Starting with Data

```
#setwd("/Users/namigabbasov/Desktop/R-Data-Carpentry")
```

```
# Libraries

# install.packages("tidyverse")
# install.packages("ggplot2")



library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(ggplot2)
```

Loading the survey data

```
#download.file(url = "https://ndownloader.figshare.com/files/2292169",
              #destfile = "portal_data_joined.csv")
#surveys <- read_csv("portal_data_joined.csv")
```

```r
# data import from github
surveys<-read_csv("https://raw.githubusercontent.com/UnitForDataScience/RWorkshop/main/portal
```

```
Rows: 34786 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (6): species_id, sex, genus, species, taxa, plot_type
dbl (7): record_id, month, day, year, plot_id, hindfoot_length, weight

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Frames

```r
str(surveys)
```

```
spc_tbl_ [34,786 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ record_id      : num [1:34786] 1 72 224 266 349 363 435 506 588 661 ...
 $ month          : num [1:34786] 7 8 9 10 11 11 12 1 2 3 ...
 $ day            : num [1:34786] 16 19 13 16 12 12 10 8 18 11 ...
 $ year           : num [1:34786] 1977 1977 1977 1977 1977 ...
 $ plot_id        : num [1:34786] 2 2 2 2 2 2 2 2 2 2 ...
 $ species_id     : chr [1:34786] "NL" "NL" "NL" "NL" ...
 $ sex            : chr [1:34786] "M" "M" NA NA ...
 $ hindfoot_length: num [1:34786] 32 31 NA NA NA NA NA NA NA NA ...
 $ weight         : num [1:34786] NA NA NA NA NA NA NA NA 218 NA ...
 $ genus          : chr [1:34786] "Neotoma" "Neotoma" "Neotoma" "Neotoma" ...
 $ species        : chr [1:34786] "albigula" "albigula" "albigula" "albigula" ...
 $ taxa           : chr [1:34786] "Rodent" "Rodent" "Rodent" "Rodent" ...
 $ plot_type      : chr [1:34786] "Control" "Control" "Control" "Control" ...
 - attr(*, "spec")=
  .. cols(
  ..   record_id = col_double(),
  ..   month = col_double(),
  ..   day = col_double(),
  ..   year = col_double(),
  ..   plot_id = col_double(),
  ..   species_id = col_character(),
  ..   sex = col_character(),
```

```
  ..   hindfoot_length = col_double(),
  ..   weight = col_double(),
  ..   genus = col_character(),
  ..   species = col_character(),
  ..   taxa = col_character(),
  ..   plot_type = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

Inspecting data frames

```
#size
dim(surveys)
```

```
[1] 34786    13
```

```
nrow(surveys)
```

```
[1] 34786
```

```
ncol(surveys)
```

```
[1] 13
```

```
# content
head(surveys, n= 10)
```

```
# A tibble: 10 x 13
   record_id month   day  year plot_id species_id sex   hindfoot_length weight
       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>  <dbl>
 1         1     7    16  1977       2 NL         M                  32     NA
 2        72     8    19  1977       2 NL         M                  31     NA
 3       224     9    13  1977       2 NL         <NA>               NA     NA
 4       266    10    16  1977       2 NL         <NA>               NA     NA
 5       349    11    12  1977       2 NL         <NA>               NA     NA
 6       363    11    12  1977       2 NL         <NA>               NA     NA
 7       435    12    10  1977       2 NL         <NA>               NA     NA
 8       506     1     8  1978       2 NL         <NA>               NA     NA
```

```
 9       588     2    18  1978       2 NL          M                   NA    218
10       661     3    11  1978       2 NL          <NA>                NA     NA
# i 4 more variables: genus <chr>, species <chr>, taxa <chr>, plot_type <chr>
```

```
tail(surveys)
```

```
# A tibble: 6 x 13
  record_id month   day  year plot_id species_id sex   hindfoot_length weight
      <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>      <chr>           <dbl>  <dbl>
1     26787     9    27  1997       7 PL         F                  21     16
2     26966    10    25  1997       7 PL         M                  20     16
3     27185    11    22  1997       7 PL         F                  21     22
4     27792     5     2  1998       7 PL         F                  20      8
5     28806    11    21  1998       7 PX         <NA>               NA     NA
6     30986     7     1  2000       7 PX         <NA>               NA     NA
# i 4 more variables: genus <chr>, species <chr>, taxa <chr>, plot_type <chr>
```

```
# names
names(surveys)
```

```
 [1] "record_id"       "month"           "day"             "year"
 [5] "plot_id"         "species_id"      "sex"             "hindfoot_length"
 [9] "weight"          "genus"           "species"         "taxa"
[13] "plot_type"
```

```
row_names<-rownames(surveys)

# summary
summary(surveys)
```

```
   record_id         month              day             year         plot_id
 Min.   :    1   Min.   : 1.000   Min.   : 1.0   Min.   :1977   Min.   : 1.00
 1st Qu.: 8964   1st Qu.: 4.000   1st Qu.: 9.0   1st Qu.:1984   1st Qu.: 5.00
 Median :17762   Median : 6.000   Median :16.0   Median :1990   Median :11.00
 Mean   :17804   Mean   : 6.474   Mean   :16.1   Mean   :1990   Mean   :11.34
 3rd Qu.:26655   3rd Qu.:10.000   3rd Qu.:23.0   3rd Qu.:1997   3rd Qu.:17.00
 Max.   :35548   Max.   :12.000   Max.   :31.0   Max.   :2002   Max.   :24.00

  species_id            sex            hindfoot_length     weight
 Length:34786       Length:34786       Min.   : 2.00   Min.   :  4.00
```

```
 Class :character   Class :character   1st Qu.:21.00   1st Qu.: 20.00
 Mode  :character   Mode  :character   Median :32.00   Median : 37.00
                                       Mean   :29.29   Mean   : 42.67
                                       3rd Qu.:36.00   3rd Qu.: 48.00
                                       Max.   :70.00   Max.   :280.00
                                       NA's   :3348    NA's   :2503
    genus              species              taxa            plot_type
 Length:34786       Length:34786       Length:34786       Length:34786
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Indexing and subsetting data frames

```
# index by numbers
firstrow_first_column<-surveys[1,1]
firstrow_allcolumns<-surveys[1,]
allrows_firstcolumn<-surveys[,1]

first_column <-surveys[[1]]        # get first column as vector

surveys[1:6, 5:7]   # get a part of the data
```

```
# A tibble: 6 x 3
  plot_id species_id sex
    <dbl> <chr>      <chr>
1       2 NL         M
2       2 NL         M
3       2 NL         <NA>
4       2 NL         <NA>
5       2 NL         <NA>
6       2 NL         <NA>
```

```
# index by a column name

plot_id<- surveys["plot_id"]
plot_id<-surveys[, "plot_id"]
sex<- my_colum<-surveys$sex
```

Factors

```
# make factor variable from surveys dataframe
order<- factor(c("Less", "A bit more", "more", "most"))
levels(order)
```

```
[1] "A bit more" "Less"       "more"       "most"
```

```
# reorder: levels = c("male", "female")
ordered<-factor(order, levels = c("Less", "A bit more", "more", "most"))
levels(ordered)
```

```
[1] "Less"       "A bit more" "more"       "most"
```

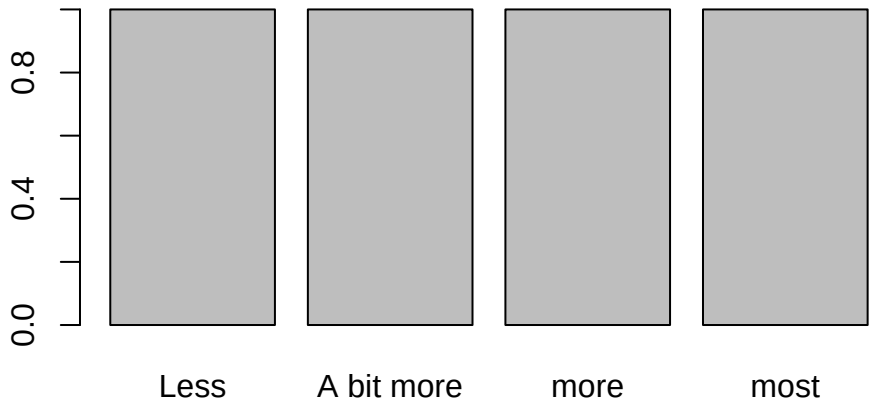
```
#converting first as character first and then as numeric
years<- factor(c(1991, 1993, 1992, 1999,1990))

as.numeric(years) # incorrect way to covert to numeric
```

```
[1] 2 4 3 5 1
```

```
as.numeric(as.character(years)) # right way to covert to numeric
```

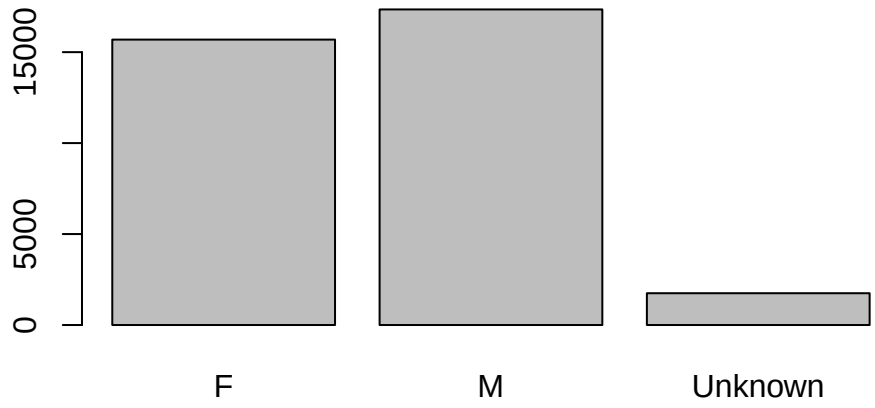
```
[1] 1991 1993 1992 1999 1990
```

```
# ploting
plot(ordered)
```

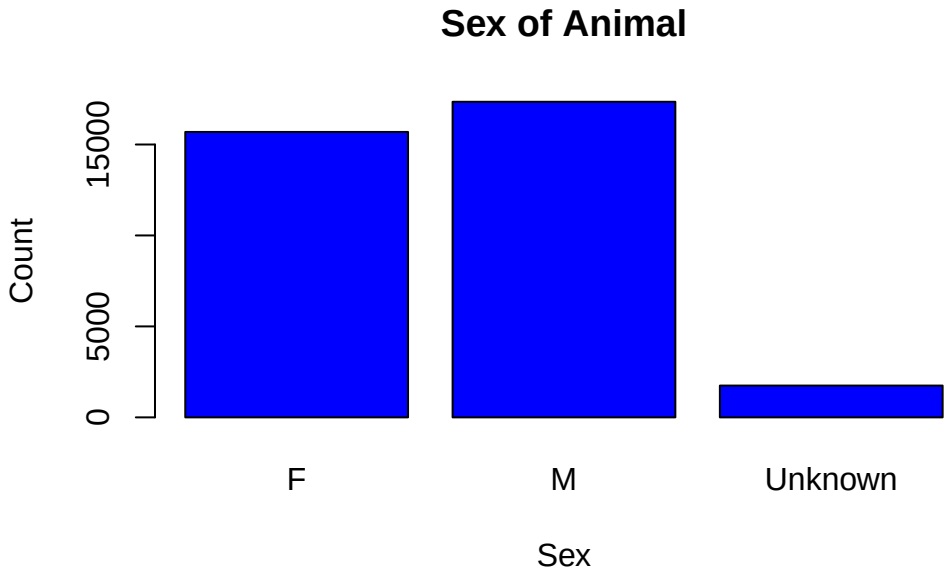



```
# Renaming factors: save a column as object, addNA(x),
sex <- surveys$sex
sex<- addNA(sex)
levels(sex)[3]<-"Unknown"
```

```
# plot again
plot(sex)
```




```
# make a barplot
sex_table<-table(sex)
barplot(sex_table, main="Sex of Animal", col="blue", xlab="Sex", ylab="Count")
```



Formatting dates

```
library(lubridate)
```

```
my_date <- ymd("2015-01-01")
str(my_date)
```

```
 Date[1:1], format: "2015-01-01"
```

```
my_date <- ymd(paste("2015", "1", "1", sep = "-"))
str(my_date)
```

```
 Date[1:1], format: "2015-01-01"
```

```
# create a data variable in survey dataset
cont_time_variables<-paste(surveys$year, surveys$month, surveys$day, sep = "-")
date_column<-ymd(paste(surveys$year, surveys$month, surveys$day, sep = "-"))
```

```
Warning: 129 failed to parse.
```

```
surveys$date<-ymd(paste(surveys$year, surveys$month, surveys$day, sep = "-"))
```

```
Warning: 129 failed to parse.
```

```
na_day<-is.na(surveys$day)
surveys$date[na_day]
```

```
Date of length 0
```

```
na_month<-is.na(surveys$month)
surveys$date[na_month]
```

```
Date of length 0
```

```
na_year<-is.na(surveys$year)
surveys$date[na_year]
```

```
Date of length 0
```

```
na_date<-is.na(surveys$date)
surveys$date[na_date]
```

```
  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
 [26] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
 [51] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
 [76] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[101] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
[126] NA NA NA NA
```

```
sum(is.na(surveys$date))
```

```
[1] 129
```

```
missing_dates <- surveys[is.na(surveys$date), c("year", "month", "day")]
head(missing_dates)
```

```
# A tibble: 6 x 3
   year month   day
  <dbl> <dbl> <dbl>
1  2000     9    31
2  2000     4    31
3  2000     4    31
4  2000     4    31
5  2000     4    31
6  2000     9    31
```