STAT 231 Assignment 1

Analysis 1

1a: My ID number is 20936386. I will be analyzing @JohnTory from my personal accounts, and @CTVKitchener from my organizational accounts.

1b:

Day	Personal account	Organizational account
Monday	34	28
Tuesday	27	28
Wednesday	29	51
Thursday	31	24
Friday	39	18
Saturday	21	26
Sunday	30	34

1c: Barplots of `day.of.week`:

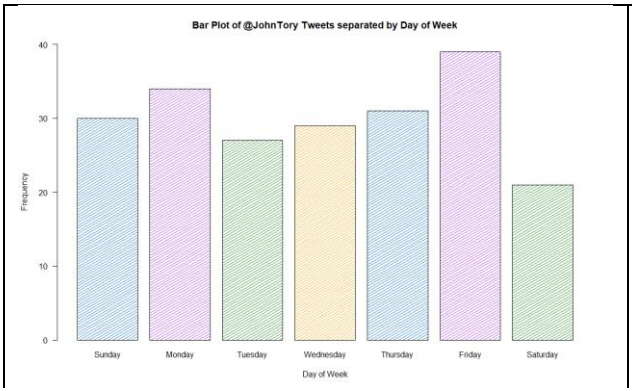


Figure 1a) Bar plot of @JohnTory tweets separated by day of week

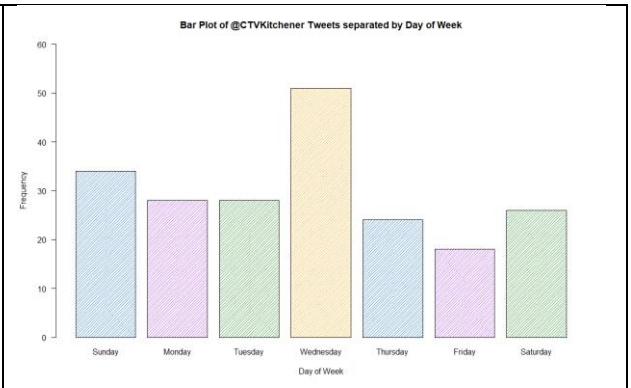


Figure 1b) Bar plot of @CTVKitchener tweets separated by day of week

1d: The distributions of `day.of.week` for @JohnTory and @CTVKitchener are somewhat similar. For @JohnTory, we can see that days share relatively similar numbers of tweets. There is increased Tweets on Monday/Friday and decreased Tweets on Saturday, as mayors also have weekends and the elevated number reflect the information that is missed on the weekends. For @CTVKitchener, most days also share relatively similar number of tweets, except for Wednesday. This is probably because Wednesdays are proven to be the best time to post for the greatest number of likes (https://blog.hootsuite.com/best-time-to-post-on-instagram/), thus incentivizing the organization to post mor often on that day.

Analysis 2

2a: My ID number is 20936386. I will be analyzing @JohnTory from my personal accounts, and @CTVKitchener from my organizational accounts.

2b:

Sample statistic	Personal account	Organizational account
Mean	11.74561	11.80785
Median	10.93222	12.00833
SD	3.650504	7.305696
Skewness	0.6818602	-0.01933564
Kurtosis	2.904695	1.674756

2c: Relative frequency histograms of `time.of.day.hour` with superimposed probability density function curves:

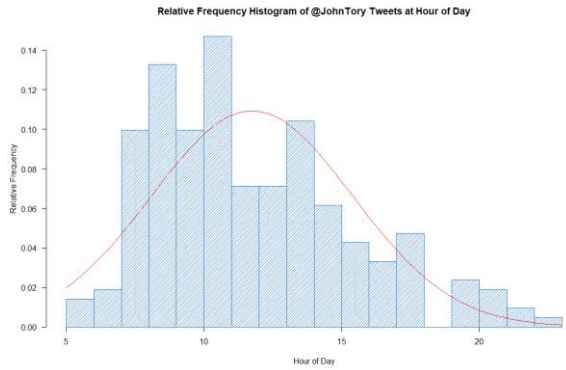


Figure 2a) Relative frequency histogram of @JohnTory tweets separated by hour of day

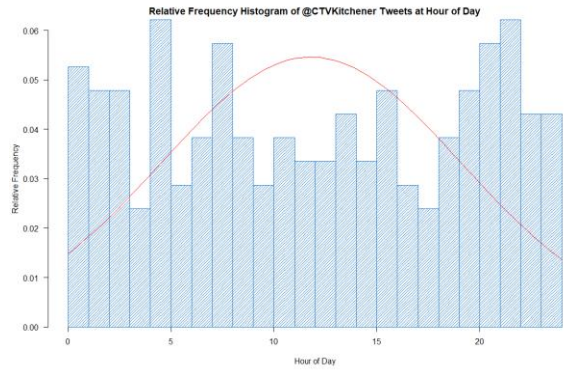
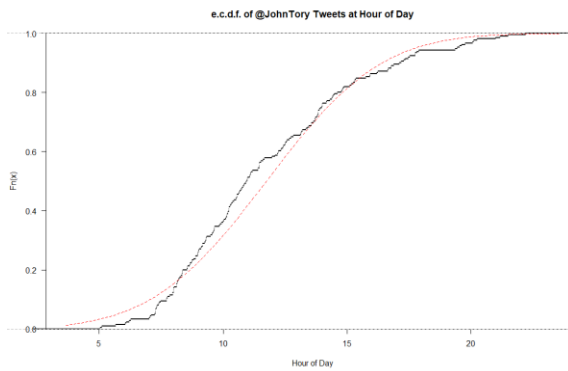


Figure 2b) Relative frequency histogram of @CTVKitchener tweets separated by hour of day

2d: Empirical cumulative distribution function plots of `time.of.day.hour` with superimposed cumulative distribution function curves:



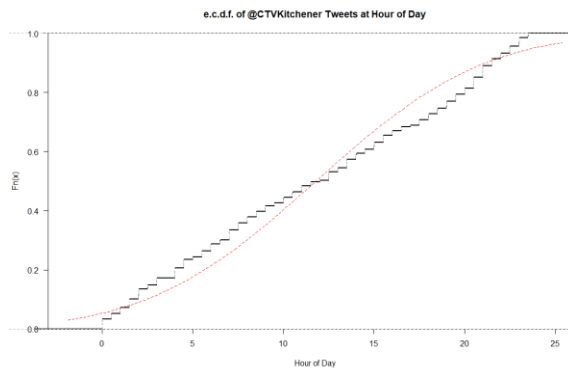


Figure 3a) Empirical cumulative distribution function plots of @JohnTory tweets separated by hour of day	Figure 3b) Empirical cumulative distribution function plots of @CTVKitchener tweets separated by hour of day

2e: In my sample, @JohnTory has tweeted approximately 10% of their tweets before 8am, and @CTVKitchener has tweeted approximately 10% of their tweets before 3am.

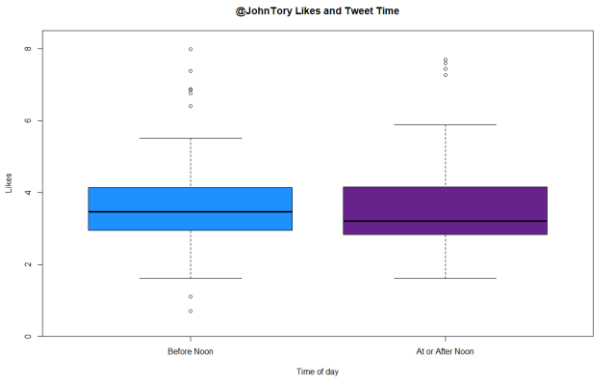
2f: **Personal account** (@JohnTory): Based on the plot in part 2c, we can see that there is only slightly correlation between the Gaussian cumulative distribution function curve and the supplied data. Our data has a large peak with a large tail. Even though the graph is positively skewed, this has some resemblance to our Gaussian curve. This makes sense as the personal account is run by one person, who needs to sleep. Thus, we can expect less tweets in early mornings and late evenings. Secondly, since John Tory is a mayor, he would most likely provide updates every morning, thus explaining the peak in tweets at that time. Our generated Gaussian distribution matches slightly to sample data and reflects our analysis. There is a lot of error, due to the low sample size of 200 tweets. Thus, there is slight correlation between our sample data and the generated Gaussian curve.

Organizational account (@CTVKitchener): Based on the plot in part 2c, we can see that the relative frequency histogram does not fit well with the suitable superimposed Gaussian c.d.f. curve. The plot distribution of the data seems to be uniform, as there is no defined peak and every hour seems to have similar number of tweets. This is most likely due to that CTV Kitchener runs 24 hours to report news at any time of day, and thus there is no Normal distribution. However, our data generated from a Gaussian distribution to converge to one strong peak, and there should be two tails on each side, as seen in the red Gaussian curve. Overall, the Gaussian model does not fit well at all with the supplied data.

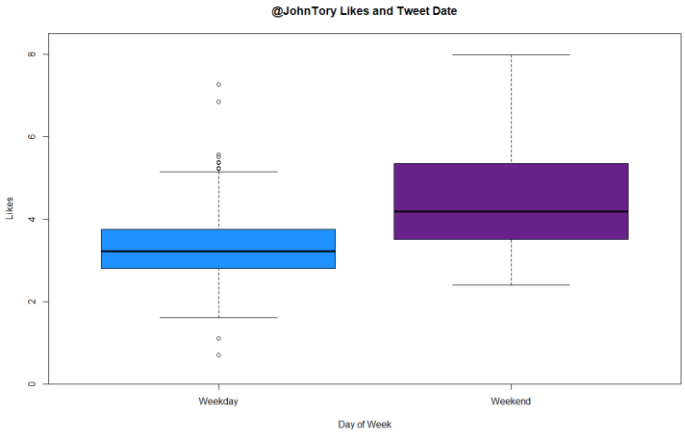
Analysis 3

3a: My ID number is 20936386. I will be analyzing @JohnTory for Analysis 3.

3b: Side-by-side boxplot of `likes` for tweets published before noon vs. after noon:



3c: Side-by-side boxplot of likes for tweets published on weekends vs. weekdays:



3d: Based on the analysis of 3b, we can conclude with only some degree of certainty that posting before noon will generate more likes. Posting in the afternoon will generate more variance in the interquartile range, as shown in the whisker lines and box, however, posting before noon has more outliers. Thus, posting before noon has a larger peak, but also has a stronger tail. Posting after noon has a lower peak, and a weaker tail. Finally, posting before noon has higher median likes.

Based on the analysis of 3c, we can conclude with a confident degree of certainty that posting on the weekend will generate more likes. Since the median is significantly higher and the interquartile range has small overlap, we can be confident. However, since the whiskers overlap between the two graphs, there is some uncertainty.

In summary, @JohnTory should post before noon on weekends.