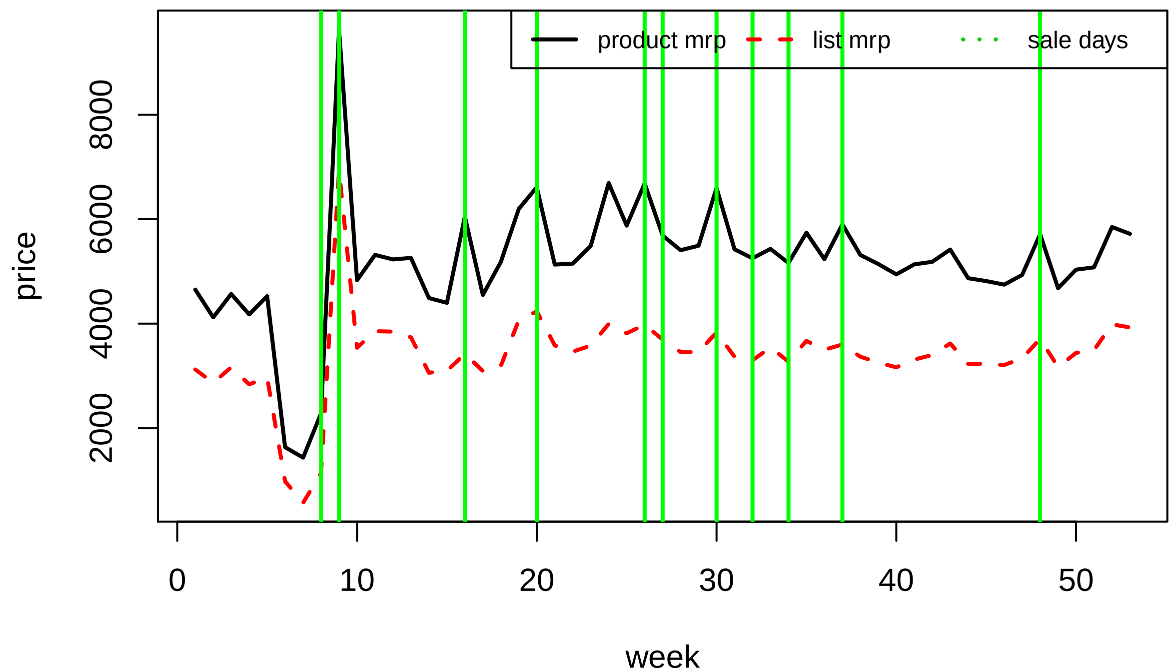


Insights: One can clearly observe the peaks in weekly `GMV` aligns with promotion markings. This clearly infers a positive impact of promotions on `GMV`. Week 16 sales is kind of outlier, could be week long promitions/sales events planned/festival sale. First 10 weeks sales were also seems outliers, but on the lower side of 25th percentile.

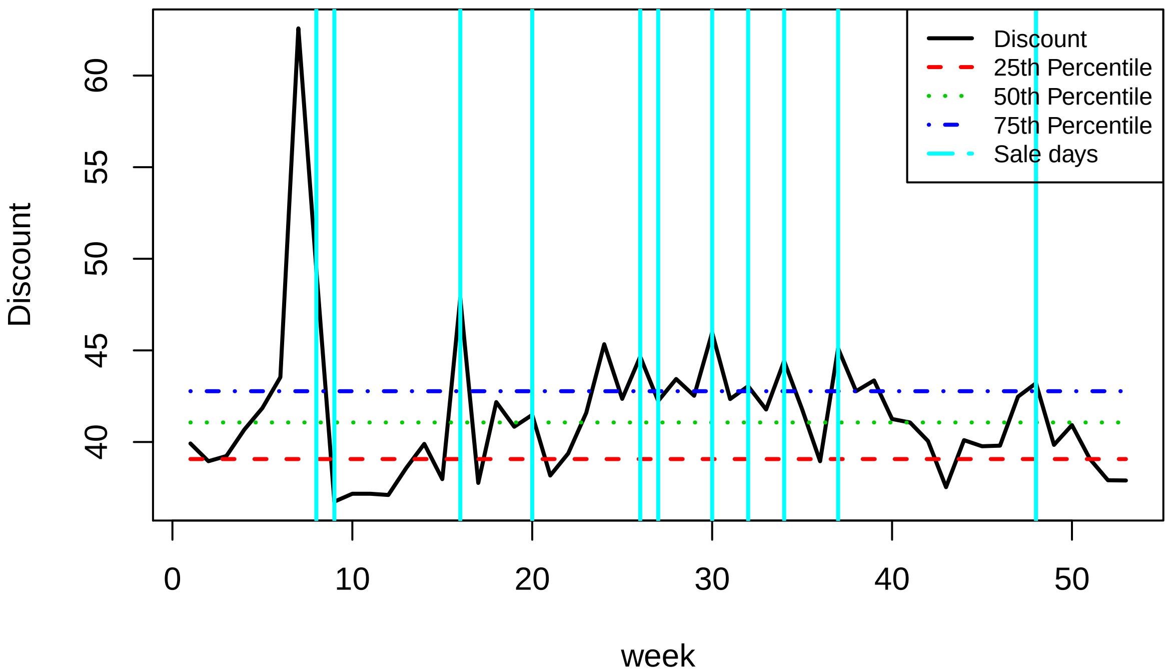
Pricing Analysis:

Will look at average weekly `Max Retail Price` and `Listed MRP` of the products sold through the website. Clearly `List MRP` is consistently lower w.r.t `product_mrp`. Next chart shows the average weekly discount.

```
matplot(data_week$week, cbind(data_week$product_mrp,data_week$list_mrp),
        type='l',lwd = 2,xlab='week',ylab='price')
abline(v=saledays,col='green',lwd=2)
legend('topright', inset = 0, legend = c('product mrp','list mrp','sale days'),
       lty = c(1:3), lwd = 2, col=c(1,2,3), horiz = TRUE, cex = 0.75)
```




```
quant <- quantile(data_week$discount,c(0.25,0.5,0.75))
matplot(data_week$week, cbind(data_week$discount,rep(quant[1],53),
                              rep(quant[2],53),rep(quant[3],53)),
        type='l',lwd=2,xlab = 'week',ylab = 'Discount')
abline(v=saledays,col='cyan',lwd=2)
legend('topright', inset = 0, legend = c('Discount','25th Percentile',
                                         '50th Percentile','75th Percentile',
                                         'Sale days'),
       lty = c(1:5), lwd = 2, col=c(1:5), cex = 0.75)
```



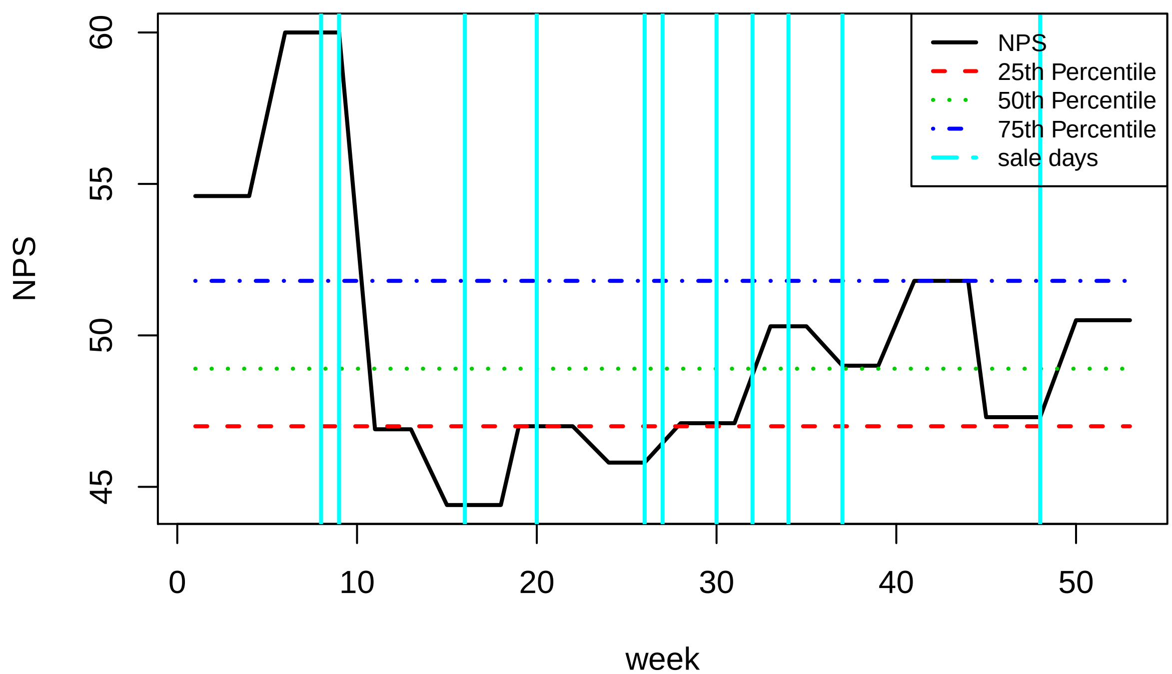
Insights:

1. Avg `MRP` of products sold during promotions is high
2. Discounts allowed throught out year
3. During Non-Promotion weeks, discounts vary by 5%
4. Discounts are higher by just 4% during promitions.

NPS(Net Promoter Score):

Will study how `NPS` tall across the weeks for a given year. `NPS` can vary anywhere between -100 and 100. `NPS` is a measure of customer satisfaction, loyalty. Basically an approapriate `NPS` is arrived through regular customer survey.

```
quant <- quantile(data_week$NPS,c(0.25,0.5,0.75))
matplot(data_week$week, cbind(data_week$NPS,rep(quant[1],53),
                              rep(quant[2],53),rep(quant[3],53)),
        type='l',lwd=2,xlab = 'week',ylab = 'NPS')
abline(v=saledays,col='cyan',lwd=2)
legend('topright', inset = 0, legend = c('NPS','25th Percentile','50th Percentile',
                                         '75th Percentile','sale days'),
       lty = c(1:5), lwd = 2, col=c(1:5), cex = 0.75)
```



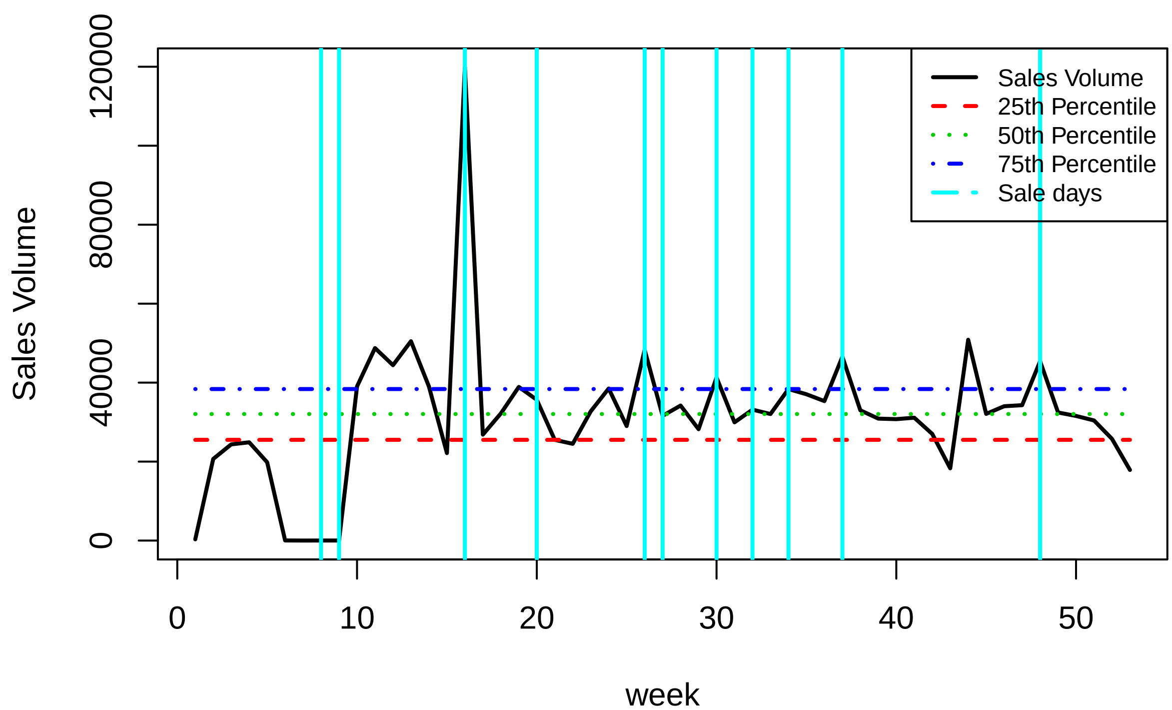
Insights:

Overall `NPS` for Eleckart is on the positive side ranging over 45-60 window. well the data around weeks 7-9 seems corrupted as seen on all previous charts, lets exclude these weeks data for our insights. I also overlay `sale/promotion` week markers, for better understanding of `NPS` during weeks where no promotions being run. Relatively `NPS` is higher during weeks where there are no promotions, kind of implying due to high sales volumes during promitons customer satisfaction might be compromised either due to delays in shipping or product quality. If we were to relates `NPS` to our sales, we see a inverse relationship between `NPS` and `GMV`, though it may not be granted, as a higher `NPS` might be helping sales during weeks with no promotions running.

Sales Volume:

Below is weekly sales volume. Its pretty much follows `GMV` pattern. Sales volume pretty much confines to a narrow band between 25th and 75th percentile, expect a peak at week 16 and a poor sales reported for first 9 weeks. we can ignore `Units` sold for model building as it doesn't sound like an attriting feature that can influence sales.

```
quant <- quantile(data_week$units,c(0.25,0.5,0.75))
matplot(data_week$week, cbind(data_week$units,rep(quant[1],53),
                              rep(quant[2],53),rep(quant[3],53)),
        type='l',lwd=2,xlab = 'week',ylab = 'Sales Volume')
abline(v=saledays,col='cyan',lwd=2)
legend('topright', inset = 0, legend = c('Sales Volume','25th Percentile',
                                         '50th Percentile','75th Percentile',
                                         'Sale days'),
       lty = c(1:5), lwd = 2, col=c(1:5), cex = 0.75)
```



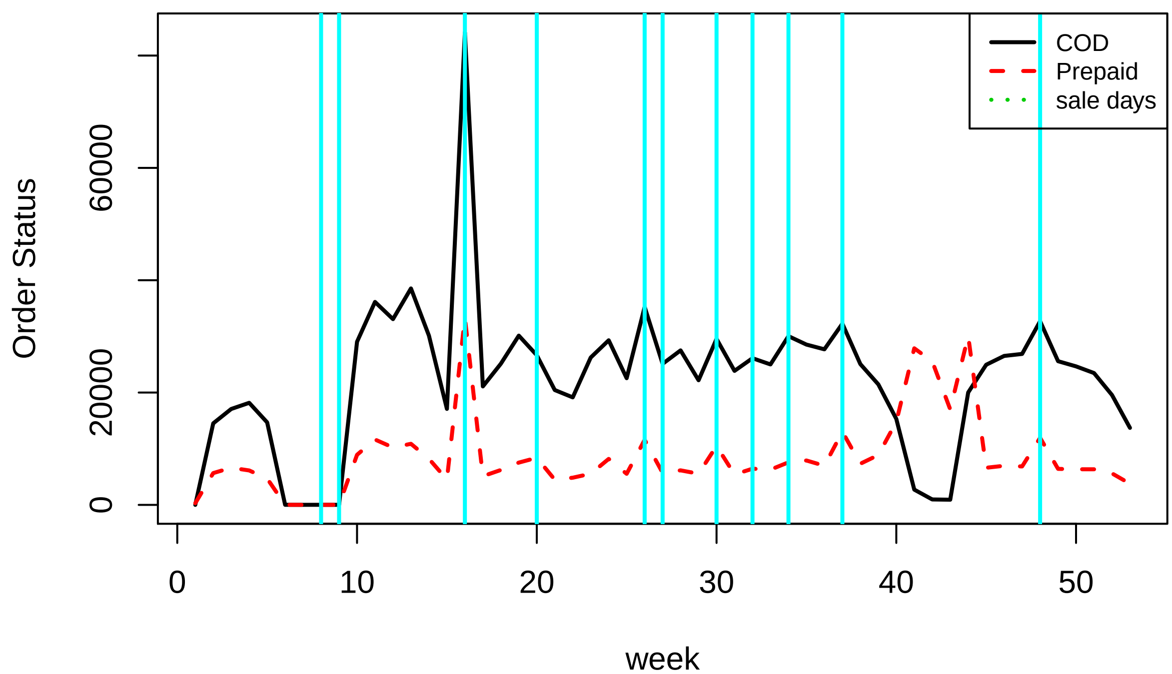
Insights:

Between weeks 10 and 48, sales volume is kind of stable, also one can observe, most of the promotions are being run during this period. By contrasting sales numbers among 10-48 weeks and rest of the weeks, we can imply pretty confidently promotions are kind of help sales.

Delivery Status:

For a developing country like India, where online transactions are not prevalent mode of payment of those goods bought online, e-commerce websites have alternate payment methods like `Cash on Delievery`. Will look at delivery status customer choose at weekly level.

```
matplot(data_week$week, cbind(data_week$COD,data_week$Prepaid), type = 'l',
        lwd = 2, xlab = 'week', ylab = 'Order Status')
abline(v=saledays,col='cyan',lwd=2)
legend('topright', inset = 0, legend = c('COD','Prepaid','sale days'),
       lty = c(1:3), lwd = 2, col=c(1:3), cex = 0.75)
```



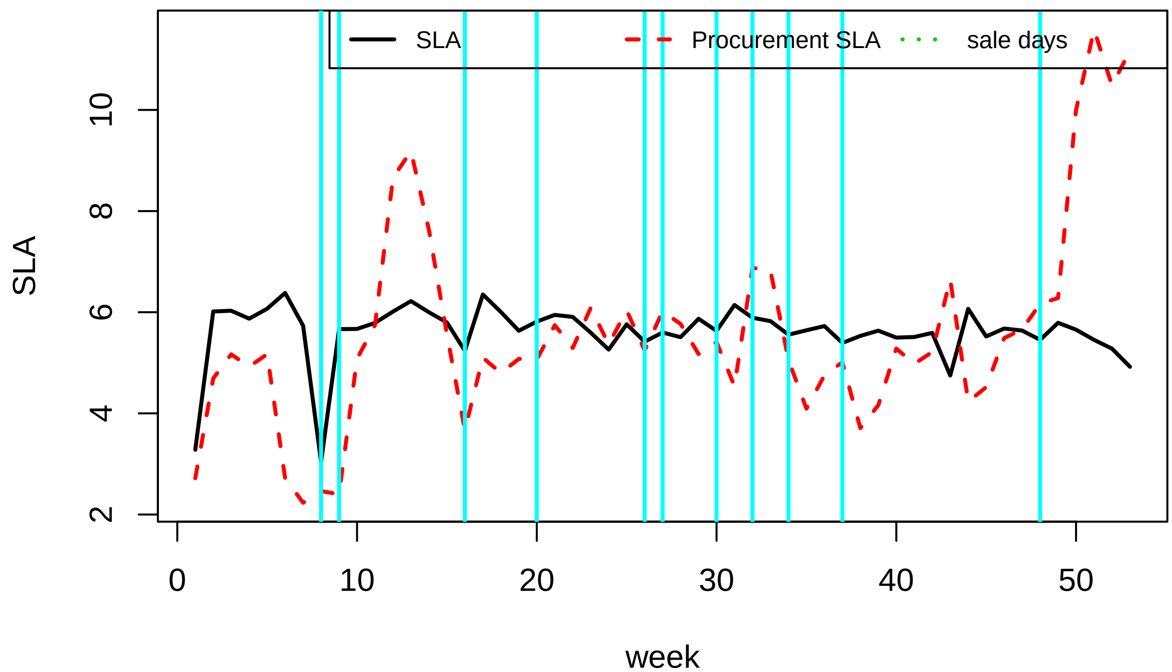
Insights:

Irrespective of promotions, majority of customers heavily utilizing `Cash On Delivery` facility consistently. we can say, there is a significant impact on the sales with this alternate payment method.

Service Level Aggrement:

`Service Level Aggrement` is a commitment to the customer to have his goods delivered in with a certain number of days. Below chart shows the weekly average `SLA` and `Procurement SLA`.

```
matplot(data_week$week, cbind(data_week$sla,data_week$procurement_sla), type = 'l',
        lwd = 2, xlab = 'week', ylab = 'SLA')
abline(v=saledays,col='cyan',lwd=2)
legend('topright', inset = 0, legend = c('SLA','Procurement SLA','sale days'),
       lty = c(1:3), lwd = 2, col=c(1:3), cex = 0.75, horiz = TRUE)
```



Insights:

`SLA` is kind of slightly higher than `Procurement SLA`, which is quiet good metric, so as, one has margin buffer. During longer promotion periods like week 16 and 20, `Procurement SLA` is lower compared to `SLA`. During weeks where no promotion running, `Procurement SLA` is higher, could this be factor, for lower sales during these periods.

Correlation:

Lets see how Non-Advertising KPIs were correlated

```
pairs.panels(data_week[-c(8:16)])
```

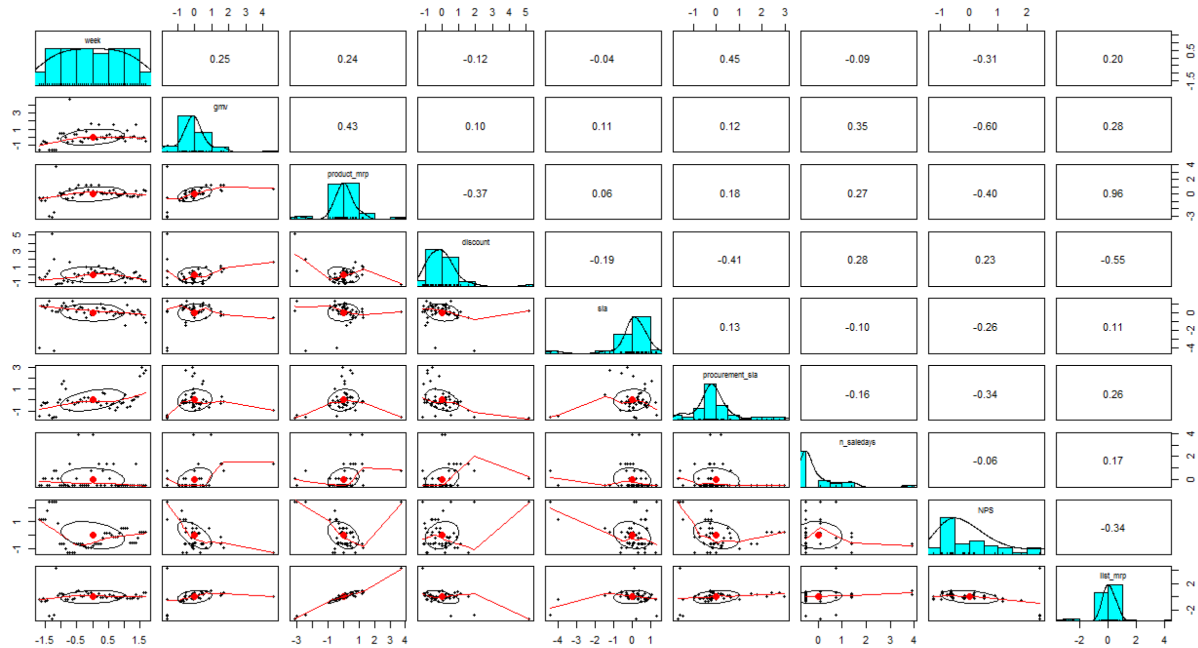


Figure 1:

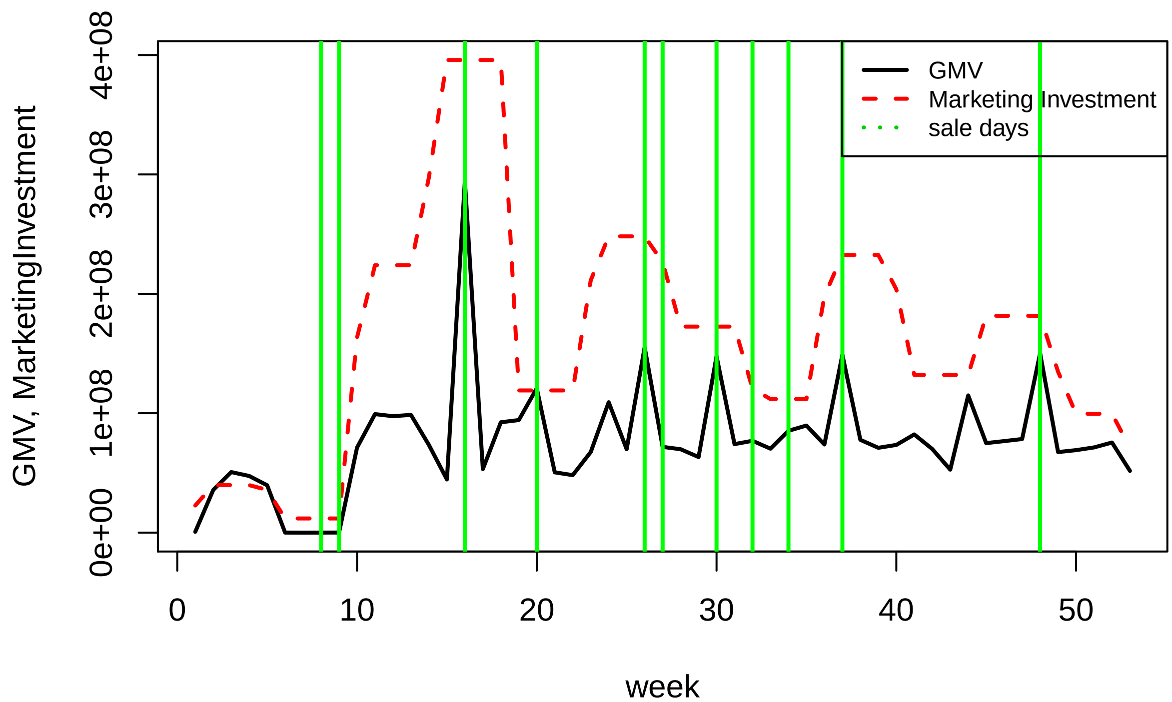
Insights:

`product_mrp` and `list_mrp` are highly correlated, this is kind of expected as we observed, for most of the weeks the range of `discount` offered on the products was with in narrow band.

GMV vs Advertising Spend:

Here is the comparision of weekly `GMV` against `Advertising` spend.

```
matplot(data_week$week, cbind(data_week$gmv, data_week$TotalInvestment),
        type='l',lwd = 2,xlab = 'week',ylab = 'GMV, MarketingInvestment')
abline(v=saledays,col='green',lwd=2)
legend('topright', inset = 0, legend = c('GMV','Marketing Investment','sale days'),
       lty = c(1:3), lwd = 2, col=c(1,2,3), cex = 0.75)
```



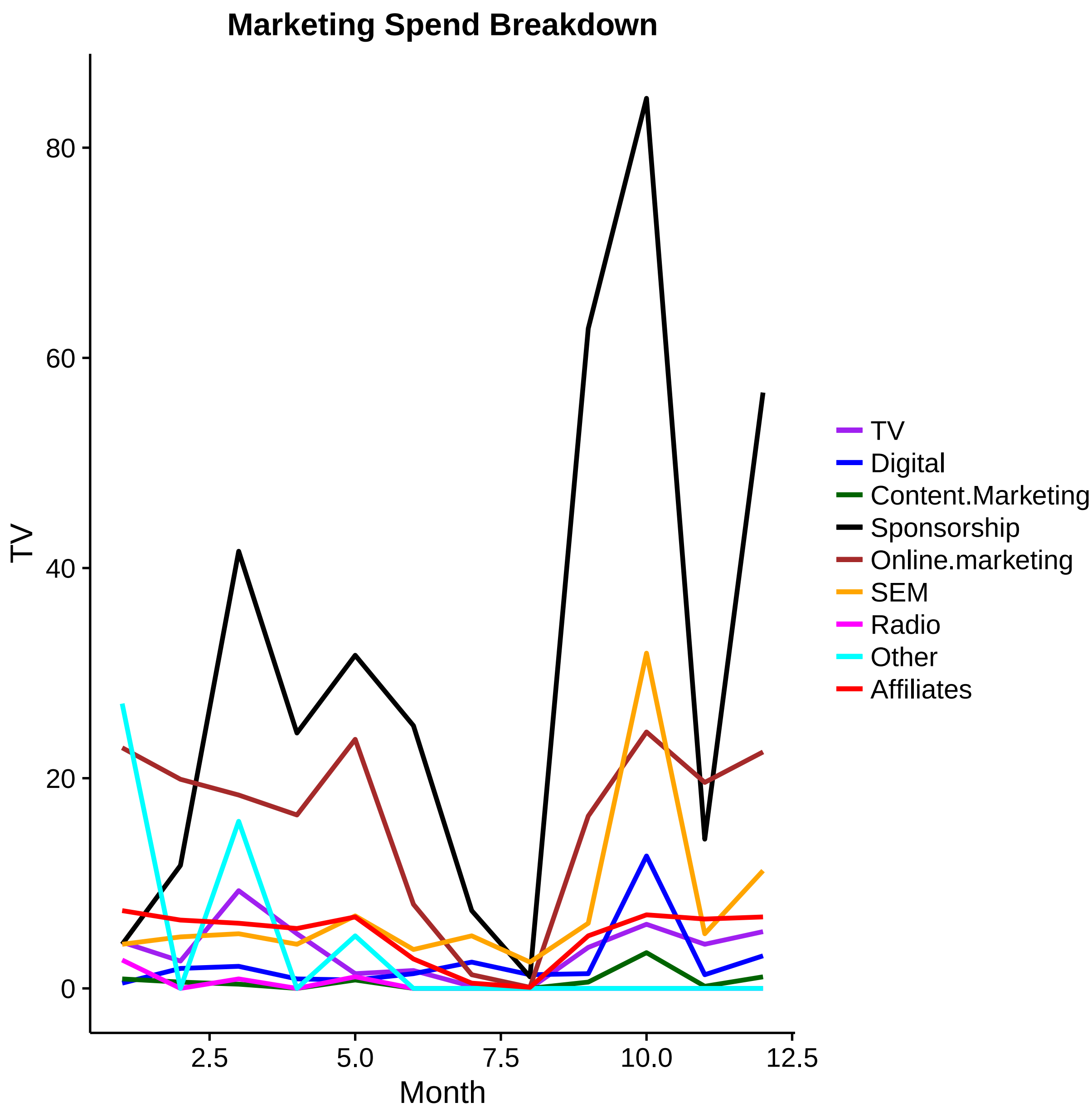
Insights:

Total weekly advertising amount spend was consistently high compared to `GMV` for all weeks. Well, either they were spending lot more on advertising than required, (or) one really need to spend higher than `GMV` possible.

Marketing Spend Across Channels:

Weekly aggregated marketing spend is confusing to draw clear insights, will analyze marketing spend at monthly level.

```
matplot(data_week$week, cbind(data_week$Digital, data_week$TV, data_week$Sponsorship,
                              data_week$ContentMarketing, data_week$OnlineMarketing,
                              data_week$Affiliates, data_week$SEM, data_week$Radio,
                              data_week$Other),
        type='l',lwd = 2,xlab = 'week',ylab = 'Marketing Spend')
legend('topright', inset = 0, legend = colnames(data_week[,c(8:16)]),
       lty = c(1:9), lwd = 2, col=c(1:9), cex = 0.5)
```

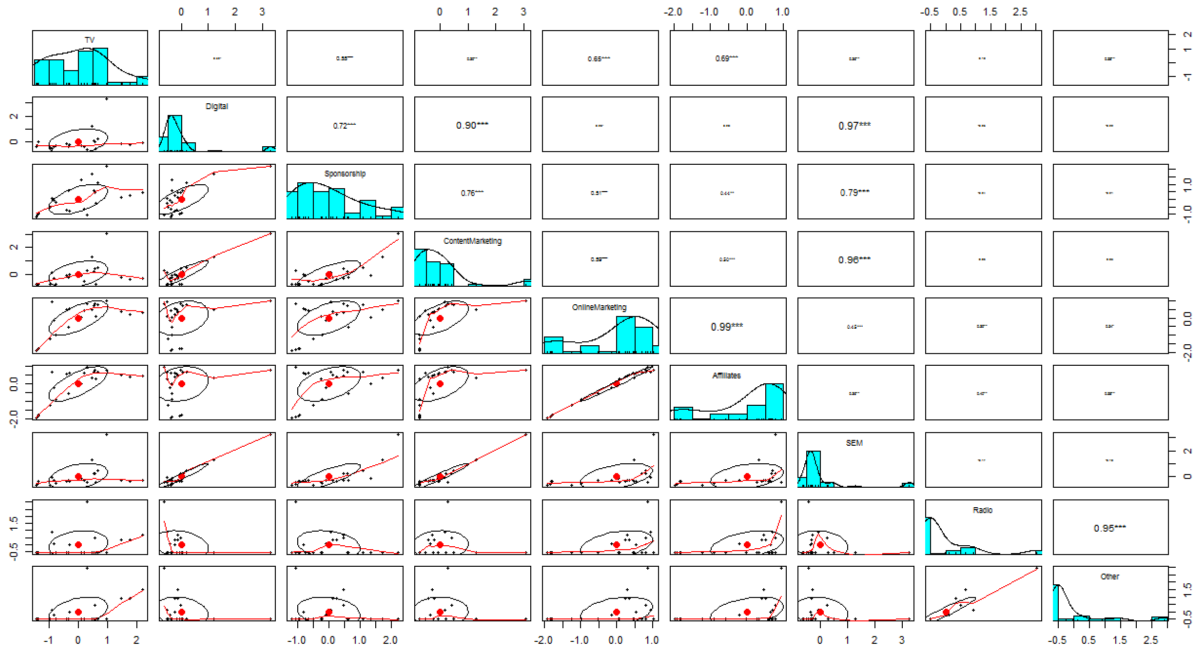


Insights:

A large portion of marketing spend goes towards sponsorships followed by online marketing. On annaul basis, `SEM`, `Digital` channels receives significant funding during month of October. `March`, `May` and `October` are the months when large portion of money spent on marketing (or) running promotions.

Correlation Among Advertising Channels:

```r
pairs.panels(data_week[8:16])
```



Insights:

There were multiple pairs of correlations exists among Advertising channel spends. 1. `Radio` and `Other` Advertising channel spends are highly correlated. 2. `SEM`, `Digital` and `Content Marketing` spends are highly correlated. 3. `Affiliates` and `Online Marketing` are highly correlated.