Alex Gibbons

March 18, 2024

CMPT 363 – Data Mining

Tidyverse code and output

# install.packages("tidyverse")

library("tidyverse")

library("readxl")

mortality\_data <- read\_excel("C:/Users/Kelly/Downloads/child\_mortality\_rates.xlsx",sheet=1)

# printing tibble displays first 10 rows of tibble by default

mortality\_dataA screenshot of a computer screen

Description automatically generated

# use print function to print specific num rows

print(mortality\_data,n=150)

# to view top and bottom of dataset use head and tail function

head(mortality\_data,10) # if n is not specified, 6 is default

tail(mortality\_data)

A screenshot of a computer screen

Description automatically generated

# summary statistics of data set

summary(mortality\_data)A screenshot of a computer

Description automatically generated

# to only get certain columns use select function

select(mortality\_data, Year, "01-04 Years")

A screenshot of a computer program

Description automatically generated

# to select all BUT certain columns, preceded column name with -

select(mortality\_data, -Year)

A screenshot of a computer code

Description automatically generated

# combine with summary function to get a summary of column data

summary(select(mortality\_data,"01-04 Years"))

A blue screen with orange text

Description automatically generated

# may need to melt data from wide to long

# use pivot\_longer function

mortality\_long <- pivot\_longer(mortality\_data,cols=c("01-04 Years","05-09 Years","10-14 Years","15-19 Years"),names\_to = "AgeGroup", values\_to = "DeathRate")

mortality\_long

A computer screen shot of a blue screen

Description automatically generated

# use mutate and rename functions to directly modify data within the tibble

mortality\_long<-mutate(mortality\_long,DeathRate=DeathRate \* 100000)

# death rate multiplied by 100,000 to get num people dead

A computer screen shot of a number

Description automatically generated

# add calculated column for respective decade

mortality\_long <- mutate(mortality\_long, Decade = (Year%/%10) \*10)

A screenshot of a computer screen

Description automatically generated

# modify two columns in a single statement

# combine two previous commands

mortality\_long <- mutate(mortality\_long,DeathRate=DeathRate\*100000,Decade=(Year%/%10)\*10)

# rename DeathRate column

mortality\_long<-rename(mortality\_long, DeathsPer100K=DeathRate)

A screenshot of a computer screen

Description automatically generated

# saving tibble to file

# use saveRDS function

saveRDS(mortality\_long,"C:/Users/Kelly/Downloads/mortality\_long.rds")

# use readRDS funcition to read tibble from file

readRDS("C:/Users/Kelly/Downloads/mortality\_long.rds")

A screenshot of a computer

Description automatically generated

# common functions for calculating summary columns for a tibble:

# group\_by, summarize, n, mean

# pipe operator is also common when summarizing grouped data

# Ctrl+Shift+M - pipe operator shortcut

# calculate mean num death per 100K for each decade and find num rows each decade

mortality\_long %>% group\_by(Decade) %>% summarize(MeanDeaths = mean(DeathsPer100K),Count = n())

A blue screen with yellow text

Description automatically generated

# visualizing data using ggplot

# ggplot(data,mapping)

# aes(x,y,color,fill)

# geom\_line()

ggplot(mortality\_long, aes(x=Year,y=DeathsPer100K,color=AgeGroup)) + geom\_line()

