EEEB UN3005/GR5005
Lab - Week 02 - 04 and 06 February 2019

USE YOUR NAME HERE

Data Cleaning

To practice data cleaning, in this week's lab, we'll be using published data on RNA viruses collated by Mark Woolhouse and Liam Brierley. This dataset contains trait information gathered from the scientific literature on 214 RNA viruses that are known to infect humans. See the "Data Records" section of the published paper for information on the variables included in the dataset. I've downloaded the data and converted it to a CSV file for your ease of use. Find the data on either the class CourseWorks page or the class GitHub repository as `Woolhouse_and_Brierley_RNA_virus_database.csv`.

Exercise 1: Data Import

Download the Woolhouse and Brierley data, and import it into R, assigning it to an object named `viruses`. Run `summary()` on this object. You'll get a load of information in return, but this is just to familiarize yourself broadly with the dataset.

```r
setwd('D:/Codes/ClassCodes/')
viruses = read.csv('Woolhouse_and_Brierley_RNA_virus_database.csv')
summary(viruses)
```

```
##                                     Species                Genus
##  African green monkey simian foamy virus:  1   Flavivirus     : 27
##  Aichivirus A                           :  1   Alphavirus     : 21
##  Alagoas vesiculovirus                  :  1   Orthobunyavirus: 17
##  Alphacoronavirus 1                     :  1   Orthohantavirus: 14
##  Andes orthohantavirus                  :  1   Mammarenavirus : 11
##  Aroa virus                             :  1   Enterovirus    :  8
##  (Other)                                :208   (Other)        :116
##               Family      Envelope           Genome    ICTV.history.URL
##  Flaviviridae    :30   Min.   :0.0000   (-)ssRNA:100   Min.   :20160985
##  Picornaviridae  :23   1st Qu.:1.0000   (+)ssRNA: 91   1st Qu.:20161326
##  Togaviridae     :22   Median :1.0000   dsRNA   : 14   Median :20162158
##  Rhabdoviridae   :18   Mean   :0.7897   ssRNA-RT:  9   Mean   :20162511
##  Peribunyaviridae:17   3rd Qu.:1.0000                  3rd Qu.:20164116
##  Hantaviridae    :14   Max.   :1.0000                  Max.   :20165193
##  (Other)         :90
##  Discovery.year                   Reference..discovery.
##  Min.   :1901   Available from external site: 20
##  1st Qu.:1959   13691675                    :  5
##  Median :1974   19033469                    :  3
##  Mean   :1975   1124969                     :  2
##  3rd Qu.:1995   13504446                    :  2
##  Max.   :2015   15391667                    :  2
##                 (Other)                     :180
##  Serological.detection.only Vector   Inhalation Ingestion Sexual
##  N:178                      ? :  2   ? :  2     ? :  2    ?:  2
##  Y: 36                      0 :121   0 :145     0 :179    0:202
##                             1 : 90   1 : 47     1 : 24    1: 10
##                             1*:  1   1*: 20     1*:  9
##
##
##
##  Iatrogenic..inc..blood. Fomites  Broken.Skin Maternal Direct.Contact
##  ?:  2                   ? :  2   ? :  2      ?:  2    ? :  2
##  0:203                   0 :193   0 :187      0:184    0 :144
##  1:  9                   1 : 17   1 : 24      1: 28    1 : 53
##                          1*:  2   1*:  1               1*: 15
##
##
##
##  Reference..Transmission.route.   Ref2..TR.        Transmission.level
##  Min.   :    3981               Min.   :  205151   2 :123
##  1st Qu.:15610719               1st Qu.:13333377   3 : 31
##  Median :21327768               Median :19033469   4a: 34
##  Mean   :19032533               Mean   :18622755   4b: 26
##  3rd Qu.:24128509               3rd Qu.:24508858
##  Max.   :28653496               Max.   :28302313
##                                 NA's   :180
##  Person.to.person   Host.range  Human.only Non.human.primate Other.mammals
##  ?:  2             broad  :161   ?:  6      ?:  6             ?:  6
##  0:121             narrow : 47   0:182      0:147             0: 50
##  1: 91             unknown:  6   1: 26      1: 61             1:158
##
##
##
##
##  Birds   Reptiles Fish    Reference..Host.range.   Ref2..HR.
##  ?:  6   ?:  6    ?:  6   Min.   :    3981       Min.   : 1325663
##  0:171   0:201    0:207   1st Qu.:12145673       1st Qu.:15598044
##  1: 37   1:  7    1:  1   Median :19486058       Median :21106767
##                           Mean   :17593006       Mean   :18692212
##                             3rd Qu.:22936195       3rd Qu.:23155478  
##                             Max.   :28653496       Max.   :28549438  
##                                                    NA's   :160
```

Exercise 2: Code Translation

For this series of exercises, you'll be given a chunk of code that does some data manipulation in base R. Your goal is to describe what this code is doing (in text below the code) and then translate that data manipulation operation using `dplyr` functions (in the empty code chunks). The `dplyr` solution will hopefully be simpler and more intuitive to you (which is why I'm encouraging you to learn `dplyr`). However, as an R user, you'll also be seeing lots of code written with base R functions, so best to be able to understand the basics of data manipulation with these built-in functions as well.

a)

The codes below are trying to find and output rows (observations) whose `Family` column is `Coronaviridae`.

- Base R code:

```
viruses[viruses$Family == "Coronaviridae", ]
```

```
##                                                   Species            Genus
## 21                                     Alphacoronavirus 1 Alphacoronavirus
## 22                               Human coronavirus 229E  Alphacoronavirus
## 23                                Human coronavirus NL63 Alphacoronavirus
## 24                                      Betacoronavirus 1  Betacoronavirus
## 25                                Human coronavirus HKU1  Betacoronavirus
## 26  Middle East respiratory syndrome-related coronavirus  Betacoronavirus
## 27 Severe acute respiratory syndrome-related coronavirus  Betacoronavirus
## 28                                        Human torovirus        Torovirus
##           Family Envelope   Genome ICTV.history.URL Discovery.year
## 21 Coronaviridae        1 (+)ssRNA         20161200           2007
## 22 Coronaviridae        1 (+)ssRNA         20161203           1966
## 23 Coronaviridae        1 (+)ssRNA         20161204           2004
## 24 Coronaviridae        1 (+)ssRNA         20161212           1967
## 25 Coronaviridae        1 (+)ssRNA         20161214           2005
## 26 Coronaviridae        1 (+)ssRNA         20161215           2012
## 27 Coronaviridae        1 (+)ssRNA         20161219           2003
## 28 Coronaviridae        1 (+)ssRNA         20161240           1984
##    Reference..discovery. Serological.detection.only Vector Inhalation
## 21              17447647                          Y      0          0
## 22               4285768                          N      0          1
## 23              15073334                          N      0          1
## 24               5231356                          N      0          1
## 25              15613317                          N      0          1
## 26              23075143                          N      0          1
## 27              12711465                          N      0          1
## 28               6143978                          N      0          0
##    Ingestion Sexual Iatrogenic..inc..blood. Fomites Broken.Skin Maternal
## 21        1*      0                      0       0           0        0
## 22         0      0                      0       1           0        0
## 23         0      0                      0       1           0        0
## 24         0      0                      0       1           0        0
## 25         0      0                      0       1           0        0
## 26         0      0                      0       0           0        0
## 27         0      0                      0       0           0        0
## 28         1      0                      0       0           0        0
##    Direct.Contact Reference..Transmission.route. Ref2..TR.
## 21              0                       22320357        NA
## 22              0                       26556276        NA
## 23              0                       21366416        NA
## 24              0                       28549438        NA
## 25              0                       23161446        NA
## 26              1                       28653496        NA
## 27              0                       15018126        NA
## 28              0                        9426455        NA
##    Transmission.level Person.to.person Host.range Human.only
## 21                  2                0      broad          0
## 22                 4b                1     narrow          1
## 23                 4b                1     narrow          1
## 24                 4a                1      broad          0
## 25                 4b                1     narrow          1
## 26                  3                1      broad          0
## 27                 4a                1      broad          0
## 28                 4b                1     narrow          1
##    Non.human.primate Other.mammals Birds Reptiles Fish
## 21                 0             1     0        0    0
## 22                 0             0     0        0    0
## 23                 0             0     0        0    0
## 24                 0             1     0        0    0
## 25                 0             0     0        0    0
## 26                 0             1     0        0    0
## 27                 0             1     0        0    0
## 28                 0             0     0        0    0
##    Reference..Host.range. Ref2..HR.
## 21               22320357  28549438
## 22               12551991        NA
## 23               21366416        NA
## 24               22362949  28549438
## 25               23161446        NA
## 26               28653496        NA
## 27               12939793        NA
## 28                9426455        NA
```

- `dplyr` equivalent:

```
filter(viruses, Family == 'Coronaviridae')
```

```
##                                                Species            Genus
## 1                                   Alphacoronavirus 1 Alphacoronavirus
## 2                               Human coronavirus 229E Alphacoronavirus
## 3                               Human coronavirus NL63 Alphacoronavirus
## 4                                    Betacoronavirus 1  Betacoronavirus
## 5                               Human coronavirus HKU1  Betacoronavirus
## 6  Middle East respiratory syndrome-related coronavirus  Betacoronavirus
## 7 Severe acute respiratory syndrome-related coronavirus  Betacoronavirus
## 8                                      Human torovirus        Torovirus
##          Family Envelope   Genome ICTV.history.URL Discovery.year
## 1 Coronaviridae        1 (+)ssRNA         20161200           2007
## 2 Coronaviridae        1 (+)ssRNA         20161203           1966
## 3 Coronaviridae        1 (+)ssRNA         20161204           2004
## 4 Coronaviridae        1 (+)ssRNA         20161212           1967
## 5 Coronaviridae        1 (+)ssRNA         20161214           2005
## 6 Coronaviridae        1 (+)ssRNA         20161215           2012
## 7 Coronaviridae        1 (+)ssRNA         20161219           2003
## 8 Coronaviridae        1 (+)ssRNA         20161240           1984
##   Reference..discovery. Serological.detection.only Vector Inhalation
## 1              17447647                          Y      0          0
## 2               4285768                          N      0          1
## 3              15073334                          N      0          1
## 4               5231356                          N      0          1
## 5              15613317                          N      0          1
## 6              23075143                          N      0          1
## 7              12711465                          N      0          1
## 8               6143978                          N      0          0
##   Ingestion Sexual Iatrogenic..inc..blood. Fomites Broken.Skin Maternal
## 1        1*      0                       0       0           0        0
## 2         0      0                       0       1           0        0
## 3         0      0                       0       1           0        0
## 4         0      0                       0       1           0        0
## 5         0      0                       0       1           0        0
## 6         0      0                       0       0           0        0
## 7         0      0                       0       0           0        0
## 8         1      0                       0       0           0        0
##   Direct.Contact Reference..Transmission.route. Ref2..TR.
## 1              0                       22320357        NA
## 2              0                       26556276        NA
## 3              0                       21366416        NA
## 4              0                       28549438        NA
## 5              0                       23161446        NA
## 6              1                       28653496        NA
## 7              0                       15018126        NA
## 8              0                        9426455        NA
##   Transmission.level Person.to.person Host.range Human.only
## 1                  2                0      broad          0
## 2                 4b                1     narrow          1
## 3                 4b                1     narrow          1
## 4                 4a                1      broad          0
## 5                 4b                1     narrow          1
## 6                  3                1      broad          0
## 7                 4a                1      broad          0
## 8                 4b                1     narrow          1
##   Non.human.primate Other.mammals Birds Reptiles Fish
## 1                 0             1     0        0    0
## 2                 0             0     0        0    0
## 3                 0             0     0        0    0
## 4                 0             1     0        0    0
## 5                 0             0     0        0    0
## 6                 0             1     0        0    0
## 7                 0             1     0        0    0
## 8                 0             0     0        0    0
##   Reference..Host.range. Ref2..HR.
## 1               22320357  28549438
## 2               12551991        NA
## 3               21366416        NA
## 4               22362949  28549438
## 5               23161446        NA
## 6               28653496        NA
## 7               12939793        NA
## 8                9426455        NA
```

b)

The codes below are trying to slice the origin dataframe's 1-10 rows and 1, 2, 3, 7 columns, and output the result.

- Base R code:

```
viruses[1:10, c(1, 2, 3, 7)]
```

```
##                                         Species          Genus       Family
## 1                        Chapare mammarenavirus Mammarenavirus Arenaviridae
## 2                     Guanarito mammarenavirus Mammarenavirus Arenaviridae
## 3                         Junín mammarenavirus Mammarenavirus Arenaviridae
## 4                         Lassa mammarenavirus Mammarenavirus Arenaviridae
## 5                          Lujo mammarenavirus Mammarenavirus Arenaviridae
## 6  Lymphocytic choriomeningitis mammarenavirus Mammarenavirus Arenaviridae
## 7                       Machupo mammarenavirus Mammarenavirus Arenaviridae
## 8                        Mobala mammarenavirus Mammarenavirus Arenaviridae
## 9                      Pichindé mammarenavirus Mammarenavirus Arenaviridae
## 10                        Sabiá mammarenavirus Mammarenavirus Arenaviridae
##    Discovery.year
## 1            2008
## 2            1991
## 3            1958
## 4            1970
## 5            2009
## 6            1934
## 7            1964
## 8            1985
## 9            1974
## 10           1994
```

Hint: Look at the `dplyr` function called `slice()` using `?slice()`.

- `dplyr` equivalent:

```
select(slice(viruses, 1:10), c(1, 2, 3, 7))
```

```
##                                        Species          Genus       Family
## 1                       Chapare mammarenavirus Mammarenavirus Arenaviridae
## 2                     Guanarito mammarenavirus Mammarenavirus Arenaviridae
## 3                         Junín mammarenavirus Mammarenavirus Arenaviridae
## 4                         Lassa mammarenavirus Mammarenavirus Arenaviridae
## 5                          Lujo mammarenavirus Mammarenavirus Arenaviridae
## 6  Lymphocytic choriomeningitis mammarenavirus Mammarenavirus Arenaviridae
## 7                       Machupo mammarenavirus Mammarenavirus Arenaviridae
## 8                        Mobala mammarenavirus Mammarenavirus Arenaviridae
## 9                      Pichindé mammarenavirus Mammarenavirus Arenaviridae
## 10                        Sabiá mammarenavirus Mammarenavirus Arenaviridae
##    Discovery.year
## 1            2008
## 2            1991
## 3            1958
## 4            1970
## 5            2009
## 6            1934
## 7            1964
## 8              1985
## 9              1974
## 10             1994
```

c)

The codes below are trying to get rows whose `Envelope` feature equal to 0 and get `Species` column from these rows, and then sort the single column Alphabetically.

- Base R code:

```
sort(viruses$Species[viruses$Envelope == 0])
```

```
##  [1] Aichivirus A                          Banna virus
##  [3] Cardiovirus A                         Cardiovirus B
##  [5] Colorado tick fever virus             Corriparta virus
##  [7] Cosavirus A                           Cosavirus B
##  [9] Cosavirus D                           Cosavirus E
## [11] Cosavirus F                           Enterovirus A
## [13] Enterovirus B                         Enterovirus C
## [15] Enterovirus D                         Enterovirus E
## [17] Equine rhinitis A virus               Erbovirus A
## [19] Eyach virus                           Foot-and-mouth disease virus
## [21] Great Island virus                    Hepatovirus A
## [23] Human picobirnavirus                  Lebombo virus
## [25] Mamastrovirus 1                       Mamastrovirus 6
## [27] Mamastrovirus 8                       Mamastrovirus 9
## [29] Mammalian orthoreovirus               Nelson Bay orthoreovirus
## [31] Norwalk virus                         Orthohepevirus A
## [33] Orungo virus                          Parechovirus A
## [35] Parechovirus B                        Rhinovirus A
## [37] Rhinovirus B                          Rhinovirus C
## [39] Rotavirus A                           Rotavirus B
## [41] Rotavirus C                           Rotavirus H
## [43] Salivirus A                           Sapporo virus
## [45] Vesicular exanthema of swine virus
## 214 Levels: African green monkey simian foamy virus ... Zika virus
```

- `dplyr` equivalent:

```
viruses %>%
  filter(Envelope == 0) %>%
  arrange(Species) %>%
  select(Species)
```

```
##                               Species
## 1                        Aichivirus A
## 2                         Banna virus
## 3                      Cardiovirus A
## 4                      Cardiovirus B
## 5          Colorado tick fever virus 
## 6                   Corriparta virus
## 7                        Cosavirus A
## 8                        Cosavirus B
## 9                        Cosavirus D
## 10                       Cosavirus E
## 11                       Cosavirus F
## 12                     Enterovirus A
## 13                     Enterovirus B
## 14                     Enterovirus C
## 15                     Enterovirus D
## 16                     Enterovirus E
## 17          Equine rhinitis A virus 
## 18                       Erbovirus A
## 19                       Eyach virus
## 20     Foot-and-mouth disease virus 
## 21               Great Island virus 
## 22                     Hepatovirus A
## 23             Human picobirnavirus 
## 24                    Lebombo virus 
## 25                   Mamastrovirus 1
## 26                   Mamastrovirus 6
## 27                   Mamastrovirus 8
## 28                   Mamastrovirus 9
## 29           Mammalian orthoreovirus
## 30          Nelson Bay orthoreovirus
## 31                    Norwalk virus 
## 32                  Orthohepevirus A
## 33                     Orungo virus 
## 34                    Parechovirus A
## 35                    Parechovirus B
## 36                      Rhinovirus A
## 37                      Rhinovirus B
## 38                      Rhinovirus C
## 39                      Rotavirus A 
## 40                      Rotavirus B 
## 41                      Rotavirus C 
## 42                       Rotavirus H
## 43                       Salivirus A
## 44                     Sapporo virus
## 45 Vesicular exanthema of swine virus
```

Exercise 3: Code Annotation

In the following series of exercises, you will be provided with functioning R code of `dplyr` data manipulation pipelines. Your goal is to comment these code blocks line-by-line, describing what each function is doing to create the final output. Please note, if you're not sure how a given line is functioning within the whole code block, this type of code is easily run in successively larger chunks. In other words, start by running the first line, then the first two lines, then the first three lines, etc. in order to see how the output changes. Additionally, reviewing function help files (e.g., `?some_function()`) may shed light on what's happening.

a)

```r
viruses %>%
#Input the viruses dataset using pipeline
  mutate(Envelope_mod = ifelse(Envelope == 1, "enveloped", "not enveloped")) %>%
  #Create a new column named 'Envelope_mod'. If the row's 'Envelope' feature
  #equal 1, assign its 'Envelope_mod' with 'enveloped'. Otherwise,
  #assign with 'not envepoed'.
  filter(Discovery.year >= 1990) %>%
  #Find rows with 'Discovery.year' feature larger than or equal to 1990.
  filter(Transmission.level %in% c("3", "4a", "4b")) %>%
  #Find rows with 'Transmission.level' feature equal to '3' or '4a' or '4b'.
  select(Family, Species, Envelope_mod) %>%
  #Select 'Family', 'Species', 'Envelope_mod' three columns.
  #Sort first by 'Family' column, then, by 'Species' column.
  arrange(Family, Species)
```

```
##             Family                                               Species
## 1     Arenaviridae                              Guanarito mammarenavirus
## 2     Arenaviridae                                   Lujo mammarenavirus
## 3     Arenaviridae                                  Sabiá mammarenavirus
## 4     Astroviridae                                       Mamastrovirus 6
## 5     Astroviridae                                       Mamastrovirus 8
## 6     Astroviridae                                       Mamastrovirus 9
## 7    Coronaviridae                                Human coronavirus HKU1
## 8    Coronaviridae                                Human coronavirus NL63
## 9    Coronaviridae  Middle East respiratory syndrome-related coronavirus
## 10   Coronaviridae Severe acute respiratory syndrome-related coronavirus
## 11     Filoviridae                                 Bundibugyo ebolavirus
## 12    Flaviviridae                                           Pegivirus A
## 13    Flaviviridae                                           Usutu virus
## 14    Hantaviridae                                 Andes orthohantavirus
## 15 Paramyxoviridae                                    Nipah henipavirus 
## 16    Phenuviridae                                      SFTS phlebovirus
## 17  Picornaviridae                                          Aichivirus A
## 18  Picornaviridae                                           Cosavirus A
## 19  Picornaviridae                                           Cosavirus B
## 20  Picornaviridae                                           Cosavirus D
## 21  Picornaviridae                                           Cosavirus E
## 22  Picornaviridae                                           Cosavirus F
## 23  Picornaviridae                                          Rhinovirus C
## 24  Picornaviridae                                           Salivirus A
## 25   Pneumoviridae                                 Human metapneumovirus
## 26      Reoviridae                              Nelson Bay orthoreovirus
## 27      Reoviridae                                           Rotavirus H
## 28    Retroviridae                        Primate T-lymphotropic virus 3
## 29   Rhabdoviridae                                  Bas-Congo tibrovirus
##     Envelope_mod
## 1       enveloped
## 2       enveloped
## 3       enveloped
## 4  not enveloped
## 5  not enveloped
## 6  not enveloped
## 7       enveloped
## 8       enveloped
## 9       enveloped
## 10     enveloped
## 11     enveloped
## 12     enveloped
## 13     enveloped
## 14     enveloped
## 15     enveloped
## 16     enveloped
## 17 not enveloped
## 18 not enveloped
## 19 not enveloped
## 20 not enveloped
## 21 not enveloped
## 22 not enveloped
## 23 not enveloped
## 24 not enveloped
## 25     enveloped
## 26 not enveloped
## 27 not enveloped
## 28     enveloped
## 29     enveloped
```

b)

```r
viruses %>%
#Input the viruses dataset using pipeline
  filter(Discovery.year >= 1990) %>%
  #Find rows with 'Discovery.year' larger or equal to 1990
  filter(Transmission.level %in% c("3", "4a", "4b")) %>%
  #Find rows with 'Transmission.level' feature equal to '3' or '4a' or '4b'.
  group_by(Family) %>%
  #Classify the dataset into several groups by their 'Family' column.
  #Rows with same 'Family' will be in same group.
  summarize(
  #Create a new dataframe to store data below
    n = n(),
    #Get number of rows of each group.
    n_enveloped = sum(Envelope),
    #Get the sum of 'Envelope' of each group.
    proportion_enveloped = (n_enveloped/n)*100
    #Get the ratio of envelope species to all the species in a group
  ) %>%
  #Sort the summary dataframe by the 'n' column,
  #from the largest to the smallest.
  arrange(desc(n))
```

```
## # A tibble: 13 x 4
##    Family              n n_enveloped proportion_enveloped
##    <fct>           <int>       <int>                <dbl>
##  1 Picornaviridae      8           0                    0
##  2 Coronaviridae       4           4                  100
##  3 Arenaviridae        3           3                  100
##  4 Astroviridae        3           0                    0
##  5 Flaviviridae        2           2                  100
##  6 Reoviridae          2           0                    0
##  7 Filoviridae         1           1                  100
##  8 Hantaviridae        1           1                  100
##  9 Paramyxoviridae     1           1                  100
## 10 Phenuviridae        1           1                  100
## 11 Pneumoviridae       1           1                  100
## 12 Retroviridae        1           1                  100
## 13 Rhabdoviridae       1           1                  100
```

What do you notice about the `proportion_enveloped` column?

Answer: It is either 0% or 100%.

c)

```
#Input the viruses dataset using pipeline
viruses %>%
  #Classify the dataset into several groups by their 'Family' column.
  #Rows with same 'Family' will be in same group.
  group_by(Family) %>%
  #Get a summary dataset. Write it with a column named 'n_envelope_types'
  #whose values are the number of the kinds of 'enveloped'
  #or 'non-enveloped', namely, 1 or 2.
  summarize(n_envelope_types = n_distinct(Envelope)) %>%
  #Sort the summary dataset by decreasing n_envelope_types number.
  arrange(desc(n_envelope_types))
```

```
## # A tibble: 22 x 2
##    Family        n_envelope_types
##    <fct>                    <int>
##  1 Arenaviridae                 1
##  2 Astroviridae                 1
##  3 Bornaviridae                 1
##  4 Caliciviridae                1
##  5 Coronaviridae                1
##  6 Filoviridae                  1
##  7 Flaviviridae                 1
##  8 Hantaviridae                 1
##  9 Hepeviridae                  1
## 10 Nairoviridae                 1
## # ... with 12 more rows
```

What do you learn from this data summary about the number of distinct envelope types per viral family?

Answer: The viruses in one family have the same enveloped feature. That is, either all are enveloped, or all are non-enveloped.

Bonus Exercise: Install `rethinking`

If you have not yet installed the `rethinking` package, now would be a good time to try to do so, using the instructions at https://github.com/rmcelreath/rethinking.