01_Data management_02_visualization

Hyemin Park(hyemin.park@snu.ac.kr)

14 January 2022

0. check data

00.data set

```
df <- df[,-1]
df
```

```
## # A tibble: 177 x 127
##       ID BIRTH_YEAR LATINO_DUM RACE  RACE_TEXT SEX    SEX_ORIENT LI1_POINTLESS
##    <dbl>      <dbl> <chr>      <chr> <chr>     <chr>  <chr>      <chr>
##  1  4138       1998 N_latino   Black <NA>      Male   straight   Vdisagree
##  2  4534       1999 N_latino   Black <NA>      Female straight   Vdisagree
##  3  4552       1997 N_latino   Asian <NA>      Male   gay        disagree
##  4  4393       2000 Y_latino   White <NA>      Female straight   disagree
##  5  4483       2000 N_latino   White <NA>      Female straight   Vdisagree
##  6  3982       2000 N_latino   White <NA>      Female straight   Vdisagree
##  7  3865       1999 N_latino   White <NA>      Female straight   Vdisagree
##  8  4630       2000 N_latino   White <NA>      Female straight   Vdisagree
##  9  3961       2000 N_latino   Black <NA>      Female straight   disagree
## 10  4183       1999 N_latino   White <NA>      Female straight   disagree
## # ... with 167 more rows, and 119 more variables: LI2_PARENTSSAIDSO <chr>,
## #   LI3_FORCEDBYPARENTS <chr>, LI4_FORCEDBYTEACHERS <chr>, LI5_NOSCHOOL <chr>,
## #   LI6_PROMISEDGIFTS <chr>, LI7_AVOIDSCOLDING <chr>, PEI1_LOOKBETTER <chr>,
## #   PEI2_GETBETTERGRADE <chr>, PEI3_LOOKHARDWORKER <chr>,
## #   PEI4_PERFORMBETTER <chr>, PEI5_NOTLOOKINCOMPETENT <chr>,
## #   PEI6_PROVESMART <chr>, PEI7_NOTFEELFAILURE <chr>, PGI1_NOTLEARNWELL_ <chr>,
## #   PGI2_MISTAKE <chr>, PGI3_NOTFINISHHW <chr>, IDI1_VALUELEARNING <chr>, ...
```

01. missing values

```
sum(is.na(df))
```

```
## [1] 456
```

```
sum(is.na(df[1:7]))
```

```
## [1] 183
```

```r
sum(is.na(df[8:99]))
```

```
## [1] 0
```

```r
sum(is.na(df[100:127]))
```

```
## [1] 273
```

```r
colSums(is.na(df[1:7]))
```

```
##         ID BIRTH_YEAR LATINO_DUM       RACE  RACE_TEXT        SEX SEX_ORIENT
##          0          0          0          7        176          0          0
```

```r
colSums(is.na(df[100:127]))
```

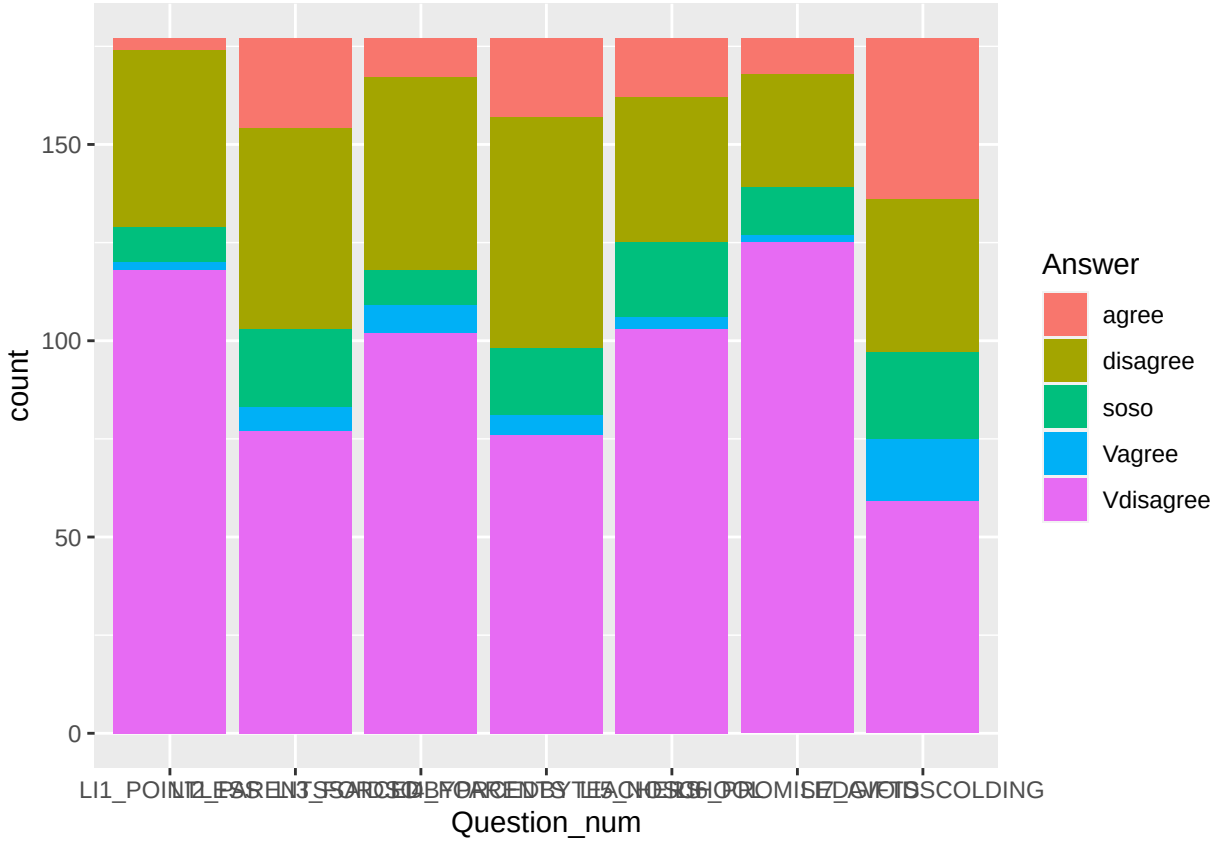
```
##             TYPE_OF_TASK     ATITUDE_TOWARDS_MATH             GOOD_AT_MATH
##                        0                        3                        3
##                  JOY_MIN                  JOY_MAX                 JOY_MEAN
##                        1                        1                        1
##                  JOY_FRE                  ANX_MIN                  ANX_MAX
##                        1                        1                        1
##                 ANX_MEAN                  ANX_FRE           JOY_PERCENTAGE
##                        1                        1                        0
##       ANXIETY_PERCENTAGE                      N_H                      N_M
##                        0                        6                        6
##                      N_L                  N_WRONG                 AVGSCORE
##                        6                        6                        6
##               TOTALSCORE                    STATE ATITUDE_TOWARDS_LITERACY
##                        6                        0                        8
##         GOOD_AT_LITERACY                 SUMSCORE                       RT
##                        8                        6                        6
##       TIME_SELF_RELIENCE       TIME_HELP_SEEKING          TIME_LOW_PERSIS
##                        6                        6                        6
##            VEC_VARIABLES
##                      177
```

1. bar charts

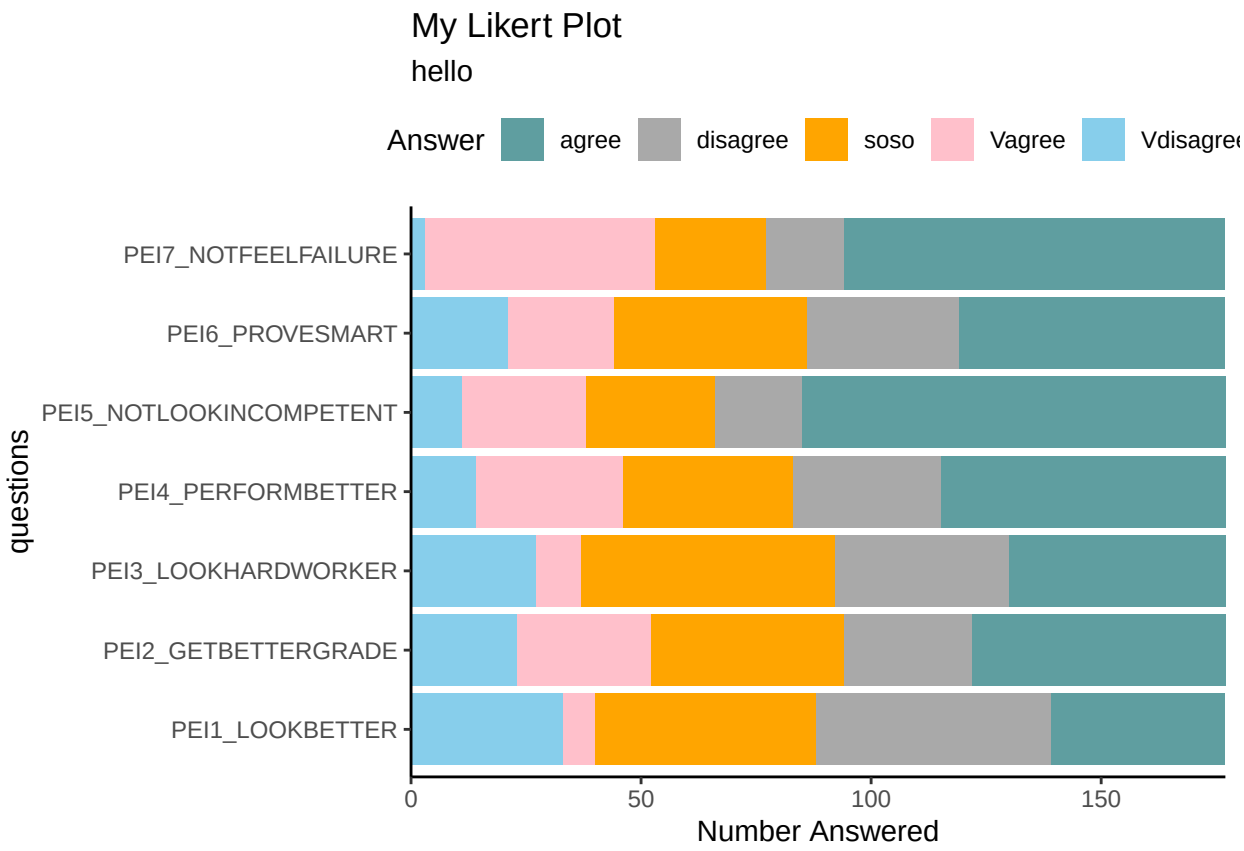
```r
### select items: Internalization
items1 <- select(df, c("ID",starts_with("LI")))
items2 <- select(df, c("ID",starts_with("PEI")))
items3 <- select(df, c("ID",starts_with("PGI")))
items4 <- select(df, c("ID",starts_with("IDI")))
items5 <- select(df, c("ID","RACE","SEX","SEX_ORIENT", starts_with("INI")))


### Transform the items
items1_bar <- items1 %>% gather(key='Question_num', value='Answer', -ID)
items2_bar <- items2 %>% gather(key='Question_num', value='Answer', -ID)
```

```r
items3_bar <- items3 %>% gather(key='Question_num', value='Answer', -ID)
items4_bar <- items4 %>% gather(key='Question_num', value='Answer', -ID)
items5_bar <- items5 %>% gather(key='Question_num', value='Answer', -ID)


### Draw a stacked bar chart with items2
plot1 <- ggplot(items1_bar, aes(x=Question_num)) +
  geom_bar(aes(fill=Answer))


### Change options wth items2
plot2 <- ggplot(items2_bar, aes(x=Question_num)) +
  geom_bar(aes(fill=Answer)) +
  scale_fill_manual(values= c ("Cadetblue", "Dark gray","orange","Pink","Sky Blue")) +
  theme(axis.title = element_text (face = "plain", size = 15, color = "black"),
        axis.text.x = element_text(size=50),
        strip.text.x = element_text(size = 5)) +
  scale_y_continuous(expand=expansion(0)) +
  labs(
    title='My Likert Plot', subtitle='hello',
    x='questions', y='Number Answered'
  ) +
  theme_classic() +
  theme(legend.position='top') +
  coord_flip()


### Change to percentage with items3 and items4
items3.4_bar <- rbind(items3_bar, items4_bar)
plot3 <- ggplot(items3.4_bar, aes(x=Question_num)) +
  geom_bar(aes(fill=Answer), position="fill") +
  scale_fill_brewer(palette='Spectral', direction=-1) +
  scale_y_continuous(expand=expansion(0), labels=scales::percent_format()) +
  labs(
    title='My Likert Plot', subtitle='Twenty Questions!',
    x='Questions', y='Number Answered'
  ) +
  theme_classic() +
  theme(legend.position='top') +
  coord_flip()


###  items5
plot4 <- ggplot(items5, aes(x=RACE, y=INI1_KNOWPURPOSE))+
  geom_bar(position="stack", stat="identity")

plot5 <- ggplot(items5, aes(x=SEX, y=INI1_KNOWPURPOSE))+
  geom_bar(position="stack", stat="identity")

plot6 <- ggplot(items5, aes(x=SEX_ORIENT, y=INI1_KNOWPURPOSE))+
  geom_bar(position="stack", stat="identity")
```

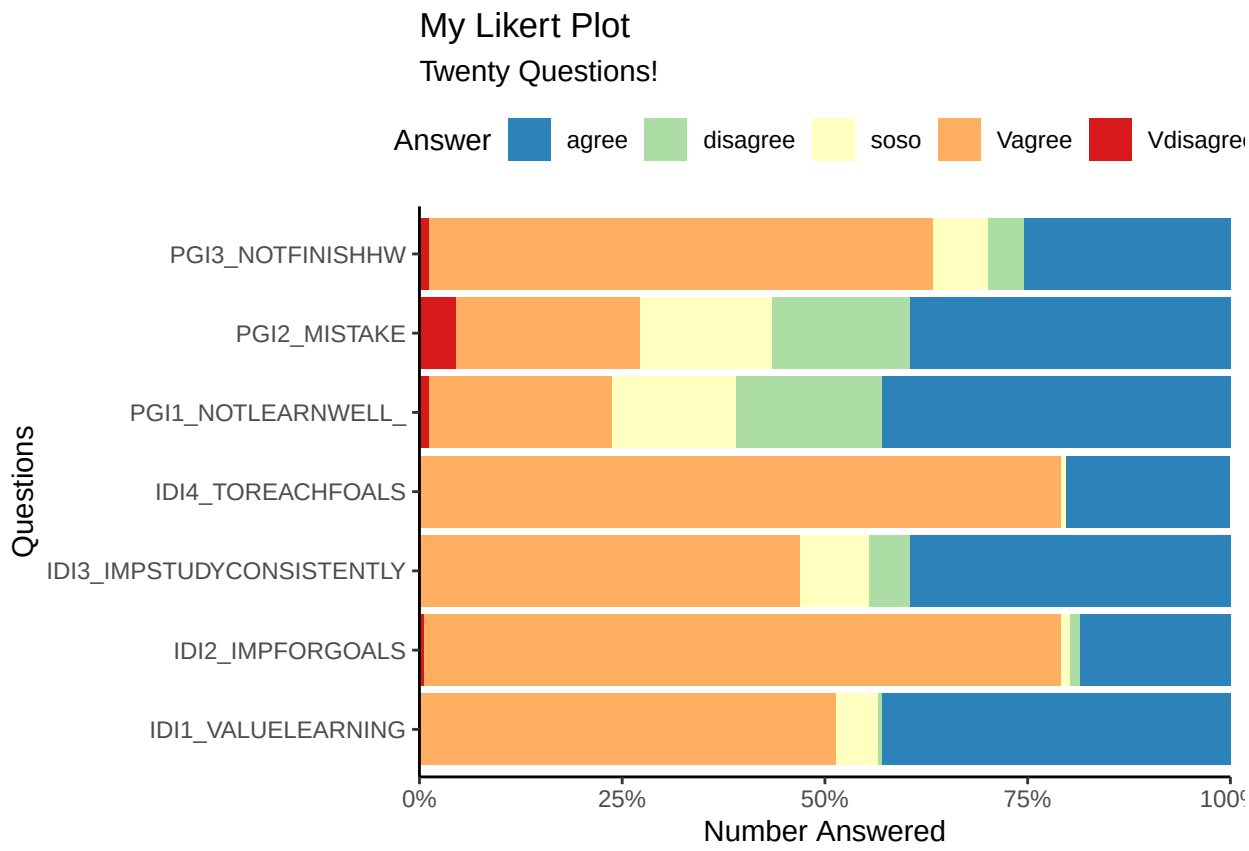
```
### show plots
plot1
```



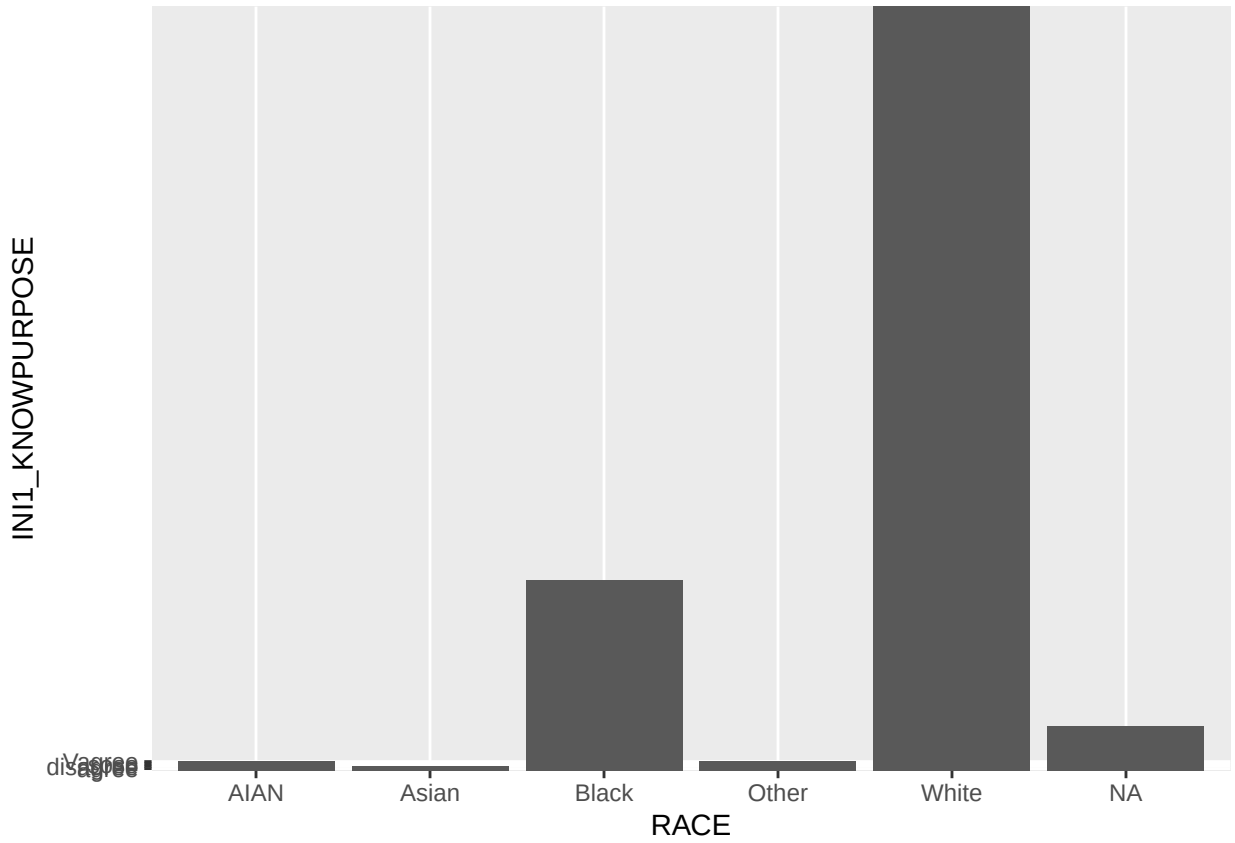

```
plot2
```


My Likert Plot
hello
Answer
agree
disagree
soso
Vagree
Vdisagree
PEI7_NOTFEELFAILURE
PEI6_PROVESMART
PEI5_NOTLOOKINCOMPETENT
PEI4_PERFORMBETTER
PEI3_LOOKHARDWORKER
PEI2_GETBETTERGRADE
PEI1_LOOKBETTER
questions
0
50
100
150
Number Answered

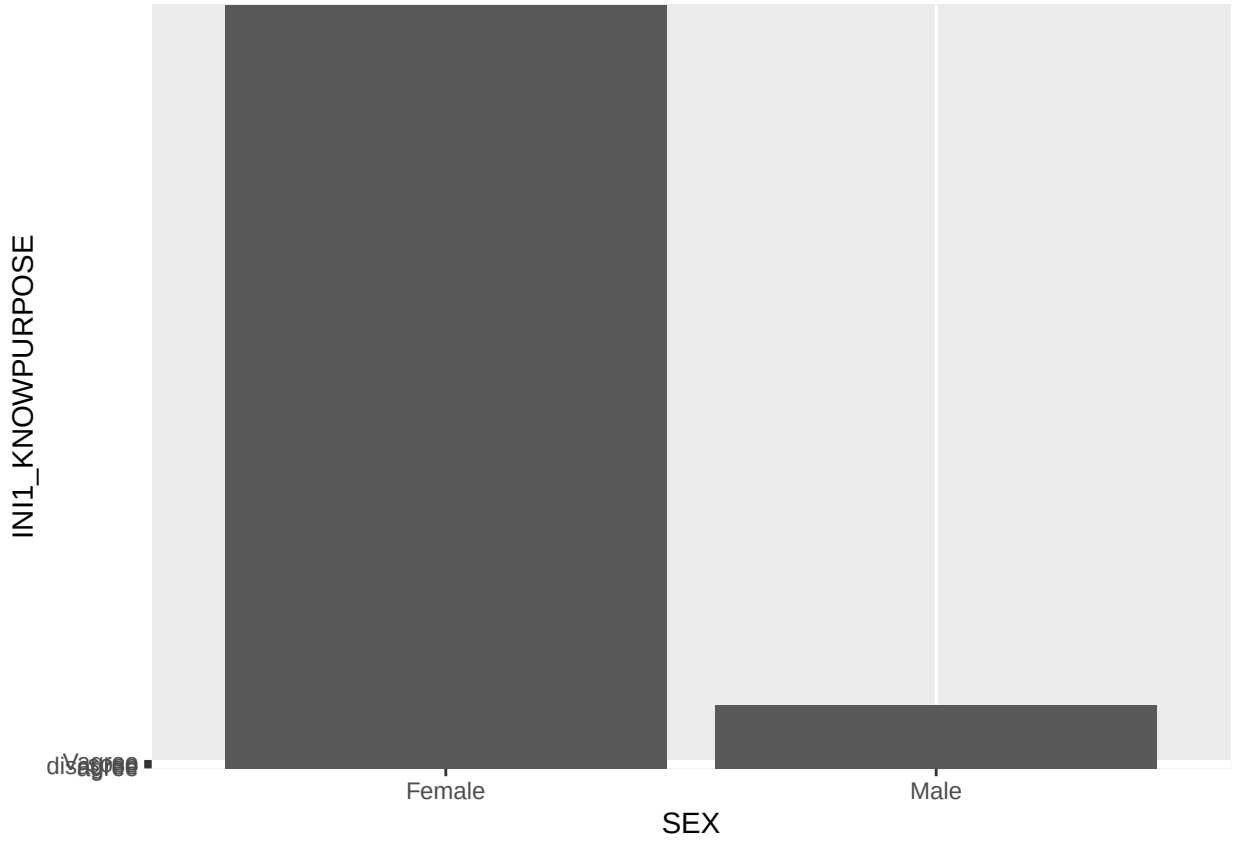
```
plot3
```


My Likert Plot
Twenty Questions!
Answer
agree
disagree
soso
Vagree
Vdisagre
PGI3_NOTFINISHHW
PGI2_MISTAKE
PGI1_NOTLEARNWELL_
IDI4_TOREACHFOALS
IDI3_IMPSTUDYCONSISTENTLY
IDI2_IMPFORGOALS
IDI1_VALUELEARNING
Questions
0%
25%
50%
75%
100
Number Answered

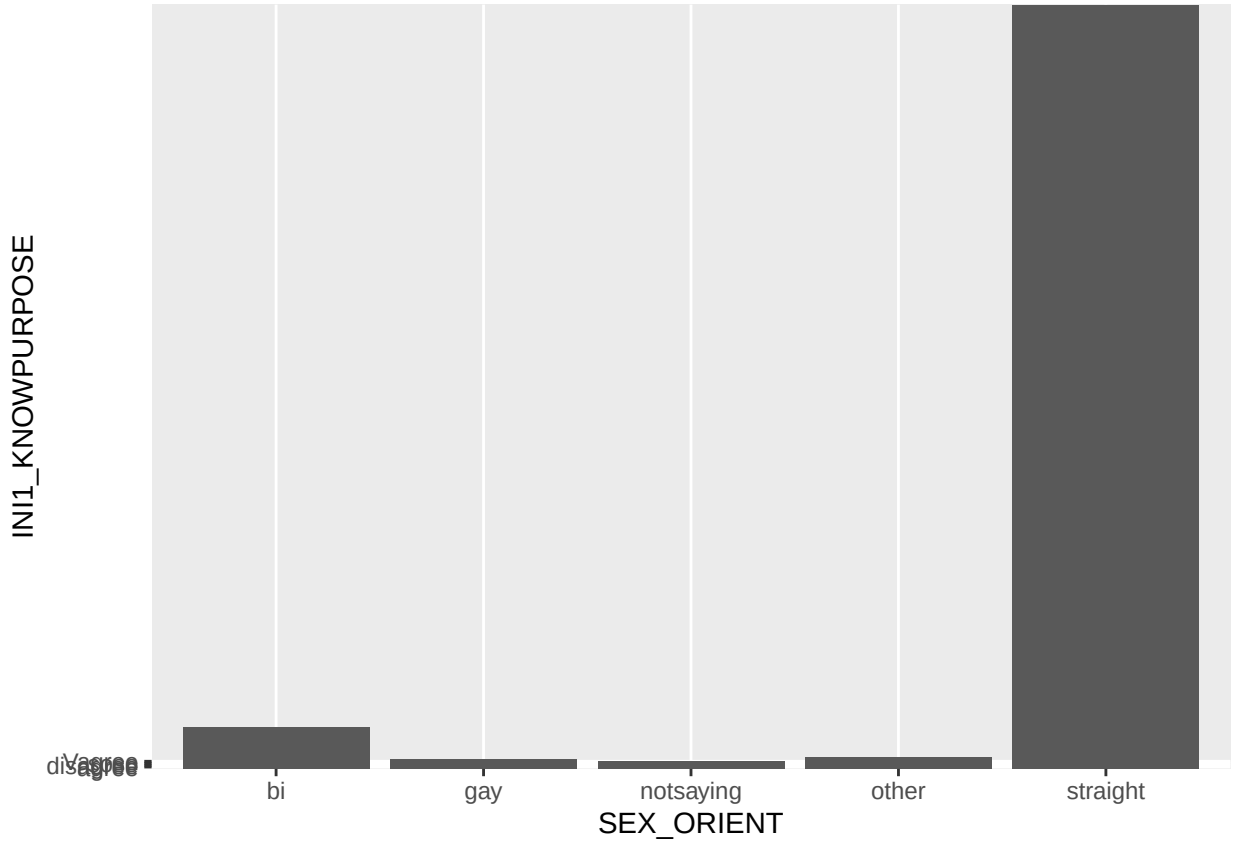
```
plot4
```




```
plot5
```




```
plot6
```




```
## LESSONS

  1. basics
  - Shortcut keys of a code chunk in window is ‘Ctrl’+’Alt’+’i‘
  - Rmarkdown used grammar of markdown
    + ’‘words‘’: empathize words
    + ’*italic*’ or ’_italics_’: italic
    + ’**bold**’ or ’__bold__’: bold
    + ’~~strikethrough~~’: strikethrough
    + ’[insert link, name](https://www.linkedin.com/in/hyemin-park-285a43204/)’: insert hiperlink. put a
    + ’![insert figure, name if wanted](https://www.google.com/url?sa=i&url=https%3A%2F%2Fblankspace-dev

  2. upgrades and data analysis
  - code chunk option -> {r setup, ‘include=FALSE‘}
  this chunk is included in the R documents but not printed usually when we import libraries and data
  - use cheat sheet
```