

# Statistics 4004: Homework 3

Due Thursday Feb 7, start of class

*2019-01-31*

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Reproducible Research, R and version control, getting, cleaning and munging data and finally, summarizing data. Again, we are focusing on Reproducible Analysis which, for us, is accomplished by mixing code, figures and text into a cohesive document that fully describes both the process we took to go from data to results and the rationale behind our data driven conclusions. This week we begin creating tidy data sets. While others have proposed standards for sharing data with statisticians, as practicing data scientists, we realize the often onerous task of getting, cleaning and formatting data is usually in our hands. Again, we will use BitBucket to turn in and retrieve the homework assignments.

## Problem 1

Work through the “Getting\_and\_Cleaning\_Data” *swirl* lesson part 3 and “Exploratory\_Data\_Analysis” *swirl* lesson parts 2, 5 and 7.

From the R command prompt:

```
install.packages("swirl")  
library(swirl)  
install_course("Getting_and_Cleaning_Data")  
swirl()
```

*Nothing to turn in*

## Problem 2

Read through the Git help Chapters 1 and 2. <https://git-scm.com/book/en/v2>

Just good stuff to know, you will NOT be tested on it but may find it helpful if you want a little more Git.

*Nothing to turn in*

## Problem 3

Create a new R Markdown file (file->new->R Markdown; author = you, choose PDF).

The filename should be: HW3\_pid, i.e. for me it would be HW3\_rsettag.

You will use this new R Markdown file to solve problems 4-7.

## Problem 4 [1 point]

Scenario: You are given a dataset and being a good data scientist, import, munge and summarize the data. The summary stats are given in Table 1:

Table 1: summary of Anscombe dataset

	x1	x2	x3	x4	y1	y2	y3	y4
mean	9.00	9.00	9.00	9.00	7.50	7.50	7.50	7.50
sd	3.32	3.32	3.32	3.32	2.03	2.03	2.03	2.03

## Part A.

What are your initial thoughts?

You proceed with creating a linear model as per your collaborators requirements and get the following coefficients:

Table 2: linear model for x1 vs y1, x2 vs y2, etc in Anscombe data

	lm1	lm2	lm3	lm4
(Intercept)	3.0000909	3.000909	3.0024545	3.0017273
x1	0.5000909	0.500000	0.4997273	0.4999091

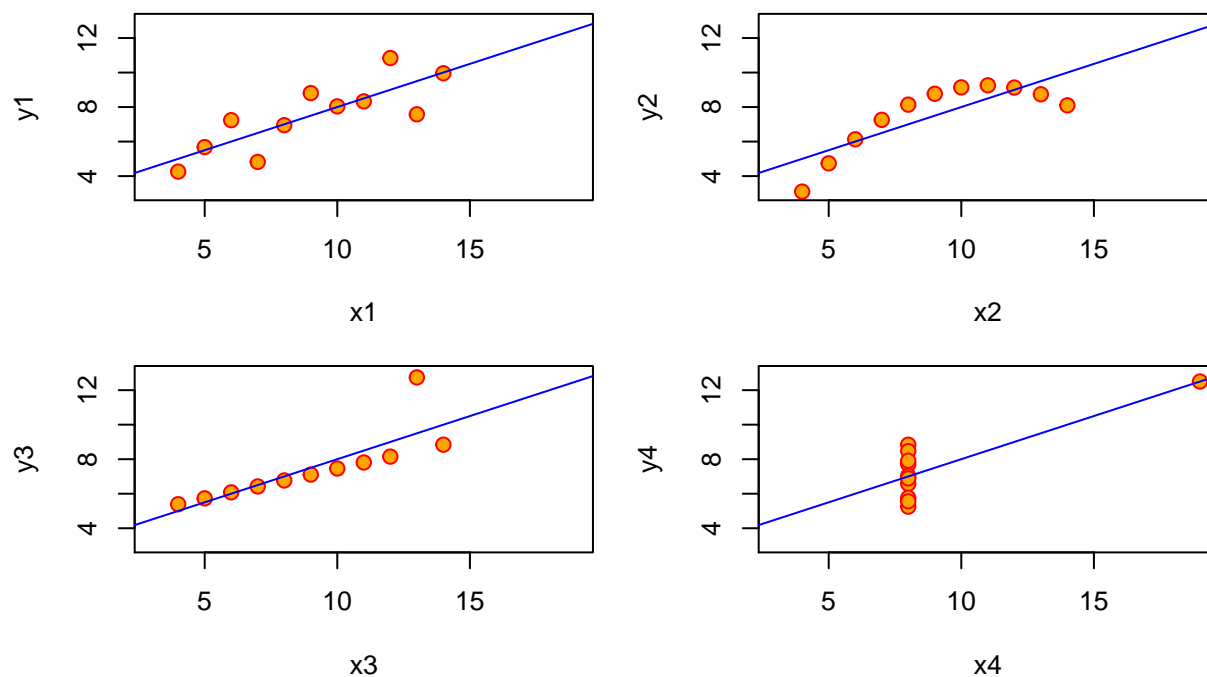
## Part B.

What are your thoughts after seeing the regression coefficient output?

## Part C.

You then plot the data to obtain the result shown in Figure 1 below. What is the lesson here?

### Linear model for Anscombe's data



## Problem 5

In these exercises, you will import, munge, clean and summarize datasets from Wu and Hamada's *Experiments: Planning, Design and Analysis* book used in the graduate experiment design class. For each dataset, please weave your code and text to describe both your process and observations. Make sure you create a tidy dataset describing the variables, create a summary table of the data, note any issues with the data, and create a single plot of each dataset highlighting some feature of the data. For full credit, you will use the dplyr and tidyr packages for munging and summarizing the data. Remember to label your plots including axis and any necessary legends.

### Part A [3 points]

Ten parts were selected randomly from a line and duplicate measurements of the part's wall thickness were taken by each of three operators of the measurement apparatus.

<https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/ThicknessGauge.dat>

### Part B [3 points]

Larvae counts at two ages given 5 different treatments in 8 blocks.

<https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LarvaeControl.dat>

## Problem 6 [1 point]

Please knit this document to PDF (name should be HW3\_pid) and push to BitBucket:

In the R Terminal, type:

1. git pull
2. git add HW3\_pid.[pR]\* (NOTE: this should add two files)
3. git commit -m "final HW3 submission"
4. git push

## Optional Extra Credit [1 point]

Tomato yield data. Two tomato varieties were planted at 3 different densities. The experiment was run in triplicate. I have created a tidy dataset using tidyr and dplyr. Your job is to duplicate this using base R functions.

The finished product is given in Table 1 below.

Table 3: Tomato data summary

Clone	Replicate	value	Variety
Length:18	Length:18	Length:18	Length:18
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

## Grading Rubric:

- 1 point: successfully submitted to BitBucket
- 2 points: Neat, well written document
- 1 point: reasonable response to question 4
- 6 points: correct answers to problem 5 as given
- 1 bonus point