Ryan Kaplan
Data Analytics Assignment 3

NYT3:

1a) We can see that the average age is 29, and the median is 31, though it ranges from 0 to 109. Although the range is quite large, the average age is only 29, which is quite young. Here, gender is a binary 0 or 1, though its difficult to tell which represents male and which is female. Whichever is 0 seems to be more common, as the "average" is 0.3668

1b) I did the Anderson Darling test, and each of them were far below the 0.05 threshold. Were it not for the large amount of 0's, age looks like a gamma distribution. Since gender category is a binary here, you could call it a Bernoulii normal distribution.

1c) For ecdf, both are about what you'd expect, showing a particularly "normal" distribution for age, with the exception of the 0 case. The gender ecdf is also tricky to interpret, since its only a binary option. The QQ plot looks correct for the gender, and for age its only a little off, again likely owing to the massive amount of 0's at the beginning.

1d) we got a correlation score of .366, so it's unlikely that age and gender are strongly correlated here. That makes sense intuitively, as its not like older men are much more likely to read newspapers than older women, for example. I'm sure that the many age-zeros are throwing it off as well.

1e) I wish I knew whether the 0 gender meant male or female. With age. I suspect the many zeros in the age represent a null-value, and perhaps would be worth getting rid of. 1 year olds and onward do seem to be a legitimate value however, so I would guess that there would be many valid 0's, unless they start counting at 1, at which case they would all indeed be null.

NYT4

1a) The average and median age is still 29 and 31, though this time it ranges from 0 to 108. The most common is again 0. The "average" gender is close, at .3695 this time, so slightly more of 1 than 0 compared to last time.

1b) I again did the Anderson Darling test, and got the same result as before. Like the first time, except for the 0's, the distribution is broadly normal. Gender is still a binary, and seems to be essentially unchanged, so its a Bernoulii normal distribution.

1c) The ecdf results are basically the same as before, showing the age as mostly fitting. Gender is still normal looking, albeit in binary form. Same for the qq plots, both showing basically the same results as before.

1d) we got a correlation score of .369, so although it's slightly less correlated than before, we have the same broad result. This should be consistent, given their previous similarity. It would be very strange to find them wildly correlated dispite everything else remaining the same

1e) I'm curious what this is showing. The data is so similar between the two sets, even more so than simply carrying out the same study a year apart or something.

NYT5

1a) The median impressions are 5, with a very close average of 4.999. They range from 0 to 18. The "average" gender is close again, with .3683, basically the same result.

1b) Anderson Darling again gives us a p-value well below the threshold. While gender is unchanged, the impressions are more normally distributed than age was. This is certainly due to the many 0's in age.

1c) The ecdf for gender is the same, and is quite normal for impressions. The QQ plot for impressions is much closer to the normal distribution. As explained before, this is what we'd expect, and gender is unchanged.

1d) Interestingly, it seems that they are very slightly negatively correlated. The correlation is -0.0005, so it would seem that a gender of 1 has fewer average impressions, 0 has more. This does seem like the kind of thing that could be affected by gender, so that makes sense.

1e) I was surprised to see impressions so normally distributed. I would've guess that most people would be either high or low.

NYT6

1a) The median impressions are again 5, with an almost identical average of 4.995, ranging from 0 to 20 this time. The clicks range from 0 to 4, with an average of 0.09. Interestingly, each quartile is exactly 0.

1b) Anderson Darling once again gives us a p-value well below the threshold, for both of them. Impressions are unchanged, and it seems that the vast majority of clicks are 0. This seems like a good example of an exponential distribution

1c) The ecdf for impressions is the same, and is quite normal for clicks, with a strong exponential curve. The QQ plot for impressions is basically identical. For clicks, we see scaling results for increasing numbers of clicks, but its more akin to a binary choice like gender was than a true range of results.

1d) Impressions and clicks seem to be positively correlated. The correlation is .132, so more impressions correlates to more clicks. This makes sense, as impressions and clicks are usually tracked together, and a 0 impressions almost always means no clicks.

1e) I'm interested at the difference between clicks and impressions. It seems that 0 impressions means 0 clicks, but not vice-versa.

NYT7

1a) The clicks are virtually unchanged, and sign-in is a binary, like gender was. The average sign-in was 0.7, so more people did so than not. The first and second quartiles for sign-in are 0, and the last two are 1.

1b) Anderson Darling shows a p-value far below the threshold, for both of them. Clicks are unchanged, and it seems that again most sign-ins are 1. Like gender was, this seems like a good example of a bernoulli-normal distribution.

1c) We have the same ecdf for clicks, and the sign-on is reminiscent of gender, since we have a binary result here. The QQ plot for sign-ons is also similar to what gender was, which makes sense. The QQ plot for clicks is again identical.

1d) Clicks and sign-ins seem to be negatively correlated. The correlation is -.105, so more clicks corresponds to less sign-ins. Much like impressions, clicks and sign-ons should be pretty interconnected, so this runs contrary to what I would expect

1e) I would assume that clicks are necessary for sign-ons, so it's counter-intuitive for me to see them negatively correlated. I'm not sure why this is the case

NYT8

1a) We see the same results as before, for each of them. Interestingly, the boxplots for the two of them are nearly identical, which makes sense. The only difference being signed_in is closer to 1, and gender is closer to 0.

1b) Anderson Darling again give a p-value well under the threshold, for both of them. Their histograms are nearly identical, which makes sense. They are almost mirror images, which fits since they're closer to one and zero, respectively.

1c) Both of our plots are the same as before, which makes sense. Their ecdf plots are identical, just one is shifted down. Similarly, their qq plots are the same, just with one shifted left.

1d) There is actually a strong correlation between sign-ins and gender. The correlation is .5011, so gender 0 has less sign-ins than gender 1. Sign-in habits seem logical to differ between men and women, so this isn't entirely unexpected.

1e) I would like to know the specific details of the study, to see how much I would expect sign-in habits to change by gender.

NYT9

1a) The results for both are the same as before. There are slight differences, for example
the click average is now 0.09268, and the age average is now 29.44991. The median age is still 31, and for clicks its still 0.

1b) The Anderson Darling test consistently gives them scores well below the threshold. Their histograms are basically the same as before, which makes sense.
We again see clicks with only a few possible results, going down with time, and age largely normal, with the exception of the large amount of 0's.

1c) The two plots are basically the same as before, as we would expect. Clicks again acts mostly as a binary, while age is mostly normally distributed. It is no surprise at this point to see such similar results between datasets.

1d) We see a slight negative correlation between age and clicks. The correlation is -0.061, so the lower the age, the more average clicks.
Intuitively this makes sense since young people tend to be more tech-savvy

1e) I wonder how much the correlation between clicks and age is skewed by the large number of 0's. This would probably depend greatly on
whether all of the 0's are illegitimate or not.