



# FINAL PROJECT

## Forecasting U.S. Flight Traffic

Business Forecasting 22:544:608:60

2024 FALL

Sumin Oh

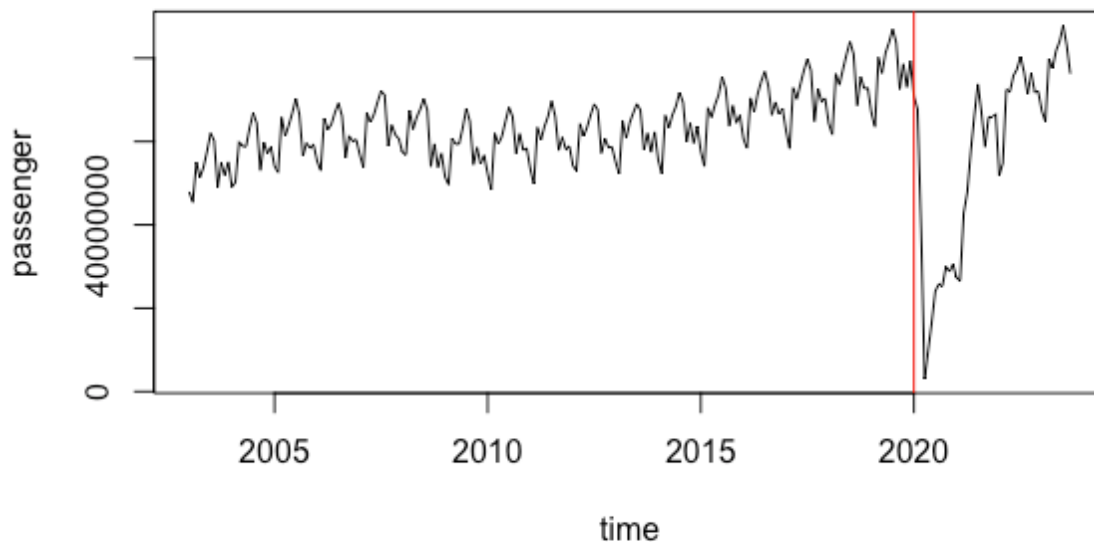
So518@scarletmail.rutgers.edu

## 1. Purpose of Forecasting

In this project, the number of passengers for the next 1 year, which is Y2024, will be predicted. Covid restricted passenger to use flights, but the number is recovering gradually after its drastic drop in 2020. This project is to forecast the total number of passengers using various forecast methods and compare the accuracy of those models to find out the best methods.

## 2. About this Data

This data provides U.S. monthly airline traffic from 2003 to 2023. Since it experienced significant change since 2020 due to the outbreak of Covid 19, data was cut from 2020 to 2023.



## 3. Exploratory Data Analysis

### (1) Box Plot & Histogram

Data is slightly skewed to left, with most of the values located between 60M to 80M. No outlier is detected. Through the Maximum and Minimum values, we can find out the values are quite large in size.

### (2) ACF

Starting from Lag1, the value far exceeds the confidence interval range. Values gradually decrease step by step, and from Lag 9, value goes under the confidence interval range. This figure implicate that there exists trend in the data. As for the seasonality, if this monthly data had seasonality, it should show high values around lag 12 or 24, however it isn't. So from this ACF analysis, we can assume that seasonality is not big in this data.

### (3) Decomposition

By doing Decomposition with `stl()` function, it showed there is a seasonality in the data. But considering that it wasn't stand out with ACF analysis, we can assume that there exists some seasonality in data but is not very central to explain patterns in data.

(4) Accuracy Measure Decision

Considering the type of the data(trend, weak seasonality, large values), MAE, RMSE, and MAPE were chosen for Accuracy measure.

#### 4. Forecasting & Residual Analysis

(1) Naïve Method

- Naïve predicted the values to be constantly 76M over the period.
- Residuals were randomly located without patterns.
- Residuals were not normally distributed, meaning that it has some limitations in explaining the patterns.

(2) Simple Moving Average Method

- Experiment was done under order 3,6,9,12 and accuracy measures were compared. As a result, order 12 turned out to be having the highest accuracy. Forecast was done using Moving Average 12 months.
- Residuals were very close to zeros.
- Residuals were quite normally distributed.
- Residuals didn't show any autocorrelations.

(3) Simple Smoothing Method

- `ETS(data, model="ANN")` was used to give a model additive error, no trend, and no seasonality.
- Value of Alpha is 0.9999, meaning that the model is giving almost all of its weight to recent data.
- Sigma is 1,114,981. Considering the size of the values, it seems to be not too much.
- Residuals were randomly located without patterns.
- Residuals were quite normally distributed.
- Residuals showed no autocorrelations, except for the lag 3.

(4) Holt-Winters Method

- It was done in "Additive" method
- Value of Alpha is **0.9999**, meaning that level the model gives almost all of its weight to recent data when determining the **level**.
- Value of Beta is **0.0001**, meaning that model barely updates the **trend**. This indicates that the data is likely to move around essentially stable levels and not have a clear trend.

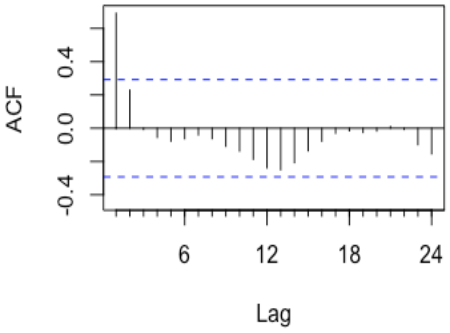
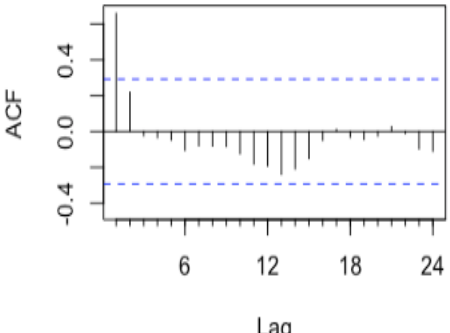
- Value of Gamma is **0.0001**, meaning that the model does not reflect **seasonality** at all, and that seasonality in the data is very weak or negligible. This suggests that there is little seasonal variation in the data, or that the model does not need to take seasonality into account to a large extent.
- Residuals were close to zero without patterns.
- Residuals were normally distributed.
- Residuals showed no autocorrelations, except for the lag 2 and 3.

(5) ARIMA method

- Auto.arima() function automatically generated a SARIMA model with two components, as seasonality was detected.
- ARIMA(0,1,1)(1,0,0)
- Non seasonal component : needs 1 differencing and needs to remove autocorrelation of residuals using Moving Average 1
- Seasonal component : Using 1 Seasonal Autoregressive(AR)
- Box-Ljung test : p-value is 0.7808, much larger than 0.05, meaning that residuals are random.
- Residuals are random, normally distributed, and not autocorrelated.

(6) Time series Regression method

- From the previous analysis, the data is assumed to have seasonality, but very week. To decide whether to include seasonality component to the model or not, both were conducted.

	With Seasonality	<b>Without Seasonality</b>
Adjusted $R^2$	0.5041	<b>0.6065</b>
Residual SE	16,150,000	<b>14,390,000</b>
AIC	1634.148	<b>1615.04</b>
Accuracy Measures	<b>MAE : 10709963</b> <b>RMSE : 13621321</b> <b>MAPE : 43.16</b>	MAE : 10744905 RMSE : 14066104 MAPE : 47.53
Residual ACF		

- Considering all those above data, regression was done **without seasonality**.
- Residuals are random, normally distributed, and not autocorrelated except for the lag 1.

## 5. Accuracy Summary & Decision

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Naive	124629.704545454544132	9557930	6409862	-23.7953910	39.621067	0.29364198	0.27265907
Moving Average	-346810.865830041118897	1836406	1117562	-0.7095576	3.096194	0.05405538	0.06742000
Simple Smoothing	931148.715920547721907	10899225	7076506	-22.1255857	39.885373	0.32418161	0.18433308
Holt-Winters	-343115.008365695422981	12490206	5688546	-26.3570402	36.995191	0.26059782	0.23675649
ARIMA	129193.222938643972157	8183899	5686949	-10.1949093	28.268650	0.26052462	-0.02593984
Regression	0.000000001490116	14066104	10744905	-31.7477090	47.534800	0.49223452	0.65779934

- Moving Average : showed the best performance in RMSE, MAE, and MAPE. Also the ACF1 is the lowest, meaning that residuals are very random.
- ARIMA : second best.
- Holt-Winters : Seasonality was reflected, however, accuracy was less than Moving Average or ARIMA.
- When deciding, since the seasonality is not large, a complex seasonal model may not be necessary. Also need to observe the models perform well in chosen measures such as MAE, RMSE, and MAPE.
- Final chosen method is : Moving Average.

## 6. To improve the result..

Even though Moving Average was chose, but at the same time always ready to use ARIMA or Holt-Winters as a complement when data patterns change or the forecast period becomes longer.