TimeFlight

Time Series & Forecasting

Final Project Report

Ayushi Shah

December 1st, 2025

1. Executive Summary

Flight delays impose heavy and recurring costs on the U.S. aviation system , often exceeding $30 billion annually, yet airlines and airports still struggle to anticipate delay pressure early enough for meaningful operational planning. TimeFlight addresses this challenge by constructing a regional forecasting framework using twenty-one years of historical delay data from the U.S. Bureau of Transportation Statistics.

The dataset, containing over 400,000 monthly airport-level observations, is aggregated into five U.S. regions to study how delay minutes evolve geographically and temporally. To forecast these patterns, we apply a suite of univariate time-series models, including Double and Triple ETS, a hand-specified ARIMA(1,0,1), and both seasonal and non-seasonal Auto-ARIMA variants. Models are evaluated under multiple cutoff windows , 2009, 2016, 2020, and 2024 , to reflect different operational eras and structural changes, using MAE, MSE, R^2, and mean error as performance metrics.

The analysis reveals that delays vary sharply by region, with the South East and West experiencing the highest and most volatile delay levels, while the North East and South West remain comparatively stable. Strong seasonal patterns appear nationwide, especially summer peaks and the sharp dip during early 2020.

Across regions and time periods, Seasonal Auto-ARIMA consistently provides the most accurate and stable forecasts, particularly in high-volatility regions. Even when magnitude errors occur, the models reliably signal upcoming high-delay periods, offering practical value for staffing, scheduling, and contingency planning.

Overall, TimeFlight shows that simple and transparent univariate models, when applied at the regional level, can produce meaningful delay-risk forecasts. Regions should be modeled independently, seasonal components are essential, and the South East and West require special operational attention. Future work will incorporate exogenous drivers such as weather and traffic, expand to airport-level modeling, and explore advanced forecasting architectures capable of addressing nonlinear patterns and structural shocks.

2. Problem Context and Importance

2.1 Economic and operational impact of flight delays

Flight delays are a visible, frustrating part of air travel, but their impact extends far beyond inconvenienced passengers. Delays can:

- Consume valuable aircraft and crew time;
- Force airlines to pay for rebooking, accommodation, and crew overtime;
- Reduce airport capacity when gates, runways, and taxiways back up;
- Ripple through interconnected hub-and-spoke networks, turning a localized storm into a nationwide schedule disruption.

Industry estimates place the annual economic cost of delays in the tens of billions of dollars, when both direct and indirect effects are considered. For airlines, the difference between a slightly under-staffed operation and a well-planned one is measured in millions of dollars and customer satisfaction scores.

2.2 Why this problem and dataset were selected

We selected this problem because flight delays represent a real and persistent challenge for passengers, airlines, and regulators, making accurate forecasting both impactful and highly relevant. The U.S. Bureau of Transportation Statistics provides rich, high-quality public data, including detailed monthly delay-cause records for every major airport, enabling large-scale analysis across nearly two decades. The dataset also has a naturally interesting time-series structure: airline operations follow strong seasonal patterns , such as holiday surges and summer peaks , and experience occasional structural shocks, including economic downturns and the COVID-19 collapse in flight activity. These characteristics make flight delays an ideal domain for studying classical time-series models like ETS and ARIMA. Finally, the project aligns closely with the themes of our AI course, connecting data representation, learning, and prediction to demonstrate how AI-driven forecasting can support decision-making in complex, real-world operational systems.

2.3 Target audience

The intended audience for TimeFlight includes:

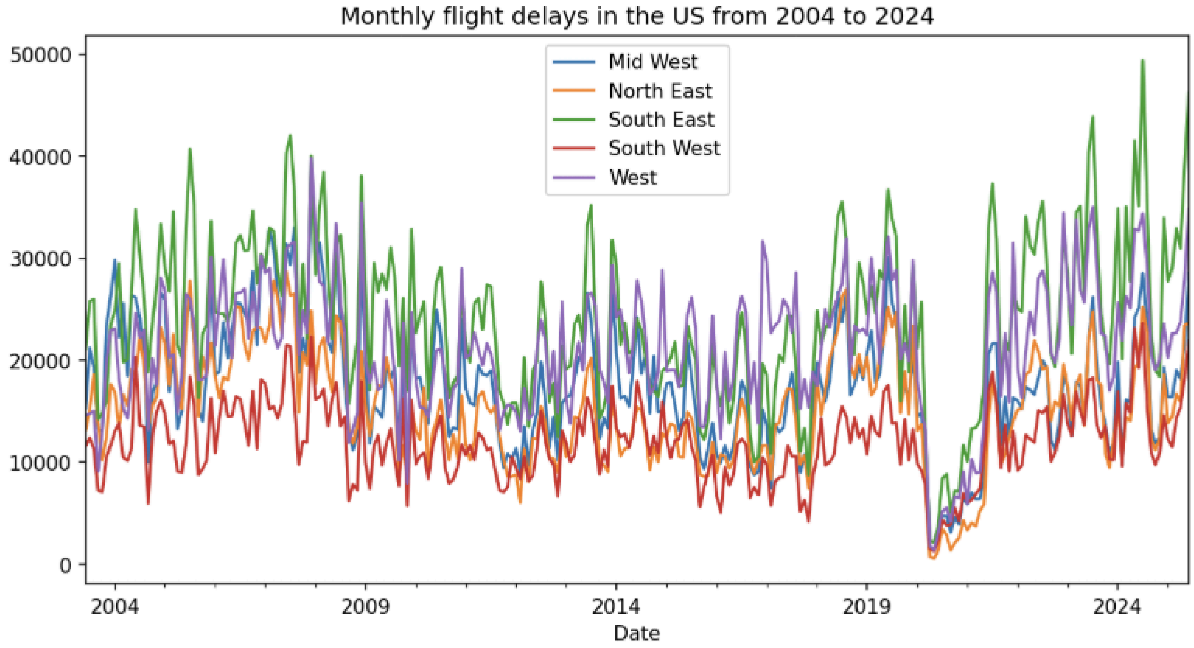
1. Airline operations and network planning teams, who must align fleet assignments, crew rotations, and schedule buffers with expected delay pressure;
2. Airport operations planners, who coordinate gate assignments, ground handling, and staffing;
3. Transportation analysts and regulators, who monitor systemic reliability and regional disparities;
4. Travel platforms and consumer-facing tools, which can surface delay-risk scores to help travelers choose routes, layover durations, or trip dates.

We assume a data-literate but non-specialist audience , people comfortable reading charts and summary statistics, but not necessarily familiar with the details of ETS or ARIMA. Accordingly, we focus on interpretable visualizations and high-level model comparisons rather than deep mathematical derivations.

3. Problem Statement and Objectives

We frame TimeFlight as a regional delay-risk forecasting problem:

> Given historical monthly arrival delay minutes for U.S. airports, aggregated into five U.S. regions, can we build simple, robust time-series models that forecast delay pressure for each region over the next 12 months and correctly identify high-risk periods?



plot 1

More formally:

- Input: A time series of monthly total arrival delay minutes for each of five regions (North East, Mid West, South East, South West, West), from 2004–2024.

- Output: For each region, a 12-month ahead forecast of total delay minutes, including a central prediction and an uncertainty band for risk planning.

Our high-level objectives are to characterize regional delay behavior through exploratory analysis , examining levels, variability, and seasonal patterns , and to build and compare a set of classical univariate time-series models across regions and cutoff regimes. We evaluate these models using consistent metrics such as MAE, MSE, R², and mean error to determine which approaches generalize best under different historical conditions.

Beyond numerical accuracy, a central goal is to translate model outputs into actionable operational recommendations, including identifying which regions should receive heightened staffing attention and when seasonal delay spikes are likely to occur. Ultimately, success is not defined solely by minimizing error metrics; rather, it is measured by whether the models deliver stable, directional signals that are practically useful, particularly in identifying high-delay months ahead of time.

4. Current Approaches and Positioning

4.1 Existing approaches to delay forecasting

In practice, flight delay forecasting draws on several major methodological families:

Operational heuristics and historical averages.
Operations teams often rely on past patterns and domain expertise (e.g., “July is storm-heavy in the South East”) to plan buffers and staffing. This approach is simple and interpretable but fails to adapt to structural shifts such as COVID-19 and offers no uncertainty estimates.

Classical time-series models.
Airlines and regulators widely use ARIMA-family models and exponential smoothing, often with seasonal components to capture recurring monthly or quarterly patterns. These models are efficient and transparent but must be retrained after major regime changes and typically focus on a single aggregate series.

Regression and machine-learning models with exogenous features.
More advanced methods incorporate weather, schedules, and network structure into regression, tree-based, or neural models. They can capture nonlinearities and richer interactions but are more complex, require extensive data, and are harder for stakeholders to interpret.

Simulation and network models.
Queuing systems and discrete-event simulations model gates, runways, and delay propagation across hub-and-spoke networks. While powerful, they demand detailed operational inputs and are less suitable for rapid, high-level regional forecasting.

4.2 How TimeFlight is differentiated

TimeFlight adopts a deliberately lightweight and regional perspective by avoiding exogenous variables such as weather or scheduling data and instead relying solely on historical delay time series to establish a lower bound on predictive power. Rather than modeling a single airport or a national aggregate, we forecast all five U.S. regions in parallel, allowing regional differences to emerge and enabling more targeted operational planning.

The approach emphasizes simple, classical models like ETS and ARIMA that are transparent, reproducible, and interpretable to non-experts, avoiding the opacity of deep learning methods. To assess robustness, we evaluate models across multiple cutoff windows , pre-2010, post-2016, post-COVID, and recent high-volatility periods , capturing structural changes that many forecasting workflows overlook. Ultimately, TimeFlight seeks to understand how effectively regional delay forecasts can be generated using classical univariate models applied to a carefully constructed regional dataset.

5. Data Description and Preparation

5.1 Raw data source

We use the Airline On-Time Performance / Cause of Delay dataset from the U.S. Bureau of Transportation Statistics (BTS).

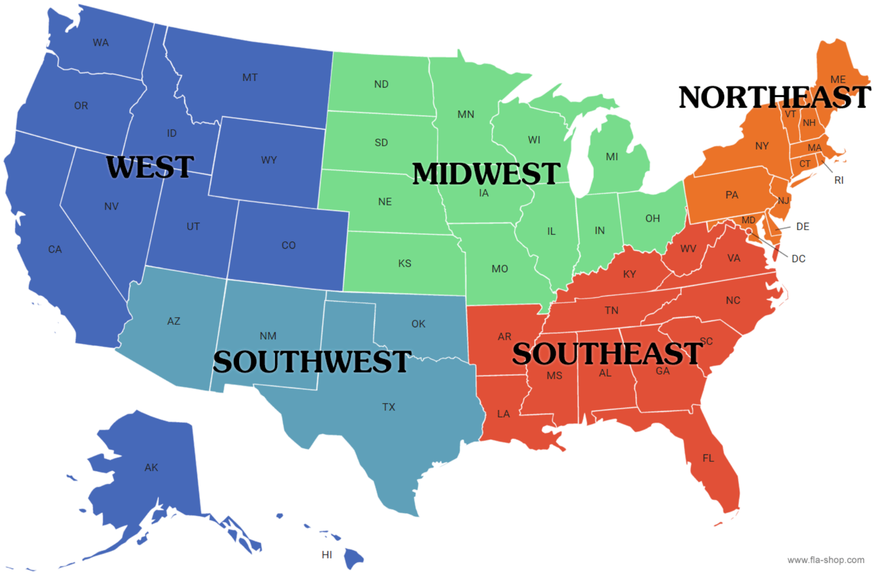
- Source: BTS TranStats – Airline On-Time Performance, Cause of Delay tables.
- Time period: 2004–2024 (21 years).
- Granularity: Each row is a single airport in a given month.
- Size: 407,721 rows and 21 variables.

Key fields include:

- Temporal identifiers: year, month;
- Airport identifiers: airport code and related metadata;
- Operational counts: total arrivals, flights;
- Delay metrics: total delay minutes, counts by cause category (e.g., weather, security, carrier).

5.2 Regional aggregation and derived dataset

To align with our regional forecasting focus, we map each airport to one of five U.S. regions: North East, Mid West, South East, South West and West



Pic 2

This mapping is done manually based on geographic location and state membership. After this step, we aggregate monthly airport-level delay metrics up to the region level, resulting in a dataset where:

- Each row corresponds to one region in one month;
- The primary target variable is the total arrival delay minutes for that region and month;
- The time index is a proper datetime object constructed from `year` and `month` for easier time-series modeling.

5.3 Data cleaning and assumptions

Several important data preparation steps are applied:

1. Datetime conversion: We convert (year, month) pairs into a single monthly datetime index, ensuring that models can interpret temporal ordering and calculate lags correctly.

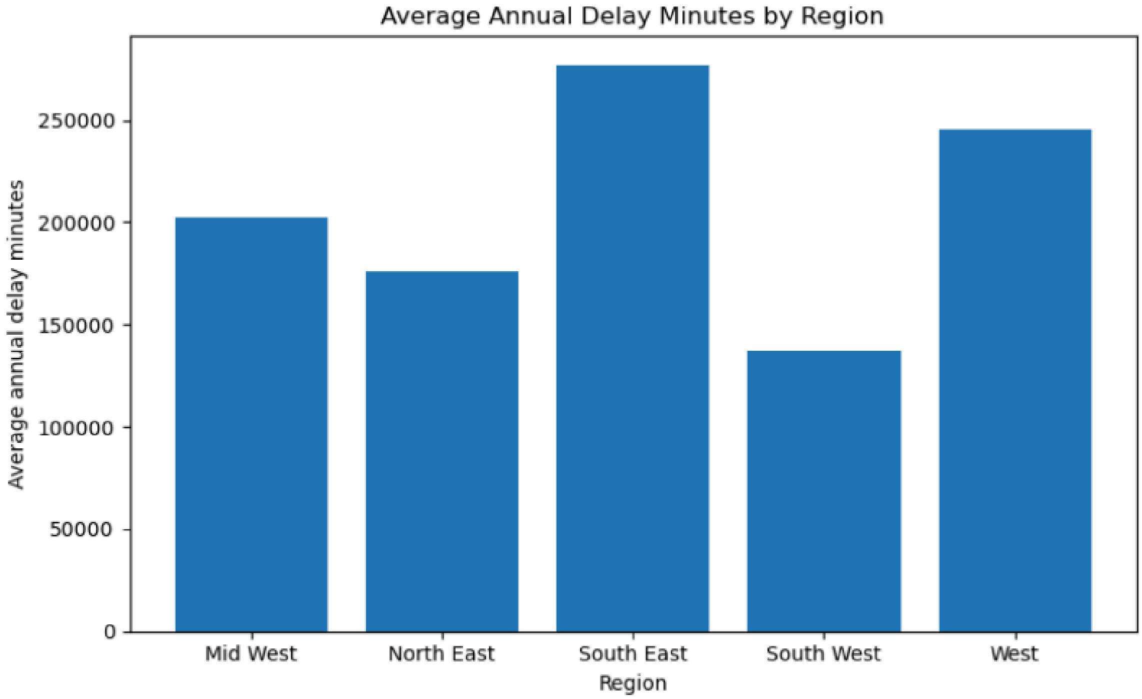
2. Definition of delay : In the BTS specification, a delay is defined as an arrival delay greater than 15 minutes, and total delay minutes reflect the aggregate of such events. We adopt this definition without modification.
3. Missing values : One advantage of this dataset is that it is effectively complete, with no missing values for the primary fields we use. A simple missing-value check confirms that delay values and key identifiers are fully populated.
4. Final modeling dataset: After cleaning and aggregation, we obtain a region-level monthly dataset, where each row represents one region and its total arrival delay minutes for that month. This becomes the input to our forecasting pipeline.

6. Exploratory Data Analysis (EDA)

EDA serves two main purposes: to understand how delay patterns differ across regions and to inform our modeling choices, particularly regarding seasonality and volatility.

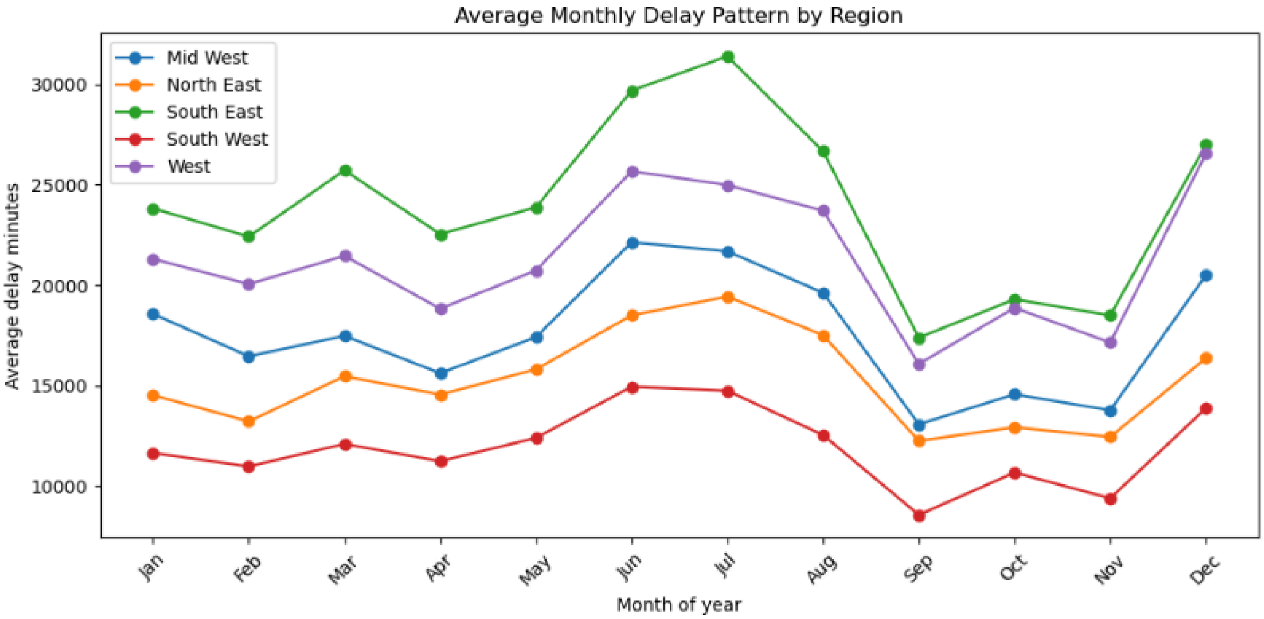
6.1 Overall distribution of delays

We first examine the distribution of monthly total delay minutes by region using histograms and boxplots. The histograms show that the South East and West have long right tails, reflecting months with severe congestion or extreme weather events, while the Mid West displays a more compact distribution with fewer extreme months. The North East and South West lie on the lower end, exhibiting fewer high-delay observations overall. Boxplots reinforce these findings: the South East has the highest median delays and the widest interquartile range, indicating substantial variability, and the West similarly shows a broad spread with multiple outliers driven by large hub airports prone to weather and traffic shocks. The Mid West falls in the mid-range, whereas the North East and South West maintain lower medians with tighter distributions. These differences suggest that regions with heavier traffic and complex hub structures tend to experience greater volatility in delays, which has direct implications for model selection and forecast difficulty.



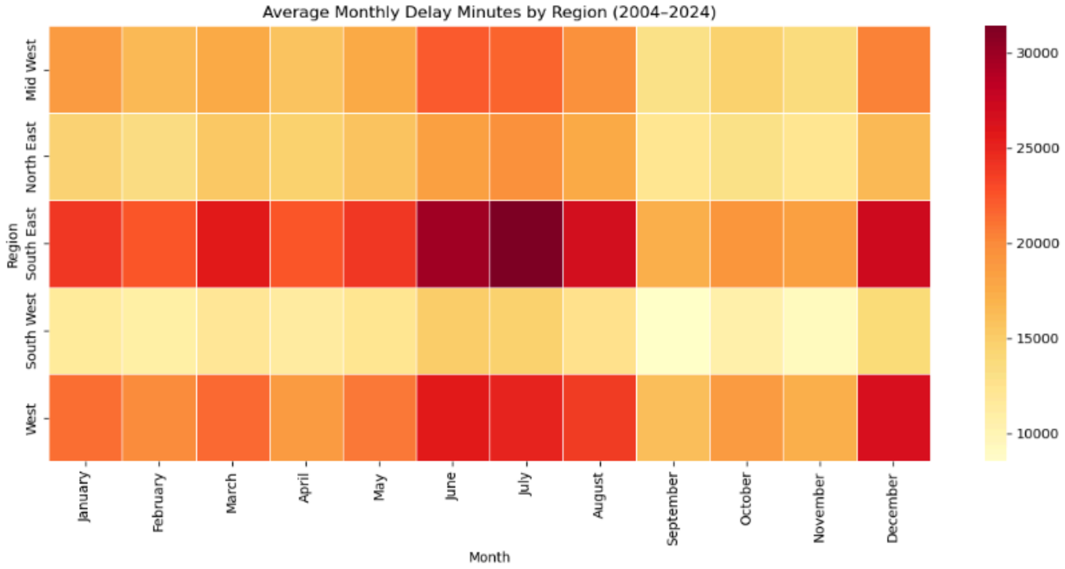
6.2 Long-term monthly delay trends

Next, we examine monthly delay trajectories from 2004 to 2024. The South East and West consistently show higher delay levels than the other regions, marked by pronounced peaks and troughs, while the North East and South West remain comparatively stable with lower overall peaks. All regions show a sharp decline in early 2020, reflecting the collapse in flight activity during the COVID-19 pandemic. These long-term patterns highlight that delays are not evenly distributed across the country, that major structural events such as COVID-19 can reshape baseline trends, and that clear seasonal cycles are present , suggesting that models with explicit seasonality are likely to perform better.



6.3 Seasonality and month-of-year effects

To quantify seasonality, we compute average delay minutes by region and month of year and visualize them using heatmaps. The results show strong summer peaks (June–August) across all regions, with the South East and West experiencing the most pronounced congestion. The North East exhibits a milder seasonal pattern, while the South West remains consistently low and stable throughout the year.



6.4 Hypotheses confirmed or revised

At the outset, we proposed three hypotheses: that regions with more traffic would experience greater volatility in delay volumes, that delays would display consistent seasonal structure across all regions, and that regions with stronger seasonality would be easier to forecast. EDA strongly supports the first two hypotheses. High-traffic regions such as the South East and West clearly show higher medians and wider spreads, with long right tails and numerous outliers, and all regions exhibit pronounced seasonal cycles with reliable summer peaks and the distinct COVID-era dip. The third hypothesis is only partially supported. While regions with strong seasonality are indeed captured more effectively by seasonal models, high volatility , such as sudden spikes caused by rare storms , still limits predictive precision and makes accurate forecasting more challenging.

7. Modeling Approach

7.1 Forecasting task and setup

We treat each region’s monthly delay series as a separate univariate time series and aim to forecast the next 12 months of total delay minutes based on historical values alone.

For each region, we:

1. Extract the historical series of monthly total delay minutes;
2. Specify one or more cutoff dates that define the end of the training period;
3. Fit multiple candidate models to the training segment;
4. Generate 12-month forecasts and compute error metrics against observed values in the forecast window;
5. Compare models across regions and cutoffs.

7.2 Models considered

We evaluate five classical time-series models, chosen to span a range of complexity and seasonal structure.

1. Double ETS (additive trend): Captures level and linear trend, but no explicit seasonality. Suitable for regions with smoother patterns or where seasonality is weak or irregular

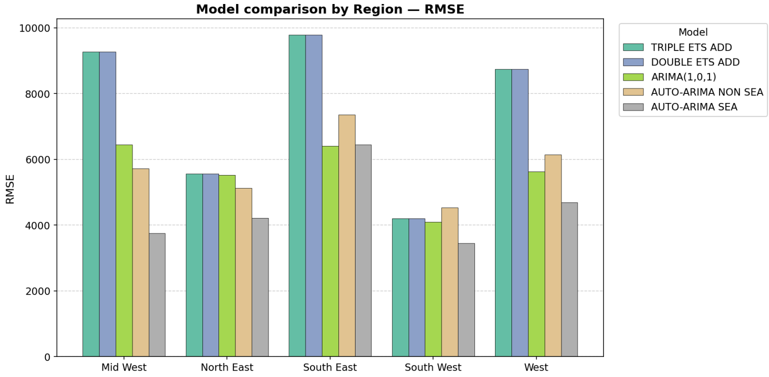
2. Triple ETS (additive trend + additive seasonality): Adds an explicit 12-month seasonal component, decomposing the series into level, trend, and seasonality. Appropriate for regions where delays rise and fall in predictable annual cycles.

3. ARIMA(1,0,1): A hand-specified non-seasonal ARIMA model capturing short-term autocorrelation via one autoregressive and one moving-average term. Serves as a simple baseline for ARIMA-family models.

4. Auto ARIMA (non-seasonal): Automatically selects ARIMA orders (p, d, q) based on information criteria, without seasonal terms. Useful when the series is stationary or seasonality is mild or captured indirectly.

5. Seasonal Auto ARIMA: Extends Auto ARIMA to include seasonal components, allowing the model to leverage monthly seasonal patterns (period = 12) explicitly. Particularly suited for regions with strong, stable seasonal cycles and structural shifts.

These models are standard in transportation and demand-forecasting applications and provide a good balance between interpretability, flexibility, and computational efficiency.



7.3 Cutoff-based forecasting design

Forecasts are sensitive to the time period used for training, especially when structural breaks occur. To assess robustness, we define four main cutoff windows:

1. 2009-01-01 (early period):
   - Focuses on pre-2010 delay patterns, capturing older operational regimes.
2. 2016-01-01 (mid-period operational change):
   - Reflects more modern post-2016 operations, after many schedule optimizations and traffic shifts.
3. 2020-01-01 (COVID-19 structural break):
   - Tests model behavior across the onset of COVID-19, when flight volumes and delay patterns changed dramatically.
4. 2024-07-01 (most recent high-volatility period):
   - Uses the smallest training window, focusing on recent years with complex, high-volatility behavior, and is arguably the most operationally realistic.

For each cutoff, we train models on data up to the cutoff date and evaluate forecasts on the subsequent months. This design reveals whether models that perform well in one historical regime continue to do so under different conditions.

8. Evaluation Methodology

8.1 Metrics

We compare models using four standard forecasting metrics:

Mean Absolute Error (MAE): Measures the average magnitude of forecast errors, regardless of direction. Easy to interpret in units of delay minutes.

Mean Squared Error (MSE): Squares errors before averaging, penalizing large mistakes more heavily. Useful when large spikes (major storms) are particularly costly.

R² (coefficient of determination): Captures the fraction of variance explained by the model. Higher R² indicates better fit.

Mean Error (ME): Reflects systematic bias (over- or under-prediction) over the forecast window.

Rather than focusing on a single metric, we look for consistent performance across MAE, MSE, and R², and we monitor ME to ensure that models are not systematically biased.

8.2 Baseline and comparison

Our evaluation is primarily relative, comparing ETS, ARIMA(1,0,1), Auto ARIMA, and Seasonal Auto ARIMA against each other. Implicitly, these models also outperform a naïve baseline that either: predicts that each month will be equal to the most recent observed value, or uses a simple seasonal naïve forecast (e.g., next July ≈ last July). We do not treat these naïve baselines as full-fledged models but use them informally to check that our methods produce meaningfully better forecasts than "no-model" strategies.

8.3 Evaluation strategy

For each region and each cutoff:

1. Split the time series into training (up to cutoff) and forecast window (after cutoff).
2. Fit all five models on the training segment.
3. Generate 12-month ahead forecasts for each model.
4. Compute MAE, MSE, R², and ME over the forecast window.
5. Compare models across metrics and identify the best-performing model for that region and cutoff.

We then examine patterns such as:

- Which models dominate across multiple regions?
- Do different models perform better in stable vs. volatile regions?
- How do results change between pre-COVID and post-COVID windows?

9. Results

9.1 Overall model performance

Across regions and cutoffs, Seasonal Auto ARIMA emerges as the strongest overall model, consistently delivering:

- Lower MAE and MSE relative to non-seasonal models;
- Higher R^2, especially in regions with strong, regular seasonality;
- Better alignment with observed peaks and troughs, including summer spikes.

ETS models provide reasonable performance in more stable regions but struggle to fully capture the complex seasonal and structural dynamics in the South East and West. Non-seasonal ARIMA and Auto ARIMA models underperform when delays show strong yearly cycles.

9.2 Regional differences

South East and West

These regions have the highest average annual delays and the broadest variability, driven by large hub airports, weather, and traffic complexity. Seasonal Auto ARIMA significantly outperforms ETS and non-seasonal ARIMA variants, as it can leverage strong seasonal structure and adapt to trend changes. In these regions, univariate models still struggle with extreme spikes caused by rare events, but they accurately flag periods (e.g., summer months) where delays are likely to be high.

Mid West and North East

These regions occupy the middle tier in terms of delay levels and variability. Seasonality is present but less extreme; both Triple ETS and Seasonal Auto ARIMA perform competitively. Because volatility is lower than in the South East and West, models are better able to track both seasonal cycles and trend shifts, resulting in relatively stable performance across cutoffs.

South West

The South West has the lowest overall delay burden, with narrower distributions and fewer extreme months.Here, simpler models (Double ETS, non-seasonal Auto ARIMA) can perform reasonably well, as the series is relatively smooth. Seasonal components still improve fit slightly, but the performance gap to Seasonal Auto ARIMA is smaller than in high-volatility regions.

9.3 Impact of cutoff windows

Evaluating models across different cutoff points reveals several patterns:

- Early cutoff (2009) and mid-period cutoff (2016):
    - Models trained on long histories (pre-2010 or including early years) can capture long-term trends but may underperform if structural changes after 2016 alter the baseline level of delays.

- COVID cutoff (2020):
    - Including the onset of COVID-19 in the training data forces models to confront a structural break. Some models misinterpret the 2020 collapse in flight activity as a new normal rather than a transient shock.

- Recent cutoff (2024-07-01):
    - The training window is shorter but more reflective of current operational regimes. Seasonal Auto ARIMA remains robust even in this high-volatility environment, whereas ETS models occasionally lag in adapting to rapid level changes.

Overall, Seasonal Auto ARIMA shows the most stable performance across all four cutoff regimes, reinforcing our conclusion that it is the best candidate for operational use.

9.4 Qualitative effectiveness

Even when absolute forecast errors are non-trivial , especially in months with extreme weather , the models provide several forms of operationally useful signal:

- Consistently marking summer months as higher risk;
- Highlighting that South East and West regions are more likely to experience severe delay spikes;
- Reflecting post-COVID increases in volatility and shifts in baseline delay levels.

This directional accuracy is valuable for planning, even when point forecasts are not perfect.

10. Effectiveness, Limitations, and Lessons Learned

10.1 Effectiveness

From an AI and analytics perspective, TimeFlight demonstrates that simple, classical time-series models can deliver useful delay-risk forecasts when supported by thoughtful data engineering (regional aggregation, seasonality analysis). Region-specific modeling is crucial; a single national model would obscure meaningful differences in level, volatility, and seasonal structure. Seasonal Auto ARIMA effectively captures recurring patterns and underlying trends, delivering strong performance across both stable and volatile regions.

10.2 Limitations

Several limitations remain:

1. Univariate models only.
   Our models rely solely on historical delay values; they do not incorporate exogenous drivers like weather, traffic volume, or schedule complexity. As a result, they cannot “see” incoming storms or holiday spikes except through patterns in historical data.

2. Regional aggregation hides local structure.
   Aggregating to the regional level improves signal-to-noise ratios and reduces dimensionality, but it hides airport-specific dynamics, such as chronic bottlenecks at individual hubs.

3. Structural shocks remain challenging.
   Events like COVID-19 produce abrupt, large-scale shifts that violate the stationary assumptions underpinning many time-series models.While models can eventually adapt, performance during the shock periods is limited.

10.3 Lessons learned

Working through this project, we learned several important lessons:

- Seasonality matters more than expected.
  Moving from non-seasonal models to explicitly seasonal Auto ARIMA produced a larger performance jump than we initially anticipated, especially in regions with strong summer peaks.

- Model choice interacts with regional volatility.
  ETS models that work well in smoother regions can fail badly in high-volatility settings, while Seasonal Auto ARIMA remains robust.

- Evaluation design is as important as model choice.
  Designing the cutoff-based evaluation to span pre-2010, post-2016, and post-COVID regimes required careful thought but yielded rich insights about model robustness over time.

- Simple models can go surprisingly far.
  Despite the popularity of complex ML architectures, well-tuned classical models plus strong EDA can provide transparent, actionable forecasts without requiring huge computational resources.

If we were to repeat the project, we would invest more time in:

- Designing baseline models explicitly (e.g., naïve seasonal, simple regression with trend) for clearer benchmarking;

- Systematically evaluating model performance under rolling forecast origins, rather than a small set of fixed cutoffs.

11. Recommendations and Future Work

11.1 Operational recommendations

Based on TimeFlight's findings, we recommend that airlines, airports, and analysts:

1. Treat each region separately: Do not assume a uniform national delay pattern. Regional differences in delay level and volatility are large and persistent.

2. Prioritize the South East and West: These regions consistently show higher and more volatile delays and should be monitored more closely. Staffing and gate planning should account for earlier and larger summer spikes in these regions.

3. Leverage seasonal forecasts for staffing and scheduling: Use seasonal Auto ARIMA forecasts as early-warning signals of high-risk months. Lock in staffing plans and contingency resources ahead of summer, when delays reliably jump.

4. Use stable regions as baselines: The North East and South West, with more stable patterns, can serve as reference baselines for detecting unusual shifts elsewhere.

11.2 Technical roadmap

Future technical enhancements for TimeFlight include:

1. Incorporate exogenous variables.
    - Integrate real-time weather data, traffic volumes, and schedule information as exogenous regressors in ARIMAX or other models.
    - This could improve responsiveness to upcoming storms and holidays.

2. Airport-level and multi-scale modeling.
    - Drill down to airport-specific time series, possibly using hierarchical or multi-level models to reconcile region- and airport-level forecasts.

3. Modeling other delay outcomes.
   - Extend the framework to forecast delay frequency, average delay per flight, or delay duration distribution, not just total delay minutes.

4. Exploring more advanced architectures.
   - Investigate Prophet, temporal convolutional networks, and transformer-based time-series models to better handle complex seasonality and regime shifts.

5. Integration into decision support tools.
   - Embed TimeFlight into operational dashboards or APIs used by airlines and travel platforms, turning regional forecasts into actionable alerts and recommendations.

12. Conclusion and Key Takeaways

The United States does not experience flight delays uniformly. Some regions face persistent, predictable surges, while others operate within more stable ranges. From both public policy and operational perspectives, this argues for region-specific reliability planning rather than a one-size-fits-all national strategy.

Our analysis shows that:

- The South East and West bear the heaviest and most volatile delay burdens, with large seasonal peaks and frequent outliers.

- The North East and South West are relatively more stable, providing baselines against which unusual behavior can be detected.

- Delays across all regions follow a consistent seasonal rhythm, with pronounced summer peaks and a clear COVID-era disruption in early 2020.

Forecasting results confirm that Seasonal Auto ARIMA, applied separately to each region, offers the strongest balance of accuracy and interpretability. These models reliably indicate high-delay months in advance, providing directional, actionable signals for staffing, gate allocation, and contingency planning. While univariate models cannot fully capture the complexity of weather and operational dynamics, they demonstrate that even simple approaches can meaningfully reduce uncertainty around future delay pressure.

Looking forward, enriching TimeFlight with exogenous variables, airport-level granularity, and more advanced architectures will likely yield further gains. But even in its current form, TimeFlight illustrates the value of combining high-quality public data, careful exploratory analysis, and classical AI methods to tackle a real-world, economically significant problem in air transportation.