Exam 1, STAT 450

Due: Friday, October 11

Directions:

- This exam should be completed using Quarto and submitted to Canvas as self-contained HTML or PDF file.
- Your solutions to this exam must be your own work.
- Make sure your Quarto document is well-formatted: label each exercise with a header, use separate code chunks for your answers to each exercise, and any written analysis should be formatted as plain text outside of the code chunks. Points may be deducted for poor formatting.

First, load the following R packages:

```
library(tidyverse)
library(nycflights13)
```

All questions use the `flights` data set.

```
flights
```

```
# A tibble: 336,776 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1      517            515         2      830            819
 2  2013     1     1      533            529         4      850            830
 3  2013     1     1      542            540         2      923            850
 4  2013     1     1      544            545        -1     1004           1022
 5  2013     1     1      554            600        -6      812            837
 6  2013     1     1      554            558        -4      740            728
 7  2013     1     1      555            600        -5      913            854
 8  2013     1     1      557            600        -3      709            723
 9  2013     1     1      557            600        -3      838            846
10  2013     1     1      558            600        -2      753            745
# i 336,766 more rows
# i 11 more variables: 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>
```

Type `help(flights)` to read the documentation on this data set in the help menu.

Question 1 (30 points)

Use `filter()` to find all flights that

(a) Flew to San Francisco International Airport (SFO).

(b) Departed during the summer months (June, July, August).

(c) Departed from LaGuardia Airport (LGA) on October 31, 2013.

(d) Were operated by Southwest Airlines (WN), and had departure delays that were 30 or more minutes.

(e) Were operated by Southwest Airlines (WN), and have missing values for the departure time. What do these rows represent?

Question 2 (25 points)

Use `group_by()` and `summarize()` to create a data frame with the following columns:

- Count of the number of flights for each carrier.
- Mean departure delay for each carrier.
- Mean arrival delay for each carrier.
- Count of the number of canceled flights for each carrier.

Your code should recreate the following table:

```
# A tibble: 16 x 5
   carrier count dep_delay_mean arr_delay_mean canceled
   <chr>   <int>          <dbl>          <dbl>    <int>
 1 9E      18460          16.7            7.38      1044
 2 AA      32729           8.59           0.364      636
 3 AS        714           5.80          -9.93         2
 4 B6      54635          13.0            9.46       466
 5 DL      48110           9.26           1.64       349
 6 EV      54173          20.0           15.8       2817
 7 F9        685          20.2           21.9          3
 8 FL       3260          18.7           20.1          73
 9 HA        342           4.90          -6.92         0
10 MQ      26397          10.6           10.8       1234
11 OO         32          12.6           11.9          3
12 UA      58665          12.1            3.56       686
13 US      20536           3.78           2.13       663
14 VX       5162          12.9            1.76        31
15 WN      12275          17.7            9.65       192
16 YV        601          19.0           15.6         56
```

Question 3 (10 points)

Refer to the data frame of grouped summary statistics that you created in Question 2. Use `ggplot2` to make a scatter plot that shows the relationship between the mean departure delay and arrival delay for the different airline carriers. Map either the color or size of the points to the number of canceled flights. Write 2-3 sentences providing your interpretation of this plot. Bonus points may be awarded for including the labels for the different carriers in the scatter plot (Hint: use `geom_text()` and adjust the position of the labels so they don't overlap with the points).

Question 4 (10 points)

(a) Which carriers had the greatest number of flights departing from New York City?

(b) Which carriers had the longest average arrival delays?

(c) Which carriers had the shortest average arrival delays? What do negative values represent?

(d) Which carriers had the greatest number of canceled flights?

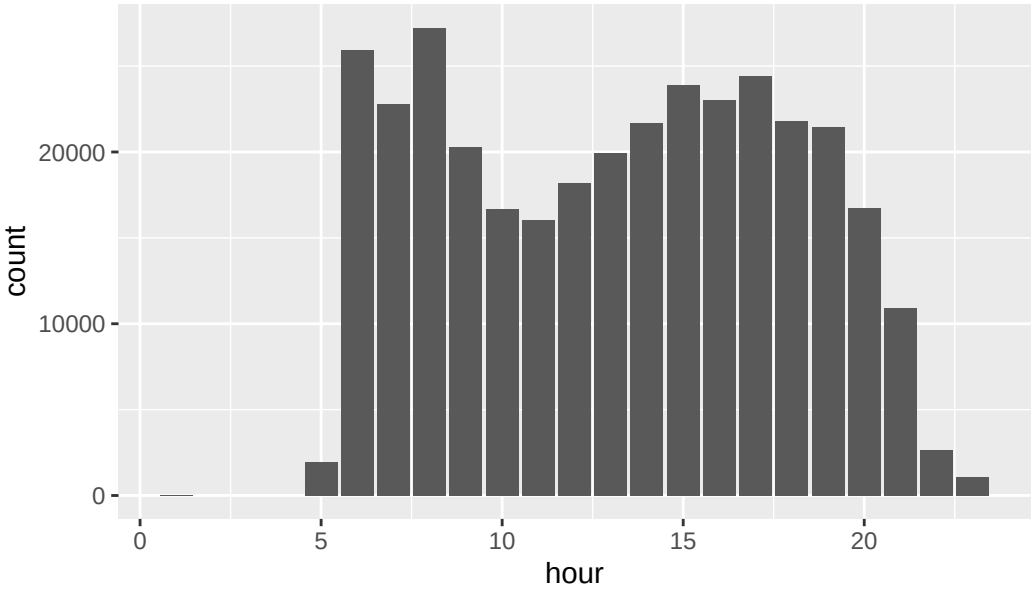
Hint: Use the pipe with `arrange()`

Question 5 (25 points)

Recreate the R code necessary to make the following graphs. In your submission, show both the R code and the graphs.

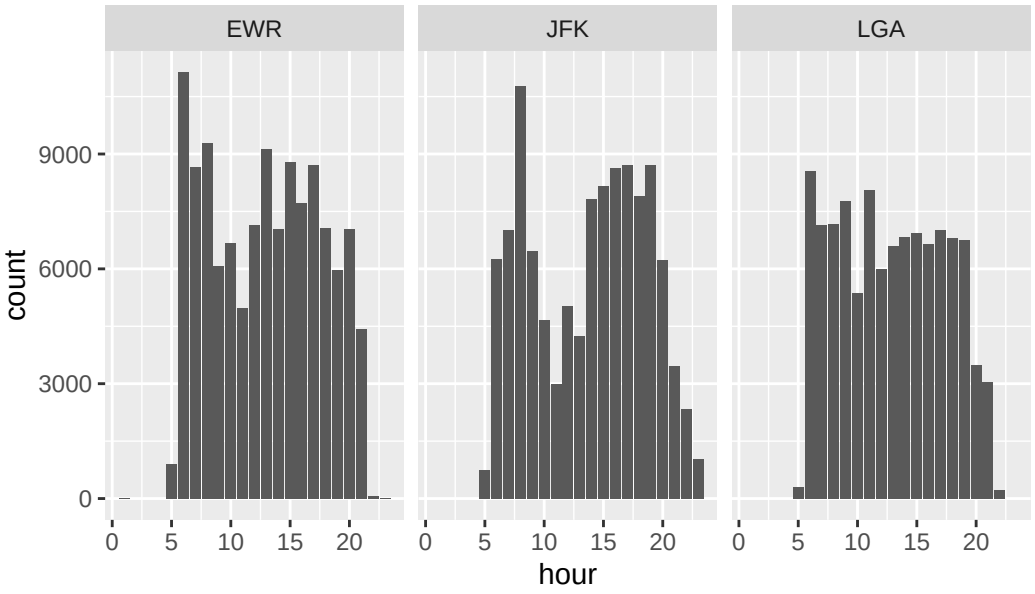
a

Bar plot of number of flights departing each hour of the day.



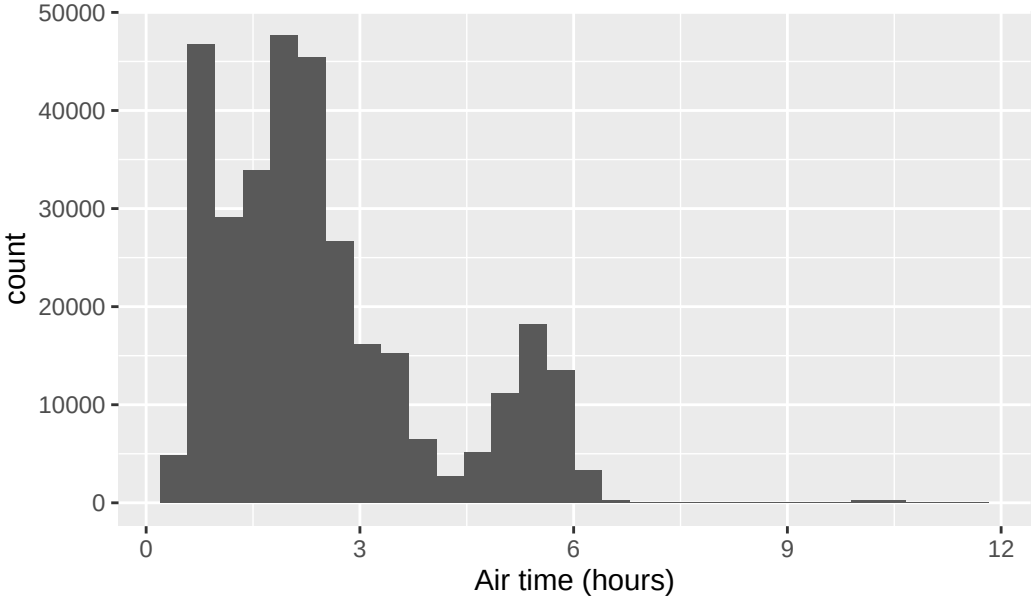
b

For each origin airport, bar plot of the number of flights departing each hour of the day.



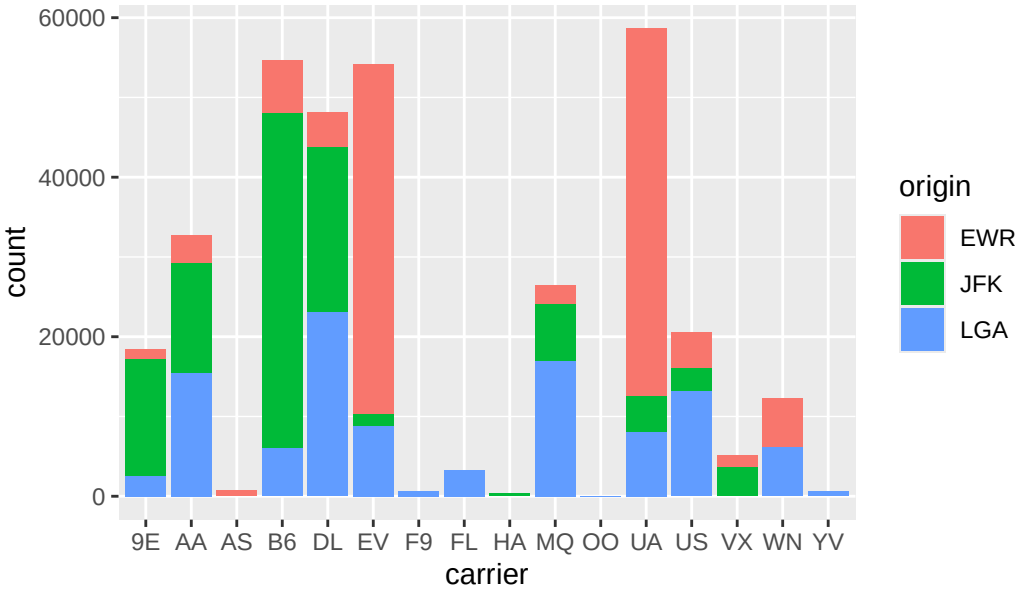
c

Histogram of air time (in hours).



d

Stacked bar plot of number of flights for each carrier, with fill color corresponding to origin airport.



e

Stacked bar plot that shows proportions instead of counts.