LAB_DA_2-PART-I

20BDS0354 A SAI CHARAN

2023-09-04

Importing Libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Read 50% of records from file “data_cat1.csv” with the following columns only

```
TotalRows<-nrow(read.csv("C:/Users/saic3/CSE3046-F2-LAB_SLOT_L3+L4/Datasets/data_cat1.csv"))


df<-read.csv("C:/Users/saic3/CSE3046-F2-LAB_SLOT_L3+L4/Datasets/data_cat1.csv",nrows = TotalRows/2)%>%se
"origin","dest")


head(df)
```

```
##   dep_time sched_dep_time arr_time sched_arr_time carrier flight origin dest
## 1      517            515      830            819      UA   1545    EWR  IAH
## 2      533            529      850            830      UA   1714    LGA  IAH
## 3      542            540      923            850      AA   1141    JFK  MIA
## 4      544            545     1004           1022      B6    725    JFK  BQN
## 5      554            600      812            837      DL    461    LGA  ATL
## 6      554            558      740            728      UA   1696    EWR  ORD
```

```
cat("Taken Rows are",nrow(df))
```

```
## Taken Rows are 168388
```

```
cat(" \nTotal Rows are",TotalRows)
```

```
##
```

```
## Total Rows are 336776
```

2. How many variables have missing values?

```
colnames(df)[colSums(is.na(df)) > 0]
```

```
## [1] "dep_time" "arr_time"
```

Remove all missing records and save them in another csv file.

```
df<-drop_na(df)
library(readxl)
write.csv(df,"ModifiedDataCAT-I")
apply(df,2,anyNA)
```

```
##       dep_time sched_dep_time       arr_time sched_arr_time        carrier
##          FALSE          FALSE          FALSE          FALSE          FALSE
##         flight         origin           dest
##          FALSE          FALSE          FALSE
```

3.Create new attributes for departure delay, and arrival delay.

```
class(df$dep_time)
```

```
## [1] "integer"
```

```
class(df$sched_dep_time)
```

```
## [1] "integer"
```

```
class(df$arr_time)
```

```
## [1] "integer"
```

```
class(df$sched_arr_time)
```

```
## [1] "integer"
```

3.Create new attributes for departure delay, and arrival delay.

```
C<-ifelse(df$dep_time<1000,paste0("0",df$dep_time),df$dep_time)
  D<-ifelse(df$sched_dep_time<1000,paste0("0",df$sched_dep_time),df$sched_dep_time)
  h1<-as.integer(substring(C,1,2))-as.integer(substring(D,1,2))
  m1<-as.integer(substring(C,3,4))-as.integer(substring(D,3,4))
  Out<-h1*60+m1
  df<-df%>%mutate(dep_delay=Out)
```

```
 C<-ifelse(df$arr_time<1000,paste0("0",df$arr_time),df$arr_time)
  D<-ifelse(df$sched_arr_time<1000,paste0("0",df$sched_arr_time),df$sched_arr_time)
  h1<-as.integer(substring(C,1,2))-as.integer(substring(D,1,2))
  m1<-as.integer(substring(C,3,4))-as.integer(substring(D,3,4))
  Out<-h1*60+m1
```

```
df<-df%>%mutate(arr_delay=Out)
head(df)
```

```
##   dep_time sched_dep_time arr_time sched_arr_time carrier flight origin dest
## 1      517            515      830            819      UA   1545    EWR  IAH
## 2      533            529      850            830      UA   1714    LGA  IAH
## 3      542            540      923            850      AA   1141    JFK  MIA
## 4      544            545     1004           1022      B6    725    JFK  BQN
## 5      554            600      812            837      DL    461    LGA  ATL
## 6      554            558      740            728      UA   1696    EWR  ORD
##   dep_delay arr_delay
## 1         2        11
## 2         4        20
## 3         2        33
## 4        -1       -18
## 5        -6       -25
## 6        -4        12
```

4.a Had an arrival delay of two or more hours

```
head(df%>%filter(arr_delay>=120),n = 50)
```

```
##    dep_time sched_dep_time arr_time sched_arr_time carrier flight origin dest
## 1       811            630     1047            830      MQ   4576    LGA  CLT
## 2       957            733     1056            853      UA    856    EWR  BOS
## 3      1114            900     1447           1222      UA   1086    LGA  IAH
## 4      1505           1310     1638           1431      EV   4497    EWR  RIC
## 5      1525           1340     1831           1626      B6    525    EWR  MCO
## 6      1528           1459     2002           1647      EV   3806    EWR  STL
## 7      1549           1445     1912           1656      EV   4181    EWR  MCI
## 8      1558           1359     1718           1515      EV   5712    JFK  IAD
## 9      1732           1630     2028           1825      EV   4092    EWR  DAY
## 10     1803           1620     2008           1750      MQ   4622    LGA  BNA
## 11     1815           1325     2120           1542      EV   4417    EWR  OMA
## 12     1842           1422     1958           1535      EV   4633    EWR  BTV
## 13     1856           1645     2212           2005      AA    181    JFK  LAX
## 14     1934           1725     2126           1855      MQ   4255    JFK  BNA
## 15     1938           1703     2109           1823      EV   4300    EWR  RIC
## 16     1942           1705     2124           1830      MQ   4410    JFK  DCA
## 17     2006           1630     2230           1848      EV   4644    EWR  SAV
## 18     2009           1808     2145           1942      EV   4440    EWR  PIT
## 19     2103           2110     2345             17      DL   1668    JFK  LAS
## 20     2115           1700     2330           1920      9E   3347    JFK  CVG
## 21     2119           1930     2358           2136      EV   4543    EWR  DSM
## 22     2209           2145       58             37      B6     35    JFK  PBI
## 23     2221           2000     2331           2124      EV   4462    EWR  BUF
## 24      817            630     1107            845      EV   4235    EWR  IND
## 25      833            558     1018            727      UA    651    EWR  ORD
## 26      905            822     1313           1045      EV   4140    EWR  XNA
## 27      923            624     1051            758      EV   4104    EWR  PIT
## 28     1125            925     1445           1146      9E   3658    LGA  GRR
## 29     1244            900     1431           1104      EV   4412    EWR  MYR
## 30     1332            904     1616           1128      EV   4364    EWR  MCI
## 31     1412            838     1710           1147      UA    468    EWR  MCO
## 32     1451           1232     1749           1533      UA   1121    EWR  FLL
## 33     1548           1340     1710           1500      EV   4617    EWR  PIT
## 34     1607           1030     2003           1355      AA    179    JFK  SFO
## 35     1710           1526     1857           1655      EV   4502    EWR  BNA
## 36     1728           1515     1843           1632      EV   4133    EWR  IAD
## 37     1751           1450     2041           1755      AA   1813    JFK  MCO
## 38     1755           1547     1929           1708      EV   3272    EWR  BUF
## 39     1806           1629     2008           1808      EV   4308    EWR  RDU
## 40     1845           1645     2137           1859      9E   3983    EWR  CVG
## 41     1848           1701     2112           1856      EV   4202    EWR  STL
## 42     1849           1724     2235           1938      EV   4321    EWR  MCI
## 43     1909           1621     2116           1816      EV   4092    EWR  DAY
## 44     2005           1705     2126           1830      MQ   4410    JFK  DCA
## 45     2043           1759     2251           1949      EV   4581    EWR  CMH
## 46     2113           2110     2346             17      DL   1627    JFK  LAS
## 47     2115           1837     2322           2045      EV   4125    EWR  GRR
## 48     2128           2130       43             18      B6    383    LGA  FLL
## 49     2131           1512     2340           1741      UA    488    LGA  DEN
## 50      235           2359      700            437      B6    727    JFK  BQN
##    dep_delay arr_delay
## 1        101       137
## 2        144       123
## 3        134       145
## 4        115       127
## 5        105       125
## 6         29       195
## 7         64       136
## 8        119       123
## 9         62       123
## 10       103       138
## 11       290       338
## 12       260       263
## 13       131       127
## 14       129       151
## 15       155       166
## 16       157       174
## 17       216       222
## 18       121       123
## 19        -7      1358
## 20       255       250
## 21       109       142
## 22        24       121
## 23       141       127
## 24       107       142
## 25       155       171
## 26        43       148
## 27       179       173
## 28       120       179
## 29       224       207
## 30       268       288
## 31       334       323
## 32       139       136
## 33       128       130
## 34       337       368
## 35       104       122
## 36       133       131
## 37       181       166
## 38       128       141
## 39        97       120
## 40       120       158
## 41       107       136
## 42        85       177
## 43       168       180
## 44       180       176
## 45       164       182
## 46         3      1359
## 47       158       157
## 48        -2       175
## 49       379       359
## 50     -1284       143
```

```
cat("There are Totally ",nrow(df%>%filter(arr_delay>=120)))
```

```
## There are Totally  3211
```

```
cat("Rows Actually But Truncated Print to 1000")
```

```
## Rows Actually But Truncated Print to 1000
```

4.b Flew to DEN or JFK

```
head(df%>%filter(dest=="DEN"|dest=="JFK"),n = 50)
```

```
##    dep_time sched_dep_time arr_time sched_arr_time carrier flight origin dest
## 1       646            645      910            916      UA    883    LGA  DEN
## 2       727            730      959            952      UA   1162    EWR  DEN
## 3       752            750     1025           1029      UA    477    LGA  DEN
## 4       754            755     1103           1030      WN    733    LGA  DEN
## 5       813            815     1103           1056      DL    914    LGA  DEN
## 6       833            835     1134           1102      F9    835    LGA  DEN
## 7       906            843     1134           1125      UA   1643    EWR  DEN
## 8       929            925     1220           1150      WN    766    EWR  DEN
## 9      1005           1000     1239           1234      UA   1625    EWR  DEN
## 10     1123           1110     1410           1336      UA    405    LGA  DEN
## 11     1253           1212     1524           1436      UA    754    EWR  DEN
## 12     1255           1255     1540           1535      WN   1251    LGA  DEN
## 13     1506           1512     1723           1741      UA    407    LGA  DEN
## 14     1531           1520     1809           1750      UA    365    EWR  DEN
## 15     1603           1600     1839           1830      WN    591    EWR  DEN
## 16     1610           1555     1852           1834      DL   1939    LGA  DEN
## 17     1628           1630     1907           1923      DL    920    JFK  DEN
## 18     1634           1626     1913           1852      UA     69    EWR  DEN
## 19     1712           1719     1939           1958      UA    509    LGA  DEN
## 20     1716           1730     1947           1953      F9    511    LGA  DEN
## 21     1827           1829     2105           2056      UA   1139    EWR  DEN
## 22     2129           2120     2342           2351      B6     97    JFK  DEN
## 23      647            645      849            916      UA    320    LGA  DEN
## 24      728            730     1001            952      UA    311    EWR  DEN
## 25      740            740     1025           1015      WN   1581    LGA  DEN
## 26      741            740     1005           1019      UA    328    LGA  DEN
## 27      811            815     1100           1056      DL    914    LGA  DEN
## 28      827            835     1120           1102      F9    835    LGA  DEN
## 29      844            843     1114           1125      UA   1643    EWR  DEN
## 30      932            925     1157           1155      WN   1484    EWR  DEN
## 31     1004           1003     1235           1237      UA   1625    EWR  DEN
## 32     1123           1110     1434           1340      UA   1477    LGA  DEN
## 33     1209           1212     1451           1440      UA   1492    EWR  DEN
## 34     1258           1255     1537           1535      WN   1251    LGA  DEN
## 35     1521           1520     1745           1754      UA   1107    EWR  DEN
## 36     1524           1500     1803           1743      DL   1939    LGA  DEN
## 37     1616           1610     1839           1840      WN   2305    EWR  DEN
## 38     1629           1630     1905           1923      DL    920    JFK  DEN
## 39     1644           1625     1915           1851      UA   1556    EWR  DEN
## 40     1712           1700     1927           1939      UA    509    LGA  DEN
## 41     1728           1730     1952           1953      F9    511    LGA  DEN
## 42     1831           1829     2104           2056      UA   1462    EWR  DEN
## 43     1907           1815     2137           2052      DL    884    LGA  DEN
## 44     2116           2120     2342           2351      B6     97    JFK  DEN
## 45     2131           1512     2340           1741      UA    488    LGA  DEN
## 46      644            645      847            918      UA    338    LGA  DEN
## 47      702            658      906            915      UA    510    EWR  DEN
## 48      757            755     1008           1030      WN    733    LGA  DEN
## 49      801            759     1025           1025      UA    429    LGA  DEN
## 50      815            817     1048           1058      DL    914    LGA  DEN
##    dep_delay arr_delay
## 1          1        -6
## 2         -3         7
## 3          2        -4
## 4         -1        33
## 5         -2         7
## 6         -2        32
## 7         23         9
## 8          4        30
## 9          5         5
## 10        13        34
## 11        41        48
## 12         0         5
## 13        -6       -18
## 14        11        19
## 15         3         9
## 16        15        18
## 17        -2       -16
## 18         8        21
## 19        -7       -19
## 20       -14        -6
## 21        -2         9
## 22         9        -9
## 23         2       -27
## 24        -2         9
## 25         0        10
## 26         1       -14
## 27        -4         4
## 28        -8        18
## 29         1       -11
## 30         7         2
## 31         1        -2
## 32        13        54
## 33        -3        11
## 34         3         2
## 35         1        -9
## 36        24        20
## 37         6        -1
## 38        -1       -18
## 39        19        24
## 40        12       -12
## 41        -2        -1
## 42         2         8
## 43        52        45
## 44        -4        -9
## 45       379       359
## 46        -1       -31
## 47         4        -9
## 48         2       -22
## 49         2         0
## 50        -2       -10
```

```
abc<-nrow(df%>%filter(dest=="DEN"|dest=="JFK"))
cat("There are Totally ",abc)
```

```
## There are Totally  3584
```

```
cat(" Rows Actually But Truncated Result")
```

```
##  Rows Actually But Truncated Result
```

4.c Arrived more than two hours late, but didn't leave late

```
  head(df%>%filter(arr_delay>120 & dep_delay<=0),n = 50)
```

```
##    dep_time sched_dep_time arr_time sched_arr_time carrier flight origin dest
## 1      2103           2110     2345             17      DL   1668    JFK  LAS
## 2      2128           2130       43             18      B6    383    LGA  FLL
## 3       235           2359      700            437      B6    727    JFK  BQN
## 4      1025           1032     1521           1240      EV   4255    EWR  CHS
## 5      2129           2130     2326             18      B6    199    JFK  LAS
## 6      2155           2159       52            100      B6     11    JFK  FLL
## 7      2059           2100     2349             31      DL   2363    JFK  LAX
## 8      2156           2159       46            100      B6     11    JFK  FLL
## 9        37           2230      341            131      B6     11    JFK  FLL
## 10     2116           2130     2400             18      B6    383    LGA  FLL
## 11     2058           2100     2350             31      DL   2363    JFK  LAX
## 12     1433           1436     1844           1543      EV   4372    EWR  DCA
## 13     2131           2135     2358             24      B6     43    JFK  MCO
## 14     2121           2130     2354             25      B6    383    LGA  FLL
## 15     1718           1725     2339           2050      UA    512    JFK  SFO
## 16     2100           2100     2356             31      DL   2363    JFK  LAX
## 17     1504           1506     1937           1641      UA   1702    EWR  ORD
## 18      842            845     1630           1350      AA   1357    JFK  SJU
## 19      616            625     1203            937      UA    304    LGA  IAH
## 20     2054           2100     2359             31      DL   2363    JFK  LAX
## 21      644            655     1624           1030      DL   1415    JFK  SLC
## 22      829            830     1633           1146      UA   1601    EWR  FLL
## 23     1100           1100     1815           1420      DL   1275    JFK  SLC
## 24     1724           1725     2257           2050      UA    512    JFK  SFO
## 25     1419           1420     1754           1550      MQ   3728    EWR  ORD
## 26      921            929     1410           1044      EV   4636    EWR  DCA
## 27     1159           1205     1813           1520      AA    743    LGA  DFW
## 28     1234           1240     1855           1540      AA   1853    EWR  DFW
## 29     1242           1245     1907           1550      AA    745    LGA  DFW
## 30     1424           1430     2058           1735      AA    883    EWR  DFW
## 31     1451           1455     1908           1645      AA    337    LGA  ORD
## 32     2056           2100     2348             18      DL    427    JFK  LAX
## 33     2126           2137     2356             27      B6   1371    LGA  FLL
## 34     2105           2113     2358             30      UA    617    EWR  SFO
## 35      800            805     1323           1022      EV   4364    EWR  MCI
## 36     2128           2137     2356             27      B6   1371    LGA  FLL
## 37     1507           1515     1947           1700      AA    341    LGA  ORD
## 38     2119           2137     2353             27      B6   1371    LGA  FLL
## 39     2245           2255     2400             12      B6   2002    JFK  BUF
## 40     2023           2025     2359             26      UA   1071    EWR  BQN
## 41     2237           2255     2346             12      B6   2002    JFK  BUF
## 42     1350           1350     1736           1526      EV   5181    LGA  MSN
## 43     1357           1359     1858           1654      AA   1151    LGA  DFW
## 44     2055           2100     2329             20      DL    427    JFK  LAX
## 45     2105           2113     2349             30      UA    242    EWR  SFO
## 46     2056           2100     2347             20      DL    427    JFK  LAX
## 47     2108           2113     2342             30      UA    617    EWR  SFO
## 48     2248           2255     2359             12      B6   2002    JFK  BUF
## 49     2057           2100     2340             20      DL    427    JFK  LAX
## 50      758            800     1240           1003      9E   3353    JFK  DTW
##    dep_delay arr_delay
## 1         -7      1358
## 2         -2       175
## 3      -1284       143
## 4         -7       161
## 5         -1      1338
## 6         -4       242
## 7         -1      1248
## 8         -3       186
## 9      -1163       130
## 10       -14      1372
## 11        -2      1249
## 12        -3       181
## 13        -4      1314
## 14        -9      1309
## 15        -7       169
## 16         0      1255
## 17        -2       176
## 18        -3       160
## 19        -9       146
## 20        -6      1258
## 21        -11        354
## 22         -1        287
## 23          0        235
## 24         -1        127
## 25         -1        124
## 26         -8        206
## 27         -6        173
## 28         -6        195
## 29         -3        197
## 30         -6        203
## 31         -4        143
## 32         -4       1360
## 33        -11       1309
## 34         -8       1258
## 35         -5        181
## 36         -9       1309
## 37         -8        167
## 38        -18       1306
## 39        -10       1378
## 40         -2       1313
## 41        -18       1364
## 42          0        130
## 43         -2        124
## 44         -5       1289
## 45         -8       1249
## 46         -4       1307
## 47         -5       1242
## 48         -7       1377
## 49         -3       1300
## 50         -2        157
```

```
cat("There are Totally ",nrow(df%>%filter(arr_delay>120 & dep_delay<=0)))
```

```
## There are Totally  228
```

5. Which carrier has the worst delays?

#Lets Check If Null Exists Due to Variable Creation

```
apply(df,2,anyNA)
```

```
##       dep_time sched_dep_time       arr_time sched_arr_time        carrier
##          FALSE          FALSE          FALSE          FALSE          FALSE
##         flight         origin           dest      dep_delay      arr_delay
##          FALSE          FALSE          FALSE           TRUE           TRUE
```

So We Remove Or We Doesnt Consider Them While Calculating MaxDelays So na.Rm=TRUE Below

```
WorseDelays<-df%>%group_by(carrier)%>%summarise(maxarrdel=max(arr_delay,na.rm = T),maxdepdelay=max(dep_
WorseDelays
```

```
## # A tibble: 16 x 4
##    carrier maxarrdel maxdepdelay totalDelay
##    <chr>       <dbl>       <dbl>      <dbl>
##   1 DL          1378         911       2289
##   2 B6          1378         502       1880
##   3 UA          1378         408       1786
##   4 AA          1138         636       1774
##   5 F9           242         853       1095
##   6 EV           538         548       1086
##   7 FL           436         470        906
##   8 9E           396         408        804
##   9 YV           381         387        768
##  10 WN           385         378        763
##  11 MQ           366         361        727
##  12 VX           344         367        711
##  13 US           336         374        710
##  14 AS           198         222        420
##  15 HA           154         206        360
##  16 OO           107          67        174
```

```
CarrierNa<-WorseDelays%>%filter(totalDelay==max(totalDelay))%>%select(carrier)
cat("So The Carrier With Maximum/Worse Delay is",as.character(CarrierNa))
```

```
## So The Carrier With Maximum/Worse Delay is DL
```

6. Count all flights operating between JFK and MIA airports.

```
head(df%>%filter((origin=="JFK"&dest=="MIA")|(origin=="MIA"&dest=="JFK")))
```

```
##   dep_time sched_dep_time arr_time sched_arr_time carrier flight origin dest
## 1      542            540      923            850      AA   1141    JFK  MIA
## 2      759            800     1057           1127      DL   1843    JFK  MIA
## 3      826            715     1136           1045      AA    443    JFK  MIA
## 4      912            900     1241           1220      AA    647    JFK  MIA
## 5     1306           1240     1622           1555      AA   2041    JFK  MIA
## 6     1446           1455     1803           1825      AA   1769    JFK  MIA
##   dep_delay arr_delay
## 1         2        33
## 2        -1       -30
## 3        71        51
## 4        12        21
## 5        26        27
## 6        -9       -22
```

```
count(df%>%filter((origin=="JFK"&dest=="MIA")|(origin=="MIA"&dest=="JFK")))
```

```
##      n
## 1 1677
```

7. Group by column "career" and "origin" and extract group class ('UA', 'JFK')

```
  grp<-df%>%group_by(carrier,origin)%>%summarise()
```

```
## `summarise()` has grouped output by 'carrier'. You can override using the
## `.groups` argument.
```

```
class(grp%>%filter(carrier=="UA"&origin=="JFK"))
```

```
## [1] "grouped_df" "tbl_df"     "tbl"        "data.frame"
```