Exam 2 Review

1. Cleansing and Data Manipulation
2. Why data cleansing is important
3. Data cleansing/transformation is an essential (usually the most time consuming) part of a data analytics project.
4. A properly prepared dataset is the prerequisite of statistical modeling, prediction, and inference.
5. The “Garbage in, garbage out” rule applies.
6. Apply methods to manipulate strings
7. General
8. tolower(): translate to lower case
9. toupper(): translate to upper case
10. nchar(): count the number of characters
11. trimws(): trim whitespace [ \t\r\n]
12. stringi::stri\_reverse(): reverse a string
13. Pattern Matching and Replacement
14. regexpr(): match patterns
15. grep(): match patterns
16. sub(): replace the first match
17. gsub(): replace all matches
18. Substrings
19. substr(x, start, stop)
20. substring(text, first, last = 1000000L)
21. Split Strings
22. strsplit(x, split, fixed = FALSE, perl = FALSE, useBytes = FALSE)
23. Concatenate
24. paste()
25. paste0()
26. Notes:
27. To count the number of characters in a string, do NOT use length(), use nchar().
28. To find the position of matches in a string, do NOT use grep(), use regexpr() to find the position of the first match and use gregexpr() to get positions of all matches.
29. Forward Pipe Operator
30. 
31. Create, Recode, Rename, and Manipulate Variables
32. Create Variables
33. $



1. Transform()



1. dyply::mutate()



Not sure what’s the difference

1. Recode Variables
2. Base r



1. dyplyr:recode



1. Rename Variables
2. Base R



1. dplyr rename



1. Sort
2. order()



1. arrange()



1. Subset
2. which()



1. subset()



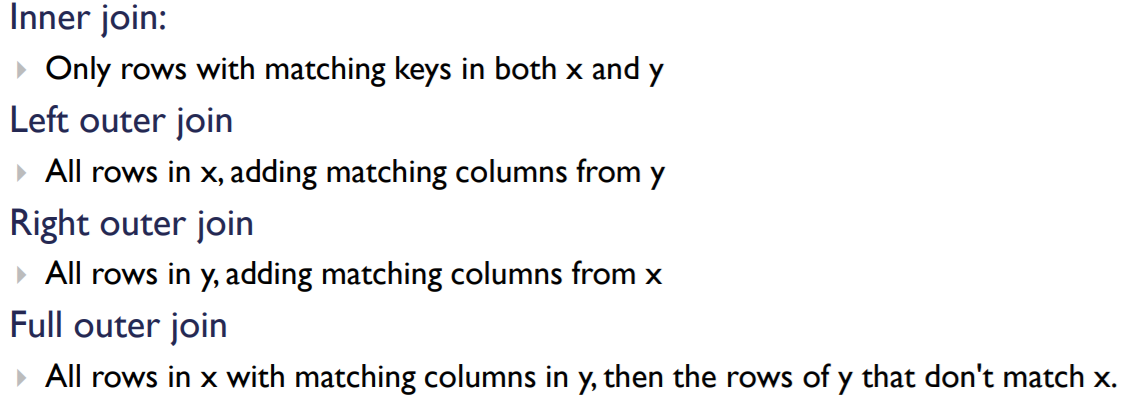
1. filter()



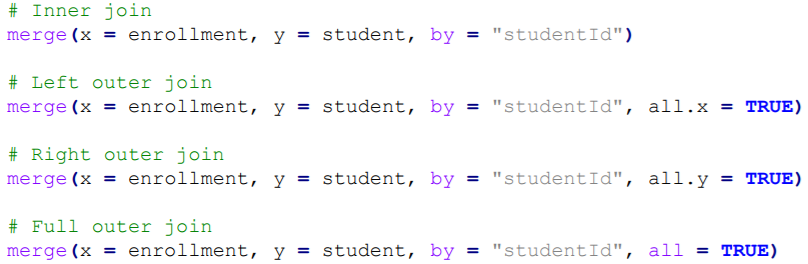
1. select()



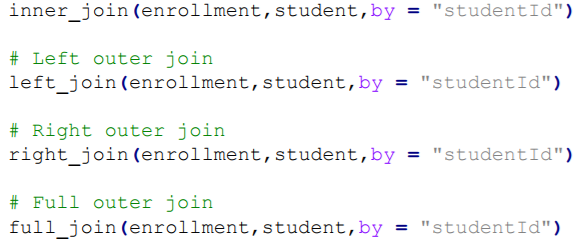
1. Join/Merge



1. Base R



1. dplyr



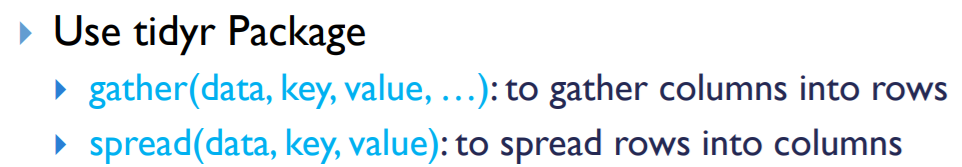
1. Aggregate
2. Aggregate(): Group data and calculate summary statistics



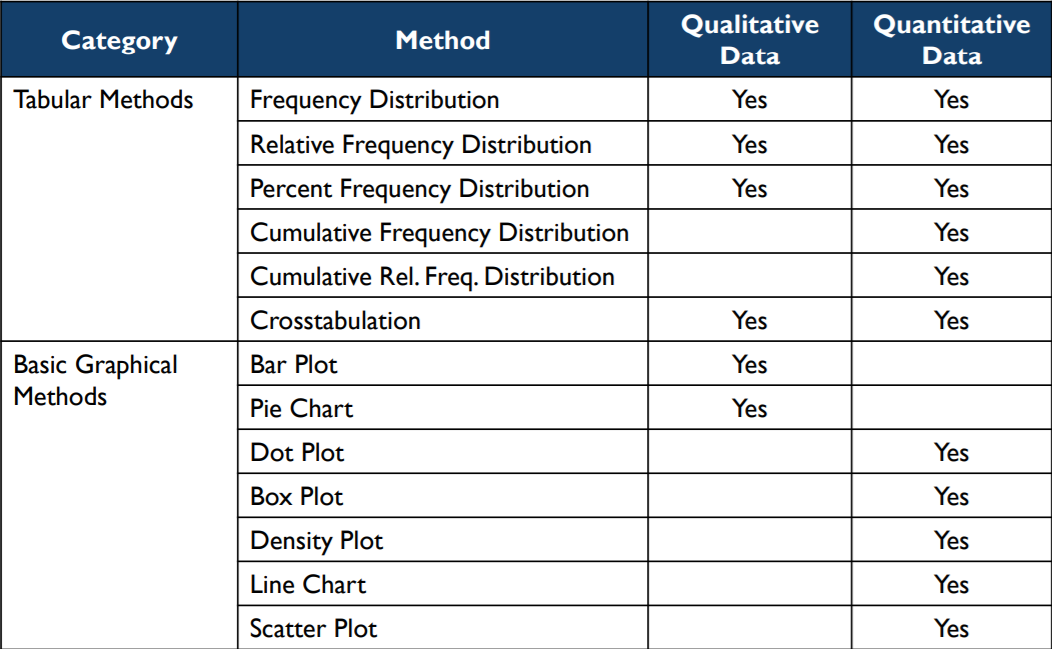
1. group\_by(): Group data
2. summarize(): Calculate summary statistics

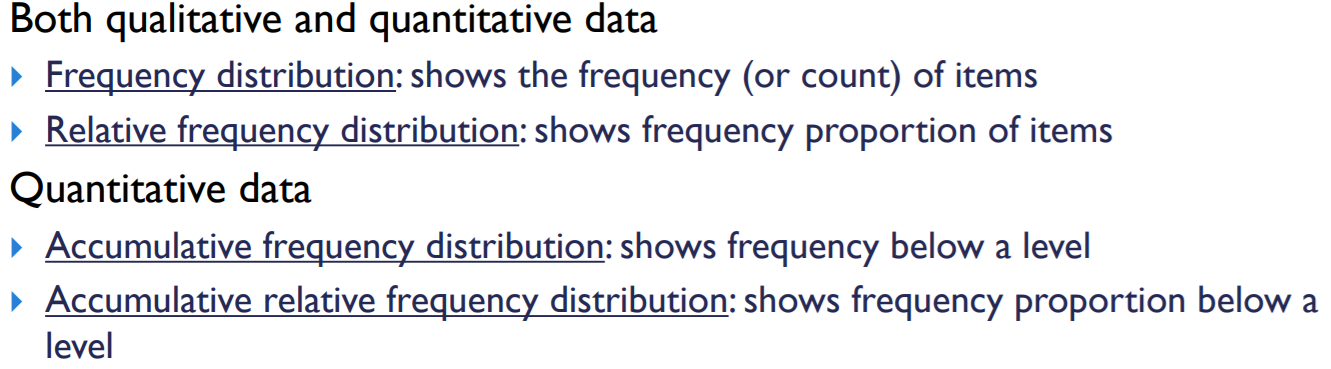


1. Reshape

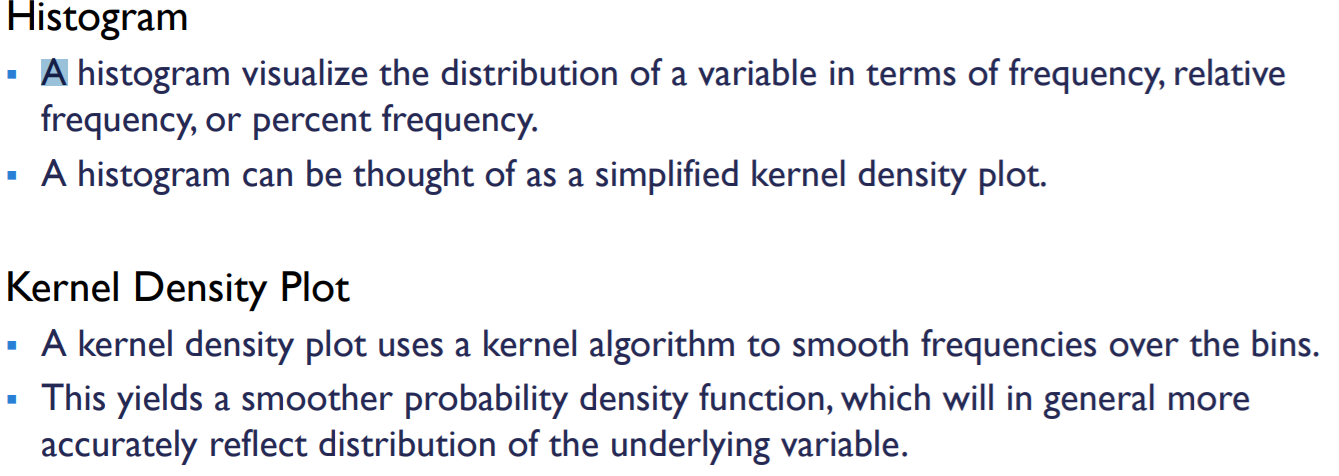


1. Data Summarization and Visualization
2. Choose appropriate tabular and basic graphic methods for different types of data (quantitative vs qualitative)

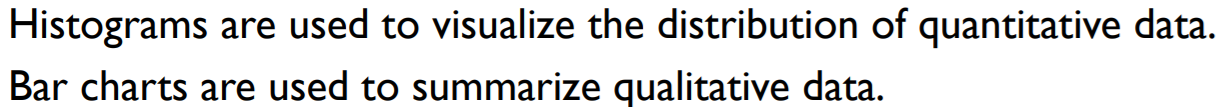




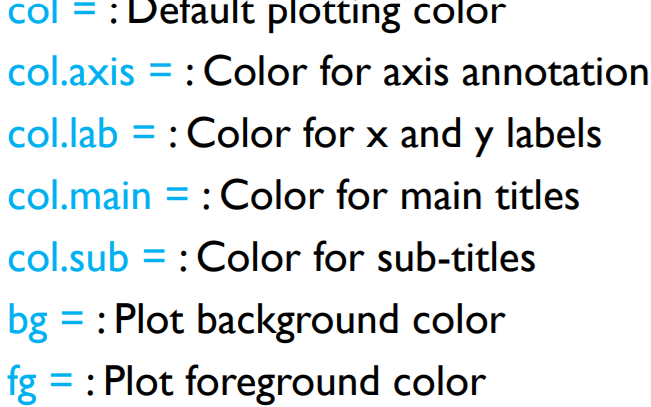
1. Histograms
2. Histograms vs density plots



1. Histograms vs bar charts



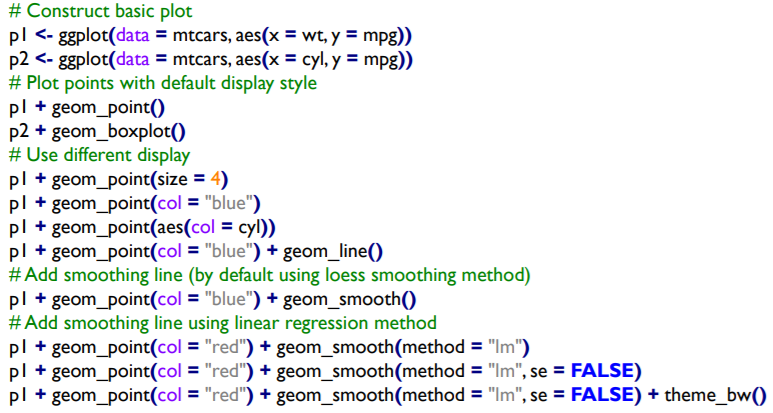
1. Line plots
2. LOWESS = Locally Weighted Scatterplot Smoothing
3. Color parameters



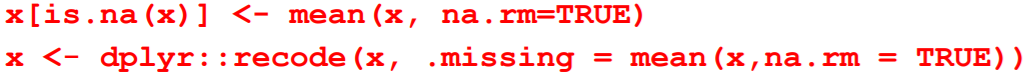
1. ggplot2
2. qplot: use quick plotting qplot() function to create basic graphs



1. ggplot:



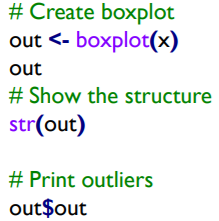
1. Plus operator: typically construct a plot incrementally, using the “+” operator to add layers to the existing ggplot object.
2. Uncommon visualization definitions
3. Spatial data: Represents location, size, and shape of physical objects by numbers in a geographic coordinate system
4. Hexagonal binning: A form of bivariate histogram used for visualizing the structure of datasets with large numbers
5. Mosaic plot: Graphical method for visualizing data from two or more qualitative variables
6. Heat map: Using color to represent a third dimension
7. Data Exploration
8. Missing data
9. Listwise deletion: Remove missing values or cases from analysis
10. is.na() determine which elements are missing
11. na.omit() Removing missing values or cases (listwise deletion)
12. na.rm = T Excludes missing values from analysis
13. Imputation: replacing missing data with substituted value
14. Replace with mean



First = Base R

Second = dplyr

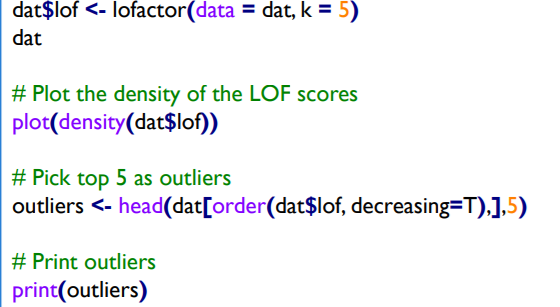
1. Outliers
2. Recode: When outliers are incorrectly recorded in the dataset
3. Remove: when outliers are incorrectly included in the dataset
4. Keep: when outliers are correct values in the dataset
5. Detecting potential outliers
6. Boxplot rule: Values beyond [Q1 – 1.5 \* IQR, Q3 + 1.5 \* IQR] are outliers



1. Z-Score: Any data value with a z-score less than -3 or greater than 3 can be regarded as an outlier



1. Density-based local outlier (LOF) = Local Outlier Factor Algorithm: If the value’s density is significantly lower than that of its neighbors (with an LOF value greater than one), the value is in a sparser region than its neighbors; suggesting the value is an outlier



1. Removal of outliers

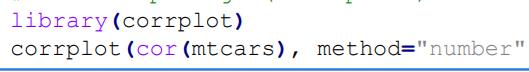


1. Correlation analysis
2. Covariance = Direction of the linear association
3. > 0 = positive relationship
4. <0 = negative relationship

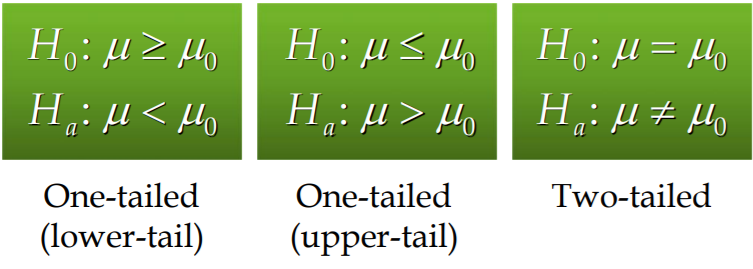


1. Correlation = Direction and strength of the linear association
2. Correlation coefficient = normalized measurement of linear association between variables (-1 to 1)

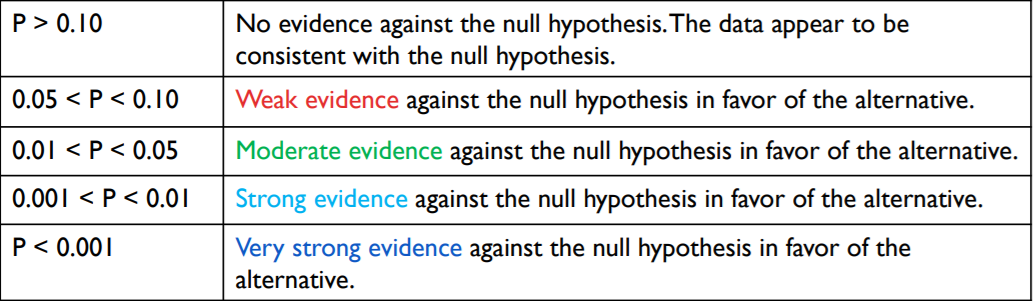




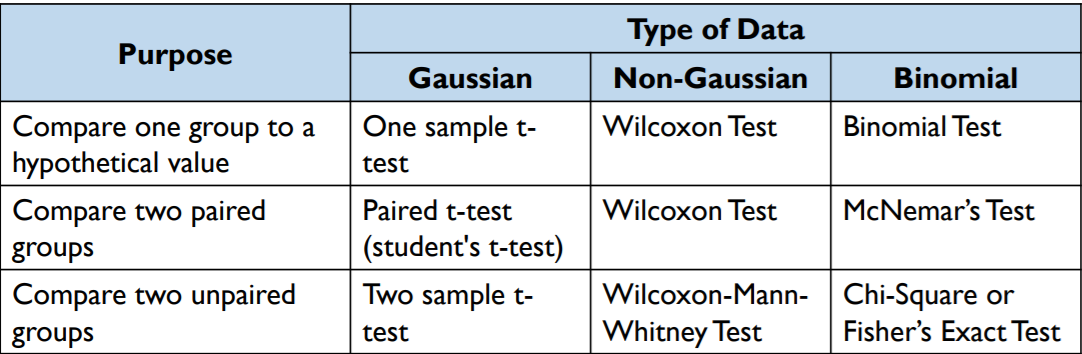
1. Pop vs. sample
2. Pop = the whole body of subjects that are of interest to us in a particular study
3. Sample = portion of the population from which we collect data
4. Hypothesis Testing:
5. Null hypothesis H0 : a tentative assumption about a population parameter
6. Alternative hypothesis Ha : the opposite of the null hypothesis

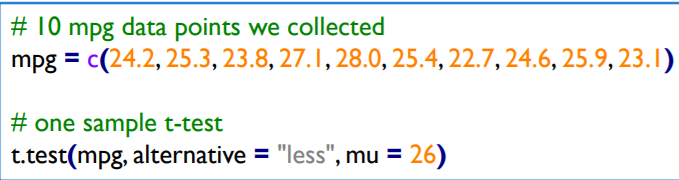


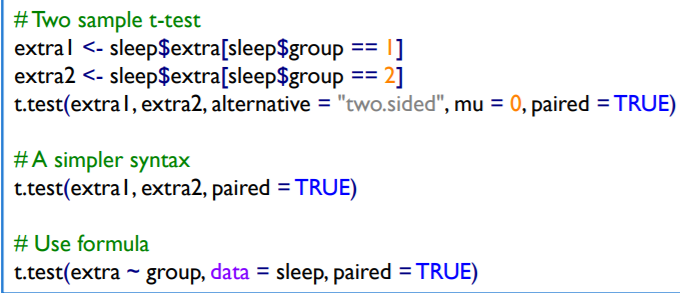
1. Pvalue: the probability of obtaining the observed or more extreme effect in your sample data, assuming that the null hypothesis is true; the smaller the pvalue the stronger the evidence against the null hypothesis



1. One and two sample tests







1. ANOVA (analysis of variance or regression): how to compare means of some variable between two or more groups
2. Regression Analysis

Regression is about estimating relationships among variables

Regression is a statistical technique that attempts to build a function of independent variables (IVs) or predictors to predict or explain a dependent variable (DV).

