The impact of meteorological factors of flight delay are analyzed by various linear methods

Zixi Song

2024-03-21

Abstract

This study examines the complex issue of flight delays, utilizing a simulated dataset to analyze how weather conditions, airline efficiency, and airport infrastructure affect flight timeliness. Through exploratory data analysis and linear regression, it assesses the impact of various factors, including wind intensity and terminal operations. Preliminary analysis offers an initial understanding, while subsequent modeling quantifies their influence. Results indicate that although weather significantly affects delays, a combination of factors related to airline and airport operations provides a fuller picture. The study proposes strategies to enhance efficiency and passenger experience and informs policy development aimed at reducing delays. It is important to note, however, that the simulated nature of the dataset and the inherent assumptions of linear modeling introduce limitations to the study's applicability in real-world settings. These findings serve as a basis for future research with more complex models and diverse data, to refine our understanding of flight delays.

Introduction

In today's globalized world, air transport is an important link between countries and regions, and its efficiency and punctuality have a profound impact on economic activity and People's Daily lives. However, as a widespread problem in the air transport system, flight delays are not only an inconvenience to passengers, but also a financial burden to airlines and airport operations[1]. Therefore, in-depth understanding and analysis of various factors affecting flight delay is of great significance for improving flight punctuality rate and optimizing the operation of air transport system[2].

Meteorological conditions such as temperature, rainfall, wind speed and other natural factors are generally considered to be important external factors affecting the on-time rate of flights. In addition, internal factors such as the operational efficiency of airlines, the modernization of airport facilities, and the physical layout of airports also have a non-negligible impact on the punctual departure of flights. Although existing studies have explored these factors from different perspectives, few studies have integrated these factors to comprehensively analyze their combined impact on flight delays using statistical methods.

Based on a flight delay record containing simulated data, this study aims to explore how meteorological factors, airline operational efficiency, airport terminals and other factors affect flight delays through exploratory data analysis, linear regression models, and multiple regression analysis. In particular, this study attempts to explore whether there are significant differences between different airlines and different airport terminals when facing the same meteorological conditions and aircraft capacity, and how these differences affect the on-time departure of flights[3].

Through a comprehensive analysis of these factors, this study aims to provide strategic recommendations for airport management and airlines to reduce flight delays and improve flight punctuality, thereby improving passenger satisfaction and operational efficiency. In addition, the findings of this study will also provide data

The project can be access via https://github.com/KristySzx/FinalPaper

support and reference for air transport management departments when formulating relevant policies and measures.

The remainder of this paper is structured as follows: Section 'Method' details the simple linear regression and multiple linear regression approaches used in our analysis. Section 'Data' describes the dataset employed in this study and presents preliminary findings on the distribution and characteristics of flight delays. This is followed by a more in-depth 'Modeling Process' section, where we construct and analyze two models to identify significant predictors of flight delays. We then perform 'Assumption Checks for Model' to validate our models' robustness and the assumptions underpinning them. The subsequent 'Model Result Discussion' section interprets our findings, discussing their implications for airport and airline operations. Finally, the 'Conclusion' consolidates our findings, acknowledges the limitations of our study, and suggests directions for future research. An 'Appendix' provides additional details on the dataset used in our analysis.

Method

In order to comprehensively evaluate the impact of meteorological factors, airline operations and airport facilities on flight delays, this study mainly adopts simple linear regression and multiple linear regression research methods to explore the impact analysis of meteorological factors, airline operations and airport facilities on flight delays. This paper will introduce our research methods from the aspects of simple linear regression and multiple regression.

Simple linear regression (SLR):

A simple linear regression model was used to explore the effect of maximum wind speed (wind_gust) as a numerical explanatory variable on flight delay time. By analyzing the model coefficient and significance test, the influence intensity and statistical significance of wind speed on flight delay were evaluated.[4]

Simple linear regression is a fundamental and powerful tool in statistics, used for analyzing the linear relationship between two continuous variables. Its core purpose is to predict the value of one variable (the dependent variable or response variable) using the value of another (the independent variable or explanatory variable). The simple linear regression model assumes a straight-line relationship between the two variables, implying that changes in the dependent variable can be explained by changes in the independent variable, and this change is consistent.

The simple linear regression model can be expressed by the Eq.(1)

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

where Y represents the dependent variable. X represents the independent variable. β_0 is the intercept term, indicating the expected value of Y when X=0. β_0 is the slope term, indicating the average change in Y for each one-unit increase in X. ϵ represents the error term, reflecting the impact of factors other than X on Y.

In practice, the true values of β_0 and β_1 are usually unknown, so we need to estimate them from the data. The most common estimation method is the Ordinary Least Squares, which aims to find the estimates of β_0 and β_1 that minimize the sum of the squared vertical distances of all data points from the regression line as the Eq.(2)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

where $\hat{\beta}_1$ is the estimated slope.$\hat{\beta}_0$ is the estimated intercept.$\bar{x}$ and $\bar{y}$ are the sample means of X and Y, respectively and n is the number of observations.

In the research topic of this paper, we choose maximum wind speed as the independent variable and flight delay as the dependent variable to carry out a simple linear regression model. For the specific modeling process and result analysis, please refer to Section Modeling process

Multiple linear regression(MLR)

Multiple linear regression model was constructed, taking flight delay time as response variables, terminal number, airline and whether it rained as categorical variables, and aircraft capacity, air temperature, wind speed and wind direction as numerical explanatory variables. In particular, categorical variables are processed and reference groups are set to analyze their effects on delay time.

Multiple linear regression is a statistical analysis method used to explore the linear relationship between two or more independent variables (explanatory variables) and one dependent variable (response variable). Similar to simple linear regression, multiple linear regression assumes that the dependent variable can be predicted through a linear combination of the independent variables. However, unlike simple linear regression, multiple linear regression considers multiple independent variables, making the analysis more complex and comprehensive [5].

The multiple linear regression model can be represented by the following mathematical equation as the Eq.(3):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon \tag{3}$$

where Y represents the dependent variable.$X_0, x_1, X_2, X_3, ..., X_k$ represent the k independent variables.β_0 is the intercept term, indicating the expected value of Y when all the independent variables are $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are the slope terms, indicating the average impact of each independent variable on Y.ϵ represents the error term.

Multiple linear regression also uses the least squares method to estimate model parameters, aiming to minimize the sum of squared residuals between actual observations and model predictions.The specific process can be seen in formula (2)

Data

Introduction to dataset

The text uses simulation data from airports to represent delay information for flights departing from individual airports. All flights are commercial flights carrying passengers.This dataset contains delay information for departing flights from a fictional airport and represents simulation data for flight delays under different conditions.Moreover This study uses dplyr [6] and knitr [7] packages for EDA work Here's a closer look at this dataset [8]:

- `delay`: the number of minutes after the scheduled departure time that the plane took off (given to 2 decimal places); negative values suggest the plane departed earlier than planned.
- `terminal`: the airport terminal from which the flight departed (either 1, 2 or 3).
- `airline`: the airline operating the flight (either `Air1`, `Air2`, `Air3` or `Air4`).
- `capacity`: the maximum number of passengers the plane can hold.
- `temperature`: the air temperature at the airport at the scheduled time of departure (in degrees Celsius).
- `rain`: whether it was raining at the airport at the scheduled time of departure (yes or no).
- `wind_gust`: the speed of the largest wind gust recorded at the airport that day (in knots).
- `wind_direct`: the direction of the largest wind gust recorded at the airport that day (in degrees, measured relative to north).

Prior to detailed statistical analysis, exploratory data analysis is a crucial step that helps researchers understand the fundamental properties of the data set, uncover potential patterns and outliers, and test

Table 1: Summary Statistics of Flight Delay Dataset

Statistic	Delay	Terminal	Capacity	Temperature	Wind Gust	Wind Direct
Min.	-8.67	1.000	60.0	-7.61	0.010	0.10
1st Qu.	43.83	1.000	200.0	11.57	1.500	43.02
Median	85.25	2.000	275.0	16.05	2.695	85.45
Mean	97.37	2.002	281.3	16.03	2.886	138.61
3rd Qu.	141.87	3.000	350.0	20.67	4.210	240.40
Max.	337.33	3.000	600.0	36.25	7.330	359.90

hypotheses. Using the above flights.csv dataset, this study aims to analyze and understand various factors affecting flight delay.The results are as the Table 1:

Moreover, ggplot2 [9] and reshape2 [10] were used to visualize the data and then perform EDA analysis [11].

Measurement

The data utilized in this study undergoes a detailed measurement process to ensure accurate representation of the flight delay phenomenon. Measurement of each variable is as follows:

1. Delay: This is measured in minutes and represents the length of time between the scheduled and actual departure time. Delays are recorded with two decimal places for precision, where negative values indicate early departures.
2. Terminal: This categorical variable identifies the airport terminal from which the flight departed, labeled as 1, 2, or 3. This metric indicates the operational zone within the airport.
3. Airline: Denoted as Air1, Air2, Air3, or Air4, this categorical variable specifies the airline company operating the flight. It captures the diversity of operational protocols and efficiency.
4. Capacity: Representing the number of passengers a plane can hold, this variable is measured as a whole number and reflects the aircraft's size and potential complexity in boarding and management.
5. Temperature: Recorded in degrees Celsius, the temperature reflects the ambient air conditions at the time of the scheduled departure.
6. Rain: This binary categorical variable indicates the presence (yes) or absence (no) of rain at the scheduled departure time, highlighting weather conditions that could influence flight punctuality.
7. Wind Gust: Measured in knots, this variable captures the intensity of the strongest wind gust recorded at the airport on the day, providing insight into weather-related disruptions.
8. Wind Direction: Represented in degrees relative to north, this variable measures the direction of the strongest wind gust, adding context to how weather patterns may impact flight operations.

To guarantee the fidelity of measurements, the study employs robust data cleaning protocols. This includes the verification of data consistency across different sources, the rectification of any discrepancies found in the data, and the treatment of outliers that could skew the analysis. Additionally, the study conducts a thorough unit consistency check to ensure that all measurements conform to international standards. These meticulous measurement procedures lay the groundwork for reliable statistical modeling and data analysis, forming the basis for the study's subsequent findings and recommendations.

Univariate analysis

In the univariate analysis stage, the distribution characteristics of numerical variables are evaluated in detail by drawing histograms and box plots. The histogram and box plot of the delay variable are drawn first. The histogram reveals the overall distribution pattern of the variable, while the box plot helps to identify potential outliers. The analysis results of numerical variables are as follows:

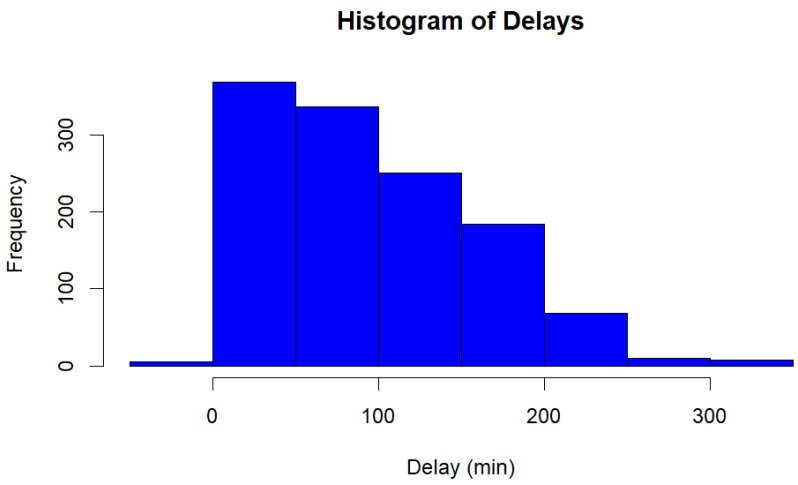

Figure 1: Frequency distribution of delays depicted in a histogram, demonstrating the occurrence of delay intervals within a range of 0 to 300 minutes.

Histogram figure1 analysis shows that flight delays are mainly concentrated in a shorter time frame, showing a clear right-skewed distribution, which indicates that while most flights are not delayed for long periods of time, there are still some flights that experience longer delays, especially the long tail area on the right side of the histogram, indicating that a small number of flights may experience significantly higher than average delays.

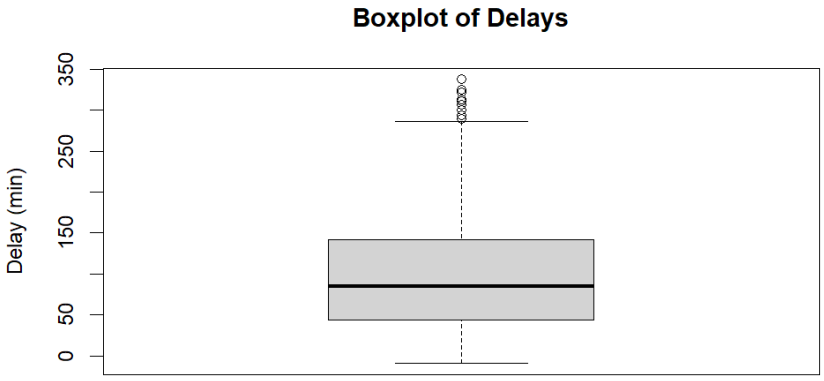

Figure 2: Boxplot illustrating the distribution and variability of delay times.

The analysis results of the boxplot 2 reveal that the median flight delay time is located in a relatively low time interval, while the extension of the upper tail and numerous anomalies distributed outside the high delay time suggest that in addition to regular flight delays, there are many extreme delay events, which may be caused by unusual special circumstances, such as bad weather conditions or mechanical failures.

Combining these two charts, we can conclude that while most flight delays were relatively short, quite a few flights experienced considerable delays. The extreme value of these delays can be caused by specific,

infrequent events, such as bad weather conditions or technical issues. These findings highlight the importance of further analysis of the causes of delays to optimize flight time management.

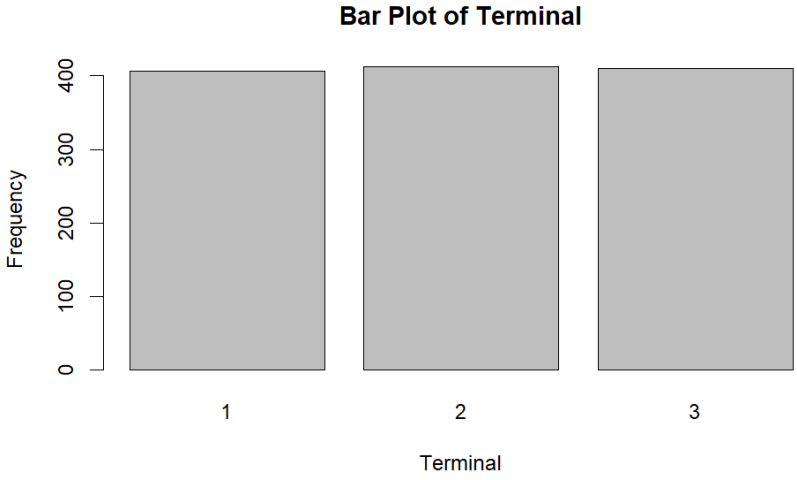

Figure 3: Bar chart representation of frequency distribution across different terminals.

According to the visualization results of Fig3, the number of flights in the three terminals is relatively evenly distributed, which means that from the perspective of data collection, the number of samples in each terminal is balanced. This provides an equal comparative basis for subsequent analysis, allowing us to conduct reasonable comparative analysis without worrying about bias due to unbalanced sample sizes. In the Fig4

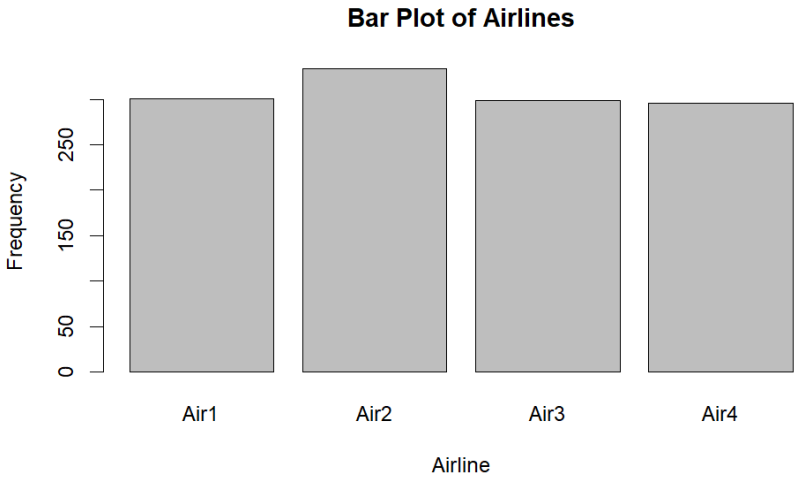

Figure 4: Bar plot displaying the frequency of flights operated by different airlines.

the number of flights across the four airlines also shows a similar uniform distribution. This indicates that the sample size of each airline is basically the same during the data collection process, so that the delay time can be effectively compared between different airlines, without being affected by too much or too little data for one airline.

Bivariate analysis

The bivariate analysis phase focuses on the relationship between variables. In this study, scatter plot is used to explore the relationship between numerical independent variables and flight delay. These diagrams can reveal whether there is an underlying linear or non-linear relationship between the variables. In the Fig5 the

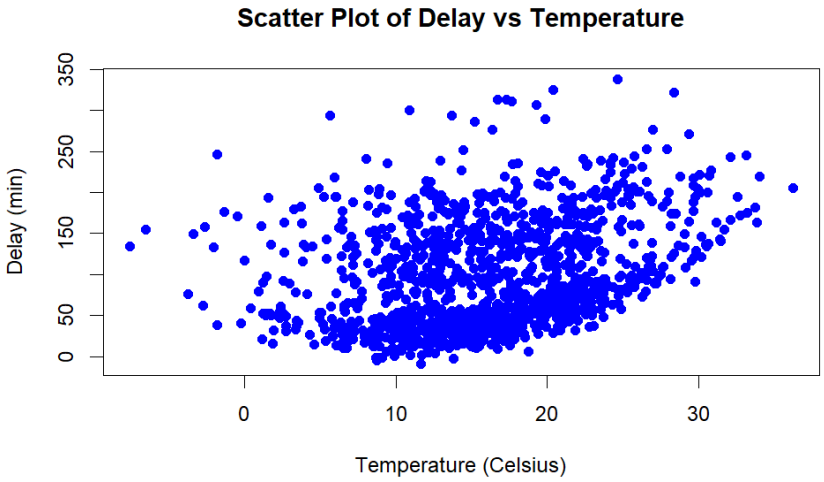

Figure 5: Scatter plot analyzing the relationship between flight delays (in minutes) and ambient temperature (Celsius).

scatter plot shows no clear pattern indicating a direct relationship between temperature and flight delays. Flight delays appear to be distributed randomly at different temperatures, suggesting that temperature may not be the primary factor affecting delays, or that its effect may be masked by other unaccounted for variables. For the Fig 6, the data points also did not show a clear correlation trend. While there was a slight upward

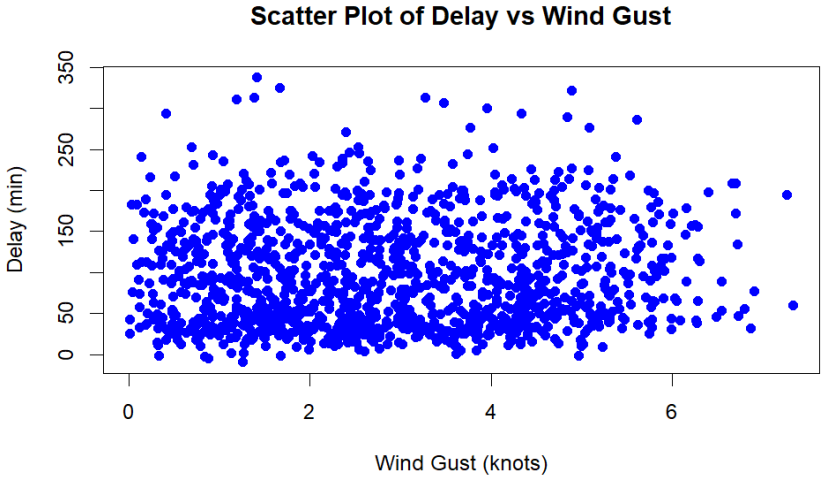

Figure 6: Scatter plot assessing the correlation between flight delays (in minutes) and wind gust intensity (knots).

trend in delay times in some wind speed bands, the relationship was inconsistent and had a large number of overlapping data points. This suggests that wind speed alone may not be enough to explain changes in delay

times, and that the interaction of wind speed with other variables may need to be considered.

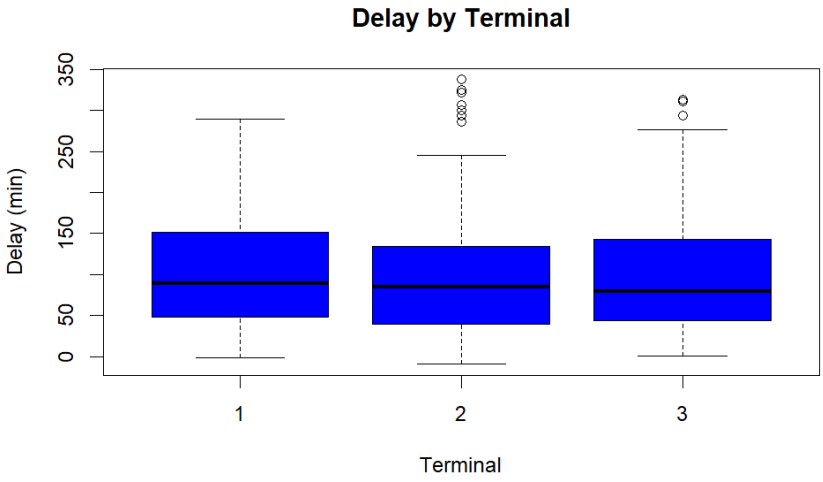

Figure 7: Boxplot analysis of flight delays by terminal, detailing the variability within and across different terminals.

In the box plot of Fig 7, we can observe that the median delay time of the three terminals is similar, but the delay time distribution range of terminal 1 is slightly wider and there are more outliers. This could imply that more flights in Terminal 1 are experiencing extreme delays, or that there are more variables in Terminal 1 operations that affect flight punctuality.

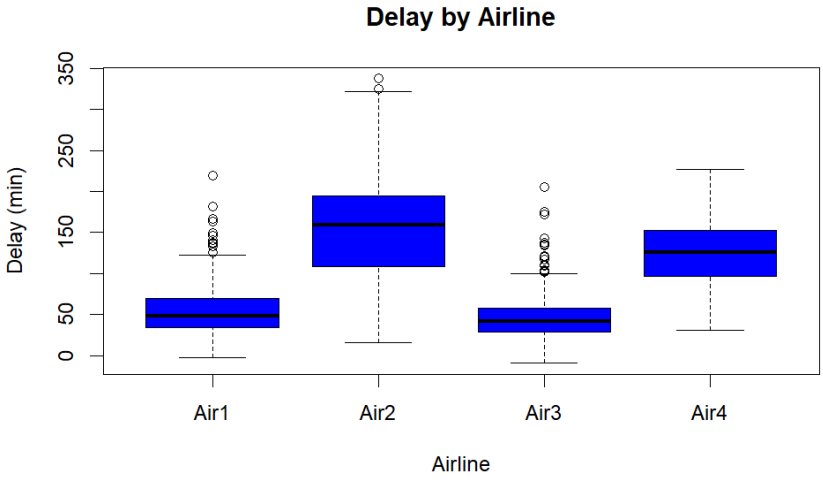

Figure 8: Boxplot analysis of flight delays by airline, highlighting the delay patterns for each carrier.

The box plot of Fig8 shows that there are certain differences in the median delay time of different airlines, among which Air1 and Air3 have lower median delay time, while Air2 has higher median delay time, and Air2 and Air4 show more abnormal delay values. This could indicate that Air1 and Air3 may be more efficient at scheduling flights, or that their flights are less affected by external influences.

Correlation analysis

Finally, the Pearson correlation coefficients between the numerical variables were calculated and visualized using heat maps to assess the strength of the linear relationship between the variables. The correlation analysis helps to identify the key factors that may affect flight delay, and also provides the basis for the variable selection of the subsequent multiple linear regression model.

Correlation heat maps show the coefficients of correlation between different variables. The depth of the color represents the size of the correlation coefficient, red indicates a positive correlation, the darker the color indicates a stronger correlation, and white indicates no significant correlation [12]. From the Fig9, we can see

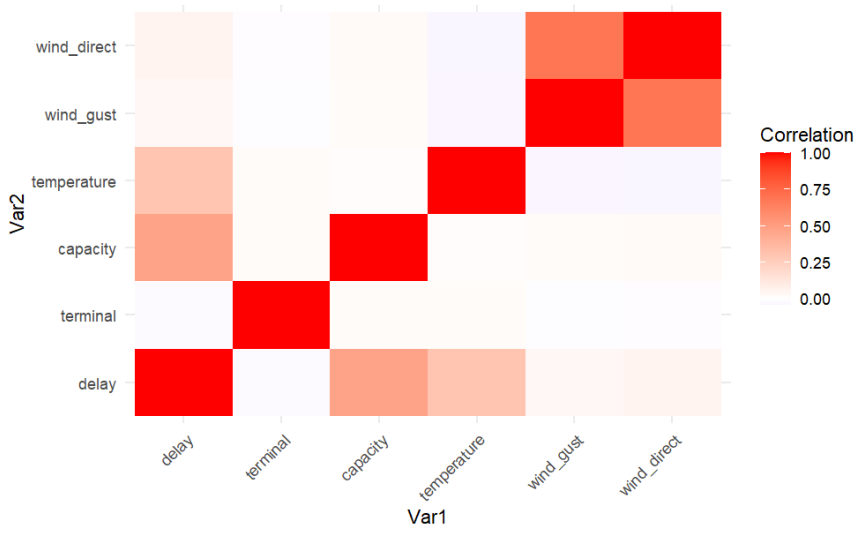

Figure 9: Heat map of the multivariable correlation coefficients, offering insights into the interdependencies among various flight and environmental factors.

that:

The delay shows a strong positive correlation with terminal and aircraft capacity. This may indicate that differences in terminals and the capacity of aircraft are partly related to flight delays. This may be because certain terminals may face greater passenger flow or operational efficiency issues, while larger capacity aircraft may require longer boarding and departure preparation times.

There is also a positive correlation between terminals and aircraft capacity, suggesting that some terminals may specialize in handling larger aircraft, or that the routes they serve may typically use larger aircraft.

The correlation between other variables such as temperature, wind speed and wind_direct and flight delay time is not very significant, indicating that they may not have a direct or strong impact on flight delay, or it is necessary to consider the role of other variables to understand their relationship with flight delay.

With respect to wind_direct, since wind direction is a directional variable and its value does not directly represent intensity or change, it is reasonable that it has a low or insignificant correlation with delays.

These findings guide us to pay special attention to two variables, terminal and aircraft capacity, and how they may interact with other variables when conducting multiple linear regression analyses.

Through the above exploratory data analysis, this study lays a solid foundation for in-depth understanding of the influencing factors of flight delay, and provides important prerequisite information for further multivariate analysis.

Modeling Process

SLR Model construction of Wind_gust and Delay

After completing exploratory data analysis of flights.csv dataset, this study further explores the relationship between wind speed (wind_gust) and flight delay time. According to the theory of the airport management team, wind speed may be a key factor in causing flight delays. To verify this theory, we construct a simple linear regression Model (Model 1), using wind speed as an independent variable (predictor) to predict flight delay time. Using the standard linear regression analysis method, the following model is constructed:

$$Delay = \beta_0 + \beta_1 \times WindGust + \epsilon \tag{4}$$

where Delay is the time of flight delay and Wind_gust is the maximum wind speed.It's the intercept,Is the prediction coefficient of wind speed to delay time, and ϵ is the error term.

The model summary is shown as the Table2:

Table 2: Summary of SLR Model1 Fit

Residuals:	Min	1Q	Median	3Q	Max
	-103.52	-53.49	-13.65	44.59	242.25
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	92.892	3.706	25.066	<2e-16 ***	
wind_gust	1.552	1.119	1.388	0.165	

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.87 on 1228 degrees of freedom
Multiple R-squared: 0.001566, Adjusted R-squared: 0.0007531
F-statistic: 1.926 on 1 and 1228 DF, p-value: 0.1654

SLR model analysis and result

1. Intercept: The model estimates an intercept of 92.892, which is the predicted average delay time when the wind speed is 0 knots. The T-value of the intercept is very large, and its p-value is less than 0.001, indicating that the intercept is statistically significant.
2. Wind_gust coefficient: The coefficient of wind speed is 1.552, meaning that the predicted average delay time increases by approximately 1.552 minutes for each additional knot of wind speed. However, the coefficient of wind speed was not statistically significant (p value 0.165), suggesting that the relationship between wind speed and flight delay time was not statistically strong.
3. R-squared: The R-squared value of the model is 0.001566, which means that the model explains only about 0.16% of the variability of the delay time. This value is very low, indicating that there is little linear relationship between wind speed and flight delay time.
4. Adjusted R-squared: The adjusted R-squared value is 0.0007531, and R-squared adjusted for degrees of freedom shows similar explanatory power.
5. F-statistic: The P-value of F-statistic is 0.1654, indicating that the entire model is not statistically significant, which further supports that wind speed is not a strong predictor of flight delay [13].
6. The simple linear regression model shows that the relationship between maximum wind speed and flight delay time is not statistically significant (p = 0.165), and the R-squared value of the model is only

0.001566, indicating that the explanatory power of wind speed on flight delay time is extremely low. This suggests that other factors besides wind speed played a role in the delay.

MLR model construction of Delay

Following the analysis of Model 1, this study further adopts the general linear Model (Model 2) to comprehensively consider more factors that may affect flight delay. In this model, flight delay time is used as a dependent variable, while terminal, airline and rainfall conditions are included as categorical variables, as well as aircraft capacity, air temperature, wind speed and wind direction as numerical variables. Model 2 aims to reveal how, after controlling for multiple variables, these variables independently and collectively affect flight delay times.

The modeling results are as follows:

$$\begin{aligned} \text{Delay} = \beta_0 &+ \beta_1 \times \text{Terminal} + \beta_2 \times \text{Airline} \\ &+ \beta_3 \times \text{Rain} + \beta_4 \times \text{Capacity} \\ &+ \beta_5 \times \text{Temperature} + \beta_6 \times \text{WindGust} \\ &+ \beta_7 \times \text{WindDirect} + \epsilon \end{aligned} \tag{5}$$

Here, terminals, airlines, and rainfall conditions are treated as factor variables, and each category is associated with a coefficient that measures the impact of that category relative to a reference group.

Table 3: Summary of MLR mode2 Fit

Residuals:	Min	1Q	Median	3Q	Max
	-60.403	-14.105	-2.912	12.842	142.810
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-72.027201	3.257775	-22.109	< 2e-16 ⋆⋆⋆	
terminal2	-10.062875	1.694169	-5.940	3.72e-09 ⋆⋆⋆	
terminal3	-7.226512	1.696220	-4.260	2.20e-05 ⋆⋆⋆	
airlineAir2	103.015518	1.929380	53.393	< 2e-16 ⋆⋆⋆	
airlineAir3	-4.719053	1.984180	-2.378	0.0175 ⋆	
airlineAir4	72.168285	1.987042	36.319	< 2e-16 ⋆⋆⋆	
rainydays	10.867492	1.387132	7.835	1.02e-14 ⋆⋆⋆	
capacity	0.277917	0.006236	44.567	< 2e-16 ⋆⋆⋆	
temperature	2.773309	0.103548	26.783	< 2e-16 ⋆⋆⋆	
wind_gust	0.018183	0.585166	0.031	0.9752	
wind_direct	0.021913	0.008790	2.493	0.0128 ⋆	

Signif. codes:	0 ⋆⋆⋆ 0.001 ⋆⋆ 0.01 ⋆ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error:	24.22 on 1219 degrees of freedom
Multiple R-squared:	0.8574, Adjusted R-squared: 0.8563
F-statistic:	733.1 on 10 and 1219 DF, p-value: < 2.2e-16

MLR model analysis and result

Overall significance of the model: the F statistic is very high (F-statistic: 733.1), and the p value is much less than 0.001, indicating that the model is statistically significant, and we are fully confident that at least one predictor has a significant impact on flight delays.

Significance of explanatory variables:

1. Terminals (terminal2 and terminal3) : These two terminals have significantly fewer delays than the reference group (terminal1).
2. Airlines (airlineAir2 and airlineAir4) : airlineAir2 has significantly longer delays than Air1, while Air4 has significantly more delays. The Air3 delays were relatively short, but the impact was minor.
3. Rain_yes: light delays increase significantly when it rains.
4. Aircraft capacity: An increase in aircraft capacity is significantly associated with an increase in flight delays.
5. temperature: Higher temperatures are significantly associated with increased flight delays.
6. Wind_gust : Although included in the model, the relationship with flight delay time is not significant.
7. Wind_direct : has some positive effect on delay time, but the effect is relatively small.

Goodness of fit:

An R-squared value of 0.8574 indicates that the model explains approximately 85.74% of the variability in flight delay times, and an adjusted R-squared value of 0.8563 indicates that the model fits very well.

Residual analysis: The range of residuals is relatively small and the standard error of residuals is 24.22 minutes.

In summary, the model points out that terminals, airlines, rainfall conditions, aircraft capacity and temperature are important factors affecting flight delays. However, the effect of wind speed is not significant in this model and may need to be analyzed in combination with other variables. Wind direction, while showing some influence, plays a relatively small role. These results provide valuable insights for airport management and airlines to reduce flight delays.

Assumption Checks for Model

Since the effect of model 2 is significantly higher than that of model 1, only the assumption of model 2 is checked in this chapter. It is discussed from three aspects of homoscedasticity, normality of errors and autocorrelation. This paper uses lmtest [14],car [15] and nortest [16] packages in R for the checks.

Checks for homoscedasticity

In the scatter plot of the provided residuals and fitted values, the residuals do not show a constant diffusion around the horizontal line, but show a tendency to increase the diffusion as the fit value increases. This pattern usually indicates the presence of heteroscedasticity, meaning that the model's homoscedasticity assumption may have been violated.

Specifically, the spread of the residuals is smaller at both ends of the fit value and larger in the middle region. This non-constant variance may indicate that the model shows varying degrees of accuracy in predictions of different values, and may require more complex transformations to the model or consider using heteroscedasticity robust standard errors to adjust the model estimates. In addition, this irregular distribution of residuals may also suggest that important predictors are missing from the model or that there are non-linear relationships that are not captured.

According to the Breusch-Pagan test results, the BP statistic is 68.266, corresponding to a P-value of 9.574e-11, which is a very small value, far smaller than any conventional significance level (such as 0.05 or 0.01) [17]. This means that we can reject the null hypothesis of homoscedasticity and assume that the data is heteroscedasticity. In other words, the variance of the residuals is not constant, but changes as the model's predicted value changes. This may require adjustments to the model or data, for example, to improve the model by transforming the response variables or using heteroscedasticity robust standard errors.

Checks for normality of errors

In the provided Q-Q plot, it can be seen that the tail of the data (especially the right tail) deviates from the ideal line of the normal distribution, which indicates that the residual distribution may have a heavier tail. The deviation from the right tail specifically points to some outliers in the model residuals that are higher than would be expected from a normal distribution.

This pattern of the residual distribution in the Q-Q plot usually indicates that the residual does not follow a perfect normal distribution and may require further model diagnosis and adjustment. This deviation may be due to nonlinear relationships in the data, outliers, or the contribution of certain variables to the prediction that is not adequately captured in the model. In actual analysis, it may be necessary to consider data transformations or use more robust statistical methods to deal with this non-normality.

Checks for autocorrelation

According to the Durbin-Watson test results, the DW value is 2.0101, which is very close to the ideal value of 2 without autocorrelation, and the P-value of 0.5711 is much higher than the usual significance level (e.g. 0.05). This shows that there is not enough evidence to reject the null hypothesis of no autocorrelation, that is, there is no autocorrelation problem in the model residuals. Therefore, we can conclude that there is no obvious autocorrelation in the residual of Model 2, which is a good feature in the analysis of time series data. This means that there does not appear to be a systematic pattern of association between adjacent observations in the model.

Model Result Discussion

Based on the comprehensive analysis of Model 2, including the estimation of model coefficients, the scatter plot of residuals, the results of the Breusch-Pagan test, the Q-Q plot, and the Durbin-Watson test, we can draw the following comprehensive conclusions:

1. Model significance:The overall F statistic of the model is very significant, which indicates that we have sufficient evidence that at least one predictor in the model has a significant effect on flight delay time.
2. Significance of variable:Most of the predictive variables, including terminals, airlines, rainfall, aircraft capacity and temperature, have a statistically significant impact on flight delay times. The effect of wind speed is not significant, while the effect of wind direction is significant but relatively small.
3. Homovariance: The Breusch-Pagan test shows that the residual of the model has heteroscedasticity, which may need to be resolved by data transformation or using heteroscedasticity robust estimation methods.
4. Normality: The Q-Q plot which is shown in the Fig11 reveals that the distribution of residuals deviates from the normal distribution at the tail, indicating that there may be nonlinear relationships or uncaptured outliers in the model [18].
5. Autocorrelation: Durbin-Watson test results show that there is no autocorrelation in the residual [19], which is one of the good features of the model.

In summary, Model 2 effectively captures multiple key factors that affect flight delays, the model accounts for a large amount of variability, and performs well in terms of autocorrelation. However, some assumptions of the model, such as variance and normality, are not satisfied. This may affect the standard error of the model coefficient and the accuracy of the prediction interval. Subsequent work may include transforming the data or using more complex models to address these issues, as well as exploring other potential predictors not included in the model.

These conclusions are valuable to airport management because they reveal potential opportunities to reduce flight delays. By optimizing terminal and airline operational processes, as well as preparing for bad weather, flight on-time performance can be significantly improved. In addition, these findings provide the aviation industry with data-driven insights to improve its services and increase customer satisfaction.

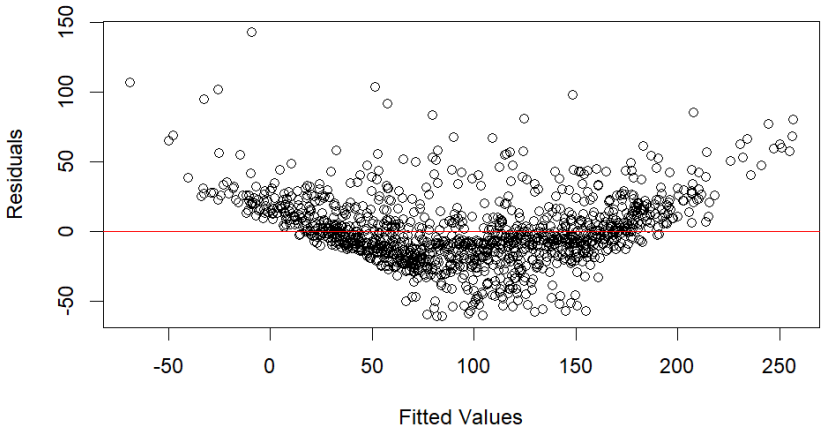

Figure 10: Residual plot for Model 2, presenting the discrepancies between observed and predicted flight delays.

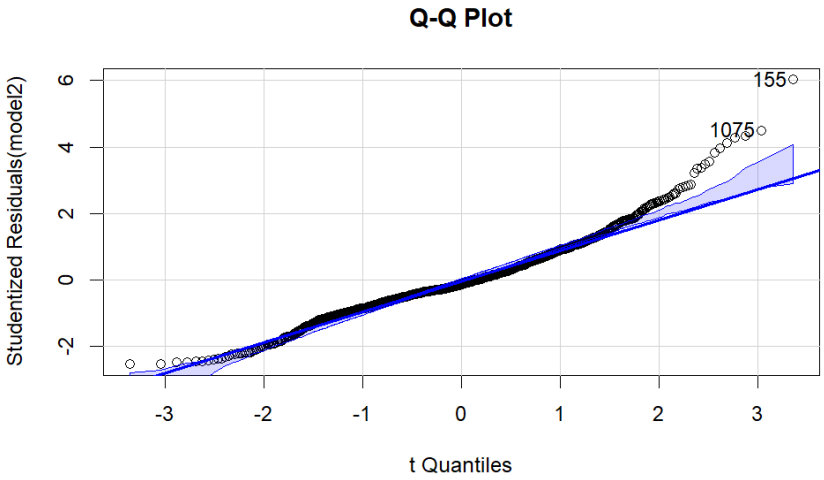

Figure 11: Quantile-Quantile plot for Model 2 residuals, to evaluate the normality of the distribution.

The study's insights into flight delays must take into account some limitations. These data, while comprehensive for a single hypothetical airport, may not reflect the full range of conditions at real airports. The linear models used may not capture the complex, non-linear nature of the actual delay factors. In addition, the lack of detailed causal data, such as exact causes of delays and excluded time series analysis, means that some effects on delays may not have been identified. In addition, due to potential model overfitting, the predictive power of the study may not fully translate to other datasets. Finally, data granularity is limited, which limits the in-depth study of a single or specific cause of delay. These limitations suggest that further study of more diverse and detailed data, possibly combined with non-linear and time series analysis, may yield a more nuanced understanding of flight delays.

Conclusion

In this study, we built and analyzed two models, one is simple linear regression Model1, the other is multiple linear regression Model2, which focuses on the study of multiple linear regression Model2, aiming to reveal the key factors affecting the flight delay time. The model combines multiple predictors, from terminal buildings to weather conditions. The results show that terminal, airline, rainfall, aircraft capacity and temperature are significant factors affecting flight delay time.

Although the overall performance of the model is very strong and can account for most of the variability of the delay time, testing of the model hypothesis shows heteroscedasticity and non-normality of the residual distribution. These findings point to areas where the model may need to be improved, such as by changing variables or using more sophisticated statistical methods to improve the robustness of the model.

In addition, the Durbin-Watson test results show that the assumption of the independence of the residual in the time series data is satisfied. This is positive for regression models because it shows that there is no problem with autocorrelation in model predictions.

Taken together, this study provides important insights into understanding the dynamics of flight delays and provides data support for airport management to help them identify possible intervention points to reduce delays and improve flight punctuality. These conclusions provide the airline industry with strategies to optimize operations and enhance passenger satisfaction. Future work could include exploring other potential contributing factors, as well as adopting more advanced models to more accurately predict and manage flight delays.

References

[1] Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*, 2017.

[2] Yi Ding. Predicting flight delay based on multiple linear regression. In *IOP conference series: Earth and environmental science*, volume 81, page 012198. IOP Publishing, 2017.

[3] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2016.

[4] Selcuk Alemdag, Zulfu Gurocak, and Candan Gokceoglu. A simple regression based approach to estimate deformation modulus of rock masses. *Journal of African Earth Sciences*, 110:75–80, 2015.

[5] Evangelos C Alexopoulos. Introduction to multivariate regression analysis. *Hippokratia*, 14(Suppl 1):23, 2010.

[6] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2020. R package version 1.0.5.

[7] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2020. R package version 1.31.

[8] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[9] Hadley Wickham. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2020. R package version 3.3.3.

[10] Hadley Wickham. *reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*, 2017. R package version 1.4.4.

[11] Tarek Azzam, Stephanie Evergreen, Amy A Germuth, and Susan J Kistler. Data visualization and evaluation. *New Directions for Evaluation*, 2013(139):7–32, 2013.

[12] Nithya J Gogtay and Urmila M Thatte. Principles of correlation analysis. *Journal of the Association of Physicians of India*, 65(3):78–81, 2017.

[13] Bruce S Weir and William G Hill. Estimating f-statistics. *Annual review of genetics*, 36(1):721–750, 2002.

[14] Achim Zeileis and Torsten Hothorn. *lmtest: Testing Linear Regression Models*, 2020. R package version 0.9-38.

[15] John Fox and Sanford Weisberg. *car: Companion to Applied Regression*, 2019. R package version 3.0-10.

[16] Stefan Th. Gries. *nortest: Tests for Normality*, 2020. R package version 1.0-4.

[17] Andreea G Halunga, Chris D Orme, and Takashi Yamagata. A heteroskedasticity robust breusch–pagan test for contemporaneous correlation in dynamic panel data models. *Journal of econometrics*, 198(2):209–230, 2017.

[18] BRSI Das and Sidney I Resnick. Qq plots, random sets and data from a heavy tailed distribution. *Stochastic Models*, 24(1):103–132, 2008.

[19] Marc Nerlove and Kenneth F Wallis. Use of the durbin-watson statistic in inappropriate situations. *Econometrica: Journal of the Econometric Society*, pages 235–238, 1966.

Appendix

Datasheet for the Dataset of 2015 Flight Delays and Cancellations

Motivation

Purpose of Dataset Creation: The 2015 Flight Delays and Cancellations dataset was compiled with the aim of providing insight into the on-time performance of domestic flights operated by large air carriers in the United States. The dataset serves as a resource for analyzing trends in flight delays and cancellations, aiding travelers in making informed decisions and supporting the aviation industry in improving service reliability.

Creators: The dataset was collected and published by the U.S. Department of Transportation's Bureau of Transportation Statistics as part of their ongoing efforts to monitor and report on the quality of air travel.

Composition

Data Instances: The dataset includes detailed records for domestic flights in 2015, featuring data on delays, cancellations, and operational aspects of the flights. Attributes include delay time, terminal, airline, capacity, temperature, rain presence, wind gust speed, and wind direction.

Number of Instances: The dataset encompasses a comprehensive collection of flight records throughout 2015, with each record corresponding to a unique flight instance.

Sampling: Data represents an exhaustive collection of reported flights, capturing a wide array of conditions and outcomes to accurately reflect the year's flight operations.

Data Type: Quantitative measurements (e.g., delay in minutes, temperature, wind gust speed) and categorical data (e.g., terminal, airline, rain presence).

Labels: Records are identifiable by attributes such as date, flight number, airline, and terminal, ensuring detailed traceability.

Data Splits: Not applicable, as the dataset provides a holistic view of the year's flight operations without division into training or testing sets.

Collection Process

Data Collection: Information was aggregated from official flight operation reports, ensuring accuracy and reliability. The collection process adhered to the DOT's standards for data reporting and integrity.

Preprocessing/Cleaning/Labeling

Preprocessing Done: Data underwent rigorous cleaning and structuring, aligning with the International Classification of Diseases for categorizing delays and ensuring uniformity across records.

Raw Data Availability: While the processed dataset is openly accessible, raw data may be obtained through official requests to the DOT.

Uses

Current Uses: The dataset is utilized by travelers, researchers, and industry analysts for understanding and predicting flight delay patterns, enhancing travel planning, and facilitating improvements in airline operational efficiency.

Potential Uses: Future applications may extend to machine learning models for delay prediction, optimization studies for airport operations, and academic research in aviation management and policy development.

Restrictions on Use: Distributed under CC0: Public Domain License, allowing unrestricted use, distribution, and modification.

Distribution

Distribution Channels: Available for download through the Bureau of Transportation Statistics website and related data platforms.

Licence: CC0: Public Domain.

Maintenance

Maintainers: Oversight is provided by the Bureau of Transportation Statistics, ensuring the dataset's relevance and accuracy.

Update Schedule: The dataset specifically covers the year 2015 and does not specify an update frequency.

Ethical Considerations

Data Privacy: Aggregated data ensures individual flight details are not personally identifiable, prioritizing privacy while offering valuable insights.

Ethical Oversight: Compiled with consideration for ethical standards in data reporting and public dissemination, focusing on the collective benefit of improved air travel understanding.

Disclaimer

Limitations: The dataset's scope is confined to flights in 2015, and interpretations should consider temporal and geographical context.

Errata and Updates: Users are encouraged to consult the Bureau of Transportation Statistics for the latest data and corrections.

This datasheet provides a comprehensive overview of the 2015 Flight Delays and Cancellations dataset, underlining its significance, composition, and practical applications. It aims to foster informed utilization of the data, supporting endeavors to advance the understanding and management of flight operations.