# MATH 216: Assignment 3

Solutions

## Problems

### Question 1

Worth 2 points. Students may have created not have printed the column for route, but it should be apparent from their code that they have created it using 'unite', 'paste', or similar.

Create a new column called `route` which contains the both the origin and destination airport codes seperated by a dash. For example, a flight that goes from JFK to MIA should display `JFK-MIA` in this new column.

```
tblflights %>%
  unite(route, c(origin, dest), sep="-") %>%
  select(year, month, day, dep_time, route) %>% #optional
  head(10) #optional
```

```
## # A tibble: 10 x 5
##     year month   day dep_time route
##    <int> <int> <int>    <int> <chr>
##  1  2013     1     1      517 EWR-IAH
##  2  2013     1     1      533 LGA-IAH
##  3  2013     1     1      542 JFK-MIA
##  4  2013     1     1      544 JFK-BQN
##  5  2013     1     1      554 LGA-ATL
##  6  2013     1     1      554 EWR-ORD
##  7  2013     1     1      555 EWR-FLL
##  8  2013     1     1      557 LGA-IAD
##  9  2013     1     1      557 JFK-MCO
## 10  2013     1     1      558 LGA-ORD
```

```
## OR
# tblflights$route <- paste(tblflights$origin, "-", tblflights$dest)
```

## Question 2

1 point for getting the answer 16. 2 points for printing out the table.

How many different airlines are there?

```
length(unique(tblflights$carrier))
```

```
## [1] 16
```

```
#OR
#length(levels(as.factor(tblflights$carrier)))

## There are lots of different ways to do this
```

Determine the mean departure delay for each of these different airlines. Note: the var `dep_delay` is the arrival delay in minutes.

```
tblflights %>%
  group_by(carrier) %>%
  summarize(mean.delay=mean(dep_delay, na.rm=T)) #na.rm=T is iport
```

```
## # A tibble: 16 x 2
##    carrier mean.delay
##    <chr>        <dbl>
##  1 9E           16.7
##  2 AA            8.59
##  3 AS            5.80
##  4 B6           13.0
##  5 DL            9.26
##  6 EV           20.0
##  7 F9           20.2
##  8 FL           18.7
##  9 HA            4.90
## 10 MQ           10.6
## 11 OO           12.6
## 12 UA           12.1
## 13 US            3.78
## 14 VX           12.9
## 15 WN           17.7
## 16 YV           19.0
```

Whoops! no question 3...

## Question 4

2 points for a the appropriate table.

Print a list of the 10 flights with the longest departure delays. Your list should include the date, schedule departure time, actual departure time, carrier, the departure airport and destination airport.

Optional: include a meme of how you would feel if you were on one of these 10 flights

```
tblflights %>%
  arrange(-dep_delay) %>%
  head(10) %>%
  select(year, month, day, sched_dep_time, dep_time, dep_delay, carrier, origin, dest)
```

```
## # A tibble: 10 x 9
##     year month   day sched_dep_time dep_time dep_delay carrier origin dest
##    <int> <int> <int>          <int>    <int>     <dbl> <chr>   <chr>  <chr>
##  1  2013     1     9            900      641      1301 HA      JFK    HNL
##  2  2013     6    15           1935     1432      1137 MQ      JFK    CMH
##  3  2013     1    10           1635     1121      1126 MQ      EWR    ORD
##  4  2013     9    20           1845     1139      1014 AA      JFK    SFO
##  5  2013     7    22           1600      845      1005 MQ      JFK    CVG
##  6  2013     4    10           1900     1100       960 DL      JFK    TPA
##  7  2013     3    17            810     2321       911 DL      LGA    MSP
##  8  2013     6    27           1900      959       899 DL      JFK    PDX
##  9  2013     7    22            759     2257       898 DL      LGA    ATL
## 10  2013    12     5           1700      756       896 AA      EWR    MIA
```

## Question 5

Worth 3 points. Take off marks for not including a title, or x and y axis labels. Students may have chosen to add a limit to the x-axis to better display the data.

Create two histograms - one for the departure delays and one for the arrival delays. Make sure they are well-labelled (have a title, x and y axis labels, etc.)

```
library(gridExtra)
```

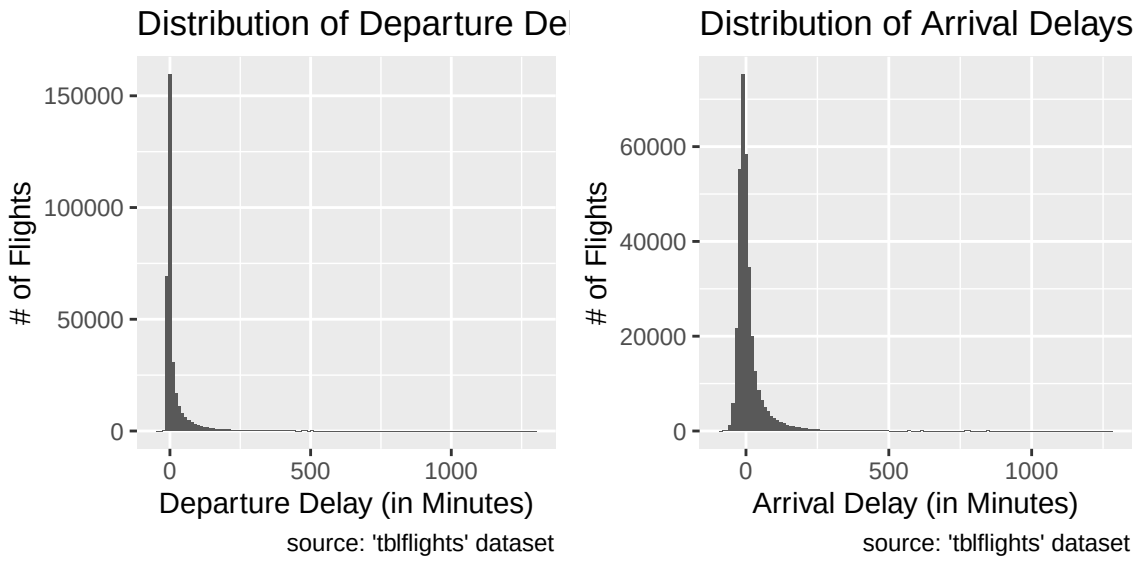
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
plot1 <- tblflights %>%
  ggplot(aes(dep_delay)) +
  geom_histogram(binwidth = 11) +
  labs(title = "Distribution of Departure Delays",
         caption = "source: 'tblflights' dataset",
         x = "Departure Delay (in Minutes)",
         y = "# of Flights")

plot2 <- tblflights %>%
  ggplot(aes(arr_delay)) +
  geom_histogram(binwidth = 11) +
  labs(title = "Distribution of Arrival Delays",
         caption = "source: 'tblflights' dataset",
         x = "Arrival Delay (in Minutes)",
         y = "# of Flights")

grid.arrange(plot1, plot2, ncol=2)
```



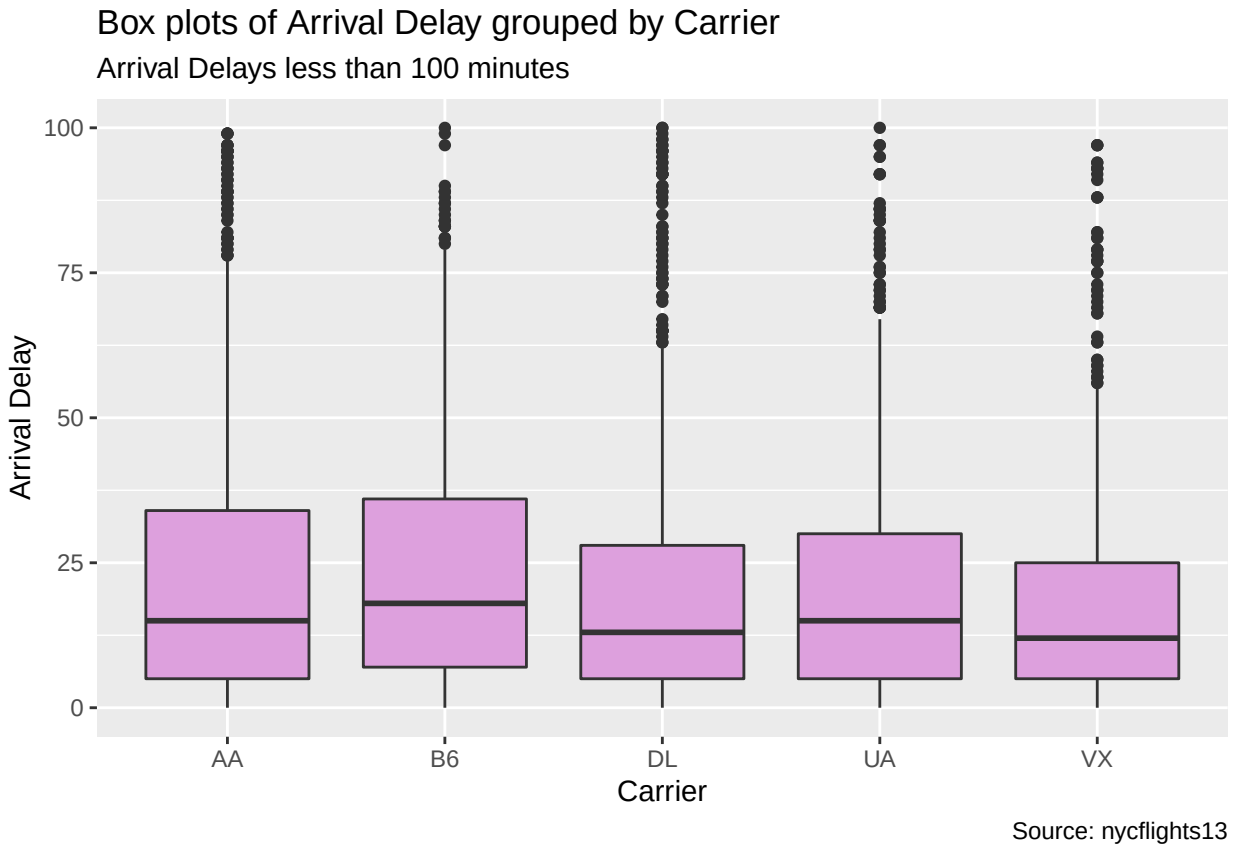
## Question 6

3 points. Take off marks for not including a title, or x and y axis labels. Students may have chosen transform the y axis or to subset the y-axis to better see the comparison. Both are okay with me.

For all the flights JFK - LAX, create a side-by-side boxplot to compare the arrival delay across all carriers. Make sure they are well-labelled (have a title, x and y axis labels, etc.)

```
JFKLAX <- tblflights %>%
  filter(origin == "JFK", dest =="LAX") %>%
  mutate(arr_delay_trans = arr_delay^(1/3))

ggplot(JFKLAX, aes(carrier, arr_delay)) +
  geom_boxplot( fill="plum") +
  labs(title="Box plots of Arrival Delay grouped by Carrier",
         subtitle="Arrival Delays less than 100 minutes",
         caption="Source: nycflights13",
         x="Carrier",
         y="Arrival Delay") +
  ylim(c(0,100))
```



## Question 7

Worth 3 points.

Create a table called carrier_flights_over_time that displays the count of the number of flights for all carriers across all months in wide format.

Hint: the first column should be all the carriers, the first row should be all the months. The data inside the table should be the number of flights.

```
carrier_flights_over_time <- tblflights %>%
  group_by(carrier, month) %>%
  summarize(num_flights = n()) %>%
  spread(month, num_flights)

carrier_flights_over_time
```

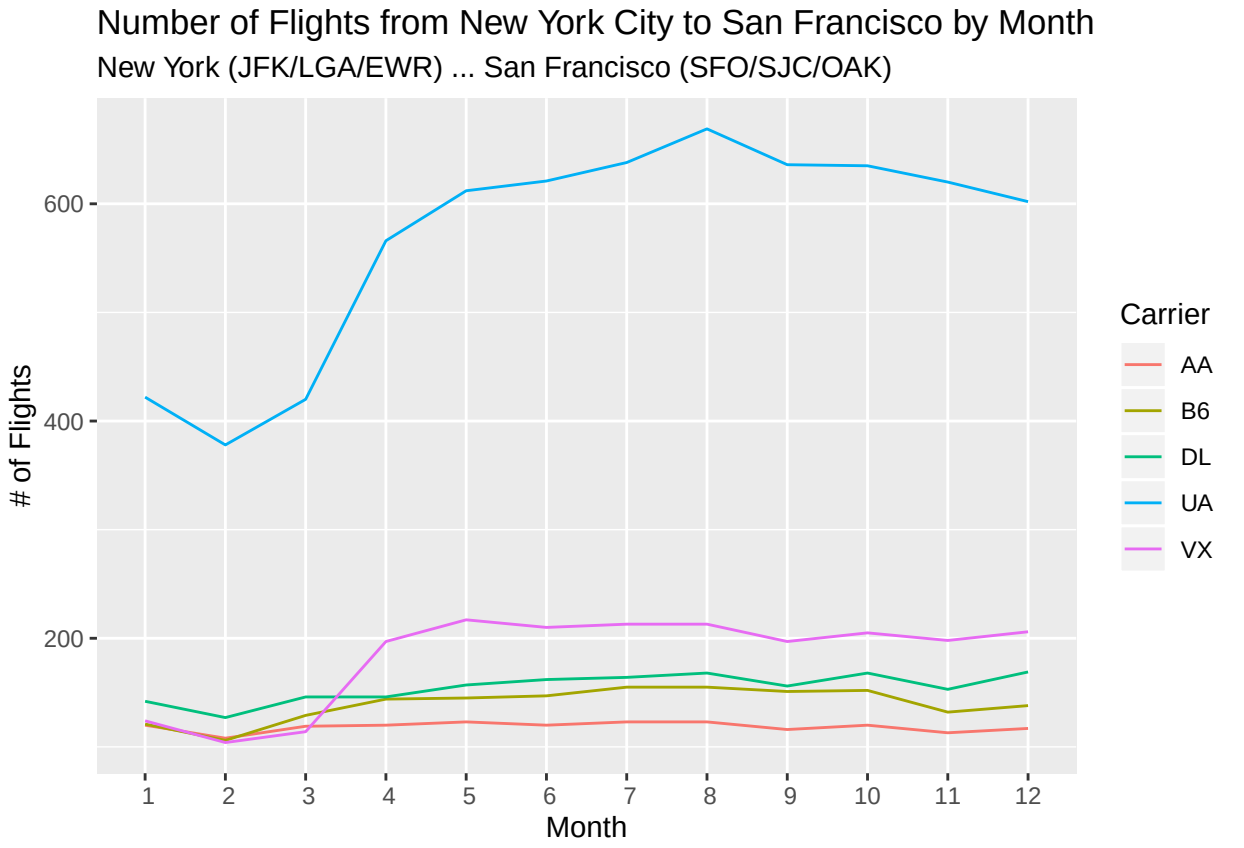
```
## # A tibble: 16 x 13
## # Groups:   carrier [16]
##    carrier   `1`   `2`   `3`   `4`   `5`   `6`   `7`   `8`   `9`  `10`
##    <chr>   <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
##  1 9E       1573  1459  1627  1511  1462  1437  1494  1456  1540  1673
##  2 AA       2794  2517  2787  2722  2803  2757  2882  2856  2614  2715
##  3 AS         62    56    62    60    62    60    62    62    60    62
##  4 B6       4427  4103  4772  4517  4576  4622  4984  4952  4291  4361
##  5 DL       3690  3444  4189  4092  4082  4126  4251  4318  3883  4093
##  6 EV       4171  3827  4726  4561  4817  4456  4641  4563  4725  4908
##  7 F9         59    49    57    57    58    55    58    55    58    57
##  8 FL        328   296   316   311   325   252   263   263   255   236
##  9 HA         31    28    31    30    31    30    31    31    25    21
## 10 MQ       2271  2044  2256  2211  2284  2178  2261  2263  2206  2228
## 11 OO          1    NA    NA    NA    NA     2    NA     4    20    NA
## 12 UA       4637  4346  4971  5047  4960  4975  5066  5124  4694  5060
## 13 US       1602  1552  1721  1727  1785  1736  1786  1779  1698  1846
## 14 VX        316   271   303   466   496   480   489   489   453   472
## 15 WN        996   911   998   980  1006  1028  1076  1047  1010  1091
## 16 YV         46    48    18    38    49    49    81    65    42    66
## # ... with 2 more variables: `11` <int>, `12` <int>
```

### Question 8: Challenge

Worth 3 points. Take off marks for not including a title, or x and y axis labels

Consider all the flights going from anywhere in the NYC area to anywhere in the NYC area (JFK, LGA or EWR) to anywhere in the San Fransisco Bay area (SFO, SJC or OAK). Sum how many flights were ran by each airline in each month and create a line plot to show the the number of flights varies over time for each airline. Make sure your plot is well-labelled (have a title, x and y axis labels, etc.)

```
tblflights %>%
  filter((origin == "JFK" | origin == "LGA" | origin == "EWR") & (dest == "SFO" | dest == "SJC" | dest =
  group_by(carrier, month) %>%
  summarize(num_flights = n()) %>%
  ggplot(aes(x = month, y = num_flights, group = carrier, color = carrier)) +
  geom_line() +
  labs(title = "Number of Flights from New York City to San Francisco by Month",
         subtitle = "New York (JFK/LGA/EWR) \U2192 San Francisco (SFO/SJC/OAK)",
         x = "Month",
         y = "# of Flights",
         color = "Carrier") + # color='Carrier' meant to change legend title
  scale_x_discrete(limits=c(1:12))
```




## Acknowledgements

Worth 1 mark. Make sure the students include at least one acknowledgement.