HW7

2024-10-27

1.

Consider the dataset txhousing which is in the ggplot2 R package. This dataset contains information about 46 Texas cities, recording the number of house sales (sales), the total volume of sales (volume), the average and median sale prices, the number of houses listed for sale (listing) and the number of months inventory (inventory). Data is recorded monthly from Jan 2000 to Apr 2015, 187 entries for each city.

(a)

Plot a time series of sales against date for each city on the same plot using ggplot2 package. Describe problems that make it hard to see the long-term trend in this plot.

(b)

We can fix one problem described in (a) by plotting sales in the log-scale. Plot a time series of sales in the log-scale against date for each city on the same plot using ggplot2.

(c)

We can fix another problem described in (a) by removing the strong seasonal trend present in the dataset. Fit a linear model with log(sales) against month for each city, then create a new column in txhousing, call it "rel_sales", which contains the residuals of the linear fits. Then plot a time series of rel sales against date for each city (on the same plot) using ggplot2

(d)

On the plot for (c), put in the mean curve using the function geom line() on the R code in (c).

(e)

Provide a page of plots of rel sales against month by year using the facet option on ggplot2 package. On each plot, put in the mean curve using geom line() options.

2.

When a cell phone relay tower is not working properly, wireless providers can lose great deal of money so it is important for them to be aware of and to be able to fix problems expeditiously. A first step toward understanding the problems involved is to collect data from designed experiments involving three factors, as was done in the experiment that gave rise to the dataset available in the file breakdown.dat. In this case, a reported problem was initially classified as low or high severity, simple or complex, and the engineer assigned was rated as relatively new (novice) or expert(guru). Two times were observed. The dataset has measurements on severity level of the problem (low/high, column 1), level of complexity of the problem (simple/complex, column 2), experience level of engineer (novice/guru, column 3), and time to assess problem (column 4), implement solution (column 5) and total resolution (column 6).

(a)

Read the data set and name each column, the variables corresponding names, problem, severity and engineer, i.e., problem for low and high, severity for simple and complex and engineer for new or expert. The names for the last three columns are given by "Assessment","Implementation","Resolution". Note that the resolution is the sum of the time assessment of the problem and the implementation.

(b)

For the engineer type as a categorical variable on the x axis, create a facet with boxplots for the variables. You may need to use the reshape2 package. Also rename the y-axis label. What difference can you tell for the engineer for each time?

(c)

The above boxplot may be hard to interpret for some. For the above, separate the levels severity and create side-by-side boxplots for each level of engineer. What conclusions can you draw from each of these plots?

(d)

Now create a new faceting plot, by faceting with "Problem" and the three variables. Discuss the plot and draw appropriate conclusions, if any.