TEAM DETAILS

GROUP-17

Name	Roll Number
Priyam Bajpai	S20190010144
Shreyash Mishra	S20190010120
Anurag Kumar	S20190010009
Piyush Kumar	S20190010141
Siddharth Pandey	S20190010163

Topic 2: Pearson's Product-Moment Correlation analysis for paired data

Date of Submission: 17-11-2021

IDA PROJECT

REPORT

TOPIC :

Pearson's Product-Moment Correlation analysis for paired data.

Pearson's correlation coefficient is the test statistic that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Assumptions:

1. **Independent of case:** Cases should be independent of each other.
2. **Linear relationship:** Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
3. **Homoscedasticity:** The residual scatterplot should be roughly rectangular-shaped.

Properties:

1. **Limit:** Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
2. **Pure number:** It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3. **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

$$r = \frac{\Sigma(x^i - \bar{x})(y^i - \bar{y})}{\sqrt{\Sigma(x^i - \bar{x})^2 \Sigma(y^i - \bar{y})^2}}$$

where,

R = Pearson Correlation Coefficient

x^i = x variable sample y^i = y variable sample

$\bar{x}$ = mean of values in x variable sample $\bar{y}$ = mean of values in y variable sample

Degree of correlation:

1. **Perfect:** If the value is near ± 1, then it is said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

2. **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
3. **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4. **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
5. **No correlation:** When the value is zero.

PROBLEM STATEMENT:

a) Find the correlation between the calories of beverages and carbs, and calories of food and carbs. Use 'starbucks-menu-nutrition-drinks.csv' for beverages and 'starbucks-menu-nutrition-food.csv' for food.
b) If correlation exists, then consult the statistical table to report the significance test. Make any reasonable assumption, if you need it in your experiment.
c) Calculate the "Coefficient of determination" and comment on your result.

DATASET :

Three datasets were given (source):

- Starbucks-menu-nutrition-drinks.csv
- Starbucks-menu-nutrition-food.csv
- Starbucks_drinkMenu_expanded.csv

1) Starbucks-menu-nutrition-drinks.csv

We used this dataset to calculate the correlation between the calories of beverages and carbs. This dataset contains missing values, denoted by "-" in the table. The dataset contained either values for all attributes of a beverage, or no value for any attribute of a beverage. Out of 177 beverages in the dataset, 85 beverages did not have any information about

them at all. So, we decided to drop all such beverages from our dataset. The remaining 92 beverages were used for the assigned task.

```
> drinks <- read.csv("starbucks-menu-nutrition-drinks.csv")
> head(drinks,6)
                                          X Calories Fat..g. Carb...g. Fiber..g. Protein Sodium
1          Cool Lime Starbucks Refreshers™ Beverage       45       0        11         0       0     10
2                             Ombré Pink Drink        -       -         -         -       -      -
3                                   Pink Drink        -       -         -         -       -      -
4     Strawberry Acai Starbucks Refreshers™ Beverage       80       0        18         1       0     10
5 Very Berry Hibiscus Starbucks Refreshers™ Beverage       60       0        14         1       0     10
6                                 Violet Drink        -       -         -         -       -      -
```

2) Starbucks-menu-nutrition-food.csv

We used this dataset to calculate the correlation between the calories of food and carbs. This dataset did not contain any missing values, and all 113 food items were used for further analysis.

```
> food <- read.csv("starbucks-menu-nutrition-food.csv",fileEncoding = "UTF-16")
> head(food,6)  #get some idea how dataset is like
                                    X Calories Fat..g. Carb...g. Fiber..g. Protein..g.
1                         Chonga Bagel      300       5        50         3          12
2                         8-Grain Roll      380       6        70         7          10
3                     Almond Croissant      410      22        45         3          10
4                        Apple Fritter      460      23        56         2           7
5                     Banana Nut Bread      420      22        52         2           6
6 Blueberry Muffin with Yogurt and Honey      380      16        53         1           6
```

3) Starbucks_drinkMenu_expanded.csv

We did not use this dataset for the assigned task, as per the problem statement.

IMPLEMENTATION :

a) First, for calculating the degree of correlation or pearson correlation coefficient, we find the variance and mean of paired data. We have created separate functions for variance and pearson's correlation coefficient. We did not use any library functions for doing so, and have verified that the output matches the library function result.

```r
# Pearson Correlation (Degree of Correlation)
pearson_corr <- function(X,Y) {
  if (length(X) != length(Y)){ #if columns are not equal, return error
    return("Error: unequal columns")
  }
  n <- length(X)  #length of column
  mX <- total(X)/n  #mean of column X
  mY <- total(Y)/n  #mean of column Y
  vX <- variance(X) #variance of column X
  vY <- variance(Y) #variance of column Y
  ssX <- (vX*(n-1))^0.5  #squared deviation of column X
  ssY <- (vY*(n-1))^0.5  #squared deviation of column Y

  tot <- 0.0
  for (i in 1:length(X)){
    tot <- tot + (X[i]-mX)*(Y[i]-mY)  #paired sum of deviation about mean
  }

  r <- tot/(ssX*ssY)  #get pearson coefficient
  return(r)
}
```

b) To test whether the association is merely apparent, and might have arisen by chance, we use the t test to check the significance of our result. We have created separate functions for degree of freedom and t value. We did not use any library functions for doing so, and used a statistical table to determine critical value.

```r
# Degree of Freedom ((m-1)*(n-1))
degree_of_freedom <- function(X,Y){
  if (length(X) != length(Y)){  #if columns are not equal, return error
    return("unequal columns")
  }
  return ((length(X)-2))  #return degree of freedom
}

# T value from degree of correlation
t_value <-function(X,Y){
  if (length(X) != length(Y)){  #if columns are not equal, return error
    return("unequal columns")
  }
  n <- length(X)  #length of column
  r <-pearson_corr(X,Y)  #get pearson correlation coefficient
  value_of_t <- r *((n-2)/(1-(r^2)))^0.5  #get t value for correlation
  return (value_of_t)

}
```

c) To calculate the coefficient of determination, we have created a separate function.

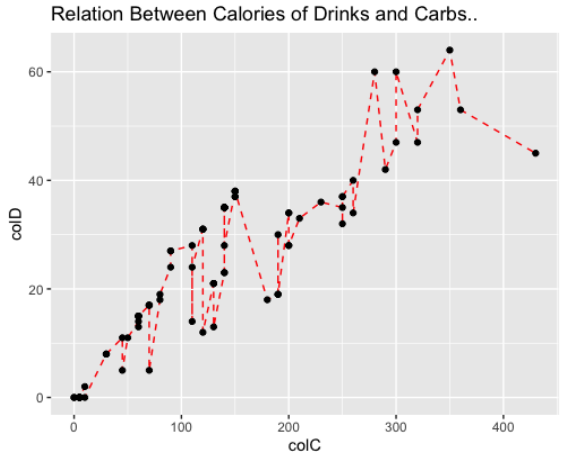
```
# Coefficient of determination
coefficient_of_detemination <-function(X,Y){
  if (length(X) != length(Y)){  #if columns are not equal, return error
    return("unequal columns")
  }
  xy <- 0.0
  x2 <-0.0
  y2 <-0.0
  n <- length(X)  #length of column
  mX <- total(X)/n  #mean of column X
  mY <- total(Y)/n  #mean of column Y
  for (i in 1:n){
        xy <- xy + ((X[i]-mX)*(Y[i]-mY))  #paired sum of deviation about mean
  }
  for (i in 1:n){
    x2 <- x2 + ((X[i]-mX)*(X[i]-mX))  #squared sum of deviation about mean of column X
  }
  for (i in 1:n){
    y2 <- y2 + ((Y[i]-mY)*(Y[i]-mY))  #squared sum of deviation about mean of column Y
  }
  r_2=xy/((x2*y2)^0.5)  #get coefficient of determination
  return (r_2^2)
}
```

EXPERIMENT :

A. Correlation between the calories of beverages and carbs:

After we dropped beverages that did not contribute to the dataset (due to missing values), we calculated the pearson correlation coefficient, using the created function, between the calories and carbs of a beverage. We found the value, r, to be **0.871**. This implies that the calories and carbs of a beverage are highly and **positively correlated**, i.e., as calories increase, the carbs increase as well.

To perform Pearson correlation analysis, we assume that the two variables (carbs and calories) should be linear. This can be verified by plotting a scatter plot of these variables.



We can see here that these variables are, more or less, linear. So Pearson correlation is valid in this case.

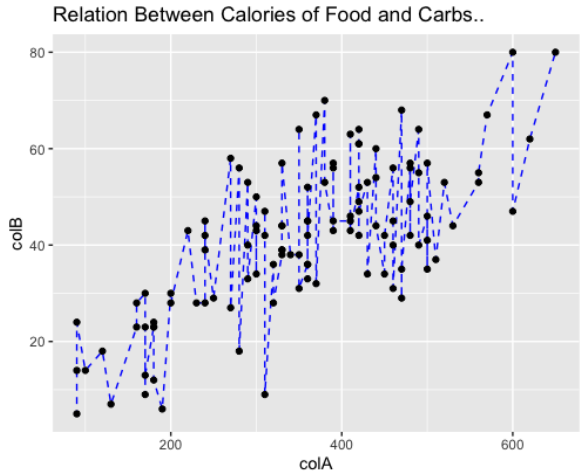
To test the significance of our result, we use the t-test. Using the created functions, we determine the degrees of freedom to be 90 and the test statistic (t value) to be 16.84 . Consulting the t-test table for single tailed test, at degrees of freedom 90 and for $\alpha = 0.05$, we find that the critical value of t is 1.6620. Now, since the critical value of t is significantly less than the test statistic, we conclude that the value of Pearson's correlation coefficient in this case may be regarded as **less significant**.

We also find out the coefficient of determination using regression analysis that depicts the quality of fit. Using the created functions, we got the value, R^2 , to be **0.7591** . This denotes a **good fit**, and that 75.91% of variance of carb can be predicted by the linear regression and calories (and vice versa). We also observe that, since there is only one predictor included in the model, the value of R^2 (0.759134) is the square of the value of the Pearson correlation coefficient (0.871283).

B. Correlation between the calories of food and carbs:

Using the created function, we calculated the pearson correlation coefficient between the calories and carbs of a beverage. We found the value, r, to be **0.708** . This implies that the calories and carbs of a beverage are highly and **positively correlated**, i.e., as calories increase, the carbs increase as well.

To perform Pearson correlation analysis, we assume that the two variables (carbs and calories) should be linear. This can be verified by plotting a scatter plot of these variables.



We can see here that these variables are, more or less, linear. So Pearson correlation is valid in this case.

To test the significance of our result, we use the t-test. Using the created functions, we determine the degrees of freedom to be 111 and the test statistic (t value) to be 10.577 . Consulting the t-test table for single tailed test, at degrees of freedom 111 and for α = 0.05, we find that the critical value of t is 1.6857. Now, since the critical value of t is significantly less than the test statistic, we conclude that the value of Pearson's correlation coefficient in this case may be regarded as **less significant**.

We also find out the coefficient of determination using regression analysis that depicts the quality of fit. Using the created functions, we got the value, R^2 , to be **0.5019** . This denotes an **average fit**, and that 50.19% of variance of carb can be predicted by the linear regression and calories (and vice versa). We also observe that, since there is only one predictor included in the model, the value of R^2 (0.501966) is the square of the value of the Pearson correlation coefficient (0.7084955).

RESULTS:

A. For the beverage dataset, we got the pearson correlation coefficient as 0.871, which indicates that, for beverages, carbs and calories are highly and positively correlated. However, after performing t-test, we observe that the value of Pearson's correlation coefficient in this case may be regarded as less significant. We also determined the coefficient of determination as 0.7591, indicating a good fit.
B. For the food dataset, we got the pearson correlation coefficient as 0.708, which indicates that, for beverages, carbs and calories are highly and positively correlated. However, after performing t-test, we observe that the value of Pearson's correlation coefficient in this case may be regarded as less significant. We also determined the coefficient of determination as 0.5019, indicating an average fit.