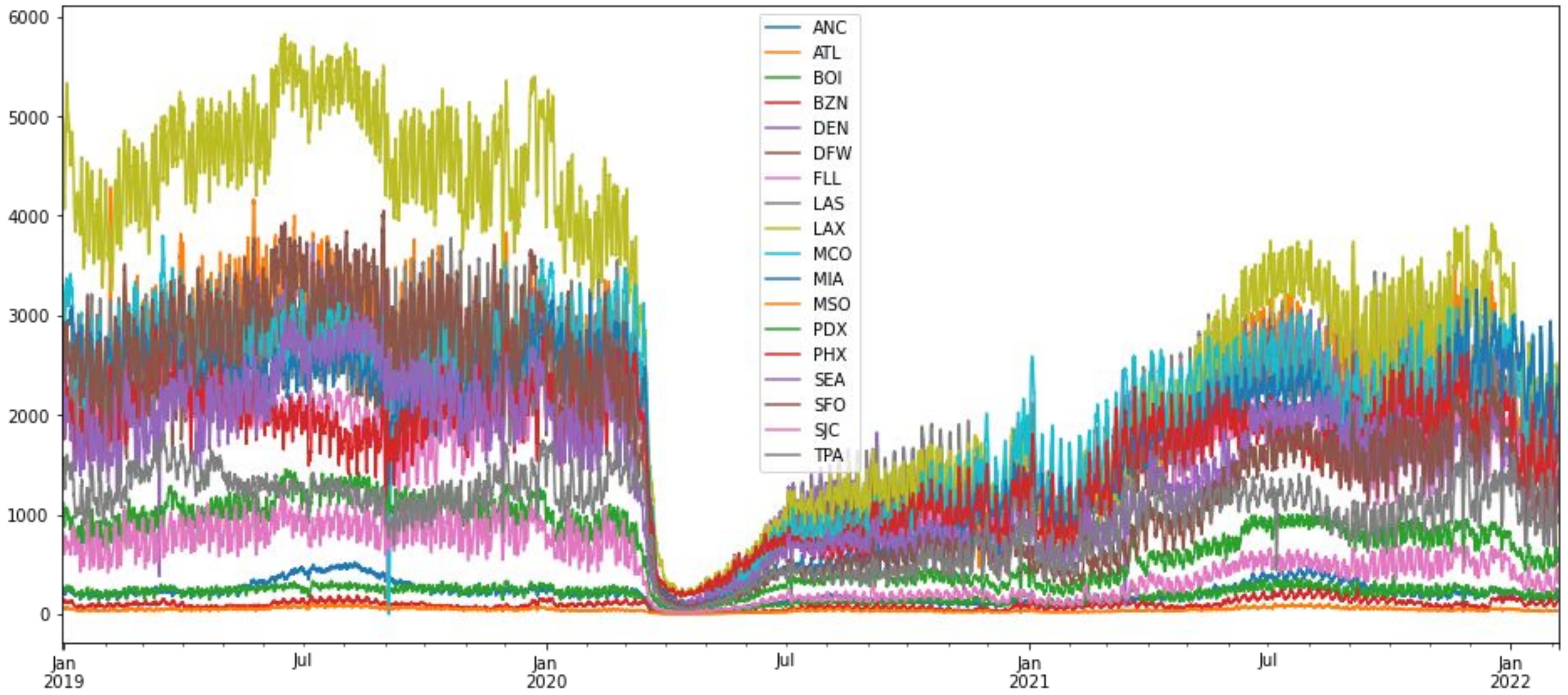
Predict TSA Throughput

A time series analysis

Problem Identification and Business Context



6000
5000
4000
3000
2000
1000
0
Jan
2019
Jul
Jan
2020
Jul
Jan
2021
Jul
Jan
2022
ANC
ATL
BOI
BZN
DEN
DFW
FLL
LAS
LAX
MCO
MIA
MSO
PDX
PHX
SEA
SFO
SJC
TPA

Predictive model - TSA Throughput

Purpose:

- Allow airports and airlines make better resource allocation and staffing decisions
- TSA efficiency and customer service
- Recovering from Covid-19 travel restrictions beginning in March 2020

Stakeholders:

- Airports
- Airlines
- Airport businesses

Challenges:

- Extreme weather
- War
- Pandemics

Data Collection

Data gathering: Github repository (source)

- Individual Csv files for each airport
- Columns for each gate
- Values: TSA throughput (number of people going through security)

Data Collection

Building final Dataframe:

- Aggregate all gates within an airport
- A column for each airport
- Datetime Index (hourly)

Data Exploration and Cleaning

Data

18 columns: US International Airports

27216 rows: Hourly Data between

December 30th 2018 – February 5th 2022

Values: Number of people going through security

Airports

ANC - Anchorage

ATL - Atlanta

BOI - Boise

BZN - Bozeman

DEN - Denver

DFW - Dallas Fort Worth

FLL - Fort Lauderdale

LAS - Las Vegas

LAX - Los Angeles

MCO - Orlando

MIA - Miami

MSO - Missoula Montana

PDX - Portland

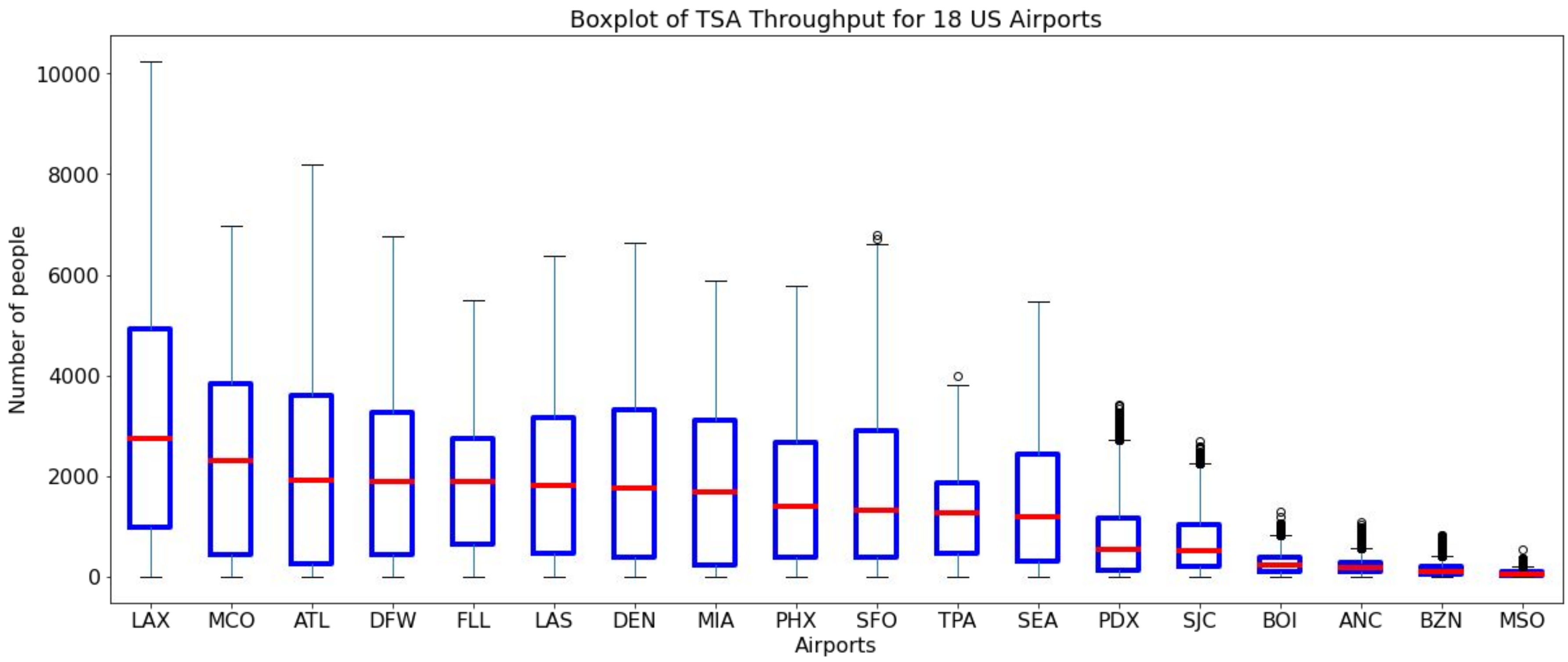
SEA - Seattle

SFO - San Francisco

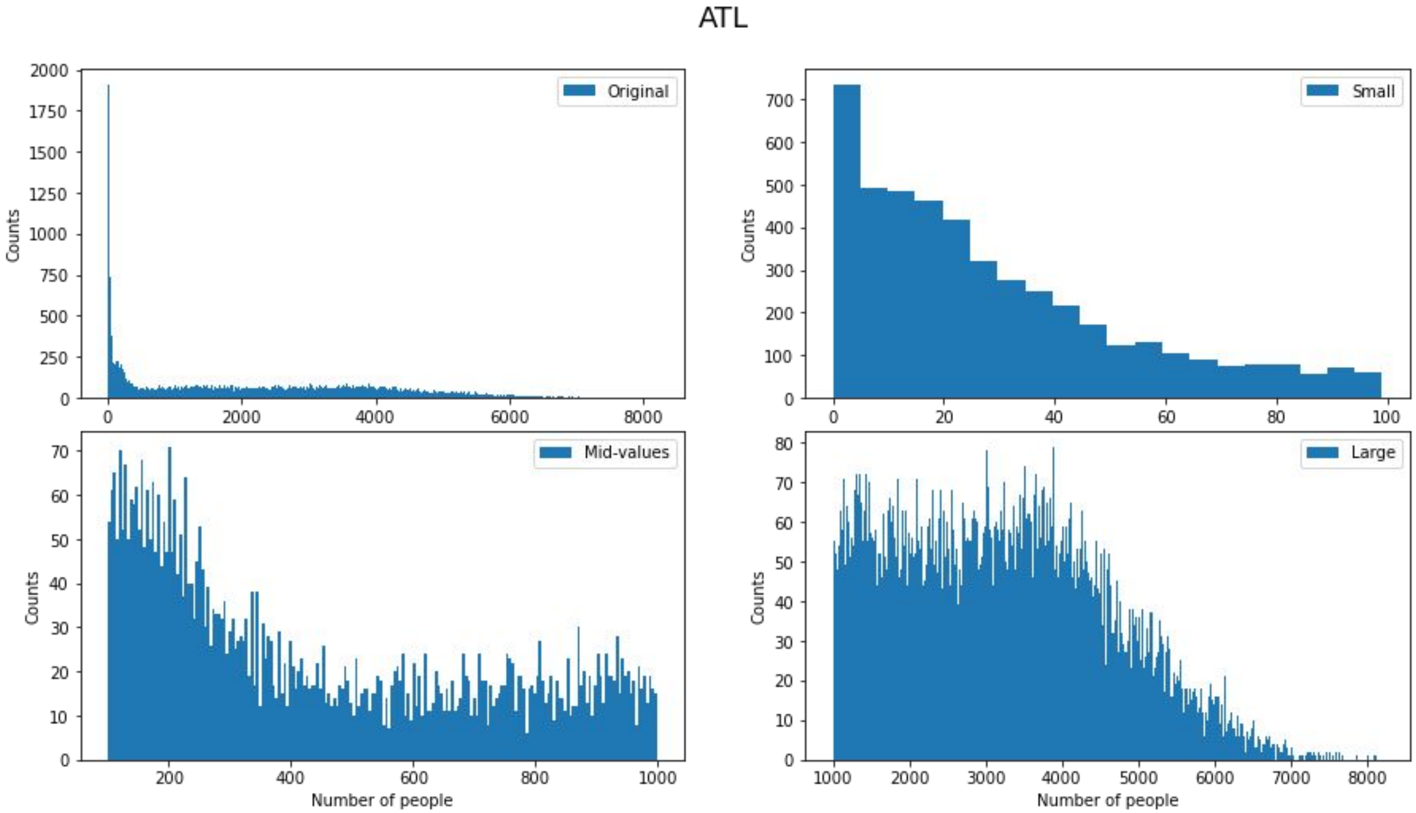
SJC - San Jose

TPA - Tampa

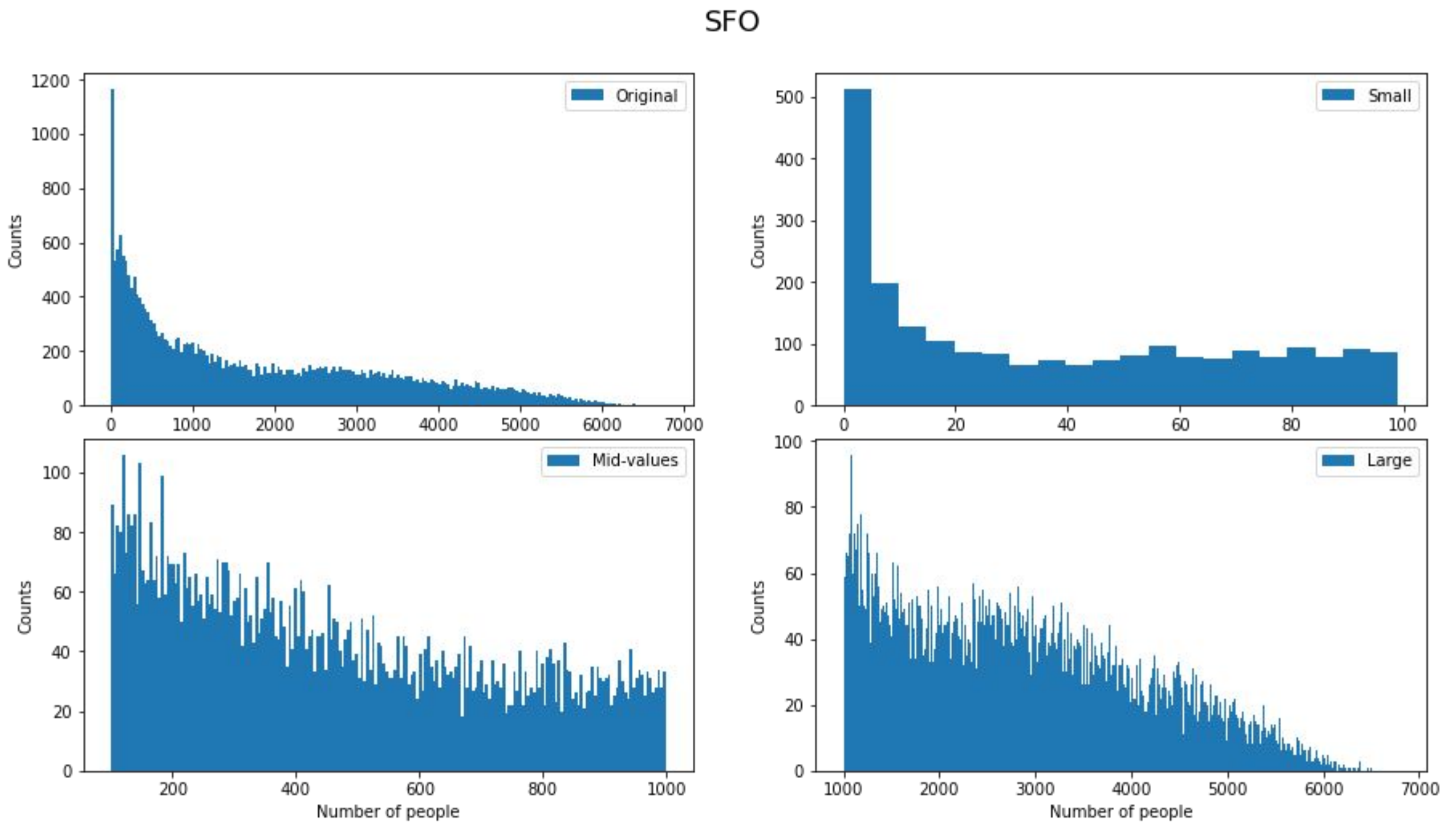
Boxplots


Boxplot of TSA Throughput for 18 US Airports
Number of people
10000
8000
6000
4000
2000
0
LAX
MCO
ATL
DFW
FLL
LAS
DEN
MIA
PHX
SFO
TPA
SEA
PDX
SJC
BOI
ANC
BZN
MSO
Airports

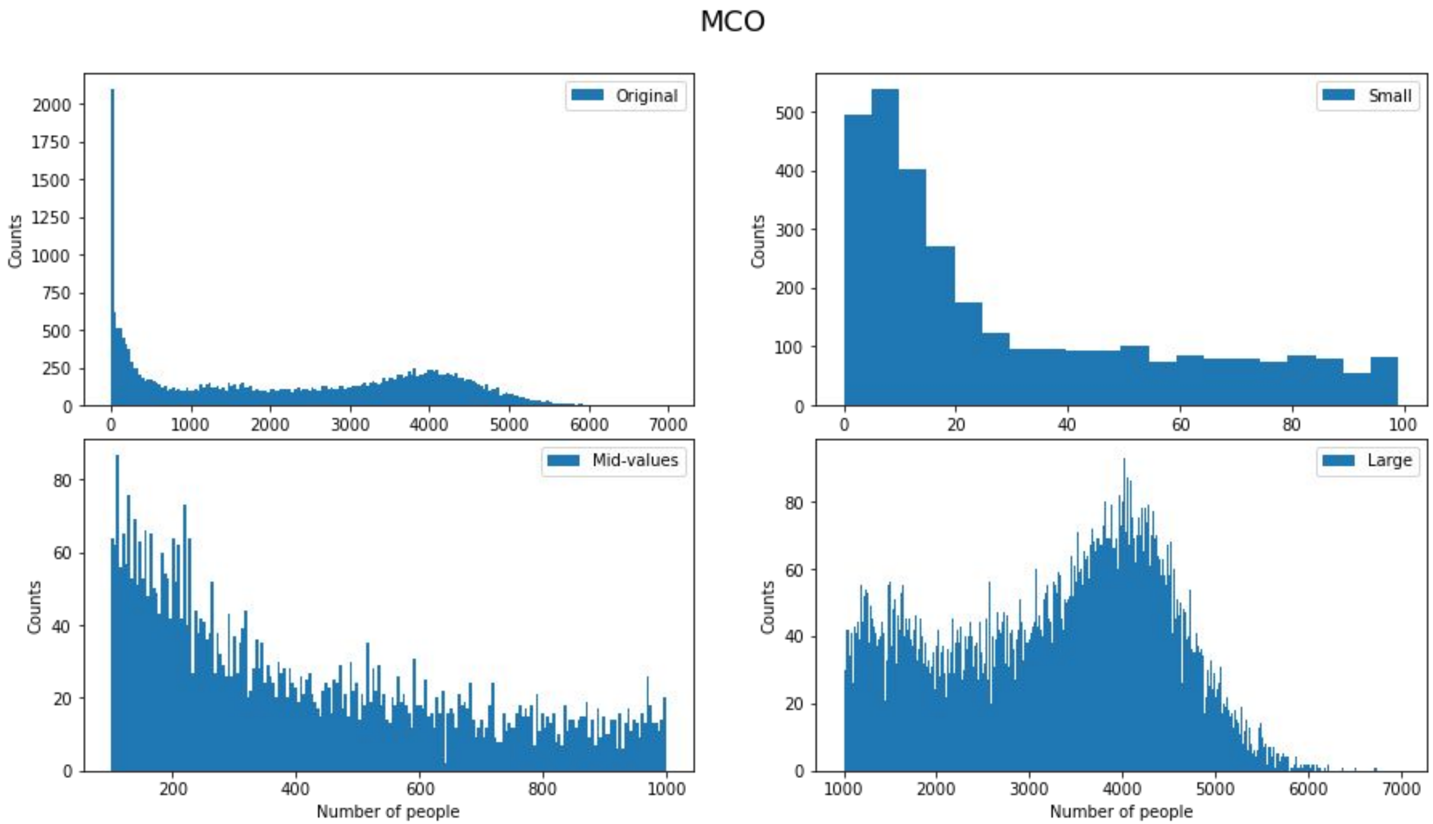
Histograms


ATL
Original
Small
Mid-values
Large
Counts
Number of people
2000
1750
1500
1250
1000
750
500
250
0
2000
4000
6000
8000
700
600
500
400
300
200
100
20
40
60
80
100
70
60
50
40
30
20
10
200
400
600
800
1000
80
3000
5000
7000

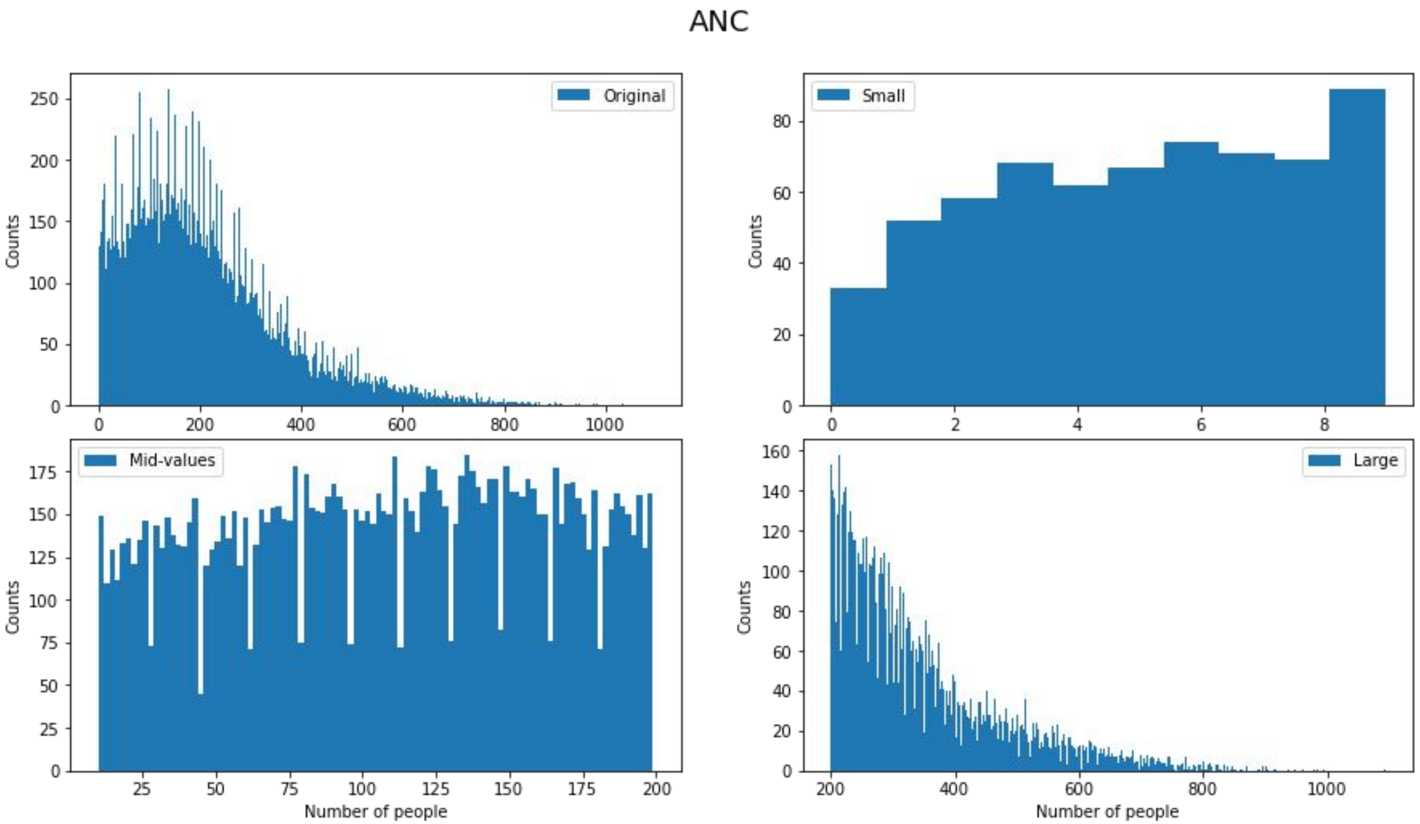
Histograms


SFO
Original
Small
Mid-values
Large
Counts
Number of people
1200
1000
800
600
400
200
0
1000
2000
3000
4000
5000
6000
7000
500
300
100
20
40
60
80

Histograms

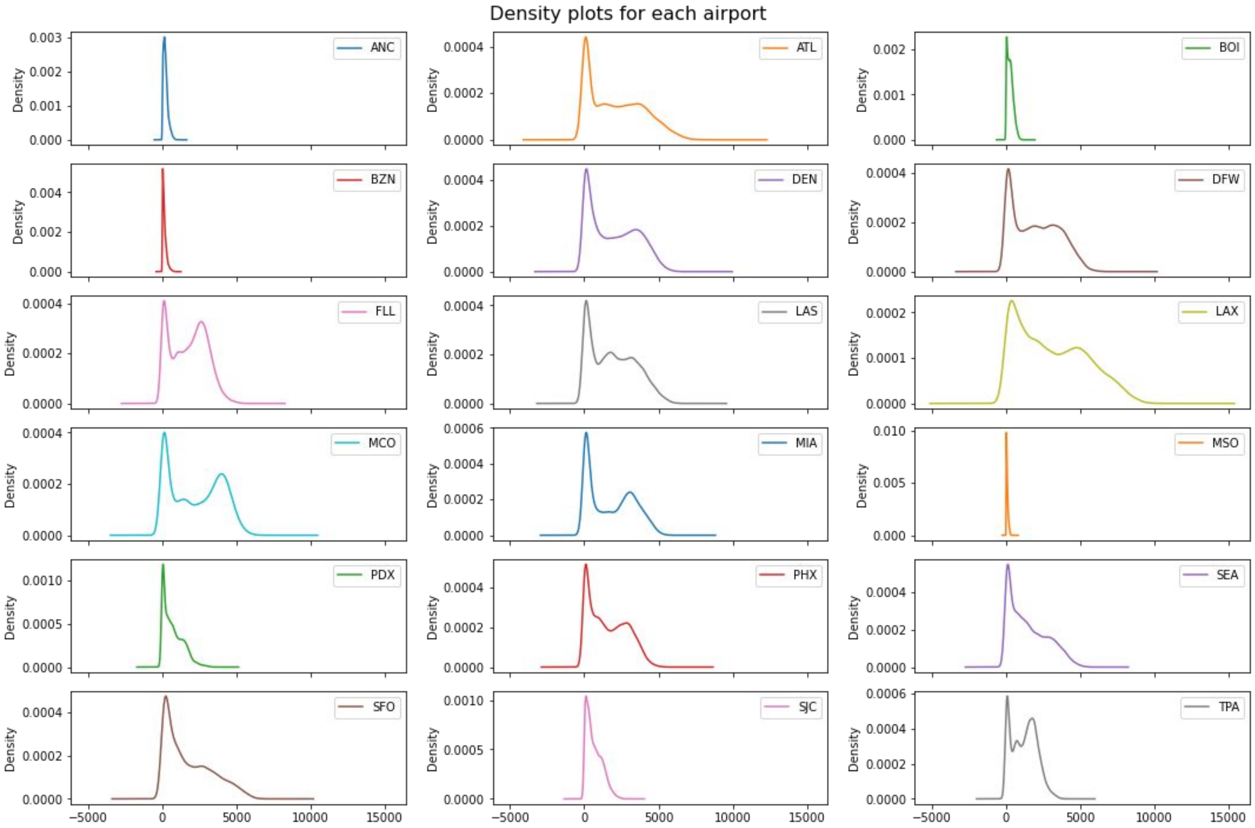

MCO
Original
Small
Mid-values
Large
Counts
Number of people
0
250
500
750
1000
1250
1500
1750
2000
1000
2000
3000
4000
5000
6000
7000
100
200
300
400
20
40
60
80
600
800

Histograms



ANC
Original
Small
Mid-values
Large
Counts
Number of people

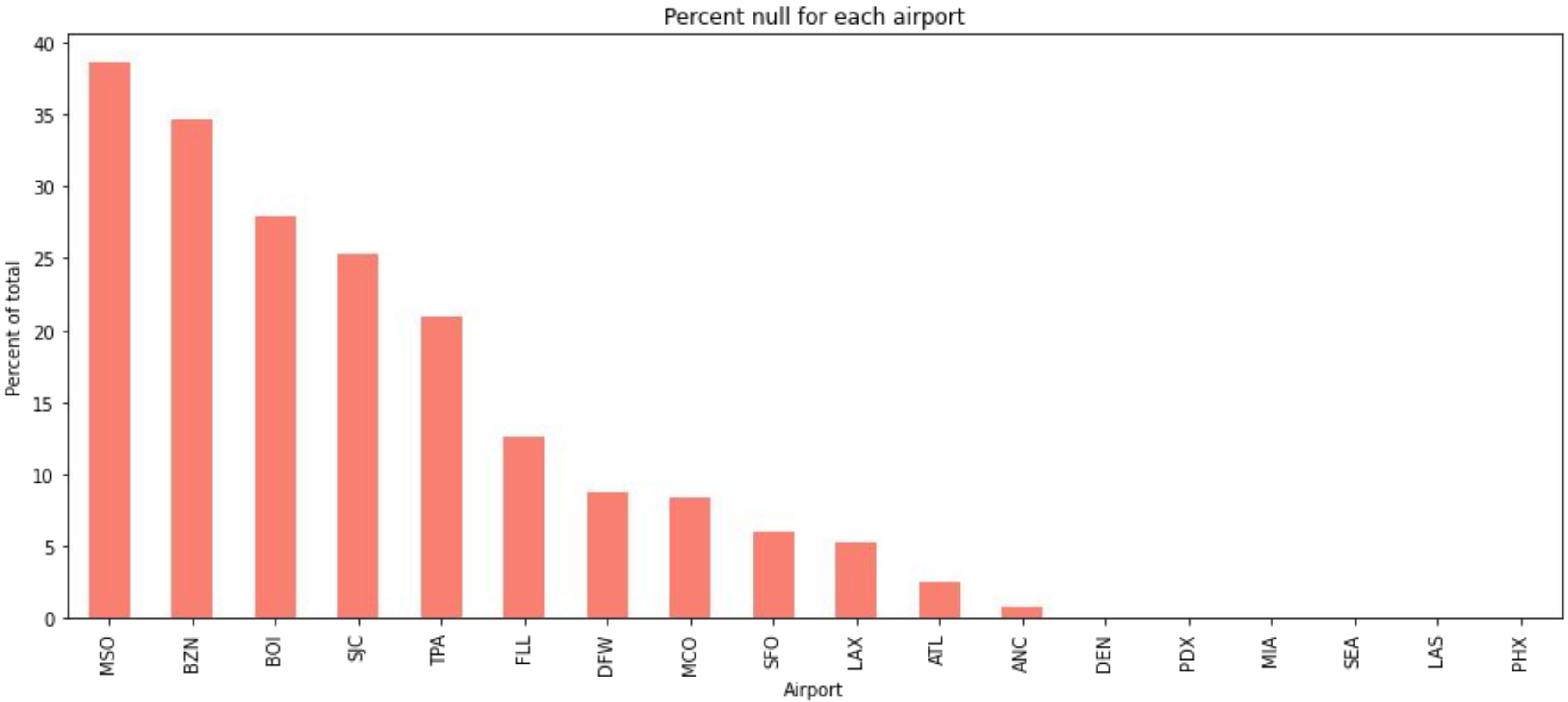
Density Plots

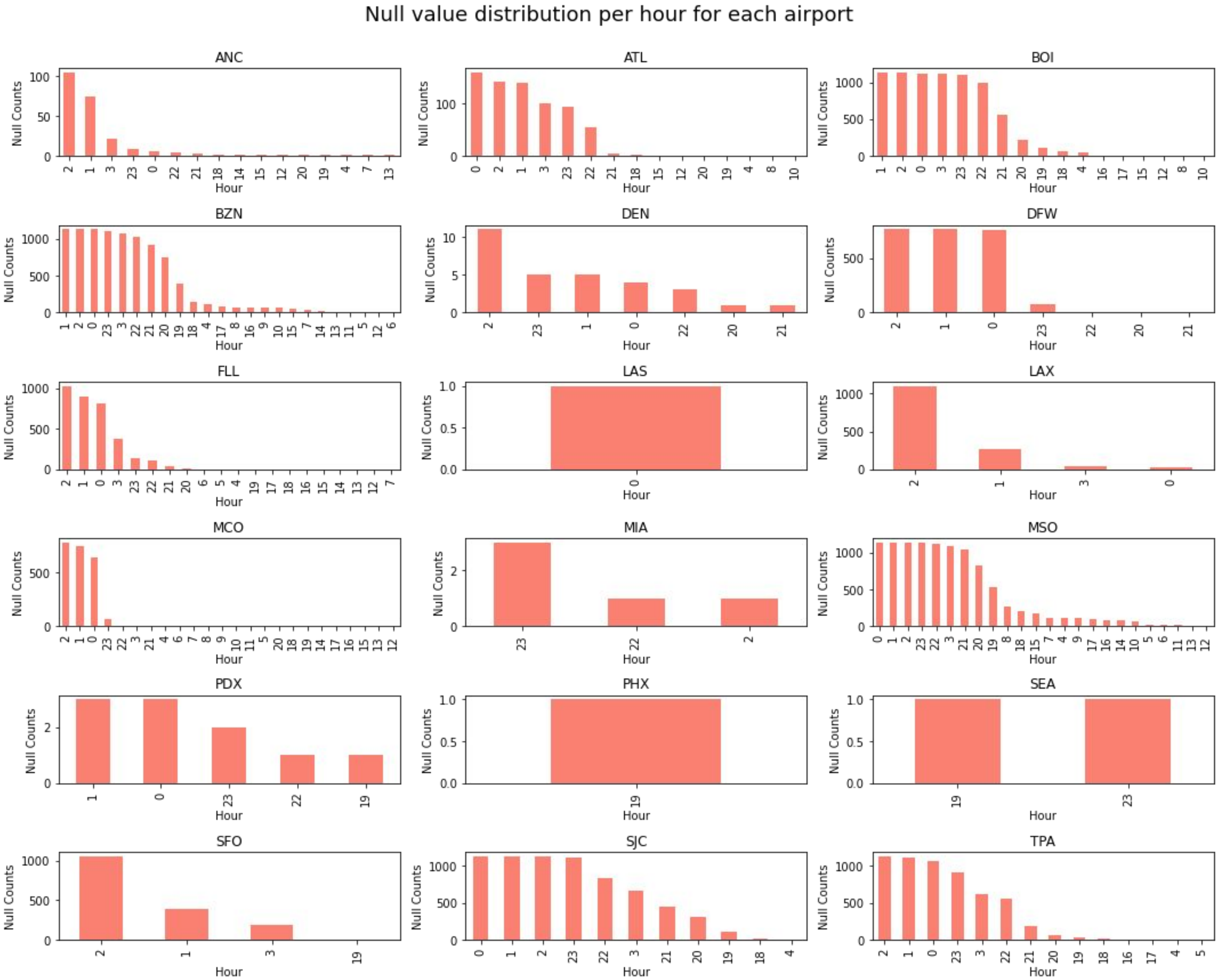


Data Distribution

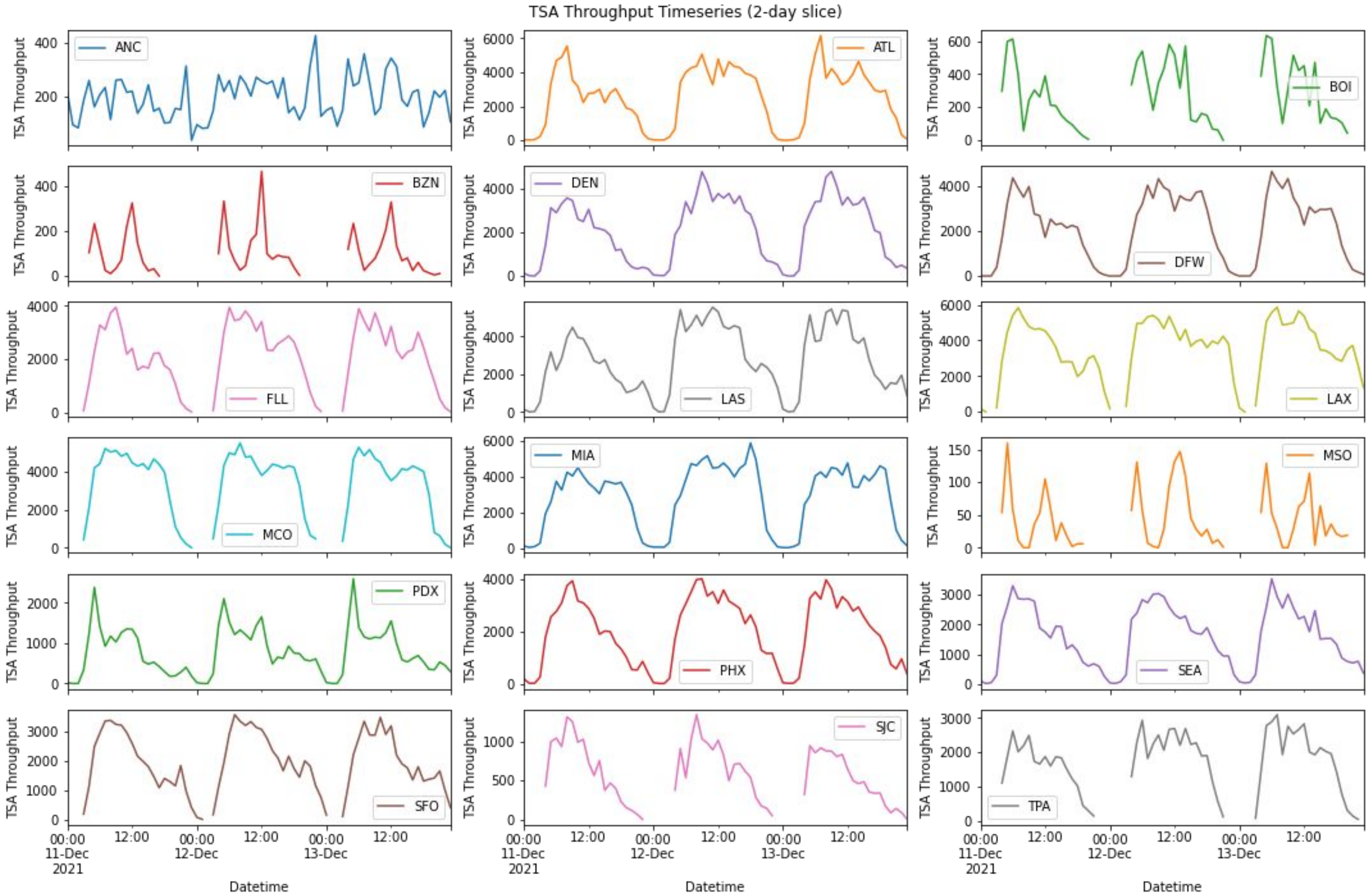
- First peak for relatively low counts
- Large airports have a second peak at higher counts
- Extremely busy times with relatively large throughput are much more rare

Null values


Percent null for each airport
Percent of total
Airport
0
5
10
15
20
25
30
35
40
MSO
BZN
BOI
SJC
TPA
FLL
DFW
MCO
SFO
LAX
ATL
ANC
DEN
PDX
MIA
SEA
LAS
PHX



Null value distribution per hour for each airport
ANC
ATL
BOI
BZN
DEN
DFW
FLL
LAS
LAX
MCO
MIA
MSO
PDX
PHX
SEA
SFO
SJC
TPA
Null Counts
Hour


TSA Throughput Timeseries (2-day slice)
TSA Throughput
Datetime
ANC
ATL
BOI
BZN
DEN
DFW
FLL
LAS
LAX
MCO
MIA
MSO
PDX
PHX
SEA
SFO
SJC
TPA
00:00
11-Dec
2021
12:00
00:00
12-Dec
12:00
00:00
13-Dec
12:00

Null value treatment

- Small airports have a higher ratio of values missing than large airports

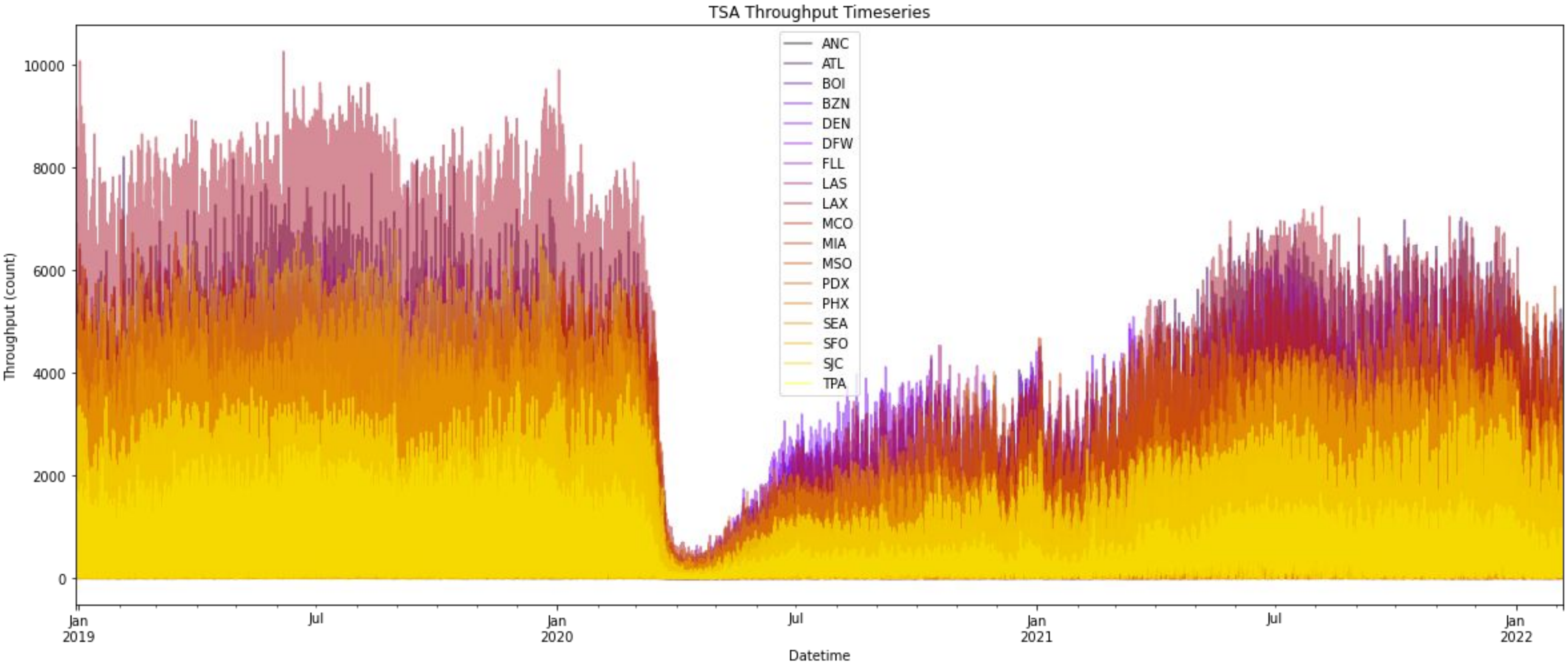
  (Outgoing flight times for small vs large airports)

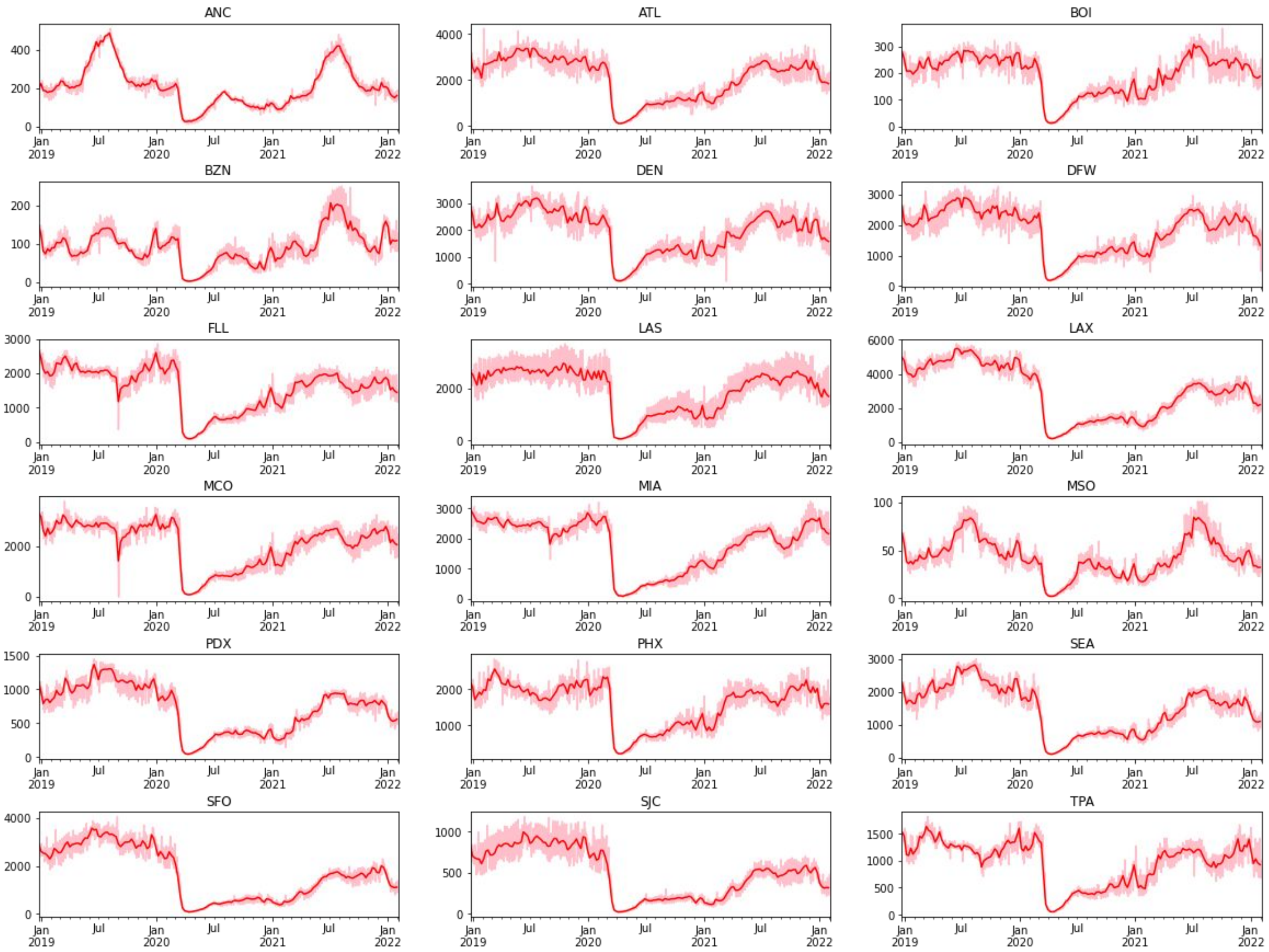
- Missing values are concentrated in late evening to early morning hours

  (Least common times for outgoing flights for large airports)

Impute null values with ZERO

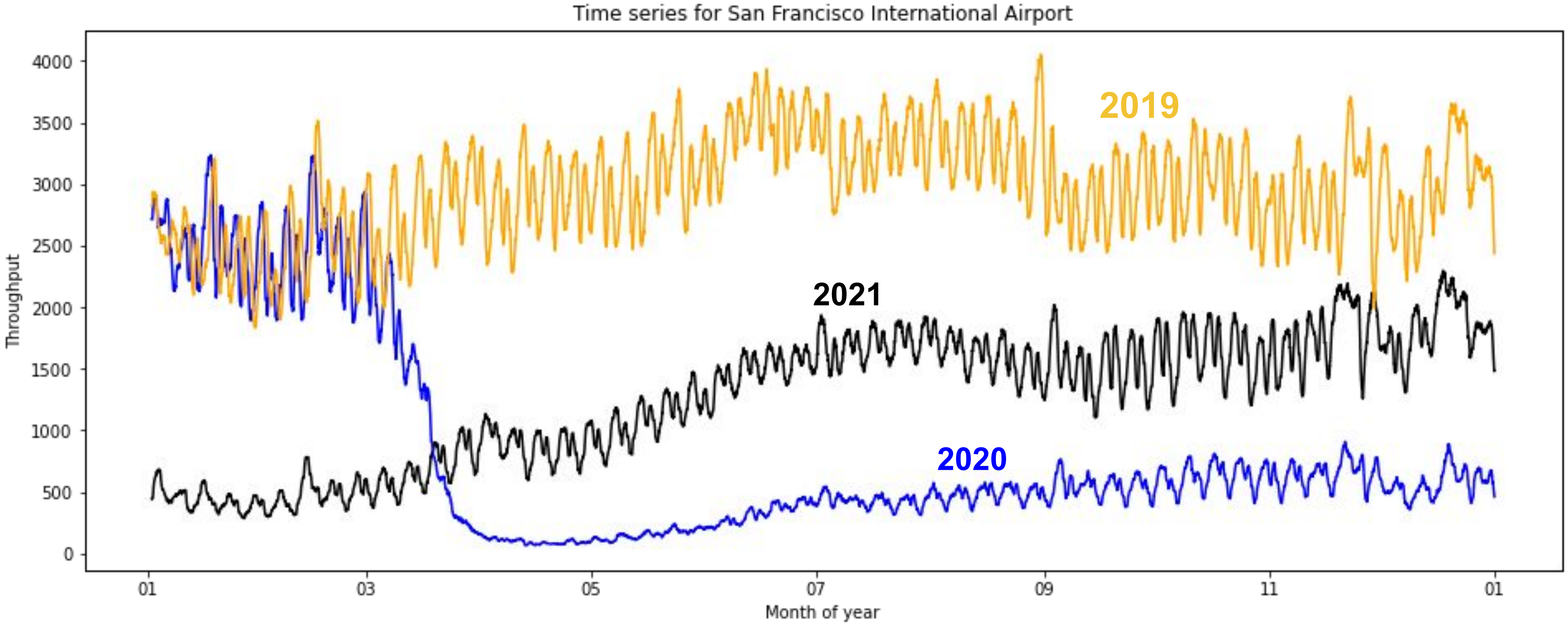
Time Series Plot - All Airports



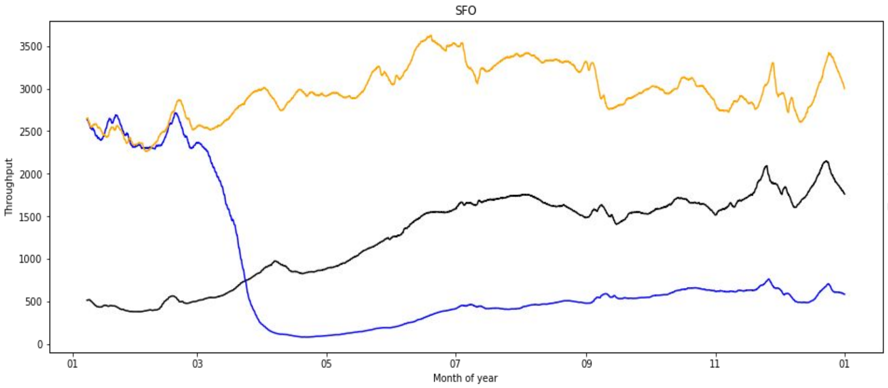


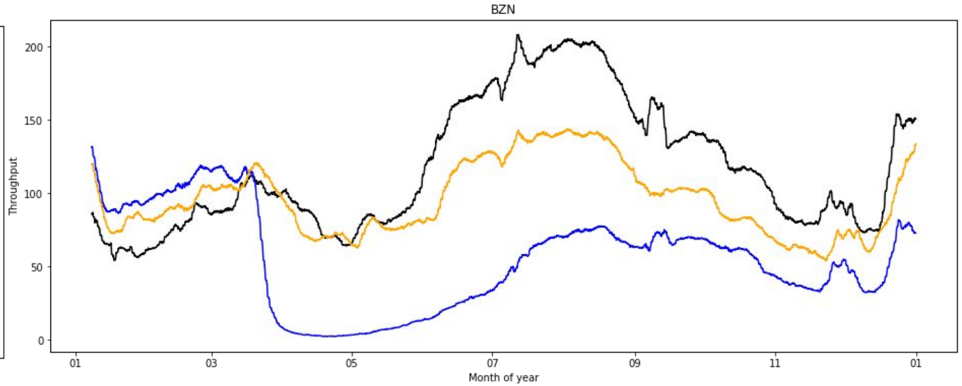
ANC
ATL
BOI
BZN
DEN
DFW
FLL
LAS
LAX
MCO
MIA
MSO
PDX
PHX
SEA
SFO
SJC
TPA
Jan 2019
Jul
Jan 2020
Jul
Jan 2021
Jul
Jan 2022

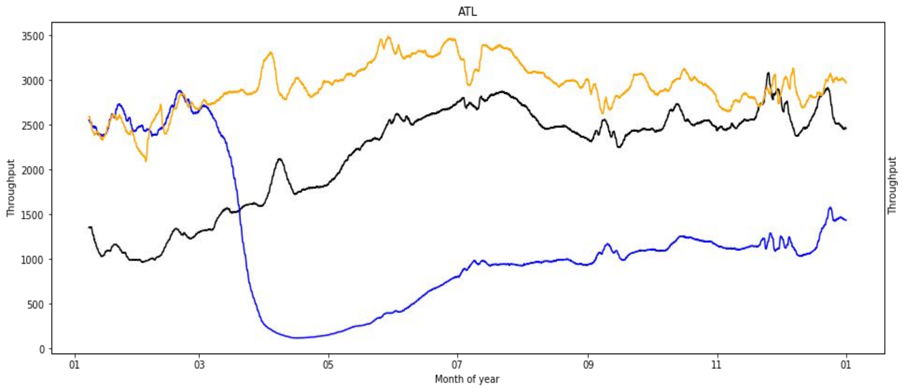
Yearly Trends - SFO

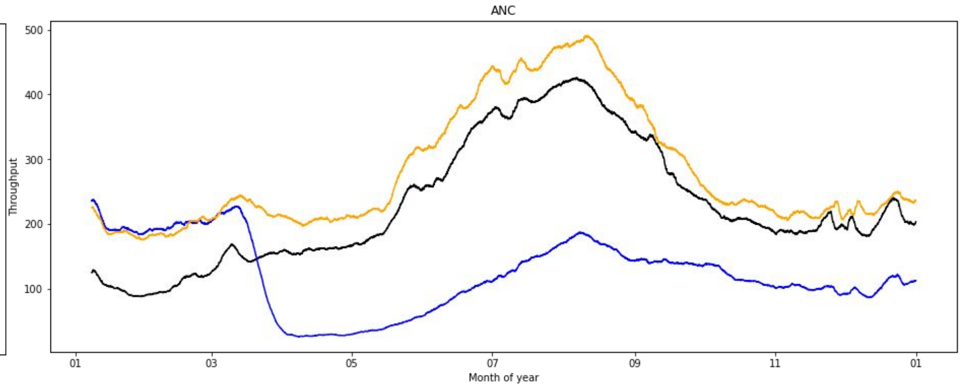


Yearly Trends

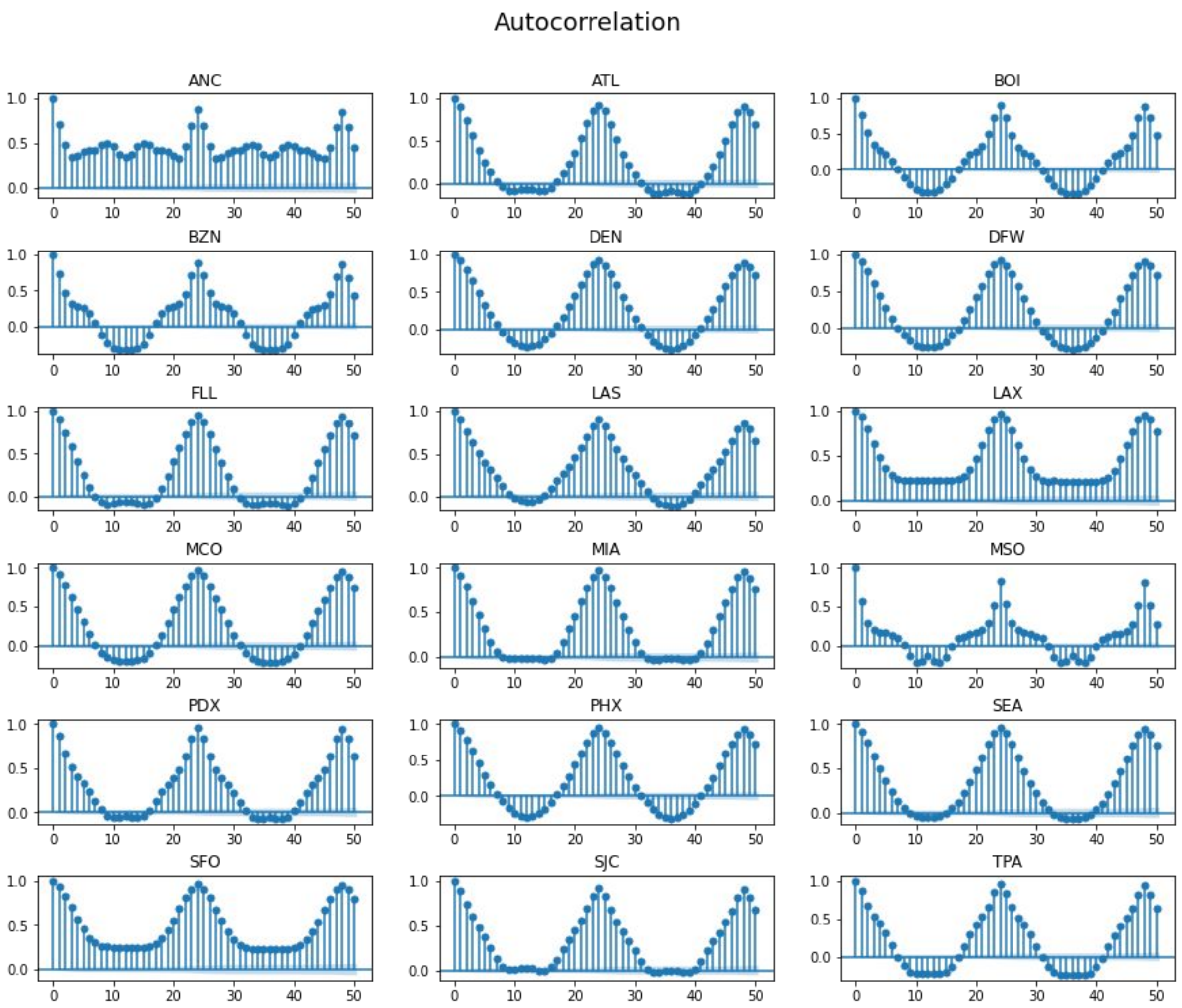






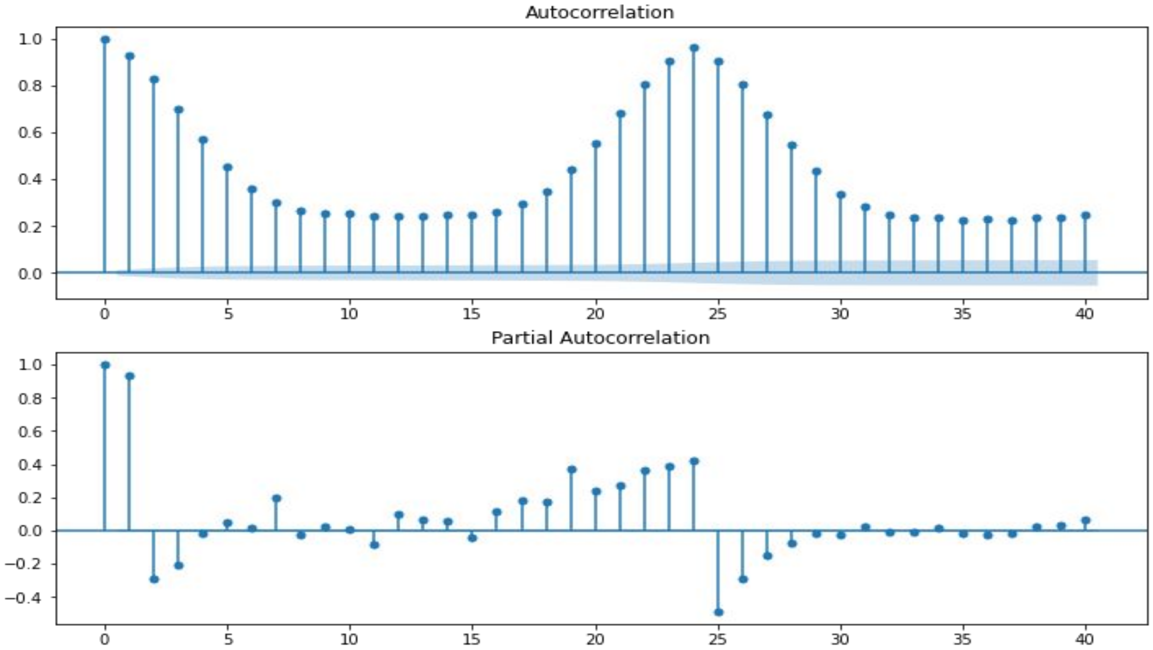


Preprocessing and Training


Autocorrelation
ANC
ATL
BOI
BZN
DEN
DFW
FLL
LAS
LAX
MCO
MIA
MSO
PDX
PHX
SEA
SFO
SJC
TPA

Seasonality - SFO

Strong seasonality with period of 24 (hours)

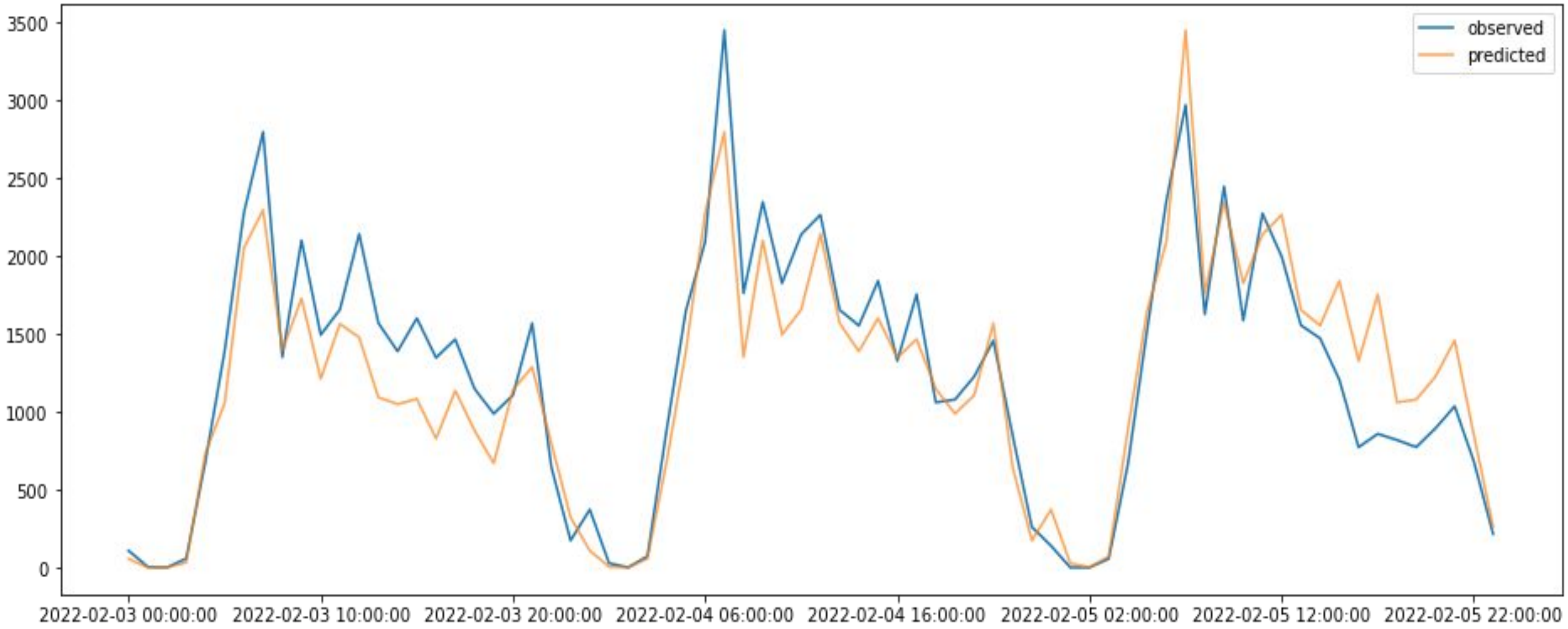


Augmented Dicky-Fuller Test

- Null hypothesis - time series is non-stationary
- Only tests for trend
- Reject null if p-value is small (less than 5%)

Baseline Model - Yesterday’s values

Mean Error = 276 people



Modeling

ARIMA Model

Grid search

(p, d, q) X (P, D, Q, S)

Evaluate Model

Prediction

Future work: