CUSP GX-7033 Final Project
Contrasting and Predicting Commercial Flight Characteristics Pre- and Post-COVID-19 Pandemic

Research Report

Zhexuan Tang (zt2339), Ruixin Gan(rg4743), Zheyan Chen (zc3053),
Yin Wang(yw7422), Weihao Xia(wx2233)

Abstract

The research investigates the global commercial aviation dynamics before, during, and after the COVID-19 pandemic. Utilizing a comprehensive dataset from the OpenSky Network, the research employs time series analysis, clustering, and Gaussian Prediction models to analyze shifts in aircraft types and flight patterns. The time series analysis reveals that the industry's growth, lockdown, and recovery phases correlate closely with travel seasonality. Clustering analysis indicates shifts in aircraft types and route preferences. The Gaussian Prediction models provide insights into future trends, suggesting varied recovery paths for different aircraft types and routes. This analysis helps understand the industry's resilience status and adaptive strategies during pandemic disruptions, providing a reference for future growth and operational optimization.

Introduction

In recent years, the dynamics of the global commercial airline industry have undergone significant shifts of continuous growth and sudden declines. Shortly before COVID-19 pandemic, the industry was reaching new heights, with 2018-2019 marking a period of robust expansion in air travel. Airlines were not only increasing their fleets but also adding new routes to accommodate the growing demand for passenger services across various continents. (Mutzabaugh, 2019 & Morris, 2018) However, the COVID-19 pandemic by the World Health Organization abruptly halted this growth trajectory (nfid.org, 2019), as the plethora of travel restrictions and stringent social distancing measures lead to a steep decline in flight operations worldwide.

As the industry has started to recover, it has also begun to evolve in response to the new post-pandemic realities. Due to the shifting commercial landscape, there might have been changes in the types of aircraft used for certain origin-destination pairs, reflecting shifts in airline strategies and passenger preferences. This adjustment suggests a strategic realignment by airlines to optimize operations and adapt to the altered market conditions. The potential contrast in ridership trends, the types of planes used, and the popularity of specific routes before and after the pandemic underscores the need for a comprehensive analysis. The project aims to delve into these aspects, providing a detailed examination of the shifts in the global commercial airline industry brought about by the COVID-19 pandemic and its aftermath. By analyzing these changes, we can gain insights into the resilience and adaptability of this vital industry.

Research Question and Hypothesis

Our research question is as follows: How did the respective ridership differ in these flight clusters in the three timeframes of the pandemic, and how have the past ridership recovery trends predict future ridership recovery in these clusters?

Our hypothesis to these questions are as follows:

1. Narrow bodied aircrafts will remain as the most predominant plane type in all examined timeframes due to its wider applicability, but post-pandemic would see more of them composing of total plane types flown.
2. Post-pandemic recovery would continue in a similar manner to how the pre-pandemic ridership grew, but would not recover in absolute value in the timeframe we are observing.

Literature Review

Time series analysis is an important part of historical flight data analysis and recovery of COVID-19-affected flight traffic forecast. In the analysis of a single airport flight, Schneider and Chen (2020) applied ARIMA and Holt-Winters exponential smoothing methods, highlighting the importance of stationarity and seasonality. For the recovery of flight volumes after the pandemic, Barczak et al. (2022) used seasonal indicators and Fourier spectral analysis for EU airports' passenger traffic; Tolcha (2023) utilized intervention analysis and SARIMAX to analyze monthly time-series data of flight frequencies in Africa; Xu et al. (2024).

Clustering is a common and basic method as part of model predicting flight demand, Güvercin et al. (2004) adopt a cluster demand forecasting method to forecast flight demand from historical booking data. Yi Gao (2022) used TSA passenger numbers to measure air travel demand at U.S. airports during the pandemic, employing a k-shape clustering algorithm and dynamic time warping.

Machine learning research in aviation primarily focuses on prediction and network analysis, especially flight fares, delays, and travel demand. Fewer studies target flight volume forecasting, particularly pre- and post-event comparisons. Firat et al. (2021) utilized a comprehensive dataset with various models (ANN, LR, GB, RF) to assess pre-COVID flight capacities. Ehsani et al. (2024) compared traditional time series and deep learning models (CNN, ConvLSTM, DeepShallow Network) using seasonality, traffic data, and fare information to forecast passenger traffic, evaluating performance with the MSE metric.

Methodology and Data Sources

Data Sources

Our approach starts from utilizing an extensive dataset selected through several similar open-source platforms. The primary data source for this study was obtained from the OpenSky Network, a reputable portal providing ICAO-coded aircraft and flight data. The specific dataset can be accessed at https://opensky-network.org/data/datasets#d4. The downloaded dataset spans a critical period from December 31, 2018, to December 31, 2022, encapsulating the pre-pandemic growth phase, the collapse during the pandemic, and the subsequent recovery phase. The raw dataset retrieved from OpenSky Network contains the following: the identifier of the flight displayed on ATC screens(callsign); the commercial number of the flight; the transponder unique identification number(icao24); the aircraft tail number; the aircraft model type; a four-letter code for the origin and destination airport of the flight; the UTC timestamp of the first message received by the OpenSky Network; the UTC timestamp of the last message received by the OpenSky Network; the UTC day of the last message received by the OpenSky Network; the first and last detected position of the aircraft(latitude, longitude, altitude).

Data Cleaning

The raw data was recorded in 48 separate files based on monthly flight data collection. We removed unneeded details such as commercial number, transponder unique identification number and the aircraft tail number in each file, and then merged the files and removed the records with insufficient data. After the initial data cleaning process, the dataset needs to be filtered again to include only commercial airlines and airports serviced by the latest in-service jets to ensure no irrelevant data slips through in data analysis. Since the data is coded by ICAO conventions, we approached this step of the data clean process from three angles: ICAO airline call signs, ICAO aircraft type codes, and ICAO airports destination codes.

First, ICAO based flight callsigns for commercial civil aviation services start with three capitalized letters and three to four numbers signifying the flight number (the most common interface with passengers compared to the two digit IATA callsign in passenger terminals). This operation is performed with two python regular expressions (Regex): one for the first three letters and the second for the next three/four digits.

The next filter for aircraft type code is similarly applied, as the typecode signifies most importantly the manufacturer of the aircraft. For the sake of maintaining as much commercial aircraft type data as possible, we selected the filters based on the these five major global aircraft manufacturers with widespread adoption were selected: Airbus with their A300 series jets, Boeing with their B700 series jets, Bombardier with their A220 (formerly CS300 before sale of the IP from Bombardier to Airbus) and numerous turboprops, De Havilland with their turboprops (Dash series), Embraer with their E1 series of regional aircrafts and Comac with their ARJ series of regional aircrafts (C919 did not enter revenue service until 2023, which was out of our data range). This filter is created with a simpler string identifier "str.startswith" and the corresponding identifiers were "A3", "B7", "BCS", "DH", "E1", "CRJ", respectively.

The last filter of origin and destination airport codes were similar in vein with the previous two filters. Commercial airports that operate regular passenger services will have ICAO codes in the form of 4 letters and 4 letters only, which means another simple regular expression filter will suffice for both the origin and destination columns. After applying the three filters one after another, we now have a cleaned dataset ready for analysis. In a separate but related note, we have normalized each flight's carrying capacity to account for ridership in these flights, as each airline has different seating arrangements that results in varying levels of passenger capacity that were calculated and built to order by the aforementioned aircraft manufacturers to be the most economical and financially lucrative for airlines' daily operations. Hence, to account for the carrying capacity, an arbitrary 90% occupied rate and an 65% occupied rate is imposed on an average capacity number for a specific type of aircraft class in the typical researched arrangement. Our ICAO data also account for stratified subcategories for more popular aircrafts (such as Airbus A320 subcategories and Boeing 737 subcategories among others present in the data), and these subclasses were considered to have the same carrying capacity to simply calculation and help to create more interpretable clustering and Gaussian Processing models.

Methodology

We first start with a time series analysis to observe visually the general trends of the ebb and flow of the flight growth and whether the data corresponds with seasonality of the travel patterns and demands and whether the datapoint reflects critical junctions in travel behaviors such as global holidays. The time series analysis also serves as the starting point to divide the data into three distinct sections - pre pandemic growth, pandemic lockdown, and post pandemic recovery, which will serve as basis for the latter in this research.

The clustering analysis aims to categorize and analyze the changes in aircraft utilization before, during, and after the pandemic, crucial for identifying trends and shifts in aircraft preferences

and airline operational strategies. This involved classifying aircraft into three main types based on their body: the conventional classification of narrowbody and widebody, as well as medium (which serves as an intermediary between the first two categories). This step was. Although in itself a descriptive analysis method, we hope to examine whether specific clusters of aircraft type used between popular routes have changed before and after pandemic lockdown. Our original filtered data contained many sub-variants of most popular commercial aircraft such as "A320-214", "B739", "E175", and for clustering purposes, we have counted all plane types of the same sub-variants to be of the same aircraft, so that when we cluster the plane types the data will be representative of the larger general plane series.
Categorized by plane size into 'narrow', 'mid', 'wide'.

After aggregating total flight numbers for each general aircraft type, we use silhouette score and elbow method to find the best k-means cluster count for both pandemic and post-pandemic datasets, clustering flights by aircraft type and size. After clustering, we'll apply a Gaussian Prediction model to both pre-pandemic growth and post-pandemic recovery time series data. Then, we'll conduct a stratified analysis by aircraft type. This predictive model provides valuable insights into how different segments of the market might recover at divergent rates. Furthermore, we aim to compare the predicted post-pandemic recovery trajectory produced by Gaussian Prediction with the assumed pre-pandemic growth trajectory using the same dataset based on both observed pre- and post-pandemic trends, taking into account varying plane types and originating continents. To initialize the Gaussian Processing models, we first decomposed the all-time series into their constituent components, including trend, seasonality, and residual (or noise) to develop insight into training, testing and eventually choosing the optimal periodic kernel. The exclusion of anomalies is based on the residual mean ± 3x standard deviation. After anomaly exclusion and time series decomposition, different Gaussian process kernels are combined to fit on pre-pandemic and post-pandemic time series of all flights' daily data and flight's daily data by their body type.

Descriptive Analysis

The already processed flight data is first utilized to categorize the flight data into before covid, during covid and after covid time periods based on three time periods: 2019-1-1 to 2019-12-31, 2020-1-1 to 2021-12-31 and 2022-1-1 to 2022-12-31. The total number of passengers carried by the flights during the three epidemic time periods was then calculated. It can be calculated by multiplying the maximum capacity of different aircraft types by the load factor of the three time periods of the epidemic to find out the number of passengers of different aircraft types and then summarizing them according to the flights. Next, we calculated and sorted the total number of flights to the original destination. place. The top 20, top 20 to 50 and top 50 original places are filtered out and then the flights departing from these original places are filtered out based on these data. Finally, we calculated the total number of passengers carried by these flights and used the number to divide the total number of passengers during the three outbreak periods. The calculated passenger number for each top range over the three covid time periods is shown in Figure 0.1. And the percentage share is shown in Figure 0.2.

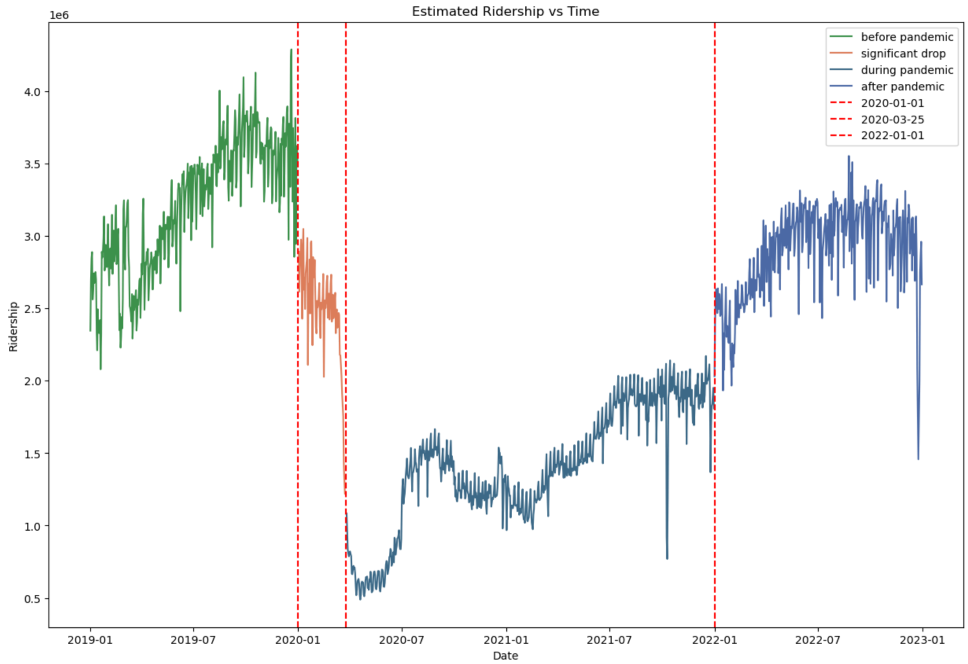


Figure 1.1 Ridership Time Series

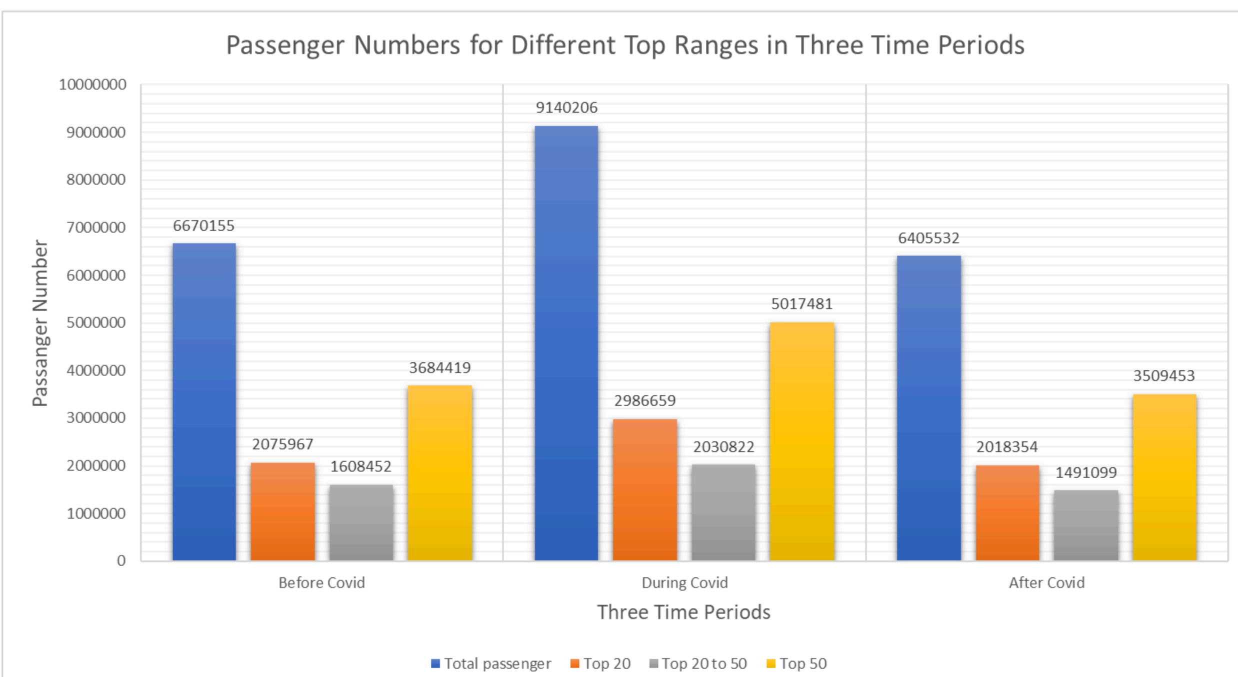


Figure 1.2.1 Passenger Numbers for Different Top Ranges in Three Time Periods

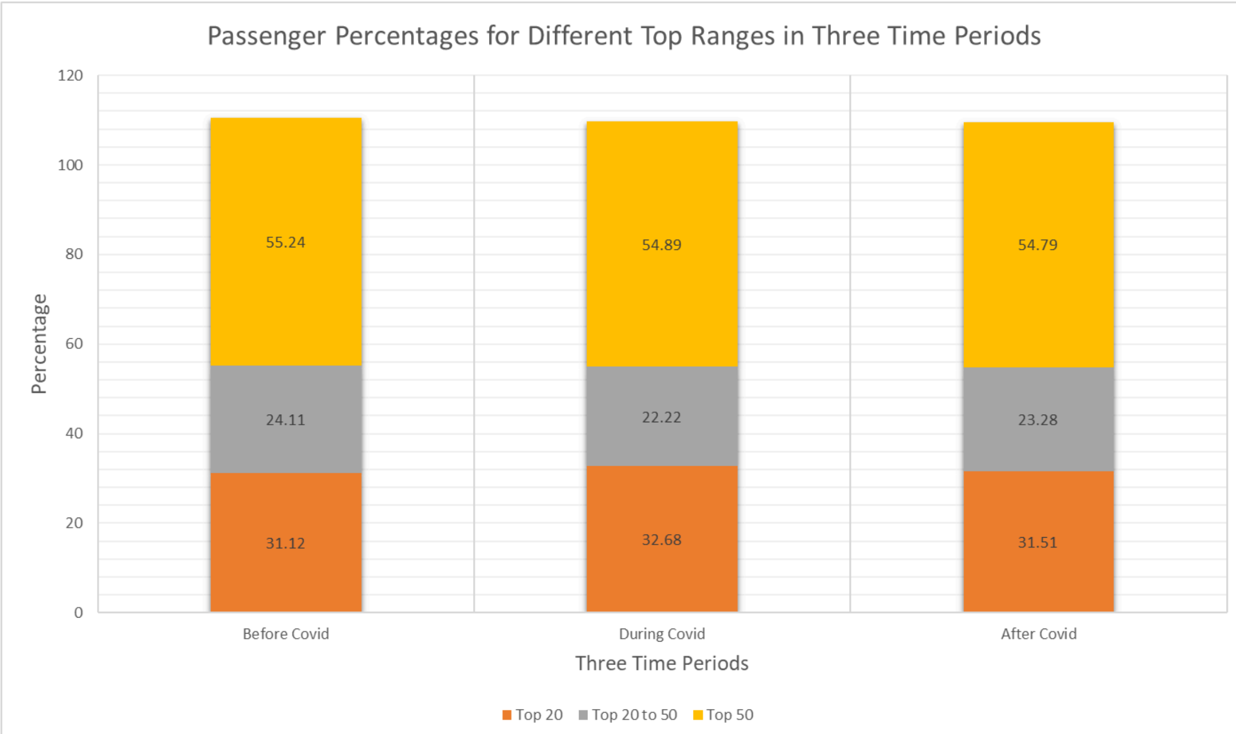


Figure 1.2.2 Passenger Percentages for Different Top Ranges in Three Time Periods

Pre-Pandemic Clustering Analysis

The silhouette score is highest for 2 clusters, suggesting good separation and cohesion, yet 3 clusters provide more detailed data segmentation with acceptable cluster definition. At the same time, the Elbow Method that plots the sum of squared distances within clusters (WCSS) against the number of clusters, indicates a sharp decrease in WCSS up to 3 clusters. This suggests that 3 clusters are optimal, as further increases yield little improvement in data modeling, reinforcing the choice of 3 as the optimal number of clusters.

	size	cluster	count
0	mid	0	4
1	narrow	0	29
2	narrow	1	3
3	narrow	2	1
4	wide	0	6

Figure 2.1.1 Number of Aircraft Type Within Cluster and Size (pre-pandemic)

	cluster	type	flight_count
13	0	B77	331943
40	1	A320	1374012
42	2	B73	2440608

Figure 2.1.2 Most Common Aircraft Type of Clusters (pre-pandemic)

Based on the Figure 2.1.2 above, Cluster 0 contains the fewest total annual flights and is primarily dominated by a significant number of narrow-bodied aircraft, which appear to be the most prevalent type across all clusters. This cluster also features a small percentage of both mid-sized and wide-bodied aircraft, hinting at a diverse but predominantly narrow-bodied fleet composition. Cluster 1, exhibits a median level of annual flight activity, and is exclusively composed of three narrow-bodied aircraft. Lastly, Cluster 2 stands out with the highest annual flight count, and interestingly, it includes only one type of aircraft, the Boeing 737 series.

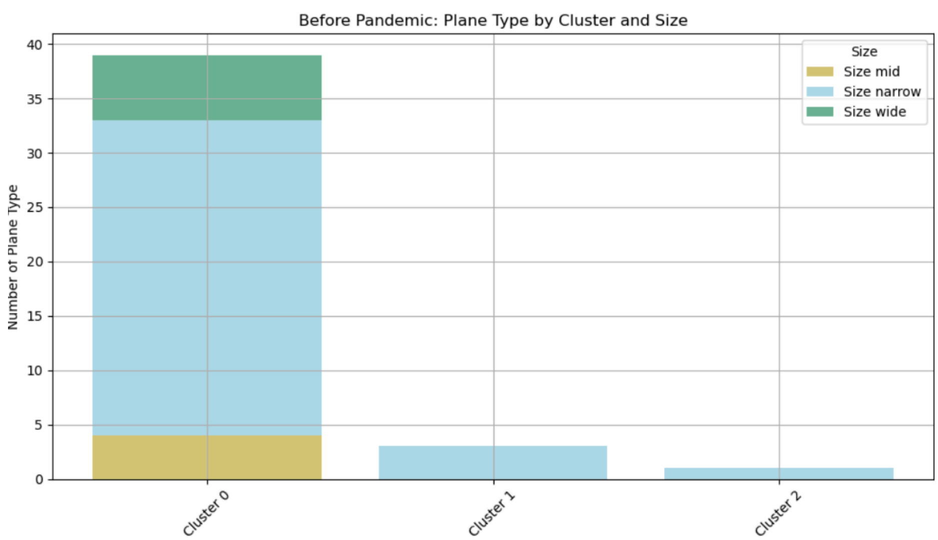


Figure 2.1.3 Total Number of Plane Type Within Cluster and Size (pre-pandemic)

Shown In Figure 1.1.3, Cluster 1's dominance in flight count suggests that operations involving narrow aircraft are crucial, possibly forming the backbone of the fleet's utilization. Analyzing the specific routes and flight frequencies within this cluster could provide insights into core revenue-generating operations. Cluster 2's lower counts could be subject to review for efficiency improvements or might represent niche markets or specific operational needs.

Mid-Pandemic Clustering Analysis

Many routes were either temporarily suspended or operated on a limited schedule to comply with government regulations and to match the reduced passenger loads. (Bouwer, 2022). This operational adjustment was crucial to mitigate financial losses during periods of low demand.

Due to the limited flight schedules during the pandemic, the number of flights operated by major scheduled narrow body aircraft significantly decreased. Looking at the silhouette score and elbow plot, the silhouette score suggests 2 clusters as optimal due to the highest average score indicating strong separation and cohesion and the Elbow method suggests either 2 or 3 clusters, due to significant bend at 3 clusters, indicating diminishing returns on variance explained beyond this point. Considering clustering different distinct groups and more granularity, we remain k = 3 as the optimal clusters.

	size	cluster	count
0	mid	0	4
1	narrow	0	29
2	narrow	1	1
3	narrow	2	3
4	wide	0	5

Figure 2.2.1 Number of Aircraft Type Within Cluster and Size (Mid-pandemic)

	cluster	type	flight_count
11	0	B77	220557
38	1	B73	1559583
40	2	A320	641611

Figure 2.2.2 Most Common Aircraft Type of Clusters (Mid pandemic)

During the pandemic, the flight activity across different aircraft types showed noticeable variations across three clusters:

Cluster 0 saw fewer flights overall, with narrow-body aircraft featuring prominently, though their numbers were significantly lower than those in Cluster 2. Wide-body aircraft were also present but had more flights than mid-sized aircraft, which recorded the lowest flight counts, indicating their limited use or specialization for certain routes.

Cluster 1 experienced a high number of flights, dominated by narrow-body aircraft. This suggests a strategic preference for using more fuel-efficient or appropriately-sized aircraft to meet the reduced demand during the pandemic, indicating a shift in operational strategy to adapt to changing conditions.

Cluster 2 had a medium number of flights, with a strong focus on narrow-body aircraft, particularly models like the A320 and A31. These aircraft are typically used for a mix of short to medium-haul flights, highlighting their continued use despite overall reduced flight activity.

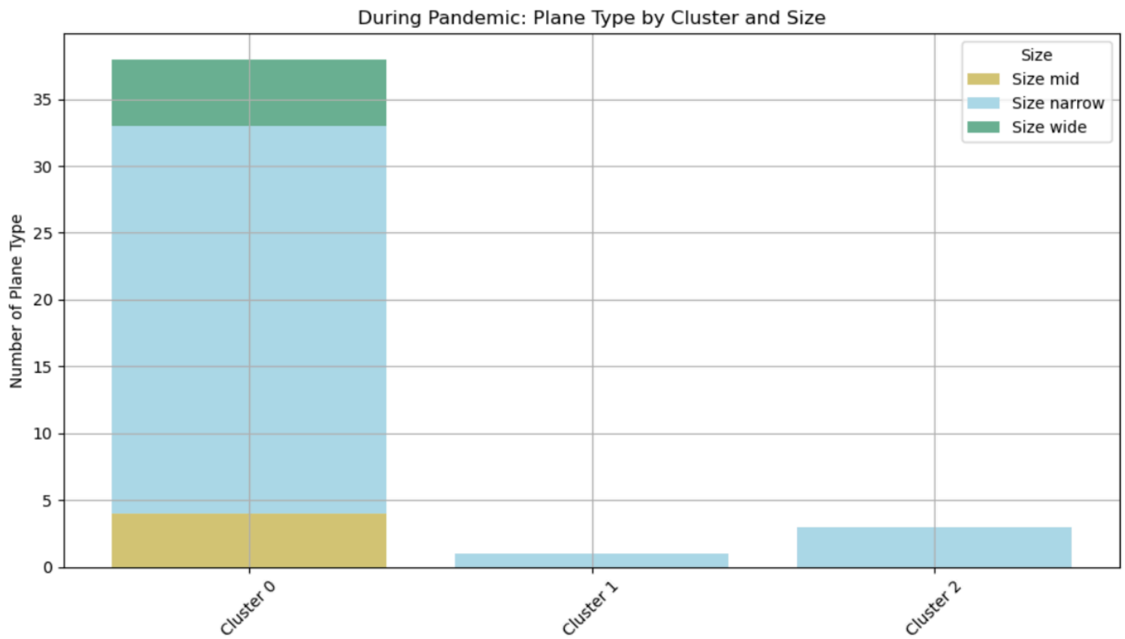


Figure 2.2.3 Total Number of Plane Type Within Cluster and Size (During pandemic)

During the pandemic, narrow body aircrafts were still significantly utilized, adjusting to the reduced passenger demand and potentially maintaining essential services. Conversely, mid-size and wide-body aircraft, which are typically employed on longer and international routes, faced more severe impacts due to travel restrictions.

Post Pandemic Clustering Analysis

After the pandemic, commercial aviation was gradually recovering from a release of government regulations and to match the increasing passenger loads of tourist travel. In analysis,The average within-cluster sum of squares decreases substantially, and the percentage of variance explained demonstrates a slight plateau after reaching 3 clusters. Consequently, 3 clusters remains advantageous, as it ensures a good balance between detailed data segmentation and manageable complexity.

	size	cluster	count
0	mid	2	4
1	narrow	0	3
2	narrow	1	1
3	narrow	2	30
4	wide	2	5

Figure 2.3.1 Number of Aircraft Type Within Size and Cluster (Post-pandemic)

	cluster	type	annual_flight
1	0	A320	892100.0
3	1	B73	2097400.0
23	2	CRJ9	263245.0

Figure 2.3.2 Most Common Aircraft Type Within each Cluster (Post-pandemic)

Cluster 0 Medium Annual Number of flights:
This cluster is dominated by narrow-body aircraft, with the Airbus A320 and its variants (A321) having a significant presence, reflecting perhaps routine short to medium-haul flights.The annual flight count of 892,100 for A320 type in Cluster 0, which suggests a robust recovery of these routes after the pandemic.

Cluster 1 (B737) Higher Annual Number of flights:
Cluster 1 prominently features the B737, with an annual flight count of 2,097,400, indicating a high volume of flights. Given that the B737 also typically serves short to medium-haul routes, this further supports a strong recovery in these sectors.

Cluster 2 Lower Annual Number of Flights:
Cluster 2 is notably diverse but has a large proportion of regional and smaller narrow-body and medium-body aircraft like the CRJ9, B75, and B76. This cluster likely represents routes or operations where smaller/more economic and well-tested aircraft are optimal, including regional or less dense routes. The CRJ900 (CRJ9) in Cluster 2, which is a regional jet, shows an annual flight count of 263,245. This suggests that domestic travel has also recovered significantly post-pandemic, though not to the extent of the more dominant narrow-body sectors in Clusters 0 and 1.

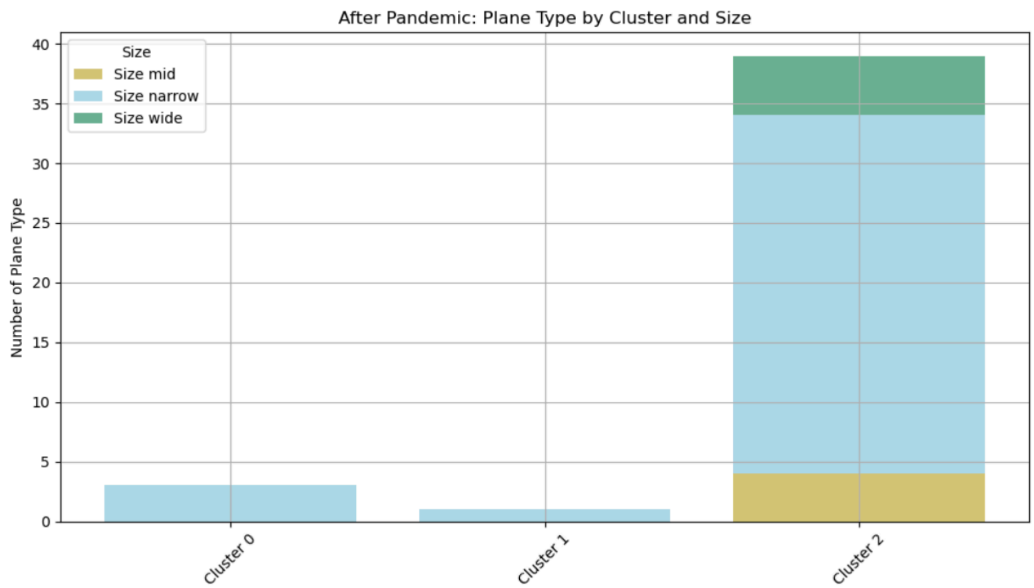


Figure 2.3.3 Total Number of Plane Type Within Cluster and Size (Post-pandemic)

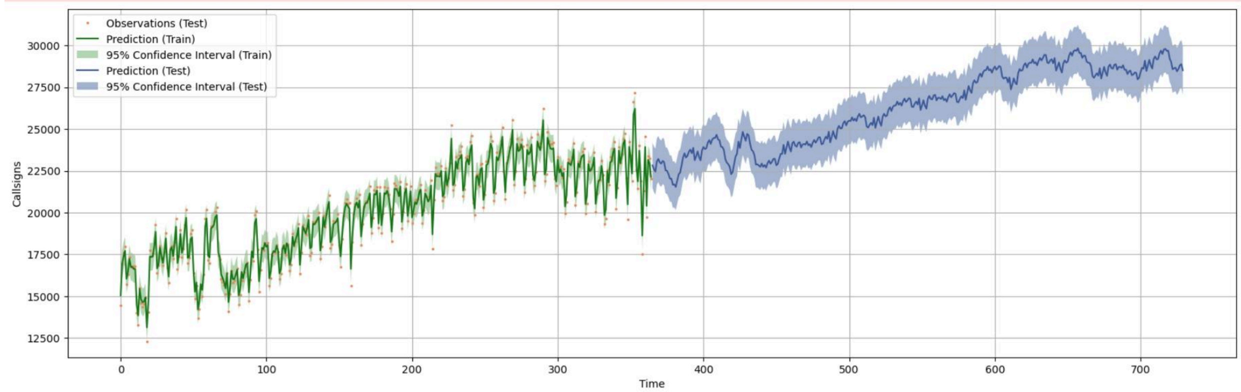
Gaussian Processing Analysis

Generally, Matern Kernel is applied for modeling roughness due to factors like daily fluctuations or operational disruptions; DotProduct kernel introduces a linearly increasing component into flight data; RationalQuadratic is useful for data with varied oscillation frequencies or amplitude changes; RBF captures trends that change slowly over time; WhiteKernel accounts for the noise in the data; ExpSineSquared kernel models weekly/monthly/yearly variations in flight patterns. Before predicting out-of-sample data, training and testing datasets are used for validating and upgrading each model (see graphs in appendices). Comparing the results, we can estimate the recovery trend of the post-pandemic flights.

(1) All Flights

The data has an obvious month-period and a smooth trend (see graphs in appendices), with very few anomalies(7).

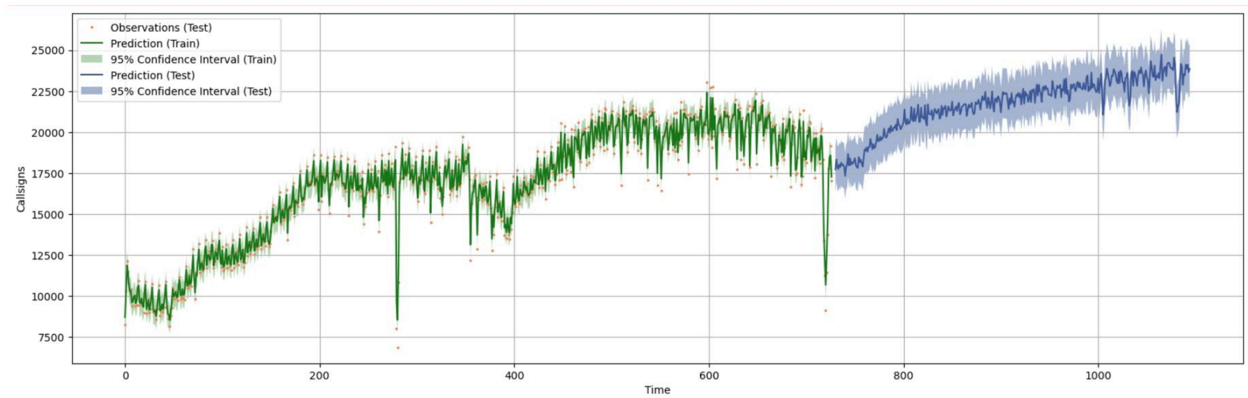
Pre-pandemic: GP model demonstrates good training data fit, but its high sensitivity to small changes could indicate overfitting, as reflected by the extreme values in the Matern and RBF kernels' length scales. The 2020 prediction suggests that flight amount per day should have reached about 30000 maximum without COVID-19.




```
Log Marginal Likelihood (optimized): -3749.474479552617
Optimized Kernel:
 316**2 * Matern(length_scale=7.27e-12, nu=1.5) + DotProduct(sigma_0=7.32e+04) + 316**2 * RationalQuadratic(alpha=7.24, length_scale=5.79e-15) + 316**2 *
RBF(length_scale=0.625) + WhiteKernel(noise_level=1e+05) + 316**2 * ExpSineSquared(length_scale=0.01, periodicity=31) + 316**2 * ExpSineSquared(length_sc
ale=0.0667, periodicity=365)
```

Figure 3.1.1 All Flights GP 2020 Prediction (pre-pandemic)

Post-pandemic: The GP model overfits the training set and diverges from the test set. The 2023 prediction shows that at the end of 2023, the flight volume is going to reach about 230,000 per day maximum, close to the average flight amount per day in the second half of 2019(post-pandemic).



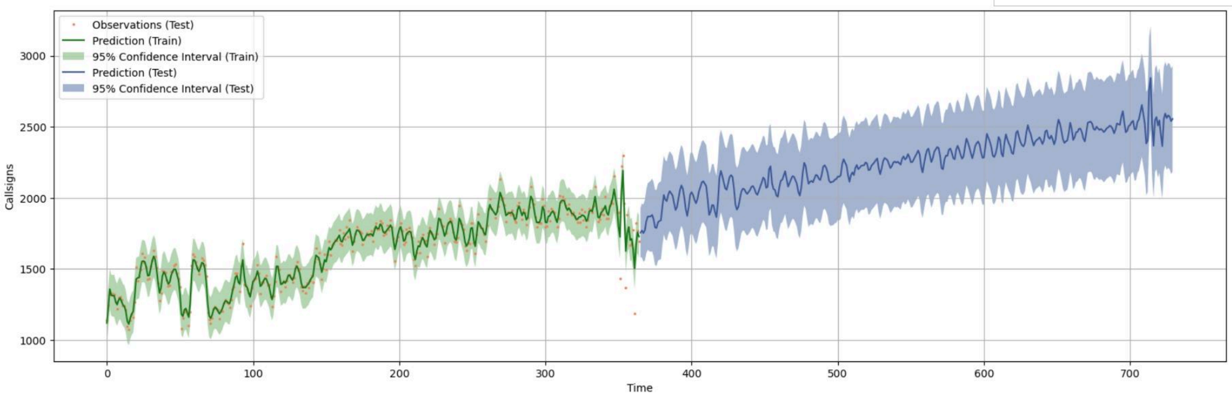

```
Log Marginal Likelihood (optimized): -6823.333660898841
Optimized Kernel:
 316**2 * Matern(length_scale=0.0848, nu=1.5) + DotProduct(sigma_0=1.15e+04) + 316**2 * RationalQuadratic(alpha=1e+05, length_scale=1.17) + 316**2 * RBF
(length_scale=35.6) + WhiteKernel(noise_level=1e+05) + 316**2 * ExpSineSquared(length_scale=0.01, periodicity=30.1) + 316**2 * ExpSineSquared(length_scal
e=0.0103, periodicity=363)
```

Figure 3.2.1 All Flights GP 2023 Prediction (post-pandemic)

(2) Wide-body Aircraft

The data has a month-period and a rough trend, with relatively more anomalies(22). Wide-body planes only take a very small proportion of all flight' data.

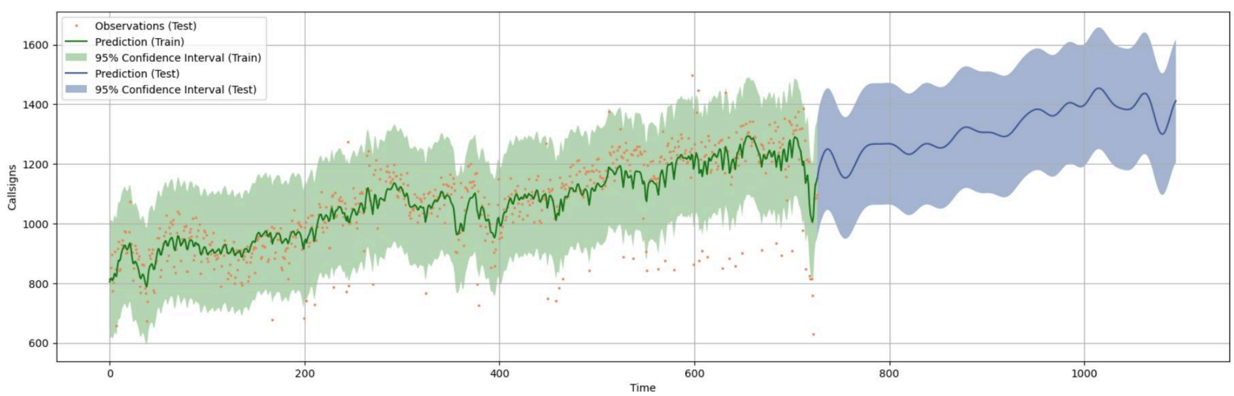
Pre-pandemic: Large confidence intervals in the testing set predictions point to potential overfitting to the training set and a high degree of uncertainty when facing testing set. The 2020 prediction suggests that flight amount per day should have boosted and reached about 2500-2800 maximum without COVID-19.




```
Log Marginal Likelihood (optimized): -2171.6893232737707
Optimized Kernel:
 10.1**2 * Matern(length_scale=20, nu=1.5) + DotProduct(sigma_0=1.29e+03) + 103**2 * RationalQuadratic(alpha=1.65, length_scale=7.1) + 17.6**2 * RBF(length_scale=20) + WhiteKernel(noise_level=3.81e+03) + 80.8**2 * ExpSineSquared(length_scale=0.0177, periodicity=361)
```

Figure 3.2.2 Wide-Body Flights GP 2020 Prediction (pre-pandemic)

Post-pandemic: the large confidence intervals in the predictions point to a high degree of uncertainty. The 2023 prediction shows that at the end of 2023, the flight volume is going to reach about 1400-1600 per day maximum, even slightly smaller than the average flight amount per day in the first season of 2019(post-pandemic). This indicates a slow recovery speed of wide-body flights.



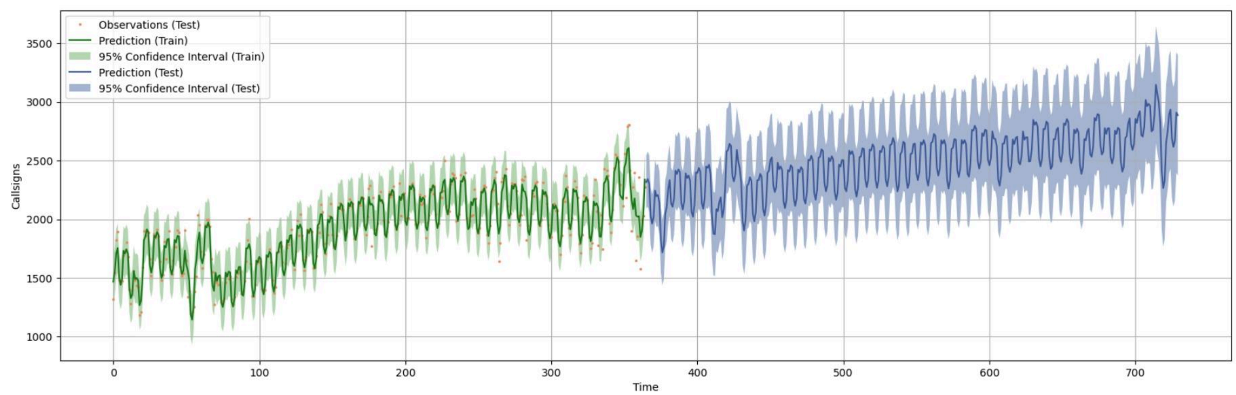

```
Log Marginal Likelihood (optimized): -4322.678825251027
Optimized Kernel:
 17.4**2 * Matern(length_scale=1, nu=1.5) + DotProduct(sigma_0=1.66e+03) + 38**2 * RBF(length_scale=1) + WhiteKernel(noise_level=8.51e+03) + 60.4**2 * ExpSineSquared(length_scale=1e+03, periodicity=30) + 70.5**2 * ExpSineSquared(length_scale=0.209, periodicity=360)
```

Figure 3.2.3 Wide-Body Flights GP 2023 Prediction (post-pandemic)

(3) Mid-body Aircraft

The data has both month- and week-period, with very few anomalies(6). Mid-body planes also only take a very small proportion of all flight' data.

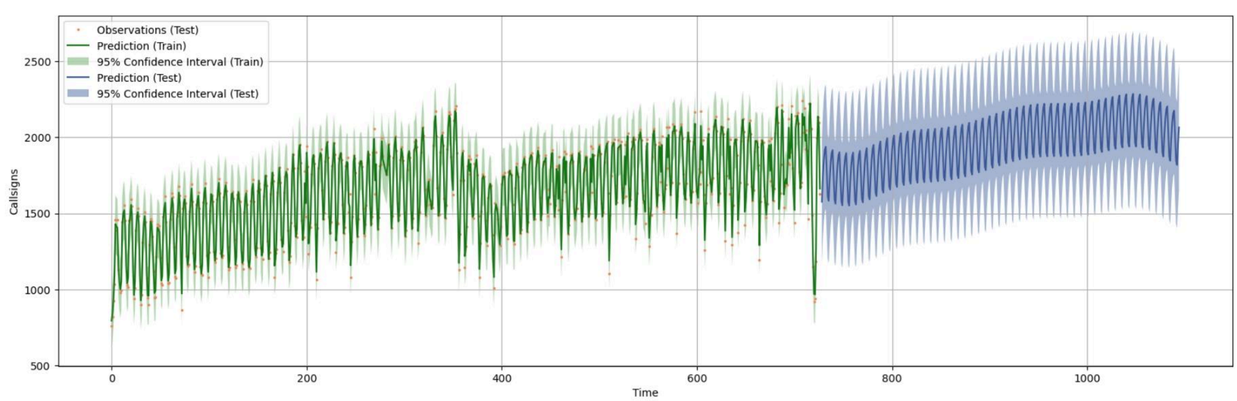
Pre-pandemic: Large confidence intervals in the testing set predictions point to potential overfitting to the training set and a high degree of uncertainty when facing the testing set. The 2020 prediction suggests that flight amount per day should have boosted and reached about 2800-3500 maximum without COVID-19.




```
Log Marginal Likelihood (optimized): -2285.920166757763
Optimized Kernel:
 5.29**2 * Matern(length_scale=20, nu=1.5) + DotProduct(sigma_0=1.52e+03) + 113**2 * RationalQuadratic(alpha=1.19, length_scale=20) + 7.61**2 * RBF(lengt
h_scale=20) + WhiteKernel(noise_level=9.09e+03) + 151**2 * ExpSineSquared(length_scale=0.108, periodicity=7) + 110**2 * ExpSineSquared(length_scale=0.043
9, periodicity=360)
```

Figure 3.3.1 Mid-Body Flights GP 2020 Prediction (pre-pandemic)

Post-pandemic: The ExpSineSquared kernels have very diverse length scales, from very short (0.117 for week-period) to longer ones (6.27 for month-period and 0.692 for year-period), suggesting the impact of time-based cycles varies significantly across different scales. The 2023 prediction shows that at the end of 2023, the flight volume is going to reach about 2300-2700 per day maximum, higher than or equal to the maximum flight amount per day in 2019(post-pandemic). This indicates a strong recovery of mid-body flights.




```
Log Marginal Likelihood (optimized): -4798.536994459281
Optimized Kernel:
 130**2 * Matern(length_scale=1, nu=1.5) + DotProduct(sigma_0=1.29e+03) + 140**2 * RBF(length_scale=1) + WhiteKernel(noise_level=5.26e+03) + 160**2 * Exp
SineSquared(length_scale=1.49, periodicity=7) + 143**2 * ExpSineSquared(length_scale=360, periodicity=30.2) + 125**2 * ExpSineSquared(length_scale=0.824,
periodicity=363)
```

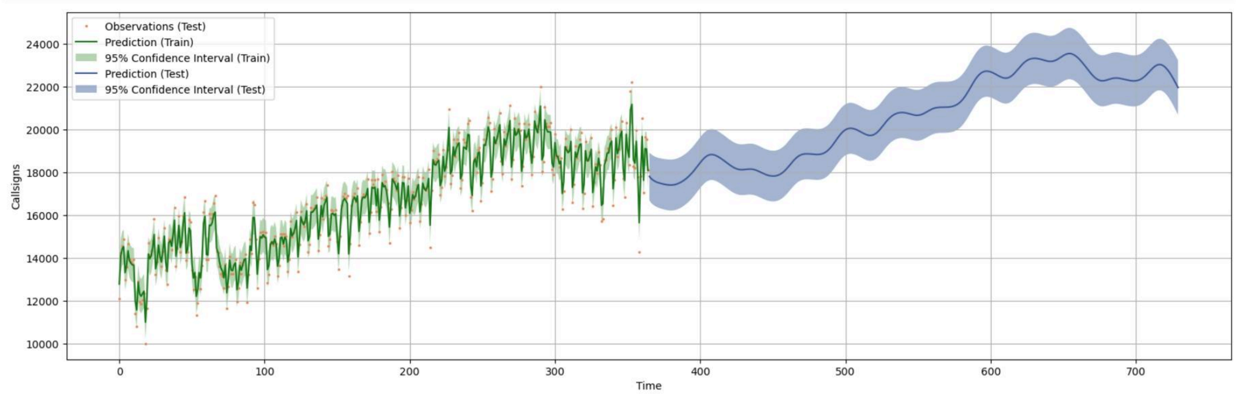
Figure 3.3.2 Mid-Body Flights GP 2023 Prediction (post-pandemic)

(4) Narrow-body Aircraft

The data has only an obvious month-period, with very few anomalies(6) interestingly all in post-pandemic. Narrow-body planes take the majority proportion of all flight' data.

Pre-pandemic: Matern Kernel has an extremely small length scale, suggesting the model is highly sensitive to minute changes in the data, potentially overfitting to the training data's noise.

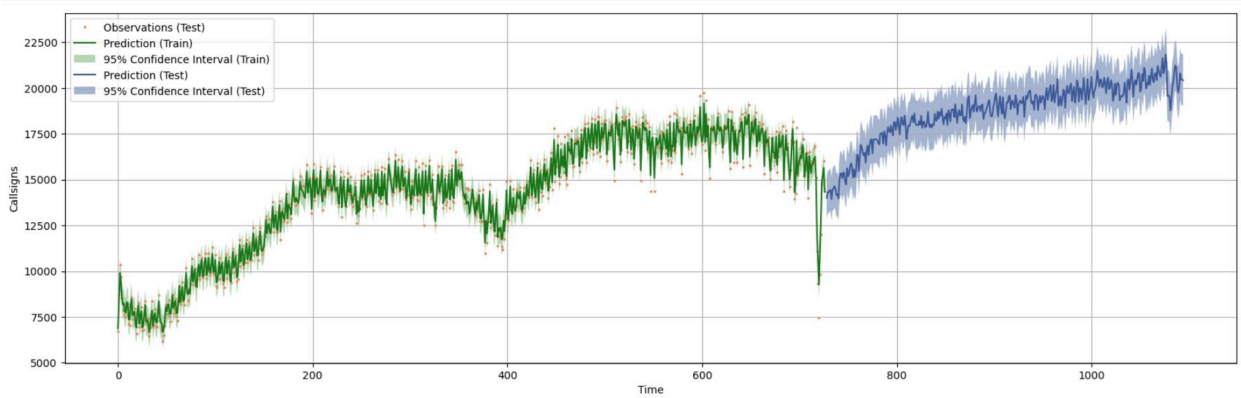
The 2020 prediction suggests that flight amount per day should have boosted and reached about 18000-20000 maximum without COVID-19.




```
Log Marginal Likelihood (optimized): -3584.470973498887
Optimized Kernel:
 316**2 * Matern(length_scale=2.7e-05, nu=1.5) + DotProduct(sigma_0=1.46e+04) + 316**2 * RationalQuadratic(alpha=1.34e-05, length_scale=9.88) + 316**2 *
RBF(length_scale=0.767) + WhiteKernel(noise_level=1e+05) + 316**2 * ExpSineSquared(length_scale=2.2, periodicity=31) + 316**2 * ExpSineSquared(length_sca
le=0.248, periodicity=365)
```

Figure 3.4.1 ***Narrow-Body Flights*** *GP 2020 Prediction* ***(pre-pandemic)***

Post-pandemic: WhiteKernel indicates a high level of noise in the data. The 2023 prediction shows that at the end of 2023, the flight volume is going to reach about 20000-22500 per day maximum, close to the maximum flight amount per day in 2019(post-pandemic).




```
Log Marginal Likelihood (optimized): -6333.147327860618
Optimized Kernel:
 316**2 * Matern(length_scale=0.00446, nu=1.5) + DotProduct(sigma_0=3.36e+03) + 316**2 * RationalQuadratic(alpha=1.32e-05, length_scale=47.9) + 316**2 *
RBF(length_scale=29.9) + WhiteKernel(noise_level=1e+05) + 316**2 * ExpSineSquared(length_scale=0.0119, periodicity=30.3) + 316**2 * ExpSineSquared(length
_scale=0.0133, periodicity=361)
```

Figure 3.4.2 ***Narrow-Body Flights*** *GP 2023 Prediction* ***(post-pandemic)***

Discussion, Limitation, and Future Research

Conclusion

Our research finds that post COVID recovery has airlines opting extremely conservative and cautious methods of resuming many flights with smaller, more economical aircrafts. And post-pandemic flight characteristics have seen the most prevalent plane type used are regional and narrow body aircrafts. We were also able to observe that widebody aircrafts (B777s, A380s, A340/350s) have potentially a faster recovery rate than midbody aircrafts, whose combine some advantages of narrow- and wide-body aircrafts but exhibits a trend of weekly seasonality, meaning the mid-body aircrafts are utilized on peak season to complement existing services, and therefore would not recover as fast (in total flight count and therefore total passenger count) as to the larger widebody aircrafts. In terms of overall flight count recovery, Gaussian Prediction methods shows that based on 2021 and 2022 global flight recovery trends, it will take another 2-3 years before total flight count and ridership would match the flight count projection based on pre-pandemic flight and ridership data that we have examined.

Limitation

The data on airline load factors before, during, and after the pandemic are approximated at 90%, 65%, and 90% respectively because passenger ridership will always remain a commercial secret where no airline is willing to release accurate passenger count per rider. 90% was chosen for pre- and post- pandemic dates because we have considered both undersold and oversold flights, and 65% was chosen as the weight for mid-pandemic flights due to varying levels of flying restrictions in different destination countries globally. With further investigation and perhaps the incorporation of a scientifically ratified load factor calculation method, we could model precise passenger account data, and ultimately improve prediction accuracy of our future estimations .

GPR Model Precision and Data Availability: Our Gaussian Process Regression (GPR) model's precision is limited due to the lack of available data from 2017-2018. As we only have data from one year prior to the pandemic, we are constrained to make predictions based on this single year. This situation leads to insufficient model training, and impact on the accuracy of our results. Need to mention, accessing more comprehensive data sets typically requires payment, unfortunately, this has been a prevention for us to achieve the ideal precision and completeness in our model.

Future research

Our future studies will focus on gathering a more extensive set of data, especially from the years 2017-2018, which are currently missing from our dataset. This will help in developing a more precise model and will enable more accurate predictions. Besides, to further refine our forecasting model, we also want to integrate geographic factors and varying degrees of pandemic impact across different regions. Data can be segmented by region according to the OD pair to analyze the specific impact of the pandemic on airline load factors in different areas, and a Severity index could also be introduced. This future direction not only aims to analyze the past of the airline industry but also contributes to the broader understanding of how global pandemics affect transportation, and makes our study more meaningful. Furthermore, our classification of datasets can be examined from the airline company perspective, where we classify and cluster based on airline alliances(SkyTeam, One World, Star Alliance), Hubs and operational headquarters, and most frequent services.

Group Responsibilities

Zhexuan Tang (zt2339)	Report Writing, editing, and Theme Selection Project Presentation (20%)
Ruixin Gan(rg4743)	Gaussian Process Analysis and Methodology (20%)
Zheyan Chen (zc3053)	Clustering Analysis and Methodology (20%)
Yin Wang(yw7422)	Data Research Cleaning & Filtering (20%)
Weihao Xia(wx2233)	Data Cleaning & Filtering, Time Series Analysis (20%)

References, Sources, and Code Base

References

Barczak, Agnieszka, et al. "The Impact of COVID-19 Pandemic on Air Transport Passenger Markets: Implications for Selected EU Airports Based on Time Series Models Analysis." *Sustainability*, vol. 14, no. 4345, 2022, doi:10.3390/su14074345.

Bouwer, J., Krishnan, V., Saxon, S., & Tufft, C. (2022). Taking stock of the pandemic's impact on Global Aviation. Retrieved from https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/taking-stock-of-the-pandemics-impact-on-global-aviation.

"COVID-19 - NFID." *Https://Www.nfid.org/*, www.nfid.org/infectious-diseases/covid-19/.

Ehsani, Sina, et al. "Predicting the Skies: A Novel Model for Flight-Level Passenger Traffic Forecasting." *arXiv preprint arXiv*:2401.03397v2, 9 Jan. 2024, arxiv.org/abs/2401.03397v2.

Firat, Murat, Yiltas-Kaplan, Derya, and Samli, Ruya. "Forecasting Air Travel Demand for Selected Destinations Using Machine Learning Methods." *Journal of Universal Computer Science*, vol. 27, no. 6, 28 Jun. 2021, doi:10.3897/jucs.68185.

Gao, Yi. "Benchmarking the Recovery of Air Travel Demands for US Airports During the COVID-19 Pandemic." *Transportation Research Interdisciplinary Perspectives*, vol. 13, 2022, doi:10.1016/j.trip.2021.100570.

Mutzabaugh, Ben. "New Airline Routes September 2019." *The Points Guy*, 30 Sept. 2019, thepointsguy.com/news/new-airline-routes-september-2019/.

Morris, Hugh. "Where You Should Be Flying in 2019: 11 Most Exciting New Routes." *Travel Agent Central*, 28 Dec. 2018, www.travelagentcentral.com/transportation/where-you-should-be-flying-2019-11-most-exciting-new-routes.

Schneider, Rose, and Xin Chen. "Predicting Flight Demand under Uncertainty." *KSCE Journal of Civil Engineering*, vol. 24, no. 2, 3 Jan. 2020, pp. 635–646, https://doi.org/10.1007/s12205-020-0857-9. Accessed 6 May 2024.

Tolcha, Tassew Dufera. "The State of Africa's Air Transport Market Amid COVID-19, and Forecasts for Recovery." *Journal of Air Transport Management,* vol. 108, 2023, doi:10.1016/j.jairtraman.2023.102380.

Wei, Fan & Chi, Hong. (2004). Cluster model for flight demand forecasting. 3170- 3173 Vol.4. 10.1109/WCICA.2004.1343106.

Data Source

IP2Location. "ip2location-iata-icao." *GitHub*, updated 28 Apr. 2024, github.com/ip2location/ip2location-iata-icao. Release note: Add Milan-Linate Airport.

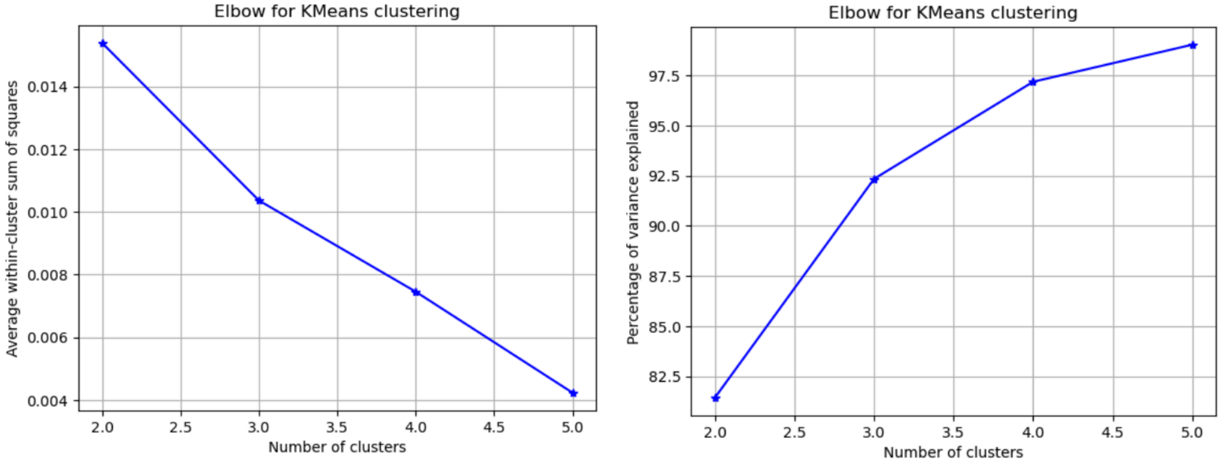
Olive, Xavier, Martin Strohmeier, and Jannis Lübbe. "Crowdsourced Air Traffic Data from The OpenSky Network 2020." *Zenodo*, 11 May 2023, zenodo.org/record/7923702.

Code Repository

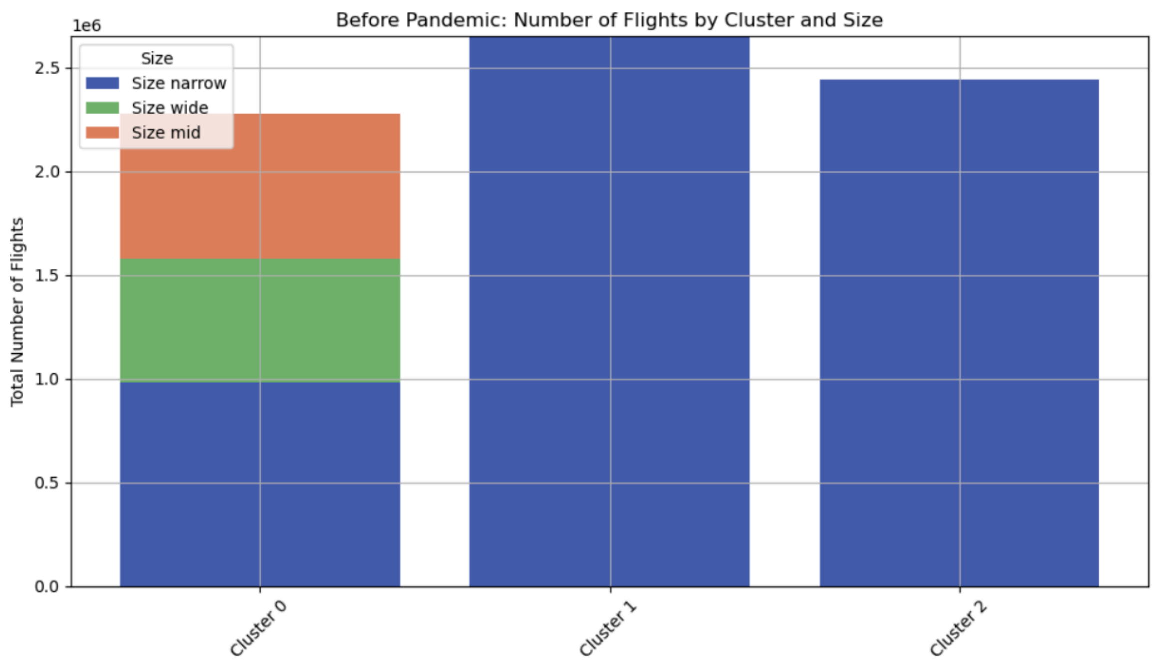
https://github.com/WeihaoXia/7033prjct

Appendices

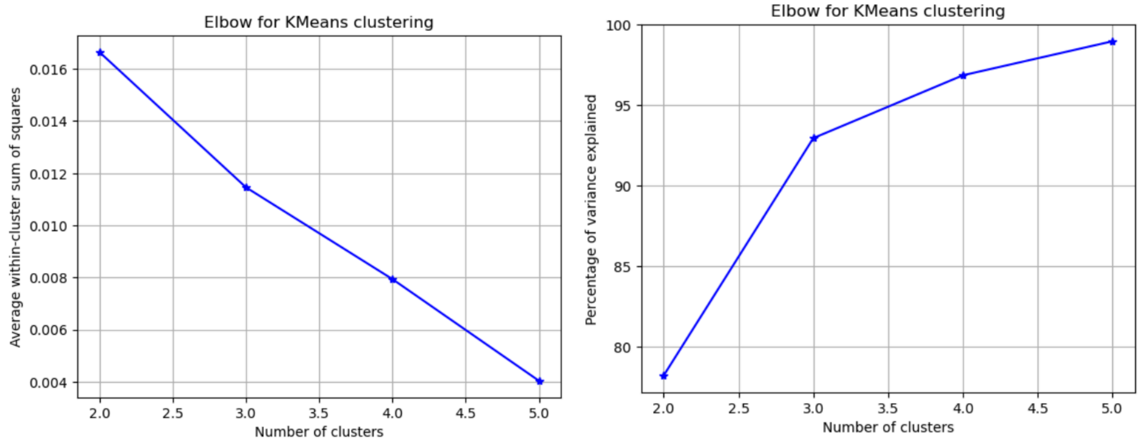
(1) Clustering



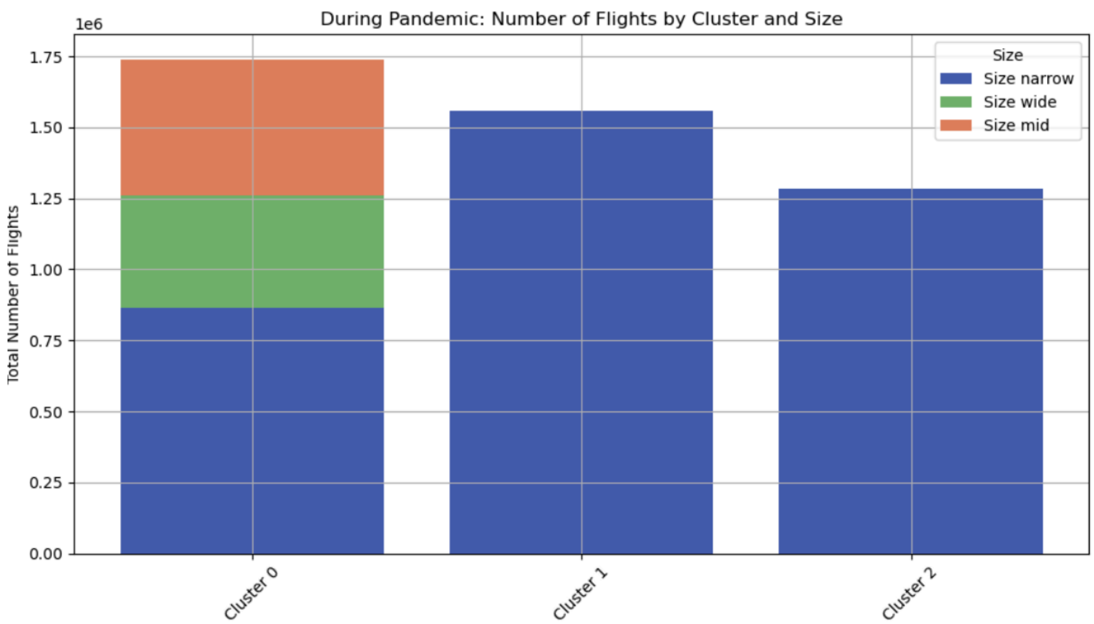
1.1.4 Elbow Method** for Optimal K cluster **(pre-pandemic)



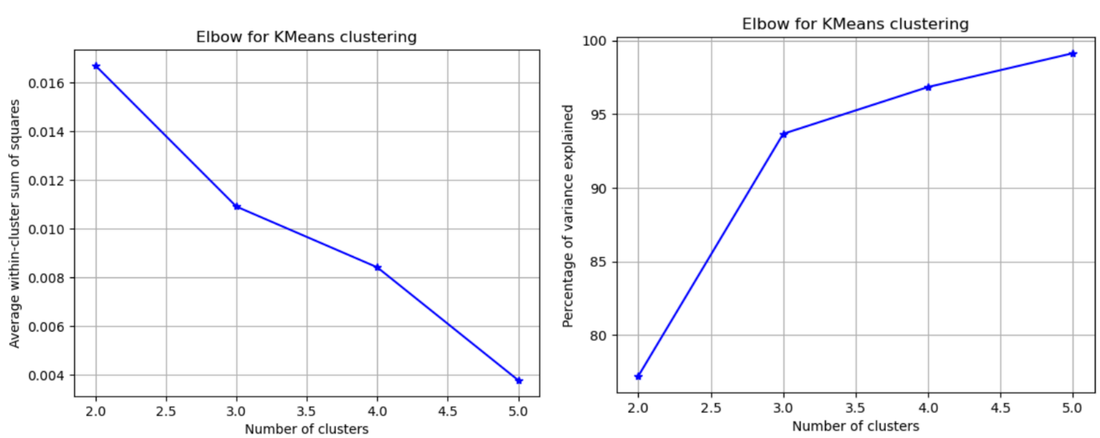
***1.1.5 Total Number of Flights** by Cluster and Size (**Pre-Pandemic**)*



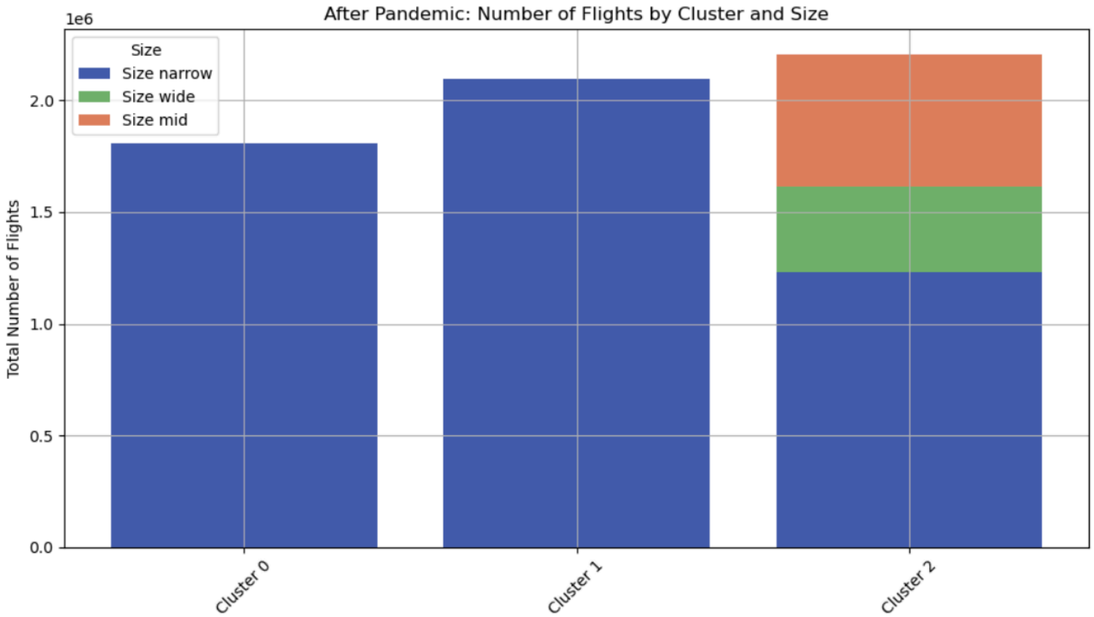
***1.2.4 Elbow Method** for Optimal K cluster (**During Pandemic**)*



***1.2.5 Total Number of Flights** by Cluster and Size (**During Pandemic**)*



***1.3.4 Elbow Method** for Optimal K cluster (**Post-pandemic**)*



***1.3.5 Total Number of Flights** by Cluster and Size (**During Pandemic**)*

(2) GP

(a) Time Series Decomposition

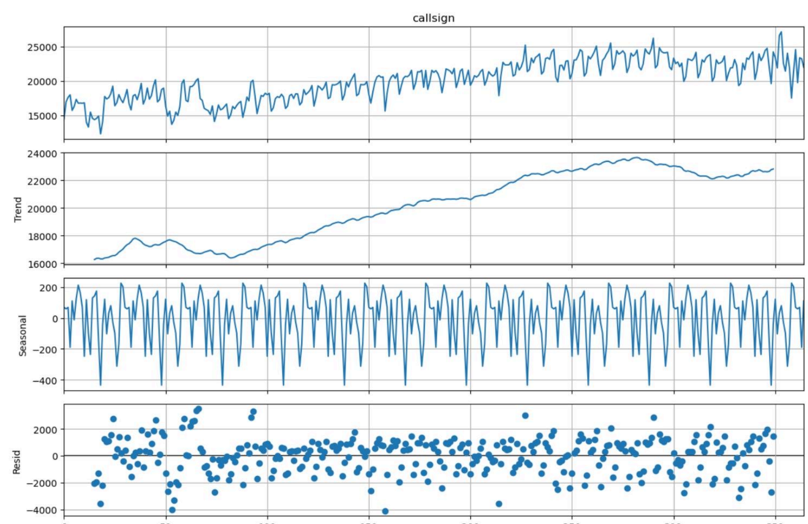


*Figure 3.5.1 **All Flights** Time Series Decomposition **(pre-pandemic)***

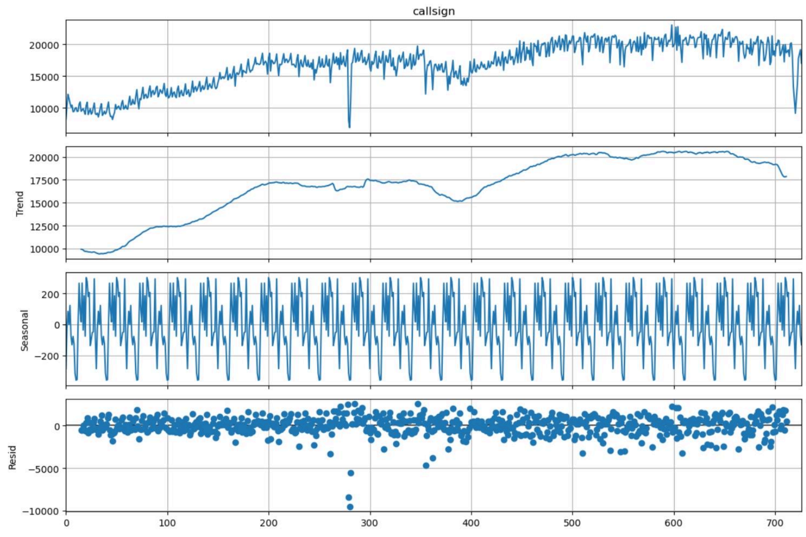


Figure 3.5.2 ***All Flights*** *Time Series Decomposition* ***(post-pandemic)***

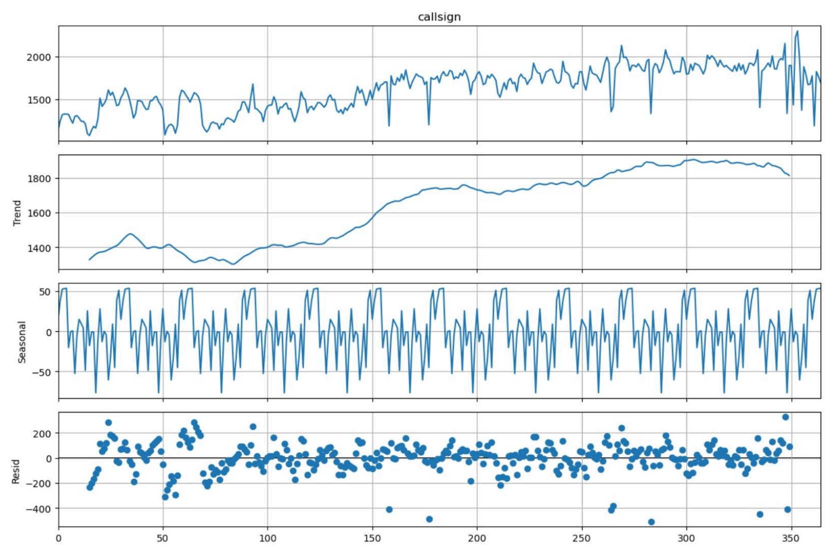


Figure 3.5.2 ***Wide-Body Flights*** *Time Series Decomposition* ***(pre-pandemic)***

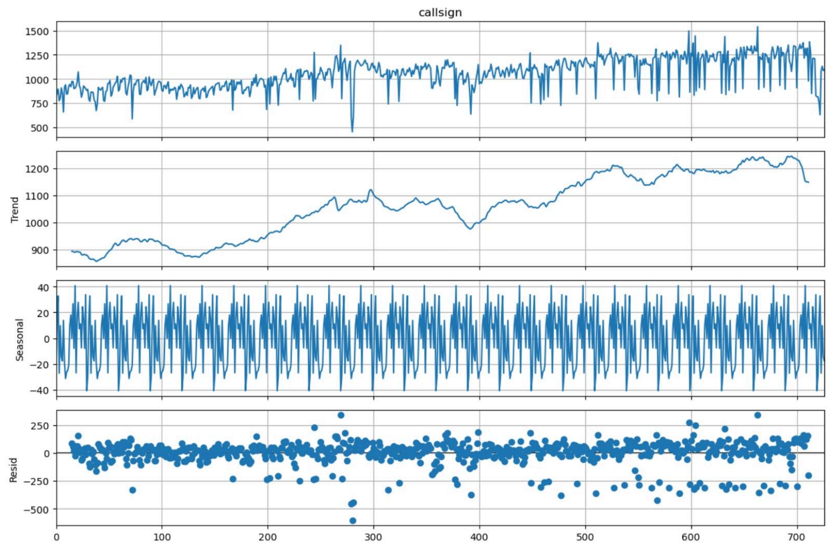


Figure 3.5.3 ***Wide-Body Flights*** *Time Series Decomposition* ***(post-pandemic)***

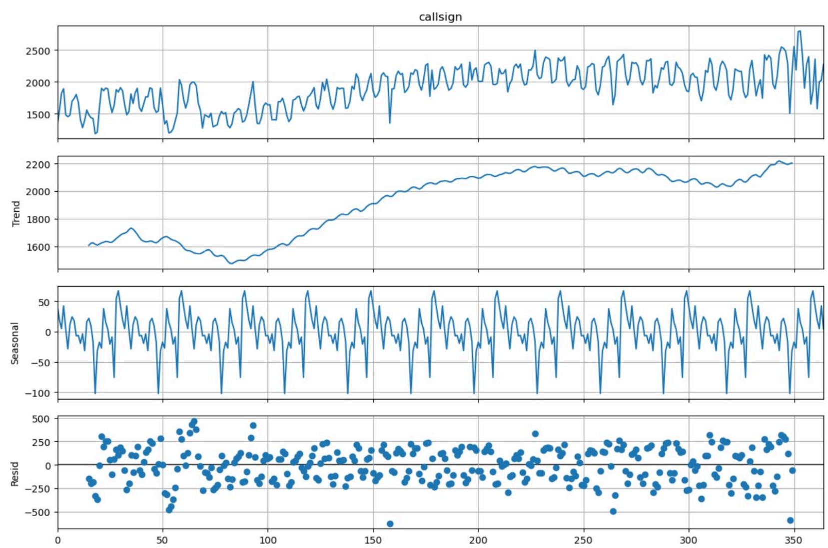


Figure 3.5.4 ***Mid-Body Flights*** *Time Series Decomposition* ***(pre-pandemic)***

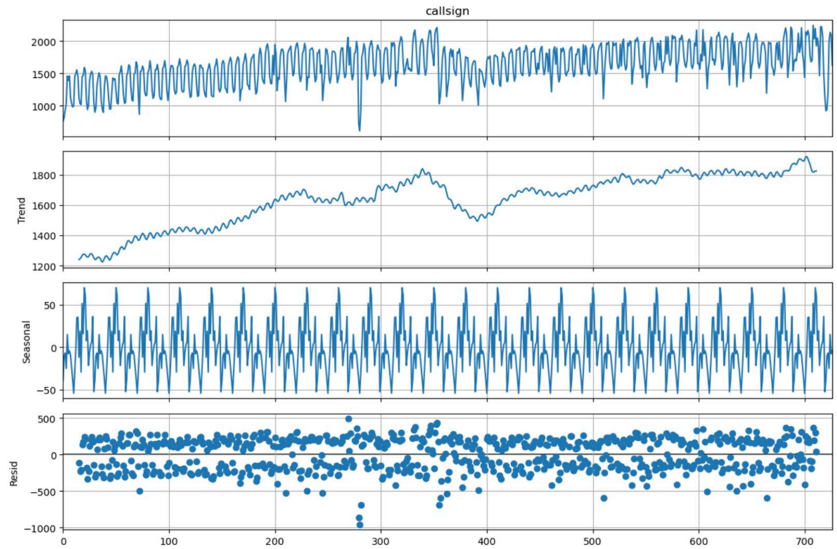


Figure 3.5.4 ***Mid-Body Flights*** *Time Series Decomposition* ***(post-pandemic)***

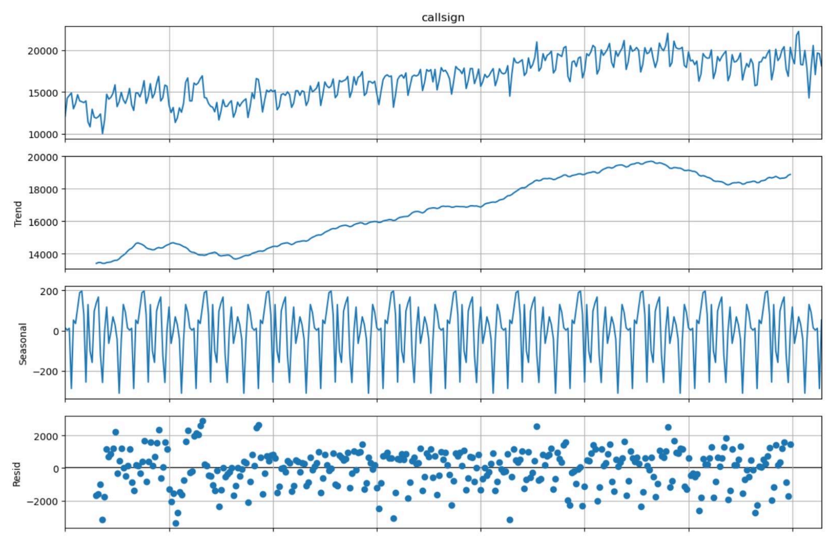


Figure 3.5.5 ***Narrow-Body Flights*** *Time Series Decomposition* ***(pre-pandemic)***

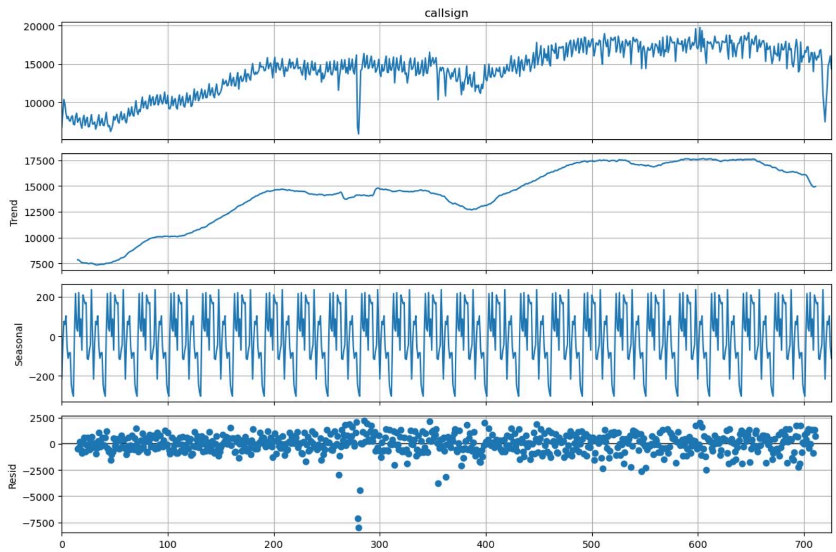
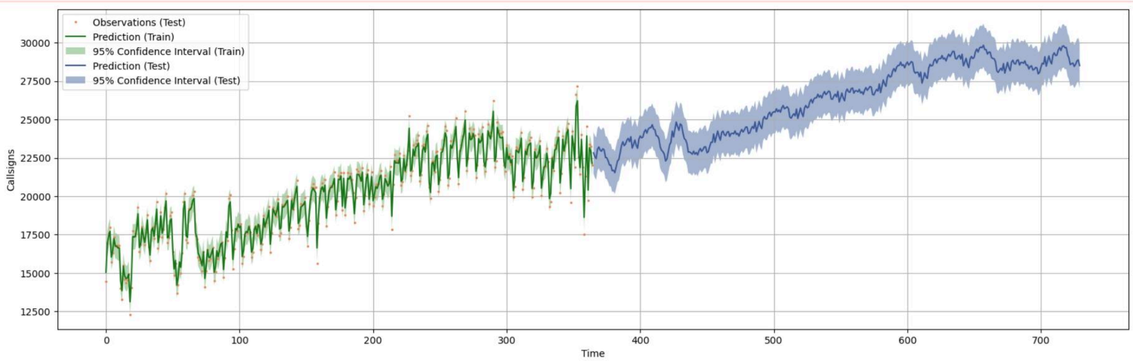


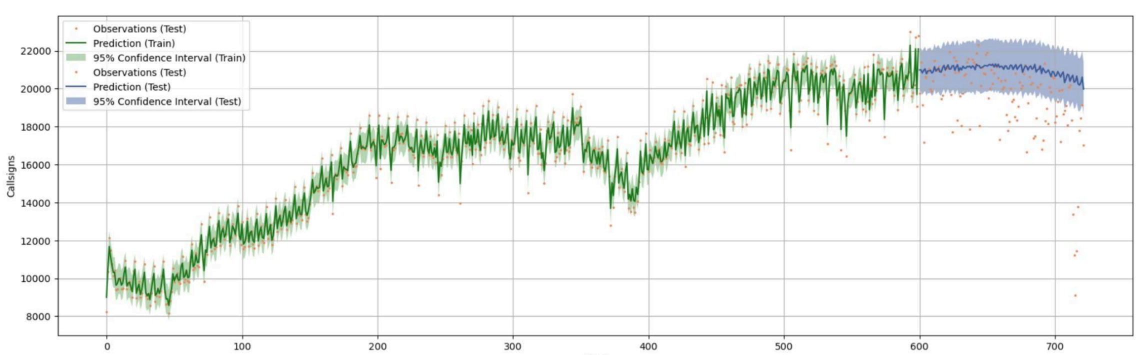
Figure 3.5.6 ***Narrow-Body Flights*** *Time Series Decomposition* ***(post-pandemic)***

(b) GP Training and Testing



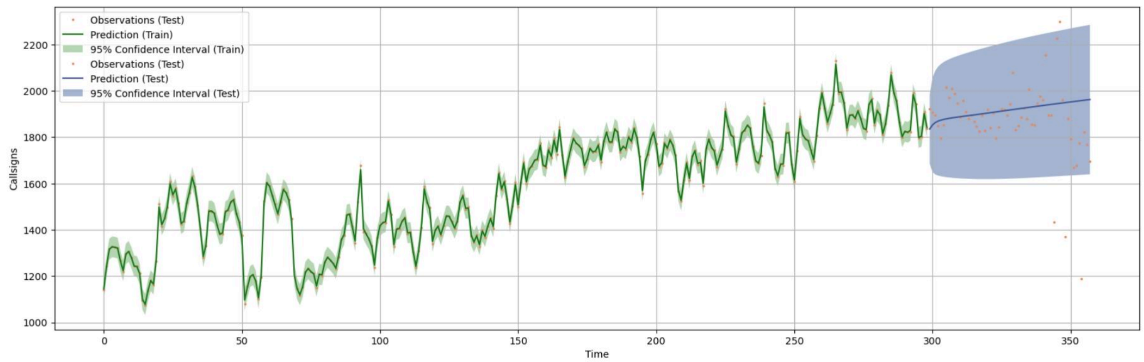

```
Log Marginal Likelihood (optimized): -3749.474479552617
Optimized Kernel:
 316**2 * Matern(length_scale=7.27e-12, nu=1.5) + DotProduct(sigma_0=7.32e+04) + 316**2 * RationalQuadratic(alpha=7.24, length_scale=5.79e-15) + 316**2 *
RBF(length_scale=0.625) + WhiteKernel(noise_level=1e+05) + 316**2 * ExpSineSquared(length_scale=0.01, periodicity=31) + 316**2 * ExpSineSquared(length_sc
ale=0.0667, periodicity=365)
```

Figure 3.6.1 ***All Flights*** *GP Training and Testing* ***(pre-pandemic)***



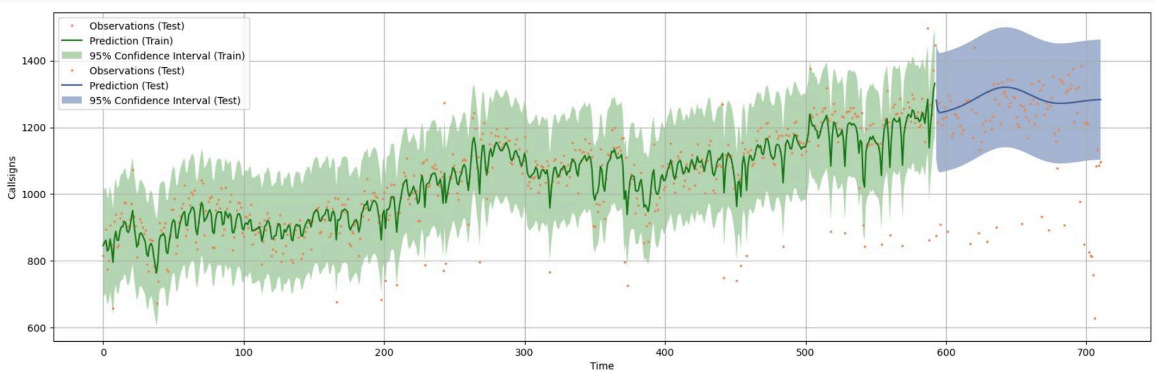

```
Log Marginal Likelihood (optimized): -5347.588616935177
Optimized Kernel:
 316**2 * Matern(length_scale=0.0136, nu=1.5) + DotProduct(sigma_0=2.32e+04) + 316**2 * RationalQuadratic(alpha=1.09, length_scale=0.518) + 316**2 * RBF
(length_scale=24.6) + WhiteKernel(noise_level=1e+05) + 128**2 * ExpSineSquared(length_scale=0.0573, periodicity=30.8) + 316**2 * ExpSineSquared(length_sc
ale=0.635, periodicity=360)
```

Figure 3.6.2 ***All Flights*** *GP Training and Testing* ***(post-pandemic)***



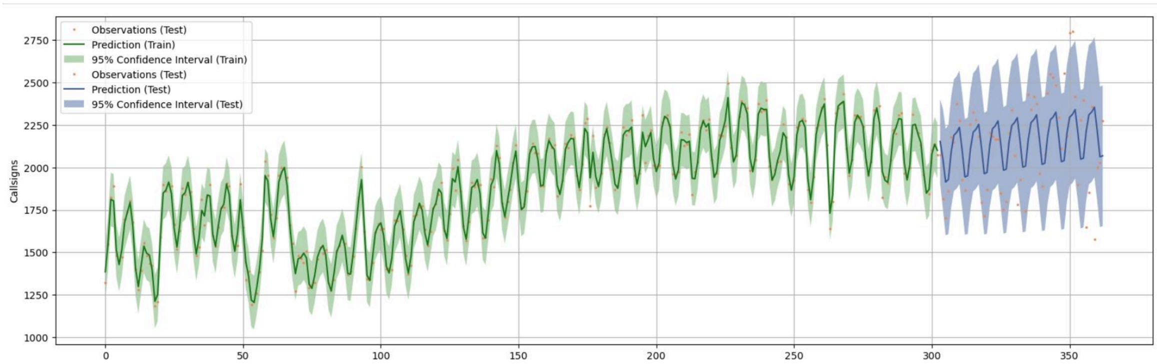

```
Log Marginal Likelihood (optimized): -1726.5773648098161
Optimized Kernel:
 1.14**2 * Matern(length_scale=20, nu=1.5) + DotProduct(sigma_0=1.27e+03) + 121**2 * RationalQuadratic(alpha=0.381, length_scale=1.26) + 55.9**2 * RBF(le
ngth_scale=20) + WhiteKernel(noise_level=275) + 287**2 * ExpSineSquared(length_scale=8.41, periodicity=365)
```

Figure 3.6.3 ***Wide-Body Flights*** *GP Training and Testing* ***(pre-pandemic)***



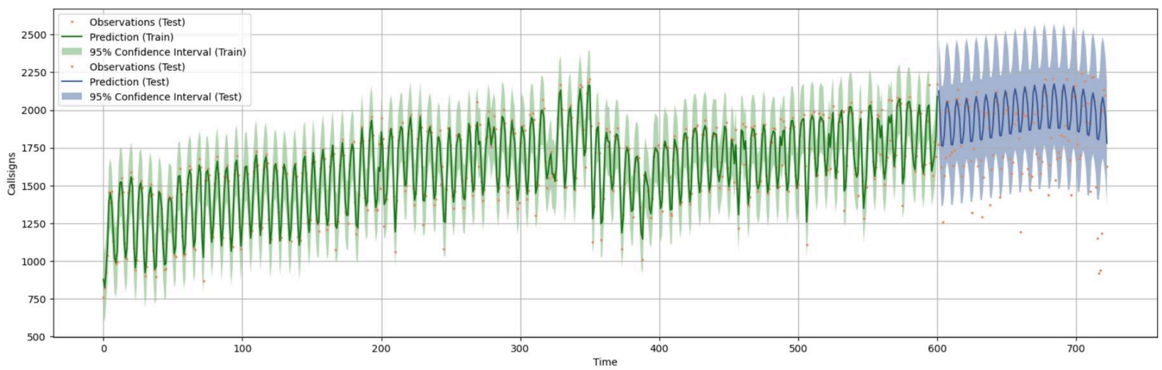

```
Log Marginal Likelihood (optimized): -3513.1271248576104
Optimized Kernel:
 56.2**2 * Matern(length_scale=1, nu=1.5) + DotProduct(sigma_0=870) + 0.141**2 * RBF(length_scale=1) + WhiteKernel(noise_level=4.84e+03) + 9.66**2 * ExpS
ineSquared(length_scale=1e+03, periodicity=30) + 37.8**2 * ExpSineSquared(length_scale=0.388, periodicity=360)
```

Figure 3.6.4 ***Wide-Body Flights*** *GP Training and Testing* ***(post-pandemic)***



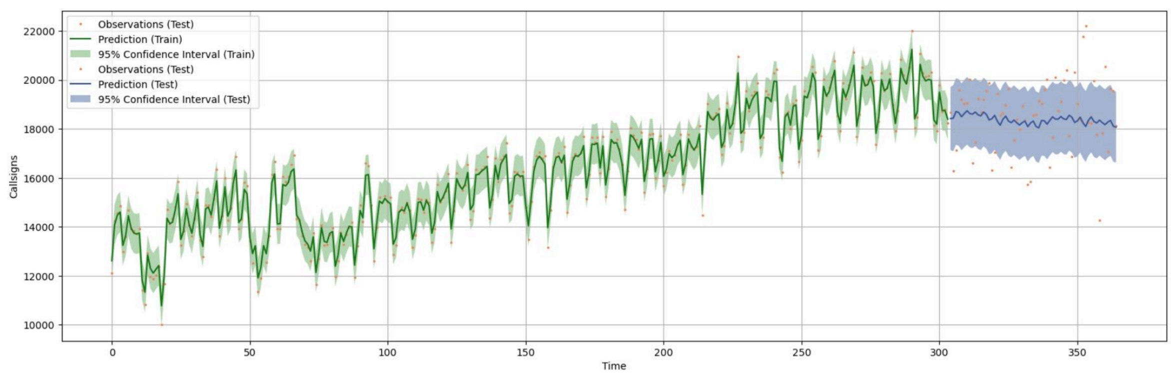

```
Log Marginal Likelihood (optimized): -1904.5408237513245
Optimized Kernel:
 131**2 * Matern(length_scale=1.57, nu=1.5) + DotProduct(sigma_0=1.62e+03) + 120**2 * RationalQuadratic(alpha=0.603, length_scale=20) + 0.00316**2 * RBF
(length_scale=20) + WhiteKernel(noise_level=3.96e+03) + 198**2 * ExpSineSquared(length_scale=1.61, periodicity=7) + 9.64**2 * ExpSineSquared(length_scale
=4.19e-09, periodicity=365)
```

Figure 3.6.5 ***Mid-Body Flights*** *GP Training and Testing* ***(pre-pandemic)***



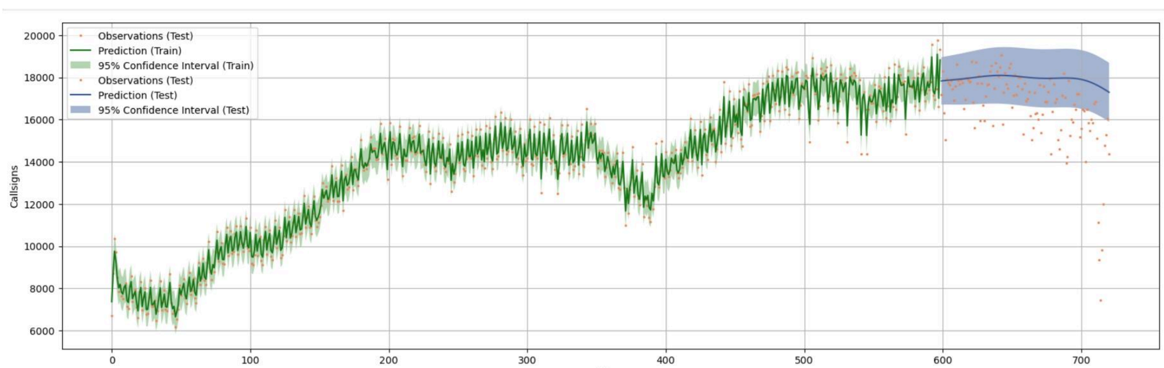

```
Log Marginal Likelihood (optimized): -3972.1414220985384
Optimized Kernel:
 0.58**2 * Matern(length_scale=1, nu=1.5) + DotProduct(sigma_0=1.39e+03) + 176**2 * RBF(length_scale=1) + WhiteKernel(noise_level=8.34e+03) + 123**2 * Ex
pSineSquared(length_scale=0.117, periodicity=7) + 0.878**2 * ExpSineSquared(length_scale=6.27, periodicity=30) + 95.4**2 * ExpSineSquared(length_scale=0.
692, periodicity=365)
```

Figure 3.6.6 ***Mid-Body Flights*** *GP Training and Testing* ***(post-pandemic)***




```
Log Marginal Likelihood (optimized): -2804.968337550533
Optimized Kernel:
 316**2 * Matern(length_scale=1.22e-06, nu=1.5) + DotProduct(sigma_0=1.5e+04) + 316**2 * RationalQuadratic(alpha=0.441, length_scale=10) + 316**2 * RBF(l
ength_scale=0.753) + WhiteKernel(noise_level=1e+05) + 203**2 * ExpSineSquared(length_scale=0.166, periodicity=31) + 316**2 * ExpSineSquared(length_scale=
0.013, periodicity=365)
```

Figure 3.6.7 ***Narrow-Body Flights*** *GP Training and Testing* ***(pre-pandemic)***




```
Log Marginal Likelihood (optimized): -5108.253460352412
Optimized Kernel:
 316**2 * Matern(length_scale=35.2, nu=1.5) + DotProduct(sigma_0=1.06e+04) + 316**2 * RationalQuadratic(alpha=1e+05, length_scale=3.89e-26) + 316**2 * RB
F(length_scale=0.246) + WhiteKernel(noise_level=1e+05) + 0.00321**2 * ExpSineSquared(length_scale=1.31e-25, periodicity=31) + 316**2 * ExpSineSquared(len
gth_scale=0.481, periodicity=360)
```

Figure 3.6.8 ***Narrow-Body Flights*** *GP Training and Testing* ***(post-pandemic)***