Paid Search Advertising Data for Hotels

An Executive Summary

Anamika Soni ~ John Montgomery ~
Mustafa Mudheher ~ Nissim Panchpor ~
Shrija Priyanil ~Snehal Patil

OVERVIEW

Our final project focused on analyzing paid search advertising data for hotels, in the US and worldwide. Paid search marketing (also known as pay per click marketing) is where an organization is able to essentially bid on keywords/key-phrases, and when a customer searches for these words, an organization's ad shows up on the page based on their initial bid. In this dataset, we had 301 keywords/key-phrases used by hotels to display their ads to potential customers. The click data is for the month of April (2004), and each set of keywords and key-phrases is associated with specific characteristics such as location, branding etc. It includes the daily frequency of clicks, accumulated cost of the clicks, average position of the ad and impression of the ads.

Paid search ads cost the advertiser a defined amount each time someone clicks on that result. Cost per click can range from 50 cents to several dollars per click, depending on various factors and search engine algorithms. Paid search ads help to increase organizations' visibility on search engines and traffic to their websites, and ultimately sales. The position of one's paid ad on Google is determined not only by the highest bid, but also by the quality score, which is a measure of how relevant your keyword is to your ad text and to what a user is searching for. Another key factor is that users who click on paid search ads are usually the ones who are actively seeking your product or service (hotel reservation in this case), since paid search ads target specific search queries and target users who are highly likely to buy the product.

Paid search advertising is a great way to leverage the success of search engines and help grow ones business. The aim of our project is to analyze the efficacy of paid search budget as well as analyze the search phrases for their efficiency and relevance, by using the methods and techniques learned in this class.

DESCRIPTION OF THE DATASET

The data was made available to us in the form of several Excel files which contained the raw data and descriptive information about the keywords/key-phrases. We were able to combine the files into a single data frame using R, and the unique ID's provided within the data. The combined dataset now contained 8498 observations, with both generic and branded type of keyword search phrases. The data included variables such as:

- **Keyword ID**: unique identifier for the keyword/key-phrase
- **Keyword Type ID**: unique identifier for the keyword type (i.e. Generic and Branded)
- **Average position**: the display position of the paid search ad on the search results page
- **Clicks**: the number of clicks received by an advertisement on a particular day
- **Impressions**: the number of times the ad was fetched and displayed to the user
- **Cost**: total cost to the advertiser on a particular day (i.e. Cost of the clicks)
- **Reservations**: this refers to the number of reservations made online through the ad (sales)
- **Category ID**: unique identifier for keyword category

- **Secondary category ID**: unique identifier for keyword's secondary category

To perform a thorough analysis of the keywords, we calculated the following data points from the existing variables available:

- **CPC (Cost Per Click)**: This was measured using the following equation:
  (Total Cost of Clicks)/(Total Number of Clicks) = Cost Per Click

- **CTR (Click Through Rate)**: This is a measure of performance, calculated in the following manner:
  (Total Clicks on Ad)/(Total Number of Impressions) = Click Through Rate
  This is an important variable since it is also used to determine the keyword's quality score which in turn enables the advertiser to achieve a higher ad position at a lower cost.

- **Conversion Rate**: Another measure of performance, conversion rate is calculated in the following manner:
  (Total Number of Reservations Made Through Ad)/(Total Number of Clicks) = Conversion Rate
  This is an important variable since it helps to measure the performance of the PPC campaign.

VARIABLE SELECTION AND REGRESSION ANALYSIS

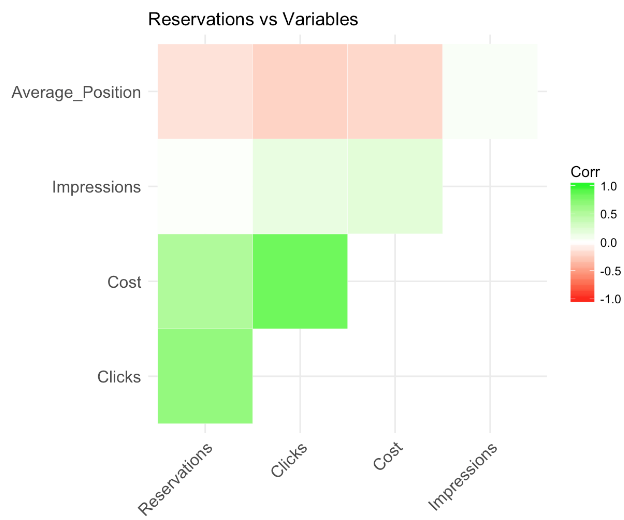

Fig. 1 Correlation Plot

To better understand how the given variables interact with each other, we created a correlation plot in R (Fig. 1 and Fig. 2) to see which variables are most correlated before going through and creating different models. We found that the most correlated variables were **Cost and Clicks**. This makes sense because search engine ads usually charge by the click. The next most correlated variables were Reservations and clicks then **Reservations and Cost**. This also makes sense because a click should lead to a reservation or else businesses would not waste their money paying for clicks. *Impressions were not correlated with reservations at all and only slightly correlated with cost and clicks*. This was somewhat surprising because one would think that an ad with lots of impressions would have some impact on reservations but this was not the case. The correlation of Average Position to Reservations, Cost, and Clicks were slightly negative. This also makes sense because the lower the average position, the higher it is on the page. A high page position will have the most visibility and receive more clicks. Now that we identified the correlations between variables, we started making models keeping in mind highly correlated variables in the same model will not add value or make it a better model.

Reservations	1.00
Avg. Position	-0.15
Impressions	0.02
Clicks	0.66
Cost	0.51
Page	-0.04

Fig. 2 Correlation Table

Linear Models

Using linear models, we wanted to find a model with a good fit for reservations being the dependent variable. To do this, we ran linear models with Reservations as the dependent variables Average Position, Impressions, Clicks, and Cost separately as the independent variables. We found that Reservations ~ Clicks had the highest R^2 at 43.7%.

Continuing with linear models, we wanted to test concave logarithmic models as well as concave quadratic models. To do this, we created new log and squared variables. With these new variables, we tested Reservations with multiple combinations of independent variables. We found that Reservations ~ Cost + $Clicks^2$ was a higher R^2 value of 47.05%. However, since Cost and Clicks are highly correlated , dependent variables, we decided to remove Cost and just run Reservations ~ $Clicks^2$ which gave us an R^2 value of 46.7%. The additional models we ran can be found in the appendix.

Furthermore, our team created a lagReservation variable with 1 and 7 days as the lag to test if there is any carryover effect. Our intuition behind this hypothesis was that potential customers might search to book hotels for a certain period of time, and then (a day or a week) they actually book their stay. Unfortunately this analysis did not enhance our R^2 and different carryover models did not provide better data explanation from the previous models.

Accordingly, our best model explaining the behavior of online reservations can be equated with the following formula:

$$Reservations = 0.018 + (0.015 * Clicks^2)$$

Further analysis took place to investigate the behavior of our sole parameter - the number of clicks. Our team followed the same approach we applied on the online reservations and applied it on the clicks variable. Several models were created to understand its behavior and we found out that click were highly correlated with log values of both average positions and impressions. It is also worth mentioning that we eliminated the option of using clicks as a significant factor because of the high dependency between both clicks and costs - the cost factor is highly dependent on number of clicks.

The clicks model had an R2 of 0.47 and is equated by the following equation

$$Clicks = 3.43 - 3.23 * \log(Average\ Position) + 0.93 * \log(Impressions)$$

DIVING DEEPER INTO THE DATA:

Our team looked deeper into the significant parameters - Clicks, Impression and Average positions in an attempt to understand the behavior of each factor, aside from the regression models provided earlier.

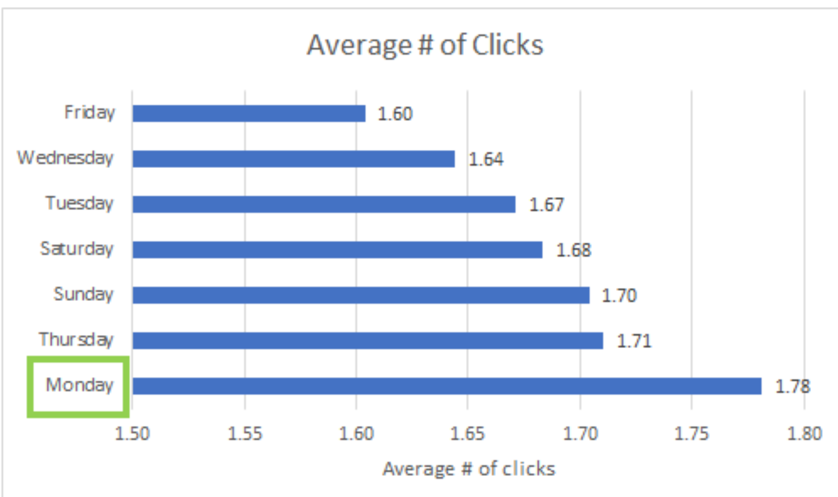

Fig. 3 Average number of clicks by day of the week

The given graphs attempted to analyze keyword trends based on the day of the week. We observed that the data was not uniformly distributed and there were several keyword phrases that did not have any observations on multiple days - probably since they did not win the bid or were not selected by the search engine. Fig 3, 4 and 5 helped us to understand that most users tend to search for hotels on Mondays and Tuesdays -- indicating users tendencies to plan their trips at the beginning of the week, and possibly after weekends when they discussed initial plans with their social groups. These graphs also helped to confirm that these hotel advertisers were performing good cost optimization, the bids were less on the most popular days i.e. days with higher number of clicks. For example, it would be most feasible that a hotel posts its ads on a Monday because it has the highest number of clicks, a significant factor to the online reservations, and it has the second to lowest number of bids by the hotels industry; possibly resulting in the lowest bidding price.

Fig. 3 Average number of clicks by day of the week

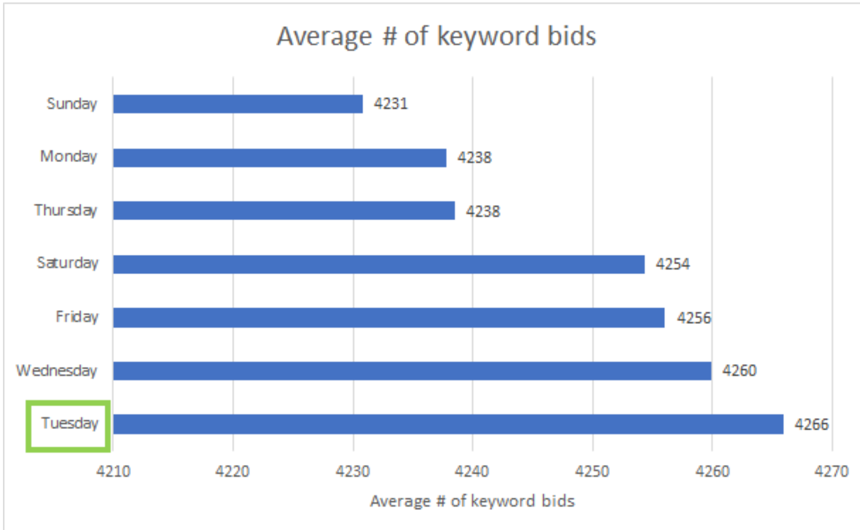

Fig. 4 Average number of keyword bids by day of the week

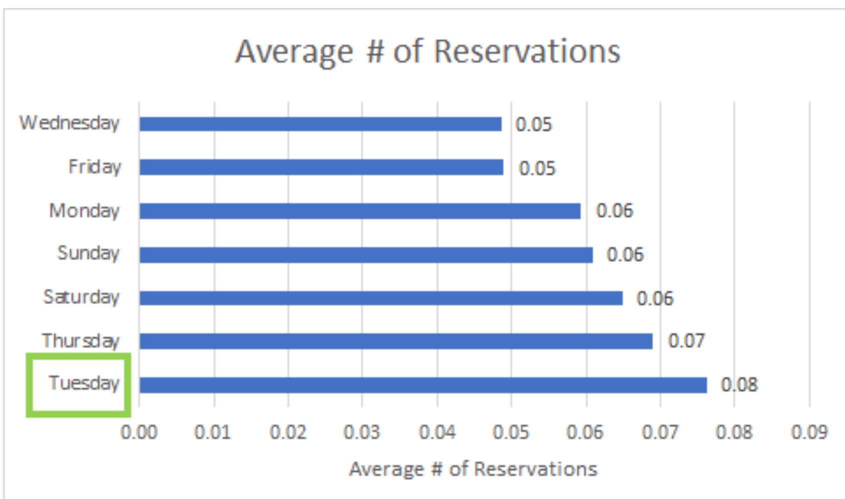

Fig. 5 Average number of reservations by day of the week

We were curious about analyzing category and secondary category of keywords by measures such as clicks, CTR and reservations. We mainly used tree maps for this analysis.

What is a tree map?
A tree map visualizes one dimension of your data with typically two of the measures in data. The dimensions are denoted by rectangles and measures are denoted by using two parameters - size and color of each rectangle.

First, we compared average CTR and count of keywords for each category using a tree map. In the tree map, color of a rectangle denotes average CTR and size of rectangle denotes count of keywords. So, larger a rectangle, higher the count of keywords for that category and darker the shade of color, higher the average CPR. As per this tree map, keywords in 'Top 25 Keywords' category have a higher CTR on an average.

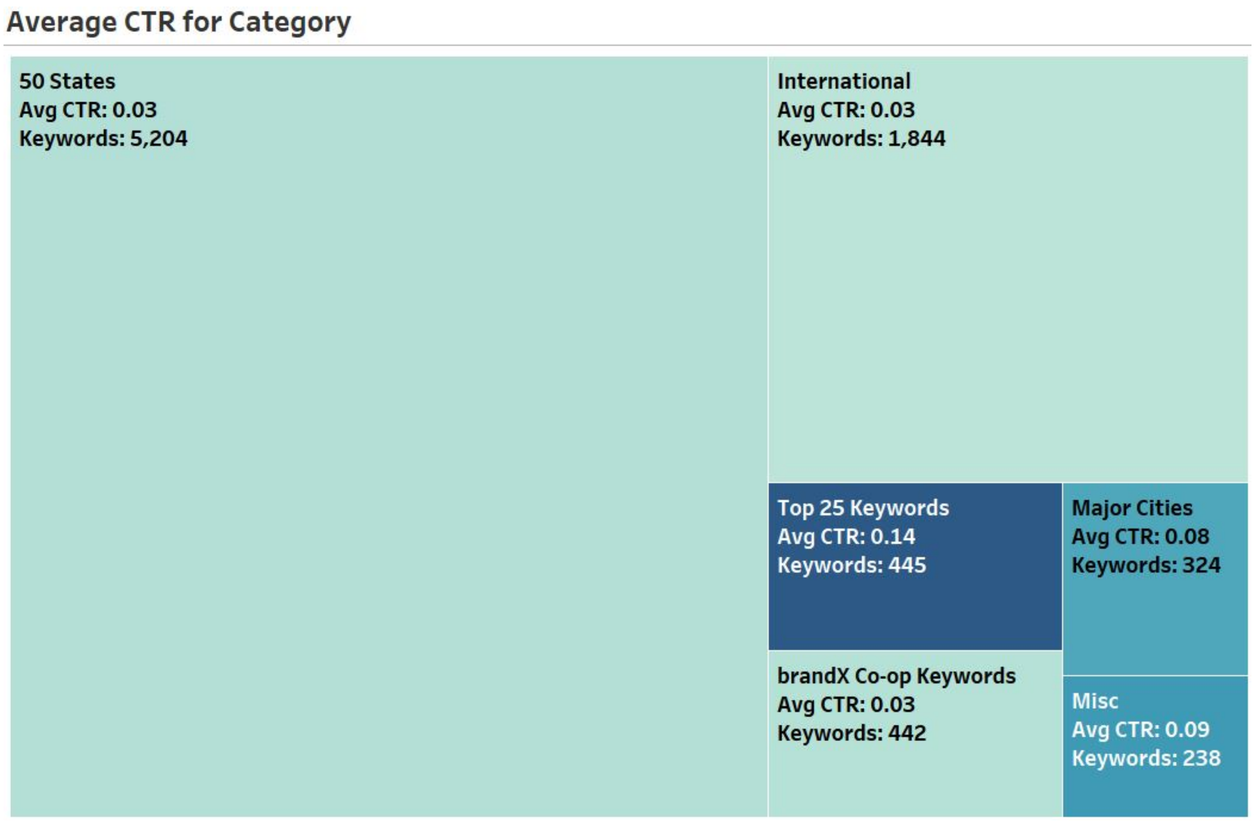

Fig. 6 Average CTR by Category

Now, if we drill down to secondary category '50 States', we can analyze hotel keywords belonging to which state are well. We created a bar plot of these keywords by sum of their respective clicks. Figure 7 shows that keywords belonging to hotels in Florida are receiving the highest average number of clicks.

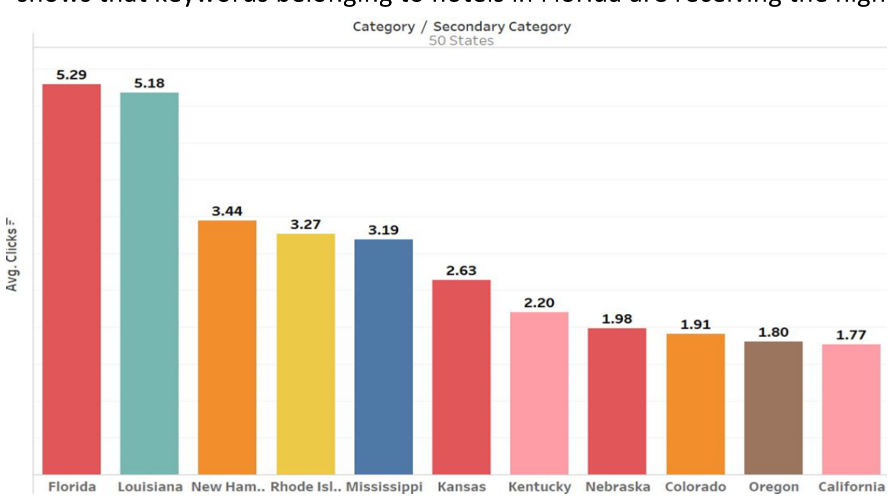

Fig. 7 Average number of clicks by Secondary Category

For further analysis, we built a tree map for secondary category '50 States' for measures average CTR and count of keywords. As seen from the tree map, California has a higher count of hotel keyword occurrences than Florida, but keywords in Florida have a higher CTR on an average. We think this is because Florida is seen as more of a tourist destination than California. People might be researching hotels in Florida more as they prepare for a summer vacation over the next quarter.

Also, keywords belonging to hotels in Alaska have the highest average CTR among all the states. However, these keywords only have 4 clicks from 16 total impressions. Thus, we think it is not a good idea to assess the performance of these keywords yet.

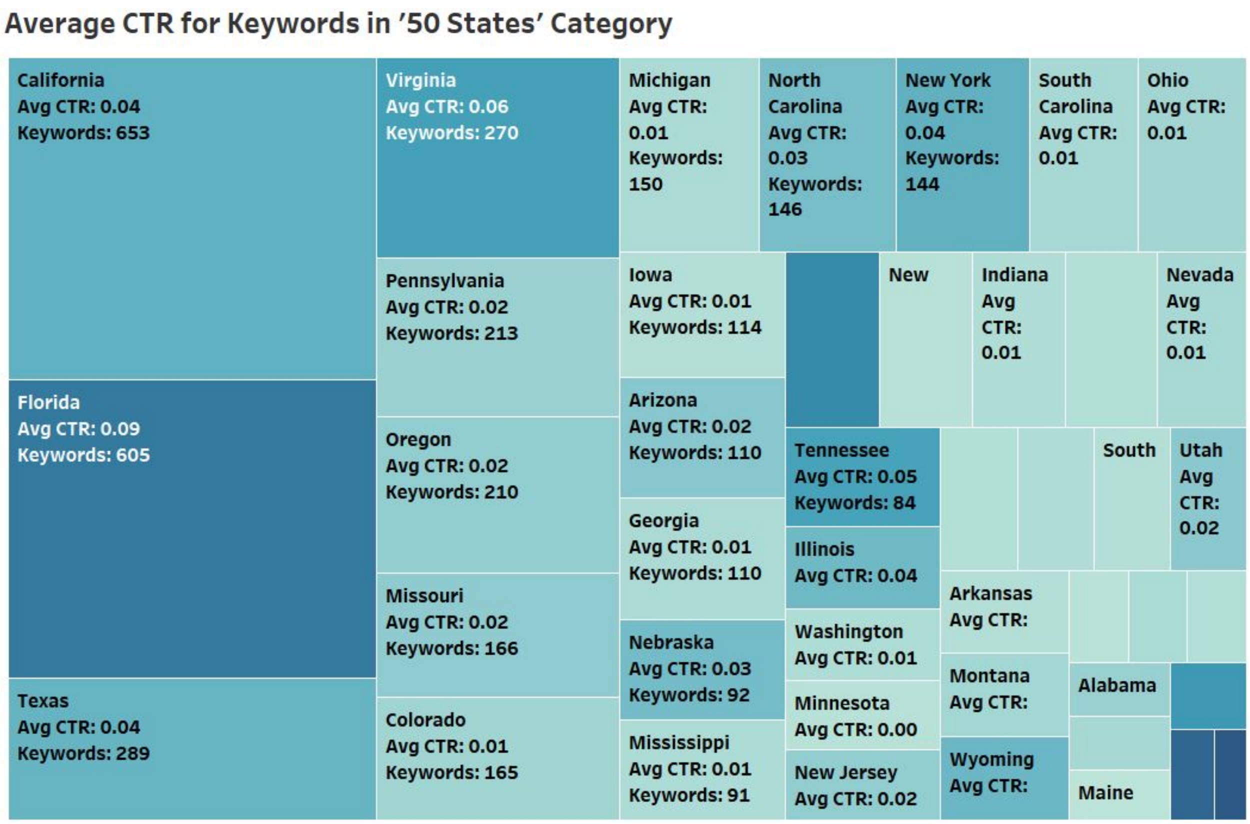

Fig. 8 Average CTR for Keywords in the ’50 States’ Category

Similarly, we created a tree map for secondary category ‘International’. Keywords belonging to hotels in New Zealand have a higher CTR on average compared to other regions. Even though hotels in Europe have a really high occurrence at 1,119 keywords over the entire month, its average CTR is lower to many other regions such as New Zealand and Canada.

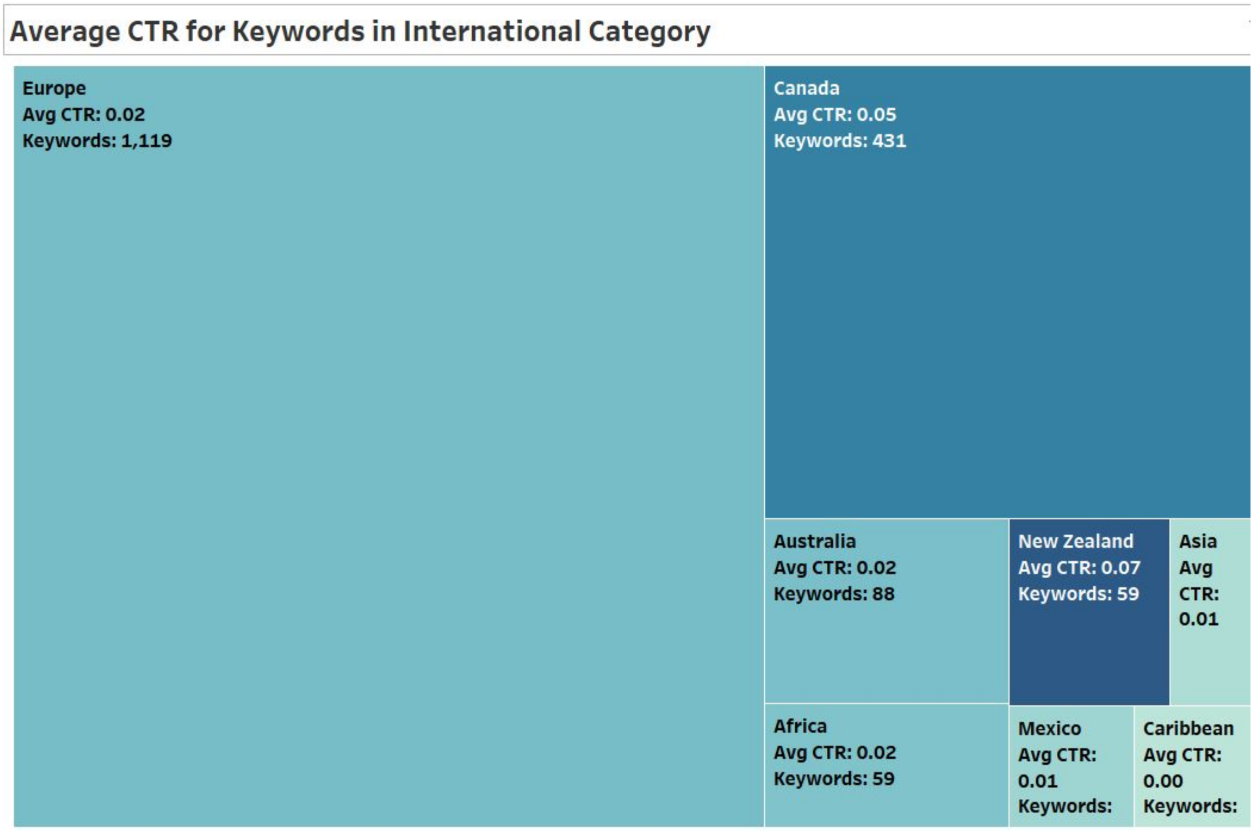

Fig. 9 Average number of clicks by Secondary Category

Keyword	Impressions	Clicks	CTR
london cheap hotels	11,325	19	0.001
innsbruck hotel	13,417	34	0.002
gent hotel	11,138	37	0.003
dortmund hotel	3009	10	0.003
canterbury hotel	2134	10	0.004

Fig. 10 Keywords with lowest CTR in Europe category

To analyze why Europe had high number of keywords but lower relative CTR, we analyzed keywords specific to Europe. Since average CTR for these keywords is not proportional to the number of keywords, it seemed that some keywords had high impressions but low clicks. Figure 9 shows the top 5 words with low CTR. For example, "london cheap hotels" had almost one click per thousand impressions. Since these keywords are shown to searchers but they are not getting as many clicks, we think that the content of the ads could be improved or , or the ads don't look good enough to attract clicks.

KEYWORD TYPE ANALYSIS:

Once we analyzed the keywords on the basis of the category they belonged to and the day of the week where they received highest number of clicks, we wanted to delve further into the keyword types and how the keywords/key-phrases performed when they belonged to one or the other keyword type. Theoretically, we know that branded keywords do better in a PPC ad campaign. Branded key-phrases contain the brand's name or some variation of it, to allow users to immediately find what they are looking for. This even allows the brand to save money since the organic links have a lower average position in the search results.

Keyword Type	Total Keywords	Total Clicks	Total Reservations	Average Conversion Rate
Generic	201	6011	100	0.56%
Branded	100	8293	418	1.77%

Fig. 11 Keywords split by type

On performing a simple analysis of the dataset, we found that one third of the keywords were of type branded. However, the performance of branded keywords was significantly better, as they resulted in 57.9% of the total number of clicks for all 301 keywords, approximately 80% of all the reservations made, and the average conversion rate was three times better than that of the generic keywords.

On further analysis we found that approximately 78.6% of the generic keywords had no reservations in the month of April, whereas only 59% of the branded keywords resulted in no reservations. The maximum number of reservations made for a generic keyword was 4, whereas for a branded keyword it was 10 reservations.

Digging deeper, we decided to perform revenue analysis on the data. Were there keywords that had clicks but did not result in reservations?

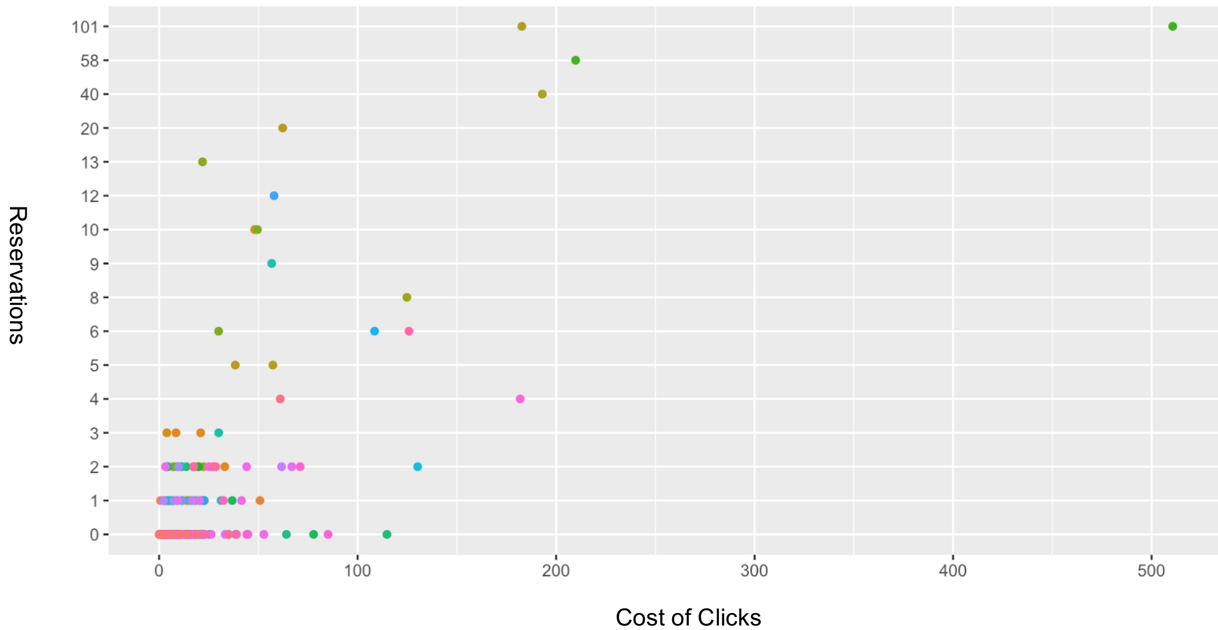

Fig. 12 Total Cost of Clicks vs Total Reservations

After plotting a graph of total cost of clicks vs total reservations (per keyword, as seen in figure 12) we saw a huge clump of keywords in the $0-$100 vs 0-2 reservations area of the graph.

Out of these keywords, we found that there were 217 keywords that were costing the bidder money, but not making them any revenue. The biggest example of this was "cocoa beach hotel" (generic type) with a total of 273 clicks in April 2004, but 0 reservations.

One of the biggest surprises with this analysis was that (as you can see in figure 13) there were a few branded words as well in this sample of keywords. This was surprising because branded key-phrases are used to attract customers who are looking for the particular brand, and one could intuitively say that if a customer searches for a particular brand, they might be pretty decided on making a reservation.

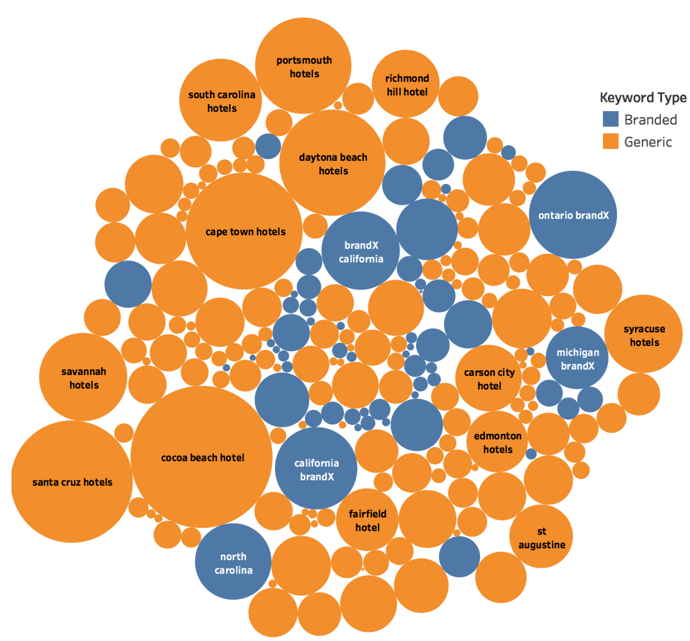

Fig. 13 Bubble chart to show keywords with 0 reservations

AREAS OF OPPORTUNITY:

To perform a full, thorough analysis on a paid search dataset, we realized we would need more data points, and so based on our analysis on the limited dataset available, we could recommend following areas of opportunity to optimize the PPC advertising results:

Identical keywords with significantly different click rate

Keywords	Cost per Click	# of Clicks
akron hotel	0.34	1
akron hotels	0.37	20
alexandria hotel	0.33	9
alexandria hotels	0.40	31
brandX hershey	0.36	1
brandX hershey pa	0.34	8
brandX inn	0.37	167
brandX inns	0.35	49
brandX motel	0.38	2
brandX motels	0.32	65
brandX reservation	0.27	16
brandX reservations	0.35	142
buena park hotel	0.26	45
buena park hotels	0.46	24
klamath falls hotel	0.40	25
klamath falls hotels	0.41	36
lincoln city hotels	0.33	174
lincoln hotel	0.50	4
lincoln hotels	0.37	4
santa cruz hotel	1.54	46
santa cruz hotels	2.30	37
Sun City hotel	0.43	6
sun city hotels	0.27	65

Fig. 14 Almost identical keywords with different click rates

Opportunity #1: Optimizing bids on similar keyword phrases -- we discovered similar keyword phrases, in terms of the words used and the category they belonged to, usually used by advertisers to widen their opportunity net, and compared their performance. Based on this analysis, we could recommend advertisers to perform similar exercise and optimize their bids for high-performing keyword phrases.

For example, if you look at figure 14, the keyword "hotels" seems to work much better than "hotel" in a search key-phrase, so instead of bidding on both, the advertisers should analyze whether both are generating comparable revenue and then try and optimize on one keyword over the other if needed. "akron hotel" received only one click in April, whereas "akron hotels" received 20 clicks.

Opportunity #2: Optimizing a keyword's quality score to secure a higher ad position – While our models showed that average position was not of statistical significance, we could clearly see our data that the lower the average position (higher up on the page) of the advertisement, the more clicks it received. Having a good quality score also lowers the cost per click as well as the cost per conversion. Cost per conversion is an important factor since its usually higher than the cost per click. Some factors that affect quality score are click through rate, relevance of the keyword/key-phrase to the ad group, and historical performance of your keyword/key-phrase.

Opportunity #3: It might be beneficial to review the 217 words that are costing the advertiser money on the clicks, but not generating any revenue. Further analysis on the customer behavior for these keywords would be useful, especially the branded words that are in this set of keywords/key-phrases since branded keywords are supposed to lead the customer to the action that you want them to perform.

Overall, we realize that this dataset needs to be more comprehensive in terms of the data points collected, and also needs to collected over the period of maybe 4-5 months for true analysis of the performance of the keywords.

REFERENCES:

Rutz, Oliver J., Bucklin, Randolph E., & Sonnier, Garrett P. (2012). A latent instrumental variables approach to modeling keyword conversion in paid search advertising.(Report). Journal of Marketing Research, 49(3), 306-319.

Marketing 562 – Consumer Analytics Slides

APPENDIX (Code for Models):

```
#######################################################################################
##########################
# Final Project- Hotel Data
#######################################################################################
##########################
#######################################################################################
##########################
# Set Path to Directory with Data
# 1) By Hand: setwd("C:/Users/orutz/Documents/2 UW/1 Data-driven Marketing (DDM)/1 Class/1
Lectures/0 CURRENT Version/2 Lecture Materials/6 Binary Logit")
# 2) Better: Set Directory to where your code is saved. I recommend saving your data in the SAME Folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# Clear All Variables & Clear the Screen
rm(list=ls())
cat("\014")

library(readxl)
rawdata_full_updated <- read_excel("C:/Users/johnw/OneDrive/Documents/Foster/MKTG 562-
Consumer Analytics/Final Project/rawdata_full_updated.xlsx")
View(rawdata_full_updated)

# Read in the Data
data.raw = rawdata_full_updated

# Explore the data
str(data.raw)
```

```
summary(data.raw)

#Create Log Data
data.raw$logAverage_Position <- log(data.raw$Average_Position)
data.raw$logImpressions <- log(data.raw$Impressions)
data.raw$logCost <- log(data.raw$Cost)
data.raw$logClicks <- log(data.raw$Clicks)
data.raw$logReservations <- log(data.raw$Reservations)

#Create Clicks^2
data.raw$Clicks2 <- (data.raw$Clicks)^2
data.raw$Average_Position2 <- (data.raw$Average_Position)^2
data.raw$Impressions2 <- (data.raw$Impressions)^2
data.raw$lagClicks2 <- (data.raw$lagClicks)^2
data.raw$lagCost2 <- (data.raw$lagCost)^2


# Create LAG Sales
data.raw$lagReservations1 <- c(NA, head(data.raw$Date, -1))
data.raw$lagReservations7 <- c(NA,NA,NA,NA,NA,NA,NA, head(data.raw$Date, -7))
data.raw$lagClicks <- c(NA,NA,NA,NA,NA,NA,NA, head(data.raw$Clicks, -7))
data.raw$lagCost <- c(NA,NA,NA,NA,NA,NA,NA, head(data.raw$Cost, -7))
data.raw$lagCTR <- c(NA, head(data.raw$CTR, -1))


#Changing variables in DataSet
data.raw[, is.Page:= ifelse(Page=="First" | Page=="Second", "Weekend", "Weekday")]

##########################################################################################
##########################
# REGRESSION MODELS

# You need: y (Dependent Variable)
# For this example, Sales is used

# You need: X (Independent Variables)
# For this example, Advertising and Lagged Sales (i.e, a model with carry-over) are used

# Run the Regression (includes an INTERCEPT)
#Short Run response
# Simple linear model- Reservations~ Impressions
lm.model11 <- lm(Reservations ~ Average_Position, data = data.raw)
```

```
lm.model12 <- lm(Reservations ~ Impressions, data = data.raw)
lm.model13 <- lm(Reservations ~ Clicks, data = data.raw)
lm.model14 <- lm(Reservations ~ Cost data = data.raw)

summary(lm.model11)
summary(lm.model12)
summary(lm.model13) #Best
summary(lm.model14)

# Clicks R2=.437, Cost R2=.2647, AVG_POS R2=.02291

# Concave logarithmic model
lm.model15 <- lm(Reservations ~ logImpressions, data = data.raw)
lm.model16 <- lm(Reservations ~ logAverage_Position, data = data.raw)

summary(lm.model15)
summary(lm.model16)

#none have good fit

# Concave quadratic model
lm.model18 <- lm(Reservations ~ Cost + Impressions2, data = data.raw)
lm.model19 <- lm(Reservations ~ Cost + Clicks2, data = data.raw)
lm.model20 <- lm(Reservations ~ Cost + Average_Position2, data = data.raw)

summary(lm.model18)
summary(lm.model19)
summary(lm.model20)
# lm.model19 R2=.4698

#Long Run response
# Simple linear model
lm.model21 <- lm(Reservations ~ Clicks +lagReservations, data = data.raw)
# Concave logarithmic model
lm.model22 <- lm(Reservations ~ logImpressions +lagReservations1, data = data.raw)
# Concave quadratic model
lm.model23 <- lm(Reservations ~ Cost + Clicks2 +lagReservations1, data = data.raw)
#Best Model with a Longer Lag- significance goes down with every additional day
lm.model70 <- lm(Reservations ~ Cost + Clicks2 +lagReservations7, data = data.raw)

lm.model24 <- lm(Reservations ~ Average_Position + lagReservations1, data = data.raw)
lm.model25 <- lm(Reservations ~ Average_Position + Clicks2+ lagReservations1, data = data.raw)
```

```
# Display Results

summary(lm.model21)
summary(lm.model22)
summary(lm.model23)
summary(lm.model24)
summary(lm.model25)

summary(lm.model70)

#lm.model23 R2=.4705
#Conversion Rate regressions
lm.model26 <- lm(Conversion_Rate)

lm.model29 <- lm(Reservations ~  Clicks2, data = data.raw)

lm.model30 <- lm(Reservations ~ Cost + , data=data.raw) #R2 .467

lm.model31 <- lm(Conversion_Rate ~ CTR + Average_Position , data =data.raw)

summary(lm.model31)
summary(Lm.model30)
summary(lm.model29)### R2=.4671 BEST
#summary(lm.model5)
#summary(lm.model6)
#summary(lm.model7)
#summary(lm.model8)

lm.mode50 <- lm(Reservations ~ Average_Position + Impressions + Clicks + Cost, data = data.raw)

lm.mode51 <- lm(Reservations ~ Average_Position + Impressions + Clicks + Cost, data = data.raw)

# Display Results
summary(lm.mode50)
summary(lm.mode51)

lm.model52 <- lm(Reservations ~ Average_Position + Keyword_Type + Impressions + Clicks + Cost , data
= data.raw)

summary(lm.model52)
```

```
#####################################################################################
##########################
#Click as dependent Variable

lm.model33 <- lm(Clicks ~ Cost + Average_Position + Page, data=data.raw)

lm.model34 <- lm(Clicks ~ Average_Position + Page + Impressions, data=data.raw)
lm.model35 <- lm(Clicks ~ Average_Position2 + Page + Impressions, data=data.raw)
lm.model36 <- lm(Clicks ~ Average_Position2 + Page + Impressions2, data=data.raw)

lm.model37 <- lm(Clicks ~ logAverage_Position + Page + Impressions, data=data.raw)
lm.model38 <- lm(Clicks ~ logAverage_Position + logImpressions, data=data.raw)


summary(lm.model33)
summary(lm.model34)# R2=.098
summary(lm.model35)
summary(lm.model36)
summary(lm.model37)
summary(lm.model38)#R2=.2206

#lm.model33 Adjusted R-squared:  0.6859

plot(data.raw$Clicks2, data.raw$Reservations)


#####################################################################################
##########################
#Impressions as dependent Variable

lm.model38 <- lm(Impressions ~ lagCTR, data=data.raw)

summary(lm.model38)

#####################################################################################
##########################
#Split Data

data.estimation <- data.raw[1:2000,]

# Run the Binary Logit Model (includes an INTERCEPT)
glm.model <- glm(Reservations ~ Clicks + Cost, data = data.estimation)
```

```
# Display Results
summary(glm.model)

# Define HOLDOUT Data (i.e, IDs 201-500)
data.holdout <- data.raw[2001:8497,]
str(data.holdout)

# Predicting Reservations

(prediction.holdout <- data.frame(Rownum = data.holdout$Rownum,
                         LogitProbability   = predict(glm.model,data.holdout,type = c("response")),
                         LogitPredict = predict(glm.model,data.holdout,type = c("response"), digits = 0)))

sum(prediction.holdout["LogitPredict"])

sum(data.holdout$Reservations)

###### Prediction for CLICKS ###########

data.estimation <- data.raw[1:2000,]

# Run the Binary Logit Model (includes an INTERCEPT)
glm.model <- glm(Clicks ~ logAverage_Position + logImpressions, data = data.estimation)

# Display Results
summary(glm.model)

# Define HOLDOUT Data (i.e, IDs 201-500)
data.holdout <- data.raw[2001:8497,]
str(data.holdout)

# Predicting Reservations

(prediction.holdout <- data.frame(Rownum = data.holdout$Rownum,
                         LogitProbability   = predict(glm.model,data.holdout,type = c("response")),
                         LogitPredict = round(predict(glm.model,data.holdout,type = c("response")), digits =
0)))

sum(prediction.holdout["LogitPredict"])

sum(data.holdout$Clicks)
```

```
#####################################################################################
##########################
lm.model53 <- lm(Reservations ~ lagClicks2, data=data.raw)

summary(lm.model53)
```