# BACS HW9 - 109006234

Credit: 109006278

April 16th 2023

### Loading The Libraries

```
library(data.table)
library(SnowballC)
library(lsa)
```

```
## Warning: package 'lsa' was built under R version 4.2.3
```

```
library(compstatslib)
```

### Loading The Datas

```
ac_bundles_dt <- fread(
        "G:/My Drive/111_2_BACS/HW9/piccollage_accounts_bundles.csv")
ac_bundles_matrix <- as.matrix(ac_bundles_dt[, -1, with=FALSE])
```

## Problem 1

**(a) Let's explore to see if any sticker bundles seem intuitively similar:**
**(i) Download PicCollage onto your mobile from the App Store and take a look at the style and content of various bundles in their Sticker Store (iOS app: can see how many recommendations does each bundle have? Android app might not have recommendations)**
For iOS devices, each bundle will have six recommendations.

**(ii) Find a single sticker bundle that is both in our limited data set and also in the app's Sticker Store. Then, use your intuition to recommend (guess!) five other bundles in our dataset that might have similar usage patterns as this bundle**
For this case, I chose `family` bundle and I think it would be related to `hellobaby`, `toMomwithLove`, `hbd2016`, `happy`, `Mom2013`.

**(b) Let's find similar bundles using geometric models of similarity:**
**(i) Let's create cosine similarity based recommendations for all bundles:**
**1. Create a matrix or data.frame ofthe top 5 recommendations for all bundles**

```
cos_sim <- cosine(ac_bundles_matrix)
cos_sim_add <- apply(cos_sim, 1, mean)
cos_sim_add_rank <- cos_sim_add[order(cos_sim_add, decreasing = TRUE)]
cos_sim_add_rank[1:5]
```

```
##     springrose eastersurprise         bemine     watercolor hipsterholiday
##      0.1578966      0.1459645      0.1383451      0.1375165      0.1368757
```

**2. Create a new function that automates the above functionality:it should take an accounts-bundles matrix as aparameter, and return a data object with the top 5 recommendations for each bundle in our dataset,using cosine similarity.**

```
get_top5 <- function (bundle_name,data) {
  reg1 <- data[bundle_name,]
  reg2 <- reg1[order(reg1, decreasing = TRUE)]
  return (reg2[2:6])
}
```

**3. What are the top 5 recommendations for the bundle you chose to explore earlier?**

```
get_top5("family", cos_sim)
```

```
##  halloweenparty     babyanimals       hellobaby   toMomwithLove hipsteroverlays
##       0.9045340       0.7454838       0.7242068       0.5048281       0.4039816
```

**(ii) Let's create correlation based recommendations.**

```
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
row.names(ac_bundles_mc_b) <- row.names(ac_bundles_dt)
cor_sim_2 <- cosine(ac_bundles_mc_b)
get_top5("family", cor_sim_2)
```

```
##  halloweenparty     babyanimals       hellobaby   toMomwithLove hipsteroverlays
##       0.9045329       0.7454603       0.7241781       0.5052539       0.4038985
```

**(iii) Let's create adjusted-cosine based recommendations.**

```
bundle_means <- apply(ac_bundles_matrix, 1, mean)
bundle_means_matrix <- replicate(ncol(ac_bundles_matrix), bundle_means)
ac_bundles_mc_b <- ac_bundles_matrix - bundle_means_matrix
cor_sim_3 <- cosine(ac_bundles_mc_b)
get_top5("family", cor_sim_3)
```

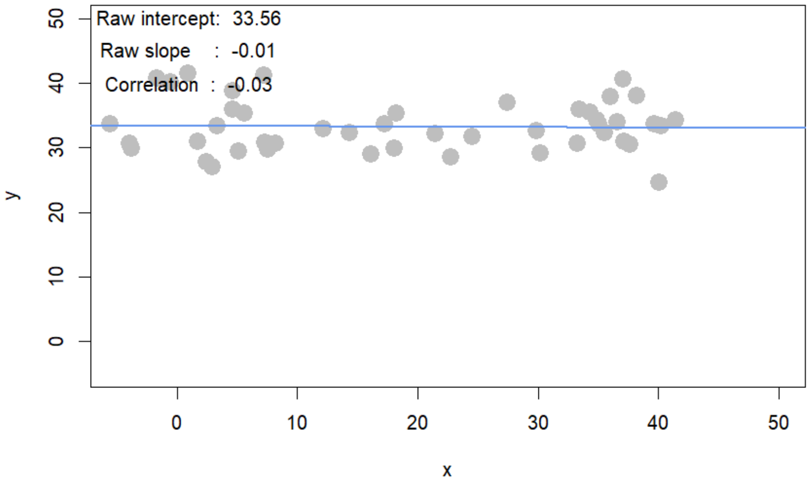
```
## floralwedding chicchristmas christmassnow       cny2017    frombierun
##     0.9923109     0.9923109     0.9907902     0.9907847     0.9907740
```

**(c) Are the three sets of geometric recommendations similar in nature (theme/keywords) to the recommendations you picked earlier using your intuition alone? What reasons might explain why your computational geometric recommendation models produce different results from your intuition?**
Only some sets that are chosen with intuition aligns with the Top 5 Recommendations. This could happen because we took the meaning of the bundles names literally. Maybe the objects within that bundle will really show for which occasion it would be suitable for.

**(d) What do you think is the conceptual difference in cosine similarity, correlation, and adjusted-cosine?**
Cosine similarity measures the cosine of the angle between the vectors, which ranges from -1 (perfectly dissimilar) to 1 (perfectly similar). Correlation, on the other hand, measures the degree to which the two variables move together in a linear fashion. Correlation ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. Adjusted-cosine similarity is a variant of cosine similarity that adjusts for differences in the mean rating given by different users in collaborative filtering.

## Problem 2

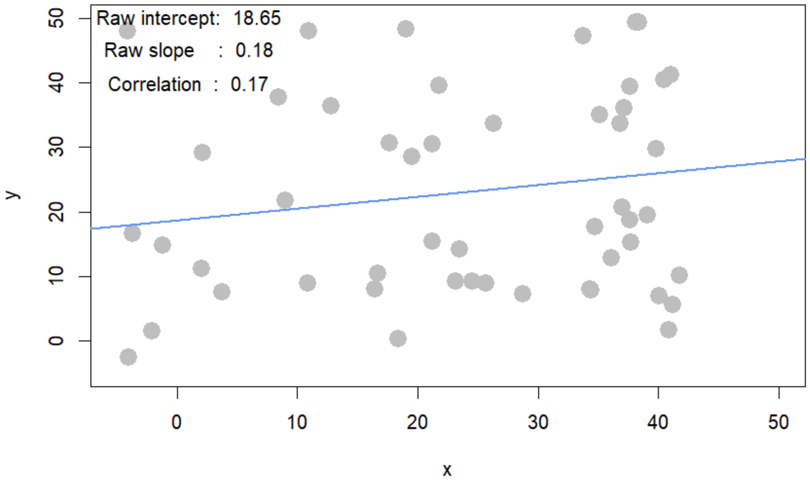
**(a) Scenario A: Create a horizontal set of random points, with a relatively narrow but flat distribution**



**(i) What raw slope of x and y would you generally expect?**
The raw slope is expected to always be close to zero.

**(ii) What is the correlation of x and y that you would generally expect?**
As we can observe, the line is horizontal, which means the correlation will also be close to zero.

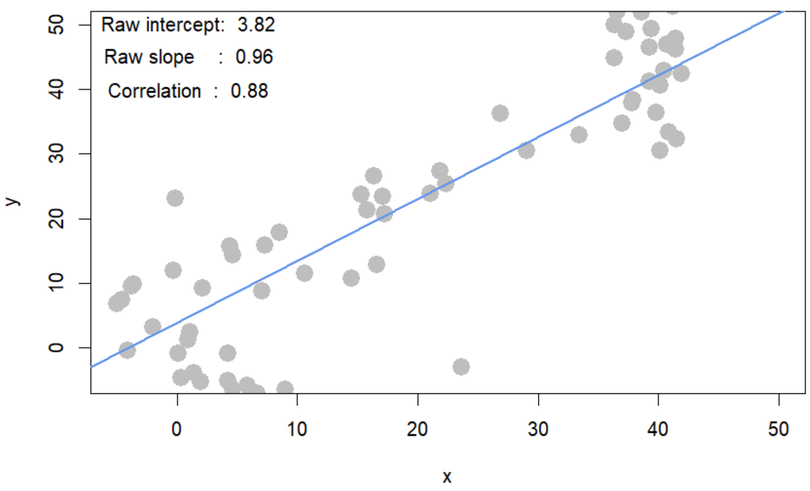
**(b) Scenario B: Create a random set of points to fill the entire plotting area, along both x-axis and y-axis**



**(i) What raw slope of x and y would you generally expect?**
The raw slope is expected to be close to zero as well.

**(ii) What is the correlation of x and y that you would generally expect?**
As both variable has no relationship whatsoever, the correlation will be close to zero.

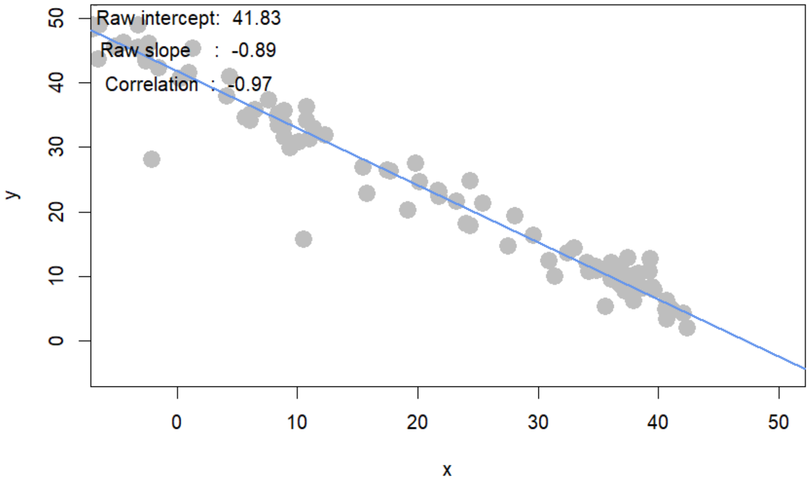
**(c) Scenario C: Create a diagonal set of random points trending upwards at 45 degrees**



**(i) What raw slope of x and y would you generally expect?**
For this scenario, the slope is expected to reach positive 1.

**(ii) What is the correlation of x and y that you would generally expect?**
From the results, we can observe that the correlation will be expected to always be close to positive 1.

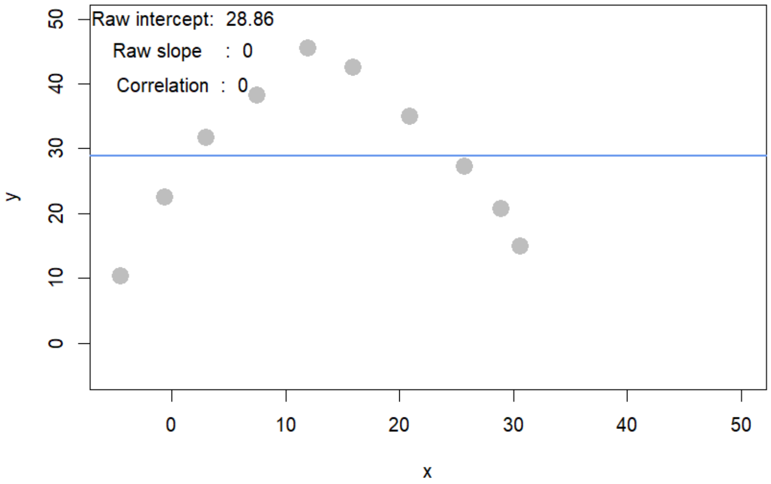
**(d) Scenario D: Create a diagonal set of random trending downwards at 45 degrees**



**(i) What raw slope of x and y would you generally expect?**
For this scenario, the slope is expected to reach negative 1.

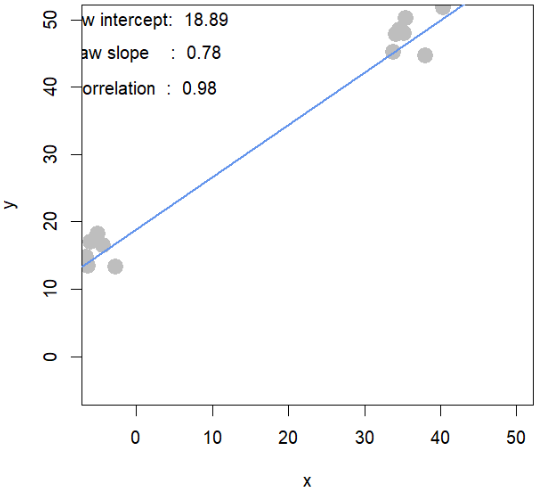
**(ii) What is the correlation of x and y that you would generally expect?**
As we can observe from the plot, the correlation between x and y is always close to negative 1 as well.

**(e) Apart from any of the above scenarios, find another pattern of data points with no correlation (r approximately 0)**



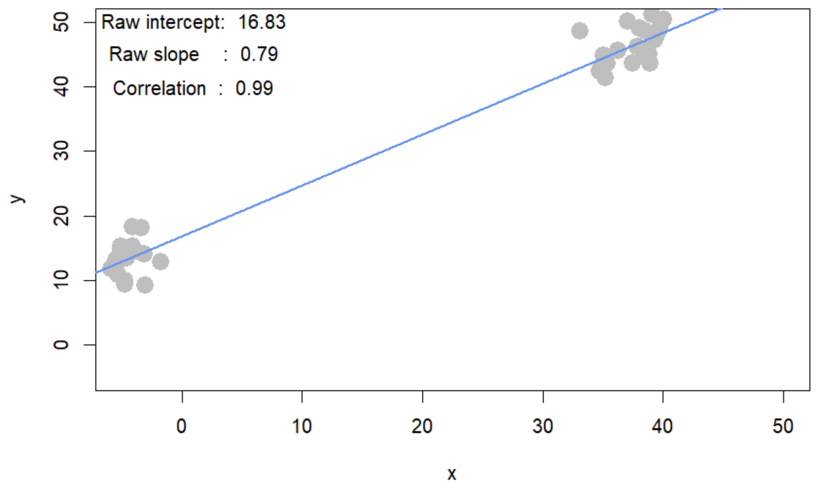
A possible scenario that will yield r = 0, but the variables are strongly correlated to each other is a scatter plot of `ages vs. time spent working`, in which `ages` is the x-axis and `time spent`is the y-axis. As both raw slope and correlation are close to zero, however the plot shows otherwise. Taking the example, we can see that in early years, time spent working are much less, compared to adulthood. After adulthood, the time spent on working drops significantly as they reach retirement age.

**(f) Apart from any of the above scenarios, find another pattern of data points with perfect correlation (r approximately 1)**



A possible scenario that will yield r = 1, but the visual suggest differently, is a scatter plot of `ages vs. the amount of diapers bought`, in which `ages` is the x-axis and `the amount of diapers bought`is the y-axis.. As both raw slope and correlation are close to one, however the plot visualization shows a big gap in between the groups of data. Taking the example, we can see that in early years, parents will buy a lot of diapers to accomodate their newborns until toddler years. After people start to enter their retirement age or simply put as they get older, they will need to buy diapers as well.

**(g) Let's see how correlation relates to simple regression, by simulating any linear relationship you wish:**



### (i) Record the points

```r
dataset <- data.frame(
  x = c(5.0815109, 15.8784436, 25.6631639, 26.9284294, 30.1337688,
        32.6642999, 38.4845214, 18.5776768, 13.1792105, 11.2391366),
  y = c(19.75412, 33.26087, 38.15892, 37.26837, 37.86207,
        40.83058, 47.06447, 31.77661, 29.25337, 28.06597) )
```

### (ii) Estimate the regression intercept and slope of pts

```r
summary(lm(dataset$y ~ dataset$x))
```

```
##
## Call:
## lm(formula = dataset$y ~ dataset$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9166 -0.9844  0.3072  1.0916  3.0531
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.1235     1.4402   13.28 9.88e-07 ***
## dataset$x     0.6981     0.0600   11.63 2.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.914 on 8 degrees of freedom
## Multiple R-squared:  0.9442, Adjusted R-squared:  0.9372
## F-statistic: 135.4 on 1 and 8 DF,  p-value: 2.711e-06
```

### (iii) Estimate the correlation of x and y

```
cor(dataset)
```

```
##           x         y
## x 1.0000000 0.9717023
## y 0.9717023 1.0000000
```

### (iv) Standardize

```
std <- data.frame(x = scale(dataset$x), y = scale(dataset$y))
summary(lm(std$y ~ std$x))
```

```
##
## Call:
## lm(formula = std$y ~ std$x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38176 -0.12884  0.04022  0.14288  0.39962
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.974e-18  7.923e-02    0.00        1
## std$x        9.717e-01  8.351e-02   11.63 2.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2505 on 8 degrees of freedom
## Multiple R-squared:  0.9442, Adjusted R-squared:  0.9372
## F-statistic: 135.4 on 1 and 8 DF,  p-value: 2.711e-06
```

```
cor(std)
```

```
##           x         y
## x 1.0000000 0.9717023
## y 0.9717023 1.0000000
```

### (v) What is the relationship between correlation and the standardized simple-regression estimates?

The relationship between correlation and standardized simple regression estimates is that the correlation coefficient between two variables is equal to the standardized regression coefficient in a simple linear regression model when both variables are standardized.