CSE3020 - Data Visualisation

Lab Assignment 5

HeatMap and nycflights13

SANJIT KUMAR
18BCE0715
DR NALINI N
LAB - L13 + L14

Question/Task

INSTALL.PACKAGES("NYCFLIGHTS13")—FLIGHTS DATASET

1. REMOVE NA VALUES

2. COMPUTE FLIGHT DELAY COST FOR EVERY FLIGHT. AND DELAY COST INTO DATASET
HINT: COST INDEX=[(NUMBER OF FLIGHTS)*MEAN(DELAY)/MEAN(DISTANCE)]

3. SELECT TOP 50 LARGEST ARRIVAL DELAYS

4. CONVERT DELAY COST DATAFRAME TO A MATRIX
HINT: DELAY_MAT<- DELAY_DF.MATRIX(TOP50)

5. VISUALIZE HEAT MAP
HINT: C("FLIGHTS","DISTANCE","DELAY","COST INDEX")

Answers

Preparing the Data

```
library("nycflights13")
library("dplyr")
# Sanjit Kumar - 18BCE0715
flights
df = flights
```

Go to file/function | Addins

frq_table | 18BCE0715_MidTerm.R | 18BCE0715_LabAssignment5.R | df4 | df2 | df | heatmap-up.R

Filter

ep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin	dest	air_time	distance	hour	minute	time_hour
900	-4	1226	1220	6	AA	1	N324AA	JFK	LAX	358	2475	9	0	2013-01-01 09:00:00
1123	30	1454	1425	29	B6	1	N552JB	JFK	FLL	167	1069	11	23	2013-01-01 11:00:00
900	-5	1225	1220	5	AA	1	N336AA	JFK	LAX	336	2475	9	0	2013-01-02 09:00:00
1123	11	1449	1425	24	B6	1	N506JB	JFK	FLL	175	1069	11	23	2013-01-02 11:00:00
900	-5	1144	1220	-36	AA	1	N327AA	JFK	LAX	323	2475	9	0	2013-01-03 09:00:00
1123	38	1510	1425	45	B6	1	N531JB	JFK	FLL	171	1069	11	23	2013-01-03 11:00:00
900	-2	1210	1220	-10	AA	1	N328AA	JFK	LAX	327	2475	9	0	2013-01-04 09:00:00
1123	7	1427	1425	2	B6	1	N659JB	JFK	FLL	156	1069	11	23	2013-01-04 11:00:00
2030	0	2313	2338	-25	UA	1	N24729	EWR	PBI	142	1023	20	30	2013-01-04 20:00:00
900	-9	1206	1220	-14	AA	1	N328AA	JFK	LAX	343	2475	9	0	2013-01-05 09:00:00
1123	10	1436	1425	11	B6	1	N653JB	JFK	FLL	168	1069	11	23	2013-01-05 11:00:00
2029	-2	2317	2330	-13	UA	1	N79279	EWR	PBI	148	1023	20	29	2013-01-05 20:00:00
900	5	1234	1220	14	AA	1	N327AA	JFK	LAX	344	2475	9	0	2013-01-06 09:00:00
1123	10	1449	1425	24	B6	1	N590JB	JFK	FLL	169	1069	11	23	2013-01-06 11:00:00
900	4	1203	1220	-17	AA	1	N335AA	JFK	LAX	316	2475	9	0	2013-01-07 09:00:00
900	-7	1209	1220	-11	AA	1	N323AA	JFK	LAX	341	2475	9	0	2013-01-08 09:00:00
900	0	1206	1220	-14	AA	1	N329AA	JFK	LAX	345	2475	9	0	2013-01-09 09:00:00
900	-4	1159	1220	-21	AA	1	N336AA	JFK	LAX	328	2475	9	0	2013-01-10 09:00:00
900	3	1231	1220	11	AA	1	N329AA	JFK	LAX	350	2475	9	0	2013-01-11 09:00:00
900	-6	1219	1220	-1	AA	1	N319AA	JFK	LAX	344	2475	9	0	2013-01-12 09:00:00
900	-9	1225	1220	5	AA	1	N323AA	JFK	LAX	344	2475	9	0	2013-01-13 09:00:00
900	-2	1227	1220	7	AA	1	N339AA	JFK	LAX	350	2475	9	0	2013-01-14 09:00:00
900	-2	1221	1220	1	AA	1	N329AA	JFK	LAX	351	2475	9	0	2013-01-15 09:00:00
900	23	1245	1220	25	AA	1	N323AA	JFK	LAX	346	2475	9	0	2013-01-16 09:00:00
900	-5	1217	1220	-3	AA	1	N332AA	JFK	LAX	346	2475	9	0	2013-01-17 09:00:00
900	-2	1213	1220	-7	AA	1	N319AA	JFK	LAX	323	2475	9	0	2013-01-18 09:00:00
900	-1	1215	1220	-5	AA	1	N335AA	JFK	LAX	342	2475	9	0	2013-01-19 09:00:00
900	-5	1202	1220	-18	AA	1	N329AA	JFK	LAX	344	2475	9	0	2013-01-20 09:00:00
900	3	1226	1220	6	AA	1	N329AA	JFK	LAX	362	2475	9	0	2013-01-21 09:00:00
900	4	1215	1220	-5	AA	1	N320AA	JFK	LAX	330	2475	9	0	2013-01-22 09:00:00
900	-2	1201	1220	-19	AA	1	N320AA	JFK	LAX	336	2475	9	0	2013-01-23 09:00:00

1) Remove NA values

```
# Removing NA
df2 = na.omit(df)
df2
```

```
# A tibble: 327,346 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>
 1  2013     1     1      517            515         2      830            819        11 UA
 2  2013     1     1      533            529         4      850            830        20 UA
 3  2013     1     1      542            540         2      923            850        33 AA
 4  2013     1     1      544            545        -1     1004           1022       -18 B6
 5  2013     1     1      554            600        -6      812            837       -25 DL
 6  2013     1     1      554            558        -4      740            728        12 UA
 7  2013     1     1      555            600        -5      913            854        19 B6
 8  2013     1     1      557            600        -3      709            723       -14 EV
 9  2013     1     1      557            600        -3      838            846        -8 B6
10  2013     1     1      558            600        -2      753            745         8 AA
# … with 327,336 more rows, and 9 more variables: flight <int>, tailnum <chr>, origin <chr>,
#   dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
>
```

2) Compute flight delay cost for every flight. And delay cost into dataset
Hint: Cost Index=[(number of flights)*mean(delay)/mean(distance)]

```
# Finding the Cost of Delay for each flight and adding it back to
the original data

df3 = df2 %>%
  group_by(flight) %>%
  summarize(mean_delay = mean(dep_delay), mean_distance =
mean(distance), freq = n())
df3$cost = (df3$freq*df3$mean_delay)/df3$mean_distance
df3
df4 = merge (df2,df3)
df4
```

```
> # Finding the Cost of Delay for each flight and adding it back to the original data
> df3 = df2 %>%
+   group_by(flight) %>%
+   summarize(mean_delay = mean(dep_delay), mean_distance = mean(distance), freq = n())
> df3$cost = (df3$freq*df3$mean_delay)/df3$mean_distance
> df3
# A tibble: 3,835 x 5
   flight mean_delay mean_distance  freq    cost
    <int>      <dbl>         <dbl> <int>   <dbl>
 1      1       5.30          1805.   697  2.05
 2      2      -0.569          309.    51 -0.0938
 3      3       3.68          2036.   628  1.13
 4      4       7.52          1024.   391  2.87
 5      5       4.43          1597.   324  0.898
 6      6       7.42          1596.   206  0.957
 7      7      15.7           2399.   236  1.54
 8      8       6.94           292.   234  5.56
 9      9      16.6            944    152  2.67
10     10      24.3            607.    61  2.44
# … with 3,825 more rows
>
```

dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	tailnum	origin	dest	air_time	distance	hour	minute	time_hour	mean_delay	mean_distance	freq	cost
1024	1029	-5	1312	1320	-8	B6	N636JB	JFK	FLL	145	1069	10	29	2013-07-03 10:00:00	5.295552	1804.544	697	2.045392
855	900	-5	1204	1220	-16	AA	N327AA	JFK	LAX	338	2475	9	0	2013-03-03 09:00:00	5.295552	1804.544	697	2.045392
1051	1029	22	1328	1320	8	B6	N634JB	JFK	FLL	141	1069	10	29	2013-07-07 10:00:00	5.295552	1804.544	697	2.045392
1058	1029	29	1350	1320	30	B6	N625JB	JFK	FLL	143	1069	10	29	2013-08-12 10:00:00	5.295552	1804.544	697	2.045392
1032	1029	3	1335	1320	15	B6	N662JB	JFK	FLL	167	1069	10	29	2013-07-05 10:00:00	5.295552	1804.544	697	2.045392
1041	1021	20	1331	1325	6	B6	N559JB	JFK	FLL	155	1069	10	21	2013-12-20 10:00:00	5.295552	1804.544	697	2.045392
1050	1055	-5	1332	1353	-21	B6	N608JB	JFK	FLL	145	1069	10	55	2013-09-11 10:00:00	5.295552	1804.544	697	2.045392
1054	1055	-1	1326	1353	-27	B6	N760JB	JFK	FLL	137	1069	10	55	2013-09-08 10:00:00	5.295552	1804.544	697	2.045392
913	915	-2	1200	1210	-10	AA	N338AA	JFK	LAX	320	2475	9	15	2013-08-04 09:00:00	5.295552	1804.544	697	2.045392
900	900	0	1235	1220	15	AA	N336AA	JFK	LAX	362	2475	9	0	2013-02-10 09:00:00	5.295552	1804.544	697	2.045392
853	900	-7	1218	1225	-7	AA	N339AA	JFK	LAX	359	2475	9	0	2013-12-03 09:00:00	5.295552	1804.544	697	2.045392
902	910	-8	1211	1220	-9	AA	N329AA	JFK	LAX	332	2475	9	10	2013-09-25 09:00:00	5.295552	1804.544	697	2.045392
1156	1120	36	1451	1422	29	B6	N562JB	JFK	FLL	159	1069	11	20	2013-03-03 11:00:00	5.295552	1804.544	697	2.045392
900	900	0	1146	1225	-39	AA	N328AA	JFK	LAX	327	2475	9	0	2013-06-01 09:00:00	5.295552	1804.544	697	2.045392
906	910	-4	1152	1220	-28	AA	N327AA	JFK	LAX	326	2475	9	10	2013-10-26 09:00:00	5.295552	1804.544	697	2.045392
1026	1029	-3	1331	1320	11	B6	N715JB	JFK	FLL	141	1069	10	29	2013-07-20 10:00:00	5.295552	1804.544	697	2.045392
1026	1029	-3	1316	1320	-4	B6	N564JB	JFK	FLL	153	1069	10	29	2013-07-10 10:00:00	5.295552	1804.544	697	2.045392
1448	1455	-7	1915	1951	-36	DL	N37700	JFK	SJU	186	1598	14	55	2013-11-11 14:00:00	5.295552	1804.544	697	2.045392
1103	1055	8	1426	1353	33	B6	N793JB	JFK	FLL	150	1069	10	55	2013-09-23 10:00:00	5.295552	1804.544	697	2.045392
856	900	-4	1143	1225	-42	AA	N338AA	JFK	LAX	323	2475	9	0	2013-04-04 09:00:00	5.295552	1804.544	697	2.045392
1119	1120	-1	1414	1422	-8	B6	N624JB	JFK	FLL	160	1069	11	20	2013-02-19 11:00:00	5.295552	1804.544	697	2.045392
1143	1125	18	1451	1421	30	B6	N594JB	JFK	FLL	158	1069	11	25	2013-04-17 11:00:00	5.295552	1804.544	697	2.045392
902	900	2	1212	1225	-13	AA	N324AA	JFK	LAX	346	2475	9	0	2013-12-18 09:00:00	5.295552	1804.544	697	2.045392
905	910	-5	1205	1220	-15	AA	N339AA	JFK	LAX	328	2475	9	10	2013-10-14 09:00:00	5.295552	1804.544	697	2.045392
1144	1125	19	1458	1421	37	B6	N646JB	JFK	FLL	171	1069	11	25	2013-04-01 11:00:00	5.295552	1804.544	697	2.045392

3) Select top 50 largest arrival delays

```
# Top 50 Largest Arrival Delay

top_50_rows_based_on_arrdelay = top_n(df4, 50, arr_delay)
top_50_arrdelay = top_50_rows_based_on_arrdelay$arr_delay
top_50_arrdelay
```

```
> top_50_rows_based_on_arrdelay = top_n(df4, 50, arr_delay)
> top_50_arrdelay = top_50_rows_based_on_arrdelay$arr_delay
> top_50_arrdelay
 [1]  676 1272  561  846  878 1007  632  852  783  612  551  572  688  769  516  773  598  834  681
[20]  847  821  744  616  614  780  571  645  783  850  802  796  895  915  674  767  595  560  784
[39]  931  856  648  989 1127 1109  875  744  851  506  538  577
>
```

4) convert delay cost dataframe to a matrix
Hint: delay_mat<- delay_df.matrix(top50)

```
# Converting df4 into matrix
keep_attributes = c("flight","cost")
delay_mat<-
data.matrix(top_50_rows_based_on_arrdelay[keep_attributes])
delay_mat
```

```
> # Converting df4 into matrix
> keep_attributes = c("flight","cost")
> delay_mat<- data.matrix(top_50_rows_based_on_arrdelay[keep_attributes])
> delay_mat
      flight       cost
 [1,]     23  2.4046285
 [2,]     51  0.4082701
 [3,]    141  4.0314130
 [4,]    172  4.0365239
 [5,]    172  4.0365239
 [6,]    177  3.5896975
 [7,]    187  1.2007797
 [8,]    257  4.7335595
 [9,]    257  4.7335595
[10,]    269  2.3586182
[11,]    349  4.1808312
[12,]    350  3.2471588
[13,]    502  1.0061292
[14,]    503  1.9641344
[15,]    515  3.6770530
[16,]    575  4.3636629
[17,]    731  2.0815939
[18,]    835  0.5773778
[19,]   1091  3.0524335
[20,]   1223  1.0318922
[21,]   1435  8.8165892
[22,]   1485  7.9582814
[23,]   1485  7.9582814
[24,]   1697  1.3829458
[25,]   1715  5.6215320
```

```
[25,]   1715  5.6215320
[26,]   1819  6.5421569
[27,]   1895  0.6699255
[28,]   1901  2.8709705
[29,]   2007  0.9678077
[30,]   2019  8.0428076
[31,]   2042 13.8297209
[32,]   2047  0.7375328
[33,]   2119  1.9421569
[34,]   2131 10.7968127
[35,]   2319  2.5450980
[36,]   2343  1.4919571
[37,]   2343  1.4919571
[38,]   2363  1.9147475
[39,]   2391  8.7401503
[40,]   2391  8.7401503
[41,]   2437  1.4197080
[42,]   3075  7.6926995
[43,]   3535 15.1325052
[44,]   3695  6.3699583
[45,]   3744  4.0065651
[46,]   3798  7.3900185
[47,]   3944  8.0757630
[48,]   4326 12.7022386
[49,]   4711  3.3102029
[50,]   5716  4.8114035
>
```

5) Visualize Heat Map
Hint: c("Flights","Distance","Delay","Cost Index")

```
#Visualising Heat Map
keep_attributes_2 = c("flight","cost","distance","mean_delay")
heat_map_mat <-
data.matrix(top_50_rows_based_on_arrdelay[keep_attributes_2])
nba_heatmap <- heatmap(heat_map_mat, Rowv=NA, Colv=NA,
                       col = terrain.colors(256), scale="column",
margins=c(5,10),
                       xlab = "Flight Delay Cost",
                       ylab = "Flights",
                       cexCol = 1,
                       main = "Flight Heat Map (cost ,delay,
distance)")
```

Flight Heat Map (cost ,delay, distance)