Homework3Q3

David Li

November 13, 2017

Stats 506: Homework 3 Question 3
David Li
Data Used For This Question:
NYCflights14 Data: https://raw.githubusercontent.com/wiki/arunsrinivasan/flights/NYCflights14/flights14.csv
Scraping URL: https://www.world-airport-codes.com/distance/
AirportCodeDists: from Course Page

```
library("data.table", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
library("ggplot2", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
library("knitr", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
library("rmarkdown", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
library("curl", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
library("rvest", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
library("tidyverse", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
```

```
# Functions Section

# Extract the miles text from a string
get_miles = function(txt){
  y = str_split(txt,'\\(')[[1]]
  z = str_split(y[2],' ')[[1]][1]
  as.numeric(z)
}
# Distance between generally two cities, a1 and a2
scrape_dist = function(a1, a2){
  url = sprintf('https://www.world-airport-codes.com/distance/?a1=%s&a2=%s',
              a1, a2)
  srch = read_html(url) #accesses the url for searching
  txt =
    srch %>%
    html_node("strong") %>%
    html_text()
  get_miles(txt) # Extract Miles
}

# Utilizes scrape_dist function for one fixed point to multiple targets, creates a tibble of results
get_dists = function(fixed, targets){
  dists = sapply(targets, function(target) scrape_dist(fixed, target))
  tibble(from=fixed, to=targets, dist=dists)
}

# Looping through the combinations within a vector and calculating their distances
inner_loop =  function(i){
  get_dists(OrigDestVec[i], OrigDestVec[{i+1}:length(OrigDestVec)])
}
# End Functions Section
```

```r
# Part A
# Import the dataset
nyc14 = fread("https://raw.githubusercontent.com/wiki/arunsrinivasan/flights/NYCflights14/flights14.csv

# Find the unique origin and destination airports, and append these together
orig_codes = unique(nyc14$origin)
dest_codes = unique(nyc14$dest)
OrigDestVec = c(orig_codes, dest_codes)

# Initializing a empty matrix to hold all the distances
orig_origdest = matrix(, ncol = 3)
colnames(orig_origdest) = c("from", "to", "dist")
for(i in 1:3){ # The first three values are the origin airports, which we want as starting targets for
  append = inner_loop(i)
  if(i == 1){ # Take the whole first iteration as starting data.frame
    orig_origdest = append
  }
  else{ # Append as we loop through the combinations
    orig_origdest = rbind(orig_origdest,append)
  }
}
head(orig_origdest)
```

```
## # A tibble: 6 x 3
##    from    to    dist
##   <chr> <chr>   <dbl>
## 1   JFK   LGA   10.69
## 2   JFK   EWR   20.75
## 3   JFK   LAX 2469.33
## 4   JFK   PBI 1029.65
## 5   JFK   MIA 1091.77
## 6   JFK   SEA 2414.93
```

```r
# Loading dataset of scraped dist between destinations, requires to be in working directory
load("AirportCodeDists.RData")

# Part B
AllDist = rbind(orig_origdest, df_dist)
# Append all of our data together, combinations between origins and destinations
AllDist_trans <- data.table(from=AllDist$to, to=AllDist$from, dist=AllDist$dist)
# Reversal of columns to count reverse routes
NewAllDist = rbind(AllDist, AllDist_trans) # All distances possible
reshaped_newalldist = dcast(NewAllDist, from ~ to) # Reshape to wide
```

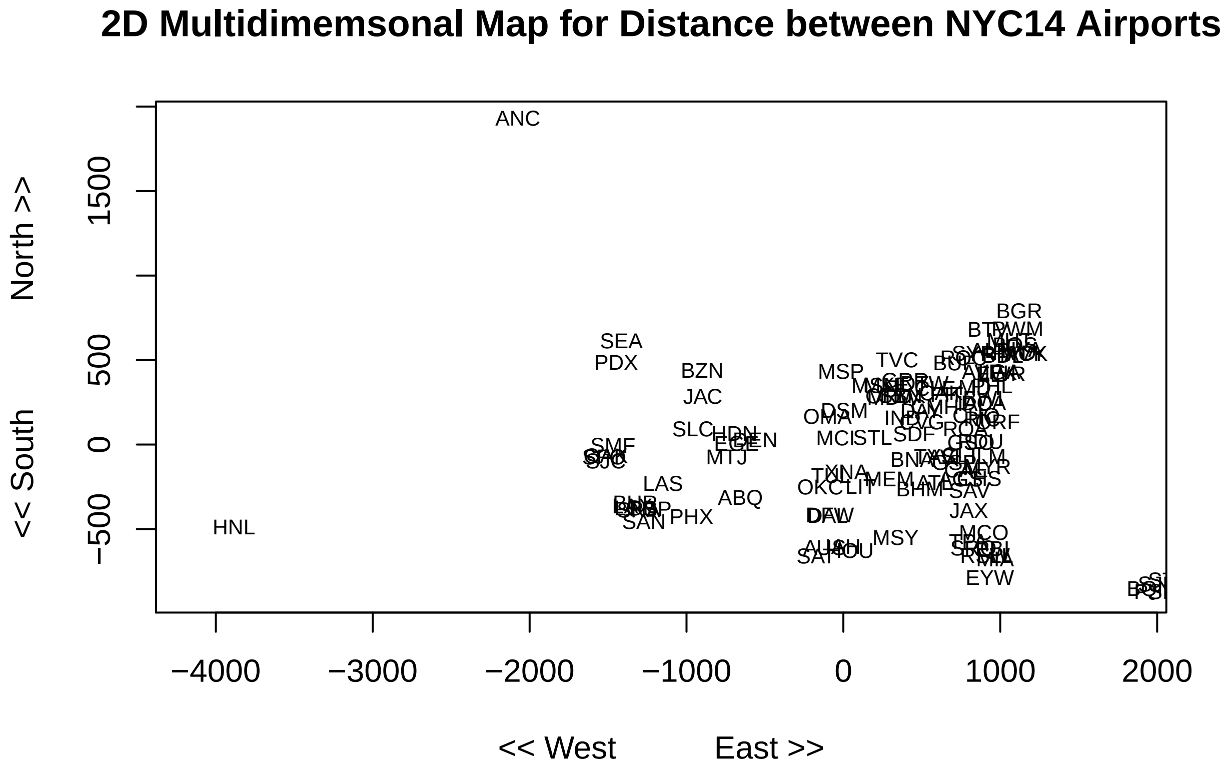
```
## Using 'dist' as value column. Use 'value.var' to override
```

```r
# The first column is the rownames, so we have to coerce that to be our rownames
reshaped_newalldist2 <- data.frame(reshaped_newalldist[,-1])
rownameslist = reshaped_newalldist[,1]
row.names(reshaped_newalldist2) = rownameslist$from
for(i in 1:112){ # Make the diagonal of NAs into 0, since MDS requires this
  reshaped_newalldist2[i,i] = 0
}
head(reshaped_newalldist2)
```

```
##         ABQ     ACK     AGS     ALB     ANC     ATL     AUS     AVL
## ABQ    0.00 2016.14 1409.32 1829.73 2613.31 1266.50  618.20 1355.02
## ACK 2016.14    0.00  849.79  218.09 3463.59  945.68 1716.66  785.57
## AGS 1409.32  849.79    0.00  784.63 3507.27  143.12  947.08  146.50
## ALB 1829.73  218.09  784.63    0.00 3262.56  852.57 1575.78  688.34
## ANC 2613.31 3463.59 3507.27 3262.56    0.00 3410.19 3174.55 3372.19
## ATL 1266.50  945.68  143.12  852.57 3410.19    0.00  811.47  164.26
##         AVP     BDL     BGR     BHM     BNA     BOS     BQN     BTV
## ABQ 1722.72 1882.17 2090.82 1135.94 1120.44 1969.60 2667.35 1876.98
## ACK  293.97  143.56  253.32 1056.37  961.73   90.90 1581.88  271.97
## AGS  648.09  778.90 1055.89  276.31  328.30  860.80 1377.22  900.23
## ALB  138.59   79.93  285.92  945.53  824.10  144.77 1720.56  123.42
## ANC 3275.33 3342.01 3312.08 3337.06 3196.19 3373.23 4867.12 3192.36
## ATL  713.98  858.89 1133.34  133.91  214.01  945.40 1494.58  960.52
##         BUF     BUR     BWI     BZN     CAE     CAK     CHO     CHS
## ABQ 1584.51  670.18 1666.46  779.06 1449.58 1423.76 1570.42 1525.01
## ACK  459.37 2649.28  376.91 2059.33  787.46  592.76  495.05  796.30
## AGS  684.00 2079.46  497.69 1760.90   62.50  522.16  383.74  115.98
## ALB  250.08 2451.38  288.58 1842.92  725.75  413.02  401.41  760.38
## ANC 3091.68 2328.34 3361.23 1876.36 3503.37 3109.58 3358.89 3597.98
## ATL  712.72 1936.68  576.31 1638.01  191.12  528.95  456.53  258.52
##         CLE     CLT     CMH     CVG     DAL     DAY     DCA     DEN
## ABQ 1408.35 1446.29 1339.60 1237.96  579.13 1269.04 1646.56  349.61
## ACK  611.26  722.21  677.80  785.30 1579.41  748.00  404.42 1807.38
## AGS  555.64  140.11  460.80  420.24  861.59  468.28  467.89 1332.31
## ALB  422.73  646.04  507.82  621.74 1425.60  574.69  317.91 1605.55
## ANC 3069.68 3437.82 3110.67 3102.81 3046.75 3071.66 3366.67 2400.27
## ATL  555.36  226.52  447.62  374.16  719.53  433.05  546.84 1197.09
##         DFW     DSM     DTW     EGE     EWR     EYW     FLL     GRR
## ABQ  567.79  831.78 1344.12  318.43 1801.12 1650.23 1685.08 1251.00
## ACK 1587.17 1219.68  687.87 1926.52  217.77 1335.47 1196.65  799.84
## AGS  871.95  852.89  615.59 1446.46  663.36  609.07  515.73  684.79
## ALB 1432.45 1018.58  487.85 1723.78  143.29 1336.05 1206.71  593.63
## ANC 3037.90 2670.27 2977.51 2341.33 3360.90 4022.28 3986.81 2870.51
## ATL  729.79  742.98  595.37 1309.86  745.18  647.65  581.85  641.41
##         GSO     GSP     HDN     HNL     HOU     HYA     IAD     IAH
## ABQ 1496.01 1377.47  377.38 3228.30  758.42 2005.09 1624.10  742.93
## ACK  640.20  792.67 1927.22 5152.89 1621.51   30.93  421.96 1610.73
## AGS  220.82  106.42 1472.62 4637.57  824.67  858.03  459.84  819.86
## ALB  563.87  706.07 1721.85 4941.82 1493.72  194.99  324.79 1480.56
## ANC 3422.12 3413.11 2285.33 2780.03 3282.90 3433.96 3347.83 3261.10
## ATL  306.08  153.22 1337.81 4494.67  694.77  950.27  533.75  688.17
##         ILM     IND     JAC     JAX     JFK     LAS     LAX     LGA
## ABQ 1626.69 1158.27  631.48 1477.35 1821.43  485.34  675.75 1816.81
## ACK  644.57  858.37 2059.97  986.21  198.62 2431.07 2659.46  201.43
## AGS  241.29  499.92 1697.90  199.35  674.83 1884.70 2085.04  678.24
## ALB  626.07  681.27 1846.46  951.83  145.68 2231.56 2462.13  136.23
## ANC 3591.81 3012.54 2004.01 3680.00 3376.46 2301.13 2342.96 3365.98
## ATL  376.46  432.62 1569.27  269.94  759.34 1742.76 1942.17  761.06
##         LGB     LIT     MCI     MCO     MDT     MDW     MEM     MHT
## ABQ  662.82  814.83  717.14 1550.28 1663.38 1118.99  939.74 1950.24
## ACK 2649.93 1283.83 1302.14 1089.60  358.44  914.34 1160.75  135.78
## AGS 2071.95  594.51  818.14  343.54  552.09  661.82  472.01  872.48
## ALB 2453.16 1133.38 1112.72 1074.32  233.77  715.54 1017.15  120.53
## ANC 2357.30 3090.41 2755.40 3811.45 3302.87 2853.74 3147.01 3328.97
## ATL 1929.04  451.81  691.79  404.33  619.56  591.49  331.05  951.32
##         MIA     MKE     MSN     MSP     MSY     MTJ     MVY     MYR
## ABQ 1686.67 1140.07 1078.94  979.88 1012.88  250.03 1987.48 1576.50
## ACK 1217.75  920.04  993.04 1191.28 1369.95 1998.93   30.33  713.55
## AGS  533.02  735.51  783.87  996.26  540.30 1487.86  832.67  176.17
## ALB 1227.62  713.26  785.07  976.91 1266.30 1798.92  188.37  685.22
## ANC 3998.54 2788.12 2730.14 2511.46 3425.41 2374.55 3438.18 3590.92
## ATL  595.91  670.03  708.02  906.98  424.73 1348.86  925.64  316.26
##         OAK     OKC     OMA     ORD     ORF     PBI     PDX     PHL
## ABQ  886.80  508.88  724.30 1115.88 1697.99 1668.85 1110.12 1743.13
## ACK 2750.13 1539.49 1336.36  921.41  446.07 1157.36 2605.97  288.13
## AGS 2265.45  901.48  939.31  677.27  406.49  475.29 2296.17  583.39
## ALB 2542.09 1368.13 1134.98  721.07  423.95 1165.17 2388.69  212.30
## ANC 2013.17 2877.37 2608.51 2838.52 3501.91 3953.12 1538.90 3370.28
## ATL 2125.22  759.06  820.82  606.63  515.53  545.54 2167.96  665.82
##         PHX     PIT     PSE     PSP     PVD     PWM     RDU     RIC
## ABQ  327.89 1482.85 2716.91  569.94 1946.42 2010.60 1561.47 1633.70
## ACK 2342.90  533.75 1619.29 2564.38   77.80  165.83  599.64  465.95
## AGS 1726.96  501.21 1426.72 1978.61  815.72  947.38  250.26  387.17
## ALB 2154.45  366.81 1761.20 2369.32  140.56  186.50  544.14  407.04
## ANC 2548.39 3172.35 4914.94 2400.02 3393.88 3326.90 3472.58 3431.20
## ATL 1583.85  526.88 1544.45 1835.64  903.07 1026.30  355.56  480.41
##         ROA     ROC     RSW     SAN     SAT     SAV     SBN     SDF
## ABQ 1488.64 1639.38 1582.38  627.09  608.48 1472.09 1188.04 1175.09
## ACK  595.09  410.28 1215.47 2633.14 1781.99  880.89  840.82  858.52
## AGS  295.37  712.42  472.31 2030.55 1007.01   96.59  623.30  393.46
## ALB  496.72  197.44 1206.46 2440.17 1641.86  837.77  643.70  701.54
## ANC 3353.01 3117.01 3906.56 2449.23 3188.08 3602.99 2905.59 3114.78
## ATL  357.27  749.75  515.88 1887.44  872.80  214.24  567.14  321.93
##         SEA     SFO     SJC     SJU     SLC     SMF     SNA     SRQ
## ABQ 1178.88  894.30  868.03 2731.21  493.09  864.58  648.69 1511.51
## ACK 2566.96 2760.58 2744.64 1594.36 2163.66 2693.79 2639.48 1190.92
## AGS 2301.89 2274.44 2251.67 1427.40 1722.24 2226.80 2057.48  414.27
## ALB 2348.99 2552.64 2537.40 1740.44 1955.09 2484.47 2443.36 1168.89
## ANC 1445.02 2015.58 2042.54 4907.91 2120.61 1969.91 2373.45 3830.04
## ATL 2177.79 2134.13 2111.10 1547.73 1586.66 2087.55 1914.52  445.37
##         STL     STT     SYR     TPA     TUL     TVC     TYS     XNA
## ABQ  931.69 2791.58 1717.23 1495.26  607.10 1294.87 1271.50  695.40
## ACK 1086.17 1611.59  335.00 1156.89 1429.46  819.05  840.33 1345.64
## AGS  598.29 1477.22  743.83  374.23  813.08  809.05  204.36  727.73
## ALB  908.40 1763.99  119.24 1131.47 1257.25  603.31  724.87 1176.87
## ANC 2929.25 4948.15 3168.52 3797.06 2887.78 2772.56 3302.97 2934.92
## ATL  484.03 1599.84  793.85  406.95  672.45  769.69  152.23  588.13
```

```r
#Part C
# Doing Multi-dimensional scaling
fit <- cmdscale(reshaped_newalldist2)
colnames(fit) = c("xvalue", "yvalue")

# Setting up variables to allow for plotting a 2D map
x = -fit[,1]
y = fit[,2]
plot(x, y, pch = 19, xlab="<< West           East >>", ylab="<< South            North >>",
    main="2D Multidimemsonal Map for Distance between NYC14 Airports", type="n")
text(x, y, pos = 4, labels = row.names(reshaped_newalldist2), cex=.7)
```


2D Multidimemsonal Map for Distance between NYC14 Airports
North >>
<< South
<< West
East >>
ANC
HNL
SEA
PDX
BZN
JAC
SLC
HDN
EGE
DEN
MTJ
SMF
SJC
LAS
ABQ
SAN
PHX
MSP
TVC
OMA
DSM
MCI
STL
SDF
IND
TYS
BNA
OKC
LIT
MEM
XNA
TUL
BHM
ATL
SAV
JAX
MSY
AUS
HOU
SAT
DFW
DAL
MCO
TPA
RSW
MIA
EYW
BGR
BTV
PWM
ROA
ORF
CHS
CAE
RDU
GSO
CLT
ILM
MYR