IMT 573: Problem Set 4 - Data Analysis

Jay Kuo

Due: Friday, November 4, 2022

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset4.Rmd` file from Canvas. Open `problemset4.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset4.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```r
a + b # these object don't exist
# if you run this on its own it will give an error
```

6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps4_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

Setup In this problem set you will need, at minimum, the following R packages.

```r
# Load standard libraries
library(tidyverse)
library(gridExtra)
```

Problem 1: 50 States in the USA In this problem we will use the `state` dataset, available as part of the R statistical computing platform. This data is related to the 50 states of the United States of America. Load the data and use it to answer the following questions. See here for more.

(a)

Describe the data and each variable it contains. Tidy the data, preparing it for a data analysis.

Question

- Describes data and each variable.

Answer

```r
# data(state) to get the data
?state
```

- Meaning of each variable:
- state.abb: character vector of 2-letter abbreviations for the state names.
- state.area: numeric vector of state areas (in square miles).
- state.center: list with components named x and y giving the approximate geographic center of each state in negative longitude and latitude. Alaska and Hawaii are placed just off the West Coast.
- state.division: factor giving state divisions (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, and Pacific).
- state.name: character vector giving the full state names.
- state.region: factor giving the region (Northeast, South, North Central, West) that each state belongs to.
- state.x77: matrix with 50 rows and 8 columns giving the following statistics in the respective columns.
- Population: population estimate as of July 1, 1975
- Income: per capita income (1974)
- Illiteracy: illiteracy (1970, percent of population)
- Life Exp: life expectancy in years (1969–71)
- Murder: murder and non-negligent manslaughter rate per 100,000 population (1976)
- HS Grad: percent high-school graduates (1970)
- Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- Area: land area in square miles

Question

- Data is tidy and exploratory analysis is conducted

Answer

```r
state_data <- data.frame(abb = state.abb,
                         area = state.area,
                         center_x = state.center[1],
                         center_y = state.center[2],
                         division = state.division,
                         name = state.name,
                         region = state.region) %>%
  cbind(state.x77) # Column bind
```

```r
# structure
str(state_data)
```

```
## 'data.frame':    50 obs. of  15 variables:
##  $ abb       : chr  "AL" "AK" "AZ" "AR" ...
##  $ area      : num  51609 589757 113909 53104 158693 ...
##  $ x         : num  -86.8 -127.2 -111.6 -92.3 -119.8 ...
##  $ y         : num  32.6 49.2 34.2 34.7 36.5 ...
##  $ division  : Factor w/ 9 levels "New England",..: 4 9 8 5 9 8 1 3 3 3 ...
##  $ name      : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ region    : Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

```
# summary statistics
summary(state_data)
```

```
##      abb                 area               x                  y
##  Length:50          Min.   :  1214   Min.   :-127.25   Min.   :27.87
##  Class :character   1st Qu.: 37317   1st Qu.:-104.16   1st Qu.:35.55
##  Mode  :character   Median : 56222   Median : -89.90   Median :39.62
##                     Mean   : 72368   Mean   : -92.46   Mean   :39.41
##                     3rd Qu.: 83234   3rd Qu.: -78.98   3rd Qu.:43.14
##                     Max.   :589757   Max.   : -68.98   Max.   :49.25
##
##                 division      name                     region     Population
##  South Atlantic    : 8   Length:50          Northeast    : 9   Min.   :  365
##  Mountain          : 8   Class :character   South        :16   1st Qu.: 1080
##  West North Central: 7   Mode  :character   North Central:12   Median : 2838
##  New England       : 6                      West         :13   Mean   : 4246
##  East North Central: 5                                         3rd Qu.: 4968
##  Pacific           : 5                                         Max.   :21198
##  (Other)           :11
##      Income       Illiteracy       Life Exp        Murder
##  Min.   :3098   Min.   :0.500   Min.   :67.96   Min.   : 1.400
##  1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12   1st Qu.: 4.350
##  Median :4519   Median :0.950   Median :70.67   Median : 6.850
##  Mean   :4436   Mean   :1.170   Mean   :70.88   Mean   : 7.378
##  3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89   3rd Qu.:10.675
##  Max.   :6315   Max.   :2.800   Max.   :73.60   Max.   :15.100
##
##     HS Grad          Frost             Area
##  Min.   :37.80   Min.   :  0.00   Min.   :  1049
##  1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
##  Median :53.25   Median :114.50   Median : 54277
##  Mean   :53.11   Mean   :104.46   Mean   : 70736
##  3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81162
##  Max.   :67.30   Max.   :188.00   Max.   :566432
##
```

(b)

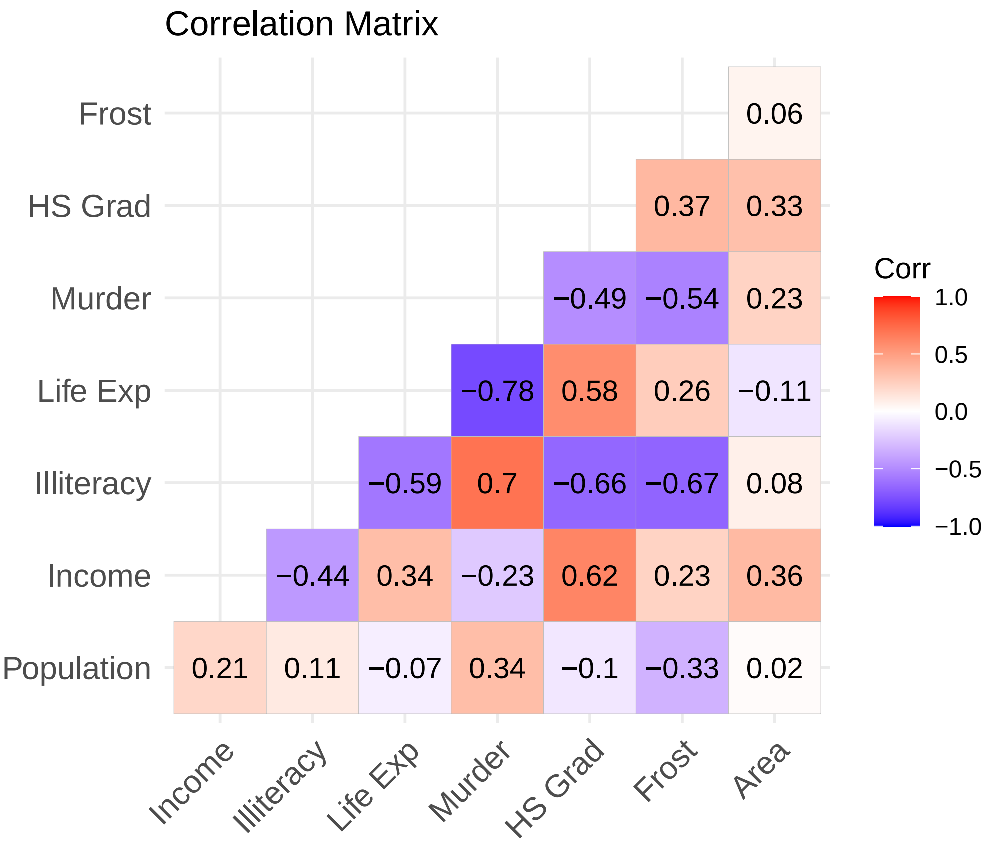
Suppose you want to explore the relationship between a state's `Murder` rate and other characteristics of the state, for example population, illiteracy rate, and more. Begin by examining the bivariate relationships associated with `Murder` rate present in the data. What does your analysis suggest might be important varibles to consider in building a model to explain variation in murder rates?

Question

- Examines bivariate relationships using charts or plots
- Visualizations have appropriate titles, labeled axes, and trend lines if applicable, printing data frames will not count

Answer

```
library(ggcorrplot)

ggcorrplot(cor(state_data[,8:15]),
           title = 'Correlation Matrix',
           type = 'lower',
           lab = TRUE)
```



Question

- Discusses important variables to consider in model building

Answer

- Life Exp: More murder, less people survived, lower life expectancy
- Illiteracy: Lower education level, higher murder rate
- Frost: When the weather is bad, fewer people will be on the street and less murder will occur.

- HS Grad: Higher the education level, lower the murder rate

(c)

> Develop a new research question of your own that you can address using the `state` dataset. Clearly state the question you are going to address. Provide at least one visualization to support your exploration of this question. Discuss what you find.

Question

- Develops research question.

Answer

- Is there the regional difference in murder rate?

Question

- Creates visualization with appropriate title, labeled axe, and trend lines if applicable, printing data frames will not count.

Answer

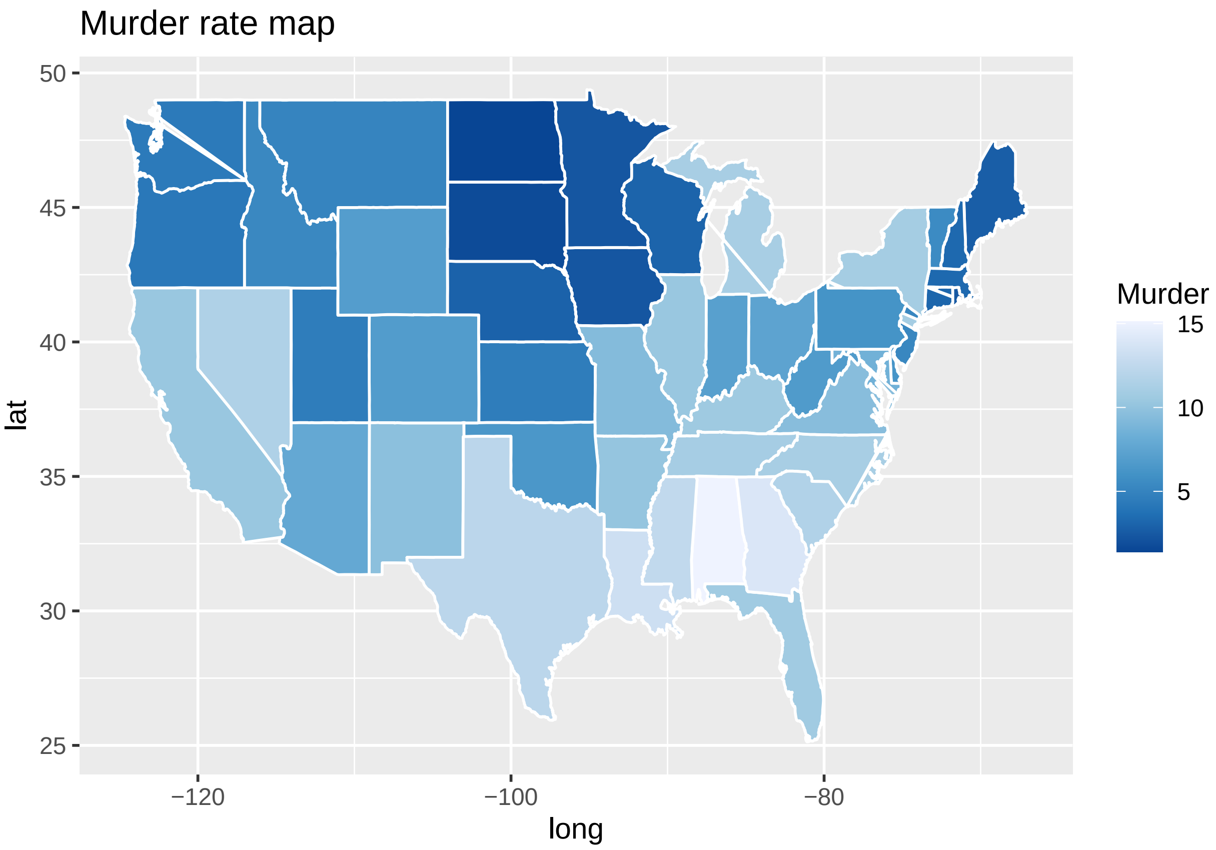
```r
library(maps)
```

```
##
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':
##
##     map
```

```r
library(mapdata)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
state_data$name %<>%
  tolower()

# combine the data with US map information

dt <- map_data("state") %>%
  left_join(state_data, by = c("region" = "name"))

ggplot(data = dt, aes(x= long, y= lat, fill = Murder, group = region)) +
  geom_polygon(color = "white") +
  scale_fill_distiller("Murder") +
  ggtitle("Murder rate map")
```



Question

- Discuss findings

Answer

- The murder rate in the northeast, north, and northwest areas is lower according to the plot.

Problem 2: Asking Data Science Questions: Crime and Educational Attainment In Problem Set 3, you joined data about crimes and educational attainment. Here you will use this new combined dataset to examine questions around crimes in Seattle and the educational attainment of people living in the areas in which the crime occurred. A standard dataset will be available on canvas after the problem set 3 due date.

(a) Develop a Data Science Question

Develop your own question to address in this analysis. Your question should be specific and measurable, and it should be able to be addressed through a basic analysis of the crime dataset you compiled in Problem Set 3.

Question

- Develop question for analysis.
- Question is specific and measurable.

Answer

```
# import the data
joint_data <- read.csv("/Users/Jay/Desktop/J/Working on/MSIM/IMT573/problem set/problemset4/joined_data
```

Answer

- What is the cumulated beat of each category in each year?

(b) Describe and Summarize

Briefly summarize the dataset, describing what data exists and its basic properties. Comment on any issues that need to be resolved before you can proceed with your analysis.

Question

- Summarizes dataset, describes what data exists and performs a descriptive analysis

Answer

```
#summary(join_data)
#str(joint_data)
colnames(joint_data)
```

```
##  [1] "Report.Number"
##  [2] "Occurred.Date"
##  [3] "Occurred.Time"
##  [4] "Reported.Date"
##  [5] "Reported.Time"
##  [6] "Crime.Subcategory"
##  [7] "Primary.Offense.Description"
##  [8] "Precinct"
##  [9] "Sector"
## [10] "Beat"
## [11] "Neighborhood"
## [12] "Year"
## [13] "censusId"
## [14] "Location.1"
## [15] "Latitude"
## [16] "Longitude"
## [17] "CensusCode"
## [18] "state"
## [19] "county"
## [20] "GEO_ID"
## [21] "Population.18.to.24.years"
## [22] "Population.18.to.24.years.Less.than.high.school.graduate"
## [23] "Population.18.to.24.years.High.school.graduate..includes.equivalency."
## [24] "Population.18.to.24.years.Some.college.or.associate.s.degree"
## [25] "Population.18.to.24.years.Bachelor.s.degree.or.higher"
## [26] "Population.25.years.and.over"
## [27] "Population.25.years.and.over.Less.than.9th.grade"
## [28] "Population.25.years.and.over.9th.to.12th.grade..no.diploma"
## [29] "Population.25.years.and.over.High.school.graduate..includes.equivalency."
## [30] "Population.25.years.and.over.Some.college..no.degree"
## [31] "Population.25.years.and.over.Associate.s.degree"
## [32] "Population.25.years.and.over.Bachelor.s.degree"
## [33] "Population.25.years.and.over.Graduate.or.professional.degree"
## [34] "Population.25.years.and.over.High.school.graduate.or.higher"
## [35] "Population.25.years.and.over.Bachelor.s.degree.or.higher"
```

- Report number [1]
- Time: [2] - [5], [12]
- Crime type: [6] - [7]
- Location of incident: [8] - [11]
- Location of population: [13] - [20]
- Population: [21], [26]

- Education level [22] - [25], [27] - [35]

Question

- Comments on any issues within data that need to be resolved

Answer

- I need to create one variables, count. In this case, one beat is one count, I then can further analyze if the number of crimes is based on the beat.

(c) Data Analysis

> Use the dataset to provide empirical evidence that helped address your question from part (a). Discuss your results. Provide at least one visualization to support your narrative.

Question

- Conducts analysis appropriate for chosen question.

Answer

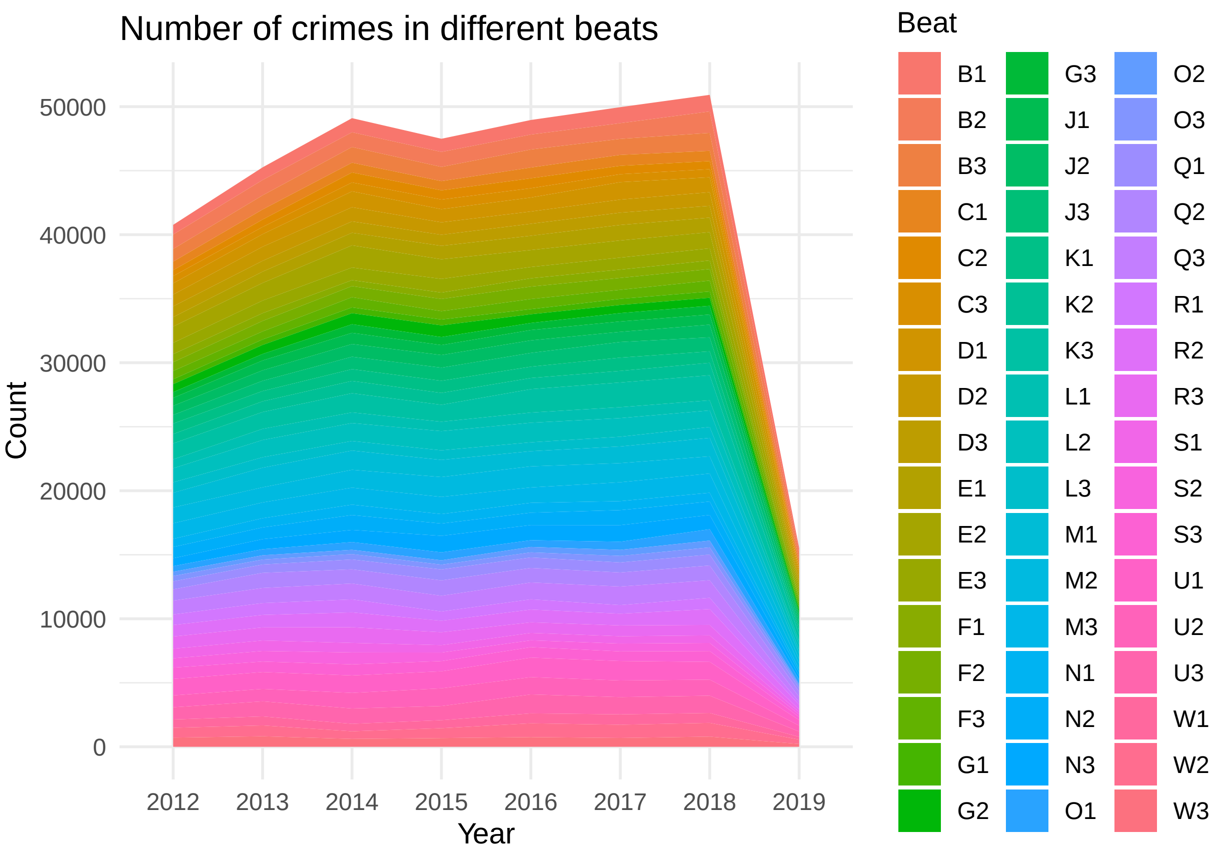
```r
joint_data$Count <- 1
joint_data$Year <- joint_data$Occurred.Date %>%
  str_sub(start = 7L, end = 10L)

tab <- aggregate(Count ~ Year + Beat, data = joint_data, FUN = sum)
```

Question

- Creates visualization with appropriate title, labeled axes, and trend lines if applicable, printing data frames will not count

Answer

```r
ggplot(tab, aes(x = Year, y = Count, fill = Beat, group = Beat))+
  geom_area()+
  theme_minimal()+
    labs(title = "Number of crimes in different beats")
```



Question

- Discussion of the result.

Answer

- Each beat has similar trend almost every year. There are more crimes in 2014-2018, and the data of 2019 is not complete. Also, from the plot, 2019 drop because the dataset doesn't have complete record of 2019.

(d) Reflect and Question

Comment on the questions (and answers) in this analysis. Were you able to answer all of these questions? Are all questions well defined? Is the data good enough to answer all these?

Question

- Reflects on question quality and analysis process with thoughtful discussion.

Answer

- Although the number of crimes of different beats is different, the time trends are similar to each other.

Question

- Answer all sections of the question.

Answer

```
library(reshape)
```

```
##
## Attaching package: 'reshape'
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

```
cast(tab, Year~Beat, value.var = "Count")
```

```
## Using Count as value column.  Use the value argument to cast to override this choice
```

```
##   Year   B1   B2   B3  C1  C2  C3   D1   D2  D3   E1   E2   E3  F1  F2  F3  G1
## 1 2012  753 1133  986 672 458 565  879  919 843  738 1270  883 630 684 648 380
## 2 2013  989 1198 1105 787 577 565  980 1102 843  883 1380  982 606 819 630 427
## 3 2014 1104 1152 1222 765 781 694 1240 1106 889  983 1734  985 465 869 795 449
## 4 2015 1029 1171 1109 721 716 755 1026  973 865 1048 1533 1068 499 955 650 455
## 5 2016 1137 1169 1408 841 774 728 1090  958 939 1115 1325  881 647 972 791 398
## 6 2017 1252 1227 1250 852 655 621 1361 1022 961 1210 1339  939 646 932 706 476
## 7 2018 1299 1675 1402 802 631 632 1167 1064 919 1139 1263  946 669 901 853 491
## 8 2019  391  528  364 254 197 166  323  347 271  343  356  348 170 297 288 181
##    G2  G3  J1  J2   J3   K1  K2   K3  L1   L2  L3   M1   M2   M3  N1   N2   N3
## 1 627 446 618 714  673  818 732 1248 695 1097 816 1148 1236 1216 637  879  615
## 2 704 507 769 819  799  803 837 1313 872 1346 817 1533 1181 1234 718  915  787
## 3 864 674 873 991  968  921 961 1499 831 1402 748 1501 1396 1327 814 1159  949
## 4 914 626 779 987 1022  949 929 1321 733 1501 747 1339 1550 1342 741  960 1289
## 5 681 549 808 980 1075  884 871 1829 793 1531 694 1183 1648 1201 770  970 1180
## 6 636 652 811 794 1220 1035 906 1931 852 1455 759 1304 1492 1475 702 1174 1294
## 7 624 684 790 990 1094  896 973 1958 787 1296 850 1442 1341 1502 687 1052 1111
## 8 209 189 220 310  352  305 351  512 260  421 279  433  366  426 234  300  292
##    O1  O2  O3  Q1   Q2   Q3   R1   R2   R3  S1  S2  S3   U1   U2   U3  W1   W2
## 1 439 330 417 623  883 1082  824  901  952 750 728 888 1282  928  951 648  771
## 2 473 334 386 651 1182 1180  925  976 1025 817 827 818 1320  969 1180 702  847
## 3 604 332 423 769 1111 1238 1015 1160 1224 742 933 867 1362 1219 1196 584  596
## 4 621 340 372 855 1203 1227  726  897 1008 579 668 798 1329 1384 1104 626  773
## 5 526 424 398 818 1135 1318  799  997  830 578 534 816 1526 1350 1480 775 1079
## 6 636 470 513 783 1095 1452  672  918  842 573 613 752 1521 1298 1346 813 1032
## 7 886 530 560 857 1151 1368  884 1238  818 590 634 846 1386 1255 1349 772 1076
## 8 338 134 170 291  399  384  235  361  210 190 202 242  431  463  413 215  340
##    W3
## 1 715
## 2 820
## 3 615
## 4 679
## 5 754
## 6 689
## 7 792
## 8 222
```

Answer

- Through this answer, we can find out exact number difference of each beat of each year compare to different year as well as each beat compare to each other.