Workbook

Part 1: Motivation

Problem Statement

- High Costs per click with SEM (Search Engine Marketing)
- Airline industry a Competitive market with Low margins

State the Questions

- Where do we allocate our marketing budget most efficiently?
- How can we reduce Cost/Click, increase revenue and optimize performance?
- Which search engine delivers the most ROI? (Manuel)
- what are the customer segments / search engine –> Specific pattern in buying behavior?

Main Objectives

- Find out profitability of campaigns / search engines / keywords
- Compare different bid strategies
- Which platform offers the most visibility?
- Find out single-click conversion rate of branded / unbranded keywords?

What could be a positive outcome?

- Minimize Cost/Click
- Maximize ROA
- Maximize Single-click conversion

Part 2: Method

What key resources do we acquire?

- Description Data:

Useful variables in the dataset (Type: xls)

- $impressions

Features of interest - costs per publisher - $Cost / Click - cost / $campaigns - costs / $bidstrategy

2. R Libraries

```
# Import Libraries
library(readxl)
library(tidyr)
library(plotly)
library(dplyr)
```

What is our approach to solve the problem?

High level process of steps

Part 3: Mechanics

Inspect & Import data

R tries to import the first sheet of the excel file which resolves in an error. This is why the argument read_excel function has to be used to specify the column.

```
# Inspect sheets of excel-file
excel_sheets('Spreadsheet_Data.xls')
```

```
## [1] "DoubleClick" "Copyright"   "Kayak"
```

```
# Import data
kayak <- read_excel("Spreadsheet_Data.xls",
                                                    sheet = "Kayak")

doubleclick <- read_excel("Spreadsheet_Data.xls",
                          sheet = "DoubleClick")
```

Massaging

```
#Convert to dataframe
doubleclick_clean <- as.data.frame(doubleclick)

#Look for weird stuff
colSums(is.na(doubleclick_clean))
```

```
##             Publisher ID           Publisher Name               Keyword ID
##                        0                        0                        0
##                  Keyword             Match Type                 Campaign
##                        0                        0                        0
##            Keyword Group                 Category             Bid Strategy
##                        0                        0                     1224
##             Keyword Type                   Status        Search Engine Bid
##                        0                        0                        0
##                   Clicks             Click Charges      Avg. Cost per Click
##                        0                        0                        0
##              Impressions      Engine Click Thru %                 Avg. Pos.
```

```
##                         0                         0                         0
##           Trans. Conv. %       Total Cost/ Trans.                    Amount
##                         0                         0                         0
##               Total Cost Total Volume of Bookings
##                         0                         0
```

```
table(doubleclick_clean$`Bid Strategy`)
```

```
##
##                    Pos 3-6              Position 1- 3
##                         45                        264
##        Position 1-2 Target  Position 1-4 Bid Strategy
##                        274                        111
##       Position 1 -2 Target  Position 2-5 Bid Strategy
##                         11                        333
## Position 5-10 Bid Strategy  Postiion 1-4 Bid Strategy
##                       2208                         40
```

```
# Replace NA entries  in bid strategy with Unassigned
doubleclick_clean$`Bid Strategy`[is.na(doubleclick_clean$`Bid Strategy`)] = "Unassigned"

# Notice how the number of rows gets reduced
print(nrow(doubleclick_clean))
```

```
## [1] 4510
```

```
# Look for Spelling mistakes
unique(doubleclick_clean $`Bid Strategy`)
```

```
## [1] "Unassigned"                 "Position 2-5 Bid Strategy"
## [3] "Position 1- 3"              "Position 1-2 Target"
## [5] "Position 5-10 Bid Strategy" "Position 1-4 Bid Strategy"
## [7] "Position 1 -2 Target"       "Postiion 1-4 Bid Strategy"
## [9] "Pos 3-6"
```

```
# Replace Typos
doubleclick_clean$`Bid Strategy` <- gsub("Postiion 1-4 Bid Strategy","Position 1-4 Bid Strategy",double

doubleclick_clean$`Bid Strategy` <- gsub("Position 1 -2 Target","Position 1-2 Target",doubleclick_clean
```

Descriptive

```
# Count of observations

# Create data set for analysis
sem <- doubleclick_clean[,c('Campaign','Keyword','Keyword Group','Publisher Name', 'Bid Strategy','Engi

# Get a big picture understanding of the data
summary(sem)
```

```
##    Campaign           Keyword          Keyword Group      Publisher Name
##  Length:4510        Length:4510        Length:4510        Length:4510
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Bid Strategy       Engine Click Thru %  Match Type        Trans. Conv. %
##  Length:4510        Min.   :  0.000     Length:4510        Min.   :  0.0000
##  Class :character   1st Qu.:  1.532     Class :character   1st Qu.:  0.0000
##  Mode  :character   Median :  4.106     Mode  :character   Median :  0.0000
##                     Mean   : 11.141                        Mean   :  0.5693
##                     3rd Qu.: 10.917                        3rd Qu.:  0.0000
##                     Max.   :200.000                        Max.   :900.0000
##  Total Cost/ Trans.  Impressions      Total Volume of Bookings
##  Min.   :   0.00    Min.   :      0   Min.   :  0.0000
##  1st Qu.:   0.00    1st Qu.:     28   1st Qu.:  0.0000
##  Median :   0.00    Median :    176   Median :  0.0000
##  Mean   :  27.61    Mean   :   9284   Mean   :  0.8734
##  3rd Qu.:   0.00    3rd Qu.:    844   3rd Qu.:  0.0000
##  Max.   :9597.17    Max.   :8342415   Max.   :439.0000
```

```
str(sem)
```

```
## 'data.frame':    4510 obs. of  11 variables:
##  $ Campaign                : chr  "Western Europe Destinations" "Geo Targeted DC" "Air France Brand &
##  $ Keyword                 : chr  "fly to florence" "low international airfare" "air discount france
##  $ Keyword Group           : chr  "Florence" "Low International DC" "France" "Air France" ...
##  $ Publisher Name          : chr  "Yahoo - US" "Yahoo - US" "MSN - Global" "Google - Global" ...
##  $ Bid Strategy            : chr  "Unassigned" "Unassigned" "Position 2-5 Bid Strategy" "Position 1-
##  $ Engine Click Thru %     : num  9.09 16.67 11.11 14.71 2.52 ...
##  $ Match Type              : chr  "Advanced" "Advanced" "Broad" "Exact" ...
##  $ Trans. Conv. %          : num  900 100 100 3.39 12.5 ...
##  $ Total Cost/ Trans.      : num  0.257 0.625 0.388 1.156 2.2 ...
##  $ Impressions             : num  11 6 9 401 318 722 13 547 448 129 ...
##  $ Total Volume of Bookings: num  9 1 1 2 1 2 1 2 1 1 ...
```

```
# Find out most frequently used bid strategy
table(sem$`Bid Strategy`)
```

```
##
##                    Pos 3-6              Position 1- 3
##                         45                        264
##        Position 1-2 Target  Position 1-4 Bid Strategy
##                        285                        151
##  Position 2-5 Bid Strategy Position 5-10 Bid Strategy
##                        333                       2208
##                 Unassigned
##                       1224
```

```
# Find out unique publishers
unique(sem$`Publisher Name`)
```

```
## [1] "Yahoo - US"        "MSN - Global"      "Google - Global"
## [4] "Overture - Global" "Google - US"       "Overture - US"
## [7] "MSN - US"
```

```
# Average out the clickthroughs per publisher
clickthrough_publisher <- aggregate(sem$`Engine Click Thru %`, by=list(sem$`Publisher Name`), FUN=mean)

# Visualize average clickthroughs per publisher
plot_ly(clickthrough_publisher, x = clickthrough_publisher$`Group.1`, y=~`x`,title = 'Average Clickthrou
        layout(title = 'Clickthrough per Publisher', plot_bgcolor = "#e5ecf6",xaxis = list(title = 'Publ
```

```
## No trace type specified:
##   Based on info supplied, a 'bar' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#bar

## Warning: 'bar' objects don't have these attributes: 'title'
## Valid attributes include:
## '_deprecated', 'alignmentgroup', 'base', 'basesrc', 'cliponaxis', 'constraintext', 'customdata', 'cus
```

```
# Sum up Transactions per publisher
transactions_publisher <- aggregate(sem$`Total Volume of Bookings`, by=list(sem$`Publisher Name`), FUN=s

# Visualize transactions per publisher
plot_ly(transactions_publisher, x = transactions_publisher$`Group.1`, y=~`x`,title = 'Bookings per publi
        layout(title = 'Bookings per publisher', plot_bgcolor = "#e5ecf6",xaxis = list(title = 'Publishe
```

```
## No trace type specified:
##   Based on info supplied, a 'bar' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#bar

## Warning: 'bar' objects don't have these attributes: 'title'
## Valid attributes include:
## '_deprecated', 'alignmentgroup', 'base', 'basesrc', 'cliponaxis', 'constraintext', 'customdata', 'cus
```

```
# What are the overall average costs / transaction
avg_costs_transaction <- print(mean(sem$`Total Cost/ Trans.`))
```

```
## [1] 27.60745
```

```
# Average out the costs per transaction per publisher
costs_publisher <- aggregate(sem$`Total Cost/ Trans.`, by=list(sem$`Publisher Name`), FUN=mean)

# Visualize average costs per transaction per engine
plot_ly(costs_publisher, x = costs_publisher$`Group.1`, y=~`x`)%>%
        layout(title = 'Average Costs per Publisher', plot_bgcolor = "#e5ecf6",xaxis = list(title = 'Pub
```

```
## No trace type specified:
##   Based on info supplied, a 'bar' trace seems appropriate.
##   Read more about this trace type -> https://plotly.com/r/reference/#bar
```

It seems like Google-US has the highest clickthrough rate and the costs / click are unusually high for Yahoo - US. One reason could be the advanced Match Type that gets Air France uses on that engine.

Yahoo-US has the highest percentage of click through rate with and impressive ~16%. What makes this output so impressive is that Yahoo-US has the second lowest cost per campaign with an average of $7.95, and Yahoo-US is still able to concure the top three Transactions per publishes with a total of 662.

```
# Total Cost per Transaction - Distribution per Publisher
plot_ly(sem,y = ~`Total Cost/ Trans.`, color = ~`Publisher Name`, type = "box")
```

```
# Visualize distribution of Bid Strategies for single Publishers
plot_ly(sem[which(sem$`Publisher Name`=='Google - US'),], x = ~`Publisher Name`, y = ~`Total Cost/ Trans
```

```
# Visualize impressions per campaign
plot_ly(doubleclick_clean, x = doubleclick_clean$`Campaign`, y=~Impressions, type='bar')
```

```
library('GGally')
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
# Select all the numerical variables
logic <- sapply(sem, is.numeric)
numerical_var <- sem[,logic]

# Select all the numerical variables
#logic <- sapply(doubleclick_clean, is.numeric)
#numerical_var <- doubleclick_clean[,logic]

numerical_var_standardized <- as.data.frame(scale(numerical_var))

p <- ggpairs(numerical_var_standardized, title="correlogram with ggpairs()")
ggplotly(p)
```

```
## Warning: Can only have one: highlight

## Warning: Can only have one: highlight

## Warning: Can only have one: highlight

## Warning: Can only have one: highlight
```

Most impressions come from unassigned keywords.

```
# Select observations with the highest total cost per transaction
sem_sub <- subset(sem,subset = `Total Cost/ Trans.` > 0)

# Visualize the costs per transactions for different Publisher
p <- plot_ly(sem_sub, y = ~`Total Cost/ Trans.`, color = I("black"),
             alpha = 0.2, boxpoints = "suspectedoutliers")
p1 <- p %>% add_boxplot(x = ~`Publisher Name`)
p1
```

```
# Visualize the converted transactions for different bid strategies
convert_bid <- plot_ly(sem_sub, y = ~`Trans. Conv. %`, color = I("black"),
              alpha = 0.2, boxpoints = "suspectedoutliers")
p2 <- p %>% add_boxplot(x = ~`Bid Strategy`)
p2
```

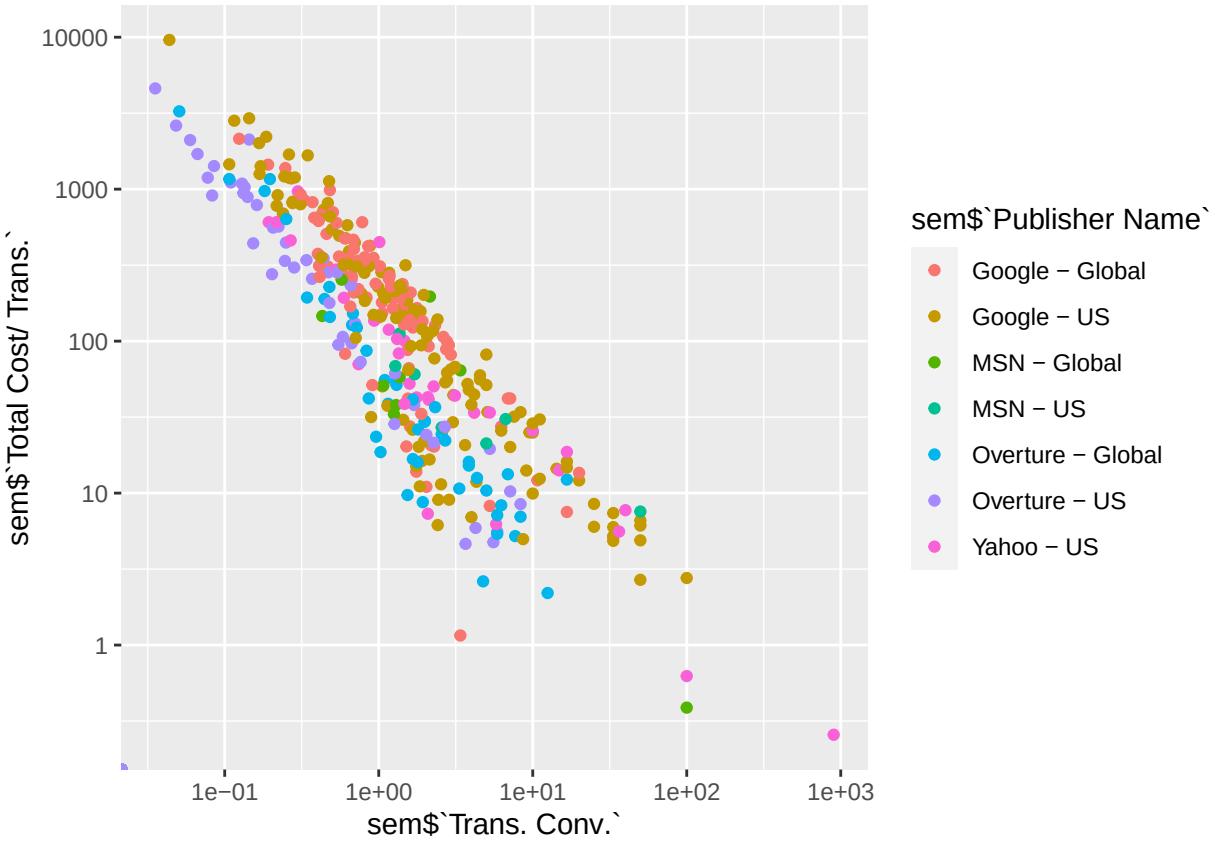
```
# Visualize the numerical variables in 3D-Space
plot_ly(sem, x = ~`Engine Click Thru %`, y = ~`Trans. Conv. %`, z =~`Total Cost/ Trans.`) %>%
  add_markers(color = ~`Trans. Conv. %`)
```

Keywords

```
ggplot(data=doubleclick_clean, aes(x=sem$`Trans. Conv.`, y=sem$`Total Cost/ Trans.`, color=sem$`Publishe
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```



Predictive

Feature Selection Model

Message

Key Findings

The C-suite of _____ face the following (problem/challenge), which is best solved with _ (solution) having an impact and/or making profits via _____ . The unique advantages/differentiators of the MVP are _____ , when comparing with the following key competitors / alternatives: _____

Next steps(What needs to be done!)

- Do branded keywords bring in more revenue?
- Are broad or focused keywords more profitable?
- Can assist keywords help increase conversion rate?