DA5020 – Practicum 2

Rebecca Weiss

11/7/2021

Clear the workspace:

```
rm(list = ls())
```

```
#load all necessary libraries
library(mongolite)
library(tidyverse)
library(lubridate)
library(RCurl)
library(rvest)
library(scrapeR)
library(usmap)
```

1. Configure the Database: MongoDB Atlas (10 points)

1.1 Create account, add users, connect IP, etc: screenshot attached

1.2 Establish mongo_url

```
mongo_url <- 'mongodb+srv://rweiss:da5020NEU@da5020-cluster.etlrb.mongodb.net/myFirstDatabase?retryWrit
```

1.3 Create the database and load the data

```
# connect to mongoDB instance and create a database called airline_performance and a collection called
mongo_connection <- mongo(collection = 'flights_2019', db = 'airline_performance', url = mongo_url)
```

```
# load  attached CSV data "2019_ONTIME_REPORTING_FSW.csv"
data <- read.csv("2019_ONTIME_REPORTING_FSW.csv")

# check to make sure data looks ok
head(data)
```

```
##      FL_DATE CARRIER_CODE TAIL_NUM FL_NUM ORIGIN ORIGIN_ST DEST DEST_ST
## 1 2019-10-01           AA   N916NN   2311    TUS        AZ  ORD      IL
## 2 2019-10-01           AA   N733UW   2315    PHX        AZ  DEN      CO
## 3 2019-10-01           AA   N140AN   2318    DFW        TX  LAX      CA
## 4 2019-10-01           AA   N925AN   2325    SNA        CA  DFW      TX
## 5 2019-10-01           AA   N143AN   2328    ATL        GA  LAX      CA
```

```
## 6 2019-10-01           AA   N816NN   2339    PHX        AZ  JFK      NY
##   DEP_TIME DEP_DELAY ARR_TIME ARR_DELAY ELAPSED_TIME DISTANCE
## 1      828         0     1353         0          205     1437
## 2     1907         0     2159         0          112      602
## 3     1904       104     2016       101          192     1235
## 4     1729         4     2215         3          166     1205
## 5      656         0      841         0          285     1947
## 6      911         0     1702         0          291     2153
```

```
#insert the data into the database
# mongo_connection$insert(data)

#verify the number of documents inserted
mongo_connection$count()
```

```
## [1] 1897503
```

2. Let's explore patterns in the region. For each of the original 3 states (i.e. AZ, NV, CA), analyze the most popular outbound/destination airports. For example, if a flight originated in CA (at any of its airports), where do they often go? Comment on your findings and visualize the top results.

```
# write a function to query the state data from mongoDB
getdata <- function(stmt)   {
  #Display only certain fields we want
  results <- mongo_connection$find(query = stmt, fields = '{"ORIGIN_ST": true,
                                    "DEST": true, "DEST_ST": true,
                                    "DISTANCE": true, "_id": false}')
  # count and sort top results
  top_dest <- results %>%
  group_by(DEST_ST) %>%
  summarise("Total" = n()) %>%
  arrange(desc(Total))

  # rename state column to work with plotting
  names(top_dest) <- c("state", "Total")
  return(top_dest)
}
```

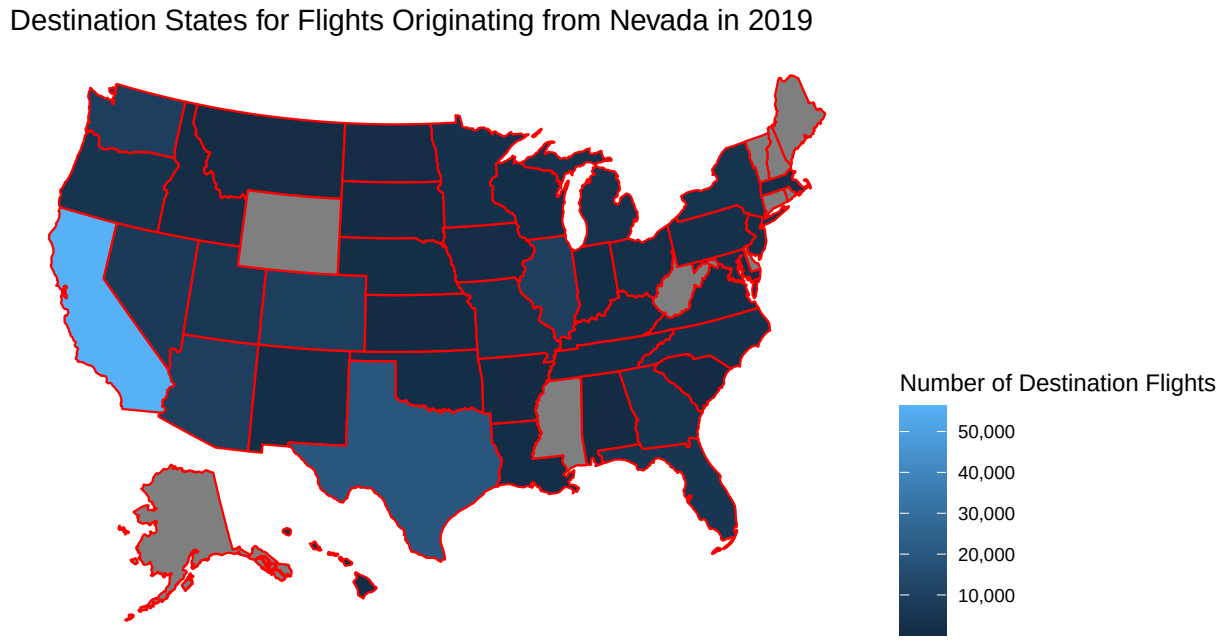
```
#  write a function to plot ALL outgoing destinations:
map_dest <- function(top_dest, state)  {
    p <- plot_usmap(data = top_dest, values = "Total", color = "red") +
    scale_fill_continuous(name = "Number of Destination Flights", label = scales::comma) +
    theme(legend.position = "right") +
    labs(caption = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D")
    if(state == "NV") {
      p <- p+labs(title = "Destination States for Flights Originating from Nevada in 2019")
    } else if(state == "CA") {
      p <- p+labs(title = "Destination States for Flights Originating from California in 2019")
    } else {
      p <- p+labs(title = "Destination States for Flights Originating from Arizona in 2019")
```

```
    }
    p
}
```

```
# write a function to visualize top 10 results
bartop_dest <- function(top_dest, state)  {
  # visualize top 10 destinations
  top10 <- top_dest %>% slice(1:10)

  # create bar chart
  b <- ggplot(data = top10, aes(x=reorder(state, Total), y=Total)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    # coord_flip() +
    labs(title = "Top 10 Destinations of Aircraft 309NV",
         subtitle = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
          y = "Number of Trips",
          x = "Destination State") +
    theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
    theme(plot.caption.position = "plot", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
    theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black"))  +
    # scale_y_continuous(labels = scales::comma) +
    theme(axis.text = element_text(size=8, hjust=0.5, color="black")) +
    geom_text(aes(label = Total), size = 2.5, position = position_stack(vjust = 0.5))
    labs(caption = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D")
    if(state == "NV") {
      b <- b+labs(title = "Destination States for Flights Originating from Nevada",
                  caption = "The 10 states in this graph account for the most frequent destination state
                  for flights originating from Nevada in 2019.")
    } else if(state == "CA") {
      b <- b+labs(title = "Destination States for Flights Originating from California",
                  caption = "The 10 states in this graph account for the most frequent destination state
                  flights originating from California in 2019.")
    } else {
      b <- b+labs(title = "Destination States for Flights Originating from Arizona in 2019",
                  caption = "The 10 states in this graph account for the most frequent destination state
                  for flights originating from Arizona in 2019.")
    }
    b
}
```

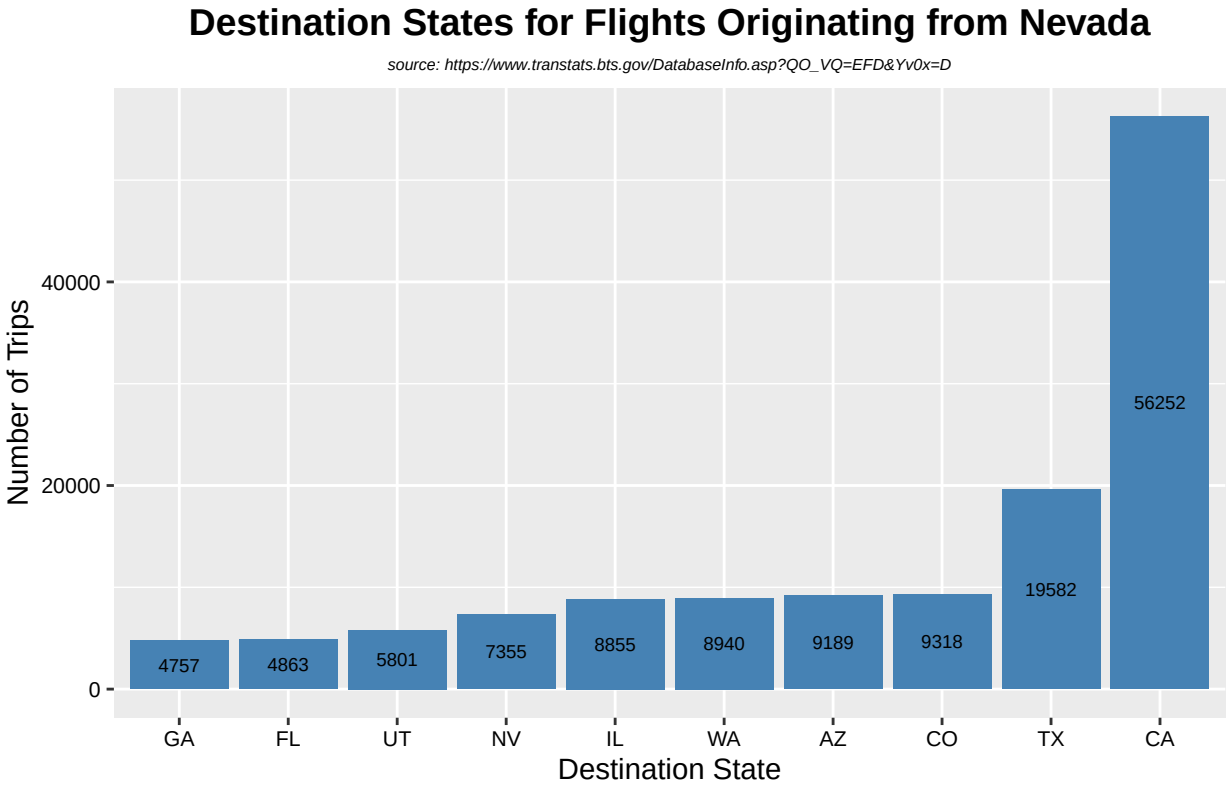
Now, we will use functions to analyze/visualize data from each origin state:

```
# Nevada:
nv <- getdata('{"ORIGIN_ST": "NV"}')
map_dest(nv, "NV")
```



source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D

```
bartop_dest(nv, "NV")
```

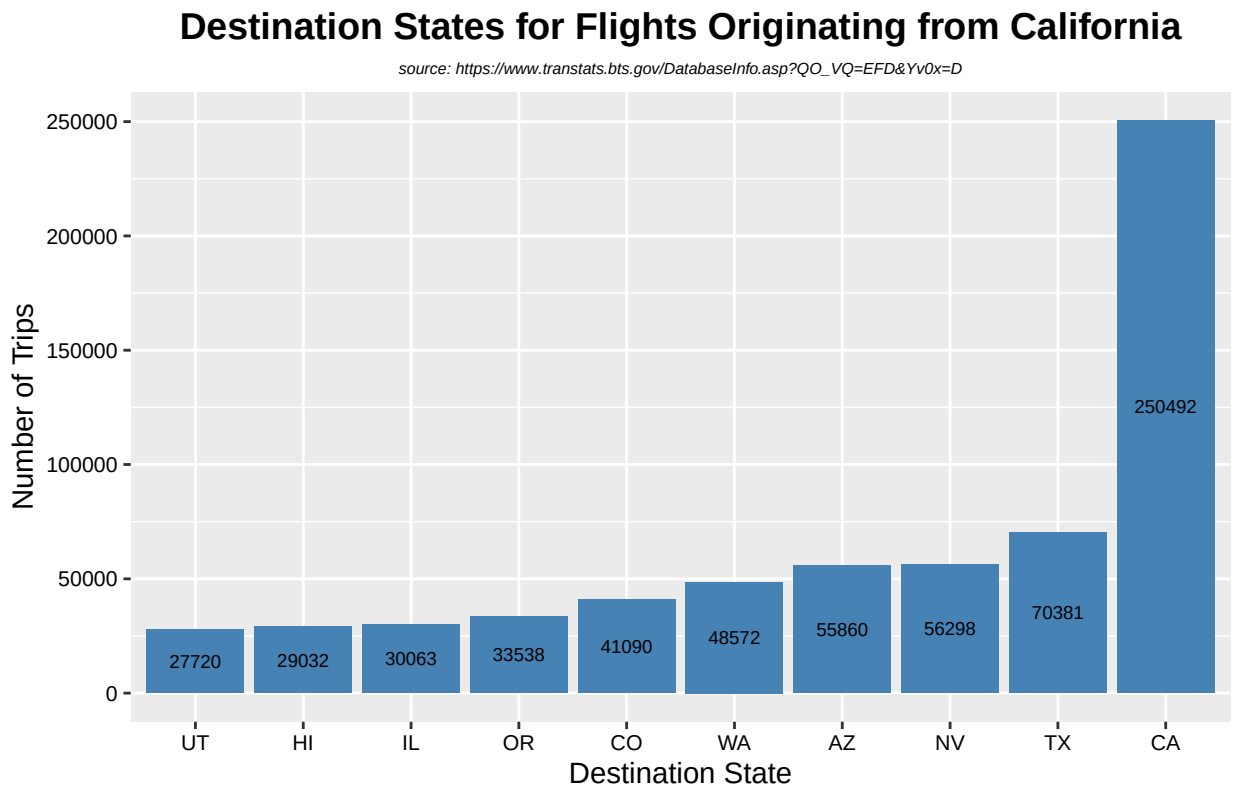


The 10 states in this graph account for the most frequent destination states for flights originating from Nevada in 2019.

```
# California:
ca <- getdata('{"ORIGIN_ST": "CA"}')
map_dest(ca, "CA")
```

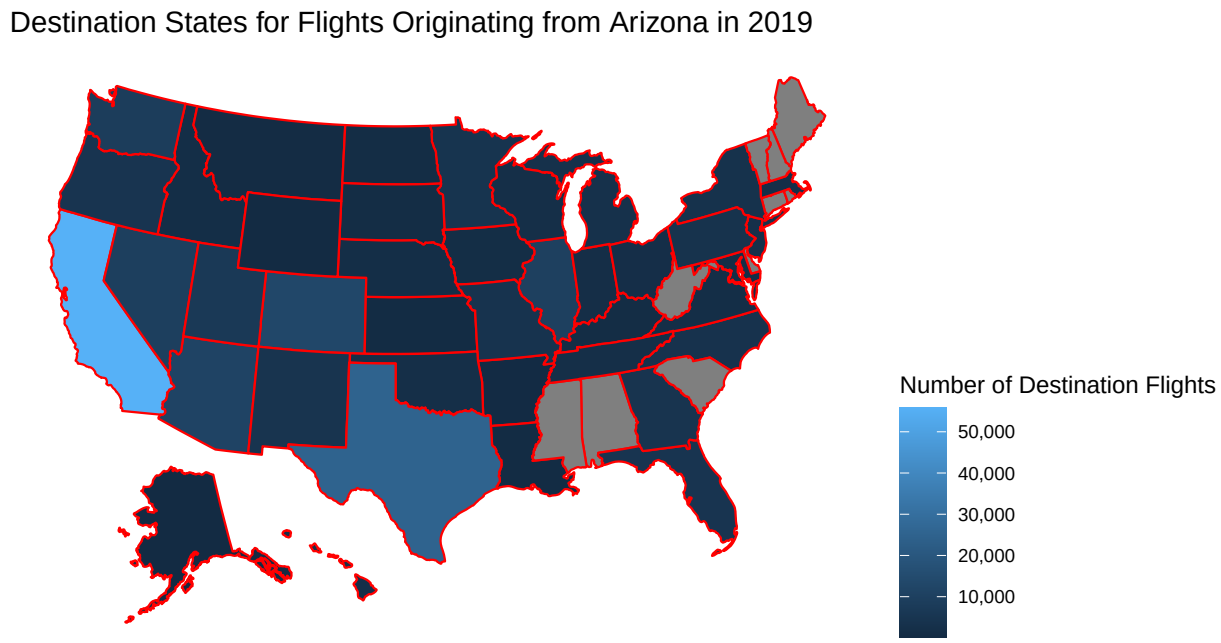
source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D

```
bartop_dest(ca, "CA")
```



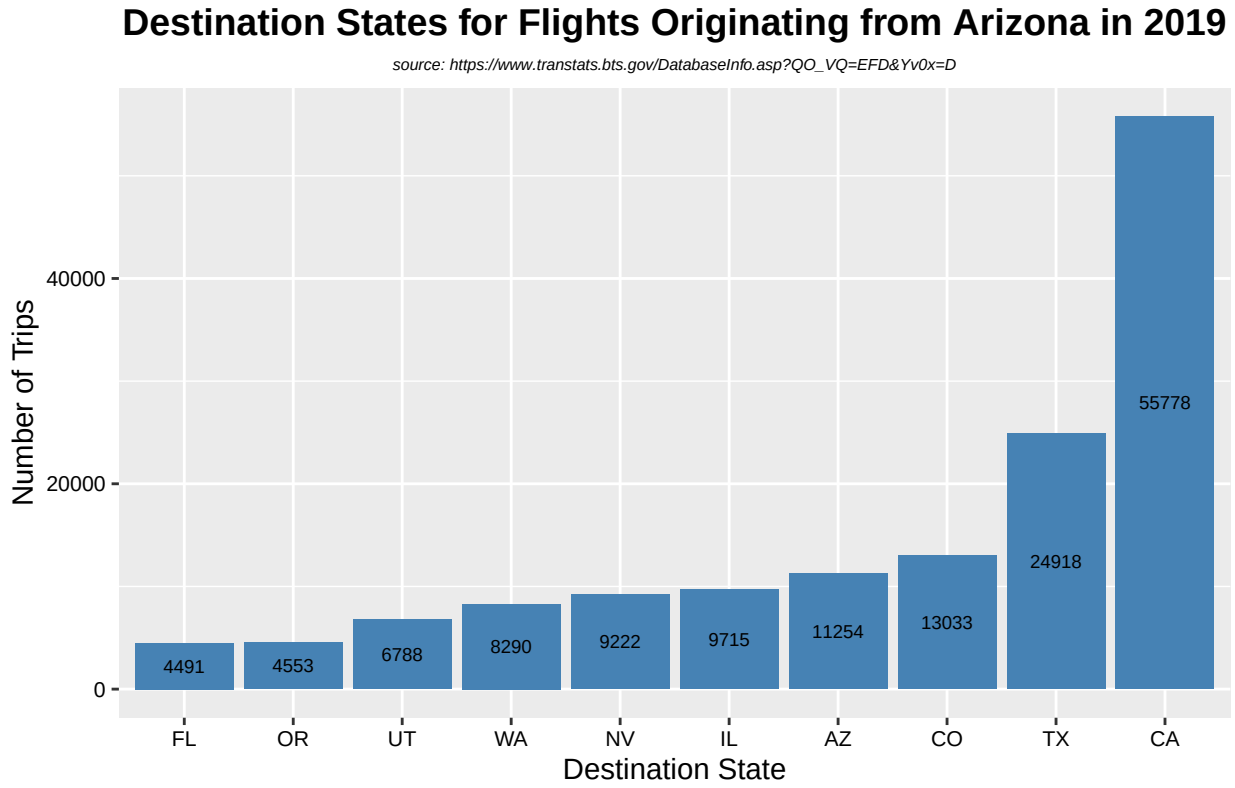
The 10 states in this graph account for the most frequent destination states for flights originating from California in 2019.

```
# Arizona
az <- getdata('{"ORIGIN_ST": "AZ"}')
map_dest(az, "AZ")
```



source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D

```
bartop_dest(az, "AZ")
```



The 10 states in this graph account for the most frequent destination states for flights originating from Arizona in 2019.

From the visualizations generated, we see quite a few trends in the region. First, that of the three origin states, all of them travel to California the most, with Texas in second place. This makes sense as California and Texas are heavily populated, are large geographically, and have many airline hubs for international/other connecting flights. Another takeaway is that most of these flights originating from these three states go all over the country, but for the most part, they have the highest frequency to travel out west. California has the most outbound states, as we can see from the maps: Arizona and Nevada have more states that they had not traveled to, which are depicted by the states shaded in gray on the map.

3. Let's explore the carriers. Calculate the total flights for each airline/operator.

a. Ensure that you indicate the full name of each carrier, in lieu of the carrier code. This will require web scraping.

```
## Web scraping:

# download HTML from URL, scrape table of airline codes and save into code_df
table_url <- "https://en.wikipedia.org/wiki/List_of_airline_codes"
html_data <- read_html(table_url)
code_df <- html_data %>%
  html_node('.wikitable') %>%
  html_table() %>%
  select(c(1, 3)) # select only relevant columns (code (IATA) + airline)
```

```
## MongoDB query:
# group by carrier name, sort, limit to 10
```

```
stmt <- '[{"$group":
          {"_id": "$CARRIER_CODE",
          "Total": {"$sum": 1}
            }
          },
  {"$sort": {"Total": -1}},
  {"$limit": 10}
]'

airline_count <- mongo_connection$aggregate(stmt)

# change column names
names(airline_count) <- c("Carrier", "Total")

# view results
airline_count
```

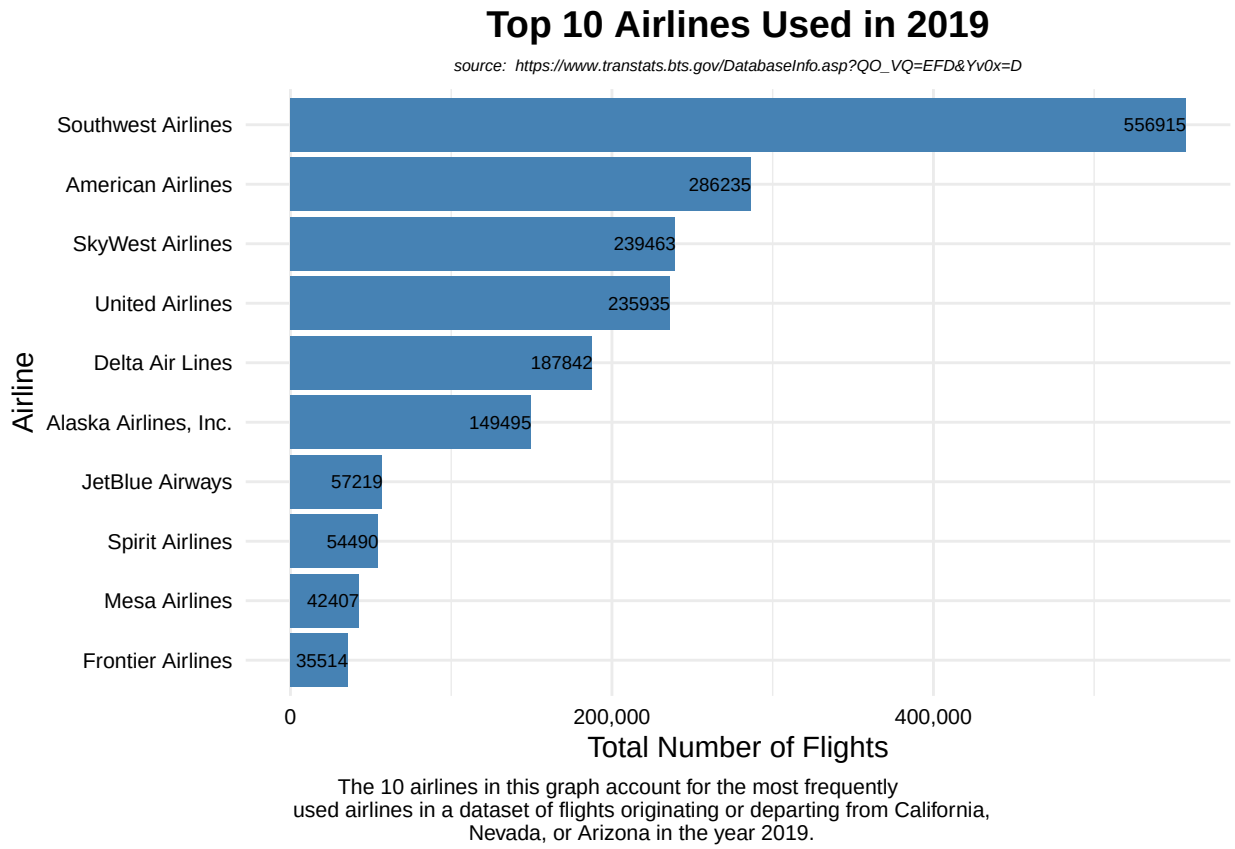
```
##    Carrier  Total
## 1       WN 556915
## 2       AA 286235
## 3       OO 239463
## 4       UA 235935
## 5       DL 187842
## 6       AS 149495
## 7       B6  57219
## 8       NK  54490
## 9       YV  42407
## 10      F9  35514
```

```
## Joining data:

# join mongoDB query with web-scraped
airline <- left_join(airline_count, code_df, by = c("Carrier" = "IATA"))

# make sure data looks OK
airline
```

```
##    Carrier  Total                Airline
## 1       WN 556915    Southwest Airlines
## 2       AA 286235     American Airlines
## 3       OO 239463      SkyWest Airlines
## 4       UA 235935       United Airlines
## 5       DL 187842       Delta Air Lines
## 6       AS 149495 Alaska Airlines, Inc.
## 7       B6  57219       JetBlue Airways
## 8       NK  54490       Spirit Airlines
## 9       YV  42407         Mesa Airlines
## 10      F9  35514     Frontier Airlines
```

b. Visualize the top 10 results and show the carrier name and the frequency. Explain the results.

```
# create bar chart
ggplot(data = airline, aes(x=reorder(Airline, Total), y=Total)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Top 10 Airlines Used in 2019",
       subtitle = "source:  https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       caption = "The 10 airlines in this graph account for the most frequently
       used airlines in a dataset of flights originating or departing from California,
       Nevada, or Arizona in the year 2019.",
       y = "Total Number of Flights",
       x = "Airline") +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
  theme(plot.caption.position = "plot", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
  theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black"))  +
  scale_y_continuous(labels = scales::comma) +
  theme(axis.text = element_text(size=8, hjust=0.5, color="black")) +
  geom_text(aes(label = Total), color = "black", size=2.5, hjust=.99)
```



From the output above, we see that Southwest Airlines has the most flights by a longshot, with almost double the next airline, American. Once we get past the top 6 airlines – Southwest, American, SkyWest, United, Delta, and Alaska – we see a pretty significant drop in the number of flights by airline. Seeing this trend, it is very likely that other airlines make up a very small portion of the total number of flights departing or originating from CA, NV, or AZ in 2019.

4. Select the top 5 airlines, from the previous question, and calculate the total flight hours for each month (grouped by airline). Explain and visualize the results. Hint: the total flight hours is not equivalent to the frequency of flights and ensure that you display the total hours and not the total minutes.

```
# query MongoDB
stmt <- '[
{"$group":
{"_id":
{"CARRIER_CODE":"$CARRIER_CODE",
"FL_DATE": "$FL_DATE"},
"Sum": {"$sum":"$ELAPSED_TIME"}}
  },
      {"$project":
          {"_id":0,
          "Carrier": "$_id.CARRIER_CODE",
          "Date": "$_id.FL_DATE",
          "Sum":1
          }
        }]'

time_date_airline <- mongo_connection$aggregate(stmt)

# make sure data looks OK
head(time_date_airline)
```

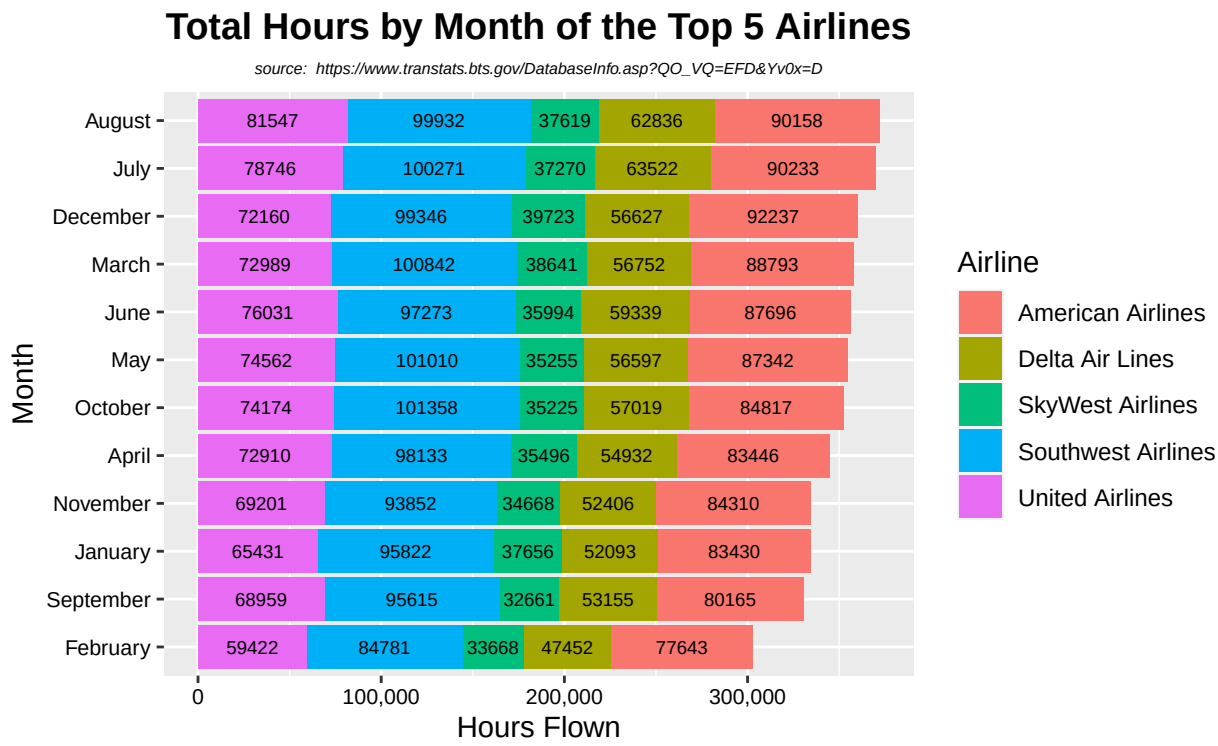
```
##      Sum Carrier       Date
## 1 152614      UA 2019-05-23
## 2 206556      WN 2019-05-24
## 3  17859      F9 2019-06-20
## 4 140491      WN 2019-11-28
## 5  17831      F9 2019-05-19
## 6 121975      UA 2019-10-31
```

```
# filter data for top 5 airlines (using top 5 from output of problem 3)
top5 <- airline_count %>% slice(1:5)

final5 <- inner_join(time_date_airline, top5, by = "Carrier")

# convert Date to month
final5$Month <- months(as.Date(final5$Date))

# get Airline name from scraped data
final5 <- left_join(final5, code_df, by = c("Carrier" = "IATA"))

# select columns of interest, get airline name, group by carrier
(final5 <- final5 %>%
  select(Sum, Airline, Month) %>%
  group_by(Month, Airline) %>%
  summarise("Hours" = (sum(Sum) / 60)))
```

```
## `summarise()` has grouped output by 'Month'. You can override using the `.groups` argument.
```

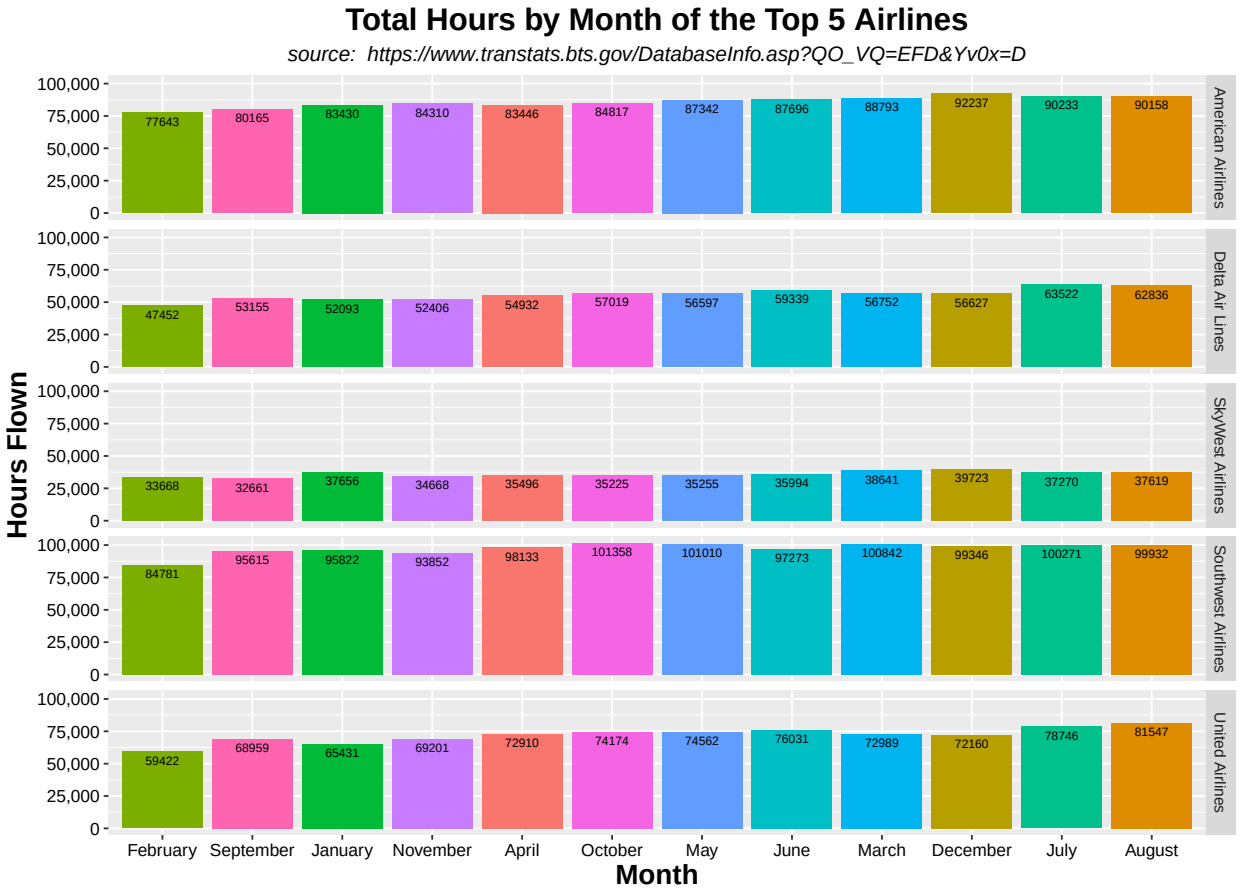
```
## # A tibble: 60 x 3
## # Groups:   Month [12]
##    Month  Airline             Hours
##    <chr>  <chr>               <dbl>
##  1 April  American Airlines  83446.
##  2 April  Delta Air Lines    54932.
##  3 April  SkyWest Airlines   35496.
##  4 April  Southwest Airlines 98133.
##  5 April  United Airlines    72910.
##  6 August American Airlines  90158.
##  7 August Delta Air Lines    62836.
##  8 August SkyWest Airlines   37619.
##  9 August Southwest Airlines 99932.
## 10 August United Airlines    81547.
## # ... with 50 more rows
```

```
# visualize top 5 airlines by month: total flights each month
ggplot(final5, aes(fill=Airline, y=Hours, x=reorder(Month, Hours),
                   label=sprintf("%0.0f", round(Hours, digits = 0)))) +
    geom_bar(position="stack", stat="identity") +
  geom_text(size = 2.5, position = position_stack(vjust = 0.5)) +
  coord_flip() +
    labs(title = "Total Hours by Month of the Top 5 Airlines",
       subtitle = "source:  https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       caption = "The 5 airlines in this graph account for the most frequently
       used airlines in a dataset of flights originating or departing from California,
       Nevada, or Arizona in the year 2019. From these 5 airlines, we show the number of hours
       each airline flew by month in 2019.",
       y = "Hours Flown",
       x = "Month") +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
  theme(plot.caption.position = "plot", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
  theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black"))  +
  scale_y_continuous(labels = scales::comma) +
  theme(axis.text = element_text(size=8, hjust=0.5, color="black"))
```



The 5 airlines in this graph account for the most frequently used airlines in a dataset of flights originating or departing from California, Nevada, or Arizona in the year 2019. From these 5 airlines, we show the number of hours each airline flew by month in 2019.

```
# visualize breakdown by airline
g <- ggplot(final5, aes(x = reorder(Month, Hours), y = Hours, fill = Month,
                        label=sprintf("%0.0f", round(Hours, digits = 0))))

g +
  geom_bar(stat = "identity") +
  facet_grid(facets = Airline ~ .) +
  theme(legend.position = "none") +
  geom_text(vjust=1.5, hjust=.5, color="black",
            position = position_dodge(0), size=2.5) +
      labs(title = "Total Hours by Month of the Top 5 Airlines",
       subtitle = "source:  https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       caption = "The 5 airlines in this graph account for the most frequently
       used airlines in a dataset of flights originating or departing from California,
       Nevada, or Arizona in the year 2019. From these 5 airlines, we show the number of hours
       each airline flew by month in 2019, further faceted by each airline.",
       y = "Hours Flown",
       x = "Month") +
  theme(plot.subtitle=element_text(size=12, hjust=0.5, face="italic", color="black")) +
  theme(plot.caption.position = "plot", plot.caption=element_text(size=12, hjust=0.5, color="black")) +
  theme(plot.title=element_text(size=18, hjust=0.5, face="bold", color="black"))  +
  scale_y_continuous(labels = scales::comma) +
  theme(axis.text = element_text(size=10, hjust=0.5, color="black"),
        axis.title = element_text(size=16, face = "bold"))
```



The 5 airlines in this graph account for the most frequently used airlines in a dataset of flights originating or departing from California, Nevada, or Arizona in the year 2019. From these 5 airlines, we show the number of hours each airline flew by month in 2019, further faceted by each airline.

Both of the outputs above show the same information, but in 2 different formats. From the first output, the total number of hours was flown each month by the 5 airlines, and colored to reflect which airline that was. In the second graph, the airlines are faceted, and looking at each column we can compare the number of hours each airline flew each month. In both graphs, it is clear that Southwest Airlines has the most hours, with American Airlines not too far behind. Interestingly, Skywest Airlines has the least number of hours despite it having more total flights than United or Delta airlines, as we saw in the output from question 3. This suggests that possibly it does more short distance flights, as it has a higher number of total flights, yet fewer hours than United or Delta.

Another interesting takeaway from these graphs is that we can see which months had the highest total hours of flights from these top 5 airlines. August and July were the top months with the most hours flown, with December not too far behind. This would make sense because people likely travel in the summer, and during the holidays. What surprised me was to see that November was not in that top tier of hours per month, but rather was 4th to last. There could be a number of reasons why this happened, such as people may live closer to home and don't need to fly as many *hours* for Thanksgiving, whereas the other months the flying done could be to further vacation destinations. Or, people may drive home for Thanksgiving, is another possible explanation. The month to come dead last in number of hours flown is February, which I think is to be expected since it is not great weather, and there are not significant holidays where people are off work /school to travel much.

5. Select any (1) aircraft, and explore the data to determine where it often travels. Calculate its average arrival and departure delays at the airports. After which analyze all the results to identify any patterns that are evident and also indicate which airline operates that aircraft. Explain your findings and visualize the results. Note: the TAIL_NUM can help you to identify each unique aircraft.

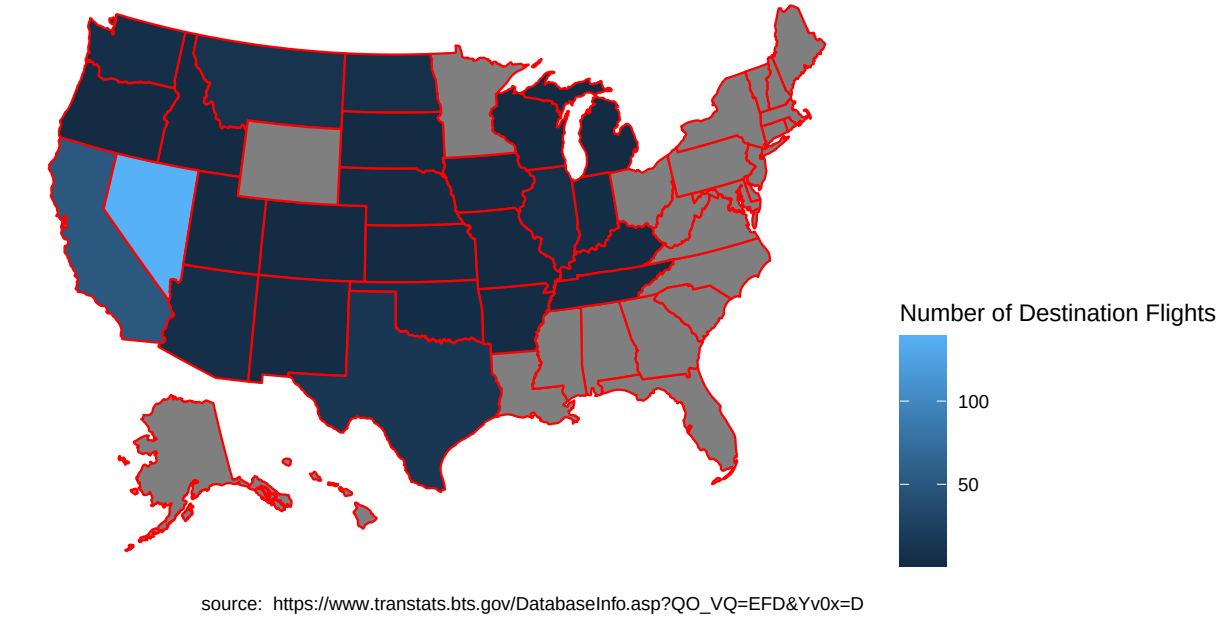
```
#first, we want to see how many distinct tail_nums we have
length(mongo_connection$distinct("TAIL_NUM"))
```

```
## [1] 4889
```

From the 4889 distinct aircrafts, I have randomly selected TAIL_NUM: 309NV

```
#the query below will get all data
stmt <- '{"TAIL_NUM": "309NV"}'

#Display only certain fields we want for '309NV'
ac_results <- mongo_connection$find(query = stmt, fields = '{"CARRIER_CODE": true, "ORIGIN_ST": true,
                                    "DEST_ST": true, "DEP_DELAY": true,
                                    "ARR_DELAY": true, "_id": false}')
# make sure to get the carrier name, so join with table in previous question:
ac_results <- left_join(ac_results, code_df, by = c("CARRIER_CODE" = "IATA"))

# calculate arrive and departure delay average (minutes)
dep_avg <- mean(ac_results$DEP_DELAY)
arr_avg <- mean(ac_results$ARR_DELAY)
```

```
### DESTINATIONS ###
# get destinations by the aircraft, sorted by most frequent
top_dest <- ac_results %>%
  group_by(DEST_ST) %>%
  summarise("Total" = n()) %>%
  arrange(desc(Total))


# rename state column to work with plotting
names(top_dest) <- c("state", "Total")
# plot all destinations, show by count
plot_usmap(data = top_dest, values = "Total", color = "red") +
  scale_fill_continuous(name = "Number of Destination Flights", label = scales::comma) +
  labs(title = "Destination States for Aircraft 309NV",
       caption = "source:  https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       subtitle = "Airline: Allegiant Air; Average Delay = 11.2 minutes arrival, 9.7 minutes departure")
  theme(legend.position = "right")
```

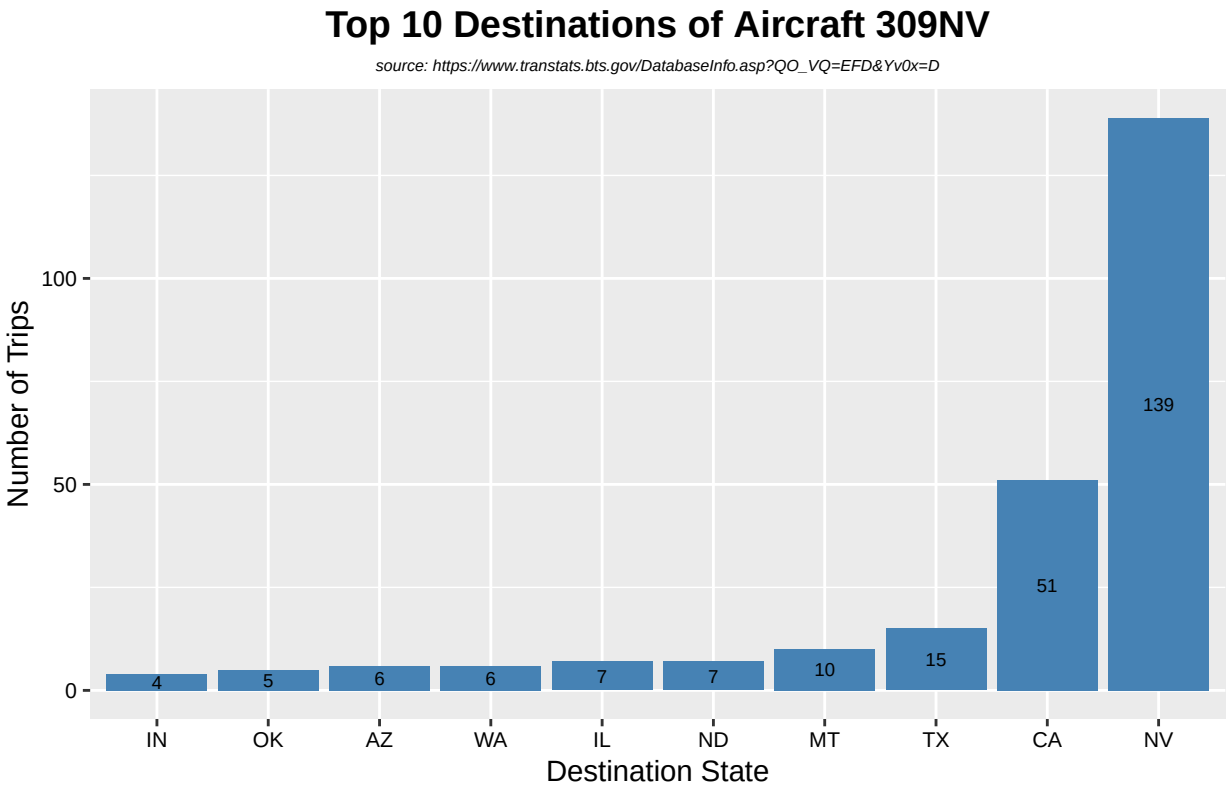



```
# visualize top 10 destinations
(top10 <- top_dest %>% slice(1:10))
```

```
## # A tibble: 10 x 2
##    state Total
##    <chr> <int>
##  1 NV      139
##  2 CA       51
##  3 TX       15
##  4 MT       10
##  5 IL        7
##  6 ND        7
##  7 AZ        6
##  8 WA        6
##  9 OK        5
## 10 IN        4
```

```
# create bar chart
ggplot(data = top10, aes(x=reorder(state, Total), y=Total)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  # coord_flip() +
  labs(title = "Top 10 Destinations of Aircraft 309NV",
       subtitle = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       caption = "The 10 states in this graph account for the most frequent destination
       states for aircraft 309NV in the year 2019.",
```

```
      y = "Number of Trips",
      x = "Destination State") +
theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
theme(plot.caption.position = "plot", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black"))  +
# scale_y_continuous(labels = scales::comma) +
theme(axis.text = element_text(size=8, hjust=0.5, color="black")) +
geom_text(aes(label = Total), size = 2.5, position = position_stack(vjust = 0.5))
```



The 10 states in this graph account for the most frequent destination states for aircraft 309NV in the year 2019.

From the output, we can see that TAIL_NUM 309NV is with Allegiant Airlines and has a average departure delay of 9.7 minutes, and an average arrival delay of 11.2 minutes. From the bar plot, we see that the most frequent destination is Nevada by far, with California next. From the map visual, we see that the aircraft stays pretty exclusively to the west region, and does not tend to fly out to the east coast. One interesting pattern seen on the map is that this aircraft does not fly to Wyoming, but flies to every other state surrounding it.

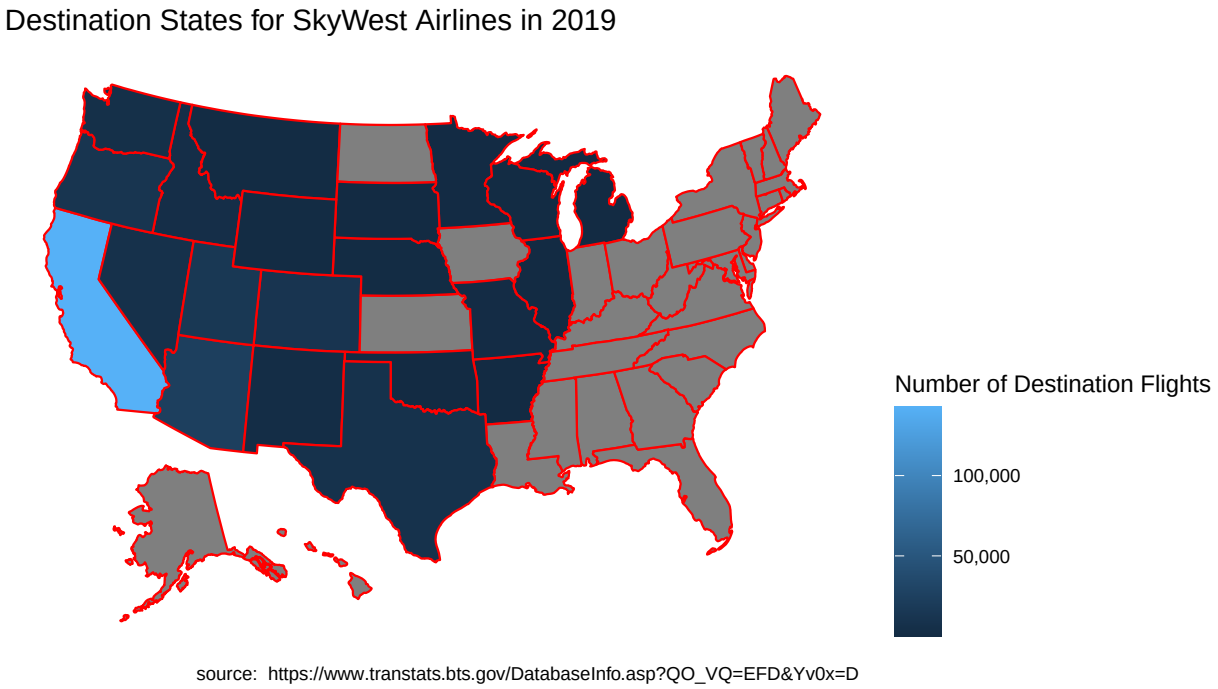
6. (Bonus). Build one additional query to test a hypothesis or answer a question that you have about the dataset. Your query should retrieve data from MongoDB and evaluate the pattern/trend. Prepare supporting visualizations for your analysis. If necessary, you can integrate any additional data that provide more details or support your analysis/findings.

A question I am curious about arises from my answer for question 4: Why is it that SkyWest airlines has fewer hours traveled than the other top 5 airlines, but is 3rd in total number of flights departing or originating in CA, NV, AZ? As I suggested in my explanation, one hypothesis I have is that SkyWest flies shorter *distances*

and *frequently*, so that is why there is a discrepancy between the number of hours flown vs the frequency of use. Since we already analyzed the frequency in question 3 and found that there were 239,463 flights in 2019, I am going to look at the distances and destination of those flights to see if my hypothesis holds.

From previous queries, I know that CARRIER_CODE of SkyWest = 'OO'. I will start by collecting the data from that code in monogdb, and plot the frequent places traveled (by looking at the destination states), and then calculate the average distance.

```
# use the function created in question 2 to get the destination data I want
sky <- getdata('{"CARRIER_CODE": "OO"}')

# plot all destinations, show by count
plot_usmap(data = sky, values = "Total", color = "red") +
  scale_fill_continuous(name = "Number of Destination Flights", label = scales::comma) +
  labs(title = "Destination States for SkyWest Airlines in 2019",
       caption = "source:  https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D") +
  theme(legend.position = "right")
```



As we can see from the map, the most frequently traveled to state is California, and the airline does not fly east past Illinois. These data make sense, however, there are still a good number of miles between the states the airline *does* fly to, so this graph alone does not fully satisfy our hypothesis. Lets take a look at the distribution of distances, to get a better understanding of how often Skywest airlines travels shorter distances

```
# get full data frame of fields we want
results <- mongo_connection$find(query = '{"CARRIER_CODE": "OO"}', fields = '{"ORIGIN_ST": true, "DEST"
                                 "DEST_ST": true, "DISTANCE": true, "_id": false}')
```

```r
# show summary data of output
summary(results)
```

```
##   ORIGIN_ST             DEST             DEST_ST             DISTANCE
##  Length:239463      Length:239463      Length:239463      Min.   :  66.0
##  Class :character   Class :character   Class :character   1st Qu.: 308.0
##  Mode  :character   Mode  :character   Mode  :character   Median : 493.0
##                                                           Mean   : 562.2
##                                                           3rd Qu.: 737.0
##                                                           Max.   :1772.0
```

```r
# Calculate values and display
print(paste("The mean distance traveled on SkyWest Airlines =", mean(results$DISTANCE)))
```

```
## [1] "The mean distance traveled on SkyWest Airlines = 562.218522276928"
```

```r
print(paste("The median distance traveled on SkyWest Airlines =", median(results$DISTANCE)))
```

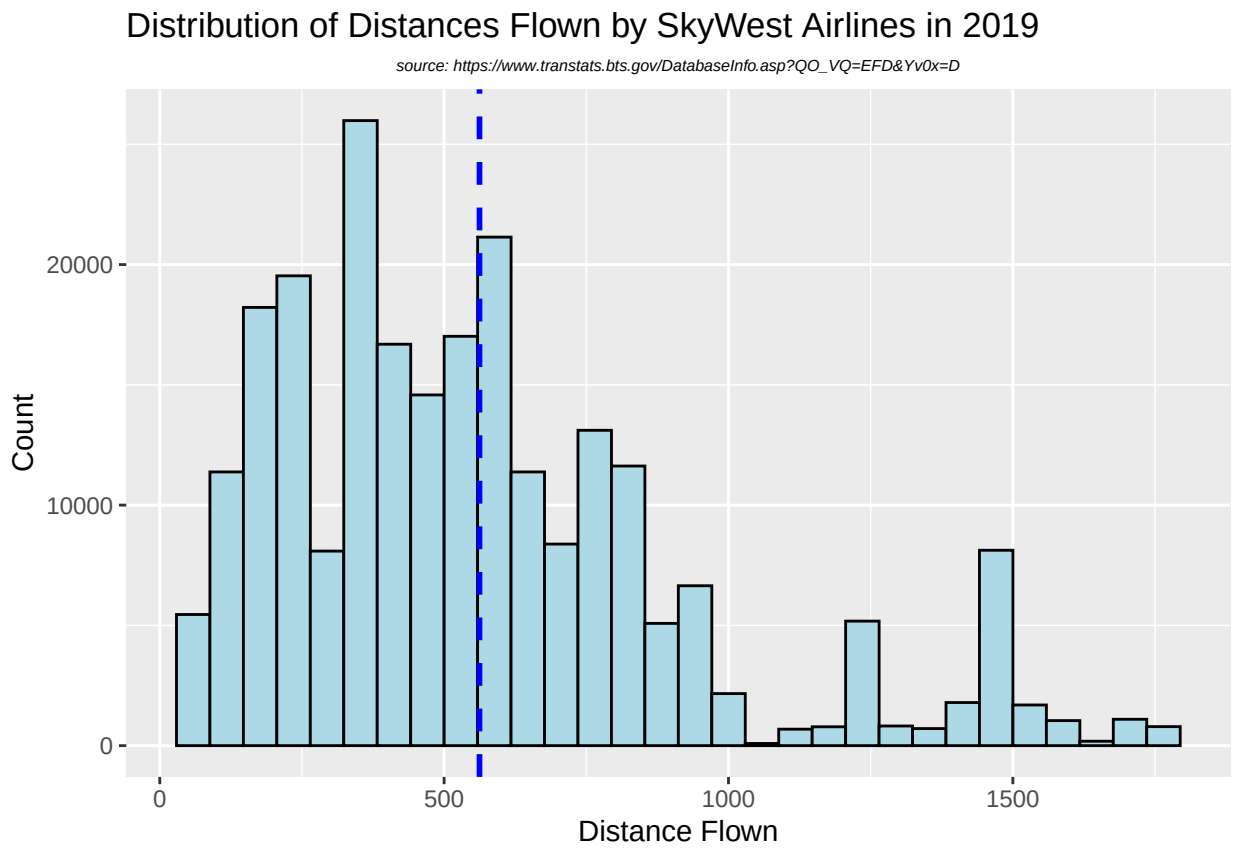
```
## [1] "The median distance traveled on SkyWest Airlines = 493"
```

```r
print(paste("The standard deviation of distance traveled on SkyWest Airlines =", sd(results$DISTANCE)))
```

```
## [1] "The standard deviation of distance traveled on SkyWest Airlines = 362.807817347324"
```

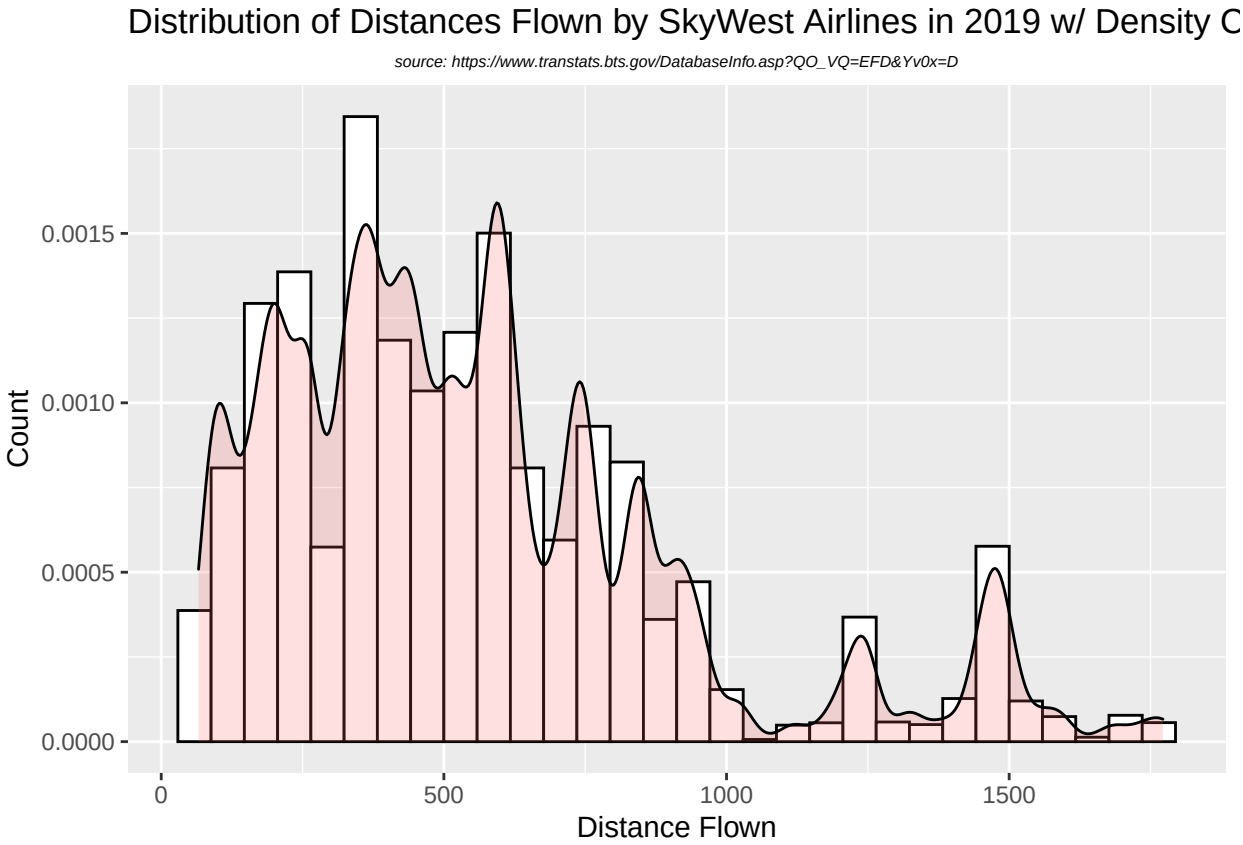
```r
# create general histogram variable
p <- ggplot(results, aes(x=DISTANCE)) +
  geom_histogram(color="black", fill="lightblue") +
  labs(title = "Distribution of Distances Flown by SkyWest Airlines in 2019",
       x = "Distance Flown",
       y = "Count",
       subtitle = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D") +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black"))


# Show plot with dotted line for mean:
p + geom_vline(aes(xintercept=mean(DISTANCE)),
            color="blue", linetype="dashed", size=1)
```

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




```
# add a density plot:
ggplot(results, aes(x=DISTANCE)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white") +
 geom_density(alpha=.2, fill="#FF6666") +
  labs(title = "Distribution of Distances Flown by SkyWest Airlines in 2019 w/ Density Curve",
       x = "Distance Flown",
       y = "Count",
       subtitle = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D") +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black"))
```

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



As we can see from the histograms and density curve graphs, the mean falls just above 500 miles traveled – which can be seen on the dotted line in the first, blue curve as well as the values printed above – and the density/rest of the curve is skewed so that many of the flights are less than that. We also see that the median (493 miles) is fairly similar to the mean (562 miles).

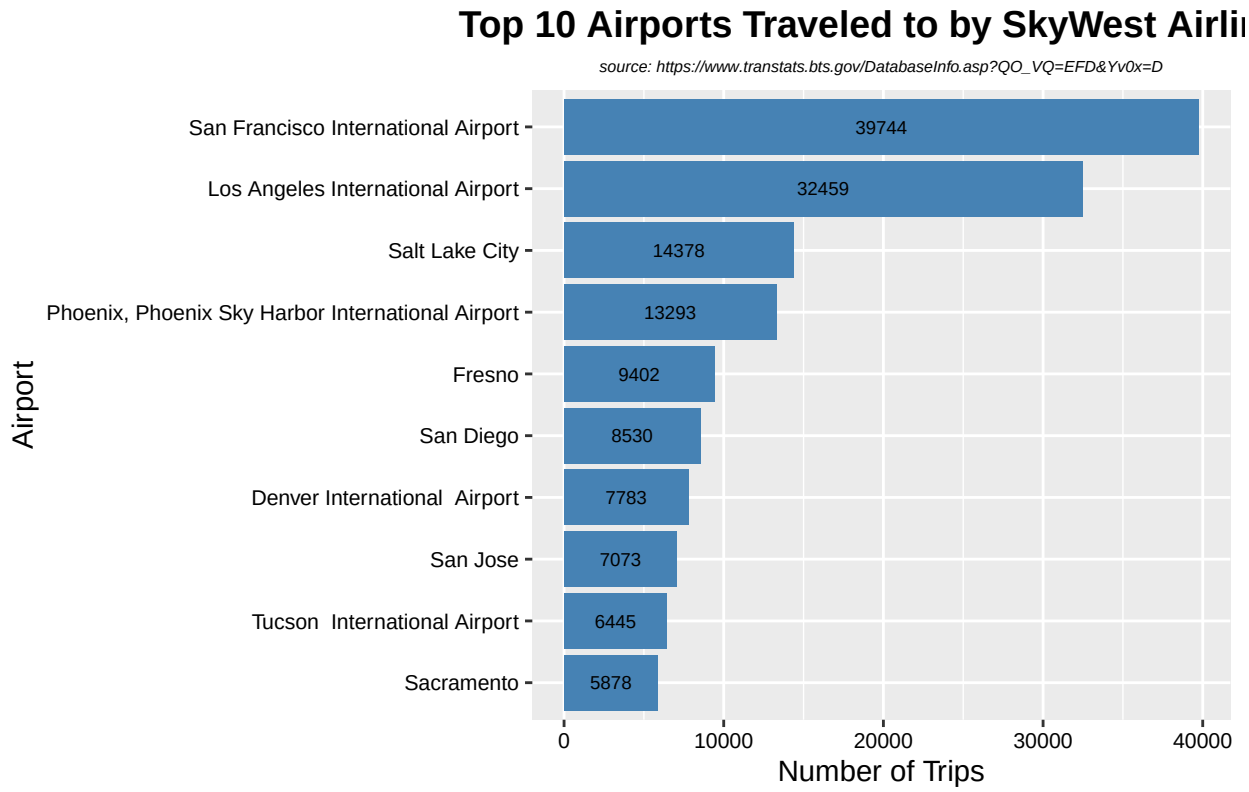
One final piece I think we can explore is how often is SkyWest Airlines going to certain airports. Which ones does it frequent the most? To analyze this, we will need to scrape data to determine what the name of the airport is from the DEST code we queried in our *results* dataframe.

```
# scrape for airport names
airport_url <- 'https://www.leonardsguide.com/us-airport-codes.shtml'
html_data <- read_html(airport_url)
airport_df <- html_data %>%
  html_node('table') %>%
  html_table()

# change names
names(airport_df) <- c("Airport", "DEST")


# join with results
aps <- left_join(results, airport_df, by = "DEST")

top_ap <- aps %>%
  drop_na() %>% # in case our large table does not have every airport
  group_by(Airport) %>%
  summarise("Total" = n()) %>%
```

```
  arrange(desc(Total))

# visualize top 10 airports
(top10_ap <- top_ap %>% slice(1:10))
```

```
## # A tibble: 10 x 2
##    Airport                                          Total
##    <chr>                                            <int>
##  1 San Francisco International Airport              39744
##  2 Los Angeles International Airport                32459
##  3 Salt Lake City                                   14378
##  4 Phoenix, Phoenix Sky Harbor International Airport 13293
##  5 Fresno                                            9402
##  6 San Diego                                         8530
##  7 Denver International  Airport                     7783
##  8 San Jose                                          7073
##  9 Tucson  International Airport                     6445
## 10 Sacramento                                        5878
```

```
# create bar chart
ggplot(data = top10_ap, aes(x=reorder(Airport, Total), y=Total)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 10 Airports Traveled to by SkyWest Airlines",
       subtitle = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       caption = "The 10 airports in this graph account for the most frequent destination
       airports for SkyWest Airlines in the year 2019.",
       y = "Number of Trips",
       x = "Airport") +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
  theme(plot.caption.position = "plot", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
  theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black"))  +
  # scale_y_continuous(labels = scales::comma) +
  theme(axis.text = element_text(size=8, hjust=0.5, color="black")) +
  geom_text(aes(label = Total), size = 2.5, position = position_stack(vjust = 0.5))
```



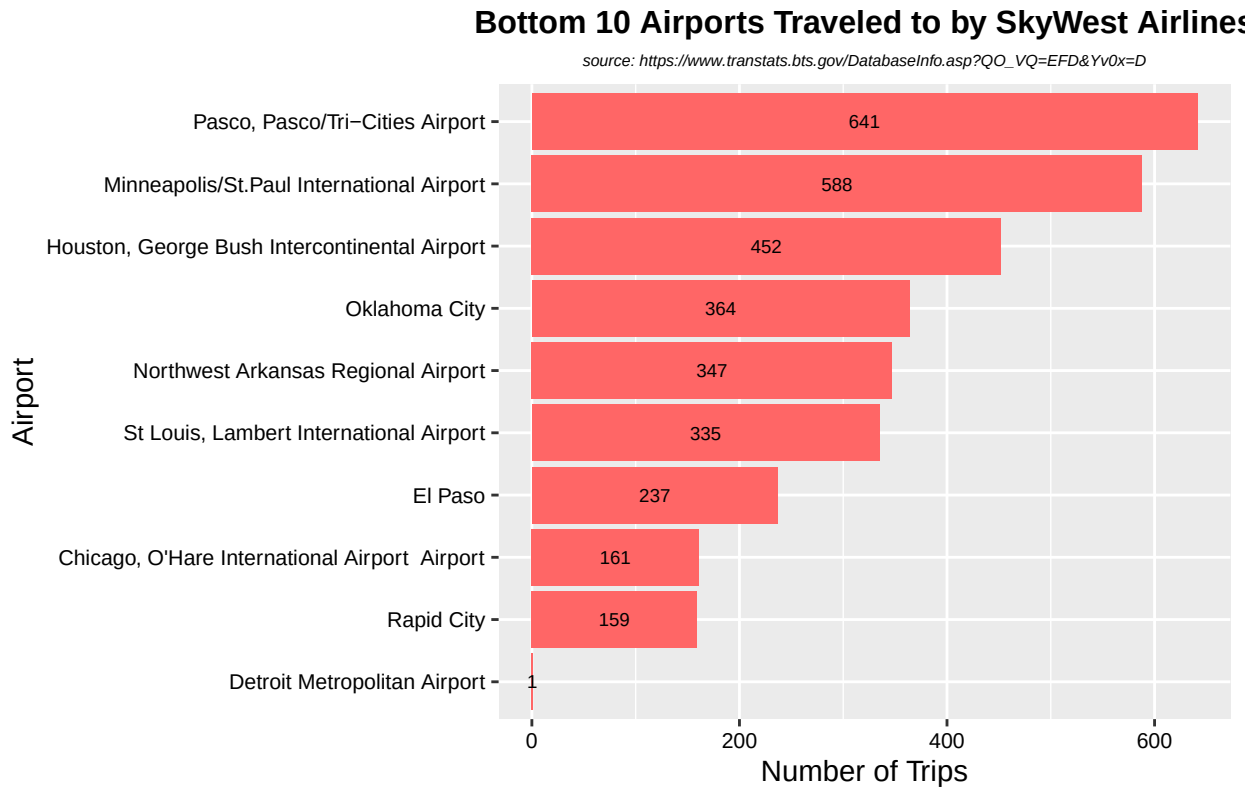
The 10 airports in this graph account for the most frequent destination airports for SkyWest Airlines in the year 2019.

```
# visualize bottom 10 airports
(bot10_ap <- top_ap %>% slice_tail(n=10))
```

```
## # A tibble: 10 x 2
##    Airport                                          Total
##    <chr>                                            <int>
##  1 Pasco, Pasco/Tri-Cities Airport                    641
##  2 Minneapolis/St.Paul International Airport          588
##  3 Houston, George Bush Intercontinental Airport      452
##  4 Oklahoma City                                      364
##  5 Northwest Arkansas Regional Airport                347
##  6 St Louis, Lambert International Airport            335
##  7 El Paso                                            237
##  8 Chicago, O'Hare International Airport  Airport     161
##  9 Rapid City                                         159
## 10 Detroit Metropolitan Airport                         1
```

```
# create bar chart
ggplot(data = bot10_ap, aes(x=reorder(Airport, Total), y=Total)) +
  geom_bar(stat = "identity", fill="#FF6666") +
  coord_flip() +
  labs(title = "Bottom 10 Airports Traveled to by SkyWest Airlines",
       subtitle = "source: https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFD&Yv0x=D",
       caption = "The 10 airports in this graph account for the least frequent destination
       airports for SkyWest Airlines in the year 2019.",
```

```
      y = "Number of Trips",
      x = "Airport") +
theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
theme(plot.caption.position = "plot", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
theme(plot.title=element_text(size=12, hjust=0.5, face="bold", color="black"))  +
# scale_y_continuous(labels = scales::comma) +
theme(axis.text = element_text(size=8, hjust=0.5, color="black")) +
geom_text(aes(label = Total), size = 2.5, position = position_stack(vjust = 0.5))
```



The 10 airports in this graph account for the least frequent destination airports for SkyWest Airlines in the year 2019.

The output of this graph tells us what we would expect: that the most frequented destination airports by SkyWest airlines are in California – San Francisco and Los Angeles. We also see that of these 10 airports, 6 of them are in California. Between this output and the data shown in the histogram, and other stats above, it seems that a plausible explanation is that the most frequented airports/states are within the region, and they are shorter distances. This is also holding up in the output for the *bottom* 10 airports, which shows airports in the midwest/southern states that were shaded in the US plot map.

Overall, it seems that the hypothesis that SkyWest Airlines is such a frequently used airline yet does not have as many traveling hours as other airlines could be explained by having more frequent flights to nearby states/airports, and based on the analysis performed, is very plausible.