Team C2

Lee

10/12/2019

Data Cleaning and EDA part

```
NumericDataRaw <- read_csv("Numeric.csv")
```

```
## Warning: Duplicated column names deduplicated: 'Q79' => 'Q79_1' [66], 'Q80'
## => 'Q80_1' [68], 'Q103' => 'Q103_1' [88]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
TextDataRaw <- read_csv("ChoiceText.csv")
```

```
## Warning: Duplicated column names deduplicated: 'Q79' => 'Q79_1' [66], 'Q80'
## => 'Q80_1' [68], 'Q103' => 'Q103_1' [88]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character()
## )
## See spec(...) for full column specifications.
```

Import Client specified IDs

```
ExcludedIDs = readxl::read_xlsx(path = "Need to be excluded IDs.xlsx",
  sheet = 3,col_names = TRUE,col_types = "text")%>%select(1)%>%pull()
```

```
## New names:
## * `` -> ...2
## * `` -> ...4
```

```
print(paste0("Number of IDs Client would like to exclude: ",
  as.character(length(ExcludedIDs))))
```

```
## [1] "Number of IDs Client would like to exclude: 87"
```

— Import ends here

Tag Original Dataset with ID valid status, Valid ID will be tagged “Valid” in the column:“ValidReponse”, Invalid ones will be tagged “Invalid”

```
TextDataRaw = TextDataRaw%>%
  mutate(ValidResponse = if_else(ResponseId %in% ExcludedIDs,
  "Invalid","Valid"), ValidResponse = as.factor(ValidResponse),Q103 = as.factor(Q103))
NumericDataRaw = NumericDataRaw%>%
  mutate(ValidResponse = if_else(ResponseId %in%
  ExcludedIDs,"Invalid","Valid"),
  ValidResponse = as.factor(ValidResponse),Q103 = as.factor(Q103))
```

Split them just in case

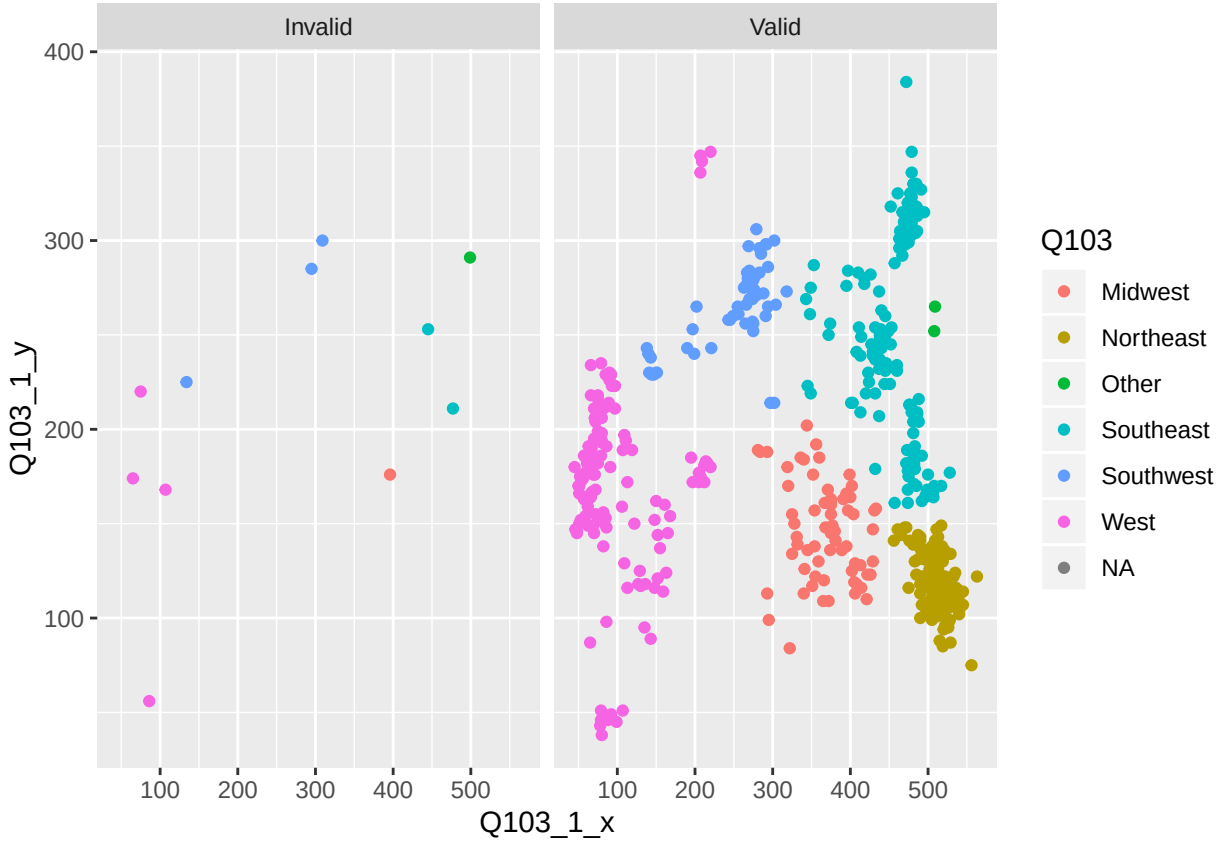
```
TextRawWithoutExIDs = filter(TextDataRaw,!ResponseId %in% ExcludedIDs)
TextRawofExIDs = filter(TextDataRaw,ResponseId %in% ExcludedIDs)
```

— Tag and split ends here

```
geo=select(TextDataRaw,Q103_1_x:Q103,ValidResponse)%>%
  slice(-(1:2))%>%
  mutate(Q103 = as.factor(Q103),Q103_1_x = as.numeric(Q103_1_x),Q103_1_y = as.numeric(Q103_1_y))

ggplot(geo)+aes(x = Q103_1_x,y = Q103_1_y,color = Q103)+
  geom_point()+facet_wrap(geo$ValidResponse)
```

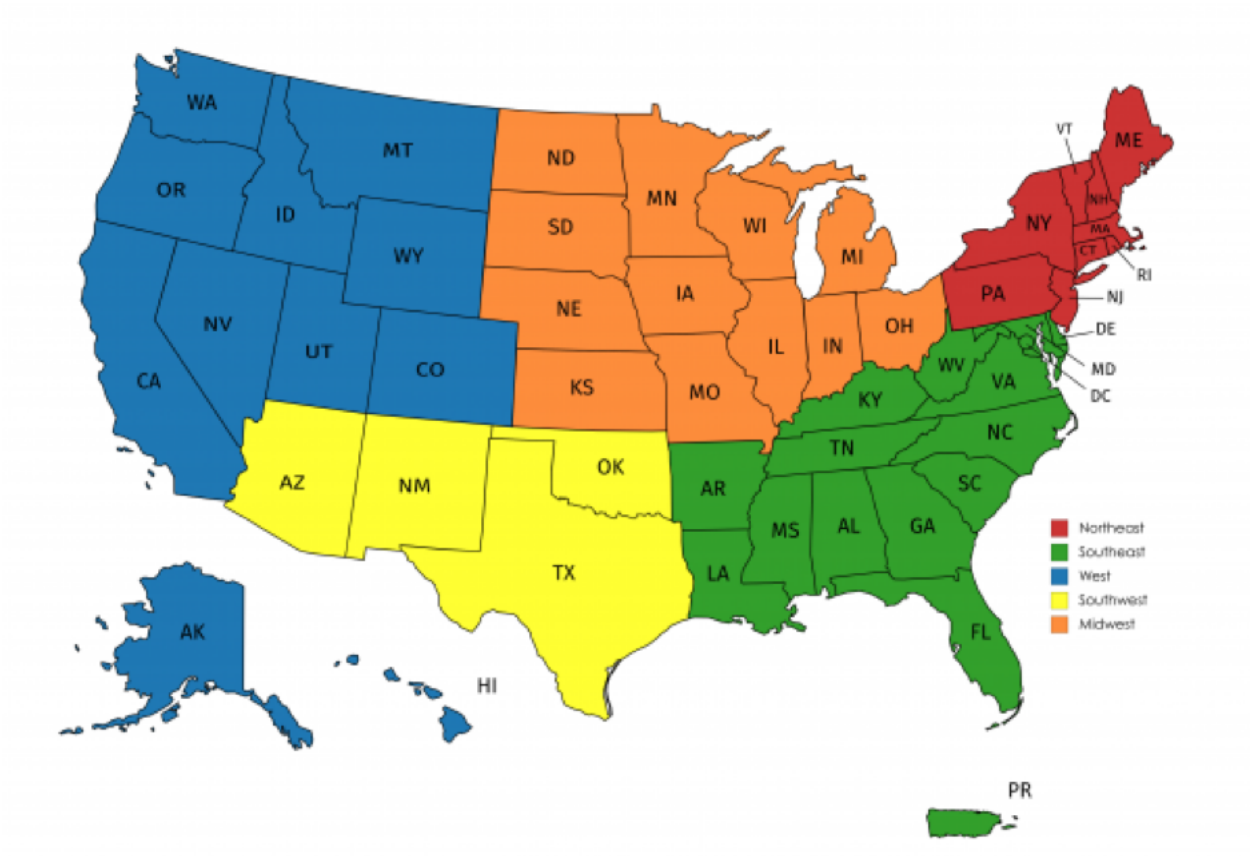
```
## Warning: Removed 76 rows containing missing values (geom_point).
```



What's the map actually used in the survey??

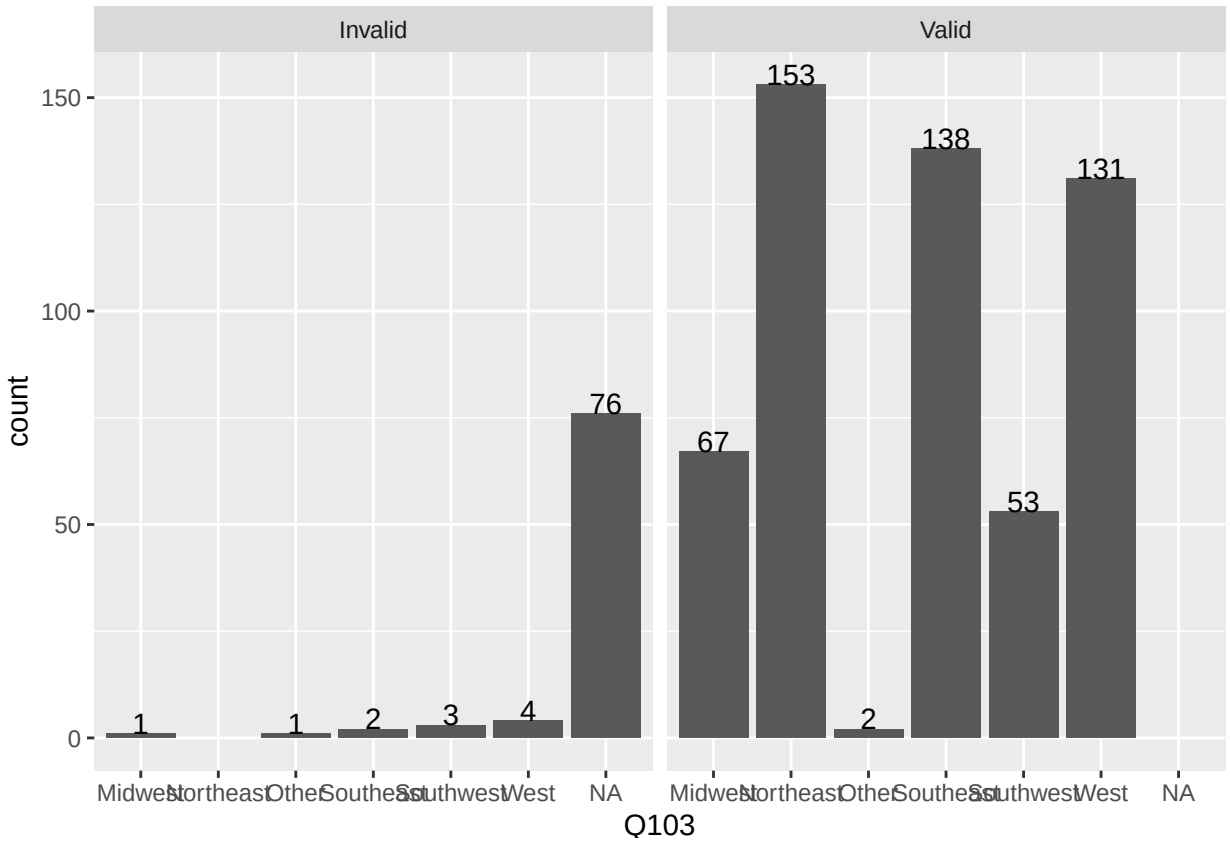
This should be the map from final survey version.

```
img <- readPNG("Picture1.png")
 grid.raster(img)
```



Check counts in each region

```
ggplot(geo) + aes(x = Q103) + geom_bar()+geom_text(stat='count',
aes(label=..count..), vjust=0)+facet_wrap(geo$ValidResponse)
```



We have 76 NA values in geo questions. the population in this category is larger than Midwest and Southwest. Can we simply excluding them?

— Initial visual ends here

— Seperate Demographics

Seperate the Demographic part, setting them as adequate type

```
DemoText = select(TextDataRaw,Q10:Q28,ValidResponse,ResponseId)%>%slice(-(1:2))%>%
  mutate_at(.vars = c(1:4,8:9,12:13,16:21), .funs = as.factor)%>%
  mutate_at(.vars = 5:6, .funs = as.numeric)%>%mutate(Q13 = as.list(strsplit(Q13, ",")),
  Q18 = (as.list(strsplit(Q18, ","))))
```

Combine “Day care” with“Daycare”

```
DemoText = DemoText%>%mutate(Q13_6_TEXT = if_else(DemoText$Q13_6_TEXT == "Daycare","Day care",
          DemoText$Q13_6_TEXT))
```

Right now, two lists are in the result, need to separate them into columns

— Separation For Q13

```
Q13Text = select(DemoText,Q13,ValidResponse,ResponseId)%>%
  unnest(Q13)%>%spread(key = Q13,value = Q13)%>%
  mutate_at(3:9, ~replace(., !is.na(.), 1.))%>%
  mutate_at(3:9, ~replace(., is.na(.), 0.))
```

— Sep for Q13 ends here

— Similarily Separation for Q18

```
Q18Text = select(DemoText,Q18,ValidResponse,ResponseId)%>%
  unnest(Q18)%>%spread(key = Q18,value = Q18)%>%select(-3)%>%
  mutate_at(3:25, ~replace(., !is.na(.), 1.))%>%
  mutate_at(3:25, ~replace(., is.na(.), 0.))
# Bird's colname is buggy replace it
colnames(Q18Text)[colnames(Q18Text)=="Bird (i.e."] <- "Bird"
```

— Sep for Q18 ends here

At this point, we will have DemoText as the demographic part cleaned. Q13Text and Q18Text as the suppliment table to those two questions.

Be aware that a lot of the input had translated into factors, which depends on the graph you would like to plot, may results in some issues.

Have fun!!