|  |  |  |
| --- | --- | --- |
| Delete | Maybe Keep | Keep |
| * RT * SERIALNO * DIVISION * REGION * STATE * ADJINC * GCM: Length of time responsible for grandchildren: 4292 missing values * GCR: Grandparents responsible for grandchildren: 4292 missing values * Decade: 3439 missing values * DRIVESP: Number of vehicles calculated from JWRI: 2549 missing values * ESP: Employment status parents: 4177 missing values * FER: Gave birth to child within the past 12 months: 3179 missing values * FOD1P: 2349 missing values * FOD2P: 4058 missing values * INDP: Industry recode for 2023 and later based on 2022 IND codes: 1077 missing values * JWAP: Time of arrival at work - hour and minute: 2188 missing values * JWDP: Time of departure 2188 * JWMNP: Travel time to work: missing 2188 * JWRIP: Vehicle occupancy: 2549 * JWTRNS: Means of transportation to work: 1729 * LANP: Language spoken at home: 3381 * MIGPUMA: 3803 * MIGSP:3803 | * SPORDER: Person number * PUMA: Public use microdata area code * CIT: Citizenship status * CITWP: Year of naturalization (lots of missing values * COW: Class of worker (lots of missing values 1077) * ENG: Ability to Speak English: 3381 missing values * GCL: Grandparents living with grandchildren * HIMRKS: Subsidized Marketplace Coverage * INTP: Interest, dividends, and net rental income past 12 months (use ADJINC to adjust to constant dollars) * JWMNP: Travel time to work * ESR: Employment status recode: only 58 missing rows * MARHD: Divorced in the past 12 months: 1542 missing * MARHM: Married in the past 12 months:1542 missing * MARHT: 1542 * MARHW: 1542 * MARHYP: 1542 | * PWGTP: Person's weight * HINS1: Insurance through a current or former employer or union * HINS2: Insurance purchased directly from an insurance company * HINS3: Medicare, for people 65 and older, or people with certain disabilities * HINS4: Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability * HINS5: TRICARE or other military health care * HINS6: TRICARE or other military health care * HINS7: Indian Health Service * LANX: Language other than English spoken at home * MAR: Marital Status * MARHT: Number of times married * MARHW: Widowed in the past 12 months * MARHW: Year last married * MIG: Mobility status * MIL: Military Status |

Record the methods used:

* Unsupervised methods
  + Near zero variance
  + Find correlation
  + 50 percent n/a
  + Column fill in categorical columns
  + Numerical find the median or the average
  + Outliers: isolation forest to review and remove outliers
  + Feature selection:
    - Correlation feature selection
    - Boruta

Tasks:

* Create a github project **done**
* Upload data **done**
* Write summary of project **done**
* Remove irrelevant attributes (e.g., tuple id, sample date) **In progress**
* Review columns that have more than 50% n/a
* Remove zero-variance or near zero-variance attributes
* Remove duplicate attributes.
* Packages in R for feature selection: Fselector, Caret, Rweka, Boruta, . . .
* Column fill in categorical columns
  + Look for imputation algorithms

# Get content into a data frame

df <- read.csv("project\_data.csv", header=TRUE, sep = ",")

# Printing content of Text File

#head(df, 5)

# find location of missing values

print("Position of missing values ")

#which(is.na(df))

# count total missing values

print("Count of total missing values ")

sum(is.na(df))

#Sum of the missing values in each column

sort(colSums(is.na(df)), decreasing=TRUE)

#Learning the summary of some of the important variables

"""

# RT is only P

"""