Indeed leads homework solution

Deepak Bastola

December 4, 2020

Let's first import the data. After a quick look, there are NAs or mising values in the revenue variable. I replaced the missing values with 0 and scaled the revenue by a million dollars for easier readibility.

```
# Import data
mydata <- read.csv("/home/deepak/Downloads/homework_data_set.csv")
# Replace NAs with 0
mydata$revenue[is.na(mydata$revenue)]=0
# Scale the revenue by millions
mydata$revenue <- mydata$revenue/1000000
```

Question 1

How many leads are represented in this dataset? Describe both the assigned and unassigned populations. What is the average revenue of each group?

```
# Number of Leads
nleads <- nrow(mydata)
nleads
```

```
## [1] 77891
```

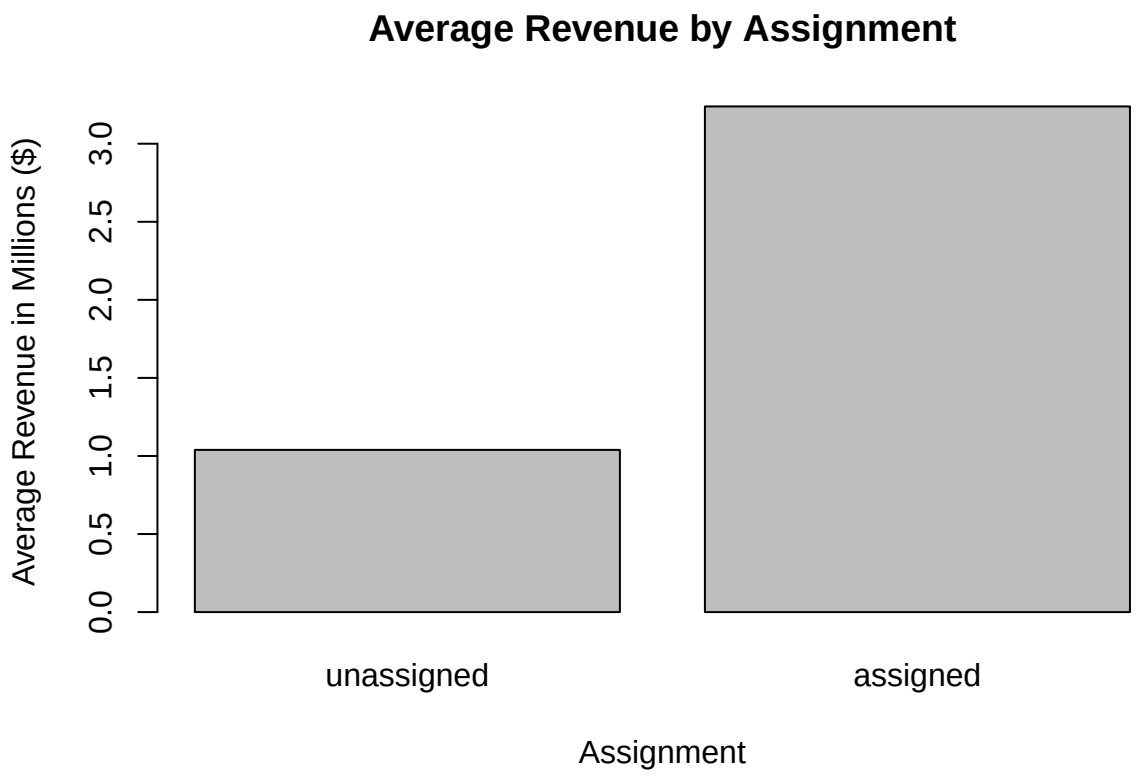
```
table(mydata$assigned)
```

```
##
##     0     1
## 40812 37079
```

```
# Average revenue by assigned
avg_revenue <- aggregate(revenue~assigned,
                         mydata,
                         FUN=mean,
                         na.rm = TRUE)[2][,1]
names(avg_revenue) <- c("unassigned", "assigned")
avg_revenue
```

```
## unassigned   assigned
##   1.039001   3.238846
```

```
# Vizualization of average revenue by assignment
barplot(avg_revenue, main = "Average Revenue by Assignment",
        xlab = "Assignment",
        ylab = "Average Revenue in Millions ($)")
```



There are 77891 leads in this dataset as represented by the number of cases or number of rows. Among these leads, 37079 were assigned and 40812 were not assigned. The average revenue under the assigned category is approximately 3.24 millions and that under unassigned category is 1.04 millions.

Question 2

What are the most important metrics to considering when answering the problem statement? Why?

```
names(mydata)
```

```
##  [1] "X"                    "advertiser_id"
##  [3] "assigned"             "date_assignment_starts"
##  [5] "date_assignment_ends" "first_revenue_date"
##  [7] "date_created"         "age"
##  [9] "assign_days"          "revenue"
```

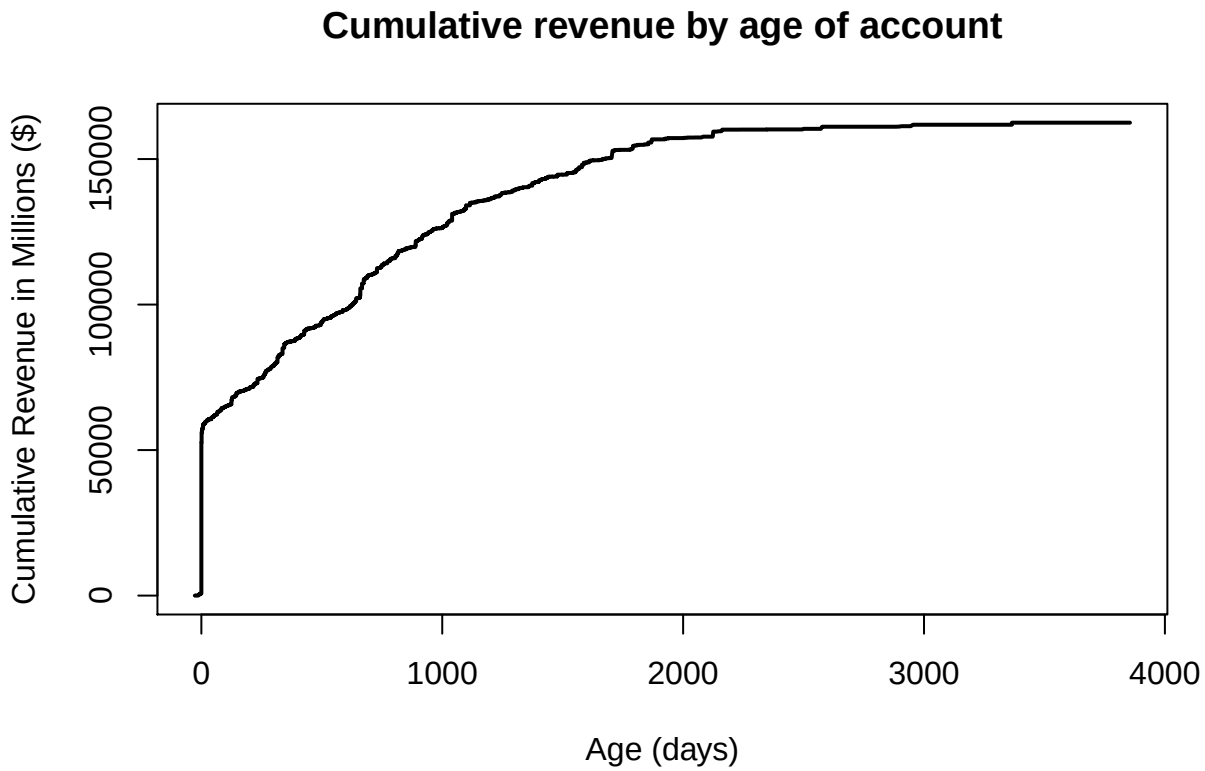
There are 9 variables in the dataset besides the lead number. We are interested in the revenue created by these leads. The good predictors of revenue are assigned and age.

Question 3

Analyze any existing relationship between account age and revenue.

```
# Order mydata with age
mydata_age_order <- mydata[order(mydata$age),]
```

```r
# Add a new column wih cumuative revenue
mydata_age_order$cum_revenue <- cumsum(mydata_age_order$revenue)
plot(mydata_age_order$age,
     mydata_age_order$cum_revenue,
     main = "Cumulative revenue by age of account",
     ylab = "Cumulative Revenue in Millions ($)",
     xlab = "Age (days)",
     type = "l",
     lwd = 2)
```



The cumulative revenue could be regarded as a good metric to look at when we want to track the revenue with age. In the data ordered with age, we can plot the cumulative revenue with age to see if there is some trend. From the figure, we see that a third of the revenue was created within a few days of lead creation. The revenue generation then slows down and plateaus at around 3000 days.

Question 4

What is the incremental value of assigning a lead to the sales team?

```r
# Increment using a linear model with useful covariates.
mod_final <- lm(revenue ~ assigned + age + assign_days,
                data = mydata)
summary(mod_final)
```

```
##
## Call:
## lm(formula = revenue ~ assigned + age + assign_days, data = mydata)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##   -9.0   -3.1   -1.9   -0.7 6530.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7009862  0.4643457   7.970 1.60e-15 ***
## assigned     1.1419559  0.3535763   3.230  0.00124 **
## age          0.0014336  0.0003605   3.976 7.01e-05 ***
## assign_days -0.0216027  0.0034419  -6.276 3.48e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.26 on 77887 degrees of freedom
## Multiple R-squared:  0.001485,   Adjusted R-squared:  0.001447
## F-statistic: 38.62 on 3 and 77887 DF,  p-value: < 2.2e-16
```

To see the incremental value of assigining a lead to the sales team, we can fit a linear model with revenue as the response, and assigned, age, assign_days as the covariates. This is the best final model among all the possible covariates. Although the R^2 value is quite low, the covariates are quite significant in this final reduced model.

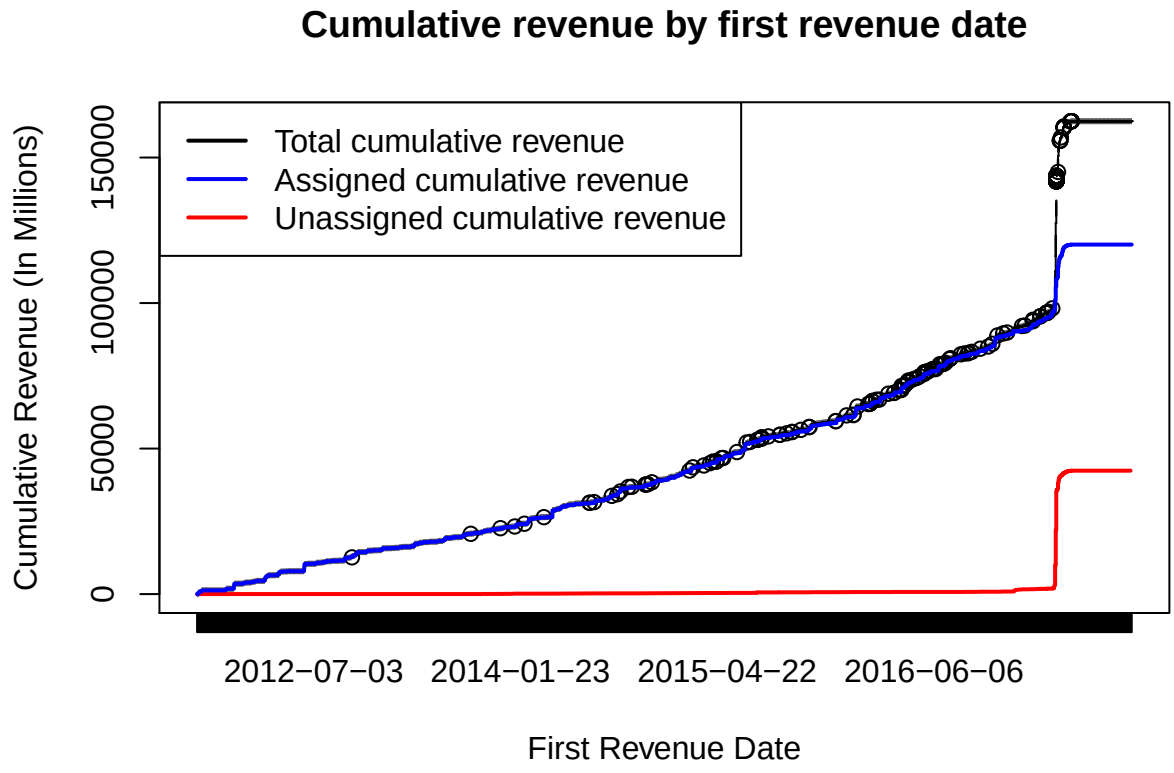
According to the model, for fixed age, and assign_days, the leads that are assigned create 1.14 millions more revenue on average than the leads that are unassigned. This is exactly the incremental value of assigning a lead to the sales team.

Question 5 (Bonus Question)

Investigate the data however you wish and discuss any interesting insights you can find in the data. Don't feel pressured to spend hours on this.

```r
# Order data by first revenue date, notice negative age values for certain leads.
mydata <- mydata[order(mydata$first_revenue_date),]
mydata$cum_revenue <- cumsum(mydata$revenue)

# Split the data into assigned and unassigned
mydata_assigned <- mydata_age_order[mydata_age_order$assigned==1,]
mydata_unassigned <- mydata_age_order[mydata_age_order$assigned==0,]

# Order data by first revenue date in unassigned category
mydata_unassigned <- mydata_unassigned[order(mydata_unassigned$first_revenue_date),]
mydata_unassigned$cum_revenue <- cumsum(mydata_unassigned$revenue)

# Order data by first revenue date in unassigned category
mydata_assigned <- mydata_assigned[order(mydata_assigned$first_revenue_date),]
mydata_assigned$cum_revenue <- cumsum(mydata_assigned$revenue)

plot(mydata$first_revenue_date,
     mydata$cum_revenue,
     main = "Cumulative revenue by first revenue date",
     xlab = "First Revenue Date",
     ylab = "Cumulative Revenue (In Millions)",
     type = "l",
     lty = 1)
lines(mydata_unassigned$first_revenue_date, mydata_unassigned$cum_revenue, col='red', lwd = 2)
lines(mydata_assigned$first_revenue_date, mydata_assigned$cum_revenue, col='blue', lwd = 2)
```

```
legend("topleft",
       c("Total cumulative revenue","Assigned cumulative revenue", "Unassigned cumulative revenue"),
       lty = c(1,1,1),
       lwd = c(2,2,2),
       col = c("black","blue","red"))
```



It is noticeable from the data that there are negative values of age. So, some leads were created after they were assigned and they generated revenues before being assigned. We can see the cumulative revenue of the leads with the first revenue date among the total, assigned, and unassigned category in the plot above. In the beginning of 2017 we see that there was a big jump in revenue from unassigned leads, more than that for assigned leads. This could be because the leads were sold before the assignment. The fact that even assigned leads generated more revenue than usual during this time-frame warants further investiagtion to find some confounding factor that is driving the sales.