(), , 1–15

ARTICLE

Analyzing Flight Delay Times at US Airports through Detection of Outliers and Anomalies

Halilou Junior Tmite

Abstract

As more and more data is collected everyday, dealing with anomalies and outliers has become crucial to perform effective analyses. Previously, poor data collection and/or data processing issues were associated with irregular/missing values. However, with the advent of big data, outliers/anomalies are being increasingly associated with criminal activities, fraud attempts, and targeted attacks. For example, major credit card companies use anomaly detection to detect unusual purchases and possibly identify if a credit card was stolen. Outliers/anomalous observations are usually extreme and irregular values which behave differently from the vast majority of observations. However, one must be careful to recognize that observations can be anomalous in one context, but not another.

In this report, we use anomaly detection and outlier analysis to understand arrival and departure delays of flights at US airports. Our approach begins by exploring the data and using plots/summary statistics of different variables to manually detect outliers/anomalies. Since we are limited with this approach, we aggregate our raw data using summary measures to derive additional variables and reduce the number of observations significantly. Feature selection is then performed to reduce the number of derived variables for analysis. In the end, different distance and density-based anomaly detection algorithms are applied to the reduced dataset and validation is performed by comparing our results across various algorithms.

DATA SET	DESCRIPTION
flights1˙2019˙1	Reporting US Carrier On-Time Performance, January 2019

Keywords: Outlier, Anomaly, Arrival/Departure delay, Feature selection

INTRODUCTION

The dataset is about flights at US airports and mainly contains information about the arrival and departure delay, along with the city of origin and destination, airport of origin and destination and the US state of destination. Information about a flight arriving later than fifteen minutes is also provided. A data dictionary for the dataset is provided along with accompanying visualizations to help the reader better understand the dataset.

Section 1: Exploratory Analysis of the report focuses on cleaning the raw dataset and using exploratory analysis to find outliers/anomalies.

Section 2: Feature Selection is about deriving variables using summary measures (mean, variance, interquartile range, maximum, minimum, etc.) to aggregate the dataset using Origin Airport and Day of Week. The derived variables are further reduced using feature selection by correlation and by variance. This reduction in dimension is necessary to effectively use the anomaly detection algorithm, especially the distance-based ones.

Section 3: Anomaly Detection Algorithms is about using different distance- and density-based anomaly detection algorithms on our reduced derived dataset. Differenty types of distances are used for each applicable algorithm and results are provided in tables.

Section 4: Validation of Results is about comparing our anomaly detection results across the different algorithms. It is expected that distance-based algorithms will be different in anomaly detection compared to density-based algorithms. Anomalous observations are further analyzed using arrival/departure delay to figure out why they were classified as such.

Data Dictionary

VARIABLE NAME	DATA TYPE	DATA FORMAT	DESCRIPTION	RANGE	EXAMPLE
YEAR	Integer	YYYY	The year in which the flight operates.	2019	1905-07-11
DAY_OF_WEEK	Integer	#	The day of week that the flight departs on.	44568	1 - Monday
FL_DATE	Date	YY-MM-DD	The date that the flight departs on.	(2019-01-01) – (2019-01-31)	2019-01-01
ORIGIN_AIRPORT_ID	Integer	#####	Origin Airport, Airport ID. An identification number assigned by US Department of Transportation (DOT) to identify a unique airport.	NA	13495
ORIGIN_AIRPORT_SEQ_ID	Integer	########	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time.	NA	1349505
ORIGIN_CITY_MARKET_ID	Integer	#####	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market.	NA	33495
ORIGIN_CITY_NAME	Text	ABC	The city name of the Origin Airport.	NA	New Orleans, LA
DEST_AIRPORT_ID	Integer	#####	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.	NA	12953
DEST_AIRPORT_SEQ_ID	Integer	########	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time.	NA	1295304
DEST_CITY_MARKET_ID	Integer	#####	Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market.	NA	31703
DEST_CITY_NAME	Text	ABC	The city name of the Destination Airport.	NA	New York, NY
DEST_STATE_ABR	Text	AB	State code of the Destination Airport.	NA	NY
DEP_DELAY	Integer	####	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.	(-47) – (1651)	-10
ARR_TIME	Time	hhmm	Actual arrival time in hours and minutes.	(1) – (2400)	1832
ARR_DELAY	Integer	####	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.	(-85) – (1638)	-25
ARR_DELAY_NEW	Integer	####	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.	(0) – (1638)	0
ARR_DEL15	Integer	0 or 1	Arrival Delay Indicator, 15 Minutes or More (1=Yes).	NA	0

FIGURE 1. Data dictionary for the flights1˙2019˙1 dataset. Missing value codes depend on the variable – they include NA.

1 Explore and Visualize the Dataset

This section goes into detail about exploring the dataset to construct new data based on the variables we derive through the raw dataset. First, we clean the dataset by removing all the missing values. Then, we explore and visualize our raw dataset.

1.1 Cleaning Raw Dataset

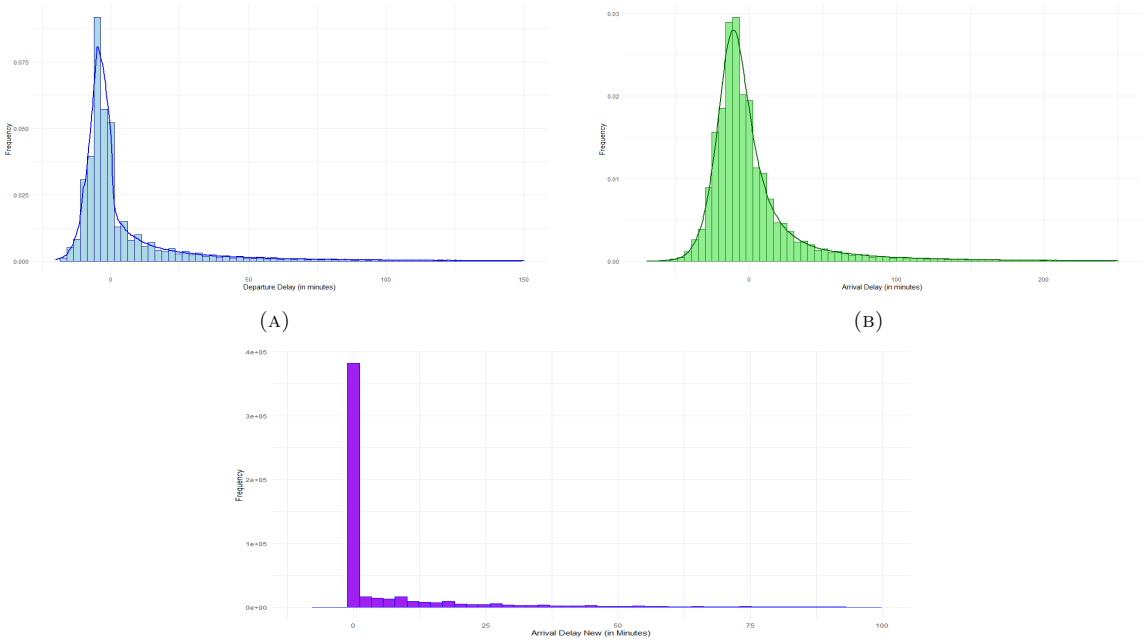
TABLE 1. Number of Observation in Dataset

	$BeforeCleaning$	$AfterCleaning$
$NumberRow$	583985	565963

Data cleansing is the first step in the data analysis process. It is part of the preparation process. The purpose of data cleaning is to identify and fix errors, duplicates, and also missing values. During the preparation of our data, we identified missing values that represent **1.57%** Table (1). We decided to remove all the rows where there is at least one missing value.
After cleaning we end up with **565 963**. observations

1.2 Explore and Visualize

1.2.1 Distribution Plots



(A) (B)

(C)

FIGURE 2. Distribution and Density (A) Departure Delay (B)Arrival Delay (C) Arrival Delay New

Figure 2 represents the distribution plots with density curve overlay of the three main numerical variables in the dataset: (A) Departure delay and (B) Arrival delay. The distribution plots are highly skewed with a "fat" tail. These plots can show that there are some anomalous/outlying observations in the dataset which may be due to a number of reasons: incorrect data entry, misclassification of flights, bad weather/turbulence, airport staff shortage, etc. However, these plots fail to identify which observations are anomalous which is something an analyst may be interested in.

1.2.2 Summary Statistics

	DEP_DELAY	ARR_DELAY	ARR_DELAY_NEW
Mean	9.68	4.26	13.65
Standard Deviation	48.42	51.16	47.49
Kurtosis	172.19	142.54	184.41
Skewness	10.06	8.84	10.52
Minimun	-47	-85	0
First Quartile	-6	-16	0
Median	-3	-7	0
Third Quartile	5	7	7.000
Maximun	1651	1638	1638
Range	(-47,1651)	(-85,1638)	(0,1638)
Interquartile Range	11	23	7
Mode	-5	-9	0
Count	565963	565963	565963

FIGURE 3. Descriptive analysis of the flights1˙2019˙1 raw dataset.

The descriptive analysis allows us to summarize our data and to describe the different variables. We have a large positive skewness which tells us that our data is skewed precisely right tail. We also have a very large kurtosis which tells us that we have a fat tail and justifies our skewness. The maximum and minimum values are extremely large and very far from our mean which leads us to question if it is not an outlier without forgetting also that we have a rather large standard deviation which gives a sense to the dispersion of our values.

1.2.3 Data Visualization

We focus on flights between several major cities. By demonstrating their average arrival (departure) time to see if we can summarize some useful information.
First, in Figure 4(A). The flights are between Los Angeles and New York. It is obvious that the average arrival delay is far smaller than the average departure delay for both directions. They are even negative, meaning that most of the flights between these two cities are earlier than scheduled. By our consideration, that may be because Los Angeles to New York is a busy and well-organized route since many passengers may transfer to other continents. In Figure 4(C), from New York to Atlanta, we have the same situation. The difference is average arrival delay is still positive. Rather than earlier than scheduled, most of them are still late as these routes are considered travel routes. The rest of them are flights between Chicago and Miami in Figure 4(B), and Los Angeles and San Francisco in Figure 4(D). We can see that in most cases, flights were actually faster since the average arrival delays are slightly smaller than average departure delays.

A flight can be delayed for departure for several reasons. These include issues with the maintenance of the plane, passengers missing in the airport, security problems, airport staffing shortages, equipment breakdown, etc. Similarly, a flight can be delayed for arrival for several reasons. These include weather problems where the pilot may have to take a longer route to avoid turbulence,

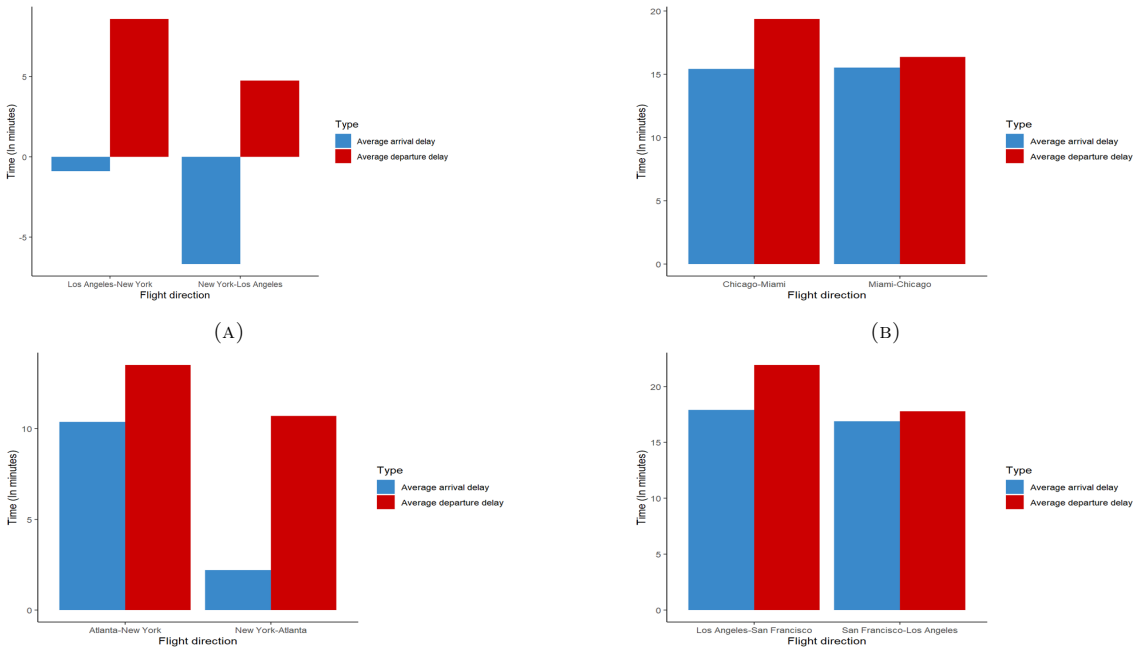


FIGURE 4. Average Arrival and Departure Delay between (A) Los Angeles and New (B) York Chicago and Miami (C) Atlanta and New York (D) LOS Angeles and San Francisco

circling above the airport in a holding pattern due to congestion and headwinds experienced during the flight which reduces the overall speed, causing delays. Comparing average (arrival departure) delay longer than 15 mins, their proportion in each weekday. We can see that in Figure (5), all proportions of the average departure delay are slightly lower than the average arrival delay. Meaning that for most of the flights, if their departure delay is more than 15 mins, there is a high probability for their arrival delay to be longer than 15 mins as well. But for those departure delays are less than 15 mins, there is only a little percentage of them, whose arrival delays are longer than 15 mins.

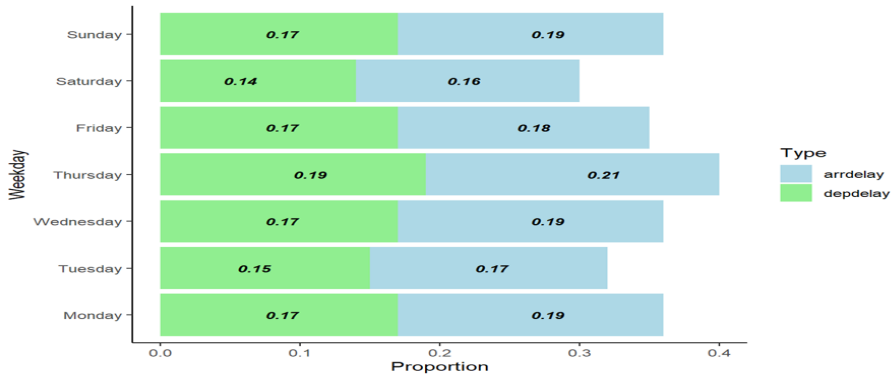


FIGURE 5. Proportion of Arrival and Departure Delay longer than 15 minutes in each weekday.

To analyze flight delays between holiday travel and normal weekday travel, we focused on days at the beginning of January. Figure (6) shows the average arrival and average departure delays for the first four days of the New Year. This visualization was created to analyze the impact of holiday travel on flight schedules. As is common knowledge, January 1st is a holiday in the U.S. and most people travel between January 1st and January 2nd so that they can get back to work during the weekday. From the visualization, we see that the average arrival and departure delays are higher on January 1st and January 2nd. This can be due to increased passenger traffic, the airport being short-staffed, a shortage of pilots, etc. because of the winter holidays. Compared to the first two days of January, the average arrival and departure delays for January 3rd and January 4th are much lower. This may be because most people who need to travel to get to work have done that by January 2nd and the airport staffing situation should also be back to normal. Less passenger traffic and adequate airport staffing result in shorter delays.

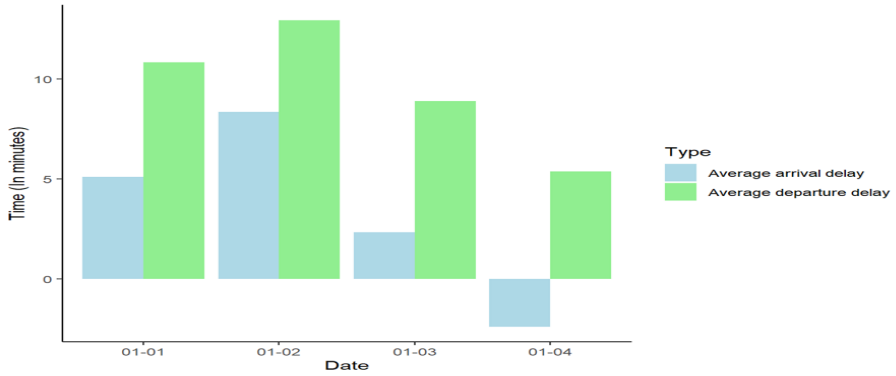
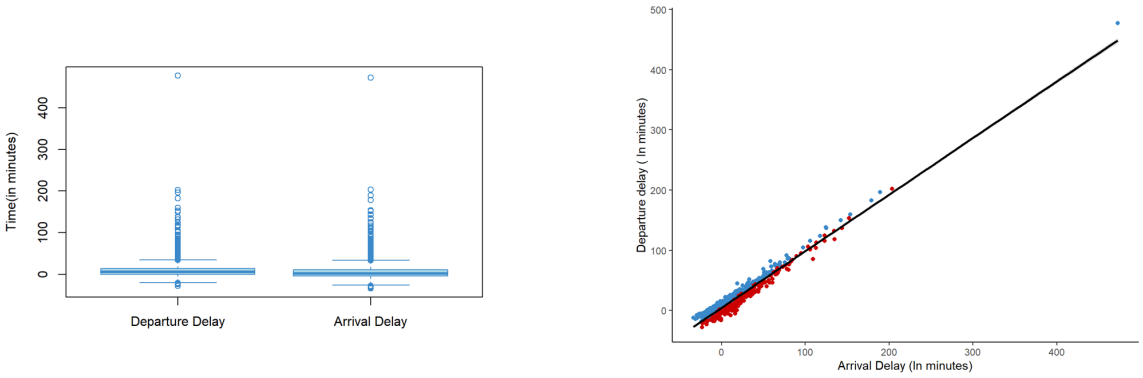


FIGURE 6. Average (departure and arrival) delay from Jan 01- Jan 04.

We keep focusing on departure delays and arrival delays. However, the original dataset is too large for analysis. Then we aggregate the day of the week and airports' ID by means to shrink the size of the data. Then we have Figure 7(A). A boxplot compares the mean of (Departure Arrival) delay in the aggregated dataset. It is clear that both of them have an extreme outlier at the top of

the plot. The rest parts are almost the same for them. Like they both have a lot of outliers above the intervals and only a little below them. After that, we use another way to demonstrate this in Figure 7(B). By drawing a scatterplot, it is also obvious that there's an extreme outlier right in the right top corner as well. The majority of points are from -20 to 80. Fitting those points into a model, we can get a fit line that clears in the middle with black color. The points in red below the line look a little bit more than the points in blue which also confirm what we got. The departure delay is smaller than the arrival delay on average.



(A) (B)

FIGURE 7. Arrival Delay and Departure Day (A) Boxplot (b) Scatterplot

2 Feature Selection

For the rest of our report, we decided to reduce the size of our data to be able to use the various anomaly detection algorithms effectively. We decide to aggregate our data by DAY˙OF˙WEEK and ORIGIN˙AIRPORT˙ID and have derived new summary statistics based on DEP˙DELAY and ARR˙DELAY. For each numerical variable(Arrival and departure delay, we calculate the mean, the interquartile range, the variance, the minimum, the maximum, and the median respectively. Our raw data had **565963** observations and **17** variables and after aggregation we ended up with **2375** observations and **14** derived variables. So we have reduced our data by more than **90%**. We should then select among the newly derived variables which are ideal for the implementation of the different anomaly detection algorithms. We decided to use two methods that will allow us to choose the final derived variables: **Variance** and **correlation** feature selection methods apply to our derived variables.

2.0.1 Feature Selection using Variance

VARIABLE	VARIANCE
ARR_DELAY_VAR	207583445.9
DEP_DELAY_VAR	202709438.2
ARR_DELAY_MAX	100229.81
DEP_DELAY_MAX	100108.346
DEP_DELAY_IQR	523.6383254
ARR_DELAY_IQR	519.5289103
ARR_DELAY_MEAN	486.0155076
DEP_DELAY_MEAN	451.7130654
ARR_DELAY_MEDIAN_var	197.6154918
ARR_DELAY_MIN	157.625161
DEP_DELAY_MEDIAN_var	154.7114803
DEP_DELAY_MIN	30.57759056

FIGURE 8. Derived Variables Variance.

Variance measures variability from the mean or average. This method will allow us to know which variables have different values from each other or are significantly different from each other. You can find in figure (8) the ranking variance of our derived variables. Based on the variance method we can conclude that the variance, the maximum, and the Interquartile range-derived variables will be optimal for the anomaly detection algorithms.

2.0.2 Feature Selection using Correlation

Variance measures variability from the mean or average. This method will allow us to know which variables have different values from each other or are significantly different from each other. You can find in the figure the ranking variance of our derived variable. Based on the variance method we can conclude the

	DEP_DELAY_MEAN	DEP_DELAY_IQR	DEP_DELAY_VAR	DEP_DELAY_MIN	DEP_DELAY_MAX	DEP_DELAY_MEDIAN	ARR_DELAY_MEAN	ARR_DELAY_IQR	ARR_DELAY_VAR	ARR_DELAY_MIN	ARR_DELAY_MAX	ARR_DELAY_MEDIAN
DEP_DELAY_MEAN	1	0.77	0.831	0.184	0.399	0.629	0.874	0.779	0.831	0.1	0.396	0.638
DEP_DELAY_IQR	0.77	1	0.523	0.098	0.093	0.622	0.774	0.837	0.522	0.19	0.093	0.664
DEP_DELAY_VAR	0.831	0.523	1	0.074	0.372	0.546	0.858	0.538	0.995	0.073	0.37	0.519
DEP_DELAY_MIN	0.184	0.098	0.074	1	0.192	0.286	0.169	0.098	0.075	0.429	0.193	0.254
DEP_DELAY_MAX	0.399	0.093	0.372	0.192	1	0.112	0.362	0.155	0.373	0.466	0.999	0.097
DEP_DELAY_MEDIAN	0.629	0.622	0.546	0.286	0.112	1	0.611	0.595	0.546	0.142	0.111	0.992
ARR_DELAY_MEAN	0.874	0.774	0.858	0.169	0.362	0.611	1	0.787	0.858	0.193	0.364	0.697
ARR_DELAY_IQR	0.779	0.837	0.538	0.098	0.155	0.595	0.787	1	0.54	0.085	0.155	0.634
ARR_DELAY_VAR	0.831	0.522	0.995	0.075	0.373	0.546	0.858	0.54	1	0.067	0.373	0.517
ARR_DELAY_MIN	0.1	0.19	0.073	0.429	0.466	0.142	0.193	0.085	0.067	1	0.464	0.257

FIGURE 9. Derived Variable Correlation Matrix.

The correlation method allows us to analyze which of our derived variables are correlated. The more the variables are correlated the more they will not be adequate for our algorithm. Figure (9) represents our correlation matrix. The light green color corresponds to the variables that have a correlation between 0.5 and 0.75 and the green color is for the variables that have a correlation

higher than 0.75. We can conclude that the mean of the arrival and departure delay are highly correlated then cannot be selected in our anomaly detection algorithms

2.0.3 Final Selection of variable

The variance and correlation methods allow us to deduce which variable will be adequate for our different algorithms. We finally decide to select the variance, interquartile range for the arrival and departure delay

3 Anomaly Detection Algorithms

We apply our anomaly detection algorithms to variables we just select.

3.0.1 Distance To All Points (DTAP)

We first start by building a DTAP anomaly detector for several types of distances on scaled artificial data. This algorithm is distance-based. It calculates the distances from one observation to all the others using different methods such as Euclidean, Chebychev, Manhattan, and Minkowski. The top v = 6 anomalous observations for each method are shown below in Figure(10).

	Euclidean	Chebychev	Manhattan	Minkowski ($p = 4$)
Obs	380	380	380	380
	2118	2118	2118	2118
	2205	2205	2205	2205
	1795	2000	1795	1795
	2000	206	2239	2239
	206	1795	2000	2000

FIGURE 10. Outliers Detection using DTAP Different Methods.

The following plot Figure(11) is for the Euclidean distances to all points. We can see that clearly that there is an outlier in 380.

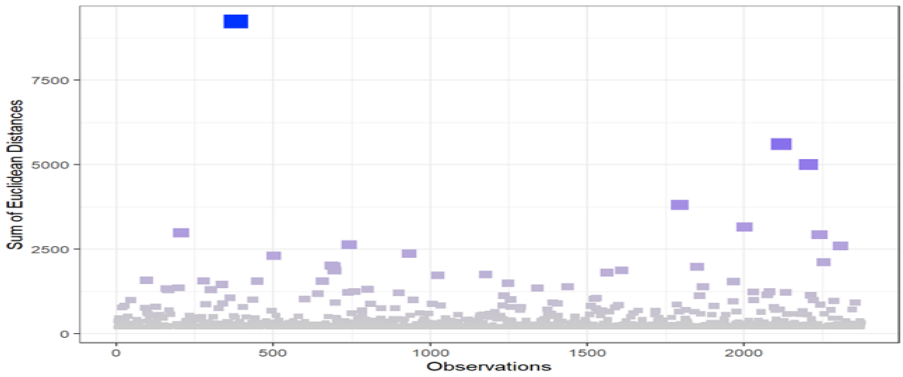


FIGURE 11. Euclidean Distances to all Points Plot.

3.0.2 Distance to Nearest Neighbour(DNN)

Then we build a DNN anomaly detector for several types of distances on the scaled artificial data as well. This algorithm is also distance-based. It calculates the distances from one observation to its nearest neighbor using different methods as we used above. We surprisingly find out their first six anomalous observations are exactly the same as shown below in Figure(12).

	Euclidean	Chebychev	Manhattan	Minkowski ($p = 4$)
Obs			380 2118 2205 1795 2000 206	

FIGURE 12. Outliers Detection using DNN Different Methods.

The following plot Figure(13) is for the Euclidean distances to nearest neighbors. We can see that clearly that there is an outlier in 380 as well.

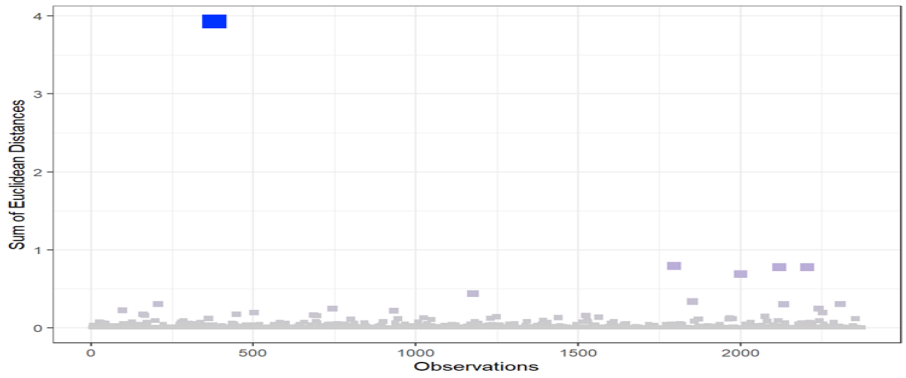
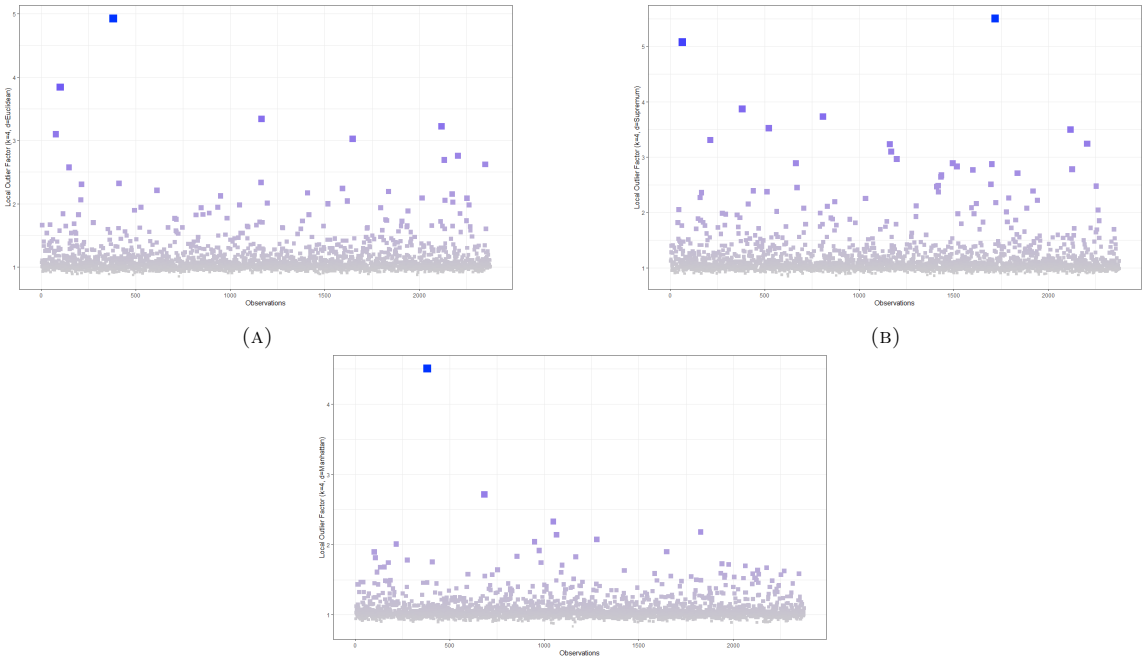


FIGURE 13. Euclidean Distance to Nearest Neighbour Plot.

3.0.3 Local Outlier Factor

This algorithm is a density-based approach for identifying local outliers. It compares the local density of a point to that of its k neighbors using a distance metric. If the density of the point is significantly lower than that of its neighbors, the point may be in a low-density region and may potentially be an outlier. Again, we perform the local outlier factor algorithm using three different distance metrics (euclidean, maximum, and manhattan) on the scaled data. For all three distances, we get observation 380 as an anomalous observation and observation 2118 is identified as anomalous for Euclidean and Chebychev distances (Figure 14 and 15). These results are somewhat different from the distance-based approaches. These observations are further described in the validation section.



(A) (B)

(C)

FIGURE 14. Local Outlier Factor Plot (A) Euclidean (B)Chebsyhev (C) Manhattan

	Euclidean	Chebychev	Manhattan
Obs	380	1718	380
	99	62	682
	1165	380	1047
	2118	806	1827
	76	521	1065
	1647	2118	1277

FIGURE 15. Outliers Detection using Local Outlier Factor Different Methods.

3.0.4 **Isolation Forest**

This is also a density-based algorithm that tries to identify outlier based on the assumptions that only a few outliers are present in the data and that these outliers have characteristics different from other "normal" points. It tries to isolate anomalous points. It does this by randomly selecting an attribute and then randomly selecting a split value between that attribute's min and max values. This recursively partitions the points until every point is isolated in its own partition. The observations that appear to be potentially anomalous or outlying are observation 380, 2118, 2205, and 206 (Figure 16). These observations are further described in the validation section.

	Isolation Forest
Obs	380
	2118
	2205
	206
	1795
	2306

FIGURE 16. Isolation Forest Result.

This plot (Figure 17) for the Isolation Forest algorithm clearly shows Observation 380 as an outlier/anomaly.

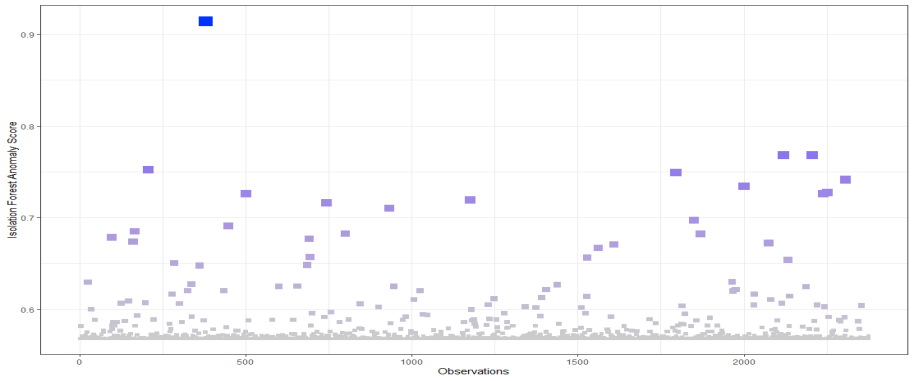


FIGURE 17. Isolation Forest Plot.

3.0.5 **Distance-Based Outlier Basis Using Neighbours (DOBIN)**

The DOBIN algorithm builds a basis which is better suited for the eventual detection of outlying observations. DOBIN's main idea is to search for nearest neighbours that are in fact relatively distant from one another. The observations that appear to be potentially anomalous or outlying are 380, 206, 1795 and 2306 (Figure 18). These observations are similar to results obtained from other algorithms (both distance-based and desnity-based) and are further described in the validation section.

	Knn_dobbin
Obs	380
	206
	1795
	2306
	501
	2205

FIGURE 18. DOBIN Result.

This plot (Figure 19) for the DOBIN algorithm clearly shows Observation 380 and 206 as an outlier/anomaly.

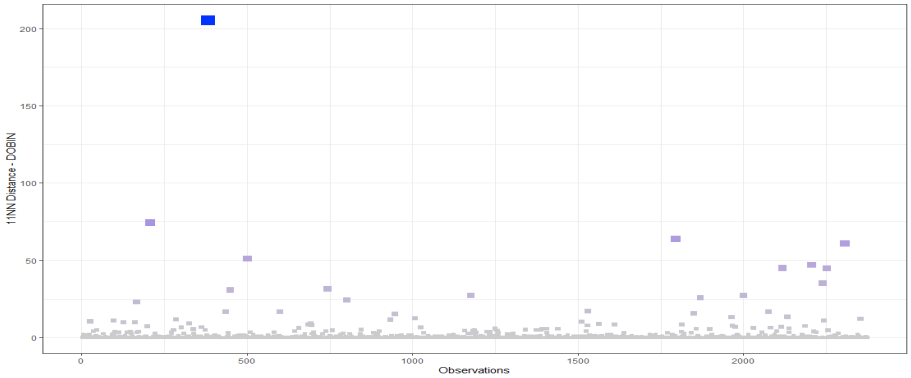


FIGURE 19. DOBIN Plot.

4 Validation of Result

The anomalous observations detected by different algorithms are quite similar, though they differ slightly between the distance-based and density-based approaches. Especially for observation 380 and observation 2118, they appear in every algorithm which clearly confirms they are both obvious outliers. To validate this, we turn back to the raw dataset and find there were only two lines of data for observation 380 which were flights from Belleville, IL to cities in Florida on Jan 01 Tuesday. Among these two, it has an outlier with 937 mins in departure delay and 936 mins in arrival delay. Observation 2118 only has four lines, they were flights from Cheyenne, WY to Dallas, TX on Jan 06 Sunday. Even if two of them were earlier than scheduled, the departure delays and arrival delays for rest two were still large enough to make it an outlier, with (354,480) mins late for departure and (371,463) mins late for arrival respectively. Then we select observations 2205 and 1795 as an example to validate our result as well. For observation 2205, they were flights from Jamestown, ND to Denver, CO, and Devils Lake, ND on Jan 06 Sunday, there are three obvious outliers with (arrival and departure) delays being more than 400 mins. Observation 1795 has four lines with one

anomaly(805 mins late for both arrival and departure). They were flights from Devils Lake, ND to Jamestown, ND on Jan 05 Saturday.

CONCLUSION

In conclusion, this project requires us to use several different methods to detect outliers and anomalies. We first started with exploring the dataset, plotting different variables and summary statistics to identify anomalous observations manually. It became clear to us that the dataset dimension was too high for this approach to work. We then reduced the dimension of the dataset by aggregating on the Day of Week and Origin Airport and reduce the number of observations from roughly 600,000 to roughly 2300 and use summary measures to derive variables. Feature selection was then used to reduce the number of derived variables and the final reduced derived variables were used for the different algorithms for anomaly detection. We then compared results from different algorithms to see if they are similar for any observations which is a way we could validate our results.

WORK COMPLETED BY EACH MEMBER

Victor and Gaurav worked together on the different algorithms. Gaurav completed the data dictionary and some of the exploratory visualizations. Gaurav completed the Abstract, Introduction, and Conclusion, and Victor did some of the visualizations. Halilou helped create the overall format of the report, and feature selection. Victor helped verify our results for different algorithms.

REFERENCES

List of References

General. BTS. (n.d.). Retrieved December 10, 2022, from https://www.transtats.bts.gov/Fields.asp?