1.Introduction

In the last ten years, according to the Bureau of Transportation Statistics (BTS), only 79.63% [1] of all flights have performed on time. Only a few remaining percentage were cancelled or diverted, less than 2%; rest of them were delayed mainly due to late arriving aircraft followed by the cause of the national aviation system and air carrier. A flight is considered delayed when a flight arrives or departs 15 or more minutes than the scheduled time. Averagely speaking, 720 million people [2] were on board and 144 million of those were affected by flight delays caused by five main reasons. Those reasons are [3]:

1. **Air Carrier:** The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
2. **Extreme Weather:** Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
3. **National Aviation System (NAS):** Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
4. **Late-arriving aircraft:** A previous flight with the same aircraft arrived late, causing the present flight to depart late.
5. **Security:** Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

These series of delays cause a serious financial burden on airlines. In 2010, the Federal Aviation Administration commission estimated that flight delays cost the airline companies $8 billion a year, most of which due to increased budget on crews, fuel and maintenance [4]. It is also worth noting that this cost adds up to $32 billion in a year by accounting for other lateral costs such as passenger cost and indirect effect on associated businesses.

By addressing delayed flights problems, it is possible to reduce its heavy financial load on airlines, passengers and GDP. That is,

(1) letting passengers know about the probability of a particular flight being delayed before booking can help them to plan on a better date and time. In addition, it will save passengers more than a million hours of unnecessary delays per year [5].

(2) booking websites, airline companies, travel agencies can provide better customer service. Providing likelihood of flight delays to clients can decrease the number of complains and increase customer satisfaction.

(3) airline companies can pinpoint important factors that cause flight delays and take precautions depending on this information. The principal benefit of this study to airlines is the reduction in additional operating cost caused by delays.

Overall, the findings of this study can provide a high-profile achievement by addressing aviation delay problems with a robust prediction model and helping people and businesses better on planning their flights

2.Data Acquiring & Wrangling

In this section, I will provide information about what data acquiring and wrangling processes were taken.

How data is gathered: For this study, I used four different datasets from various sources.

1) Airline on-time performance dataset is available on The Bureau of Transportation Statistics' website. In the process of data scraping, I used the selenium library in Python. This library works with a specific version of a Chrome driver and can be downloaded from this page. It downloads the monthly data to the ~/data/flight_data folder. From then on, a shell script concatenates these monthly flight data into a single csv file under the same folder.

2) Weather data was obtained from Iowa State University's Environmental Mesonet Platform. This platform works like an API service which needs modification of a requested URL for each inquiry. The python script for weather data can be found on this GitHub repo.

3) Airport data was obtained from The Bureau of Transportation Statistics' website manually. This dataset has a substantial amount of information such as airport names, location, timezone etc.

4) ICAO data was acquired from OpenFlight dataset. This data is needed because The Environmental Mesonet Platform of Iowa State University requires ICAO codes as station IDs. This information does not exist in the airport dataset obtained from BTS.

How data is cleaned: Airline on-time performance dataset have redundant out-of-scope information, so I only kept 30 necessary features in the data gathering process. I created a dictionary called flag to identify null entries (condition 1), out of range entries (condition 3), and both null and out of range entries (condition 2). Certain implementations have been performed based on flag values. For instance, a week day column has to be in between 1 and 7. Any entry that has null in that column needs to be imputed using pandas dt accessor. Here is the list of ranges of certain columns.

Column	**Minimum**	**Maximum**	**Treatment (or Remarks)**
Date	Given	Given	Any date entries out of a given range is removed.
Week Day	1	7	Imputed if possible (using dt accessor of pandas)
IATA	N/A	N/A	Any null entries are removed.
TailNum	N/A	N/A	Any null entries are removed.
OrgAirID	10001	16878	Any out of range entries are removed.
DestAirID	10001	16878	Any out of range entries are removed.
OrgMarID	30001	36845	Imputed if possible. Otherwise removed.
DestMarID	30001	36845	Imputed if possible. Otherwise removed.
Div	0	1	Any wrong entries are imputed based on certain columns.
Cncl	0	1	Any wrong entries are imputed based on certain columns.
CnclCd	1	4	Any out of range entries are removed.
ScDepTime	0	2400	Any out of range entries are removed.
ScArrTime	0	2400	Any out of range entries are removed.
ScElaTime	N/A	N/A	Any null entries are removed.
DepTime	0	2400	Imputed if certain conditions met.
DepDelay	N/A	N/A	Imputed if certain conditions met.
TxO	N/A	N/A	Imputed if certain conditions met.
TxI	N/A	N/A	Imputed if certain conditions met.
WhOff	0	2400	Imputed if certain conditions met.
WhOn	0	2400	Imputed if certain conditions met.
ArrTime	0	2400	Imputed if certain conditions met.
ArrDelay	N/A	N/A	Imputed if certain conditions met.

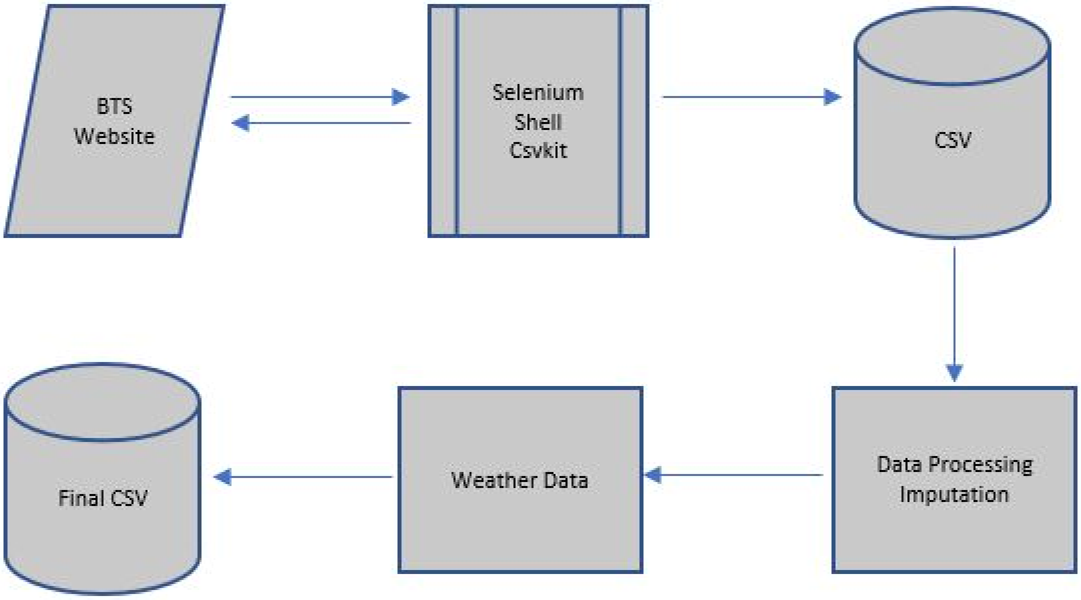
In the final form of the data, null entries are only allowed when a flight is cancelled or diverted. The columns need to be null when a flight is cancelled are:

DepTime, DepDelay, TxO, WhOff, WhOn, TxI, ArrTime, ArrDelay, AcElaTime, AirTime

and when a flight is diverted are:

ArrTime, ArrDelay, AcElaTime, AirTime, TxI, WhOn

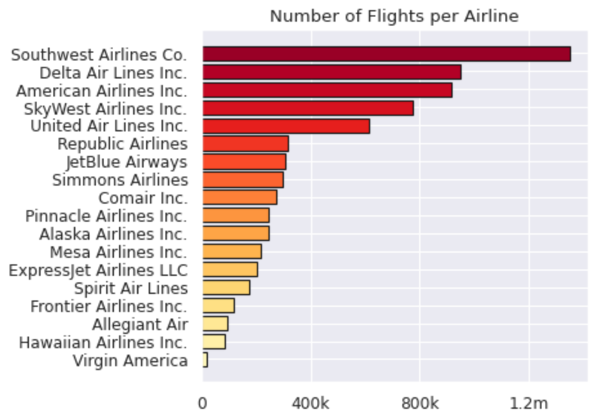
Here is the general flowchart of data collection, cleaning, processing, and imputation.

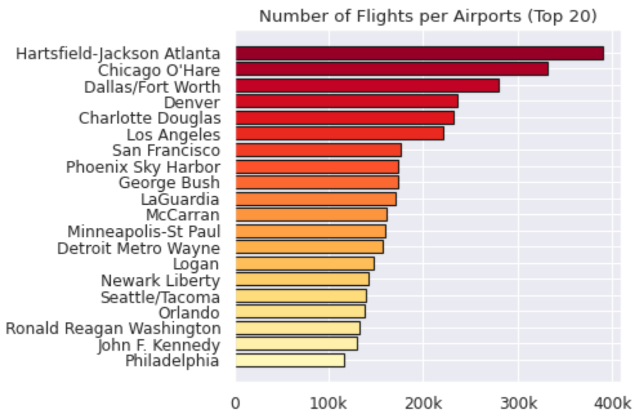


3.Exploratory Data Analysis

In this section, I will investigate the data to get meaningful insights, relations between flight delays and several other factors. This could be statistical correlation or patterns.

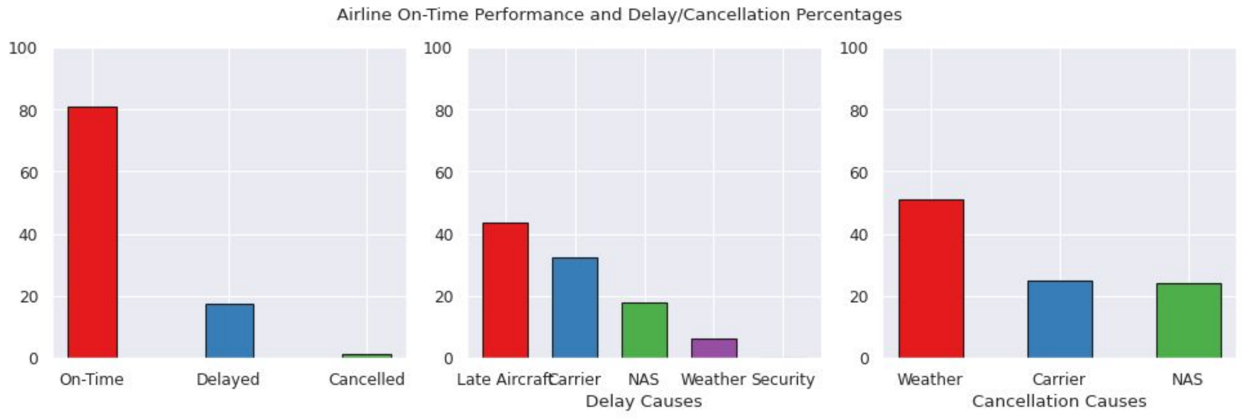
3.1.Busiest Airlines and Airports





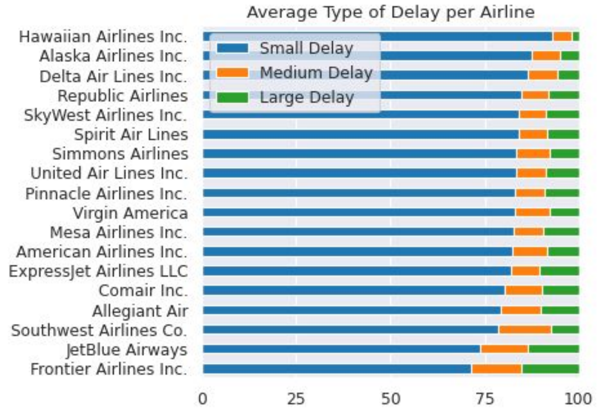
The figures above show the most flying 20 airlines in the US on the left and the busiest 20 US airports on the right. With more than 1.2 million flights during 2018, Southwest Airlines is the most flying airline that is followed by Delta and American Airlines with 900k+ flights. Furthermore, Virgin America is the least flying airline with less than 50k flights. On the other hand, Hartsfield-Jackson Atlanta Airport is the busiest one with almost 400k flights during 2018 that is followed by Chicago O'hare with 330k flights. They are the only two airports passing the 300k flight boundary. There are only 4 airports (Dallas, Denver, Charlotte, and Los Angeles) that have 200k-300k flights in a year.

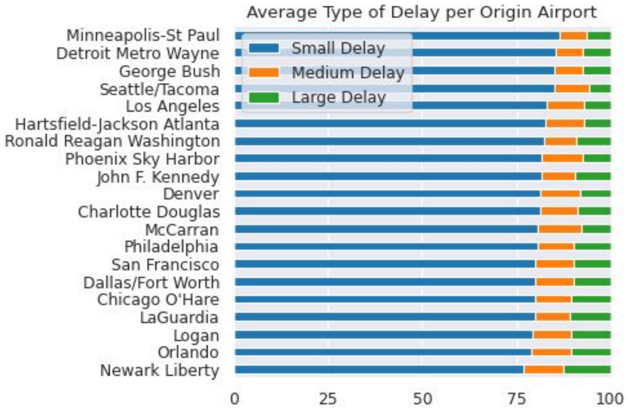
3.2.Airline On-Time Performance (General)



There are several causes for a flight to be delayed. It is interesting to look at the proportions or chances of encountering a certain delay type. Almost 20% of the fights tend to be late (this taking account delays more than 15 mins). Most planes are late due to late aircraft and/or airline delays. This is interesting because it means that it is an area of the airline industry that could be optimized. Supporting our argument, one notices that the lower amount of delays are caused due to weather and security issues. Showing us that security for example is optimized to a very large level and other types of delays could eventually be avoided by optimizing organization in airports or establishing a delay prevention schedule. Weather delays are rather rare and are not primary reasons for flights to be delayed.

3.3.Airline On-Time Performance per Airline Companies and Airports



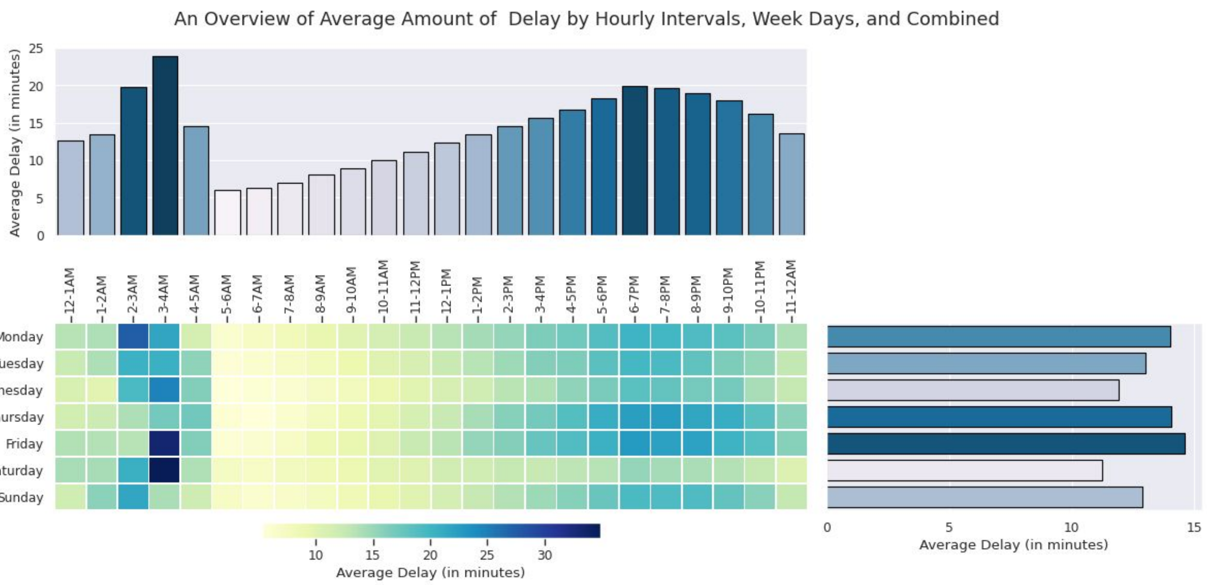


Delays are divided into three categories, namely "On Time or Small delay" (up to 15 minutes delay), "Medium delay" (15 – 45 minutes delay) and "Large delay" (45 minutes delay). In this way the graphic representation is more understandable as well as the possibility of directly comparing the variables related with delays. As represented, one can observe that Airlines that were particularly good at being on time compared to other airlines were Hawaiian Airlines, Alaska Airlines, and Delta Airlines. Those that tend to have a larger delay are Frontier Airlines, JetBlue Airways, and Southwest Airlines.

The same type of analysis can be done for different airports. In this case one can observe that airports such as Minneapolis and Detroit stand out positively, while the one with the worst frequency of delays is Newark. However, the differences are less obvious and, generally speaking, one can conclude that airlines play a more significant role than the airports of departure.

Since, all of the airlines are operating in these top 20 airports, the delay variation due to a specific airline is averaged out. Also, the busiest airports shown in Fig. 1 are not the best performers due to the domino effect of a delay caused by environmental conditions.

3.4.Effect of departure time and week days of a flight on departure delay



The figure above summarizes the average amount of delay in minutes by daily, hourly, and combined time intervals. One can realize that the two histograms around the heat map represent average delay with respect to daily (on the right) and hourly (on the top) axes. It is clearly seen that the 3:00AM-4:00AM time slot is the worst time to fly, especially on Friday and Saturday nights, due to long average delays (>30mins). On the other hand, it is best to schedule a flight in the early morning (5:00AM - 10:00AM), because the average delays are way below the yearly average (10 mins). Towards the evening, the delays are monotonously increasing and reaching its peak at 6:00PM-7:00PM. One can interpret this graph as the sum of two gaussian distributions centered at the two peak delay times mentioned above. The standard deviation during the night is much smaller than the evening. I believe this is related to the low density of flights during the night.

Day to day variation of the delays is much smaller (< 4mins) than the variation during the day. The best days to travel are Saturday and Wednesday that can be associated with weekend oriented travels where Friday evening and Monday morning are busier. Due to seat availability and cost, the increase in mobility is extending to Thursday and Tuesday. Long weekends may also have some effect in this increase.

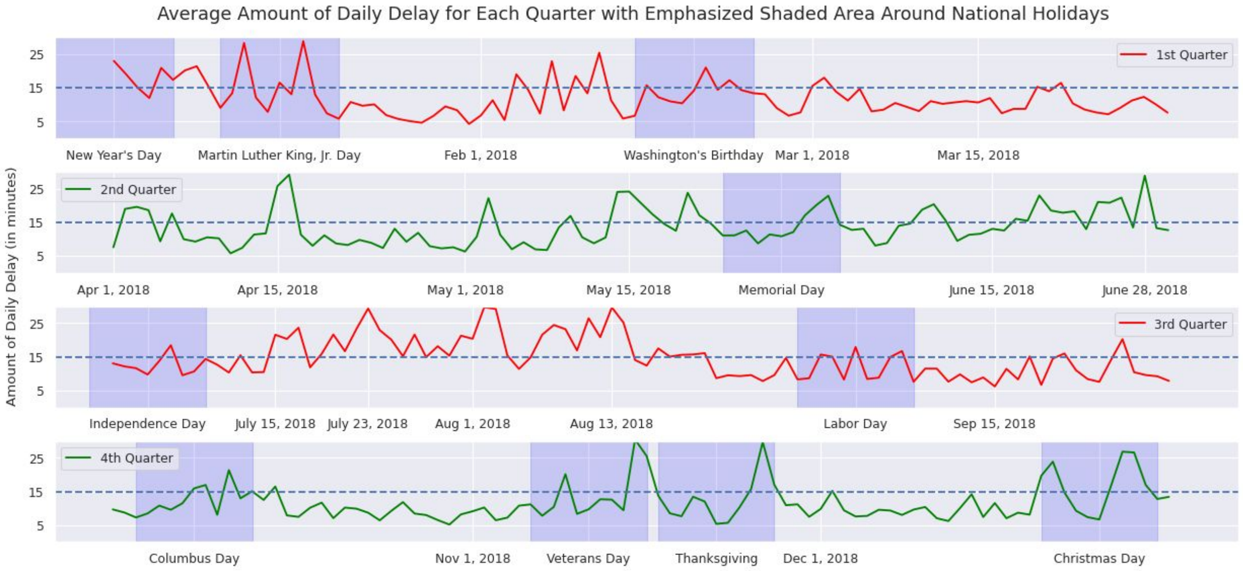
From hourly interval perspective,

- delays are higher between 2:00 AM - 4:00 AM,
- early morning flights have less probability to be delayed (5:00 AM - 10:00 AM),
- probability of delay constantly increases between 5:00 AM and 7:00 PM,
- its relative peak is at between 6:00 PM - 7:00 PM.

From daily interval perspective,

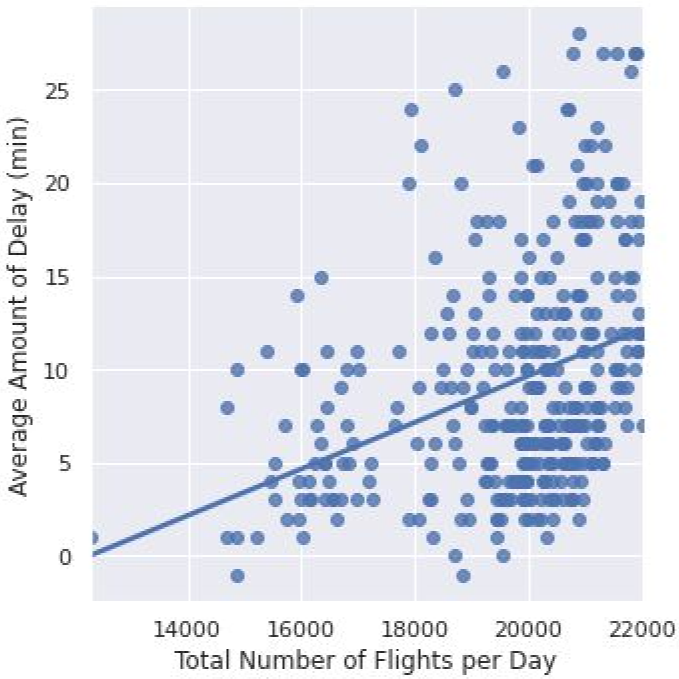
- average amount of delay per day are so close to each other,
- the minimum delay happens on Saturdays,
- Saturdays happen to have minimum delays in evening flights,
- Thursdays and Fridays seem to have a higher chance of having delay.

3.5.Trends by Quarters



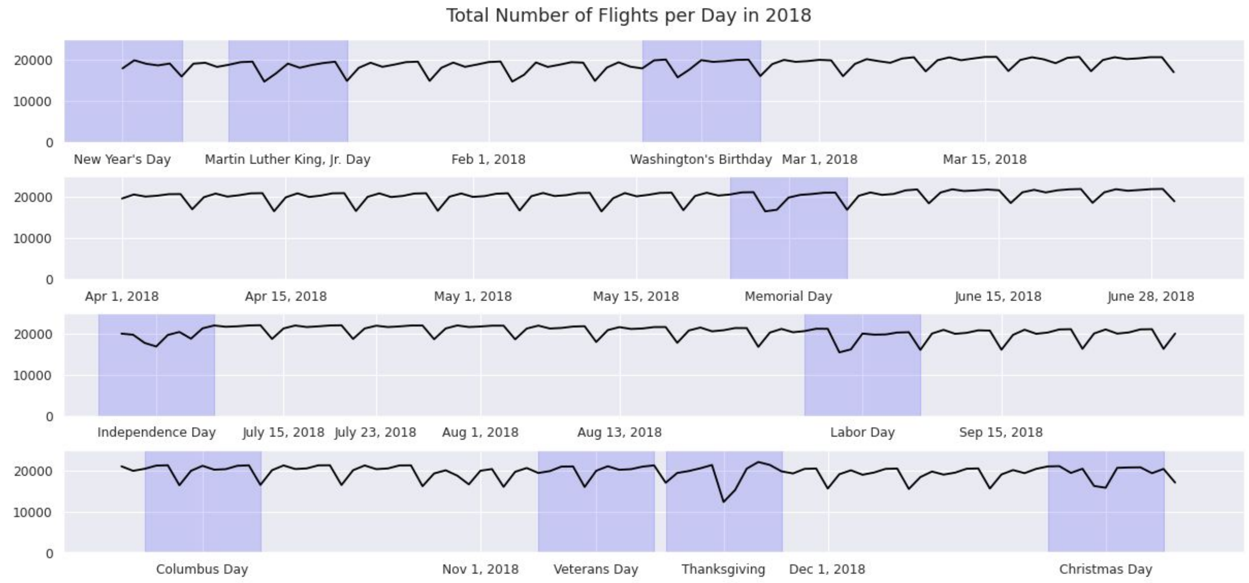
The graph above shows the average amount of delay per day in four quarters of a year. Blue shaded area represents plus and minus five days from national holidays (in total 10 days). The Federal Aviation Administration (FAA) considers any flight that is late more than 15 minutes **as a delay**. That's why I also emphasized 15 minutes with a dashed line in the plot. The increase in average delays before and after the national holidays can be associated with higher traffic density of the airports. This can be seen in the following figure.

3.6.Efect of Total Number of Flights on Average Amount of Delay



The figure shows the trend analysis that investigates the effect of airport traffic density on average delay. The solid blue line shows the trend while dots represent the individual days of year 2018. It is clearly seen that there is a linearly increasing trend as expected.

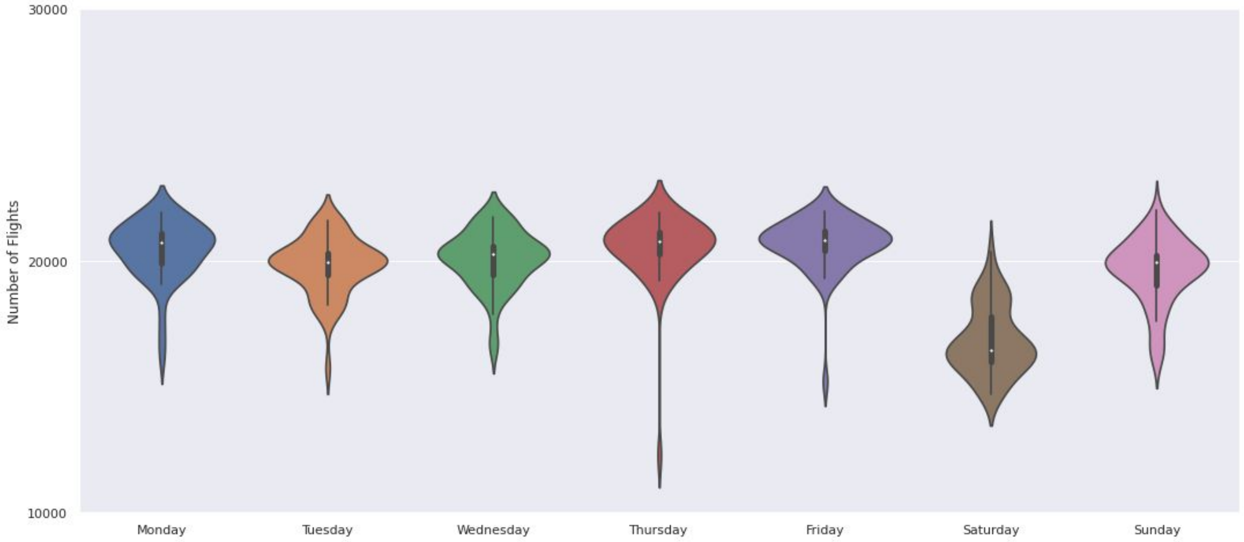
3.7.National Holidays’ Effect on Total Number Flights



The number of flights during 2018 are plotted daily to understand is there any density increase during the national holidays. It seems that during the national holidays the flight distribution is redefined, but there is no visible increase in the number of flights.

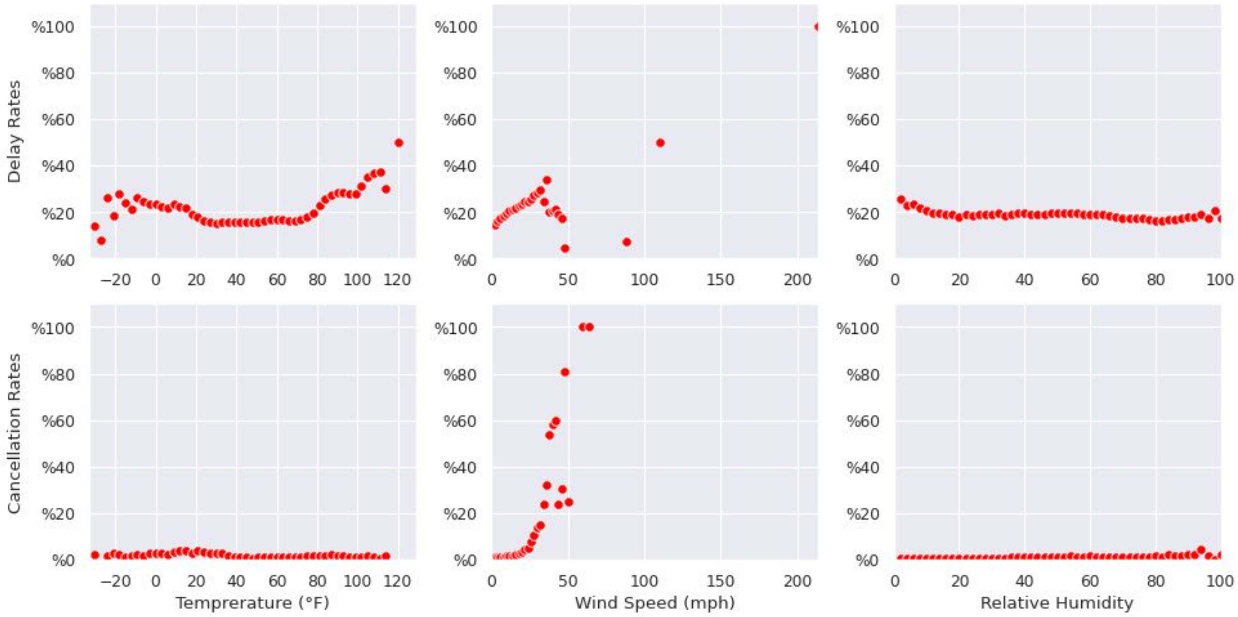
During normal times, there is a pattern that repeats itself weekly with a valley on Saturdays. However, the pattern is changed during national holidays according to the day of the holiday. The most drastic change happened during Thanksgiving holiday where Thanksgiving day (Thursday) has the lowest number of flights of the entire year and Saturday of that week has way more flights than average number of flights on Saturdays. Additionally, the number of flights is varying with respect to the seasons. During autumn and winter they are generally low and increase in Spring and reach their maximum point in mid-summer. This is probably related to summer break of students and vacation preference of workers.

3.8.Total Number of Flights (Daily) Distribution per Week Day



The figure shows the density of the flights for each day of the week for entire year. Therefore, each day has 52 data points demonstrated as both a boxplot and a violin plot that is symmetrically plotted Gaussian distribution of points. It is clearly seen that, Saturday is the least dense day of the week. Also, the low end of the Thursday and Friday are related to the Thanksgiving week. Additionally, low and high ends of violin plots are probably related to long weekends where some travel is shifted to the other days. On average, Monday is denser than Sunday that is probably related to the higher number of short flights than long ones during 2018. Since they are weekend travels, people prefer early morning flights on Monday to Sunday when they can arrive to work on time.

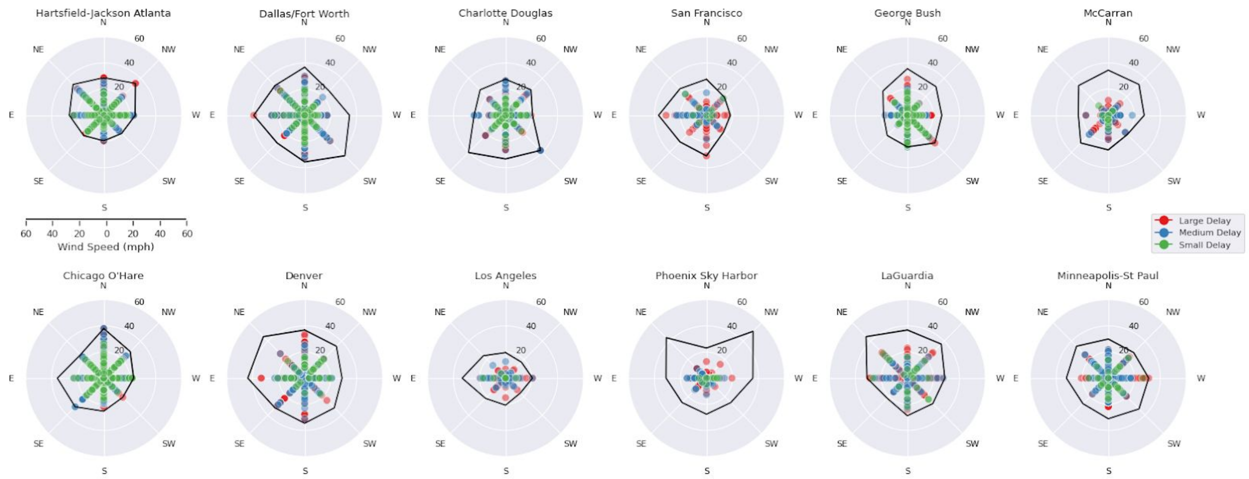
3.9.Effect of Weather Conditions on Delays and Cancellations



In this figure, the effect of different weather parameters such as temperature, wind speed and humidity on delays and cancellations is demonstrated. It seems that the delays and cancellations are almost indifferent for increase in temperature and humidity. High temperatures above 80 causes more delays than base delay of 15-20% that is due to the other delays unrelated to weather. On the other hand, delays are not monotonously increasing in low temperatures. The variation is probably due to the experience and presence of the equipment to land and take off planes in the airports where freezing weather is expected.

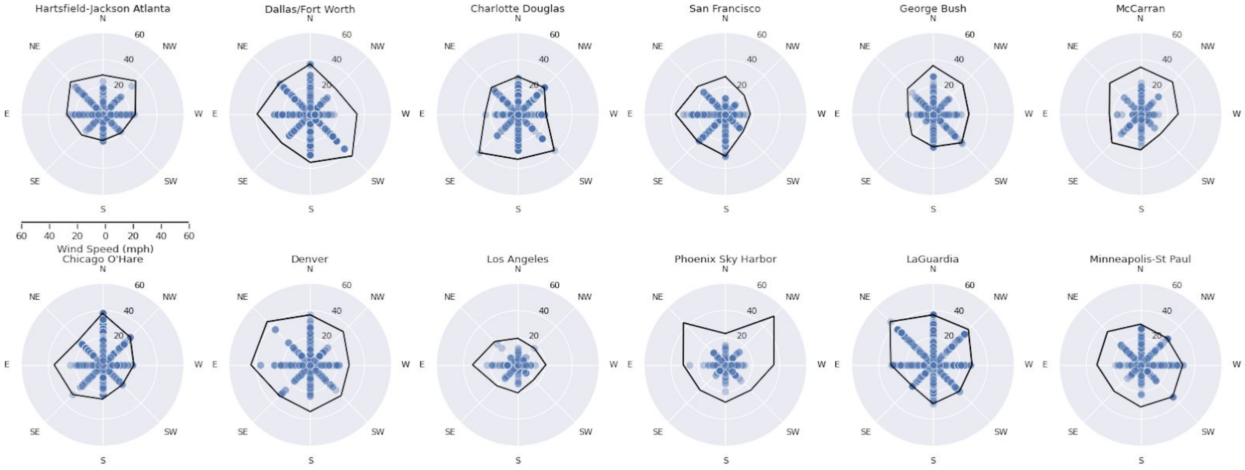
It is clearly seen that wind speed is directly influencing the delays and cancellations. The delays are increasing linearly up to 40mph. Then, the delays are behaving different. Main reason is the exponential increase of cancellations with the wind speed. Since most of the flights above 40mph are cancelled, the delay rate is decreased. Also, there are delays at higher speeds where no cancellation data is given that is probably due to the short duration of such high speed winds.

3.10.Effect of Wind Speed and Direction on Weather Caused Delays



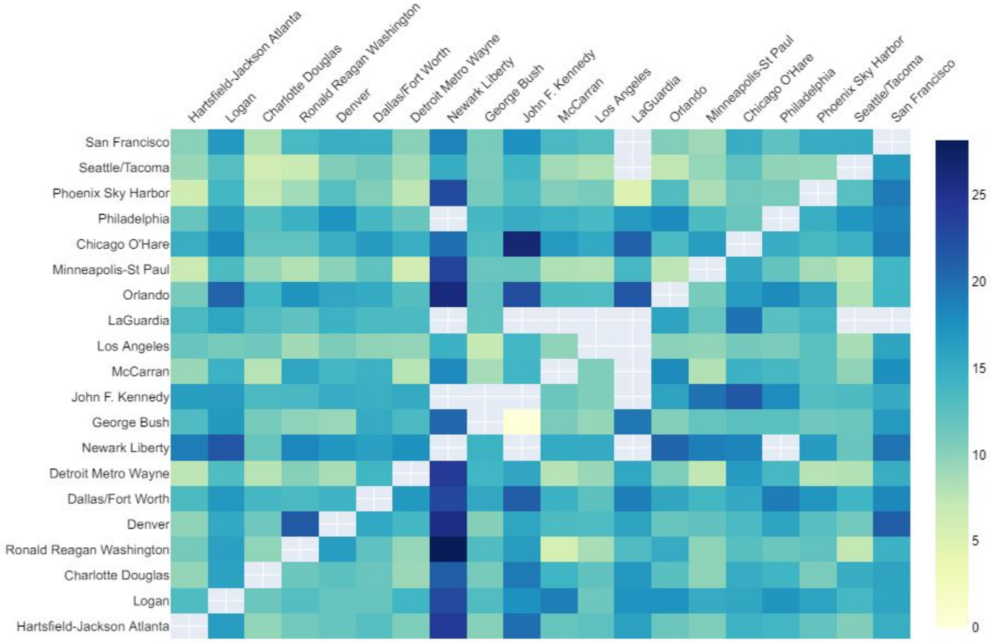
In this figure, the maximum wind speeds for each direction observed in top 12 US airports are shown as a black solid line. The colored dots represent the amount of weather caused delay for corresponding wind speed and direction measured. In this figure, green dominant areas show that airport operation continues well under corresponding weather. Therefore, airports like Chicago O'Hare, Hartsfield-Jackson Atlanta, Dallas Fort Worth and George Bush are operating well for most of the wind speeds. On the other hand, Los Angeles, San Francisco, Phoenix Sky Harbor and McCarran are affected by even relatively low speeds. This difference is related to the design of the airport including its position, direction and length of the airport runways, type of the winds, and technological advancement of the airport. For example, Los Angeles and San Francisco are placed near the ocean, therefore the winds from the ocean can cause more problems even at low speeds due to immediate weather change. On the other hand, Chicago O'Hare is a large airport with several runways and still expanding. Its history shows that runway lengths and directions are changed to operate at high speed winds from North.

3.11.Effect of Wind Speed and Direction on Weather Caused Cancellations



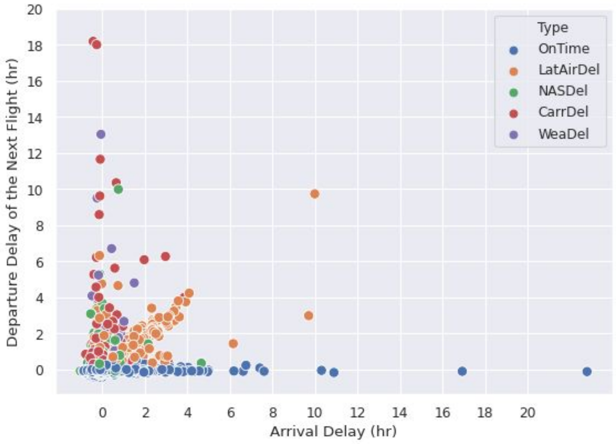
The figure shows the weather related cancellations similar to the previous one. There are cancellations for every wind speed and direction for each airport. The distribution is similar to the delays with extra points at high wind speeds. As an example, all flights in LaGuardia airport are cancelled when the wind from the North East direction exceeds 30mph. It is not easy to comment on the cancellations at low speeds where normally small delays are expected. Because, the cancellation rate is approximately 1-2% as shown in Figure 10.

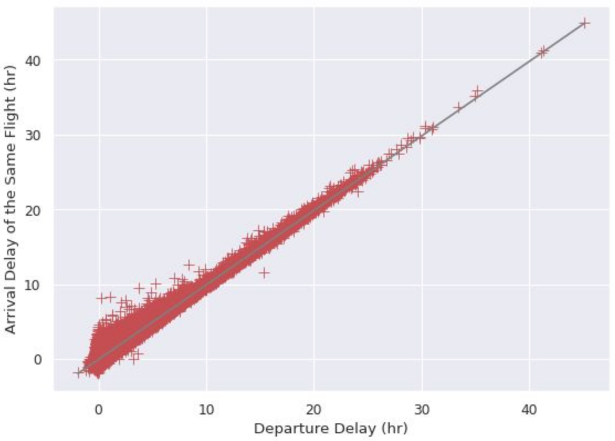
3.12.Origin-Destionation Pair Average Amount of Delay



Average amount of delay of the flights operated between top 20 airports are shown as a heat map where rows are labeled according to the origin airport. Grey cells show NaN values meaning no flight data is available between corresponding airports during 2018. It is obvious that Newark Liberty is the worst destination airport. Additionally, the flights between Chicago O'Hare and John F. Kennedy airports have larger delays than the average in both directions. With using a heat map, one can compare on time performance of nearby airports and, if possible, select destination and origin airports accordingly.

3.13.Late Aircraft Delays





The graph on the left demonstrates how the delay on the arrival affects the departure of the next flight of the same aircraft. The reason for the departure delay is color coded. On time departures, shown as orange, occur even when there are large arrival delays. It is probably related to scheduling where the plane has several hours before the next flight. Furthermore, Late Aircraft Delay (shown as red) influences the next flight in a linear fashion as expected. Moreover, weather and carrier related delays, shown as green and red respectively, are the main reasons for delays when the aircraft has completed its previous flight on time.

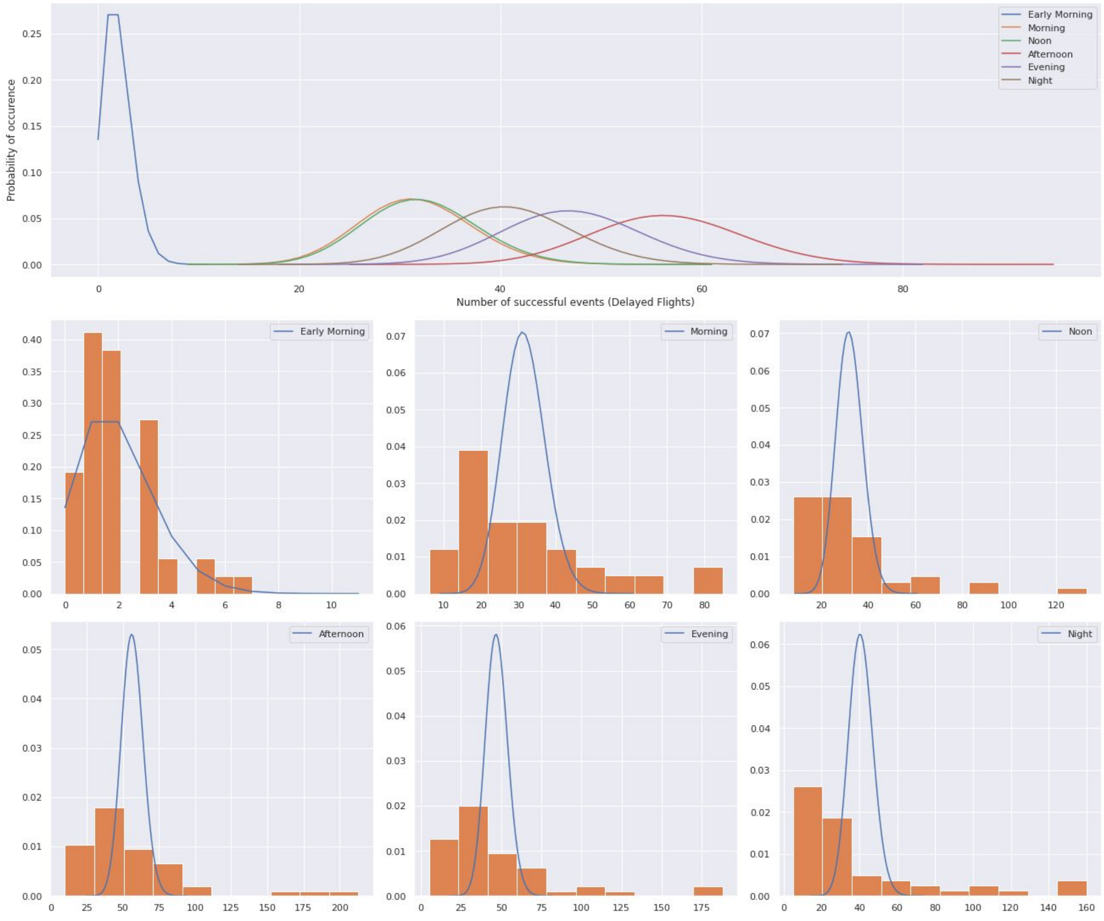
On the other hand, the right figure shows the effect of departure delay from the origin airport on the arrival of the same flight to the destination airport. The data behaves linear as expected, because the travel speeds and routes of flights are well regulated. Still they can speed up or take different routes with the permission of aviation authorities to go faster than planned. In contrast, they may need to slow down, take longer paths due to weather or wait some time to land due to heavy airport traffic that will cause extra delays. The standard deviation of the delays from the linear line is 13 minutes.

4.Statistical Analysis

4.1. Poisson Model

Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

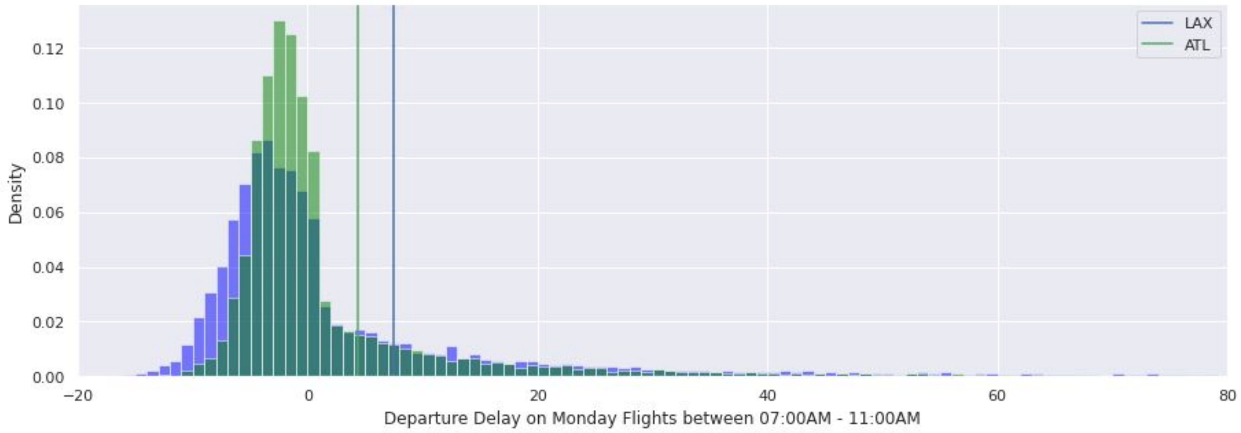
In this example, we will investigate the poisson model for flight from Atlanta Hartsfield Airport where number of successful events would be the number of delayed flights. To satisfy assumptions of the poisson model, independence, homogeneity, and fixed time, we will divide time of the day first into 7 categories. Assuming that one delayed event does not affect the next event, we can compute the number of delayed flights, which is λ (np).



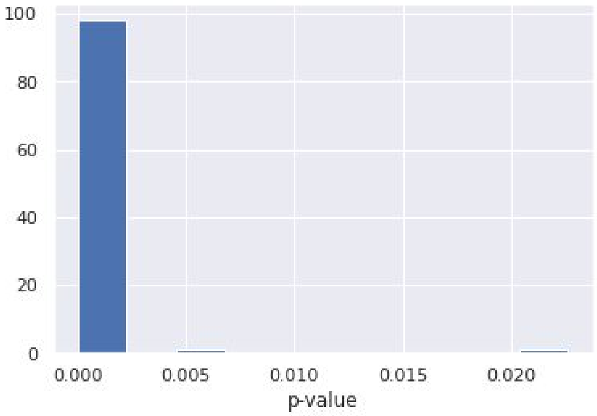
Based on six different poisson models, one can conclude that only early morning flights better represent a poisson model. Estimating the number of delayed flights during other times of the day can be predicted using different models such as normal, lognormal, and exponential.

4.2.Hypothesis Testing

In this example, we will compare departure delay distributions of morning flights on Monday between Los Angeles Airport (LAX) and Atlanta-Hartsfield Airport (ATL) using t-test. The null hypothesis is that the average amount of delay is equal or that the difference in average amount of delay is statistically not significant.



The green distribution, ATL, looks shifted to the right and its peak is higher around zero. This can be explained by the number of flights since ATL has more flight operations than LAX. The green and blue bars show the theoretical mean of ATL and LAX correspondingly.



T-test is performed with 5000 resamples 100 times. All samples show that p-value is smaller than significance level, 0.05 (95%). Therefore, we reject null-hypothesis which is that means of two distributions are equal or that the difference in means of two distributions is not significant.

References

[1] https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1
[2] https://www.transtats.bts.gov/Data_Elements.aspx?Data=1
[3] https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
[4] https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf
[5] Sud, V. P., Tanino, M., Wetherly, J., Brennan, M., Lehky, M., Howard, K., & Oiesen, R. (2009). Reducing flight delays through better traffic management. *Interfaces*, *39*(1), 35-45.