Objective

The objective of this session is to get familiar with 'dplyr' library which makes woring with dataframes a lot easier and clean.

Key takeaways

Exploring the below functions in dplyr

1. select() - select columns
2. filter() - filter rows
3. arrange() - re-order or arrange rows
4. mutate() - create new columns
5. summarise() - summarise values
6. group_by() - allows for group operations in the "split-apply-combine" concept

'dplyr' Package

Read the dataset.

```
msleep <- read.csv("msleep_ggplot2.csv")
head(msleep, 5)

##                          name       genus  vore        order conservation
## 1                    Cheetah   Acinonyx carni    Carnivora           lc
## 2                 Owl monkey      Aotus  omni     Primates         <NA>
## 3            Mountain beaver Aplodontia herbi     Rodentia           nt
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
## 5                        Cow        Bos herbi Artiodactyla domesticated
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt
## 1        12.1        NA          NA  11.9      NA  50.000
## 2        17.0       1.8          NA   7.0 0.01550   0.480
## 3        14.4       2.4          NA   9.6      NA   1.350
## 4        14.9       2.3   0.1333333   9.1 0.00029   0.019
## 5         4.0       0.7   0.6666667  20.0 0.42300 600.000

dim(msleep)

## [1] 83 11

# install.packages("dplyr")
library(dplyr)

##
## 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
```

Function1: select

```
sleepData <- select(msleep, name, sleep_total)
head(sleepData)
```

```
##                          name sleep_total
## 1                    Cheetah        12.1
## 2                 Owl monkey        17.0
## 3            Mountain beaver        14.4
## 4 Greater short-tailed shrew        14.9
## 5                        Cow         4.0
## 6           Three-toed sloth        14.4
```

```
# Select all, except specific solumns
head(select(msleep, -name))
```

```
##        genus  vore        order conservation sleep_total sleep_rem
## 1   Acinonyx carni    Carnivora           lc        12.1        NA
## 2      Aotus  omni     Primates         <NA>        17.0       1.8
## 3 Aplodontia herbi     Rodentia           nt        14.4       2.4
## 4    Blarina  omni Soricomorpha           lc        14.9       2.3
## 5        Bos herbi Artiodactyla domesticated         4.0       0.7
## 6   Bradypus herbi       Pilosa         <NA>        14.4       2.2
##   sleep_cycle awake brainwt  bodywt
## 1          NA  11.9      NA  50.000
## 2          NA   7.0 0.01550   0.480
## 3          NA   9.6      NA   1.350
## 4   0.1333333   9.1 0.00029   0.019
## 5   0.6666667  20.0 0.42300 600.000
## 6   0.7666667   9.6      NA   3.850
```

```
# Look at the column names
names(msleep)
```

```
##  [1] "name"         "genus"        "vore"         "order"
##  [5] "conservation" "sleep_total"  "sleep_rem"    "sleep_cycle"
##  [9] "awake"        "brainwt"      "bodywt"
```

```
# Select a series of columns
head(select(msleep, name:order))
```

```
##                          name      genus  vore        order
## 1                    Cheetah   Acinonyx carni    Carnivora
## 2                 Owl monkey      Aotus  omni     Primates
## 3            Mountain beaver Aplodontia herbi     Rodentia
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha
## 5                        Cow        Bos herbi Artiodactyla
## 6           Three-toed sloth   Bradypus herbi       Pilosa
```

```
# Select columns starting with specific names
head(select(msleep, starts_with("sl")))
```

```
##   sleep_total sleep_rem sleep_cycle
## 1        12.1        NA          NA
## 2        17.0       1.8          NA
## 3        14.4       2.4          NA
## 4        14.9       2.3   0.1333333
## 5         4.0       0.7   0.6666667
## 6        14.4       2.2   0.7666667
```

Function2: filter

```
# Select rows w.r.t a particular column values
f1 = filter(msleep, sleep_total >= 16)
head(f1, 5)
```

```
##                      name      genus    vore           order conservation
## 1              Owl monkey      Aotus    omni        Primates         <NA>
## 2    Long-nosed armadillo    Dasypus   carni       Cingulata           lc
## 3 North American Opossum  Didelphis    omni Didelphimorphia           lc
## 4           Big brown bat  Eptesicus insecti      Chiroptera           lc
## 5    Thick-tailed opposum Lutreolina   carni Didelphimorphia           lc
##   sleep_total sleep_rem sleep_cycle awake brainwt bodywt
## 1        17.0       1.8          NA   7.0  0.0155  0.480
## 2        17.4       3.1   0.3833333   6.6  0.0108  3.500
## 3        18.0       4.9   0.3333333   6.0  0.0063  1.700
## 4        19.7       3.9   0.1166667   4.3  0.0003  0.023
## 5        19.4       6.6          NA   4.6      NA  0.370
```

```
# Select rows which satifty multiple conditions
f2 = filter(msleep, sleep_total >= 16, bodywt >= 1)
head(f2, 5)
```

```
##                      name      genus    vore           order conservation
## 1    Long-nosed armadillo    Dasypus   carni       Cingulata           lc
## 2 North American Opossum  Didelphis    omni Didelphimorphia           lc
## 3         Giant armadillo Priodontes insecti       Cingulata           en
##   sleep_total sleep_rem sleep_cycle awake brainwt bodywt
## 1        17.4       3.1   0.3833333   6.6  0.0108    3.5
## 2        18.0       4.9   0.3333333   6.0  0.0063    1.7
## 3        18.1       6.1          NA   5.9  0.0810   60.0
```

```
# Select rows, according to specific values of an attribute
f3 = filter(msleep, order %in% c("Perissodactyla", "Primates"))
head(f3, 5)

##           name         genus  vore          order conservation sleep_total
## 1   Owl monkey         Aotus  omni       Primates         <NA>        17.0
## 2       Grivet Cercopithecus  omni       Primates           lc        10.0
## 3        Horse         Equus herbi Perissodactyla domesticated         2.9
## 4       Donkey         Equus herbi Perissodactyla domesticated         3.1
## 5 Patas monkey  Erythrocebus  omni       Primates           lc        10.9
##   sleep_rem sleep_cycle awake brainwt bodywt
## 1       1.8          NA   7.0  0.0155   0.48
## 2       0.7          NA  14.0      NA   4.75
## 3       0.6           1  21.1  0.6550 521.00
## 4       0.4          NA  20.9  0.4190 187.00
## 5       1.1          NA  13.1  0.1150  10.00
```

Function3: arrange

Arrange the rows according to specific order of an attribute

```
a1 = arrange(msleep, sleep_total)
head(a1, 5)

##          name         genus  vore          order conservation sleep_total
## 1     Giraffe       Giraffa herbi   Artiodactyla           cd         1.9
## 2 Pilot whale Globicephalus carni        Cetacea           cd         2.7
## 3       Horse         Equus herbi Perissodactyla domesticated         2.9
## 4    Roe deer     Capreolus herbi   Artiodactyla           lc         3.0
## 5      Donkey         Equus herbi Perissodactyla domesticated         3.1
##   sleep_rem sleep_cycle awake brainwt  bodywt
## 1       0.4          NA 22.10      NA 899.995
## 2       0.1          NA 21.35      NA 800.000
## 3       0.6           1 21.10  0.6550 521.000
## 4        NA          NA 21.00  0.0982  14.800
## 5       0.4          NA 20.90  0.4190 187.000

a2 = arrange(msleep, -sleep_total)
head(a2, 5)

##                     name      genus    vore           order conservation
## 1       Little brown bat     Myotis insecti      Chiroptera         <NA>
## 2          Big brown bat  Eptesicus insecti      Chiroptera           lc
## 3   Thick-tailed opposum Lutreolina   carni Didelphimorphia           lc
## 4        Giant armadillo Priodontes insecti       Cingulata           en
## 5 North American Opossum  Didelphis    omni Didelphimorphia           lc
##   sleep_total sleep_rem sleep_cycle awake brainwt bodywt
## 1        19.9       2.0   0.2000000   4.1 0.00025  0.010
## 2        19.7       3.9   0.1166667   4.3 0.00030  0.023
## 3        19.4       6.6          NA   4.6      NA  0.370
## 4        18.1       6.1          NA   5.9 0.08100 60.000
## 5        18.0       4.9   0.3333333   6.0 0.00630  1.700
```

```
a3 = arrange(msleep, order, sleep_total)
head(a3, 5)
```

```
##       name     genus  vore        order conservation sleep_total sleep_rem
## 1   Tenrec    Tenrec  omni Afrosoricida         <NA>        15.6       2.3
## 2  Giraffe   Giraffa herbi Artiodactyla           cd         1.9       0.4
## 3 Roe deer Capreolus herbi Artiodactyla           lc         3.0        NA
## 4    Sheep      Ovis herbi Artiodactyla domesticated         3.8       0.6
## 5      Cow       Bos herbi Artiodactyla domesticated         4.0       0.7
##   sleep_cycle awake brainwt  bodywt
## 1          NA   8.4  0.0026   0.900
## 2          NA  22.1      NA 899.995
## 3          NA  21.0  0.0982  14.800
## 4          NA  20.2  0.1750  55.500
## 5   0.6666667  20.0  0.4230 600.000
```

```
a4 = arrange(msleep, desc(order), sleep_total)
head(a4, 5)
```

```
##                         name     genus    vore        order conservation
## 1      Eastern american mole  Scalopus insecti Soricomorpha           lc
## 2  Lesser short-tailed shrew Cryptotis    omni Soricomorpha           lc
## 3            Star-nosed mole Condylura    omni Soricomorpha           lc
## 4                 Musk shrew    Suncus    <NA> Soricomorpha         <NA>
## 5 Greater short-tailed shrew   Blarina    omni Soricomorpha           lc
##   sleep_total sleep_rem sleep_cycle awake brainwt bodywt
## 1         8.4       2.1   0.1666667  15.6 0.00120  0.075
## 2         9.1       1.4   0.1500000  14.9 0.00014  0.005
## 3        10.3       2.2          NA  13.7 0.00100  0.060
## 4        12.8       2.0   0.1833333  11.2 0.00033  0.048
## 5        14.9       2.3   0.1333333   9.1 0.00029  0.019
```

Function4: mutate

```
# Create a new column using other columns
m1 = mutate(msleep, rem_proportion = sleep_rem / sleep_total)
head(m1, 5)
```

```
##                         name      genus  vore        order conservation
## 1                    Cheetah   Acinonyx carni    Carnivora           lc
## 2                 Owl monkey      Aotus  omni     Primates         <NA>
## 3            Mountain beaver Aplodontia herbi     Rodentia           nt
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
## 5                        Cow        Bos herbi Artiodactyla domesticated
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt rem_proportion
## 1        12.1        NA          NA  11.9      NA  50.000             NA
## 2        17.0       1.8          NA   7.0 0.01550   0.480      0.1058824
## 3        14.4       2.4          NA   9.6      NA   1.350      0.1666667
## 4        14.9       2.3   0.1333333   9.1 0.00029   0.019      0.1543624
## 5         4.0       0.7   0.6666667  20.0 0.42300 600.000      0.1750000

m2 = mutate(msleep, rem_proportion = sleep_rem / sleep_total,
       bodywt_grams = bodywt * 1000)
head(m2, 5)

##                        name      genus  vore        order conservation
## 1                   Cheetah   Acinonyx carni    Carnivora           lc
## 2                Owl monkey      Aotus  omni     Primates         <NA>
## 3           Mountain beaver Aplodontia herbi     Rodentia           nt
## 4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
## 5                       Cow        Bos herbi Artiodactyla domesticated
##   sleep_total sleep_rem sleep_cycle awake brainwt  bodywt rem_proportion
## 1        12.1        NA          NA  11.9      NA  50.000             NA
## 2        17.0       1.8          NA   7.0 0.01550   0.480      0.1058824
## 3        14.4       2.4          NA   9.6      NA   1.350      0.1666667
## 4        14.9       2.3   0.1333333   9.1 0.00029   0.019      0.1543624
## 5         4.0       0.7   0.6666667  20.0 0.42300 600.000      0.1750000
##   bodywt_grams
## 1        50000
## 2          480
## 3         1350
## 4           19
## 5       600000
```

Function5: summarise

```
names(msleep)

##  [1] "name"         "genus"        "vore"         "order"
##  [5] "conservation" "sleep_total"  "sleep_rem"    "sleep_cycle"
##  [9] "awake"        "brainwt"      "bodywt"

summarise(msleep, avg_sleep = mean(sleep_total))

##   avg_sleep
## 1  10.43373

summarise(msleep,
          avg_sleep = mean(sleep_total),
          min_sleep = min(sleep_total),
          max_sleep = max(sleep_total),
          total = n())

##   avg_sleep min_sleep max_sleep total
## 1  10.43373       1.9      19.9    83
```

Function6: Group_by

```
a = summarise(group_by(msleep, order),
          avg_sleep = mean(sleep_total),
          min_sleep = min(sleep_total),
          max_sleep = max(sleep_total),
          total = n())
```

Function7: Chaining

```
msleep %>%
  select(name, sleep_total) %>%
  head

##                         name sleep_total
## 1                    Cheetah        12.1
## 2                 Owl monkey        17.0
## 3            Mountain beaver        14.4
## 4 Greater short-tailed shrew        14.9
## 5                        Cow         4.0
## 6           Three-toed sloth        14.4

msleep %>%
  select(name, order, sleep_total) %>%
  arrange(order, sleep_total) %>%
  head

##       name        order sleep_total
## 1   Tenrec Afrosoricida        15.6
## 2  Giraffe Artiodactyla         1.9
## 3 Roe deer Artiodactyla         3.0
## 4    Sheep Artiodactyla         3.8
## 5      Cow Artiodactyla         4.0
## 6     Goat Artiodactyla         5.3

msleep %>%
  group_by(order) %>%
  summarise(avg_sleep = mean(sleep_total),
            min_sleep = min(sleep_total),
            max_sleep = max(sleep_total),
            total = n())

## Source: local data frame [19 x 5]
##
##              order avg_sleep min_sleep max_sleep total
##             (fctr)     (dbl)     (dbl)     (dbl) (int)
## 1     Afrosoricida 15.600000      15.6      15.6     1
## 2     Artiodactyla  4.516667       1.9       9.1     6
## 3        Carnivora 10.116667       3.5      15.8    12
## 4          Cetacea  4.500000       2.7       5.6     3
## 5       Chiroptera 19.800000      19.7      19.9     2
## 6        Cingulata 17.750000      17.4      18.1     2
## 7  Didelphimorphia 18.700000      18.0      19.4     2
## 8    Diprotodontia 12.400000      11.1      13.7     2
## 9   Erinaceomorpha 10.200000      10.1      10.3     2
## 10      Hyracoidea  5.666667       5.3       6.3     3
## 11      Lagomorpha  8.400000       8.4       8.4     1
## 12     Monotremata  8.600000       8.6       8.6     1
## 13  Perissodactyla  3.466667       2.9       4.4     3
## 14          Pilosa 14.400000      14.4      14.4     1
## 15        Primates 10.500000       8.0      17.0    12
## 16     Proboscidea  3.600000       3.3       3.9     2
## 17        Rodentia 12.468182       7.0      16.6    22
## 18      Scandentia  8.900000       8.9       8.9     1
## 19    Soricomorpha 11.100000       8.4      14.9     5
```

Excercise

Load mtcars dataset into R and solve the following using dplyr functions

1. Create a dataframe which has columns 'mpg', 'cyl', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'am', 'gear'
2. Subset the dataframe which has cars with mpg above 20
3. Create a new column which gives hp/wt ratio
4. What is the mean hp/wt ratio of manual and automatic transmission cars
5. Perform all the above steps using chaining operation