506 Problem Set 1

Zimo Shu

Problem 1 - Abalone Data

a. Import the data into a data.frame in R. Use the information in the "abalone.names" file to give appropriate column names.

```
# Confirm the files' path in my directory
getwd()
```

```
[1] "/Users/amanda/stats506coursework"
```

```
dir("abalonedata")
```

```
[1] "abalone.data"  "abalone.names"
```

```
# Import the data into a data.frame
df_data <- read.table("~/stats506coursework/abalonedata/abalone.data", sep = ",", header = FA
```

```
# Take a look at the data
head(df_data)
```

```
  V1    V2    V3    V4     V5     V6     V7    V8 V9
1  M 0.455 0.365 0.095 0.5140 0.2245 0.1010 0.150 15
2  M 0.350 0.265 0.090 0.2255 0.0995 0.0485 0.070  7
3  F 0.530 0.420 0.135 0.6770 0.2565 0.1415 0.210  9
4  M 0.440 0.365 0.125 0.5160 0.2155 0.1140 0.155 10
5  I 0.330 0.255 0.080 0.2050 0.0895 0.0395 0.055  7
6  I 0.425 0.300 0.095 0.3515 0.1410 0.0775 0.120  8
```

```r
head(readLines("abalonedata/abalone.names"), 10) # Source: googled how to properly use readLi
```

```
 [1] "1. Title of Database: Abalone data"
 [2] ""
 [3] "2. Sources:"
 [4] ""
 [5] "   (a) Original owners of database:"
 [6] "\tMarine Resources Division"
 [7] "\tMarine Research Laboratories - Taroona"
 [8] "\tDepartment of Primary Industry and Fisheries, Tasmania"
 [9] "\tGPO Box 619F, Hobart, Tasmania 7001, Australia"
[10] "\t(contact: Warwick Nash +61 02 277277, wnash@dpi.tas.gov.au)"
```

According to abalone.names, the eight attributes are **Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight and Rings.**

```r
# Use the information in the "abalone.names" file to give appropriate column names
colnames(df_data) <- c("Sex","Length","Diameter","Height",
                       "WholeWeight","ShuckedWeight",
                       "VisceraWeight","ShellWeight","Rings")
head(df_data)
```

```
  Sex Length Diameter Height WholeWeight ShuckedWeight VisceraWeight
1   M  0.455    0.365  0.095      0.5140        0.2245        0.1010
2   M  0.350    0.265  0.090      0.2255        0.0995        0.0485
3   F  0.530    0.420  0.135      0.6770        0.2565        0.1415
4   M  0.440    0.365  0.125      0.5160        0.2155        0.1140
5   I  0.330    0.255  0.080      0.2050        0.0895        0.0395
6   I  0.425    0.300  0.095      0.3515        0.1410        0.0775
  ShellWeight Rings
1       0.150    15
2       0.070     7
3       0.210     9
4       0.155    10
5       0.055     7
6       0.120     8
```

b. The data contains information on three different sexes of abalone. Report the number of observations belonging to each sex.

```
# Report the number of observations belonging to each sex
sex_num <- table(df_data$Sex)
sex_num
```

```
   F    I    M
1307 1342 1528
```

Hence, there are 1307 Female observations, 1342 Infant observations, and 1528 Male observations.

c. Use the data to answer the following questions:

1. Which weight has the highest correlation with rings?

```
weights <- df_data[, c("WholeWeight", "ShuckedWeight", "VisceraWeight", "ShellWeight")]
# The correlation with rings for different sorts of weight
cor_weightrings <- cor(weights, df_data$Rings)
cor_weightrings
```

```
                   [,1]
WholeWeight   0.5403897
ShuckedWeight 0.4208837
VisceraWeight 0.5038192
ShellWeight   0.6275740
```

Hence, ShellWeight has the highest correlation (0.6275740) with rings.

2. For that weight, which sex has the highest correlation?

```
# Subset data by sex
df_f <- subset(df_data, Sex == "F")
df_m <- subset(df_data, Sex == "M")
df_i <- subset(df_data, Sex == "I")

# For ShellWeight, correlations with Rings by sex
cor(df_f$ShellWeight, df_f$Rings)
```

```
[1] 0.405907
```

```r
cor(df_m$ShellWeight, df_m$Rings)
```

```
[1] 0.5109967
```

```r
cor(df_i$ShellWeight, df_i$Rings)
```

```
[1] 0.7254357
```

So for ShellWeight, the sex I (infants) has the highest correlation.

3. What are the weights of the abalone with the most rings?

```r
# Find the most rings
max_ring <- max(df_data$Rings)
most_rings <- subset(df_data, Rings == max_ring)
# Find the weights of the abalone with the most rings
most_rings[, c("WholeWeight","ShuckedWeight","VisceraWeight","ShellWeight")]
```

```
    WholeWeight ShuckedWeight VisceraWeight ShellWeight
481      1.8075        0.7055        0.3215       0.475
```

The weights of the abalone with the most rings are shown as above.

WholeWeight: 1.8075

ShuckedWeight: 0.7055

VisceraWeight: 0.3215

ShellWeight: 0.475

4. What percentage of abalones have a viscera weight larger than their shell weight?

```r
# Viscera weight larger than their shell weight  df_data$VisceraWeight > df_data$ShellWeight
mean(df_data$VisceraWeight > df_data$ShellWeight) * 100
```

```
[1] 6.511851
```

Therefore, about 6.512% of abalones have a viscera weight larger than their shell weight.

d. Create a table of correlations between weights and rings, within each sex. The columns should be the four weights, and the rows should be the sexes. (This table does not need to be "fancy" but should clearly identify what each value represents.)

```
# Get three correlations between weights and rings for sex
# F
cor_f <- cor(subset(df_data, Sex == "F")$Rings, subset(df_data, Sex == "F")[, c("WholeWeight"
# M
cor_m <- cor(subset(df_data, Sex == "M")$Rings, subset(df_data, Sex == "M")[, c("WholeWeight"

# I
cor_i <- cor(subset(df_data, Sex == "I")$Rings, subset(df_data, Sex == "I")[, c("WholeWeight"

# Combine the correlations into one table
correlation_table <- rbind(Female = cor_f, Male = cor_m, Infant = cor_i)
rownames(correlation_table) <- c("Female", "Male", "Infant")
correlation_table
```

```
       WholeWeight ShuckedWeight VisceraWeight ShellWeight
Female   0.2667585    0.09484802     0.2116154   0.4059070
Male     0.3721966    0.22239382     0.3209535   0.5109967
Infant   0.6963268    0.62024577     0.6732727   0.7254357
```

The correlation table is as above.

e. Carry out a series of t-tests to examine whether the number of rings differs across the three sexes. Present the R output and interpret the results.

```
# Subset rings from the data frame in terms of different sexes
ring_F <- subset(df_data, Sex == "F")$Rings
ring_M <- subset(df_data, Sex == "M")$Rings
ring_I <- subset(df_data, Sex == "I")$Rings
```

```
# t tests between Female and Male
test_FM <- t.test(ring_F, ring_M)
test_FM
```

```

    Welch Two Sample t-test

data:  ring_F and ring_M
```

```
t = 3.6657, df = 2742.4, p-value = 0.0002514
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1971045 0.6505082
sample estimates:
mean of x mean of y
  11.1293   10.7055
```

```r
# t tests between Female and Infants
test_FI <- t.test(ring_F, ring_I)
test_FI
```

```
    Welch Two Sample t-test

data:  ring_F and ring_I
t = 29.477, df = 2508.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.023380 3.454304
sample estimates:
mean of x mean of y
11.129304  7.890462
```

```r
# t tests between Male and Infants
test_MI <- t.test(ring_M, ring_I)
test_MI
```

```
    Welch Two Sample t-test

data:  ring_M and ring_I
t = 27.221, df = 2859, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.612263 3.017808
sample estimates:
mean of x mean of y
10.705497  7.890462
```

According to the outputs, the p-values from three t-test results are all pretty small (less than 0.05 and 0.001), which implies that there is sufficient evidence for us to reject the null hypothesis of equal mean rings for each pair of sexes. Specifically, we could see that females have slightly more rings than males, and females and males have more rings than infants.

Problem 2 - Food Expenditure Data

a. Import the data into a data.frame in R.

```r
df_food <- read.csv("~/Downloads/food_expenditure.csv")
head(df_food)
```

```
  ID What.is.your.age.
1  1                68
2  2                88
3  3                82
4  4                73
5  5                89
6  6                18
  How.many.individuals.live.in.your.household.for.which.you.are.responsible.for.food.expendi
1
2
3
4
5
6
  What.state.do.you.live.in.
1                         LA
2                         WA
3                         MS
4                         AK
5                         IN
6                         WI
  What.currency.are.you.reporting.your.food.expenditures.in.
1                                                        USD
2                                                        USD
3                                                        USD
4                                                        USD
5                                                        USD
6                                                        EUR
  What.was.your.total.food.expenditure.in.the.last.week.
1                                                 436.35
2
```

```
3                                                279.1
4                                               -20.98
5                                               494.87
6                                               276.32
  What.was.your.total.food.expenditures.at.grocery.stores.in.the.last.week.
1                                                                   168.59
2                                                                   452.10
3                                                                   301.66
4                                                                   139.66
5                                                                       NA
6                                                                   394.44
  What.was.your.food.expenditure.while.dining.out.in.the.last.week.
1                                                           140.71
2                                                           192.94
3                                                           239.84
4                                                            69.19
5                                                           191.72
6                                                           283.20
  What.was.your.food.expenditure..miscellaneous..in.the.last.week.
1                                                          109.77
2                                                              NA
3                                                          103.94
4                                                           44.84
5                                                          172.31
6                                                          114.06
  How.many.times.did.you.dine.out.last.week.
1                                          4
2                                          1
3                                          9
4                                          2
5                                          3
6                                          6
  Are.you.including.alcohol.in.your.food.expenditures.
1                                                  Yes
2                                              Unknown
3                                                  Yes
4                                              Unknown
5                                                  Yes
6                                              Unknown
  What.food.assistance.programs..if.any..did.you.use.for.your.food.expenditures.last.week.
1                                                                                     None
2                                                                                     SNAP
3                                                                                     None
```

```
4                                                                                 None
5                                                                                 None
6                                                                          Food Pantry
```

b. Clean up the variable names. Simplify them.

```r
names(df_food)
```

```
 [1] "ID"
 [2] "What.is.your.age."
 [3] "How.many.individuals.live.in.your.household.for.which.you.are.responsible.for.food.expe
 [4] "What.state.do.you.live.in."
 [5] "What.currency.are.you.reporting.your.food.expenditures.in."
 [6] "What.was.your.total.food.expenditure.in.the.last.week."
 [7] "What.was.your.total.food.expenditures.at.grocery.stores.in.the.last.week."
 [8] "What.was.your.food.expenditure.while.dining.out.in.the.last.week."
 [9] "What.was.your.food.expenditure..miscellaneous..in.the.last.week."
[10] "How.many.times.did.you.dine.out.last.week."
[11] "Are.you.including.alcohol.in.your.food.expenditures."
[12] "What.food.assistance.programs..if.any..did.you.use.for.your.food.expenditures.last.weel
```

I would use ID, Age, Size, State, Currency, Total_exp, Grocery_exp, Dine_exp, Misc, Dine_freq, Alcohol and Assistance for variable names.

```r
names(df_food) <-c("ID","Age","Size","State","Currency","Total_exp","Grocery_exp","Dine_exp"

# Check the table
df_food
```

```
   ID Age Size State Currency Total_exp Grocery_exp Dine_exp   Misc Dine_freq
1   1  68    7    LA      USD    436.35      168.59   140.71 109.77         4
2   2  88    5    WA      USD                452.10   192.94     NA         1
3   3  82    3    MS      USD     279.1      301.66   239.84 103.94         9
4   4  73    8    AK      USD    -20.98      139.66    69.19  44.84         2
5   5  89    0    IN      USD    494.87          NA   191.72 172.31         3
6   6  18    6    WI      EUR    276.32      394.44   283.20 114.06         6
7   7  38    4    DC      USD    318.79      153.49   104.05  39.21         1
8   8  28    8    ID      USD    304.52      286.70       NA  24.61        10
9   9  16    4    SC      USD    325.71      484.22   289.89 145.01         1
10 10  84    1    HI      USD    332.08      236.68   105.59  38.86         9
11 11  46    1    ND      USD   -201.52       40.54    10.57  40.24         1
```

12	12	29	5	UT	USD	622.58	144.16	58.50	14.73	5
13	13	90	7	DC	USD	292.08	168.88	64.77	29.12	8
14	14	22	5	NV	USD	505.11	381.19	121.20	76.45	8
15	15	70	6	WA	USD	311.84	212.63	93.26	111.79	30
16	16	3	1	ME	USD	555.39	280.91	63.78	NA	0
17	17	31	8	WI	USD	529.38	139.08	NA	24.01	2
18	18	23	12	VA	USD	404.6	243.86	159.71	94.78	10
19	19	86	6	RI	USD	561.31	444.56	186.67	NA	6
20	20	81	7	MT	USD	604.2	254.94	152.27	145.16	10
21	21	51	1	AK	USD	794.98	409.40	123.04	132.14	1
22	22	70	4	IL	USD	284.72	197.66	122.72	66.61	2
23	23	39	3	PA	EUR	184.15	280.08	149.95	76.74	5
24	24	79	6	OR	USD	789.86	269.49	199.48	154.01	9
25	25	33	7	WA	USD	141.85	265.90	58.71	203.34	9
26	26	19	1	CT	USD	865.36	438.04	148.28	NA	10
27	27	47	12	GA	USD	457.64	220.24	63.68	181.99	1
28	28	42	5	NV	EUR	762.41	266.37	142.55	184.82	9
29	29	48	12	MT	USD	514.22	105.19	48.59	47.80	15
30	30	85	12	NC	USD	431.93	244.75	130.94	131.07	20
31	31	59	6	TX	USD	44.13	347.12	58.75	146.84	8
32	32	48	12	NE	USD	~350	227.74	78.38	55.25	7
33	33	37	7	NY	USD	-25	66.24	30.79	NA	7
34	34	58	6	CO	USD	398.68	186.02	123.60	153.23	10
35	35	79	0	DC	USD	477.72	237.40	NA	32.02	20
36	36	75	8	WY	USD	346.09	429.12	240.54	208.01	5
37	37	150	0	IL	USD	-25	64.36	13.93	59.46	6
38	38	58	8	IA	CAD	852.93	216.58	95.28	125.48	8
39	39	23	1	NV	USD	366	248.39	61.93	123.18	30
40	40	61	0	DC	CAD	156.05	388.40	162.57	81.65	9
41	41	68	4	FL	USD	113.97	139.72	67.64	17.44	7
42	42	58	7	SC	CAD	366.45	79.78	39.53	52.27	5
43	43	62	2	AZ	USD	358.13	161.24	118.86	33.27	3
44	44	77	1	NM	USD	654.01	278.24	185.93	77.01	30
45	45	46	0	WA	EUR	364.97	458.23	186.53	135.61	1
46	46	51	0	AL	CAD		350.09	81.04	327.59	2
47	47	30	5	MI	USD	0.46	198.30	56.36	184.63	2
48	51	21	1	NV	USD	0	175.13	101.38	155.85	4
49	52	31	5	IL	USD	574.63	135.49	99.38	20.42	9
50	53	27	1	MN	USD	481.04	381.95	207.97	57.90	1
51	54	67	4	NV	USD	-25	373.16	184.65	-22.47	2
52	55	24	3	WI	USD	257.59	62.73	34.88	52.37	7
53	56	15	2	WV	USD	537.65	226.99	131.19	-19.81	20
54	57	38	1		USD	434.91	253.20	45.38	70.11	0

55	58	67	8	OH	USD	427.24	34.22	21.13	5.64	10
56	59	42	0	LA	USD	-25	155.34	98.89	80.35	15
57	60	65	7	NJ	CAD	200.74	213.79	175.49	94.52	10
58	61	22	3	OK	USD	253.33	360.63	103.60	-0.02	3
59	62	21	3	ME	USD	151.63	184.29	163.52	66.11	8
60	64	74	6	RI	USD	331.69	172.17	76.16	60.66	15
61	65	34	1	AL	USD	377.16	123.85	66.73	53.43	0
62	66	44	5	WV	USD	808.2	453.84	134.64	NA	10
63	67	81	0	FL	CAD	299.44	152.47	70.78	103.36	2
64	68	73	0	CA	USD	903.34	535.71	177.89	NA	15
65	69	65	4		USD	300.85	215.11	162.36	127.87	4
66	70	74	8	AL	USD	554.75	101.25	32.45	30.16	4
67	71	NA	0	XX	USD	0	6.92	4.02	5.22	3
68	72	66	8	NE	CAD	438.77	424.70	104.74	89.23	7
69	73	42	0	KS	USD	10.69	429.88	114.11	NA	5
70	74	18	2	UT	USD	-25	312.83	174.82	-1.68	8
71	75	150	6	PA	USD	182.65	403.44	302.01	64.67	9
72	77	70	6	WY	USD	659.87	77.55	25.79	10.18	9
73	78	83	5	VT	USD	779.49	226.23	145.32	44.29	10
74	79	71	0	DE	CAD	596.91	74.52	51.07	19.85	7
75	80	61	4	XX	USD	350.89	155.17	32.77	128.25	9
76	81	69	0	IL	USD	-113.7	189.40	87.00	196.43	10
77	82	45	2	PA	USD	0	192.51	60.96	NA	2
78	83	18	6	XX	USD		167.49	67.61	47.66	7
79	84	NA	12	ND	USD	631.88	234.08	54.77	NA	6
80	85	80	6	MN	USD	338.99	392.50	190.88	257.16	3
81	86	34	2	KY	EUR	346.87	196.75	126.31	158.98	4
82	87	66	4	ND	USD	510.02	655.63	116.01	128.92	6
83	88	84	3	CO	USD	291.98	234.25	48.84	160.69	3
84	89	27	1	IL	USD	476.51	287.35	142.44	256.41	4
85	90	81	4	PR	USD	669.93	378.69	263.26	49.67	7
86	92	3	3	AL	EUR	360.95	214.35	37.66	82.36	3
87	93	19	8	MO	USD	232.01	582.38	224.92	NA	5
88	94	7	12	PR	USD	796.02	216.26	89.80	63.43	2
89	95	30	12	TX	USD	68.05	230.87	136.63	72.37	6
90	96	6	12	NE	USD		452.31	NA	113.17	3
91	97	59	3	MA	CAD	468.61	141.87	89.95	99.39	6
92	98	17	1	HI	USD	312.88	462.08	217.19	7.54	3
93	99	85	6	NH	USD	628.59	190.77	71.82	80.38	15
94	100	150	12	IA	USD	474.25	21.99	4.71	15.93	9
95	101	20	12	CA	USD	554.42	380.56	147.83	62.63	10
96	102	8	5	TN	USD	-28.59	327.06	98.62	89.20	7
97	103	43	12	KS	CAD	375.79	112.27	56.91	23.37	5

98	104	46	6	MA	USD	358.42	195.66	130.40	43.42	3
99	105	81	1	IL	USD	493.97	313.15	149.92	238.45	1
100	106	79	1	MD	USD	674.45	249.78	200.70	149.13	5
101	108	17	1	IL	CAD	412.68	327.76	158.38	NA	10
102	109	34	1	MS	USD	357.18	181.17	130.43	93.66	9
103	110	83	3	AZ	USD	289.94	26.66	18.79	10.31	3
104	111	63	6	VA	USD		313.34	73.57	124.93	3
105	112	88	0	OH	USD	38.9	231.06	72.45	NA	5
106	113	77	0	TN	USD	80.85	304.51	194.44	78.32	5
107	114	35	1	PA	USD	348.55	222.32	73.81	99.41	1
108	115	47	12	XX	USD	602.64	250.90	125.99	120.22	10
109	116	23	12	VT	USD	816.52	233.00	98.00	196.39	4
110	117	57	6	WA	USD	43.14	289.77	115.19	81.82	1
111	118	49	8	AL	USD	-51.81	273.09	74.79	83.01	4
112	119	76	1	AK	USD	365.79	217.12	58.61	NA	15
113	120	8	8	CO	EUR	222.67	189.21	64.20	100.95	5
114	121	30	2	OR	USD	446.7	67.61	13.85	40.32	20
115	122	51	6	CA	USD	-25	64.35	40.28	23.40	15
116	123	80	3	CT	USD	479.69	155.32	100.28	106.35	7
117	124	28	5	PR	USD	-25	103.58	18.38	29.25	8
118	125	19	8	DE	USD	0	255.80	49.53	64.36	4
119	126	45	6	GA	USD	831.67	223.32	85.16	77.38	5
120	127	6	4	GA	USD	360.1	NA	8.09	15.91	5
121	130	30	7	OR	USD	0	174.57	128.09	19.06	4
122	131	27	5	LA	USD	763.08	131.63	63.33	-6.85	8
123	132	77	4	CT	USD	-25	24.90	12.24	9.10	7
124	133	78	1	LA	USD	152.31	NA	65.87	45.97	10
125	134	85	12	SC	USD	301.83	236.43	143.16	186.30	1
126	135	19	0	IA	USD	191.01	140.44	61.75	120.73	4
127	136	2	2	NC	USD	445.18	NA	47.02	141.64	5
128	137	31	8	CT	USD	428.16	308.43	226.07	96.81	7
129	138	28	8	MA	USD	594.57	322.71	214.61	209.08	5
130	139	74	5	VT	USD	-88.1	311.37	215.30	206.58	8
131	140	19	5	MD	USD	701.31	292.02	102.61	232.53	8
132	141	60	4	MD	USD	394.33	75.19	21.87	11.86	8
133	142	82	1	VT	USD	980.51	399.70	136.78	316.47	20
134	143	73	7	AL	USD	1049.19	335.26	95.41	311.18	0
135	144	NA	8	AZ	USD	9999999	341.89	150.29	89.82	9
136	145	67	1	WV	USD	469.82	288.69	152.09	48.47	4
137	146	58	2	TX	USD	273.56	121.73	44.72	105.44	9
138	147	67	1		USD	93.39	305.19	61.04	38.13	15
139	149	44	12	MA	USD	0	99.79	51.68	13.55	6
140	150	42	8	ID	USD	519.49	138.13	63.15	34.12	4

141	151	51	12	MS	CAD	279.85	133.26	60.22	11.06	6
142	152	61	3	PA	USD	866.78	136.25	75.12	20.63	15
143	153	61	4	RI	USD	244.06	106.52	38.18	69.52	2
144	154	48	6	OH	USD	271.92	143.34	91.71	8.83	3
145	155	15	8	OK	USD	288.34	292.15	96.25	41.77	2
146	156	46	7	WI	CAD	172.83	191.59	100.21	123.31	6
147	157	58	4	AL	EUR		165.44	43.97	69.92	4
148	161	67	1	VT	USD	458.84	323.71	168.76	96.45	5
149	163	64	6	MI	USD	517.2	181.40	86.95	NA	7
150	164	65	3	ID	USD	125.16	9.45	4.33	4.09	7
151	166	90	6		USD	277.78	261.15	131.70	16.93	7
152	167	30	7	MD	USD	159.28	51.79	39.51	18.66	7
153	168	8	7	NH	USD	469.54	53.47	11.10	21.34	5
154	169	83	1	WI	USD	177.01	458.42	137.36	146.98	5
155	170	40	2	IA	CAD	638.23	523.44	115.95	131.59	0
156	171	84	2	WI	USD	831.7	154.98	138.23	78.54	2
157	172	31	6	CO	USD	0	25.42	10.64	NA	10
158	173	33	3	FL	USD	222.07	27.85	13.32	13.44	9
159	174	NA	1	LA	USD	339.24	392.50	157.53	298.49	20
160	175	39	1	AZ	USD	447.78	155.02	123.40	108.53	10
161	176	16	3	TX	USD	454.66	100.61	67.30	34.32	15
162	177	90	6	IN	USD		356.81	213.07	77.22	0
163	178	68	8	PR	USD	419.23	51.60	42.67	30.30	20
164	179	49	8		USD	102.96	167.14	100.07	-1.66	8
165	180	52	1	WY	USD	-28.99	240.29	63.20	49.79	10
166	181	53	7	AZ	USD	520.64	271.85	68.08	130.76	15
167	182	18	7	DE	EUR	191.02	282.43	198.83	107.13	3
168	183	47	4	AL	USD	476.38	378.96	146.60	16.33	7
169	184	47	8	VT	USD	575.32	167.92	47.37	NA	2
170	185	90	5	AK	USD	149.26	-73.48	NA	NA	4
171	187	52	1	TN	USD	730.91	231.92	42.75	76.35	10
172	188	66	8	NJ	USD	545.7	40.75	19.38	39.40	2
173	189	NA	0	OK	USD	629.39	356.17	127.45	181.37	15
174	190	16	12	ID	USD	430.66	222.54	136.85	NA	8
175	191	68	2	ME	CAD	472.25	347.08	157.12	NA	6
176	192	54	5	NY	USD	523.62	154.13	50.25	24.81	3
177	193	NA	12	DC	USD	281.21	102.73	12.20	NA	6
178	194	68	0	ID	USD	299.39	223.92	168.69	135.80	15
179	195	89	8	ID	USD	628.22	211.46	92.17	65.54	2
180	197	86	4	MT	USD	926.01	149.17	66.20	5.44	4
181	198	30	2	VA	USD	408.82	229.45	133.52	125.36	4
182	199	90	0	VT	USD	109.44	115.02	32.37	96.87	15
183	200	4	4	PA	EUR	663.37	318.27	129.63	51.01	3

184	201	77	3	DC	USD	394.14	198.25	109.50	31.98	1
185	202	52	7	KS	USD		167.71	51.24	37.59	4
186	203	17	0	WV	USD	-31.23	269.22	151.78	74.02	20
187	204	NA	5	WY	USD	418.79	308.26	99.96	145.00	8
188	205	54	0	SC	USD	337.2	207.84	107.22	-15.63	0
189	207	64	1	FL	USD	0	304.87	87.67	86.73	15
190	208	54	7	WI	USD	24.86	230.58	118.06	73.83	20
191	209	68	2	VT	USD	469.73	362.11	202.42	-22.92	7
192	210	70	12	MN	USD		164.08	80.31	58.32	15
193	211	3	0	MT	USD	914.78	283.95	90.40	269.16	2
194	212	50	4	WI	USD	503.33	448.04	191.17	3.50	1
195	213	16	3	NM	USD	330.42	130.88	99.86	36.73	4
196	215	38	1	AZ	USD	504.49	389.02	102.25	228.84	2
197	217	56	2	OH	USD	706.36	117.70	57.16	41.69	7
198	218	3	0	NC	USD	0	421.68	254.45	39.58	8
199	219	42	2	PA	USD	338.39	358.79	209.13	225.76	15
200	220	51	7	NJ	USD	13.88	390.23	187.16	-7.62	6
201	221	85	5	FL	USD	526.39	300.39	178.98	75.05	9
202	222	21	4	MS	USD	458.28	143.68	66.84	44.52	15
203	224	26	4	WI	USD	296.91	415.10	152.52	NA	2
204	225	28	1	AZ	CAD	157.65	384.96	178.98	58.50	8
205	226	38	1	WI	USD	439.92	129.28	61.36	NA	10
206	227	5	8	SD	USD	554.18	651.01	341.93	134.28	10
207	228	51	12	CO	USD	407.43	155.91	40.42	178.87	6
208	230	83	1	WV	USD	413.68	239.02	88.30	118.91	7
209	231	74	12	MS	USD	206.1	416.60	171.82	100.48	0
210	232	76	3	SC	USD	702.22	354.07	110.85	206.87	0
211	233	77	1	WI	USD	614.19	182.34	159.21	62.81	8
212	234	25	6	RI	USD	503.78	248.54	40.31	97.45	4
213	235	45	12	WA	USD	99.47	168.89	66.95	149.65	1
214	236	37	0	IA	USD	194.13	122.03	83.11	58.90	4
215	237	83	2	IA	USD	440.67	310.17	108.71	40.97	10
216	238	71	4	VT	USD	553.9	305.57	70.33	73.73	1
217	239	72	5	AZ	CAD	249.49	239.71	189.22	154.34	15
218	240	4	6	NY	USD		164.02	141.08	89.56	15
219	241	16	8	TX	USD	570.85	104.97	38.44	55.39	6
220	242	NA	5	ID	USD	334.96	NA	52.73	126.68	2
221	243	78	12	SD	USD	372.02	14.68	5.78	NA	5
222	244	77	8	WA	USD	449.52	25.66	6.59	NA	4
223	245	16	7	DC	USD	695.13	285.61	136.73	109.09	8
224	246	18	12	XX	USD	829.06	353.68	256.65	179.00	8
225	247	48	6	OR	USD	677.2	193.65	100.05	115.74	2
226	248	54	1	LA	USD	645.8	502.64	231.80	93.19	3

227	249	23	6	AR	USD	531.27	191.46	91.23	6.84	4
228	250	82	6	LA	USD	318.53	301.81	106.93	235.96	8
229	252	45	1	WV	USD	326.43	607.20	221.89	NA	0
230	253	26	2	WV	CAD	283.38	187.75	80.25	67.31	7
231	254	58	12	DC	USD	171.17	500.33	335.12	190.55	20
232	255	81	7	PA	USD	191.14	62.87	22.55	49.63	2
233	256	48	7	ND	USD	-101.25	NA	63.64	57.85	6
234	257	32	4	KS	USD	285.71	292.33	59.80	208.02	8
235	258	44	8	RI	USD	619.39	110.08	24.62	109.15	1
236	259	16	12	GA	USD	687.45	191.32	69.49	163.87	5
237	260	88	0	MS	USD	393.96	238.79	127.89	99.20	20
238	261	81	2	WY	USD	284.15	320.24	192.50	24.91	8
239	262	44	2	IN	CAD	346.25	345.89	83.07	79.68	7
240	263	44	0	CT	USD	804.71	90.24	19.23	NA	4
241	264	87	12	OK	USD	0	119.05	40.16	24.78	9
242	265	40	8	PR	USD	405.04	286.88	111.79	198.15	20
243	266	57	1	UT	USD	797.61	12.54	5.53	4.63	15
244	267	76	12	XX	USD	180.68	358.72	172.49	-15.50	9
245	268	89	7	NY	USD	298.1	439.47	88.47	NA	0
246	269	50	5	DC	USD	643.22	296.92	234.98	109.39	1
247	270	36	12	MS	USD	0	154.38	90.03	23.13	20
248	271	84	12	KY	USD	84.8	361.79	210.05	-5.97	9
249	272	35	8	KS	USD	350.91	85.39	74.27	33.44	6
250	273	54	8	MN	USD	524.49	150.62	32.44	NA	20
251	274	85	8	TN	USD	394.3	209.59	115.50	123.41	8
252	275	83	4	AR	USD	0	61.57	22.05	54.88	4
253	276	79	0	MN	USD	455.1	190.14	106.51	123.85	4
254	277	78	12	VA	CAD	864.72	407.74	283.14	122.20	9
255	278	55	12	MA	USD	566.29	360.76	186.40	96.85	15
256	279	34	12	LA	USD	70.29	270.44	NA	62.93	20
257	280	47	3	ND	USD	316.39	517.47	111.68	NA	2
258	281	29	5	KY	USD	635.88	658.51	238.98	NA	6
259	283	6	0	NC	CAD	434.91	162.45	55.88	97.09	9
260	284	NA	5	AK	USD	423.81	15.22	9.26	3.81	3
261	286	33	1	AZ	USD	0	387.72	106.12	NA	2
262	287	59	7	CA	EUR	469.17	37.74	13.59	10.23	0

	Alcohol	Assistance
1	Yes	None
2	Unknown	SNAP
3	Yes	None
4	Unknown	None
5	Yes	None
6	Unknown	Food Pantry

7	Yes	SNAP
8	No	None
9		WIC
10	Unknown	None
11	No	None
12	No	None
13	Yes	SNAP
14	No	None
15	Yes	SNAP
16	N	None
17	Unknown	None
18	No	None
19	Unknown	Food Pantry
20	No	None
21	No	Food Pantry
22	No	None
23	Unknown	None
24	Yes	School Meals
25	No	None
26	Yes	Food Pantry
27	Yes	None
28	No	None
29	Yes	None
30	Yes	Food Pantry
31	Yes	SNAP
32	No	None
33	No	None
34	Yes	SNAP
35	Yes	None
36	Unknown	WIC
37		None
38	N	None
39	Yes	None
40	Yes	?
41		None
42	No	SNAP
43	Y	None
44		None
45		None
46	Yes	None
47	No	None
48	No	None
49	No	None

50	Yes	None
51	N	None
52	No	SNAP
53	Yes	None
54	No	None
55	Yes	Food Pantry
56	Yes	None
57	No	None
58	No	School Meals
59	Yes	None
60	Yes	None
61	No	School Meals
62	Yes	None
63	Y	None
64	Yes	None
65	Yes	None
66	Yes	None
67	No	None
68	Unknown	Food Pantry
69	No	None
70	Yes	?
71	Yes	None
72	Yes	?
73	N	WIC
74	Yes	None
75	No	None
76	Yes	None
77	Yes	None
78	Yes	School Meals
79	No	None
80	No	Food Pantry
81		None
82	No	WIC
83	No	None
84	Yes	None
85	Yes	None
86	Yes	Food Pantry
87	Unknown	None
88	No	School Meals
89	Unknown	None
90	N	None
91	Yes	None
92	No	School Meals

93	Unknown	None
94	No	None
95	Y	None
96	Unknown	None
97		?
98	No	None
99	Unknown	SNAP
100		None
101	Unknown	None
102		SNAP
103	No	None
104	Unknown	Food Pantry
105	No	None
106	No	None
107	No	None
108	No	School Meals
109	No	None
110	No	None
111	No	None
112	Yes	None
113	Y	None
114	Yes	None
115	No	None
116		None
117	No	None
118	Unknown	None
119	Yes	None
120	Unknown	?
121	No	None
122	No	Food Pantry
123	No	SNAP
124	Yes	None
125	Yes	None
126	No	None
127	Unknown	?
128	No	None
129	No	None
130	Yes	None
131	No	None
132	N	Food Pantry
133	Yes	None
134	Yes	None
135	No	None

136	Yes	None
137	Yes	School Meals
138	Y	None
139	No	None
140	Yes	None
141	Yes	None
142	No	None
143	Yes	None
144	N	Food Pantry
145	Unknown	None
146	Unknown	SNAP
147	No	None
148	Y	None
149	Yes	None
150	No	SNAP
151	N	None
152	No	SNAP
153	Unknown	None
154	No	None
155	N	None
156	No	School Meals
157		None
158	Yes	SNAP
159	Yes	None
160	Yes	WIC
161	No	None
162	Yes	None
163	Yes	Food Pantry
164	No	None
165		?
166		None
167	No	None
168	No	None
169	Yes	None
170	Yes	Food Pantry
171	No	WIC
172	Yes	?
173	No	None
174	No	None
175	Yes	None
176	N	Food Pantry
177	No	None
178		None

179	No	None
180	No	School Meals
181	No	WIC
182	N	None
183	Unknown	None
184	Yes	None
185	No	Food Pantry
186	No	None
187	Yes	None
188	No	None
189	No	None
190	Yes	None
191	Yes	?
192		None
193	Yes	None
194	N	None
195	No	None
196	Yes	None
197	Yes	None
198	No	SNAP
199	No	None
200	Yes	School Meals
201	No	Food Pantry
202	Yes	None
203		SNAP
204	No	None
205	Unknown	None
206	Yes	None
207	No	None
208	No	WIC
209	Yes	None
210	N	SNAP
211	No	WIC
212	No	None
213	No	None
214	Yes	None
215	No	None
216	Y	SNAP
217	Yes	Food Pantry
218	No	None
219	No	Food Pantry
220	No	None
221	Yes	SNAP

```
222     Yes         None
223     Yes         None
224     Yes         None
225     Yes         SNAP
226      No            ?
227                 None
228     Yes         None
229     Yes         None
230     Yes  Food Pantry
231      No         None
232     Yes         None
233     Yes         None
234       N         None
235     Yes         None
236 Unknown         None
237      No         None
238      No         None
239     Yes         None
240     Yes         None
241     Yes         None
242      No         SNAP
243      No         None
244      No          WIC
245      No         None
246     Yes         None
247      No School Meals
248     Yes         None
249     Yes         None
250     Yes         SNAP
251     Yes  Food Pantry
252 Unknown         None
253      No         SNAP
254     Yes         None
255     Yes         None
256     Yes         SNAP
257      No         None
258      No         SNAP
259 Unknown         None
260     Yes         None
261 Unknown         None
262     Yes         None
```

c. Restrict the data to those paying in US dollars (USD). Show that it worked by confirming the number of observations before and after restricting the data.

```
# The number of obervations before
obs <- nrow(df_food)
obs
```

```
[1] 262
```

```
# Restrict the data
df_food <- subset(df_food, Currency == "USD")

# The number of obervations after
obs_after <- nrow(df_food)
obs_after
```

```
[1] 230
```

It worked, since the observations becomes 230 from 262.

There are a number of issues with this data, likely due to the self-reported nature. For each of the following variables, clean them by removing any row with inappropriate data. For each variable, explain your rules for eliminating rows. For example, for the age variable, you might state "Excluded all minors under the age of 18". (Note that there is no "right" answer here, the goal is to i) choose reasonable rules and ii) carry out the corresponding code.)

I will also drop all na values.

d. The variable related to age.

My rule: Excluded all minors under the age of 18 and all seniors above the age of 100.

```
df_food <- subset(df_food, !is.na(Age) & Age >= 18 & Age <= 100)
```

e. The variable related to state.

My rule: Excluded all that are not US states.

```
states <- c(
  "AL","AK","AZ","AR","CA","CO","CT","DE","FL","GA","HI","ID","IL","IN","IA",
  "KS","KY","LA","ME","MD","MA","MI","MN","MS","MO","MT","NE","NV","NH","NJ",
  "NM","NY","NC","ND","OH","OK","OR","PA","RI","SC","SD","TN","TX","UT","VT",
  "VA","WA","WV","WI","WY"
)
df_food <- subset(df_food, !is.na(State) & State %in% states)
```

f. The four variables related to food expenditures.

My rule: Excluded all that are negative or zero.

```
df_food <- subset(df_food,
              !is.na(Total_exp)   & Total_exp   > 0 &
              !is.na(Grocery_exp) & Grocery_exp > 0 &
              !is.na(Dine_exp)    & Dine_exp    > 0 &
              !is.na(Misc)        & Misc        > 0)
```

g. The variable related to number of times dining out.

My rule: Excluded all that are more than 7.

```
df_food <- subset(df_food, !is.na(Dine_freq) & Dine_freq <= 7)
```

h. Report your final number of observations after this cleaning.

```
nrow(df_food)
```

```
[1] 66
```

My final number of observations after this cleaning is 66.

Problem 3 - Collatz conjecture

a. Write function nextCollatz that given a positive integer, computes the next number in its Collatz sequence. Be sure to provide a reasonable error on an invalid input. Be sure to document your function (see instructions above).

Input: A positive integer Output: A positive integer

```
#' Function to compute the next number in the Collatz sequence
#'
#' @param x a positive integer
#'
#' @return The next number in its Collatz sequence for `x`
nextCollatz <- function(x) {
  if (x %% 2 == 0) { # x is even
    return(x/2)
  }
  if (x %% 2 != 0) { # x is odd
```

```r
    return(3*x + 1)
  }
  if (is.na(x) || !is.numeric(x) || x <= 0) {
    stop("The input must be a positive integer.")
  }
}
```

```r
# Reproducing the examples
nextCollatz(5)
```

```
[1] 16
```

```r
nextCollatz(16)
```

```
[1] 8
```

b. Create a function collatzSequence that returns the Collatz sequence for a given input. Use your nextCollatz function to perform the calculation. Be sure to provide a reasonable error on an invalid input. Be sure to document your function (see instructions above).

Input: A positive integer Output: A list containing the vector of the entries in the Collatz sequence, beginning at the input and ending at 1; and the length of the Collatz sequence.

```r
#' Function that returns the Collatz sequence
#'
#' @param y a positive integer
#'
#' @return A list containing the vector of the entries in the Collatz sequence
collatzSequence <- function(y) {
  if (is.na(y) || !is.numeric(y) || y <= 0) {
    stop("The input must be a positive integer.")
  }
  collatz_seq <- y
  while (y != 1) {
    y <- nextCollatz(y)
    collatz_seq <- c(collatz_seq, y) # Source: Here I first got an error by mistakenly using
  }
  return(collatz_seq)
}
```

```r
# Reproducing the examples
collatzSequence(5)
```

```
[1]  5 16  8  4  2  1
```

```r
collatzSequence(19)
```

```
 [1] 19 58 29 88 44 22 11 34 17 52 26 13 40 20 10  5 16  8  4  2  1
```

c. Use these functions to find the shortest and longest Collatz sequence starting with values between 100 and 500, inclusive. In the case of ties, report the lowest starting value.

```r
# Starting with values between 100 and 500
starting_value <- 100:500
# The length list
lens <- numeric(0)
for (i in starting_value) {
  lens <- c(lens, length(collatzSequence(i)))
}

max(lens)
```

```
[1] 144
```

```r
min(lens)
```

```
[1] 8
```

```r
starting_value[which.max(lens)]
```

```
[1] 327
```

```r
starting_value[which.min(lens)]
```

```
[1] 128
```

Therefore, I got the longest sequence length is 144 (starting value is 327) and the shortest is 8 (starting value is 128).