Project 3 Report

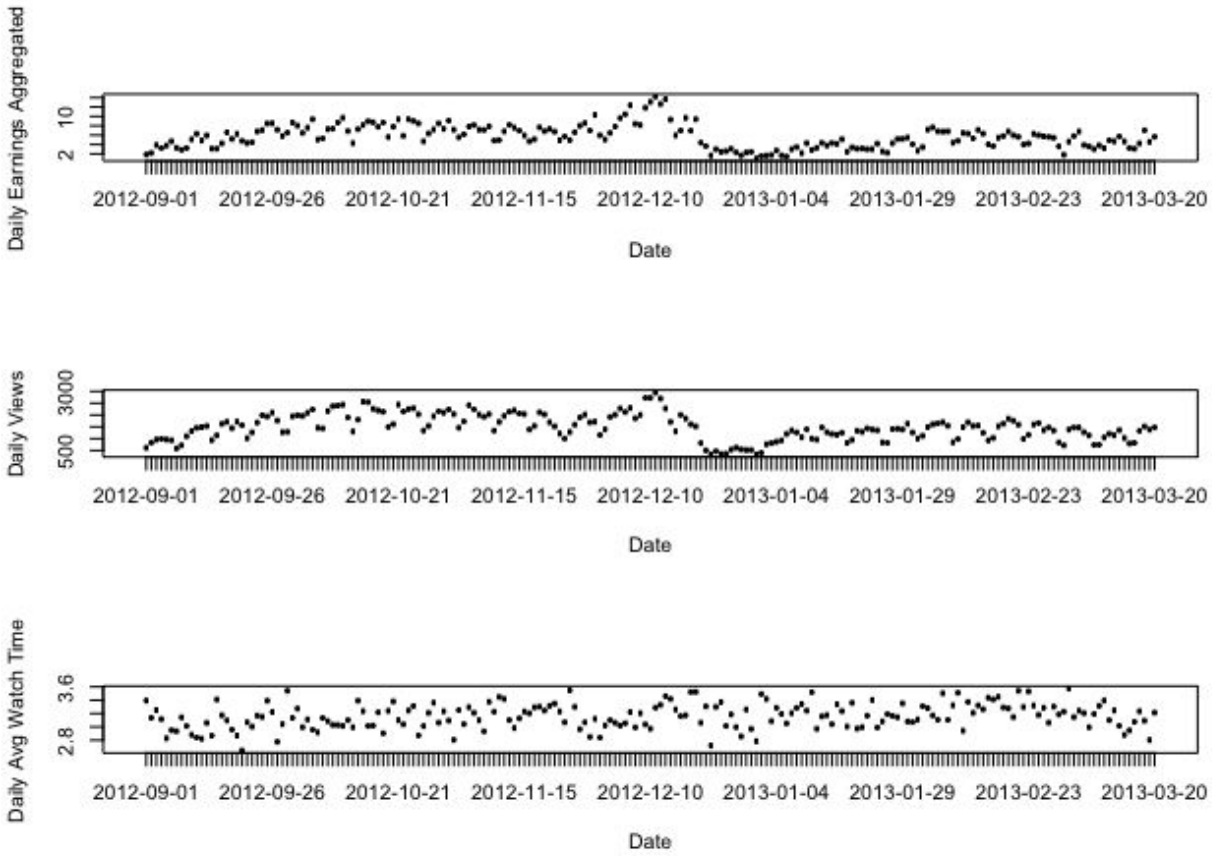
Tyler Chambers, Lakshay Wadhwa, Rachel Bauer, Christine Ottinger, Michael Gambee

Introduction

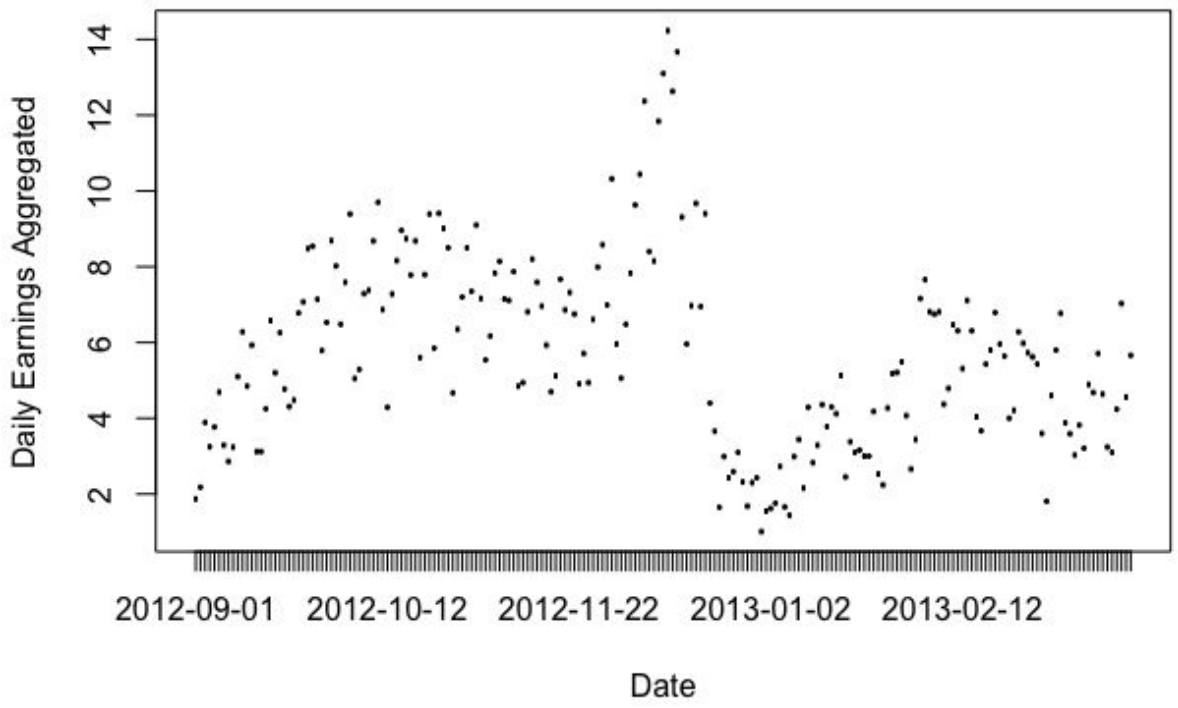
Creating and correctly monetizing content in the digital space becomes a real puzzle when all of the interchangeable factors are taken into consideration. Thankfully this data set really possesses one main question; whether or not the running of Ad pre-roll provides enough increased monetization to justify its intrusive nature. While there is certainly a bit of a judgement call to be made here, we can at least use this data set to provide some numeric measures that might help us calculate our tradeoff.

Examining KPIs and Plots

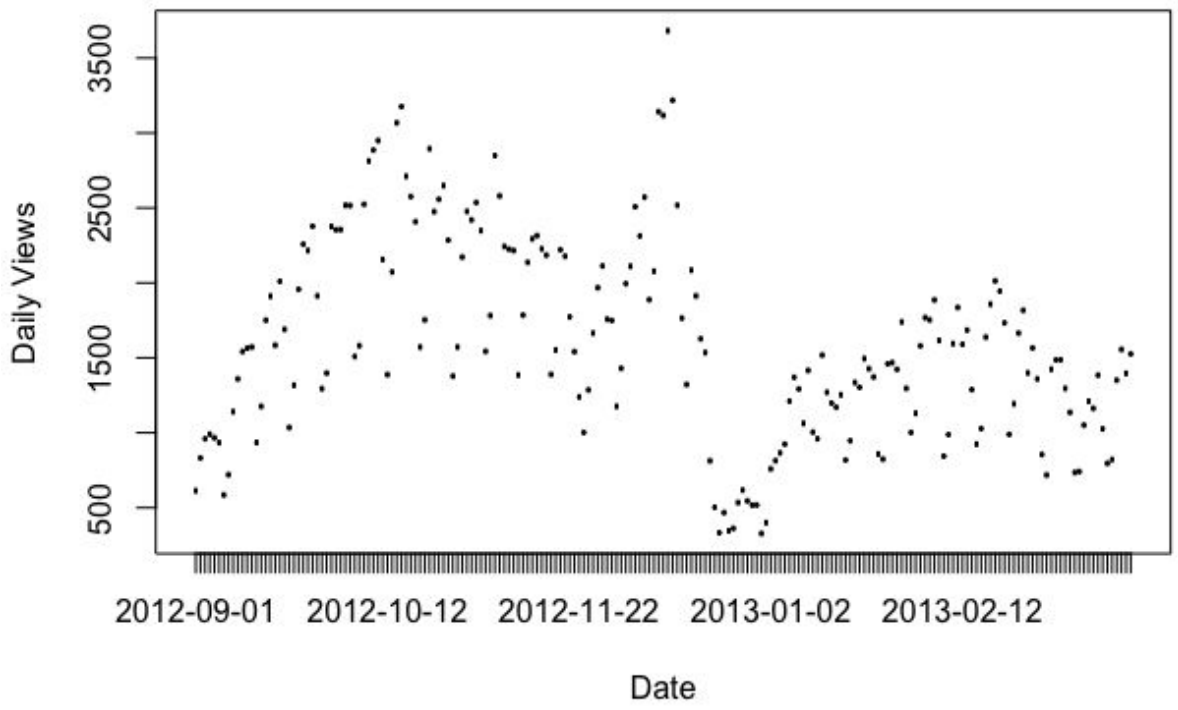
First it is important to develop a baseline to evaluate performance against. We have three main key performance indicators (KPIs) we will be looking at for this data set: Daily Earnings, which measures the daily monetary contribution gained per video from the channel; Views, which measures the daily views gained per video on the channel and can show the channels reach; Finally Average Watch Time, which shows the average amount of time spent watching a video per view for a day, which can be a way of measuring audience attention. Each of these statistics are important, however, the crux of our original question can be sort of boiled down to whether or not an increase of Daily Earnings from Ad Pre-Roll is worth a potential loss in Average Watch Time, which is what we see as audience attention. To help us visualize the landscape here are three scatterplots showing our KPI's throughout our experiment time frame. It should be noted that we winsorized all of the data at a 95^{th} percentile level, as we felt it necessary to balance out the heavy outliers present at points in this dataset. Winsorizing allows us to pull these heavy outliers in a bit in order to balance the data, while still allowing them to impact the data summary and analysis to some degree. Additionally, these scatterplots are measuring the variables aggregated on dates, meaning that for Daily Earnings its the sum of all video earnings for a day, Daily views is the sum of views across all videos for a day, and Daily Avg. Watch Time is the average daily watch time per video across all videos for a day.



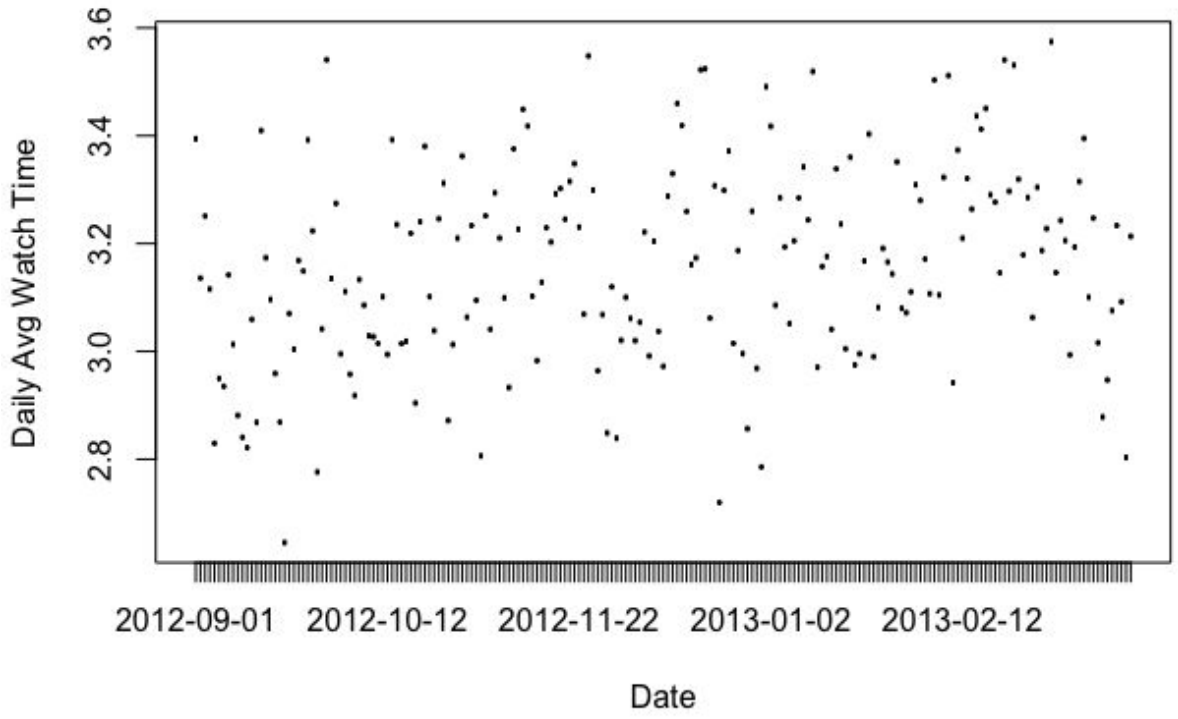
Before we dive into each scatterplot on its own, it should be noted how Daily Earnings Aggregated and Daily Views follow very similar patterns. This makes logical sense, as views are what primarily garner revenue. Daily Average Watch Time seems to follow much more of its own trend, however.



Taking a better look at Daily Earnings Aggregated reveals some interesting trends. First, it seems earnings peak at the times of year when major tests, like midterms, and especially finals, are upcoming. We can see this in beginning to mid October, as well as in mid December. We notice a dramatic decrease in earnings during most schools Winter break period, and then a slow climb again once school enters session in mid January; This mirrors a similar phenomena we see building on the left side of the graph starting at the beginning of September.



Again, we can see that Daily Views follows a similar trend to Daily Earnings, proving the closeness of that relationship. One of the main differences here comes from the units of measurement. The larger counts of views allow the differences in certain areas to look slightly more pronounced. For example, The valley seen around New Years is a bit more distanced in this graph then in Daily Earnings. Additionally the peak during mid October, which would correspond to mid-terms, is more pronounced in this graph.



This view makes the slow positive linear trend present in Daily Avg. Watch Time more apparent. Indeed, Daily Average Watch Time seems to be less hampered by seasonal effects, and instead is seeing a general rise; that is, until the switch of Ad-Pre roll at the beginning of February. This switch seems to have sent the KPI on a downward negative trend. That being said, with only a few weeks of data after this switch, it is hard to say for sure if this trend would continue. Additionally, it should be noted that a linear positive trend does seem to be present, it is likely very weak in nature.

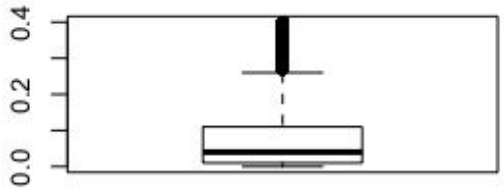
	variables	means	standard_deviations	medians
1	daily earnings	5.716915	2.5047560	5.620000
2	daily views	1486.472637	546.3333877	1471.000000
3	daily average watch time	3.161842	0.1854914	3.167761

Above is a table to lay out some useful metrics for our Daily aggregated KPIs. Small things that are interesting to note is how much the mean differs from the median for daily views. This is likely due to the peak times mentioned above, which skew the mean for the dataset to the right a bit. This was present in daily earnings as well until we winsorized the dataset, which seemed to really help counter this skew. Its effect was notable but not as drastic in the daily

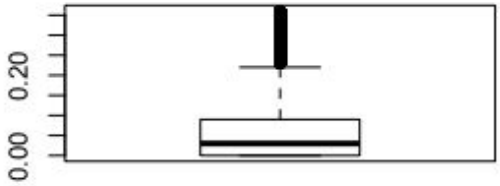
views dataset. We do not see that near as much in the daily average watch time variable, as fewer peaks and potential outliers seem present.

	variables	means	standard_deviations	medians
1	Daily Earnings	0.08189723	0.1083833	0.04
2	Views	21.29434823	17.1336251	16.00
3	Average Watch Time	3.16071485	1.2482459	3.12

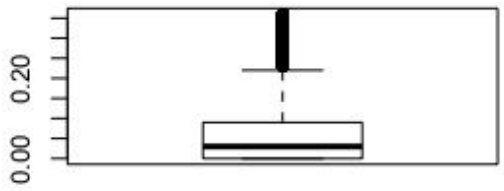
The above table is the descriptive statistics for our KPIs at the per video per day level. Here we can see that the average daily earning for a video is approximately eight cents. The much lower median at four cents tells us that this mean is likely being heavily inflated by some outlier data points, which we can observe in the boxplots below. A similar occurrence is happening with Views, which has a mean around 21.3 views per video per day, but a much lower median at 16. Average Watch Time hovers around 3.16 minutes, and is only slightly lower than its daily aggregate counterpart analyzed in the above table.



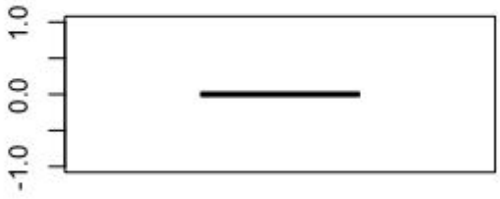
Daily_Earnings per Video per Day



AFV_Earnings per Video per Day

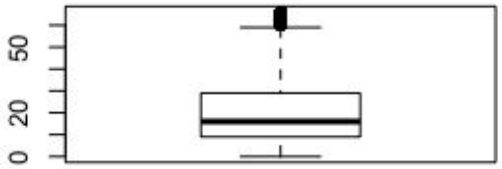


YouTube_Earnings per Video per Day

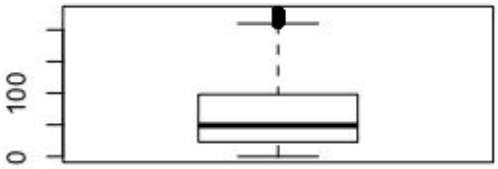


Transactions_Earnings per Video per Day

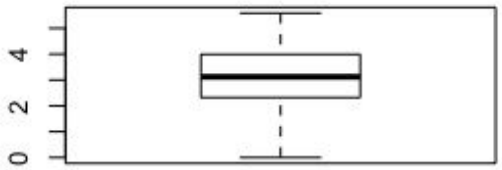
These boxplots show us that on a per video, per day basis, earnings are quite low, with high outliers hitting the $0.40 mark. Most ads seem to bring in on average around 8 cents. You will notice that AFV earnings is a similar plot to Daily Earnings, but just shifted down a bit. This is because this AFV value is based off of Youtube's Ad-Sense program, which tries to attach earnings to actual visitors. We can see from the Youtube earnings boxplot that Youtube lags just a little behind the Daily Earnings boxplot, which makes sense, as Youtube takes portions of your ad revenue. Finally, Transactions is fully zero, which is expected, as you do not have any paid content or super chat functionality.



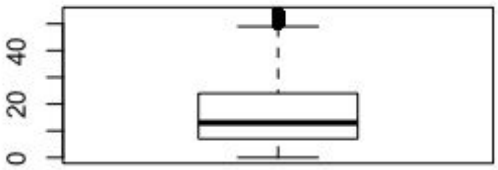
Views per Video per Day



Estimated Minutes Watched per Video per Day

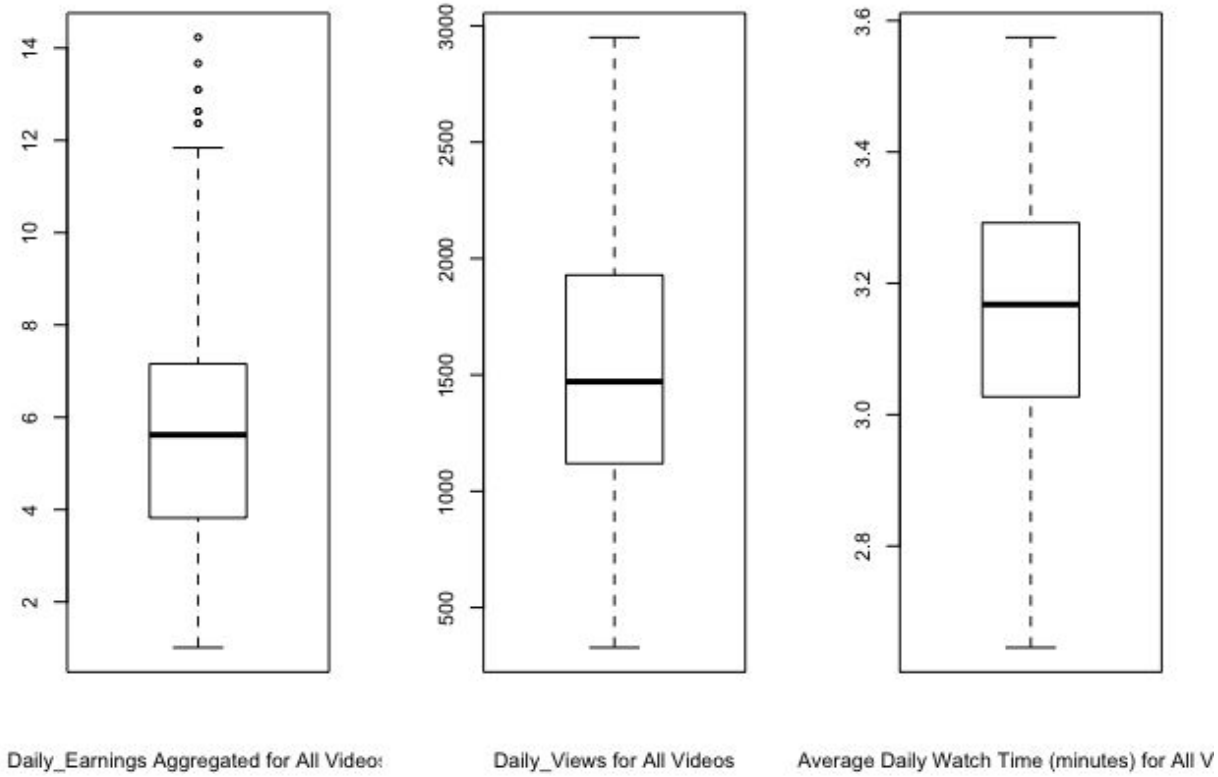


Average Minutes Viewed per Video per Day



Unique Cookies per Video per Day

Our views per video per day box plot shows us that most videos gain anywhere from 0 to 50 views a day, with outliers extending beyond that point. On average though, we expect around 15 views. Unique Cookies shows a similar but stunted plot, as this measures unique visitors. Thus people who watch the same video twice in one day, bots, and those who are able to stop Youtube from placing a cookie will not be shown in this boxplot. The estimated minutes watched per video per day boxplot has a very large range, extending from zero to beyond 200 minutes. This makes sense though, as on busy days some videos are hitting beyond 50 views, and at an average of a little higher than 3 minutes per view, these minutes watched don't seem too outlandish. Finally, it is worth taking a look at the average minutes viewed per video per day boxplot, which shows us that our average view time is a little above 3 minutes per video view. There are actually no outliers in this set, as the videos most likely set a reasonable cap of length themselves.



The Daily_Earnings Aggregated boxplot looks at how much is earned across all videos for a day. We can see that you are approaching an average of six dollars a day, although a handful of outlier days exist extending from twelve dollars to past fourteen. The interquartile range tells to generally expect somewhere between roughly four to seven dollars.The Daily_Views boxplot highlights our daily views across all of our videos. We can see that you are hitting just under an impressive 1500 views per day! The interquartile range says that on most days we would expect around 1100 to 1900 views a day. However, there seems to be a lot of variability present, as our total boxplot extends from below 500 views to almost 3000 views. Interestingly enough, there are no outliers present in our daily views data. Average daily watch time across all videos is a fairly controlled variable. You are currently averaging just under 3.2 minutes, with most data points coming in at 3 to 3.3 minutes. There are no outliers in this variable.

Hypothesis Testing and Regressions

Now that we have looked at most of our base metrics and statistics, we can turn to hypothesis testing and regression analysis to see how our variables are impacting one another, and how you might be able to affect your KPIs with your decision making. First we wanted to run a few two-sample t-tests to see if there are is a difference statistically significant difference between videos that ran pre-roll ads and those that do not. To do this test, we first subsetted the data to only include dates before the February switch, as this was the larger date window in which to run this comparison.

```
        Welch Two Sample t-test

data:  pre.ad$W_Avg_View and non.pre.ad$W_Avg_View
t = -9.6179, df = 10817, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2745897 -0.1816126
sample estimates:
mean of x mean of y
 3.021480  3.249582
```

First, we tested to see if average watch time was different between our two sets of videos. With a P-Value approaching zero, we can conclude that these two sets of videos are indeed statistically significantly different from one another at the 95% confidence level. Additionally, we can see that our mean of non-pre roll ads videos was higher at a level of 3.249582. This means that we can conclude that during the pre-February period, videos that did not have a pre-roll ad had a higher average watch time.

```
        Welch Two Sample t-test

data:  pre.ad$W_Views and non.pre.ad$W_Views
t = -9.4844, df = 10562, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.745194 -2.462264
sample estimates:
mean of x mean of y
 20.29876  23.40249
```

Next we tested to see if views was different between our two sets of videos. With a P-Value again approaching zero, we can conclude that these two sets of videos are indeed statistically significantly different from one another at the 95% confidence level. Additionally, we can see that our mean of non-pre roll ads videos was higher at a value of 23.40249 views. This means that we can conclude that during the pre-February period, videos that had did not have pre-roll ad on average attained higher view counts.

Moving on, we want to run some regressions where we compare our KPIs directly to one another, to see how they may or may not relate.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0547185  0.0024789   22.07   <2e-16 ***
s$W_Avg_View 0.0085989  0.0007295   11.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1079 on 14029 degrees of freedom
Multiple R-squared:  0.009808,  Adjusted R-squared:  0.009737
F-statistic:   139 on 1 and 14029 DF,  p-value: < 2.2e-16
```

This first regression looks at the relationship between Daily Earnings and Average Watch Time. The P-Value for Average Watch Time is approaching zero, meaning that its relationship with Daily Earnings in this model is statistically significant. That being said, the low R-Squared value and small coefficient of .0085989 tells us that while significant, Average Watch Time does not have a large impact on Daily Earnings.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.134e-03  1.207e-03   5.082 3.77e-07 ***
s$W_Views   3.558e-03  4.416e-05  80.570  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08962 on 14029 degrees of freedom
Multiple R-squared:  0.3163,    Adjusted R-squared:  0.3163
F-statistic:  6492 on 1 and 14029 DF,  p-value: < 2.2e-16
```

The next regression looks at the relationship between Daily Earnings and Views. The P-Value for Views is approaching zero, meaning that its relationship with Daily Earnings in this model is statistically significant. Here we are observing a R-Squared value of .3163, meaning that Views are accounting for 31.63% of the variation in Daily Earnings. This means that Views are a more crucial predictive element for estimating Daily Earnings. From our coefficient in our model we can say that each additional view adds on average .36 cents to our Daily Earnings.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.0141308  0.0167358  180.10   <2e-16 ***
s$W_Views   0.0068837  0.0006123   11.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.243 on 14029 degrees of freedom
Multiple R-squared:  0.008928,  Adjusted R-squared:  0.008857
F-statistic: 126.4 on 1 and 14029 DF,  p-value: < 2.2e-16
```

This final KPI comparison regression observes Views relationship with Average Watch Time. Once again, the P-Value for Views is approaching zero, meaning that its relationship with Average Watch Time is statistically significant. However, it should be noted that the R-Squared value is quite low at .008928, meaning that views does not account for much of Average Watch Time's variability. The positive coefficient of .008928 tells us that on average, as views increase by one, we would expect Average Watch Time to also increase by .0089 minutes.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.083926   0.001031  81.370  < 2e-16 ***
s$post      -0.009482   0.002230  -4.253 2.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1083 on 14029 degrees of freedom
Multiple R-squared:  0.001287,  Adjusted R-squared:  0.001216
F-statistic: 18.08 on 1 and 14029 DF,  p-value: 2.128e-05
```

This next regression looks at something a bit different; the relationship between Daily Earnings and video-date pairs after our February 5th change. The post variable is dummy coded, meaning that dates after February 4th are coded as 1, and dates on and before as 0. With that in mind, we can see that post's P-Value indicates that it is significant at the 95% confidence level. The negative coefficient for post tells us that if a video was shown after the February 5th change, on average, it made .95 cents less than the same video shown before this change. This tells me that the previous ad-roll video selection was better. That being said, from our graphs earlier, we know that there are some really strong data points in the pre-February dates, specifically around finals and midterms, so this relationship should be taken with some skepticism.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.059918   0.001301   46.06   <2e-16 ***
s$annoy_later 0.041964   0.001798   23.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1063 on 14029 degrees of freedom
Multiple R-squared:  0.0374,    Adjusted R-squared:  0.03733
F-statistic:   545 on 1 and 14029 DF,  p-value: < 2.2e-16
```

This next regression looks at a variable called annoy_later and its relationship with Daily Earnings. Annoy Later codes videos that switched from not running pre-roll ads to running pre-roll ads after February 5th. We can see that like all other relationships we have observed so far, this one is also statistically significant at the 95% confidence level. Thus its coefficient tells us that videos that fit our above description are predicted to make, on average, 4.2 more cents a day than others.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.056820   0.001462   38.87  < 2e-16 ***
s$annoy_later        0.051686   0.002019   25.61  < 2e-16 ***
s$post               0.014405   0.003152    4.57 4.93e-06 ***
s$annoy_later:s$post -0.045512   0.004363  -10.43  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1059 on 14027 degrees of freedom
Multiple R-squared:  0.04605,   Adjusted R-squared:  0.04584
F-statistic: 225.7 on 3 and 14027 DF,  p-value: < 2.2e-16
```

This final Daily Earnings regression takes our above Post and Annoy Later variables and places them both in the same model, along with their interaction variable. First we should note that all three of these predictors are statistically significant at the 95% confidence level, meaning that they are truly impactful on Daily Earnings. First, our positive coefficient on Annoy Later tells us that on average, if we have a video that changed to running pre-roll after February 5th, it makes 5.2 cents more per day. In a similar vein, videos running after the February 5th date made on average 1.4 cents more per-day. Our interaction variable, however, tells us that if the

video fits switched to running pre-roll ads after February 5th and is running on a date after that time, on average we would expect to make 4.6 cents less. Now that we have they variables all together in the same model, we can garner more actionable insights. Perhaps what is most interesting is that the group of videos that make up annoy later seem to make more ad revenue on average compared to the other videos. We can tell this from that fact that even in pre and post-February environments these videos would be predicted to make more revenue. In the pre period they would make (.056820+.051686)= .108506 dollars per video per day, compared to the other videos that would make .05682 dollars per video per day. In the post period these annoy later videos would make (.056820+.051686+.014405-.045512)= .077399 dollars compared to the other videos which would make (.056820+.014405)=.071225 dollars. This in mind, it is clear from both groups that when they did not run pre-roll ads, they actually made more revenue! We can tell this from the net negative movement the annoy later ads receive when they move into the post February 5th dates, and the net positive movement the other ads receive when moving into the post February 5th dates. All this being said, the lack of data in the post February 5th dates makes it hard to really justify all these conclusions. We know our dataset features many outliers, specifically on the upper end, that could really be biasing the Pre-February dates.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.14110    0.01188 264.382  < 2e-16 ***
s$post       0.09166    0.02569   3.568  0.00036 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.248 on 14029 degrees of freedom
Multiple R-squared:  0.0009068, Adjusted R-squared:  0.0008356
F-statistic: 12.73 on 1 and 14029 DF,  p-value: 0.0003604
```

We also looked at Average Watch Time as a function of our Post and Annoy Later variables. In regards to Post, the P-value shows that the relationship is significant at a 95% confidence level. The coefficient states that the videos after the February 4th cut off date had on average 0.092 more minutes of daily average watch time. Thus, the audience of these videos was slightly more engaged in February and March than in September through January. However, the R-squared value is very low, which means that there is still a lot of variance in the Average Watch Time data that is not accounted for by Post and therefore should not be recommended as a significant predictor.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.05153    0.01522 200.526   <2e-16 ***
s$annoy_later  0.20845    0.02103   9.914   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.244 on 14029 degrees of freedom
Multiple R-squared:  0.006957,  Adjusted R-squared:  0.006886
F-statistic: 98.28 on 1 and 14029 DF,  p-value: < 2.2e-16
```

Looking at the relationship between Average Watch Time and Annoy Later, once again the P-value implies statistical significance, though the R-squared value is also once again very small. According to the coefficient, videos that switched from having no ads before February 5 to having ads after had on average 0.208 more minutes of daily average watch time than those that did not make that change. Based on the conclusions from our earlier comparison between videos with and without an ad before February 5th, it is possible that this increase in average watch time is due to higher watch times during the period before an ad was placed on the video. The "pre-period" is a longer length of time, so it would be weighted more heavily than the "post-period" in which the ads were present.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.02148    0.01717 176.000  < 2e-16 ***
s$annoy_later         0.22810    0.02371   9.622  < 2e-16 ***
s$post                0.13974    0.03702   3.775 0.000161 ***
s$annoy_later:s$post -0.09088    0.05124  -1.774 0.076139 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.243 on 14027 degrees of freedom
Multiple R-squared:  0.008099,  Adjusted R-squared:  0.007887
F-statistic: 38.18 on 3 and 14027 DF,  p-value: < 2.2e-16
```

After running a multiple regression on these three variables, we can see that both Annoy Later and Post have statistically significant relationships with Average Watch Time while the other is taken out of the equation, though the combination of both surprisingly does not have any significance. This may be because of multicollinearity. The R-squared value is still very low, showing that the use of these two variables to attempt to predict Average Watch Time would not be ideal.

Conclusions

In summary, we have concluded that while most of our models and regressions are statistically significant at a 95% confidence level we can not truly say that these are valid interactions. This feeling for doubt can be attributed to clear seasonal effects happening around the finals time period, as students begin studying, that create significant outliers as viewed in our KPI section of the report. The small sample size in post February 5th data also adds to our suspicion as in class we observed much different numbers when only testing on the three weeks of data before February 5th and the three weeks after. Looking at our T-tests we can say that Views help explain about 31.6% of daily earnings and adds an increase in value of 31 cents per view. It is important to note that we're choosing not to label Avg Watch Time as significant, despite it having a P-value nearing zero, due to its poor R-squared value (.0098) and small coefficient (.0086). Multicollinearity may be occurring between Avg Watch Time and Views leading to deceptively significant P-value. When discerning the difference between annoy later and post ads we concluded that the annoy later ads can be perceived as the better of the two but this can be connected to the fact those ads had the luxury of running without pre-roll for all of Fall, the most optimal time we observed.

According to the two-sample t-test we conducted, there is a significant negative impact of ads on average watch time. The delta between the means of the average watch times, however, was only around 0.23 minutes. This decrease could add up over many views, but it depends on

the YouTuber's priorities and standards whether this is a practically significant difference. Data from the regressions cannot validly be used to further these observations, as the Post and Annoy Later variables compare unequal amounts of data and include a seasonal bias. As far as monetary compensation for this loss in attention, we found that videos that switched in February from not having ads to having ads had an increase in daily earnings of about 4.2 cents. Once again, it is up to the YouTuber's priorities whether these balance each other out. If he prioritizes monetary gain, he should put ads on most, if not all, of his videos. If he prioritizes the popularity of his channel, or considers both aspects to be equal, then we believe the loss of watch time is greater than the gain of earnings, especially in the long run.