BUDT 758Q – Data Models and Decisions

HW 1

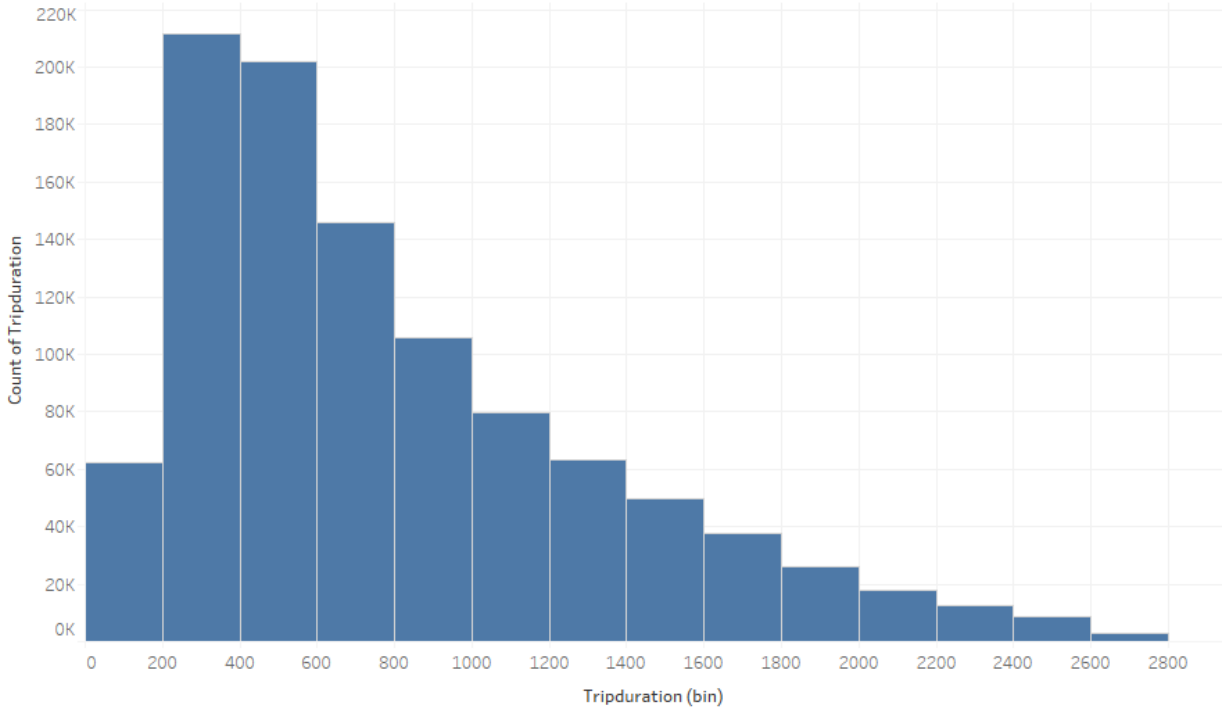
Last Name: Dighe
First Name: Madhura

Section: MB11

Q.1 City bike Trip Data

a) Figure a shows histogram of trip duration filtered for records having trip duration from 61 sec to 2700 sec. Histogram is **positively skewed** with long tail on the right-hand side of the graph.
Trip duration data for most of the trips fall in the lower range of histogram whereas we have some trips which have trip duration on higher side, resulting into long tail on right hand side as shown in figure.
For this positively skewed data, mean value of trip duration will be greater than median and mode.

Figure a

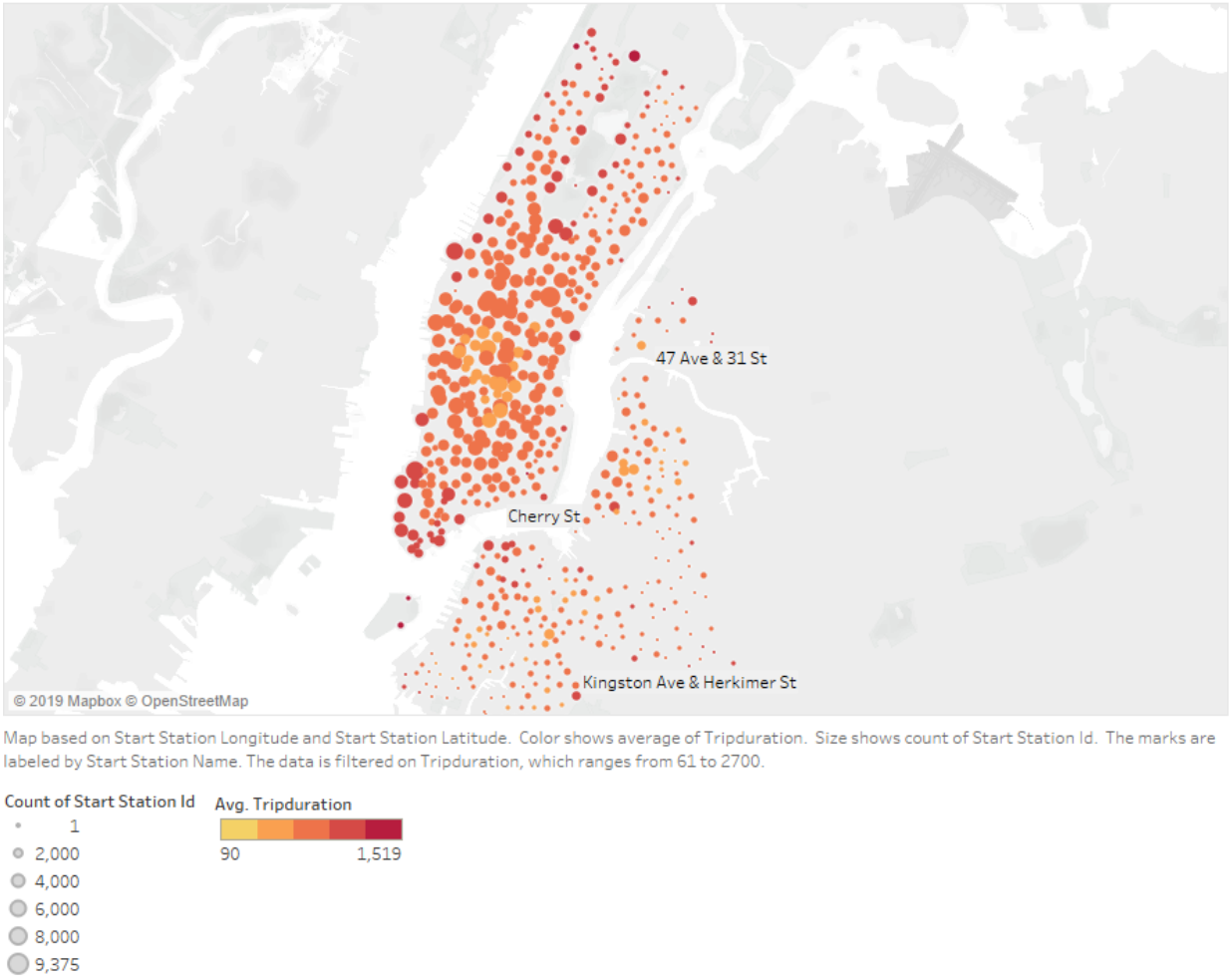


The trend of count of Tripduration for Tripduration (bin). The data is filtered on Tripduration, which ranges from 61 to 2700.

b) Figure b shows Map based on Start Station Longitude and Start Station Latitude.
Color of each dot (start station) shows average of Trip duration (gold being lowest and red being highest).

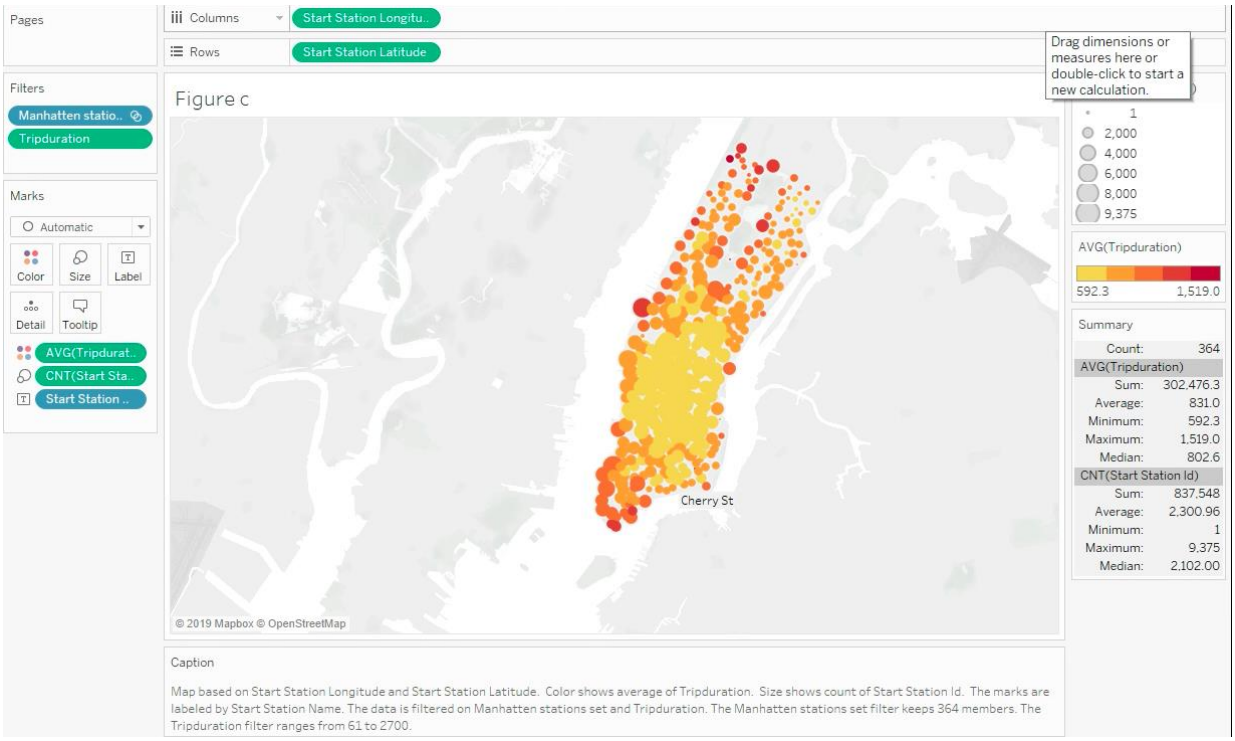
Size of the dots shows how many trips start from that station ID. More the number of trips starting from station will have bigger dot size in the plot than other start stations.

Figure b



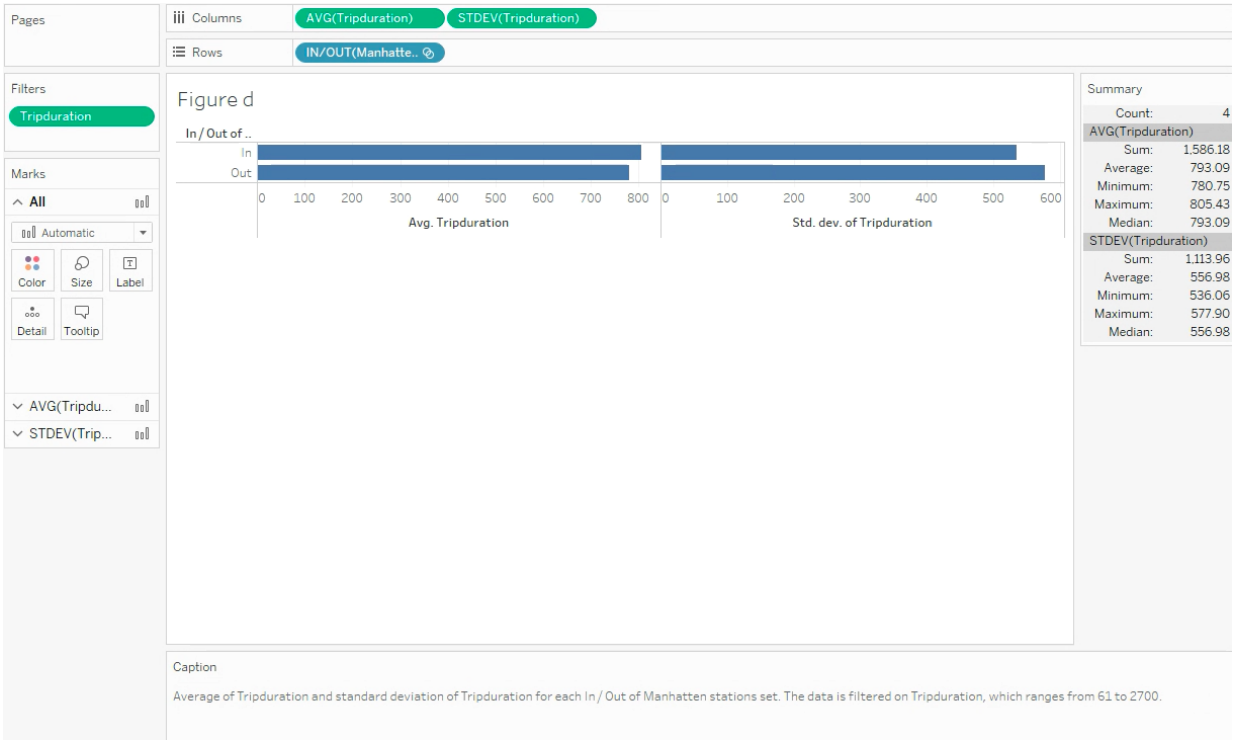
The marks are labeled by Start Station Name. The data is filtered on Trip duration, which ranges from 61 to 2700 sec.

c) Figure c shows the island of Manhattan as a set using ‘lasso selection’.



d) Average of trip duration that starts on island of Manhattan: **805.43 sec**

Average of trip duration that starts outside island of Manhattan: **780.75 sec**



Standard Deviation of trip duration that starts on island of Manhattan: **536.06 sec**

Standard Deviation of trip duration that starts outside island of Manhattan: **577.90 sec**

Just from these average trip durations, we can only say that people on Manhattan island are using city bikes trips for longer duration travel than people using city bike trips outside Manhattan.

Also, for trips on Manhattan standard deviation is less suggests there is less difference between mean of trip duration and all trips observations. And more standard deviation suggests more difference between mean of trip and all trip durations.

But, **from difference in standard deviations of both, we cannot get any significant data interpretations between two sets for now**.

e) Part-1: figure e-1

Figure e-1

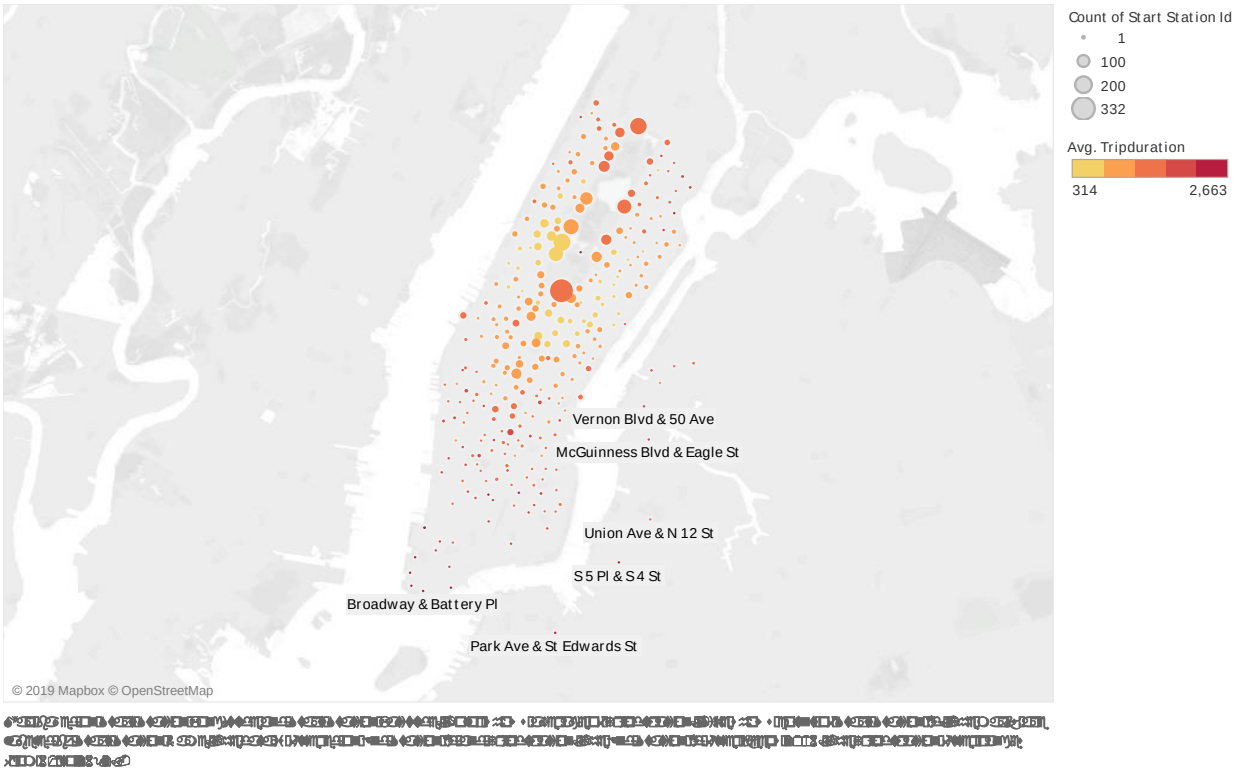


Figure e-1 shows, number of trips that are ending at station #2600 with trip duration between 61 to 2700 seconds.

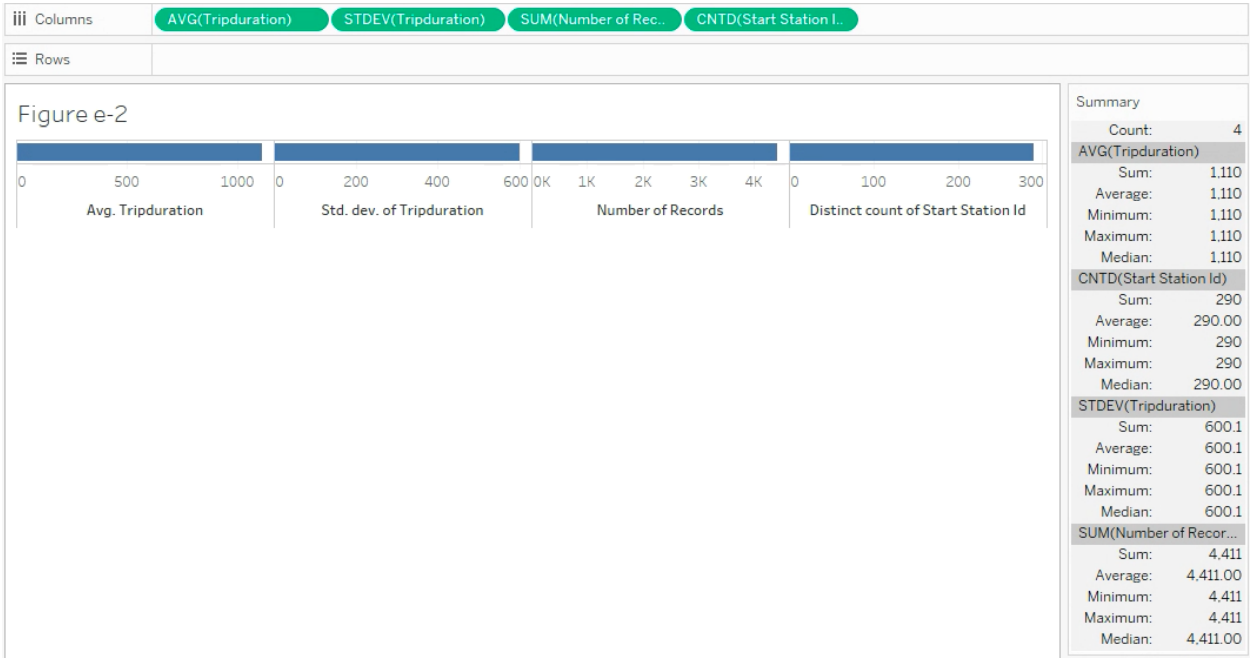
Overall, we have total of 4411 trips ending at #2600.
Color of each dot in the figure shows average trip duration (gold being minimum increasing to red being maximum) and size of the dots shows count of how many numbers of trips starts at particular start station to end at #2600.
Also, tags are included to show where the trips start to end at #2600 station.

Trips, that are ending at End station #2006 usually takes **1110 seconds** on average.

Central Park S & 6 Ave has the greatest number of trips which are ending at Station #2006.

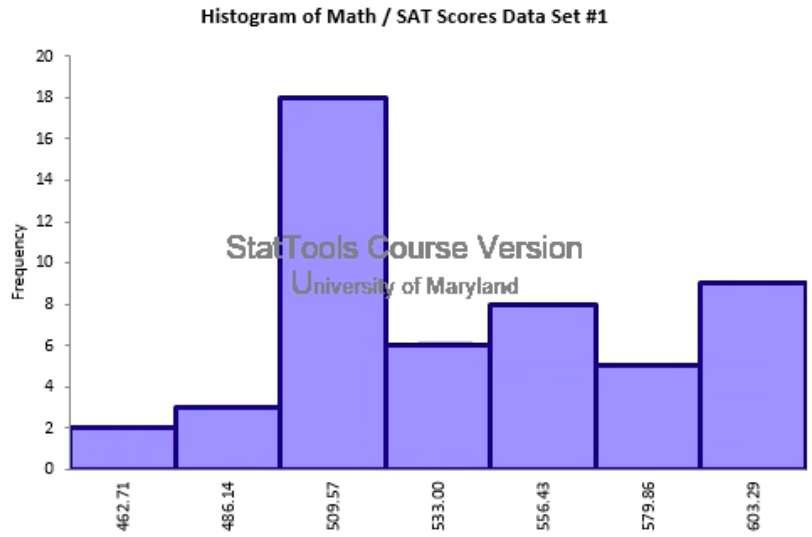
Part-2: figure e-2

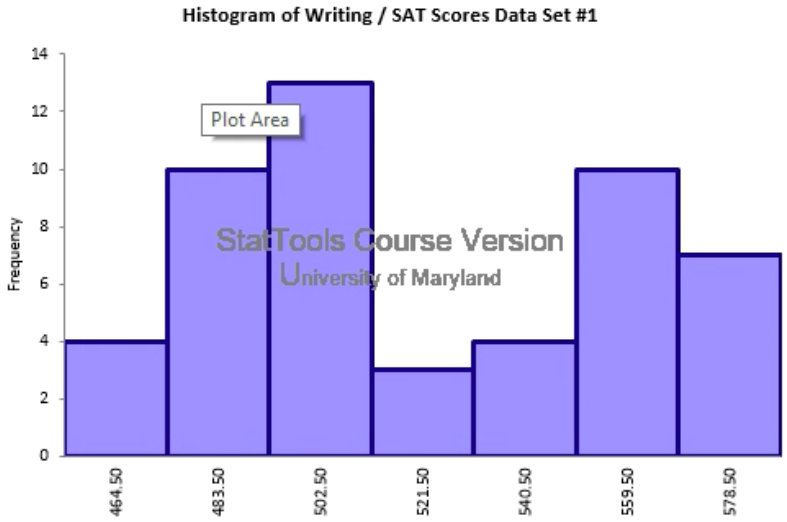
I. Average trip duration of trips ending at #2600: 1100 sec
II. Standard Deviation of trip duration: 600.1
III. Trips in June 2007 ending at #2600: 4411 trips
IV. How many different stations people picked there bike to end at #2600: 290

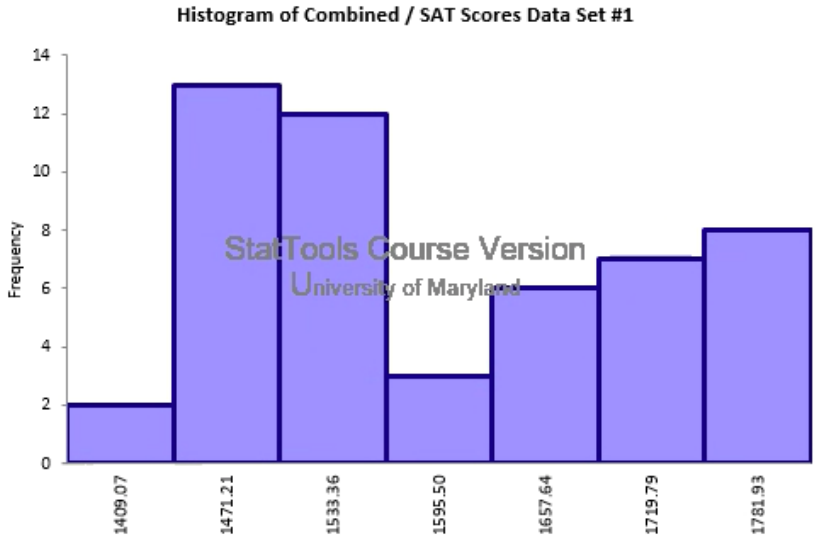


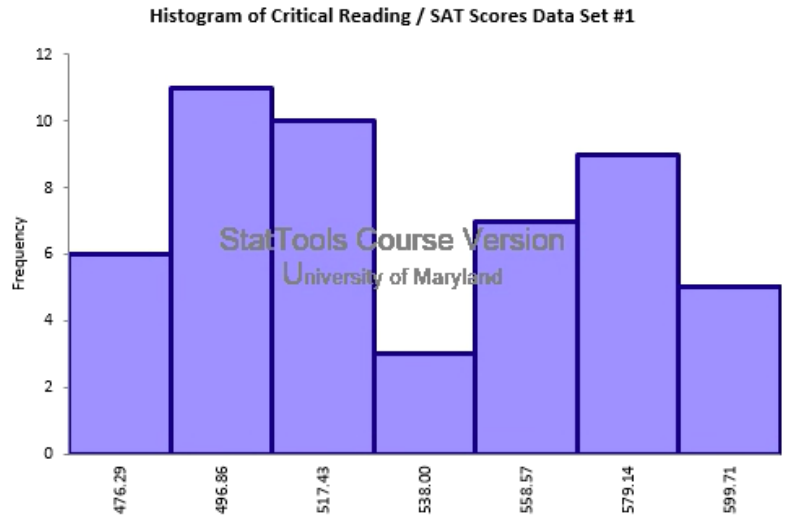
Q.2 SAT Test score data set

a) Histogram shows that the Distribution of 5 numerical variables that are **positively skewed** and **not symmetric.**









b) Mean of the combined variable is 1594.59. If we add mean of each separate variable i.e. writing, critical reading, math variables, it is equal to 1594.58.

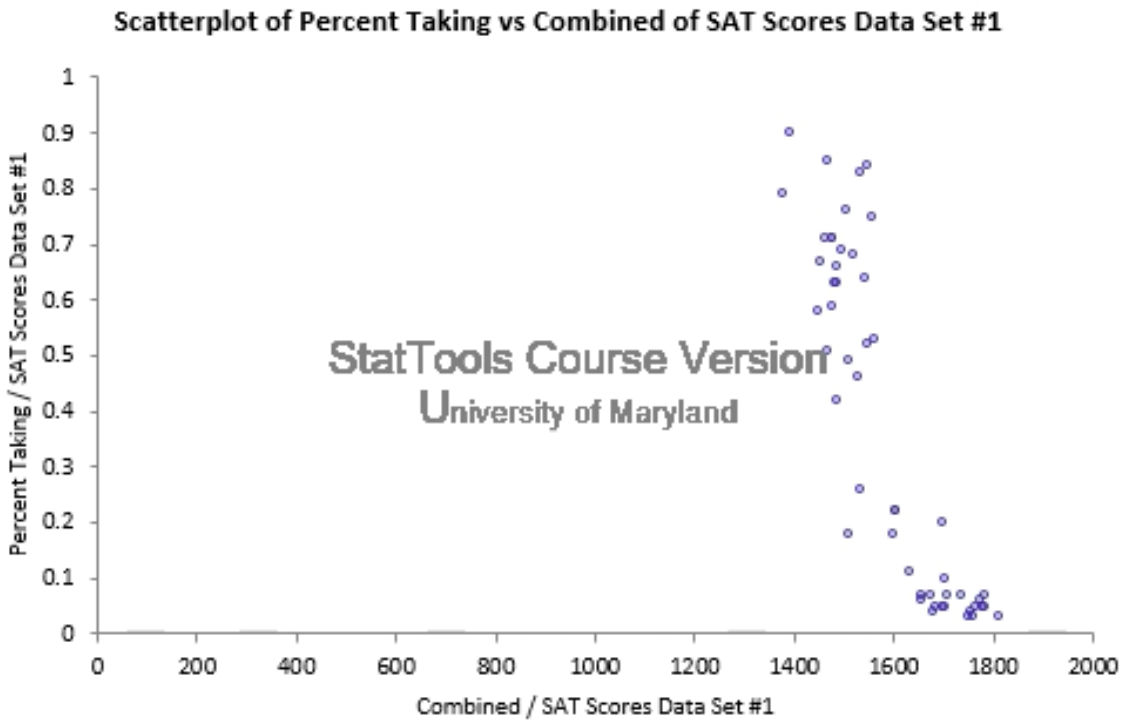
Similarly, Median of combined variable is 1556 and if we add median of each separate variable i.e. writing, critical reading, math variables, it is 1558.

The relation between Mean and Median is as follows:
Mean: writing < critical Reading < Math < combined
Median: writing < critical Reading < Math < Combined

Hence, the relationship between Means of variables holds true for Medians as well.

c) Combined SAT score and percentage taking the exam are **negatively correlated**. From the excel, we can see the **correlation coefficient as -0.879**.

Comments: As the percentage of taking exam increases, combined SAT score is going to decrease.



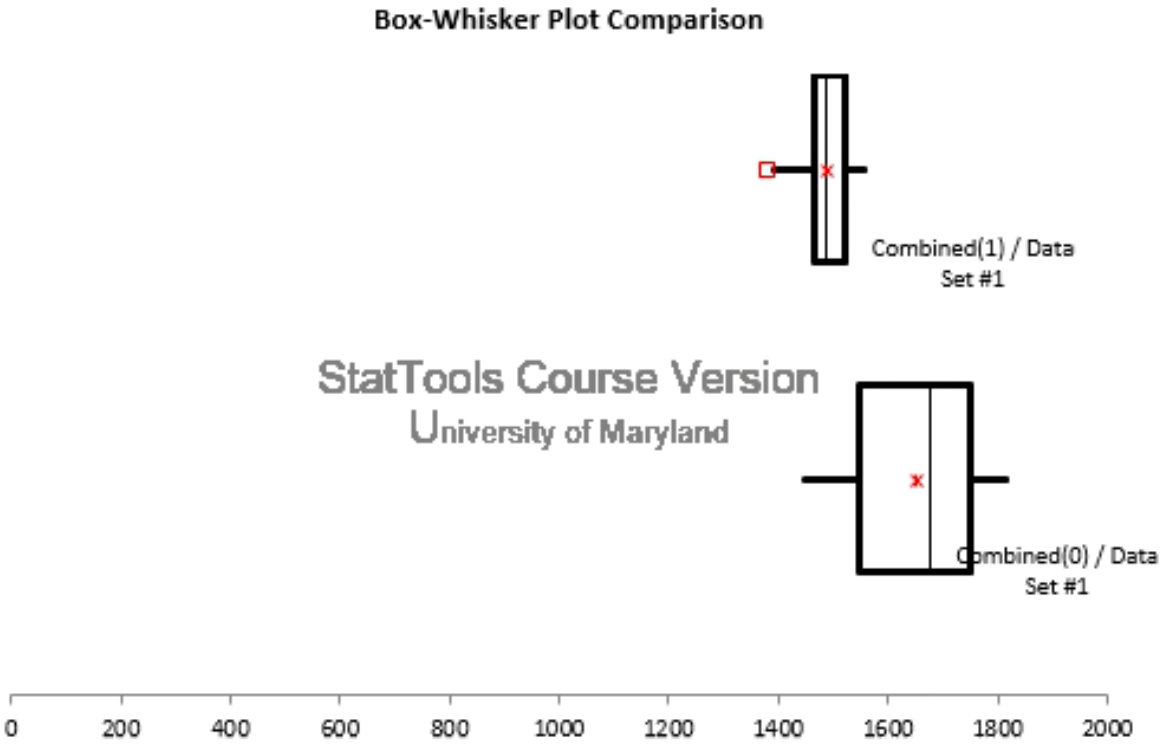
Correlation	-0.879

d) The variables mentioned in question are highly correlated not because the sum of three components writing, critical reading, math is equal to combined score; but because score increase in each separate component (writing, critical reading, Math) would essentially increase the overall combined SAT score.

e) Box-Whisker plot Comparison:

1. Median of combined score for percentage taking < 60% is more than the median combined score with >60% taking.
2. Combined (1)/ data plot is right-skewed whereas combined (0)/ data plot is left-skewed.
3. Shorter box for the first case (percentage taking > 60%) simply indicates that it is reliable and median is good indicator to describe center tendency of the combined score, as almost all the observations are located nearby median. We also can spot one outlier on the left side of plot.

   For second plot, it simply has longer box plot with many observations located away from the median, thus it is unreliable to use median and define center tendency of the combined score for percentage taking <60%.


Box-Whisker Plot Comparison
Combined(1) / Data Set #1
StatTools Course Version
University of Maryland
Combined(0) / Data Set #1
0
200
400
600
800
1000
1200
1400
1600
1800
2000