```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(reshape2)
```

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

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

```
library(readr)
library(forcats)
library(ggthemes)
```

#1 Using appropriate r code, read in the emailed excel spread sheet

```
college <- read_csv("college_score.csv")
```

```
## Parsed with column specification:
## cols(
##   UNITID = col_double(),
##   OPEID = col_double(),
##   MN_EARN_WNE_P6 = col_character(),
##   INSTNM = col_character(),
##   SAT_AVG = col_double(),
##   ADM_RATE = col_double(),
##   UGDS = col_double(),
##   COSTT4_A = col_double(),
##   AVGFACSAL = col_double(),
##   GRAD_DEBT_MDN = col_character(),
##   AGE_ENTRY = col_character(),
##   ICLEVEL = col_double()
## )
```

```r
str(college)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 7175 obs. of  12 variables:
##  $ UNITID        : num  100654 100663 100690 100706 100724 ...
##  $ OPEID         : num  100200 105200 2503400 105500 100500 ...
##  $ MN_EARN_WNE_P6: chr  "27800" "37600" "39400" "41300" ...
##  $ INSTNM        : chr  "Alabama A & M University" "University of Alabama at Birmingham" "Amridge Uni
##  $ SAT_AVG       : num  849 1125 NA 1257 825 ...
##  $ ADM_RATE      : num  0.874 0.581 NA 0.763 0.459 ...
##  $ UGDS          : num  4616 12047 293 6346 4704 ...
##  $ COSTT4_A      : num  22667 22684 13380 22059 19242 ...
##  $ AVGFACSAL     : num  7028 10517 3857 9463 7952 ...
##  $ GRAD_DEBT_MDN : chr  "32750" "21833" "22890" "22647" ...
##  $ AGE_ENTRY     : chr  "20.28374137" "23.60797466" "33.6722973" "22.72791963" ...
##  $ ICLEVEL       : num  1 1 1 1 1 1 2 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   UNITID = col_double(),
##   ..   OPEID = col_double(),
##   ..   MN_EARN_WNE_P6 = col_character(),
##   ..   INSTNM = col_character(),
##   ..   SAT_AVG = col_double(),
##   ..   ADM_RATE = col_double(),
##   ..   UGDS = col_double(),
##   ..   COSTT4_A = col_double(),
##   ..   AVGFACSAL = col_double(),
##   ..   GRAD_DEBT_MDN = col_character(),
##   ..   AGE_ENTRY = col_character(),
##   ..   ICLEVEL = col_double()
##   .. )
```

```r
college$MN_EARN_WNE_P6<- as.numeric(as.character(college$MN_EARN_WNE_P6))
```

```
## Warning: NAs introduced by coercion
```

```r
college$GRAD_DEBT_MDN<- as.numeric(as.character(college$GRAD_DEBT_MDN))
```

```
## Warning: NAs introduced by coercion
```

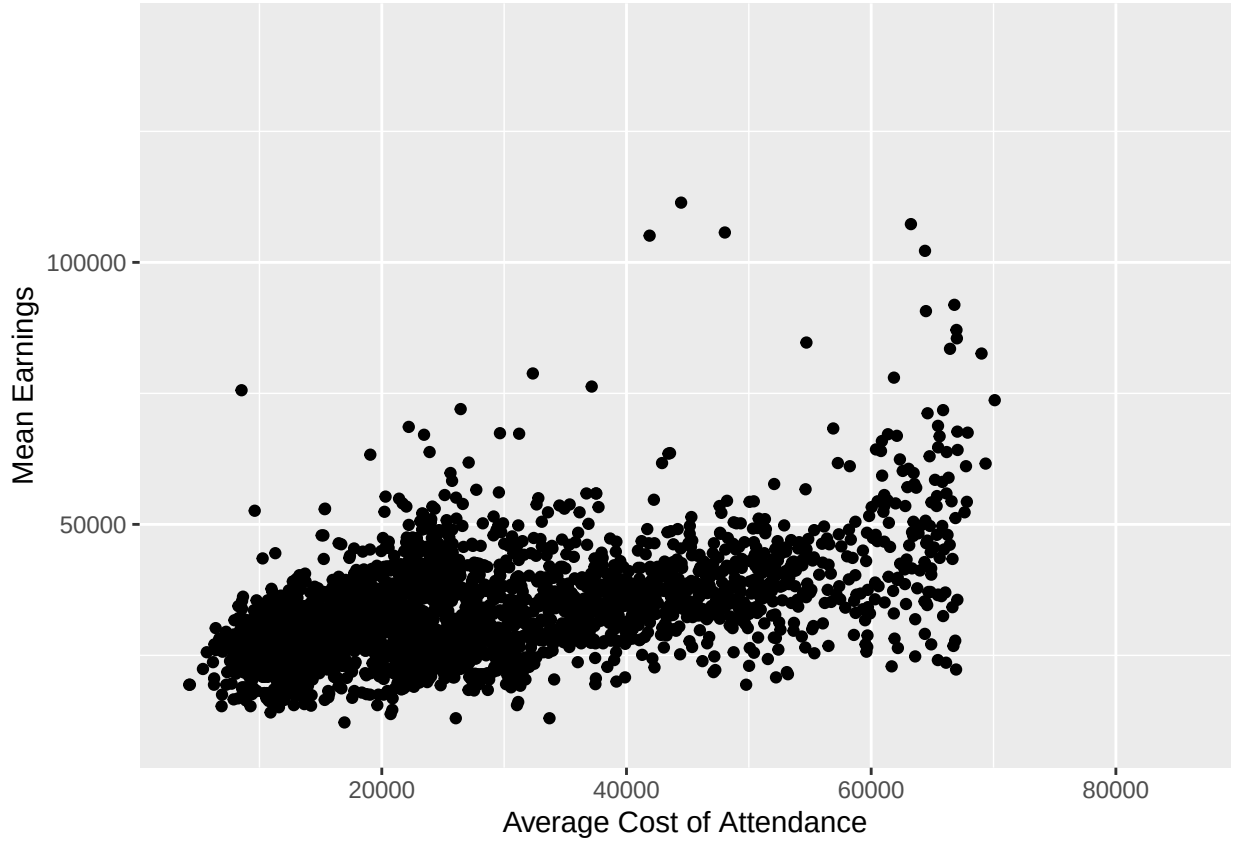
```r
college$AGE_ENTRY<- as.numeric(as.character(college$AGE_ENTRY))
```

```
## Warning: NAs introduced by coercion
```

#2 Given the level of the institution, does there appear to be an association between the average cost of attendance and the mean earnings of students six years after graduation? Make an appropriate plot to justify your response. You will be evaluated on the appropriateness of the plot and the aesthetics of the plot. (Hint: Generate two plots to make your decision, first a standard scatter plot involving the two continuous variables mentioned and then a facet plot over the appropriate categorical variable)

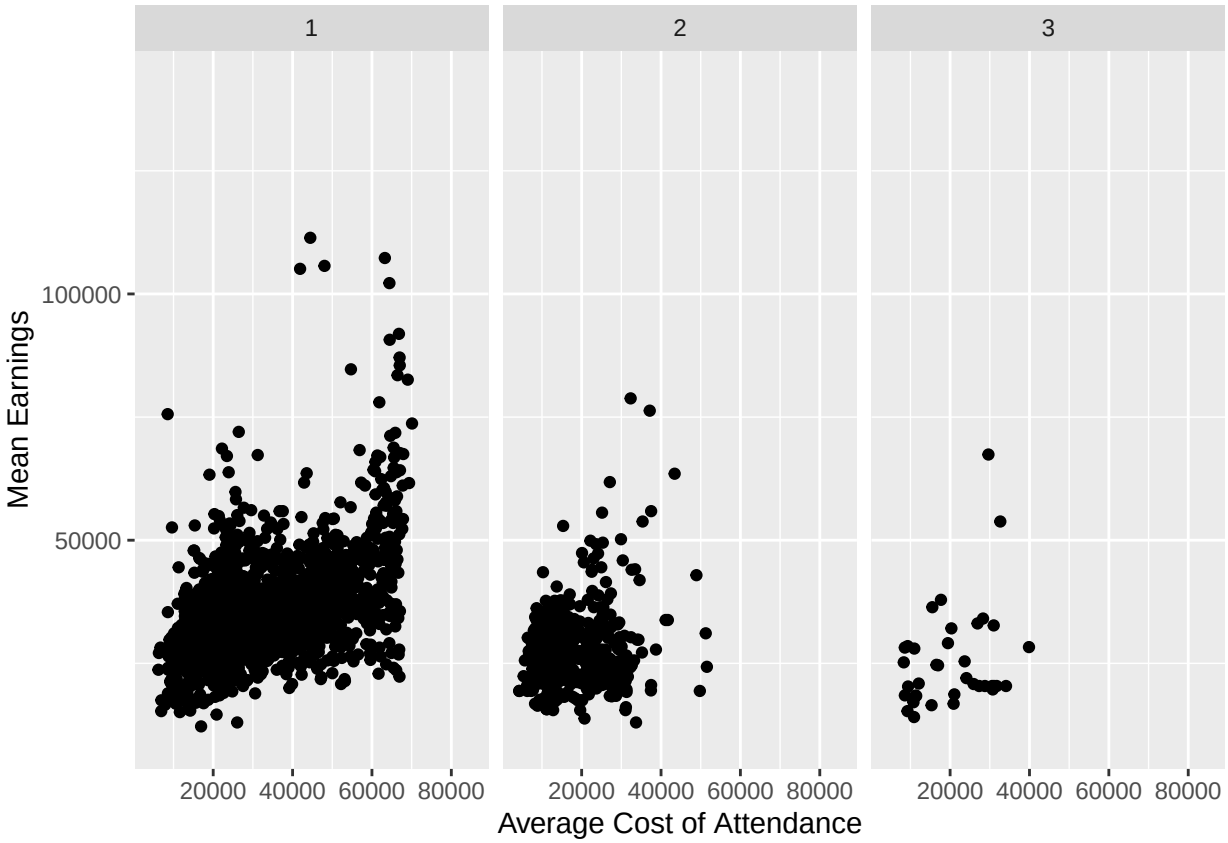
```
college %>%
  ggplot(aes(COSTT4_A, MN_EARN_WNE_P6)) +
  geom_point() +
  labs(x = "Average Cost of Attendance", y = "Mean Earnings", main = "Relation between Attendance cost a
```

```
## Warning: Removed 4112 rows containing missing values (geom_point).
```




```
college %>%
  ggplot(aes(COSTT4_A, MN_EARN_WNE_P6)) +
  geom_point() +
  labs(x = "Average Cost of Attendance", y = "Mean Earnings", main = "Relation between Attendance cost a
  facet_wrap(~ICLEVEL)
```

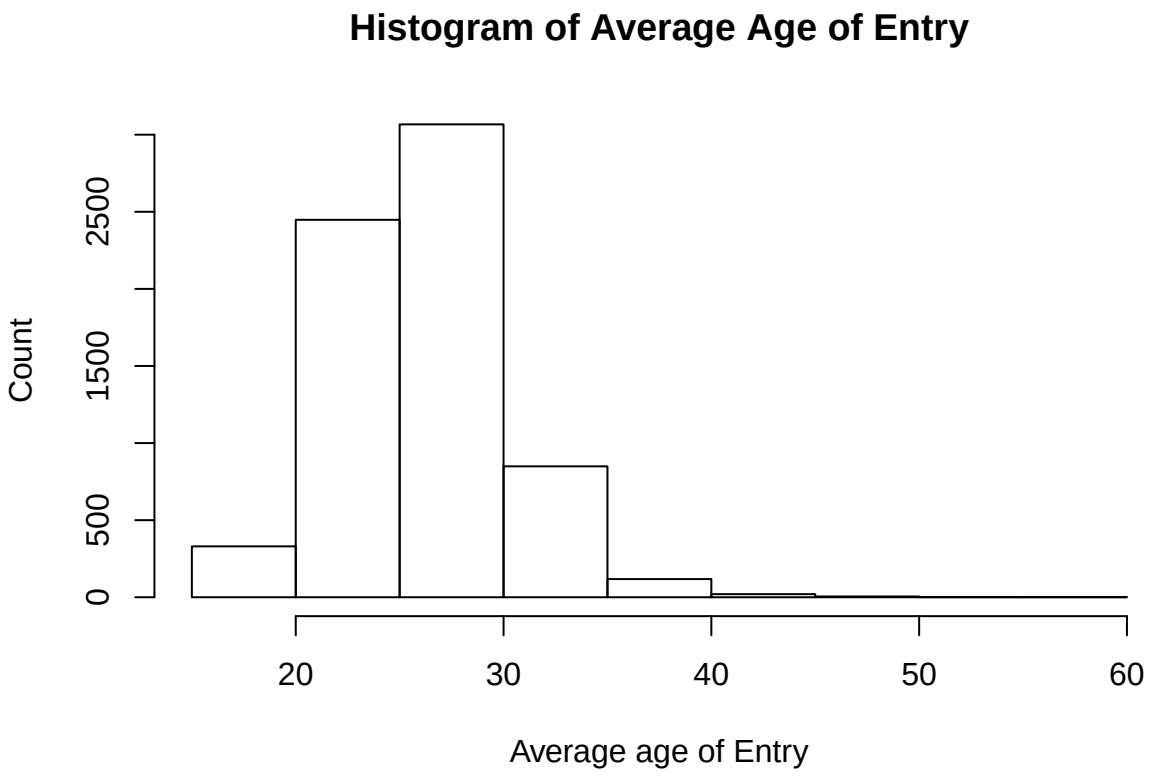
```
## Warning: Removed 4112 rows containing missing values (geom_point).
```



Looking at the 1st plot we can see that there is a slight positive association between mean earning and average cost of attendance. In the 2nd plot, the level 1 and 2 institutes have a positive association, whereas level 3 institutes have nearly 0 association between the two variables.

#3 Use r code to produce a histogram of the average age of entry. Comment on the distribution of this variable.

```
hist(college$AGE_ENTRY, xlab = "Average age of Entry", ylab = "Count", main = "Histogram of Average Age
```



The distribution of Age of Entry is right skewed.

#4 Use r code that will produce output that shows the 10 institutions that have the highest average age of entry?

```
college %>%
  select(INSTNM, AGE_ENTRY) %>%
  arrange(desc(AGE_ENTRY)) %>%
  head(n = 10)
```

```
## # A tibble: 10 x 2
##    INSTNM                                    AGE_ENTRY
##    <chr>                                         <dbl>
##  1 Advanced Technical Centers                     58.9
##  2 World Mission University                       51.6
##  3 Carolina Christian College                     49.1
##  4 Grace Mission University                       48.7
##  5 Prestige Health & Beauty Sciences Academy      48.1
##  6 Apex School of Theology                        46.0
##  7 Georgia Christian University                   45.7
##  8 Beulah Heights University                      43.6
##  9 Taft University System                         43.6
## 10 Cosmopolitan Beauty and Tech School            42.7
```

#5 There are many universities with "American University" in the name. E.g. "American University of Puerto Rico" and "National American University-Ellsworth AFB Extension". Use r code to create a data

frame, called `americandf`, that contains just the data from universities with "American University" in the name.

```
americandf <- college %>%
  filter(str_detect(INSTNM, "American University"))
americandf
```

```
## # A tibble: 47 x 12
##    UNITID  OPEID MN_EARN_WNE_P6 INSTNM SAT_AVG ADM_RATE  UGDS COSTT4_A
##     <dbl>  <dbl>          <dbl> <chr>    <dbl>    <dbl> <dbl>    <dbl>
##  1 127680 4.06e5          35200 Natio~      NA   NA       194       NA
##  2 131159 1.43e5          46800 Ameri~    1262    0.259  7276    60501
##  3 174385 4.06e5          35200 Natio~      NA   NA       152    18707
##  4 219204 4.06e5          35200 Natio~      NA   NA      1537    20069
##  5 219213 4.06e5          35200 Natio~      NA   NA       365    15870
##  6 241100 1.19e6          20700 Ameri~      NA   NA       583    16265
##  7 241128 1.19e6          20700 Ameri~      NA   NA       742    16393
##  8 242617 4.25e6             NA Inter~      NA    0.691  3972    12881
##  9 242626 3.94e5             NA Inter~      NA    0.466  3912    11158
## 10 242635 5.03e5             NA Inter~      NA    0.465  3841    12672
## # ... with 37 more rows, and 4 more variables: AVGFACSAL <dbl>,
## #   GRAD_DEBT_MDN <dbl>, AGE_ENTRY <dbl>, ICLEVEL <dbl>
```

#6 Provide r code that will produce the number of colleges from the College Score data frame that have average SAT scores that are above 1000. (Do not produce the data frame. Your code should only yield the number)

```
college %>%
  filter(SAT_AVG > 1000) %>%
  nrow()
```

```
## [1] 849
```

#7 Provide r code that will show a data frame that lists the 10 highest Average SAT scores in decreasing order. A partial data frame is given below.

```
college %>%
  arrange(desc(SAT_AVG)) %>%
  head(n = 10)
```

```
## # A tibble: 10 x 12
##    UNITID  OPEID MN_EARN_WNE_P6 INSTNM SAT_AVG ADM_RATE  UGDS COSTT4_A
##     <dbl>  <dbl>          <dbl> <chr>    <dbl>    <dbl> <dbl>    <dbl>
##  1 110404 113100          59800 Calif~    1555   0.0807   979    63471
##  2 166683 217800         107300 Massa~    1519   0.0794  4489    63250
##  3 144050 177400          73700 Unive~    1508   0.0794  5978    70100
##  4 166027 215500         102200 Harva~    1506   0.054   7447    64400
##  5 130794 142600          83500 Yale ~    1502   0.0632  5471    66445
##  6 115409 117100          61600 Harve~    1496   0.129    829    69355
##  7 190150 270700          82600 Colum~    1496   0.0683  8124    69021
##  8 221999 353500          57700 Vande~    1495   0.108   6844    63532
##  9 186131 262700          78000 Princ~    1493   0.0652  5236    61860
```

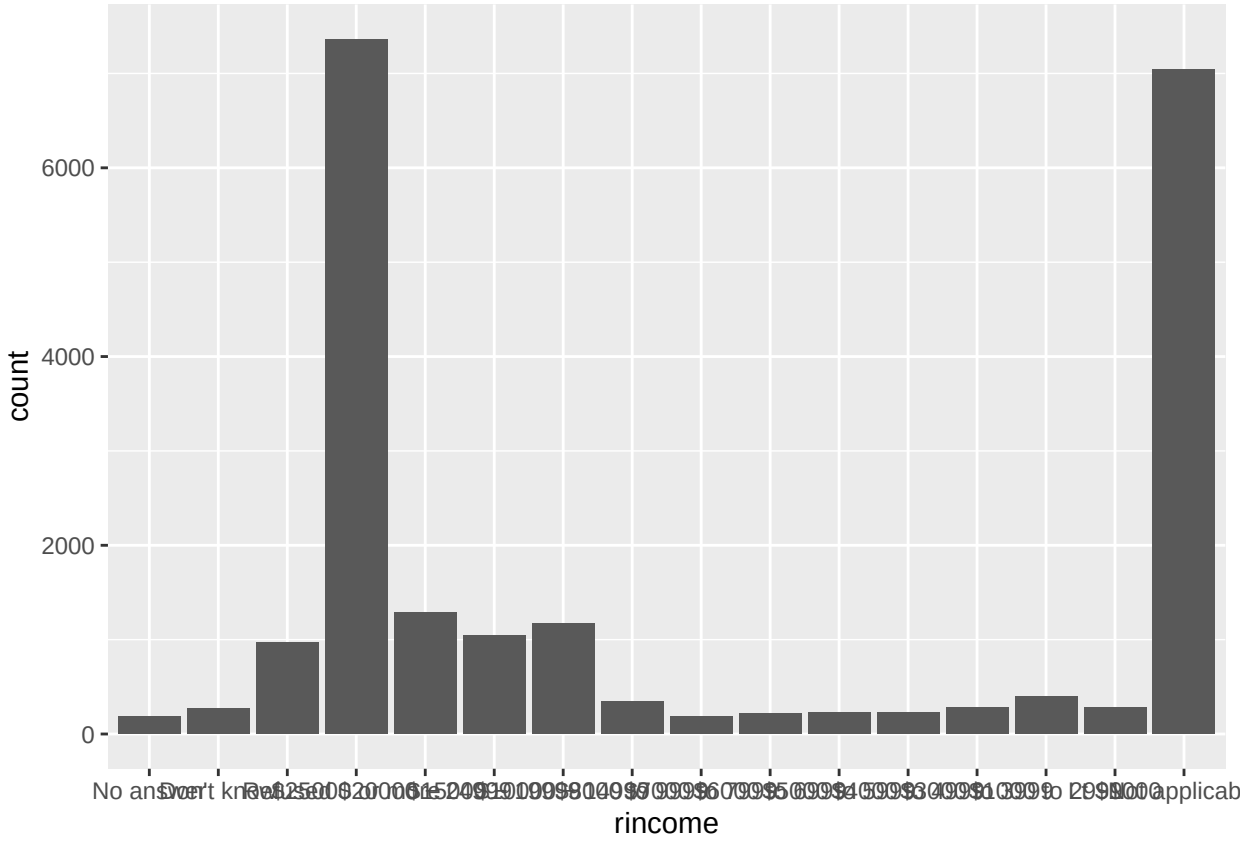
```
## 10 227757 360400          61100 Rice ~    1490   0.153   3879    58253
## # ... with 4 more variables: AVGFACSAL <dbl>, GRAD_DEBT_MDN <dbl>,
## #   AGE_ENTRY <dbl>, ICLEVEL <dbl>
```

#8 Using the gss_cat data frame, write r code that will produce the bar graph below. And explain in one or two sentences why the bar graph is difficult to interpret.

```
class(gss_cat$rincome)
```

```
## [1] "factor"
```

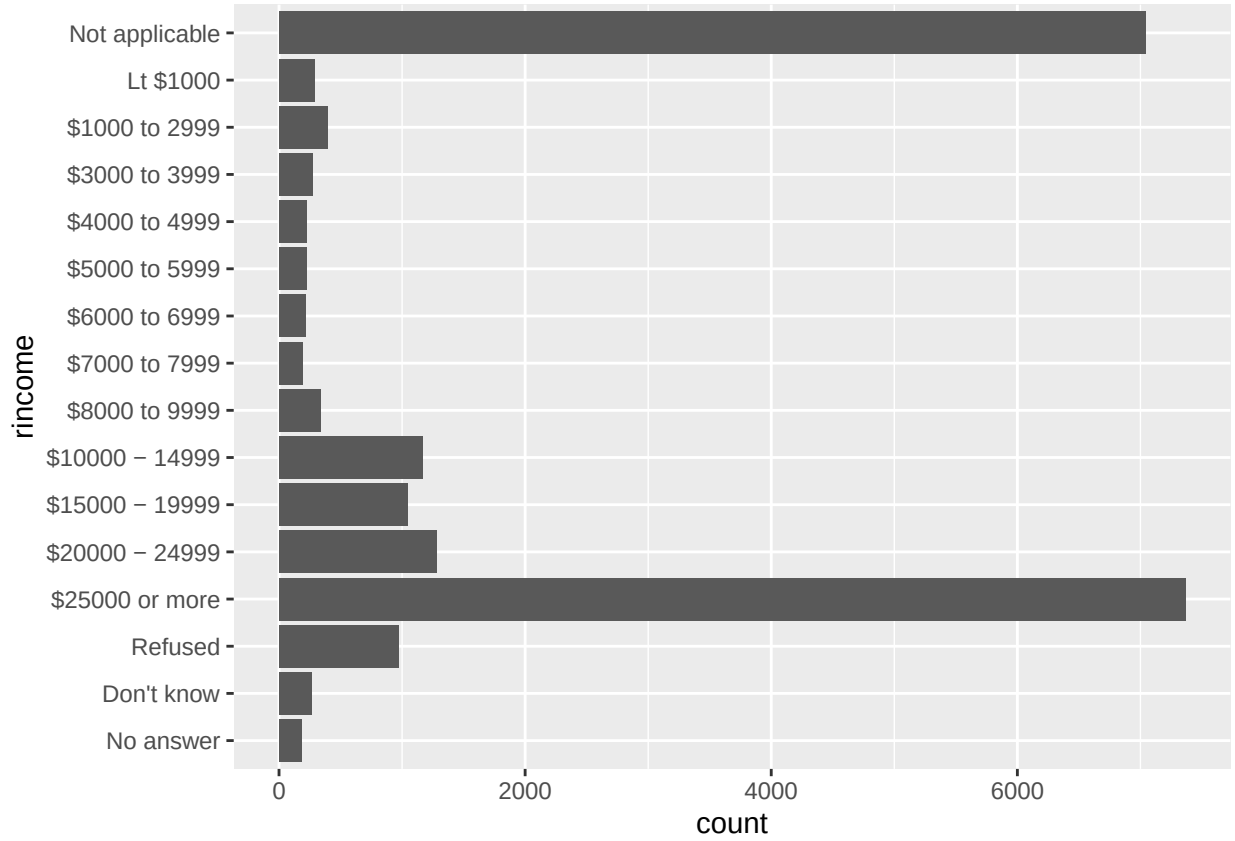
```
ggplot(gss_cat, aes(rincome)) +
  geom_bar()
```



The difficulty in interpretting the bar graph is due to the overlap of value of the x axis which makes it unreadable.

#9 Now write r code from the same data set that produce the transformed bar graph and comment on why it is an improvement

```
ggplot(gss_cat, aes(rincome)) +
  geom_bar() +
  coord_flip()
```



#Use r code to produce the tips data frame from the reshape2 package. Name three categorical variables in the data frame.

```
str(tips)
```

```
## 'data.frame':    244 obs. of  7 variables:
##  $ total_bill: num  17 10.3 21 23.7 24.6 ...
##  $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
##  $ sex       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

Sex, Smoker, Day and Time are the categorial variables in the data frame.

#10 Use r code to indicate how many levels exist for the factor day in the tips data frame and determine the frequency of each level.

```
levels(tips$day)
```

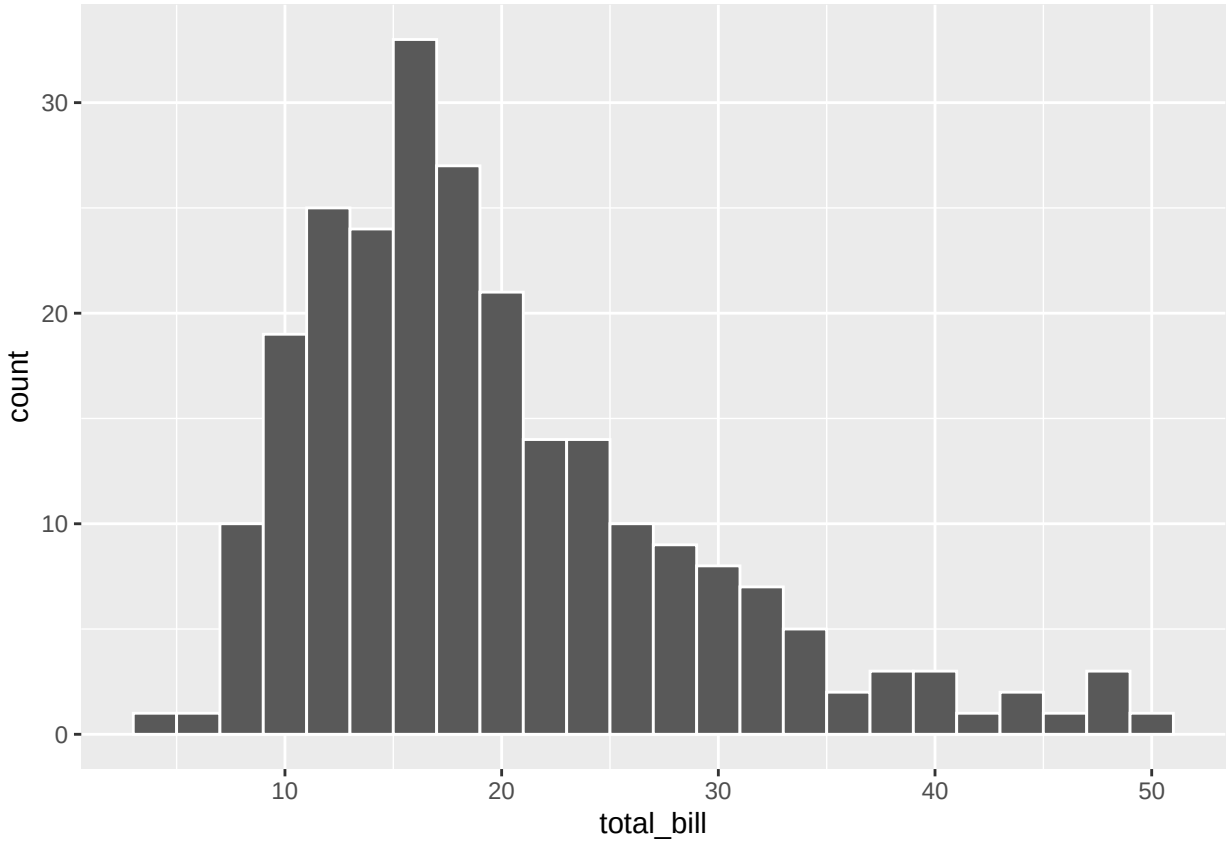
```
## [1] "Fri"  "Sat"  "Sun"  "Thur"
```

```r
table(tips$day)
```

```
##
##  Fri  Sat  Sun Thur
##   19   87   76   62
```

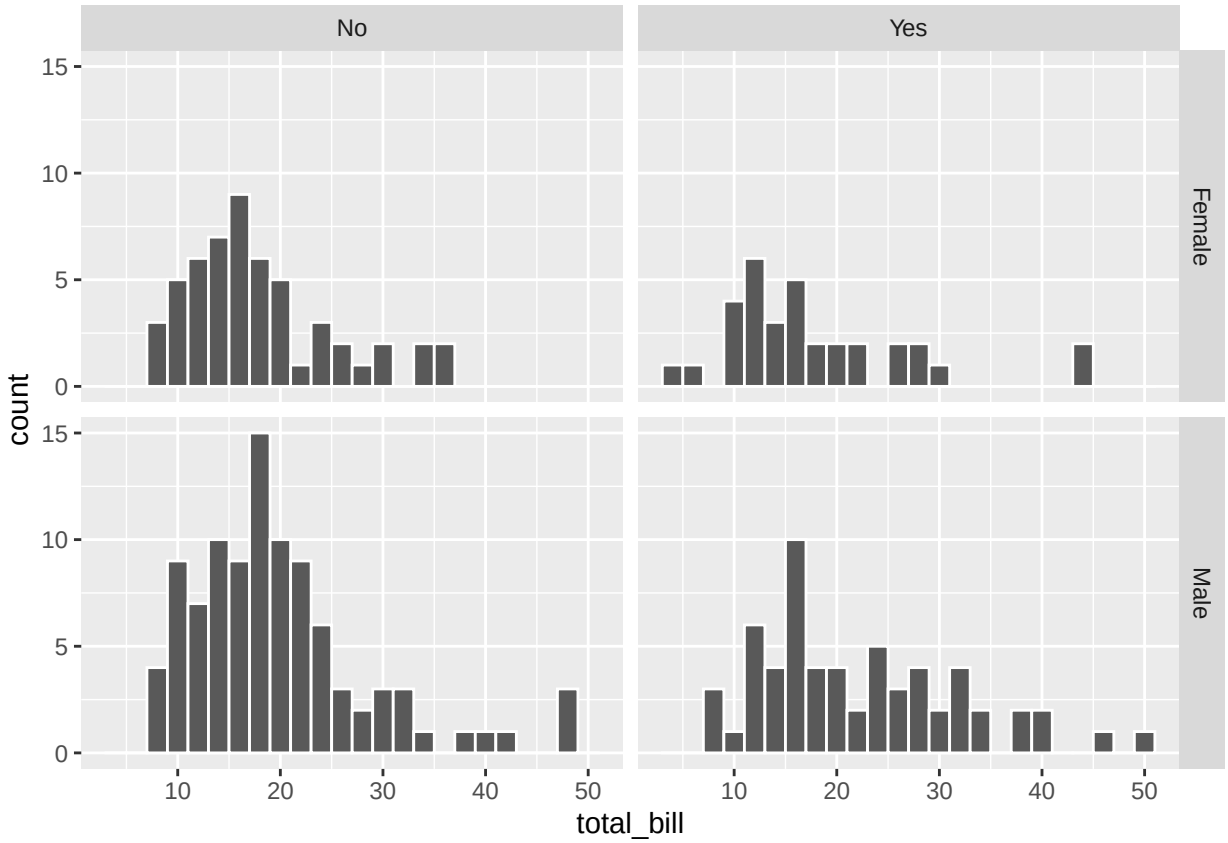
#11 Produce r code that will produce the following histogram from the tips data frame

```r
tips %>%
  ggplot() +
  geom_histogram(aes(total_bill), color = "white", bins = 30, binwidth = 2)
```



#12 Write r code that will produce the following histograms from the tips data frame

```r
tips %>%
  ggplot() +
  geom_histogram(aes(total_bill), color = "white", bins = 30, binwidth = 2) +
  facet_grid(sex ~ smoker)
```



#13 Using stringr::words, produce r code that will show all words that end with tion or ing

```
str_subset(stringr::words, "tion|ing$")
```

```
##  [1] "bring"     "condition" "during"    "evening"   "function"
##  [6] "king"      "meaning"   "mention"   "morning"   "motion"
## [11] "nation"    "position"  "question"  "relation"  "ring"
## [16] "section"   "sing"      "station"   "thing"
```