Risto B. Rushford
GSCM 451-Spr 2019
HW02

Short-Answer Problems

These concepts can appear on the optional short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. Sample vs. Population
    a. What is the distinction between a sample and a population?

A population is all of the data points in a set of data. A sample is a subset of data points. Some populations are too numerous to do any meaningful analysis, and so samples are used as a substitute.

    b. Do we observe the sample or the population?

Unless the population is particularly small, we typically observe samples. Samples are subsets of the population.

    c. Which is more important: knowledge of the sample or of the population? Why?

Knowledge of the population is what is sought, and this is done through analysis of the sample. A sample of sufficient size taken of random data points from the population is meant to provide a sufficient representation of the population it is taken from.

2. Explain the concept of sampling variation of a statistic such as the average (arithmetic mean).

No sample is perfectly representative of its population. Because of this, each sample will have variations in the statistics provided by analysis.

3. What are descriptive (or summary) statistics? Provide at least one example.

Descriptive statistics are coefficients used to summarize a data set. This includes what are called "measures of central tendency" and "measures of variability". An example of a measure of central tendency is the mean; an example of a measure of variability is the standard deviation.

4. What is the law of large numbers? What is the implication for forecasting?

The Law of Large Numbers in probability theory essentially states that the descriptive statistics of a *sufficiently large* number of datasets will converge on the true statistics of a population.

5. What is a run chart?

A run chart is a line graph of data points as measured from the output of a process. Run charts are used to find trends and patterns in the data.

6. Stable process.
    a. Define a stable process.

In Six Sigma, process stability is a measure of the consistency of a set of measurements. While variability is inherent to any set of measurements, a highly stable process will produce results consistently within specified limits.

    b. What does the run chart of a stable process look like?

The run chart of a stable process will look like a squiggly line centered on a specified value and within specified tolerances. A zoomed out view will look like a straight line. Each data point will fall within a normal distribution centered on the expected value.

    c. For a stable process, what is the forecast for the next value in time?

For a stable process, the next value forecasted will be within the normal distribution. For a tightly controlled process, the next data point will be very nearly on the line extrapolated from previous data points.

7. Define trend:

A trend is the existence of a long-term increase or decrease in a data set.

8. Define seasonality.

Seasonality is a fixed pattern of change in the values with a fixed and known frequency

9. What is a time series chart? How is it similar and yet different from a run chart?

A time series chart is a series of data points recorded periodically over time. It is similar to a run chart in that it is sequential in nature. However while time series charts are made from data points measured over a consistent frequency of measurement, run charts do not require measurements to be time consistent unless specified.

10. What are the two types of visualizations for displaying multiple time series as part of the same visualization?

One type of data visualization for displaying multiple time series is a stacked multiple time series plot. Another is to plot the time series in adjoining (usually stacked) panels.

11. How does the concept of a serial number relate to store dates in Excel and R?

Serial numbers are dates that are sequentially numbered by day. Sequential numbers in Excel use the origin "1900-01-01" while R uses sequential numbers with origin "1970-01-01"

Worked Problems

Graphics

When asked to create any output, text or graph, copy and paste that text or graph into your homework document. That is, include it as part of your answer. For the following questions, I often provide prompts for this display, but they will generally not appear in subsequent homework assignments. Just always provide the data visualizations (graphs) you are asked to generate.

Relevant Output

When you provide an answer in your homework, always provide the relevant piece of output – not all of the output, only the relevant piece.

Displaying R Output

The input and output text in the R console are just simple text. Copy and insert the text into the same, single document in which your interpretation and other commentary appear. Display copied text output from a program such as R in a monospaced font, such as Courier New, usually of size 9 pt. The same applies to the figures. Just copy and paste into your Word doc, but often re-size smaller after pasting because the default size is too large. Turn in this single document as your homework assignment via the class website Dropbox.

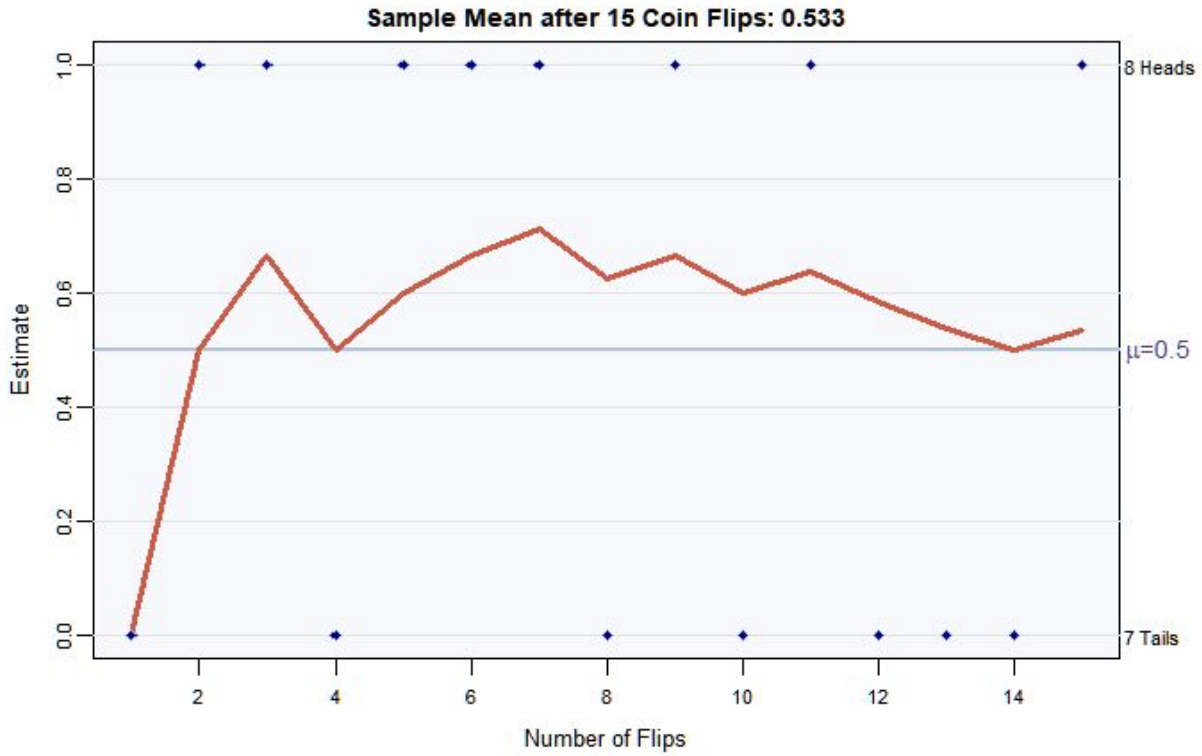
Getting Help

Because the R console displays simple text, if you have a problem and wish for help: (a) copy the R problem direct from the console, (b) include the output of the Read function to show your data that was read, and (c) forward in an email to gerbing@pdx.edu. Do not do a screenshot, just a simple copy and paste.

1. Coin flip simulation and forecast

Here we perform the simulation of coin flips where we know the true population value for the probability of the event Heads. For a. and b. the coin is fair, that is, a probability of Heads is 0.5, and for c. the coin is biased. Now look at the data from the coin flipping, pretending not to know the true underlying probability, the case of real life.

a. With the lessR function simFlips(), simulate flipping a fair coin 15 times.
   i. Show the resulting visualization.

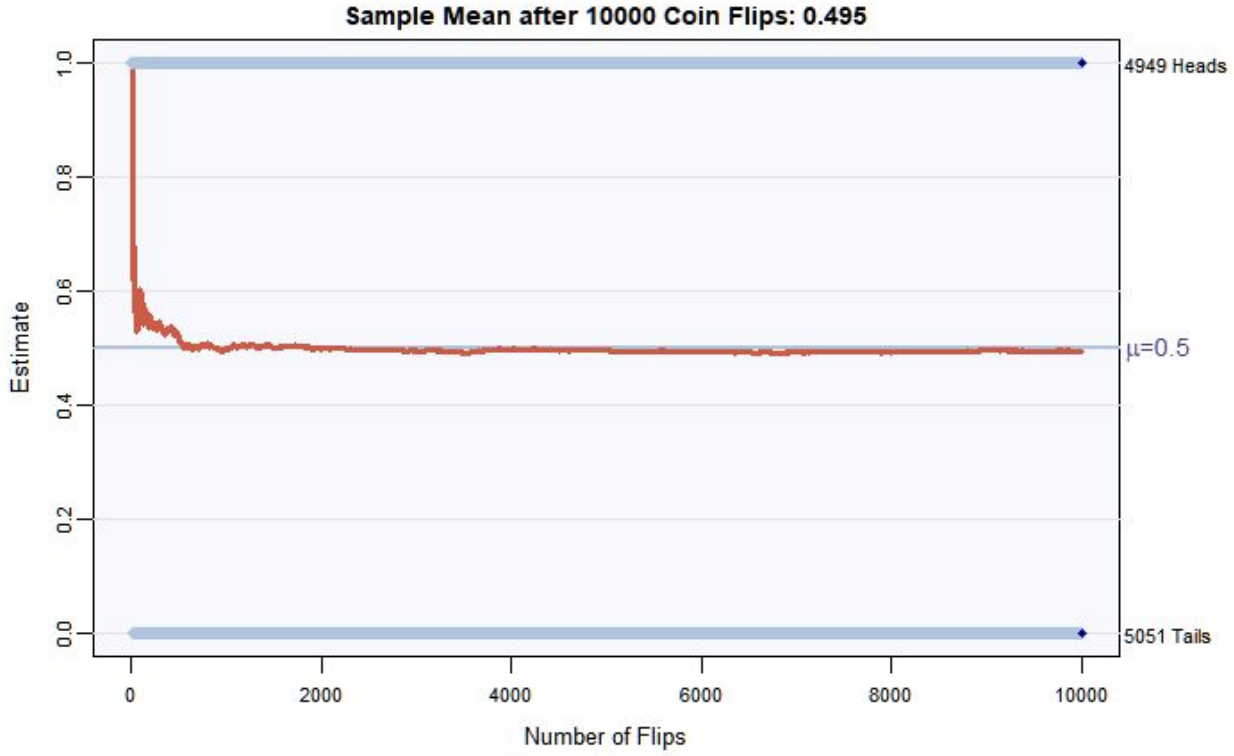


   ii. If you use the mean as the basis for the forecast (i.e., a stable process), what is the forecasted mean of another 15 flips?

Assuming a stable mean, the forecasted mean of another 15 flips is 0.533

b. Now simulate flipping a fair coin 10000 times.

i. Show the resulting visualization.



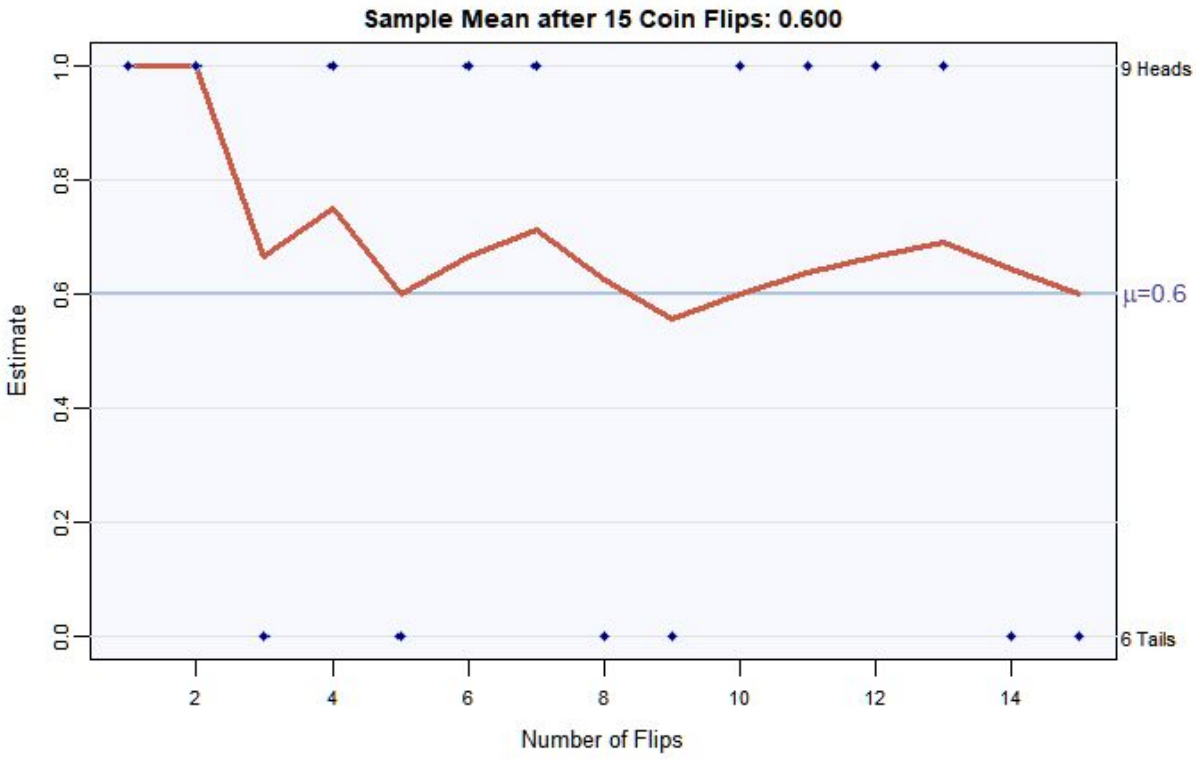
ii. Describe the visualization in terms of its overall shape?

This time the visualization resembles an initially squiggly line that converges toward the mean of 0.5.

iii. If you use the mean as the basis for the forecast (i.e., a stable process), what is the forecasted mean of another 15 flips?

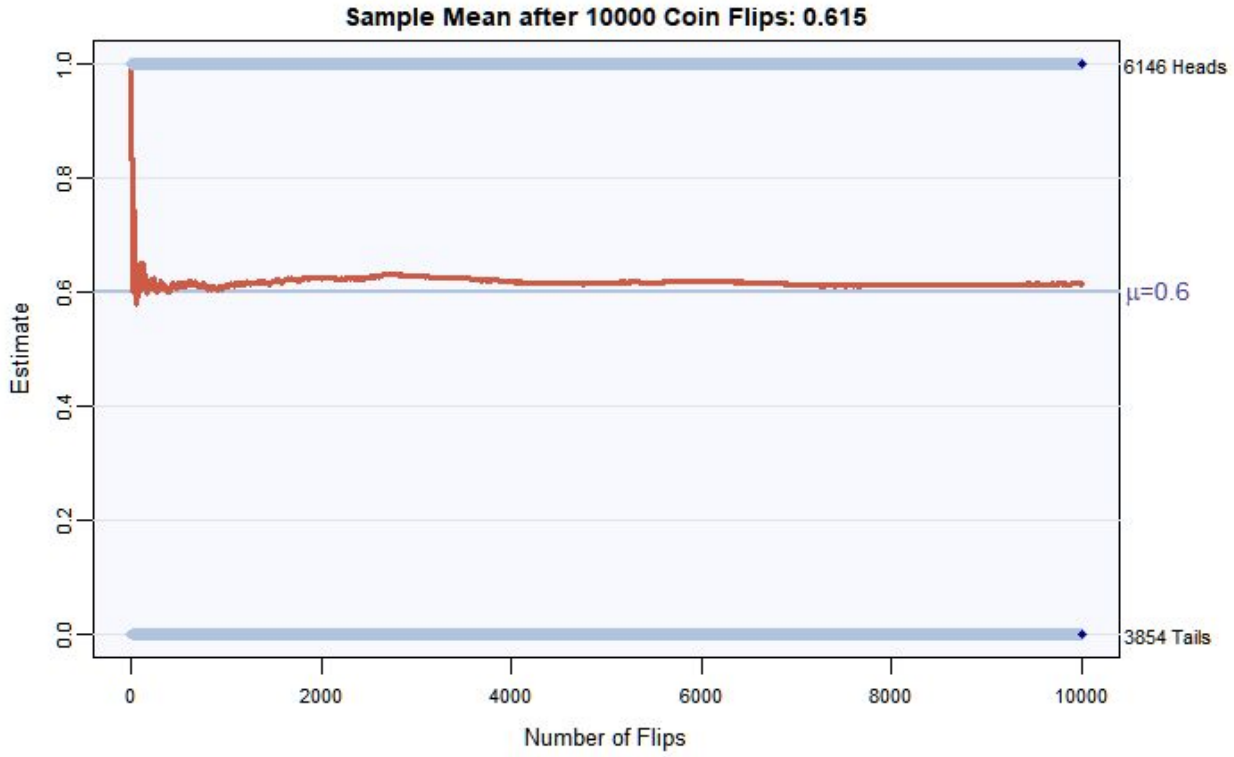
The answer is the same as for the previous simFlip() exercise (mean = 0.5) but with much more statistical significance and probability,

c. Now simulate flipping a biased coin, with a probability of a Head of 0.6.
    i. Show the resulting visualization of 15 flips? What is the forecasted mean? Can you detect the bias with 15 flips?



Now the forecasted mean is 0.6, with a slight bias toward landing on heads.

ii. Now flip the biased coin 10000 times? Can you detect bias now?



Here we clearly see the 0.6 bias as the line converges toward the expected mean of 0.6 with a measured mean of 0.615.

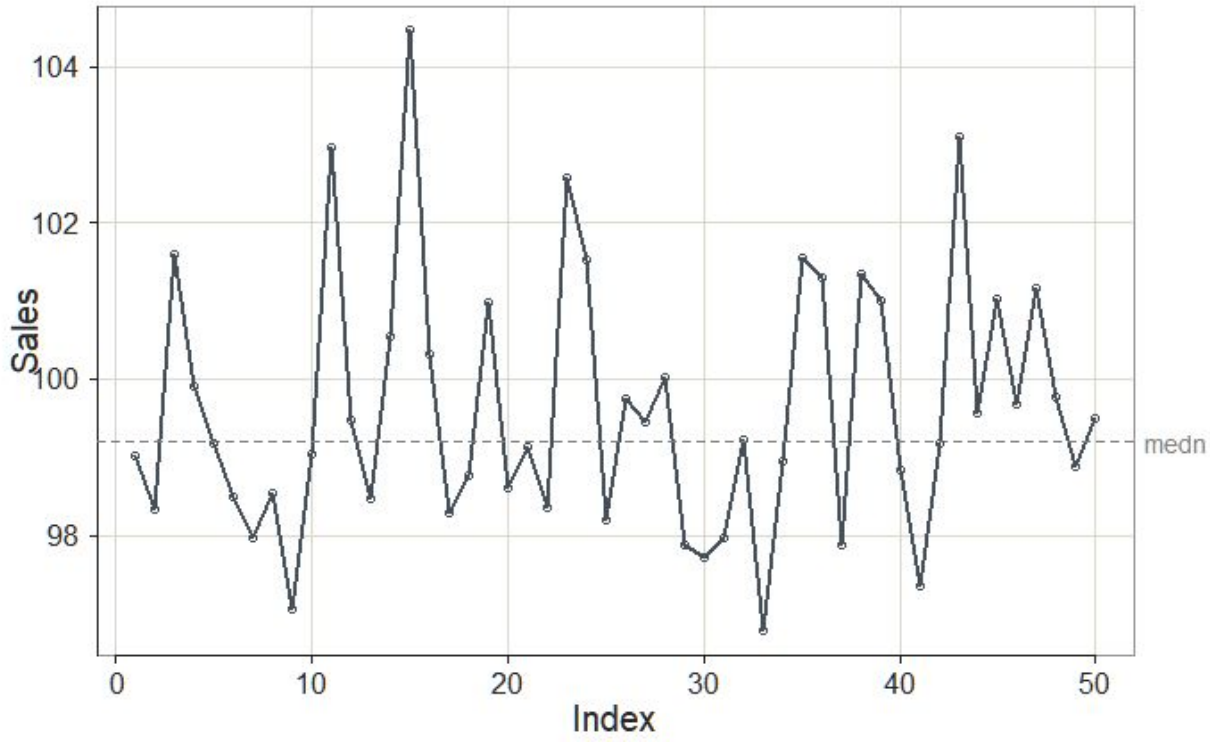
d. What do you conclude about the Law of Large Numbers in terms of forecasting?

Based on these data visualizations, I conclude that the Law of Large Numbers is quite apt in describing the tendency of a sufficiently large sample set to converge on the expected mean.

2. Run chart

Sales Data: http://web.pdx.edu/~gerbing/451/Data/HW2_2.xlsx

a. Plot and show the run chart.



Search for patterns:

b. Does the data appear to exhibit a trend? Why or why not?

The data does not appear to exhibit any trend. A trend can only become apparent with a long-term movement in a particular direction.

c. Does the data appear to exhibit seasonality? Why or why not?

The data might be exhibiting some seasonality, as at index points 10, 30 and 50 the data points are at relatively low points compared to surrounding data.

Forecast, based on your answers to b and c:

d. What is your forecast for Time #51? Why?

My forecast for time # 51 is 104. I see that after the low points there is often an intermediate point slightly higher and then a spike. I predict the time #51 will be the spike.

e. What is your forecast for Time #52? Why?

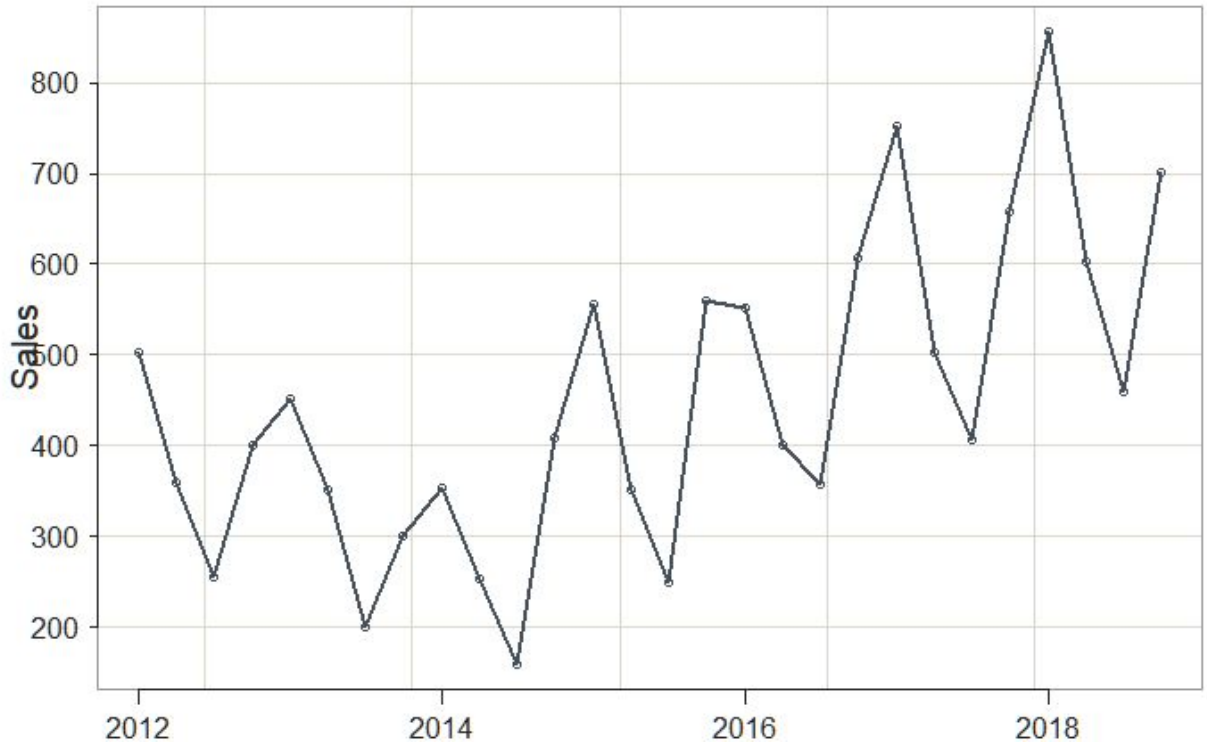
My forecast for time #52 is 101, back down toward but not yet down to the mean.

3. Time series visualization

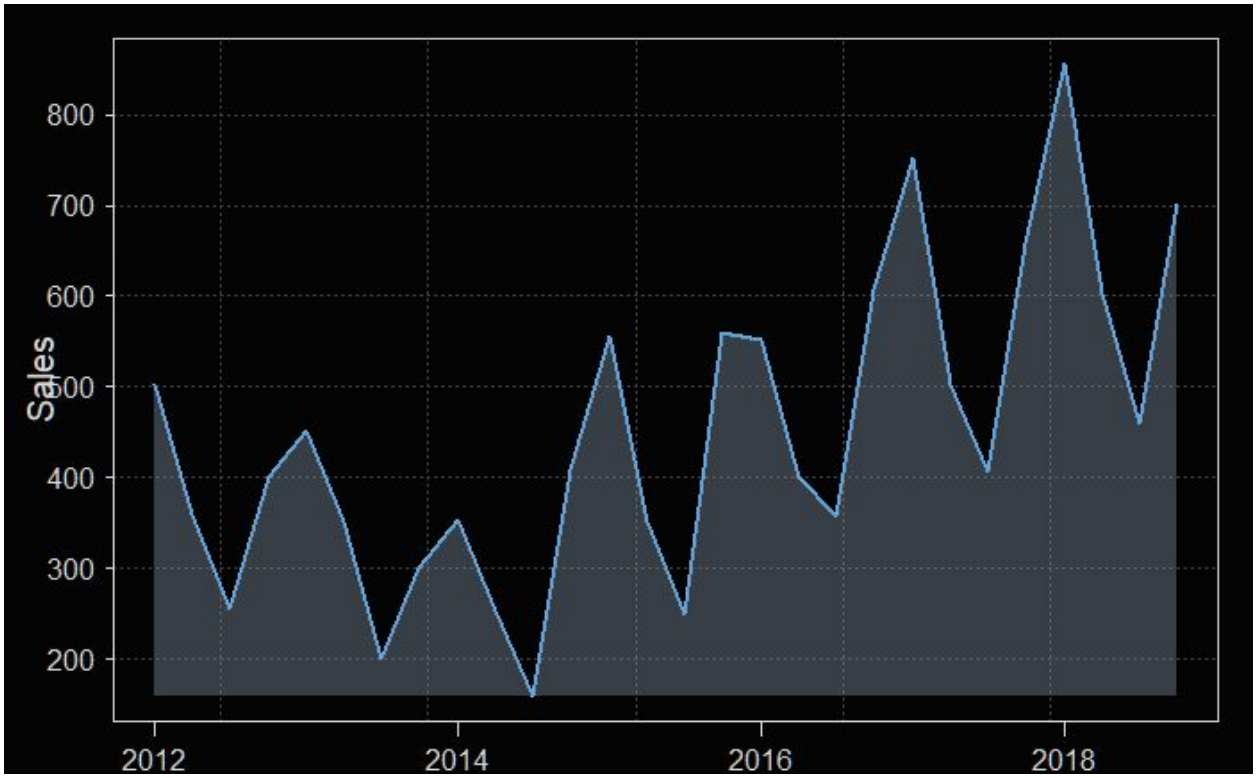
The following data file contains sales information for two products: Widgets and Openers.

Data: http://web.pdx.edu/~gerbing/451/Data/HW2_3.xlsx

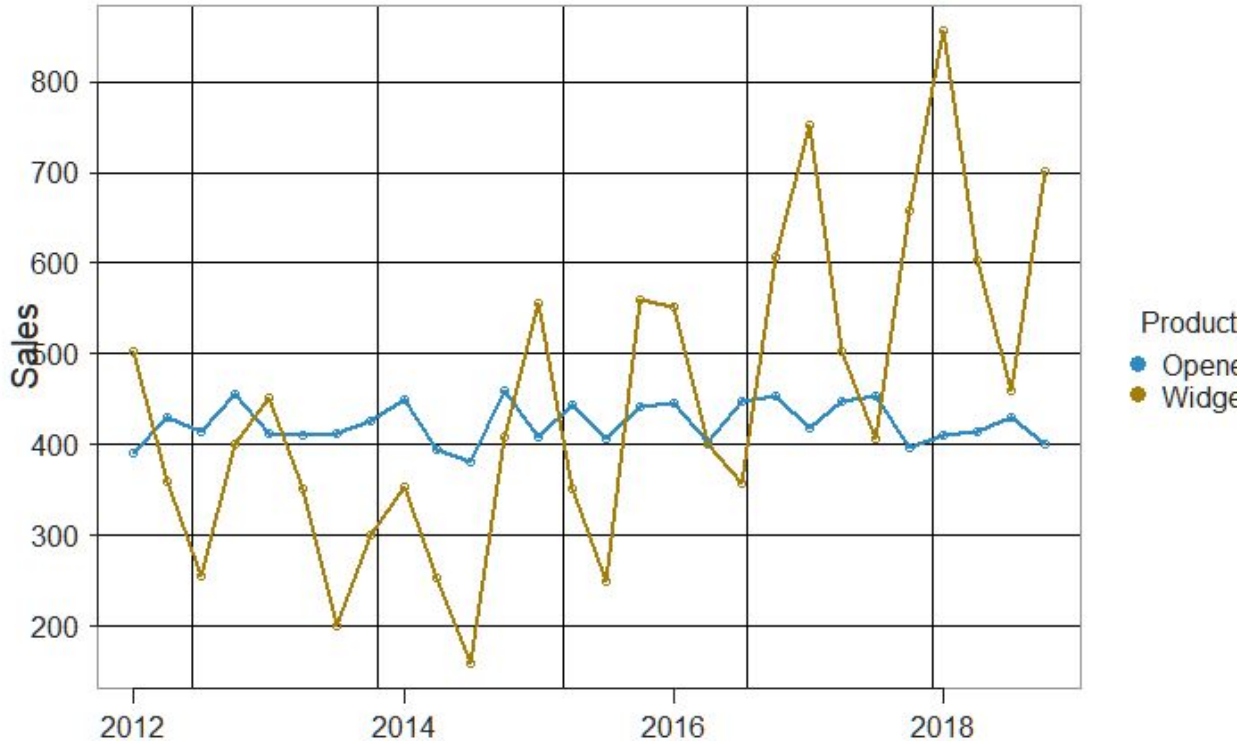
a. Plot the time series just for Widgets, default format.



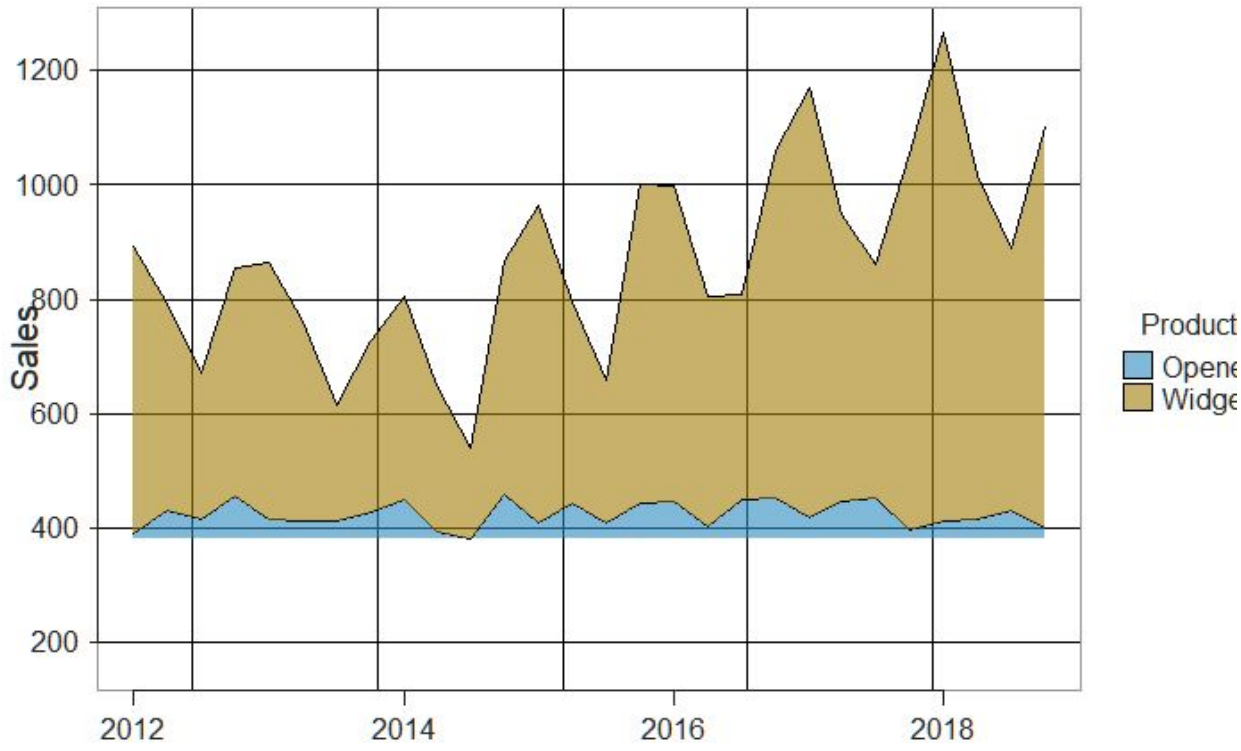
b. Plot the time series just for Widgets, WSJ (or similar color) format.



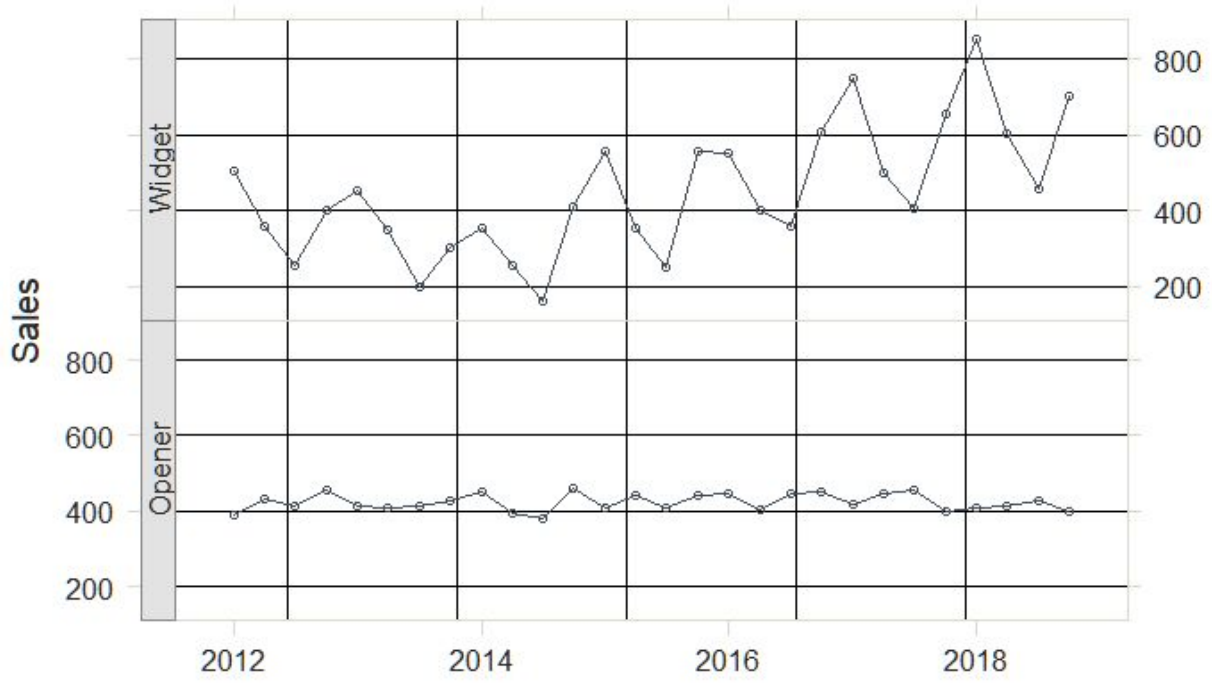
c. Plot the time series for Widgets and Openers on the same panel, unstacked.



d. Plot the time series for Widgets and Openers on the same panel, stacked.



e. Plot the time series for Widgets and Openers on two different panels, that is, a Trellis plot.



f. Forecast the next quarter's sales for Widgets and for Openers based on viewing the time series plots.

Based on the plots, I would forecast about 1300 in sales from Widgets, and about 450 in sales from Openers.

4. Time series visualization with Date conversion

The following data file contains sales information for two products: Widgets and Openers.

Data: http://web.pdx.edu/~gerbing/451/Data/HW2_3.csv

Note that now in the .csv file the date variable is stored as a character string, and so must be converted to an R Date type to plot the time series.

a. What is the conversion to obtain the date as an R Date type?

To convert a character string to an R Date type, the conversion is the as.Date() function.

b. Verify (show on your homework document) that the date variable is now a Date type.

```
> str(d$..Month)
Date[1:56], format: "2012-01-01" "2012-04-01" "2012-07-01" "2012-10-01" "2013-01-01"
"2013-04-01" …
```

c. Plot the time series just for Widgets, default format.