**R Basics and R Markdown**

[Base R cheat sheet](https://www.rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf)

[R Markdown cheat sheet](https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf)

[Homework 1](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_1_Rino_Gwen.html) unique()

[Lesson 3.2](https://r-clements.github.io/slides/session_3/intro_to_data_science_session_3_2.html#7)

[Lesson 3.3](https://r-clements.github.io/slides/session_3/intro_to_data_science_session_3_3.html#4)

* Viewing data (View, glimpse, head, tail, names, class)

[Lesson 4.3](https://r-clements.github.io/slides/session_4/intro_to_data_science_session_4_3.html#11)

* Data types

**Data Import/Export**

[Lesson 4.2](https://r-clements.github.io/slides/session_4/intro_to_data_science_session_4_2.html#9)

[Homework 4](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_4_Rino_Gwen.html)

* readr package
* csv, read\_delim
* R-specific formats

**Ethics**

[Lesson 5.1](https://r-clements.github.io/slides/session_5/intro_to_data_science_session_5_1.html#13)

* p-hacking
* reproducibility
* consent

**Visualization with ggplot2**

[ggplot2 cheatsheet](https://drive.google.com/drive/folders/0BxlAAp9pz8ArUTNsU0hUMkYwbW8)

[Homework 2](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_2_Rino_Gwen.html)

[Lesson 2.3](https://r-clements.github.io/slides/session_2/intro_to_data_science_session_2_3.html" \l "7)

* Aesthetics
* Geoms (point, smooth)
* Statistical transformations (geom\_bar with count and proportion)
* Labels and themes

[Lesson 3.1](https://r-clements.github.io/slides/session_3/intro_to_data_science_session_3_1.html#8)

* Position (identity, dodge, fill, jitter)
* coord\_flip
* Geographic coordinates
* Canonical data graphics - univariate
  + Bar chart (categorical variables)
  + Histogram (continuous variables)
  + Density plot (continuous variables)
* Canonical data graphics - multivariate
  + Scatterplot (geom\_point) x = continuous, y = continuous
  + Boxplot (geom\_boxplot) x = categorical, y = continuous
  + Column (geom\_col) x = categorical, y = continuous (see [Homework 2](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_2_Rino_Gwen.html))

[Homework 3](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_3_Rino_Gwen.html) dealing with overplotting with geom\_jitter and geom\_count

**Data Wrangling**

[Cheatsheet](https://drive.google.com/drive/folders/1jJbNMxP_2JErp5akRDlSl3kq5w9ZRlYb)

[Lesson 3.3](https://r-clements.github.io/slides/session_3/intro_to_data_science_session_3_3.html#4)

* Tibbles (subsetting at [Lesson 4.3](https://r-clements.github.io/slides/session_4/intro_to_data_science_session_4_3.html#11) and [Homework 4](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_4_Rino_Gwen.html))
* dplyr
  + select (subset of columns) see [Homework 3](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_3_Rino_Gwen.html)
    - ends\_with
    - contains
    - matches
    - num\_range
    - everything
  + filter (subset of rows) see [Homework 3](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_3_Rino_Gwen.html)
    - comparisons
    - logicals
    - is.na (see also [Lesson 4.1](https://r-clements.github.io/slides/session_4/intro_to_data_science_session_4_1.html#7))
  + mutate (add or modify columns)
  + arrange (sort the rows)
  + summarize (aggregate the data by groups)
    - see also [Lesson 4.1](https://r-clements.github.io/slides/session_4/intro_to_data_science_session_4_1.html#7) and [Homework 4](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_4_Rino_Gwen.html)
      * counts with and without summarize
      * measures of center, spread, rank, position
      * proportion of a logical True = 1, False = 0

[Lesson 4.3](https://r-clements.github.io/slides/session_4/intro_to_data_science_session_4_3.html#11)

* tidyr (see [Homework 4](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_4_Rino_Gwen.html))
  + gather
  + spread
  + separate
  + unite

[Lesson 6.3](https://r-clements.github.io/slides/session_6/intro_to_data_science_session_6_3.html#3)

* recoding variables (character -> factor, e.g.)
* ordered and unordered factors – forcats package
* case\_when
* lubridate package (numbers -> dates)

**Exploratory Data Analysis** (EDA)

[Lesson 5.2](https://r-clements.github.io/slides/session_5/intro_to_data_science_session_5_2.html#9)

* robust (resistant to outliers)
* variation/spread
  + variance
  + standard deviation
  + inter-quartile range
  + visualization with histograms or bar graphs
  + measures of center
  + shape/distribution
* outliers

[Lesson 5.3](https://r-clements.github.io/slides/session_5/intro_to_data_science_session_5_3.html#4)

* covariation and correlation
  + one categorical and one continuous variable (boxplot)
  + two categorical variables (heat map)
  + two continuous variables (scatterplot)
* patterns and models

**Statistics**

[Lesson 6.2](https://r-clements.github.io/slides/session_6/intro_to_data_science_session_6_2.html#2)

* populations and samples
* uncertainty and standard error

[Homework 6](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_6_Rino_Gwen.html): central limit theorem

[Lesson 6.3](https://r-clements.github.io/slides/session_6/intro_to_data_science_session_6_3.html#3)

* Simpson’s Paradox ([Homework 6](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_6_Rino_Gwen.html))

[Lesson 6.4](https://r-clements.github.io/slides/session_6/intro_to_data_science_session_6_4.html#6)

* Statistical learning
  + Inference
  + Prediction
* Simple linear regression
  + Sum of residuals (squared)
  + Interpreting coefficients

[Lesson 7.1](https://r-clements.github.io/slides/session_7/intro_to_data_science_session_7_1.html#5) ([Homework 7](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_7_Rino_Gwen.html))

* More simple linear regression
  + Significance of coefficients
  + Accuracy of coefficients (p-value)
  + Model accuracy (R2)
* broom package
* Multiple linear regression
  + Interpreting coefficients of categorical predictor variables

[Lesson 7.2](https://r-clements.github.io/slides/session_7/intro_to_data_science_session_7_2.html#4)

* Supervised vs unsupervised learning
* Regression (when Y is numeric) vs classification (when Y is categorical)
  + NB use regression to find *probability* of a classification vs classification to predict class

[Lesson 7.3](https://r-clements.github.io/slides/session_7/intro_to_data_science_session_7_3.html#6)

* Evaluating predictive regression models
  + Root mean squared error ([Homework 7](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_7_Rino_Gwen.html))
  + Mean absolute error
* The bias-variance trade off

[Homework 7](file:///Users/gwenrino/Documents/Data_Science_Learning/compscix-415-2-assignments/Homework_7_Rino_Gwen.html)

* Create training set and test set