Analyzing U.S flight delays across major New York airports Using descriptive and statistical tools and methods

Abstract: This paper focuses on airline delays, airline cancelations, and airline diversions. The data consists of three major airports in the New York area from September 2010 to September 2019. This paper will explore five major causes that lead to airline delays. By using statistical methods such as; random sampling and estimation, t-test and regression analysis, it will discuss the trends of airline delays, cancelations and diversions. In addition, it will also address the proportion of different causes leading to flight delay, the delay rate of various airline carriers and the correlation between delay and cancelation, as well as delay and diversion. We aim to provide helpful suggestions to travelers. We also aim to prepare insights for the aviation industry, enabling them to take measures to reduce airlines delays, cancellations, diversions whilst providing a better consumer and employee experience.

1. Introduction

Flight delays, cancellation and diversions across Newark International Airport, John F. Kennedy international Airport and LaGuardia Airport will be the main objective of this study. Although flight delays have majorly improved over the years they are still a major cause of concern for travelers. Throughout this study we will create statistical models to assess the current position of airline carriers, airports and the major causes of delay. The dataset utilized is extracted from the Bureau of Transportation Statistics. The dataset provides information on Airline On-Time Statistics and Delay causes for EWR, JFK, and LGA.

1.1 Background and Motivation

The transportation industry has been rapidly growing over the past few years, especially air travel. However, this growth has been plagued by flight delays and cancellations, leading to extreme losses to the airline industry. Flight delays are not only expensive to airlines but extremely inconvenient for travelers. Yet, this year it has been unavoidable. According to CNBC, American passengers in 2019 had a “rocky summer travel season”. U.S airline cancellation jumped from 1.7 % in 2018 to 2.4% in 2019. In addition to this, in June 2019 on- time arrivals fell down by 3.1% when compared to 2018 [1] According to the *Air Travel Consumer Report* issued in November 2019 by the United States department of transportation [2], flight delay has become one of the main reasons why customers are dissatisfied with airlines. In the meantime, according to the research by Ball Et all., flight delays in the United States have resulted in a loss of around 4 billion dollars in terms of domestic GDP, there’s also an 8.3-billion-dollar loss incurred by airlines as well as a 16.7-billion-dollar cost to travelers [3]. Generally speaking, frequent flight delays have had a negative impact on travelers and airlines. As a consequence, we saw the need to look deeper into this data and further understand and analyze the current issue.

Individuals use commercial airlines to travel throughout the year. They do so for vacation, business and more. Vacations are supposed to be a joyful and relaxing affair, yet air travel many times puts a halt to that stress free environment. Most travelers believe that the culprit is airport security as this is often extensively portrayed on the news. Typically, the only complaints publicized in regard to airlines specifically are for unfair treatment. This paper will study and analyze the following.

- Identification of average flight delay times and delay times of individual airlines over the past nine years.
- Identification of all major reasons behind flight delays and cancellations.
- Analysis of the impacts of particular causes leading to flight delays or cancellations.
- Identification of coefficients between airline delays, cancellations and diversions.

1.2 Causes of delay

According to statistics provided by The Bureau of Transportation, the causes of delays or cancellation can be divided into the following broad categories.

Weather: Significant meteorological conditions that are present or forecasted. Based on the judgment of the carrier these conditions may delay or prevent flight operations.

Security: Delays or cancellations caused by emergency evacuation of a terminal or concourse, boarding interruptions or cancellations due to a security breach, inoperative screening equipment and/or lines in excess of 29 minutes at screening areas. Travelers have a higher probability of having their flights cancelled (1.5%) or diverted (1.2%) than to be delayed by security (.5%).

National Aviation System Delay: These are delays and cancellations attributable to the national aviation system. They refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control. The last time the National Aviation System led the Nation in causes for Airline delays was in 2003. Since then it has dropped to the 3rd key factor for delays across the country. This is still astonishing considering that the software controlling the National Aviation System is 40 years old. The system which is called Host is considered to be a safe program to help get planes from point A to Point B. The system is considered entirely inefficient and is unable to handle a large amount of traffic [4] It is said to be, "still safe, in terms of getting planes from point A to point B. But it's unbelievably inefficient. It can handle a limited amount of traffic, and controllers can't see anything outside of their own airspace".

Airline Delays: These delays occur when the cause is due to a circumstance under the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.). One of these circumstances is visible in the industry today. Airlines are currently dealing with a global shortage of pilots, its most important position. In fact, the shortage of pilots has been explained by many as a "Crisis". Boeing's CEO Denis Muilenburg has recently stated, "Air travel is growing so rapidly that 800,000 new pilots will be needed over the next 20 years". However, these predictions aren't just based on the amount of aircrafts Boeing will need once they increase their fleet. A good portion of these new pilots will be replacing an old and diminished workforce. The average age of Pilots in the United States is currently 48 years old, with a mandatory retirement age of 65. There has been an extreme dip in the number of new pilots since the 1990's due to the airline industry being deregulated in 1974. Succeeding deregulation, airlines were able to control their own prices, schedules, and miles. Due to this, airlines eventually dropped airfare prices. This resulted in a decrease in starting salary for positions within the airline industry, this then pushed away young professionals that may have been interested initially. As of now a majority of the pilots within the workforce are comprised of Baby Boomers that are within 20 years of being forced into retirement.

Flying is statistically the safest way to travel, however Pilots have an extremely high occupational fatalities ratio. Studies in 2017 have proven that being a Pilot is the 4th most dangerous job in America, one spot below Sailors and Marine oilers and one spot above Paving, surfacing, and tampering equipment operators. Pilots had the fourth most total occupational deaths in 2017, which results in about 59 deaths. This means that the fatal injury rate per 100,000 workers is 50.4 this rate is extremely high for an industry already facing a shortage of employees. Yet, there is only and one in an eleven million chance of the plane crashing. These statistics are very important to the airline industry as absenteeism and increased turnover rate have always been directly correlated with employee morale. Most businesses have re-created corporate culture as a way to increase morale. However, this is difficult to accomplish in the airline industry where the workforce is diminishing due to age and health concerns. In addition, the demand for flights is estimated to reach an all-time high this year, after topping 1 Billion flights in 2018 (Foreign and Domestic) with an increase of about 4.8% from 2019, data is proving that over time this deficit will increase. Currently commercial pilots are scheduled to work "On average" 75 hours per month, with an additional 150 hours spent on pre planning and reviewing flight plans. Being a pilot is an extremely demanding job, external factors such as culture are no incentive as the position itself is becoming overlooked by potential employees.

Aircraft arriving late: An example of this situation is when a delayed flight will cause the next scheduled flight to depart late. This situation causes what most industries refer to as "The domino or Snowball effect." It's defined as one small action that automatically sets off a chain reaction that continues to roll downhill. It's a set of events affects thousands of people, across the country. It is usually triggered by a small error. When a carrier is delayed each trip, passenger, or staff member's day is shifted to the right. For a traveler flying from New York City earlier in the day, their flight might be delayed temporarily by 15 or more minutes. However, for travelers planning to catch the last flight of the night, the snowball effect can lead to cancelled or diverted flights. This directly impacts passengers, who then might be forced to change their itinerary.

1.3 Chosen Data

a) Dataset

The dataset extracted illustrates All the Major Airports on Time Arrival Performance across the Nation from September 2010 – September 2019. By analyzing the data, it is understood that over the past 10 years there were exactly 56 Million flight operations within the nation. Within those 10 years only 79.5% percent of flights arrived without a delay. This results in a 20.5% chance of having to stay in the airport longer than necessary. In order to create a sample of the dataset, the data will be further explored to focus on the 3 Major Airports; JFK, EWR, and LGA. The dataset incorporates more than 3,300 rows providing an adequate amount necessary to conduct the study.

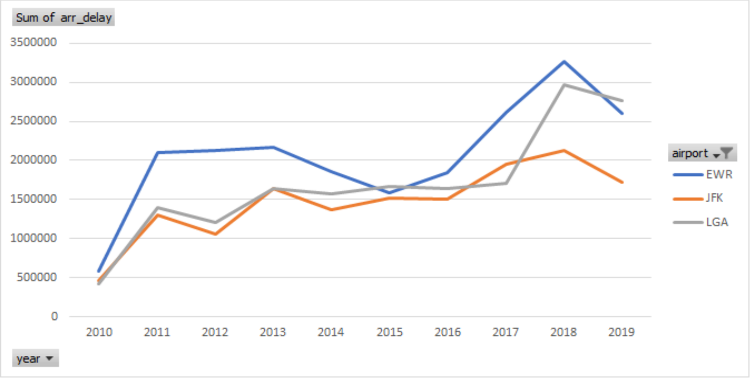
b) Tools incorporated

Microsoft Excel features calculation, graphic tools, pivot tables, and a macro programming language called Visual Basic for Applications. It can display data as line graphs, histograms and charts [5]. This study utilizes multiple functions of Excel including descriptive statistics analysis, pivot tables, t-test, regression etc.

2. Data Visualization and Exploration

2.1 Classification by different airports

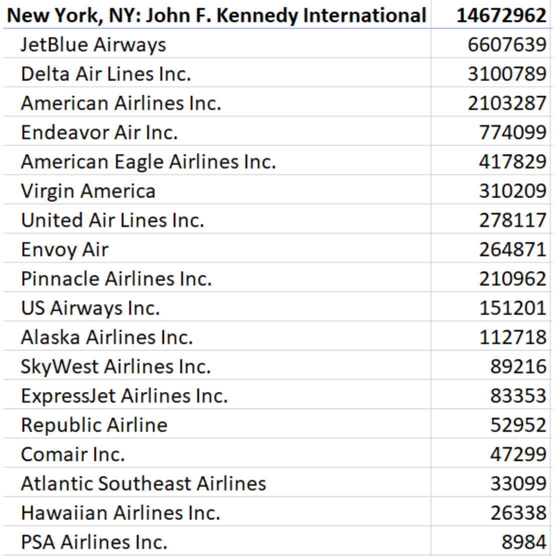
Based on the dataset, it is visible that flight delays in major New York airports have been on the rise since September 2010. Yet, between 2018 and 2019 there was a sharp decline. By further analysis and exploration of the data set and additional contributing factors, we may be able to understand the trend visible.



Annual Flight Delays number in major New York airports (2010.09-2019.09)

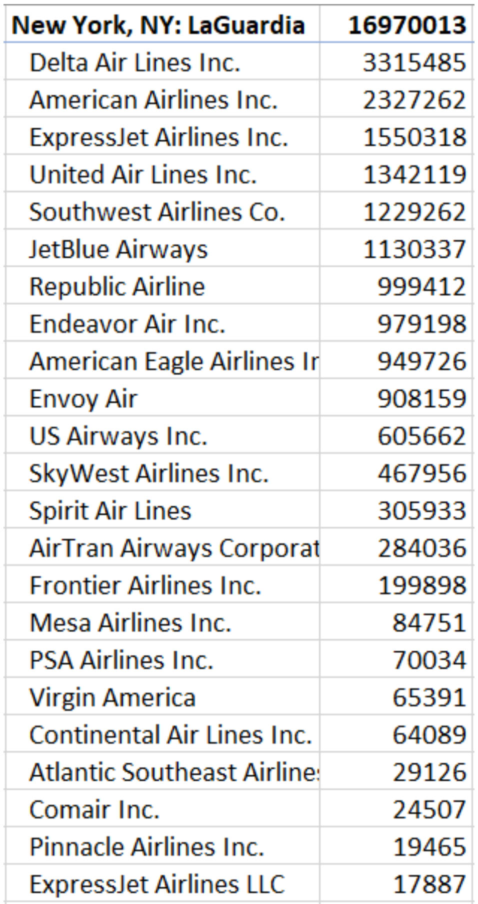
2.2 Clarification by different airlines companies

By the annual flight delays, the annual airline delays at EWR almost surpassed JFK and LGA in the past ten years. We assume that the airline companies in EWR airport are different from airline companies in JFK airport and LGA airport, and it is possible that the airline companies in EWR airport may be easier to delay. To testify this assumption, we can classify airlines in the three airports.



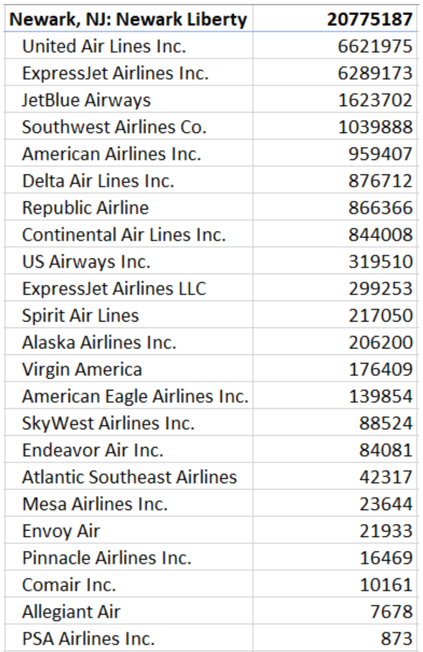
New York, NY: John F. Kennedy International	14672962
JetBlue Airways	6607639
Delta Air Lines Inc.	3100789
American Airlines Inc.	2103287
Endeavor Air Inc.	774099
American Eagle Airlines Inc.	417829
Virgin America	310209
United Air Lines Inc.	278117
Envoy Air	264871
Pinnacle Airlines Inc.	210962
US Airways Inc.	151201
Alaska Airlines Inc.	112718
SkyWest Airlines Inc.	89216
ExpressJet Airlines Inc.	83353
Republic Airline	52952
Comair Inc.	47299
Atlantic Southeast Airlines	33099
Hawaiian Airlines Inc.	26338
PSA Airlines Inc.	8984

Annual Airline Delays number by airports – JKF airport



New York, NY: LaGuardia	16970013
Delta Air Lines Inc.	3315485
American Airlines Inc.	2327262
ExpressJet Airlines Inc.	1550318
United Air Lines Inc.	1342119
Southwest Airlines Co.	1229262
JetBlue Airways	1130337
Republic Airline	999412
Endeavor Air Inc.	979198
American Eagle Airlines Ir	949726
Envoy Air	908159
US Airways Inc.	605662
SkyWest Airlines Inc.	467956
Spirit Air Lines	305933
AirTran Airways Corporat	284036
Frontier Airlines Inc.	199898
Mesa Airlines Inc.	84751
PSA Airlines Inc.	70034
Virgin America	65391
Continental Air Lines Inc.	64089
Atlantic Southeast Airline:	29126
Comair Inc.	24507
Pinnacle Airlines Inc.	19465
ExpressJet Airlines LLC	17887

Annual Airline Delays number by airports – LGA airport

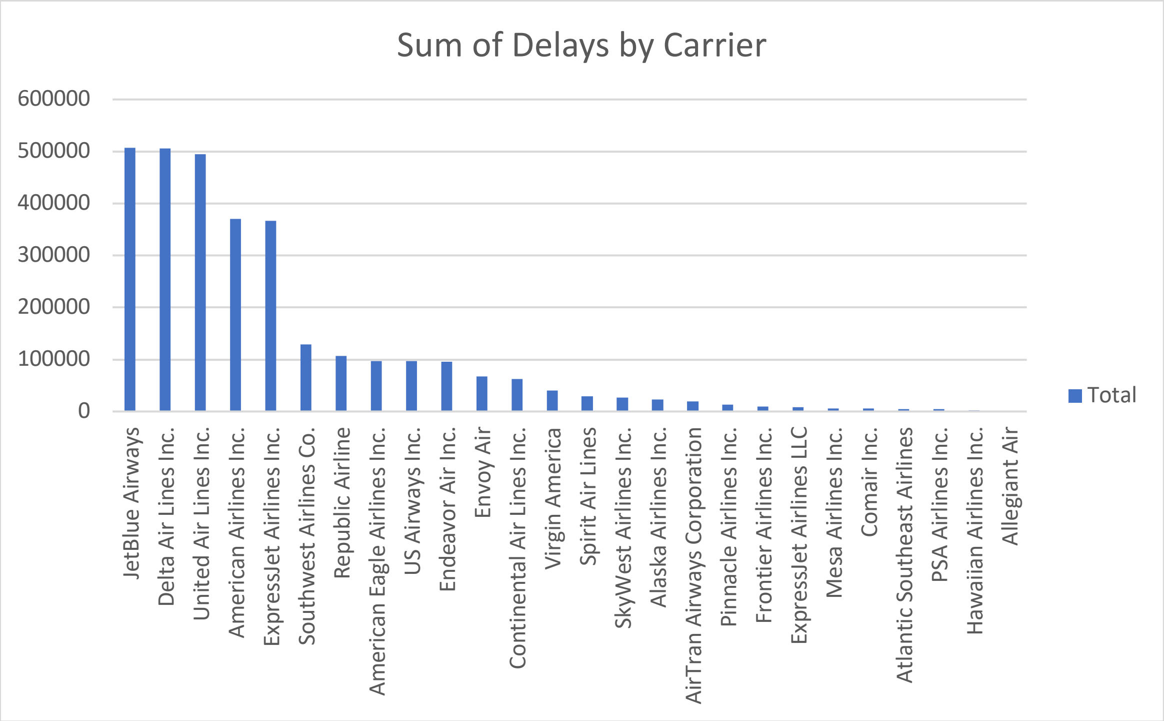


Newark, NJ: Newark Liberty	20775187
United Air Lines Inc.	6621975
ExpressJet Airlines Inc.	6289173
JetBlue Airways	1623702
Southwest Airlines Co.	1039888
American Airlines Inc.	959407
Delta Air Lines Inc.	876712
Republic Airline	866366
Continental Air Lines Inc.	844008
US Airways Inc.	319510
ExpressJet Airlines LLC	299253
Spirit Air Lines	217050
Alaska Airlines Inc.	206200
Virgin America	176409
American Eagle Airlines Inc.	139854
SkyWest Airlines Inc.	88524
Endeavor Air Inc.	84081
Atlantic Southeast Airlines	42317
Mesa Airlines Inc.	23644
Envoy Air	21933
Pinnacle Airlines Inc.	16469
Comair Inc.	10161
Allegiant Air	7678
PSA Airlines Inc.	873

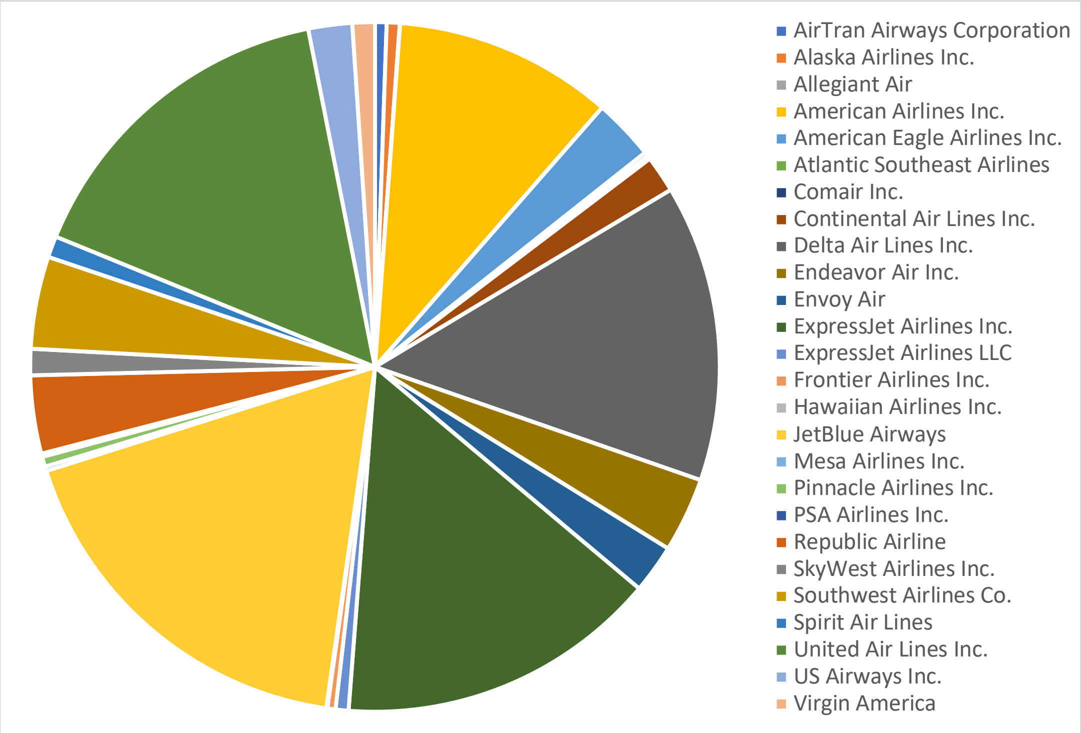
Annual Airline Delays number by airports – EWR airport

By observing the tables, it is visible that the top three airlines airlines with the most delays at EWR are United Air Lines Inc., ExpressJet Airlines Inc., JetBlue Airways. These three airlines were also amongst the top airlines for the worst delays at JFK and LGA. Yet, cannot conclude that EWR witnessed a larger number of delays just due to the top three airlines ranking poorly, there are many other factors that contribute to a high volume of delays at EWR.

By classifying the airlines, it is noticeable that certain companies have higher delays times in all the three airports, and certain companies have less airline delays times in all the three airports. As shown below, when the data of all the airports are taken into consideration, the rankings of all based on delays are visible. JetBlue Airways being the worst and Allegiant air being the best.



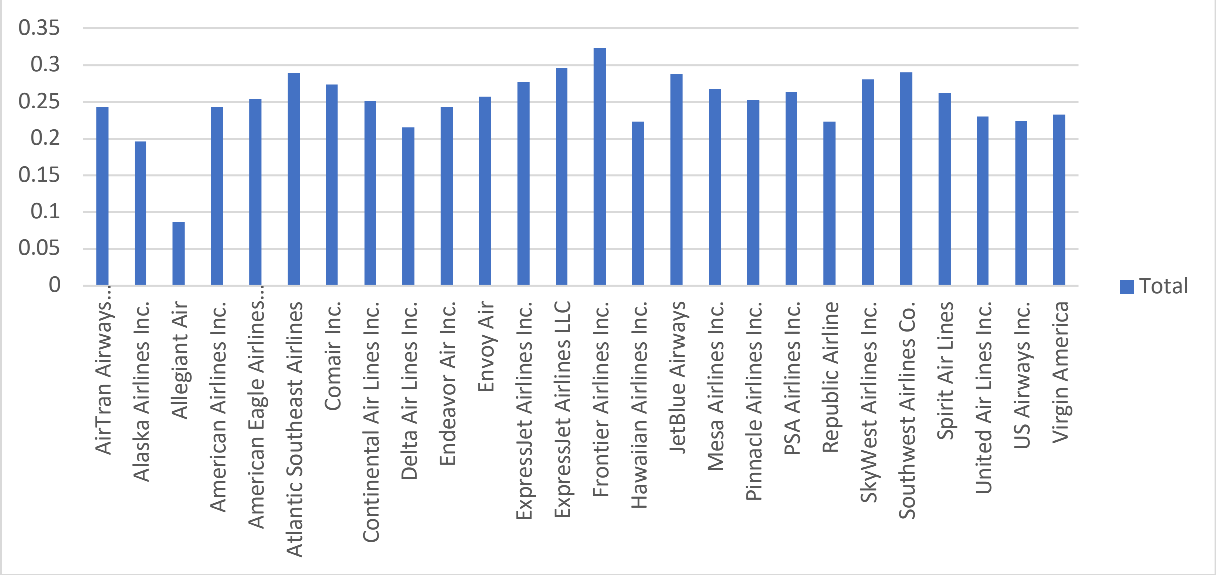
Column chart - Annual Airline Delays divided by airline companies in the past ten years



Pie chart - Annual Airline Delays divided by airline companies in the past nine years

Based on the above it is visible that United Air Lines Inc., JetBlue Airways, ExpressJet Airlines Inc., Delta Air Lines Inc have the larger proportion of delays. This proves that these airline companies have more annual airline delays times than other airlines companies.

Yet, we have to consider that the airlines with the most delays might be the airlines with the highest annual volume of flights. To further study this, we made a slight change to the original data [6], This was done in order to calculate the delays in proportion to annual flights.

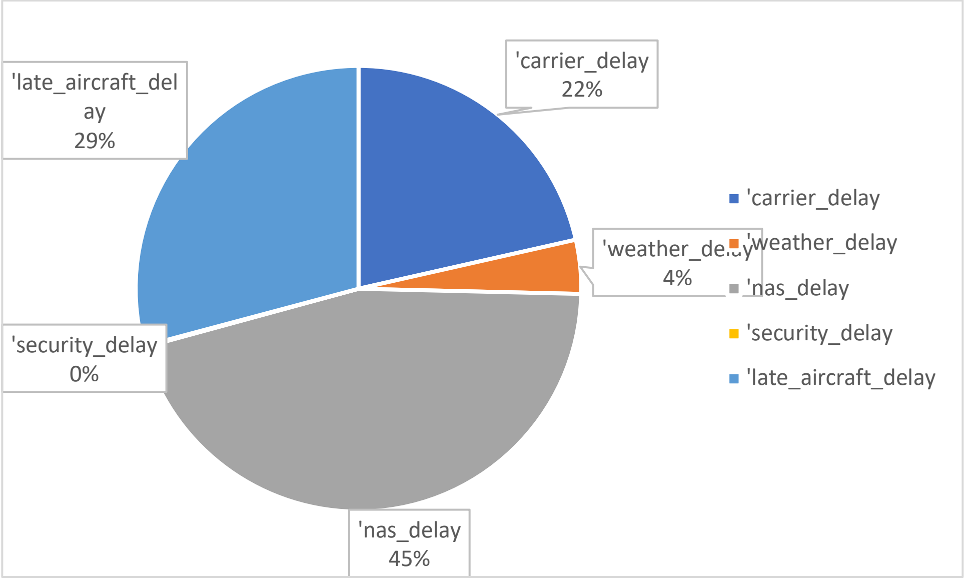


Column chart – Airline delays proportion divided by airline companies in the past nine years

Even though the numbers of total delays for each individual airline are vastly different from each other, the general delay rate does not vary much. Frontier Airlines Inc. has the highest delay rate, and ExpressJet Airlines Inc., JetBlue Airways fall almost just as high. However, United Air Lines Inc., which is known for having the largest delay times, has moderate delays rate. This shows that the volume of flights per carrier has a direct relation to their rate of delay. If an airline has a high volume of flights, especially during high traffic seasons they are prone to have a higher rate of delay.

2.3 Clarification by different causes

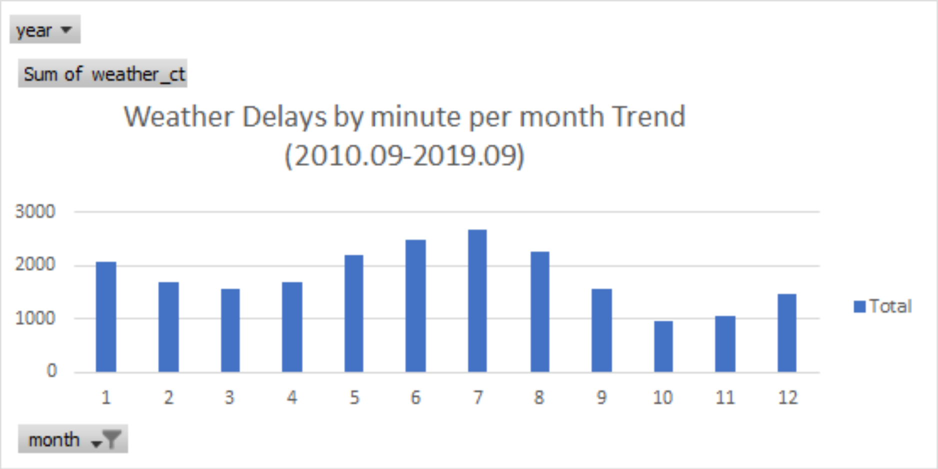
Taking a look at our major causes of airline delay; security, national aviation system delays, carrier delays, aircraft arriving late and extreme weather.



Proportion of five main causes of airline delays

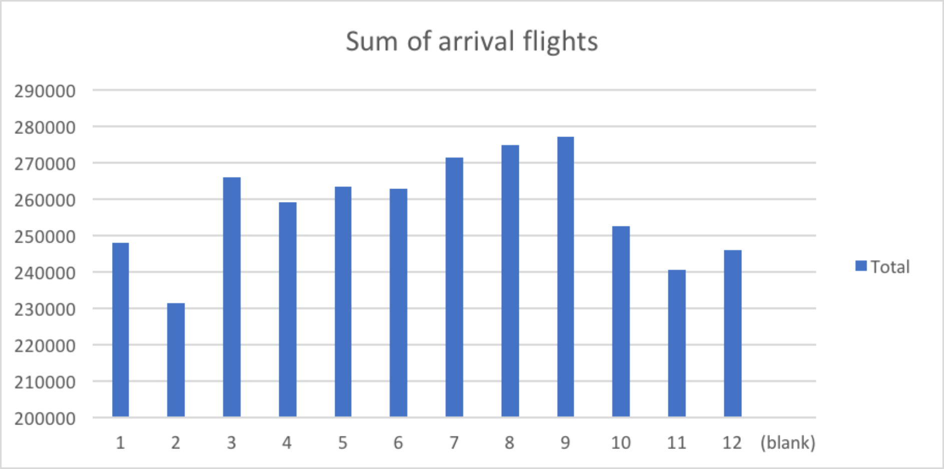
Based on the figure above, it is apparent that National Aviation System delays is the most significant reason resulting in airline delays, with a majority portion of 45%. Late arrival and delays are play an important with 29% and 22%. Extreme weather, which seems to be the most common reason of airline delays, merely constitutes 4%. Based on this information it is visible that the biggest concern for airlines are controllable factors, yet the delays are still on the rise. The commercial industry is a long way from reducing delays.

Even though extreme weather has a low rate of impact, every year the summer season in New York brings chaos at the airports. Just based on general inference travelers may believe that airline delays due to weather will most commonly occur in the winter, between the months of November to March where harsh conditions such as heavy snow storms may be present. Our data clearly denies that, it proves that the most delays due to weather in New York occur in July. There is a peak in the summer months and then a decline in winter months. According to Fortune’s Executive Travelers report, JFK and in general New York City airports are a part of the top 10 airports across the nation where summer delays are worse than winter delays. This is because these airports see heavy traffic during the warm months, and weather conditions such as thunderstorms are “a recipe for missed connections”. In addition, “JFK has a 71% on time rating in the winter months, in the summer it drops by almost 4%”. Lastly, New York is one of the wettest cities in the U.S during the summer months, so thunderstorms are bound to drive up the delays. The graph below shows the average of weather delays by minute for each month for the ten-year period.



Weather Delays by minute per month Trend (2010.09 – 2019.09)

The number of flights arriving at JFK, EWR and LGA are relatively similar throughout the months, the summer months still remain on the top. As discussed earlier, there is a high rate of individuals travelling from New York to other locations in the summer, as compared to the winter months.

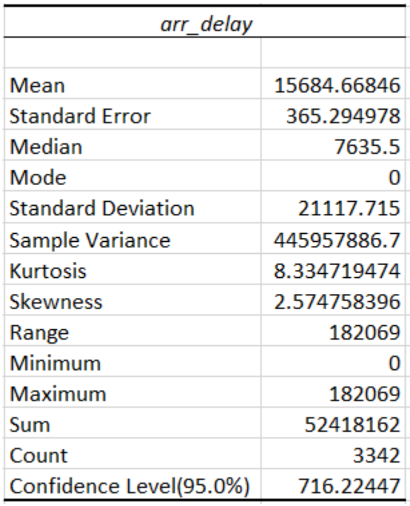


Summary of arrival flights (2010.09 -2019.09)

3. Descriptive Statistical Measures and Statistical Inference

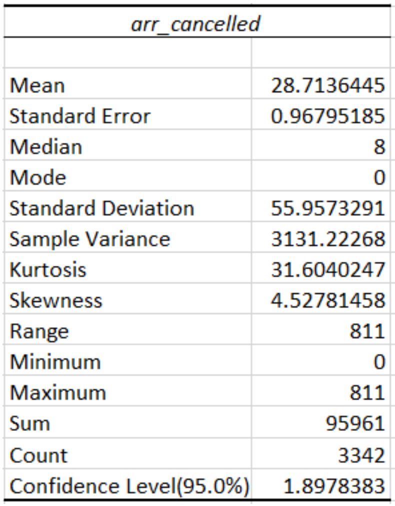
3.1 Descriptive Statistical Measures

Being aware of the average delays times, cancelation times as well as diversion information can be of great significance for travelers. Therefore, we used descriptive analytics to study the descriptive statistical figures.



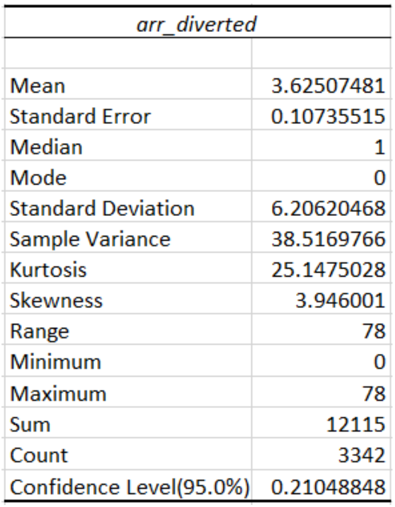
arr_delay	
Mean	15684.66846
Standard Error	365.294978
Median	7635.5
Mode	0
Standard Deviation	21117.715
Sample Variance	445957886.7
Kurtosis	8.334719474
Skewness	2.574758396
Range	182069
Minimum	0
Maximum	182069
Sum	52418162
Count	3342
Confidence Level(95.0%)	716.22447

Airline Delays



arr_cancelled	
Mean	28.7136445
Standard Error	0.96795185
Median	8
Mode	0
Standard Deviation	55.9573291
Sample Variance	3131.22268
Kurtosis	31.6040247
Skewness	4.52781458
Range	811
Minimum	0
Maximum	811
Sum	95961
Count	3342
Confidence Level(95.0%)	1.8978383

Airline Cancelations



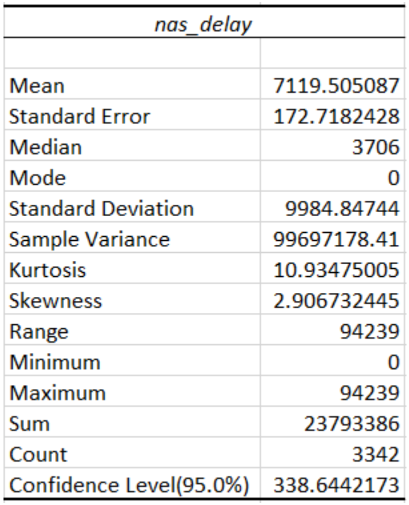
arr_diverted	
Mean	3.62507481
Standard Error	0.10735515
Median	1
Mode	0
Standard Deviation	6.20620468
Sample Variance	38.5169766
Kurtosis	25.1475028
Skewness	3.946001
Range	78
Minimum	0
Maximum	78
Sum	12115
Count	3342
Confidence Level(95.0%)	0.21048848

Airline Diversions

From the figures above it is established that the median is less than the mean, and standard deviation of three tables is large. We can assume that in most years covered by the dataset, airline delays, cancelations and diversions happened less than 15685, 29, and 4 times. However, in several years, the delays, cancelation and diversions are abnormal and affect the overall data.

3.2 Statistical Inference

From the proportion of flight delay causes, it is visible that majority of the cause of delay is due to National Aviation Delays. Since this information is crucial to the delay statistics, we will further study and analyze this information.



nas_delay	
Mean	7119.505087
Standard Error	172.7182428
Median	3706
Mode	0
Standard Deviation	9984.84744
Sample Variance	99697178.41
Kurtosis	10.93475005
Skewness	2.906732445
Range	94239
Minimum	0
Maximum	94239
Sum	23793386
Count	3342
Confidence Level(95.0%)	338.6442173

National Aviation Delays

Through the use of descriptive analytics, the calculated mean of national aviation system delays is 7119.51, which reveals that in the past ten years, national aviation system delays happened 7,119.51 times on average per year. Due to the calculated mean, it is understood that some years the delay on average was greater than 7,119.51 and some were less. Based on this deduction and the existence of standard error, we can test the following hypothesis.

H_0: average times of national aviation system delays per year ≤ 7,300

H_1: average times of national aviation system delays per year > 7,300

The t-value is calculated using the following formula:

$$t = \frac{\underline{x} - u_0}{s/\sqrt{\underline{n}}}$$

According to the formula, t = -1.45

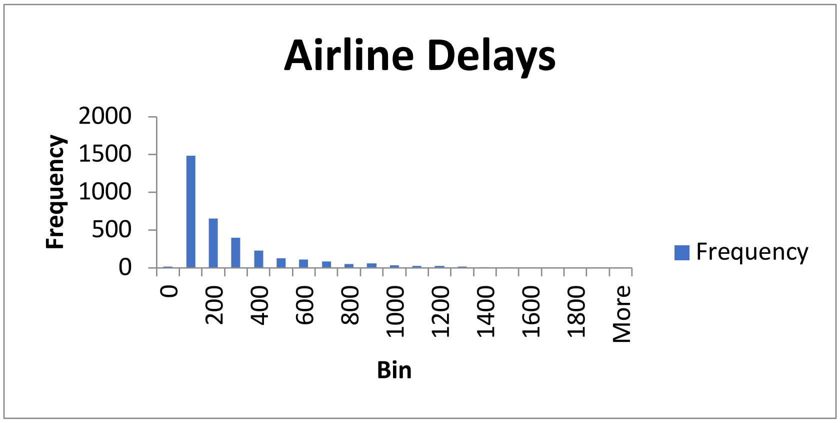
Critical values are ±1.96 according to excel.

Since the t-test statistic falls between these values, we cannot reject H_0. This creates an indicator that the conclusion, average national system delay times between September 2010 to September 2019 of all airports is larger than 7,300 cannot be made.

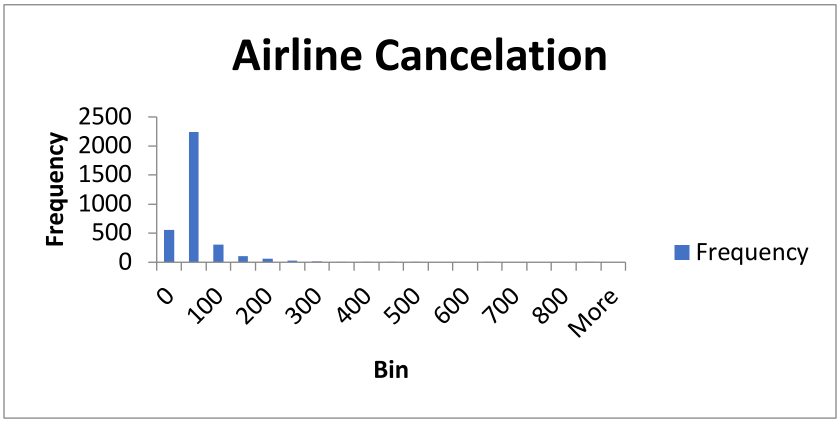
4. Random Sampling and Sampling Estimations

4.1 Random Sampling

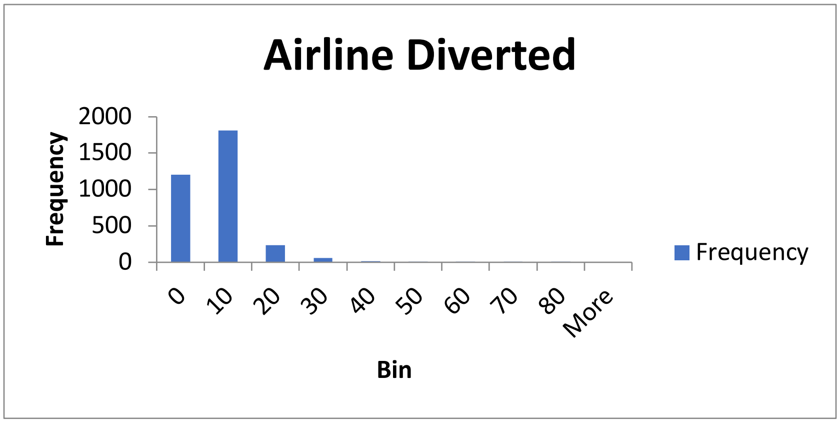
The following histograms will aid in a better understanding of the distribution of airline delays, cancellations and diversions.



Histogram of Airline Delays



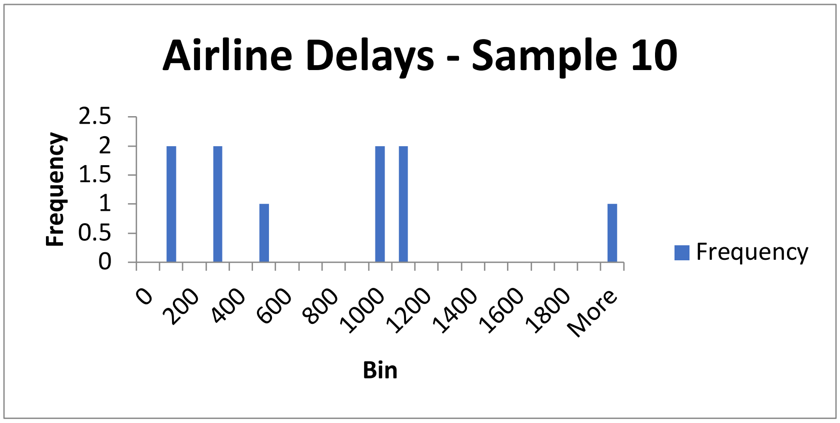
Histogram of Airline Cancelations



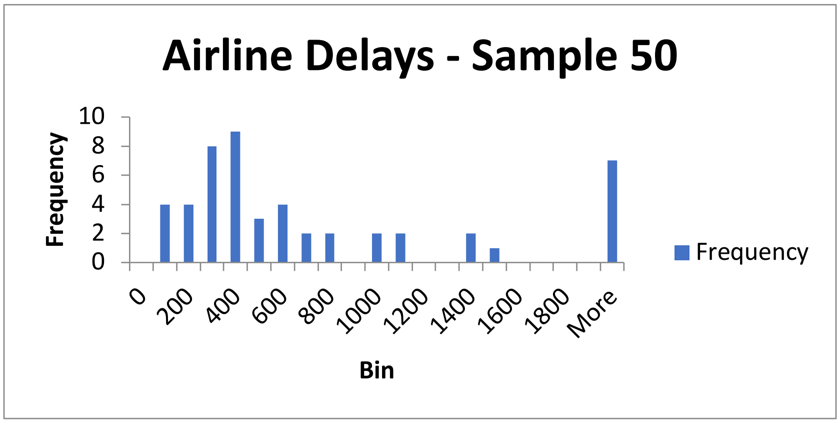
Histogram of Airline Diversions

As shown above, the distribution of airline delays, cancellations and diversions is similar to binomial distribution, and concentrated between 100 to 200 for delays, between 100 to 200 for cancellations, and between 10 to 20 for diversions.

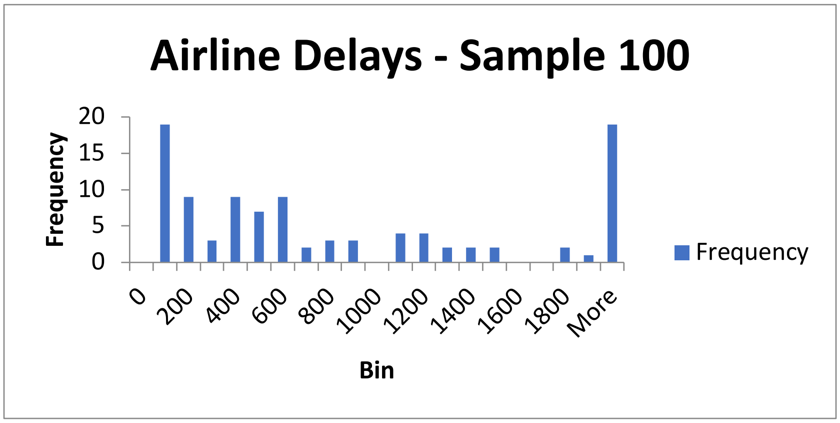
One interesting aspect that brought out curiosity was whether the samples will represent the same information as the histograms and to what extent can samples represent the whole population. As consequence, we decided to select 10, 50, 100 randomly by using sampling tools in the excel. We believe that randomization is a good way to select appropriate samples.



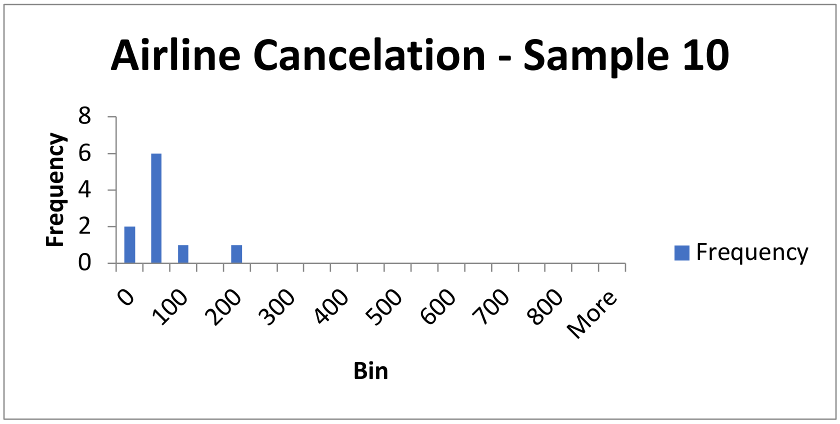
Histogram of Airline Delays – Sample Size 10



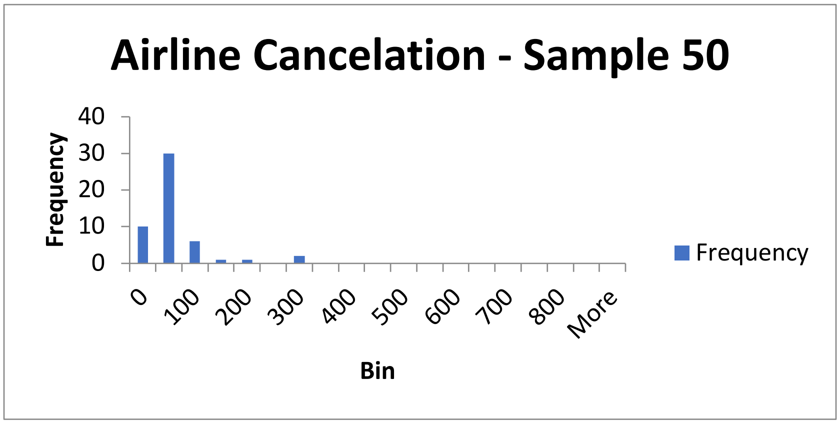
Histogram of Airline Delays – Sample Size 50



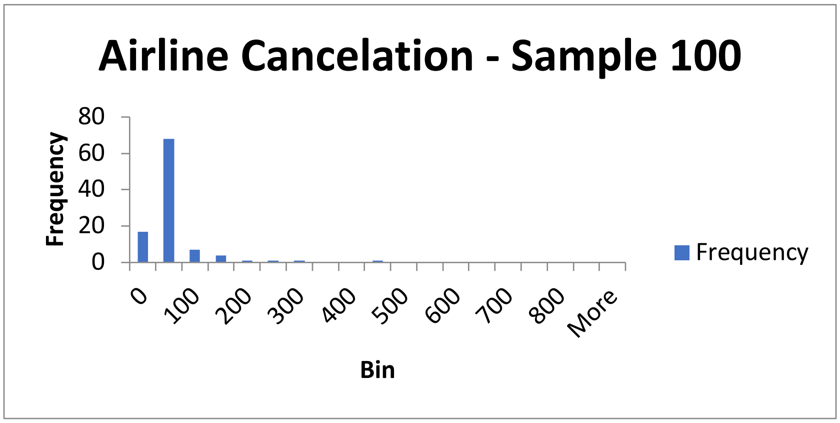
Histogram of Airline Delays – Sample Size 100



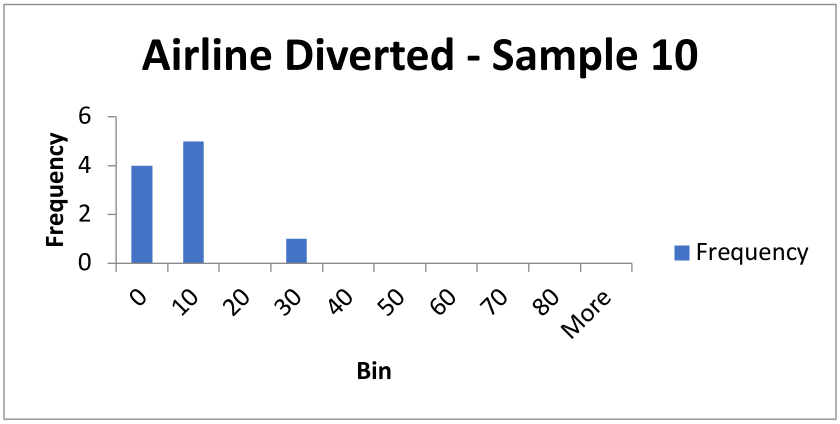
Histogram of Airline Cancelation – Sample Size 10



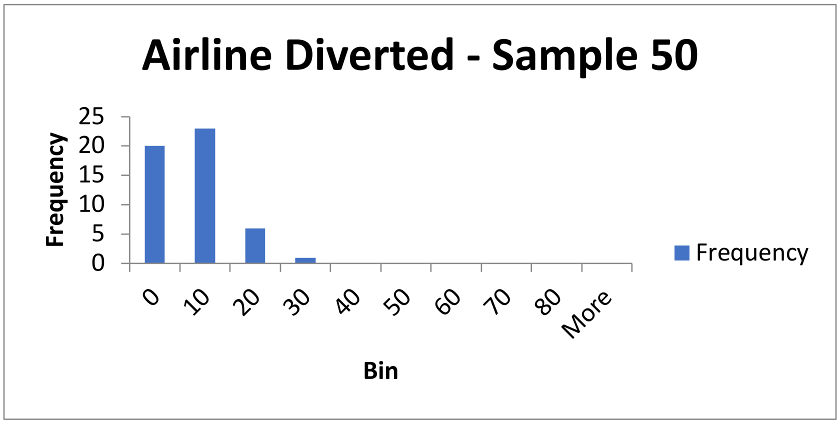
Histogram of Airline Cancelation – Sample Size 50



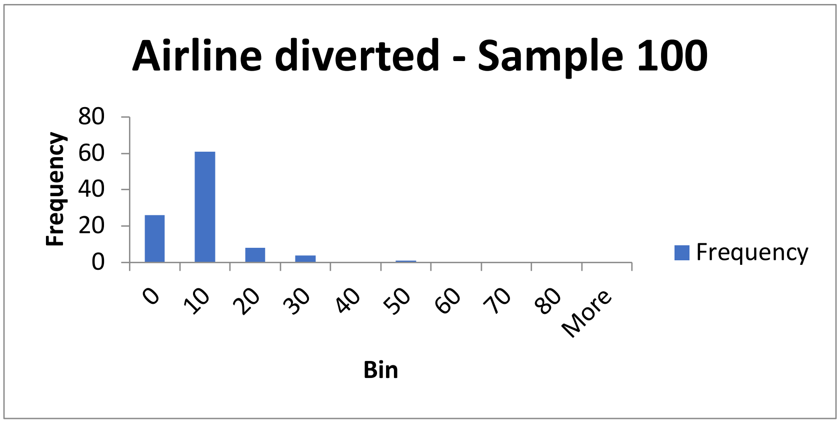
Histogram of Airline Cancelation – Sample Size 100



Histogram of Airline Diverting – Sample Size 10



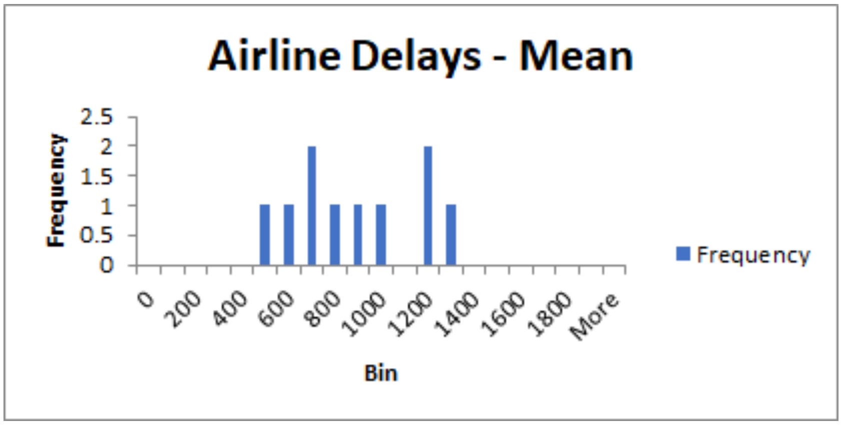
Histogram of Airline Diverting – Sample Size 50



Histogram of Airline Diverting -Sample Size 100

4.2 Sampling Estimations

As displayed in the sample histograms above, when the sample number increases, the distribution begins to represent the population histogram. However, as they are randomly selected they can only be used as estimations. To create a better representation of the whole population by using samples. The next step was to randomly select 10 groups of samples, where each group contains 10 sample sizes. The histogram is shown below.



Histogram of Airline Delays - Mean

The histogram above is not similar to airline delays histogram, consisting all of the data. This is because the population is not normally distributed. Due to this, we were unable to provide an illustration using the sample means.

The dataset covers 3 major airports (JKF, LGA, EWR). Due to this the standard deviation cannot be used as the population standard deviation. Instead of the three major airports, we used the initial dataset, including all airports across the united states to calculate the standard deviation. This was done for the same number of years, September 2010 to September 2019. The result of this calculation was 271.42. [7]. It was calculated using the following formula:

$$n \geq \left(z_{\frac{\alpha}{2}}\right)^2 \frac{\sigma^2}{E^2}$$

The expected the margin of error is ±2%. If this is true, the result will be 18,578.13, we should at least get equal to or greater than 18,579 sample sizes. For the population, the data greater than 140,000, is acceptable. Yet for our sample dataset, we have about 3,300 rows, due to this we will not be able to adapt this.

Instead we will calculate a more appropriate sample size by using the following formula:

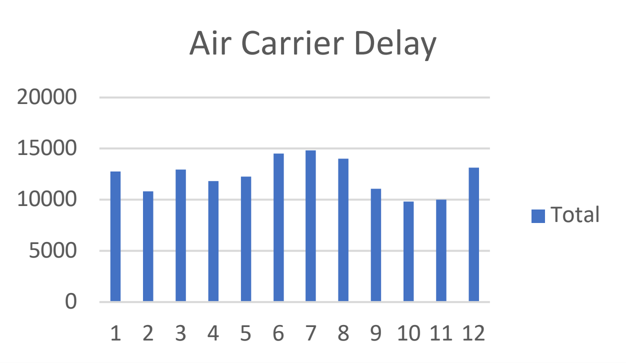
$$n \geq \left(z_{\frac{\alpha}{2}}\right)^2 \frac{\pi(1-\pi)}{E^2}$$

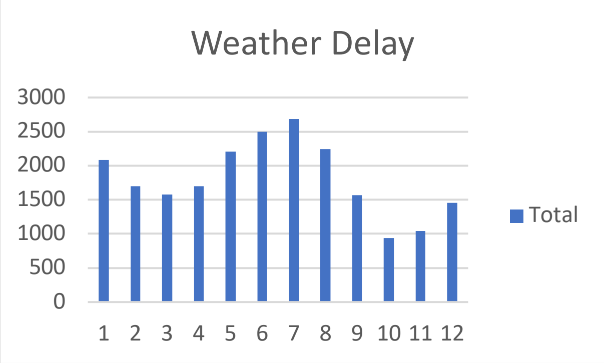
By ensuring that the margin of error is ±2%, and 95% confidence intervals.

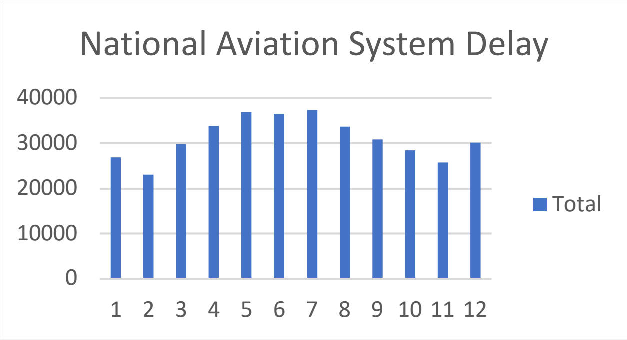
As a result, the n should be more than 2,401, which means we should at least 2,401 sample sizes from our dataset to represent the population.

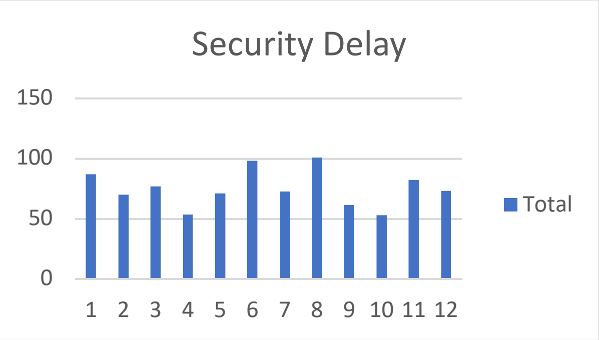
5. Regression Analysis

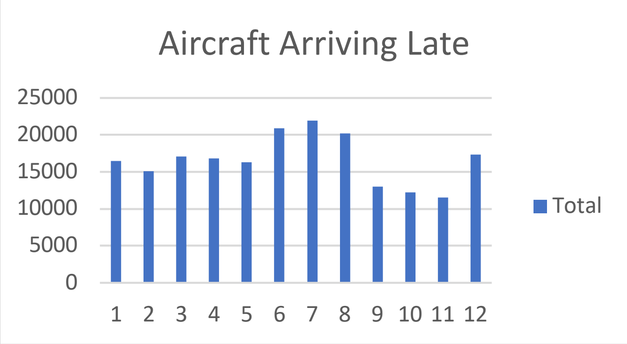
The main reasons negatively affecting airline on-time performance are divided into; air carrier delay, weather delay, National Aviation System Delay, security delay, aircraft arriving late, flight cancellation, as well as airline diversion [8]. The visualizations below can better aid understanding of the distribution of each specific reason throughout each month during past nine years (September 2010 – September 2019)

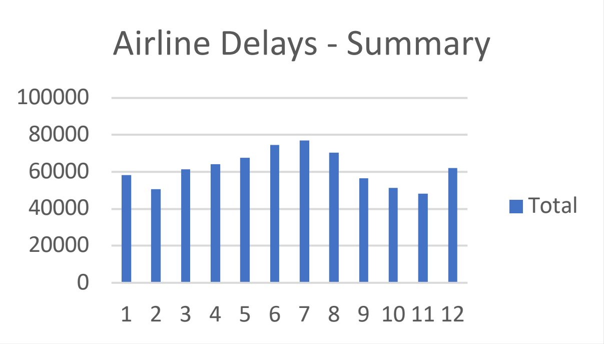


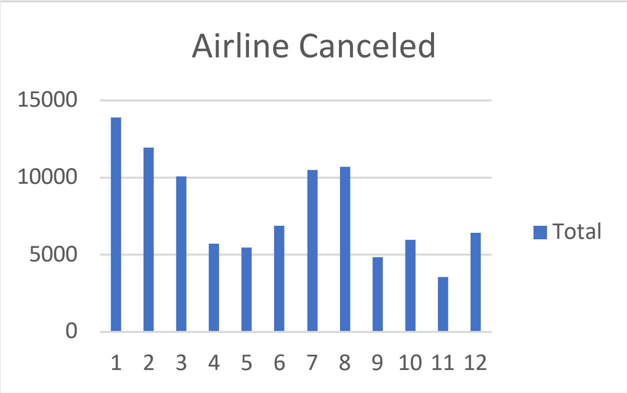


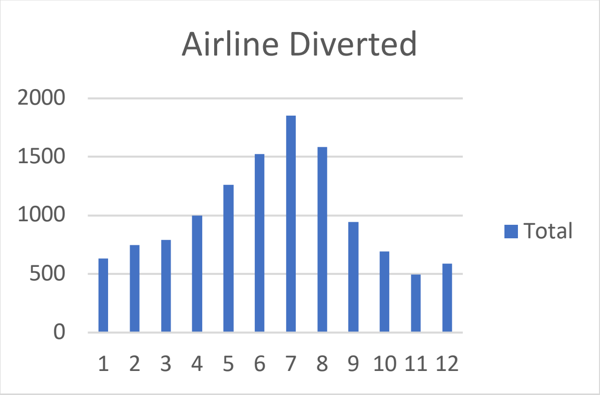










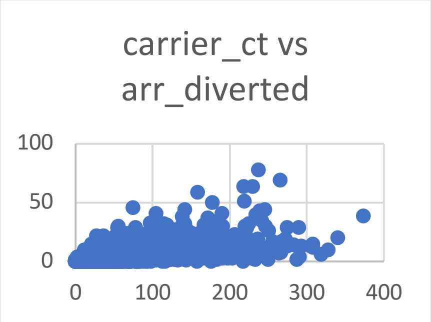


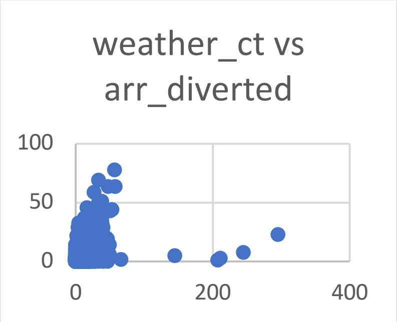
As displayed in the column charts above, the most air carrier delays happened in July, approximately 15,000. June and August also witnessed high air carrier delays. Similarly, weather delays, national aviation system delays, aircraft arriving late mostly happed in June, July and August, with July experiencing highest weather delays, national aviation system delays, and late aircraft arrival. However, security delays are not as high in July but August and June still hold the first position and second position.

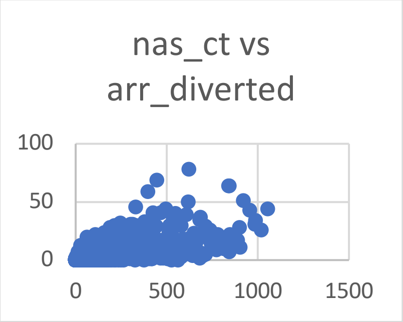
Airline cancelation, however, is colossally different from airline delays, with January being the highest, almost reaching 15,000. The column graphs of airline diversions are similar to normal distributions, with June, July and August being the top three months.

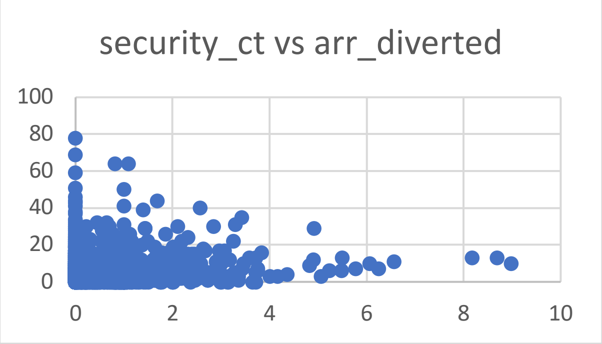
As a consequence of the analysis above, we can assume that airline diversions may be corelated to airline delays, especially positively corelated to weather delays, national aviation system delays and aircraft arriving late. Yet, it may be irrelevant to air carrier delay and security delay. Airline cancellations are likely to be irrelevant to airline delays. To testify our hypothesis, we assume that X_1 = carrier_ct, X_2 = weather_ct, X_3 = nas_ct, X_4 = security_ct, X_5 = late_aircraft_ct, Y_1 = arr_diverted, Y_2 = air_canceled. [9]

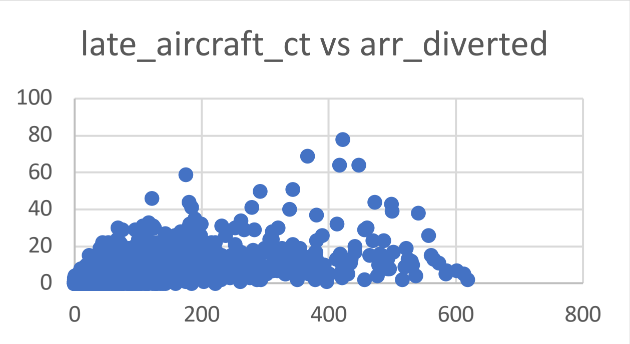
The first step is to use scatter charts to predict the relation between Y_1 and X_1, X_2, X_3, X_4, X_5, Y_2 and X_1, X_2, X_3, X_4, X_5.

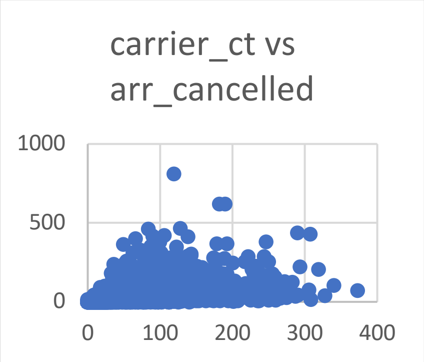


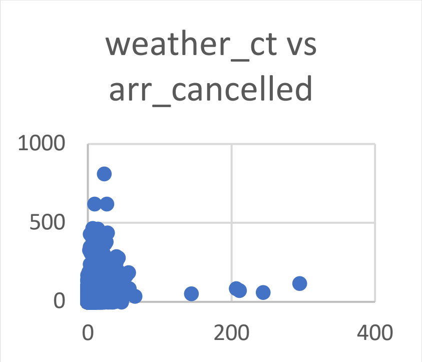


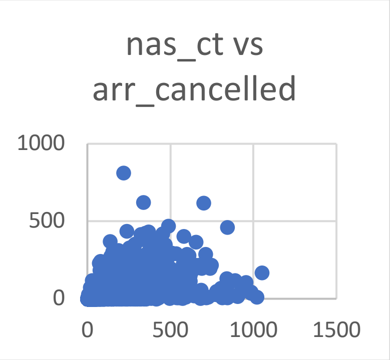


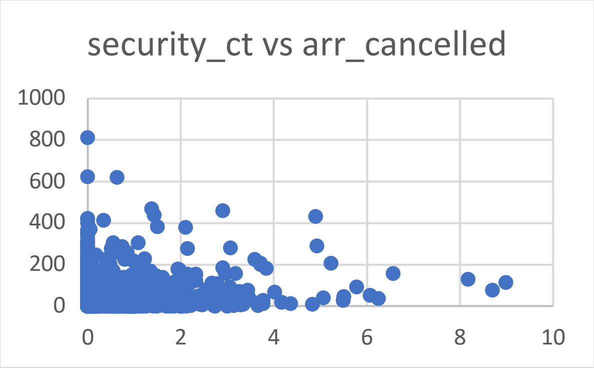


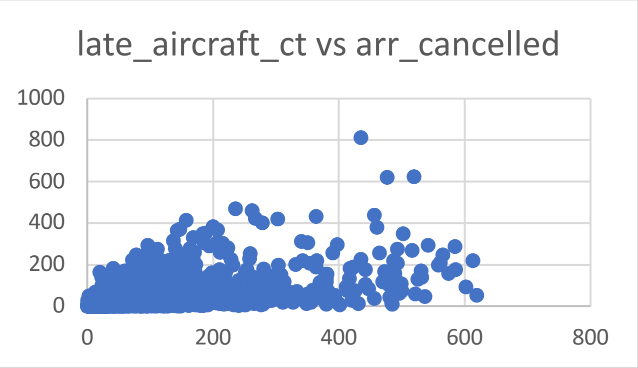












According to the scatter charts, we can assume that the Y_1 has a linear correlation with X_1, X_2, X_3, X_4, X_5, Y_2 has a linear correlation with X_1, X_2, X_3, X_4, X_5. As a result, we will be using the following model.

$$Y1 = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \beta4X4 + \beta5X5$$

$$Y2 = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \beta4X4 + \beta5X5$$

By using data analysis in Excel, we conducted the following result:

Regression Statistics	
Multiple R	0.69178272
R Square	0.47856333
Adjusted R Square	0.4777818
Standard Error	4.48489312
Observations	3342

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	61584.02629	12316.81	612.3418	0
Residual	3336	67101.19244	20.11427		
Total	3341	128685.2187			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.140097	0.105989838	-1.3218	0.186327	-0.34790863	0.0677147	-0.34790863	0.0677147
carrier_ct	0.03793813	0.003528525	10.75184	1.57E-26	0.031019835	0.0448564	0.03101983	0.04485642
weather_ct	0.05775445	0.007521914	7.67816	2.11E-14	0.043006424	0.0725025	0.04300642	0.07250249
nas_ct	0.01869055	0.000955564	19.55971	9.65E-81	0.016817004	0.0205641	0.016817	0.02056411
security_ct	-0.5123152	0.128542469	-3.98557	6.88E-05	-0.76434521	-0.260285	-0.76434521	-0.2602851
late_aircraft_ct	-0.0040513	0.002107353	-1.92248	0.054631	-0.00818317	8.05E-05	-0.00818317	8.0497E-05

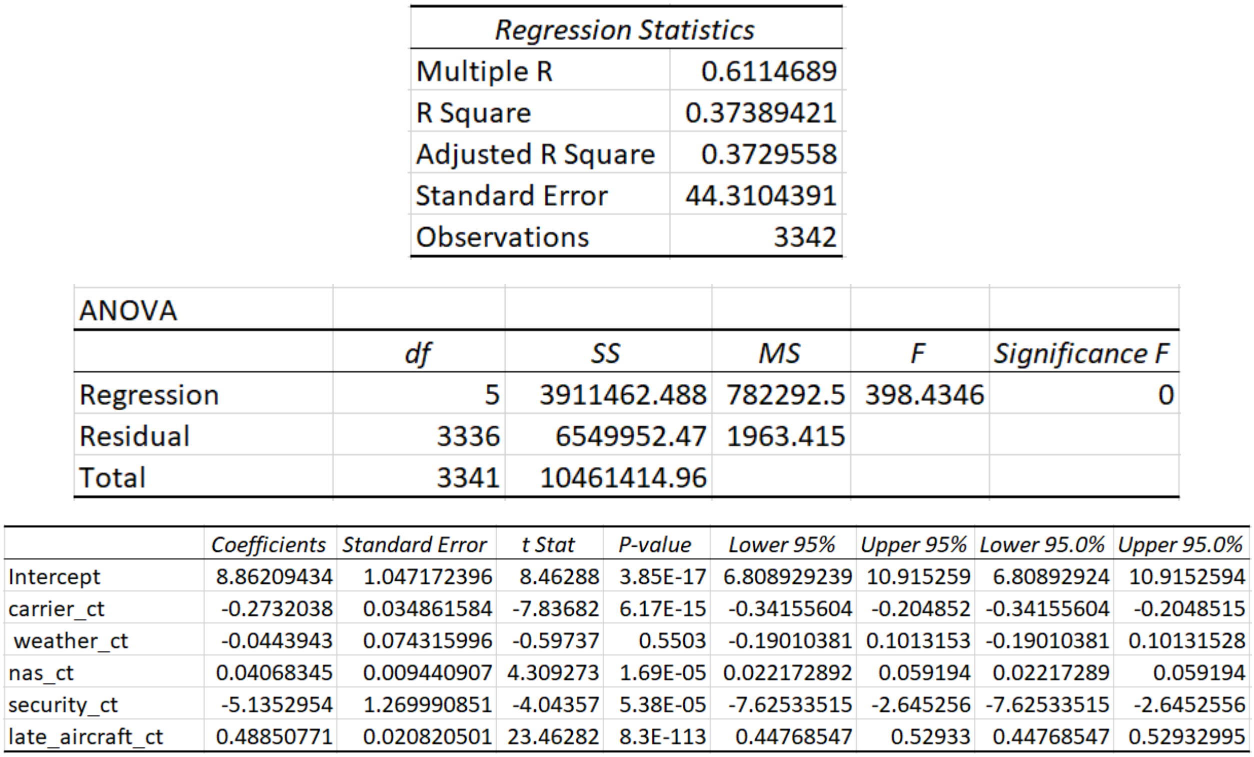
According to the tables above, the equation formula can be written as:

$$Y_1 = 0.03793813X_1+0.05775445X_2+0.01869055X_3-0.5123152X_4-0.0040513X_5-0.140097$$

We can deduce that air carrier delay, weather delay and national aviation system delays are weak-positively corelated to airline diversions. We can also deduce that security delays are negatively correlated to airline diversions. Lastly, aircraft arriving late is an irrelevant independent variable.

By focusing on the P-value of each individual variable, the P-value of carrer_ct, P-value of weather_ct, P-value of nas_ct, and P-value of security_ct are smaller than 0.05, and P-value of late_aircraft_ct is larger than 0.05, therefore we consider that the equation above can be revised as below.

$$Y_1 = 0.03793813X_1+0.05775445X_2+0.01869055X_3-0.5123152X_4-0.140097$$



Regression Statistics	
Multiple R	0.6114689
R Square	0.37389421
Adjusted R Square	0.3729558
Standard Error	44.3104391
Observations	3342

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	3911462.488	782292.5	398.4346	0
Residual	3336	6549952.47	1963.415		
Total	3341	10461414.96			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.86209434	1.047172396	8.46288	3.85E-17	6.808929239	10.915259	6.80892924	10.9152594
carrier_ct	-0.2732038	0.034861584	-7.83682	6.17E-15	-0.34155604	-0.204852	-0.34155604	-0.2048515
weather_ct	-0.0443943	0.074315996	-0.59737	0.5503	-0.19010381	0.1013153	-0.19010381	0.10131528
nas_ct	0.04068345	0.009440907	4.309273	1.69E-05	0.022172892	0.059194	0.02217289	0.059194
security_ct	-5.1352954	1.269990851	-4.04357	5.38E-05	-7.62533515	-2.645256	-7.62533515	-2.6452556
late_aircraft_ct	0.48850771	0.020820501	23.46282	8.3E-113	0.44768547	0.52933	0.44768547	0.52932995

We originally assumed that airline cancellations are irrelevant to these variables, after using regression analysis method, we can deduce that

$$Y_2 = -0.2732038X_1-0.0443943X_2+0.04068345X_3-5.1352954X_4+0.48850771X_5+8.86209434$$

According to the result of the regression analysis, the P-value of X_1, X_3, X_4, X_5 is smaller than 0.05, hence we are able to use these variables. However, the P-value of X_2 is larger than 0.05, therefore we can conclude that it is an irrelevant variable.

Consequently, we modify our equation formula as below,

$$Y_2 = -0.2732038X_1+0.04068345X_3-5.1352954X_4+0.48850771X_5+8.86209434$$

From the equation, we can tell that security has a strongly negative correlation with airline cancelation, which is reasonable because it is safer to cancel a flight once safety has become a concerning issue.

6. Conclusion

Airline delays, cancellations, and diversions have become a common event at New York airports. On a macro level, they have negative impacts on industry's economy. This may then result in a passive influence on the countries GDP in the long run. To a micro level, travelers find it disturbing that their flight is delayed, canceled or diverted. In addition, airline delays, cancellations and diversions will hurt reputation of the airlines and lower customer as well as employee satisfaction rates.

Based on the analysis conducted throughout the study it is visible that the majority of airline delay causes occur the most during the summer season, between June and August. It would be beneficial for customers to understand this information and be better prepared as there is a high volume of individuals who fly out during this season. In addition, if anyone is travelling for any urgent requirement or business it is best to avoid Newark Aiports as well as Delta Air Lines, JetBlue Airways, United Air Lines, and ExpressJet Airlines in the upcoming summer season. In addition, travelers should double check weather forecasts and flight status' before leaving their home to avoid unnecessary congestion at the airports, especially in January where flight cancellations are at their highest. We believe that if travelers are equipped and able to make better decisions regarding their flights it will be easier for flights to be on time. If such small errors and minimal delays are avoided, the airlines will be able to save large amounts of money which may then allow them to combat National Aviation delays and system improvements which they still have not been able to do. This will then allow better infrastructure and increased airport employee moral who will no longer be worn down by frustrated crew members and passengers, leading to shorter delay counts and times. Even though certain influential factors are unavoidable, this paper provides useful insights for the aviation industry and passengers to take actions in the future in order to manage and avoid delays.

Bibliography

[1] https://www.cnbc.com/2019/08/15/more-us-airline-passengers-are-facing-canceled-and-oversold-flights.html

[2] Air Travel Consumer Report, Issued: November 2019, https://www.transportation.gov/sites/dot.gov/files/docs/resources/individuals/aviation-consumer-protection/357186/november-2019-atcrr.pdf

[3] Ann, B.G. (2010). Flight delays cost $32.9 billion, passengers foot half the bill. Berkeley News October 18, 2010.

[4] Why 40 Year Old Tech is Still Running Americas Air Traffic Control - https://www.wired.com/2015/02/air-traffic-control/

[5] Wikipedia, https://en.wikipedia.org/wiki/Microsoft_Excel

[6] Refer to "Dataset – Overall Population"

[7] Refer to Excel File "Dataset - 2010.09~2019.09 All Airport"

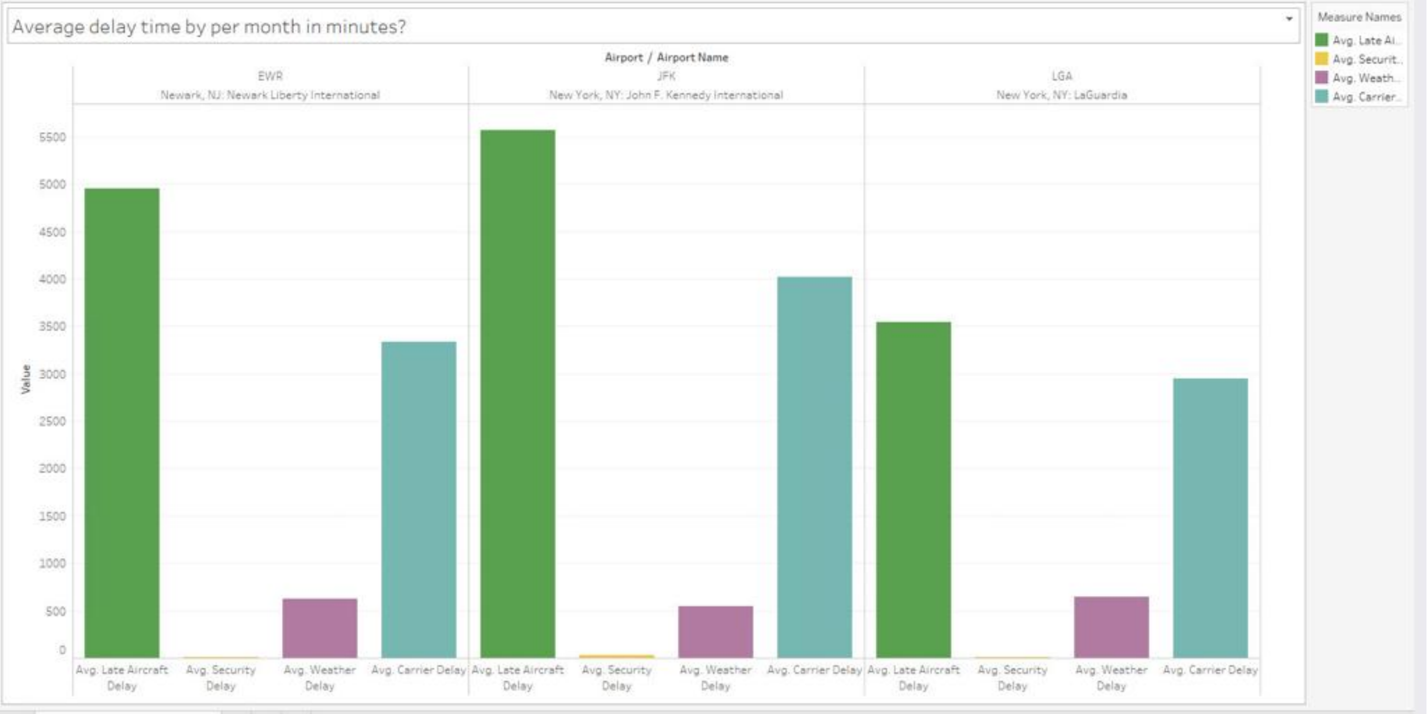
[8] Bureau of Transportation Statistics, https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1

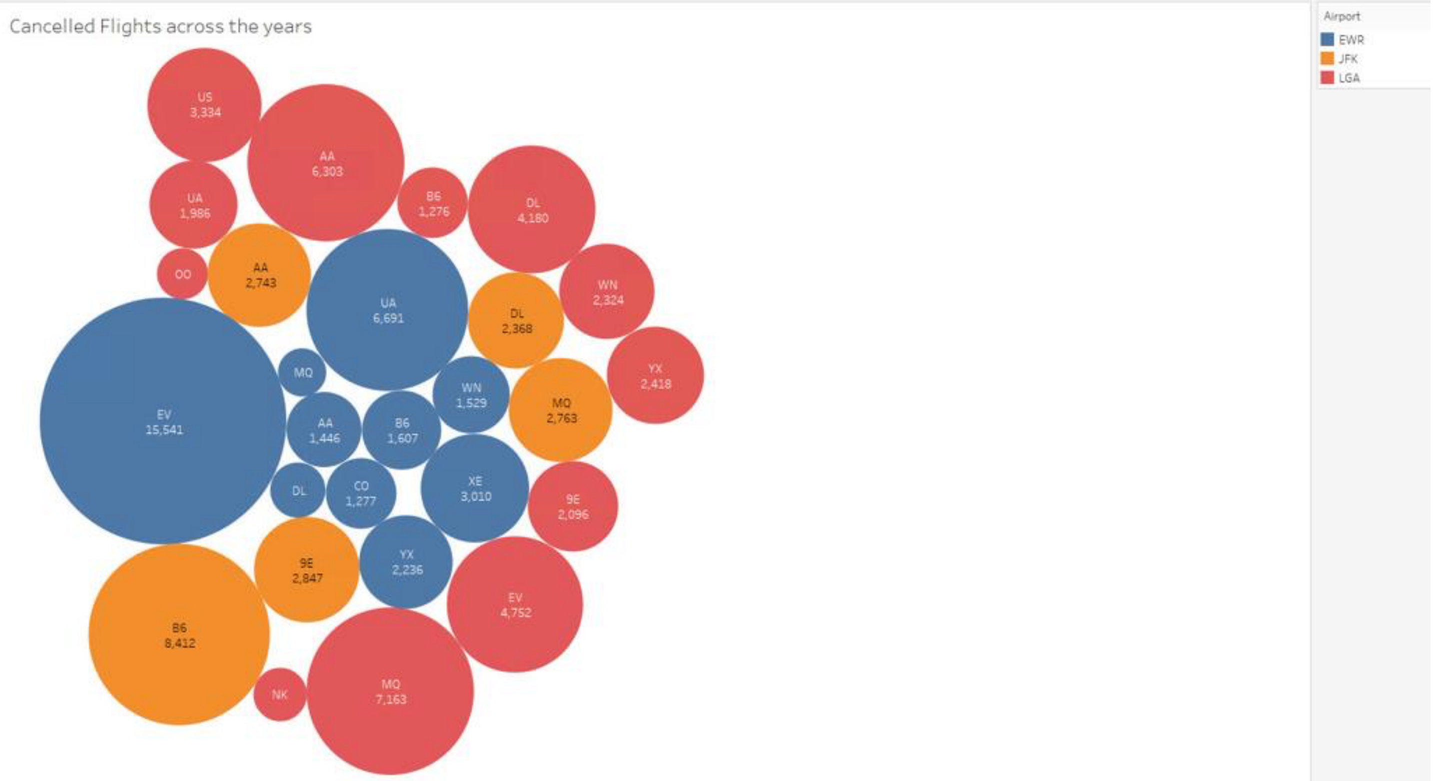
[9] Refer to Excel File "Dataset – Overall population"

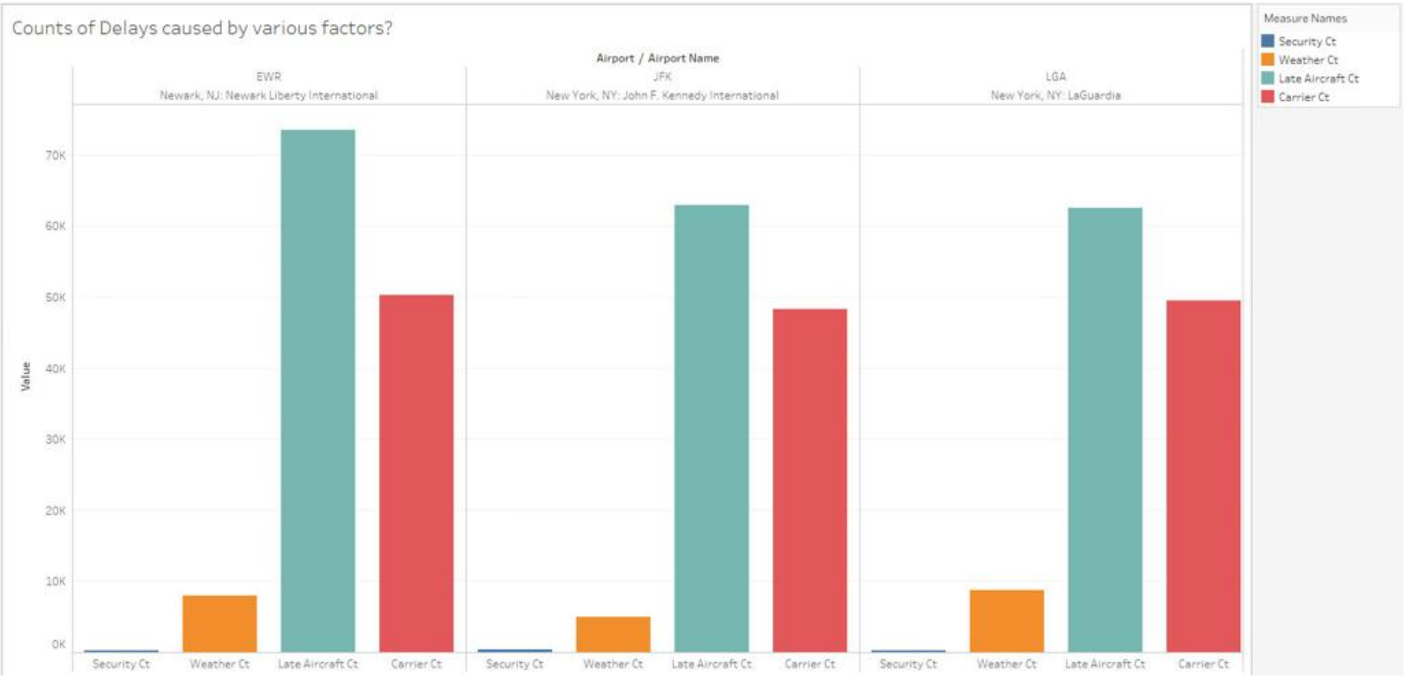
Dataset Dictionary

1. Carrier - carrier number
2. Carrier_name – name of airline
3. Airport - name of airport
4. Airport_name - location of airport
5. arr_flights - no. of arrival flights
6. arr_del15 - Arrival Delay Indicator, 15 Minutes or More (1=Yes)
7. weather_ct – count of weather delays
8. nas_ct – count of national aviation system delays
9. security_ct – count of security delays
10. Late_aircraft_ct – count of delays due to a late aircraft
11. Arr_cancelled - arrival cancellation
12. arr_diverted - arrival diverted
13. arr_delay - arrival delay in minutes
14. carrier_delay - Carrier Delay, in Minutes
15. Weather_delay - Weather delay, min
16. nas_delay - National Air System Delay, in minutes
17. security_delay - security delay, min
18. late_aircraft_delay - late aircraft delay, min

Additional Explorations






Counts of Delays caused by various factors?
Airport / Airport Name
EWR
Newark, NJ: Newark Liberty International
JFK
New York, NY: John F. Kennedy International
LGA
New York, NY: LaGuardia
Value
70K
60K
50K
40K
30K
20K
10K
0K
Security Ct
Weather Ct
Late Aircraft Ct
Carrier Ct
Measure Names
Security Ct
Weather Ct
Late Aircraft Ct
Carrier Ct