A

Group Project Report on

Analyzing the flight delays in United States and providing insights to passengers for better planning their travel

By
(Team-Y)

Chandu Yerragopu
Saran Prasad Balasubramaniam
Wen Xie

Under the esteemed guidance of

Dr. Jaeki Song, Ph.D.
Course: ISQS 6337 Scripting Languages

Table of contents

1 Introduction

Travel industry is one of the rapidly growing industry in 21st century. Air-travel is one of the widely used means of commuting these days and especially when you consider a vast country like united states. Although Air travel is the fastest means of transport, we often have problem with delay of the air-crafts for various reasons. This project titled "Analyzing the flight delays in United States and providing insights to passengers for better planning their travel" is aimed at finding solutions to such problems and better assist the passenger in choosing a right flight on a right day and time and based on the origin and destination airport.

We started our project by asking some of the common questions any passenger would generally ask before he/she plans his/her travel. Then we have spent efforts in identifying the suitable and accurate data available which will help us answer the questions that we initially planned to answer. In the next sections of this document, we explained where we obtained data from, how data looks like initially, what are the planned steps for analysis and lists various visualizations in the process of answering these questions.

We have primarily used the two very popular open source programming languages for Data Analysis namely R and Python. Python is mainly used for the initial steps such as data cleaning and pre-processing. Later using R, various visualizations were plotted which gives a clear idea of the trends in data by looking at a picture. By drawing inferences from the visualization, we tried to relate them to the real world situations which seems to be correct in most of the cases.

2 Data, Analysis and Results

2.1 Data Source

One of the primary and important step in answering any problem is to identify the right data for the same. In our project, we have obtained data from the below source:

https://catalog.data.gov

Below are the brief details about the data.

- Type of data file: .csv
- Size: 260 MB
- No.of attributes/columns: 30
- No.of records: 104857

2.2 Description of Data

Column Name	Description
Year	Time related attributes
Month	-do-
DayofMonth	-do-
DayOfWeek	-do-
DepTime	Actual Departure time
CRSDepTime	Computer Reservation System Departure Time
ArrTime	Actual Arrival Time
CRSArrTime	Computer Reservation System Arrival Time
UniqueCarrier	Unique carrier code
FlightNum	Flight Number
TailNum	Tail Number of the flight
ActualElapsedTime	Duration of Flight
CRSElapsedTime	Scheduled Duration of Flight
AirTime	Time of flight in air
ArrDelay	Delay in Arrival
DepDelay	Delay in Departure
Origin	Origin Airport Code
Dest	Destination Airport Code
Distance	
TaxiIn	

TaxiOut	
Cancelled	Whether the flight is cancelled
CancellationCode	
Diverted	
CarrierDelay	
WeatherDelay	
NASDelay	National Airspace System Delay
SecurityDelayn	
LateAircraftDelay	

2.3 Assumptions

- Monday is taken as the first day of the week.
- As there is no data about the number of passengers on a given day, inferences related to that part were not covered.
- The data is obtained from the data.gov , a US government website and we believe the data is true and accurate to be used in analysis.

2.4 Initial Steps

The first step in analyzing any data is to start understanding the attributes and identifying the required columns to use in our analysis. Here is the list of steps followed for analysis:

a) Understanding the dataset

b) Cleaning the dataset

c) Identifying the required columns for analysis and suitable models

2.4.1 Understanding the dataset

The dataset has around 30 attributes of various kind. We have gone through the data to identify how exactly data looks like. We are interested in answering the questions on two broad categories such as based on time and based on airports.

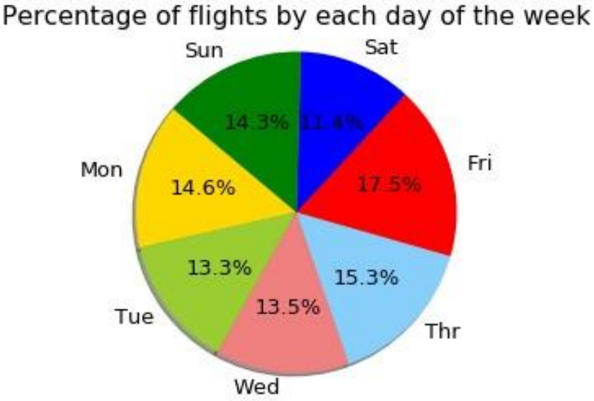
2.4.2 Identifying the required columns for analysis and cleaning the dataset

From the list of columns available in the data, we have obtained a subset of the data by slicing and remove the unrequired columns using Python. Below is the list of columns used for our analysis:

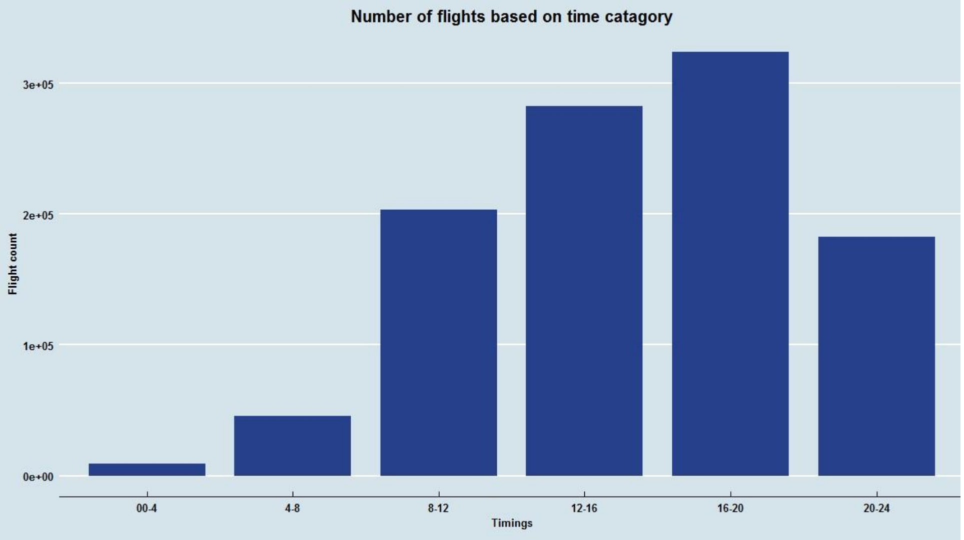
Year, Month, DayofMonth, DayOfWeek, DepTime, ArrTime, ActualElapsedTime, CRSElapsedTime , ArrDelay, DepDelay, Origin, Dest, CarrierDelay, WeatherDelay NASDelay, SecurityDelay, LateAircraftDelay

2.5 Data Analysis and answering of various questions which will collectively help in better decision making

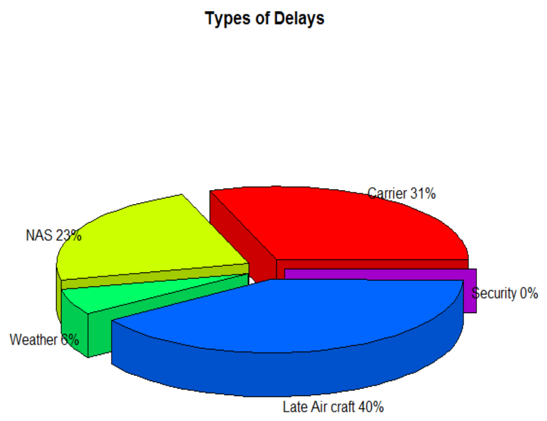
1. How many flights were there for each weekday?



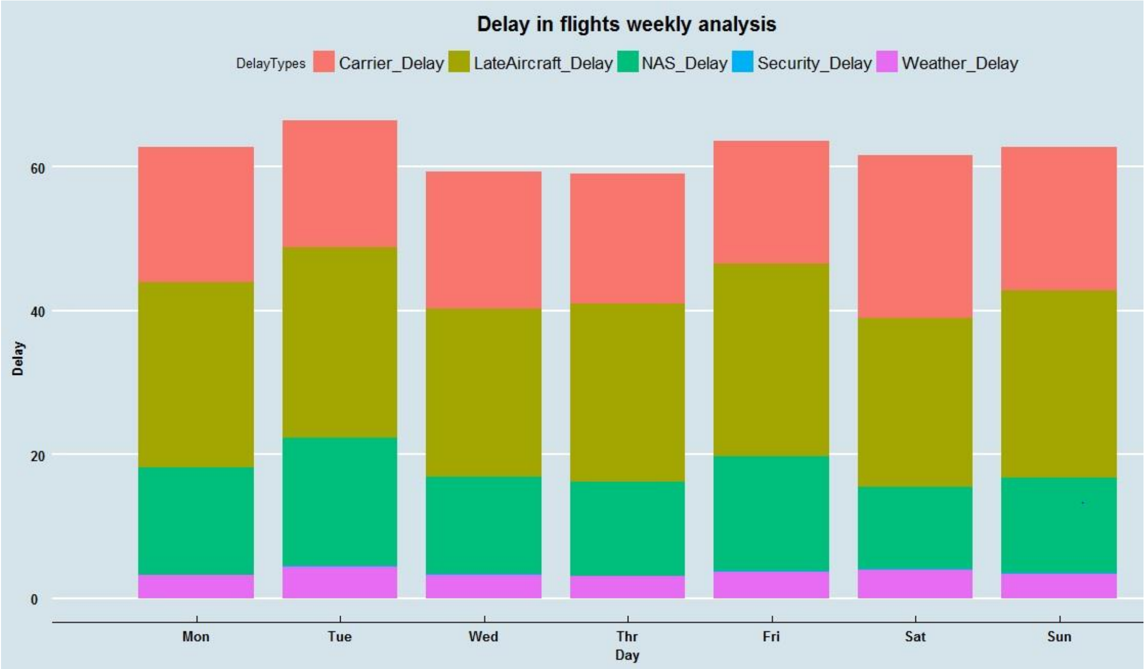
2. Which time of the day has more no.of flights?



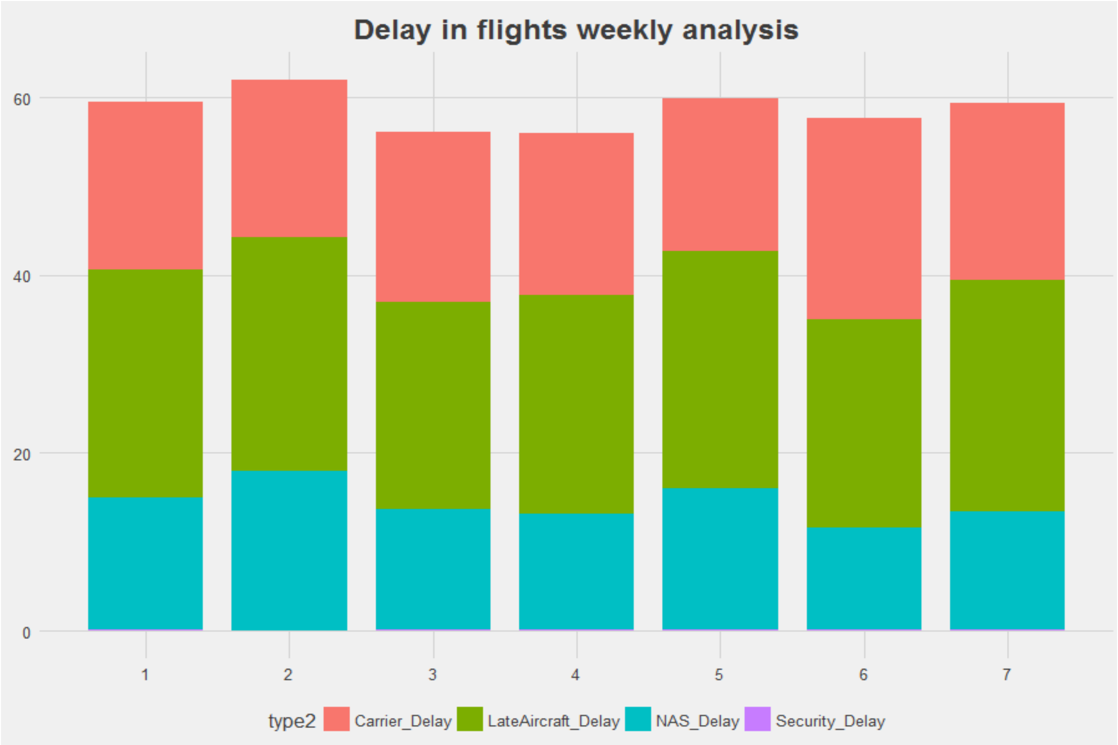
3. What are the variables and how much they contribute to delays?



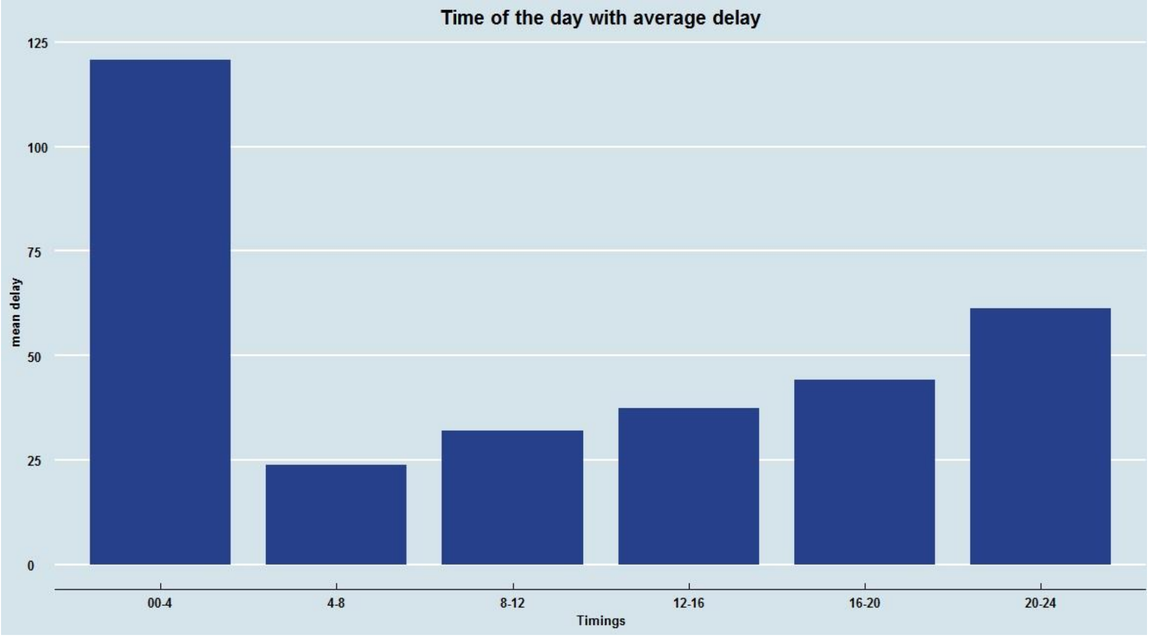
4. Which weekday had the highest proportion of flights that were late and due to what type of delay?



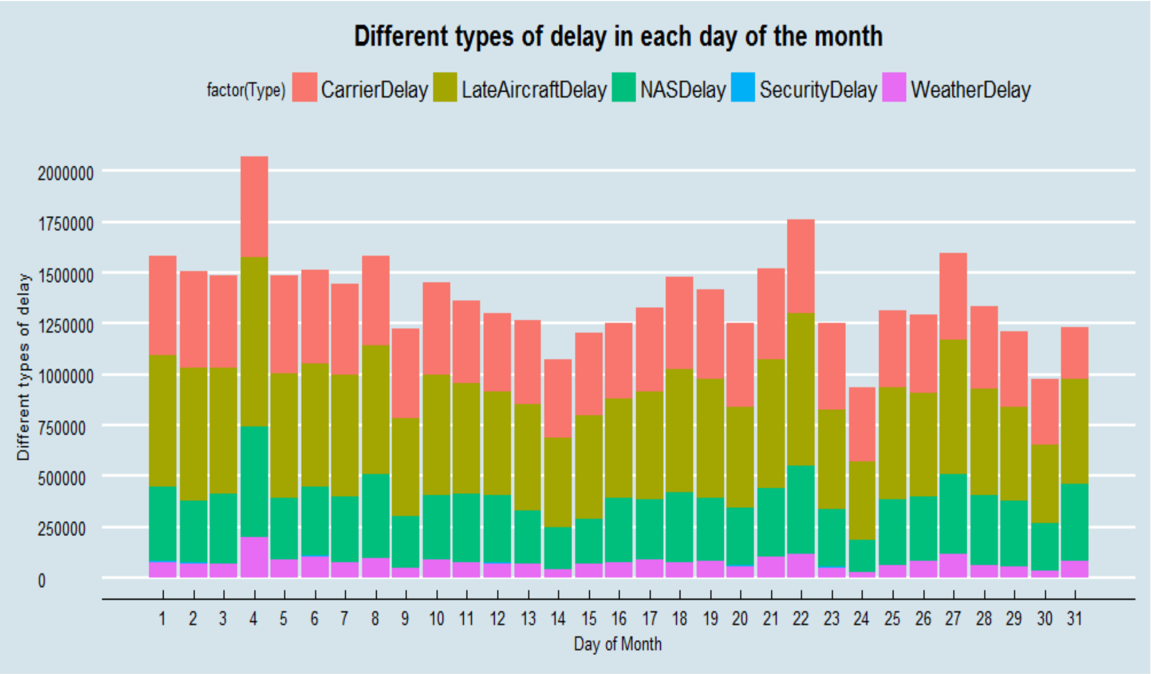
5. Which day of the week is best for minimizing departure delays for customers?



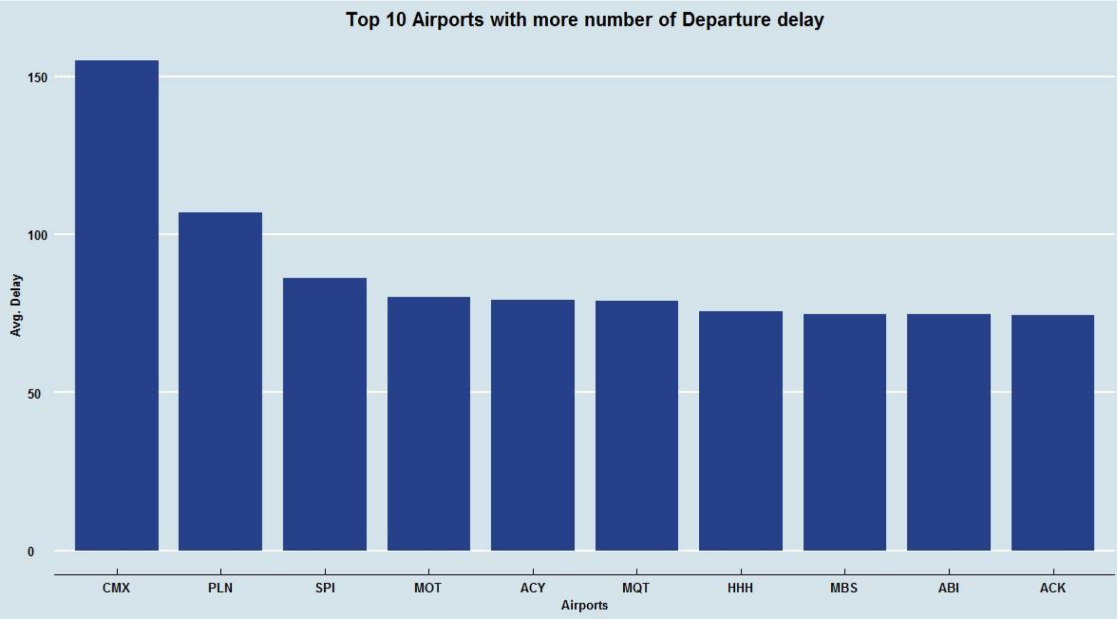
6. Which time of day is best for minimizing the departure delay?



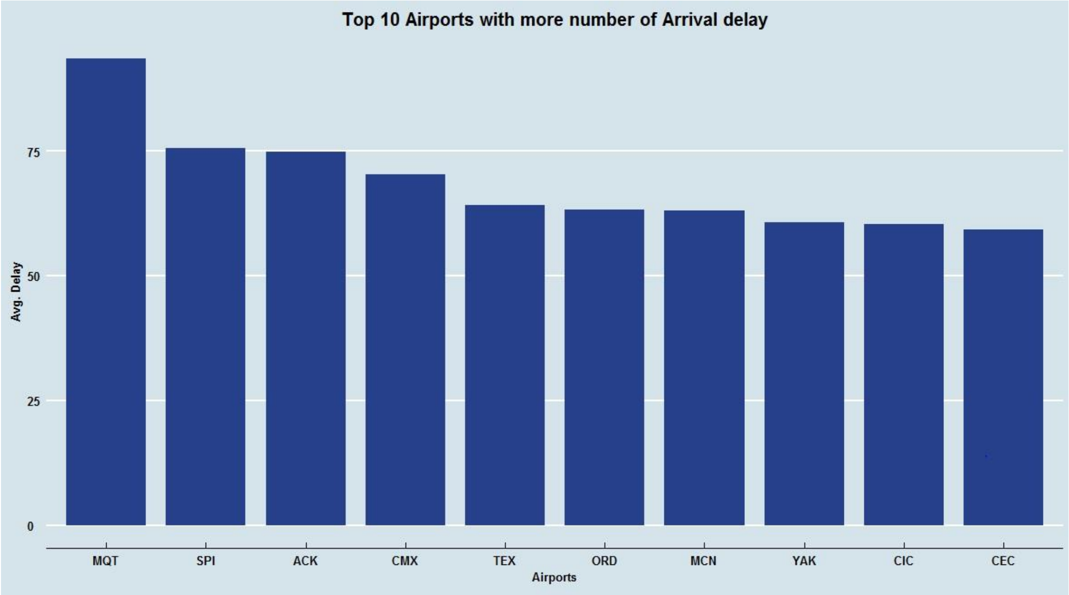
7. Which is the best day of the month to fly?



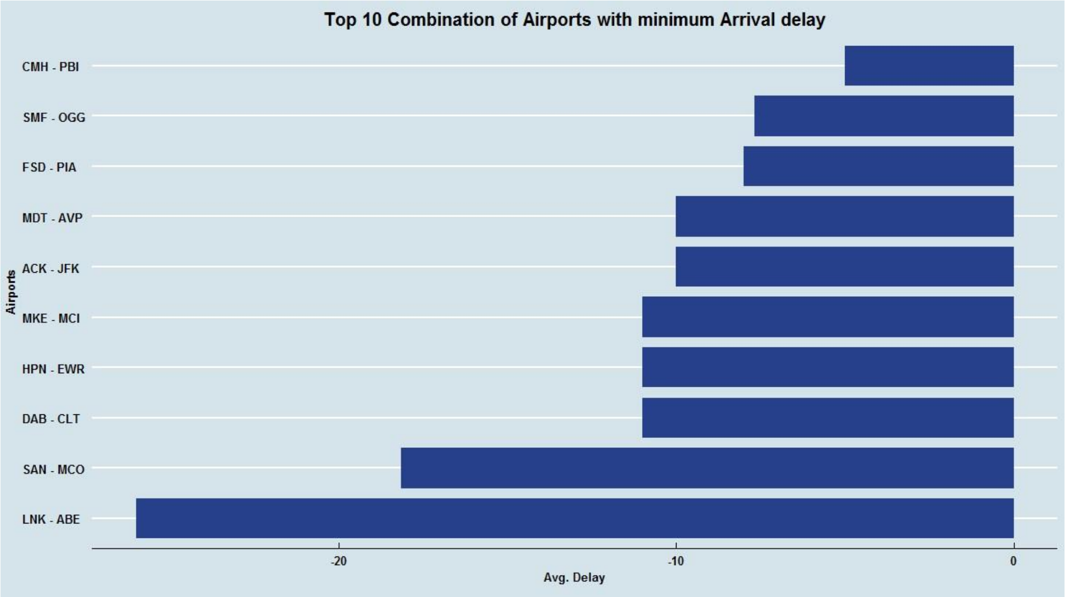
8. What is the list of Top 10 airports in US with more departure delay?



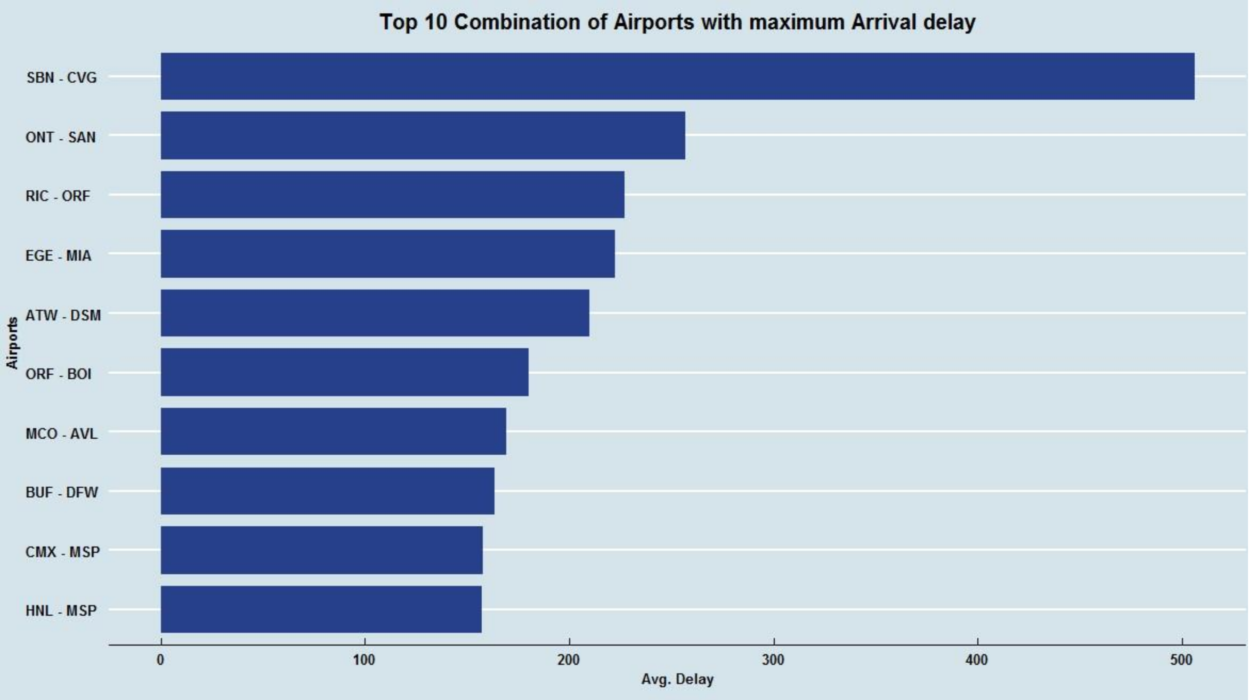
9. What is the list of Top 10 airports in US with more Arrival delay?



10. Which combination of arrival and departure airports is associated with the minimum delays?



11. Which combination of arrival and departure airports is associated with the maximum delays?


Top 10 Combination of Airports with maximum Arrival delay
Airports
SBN - CVG
ONT - SAN
RIC - ORF
EGE - MIA
ATW - DSM
ORF - BOI
MCO - AVL
BUF - DFW
CMX - MSP
HNL - MSP
Avg. Delay