Digital Attribution at W.M. Winters

Winters Attribution

2/20/2020

Pairs: Cohort B: Sherry Zuo, Yishuang Song

The assignment uses the "Winters_Attribution-TERM-SECTION" data. The Winters case study uses a unique dataset was collected in collaboration with a large online media analytics and optimization platform company. The online media company managed the entire campaign of a U.S.-based retailer. The individual-level data set consists of advertising exposures and user-initiated actions, with users tracked across different advertising channels and media. Note that all observations relate to an order (touchpoints that do not lead to a purchase are absent). The unit of observation is an order-touchpoint, so that the same order is repeated by the number of touches.

```
load("~/Desktop/BU/BA 860/Winters_Attribution-W20-MSBA-TTh.RData")
glimpse(data)
```

```
## Observations: 7,597
## Variables: 12
## $ Orderid          <int> 11634052, 11634052, 11634059, 11634059, 11634...
## $ Orderdatetime    <chr> "2012-05-01 4:24", "2012-05-01 4:24", "2012-0...
## $ Saleamount       <dbl> 341.50, 341.50, 339.00, 339.00, 339.00, 101.7...
## $ Newcustomer      <chr> "Y", "Y", "Y", "Y", "Y", "N", "N", "N", "N", ...
## $ Position         <int> 1, 0, 2, 1, 0, 7, 6, 5, 4, 3, 2, 1, 0, 7, 6, ...
## $ Positiondatetime <chr> "2012-05-01 3:49", "2012-05-01 3:47", "2012-0...
## $ Groupname        <chr> "BUZZ AFFILIATE", "SEARCH GOOGLE BRAND", "PRI...
## $ Networkname      <chr> "Buzz CPA Affiliate", "G: Medifast Brand Term...
## $ Networkid        <chr> "buzz23", "g000793", "medifastok.com", "nar74...
## $ Brand            <chr> "N", "Y", "N", "N", "N", "N", "N", "N", "N", ...
## $ Positionname     <chr> "CONVERTER", "ORIGINATOR", "CONVERTER", "ASSI...
## $ DaysToConvert    <dbl> 0, 0, 2, 7, 8, 2, 2, 2, 2, 2, 2, 3, 3, 67, 67...
```

Q1. (30 pts) Compare first-touch vs. last-touch attribution models

a) By the different media channels, what is the total number of orders by last-touch ("converter") and by first-touch ("originator") attribution? What is the corresponding share of credit for the two attribution models?

```
#Originator->Roaster->Assist->Converter
## table
channel <- data %>%
  select(Groupname, Positionname) %>%
  filter(Positionname == 'CONVERTER' | Positionname == 'ORIGINATOR')
```

```
a <- table(channel$Groupname, channel$Positionname)
a
```

```
##
```

```
##                            CONVERTER ORIGINATOR
##   BUZZ AFFILIATE                 466        204
##   CJ                             237         76
##   CPM                            839        609
##   DIRECT MAIL                      0          1
##   OTHER                            5         19
##   PRINT - MAGAZINES                5          2
##   SEARCH GOOGLE BRAND              0        508
##   SEARCH GOOGLE NON-BRAND         32         47
##   SEARCH MSN BRAND                 0        128
##   SEARCH MSN NON-BRAND             5          6
##   Social                           0          1
##   TV                              19         16
##   Uncategorized                   19         10
```

```
b <- prop.table(a,2)*100
colnames(b) <- c('CONVERTER %','ORIGINATOR %')
b
```

```
##
##                            CONVERTER % ORIGINATOR %
##   BUZZ AFFILIATE           28.64167179  12.53841426
##   CJ                       14.56668715   4.67117394
##   CPM                      51.56730178  37.43085433
##   DIRECT MAIL               0.00000000   0.06146281
##   OTHER                     0.30731407   1.16779348
##   PRINT - MAGAZINES         0.30731407   0.12292563
##   SEARCH GOOGLE BRAND       0.00000000  31.22311002
##   SEARCH GOOGLE NON-BRAND   1.96681008   2.88875230
##   SEARCH MSN BRAND          0.00000000   7.86724032
##   SEARCH MSN NON-BRAND      0.30731407   0.36877689
##   Social                    0.00000000   0.06146281
##   TV                        1.16779348   0.98340504
##   Uncategorized             1.16779348   0.61462815
```

```
c <- cbind(a,b)
c
```

```
##                          CONVERTER ORIGINATOR CONVERTER % ORIGINATOR %
## BUZZ AFFILIATE                 466        204  28.6416718  12.53841426
## CJ                             237         76  14.5666872   4.67117394
## CPM                            839        609  51.5673018  37.43085433
## DIRECT MAIL                      0          1   0.0000000   0.06146281
## OTHER                            5         19   0.3073141   1.16779348
## PRINT - MAGAZINES                5          2   0.3073141   0.12292563
## SEARCH GOOGLE BRAND              0        508   0.0000000  31.22311002
## SEARCH GOOGLE NON-BRAND         32         47   1.9668101   2.88875230
## SEARCH MSN BRAND                 0        128   0.0000000   7.86724032
## SEARCH MSN NON-BRAND             5          6   0.3073141   0.36877689
## Social                           0          1   0.0000000   0.06146281
## TV                              19         16   1.1677935   0.98340504
## Uncategorized                   19         10   1.1677935   0.61462815
```

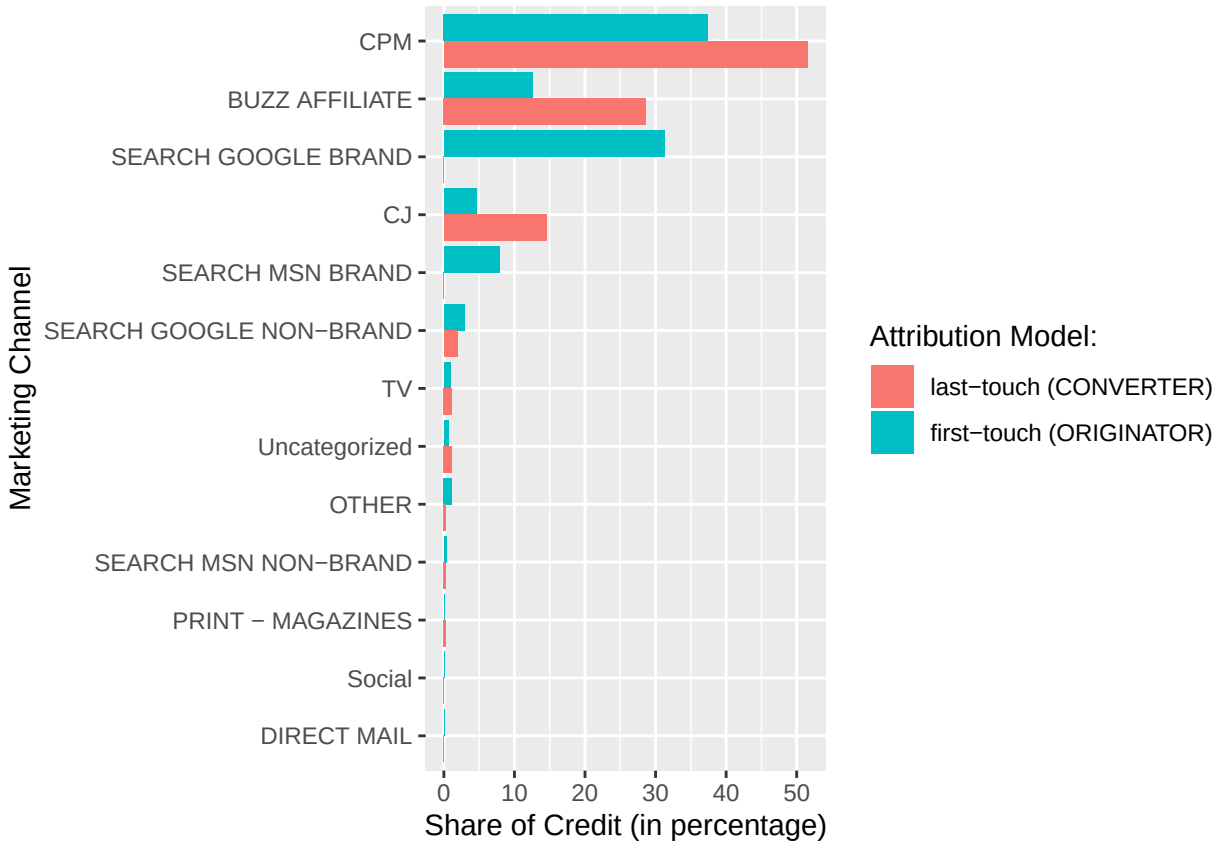
```
TOTAL <- colSums(c)
d <- kable(rbind(c,TOTAL))
d
```

	CONVERTER	ORIGINATOR	CONVERTER %	ORIGINATOR %
BUZZ AFFILIATE	466	204	28.6416718	12.5384143
CJ	237	76	14.5666872	4.6711739
CPM	839	609	51.5673018	37.4308543
DIRECT MAIL	0	1	0.0000000	0.0614628
OTHER	5	19	0.3073141	1.1677935
PRINT - MAGAZINES	5	2	0.3073141	0.1229256
SEARCH GOOGLE BRAND	0	508	0.0000000	31.2231100
SEARCH GOOGLE NON-BRAND	32	47	1.9668101	2.8887523
SEARCH MSN BRAND	0	128	0.0000000	7.8672403
SEARCH MSN NON-BRAND	5	6	0.3073141	0.3687769
Social	0	1	0.0000000	0.0614628
TV	19	16	1.1677935	0.9834050
Uncategorized	19	10	1.1677935	0.6146281
TOTAL	1627	1627	100.0000000	100.0000000

Answer: the total number of orders by last-touch ("converter") attribution model is **1627**, and by first-touch ("originator") attribution model is **1627**, the corresponding share of credit for the two attribution models see the table above.

b) In a single bar plot, plot the share of credit (in percentage) for the first- and last-touch attribution models by marketing channel.

```
E <- as.data.frame(b)
ggplot(E,aes(x=reorder(Var1,Freq),y=Freq,fill=Var2))+
  geom_bar(position = 'dodge', stat = 'identity')+
  labs(x="Marketing Channel", y="Share of Credit (in percentage)")+
  coord_flip()+
  scale_fill_discrete(name="Attribution Model:",
                labels = c("last-touch (CONVERTER)", "first-touch (ORIGINATOR)"))
```



c) Compare and contrast the two attribution model results. What would be the consequence to Winters if it allocated its marketing budget entirely on the basis of the last-touch attribution model?

Answer: Based on the two attribution model results shown in the graph above, both first-touch and last-touch position are important, because for the top two channels "CPM" and "BUZZ AFFILIATE", both converter and originator count for more than 50% of share of credit. If Winters allocates all marketing budget to "Converter", then channel "CPM", "BUZZ AFFILIATE", and "CJ" will be beneficial. However, it will hurt the channel "Search Google Brand" and "Search MSN Brand". Therefore, it's better to allocate marketing budget on the basis of both first-touch and last-touch attribution models, and allocate more budgets to media channels "BUZZ AFFILIATE", "CJ", "CPM", especially for channel "CPM"

Q2. (20 pts) Compare new customers and old customers

Hint: The data is structured at the order-touch level, but question 2 asks you to examine the data at the order level. You can do so by filtering the data to only examine "ORIGINATOR" touches.

a) Using the DaysToConvert variable, what is the average number of days that it takes for a new customer to convert (from the first touchpoint)? What is the average number of days that it takes for an old customer to convert?

```
average_DaysToConvert<-data%>%
  filter(Positionname == 'ORIGINATOR')%>%
  group_by(Newcustomer)%>%
  summarize(average_DaysToConvert=mean(DaysToConvert))
average_DaysToConvert
```

```
## # A tibble: 2 x 2
##   Newcustomer average_DaysToConvert
##   <chr>                       <dbl>
## 1 N                           29.0
## 2 Y                            6.32
```

Answer: The average number of days that it takes for a **new customer** to convert is **6.32**, the average number of days that it takes for an **old customer** to convert is **29.02**

b) What is the average number of touchpoints by new versus old customer's orders?
Hint: Use the Touches variable if available. If not, create the Touches variable for the number of touchpoints per order. R users can use the add_count() function. To do so correctly, you need to count touches per orderid before you filter (else, filtering out "ORIGINATOR" touches trivially implies 1 touch per order).

```
average_touchpoints<-data%>%
  add_count(Orderid, name="Touches")%>%
  filter(Positionname == 'ORIGINATOR')%>%
  group_by(Newcustomer)%>%
  summarize(average_touchpoints=mean(Touches))
average_touchpoints
```

```
## # A tibble: 2 x 2
##   Newcustomer average_touchpoints
##   <chr>                     <dbl>
## 1 N                          5.14
## 2 Y                          4.34
```

Answer: The average number of touchpoints by new customers is **4.34**, the average number of touchpoints by old customers is **5.14**

c) What is the average order sales amount by new versus old customer's orders?

```
average_sales<-data%>%
  filter(Positionname == 'ORIGINATOR')%>%
  group_by(Newcustomer)%>%
  summarize(average_sales=mean(Saleamount))
average_sales
```

```
## # A tibble: 2 x 2
##   Newcustomer average_sales
##   <chr>               <dbl>
## 1 N                    205.
## 2 Y                    267.
```

Answer: The average order sales amount by new customers is **266.52**, the average order sales amount by old customers is **205.30**

d) Summarize how new and old customers differ along these three variables.

```
kable(cbind(average_DaysToConvert, average_touchpoints[,2], average_sales[,2]))
```

Newcustomer	average_DaysToConvert	average_touchpoints	average_sales
N	29.016418	5.144776	205.3025
Y	6.321839	4.336468	266.5208

Answer: For **new customers**, they need 6 days to convert from the first touchpoint, 4 touches on average per order and generate 267 dollars for sales. On contrast, the **old customers** need 29 days to convert, 5 touches on average and generate 205 dollars per sales. **Comparing new and old customers**, we can conclude that new customers need less days to convert, less touches and provide higher sales revenue, while old customers need more days to convert, more touches but generate lower sales revenue. The reason can be that compared to new customers who are unfamiliar with the company, old customers are already aware of all products and therefore advertising seems less attractive to old customers.

Q3. (20 pts) Consider the revenue per marketing channel. For this question, focus on the originators (first-touch attribution).

a) Create a table (as in Q1) containing the average sales per order as well as the total revenue by originator channel.

```
astroc<-data%>%
  filter(Positionname == 'ORIGINATOR')%>%
  group_by(Groupname)%>%
  summarize(average_sales=mean(Saleamount), total_revenue=sum(Saleamount))
TOTAL<-list('TOTAL', colSums(astroc[2]), colSums(astroc[3]))
kable(rbind(astroc,TOTAL))
```

Groupname	average_sales	total_revenue
BUZZ AFFILIATE	254.0873	51833.81
CJ	258.2901	19630.05
CPM	230.2288	140209.31
DIRECT MAIL	170.9800	170.98
OTHER	238.1832	4525.48
PRINT - MAGAZINES	300.0300	600.06
SEARCH GOOGLE BRAND	250.0232	127011.81
SEARCH GOOGLE NON-BRAND	235.9751	11090.83
SEARCH MSN BRAND	229.1580	29332.23
SEARCH MSN NON-BRAND	235.2483	1411.49
Social	561.0800	561.08
TV	280.4294	4486.87
Uncategorized	174.9050	1749.05
TOTAL	3418.6185	392613.05

b) What is the total incremental gross revenue accruing to Winters by originator channel? Express your answer in a table. Assume that Winters has a gross margin of 40%. Also assume an incrementality factor of 5% for branded search and 10% for the remaining channels. Note: An incrementality factor refers to the share of sales that are assumed to be incremental or caused by the channel. For instance, an incrementality factor of 20% implies that 0.20 dollars of every 1 dollar in sales is incremental.

```
incremental_gross<-data%>%
  filter(Positionname == 'ORIGINATOR')%>%
  group_by(Groupname, Brand)%>%
  summarize(total_revenue=sum(Saleamount))%>%
  mutate(incremental_gross=total_revenue*0.4*ifelse(Brand=="Y",0.05, 0.1))
kable(incremental_gross)
```

Groupname	Brand	total_revenue	incremental_gross
BUZZ AFFILIATE	N	51833.81	2073.3524
CJ	N	19630.05	785.2020
CPM	N	140209.31	5608.3724
DIRECT MAIL	NULL	170.98	6.8392
OTHER	N	1054.27	42.1708
OTHER	NULL	681.75	27.2700
OTHER	Y	2789.46	55.7892
PRINT - MAGAZINES	N	600.06	24.0024
SEARCH GOOGLE BRAND	Y	127011.81	2540.2362
SEARCH GOOGLE NON-BRAND	N	11090.83	443.6332
SEARCH MSN BRAND	Y	29332.23	586.6446
SEARCH MSN NON-BRAND	N	1411.49	56.4596
Social	N	561.08	22.4432
TV	NULL	4486.87	179.4748
Uncategorized	N	1749.05	69.9620

```
incremental_gross%>%
  group_by(Brand)%>%
  summarize(total_incremental_gross=sum(incremental_gross))
```

```
## # A tibble: 3 x 2
##   Brand total_incremental_gross
##   <chr>                   <dbl>
## 1 N                       9126.
## 2 NULL                     214.
## 3 Y                       3183.
```

Answer: The total incremental gross revenue accruing to Winters by originator channel for each channel shows above, and it is **$3182.67** for branded search channels, **$9125.60** for unbranded search channels, and **$213.58** for Null channels.

c) You just found out that Winters search ad team spent $4,200 on the branded search advertising covered in the data (e.g. during the time period in the data). What would you advise the search team based on your calculation directly above?

Answer: Base on the calculation, the total incremental gross of branded search is **3182.67** which is less than what Winters search ad-team spent (**$4200**) on branded search advertising, so we advise the search team still loading the branded search advertising but spend less than 3182.67 in case competitors "stolen" touches.

Q4. (25 pts) Linear/uniform attribution
Hint: The linear attribution model divides the attribution share between touches equally. For example, an order with one CPM, one CJ, and one TV touchpoint will have place one third attribution share on each touch. This can be accomplished simply by using the Touches variable (see Q2) to define the new variable: LinearAttributionShare = 1 / Touches

a) By the different marketing channels, what is the total of the linear attribution shares? What is the corresponding share of credit (in percentage) according to the linear attribution model? Express your answer in a table like in Q1. Hint: By construction, the total linear attribution shares must sum to the total number of orders.

```
linear_attribution_shares<-data%>%
  add_count(Orderid)%>%
  mutate(LinearAttributionShare=1/n)%>%
  group_by(Groupname)%>%
  summarize(Total_Linear_Attribution_Shares=sum(LinearAttributionShare))
linear_attribution_shares
```

```
## # A tibble: 13 x 2
##    Groupname               Total_Linear_Attribution_Shares
##    <chr>                                             <dbl>
##  1 BUZZ AFFILIATE                                    319.
##  2 CJ                                                139.
##  3 CPM                                               829.
##  4 DIRECT MAIL                                         0.333
##  5 OTHER                                               7.89
##  6 PRINT - MAGAZINES                                   3.02
##  7 SEARCH GOOGLE BRAND                               211.
##  8 SEARCH GOOGLE NON-BRAND                            29.1
##  9 SEARCH MSN BRAND                                   53.0
## 10 SEARCH MSN NON-BRAND                                4.78
## 11 Social                                              0.625
## 12 TV                                                 14.3
## 13 Uncategorized                                      16.6
```

```
length(unique(data$Orderid))
```

```
## [1] 1627
```

```
total_shares<-sum(linear_attribution_shares$Total_Linear_Attribution_Shares)
total_shares
```

```
## [1] 1627
```

```
total_shares==length(unique(data$Orderid))
```

```
## [1] TRUE
```

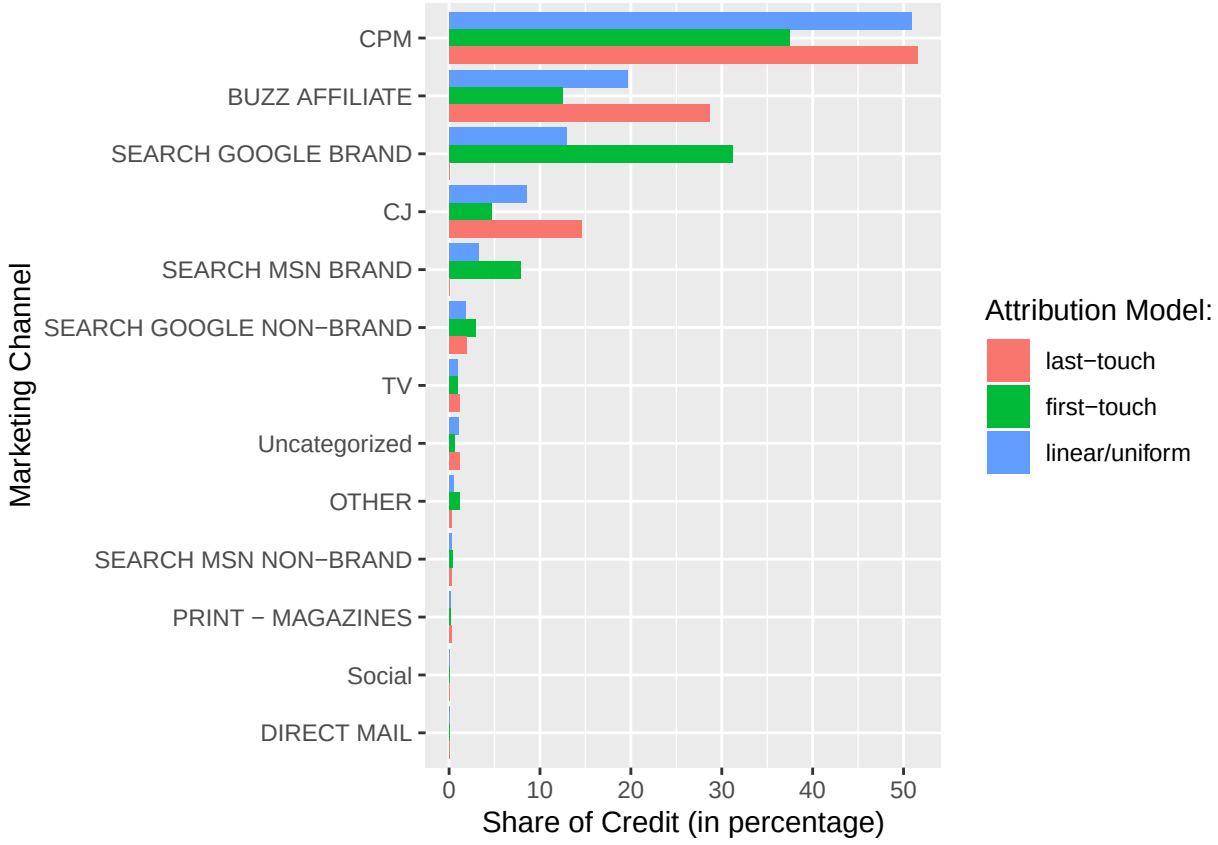
```
percentage_share_of_credit<-linear_attribution_shares%>%
  mutate('Share of Credit %' = Total_Linear_Attribution_Shares*100/total_shares)
TOTAL<-list('TOTAL', total_shares, 100)
linear_attribution_model<-rbind(percentage_share_of_credit, TOTAL)
kable(linear_attribution_model)
```

Groupname	Total_Linear_Attribution_Shares	Share of Credit %
BUZZ AFFILIATE	319.0964286	19.6125648
CJ	138.6797619	8.5236485
CPM	828.7468254	50.9371128
DIRECT MAIL	0.3333333	0.0204876
OTHER	7.8892857	0.4848977
PRINT - MAGAZINES	3.0150794	0.1853153
SEARCH GOOGLE BRAND	210.7948413	12.9560443
SEARCH GOOGLE NON-BRAND	29.1392857	1.7909825
SEARCH MSN BRAND	52.9968254	3.2573341
SEARCH MSN NON-BRAND	4.7833333	0.2939971
Social	0.6250000	0.0384143
TV	14.3285714	0.8806743
Uncategorized	16.5714286	1.0185266
TOTAL	1627.0000000	100.0000000

Answer: The total of the linear attribution shares for all channels is **1627**, and the linear attribution shares of each channel shows in the table above. The corresponding share of credit (in percentage) according to the linear attribution model also shows in the table above.

b) In a single bar plot, plot the share of credit (in percentage) for all three attribution models: first-touch, last-touch and linear/uniform.

```
ln=data.frame(Var1=percentage_share_of_credit$Groupname,
              Var2="linear/uniform",
              Freq=percentage_share_of_credit$`Share of Credit %`)
models=rbind(E,ln)
ggplot(models,aes(x=reorder(Var1,Freq),y=Freq,fill=Var2))+
  geom_bar(position = 'dodge', stat = 'identity')+
  labs(x="Marketing Channel", y="Share of Credit (in percentage)", fill = "Position")+
  coord_flip()+
  scale_fill_discrete(name="Attribution Model:",
                      labels = c("last-touch", "first-touch", "linear/uniform"))
```



c) Compare the linear model to the first-touch and last-touch models.

Answer: Base on the graph, the linear/uniform model is usually between the last-touch model and the first-touch model. The reason is that first-touch and last-touch attribution models only consider one touchpoint through the buying process, while linear/uniform model considers all touchpoints equally in one buying process. It is more reasonable to use the linear/uniform model as it is more comprehensive considering all aspects of users' behaviors during the entire buying process.

Q5. (30 pts) Examine the role of the intermediate (Roster and Assist) touch points.

a) Focusing on the top channels listed below, what is the proportion of each channel's touchpoints by position name: 1) Originator, 2) Roster, 3) Assist, and 4) Converter. For full credit, the rows must be listed in that order.
Top channels: "Buzz Affiliate, CJ, CPM, Search Google Brand, Search Google Non-Brand, Search MSN Brand, TV"

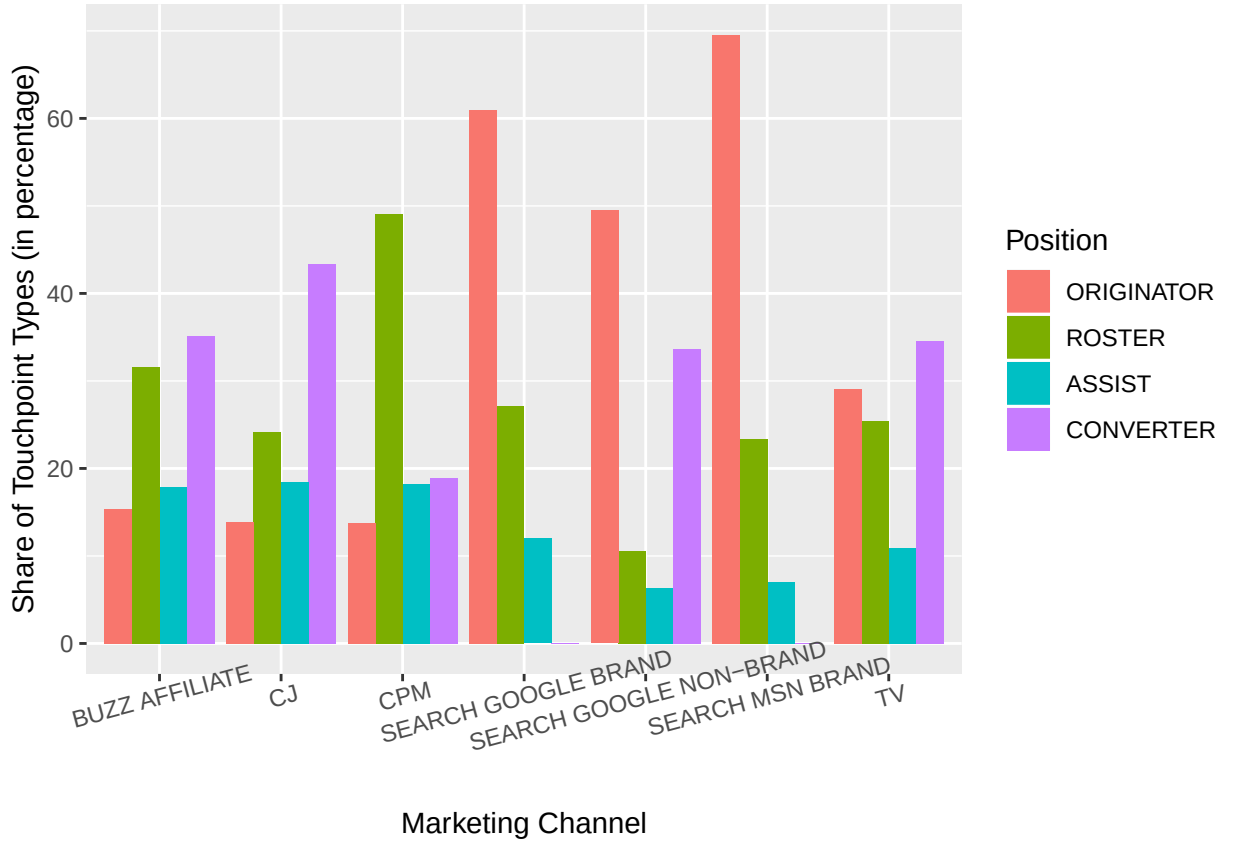
```
top_channel <- data%>%
  add_count(Orderid, name="Touches")%>%
  filter(Groupname=="BUZZ AFFILIATE" | Groupname=="CJ" | Groupname=="CPM"
         |Groupname=="SEARCH GOOGLE BRAND" | Groupname=="SEARCH GOOGLE NON-BRAND"
         | Groupname=="SEARCH MSN BRAND" | Groupname=="TV")

f=prop.table(table(top_channel$Groupname, top_channel$Positionname), 1)
TOTAL <- rowSums(f)

g<-cbind(f, TOTAL)
```

```r
#1) Originator, 2) Roster, 3) Assist, and 4) Converter
h<-g[, c(3,4,1,2,5)]
kable(h)
```

	ORIGINATOR	ROSTER	ASSIST	CONVERTER	TOTAL
BUZZ AFFILIATE	0.1536145	0.3162651	0.1792169	0.3509036	1
CJ	0.1391941	0.2417582	0.1849817	0.4340659	1
CPM	0.1374718	0.4909707	0.1821670	0.1893905	1
SEARCH GOOGLE BRAND	0.6091127	0.2709832	0.1199041	0.0000000	1
SEARCH GOOGLE NON-BRAND	0.4947368	0.1052632	0.0631579	0.3368421	1
SEARCH MSN BRAND	0.6956522	0.2336957	0.0706522	0.0000000	1
TV	0.2909091	0.2545455	0.1090909	0.3454545	1

b) In a single bar plot, plot the share in percentage (y-axis) of touchpoint types by marketing channels (x-axis).

```r
F<-as.data.frame(f)
F$Var2<-factor(F$Var2, levels=c("ORIGINATOR", "ROSTER", "ASSIST", "CONVERTER"))
ggplot(F,aes(x=reorder(Var1,Freq),y=Freq*100,fill=Var2))+
  geom_bar(position = 'dodge', stat = 'identity')+
  labs(x="Marketing Channel", y="Share of Touchpoint Types (in percentage)", fill = "Position")+
  theme(axis.text.x=element_text(angle=15))
```



c) Summarize the touch-point type results. Which channels seem to have relatively more or less of its

touchpoints as rosters and assist? As a consequence, which of these channels would receive too much or too little credit under first- and last-touch attribution?

Answer:
Different channels have different distribution for the shares within the buying process of 4 positions. Considering each channel separately, the channel BUZZ AFFILIATE, CJ and TV have the largest shares on "CONVERTER"; channel CPM has the largest shares on "ROSTER"; channel SEARCH GOOGLE BRAND, SEARCH GOOGLE NON-BRAND, and SEARCH MSN BRAND have the largest shares on "ORIGINATOR".
Channel **'SEARCH GOOGLE NON-BRAND' and 'SEARCH MSN BRAND'** seem to have **relatively less** of its touchpoints as rosters and assist. Channel **'CPM' and 'BUZZ AFFILIATE'** seem to have **relatively more** of its touchpoints as rosters and assist. As a **consequence**, combining the results with absolute difference between 'first- and last-touch' attribution and 'roster and assist' attribution, we think channel CPM would receive too little credit under first- and last-touch attribution, and channel SEARCH GOOGLE NON-BRAND would receive too much credit under first- and last-touch attribution. Therefore, as data analysts, we need to carefully choose the appropriate driving factors in various situation before we determine how to allocate budget for advertising channels and activities.