Assignment 1 Aaron Krishnapillai

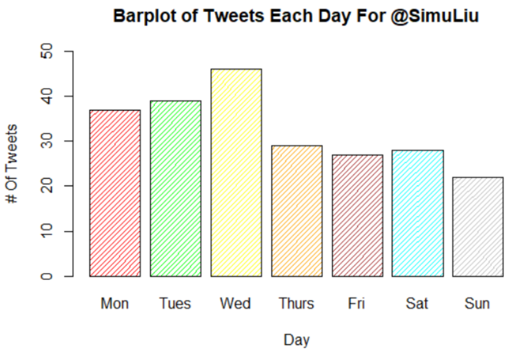
Analysis 1

1a: My ID number is [20892405]. I will be analyzing [@SimuLiu] from my personal accounts, and [@FANEXPOCANADA] from my organizational accounts.

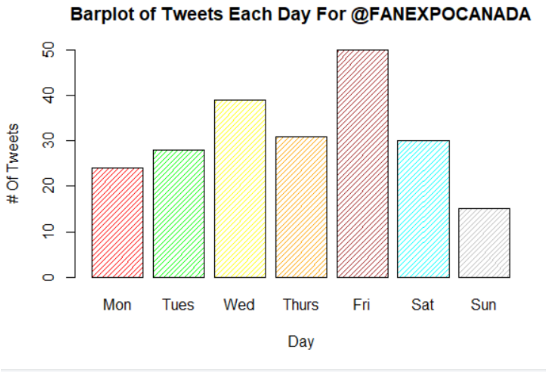
1b:

Day	Personal account	Organizational account
Monday	37	24
Tuesday	39	28
Wednesday	46	39
Thursday	29	31
Friday	27	50
Saturday	28	30
Sunday	22	15

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



(a) @SimuLiu



(b) @FANEXPOCANADA

1d: The distributions of `day.of.week` for @SimuLiu and @FANEXPOCANADA are somewhat similar. For @SimuLiu, we can see that he frequently tweets at the beginning of the week and progressively tweets more each day until Thursday. From Thursday we see a decline in production of tweets for the rest of the week. While for @FANEXPOCANADA, we can see that the same pattern occurs where he increases in tweets each day but there is a dip on Thursday. After Thursday, it spikes to its maximum on Friday and progressively declines for the rest of the week.

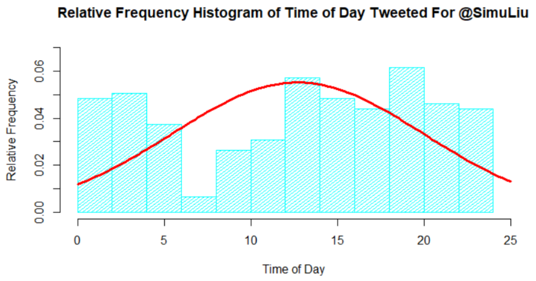
Analysis 2

2a: My ID number is 20892405. I will be analyzing @SimuLiu from my personal accounts, and @FANEXPOCANADA from my organizational accounts.

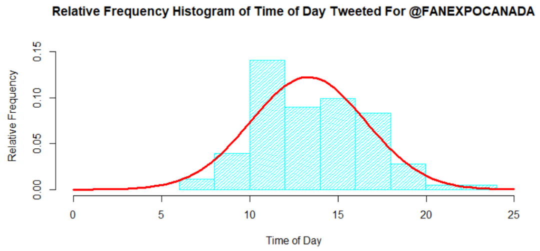
2b:

Sample statistic	Personal account	Organizational account
Mean	12.710	13.324
Median	13.789	13.001
SD	7.230	3.263
Skewness	-0.311	0.254
Kurtosis	1.778	2.546

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

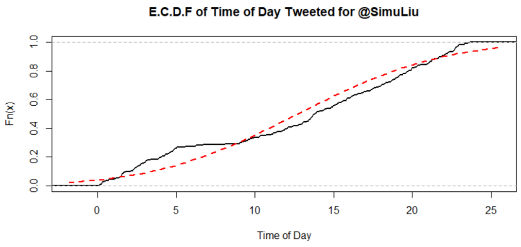


(a) @SimuLiu

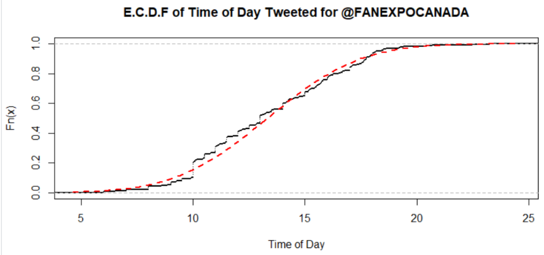


(b) @FANEXPOCANADA

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



(a) @SimuLiu



(b) @FANEXPOCANADA

2e: In my sample, @SimuLiu has tweeted approximately 10% of their tweets before 2am, and @FANEXPOCANADA has tweeted approximately 10% of their tweets before 10AM.

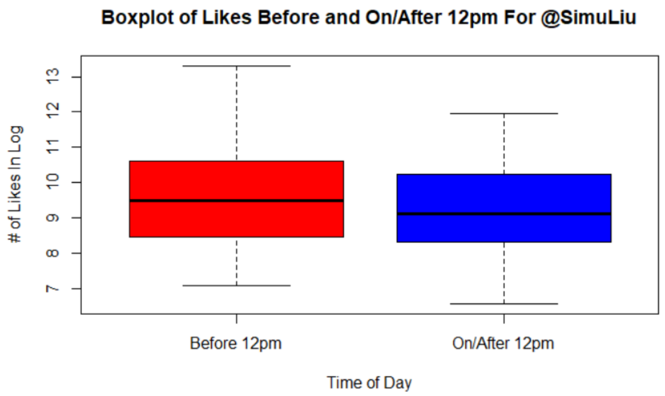
2f: **Personal account (@SimuLiu))**: Based on the plot in part 2c, we can see our histogram that majority of tweets are past 3pm while for data generated from a Gaussian distribution we would expect to see majority of tweets between 5 and 10. We can see a discrepancy between the Gaussian and barplot between 5am and 12pm. On the Gaussian model we expect to see him tweet more often during that period of time. The Gaussian model shows that the number of tweets decline after 12pm but the data seam to stay above the Gaussian curve. Between 8pm and 12pm, the data is above the curve. From 2a we can see that the data is skewed to right as the skewness ¡ 0. The kurtosis ¡ 3 meaning that the peak of the Gaussian model is board. This seems to match the plot has the points seems to be evening distributed during the day. There is a discrepancy in the histogram between 5am to 7am. Overall does not fit well as there are many discrepancy between the curve and the histogram

Organizational account (@FANEXPOCANADA): Based on the plot in part 2c, we can see an increasing and then deceasing trend in tweets through out the day. On the Gaussian model we see a similar trend and a peak close to the normal distribution. We can see that skewness ¿ 0 as there is a discrepancy in between the curve and plot between 10am and 12pm. What we expect to see on our Gaussian is to see majority of our tweets within 1 standard deviation and this seems to be the case. The curve fits the plot well with a slight discrepancy stated above. The height of most powers see the line up with the graph fairly well and the curve is an accurate representation on the relative frequency through out the day .

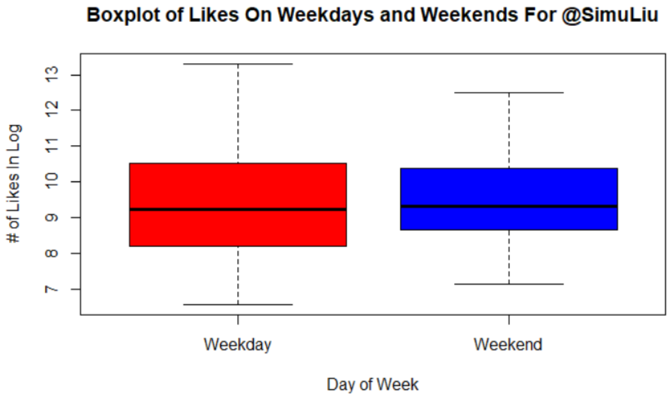
Analysis 3

3a: My ID number is 20892405. I will be analyzing @SimuLiu 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 results of Analysis 3b, we would recommend publishing tweets before 12pm and the median number of likes before 12pm is higher than the median number of like after 12pm. Both plots have a similar spread in data but the data below 12pm is placed higher that the data after 12pm. This tells us that the average number of likes is higher before 12pm.

Based on the results of Analysis 3c, we would recommend publishing tweets on the weekend. Even though the median is about the same the range of the data seems to be closer the median. We will see more consistent likes. There may be uncertainty and the median for the weekend is placed lower within the interquartile range meaning it might not be the true median.