Executive Summary:

We conduct a thorough investigation of flight delays at LaGuardia Airport in New York, NY, in this study. Our goal is to identify the main causes of delays and offer helpful suggestions to the airport's management so they can improve operational efficiency. We address important concerns about airline performance, types of delays, and seasonal fluctuations by looking at a year's worth of data from the Bureau of Transportation Statistics. The dataset is a monthly data that consists of 15 variables, which are taken from January 2022 to December 2022. The dataset is taken for the Airport LaGuardia, New York, and the description of the dataset is given below.

Variable	***Description***
year	Year of the Data Record
month	Month of the Data Record
carrier	Carrier Code Representing the Airline
carrier_name	Name of the Airline Carrier
airport	Airport Code Representing the Airport
airport_name	Name of the Airport
arr_flights	Total Number of Arrival Flights
arr_del15	Number of Arrival Flights Delayed by 15 Minutes or More.
carrier_ct	Number of Arrival Flights Delayed Due to Carrier-related Issues
weather_ct	Number of Arrival Flights Delayed Due to Weather-related Issues
nas_ct	Number of Arrival Flights Delayed Due to NAS Issues
security_ct	Number of Arrival Flights Delayed Due to Security-related Issues
late_aircraft_ct	Number of Arrival Flights Delayed Due to Late Aircraft Arrival
arr_cancelled	Number of Arrival Flights Canceled
arr_diverted	Number of Arrival Flights Diverted to Another Airport
arr_delay	Total Minutes of Arrival Delay for all Flights
carrier_delay	Total Minutes of Delay Attributed to Carrier-related Issues
weather_delay	Total Minutes of Delay Attributed to Weather-related Issues
nas_delay	Total Minutes of Delay Attributed to NAS Issues
security_delay	Total Minutes of Delay Attributed to Security-related Issues
late_aircraft_delay	Total Minutes of Delay Attributed to Late Aircraft Arrival.

For this purpose, we used some statistical techniques related to this study. First represent the dataset by graphical representation, which consists of the histogram, which tells us about the distribution of the variables. The main characteristics of the dataset are checked by using descriptive statistics. There is some relationship that exists between the variables, so we apply correlation coefficients with their scatter plot to check this relationship. The main objective of this study is that there are a lot of factors that influence on delays flights so the most important statistical technique which is used in this study is the regression model which tells us the impact

of the factors on flight delays. At the end check the most important assumptions of the regression model by using some statistical tests.

Discussion:

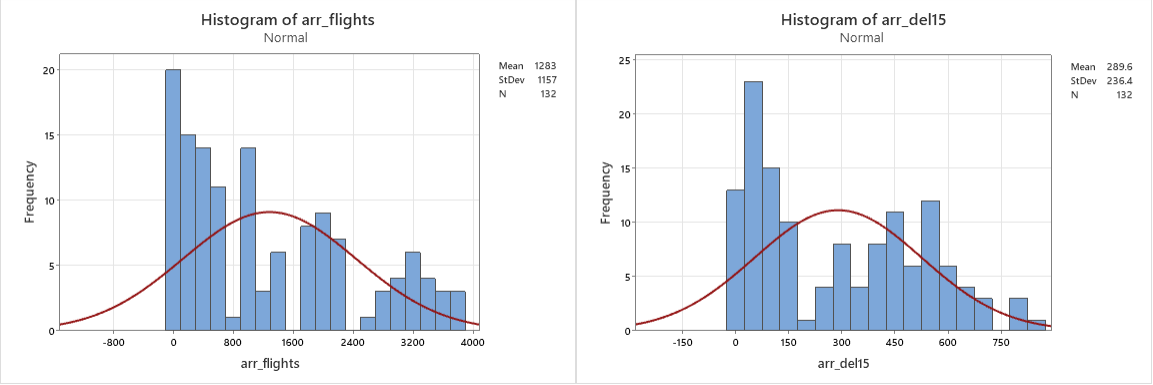
There are 15 variables in this study but we used in which 14 are independent variables, and one dependent variable, but out of 14 there are 11 independent variables, which have a statistically significant effect on the dependent variable “Total Number of Arrival Flights”, so we used total 12 variables in this study, which are given below.

Graphical Representations:

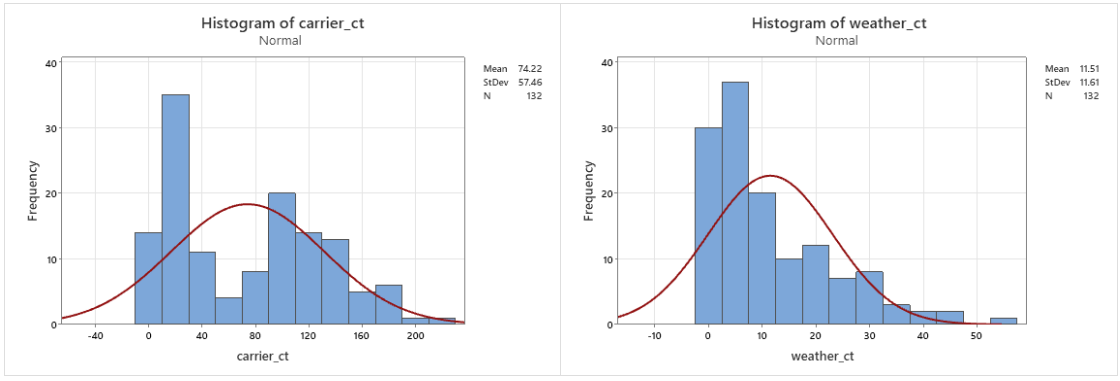
To check the nature of the variables by using graphical representations. These variables are quantitative variables so one of the most familiar statistical graphs is a histogram which is used to check the distribution of the dataset.

1) Histogram:

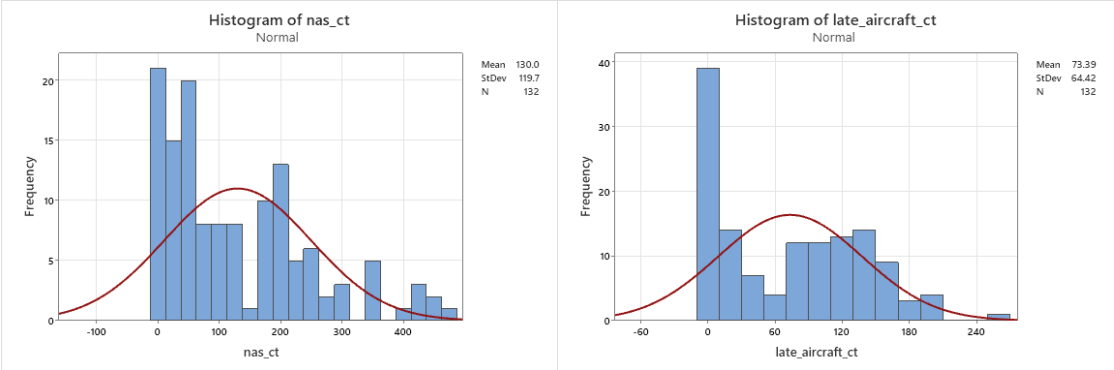
One of the most familiar distributions in statistics is called normal distribution, which is also called the “Mother of Statistics”. If any variable follows normal distribution then it will give a more significant result in Statistics. To check this normality assumption for each variable use the following histogram.



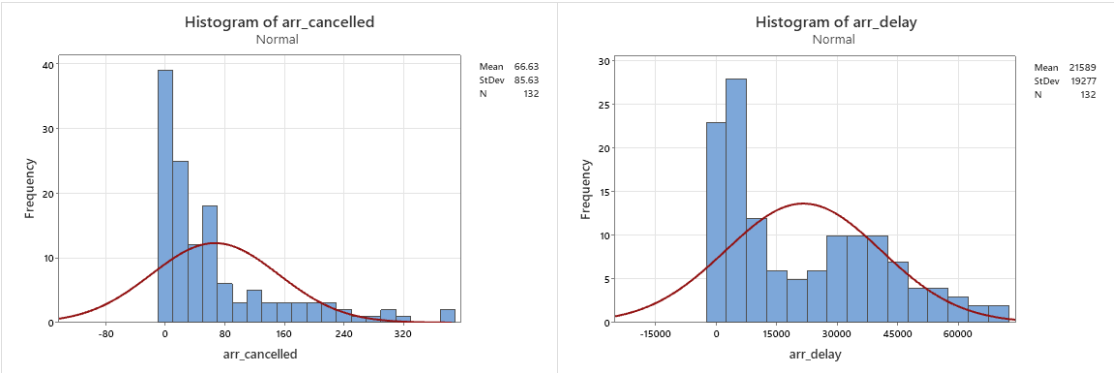
The histogram of “arr_flights” shows that the data is positively skewed which means that the variable “arr_flights” does not follow normal distribution. The histogram of “arr_dell5” indicates that the data is also positively skewed which means that the variable “arr_dell5” also does not follow the normal distribution.



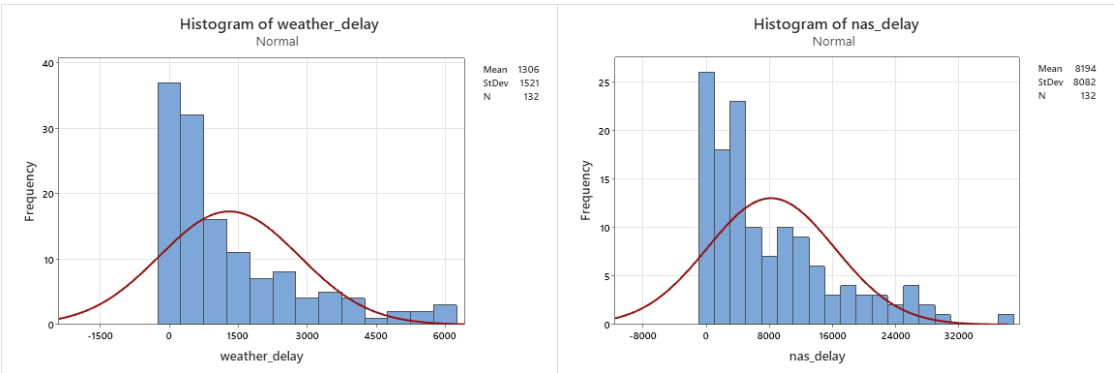
The histogram of “carrier_ct” shows that the data is positively skewed which means that the variable “carrier_ct” does not follow normal distribution. The histogram of “weather_ct” indicates that the data is also positively skewed which means that the variable “weather_ct” also does not follow the normal distribution.



The histogram of “nas_ct” shows that the data is positively skewed which means that the variable “nas_ct” does not follow normal distribution. The histogram of “late_aircraft_ct” indicates that the data is also positively skewed which means that the variable “late_aircraft_ct” also does not follow the normal distribution.

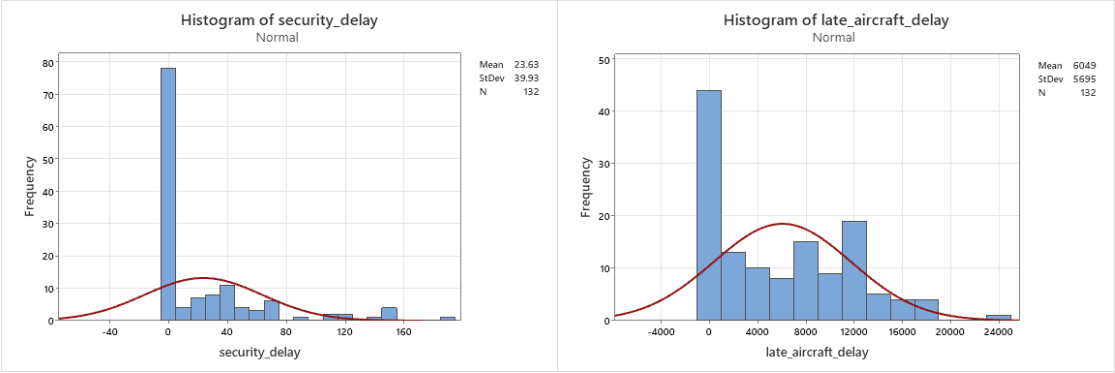


The histogram of “arr_cancelled” shows that the data is positively skewed which means that the variable “arr_cancelled” does not follow normal distribution. The histogram of “arr_delay” indicates that the data is also positively skewed which means that the variable “arr_delay” also does not follow the normal distribution.



The histogram of “weather_delay” shows that the data is positively skewed which means that the variable “weather_delay” does not follow normal distribution. The histogram of “nas_delay”

indicates that the data is also positively skewed which means that the variable “ns_delay” also does not follow the normal distribution.



The histogram of “security_delay” shows that the data is positively skewed which means that the variable “security_delay” does not follow the normal distribution. The histogram of “late_aircraft_delay” indicates that the data is also positively skewed which means that the variable “late_aircraft_delay” also does not follow the normal distribution.

2) Descriptive Statistics:

The main characteristics of each variable using the following descriptive statistics by Minitab.

Statistics

Variable	Total Count	Mean	StDev	Minimum	Median	Maximum
arr_flights	132	1283	1157	43	1031	3885
arr_del15	132	289.6	236.4	6.0	272.0	840.0
carrier_ct	132	74.22	57.46	3.00	79.82	217.81
weather_ct	132	11.51	11.61	0.00	7.02	53.69
nas_ct	132	130.0	119.7	0.0	90.0	474.3
late_aircraft_ct	132	73.39	64.42	0.00	78.92	265.79
arr_cancelled	132	66.63	85.63	0.00	32.00	386.00
arr_delay	132	21589	19277	403	16720	71060
weather_delay	132	1306	1521	0	673	6205
nas_delay	132	8194	8082	0	4914	38291
security_delay	132	23.63	39.93	0.00	0.00	193.00
late_aircraft_delay	132	6049	5695	0	4525	23544

The interpretation of each variable is given below.

arr_flights: This variable likely represents the total number of flights that arrived during the given period. The mean (average) number of flights per period is 1283, with a standard deviation of 1157. The minimum number of flights in a period is 43, while the maximum is 3885.

arr_del15: This variable appears to represent the total number of flights that arrived at least 15 minutes late. The mean number of delayed flights per period is 289.6, with a standard deviation of 236.4. The minimum number of delayed flights in a period is 6.0, while the maximum is 840.0.

carrier_ct: This variable could represent the number of flights delayed due to issues with the carrier (e.g., airline-related issues). The mean number of carrier-related delays per period is 74.22, with a standard deviation of 57.46. The minimum is 3.00, and the maximum is 217.81.

weather_ct: This likely represents the number of flights delayed due to weather conditions. The mean number of weather-related delays per period is 11.51, with a standard deviation of 11.61. The minimum is 0.00, and the maximum is 53.69.

nas_ct: This variable may represent delays attributed to the National Airspace System (NAS), such as air traffic control issues. The mean number of NAS-related delays per period is 130.0, with a standard deviation of 119.7. The minimum is 0.0, and the maximum is 474.3.

late_aircraft_ct: This variable likely represents delays caused by late-arriving aircraft. The mean number of delays per period due to late aircraft is 73.39, with a standard deviation of 64.42. The minimum is 0.00, and the maximum is 265.79.

arr_cancelled: This variable represents the total number of flights that were cancelled. The mean number of cancelled flights per period is 66.63, with a standard deviation of 85.63. The minimum is 0.00, and the maximum is 386.00.

arr_delay: This variable represents the total number of minutes of delay for all flights in the period. The mean delay per period is 21589 minutes, with a standard deviation of 19277 minutes. The minimum delay is 403 minutes, and the maximum delay is 71060 minutes.

weather_delay: This likely represents the total number of minutes of delay specifically attributed to weather conditions. The mean weather-related delay per period is 1306 minutes, with a standard deviation of 1521 minutes. The minimum delay is 0 minutes, and the maximum delay is 6205 minutes.

nas_delay: This variable represents the total number of minutes of delay attributed to NAS issues. The mean NAS-related delay per period is 8194 minutes, with a standard deviation of 8082 minutes. The minimum delay is 0 minutes, and the maximum delay is 38291 minutes.

security_delay: This likely represents the total number of minutes of delay attributed to security issues. The mean security-related delay per period is 23.63 minutes, with a standard deviation of 39.93 minutes. The minimum delay is 0.00 minutes, and the maximum delay is 193.00 minutes.

late_aircraft_delay: This variable represents the total number of minutes of delay specifically attributed to late-arriving aircraft. The mean delay per period due to late aircraft is 6049 minutes, with a standard deviation of 5695 minutes. The minimum delay is 0 minutes, and the maximum delay is 23544 minutes.

Correlation Matrix:

To check the relationship between the variables using the following correlation matrix.

Correlations

	arr_flights	arr_del15	carrier_ct	weather_ct	nas_ct	late_aircraft_ct
arr_del15	0.953					
carrier_ct	0.804	0.893				
weather_ct	0.868	0.818	0.619			
nas_ct	0.938	0.952	0.731	0.818		
late_aircraft_ct	0.876	0.953	0.911	0.750	0.832	
arr_cancelled	0.771	0.714	0.694	0.738	0.631	0.693
arr_delay	0.942	0.964	0.836	0.861	0.921	0.923
weather_delay	0.806	0.757	0.561	0.919	0.780	0.660
nas_delay	0.902	0.900	0.671	0.842	0.947	0.790
security_delay	0.239	0.316	0.413	0.078	0.237	0.329
late_aircraft_delay	0.868	0.936	0.863	0.768	0.833	0.974

	arr_cancelled	arr_delay	weather_delay	nas_delay	security_delay
arr_del15					
carrier_ct					
weather_ct					
nas_ct					
late_aircraft_ct					
arr_cancelled					
arr_delay	0.762				
weather_delay	0.738	0.830			
nas_delay	0.669	0.944	0.825		
security_delay	0.179	0.235	0.055	0.179	
late_aircraft_delay	0.685	0.941	0.684	0.815	0.266

The interpretations of the correlation matrix are given below.

arr_flights vs. Others: There is a strong positive correlation between the total number of flights (arr_flights) and variables related to delays (arr_del15, arr_delay), as well as individual delay types (carrier_ct, weather_ct, nas_ct, late_aircraft_ct). This suggests that as the total number of flights increases, the number of delays tends to increase as well.

arr_del15 vs. Others: High positive correlations are observed between the number of flights delayed by 15 minutes or more (arr_del15) and various delay-related variables, including carrier_ct, weather_ct, nas_ct, late_aircraft_ct, arr_delay, weather_delay, nas_delay, and late_aircraft_delay. This indicates that when there are more flights delayed, there tend to be more delays in each category and overall.

arr_cancelled vs. Others: There are positive correlations between the number of cancelled flights (arr_cancelled) and variables related to delays (arr_del15, arr_delay), suggesting that when there are more cancelled flights, there tend to be more delays as well.

arr_delay vs. Others: Strong positive correlations exist between the total delay time (arr_delay) and variables related to delays (arr_del15, carrier_ct, weather_ct, nas_ct, late_aircraft_ct,

weather_delay, nas_delay, late_aircraft_delay), indicating that as the total delay time increases, the number of delays and their duration in each category also increase.

Security Delay vs. Others: There are relatively weak positive correlations between security delays and other variables, indicating that security delays have less influence compared to other types of delays.

Regression Analysis:

To check the impact of the factors on the flight delays using the following multiple linear regression model in which arr_flights is used as a dependent variable, arr_del15, carrier_ct, weather_ct, nas_ct, late_aircraft_ct, arr_cancelled, arr_delay, weather_delay, nas_delay, security_delay, and late_aircraft_delay are used as independent variables. The result of Minitab is given below.

Regression Equation

arr_flights = 4.9 - 122.0 arr_del15 + 120.9 carrier_ct + 150.2 weather_ct + 129.4 nas_ct
+ 122.5 late_aircraft_ct + 2.409 arr_cancelled + 0.0603 arr_delay
- 0.1984 weather_delay - 0.0852 nas_delay + 1.83 security_delay
- 0.0717 late_aircraft_delay

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4.9	38.3	0.13	0.899	
arr_del15	-122.0	59.4	-2.05	0.042	426270.35
carrier_ct	120.9	59.4	2.03	0.044	25187.32
weather_ct	150.2	58.9	2.55	0.012	1009.18
nas_ct	129.4	59.6	2.17	0.032	109872.12
late_aircraft_ct	122.5	59.2	2.07	0.041	31406.69
arr_cancelled	2.409	0.471	5.12	0.000	3.51
arr_delay	0.0603	0.0140	4.31	0.000	157.50
weather_delay	-0.1984	0.0479	-4.15	0.000	11.45
nas_delay	-0.0852	0.0196	-4.34	0.000	54.43
security_delay	1.83	1.07	1.70	0.091	3.95
late_aircraft_delay	-0.0717	0.0261	-2.75	0.007	47.75

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
246.233	95.85%	95.47%	94.44%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	11	168184903	15289537	252.18	0.000
arr_del15	1	255810	255810	4.22	0.042
carrier_ct	1	250962	250962	4.14	0.044
weather_ct	1	394342	394342	6.50	0.012

nas_ct	1	285904	285904	4.72	0.032
late_aircraft_ct	1	259899	259899	4.29	0.041
arr_cancelled	1	1589278	1589278	26.21	0.000
arr_delay	1	1124092	1124092	18.54	0.000
weather_delay	1	1041965	1041965	17.19	0.000
nas_delay	1	1141166	1141166	18.82	0.000
security_delay	1	176068	176068	2.90	0.091
late_aircraft_delay	1	457131	457131	7.54	0.007
Error	120	7275674	60631		
Total	131	175460577			

For each additional delayed flight, the number of flights decreases by 122.0 on average. This suggests that delays tend to correlate with fewer. For each unit increase in carrier-related delays, the number of flights increases by 120.9 on average. This implies that carrier-related delays might be associated with more flights arriving. For each unit increase in weather-related delays, the number of flights increases by 150.2 on average. This suggests that weather-related delays may coincide with more flights arriving.

For each unit increase in NAS-related delays, the number of flights increases by 129.4 on average. This implies that NAS-related delays might be associated with more flights arriving. For each unit increase in late aircraft-related delays, the number of flights increases by 122.5 on average. This suggests that delays due to late-arriving aircraft may coincide with more flights arriving. For each additional canceled flight, the number of flights increases by 2.409 on average. This might seem counterintuitive but could suggest that when flights are canceled, there may be compensatory scheduling or rescheduling of additional flights.

For each additional minute of delay, the number of flights increases by 0.0603 on average. This suggests a slight positive correlation between total delay time and the number of flights. For each additional minute of delay due to weather, the number of flights decreases by 0.1984 on average. This implies that delays specifically due to weather might coincide with fewer overall flights. For each additional minute of delay due to NAS issues, the number of flights decreases by 0.0852 on average.

This suggests that delays attributed to NAS issues might coincide with fewer overall flights. For each additional minute of delay due to security reasons, the number of flights increases by 1.83 on average. This might seem unexpected and could suggest potential complexities in the relationship between security delays and flight volume. For each additional minute of delay due to late aircraft, the number of flights decreases by 0.0717 on average. This suggests that delays specifically due to late-arriving aircraft might coincide with fewer overall flights.

The P-values of all independent variables are less than the level of significance of 0.05, which means that all independent variables have a significant effect on the number of arrival flights. The P-value of the F-test is also less than the significance level $\alpha = 0.05$, which means the overall model is statistically significant. The coefficient of determination (R^2) value is 95.85% which means that the variation in the number of total flights is explained by 96% by the variation in all independent variables, which concludes that the overall model is well fitted for the future prediction.

Diagnostic Testing:

There are some most important assumptions of the regression model, which should be fulfilled. To detect these assumptions using the following statistical tests.

1) **No Multicollinearity:**

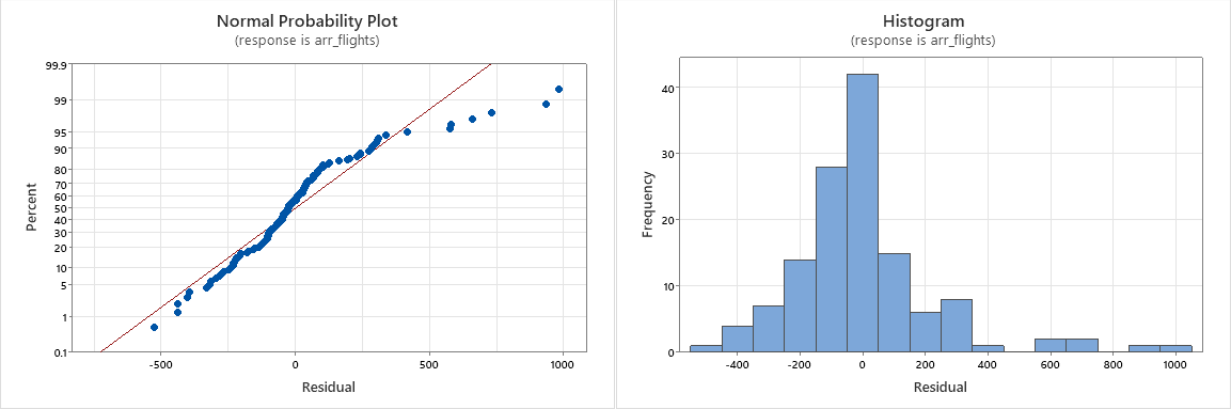
To check the assumption of no-multicollinearity using the variance inflation factor (VIF) test. The last column of the coefficient table shows VIF values.

Conclusion:

There are maximum VIF values are greater than 10 so according to the rule of thumb, there is a multicollinearity problem exists in the model.

2) **Normality:**

To check the assumption of normality use the following normal probability plot and histogram.



The normal probability plot of residuals shows that there are some observations away from the normal line, and the histogram indicates that the data is positively skewed which concludes that the model does not follow the normal distribution.

3) **No Autocorrelation:**

To check this assumption, use the following Durbin-Watson test.

Durbin-Watson Statistic

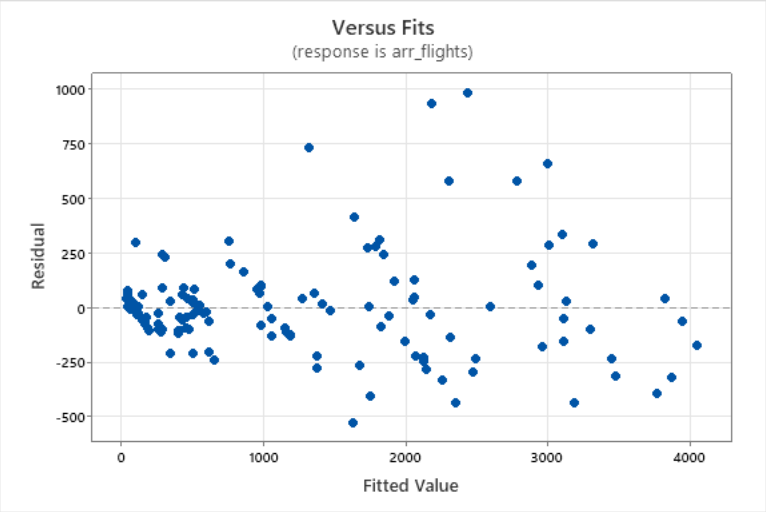
Durbin-Watson Statistic = 1.92434

Conclusion:

The Durbin-Watson value is 1.92, so according to the rule of thumb there is a positive autocorrelation, and we conclude that the model has the problem of autocorrelation.

4) **No Heteroscedasticity:**

To test the no heteroscedasticity assumption using the following scatter plot of residuals.



The scatter plot of residuals shows that there are a lot of observations are deviated from the mean which concludes that the model does not fulfill the assumption of homoscedasticity.

Conclusion:

According to the regression model, the flight days are due to the delayed 15 minutes or more, due to carrier-related issues, weather-related issues, NAS, late aircraft arrival, flight cancellation, and due to attributed to late aircraft arrival. The regression model shows that there is a positive impact of arrival flights delayed due to carrier-related issues, due to weather-related issues, due to NAS issues, due to late aircraft arrival, canceled arrival flights, total minutes of arrival delay for all flights, total minutes of delay attributed to weather-related issues, and total minutes of delay attributed to security-related issues on total delay flights, and there is a negative impact of the number of arrival flights delayed by 15 minutes or more, total minutes of delay attributed to weather-related issues, total minutes of delay attributed to NAS issues, and total minutes of delay attributed to late aircraft arrival on total delay flights.

The limitations of this study are that the regression model does not fulfill the assumptions of normality, no multicollinearity, homoscedasticity, and no autocorrelation, so we will recommend for this future study to remove these problems using alternative methods of fixing. According to the regression model result, we suggest that the management try to fix the issues related to carrier, weather, NAS, late aircraft arrival, security, and the total minutes of delay attributed to these issues because these all factors cause delays in the arrival flights.