Homework 2 STAT 5014

Shane Bookhultz

September 11, 2017

Problem 4

Version control can be very useful in the classroom by letting me revert back to older versions of code and allowing me to experiment with new features. Additionally, using version control, I can share code with my classmates without having to physically be there, and we can work on different versions of our code. Lastly, version control can let me undo any changes I make so instead of making a huge mistake without VC, I can revert my mistakes.

Problem 5

a. Sensory data from five operators

Table 1: Brain and body weight data summary

Operator 1	Operator 2	Operator 3	Operator 4	Operator 5
Min. :0.900	Min. :1.500	Min. :0.800	Min. :0.900	Min. :0.700
1st Qu.:2.850	1st Qu.:3.450	1st Qu.:2.650	1st Qu.:3.925	1st Qu.:2.250
Median :4.550	Median :4.950	Median :4.150	Median :5.400	Median :4.600
Mean :4.593	Mean :5.063	Mean :4.167	Mean :5.193	Mean :4.267
3rd Qu.:5.950	3rd Qu.:6.225	3rd Qu.:5.400	3rd Qu.:6.275	3rd Qu.:5.800
Max. :9.000	Max. :9.200	Max. :9.000	Max. :9.400	Max. :8.800

For this dataset, I removed the first two rows in the original dataset as both rows had either "Operator" in them or a non-descriptive header. From there, I converted the Item column from a factor into a numeric variable to make further cleaning easier. Next, I moved over rows that didn't have an item number over 1 column, and then correctly renumbered the row. Lastly, I implemented an ID variable to identify each result, and moved the rows so the ID was in front, being the identifing variable.

Some issues with the data include missing data, incorrect data in columns, and a mixture of number and character headers.

b. Gold Medal performance for Olympic Men's Long Jump

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Table 2: Long Jump data summary

Year	Long Jump
Min. :-4.00	Min. :249.8
1st Qu.:21.00	1st Qu.:295.4
Median :50.00	Median :308.1
Mean :45.45	Mean :310.3
3rd Qu.:71.00	3rd Qu.:327.5
Max. :92.00	Max. :350.5
NA's :2	NA's :2

In this Gold Medal dataset, I had seperate the columns of the data, as the variables year and long jump were repeated horizontally. I also had to convert the variables to numeric so I could create a summary table. After I seperated the columns, I then full joined them so I could create a dataset with only 2 variables.

One issue with the data was multiple repeated variables.

c. Brain weight (g) and body weight (kg) for 62 species.

Table 3: Brain and body weight data summary

Body Weight	Brain Weight
Min. : 0.005	Min. : 0.10
1st Qu.: 0.600	1st Qu.: 4.25
Median : 3.342	Median : 17.25
Mean : 198.790	Mean : 283.13
3rd Qu.: 48.203	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00

In the Brain and body weight dataset, the main issue was again, like the last one, was that the variables were seperated into 3 columns of the same variable. So again I seperated the columns and then binded the rows to create a dataset of 2 variables.

Main issue was repeated variables

d. Triplicate measurements of tomato yield for two varieties of tomatoes at three planting densities

```
## [1] "character"
```

Table 4: Brain and body weight data summary

Tomato_Species	10000_density	20000_density	30000_density
Ife#1 :3	10.1:1	11.5:1	13.7:1
PusaEarlyDwarf:3	15.3:1	12.7:1	14.4:1
NA	16.1:1	13.7:1	15.4:1
NA	17.5:1	16.6:1	18 :1
NA	8.1 :1	18.5:1	20.8:1
NA	8.6 :1	19.2:1	21 :1

This dataset was the trickiest in terms of data manipulation. The problems with this dataset were creating consistent density columns, creating a tomato species column, and splitting up the comma seperated cells. I took apart the comma seperated cells, and put them into a numeric vector. I created a tomato matrix and then iterating through a for loop I put each of the seperate components of the numeric vectors and put them into the tomato matrix, corresponding the correct species and density.

The issues with this dataset were correcting rows and columns, multiple values in a single cell, and indentifying unique cases.

Problem 6

Find a plant dataset

```
##
## | Hi! I see that you have some variables saved in your workspace. To keep
## | things running smoothly, I recommend you clean up before starting swirl.
##
## | Type ls() to see a list of the variables in your workspace. Then, type
## | rm(list=ls()) to clear your workspace.
##
## | Type swirl() when you are ready to begin.
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
##
## =====================================================
##                               Dependent variable:
##                           ---------------------------
##                                    pH_Median
## -----------------------------------------------------
## Foliage_ColorGray-Green            0.413***
##                                     (0.123)
##
## Foliage_ColorGreen                 0.185***
##                                     (0.063)
##
## Foliage_ColorRed                     0.163
##                                     (0.276)
##
## Foliage_ColorWhite-Gray             0.445**
##                                     (0.189)
##
## Foliage_ColorYellow-Green           -0.062
##                                     (0.134)
##
## Constant                           5.999***
##                                     (0.060)
##
## -----------------------------------------------------
## Observations                          832
## R2                                   0.023
## Adjusted R2                          0.017
## Residual Std. Error            0.539 (df = 826)
## F Statistic                 3.958*** (df = 5; 826)
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

```
## [1] "                           Df Sum Sq Mean Sq F value  Pr(>F)   "
## [2] "tidy_tomato$Foliage_Color   5   5.75  1.1495   3.958 0.00149 **"
## [3] "Residuals                 826 239.88  0.2904                   "
## [4] "---"
## [5] "Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1"
```

Problem 7

```
## [1] "Gebreken"
```

```
##    Gebrek.identificatie     Ingangsdatum.gebrek        Einddatum.gebrek
##             "character"               "integer"               "integer"
## Gebrek.paragraaf.nummer   Gebrek.artikel.nummer     Gebrek.omschrijving
##               "integer"             "character"             "character"
```

```
## [1] "Geconstat"
```

```
##                          Kenteken Soort.erkenning.keuringsinstantie
##                       "character"                       "character"
## Meld.datum.door.keuringsinstantie  Meld.tijd.door.keuringsinstantie
##                         "integer"                         "integer"
##              Gebrek.identificatie      Soort.erkenning.omschrijving
##                       "character"                       "character"
##     Aantal.gebreken.geconstateerd
##                         "integer"
```

```
## [1] "Personen"
```

```
##                          Kenteken                     Voertuigsoort
##                       "character"                       "character"
##                              Merk                   Handelsbenaming
##                       "character"                       "character"
##              Datum.tenaamstelling                         Bruto.BPM
##                       "character"                         "integer"
##                    Cilinderinhoud              Massa.ledig.voertuig
##                         "integer"                         "integer"
## Toegestane.maximum.massa.voertuig            Datum.eerste.toelating
##                         "integer"                       "character"
##    Datum.eerste.afgifte.Nederland                    Catalogusprijs
##                       "character"                         "integer"
##                     WAM.verzekerd
##                       "character"
```