# Data Challenge

Donovan Johnson

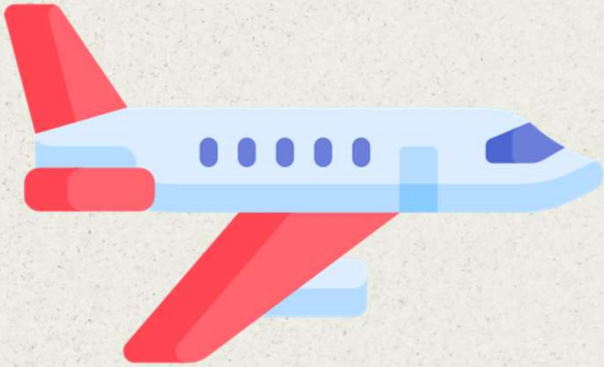
# SECTION

# 01.

Overview

# Problem Statement

Consulting for an airline company looking to enter the United States domestic market. The airline is looking to identify medium and large airports as their desired operating locations. The company believes that it has a competitive advantage in maintaining punctuality, so it plans on making this a big part of its brand image with a motto, “On time, for you.” To kick start operations, the company is seeking to start its entrance into the US market with 5 round trip routes.



# Data Set Overview

- **Airport_Codes**: Information on airports and includes airport code, city, country, and coordinates
  - Rows of data: 55,369
- **Flights**: Flights data for Q1 2019 and includes date, origin, destination, distance, flight number, and occupancy rate
  - Rows of data: 1,048,513
- **Tickets**: Sample tickets data for Q1 2019 and includes itinerary details and fare information.
  - Rows of data: 1,167,202

# Assumptions

**Costs:**

○ Fuel, Oil, Maintenance, Crew – $8 per mile total

○ Depreciation, Insurance, Other – $1.18 per mile total

○ Airport operational costs for the right to use the airports and related services are fixed at $5,000 for medium airports and $10,000 for large airports. There is one charge for each airport where a flight lands. Thus, a round trip flight has a total of two airport charges.

○ Delays that are 15 minutes or less are free, however each additional minute of delay costs the airline $75 in added operational costs. This is charged separately for both arrival and departure delays.

○ Each airplane will cost $90 million

**Revenue:**

○ Each plane can accommodate up to 200 passengers and each flight has an associated occupancy rate provided in the Flights data set. Do not use the Tickets data set to determine occupancy.

○ Baggage fee is $35 for each checked bag per flight. We expect 50% of passengers to check an average of 1 bag per flight. The fee is charged separately for each leg of a round trip flight, thus 50% of passengers will be charged a total of $70 in baggage fees for a round trip flight.

○ Disregard seasonal effects on ticket prices (i.e. ticket prices are the same in April as they are on Memorial Day or in December)

**Round trip:**

Each airplane is dedicated to one round trip route between the 2 airports

# SECTION

# 02.

Quality Check

# Issues with Data Set

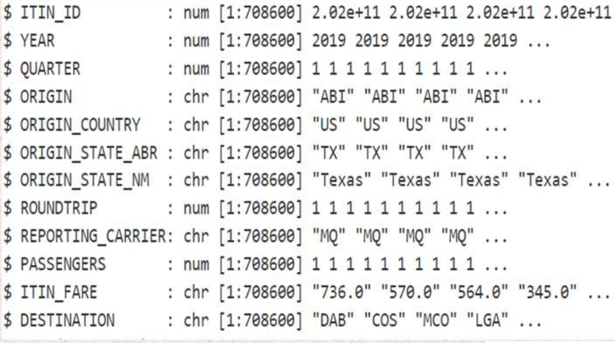
In order to have a tidy data collection, some critical quality issues with the dataset needed to be solved.

**Unfiltered Data**: The data sets have not been filtered, doing so will allow us to examine the columns that are relevant to our investigation.

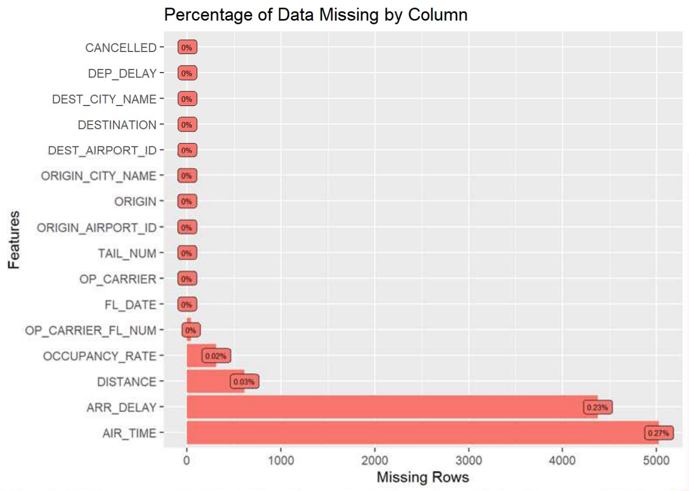
**Missing Data**: There are missing values in the data set's columns.

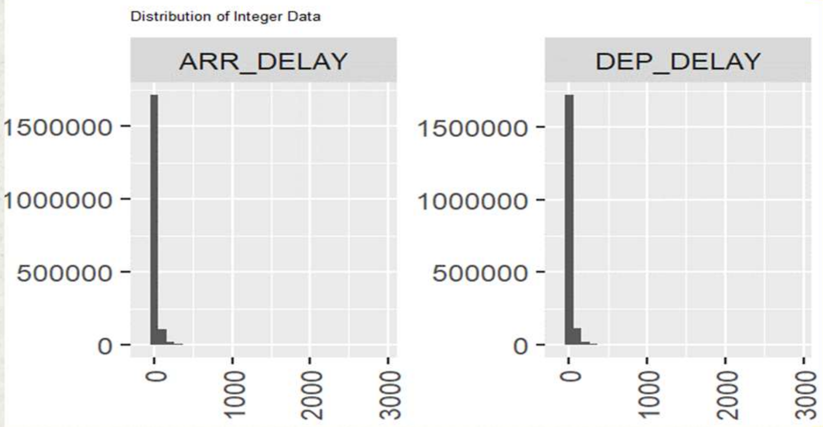
**Incorrect Data Types**: Some variables are in 'chr' format when they should be numeric.

**Outliers**: Existing outliers should be addressed in order to have more reliable data. Outliers can skew our results if not addressed.

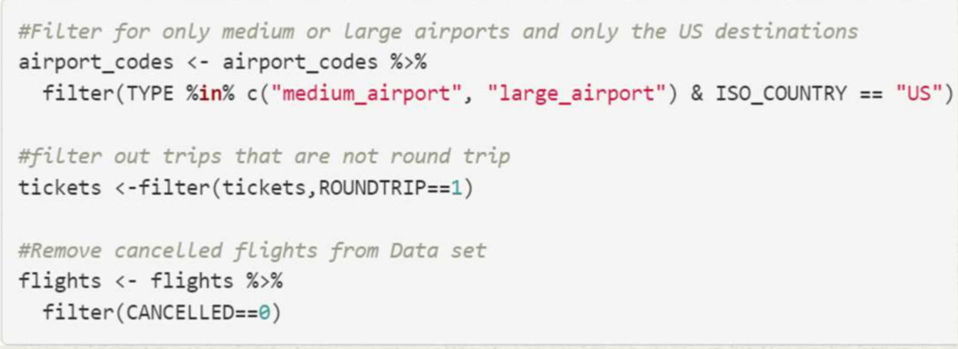


```
$ ITIN_ID          : num [1:708600] 2.02e+11 2.02e+11 2.02e+11 2.02e+11
$ YEAR             : num [1:708600] 2019 2019 2019 2019 2019 ...
$ QUARTER          : num [1:708600] 1 1 1 1 1 1 1 1 1 1 ...
$ ORIGIN           : chr [1:708600] "ABI" "ABI" "ABI" "ABI" ...
$ ORIGIN_COUNTRY   : chr [1:708600] "US" "US" "US" "US" ...
$ ORIGIN_STATE_ABR : chr [1:708600] "TX" "TX" "TX" "TX" ...
$ ORIGIN_STATE_NM  : chr [1:708600] "Texas" "Texas" "Texas" "Texas" ...
$ ROUNDTRIP        : num [1:708600] 1 1 1 1 1 1 1 1 1 1 ...
$ REPORTING_CARRIER: chr [1:708600] "MQ" "MQ" "MQ" "MQ" ...
$ PASSENGERS       : num [1:708600] 1 1 1 1 1 1 1 1 1 1 ...
$ ITIN_FARE        : chr [1:708600] "736.0" "570.0" "564.0" "345.0" ...
$ DESTINATION      : chr [1:708600] "DAB" "COS" "MCO" "LGA" ...
```





# Solutions to Data Set Issues



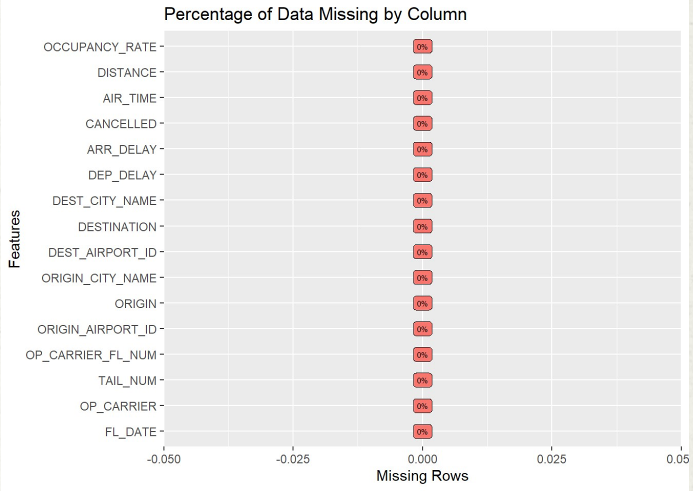
```
#Filter for only medium or large airports and only the US destinations
airport_codes <- airport_codes %>%
  filter(TYPE %in% c("medium_airport", "large_airport") & ISO_COUNTRY == "US")

#filter out trips that are not round trip
tickets <-filter(tickets,ROUNDTRIP==1)

#Remove cancelled flights from Data set
flights <- flights %>%
  filter(CANCELLED==0)
```

## Filter

Filter for medium or large aiports and US only in airport_codes data set.

Filter for only roundtrip flights in tickets data set.

Filter for non-cancelled flights in flights data set.

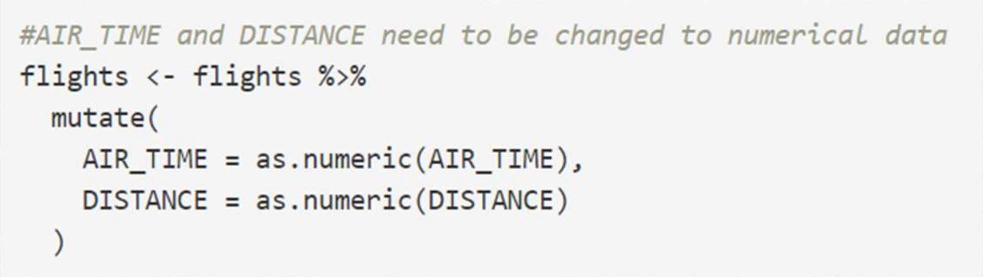


## Impute Missing Data

Impute NAs for integer variables using the median.

Delete NAs for character variables because imputing is not possible for the affected character variables.

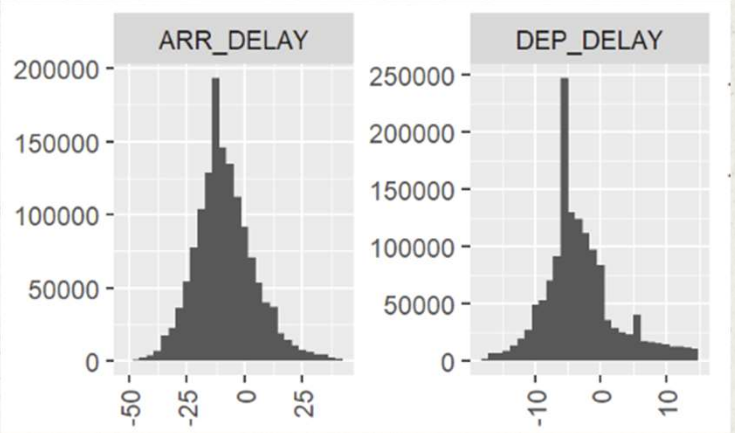
# Solutions to Data Set Issues Cont.



```
#AIR_TIME and DISTANCE need to be changed to numerical data
flights <- flights %>%
  mutate(
    AIR_TIME = as.numeric(AIR_TIME),
    DISTANCE = as.numeric(DISTANCE)
  )
```

## Transform to Numeric

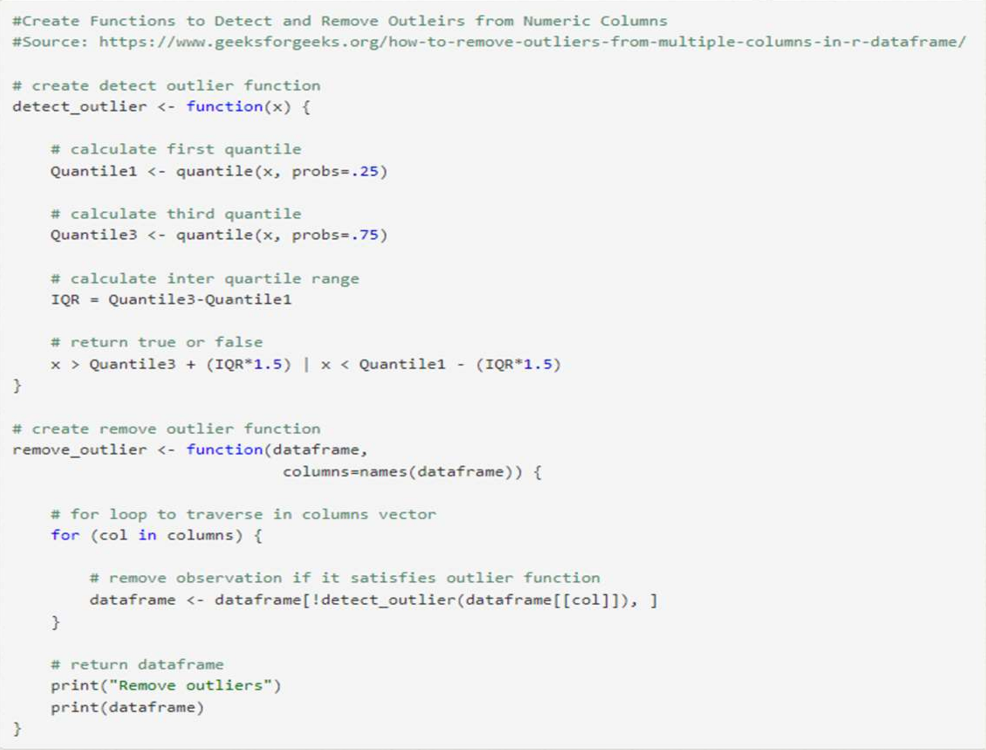
Variables of type character must be converted to numeric in order to be graphed and used in Exploratory Data Analysis.



## Remove Outliers

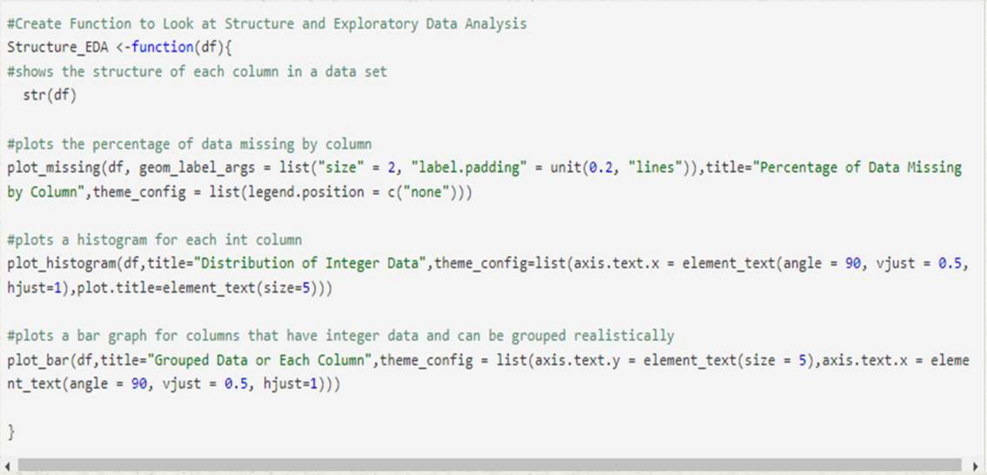
Remove Outliers that are not in the data range: (Q1-1.5*IQR , Q3+1.5*IQR). Any data point outside this range is considered as an outlier and is deleted.

# Function for EDA and Removing Outliers



```r
#Create Functions to Detect and Remove Outleirs from Numeric Columns
#Source: https://www.geeksforgeeks.org/how-to-remove-outliers-from-multiple-columns-in-r-dataframe/

# create detect outlier function
detect_outlier <- function(x) {

    # calculate first quantile
    Quantile1 <- quantile(x, probs=.25)

    # calculate third quantile
    Quantile3 <- quantile(x, probs=.75)

    # calculate inter quartile range
    IQR = Quantile3-Quantile1

    # return true or false
    x > Quantile3 + (IQR*1.5) | x < Quantile1 - (IQR*1.5)
}

# create remove outlier function
remove_outlier <- function(dataframe,
                            columns=names(dataframe)) {

    # for loop to traverse in columns vector
    for (col in columns) {

        # remove observation if it satisfies outlier function
        dataframe <- dataframe[!detect_outlier(dataframe[[col]]), ]
    }

    # return dataframe
    print("Remove outliers")
    print(dataframe)
}
```

EDA and Structures



```r
#Create Function to Look at Structure and Exploratory Data Analysis
Structure_EDA <-function(df){
#shows the structure of each column in a data set
  str(df)

#plots the percentage of data missing by column
plot_missing(df, geom_label_args = list("size" = 2, "label.padding" = unit(0.2, "lines")),title="Percentage of Data Missing
by Column",theme_config = list(legend.position = c("none")))

#plots a histogram for each int column
plot_histogram(df,title="Distribution of Integer Data",theme_config=list(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1),plot.title=element_text(size=5)))

#plots a bar graph for columns that have integer data and can be grouped realistically
plot_bar(df,title="Grouped Data or Each Column",theme_config = list(axis.text.y = element_text(size = 5),axis.text.x = eleme
nt_text(angle = 90, vjust = 0.5, hjust=1)))

}
```

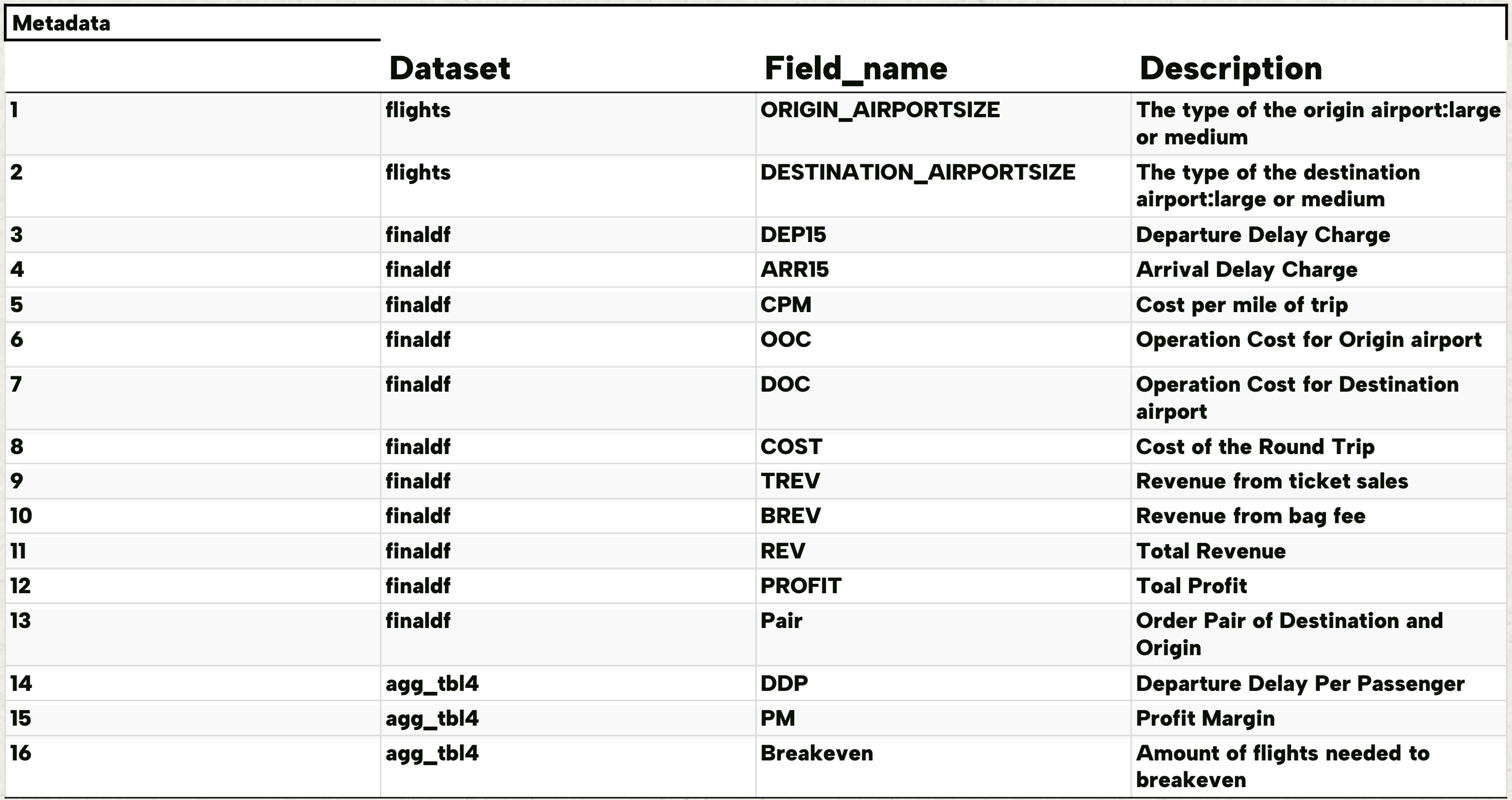
Remove Outliers

# SECTION

# 03.

Data Munging

# Meta Data



**Metadata**

| | Dataset | Field_name | Description |
|---|---|---|---|
| 1 | flights | ORIGIN_AIRPORTSIZE | The type of the origin airport:large or medium |
| 2 | flights | DESTINATION_AIRPORTSIZE | The type of the destination airport:large or medium |
| 3 | finaldf | DEP15 | Departure Delay Charge |
| 4 | finaldf | ARR15 | Arrival Delay Charge |
| 5 | finaldf | CPM | Cost per mile of trip |
| 6 | finaldf | OOC | Operation Cost for Origin airport |
| 7 | finaldf | DOC | Operation Cost for Destination airport |
| 8 | finaldf | COST | Cost of the Round Trip |
| 9 | finaldf | TREV | Revenue from ticket sales |
| 10 | finaldf | BREV | Revenue from bag fee |
| 11 | finaldf | REV | Total Revenue |
| 12 | finaldf | PROFIT | Toal Profit |
| 13 | finaldf | Pair | Order Pair of Destination and Origin |
| 14 | agg_tbl4 | DDP | Departure Delay Per Passenger |
| 15 | agg_tbl4 | PM | Profit Margin |
| 16 | agg_tbl4 | Breakeven | Amount of flights needed to breakeven |

# Join Data

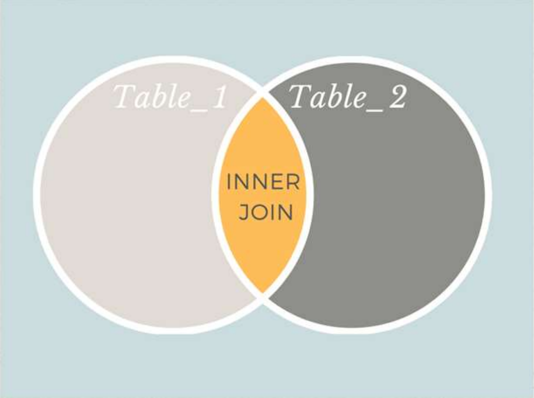
Initially, non-relevant ORIGIN and DESTINATION values, pertaining to airports of sizes other than medium or large, were excluded from the flights dataset.

Subsequently, the dataset was expanded with the inclusion of two supplementary columns, namely ORIGIN_AIRPORTSIZE and DESTINATION_AIRPORTSIZE. These additions serve the purpose of denoting whether the respective ORIGIN or DESTINATION airport falls within the medium or large category.

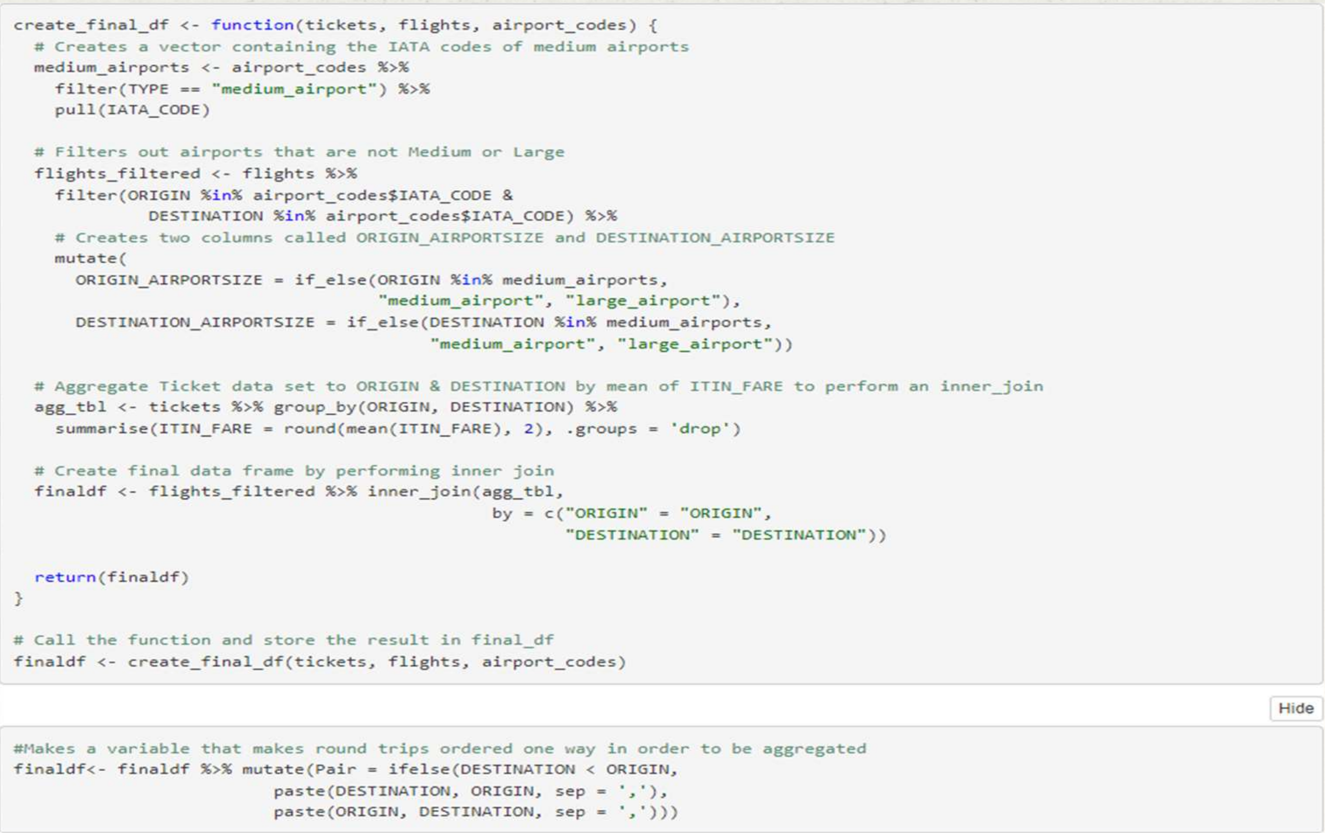
To optimize the analytical process, an alternative approach was adopted wherein ticket data was aggregated based on the ORIGIN and DESTINATION parameters. The mean value of the ITIN_FARE variable was the sole focus, rendering the comprehensive array of variables associated with the flights dataset unnecessary for this specific analysis.

The integration of the tickets and flights datasets was achieved through an inner join operation, conducted with respect to the ORIGIN and DESTINATION attributes. This selection was made to automatically refine the dataset, retaining exclusively the entries corresponding to medium and large airports from the tickets dataset, as well as excluding non-round trip entries from the flights dataset.

Concluding the data manipulation process, a new column was generated to combine the DESTINATION and ORIGIN values, arranging them in alphabetical order. This additional feature was devised to establish a distinctive identifier for round trips that share the same route but initiate from distinct origins.



# Data Munging Code



```
create_final_df <- function(tickets, flights, airport_codes) {
  # Creates a vector containing the IATA codes of medium airports
  medium_airports <- airport_codes %>%
    filter(TYPE == "medium_airport") %>%
    pull(IATA_CODE)

  # Filters out airports that are not Medium or Large
  flights_filtered <- flights %>%
    filter(ORIGIN %in% airport_codes$IATA_CODE &
             DESTINATION %in% airport_codes$IATA_CODE) %>%
    # Creates two columns called ORIGIN_AIRPORTSIZE and DESTINATION_AIRPORTSIZE
    mutate(
      ORIGIN_AIRPORTSIZE = if_else(ORIGIN %in% medium_airports,
                                   "medium_airport", "large_airport"),
      DESTINATION_AIRPORTSIZE = if_else(DESTINATION %in% medium_airports,
                                        "medium_airport", "large_airport"))

  # Aggregate Ticket data set to ORIGIN & DESTINATION by mean of ITIN_FARE to perform an inner_join
  agg_tbl <- tickets %>% group_by(ORIGIN, DESTINATION) %>%
    summarise(ITIN_FARE = round(mean(ITIN_FARE), 2), .groups = 'drop')

  # Create final data frame by performing inner join
  finaldf <- flights_filtered %>% inner_join(agg_tbl,
                                              by = c("ORIGIN" = "ORIGIN",
                                                     "DESTINATION" = "DESTINATION"))

  return(finaldf)
}

# Call the function and store the result in final_df
finaldf <- create_final_df(tickets, flights, airport_codes)
```

```
#Makes a variable that makes round trips ordered one way in order to be aggregated
finaldf<- finaldf %>% mutate(Pair = ifelse(DESTINATION < ORIGIN,
                         paste(DESTINATION, ORIGIN, sep = ','),
                         paste(ORIGIN, DESTINATION, sep = ',')))
```

# SECTION

# 04.

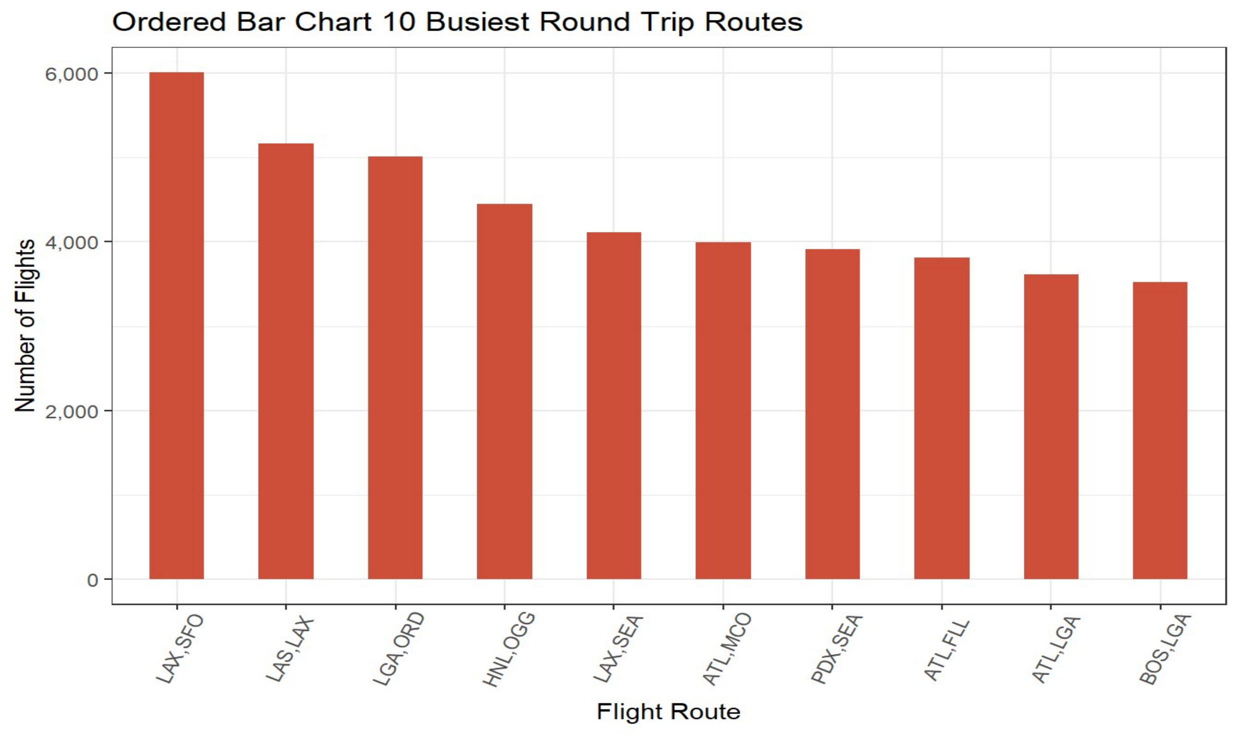
Craft a Visual Data Narrative

# 1. The 10 Busiest Round Trips

The most extensively traveled round trip route was identified as (LAX, SFO), comprising 6014 round trips during the initial quarter of 2019.

Airports LAX, LGA, and ATL prominently feature, each accounting for three instances among the busiest round trips during the same quarter of 2019.

Regions that present noteworthy potential for our client encompass California, New York, and Georgia.

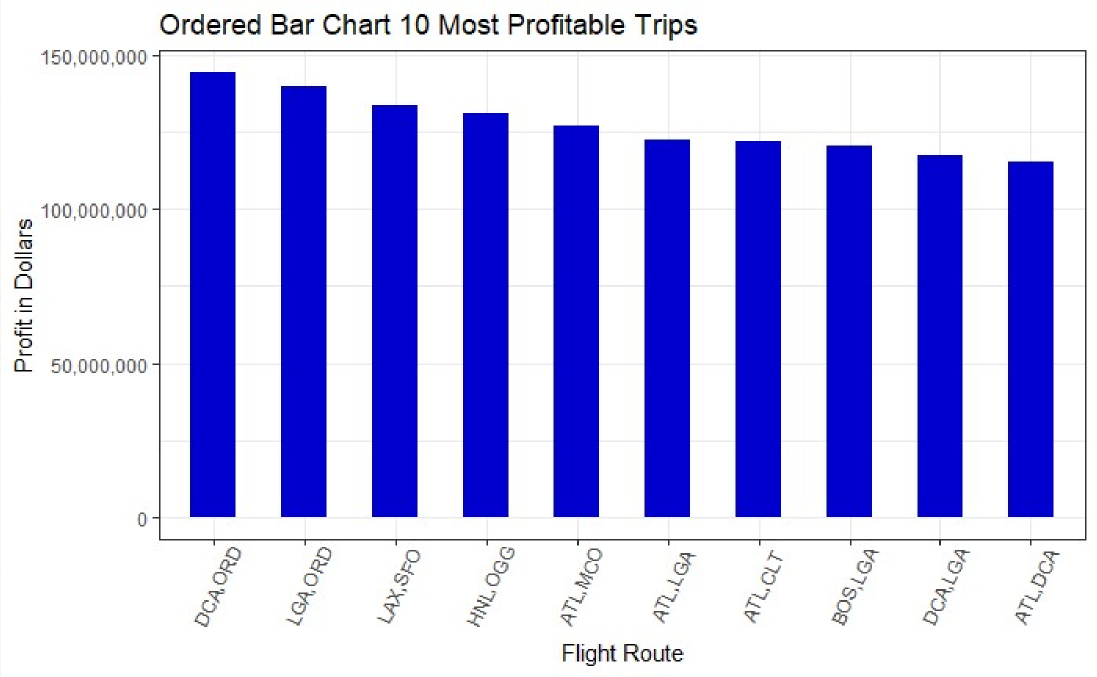


# 2. The 10 Most Profitable Round Trips

The leading round trip flight in terms of profitability within the United States during Q1 2019 is unequivocally identified as (DCA, ORD), generating a substantial total profit of $144,310,308.08 for the respective airliners.

Following closely, the (LGA, ORD) and (LAX, SFO) routes claim the positions of second and third most profitable trips, consecutively.

Notably, an observation can be made that the decline in profitability subsequent to the initial top three most lucrative trips is not significant.

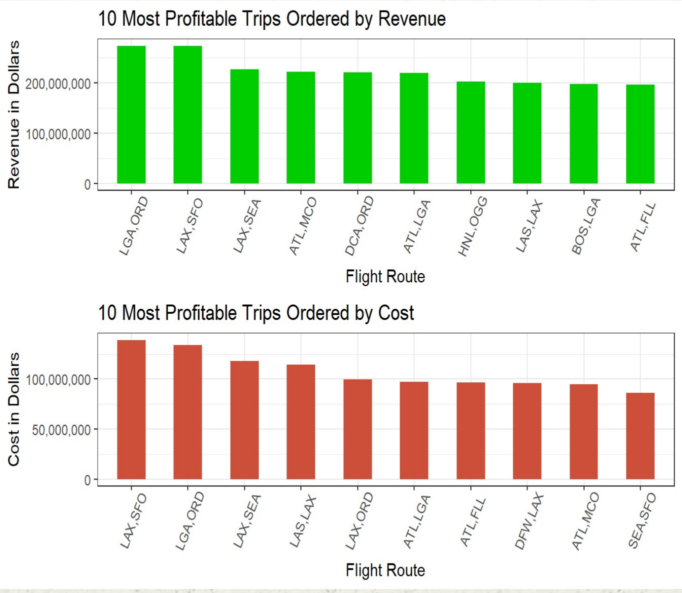


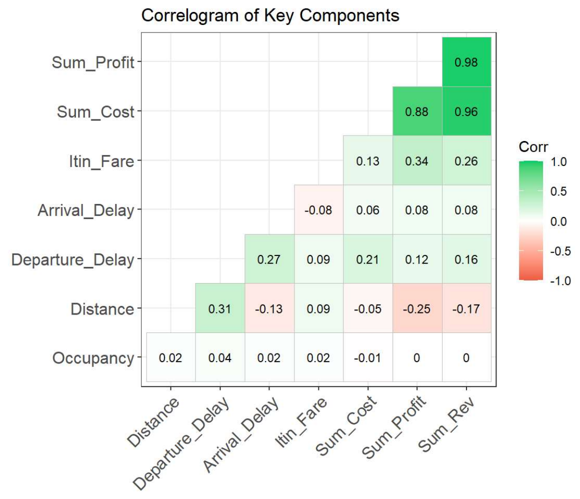
# 2. The 10 Most Profitable Round Trips cont.

Among the top ten most profitable flight routes, (LGA, ORD) emerges as the highest revenue generation.

Conversely, within the same set of top ten profitable routes, (LAX, SFO) exhibits the highest incurred costs.

An apparent correlation surfaces, suggesting that elevated itinerary fare rates coupled with abbreviated flight durations correlate with heightened profitability.




Correlogram of Key Components

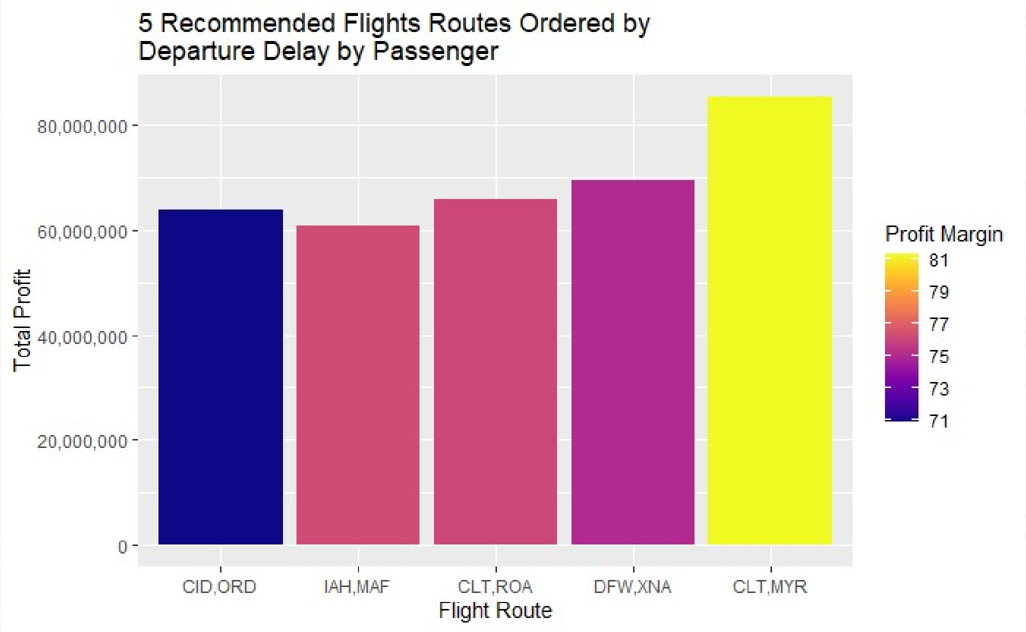
| | Distance | Departure_Delay | Arrival_Delay | Itin_Fare | Sum_Cost | Sum_Profit | Sum_Rev |
|---|---|---|---|---|---|---|---|
| Sum_Profit | | | | | | | 0.98 |
| Sum_Cost | | | | | | 0.88 | 0.96 |
| Itin_Fare | | | | | 0.13 | 0.34 | 0.26 |
| Arrival_Delay | | | | -0.08 | 0.06 | 0.08 | 0.08 |
| Departure_Delay | | | 0.27 | 0.09 | 0.21 | 0.12 | 0.16 |
| Distance | | 0.31 | -0.13 | 0.09 | -0.05 | -0.25 | -0.17 |
| Occupancy | 0.02 | 0.04 | 0.02 | 0.02 | -0.01 | 0 | 0 |

# 3. Five Recommended Round Trips

The selection of these specific five round-trip flights is constructed by meticulous criteria:

They encapsulate the 95th percentile of profit as well as the 90th percentile of profit margin. This combined evaluation not only signifies immediate profitability but also underlines the potential for enduring sustainable profits over the long term.

Additionally, a meticulous sorting approach has been ordered to prioritize flights with the best departure delay per passenger. This alignment with the company's business ethos of "On time, for you" shows a commitment to punctuality and customer satisfaction.

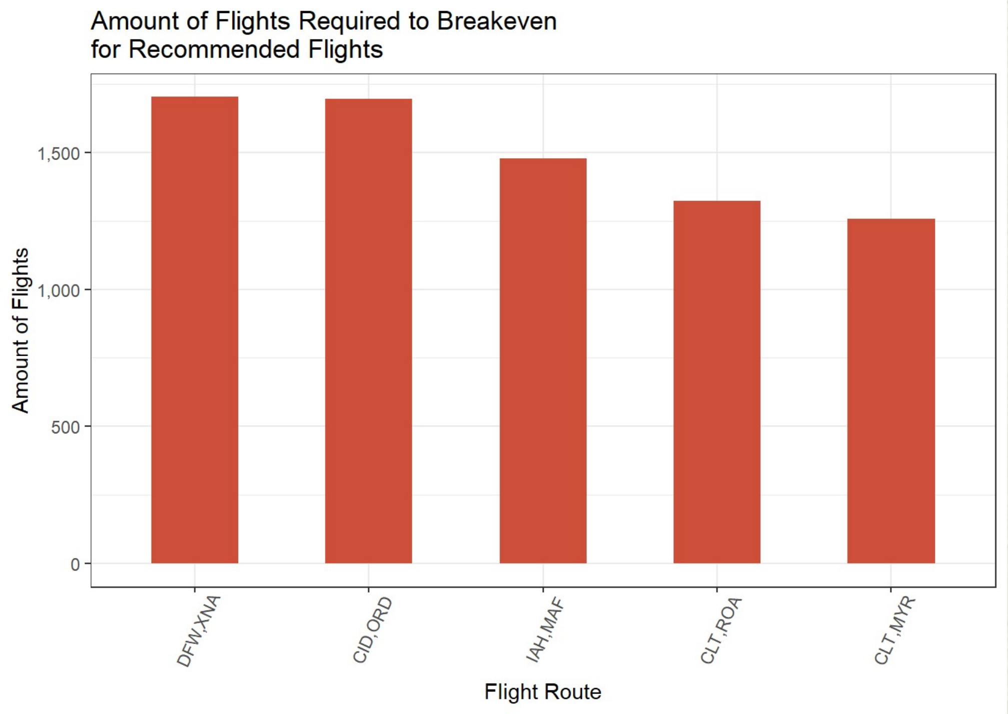


# 4. Number of round trips flight it will take to break even

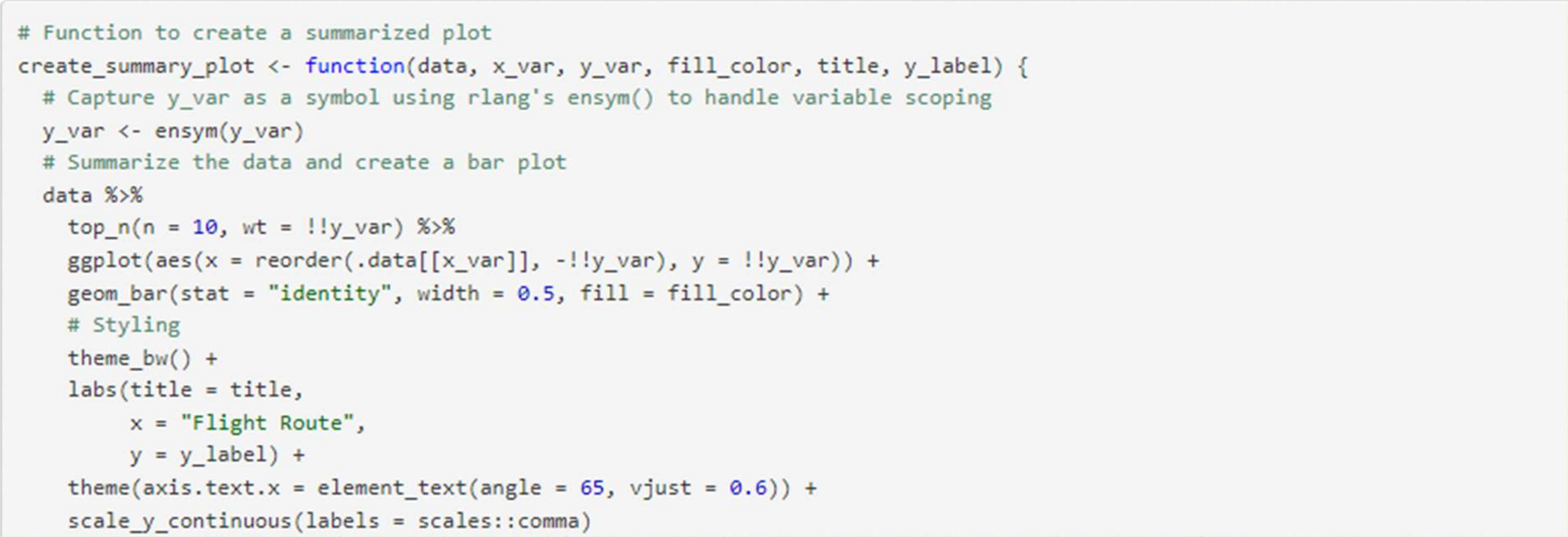
Drawing insights from my collection of five recommended round-trip routes, a comprehensive overview of the associated breakeven costs is as follows:

Impressively, the (CLT, MYR) route stands out for its efficiency in attaining the breakeven point, necessitating the fewest flights to achieve this critical milestone.

Worth highlighting is the remarkable performance of the very same route, (CLT, MYR), which not only secures the position of highest profitability but also boasts the most robust profit margin among the selections. This strategic alignment shows its potential as a lucrative avenue for sustainable growth and financial success.



# Function to Create Summarized Plot



```r
# Function to create a summarized plot
create_summary_plot <- function(data, x_var, y_var, fill_color, title, y_label) {
  # Capture y_var as a symbol using rlang's ensym() to handle variable scoping
  y_var <- ensym(y_var)
  # Summarize the data and create a bar plot
  data %>%
    top_n(n = 10, wt = !!y_var) %>%
    ggplot(aes(x = reorder(.data[[x_var]], -!!y_var), y = !!y_var)) +
    geom_bar(stat = "identity", width = 0.5, fill = fill_color) +
    # Styling
    theme_bw() +
    labs(title = title,
         x = "Flight Route",
         y = y_label) +
    theme(axis.text.x = element_text(angle = 65, vjust = 0.6)) +
    scale_y_continuous(labels = scales::comma)
```

# 5. KPI to track for recommended flights

Aligned with the company's motto of "On time, for you," the selection of key performance indicators (KPIs) reflects a strategic focus on punctuality and client satisfaction.

**Delay Reasons:** By meticulously tracking causes of delays, valuable insights are gained into recurrent issues affecting specific flight routes. Addressing these factors enhances the ability to optimize operations for timely departures.

**Cancellation Rates:** Similar to the approach taken with Delay Reasons, a comprehensive understanding of cancellations aids in devising strategies to minimize their occurrence. Identifying cancellation triggers empowers the organization to bolster operational reliability.

**Arrival Punctuality:** Calculating the ratio of delayed flights shows the organization's commitment to upholding a superior standard of passenger contentment through consistently punctual flights.

**Satisfied Passenger Quota:** The measurement of the proportion of passengers arriving on time reinforces the company's unwavering dedication to ensuring the highest levels of customer satisfaction throughout their travel experience.

**Jet Fuel Price Index as Cost Percentage:** Given the substantial impact of jet fuel costs on operations, vigilant monitoring of this index is imperative. The volatility of fuel prices, influenced by external factors, emphasizes the importance of meticulous management of this expense, which can constitute a substantial portion of operating costs, ranging between 30% to 60%.

# SECTION

# 05.

Final Recommendations

# Final Recommendations

The recommended selection of five round-trip flight options includes:

1. **(DFW, XNA)**
2. **(CID, ORD)**
3. **(IAH, MAF)**
4. **(CLT, ROA)**
5. **(CLT, MYR)**

These choices are grounded in three fundamental metrics: Profit, Profit Margin, and Departure Delay per Passenger. While the incorporation of the first two metrics is inherent to informed decision-making, signifying the bedrock of profitability and financial viability for the client's entry into the US market, the addition of the third metric reflects the client's unwavering commitment to ensuring punctual passenger transportation. This metric aligns with the client's conviction in holding a competitive edge and, thus, highlights the significance of prioritizing flight routes based on departure delay.

# SECTION

# 06.

What’s Next?

# What’s Next

To optimize our analysis, we can begin by establishing a test and training set from our substantial and reliable data collection. This will enable us to leverage supervised machine learning techniques to uncover the major influences on DEP_Delay, ARR_Delay, Profit, and Profitability within the dataset.

Expanding our analytical arsenal, geographic visualizations present a valuable avenue. By employing these visualizations, we can delve deeper into groups of round-trip flights, potentially revealing geographical patterns that correlate with higher profitability.

Another pivotal step involves delving into our cancelled flight dataset to discern strong correlations with variables that contribute to flight cancellations. This analysis holds the key to understanding the determinants of this operational challenge.

Considering the potential benefits of extended timeframes, integrating machine learning techniques for imputing missing data becomes an intriguing proposition. Although computationally demanding, opting for median imputation offers a robust solution to address data gaps.

Concluding our strategic considerations, a nuanced approach is at play when utilizing the 95th percentile for profit and the 90th percentile for profit margin. In a departure from traditional ascending arrangements, our unique approach involves sorting recommendations in descending order based on departure delay per passenger. This approach aims to focus on markets where flights have a high profit/profit margin
per roundtrip but will sort our recommendation by routes that have the most delays.