Week 3 Notes

Week 3

We want to be able to read in raw data and manipulate it, combine data sources (through SQL style joins), summarize data to glean insights, apply common analysis methods (predictive modeling), and communicate effectively (through dashboards).

R Packages

R packages already loaded (also referred to as libraries, modules, etc)
-CRAN houses all of the approved R Packages
-plenty of other packages on places like GitHub

"Base R" - package that come by default in RStudio (Global Environment)
-If there are the same functions in multiple packages, R will default to the most recent one
-base package has c(), data.frame(), list(),...

Installing an R Package
-Install package using code, menus, or Packages Tab
-Tidyverse
-install.packages("dplyr") –> dplyr is part of the tidyverse
-download file from internet to local machine (typically only one time) and then bring into R
-once downloaded, use the library() or require() to access it, library("dplyr")
-library and require are very similar by library throws an error if no package and require returns FALSE

Set Packages to Load Automatically
-access .Rprofile file

-not recommended to do this for collaboration

Accessing a Package in R Session
- to see everything –> ls("package:dplyr")

Call Functions from Library
-Call without loading the full library with ::
-if you just want one particular dataset
-dplyr::filter(iris, Species == "virginica")
-helps so you don't overwrite duplicate functions with another library

Example:
-Package to create a .pdf from .qmd
-You can download repo locally using the terminal by doing >git clone https://
-switch format at the top of the .qmd file to pdf instead of html
-install package in console install.packages("tinytex")
-run library("tinytex") to access, will now be in environment
-run install_tinytex - downloads a minimal tex so you can output to pdf
-cntrl+shift+k to export

terminal to push to git
git add .
then git commit -m "commit message"
then git push

Reading Delimited Data

Reading a CSV file:

```r
library(readr)
air_qaulity_data <- read_csv("https://www4.stat.ncsu.edu/~online/datasets/AirQuality.csv")
```

```
New names:
Rows: 9471 Columns: 18
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(2): Date, Time dbl (14): ...1, CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT),
PT08.S2(NMHC), NOx(... lgl (2): ...17, ...18
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.
* `` -> `...1`
* `...16` -> `...17`
* `...17` -> `...18`
```

```r
air_qaulity_data
```

```
# A tibble: 9,471 x 18
    ...1 Date       Time     `CO(GT)` `PT08.S1(CO)` `NMHC(GT)` `C6H6(GT)`
   <dbl> <chr>      <chr>       <dbl>         <dbl>      <dbl>      <dbl>
 1     1 10/03/2004 18.00.00      2.6          1360        150       11.9
 2     2 10/03/2004 19.00.00      2            1292        112        9.4
 3     3 10/03/2004 20.00.00      2.2          1402         88        9
 4     4 10/03/2004 21.00.00      2.2          1376         80        9.2
 5     5 10/03/2004 22.00.00      1.6          1272         51        6.5
 6     6 10/03/2004 23.00.00      1.2          1197         38        4.7
 7     7 11/03/2004 00.00.00      1.2          1185         31        3.6
 8     8 11/03/2004 01.00.00      1            1136         31        3.3
 9     9 11/03/2004 02.00.00      0.9          1094         24        2.3
10    10 11/03/2004 03.00.00      0.6          1010         19        1.7
# i 9,461 more rows
# i 11 more variables: `PT08.S2(NMHC)` <dbl>, `NOx(GT)` <dbl>,
#   `PT08.S3(NOx)` <dbl>, `NO2(GT)` <dbl>, `PT08.S4(NO2)` <dbl>,
#   `PT08.S5(O3)` <dbl>, T <dbl>, RH <dbl>, AH <dbl>, ...17 <lgl>, ...18 <lgl>
```

```
air_qaulity_data$`CO(GT)`[1:10]
```

```
 [1] 2.6 2.0 2.2 2.2 1.6 1.2 1.2 1.0 0.9 0.6
```

Reading in a Fixed Width Field (FWF)

```
cigarettes_data <-
read_fwf("https://www4.stat.ncsu.edu/~online/datasets/cigarettes.txt",fwf_widths(c(17, 5, 9,
                                                                               c("brand","tar
```

```
Rows: 23 Columns: 5
-- Column specification --------------------------------------------------------

chr (1): brand
dbl (4): tar, nicotine, weight, co

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.
```

```
#widths by counting: Alpine           14.1 0.86     0.9853 13.6
```