Jessica Ghai's Counsulting Project

Team C2

10/15/2019

Data Cleaning and EDA part

There are 3 files, `Numeric.csv`,`ChoiceText.csv` & `Need to be excluded IDs.xlsx`. The 1st two files are the results from the Survey, one in numeric form and the other lays out the actual choices the respondent made.

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

Import "invalid" responses IDs, for the purpose of EDA, we will keep these response in the data and analyze them to determine the conditions or restrictions which would be post on our analysis if we take them out.

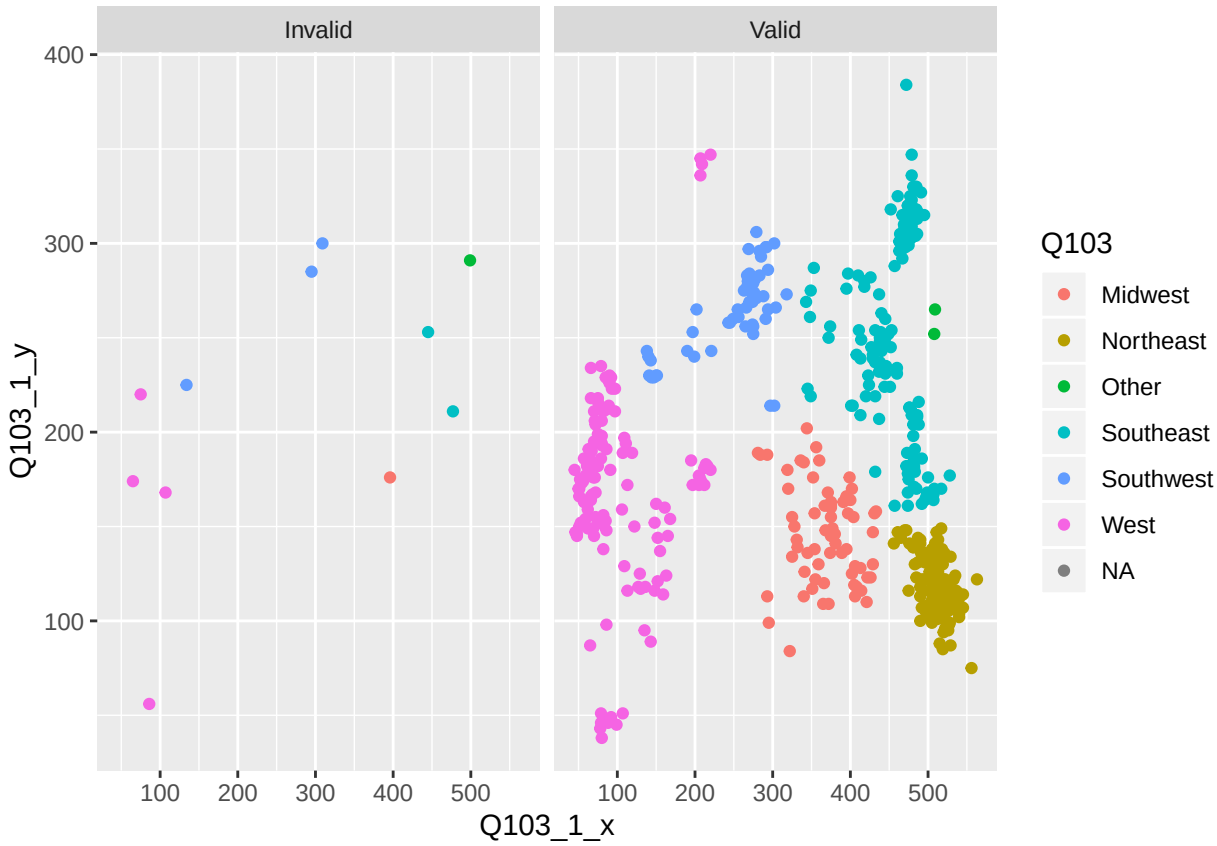
```
ExcludedIDs = readxl::read_xlsx(path = "Need to be excluded IDs.xlsx",
  sheet = 3,col_names = TRUE,col_types = "text")%>%select(1)%>%pull()
print(paste0("Number of IDs Client would like to exclude: ",
  as.character(length(ExcludedIDs))))
```

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

Import ends here.

Initial visualization of the demographic data

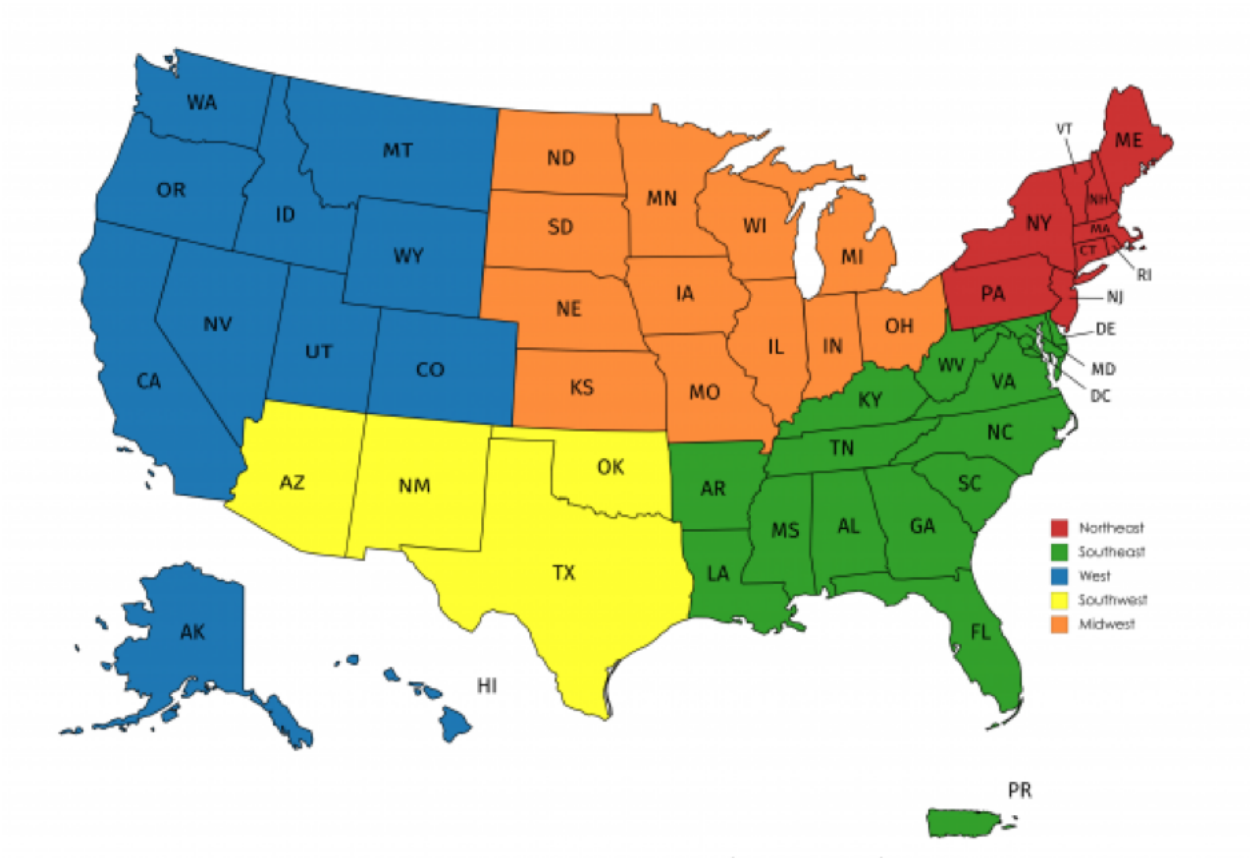
```
geo=select(TextDataRaw,Q103_1_x:Q103,ValidResponse)%>%
  slice(-(1:2))%>%
  mutate(Q103 = as.factor(Q103),Q103_1_x = as.numeric(Q103_1_x),Q103_1_y = as.numeric(Q103_1_y))

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



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

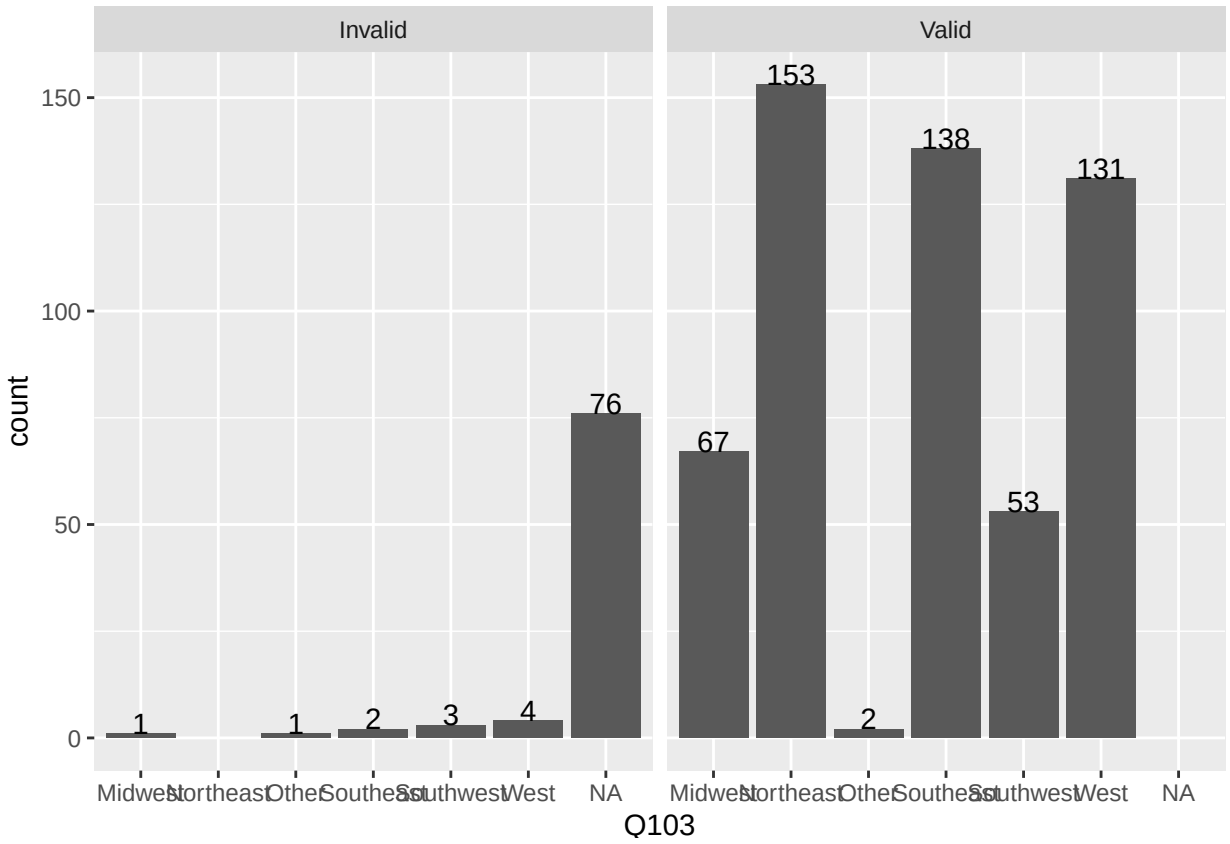
This should be the map from final survey version.

```
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

Correcting column names and factor codings for `DemoText`

```
## check levels:
print(levels(DemoText$Q10))
```

```
## [1] "Board Certified Assistant Behavior Analyst (BCaBA)"
## [2] "Board Certified Behavior Analyst (BCBA)"
```

```
## [3] "Board Certified Behavior Analyst- Doctoral (BCBA-D)"
## [4] "Not certified"
## [5] "Registered Behavior Technician (RBT)"
```

```
## Relevel
DemoText$Q10 = relevel(DemoText$Q10,"Registered Behavior Technician (RBT)")
## Do the same to all demo questions
DemoText$Q104 <-  factor(DemoText$Q104,levels = rev(levels(DemoText$Q104)))

DemoText$Q58_1 <- factor(DemoText$Q58_1, levels = c("< 2 years","2 to 5 years", "6 to 10 years","11 to

DemoText$Q58_2 <-  relevel(DemoText$Q58_2,"Yes")

DemoText$Q19 <-  relevel(DemoText$Q19,"Yes")

DemoText$Q17 <- factor(DemoText$Q17, levels = c("1",    "2",    "3-5", "+5", "None"))

DemoText$Q105 <- factor(DemoText$Q105, levels = c("Never"    ,      "Seldom" ,   "Sometimes","Often"))

DemoText$Q22 <- factor(DemoText$Q22, levels = c("Never"    ,      "Seldom" ,   "Sometimes","Often"))

DemoText$Q29 <-  relevel(DemoText$Q29,"Yes")

DemoText$Q30 <- factor(DemoText$Q30, levels = c("Never"    ,      "Seldom" ,   "Sometimes","Often"))

DemoText$Q24 <-  relevel(DemoText$Q24,"Yes")

DemoText$Q28 <-  relevel(DemoText$Q28,"Yes")

## Rename the columns so the number match the number on the survey
colnames(DemoText) <- c("Q1","Q2","Q3","Q3_text","Q4_x","Q4_y","Q4_Region","Q5a","Q5b","Q6","Q6_text","
```

Separation For Q6

```
Q6Text = select(DemoText,Q6,ValidResponse,ResponseId)%>%
  unnest(Q6)%>%spread(key = Q6,value = Q6)%>%
  mutate_at(3:9, ~replace(., !is.na(.), 1.))%>%
  mutate_at(3:9, ~replace(., is.na(.), 0.))
```

Similarily Separation for Q9

```
Q9Text = select(DemoText,Q9,ValidResponse,ResponseId)%>%
  unnest(Q9)%>%spread(key = Q9,value = Q9)%>%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(Q9Text)[colnames(Q9Text)=="Bird (i.e."] <- "Bird"
```

At this point, we will have DemoText as the demographic part cleaned. Q6Text and Q9Text as the suppliment table to those two questions. What's missing here is the EDA on the valid and invalid groups to see if they belongs to the same population or not. In other words, what and how much bias will be introduced into the sample if we exclude them?

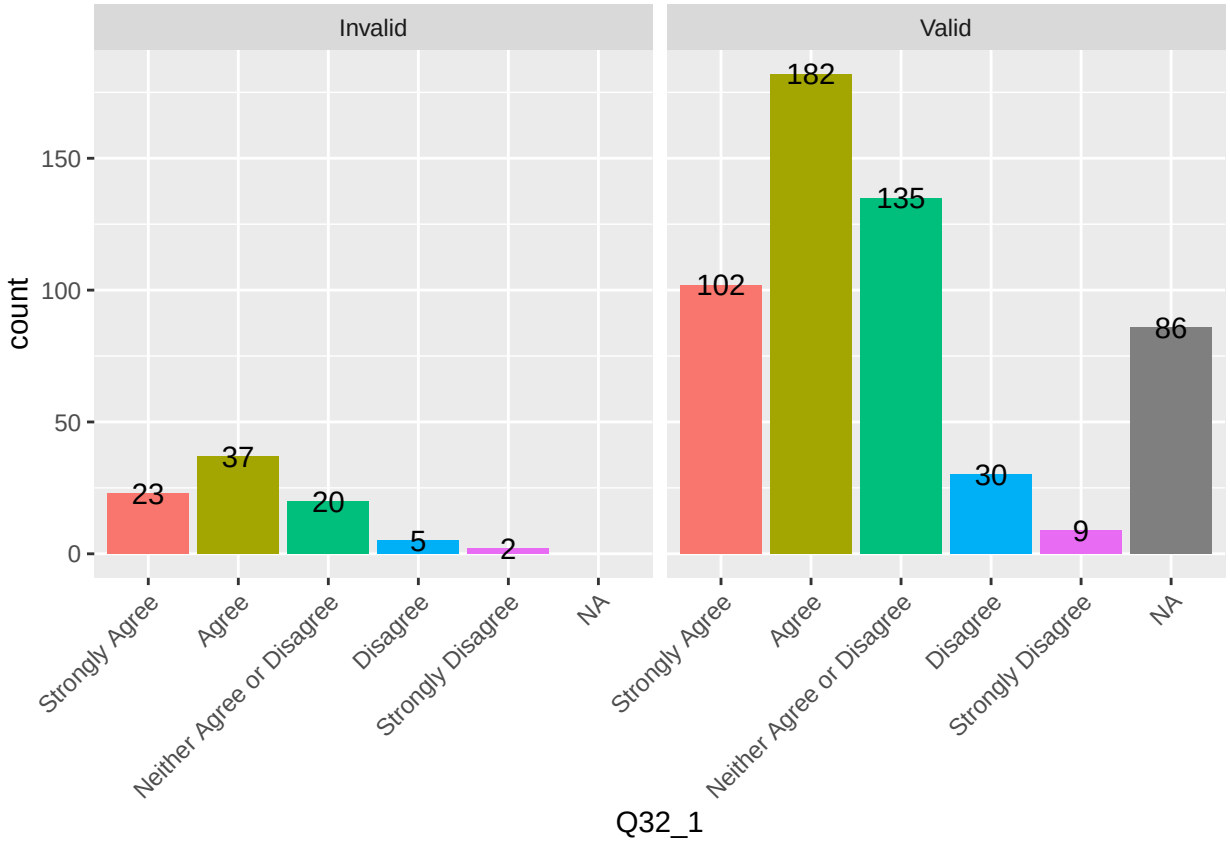
Last Question Data Cleaning

Preparing the data

```
LastQText = select(TextDataRaw,Q32_1:Q32_4,ValidResponse,ResponseId)%>%slice(-(1:2))

Q32_1Text = select(LastQText,Q32_1,ValidResponse,ResponseId)%>%unnest(Q32_1)%>%spread(key = Q32_1,value
Q32_1Text = Q32_1Text[,c(1,2,6,3,5,4,7,8)]

Q32_2Text = select(LastQText,Q32_2,ValidResponse,ResponseId)%>%unnest(Q32_2)%>%spread(key = Q32_2,value
Q32_2Text = Q32_2Text[,c(1,2,6,3,5,4,7,8)]

Q32_3Text = select(LastQText,Q32_3,ValidResponse,ResponseId)%>%unnest(Q32_3)%>%spread(key = Q32_3,value
Q32_3Text = Q32_3Text[,c(1,2,6,3,5,4,7,8)]

Q32_4Text = select(LastQText,Q32_4,ValidResponse,ResponseId)%>%unnest(Q32_4)%>%spread(key = Q32_4,value
Q32_4Text = Q32_4Text[,c(1,2,6,3,5,4,7,8)]
```

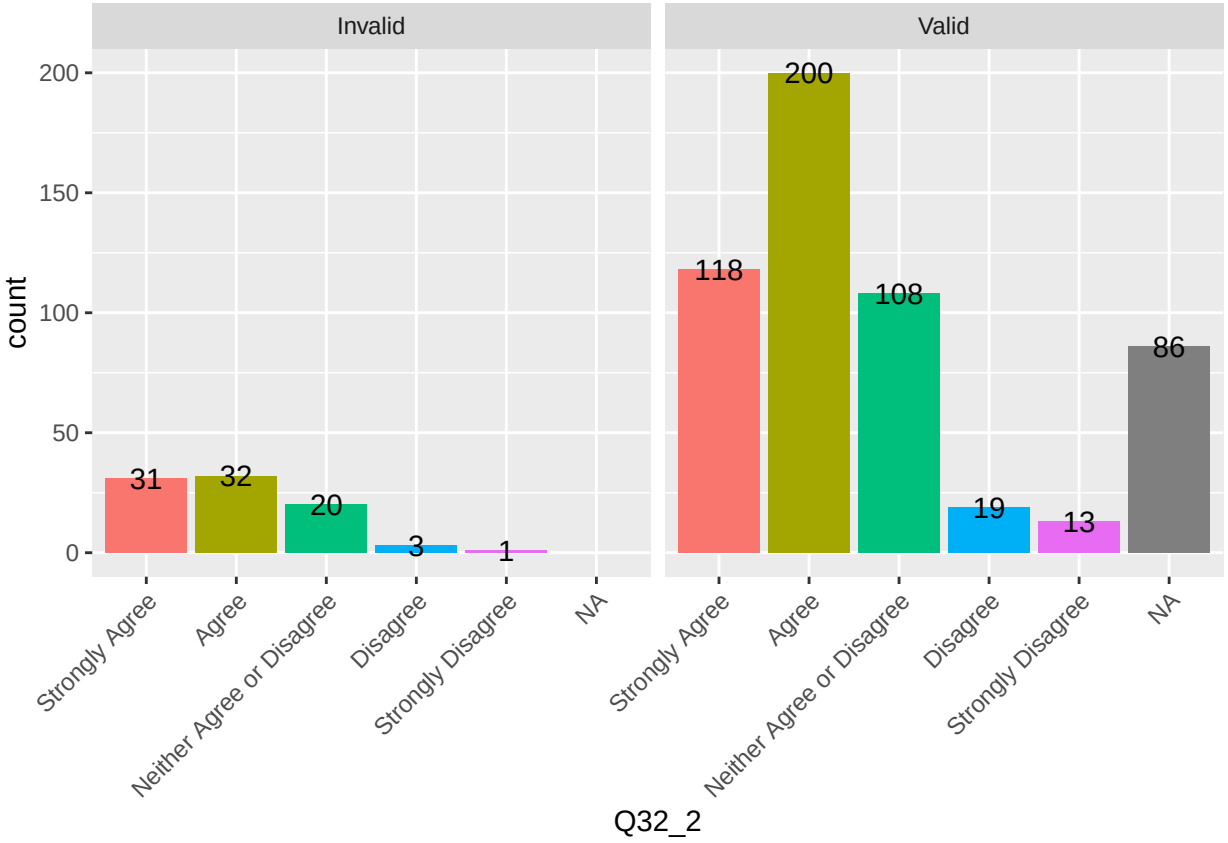
Q55a.1.I believe a majority of the target population outlined above would respond positively to therapeutic services (e.g., ABA services) that incorporate animals.

```
LastQText$Q32_1 <- factor(LastQText$Q32_1,levels = c("Strongly Agree", "Agree", "Neither Agree or Disag
ggplot(LastQText) + aes(Q32_1,fill = Q32_1)+geom_bar(show.legend = FALSE)+geom_text(stat='count', aes(la
```



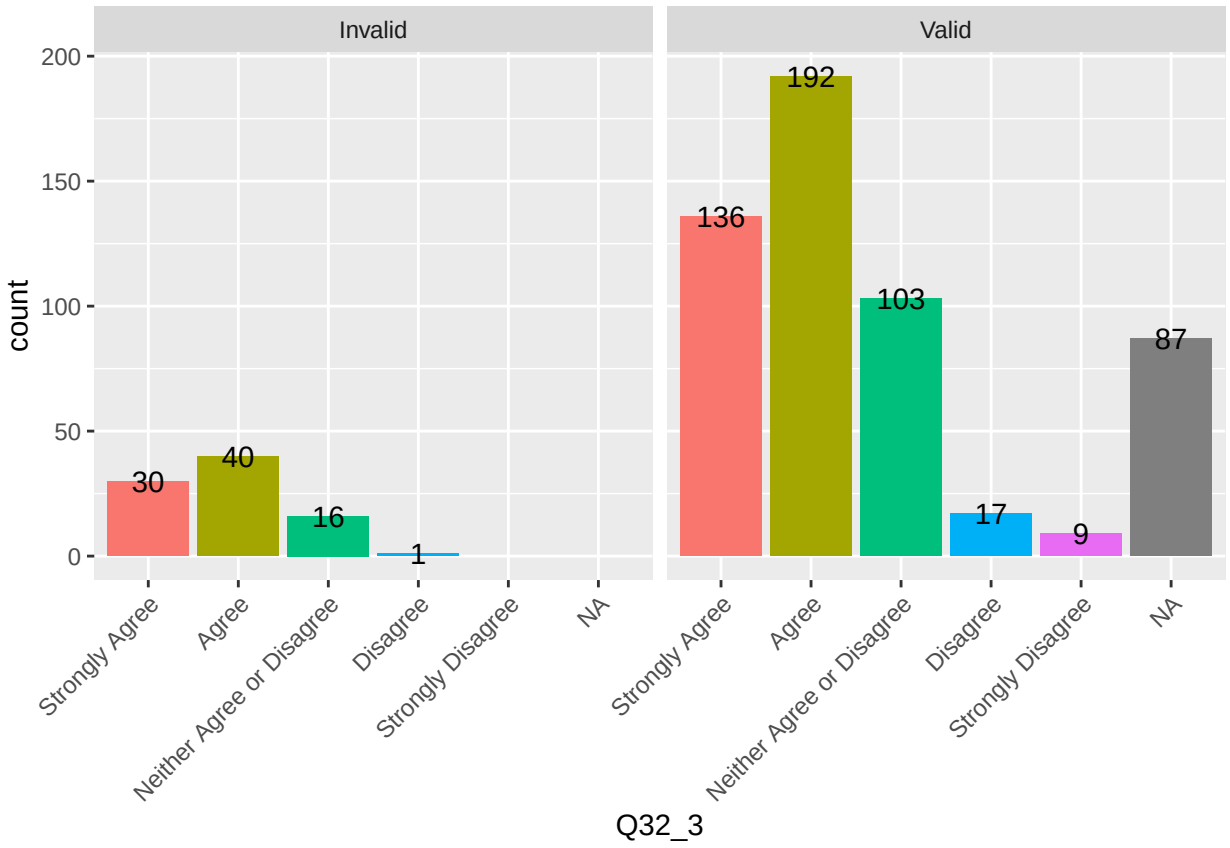
55b.2.I believe the presence of an animal can have a calming effect on many individuals within the target population outlined above

```
LastQText$Q32_2 <- factor(LastQText$Q32_2,levels = c("Strongly Agree", "Agree", "Neither Agree or Disagr
ggplot(LastQText) + aes(Q32_2,fill = Q32_2)+geom_bar(show.legend = FALSE)+geom_text(stat='count', aes(la
```



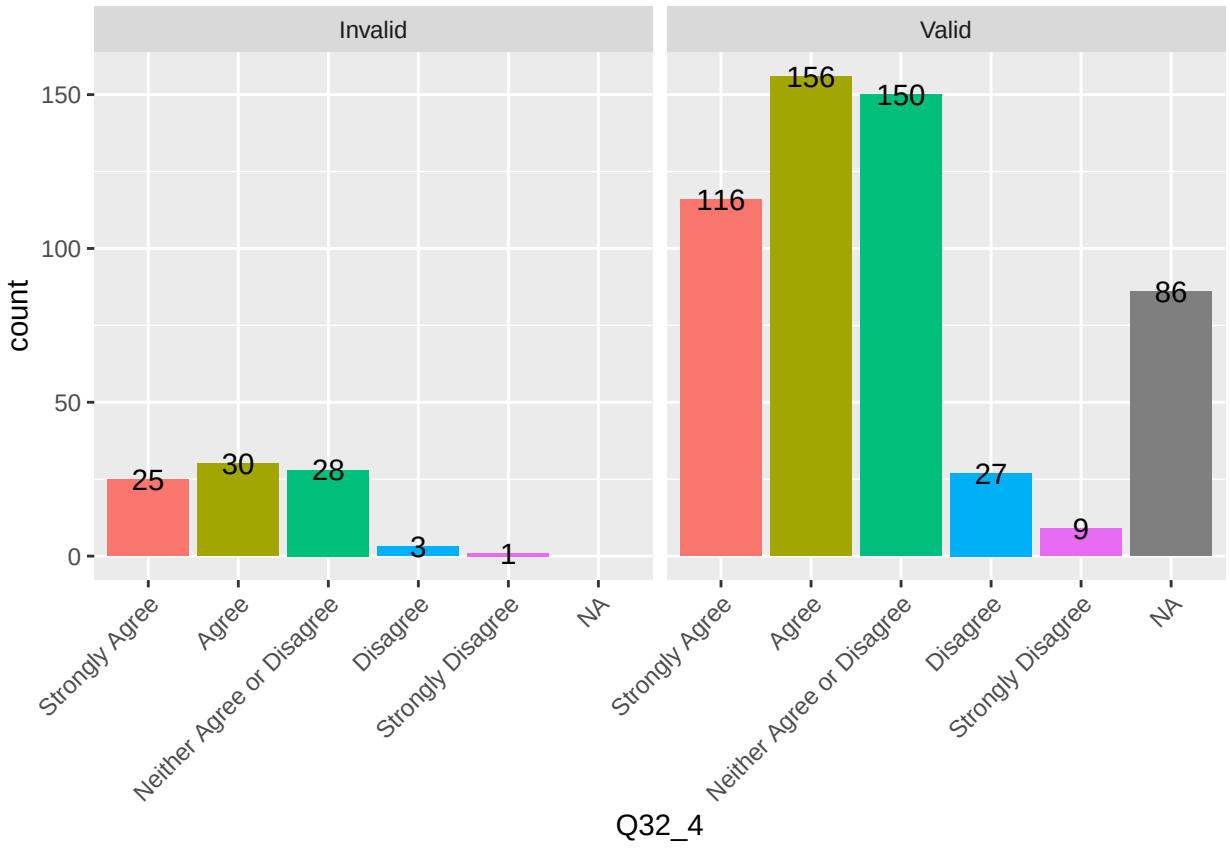
55c.3.I believe the presence of an animal can increase the frequency of social interaction opportunities between individuals within the target population outlined above and other people.

```
LastQText$Q32_3 <- factor(LastQText$Q32_3,levels = c("Strongly Agree", "Agree", "Neither Agree or Disagr
ggplot(LastQText) + aes(Q32_3,fill = Q32_3)+geom_bar(show.legend = FALSE)+geom_text(stat='count', aes(la
```



55d.4.I believe animals can aid in effectively teaching social interaction and communication skills associated with ASD symptomatology for a majority of the target population outlined above.

```
LastQText$Q32_4 <- factor(LastQText$Q32_4,levels = c("Strongly Agree", "Agree", "Neither Agree or Disagr
ggplot(LastQText) + aes(Q32_4,fill = Q32_4)+geom_bar(show.legend = FALSE)+geom_text(stat='count', aes(la
```



Testing the Invalid IDs' Responses

Check if the factor is adequatly transfered into integers

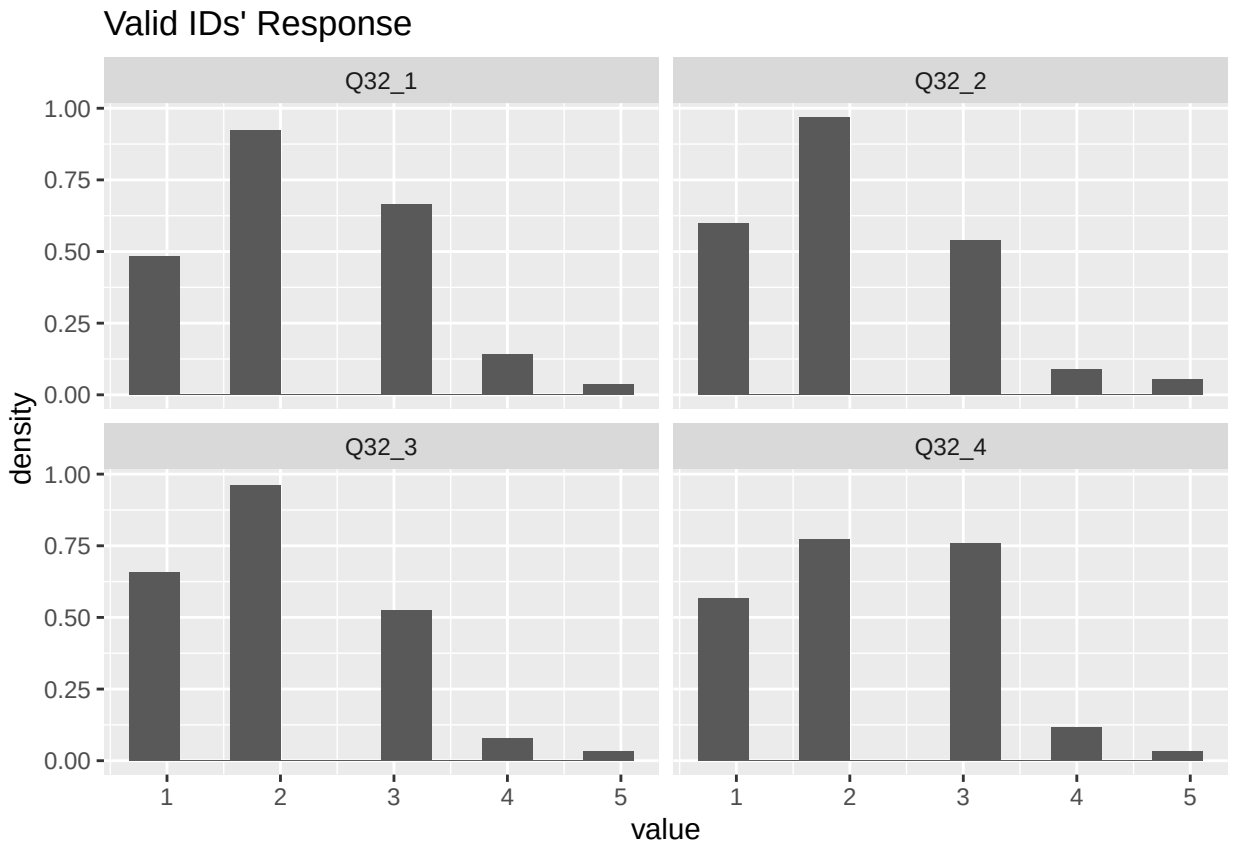
```
LastQTextTEST <- LastQText%>%mutate_at(.var = 1:4,.funs = as.factor)%>%mutate_at(.var = 1:4,.funs = as.
```

Create separate tables for Valid and Invalid Responses

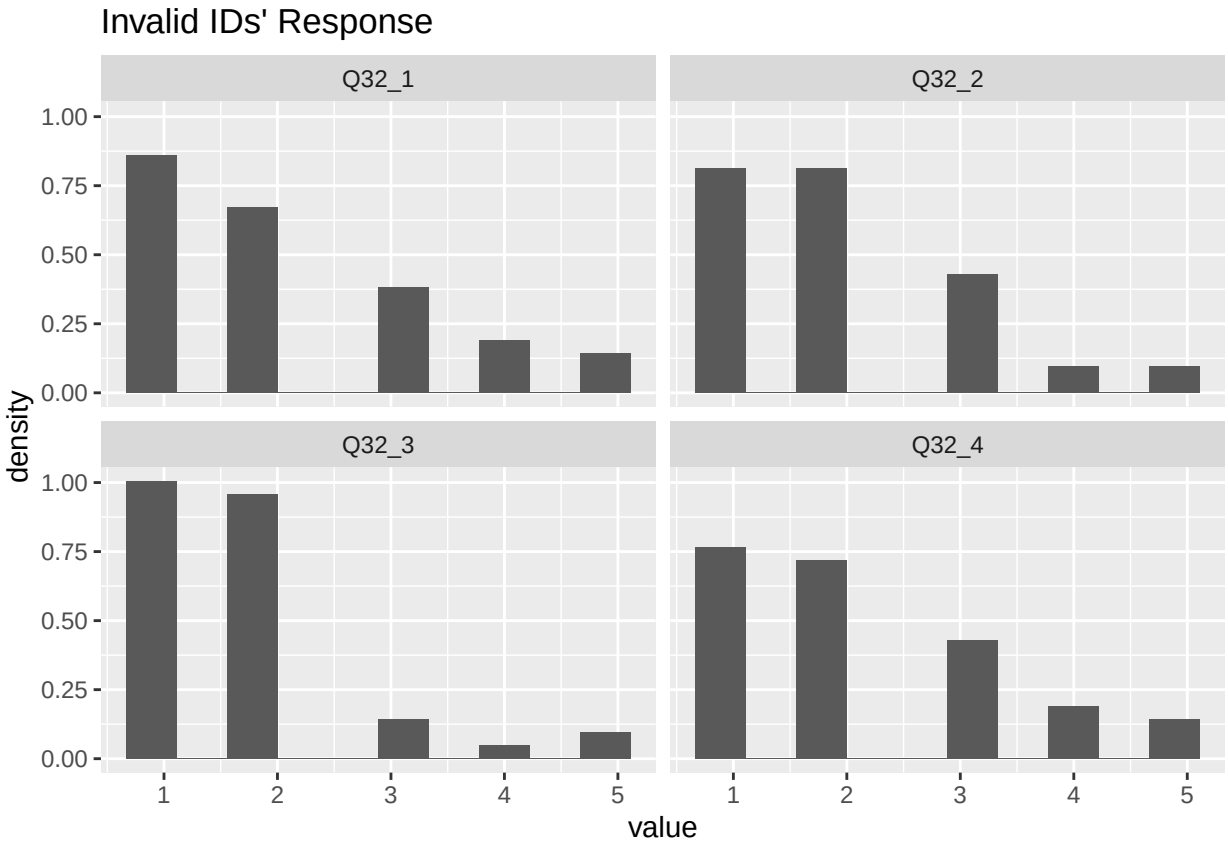
```
LQvalid <-  LastQTextTEST%>%filter(ValidResponse == "Valid")%>%select(1:4)%>%drop_na()
LQInvalid<-  LastQTextTEST%>%filter(ValidResponse == "Invalid")%>%select(1:4)%>%drop_na()
```

Plot the Histogram, but 1st needs to rearrange the data frame

```
LQvalid1 <- LQvalid %>% melt(measure.vars = 1:4)
ggplot(LQvalid1)+aes(x = value)+geom_histogram(bins = 10,aes(y=..density..))+facet_wrap(LQvalid1$variabl
```




```
LQInvalid1 <- LQInvalid %>% melt(measure.vars = 1:4)
ggplot(LQInvalid1)+aes(x = value)+geom_histogram(bins = 10,aes(y=..density..))+facet_wrap(LQInvalid1$va
```



The data does not seem to have a normal distribution, can we still use t test to test these two sample means?

In case we can, below is the t test results.

```
for(i in 1:4){
  print(t.test(select(LQInvalid,i)%>%pull(),select(LQvalid,i)%>%pull()))
}
```

```
##
##  Welch Two Sample t-test
##
## data:  select(LQInvalid, i) %>% pull() and select(LQvalid, i) %>% pull()
## t = -0.5746, df = 51.084, p-value = 0.5681
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4699758  0.2608059
## sample estimates:
## mean of x mean of y
##  2.148936  2.253521
##
##
##  Welch Two Sample t-test
##
## data:  select(LQInvalid, i) %>% pull() and select(LQvalid, i) %>% pull()
## t = -0.52465, df = 52.917, p-value = 0.602
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.4061454  0.2377307
## sample estimates:
## mean of x mean of y
##  2.042553  2.126761
##
##
##  Welch Two Sample t-test
##
## data:  select(LQInvalid, i) %>% pull() and select(LQvalid, i) %>% pull()
## t = -1.7922, df = 53.435, p-value = 0.07877
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.56169191  0.03153223
## sample estimates:
## mean of x mean of y
##  1.787234  2.052314
##
##
##  Welch Two Sample t-test
##
## data:  select(LQInvalid, i) %>% pull() and select(LQvalid, i) %>% pull()
## t = -0.11495, df = 51.44, p-value = 0.9089
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3809444  0.3396755
## sample estimates:
## mean of x mean of y
##  2.212766  2.233400
```

Judging from the t statistics, it seems that the id excluded were appeared to have high probability that they comes from the same population? Can we safely exclude them? Or does it matters if we exclude them or not?