BA860 Assignment 3

Qiaoling Huang (U20421641)

2/13/2020

Pairs with Salina(Ziqin Ma)

```
load("Downloads/Winters_Attribution-W20-MSBA-MW.RData")
glimpse(data)
```

```
## Observations: 7,004
## 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, 174.6...
## $ Newcustomer      <chr> "Y", "Y", "Y", "Y", "Y", "N", "N", "Y", "Y", ...
## $ Position         <int> 1, 0, 2, 1, 0, 1, 0, 3, 2, 1, 0, 2, 1, 0, 3, ...
## $ 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, 0, 1, 1, 1, 5, 8, 1, 1, 1, 0, ...
```

1_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?

```
## calculate the number of first touch and last touch
first_touch <- data %>%
  filter(Positionname == "ORIGINATOR") %>%
  select(Groupname, Positionname) %>%
  group_by(Groupname) %>%
  summarise(Originator = n())

last_touch <- data %>%filter(Positionname == "CONVERTER") %>%
  select(Groupname, Positionname) %>%
  group_by(Groupname) %>%
  summarise(Converter = n())
```

```
## merge first touch and last touch into one table
tab1 <- left_join(first_touch, last_touch)
```

```
## Joining, by = "Groupname"
```

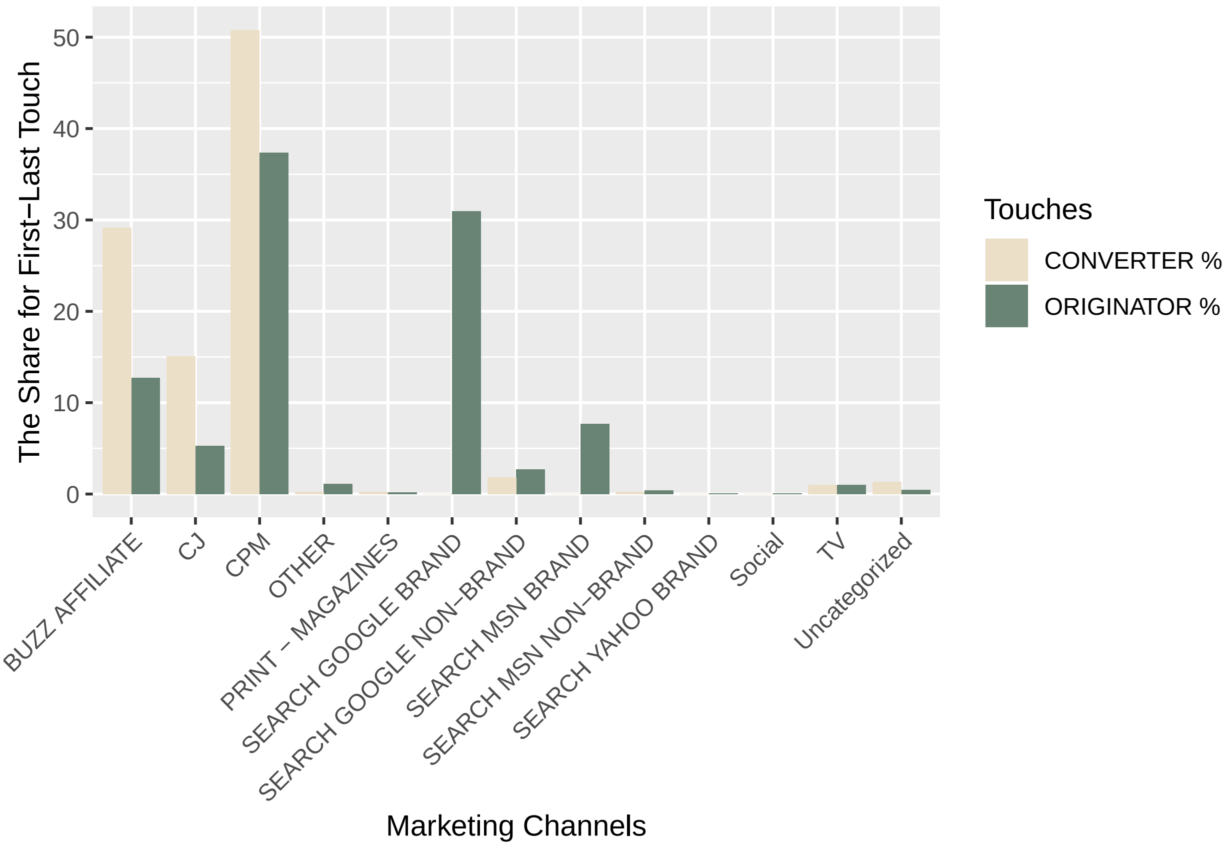
```
## calculate the persentage for converter and originator
tab2<-tab1 %>% mutate(Originator_per = Originator/sum(Originator) *100,
                Converter_per = Converter/sum(Converter, na.rm = T)*100)
## reorder the column
```

```r
tab2 <- tab2 %>%
  select(Groupname, Converter, Originator, Converter_per, Originator_per)
tab2[is.na(tab2)] <- 0
colnames(tab2) <- c("Groupname", "CONVERTER", "ORIGINATOR", "CONVERTER %", "ORIGINATOR %")
TOT <- list("TOTAL", sum(tab2$CONVERTER), sum(tab2$ORIGINATOR),
            sum(tab2$`CONVERTER %`), sum(tab2$`ORIGINATOR %`))

table <- rbind(tab2, TOT)
kable(table)
```

Groupname	CONVERTER	ORIGINATOR	CONVERTER %	ORIGINATOR %
BUZZ AFFILIATE	443	193	29.1831357	12.7140975
CJ	229	80	15.0856390	5.2700922
CPM	771	567	50.7905138	37.3517787
OTHER	4	17	0.2635046	1.1198946
PRINT - MAGAZINES	4	3	0.2635046	0.1976285
SEARCH GOOGLE BRAND	0	470	0.0000000	30.9617918
SEARCH GOOGLE NON-BRAND	28	41	1.8445323	2.7009223
SEARCH MSN BRAND	0	117	0.0000000	7.7075099
SEARCH MSN NON-BRAND	4	6	0.2635046	0.3952569
SEARCH YAHOO BRAND	0	1	0.0000000	0.0658762
Social	0	1	0.0000000	0.0658762
TV	15	15	0.9881423	0.9881423
Uncategorized	20	7	1.3175231	0.4611331
TOTAL	1518	1518	100.0000000	100.0000000

1_b: In a single bar plot, plot the share of credit (in percentage) for the first- and last-touch attribution models by marketing channel.

```r
library(ggplot2)
## Wrangling data for plot
tab3<-tab2 %>% select(Groupname, 'CONVERTER %', 'ORIGINATOR %') %>%
  gather(Touches, Number_of_touches, 'CONVERTER %':'ORIGINATOR %')
## plot
tab3 %>%
  ggplot(aes(Groupname, Number_of_touches, fill = Touches ))+
  geom_col(position = "dodge") +
  scale_fill_manual(values=c("#ecdfc8", "#698474"))+
  xlab("Marketing Channels")+
  ylab("The Share for First-Last Touch")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



1_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: The last-touch mostly distributes on the channels like Buzz Affiliate, CJ, CPM, but there are also a lot of first-touch happens among these channels. Meaning that these last-touches somehow are correlated with first-touch. Moreover, there are some search engine channel generate only first-touch in this case such as search google brand and search MSN brand. One of my assumption is people search the Winters brand from these search engine channels which bring them to the other channels and end up purchase. Therefore, The consequence is that Winters may not be able to generate sales as much as now if it allocated its marketing budget entirely on the basis of the last-touch attribution model. Because a lot of time last-touch rely on first-touch and first-touch spur customer to the website and eventually convert.

2-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?

```
data %>% filter(Positionname == "ORIGINATOR") %>%
  select(Newcustomer, DaysToConvert) %>%
  group_by(Newcustomer) %>%
  summarise(avg_DayToConvert = mean(DaysToConvert))
```

```
## # A tibble: 2 x 2
##   Newcustomer avg_DayToConvert
##   <chr>                  <dbl>
## 1 N                      29.0
## 2 Y                       6.14
```

Answer: The average number of days that it takes for a new customer to convert is about 6 days. And it takes about 29 days for a old customer to convert.

2_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.

```
## create the Touches variable
new_data <- data %>% group_by(Orderid) %>% mutate(Touches = n())
new_data %>% filter(Positionname == "ORIGINATOR") %>%
  select(Newcustomer, Touches) %>%
  group_by(Newcustomer) %>%
  summarise(avg_touches = mean(Touches))
```

```
## Adding missing grouping variables: `Orderid`
```

```
## # A tibble: 2 x 2
##   Newcustomer avg_touches
##   <chr>             <dbl>
## 1 N                  5.18
## 2 Y                  4.25
```

Answer: The average number of touchpoints by new customer's order is about 4 times versus by old customer's orders is about 5 times.

2_c: What is the average order sales amount by new versus old customer's orders?

```
## calculate the average sales for new and old customers
avg_sales<- data %>% filter(Positionname == "ORIGINATOR") %>%
  group_by(Newcustomer) %>%
  summarise(avg_sales = mean(Saleamount))
print(avg_sales)
```

```
## # A tibble: 2 x 2
##   Newcustomer avg_sales
##   <chr>           <dbl>
## 1 N                207.
## 2 Y                272.
```

```
## calculate the different
diff <- avg_sales[2,2]-avg_sales[1,2]
print(diff)
```

```
##   avg_sales
## 1  64.54152
```

Answer: The average order sales amount by new customers is 271.9467 in dollars and by old customer is 207.4052 in dollars. New customers generate 64.54152 in dollars more in average sales than old customers.

2_d: Summarize how new and old customers differ along these three variables.

Answer: It usually takes 23 days more to get old customer to convert than new customers and new customers generate more 64.54152 in dollars sales than old customers. And, it takes 4 times as average number of touchpoints for new customers to convert, but it takes 5 times for old customers to convert. As a result, new cutsomers are more effient to convert than old customers and also new customers generate more sales.

3_a: Create a table (as in Q1) containing the average sales per order as well as the total revenue by originator channel.

```
## calculate average sales and the total sales by orginator channel
tab4<-data %>% filter(Positionname == "ORIGINATOR") %>%
  select(Orderid, Saleamount, Groupname, Positionname) %>%
  group_by(Groupname) %>%
  summarise(Average_sales = mean(Saleamount, na.rm = T),
            Total_sales = sum(Saleamount))
Total4<-list("TOTAL", sum(tab4$Average_sales), sum(tab4$Total_sales))
tab5 <-rbind(tab4, Total4)
kable(tab5)
```

Groupname	Average_sales	Total_sales
BUZZ AFFILIATE	258.0513	49803.90
CJ	262.3790	20990.32
CPM	242.2739	137369.32
OTHER	227.4729	3867.04
PRINT - MAGAZINES	324.6533	973.96
SEARCH GOOGLE BRAND	250.3493	117664.16
SEARCH GOOGLE NON-BRAND	234.5817	9617.85
SEARCH MSN BRAND	229.0900	26803.53
SEARCH MSN NON-BRAND	274.9000	1649.40
SEARCH YAHOO BRAND	258.4900	258.49
Social	165.0000	165.00
TV	239.2540	3588.81
Uncategorized	200.4157	1402.91
TOTAL	3166.9112	374154.69

3_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 of every $1 in sales is incremental.

```
## calculate the total sales for unbrand and brand

tab6<-data %>% filter(Positionname == "ORIGINATOR") %>%
  group_by(Groupname, Brand) %>%
  summarise(Total_sales = sum(Saleamount)) %>%
  mutate(Incremental_Gross = Total_sales * 0.4*ifelse(Brand == "Y", 0.05, 0.1))

kable(tab6)
```

Groupname	Brand	Total_sales	Incremental_Gross
BUZZ AFFILIATE	N	49803.90	1992.1560
CJ	N	20990.32	839.6128
CPM	N	137369.32	5494.7728
OTHER	N	630.32	25.2128
OTHER	NULL	677.13	27.0852
OTHER	Y	2559.59	51.1918
PRINT - MAGAZINES	N	973.96	38.9584
SEARCH GOOGLE BRAND	Y	117664.16	2353.2832
SEARCH GOOGLE NON-BRAND	N	9617.85	384.7140
SEARCH MSN BRAND	Y	26803.53	536.0706
SEARCH MSN NON-BRAND	N	1649.40	65.9760
SEARCH YAHOO BRAND	Y	258.49	5.1698
Social	N	165.00	6.6000
TV	NULL	3588.81	143.5524
Uncategorized	N	1402.91	56.1164

```r
## Compare the incremental gross for branded search or else
c<-tab6 %>% group_by(Brand) %>%
  summarise(Incremental_Gross_by_Brand = sum(Incremental_Gross))
c
```

```
## # A tibble: 3 x 2
##   Brand Incremental_Gross_by_Brand
##   <chr>                      <dbl>
## 1 N                          8904.
## 2 NULL                        171.
## 3 Y                          2946.
```

Answer: The total incremental gross revenue by originator channel for branded search is 2945.7154, and 8904.1192 for the Non-brand search channels and 170.6376 for the remaining channels.

3_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?

```r
## calculate the loss on ads on brand search advertising
Loss_on_brandsearch <- c$Incremental_Gross_by_Brand[c$Brand == "Y"] - 4200
Loss_on_brandsearch
```

```
## [1] -1254.285
```

Answer: The loss on brand search is -1254.285 in dollars. I would advise the search team not to put too much efforts and spend less money on brand search advertising based on the result.

4_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.

```r
## Calculate the total of the linear attribution shares
Attribution_share <- data %>% add_count(Orderid) %>%
  mutate(Linear_attribution_share = 1/n) %>%
```

```r
  group_by(Groupname) %>%
  summarise(Total_Linear_Attribution_Share = sum(Linear_attribution_share))

Attribution_share
```

```
## # A tibble: 13 x 2
##    Groupname                Total_Linear_Attribution_Share
##    <chr>                                             <dbl>
##  1 BUZZ AFFILIATE                                  303.
##  2 CJ                                              135.
##  3 CPM                                             771.
##  4 OTHER                                             7.00
##  5 PRINT - MAGAZINES                                 2.93
##  6 SEARCH GOOGLE BRAND                             194.
##  7 SEARCH GOOGLE NON-BRAND                          24.1
##  8 SEARCH MSN BRAND                                 49.1
##  9 SEARCH MSN NON-BRAND                              4.11
## 10 SEARCH YAHOO BRAND                                0.429
## 11 Social                                            0.375
## 12 TV                                               12.4
## 13 Uncategorized                                    14.6
```

```r
## validate the total shares
Total_shares <- sum(Attribution_share$Total_Linear_Attribution_Share)
Total_shares
```

```
## [1] 1518
```

```r
Total_shares == length(unique(data$Orderid))
```

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

```r
## Calculate the corresponding share of construction
Share_credit <- Attribution_share %>%
  mutate('Share of Credit %' = (Total_Linear_Attribution_Share/Total_shares * 100))

Total<-list("TOTAL", Total_shares, sum(Share_credit$`Share of Credit %`))
Linear_attribution_model <- rbind(Share_credit, Total)

## aggregate to a table
kable(Linear_attribution_model)
```

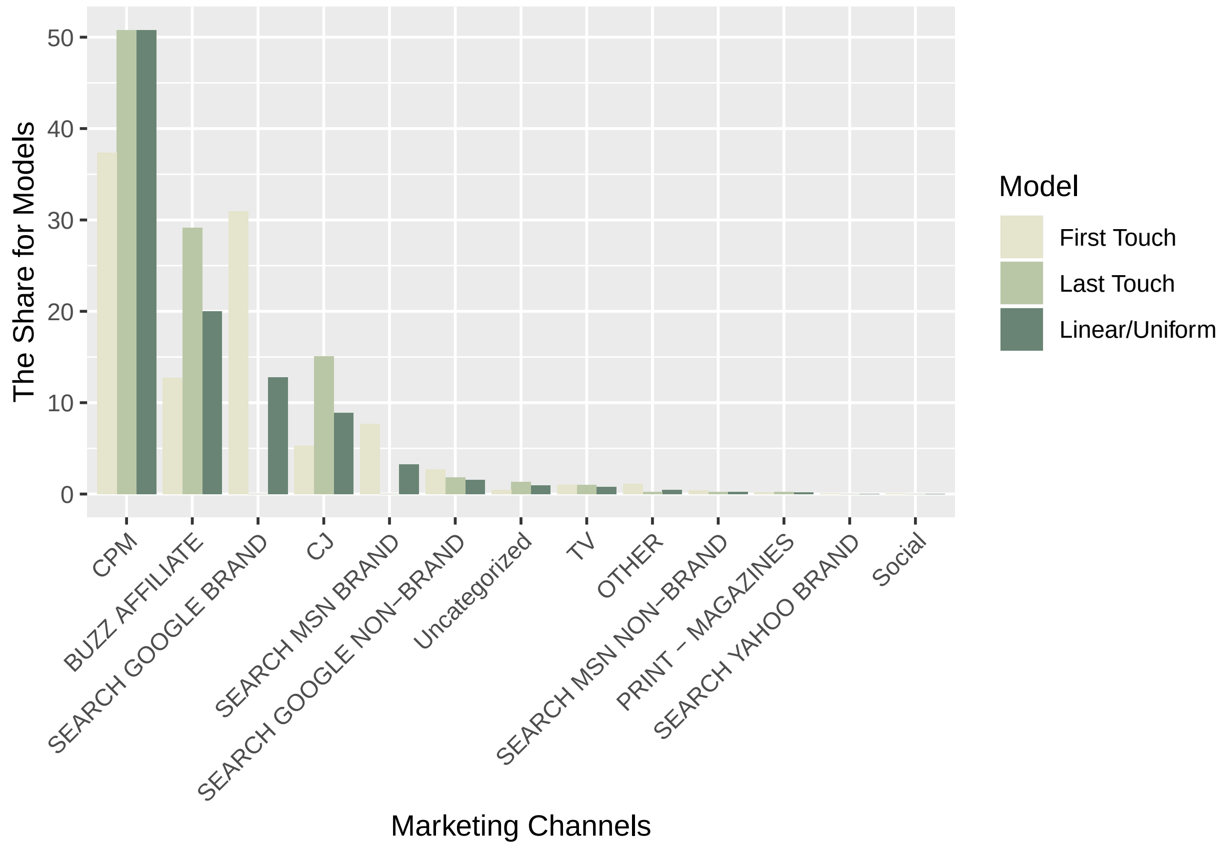
Groupname	Total_Linear_Attribution_Share	Share of Credit %
BUZZ AFFILIATE	303.4789683	19.9920269
CJ	134.6825397	8.8723676
CPM	770.7567460	50.7744892
OTHER	6.9976190	0.4609762
PRINT - MAGAZINES	2.9261905	0.1927662
SEARCH GOOGLE BRAND	194.0670635	12.7843915
SEARCH GOOGLE NON-BRAND	24.0896825	1.5869356
SEARCH MSN BRAND	49.1194444	3.2358000
SEARCH MSN NON-BRAND	4.1055556	0.2704582
SEARCH YAHOO BRAND	0.4285714	0.0282326
Social	0.3750000	0.0247036
TV	12.3952381	0.8165506
Uncategorized	14.5773810	0.9603018
TOTAL	1518.0000000	100.0000000

4_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.

```
## Combine first_touch, last_touch and linear/uniform in one table
tab7<-tab2 %>% select(Groupname, 'CONVERTER %', 'ORIGINATOR %')
tab8<-inner_join(tab7,Share_credit[,-2])
```

```
## Joining, by = "Groupname"
```

```
## plot
tab9<-tab8 %>% gather(Model, Share_percentage, 'CONVERTER %':`Share of Credit %`, na.rm = T)

tab9$Model[tab9$Model == "CONVERTER %"] <- "Last Touch"
tab9$Model[tab9$Model == "ORIGINATOR %"] <- "First Touch"
tab9$Model[tab9$Model == 'Share of Credit %'] <- "Linear/Uniform"

tab9 %>%
  ggplot(aes(x = reorder(Groupname, -Share_percentage*(Model == "Linear/Uniform")),
             y = Share_percentage, fill = Model))+
  geom_col(position = "dodge")+
  xlab("Marketing Channels")+
  ylab("The Share for Models")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  scale_fill_manual(values = c("#e5e4cc", "#bac7a7","#698474"))
```



4_c:Compare the linear model to the first-touch and last-touch models.

Answer: The linear/uniform model is more reasonable to use measuring campaigns as it appears among all the marketing channels based on the graph above. On the other hand, it's hard to measure campaigns based on comparing first touch and last touch because not all the channels appear both first and last touches.

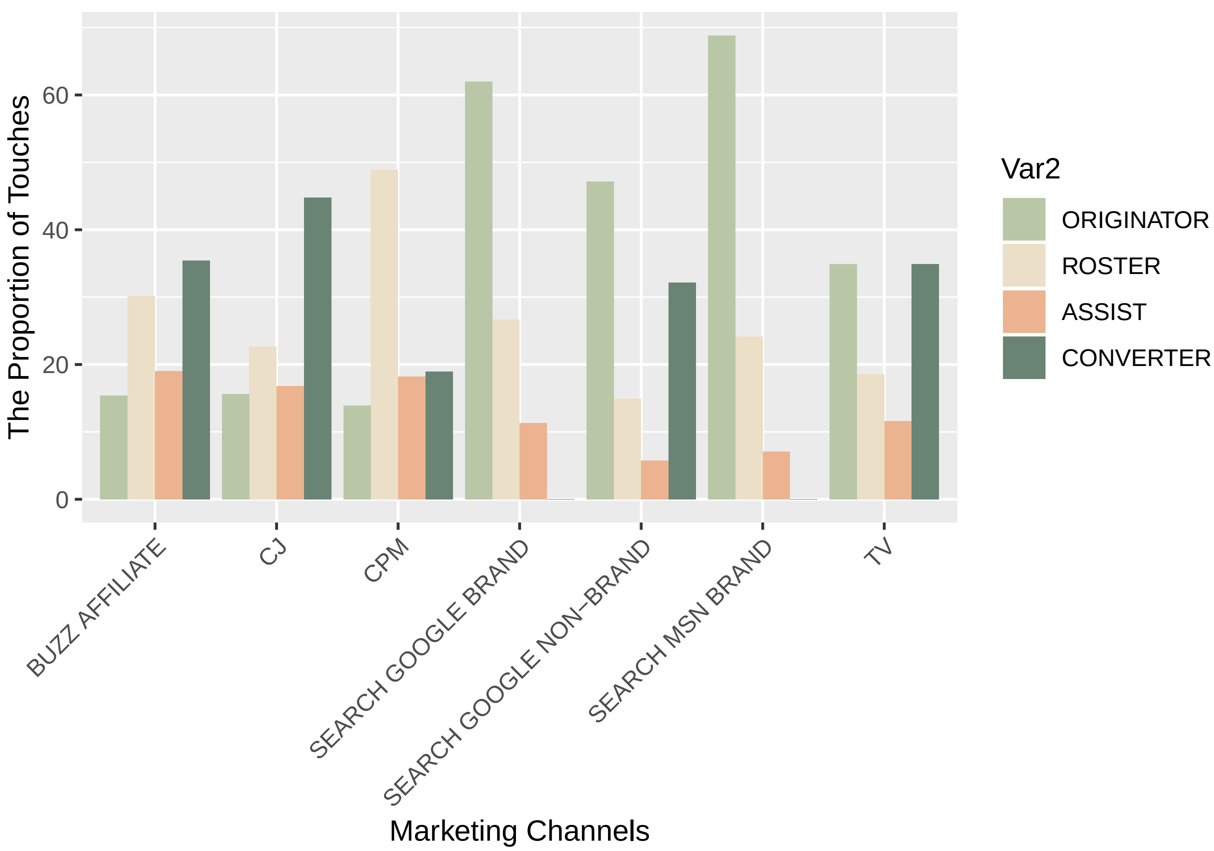
5_a:Focusing on the top channels listed below, what is the proportion of each channel's touchpoints by position name: 1) Originator, 2) Roster, 3) Assist, and 4) Converter. For full credit, the rows must be listed in that order.

```
top_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") %>%
  select(Groupname, Positionname, Touches)

tab10<-prop.table(table(top_channel$Groupname, top_channel$Positionname),1) *100
TOTAL <- rowSums(tab10, na.rm = T)
tab11<-cbind(tab10, TOTAL)
tab11<-tab11[,c(3,4,1,2,5)]
kable(tab11)
```

	ORIGINATOR	ROSTER	ASSIST	CONVERTER	TOTAL
BUZZ AFFILIATE	15.42766	30.13589	19.024780	35.41167	100
CJ	15.65558	22.70059	16.829746	44.81409	100
CPM	13.92778	48.90690	18.226480	18.93884	100
SEARCH GOOGLE BRAND	62.00528	26.64908	11.345646	0.00000	100
SEARCH GOOGLE NON-BRAND	47.12644	14.94253	5.747126	32.18391	100
SEARCH MSN BRAND	68.82353	24.11765	7.058823	0.00000	100
TV	34.88372	18.60465	11.627907	34.88372	100

5_b: In a single bar plot, plot the share in percentage (y-axis) of touchpoint types by marketing channels (x-axis).

```
tab12<- as.data.frame(tab10)
tab12$Var2 <- factor(tab12$Var2,
                     levels = c("ORIGINATOR", "ROSTER", "ASSIST", "CONVERTER"))

tab12 %>% ggplot(aes(Var1, Freq, fill = Var2))+
  geom_col(position = "dodge")+
  xlab("Marketing Channels")+
  ylab("The Proportion of Touches")+
  scale_fill_manual(values = c("#bac7a7","#ecdfc8","#ecb390", "#698474"))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



5_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: Based on the graph above, BUZZ AFFILIATE, CJ, and CPM seem to have relatively more of its touchpoints as rosters and assist, and SEARCH GOOD NON-BRAND seems to have relatively less.

SEARCH GOOGLE BRAND and SEARCH MSN BRAND would receive too much credit under first touch but zeon credit under last touch. Also, BUZZ ADDILIATE and CJ would receive more credit under last touch as apposed to other touchpoints.