Do Flight Delays Increase the Likelihood of Baggage Mishandling?

Jessica Weeks and Tracey Zicherman[1]
Merrimack College Graduate Programs
DSE6311 Data Science Capstone
Dr. Geist
August 25, 2024

[1] Each author contributed equally to the design, coding & development, analysis, and writing of this project.

Abstract

This study investigated the relationship between flight delays and the likelihood of baggage mishandling, a concern that affected millions of airline passengers. Given the prevalence of both issues, understanding their potential connection was crucial for improving the air travel experience. Utilizing datasets from the Bureau of Transportation Statistics, we aggregated flight-level data to the carrier-month level and developed predictive models, including a linear regression model, to explore this relationship.

Our analysis revealed a significant positive correlation between flight delays and baggage mishandling. Specifically, carriers with a higher proportion of late aircraft delayed flights were found to have a substantially increased average mishandled baggage ratio. Additionally, longer maximum departure delays were associated with higher mishandling rates. These findings suggested that operational delays could exacerbate baggage handling inefficiencies, highlighting the importance of timely operations in minimizing baggage mishandling. For airlines, these insights underscored the need for strategies that focused on punctuality and addressed the broader operational impacts of delays.

Background & Research Question

With air travel at an all-time high, the aviation industry is facing a surge in some of the most common and frustrating problems that affect passengers: flight delays and baggage mishandling. Millions of travelers in the U.S. are familiar with the inconvenience of flight delays, which can lead to missed connections and disrupted plans. Baggage mishandling, on the other hand, can leave passengers stranded without their essentials or cherished belongings. The fear of arriving at a destination without one's baggage is so pervasive that many passengers go to great lengths to avoid checking luggage altogether (Dawes, 2023). In 2022, the U.S. Department of Transportation reported a record number of customer complaints, with flight delays and mishandled baggage ranking first and third, respectively (Murray, 2023). These issues cost customers time and money and significantly undermine the travel experience.

Given that nearly 90% of Americans travel by air (Airlines For America, 2023), understanding the potential connection between flight delays and baggage mishandling is highly relevant. Despite the prevalence of these issues, the relationship between them has not been thoroughly explored. This project sought to investigate whether there was a correlation between flight delays and the likelihood of baggage mishandling. By uncovering patterns or connections between these two problems, we aimed to provide insights that could help travelers make more informed decisions and better prepare for potential disruptions, ultimately reducing stress and unexpected costs.

Data Acquisition

Data on monthly air carrier mishandled baggage claims and flight level on-time performance were obtained from the Bureau of Transportation Statistics.

These datasets were chosen due to the presence of critical features that supported our analysis. The on-time flight datasets included various types of delays, their duration in minutes, and binary classifications for canceled or diverted flights. The mishandled baggage datasets included the number of mishandling incidents and the number of passengers.

Data had to be downloaded manually – the mishandled baggage data as a single .csv file, and the on-time performance data as 36 .csv files, comprising the three years from 2016-2018. Scripts were written to clean and aggregate the on-time performance data resulting in a large dataset of roughly 1.4 GB, then merged with the mishandled baggage data resulting in a single dataset of monthly aggregate statistics for the years 2016-2018.

The monthly aggregate dataset was subjected to further cleaning, exploratory analysis, preprocessing, and feature selection, resulting in several intermediate datasets. As the initial dataset contained many features, most of which did not make it through the data processing pipeline, we do not include a data dictionary for all features. We do, however, present dictionaries for the features in the initial monthly aggregate dataset, as well as the features in the final modeling dataset, some of which were engineered (see Appendix).

Data Cleaning

All data manipulation was performed using the R programming language. Some minimal cleaning was performed and the on-time flight performance datasets were aggregated into a single .csv of approximately 18 million rows. We note that to preserve as much information as possible, the on-time flight data was not aggregated before initial exploratory data analysis. In particular, this initial analysis was used to inform any further cleaning and preprocessing steps that should be performed, and which aggregate functions should be chosen.

The following preliminary steps were performed sequentially:

1. *Preliminary wrangling and cleaning*: The on-time flight performance data was loaded into a tibble of approximately 18.5 million rows from 36 .csv files covering 2016-2018. A subset of relevant features was selected from the raw data files. Of these features, those related to delay types (for example, carrier versus weather delay) were missing values for non-delayed flights, so these were re-coded as 0. A small number of flights missing delay times were dropped. The resulting single tibble was stored for preliminary EDA. The mishandled baggage data for 2016-2021 was loaded into a single tibble of approximately 1000 rows, and a subset of relevant features was selected.
2. *Preliminary data analysis*: Summary statistics were reported and missing and problematic values were investigated, including outliers. Histograms of arrival and departure delay features were produced, as well as additional visualizations. Scatterplots and correlation matrices were generated to assess collinearity.

A secondary EDA was conducted to refine feature selection, updating the target feature to `AVG_MISHAND_RATIO`. Two new features, `AIRCRAFT_UTIL_RATIO (NUM_FLIGHTS/PASSENGERS)` and `HIGH_UTIL_RATIO (AIRCRAFT_UTIL_RATIO > 0.015)`, were created. Initial regression models were re-fit with these updates, improving the fit, and additional feature selection followed. The preprocessing pipeline was updated to include scaling for regularized models and K-means, as well as one-hot encoding of the categorical feature `CARRIER_NAME`. Hyperparameters for principal components analysis and elastic net regression models were tuned for better performance, and an additional random forest model was fit and trained.

Data Exploration

After preliminary and secondary EDA and preprocessing, the final EDA was performed. Single and paired variable plots of important features (including the target feature `AVG_MISHAND_RATIO`) were produced and are detailed below.

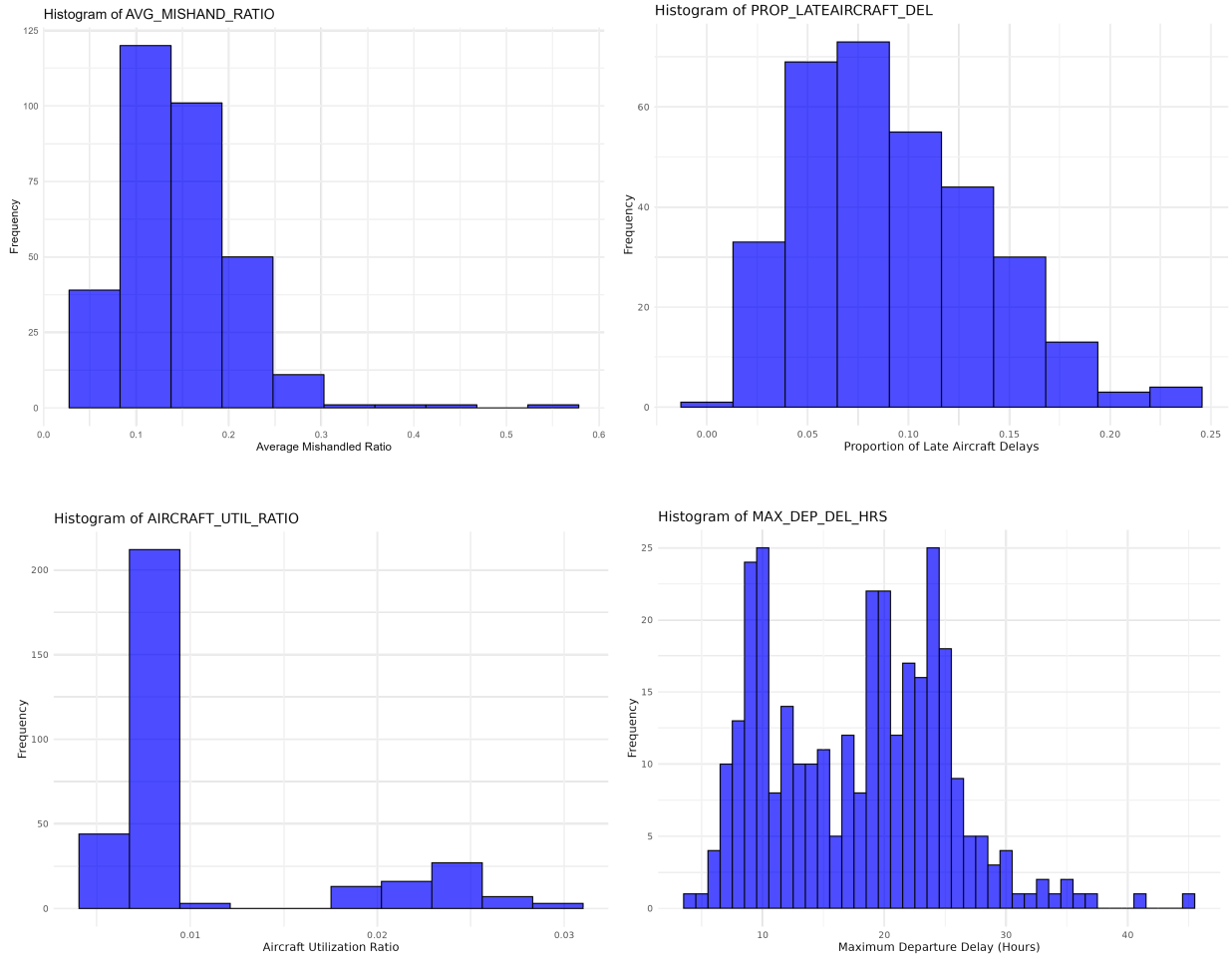

Fig 1: Histograms of Continuous Features in Final Dataset

The distribution of `AVG_MISHAND_RATIO` has a strong right skew, indicating most carriers have small values and a few have large ones. The proportion of late aircraft delayed `PROP_LATEAIRCRAFT_DEL` and maximum departure delay time `MAX_DEP_DEL_HRS` are

also right skewed but only slightly, and MAX_DEP_DEL_HRS has two peaks, suggesting a bimodal distribution. Finally, note that there is a large gap between low and high values of AIRCRAFT_UTIL_RATIO. These histograms indicate good distributions for modeling, without requiring further feature transformations.

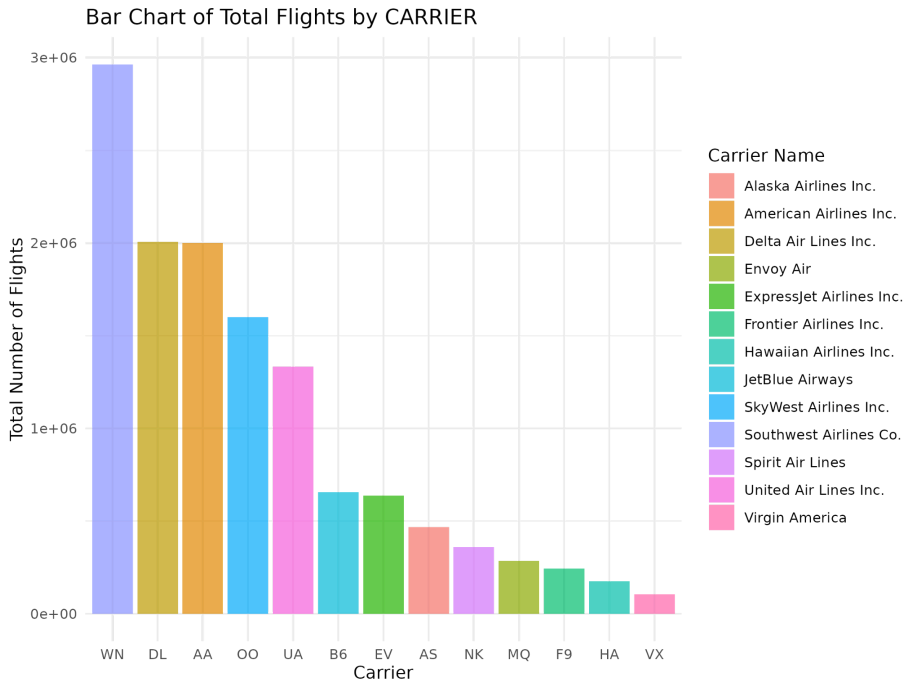

Fig 2: Total Flights by Carrier

The bar chart above shows the large differences between carriers' total number of flights, which has strong implications for the number of mishandled baggage items. For this reason, the number of carrier flights (NUM_FLIGHTS) was not included in the final model, though we included this plot for illustrative purposes.

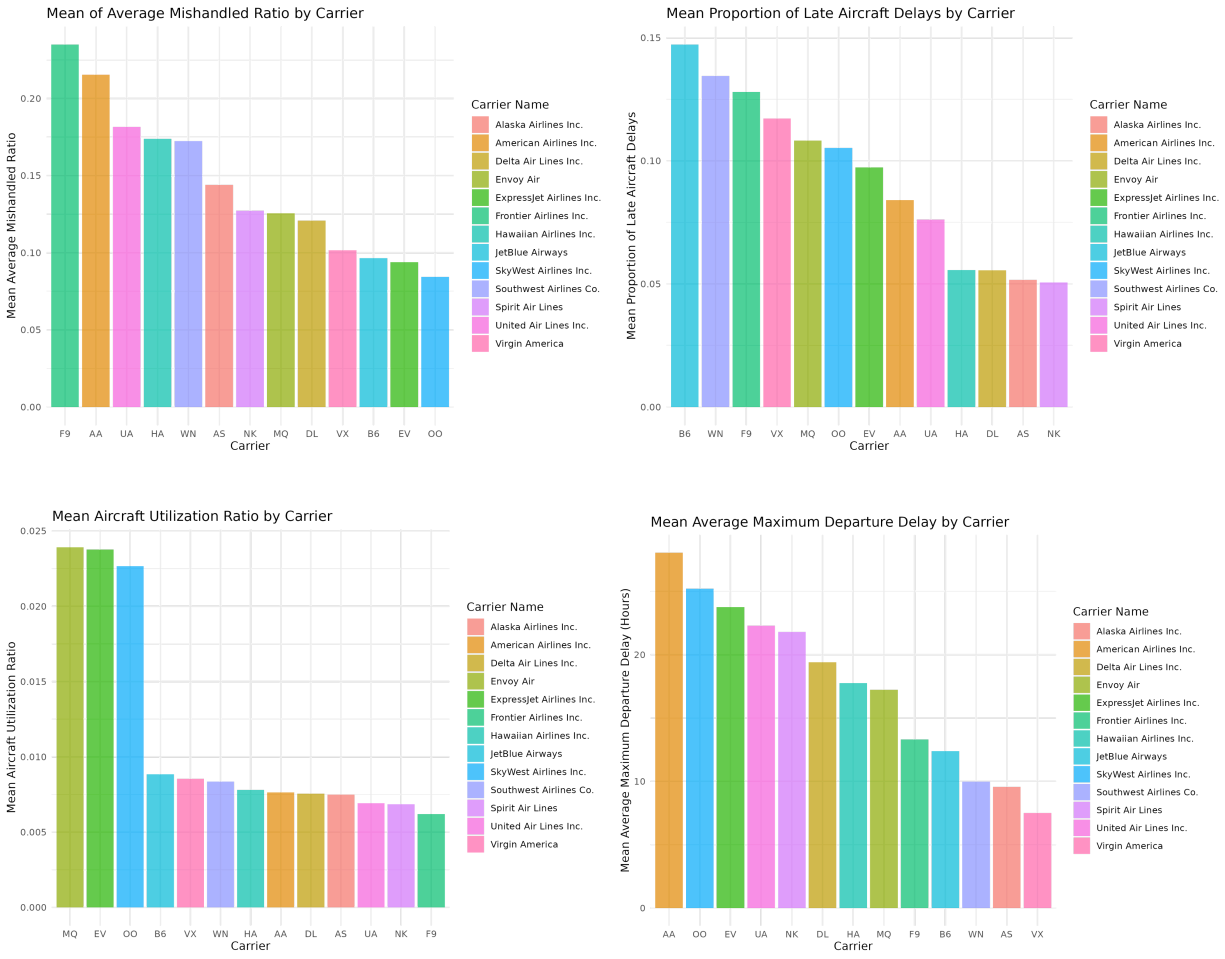

Fig 3: Mean Of Continuous Features by Carrier

The bar charts above show the mean values of each continuous feature over all months in the data set, colored by carrier. Note there is considerable variation between these. Also note that there is a large gap in aircraft utilization ratio between the carriers with the three largest values, and the rest, which is consistent with the gap observed in the histogram above.

Fig 4: Scatterplots of Pairs of Continuous features

A few patterns are suggested from these scatterplots. There is a clear positive correlation between the proportion of both late aircraft delayed flights and maximum departure delay and the average mishandled baggage ratio. The clear gap in aircraft utilization ratio between a cluster of several carriers with a high aircraft utilization ratio and the rest is evident, and this cluster of carriers has a much lower mean average mishandled ratio than the rest. Furthermore, the indicator variables for these carriers all appear in the final model showing a significant negative association with the average mishandled baggage ratio.

Models

Pre-Processing & Feature Engineering

The following preprocessing steps were performed sequentially:

1. *Custom subset feature selection*. A final set of relevant features was chosen from the preliminary steps above. Arrival and departure delay times, binary features encoding flight cancellation and diversion, as well as delay time features related to the type of delay (e.g. weather versus security) were chosen from the on-time performance dataset. The number of passengers and the number of mishandled baggage pieces were chosen from the mishandled baggage dataset. Year, month, and carrier were chosen from both datasets. These feature selections were guided primarily by domain and statistical knowledge. The years 2019-2021 were dropped from the mishandled baggage dataset, given that the number of passengers was missing for those years. It was determined that this was an important control feature, given the high probability of a strong but confounding positive association between the number of passengers (and hence airline carrier size) and the number of mishandled pieces of baggage.
2. The arrival delay feature was dropped because of low predictive value and potential confounding, given its extremely high correlation with departure delay ($\boldsymbol{\rho} = 0.95$) and redundancy.

The flight-level on-time performance data was aggregated by month and carrier to enable joining with the monthly carrier-level data in the mishandled baggage dataset. After all additional features were engineered, the on-time and mishandled baggage datasets were joined on month, year, and carrier features, to create a final single dataset.

Before aggregating by month, the following features were created in the on-time flight performance dataset.

- `DELAYED`: Indicator for whether flight is delayed on departure.
- `IS_CARRIER_DEL, IS_WEATHER_DEL, IS_NAS_DEL, IS_LATEAIRCRAFT_DEL`: Indicators for carrier, weather, national air system [NAS], security, and late aircraft delay.

During aggregation, the following additional monthly features were created.

- `NUM_FLIGHTS`: Number of carrier flights.
- `PROP_DELAYED, PROP_CARRIER_DEL, PROP_WEATHER_DEL, PROP_LATEAIRCRAFT_DEL`: The proportion of flights satisfying the corresponding indicator variable (mean of binary 0/1 coded feature `CANCELLED`).
- `MIN_DEP_DEL, Q1_DEP_DEL, MED_DEP_DEL, Q3_DEP_DEL, MAX_DEP_DEL`: Five summary statistics of departure delay time.

After merging, the following features were created.

- `MISHAND_PASS_RATIO`: Ratio of mishandled baggage to the number of passengers.
- `MISHAND_FLIGHTS_RATIO`: Ratio of mishandled baggage to the number of flights.
- `AVG_MISHAND_RATIO`: Average ratio of `MISH_PASS_RATIO` and `MISHAND_FLIGHTS_RATIO`.

The above ratio variables were created to measure carrier “efficiency” or “accuracy” concerning baggage handling. In this case, efficiency is defined as low numbers of mishandled baggage relative to the size of airline operations, as measured by passenger and flight numbers.

After further deliberation, it was determined that one of the target derivative features `MISHAND_PASS_RATIO, MISHAND_FLIGHTS_RATIO, AVG_MISHANDLED_RATIO` will be a better target feature than the original chosen target `MISHANDLED_BAGGAGE`(number of mishandled baggage claims) since these are less sensitive to carrier size. The result of this feature engineering was a total of 17 numeric features, of which 11 were selected after diagnostics and experimentation to be used for preliminary modeling.

The initial research question we proposed was to investigate the relationship between the aircraft departure delay and mishandled baggage. As it was phrased, the initial analysis plan assumed this relationship could be investigated at the flight level. However, gaining a better understanding of the datasets led to an adjustment in the plan. The mishandled baggage dataset was already aggregated at the monthly carrier level, and the on-time flight performance dataset needed to be aggregated to match. Subsequently, the research question was amended to hypothesize a relationship between the number of mishandled baggage pieces and aggregate monthly delay statistics.

Furthermore, data wrangling and preprocessing proved to be more challenging than expected. Preliminary EDA revealed that the number of passengers was only present in the mishandled baggage dataset for 2016-2018, while the original on-time performance dataset only began in 2018. The resulting overlapping joint dataset had only 140 rows which was deemed too small of a sample size for reliable statistical inference and regression analysis. Therefore, more data was needed, and fortunately, another on-time performance dataset on the BTS website was found that provided the same features for the years 2016-2018. After aggregating the on-time performance dataset and joining with the mishandled baggage dataset, the resulting dataset had 325 rows, sufficient to support robust regression analysis.

Preprocessing and feature engineering resulted in several useful features as measured by the significance of the coefficient estimates[2] in the initial regression model. The principal component analysis provided a possible path toward dimensional reduction, however, it was decided to skip this step, given the project objective of interpretability. K-means clustering indicated no useful cluster structure in the data, so no cluster label feature will be used in modeling.

We used a final set of 9 unstandardized features for final model selection, evaluation, and interpretation. In initial modeling and tuning, we used a set of 11 standardized features, but further investigation revealed a relatively high variance inflation factor ($VIF \approx 10.12$) for the proportion of carrier delayed flights (`PROP_CARRIER_DELAY`) and a relatively low significance ($p \approx 0.012$) for the Alaska Airlines indicator feature (`CARRIER_NAMEAlaska_Airlines_Inc`) so these features were dropped in the final model.

The primary reason for using unstandardized features was the overall project objective of interpretability, and it was determined that unstandardized features would provide more insight into the relationship between flight delay times and mishandled baggage than standardized features, particularly because unstandardized features retain their original units. We also converted the maximum monthly departure delay from minutes (`MAX_DEP_DELAY`) to (`MAX_DEP_DELAY_HRS`) since the latter was a more natural scale.

The target feature, average mishandled baggage ratio (`AVG_MISHAND_RATIO`), was engineered. This feature was calculated as the average of two ratios, the number of mishandled

[2] Hence, correlation, or equivalently, degree of linear association.

baggage items per number of flights (`MISHAND_BAGGAGE_RATIO`) and per number of passengers (`MISHAND_PASSENGER_RATIO`):

$$AVG\_MISHANDLED\_RATIO = \frac{1}{2}(MISHAND\_BAGGAGE\_RATIO + MISHAND\_PASSENGER\_RATIO)$$

$$= \frac{1}{2}\left(\frac{MISHANDLED\_BAGGAGE}{NUM\_FLIGHTS} + \frac{MISHANDLED\_BAGGAGE}{PASSENGERS}\right)$$

The ratios were chosen due to the obvious (and uninteresting) natural association between the number of flights and passengers and the number of mishandled baggage items – the more flights taken and the more passengers carried, the more baggage items that will be processed, and the more that will be mishandled. Both ratios ignore this association by focusing on how many baggage items are mishandled per flight or passenger. Finally, the average of the two is taken to provide balance (since it was unclear which ratio was preferred).

Algorithms Selection

Unsupervised Methods

After feature engineering, unsupervised methods were used on the resulting features to aid in pattern discovery. *Principal component analysis* was performed on 19 numerical features. The cumulative variances of the principal components and their differences were calculated and plotted to facilitate the selection of the number of principal components for possible dimensional reduction.

The cumulative variance plot did not show a clear elbow, but the difference in cumulative variances revealed a substantial flattening at 12 components. This will be investigated further in future work, specifically whether the initial model fit on the full dataset can be improved by dimensional reduction to the first 12 PCA components.

After performing PCA to simplify the data, *k-means clustering* was investigated to discover potential clustering structures. The algorithm was run multiple times with different numbers of groups (from 2 to 20) and measured how well the groups were separated. Three metrics were evaluated. WCSS demonstrated how similar the points are within each group, BCSS for how different the points are between groups, and Silhouette scores for how well each point fits into its assigned group. This was done for the original dataset (with 19 features) and the simplified dataset (with 12 PCA components). The lack of elbows in all WCSS-BCSS plots and the relatively low silhouette scores indicated that a good cluster structure was not evident in this data set.

Feature selection and engineering resulted in a final dataset suitable for modeling which contained 11 numeric features. Preprocessing and feature engineering resulted in several useful features as measured by the significance of the coefficient estimates[3] in the initial regression model. The principal component analysis provided a possible path toward dimensional reduction, however, it was decided to skip this step, given the project objective of interpretability. K-means clustering indicated no useful cluster structure in the data, so no cluster label feature will be used in modeling.

Supervised Methods

While statistical models such as gradient-boosted trees or neural networks can provide greater predictive accuracy, they are "black-box" methods that provide little information to be understood by analysts or acted on by stakeholders. With that in mind, it was determined that insight into the research question was preferable over predictive accuracy, and the family of regression models was selected.

[3] Hence, correlation, or equivalently, degree of linear association.

An initial linear regression model was fit and diagnostic plots were generated to test the model assumptions, namely that the target is a linear function of the feature plus normally distributed errors with constant variance. Model diagnostics indicated possible deviation from modeling assumptions.

The residuals-versus-fitted and scale-location plots indicated non-constant error variance of the residuals especially for extreme values of the target – in particular, for lower values, there was a negative bias in the residuals, while variance seemed to increase as the fitted values increased.

The Q-Q residuals plot showed relatively good support for the normality of the residuals, albeit with some deviation near the ends, indicating larger tails of the error distribution. Finally, the residuals-vs-leverage plot showed only one potentially high residual/leverage point, which appeared close to Cook's distance contour of 0.5 (possible outlier).

Taken together, these diagnostic plots showed that the data doesn't perfectly fit the assumptions of a linear regression model. Specifically, the errors don't exhibit constant variance and aren't normally distributed, which could affect the accuracy of our initial model. Later steps were taken to improve the final model fit, causing the resulting diagnostics to improve dramatically.

The F-test for overall model fit was highly significant ($p < 2.2e\text{-}16$), providing strong evidence that at least one of the feature coefficients was non-zero. Multiple R-squared and adjusted R-squared values (0.8993 and 0.8958 respectively) indicated a good fit.

Coefficient estimates[4] of the number of passengers, the proportion of delayed flights, the first and third quartiles of departure delay time, and the maximum delay time were all highly significant (features `PASSENGERS, PROP_DELAYED, Q1_DEP_DEL, Q3_DEP_DEL,`

[4] That is, the estimated degree of linear relationship.

MAX_DEP_DELAY). The number of flights and proportion of weather-delayed flights (NUM_FLIGHTS, PROP_WEATHER_DEL) were also significant at the 0.05 significance level. The proportion of carrier, late aircraft delayed flights, minimum delay time, and median delay time (PROP_CARRIER_DEL, PROP_LATEAIRCRAFT_DEL, MIN_DEP_DEL, MED_DEP_DELAY) were not found to be significant.

Given the potential of even a simple model such as linear regression to overfit given enough data and features, model fit can often be improved by regularization methods, which artificially constrain model complexity. For this reason, an initial elastic net regression model was also investigated. Its regularization penalty is a combination of the L_1 (lasso) and L_2 (ridge) regularization penalties, balancing out the relative advantages and disadvantages of both models (Zou & Hastie, 2005). Specifically, the elastic regularization penalty is a weighted average of L_1 and L_2 regularization penalties, with the weight parameter α which can be tuned to optimize the penalty.

We tuned the elastic net hyperparameters, first by using an initial α value of 0.5 (average of lasso and ridge) and 10-fold cross-validation to tune the shrinkage parameter λ which governs the strength of the L_1 and L_2 regularization[5]. The weight parameter α value was subsequently tuned, and an optimal elastic regression model was fit.

Hyperparameter	Tuned Value
alpha	0.8
lambda	0.002

Fig 5: Elastic Net Hyperparameter Tuned Values

[5] Note the shrinkage parameter λ is distinct from the weight parameter α. The shrinkage parameter controls the degree of restriction of the parameter space, while the weight parameter controls the relative balance of lasso and ridge penalties, themselves governed by shrinkage.

We also fit a random forest model and tuned the number of trees and the number of features used at each tree split.

Hyperparameter	Tuned Value
mtry	3
ntree	200

Fig 6: Random Forest hyperparameter tuning results.

To facilitate direct comparison with all tuned models, we ran a 10-fold CV estimation for several regression metrics on the $d = 9$ unstandardized features. We chose CV as it is known to be a less biased measure of generalization performance than the point estimates obtained from the train-test split. The selected CV regression metrics – root-mean-squared-error (RMSE), mean absolute error (MAE), and R-squared – were chosen for interpretability on the unstandardized feature because they are measured on the same scale[6]. RMSE and MAE measure predictive accuracy while the R-squared measures goodness-of-fit.

Model	RMSE	MAE	R-squared
Linear Regression	0.0335	0.0253	0.595
Elastic Net	0.0337	0.0253	0.592
Random Forest	0.211	0.0162	0.839

Fig 7: 10-fold CV estimates of evaluation metrics.

We note that by these CV estimates, the random forest has both better predictive accuracy and better goodness of fit. That said, the numbers for linear regression and elastic net are respectable,

[6] For example, root-mean-squared-error RMSE is in the same units as the target feature. So for example, in our case, MSE is in units of the square of the average mishandled ratio, while RMSE is just the same units of average mishandled ratio (which are dimensionless, i.e. purely numerical).

and we note that random forest models are easily known to overfit.[7] Moreover, random forest models are essentially black-box models, or close to it. Feature importance estimates do not provide a clear interpretation, though they can be helpful and suggestive of which features the model determines have a strong association with the target.

Final Model

As mentioned previously, the main objective of this project, with the specific research question we had in mind, was to understand the relationship between flight delay times and the number of mishandled baggage items. Thus, model interpretability was deemed more important than predictive accuracy.

Furthermore, we specifically hypothesized a positive relationship between flight delays and mishandled baggage, so we preferred an interpretable model that allowed us to test this hypothesis.

Given the main project objective of interpretability and the possibility of random forest overfitting, the decision was thus made to avoid the random forest regression model. Moreover, given that the elastic net and linear regression model RMSE, MAE and r-squared CV estimates were essentially identical (if not slightly better) we chose the linear regression for our final model.

Model Interpretation

The final linear regression model was chosen for interpretability, and we will interpret the coefficient estimates shortly. First, we say a few words about specification and goodness of fit.

[7] We did not have a large enough number of observations (n = 322) to provide a reliable measure of overfitting through train-test splits, and since we prioritized interpretability, it wasn't deemed necessary.

The final model diagnostic plots reveal excellent support for the underlying model assumptions of least-squared multiple regression, namely a linear functional relationship between features and target, as well as constant variance and normal distribution of residuals. Moreover, the residuals vs. leverage plot does not display the Cook's distance contours, indicating the absence of outliers. From this perspective, the model is well-specified.

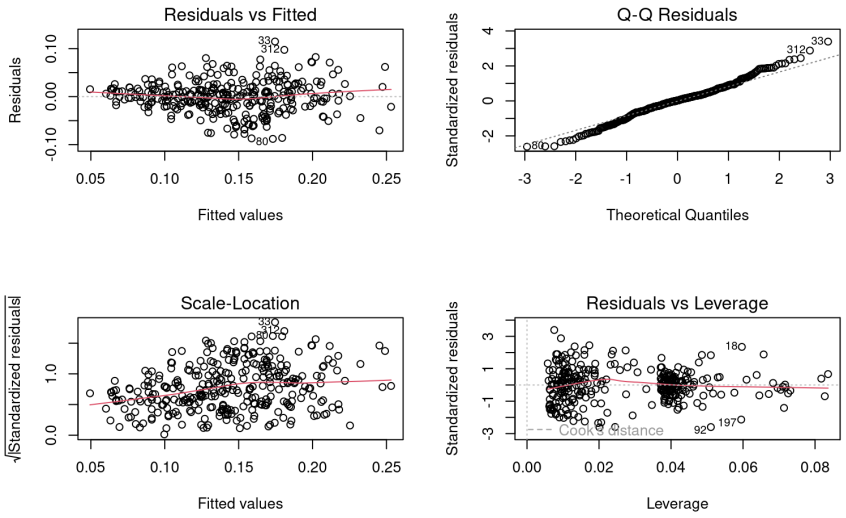

Fig 8: Final Linear Regression Model Diagnostic Plots

The F-statistic is extremely significant ($p < 2.2e\text{-}16$) providing strong evidence for the alternative hypothesis of at least one non-zero coefficient. The R-squared estimate of 0.592 is respectable.

We also note that usage of the R-squared statistic to assess goodness of fit can be controversial. For our purposes, it is worth mentioning that R-squared increases with the number of features, so it can mask overfitting, as the model may be capturing noise rather than an underlying true association (Goyal, 2021). This is consistent with the reduction in R-squared from an initial estimate of 0.6612 using $d = 11$. A simpler model with fewer features generally has a lower R-squared and still may be preferred from this perspective.

Finally, all model features were highly significant, with very small p-values. Thus we find the model is a good fit.

Statistic	Value
Residual Standard Error	0.03399
Degrees of Freedom	313
Multiple R-squared	0.5953
Adjusted R-squared	0.585
F-statistic	57.55
p-value	< 2.2e-16

Fig 9: Summary Statistics for Final Linear Regression Model (Unstandardized Data)

Now we report our main results, in the form of coefficient estimates, some of which directly pertain to our original research hypothesis. Recall that each observation in the final dataset corresponds to aggregate carrier statistics by month.

Feature	Coefficient	Standard Error	p-value
(Intercept)	0.1055484	0.0085363	< 2e-16
PROP_LATEAIRCRAFT_DEL	0.352310	0.0508868	2.50e-11
MAX_DEP_DEL_HRS	0.0028222	0.0003031	< 2e-16
AIRCRAFT_UTIL_RATIO	-2.6053639	0.5852653	1.19e-05
CARRIER_NAMEExpressJet Airlines Inc	-0.0510089	0.0113873	1.05e-05
CARRIER_NAMEFrontier Airlines Inc	0.0356716	0.0079762	1.08e-05
CARRIER_NAMEJetBlue Airways	-0.0729467	0.0079756	< 2e-16
CARRIER_NAMESkyWest Airlines Inc.	-0.0702952	0.0109960	5.91e-10
CARRIER_NAMESpirit Air Lines	-0.0398813	0.0073477	1.15e-07

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

Fig 10: Coefficient Estimates for Final Linear Regression Model

First, note all estimated *p*-values are extremely small. Given the relatively small sample size ($n = 322$) this is very strong evidence of statistical significance for all features.

Of note, the coefficient estimate (`0.2356`) for maximum departure delay (`MAX_DEP_DEL_HRS`) indicates a positive linear relationship with average mishandled baggage ratio (`AVG_MISHANDLED_RATIO`) – for every increase in 1 hr of maximum delayed flight for the month, a carrier's average mishandled baggage ratio increased by roughly 0.2%. The coefficient estimate for the proportion of late aircraft delayed flights (`PROP_CARRIER_DELAY`) indicates a *strong* positive linear relationship with the average mishandled baggage ratio – for every increase in 10% of the proportion of monthly flights delayed due to a late aircraft, a carrier's average mishandled baggage ratio increased by roughly 35%, a considerable increase.

Taken together, these results provide strong evidence in support of the research hypothesis that carriers with more delayed flights tend to have more mishandled baggage.

Conclusions

This study aimed to explore the relationship between flight delays and baggage mishandling, a significant concern for millions of airline passengers. The research was motivated by the hypothesis that increased delays would correlate with a higher incidence of mishandled baggage. To investigate this, we utilized datasets from the Bureau of Transportation Statistics, specifically focusing on on-time flight performance and baggage mishandling reports. Through rigorous data cleaning, feature engineering, and exploratory data analysis, we aggregated flight-level data to the carrier-month level. We also developed several predictive models, including linear regression and elastic net regression, to examine these relationships.

Our analysis yielded several key insights. First, the linear regression model revealed a statistically significant positive relationship between the proportion of late aircraft delayed flights (`PROP_LATEAIRCRAFT_DEL`) and the average mishandled baggage ratio (`AVG_MISHANDLED_RATIO`). Specifically, a 10% increase in delayed flights due to late

aircraft was associated with a 35% increase in the average mishandled baggage ratio. Similarly, maximum departure delay hours (`MAX_DEP_DEL_HRS`) also showed a positive relationship with mishandled baggage, supporting our hypothesis that longer delays contribute to increased baggage mishandling. Including carrier-specific variables further underscored the variability in performance across different airlines, with some carriers exhibiting significantly lower mishandling rates even with similar delay profiles.

These findings provide strong evidence supporting the initial hypothesis that flight delays and baggage mishandling are indeed interconnected. The results suggest that operational inefficiencies leading to delays may spill over into baggage handling processes, thus increasing the likelihood of mishandling. For passengers, this highlights the importance of considering an airline's delay history when planning travel. For airlines, the study underscores the need to address not just punctuality, but also the broader operational impacts of delays to improve customer satisfaction. Future research could extend these findings by exploring additional factors, such as seasonal variations or the impacts of specific delay causes, to develop more comprehensive strategies for mitigating baggage mishandling.

Discussion and Next Steps

Summary of Key Takeaways

This study set out to explore the relationship between flight delays and baggage mishandling, a pressing issue for both passengers and airlines. Our primary research question focused on whether there is a positive correlation between flight delays and the frequency of mishandled baggage. Through detailed data acquisition, cleaning, and feature engineering, we developed a robust dataset that allowed us to perform a thorough analysis. The best model for

addressing our research question was a well-specified and well-fit linear regression model, which provided clear insights into the relationship between delays and baggage mishandling.

The final linear regression model, chosen for its interpretability, revealed strong, statistically significant associations between key delay metrics and baggage mishandling. Specifically, we found that an increase in the proportion of late aircraft delayed flights (`PROP_LATEAIRCRAFT_DEL`) is associated with a substantial rise in the average mishandled baggage ratio (`AVG_MISHANDLED_RATIO`). Additionally, longer maximum departure delays (`MAX_DEP_DEL_HRS`) were linked to higher mishandling rates. These results, supported by the regression coefficients and the low *p*-values (Fig 3, Fig 6), provide compelling evidence that operational delays significantly impact baggage handling efficiency. This aligns with our initial hypothesis and underscores the utility of the chosen linear regression model in answering our research questions.

Recommendations and Future Directions

Based on our findings, we recommend that airlines and airport management focus on minimizing delays, particularly those related to late aircraft, as a strategy to reduce baggage mishandling. Given the strong correlation between delay metrics and mishandled baggage, improving on-time performance could directly enhance customer satisfaction and reduce complaint volumes. Specifically, management could invest in better logistical coordination, more efficient ground operations, and proactive delay management strategies to mitigate the downstream effects on baggage handling.

For future analyses, several extensions of this work could be beneficial. First, incorporating more granular data, such as flight-level handling procedures or passenger feedback, could provide deeper insights into specific operational shortcomings. Additionally,

analyzing seasonal trends or the impact of external factors, such as weather events or peak travel periods, could reveal more about the conditions under which delays most severely affect baggage handling. Another promising direction would be to explore machine learning models with a focus on interpretability, such as decision trees or rule-based systems, to identify non-linear relationships that might be missed by linear regression.

Caveats and Concerns

While our analysis provides strong evidence of the relationship between flight delays and baggage mishandling, several caveats should be noted. First, the dataset was aggregated at the monthly carrier level, which, while necessary for merging with the mishandled baggage dataset, may obscure flight-level variations and nuances. This aggregation could potentially dilute the strength of individual associations or mask outliers that could be significant in a more granular analysis.

Additionally, the reliance on a linear regression model, while justified for interpretability, inherently assumes linear relationships between variables, which might oversimplify more complex interactions. While model diagnostics indicated a good fit and well-specified model (Fig 2, Fig 6), the possibility of omitted variable bias remains, especially given the complexity of airline operations.

In conclusion, while the findings are robust and provide actionable insights, future work should aim to address these limitations by exploring more detailed datasets and considering alternative modeling approaches that can capture non-linear relationships and interactions.

Code Availability

To access the data and code used for this project, visit this GitHub repository.

References

Airlines for America. (2023, March 22). *New survey: Nearly 90 percent of Americans have flown commercially.* https://www.airlines.org/new-survey-nearly-90-percent-of-americans-have-flown-commercially/

Dawes, J. (2023, March 20). *The real reasons behind air travel baggage delays.* Skift. https://skift.com/2023/03/20/the-real-reasons-behind-air-travel-baggage-delays

Goyal, C. (2021, May 15). *The game of increasing R squared in a regression model.* Analytics Vidya. https://www.analyticsvidhya.com/blog/2021/05/the-game-of-increasing-r-squared-in-a-regression-model/#:~:text=Drawbacks%20of%20using%20R%20Squared%20:&text=%F0%9F%91%89%20R2%20assumes%20that,modified%20version%20of%20R2.

Murray, T. (2023, December 14). *The plane truth 4.* PIRG. https://pirg.org/edfund/resources/the-plane-truth-4/

Bureau of Transportation Statistics. (2024). United States Department of Transportation. https://www.transtats.bts.gov/

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301-320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

Data Dictionary

Feature	Description	Type
PASSENGERS	Number of passengers	Numeric [integer]
NUM_FLIGHTS	Number of carrier flights	Numeric [integer]
PROP_DELAYED	Proportion of delayed flights	Numeric [float]
PROP_CARRIER_DEL	Proportion of carrier delayed flights	Numeric [float]
PROP_WEATHER_DEL	Proportion of weather delayed flights	Numeric [float]
PROP_LATEAIRCRAFT_DEL	Proportion of late aircraft delayed flights	Numeric [float]
MIN_DEP_DEL	Minimum flight delay [minutes]	Numeric [float]
Q1_DEP_DEL	First quartile flight delay [minutes]	Numeric [float]
MED_DEP_DEL	Median flight delay [minutes]	Numeric [float]
Q3_DEP_DEL	Third quartile flight delay [minutes]	Numeric [float]
MAX_DEP_DEL	Maximum flight delay [minutes]	Numeric [float]

Fig 11: Initial Monthly Aggregate Features

Feature	Description	Type
PROP_LATEAIRCRAFT_DEL	Proportion of late aircraft delayed flights	Numeric [float]
MAX_DEP_DEL_HRS	Maximum flight delay [hours]	Numeric [float]
AIRCRAFT_UTIL_RATIO	Number of flights / passengers	Numeric [float]
CARRIER_NAMEExpressJet Airlines Inc	Express Jet Airlines Indicator	Numeric [integer]
CARRIER_NAMEFrontier Airlines Inc	Frontier Jet Airlines Indicator	Numeric [integer]
CARRIER_NAMEJetBlue Airways	JetBlue Airways Indicator	Numeric [integer]
CARRIER_NAMESkyWest Airlines Inc.	SkyWest Airlines Indicator	Numeric [integer]
CARRIER_NAMESpirit Air Lines	Spirit Airlines Indicator	Numeric [integer]

Fig 12: Final Monthly Aggregate Features