Zewail City of Science and Technology University of Science and Technology CIE 427- fall 2021 

Flight Records Analysis

Asmaa Ismail 201600780

Anhar Hassan 201601171

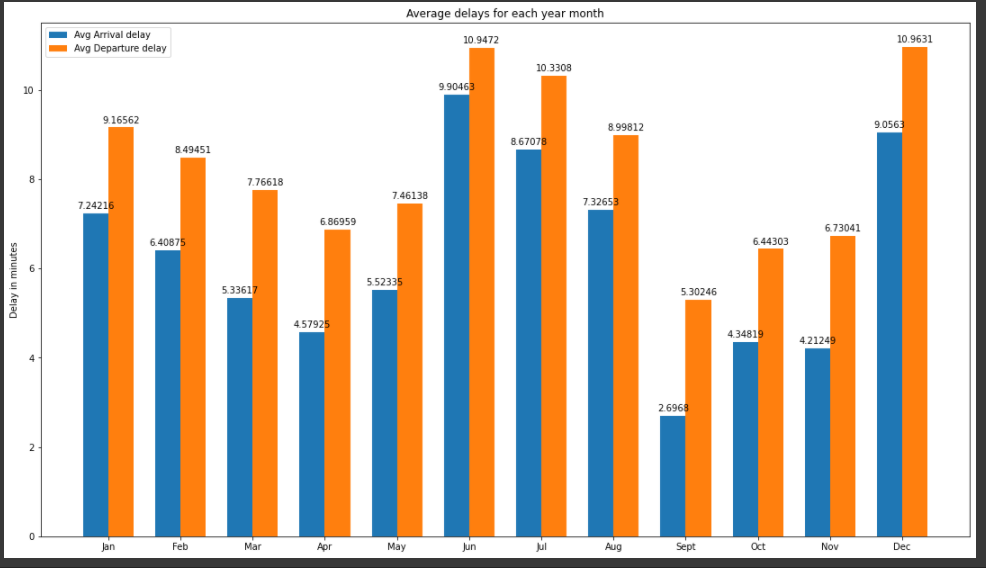
Ahmed Elgabry 20170100

# ***\* Note: a better quality version of all visualizations can be found in visualisations.ipynb in the*** [***project’s drive***](https://drive.google.com/drive/folders/1lYn4RJa2gngVxKmW0Col-vgad7GK8CHG?usp=sharing)***.***

# *Question one:* what is the best month, day of month/ of week to minimize delay?

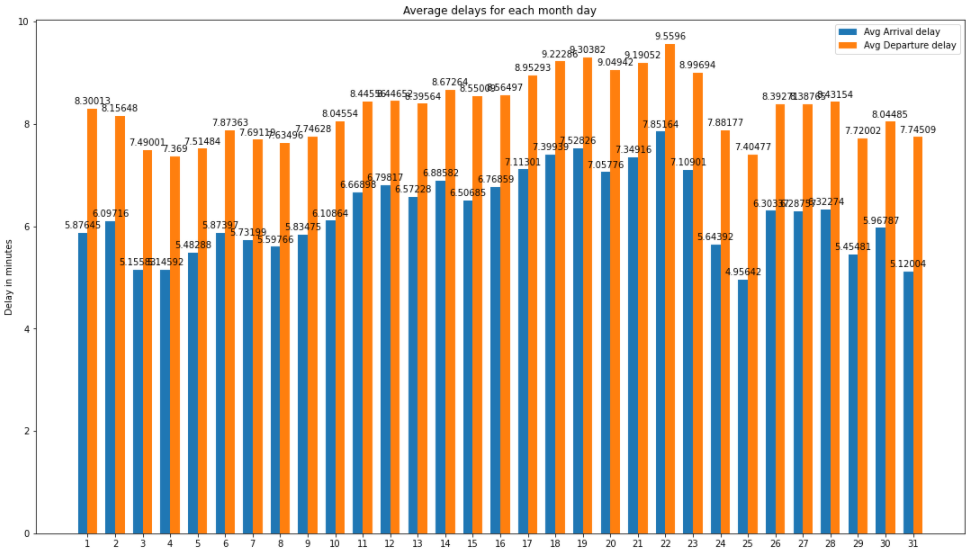
**The following figures contains a distribution of average delays across: months of year, day of month, and day of week**

**1- Delay Per Month:**

****

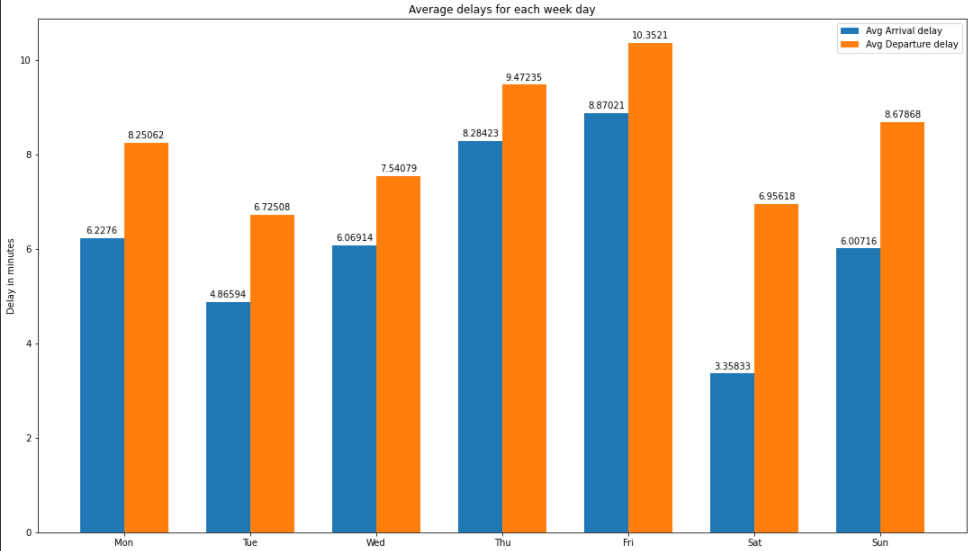
***Figure 1: the average arrival and departure delay for each month***

**2- Average Delays For each month day**

****

***Figure 2: the average arrival and departure delay for each day on the month***

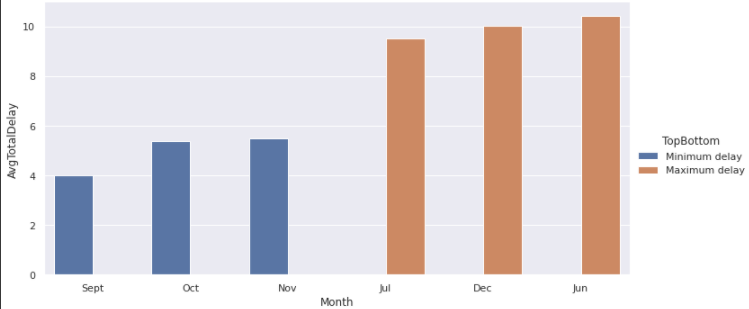
**3- average delays for each week day**

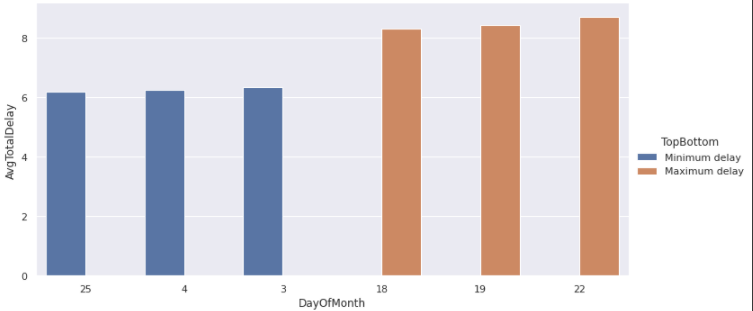
****

***Figure 3: the average arrival and departure delay for each day of the week***

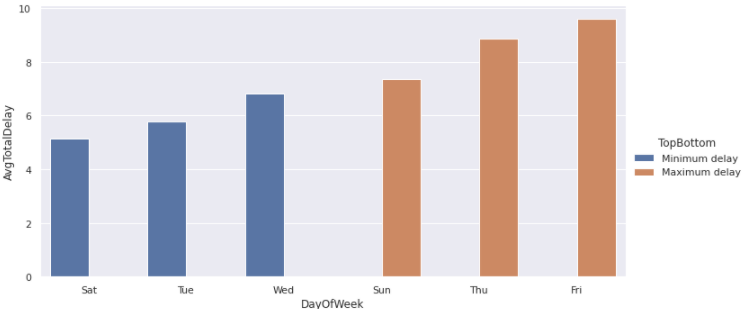
**The Month, day of month, and week day associated with the most and least delay:**

The following figures describe the top and bottom 3 month or day in terms of delay.

***Figure 4: top and bottom 3 months of year in terms of average delay***

****

***Figure 5: top and bottom 3 days of the month in terms of average delay***

***Figure 6: top and bottom 3 days of week in terms of average delay***

**We noticed that**

**⇒** The best month to travel in to minimize delay is September while the worst months (almost all the time) are June and december. We think this makes sense when considering that December is the holiday season (christmas and new year) where airports might be unusually crowded causing higher than average delays. June is the first month of summer , and school summer holidays, therefore; many people go on vacations which may cause airport crowding and explain the high delay. September is associated with the minimum delays, we suspect because it's not associated with any holiday and it's the start of school season in the US so many people stay at home .

***⇒***the worst day of week to travel on in terms of delay is friday; which makes sense because the work week in the US ends on friday; and the weekend is on saturday and sunday for most workers, therefore, we would expect more crowding (and more delays) on friday as people travel for the weekend. The best day to travel (minimum delay) on the other hand is saturday, which, again, makes sense because it's in the middle of the weekend and not many people would be booking flights then.

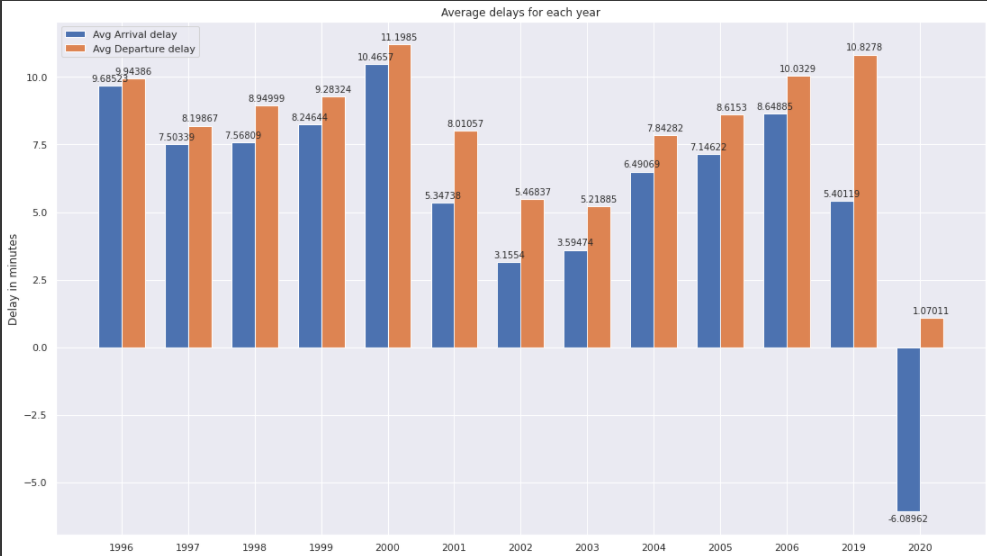
⇒ so the best time to travel to minimize delays would be september 25th on a Saturday.

# 

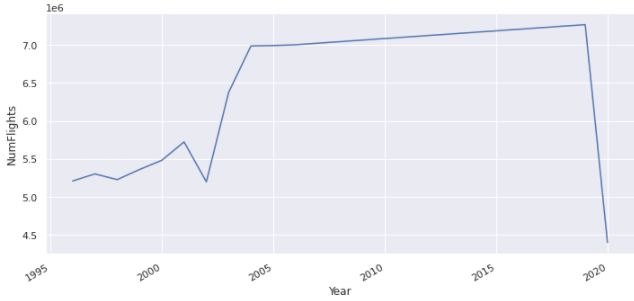
# Question two: compare 2020 travel patterns vs prior years (effect of covid19 and lockdown):

**The following figures describe the number of flights, average delay, and number of cancelled flights across all years.**

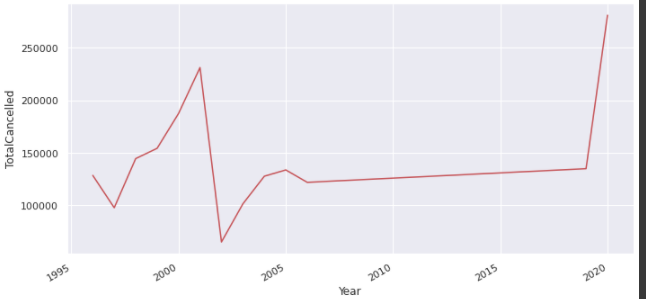
**1- Average Delay for each year**



***Figure 7 : The average Delay for each year***

**2- Number of Flights in each year**

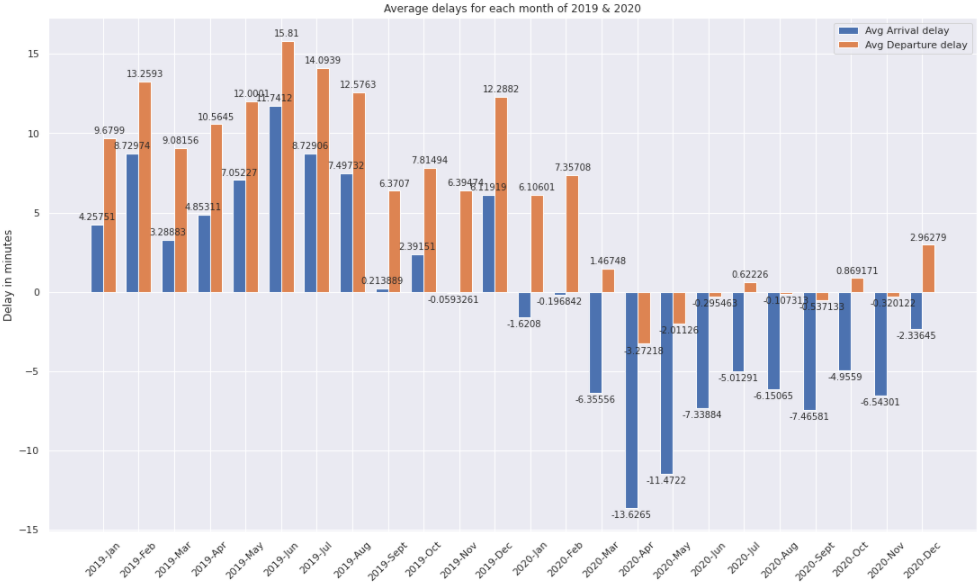
***Figure 8: The Number of Flights in each year***

**3- Total Cancelled Flights in each year**

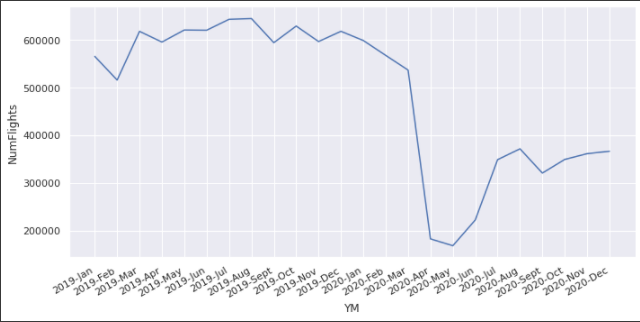
***Figure 9: Total Cancelled Flights in each year***

**focusing on the months of 2019/2020:**

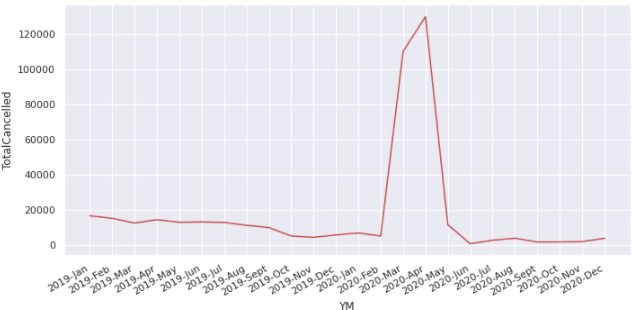
**Average Delay for each month of 2019 & 2020**

***Figure 10: Average Delay for each month of 2019 & 2020***

**Number of Flights in 2019 and 2020**

***Figure 11: Number of Flights in the months of 2019 and 2020***

**Total Cancelled Flights During 2019 & 2020**

***Figure 12: Total Cancelled Flights During 2019 & 2020***

**We noticed that**

**⇒** we notice a very interesting observation in the yearly delays data: 2020 is associated with by far the smallest delay values with the mean arrival delay actually having a negative value (early arrival). We think this makes sense given the 2020 lockdown and flight restrictions which caused a significant reduction in the number of flights being dispatched thus reducing the crowding and workload delays (logistical and organizational) that come with it.

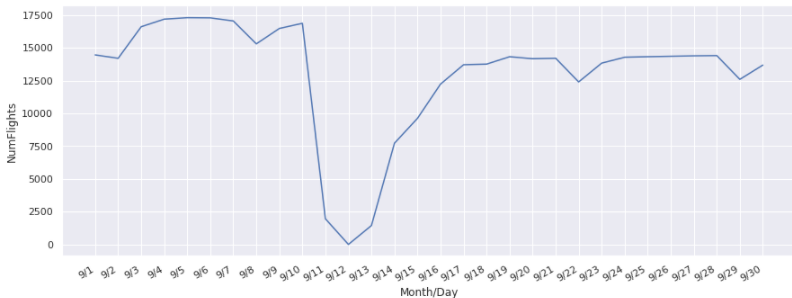
***⇒*** from the yearly dispatched and cancelled flights data, we notice that the minimum number of flights dispatched as well as the highest number of flights cancelled was in 2020. This is expected because of covid lockdown.

⇒ when analyzing the monthly trends in 2020, we found that the sharpest decrease in dispatched flights and the sharpest increase in flight cancellation happened in april of 2020. This matches the real timeline of covid lockdown since it was in the end of march and start of april that total lockdown and travel restriction policies were being implemented in the US.

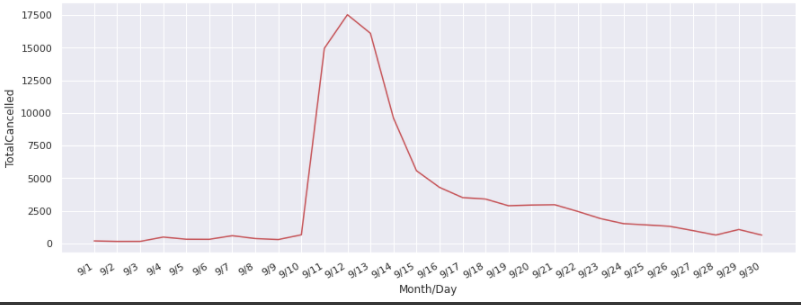
# 

# *Question Three:* flight patterns pre-and-post 9/11:

**Number of flights during the month of september**

***Figure 13: Number of flights Before and After 11 September (month of sept)***

**Total Canceled Flights Before and After 11 September (month of september)**



***Figure 14: Total Canceled Flights Before and After 11 September (month of sept)***

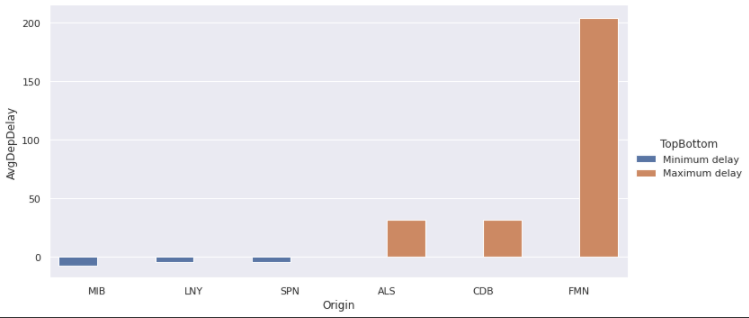
**We noticed that**

**⇒** by analysing the number of dispatched and cancelled flights per day during the month of september 2001, we see a very clear trend around the 11th day of september. The number of dispatched flights took a very sudden sharp fall on september 11, 12, and 13 with the minima being on sept 12 (the day after the twin tower attack).

⇒ Similarly the number of cancelled flights sharply and suddenly increased around the same dates before slowly dropping back to normal in the days that follow.

# *Question Four:* what are the delay patterns per airport - the airports associated with the highest departure delays?

The following figure describes the top and bottom 3 airports in terms of delay.



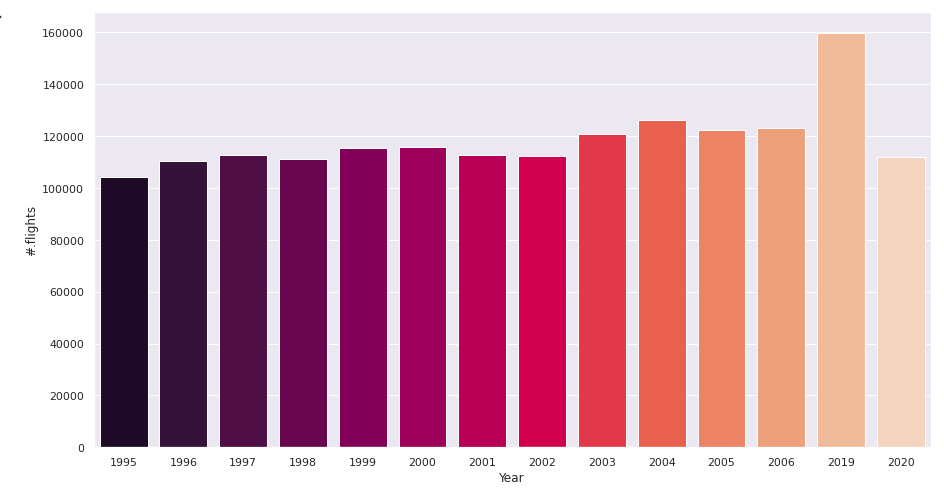
***Figure 15: the top and bottom three Average Delay Airports***

**We noticed that**

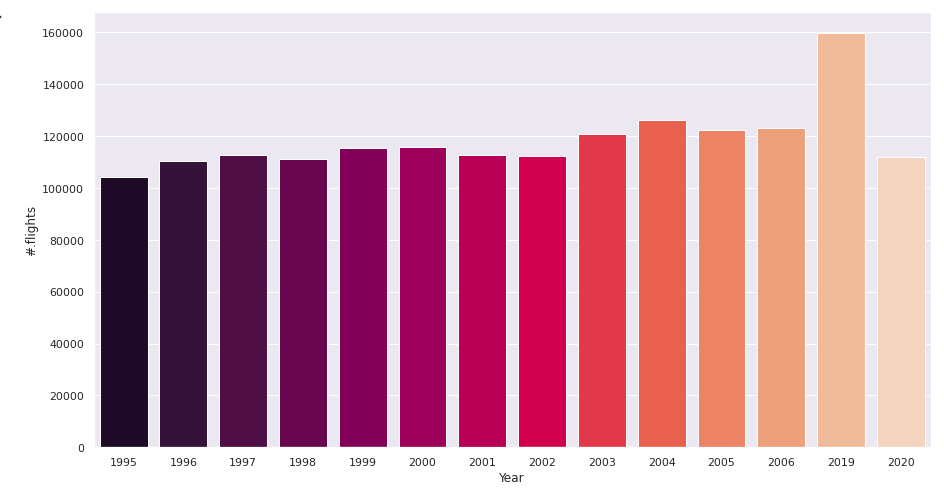
**⇒** Four Corners Regional Airport (FMN) has The maximum delay.

Minot Air Force Base (MIB) in North Dakota has the minimum delay. Considering that it’s an airforce establishment it makes sense that it would have the least delay.

# Question Five: How does the number of flights between different locations change over time?



***Figure 16: the # of flights over the years coming out of Washington DC Airport.***



***Figure 17: the # of flights over the years between the local New York airports.***

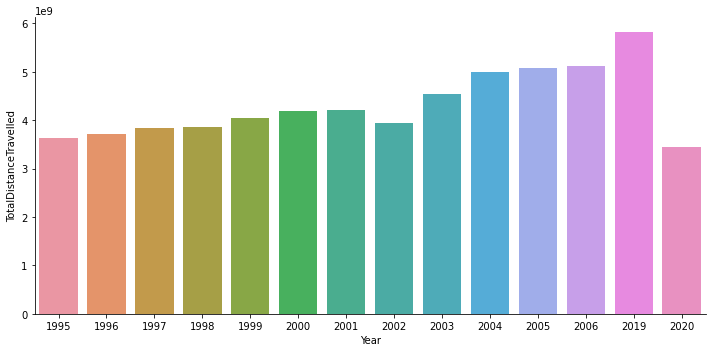
**We observe that**

**⇒** When plotting states like Massachusetts either local flights or flights to other states, we only get the last two years, this was quite odd since Massachisetts is actually considered an old state with really old airports. We conclude it is due to the statistical center administration itself there. We noticed other patterns where the local flights are sometimes missed in comparison to flights with other states, like NJ to NY, this is expected since NJ used to have only 2 airports till very recently, and became 3.

* The odd thing is no matter what states we change, it is always that 2007 with the least significant number of flights and 2008 totally missing. We first speculated it may be due to the great recession in 2008, we checked the online data from cnc, it said it only dropped 30%, but in our dataset it dropped much more than that. We think it is due to some logistic issues during such a hard time that made them document very few of the flights data by then. Then later we drop all of that when we remember the 2007 issue explained in the technical report.

# Question six: the total distance travelled in each year:

The following is a distribution of the total travelled distance per year.

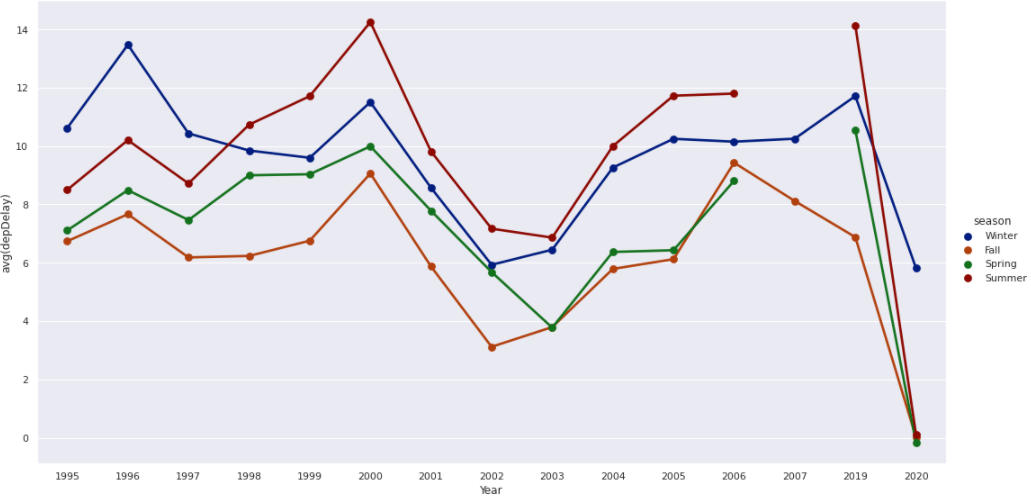


***Figure 18 : a distribution of the total travelled distance per year***

We observe that:

more recent years tend to have higher distances with the travelled distance climaxing in 2019. However, only 2020 significantly breaks this trend by having the lowest distance of all (due to covid lockdown). Generally we expect the total travelled distance to increase in recent years; since air travel is getting more affordable and the world experiences more globalization, more and more people are travelling by air. Not to mention that the technology and capacity of airports have gotten better as time goes on and more flights can be dispatched serving more people and covering higher distances.

# REQUIREMENT 7: Seasons.

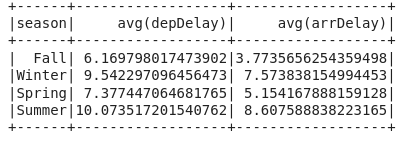
***Figure 19 : The trend of the average Departure delay for the four seasons over the years.***

******

***Figure 19 : The distribution of the average Arrival delay for the four seasons over the years***

**We observe that**

**⇒**



This matrix shows summer to be the season with the highest delays.

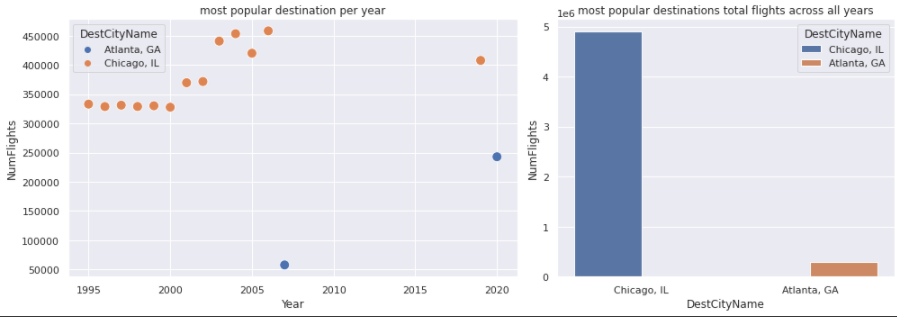
A quite more scrutiny shows that summer has dominated the Departure delay since 1998, after it was the winter with the highest average dep delay, and Fall is always with the lowest average departure delay.

In the arrival delay, the summer still dominated it, however there was a very nice observation at the year of 2020 where the average arrival delay was below zero, which means it arrives before its expected time, at first we thought it is due to advance in aerodynamics or some similar explanation, but it is not enough to explain such sudden change over a year. The reasonable explanation is actually it is due to the pandemic which decreased the number of the flights radically, and this definitely has a significant effect on the arrival time.

# Another insightful observation is how radically the number of flights reduced in 2002, especially in the Fall, which is the first anniversary of the 9 eleven attack.

# *Question Eight:* what are the Most popular destinations across the years?

**The following figures describe the number of Flights in each year per destination city. On the right are the two most popular destinations and their number of received flights.**

****

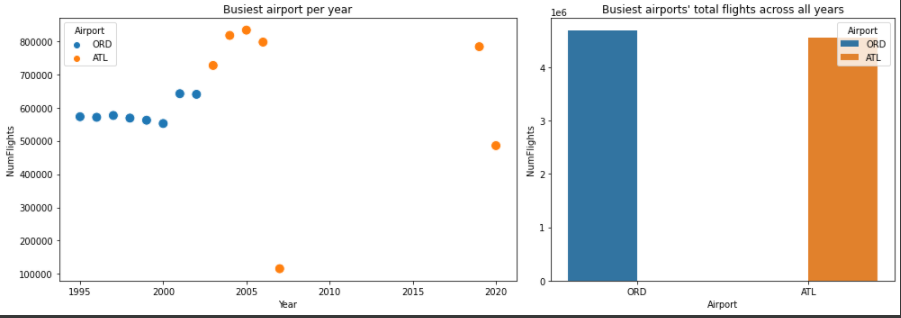
***Figure 25: popular Destinations per year***

**We noticed that**

**⇒**the most popular destination has been overwhelmingly Chicago, Illinois since 1995.

# *Question Nine:* what are the busiest airports across years?

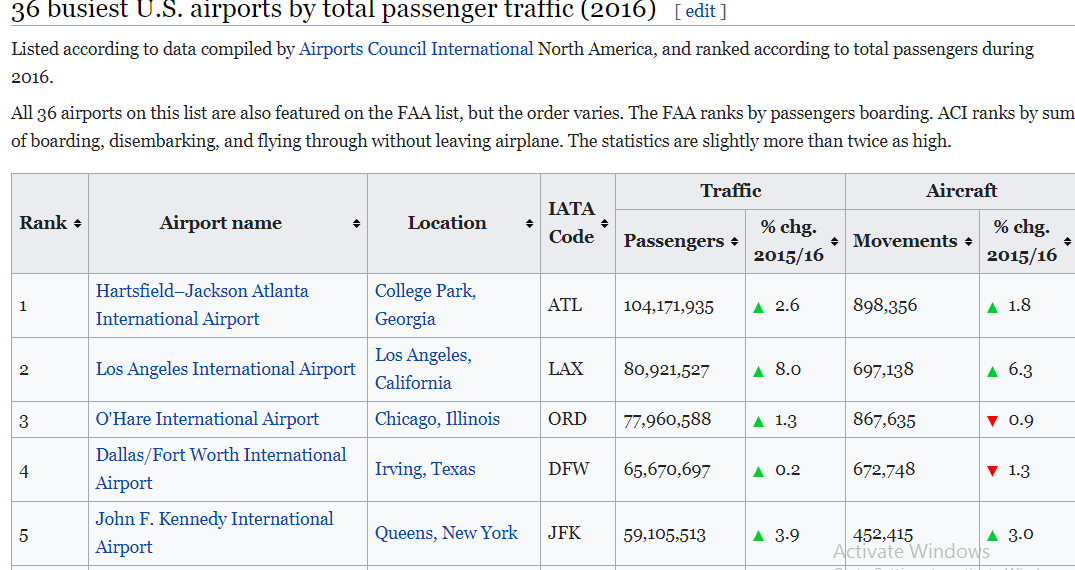
**The following figure describes the number of Flights in Each Year per airport. On the right are two busiest airports as per our analysis.**

****

***Figure 26: popular Airport per year***

**We noticed that**

**⇒**we note that the busiest airports across the years are O'Hare International Airport (ORD) in Chicago Illinois, and Hartsfield-Jackson International Airport (ATL) in Atlanta Georgia. This is consistent with real documented trends such as the reference image below: we note that both ORD and ATL are in the top 3 busiest US airports in 2016 as per this wikipedia page. It is to be noted however that this reference measure total passenger traffic while we measure number of incoming and outgoing flights so the results will not perfectly align but it's an interesting analogy nonetheless



# Question Ten: ML analysis:

1- predicting departure delay based on time of year, month, and week, as well as source, destination and distance:

We used a regression model to estimate a predicted delay given the aforementioned features, however, we couldn’t achieve a decent prediction accuracy. We calculated the correlation between the delay and every other variable in the data and the found correlation coefficients to be all +-0.2 or less so we conclude that there isn’t really much of a linear correlation between them. We suspect it’s because the data is full of outlier values because each year, month or even week has its own unexpected events that may affect the delay in an unpredictable way. Bad weather for example, or any such small unpredictable factors can dramatically affect the delay, not to mention major events like 9/11 or the covid pandemic.

2- predicting flight cancellation based on the same features:

We tried a random forest classifier to predict flight cancellation patterns with mild success. We attribute the low accuracy of this model to the same factors that limited the other one.