For the following exercises, please include a single .zip file that contains your source files, data output, etc. Also include (outside of the zip file) an MS-Word or PDF file containing your responses to the individual questions.

In class, we discussed comparing data against averages to identify interesting or anomalous events and extended the idea to moving averages. For data reported on a daily basis, the 30-day moving average for a given day is computed as the average of the values for the current day and the prior 29 days (30 days total). See https://en.wikipedia.org/wiki/Moving_average#Simple_moving_average (Links to an external site.)Links to an external site.. The 30-day moving standard deviation is the standard deviation of the values over those same 30 days. We can then normalize values using the formula NormalizedValue = (Value - Mean) / StdDev. The result is a representation of each value measured in standard deviations from the mean relative to the 30-day period. See https://en.wikipedia.org/wiki/Standard_score (Links to an external site.)Links to an external site..

You will find daily and hourly session data for the One-Stop-Shop from November 2011 to October 2017 in the files section on canvas. Use that data to complete the following exercises.

1.

Using the daily session data, implement Python code to compute 30-day and 365-day moving averages, standard deviations and normalized values. Please include your Python code and an output file (CSV) containing the computed values as well as the dates and original values.

2.

Using your data from Exercise 1, identify the days that stand out as anomalous during the following time periods:

a. June 1st, 2017– August 31st, 2017

b. December 1st 2016 – January 31st, 2017

c. Indicate the criteria you used to determine the anomalous days for Parts a and b.

3.

We also discussed the comparison of moving averages as potential indications of broader changes in activity. For instance, if analyzing stock market data, a period in which the 30-day moving average crosses from being less than the 365-day moving to being greater might indicate that the stock is being broadly purchased where-as a cross-over from greater than the 365-day moving average to less than could indicate broad sales. Using your data from Exercise 1, answer the following:

a. On what days did the One-Stop-Shop session data 30-day moving average cross from below to above the 365-day moving average.

b. On what days did the One-Stop-Shop session data 365-day moving average cross from above to below the 365-day moving average.

c. What do these cross-over points mean in terms of system usage?

d. Explain how you identified the days for Parts a and b.

4.

Repeat Exercises 1 and 2 using the hourly data to identify anomalous events in the hourly data. Again, use 365-day and 30-day moving averages. (To reiterate, I am asking for you to use moving averages measured in days on data that is reported hourly. And, I am asking you to identify which hours were anomalous.) Please provide code and responses as requested in Exercises 1 and 2.

5.

Are your results from Exercise 4 consistent with your results from Exercises 1 and 2? Please explain.

6.

Moving averages and standard deviations may be adversely impacted by relatively large data values. The median and inter-quartile range are not as susceptible to extreme data. You can compute a moving median, moving inter-quartile range, and normalized values using the formula NormalizedValue = (Value - Median) / InterQuartileRange in a similar manner so-as to avoid the impact of extreme data. Repeat Exercise 1 and 2 using median, inter-quartile range and normalized values. Please provide code and responses as requested in Exercises 1 and 2.

7.

Are your results from Exercise 6 consistent with your results from Exercises 1 and 2? Please explain.

In class we simulated the birthday problem using the assumption of 365 days in a year and uniform probabilities of birthdates. It was observed that this may not be realistic because of leap years and non-uniform probabilities of birthdates. A distribution of birthdates gathered from insurance records can be found here: http://www.panix.com/~murphy/bdata.txt (Links to an external site.)Links to an external site.. Let's assume this distribution is representative of the general population. Use this as well as my code from class to answer the following:

8.

Implement a random birthday function to randomly select a birthday day of year (an integer between 1 and 366, inclusive) according to the distribution referenced above. Test your function by randomly generating "a lot" of random birthdays and comparing against the distribution. Please include your source code and an explanation of your testing approach.

9.

Using your random birthday function, modify my code to generate random birthdays for various class/group sizes and compute probabilities for various outcomes: no birthdays the same, two birthdays the same, etc. Compare your results to those using the original assumptions of 365 days and uniform probabilities for birthdates. Are the results roughly the same or "significantly" different? Please explain.

Using the coin toss simulation code from class, answer the following:

10.

Determine the number of tosses that can be made and tabulated within one-minute on your computer. Please indicate this number and your method for determining it.

11.

Using a simulated “fair” coin, what is the longest streak you observe in 10 separate runs using the number of tosses identified in Exercise 10 each time?

12.

Change the coin toss to have a 60% chance of landing heads and 40% chance of landing tails. What is the longest streak you observe in 10 separate runs using the number of tosses identified in Exercise 10 each time?

13.

Increasing the probability of heads further what is the least probability of heads (>50%) you can find that consistently produces streaks of length 100 with the number of tosses identified in Exercise 10? Explain how you found your answer.

14.

Estimate and check the following: Increasing the probability of heads further, what is the least probability of heads (> 50%) you can find that consistently produces streaks of length 200 with the number of tosses identified in Exercise 10? Was your estimate accurate?

15.

Could the longest streak in a given number of tosses be used to test a hypothesis about the probability of heads in a single toss? Please explain.