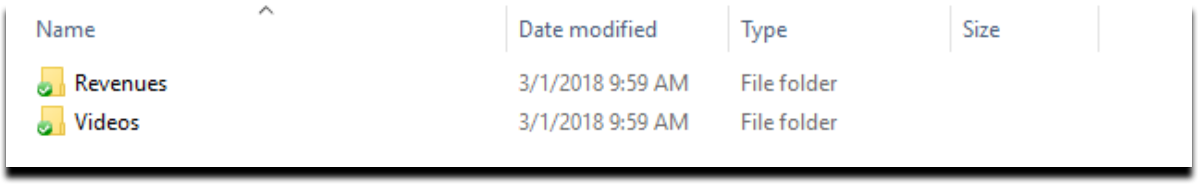
Project 3 – Due November 16 at 5pm.

This is an exercise in merging disparate data sources into a useful data set, and using the merged data set to perform some useful analysis.

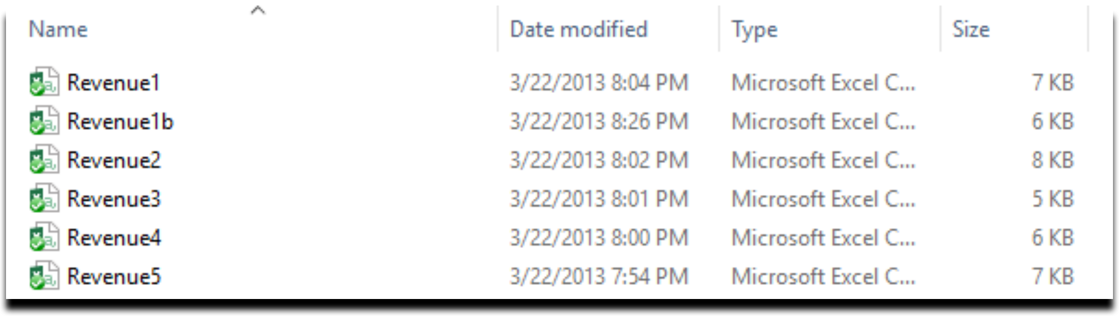
Part 1. Assembling the data.

The zipped folder YouTubeAttention2013.zip contains two subfolders: Revenues and Videos.



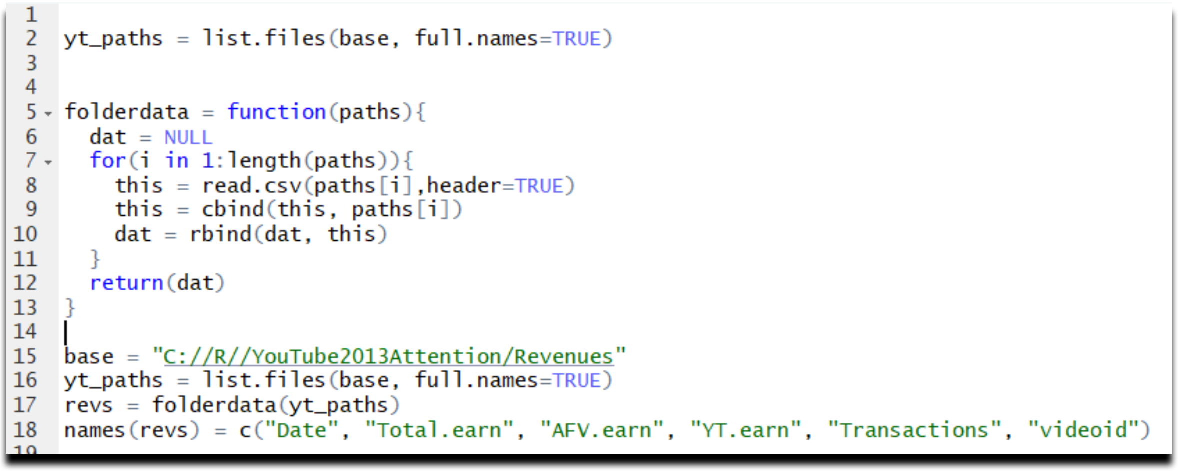
Name	Date modified	Type	Size
Revenues	3/1/2018 9:59 AM	File folder	
Videos	3/1/2018 9:59 AM	File folder	

Within each folder, you will find 98 comma delimited files in which, the name of the files is of the format RevenueXXX.csv or VideoXXX.csv, where "XXX" is a video identifier. For example, in the screenshot below, "Revenue1.csv" is a data set of daily revenues for Video 1, whereas "Revenue1b.csv" is a data set of daily revenues for Video 1b.



Name	Date modified	Type	Size
Revenue1	3/22/2013 8:04 PM	Microsoft Excel C...	7 KB
Revenue1b	3/22/2013 8:26 PM	Microsoft Excel C...	6 KB
Revenue2	3/22/2013 8:02 PM	Microsoft Excel C...	8 KB
Revenue3	3/22/2013 8:01 PM	Microsoft Excel C...	5 KB
Revenue4	3/22/2013 8:00 PM	Microsoft Excel C...	6 KB
Revenue5	3/22/2013 7:54 PM	Microsoft Excel C...	7 KB

To begin assembling the data, you'll need to find a systematic way to read the data into R. The following segment of code will be useful (once adapted) to organize the Revenue data. You'll need to do the same with the Video ("Attention") data, and you'll probably want to strip out some of common formatting in the "videoid" field (there's more code beyond line 18). The idea of this is to give you a forceful hint regarding what you need to do to read the data in.



```
1
2   yt_paths = list.files(base, full.names=TRUE)
3
4
5   folderdata = function(paths){
6     dat = NULL
7     for(i in 1:length(paths)){
8       this = read.csv(paths[i],header=TRUE)
9       this = cbind(this, paths[i])
10      dat = rbind(dat, this)
11    }
12    return(dat)
13  }
14
15  base = "C://R//YouTube2013Attention/Revenues"
16  yt_paths = list.files(base, full.names=TRUE)
17  revs = folderdata(yt_paths)
18  names(revs) = c("Date", "Total.earn", "AFV.earn", "YT.earn", "Transactions", "videoid")
```

Once you have adapted the code, read in the revenue and the video data into separate objects (e.g., revs and vids), and then merge the revenue and video data into a common data frame that day x video observations on revenues and video characteristics for all of the videos (98 in total) and all of the days in the sample (from September 1, 2011 to March 20, 2012).

By the end of "Part 1," you should have a fully assembled data set with the following characteristics:

- Each observation is at the date-video level.
- There should be fields for "daily_earnings" ("Total estimated earnings" renamed), "views", "average_watch_time" ("Average view duration" renamed), "video_id" (this should match the format of the "video.id" field from "adinstream.csv," which we went over in class on Oct 8/9), and "DATE" (the date field, formatted as a "Date" object; see help(as.Date()), or Google information on Dates in R).
- Please also create variables within this data frame for "earnings_per_view" (daily earnings / views).
- You should also merge onto this dataset the "instream.csv" data set, which contains information about the experiment I ran regarding pre-roll ads placed on these videos. We discussed this in class (Experimenting

As a summary, produce both tabular and graphical summary information on daily_earnings, views, average_watch_time. Specifically, it would be useful to report the mean, standard deviation, and median of each of these variables. Also, produce a boxplot for each variable to summarize the distribution of values these variables takes on.

<u>Part 2. Analysis</u>.

Use the merged data set you constructed in Part 1 to answer the following questions:

1. During the pre-period (before Feb 5th),
   - Is average watch time different for videos that have pre-roll ads versus not?
   - Is the number of views different for videos that have pre-roll ads versus not?

   For each comparison, conduct an appropriate hypothesis test, and offer an interpretation in the context of the problem.

2. In the full sample,
   - How does attention (measured by average watch time) relate to daily earnings on a video? Is this relationship statistically significant?
   - How do views relate to daily earnings on a video? Is this relationship statistically significant?
   - How does attention relate to views on a video? Is this relationship statistically significant?

3. Define a date-level variable (named "post") that equals whether the date is during the post period (Feb 5 or later, = 1) or during the pre period (Feb 4 or earlier, =0). Using this variable and regression tools, estimate the specification:

$$daily\_earnings_{it} = \beta_0 + \beta_1 post_t + \epsilon_{it}$$

For your estimate of β_1, is this estimate statistically significant? Interpret this estimate in the context of the problem.

4. Define a video-level variable (named "annoy_later") that equals whether a video went from no pre-roll advertisements to pre-roll advertisements in the post period (=1 if no pre-roll → pre-roll, =0 otherwise). Using this variable and regression tools, estimate the specification:

$$daily\_earnings_{it} = \beta_0 + \beta_1 annoy\_later_i + \epsilon_{it}$$

For your estimate of β_1, is this estimate statistically significant? Interpret this estimate in the context of the problem.

5. Putting the previous two parts together, estimate the following multiple regression equation specification:

$$daily\_earnings_{it} = \beta_0 + \beta_1 annoy\_later_i + \beta_2 post_t + \beta_3 annoy\_later_i \times post_t + \ \epsilon_{it}$$

For your estimate of β_3, is this estimate statistically significant? Interpret this estimate in the context of the problem.

6. For Questions 3, 4, and 5, please repeat the analysis, but using $average\_watch\ \_time$ as the dependent variable, instead of daily earnings.

Part 3. Report. Your report should provide the relevant tradeoffs for a YouTube content provider who is interested in knowing whether pre-roll advertisements will increase daily earnings, but is worried that these advertisements might reduce the amount of attention that users pay to the videos the content provider posts. If there is a significant loss of attention to the content, how much additional revenue can the content provider expect to receive to compensate for this? Please organize your discussion, graphs, and tables in a manner that is straightforward to understand for this YouTuber (who remembers loosely what a p-value is from a years-ago stats class).