Exercise_2_DataAnalysis

Ulin

2024-03-19

目录

1 数据预处理

1.1 包导入

```
library(nycflights13)
```

```
## Warning: 程辑包'nycflights13'是用R版本4.3.3 来建造的
```

```
library(dplyr)
```

```
##
## 载入程辑包: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1.2 数据导入

```
flights <- nycflights13::flights
summary(flights)
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
```

```
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00.00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00.00
##  Median :29.00   Median :2013-07-03 10:00:00.00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54.64
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00.00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00.00
##
```

```
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
```

```
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

查看 flights 数据集的数据，共有 19 个维度的数据项。

2 Problem 1: Origin~Dest

问题描述：统计一下，共有多少个不同的往返地组合，并将这些往返地组合抽取出来，构造成一个名为 Ori.Dest 的新数据集。

2.1 数据处理

```
# 使用管道命名符操作
flights %>%
  select('origin','dest') %>% # 选取变量
  distinct() -> Ori.Dest # 去除重复组合
print(dim(Ori.Dest)[1]) # 输出结果
```

```
## [1] 224
```

```
head(Ori.Dest,20)
```

```
## # A tibble: 20 x 2
##    origin dest
##    <chr>  <chr>
##  1 EWR    IAH
##  2 LGA    IAH
##  3 JFK    MIA
##  4 JFK    BQN
##  5 LGA    ATL
##  6 EWR    ORD
##  7 EWR    FLL
##  8 LGA    IAD
##  9 JFK    MCO
```

```
## 10 LGA    ORD
## 11 JFK    PBI
## 12 JFK    TPA
## 13 JFK    LAX
## 14 EWR    SFO
## 15 LGA    DFW
## 16 JFK    BOS
## 17 EWR    LAS
## 18 LGA    FLL
## 19 EWR    PBI
## 20 LGA    MSP
```

2.2 结果

问题 1，共有 224 种不同的往返地组合

3 Problem 2: Longitude & Latitude

问题描述：请将 flights 数据集进行扩充，增加 4 列，这 4 列分别是 Ori.Lat, Ori.Lon,Dest.Lat,Dest.Lon，分别表示这起飞地的经纬度和到达地的经纬度。

3.1 数据查看

```
airports <- nycflights13::airports # 导入 airports 数据集
summary(airports) # 查看数据
```

```
##      faa                name                lat             lon
##  Length:1458        Length:1458        Min.   :19.72   Min.   :-176.65
##  Class :character   Class :character   1st Qu.:34.26   1st Qu.:-119.19
##  Mode  :character   Mode  :character   Median :40.09   Median : -94.66
##                                        Mean   :41.65   Mean   :-103.39
##                                        3rd Qu.:45.07   3rd Qu.: -82.52
##                                        Max.   :72.27   Max.   : 174.11
##       alt                tz              dst               tzone
##  Min.   : -54.00   Min.   :-10.000   Length:1458        Length:1458
##  1st Qu.:  70.25   1st Qu.: -8.000   Class :character   Class :character
##  Median : 473.00   Median : -6.000   Mode  :character   Mode  :character
```

```
##  Mean   :1001.42   Mean   : -6.519
##  3rd Qu.:1062.50   3rd Qu.: -5.000
##  Max.   :9078.00   Max.   :  8.000
```

```
head(airports) # 查看数据示例
```

```
## # A tibble: 6 x 8
##   faa   name                            lat   lon   alt    tz dst   tzone
##   <chr> <chr>                         <dbl> <dbl> <dbl> <dbl> <chr> <chr>
## 1 04G   Lansdowne Airport              41.1 -80.6  1044    -5 A     America/Ne~
## 2 06A   Moton Field Municipal Airport  32.5 -85.7   264    -6 A     America/Ch~
## 3 06C   Schaumburg Regional            42.0 -88.1   801    -6 A     America/Ch~
## 4 06N   Randall Airport                41.4 -74.4   523    -5 A     America/Ne~
## 5 09J   Jekyll Island Airport          31.1 -81.4    11    -5 A     America/Ne~
## 6 0A9   Elizabethton Municipal Airport 36.4 -82.2  1593    -5 A     America/Ne~
```

根据搜索，FAA 代码即由美国联邦航空管理局指定的位置代码。经检查，'faa'列能够与 flights 数据集的 'origin' 列和 'dest' 列匹配。

3.2 数据处理

```
# 进行连接
airports %>%
  select('faa','lat','lon') %>%
  merge(flights,.,all.x='TRUE',by.x = 'origin', by.y = 'faa') %>%
  rename('Ori.Lat'= 'lat', 'Ori.Lon'='lon') -> flights

airports %>%
  select('faa','lat','lon') %>%
  merge(flights,.,all.x='TRUE',by.x = 'dest', by.y = 'faa') %>%
  rename('Dest.Lat'= 'lat', 'Dest.Lon'='lon') -> flights

head(flights)
```

```
##   dest origin year month day dep_time sched_dep_time dep_delay arr_time
## 1  ABQ    JFK 2013    11   1     1950           2000       -10     2226
## 2  ABQ    JFK 2013     4  22     1712           1630        42     1946
## 3  ABQ    JFK 2013    11  24     2000           2000         0     2252
```

```
## 4  ABQ    JFK 2013    11   6     1952           2000        -8     2309
## 5  ABQ    JFK 2013     4  27     2020           2025        -5     2245
## 6  ABQ    JFK 2013     7  31     2002           2007        -5     2219
##   sched_arr_time arr_delay carrier flight tailnum air_time distance hour minute
## 1           2303       -37      B6     65  N659JB      253     1826   20      0
## 2           1915        31      B6   1505  N821JB      256     1826   16     30
## 3           2303       -11      B6     65  N633JB      263     1826   20      0
## 4           2303         6      B6     65  N661JB      279     1826   20      0
## 5           2304       -19      B6   1505  N633JB      246     1826   20     25
## 6           2259       -40      B6   1505  N763JB      237     1826   20      7
##             time_hour  Ori.Lat   Ori.Lon Dest.Lat  Dest.Lon
## 1 2013-11-01 20:00:00 40.63975 -73.77893 35.04022 -106.6092
## 2 2013-04-22 16:00:00 40.63975 -73.77893 35.04022 -106.6092
## 3 2013-11-24 20:00:00 40.63975 -73.77893 35.04022 -106.6092
## 4 2013-11-06 20:00:00 40.63975 -73.77893 35.04022 -106.6092
## 5 2013-04-27 20:00:00 40.63975 -73.77893 35.04022 -106.6092
## 6 2013-07-31 20:00:00 40.63975 -73.77893 35.04022 -106.6092
```

3.3 结果展示

```
print(names(flights))
```

```
##  [1] "dest"          "origin"        "year"          "month"
##  [5] "day"           "dep_time"      "sched_dep_time" "dep_delay"
##  [9] "arr_time"      "sched_arr_time" "arr_delay"     "carrier"
## [13] "flight"        "tailnum"       "air_time"      "distance"
## [17] "hour"          "minute"        "time_hour"     "Ori.Lat"
## [21] "Ori.Lon"       "Dest.Lat"      "Dest.Lon"
```

已成功添加了出发地和目的地的经纬度

4 Problem 3: Define Function_CalcualteDistance

问题描述：构造一个函数，这个函数可以传递 4 个参数，也就是上题的那四个参数，然后根据这 4 个参数，计算出起飞地点和到达地点之间的距离。函数名为 Calculate_Distance。

4.1 函数定义

```
Calculate_Distance <- function(Ori.lat, Ori.lon, Des.lat, Des.lon){
  x = Des.lon-Ori.lon # 经度差
  y = Des.lat-Ori.lat # 维度差
  distance <- sqrt((x^2) + (y^2)) # 根据目的地和出发地的欧氏距离计算
  return(distance)
}
```

5 Problem 4: Call Function_CalcualteDistance

问题描述：利用上述函数，为 flights 增加一列，此列名为 Cal.Distance，这一新的变量是根据上述函数计算出的起飞地和到达地之间的距离。

5.1 函数调用

```
distance_cal <- Calculate_Distance(flights$Ori.Lat, flights$Ori.Lon, flights$Dest.Lat, flights$Des
flights <- mutate(flights, Cal.Distance=distance_cal)
head(flights)
```

```
##   dest origin year month day dep_time sched_dep_time dep_delay arr_time
## 1  ABQ    JFK 2013    11   1     1950           2000       -10     2226
## 2  ABQ    JFK 2013     4  22     1712           1630        42     1946
## 3  ABQ    JFK 2013    11  24     2000           2000         0     2252
## 4  ABQ    JFK 2013    11   6     1952           2000        -8     2309
## 5  ABQ    JFK 2013     4  27     2020           2025        -5     2245
## 6  ABQ    JFK 2013     7  31     2002           2007        -5     2219
##   sched_arr_time arr_delay carrier flight tailnum air_time distance hour minute
## 1           2303       -37      B6     65  N659JB      253     1826   20      0
## 2           1915        31      B6   1505  N821JB      256     1826   16     30
## 3           2303       -11      B6     65  N633JB      263     1826   20      0
## 4           2303         6      B6     65  N661JB      279     1826   20      0
## 5           2304       -19      B6   1505  N633JB      246     1826   20     25
## 6           2259       -40      B6   1505  N763JB      237     1826   20      7
##             time_hour  Ori.Lat   Ori.Lon Dest.Lat  Dest.Lon Cal.Distance
## 1 2013-11-01 20:00:00 40.63975 -73.77893 35.04022 -106.6092     33.30437
```

```
## 2 2013-04-22 16:00:00 40.63975 -73.77893 35.04022 -106.6092     33.30437
## 3 2013-11-24 20:00:00 40.63975 -73.77893 35.04022 -106.6092     33.30437
## 4 2013-11-06 20:00:00 40.63975 -73.77893 35.04022 -106.6092     33.30437
## 5 2013-04-27 20:00:00 40.63975 -73.77893 35.04022 -106.6092     33.30437
## 6 2013-07-31 20:00:00 40.63975 -73.77893 35.04022 -106.6092     33.30437
```