Module 1 - Data Maniplulation

Importing Data

Here we read csv files `lowbwt_Low.csv` and `lowbwt_Normal.csv`

You will need to change the working directory to your personal file location.

```
# read and name data
low_birth = read.csv("./lowbwt_Low.csv")
norm_birth = read.csv("./lowbwt_Normal.csv")
```

Examine Data Attributes

```
# Variable names
names(low_birth)
```

```
[1] "id"    "smoke" "age"
```

```
# Data dimension: rows x columns; here: 59 rows and 3 columns
dim(low_birth)
```

```
[1] 59  3
```

```
# Number of rows
nrow(low_birth)
```

```
[1] 59
```

```r
# Number of columns
ncol(low_birth)
```

```
[1] 3
```

```r
# Head and Tail observations
head(low_birth)
```

```
  id smoke age
1 31     0  20
2 76     0  20
3 44     1  20
4 68     1  17
5 23     1  19
6 45     1  17
```

```r
tail(low_birth)
```

```
   id smoke age
54 19     0  24
55 11     1  34
56 56     1  31
57 65     1  30
58 10     0  29
59 22     1  32
```

```r
# Check for number of missing values
sum(is.na(low_birth))
```

```
[1] 0
```

```r
# Examine the classes of each column
str(low_birth)
```

```
'data.frame':   59 obs. of  3 variables:
 $ id   : int  31 76 44 68 23 45 51 49 71 83 ...
 $ smoke: int  0 0 1 1 1 1 1 0 0 0 ...
 $ age  : int  20 20 20 17 19 17 20 18 17 17 ...
```

```
# Tabulate variable smoke
table(low_birth$smoke)
```

```
 0  1
29 30
```

Data Manipulation using `dplyr`

Note: to apply these changes to the existing data, you must reassign the change.

i.e. `low_birth = filter(low_birth, age < 20)`

NOTE: you will need to install the `{tidyverse}` package. Run `install.packages("tidyverse")` in the Console.

```
# install and load tidyverse (contains dplyr)
# install.packages("tidyverse")
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.0      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
# Select only column/variable age
dplyr::select(low_birth, age)
```

	age
1	20
2	20
3	20
4	17
5	19
6	17
7	20
8	18
9	17
10	17
11	18
12	20
13	19
14	20
15	20
16	20
17	14
18	17
19	19
20	15
21	15
22	16
23	14
24	23
25	21
26	23
27	27
28	24
29	23
30	26
31	22
32	25
33	21
34	25
35	26
36	21
37	22
38	28
39	27
40	23
41	24
42	21

```
43  25
44  24
45  23
46  24
47  25
48  25
49  21
50  28
51  26
52  25
53  26
54  24
55  34
56  31
57  30
58  29
59  32
```

```r
# Keep only rows where 'age' is less than 20
filter(low_birth, age < 20)
```

```
   id smoke age
1  68     1  17
2  23     1  19
3  45     1  17
4  49     0  18
5  71     0  17
6  83     0  17
7  50     1  18
8  33     0  19
9  78     1  14
10 37     1  17
11 34     1  19
12 57     0  15
13 62     0  15
14 25     0  16
15 81     0  14
```

```r
# Select rows that contain missing data
filter(low_birth, is.na(age))
```

```
[1] id    smoke age
<0 rows> (or 0-length row.names)
```

```r
# Remove column age
dplyr::select(low_birth, -age)
```

```
   id smoke
1  31     0
2  76     0
3  44     1
4  68     1
5  23     1
6  45     1
7  51     1
8  49     0
9  71     0
10 83     0
11 50     1
12 27     1
13 33     0
14 47     0
15 40     1
16 60     1
17 78     1
18 37     1
19 34     1
20 57     0
21 62     0
22 25     0
23 81     0
24 17     0
25 20     1
26 82     1
27 43     0
28 61     1
29 59     1
30 77     1
31 42     1
32 13     0
33 30     0
34 26     1
```

```
35 35     1
36 28     0
37 67     1
38  4     1
39 16     0
40 69     1
41 29     1
42 84     1
43 32     0
44 36     0
45 63     0
46 18     0
47 46     0
48 15     0
49 52     0
50 79     1
51 54     0
52 24     0
53 75     0
54 19     0
55 11     1
56 56     1
57 65     1
58 10     0
59 22     1
```

```r
# Filter rows: select all 25+ yrs old, smokers
filter(low_birth, age > 25 & smoke == "1")
```

```
  id smoke age
1 77     1  26
2 35     1  26
3  4     1  28
4 79     1  28
5 11     1  34
6 56     1  31
7 65     1  30
8 22     1  32
```

```r
# Ordering data by variable/column 'id'
arrange(low_birth, id)
```

```
   id smoke age
1   4     1  28
2  10     0  29
3  11     1  34
4  13     0  25
5  15     0  25
6  16     0  27
7  17     0  23
8  18     0  24
9  19     0  24
10 20     1  21
11 22     1  32
12 23     1  19
13 24     0  25
14 25     0  16
15 26     1  25
16 27     1  20
17 28     0  21
18 29     1  24
19 30     0  21
20 31     0  20
21 32     0  25
22 33     0  19
23 34     1  19
24 35     1  26
25 36     0  24
26 37     1  17
27 40     1  20
28 42     1  22
29 43     0  27
30 44     1  20
31 45     1  17
32 46     0  25
33 47     0  20
34 49     0  18
35 50     1  18
36 51     1  20
37 52     0  21
```

```
38 54     0  26
39 56     1  31
40 57     0  15
41 59     1  23
42 60     1  20
43 61     1  24
44 62     0  15
45 63     0  23
46 65     1  30
47 67     1  22
48 68     1  17
49 69     1  23
50 71     0  17
51 75     0  26
52 76     0  20
53 77     1  26
54 78     1  14
55 79     1  28
56 81     0  14
57 82     1  23
58 83     0  17
59 84     1  21
```

```
# Arrange by id in descending order
arrange(low_birth, desc(id))
```

```
   id smoke age
1  84     1  21
2  83     0  17
3  82     1  23
4  81     0  14
5  79     1  28
6  78     1  14
7  77     1  26
8  76     0  20
9  75     0  26
10 71     0  17
11 69     1  23
12 68     1  17
13 67     1  22
14 65     1  30
```

15	63	0	23
16	62	0	15
17	61	1	24
18	60	1	20
19	59	1	23
20	57	0	15
21	56	1	31
22	54	0	26
23	52	0	21
24	51	1	20
25	50	1	18
26	49	0	18
27	47	0	20
28	46	0	25
29	45	1	17
30	44	1	20
31	43	0	27
32	42	1	22
33	40	1	20
34	37	1	17
35	36	0	24
36	35	1	26
37	34	1	19
38	33	0	19
39	32	0	25
40	31	0	20
41	30	0	21
42	29	1	24
43	28	0	21
44	27	1	20
45	26	1	25
46	25	0	16
47	24	0	25
48	23	1	19
49	22	1	32
50	20	1	21
51	19	0	24
52	18	0	24
53	17	0	23
54	16	0	27
55	15	0	25
56	13	0	25
57	11	1	34

```
58 10     0  29
59  4     1  28
```

```r
# Order by multiple columns/variables
arrange(low_birth, smoke, desc(age))
```

```
   id smoke age
1  10     0  29
2  43     0  27
3  16     0  27
4  54     0  26
5  75     0  26
6  13     0  25
7  32     0  25
8  46     0  25
9  15     0  25
10 24     0  25
11 36     0  24
12 18     0  24
13 19     0  24
14 17     0  23
15 63     0  23
16 30     0  21
17 28     0  21
18 52     0  21
19 31     0  20
20 76     0  20
21 47     0  20
22 33     0  19
23 49     0  18
24 71     0  17
25 83     0  17
26 25     0  16
27 57     0  15
28 62     0  15
29 81     0  14
30 11     1  34
31 22     1  32
32 56     1  31
33 65     1  30
34  4     1  28
```

```
35 79      1   28
36 77      1   26
37 35      1   26
38 26      1   25
39 61      1   24
40 29      1   24
41 82      1   23
42 59      1   23
43 69      1   23
44 42      1   22
45 67      1   22
46 20      1   21
47 84      1   21
48 44      1   20
49 51      1   20
50 27      1   20
51 40      1   20
52 60      1   20
53 23      1   19
54 34      1   19
55 50      1   18
56 68      1   17
57 45      1   17
58 37      1   17
59 78      1   14
```

```r
# Rename variable 'smoke' to 'Smoking_Status'
rename(low_birth, Smoking_Status = smoke)
```

```
   id Smoking_Status age
1  31              0  20
2  76              0  20
3  44              1  20
4  68              1  17
5  23              1  19
6  45              1  17
7  51              1  20
8  49              0  18
9  71              0  17
10 83              0  17
11 50              1  18
```

12	27	1	20
13	33	0	19
14	47	0	20
15	40	1	20
16	60	1	20
17	78	1	14
18	37	1	17
19	34	1	19
20	57	0	15
21	62	0	15
22	25	0	16
23	81	0	14
24	17	0	23
25	20	1	21
26	82	1	23
27	43	0	27
28	61	1	24
29	59	1	23
30	77	1	26
31	42	1	22
32	13	0	25
33	30	0	21
34	26	1	25
35	35	1	26
36	28	0	21
37	67	1	22
38	4	1	28
39	16	0	27
40	69	1	23
41	29	1	24
42	84	1	21
43	32	0	25
44	36	0	24
45	63	0	23
46	18	0	24
47	46	0	25
48	15	0	25
49	52	0	21
50	79	1	28
51	54	0	26
52	24	0	25
53	75	0	26
54	19	0	24

```
55 11              1  34
56 56              1  31
57 65              1  30
58 10              0  29
59 22              1  32
```

```r
# Create a variable for log of 'age'
mutate(low_birth, log_age = log(age))
```

```
   id smoke age  log_age
1  31     0  20 2.995732
2  76     0  20 2.995732
3  44     1  20 2.995732
4  68     1  17 2.833213
5  23     1  19 2.944439
6  45     1  17 2.833213
7  51     1  20 2.995732
8  49     0  18 2.890372
9  71     0  17 2.833213
10 83     0  17 2.833213
11 50     1  18 2.890372
12 27     1  20 2.995732
13 33     0  19 2.944439
14 47     0  20 2.995732
15 40     1  20 2.995732
16 60     1  20 2.995732
17 78     1  14 2.639057
18 37     1  17 2.833213
19 34     1  19 2.944439
20 57     0  15 2.708050
21 62     0  15 2.708050
22 25     0  16 2.772589
23 81     0  14 2.639057
24 17     0  23 3.135494
25 20     1  21 3.044522
26 82     1  23 3.135494
27 43     0  27 3.295837
28 61     1  24 3.178054
29 59     1  23 3.135494
30 77     1  26 3.258097
31 42     1  22 3.091042
```

```
32 13     0  25 3.218876
33 30     0  21 3.044522
34 26     1  25 3.218876
35 35     1  26 3.258097
36 28     0  21 3.044522
37 67     1  22 3.091042
38  4     1  28 3.332205
39 16     0  27 3.295837
40 69     1  23 3.135494
41 29     1  24 3.178054
42 84     1  21 3.044522
43 32     0  25 3.218876
44 36     0  24 3.178054
45 63     0  23 3.135494
46 18     0  24 3.178054
47 46     0  25 3.218876
48 15     0  25 3.218876
49 52     0  21 3.044522
50 79     1  28 3.332205
51 54     0  26 3.258097
52 24     0  25 3.218876
53 75     0  26 3.258097
54 19     0  24 3.178054
55 11     1  34 3.526361
56 56     1  31 3.433987
57 65     1  30 3.401197
58 10     0  29 3.367296
59 22     1  32 3.465736
```

```
# Centering the data by subtracting the mean from variable 'age'
mutate(low_birth, center_age = age - mean(age))
```

```
   id smoke age center_age
1  31     0  20 -2.3050847
2  76     0  20 -2.3050847
3  44     1  20 -2.3050847
4  68     1  17 -5.3050847
5  23     1  19 -3.3050847
6  45     1  17 -5.3050847
7  51     1  20 -2.3050847
8  49     0  18 -4.3050847
```

```
9  71     0  17 -5.3050847
10 83     0  17 -5.3050847
11 50     1  18 -4.3050847
12 27     1  20 -2.3050847
13 33     0  19 -3.3050847
14 47     0  20 -2.3050847
15 40     1  20 -2.3050847
16 60     1  20 -2.3050847
17 78     1  14 -8.3050847
18 37     1  17 -5.3050847
19 34     1  19 -3.3050847
20 57     0  15 -7.3050847
21 62     0  15 -7.3050847
22 25     0  16 -6.3050847
23 81     0  14 -8.3050847
24 17     0  23  0.6949153
25 20     1  21 -1.3050847
26 82     1  23  0.6949153
27 43     0  27  4.6949153
28 61     1  24  1.6949153
29 59     1  23  0.6949153
30 77     1  26  3.6949153
31 42     1  22 -0.3050847
32 13     0  25  2.6949153
33 30     0  21 -1.3050847
34 26     1  25  2.6949153
35 35     1  26  3.6949153
36 28     0  21 -1.3050847
37 67     1  22 -0.3050847
38  4     1  28  5.6949153
39 16     0  27  4.6949153
40 69     1  23  0.6949153
41 29     1  24  1.6949153
42 84     1  21 -1.3050847
43 32     0  25  2.6949153
44 36     0  24  1.6949153
45 63     0  23  0.6949153
46 18     0  24  1.6949153
47 46     0  25  2.6949153
48 15     0  25  2.6949153
49 52     0  21 -1.3050847
50 79     1  28  5.6949153
51 54     0  26  3.6949153
```

```
52 24     0  25  2.6949153
53 75     0  26  3.6949153
54 19     0  24  1.6949153
55 11     1  34 11.6949153
56 56     1  31  8.6949153
57 65     1  30  7.6949153
58 10     0  29  6.6949153
59 22     1  32  9.6949153
```

```
# Use case_when function to create new age categories
# Cat 1: Age < 25; Cat 2: 25 < Age < 30. Cat 3: Age > 30
mutate(low_birth, new_age = case_when(age < 25 ~ 1,
                                      age >= 25 & age < 30 ~ 2,
                                      age > 30 ~ 3))
```

```
   id smoke age new_age
1  31     0  20       1
2  76     0  20       1
3  44     1  20       1
4  68     1  17       1
5  23     1  19       1
6  45     1  17       1
7  51     1  20       1
8  49     0  18       1
9  71     0  17       1
10 83     0  17       1
11 50     1  18       1
12 27     1  20       1
13 33     0  19       1
14 47     0  20       1
15 40     1  20       1
16 60     1  20       1
17 78     1  14       1
18 37     1  17       1
19 34     1  19       1
20 57     0  15       1
21 62     0  15       1
22 25     0  16       1
23 81     0  14       1
24 17     0  23       1
25 20     1  21       1
```

```
26 82     1  23       1
27 43     0  27       2
28 61     1  24       1
29 59     1  23       1
30 77     1  26       2
31 42     1  22       1
32 13     0  25       2
33 30     0  21       1
34 26     1  25       2
35 35     1  26       2
36 28     0  21       1
37 67     1  22       1
38  4     1  28       2
39 16     0  27       2
40 69     1  23       1
41 29     1  24       1
42 84     1  21       1
43 32     0  25       2
44 36     0  24       1
45 63     0  23       1
46 18     0  24       1
47 46     0  25       2
48 15     0  25       2
49 52     0  21       1
50 79     1  28       2
51 54     0  26       2
52 24     0  25       2
53 75     0  26       2
54 19     0  24       1
55 11     1  34       3
56 56     1  31       3
57 65     1  30      NA
58 10     0  29       2
59 22     1  32       3
```

Combine Data Sets

```
# stack low_birth & norm_birth
low_and_norm = rbind(low_birth, norm_birth)

# combine by specific variable
```

```r
admin_birth = read.csv("./lowbwt_Admin.csv")
birth_final = full_join(admin_birth, low_and_norm, by = "id")

# export data
write.csv(birth_final, file = "./birth_final.csv")
```