Return On Ad Spending [ROAS]

Raj Gopalan

Project done as part of UCLA MBA, Marketing Analytics course

An advertiser conducted a randomized experiment (A/B test) using cities as the unit of measure, half the cities got the ads and half didn't. Now the advertiser wants to measure the return on ad spend (ROAS). ROAS is calculated as:

ROAS = Total Incremental Sales / Total Advertising Spend

Example - An ROAS of 3 indicates that the advertiser made $2 in sales for every $1 in ad spend.

The A/B test lasted 30 days. Daily sales by city for 60 days prior to the test and for the 30 days during the test is provided. The ad spend for each day is provided as well (there was no advertising before the test started). 10,000 cities were randomly chosen and randomly assigned to a test group (5000) and control group (5000).

The question that needs to be answered: Does advertising impact sales? If so can you quantify this effect?

Disclaimer:

I can't provide the data files as per my professor. But, the intent is to analyze the files and bring out the insights from the data, which can help answer the question: what is the ROAS? So, let's go!

```
adspend <- read.csv("data/adSpend.csv", header = TRUE)
assignments <- read.csv("data/gmaAssignment.csv", header = TRUE)
sales <- read.csv("data/salesData.csv", header = TRUE)

# Use install_bitbucket('perossichi/DataAnalytics') to download the
# DataAnalytics library
library(DataAnalytics)
library(ggplot2)
library(reshape)

# Merge the three sources of data using gma as the key
roas.ads <- merge_recurse(list(adspend, assignments, sales))
sum(is.na(roas.ads))
```

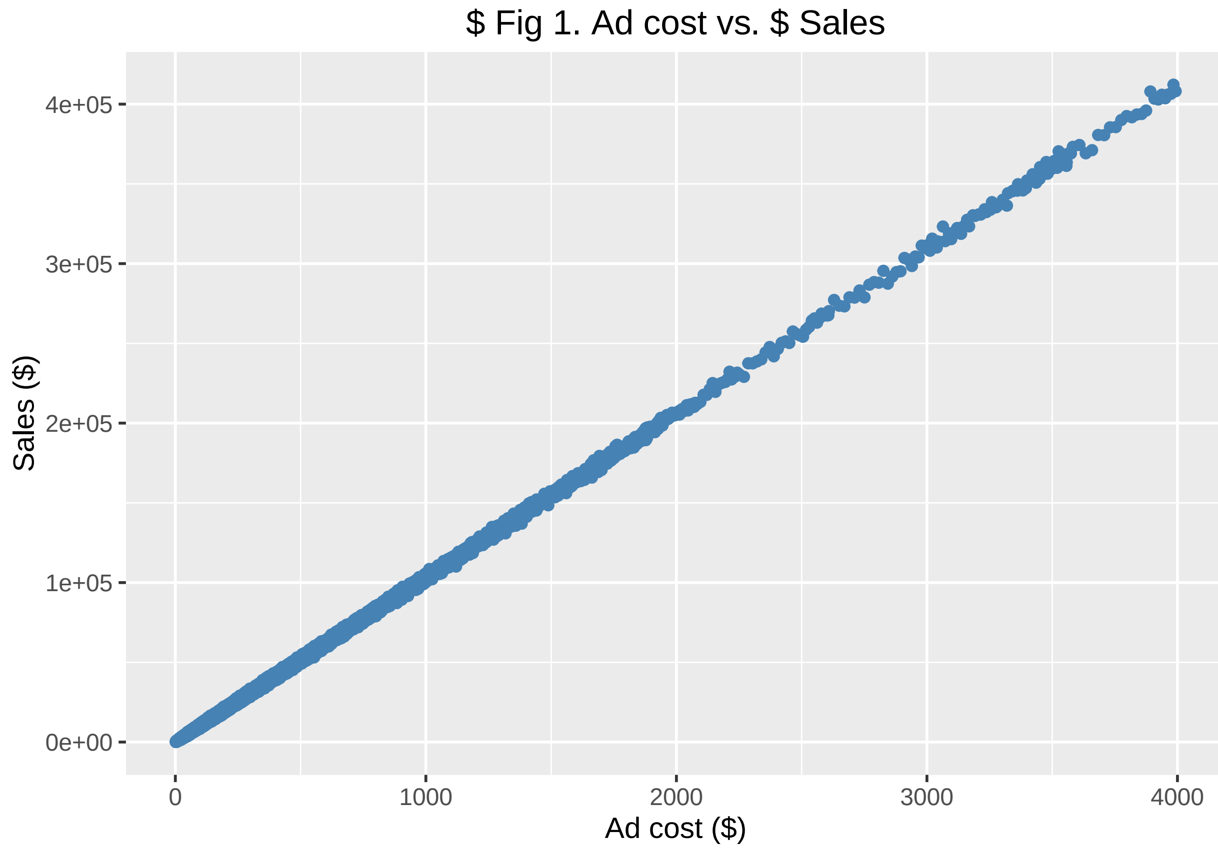
```
## [1] 0
```

Great, **no** NAs!

Let's visualize the data.

Let's start by doing a quick **scatter plot** of *Ad cost vs. Sales* to see if there is a correlation between the two. I will be using data which contains ad cost information, ignoring the data when the ads were not shown. The expectation is that we should see a positive correlation.

```
fig1 <- ggplot(roas.ads[roas.ads$ad.clicks > 0, ], aes(cost.clicks, sales))
fig1 + geom_point(color = "steel blue") + ggtitle("$ Fig 1. Ad cost vs. $ Sales") +
    labs(x = "Ad cost ($)", y = "Sales ($)")
```



The **positive correlation** between sales and ads is clearly visible in *Fig 1.* above, i.e. as the cost of ads increases, the sales seems to be increasing as well. Now that I'm seeing a correlation between sales and advertisement spend, the approach I would like to take, is to first break down the problem that needs to be solved.

I know the A/B test lasted 30 days and that the daily sales by city for 60 days prior and 30 days during the test is provided. This means that there are three months worth of data. A quick summary of the *date* column should validate this.

```
summary(roas.ads$date)
```

```
## 2014-01-01 2014-01-02 2014-01-03 2014-01-04 2014-01-05 2014-01-06
##      10000      10000      10000      10000      10000      10000
## 2014-01-07 2014-01-08 2014-01-09 2014-01-10 2014-01-11 2014-01-12
##      10000      10000      10000      10000      10000      10000
## 2014-01-13 2014-01-14 2014-01-15 2014-01-16 2014-01-17 2014-01-18
##      10000      10000      10000      10000      10000      10000
## 2014-01-19 2014-01-20 2014-01-21 2014-01-22 2014-01-23 2014-01-24
##      10000      10000      10000      10000      10000      10000
## 2014-01-25 2014-01-26 2014-01-27 2014-01-28 2014-01-29 2014-01-30
##      10000      10000      10000      10000      10000      10000
## 2014-01-31 2014-02-01 2014-02-02 2014-02-03 2014-02-04 2014-02-05
##      10000      10000      10000      10000      10000      10000
## 2014-02-06 2014-02-07 2014-02-08 2014-02-09 2014-02-10 2014-02-11
##      10000      10000      10000      10000      10000      10000
## 2014-02-12 2014-02-13 2014-02-14 2014-02-15 2014-02-16 2014-02-17
##      10000      10000      10000      10000      10000      10000
## 2014-02-18 2014-02-19 2014-02-20 2014-02-21 2014-02-22 2014-02-23
```

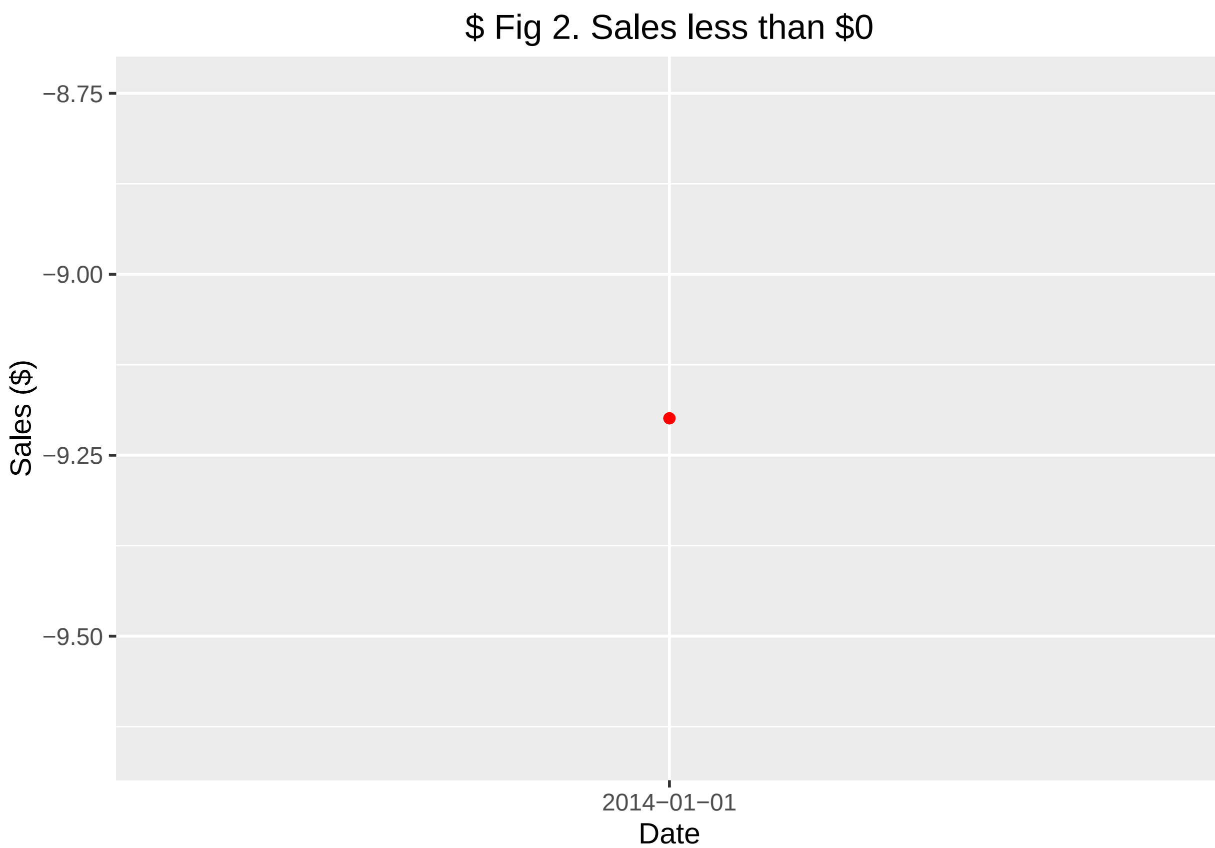
```
##      10000      10000      10000      10000      10000      10000
## 2014-02-24 2014-02-25 2014-02-26 2014-02-27 2014-02-28 2014-03-01
##      10000      10000      10000      10000      10000      10000
## 2014-03-02 2014-03-03 2014-03-04 2014-03-05 2014-03-06 2014-03-07
##      10000      10000      10000      10000      10000      10000
## 2014-03-08 2014-03-09 2014-03-10 2014-03-11 2014-03-12 2014-03-13
##      10000      10000      10000      10000      10000      10000
## 2014-03-14 2014-03-15 2014-03-16 2014-03-17 2014-03-18 2014-03-19
##      10000      10000      10000      10000      10000      10000
## 2014-03-20 2014-03-21 2014-03-22 2014-03-23 2014-03-24 2014-03-25
##      10000      10000      10000      10000      10000      10000
## 2014-03-26 2014-03-27 2014-03-28 2014-03-29 2014-03-30 2014-03-31
##      10000      10000      10000      10000      10000      10000
```

As expected, the summary identifies the start and end dates. There is information on January, February and March. I am converting the dates into months, with the intent of identifying seasonal effects, if any. I would like to ensure that the months are added as categorical variables so that running a regression becomes straighforward.

```r
roas.ads$January <- ifelse((months(as.Date(roas.ads$date, "%Y-%m-%d")) == "January"),
    1, 0)
roas.ads$February <- ifelse((months(as.Date(roas.ads$date, "%Y-%m-%d")) == "February"),
    1, 0)
roas.ads$March <- ifelse((months(as.Date(roas.ads$date, "%Y-%m-%d")) == "March"),
    1, 0)
```

I first need to ensure that all sales values are greater than 0. A quick plot of sales values less than 0 would help check this:

```r
fig2 <- ggplot(roas.ads[roas.ads$sales < 0, ], aes(date, sales))
fig2 + geom_point(color = "red") + ggtitle("$ Fig 2. Sales less than $0") +
    labs(x = "Date", y = "Sales ($)")
```



Observations: It looks like there is one sales value less than $0. A logical interpretation could be that it was a return that happened. So, for my analysis, I will exclude this row from the data set.

```
roas.validSales <- roas.ads[roas.ads$sales > 0, ]
```

I will set the *sales* as the dependent variable, y (a proxy for incremental sales), and *cost.clicks* as the independent variable, x. Then, the regression coefficient (β) obtained from running a regression of the sales on the log transform of cost.clicks, would ideally give ROAS, i.e. ROAS = β / 100. This is based on the interpretation of the results of running a regression of a dependent variable on the log transform of an independent variable.

Here is what the regression equation would looke like:
$y = \alpha + \beta * log(x) + \epsilon$
where β = ROAS * 100

I will create a column for the log transform of ad spend (cost.clicks) in the existing dataframe. Since there are values of *0* for cost.clicks, I will add one to cost.clicks before taking the log.

```
roas.validSales$logAdSpend = log(1 + roas.validSales$cost.clicks)
```

In order to ensure that the regression is run using relevant data, I will be using the data from March time-frame. I have test group and control group data, which helps remove sampling bias.

```
roas.validSales.March = roas.validSales[roas.validSales$March == 1,]
lm.1.Results <- lm(data = roas.validSales.March,
                   sales ~ logAdSpend
                   )
lmSumm(lm.1.Results)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 310000 observations
##
## lm(formula = sales ~ logAdSpend, data = roas.validSales.March)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)    10490     45.21  231.95       0
## logAdSpend      1627     14.29  113.90       0
## ---
## Standard Error of the Regression:  18490
## Multiple R-squared:  0.04  Adjusted R-squared:  0.04
## Overall F stat: 12972.39 on 1 and 309998 DF, pvalue= 0
```

Observations:

1. As expected, logAdSpend is statistically significant.
2. $1 increase in ad spend now lifts sales by 1627 / 100 = $16.27.

Hence, **ROAS = 16.27**. This means that the advertiser got **$16.27 in sales, for every $1 spent on advertising**!