Profile DSS Customers By Year

library(inspectdf)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

library(here)

## here() starts at /home/ads7fg/git/vecf\_counts

library(maditr)

##   
## Use magrittr pipe '%>%' to chain several operations:  
## mtcars %>%  
## let(mpg\_hp = mpg/hp) %>%  
## take(mean(mpg\_hp), by = am)  
##

##   
## Attaching package: 'maditr'

## The following objects are masked from 'package:data.table':  
##   
## dcast, melt

## The following objects are masked from 'package:dplyr':  
##   
## between, coalesce, first, last

library(dataplumbr)

## Warning: replacing previous import 'data.table::month' by  
## 'lubridate::month' when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::hour' by 'lubridate::hour'  
## when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::quarter' by  
## 'lubridate::quarter' when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::week' by 'lubridate::week'  
## when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::year' by 'lubridate::year'  
## when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::wday' by 'lubridate::wday'  
## when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::second' by  
## 'lubridate::second' when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::minute' by  
## 'lubridate::minute' when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::mday' by 'lubridate::mday'  
## when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::yday' by 'lubridate::yday'  
## when loading 'dataplumbr'

## Warning: replacing previous import 'data.table::isoweek' by  
## 'lubridate::isoweek' when loading 'dataplumbr'

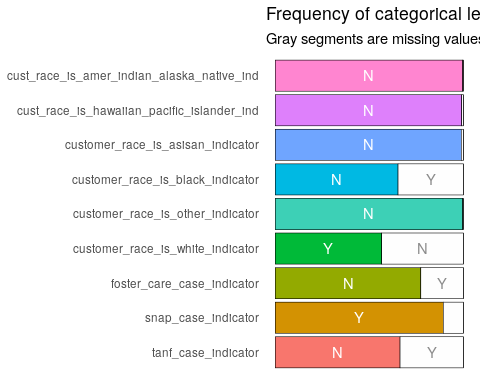
library(ggplot2)

dt <- fread(here("data/working/DSS/dss\_customers\_by\_year.csv"))

## Value Frequencies

### Categorical variables

cat\_chart <-   
 dt %>%   
 dt\_select(-1) %>%   
 inspect\_cat() %>%   
 show\_plot(high\_cardinality = 1)



## Completeness

# Number of cell values missing per row  
row\_empties <- rowSums(var.is\_blank(dt))  
  
# Create better visualization  
row\_empties\_dt <- as.data.table(row\_empties)  
records\_with\_empties <- row\_empties\_dt[row\_empties > 0, .N]  
records\_with\_empties

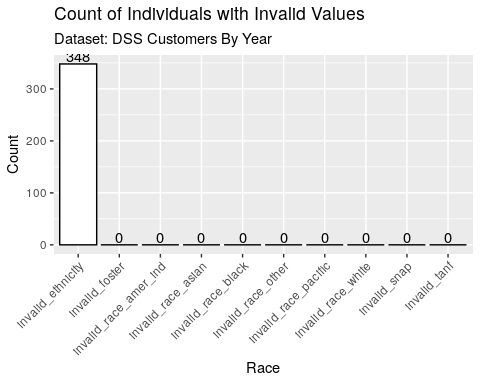
## [1] 0

# Number of cell values missing per column  
col\_empties <- colSums(var.is\_blank(dt))  
  
# Create better visualization  
  
col\_empties\_dt <- as.data.table(col\_empties, keep.rownames = T)  
#dtt <- setDT(df, keep.rownames=TRUE)  
colnames(col\_empties\_dt) <- c("item", "empties")  
item\_empties <- col\_empties\_dt[order(-empties)]  
item\_empties

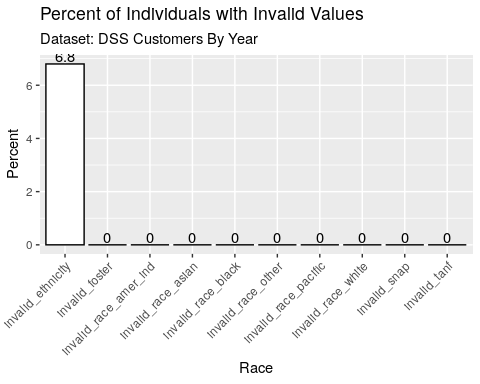
## item empties  
## 1: unique\_id 0  
## 2: age\_class\_code 0  
## 3: age\_group\_code 0  
## 4: age\_type\_code 0  
## 5: calendar\_year\_number 0  
## 6: cust\_race\_is\_amer\_indian\_alaska\_native\_ind 0  
## 7: cust\_race\_is\_hawaiian\_pacific\_islander\_ind 0  
## 8: customer\_race\_is\_asisan\_indicator 0  
## 9: customer\_race\_is\_black\_indicator 0  
## 10: customer\_race\_is\_other\_indicator 0  
## 11: customer\_race\_is\_white\_indicator 0  
## 12: ethnicity\_code 0  
## 13: foster\_care\_case\_indicator 0  
## 14: gender\_code 0  
## 15: month\_of\_birth 0  
## 16: snap\_case\_indicator 0  
## 17: tanf\_case\_indicator 0  
## 18: year\_of\_birth 0

## Valid Values

#. gender ----  
invalid\_gend <- c("1", "2")  
gender\_invalid <- dt[gender\_code %in% invalid\_gend, .N]  
  
# ethnicity  
invalid\_eth <- c("0", "3", "90", "93")  
invalid\_ethnicity <- data.table(invalid\_ethnicity = dt[ethnicity\_code %in% invalid\_eth, .N])  
  
# race  
invalid\_race\_amer\_ind <- data.table(invalid\_race\_amer\_ind = dt[!cust\_race\_is\_amer\_indian\_alaska\_native\_ind %in% c("Y", "N"), .N])  
invalid\_race\_asian <- data.table(invalid\_race\_asian = dt[!customer\_race\_is\_asisan\_indicator %in% c("Y", "N"), .N])  
invalid\_race\_black <- data.table(invalid\_race\_black = dt[!customer\_race\_is\_black\_indicator %in% c("Y", "N"), .N])  
invalid\_race\_pacific <- data.table(invalid\_race\_pacific = dt[!cust\_race\_is\_hawaiian\_pacific\_islander\_ind %in% c("Y", "N"), .N])  
invalid\_race\_other <- data.table(invalid\_race\_other = dt[!customer\_race\_is\_other\_indicator %in% c("Y", "N"), .N])  
invalid\_race\_white <- data.table(invalid\_race\_white = dt[!customer\_race\_is\_white\_indicator %in% c("Y", "N"), .N])  
  
# year  
invalid\_year <- data.table(invalid\_year = dt[!as.integer(calendar\_year\_number) %in% 2013:2016, .N])  
  
# services  
invalid\_foster <- data.table(invalid\_foster = dt[!foster\_care\_case\_indicator %in% c("Y", "N"), .N])  
invalid\_snap <- data.table(invalid\_snap = dt[!snap\_case\_indicator %in% c("Y", "N"), .N])  
invalid\_tanf <- data.table(invalid\_tanf = dt[!tanf\_case\_indicator %in% c("Y", "N"), .N])  
  
  
vv\_multi <- data.table(invalid\_ethnicity,   
 invalid\_race\_amer\_ind,   
 invalid\_race\_asian,   
 invalid\_race\_black,   
 invalid\_race\_pacific,   
 invalid\_race\_other,   
 invalid\_race\_white,  
 invalid\_foster,  
 invalid\_snap,  
 invalid\_tanf)  
vv\_multi\_t <- as.data.table(t(vv\_multi), keep.rownames = T)  
  
colnames(vv\_multi\_t) <- c("item", "count")  
vv\_multi\_t[, pct := round(100\*(count/nrow(dt)), 1)]  
  
g <- ggplot(vv\_multi\_t, aes(x=item, y=count)) +  
 geom\_bar(stat="identity", colour="black", fill="white") +   
 xlab("Race") + ylab("Count") +  
 labs(title = "Count of Individuals with Invalid Values",  
 subtitle = "Dataset: DSS Customers By Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 geom\_text(aes(label=count), position=position\_dodge(width=0.9), vjust=-0.2)  
g



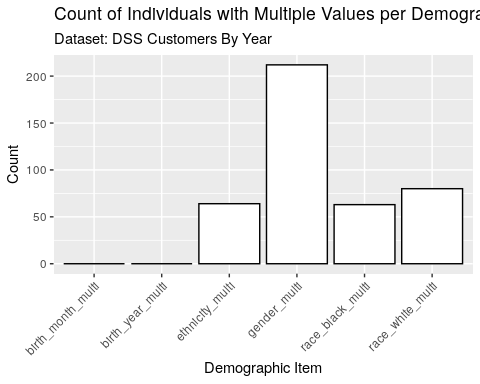
gp <- ggplot(vv\_multi\_t, aes(x=item, y=pct)) +  
 geom\_bar(stat="identity", colour="black", fill="white") +   
 xlab("Race") + ylab("Percent") +  
 labs(title = "Percent of Individuals with Invalid Values",  
 subtitle = "Dataset: DSS Customers By Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 geom\_text(aes(label=pct), position=position\_dodge(width=0.9), vjust=-0.2)  
gp



## Longitudinal Consistency

### Does any demographic have multiple values over time?

#final\_dt[, .(cnt = length(unique(`Year Built`))), .(`Parcel ID`)][order(-cnt)]  
  
gender\_multi <-  
 data.table(gender\_multi = dt[!is.na(gender\_code), .(cnt = length(unique(gender\_code)), dmg = "multi gender"), unique\_id][cnt > 1, .N])  
  
birth\_year\_multi <-  
 data.table(birth\_year\_multi = dt[!is.na(year\_of\_birth), .(cnt = length(unique(year\_of\_birth)), dmg = "multi birth yr"), unique\_id][cnt > 1, .N])  
  
birth\_month\_multi <-  
 data.table(birth\_month\_multi = dt[!is.na(month\_of\_birth), .(cnt = length(unique(month\_of\_birth)), dmg = "multi birth mo"), unique\_id][cnt > 1, .N])  
  
ethnicity\_multi <-  
 data.table(ethnicity\_multi = dt[!is.na(ethnicity\_code) & ethnicity\_code != 0, .(cnt = length(unique(ethnicity\_code)), dmg = "multi ethnicity"), unique\_id][cnt > 1, .N])  
  
race\_white\_multi <-  
 data.table(race\_white\_multi = dt[!is.na(customer\_race\_is\_white\_indicator) &  
 customer\_race\_is\_white\_indicator %in% c("Y", "N"), .(cnt = length(unique(customer\_race\_is\_white\_indicator)), dmg = "multi race white"), unique\_id  
 ][cnt > 1, .N])  
  
race\_black\_multi <-  
 data.table(race\_black\_multi = dt[!is.na(customer\_race\_is\_black\_indicator) &  
 customer\_race\_is\_black\_indicator %in% c("Y", "N"), .(cnt = length(unique(customer\_race\_is\_black\_indicator)), dmg = "multi race black"), unique\_id  
 ][cnt > 1, .N])  
  
dmg\_multi <- data.table(gender\_multi, birth\_year\_multi, birth\_month\_multi, ethnicity\_multi, race\_white\_multi, race\_black\_multi)  
dmg\_multi\_t <- as.data.table(t(dmg\_multi), keep.rownames = T)  
colnames(dmg\_multi\_t) <- c("item", "count")  
dmg\_multi\_t[, pct := round(100\*(count/nrow(dt)), 1)]  
  
g <- ggplot(dmg\_multi\_t, aes(x=item, y=count)) +  
 geom\_bar(stat="identity", colour="black", fill="white") +   
 xlab("Demographic Item") + ylab("Count") +  
 labs(title = "Count of Individuals with Multiple Values per Demographic Item",  
 subtitle = "Dataset: DSS Customers By Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
g



gp <- ggplot(dmg\_multi\_t, aes(x=item, y=pct)) +  
 geom\_bar(stat="identity", colour="black", fill="white") +   
 xlab("Demographic Item") + ylab("Percent") +  
 labs(title = "Percent of Individuals with Multiple Values per Demographic Item",  
 subtitle = "Dataset: DSS Customers By Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
gp

