Bilge K., David L., Iryna T., Liam D.

Introduction

In data science, statistical analysis often reveals patterns that intuition alone would miss. However, real world datasets like those on U.S. domestic flights are rarely perfect, with issues such as bias, missing values, and errors distorting conclusions. This report focuses on anomaly detection to identify outlier flight instances, routes, and locations that deviate significantly from typical behavior.

Amid rising public concern over inconsistent passenger compensation for delays and cancellations, this analysis evaluates which flight routes and cities would be most affected if the European Union's Regulation (EC) No. 261/2004 were to be applied to U.S. domestic flights. Under EU Regulation (EC) No. 261/2004, passengers are entitled to compensation for flight delays, cancellations, and denied boarding. For delays of 3+ hours, compensation varies by distance: €250 for flights up to 1,500 km, €400 for 1,500–3,500 km, and €600 for flights over 3,500 km (between the EU and non-EU countries). For delays of 2+ hours, airlines must provide meals, communication, and overnight accommodation if needed. In the case of cancellations within 14 days of departure, passengers are entitled to compensation and can choose between a full refund or re-routing. Passengers with missed connections on a single booking and delays of 3+ hours are also eligible for compensation. Reimbursement for hotel stays, meals, and transport is required during significant delays, and compensation applies per passenger, including children with paid seats. By measuring anomalous observed delays against EU compensation thresholds, specific patterns, locations, and operational conditions were identified where such policies would have the greatest impact, helping to inform the feasibility and fairness of adopting similar standards domestically in the United States.

Part I – Preliminary Data Exploration

The data dictionary provides a brief description of the variables of the dataset. Variables are in numerical, float, binary, and date forms. Numeric columns that are not used for indexing are of greatest use for anomaly detection. Since the original dataset lacked sufficient information to effectively identify anomalies on its own, the dataset has been augmented with of additional variables such as precipitation, distance between cities, flight routes, cancellation, and flags for the policy, were crucial. These supplementary features provide broader contextual insight, allowing to capture irregularities that would otherwise go undetected.

Variable	Type	Range	Description	Num NAs	Examples	Data Manipulation
FL_DATE	date	2019-01-01 – 2019-01-31	Date of flight	0	2019-01-21	NA
ORIGIN_AIRPORT_ID	int	10135 – 16218	ID for airport in origin city	0	13495	NA
DEST_AIRPORT_ID	int	10135 – 16218	ID for airport in destination city	0	11194	NA
DEP_DELAY	int	-47 – 1651	How early/late a flight was at departing compared to scheduled departure time	16,355	10	NA
ARR_DELAY	int	-85 – 1638	How early/late a flight was arriving compared to schedule arrival time	18,022	11	NA
city_origin	char	Aberdeen – Yuma	The city a flight is departing from	0	Chicago	Removed everything past the "," in ORIGIN_CITY_NAME in the original datafile
state_id_origin	char	AK – WY	The state corresponding to the city a flight is departing from	0	CA	Manually added state based on city_origin
ORIGIN_LAT	float	-14.28 – 71.29	Latitude coordinate of the flight origin	0	33.57	Left joined origin city latitude from external datafile
ORIGIN_LONG	float	-176.63 – 145.75	Longitude coordinate of the flight origin	0	-112.09	Left joined origin city longitude from external datafile

city_dest	char	Aberdeen – Yuma	The city a flight is arriving into	0	New York	Removed everything past the "," in DEST_CITY_NAME in the original datafile
state_id_dest	char	AK – WY	The state corresponding to the city a flight is arriving into	0	NY	Manually added state based on city_dest
DEST_LAT	float	-14.28 – 71.29	Latitude coordinate of the flight destination	0	42.96	Left joined destination city latitude from external datafile
DEST_LONG	float	-176.63 – 145.75	Longitude coordinate of the flight destination	0	-85.66	Left joined destination city longitude from external datafile
Distance_km	float	70.35 – 7986.36	Distance in kilometers between ORIGIN_LAT/ ORIGIN_LONG and DEST_LAT/ DEST_LONG	0	542.97	Calculation of the geodesic along the earth curve. Not this may not be reflective of ACTUAL flight distance.
Cancelled	binary	0, 1	If a flight was cancelled	0	0	1 if ARR_DELAY is NA, 0 otherwise
Percipitation_mm	float	0 – 107.7	Total liquid-equivalent precipitation (rain + melted snow / freezing rain) at nearest recording station to the origin city latitude/longitude	0	95.0	Left joined from NOAA datafile on weather stations closest to ORIGIN_LAT/ ORIGIN_LONG with precipitation data for all days of the original datafile
FLIGHT_ROUTE_NUM	int	1 – 2778	Unique number assigned for flight routes	0	2747	
FLIGHT_ROUTE_CITIES	char	NA	The two cities a flight goes between	0	atlanta-elmira	Concatenation between origin and destination city, listing the one first that comes first alphabetically
CANCEL_1500	binary	0, 1	Binary flag to bucket cancelled flights into appropriate policy bin	0	1	if cancelled = 1 and distance_km <1500, then = 1, otherwise 0
CANCEL_3500	binary	0, 1	Binary flag to bucket cancelled flights into appropriate policy bin	0	0	if cancelled = 1 and distance_km >= 1500, then 1, else 0
COMP_1500	binary	0, 1	Binary flag to bucket delayed flights into appropriate policy bin	0	1	if ARR_DELAY > 180 & distance_km < 1500, then 1, else 0
COMP_3500	binary	0, 1	Binary flag to bucket delayed flights into appropriate policy bin	0	0	if ARR_DELAY > 180 & distance_km >= 1500, then 1, else 0
ASSIST_1500	binary	0, 1	Binary flag to bucket delayed flight where passengers are entitled to care	0	1	if DEP_DELAY > 120 & distance_km < 1500, then 1, else 0
ASSIST_3500	binary	0, 1	Binary flag to bucket delayed flight where passengers are entitled to care	0	0	if DEP_DELAY > 120 & distance_km >= 1500, then 1, else 0

Table 1. Data dictionary for flights1_2019_1 reduced dimension dataset, with created variables.

In considering compensation policies for missed connections, it is important to evaluate whether arrival delays are distinct from departure delays or if flights can recover lost time on route. However, this is not the case. The Pearson correlation coefficient between the two is 0.96, indicating an almost one-to-one correspondence. This high correlation suggests that flights never make up for departure delays in any significant way, and as such, separate policies for arrival delays may be redundant. Instead, departure delay metrics alone provide a reliable basis for assessing compensation eligibility. As shown in **Figure 2**, there are no outliers of flight instances where arrival delay was reduced with respect to its departure delay.

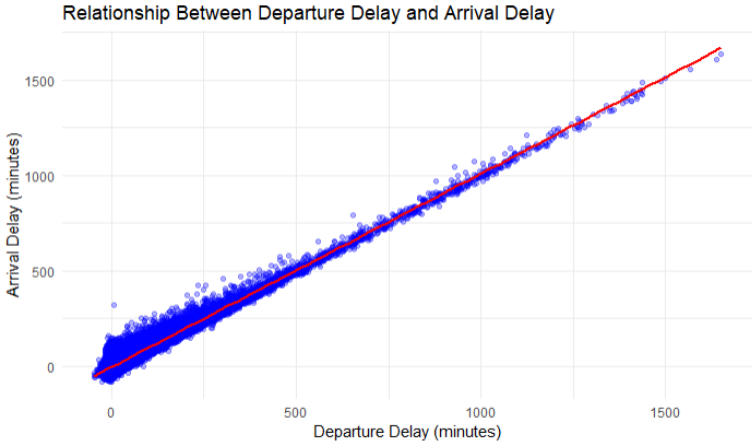

Figure 2. Scatterplot and Line of Best Fit for Arrival Delays vs Departure Delays

Following, ran IQR tests. The IQR test identifies potential outliers by measuring the spread of the middle 50% of the data and flagging any values that fall below the first quartile minus 3 times the IQR or above the third quartile plus 3 times the IQR. These tests are summarized in the table below.

Variable/Column	Q1	Q3	IQR	# of Outliers	% of Dataset
DEP_DELAY	-6	5	11	52,360	8.97%
ARR_DELAY	-16	7	23	26,456	4.46%
Distance_km	575.5574	1,671.429	1,095.872	960	0.16%

Table 2. IQR Results for Departure Delays, Arrival Delays, and Flight Distance.

In the IQR test of each of these variables, as well as Mahalanobis tests ran on the data, the reoccurring issue of too many potential outliers came up. Each gave over tens of thousands, which points out an issue addressed throughout the report: volume. Because of the large size of the dataset, it becomes a hurdle when observations that fit the respective description of the anomaly also comes in large numbers, making it difficult to parse for better potential anomalies, which is further explored throughout the report.

Part II – Outlying Observations and Anomaly Detection Methods

To deepen anomaly investigation, how flight instances align with the European Union's compensation thresholds, focusing on the intersection between departure delays and flight distance, two key factors in determining passenger eligibility for compensation were examined. As shown in the DBSCAN clustering output (**Figure 1**), the most prominent anomalies correspond to departure delays exceeding five hours, which are clearly distinguishable as outliers. To generate this plot, a representative subsample of 20,000 observations to reduce dimensionality and computational time was selected. This subsample maintains the distributional characteristics of the full dataset. For DBSCAN, the epsilon parameter was set to 0.5 and minimum points were set to 10 , values commonly used as starting points in standardized data to balance sensitivity and noise. These parameters provided a clear separation between dense clusters and meaningful outliers without overfitting to local variations.

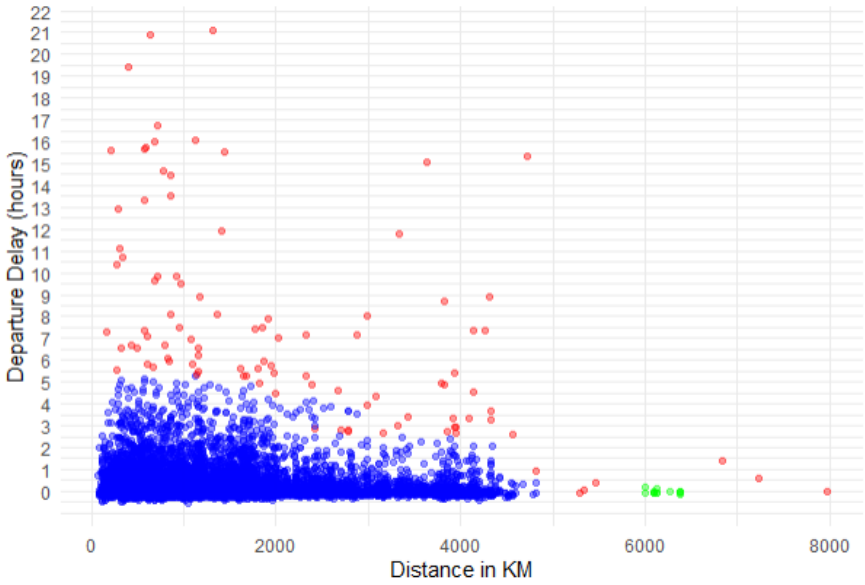

Figure 1. Density-Based Clustering of Departure Delays and Distance in KM

Delay Duration	$\leq 1500\ km$	$1500 - 3500\ km$	$> 3500\ km$	Total (By Delay)
Between 2 and 3 Hours	1890	379	4692	6961
3 Hours and More	1915	385	4930	7230
Total (By Distance)	3805	764	9622	14191

Table 3. Amount of Flight Instances in Each Delay and Distance Threshold Category

Table 3 above summarizes the number of all flights falling into each delay duration and distance category, illustrating how compensation-relevant thresholds intersect in the dataset. While the number of flights meeting these compensation-relevant thresholds is relatively low, only 14,191 out of 583,985 flights, this suggests that the European compensation model is generally manageable for airlines in terms of financial burden. However, U.S. domestic flights show a substantial portion of extreme delays exceeding three and even four hours, which are not currently addressed with standardized compensation policies. Based on these findings, it is recommended that the U.S. adopt structured delay thresholds similar the EU model, particularly by introducing explicit guidelines for delays beyond three and four hours to ensure fairness and accountability.

Keeping the reasoning on flight instances, it revealed flights experiencing departure delays longer than three hours shared several notable characteristics. On average, these flights occurred during 4.5 mm of precipitation in contrast to the mean precipitation of all flights which was 2.56mm, suggesting weather played a significant role. Interestingly, both the most common origin and destination for these delayed flights was Chicago, highlighting it as a critical node in the network of severe delays. These flights also tended to cover moderate distances, averaging around 1,293 kilometers, and ultimately arrived over 5 hours late on average.

To identify disproportionately delayed routes, flight paths with 100 or fewer instances were excluded, as they made up a small portion, 5 flights or less, and could show inflated delay proportions due to just one or two delays. Focusing on routes with over 100 flights and where at least 5% were delayed by more than three hours, certain cities appeared far more frequently. As shown in **Figure 3**, Chicago, San Francisco, and Boston stood out as the most involved in these high-delay routes, suggesting a disproportionate impact on flights connected to these hubs.

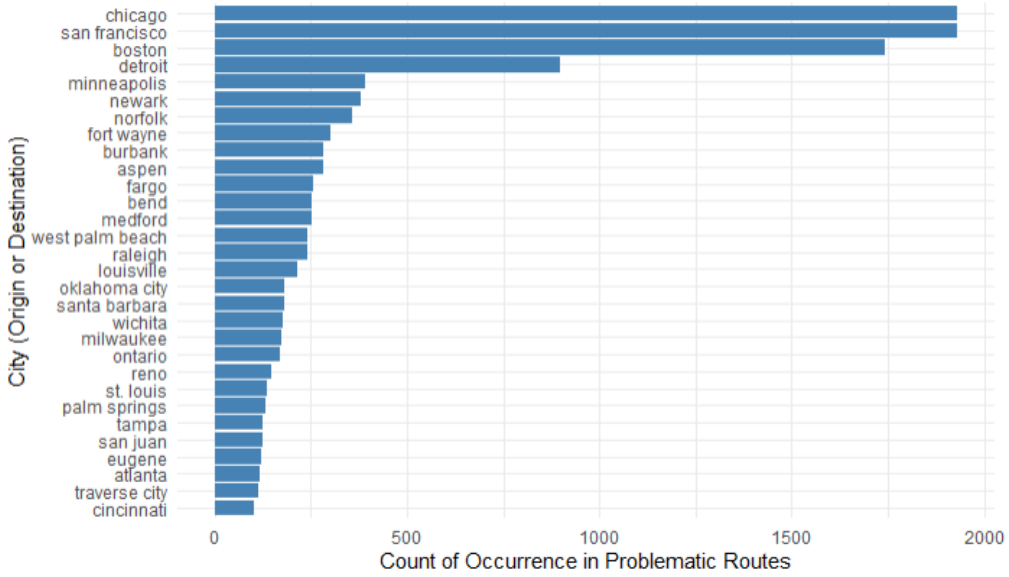

Figure 3. Cities of origin that are most affected by 3+ hour delays.

It was decided to add precipitation data to the dataset to examine whether weather was affecting flights, especially for delays and cancellations. In **Figure 4** it is noticeable that on some days where there is a relative increase in average daily precipitation, there's also an increase in avg delay length and number of cancellations, especially around January 20th-24th. On the contrary, there appears to be a spike of both January 28th onwards but average precipitation is minimal. When looking at correlations between precipitation and other numeric variables in **Figure 5a**, the correlations were negligible for variables such as arrival delay, departure delay, and cancellations, signalling that factors other than precipitation are contributing to delays and cancellations. **Figure 5b** shows the correlation between the average number of flights per day from a given origin city shows strong correlation with variables that measure the assistance and compensation that would be required under the proposed guidelines, meaning that traffic intesity at a given airport is correlated with likelihood of giving out assistance/compensation to passengers.

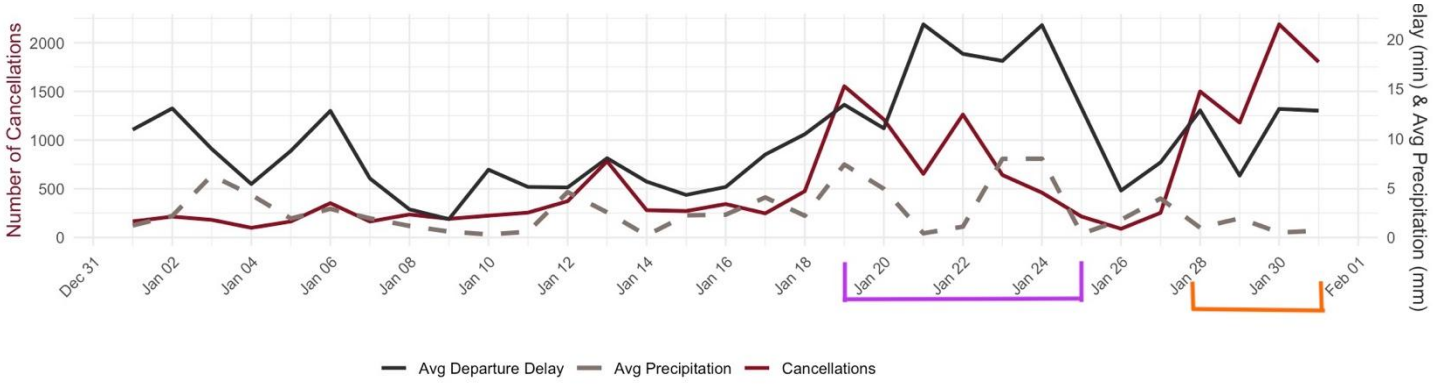

Figure 4. Cancellations, average daily departure delay, and average daily precipitation per date in dataset

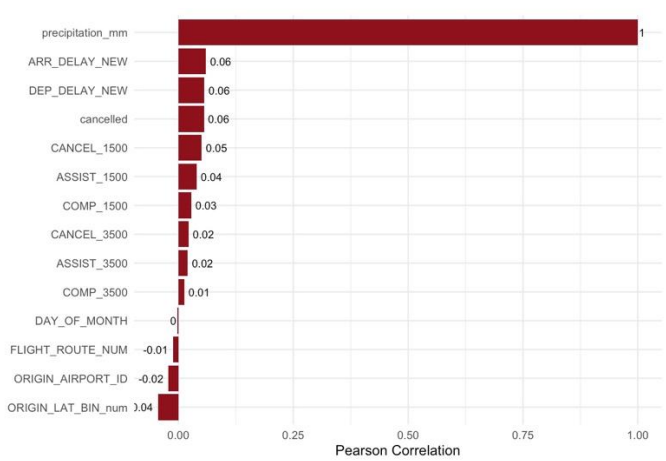

Figure 5a. Correlations between precipitation and other numeric variables.

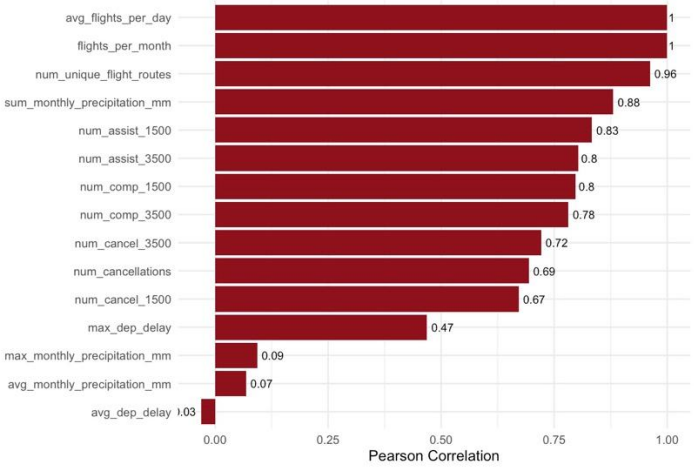

Figure 5b. Correlations between average number of flights per day and other numeric variables.

The mean Mahalanobis distances between each pair in the empirical distribution is displayed in **Figure 6**. This graph is a visualization of how unusual some data points are, when taking into account all of their flight-related characteristics, while also taking into account and adjusting for correlations between the variables. Points with high Mahalanobis distances are outliers. In this figure, the point corresponding to Chicago is noticeably distant from others, as well as the point corresponding to Atlanta.

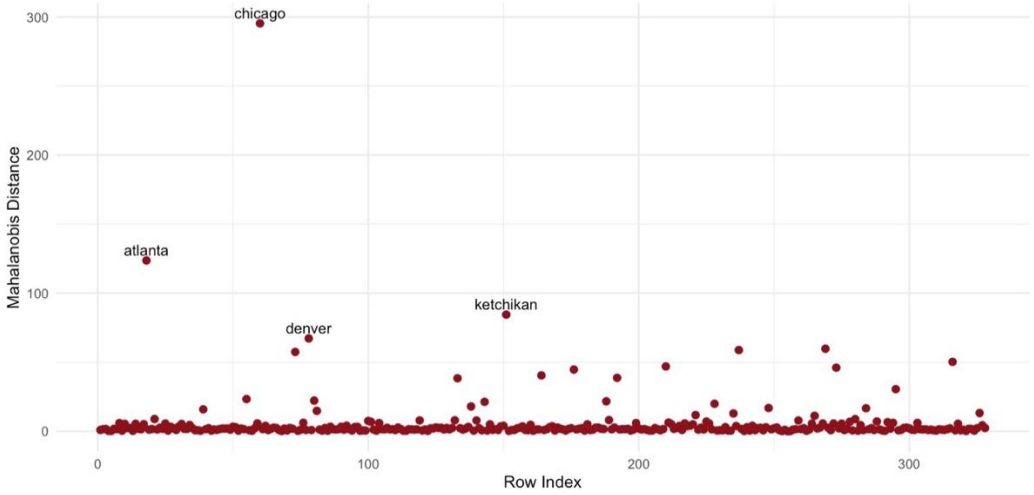

Figure 6. Mahalanobis distance to empirical distribution

The distance to all points (DTAP) anomaly detection algorithm was used to generate **Figures 7a** and **7b**. The empirical distance was measured between origin city and other origin cities, then separately between flight route and other flight routes, using features like flights per month, average flights per day, monthly precipitation mm, average daily precipitation, average departure delay, number of cancellations, etc. To ensure that each variable contributes equally to the distance metric, the data was standardized using z-score normalization, transforming each variable to have a mean of 0 and a standard deviation of 1. This step eliminates differences in units and ranges across the variables. The sum of these empirical distances were divided by the number observations of the main feature, respectively, to get the mean DTAP values. Whether a point is an outlier was determined by looking at what points are past the outer fence of a boxplot, which was defined as quartile 3 plus 3 times the inter-quantile-range. These outlying points are highlighted in garnet. Among city origins in **Figure 7a**, Chicago is once again a drastic outlier, suggesting something about the features unique to Chicago is

highly distinct from other origin cities. New York has the next highest mean DTAP. Atlanta makes an appearance again. Overall, 20 origin cities of 328 were determined to be outliers. Among flight routes in **Figure 7b**, the route New York – Chicago is a drastic outlier, followed by Los Angeles – San Francisco. 156 flight routes of 2778 were identified as outliers.

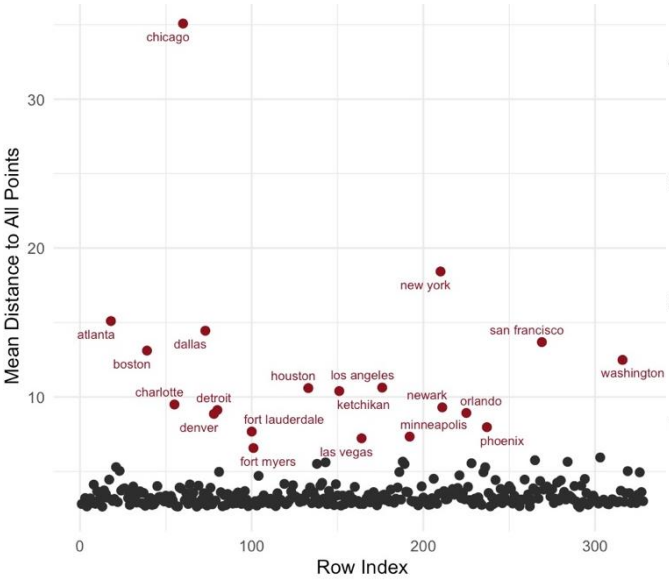

Figure 7a. Mean feature distance to all points, showing origin city outliers

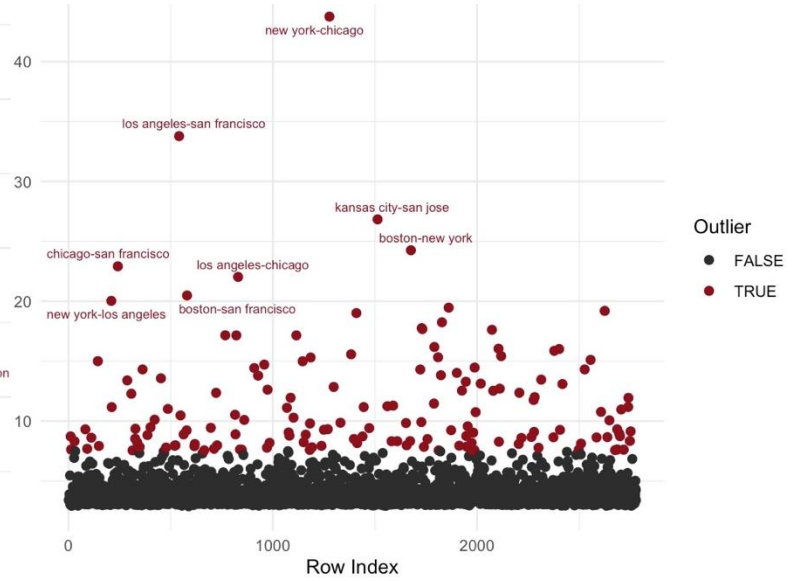

Figure 7b. Mean feature distance to all points, showing flight route outliers

The distance to nearest neighbours anomaly detection algorithm was used to generate **Figures 8a** and **8b**. As above, the distance was measured between origin city and other origin cities, then separately between flight route and other flight routes, using features like flights per month, average flights per day, monthly precipitation mm, average daily precipitation, average departure delay, number of cancellations, etc. These features were scaled to ensure all variables contributed equally, and to remove bias. The Euclidean distance measurement was used to determine similarity between points, and the nearest neighbour determined. To flag anomalies, the outer fence boxplot rule was used once again. Among city origins in **Figure 8a**, Chicago is a drastic outlier, its neighbor distance being substantially higher than other city origin points, suggesting something about the features unique to Chicago is highly distinct from other origin cities. New York was the next most distinct origin city. In total 20 out of 328 origin cities were flagged as anomalous. Among flight routes in **Figure 8b**, New York – Chicago is once again a drastic outlier, followed by Kansas City – San Jose and Los Angeles – San Francisco. With this method, 68 flight routes of 2778 were identified as outliers.

The points indicated as outliers using the DTAP and distance to nearest neighbour indicate that the subset of anomalous origin cities and flight routes are likely to be impacted by the proposed regulations. Identification of these outliers will allow flagged origin cities and flight routes to modify their resource allocation to compensate for delays and cancellations as outlined in the proposed regulations.

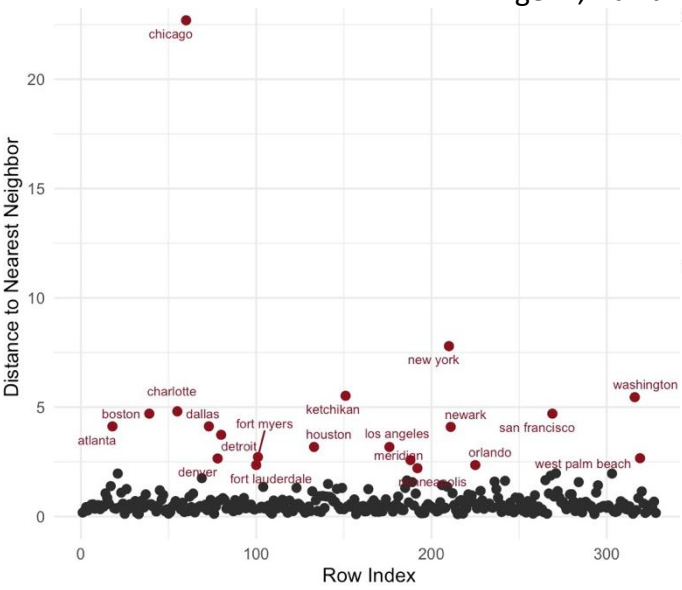

Figure 8a. Nearest neighbor feature distance, showing origin city outliers.

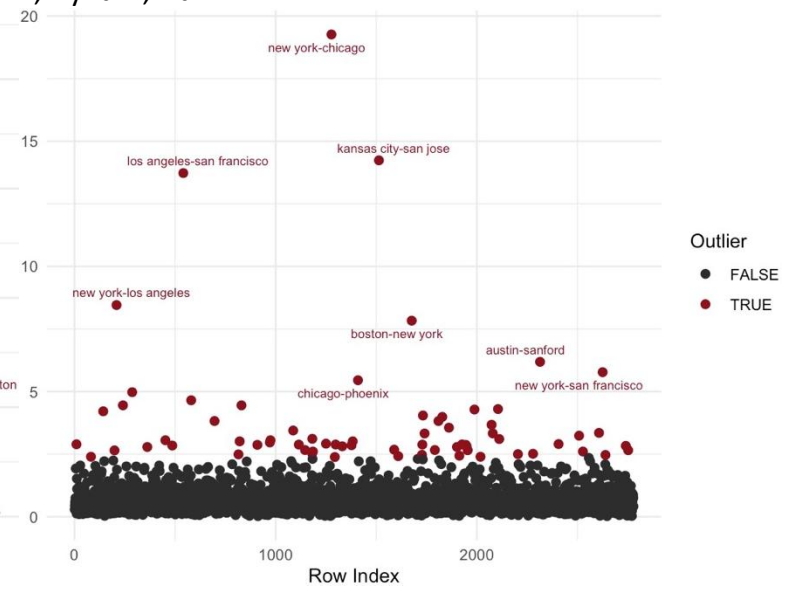

Figure 8b. Nearest neighbor feature distance, showing flight route outliers.

To further explore structural delay symmetry, flights were grouped into route-pairs (e.g., City A → City B and City B → City A). This allowed the evaluation of if delays were consistently directional or symmetric. Routes were defined as bidirectional if they met all the following conditions:

- Absolute difference in average departure delay ≤ 5 minutes
- Absolute difference in average arrival delay ≤ 5 minutes
- Absolute difference in proportion delayed ≤ 0.05
- Difference in flight count ≤ 2

These thresholds reflect distribution-based cutoffs and can serve as a baseline for assessing when route-level compensation metrics may diverge based on direction. **Figure 9** plots the difference in average departure delay against difference in proportion delayed. Most points cluster near the origin, indicating symmetric route performance. However, some extreme cases, such as the Newark, NJ <-> South Bend, IN route, show disproportionately high differences. In such cases, especially for longer routes or those with infrequent service, EU-style compensation schemes could be triggered more frequently in one direction than the other. Identifying such thresholds helps justify compensation criteria based not only on distance and delay but also on asymmetric route behavior, making the policy more robust and equitable.

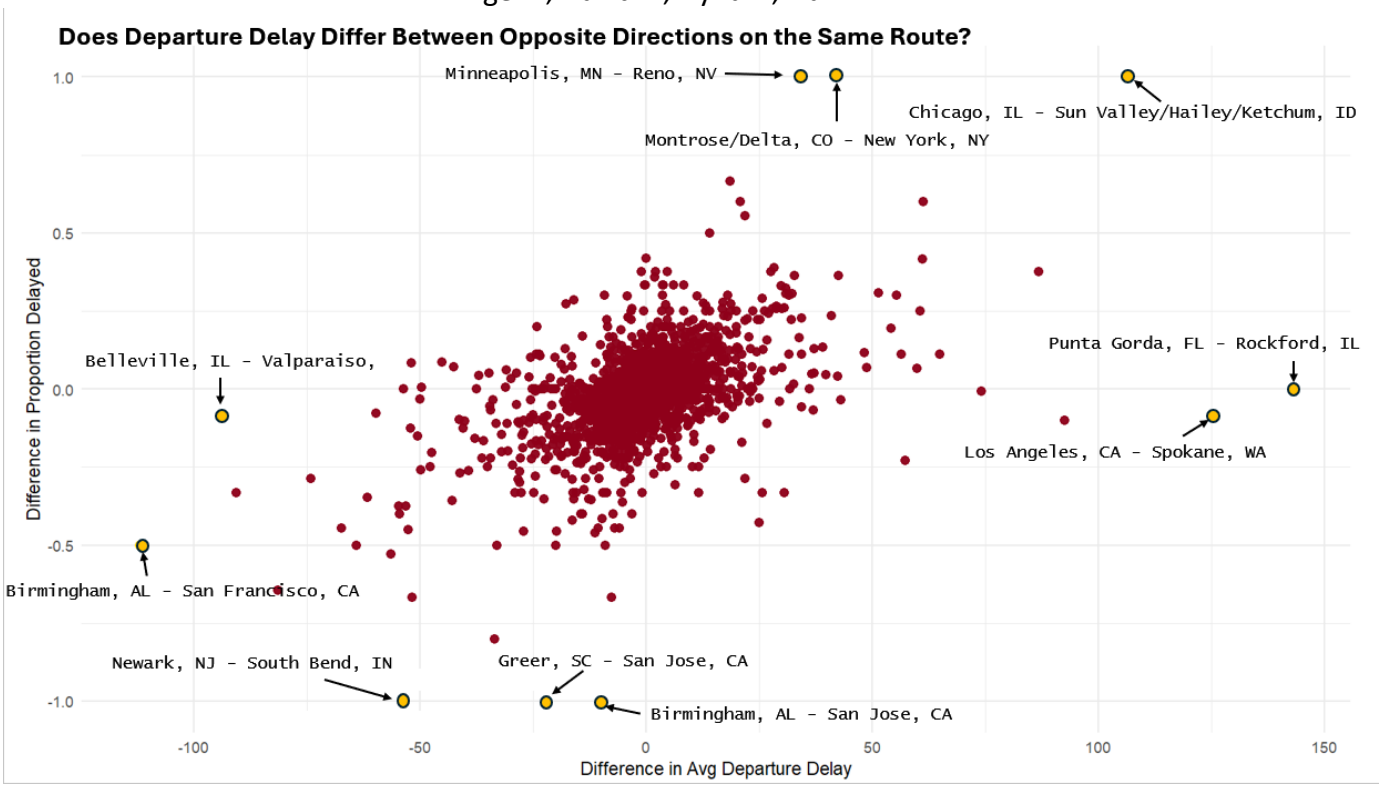

Figure 9. Density-Based Clustering of Arrival Delays and Distance in KM

Conclusion

This analysis focusses on the potential benefits and challenges of implementing an EU- style flight delay compensation policy in the United States. Through the identification of anomalies and outliers in domestic U.S flights from January 2019, and examining the interaction of delay patterns, route characteristics and airport conditions this analysis highlights the impact of Regulation (EC) No. 261/2004. Findings suggest that a small portion of flight, a mere 2.4%, would qualify for compensation under EU regulations. These compensations would be concentrated on routes and cities, particularly those including major hubs like Chicago, San Francisco and Boston.

Benefitting from multiple anomaly detections methods, including IQR analysis, Mahalanobis distance, DBSCAN clustering and distance-based outlier detection, concluded in finding certain cities and routes to exhibit significant high delays and cancellation rates. Chicago recurrently emerged as an outlier, both in terms of delays and unique route behaviour, suggesting the implementation of EU compensation policies would have a large financial and operational impact in small but significant number of areas. This allows airlines to allocate recourses accordingly.

In addition, the high correlation between departure and arrival delays (r=0.96) indicates, passengers rarely regain the lost time in flight. This supports the use of departure delays as a reliable source in assessing compensation eligibility. Analysis into weather, specifically precipitation, presented some delay clusters, but was not a major predictive factor overall in suggesting delays. This suggests operational and logistical factors to play larger rolls in delays.

The analysis of route symmetry reveals that most round-trip routes perform similarly, with a minority exhibiting substantial asymmetry in delay rates and severity. These findings provide further justification for directionally sensitive compensation policies that account for operational mishaps at both origin and destination airports.

In conclusion, applying EU delay compensations in the U.S would not only be reasonable but also necessary in certain areas. The policy would address concerns from passengers while promoting accountability among airlines. Any

implementation should be guided by a vigorous data analysis to ensure the policy is focussed, unbiased and appliable across the U.S.

Appendix

References

Chamberlain, S. (2023). *rnoaa: 'NOAA' Weather Data from R* (R package version 1.3.6) [Computer software]. https://github.com/ropensci/rnoaa

National Centers for Environmental Information. (n.d.). Climate Data Online (CDO). National Oceanic and Atmospheric Administration. https://www.ncdc.noaa.gov/cdo-web/

National Centers for Environmental Information. (2019, January). National Climate Report - January 2019. NOAA. https://www.ncei.noaa.gov/access/monitoring/monthly-report/national/201901

Team-mate Contributions

Teammate	**Contributions**
Bilge Kirilmis	Figure 9 and associated write up, conclusion, editing
David Liu	Data Exploration, DBSCAN and IQR anomaly detection methods at the flight instance level, outlining European Union Compensation standards, writing
Iryna Tkachenko-Riek	Data dictionary, dimension reduction, variable creation relating to precipitation, distance to all points & distance to nearest neighbour anomaly detection (Figures 4, 5a, 5b, 6, 7a, 7b, 8a, 8b, and associated write-up), editing
Liam Dubé	Data cleaning, dimension reduction, geodesic calculation, variable creation relating to EU standards