FIT3152 Data analytics. Tutorial 03: Manipulating data

Pre-tutorial Activity

1 Read in the data from the file "Ped_Count_December_Long.csv" that we also used for the Week 2 pre-tutorial activities. Answer the following questions performing required data manipulations in R.

- Find the five sensor locations having the highest average pedestrian activity, and the five locations having the lowest average pedestrian activity over the month.
- Find the five one-hour periods having the highest average pedestrian activity and the five one-hour periods where pedestrian activity is at its lowest during the month.
- Find the hour and location at which the average pedestrian activity is at its lowest and the hour and location at which it is at its highest over the month.

Tutorial Activities

1. Try and reproduce the summary tables from Lecture 3.

Tips

2. (a) Using the 'iris' data set calculate the area of each sepal and petal using the following approximation: Area $\approx$ `Length * Width *` $\pi/4$. Create a new column for each area measurement.

(b) Draw a scatterplot showing petal area vs sepal area, identifying each species.

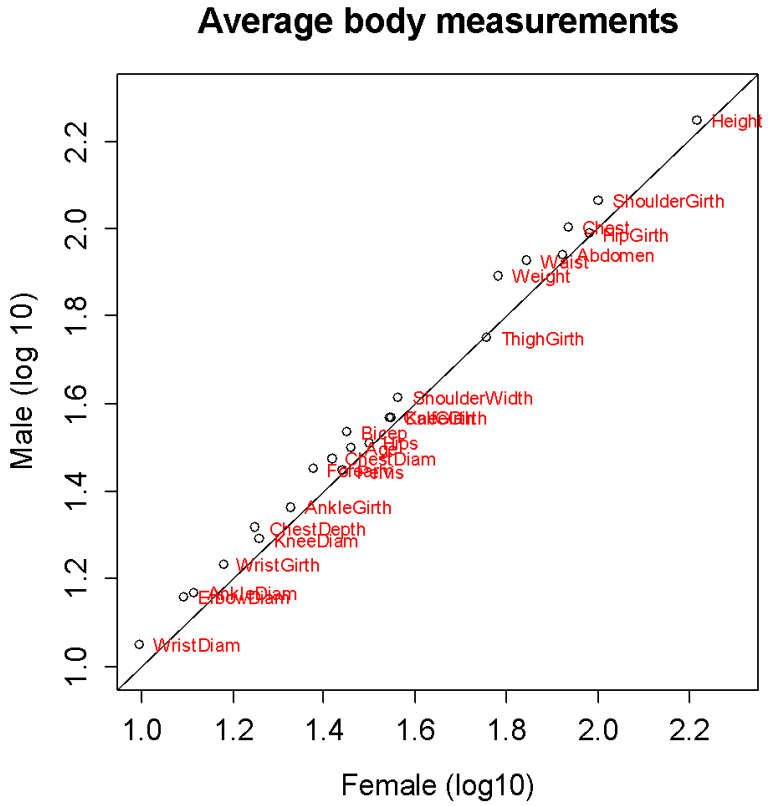
(c) Report the measurements of the plant in each species having (i) the largest petal and (ii) the largest sepal as a data frame. Your data frame should contain the original variables data along with the new columns showing sepal and petal area.

Tips

3. The file "body.dat.csv" contains data from a study on the relationship between body dimensions. The study measured 500+ active individuals.

The data was obtained from http://www.amstat.org/publications/jse/jse_data_archive.htm
A related article is http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html

Using the data draw the following plot using only R. The plot shows (on a log10 scale) the average measurements for males and females. Body parts have been identified, and an x = y line has been overplotted.



To construct the plot, I first created a data frame using a variety of functions to: aggregate and calculate the mean; transpose the data (from rows to columns); assign column names; remove rows and convert class of the data to numeric. You will also need to investigate plotting with the base graphics, (? plot) and parameter (? par) settings.

Record the commands required to produce the plot. Make any improvements you can.

Tips

4. The data file “Dunnhumby1-20.csv” is a cut down and modified set of test data from the Kaggle competition to predict when consumers would next visit a Dunnhumby supermarket and how much they would spend. See: http://www.kaggle.com/c/dunnhumbychallenge for more information. The current modified data set contains the customer ID, Date of visit, Date since last visit, and Spend for 20 customers from the test set.

   Summarize the following information for each customer in a similar format to table below. Create a data frame showing: Customer ID; average time between visits (Delta); average spent each time; correlation between delta and spend; the number of observations. *As a challenge also try and report the slope and intercept of the least squares regression of spend vs delta.*

CustomerID	AveDelta	AveSpend	CorDeltaSpend	N	RegSlope	RegInt
40	4.86	36.71	-0.003	73	-0.030	36.745
…						
…						

Tips

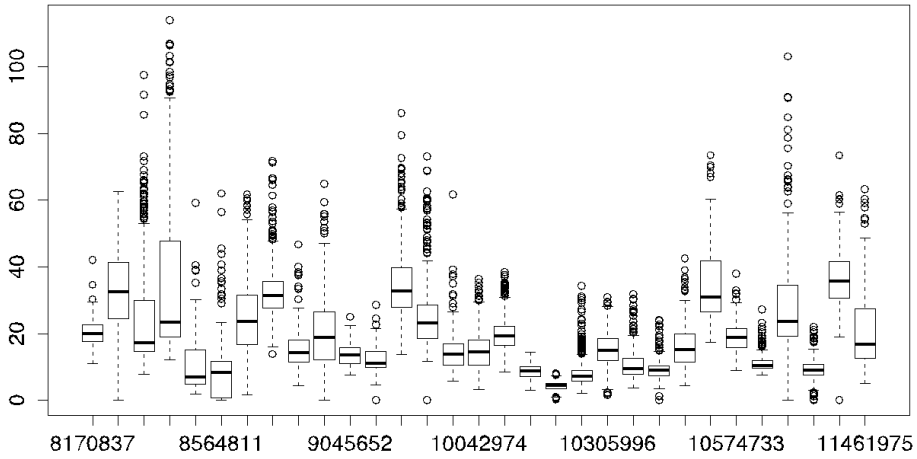
5. The data file “govhackelectricitytimeofusedataset.csv” has been created from the .txt file originally available as part of the Australian Government’s data resources. See link at: https://data.gov.au/dataset/sample-household-electricity-time-of-use-data. The file contains the smart meter records for a number of households recorded at 30 minute intervals over varying periods of time. The first few rows of the csv file are below.

CUSTOMER_KEY	End Datetime	General Supply KWH	Off Peak KWH	Gross Generation KW	Net Generation KWH
8170837	4/04/2013 11:59	0.137	0	0	0
8170837	4/04/2013 12:29	0.197	0	0	0
8170837	4/04/2013 12:59	0.296	0	0	0
8170837	4/04/2013 13:29	0.24	0	0	0
8170837	4/04/2013 13:59	0.253	0	0	0
8170837	4/04/2013 14:29	0.24	0	0	0
8170837	4/04/2013 14:59	0.238	0	0	0
8170837	4/04/2013 15:29	0.225	0	0	0
8170837	4/04/2013 15:59	0.246	0	0	0

The columns of interest are “Customer_Key” (meter), “End Datetime”, and “General SupplyKWH” (power used each 30 mins).

Using the 30 minute general supply, calculate the daily supply for each meter for every day there is data available. Because the number of records is unreliable you will also need to count the number of daily observations for each (day, meter). You should then discard any (day, meter) readings that do not have the complete number of observations.

Once you have this list you should be able to draw a box plot showing daily power consumption for each of the households. Your boxplot should look something like this:



Some functions you might try are: read.csv, by, length, sum, cbind, colnames, subset.

Extension, calculate the minimum, average and maximum daily consumption for each customer for each month you have data.

Tips