Aloha Analytics: Six-month Honolulu Travel Forecast

Nicholas Consiglio, Matthew Schmitt, and Pawarisa Sears

Abstract

Our study explores predictive analytics for air traffic management at Hawaii's airports, particularly focusing on Honolulu Airport (HNL), the state's primary aviation gateway. Using time series forecasting, we analyzed historical data on Hawaii air passenger traffic to predict activity for the initial months of 2020 (January to June). Employing diverse modeling techniques including Naive, Drift, Seasonal Naive, ETS, ARIMA, and Dynamic Regression, we sought to identify the most accurate forecasting method. Through data manipulation, model creation, and selection processes, Dynamic Regression emerged as the preferred model, exhibiting the lowest RMSE of 28769.69. Diagnostic assessments revealed that the residuals of the Dynamic Regression model closely resemble white noise, indicating its effectiveness in capturing underlying data patterns. The insights gleaned from our analysis provide valuable guidance for airport authorities, aiding strategic planning and infrastructure development to accommodate the anticipated surge in passenger arrivals, thereby optimizing Hawaii's aviation infrastructure for future growth and efficiency.

Introduction

For this project report, our goal was to analyze flight data based out of the United States of America beginning in January 2000 to December 2019, where each row is represented as a flight. Our group was interested in analyzing flights going into and out of Hawaii and decided to focus on flights flying into the state. We acted as a team of analysts where we each had separate roles, Matt was the Spatial Expert Intern, Pawarisa was the Data Science Team Manager, and Nick was a Data Analyst. We had a “fake” client who was portrayed as Professor Lammers who had an interest in going on a trip to Hawaii where he asked us a series of questions that form the outline of our project. This outline consists of Exploratory Data Analysis, Spatial Visualizations and forecasting a model that can predict the number of passengers heading to Hawaii within the next six months from January 2020 to June 2020.

Methodology

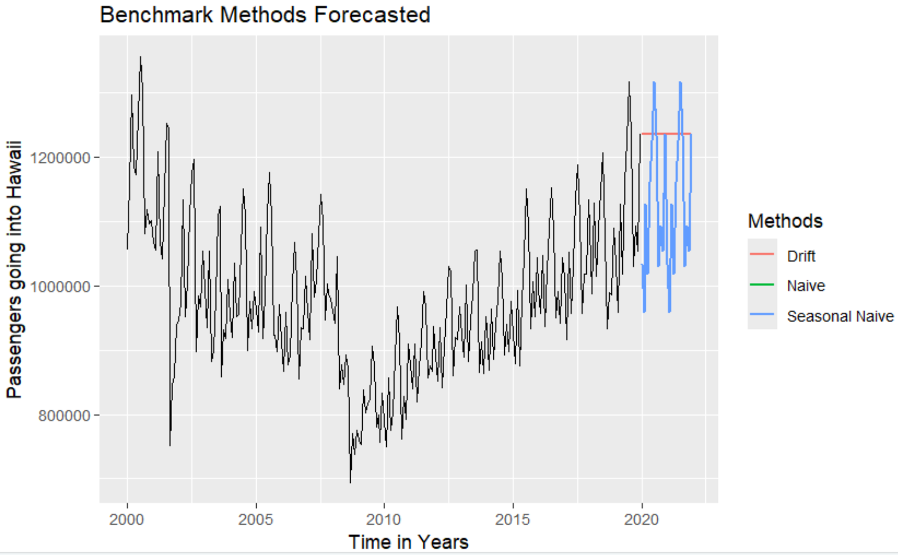
We used various R programming language data analyses to develop a predictive model to forecast airline passengers for the next six months going into the state of Hawaii. The methodology can be divided into three main steps: data manipulation, model creation, and model selection. This thorough process ensures that our predictive model is accurate and dependable. It helps airport managers make smart decisions and plan strategically based on reliable information.

Data Manipulation:

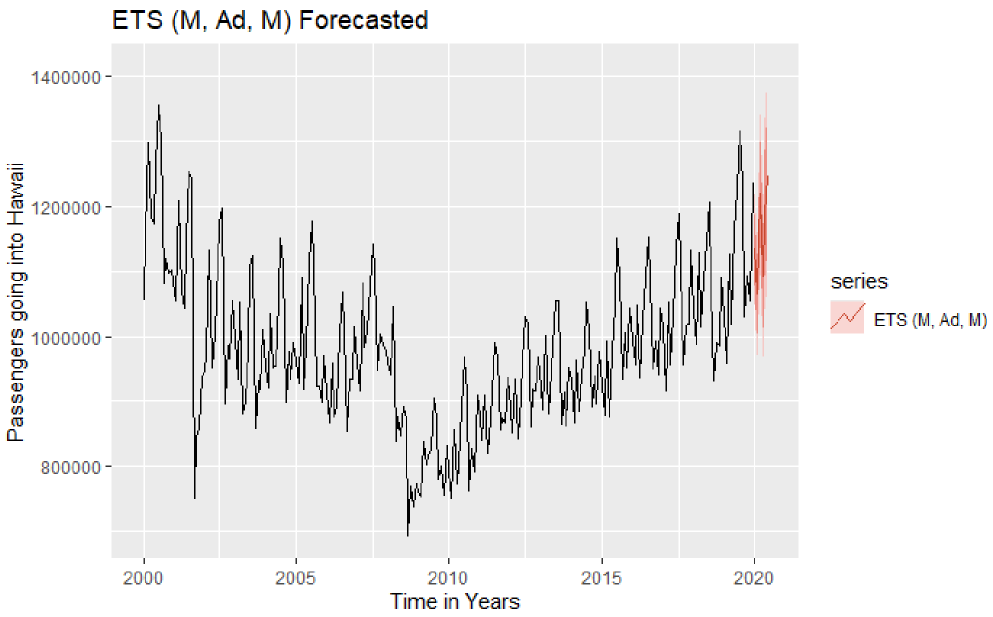
We began by importing the original dataset CSV file: US_Monthly_Air_Passengers00_19.csv, which contained U.S. flight information from January 2000 to December 2019. This dataset contained a total of 1,978,085 observations and we kept the following 12 variables: Sum_PASSENGERS, CARRIER_NAME, ORIGIN, ORIGIN_CITY_NAME, ORIGIN_STATE_NM, ORIGIN_COUNTRY_NAME, DEST, DEST_CITY_NAME, DEST_STATE_NM, DEST_COUNTRY_NAME, YEAR, and MONTH. Since we will be focusing on the flight data where the destination is Hawaii, we will filter our dataset accordingly. This new dataset contains the same variables but now has 25,828 observations. We also decided to create an additional dataset. This data set contained information about flights to the HNL airport. (HNL is the airport for Honolulu, which is the capital of Hawaii and is the biggest airport in Hawaii accounting for more than 80% of all air traffic into the state.) This new dataset contained 145 observations and variables such as flight duration in minutes and the price of a ticket for the flight to HNL. Adding this new data was fruitful for our analysis and allowed us to take unique spin on this project opening the door for a real-world situation where we as a team of analysts help a client through data analysis. When developing our forecasting models, we set aside 90% of our Hawaiian flight data to represent our training data and had the remaining 10% of our data represent the testing data.

Model Creation:

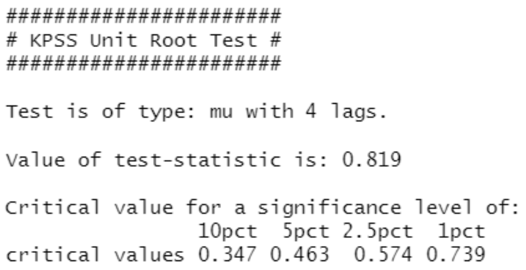
We created six types of models: forecasting methods on Naïve, Drift, Seasonal Naive, ETS, ARIMA, and a Dynamic Regression model. For each of these models we wanted to determine which model resulted in being the most accurate in forecasting the number of passengers going into Hawaii. To do this we began with benchmark models. Our benchmark models include Naive, Drift, and Seasonal Naive.



Next, we wanted to look at an ETS forecasting model. ETS stands for Error, Trend, and Seasonality. This model along with ARIMA are two of the most popular and reliable ways to forecast data. For our ETS model, we decided to let the ets() function select the best model for us. When plugging in our passenger time series, and reflecting on the results, we found the best model chosen for our time series was an (M, Ad, M) model. This model means that it contains a Multiplicative Error Term, Additive Damped Trend, and a Multiplicative Seasonality element. This model is suitable for time series data that exhibits trends that increase and decrease at a constant rate over time, with seasonal patterns and errors that are proportional to the level of the series.

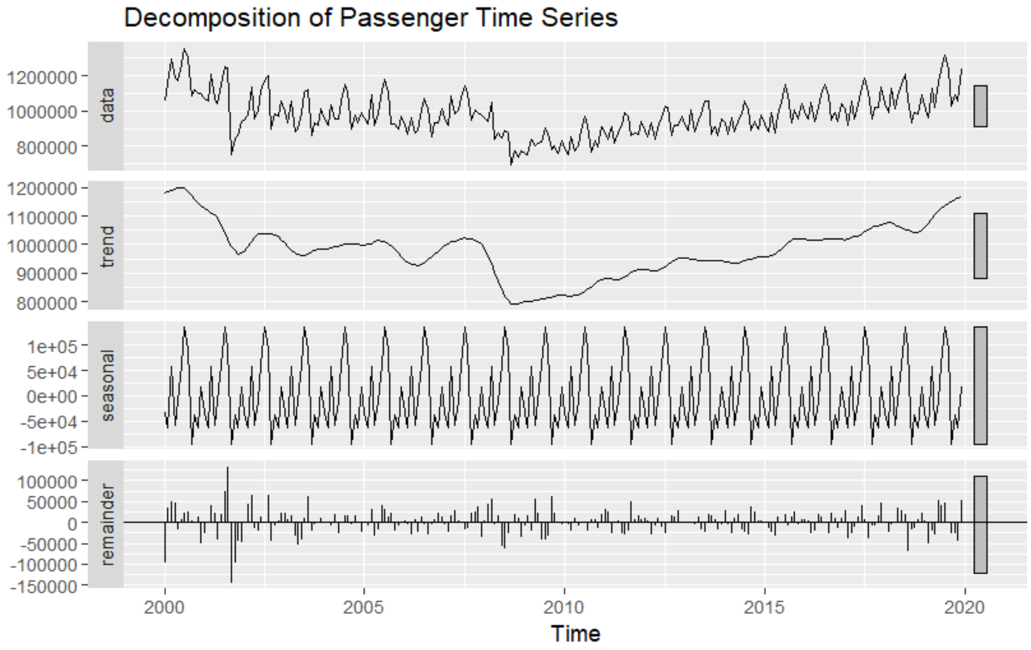


Our next model we wanted to analyze was the ARIMA model. First, we needed to check and determine if our time series was stationary. To do this we are going to use the ur.kpss() function.

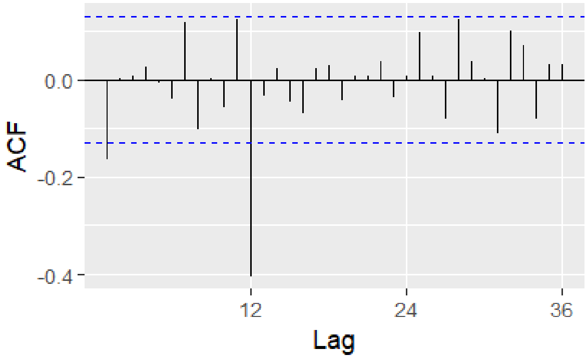


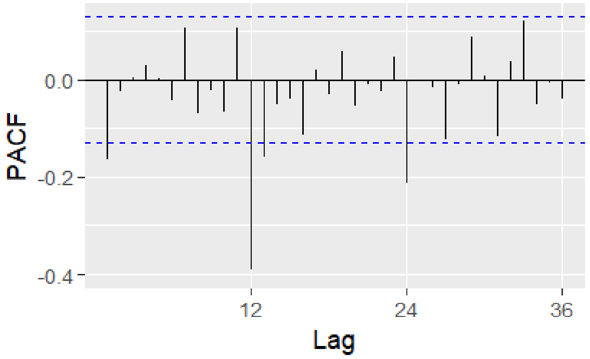
```
#######################
# KPSS Unit Root Test #
#######################

Test is of type: mu with 4 lags.

Value of test-statistic is: 0.819

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

As we can see above, the value of our Test-Statistic is 0.819. This is greater than the value of the critical value at 5% which is 0.463. This means that our time series is not stationary. To combat this, we are going to do seasonal differencing and first difference. We will conduct seasonal differencing because our time series is seasonal, and it needs to become stationary. We can tell the time series is stationary by decomposing our time series. When we decompose our time series and analyze the seasonality component, we can clearly see a pattern occurring, indicating that seasonality is present.

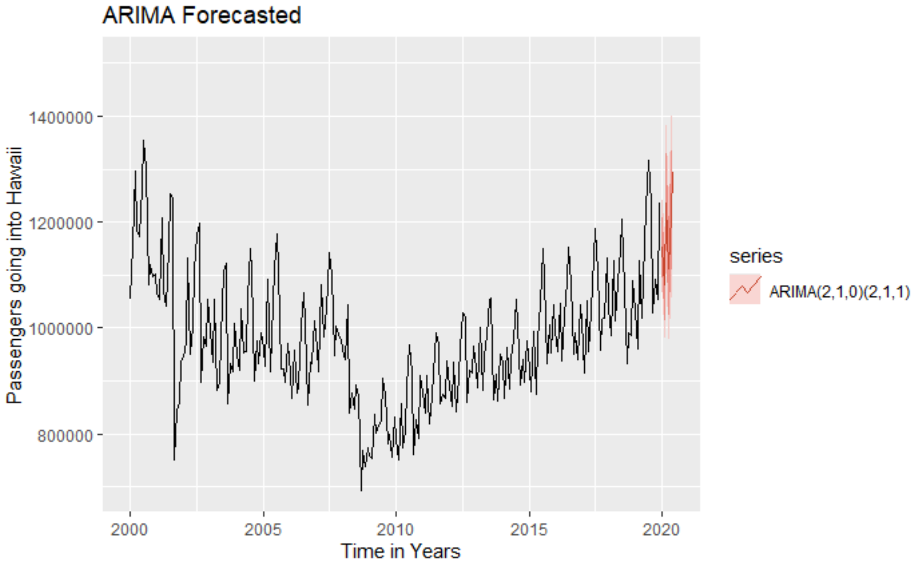


Now that we know our time series needs to be seasonally different, we next need to check how many autoregressive and moving average elements we should use in our ARIMA model. To do this we will analyze the ACF and PACF plots associated with our time series.

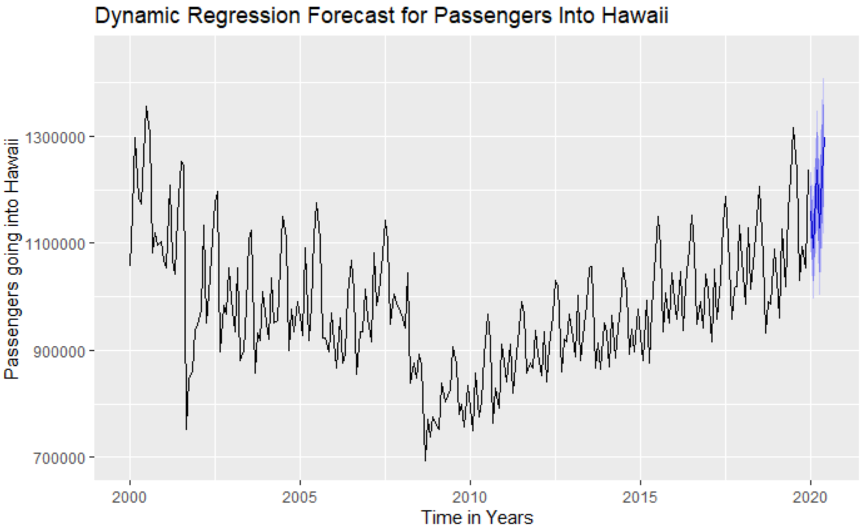




When analyzing the above ACF plot above, there is a noticeable spike for the 1st and 12th lags, resulting in one non-seasonal and one seasonal spike. For the PACF plot we can see there is a spike for the 1st, 12th, 13th, and 24th lags, resulting in a total of two non-seasonal spikes and two-seasonal spikes. This gives us an idea of what we should set for the max values for the auto.arima() function, which we will be using to forecast our time series. For our auto.arima() function we will be setting the following values d = 1, D = 1, max.p = 3, max.q = 3, max.p = 2 and max.Q = 2. We do this to let the model try and obtain the best model that it can. When plugging in our passenger time series into the auto arima function, we get a model represented as ARIMA(2,1,0)(2,1,2) as being the best-fitted model. This results in two non-seasonal autoregressive values, two seasonal autoregressive values and one seasonal moving average value.

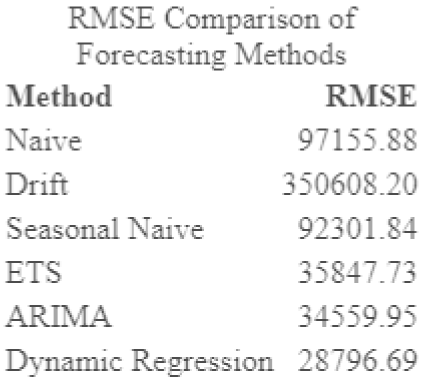


For our final modeling approach, we opted for dynamic regression. To implement this method, we indexed our monthly time series data, designating each index to represent one month, with every 12 indices representing a year. Our indexing strategy involved identifying the lowest point in the dataset and setting all values preceding this point to zero, while assigning a value of one to every data point thereafter. This indexing technique aims to prioritize recent data over historical data, thereby mitigating the influence of past periods, such as the recession, on the model's forecasts. Using the same auto.arima as above we then added our external regressor (xreg) which were the set index values mentioned above.



Model Selection:

The RMSE table below provides a comprehensive comparison of the forecasting methods employed in our analysis. By examining the RMSE values for each method, we gain insights into their predictive accuracy. The table reveals that while the benchmark methods (Naive, Drift, and Seasonal Naive) exhibit a relatively high RMSE values, indicating less accurate predictions, the ETS, ARIMA, and Dynamic Regression methods demonstrate lower RMSE values, suggesting better performance in forecasting the number of passengers arriving in Hawaii. Notably, Dynamic Regression stands out with the lowest RMSE value among the methods considered, signifying its effectiveness in capturing the underlying patterns and dynamics of the passenger data. Therefore, based on the RMSE table, we can confidently conclude that Dynamic Regression emerges as the preferred forecasting model for our analysis.

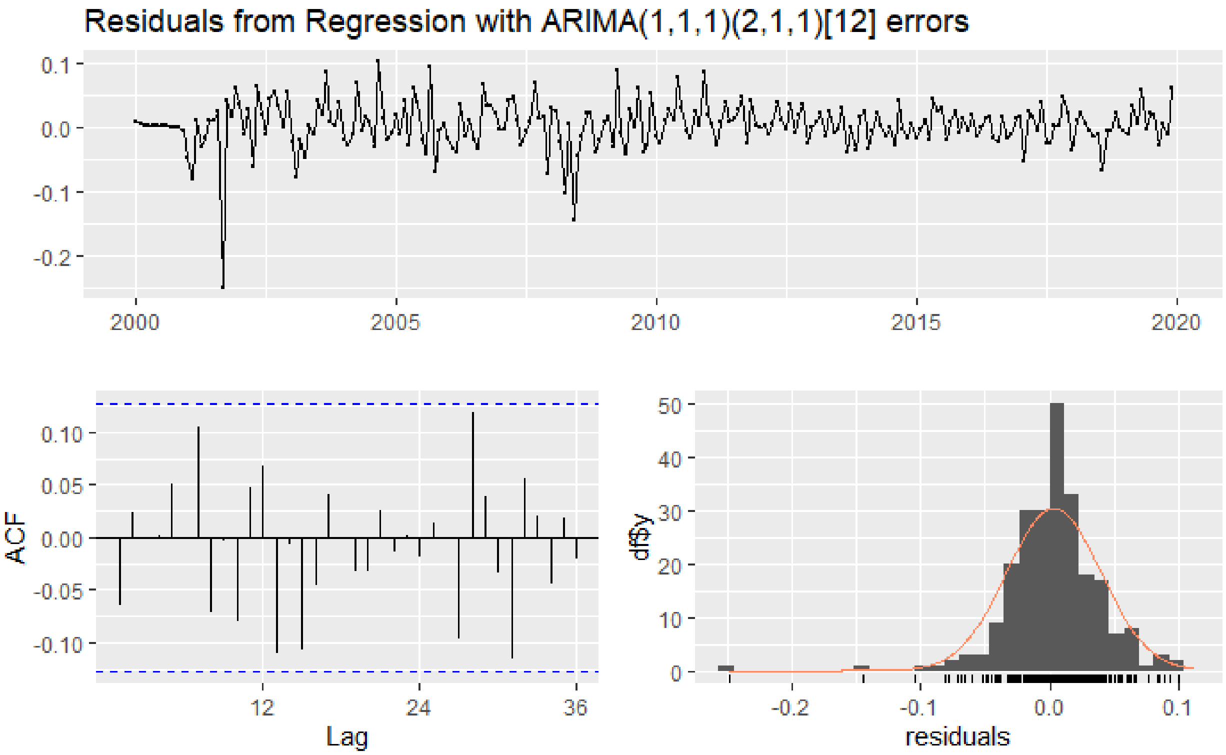


RMSE Comparison of Forecasting Methods

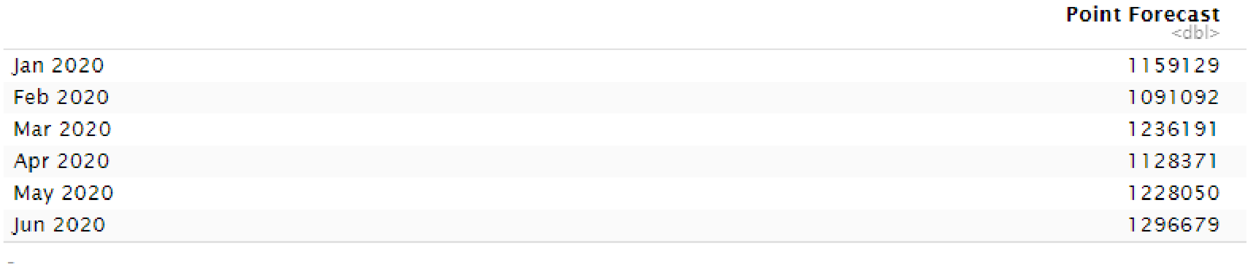
Method	RMSE
Naive	97155.88
Drift	350608.20
Seasonal Naive	92301.84
ETS	35847.73
ARIMA	34559.95
Dynamic Regression	28796.69

Analysis

Our dynamic regression model demonstrates a robust predictive capability by effectively prioritizing recent data over historical trends, mitigating the influence of past periods, such as the recession. The model's diagnostic assessment, including the Ljung-Box test, reveals that its residuals behave like white noise, indicating a lack of autocorrelation and suggesting the model's appropriateness for capturing the underlying patterns in the data.



Additionally, the model generates insightful point forecast information for the next 6 months, offering valuable insights into expected passenger arrivals. For January 2020, the forecast anticipates approximately 1,159,129 passengers, with confidence intervals at 80% ranging from 1,110,683 to 1,207,575, and at 95% spanning from 1,085,037.7 to 1,233,221. The trend continues with projections for February, March, April, May, and June 2020, indicating expected passenger counts of approximately 1,091,092, 1,236,191, 1,128,371, 1,228,050, and 1,296,679, respectively, each accompanied by corresponding confidence intervals at 80% and 95% levels.



	Point Forecast <dbl>
Jan 2020	1159129
Feb 2020	1091092
Mar 2020	1236191
Apr 2020	1128371
May 2020	1228050
Jun 2020	1296679

Conclusion

In summary, our flight project evaluated various forecasting models to predict passenger arrivals in Hawaii. We assessed each model's performance using cross validation, RMSE and diagnostic tests like residual analysis to ensure reliability. Models demonstrating white noise-like residuals, such as ARIMA and dynamic regression, were favored for their ability to capture data variability effectively. Ultimately, dynamic regression emerged as the optimal choice because of its lowest RMSE score of 28769.69. These findings suggest that employing more sophisticated forecasting techniques can lead to more accurate predictions, contributing to better decision-making processes in managing passenger traffic coming to Hawaii. With this information the airports of Hawaii could suggest expanding other airports to allow for more air travel to account for more passengers coming to one airport, as well as the potential to expand the currently heavy trafficked HNL airport.

Future Work

Exploring analogous tourist destinations to Hawaii presents an avenue for future research. Analyzing air traffic data from regions with similar characteristics, such as prominent tourist attractions and island-based geography, could provide valuable insights into passenger travel patterns and trends. Understanding how external factors, such as economic downturns like the recession, impact travel to these destinations would be particularly insightful. Investigating whether these locations experienced similar fluctuations in air passenger traffic during economic downturns could shed light on the broader effects of economic conditions on tourism. Moreover, examining how various forecasting models perform in predicting passenger arrivals to these destinations under different economic scenarios could enhance our understanding of the dynamics of air travel demand in tourism-dependent regions.