Student ID: 112077423

```
library(ggplot2)
library(compstatslib)
library(data.table)
library(tidyr)
library(lsa)
```

```
## Warning:     'lsa'           R      4.3.3
```

Question 1(a)

```
# alternative method to read data:
ac_bundles_dt <- fread("piccollage_accounts_bundles.csv")
ac_bundles_matrix <- as.matrix(ac_bundles_dt[, -1, with=FALSE])
```

In the app at the bottom of each sticker pack, we can find 6 recommendations.

Possible recommendations for *eastersurprise*:

- happyeaster2016
- hellospring
- holidaycheers
- springrose
- HeartStickerPack

Question 1(b)

(i)

cosine similarity

```
cosine_func <- function(mt) {

  cosine_matrix <- cosine(mt)

  cnt <- 1

  top <- matrix(ncol = 5, nrow=165)
  dimnames(top) <- list(colnames(cosine_matrix), c(1:5))

  while (cnt <= dim(cosine_matrix)[1]) {
    new_row <- tail(sort(cosine_matrix[cnt,]),6)
    top[cnt,] <- names(new_row[c(-6)])
    cnt <- cnt + 1
  }
```

```
  return(top)

}

#cos_top_5 <- cosine_func(ac_bundles_matrix)


cosine_recos <- function(items_matrix) {
  cos_sim_matrix <- qlcMatrix::cosSparse(items_matrix)
  bundle_names <- colnames(items_matrix)
  dimnames(cos_sim_matrix) <- list(bundle_names, bundle_names)
  diag(cos_sim_matrix) <- 2
  row_recos <- function(cos_sim_row) {
    names(sort(cos_sim_row, decreasing = TRUE))
  }
  all_recos <- t(apply(cos_sim_matrix, 2, row_recos))
  final_recos <- all_recos[, -1]
  return(final_recos[, 1:5])
}

cos_top_5 <- cosine_recos(ac_bundles_matrix)
```

```
cos_top_5['eastersurprise',]
```

```
## [1] "cutoutluv"      "bemine"         "watercolor"     "hipsterholiday"
## [5] "mmlm"
```

(ii)

correlation

```
bundle_means <- apply(ac_bundles_matrix, 2, mean)
bundle_means_matrix <- t(replicate(nrow(ac_bundles_matrix), bundle_means))
ac_bundles_mc_b <- ac_bundles_matrix - bundle_means_matrix
corr_top_5 <- cosine_recos(ac_bundles_mc_b)
```

```
corr_top_5['eastersurprise',]
```

```
## [1] "cutoutluv"        "bemine"           "watercolor"       "hipsterholiday"
## [5] "tropicalparadise"
```

(iii)

adjusted-cosine

```
bundle_means <- apply(ac_bundles_matrix, 1, mean)
ac_bundles_mc_b <- ac_bundles_matrix - bundle_means
adj_top_5 <- cosine_recos(ac_bundles_mc_b)
```

```
adj_top_5['eastersurprise',]
```

```
## [1] "cutoutluv"    "bemine"       "washiholiday" "ladolcevita"  "happy"
```

Question 1(c)

It is similar in a sense but not the same. Humans have biases when evaluating something.

Question 1(d)

Cosine similarity and correlation focus on bundle similarity whereas adjusted-cosine considers similarity between individuals.

Question 2(a)

(i) raw slope of x and y that is around 0

(ii) correlation of x and y that is around 0

Question 2(b)

(i) raw slope of x and y that is around 0

(ii) correlation of x and y that is around 0

Question 2(c)

(i) raw slope of x and y that is close to 1

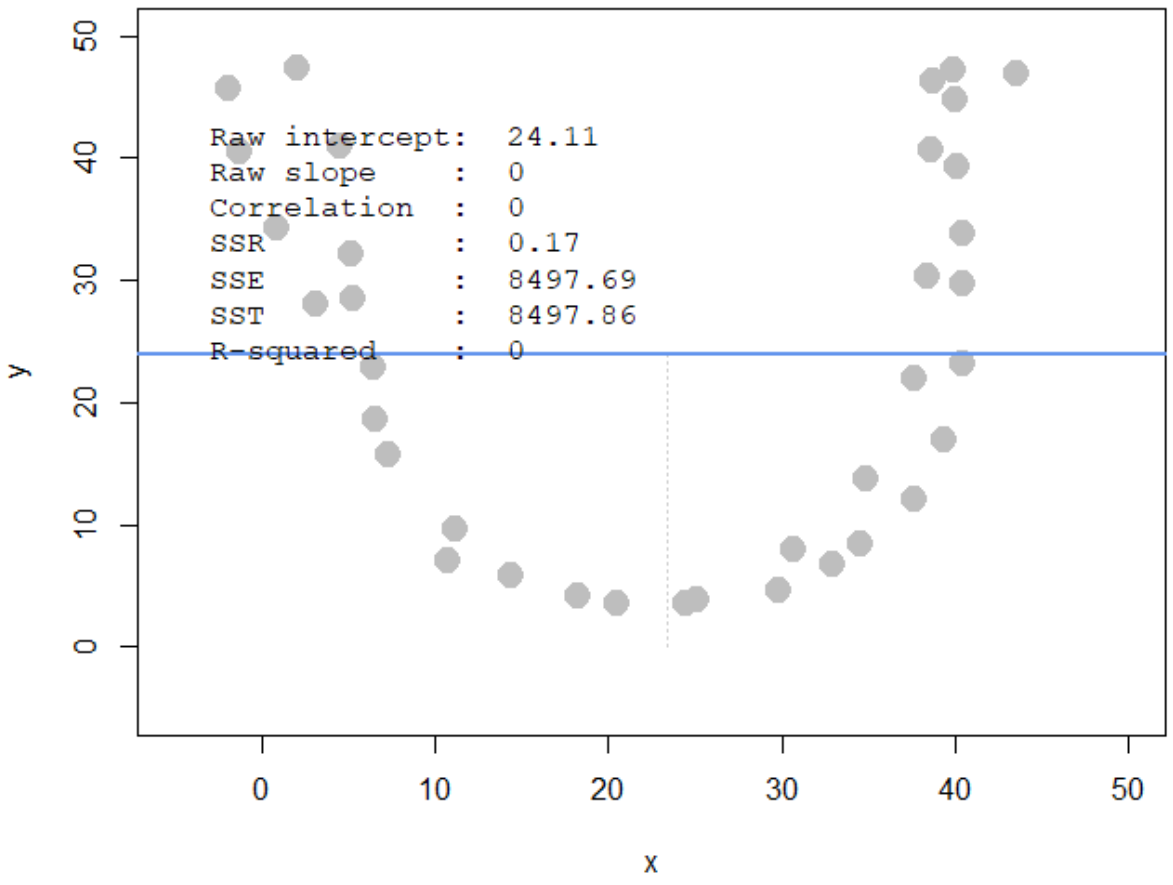
(ii) correlation of x and y that is close to 1

Question 2(d)

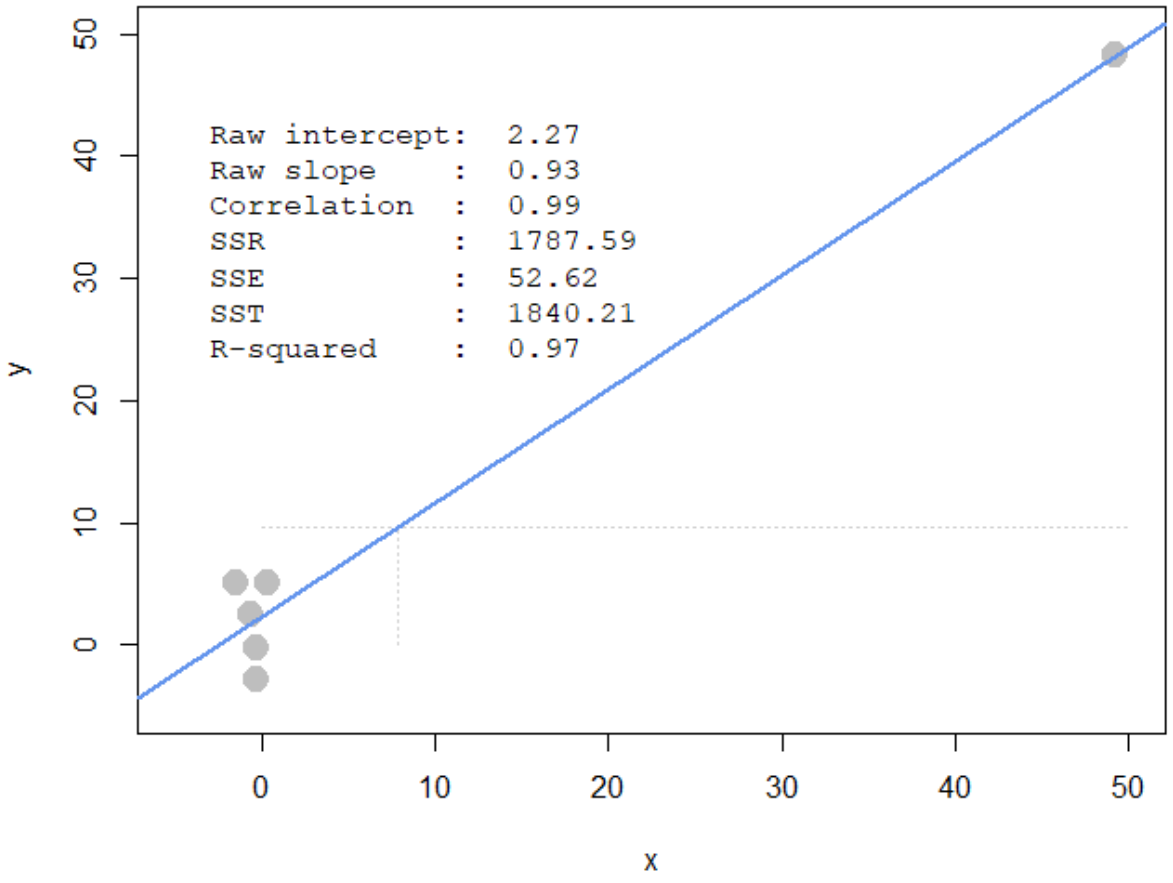
(i) raw slope of x and y that is close to -1

(ii) correlation of x and y that is close to -1

Question 2(e)


Raw intercept: 24.11
Raw slope : 0
Correlation : 0
SSR : 0.17
SSE : 8497.69
SST : 8497.86
R-squared : 0
y
x
0
10
20
30
40
50

Question 2(f)



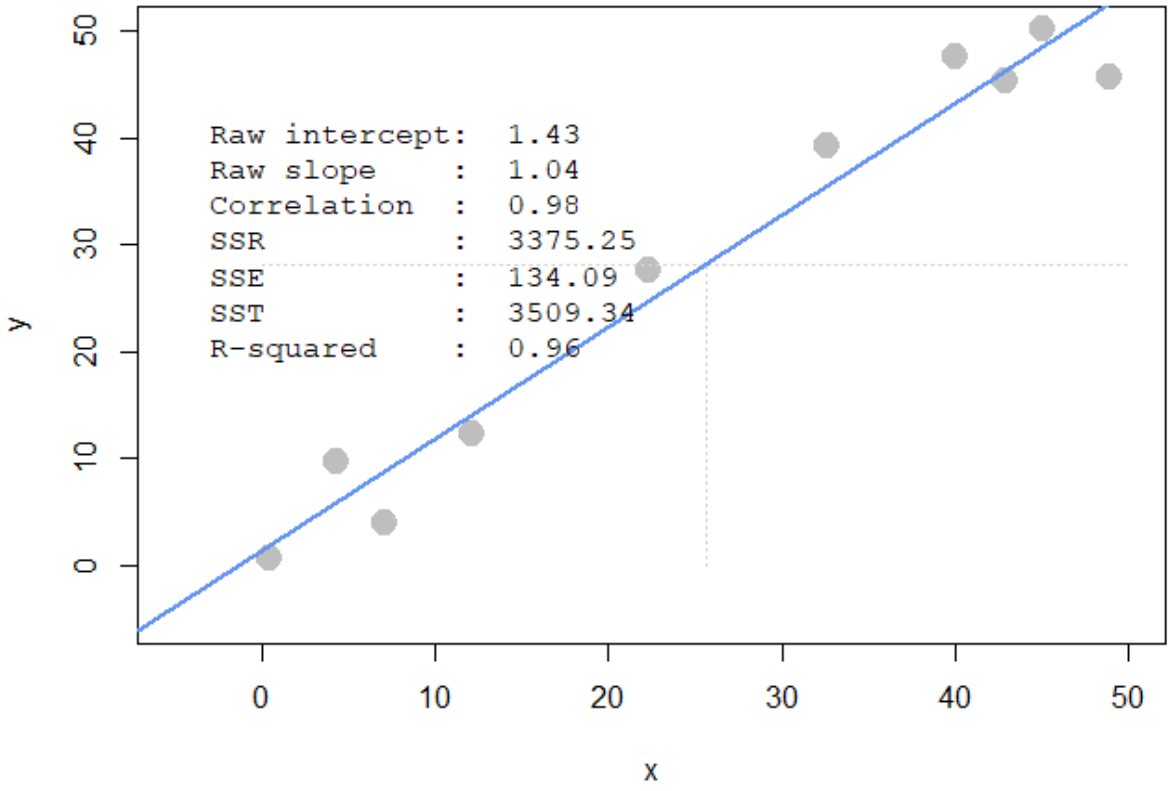
Question 2(g)

(i)

```
#pts <- interactive_regression()
# copy interactive_regression() results to pts since not able to render file that requires input
pts <- data.frame(x=c(0.4570313, 12.1660156, 22.3710938, 42.9960938,
                      49.0117188, 45.1445313,  4.3242188,  7.1171875,
                      32.6835938, 40.0957031),
                  y=c(0.6214511, 12.2460568, 27.5141956, 45.2113565,
                      45.5583596, 50.0694006,  9.6435331,  3.9179811,
                      39.1388013, 47.4668770))
pts
```

```
##           x         y
## 1 0.4570313 0.6214511
```

```
## 2  12.1660156 12.2460568
## 3  22.3710938 27.5141956
## 4  42.9960938 45.2113565
## 5  49.0117188 45.5583596
## 6  45.1445313 50.0694006
## 7   4.3242188  9.6435331
## 8   7.1171875  3.9179811
## 9  32.6835938 39.1388013
## 10 40.0957031 47.4668770
```



(ii)

```
pts_regr <- lm(pts$y ~ pts$x)
summary(pts_regr)
```

```
##
## Call:
## lm(formula = pts$y ~ pts$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9305 -1.7169  0.2991  3.4387  4.2660
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.43270    2.28428   0.627    0.548
## pts$x        1.04171    0.07341  14.191 5.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.094 on 8 degrees of freedom
## Multiple R-squared:  0.9618, Adjusted R-squared:  0.957
## F-statistic: 201.4 on 1 and 8 DF,  p-value: 5.92e-07
```

Regression intercept = 1.43

Slope = 1.04

(iii)

```
cor(pts)
```

```
##           x         y
## x 1.0000000 0.9807091
## y 0.9807091 1.0000000
```

Values seem to be the same.

(iv)

```
pts_regr <- lm(scale(pts$y) ~ scale(pts$x))
summary(pts_regr)
```

```
##
## Call:
## lm(formula = scale(pts$y) ~ scale(pts$x))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35097 -0.08695  0.01515  0.17414  0.21604
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.157e-17  6.556e-02    0.00        1
## scale(pts$x) 9.807e-01  6.911e-02   14.19 5.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2073 on 8 degrees of freedom
## Multiple R-squared:  0.9618, Adjusted R-squared:  0.957
## F-statistic: 201.4 on 1 and 8 DF,  p-value: 5.92e-07
```

Regression intercept ~ 0

Slope ~ 0.98

(v) The standardized regression coefficient (slope) is equal to the correlation.