# GrizzHacks 2020: Data Cleaning

Dillon Jaghory

9/18/2020

## Cleaning the Data

This R markdown file will document the process of cleaning the transportation/mobility, demographic and unemployment data for our GrizzHacks 5 project.

## Transportation and Mobility

Data Source: https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

```
travel <- read.csv("Trips_by_Distance.csv")

mi_travel <- subset(travel, State.FIPS==26)

mi_travel$Date <- as.Date(mi_travel$Date)

rm(travel)

mi_travel <- subset(mi_travel, Date >= "2020-03-01")

mi_travel$County.Name <- gsub(" County", "", mi_travel$County.Name)

write.csv(mi_travel,"travelData.csv")
```

## Census Demographics

Data Source: https://data.census.gov/cedsci/?q=United%20States

```
#YEAR 12 = 7/1/2019 population estimate
#AGEGRP 0 = Total
#AGEGRP 17 = Age 80 to 84 years
#AGEGRP 18 = Age 85 years or older
#TOT_POP, TOT_MALE/FEMALE = total
#IA = Native American
#WA = White
#AA = Asian
#BA = Black
#NA = Native Hawaiian or Pacific Islander
#TOM = Two or More
#H = Hispanic

demographics <- read.csv("cc-est2019-alldata-26.csv")

demographics <- subset(demographics, select = c(COUNTY, CTYNAME, YEAR, AGEGRP,
                                                TOT_POP, TOT_MALE, TOT_FEMALE,
```

```
                                                WA_MALE, WA_FEMALE,
                                                BA_MALE, BA_FEMALE,
                                                AA_MALE, AA_FEMALE,
                                                IA_MALE, IA_FEMALE,
                                                NA_MALE, NA_FEMALE,
                                                H_MALE, H_FEMALE,
                                                TOM_MALE, TOM_FEMALE) )

demographics <- subset(demographics, AGEGRP==0 | AGEGRP==17 | AGEGRP==18)
demographics <- subset(demographics, YEAR==12)

demographics$TOT_WA <- demographics$WA_MALE + demographics$WA_FEMALE
demographics$TOT_BA <- demographics$BA_MALE + demographics$BA_FEMALE
demographics$TOT_AA <- demographics$AA_MALE + demographics$AA_FEMALE
demographics$TOT_IA <- demographics$IA_MALE + demographics$IA_FEMALE
demographics$TOT_NA <- demographics$NA_MALE + demographics$NA_FEMALE
demographics$TOT_H <- demographics$H_MALE + demographics$H_FEMALE
demographics$TOT_TOM <- demographics$TOM_MALE + demographics$TOM_FEMALE

demographics <- subset(demographics, select = -c(YEAR))

demographics$CTYNAME <- gsub(" County", "", demographics$CTYNAME)

write.csv(demographics,"demographics.csv")
```

## Unemployment

Data Source: https://www.bls.gov/lau/home.htm

```
url <- "https://download.bls.gov/pub/time.series/la/la.area"

county_codes <- download.file(url, destfile = "counties.txt")

counties <- read.table("counties.txt",
               sep="\t",
               col.names=c("area_type_code", "area_code",
                           "area_text", "display_level",
                           "selectable", "sort_sequence"),
               skip = 1,
               fill=FALSE,
               strip.white=TRUE)
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
## EOF within quoted string
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
## number of items read is not a multiple of the number of columns
```

```
unemployment <- read.csv("mi_unemployment.csv")
unemployment$Series.ID <- gsub("LAU", "", unemployment$Series.ID)
unemployment$Series.ID <- gsub("03$", "", unemployment$Series.ID)


unemployment$Series.ID <- counties$area_text[match(unemployment$Series.ID, counties$area_code)]
```

```r
unemployment <- subset(unemployment, select = -c(Year, Label) )

unemployment$Series.ID <- gsub(" County, MI", "", unemployment$Series.ID)

unemployment$Period <- gsub("M0", "", unemployment$Period)

colnames(unemployment) <- c("county", "month", "rate")

write.csv(unemployment,"unemployment.csv")
```