Assignment 5: Statistics, Statistics, Statistics and more Statistics

Austin Derbique A01967241 10/9/17 CS5665

1. Let's say that you access the data set num_friends for each function call and num_friends is the original data set that is not sorted. Every time you call a function such as median, first quartile, third quartile, variance, standard deviation, or inner quartile range, you need a sorted list. The others can be accessed by searching through the list in one go, but as for the functions previously mentioned, can be optimized by calling them on an already sorted list.

2. Done. Look at SummaryStatistics.py and statistics.py. linear_algebra.py is from the author's github repo.

3. Look at prob3_results.csv for the table located in the results directory.

	Day User	Day Session	Day Event	Hour User	Hour Session	Hour Event
Count	639	639	639	14759	14759	14759
Q1	234.5	302	12733	9	9	245
Q3	750	1007	48695	44	41	1902.5
Min	95	130	4662	1	0	0
Sum	426712	608477	33360139	651701	608222	33388629
Median	366	499	20175	21	19	800
IQR	515.5	705	35962	35	32	1657.5
Range	6299	9098	700005	1406	1347	63814
Mode	223	277	11177	2	1	4
Vari-ance	600685.292	1480818.107	6885518878	5768.03747	4902.743588	23905753.3
Mean	667.780908	952.2331768	52206.7903	44.1561759	41.2102446	2262.255505
Max	6394	9228	704667	1407	1347	63814
StdDev	775.646056	1217.842001	83044.03186	75.9501699	70.02196653	4889.516659

4. These values are strange because they do not add up to the correct amounts. Although the two files are recording in hours and days, respectively, they should still sum up to the same amount. There is also a similar oddity with session and event data. There seems to be a more dramatic error between day user sum and hour user sum. The reason for this could be because whatever was logging the file did not catch all the data or potentially the timezones shifted the measurement times. (Resulting in shifted results therefore throwing off the accuracy).

5. Look at prob5_results.csv for the table located in the results directory.

6. Observation 1: It appears that the user count drops off rather dramatically between certain months.

For instance, from February to March the count goes from 1240 to 293. This could likely be because of improving weather and therefore people check the OSS less often.

Observation 2: The mean appears to be consistent through the year with only small variations from month to month. This is likely because the mean is the average and on average, the data is not changing a whole lot from month to month.

7. Look at prob5_results.csv for the table located in the results directory.

8. Observation 1: Most users only have one session. I believe this is because in one hour's time, a user uses the application only once.

Observation 2: Significantly more users, sessions, and events are logged during day hours while there is a significant drop off in logged data during the middle of the night. This is most likely because people are sleeping during the night.

9. The plots can be found in results_with_graphs.xlsx in the results directory or in the prob9 directory with contained python scripts using matplotlib.

10. Observation 1:The box and whisker diagrams for the median and 1st quartile seemed to be very similar. The 3rd quartile was quite different from the other two.

Observation 2: The 3rd quartile was significantly larger than the previous two with larger bounds as well. 130 to 512 whereas the 1st quartile is merely 4 to 10. I am not sure why this is the case.

11. Work for this problem can be found in the prob11 directory. I was unable to get a working stem and leaf plot displayed.

12. Work for this problem can be found in the prob11 directory.

13. Based on what I understand of stem and leaf plots and the times of interest for problem 12, the graphs would show variations of a larger lead surround the time of the eclipse. I imagine that the eclipse day would stand out from the other days because of the heightened user, session, and event data logged.