p8105\_hw1\_yx2507

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 0.8.3 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()

# problem 1

## making data frame

Create a data frame comprised of: a random sample of size 8 from a standard Normal distribution; a logical vector indicating whether elements of the sample are greater than 0; a character vector of length 8; a factor vector of length 8, with 3 different factor “levels”. Below shows the code and the result:

problem\_1=tibble(  
 vec\_numeric=rnorm(8),  
 vec\_logical=vec\_numeric>0,  
 vec\_char=c("a","b","c","d","e","f","g","h"),  
 vec\_factor = factor(c("M","F","M","F","M","F","M","F"))  
  
 )   
#CHECK DATA FRAME  
problem\_1

## # A tibble: 8 x 4  
## vec\_numeric vec\_logical vec\_char vec\_factor  
## <dbl> <lgl> <chr> <fct>   
## 1 -0.997 FALSE a M   
## 2 0.856 TRUE b F   
## 3 0.115 TRUE c M   
## 4 -0.443 FALSE d F   
## 5 -0.915 FALSE e M   
## 6 0.319 TRUE f F   
## 7 0.0651 TRUE g M   
## 8 0.0694 TRUE h F

## now take the mean

The result shows that numeric vairable and logical variable work while factor and character vairable can’t get the mean.

mean(problem\_1$vec\_numeric)

## [1] -0.1164166

mean(problem\_1$vec\_logical)

## [1] 0.625

mean(problem\_1$vec\_char)

## Warning in mean.default(problem\_1$vec\_char): argument is not numeric or  
## logical: returning NA

## [1] NA

mean(problem\_1$vec\_fator)

## Warning: Unknown or uninitialised column: 'vec\_fator'.

## Warning in mean.default(problem\_1$vec\_fator): argument is not numeric or  
## logical: returning NA

## [1] NA

## Conercion

Tn the below steps: \* convert the logical vector to numeric, and multiply the random sample by the result \* convert the logical vector to a factor, and multiply the random sample by the result \* convert the logical vector to a factor and then convert the result to numeric, and multiply the random sample by the result

as.numeric(problem\_1$vec\_char)

## Warning: NAs introduced by coercion

## [1] NA NA NA NA NA NA NA NA

as.numeric(problem\_1$vec\_logical)

## [1] 0 1 1 0 0 1 1 1

as.numeric(problem\_1$vec\_factor)

## [1] 2 1 2 1 2 1 2 1

coercision can change charactor vairable into N/A but can’t convert to numeric variable. So it can’t be calculated mean. While for logical variable, when converting to numerica vairables, “True” and “False” can be covert to 1 and 0 respectively. For factor variable, they can convert to the number indicating its level.

## multiply by random sample

as.numeric(problem\_1$vec\_logical)\*rnorm(8)

## [1] 0.0000000 -0.6706991 -1.1794537 0.0000000 0.0000000 0.8443095  
## [7] -0.1427310 0.5461193

as.factor(problem\_1$vec\_logical)\*rnorm(8)

## Warning in Ops.factor(as.factor(problem\_1$vec\_logical), rnorm(8)): '\*' not  
## meaningful for factors

## [1] NA NA NA NA NA NA NA NA

as.numeric(as.factor(problem\_1$vec\_logical))\*rnorm(8)

## [1] 0.8163104 -0.1836304 -1.6689403 -2.0147178 -0.1458189 0.1565454  
## [7] -1.8536203 2.0245858

# Problem 2 Including Plots

## Create data frame

Create a data frame comprised of: x: a random sample of size 500 from a standard Normal distribution y: a random sample of size 500 from a standard Normal distribution A logical vector indicating whether x + y > 1 A numeric vector created by coercing the above logical vector A factor vector created by coercing the above logical vector

set.seed(2)  
 x=rnorm(500)  
 y=rnorm(500)  
 vec\_logical\_plot= x + y > 1  
 vec\_num = as.numeric(vec\_logical\_plot)  
 vec\_fact = as.factor(vec\_logical\_plot)  
   
plot\_df=tibble(x,y,vec\_logical\_plot,vec\_fact,vec\_num)  
  
head(plot\_df)

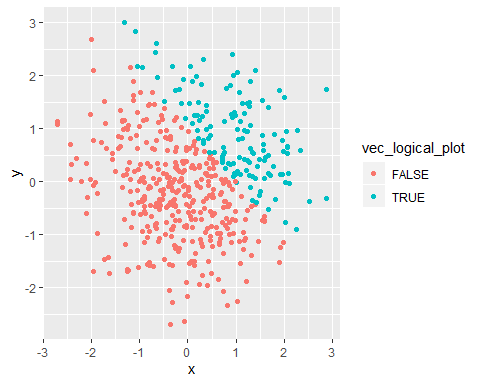
## # A tibble: 6 x 5  
## x y vec\_logical\_plot vec\_fact vec\_num  
## <dbl> <dbl> <lgl> <fct> <dbl>  
## 1 -0.897 -0.460 FALSE FALSE 0  
## 2 0.185 0.618 FALSE FALSE 0  
## 3 1.59 -0.720 FALSE FALSE 0  
## 4 -1.13 -0.584 FALSE FALSE 0  
## 5 -0.0803 0.216 FALSE FALSE 0  
## 6 0.132 1.24 TRUE TRUE 1

the number of rows in plot\_df is 500 columns is 5. The mean of x in the dataframe is 0.0616923. The median of the sample is 0.0439172.The standard deviation of the sample is1.0323776 The proportion of x+y>1 is 0.266

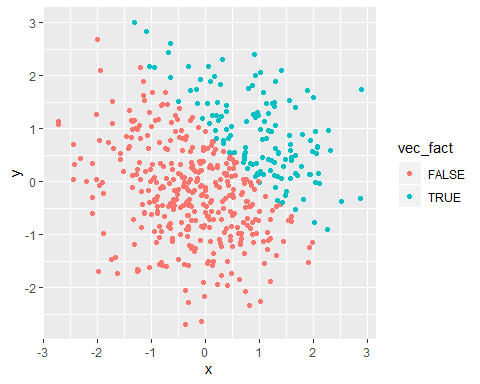
## create scatterplot

color the logical vairable

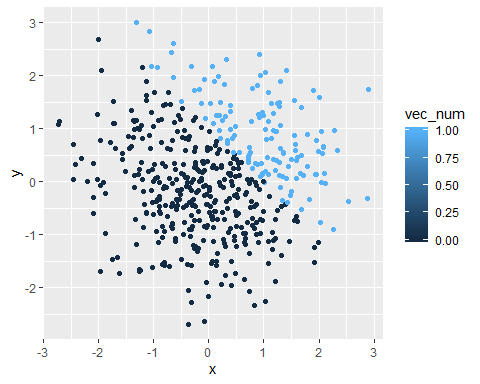
ggplot(plot\_df, aes(x = x, y = y, color=vec\_logical\_plot)) + geom\_point()

 color the factor variable

ggplot(plot\_df, aes(x = x, y = y, color=vec\_fact)) + geom\_point()

 colort the numeric varaible

ggplot(plot\_df, aes(x = x, y = y, color=vec\_num)) + geom\_point()



## export the this scatterplot

ggsave("scatter\_plot.pdf",height=4,width=6)