```
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

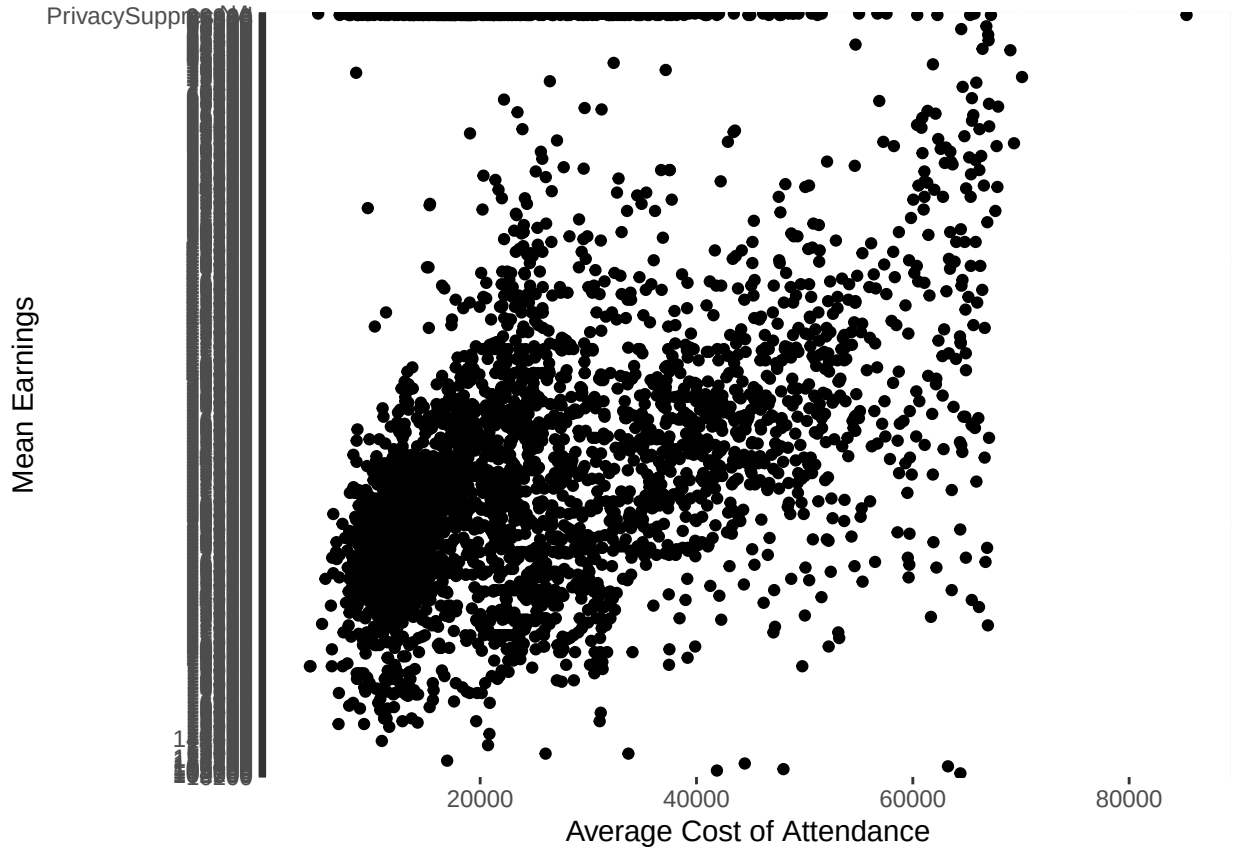
```
college <- read_csv("college_score.csv")
```

```
## 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()
## )
```

#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 3520 rows containing missing values (geom_point).
```



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

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

#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.

#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)

#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.

#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.

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

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

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

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

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