# Capstone Ecommerce

Initial understanding of the problem statement

## Problem Statement

Eleckart is an e-commerce firm specializing in online selling of electronic goods.

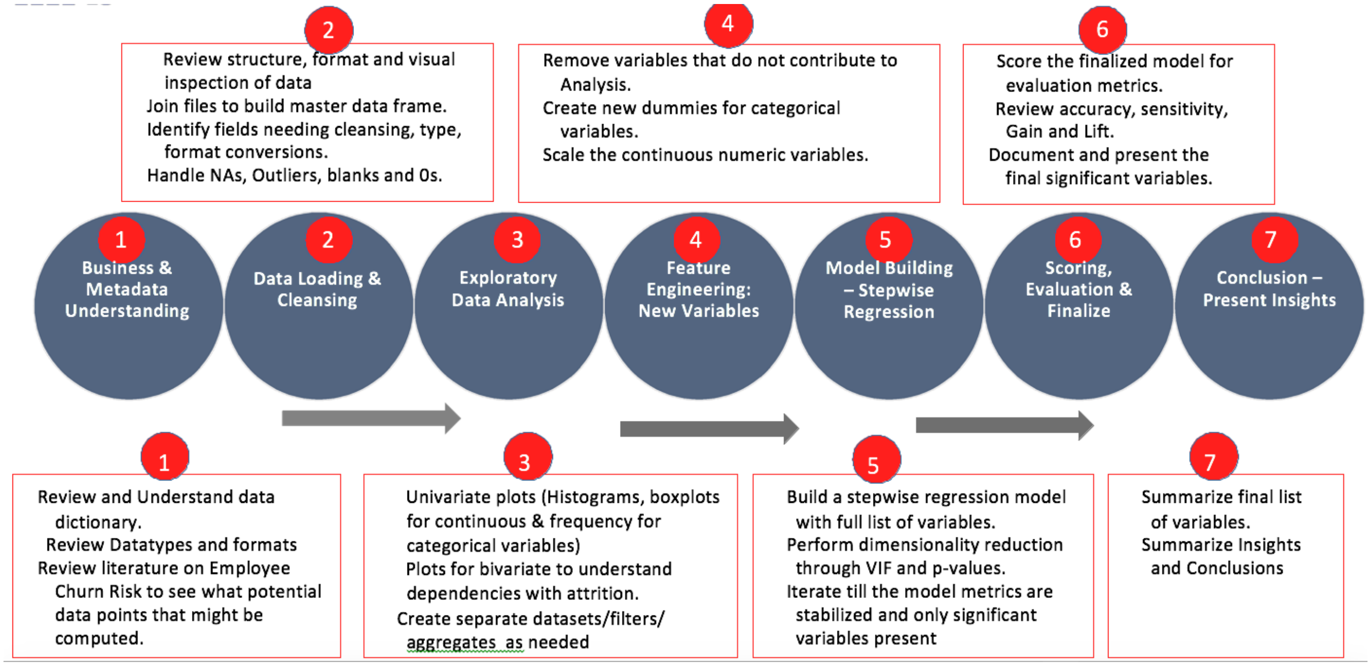
In the past year they have spent a significance amount of money on advertising their online store, which included giving offers on certain days.

CFO feels that the amount spent on advertising has not yielded sufficient results in terms of increased sale and market share.

Eleckart is in the process of charting advertising and marketing budget for the coming year and want to spend wisely this time.

They are thinking about cutting the budget of adverting where they feel return on spending is not enough and/or allocate the budget more optimally in areas where revenue response has been good and can be further improved.

## Solution Methodology



2
Review structure, format and visual inspection of data
Join files to build master data frame.
Identify fields needing cleansing, type, format conversions.
Handle NAs, Outliers, blanks and 0s.

4
Remove variables that do not contribute to Analysis.
Create new dummies for categorical variables.
Scale the continuous numeric variables.

6
Score the finalized model for evaluation metrics.
Review accuracy, sensitivity, Gain and Lift.
Document and present the final significant variables.

1
Business & Metadata Understanding
2
Data Loading & Cleansing
3
Exploratory Data Analysis
4
Feature Engineering: New Variables
5
Model Building – Stepwise Regression
6
Scoring, Evaluation & Finalize
7
Conclusion – Present Insights

1
Review and Understand data dictionary.
Review Datatypes and formats
Review literature on Employee Churn Risk to see what potential data points that might be computed.

3
Univariate plots (Histograms, boxplots for continuous & frequency for categorical variables)
Plots for bivariate to understand dependencies with attrition.
Create separate datasets/filters/ aggregates as needed

5
Build a stepwise regression model with full list of variables.
Perform dimensionality reduction through VIF and p-values.
Iterate till the model metrics are stabilized and only significant variables present

7
Summarize final list of variables.
Summarize Insights and Conclusions

## 1. Business Understanding and Data Available

**Customer Transaction data:**

We have data over 1 year sales from July 2015 to June 2016, over which we have every day sales of the electronic items across different SKUs. These data have income from sales, price and number of units ordered, SLA of the orders etc. Below is in detail explanation of each field

Attributes that describe a data point in the sales are as follows:

1. FSN_ID - Unique identifier for a brand of goods. It's an SKU.
2. ORDER_DT - Time stamped date on which the order was placed.
3. ORDER_ID - A system generated order identifier, uniquely identifying an order.
4. ORDER_ITEM_ID - Identifier for items in the order. If an order consists of more than one item, then each item in the order gets the same ORDER_ID but different ORDER_ITEM_ID.
5. GMV: Total value of the goods sold. This is essentially selling price*units sold.
6. UNITS: Number of units of a particular SKU sold in that order.
7. DELIVERYBDAYS: Delivery days from warehouse to distribution
8. DELIVERYCDAYS: Delivery days from distribution to actual end customer
9. ORDER_PAYMENT_TYPE: Type of payment. Either COD or Prepaid.
10. SLA: Number of days it typically takes to deliver the product
11. CUST_ID: Unique customer id of the customer who made the purchase.
12. PINCODE: Pin code from where the purchase made.

Apart from the above attributes, there are attributes provided which categorize the items into these hierarchical categories:

1. PRODUCT_ANALYTIC_SUPER_CATEGORY
2. PRODUCT_ANALYTIC_CATEGORY
3. PRODUCT_ANALYTIC_SUB_CATEGORY
4. PRODUCT_ANALYTIC_VERTICAL

**Montly Advertisment Spendings**

Month wise spend under various advertising channels. Namely:

TV
Digital Sponsorship
Content Marketing
Online marketing
Affiliates
SEM
Radio
Other

**Special Sale days**

Occasions on which special sale was held. These occasions can spread over multiple days.

**Monthly NPS score**

Monthly Net Promoter Score(NPS) score. This is a measure of brand loyalty of Eleckart.

## Using the data for analysis

As mentioned above, we have 1.6 million data points to do our analytical study of how Eleckart has been spending their advertising budget and how effective or ineffective it has been in terms of growth in revenue/sales. Aim of this project is to find a relation between the advertising spend via various channels and increase in sales of various category of goods. Once the relation has been established, we should be able to make some recommendations to the management on:

1. Where the advertising spending has been effective and should be continued or increased.
2. Where the advertising has not been very effective and should be cut.

To keep things simple and effective, the analysis will done for product sub categories of camera accessory, home audio and gaming accessory. The granularity of analysis will be weekly i.e. the comparison of growth/decline of any kind (sales, advertising spend etc.) will be on week on week basis.

## 2. Data Cleaning

1. NA and NULL treatment
   a. We see that 4904 rows of data do not have GMV values. So we are chose to ignore them as it is about .4% (4904 of 1.6 million) of data.
   b. We also observe that, there are rows which have MRP as Zero rupees. We chose to ignore these rows as these might be data quality issues or freebies.
   c. \N and negative values in delivery days for both deliverybdays and deliverycdays columns are set to zero.

2. Treatment of data outside the analysis window
   We are supposed to work with data from July 2015 to June 2016. So, we will remove the records that do not fit this criteria.

3. Treatment of negative values in SLA
   Negative values in SLA are set to 0

4. Treatment of outliers
   a. We see that 99.9% of the data do not have service level agreement (SLA column in data) of more than 16 days. Hence, we capped SLA at 16 days.
   b. Similarly, we see that 99.7% of product procurement SLA (product_procurement_sla column in data) values are less or equal to 15days, so we will cap the values at 15 days.
   c. In the units ordered column, we see that there are very few items (0.092%) that have been ordered in higher quantities (>4). We considered them as outliers therefore we shall drop them.

5. Checking for correlation
   Also we want to check if there is any correlation between procurement SLA and SLA. But there was no correlation observed and we concluded that these are individual metrics and are not influenced by one another.

**2. Feature engineering:**

***<u>Extract Week from date timestamp</u>***

1. Since we need to do analysis weekly, we extracted week from data across july 2015 to june 2016. Also we extracted day of the week to check if day has impact on revenue.

2. This week also helps is merging advertisement data fame with consumer sales dataframe.

***<u>Created deliveryDelay column as sum of deliverybdays and deliverycdays values</u>***

Created deliveryDelay as number of days required to deliver the order to customer. This is combination of deliverybdays (time taken to reach warehouse) and deliverycdays (time taken reach customer premise). So, now since max delay is 13 days, we have taken 7 days as cutoff. Converted to categorical variable where delay is < 7, we treat 0 as no delay and >7 as 1 delayed.

***<u>Created Binning for discount promotions</u>***

Calculate list price of product as gmv /units sold. We can now calculate discount as (product mrp – list price), this discount percentage will have impact on sales. So now we tried to bin this discount range into 5 different ranges.

NoProfitDiscount – when discount is < 0

lessThan25pcDiscount - when discount >0 and < 25 %

25to50pcDiscount – when discount >25 and <50%

50to75pcDiscount – when discount is >50 and < 75%

75to100pcDiscount – when discount is >75 and < 100%

***<u>Created expectedSla column as sum of sla and product_procurement_sla values</u>***

Customer when ordered product, SLA will be provided. Based on this sla customer will perceive something about delivery of the product. So here we have total SLA is sum of (SLA + product procurement SLA). We now need to make this SLA into bins of different range for ease analysis. We have maximum expected SLA as 16days

fastDeliveryPerception – if SLA > 0 and Less than 5 days
tolerableDeliveryPerception – if SLA > 5 and less than 10 days
delayedDeliveryPerception - if SLA > 10 days

### ***Created holidays per week based on promotional data***

There promotional weeks like rakshabandan sale, Diwali sale and new year sale etc. These are the special days in the week which should be treated differently. So we calculate them as number of special days(holidays) in week.

### ***Create P_tag and group the products using K means***

We have different range of products types among different sub categories under product verticals. Not all products under sub category have same impact on sales. There will be mass products like fast moving goods, some premium products and some average sold products. We decided to category products based on list price, units sold and product type from vertical. We give outcome as 3 clusters for k means algorithm so that we can get 3different categories

p_tags : mass moving , premium and middle

### ***Create incremental Moving Average for list price and discount***

Moving average gives average overall trend in data, so we can calculate it for list price and discount to see their trend and impact on sales. Here we chose to calculate moving average for 1,2,3 weeks. Now we calculate incremental moving average by subtracting it from base.

For example: (list price – list price with MA with 1week delay) / (listprice withdelay)

inc_LP_MA1, inc_LP_MA2, inc_LP_MA3

inc_PO_MA1, inc_PO_MA2, inc_PO_MA3

### ***Create incremental Lag for list price and discount***

We create lag variable to check impact of current price on previous price. Here we planned to create lag for both list price by 1, 2 and 3 weeks.

LP_1week, LP_2week, LP_3week

### ***Building adstock data for different advertisement channels***

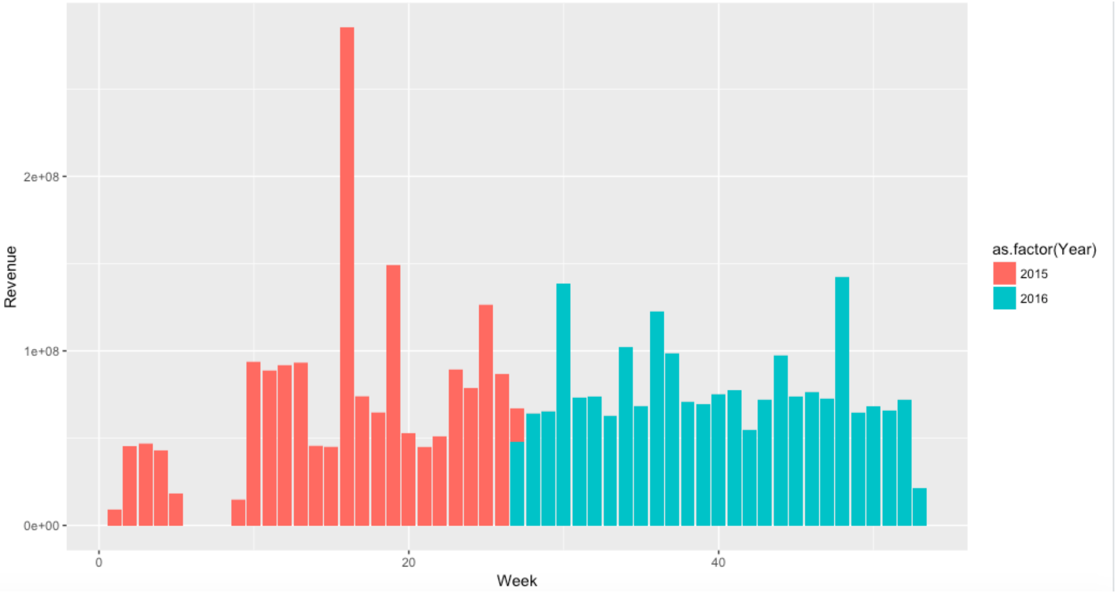
Divide the expenditure on advertisement for a channel evenly across 4 or 5 weeks depending on how many weeks are in that month. Consolidated sheet of advertisement spent across all channels is prepared per week of the period (week number) under consideration (July 2015 - June 2016).

To calculate adstock, assumption has been made that adstock rate is 0.5. i.e. for an advertisement broadcast in a week, in the following week only 50% of impact remains.

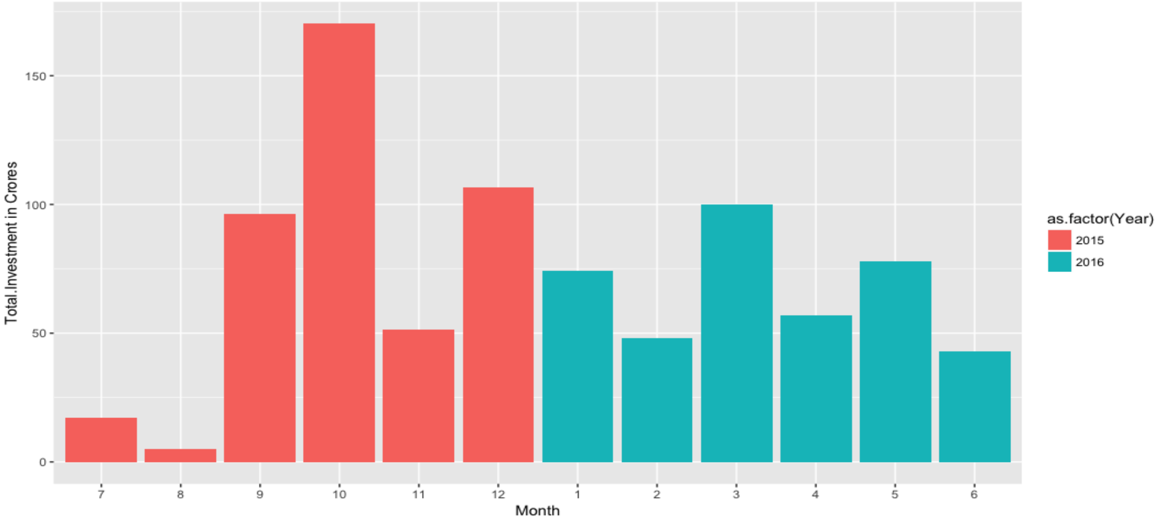
Merging the adstock data with sales data on week number (week of the year) column engineered in from order date.

***Exploratory Data Analysis – Uni-varient and Bi-variant Analysis:***

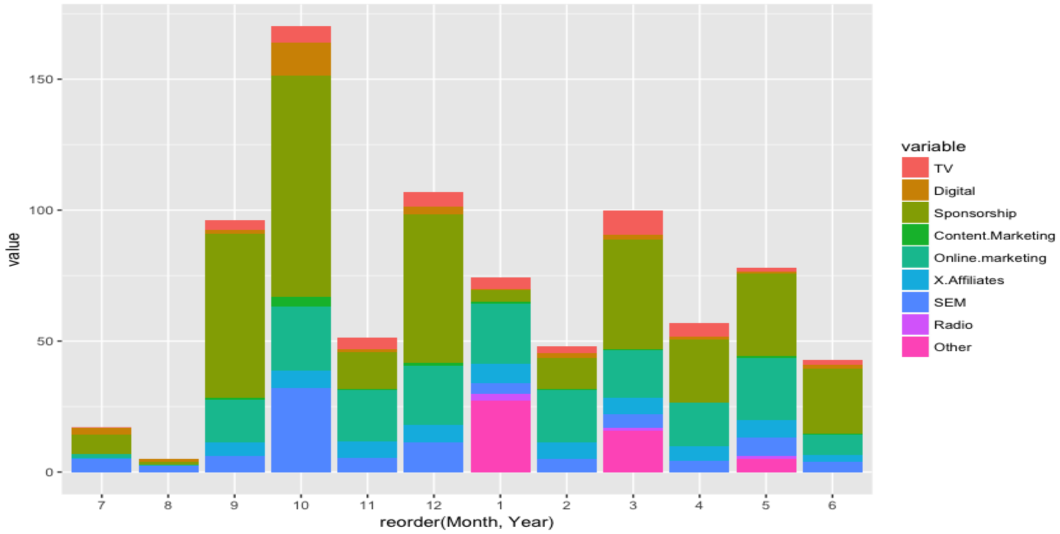
1. Variation of weekly total revenue of the company.



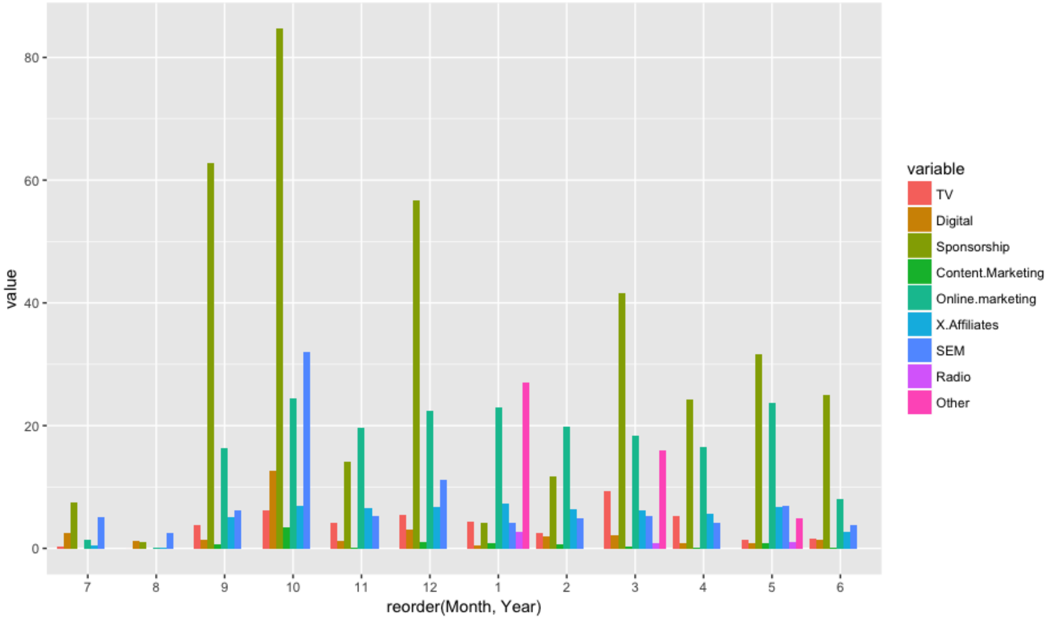
2. Total monthly advertising investments. As we see October has highest investments, this may be because of special sale weeks in that month they spent more for promotions



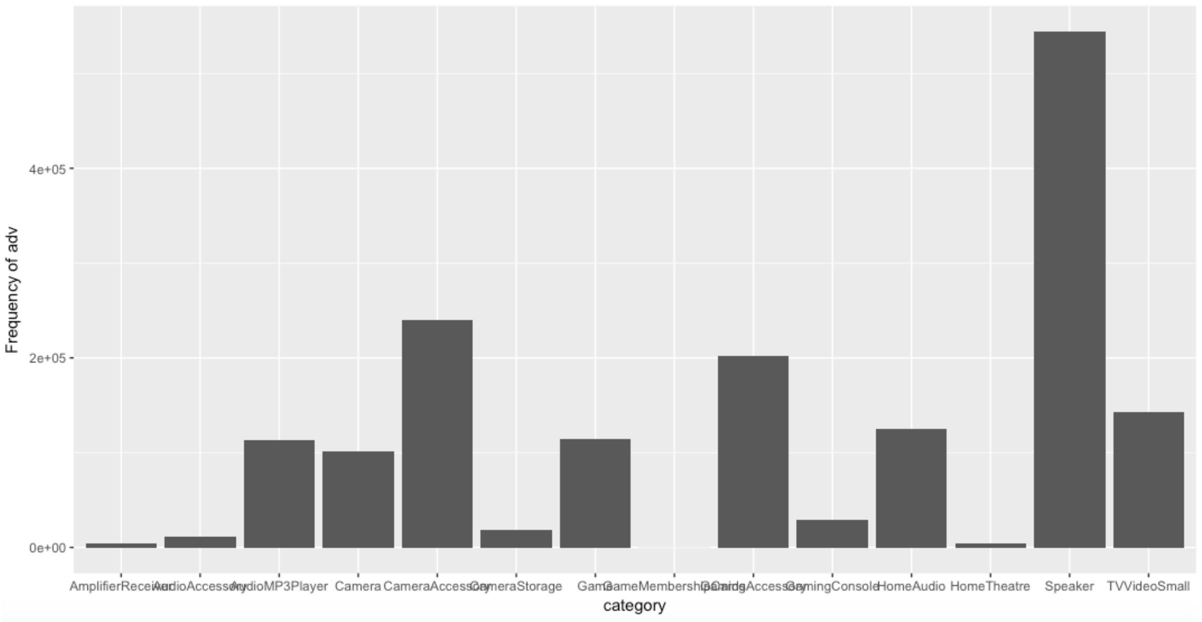
3. Proportional advertising spending's for each channel.
   As you see October month has highest advertisement spendings



dodge modal- we can clearly see sponsorship has highest investments.

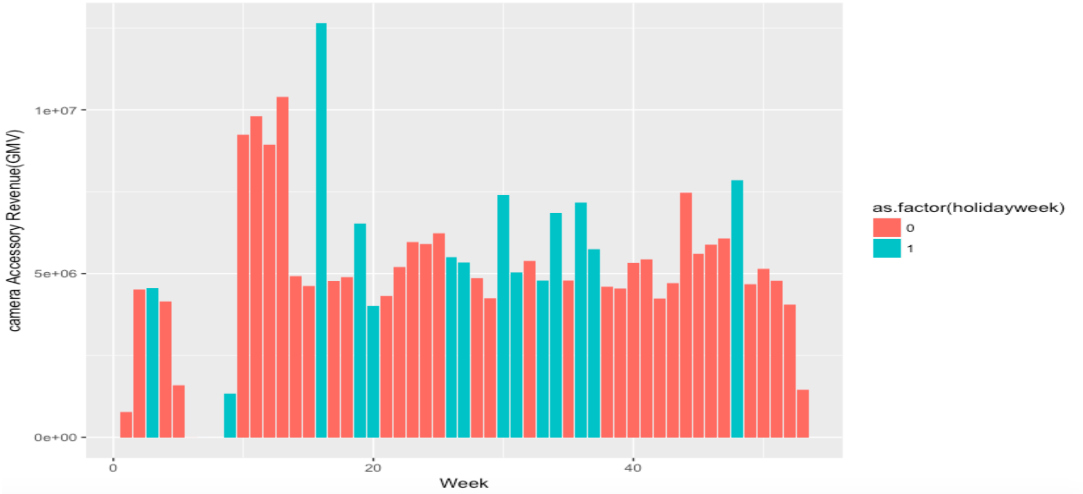


4. Frequency of advertisements based on sub category – clearly speakers have highest frequency of advertisements.

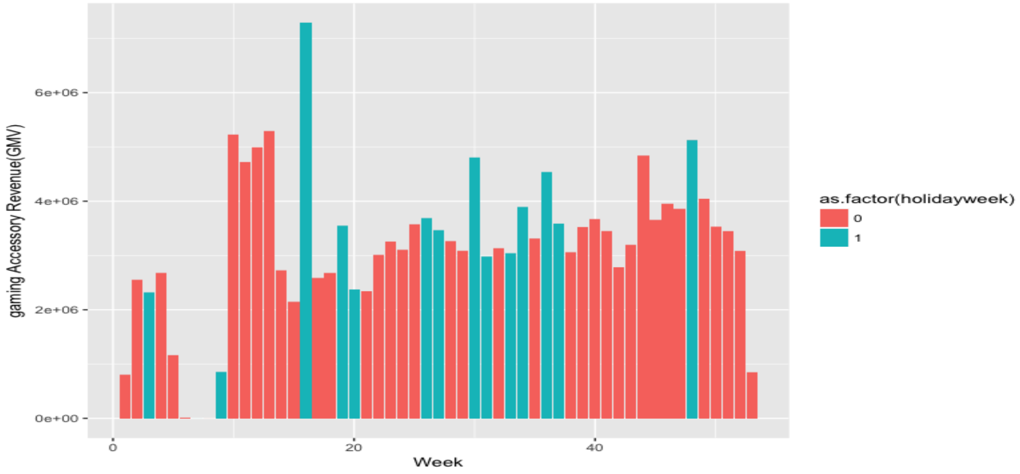


***Subcategory wise revenue***

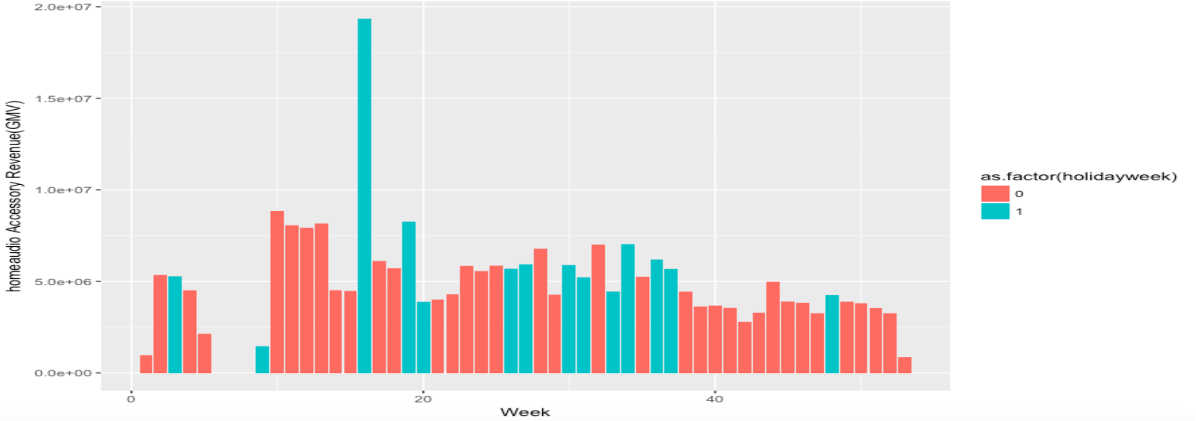
a. Total revenue for camera accessory, clearly on weekly holiday special sale revenue is more.



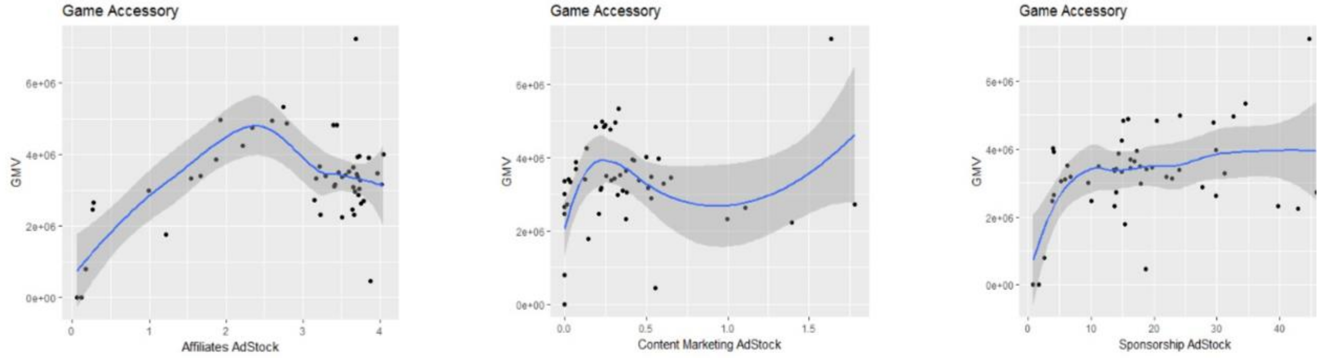
b. Gaming Accessory revenue, clearly it is high during special sale.



c.Home accessory revenue also grew much in holiday weeks.



Adstock effect on GMV for each sub category



***Preparing data for Modelling***

Now we can model with only required columns. All columns will not haveimpact on sales and revenue. We deleted few of the columns which are not for modelling. These columns are already featured onto another derived variables.

```
consumer_df$fsn_id <- NULL
consumer_df$order_date <- NULL
consumer_df$Year <-NULL
consumer_df$Month <- NULL
consumer_df$order_id<- NULL
consumer_df$order_item_id <-NULL
consumer_df$deliverybdays <- NULL
consumer_df$deliverycdays <- NULL
consumer_df$product_analytic_category <- NULL
consumer_df$product_analytic_super_category<- NULL
consumer_df$sla<-NULL
```

Factor columns in data frame for modelling need to be converted to dummy variables. Such columns are:

p_tag
deliveryPerception
promotion_range
dayOfweek
holidays

## 4. Modelling

We now should create model for three sub-categories namely:

1. Camera Accessory
2. Gaming Accessory
3. HomeAudio

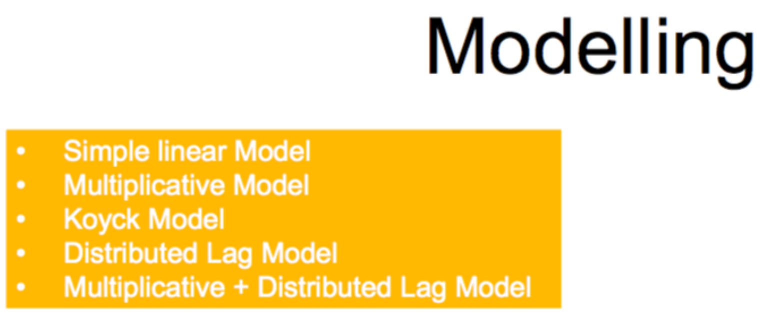
So we create 3 subsets of data from main data frame and divide the data into 7:3 ratio for training and testing.

Engineered KPIs:

So we have engineered new features, and based on that the KPIs we will base our model on are

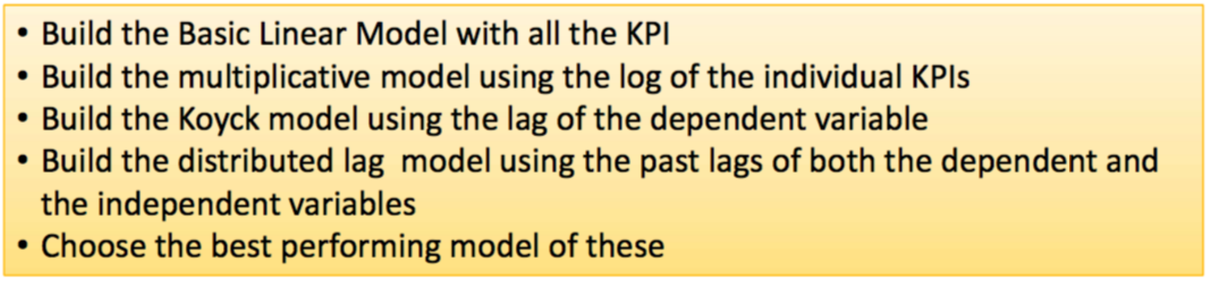
1. Adstock for each channel of advertisement spending.
2. Incremental Lag for price and promotions (discounts)
3. Payment Type
4. Holiday (promotional sale)
5. Moving average price impact
6. Product category

Using the lm function in R, we define predicted columns as GMV which is to be predicted using all remaining attributes. Once the model is built we see the summary and try to eliminate the least significant attributes (high P-value) and most co-related (High VIF).



Modelling

- Simple linear Model
- Multiplicative Model
- Koyck Model
- Distributed Lag Model
- Multiplicative + Distributed Lag Model



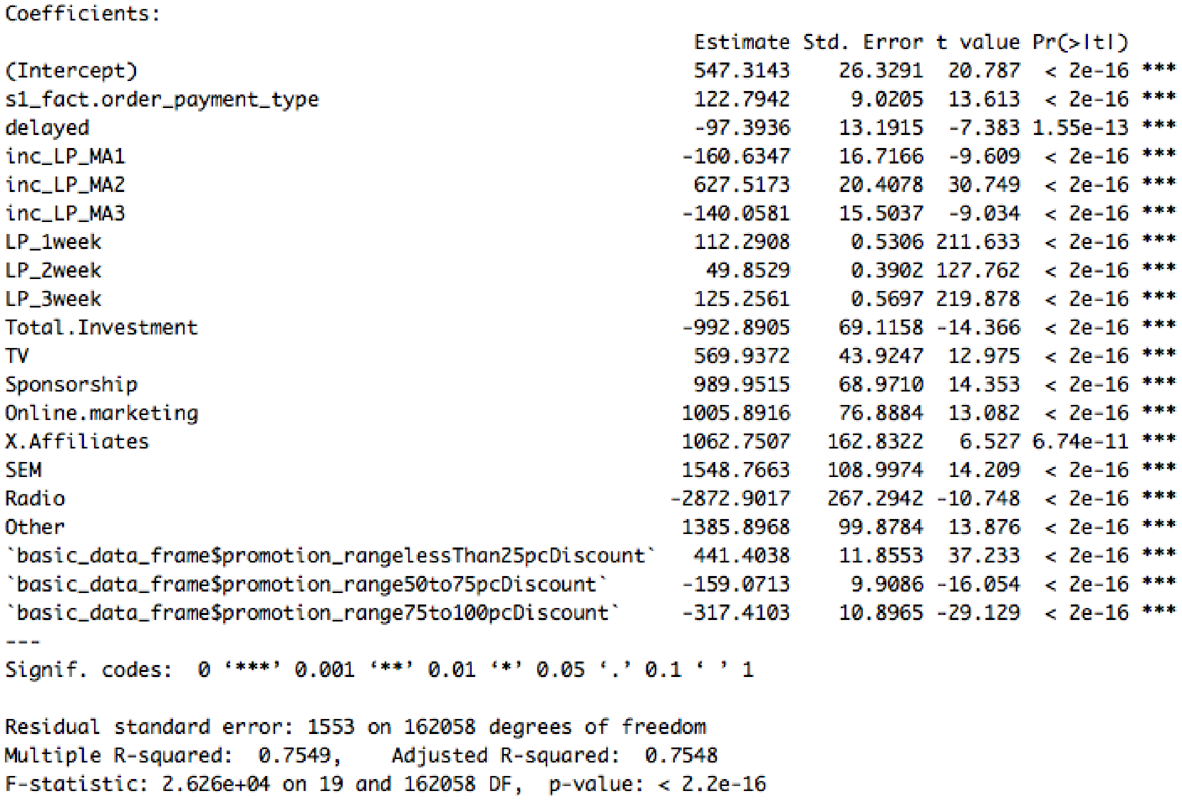
- Build the Basic Linear Model with all the KPI
- Build the multiplicative model using the log of the individual KPIs
- Build the Koyck model using the lag of the dependent variable
- Build the distributed lag model using the past lags of both the dependent and the independent variables
- Choose the best performing model of these

# Basic Linear Regression

$$Y = \alpha + \beta_1 A_t + \beta_2 P_t + \beta_3 D_t + \beta_4 Q_t + \beta_5 T_t + \epsilon$$

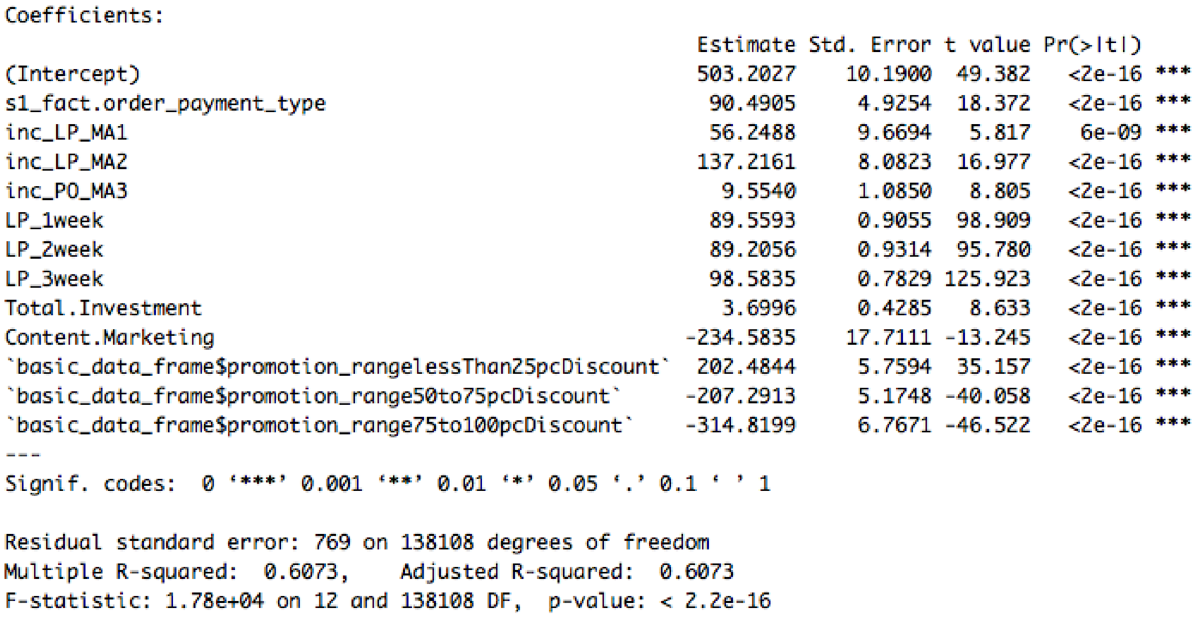
- A = KPIs related to Advertising
- P = KPIs related to pricing
- D = KPIs related to Promotions/Discounts
- Q = KPIs related to Product Assortment
- T = KPIs related to industry trend, seasonality etc

**Linear Regression Model- Camera Accessories**



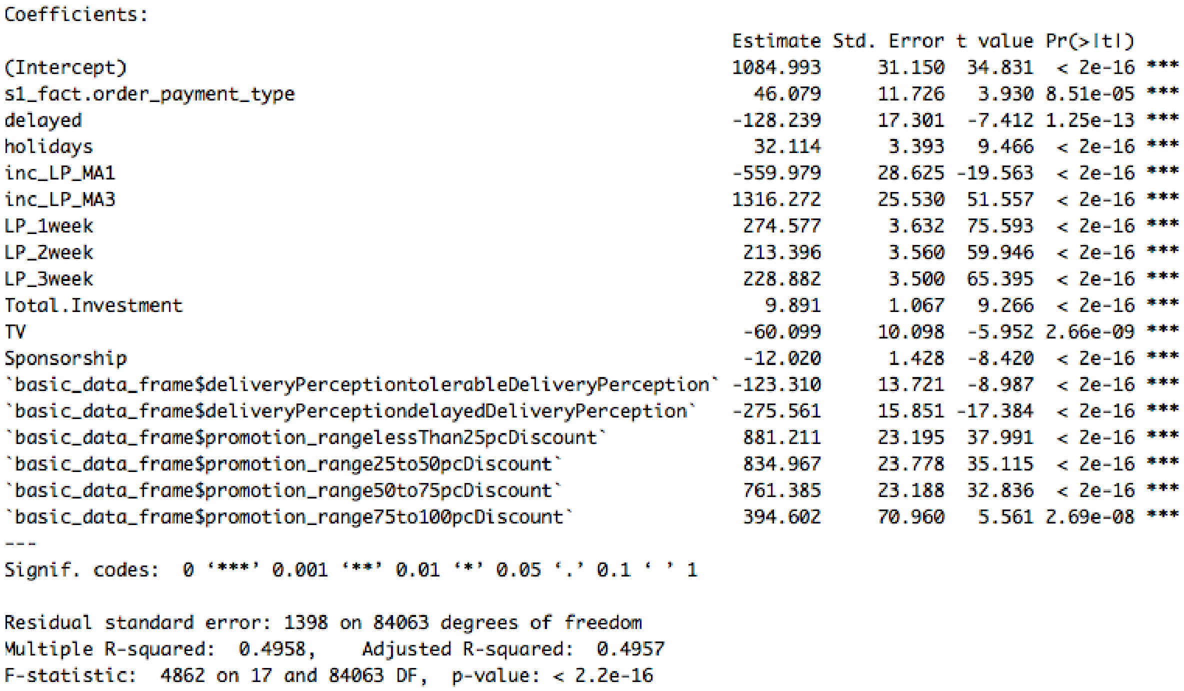
```
Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                               547.3143    26.3291  20.787  < 2e-16 ***
s1_fact.order_payment_type                                122.7942     9.0205  13.613  < 2e-16 ***
delayed                                                   -97.3936    13.1915  -7.383 1.55e-13 ***
inc_LP_MA1                                               -160.6347    16.7166  -9.609  < 2e-16 ***
inc_LP_MA2                                                627.5173    20.4078  30.749  < 2e-16 ***
inc_LP_MA3                                               -140.0581    15.5037  -9.034  < 2e-16 ***
LP_1week                                                  112.2908     0.5306 211.633  < 2e-16 ***
LP_2week                                                   49.8529     0.3902 127.762  < 2e-16 ***
LP_3week                                                  125.2561     0.5697 219.878  < 2e-16 ***
Total.Investment                                         -992.8905    69.1158 -14.366  < 2e-16 ***
TV                                                        569.9372    43.9247  12.975  < 2e-16 ***
Sponsorship                                               989.9515    68.9710  14.353  < 2e-16 ***
Online.marketing                                         1005.8916    76.8884  13.082  < 2e-16 ***
X.Affiliates                                             1062.7507   162.8322   6.527 6.74e-11 ***
SEM                                                      1548.7663   108.9974  14.209  < 2e-16 ***
Radio                                                   -2872.9017   267.2942 -10.748  < 2e-16 ***
Other                                                    1385.8968    99.8784  13.876  < 2e-16 ***
`basic_data_frame$promotion_rangelessThan25pcDiscount`    441.4038    11.8553  37.233  < 2e-16 ***
`basic_data_frame$promotion_range50to75pcDiscount`       -159.0713     9.9086 -16.054  < 2e-16 ***
`basic_data_frame$promotion_range75to100pcDiscount`      -317.4103    10.8965 -29.129  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1553 on 162058 degrees of freedom
Multiple R-squared:  0.7549,	Adjusted R-squared:  0.7548 
F-statistic: 2.626e+04 on 19 and 162058 DF,  p-value: < 2.2e-16
```

**Linear Regression Model – Gamming**

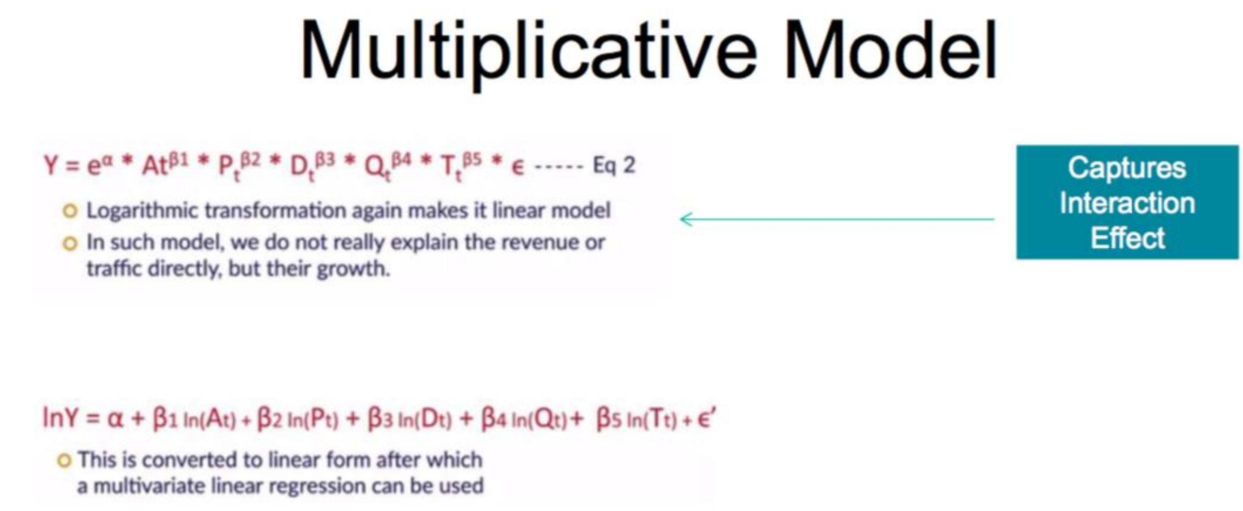


```
Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                             503.2027    10.1900  49.382   <2e-16 ***
s1_fact.order_payment_type                               90.4905     4.9254  18.372   <2e-16 ***
inc_LP_MA1                                               56.2488     9.6694   5.817    6e-09 ***
inc_LP_MA2                                              137.2161     8.0823  16.977   <2e-16 ***
inc_PO_MA3                                                9.5540     1.0850   8.805   <2e-16 ***
LP_1week                                                 89.5593     0.9055  98.909   <2e-16 ***
LP_2week                                                 89.2056     0.9314  95.780   <2e-16 ***
LP_3week                                                 98.5835     0.7829 125.923   <2e-16 ***
Total.Investment                                          3.6996     0.4285   8.633   <2e-16 ***
Content.Marketing                                      -234.5835    17.7111 -13.245   <2e-16 ***
`basic_data_frame$promotion_rangelessThan25pcDiscount`  202.4844     5.7594  35.157   <2e-16 ***
`basic_data_frame$promotion_range50to75pcDiscount`     -207.2913     5.1748 -40.058   <2e-16 ***
`basic_data_frame$promotion_range75to100pcDiscount`    -314.8199     6.7671 -46.522   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 769 on 138108 degrees of freedom
Multiple R-squared:  0.6073,	Adjusted R-squared:  0.6073 
F-statistic: 1.78e+04 on 12 and 138108 DF,  p-value: < 2.2e-16
```

**Linear Regression Model - HomeAudio**



```
Coefficients:
                                                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                                      1084.993     31.150  34.831  < 2e-16 ***
s1_fact.order_payment_type                                         46.079     11.726   3.930 8.51e-05 ***
delayed                                                          -128.239     17.301  -7.412 1.25e-13 ***
holidays                                                           32.114      3.393   9.466  < 2e-16 ***
inc_LP_MA1                                                       -559.979     28.625 -19.563  < 2e-16 ***
inc_LP_MA3                                                       1316.272     25.530  51.557  < 2e-16 ***
LP_1week                                                          274.577      3.632  75.593  < 2e-16 ***
LP_2week                                                          213.396      3.560  59.946  < 2e-16 ***
LP_3week                                                          228.882      3.500  65.395  < 2e-16 ***
Total.Investment                                                    9.891      1.067   9.266  < 2e-16 ***
TV                                                                -60.099     10.098  -5.952 2.66e-09 ***
Sponsorship                                                       -12.020      1.428  -8.420  < 2e-16 ***
`basic_data_frame$deliveryPerceptiontolerableDeliveryPerception` -123.310     13.721  -8.987  < 2e-16 ***
`basic_data_frame$deliveryPerceptiondelayedDeliveryPerception`   -275.561     15.851 -17.384  < 2e-16 ***
`basic_data_frame$promotion_rangelessThan25pcDiscount`            881.211     23.195  37.991  < 2e-16 ***
`basic_data_frame$promotion_range25to50pcDiscount`                834.967     23.778  35.115  < 2e-16 ***
`basic_data_frame$promotion_range50to75pcDiscount`                761.385     23.188  32.836  < 2e-16 ***
`basic_data_frame$promotion_range75to100pcDiscount`               394.602     70.960   5.561 2.69e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1398 on 84063 degrees of freedom
Multiple R-squared:  0.4958,	Adjusted R-squared:  0.4957 
F-statistic:  4862 on 17 and 84063 DF,  p-value: < 2.2e-16
```



# Multiplicative Model

$Y = e^{\alpha} * A_t^{\beta 1} * P_t^{\beta 2} * D_t^{\beta 3} * Q_t^{\beta 4} * T_t^{\beta 5} * \epsilon$ ----- Eq 2

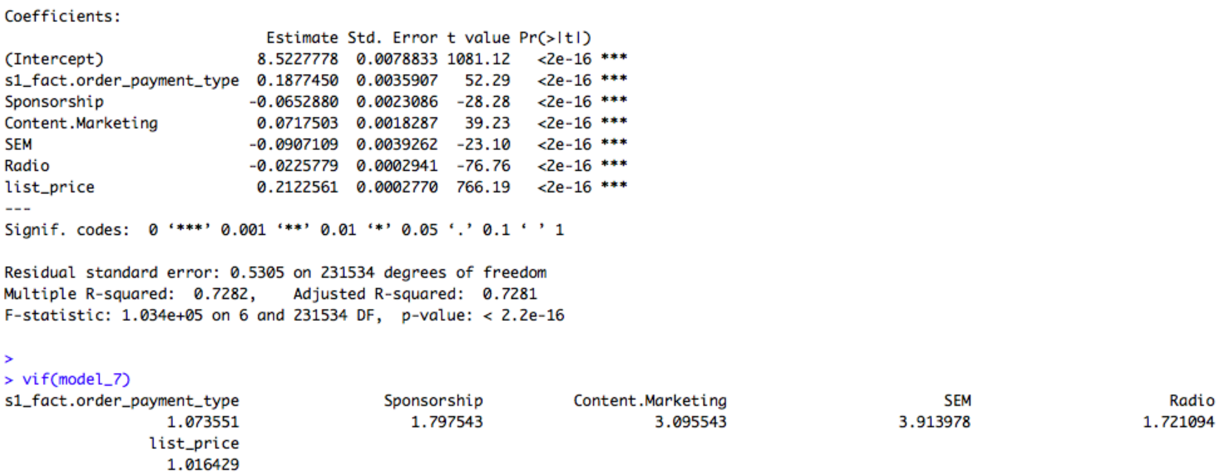
- Logarithmic transformation again makes it linear model
- In such model, we do not really explain the revenue or traffic directly, but their growth.

Captures Interaction Effect

$\ln Y = \alpha + \beta 1 \ln(A_t) + \beta 2 \ln(P_t) + \beta 3 \ln(D_t) + \beta 4 \ln(Q_t) + \beta 5 \ln(T_t) + \epsilon'$

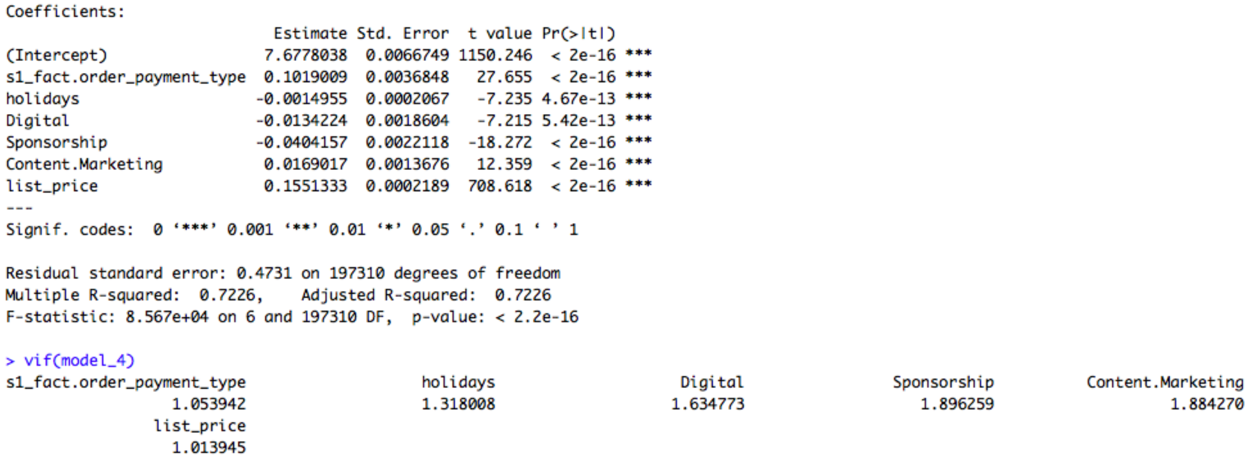
- This is converted to linear form after which a multivariate linear regression can be used

## Multiplicative Model - Camera



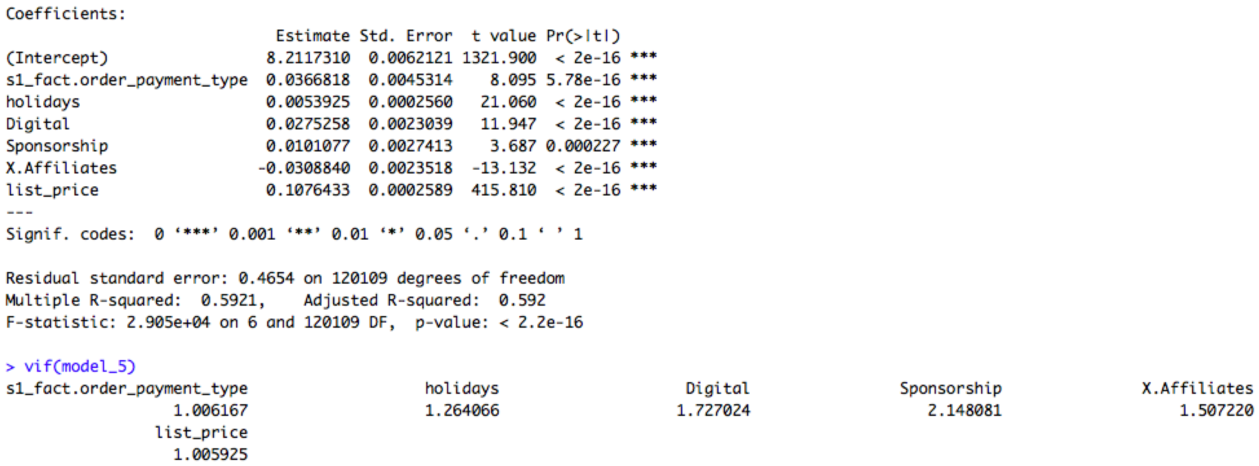
```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                8.5227778  0.0078833 1081.12   <2e-16 ***
s1_fact.order_payment_type 0.1877450  0.0035907   52.29   <2e-16 ***
Sponsorship               -0.0652880  0.0023086  -28.28   <2e-16 ***
Content.Marketing          0.0717503  0.0018287   39.23   <2e-16 ***
SEM                       -0.0907109  0.0039262  -23.10   <2e-16 ***
Radio                     -0.0225779  0.0002941  -76.76   <2e-16 ***
list_price                 0.2122561  0.0002770  766.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5305 on 231534 degrees of freedom
Multiple R-squared:  0.7282,    Adjusted R-squared:  0.7281 
F-statistic: 1.034e+05 on 6 and 231534 DF,  p-value: < 2.2e-16

> 
> vif(model_7)
s1_fact.order_payment_type                Sponsorship          Content.Marketing                        SEM                      Radio 
                  1.073551                   1.797543                   3.095543                   3.913978                   1.721094 
                list_price 
                  1.016429 
```

## Multiplicative Model -Gaming

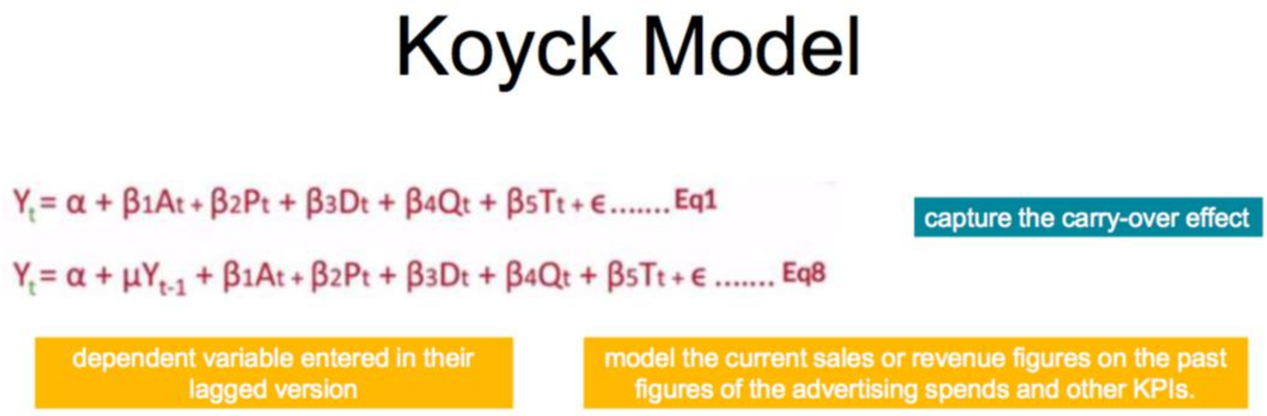


```
Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                 7.6778038  0.0066749 1150.246  < 2e-16 ***
s1_fact.order_payment_type  0.1019009  0.0036848   27.655  < 2e-16 ***
holidays                   -0.0014955  0.0002067   -7.235 4.67e-13 ***
Digital                    -0.0134224  0.0018604   -7.215 5.42e-13 ***
Sponsorship                -0.0404157  0.0022118  -18.272  < 2e-16 ***
Content.Marketing           0.0169017  0.0013676   12.359  < 2e-16 ***
list_price                  0.1551333  0.0002189  708.618  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4731 on 197310 degrees of freedom
Multiple R-squared:  0.7226,    Adjusted R-squared:  0.7226 
F-statistic: 8.567e+04 on 6 and 197310 DF,  p-value: < 2.2e-16

> vif(model_4)
s1_fact.order_payment_type                   holidays                    Digital                Sponsorship          Content.Marketing 
                  1.053942                   1.318008                   1.634773                   1.896259                   1.884270 
                list_price 
                  1.013945 
```

### Multiplicative Model – Home Audio



```
Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)    
(Intercept)              8.2117310  0.0062121 1321.900  < 2e-16 ***
s1_fact.order_payment_type  0.0366818  0.0045314    8.095 5.78e-16 ***
holidays                 0.0053925  0.0002560   21.060  < 2e-16 ***
Digital                  0.0275258  0.0023039   11.947  < 2e-16 ***
Sponsorship              0.0101077  0.0027413    3.687 0.000227 ***
X.Affiliates            -0.0308840  0.0023518  -13.132  < 2e-16 ***
list_price               0.1076433  0.0002589  415.810  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4654 on 120109 degrees of freedom
Multiple R-squared:  0.5921,	Adjusted R-squared:  0.592 
F-statistic: 2.905e+04 on 6 and 120109 DF,  p-value: < 2.2e-16

> vif(model_5)
s1_fact.order_payment_type                   holidays                    Digital                Sponsorship               X.Affiliates 
                  1.006167                   1.264066                   1.727024                   2.148081                   1.507220 
                list_price 
                  1.005925 
```



Koyck Model

$Y_t = \alpha + \beta_1 A_t + \beta_2 P_t + \beta_3 D_t + \beta_4 Q_t + \beta_5 T_t + \epsilon$ .......Eq1

$Y_t = \alpha + \mu Y_{t-1} + \beta_1 A_t + \beta_2 P_t + \beta_3 D_t + \beta_4 Q_t + \beta_5 T_t + \epsilon$ ....... Eq8

capture the carry-over effect

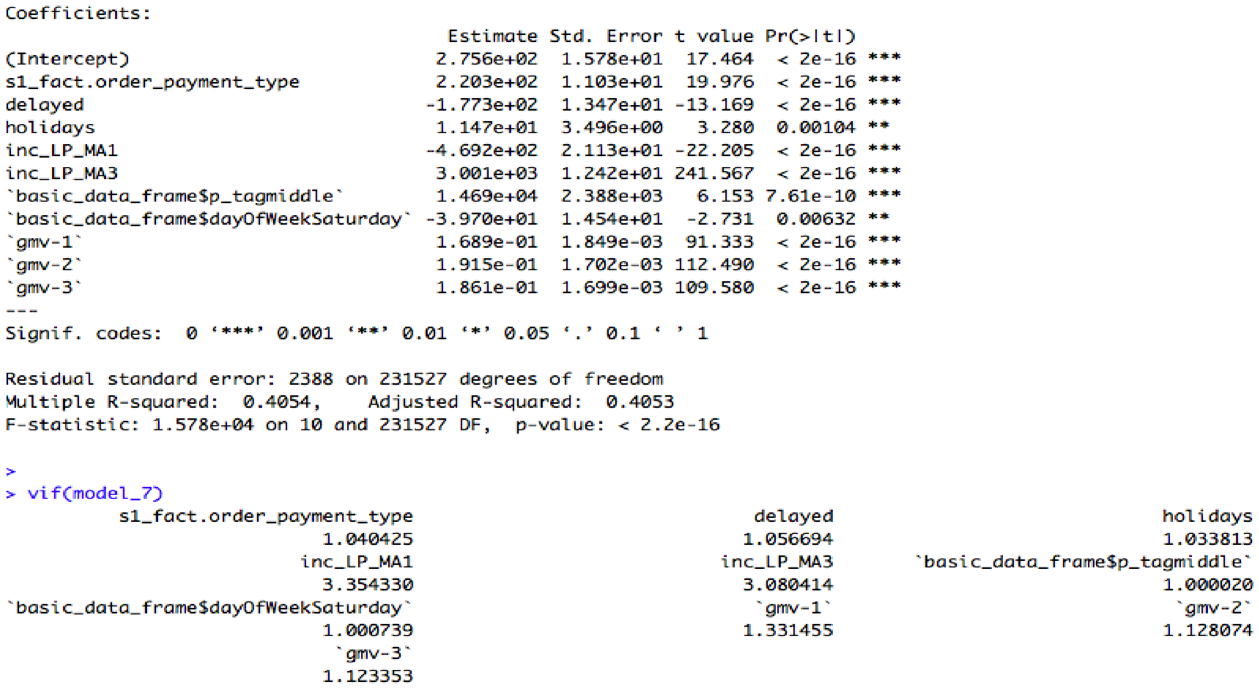
dependent variable entered in their lagged version

model the current sales or revenue figures on the past figures of the advertising spends and other KPIs.

In koyck model we create lag variables for dependent variables, here for gmv. We create lag for 3 weeks.  In some scenarios the sale of past weeks also have impact on this week sale. So we consider that for building the model.

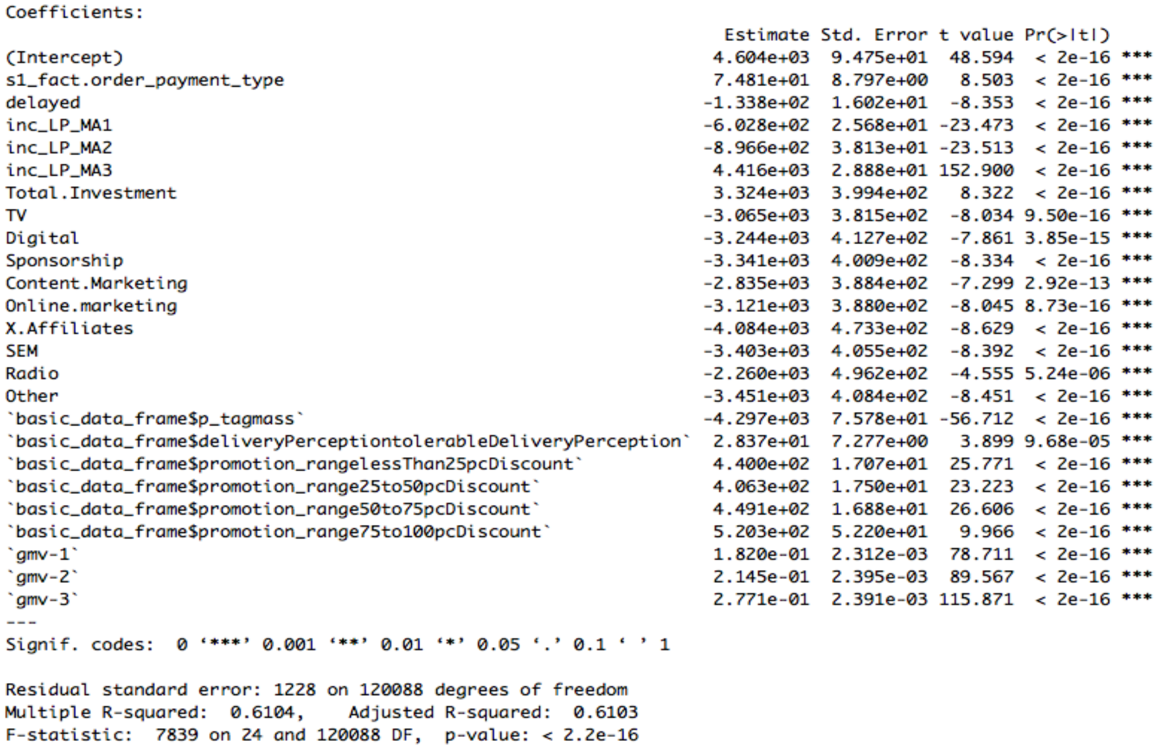
In distributed lag model, along with dependent variables we do create lag for independent variables too  like list price and discount offered.

## Koyck Model – Camera category



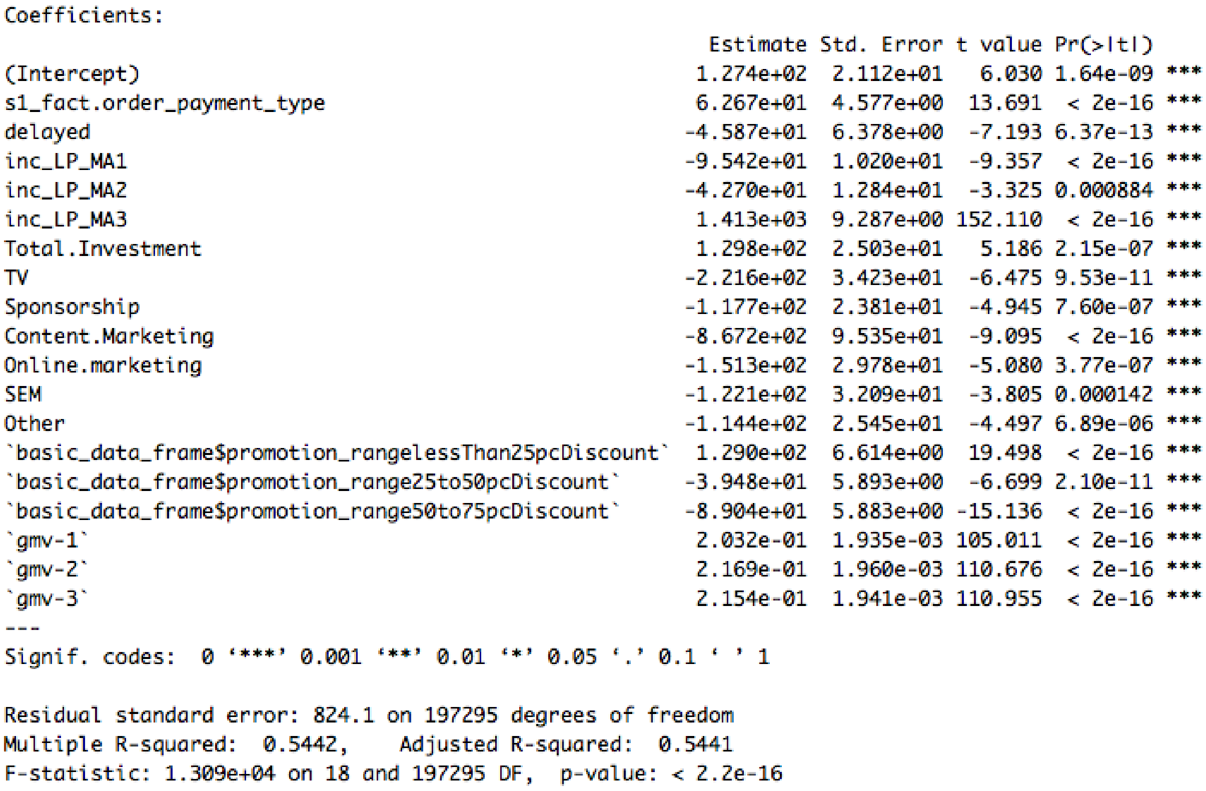
```
Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            2.756e+02  1.578e+01  17.464  < 2e-16 ***
s1_fact.order_payment_type             2.203e+02  1.103e+01  19.976  < 2e-16 ***
delayed                               -1.773e+02  1.347e+01 -13.169  < 2e-16 ***
holidays                               1.147e+01  3.496e+00   3.280  0.00104 ** 
inc_LP_MA1                            -4.692e+02  2.113e+01 -22.205  < 2e-16 ***
inc_LP_MA3                             3.001e+03  1.242e+01 241.567  < 2e-16 ***
`basic_data_frame$p_tagmiddle`         1.469e+04  2.388e+03   6.153 7.61e-10 ***
`basic_data_frame$dayOfWeekSaturday` -3.970e+01  1.454e+01  -2.731  0.00632 ** 
`gmv-1`                                1.689e-01  1.849e-03  91.333  < 2e-16 ***
`gmv-2`                                1.915e-01  1.702e-03 112.490  < 2e-16 ***
`gmv-3`                                1.861e-01  1.699e-03 109.580  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2388 on 231527 degrees of freedom
Multiple R-squared:  0.4054,	Adjusted R-squared:  0.4053 
F-statistic: 1.578e+04 on 10 and 231527 DF,  p-value: < 2.2e-16

> 
> vif(model_7)
          s1_fact.order_payment_type                              delayed                             holidays 
                            1.040425                             1.056694                             1.033813 
                          inc_LP_MA1                           inc_LP_MA3       `basic_data_frame$p_tagmiddle` 
                            3.354330                             3.080414                             1.000020 
`basic_data_frame$dayOfWeekSaturday`                              `gmv-1`                              `gmv-2` 
                            1.000739                             1.331455                             1.128074 
                             `gmv-3` 
                            1.123353 
```

## Koyck Model - Home Audio

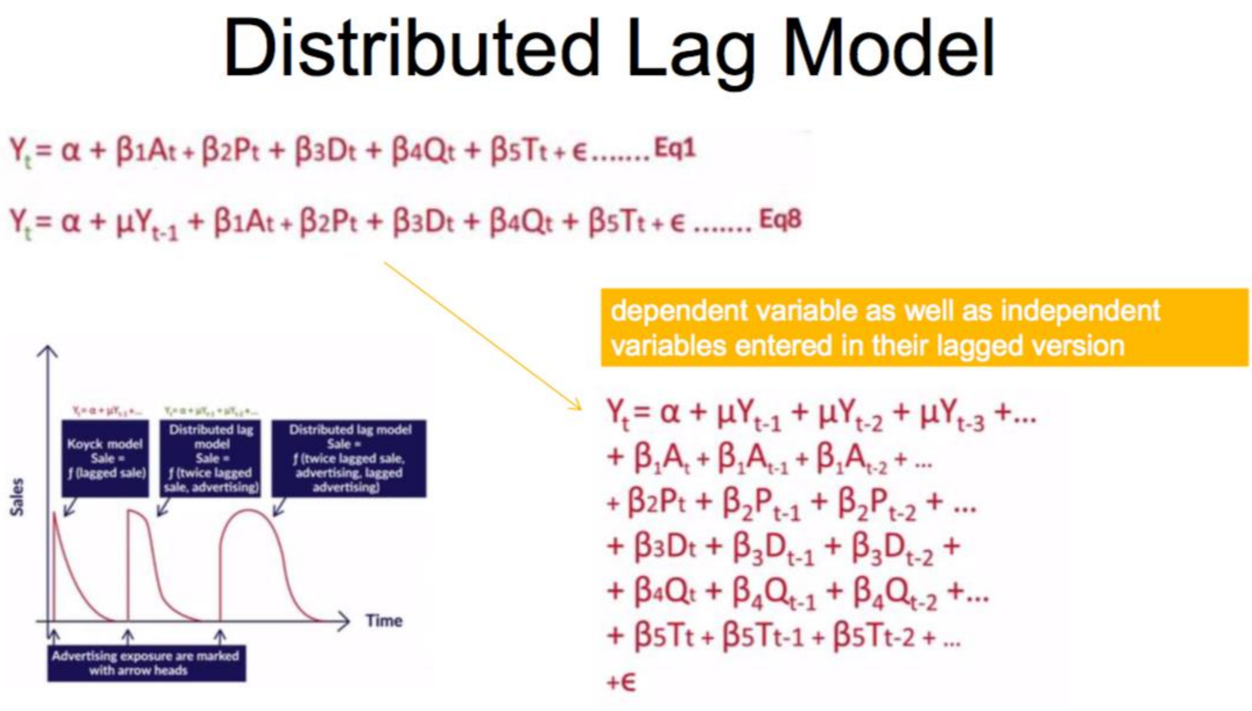


```
Coefficients:
                                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                                 4.604e+03  9.475e+01  48.594  < 2e-16 ***
s1_fact.order_payment_type                                  7.481e+01  8.797e+00   8.503  < 2e-16 ***
delayed                                                    -1.338e+02  1.602e+01  -8.353  < 2e-16 ***
inc_LP_MA1                                                 -6.028e+02  2.568e+01 -23.473  < 2e-16 ***
inc_LP_MA2                                                 -8.966e+02  3.813e+01 -23.513  < 2e-16 ***
inc_LP_MA3                                                  4.416e+03  2.888e+01 152.900  < 2e-16 ***
Total.Investment                                            3.324e+03  3.994e+02   8.322  < 2e-16 ***
TV                                                         -3.065e+03  3.815e+02  -8.034 9.50e-16 ***
Digital                                                    -3.244e+03  4.127e+02  -7.861 3.85e-15 ***
Sponsorship                                                -3.341e+03  4.009e+02  -8.334  < 2e-16 ***
Content.Marketing                                          -2.835e+03  3.884e+02  -7.299 2.92e-13 ***
Online.marketing                                           -3.121e+03  3.880e+02  -8.045 8.73e-16 ***
X.Affiliates                                               -4.084e+03  4.733e+02  -8.629  < 2e-16 ***
SEM                                                        -3.403e+03  4.055e+02  -8.392  < 2e-16 ***
Radio                                                      -2.260e+03  4.962e+02  -4.555 5.24e-06 ***
Other                                                      -3.451e+03  4.084e+02  -8.451  < 2e-16 ***
`basic_data_frame$p_tagmass`                               -4.297e+03  7.578e+01 -56.712  < 2e-16 ***
`basic_data_frame$deliveryPerceptiontolerableDeliveryPerception`  2.837e+01  7.277e+00   3.899 9.68e-05 ***
`basic_data_frame$promotion_rangelessThan25pcDiscount`      4.400e+02  1.707e+01  25.771  < 2e-16 ***
`basic_data_frame$promotion_range25to50pcDiscount`          4.063e+02  1.750e+01  23.223  < 2e-16 ***
`basic_data_frame$promotion_range50to75pcDiscount`          4.491e+02  1.688e+01  26.606  < 2e-16 ***
`basic_data_frame$promotion_range75to100pcDiscount`         5.203e+02  5.220e+01   9.966  < 2e-16 ***
`gmv-1`                                                     1.820e-01  2.312e-03  78.711  < 2e-16 ***
`gmv-2`                                                     2.145e-01  2.395e-03  89.567  < 2e-16 ***
`gmv-3`                                                     2.771e-01  2.391e-03 115.871  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1228 on 120088 degrees of freedom
Multiple R-squared:  0.6104,	Adjusted R-squared:  0.6103 
F-statistic:  7839 on 24 and 120088 DF,  p-value: < 2.2e-16
```

**Koyck Model - Gaming**



```
Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                         1.274e+02  2.112e+01   6.030 1.64e-09 ***
s1_fact.order_payment_type                          6.267e+01  4.577e+00  13.691  < 2e-16 ***
delayed                                            -4.587e+01  6.378e+00  -7.193 6.37e-13 ***
inc_LP_MA1                                         -9.542e+01  1.020e+01  -9.357  < 2e-16 ***
inc_LP_MA2                                         -4.270e+01  1.284e+01  -3.325 0.000884 ***
inc_LP_MA3                                          1.413e+03  9.287e+00 152.110  < 2e-16 ***
Total.Investment                                    1.298e+02  2.503e+01   5.186 2.15e-07 ***
TV                                                 -2.216e+02  3.423e+01  -6.475 9.53e-11 ***
Sponsorship                                        -1.177e+02  2.381e+01  -4.945 7.60e-07 ***
Content.Marketing                                  -8.672e+02  9.535e+01  -9.095  < 2e-16 ***
Online.marketing                                   -1.513e+02  2.978e+01  -5.080 3.77e-07 ***
SEM                                                -1.221e+02  3.209e+01  -3.805 0.000142 ***
Other                                              -1.144e+02  2.545e+01  -4.497 6.89e-06 ***
`basic_data_frame$promotion_rangelessThan25pcDiscount`  1.290e+02  6.614e+00  19.498  < 2e-16 ***
`basic_data_frame$promotion_range25to50pcDiscount`  -3.948e+01  5.893e+00  -6.699 2.10e-11 ***
`basic_data_frame$promotion_range50to75pcDiscount`  -8.904e+01  5.883e+00 -15.136  < 2e-16 ***
`gmv-1`                                             2.032e-01  1.935e-03 105.011  < 2e-16 ***
`gmv-2`                                             2.169e-01  1.960e-03 110.676  < 2e-16 ***
`gmv-3`                                             2.154e-01  1.941e-03 110.955  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 824.1 on 197295 degrees of freedom
Multiple R-squared:  0.5442,	Adjusted R-squared:  0.5441 
F-statistic: 1.309e+04 on 18 and 197295 DF,  p-value: < 2.2e-16
```



# Distributed Lag Model

$$Y_t = \alpha + \beta_1 A_t + \beta_2 P_t + \beta_3 D_t + \beta_4 Q_t + \beta_5 T_t + \epsilon \text{.......Eq1}$$

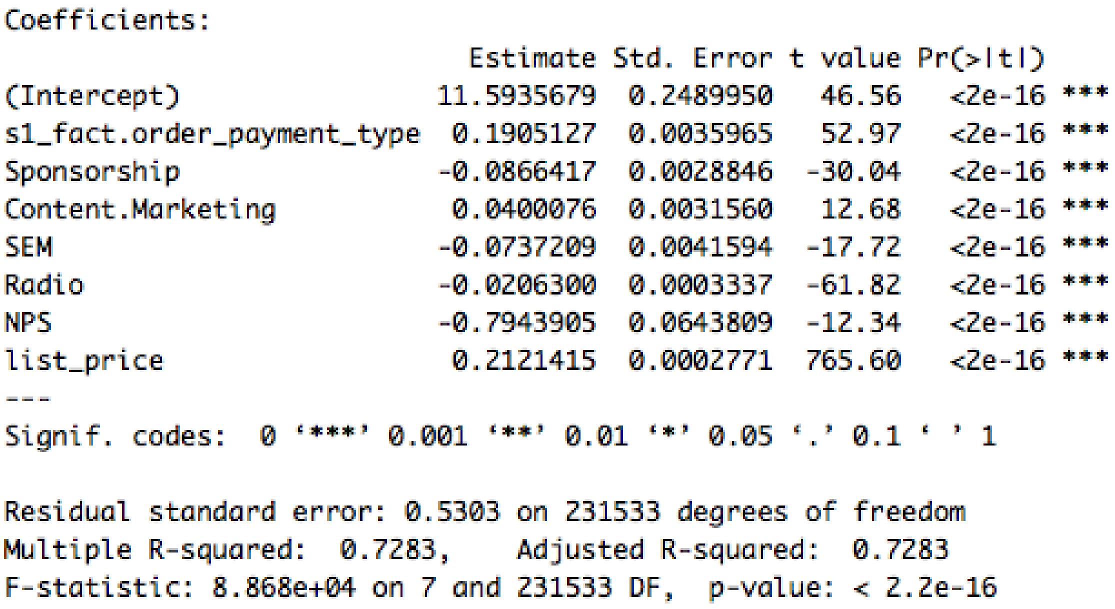
$$Y_t = \alpha + \mu Y_{t-1} + \beta_1 A_t + \beta_2 P_t + \beta_3 D_t + \beta_4 Q_t + \beta_5 T_t + \epsilon \text{ ....... Eq8}$$

dependent variable as well as independent variables entered in their lagged version

$$\begin{aligned} Y_t = \alpha &+ \mu Y_{t-1} + \mu Y_{t-2} + \mu Y_{t-3} + \ldots \\ &+ \beta_1 A_t + \beta_1 A_{t-1} + \beta_1 A_{t-2} + \ldots \\ &+ \beta_2 P_t + \beta_2 P_{t-1} + \beta_2 P_{t-2} + \ldots \\ &+ \beta_3 D_t + \beta_3 D_{t-1} + \beta_3 D_{t-2} + \\ &+ \beta_4 Q_t + \beta_4 Q_{t-1} + \beta_4 Q_{t-2} + \ldots \\ &+ \beta_5 T_t + \beta_5 T_{t-1} + \beta_5 T_{t-2} + \ldots \\ &+ \epsilon \end{aligned}$$

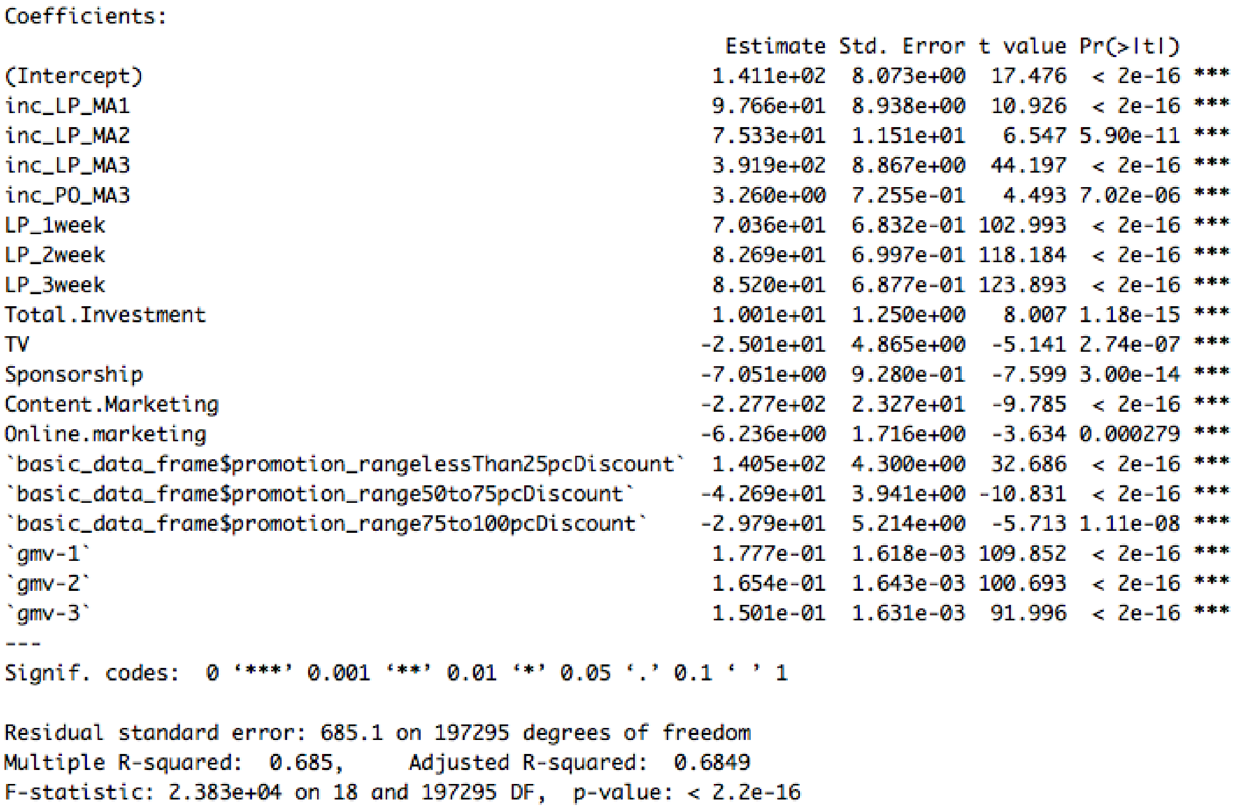
Sales
Time
$Y_t = \alpha + \mu Y_{t-1} + \ldots$
$Y_t = \alpha + \mu Y_{t-1} + \mu Y_{t-2} + \ldots$
Koyck model Sale = ƒ (lagged sale)
Distributed lag model Sale = ƒ (twice lagged sale, advertising)
Distributed lag model Sale = ƒ (twice lagged sale, advertising, lagged advertising)
Advertising exposure are marked with arrow heads

**Distributed Lag Model – Camera**



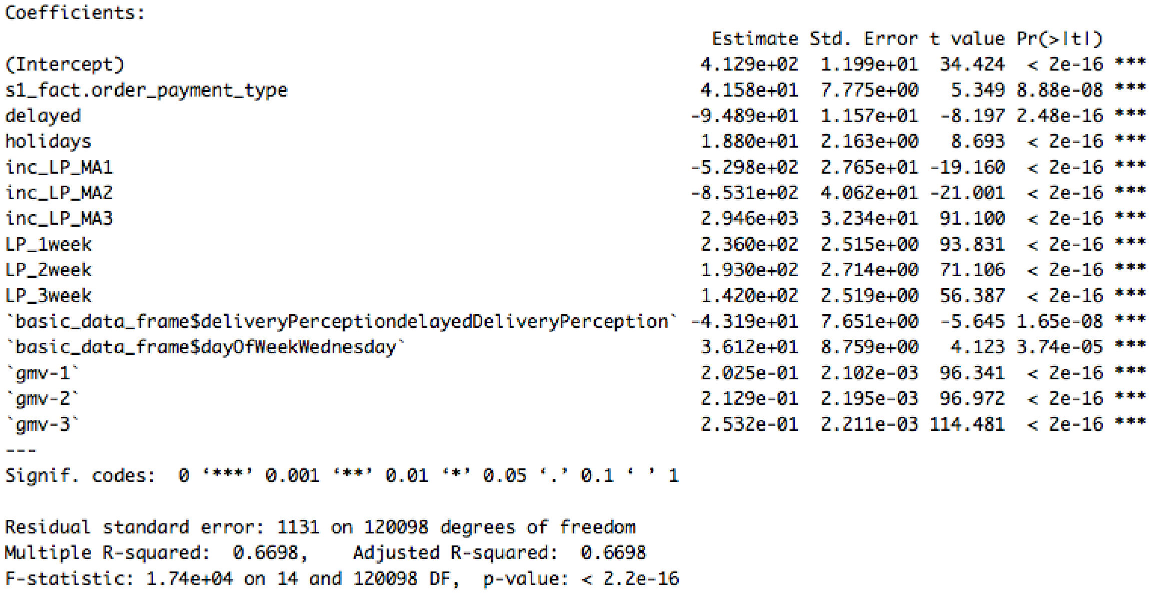
```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                11.5935679  0.2489950   46.56   <2e-16 ***
s1_fact.order_payment_type  0.1905127  0.0035965   52.97   <2e-16 ***
Sponsorship                -0.0866417  0.0028846  -30.04   <2e-16 ***
Content.Marketing           0.0400076  0.0031560   12.68   <2e-16 ***
SEM                        -0.0737209  0.0041594  -17.72   <2e-16 ***
Radio                      -0.0206300  0.0003337  -61.82   <2e-16 ***
NPS                        -0.7943905  0.0643809  -12.34   <2e-16 ***
list_price                  0.2121415  0.0002771  765.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5303 on 231533 degrees of freedom
Multiple R-squared:  0.7283,	Adjusted R-squared:  0.7283 
F-statistic: 8.868e+04 on 7 and 231533 DF,  p-value: < 2.2e-16
```

**Distributed Lag Model – Gamming**



```
Coefficients:
                                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                             1.411e+02  8.073e+00  17.476  < 2e-16 ***
inc_LP_MA1                                              9.766e+01  8.938e+00  10.926  < 2e-16 ***
inc_LP_MA2                                              7.533e+01  1.151e+01   6.547 5.90e-11 ***
inc_LP_MA3                                              3.919e+02  8.867e+00  44.197  < 2e-16 ***
inc_PO_MA3                                              3.260e+00  7.255e-01   4.493 7.02e-06 ***
LP_1week                                                7.036e+01  6.832e-01 102.993  < 2e-16 ***
LP_2week                                                8.269e+01  6.997e-01 118.184  < 2e-16 ***
LP_3week                                                8.520e+01  6.877e-01 123.893  < 2e-16 ***
Total.Investment                                        1.001e+01  1.250e+00   8.007 1.18e-15 ***
TV                                                     -2.501e+01  4.865e+00  -5.141 2.74e-07 ***
Sponsorship                                            -7.051e+00  9.280e-01  -7.599 3.00e-14 ***
Content.Marketing                                      -2.277e+02  2.327e+01  -9.785  < 2e-16 ***
Online.marketing                                       -6.236e+00  1.716e+00  -3.634 0.000279 ***
`basic_data_frame$promotion_rangelessThan25pcDiscount`  1.405e+02  4.300e+00  32.686  < 2e-16 ***
`basic_data_frame$promotion_range50to75pcDiscount`     -4.269e+01  3.941e+00 -10.831  < 2e-16 ***
`basic_data_frame$promotion_range75to100pcDiscount`    -2.979e+01  5.214e+00  -5.713 1.11e-08 ***
`gmv-1`                                                 1.777e-01  1.618e-03 109.852  < 2e-16 ***
`gmv-2`                                                 1.654e-01  1.643e-03 100.693  < 2e-16 ***
`gmv-3`                                                 1.501e-01  1.631e-03  91.996  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 685.1 on 197295 degrees of freedom
Multiple R-squared:  0.685,	Adjusted R-squared:  0.6849 
F-statistic: 2.383e+04 on 18 and 197295 DF,  p-value: < 2.2e-16
```

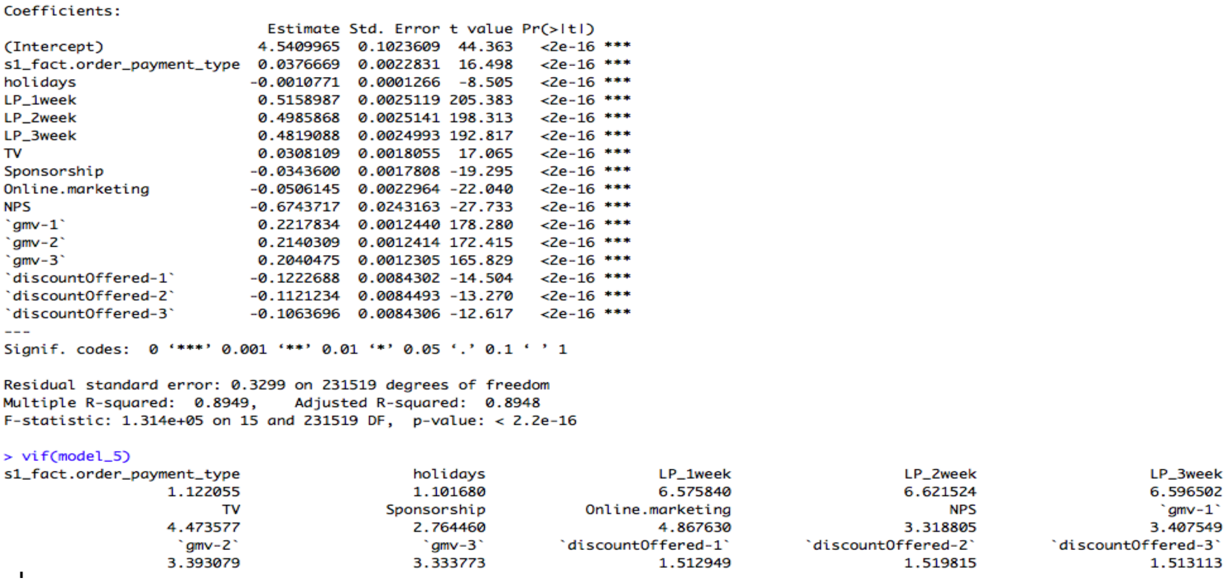
## Distributed Lag Model – Home Audio



```
Coefficients:
                                                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                                  4.129e+02  1.199e+01  34.424  < 2e-16 ***
s1_fact.order_payment_type                                   4.158e+01  7.775e+00   5.349 8.88e-08 ***
delayed                                                     -9.489e+01  1.157e+01  -8.197 2.48e-16 ***
holidays                                                     1.880e+01  2.163e+00   8.693  < 2e-16 ***
inc_LP_MA1                                                  -5.298e+02  2.765e+01 -19.160  < 2e-16 ***
inc_LP_MA2                                                  -8.531e+02  4.062e+01 -21.001  < 2e-16 ***
inc_LP_MA3                                                   2.946e+03  3.234e+01  91.100  < 2e-16 ***
LP_1week                                                     2.360e+02  2.515e+00  93.831  < 2e-16 ***
LP_2week                                                     1.930e+02  2.714e+00  71.106  < 2e-16 ***
LP_3week                                                     1.420e+02  2.519e+00  56.387  < 2e-16 ***
`basic_data_frame$deliveryPerceptiondelayedDeliveryPerception` -4.319e+01  7.651e+00  -5.645 1.65e-08 ***
`basic_data_frame$dayOfWeekWednesday`                        3.612e+01  8.759e+00   4.123 3.74e-05 ***
`gmv-1`                                                      2.025e-01  2.102e-03  96.341  < 2e-16 ***
`gmv-2`                                                      2.129e-01  2.195e-03  96.972  < 2e-16 ***
`gmv-3`                                                      2.532e-01  2.211e-03 114.481  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1131 on 120098 degrees of freedom
Multiple R-squared:  0.6698,	Adjusted R-squared:  0.6698 
F-statistic: 1.74e+04 on 14 and 120098 DF,  p-value: < 2.2e-16
```

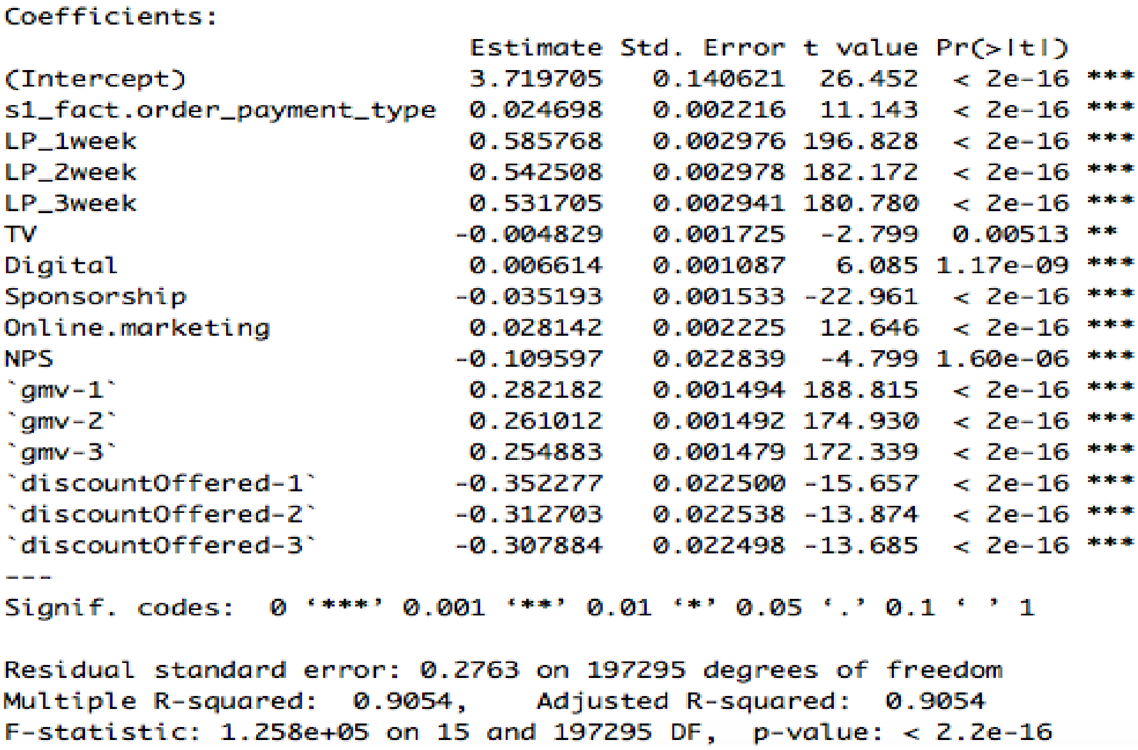
# Multiplicative + Distributed Model

### Multiplicative Distributed Model - Camera Model



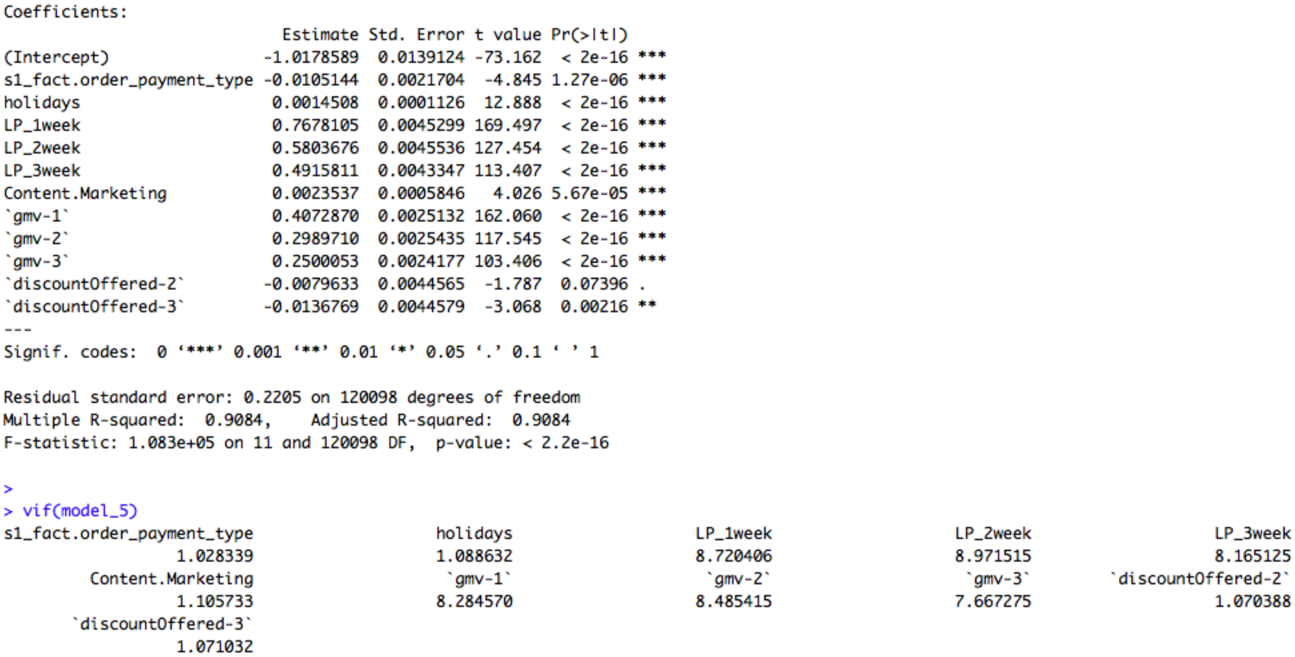
```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 4.5409965  0.1023609  44.363   <2e-16 ***
s1_fact.order_payment_type  0.0376669  0.0022831  16.498   <2e-16 ***
holidays                   -0.0010771  0.0001266  -8.505   <2e-16 ***
LP_1week                    0.5158987  0.0025119 205.383   <2e-16 ***
LP_2week                    0.4985868  0.0025141 198.313   <2e-16 ***
LP_3week                    0.4819088  0.0024993 192.817   <2e-16 ***
TV                          0.0308109  0.0018055  17.065   <2e-16 ***
Sponsorship                -0.0343600  0.0017808 -19.295   <2e-16 ***
Online.marketing           -0.0506145  0.0022964 -22.040   <2e-16 ***
NPS                        -0.6743717  0.0243163 -27.733   <2e-16 ***
`gmv-1`                     0.2217834  0.0012440 178.280   <2e-16 ***
`gmv-2`                     0.2140309  0.0012414 172.415   <2e-16 ***
`gmv-3`                     0.2040475  0.0012305 165.829   <2e-16 ***
`discountOffered-1`        -0.1222688  0.0084302 -14.504   <2e-16 ***
`discountOffered-2`        -0.1121234  0.0084493 -13.270   <2e-16 ***
`discountOffered-3`        -0.1063696  0.0084306 -12.617   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3299 on 231519 degrees of freedom
Multiple R-squared:  0.8949,	Adjusted R-squared:  0.8948 
F-statistic: 1.314e+05 on 15 and 231519 DF,  p-value: < 2.2e-16

> vif(model_5)
s1_fact.order_payment_type                   holidays                   LP_1week                   LP_2week                   LP_3week 
                  1.122055                   1.101680                   6.575840                   6.621524                   6.596502 
                        TV                Sponsorship           Online.marketing                        NPS                    `gmv-1` 
                  4.473577                   2.764460                   4.867630                   3.318805                   3.407549 
                   `gmv-2`                    `gmv-3`        `discountOffered-1`        `discountOffered-2`        `discountOffered-3` 
                  3.393079                   3.333773                   1.512949                   1.519815                   1.513113 
```

### Multiplicative Distributed Model - Gamming Model



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 3.719705   0.140621  26.452  < 2e-16 ***
s1_fact.order_payment_type  0.024698   0.002216  11.143  < 2e-16 ***
LP_1week                    0.585768   0.002976 196.828  < 2e-16 ***
LP_2week                    0.542508   0.002978 182.172  < 2e-16 ***
LP_3week                    0.531705   0.002941 180.780  < 2e-16 ***
TV                         -0.004829   0.001725  -2.799  0.00513 ** 
Digital                     0.006614   0.001087   6.085 1.17e-09 ***
Sponsorship                -0.035193   0.001533 -22.961  < 2e-16 ***
Online.marketing            0.028142   0.002225  12.646  < 2e-16 ***
NPS                        -0.109597   0.022839  -4.799 1.60e-06 ***
`gmv-1`                     0.282182   0.001494 188.815  < 2e-16 ***
`gmv-2`                     0.261012   0.001492 174.930  < 2e-16 ***
`gmv-3`                     0.254883   0.001479 172.339  < 2e-16 ***
`discountOffered-1`        -0.352277   0.022500 -15.657  < 2e-16 ***
`discountOffered-2`        -0.312703   0.022538 -13.874  < 2e-16 ***
`discountOffered-3`        -0.307884   0.022498 -13.685  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2763 on 197295 degrees of freedom
Multiple R-squared:  0.9054,	Adjusted R-squared:  0.9054 
F-statistic: 1.258e+05 on 15 and 197295 DF,  p-value: < 2.2e-16
```

## Multiplicative Distributed Model – Home Audio



```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -1.0178589  0.0139124 -73.162  < 2e-16 ***
s1_fact.order_payment_type -0.0105144  0.0021704  -4.845 1.27e-06 ***
holidays                   0.0014508  0.0001126  12.888  < 2e-16 ***
LP_1week                   0.7678105  0.0045299 169.497  < 2e-16 ***
LP_2week                   0.5803676  0.0045536 127.454  < 2e-16 ***
LP_3week                   0.4915811  0.0043347 113.407  < 2e-16 ***
Content.Marketing          0.0023537  0.0005846   4.026 5.67e-05 ***
`gmv-1`                    0.4072870  0.0025132 162.060  < 2e-16 ***
`gmv-2`                    0.2989710  0.0025435 117.545  < 2e-16 ***
`gmv-3`                    0.2500053  0.0024177 103.406  < 2e-16 ***
`discountOffered-2`       -0.0079633  0.0044565  -1.787  0.07396 .  
`discountOffered-3`       -0.0136769  0.0044579  -3.068  0.00216 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2205 on 120098 degrees of freedom
Multiple R-squared:  0.9084,	Adjusted R-squared:  0.9084 
F-statistic: 1.083e+05 on 11 and 120098 DF,  p-value: < 2.2e-16

> 
> vif(model_5)
s1_fact.order_payment_type                   holidays                   LP_1week                   LP_2week                   LP_3week 
                  1.028339                   1.088632                   8.720406                   8.971515                   8.165125 
         Content.Marketing                    `gmv-1`                    `gmv-2`                    `gmv-3`        `discountOffered-2` 
                  1.105733                   8.284570                   8.485415                   7.667275                   1.070388 
       `discountOffered-3` 
                  1.071032 
```

## Comparison of performance of Models for Camera Categories

- Adjusted R Square figures are based on the performance of the model on the training data.
- Clearly Multiplicative distributed model has better adj R square compared to others. This takes into account the lag of discount, price and gmv.
- Other model except koyck also decent enough.
- Clearly sponsorship is key for camera model. It is resembled in all the models.
- Another key variable is list price which has impact on model and as per business also this is key parameter.
- SSE when calculated on 10 fold data cross validation is low enough of multiplicative and multiplicative distributed models. So we can chose any of those model for camera accessory category.

| Model | Important Variables | Adjusted R Square |
|---|---|---|
| **Linear Model** | Moving average of list price + TV+ sponsorship + discount | 0.75 |
| **Multiplicative Model** | Payment type + sponsorship+Radio +list price | 0.7281 |
| **Koyck Model** | gmv lag + moving average of list price + day of the week | 0.4053 |
| **Distributed Model** | Payment type + sponsorship+ radio +list price | 0.7283 |
| **Multiplicative Distributed Model** | Lag of gmv + discount+ sponsorship +lag of list price | 0.8948 |

# Comparison of performance of Models for Gamming Categories

- Adjusted R Square figures are based on the performance of the model on the training data.
- Clearly Multiplicative distributed model has better adj R square compared to others. This takes into account the lag of discount, price and gmv.
- No model is decent when compared to multi distributed model.
- Clearly Content marketing is key for gaming model. It is resembled in all the models.
- Another key variable is list price and discount which has impact on model and as per business also this is key parameter.
- SSE when calculated on 10 fold data cross validation is low enough of multiplicative and multiplicative distributed models. So we can chose any of those model for gaming accessory category.

| Model | Important Variables | Adjusted R Square |
|---|---|---|
| **Linear Model** | Moving average of list price+ content marketing+ payment type + discount | 0.6073 |
| **Multiplicative Model** | Payment type + content marketing +Digital | 0.7226 |
| **Koyck Model** | gmv lag + moving average of list price + TV+ Content Marketing | 0.5441 |
| **Distributed Model** | Payment type + TV + sponsorship+ radio +list price lag | 0.6848 |
| **Multiplicative Distributed Model** | Lag of gmv + discount+ sponsorship +lag of list price + payment type | 0.9054 |

# Comparison of performance of Models for Home Audio Categories

- Adjusted R Square figures are based on the performance of the model on the training data.
- Clearly Multiplicative distributed model has better adj R square compared to others. This takes into account the lag of discount, price and gmv.
- No model is decent when compared to multi distributed model.
- Clearly holiday/big billion day sale is key for home audio model. It is resembled in all the models.
- Another key variable is list price and discount which has impact on model and as per business also this is key parameter.
- SSE when calculated on 10 fold data cross validation is low enough of multiplicative and multiplicative distributed models. So we can chose any of those model for home audio accessory category.

| Model | Important Variables | Adjusted R Square |
|---|---|---|
| **Linear Model** | Holidays sale + discount + delivery time | 0.4957 |
| **Multiplicative Model** | Payment type + holiday sale+ price +digital marketing | 0.592 |
| **Koyck Model** | gmv lag + discount+ All media + Marketing | 0.6103 |
| **Distributed Model** | Holiday sale +Payment type + list price lag + gmv lag | 0.6698 |
| **Multiplicative Distributed Model** | Lag of gmv + holidays+ content marketing+ lag of list price | 0.9084 |