CT5102: Programing for Data Analytics 2018/19

Assignment 6: Processing Data with dplyr (10 marks)

Based on the flights data set, perform the following tasks:

(1) Include the following libraries

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

(2) Create a local copy of the flights tibble

```
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>
```

(3) 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*

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

(3) Add columns for the day of the week and for the hour of the day (use `wday(time_hour)` and `hour(time_hour)` from **lubridate**.

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

(4) Average departure delay statistics by hour of day, ordered by delay.

```
> delay_hourly
# A tibble: 19 x 6
   HourOfDay AvrDepDelay    SD MinDelay MaxDelay MaxDelayHours
       <int>       <dbl> <dbl>    <dbl>    <dbl>         <dbl>
 1        19       24.7   52.7      -30     1137          19.0
 2        20       24.2   48.5      -33      878          14.6
 3        21       24.2   46.7      -43      800          13.3
 4        17       21.0   49.8      -27      896          14.9
 5        18       21.0   49.2      -22     1014          16.9
 6        22       18.7   40.5      -22      276           4.6
 7        16       18.6   46.8      -24     1126          18.8
 8        15       16.8   42.5      -22      483           8.05
 9        23       14.0   35.3      -18      245           4.08
10        14       13.7   40.6      -32      602          10.0
11        13       11.3   35.6      -20      533           8.88
12        12        8.52  33.0      -22      636          10.6
13        11        7.15  32.2      -20      437           7.28
14        10        6.45  33.6      -22      788          13.1
15         9        4.54  30.3      -24     1301          21.7
16         8        4.11  29.2      -22      911          15.2
17         7        1.91  23.4      -26      898          15.0
18         6        1.60  23.3      -21      786          13.1
19         5        0.689 15.9      -15      201           3.35
```

(5) Average departure delay statistics by month, ordered by delay.

```
> delay_monthly
# A tibble: 12 x 6
   Month AvrDepDelay    SD MinDelay MaxDelay MaxDelayHours
   <ord>       <dbl> <dbl>    <dbl>    <dbl>         <dbl>
 1 Jul         21.5   51.2      -22     1005          16.8
 2 Jun         20.7   51.3      -21     1137          19.0
 3 Dec         16.5   41.7      -43      896          14.9
 4 Apr         13.8   42.9      -21      960          16
 5 Mar         13.2   40.0      -25      911          15.2
 6 May         12.9   39.2      -24      878          14.6
 7 Aug         12.6   37.6      -26      520           8.67
 8 Feb         10.8   36.2      -33      853          14.2
 9 Jan          9.99  36.3      -30     1301          21.7
10 Sep          6.63  35.5      -24     1014          16.9
11 Oct          6.23  29.7      -25      702          11.7
12 Nov          5.42  27.6      -32      798          13.3
```

(6) Average departure delay statistics by carrier, ordered by delay.

```
> delay_carrier
# A tibble: 16 x 7
   carrier AvrDepDelay    SD MinDelay MaxDelay MaxDelayHours  NObs
   <chr>         <dbl> <dbl>    <dbl>    <dbl>         <dbl> <int>
 1 F9             20.2  58.4      -27      853          14.2    681
 2 EV             19.8  46.4      -32      548           9.13 51108
 3 YV             18.9  49.2      -16      387           6.45   544
 4 FL             18.6  52.5      -22      602          10.0   3175
 5 WN             17.7  43.2      -13      471           7.85 12044
 6 9E             16.4  45.5      -24      747          12.4  17294
 7 B6             13.0  38.4      -43      502           8.37 54049
 8 VX             12.8  44.0      -20      653          10.9   5116
 9 OO             12.6  43.1      -14      154           2.57    29
10 UA             12.0  35.5      -20      483           8.05 57782
11 MQ             10.4  39.0      -26     1137          19.0  25037
12 DL              9.22 39.7      -33      960          16    47658
13 AA              8.57 37.4      -24     1014          16.9  31947
14 AS              5.83 31.4      -21      225           3.75   709
15 HA              4.90 74.1      -16     1301          21.7    342
16 US              3.74 27.9      -19      500           8.33 19831
```

(7) Average departure delay statistics by airport by month, ordered by delay.

```
> delay_airport_month
# A tibble: 36 x 8
# Groups:   origin [3]
   origin Month AvrDepDelay    SD MinDelay MaxDelay MaxDelayHours
   <chr>  <ord>       <dbl> <dbl>    <dbl>    <dbl>         <dbl>
 1 EWR    Jan          14.9  40.8      -21     1126          18.8
 2 JFK    Jan           8.56 35.9      -17     1301          21.7
 3 LGA    Jan           5.61 29.6      -30      478           7.97
 4 EWR    Feb          13.0  37.1      -21      786          13.1
 5 JFK    Feb          11.7  37.3      -22      747          12.4
 6 LGA    Feb           6.92 33.3      -33      853          14.2
 7 EWR    Mar          18.1  44.1      -22      443           7.38
 8 JFK    Mar          10.7  35.2      -24      800          13.3
 9 LGA    Mar          10.2  39.6      -25      911          15.2
10 EWR    Apr          17.3  43.7      -21      545           9.08
# ... with 26 more rows, and 1 more variable: NObs <int>
```

(8) Average departure delay statistics by airport by hour, ordered by hour.

```
> delay_airport_time
# A tibble: 56 x 8
# Groups:   HourOfDay [19]
   HourOfDay origin AvrDepDelay    SD MinDelay MaxDelay MaxDelayHours
       <int> <chr>        <dbl> <dbl>    <dbl>    <dbl>         <dbl>
 1         5 EWR          0.656  15.8      -15      188          3.13
 2         5 JFK          0.505  16.1      -11      201          3.35
 3         5 LGA          1.23   15.6      -11      142          2.37
 4         6 EWR          3.44   26.8      -21      786         13.1
 5         6 JFK          1.12   20.6      -17      536          8.93
 6         6 LGA         -0.463  19.9      -18      419          6.98
 7         7 EWR          4.04   25.3      -21      382          6.37
 8         7 JFK          1.34   19.5      -16      364          6.07
 9         7 LGA         -0.123  24.3      -26      898         15.0
10         8 EWR          5.46   29.3      -18      502          8.37
# ... with 46 more rows, and 1 more variable: NObs <int>
```

(9) Add a new category, which divides each day into three sections (use case_when)

- Morning 5 <= time < 12
- Afternoon 12 <= time < 18
- Evening > =18

```
> select(my_flights,DaySection,everything())
# A tibble: 327,346 x 12
   DaySection time_hour           origin dest  carrier dep_delay
   <chr>      <dttm>              <chr>  <chr> <chr>       <dbl>
 1 Morning    2013-01-01 05:00:00 EWR    IAH   UA              2
 2 Morning    2013-01-01 05:00:00 LGA    IAH   UA              4
 3 Morning    2013-01-01 05:00:00 JFK    MIA   AA              2
 4 Morning    2013-01-01 05:00:00 JFK    BQN   B6             -1
 5 Morning    2013-01-01 06:00:00 LGA    ATL   DL             -6
 6 Morning    2013-01-01 05:00:00 EWR    ORD   UA             -4
 7 Morning    2013-01-01 06:00:00 EWR    FLL   B6             -5
 8 Morning    2013-01-01 06:00:00 LGA    IAD   EV             -3
 9 Morning    2013-01-01 06:00:00 JFK    MCO   B6             -3
10 Morning    2013-01-01 06:00:00 LGA    ORD   AA             -2
# ... with 327,336 more rows, and 6 more variables:
#   arr_delay <dbl>, air_time <dbl>, distance <dbl>, Month <ord>,
#   DayOfWeek <ord>, HourOfDay <int>
```

(10) Create a sample dataset (using sample_n()), and remove all departure delay values greater that 180 minutes.

```
set.seed(99)
> myf_sample
# A tibble: 9,875 x 12
   time_hour           origin dest  carrier dep_delay arr_delay
   <dttm>              <chr>  <chr> <chr>       <dbl>     <dbl>
 1 2013-05-04 14:00:00 LGA    PBI   DL              2       -11
 2 2013-10-12 18:00:00 LGA    MIA   AA              1       -31
 3 2013-06-09 18:00:00 JFK    MSP   9E             -1        20
 4 2013-09-28 08:00:00 EWR    LAX   UA             -7       -21
 5 2013-04-16 18:00:00 EWR    CLT   US             -8       -24
 6 2013-09-19 06:00:00 JFK    LAS   B6             -3        -9
 7 2013-06-05 07:00:00 JFK    SLC   DL             -3         8
 8 2013-12-17 14:00:00 JFK    MSY   B6            112       118
 9 2013-02-11 21:00:00 LGA    BOS   US             -4       -11
10 2013-11-03 17:00:00 EWR    MSP   EV             95       121
# ... with 9,865 more rows, and 6 more variables: air_time <dbl>,
#   distance <dbl>, Month <ord>, DayOfWeek <ord>, HourOfDay <int>,
#   DaySection <chr>
```

(11) Use a boxplot to visualise the departure delay by the three different time sections.