Untitled

Daniel Zoleikhaeian

2023-06-21

Setup

Importing data

```
library(grid)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.3.1

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
df_COGS2 <- read.csv('C:\\Users\\danie\\Documents\\Joshi Lab Materials\\3 Studies Dataset\\Dataset Merg
df_COGS2 <- df_COGS2[df_COGS2$cStudy == 'COGS2', ]
nrow(df_COGS2)
```

```
## [1] 2477
```

```
head(df_COGS2)
```

```
##           X cStudy cAge cDiagnosis cEnrollmentDateYear cGender cRace
## 33535 34651  COGS2   25       CTRL          2010-07-07       M    AS
## 33536 34652  COGS2   43         SZ          2010-07-14       M    CA
## 33537 34653  COGS2   44       CTRL          2010-07-15       M    AA
## 33538 34654  COGS2   50         SZ          2010-07-16       M    CA
## 33539 34655  COGS2   49       CTRL          2010-07-19       M    CA
## 33540 34656  COGS2   43         SZ          2010-07-20       M    AA
##       cHispanicorLatino cLocationInstitution cLocationCity cLocationState
## 33535                No                 UCSD     San Diego             CA
## 33536               Yes                 UCSD     San Diego             CA
## 33537                No                 UCSD     San Diego             CA
## 33538                No                 UCSD     San Diego             CA
## 33539                No                 UCSD     San Diego             CA
## 33540                No                 UCSD     San Diego             CA
##       cLocationCounty cDiagnosis2 cDiagnosis3 cDiagnosis4
## 33535       San Diego          CS          CS          CS
## 33536       San Diego          SZ      SZSAFD      SZSAFD
## 33537       San Diego          CS          CS          CS
## 33538       San Diego          SZ      SZSAFD      SZSAFD
## 33539       San Diego          CS          CS          CS
## 33540       San Diego          SZ      SZSAFD      SZSAFD
```

```
# adding year column
df_COGS2$cEnrollmentYear <- as.numeric(substr(df_COGS2$cEnrollmentDateYear, 1,4))
head(df_COGS2)
```

```
##           X cStudy cAge cDiagnosis cEnrollmentDateYear cGender cRace
## 33535 34651  COGS2   25       CTRL          2010-07-07       M    AS
## 33536 34652  COGS2   43         SZ          2010-07-14       M    CA
## 33537 34653  COGS2   44       CTRL          2010-07-15       M    AA
## 33538 34654  COGS2   50         SZ          2010-07-16       M    CA
## 33539 34655  COGS2   49       CTRL          2010-07-19       M    CA
## 33540 34656  COGS2   43         SZ          2010-07-20       M    AA
##       cHispanicorLatino cLocationInstitution cLocationCity cLocationState
## 33535                No                 UCSD     San Diego             CA
## 33536               Yes                 UCSD     San Diego             CA
## 33537                No                 UCSD     San Diego             CA
## 33538                No                 UCSD     San Diego             CA
## 33539                No                 UCSD     San Diego             CA
## 33540                No                 UCSD     San Diego             CA
##       cLocationCounty cDiagnosis2 cDiagnosis3 cDiagnosis4 cEnrollmentYear
## 33535       San Diego          CS          CS          CS            2010
## 33536       San Diego          SZ      SZSAFD      SZSAFD            2010
## 33537       San Diego          CS          CS          CS            2010
## 33538       San Diego          SZ      SZSAFD       SZSAFD            2010
## 33539       San Diego          CS          CS           CS            2010
## 33540       San Diego          SZ      SZSAFD       SZSAFD            2010
```

```
unique(df_COGS2$cRace)
```

```
## [1] "AS"  "CA"  "AA"  "MR"  "NH"  "AE"  "UNK"
```

```
# re-encoding OT/UNK/MR into one group
df_COGS2$cRace2 <- df_COGS2$cRace
df_COGS2$cRace2[df_COGS2$cRace2 %in% c('OT', 'OT/UNK', 'MR', 'UNK')] <- 'OT/MR'
nrow(df_COGS2)
```

```
## [1] 2477
```

Helper function: Diversity Index

```
mult_ent <- function(race_prop_vec) {

  tot <- 0

  for (i in 1:length(race_prop_vec)) {
    if (race_prop_vec[i] != 0) {
      tot <- tot + race_prop_vec[i] * log(1/race_prop_vec[i])
    }
  }
  return(tot)
}
```

COGS2 Analysis + Plots

COGS2: Aggregate and By-City DI

Aggregate DI

```
pdf('40. Census_vs_Study_div_index.pdf')
head(df_COGS2)
```

```
##           X cStudy cAge cDiagnosis cEnrollmentDateYear cGender cRace
## 33535 34651  COGS2   25       CTRL          2010-07-07       M    AS
## 33536 34652  COGS2   43         SZ          2010-07-14       M    CA
## 33537 34653  COGS2   44       CTRL          2010-07-15       M    AA
## 33538 34654  COGS2   50         SZ          2010-07-16       M    CA
## 33539 34655  COGS2   49       CTRL          2010-07-19       M    CA
## 33540 34656  COGS2   43         SZ          2010-07-20       M    AA
##       cHispanicorLatino cLocationInstitution cLocationCity cLocationState
## 33535                No                 UCSD     San Diego             CA
## 33536               Yes                 UCSD     San Diego             CA
## 33537                No                 UCSD     San Diego             CA
## 33538                No                 UCSD     San Diego             CA
## 33539                No                 UCSD     San Diego             CA
## 33540                No                 UCSD     San Diego             CA
##       cLocationCounty cDiagnosis2 cDiagnosis3 cDiagnosis4 cEnrollmentYear
## 33535       San Diego          CS          CS          CS            2010
## 33536       San Diego          SZ      SZSAFD      SZSAFD            2010
## 33537       San Diego          CS          CS          CS            2010
## 33538       San Diego          SZ      SZSAFD      SZSAFD            2010
## 33539       San Diego          CS          CS          CS            2010
## 33540       San Diego          SZ      SZSAFD      SZSAFD            2010
##       cRace2
## 33535     AS
## 33536     CA
## 33537     AA
## 33538     CA
## 33539     CA
## 33540     AA
```

```
group_ct <- plyr::count(df_COGS2, c('cRace2', 'cGender', 'cHispanicorLatino'))
group_ct$prop <- group_ct$freq/nrow(df_COGS2)
prop_vec <- group_ct$prop
prop_vec
```

```
##  [1] 0.1142511102 0.0052482842 0.1978199435 0.0080742834 0.0004037142
##  [6] 0.0004037142 0.0028259992 0.0012111425 0.0234154219 0.0306822769
## [11] 0.0004037142 0.1699636657 0.0246265644 0.2587807832 0.0395639887
## [16] 0.0036334275 0.0004037142 0.0064594267 0.0004037142 0.0278562778
## [21] 0.0238191361 0.0335082761 0.0262414211
```

```
agg_m_ent <- mult_ent(prop_vec)
agg_m_ent
```

```
## [1] 2.19125
```

```
agg_df <- data.frame(City = 'COGS2 Aggregate', mult_ent = agg_m_ent)

# same thing but without SD
no_sd <- df_COGS2[df_COGS2$cLocationCity != 'San Diego', ]
group_ct <- plyr::count(no_sd, c('cRace2', 'cGender', 'cHispanicorLatino'))
group_ct$prop <- group_ct$freq/nrow(no_sd)
prop_vec <- group_ct$prop
prop_vec
```

```
##  [1] 0.1309836928 0.0057864282 0.2225144661 0.0105207785 0.0005260389
##  [6] 0.0021041557 0.0005260389 0.0215675960 0.0315623356 0.0005260389
## [11] 0.1657022620 0.0252498685 0.2467122567 0.0441872699 0.0031562336
## [16] 0.0005260389 0.0063124671 0.0005260389 0.0215675960 0.0105207785
## [21] 0.0352446081 0.0136770121
```

```
agg_m_ent_nosd <- mult_ent(prop_vec)
agg_m_ent_nosd
```

```
## [1] 2.1343
```

```
agg_df_nosd <- data.frame(City = 'COGS2 No SD', mult_ent = agg_m_ent_nosd)
```

By-City DI

```
cities <- unique(df_COGS2$cLocationCity)
n <- length(cities)
df_di <- data.frame(City = rep('',n),
                    mult_ent = rep(0,n)
                    )
head(df_di)
```

```
##   City mult_ent
## 1             0
## 2             0
## 3             0
## 4             0
## 5             0
```

```
for (i in 1:length(cities)) {
  df_sub <- df_COGS2[df_COGS2$cLocationCity == cities[i],]

  group_ct <- plyr::count(df_sub, c('cRace2', 'cGender', 'cHispanicorLatino'))
  group_ct$prop <- group_ct$freq/nrow(df_sub)

  prop_vec <- group_ct$prop
  m_ent <- mult_ent(prop_vec)

  df_di[i,1] <- cities[i]
  df_di[i,2] <- m_ent
}
head(df_di)
```

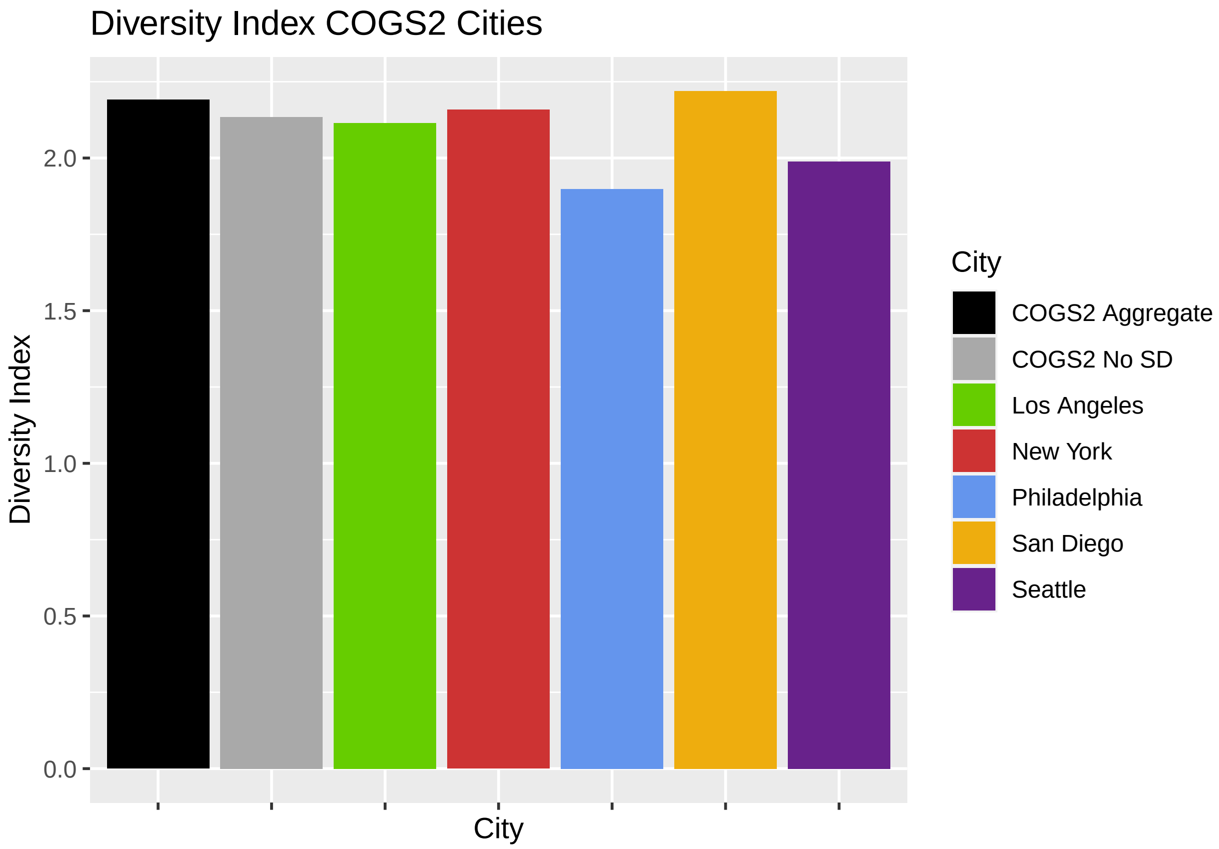
```
##          City mult_ent
## 1    San Diego 2.219907
## 2  Los Angeles 2.114581
## 3     New York 2.158019
## 4 Philadelphia 1.899062
## 5      Seattle 1.988777
```

```
df_di <- rbind(df_di,agg_df, agg_df_nosd)
df_di
```

```
##             City mult_ent
## 1       San Diego 2.219907
## 2     Los Angeles 2.114581
## 3        New York 2.158019
## 4    Philadelphia 1.899062
## 5         Seattle 1.988777
## 6 COGS2 Aggregate 2.191250
## 7     COGS2 No SD 2.134300
```

Barplot comparing results

```
my_cols <- c('black', 'darkgray', 'chartreuse3', 'brown3', 'cornflowerblue', 'darkgoldenrod2', 'darkorc

cogs2_bar <- ggplot(data = df_di, aes(x = City, y = mult_ent, fill = City)) + geom_bar(stat = 'identity
  scale_fill_manual(values=my_cols)
cogs2_bar
```

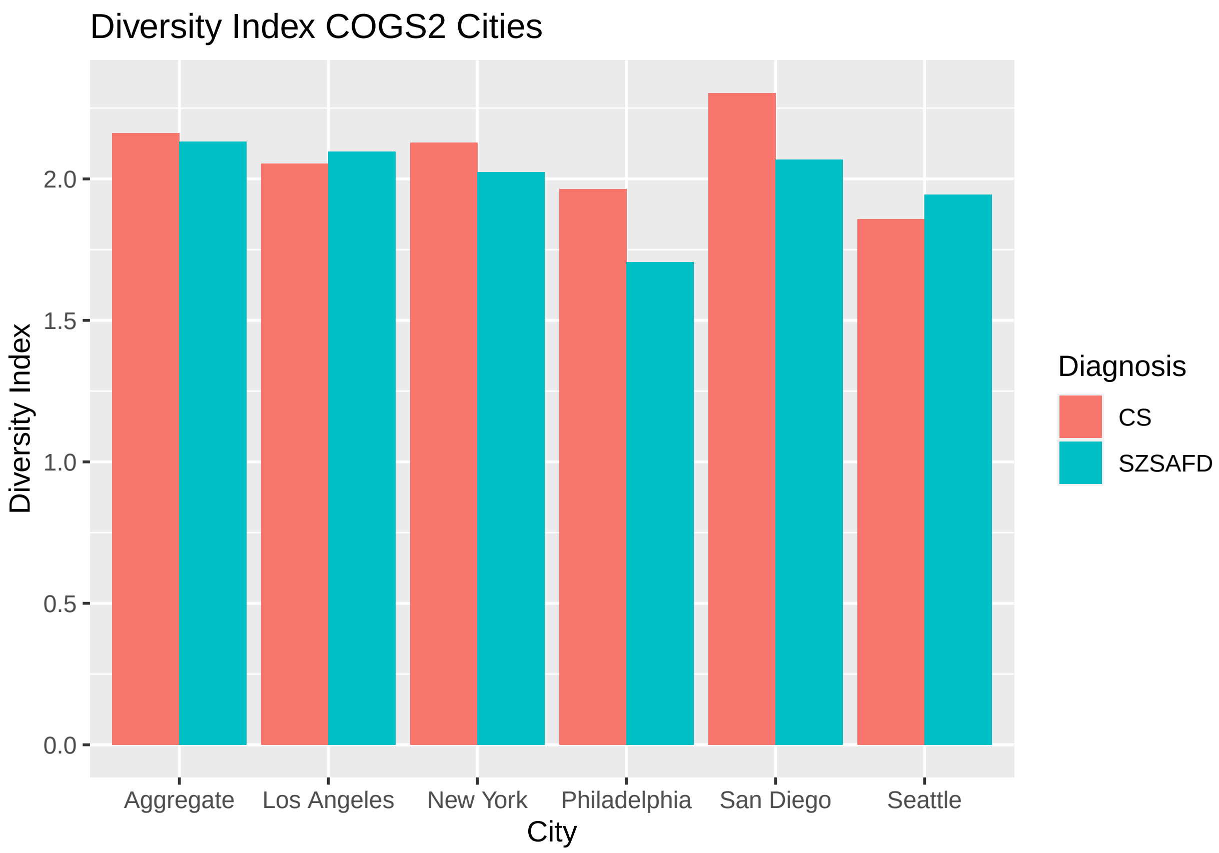


COGS2: CS and SZSAFD split

```
diagnoses <- c('CS', 'SZSAFD')
cities2 <- c(cities, 'Aggregate')
n <- 2 * length(cities2)

df_di2 <- data.frame(City = rep('',n),
                    Diagnosis = rep('',n),
                    mult_ent = rep(0,n)
                    )
n_track <- 1
for (d in 1:length(diagnoses)) {
    for (j in 1:length(cities2)) {
      if (cities2[j] != 'Aggregate') {
        df_sub <- df_COGS2[df_COGS2$cDiagnosis3 == diagnoses[d] & df_COGS2$cLocationCity == cities2[j],]
      } else {
        df_sub <- df_COGS2[df_COGS2$cDiagnosis3 == diagnoses[d],]
      }


      # if the dataframe is empty, skip this iteration
      if (nrow(df_sub) == 0) {
        df_di2[n_track,1] <- cities2[j]
        df_di2[n_track,2] <- diagnoses[d]
        df_di2[n_track,3] <- -99 # code for no data
        n_track <- n_track + 1
        next
      }

      group_ct <- plyr::count(df_sub, c('cRace2', 'cGender', 'cHispanicorLatino'))
      group_ct$prop <- group_ct$freq/nrow(df_sub)

      prop_vec <- group_ct$prop
      m_ent <- mult_ent(prop_vec)

      df_di2[n_track,1] <- cities2[j]
      df_di2[n_track,2] <- diagnoses[d]
      df_di2[n_track,3] <- m_ent
      n_track <- n_track+1
    }
}

#View(df_di2)
```

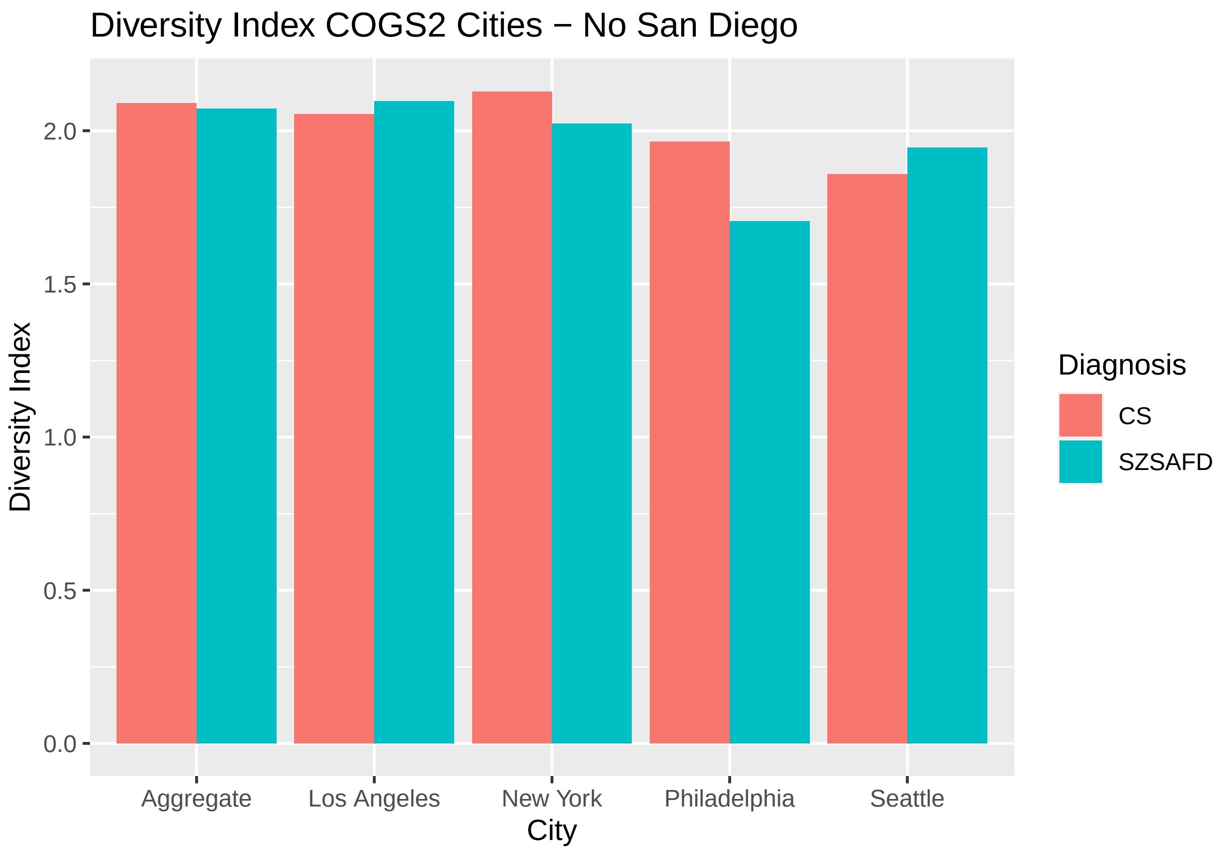
Barplot

```
bar2_COGS2 <- ggplot(data = df_di2, aes(x = City, y = mult_ent, fill = Diagnosis)) + geom_bar(stat = 'id
bar2_COGS2
```



Same thing but without San Diego

```
cities3 <- cities2[-1]

n <- 2 * length(cities3)

df_di3 <- data.frame(City = rep('',n),
                    Diagnosis = rep('',n),
                    mult_ent = rep(0,n)
                    )
n_track <- 1
for (d in 1:length(diagnoses)) {
    for (j in 1:length(cities3)) {
      if (cities3[j] != 'Aggregate') {
        df_sub <- no_sd[no_sd$cDiagnosis3 == diagnoses[d] & no_sd$cLocationCity == cities3[j],]
      } else {
        df_sub <- no_sd[no_sd$cDiagnosis3 == diagnoses[d],]
      }


      # if the dataframe is empty, skip this iteration
      if (nrow(df_sub) == 0) {
        df_di3[n_track,1] <- cities3[j]
        df_di3[n_track,2] <- diagnoses[d]
        df_di3[n_track,3] <- -99 # code for no data
        n_track <- n_track + 1
        next
      }

      group_ct <- plyr::count(df_sub, c('cRace2', 'cGender', 'cHispanicorLatino'))
      group_ct$prop <- group_ct$freq/nrow(df_sub)

      prop_vec <- group_ct$prop
      m_ent <- mult_ent(prop_vec)

      df_di3[n_track,1] <- cities3[j]
      df_di3[n_track,2] <- diagnoses[d]
      df_di3[n_track,3] <- m_ent
      n_track <- n_track+1
    }
}

bar3_COGS2 <- ggplot(data = df_di3, aes(x = City, y = mult_ent, fill = Diagnosis)) + geom_bar(stat = 'i
bar3_COGS2
```



```
plyr::count(df_COGS2, 'cLocationCity') # losing 576 samples by ignoring SD
```

```
##   cLocationCity freq
## 1   Los Angeles  481
## 2      New York  466
## 3  Philadelphia  480
## 4     San Diego  576
## 5       Seattle  474
```

ACS Analysis

Importing and checking data

```
df_acs <- read.csv('C:\\Users\\danie\\Documents\\Joshi Lab Materials\\acs_cogs2_1014.csv')
unique(df_acs$CITY) # all cities except SD
```

```
## [1] 3730 4610 5330 6430
```

```
head(df_acs)
```

```
##   YEAR SAMPLE SERIAL CBSERIAL HHWT      CLUSTER CITY CITYPOP STRATA GQ PERNUM
## 1 2010 201001  70099      255   64 2.010001e+12 3730   37971 541806  1      1
## 2 2010 201001  70111      385   53 2.010001e+12 3730   37971 542406  3      1
## 3 2010 201001  70116      449   82 2.010001e+12 3730   37971 541106  1      1
## 4 2010 201001  70116      449   82 2.010001e+12 3730   37971 541106  1      2
## 5 2010 201001  70116      449   82 2.010001e+12 3730   37971 541106  1      3
## 6 2010 201001  70122      550   67 2.010001e+12 3730   37971 541006  1      1
##   PERWT SEX AGE RACE RACED HISPAN HISPAND
## 1    64   2  71    2   200      0       0
## 2    53   1  58    1   100      0       0
## 3    82   2  38    1   100      0       0
## 4    82   1  36    1   100      0       0
## 5    91   2   3    1   100      0       0
## 6    68   2  53    1   100      0       0
```

```
# no missing data
sum(complete.cases(df_acs)) == nrow(df_acs)
```

```
## [1] TRUE
```

Encoding new race categories, binarizing hispan category

```
df_acs$Race2 <- rep(0, nrow(df_acs))
df_acs$Hispan2 <- rep(0, nrow(df_acs))

sum(df_acs$HISPAN == 9) # everyone reported a hispanic status
```

```
## [1] 0
```

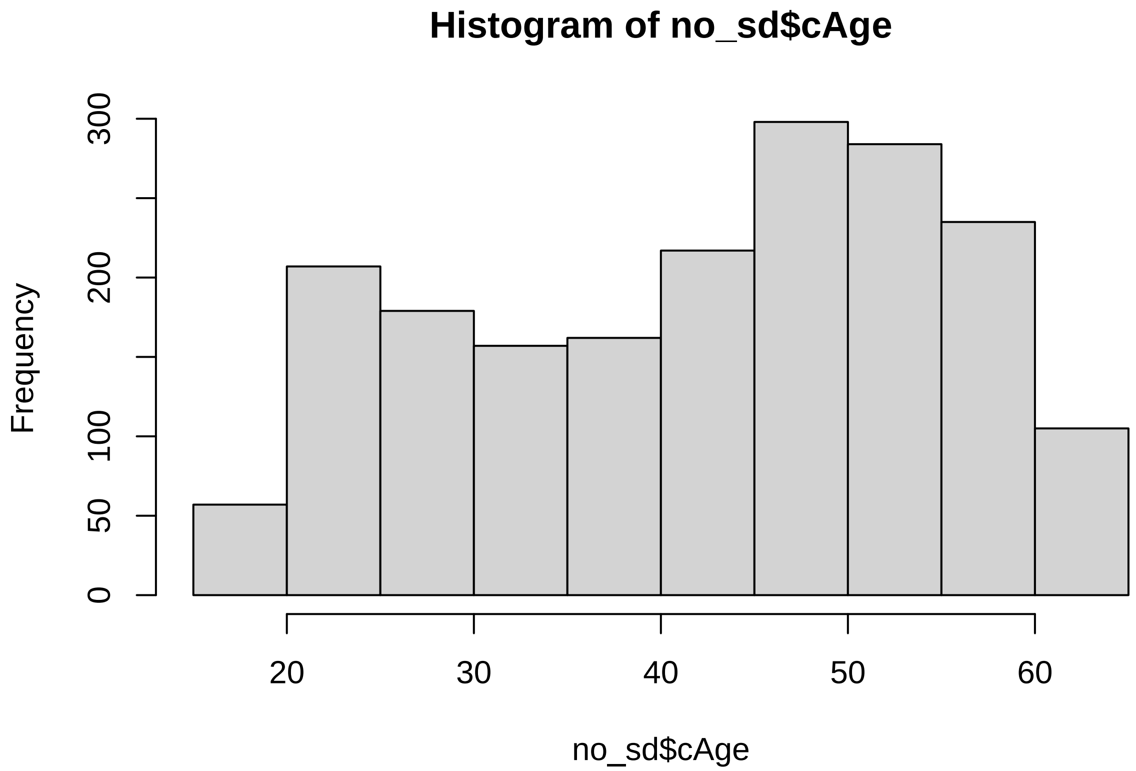
```
df_acs$Hispan2 <- as.numeric(df_acs$HISPAN != 0) # 0 for not hispanic or latino, else 1

PI_raced <- c(680:699) # PI races

df_acs$Race2[df_acs$RACE == 1 ] <- 1 # White
df_acs$Race2[df_acs$RACE == 2 ] <- 2 # Black
df_acs$Race2[df_acs$RACE == 3 ] <- 3 # American Indian or Alaska Native
df_acs$Race2[df_acs$RACE %in% 4:6 & !(df_acs$RACED %in% PI_raced) ] <- 4 # Asian
df_acs$Race2[df_acs$RACE == 6 & df_acs$RACED %in% PI_raced ] <- 5 # Pacific Islander (or Native Hawaiia
df_acs$Race2[df_acs$RACE %in% 7:9 ] <- 6 # Mixed/Other
```

Comparing ACS age ranges to COGS2 age ranges

```
# Getting the counts from no_sd
city_sam_sizes <- plyr::count(no_sd, 'cLocationCity')
city_sam_sizes
```

```
##   cLocationCity freq
## 1   Los Angeles  481
## 2      New York  466
## 3  Philadelphia  480
## 4       Seattle  474
```

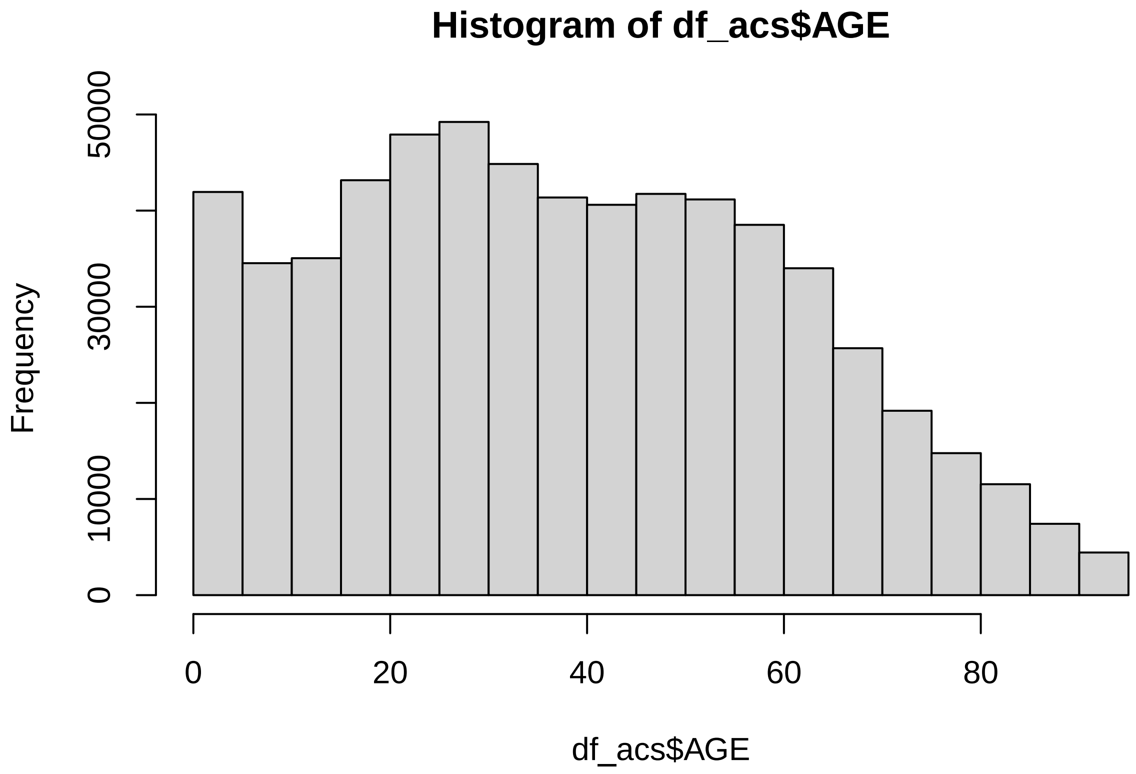
```
# age range of COGS2 was 18-65
hist(no_sd$cAge)
```



```
summary(no_sd$cAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   31.00   45.00   42.66   53.00   65.00 
```

```
# age range of the ipums sample: 0-95
hist(df_acs$AGE)
```



```
summary(df_acs$AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   20.00   37.00   38.54   56.00   95.00
```

```
# solution: truncate the acs dataframe by the age in COGS2
acs1865 <- df_acs[df_acs$AGE >= 18 & df_acs$AGE <= 65, ]
 ( nrow(df_acs) - nrow(acs1865) ) / nrow(df_acs) * 100 # lost 34% of the rows
```

```
## [1] 34.01336
```

Calculating diversity index for ACS sample

```
# effective total population
# note: due to truncation, city pop is no longer relevant

# empty dataframe to hold results
acs_di <- data.frame(CITY = rep('', 4),
                     DI = rep(0, 4))

cogs_cities_acs <- unique(acs1865$CITY)
cogs_cities_acs2 <- c(cogs_cities_acs, -1) # -1 is surrogate for aggregate
```

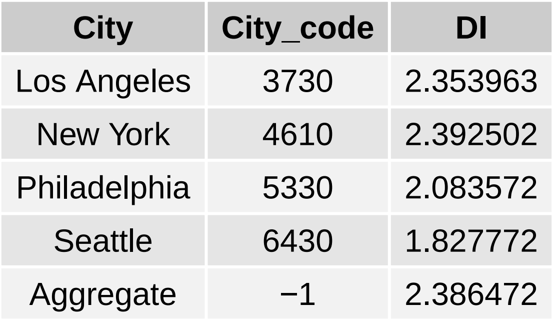
```
for (i in 1:length(cogs_cities_acs2)) {
  if (cogs_cities_acs2[i] != -1) {
    df_sub <- acs1865[acs1865$CITY == cogs_cities_acs2[i], ]
  } else {
    df_sub <- acs1865
  }

  # total effective population for that city
  # need to use this cuz subset by age
  tot <- sum(df_sub$PERWT)

  weighted_cts <- plyr::count(df_sub, c('Race2', 'Hispan2', 'SEX'), wt_var = 'PERWT')

  props <- weighted_cts$freq / tot
  acs_di[i,1] <-  cogs_cities_acs2[i]
  acs_di[i,2] <- mult_ent(props)
}

#View(acs_di)

fac_test <- factor(acs_di$CITY)
levels(fac_test) <- c('Aggregate', 'Los Angeles', 'New York', 'Philadelphia', 'Seattle')
fac_test
```

```
## [1] Los Angeles  New York     Philadelphia Seattle      Aggregate
## Levels: Aggregate Los Angeles New York Philadelphia Seattle
```

```
acs_di$City <- fac_test
acs_di <- acs_di[, c('City', 'CITY', 'DI')]
colnames(acs_di)[2] <- 'City_code'

knitr::kable(acs_di)
```

City	City_code	DI
Los Angeles	3730	2.353963
New York	4610	2.392502
Philadelphia	5330	2.083572
Seattle	6430	1.827772
Aggregate	-1	2.386472

```
grid.newpage()
grid.table(acs_di, rows = NULL)
```



City	City_code	DI
Los Angeles	3730	2.353963
New York	4610	2.392502
Philadelphia	5330	2.083572
Seattle	6430	1.827772
Aggregate	−1	2.386472

```
# showing populations
pop_counts <- plyr::count(acs1865, 'CITY', 'PERWT')
pop_counts$City <- c('Los Angeles', 'New York', 'Philadelphia', 'Seattle')
pop_counts[, c('City', 'CITY', 'freq')]
```

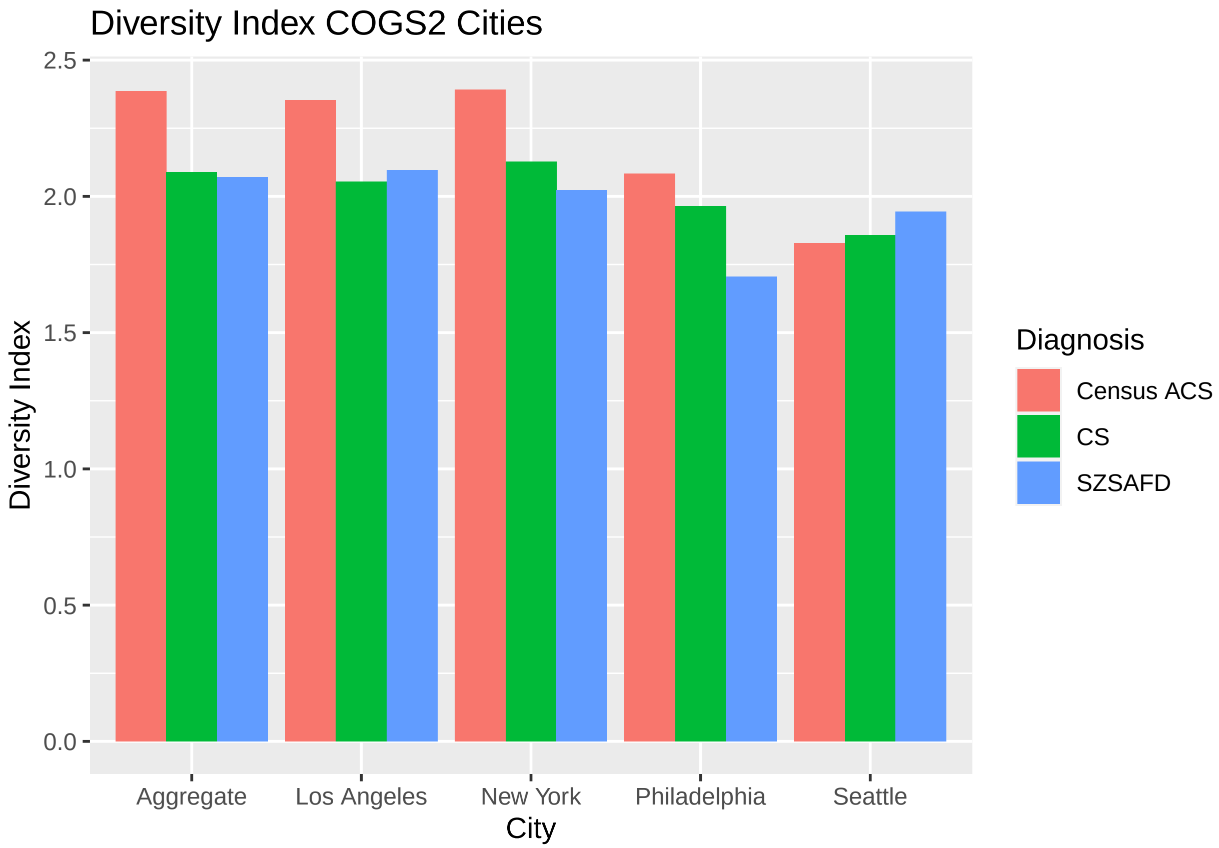
```
##           City CITY     freq
## 1  Los Angeles 3730 13007366
## 2     New York 4610 27946386
## 3 Philadelphia 5330  5126426
## 4      Seattle 6430  2367072
```

```
# LA and NY have highest counts
```

Analysis: Comparing ACS DI to study DI

```
# putting the dataframes on top of each other
cs_cogs2 <- df_di3[df_di3$Diagnosis == 'CS',]
szsafd_cogs2 <- df_di3[df_di3$Diagnosis == 'SZSAFD',]
acs_di$Diagnosis <- 'Census ACS'
colnames(cs_cogs2)[3] = colnames(szsafd_cogs2)[3] = 'DI'
di_collection <- rbind(acs_di[, c(1, 4, 3)], cs_cogs2, szsafd_cogs2)
di_collection <- di_collection[order(di_collection$City), ]
```

```
#View(di_collection)

# barplot of the dataframe
di_bars <- ggplot(data = di_collection, aes(x = City, y = DI, fill = Diagnosis)) + geom_bar(stat = 'ider

di_bars
```



Hypothesis Testing: Monte-Carlo Simulation

Getting sample sizes

```
# by-city and diagnosis sample sizes
city_diag_sam_sizes <- plyr::count(no_sd, c('cLocationCity', 'cDiagnosis4'))
city_diag_sam_sizes
```

```
##   cLocationCity cDiagnosis4 freq
## 1   Los Angeles          CS  217
## 2   Los Angeles      SZSAFD  264
## 3      New York          CS  196
## 4      New York      SZSAFD  270
## 5  Philadelphia          CS  207
## 6  Philadelphia      SZSAFD  273
## 7       Seattle          CS  221
## 8       Seattle      SZSAFD  253
```

```
# by-diagnosis sample size (aggregate)
city_diag_agg_sam_size <-  plyr::count(no_sd, 'cDiagnosis4')
city_diag_agg_sam_size
```

```
##   cDiagnosis4 freq
## 1          CS  841
## 2      SZSAFD 1060
```

Simulation Methodology

Methodology: 1) Follow same for loop structure as in generation of diversity index 2) After generating the proportions vector for use in the DI calculation: - Randomly sample n from the rows of the weighted_cts data frame - Choose n based on what COGS2's counts for CS or SZSAFD within the city of interest - Then re-generate diversity index 3) Store results in a matrix - 1 row per city per diagnosis - 10 rows total (include the aggregate) - 1000 columns

Generating the Results Matrix

```
city_diag_sam_sizes
```

```
##   cLocationCity cDiagnosis4 freq
## 1   Los Angeles          CS  217
## 2   Los Angeles      SZSAFD  264
## 3      New York          CS  196
## 4      New York      SZSAFD  270
## 5  Philadelphia          CS  207
## 6  Philadelphia      SZSAFD  273
## 7       Seattle          CS  221
## 8       Seattle      SZSAFD  253
```

```
city_diag_agg_sam_size
```

```
##   cDiagnosis4 freq
## 1          CS  841
## 2      SZSAFD 1060
```

```
agg_info <- data.frame(cLocationCity = 'Aggregate',
                       cDiagnosis4 = c('CS', 'SZSAFD'),
                       freq = city_diag_agg_sam_size$freq)
city_ns <- rbind(city_diag_sam_sizes, agg_info)
city_ns
```

```
##    cLocationCity cDiagnosis4 freq
## 1    Los Angeles          CS  217
## 2    Los Angeles      SZSAFD  264
## 3       New York          CS  196
## 4       New York      SZSAFD  270
## 5   Philadelphia          CS  207
## 6   Philadelphia      SZSAFD  273
## 7        Seattle          CS  221
## 8        Seattle      SZSAFD  253
## 9      Aggregate          CS  841
## 10     Aggregate      SZSAFD 1060
```

```
city_ns
```

```
##    cLocationCity cDiagnosis4 freq
## 1    Los Angeles          CS  217
## 2    Los Angeles      SZSAFD  264
## 3       New York          CS  196
## 4       New York      SZSAFD  270
## 5   Philadelphia          CS  207
## 6   Philadelphia      SZSAFD  273
## 7        Seattle          CS  221
## 8        Seattle      SZSAFD  253
## 9      Aggregate          CS  841
## 10     Aggregate      SZSAFD 1060
```

```
N_sim <- 1000
sim_mat <- matrix(0, nrow = 10, ncol = N_sim)

# Same for-loop structure as before
n_track <- 1

for (d in 1:length(diagnoses)) {
  for (i in 1:length(cogs_cities_acs2)) {

    if (cogs_cities_acs2[i] != -1) {
      df_sub <- acs1865[acs1865$CITY == cogs_cities_acs2[i], ]
    } else {
      df_sub <- acs1865
    }

    # total effective population for that city
    # need to use this cuz subset by age
    tot <- sum(df_sub$PERWT)

    weighted_cts <- plyr::count(df_sub, c('Race2', 'Hispan2', 'SEX'), wt_var = 'PERWT')

    props <- weighted_cts$freq / tot

    # begin random sampling

    for (N in 1:N_sim) {
      row_samples <- sample(1:nrow(weighted_cts), size = city_ns$freq[i], p = props, replace = TRUE)

      # generate proportions for the 24 groups
      row_sam_cts <- plyr::count(row_samples)
      prop_for_DI <- row_sam_cts$freq/sum(row_sam_cts$freq)
      # generate diversity index
      DI_sam <- mult_ent(prop_for_DI)

      # store in matrix
      sim_mat[n_track, N] <- DI_sam
    }
    n_track <- n_track + 1
  }
}
```

Calculating the 95% CI

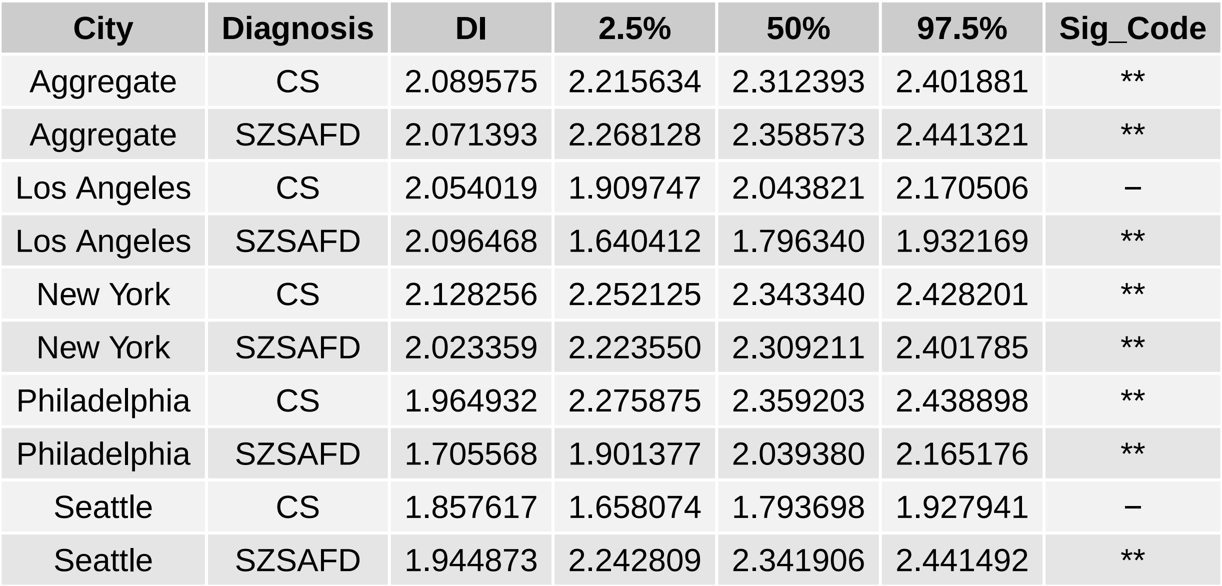
```
CI_df <- as.data.frame(t(apply(sim_mat, MARGIN = 1, FUN = quantile, prob = c(0.025, 0.50, 0.975), simpli
CI_df
```

```
##        2.5%      50%    97.5%
## 1  2.215634 2.312393 2.401881
## 2  2.268128 2.358573 2.441321
## 3  1.909747 2.043821 2.170506
## 4  1.640412 1.796340 1.932169
## 5  2.252125 2.343340 2.428201
## 6  2.223550 2.309211 2.401785
## 7  2.275875 2.359203 2.438898
## 8  1.901377 2.039380 2.165176
## 9  1.658074 1.793698 1.927941
## 10 2.242809 2.341906 2.441492
```

```
di_cogs2 <- di_collection[di_collection$Diagnosis!='Census ACS',]
rownames(di_cogs2) <- NULL
di_sig <- cbind(di_cogs2, CI_df)
di_sig$Significant <- di_sig$DI < di_sig$`2.5%` | di_sig$DI > di_sig$`97.5%`
di_sig$Sig_Code <- ifelse(di_sig$Significant, '**', '-')
#View(di_sig)

knitr::kable(di_sig)
```

City	Diagnosis	DI	2.5%	50%	97.5%	Significant	Sig_Code
Aggregate	CS	2.089575	2.215635	2.312393	2.401881	TRUE	**
Aggregate	SZSAFD	2.071393	2.268128	2.358573	2.441320	TRUE	**
Los Angeles	CS	2.054019	1.909747	2.043821	2.170506	FALSE	-
Los Angeles	SZSAFD	2.096468	1.640412	1.796340	1.932169	TRUE	**
New York	CS	2.128256	2.252125	2.343340	2.428201	TRUE	**
New York	SZSAFD	2.023359	2.223550	2.309211	2.401785	TRUE	**
Philadelphia	CS	1.964932	2.275875	2.359203	2.438898	TRUE	**
Philadelphia	SZSAFD	1.705568	1.901377	2.039380	2.165176	TRUE	**
Seattle	CS	1.857617	1.658074	1.793698	1.927941	FALSE	-
Seattle	SZSAFD	1.944873	2.242809	2.341906	2.441492	TRUE	**

```
grid.newpage()
grid.table(di_sig[, -7], rows = NULL)
```



City	Diagnosis	DI	2.5%	50%	97.5%	Sig_Code
Aggregate	CS	2.089575	2.215634	2.312393	2.401881	**
Aggregate	SZSAFD	2.071393	2.268128	2.358573	2.441321	**
Los Angeles	CS	2.054019	1.909747	2.043821	2.170506	–
Los Angeles	SZSAFD	2.096468	1.640412	1.796340	1.932169	**
New York	CS	2.128256	2.252125	2.343340	2.428201	**
New York	SZSAFD	2.023359	2.223550	2.309211	2.401785	**
Philadelphia	CS	1.964932	2.275875	2.359203	2.438898	**
Philadelphia	SZSAFD	1.705568	1.901377	2.039380	2.165176	**
Seattle	CS	1.857617	1.658074	1.793698	1.927941	–
Seattle	SZSAFD	1.944873	2.242809	2.341906	2.441492	**

Barplot of results

```
library(ggsignif)
```

```
## Warning: package 'ggsignif' was built under R version 4.3.1
```

```
census_medians <- di_sig[, c('City', 'Diagnosis', '50%')]
census_medians$Diagnosis <- paste(census_medians$Diagnosis, 'ACS')
census_medians
```

```
##             City  Diagnosis      50%
## 1      Aggregate     CS ACS 2.312393
## 2      Aggregate SZSAFD ACS 2.358573
## 3    Los Angeles     CS ACS 2.043821
## 4    Los Angeles SZSAFD ACS 1.796340
## 5       New York     CS ACS 2.343340
## 6       New York SZSAFD ACS 2.309211
## 7   Philadelphia     CS ACS 2.359203
## 8   Philadelphia SZSAFD ACS 2.039380
## 9       Seattle     CS ACS 1.793698
## 10      Seattle SZSAFD ACS 2.341906
```

```
head(census_medians)
```

```
##          City  Diagnosis      50%
## 1   Aggregate     CS ACS 2.312393
## 2   Aggregate SZSAFD ACS 2.358573
## 3 Los Angeles     CS ACS 2.043821
## 4 Los Angeles SZSAFD ACS 1.796340
## 5    New York     CS ACS 2.343340
## 6    New York SZSAFD ACS 2.309211
```

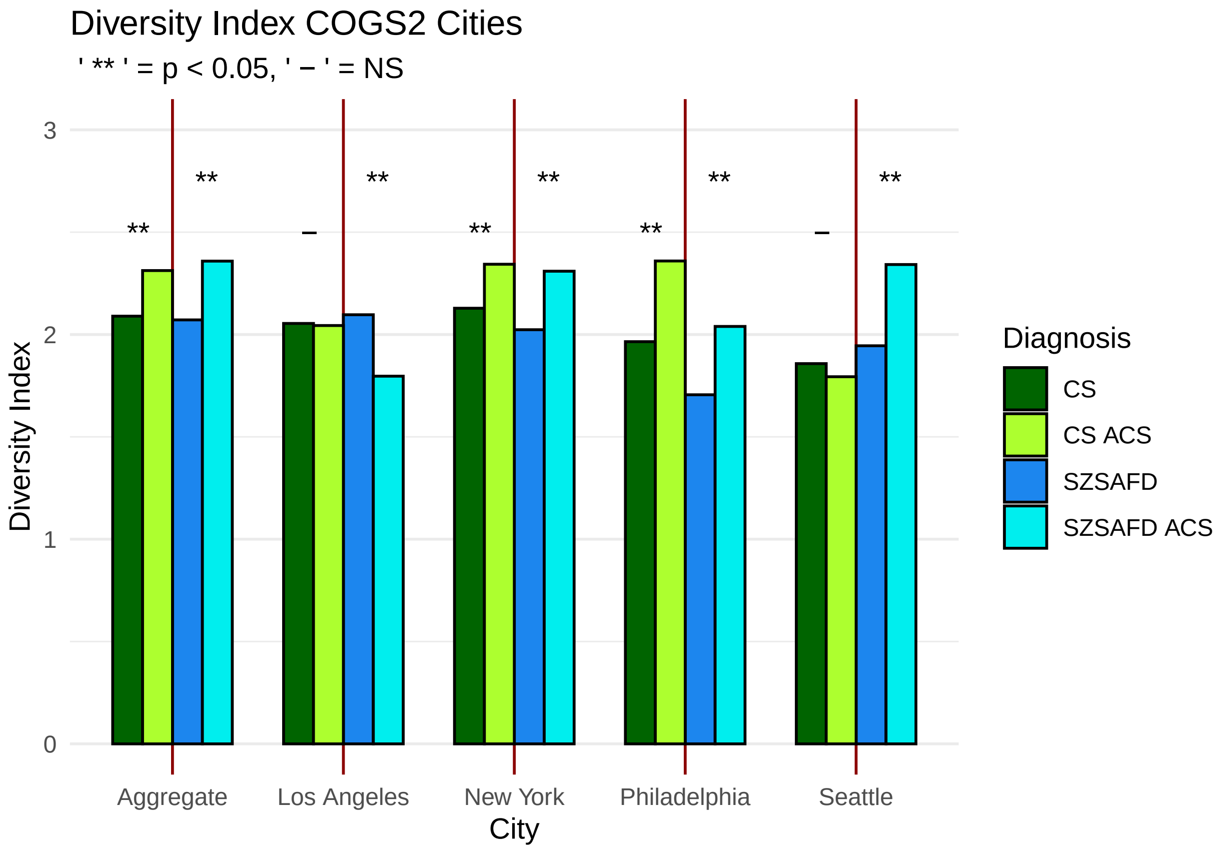
```
colnames(census_medians)[3] <- 'DI'

head(di_sig)
```

```
##          City Diagnosis       DI     2.5%      50%    97.5% Significant
## 1   Aggregate        CS 2.089575 2.215634 2.312393 2.401881        TRUE
## 2   Aggregate    SZSAFD 2.071393 2.268128 2.358573 2.441321        TRUE
## 3 Los Angeles        CS 2.054019 1.909747 2.043821 2.170506       FALSE
## 4 Los Angeles    SZSAFD 2.096468 1.640412 1.796340 1.932169        TRUE
## 5    New York        CS 2.128256 2.252125 2.343340 2.428201        TRUE
## 6    New York    SZSAFD 2.023359 2.223550 2.309211 2.401785        TRUE
##   Sig_Code
## 1       **
## 2       **
## 3        -
## 4       **
## 5       **
## 6       **
```

```
di_sig_plot <- rbind(di_sig[, 1:3], census_medians)
di_sig_plot
```

```
##            City  Diagnosis       DI
## 1     Aggregate         CS 2.089575
## 2     Aggregate     SZSAFD 2.071393
## 3   Los Angeles         CS 2.054019
## 4   Los Angeles     SZSAFD 2.096468
## 5      New York         CS 2.128256
## 6      New York     SZSAFD 2.023359
## 7  Philadelphia         CS 1.964932
## 8  Philadelphia     SZSAFD 1.705568
## 9       Seattle         CS 1.857617
## 10      Seattle     SZSAFD 1.944873
## 11    Aggregate     CS ACS 2.312393
## 12    Aggregate SZSAFD ACS 2.358573
## 13  Los Angeles     CS ACS 2.043821
## 14  Los Angeles SZSAFD ACS 1.796340
## 15     New York     CS ACS 2.343340
## 16     New York SZSAFD ACS 2.309211
## 17 Philadelphia     CS ACS 2.359203
## 18 Philadelphia SZSAFD ACS 2.039380
## 19      Seattle     CS ACS 1.793698
## 20      Seattle SZSAFD ACS 2.341906
```

```
hyp_test <- ggplot(data = di_sig_plot, aes(City, DI)) + geom_bar(aes(fill = Diagnosis), width = 0.7, sta
  ylim(0, 3.0) +
  ylab('Diversity Index') +
  ggtitle('Diversity Index COGS2 Cities') +
  theme_minimal() + theme(panel.grid.major.x = element_line(color = 'darkred'))
  #                         panel.grid.minor.x = element_line(color = 'grey68'))

cslabel.df <- data.frame(City = 0.8 + 0:4,
                         DI = rep(2.5, 5))

szsafdlabel.df <- data.frame(City = 1.2 + 0:4,
                             DI = rep(2.75, 5))

my_colors <- c('darkgreen', 'greenyellow', 'dodgerblue2', 'cyan2')

# hyp_test + scale_fill_brewer(palette="Blues") + geom_text(data = cslabel.df, label = c('**', '-', '**
#   geom_text(data= szsafdlabel.df, label = c('**', '**', '**', '**', '**'))
hyp_test + scale_fill_manual(values = my_colors) + geom_text(data = cslabel.df, label = c('**', '-', '*
  geom_text(data= szsafdlabel.df, label = c('**', '**', '**', '**', '**')) + labs(subtitle = " ' ** ' =
```



```
dev.off()
```

```
## pdf
##   3
```