A1

Yuqi Gu

QUESTION 1: Basic R Calculations

(a)

```
5^7
```

```
## [1] 78125
```

(b)

```
log(21, base = 2)
```

```
## [1] 4.392317
```

(c)

```
sum(sapply(1:1000, function(x) {cos(x + pi) / (x^2)} ))
```

```
## [1] -0.3241388
```

(d)

```
21 %% 2
```

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

(e)

```
A <- matrix(data = c(5, 6, 6, 3, 4, 6, 7, 10, 6, 4, 4, 5, 6, 2, 7, 9),
            nrow = 4, byrow = FALSE)
apply(A, 1, median)
```

```
## [1] 5.5 5.0 6.5 7.0
```

```r
apply(A, 2, mean)
```

```
## [1] 5.00 6.75 4.75 6.00
```

(f)

```r
apply(A, 1, function(x) {sum((x%%5)==0)} )
```

```
## [1] 1 0 0 2
```

(g)

```r
x <- seq(-2, 2, by=0.00001)
f <- x^4+x^3+2*x^2+3*x+4
c(x[which.min(f)], min(f))
```

```
## [1] -0.750000  2.769531
```

(h)

```r
nearest_neighbour <- function(v, x) {
  differences <- abs(v - x)
  positions <- which(differences == min(differences))
  return(positions)
}

nearest_neighbour(v = c(7, 10, 5, 10, 14, 2, 11, 8, 13, 8), x = 9)
```

```
## [1]  2  4  8 10
```

QUESTION 2: Investigating the Proportion of Non-Negatives

(a)

For any $b \in \mathbb{R}$,

$$\begin{aligned} a(y_1+b, y_2+b, \ldots, y_N+b) &= \frac{1}{N}\sum_{u\in\mathcal{P}} I_{[0,\infty)}(y_u+b) \\ &\neq \frac{1}{N}\sum_{u\in\mathcal{P}} I_{[0,\infty)}(y_u) \\ &\neq \frac{1}{N}\sum_{u\in\mathcal{P}} I_{[0,\infty)}(y_u)+b \end{aligned}$$

Thus, the proportion attribute $a(\mathcal{P})$ is neither location invariant nor location equivariant.

(b)

For any m > 0,

$$\begin{aligned} a(m\times y_1, m\times y_2, \ldots, m\times y_N) &= \frac{1}{N}\sum_{u\in\mathcal{P}} I_{[0,\infty)}(m\times y_u) \\ &= \frac{1}{N}\sum_{u\in\mathcal{P}} I_{[0,\infty)}(y_u) \\ &= a(y_1, y_2, \ldots, y_N) \\ &= a(\mathcal{P}) \end{aligned}$$

Thus, the proportion attribute $a(\mathcal{P})$ is scale invariant.

(c)

Let

$$\mathcal{P}^k = \{\underbrace{y_1, y_1, \ldots, y_1}_{k}, \underbrace{y_2, y_2, \ldots, y_2}_{k}, \ldots, \underbrace{y_N, y_N, \ldots, y_N}_{k}\} = \{x_1, x_2, \ldots, x_{Nk}\}$$

$$\begin{aligned} a(\mathcal{P}^k) &= \frac{1}{Nk}\sum_{j=1}^{Nk} I_{[0,\infty)}(x_j) \\ &= \frac{1}{Nk}\sum_{i=1}^{N} k\times I_{[0,\infty)}(y_i) \\ &= \frac{1}{Nk}\times k\sum_{i=1}^{N} I_{[0,\infty)}(y_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} I_{[0,\infty)}(y_i) \\ &= a(\mathcal{P}) \end{aligned}$$

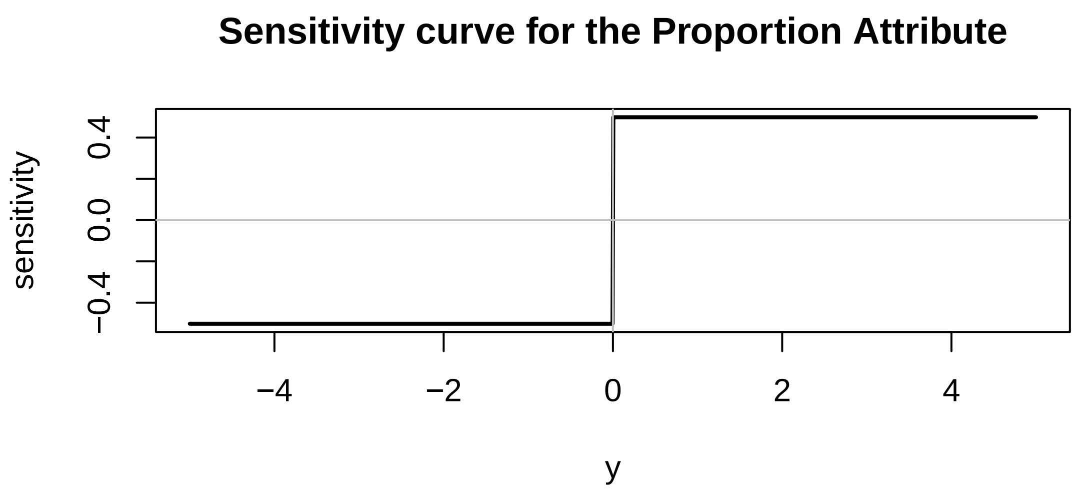
Thus, the proportion attribute $a(\mathcal{P})$ is replication invariant.

(d)

$$
\begin{aligned}
SC(y; a(\mathcal{P})) &= N[a(y_1, \dots, y_N - 1, y) - a(y_1, \dots, y_{N-1})] \\
&= N\left[\frac{1}{N}\left(\sum_{i=1}^{N-1} I_{[0,\infty)}(y_i) + I_{[0,\infty)}(y)\right) - \frac{1}{N-1}\sum_{i=1}^{N-1} I_{[0,\infty)}(y_i)\right] \\
&= \left(\sum_{i=1}^{N-1} I_{[0,\infty)}(y_i) + I_{[0,\infty)}(y)\right) - \frac{N}{N-1}\sum_{i=1}^{N-1} I_{[0,\infty)}(y_i) \\
&= I_{[0,\infty)}(y) - \frac{1}{N-1}\sum_{i=1}^{N-1} I_{[0,\infty)}(y_i) \\
&= I_{[0,\infty)}(y) - a(\mathcal{P})
\end{aligned}
$$

(e)

```r
sc = function(y.pop, y, attr, ...) {
  N <- length(y.pop) +1
  sapply( y, function(y.new) {  N*(attr(c(y.new, y.pop),...) - attr(y.pop,...))  } )
}

set.seed(341)
P <- rnorm(1000)
y <- seq(-5,5, length.out=1000)

calculate_proportion <- function(y){
  mean(y>=0)
}

plot(y, sc(P, y, calculate_proportion), type="l", lwd = 2,
     main="Sensitivity curve for the Proportion Attribute",
     ylab="sensitivity")
abline(h=0, v=0, col="grey")
```



QUESTION 3: k-Nearest Neighbour Classifier

```r
setwd('/Users/yuki/Desktop/STAT 341/Assignments/A1')
data <- read.csv("q3data.csv")

knn <- function(x, y, k) {
  n <- length(x)
  y_pred <- numeric(n)

  for (i in 1:n) {
    x_remain <- x[-i]
    distances <- abs(x_remain - x[i])
    y_remain <- y[-i]
    k_nearest <- sort(distances)[k]
    y_pred[i] <- mean(y_remain[distances <= k_nearest])>=0.5
  }
  return(y_pred)
}
knn(data$x, data$y, 3)
```

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

QUESTION 4: Kendrick vs. Drake

(a)

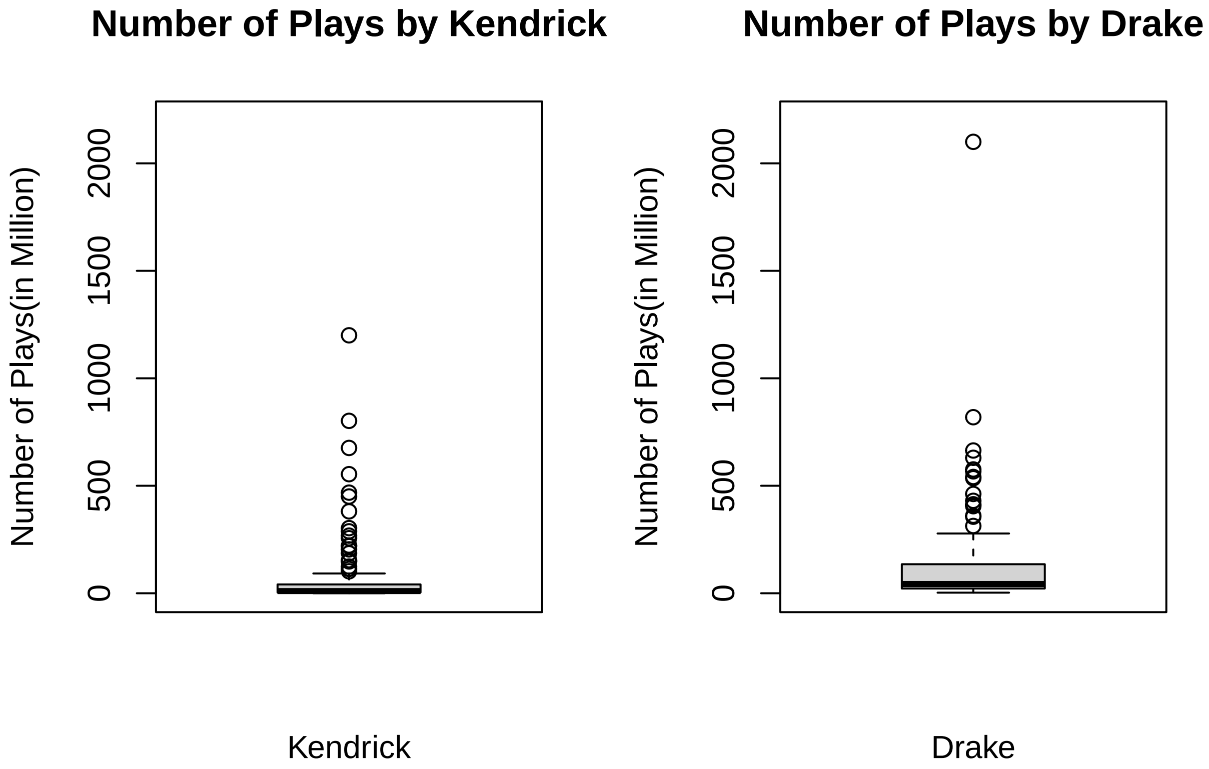
```
kvd <- read.csv("kendrick_v_drake.csv")
kvd$KorD <- c(rep(1,128), rep(0,141))
summary(kvd$Plays[1:128])
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 6.800e+04 3.550e+06 1.000e+07 6.935e+07 3.950e+07 1.200e+09
```

```
summary(kvd$Plays[129:269])
```

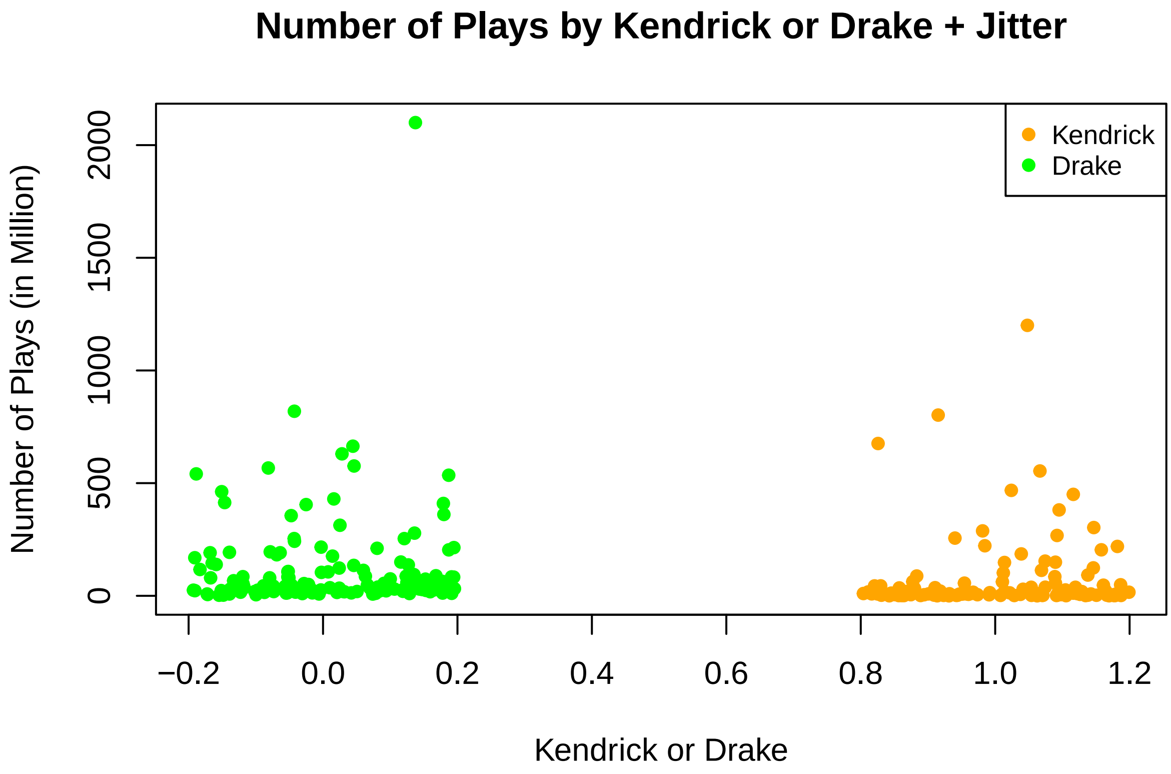
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 2.700e+06 2.200e+07 4.300e+07 1.256e+08 1.350e+08 2.100e+09
```

```
par(mfrow=c(1,2))
boxplot((kvd$Plays[1:128]/1000000), data = kvd, xlab="Kendrick", ylab =
          "Number of Plays(in Million)", ylim = c(0, 2200),
        main = "Number of Plays by Kendrick")
boxplot((kvd$Plays[129:269]/1000000), data = kvd, xlab="Drake", ylab =
          "Number of Plays(in Million)", ylim = c(0, 2200),
        main = "Number of Plays by Drake")
```



(b)

```
plot(jitter(kvd$KorD, factor = 1), jitter(kvd$Plays/1000000, factor = 1),
     main = "Number of Plays by Kendrick or Drake + Jitter",
     pch = 19, cex = 0.8,
     col = ifelse(kvd$KorD == 1, adjustcolor("orange", alpha = 3),
                  adjustcolor("green", alpha = 3)),
     xlab = "Kendrick or Drake",
     ylab = "Number of Plays (in Million)",
     type = "p"
)
legend("topright", c("Kendrick","Drake"), pch = 19, cex = 0.8, col=c("orange", "green"))
```



(c)

From part (a), comparing the summary of Plays for Drake and Kendrick, we find that IQR of Plays for Drake is larger than the IQR of Plays for Kendrick. Furthermore, the box plots also show the same result. Additionally, the largest number of Plays for Drake is much larger than the largest number of Plays for Kendrick. These indicate that Drake is more popular than Kendrick. From part (b), we observe that the points for Drake centers around 0 and the points for Kendrick centers around 1. This is because we use KorD = 1 to represent "Kendrick" and KorD = 0 to represent "Drake". Generally, the number of Plays for Drake and Kendrick are similar by observing the scatter plots in part (b). There is a large outlier for the number of Plays for Drake at the top left, which corresponds to the box plots in part (a) and the summary output.

(d)

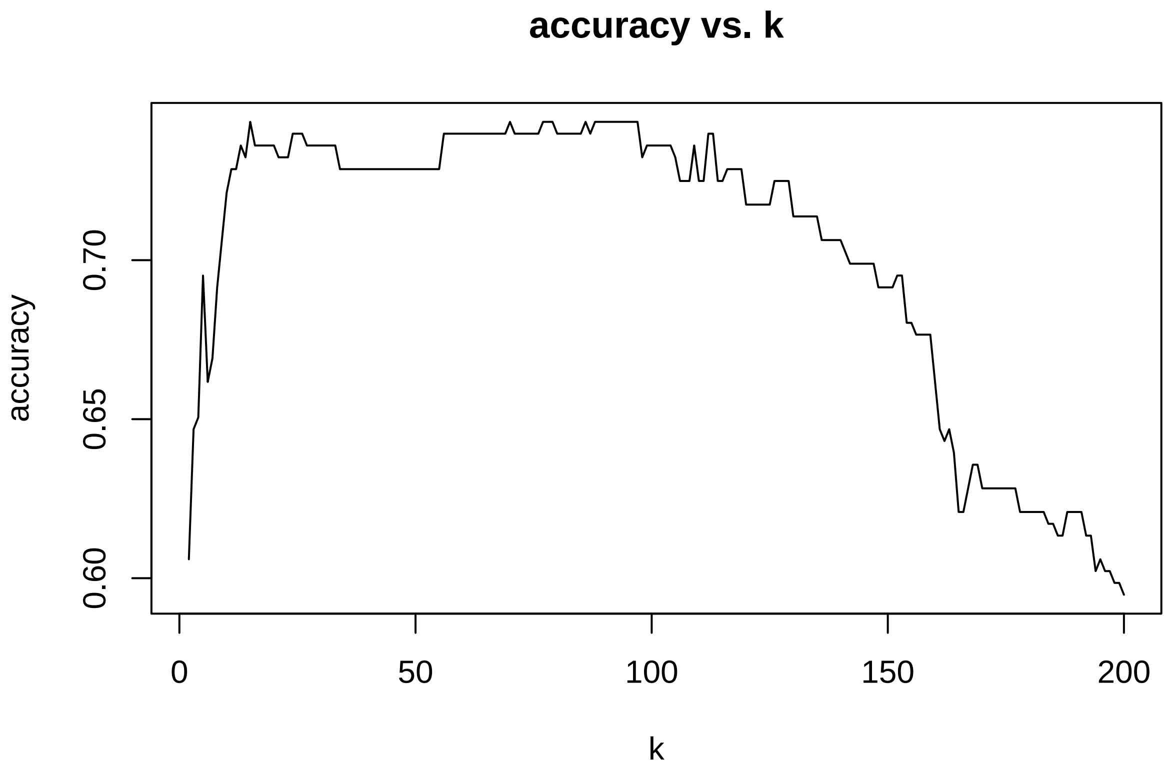
```r
y_new <- knn(kvd$Plays/1000000, kvd$KorD, 5)

TP <- sum((kvd$KorD==1)&(y_new==1))
FP <- sum((kvd$KorD==0)&(y_new==1))
FN <- sum((kvd$KorD==1)&(y_new==0))
TN <- sum((kvd$KorD==0)&(y_new==0))
cat("TP:", TP, "FP:", FP, "FN:", FN, "TN:", TN)
```

```
## TP: 84 FP: 38 FN: 44 TN: 103
```

		Prediction	
		$\hat{y}_u = 1$	$\hat{y}_u = 0$
Truth	$y_u = 1$	84	44
	$y_u = 0$	38	103

(e)

```r
accuracy <- rep(0, 199)
x <- kvd$Plays/1000000
y <- kvd$KorD
for (k in 2:200){
  y_new <- knn(x, y, k)
  accuracy[k-1] <- sum(y_new == y)/269
}
plot(2:200, accuracy, xlab="k", ylab="accuracy",
     main="accuracy vs. k", type="l")
```



```r
k <- 2:200
k_max_accuracy <- k[accuracy==max(accuracy)]
print(k_max_accuracy)
```

```
##  [1] 15 70 77 78 79 86 88 89 90 91 92 93 94 95 96 97
```

The values of k that appears to maximize the accuracy of the classification are 15, 70, 77, 78, 79, 86, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97.

(f)

The range of values of k appear optimal is roughly from 15 to 97. For the values of k that are between these two values, the trend of accuracy is generally stable and it is higher than the other parts when k is smaller than 15 or k is larger than 97. The impact of considering too many and too few neighbors is that the accuracy is low as shown in the plot from part (e) and it is not stable. There seems to be a noticeable difference in song plays between the two artists since the accuracy is above 0.6 generally overall and it is above 0.7 when we use optimal values of k. Because there are noticeable difference in song plays between Drake and Kendrick, the accuracy of using kNN to classify songs as “Kendrick” or “Drake” based on Plays/1000000 is generally high. This high accuracy shows that our algorithm is good at distinguishing which artist sang a given song just based on the number of plays of that song.