data exploration

list.files()

## [1] "abbrv\_headscan\_demo.xlsx" "data cleaning code.Rmd"   
## [3] "demographic-data-exploration.docx" "demographic-data-exploration.Rmd"   
## [5] "demographic data exploration.Rmd" "Headscan Demographics.Rproj"   
## [7] "headscan\_demographics.xlsx" "Kayna sample data.xlsx"   
## [9] "sample\_demographics\_draftgraph.jpeg"

setwd("C:\\Users\\19177\\OneDrive - Colostate\\Desktop\\Dissertation\\data analysis\\data analysis R")  
library(readxl)  
demographics<-read\_excel("headscan\_demographics.xlsx")  
head(demographics)

## # A tibble: 6 × 33  
## ID Age Gender Ethnicity `Annual househ…` `Highest level…`  
## <chr> <dbl> <chr> <chr> <chr> <chr>   
## 1 400-20201012-001 54 Male Asian/Asian A… Between $150,00… Completed a pos…  
## 2 400-20201012-002 31 Male Black, Africa… Between $50,000… Graduated from …  
## 3 400-20201012-003 49 Female White/Caucasi… Between $100,00… Completed a pos…  
## 4 400-20201012-004 49 Male White/Caucasi… Between $100,00… Completed a pos…  
## 5 400-20201012-005 34 Male White/Caucasi… Between $75,000… Graduated from …  
## 6 400-20201012-006 49 Male White/Caucasi… Between $100,00… Graduated from …  
## # … with 27 more variables: `Employment status` <chr>, `Zip Code` <dbl>,  
## # `Diagnosis: Cataracts` <chr>, `Diagnosis: Glaucoma` <chr>,  
## # `Diagnosis: Macular Degeneration (AMD - wet or dry types)` <chr>,  
## # `Diagnosis: Strabismus (eye-turn, lazy eye)` <chr>,  
## # `Diagnosis: Amblyopia` <chr>,  
## # `Diagnosis: Cornea Disease (keratoconus, scarring, severe dry eye)` <chr>,  
## # `Diagnosis: Other disease not listed` <chr>, …

colnames(demographics)

## [1] "ID"   
## [2] "Age"   
## [3] "Gender"   
## [4] "Ethnicity"   
## [5] "Annual household income"   
## [6] "Highest level of education"   
## [7] "Employment status"   
## [8] "Zip Code"   
## [9] "Diagnosis: Cataracts"   
## [10] "Diagnosis: Glaucoma"   
## [11] "Diagnosis: Macular Degeneration (AMD - wet or dry types)"   
## [12] "Diagnosis: Strabismus (eye-turn, lazy eye)"   
## [13] "Diagnosis: Amblyopia"   
## [14] "Diagnosis: Cornea Disease (keratoconus, scarring, severe dry eye)"   
## [15] "Diagnosis: Other disease not listed"   
## [16] "Frequency usage corrective lenses"   
## [17] "Laser eye surgery or other surgeries"   
## [18] "Type of correction for distance visual tasks"   
## [19] "Type of correction for near visual tasks"   
## [20] "Preferred type of vision correction"   
## [21] "Frequent use: Desktop Computer"   
## [22] "Frequent use: Laptop Computer"   
## [23] "Frequent use: Tablet (e.g., iPad, Surface)"   
## [24] "Frequent use: Smartphone"   
## [25] "Frequent use: Smart watch"   
## [26] "Frequent use: Fitness tracker (e.g., Fitbit, Garmin)"   
## [27] "Frequent use: Video game console (e.g., Xbox, Playstation)"   
## [28] "Frequent use: Handheld gaming device (e.g., PSP, Nintendo 3DS)"   
## [29] "Frequent use: Smart home assistant (e.g., Amazon Echo, Google Home, Facebook Portal)"  
## [30] "Frequent use: Video streaming device (e.g., Apple TV, Fire TV, Chromecast, Roku)"   
## [31] "Frequent use: VR headset (e.g., Oculus, HTC Vive)"   
## [32] "Frequent use: Smart glasses (e.g., Snap Spectacles, Bose Frames, Amazon Echo Frames)"  
## [33] "Using new technology"

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

demographics<-select(.data=demographics, 1:4)  
  
demographics<-rename(.data=demographics,  
 age=Age,  
 gender=Gender,  
 ethnicity=Ethnicity)  
  
demographics<-mutate(.data=demographics,  
 age = as.numeric(age),  
 gender= as.factor(gender),  
 ethnicity=as.factor(ethnicity))  
  
demographics<-filter(.data=demographics,  
 !is.na(age),  
 !is.na(ethnicity),  
 !is.na(gender))  
   
mean(demographics$age, na.rm=TRUE)

## [1] 36.39542

median(demographics$age, na.rm=TRUE)

## [1] 37

sd(demographics$age, na.rm=TRUE)

## [1] 11.52018

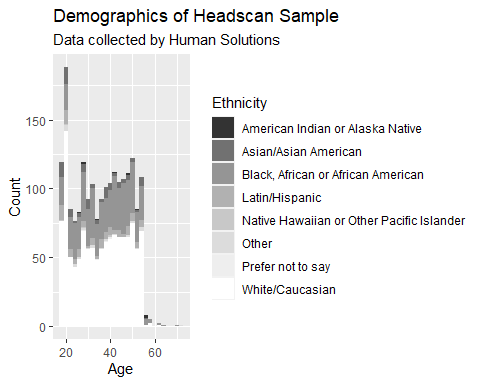
min(demographics$age, na.rm=TRUE)

## [1] 18

max(demographics$age, na.rm=TRUE)

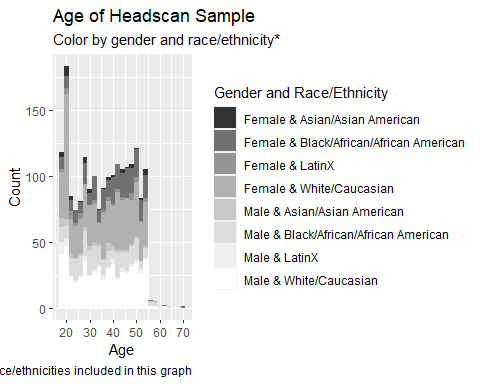
## [1] 72

ggplot(data=demographics)+  
 geom\_histogram(binwidth=2, aes(x=age, fill=ethnicity))+  
 scale\_fill\_grey(start = 0.2, end = 1)+  
 labs(title= "Demographics of Headscan Sample",  
 subtitle = "Data collected by Human Solutions",  
 fill= "Ethnicity",  
 y="Count",  
 x="Age")

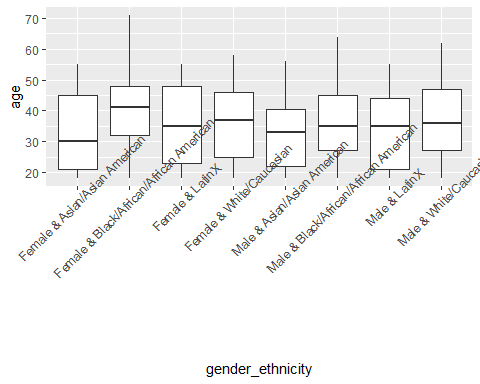


asian\_women<-filter(.data=demographics,  
 gender == "Female",  
 ethnicity == "Asian/Asian American")  
asian\_women<-select(.data=asian\_women, c(2))  
asian\_women$gender\_ethnicity <- "Female & Asian/Asian American"  
   
asian\_men<-filter(.data=demographics,  
 gender == "Male",  
 ethnicity == "Asian/Asian American")  
asian\_men<-select(.data=asian\_men, c(2))  
asian\_men$gender\_ethnicity <- "Male & Asian/Asian American"  
  
black\_women<-filter(.data=demographics,  
 gender == "Female",  
 ethnicity == "Black, African or African American")  
black\_women<-select(.data=black\_women, c(2))  
black\_women$gender\_ethnicity <- "Female & Black/African/African American"  
  
black\_men<-filter(.data=demographics,  
 gender == "Male",  
 ethnicity == "Black, African or African American")  
black\_men<-select(.data=black\_men, c(2))  
black\_men$gender\_ethnicity <- "Male & Black/African/African American"  
  
white\_women<-filter(.data=demographics,  
 gender == "Female",  
 ethnicity == "White/Caucasian")  
white\_women<-select(.data=white\_women, c(2))  
white\_women$gender\_ethnicity <- "Female & White/Caucasian"  
  
white\_men<-filter(.data=demographics,  
 gender == "Male",  
 ethnicity == "White/Caucasian")  
white\_men<-select(.data=white\_men, c(2))  
white\_men$gender\_ethnicity <- "Male & White/Caucasian"  
  
latinx\_women<-filter(.data=demographics,  
 gender == "Female",  
 ethnicity == "Latin/Hispanic")  
latinx\_women<-select(.data=latinx\_women, c(2))  
latinx\_women$gender\_ethnicity <- "Female & LatinX"  
  
latinx\_men<-filter(.data=demographics,  
 gender == "Male",  
 ethnicity == "Latin/Hispanic")  
latinx\_men<-select(.data=latinx\_men, c(2))  
latinx\_men$gender\_ethnicity <- "Male & LatinX"  
  
merge\_cols<-c("age","gender\_ethnicity")  
  
demographics2 <- merge(asian\_men, asian\_women, by = merge\_cols, all=TRUE)  
demographics3 <- merge(demographics2, black\_men, by = merge\_cols, all=TRUE)  
demographics4 <- merge(demographics3, black\_women, by = merge\_cols, all=TRUE)  
demographics5 <- merge(demographics4, white\_men, by = merge\_cols, all=TRUE)  
demographics6 <- merge(demographics5, white\_women, by = merge\_cols, all=TRUE)  
demographics7 <- merge(demographics6, latinx\_men, by = merge\_cols, all=TRUE)  
demographics\_final <- merge(demographics7, latinx\_women, by = merge\_cols, all=TRUE)

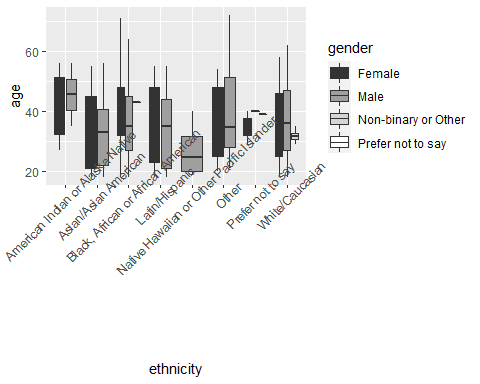
ggplot(data=demographics\_final, aes(x=age, fill=gender\_ethnicity))+  
 scale\_fill\_grey(start = 0.2, end = 1)+  
 geom\_histogram(binwidth = 2, position="stack")+  
 labs(title= "Age of Headscan Sample",  
 subtitle = "Color by gender and race/ethnicity\*",  
 y = "Count",  
 x= "Age",  
 fill="Gender and Race/Ethnicity",  
 caption="\*only male/female gender and four most common race/ethnicities included in this graph")



ggplot(data=demographics\_final)+  
 geom\_boxplot(aes(x=gender\_ethnicity, y=age))+  
 theme(axis.text.x = element\_text(angle = 45))



ggplot(data=demographics)+  
 geom\_boxplot(aes(x=ethnicity, y=age, fill=gender))+  
 scale\_fill\_grey(start = 0.2, end = 1)+  
 theme(axis.text.x = element\_text(angle = 45))



demographics$ethnicity <- recode(demographics$ethnicity,  
 "American Indian or Alaska Native" ='AI/AN',  
 "Asian/Asian American" = 'Asian',  
 "Black, African or African American" = 'Black',  
 "Latin/Hispanic" = 'LatinX',  
 "Native Hawaiian or Other Pacific Islander" = 'NH/OPI',  
 "Prefer not to say" = 'PNTS',  
 "White/Caucasian" = 'white')

freq\_table <- demographics %>% group\_by(ethnicity) %>% summarise(freq=n())  
as.data.frame(freq\_table)

## ethnicity freq  
## 1 AI/AN 8  
## 2 Asian 92  
## 3 Black 545  
## 4 LatinX 98  
## 5 NH/OPI 4  
## 6 Other 21  
## 7 PNTS 5  
## 8 white 1235

demographics$ethnicity <- factor(demographics$ethnicity, levels=freq\_table$ethnicity[order(freq\_table$freq, decreasing = T)])

ggplot(data=demographics, aes(x= age, fill=gender))+  
 scale\_fill\_grey(start = 0.2, end = 1)+  
 geom\_histogram(binwidth = 3)+  
 labs(title="Headscan Sample Demographics: Age",  
 subtitle="sorted by Race/Ethnicity, color by Gender",  
 y="Count (variable scale per group)",  
 x="Age",  
 fill="Gender")+  
 facet\_wrap(~ethnicity, scales="free\_y")+  
 scale\_y\_continuous(breaks = function(x) unique(floor(pretty(seq(0, (max(x) + 1) \* 1.1)))))

