P8105_HW1_ml4418_Rmd

Mengyuan Li

9/15/2019

problem 1-part 1

```
data1_df = tibble(
  sample = rnorm(8),
  logic = c(sample > 0),
  cha_vector = c('I' , 'am', 'the', 'best', 'girl', 'in', 'the','world'),
  factor_vector = factor(c('good', 'best', 'fair', 'fair', 'good', 'good', 'best', 'best'))
)
```

Answer: When I take mean of each variable in data1_df, only varible of sample works, but logical, character and factor variables do not work.

Answer: When I covert each variable in data1_df to numerical variable, logical variable can be converted to 0 or 1; Character variable cannot be converted; Factor variables can be converted to numbers, which are denoted as levels of factor variables.

problem 1-part 2

problem 2-part 1

```
data2_df = tibble(
  x = rnorm(500),
  y = rnorm(500),
  logic = c(x + y > 1),
  number = as.numeric(logic),
  factor = as.factor(logic)
)
nrow(data2_df)
```

```
## [1] 500
```

```
ncol(data2_df)
```

```
## [1] 5
```

```
mean(pull(data2_df, x ))
```

```
## [1] -0.02736575
```

```r
median(pull(data2_df,x ))
```

```
## [1] -0.03651225
```

```r
sd(pull(data2_df, x ))
```

```
## [1] 1.045611
```

```r
count = sum(pull(data2_df, logic), na.rm = TRUE)
proportion = count/500
```

Answer: The size of dataset is (500,5); The mean of x is 0.009588525; The median of x is 0.001172065; The standard deviation of x is 1.008548; The proportion of cases for which x+y>1 is 0.23.

problem 2- part 2

```r
# use logical variable
plot1_df = ggplot(data2_df, aes(x=x, y=y, color=logic)) + geom_point(stat="identity")
# use numeric variable
plot2_df = ggplot(data2_df, aes(x=x, y=y, color=number)) + geom_point(stat="identity")
# use factor varaible
plot3_df = ggplot(data2_df, aes(x=x, y=y, color=factor)) + geom_point(stat="identity")
# save first plot
ggsave('plot1_df.pdf', height = 6, width = 7)
```

Answer: In the first plot, colors are decided by logical variables. More specifically, we use different colors to represent different logical variables. For example, in my case, red color represents 'false' and blue color represents "true". In the second plot, colors are decided by numeric variables. More specifically, we use different colors to represent different numbers. In my case, I use light blue and dark blue to represent 0 and 1. In the third plot, colors are decided by factor variables. More specifically, we use different colors to represent different factor variables. In my exmaple, I use red color to represent factor of "false", and use blue color to represent factor of "true".

‘‘‘