Exploratory Data Analysis Example

2/26/2022

This data set shows sales leads that were assigned or not assigned a sales rep. The purpose of the data set is to estimate the incremental impact that a sales lead had on revenue. This is also a good example of ambiguous data, fuzzy column titles, and handling negative values. In addition, I'll discuss problems with drawing conclusions from observational data. As a first step, I'll start out exploring the data, which involves a combination of understanding the data, creating summary statistics tables, and producing basic plots.

Understanding the Data

First, I'll load in the data and take a look at the overall dataset and the type of each variable.

```
sales_leads <- read.csv("example_data_set.csv")
str(sales_leads)
```

```
## 'data.frame':    77891 obs. of  10 variables:
##  $ X                     : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ advertiser_id         : int  485 598 673 813 1132 1181 1183 1240 1339 1395 ...
##  $ assigned              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ date_assignment_starts: chr  "2017-02-01 00:00:00.000000" "2017-02-01 00:00:00.000000" "2017-02-0
##  $ date_assignment_ends  : chr  "2017-06-19 12:12:37.888680" "2017-06-19 12:12:37.888680" "2017-04-28
##  $ first_revenue_date    : chr  "" "" "" "" ...
##  $ date_created          : chr  "2006-07-14" "2006-08-02" "2006-08-17" "2006-09-12" ...
##  $ age                   : int  3855 3836 3821 3795 3744 3738 3737 3725 3705 3694 ...
##  $ assign_days           : int  138 138 86 138 138 138 138 138 138 138 ...
##  $ revenue               : num  NA NA NA NA NA NA NA NA NA NA ...
```

The overall data set has 77,891 observations and 10 variables. I offer brief comments about each variable below:

-*X*: Since this variable did not have a variable name, R assigned the generic variable name *X*. This appears to be an integer variable that is numbered consecutively starting at 0. Every number is a distinct number, so there are no duplicates. In the data cleaning phase, I'll rename this variable lead_id and start numbering at 1.

-*advertiser_id*: Another integer variable that likely matches a unique advertiser_id from another table. Without another table to match, there's not much to do with this variable. Every number is a distinct number, so there are no duplicates.

-*assigned*: An integer that is equal to 1 if a lead was assigned to a sales rep and equal to 0 if a lead was never assigned.0 and 1 are the only values contained in this variable.

-*date_assignment_starts*: A character variable that appears to be of the format year-month-day, hour:minute:second. Every observation is coded as either February, 1, 2, or 3, 2017 and 00:00:00.0, so it doesn't appear to contain much information. I would need to inquire to find out more about this variable, and what exactly it represents. I'll change this to a date variable of the format year-month-day, hour:minute:second that can be read by the computer.

-*date_assignment_ends*: Another character variable that has the same year-month-day, hour:minute:second format. Much like the previous variable, it's not exactly clear what this variable represents, so I'll also change this variable to the same format of year-month-day, hour:minute:second.

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
assigned	77891	0.48	0.5	0	0	1	1
age	77891	309.96	488.8	-27	0	523	3855
revenue	3340	48651766.73	173689100.35	12000	6398444.5	37049593.25	6533791000

-*first_revenue_date*: A third character variable, this one appears to be of the format year-month-day. There are a lot of missing observations and each date doesn't necessarily include a revenue number with it. I'll change the format of this variable to year-month-day.

-*date_created*: Also of the format year-month-day. Earliest date is 2006-07-14, latest date is 2017-02-28. It appears to be the date the account was opened. I'll change the format to year-month-day.

-*age*: An integer variable that has a low value of -27 and a high value of 3855. Based on the date_created variable, it appears to be the age of the account in days, although it seems odd to have a negative number of days. I'll figure out the best way to handle this variable in the data cleaning section.

-*assign_days*: Another integer value that appears to measure the number of days since the account was assigned a sales lead. However, there are a few negative observations and a lot of numbers around 136, 137, and 138.It's not obvious exactly what this variable measures.

-*revenue*: A numeric variable that appears to have a number of missing observations. I assume the variable is measured in U.S. dollars. It's not obvious if revenue is measured over the entire life of the account or just at a certain point in time.

Table 1 presents summary statistics for our variables of interest.

```
library(vtable)
sumtable(sales_leads, vars = c('assigned','age','revenue'), digits = 2)
```

Let's take a look at Table 1. Our main variable of interest *assigned* appears to have no missing values and only zeros and ones. The mean value tells us the number of sales leads that were assigned a sales rep, so approximately 48% of sales leads were assigned a sales rep. The variable *revenue* only has 3,340 observations, which means approximatley 96% of the observations are missing. As mentioned above, the *age* variable contains a number of negative numbers which seems odd to say that the number of days is negative.

Data Cleaning

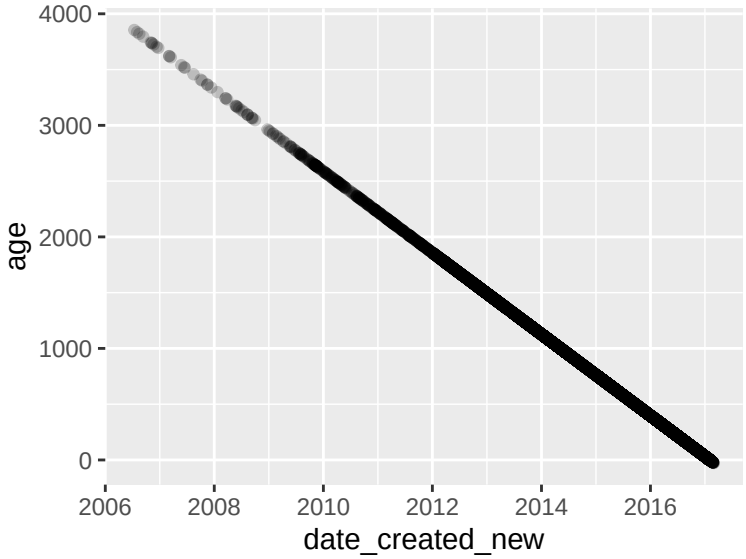
Next, I'll clean the data based on the issues identified above. First, I'll create a new variable, *lead_id*, which just starts counting the accounts at 1 as opposed to 0, like variable X. Next, I'll change the time variables so they can be read by the computer. Finally, I'll create a new variable that takes the difference between the start and end dates provided in the variables *date_assignment_starts* and *date_assignment_ends*.

```
library(lubridate)
library(magrittr)
library(tidyverse)
sales_dates_cleaned <- sales_leads %>%
  mutate(
    lead_id = X + 1,
    date_created_new = ymd(date_created),
    first_revenue_date_new = ymd(first_revenue_date),
    date_assignment_starts_new = ymd_hms(date_assignment_starts),
    date_assignment_ends_new = ymd_hms(date_assignment_ends),
    start_to_end = date_assignment_ends_new - date_assignment_starts_new
```

```
  )
str(sales_dates_cleaned)
```

Now I need to figure out what to do with the negative values in the variables *age* and *assign_days*. It appears that *age* and *date_created* are capturing similar information. Let's investigate with a scatter plot of the two variables.

```
library(ggplot2)
ggplot(data = sales_dates_cleaned) +
  aes(x = date_created_new, y = age) +
  geom_point(alpha = 0.2)
```



It appears the variables *date_created* and *age* are very closely related, which means that the negative values may not necessarily be a mistake. My initial thought was to make all of the negative values zero or assume that someone had mistakenly entered a negative sign, so making all of the negative numbers positive would solve the issue. However, that appears to not be the case, since the negative values continue on the straight line. For some reason the dates 2/1/17 through 2/3/17 are coded as 0, and the date 2/4/17 is coded as -3, with 2/5/17 coded as -4, etc. Dates starting at 1/31/17 are coded as 1 and count forward from there. There are a number of observations that have the same dates but are coded differently, so the relationship is not a perfect linear relationship, but it's obviously pretty close. It's not obvious how to modify this variable; one option would be to add 27 to each observation, but it appears it was a choice to enter negative numbers, so I would enquirer about the scale of this data before making any changes.

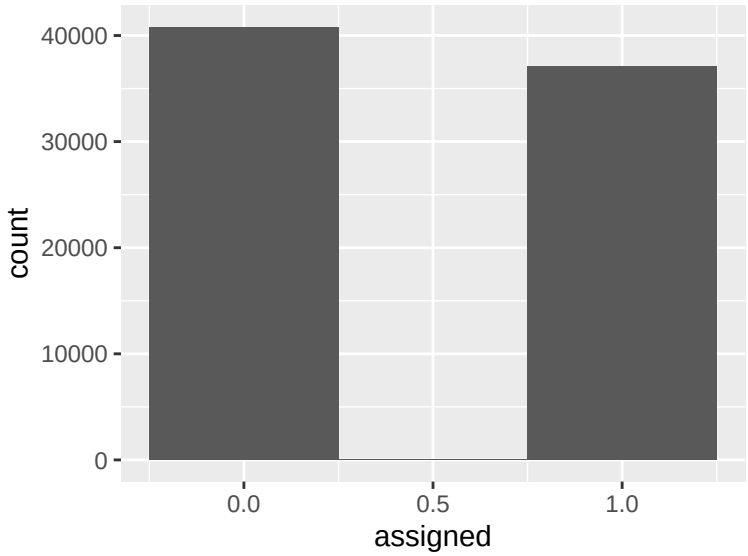
The *assign_days* variable appears to share no discernible relationship to the *date_created* or the *age* variable. As mentioned above there are a few negative numbers which do not appear to make since. The number -2 corresponds to only two of the observations from 2/1/17 and the number -1 corresponds to some of the observations on 1/31/17. It's not obvious if this is a mistake or what the mistake could be. The highest number is 138, and most of the observations are coded either 136, 137, or 138. These three numbers make up approximately 83% of the observations, with 138 accounting for approximately 42% of the observations.

Basic Plots

The final step to understanding the data is to show some basic plots, which I mostly use to identify possible outliers and investigate possible relationships between variables. Let's take a look at a basic histogram of our main variable of interest, *assigned*.

```
assigned_hist <- ggplot(data = sales_dates_cleaned) +
  geom_histogram(mapping = aes(x = assigned), binwidth = 0.5)
```

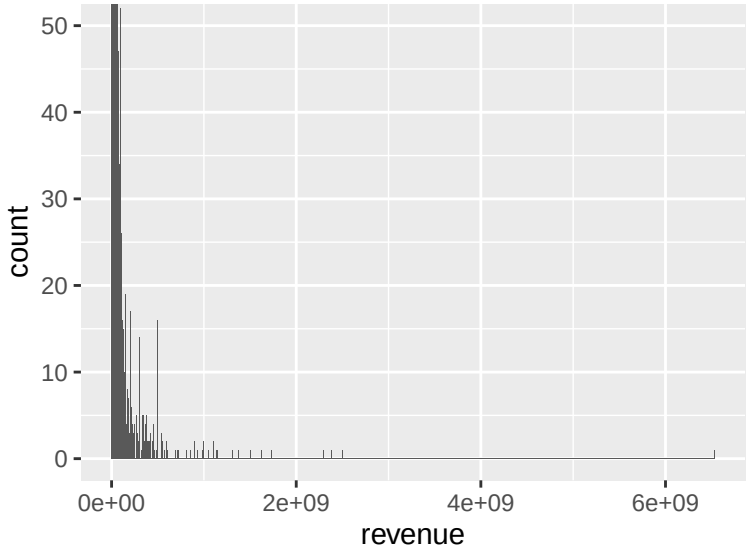
```
assigned_hist
```



This graph matches perfectly with our observation that approximately 48% of the sales leads were assigned to a sales rep, since slightly more than half of the observations equal zero. There also do not appear to be any outliers or miscoded observations.

Next, let's check out a basic histogram of the variable *revenue*.

```
rev_hist <- ggplot(data = sales_leads) +
  geom_histogram(mapping = aes(x = revenue), binwidth = 10000000) +
  coord_cartesian(ylim = c(0, 50))
rev_hist
```



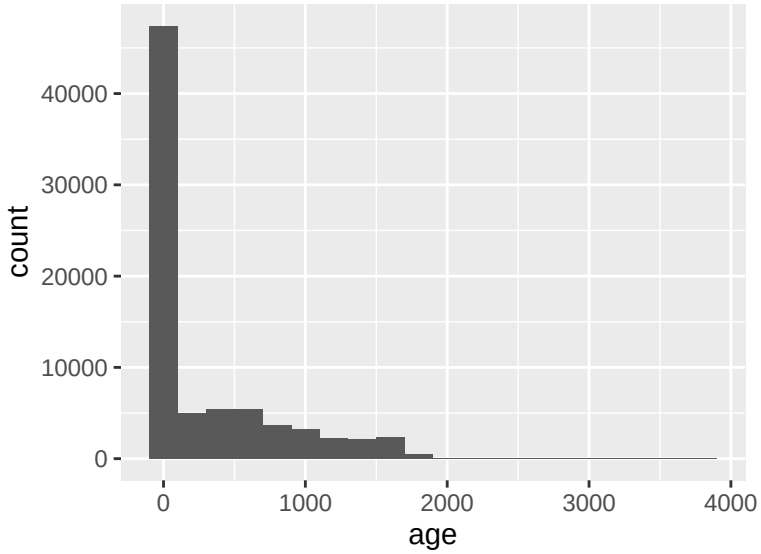
The distribution for the variable *revenue* is heavily skewed to the right, which is common for distributions that are limited to positive numbers. One option is to take the natural logartihm of *revenue* to reduce the skew of the distribution.

One observation with a value of 6,533,791,000 is more than 2.5 times as large as any other observation. I'm doubtful that this number is correct since the account was only created on 2/3/17 and the last date in the data set is 2/28/17. This seems like an unbelievably high number for an account that has been opened for less

than a month. This seems to be the only obvious outlier, but instead of completely removing the observation, I'll perform the analysis with and without the outlier and see if it's inclusion makes a difference.

Since *age* was shown to be a good measure of time and it's possible that *age* could be a confounding variable—since it's likely related to *revenue* and *assign*—let's take a look at the histogrom for *age*.

```
age_hist <- ggplot(data = sales_leads) +
  geom_histogram(mapping = aes(x = age), binwidth = 200)
age_hist
```



The distribution for *age* is also skewed to the right, with the largest percentage of the observations being coded as 0. From our previous analysis, this corresponds to the dates 2/1/17 to 2/3/17. It appears that the largest number of accounts were started on these dates. Most of the accounts are less than 2000 days old, with the oldest account being 3855 days old, which corresponds to the first account date of 7/14/06. Since these numbers represent time, this does not appear to be an outlier, since it is the oldest account.

That concludes our initial data exploration. Now on to the questions!

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?

The number of leads in the dataset should be equal to the number of observations, which is 77,891. Below I split the dataset into assigned and unassigned groups and then find the average revenue of each group, as well as, the minimum and maximum revenue for each group. I also investigate the relationship between *assigned* and the possible confounding variable, *age*.

```
sales_grouped <- sales_dates_cleaned %>%
  group_by(assigned) %>%
  summarize(
    N = n(),
    rev_min = min(revenue, na.rm = TRUE),
    rev_max = max(revenue, na.rm = TRUE),
    rev_average = mean(revenue, na.rm = TRUE),
    age_min = min(age),
    age_max = max(age),
    age_average = mean(age),
  )
sales_grouped
```

```
## # A tibble: 2 x 8
##   assigned     N rev_min    rev_max rev_average age_min age_max age_average
##      <int> <int>   <dbl>      <dbl>       <dbl>   <int>   <int>       <dbl>
## 1        0 40812   13000 6533791000   23889416.       0    3097        11.9
## 2        1 37079   12000 2500000000   76736861.     -27    3855       638.
```

As I noted earlier, the unassigned population is slightly larger than the assigned population with 40,812 observations in the unassigned population and 37,079 observations in the assigned population.

The average revenue for the assigned group is much larger than the average revenue for the unassigned group—76,736,861 for the assigned group compared to 23,889,416 for the unassigned group. The revenue minimum is similar between the two groups, and the unassigned group contains the maximum value that is more than 2.5 times larger than any other revenue observation. For reasons mentioned previously, this number is a potential outlier.

A couple of comments about the relationship between *assigned* and *age*. All of the ages with negative numbers, meaning all of the recently opened accounts are in the assigned population, as well as, the accounts with the highest numbers, meaning the oldest accounts. In fact, the sixty-eight newest and thirty-two oldest accounts are all in the assigned population. The average age of the accounts in the assigned population is also much larger than the average age of the accounts in the unassigned population. This large difference in average age is a potential warning sign for trying to draw conclusions about the relationship between *assigned* and *revenue*, since it appears there's a relationship between *assigned* and *age*. We also need to be concerned about the possibility of selection bias in the *assigned* variable.

Question 2

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

We are trying to estimate the incremental impact the sales representatives had on revenue. While this appears to be a simple question to answer on the surface, it can actually be much harder to answer because of confounding variables. The naive estimate would simply take the difference in the averages from question 1—76,736,861 - 23,889,416 = 52,847,445. However, there are likely lots of factors that affect *revenue*, if these factors are also related to *assign*, then our naive estimate of 52,847,445 could be incorrect.

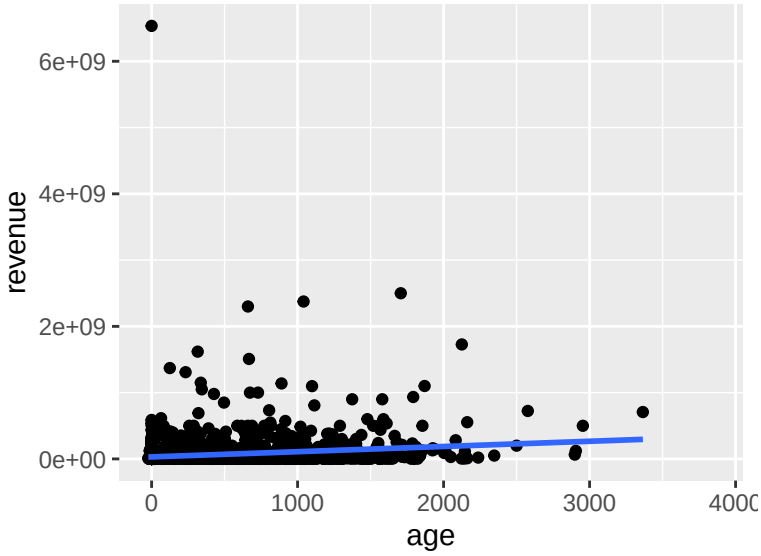
We've already started to investigate the relationship between *assign* and *age*, which we'll continue in question 3, but there could also be other factors that fit the same description—in other words, factors that affect *revenue* that are also related to *assigned*. The problem is we may not have data for all of the possible confounding variables or they may even be unobservable factors. In economics terminology, we call these factors omitted variables and the bias that these factors cause, omitted variable bias. Another possible issue is the problem of selection bias. Selection bias can be an issue when our main variable of interest, *assigned*, in this case, is not randomly determined. Possible solutions will be further discussed in the summary at the end.

Question 3

Analyze any existing relationship between account age and revenue.

To analyze the relationship, we'll look at basic scatter plots and linear regressions. I decided on a linear model for the ease of interpretability and because the issues of omitted variable bias and selection bias are not going to be solved by complicated modeling techniques. If we were purely using our model for prediction, I would entertain other models, but since we are interested in a causual question, a linear model gives us easy interpretations. Let's take a look at a basic scatter plot.

```
revenue_scatter <- ggplot(data = sales_dates_cleaned) +
  (aes(x = age, y = revenue)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
revenue_scatter
```



With the regression line added, there appears to be a positive relationship between *age* and *revenue*. Without the regression line, it would not be quite as obvious. What is obvious is how much larger the possible *revenue* outlier is compared to other revenue values. Let's run a basic linear regression with and without this point to see how it affects the result. First, the regression including the possible outlier.

```
linear_reg1 <- lm(revenue ~ age, sales_dates_cleaned)
summary(linear_reg1)
```

```
##
## Call:
## lm(formula = revenue ~ age, data = sales_dates_cleaned)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -193772935  -28443356  -19588896   -2113872 6503551144
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30239856    3338073   9.059   <2e-16 ***
## age            78672       6709  11.727   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 170200000 on 3338 degrees of freedom
##   (74551 observations deleted due to missingness)
## Multiple R-squared:  0.03957,    Adjusted R-squared:  0.03928
## F-statistic: 137.5 on 1 and 3338 DF,  p-value: < 2.2e-16
```

As we saw in the graph, the relationship is positive, and the regression results show us it's also highly statistically significant. A one day increase in *age* is associated with an increase in *revenue* of $78,672—assuming revenue is measured in dollars.We should also note that 74,551 observations out of a possible 77,891, were dropped because they are missing from the *revenue* variable. That leaves us with only 3,340 observations. We'll come back to this issue in question 5. For now, let's see how removing the outlier affects the relationship.

```
linear_reg2 <- lm(revenue ~ age, sales_dates_cleaned, subset = revenue < 6533791000)
summary(linear_reg2)
```

```
##
## Call:
## lm(formula = revenue ~ age, data = sales_dates_cleaned, subset = revenue <
##     6533791000)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -197479421  -26101317  -17199552     163448 2333931521
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27738552    2504722   11.07   <2e-16 ***
## age            81037       5033   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 127700000 on 3337 degrees of freedom
##   (74551 observations deleted due to missingness)
## Multiple R-squared:  0.07208,    Adjusted R-squared:  0.0718
## F-statistic: 259.2 on 1 and 3337 DF,  p-value: < 2.2e-16
```

The slope coefficient is larger, which is what we would expect based on the graph. However, the coefficient is still approximately 80,000. Averaged over 3,340 observations, the possible outlier doesn't have much of an effect on the results. Overall, there does appear to be a strong positive relationship between *age* and *revenue*. This is the expected result since accounts that have been opened for longer would be expected to produce more revenue.

Question 4

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

Since we established a relationship between *assigned* and *age* in question 1 and a relationship between *revenue* and *age* in question 3, including *age* as a control variable in a regression between *revenue* and *assigned*, will lead us to a better estimate of the incremental value of assigning a lead to the sales team.

```
linear_reg3 <- lm(revenue ~ assigned + age, sales_dates_cleaned)
summary(linear_reg3)
```

```
##
## Call:
## lm(formula = revenue ~ assigned + age, data = sales_dates_cleaned)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -178007358  -37876796  -15732064     550625 6511416436
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22374564    4038360   5.541 3.25e-08 ***
## assigned    23645380    6856636   3.449 0.000571 ***
## age            64939       7792   8.334  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7e+08 on 3337 degrees of freedom
##   (74551 observations deleted due to missingness)
## Multiple R-squared:  0.04298,    Adjusted R-squared:  0.0424
## F-statistic: 74.93 on 2 and 3337 DF,  p-value: < 2.2e-16
```

From question 1, without controlling for age, the difference in the assigned and un-assigned groups is 52,847,445. Once we control for age, our estimate for the incremental value of assigning a lead to the sales team decreases to 23,645,380 on average. We can see that our estimate is more than cut in half. We could imagine that controlling for additional factors would likely continue to decrease our estimate. This is why answering causal questions from observational data is so difficult. Although the results are not shown, removing the outlier in *revenue* increases the effect to 27,172,204 on average. Both answers show a large, positive, statistically significant effect of the incremental value of assigning a lead to the sales team. The questions we are left with are whether to trust the results and what could be done to increase our confidence in the results. I'll address both of these questions in the summary at the end.

An alternative estimation strategy takes into account the relationship between *assign* and *age* seen in question 1. We found the average age is equal to 638, when *assign* equals 1, while the average age is only 11.9, when *assign* equals 0. The below estimation, which includes an interaction term—in other words, a new variable *assign* times *age*—allows the effect of *assign* to depend on the value of *age*.

```
linear_reg4 <- lm(revenue ~ assigned + age + assigned:age, sales_dates_cleaned)
summary(linear_reg4)
```

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

```
##        Min         1Q     Median         3Q        Max
## -187423029  -33815946  -16427107    -469857 6509863893
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23927107    4082577   5.861 5.06e-09 ***
## assigned     19393589    7058692   2.747  0.00604 **
## age             -1616      27708  -0.058  0.95350
## assigned:age    72261      28871   2.503  0.01237 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 169800000 on 3336 degrees of freedom
##   (74551 observations deleted due to missingness)
## Multiple R-squared:  0.04477,    Adjusted R-squared:  0.04391
## F-statistic: 52.12 on 3 and 3336 DF,  p-value: < 2.2e-16
```

The resulting equation is shown below:

```
library(equatiomatic)
extract_eq(linear_reg4, use_coefs = TRUE)
```

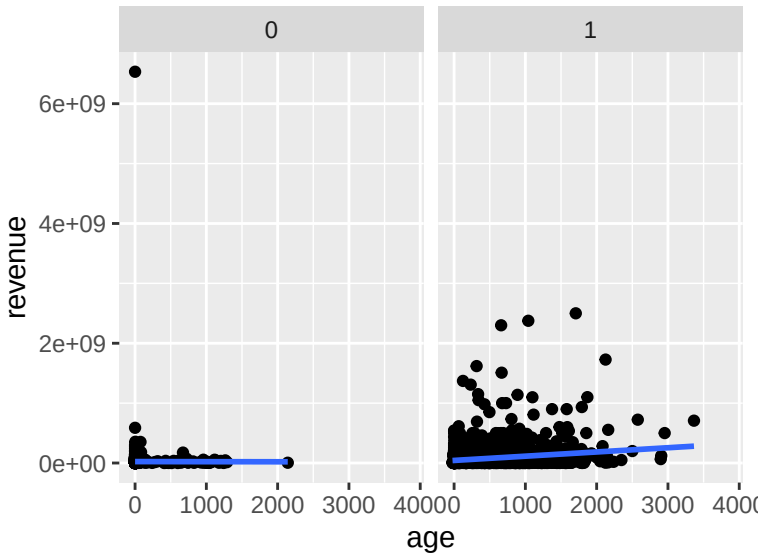
$$\widehat{\text{revenue}} = 23927107.23 + 19393588.69(\text{assigned}) - 1615.75(\text{age}) + 72261.22(\text{assigned} \times \text{age}) \quad (1)$$

According to this equation the effect of *assigned* equals:

$$\frac{\partial \widehat{revenue}}{\partial assigned} = 19393589 + 72261 * age,$$

According to this model, the biggest differences between assigned and non-assigned accounts comes when the age of the accounts is large. For example, plug-in *age* equals 0, and the difference is 19,393,589, but plug-in *age* equals 10 and the difference is now 193,935,890. Plug-in *age* equals 100 and add another zero to the difference or *age* equals 1000 for a difference of approximately 19 billion dollars. Let's see if we can picture what is happening graphically:

```
revenue_scatter <- ggplot(data = sales_dates_cleaned) +
  (aes(x = age, y = revenue)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ assigned)
revenue_scatter
```



The largest effects of *assigned* are obviously coming when the value of *age* is also large. The question that remains unanswered is whether the accounts that have a sales lead assigned to them are causing accounts to remain open longer thus leading to larger values of *age* or are accounts assigned to accounts that have been opened longer, meaning accounts are not randomly assigned leading to selection bias. I would guess the latter, which means selection bias could be a problem. I'll briefly address this issue in the summary at the end.

Question 5 (Bonus Question)

Investigate the data however you wish and discuss any interesting insights you can find in the data.

The number of missing observations in *revenue* is a potential problem that we identified during our data exploration. Let's identify the missing data and see if the summary statistics differ between the *age* and *assigned* variables. First, we'll create a variable that is TRUE if *revenue* is NA and FALSE otherwise. Then, we'll group the data into two groups and calculate summary statistics.

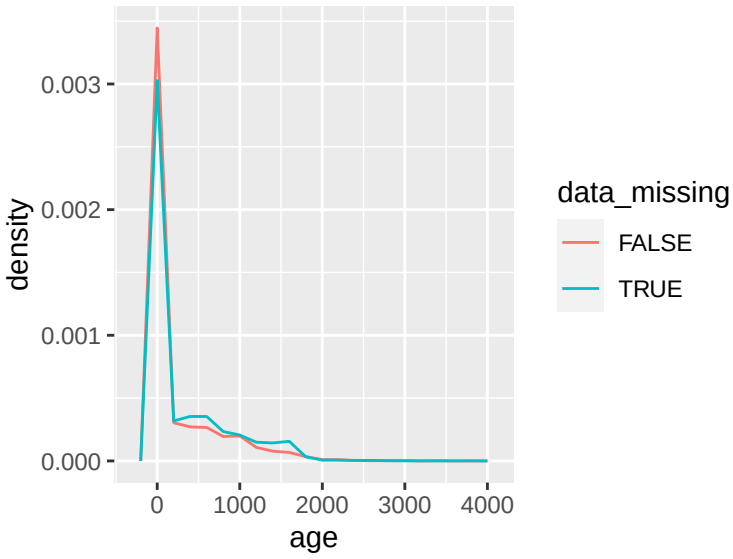
```
sales_leads_missing <- sales_leads %>%
  mutate(
    data_missing = is.na(revenue)
  )

sales_missing_data <- sales_leads_missing %>%
  group_by(data_missing) %>%
  summarize(
    N = n(),
    age_min = min(age),
    age_max = max(age),
    age_average = mean(age),
    assigned_min = min(assigned, na.rm = TRUE),
    assigned_max = max(assigned, na.rm = TRUE),
    assigned_average = mean(assigned, na.rm = TRUE)
  )
sales_missing_data
```

```
## # A tibble: 2 x 8
##   data_missing     N age_min age_max age_average assigned_min assigned_max
##   <lgl>        <int>   <int>   <int>       <dbl>        <int>        <int>
## 1 FALSE         3340     -22    3365        234.            0            1
## 2 TRUE         74551     -27    3855        313.            0            1
## # ... with 1 more variable: assigned_average <dbl>
```

As discussed in the date exploration section, 74,551 out of a possible 77,891 observations are missing. We would expect the minimum *assigned* value in both groups to be 0 and the maximum *assigned* value to be 1 for both groups. The average *assigned* value is also close, 0.469 compared to 0.476. This means the *assigned* value looks pretty similar across missing and non-missing data. The minimum and maximum value for *age* is fairly similar between the missing group and the non-missing group. The average *age* is higher among the missing group compared to the average *age* for the non-missing group. We'll investigate the shape of the distribution further with PDF graphs below:

```
age_poly <- ggplot(data = sales_leads_missing, mapping = aes(x = age, y = ..density..)) +
  geom_freqpoly(mapping = aes(color = data_missing), binwidth = 200)
age_poly
```



The PDFs between the two groups match up pretty closely. The missing data appears to contain more ages in the higher range and less zeros, which is what is causing the mean to be equal to 313 as opposed to 234 for the non-missing data, but overall the PDFs look very similar. This means that the missing data looks pretty close to our non-missing data across both *assign* and *age*. Obviously, we would prefer to have more than approximately 4% of our *revenue* data, but it's likely not worth the additional time and cost to fill in the missing data with our observed data. Especially since the PDFs match up well, our conclusions from imputation methods would likely produce similar conclusions to our current analysis.

Summary

Based on the current data and observed variables, the effect of the incremental value of assigning a lead to the sales team is large, statistically significant and equal to approximately $24,000,000 on average. However, there are a couple of questions that remain.

1. Should we trust the results?

There is no doubt that the average revenue from the assigned group is higher than the average revenue from the non-assigned group. However, this is just correlational evidence and we are interested in the causal relationship. Unfortunately, causal questions can be tough to answer with observational data. Ideally the two groups would be determined randomly. Our analysis shows that with regards to time captured by the *age* variable, *assigned* was not randomly assigned between the two groups. This could also be the case between *assigned* and other variables, as well. Without randomness in the *assigned* variable, we should be cautious with the results. It appears older accounts were more likely to be assigned a lead to the sales team, and it's hard to disentangle the effects of these two variables.

2. What can be done to increase our confidence in the results?

The gold standard in observational studies for determining causality is randomization. Ideally, the *assigned* variable would be randomly assigned, so that the resulting differences in *revenue* definitively come from *assigned*. Without this randomization, we need to be worried about the problems of omitted variable bias and selection bias. In many contexts, randomization may be unrealistic or too costly to implement. In this case, a more realistic strategy may be to collect *revenue* data at different points in time before and after the assignment of the *assigned* group. This type of data can be analyzed with a difference-in-differences modeling technique. The idea is that the only change in the *revenue* variable comes from this assignment of the *assigned* variable, as opposed to other factors that affect both *assigned* and *revenue*. The name comes from the double difference between the treatment and control groups, as well as, the difference between before and after. This additional data collection and modeling technique would lead to greater confidence in our results.