ACADGILD

SESSION 12: Generalized Linear Models

Assignment 2

Submitted by: Udaykumar
Udaykumarr019@gmail.com
(M):+91-8123431864

Data Analytics

Table of Contents

1. Problem Statement

1. Use the given link below:
https://archive.ics.uci.edu/ml/machine-learning-databases/communities/

Perform the below operations:

a) Visualize the correlation between all variable in a meaningful way, clear representation of correlations. Find out top 3 reasons for having more crime in a city.

b) What is the difference between covariance and correlation, take an example from this dataset and show the differences if any?

2. Solution

a. Visualize the correlation between all variable in a meaningful way, clear representation of correlations. Find out top 3 reasons for having more crime in a city.

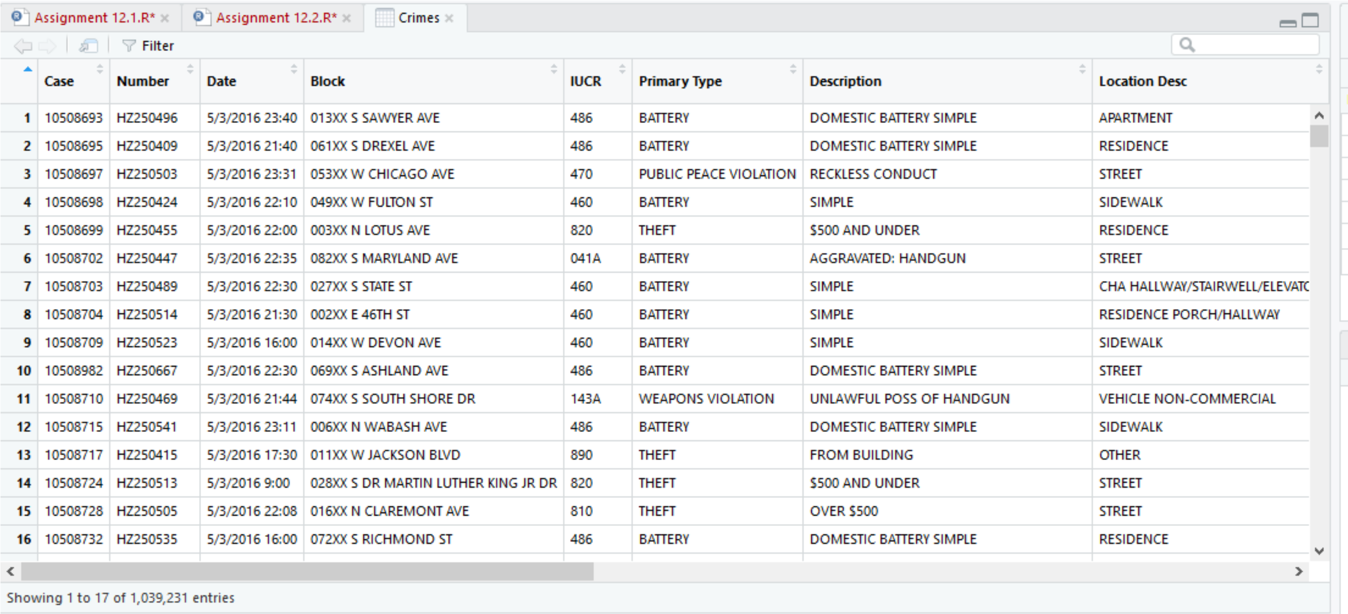
The R-script for the given problem is as follows:

```
library(readr)
Crimes <- read_csv("E:/uday/acadgild data analytics/supporting
files/communities.csv ")
View(Crimes)

names(Crimes) <- c("Case", "Number", "Date", "Block", "IUCR", "Primary Type",
"Description", "Location Desc", "Arrest", "Domestic", "Beat", "District", "Ward",
"Community Area","FBI Code", "X Coordinate", "Y Coordinate", "Year", "Updated On",
"Latitude", "Longitude", "Location")
head(Crimes)
str(Crimes)

#a. Visualize the correlation between all variables in a meaningful and clear way
# of representing.

library(dplyr)
Crimes <- na.omit(Crimes)
names(Crimes)
```

```
c <- cor(Crimes[c(11,12,13,14,18,20,21)])
c

psych::cor.plot(c)

# a.Find out top 3 reasons for having more crime in a city.
x <- as.data.frame(table(Crimes$Description))
x[order(x$Freq, decreasing = T)[1:3],]
```

The output of the R-Script (from Console window) is given as follows:

```
> library(readr)
> Crimes <- read_csv(″E:/uday/acadgild data analytics/supporting files/communities.csv ″)
Parsed with column specification: cols(
  .default = col_character(), ID =
  col_double(),
  Arrest = col_logical(),
  Domestic = col_logical(), Beat
  = col_double(), District =
  col_double(), Ward =
  col_double(),
  `Community Area` = col_double(),
  `X Coordinate` = col_double(),
  `Y Coordinate` = col_double(), Year
  = col_double(),
  Latitude = col_double(),
  Longitude = col_double()
)
See spec(...) for full column specifications.
|===================================================================================
================| 100%      216 MB
> View(Crimes)
```

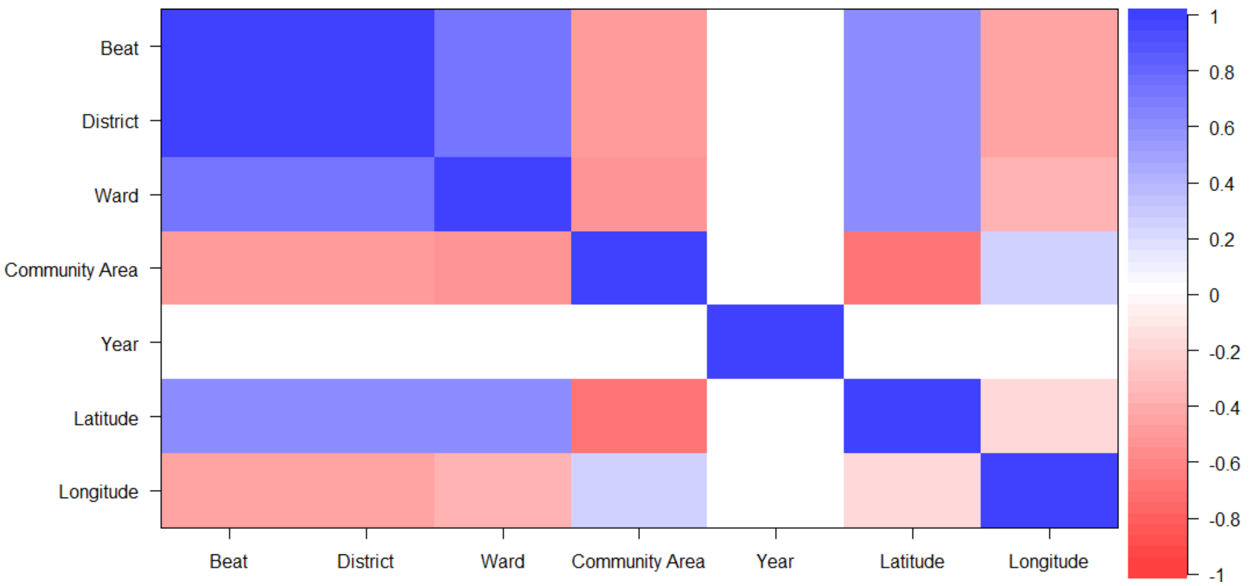


	Case	Number	Date	Block	IUCR	Primary Type	Description	Location Desc
1	10508693	HZ250496	5/3/2016 23:40	013XX S SAWYER AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT
2	10508695	HZ250409	5/3/2016 21:40	061XX S DREXEL AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE
3	10508697	HZ250503	5/3/2016 23:31	053XX W CHICAGO AVE	470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET
4	10508698	HZ250424	5/3/2016 22:10	049XX W FULTON ST	460	BATTERY	SIMPLE	SIDEWALK
5	10508699	HZ250455	5/3/2016 22:00	003XX N LOTUS AVE	820	THEFT	$500 AND UNDER	RESIDENCE
6	10508702	HZ250447	5/3/2016 22:35	082XX S MARYLAND AVE	041A	BATTERY	AGGRAVATED: HANDGUN	STREET
7	10508703	HZ250489	5/3/2016 22:30	027XX S STATE ST	460	BATTERY	SIMPLE	CHA HALLWAY/STAIRWELL/ELEVATC
8	10508704	HZ250514	5/3/2016 21:30	002XX E 46TH ST	460	BATTERY	SIMPLE	RESIDENCE PORCH/HALLWAY
9	10508709	HZ250523	5/3/2016 16:00	014XX W DEVON AVE	460	BATTERY	SIMPLE	SIDEWALK
10	10508982	HZ250667	5/3/2016 22:30	069XX S ASHLAND AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	STREET
11	10508710	HZ250469	5/3/2016 21:44	074XX S SOUTH SHORE DR	143A	WEAPONS VIOLATION	UNLAWFUL POSS OF HANDGUN	VEHICLE NON-COMMERCIAL
12	10508715	HZ250541	5/3/2016 23:11	006XX N WABASH AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	SIDEWALK
13	10508717	HZ250415	5/3/2016 17:30	011XX W JACKSON BLVD	890	THEFT	FROM BUILDING	OTHER
14	10508724	HZ250513	5/3/2016 9:00	028XX S DR MARTIN LUTHER KING JR DR	820	THEFT	$500 AND UNDER	STREET
15	10508728	HZ250505	5/3/2016 22:08	016XX N CLAREMONT AVE	810	THEFT	OVER $500	STREET
16	10508732	HZ250535	5/3/2016 16:00	072XX S RICHMOND ST	486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE

Showing 1 to 17 of 1,039,231 entries

```
> names(Crimes) <- c(″Case″, ″Number″, ″Date″, ″Block″, ″IUCR″, ″Primary Type″, ″Description″,
+                    ″Location Desc″, ″Arrest″, ″Domestic″, ″Beat″, ″District″, ″Ward″,
″Community Area″,
+                    ″FBI Code″, ″X Coordinate″, ″Y Coordinate″, ″Year″, ″Updated On″,
+                    ″Latitude″, ″Longitude″, ″Location″)
```

```
> head(Crimes)
# A tibble: 6 x 22
     Case Number Date      Block IUCR    `Primary Type` Description `Location Desc` Arrest Domestic
            Beat District  Ward
    <dbl> <chr>   <chr>  <chr> <chr> <chr>            <chr>         <chr>              <lgl>
<lgl>      <dbl>      <dbl> <dbl>
1  1.05e7 HZ250~ 5/3/~ 013X~ 486 BATTERY  DOMESTIC B~ APARTMENT                              TRUE
TRUE      1022         10     24
2  1.05e7 HZ250~ 5/3/~ 061X~ 486 BATTERY  DOMESTIC B~ RESIDENCE                              FALSE
TRUE        313          3     20
3  1.05e7 HZ250~ 5/3/~ 053X~ 470 PUBLIC PEACE ~ RECKLESS C~ STREET                           FALSE
FALSE     1524         15     37
4  1.05e7 HZ250~ 5/3/~ 049X~ 460 BATTERY  SIMPLE          SIDEWALK      FALSE FALSE        1532
  15       28
5  1.05e7 HZ250~ 5/3/~ 003X~ 820 THEFT    $500 AND U~ RESIDENCE                              FALSE
TRUE      1523         15     28
6  1.05e7 HZ250~ 5/3/~ 082X~ 041A  BATTERYAGGRAVATED~ STREETFALSE FALSE                      631
  6        8
# ... with 9 more variables: `Community Area` <dbl>, `FBI Code` <chr>, `X  Coordinate`
<dbl>,  `Y Coordinate` <dbl>,
#    Year <dbl>,  `Updated On` <chr>,  Latitude <dbl>,  Longitude <dbl>,  Location <chr>
> str(Crimes)
Classes  'spec_tbl_df' ,  'tbl_df' ,  'tbl'  and 'data.frame':          1048575 obs. of  22
variables:
 $ Case            : num   10508693 10508695 10508697 10508698 10508699 ...
 $ Number          : chr   "HZ250496" "HZ250409" "HZ250503" "HZ250424" ...
 $ Date            : chr   "5/3/2016 23:40" "5/3/2016 21:40" "5/3/2016 23:31" "5/3/2016 22:10" ...
 $ Block           : chr   "013XX S SAWYER AVE" "061XX S DREXEL AVE" "053XX W CHICAGO AVE"
"049XX W FULTON ST" ...
 $ IUCR            : chr   "486" "486" "470" "460" ...
 $ Primary Type    : chr   "BATTERY" "BATTERY" "PUBLIC PEACE VIOLATION" "BATTERY" ...
 $ Description     : chr   "DOMESTIC BATTERY SIMPLE" "DOMESTIC BATTERY SIMPLE" "RECKLESS CONDUCT"
"SIMPLE" ...
 $ Location Desc : chr     "APARTMENT" "RESIDENCE" "STREET" "SIDEWALK" ...
 $ Arrest          : logi   TRUE FALSE FALSE FALSE FALSE FALSE ...
 $ Domestic        : logi   TRUE TRUE FALSE FALSE TRUE FALSE ...
 $ Beat            : num   1022 313 1524 1532 1523 ...
 $ District        : num   10 3 15 15 15 6 1 2 24 7 ...
 $ Ward            : num   24 20 37 28 28 8 3 3 40 17 ...
 $ Community Area: num     29 42 25 25 25 44 35 38 1 67 ...
 $ FBI Code        : chr   "08B" "08B" "24" "08B" ...
 $ X Coordinate    : num   1154907 1183066 1140789 1143223 1139890 ...
 $ Y Coordinate    : num   1893681 1864330 1904819 1901475 1901675 ...
 $ Year            : num   2016 2016 2016 2016 2016 ...
 $ Updated On      : chr   "5/10/2016 15:56" "5/10/2016 15:56" "5/10/2016 15:56"
"5/10/2016 15:56" ...
 $ Latitude        : num   41.9 41.8 41.9 41.9 41.9 ...
 $ Longitude       : num   -87.7 -87.6 -87.8 -87.7 -87.8 ...
 $ Location        : chr   "(41.864073157, -87.706818608)" "(41.782921527, -
87.60436317)" "(41.894908283, -87.758371958)" "(41.885686845, -87.749515983)" ...
 - attr( ,*"spec")=
  .. cols(
  ..    ID = col_double(),
  ..    `Case Number` = col_character(),
  ..    Date = col_character(),
  ..    Block = col_character(),
  ..    IUCR = col_character(),
  ..    `Primary Type` = col_character(),
  ..    Description = col_character(),
  ..    `Location Description` = col_character(),
  ..    Arrest = col_logical(),
  ..    Domestic = col_logical(),
  ..    Beat = col_double(),
  ..    District = col_double(),
  ..    Ward = col_double(),
  ..    `Community Area` = col_double(),
  ..    `FBI Code` = col_character(),
  ..    `X Coordinate` = col_double(),
  ..    `Y Coordinate` = col_double(),
  ..    Year = col_double(),
  ..     `Updated On` = col_character(),
  ..     Latitude = col_double(),
  ..     Longitude = col_double(),
  ..     Location = col_character()
  .. )
> library(dplyr)
> Crimes <- na.omit(Crimes)
> names(Crimes)
 [1]  "Case"                "Number"             "Date"               "Block"              "IUCR"
"Primary Type"
 [7] "Description"          "Location Desc"      "Arrest"             "Domestic"           "Beat"
"District"
[13] "Ward" Coordinate"     "Community Area"     "FBI Code"           "X Coordinate"       "Y
                 "Year"
[19] "Updated On"           "Latitude"           "Longitude"          "Location"
> c <- cor(Crimes[c(11,12,13,14,18,20,21)])
> c
                          Beat        District            Ward  Community Area            Year
Latitude       Longitude
Beat                1.00000000     0.996402087     0.687144016     -0.49621344  -0.012652765
0.575284245    -0.479976546
District            0.99640209     1.000000000     0.691655842     -0.49621461  -0.008529942
0.576344843    -0.483244475
Ward                0.68714402     0.691655842     1.000000000     -0.54302431  -0.004215319
0.592008238    -0.397964013
Community Area -0.49621344      -0.496214608    -0.543024307      1.00000000   0.001632430      -
0.691892413     0.221028077
Year               -0.01265277    -0.008529942    -0.004215319      0.00163243   1.000000000      -
0.002721412 -0.004346718
Latitude            0.57528424     0.576344843     0.592008238     -0.69189241  -0.002721412
1.000000000 -0.209999084
Longitude          -0.47997655    -0.483244475    -0.397964013      0.22102808  -0.004346718      -
0.209999084     1.000000000
>
> psych::cor.plot(c)
```



```
> # Find out top 3 reasons for having more crime in a city.
> x <- as.data.frame(table(Crimes$Description))
> x[order(x$Freq, decreasing = T)[1:3],]
                           Var1     Freq
279                      SIMPLE 107887
```

```
1               $500 AND UNDER       97476
118 DOMESTIC BATTERY SIMPLE          93001
```

Conclusion/Interpretation:

Simple, **$500 and under** and Domestic **Battery Simple** are the top 3 reasons for having more crime

b. What is the difference between covariance and correlation, take an example from this dataset and show the differences if any?

The table showing the difference is shown below:

SR.NO.	BASIS FOR COMPARISON	COVARIANCE	CORRELATION
1	Meaning	Covariance is a measure indicating the extent to which two random variables change in tandem.	Correlation is a statistical measure that indicates how strongly two variables are related.
2	What is it?	Measure of correlation	Scaled version of covariance
3	Values	Lie between -∞ and +∞	Lie between -1 and +1
4	Change in scale	Affects covariance	Does not affects correlation
5	Unit free measure	No	Yes

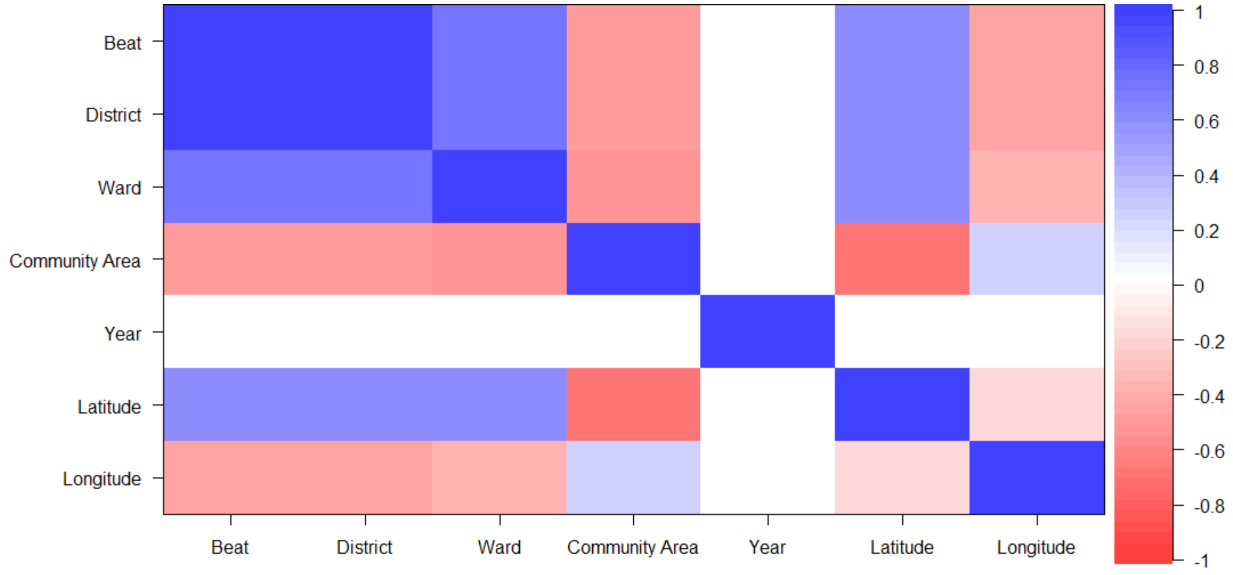
The R-script for the given problem is as follows:

```
correlation <- cor(Crimes[c(11,12,13,14,18,20,21)])
correlation
psych::cor.plot(correlation)

covariance <- cov(Crimes[c(11,12,13,14,18,20,21)])
covariance
psych::cor.plot(covariance)

#or
correlation1 <- cor(Crimes[c(11,12)])
correlation1
covariance1 <- cov(Crimes[c(11,12)])
covariance1

#or
correlation1 <- cor(Crimes[c(14,18)])
correlation1
```

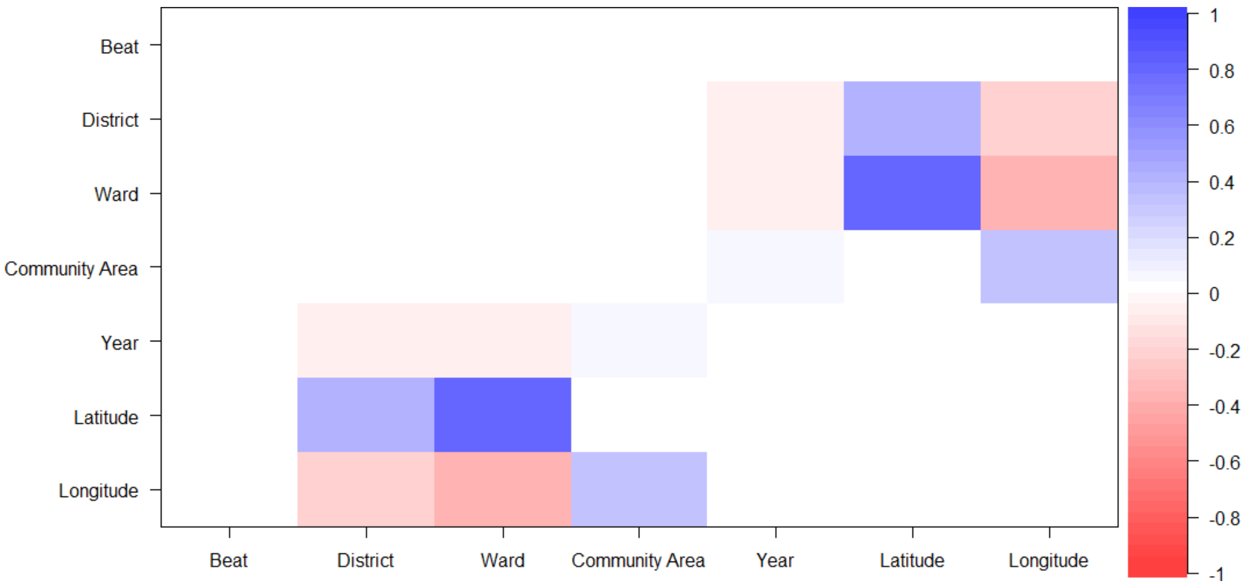
covariance1 <- cov(Crimes[c(14,18)])
covariance1

The output of the R-Script (from Console window) is given as follows:

```
> correlation <- cor(Crimes[c(11,12,13,14,18,20,21)])
> correlation
                      Beat      District          Ward  Community Area
Year
Beat            1.00000000   0.996402087   0.687144016     -0.49621344 -
0.012652765
District        0.99640209   1.000000000   0.691655842     -0.49621461 -
0.008529942
Ward            0.68714402   0.691655842   1.000000000     -0.54302431 -
0.004215319
Community Area -0.49621344  -0.496214608  -0.543024307      1.00000000
0.001632430
Year           -0.01265277  -0.008529942  -0.004215319      0.00163243
1.000000000
Latitude        0.57528424   0.576344843   0.592008238     -0.69189241 -
0.002721412
Longitude      -0.47997655  -0.483244475  -0.397964013      0.22102808 -
0.004346718
                   Latitude     Longitude
Beat            0.575284245  -0.479976546
District        0.576344843  -0.483244475
Ward            0.592008238  -0.397964013
Community Area -0.691892413   0.221028077
Year           -0.002721412  -0.004346718
Latitude        1.000000000  -0.209999084
Longitude      -0.209999084   1.000000000
> psych::cor.plot(correlation)
```



```
> covariance <- cov(Crimes[c(11,12,13,14,18,20,21)])
> covariance
                              Beat         District              Ward    Community Area
Year
Beat              478745.868597       4760.82948868    6540.34371670     -7.363621e+03 -
9.4366087362
District            4760.829489         47.68600698      65.70309277     -7.349121e+01 -
0.0634920734
Ward                6540.343717         65.70309277     189.23460975     -1.602101e+02 -
0.0625041296
Community Area     -7363.621268        -73.49121476    -160.21012410      4.599820e+02
0.0377383498
Year                  -9.436609         -0.06349207      -0.06250413      3.773835e-02
1.1618657281
Latitude              38.573554          0.38568482       0.78919204     -1.438016e+00 -
0.0002842673
Longitude            -22.838536         -0.22948700      -0.37647818      3.259970e-01 -
0.0003222071
                          Latitude        Longitude
Beat              38.5735544021       -2.283854e+01
District           0.3856848236       -2.294870e-01
Ward               0.7891920358       -3.764782e-01
Community Area -1.4380157084           3.259970e-01
Year              -0.0002842673       -3.222071e-04
Latitude           0.0093909455       -1.399483e-03
Longitude         -0.0013994835        4.729241e-03
> psych::cor.plot(covariance)
```



```
> correlation1 <- cor(Crimes[c(11,12)])
> correlation1
                 Beat   District
Beat        1.0000000 0.9964021
District 0.9964021 1.0000000
```

```
> covariance1 <- cov(Crimes[c(11,12)])
> covariance1
                  Beat    District
Beat       478745.869 4760.82949
District     4760.829   47.68601
>
> #or
>
> correlation1 <- cor(Crimes[c(14,18)])
> correlation1
                  Community Area          Year
Community Area        1.00000000 0.00163243
Year                  0.00163243 1.00000000
>
> covariance1 <- cov(Crimes[c(14,18)])
> covariance1
                  Community Area          Year
Community Area      459.98196498 0.03773835
Year                  0.03773835 1.16186573
```

Conclusion/Interpretation:

Co-Variance is a systematic relationship between a pair of random variables wherein a change in one variable reciprocated by an equivalent change in another variable. Measure of correlation, Lie between -∞ and +∞. Change in scale affects covariance

Correlation is statistical measure that indicates how strongly two variables are related. Scaled version of covariance, Lie between -1 and +1, Change in scale does not affect the correlation. Unit free measure

Correlation is a special case of covariance which can be obtained when the data is standardized.