R CRASH COURSE

Missing data

\*Options for dealing with missing data (NAs)

-Remove any row that contains NA in it

#Command: drop\_na(data)

#Downside: does it to all NAs, regardless of where they are in the dataset

\*Even if it is data that isn’t needed for calculations, is duplicated, etc

-Remove rows that contain NAs in only certain columns

#Command: filter(data, !in.na(colname))

#Pros: control exactly which NAs you want to remove, not globally

-Change your NAs to something else (in this case zero)

#Command: data %>% mutate(colname = ifelse(is.na(colname), 0, colname)

#Downside: NAs and 0 are not always the same thing, must have a very good

reason for doing this

\*Reasons for missing data

-Random missing data

#Hypothesis: data is missing at random

#If this is the case and the dataset it large, it is safe to remove missing data

-There is a correlation between missing data and data count

#removing this data may inadvertently bias the data

Intro to plotting

\*ggplot

-putting things inside the aes() parameters will do that thing to everything

-for example, putting color=variable inside will maintain that grouping even when

adding line of best fit

-You will get a separate line of best fit for each level of that variable

Data wrangling