Flight Data Analysis

Student number: 220692829

2024-07-29

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Use Ctrl + Alt + I (Windows/Linux) to insert a new code chunk in your RMarkdown document.

```
# Load Necessary Libraries
suppressPackageStartupMessages({
  library(data.table)
  library(dplyr)
  library(tidyr)
  library(lubridate)
  library(arrow)
  library(knitr)
  library(maditr)
  library(xtable)
  library(ggplot2)
  library(RColorBrewer)
})
setwd("D:/University of London/Programming for Data Science ST2195/ST2195_coursework_2023-24/Part 2")
```

Question 2

The Data Expo 2009: Airline On-Time Dataset provides detailed flight arrival and departure information for commercial flights within the USA from October 1987 to April 2008. With nearly 120 million records, this dataset includes variables such as departure and arrival delays, flight cancellations, and diversion indicators, offering a comprehensive view of on-time performance and operational disruptions across nearly two decades. For this analysis, we have selected a subset of the data covering the years 1998 to 2007 to focus on a decade of flight performance.

Parquet is a custom binary format designed specifically for the needs of big data. ***Source this script once to convert CSV Files to Parquet Format.***

```
#source("convert_to_parquet.R")

# Directory containing the Parquet files
parquet_dir <- "dataverse_files/parquet_files"

# Get a list of Parquet files
parquet_files <- list.files(parquet_dir, pattern = "\\.parquet$", full.names = TRUE)
```

(a) What are the best times and days of the week to minimise delays each year?

We may define time blocks as per the below, categorizing time of day by peak vs off-peak, providing insight into how traffic volume impacts delays.

Category	Hourly Range
Early Morning	0:00 - 5:59
Morning Peak	6:00 - 9:59
Midday	10:00 - 13:59
Afternoon Peak	14:00 - 17:59
Evening	18:00 - 23:59

Table 1: Time Blocks Categorization

```
# Define time blocks
time_block <- function(hour) {
  if (hour >= 0 & hour < 6) {
    return("Early Morning")
  } else if (hour >= 6 & hour < 10) {
    return("Morning Peak")
  } else if (hour >= 10 & hour < 14) {
    return("Midday")
  } else if (hour >= 14 & hour < 18) {
    return("Afternoon Peak")
  } else {
    return("Evening")
  }
}
```

```
# Function to process year data
process_year_data <- function(file_path, year) {
  data <- read_parquet(file_path)

  # Extract necessary features and keep only relevant columns
  data <- data %>%
    mutate(
      DayOfWeek = wday(ymd(paste(Year, Month, DayofMonth, sep = "-")), label = TRUE),
      HourOfDay = floor(CRSDepTime / 100),
      TimeBlock = sapply(HourOfDay, time_block),
      ArrivalDelay = ArrDelay,
      DepartureDelay = DepDelay
    ) %>%
    filter(!is.na(ArrivalDelay) & !is.na(DepartureDelay)) %>%
    select(DayOfWeek, TimeBlock, ArrivalDelay, DepartureDelay)

  print(summary(data))

  # Frequency table
  frequency_table <- data %>%
    count(DayOfWeek, TimeBlock) %>%
    rename("Day of Week" = DayOfWeek, "Time Block" = TimeBlock, "Frequency" = n) %>%
    pivot_wider(names_from = `Time Block`, values_from = Frequency, values_fill = 0)

  # Create Contingency Table for Arrival Delay
```

```r
  # Calculate the means for each combination of DayOfWeek and TimeBlock
  contingency_arrival <- data %>%
    group_by(DayOfWeek, TimeBlock) %>%
    summarize(ArrivalDelay = mean(ArrivalDelay)) %>%
    ungroup()

  # Round the ArrivalDelay values
  contingency_arrival$ArrivalDelay <- round(contingency_arrival$ArrivalDelay, 2)

  # Reshape the data to wide format to create the contingency table
  contingency_arrival <- dcast(contingency_arrival, DayOfWeek ~ TimeBlock, value.var = "ArrivalDelay")

  # Create Contingency Table for Departure Delay
  # Calculate the means for each combination of DayOfWeek and TimeBlock
  contingency_departure <- data %>%
    group_by(DayOfWeek, TimeBlock) %>%
    summarize(DepartureDelay = mean(DepartureDelay)) %>%
    ungroup()

  # Round the DepartureDelay values
  contingency_departure$DepartureDelay <- round(contingency_departure$DepartureDelay, 2)

  # Reshape the data to wide format to create the contingency table
  contingency_departure <- dcast(contingency_departure, DayOfWeek ~ TimeBlock, value.var = "DepartureDel

  # Return the tables
  list(frequency_table = frequency_table, contingency_arrival = contingency_arrival, contingency_departu
}
```

```r
# Define the function
plot_contingency_table <- function(data) {
  # Convert the data to a table
  my.tab <- as.table(as.matrix(data[, -1]))
  rownames(my.tab) <- data$DayOfWeek

  # Plot settings
  par(mgp = c(1.5, .3, 0))
  plot(
    0,
    0,
    pch = "",
    xlim = c(0.5, 5.5),
    ylim = c(0.5, 7.5),
    axes = FALSE,
    xlab = "Time of Day",
    ylab = ""
  )

  # Add the bubbles to the plot
  for (i in 1:nrow(my.tab)) {
    symbols(
      x = 1:ncol(my.tab),
      y = rep(i, ncol(my.tab)),
```

```
      circles = sqrt(my.tab[i, ] / 200 / pi),
      add = TRUE,
      inches = FALSE,
      fg = "lightblue",
      bg = "lightblue"
    )
  }

  # Add axes
  axis(1, col = "white", col.axis = "black", at = 1:5, labels = colnames(my.tab))
  axis(2, at = 1:7, labels = rownames(my.tab), las = 1, col.axis = "black", col = "white")

  # Add numbers to plot
  for (i in 1:nrow(my.tab)) {
    text(1:ncol(my.tab), rep(i, ncol(my.tab)), labels = round(my.tab[i, ], 2))
  }
}

#https://www.mzes.uni-mannheim.de/socialsciencedatalab/article/datavis/#:~:text=Code%3A%20Advanced%20co
```

```
# Define file paths and years
years <- 1998:2007
file_paths <- paste0("dataverse_files/parquet_files/flights_data_", years, ".parquet")

# Process each year and store results
results <- list()
for (i in seq_along(years)) {
  year <- years[i]
  message("Reviewing data for ", years[i])

  results[[as.character(years[i])]] <- suppressMessages(process_year_data(file_paths[i], years[i]))

  message("## Frequency Table for ", year, "\n")
  results[[as.character(years[i])]]$frequency_table <-
    results[[as.character(years[i])]]$frequency_table[, c('Day of Week', 'Early Morning',
                                                          'Morning Peak', 'Midday',
                                                          'Afternoon Peak', 'Evening')]
  print.data.frame(results[[as.character(years[i])]]$frequency_table)

  message("## Contingency Table for Arrival Delays in ", year, "\n")
  results[[as.character(years[i])]]$contingency_arrival <-
    results[[as.character(years[i])]]$contingency_arrival[, c('DayOfWeek', 'Early Morning',
                                                              'Morning Peak', 'Midday',
                                                              'Afternoon Peak', 'Evening')]

  #print.data.frame(results[[as.character(years[i])]]$contingency_arrival)
  plot_contingency_table(results[[as.character(years[i])]]$contingency_arrival)

  message("## Contingency Table for Departure Delays in ", year, "\n")
  results[[as.character(years[i])]]$contingency_departure <-
    results[[as.character(years[i])]]$contingency_departure[, c('DayOfWeek', 'Early Morning',
                                                              'Morning Peak', 'Midday',
                                                              'Afternoon Peak', 'Evening')]
```

```
  #print.data.frame(results[[as.character(years[i])]]$contingency_departure)
  plot_contingency_table(results[[as.character(years[i])]]$contingency_departure)

  message("======================================================================")
}
```

```
## Reviewing data for 1998
```

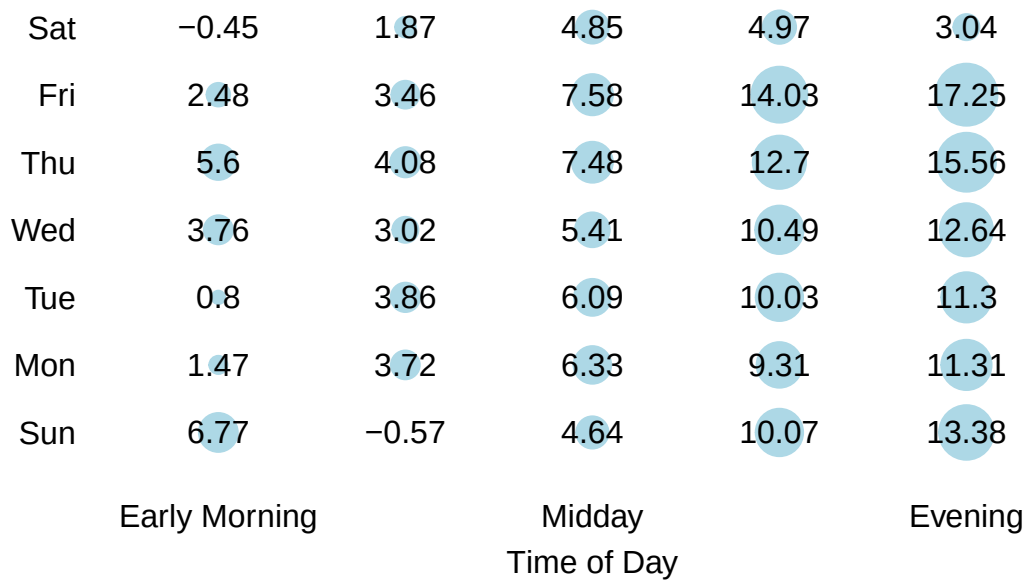
```
##  DayOfWeek     TimeBlock          ArrivalDelay       DepartureDelay
##  Sun:728289   Length:5227051     Min.   :-1233.000   Min.   :-111.000
##  Mon:761314   Class :character   1st Qu.:   -7.000   1st Qu.:  -2.000
##  Tue:761553   Mode  :character   Median :    0.000   Median :   0.000
##  Wed:763501                      Mean   :    7.587   Mean   :   8.967
##  Thu:773397                      3rd Qu.:   11.000   3rd Qu.:   7.000
##  Fri:761774                      Max.   : 1808.000   Max.   :1800.000
##  Sat:677223
```

```
## ## Frequency Table for 1998
```

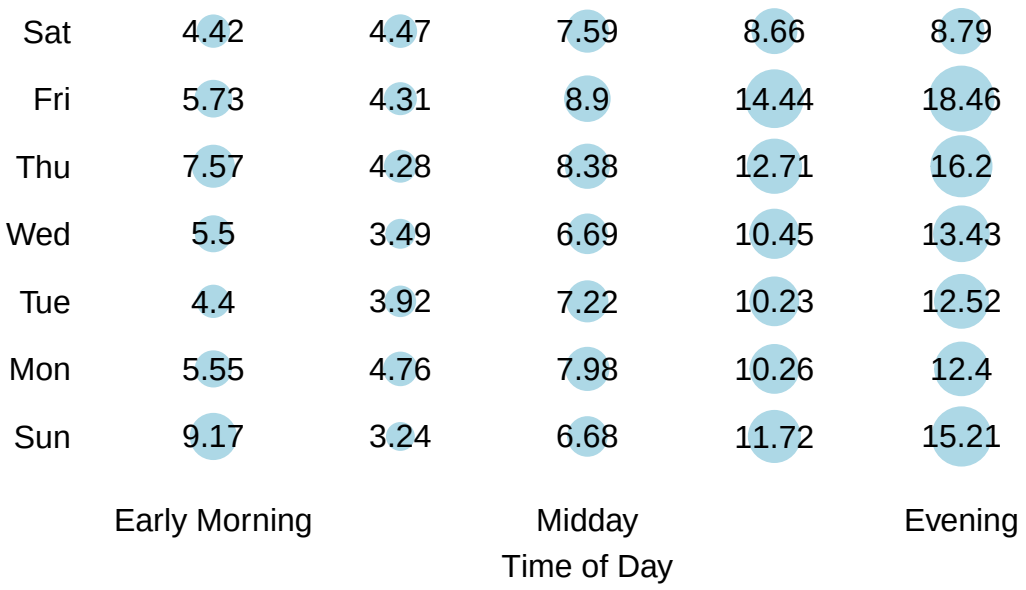
```
##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun         20195       167999 183519         182893  173683
## 2         Mon         23567       201041 184124         181640  170942
## 3         Tue         19885       202442 184695         182863  171668
## 4         Wed         17854       203178 185425         184041  173003
## 5         Thu         19634       205688 188460         186160  173455
## 6         Fri         19855       201164 185616         183146  171993
## 7         Sat         18790       186630 178083         165859  127861
```

```
## ## Contingency Table for Arrival Delays in 1998
```

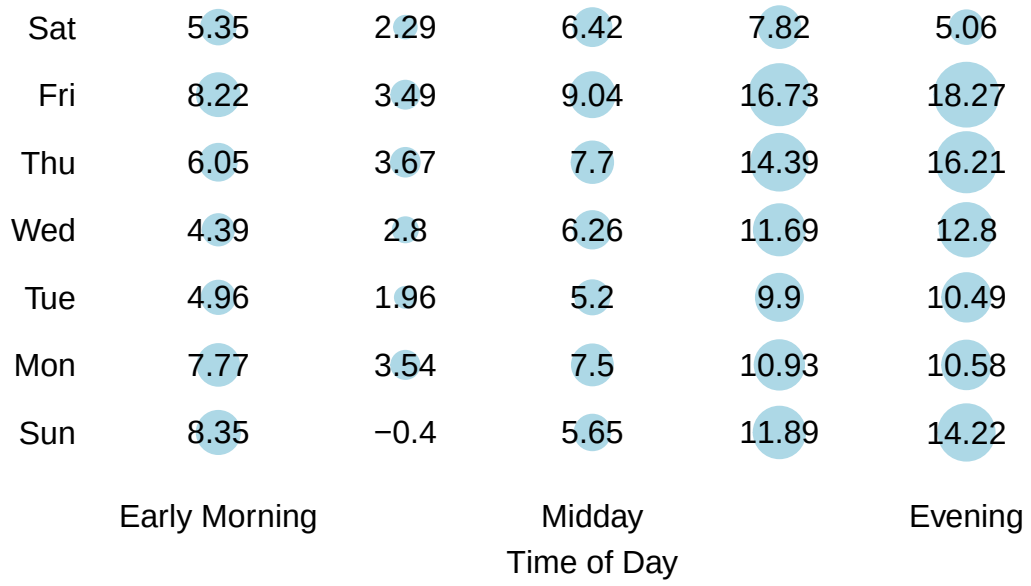
```
## ## Contingency Table for Departure Delays in 1998
```



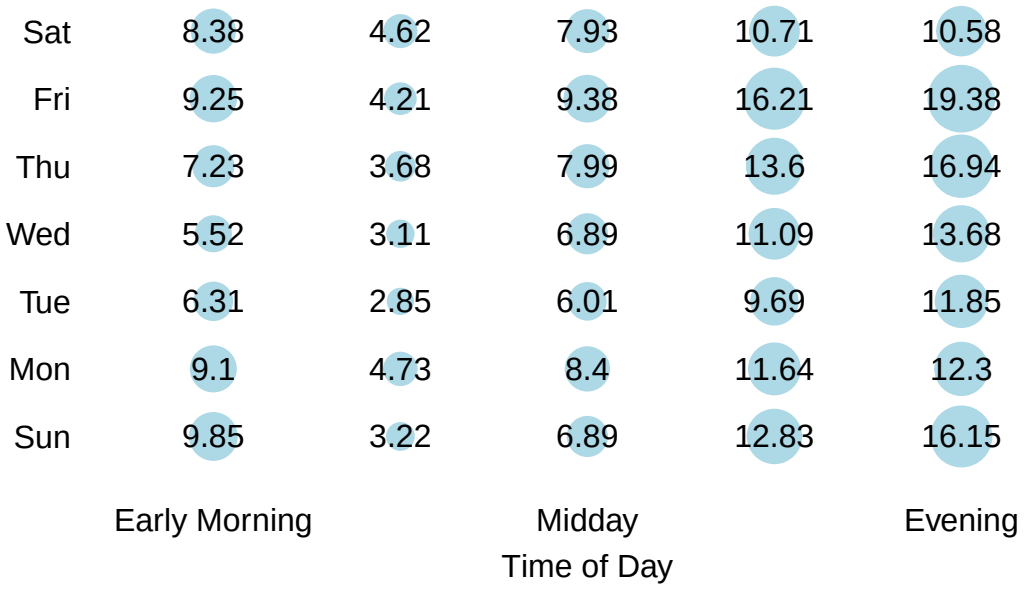

```
## ================================================================

## Reviewing data for 1999

##  DayOfWeek     TimeBlock          ArrivalDelay      DepartureDelay
##  Sun:745731   Length:5360018     Min.   :-194.000   Min.   : -76.000
##  Mon:785296   Class :character   1st Qu.:  -7.000   1st Qu.:  -2.000
##  Tue:786369   Mode  :character   Median :   0.000   Median :   0.000
##  Wed:787509                      Mean   :   8.247   Mean   :   9.277
##  Thu:781895                      3rd Qu.:  12.000   3rd Qu.:   7.000
##  Fri:791192                      Max.   :1724.000   Max.   :1740.000
##  Sat:682026

## ## Frequency Table for 1999

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun         36455       168494 182826         183342  174614
## 2         Mon         44168       202087 183100         182476  173465
## 3         Tue         38954       204565 184315         183970  174565
## 4         Wed         37691       204648 184744         184976  175450
## 5         Thu         36399       204916 184107         182923  173550
## 6         Fri         38904       205426 187390         185372  174100
## 7         Sat         35146       182990 175684         163997  124209

## ## Contingency Table for Arrival Delays in 1999
```



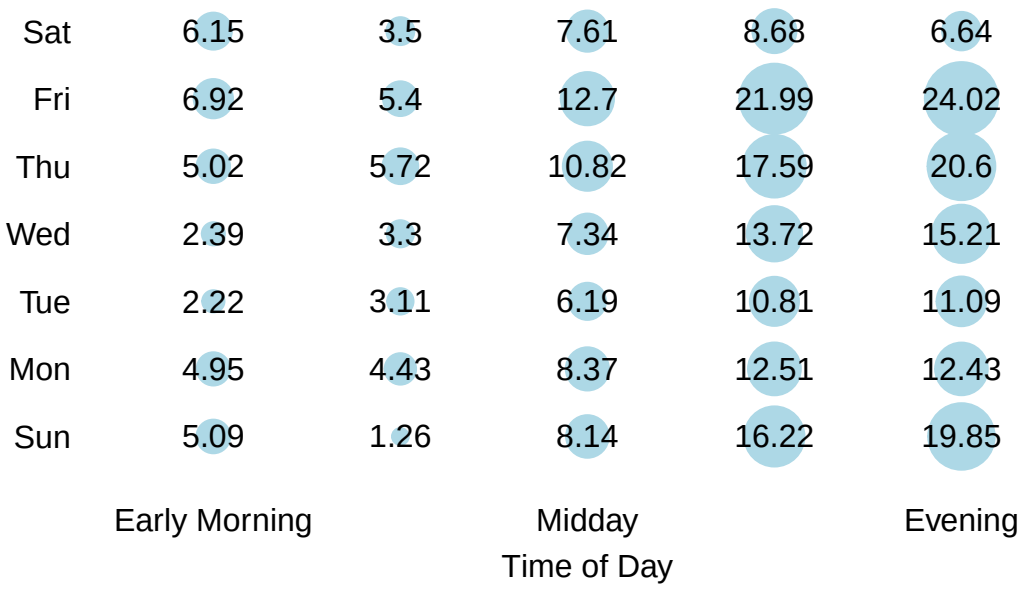

```
## ## Contingency Table for Departure Delays in 1999
```



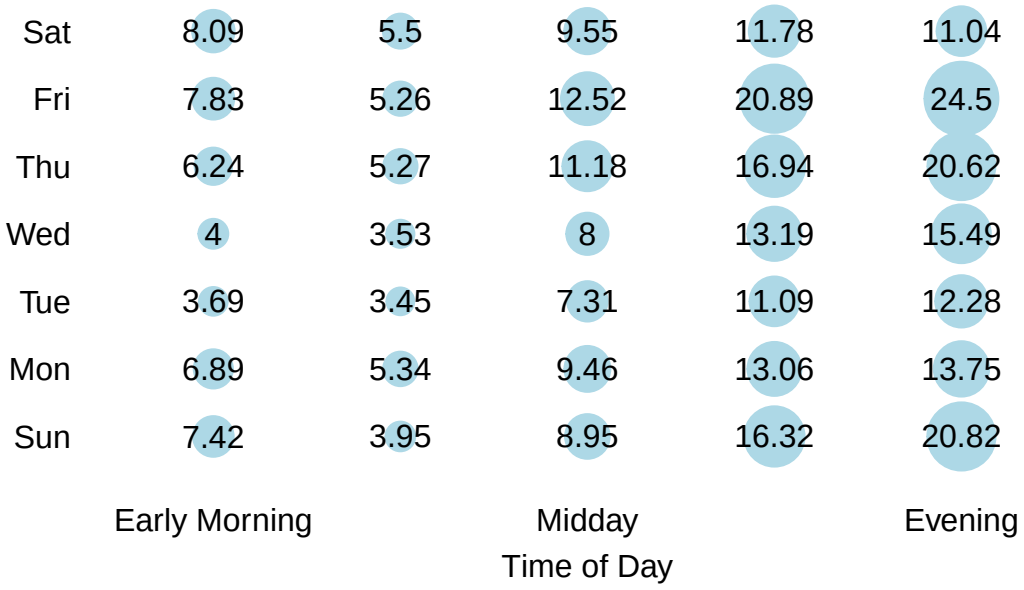

```
## ================================================================

## Reviewing data for 2000

##  DayOfWeek     TimeBlock          ArrivalDelay      DepartureDelay
##  Sun:777257   Length:5481303     Min.   :-1298.00   Min.   :-990.00
##  Mon:795489   Class :character   1st Qu.:   -7.00   1st Qu.:  -2.00
##  Tue:794933   Mode  :character   Median :    1.00   Median :   0.00
##  Wed:799738                      Mean   :   10.47   Mean   :  11.21
##  Thu:797754                      3rd Qu.:   14.00   3rd Qu.:  10.00
##  Fri:799240                      Max.   : 1441.00   Max.   :1435.00
##  Sat:716892

## ## Frequency Table for 2000

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          7142       190379 196190          196807  186739
## 2         Mon          8198       216350 194361          193055  183525
## 3         Tue          8123       216595 193444          193030  183741
## 4         Wed          8174       218349 194793          193677  184745
## 5         Thu          8250       218059 193934          193223  184288
## 6         Fri          8385       217579 194921          193372  184983
## 7         Sat          7922       200781 189945          179295  138949

## ## Contingency Table for Arrival Delays in 2000
```



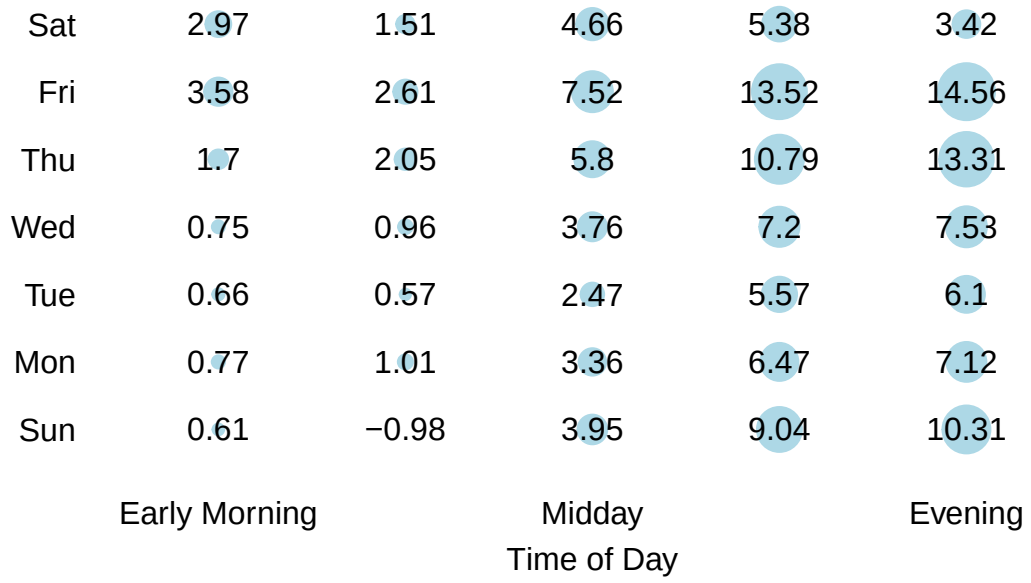

```
## ## Contingency Table for Departure Delays in 2000
```



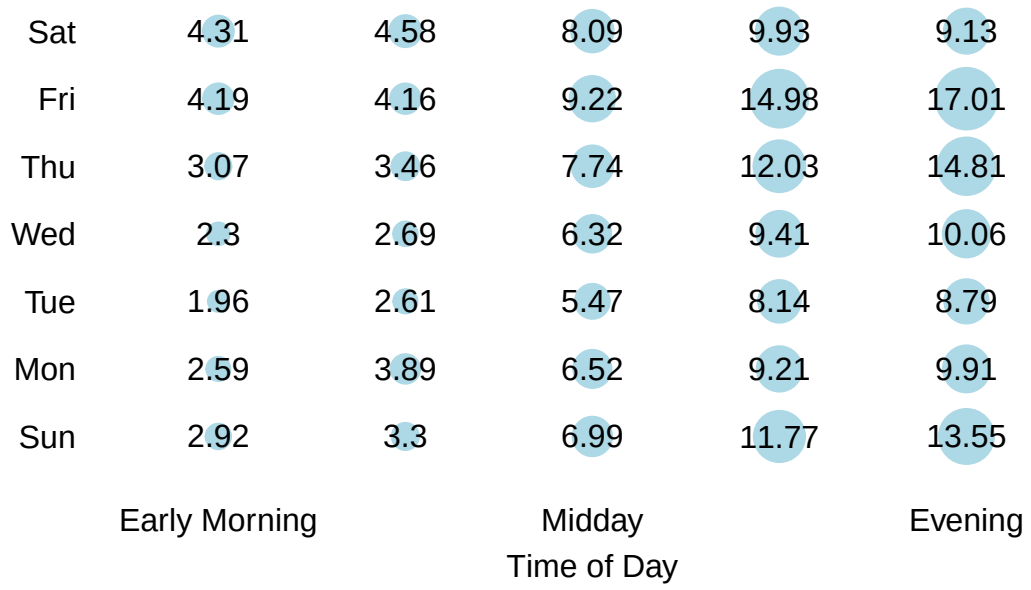

```
## ================================================================

## Reviewing data for 2001

##  DayOfWeek     TimeBlock          ArrivalDelay       DepartureDelay
##  Sun:805349   Length:5723673     Min.   :-1116.000   Min.   :-204.000
##  Mon:850475   Class :character   1st Qu.:   -9.000   1st Qu.:  -3.000
##  Tue:827109   Mode  :character   Median :   -2.000   Median :   0.000
##  Wed:827445                      Mean   :    5.528   Mean   :   8.115
##  Thu:830330                      3rd Qu.:   10.000   3rd Qu.:   6.000
##  Fri:836054                      Max.   : 1688.000   Max.   :1692.000
##  Sat:746911

## ## Frequency Table for 2001

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          5970       200955 202857         209032  186535
## 2         Mon          7443       234174 208909         212061  187888
## 3         Tue          7260       228790 201279         205740  184040
## 4         Wed          7146       227676 201593         206448  184582
## 5         Thu          7186       229277 202524         206846  184497
## 6         Fri          7281       229434 204165         208584  186590
## 7         Sat          6626       210380 195985         191102  142818

## ## Contingency Table for Arrival Delays in 2001
```



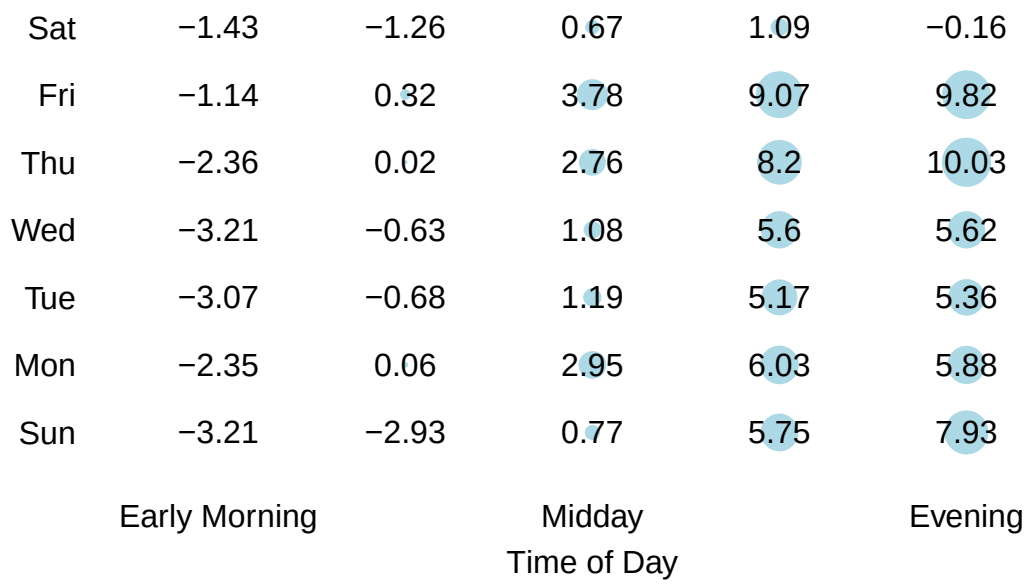

```
## ## Contingency Table for Departure Delays in 2001
```



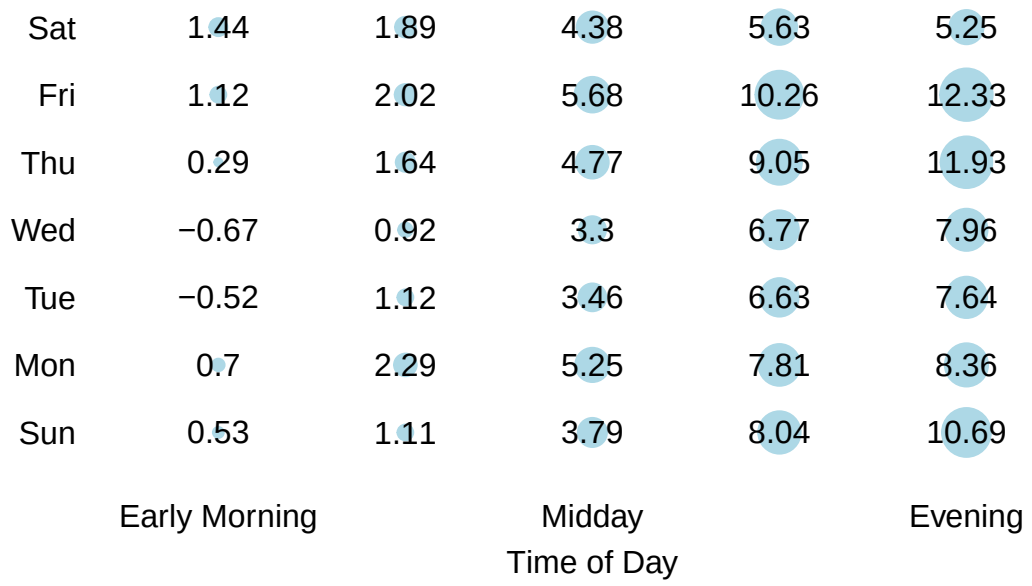

```
## ================================================================

## Reviewing data for 2002

##  DayOfWeek     TimeBlock          ArrivalDelay      DepartureDelay
##  Sun:725917   Length:5197860     Min.   :-987.000   Min.   :-1370.000
##  Mon:763569   Class :character   1st Qu.: -10.000   1st Qu.:   -4.000
##  Tue:767124   Mode  :character   Median :  -3.000   Median :    0.000
##  Wed:757254                      Mean   :   3.191   Mean   :    5.498
##  Thu:755108                      3rd Qu.:   8.000   3rd Qu.:    4.000
##  Fri:761622                      Max.   :2137.000   Max.   : 2119.000
##  Sat:667266

## ## Frequency Table for 2002

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          4594       175760 189317         194398  161848
## 2         Mon          5739       208170 192629         195505  161526
## 3         Tue          5840       210348 193707         196698  160531
## 4         Wed          5691       206483 190930         194495  159655
## 5         Thu          5679       206888 190829         193103  158609
## 6         Fri          5737       207288 192282         195122  161193
## 7         Sat          5298       186770 182347         175013  117838

## ## Contingency Table for Arrival Delays in 2002
```



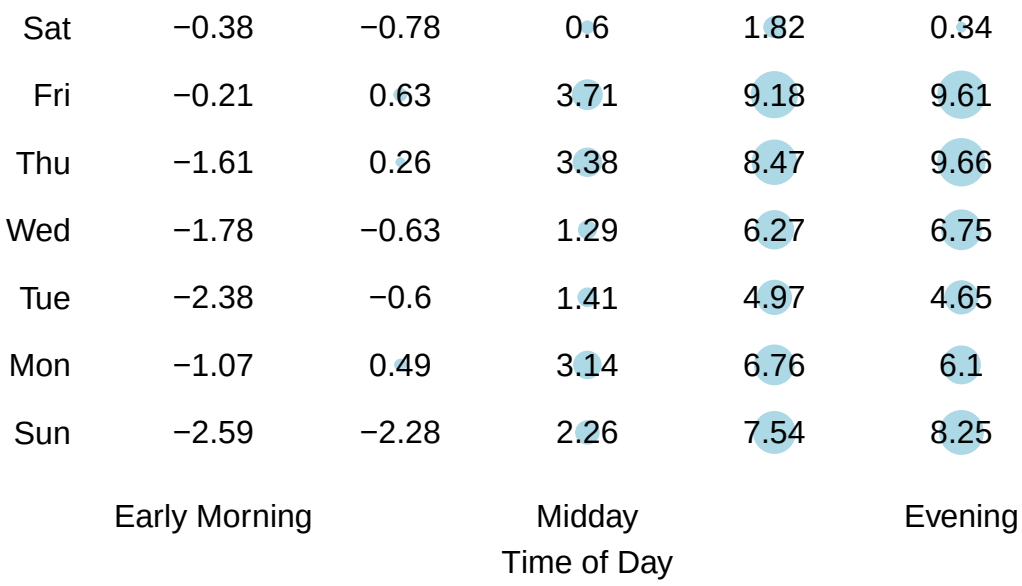

```
## ## Contingency Table for Departure Delays in 2002
```



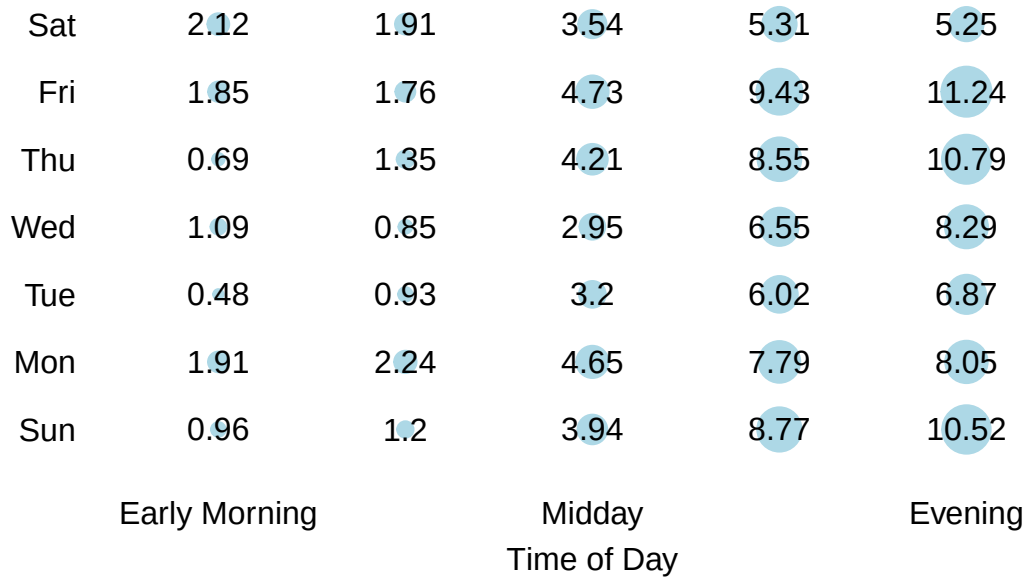

```
## ================================================================

## Reviewing data for 2003

##  DayOfWeek     TimeBlock          ArrivalDelay      DepartureDelay
##  Sun:886483   Length:6375689     Min.   :-937.000   Min.   :-1410.000
##  Mon:938114   Class :character   1st Qu.: -10.000   1st Qu.:   -4.000
##  Tue:928988   Mode  :character   Median :  -3.000   Median :    0.000
##  Wed:947120                      Mean   :   3.597   Mean   :    5.223
##  Thu:935043                      3rd Qu.:   7.000   3rd Qu.:    2.000
##  Fri:938553                      Max.   :1612.000   Max.   : 1582.000
##  Sat:801388

## ## Frequency Table for 2003

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          6338       204299 236308         239409  200129
## 2         Mon          7765       249343 240761         241183  199062
## 3         Tue          7761       246972 234760         241467  198028
## 4         Wed          7867       251221 241233         246165  200634
## 5         Thu          7731       249391 240526         240128  197267
## 6         Fri          7682       248661 241460         241595  199155
## 7         Sat          7140       217570 224949         211118  140611

## ## Contingency Table for Arrival Delays in 2003
```



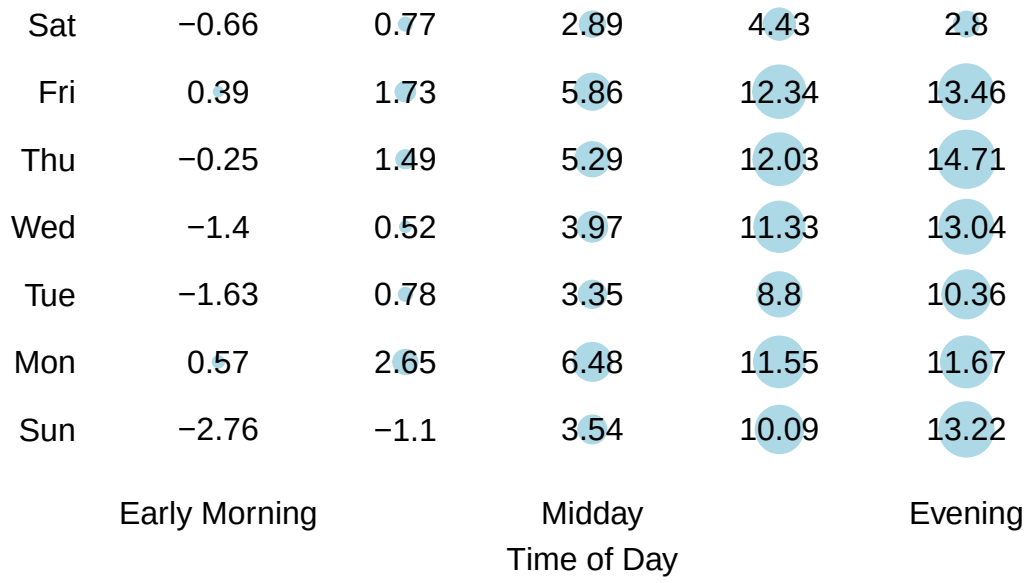

```
## ## Contingency Table for Departure Delays in 2003
```



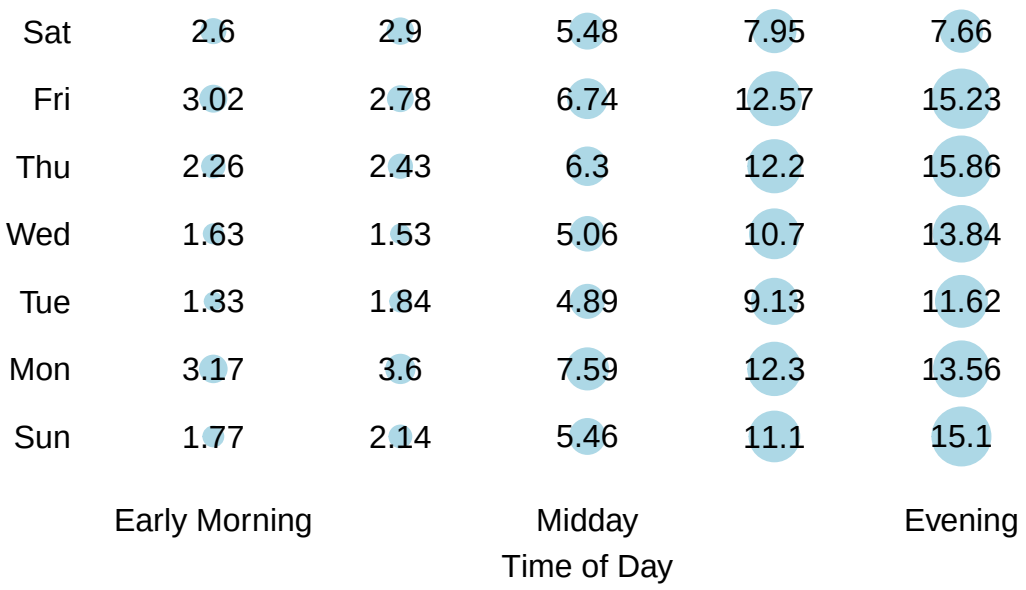

```
## ================================================================

## Reviewing data for 2004

##  DayOfWeek      TimeBlock          ArrivalDelay       DepartureDelay
##  Sun: 970075   Length:6987729     Min.   :-1302.000   Min.   :-1197.000
##  Mon:1022398   Class :character   1st Qu.:   -9.000   1st Qu.:   -4.000
##  Tue:1013439   Mode  :character   Median :   -2.000   Median :    0.000
##  Wed:1014993                      Mean   :    6.508   Mean   :    7.859
##  Thu:1038525                      3rd Qu.:   10.000   3rd Qu.:    6.000
##  Fri:1040966                      Max.   : 1879.000   Max.   : 1882.000
##  Sat: 887333

## ## Frequency Table for 2004

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          8830       231799 255087          256397  217962
## 2         Mon         10299       272823 260369          259502  219405
## 3         Tue         10204       271665 254376          259034  218160
## 4         Wed         10126       271558 255715          258808  218786
## 5         Thu         10310       278368 264655          263039  222153
## 6         Fri         10327       277864 265455          264283  223037
## 7         Sat          9779       243007 242289          227629  164629

## ## Contingency Table for Arrival Delays in 2004
```



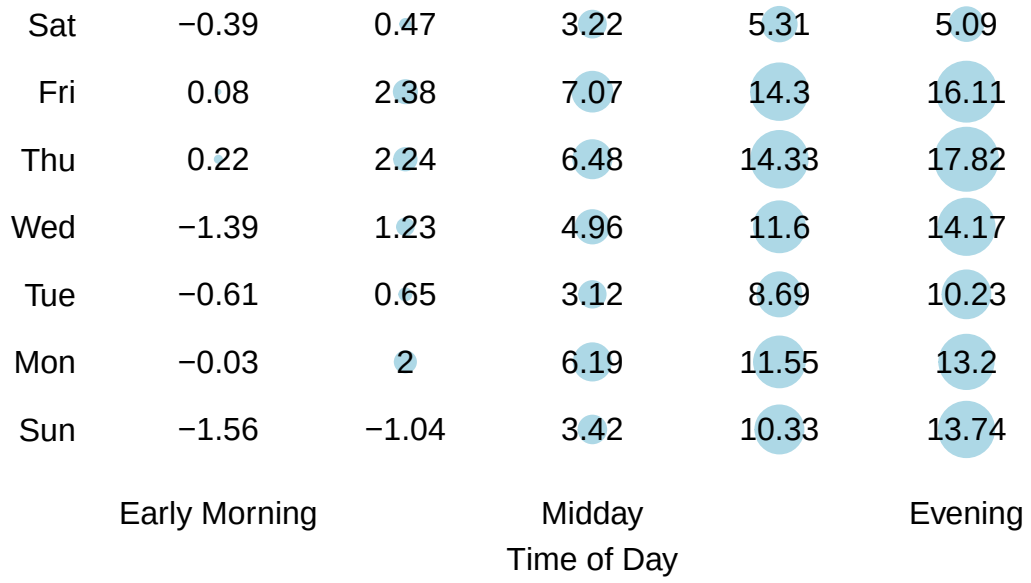

```
## ## Contingency Table for Departure Delays in 2004
```



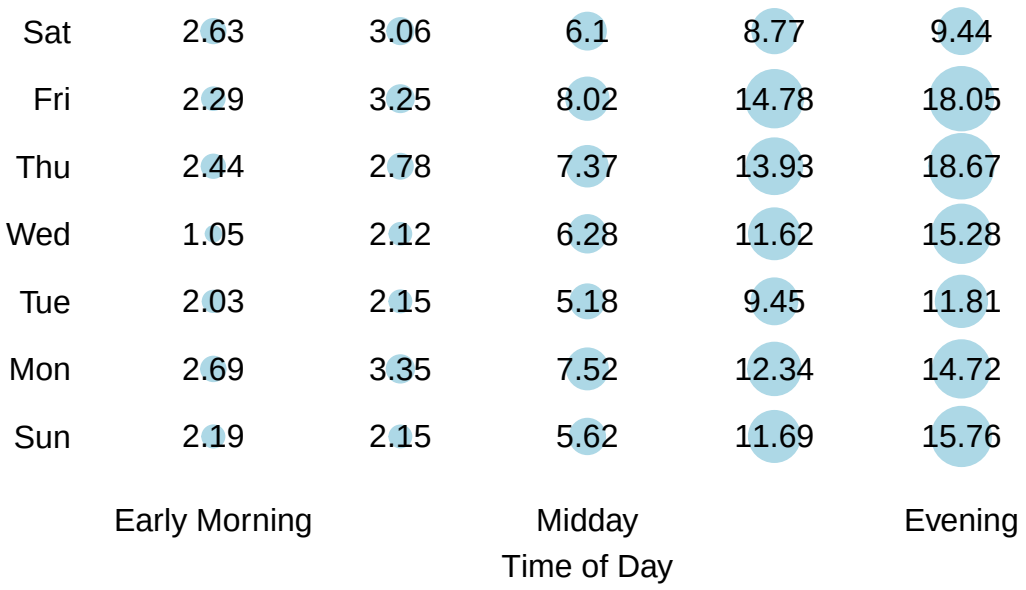

```
## ================================================================

## Reviewing data for 2005

##  DayOfWeek      TimeBlock          ArrivalDelay      DepartureDelay
##  Sun: 973742   Length:6992838     Min.   :-939.000   Min.   :-1199.000
##  Mon:1026555   Class :character   1st Qu.:  -9.000   1st Qu.:   -4.000
##  Tue:1014880   Mode  :character   Median :  -1.000   Median :    0.000
##  Wed:1019288                      Mean   :   7.181   Mean   :    8.646
##  Thu:1026029                      3rd Qu.:  11.000   3rd Qu.:    7.000
##  Fri:1031735                      Max.   :1925.000   Max.   : 1930.000
##  Sat: 900609

## ## Frequency Table for 2005

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          8887       233976 254557          255930  220392
## 2         Mon         11113       275376 259935          259478  220653
## 3         Tue         11004       273311 253380          258224  218961
## 4         Wed         11019       274608 255575          258065  220021
## 5         Thu         11114       277176 259673          258141  219925
## 6         Fri         11146       276674 260547          260663  222705
## 7         Sat         10693       248407 244689          229000  167820

## ## Contingency Table for Arrival Delays in 2005
```



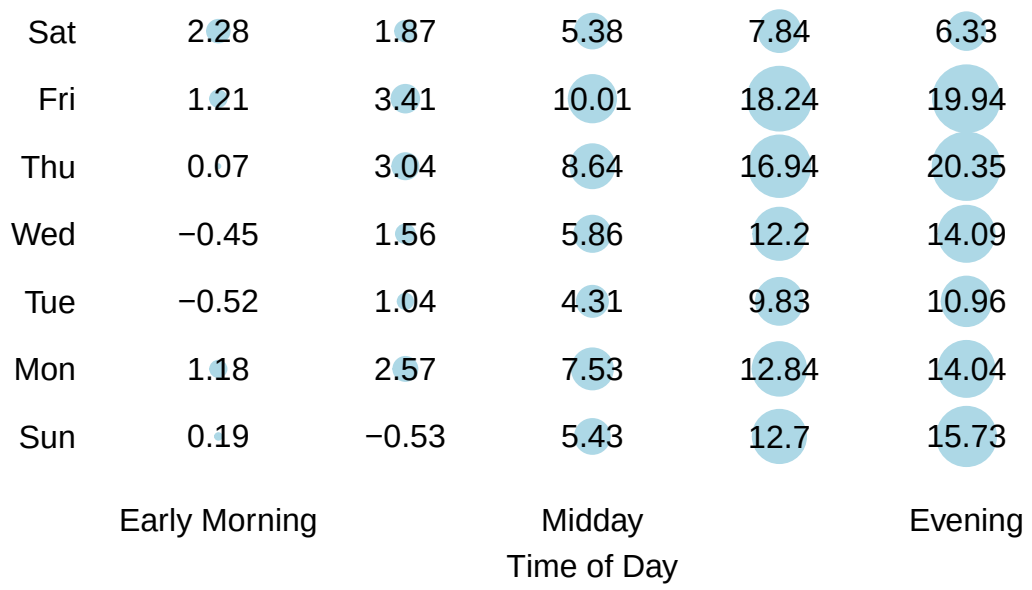

```
## ## Contingency Table for Departure Delays in 2005
```



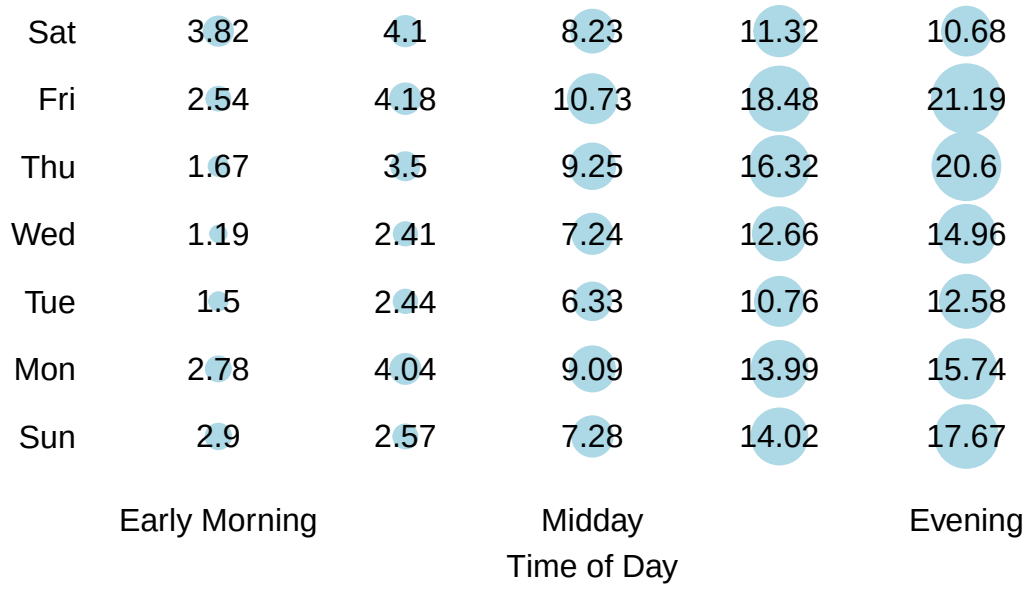

```
## ================================================================

## Reviewing data for 2006

##  DayOfWeek      TimeBlock          ArrivalDelay      DepartureDelay
##  Sun: 993653   Length:7003802     Min.   :-592.000   Min.   :-1200.00
##  Mon:1027988   Class :character   1st Qu.:  -9.000   1st Qu.:   -4.00
##  Tue:1012483   Mode  :character   Median :  -1.000   Median :    0.00
##  Wed:1022327                      Mean   :   8.683   Mean   :   10.06
##  Thu:1028443                      3rd Qu.:  13.000   3rd Qu.:    9.00
##  Fri:1032497                      Max.   :1779.000   Max.   : 1752.00
##  Sat: 886411

## ## Frequency Table for 2006

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          6322       243573 260569          264224  218965
## 2         Mon          8506       280838 260417          262530  215697
## 3         Tue          8413       278256 251607          260648  213559
## 4         Wed          8401       281409 255986          261524  215007
## 5         Thu          8595       283861 260136          260717  215134
## 6         Fri          8629       282471 260767          262111  218519
## 7         Sat          7921       249328 241351          227934  159877

## ## Contingency Table for Arrival Delays in 2006
```



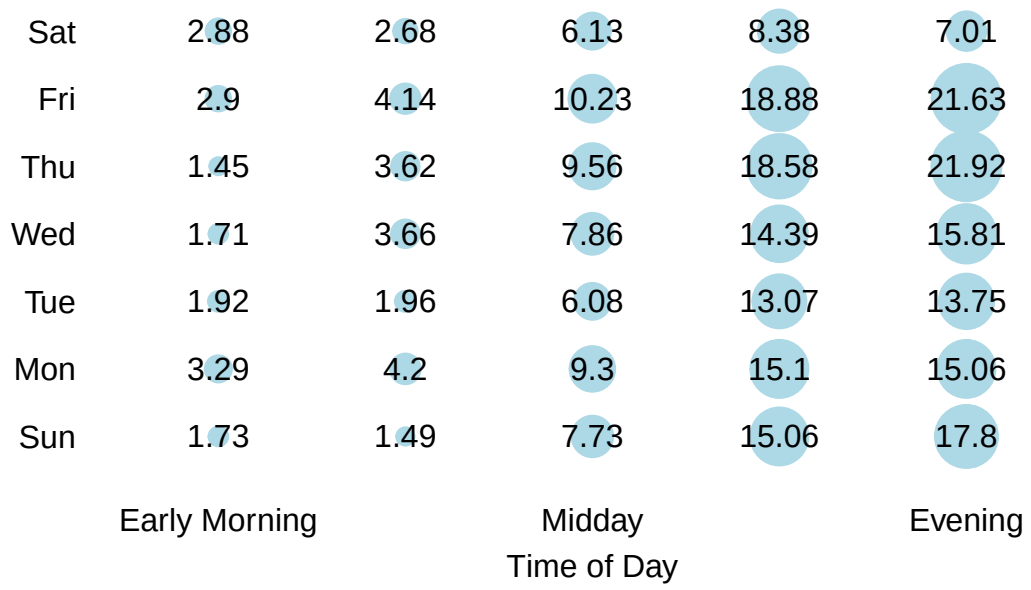

```
## ## Contingency Table for Departure Delays in 2006
```



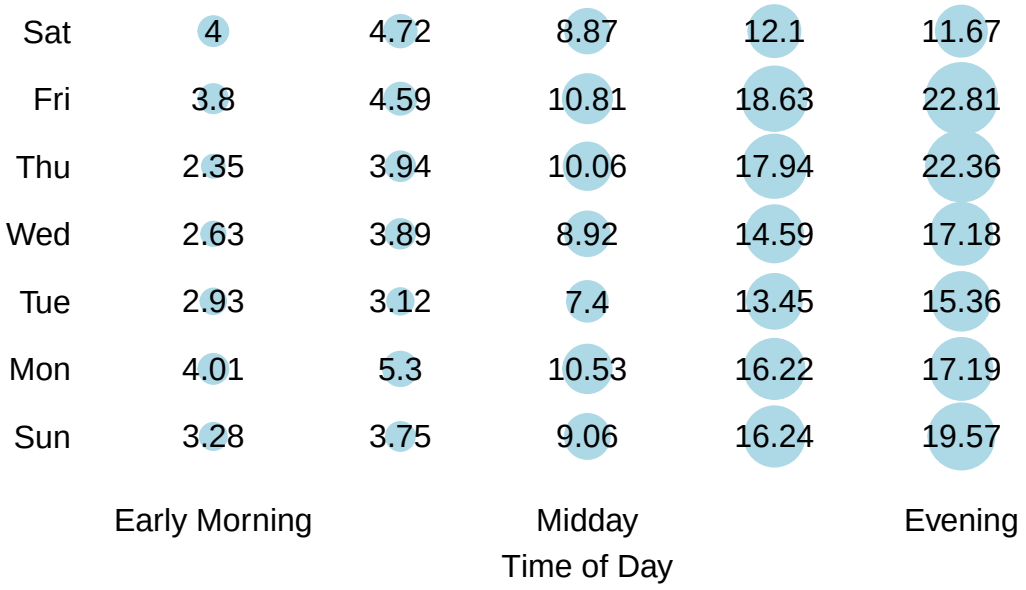

```
## ================================================================

## Reviewing data for 2007

##  DayOfWeek      TimeBlock          ArrivalDelay     DepartureDelay
##  Sun:1017148   Length:7275288     Min.   :-312.00   Min.   :-305.00
##  Mon:1087794   Class :character   1st Qu.:  -9.00   1st Qu.:  -4.00
##  Tue:1050354   Mode  :character   Median :   0.00   Median :   0.00
##  Wed:1059851                      Mean   :  10.19   Mean   :  11.36
##  Thu:1069805                      3rd Qu.:  14.00   3rd Qu.:  11.00
##  Fri:1074710                      Max.   :2598.00   Max.   :2601.00
##  Sat: 915626

## ## Frequency Table for 2007

##   Day of Week Early Morning Morning Peak Midday Afternoon Peak Evening
## 1         Sun          6745       250489 265602          266485  227827
## 2         Mon          9333       299448 276669          273913  228431
## 3         Tue          9113       292309 264628          264765  219539
## 4         Wed          9091       293062 267690          266858  223150
## 5         Thu          9249       297436 271462          266768  224890
## 6         Fri          9225       296493 271920          269158  227914
## 7         Sat          8390       260478 250375          233379  163004

## ## Contingency Table for Arrival Delays in 2007
```




```
## ## Contingency Table for Departure Delays in 2007
```


Sat
Fri
Thu
Wed
Tue
Mon
Sun
2.88
2.68
6.13
8.38
7.01
2.9
4.14
10.23
18.88
21.63
1.45
3.62
9.56
18.58
21.92
1.71
3.66
7.86
14.39
15.81
1.92
1.96
6.08
13.07
13.75
3.29
4.2
9.3
15.1
15.06
1.73
1.49
7.73
15.06
17.8
Early Morning
Midday
Evening
Time of Day

```
## ============================================================
```



A contingency table displays frequencies for combinations of two categorical variables in a matrix format to analyze the relationship between them. In this case, contingency tables were used to analyse the mean Arrival and Departure Delays. It is visually evident that flights before 10am (Early Morning; Morning Peak) are more punctual, experiencing less delays. There is less difference amongst days of the week, although flights have the worst delays on Friday, with Saturdays and Tuesdays performing better.