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))
```

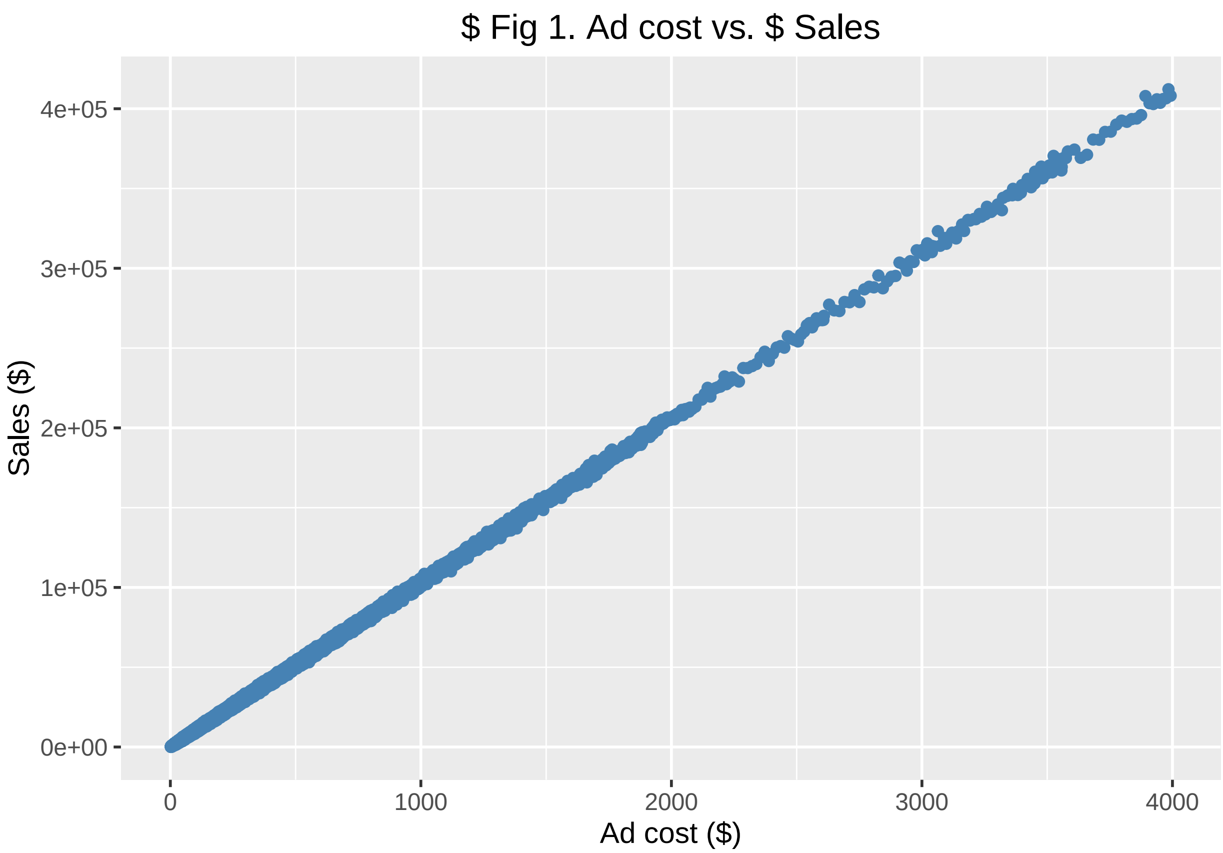
```
## [1] 0
```

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
```

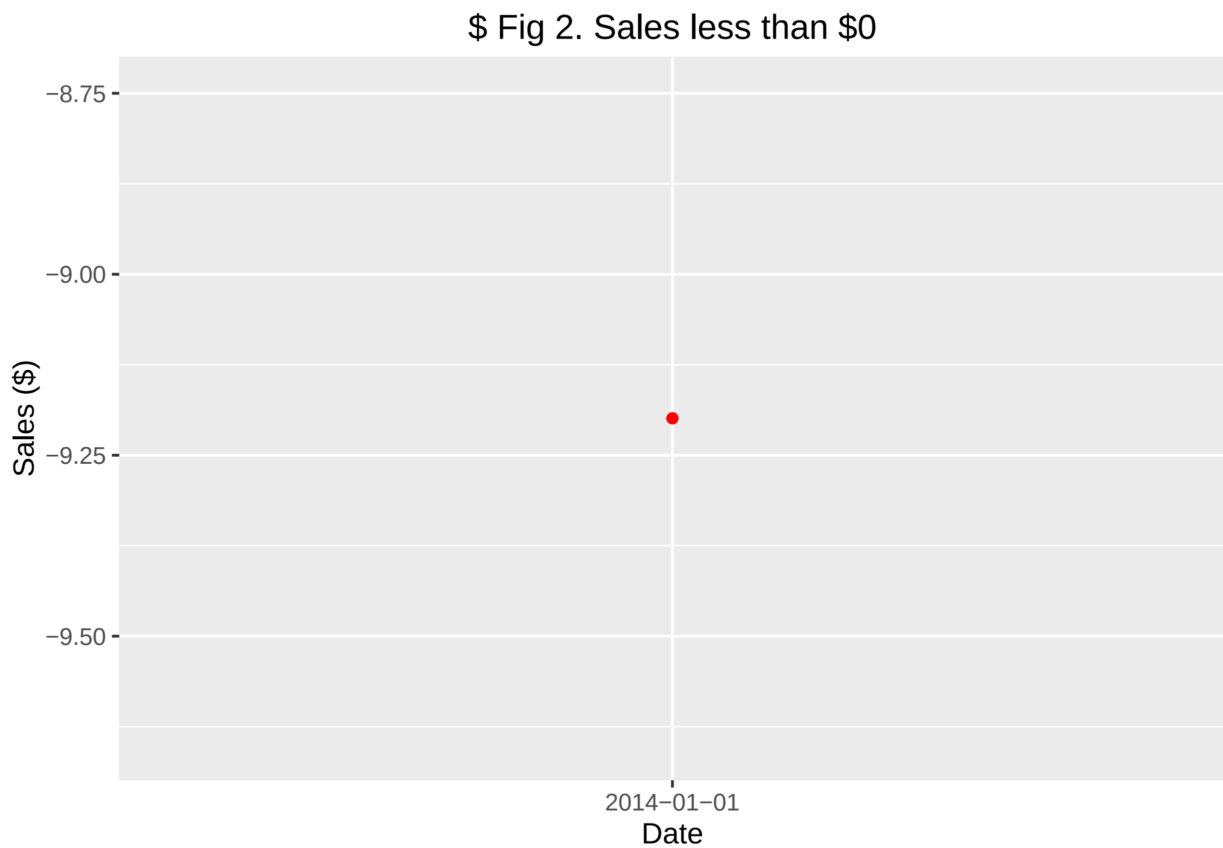
```
##      10000      10000      10000      10000      10000      10000
## 2014-02-24 2014-02-25 2014-02-26 2014-02-27 2014-02-28 2014-03-01
##      10000      10000      10000      10000      10000      10000
## 2014-03-02 2014-03-03 2014-03-04 2014-03-05 2014-03-06 2014-03-07
##      10000      10000      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.

```
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:

```
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, ]
```

The first intuition I would like to check, is if there are any time series trends affecting the sales. Running a *linear regression* of sales on February and March, could help identify monthly trends, if any. I am splitting the data based on the test and control groups, to check if the effects are the same across the groups. The expectation is that, since the test group vs control group assignment was random, the effects should be the same in both data sets.

```
roas.cg <- roas.validSales[roas.validSales$gma.group == 0, ]
roas.cg$gma.group = NULL
roas.tg <- roas.validSales[roas.validSales$gma.group == 1, ]
roas.tg$gma.group = NULL
lm.cg.Results <- lm(data = roas.cg, sales ~ February + March)
lmSumm(lm.cg.Results)
```

```
## Multiple Regression Analysis:
##     3 regressors(including intercept) and 450000 observations
##
## lm(formula = sales ~ February + March, data = roas.cg)
##
## Coefficients:
##               Estimate Std Error t value p value
## (Intercept)    10710     39.59  270.43       0
## February        3279     57.46   57.06       0
```

```
## March            2679     55.98   47.85       0
## ---
## Standard Error of the Regression:  15580
## Multiple R-squared:  0.008  Adjusted R-squared:  0.008
## Overall F stat: 1891.4 on 2 and 449997 DF, pvalue= 0
```

```
lm.tg.Results <- lm(data = roas.tg, sales ~ February + March)
lmSumm(lm.tg.Results)
```

```
## Multiple Regression Analysis:
##     3 regressors(including intercept) and 449999 observations
##
## lm(formula = sales ~ February + March, data = roas.tg)
##
## Coefficients:
##              Estimate Std Error t value p value
## (Intercept)    11330     50.18  225.84       0
## February        3471     72.84   47.65       0
## March           3246     70.96   45.74       0
## ---
## Standard Error of the Regression:  19760
## Multiple R-squared:  0.006  Adjusted R-squared:  0.006
## Overall F stat: 1468.15 on 2 and 449996 DF, pvalue= 0
```

Observations:

1. Both February and March are statistically significant at the 1% level.
2. In the control group, average March sales were \$3279 - \$2679 = **\$600** less than average February sales. Whereas, in the test group, average March sales were \$3471 - \$3246 = **\$225** less than average February sales.
3. In March, when the A/B testing was in progress, **the average sales in the test group were *higher* on than those of the control group by \$3246 - \$2679 = \$567**.
4. In both groups, average January sales were lower than in February or March.

Next hypothesis that I would like to test is, the day of the week made a difference in sales numbers. For this purpose, I am creating a set of categorical variables for the weekday corresponding to the date of each observation. I will then run a regression of the test group sales in March against days of the week, to see if there is a statistical significance.

```
roas.tg$Monday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) == "Monday"),
    1, 0)
roas.tg$Tuesday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) == "Tuesday"),
    1, 0)
roas.tg$Wednesday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) ==
    "Wednesday"), 1, 0)
roas.tg$Thursday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) == "Thursday"),
    1, 0)
roas.tg$Friday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) == "Friday"),
    1, 0)
roas.tg$Saturday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) == "Saturday"),
    1, 0)
roas.tg$Sunday <- ifelse((weekdays(as.Date(roas.tg$date, "%Y-%m-%d")) == "Sunday"),
```

```
    1, 0)

roas.tg.March <- roas.tg[roas.tg$March == 1, ]
lm.1.Results <- lm(data = roas.tg.March, sales ~ Monday + Tuesday + Wednesday +
    Thursday + Friday + Saturday)
lmSumm(lm.1.Results)
```

```
## Multiple Regression Analysis:
##     7 regressors(including intercept) and 155000 observations
##
## lm(formula = sales ~ Monday + Tuesday + Wednesday + Thursday +
##     Friday + Saturday, data = roas.tg.March)
##
## Coefficients:
##              Estimate Std Error t value p value
## (Intercept) 14570.00     133.4  109.24   0.000
## Monday        -79.91     188.6   -0.42   0.672
## Tuesday       168.30     200.0    0.84   0.400
## Wednesday      89.20     200.0    0.45   0.656
## Thursday       10.18     200.0    0.05   0.959
## Friday        -78.43     200.0   -0.39   0.695
## Saturday      -14.24     188.6   -0.08   0.940
## ---
## Standard Error of the Regression:  21090
## Multiple R-squared:  0  Adjusted R-squared:  0
## Overall F stat: 0.38 on 6 and 154993 DF, pvalue= 0.895
```

Observations:

Unfortunately, **none** of the days of the week are statistically significant. So, I will ignore days of the week in my model.

I had noted earlier that when the A/B testing was in progress in March, the sales in test group were higher on average than those of control group. Now I would like to run a linear regresion of the sales in **March** in the test group against the ad spending. The hypothesis is that, if ad spending improved sales, I should see a statistical significance of the ad spending.

```
lm.2.Results <- lm(data = roas.tg.March, sales ~ cost.clicks)
lmSumm(lm.2.Results)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 155000 observations
##
## lm(formula = sales ~ cost.clicks, data = roas.tg.March)
##
## Coefficients:
##              Estimate Std Error t value p value
## (Intercept)    710.0  14.59000   48.67       0
## cost.clicks    101.4   0.05967 1699.08       0
## ---
## Standard Error of the Regression:  4760
## Multiple R-squared:  0.949  Adjusted R-squared:  0.949
## Overall F stat: 2886867 on 1 and 154998 DF, pvalue= 0
```

Observations:

1. As expected cost.clicks are statistically significant
2. $1 increase in ad spend is indicating a lift in sales by **$100.4**. But, this $100.4 lift **includes the sales in the absence of an ad campaign**, which I will find next, by running a regression on the control group.

Now, to determine the component of sales in the absence of an ad acampaign, I am including the sales from the control group in March.

```
roas.March <- roas.validSales[roas.validSales$March == 1, ]
lm.3.Results <- lm(data = roas.March, sales ~ cost.clicks)
lmSumm(lm.3.Results)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 310000 observations
##
## lm(formula = sales ~ cost.clicks, data = roas.March)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)  8223.00   26.0600  315.52       0
## cost.clicks    84.18    0.1508  558.42       0
## ---
## Standard Error of the Regression:  13330
## Multiple R-squared:  0.501  Adjusted R-squared:  0.501
## Overall F stat: 311837.7 on 1 and 309998 DF, pvalue= 0
```

Observations:

1. As expected, cost.clicks is still statistically significant.
2. $1 increase in ad spend now lifts sales by $84.18.

So, the incremental lift in sales for the month of March, for $1 increase in ad spend = $100.4 - $84.18 = $16.22, all else remaining the same.

Hence, **ROAS = 16.22**. This means that the advertiser got **$16.22 in sales, for every $1 spent on advertising**!