HW3_Madhu

Madhu Jagdale

2/15/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

Part One - theoretical

1. Choice of a location statistic

Tasks

For each of the following random variables, discuss whether the mean or the median would be a more useful measure of centrality.

a.The annual income of U.S. households.

Answer:

Median : will be a usefull measure of centrality for the annual income. In the toal US population high annual income people will be less than low annual income. If we calculate the mean of the annual income then it will show high income for low income people. Mean is depends on outliers of dataset and median is not much sensitive to outlier data.

b.The lifetime of a 75-watt light bulbs.

Answer:

Mean : The lifetime of a 75-watt light bulbs will be around the same value.

c.The Math SAT score of entering IIT freshmen.

Answer:

Mean : Math SAT score of entering IIT freshman will be the normal distribution. And most of the students will have the same score as it is a recommended score.

d.The age of participants in the Bay to Breakers 12K road race.

Answer:

Median : In the race all age of participiants are allowed. There will be younger and older people in race. Median is used when data is uneven and it is not affected by outlier data.

1.What are (non-normalized) bar graphs useful for?

Answer: Bar graphs are used to compare things between different groups and track data over time. Bar graphs are convinient when changes are larger. Bar graphs are more easy than tables to read a data.

2.State one advantage and one disadvantage of using a normalized bar graph.

Answer: Advantages:

1)Summarizes large dataset into visuale form.

2) easy to understand data in bar graphs than tables.

Disadvantages:

1) Requires more explanation as data is in visual form.

2) Can be difficult to rad when data is large.

3.What is a histogram?

Answer: Histogram is a graphical represantation of continuous data. It is used to show frequency distribution.

4.Describe one advantage and one disadvantage of using a normalized histogram.

Answer:Advantages:

1) Helps to visualize the distribution of data.

2) Can easily show mean, median and mode.

Disadvantages:

1) Can present data which is misleading.

2) Can not read values because data is grouped into categories.

Part Two - applied

Working With the Data

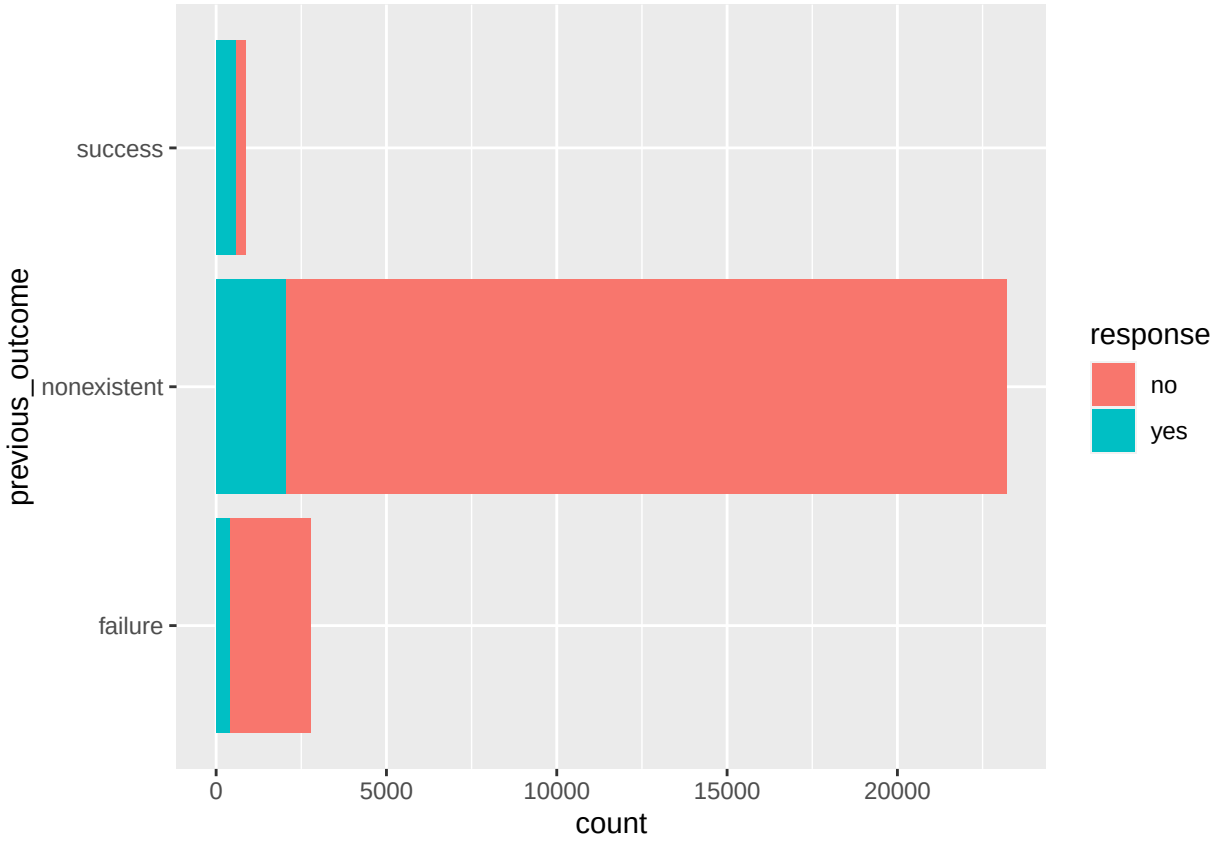
5. Create a bar graph of the *previous_outcome* variable, with *response* overlay.

```
dataset <- read_csv("https://campuspro-uploads.s3-us-west-2.amazonaws.com/a9d789c2-6b5e-4020-a941-699849
```

Answer:

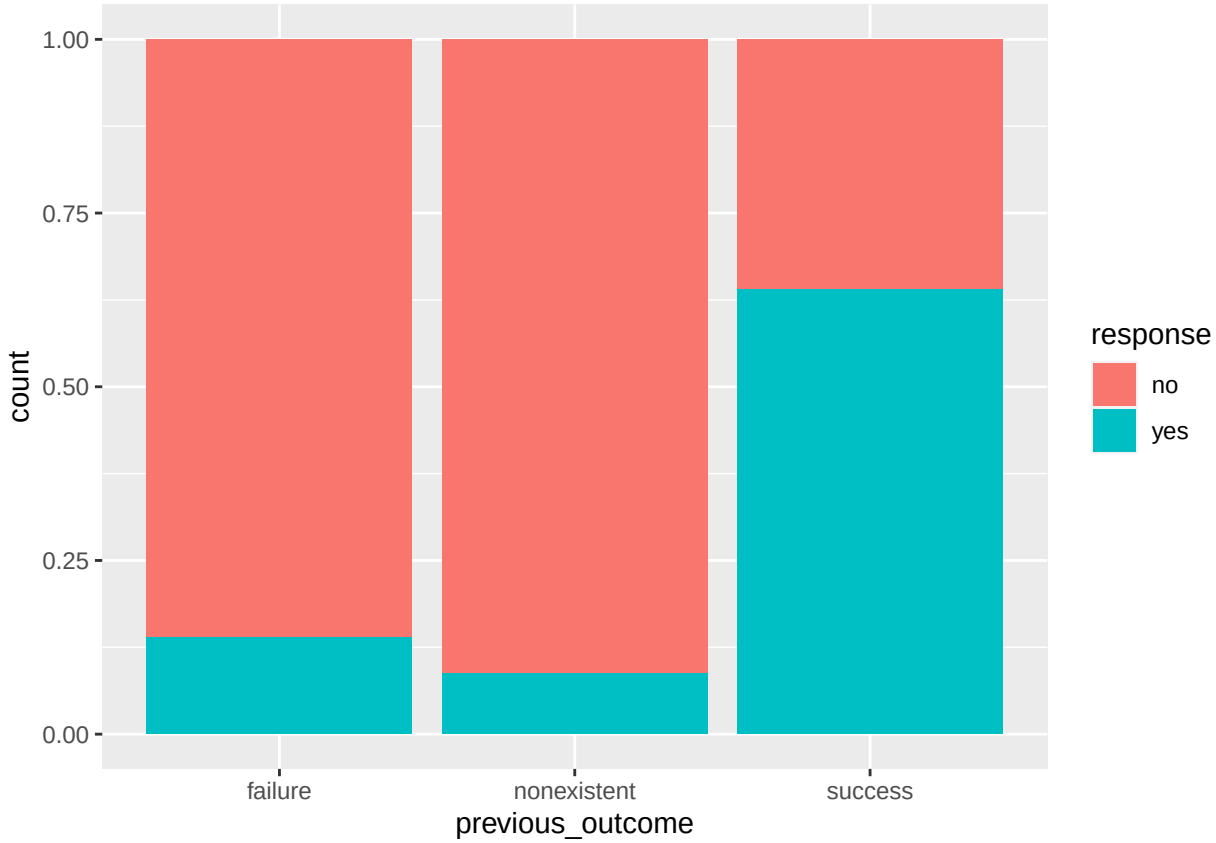
```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   age = col_double(),
##   duration = col_double(),
##   campaign = col_double(),
##   days_since_previous = col_double(),
##   previous = col_double(),
##   emp.var.rate = col_double(),
##   cons.price.idx = col_double(),
##   cons.conf.idx = col_double(),
##   euribor3m = col_double(),
##   nr.employed = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```
bank_train <- dataset
library(ggplot2)
ggplot(bank_train, aes(previous_outcome)) + geom_bar(aes(fill = response))+ coord_flip()
```



6.Create a normalized bar graph of previous_outcome variable with response overlay. Describe the relationship between previous_outcome and response.

```
ggplot(bank_train, aes(previous_outcome)) + geom_bar(aes(fill = response),position = "fill")
```



Answer:

The relationship between previous_outcome and response: Failures and nonexistents are not more likely to respond but success people are more likely to respond.

7.[optional!] Create a contingency table of previous_outcome and response. Compare the contingency table with the non-normalized bar graph and the normalized bar graph.

```
t1<-table(bank_train$response, bank_train$previous_outcome)
print(t1)
```

Answer:

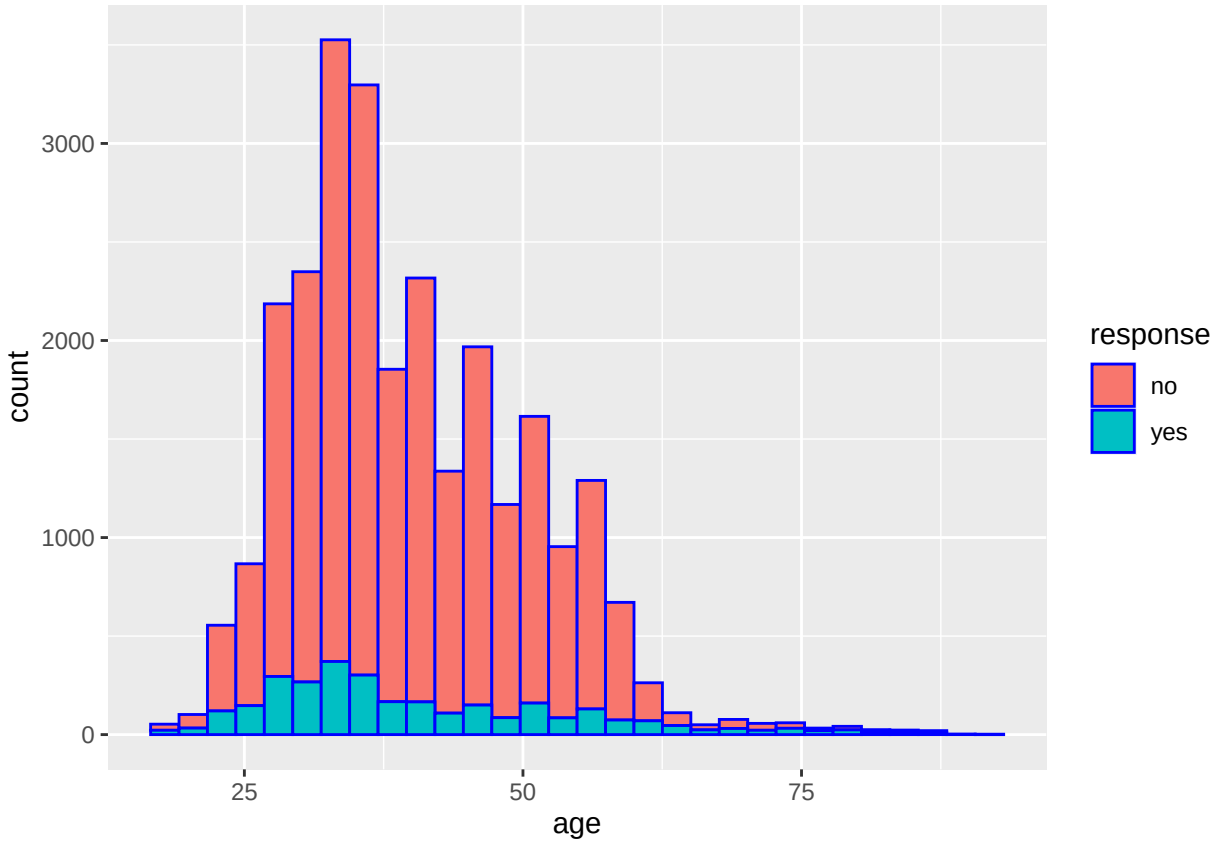
```
##
##       failure nonexistent success
##   no     2390       21176     320
##   yes     385        2034     569
```

8.Create a histogram of age with response overlay.

```
ggplot(bank_train, aes(age)) + geom_histogram(aes(fill = response), color="blue")
```

Answer:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

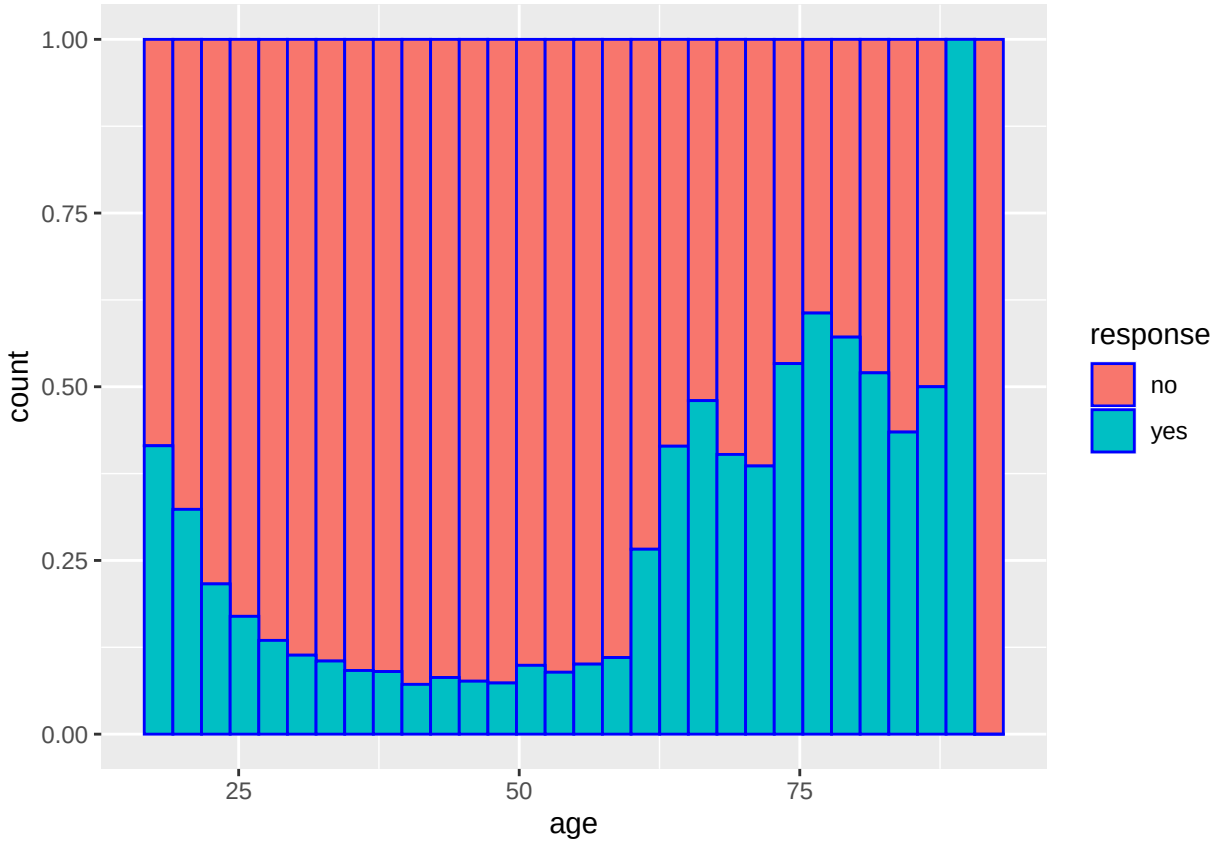


9.Create a normalized histogram of age with response overlay. Describe the relationship between age and response.

```
ggplot(bank_train, aes(age)) + geom_histogram(aes(fill = response), color="blue", position = "fill")
```

Answer:

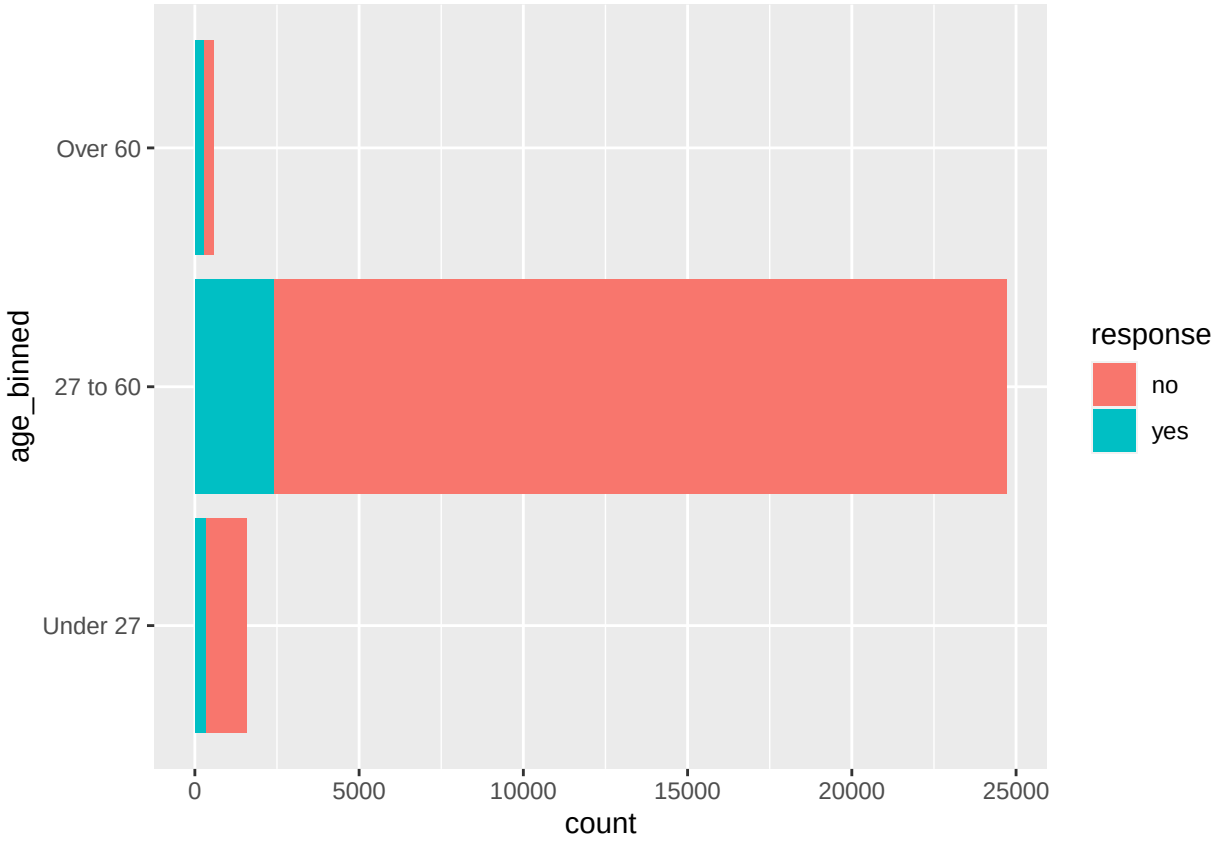
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The relationship between age and response: People whose age is less than 22 and more than 61 responses more than people age between 22 and 61.

10.Bin the age variable using the bins specified in the handout linked above and create a bar chart of the binned age variable with response overlay.

```
bank_train$age_binned <- cut(x = bank_train$age, breaks = c(0, 27, 60.01, 100), right = FALSE, labels =
ggplot(bank_train, aes(age_binned)) + geom_bar(aes(fill = response)) + coord_flip()
```



Answer:

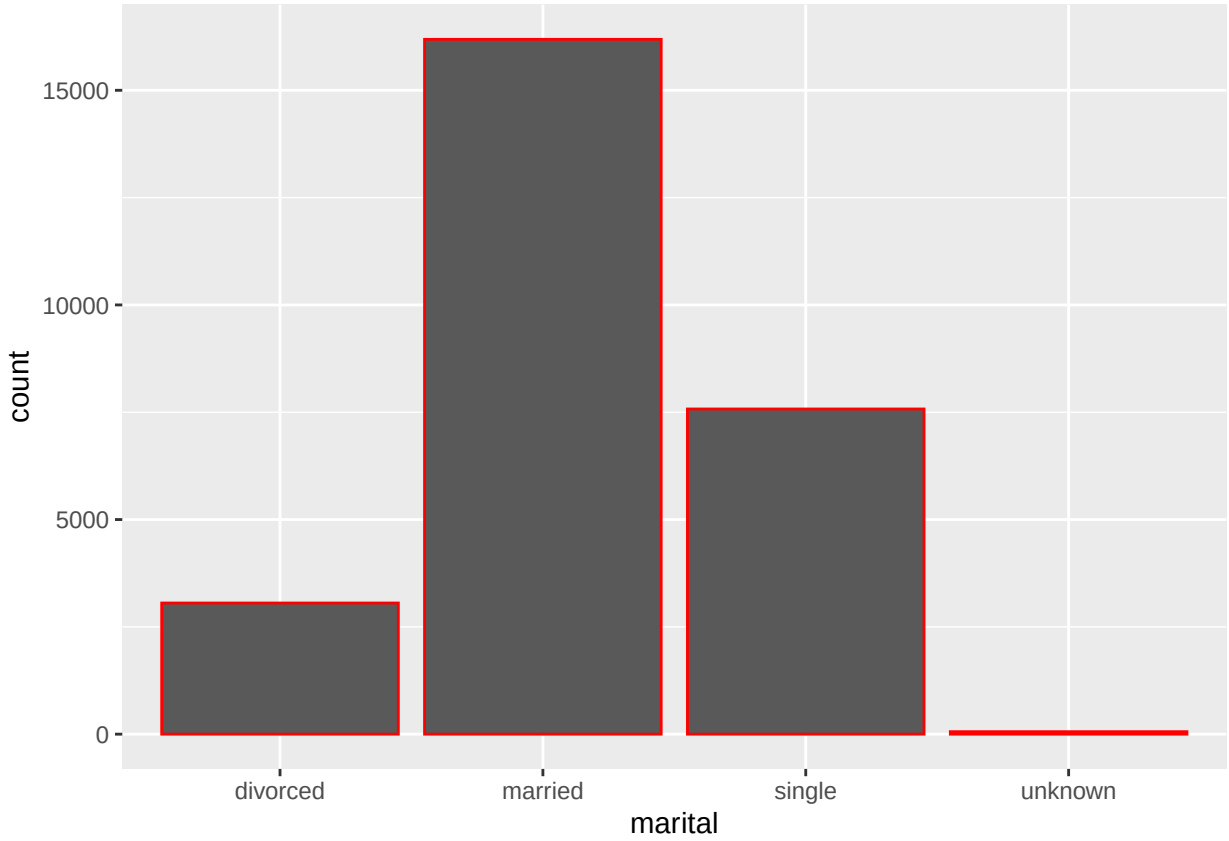
Hands-on Analysis

For Exercises in this section, continue working with the bank_marketing_training data set.

11.Produce the following graphs. What is the strength of each graph? Weakness?

a.Bar graph of marital.

```
ggplot(bank_train, aes(marital)) + geom_bar(color="red")
```

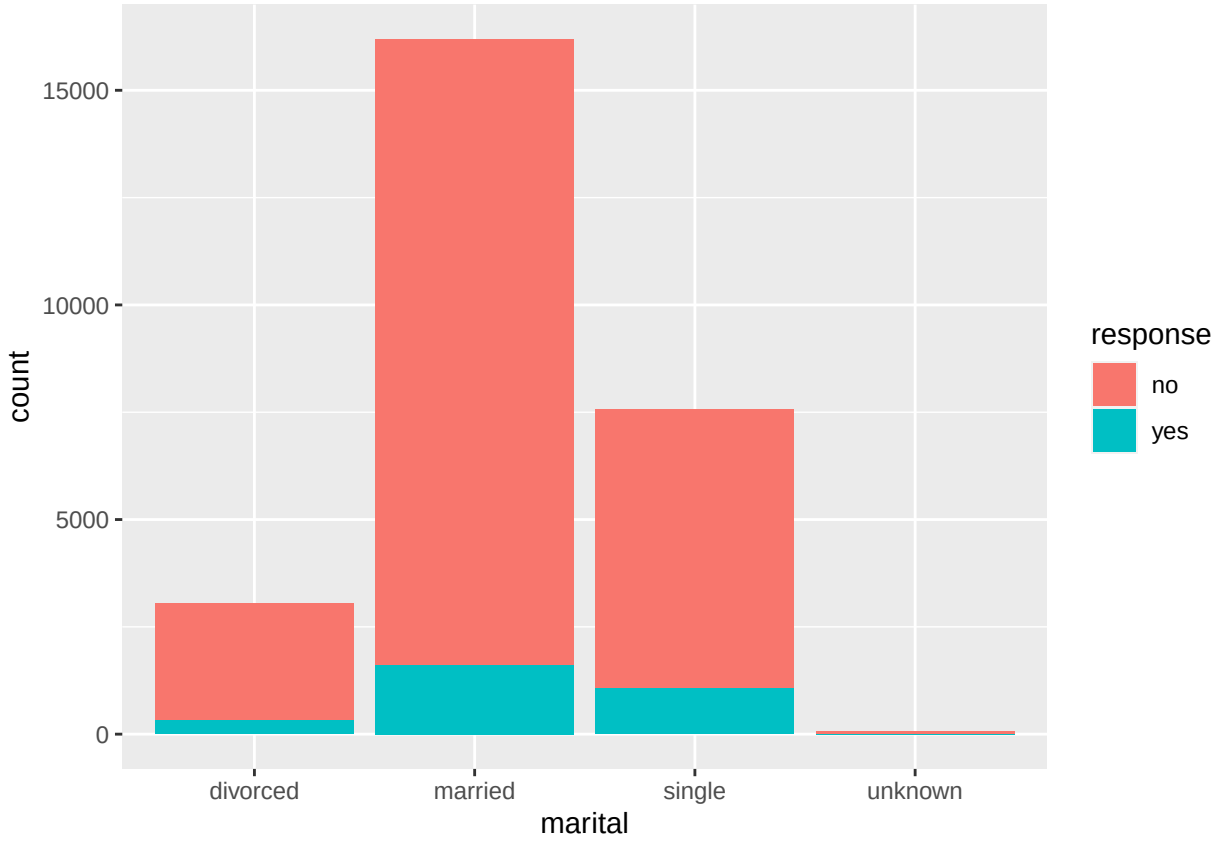


Answer:

In this bar graph of marital we get the information about the number of people under each category like divorced, married and single. Also, because the graph is not plotted with overlay of response we do not know how many people responded for yes or no under each category.

b.Bar graph of marital, with overlay of response.

```
ggplot(bank_train, aes(marital)) + geom_bar(aes(fill = response))
```

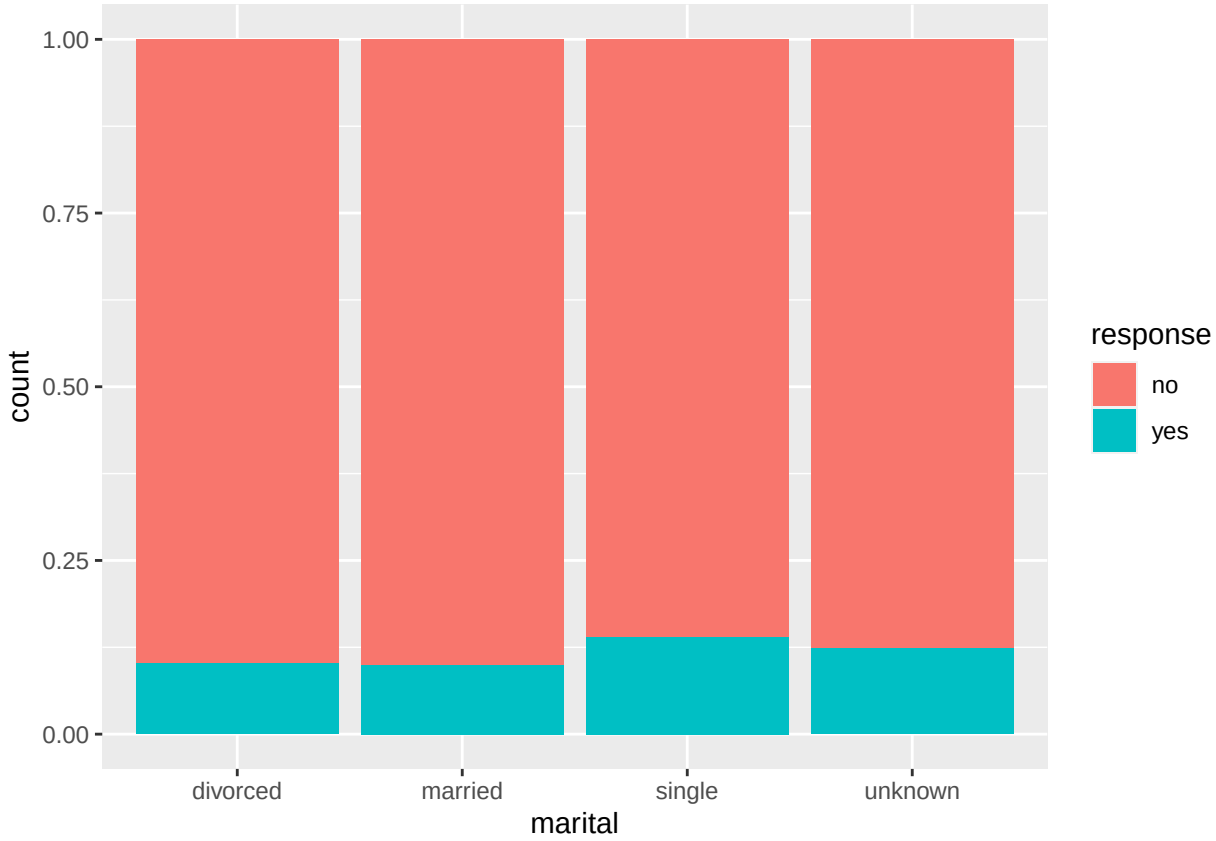


Answer:

The graph is plotted with overlay of response we do know how many people responded for yes or no under each category.Mean,median and mode is not clearly depicted from the graph.

c.Normalized bar graph of marital, with overlay of response.

```
ggplot(bank_train, aes(marital)) + geom_bar(aes(fill = response), position = "fill")
```



Answer:

The Normalized graph is plotted with overlay of response we depict Mean,median and mode from the graph.

12.Using the graph from Exercise 11c, describe the relationship between marital and response.

Answer: This is the normalized graph of response with respect to married, single, divorced and unknown category. By looking at the graph Single people responds more than other three category people.

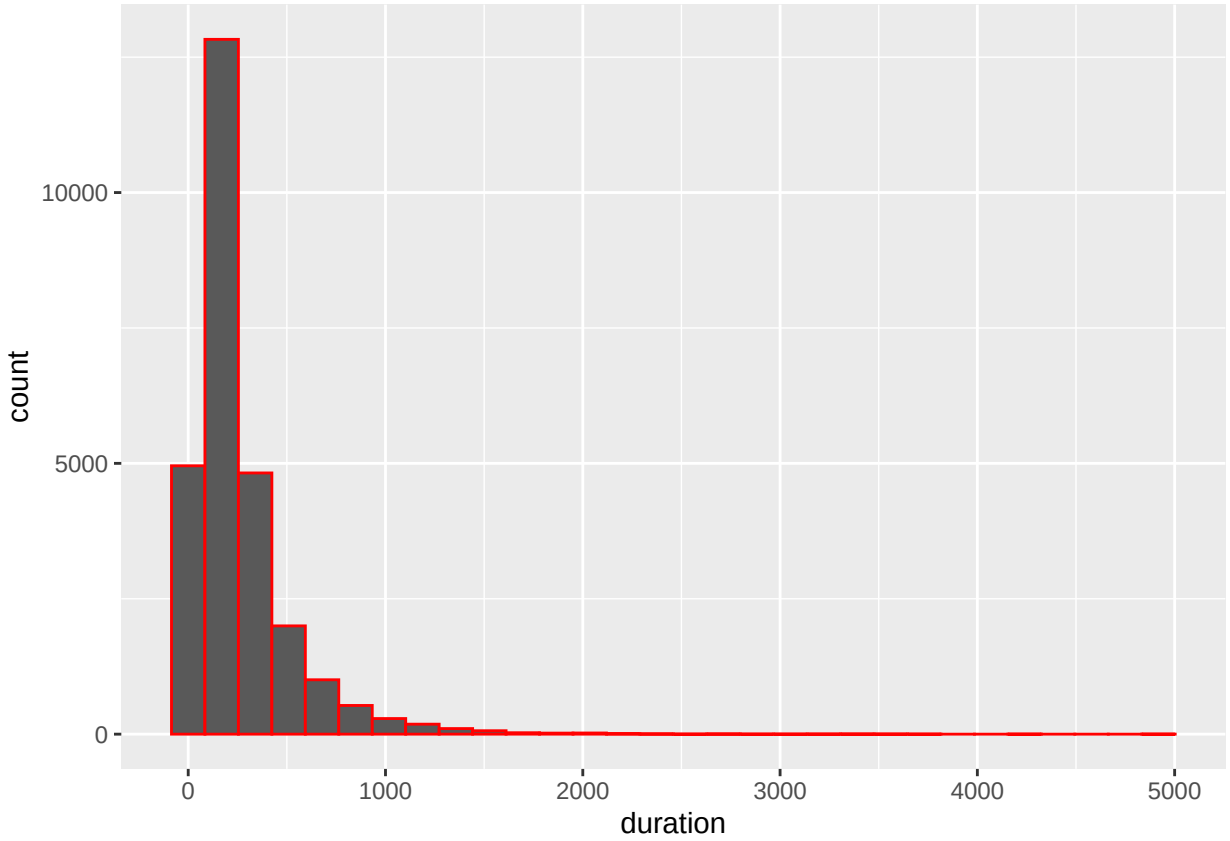
13.Produce the following graphs. What is the strength of each graph? Weakness?

a.Histogram of duration.

```
ggplot(bank_train, aes(duration)) + geom_histogram(aes(duration), color="red")
```

Answer:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



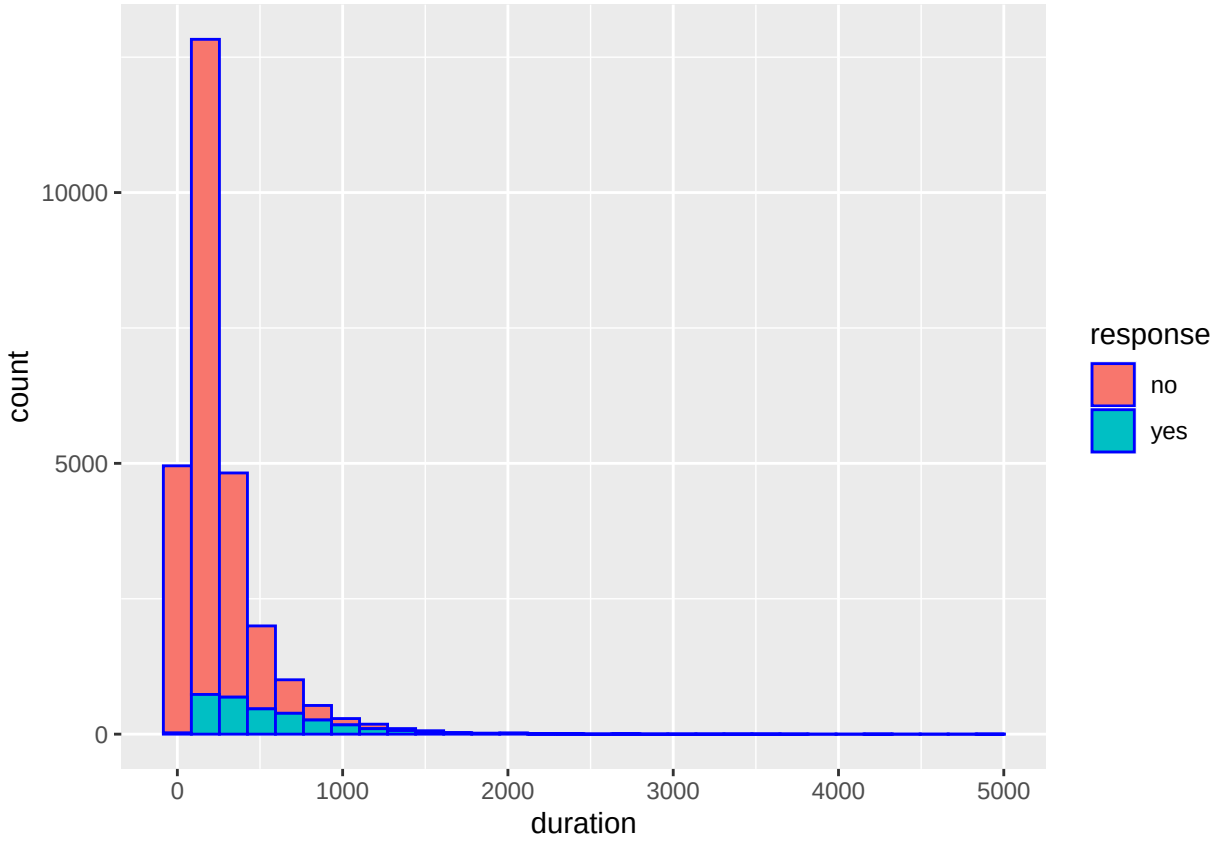
This graph shows the in how much duration people responded but there is no specific information about who responded yes or no.

b.Histogram of duration, with overlay of response.

```
ggplot(bank_train, aes(duration)) + geom_histogram(aes(fill=response), color="blue")
```

Answer:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



By looking at this histogram we can say that we have an information about the number of people responded for yes or no in what duration. Also the data is skewed distribution.

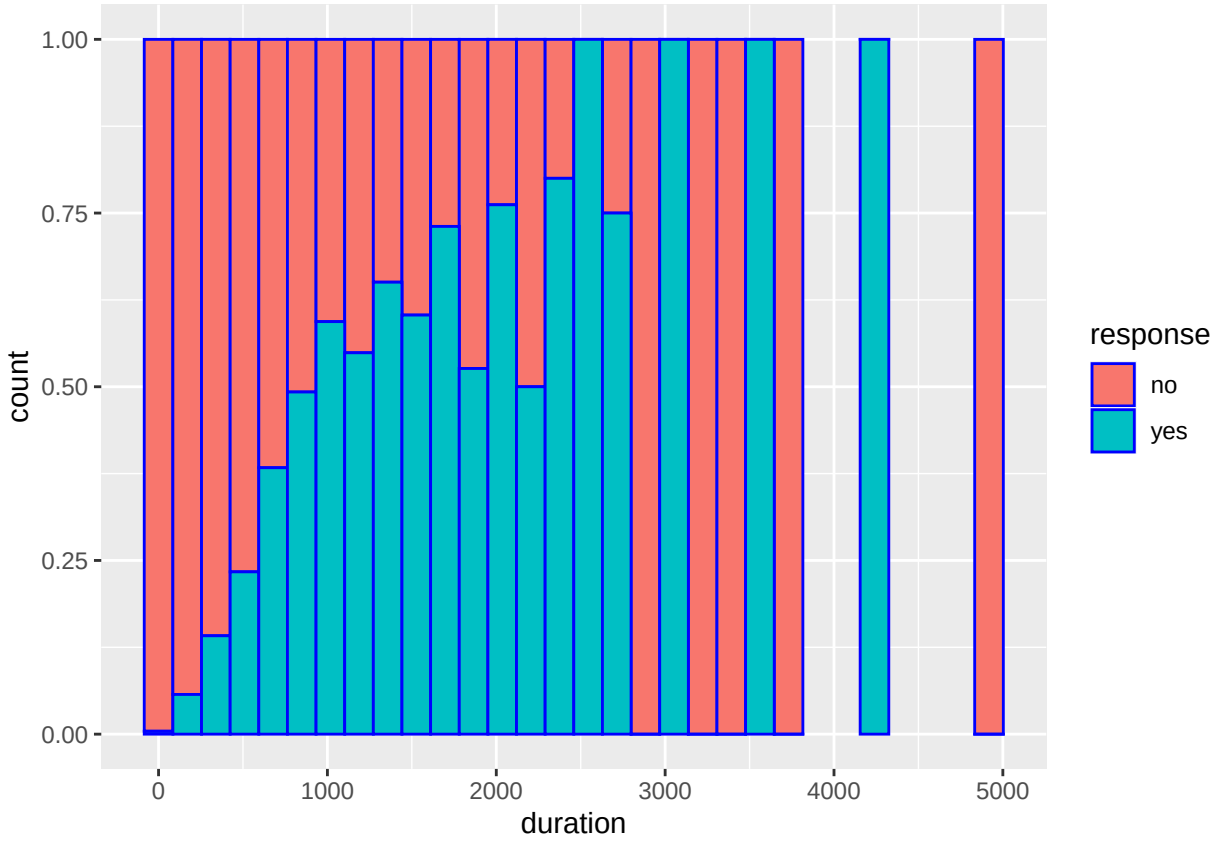
c.Normalized histogram of duration, with overlay of response.

```
ggplot(bank_train, aes(duration)) + geom_histogram(aes(fill=response), color="blue", position="fill")
```

Answer:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



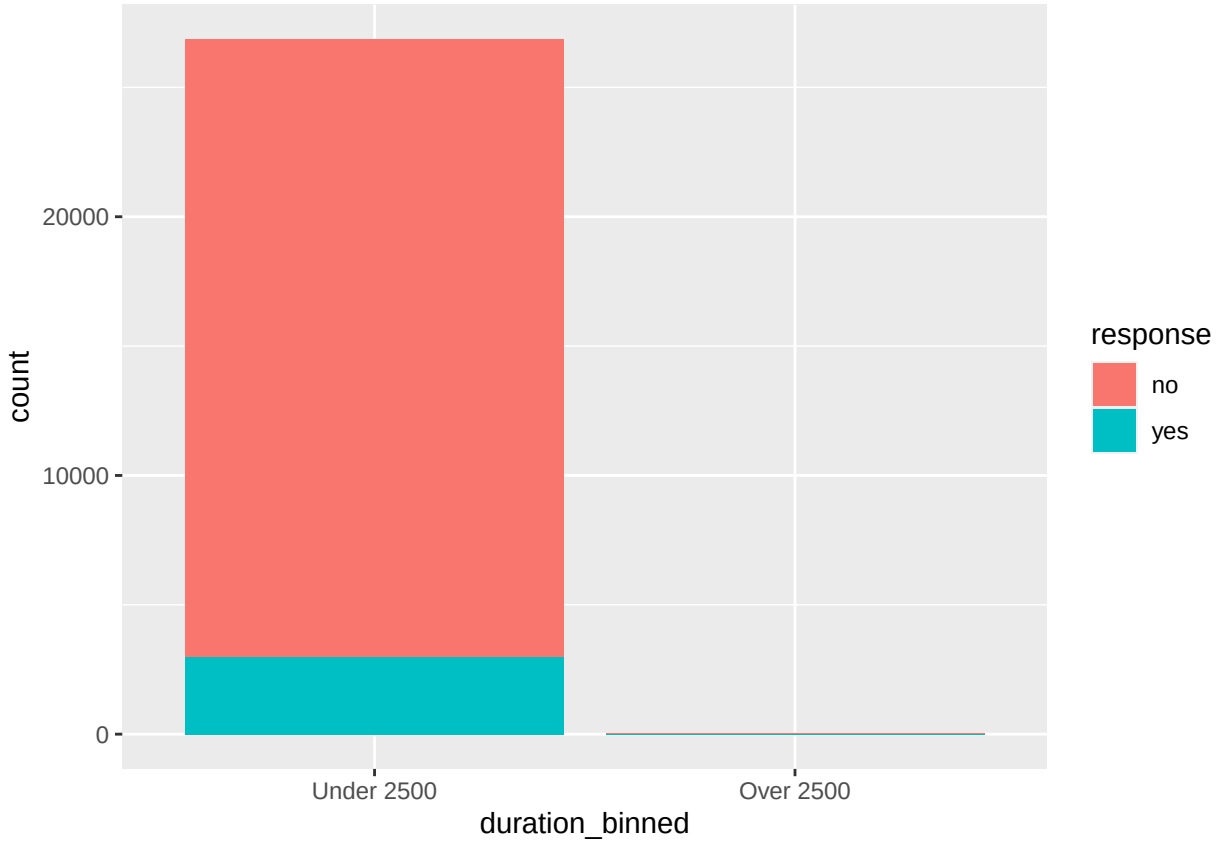
This histogram clearly shows data for all the duration counts. The normalized graph gives more clarity about the mean, median of the data.

14.Using the graph from Exercise 13c, describe the relationship between duration and response.

Answer: When duration is more the number of people responded more.

15.Examine the non-normalized and normalized histograms of duration, with overlay of response. Identify cutoff point(s) for duration, which separate low values of response from high values. Define a new categorical variable, duration_binned, using the cutoff points you identified.

```
bank_train$duration_binned <- cut(x = bank_train$duration, breaks = c(0, 2500, 5000),
right = FALSE, labels = c("Under 2500", "Over 2500"))
ggplot(bank_train, aes(duration_binned)) + geom_bar(aes(fill = response))
```



Answer:

The non-normalized histogram is skewed distribution. The normalized histogram clearly shows data for all the duration counts. Duration under 2500 is shown but there is a no duration for over 2500. Categorical variables are yes and no. Duration_binned are uder 2500 and over 2500. There are two cutoffs: 0-2500 and 2500-5000.