Homework 2: Due Friday, October 25 at 11:59pm

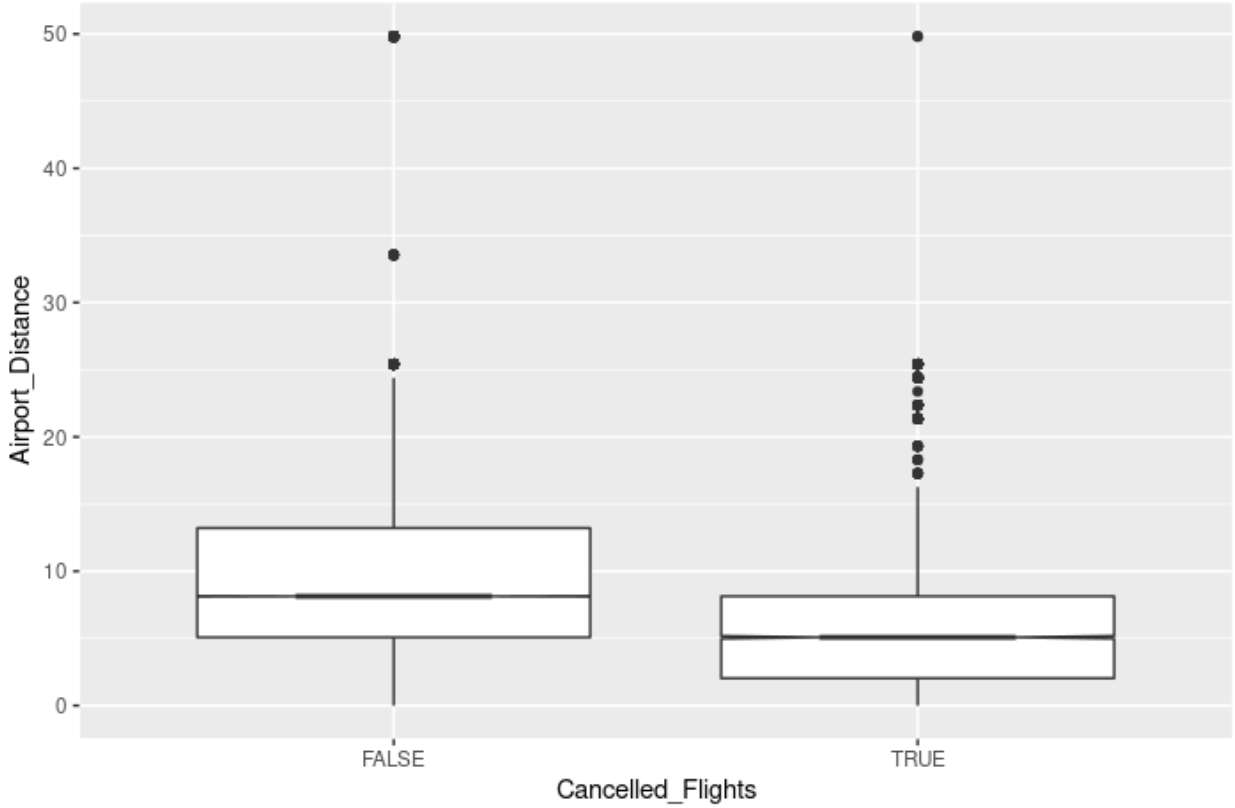
For this assignment you will be analyzing the **`flights`** data frame included in the **`nycflights13`** package, which contains 19 features for 336,776 flights that departed from New York City in 2013. The goal of this assignment is to become more familiar with data transformations and exploratory data analysis. Provide all responses in the designated spaces in this Word document, then save it as a pdf and upload it to Canvas.

1. **[25%]** Generate notched boxplots to compare distributions of airport distances between canceled and non-canceled flights. Is the median airport distance for canceled flights shorter, longer, or roughly the same as for non-canceled flights? Justify your answer, and then provide a possible explanation for this finding.

   Provide code below:

```
flights %>%mutate(cancelled = is.na(dep_time),
hour=distance%/%100,
min=distance%/%100,
distance=hour+min/60)%>%
  ggplot(mapping=aes(x=Cancelled_Flights,y=Airport_Distance))+
  geom_boxplot(mapping=aes(x=cancelled, y=distance), size=.5,
  notch=TRUE)
```

   Provide figure below:



Provide answers to questions below:

The median airport distance for cancelled flights seems to be longer than non-cancelled flights. The density of the cancelled flights box plot occurs at a greater distance such as 8-10, whereas for non-cancelled is more around 5-8.

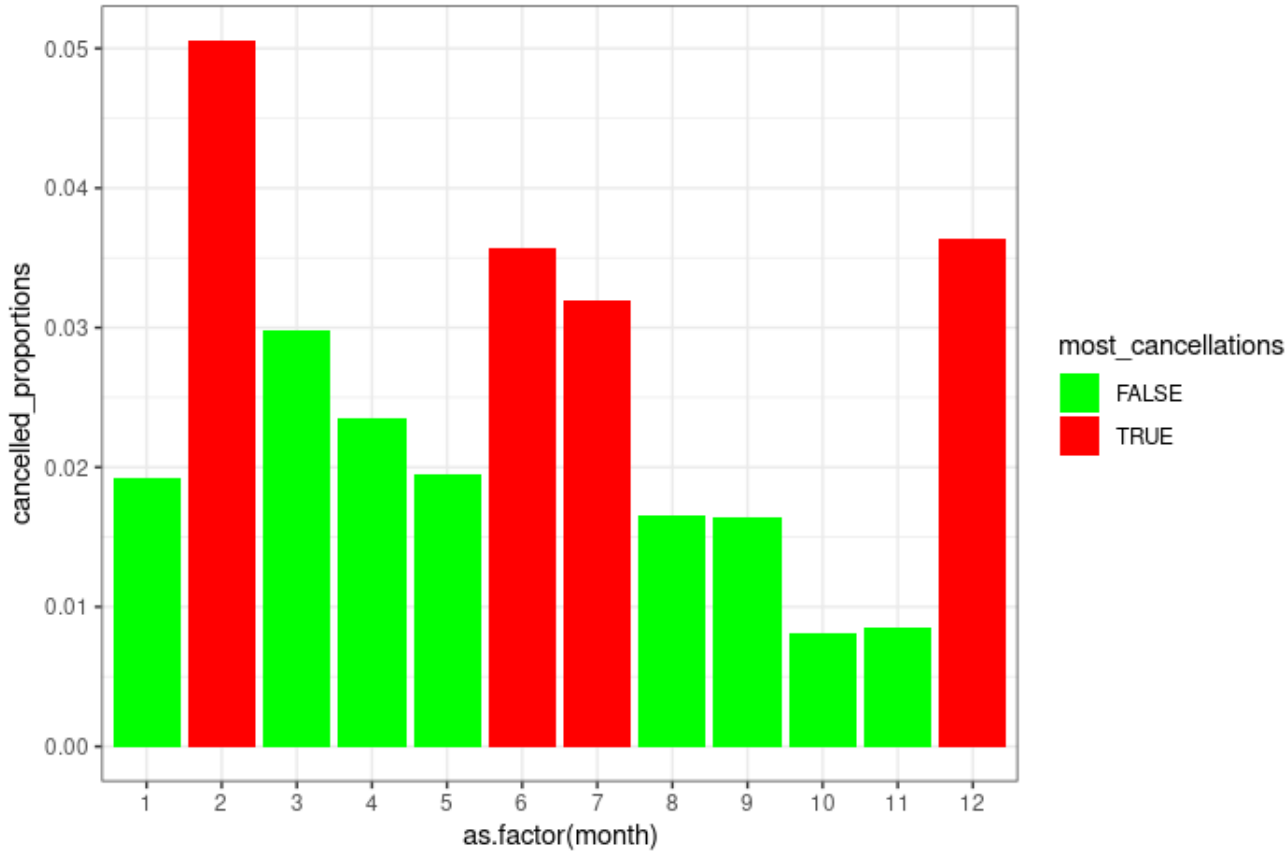
2. **[40%]** Generate a bar plot with month on the x axis and the proportion of the month's flights that are canceled on the y axis. Which *four* months of the year had the highest proportions of flight cancelations? Provide a possible explanation for this finding.

   Note: You will need to use the function **`geom_col()`** to generate a bar plot for which you provide the x and y axis features. Because **`geom_col()`** expects that the feature on the x axis is categorical, you must also use the code **`as.factor(month)`** to convert the **`month`** feature to a categorical variable taking 12 values (1, 2, …, 12).

Provide code below:

```
flights %>% group_by(month)
summarise(by_month,cancelled_proportions=mean(is.na(dep_time))) %>%
mutate(most_cancellations = ifelse(cancelled_proportions>0.03, T, F)) %>%
ggplot(mapping = aes(x=as.factor(month), y=cancelled_proportions)) +
  geom_col(aes(fill = most_cancellations))+theme_bw()+
  scale_fill_manual(values = c('green', 'red'))
```

Provide figure below:



Provide answers to questions below:

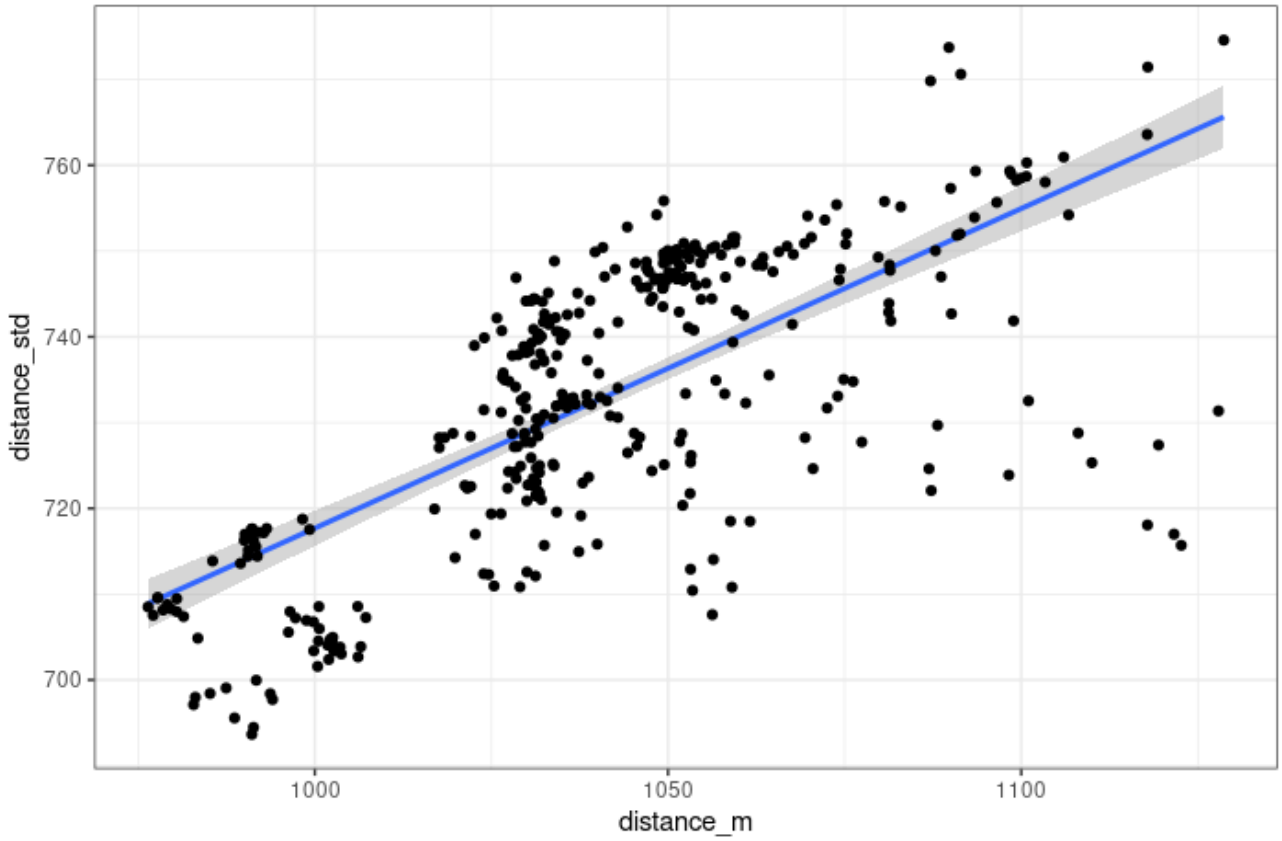
The four months that had the highest proportions of flight cancellation was February, December, June and July. The most likely reason for high cancellations is weather in the region. Winter months bring lots of snowfall, snow storms and blizzards that grounds many planes. Summer months bring about lots of rain, wind and thunderstorms. Could also be holidays.

3. **[35%]** Generate a scatterplot displaying the relationship between the mean and the standard deviation of airport distance for flights on each of the 365 days of the year. Use the **`method`** argument of the **`geom_smooth()`** function to overlay your scatterplot with a fitted straight line with confidence intervals. Is there a relationship between the mean and standard deviation of airport distance for flights on each day of the year? Describe the relationship (or lack of one), and then provide a possible explanation for this finding.

Provide code below:

```
flights %>% group_by(month, day) %>%
  summarise(distance_std=sd(distance),distance_m=mean(distance)) %>%
  ggplot(mapping=aes(x=distance_m,y=distance_std))+
  geom_smooth(method=lm)+geom_point()+theme_bw()
```

Provide figure below:



<u>Provide answers to questions below</u>:

There is a slight relationship, the mean shows the middle of the dataset and the standard deviation shows the average distance between actual data and mean. There seems to be a dense area in the middle of the scatterplot where the confidence interval is also tighter meaning the data is precise, so perhaps little variance between data.