STAT 5014 Homework 2

Samantha Sunshine

9/11/17

Problem 4

Version control can be beneficial in the classroom if I want to allow other people to work on a piece of code that I share with them. It would also help if I have made a mistake and realize I need to go back to a previous version of the code. Another benefit would be that I can access my files from multiple computers if I need to work on something and do not have my own computer.

Problem 5

a. Sensory Data

Table 1: Sensory Data

Item	Person	value
1	1	4.3
1	1	4.3
1	1	4.1
1	2	4.9
1	2	4.5
1	2	5.3
1	3	3.3
1	3	4.0
1	3	3.4
1	4	5.3
1	4	5.5
1	4	5.7
1	5	4.4
1	5	3.3
1	5	4.7
10	1	5.0
10	1	5.4
10	1	2.8
10	2	4.8
10	2	5.0

The first step was to create a table names Sensosry_raw that included the data from the url with no header and skipped the first line of original data.

```
Sensory_raw<-read.table(url, header=F, skip=1, fill=T, stringsAsFactors = F)
```

Then I created Sensory_tidy, which was the same as Sensory_raw but without the first row.

```
Sensory_tidy<-Sensory_raw[-1,]
```

The next line of code named a new dataset, Sensory_tidy_a, with the data from Sensory_tidy. This filtered the column V1 with all the nunmbers between 1 and 10, and it renamed the columns.

```
Sensory_tidy_a<-filter(.data = Sensory_tidy,V1 %in% 1:10) %>%
  rename(Item=V1,V1=V2,V2=V3,V3=V4,V4=V5,V5=V6)
```

The next chunk of code created another dataset, Sensory_tidy_b, which took the data from column V1 that was not between 1 and 10. Then the mutate function made the "Item" column repeat the numbers 1 through 10, each two times. Column V1 is now numeric factors, and I selected the data in "Item", and columns V1 through V5.

```
Sensory_tidy_b<-filter(.data = Sensory_tidy,!(V1 %in% 1:10)) %>%
                    mutate(Item=rep(as.character(1:10),each=2)) %>%
                    mutate(V1=as.numeric(V1)) %>%
                    select(c(Item,V1:V5))
```

The following line combined the two datasets as rows, one on top of the other.

```
Sensory_tidy<-bind_rows(Sensory_tidy_a,Sensory_tidy_b)
```

By renaming the column names of Sensory_tidy, I now have columns "Item", "Person_1", "Person_2", etc.

```
colnames(Sensory_tidy)<-c("Item",paste("Person",1:5,sep="_"))
```

The last chunk of code changed Sensory_tidy so that it created a column named "Person", which contained "Person_1", "Person_2", … corresponding to each data point. The corresponding data points were gathered into the new column "value". The mutate function changed the "Person" column so that we subsituted a blank space for "Person_", which left just the numbers, "1, 2, …". The last step was to arrange the data by the "Item" column in ascending order.

```
Sensory_tidy<-Sensory_tidy %>%
        gather(Person,value,Person_1:Person_5) %>%
        mutate(Person = gsub("Person_","",Person)) %>%
        arrange(Item)
```

Table 2: Sensory data summary

Item	Person	value
Length:150	Length:150	Min. :0.700
Class :character	Class :character	1st Qu.:3.025
Mode :character	Mode :character	Median :4.700
NA	NA	Mean :4.657
NA	NA	3rd Qu.:6.000
NA	NA	Max. :9.400

b. Long Jump Data

Table 3: Long Jump Data

YearCode	Year	dist
-4	1896	249.75
0	1900	282.88
4	1904	289.00
8	1908	294.50
12	1912	299.25

YearCode	Year	dist
20	1920	281.50
24	1924	293.13
28	1928	304.75
32	1932	300.75
36	1936	317.31
48	1948	308.00
52	1952	298.00
56	1956	308.25
60	1960	319.75
64	1964	317.75
68	1968	350.50
72	1972	324.50
76	1976	328.50
80	1980	336.25
84	1984	336.25
88	1988	343.25
92	1992	342.50

The first step was to create a table named LongJump_raw from the url data with no header and by skipping the first line.

```
LongJump_raw<-read.table(url,header = F,skip = 1,fill = T,stringsAsFactors = F)
```

Next, I renamed the column names of LongJump_raw as V1 and V2, repeated 4 times.

```
colnames(LongJump_raw)<-rep(c("V1","V2"),4)
```

The next line of code named a new dataset, LongJump_tidy, which combined the data in LongJump_raw. Columns 1 and 2 were combined, as were 3 and 4, 5 and 6, and finally 7 and 8.

```
LongJump_tidy<-rbind(LongJump_raw[ ,1:2],LongJump_raw[ ,3:4],LongJump_raw[ ,5:6],LongJump_raw[ ,7:8])
```

The last chunk of code changed LongJump_tidy to filter the data in column V1 that is not missing. The mutate function created new columns YearCode, which was the same as V1, Year, which added 1900 to V1 so it was an actual year, and dist, which was the same as V2. Then I selected the data set without columns V1 and V2.

```
LongJump_tidy<-LongJump_tidy %>%
  filter(!(is.na(V1))) %>%
  mutate(YearCode=V1,Year=V1+1900,dist=V2) %>%
  select(-V1,-V2)
```

Table 4: Long Jump data summary

YearCode	Year	dist
Min. :-4.00	Min. :1896	Min. :249.8
1st Qu.:21.00	1st Qu.:1921	1st Qu.:295.4
Median :50.00	Median :1950	Median :308.1
Mean :45.45	Mean :1945	Mean :310.3
3rd Qu.:71.00	3rd Qu.:1971	3rd Qu.:327.5
Max. :92.00	Max. :1992	Max. :350.5

c. Brain and Body Data

Table 5: Brain and Body Data

Brain	Body
3.385	44.5
0.480	15.5
1.350	8.1
465.000	423.0
36.330	119.5
27.660	115.0
14.830	98.2
1.040	5.5
4.190	58.0
0.425	6.4
0.101	4.0
0.920	5.7
1.000	6.6
0.005	0.1
0.060	1.0
3.500	10.8
2.000	12.3
1.700	6.3
2547.000	4603.0
0.023	0.3

First, I created a table named BrainBody_raw from the url data, which had no header, and skipped the first line of data.

```r
BrainBody_raw<-read.table(url,header=F,skip = 1,fill = T,stringsAsFactors = F)
```

The next step was to rename the columns as “Brain” and “Body” repeated 3 times.

```r
colnames(BrainBody_raw)<-rep(c("Brain","Body"),3)
```

I created a new dataset named BrainBody_tidy, which combined the columns from BrainBody_raw. Columns 1 and 2 were now combined, as were columns 3 and 4, and columns 5 and 6.

```r
BrainBody_tidy<-rbind(BrainBody_raw[ ,1:2],BrainBody_raw[ ,3:4],BrainBody_raw[ ,5:6])
```

Finally, BrainBody_tidy was filtered by the data that was not missing from the “Brain” column.

```r
BrainBody_tidy<-BrainBody_tidy %>%
  filter(!(is.na(Brain)))
```

Table 6: Brain/Body weight data summary

Brain	Body
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

d. Tomato Data

Table 7: Tomato Data

Clone	Replicate	value	Variety
10000	1	16.1	Ife 1
10000	2	15.3	Ife 1
10000	3	17.5	Ife 1
20000	1	16.6	Ife 1
20000	2	19.2	Ife 1
20000	3	18.5	Ife 1
30000	1	20.8	Ife 1
30000	2	18.0	Ife 1
30000	3	21.0	Ife 1
10000	1	8.1	PusaEarlyDwarf
10000	2	8.6	PusaEarlyDwarf
10000	3	10.1	PusaEarlyDwarf
20000	1	12.7	PusaEarlyDwarf
20000	2	13.7	PusaEarlyDwarf
20000	3	11.5	PusaEarlyDwarf
30000	1	14.4	PusaEarlyDwarf
30000	2	15.4	PusaEarlyDwarf
30000	3	13.7	PusaEarlyDwarf

First, I made a table named Tomato_raw from the url data, which had no header and skipped the first two lines.

```r
Tomato_raw<-read.table(url,header = F,skip = 2,fill = T,stringsAsFactors = F,comment.char = "")
```

Tomato_tidy is the new dataset that represents Tomato_raw, except I separated V2 into 3 columns, “C10000_1”, “C10000_2”, and “C10000_3”. I also removed the comma that was originally separating the data.
I did the same thing for V3 and V4, which split into 3 columns each, “C20000_1”, “C20000_2”, and “C20000_3”, and “C30000_1”, “C30000_2”, and “C30000_3”, respectively.
The mutate function eliminated the extra comma in C10000_3.
The gather function created a new column named “Clone”, which contained “C10000_1”, “C10000_2”, … corresponding to each data point. The corresponding data points were arranged into a new column named “value”.
Mutate created a new column, “Variety”, which was the same as column V1. It also substituted a blank space for “C” in the “Clone” column, which left “10000_1”, “10000_2”, etc. Mutate also substituted a blank space for “\#” in the “Variety” column.
The separate function broke the “Clone” column into two by splitting “10000_1” into “10000” in the “Clone” column and “1” in the new “Replicate” column.
The last two parts were to select the data without column V1, and with columns Variety, Clone, and value. Then I arranged the data by Variety.

```r
Tomato_tidy<-Tomato_raw %>%
  separate(V2,into=paste("C10000",1:3,sep="_"),sep=",",remove=T,extra="merge") %>%
  separate(V3,into=paste("C20000",1:3,sep="_"),sep=",",remove=T,extra="merge") %>%
  separate(V4,into=paste("C30000",1:3,sep="_"),sep=",",remove=T,extra="merge") %>%
  mutate(C10000_3=gsub(",","",C10000_3)) %>%
  gather(Clone,value,C10000_1:C30000_3) %>%
  mutate(Variety=V1,Clone=gsub("C","",Clone)) %>%
  mutate(Variety=gsub("\\\\#"," ",Variety)) %>%
  separate(Clone,into=c("Clone","Replicate")) %>%
  select(-V1,Variety,Clone,value) %>%
  arrange(Variety)
```

Table 8: 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

Problem 6

Table 9: Plants Data

Scientific_Name	Foliage_Color	pH_Average
Pinus rigida	Yellow-Green	4.30
Gaylussacia frondosa	Green	4.65
Lycopodium annotinum	Green	4.65
Betula nigra	Green	4.75
Osmunda regalis var. spectabilis	Green	4.75
Rhododendron maximum	Dark Green	4.75
Chamaecyparis thyoides	Green	4.90
Picea rubens	Green	4.90
Rhododendron periclymenoides	Green	4.90
Ammannia coccinea	Green	4.95
Rhododendron arborescens	Dark Green	4.95
Rhododendron atlanticum	Dark Green	4.95
Tsuga canadensis	Dark Green	4.95
Abies balsamea	Green	5.00
Betula populifolia	Green	5.00
Dryopteris cristata	Dark Green	5.00
Ilex decidua	Dark Green	5.00
Kalmia latifolia	Green	5.00
Ludwigia decurrens	Green	5.00
Osmunda claytoniana	Dark Green	5.00

Table 10: Plants Data Summary

Scientific_Name	Foliage_Color	pH_Average
Abies balsamea : 1	Dark Green : 82	Min. :4.30
Acacia constricta : 1	Gray-Green : 25	1st Qu.:5.80
Acalypha virginica: 1	Green :692	Median :6.15
Acer negundo : 1	Red : 4	Mean :6.17
Acer nigrum : 1	White-Gray : 9	3rd Qu.:6.50
Acer pensylvanicum: 1	Yellow-Green: 20	Max. :8.20
(Other) :826	NA	NA

```
Warning in model.response(mf, "numeric"): using type = "numeric" with a
factor response will be ignored
```

```
Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

```

Call:
lm(formula = plants_tidy)

Coefficients:
              (Intercept)    Foliage_ColorGray-Green
                  305.302                    -14.734
       Foliage_ColorGreen           Foliage_ColorRed
                   -8.101                     87.318
  Foliage_ColorWhite-Gray  Foliage_ColorYellow-Green
                  -84.281                     25.232
               pH_Average
                   19.169
```

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"
```

Table 11: Gebreken Data

Gebrek.identificatie	Ingangsdatum.gebrek	Einddatum.gebrek	Gebrek.paragraaf.nummer	Gebrek.artikel.nummer	Ge
O01	20100227	20160103	14	REK 24&25	Ke
D05	20100227	20150331	3	5.5.11	G
O04	20100227	20160103	14	REK 24&25	Br
O05	20100227	20160103	14	REK 24&25	Tr
O03	20100227	20160103	14	REK 24&25	W
O02	20100227	20160103	14	REK 24&25	M
J13	20120401	20170402	9	5.*.41	Be
H08	20120401	20170402	7	5.*.29	St
O06	20150401	20170402	14	REK 24&25	K
H04	20120401	20170402	7	5.*.29	W
J16	20120401	20170402	9	5.*.48	W
H17	20120401	20170402	7	5.5.30	Vo
I27	20120401	20170402	8	5.*.31	Al
F13	20120401	20170402	5	5.*.19	Ot
E05	20120401	20170402	4	5.*.15	Ve
I33	20120401	20170402	8	5.*.37	Ee
I13	20120401	20170402	8	5.*.31	O
B12	20120401	20170402	1	5.12.5	Ac
J03	20100227	20170402	9	5.*.43	Rt
H16	20120401	20170402	7	5.*.29	Sl

Table 12: Gebreken Summary Data

Gebrek.identificatie	Ingangsdatum.gebrek	Einddatum.gebrek	Gebrek.paragraaf.nummer	Gebrek.artikel.nummer
Length:20	Min. :20100227	Min. :20150331	Min. : 1.0	Length:20
Class :character	1st Qu.:20100227	1st Qu.:20160103	1st Qu.: 7.0	Class :character
Mode :character	Median :20120401	Median :20170402	Median : 8.0	Mode :character
NA	Mean :20114840	Mean :20166824	Mean : 8.8	NA
NA	3rd Qu.:20120401	3rd Qu.:20170402	3rd Qu.:14.0	NA
NA	Max. :20150401	Max. :20170402	Max. :14.0	NA

Table 13: Geconstat Data

Kenteken	Soort.erkenning.keuringsinstantie	Meld.datum.door.keuringsinstantie	Meld.tijd.door.keuringsinstantie	Geb
XPGN96	AL	20150304	1554	K04
04JKFV	AL	20160118	1433	I12
02BRDX	AL	20160707	1659	RA1
08BJPV	AL	20160119	941	D14
TJPX77	AL	20160510	1147	J03
LVRG41	AL	20161028	1038	K04
VJ67TN	AL	20141218	1549	K04
83PZVF	AL	20161222	932	K07
77LZPS	AL	20160323	1327	RA1
OK53YT	AZ	20150122	1144	I10
95TDSK	AL	20160114	1543	AC1
40FRZN	AL	20150529	1436	J03
72BFPR	AL	20150210	1502	G11
92PPRL	AL	20151209	1120	J03

Kenteken	Soort.erkenning.keuringsinstantie	Meld.datum.door.keuringsinstantie	Meld.tijd.door.keuringsinstantie	Geb
95LGNH	AL	20150731	1436	I21
20XPDV	AL	20161220	1516	G05
49HVXZ	AL	20150213	1518	G05
11ZJZJ	AL	20160512	1513	K05
33PPPP	AL	20160414	1531	K04
52PHSV	AL	20161102	1023	I20

Table 14: Geconstat Summary Dat

Kenteken	Soort.erkenning.keuringsinstantie	Meld.datum.door.keuringsinstantie	Meld.tijd.door.keuringsinstan
Length:20	Length:20	Min. :20141218	Min. : 932
Class :character	Class :character	1st Qu.:20150473	1st Qu.:1138
Mode :character	Mode :character	Median :20160118	Median :1436
NA	NA	Mean :20156096	Mean :1343
NA	NA	3rd Qu.:20160561	3rd Qu.:1521
NA	NA	Max. :20161222	Max. :1659

Kenteken	Voertuigsoort	Merk	Handelsbenaming	Datum.tenaamstelling	Bruto.BPM	Cilinder
75JGGT	Personenauto	SUZUKI	WAGON R; + 1.3	17/02/2012	2071	
53XBZF	Personenauto	MITSUBISHI	MITSUBISHI COLT	31/03/2010	3266	
85JDV8	Personenauto	SUZUKI	SPLASH	16/12/2015	2888	
2KRS53	Personenauto	BMW	118I	10/02/2017	4875	
JS119V	Personenauto	RENAULT	MEGANE	29/04/2017	4124	
9KGD19	Personenauto	SUZUKI	ALTO	13/04/2013	2875	
05ZRFT	Personenauto	SAAB	SAAB 9-3	31/10/2016	22774	
24SGGP	Personenauto	BMW	1ER REIHE; 118I	03/03/2017	8318	
12NZDH	Personenauto	FIAT	FIAT IDEA; 1.3 JTD	04/10/2016	6032	
28GLB3	Personenauto	PEUGEOT	1007	26/01/2016	4214	
19NJPV	Personenauto	AUDI	AUDI A3; 74 KW	02/12/2005	5390	
PK535K	Personenauto	HYUNDAI	I10	11/07/2017	3091	
14PSX7	Personenauto	CITROEN	C3	10/08/2016	NA	
24PHHL	Personenauto	HONDA	CIVIC 5 DR; 1.6I AT	25/06/2004	6931	
47JNKL	Personenauto	RENAULT	CLIO; 1.6 16V S2005	02/01/2014	2973	
76DDGS	Personenauto	NISSAN	NISSAN MICRA; 1.0 3HB	28/03/2013	1935	
19TGB3	Personenauto	SKODA	FABIA	18/03/2017	2172	
89ZTF1	Personenauto	MINI	MINI COOPER; COOPER	03/03/2017	4335	
50GSB1	Personenauto	TOYOTA	TOYOTA AYGO	14/05/2016	128	
NR936X	Personenauto	ALFA ROMEO	ALFA GIULIETTA	31/03/2017	5846	

Kenteken	Voertuigsoort	Merk	Handelsbenaming	Datum.tenaamstelling	Bruto.BPM	Ci
Length:20	Length:20	Length:20	Length:20	Length:20	Min. : 128	M
Class :character	Class :character	Class :character	Class :character	Class :character	1st Qu.: 2882	1s
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Median : 4124	M
NA	NA	NA	NA	NA	Mean : 4960	M

Kenteken	Voertuigsoort	Merk	Handelsbenaming	Datum.tenaamstelling	Bruto.BPM	Ci
NA	NA	NA	NA	NA	3rd Qu.: 5618	3r
NA	NA	NA	NA	NA	Max. :22774	M
NA	NA	NA	NA	NA	NA's :1	N