Data Manipulation

```
#install.packages("tidyverse")
#install.packages("tinytex")
#install.packages("lubridate")
#install.packages("readxl")
#install.packages("readr")
library(tidyverse)
library(tinytex)
library(lubridate)
library(readxl)
library(readr)
```

Task 1: Conceptual Questions

1. If your working directory is myfolder/homework/, what relative path would you specify to get the file located at myfolder/MyData.csv?

   The realtive path would be ../MyData.csv

2. What are the major benefits of using R projects?

   R projects are helpful becuase they're their own wokring environment, they connect directly with Github, and it makes projects easy to share or colaborate on becuase everything is in one place.

3. What is git and what is github?

   Git is what lives on your local computer and tracks changes locally. It then allows you to upload and merge uploaded copies to github. Github is the website where we create repsoitories and keeps the revision history of code.

4. What are the two main differences between a tibble and a data.frame?

A tibble prints output in a uniform and concise way that. Additionally, they do not coerce down to a vector when you subset to only one column using

5. Rewrite the following nested function call using BaseR's chaining operator:

   as_tibble(iris) |> select(starts_with("Petal"),Species) |> filter(Petal.Length < 1.55) |> arrange(Species)

Task 2: Reading Delimited Data

Glass Data

1. Reading glass data from URL

```
glass_data <-read_csv("https://www4.stat.ncsu.edu/~online/datasets/glass.data",
         col_names = c("ID", "RI", "Na", "Mg", "Al", "Si", "K", "Ca", "Ba", "Fe", "Type.of.G
```

```
Rows: 214 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (11): ID, RI, Na, Mg, Al, Si, K, Ca, Ba, Fe, Type.of.Glass

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.
```

```
glass_data
```

```
# A tibble: 214 x 11
      ID    RI    Na    Mg    Al    Si     K    Ca    Ba    Fe Type.of.Glass
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>         <dbl>
 1     1  1.52  13.6  4.49  1.1   71.8  0.06  8.75     0  0                 1
 2     2  1.52  13.9  3.6   1.36  72.7  0.48  7.83     0  0                 1
 3     3  1.52  13.5  3.55  1.54  73.0  0.39  7.78     0  0                 1
 4     4  1.52  13.2  3.69  1.29  72.6  0.57  8.22     0  0                 1
 5     5  1.52  13.3  3.62  1.24  73.1  0.55  8.07     0  0                 1
 6     6  1.52  12.8  3.61  1.62  73.0  0.64  8.07     0  0.26              1
 7     7  1.52  13.3  3.6   1.14  73.1  0.58  8.17     0  0                 1
 8     8  1.52  13.2  3.61  1.05  73.2  0.57  8.24     0  0                 1
 9     9  1.52  14.0  3.58  1.37  72.1  0.56  8.3      0  0                 1
10    10  1.52  13    3.6   1.36  73.0  0.57  8.4      0  0.11              1
# i 204 more rows
```

2. Starting a chain to overwrite data

```
glass_data_mutate <- glass_data |>
  mutate(Type.of.Glass = factor(Type.of.Glass,
                        levels = c(1, 2, 3, 4, 5, 6, 7),
                        labels = c("building_windows_float_processed",
                                   "building_windows_non_float_processed",
                                   "vehicle_windows_float_processed",
                                   "vehicle_windows_non_float_processed",
                                   "containers",
                                   "tableware",
                                   "headlamps")))
glass_data_mutate
```

```
# A tibble: 214 x 11
      ID    RI    Na    Mg    Al    Si     K    Ca    Ba    Fe Type.of.Glass
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct>
 1     1  1.52  13.6  4.49  1.1   71.8  0.06  8.75     0  0    building_windows~
 2     2  1.52  13.9  3.6   1.36  72.7  0.48  7.83     0  0    building_windows~
 3     3  1.52  13.5  3.55  1.54  73.0  0.39  7.78     0  0    building_windows~
 4     4  1.52  13.2  3.69  1.29  72.6  0.57  8.22     0  0    building_windows~
 5     5  1.52  13.3  3.62  1.24  73.1  0.55  8.07     0  0    building_windows~
 6     6  1.52  12.8  3.61  1.62  73.0  0.64  8.07     0  0.26 building_windows~
 7     7  1.52  13.3  3.6   1.14  73.1  0.58  8.17     0  0    building_windows~
 8     8  1.52  13.2  3.61  1.05  73.2  0.57  8.24     0  0    building_windows~
 9     9  1.52  14.0  3.58  1.37  72.1  0.56  8.3      0  0    building_windows~
10    10  1.52  13    3.6   1.36  73.0  0.57  8.4      0  0.11 building_windows~
# i 204 more rows
```

3. Continuing the chain to filter down the data

```
glass_data_filter <- glass_data |>
  mutate(Type.of.Glass = factor(Type.of.Glass,
                        levels = c(1, 2, 3, 4, 5, 6, 7),
                        labels = c("building_windows_float_processed",
                                   "building_windows_non_float_processed",
                                   "vehicle_windows_float_processed",
                                   "vehicle_windows_non_float_processed",
                                   "containers",
                                   "tableware",
                                   "headlamps"))) |>
  filter(Fe < 0.2 & Type.of.Glass %in% c("tableware", "headlamps"))
```

```
glass_data_filter
```

```
# A tibble: 38 x 11
      ID    RI    Na    Mg    Al    Si     K    Ca    Ba    Fe Type.of.Glass
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <fct>
 1   177  1.52  14    2.39  1.56  72.4  0     9.57  0        0 tableware
 2   178  1.52  13.8  2.41  1.19  72.8  0     9.77  0        0 tableware
 3   179  1.52  14.5  2.24  1.62  72.4  0     9.26  0        0 tableware
 4   180  1.52  14.1  2.19  1.66  72.7  0     9.32  0        0 tableware
 5   181  1.51  14.4  1.74  1.54  74.6  0     7.59  0        0 tableware
 6   182  1.52  15.0  0.78  1.74  72.5  0     9.95  0        0 tableware
 7   183  1.52  14.2  0     2.09  72.7  0    10.9   0        0 tableware
 8   184  1.52  14.6  0     0.56  73.5  0    11.2   0        0 tableware
 9   185  1.51  17.4  0     0.34  75.4  0     6.65  0        0 tableware
10   186  1.51  13.7  3.2   1.81  72.8  1.76  5.43  1.19     0 headlamps
# i 28 more rows
```

Yeast Data

1. Reading yeast data from URL

```
yeast_data <- read_delim("https://www4.stat.ncsu.edu/online/datasets/yeast.data",
                         delim = "  ",
                         col_names = c("seq_name", "mcg", "gvh", "alm", "mit", "erl", "pox", "
```

```
Warning: One or more parsing issues, call `problems()` on your data frame for details,
e.g.:
  dat <- vroom(...)
  problems(dat)

Rows: 1484 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: "  "
chr (2): seq_name, class
dbl (8): mcg, gvh, alm, mit, erl, pox, vac, nuc

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.
```

```
yeast_data
```

```
# A tibble: 1,484 x 10
   seq_name     mcg   gvh   alm   mit   erl   pox   vac   nuc class
   <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
 1 ADT1_YEAST  0.58  0.61  0.47  0.13   0.5   0    0.48  0.22 MIT
 2 ADT2_YEAST  0.43  0.67  0.48  0.27   0.5   0    0.53  0.22 MIT
 3 ADT3_YEAST  0.64  0.62  0.49  0.15   0.5   0    0.53  0.22 MIT
 4 AAR2_YEAST  0.58  0.44  0.57  0.13   0.5   0    0.54  0.22 NUC
 5 AATM_YEAST  0.42  0.44  0.48  0.54   0.5   0    0.48  0.22 MIT
 6 AATC_YEAST  0.51  0.4   0.56  0.17   0.5   0.5  0.49  0.22 CYT
 7 ABC1_YEAST  0.5   0.54  0.48  0.65   0.5   0    0.53  0.22 MIT
 8 BAF1_YEAST  0.48  0.45  0.59  0.2    0.5   0    0.58  0.34 NUC
 9 ABF2_YEAST  0.55  0.5   0.66  0.36   0.5   0    0.49  0.22 MIT
10 ABP1_YEAST  0.4   0.39  0.6   0.15   0.5   0    0.58  0.3  CYT
# i 1,474 more rows
```

2. Starting a chain to remove columns

```
yeast_data_select <- yeast_data |>
  select(-seq_name, -nuc)

yeast_data_select
```

```
# A tibble: 1,484 x 8
     mcg   gvh   alm   mit   erl   pox   vac class
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
 1  0.58  0.61  0.47  0.13   0.5   0    0.48 MIT
 2  0.43  0.67  0.48  0.27   0.5   0    0.53 MIT
 3  0.64  0.62  0.49  0.15   0.5   0    0.53 MIT
 4  0.58  0.44  0.57  0.13   0.5   0    0.54 NUC
 5  0.42  0.44  0.48  0.54   0.5   0    0.48 MIT
 6  0.51  0.4   0.56  0.17   0.5   0.5  0.49 CYT
 7  0.5   0.54  0.48  0.65   0.5   0    0.53 MIT
 8  0.48  0.45  0.59  0.2    0.5   0    0.58 NUC
 9  0.55  0.5   0.66  0.36   0.5   0    0.49 MIT
10  0.4   0.39  0.6   0.15   0.5   0    0.58 CYT
# i 1,474 more rows
```

3. Continuing the chain to add grouping and columns

```r
yeast_data_stats <- yeast_data |>
  select(-seq_name, -nuc) |>
  group_by(class) |>
  mutate(across(where(is.numeric), list(mean = mean, median = median), .names = "{.col}_{.fn}

yeast_data_stats
```

```
# A tibble: 1,484 x 22
# Groups:   class [44]
     mcg   gvh   alm   mit   erl   pox   vac class mcg_mean mcg_median gvh_mean
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>    <dbl>      <dbl>    <dbl>
 1  0.58  0.61  0.47  0.13   0.5   0    0.48 MIT      0.521       0.51    0.531
 2  0.43  0.67  0.48  0.27   0.5   0    0.53 MIT      0.521       0.51    0.531
 3  0.64  0.62  0.49  0.15   0.5   0    0.53 MIT      0.521       0.51    0.531
 4  0.58  0.44  0.57  0.13   0.5   0    0.54 NUC      0.453       0.45    0.458
 5  0.42  0.44  0.48  0.54   0.5   0    0.48 MIT      0.521       0.51    0.531
 6  0.51  0.4   0.56  0.17   0.5   0.5  0.49 CYT      0.480       0.48    0.469
 7  0.5   0.54  0.48  0.65   0.5   0    0.53 MIT      0.521       0.51    0.531
 8  0.48  0.45  0.59  0.2    0.5   0    0.58 NUC      0.453       0.45    0.458
 9  0.55  0.5   0.66  0.36   0.5   0    0.49 MIT      0.521       0.51    0.531
10  0.4   0.39  0.6   0.15   0.5   0    0.58 CYT      0.480       0.48    0.469
# i 1,474 more rows
# i 11 more variables: gvh_median <dbl>, alm_mean <dbl>, alm_median <dbl>,
#   mit_mean <dbl>, mit_median <dbl>, erl_mean <dbl>, erl_median <dbl>,
#   pox_mean <dbl>, pox_median <dbl>, vac_mean <dbl>, vac_median <dbl>
```

Task 3: Combining Excel & Delimited Data

1. Reading in white wine data from first sheet

```r
white_wine <- read_excel("white-wine.xlsx",
                         sheet = excel_sheets("white-wine.xlsx")[1])
white_wine
```

```
# A tibble: 4,898 x 12
   `fixed acidity` `volatile acidity` `citric acid` `residual sugar` chlorides
             <dbl>              <dbl>         <dbl>            <dbl>     <dbl>
 1               7               0.27          0.36             20.7     0.045
 2              63               0.3           0.34              1.6     0.049
 3              81               0.28          0.4               6.9     0.05
```

```
 4              72               0.23          0.32              8.5     0.058
 5              72               0.23          0.32              8.5     0.058
 6              81               0.28          0.4               6.9     0.05
 7              62               0.32          0.16              7       0.045
 8               7               0.27          0.36             20.7     0.045
 9              63               0.3           0.34              1.6     0.049
10              81               0.22          0.43              1.5     0.044
# i 4,888 more rows
# i 7 more variables: `free sulfur dioxide` <dbl>,
#   `total sulfur dioxide` <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
#   alcohol <dbl>, quality <dbl>
```

2. Reading in variable names from second sheet and replacing col names

```
white_wine_var <- read_excel("white-wine.xlsx",
                          sheet = excel_sheets("white-wine.xlsx")[2])
#white_wine_var

wine_names <- white_wine_var[[1]]

colnames(white_wine) <- wine_names

white_wine
```

```
# A tibble: 4,898 x 12
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1             7             0.27        0.36           20.7     0.045
 2            63             0.3         0.34            1.6     0.049
 3            81             0.28        0.4             6.9     0.05
 4            72             0.23        0.32            8.5     0.058
 5            72             0.23        0.32            8.5     0.058
 6            81             0.28        0.4             6.9     0.05
 7            62             0.32        0.16            7       0.045
 8             7             0.27        0.36           20.7     0.045
 9            63             0.3         0.34            1.6     0.049
10            81             0.22        0.43            1.5     0.044
# i 4,888 more rows
# i 7 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>
```

3. Adding in a white wine col

```
white_wine_final <- white_wine |>
  mutate(color = "white")

white_wine_final
```

```
# A tibble: 4,898 x 13
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1             7             0.27        0.36           20.7     0.045
 2            63             0.3         0.34            1.6     0.049
 3            81             0.28        0.4             6.9     0.05
 4            72             0.23        0.32            8.5     0.058
 5            72             0.23        0.32            8.5     0.058
 6            81             0.28        0.4             6.9     0.05
 7            62             0.32        0.16            7       0.045
 8             7             0.27        0.36           20.7     0.045
 9            63             0.3         0.34            1.6     0.049
10            81             0.22        0.43            1.5     0.044
# i 4,888 more rows
# i 8 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>,
#   color <chr>
```

4. Reading in the red wine data and adding color column

```
red_wine <- read_delim("https://www4.stat.ncsu.edu/~online/datasets/red-wine.csv",
                       delim = ";")
```

```
Rows: 1599 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: ";"
dbl (12): fixed acidity, volatile acidity, citric acid, residual sugar, chlo...

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.
```

```
#red_wine

colnames(red_wine) <- wine_names
red_wine
```

```
# A tibble: 1,599 x 12
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1           7.4             0.7         0              1.9     0.076
 2           7.8             0.88        0              2.6     0.098
 3           7.8             0.76        0.04           2.3     0.092
 4          11.2             0.28        0.56           1.9     0.075
 5           7.4             0.7         0              1.9     0.076
 6           7.4             0.66        0              1.8     0.075
 7           7.9             0.6         0.06           1.6     0.069
 8           7.3             0.65        0              1.2     0.065
 9           7.8             0.58        0.02           2       0.073
10           7.5             0.5         0.36           6.1     0.071
# i 1,589 more rows
# i 7 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>
```

```
red_wine_final <- red_wine |>
  mutate(color = "red")

red_wine_final
```

```
# A tibble: 1,599 x 13
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1           7.4             0.7         0              1.9     0.076
 2           7.8             0.88        0              2.6     0.098
 3           7.8             0.76        0.04           2.3     0.092
 4          11.2             0.28        0.56           1.9     0.075
 5           7.4             0.7         0              1.9     0.076
 6           7.4             0.66        0              1.8     0.075
 7           7.9             0.6         0.06           1.6     0.069
 8           7.3             0.65        0              1.2     0.065
 9           7.8             0.58        0.02           2       0.073
10           7.5             0.5         0.36           6.1     0.071
# i 1,589 more rows
# i 8 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>,
#   color <chr>
```

5. Combining wine datasets

```r
wine_data <- dplyr::bind_rows(white_wine_final, red_wine_final)
wine_data
```

```
# A tibble: 6,497 x 13
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1             7             0.27        0.36           20.7     0.045
 2            63             0.3         0.34            1.6     0.049
 3            81             0.28        0.4             6.9     0.05
 4            72             0.23        0.32            8.5     0.058
 5            72             0.23        0.32            8.5     0.058
 6            81             0.28        0.4             6.9     0.05
 7            62             0.32        0.16            7       0.045
 8             7             0.27        0.36           20.7     0.045
 9            63             0.3         0.34            1.6     0.049
10            81             0.22        0.43            1.5     0.044
# i 6,487 more rows
# i 8 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>,
#   color <chr>
```

6. Starting chain to filter data

```r
wine_data_filter <- wine_data |>
  filter(quality > 6.5 & alcohol < 132)

wine_data_filter
```

```
# A tibble: 1,206 x 13
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1            66             0.16        0.4             1.5     0.044
 2            66             0.17        0.38            1.5     0.032
 3            62             0.66        0.48            1.2     0.029
 4            62             0.66        0.48            1.2     0.029
 5            64             0.31        0.38            2.9     0.038
 6            68             0.26        0.42            1.7     0.049
 7            72             0.32        0.36            2       0.033
 8            74             0.18        0.31            1.4     0.058
 9            66             0.25        0.29            1.1     0.068
10            62             0.16        0.33            1.1     0.057
```

```
# i 1,196 more rows
# i 8 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>,
#   color <chr>
```

7. Continuing chain to sort

```
wine_data_filter <- wine_data |>
  filter(quality > 6.5 & alcohol < 132) |>
  arrange(desc(quality))

wine_data_filter
```

```
# A tibble: 1,206 x 13
   fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
 1            91             0.27        0.45           10.6     0.035
 2            66             0.36        0.29            1.6     0.021
 3            74             0.24        0.36            2       0.031
 4            69             0.36        0.34            4.2     0.018
 5            71             0.26        0.49            2.2     0.032
 6            62             0.66        0.48            1.2     0.029
 7            62             0.66        0.48            1.2     0.029
 8            68             0.26        0.42            1.7     0.049
 9            67             0.23        0.31            2.1     0.046
10            67             0.23        0.31            2.1     0.046
# i 1,196 more rows
# i 8 more variables: free_sulfur_dioxide <dbl>, total_sulfur_dioxide <dbl>,
#   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <dbl>,
#   color <chr>
```

8. Continuing chain to filter

```
wine_data_filter <- wine_data |>
  filter(quality > 6.5 & alcohol < 132) |>
  arrange(desc(quality)) |>
  select(matches("acid"), alcohol, color, quality)

wine_data_filter
```

```
# A tibble: 1,206 x 6
   fixed_acidity volatile_acidity citric_acid alcohol color quality
```

```
           <dbl>            <dbl>       <dbl>   <dbl> <chr>   <dbl>
 1            91             0.27        0.45     104 white       9
 2            66             0.36        0.29     124 white       9
 3            74             0.24        0.36     125 white       9
 4            69             0.36        0.34     127 white       9
 5            71             0.26        0.49     129 white       9
 6            62             0.66        0.48     128 white       8
 7            62             0.66        0.48     128 white       8
 8            68             0.26        0.42     105 white       8
 9            67             0.23        0.31     107 white       8
10            67             0.23        0.31     107 white       8
# i 1,196 more rows
```

9. Continuing chain to add columns

```
wine_data_filter <- wine_data |>
  filter(quality > 6.5 & alcohol < 132) |>
  arrange(desc(quality)) |>
  select(matches("acid"), alcohol, color, quality) |>
  group_by(quality) |>
  mutate(alcohol_mean = mean(alcohol), alcohol_sd = sd(alcohol))

wine_data_filter
```

```
# A tibble: 1,206 x 8
# Groups:   quality [3]
   fixed_acidity volatile_acidity citric_acid alcohol color quality alcohol_mean
           <dbl>            <dbl>       <dbl>   <dbl> <chr>   <dbl>        <dbl>
 1            91             0.27        0.45     104 white       9        122.
 2            66             0.36        0.29     124 white       9        122.
 3            74             0.24        0.36     125 white       9        122.
 4            69             0.36        0.34     127 white       9        122.
 5            71             0.26        0.49     129 white       9        122.
 6            62             0.66        0.48     128 white       8         94.1
 7            62             0.66        0.48     128 white       8         94.1
 8            68             0.26        0.42     105 white       8         94.1
 9            67             0.23        0.31     107 white       8         94.1
10            67             0.23        0.31     107 white       8         94.1
# i 1,196 more rows
# i 1 more variable: alcohol_sd <dbl>
```