# Homework 3

*Akshay Kumar Varanasi (av32826)*

**This homework is due on Feb. 12, 2019 at 4:00pm. Please submit as a PDF file on Canvas.**

In this homework, you are asked to evaluate two data sets and determine if they are tidy data sets. *We are referring to a very specific definition of "tidy", so if this term is unfamiliar to you, please review the lecture materials*.

**Problem 1: (2 pts)** The dataset `USAccDeaths` built into R contains accidental deaths in the US 1973-1978. You can run `?USAccDeaths` to learn more about this data set.

```
USAccDeaths
```

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov
## 1973  9007  8106  8928  9137 10017 10826 11317 10744  9713  9938  9161
## 1974  7750  6981  8038  8422  8714  9512 10120  9823  8743  9129  8710
## 1975  8162  7306  8124  7870  9387  9556 10093  9620  8285  8466  8160
## 1976  7717  7461  7767  7925  8623  8945 10078  9179  8037  8488  7874
## 1977  7792  6957  7726  8106  8890  9299 10625  9302  8314  8850  8265
## 1978  7836  6892  7791  8192  9115  9434 10484  9827  9110  9070  8633
##        Dec
## 1973  8927
## 1974  8680
## 1975  8034
## 1976  8647
## 1977  8796
## 1978  9240
```

Explain the variables present in this dataset. Using the variables in this dataset and the formal definition of tidy data that we learned in lecture, is this data set tidy? Explain why or why not.

Variables present in this dataset are Year, Month and Number. So instead of columns having name of the month and row names as years, there should be a column for months and a column for year. Observations should be death of each person. Thus, it is not a tidy dataset.

The dataset `CO2` built into R contains data on the carbon dioxide uptake in grass plants. You can run `?CO2` to learn more about this data set.

```
head(CO2)
```

```
##   Plant   Type  Treatment conc uptake
## 1   Qn1 Quebec nonchilled   95   16.0
## 2   Qn1 Quebec nonchilled  175   30.4
## 3   Qn1 Quebec nonchilled  250   34.8
## 4   Qn1 Quebec nonchilled  350   37.2
## 5   Qn1 Quebec nonchilled  500   35.3
## 6   Qn1 Quebec nonchilled  675   39.2
```

Explain the variables present in this dataset. Using the variables in this dataset and the formal definition of tidy data that we learned in lecture, is this data set tidy? Explain why or why not.

Variables present in this are Plant, Type, Treatment, concentration and uptake. Each observation is of a plant. Therefore it satisfies the rules of tidy dataset.

**Problem 2: (5 pts)** Listed below are three examples of code that violate the rules in section 2 (https://style.tidyverse.org/syntax.html) of the tidyverse style guide. Which tidyverse style guidelines are violated in these example?

```
iris %>% filter(Species=="versicolor") %>% head()
```

Violates 2.2.3 Infix operators. The operator == should be surrounded by spaces.

```
iris[50,]
```

It violates 2.2.1 Commas guideline. It is always good to put a space after a comma, never before, just like in regular English.

```
boxplot (len ~ dose, data = ToothGrowth, range = 1, width = c(2, 2, 2), varwidt
h = TRUE, notch = FALSE, outline = TRUE)
```

Here it violates 2.2.2 Parentheses. It is not good to put spaces inside or outside parentheses for regular function calls.

**Problem 3: (3 pts)** The `NCbirths` contains 1409 birth records from North Carolina in 2001. The column contents are as follows:

- **Plural**: 1=single birth, 2=twins, 3=triplets.
- **Sex**: Sex of the baby 1=male 2=female.
- **MomAge**: Mother’s age (in years).
- **Weeks**: Completed weeks of gestation.
- **Gained**: Weight gained during pregnancy (in pounds).
- **BirthWeightGm**: Birth weight in grams.
- **Low**: Indicator for low birth weight, 1=2500 grams or less, 0=otherwise.
- **Premie**: Indicator for premature birth, 1=36 weeks or sooner, 0=otherwise.
- **Marital**: Marital status: 0=married or 1=not married.

```
NCbirths <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/NCbirths.csv
")
head(NCbirths)
```

```
##   Plural Sex MomAge Weeks Gained Smoke BirthWeightGm Low Premie Marital
## 1      1   1     32    40     38     0       3146.85   0      0       0
## 2      1   2     32    37     34     0       3288.60   0      0       0
## 3      1   1     27    39     12     0       3912.30   0      0       0
## 4      1   1     27    39     15     0       3855.60   0      0       0
## 5      1   1     25    39     32     0       3430.35   0      0       0
## 6      1   1     28    43     32     0       3316.95   0      0       0
```

For single births, what are the **max** completed weeks of gestation and the **mean** birth weight for babies that were born prematurely and for babies that were carried to term? State your answer in a sentence. **HINT:** Use the function `max()` to determine the maximum completed weeks of gestation.

```
NCbirths %>% filter(Plural == "1") %>% group_by(Premie) %>% summarize(max = max
(Weeks), mean = mean(BirthWeightGm))
```

```
## # A tibble: 2 x 3
##   Premie   max  mean
##    <int> <dbl> <dbl>
## 1      0    45 3431.
## 2      1    36 2616.
```

For single birth babies, the max completed weeks of gestation and mean birth weight for babies is 45 weeks and 3430.724 gms in case of babies carried to term and 36 weeks and 2615.760 gms in case of premature birth.