analyticsedge

Mridul Jain

07/02/2023

Table of contents

1

2 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

Assignment 1 Answers

An Analytical Detective

There are two main types of crimes: violent crimes, and property crimes. In this problem, we'll focus on one specific type of property crime, called "motor vehicle theft" (sometimes referred to as grand theft auto). This is the act of stealing, or attempting to steal, a car. In this problem, we'll use some basic data analysis in R to understand the motor vehicle thefts in Chicago.

Please download the file mvtWeek1.csv for this problem (do not open this file in any spreadsheet software before completing this problem because it might change the format of the Date field).

Start:
Read the dataset mvtWeek1.csv into R, using the read.csv function, and call the data frame "mvt".

```
mvt <- read.csv("mvtWeek1.csv")
```

1.1: How many rows of data (observations) are in this dataset?
Answer: 191641

```
nrow(mvt)
```

```
[1] 191641
```

1.2: How many variables are in this dataset?
Answer: 11

```
ncol(mvt)
```

```
[1] 11
```

1.3: Using the "max" function, what is the maximum value of the variable "ID"?
Answer: 9181151

```
max(mvt$ID)
```

```
[1] 9181151
```

1.4: What is the minimum value of the variable "Beat"?
Answer: 111

```
min(mvt$Beat)
```

```
[1] 111
```

1.5: How many observations have value TRUE in the Arrest variable (this is the number of crimes for which an arrest was made)?
Answer: 15536

```
sum(mvt$Arrest)
```

```
[1] 15536
```

1.6: How many observations have a LocationDescription value of ALLEY?
Answer: 2308

```
sum(mvt$LocationDescription == "ALLEY")
```

```
[1] 2308
```

2.1: In what format are the entries in the variable Date?
Answer: Month/Day/Year Hour:Minute

```
mvt$Date[1]
```

```
[1] "12/31/12 23:15"
```

2.2: What is the month and year of the median date in our dataset? Enter your answer as "Month Year", without the quotes.
Answer: May 2006

```
DateConvert = as.Date(strptime(mvt$Date, "%m/%d/%y %H:%M"))
#summary(DateConvert)
median(DateConvert)
```

```
[1] "2006-05-21"
```

2.3: In which month did the fewest motor vehicle thefts occur?
Answer: February

```
mvt$Month = months(DateConvert)
mvt$Weekday = weekdays(DateConvert)
mvt$Date = DateConvert
table(mvt$Month)
```

```
  April    August  December  February   January      July      June     March
  15280     16572     16426     13511     16047     16801     16002     15758
    May  November   October September
  16035     16063     17086     16060
```

2.4: On which weekday did the most motor vehicle thefts occur?
Answer: Friday

```
table(mvt$Weekday)
```

```
 Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
  29284     27397     27118     26316     27319     26791     27416
```

2.5: Which month has the largest number of motor vehicle thefts for which an arrest was made?
Answer: January

```
table(mvt$Month, mvt$Arrest)
```

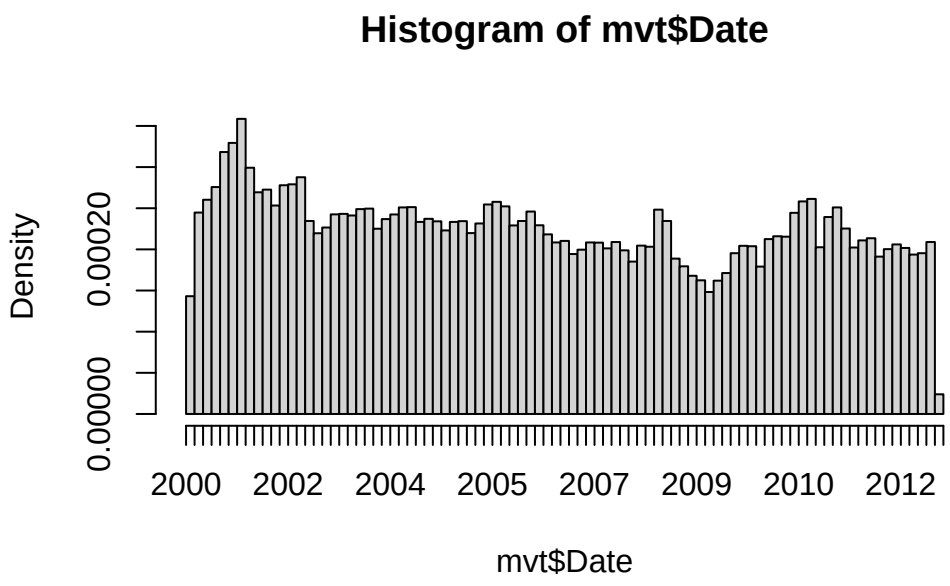
```
           FALSE  TRUE
April      14028  1252
August     15243  1329
December   15029  1397
February   12273  1238
January    14612  1435
July       15477  1324
June       14772  1230
March      14460  1298
May        14848  1187
November   14807  1256
October    15744  1342
September  14812  1248
```

3.1.1: In general, does it look like crime increases or decreases from 2002 - 2012?
Answer: Decreases

3.1.2: In general, does it look like crime increases or decreases from 2005 - 2008?
Answer: Decreases

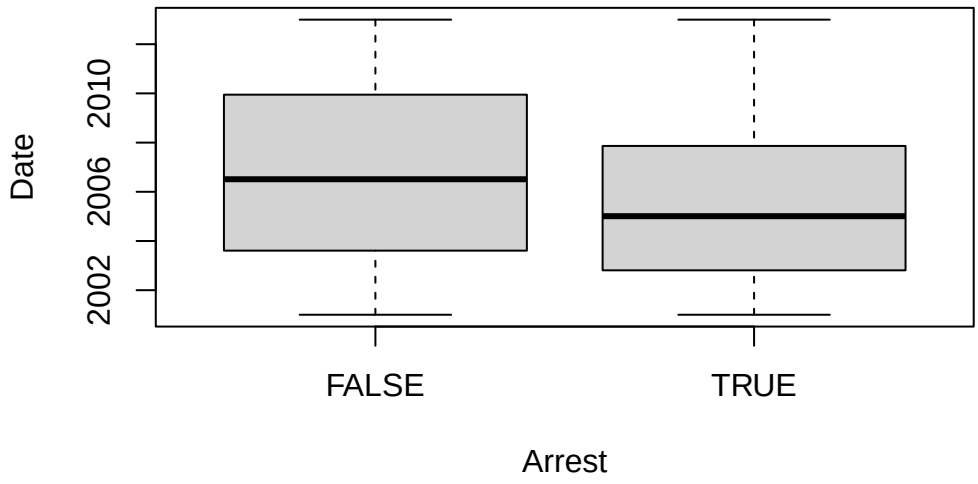
3.1.3: In general, does it look like crime increases or decreases from 2009 - 2011?
Answer: Increases

```r
hist(mvt$Date, breaks=100)
```



3.2: Does it look like there were more crimes for which arrests were made in the first half of the time period or the second half of the time period?
Answer: First half

```
boxplot(Date ~ Arrest, data = mvt)
```



3.3: For what proportion of motor vehicle thefts in 2001 was an arrest made?
Answer: 0.1041173

```
table(mvt$Year, mvt$Arrest)
```

```
      FALSE  TRUE
2001 18517  2152
2002 16638  2115
2003 14859  1798
2004 15169  1693
2005 14956  1528
2006 14796  1302
2007 13068  1212
2008 13425  1020
2009 11327   840
2010 14796   701
2011 15012   625
2012 13542   550
```

```
2152/(18517+2152)
```

```
[1] 0.1041173
```

3.4: For what proportion of motor vehicle thefts in 2007 was an arrest made?
Answer: 0.08487395

```
table(mvt$Year, mvt$Arrest)
```

```
      FALSE  TRUE
  2001 18517  2152
  2002 16638  2115
  2003 14859  1798
  2004 15169  1693
  2005 14956  1528
  2006 14796  1302
  2007 13068  1212
  2008 13425  1020
  2009 11327   840
  2010 14796   701
  2011 15012   625
  2012 13542   550
```

```
1212/(13068+1212)
```

```
[1] 0.08487395
```

3.5: For what proportion of motor vehicle thefts in 2012 was an arrest made?
Answer: 0.03902924

```
table(mvt$Year, mvt$Arrest)
```

```
      FALSE  TRUE
  2001 18517  2152
  2002 16638  2115
```

```
2003 14859  1798
2004 15169  1693
2005 14956  1528
2006 14796  1302
2007 13068  1212
2008 13425  1020
2009 11327   840
2010 14796   701
2011 15012   625
2012 13542   550
```

```
550/(13542+550)
```

```
[1] 0.03902924
```

4.1: Which locations are the top five locations for motor vehicle thefts, excluding the "Other" category?
Answer: STREET, PARKING LOT/GARAGE(NON.RESID.), ALLEY, GAS STATION, DRIVEWAY - RESIDENTIAL

```
sort(table(mvt$LocationDescription), decreasing = TRUE)
```

```
                          STREET
                          156564
PARKING LOT/GARAGE(NON.RESID.)
                           14852
                           OTHER
                            4573
                           ALLEY
                            2308
                     GAS STATION
                            2111
          DRIVEWAY - RESIDENTIAL
                            1675
 RESIDENTIAL YARD (FRONT/BACK)
                            1536
                       RESIDENCE
                            1302
                RESIDENCE-GARAGE
```

	1176
VACANT LOT/LAND	985
VEHICLE NON-COMMERCIAL	817
SIDEWALK	462
CHA PARKING LOT/GROUNDS	405
AIRPORT/AIRCRAFT	363
POLICE FACILITY/VEH PARKING LOT	266
PARK PROPERTY	255
SCHOOL, PUBLIC, GROUNDS	206
APARTMENT	184
SPORTS ARENA/STADIUM	166
CTA GARAGE / OTHER PROPERTY	148
COMMERCIAL / BUSINESS OFFICE	126
HOTEL/MOTEL	124
SCHOOL, PUBLIC, BUILDING	114
HOSPITAL BUILDING/GROUNDS	101
GROCERY FOOD STORE	80
CHURCH/SYNAGOGUE/PLACE OF WORSHIP	56
RESTAURANT	49
GOVERNMENT BUILDING/PROPERTY	48
COLLEGE/UNIVERSITY GROUNDS	47
CAR WASH	44

CONSTRUCTION SITE	35
SMALL RETAIL STORE	33
OTHER RAILROAD PROP / TRAIN DEPOT	28
AIRPORT EXTERIOR - NON-SECURE AREA	24
SCHOOL, PRIVATE, GROUNDS	23
VEHICLE-COMMERCIAL	23
DEPARTMENT STORE	22
HIGHWAY/EXPRESSWAY	22
NURSING HOME/RETIREMENT HOME	21
TAXICAB	21
MOVIE HOUSE/THEATER	18
RESIDENCE PORCH/HALLWAY	18
BAR OR TAVERN	17
WAREHOUSE	17
FACTORY/MANUFACTURING BUILDING	16
SCHOOL, PRIVATE, BUILDING	14
TAVERN/LIQUOR STORE	14
AIRPORT PARKING LOT	11
AIRPORT VENDING ESTABLISHMENT	10
ATHLETIC CLUB	9
DRUG STORE	8
OTHER COMMERCIAL TRANSPORTATION	

8
BANK
7
CONVENIENCE STORE
7
FOREST PRESERVE
6
AIRPORT TERMINAL UPPER LEVEL - NON-SECURE AREA
5
CHA APARTMENT
5
DAY CARE CENTER
5
FIRE STATION
5
ABANDONED BUILDING
4
AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA
4
BARBERSHOP
4
LAKEFRONT/WATERFRONT/RIVERBANK
4
LIBRARY
4
SAVINGS AND LOAN
4
BOWLING ALLEY
3
CLEANING STORE
3
MEDICAL/DENTAL OFFICE
3
BRIDGE
2
COLLEGE/UNIVERSITY RESIDENCE HALL
2
CURRENCY EXCHANGE
2
AIRPORT BUILDING NON-TERMINAL - SECURE AREA
1
AIRPORT EXTERIOR - SECURE AREA
1

```
                    ANIMAL HOSPITAL
                                  1
                    APPLIANCE STORE
                                  1
                          CTA TRAIN
                                  1
         JAIL / LOCK-UP FACILITY
                                  1
                          NEWSSTAND
                                  1
```

```
#only show top 6?
```

Create a subset of your data, only taking observations for which the theft happened in one of these five locations, and call this new data set "Top5".

```
Top5 <- subset(mvt, mvt$LocationDescription == "STREET"
       | mvt$LocationDescription == "PARKING LOT/GARAGE(NON.RESID.)"
       | mvt$LocationDescription == "ALLEY"
       | mvt$LocationDescription == "GAS STATION"
       | mvt$LocationDescription == "DRIVEWAY - RESIDENTIAL")
```

4.2: How many observations are in Top5?
Answer: 177510

```
nrow(Top5)
```

```
[1] 177510
```

4.3: One of the locations has a much higher arrest rate than the other locations. Which is it?
Answer: Gas Station (Check percentages)

```
Top5$LocationDescription = factor(Top5$LocationDescription)
table(Top5$LocationDescription, Top5$Arrest)
```

```
                                FALSE   TRUE
  ALLEY                          2059    249
```

```
DRIVEWAY - RESIDENTIAL           1543    132
GAS STATION                      1672    439
PARKING LOT/GARAGE(NON.RESID.)  13249   1603
STREET                         144969  11595
```

4.4: On which day of the week do the most motor vehicle thefts at gas stations happen?
Answer: Saturday

```
table(Top5$LocationDescription == "GAS STATION", Top5$Weekday)
```

```
       Friday Monday Saturday Sunday Thursday Tuesday Wednesday
FALSE  26746  25008    24917  24220    24956   24527     25025
TRUE     332    280      338    336      282     270       273
```

4.5: On which day of the week do the fewest motor vehicle thefts in residential driveways happen?
Answer: Saturday

```
table(Top5$LocationDescription == "DRIVEWAY - RESIDENTIAL", Top5$Weekday)
```

```
       Friday Monday Saturday Sunday Thursday Tuesday Wednesday
FALSE  26821  25033    25053  24335    24975   24554     25064
TRUE     257    255      202    221      263     243       234
```

Stock Dynamics

A stock market is where buyers and sellers trade shares of a company, and is one of the most popular ways for individuals and companies to invest money. The size of the world stock market is now estimated to be in the trillions. The largest stock market in the world is the New York Stock Exchange (NYSE), located in New York City. About 2,800 companies are listed on the NYSE. In this problem, we'll look at the monthly stock prices of five of these companies: IBM, General Electric (GE), Procter and Gamble, Coca Cola, and Boeing. The data used in this problem comes from Infochimps.

Please download the following files: IBMStock.csv, GEStock.csv, ProcterGambleStock.csv, CocaColaStock.csv, BoeingStock.csv (do not open these files in any spreadsheet software before completing this problem because it might change the format of the Date field).

Start:
1. Read the datasets into R, using the read.csv function, and call the data frames "IBM", "GE", "ProcterGamble", "CocaCola", and "Boeing", respectively.

```
IBM <- read.csv("IBMStock.csv")
GE <- read.csv("GEStock.csv")
ProcterGamble <- read.csv("ProcterGambleStock.csv")
CocaCola <- read.csv("CocaColaStock.csv")
Boeing <- read.csv("BoeingStock.csv")
```

2. Before working with these data sets, we need to convert the dates into a format that R can understand. Take a look at the structure of one of the datasets using the str function. Right now, the date variable is stored as a factor. We can convert this to a "Date" object in R by using the following five commands (one for each data set):

```
IBM$Date = as.Date(IBM$Date, "%m/%d/%y")
GE$Date = as.Date(GE$Date, "%m/%d/%y")
CocaCola$Date = as.Date(CocaCola$Date, "%m/%d/%y")
ProcterGamble$Date = as.Date(ProcterGamble$Date, "%m/%d/%y")
Boeing$Date = as.Date(Boeing$Date, "%m/%d/%y")
```

1.1: Our five datasets all have the same number of observations. How many observations are there in each data set?
Answer: 480

```
nrow(IBM)
```

```
[1] 480
```

```
# According to the assignment, use: str(IBM)
# We only need to use the command for one of the datasets, since they all have the same nu
```

1.2: What is the earliest year in our datasets?
Answer: 1970

```
min(IBM$Date)
```

```
[1] "1970-01-01"
```

```
# According to the assignment, use: summary(IBM$Date)
# Again, we only need to use the command for one of the datsets, since the observations st
```

1.3: What is the latest year in our datasets?
Answer: 2009

```
max(IBM$Date)
```

```
[1] "2009-12-01"
```

```
# According to the assignment, use: summary(IBM$Date)
# Again, we only need to use the command for one of the datsets, since the observations en
```

1.4: What is the mean stock price of IBM over this time period?
Answer: 144.375

```
mean(IBM$StockPrice)
```

```
[1] 144.375
```

```
# According to the assignment, use: summary(IBM$StockPrice)
```

1.5:
Answer: 9.293636

```
min(GE$StockPrice)
```

```
[1] 9.293636
```

```
# According to the assignment, use: summary(GE$StockPrice)
```

1.6:
Answer: 146.5843

```
max(CocaCola$StockPrice)
```

```
[1] 146.5843
```

```
# According to the assignment, use: summary(CocaCola$StockPrice)
```

1.7:
Answer:

```
median(Boeing$StockPrice)
```

```
[1] 44.8834
```

```
# According to the assignment, use: summary(Boeing$StockPrice)
```

1.8:
Answer: 18.19414

```
sd(ProcterGamble$StockPrice)
```

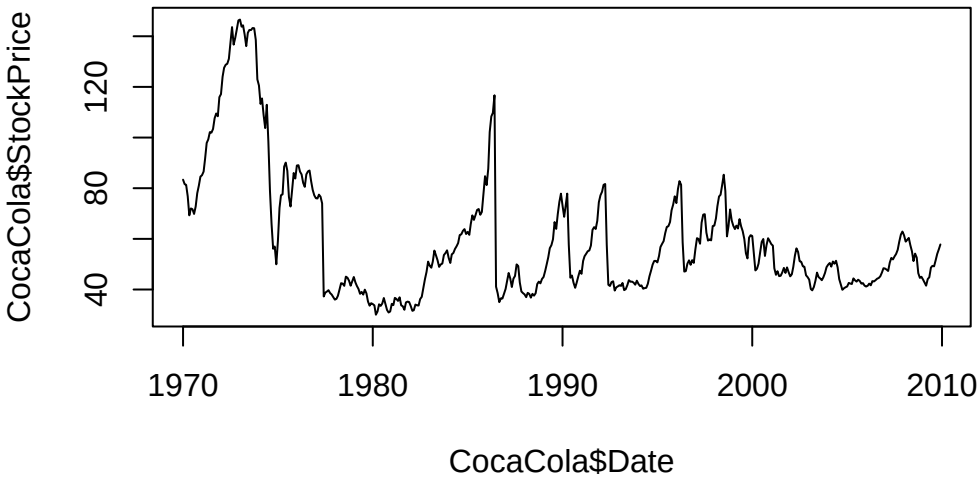
```
[1] 18.19414
```

Side note: According to the assignment, questions 1.2 - 1.7 should've been solved using the summary function. However, I used commands that would give more accurate answer. Along with the commands I used, I also wrote how the assignment could be solved using the summary function.

2.1.1: Around what year did Coca-Cola has its highest stock price in this time period?
Answer: 1973

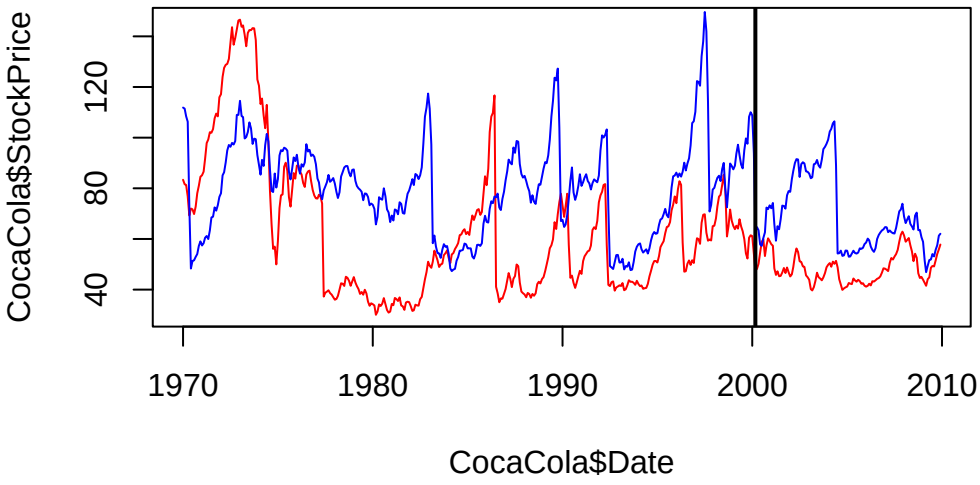
2.1.2: Around what year did Coca-Cola has its lowest stock price in this time period?
Answer: 1980

```
plot(CocaCola$Date, CocaCola$StockPrice, "l")
```



2.2: In March of 2000, the technology bubble burst, and a stock market crash occurred. According to this plot, which company's stock dropped more?
Answer: Procter and Gamble

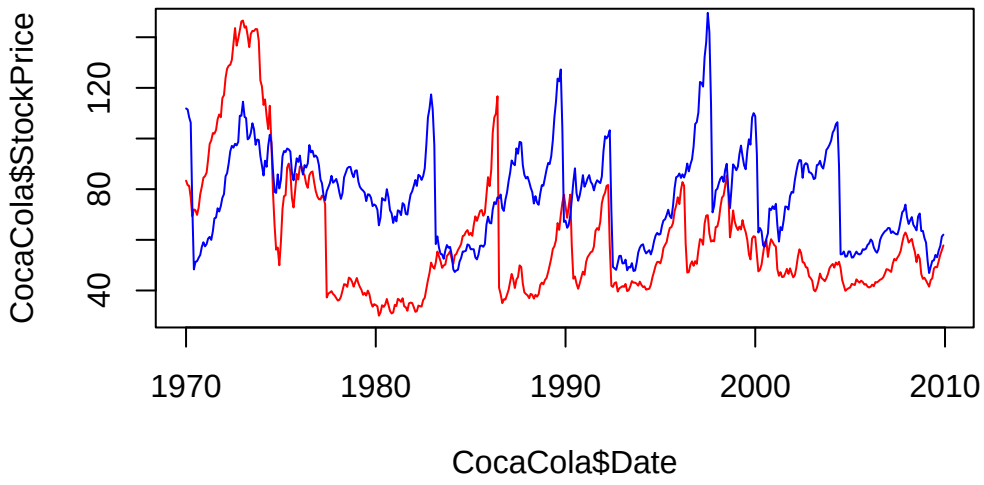
```
plot(CocaCola$Date, CocaCola$StockPrice, "l", col = "red")
lines(ProcterGamble$Date, ProcterGamble$StockPrice, col = "blue")
abline(v=as.Date(c("2000-03-01")), lwd=2)
```



2.3.1: Around 1983, the stock for one of these companies (Coca-Cola or Procter and Gamble) was going up, while the other was going down. Which one was going up?
Answer: CocaCola

2.3.1: In the time period shown in the plot, which stock generally has lower values?
Answer: CocaCola

```
plot(CocaCola$Date, CocaCola$StockPrice, "l", col = "red")
lines(ProcterGamble$Date, ProcterGamble$StockPrice, col = "blue")
```



Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.