# The Interference of Pandemic to Aviation Performance

Qilun Lyu

Rutgers University

05/03/2021

ql193@rutgers.edu
(908)930-8642

# Introduction

- Delay or cancellation of flights leads to time and profit loss for all the parties engaged: passengers, airport authorities, airline companies.
- On-Time Performance (OTP) or punctuality is regarded by the majority as the most primary Key Performance Indicators (KPI) of an airline's rating on the top of others (e.g. quality, price, etc).
- COVID-19 pandemic is still burning, bringing less profit to airports and airlines. The gradually formed new norm make forecasting necessary since it broke out.

# Introduction

- All models are wrong! But we can extract useful information for forecasting, given a decade of records.
- The project aims to forecast **Cancellation Rates(CR)**, **On-Time Rates(OTR)**, and **Average Delay Minutes(ADM)** through leveraging ARIMA models from time-series approaches.
- To make better comparison, we regard *the pandemic* as our only controlled factor, rid of seasonal effects. We left the last few months of records to evaluate the model fit.

# Data Description

**Source** Bureau of Transportation Statistics of United State (BTS) [1]

**Scope** From Jan 2011 to Dec 2020, 120 files in total, one month each; Covering US domestic flights

**Snippet** Shown below

| Date | Airline | Num | Origin | Dest | DepDelayMin | ArrDelayMin | Cancelled | Diverted |
|---|---|---|---|---|---|---|---|---|
| 2020/05/03 | AA | 1 | JFK | LAX | 1 | 16 | 0 | 0 |
| 2020/05/05 | DL | 725 | EWR | ATL | 40 | 3 | 0 | 0 |
| 2020/05/07 | AS | 15 | BOS | SEA | 0 | 0 | 1 | 0 |

Every flight is either on-time, delay, canceled, or diverted. From observations, the portion of diverted flights can be ignored. Thus, cancellation and execution (normal or delayed) are strongly negatively correlated.

[1] https://transtats.bts.gov/Fields.asp?gnoyr_VQ=FGJ

# Logic of Processing

1. Data Wrangling and Exploration
   Removing NAs, Unifying types and Converting to data frames
2. Model Fitting and Tuning
   Find a proper set of coefficients for ARIMA(p,d,q)
3. Residual Analysis
   Draw residual plots and normality checks
4. Forecasting and Verification
   Create prediction intervals and compare to the true data

# Data Wrangling

Though data is in csv format, using *read.csv* might mess up the structure, especially when quotation mark appears. Instead, we adopted

- fread(file=file, sep = ",", stringsAsFactors = FALSE, header = TRUE)

The header of the files before 2020 are of camel-back style, but fully capitalized after that with naming discrepancy.

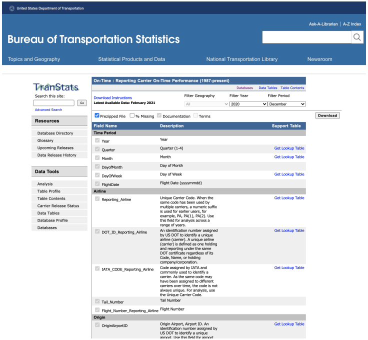



Figure 1: Data frame

# Data Wrangling

The following columns are selected

- FlightDate, IATA_ CODE_ Reporting_ Airline, Origin, Dest, DepDelayMinutes, DepBlk, ArrDelayMinutes, ArrBlk, Cancelled, Diverted

```
Columns: 110
$ Year                        <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020,
2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020…
$ Quarter                     <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4…
$ Month                       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,
10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10…
$ DayofMonth                  <int> 8, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 25,
26, 27, 28, 29, 30, 8, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23…
$ DayOfWeek                   <int> 4, 5, 7, 1, 2, 3, 4, 5, 7, 1, 2, 3, 4, 5, 7, 1, 2, 3, 4,
5, 4, 5, 7, 1, 2, 3, 4, 5, 7, 1, 2, 3, 4, 5, 7, 1, 2, 3, 4, 5, 4, 5, 7, 1…
$ FlightDate                  <chr> "2020-10-08", "2020-10-09", "2020-10-11", "2020-10-12",
"2020-10-13", "2020-10-14", "2020-10-15", "2020-10-16", "2020-10-18", "202…
$ Reporting_Airline           <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA",
"AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA"…
$ DOT_ID_Reporting_Airline    <int> 19805, 19805, 19805, 19805, 19805, 19805, 19805, 19805,
19805, 19805, 19805, 19805, 19805, 19805, 19805, 19805, 19805, 19805, 1980…
$ IATA_CODE_Reporting_Airline <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA",
"AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA"…
$ Tail_Number                 <chr> "N932AM", "N934AA", "N992AU", "N132AN", "N139AN",
"N993AN", "N166NN", "N930AU", "N992AU", "N151AN", "N143AN", "N156AN", "N165NN", …
$ Flight_Number_Reporting_Airline <int> 2259, 2259, 2259, 2259, 2259, 2259, 2259, 2259, 2259,
2259, 2259, 2259, 2259, 2259, 2259, 2259, 2259, 2259, 2259, 2259, 2260, 2260…
$ OriginAirportID             <int> 11298, 11298, 11298, 11298, 11298, 11298, 11298, 11298,
11298, 11298, 11298, 11298, 11298, 11298, 11298, 11298, 11298, 11298, 1129…
```

Figure 2: Data frame

# Data Wrangling

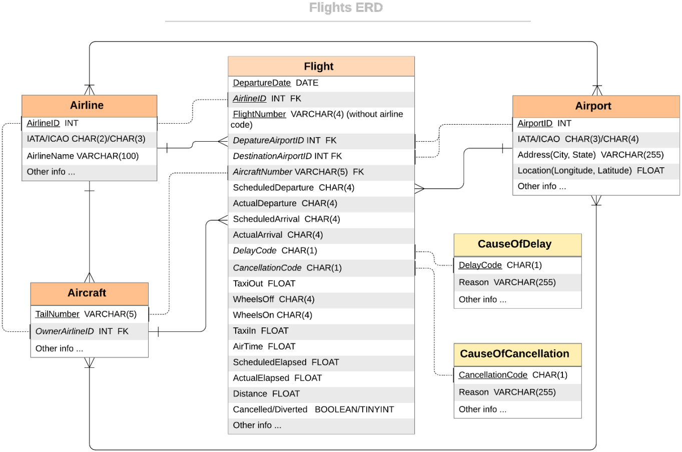



Figure 3: Entity Relation Diagram

# Interesting Findings

Pareto principle(20-80 rule): we found top 10% of airports with the most significant annual number of scheduled flights own more than 94.8% of flights. Ridiculously, the top 10% of airlines take up over 97.7% of all flights. Monopoly?
Indicating large airports and airlines by half-normal plots:

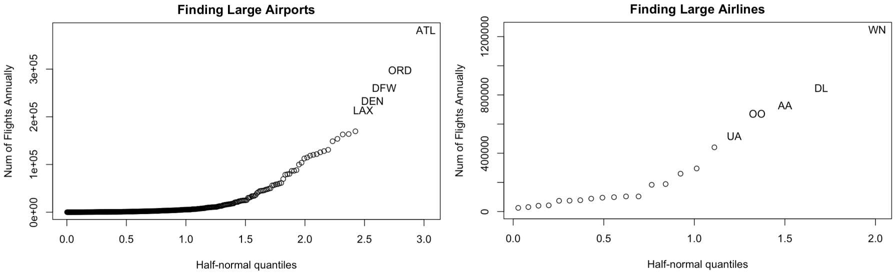



Figure 4: Indicating leverages

# Interesting Findings

Focusing on number of flights within a day, we have

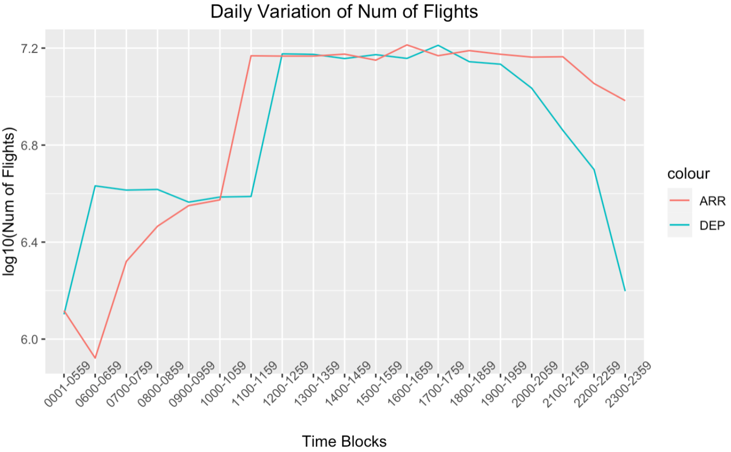



Figure 5: Compare departure and arrival traffic flows by time block

# Interesting Findings

- Mondays, Thursdays and Fridays have the most delays.
- June, July and August have the most delays.
- Both airport locations and flight routes are denser in the east coast.

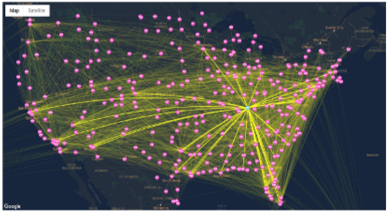


Figure 6: Distribution of airports and their interconnectivity

# Interesting Findings

## Anomaly Detections


Figure 7: Find an Incident from interactive series

# Overall Series

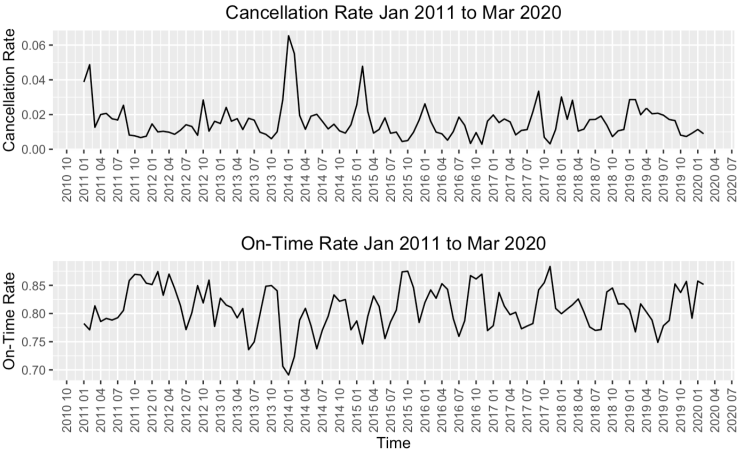



Figure 8: Compare monthly CR and OTR. Note the scale!

# Overall Series

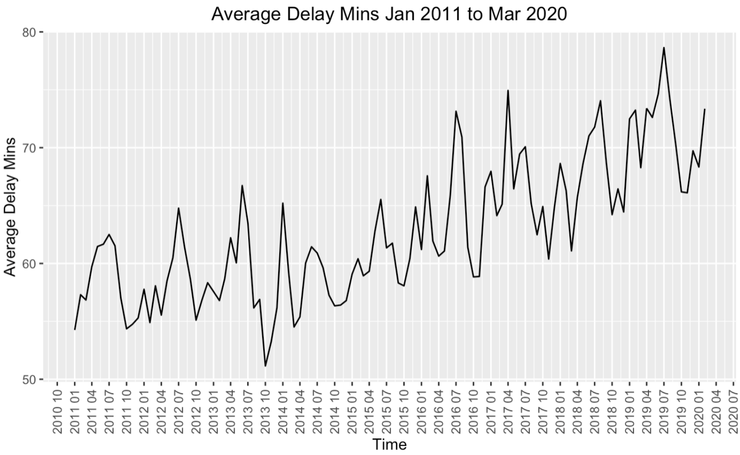



Figure 9: Monthly ADM. Suggest a Moving Average model

# Series Decomposition

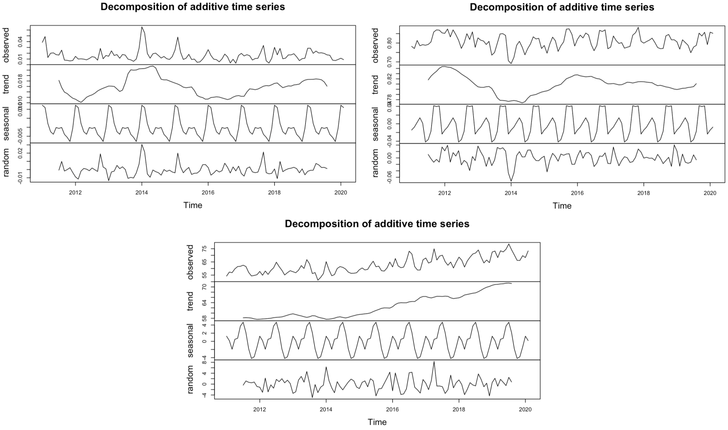



Figure 10: Decomposition of 3 series

# Essentials of Seasonal ARIMA Model

Assume we have a seasonal time series $\{Y_t\}_{t=1}^{T}$, which is fitted by a seasonal model ARIMA$(p, d, q)(P, D, Q)_m$, where $(p, d, q)$ refers to non-seasonal part, $(P, D, Q)$ refers to seasonal one and $m =$ the number of observations each year.
Define the seasonal differencing:

$$(1 - B^S)Y_t = Y_t - Y_{t-S}$$

and non-seasonal differencing

$$(1 - B)Y_t = Y_t - Y_{t-1}$$

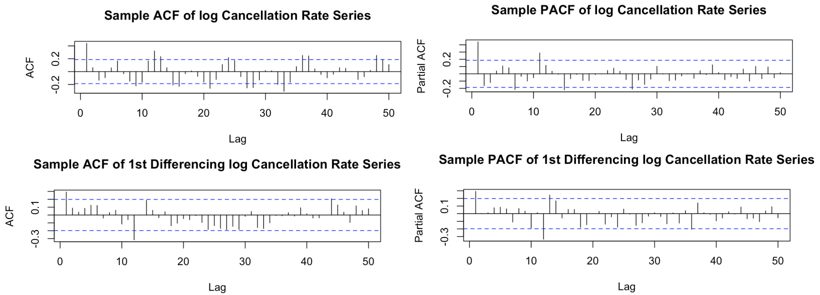
for the trend. So, we can examine the model through

$$(1 - B^{12})(1 - B)Y_t = (Y_t - Y_{t-12}) - (Y_t - Y_{t-1})$$

.

# Stationarity Testing

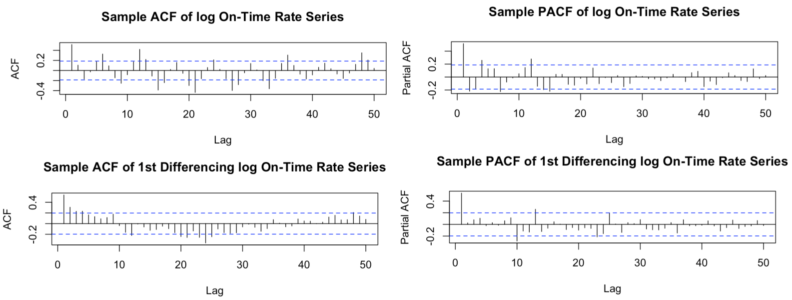
Evidently, there is a seasonal pattern, which indicates a nonstationarity before seasonal differencing.




Nonseasonal behavior: The PACF of 1st diff shows a clear spike at lag 1 and not much else until lag 36. Try AR(2) or AR(3). Seasonal behavior: In the PACF, there's a cluster of spikes around lag 12 and then not much else. Try SAR(1).

Figure 11: ACF plots and PACF plots of CR series

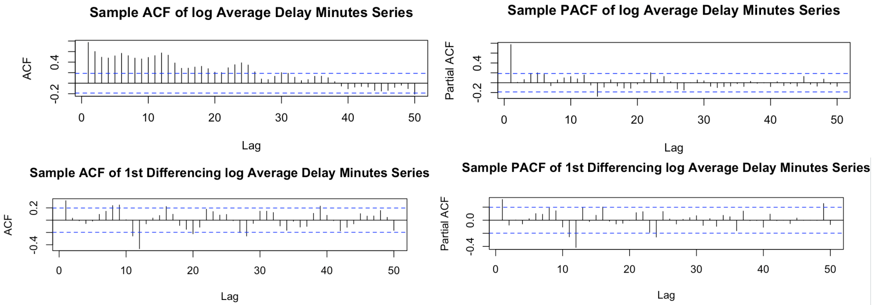
# Stationarity Testing




Nonseasonal behavior: The PACF of 1st diff shows a clear spike at lag 1 and not much else until lag 23. Try AR(1). Seasonal behavior: In the PACF of 1st diff, there's a cluster of spikes around lag 12,24 and then not much else. Try from SAR(2,0) to SARI(2,3), etc.

Figure 12: ACF plots and PACF plots of OTR series

# Stationarity Testing




Nonseasonal behavior: The ACF of 1st diff shows spikes at lag 12, and not much else until lag 28. Try from MA(2) to IMA(3,2). Seasonal behavior: In the ACF of 1st diff, there's a cluster of spikes around lag 12,28 and then not much else. Try MA(2).

Figure 13: ACF plots and PACF plots of ADM series

# Stationarity Testing

The augmented Dickey-Fuller (ADF) test statistic is the t-statistic of the estimated coefficient of $\alpha$ from the method of least squares regression. [2]
Lag order = 12,
p-value of log CR series = 0.6401,
p-value of log OTR series = 0.6764,
p-value of log ADM series = 0.6241

[2] SHUMWAY, R. H., & STOFFER, D. S. (2006). Time series analysis and its applications: with R examples. New York, Springer.

# Model Specification

Suppose a ARIMA($p, d, q$) model, where $p, q$ are determined by minimum AIC(find a good model to predict) and BIC(find a best fit to the data). We find

**CR series** ARIMA(2,0,0)×(1,0,0)[12]
with AIC=-725.24, BIC=-711.74 and $\sigma^2 < 1e-4$.

**OTR series** ARIMA(1,0,0)×(2,1,0)[12]
with AIC=-411.26, BIC=-400.92 and $\sigma^2 < 1e-3$.

**ADM series** ARIMA(0,1,2)×(0,0,2)[12]
with AIC=574 BIC=590.15 and $\sigma^2 = 10.36$.

# Residual Analysis

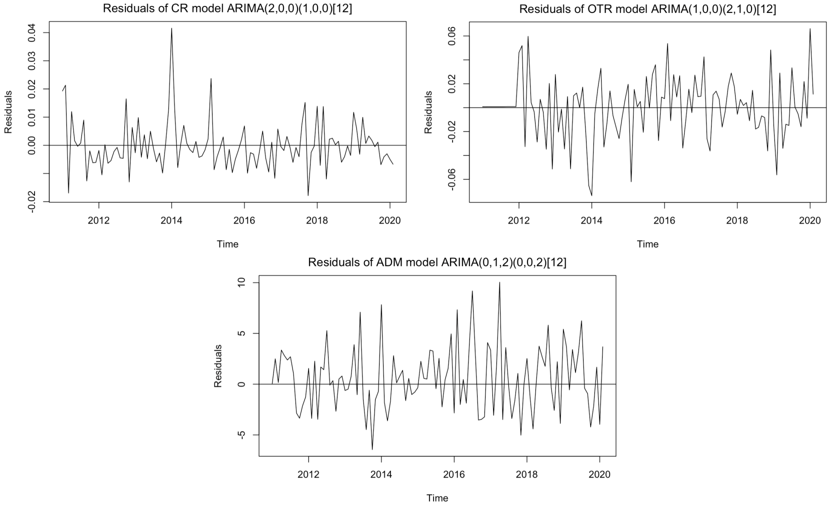



Figure 14: Residual Plots

# Ljung-Box Test and Normality Assumption

According to Ljung-Box Test and Shapiro-Wilk normality test, we have p-values of three residuals.

- **CR series** LB:0.5767, SW: 9.682e-07, reject normality.
- **OTR series** LB:0.1702, SW:0.1446.
- **ADM series** LB:0.8637, SW:0.1287.

# Forecasting

What if no pandemic,...

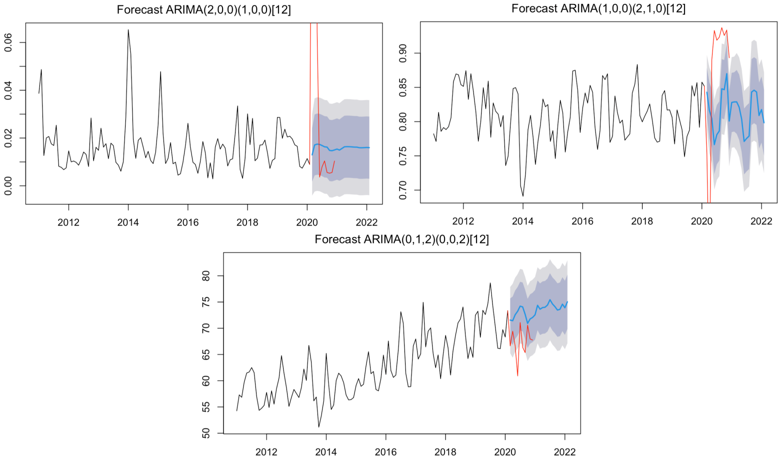
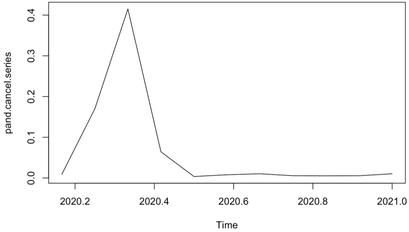
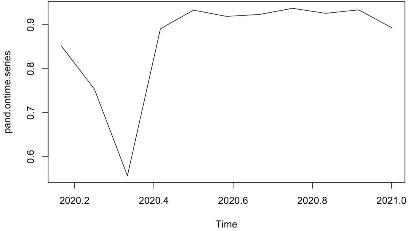


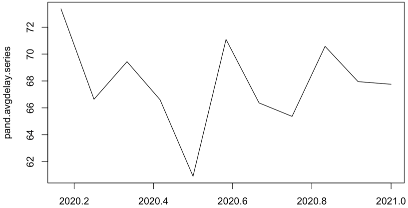

Figure 15: Forecasting Plots

# Forecasting

The series after pandemic: CR and OTR model: replace March-April 2020 with the past average; ADM model: fit a new model ARIMA(1,1,0)(0,1,0)[12] from recent 2 years data.



# Forecasting and Verification

Given 5 more months of data, we have

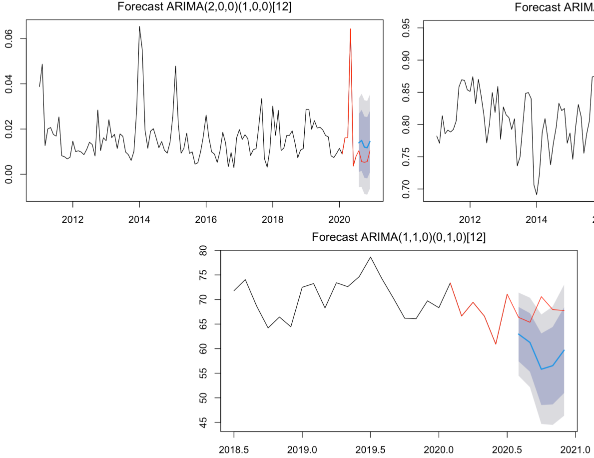



Figure 17: Forecasting Plots

# Future work

Like Gaussian Mixture Model, can we mix the two models here?
Other Questions?

# Reference


- B. Walther, "6 Most Important KPIs For Airline Operations," Information Design, 3 2021. [Online]. Available: https://www.id1.de/2019/10/25/6-most-important-kpis-for-airline-operations-and-performance-analysis/.
- "ON-TIME PERFORMANCE OTP IS BECOMING INCREASINGLY IMPORTANT TO AN AIRLINES AND AIRPORTS," OAG Aviation Worldwide Limited, 2021. [Online]. Available: https://www.oag.com/on-time-performance-airlines-airports.
- G. D. Gosling, "Aviation System Performance Measures," UC Berkeley, 1999. [Online]. Available: https://escholarship.org/uc/item/2xw9204x.
- G. a. J. G. Box, Time Series Analysis: Forecasting and Control, San Francisco: Holden-Day, 1970. S. Chatterjee, "Time Series Analysis Using ARIMA Model In R," 05 02 2018. [Online]. Available: https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/.