STAT 630: Homework 3

Due: September 22nd, 2022

Concept Questions

The following questions relate to the `nycflights` dataset found in the `openintro` package. *Note, you may have to coerce some of the columns to factors before your start.*

Find more information on the dataset here

```
library(openintro)
library(dplyr)
library(ggplot2)
data("nycflights")
?nycflights # Description of the dataset (comment out before knitting)
```

For as many questions as you can, practice using inline R code to display your answers. For example:

The number of flights in the dataset is $n = 32735$.

I will put a * next to questions you might want to try this with.

1. Pick an airline carrier (see link above for carrier codes) and a month of the year (1-12). Input your choices into the code below. The R chunk will create a histogram of `arr_delay` (arrival delay in minutes) for the airline and month you chose. Then, using `stat_function` ggplot2 will overlay a true normal distribution with the mean and standard deviation of `arr_delay` for your carrier/month.

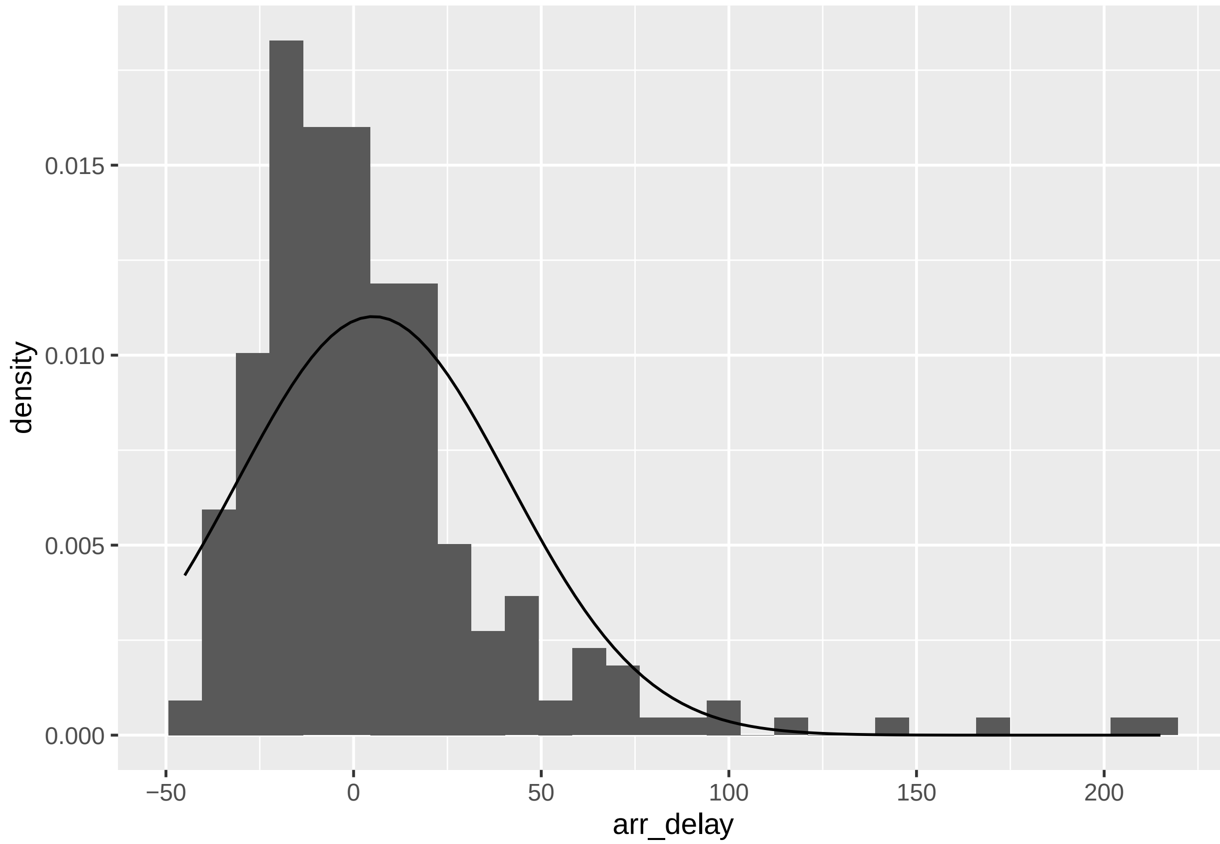
```
# Subset the data according to your chosen/carrier/month
my_flights <- nycflights %>%
  filter(carrier == "AA", month == 12)

# Calculate mean/sd of arr_delay
arr_delay_stats <- my_flights %>%
  summarise(x_bar = mean(arr_delay),
            s = sd(arr_delay))

# Plot histogram of arr_delay, overlaying true normal curve
my_flights %>%
  ggplot(aes(arr_delay)) +
  geom_histogram(aes(y = ..density..)) + # relative frequency
  stat_function(fun = dnorm, # overlay a stat function (a normal dist)
                args = list(mean = arr_delay_stats$x_bar, # let mu = x_bar
                            sd = arr_delay_stats$s)) # let sigma = s
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. Comment on the normality of `arr_delay` compared to the true normal curve.

The normality of arr_delay is not a normal distribution. This is a **right-skewed distribution**.

For questions 3 & 4, assume the distribution of `arr_delay` for your chosen carrier "AA" and month "12" is normally distributed with the mean and standard deviation found in Question 1.

3. *Using R functions, find the probability that the flight on [AA] during [12] was early.

The probability that the flight on [AA] during [12] was early means that the arr_delay < 0

Firstly, we calculate the z-score when the observed value = 0: -0.145

Then, we calculate the probability that the flight on [AA] during [12] was early P(X<0): 0.442184.

```
my_flights <- nycflights %>%
  filter(carrier == "AA", month == 12)
z_0 = (0 - arr_delay_stats$x_bar)/arr_delay_stats$s
pnorm(z_0)
```

```
## [1] 0.442184
```

4. *Using R functions, calculate the probability that a randomly selected flight on [AA] in [12] is between 30 to 60 minutes late.

z-score when X = 60: **1.51149**

The probability that a randomly selected flight on [AA] in [12] is 60 minutes late

$P(X \leq 60) = pnorm(z\_60) = 0.9346691$

z-score when X = 30: 0.683

The probability that a randomly selected flight on [AA] in [12] is 30 minutes late

$P(X \leq 30) = pnorm(z\_30) = 0.7527065$

So, the probability that a randomly selected flight on [AA] in [12] is between 30 to 60 minutes late.

$P(30 \leq X \leq 60) = P(X \leq 60) - P(X \leq 30) = pnorm(z\_60) - pnorm(z\_30) = 0.1819626$

```
z_60 = (60 - arr_delay_stats$x_bar)/arr_delay_stats$s
pnorm(z_60)
```

```
## [1] 0.9346691
```

```
z_30 = (30 - arr_delay_stats$x_bar)/arr_delay_stats$s
pnorm(z_30)
```

```
## [1] 0.7527065
```

```
pnorm(z_60)-pnorm(z_30)
```

```
## [1] 0.1819626
```

5. *Use`quantile()` to calculate the actual 95th percentile (0.95 quantile) from the `arr_delay` data on [AA] in [12]. Interpret this value in the context of the problem.

The actual 95th percentile (0.95 quantile) from the `arr_delay` data on [AA] in [12]: 67.7. It means that 95% of the random selected flights of AA in December was not late than 67.7 minutes.

```
quantile(my_flights$arr_delay, 0.95)
```

```
##  95%
## 67.7
```

```
?quantile
```

6. *Now, use `qnorm()` to find the top 5% according to the *true* normal distribution with the same mean and standard deviation as you found in Question 1. Then, explain how this compares to your answer in question 5? Why are they different or the same?

The top 5% according to the *true* normal distribution with the same mean and standard deviation: **64.82905**

This result is different from than that of question 5.

I think the actual 95th percentile is specific to the dataset and reflects the position of a value within your actual data. While **qnorm(0.05, mean, sd, lower.tail = FALSE)** provides a value based on a theoretical normal distribution, representing the point above which 5% of the data would fall in a normal distribution. These two values may or may not be the same, depending on the nature of your data and how well it approximates a normal distribution. So, for this case, which is not a normal distribution we should apply to use **quantile()** to calculate the actual 95th percentile (0.95 quantile).

```
qnorm(0.05, arr_delay_stats$x_bar, arr_delay_stats$s,  lower.tail = FALSE)
```

```
## [1] 64.82905
```

```
?nycflights
```

Exploratory Data Analysis

Research Question: Do flights that are longer in distance have more arrival delays? *Note: for the remaining questions, please use the full dataset `nycflights`.

7. What is the response variable and what is the predictor of interest? Specify the type of each variable: quantitative (discrete or continuous) or categorical (ordinal or nominal).

Response Variables:

Arrival Delay: This is likely to be a continuous quantitative variable, representing the amount of delay in minutes. It's a measure of how late or early a flight arrives.

Departure Delay: Similar to arrival delay, this is likely a continuous quantitative variable representing the amount of delay in minutes at departure.

Distance: The distance the flight covers. This is a continuous quantitative variable, typically measured in miles.

Airline Carrier: This could be a categorical nominal variable representing the airline carrier (e.g., "Delta", "American Airlines", etc.).

Predictor Variables:

Month: This is a categorical ordinal variable, as it represents different months of the year. Although represented as numbers (1 for January, 2 for February, etc.), it's not a continuous variable because there is no inherent "distance" between the months.

Day of the Month: Similar to month, this is a categorical ordinal variable.

Time of Day: This could be a categorical variable if you divide the day into categories like "morning", "afternoon", "evening", etc.

Origin: The airport from which the flight departs. This is a categorical nominal variable.

Flight Number: This is a categorical nominal variable.

8. What are two potential confounding variables that could influence the relationship between departure delays and arrival delays. At least one should be from the dataset. Cite any sources used. Explain how the variable could be associated with the predictor of interest *and* the response.

These two variables, weather conditions and distance, can potentially confound the relationship between departure delays and arrival delays. They are associated with both the predictor (departure delays) and the response (month).

Weather Conditions: Weather conditions at both the departure and arrival airports can significantly impact flight delays. Adverse weather such as thunderstorms, fog, or snow can lead to both departure and arrival delays. Poor visibility or strong winds can affect the efficiency of take-offs and landings, potentially leading to longer delays. Additionally, weather conditions along the flight path can also influence the overall travel time. In addition, every season has different weather conditions. For example, in winter, from November to February, it is snowy in somewhere, so it is possible to handle arrival and departure delay at airports.

Source: Rodríguez-Sanz, Á., Cano, J., & Fernández, B. R. (2021). Impact of weather conditions on airport arrival delay and throughput. IOP Conference Series, 1024(1), 012107. https://doi.org/10.1088/1757-899x/1024/1/012107

Distance: Longer distances typically involve longer flight times. Airlines may allocate more time for flights with longer distances to account for the travel duration. This can potentially lead to longer departure delays as they might wait for passengers or complete other pre-flight tasks. Longer distances can result in extended travel times. This can be due to factors like air traffic congestion, weather conditions, or routing adjustments. These factors can contribute to increased arrival delays.

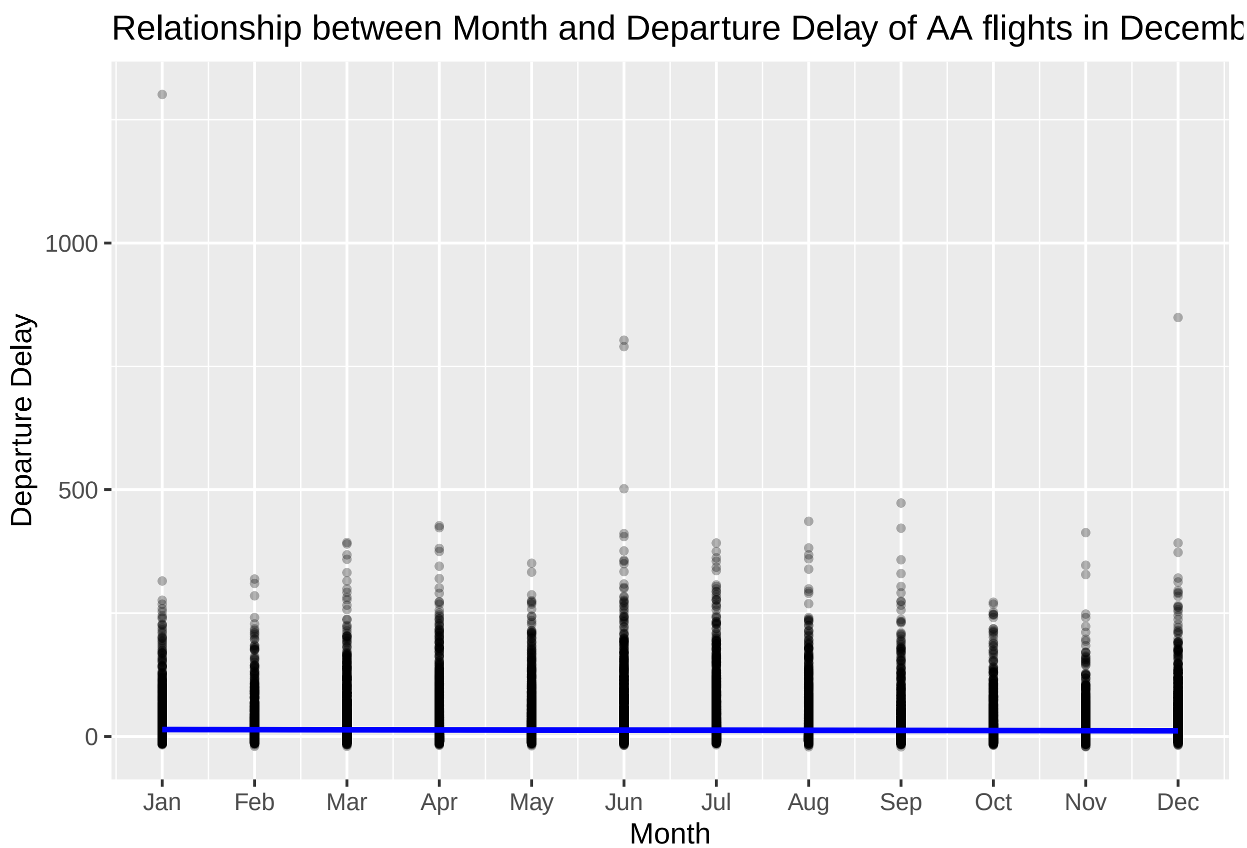
Source: Mitsokapas, E., Schäfer, B., Harris, R. J., & Beck, C. (2021). Statistical characterization of airplane delays. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-87279-8

9. Make a well-labeled plot (using your choice of R package), to display the relationship between the predictor of interest and the response. "

```r
month_names <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                 "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")


ggplot(nycflights, aes(x = month, y = dep_delay)) +
  geom_point(alpha = 0.3, size = 1) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Relationship between Month and Departure Delay of AA flights in December 2013",
       x = "Month",
       y = "Departure Delay") +
  scale_x_continuous(breaks = 1:12, labels = month_names)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



10. What do you notice about the relationship?

As we can see on the point plot, in general, the departure delay for 12 months in 2013 was around from 1 to 250 minutes. In summer, from Jun to Sep, the departure delay situation tended to be longer. There were some flights whose delay time was up to more than 750 minutes. In addition, December was also the time when it had long departure delay flights. So, it can be seen that month can possibly affect to departure delay.

11. Adjust the plot you made in Question 9 (use a new aesthetic) to also include one of the confounding variables you mentioned in Question 8. Comment on any patterns.

```r
ggplot(nycflights, aes(x = month, y = dep_delay, color = distance)) +
  geom_point(size = 2, alpha = 0.6) +
  labs(title = "Relationship between Month, Departure Delay and Distance of AA in December 2013",
       x = "Month",
```

```
       y = "Departure Delay",
       color = "Distance") +
  scale_x_continuous(breaks = 1:12, labels = month_names) +
  scale_color_gradient(low = "blue", high = "red")
```