Project 7

Meera, Lauren, Namaluba

December 1, 2017

# 1

Start by uploading the LA metro data

Q2 <- read.csv("~/290/gp04g4/scratch/nmm/la\_metro\_gbfs\_trips\_Q2\_2017.csv", header=T)  
head(Q2)

## trip\_id duration start\_time end\_time start\_station  
## 1 32815764 18 2017-06-25 19:53:00 2017-06-25 20:11:00 3047  
## 2 32821341 25 2017-06-25 20:35:00 2017-06-25 21:00:00 3005  
## 3 31652471 10 2017-06-16 08:41:00 2017-06-16 08:51:00 3023  
## 4 31700167 4 2017-06-16 17:39:00 2017-06-16 17:43:00 3005  
## 5 31717085 17 2017-06-16 20:17:00 2017-06-16 20:34:00 3051  
## 6 31786532 6 2017-06-17 09:20:00 2017-06-17 09:26:00 3005  
## start\_lat start\_lon end\_station end\_lat end\_lon bike\_id plan\_duration  
## 1 34.03998 -118.2664 3005 34.04850 -118.2585 4727 30  
## 2 34.04850 -118.2585 3020 34.03105 -118.2671 4727 30  
## 3 34.05091 -118.2410 3005 34.04850 -118.2585 4727 30  
## 4 34.04850 -118.2585 3051 34.04542 -118.2535 4727 30  
## 5 34.04542 -118.2535 3005 34.04850 -118.2585 4727 0  
## 6 34.04850 -118.2585 3007 34.05048 -118.2546 4727 30  
## trip\_route\_category passholder\_type  
## 1 One Way Monthly Pass  
## 2 One Way Monthly Pass  
## 3 One Way Monthly Pass  
## 4 One Way Monthly Pass  
## 5 One Way Walk-up  
## 6 One Way Monthly Pass

# 1.a

startstop <- data.frame(Q2$start\_station)  
startstop$Q2.end\_station <- Q2$end\_station  
nrow(unique(startstop))

## [1] 3139

There are 3,139 unique start/end station pairs.

# 1.b

head(sort(table(paste(Q2$start\_station, Q2$end\_station)), decreasing = T))

##   
## 3030 3014 3031 3005 3082 3082 3014 3030 3066 3066 3048 3048   
## 282 280 272 260 216 214

Station 3030 to 3014 is the most popular start/stop pair.

# 1.c

sum(table(paste(Q2$start\_station, Q2$end\_station) == 1))

## [1] 51918

51,918 start/end station pairs have only been used once.

# 2

Start by loading the co2 data, and putting it into a matrix

data("co2")  
co2M <- matrix((co2), ncol = 12, byrow = T)  
head(co2M)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]  
## [1,] 315.42 316.31 316.50 317.56 318.13 318.00 316.39 314.65 313.68 313.18  
## [2,] 316.27 316.81 317.42 318.87 319.87 319.43 318.01 315.74 314.00 313.68  
## [3,] 316.73 317.54 318.38 319.31 320.42 319.61 318.42 316.63 314.83 315.16  
## [4,] 317.78 318.40 319.53 320.42 320.85 320.45 319.45 317.25 316.11 315.27  
## [5,] 318.58 318.92 319.70 321.22 322.08 321.31 319.58 317.61 316.05 315.83  
## [6,] 319.41 320.07 320.74 321.40 322.06 321.73 320.27 318.54 316.54 316.71  
## [,11] [,12]  
## [1,] 314.66 315.43  
## [2,] 314.84 316.03  
## [3,] 315.94 316.85  
## [4,] 316.53 317.53  
## [5,] 316.91 318.20  
## [6,] 317.53 318.55

# 2.a

YearlyAvg <- apply(co2M, 1, mean)  
YearlyAvg

## [1] 315.8258 316.7475 317.4850 318.2975 318.8325 319.4625 319.8725  
## [8] 321.2100 322.0200 322.8900 324.4592 325.5175 326.1550 327.2933  
## [15] 329.5117 330.0792 330.9858 331.9858 333.7300 335.3358 336.6808  
## [22] 338.5150 339.7608 340.9592 342.6083 344.2467 345.7258 346.9750  
## [29] 348.7508 351.3133 352.7542 354.0367 355.4783 356.2917 356.9958  
## [36] 358.8800 360.9142 362.6867 363.8175

Year <- 1959:1997  
YearDF <- data.frame(Year, YearlyAvg)  
YearDF

## Year YearlyAvg  
## 1 1959 315.8258  
## 2 1960 316.7475  
## 3 1961 317.4850  
## 4 1962 318.2975  
## 5 1963 318.8325  
## 6 1964 319.4625  
## 7 1965 319.8725  
## 8 1966 321.2100  
## 9 1967 322.0200  
## 10 1968 322.8900  
## 11 1969 324.4592  
## 12 1970 325.5175  
## 13 1971 326.1550  
## 14 1972 327.2933  
## 15 1973 329.5117  
## 16 1974 330.0792  
## 17 1975 330.9858  
## 18 1976 331.9858  
## 19 1977 333.7300  
## 20 1978 335.3358  
## 21 1979 336.6808  
## 22 1980 338.5150  
## 23 1981 339.7608  
## 24 1982 340.9592  
## 25 1983 342.6083  
## 26 1984 344.2467  
## 27 1985 345.7258  
## 28 1986 346.9750  
## 29 1987 348.7508  
## 30 1988 351.3133  
## 31 1989 352.7542  
## 32 1990 354.0367  
## 33 1991 355.4783  
## 34 1992 356.2917  
## 35 1993 356.9958  
## 36 1994 358.8800  
## 37 1995 360.9142  
## 38 1996 362.6867  
## 39 1997 363.8175

# 2.b

Month <- 1:12  
MonthlyAvg <- apply(co2M, 2, mean)  
MonthlyAvg

## [1] 336.4308 337.2033 338.0546 339.2944 339.8821 339.3282 337.9164  
## [8] 335.9579 334.2428 334.1692 335.4679 336.6946

MonthDF <- data.frame(Month, MonthlyAvg)  
MonthDF

## Month MonthlyAvg  
## 1 1 336.4308  
## 2 2 337.2033  
## 3 3 338.0546  
## 4 4 339.2944  
## 5 5 339.8821  
## 6 6 339.3282  
## 7 7 337.9164  
## 8 8 335.9579  
## 9 9 334.2428  
## 10 10 334.1692  
## 11 11 335.4679  
## 12 12 336.6946

# 3

Start by loading the 2008 Airport data.

"2008" <- read.csv("~/290/gp01g7/scratch/nmmfolder/2008.csv", header=T)  
head(`2008`)

## Year Month DayofMonth DayOfWeek DepTime CRSDepTime ArrTime CRSArrTime  
## 1 2008 1 3 4 2003 1955 2211 2225  
## 2 2008 1 3 4 754 735 1002 1000  
## 3 2008 1 3 4 628 620 804 750  
## 4 2008 1 3 4 926 930 1054 1100  
## 5 2008 1 3 4 1829 1755 1959 1925  
## 6 2008 1 3 4 1940 1915 2121 2110  
## UniqueCarrier FlightNum TailNum ActualElapsedTime CRSElapsedTime AirTime  
## 1 WN 335 N712SW 128 150 116  
## 2 WN 3231 N772SW 128 145 113  
## 3 WN 448 N428WN 96 90 76  
## 4 WN 1746 N612SW 88 90 78  
## 5 WN 3920 N464WN 90 90 77  
## 6 WN 378 N726SW 101 115 87  
## ArrDelay DepDelay Origin Dest Distance TaxiIn TaxiOut Cancelled  
## 1 -14 8 IAD TPA 810 4 8 0  
## 2 2 19 IAD TPA 810 5 10 0  
## 3 14 8 IND BWI 515 3 17 0  
## 4 -6 -4 IND BWI 515 3 7 0  
## 5 34 34 IND BWI 515 3 10 0  
## 6 11 25 IND JAX 688 4 10 0  
## CancellationCode Diverted CarrierDelay WeatherDelay NASDelay  
## 1 0 NA NA NA  
## 2 0 NA NA NA  
## 3 0 NA NA NA  
## 4 0 NA NA NA  
## 5 0 2 0 0  
## 6 0 NA NA NA  
## SecurityDelay LateAircraftDelay  
## 1 NA NA  
## 2 NA NA  
## 3 NA NA  
## 4 NA NA  
## 5 0 32  
## 6 NA NA

# 3.a

`2008`$Date <- as.Date(with(`2008`, paste(Year, Month, DayofMonth, sep="-")), "%Y-%m-%d")  
head(`2008`)

## Year Month DayofMonth DayOfWeek DepTime CRSDepTime ArrTime CRSArrTime  
## 1 2008 1 3 4 2003 1955 2211 2225  
## 2 2008 1 3 4 754 735 1002 1000  
## 3 2008 1 3 4 628 620 804 750  
## 4 2008 1 3 4 926 930 1054 1100  
## 5 2008 1 3 4 1829 1755 1959 1925  
## 6 2008 1 3 4 1940 1915 2121 2110  
## UniqueCarrier FlightNum TailNum ActualElapsedTime CRSElapsedTime AirTime  
## 1 WN 335 N712SW 128 150 116  
## 2 WN 3231 N772SW 128 145 113  
## 3 WN 448 N428WN 96 90 76  
## 4 WN 1746 N612SW 88 90 78  
## 5 WN 3920 N464WN 90 90 77  
## 6 WN 378 N726SW 101 115 87  
## ArrDelay DepDelay Origin Dest Distance TaxiIn TaxiOut Cancelled  
## 1 -14 8 IAD TPA 810 4 8 0  
## 2 2 19 IAD TPA 810 5 10 0  
## 3 14 8 IND BWI 515 3 17 0  
## 4 -6 -4 IND BWI 515 3 7 0  
## 5 34 34 IND BWI 515 3 10 0  
## 6 11 25 IND JAX 688 4 10 0  
## CancellationCode Diverted CarrierDelay WeatherDelay NASDelay  
## 1 0 NA NA NA  
## 2 0 NA NA NA  
## 3 0 NA NA NA  
## 4 0 NA NA NA  
## 5 0 2 0 0  
## 6 0 NA NA NA  
## SecurityDelay LateAircraftDelay Date  
## 1 NA NA 2008-01-03  
## 2 NA NA 2008-01-03  
## 3 NA NA 2008-01-03  
## 4 NA NA 2008-01-03  
## 5 0 32 2008-01-03  
## 6 NA NA 2008-01-03

# 3.b

head(sort(tapply(`2008`$DepDelay, `2008`$Date, mean, na.rm=T), decreasing=T))

## 2008-12-21 2008-12-19 2008-12-23 2008-12-27 2008-11-30 2008-12-20   
## 38.86555 38.84811 35.49315 34.70340 34.19139 34.15359

The highest average departure delay was 38.87 minutes on December 21st.

# 4

"2015" <- read.csv("~/290/gp02g4/scratch/nmmfolder/2015.csv", header=T)  
`2015`$date <- format(as.POSIXct(`2015`$tpep\_pickup\_datetime, format="%Y-%m-%d %H:%M:%S"), "%Y-%m-%d")  
`2015`$passengers <- as.numeric(levels(`2015`$passenger\_count))[`2015`$passenger\_count]

## Warning: NAs introduced by coercion

head(tapply(`2015`$passengers, `2015`$date, mean, na.rm=T))

## 2015-01-01 2015-01-02 2015-01-03 2015-01-04 2015-01-05 2015-01-06   
## 1.807282 1.757090 1.775246 1.720275 1.652919 1.645666

5ab.) The data was imported, and a header was added using the metadata. The header of the file is shown below.

campdata <- read.delim("/home/washin41/290/gp07g3/scratch/washin41/camp.txt", sep = "|", header = FALSE, col.names = c("CMTE\_ID","AMNDT\_IND","RPT\_TP","TRANSACTION\_PGI","IMAGE\_NUM","TRANSACTION\_TP","ENTITY\_TP","NAME","CITY","STATE","ZIP\_CODE","EMPLOYER","OCCUPATION","TRANSACTION\_DT","TRANSACTION\_AMT","OTHER\_ID","TRAN\_ID","FILE\_NUM","MEMO\_CD","MEMO\_TEXT","SUB\_ID"))  
head(campdata)

## CMTE\_ID AMNDT\_IND RPT\_TP TRANSACTION\_PGI IMAGE\_NUM TRANSACTION\_TP  
## 1 C00581397 A YE P 2.016023e+17 15  
## 2 C00581397 A YE P 2.016023e+17 15  
## 3 C00575233 A YE P 2.017013e+17 15E  
## 4 C00573758 A YE P 2.017020e+17 15  
## 5 C00031054 A YE P 2.016051e+17 15  
## 6 C00276311 N YE P 2.016013e+17 15  
## ENTITY\_TP NAME CITY STATE ZIP\_CODE  
## 1 IND KONKUS, JOHN SR. POOLESVILLE MD 20837  
## 2 IND PERRY, PAMELA PANAMA CITY FL 32408  
## 3 IND SIEGEL, SHIELA SEMINOLE FL 337722012  
## 4 IND PEW, ROBERT D RESTON VA 201911342  
## 5 ORG DNC SERVICES CORP. WASHINGTON DC 20003  
## 6 IND KENNEDY, CHARLIE SCOTTSBLUFF NE 69361  
## EMPLOYER OCCUPATION  
## 1 RETIRED RETIRED  
## 2 SELF BEACHSIDE INTERIORS  
## 3   
## 4 N/A RETIRED  
## 5   
## 6 BLUE CROSS BLUE SHIELD OF NE PROVIDER RELATIONSHIP MANAGER  
## TRANSACTION\_DT TRANSACTION\_AMT OTHER\_ID TRAN\_ID FILE\_NUM  
## 1 12232016 500 SA11AI.5882 1052510  
## 2 12142016 1054 SA11AI.5884 1052510  
## 3 12312016 2500 C00401224 VPFF8GFPG72 1142878  
## 4 12182016 50 SA0207172822526 1147915  
## 5 12312016 15158 11AI-000155577 1071844  
## 6 12312016 264 SA030816270932 1054227  
## MEMO\_CD MEMO\_TEXT SUB\_ID  
## 1 X 4.02292e+18  
## 2 X 4.02292e+18  
## 3 \* EARMARKED CONTRIBUTION: SEE BELOW 4.01302e+18  
## 4 2.02082e+18  
## 5 X MI PARTY VICTORY FUND 4.05132e+18  
## 6 2.03082e+18

5c.) First, tapply was used to sum the transactions and index them by state. Then, the grep was used to find the max, and the results were used to extract the cell containing the max. California had the most contributions, contributing 6134211 dollars.

fundtab <- tapply(campdata$TRANSACTION\_AMT, campdata$STATE, sum)  
fundtab[grep(max(fundtab),fundtab)]

## CA   
## 6134211

5d.) AWK was used to double check the sum obtained in question 5c. The command used was the following:

# cat camp.txt | cut -d"|" -f15,10 | awk -F"|" '{myarray[$1]+=$2} END {for (n in myarray) {print n,myarray[n]}}' | sort -k2 -n

6a.)

# First, the database was connected to  
  
library("RMySQL")

## Loading required package: DBI

myDriver <- dbDriver("MySQL")  
con <- dbConnect(myDriver, host="mydb.ics.purdue.edu", username="washin41",dbname="washin41")  
  
# Then, SQL was used to find the data needed  
  
q6 <- dbSendQuery(con, "SELECT b.H, b.2B, b.3B, b.HR, b.playerID, b.yearID FROM Batting b WHERE b.playerID='aaronha01';")  
q6out <- fetch(q6)  
q6out

## H 2B 3B HR playerID yearID  
## 1 131 27 6 13 aaronha01 1954  
## 2 189 37 9 27 aaronha01 1955  
## 3 200 34 14 26 aaronha01 1956  
## 4 198 27 6 44 aaronha01 1957  
## 5 196 34 4 30 aaronha01 1958  
## 6 223 46 7 39 aaronha01 1959  
## 7 172 20 11 40 aaronha01 1960  
## 8 197 39 10 34 aaronha01 1961  
## 9 191 28 6 45 aaronha01 1962  
## 10 201 29 4 44 aaronha01 1963  
## 11 187 30 2 24 aaronha01 1964  
## 12 181 40 1 32 aaronha01 1965  
## 13 168 23 1 44 aaronha01 1966  
## 14 184 37 3 39 aaronha01 1967  
## 15 174 33 4 29 aaronha01 1968  
## 16 164 30 3 44 aaronha01 1969  
## 17 154 26 1 38 aaronha01 1970  
## 18 162 22 3 47 aaronha01 1971  
## 19 119 10 0 34 aaronha01 1972  
## 20 118 12 1 40 aaronha01 1973  
## 21 91 16 0 20 aaronha01 1974  
## 22 109 16 2 12 aaronha01 1975  
## 23 62 8 0 10 aaronha01 1976

6b.) The desired columns were extracted

q6matrix <- q6out[,c(1:4)]

6c.) An sapply was used to find all of the maxes for every column.

sapply(q6matrix, max)

## H 2B 3B HR   
## 223 46 14 47

7a

data = read.table("ftp://ftp.cmdl.noaa.gov/data/meteorology/in-situ/sum/met\_sum\_insitu\_1\_obop\_hour\_2016.txt",header=FALSE)  
head(data)

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14  
## 1 SUM 2016 1 1 0 162 9.4 100 647.04 -33.4 -32.7 -32.6 50 -99  
## 2 SUM 2016 1 1 1 157 9.3 100 648.08 -36.0 -35.2 -35.0 48 -99  
## 3 SUM 2016 1 1 2 152 8.9 100 648.92 -38.1 -37.1 -36.9 46 -99  
## 4 SUM 2016 1 1 3 138 8.1 100 649.30 -40.0 -38.3 -37.9 44 -99  
## 5 SUM 2016 1 1 4 135 8.8 100 649.72 -40.3 -38.7 -38.0 41 -99  
## 6 SUM 2016 1 1 5 129 8.8 100 650.10 -40.3 -38.5 -37.6 42 -99

7b

data[data=="-99" | data == "99" | data$V7<0] = NA  
head(data)

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14  
## 1 SUM 2016 1 1 0 162 9.4 100 647.04 -33.4 -32.7 -32.6 50 NA  
## 2 SUM 2016 1 1 1 157 9.3 100 648.08 -36.0 -35.2 -35.0 48 NA  
## 3 SUM 2016 1 1 2 152 8.9 100 648.92 -38.1 -37.1 -36.9 46 NA  
## 4 SUM 2016 1 1 3 138 8.1 100 649.30 -40.0 -38.3 -37.9 44 NA  
## 5 SUM 2016 1 1 4 135 8.8 100 649.72 -40.3 -38.7 -38.0 41 NA  
## 6 SUM 2016 1 1 5 129 8.8 100 650.10 -40.3 -38.5 -37.6 42 NA

8a

sapply(data[6:14],summary)

## V6 V7 V8 V9 V10 V11  
## Min. -999.0000 0.000000 -9.0000 -999.9000 -999.90000 -999.90000  
## 1st Qu. 118.0000 3.800000 100.0000 660.6900 -36.30000 -32.20000  
## Median 173.0000 5.400000 100.0000 669.9800 -25.30000 -23.70000  
## Mean 168.0523 5.814446 98.9886 642.6798 -54.15602 -24.36172  
## 3rd Qu. 230.0000 7.200000 100.0000 677.7300 -16.80000 -16.10000  
## Max. 360.0000 25.200000 100.0000 692.5100 -2.10000 -2.40000  
## NA's 185.0000 166.000000 1197.0000 166.0000 166.00000 166.00000  
## V12 V13 V14  
## Min. -54.80000 15.00000 NA  
## 1st Qu. -31.60000 47.00000 NA  
## Median -23.30000 58.00000 NA  
## Mean -23.72994 56.73001 NaN  
## 3rd Qu. -15.80000 66.00000 NA  
## Max. -2.10000 88.00000 NA  
## NA's 166.00000 465.00000 8743

8b

sapply(data[6:14],max)

## V6 V7 V8 V9 V10 V11 V12 V13 V14   
## NA NA NA NA NA NA NA NA NA

8c

tapply(data$V7, data$V3, FUN=mean)

## 1 2 3 4 5 6 7 8   
## 6.204722 5.296870 6.054817 7.215385 5.917070 5.668472 3.734641 4.546640   
## 9 10 11 12   
## 4.555114 7.626467 6.460159 6.575556

9a

data2 = c(seq(as.Date("1958-08-09"),as.Date("2017-11-11"),by=1))  
head(data2)

## [1] "1958-08-09" "1958-08-10" "1958-08-11" "1958-08-12" "1958-08-13"  
## [6] "1958-08-14"

9b

entry= 'wget billboard.com/charts/hot-100/'  
data3 = paste(entry,data2,sep="")  
head(data3)

## [1] "wget billboard.com/charts/hot-100/1958-08-09"  
## [2] "wget billboard.com/charts/hot-100/1958-08-10"  
## [3] "wget billboard.com/charts/hot-100/1958-08-11"  
## [4] "wget billboard.com/charts/hot-100/1958-08-12"  
## [5] "wget billboard.com/charts/hot-100/1958-08-13"  
## [6] "wget billboard.com/charts/hot-100/1958-08-14"

10

sapply(head(data3), system)

setwd(“~/290/gp07g3/scratch”)

## "wget billboard.com/charts/hot-100/1958-08-09"

billboard.com/charts/hot-100/2017-12-11964-01-06  
## "wget billboard.com/charts/hot-100/1958-08-10"

billboard.com/charts/hot-100/2017-12-11964-01-06  
## "wget billboard.com/charts/hot-100/1958-08-11"

billboard.com/charts/hot-100/2017-12-11964-01-06  
## "wget billboard.com/charts/hot-100/1958-08-12"

billboard.com/charts/hot-100/2017-12-11964-01-06  
##"wget billboard.com/charts/hot-100/1958-08-13"

billboard.com/charts/hot-100/2017-12-11964-01-06  
## "wget billboard.com/charts/hot-100/1958-08-14"

billboard.com/charts/hot-100/2017-12-11964-01-06