CSCI-B565: Data Mining

Homework 3

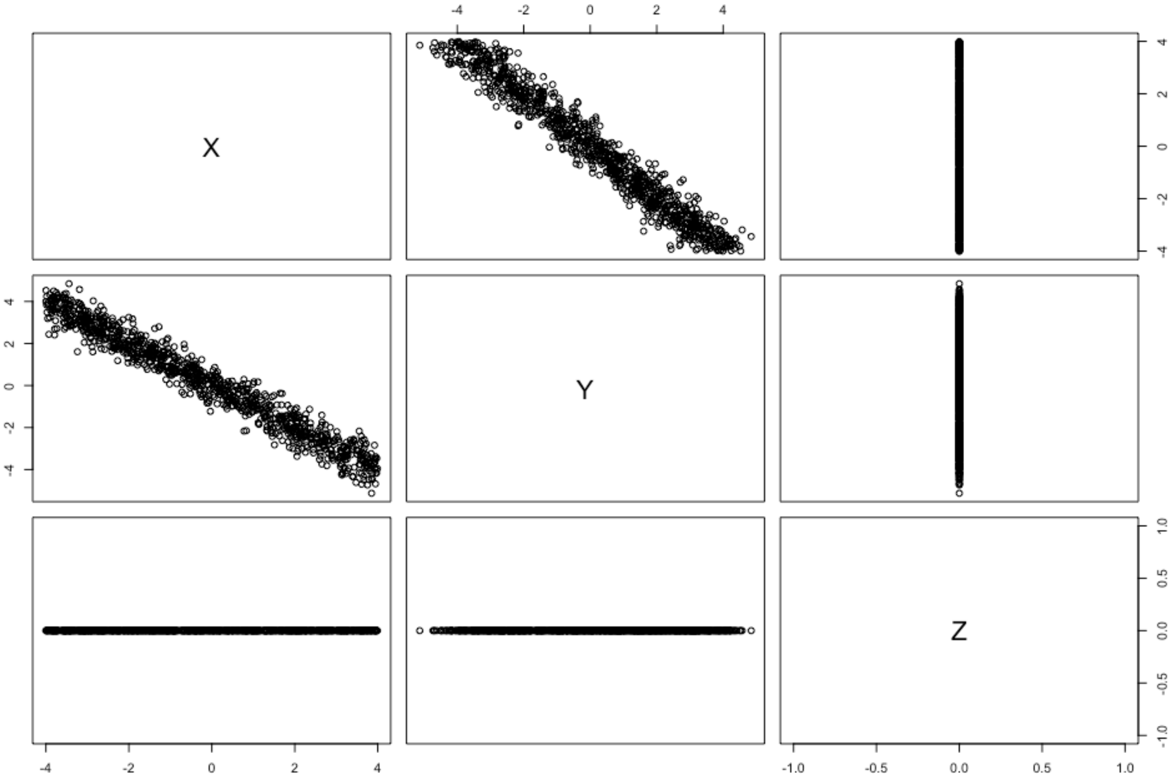
Shubhavi Arya

Instructor’s Name: Christopher Raphael

March 2, 2022

Problem 1:

```
# set random seed
set.seed(123)

# generate dataset
X <- runif(n = 1000,
        min = -4,
        max = 4)

Y <- -X + rnorm(n=1000, 0, 0.5)

Z<- rnorm(n=1000, 0, 0)

data <- data.frame(X,Y,Z)
# plot observations
plot(X,
    Y,
    type = "p",
    main = "A Scatterplot of X and Y",
    col = "steelblue",
    pch = 19)

pairs(~X+Y+Z, data)
```



```
> cov(data)   ###Covariance matrix
          X         Y Z
X  5.318915 -5.323203 0
Y -5.323203  5.580256 0
Z  0.000000  0.000000 0
```

Diagonalization of matrix:

$$U = \begin{pmatrix} 0 & -0.97575\ldots & 1.02484\ldots \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{10.899171 + \sqrt{113.41425\ldots}}{2} & 0 \\ 0 & 0 & \frac{10.899171 - \sqrt{113.41425\ldots}}{2} \end{pmatrix}$$

$$U^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ -0.49984\ldots & 0.51226\ldots & 0 \\ 0.49984\ldots & 0.48773\ldots & 0 \end{pmatrix}$$

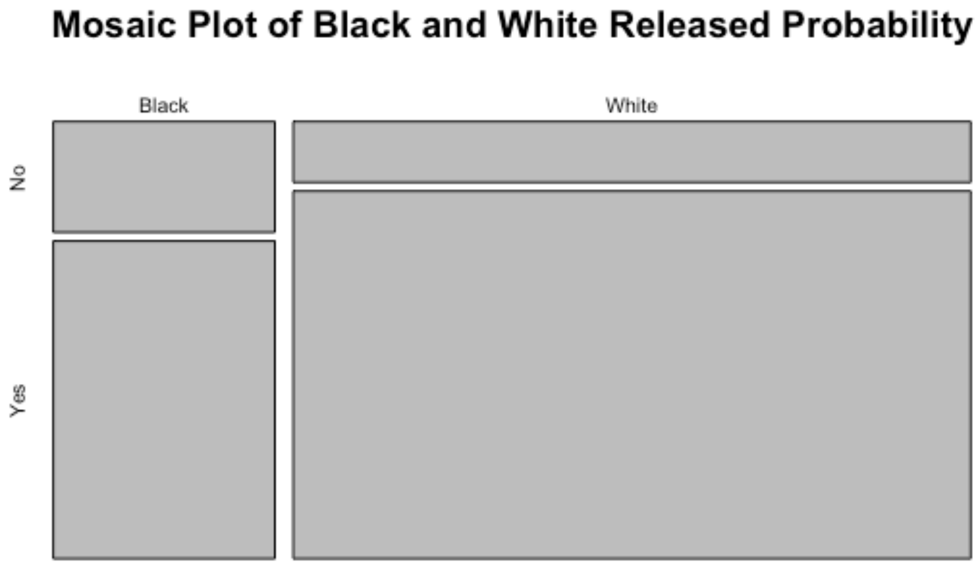
Problem 2:

Part a

```
###Q2


arrests = as.matrix(read.csv("Arrests.csv"))


###Part a
white <- arrests[which(arrests[,3] == "White"), ]
wmat <- data.matrix(white)
wh_released<- wmat[,c(2)]


black <- arrests[which(arrests[,3] == "Black"), ]
bmat <- data.matrix(black)
bl_released<- bmat[,c(2)]
```

```
White_Yes <- which(wh_released== "Yes")
Black_Yes <- which(bl_released== "Yes")

White_prob = length(White_Yes)/length(wh_released)
Black_prob = length(Black_Yes)/length(bl_released)

arrests_mos = read.csv("Arrests.csv")
table_prob <- table(arrests_mos$colour, arrests_mos$released)
mosaicplot(table_prob)
```

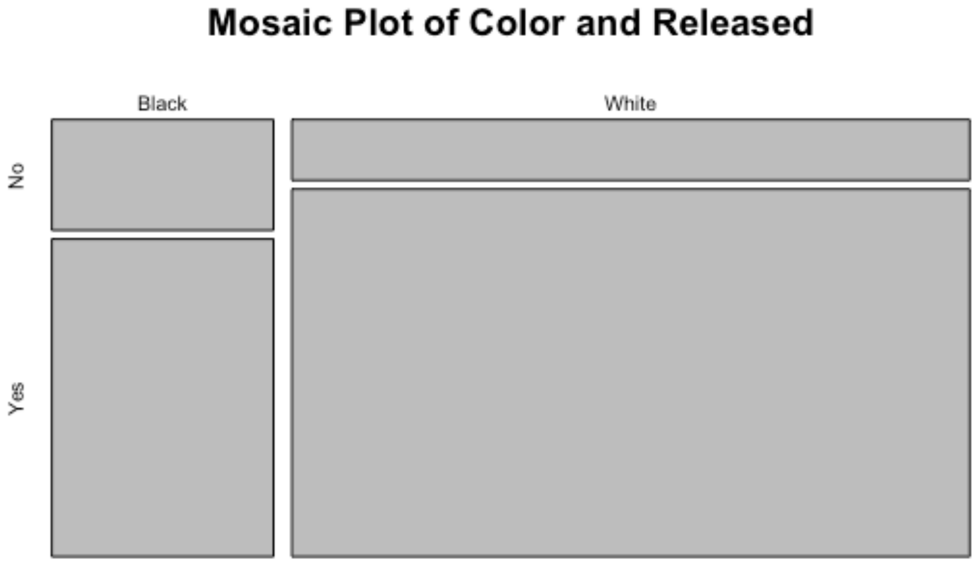
```
> print(White_prob)
[1] 0.8580498
> print(Black_prob)
[1] 0.7414596
```

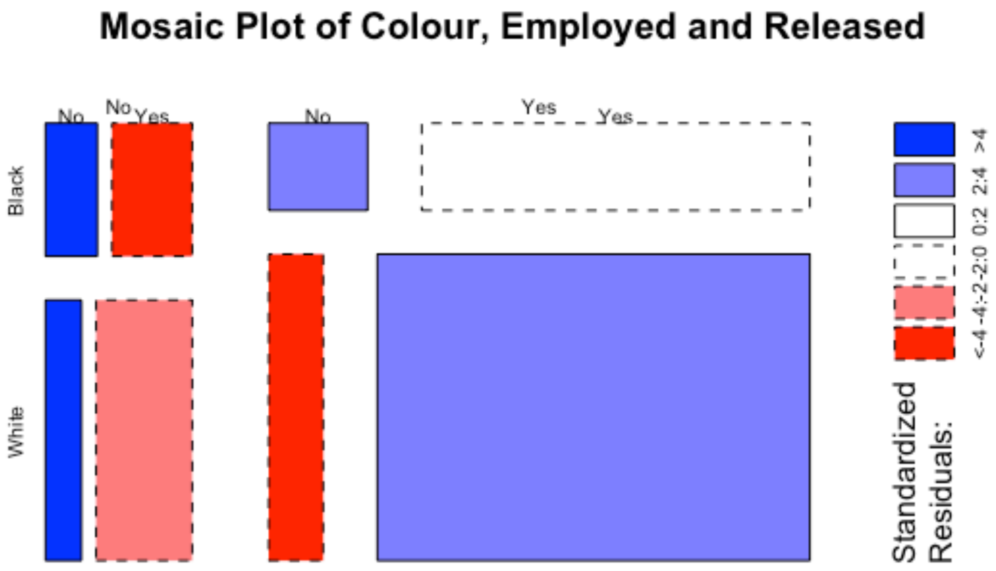


Summary: As seen in the mosaic plot and calculations with R above, White population has a slightly higher chance (0.86) of getting released than black population (0.74).

Part b

```
table_col_rel_emp <- table(arrests_mos$employed, arrests_mos$colour, arrests_mos$released )
mosaicplot(table_col_rel_emp, shade = TRUE, margin = list(1:2, 3), main="Mosaic Plot of Colour, Employed and Released")
> fm <- loglin(table_col_rel_emp, list(1:2, 3))
2 iterations: deviation 0
> pchisq(fm$pearson, fm$df, lower.tail = FALSE)
[1] 3.146218e-58
```

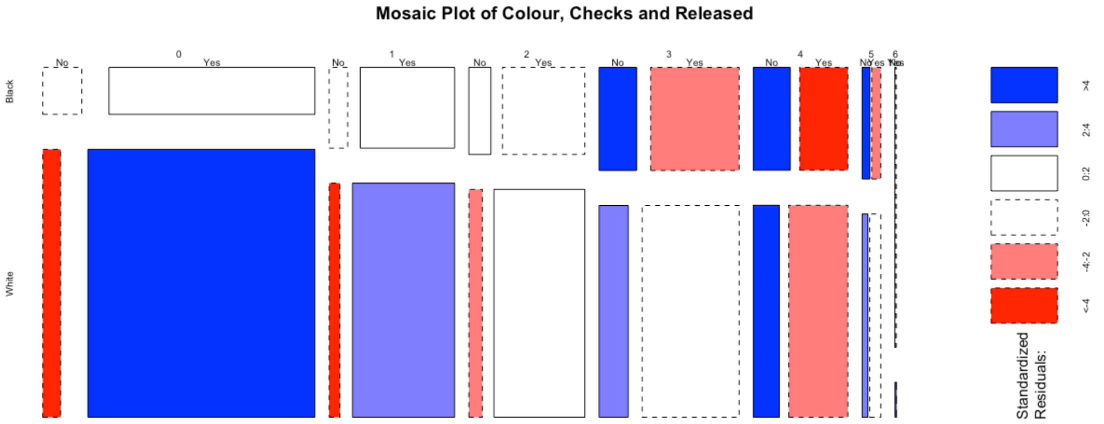




We can see that there is a difference in the release rates between the “Blacks” and “Whites” when mosaic plot is plotted without the employment categories. But as we can see from the mosaic plot above, there is no difference among the variable outcomes in each employment category when they are taken into account. Specifically, there is no difference in release rate between “Blacks” and “Whites” among **each employment category** according to the mosaic plot which suggests that there is not a racial bias in the release decisions and the variables “released” and “color” are conditionally independent given employment.

Part c

```
> table_col_rel_check <- table(arrests_mos$checks, arrests_mos$colour, arrests_mos$released )
> mosaicplot(table_col_rel_check, shade = TRUE, margin = list(1:2, 3),
main="Mosaic Plot of Colour, Checks and Released")
> fmb <- loglin(table_col_rel_check, list(1:2, 3))
2 iterations: deviation 5.684342e-14
> pchisq(fmb$pearson, fmb$df, lower.tail = FALSE)
[1] 1.941944e-77
```


Mosaic Plot of Colour, Checks and Released

As we can see from the mosaic plot above, there appears to little difference among the variable outcomes in each Check category when they are taken into account. Specifically, there is no difference in release rate between “Blacks” and “Whites” among **each check category** according to the mosaic plot which suggests that there is not a racial bias in the release decisions and the variables “released” and “color” are conditionally independent given checks.

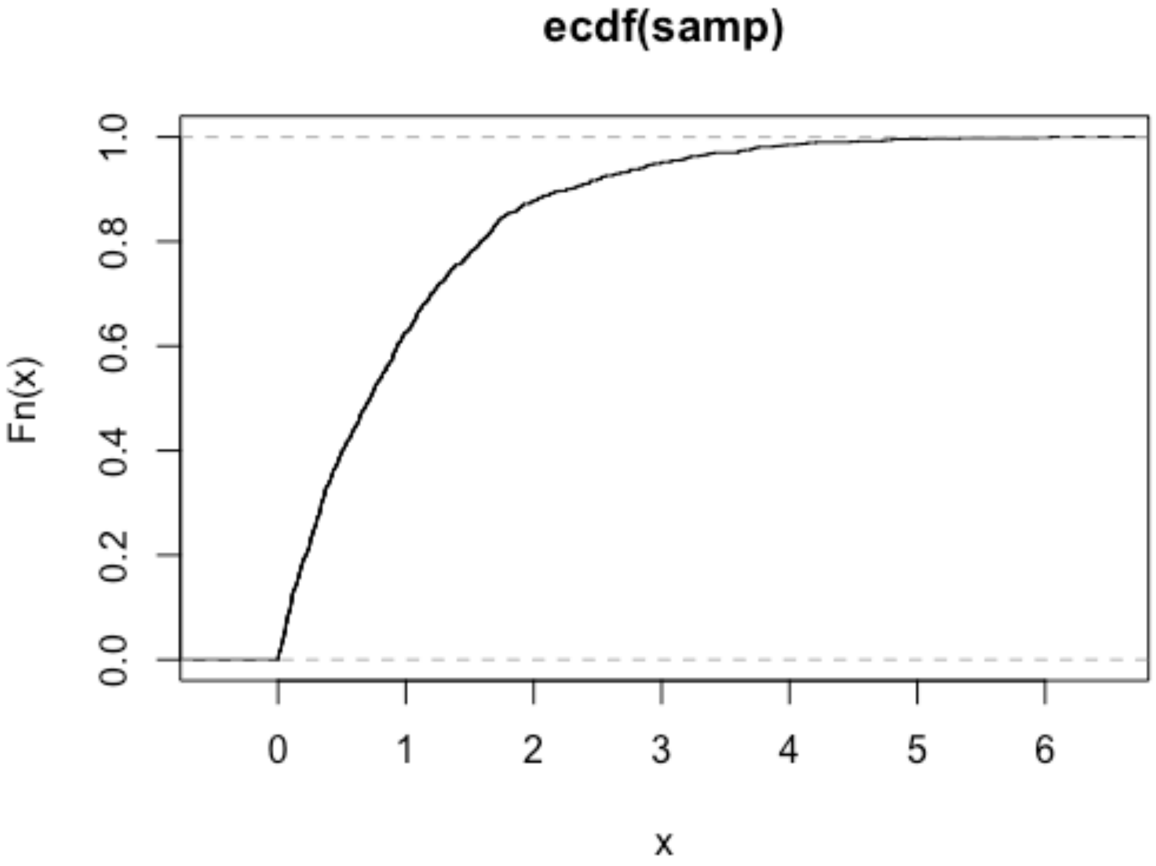
Part d

The data does not appear to be consistent with racial bias when other factors are taken into account. However when the racial identity/ color and release decision are solely correlated there is some racial bias visible. What we can see is that there is not a clear differentiation or inclination towards one specific race when other factors such as checks and employment category are considered - making release decision and color conditionally independent. What we can also see is that there is not a large difference in the probability of a person getting released among racial identities (<0.2). It can be inferred that therefore, the data is not fully consistent with racial bias however there may be some patterns/ cases that could be looked further to investigate it more.

Problem 3:

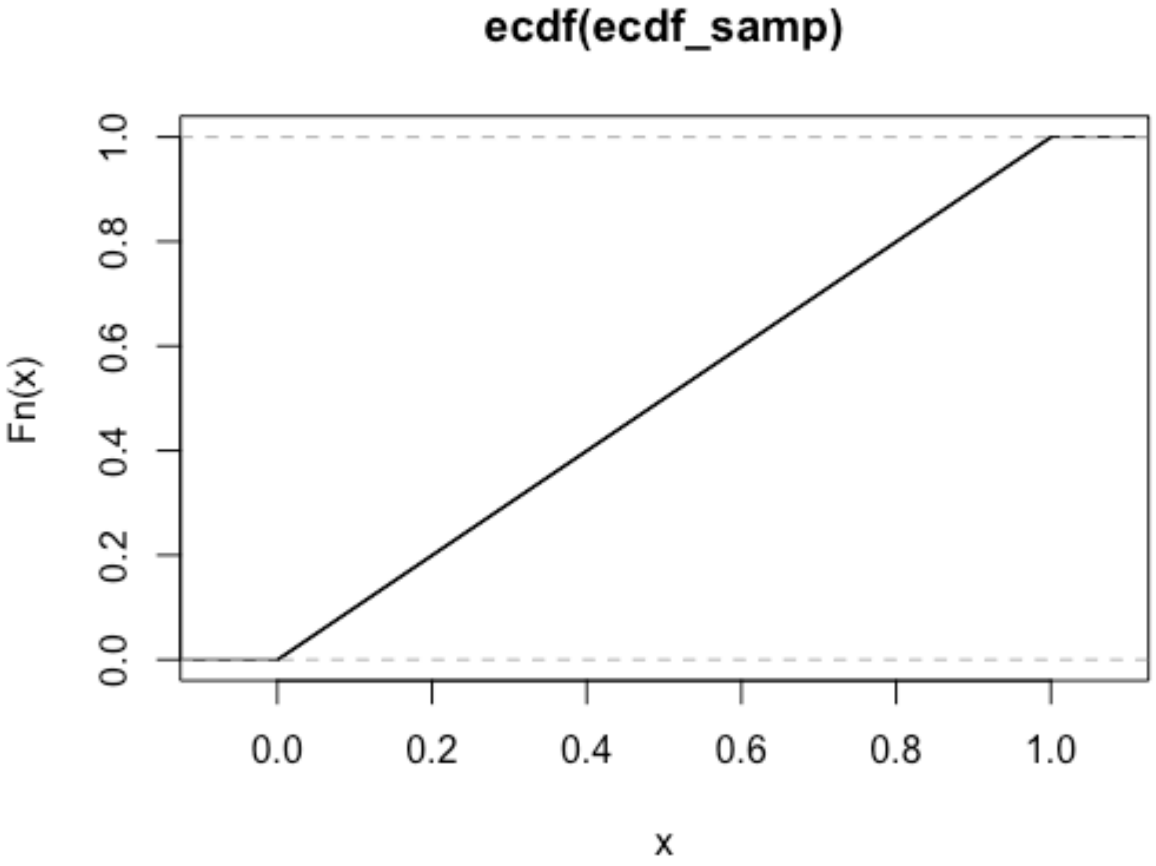
Part a

```
samp <- rexp(1000)
plot(ecdf(samp))
```



Part b

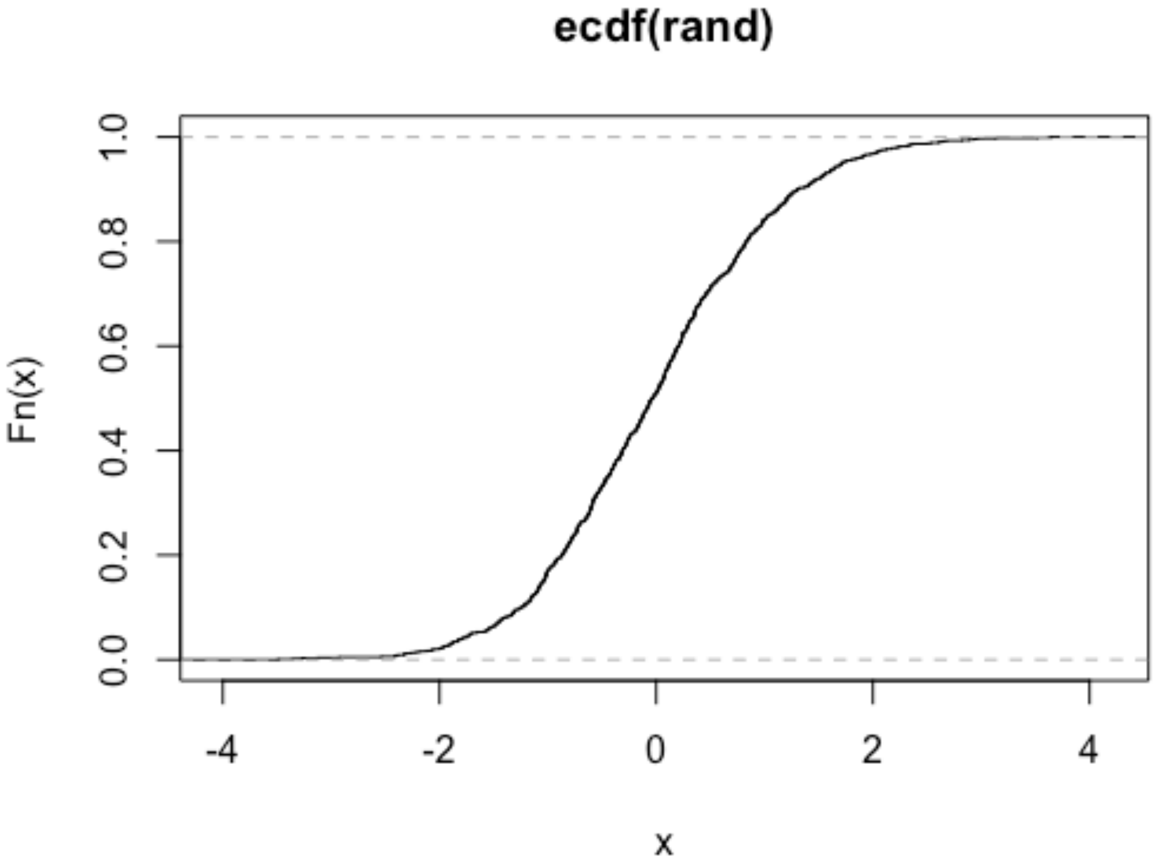
```
ecdf_func <- ecdf(samp)


## New data set
ecdf_samp <- data.frame(lapply(samp,ecdf_func))


## CDF of new data set
ecdf_samp <- as.vector(ecdf_samp)
mynewcdf <-ecdf(ecdf_samp)
plot(mynewcdf)
```



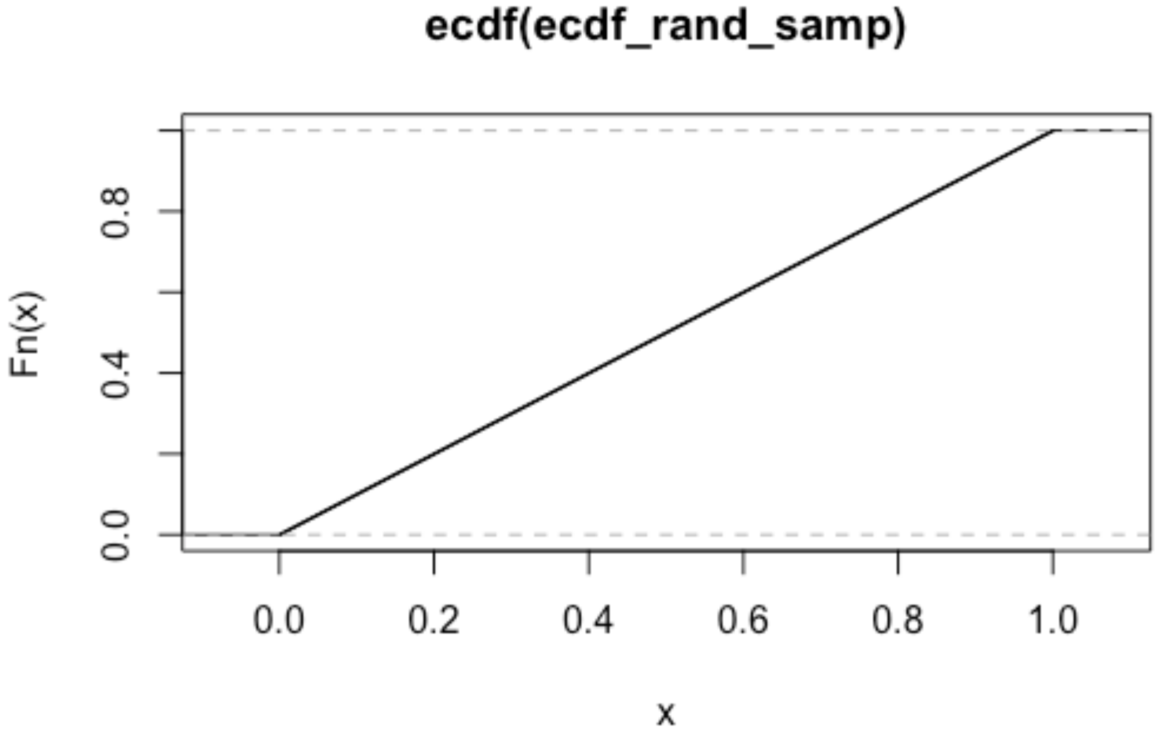
Part c

The results of the transformed points will still have the same distribution as in the previous part even if the numbers are random from a different distribution because applying ecdf(ecdf(x)) = ecdf(ecdf(y)). Applying CDF to CDF of its previous results gives the same result, no matter what sample we take. We can prove it with a random sample from a normal distribution below:

```
rand <- rnorm(1000)
plot(ecdf(rand))
```



```
ecdf_func_rand <- ecdf(rand)
ecdf_rand_samp <- data.frame(lapply(rand,ecdf_func_rand))

ecdf_rand_samp <- as.matrix(ecdf_rand_samp)
ecdf_rand_samp <- ecdf_rand_samp[1,]
ecdf_rand_samp <- as.numeric(ecdf_rand_samp)
mynewcdf_rand <- ecdf(ecdf_rand_samp)
plot(mynewcdf_rand)
```



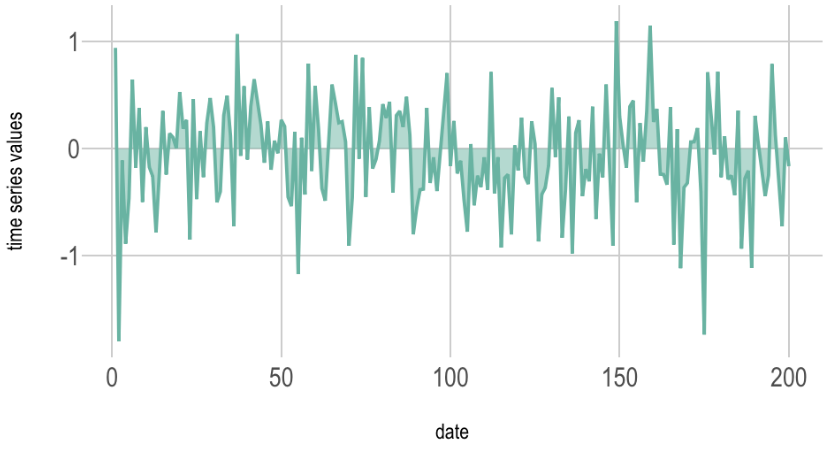
As we can see above, the final plot is the same as the final plot in Part (b). Hence proved, both verbally and graphically.

Problem 4:

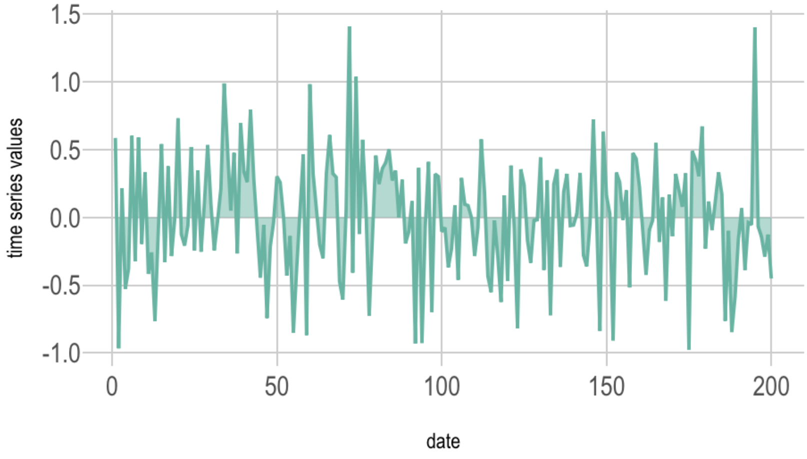
Part a

```
time_series <- read.csv("time_series.csv")
# Libraries
library(ggplot2)
library(dplyr)
library(plotly)
library(hrbrthemes)
```

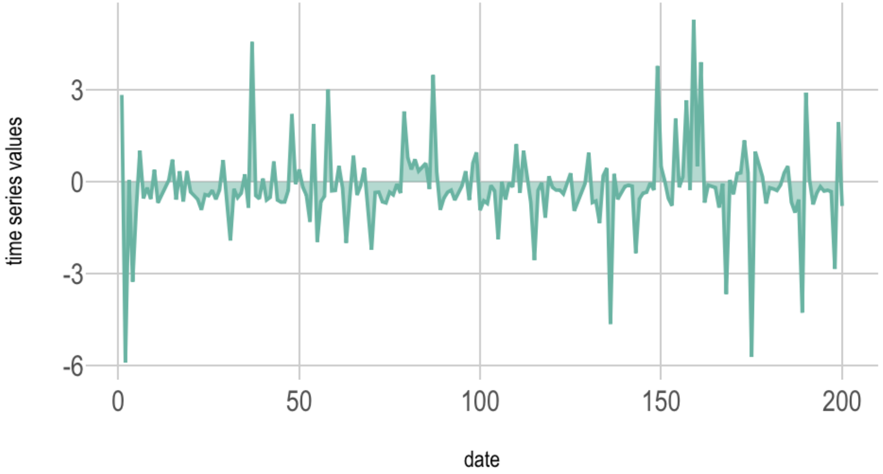
```
# Load dataset
data <- time_series
date <- (1:200)
z <- sample(date,1)
var_names <- names(data)
a<- get(var_names[z], data)


# Usual area chart
p <- data %>%
  ggplot( aes(x=date, y=a)) +
  geom_area(fill="#69b3a2", alpha=0.5) +
  geom_line(color="#69b3a2") +
  ylab(“time series values")+
  theme_ipsum()

# Turn it interactive with ggplotly
p <- ggplotly(p)
p
```



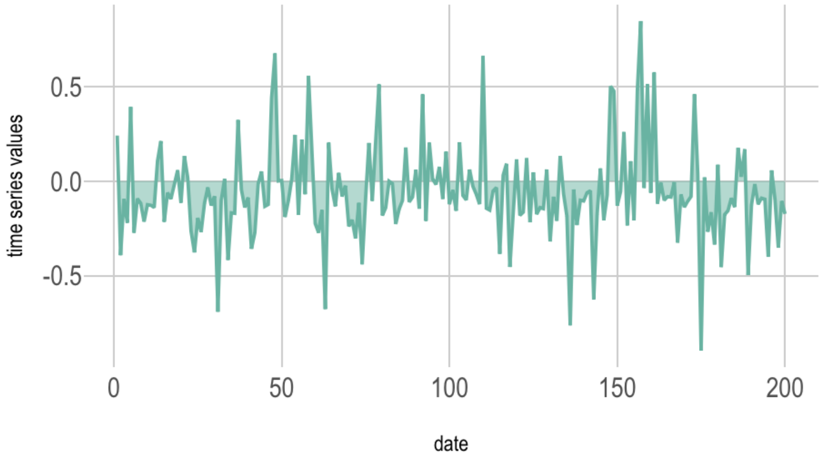
timeSer1.html : V30



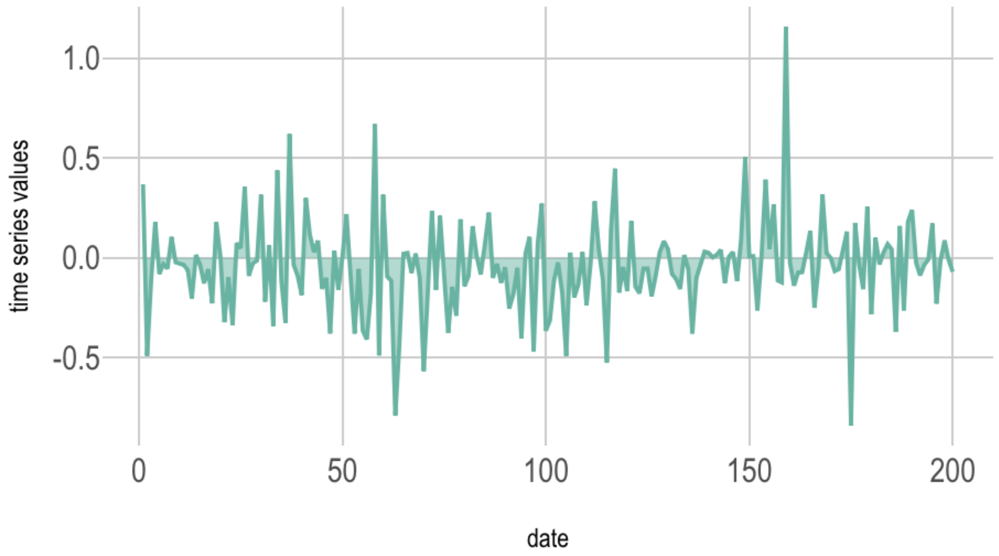
timeSer5.html : V32



timeSer2.html : V15



timeSer3.html : V72



timeSer4.html : V90

We can see above that models V72 and V90 follow the same pattern. The rest of the time series all follow a different pattern and form a different category.

The four categories of time series would be:

1. Irregular time series
2. Time series with a seasonality pattern
3. Time Series with a long term trend without calendar related or irregular effects
4. Time Series with a cyclical pattern i.e. periodical but not seasonal variations/ patterns

Part b

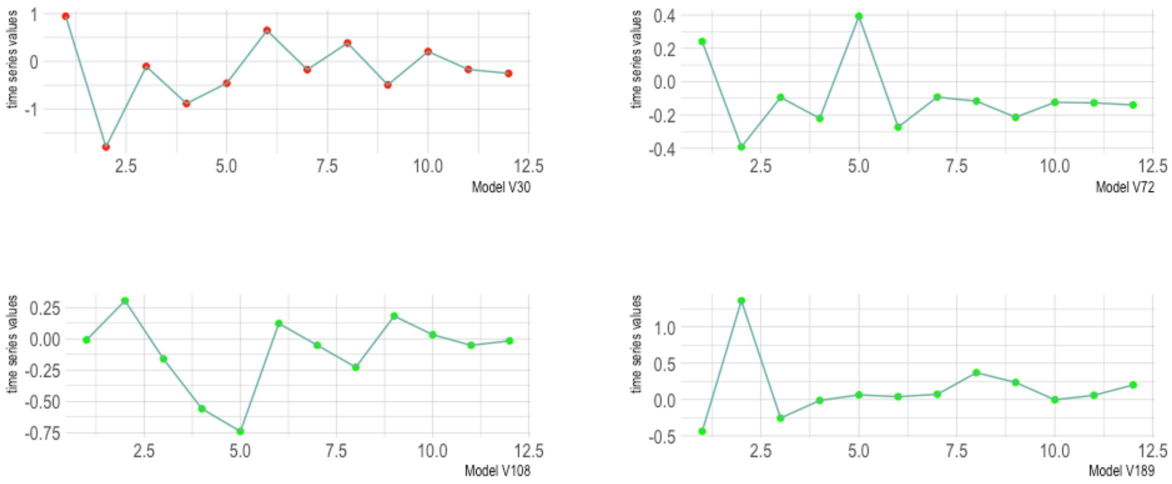
Two features that effectively separate the time series into four categories when shown in a scatterplot are:

1. Seasonality

Here we look at the patterns of time series models in 12 time stamps (observations)

Code

```
####Seasonality (12 months)
library(gridExtra)

#V30
ndata <- time_series[1:12,]
newdate <- (1:12)
py<- ndata$V30
p1 <- ndata %>%
  ggplot( aes(x=newdate, y=py)) +
  geom_point(color="red") + geom_line(color="#69b3a2")+
  ylab("time series values")+xlab("Model V30")+
  theme_ipsum()
p1

##V72
newdate <- (1:12)
py4<- ndata$V72
p4 <- ndata %>%
  ggplot( aes(x=newdate, y=py4)) +
  geom_point(color="green") + geom_line(color="#69b3a2")+
  ylab("time series values")+xlab("Model V72")+
  theme_ipsum()
p4

##V108
newdate <- (1:12)
py5<- ndata$V108
p5 <- ndata %>%
  ggplot( aes(x=newdate, y=py5)) +
  geom_point(color="green") + geom_line(color="#69b3a2")+
  ylab("time series values")+xlab("Model V108")+
  theme_ipsum()
p5

##V108
newdate <- (1:12)
py6<- ndata$V59
p6 <- ndata %>%
  ggplot( aes(x=newdate, y=py6)) +
  geom_point(color="green") + geom_line(color="#69b3a2")+
  ylab("time series values")+xlab("Model V189")+
  theme_ipsum()
p6

grid.arrange(p1,p4,p5,p6, nrow = 3)
```



We can see from the scatter plot (joined with a line curve) above over a 12 time stamp window that they all have different seasonal pattern, hence differentiating them by seasonality. All model time series in the dataset seem to follow these 4 model patterns closely.

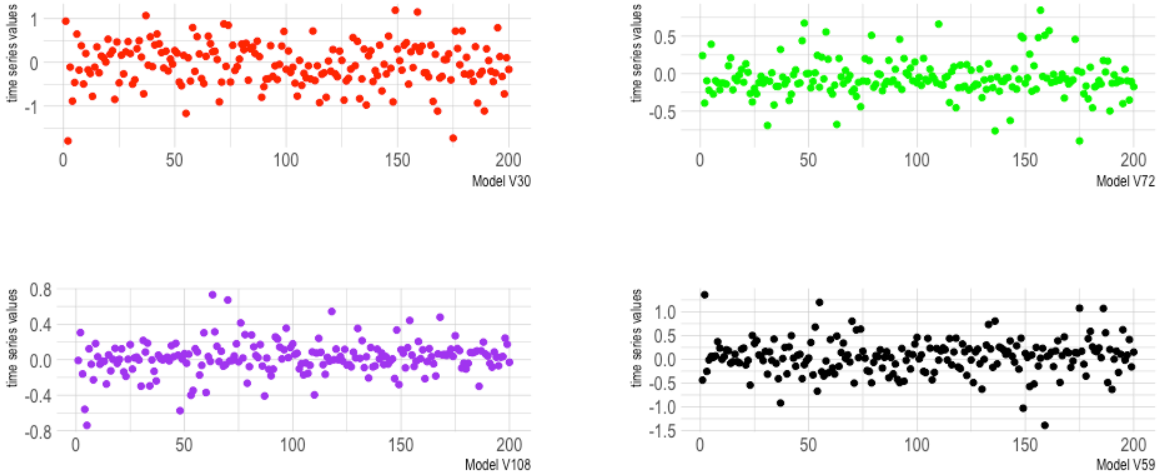
2. Noise (Here we look at random, unexplainable variation in time patterns) and Trends (Here we look at increasing/ decreasing patterns in time series)

Code

```
#V30
data <- time_series
newdate <- (1:200)
```

```
py<- data$V30
p1 <- data %>%
  ggplot( aes(x=newdate, y=py)) +
  geom_point(color="red") +
  ylab("time series values")+xlab("Model V30")+
  theme_ipsum()
p1

##V72
py4<- data$V72
p4 <- data %>%
  ggplot( aes(x=newdate, y=py4)) +
  geom_point(color="green")+
  ylab("time series values")+xlab("Model V72")+
  theme_ipsum()
p4

##V108
py5<- data$V108
p5 <- data %>%
  ggplot( aes(x=newdate, y=py5)) +
  geom_point(color="purple") +
  ylab("time series values")+xlab("Model V108")+
  theme_ipsum()
p5

##V59
```

```
py6<- data$V59

p6 <- data %>%

  ggplot( aes(x=newdate, y=py6)) +

  geom_point(color="black") +

  ylab("time series values")+xlab("Model V59")+

  theme_ipsum()

p6


grid.arrange(p1,p4,p5,p6, nrow = 3)
```

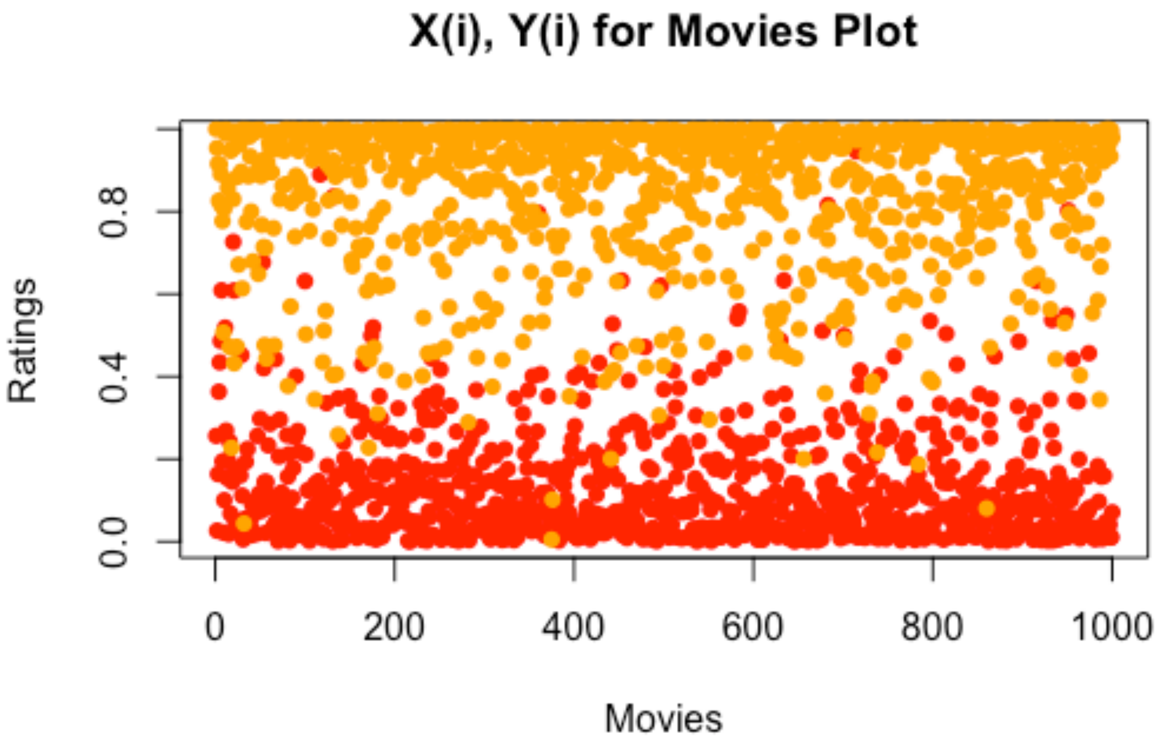


We can see from the above scatter plot how the 4 model time series are different in terms of their variation in noise as well as increasing/ decreasing trends, thus differentiating them by those two factors as well.

Problem 5

Part a

```
left_x <- rexp(1000,rate=8)
hist(left_x)
right_y <- rbeta(1000,2,0.5,ncp=2)
hist(right_y)

plot(left_x, col="red", pch=16, ylab = "Ratings", xlab = "Movies", main="X(i), Y(i) for Movies Plot")
points(right_y, pch=16, col=“orange”)
```



Part b

Two variables X and Y are independent if P(X=x, Y=y) = P(X=x)P(Y=y), for all x,y.

This is true for the two reviewers' score according to me since knowing the value of one reviewer's score does not influence or change the probability of the score value from the other reviewer. Therefore, the two reviewers do not have any interactions with each other when assigning the scores to the movies and hence their scores are independent of each other. So, it would be reasonable to model the two reviewers score as independent random variables.

Part c

Xi = score of reviewer 1

Yi = score of reviewer 2

Let us assume that for each movie,

$P(Xi > Yi)$

Then,

$P(Xi > Yi) = P(Xi) * (1-P(Yi))$

Xi needs to be higher and Yi should be lower every time.

When we repeat this for 1000 times, since both of our samples are random normal distributions, we will find that $P(Xi) - P(Yi) = 0.5$ which means that $P(Xi) > P(Yi)$. So this is possible that one reviewer always gives a higher percentile score than the other one.

Part d

Xi = score of reviewer 1

Yi = score of reviewer 2

As shown above, let us assume that for each movie,

$P(Xi > Yi)$ so that $Xi-Yi > 0$

Then,

$P(Xi > Yi) = P(Xi) * (1-P(Yi))$

Difference will be positive only when Xi will be higher than Yi.

When we repeat this for 1000 times, since both of our samples are random normal distributions, we will find that $P(Xi) - P(Yi) = 0.5$ on average

Part e

$P(Xi>Yi) + P(Yi>Xi)$

$= P(Xi>Yi)+ P(Yi>Xi)$

On average $P(Xi) - P(Yi) = 0.5$,

So on average, $P(Xi>Yi)= 0.5$

Now Let us take Yi to be 0.6 and Xi to be 0.2 in the case when Yi is > Xi

So,

P(Yi>Xi) = P(Yi)*P(1-P(Xi)) = 0.6*0.8 = 0.48

Therefore,

P(Xi>Yi)+ P(Yi>Xi)

= 0.5 + 0.48 = 0.98

So, it is possible that xi > yi for all but 1 movie.

Problem 6

It is given that D(ij) is the distance between (xi , yi) and (xj , yj).

We are also given that distance between a point and itself is 0.

D(ii) = distance between (xi , yi) and (xi , yi) = 0

Which means that for any value i, the value fo D(ii) = 0

Hence, the diagonal elements of the matrix should be 0.

0 is not positive and also not negative.

But we are given that D consists of only positive numbers and therefore it is not possible to create a matrix D such that D(ii) = 0

Hence, it is not possible to find a collection of n points as described that the Euclidean distance between (xi , yi) and (xj , yj) is Dij.

Problem 7

Multivariate Data using BoxPlots

Age: Can be plotted effectively using box plots. We can see easily the youngest to the oldest individuals as well as the all the quartiles and median age. It is also helpful to observe the relationship of age with different variables using box plots as well. The pattern with age using box plot is robust for different age groups and the differences are very well highlighted.

Weight: Weight can also be effectively plotted using box plots. We can see all the percentiles from 25th to 75th effectively. This is especially helpful to look at new born infants and sick individuals to effectively identify patterns and differences.

Height: Height can be plotted effectively using box plots as well. We can compare the heights of males and females as well as the distribution of heights by various groups. We can also look at the median, IQR and range of the heights of various groups effectively with a box plot.

Income: Income can be communicated very well using box plots. We can look at the IQR, range and median of income by various regions, groups and occupations clearly; also identify any outliers. It is also helpful to guide further multivariate analysis using procedures such as PCA.