Analysis and Modeling of the Relationship Between COVID-19 Data and Air Traffic

Members & Student numbers

- Minghao Li 6212999
- Xinyu Yang 6302750
- Yilin Shi 6140343
- Yue Guo 6147275
- Yumeng Pan 6134130

Introduction

The COVID-19 pandemic significantly influenced global air traffic, including in the Netherlands, where travel restrictions and behavioral shifts caused substantial fluctuations in air traffic volume. As the country reopened in phases, the recovery process unveiled diverse patterns among various flight types and traveler segments. This study aims to analyze historical traffic data to investigate the disparities across distinct flight categories. Insight into these dynamics is essential for shaping effective air traffic management strategies and policy responses to future pandemics or similar disruptions.

Research Objectives

This project aims to analyze the relationship between COVID-19 data and air traffic. It involves examining traffic volumes of various flight types such as domestic, international, passenger, and cargo during the pandemic. The study seeks to understand how these flights were impacted by COVID-19 and to provide insights into the correlation between COVID-19 data and air traffic patterns.

Research Questions (RQs)

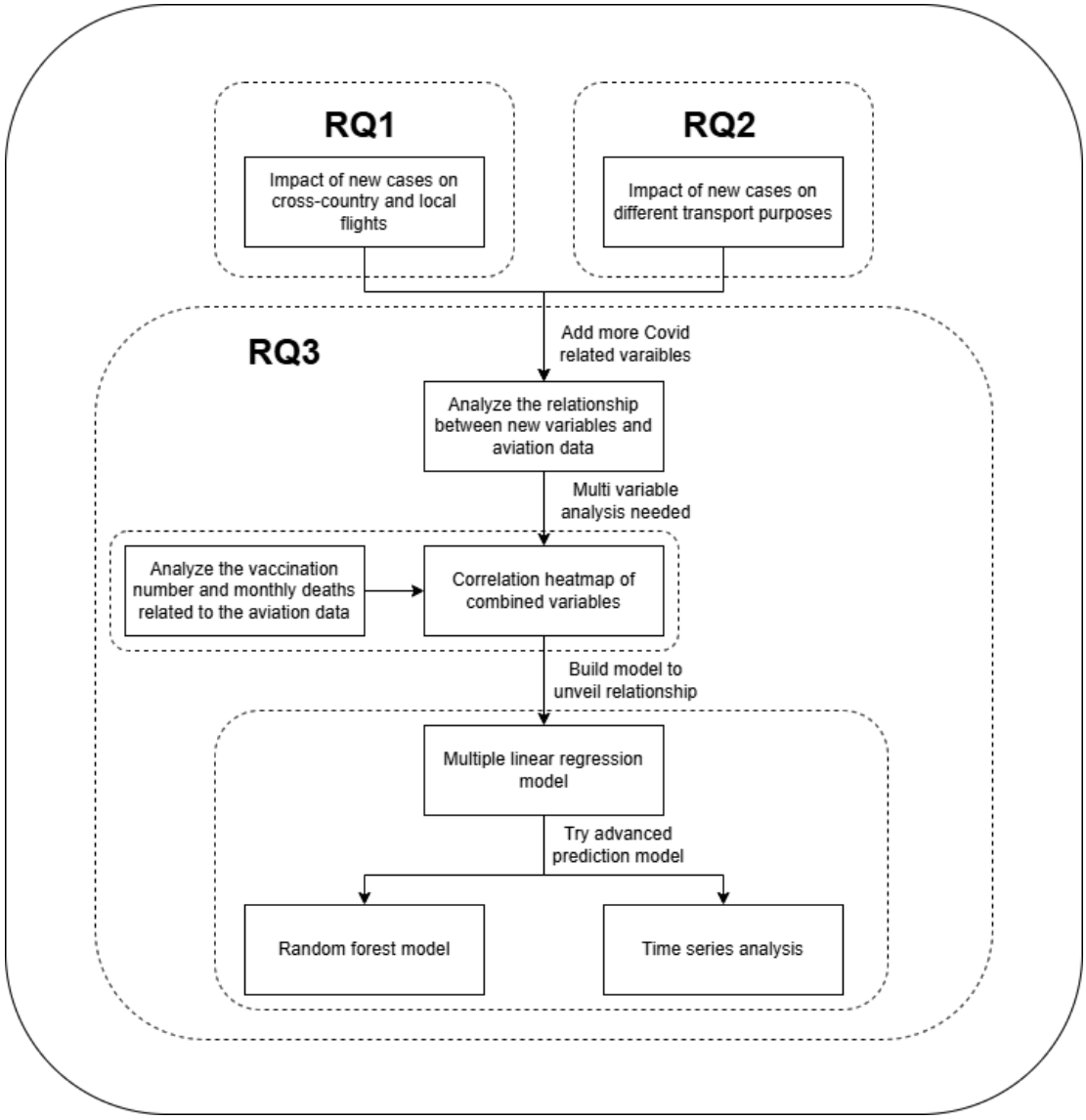
- RQ1: What is the impact of the monthly COVID-19 new cases on different flight types (e.g. cross-country and local flights) in the Netherlands?
- RQ2: What is the impact of the monthly COVID-19 new cases on different transportation purposes (e.g. numbers of passengers, cargo, and mail) in air traffic in the Netherlands?
- RQ3: The modeling of relationship between aviation operation and Covid-19 data.

Data Sources

- Our World in Data (OWID) COVID-19 Dataset:
  - https://ourworldindata.org/coronavirus

    - Pandemic-related data including new cases in each month will be used to correlate air traffic pattern variation.
- CBS (Statistics Netherlands) Open Data:
    - CBS Transport and Mobility Dataset:
    - https://opendata.cbs.nl/statline/#/CBS/en/dataset/37478eng/table?ts=1728287180831
    - This source is used to research the variation across different flight modes and the numbers of passengers, cargo, and mail through the COVID period.

Mindmap



Analysis and Visualization of Question 1 and 2

In this section, we analyze the global COVID-19 infection data and its impact on aviation activities. The analysis includes the following steps:

1. **Data Reading and Preprocessing**:

    - Read the global COVID-19 infection data from an Excel file.

- Convert the 'date' column to datetime format.
- Group the data by month and calculate the total number of new cases for each month.

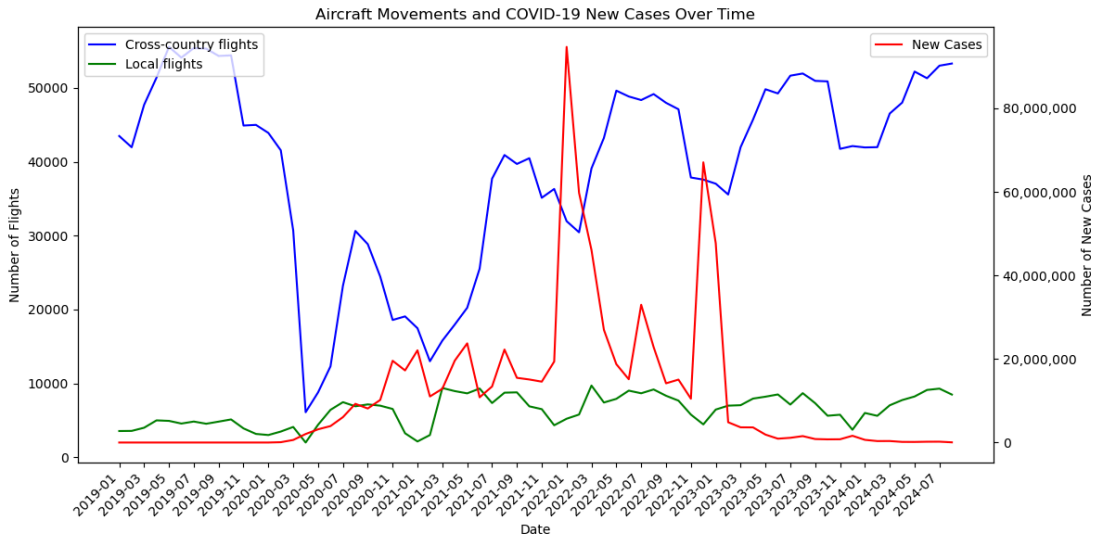
2. **Merging with Aviation Data**:
   - Read the monthly aviation data from a CSV file.
   - Parse the date column and group the data by month.
   - Merge the COVID-19 data with the aviation data on the 'month' column.
3. **Visualization**:
   - Plot the number of new COVID-19 cases and various aviation-related variables over time.
   - Use line charts to visualize the trends and relationships between the variables.

The following plots are generated to visualize the data:

1. **Aircraft Movements and COVID-19 New Cases Over Time**:
   - Cross-country flights and local flights are plotted on the primary y-axis.
   - New COVID-19 cases are plotted on the secondary y-axis.
2. **Commercial Air Traffic and COVID-19 New Cases Over Time**:
   - Total passengers, total cargo, and total mail are plotted on the primary y-axis.
   - New COVID-19 cases are plotted on the secondary y-axis.

These visualizations help us understand the impact of the COVID-19 pandemic on aviation activities and identify any potential correlations between the variables.

Research Question 1: The Impact of COVID-19 on Cross-country and Local Flight Trends



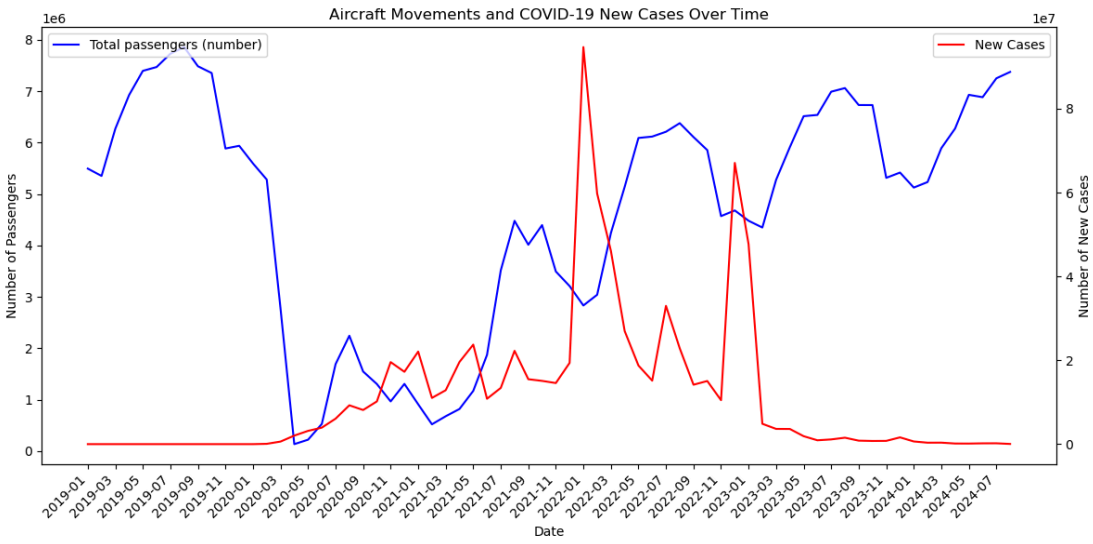
The image shows that during non-pandemic periods, both cross-country and local flights follow a seasonal pattern, with more flights in the summer and fewer

in the winter. Cross-country flights were hardest hit during the early stages of the pandemic, with numbers sharply declining from November 2019 and reaching a low in May 2020, dropping below 10,000. As the pandemic eased, cross-country flights gradually recovered and returned to near pre-pandemic levels by 2023.

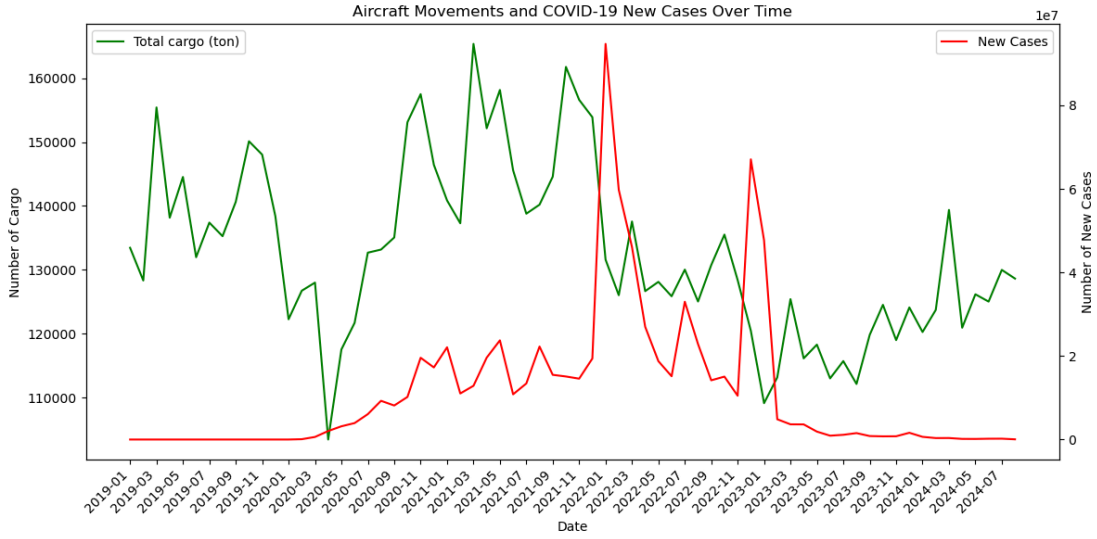
Local flights also declined, though less dramatically, and recovered more slowly, with overall numbers remaining lower. The image suggests no clear relationship between global COVID-19 cases and flight trends. This is likely because flight numbers were more influenced by government policies than the direct rise in cases. Many countries imposed strict travel restrictions, including border closures and flight cancellations, early in the pandemic, often before case numbers peaked.

In conclusion, the pandemic significantly impacted both cross-country and local flight numbers. However, the impact might be more closely tied to government policies and travel restrictions than to the rise in case numbers. Flight trends aligned more with the timing and adjustments of these preventive measures.

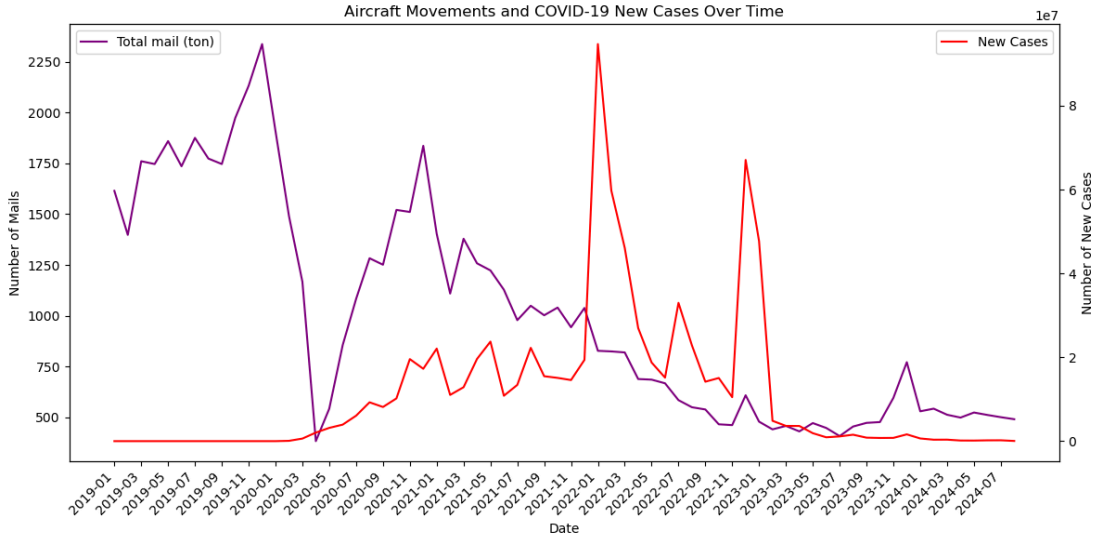
Research Question 2: The Impact of COVID-19 on the Numbers of Passengers, Cargo, and Mail



Before the pandemic in 2019, the number of passengers fluctuated between 5 million and 8 million, with higher numbers during the summer, likely due to tourism. However, by April 2020, passenger numbers plummeted to around 133,752 due to the outbreak of COVID-19. A slight recovery occurred between July and September 2020, reaching approximately 2 million passengers, although still lower than pre-pandemic levels. This period showed a consistent seasonal pattern of higher passenger numbers in the summer. Following this, passenger numbers declined again from November 2020 to March 2021, then rebounded to around 4 million by September 2021. After that, the number of passengers gradually increased yearly, but the cyclical trend of more passengers in summer and fewer in winter persisted.



Before the pandemic, cargo volumes fluctuated between 130,000 and 160,000 tons, with relatively stable cargo operations. Following the outbreak of COVID-19 in early 2020, cargo volumes dropped significantly to a low of 103,420 tons in May, likely due to reduced flights and logistical constraints. From mid-2020 to early 2021, cargo volumes started to recover but continued to fluctuate significantly, possibly due to the resurgence of the pandemic and varying lockdown measures across regions. By 2022 and 2023, cargo volumes had decreased slightly, but after 2023, a gradual increase was observed, although not reaching pre-pandemic levels.



From 2019 to January 2020, the number of mail showed considerable fluctuations but remained relatively high, between 1,250 and 2,000 tons. With the onset of the pandemic in early 2020, the number of mail fell sharply, reaching a low of 382 tons between March and May 2020, likely due to global postal disruptions and international lockdowns. From May 2020 onward, mail shipments gradually increased, reflecting a recovery in air mail transport. However, after May 2021, the number of mail began to decline again, showing some fluctuations. By 2023, mail shipments had stabilized but remained well below pre-pandemic levels.

Summary

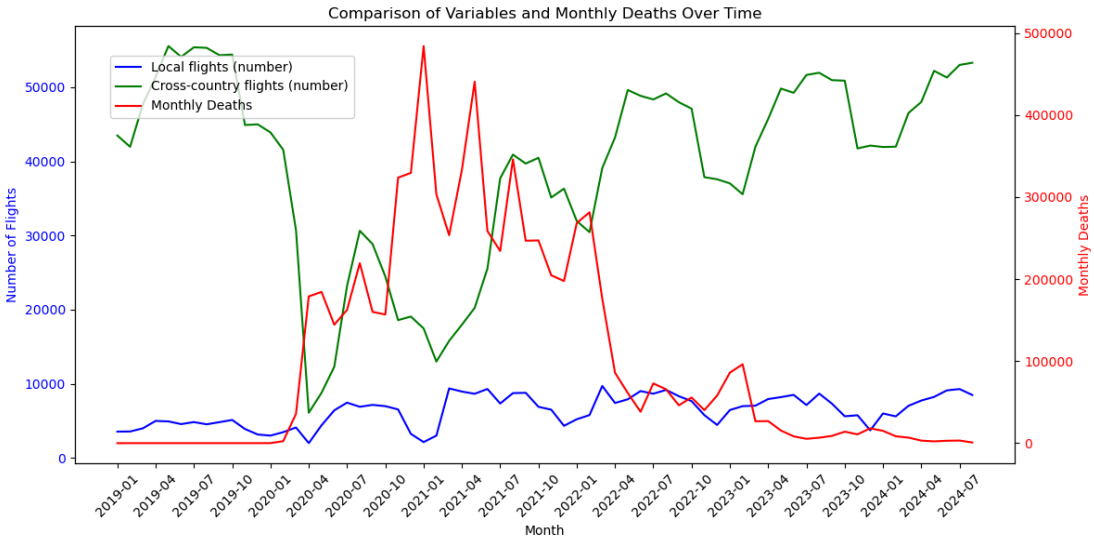
Both cross-country and local flights were heavily impacted by the pandemic. However, the recovery and trends in air traffic were more likely linked to governmental actions and travel policies rather than the mere progression of COVID-19 cases globally.In the early stages of the pandemic, despite the number of new cases being relatively low, the number of passengers, cargo, and mail dropped significantly. From mid-2020 to mid-2021, even as the number of new cases rose, overall air transport volumes began to recover. In early 2022, as new case numbers spiked, cargo volumes showed a notable decrease, while the number of passengers and mail experienced smaller drops. Similarly, during subsequent peaks in new cases, air transport volumes saw little variation.

The analysis suggests a weak correlation between air transport volumes and global new COVID-19 cases. Therefore, we will introduce more Covid related variables to further investigate the relationship between COVID-19 data and air traffic in the next part.

Research Question 3: Analysis and Modeling

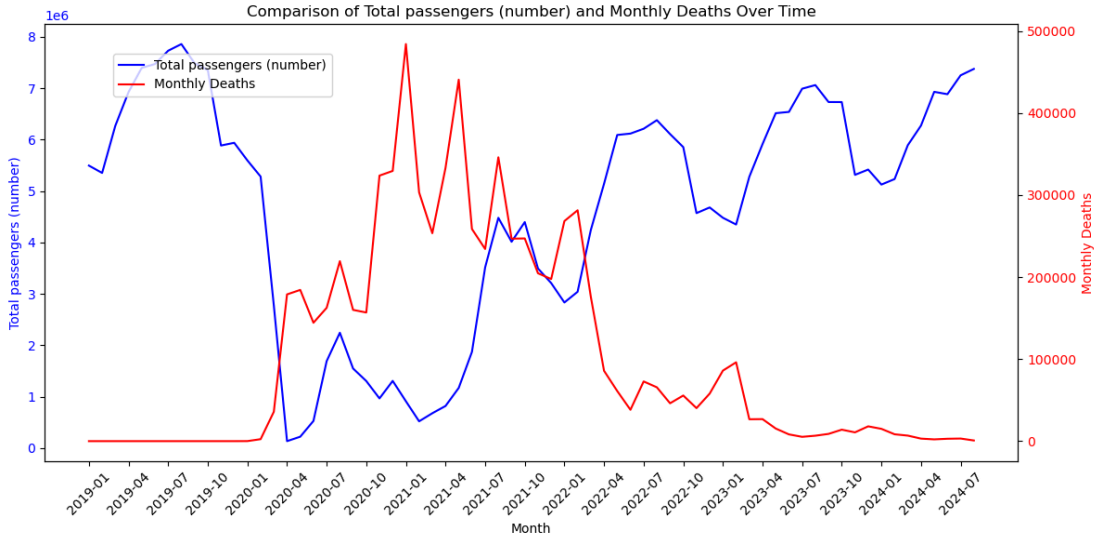
We introduced two new COVID-19 data variables: monthly deaths and monthly vaccinations. Next, we will plot the five aviation-related variables (cross-country flights, local flights, number of passengers, amount of cargo, and amount of mail) alongside the monthly deaths and monthly vaccinations respectively. These plots will help visualize the trends and relationships between the aviation variables and the COVID-19 data over time.

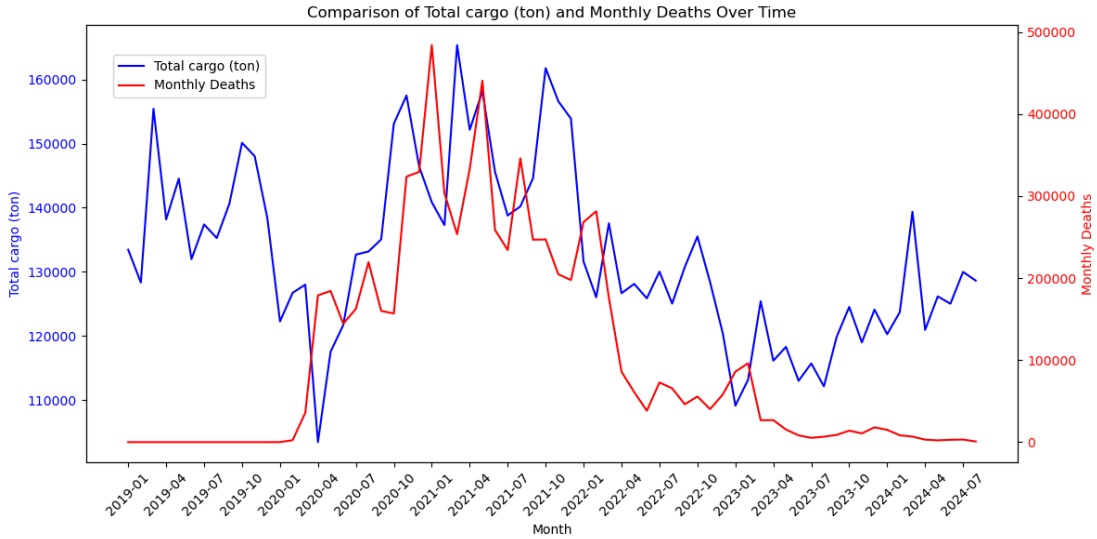
The first part is to analyze the relationship between monthly deaths and aviation data.

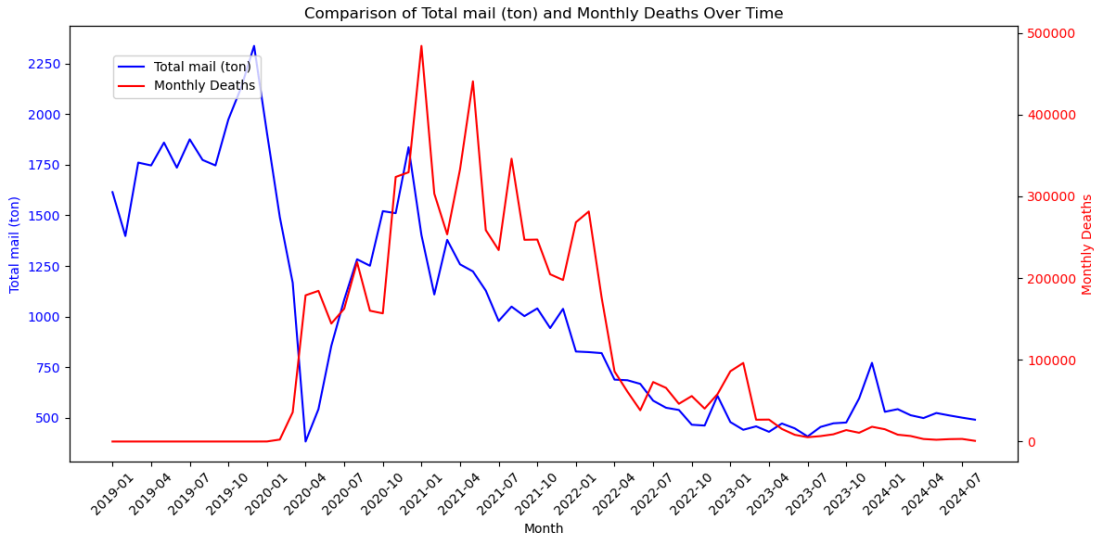


After the outbreak of the pandemic in 2020, monthly deaths surged rapidly, reaching a peak in early 2021. Subsequently, the number of deaths gradually decreased and stabilized by 2022, almost reaching zero. Local flights and cross-

country flights saw significant declines during the outbreak, especially cross-country flights (green line). As monthly deaths decreased in early 2022, cross-country flight numbers began to recover. However, once the death rate approached zero, the fluctuations in flight numbers did not directly correlate with it. Thus, there is no evident direct relationship between monthly deaths and local flights or cross-country flights.

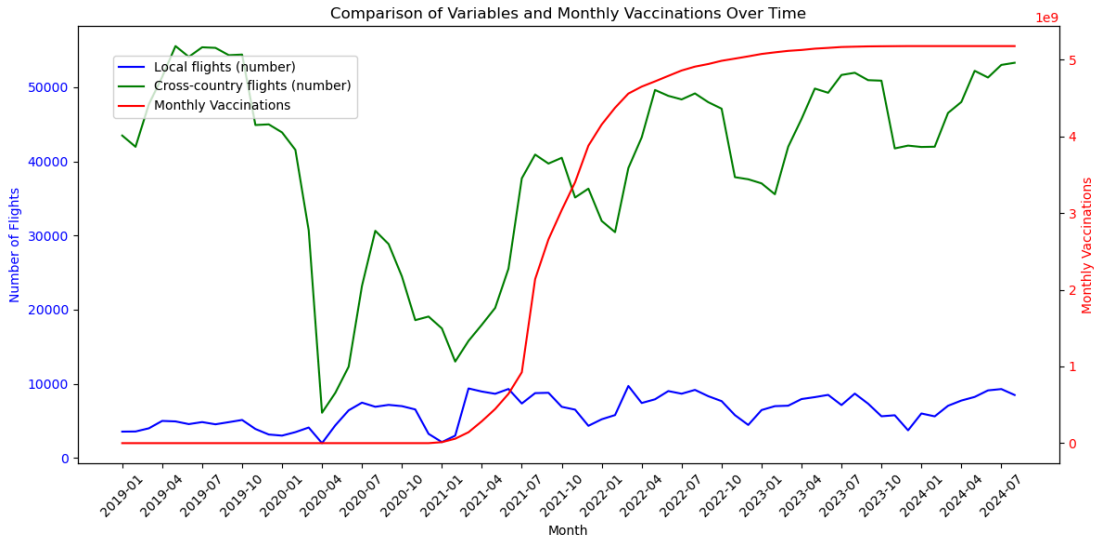




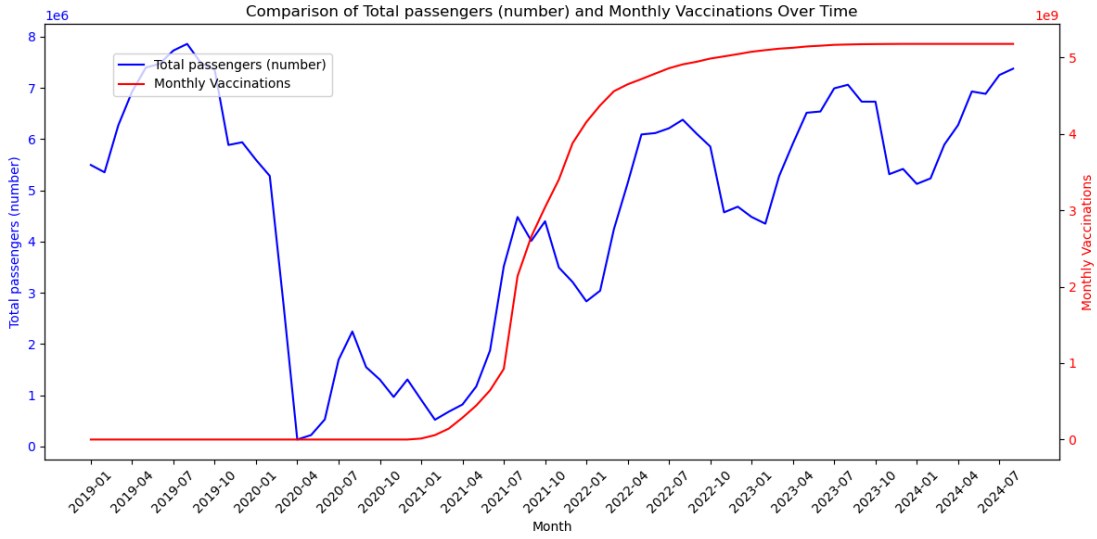


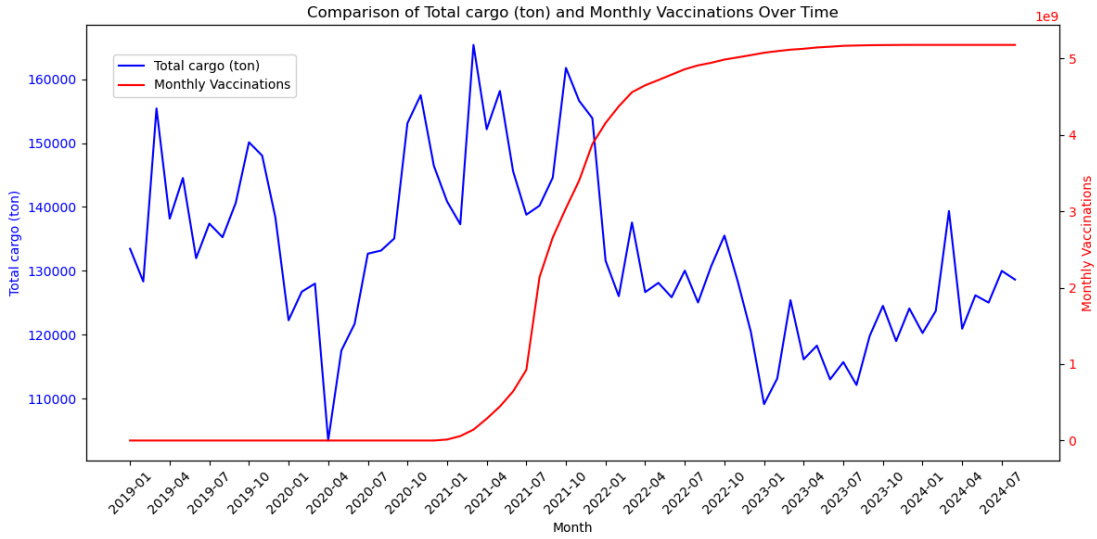
Passenger numbers dropped sharply during the initial COVID-19 outbreak and recovered as deaths decreased, but there is no strong correlation between the two in the later stages. Cargo volume showed smaller fluctuations compared to passenger numbers and remained relatively stable, with no clear relationship to monthly deaths. Mail volume showed larger fluctuations during the initial stages of the pandemic, especially when the number of deaths peaked. However, mail transport gradually returned to normal while monthly deaths dropped and remained stable. In all cases, monthly deaths did not show a clear long-term direct correlation with the aviation variables.

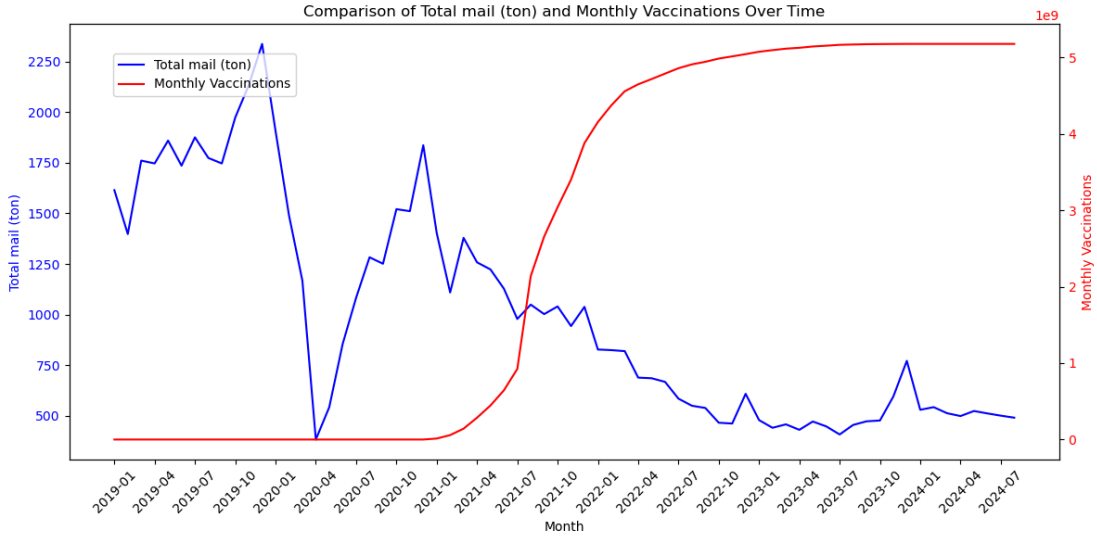
Next part is to analyze the relationship between monthly cumulative vaccinations and aviation data.



Monthly cumulative vaccinations (red line) began to rise rapidly in mid-2021 and continued steadily, reaching a stable level by 2022. This indicates the widespread rollout and coverage of vaccinations. The increase in cumulative vaccinations is somewhat correlated with the recovery of cross-country flights over time because the number of cross-country flights rose during the period of the fastest vaccination growth. However, there is no strong direct relationship. Over the entire pandemic period, cumulative vaccinations and flight numbers show no clear connection.







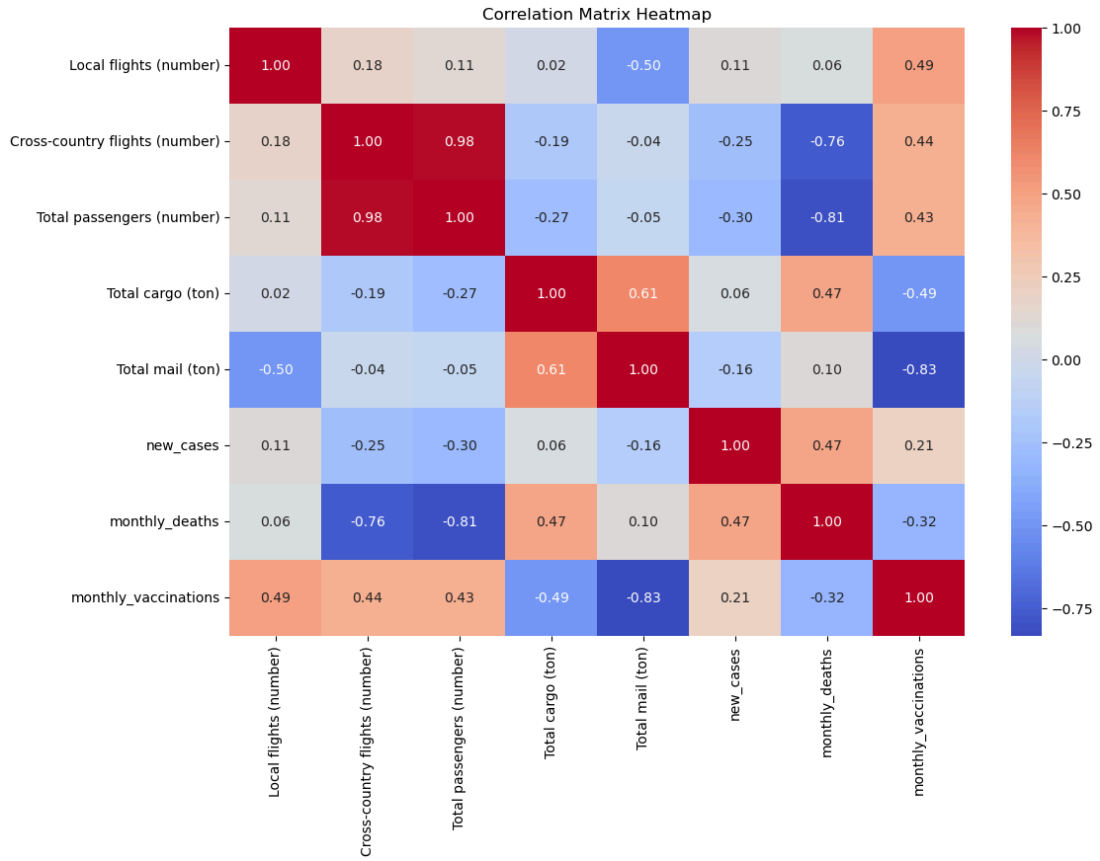
Passenger numbers gradually recovered as vaccinations increased, indicating some correlation. Although passenger recovery showed fluctuations, it generally increased during the period of rising vaccinations. Cargo volume experienced slight fluctuations early in the pandemic, but after vaccinations began, cargo volume remained relatively stable, showing minimal impact from vaccination growth. Mail volume showed some recovery after vaccination growth, the fluctuations remained significant, and no sustained recovery trend was observed.

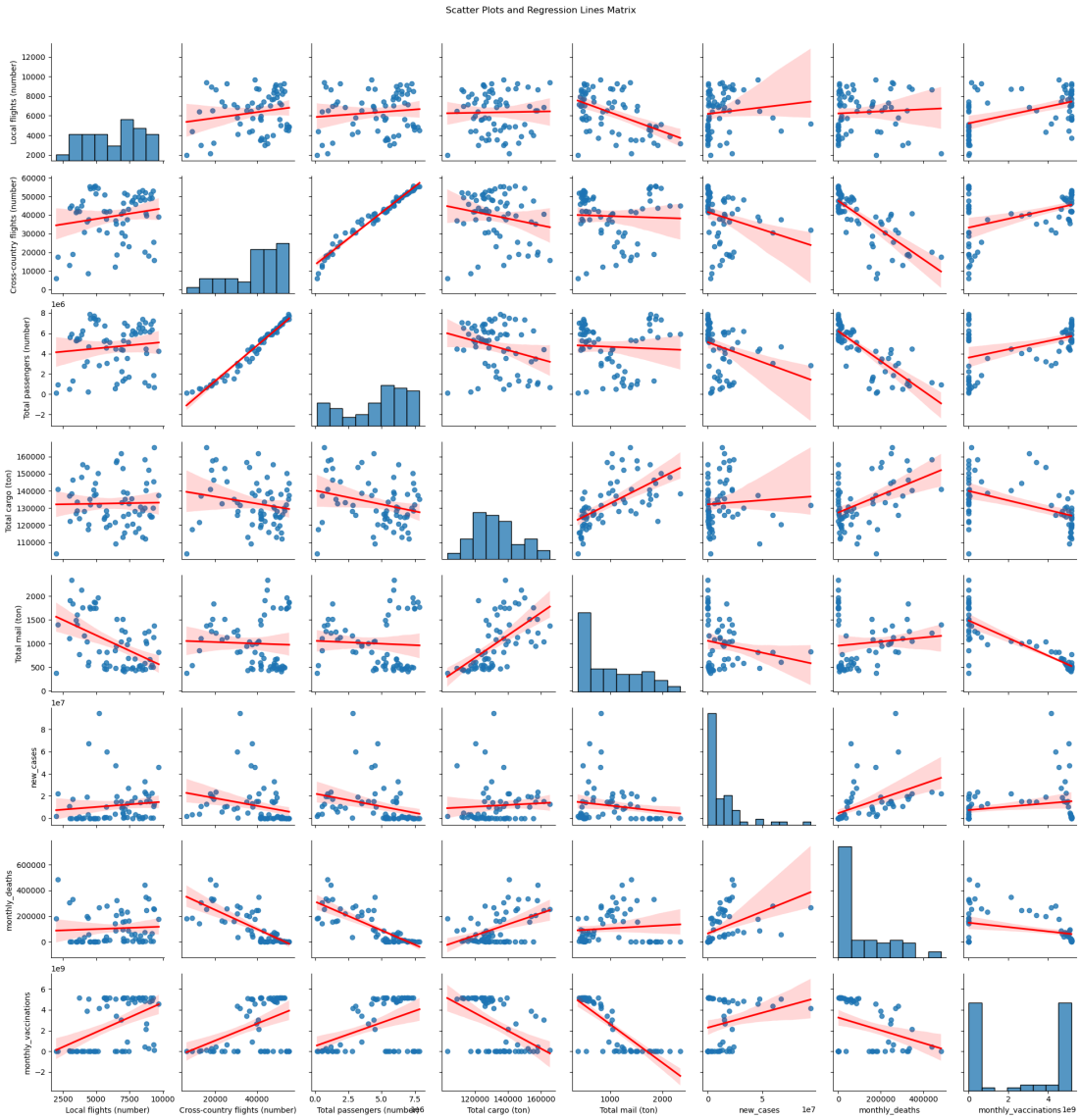
The relationship between vaccinations and mail volume is weak. Vaccinations had an impact on passenger numbers in specific period but overall, the correlation between vaccinations and aviation traffic (including passenger, cargo, and mail transport volumes) is weak.

In conclusion, the new introduced variables show weak correlationship with the air traffic data, next we will try to implement multi-linear regression models to analyze the relationship.

Correlations with new cases deaths and vaccination

In this section, we will incorporate vaccination and COVID-19 death data to explore potential correlations with aviation-related variables. By examining these correlations, we aim to uncover insights into how the pandemic and vaccination efforts have impacted aviation activities. First, we calculate the correlation coefficient between each variable and the number of deaths as well as the number of vaccinations and new cases. The variables analyzed are as follows: local flights, cross-country flights, total passengers, total cargo, and total mail. These coefficients indicate the strength and direction of the linear relationship between each variable and the number of deaths as well as the number of vaccinations.



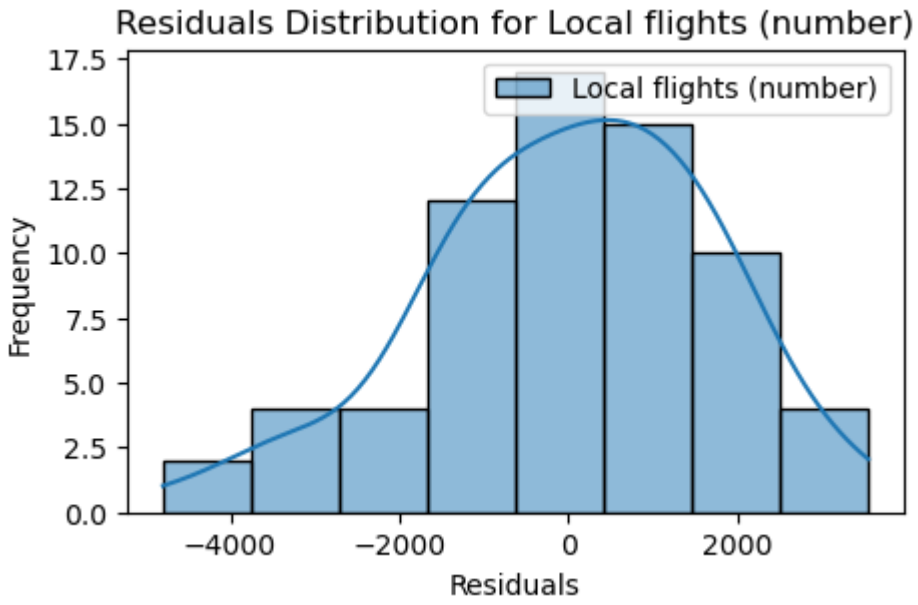


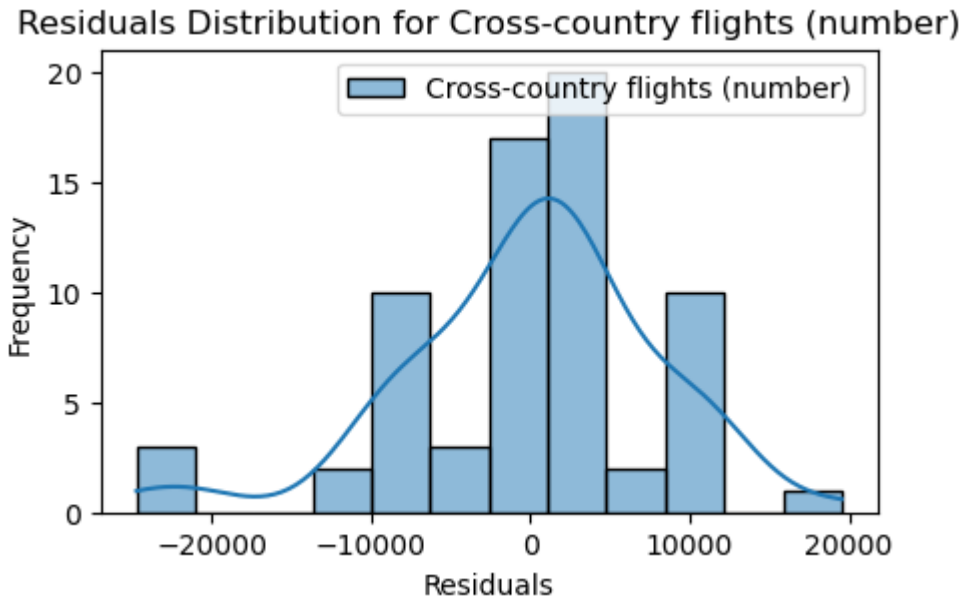
Attempting Multiple Linear Regression

In this section, we will attempt to use a Multiple Linear Regression model to predict various aviation-related variables based on the number of new COVID-19 cases, monthly deaths, and monthly vaccinations. The variables we will analyze include:

- Local flights (number)
- Cross-country flights (number)
- Total passengers (number)
- Total cargo (ton)
- Total mail (ton)

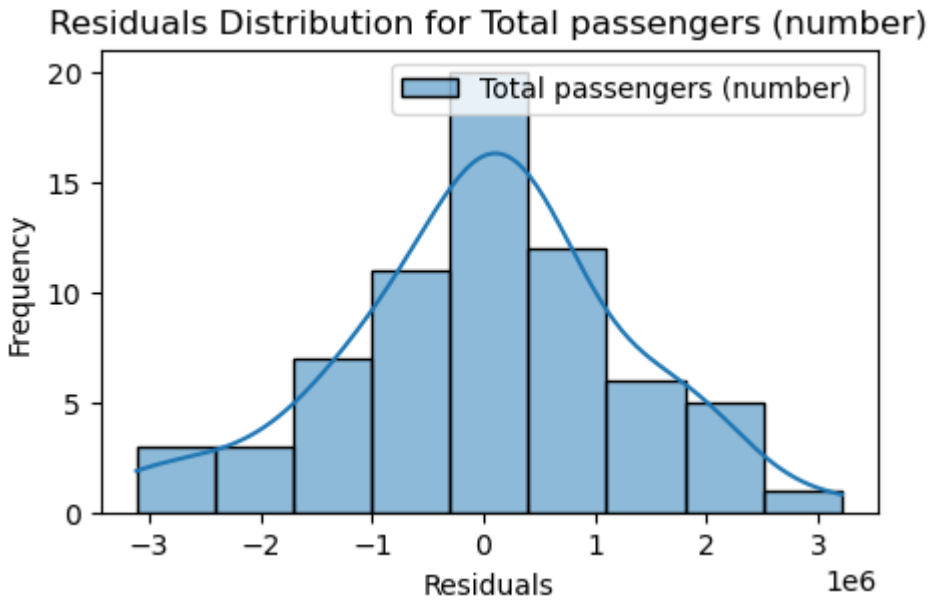
We will split the data into training and testing sets, fit the Multiple Linear Regression model, and evaluate its performance using metrics such as Mean Squared Error (MSE) and R-squared (R^2). Additionally, we will visualize the actual vs. predicted values for each variable.

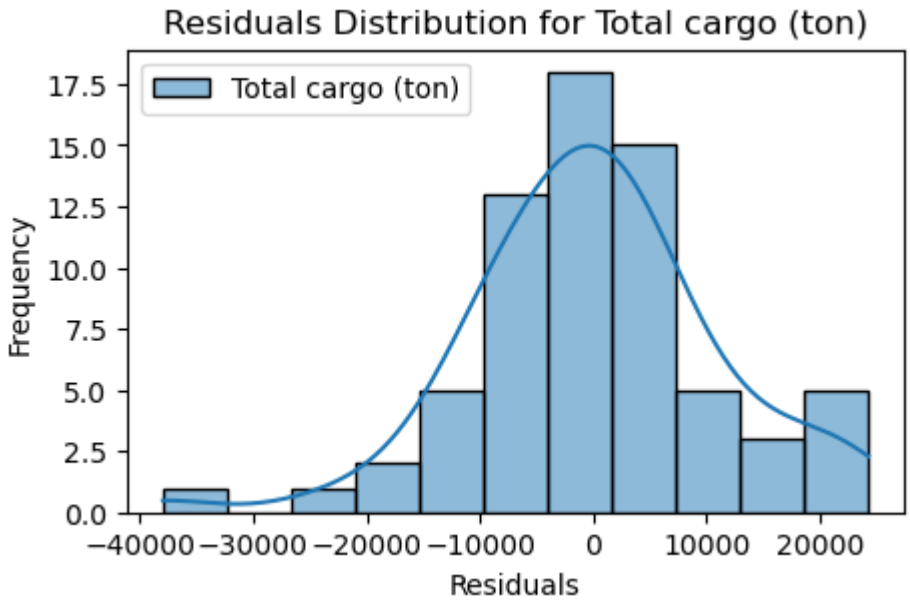


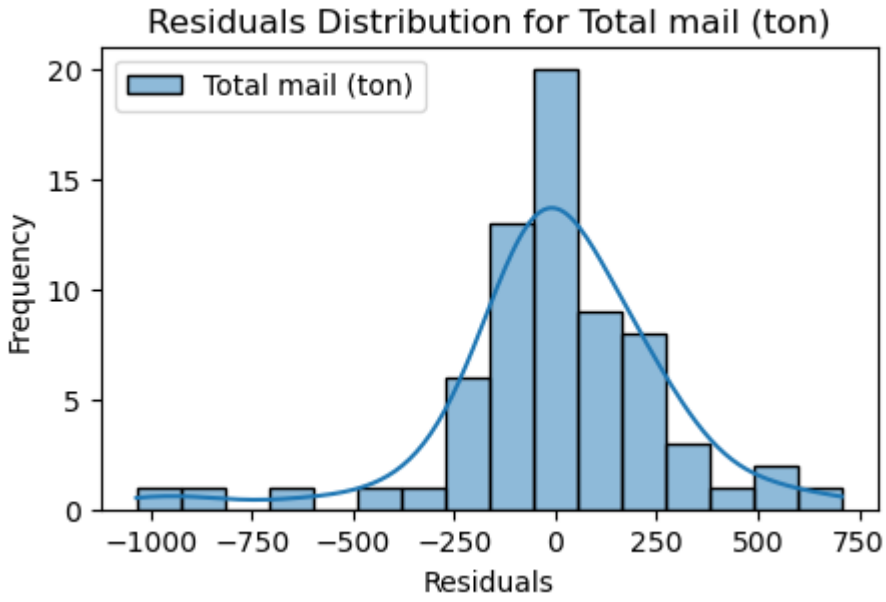


The first graph shows the residuals for the number of local flights and the quantiles of the residuals versus a theoretical normal distribution. Despite the residuals clustering around zero, the spread and presence of outliers suggest that the model may not capture some of the variability in local flights data. The low R^2 (R^2=0.324) also supports the insignificant relationship between local flights number and the variables.

The second graph, along with the R^2 value of 0.621, indicate a stronger relationship between cross-country flights number and COVID-related variables compared to domestic flights. This result is logical considering the Netherlands' significant role in international air traffic.







The first graph illustrates that total passenger numbers are closely related to COVID-related variables, as reflected by an R^2 value of 0.689. This relatively strong

correlation suggests that the pandemic has had a notable impact on passenger aviation, with fluctuations in COVID-19 cases, deaths, and vaccinations likely influencing travel demand.

In comparison, the second figure shows that total cargo has a much weaker relationship with COVID-related indicators, also indicated by an R^2 value of 0.349. This suggests that cargo transport has been less affected by pandemic factors such as case numbers, deaths, and vaccinations. The lower correlation may be due to the essential nature of cargo transport, which likely continued relatively unaffected during the pandemic to meet global supply chain demands, regardless of fluctuations in COVID-19 variables.

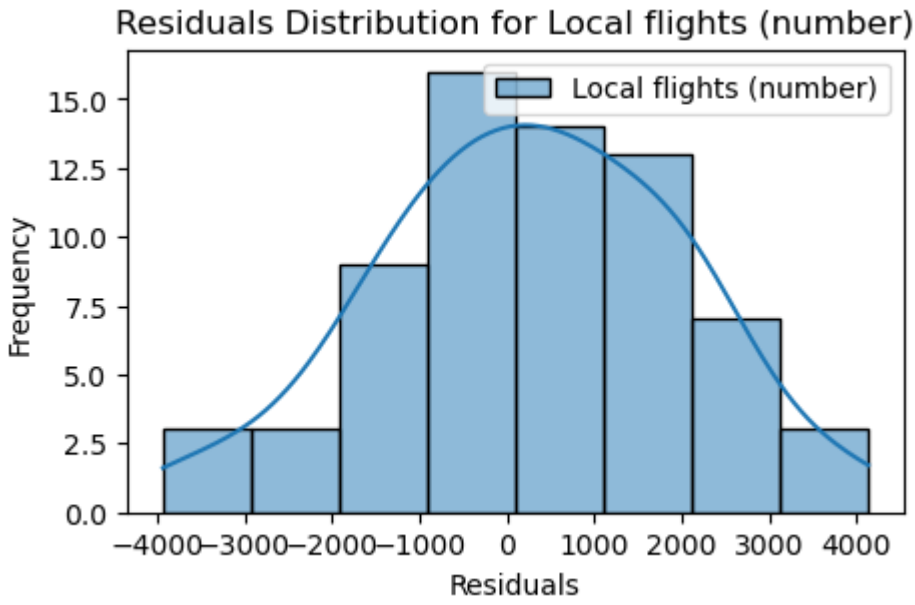
The third graph reveals a significant relationship between total mail and COVID-related variables, demonstrated by an R^2 value of 0.743. This correlation indicates that the pandemic has an influence on mail volumes, likely due to shifts in consumer behavior and increased reliance on delivery services during lockdowns.

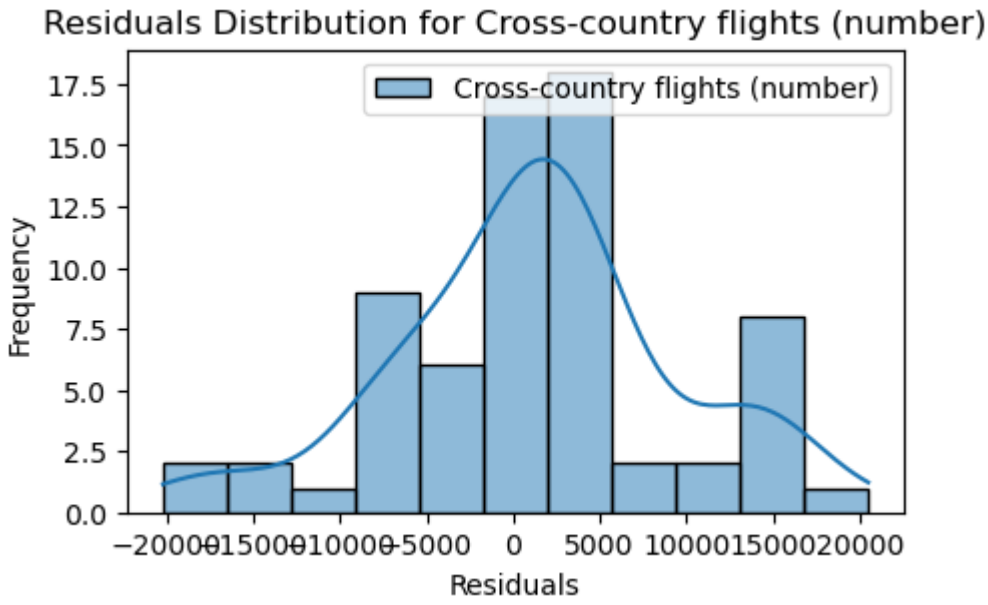
Summary

While the multiple linear regression model performs reasonably well for predicting some variables, it falls short in capturing the relationships for others. This suggests that a direct linear approach may not fully reflect the complexities within the data. To address this, we plan to apply some transformation to the variables and investigate whether this approach can reveal stronger or more linear relationships.

Attempting Multiple Linear Regression with Log Transformation

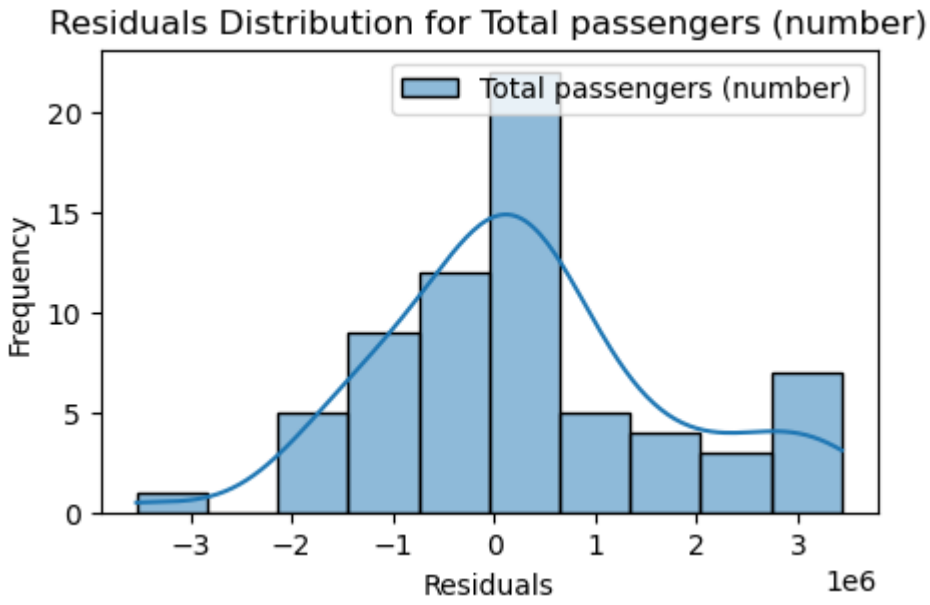
In this section, we apply a log transformation to the variables to see whether this approach can reveal stronger or more linear relationships. By transforming the data, we aim to enhance the model's accuracy and better account for potential nonlinearities, leading to more robust predictions and deeper insights.

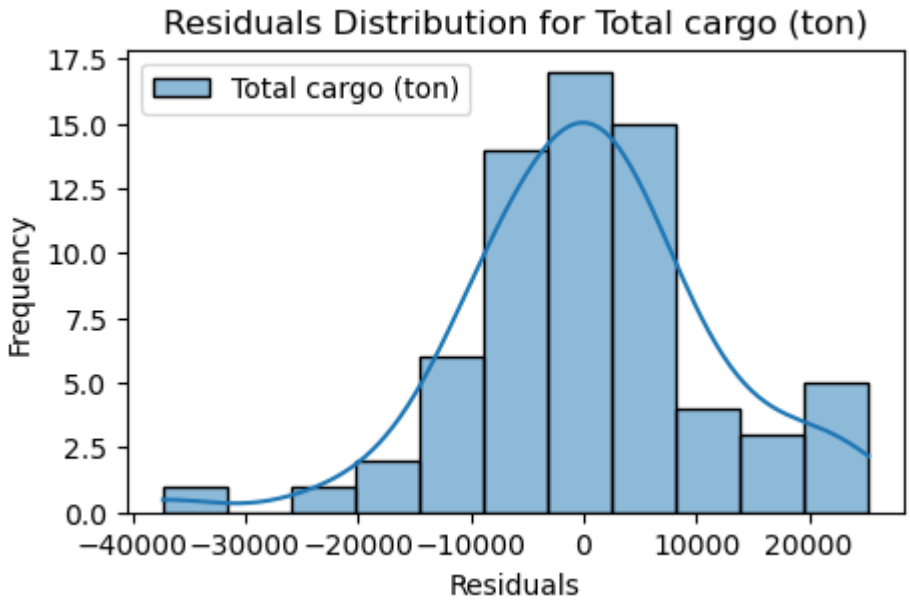


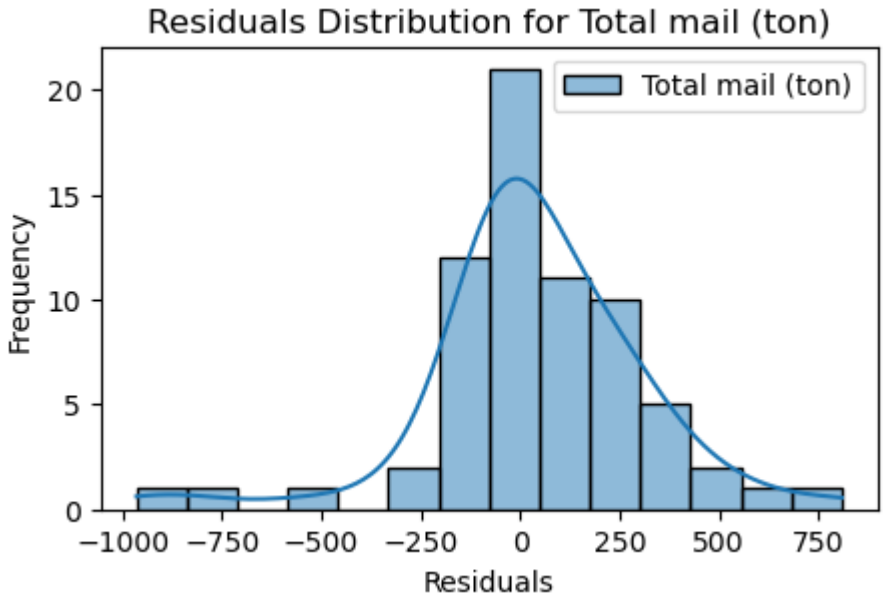


In the first graph, the log transformation does not appear to improve the relationship between local flights and the variables. This conclusion is supported by a low R^2 value of 0.305, indicating that the log transformation fails to uncover a stronger connection. The result suggests a weak or nonexistent relationship between local flights and the COVID-related factors.

Similarly, the log transformation does not appear to enhance the relationship between cross-country flights and the variables. This is indicated by the relatively low R^2 value of 0.528, which is lower than that of the original model. This finding suggests that the transformation failed to improve the model's ability to capture the underlying relationship, highlighting the limitations of both the original and transformed models in this context.







The same situation applies to the relationships between total passengers and total cargo with the variables. The R^2 values are 0.561 and 0.339, respectively, both

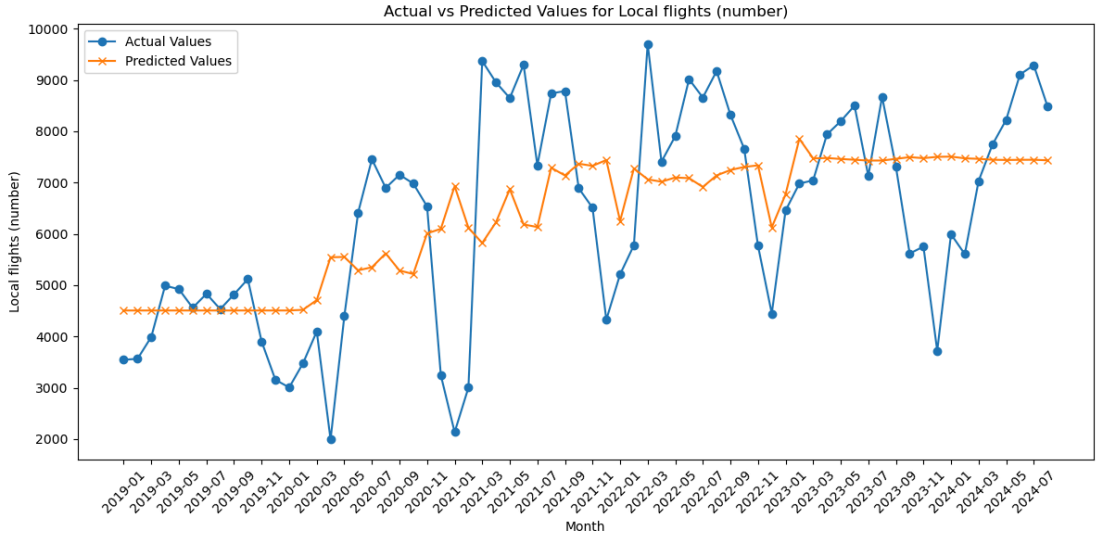
of which are lower than those of the original models. This indicates that the log transformation did not improve the model's ability to capture the underlying relationships between total passengers, total cargo, and the COVID-related factors.

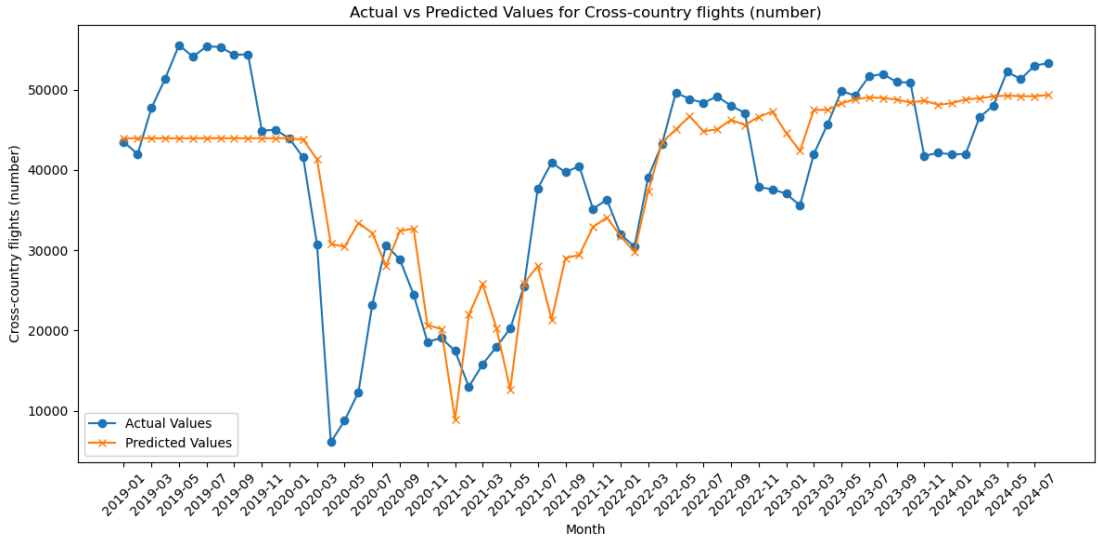
The relationship between total mail and the variables stands as the only exception. With an R^2 value of 0.750, which is slightly higher than that of the original model, this improvement may be attributed to the close relationship between total mail and the COVID-related factors.

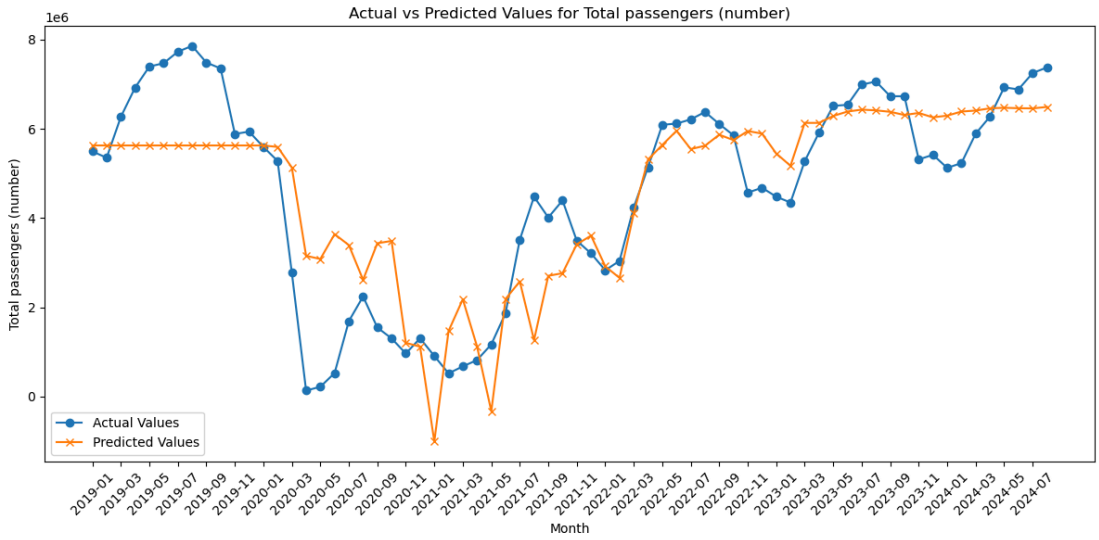
Summary

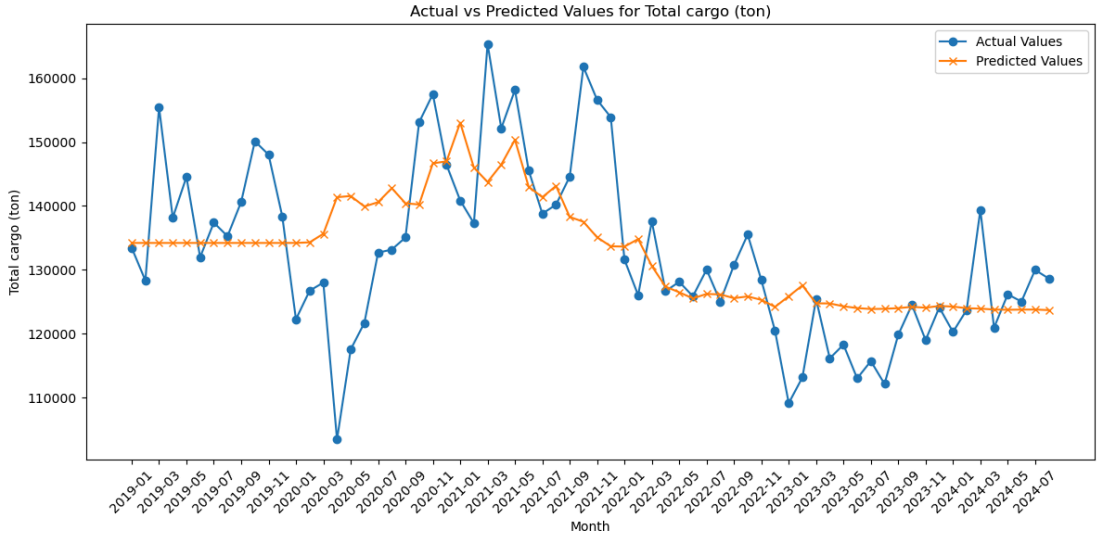
Based on these results, the log transformation does not lead to any improvement in the model's performance. As a result, we have decided to discontinue its use. Finally, we present the actual vs. predicted value graphs for the five indicators, illustrating the model's predictions compared to the observed data for each variable.

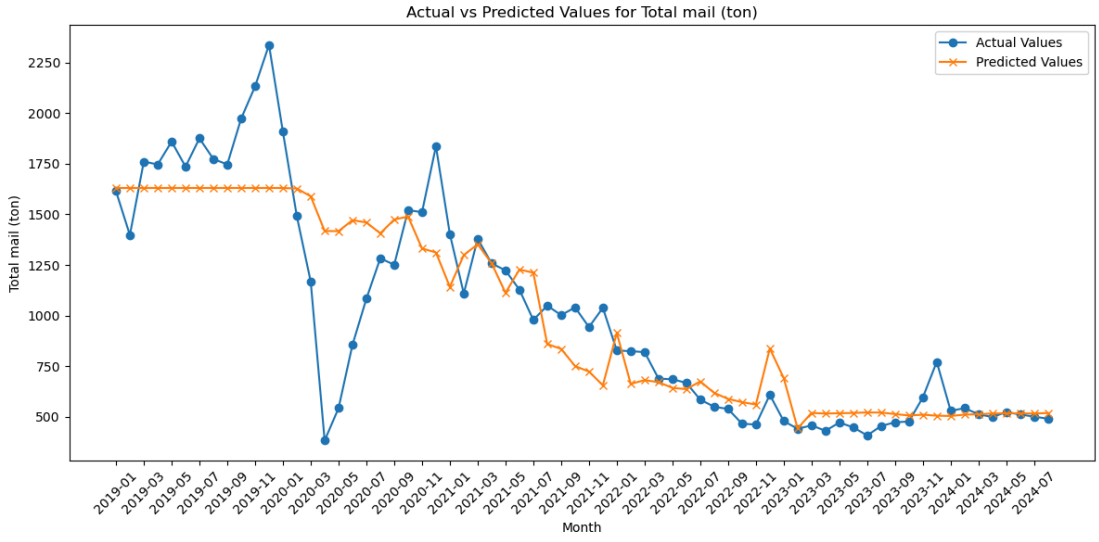
Actual vs. Predicted Value Graphs











The findings in this section reveal that while multiple linear regression offers some insights to predict aviation indicators, it does not accurately capture the relationship with COVID-related variables. To address this limitation, we plan to apply alternative predictive models, such as random forest model and time series analysis, to further investigate the connection between aviation performance and the pandemic. This change in methodology aims to provide a more

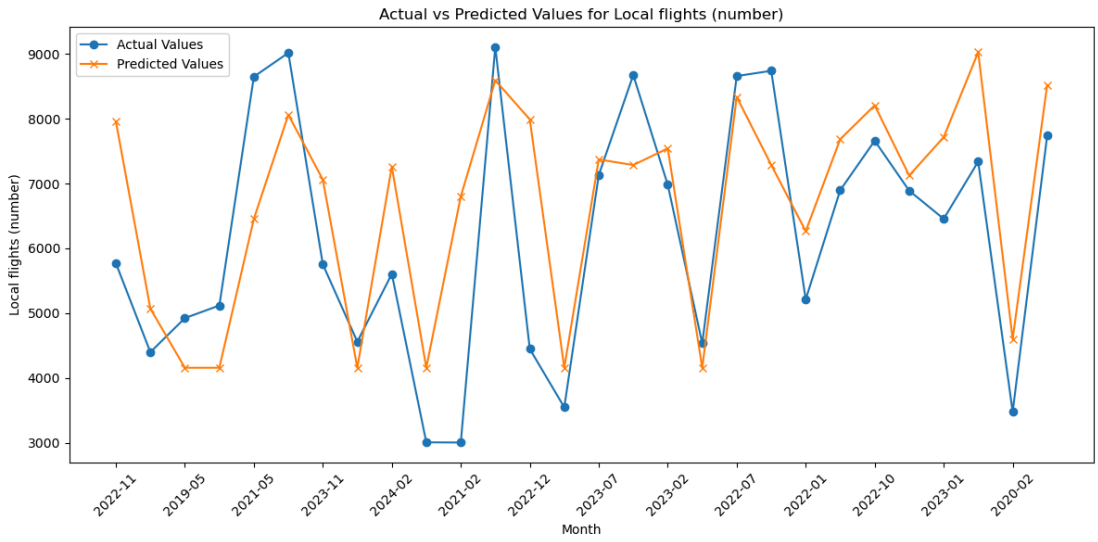
comprehensive understanding of the relationships and improve prediction accuracy.

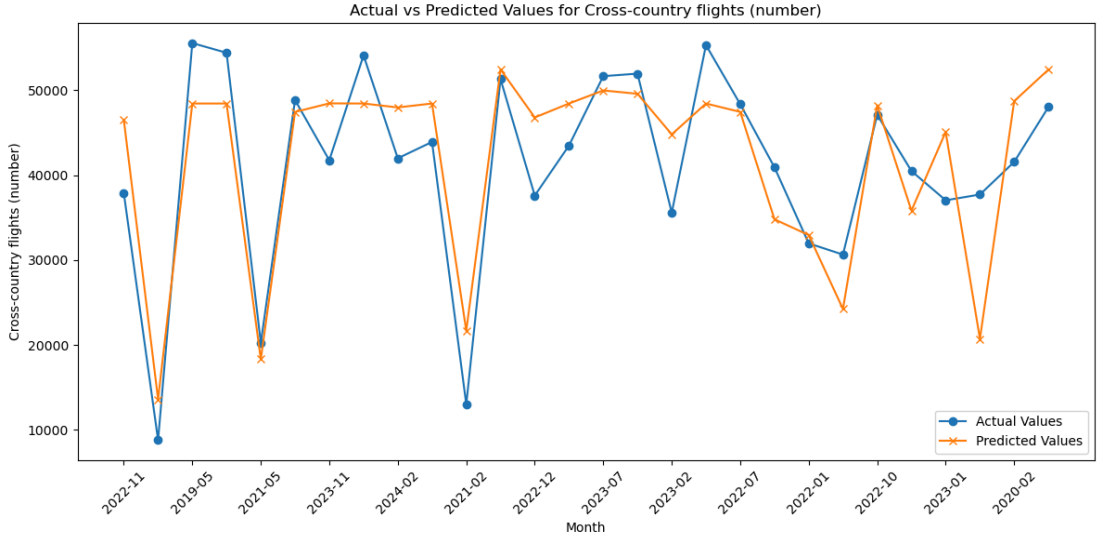
Attempting Random Forest Model

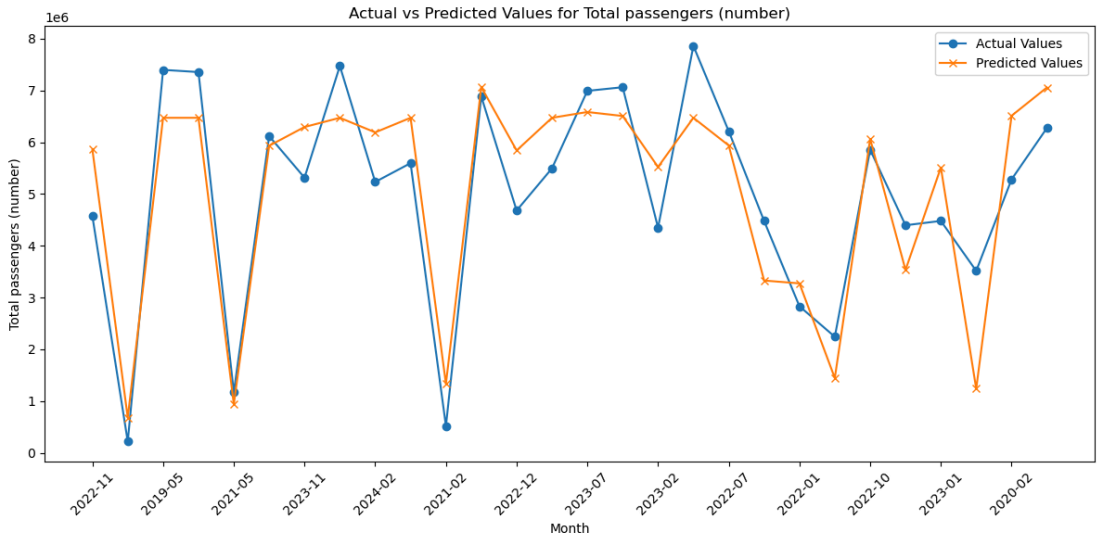
In this section, we will attempt to use a Random Forest model to predict various aviation-related variables based on the number of new COVID-19 cases. The variables we will analyze include:

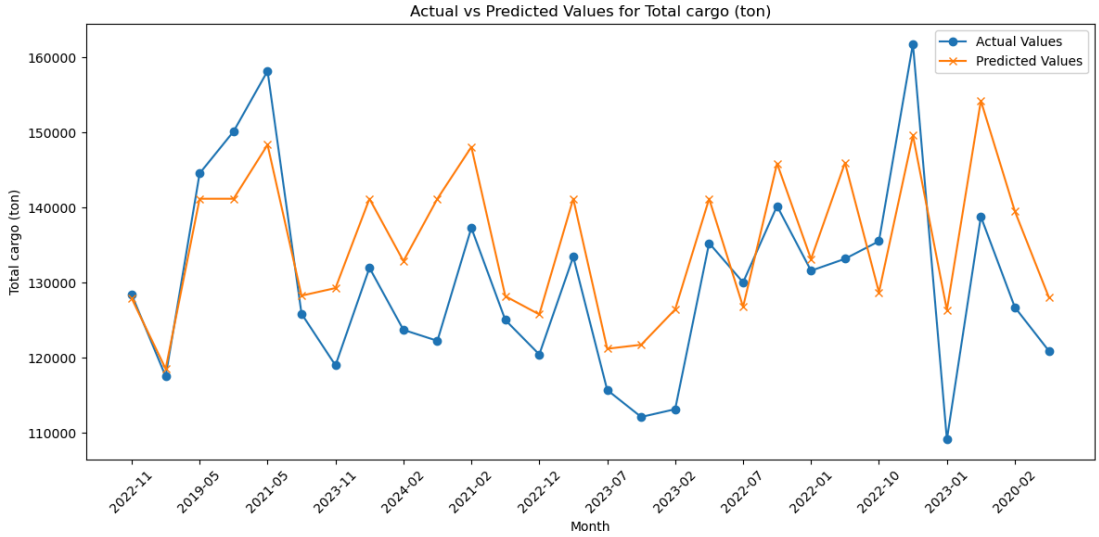
- Local flights (number)
- Cross-country flights (number)
- Total passengers (number)
- Total cargo (ton)
- Total mail (ton)

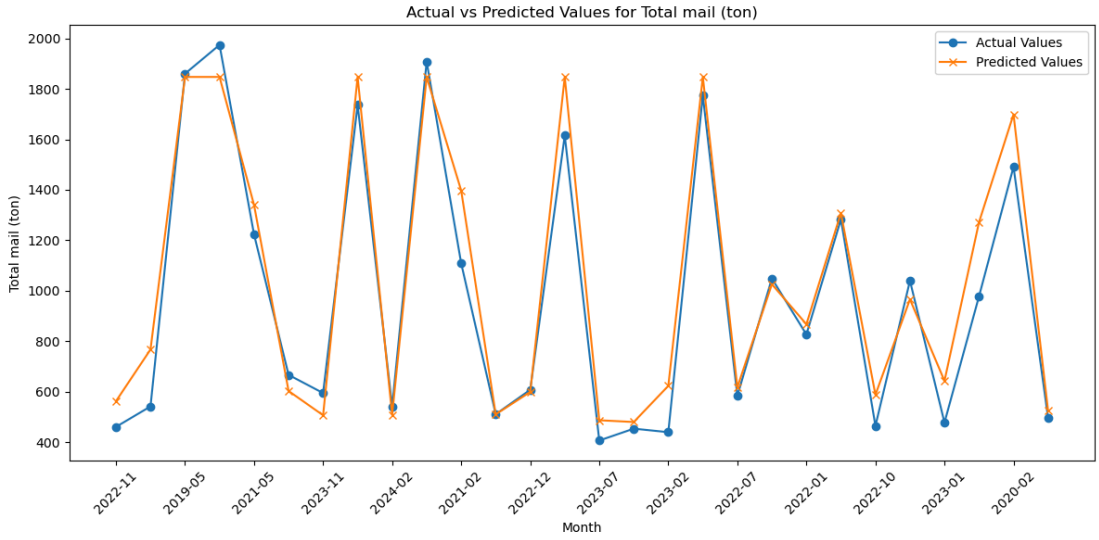
We will split the data into training and testing sets, fit the Random Forest model, and evaluate its performance using metrics such as Mean Squared Error (MSE) and R-squared (R^2). Additionally, we will visualize the actual vs. predicted values for each variable.











The Random Forest model shows an R^2 value of 0.6337, indicating that the model explains about 63.37% of the variance. While it demonstrates some predictive ability, it is not highly accurate. The model's MSE is 215,993,945,235.1363, which is quite large. Although it can be explained by the large scale of the data, it signifies high prediction error on the test set. These results suggest the model captures some patterns, but the three pandemic variables used are insufficient to fully explain the fluctuations in the aviation industry.

Feature importance analysis shows that `monthly_deaths` has the highest influence, with an importance score of 0.8393, indicating that death counts significantly affect aviation variables. In contrast, `new_cases` has a much smaller impact, with an importance of only 0.0289.

Visual comparisons reveal that the model performs well during certain periods, but shows larger discrepancies in others. To improve performance, it may be necessary to include additional variables, such as lockdown policy indices, vaccination coverage, and mobility data, to better capture the factors influencing aviation trends.

Time Series Analysis

Steps:

Stationarity Testing: - Use tests like the Augmented Dickey-Fuller (ADF) test or the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test to check the stationarity of the time series. - If the time series is not stationary, make it stationary through differencing, log transformation, etc.

Model Selection and Training: - Choose an appropriate time series model, such as ARIMA, SARIMA, Holt-Winters, etc. - Fit the model using training data and adjust model parameters for the best fit.

Model Evaluation: - Evaluate the model's predictive performance using test data. - Calculate evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc.

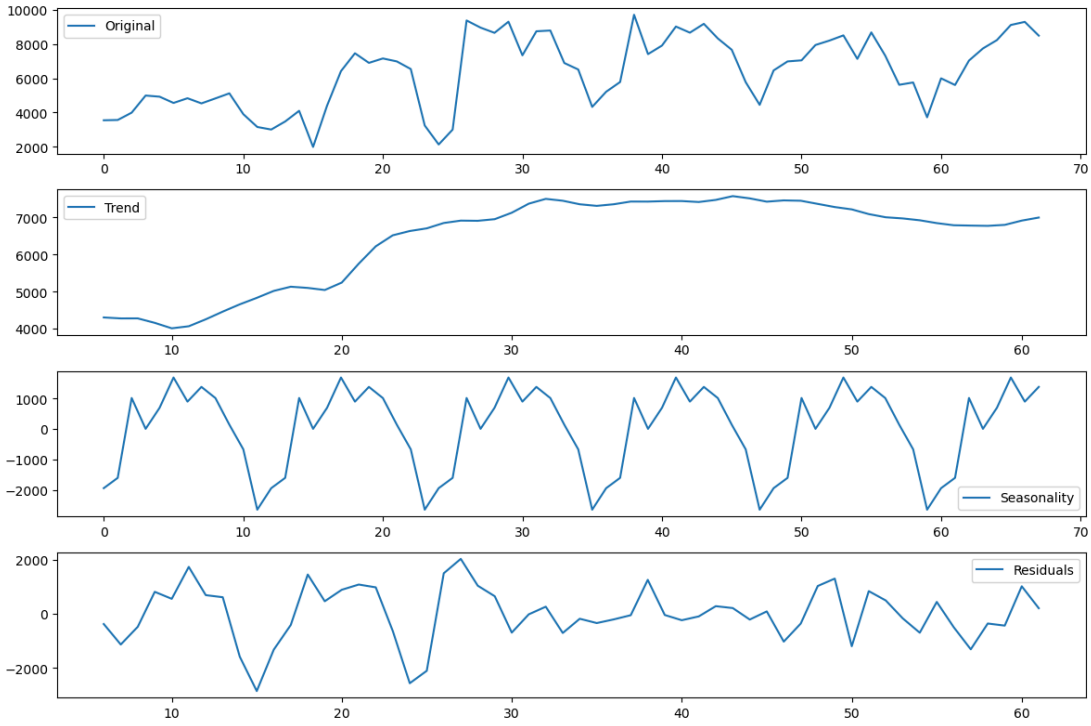
Forecasting and Validation: - Use the trained model to forecast future data. - Compare the forecasted results with actual data to validate the model's accuracy.

Local flight

This time series analysis focuses on local flight volume data, employing various methods including time series decomposition, stationarity tests, SARIMAX model fitting, and performance evaluation. Below is a detailed analysis of the results and potential improvements.

Time Series Decomposition

The original data was decomposed into trend, seasonal, and residual components.



- Trend Component: The data initially exhibited an upward trend, followed by stabilization between time points 30 and 40, likely due to supply constraints or market saturation.
- Seasonal Component: The seasonal component revealed periodic fluctuations in flight volume, such as peaks between time points 10 and 20, possibly related to holiday travel demand.
- Residual Component: Significant residual deviations were observed between time points 20 and 30, suggesting that the model did not fully capture the data dynamics, potentially indicating unconsidered exogenous variables or nonlinear characteristics.

Stationarity Test

Stationarity was tested using ADF and KPSS tests

```
ADF Test for monthly_deaths: (-2.4366970042563367, 0.13164563170113136, 9, 58, {'1%': -3.548493559596539, '5%': -2.912836594776334, '10%': -2.
594129155766944}, 1393.5546896908131)
KPSS Test for monthly_deaths: (0.292961883275794, 0.1, 5, {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739})
monthly_deaths is not stationary. Differencing applied.
                               SARIMAX Results
==============================================================================
Dep. Variable:         monthly_deaths   No. Observations:                   67
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -820.456
Date:                Wed, 09 Oct 2024   AIC                           1646.913
Time:                        11:41:59   BIC                           1653.482
Sample:                             0   HQIC                          1649.509
                                  - 67
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1251      0.145     -0.863      0.388      -0.409       0.159
ma.L1         -0.9750      0.141     -6.915      0.000      -1.251      -0.699
sigma2      4.375e+09   1.64e-11   2.67e+20      0.000    4.38e+09    4.38e+09
===================================================================================
Ljung-Box (L1) (Q):                   0.16   Jarque-Bera (JB):                18.24
Prob(Q):                              0.69   Prob(JB):                         0.00
Heteroskedasticity (H):               0.12   Skew:                            -0.24
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.53
===================================================================================
```

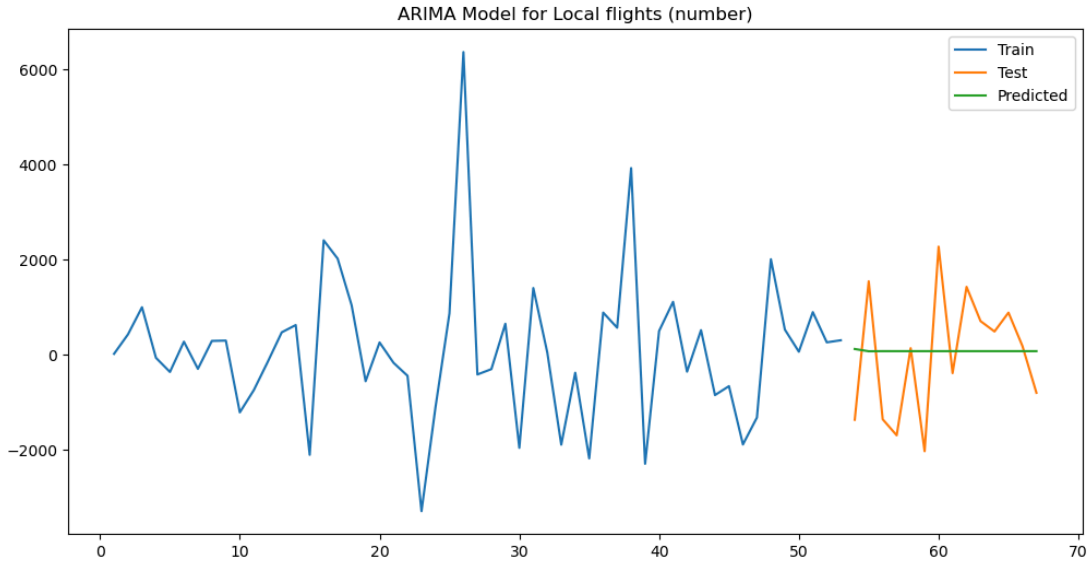
markdown

ADF and KPSS tests indicated that the original data was non-stationary, requiring differencing. After first-order differencing, stationarity improved significantly,

though residuals still exhibited some structural bias, suggesting the consideration of higher-order differencing or other transformation methods.

SARIMAX Model Evaluation

The SARIMAX model was used to fit the data, employing an ARIMA(1,1,1) structure for feature capture.



- Autoregressive Term (AR): The high p-value suggests that the AR term is not significant. It may be advisable to simplify the model by removing this component to reduce complexity and prevent overfitting.
- Moving Average Term (MA): The MA term had a significant p-value, indicating its effectiveness in reducing short-term noise and improving prediction reliability.
- Model Selection Metrics (AIC and BIC): The AIC and BIC values were relatively high, suggesting that parameter optimization via grid search could further improve model fitting and prediction performance.

Model Prediction Performance Analysis

Comparison of Prediction Results with Actual Data

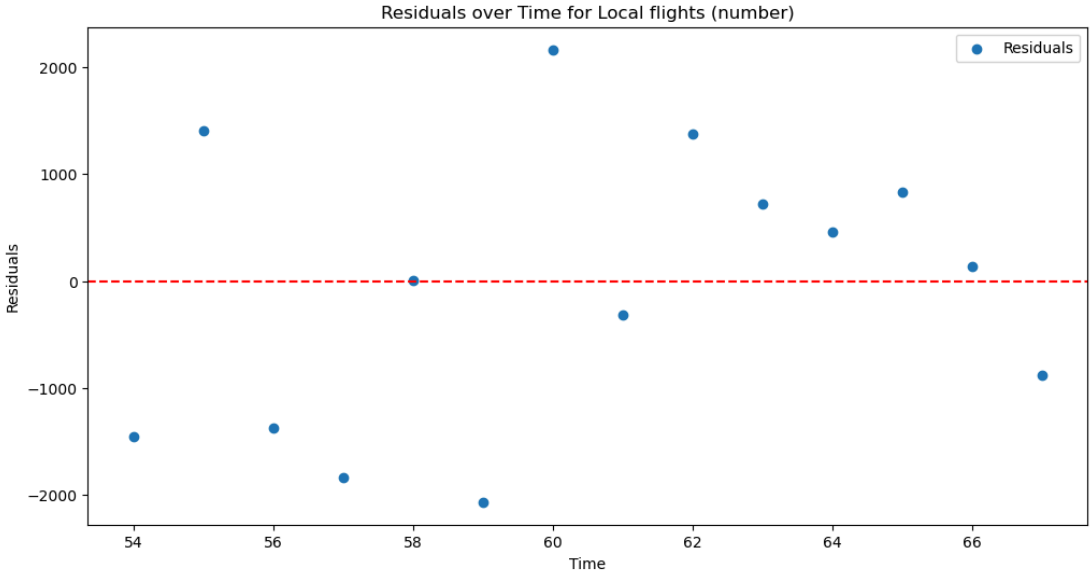
Test set performance was suboptimal, particularly between time points 50 and 60, where the model failed to capture rapid data fluctuations.

- Prediction Curve Smoothness: To improve responsiveness to short-term fluctuations, it may be worthwhile to increase model flexibility by incorporating additional MA or AR terms or including exogenous variables.
- Impact of External Variables: Flight volume is influenced by factors such as fuel prices, weather, and policy changes. Including these factors to construct a multivariate time series model could enhance predictive accuracy.

Directions for Improving Prediction Performance

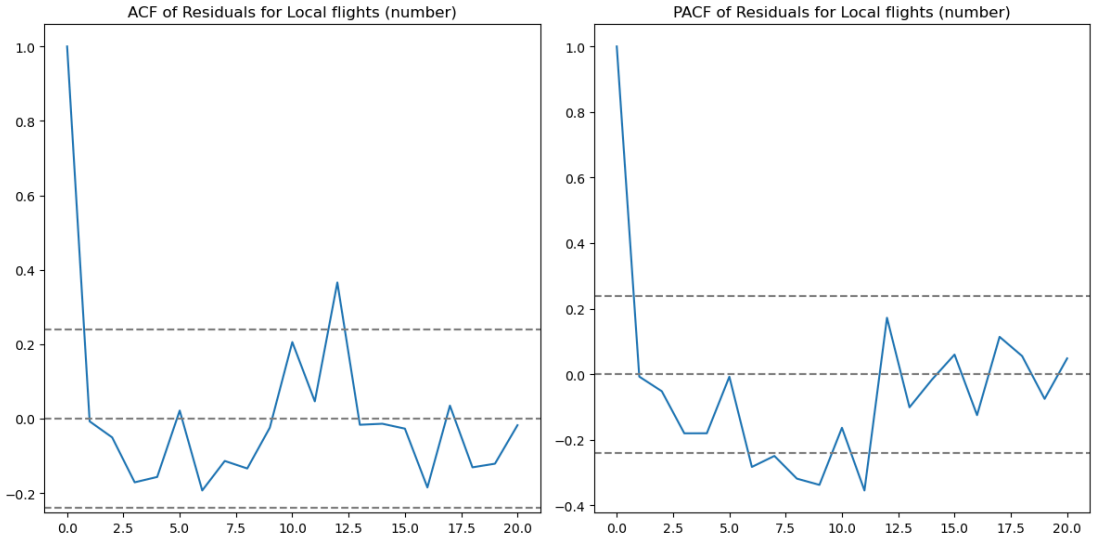
- More Complex Models: Consider employing higher-order ARIMA or nonlinear models such as LSTM to address high volatility.
- Feature Engineering: Extract additional features related to seasonality and external variables to enhance prediction accuracy.

Residual Analysis



Residual Distribution

Ideally, residuals should be randomly distributed with a mean close to zero. However, at specific time points (e.g., 55 and 65), residuals significantly deviated from zero.



Autocorrelation Analysis

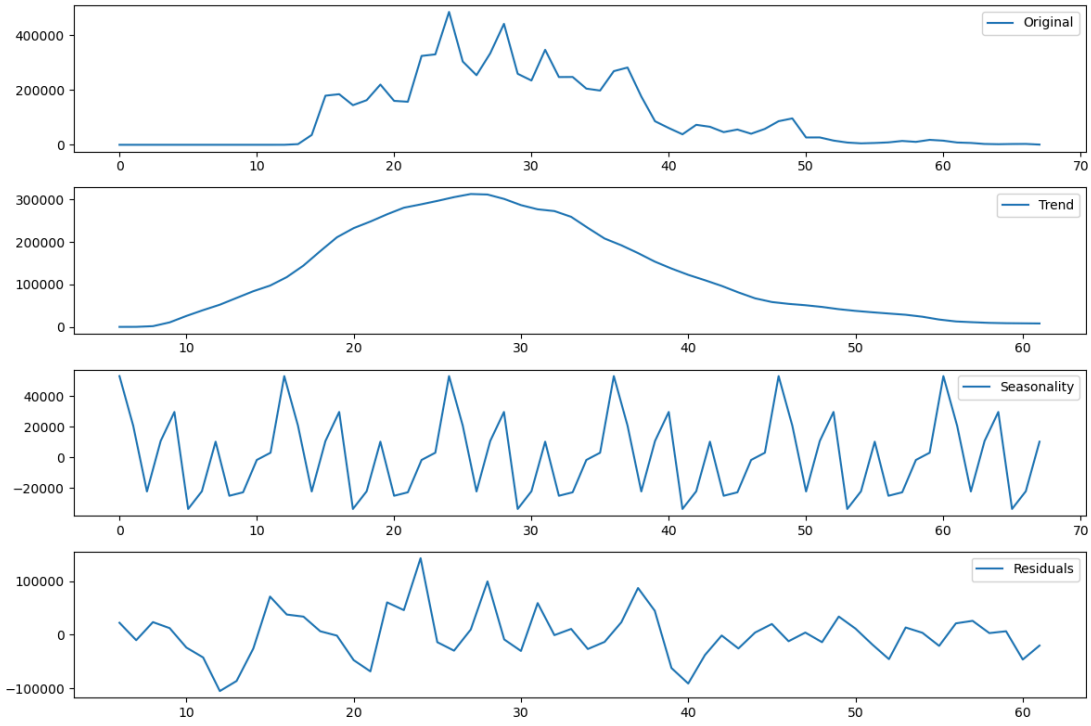
ACF and PACF: Significant autocorrelations were observed at multiple lags, suggesting the addition of lag terms or the inclusion of more explanatory variables.

Monthly death

This analysis focuses on the time series of monthly death counts, using methods such as time series decomposition, stationarity testing, SARIMAX model fitting, and prediction performance evaluation. The following sections present a concise evaluation of the results and potential improvements.

Time Series Decomposition

The monthly death data was decomposed into trend, seasonal, and residual components.



- Trend Component: The trend shows an increase in deaths, peaking around time point 30, then declining beyond time point 40, reflecting the epidemic's progression and possible interventions.
- Seasonal Component: Seasonal fluctuations indicate periods with consistently higher death counts, likely linked to seasonal factors like increased winter vulnerabilities.
- Residual Component: Significant residual fluctuations suggest exogenous factors or noise not captured by the trend and seasonal components, particularly between time points 20 to 30.

Stationarity Test

Stationarity was tested using ADF and KPSS tests

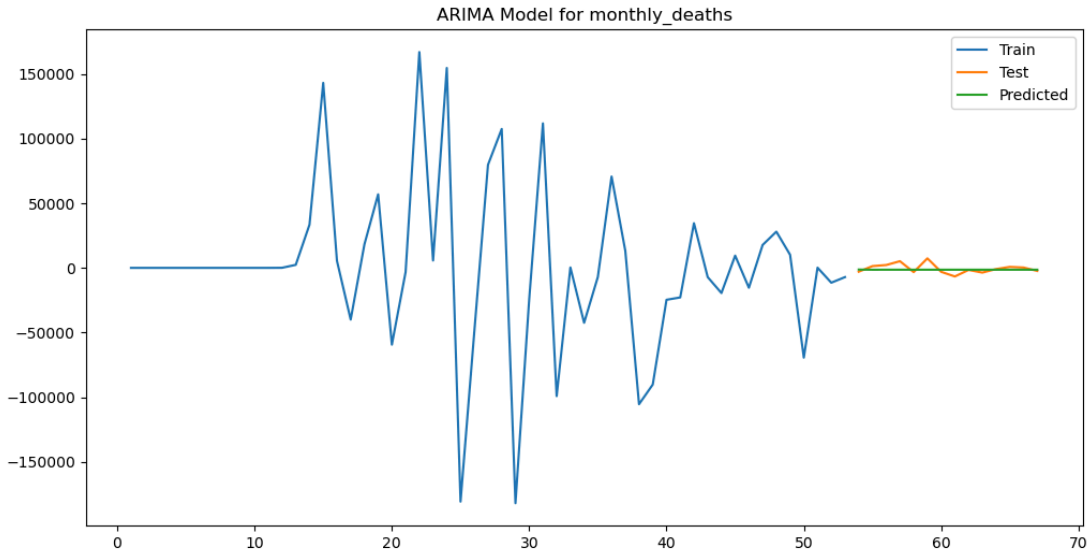
```
ADF Test for Local flights (number): (-3.177268146622313, 0.021338773288993453, 0, 67, {'1%': -3.5319549603840894, '5%': -2.905755128523123,
'10%': -2.5903569458676765}, 978.304488428362)
KPSS Test for Local flights (number): (0.7383547264004816, 0.010058661236319853, 4, {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739})
Local flights (number) is not stationary. Differencing applied.
                               SARIMAX Results
==============================================================================
Dep. Variable:     Local flights (number)   No. Observations:                   67
Model:                   ARIMA(1, 1, 1)   Log Likelihood                -577.625
Date:                  Wed, 09 Oct 2024   AIC                           1161.251
Time:                          11:41:53   BIC                           1167.820
Sample:                               0   HQIC                          1163.847
                                   - 67
Covariance Type:                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0519      0.145     -0.357      0.721      -0.337       0.233
ma.L1         -1.0000      0.108     -9.300      0.000      -1.211      -0.789
sigma2      2.216e+06   4.85e-08   4.57e+13      0.000    2.22e+06    2.22e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                55.33
Prob(Q):                              0.94   Prob(JB):                         0.00
Heteroskedasticity (H):               1.05   Skew:                             1.11
Prob(H) (two-sided):                  0.92   Kurtosis:                         6.89
===================================================================================
```

markdown

The ADF test indicates non-stationarity. The KPSS test supports this. First-order differencing improved stationarity, though higher-order differencing or alternative transformations might help further.

SARIMAX Model Evaluation

The SARIMAX model was used to fit the data, employing an ARIMA(1,1,1) structure for feature capture.



- Autoregressive Term (AR): The AR term's high p-value suggests it may not be significant, and removing it could reduce overfitting.
- Moving Average Term (MA): The MA term is significant and helps capture short-term irregularities.
- Model Selection Metrics (AIC and BIC): The high AIC and BIC values suggest room for optimization, possibly through grid search.

Model Prediction Performance Analysis

Comparison of Prediction Results with Actual Data

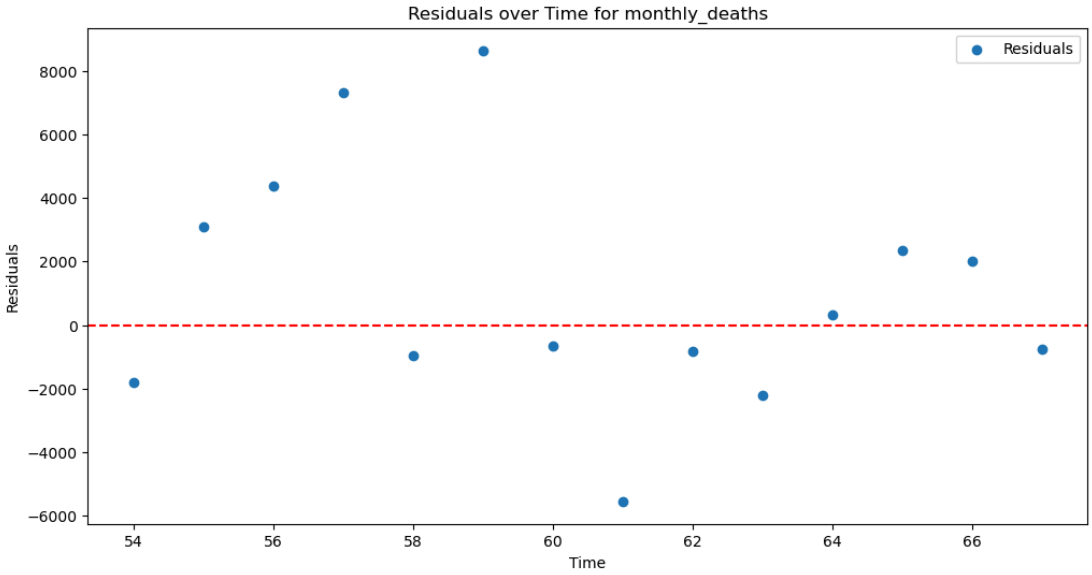
The model's predictive performance is suboptimal, particularly during the testing phase, where predictions are too flat and fail to capture actual volatility.

- Prediction Curve Smoothness: The lack of sensitivity to fluctuations suggests adding more AR or MA terms or including exogenous regressors.
- External Variables Impact: Including variables like healthcare capacity and policy interventions could improve accuracy, especially during extreme changes.

Directions for Improving Prediction Performance

- Model Complexity: Advanced models like higher-order ARIMA or LSTM could capture complex temporal patterns better.
- Feature Engineering: Include features such as temperature, healthcare utilization, and population movement.

Residual Analysis



Residual Distribution Residuals show systematic deviations between time points 54 and 64, indicating that the model has not fully captured underlying patterns.

Autocorrelation Analysis ACF and PACF: Significant autocorrelations suggest adding lagged variables or using higher-order differencing.

Conclusion

In this study, we analyzed the impact of the COVID-19 pandemic on various aspects of air traffic in the Netherlands, including cross-country flights, local flights, passenger numbers, cargo volume, and mail volume. Our analysis revealed several key insights:

1. **Impact of COVID-19 on Air Traffic**:
    - Both cross-country and local flights experienced significant declines during the early stages of the pandemic, with cross-country flights being more severely affected.
    - Passenger numbers plummeted sharply in early 2020 but showed a gradual recovery aligned with the easing of travel restrictions and the rollout of vaccinations.
    - Cargo and mail volumes also experienced fluctuations, though cargo was less affected compared to passenger traffic, likely due to the essential nature of cargo transport.
2. **Correlation with COVID-19 Variables**:
    - The correlation analysis indicated a weak relationship between air traffic variables and COVID-19 new cases, deaths, and vaccinations. This suggests that factors other than the direct progression of the pandemic, such as government policies and travel restrictions, played a more significant role in influencing air traffic trends.
3. **Modeling**:
    - Multiple linear regression models provided some insights but were limited in capturing the complexities of the data. The introduction of log transformations did not significantly improve model performance.
    - The Random Forest model showed moderate predictive ability, with monthly deaths having the highest influence on aviation variables.
    - Time series analysis, including SARIMAX models, highlighted the importance of considering external variables and more complex modeling techniques to improve prediction accuracy.
4. **Future Directions**:
    - To enhance the predictive models, incorporating additional variables such as lockdown policy indices, mobility data, and other exogenous factors is essential.
    - Exploring advanced modeling techniques, including nonlinear models like LSTM and multivariate time series models, could provide deeper insights and more accurate predictions.

Overall, this study underscores the multifaceted impact of the COVID-19 pandemic on air traffic and highlights the need for comprehensive modeling

approaches to understand and predict aviation trends in the context of global disruptions.

Contribution statement

Xinyu Yang:

- Define research question.
- Find and select available data for RQ1.
- Result analysis and report writing for RQ1 and RQ3: random forest analysis.

Yilin Shi:

- Define research question.
- Find and select available data for RQ2.
- Result analysis and report writing for RQ2 and RQ3: relationship analysis of newly introduced variables.

Yue Guo:

- Define research question.
- Proposal writing.
- Find and select available data for RQ3.
- Report integration.

Minghao Li:

- Define research question.
- Result analysis and report writing for RQ3: multi linear regression analysis.
- Code review and Github repository layout.

Yumeng Pan:

- Define research question.
- Coding for data processing, analysis and visualization.
- Result analysis and report writing for RQ3: time series analysis.
- File management and upload.