problem2

Hamda Hassan

```r
library(resampledata3)
```

```
##
## Attaching package: 'resampledata3'

## The following object is masked from 'package:datasets':
##
##     Titanic
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)
library(ggplot2)
library(nycflights13)
library(tibble)
```

Question: What are the five most common destination airports for United Airlines flights from New York City? Describe the distribution and the average gain for each of these five airports.

Part one: Figure out the five most common destinations

The most common airports fro the United Airlines are ORD, IAH, SFO, LAX, and DEN

```r
# Filter for United Airlines flights (carrier == "UA")
ua_flights <- flights %>%
  filter(carrier == "UA")
```

```r
# Count destinations and select the top 5
top5_dests <- ua_flights %>%
  count(dest, sort = TRUE) %>%
  slice_max(n, n = 5)

top5_dests_named <- top5_dests %>%
  left_join(airports, by = c("dest" = "faa")) %>%
  select(dest, name, n)

top5_dests_named
```

```
## # A tibble: 5 x 3
##   dest  name                            n
##   <chr> <chr>                       <int>
## 1 ORD   Chicago Ohare Intl           6984
## 2 IAH   George Bush Intercontinental  6924
## 3 SFO   San Francisco Intl           6819
## 4 LAX   Los Angeles Intl             5823
## 5 DEN   Denver Intl                  3796
```

Part Two: Describe the distribution and the average gain for each of these five airports.

Our gain variable which is the departure delay - the arrival delay tells us that across the 5 most frequent UA destinations, the average gain was relatively small. All the values are positive which indicates that flights arrived with less delay than it departed with. Flights to all 5 airports made up between 6.9 and 8.7 minutes on average.

```r
# Create the gain variables
ua_flights <- flights %>%
  filter(carrier == "UA") %>%
  mutate(gain = dep_delay - arr_delay)


# Filter to the top 5 destinations
top5_codes <- top5_dests$dest

ua_top5 <- ua_flights %>%
  filter(dest %in% top5_codes)
```

```r
avg_gain <- ua_top5 %>%
  group_by(dest) %>%
  summarize(
    avg_gain = mean(gain, na.rm = TRUE),
    median_gain = median(gain, na.rm = TRUE),
    sd_gain = sd(gain, na.rm = TRUE),
    n = n()
  ) %>%
  arrange(desc(n))

avg_gain_named <- avg_gain %>%
  left_join(airports, by = c("dest" = "faa")) %>%
```

```
  select(dest, name, avg_gain, median_gain, sd_gain, n)

avg_gain_named
```

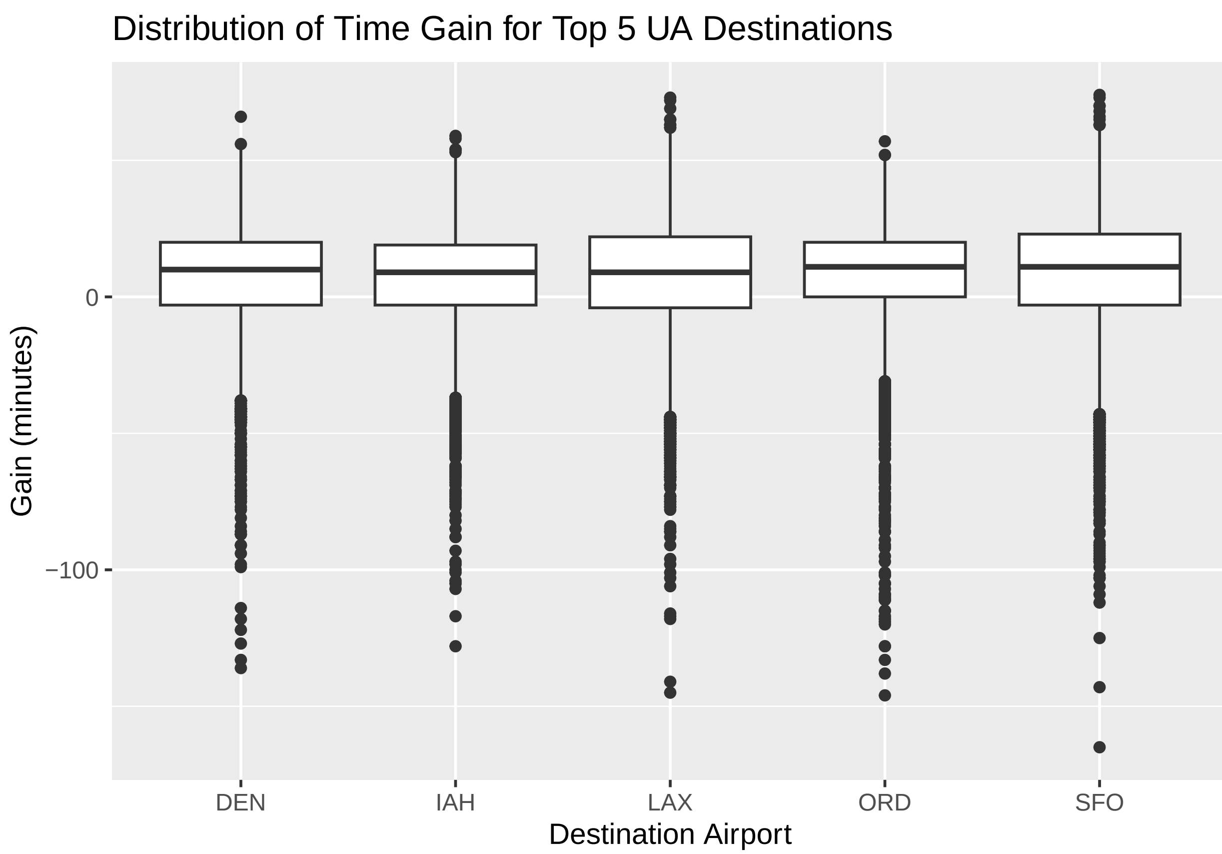
```
## # A tibble: 5 x 6
##   dest  name                           avg_gain median_gain sd_gain     n
##   <chr> <chr>                             <dbl>       <dbl>   <dbl> <int>
## 1 ORD   Chicago Ohare Intl                 7.78          11    19.2  6984
## 2 IAH   George Bush Intercontinental       6.86           9    18.4  6924
## 3 SFO   San Francisco Intl                 8.70          11    22.4  6819
## 4 LAX   Los Angeles Intl                   7.83           9    21.9  5823
## 5 DEN   Denver Intl                        7.30          10    20.0  3796
```

Plot the distribution

We also created a box-plot to visualize the flight gains and learned that the distribution of gains for all five airports were similar. Each destination had a centered distribution slightly above zero, which meant most flights made up at least a small amount of lost time. - although most UA flights made up a small time in the air, the distribution remain highly variable, and extreme delays still occur across all 5 destinations.

```
ggplot(ua_top5, aes(x = dest, y = gain)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Time Gain for Top 5 UA Destinations",
    x = "Destination Airport",
    y = "Gain (minutes)"
  )
```

```
## Warning: Removed 553 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



Confidence Interval for Average Gain

All 5 confidence intervals lie above zero, which confirms that UA flights to each of these destinations consistently make up time in the air on average. SFO, LAX, and ORD have slightly higher mean gains, but the intervals overlap across all airports. This suggests that although SFO appears to make up the most time (mean = 8.7 minutes), the differences between airports are modest and not extremely distinct.

```
ci_gain <- ua_top5 %>%
  group_by(dest) %>%
  summarize(
    mean_gain = mean(gain, na.rm = TRUE),
    sd_gain = sd(gain, na.rm = TRUE),
    n = n(),
    se = sd_gain / sqrt(n),
    lower = mean_gain - 1.96 * se,
    upper = mean_gain + 1.96 * se
  )

ci_gain
```

```
## # A tibble: 5 x 7
##   dest  mean_gain sd_gain     n    se lower upper
##   <chr>     <dbl>   <dbl> <int> <dbl> <dbl> <dbl>
## 1 DEN        7.30    20.0  3796 0.325  6.66  7.94
```

```
## 2 IAH        6.86    18.4  6924 0.222  6.43  7.30
## 3 LAX        7.83    21.9  5823 0.287  7.26  8.39
## 4 ORD        7.78    19.2  6984 0.229  7.33  8.23
## 5 SFO        8.70    22.4  6819 0.271  8.16  9.23
```

Regression Model (Gain - Destination + Distance + Flight Time)

we fit a regression model using destination, distance, and air time as predictors. Air time was the strongest and most significant predictor, with longer flights tending to make up less time on average. Distance itself was not significant after accounting for air time. The destination coefficients were all significant, indicating that even after controlling for flight duration, certain airports consistently show higher or lower gains compared to others. The model explains about 35% of the variation in gain ($R^2 = 0.35$), suggesting that both flight characteristics and destination-specific factors contribute to how much time is recovered in the air.

```r
model <- lm(gain ~ dest + distance + air_time, data = ua_top5)
summary(model)
```

```
##
## Call:
## lm(formula = gain ~ dest + distance + air_time, data = ua_top5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -146.583   -7.428    2.305   10.540   49.077
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 183.500893  17.764754   10.329   <2e-16 ***
## destIAH     -21.400433   2.280440   -9.384   <2e-16 ***
## destLAX      80.195927   9.378007    8.551   <2e-16 ***
## destORD     -86.249674   9.775281   -8.823   <2e-16 ***
## destSFO      95.076325  10.602387    8.967   <2e-16 ***
## distance     -0.002106   0.011067   -0.190    0.849
## air_time     -0.769298   0.006064 -126.862   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.44 on 29786 degrees of freedom
##   (553 observations deleted due to missingness)
## Multiple R-squared:  0.3533, Adjusted R-squared:  0.3532
## F-statistic:  2712 on 6 and 29786 DF,  p-value: < 2.2e-16
```