Lab 11 - Data, Aesthetics, & Geometries

Your Name Here

November 9, 2017

Complete the following exercises below. Knit together the PDF document and commit both the Lab 11 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Which variables in your dataset are you interested in visualizing? Describe the level of measurement of these variables and what type of geography you think is appropriate to represent these variables. Give your reasoning for choosing the `geom_()` you selected.

I'm interested in visualizing how the personality characterisitcs NOASSERT and PLEASE affect the coerced intercourse variables AUTHSI and PRESSSI. I also want to account for race in these comparisons. All of these variables are categorical and because of that I think the geom_bar() function could be the most approriate function if I want to look at multiple variables at the same time.

2. Is your data in the proper format to visualize the data in the way you want? Why or why not? *If you need/want to change the structure of your data, do it below.*

It is for some comparisons, but if I want to look at NOASSERT and PLEASE at the same time I may have to combine them into a new column.

3. Create at least two different exploratory plots of the variables you chose using the skills we covered in class today. What types of mapping aesthetics did you choose and why? What do these plots tell you about your data?

I arranged my data in two bar charts each having two variables co-ocurring which are then facetted by race. I also did another bar chart looking at NOASSERT, PLEASE, AUTHSI, and PRESSSI. These help me understand the levels of responses for each answer in the different variables, however, they are not the most meaningful plots for understanding the interaction between the different variables which is what I actually want to represent.

```r
#load haven package
#install.packages("haven")
library("haven")

#load tidyverse and dplyr
library("tidyverse")
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

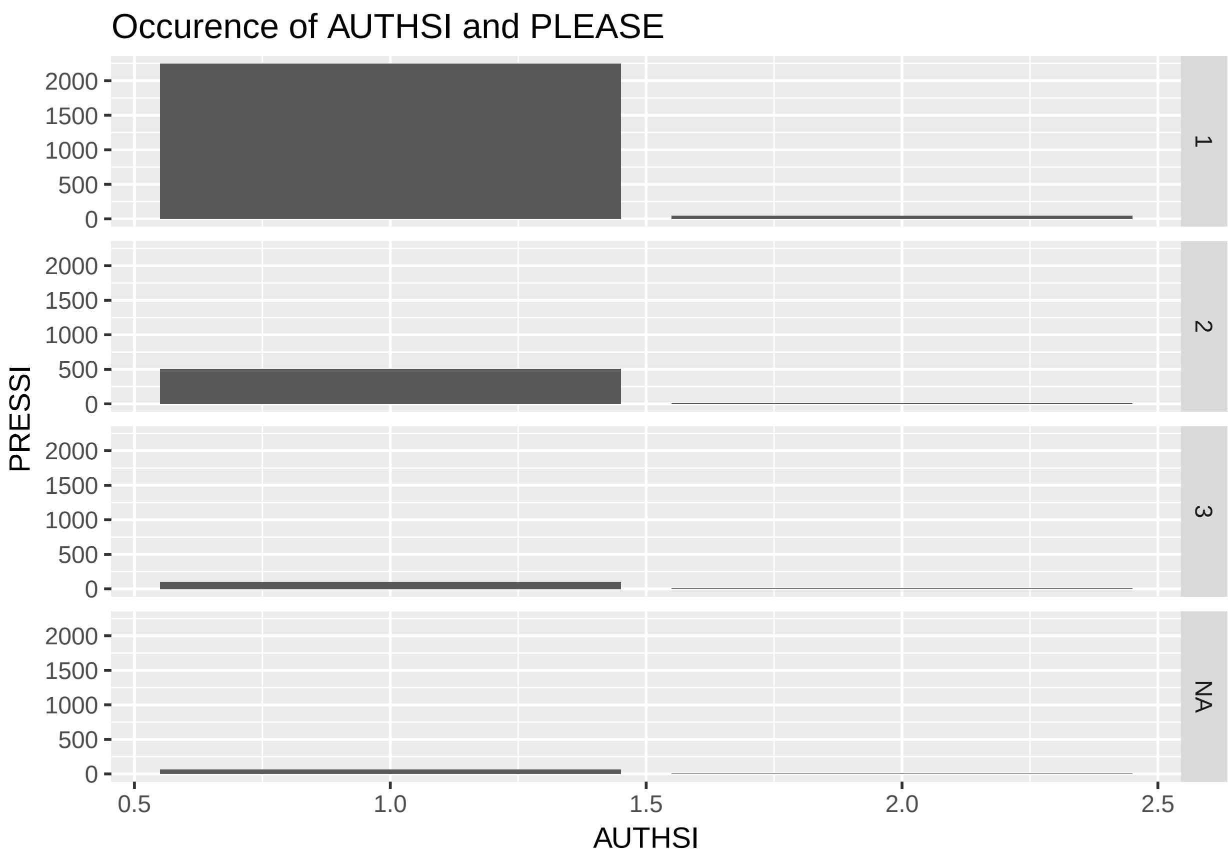
```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```r
library("dplyr")

#install.packages("reshape2")
library("reshape2")
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

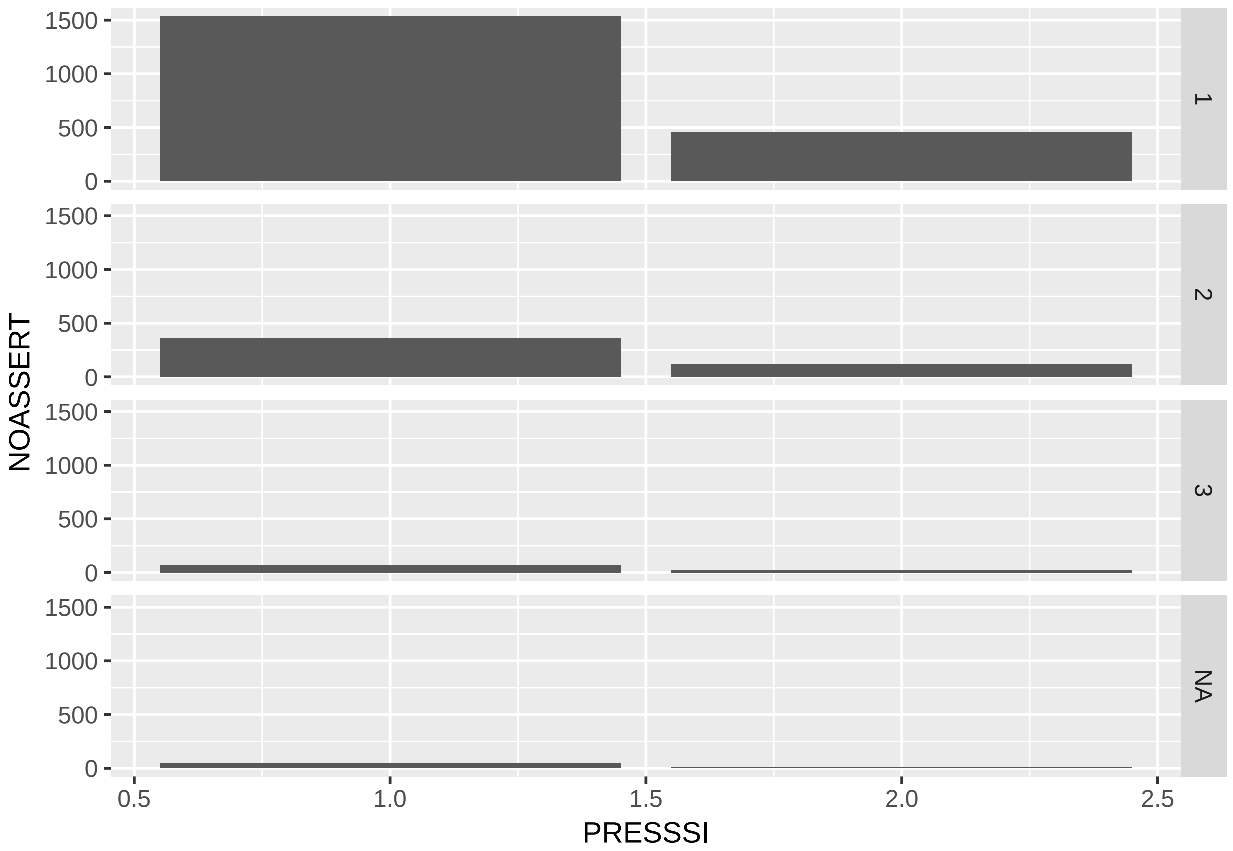
```
#load tidyr
library("tidyr")

#load full dataset
data <- read_por("~/Projects/sexual_violence_college/ICPSR_03212/DS0001/03212-0001-Data.por")


#create data_condensed
data_condensed <- data %>%
  dplyr::select(RACE, NOASSERT, PLEASE,  AUTHSI, PRESSSI)

#convert columns to numeric
data_condensed$RACE <- as.numeric(data_condensed$RACE)

data_condensed$NOASSERT<- as.numeric(data_condensed$NOASSERT)

data_condensed$PLEASE <- as.numeric(data_condensed$PLEASE)

data_condensed$AUTHSI <- as.numeric(data_condensed$AUTHSI)

data_condensed$PRESSSI <- as.numeric(data_condensed$PRESSSI)


#load ggplot2 package
library("ggplot2")


#create bar chart for AUTHSI facetted by RACE
#for real presentation I would want to figure out how to change the labels for the facets

ggplot(data_condensed, aes(AUTHSI, PLEASE)) +
  geom_bar(mapping = NULL, data = NULL, stat = "identity", position = "stack", width = NULL,     binwid
```

```
## Warning: Removed 12 rows containing missing values (position_stack).
```

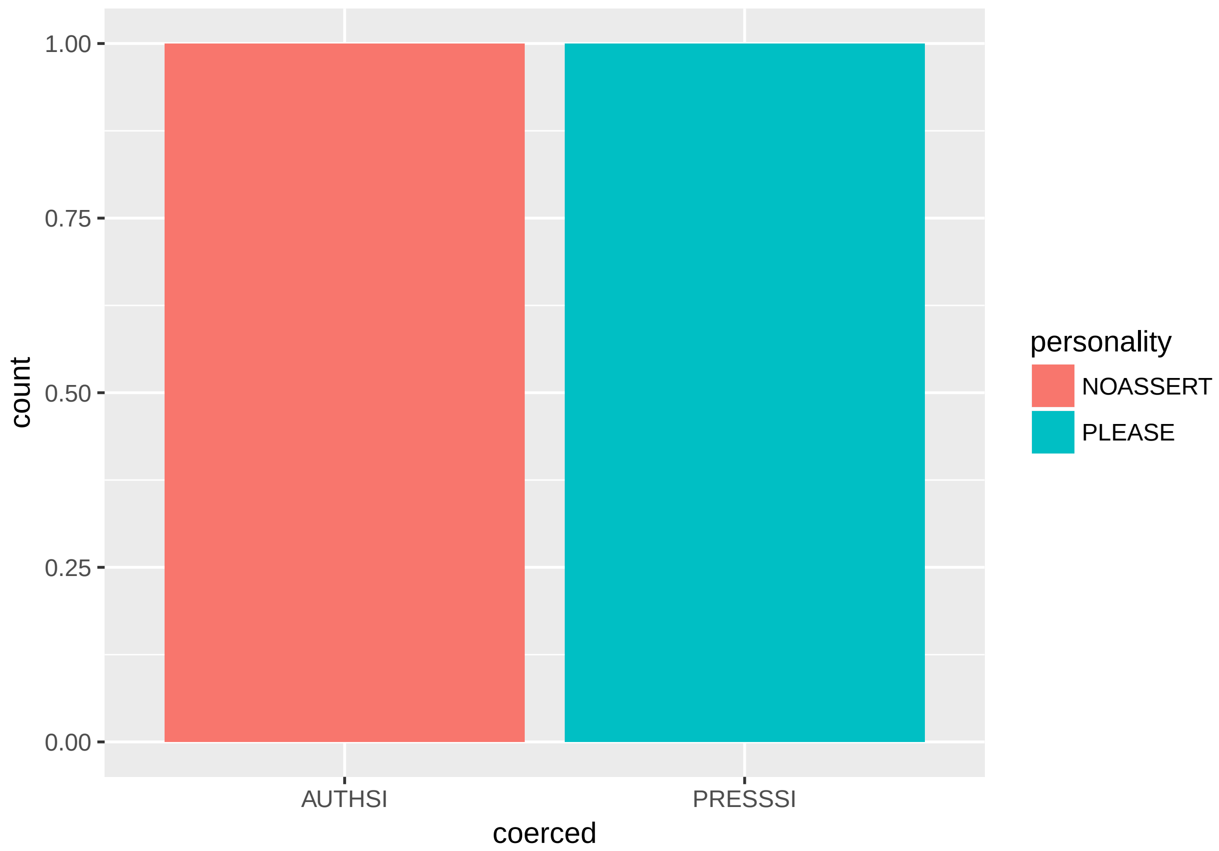


```r
#create bar chart for PRESSSI facetted by RACE

ggplot(data_condensed, aes(PRESSSI, NOASSERT)) + geom_bar(mapping = NULL, data = NULL, stat = "identity
  show.legend = NA, inherit.aes = TRUE) + facet_grid(RACE ~ .)
```

```
## Warning: Removed 13 rows containing missing values (position_stack).
```



```r
#create side by side bar chart for AUTHSI and PRESSSI for their occurences of NOASSERT and PLEASE
coerced <- c(rep("AUTHSI"),rep("PRESSSI"))
personality <- c("NOASSERT","PLEASE")

df_bar <- data.frame(coerced, personality)

ggplot(df_bar, aes(coerced)) + geom_bar(aes(fill = personality), position = "dodge")
```

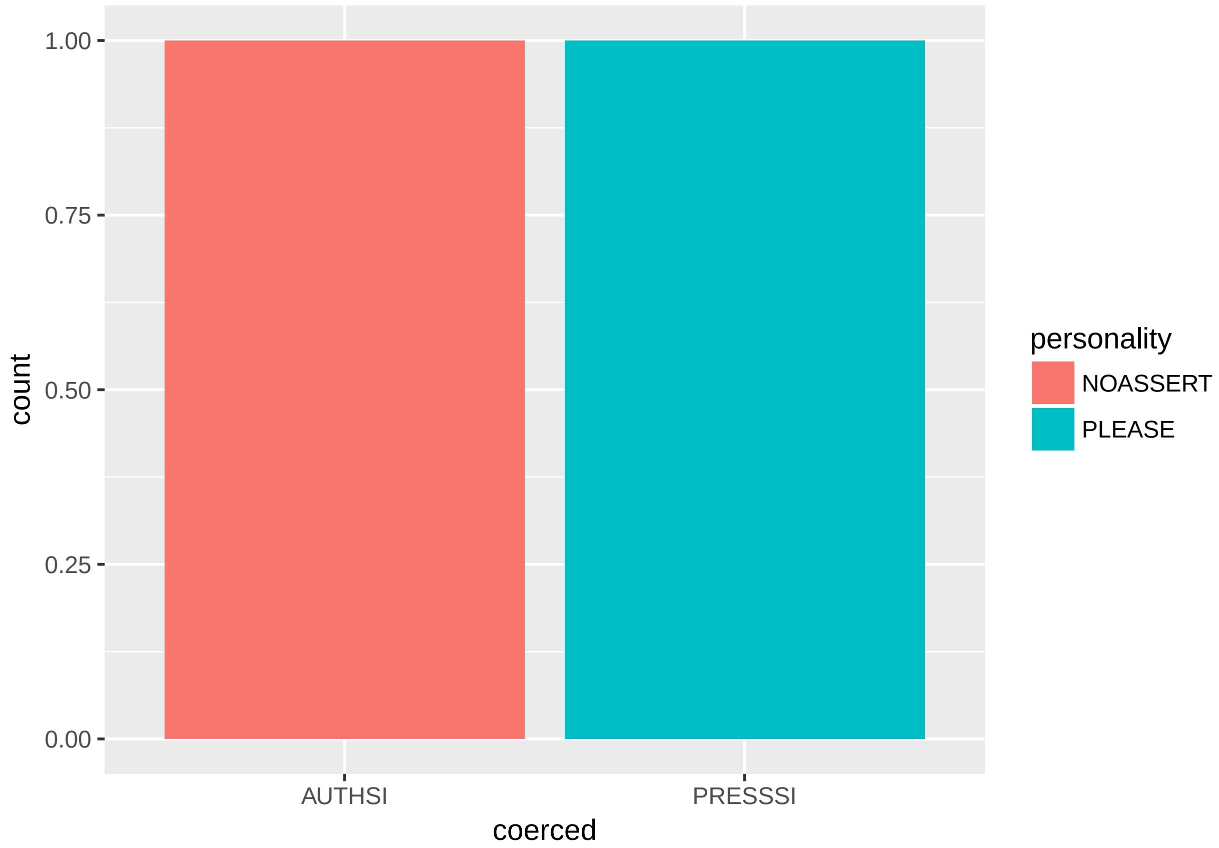


```
#ggplot(data.with.P, aes(personality)) + geom_bar(aes(fill = AUTHSI), position = "dodge")
```

4. Create at least three variations of the plots you've already made by modifying some of the arguments we covered in class (i.e. `position`, `scale`, `size`, `linetype` etc.). Do any of these modifications help you understand your data better? Why or why not? Do any of them create a misleading interpretation of the relationships between your variables? If yes, how so?

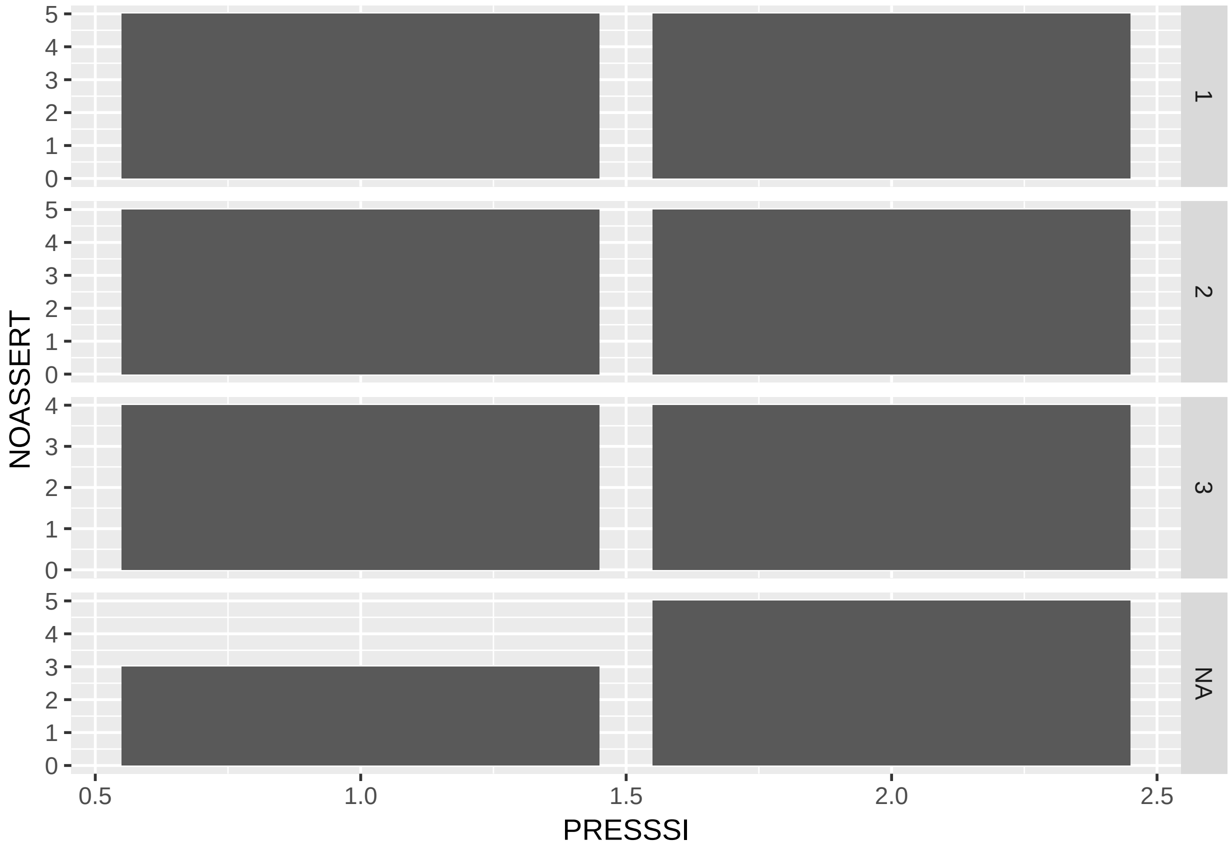
Only the first variation was actually helpful in understanding the data. I think by stacking the bars instead of placing them side by side it is easier to visualize the comparison between the variables. The second variation makes it very hard to understand the data and because of the scale and colors all of the bar charts look to similar to get a meaningful analysis from the representation. The third variation just changes the color of the bars but it demonstrates how important it is in visual representations of data to format the visual in a way that is easy to understand right away. With the last variation the colors became very similar and it takes longer to see the comparison that is supposed to be represented by the bar chart.

```
#variation 1 changing position type
ggplot(df_bar, aes(coerced)) + geom_bar(aes(fill = personality), position = "fill")
```

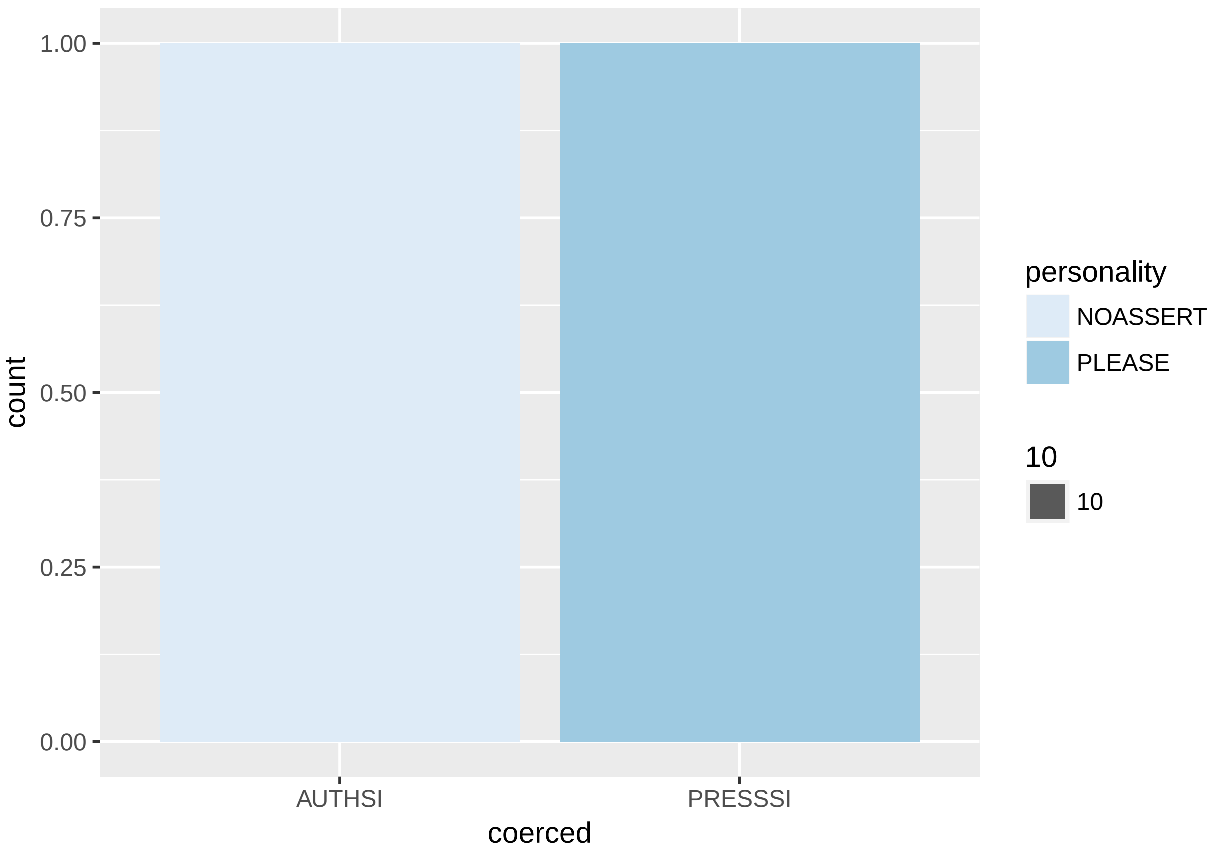


```
#variation 2 changing position type and scale for facet
ggplot(data_condensed, aes(PRESSSI, NOASSERT)) + geom_bar(mapping = NULL, data = NULL, stat = "identity
  show.legend = NA, inherit.aes = TRUE) + facet_grid(RACE ~ ., scales = "free")
```

```
## Warning: Removed 13 rows containing missing values (geom_bar).
```



```
#variation 3 changing scale of fill
ggplot(df_bar, aes(coerced, size = 10)) + geom_bar(aes(fill = personality), position = "fill") + scale_
```



5. From the plots you've created thus far, do any of them seem appropriate for a general audience? Why or why not? If so, what do you think you'd still need to do to make them more suitable as explanatory visualizations?

The stacked bar chart is the closest represenation to what I would actually want from an audience. However, I would like to organize it in a more meaningful way by placing PLEASE and NOASSERT together on the x-axis and seeing how they relate to AUTHSI and PRESSSI in separate bar charts. This will mean I need to combine PLEASE and NOASSERT into a new column named PERSONALITY, which I am currently working on.