HW3_Wei_Yanran

Yanran Wei

September 18, 2017

Problem 4

The two links describe good programming style that makes R code easier to read, share and verify. It set up a kind of fundamental standard for R users to follow to create readable R code. In these two links, rules like naming, notation, syntax, functions and organizations are introduced. For me, the way to improve my coding style is to follow above rules and practice. For example, I will name variables in similar format, leave space around all binary operators and write comments if necessary. Then my code will be more readable to other R users.

Problem 5

```
library(lintr)
lint(filename = "./02_data_munging_summarizing_R_git/HW2_Wei_Yanran.Rmd")
```

There are hundreds of suggestions on my code. The most common suggestion include *Put spaces around all infix operators*, *Commas should always have a space after* and *Variable and function names should be all lowercase*. Other suggestions are *Trailing whitespace is superfluous*, *lines should not be more than 80 characters* and *Only use double-quotes*. These are all the bad programming syles I am gonna to improve in my homework.

Problem 6

a

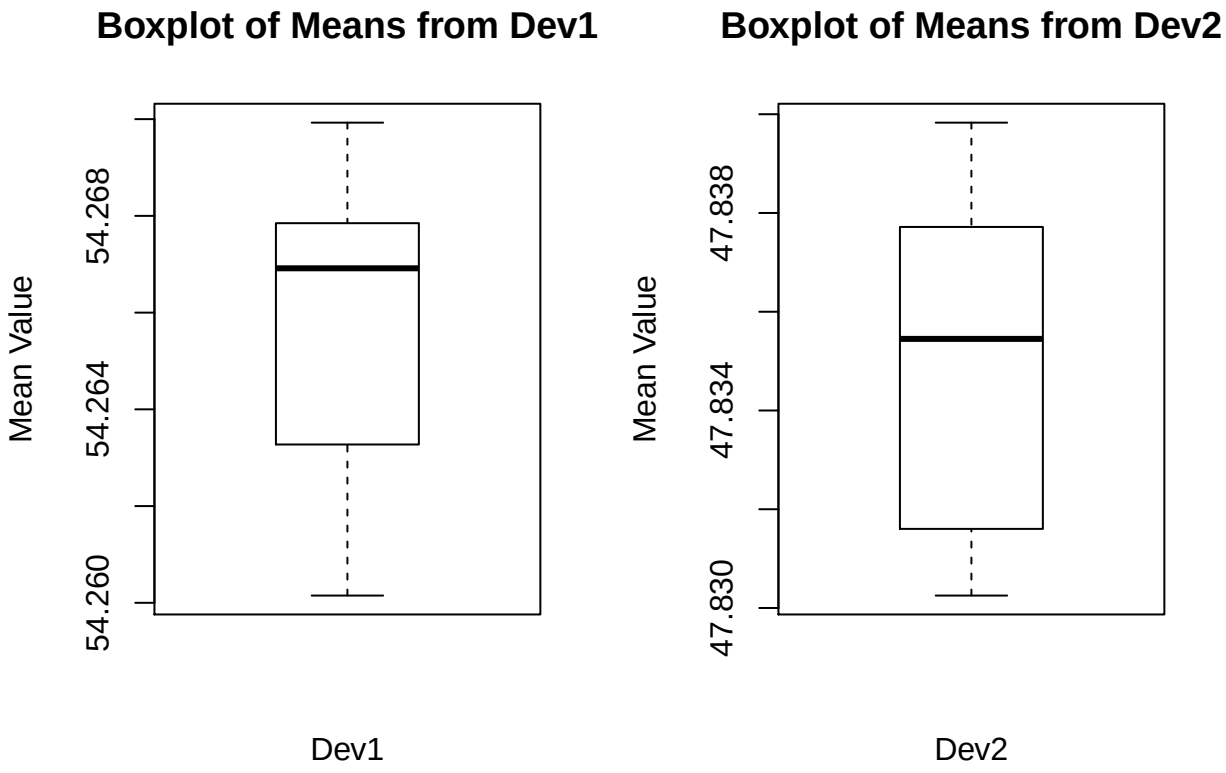
Below is the summary of 13 observers with mean, standard deviation and correlation for each observer.

Table 1: Summary of 13 Observers

Observers	mean1	mean2	sd1	sd2	correlation
1	54.26610	47.83472	16.76983	26.93974	-0.0641284
2	54.26873	47.83082	16.76924	26.93573	-0.0685864
3	54.26732	47.83772	16.76001	26.93004	-0.0683434
4	54.26327	47.83225	16.76514	26.93540	-0.0644719
5	54.26030	47.83983	16.76774	26.93019	-0.0603414
6	54.26144	47.83025	16.76590	26.93988	-0.0617148
7	54.26881	47.83545	16.76670	26.94000	-0.0685042
8	54.26785	47.83590	16.76676	26.93610	-0.0689797
9	54.26588	47.83150	16.76885	26.93861	-0.0686092
10	54.26734	47.83955	16.76896	26.93027	-0.0629611
11	54.26993	47.83699	16.76996	26.93768	-0.0694456
12	54.26692	47.83160	16.77000	26.93790	-0.0665752
13	54.26015	47.83972	16.76996	26.93000	-0.0655833

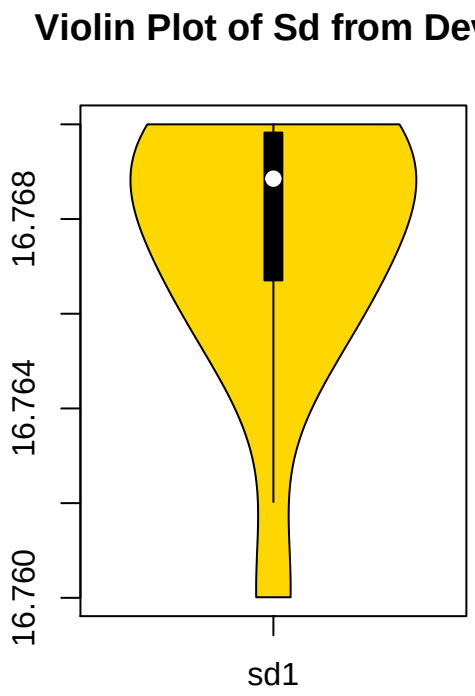
b

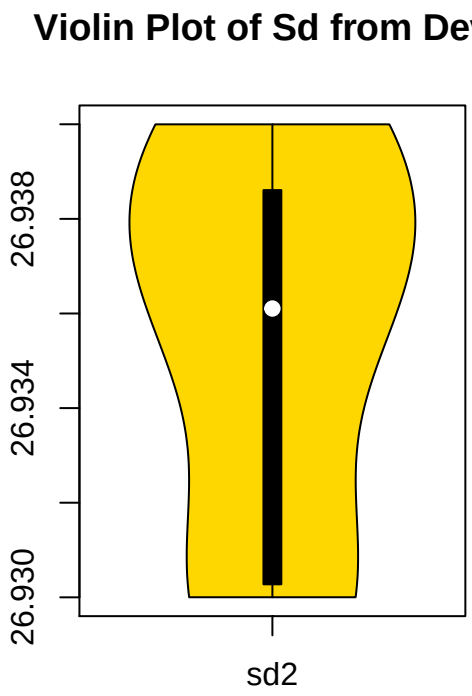
Below is the boxplot of means of 13 observers. The mean value of dv1 is larger than that of dv2.



c

Below is the violin plot of sd of 13 observers.





Problem 7

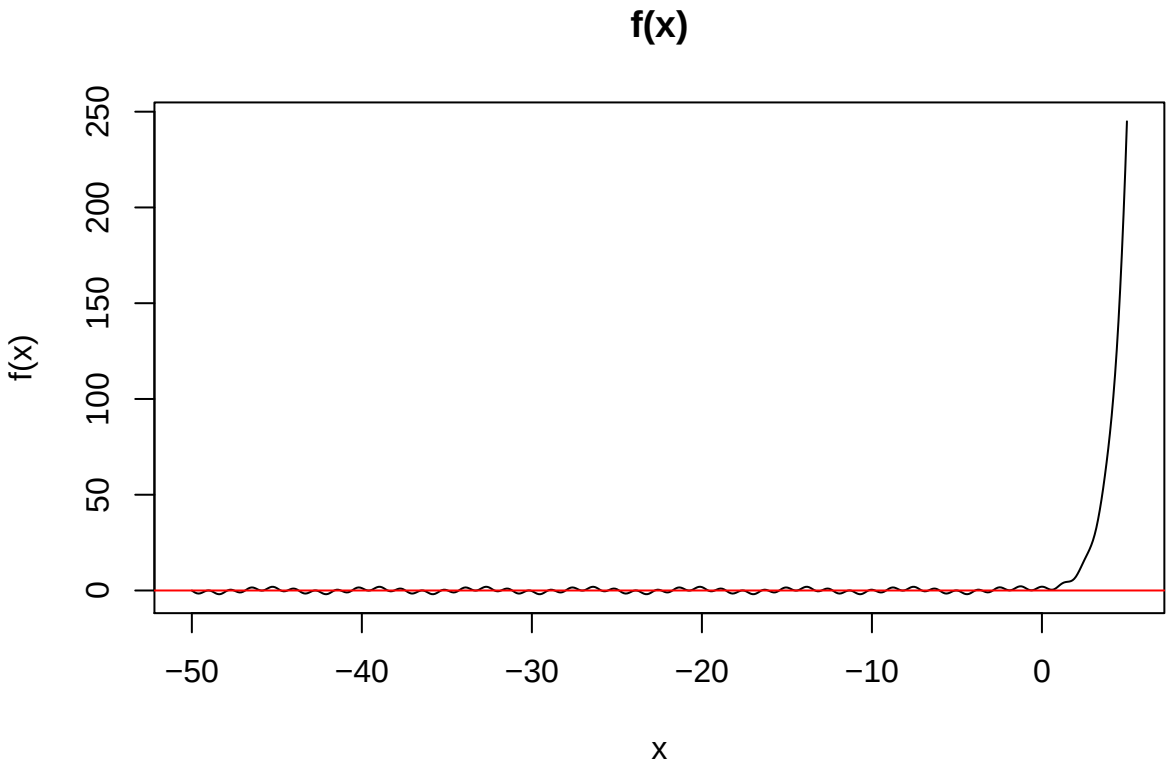
Submmary table after tidying is shown as below.

Table 2: Blood Pressure Data Summary

Day	methods	replicate	value
Length:90	Length:90	Length:90	Min. :110.8
Class :character	Class :character	Class :character	1st Qu.:125.5
Mode :character	Mode :character	Mode :character	Median :130.4
NA	NA	NA	Mean :129.0
NA	NA	NA	3rd Qu.:134.3
NA	NA	NA	Max. :139.6

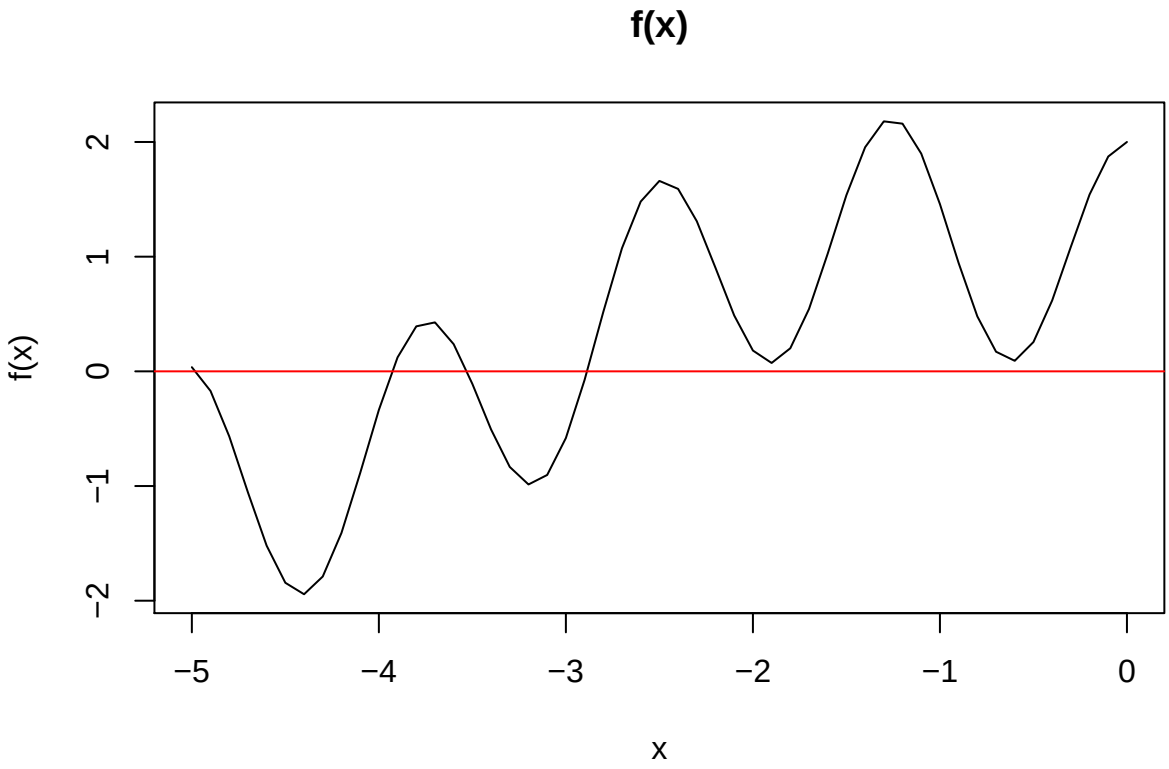
Problem 8

From the table above, we can observe that the value of the function is positive after 5 due to the large positive value of 3^x.



So we shrink the dependent variable to the data range of (-50,0). Now the plot of the function is as below. Because the value of 3ˆx becomes small as value of x decreases, the function is mainly influenced by sin(x) and cos(5x) which lead to periodically fluctuation.

To simplify the question, we only observe one 'period' which correspond to x between the range of (-5,0).




```
## [1] -2.864494 -2.886821 -2.887058
## [1] -3.528046 -3.528722
## [1] -3.935649 -3.930160 -3.930113
## [1] -4.961804 -4.970864 -4.971500
```

In the range of x between -5 and 0, ther are four solutions, -2.8870, -3.5287, -3.9301 and -4.9715.

Problem 9

a. Load databasets

Load table using function fread.

b. Merge three tables

Merge tables using merge function. 1987553 observations in the merged table.

c. Clean the data and remove NA

125122 observations remained after removing NA.

d. How many DIFFERENT makes and models of cars

35 makes and 385 models.

e. 5 most frequent defects and make/models

Select columns we need utilizaing sqldf funcion.

Table 3: 5 most frequetn defects

defect	make	count
K04	PEUGEOT	105
AC1	PEUGEOT	54
RA2	PEUGEOT	40
J03	OPEL	33
G05	SUZUKI	32

f. Relationship between number of defects and make

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.835	5.187	1.511	0.1365
make	0.1868	0.4078	0.458	0.6487

Table 5: Fitting linear model: count ~ make

Observations	Residual Std. Error	R^2	Adjusted R^2
58	16.87	0.003732	-0.01406

Table 6: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
make	1	59.68	59.68	0.2098	0.6487
Residuals	56	15931	284.5	NA	NA

g. Relationship between number of defects and model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.28	4.08	4.481	3.719e-05
model	-0.3211	0.135	-2.378	0.02083

Table 8: Fitting linear model: count ~ model

Observations	Residual Std. Error	R^2	Adjusted R^2
58	16.1	0.09174	0.07552

Observations	Residual Std. Error	R^2	Adjusted R^2

Table 9: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
model	1	1467	1467	5.656	0.02083
Residuals	56	14524	259.4	NA	NA