Assignment 6 - CT5102

Manipulating Data with dplyr

Include the following libraries.

```
library(nycflights13)
library(dplyr)
library(ggplot2)
library(lubridate)
```

Create a local copy of the flights data set

```
my_flights <- flights
my_flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   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>
```

Filter out missing values for dep_delay and arr_delay, and select the following columns from the data set: time_hour, origin, dest, carrier, dep_delay, arr_delay, air_time, distance.

```
## # A tibble: 327,346 x 8
##    time_hour           origin dest  carrier dep_delay arr_delay air_time
##    <dttm>              <chr>  <chr> <chr>       <dbl>     <dbl>    <dbl>
##  1 2013-01-01 05:00:00 EWR    IAH   UA              2        11      227
##  2 2013-01-01 05:00:00 LGA    IAH   UA              4        20      227
##  3 2013-01-01 05:00:00 JFK    MIA   AA              2        33      160
##  4 2013-01-01 05:00:00 JFK    BQN   B6             -1       -18      183
##  5 2013-01-01 06:00:00 LGA    ATL   DL             -6       -25      116
##  6 2013-01-01 05:00:00 EWR    ORD   UA             -4        12      150
##  7 2013-01-01 06:00:00 EWR    FLL   B6             -5        19      158
##  8 2013-01-01 06:00:00 LGA    IAD   EV             -3       -14       53
##  9 2013-01-01 06:00:00 JFK    MCO   B6             -3        -8      140
## 10 2013-01-01 06:00:00 LGA    ORD   AA             -2         8      138
## # ... with 327,336 more rows, and 1 more variable: distance <dbl>
```

Add columns for the day of the week, the hour of the day and month from lubridate.

```
## # A tibble: 327,346 x 11
##    time_hour           Month DayOfWeek HourOfDay origin dest  carrier
##    <dttm>              <ord> <ord>         <int> <chr>  <chr> <chr>
##  1 2013-01-01 05:00:00 Jan   Tue               5 EWR    IAH   UA
##  2 2013-01-01 05:00:00 Jan   Tue               5 LGA    IAH   UA
##  3 2013-01-01 05:00:00 Jan   Tue               5 JFK    MIA   AA
##  4 2013-01-01 05:00:00 Jan   Tue               5 JFK    BQN   B6
##  5 2013-01-01 06:00:00 Jan   Tue               6 LGA    ATL   DL
##  6 2013-01-01 05:00:00 Jan   Tue               5 EWR    ORD   UA
##  7 2013-01-01 06:00:00 Jan   Tue               6 EWR    FLL   B6
##  8 2013-01-01 06:00:00 Jan   Tue               6 LGA    IAD   EV
##  9 2013-01-01 06:00:00 Jan   Tue               6 JFK    MCO   B6
## 10 2013-01-01 06:00:00 Jan   Tue               6 LGA    ORD   AA
## # ... with 327,336 more rows, and 4 more variables: dep_delay <dbl>,
## #   arr_delay <dbl>, air_time <dbl>, distance <dbl>
```

Calculate the average departure delay statistics by hour of day, ordered by delay.

```
## # A tibble: 19 x 2
##    HourOfDay AvrDepDelay
##        <int>       <dbl>
##  1        19      24.7
##  2        20      24.2
##  3        21      24.2
##  4        17      21.0
##  5        18      21.0
##  6        22      18.7
##  7        16      18.6
##  8        15      16.8
##  9        23      14.0
## 10        14      13.7
## 11        13      11.3
## 12        12       8.52
## 13        11       7.15
## 14        10       6.45
## 15         9       4.54
## 16         8       4.11
## 17         7       1.91
## 18         6       1.60
## 19         5       0.689
```

Average departure delay statistics by month, ordered by delay.

```
## # A tibble: 12 x 2
##    Month AvrDepDelay
##    <ord>       <dbl>
##  1 Jul         21.5
##  2 Jun         20.7
##  3 Dec         16.5
##  4 Apr         13.8
##  5 Mar         13.2
##  6 May         12.9
##  7 Aug         12.6
##  8 Feb         10.8
##  9 Jan          9.99
## 10 Sep          6.63
## 11 Oct          6.23
## 12 Nov          5.42
```

Average departure delay statistics by carrier, ordered by delay.

```
## # A tibble: 16 x 2
##    carrier AvrDepDelay
##    <chr>         <dbl>
##  1 F9             20.2
##  2 EV             19.8
##  3 YV             18.9
##  4 FL             18.6
##  5 WN             17.7
##  6 9E             16.4
##  7 B6             13.0
##  8 VX             12.8
##  9 OO             12.6
## 10 UA             12.0
## 11 MQ             10.4
## 12 DL              9.22
## 13 AA              8.57
## 14 AS              5.83
## 15 HA              4.90
## 16 US              3.74
```

Average departure delay statistics by airport by month, ordered by delay.

```
## # A tibble: 36 x 3
## # Groups:   origin [3]
##    origin Month AvrDepDelay
##    <chr>  <ord>       <dbl>
##  1 JFK    Jul          23.5
##  2 EWR    Jun          22.3
##  3 EWR    Jul          21.9
##  4 EWR    Dec          20.9
##  5 JFK    Jun          20.3
##  6 LGA    Jun          19.3
##  7 LGA    Jul          18.8
##  8 EWR    Mar          18.1
##  9 EWR    Apr          17.3
## 10 EWR    May          15.2
## # ... with 26 more rows
```

Add a new category, which divides each day into three sections (use case_when) - Morning 5 <= time < 12 - Afternoon 12 <= time < 18 - Evening > =18 Average departure delay statistics by airport by month, ordered by delay.

```
## # A tibble: 327,346 x 12
##    time_hour           DaySection Month DayOfWeek HourOfDay origin dest
##    <dttm>              <chr>      <ord> <ord>         <int> <chr>  <chr>
##  1 2013-01-01 05:00:00 Morning    Jan   Tue               5 EWR    IAH
##  2 2013-01-01 05:00:00 Morning    Jan   Tue               5 LGA    IAH
##  3 2013-01-01 05:00:00 Morning    Jan   Tue               5 JFK    MIA
##  4 2013-01-01 05:00:00 Morning    Jan   Tue               5 JFK    BQN
##  5 2013-01-01 06:00:00 Morning    Jan   Tue               6 LGA    ATL
##  6 2013-01-01 05:00:00 Morning    Jan   Tue               5 EWR    ORD
##  7 2013-01-01 06:00:00 Morning    Jan   Tue               6 EWR    FLL
##  8 2013-01-01 06:00:00 Morning    Jan   Tue               6 LGA    IAD
##  9 2013-01-01 06:00:00 Morning    Jan   Tue               6 JFK    MCO
## 10 2013-01-01 06:00:00 Morning    Jan   Tue               6 LGA    ORD
## # ... with 327,336 more rows, and 5 more variables: carrier <chr>,
## #   dep_delay <dbl>, arr_delay <dbl>, air_time <dbl>, distance <dbl>
```

Use a boxplot to visualise the departure delays of less that 60 minutes by the three different time sections for each day of the week