Oxford-Comma-Tidyverse-Part-1

Michael Y.

due 12/08/2019

Contents

TidyVerse assignment, due on 12/08 4

You have two tasks: 4

1. Create an Example. 4

Oxford Comma dataset 4

Variables in the dataset 5

`readr::read_csv`: Initial attempt to load up the data from fivethirtyeight's github site: 8

Without my having made any column specifications, `read_csv` hasn't read in the data as we would like: 9

`readr::spec_csv()` can specify the type of each column 12

Check result with `readr::spec_csv` 12

`readr::read_csv`: Re-load the data, this time specifying the above column types: 14

The initial variable names were awful. 16

`dplyr::bind_cols()` as tidyverse equivalent to cbind() : 16

`dplyr::rename_all`: rename all of the variable names:: 17

Here are the first five records: 19

Here are the same, transposed for display: 20

Manipulate Data 23

Recategorize variable[2] `USES_Oxford` as "True" or "False" 23

`forcats::fct_count()`: Count the number of responses of each type 23

`forcats::fct_recode`: Recode levels as (F,T) 23

Plot result using ggplot2 23

Resequence the (unordered) levels for [4] `CARE_Oxford` to reflect the semantic ordering: 25

- `forcats::fct_count()`: Count the number of responses of each type 25
- `fct_relevel`: Resequence the levels for the factor 25
- Plot result using ggplot2 26

Recategorize the responses to [5] `DATA_Sentence` to reflect “PLURAL” or “SINGULAR” 28

- `forcats::fct_count()`: Count the number of responses of each type 28
- Replace the above sentences with the word “SINGULAR” or “PLURAL” to reflect user preference 28
- `forcats::fct_recode`: Recode levels as “SINGULAR” or “PLURAL” 28
- Plot result using ggplot2 28

Resequence the (unordered) levels for [7] `DATA_Care` to reflect the semantic ordering: 30

- `forcats::fct_count()`: Count the number of responses of each type 30
- `fct_relevel`: Resequence the levels for the factor `DATA_Care` 30
- Plot result using ggplot2 31

Resequence the (unordered) levels for [8] `Grammar_Important` to reflect the semantic ordering 33

- `forcats::fct_count()`: Count the number of responses of each type 33
- `forcats::fct_recode`: Recode level "“Neither important nor unimportant (neutral)” as “NEUTRAL” because it is too long 34
- `fct_relevel`: Resequence the (unordered) levels for [8] `Grammar_Important` to reflect the ordering: 34
- Plot result using ggplot2 35

Resequence the (unordered) levels for [10] `AgeBands` to reflect the ordering of the bands: 37

- `forcats::fct_count()`: Count the number of responses of each type 37
- Plot the Age of participants using ggplot2 38

Resequence the (unordered) levels for [11] `IncomeBands` to reflect the ordering of Income, from lowest to highest: 39

- `forcats::fct_count()`: Count the number of responses of each type 39
- Plot the Income of participants using ggplot2 40

Resequence the (unordered) levels for [12] `Education` to reflect the ordering of Educational Attainment: 42

- `forcats::fct_count()`: Count the number of responses of each type 42

Plot the Education Level of participants using ggplot2 43
Resequence the (unordered) levels for [13] `Location` to reflect to reflect geography (east coast to west coast): 44
`forcats::fct_count()`: Count the number of responses of each type 44
Plot the Location of respondents using ggplot2 45
[1] `RespondentID` should not impact the results – it is just an identifier, so drop it 47
`dplyr::select(-[columnname])`: Drop variable [1] `RespondentID` 47
References 48
End of Part 1 **48**

TidyVerse assignment, due on 12/08

In this assignment, you'll practice collaborating around a code project with GitHub.
You could consider our collective work as building out a book of examples on how to use TidyVerse functions.

GitHub repository: https://github.com/acatlin/FALL2019TIDYVERSE

FiveThirtyEight.com datasets .

Kaggle datasets .

You have two tasks:

1. Create an Example.

Using one or more TidyVerse packages, and any dataset from fivethirtyeight.com or Kaggle, create a programming sample "vignette" that demonstrates how to use one or more of the capabilities of the selected TidyVerse package with your selected dataset. **(25 points)**

Oxford Comma dataset

In June of 2014, `FiveThirtyEight.com` ran an online poll (using "surveymonkey.com") asking Americans whether they preferred the serial comma (also known as the `Oxford Comma`.)

Additional questions were posed regarding the respondents' educational level, income level, age, and what part of the country each person was from.

Additional grammatical questions which were part of the same poll concerned usage of the word "data": respondents were asked whether they considered "data" to be *singular* or *plural*.

Following conclusion of the poll, FiveThirtyEight.com published a piece ***Elitist, Superfluous, Or Popular? We Polled Americans on the Oxford Comma***[1] and made the **underlying dataset**[2] available on github .

Variables in the dataset

The raw dataset contains 1129 cases, each of which represents a response to an online poll conducted in June 2014, where participants were asked various questions, including:

1) whether they knew what the Oxford Comma is,
2) which of two sentences (one with the serial comma, and one without) they preferred, and
3) whether they believed the use of proper grammar was important.

Additionally, participants were asked questions regarding their gender, age, income, educational attainment, and geographic region.

The overall dataset includes the following variables and possible responses:

n	variable	question or description	type	data dictionary
1	`RespondentID`	numerical ID of participant	numerical, discrete	unique identifiers assigned by the survey site (surveymonkey.com)
2	`USES_Oxford`	"In your opinion, which sentence is more gramatically correct?"	categorical, nominal	1-"It's important for a person to be honest, kind and loyal." 2-"It's important for a person to be honest kind and loyal."
3	`HEARD_Oxford`	"Prior to reading about it above, had you heard of the serial (or Oxford) comma?"	categorical, binary	"No", "Yes"
4	`CARE_Oxford`	"How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?"	categorical, ordinal	"Not at All" , "Not Much" , "Some" , "A lot"
5	`DATA_Sentence`	"How would you write the following sentence?" (One uses "Data" as singular, the other as plural)	categorical, binary	Plural: "Some experts say it's important to drink milk, but the data are inconclusive." ; Singular: "Some experts say it's important to drink milk, but the data is inconclusive."

n	variable	question or description	type	data dictionary
6	DATA_Plural	“When faced with using the word ‘data’, have you ever spent time considering if the word was a singular or plural noun?”	categorical, binary	“Yes”, “No”
7	DATA_Care	“How much, if at all, do you care about the debate over the use of the word ‘data’ as a singluar or plural noun?”	categorical, ordinal	“Not at All” , “Not Much” , “Some” , “A lot”
8	Grammar_Important	“In your opinion, how important or unimportant is proper use of grammar?”	Categorical, ordinal	“Very unimportant”, “Somewhat unimportant”, “Neither important nor unimportant (neutral)”, “Somewhat important”, “Very important”
9	Gender	Participant’s gender (only “Male” and “Female” choices offered)	Categorical, binary	“Female”, “Male”
10	AgeBands	Participant’s age, in one of four bands	Categorical, ordinal	“18-29”, “30-44”, “45-60”, “> 60”
11	IncomeBands	Participant’s household income, in one of five bands	Categorical, ordinal	“$0 -$24,999” , “$25,000-$49,999” , “$50,000-$99,999” , “$100,000-$149,999” , “$150,000+”
12	Education	Participant’s level of education, in one of five categories	Categorical, ordinal	“Less than high school degree”, “High school degree”, Some college or Associate degree“,”Bachelor degree“,”Graduate degree"

n	variable	question or description	type	data dictionary
13	`Location`	Participant’s geographic location, in one of 9 regions	Categorical, nominal	“New England”,“Middle Atlantic”,“South Atlantic”," East North Central“,”East South Central“,”West North Central“,”West South Central“,”Mountain“,”Pacific"

So, other than the initial column, all the remaining columns are factors.

`readr::read_csv`: Initial attempt to load up the data from fivethirtyeight's github site:

```
commadataURL <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/comma-survey/comma-survey.csv"
### Read the data using read_csv from tidyverse package readr
tv_commadata <- read_csv(commadataURL)
```

```
## Parsed with column specification:
## cols(
##   RespondentID = col_double(),
##   `In your opinion, which sentence is more gramatically correct?` = col_character(),
##   `Prior to reading about it above, had you heard of the serial (or Oxford) comma?` = col_character(),
##   `How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?` = col_character(),
##   `How would you write the following sentence?` = col_character(),
##   `When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun?` = col_charac
##   `How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?` = col_character(),
##   `In your opinion, how important or unimportant is proper use of grammar?` = col_character(),
##   Gender = col_character(),
##   Age = col_character(),
##   `Household Income` = col_character(),
##   Education = col_character(),
##   `Location (Census Region)` = col_character()
## )
```

Without my having made any column specifications, `read_csv` hasn't read in the data as we would like:

```
# First 5 cases
tv_commadata %>% head(5)
```

```
## # A tibble: 5 x 13
##   RespondentID `In your opinio~ `Prior to readi~ `How much, if a~ `How would you ~ `When faced wit~ `How much, if a~ `In your opinio~ G
##          <dbl> <chr>            <chr>            <chr>            <chr>            <chr>            <chr>            <chr>            <
## 1   3292953864 It's important ~ Yes              Some             Some experts sa~ No               Not much         Somewhat import~ M
## 2   3292950324 It's important ~ No               Not much         Some experts sa~ No               Not much         Somewhat unimpo~ M
## 3   3292942669 It's important ~ Yes              Some             Some experts sa~ Yes              Not at all       Very important   M
## 4   3292932796 It's important ~ Yes              Some             Some experts sa~ No               Some             Somewhat import~ M
## 5   3292932522 It's important ~ No               Not much         Some experts sa~ No               Not much         <NA>             <
## # ... with 4 more variables: Age <chr>, `Household Income` <chr>, Education <chr>, `Location (Census Region)` <chr>
```

```
# First 5 cases, transposed:
tv_commadata %>% head(5) %>% t
```

```
##                                                                                                                         [,1]
## RespondentID                                                                                                            "3292953864"
## In your opinion, which sentence is more gramatically correct?                                                           "It's important
## Prior to reading about it above, had you heard of the serial (or Oxford) comma?                                         "Yes"
## How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?           "Some"
## How would you write the following sentence?                                                                             "Some experts sa
## When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun? "No"
## How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?        "Not much"
## In your opinion, how important or unimportant is proper use of grammar?                                                 "Somewhat import
## Gender                                                                                                                  "Male"
## Age                                                                                                                     "30-44"
## Household Income                                                                                                        "$50,000 - $99,9
## Education                                                                                                               "Bachelor degree
## Location (Census Region)                                                                                                "South Atlantic"
##                                                                                                                         [,2]
## RespondentID                                                                                                            "3292950324"
## In your opinion, which sentence is more gramatically correct?                                                           "It's important
## Prior to reading about it above, had you heard of the serial (or Oxford) comma?                                         "No"
## How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?           "Not much"
## How would you write the following sentence?                                                                             "Some experts sa
## When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun? "No"
## How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?        "Not much"
## In your opinion, how important or unimportant is proper use of grammar?                                                "Somewhat unimpo
## Gender                                                                                                                 "Male"
## Age                                                                                                                    "30-44"
## Household Income                                                                                                       "$50,000 - $99,9
## Education                                                                                                              "Graduate degree
## Location (Census Region)                                                                                               "Mountain"
##                                                                                                                        [,3]
## RespondentID                                                                                                           "3292942669"
## In your opinion, which sentence is more gramatically correct?                                                          "It's important 
## Prior to reading about it above, had you heard of the serial (or Oxford) comma?                                        "Yes"
## How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?           "Some"
## How would you write the following sentence?                                                                            "Some experts sa
## When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun? "Yes"
## How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?        "Not at all"
## In your opinion, how important or unimportant is proper use of grammar?                                                "Very important"
## Gender                                                                                                                 "Male"
## Age                                                                                                                    "30-44"
## Household Income                                                                                                       NA
## Education                                                                                                              NA
## Location (Census Region)                                                                                               "East North Cent
##                                                                                                                        [,4]
## RespondentID                                                                                                           "3292932796"
## In your opinion, which sentence is more gramatically correct?                                                          "It's important 
## Prior to reading about it above, had you heard of the serial (or Oxford) comma?                                        "Yes"
## How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?           "Some"
## How would you write the following sentence?                                                                            "Some experts sa
## When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun? "No"
## How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?        "Some"
## In your opinion, how important or unimportant is proper use of grammar?                                                "Somewhat import
## Gender                                                                                                                 "Male"
## Age                                                                                                                    "18-29"
## Household Income                                                                                                       NA
## Education                                                                                                              "Less than high 
## Location (Census Region)                                                                                               "Middle Atlantic
##                                                                                                                        [,5]
## RespondentID                                                                                                           "3292932522"
## In your opinion, which sentence is more gramatically correct?                                                          "It's important 
## Prior to reading about it above, had you heard of the serial (or Oxford) comma?                                        "No"
## How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?           "Not much"
## How would you write the following sentence?                                                                            "Some experts sa
## When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun? "No"
## How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?        "Not much"
## In your opinion, how important or unimportant is proper use of grammar?                                                NA
## Gender                                                                                                                 NA
## Age                                                                                                                    NA
## Household Income                                                                                                       NA
## Education                                                                                                              NA
## Location (Census Region)                                                                                               NA
```

```
tv_commadata %>%
  head(5) %>%
  t %>%
  kable() %>%
  kable_styling(c("striped", "bordered"))
```

RespondentID	3292953864
In your opinion, which sentence is more gramatically correct?	It's important for a person to be honest,
Prior to reading about it above, had you heard of the serial (or Oxford) comma?	Yes
How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?	Some
How would you write the following sentence?	Some experts say it's important to drink
When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun?	No
How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?	Not much
In your opinion, how important or unimportant is proper use of grammar?	Somewhat important
Gender	Male
Age	30-44
Household Income	$50,000 - $99,999
Education	Bachelor degree
Location (Census Region)	South Atlantic

```
tv_commadata %>% summary %>%
  kable() %>%
  kable_styling(c("striped", "bordered"))
```

	RespondentID	In your opinion, which sentence is more gramatically correct?	Prior to reading about it above, had you heard of the serial (or Oxford) c
	Min. :3288375700	Length:1129	Length:1129
	1st Qu.:3289469695	Class :character	Class :character
	Median :3290113576	Mode :character	Mode :character
	Mean :3290127075	NA	NA
	3rd Qu.:3290776606	NA	NA
	Max. :3292953864	NA	NA

It read in the first column as a double, and all of the remaining columns as characters.

In reality, we want each of the remaining columns to be a **factor**, with only a few possible responses.

`readr::spec_csv()` can specify the type of each column

```
tv_columnspec = paste0(c("n",rep("f",12)),collapse="")
tv_columnspec
```

```
## [1] "nffffffffffff"
```

This indicates that the first column is numeric, and each of the remaining 12 columns is a factor.

Check result with `readr::spec_csv`

```
spec_csv(commadataURL, col_types=tv_columnspec)
```

```
## cols(
##   RespondentID = col_number(),
##   `In your opinion, which sentence is more gramatically correct?` = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   `Prior to reading about it above, had you heard of the serial (or Oxford) comma?` = col_factor(levels = NULL, ordered = FALSE, incluc
##   `How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?` = col_factor(levels =
##   `How would you write the following sentence?` = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   `When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun?` = col_factor
##   `How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?` = col_factor(levels
##   `In your opinion, how important or unimportant is proper use of grammar?` = col_factor(levels = NULL, ordered = FALSE, include_na = F
##   Gender = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   Age = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   `Household Income` = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   Education = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE),
##   `Location (Census Region)` = col_factor(levels = NULL, ordered = FALSE, include_na = FALSE)
## )
```

readr::read_csv: Re-load the data, this time specifying the above column types:

```
tv_commadata <- read_csv(commadataURL, col_types = tv_columnspec)
tv_commadata %>% summary
```

```
##   RespondentID                                   In your opinion, which sentence is more gramatically correct?
##  Min.   :3288375700   It's important for a person to be honest, kind and loyal. :488
##  1st Qu.:3289469695   It's important for a person to be honest, kind, and loyal.:641
##  Median :3290113576
##  Mean   :3290127075
##  3rd Qu.:3290776606
##  Max.   :3292953864
##
##  Prior to reading about it above, had you heard of the serial (or Oxford) comma?
##  Yes :655
##  No  :444
##  NA's: 30
##
##
##
##
##  How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?
##  Some      :414
##  Not much  :268
##  A lot     :291
##  Not at all:126
##  NA's      : 30
##
##
##                                                        How would you write the following sentence?
##  Some experts say it's important to drink milk, but the data is inconclusive. :865
##  Some experts say it's important to drink milk, but the data are inconclusive.:228
##  NA's                                                                         : 36
##
##
##
##
##  When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun?
##  No  :547
##  Yes :544
##  NA's: 38
##
##
##
##
##  How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?
##  Not much  :403
##  Not at all:203
##  Some      :352
##  A lot     :133
##  NA's      : 38
##
##
##         In your opinion, how important or unimportant is proper use of grammar?    Gender       Age                 Household Income
##  Somewhat important                         :333                                Male  :489   30-44:254   $50,000 - $99,999  :290
##  Somewhat unimportant                       :  7                                Female:548   18-29:221   $25,000 - $49,999  :158
##  Very important                             :688                                NA's  : 92   > 60 :272   $0 - $24,999       :121
##  Very unimportant                           :  5                                             45-60:290   $150,000+          :103
##  Neither important nor unimportant (neutral): 26                                             NA's : 92   $100,000 - $149,999:164
##  NA's                                       : 70                                                         NA's               :293
##
##                              Education         Location (Census Region)
##  Bachelor degree                 :344   Pacific           :180
##  Graduate degree                 :276   East North Central:170
##  Less than high school degree    : 11   South Atlantic    :164
##  Some college or Associate degree:295   Middle Atlantic   :140
##  High school degree              :100   West South Central: 88
##  NA's                            :103   (Other)           :285
##                                         NA's              :102
```

Now the data are all recognized as (unordered) factors.

The initial variable names were awful.

dplyr::bind_cols() as tidyverse equivalent to cbind() :

```
initial_variable_names <- names(tv_commadata)

# index the variable names using bind_cols from dplyr
bind_cols(column=seq(initial_variable_names),
          InitialVariableName=initial_variable_names) %>%
  kable() %>%
  kable_styling(c("striped", "bordered"))
```

column	InitialVariableName
1	RespondentID
2	In your opinion, which sentence is more gramatically correct?
3	Prior to reading about it above, had you heard of the serial (or Oxford) comma?
4	How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?
5	How would you write the following sentence?
6	When faced with using the word "data", have you ever spent time considering if the word was a singular or plural noun?
7	How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?
8	In your opinion, how important or unimportant is proper use of grammar?
9	Gender
10	Age
11	Household Income
12	Education
13	Location (Census Region)

Because these “variables” are so long, it is difficult to display the information.
I’ll replace each column header with a succinct name, while saving the above questions in an array.

dplyr::rename_all: rename all of the variable names::

```
new_variable_names <- c("RespondentID",
                        "USES_Oxford",
                        "HEARD_Oxford",
                        "CARE_Oxford",
                        "DATA_Sentence",
                        "DATA_Plural",
                        "DATA_Care",
                        "Grammar_Important",
                        "Gender",
                        "AgeBands",
                        "IncomeBands",
                        "Education",
                        "Location")

tv_commadata <- tv_commadata %>% rename_all( function(.){new_variable_names} )
tv_commadata %>% summary
```

```
##   RespondentID                                                            USES_Oxford  HEARD_Oxford     CARE_Oxford
##  Min.   :3288375700   It's important for a person to be honest, kind and loyal. :488   Yes :655     Some      :414
##  1st Qu.:3289469695   It's important for a person to be honest, kind, and loyal.:641   No  :444     Not much  :268
##  Median :3290113576                                                                    NA's: 30     A lot     :291
##  Mean   :3290127075                                                                                 Not at all:126
##  3rd Qu.:3290776606                                                                                 NA's      : 30
##  Max.   :3292953864
##
##                                                                         DATA_Sentence DATA_Plural      DATA_Care
##  Some experts say it's important to drink milk, but the data is inconclusive. :865   No  :547    Not much  :403
##  Some experts say it's important to drink milk, but the data are inconclusive.:228   Yes :544    Not at all:203
##  NA's                                                                         : 36   NA's: 38    Some      :352
##                                                                                                  A lot     :133
##                                                                                                  NA's      : 38
##
##
##                                    Grammar_Important    Gender     AgeBands                IncomeBands
##  Somewhat important                         :333     Male  :489   30-44:254   $50,000 - $99,999  :290
##  Somewhat unimportant                       :  7     Female:548   18-29:221   $25,000 - $49,999  :158
##  Very important                             :688     NA's  : 92   > 60 :272   $0 - $24,999       :121
##  Very unimportant                           :  5                  45-60:290   $150,000+          :103
##  Neither important nor unimportant (neutral): 26                  NA's : 92   $100,000 - $149,999:164
##  NA's                                       : 70                              NA's               :293
##
##                            Education                Location
##  Bachelor degree                 :344   Pacific           :180
##  Graduate degree                 :276   East North Central:170
##  Less than high school degree    : 11   South Atlantic    :164
##  Some college or Associate degree:295   Middle Atlantic   :140
##  High school degree              :100   West South Central: 88
##  NA's                            :103   (Other)           :285
##                                         NA's              :102
```

Here are the first five records:

```
tv_commadata %>%
  head(5)
```

```
## # A tibble: 5 x 13
##   RespondentID USES_Oxford HEARD_Oxford CARE_Oxford DATA_Sentence DATA_Plural DATA_Care Grammar_Importa~ Gender AgeBands IncomeBands
##          <dbl> <fct>       <fct>        <fct>       <fct>         <fct>       <fct>     <fct>            <fct>  <fct>    <fct>
## 1   3292953864 It's impor~ Yes          Some        Some experts~ No          Not much  Somewhat import~ Male   30-44    $50,000 - ~
## 2   3292950324 It's impor~ No           Not much    Some experts~ No          Not much  Somewhat unimpo~ Male   30-44    $50,000 - ~
## 3   3292942669 It's impor~ Yes          Some        Some experts~ Yes         Not at a~ Very important   Male   30-44    <NA>
## 4   3292932796 It's impor~ Yes          Some        Some experts~ No          Some      Somewhat import~ Male   18-29    <NA>
## 5   3292932522 It's impor~ No           Not much    Some experts~ No          Not much  <NA>             <NA>   <NA>     <NA>
## # ... with 2 more variables: Education <fct>, Location <fct>
```

```
# using kable:
tv_commadata %>%
  head(5) %>%
  kable() %>%
  kable_styling(c("striped", "bordered"))
```

RespondentID	USES_Oxford	HEARD_Oxford	CARE_Oxford	DATA_Sentence
3292953864	It’s important for a person to be honest, kind and loyal.	Yes	Some	Some experts say it’s important to drink milk, bu
3292950324	It’s important for a person to be honest, kind, and loyal.	No	Not much	Some experts say it’s important to drink milk, bu
3292942669	It’s important for a person to be honest, kind, and loyal.	Yes	Some	Some experts say it’s important to drink milk, bu
3292932796	It’s important for a person to be honest, kind, and loyal.	Yes	Some	Some experts say it’s important to drink milk, bu
3292932522	It’s important for a person to be honest, kind and loyal.	No	Not much	Some experts say it’s important to drink milk, bu

Here are the same, transposed for display:

```
tv_commadata %>%
  head(5) %>%
  t
```

```
##                     [,1]
## RespondentID        "3292953864"
## USES_Oxford         "It's important for a person to be honest, kind and loyal."
## HEARD_Oxford        "Yes"
## CARE_Oxford         "Some"
## DATA_Sentence       "Some experts say it's important to drink milk, but the data is inconclusive."
## DATA_Plural         "No"
## DATA_Care           "Not much"
## Grammar_Important "Somewhat important"
## Gender              "Male"
## AgeBands            "30-44"
## IncomeBands         "$50,000 - $99,999"
## Education           "Bachelor degree"
## Location            "South Atlantic"
##                     [,2]
## RespondentID        "3292950324"
## USES_Oxford         "It's important for a person to be honest, kind, and loyal."
## HEARD_Oxford        "No"
## CARE_Oxford         "Not much"
## DATA_Sentence       "Some experts say it's important to drink milk, but the data is inconclusive."
## DATA_Plural         "No"
## DATA_Care           "Not much"
## Grammar_Important "Somewhat unimportant"
## Gender              "Male"
## AgeBands            "30-44"
## IncomeBands         "$50,000 - $99,999"
## Education           "Graduate degree"
## Location            "Mountain"
##                     [,3]
## RespondentID        "3292942669"
## USES_Oxford         "It's important for a person to be honest, kind, and loyal."
## HEARD_Oxford        "Yes"
## CARE_Oxford         "Some"
## DATA_Sentence     "Some experts say it's important to drink milk, but the data is inconclusive."
## DATA_Plural       "Yes"
## DATA_Care         "Not at all"
## Grammar_Important "Very important"
## Gender            "Male"
## AgeBands          "30-44"
## IncomeBands       NA
## Education         NA
## Location          "East North Central"
##                   [,4]
## RespondentID      "3292932796"
## USES_Oxford       "It's important for a person to be honest, kind, and loyal."
## HEARD_Oxford      "Yes"
## CARE_Oxford       "Some"
## DATA_Sentence     "Some experts say it's important to drink milk, but the data is inconclusive."
## DATA_Plural       "No"
## DATA_Care         "Some"
## Grammar_Important "Somewhat important"
## Gender            "Male"
## AgeBands          "18-29"
## IncomeBands       NA
## Education         "Less than high school degree"
## Location          "Middle Atlantic"
##                   [,5]
## RespondentID      "3292932522"
## USES_Oxford       "It's important for a person to be honest, kind and loyal."
## HEARD_Oxford      "No"
## CARE_Oxford       "Not much"
## DATA_Sentence     "Some experts say it's important to drink milk, but the data is inconclusive."
## DATA_Plural       "No"
## DATA_Care         "Not much"
## Grammar_Important NA
## Gender            NA
## AgeBands          NA
## IncomeBands       NA
## Education         NA
## Location          NA
```

```
#using kable
tv_commadata %>%
  head(5) %>%
  t %>%
  kable() %>%
  kable_styling(c("striped", "bordered"))
```

RespondentID	3292953864	3292950324
USES_Oxford	It’s important for a person to be honest, kind and loyal.	It’s important for a person to be honest, kind, and loyal.
HEARD_Oxford	Yes	No
CARE_Oxford	Some	Not much
DATA_Sentence	Some experts say it’s important to drink milk, but the data is inconclusive.	Some experts say it’s important to drink milk, but the data i
DATA_Plural	No	No
DATA_Care	Not much	Not much
Grammar_Important	Somewhat important	Somewhat unimportant
Gender	Male	Male
AgeBands	30-44	30-44
IncomeBands	$50,000 - $99,999	$50,000 - $99,999
Education	Bachelor degree	Graduate degree
Location	South Atlantic	Mountain

Despite changing the column headers, we still can’t easily see all the information because the text of certainresponses is so long.

Manipulate Data

I made various adjustments to the initial data, including:

Recategorize variable[2] `USES_Oxford` as "True" or "False"

`forcats::fct_count()`: Count the number of responses of each type

```
tv_commadata$USES_Oxford %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[2],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

In your opinion, which sentence is more gramatically correct?	Count
It's important for a person to be honest, kind and loyal.	488
It's important for a person to be honest, kind, and loyal.	641

`forcats::fct_recode`: Recode levels as (F,T)

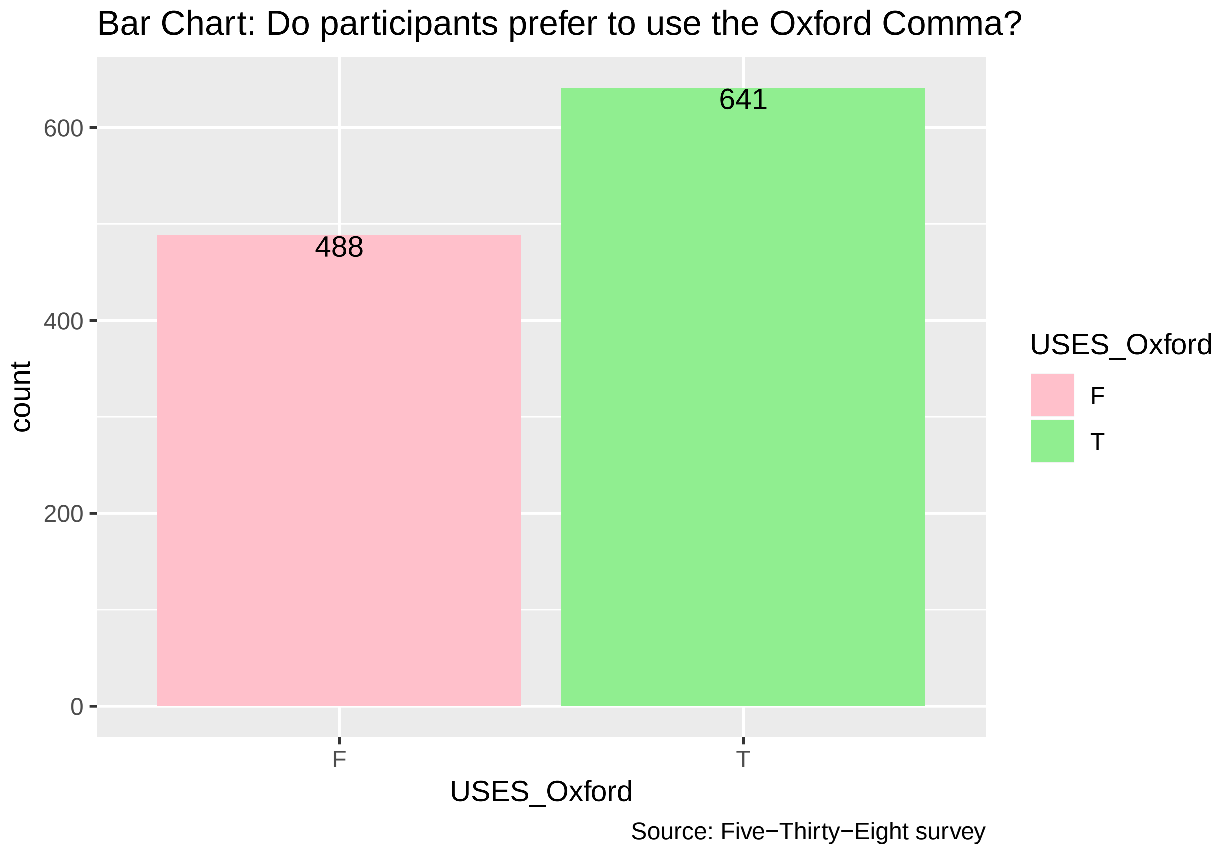
```
oldlevels2 <- tv_commadata$USES_Oxford %>% levels()
tv_commadata$USES_Oxford <- tv_commadata$USES_Oxford %>%
  fct_recode(F=oldlevels2[1],T=oldlevels2[2])

### display results
tv_commadata$USES_Oxford %>% fct_count() %>%
  kable(col.names=c(new_variable_names[2],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

USES_Oxford	Count
F	488
T	641

Plot result using ggplot2

use scale_fill_manual to specify my own color choice

```
myColors <- c("pink","lightgreen")
ggplot(tv_commadata, aes(x=USES_Oxford,fill=USES_Oxford)) +
  geom_bar()+
  scale_fill_manual(values=myColors) +
  geom_text(stat='count', aes(label=..count..), vjust=1) +
  labs(title="Bar Chart: Do participants prefer to use the Oxford Comma?",
       caption="Source: Five-Thirty-Eight survey")
```



Bar Chart: Do participants prefer to use the Oxford Comma?

Source: Five−Thirty−Eight survey

Resequence the (unordered) levels for [4] CARE_Oxford to reflect the semantic ordering:

forcats::fct_count(): Count the number of responses of each type

```
tv_commadata$CARE_Oxford %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[4],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?	Count
Some	414
Not much	268
A lot	291
Not at all	126
NA	30

Note that the sequence in which the responses are listed ("Some","Not Much","A lot","Not at all") does not reflect their semantic ordering. We would like to list the values in the sequence from worst to best, i.e.,

- "Not at all",
- "Not much",
- "Some",
- "A lot" , followed by
- NA.

fct_relevel: Resequence the levels for the factor

to reflect how much does the participant ***care about*** the Oxford Comma

```
### use fct_relevel from library `forcats` to sort the CARE_Oxford levels ordinally
tv_commadata$CARE_Oxford <- tv_commadata$CARE_Oxford %>%
  fct_relevel(levels(tv_commadata$CARE_Oxford)[c(4,2,1,3)])
### display results
tv_commadata$CARE_Oxford %>% fct_count() %>%
  kable(col.names=c(new_variable_names[4],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

CARE_Oxford	Count
Not at all	126
Not much	268
Some	414
A lot	291
NA	30

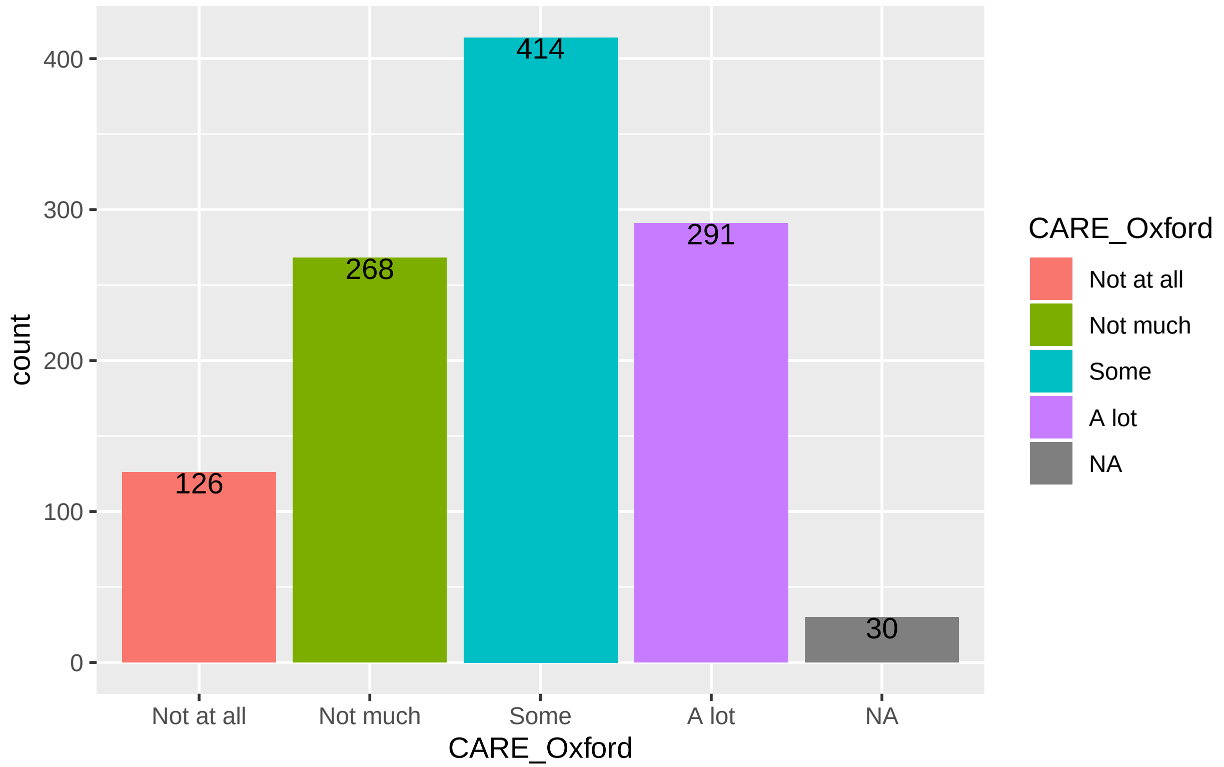
Plot result using ggplot2

Use theme(plot.title = element_text(size = 10))

to prevent long text from overflowing the page

```r
ggplot(tv_commadata, aes(x=CARE_Oxford,fill=CARE_Oxford)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=1) +
  labs(title=initial_variable_names[4],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 10))
```

ıch, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?



Source: Five−Thirty−Eight survey

Recategorize the responses to [5] `DATA_Sentence` to reflect “PLURAL” or “SINGULAR”

`forcats::fct_count()`: Count the number of responses of each type

```
tv_commadata$DATA_Sentence %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[5],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

How would you write the following sentence?	Count
Some experts say it’s important to drink milk, but the data is inconclusive.	865
Some experts say it’s important to drink milk, but the data are inconclusive.	228
NA	36

Replace the above sentences with the word “SINGULAR” or “PLURAL” to reflect user preference

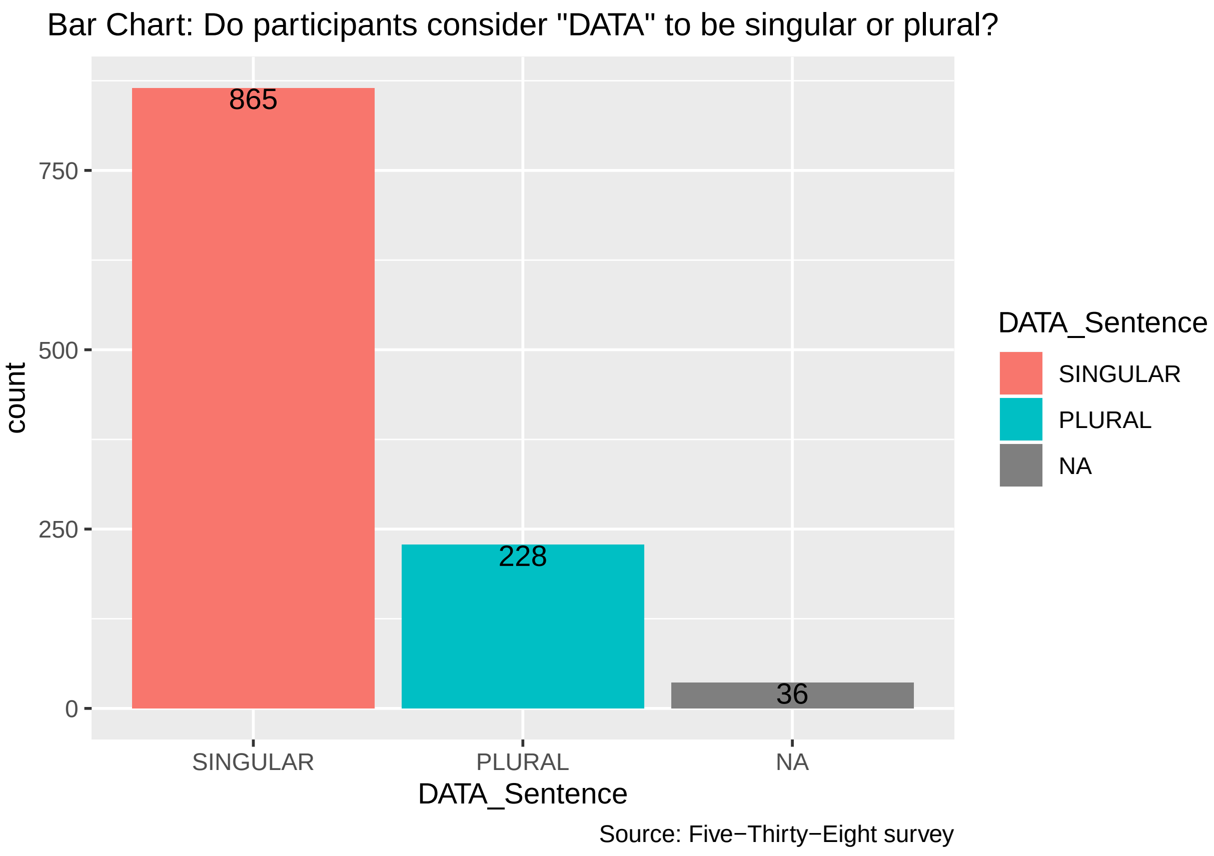
`forcats::fct_recode`: Recode levels as “SINGULAR” or “PLURAL”

```
oldlevels5 <- tv_commadata$DATA_Sentence %>% levels()
tv_commadata$DATA_Sentence <- tv_commadata$DATA_Sentence %>%
  fct_recode("SINGULAR"=oldlevels5[1],"PLURAL"=oldlevels5[2])

### display results
tv_commadata$DATA_Sentence %>% fct_count() %>%
  kable(col.names=c(new_variable_names[5],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

DATA_Sentence	Count
SINGULAR	865
PLURAL	228
NA	36

Plot result using ggplot2

```r
ggplot(tv_commadata, aes(x=DATA_Sentence,fill=DATA_Sentence)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=1) +
  labs(title='Bar Chart: Do participants consider "DATA" to be singular or plural?',
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```



Resequence the (unordered) levels for [7] `DATA_Care` to reflect the semantic ordering:

`forcats::fct_count()`: Count the number of responses of each type

```
tv_commadata$DATA_Care %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[7],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?	Count
Not much	403
Not at all	203
Some	352
A lot	133
NA	38

Note that the sequence in which the responses are listed ("Not Much","Not at all","Some","A lot") does not reflect their semantic ordering. We need to flip the sequence of the first two items, because we would like to list the values in the sequence from worst to best, i.e.,

- "Not at all",
- "Not much",
- "Some",
- "A lot" , followed by
- NA.

`fct_relevel`: Resequence the levels for the factor `DATA_Care`

to reflect how much does the participant ***care about*** care about whether "Data" is considered Singular or Plural

```
### use fct_relevel from library `forcats` to sort the DATA_Care levels ordinally
tv_commadata$DATA_Care <- tv_commadata$DATA_Care %>%
  fct_relevel(levels(tv_commadata$DATA_Care)[c(2,1,3,4)])
### display results
tv_commadata$DATA_Care %>% fct_count() %>%
  kable(col.names=c(new_variable_names[7],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

DATA_Care	Count
Not at all	203
Not much	403
Some	352
A lot	133
NA	38

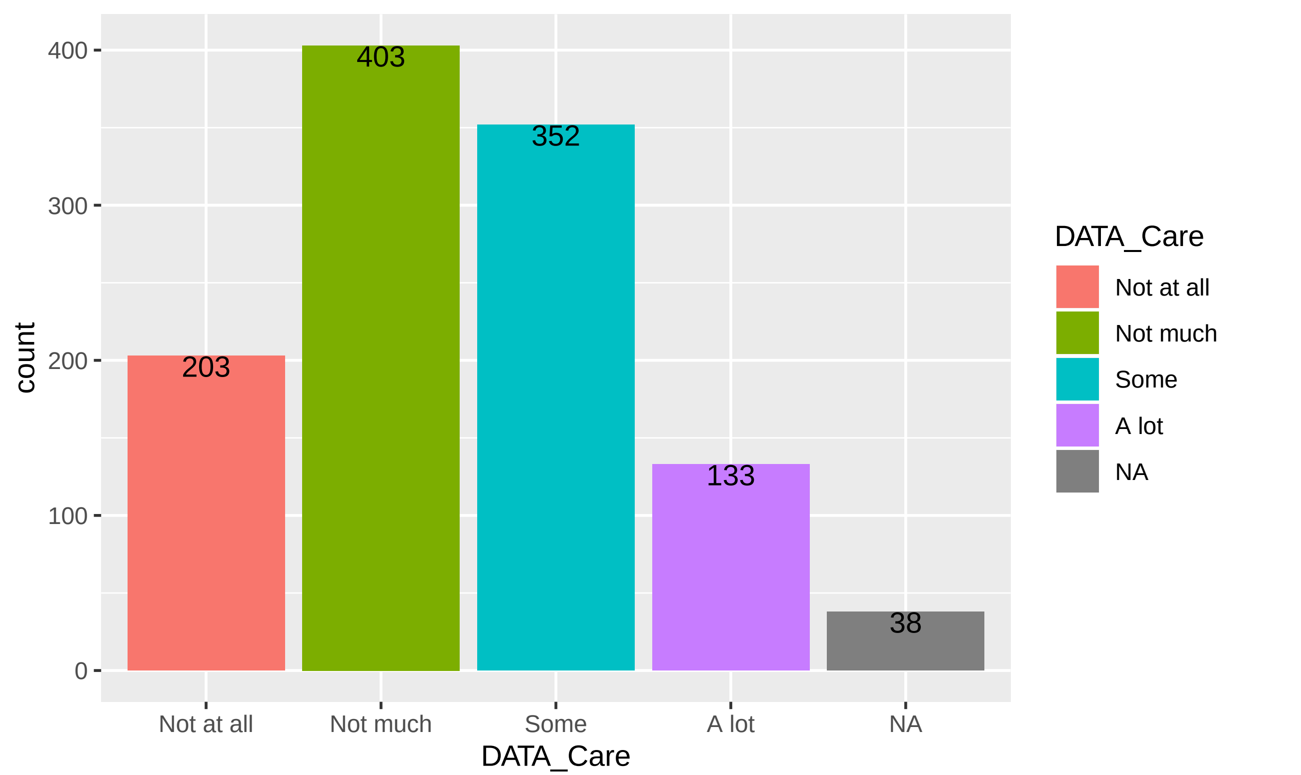
Plot result using ggplot2

Use theme(plot.title = element_text(size = 10))

to prevent long text from overflowing the page

```
ggplot(tv_commadata, aes(x=DATA_Care,fill=DATA_Care)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=1) +
  labs(title=initial_variable_names[7],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 10))
```

uch, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?



Source: Five−Thirty−Eight survey

Resequence the (unordered) levels for [8] Grammar_Important to reflect the semantic ordering

forcats::fct_count(): Count the number of responses of each type

```
tv_commadata$Grammar_Important %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[8],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

In your opinion, how important or unimportant is proper use of grammar?	Count
Somewhat important	333
Somewhat unimportant	7
Very important	688
Very unimportant	5
Neither important nor unimportant (neutral)	26
NA	70

Note that the sequence in which the responses are listed,

- "Somewhat important",
- "Somewhat unimportant",
- "Very important",
- "Very unimportant",
- "Neither important nor unimportant (neutral)"

does not reflect their semantic ordering. We need to reorder the sequence because we would like to list the values in the sequence from worst to best, i.e.,

- "Very unimportant",
- "Somewhat unimportant",
- "Neither important nor unimportant (neutral)"
- "Somewhat important",
- "Very important", followed by
- NA.

forcats::fct_recode: Recode level "“Neither important nor unimportant (neutral)” as “NEUTRAL” because it is too long

```r
oldlevels8 <- tv_commadata$Grammar_Important %>% levels()
tv_commadata$Grammar_Important <- tv_commadata$Grammar_Important %>%
  fct_recode("NEUTRAL"=oldlevels8[5])

### display results
tv_commadata$Grammar_Important %>% fct_count() %>%
  kable(col.names=c(new_variable_names[8],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Grammar_Important	Count
Somewhat important	333
Somewhat unimportant	7
Very important	688
Very unimportant	5
NEUTRAL	26
NA	70

fct_relevel: Resequence the (unordered) levels for [8] Grammar_Important to reflect the ordering:

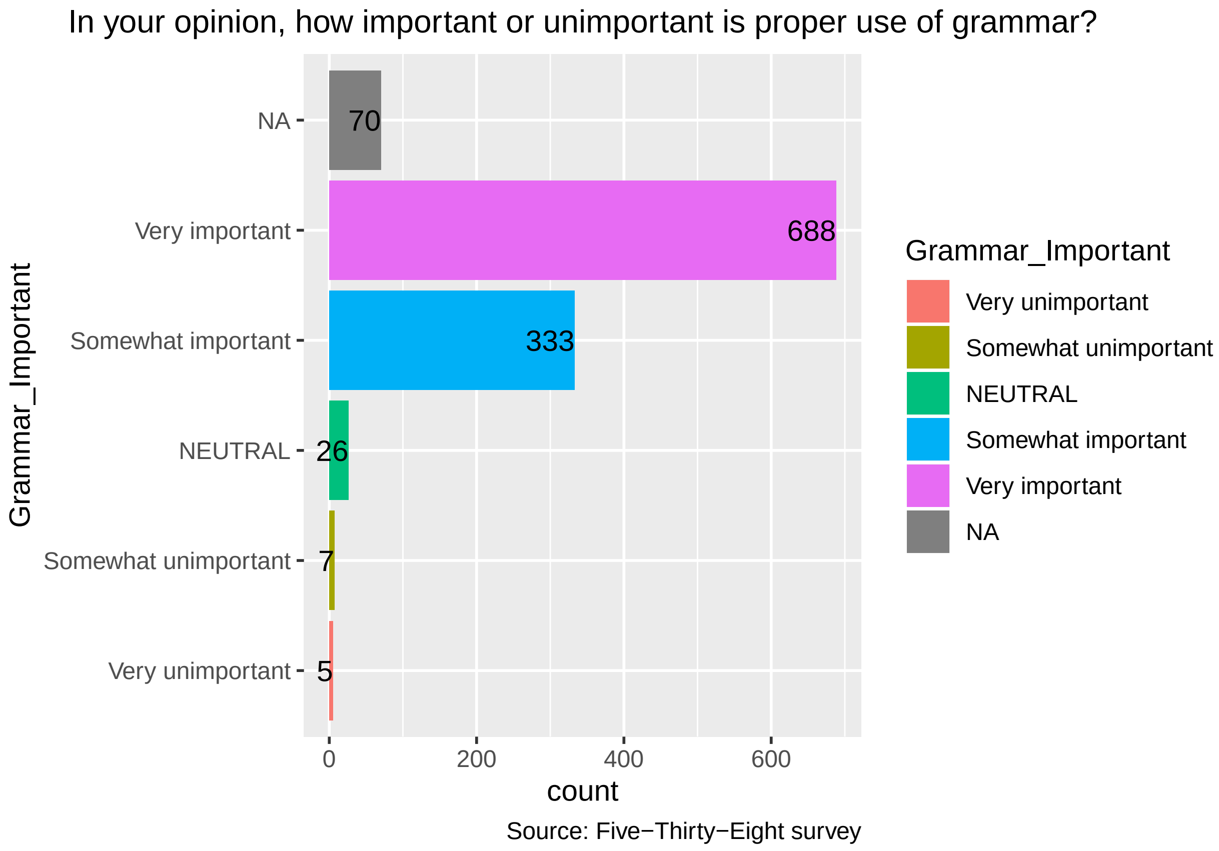
```r
### use fct_relevel from library `forcats` to sort the Grammar_Important levels ordinally
tv_commadata$Grammar_Important <- tv_commadata$Grammar_Important %>%
  fct_relevel(levels(tv_commadata$Grammar_Important)[c(4,2,5,1,3)])
### display results
tv_commadata$Grammar_Important %>% fct_count() %>%
  kable(col.names=c(new_variable_names[8],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Grammar_Important	Count
Very unimportant	5
Somewhat unimportant	7
NEUTRAL	26
Somewhat important	333
Very important	688
NA	70

Plot result using ggplot2

use coord_flip() to make the bars horizontal, to make space for the headings

```r
ggplot(tv_commadata, aes(x=Grammar_Important,fill=Grammar_Important)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), hjust=1) +
  labs(title=initial_variable_names[8],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))+
  coord_flip()
```



Resequence the (unordered) levels for [10] AgeBands to reflect the ordering of the bands:

forcats::fct_count(): Count the number of responses of each type

```
tv_commadata$AgeBands %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[10],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Age	Count
30-44	254
18-29	221
> 60	272
45-60	290
NA	92

Note that the above Age Bands are not listed in sequence from youngest to oldest.

We need to resequence the levels in order to fix this:

```
### use fct_relevel from library `forcats` to sort the AgeBands levels ordinally
tv_commadata$AgeBands <- tv_commadata$AgeBands %>%
  fct_relevel(levels(tv_commadata$AgeBands)[c(2,1,4,3,5)])
```

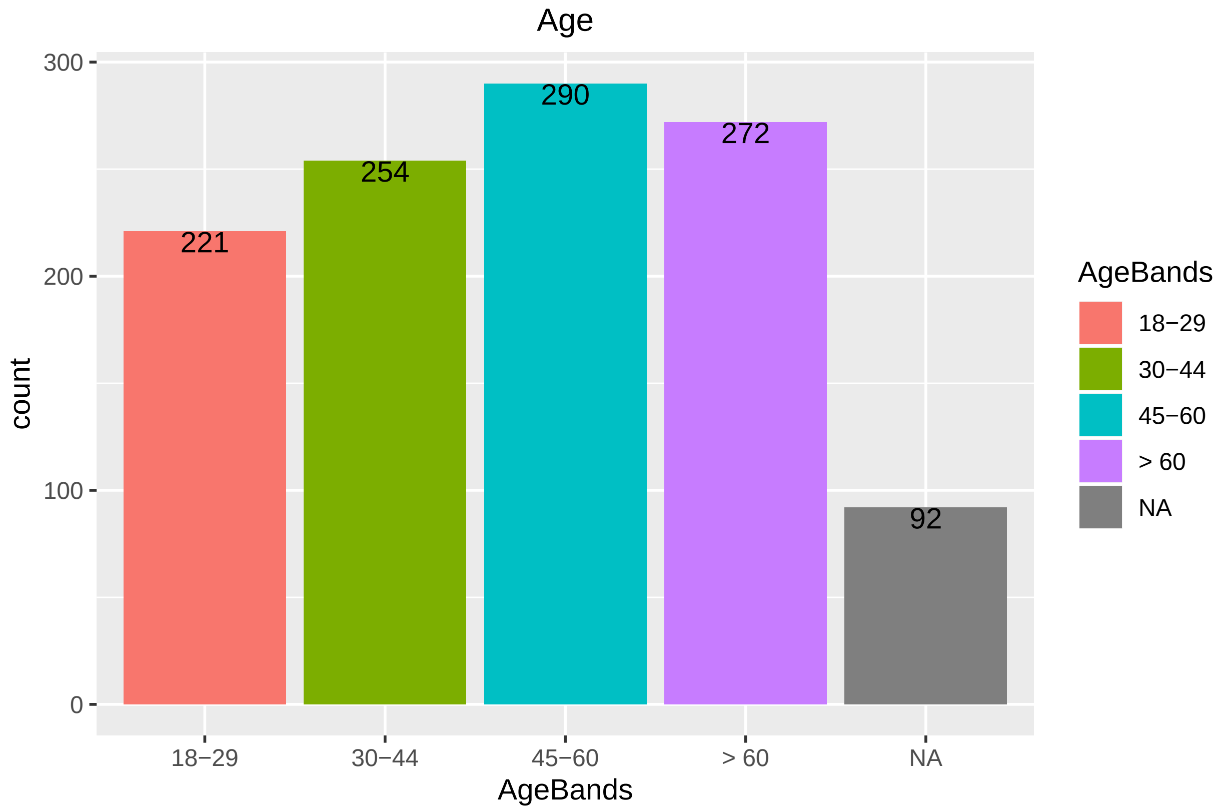
```
## Warning: Unknown levels in `f`: NA
```

```
### display results
tv_commadata$AgeBands %>% fct_count() %>%
  kable(col.names=c(new_variable_names[10],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

AgeBands	Count
18-29	221
30-44	254
45-60	290
> 60	272
NA	92

Plot the Age of participants using ggplot2

```r
ggplot(tv_commadata, aes(x=AgeBands,fill=AgeBands)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=1) +
  labs(title=initial_variable_names[10],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```



Source: Five−Thirty−Eight survey

Resequence the (unordered) levels for [11] `IncomeBands` to reflect the ordering of Income, from lowest to highest:

`forcats::fct_count()`: Count the number of responses of each type

```r
tv_commadata$IncomeBands %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[11],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Household Income	Count
$50,000 - $99,999	290
$25,000 - $49,999	158
$0 - $24,999	121
$150,000+	103
$100,000 - $149,999	164
NA	293

Note that the above Income Bands are not listed in sequence from lowest to highest.

We need to resequence the levels in order to fix this:

```r
### use fct_relevel from library `forcats` to sort the IncomeBands levels ordinally
tv_commadata$IncomeBands <- tv_commadata$IncomeBands %>%
  fct_relevel(levels(tv_commadata$IncomeBands)[c(3,2,1,5,4,6)])
```

```
## Warning: Unknown levels in `f`: NA
```

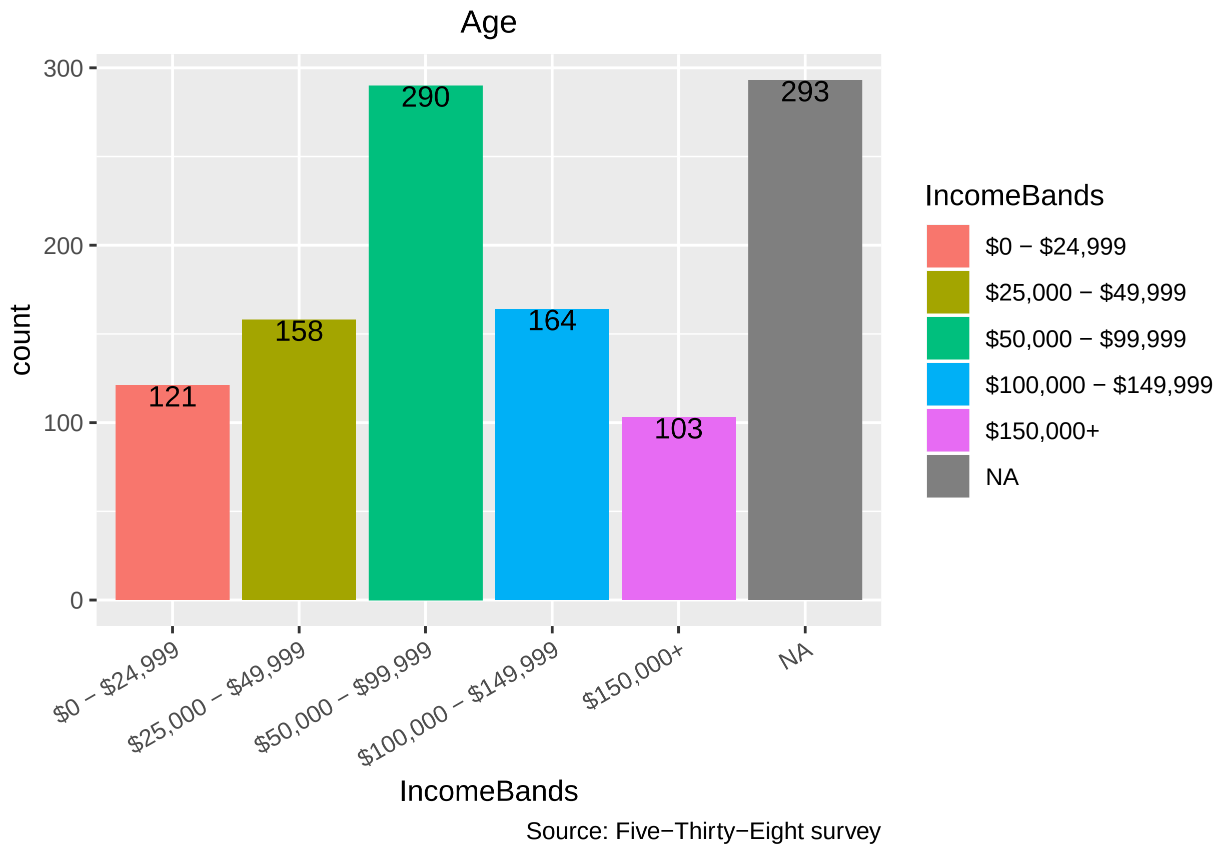
```r
### display results
tv_commadata$IncomeBands %>% fct_count() %>%
  kable(col.names=c(new_variable_names[11],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

IncomeBands	Count
$0 - $24,999	121
$25,000 - $49,999	158
$50,000 - $99,999	290
$100,000 - $149,999	164
$150,000+	103
NA	293

Plot the Income of participants using ggplot2

use axis.text.x=element_text(angle=30) to rotate the column headings

```
ggplot(tv_commadata, aes(x=IncomeBands,fill=IncomeBands)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), vjust=1) +
  labs(title=initial_variable_names[10],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.text.x=element_text(angle=30, hjust=1))
```



Source: Five−Thirty−Eight survey

Resequence the (unordered) levels for [12] `Education` to reflect the ordering of Educational Attainment:

`forcats::fct_count()`: Count the number of responses of each type

```r
tv_commadata$Education %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[12],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Education	Count
Bachelor degree	344
Graduate degree	276
Less than high school degree	11
Some college or Associate degree	295
High school degree	100
NA	103

Note that the above Education levels are not listed in sequence from lowest to highest.

We need to resequence the levels in order to obtain the desired sequence:

```r
### use fct_relevel from library `forcats` to sort the Education levels ordinally
tv_commadata$Education <- tv_commadata$Education %>%
  fct_relevel(levels(tv_commadata$Education)[c(3,5,4,1,2,6)])
```

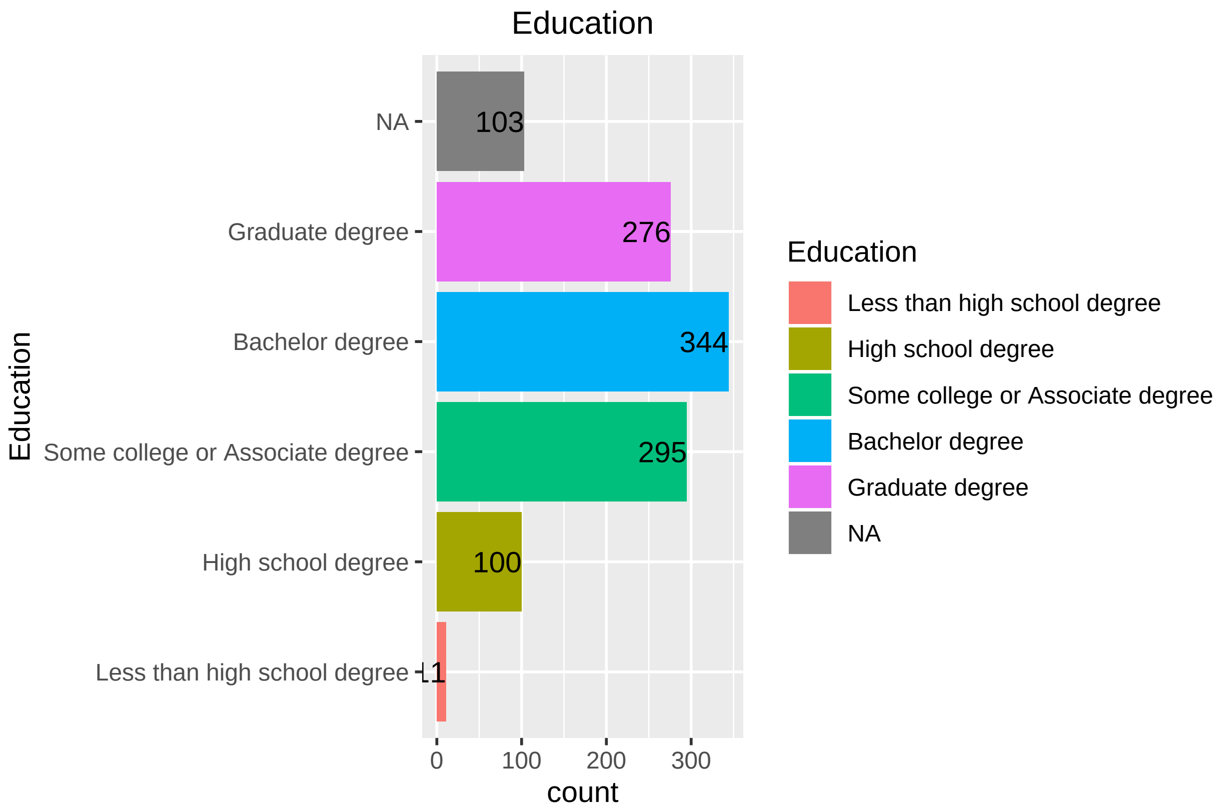
```
## Warning: Unknown levels in `f`: NA
```

```r
### display results
tv_commadata$Education %>% fct_count() %>%
  kable(col.names=c(new_variable_names[12],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Education	Count
Less than high school degree	11
High school degree	100
Some college or Associate degree	295
Bachelor degree	344
Graduate degree	276
NA	103

Plot the Education Level of participants using ggplot2

```r
ggplot(tv_commadata, aes(x=Education,fill=Education)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), hjust=1) +
  labs(title=initial_variable_names[12],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))+
  coord_flip()
```



Source: Five−Thirty−Eight survey

Resequence the (unordered) levels for [13] `Location` to reflect to reflect geography (east coast to west coast):

`forcats::fct_count()`: Count the number of responses of each type

```
tv_commadata$Location %>% fct_count() %>%
  kable(col.names=c(initial_variable_names[13],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Location (Census Region)	Count
South Atlantic	164
Mountain	87
East North Central	170
Middle Atlantic	140
New England	73
Pacific	180
East South Central	43
West North Central	82
West South Central	88
NA	102

Note that the above Locations levels are not listed to reflect geography (east coast to west coast; north to south):

We need to resequence the levels in order to obtain the desired sequence:

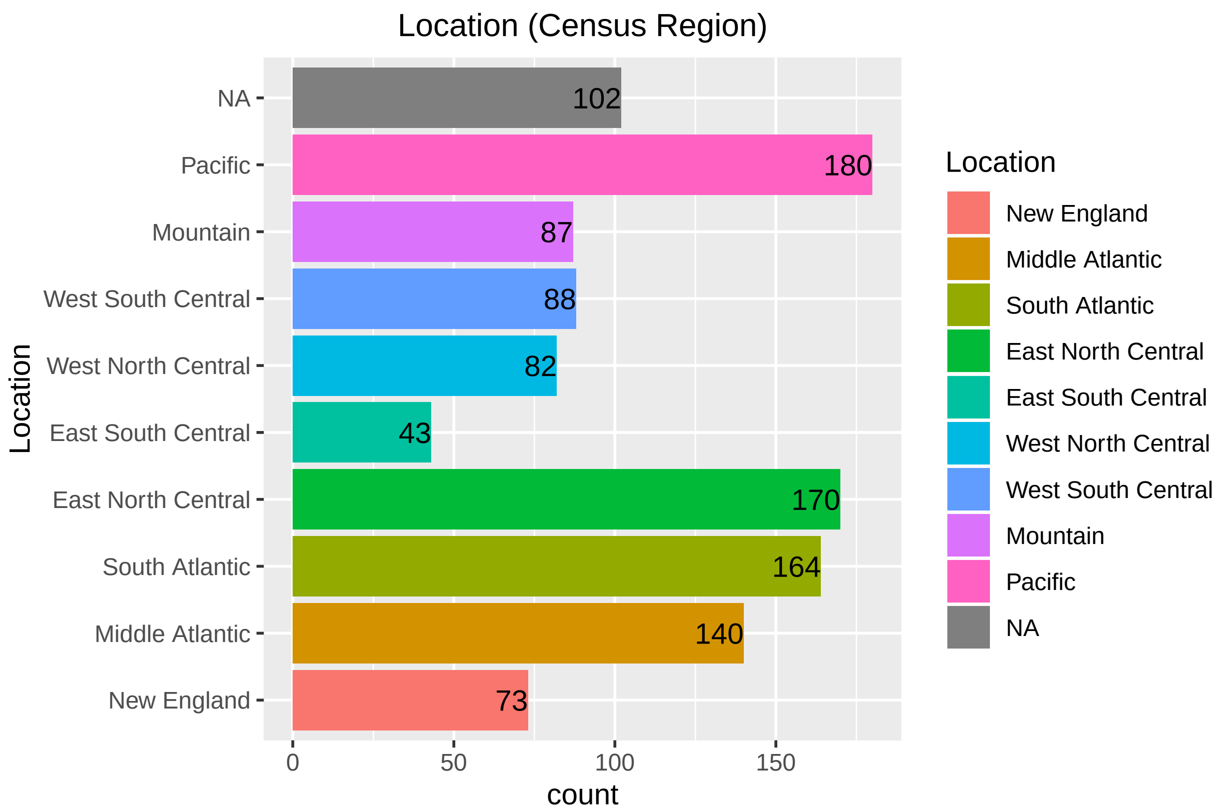
- New England,
- Middle Atlantic,
- South Atlantic,
- East North Central,
- East South Central,
- West North Central,
- West South Central,
- Mountain,
- Pacific, followed by
- NA

```
### use fct_relevel from library `forcats` to sort the Location data geographically
tv_commadata$Location <- tv_commadata$Location %>%
  fct_relevel(levels(tv_commadata$Location)[c(5,4,1,3,7,8,9,2,6)])
### display results
tv_commadata$Location %>% fct_count() %>%
  kable(col.names=c(new_variable_names[13],"Count")) %>%
  kable_styling(c("striped", "bordered"))
```

Location	Count
New England	73
Middle Atlantic	140
South Atlantic	164
East North Central	170
East South Central	43
West North Central	82
West South Central	88
Mountain	87
Pacific	180
NA	102

Plot the Location of respondents using ggplot2

```
ggplot(tv_commadata, aes(x=Location,fill=Location)) +
  geom_bar()+
  geom_text(stat='count', aes(label=..count..), hjust=1) +
  labs(title=initial_variable_names[13],
       caption="Source: Five-Thirty-Eight survey")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))+
  coord_flip()
```



Source: Five−Thirty−Eight survey

[1] `RespondentID` should not impact the results – it is just an identifier, so drop it

`dplyr::select(-[columnname])`: Drop variable [1] `RespondentID`

```
tv_commadata <- tv_commadata %>% select(-RespondentID)
```

References

1. Hickey, Walt, “Elitist, Superfluous, Or Popular? We Polled Americans on the Oxford Comma” (June 17, 2014), FiveThirtyEight.com . Retrieved December 1, 2019, from https://fivethirtyeight.com/features/elitist-superfluous-or-popular-we-polled-americans-on-the-oxford-comma/.

2. FiveThirtyEight.com survey of Oxford Comma Usage (2014, June). Retrieved December 1, 2019, from https://raw.githubusercontent.com/fivethirtyeight/data/master/comma-survey/comma-survey.csv.

End of Part 1