Problem Identification
Predict TSA throughput, using past data starting in 2019 to february 2022 to predict future values.
Why: proper staffing, infrastructure, resources,
Stakeholders: airports, airlines, airport businesses
Challenges: unpredictable events: weather, war, pandemics etc
Data: TSA throughput counts for 18 US airports, by gate. Github repo, hourly data so consider time series analysis

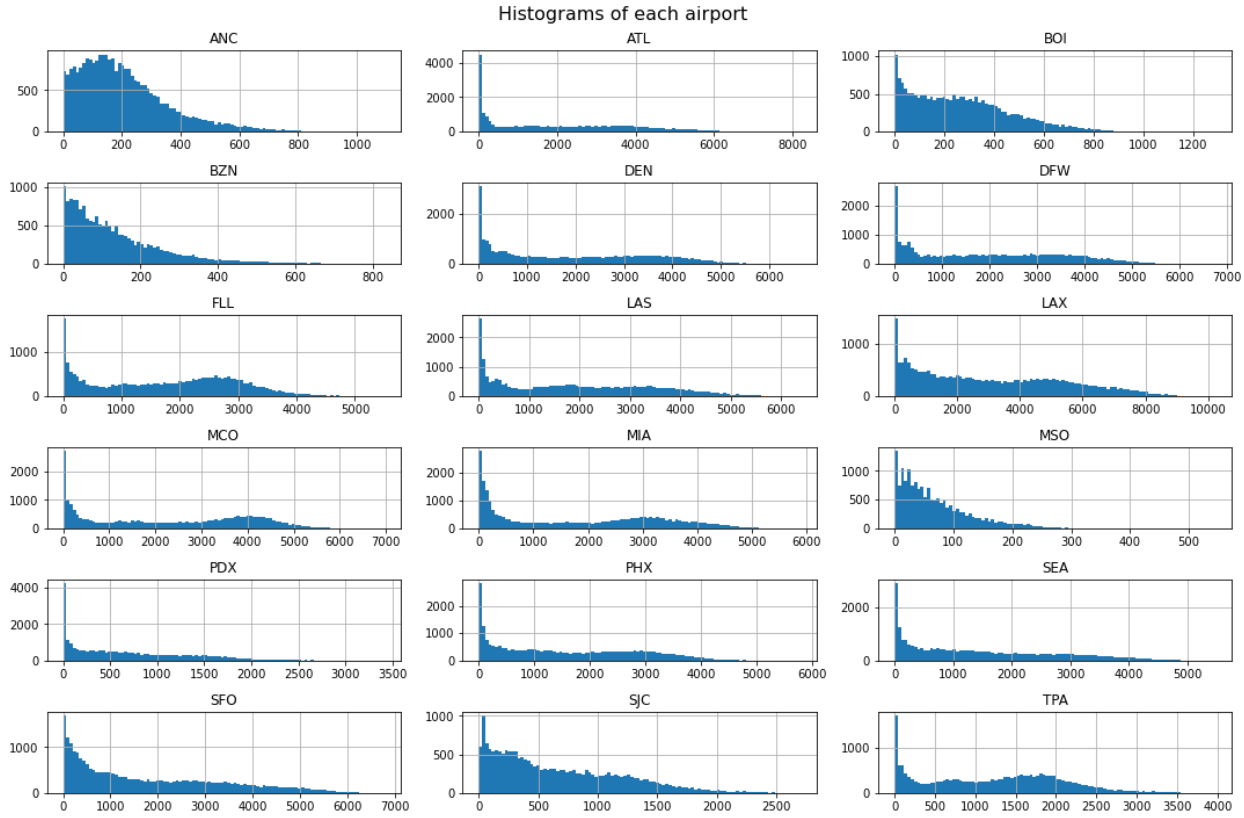
Data Wrangling
Collection - Data in the repo is individual csv files for each individual airport. Within each csv, values are divided into columns for each individual gate. My analysis will be airport-wide and thus gate level details can be aggregated to present totals for each airport.
Downloaded repository. For each individual airport, summed over gate values and joined individual airports into a dataframe with total hourly throughput for each airport. Set the index to datetime and we have hourly time series of 18 US airports between December 2018 and February 2022.

- 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
- PHX - Phoenix
- SEA - Seattle
- SFO - San Francisco
- SJC - San Jose
- TPA - Tampa

Exploration - 27216 entires and 18 columns with hourly date time index.
Data type for all columns is float64.
From the summary statistics, we can identify small vs. large airports. The highest mean and maximum belong to LAX at 3,097 and 10,250. The smallest mean and max values belong to MSO at 69 people and maximum 546. We can divide airports into small and large based on number of people going through on average. Airports with less than 1,000 individuals on average are ANC, BOI, BZN, PDX, SJC and MSO. These we classify as small. The remaining airports have mean throughput over 1,000 individuals with LAX, ATL, and MCO above 2,000.
No duplicates.

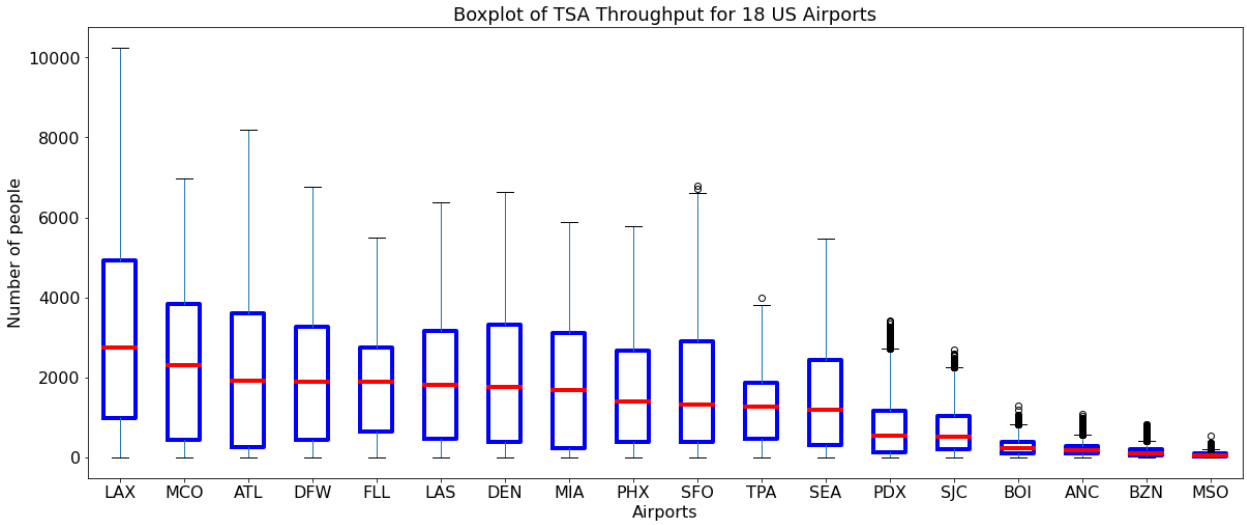
EDA

Histograms:



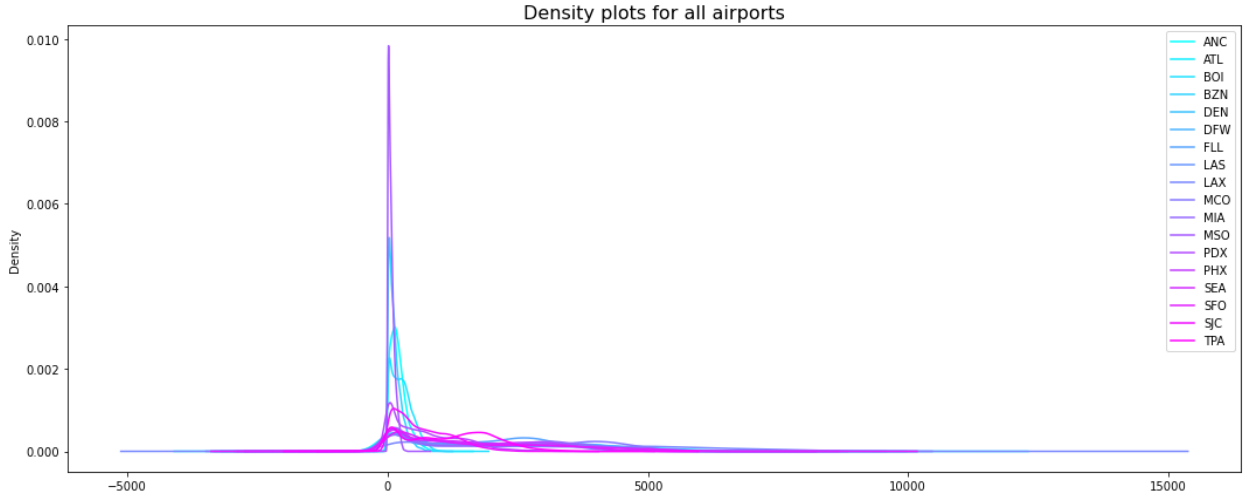
Distributions for nearly all airports are quite high for lower throughput counts and for most airports, values decrease as number of people going through for a give time period increases (quite rapidly). The large airports have a second peak where a higher numbers of people are going through TSA for a given hourly time period.

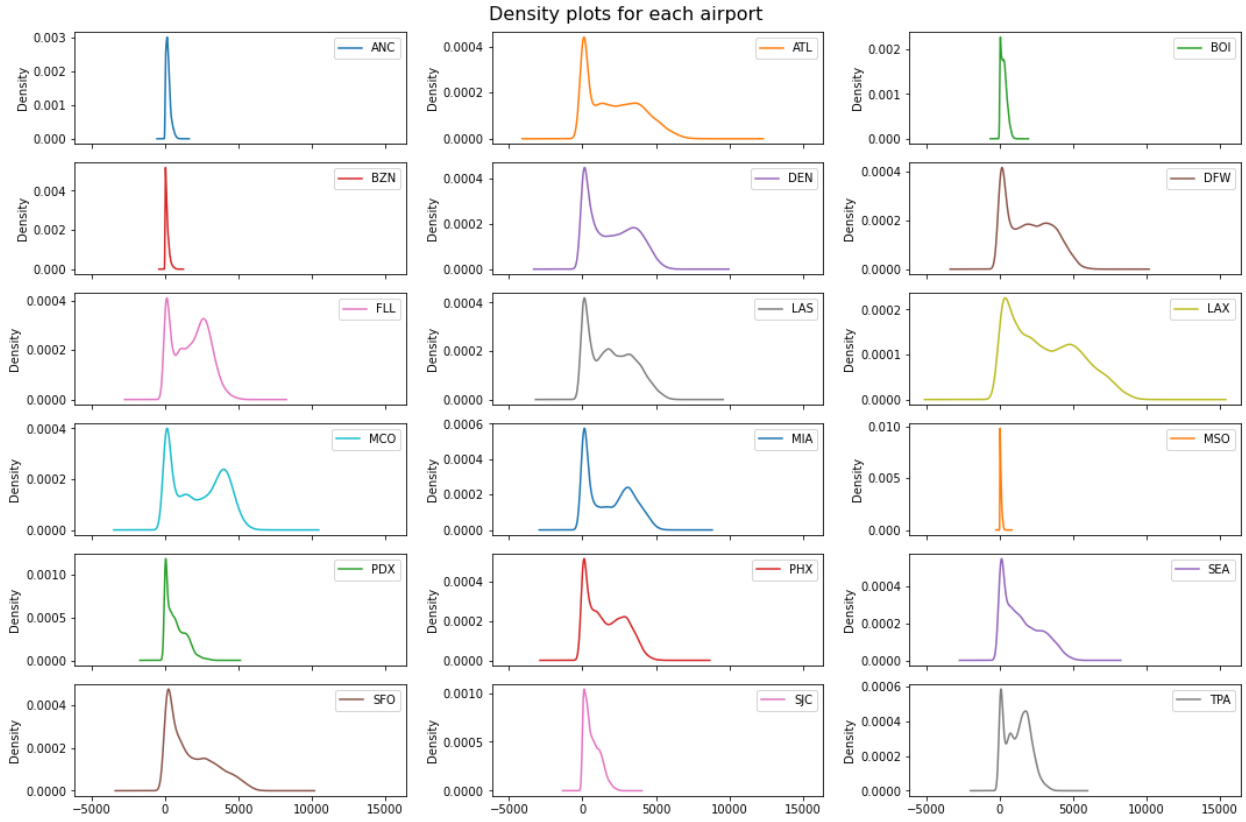
Boxplots:



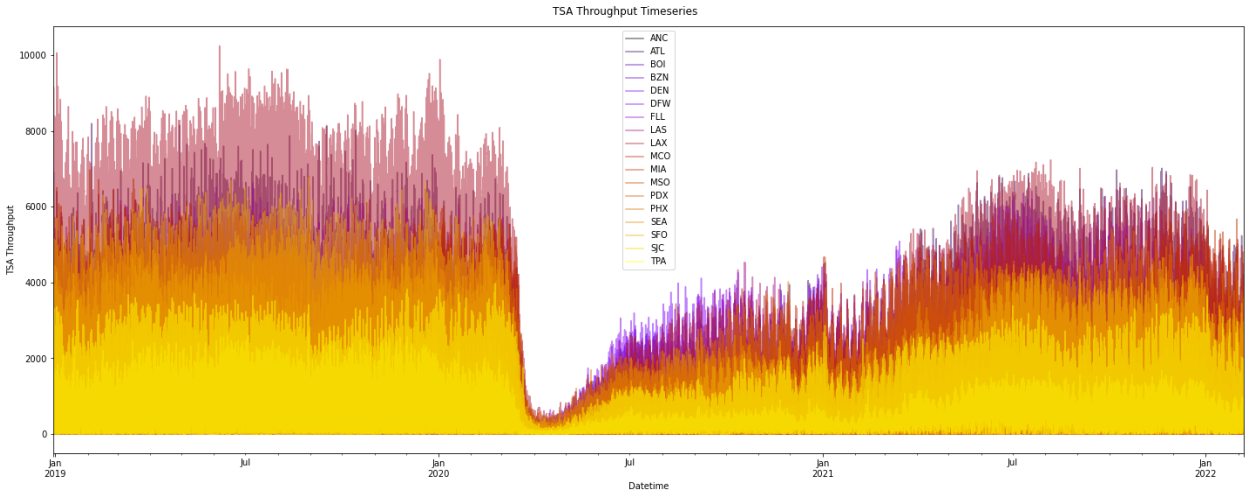
From the boxplots, it is easy to identify small vs large airports. LAX gets the highest throughput while Boise International, Anchorage, Bozeman International and Missoula Montana airports are the smallest airports on the list.

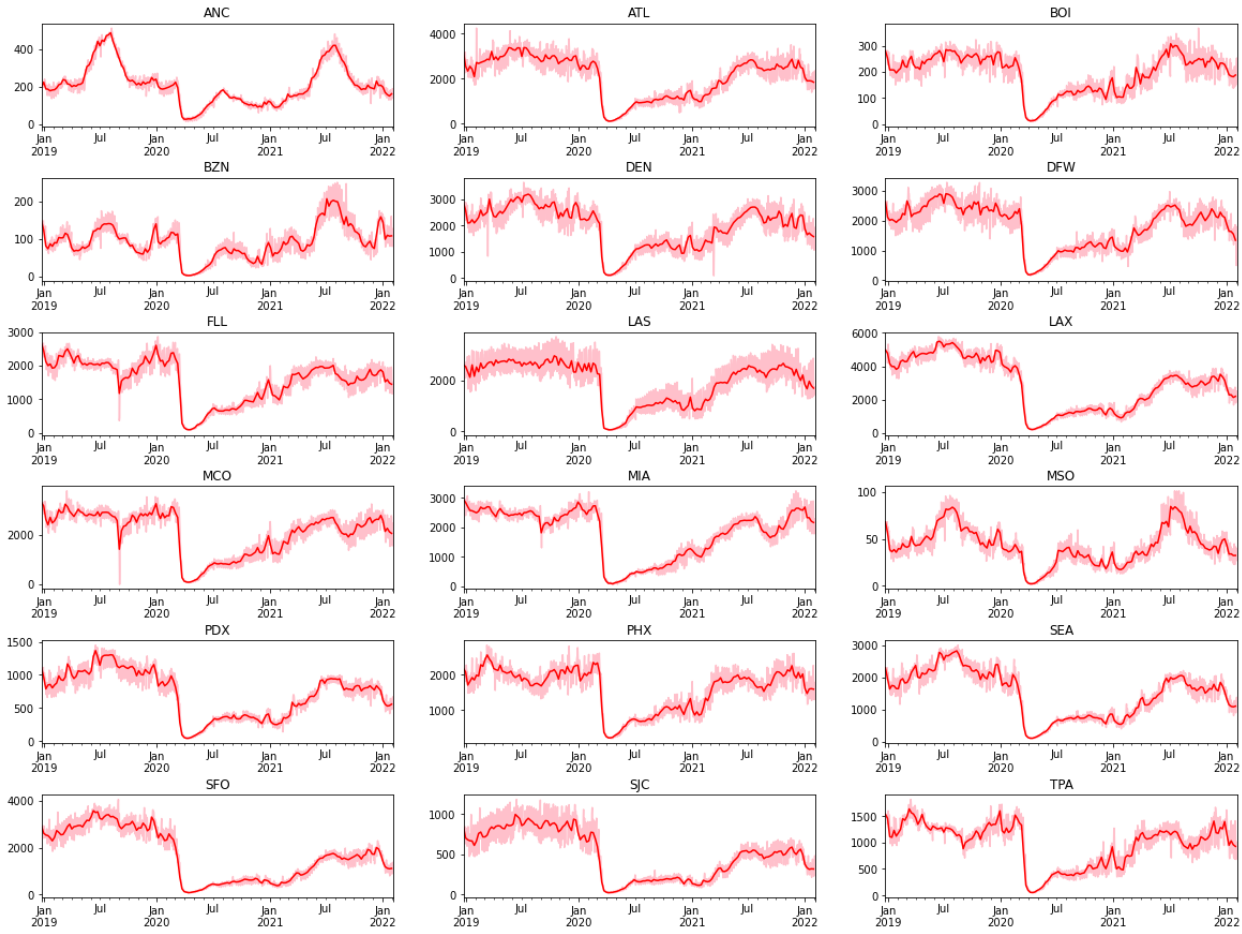
Density plots:



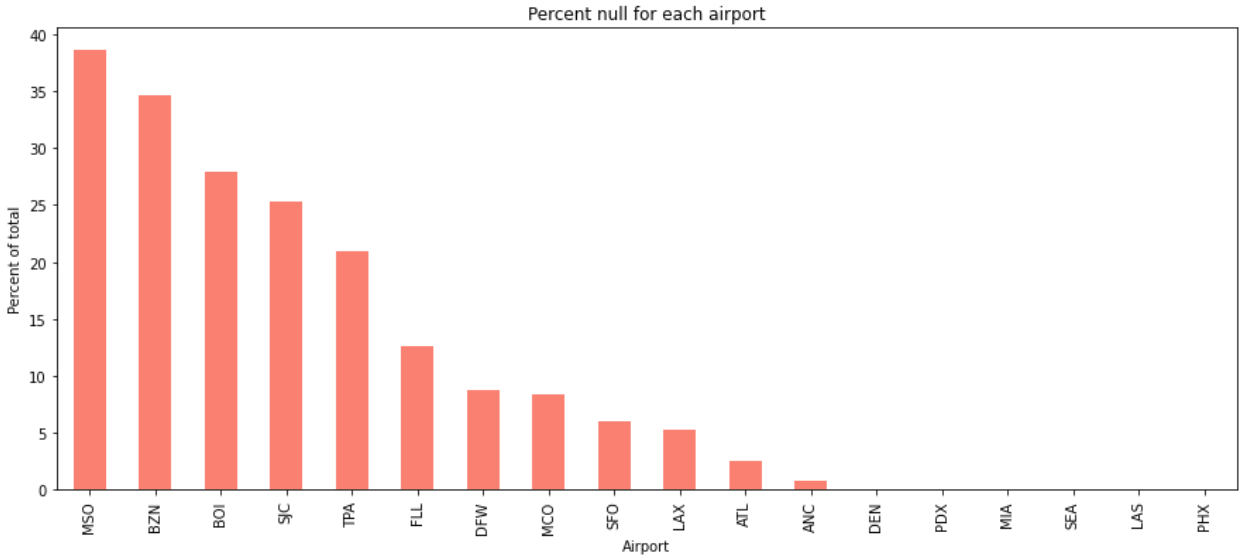


From the density plots, we can better see the second peak appearing for larger throughput values as airports become larger.
Time Series plots:

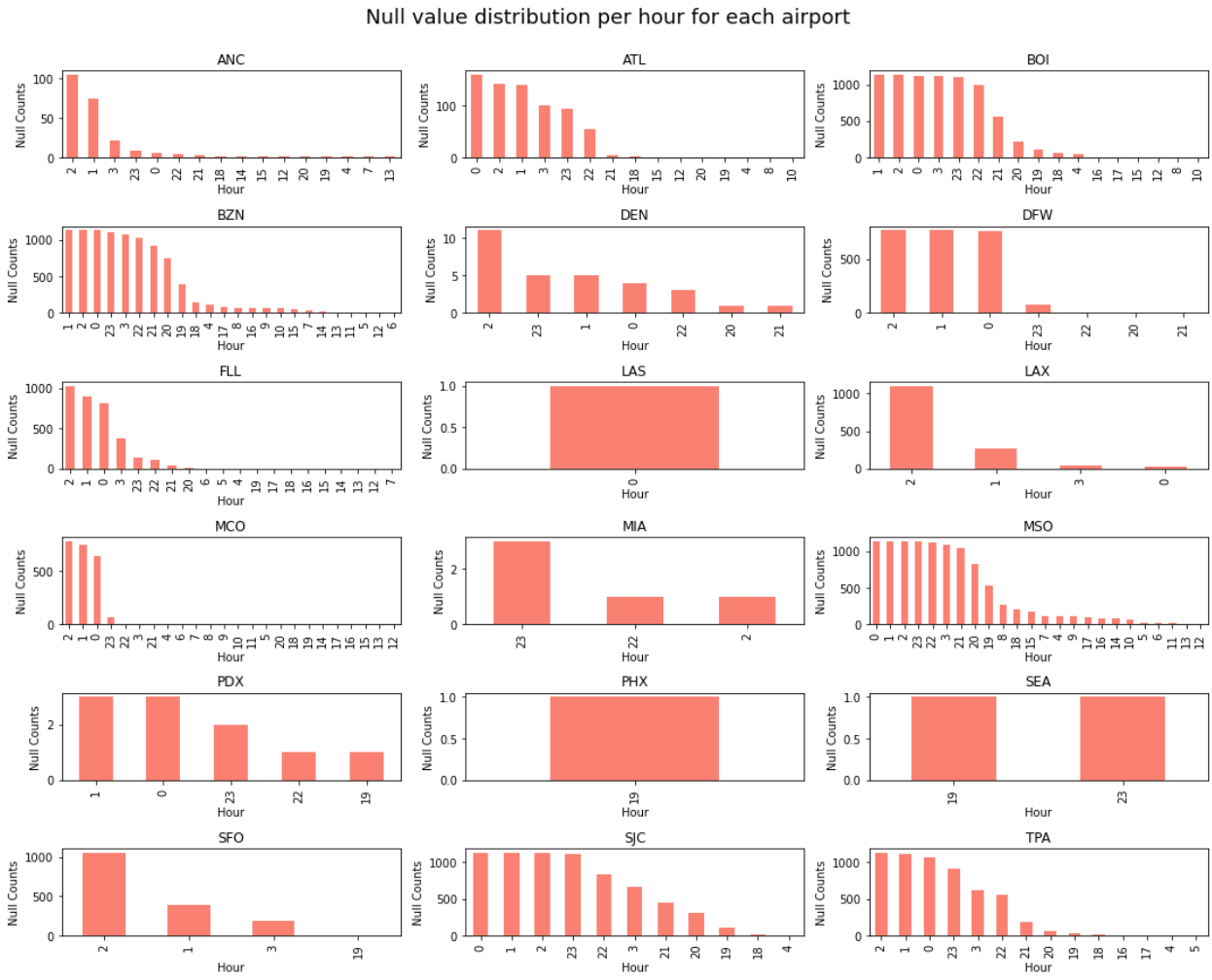




Null values:

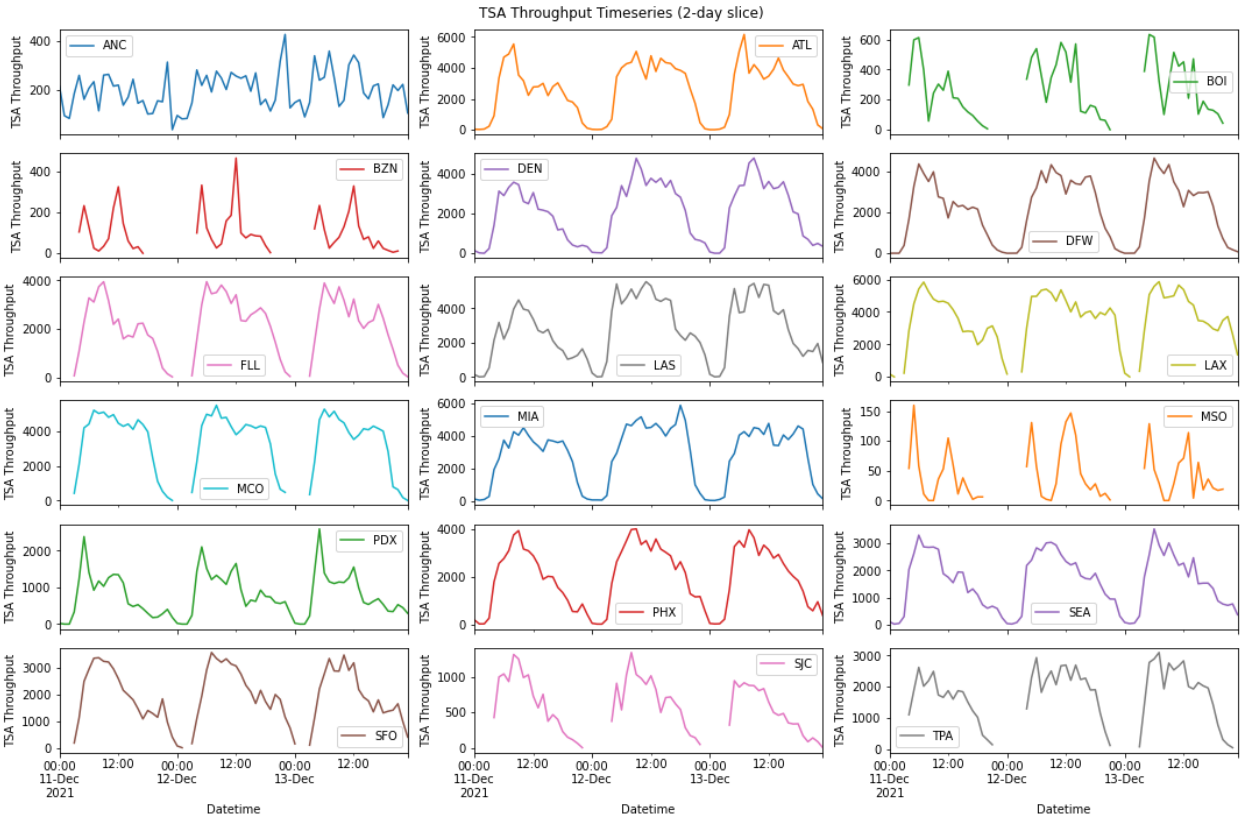


The small airports such as MSO, BZN, BOI, SJC and TPA have relatively high numbers of NA values (more than 20 percent).

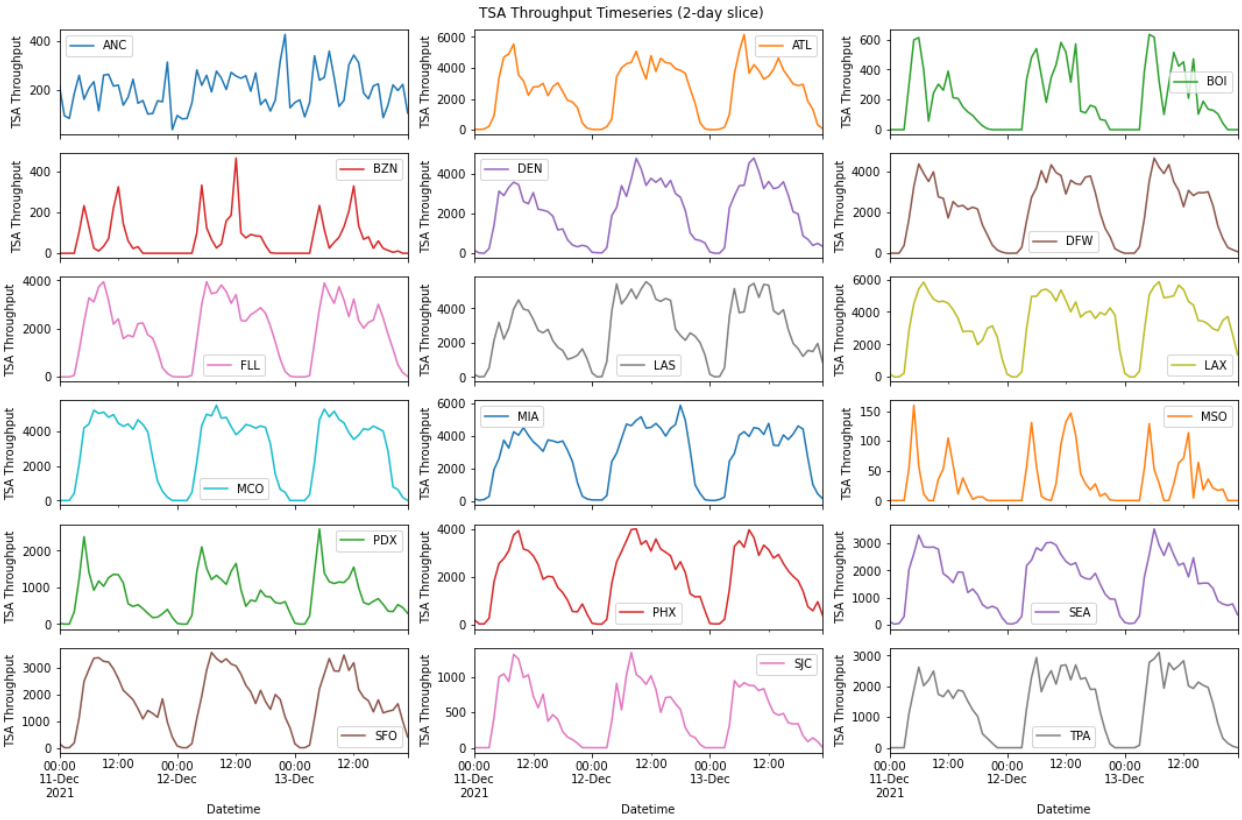


The small airports (ANC, BOI, BZN, MSO, SJC, and TPA) have a larger time range where they are missing values. Phoenix and LAS only have one missing value. Generally, it appears that all the missing values are in the middle of the night where airports are quite and not as many flights goint out and this range is much wider for smaller airports when fewer outgoing flights would be scheduled in the evening and early morning hours (between 7 pm to 4 am).

We can visualize where the null values are if we zoom into our time series plots for each airport. From these plots, it is easier to see that most missing values are in the evening or early morning hours.

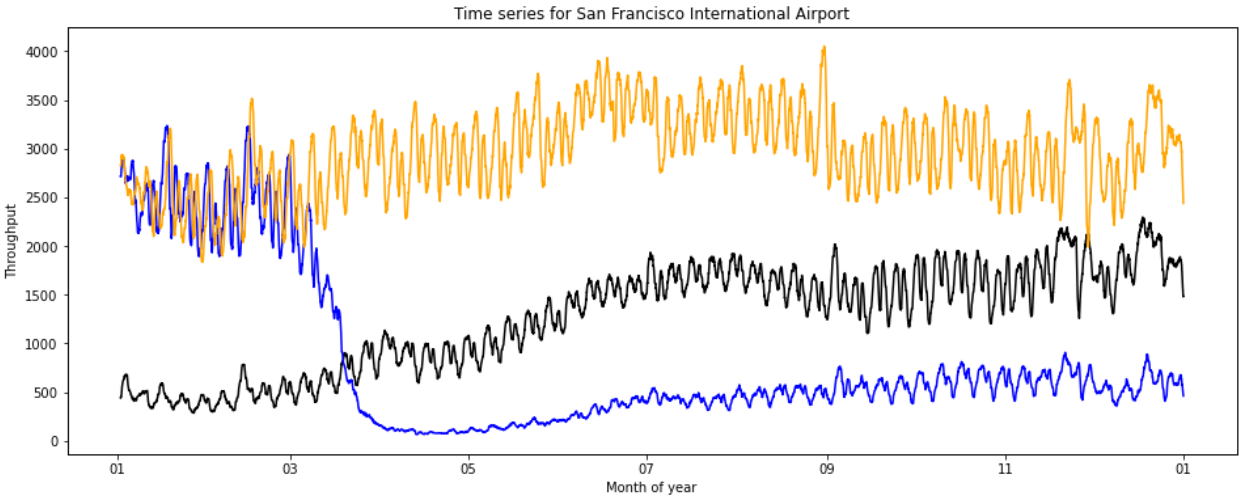


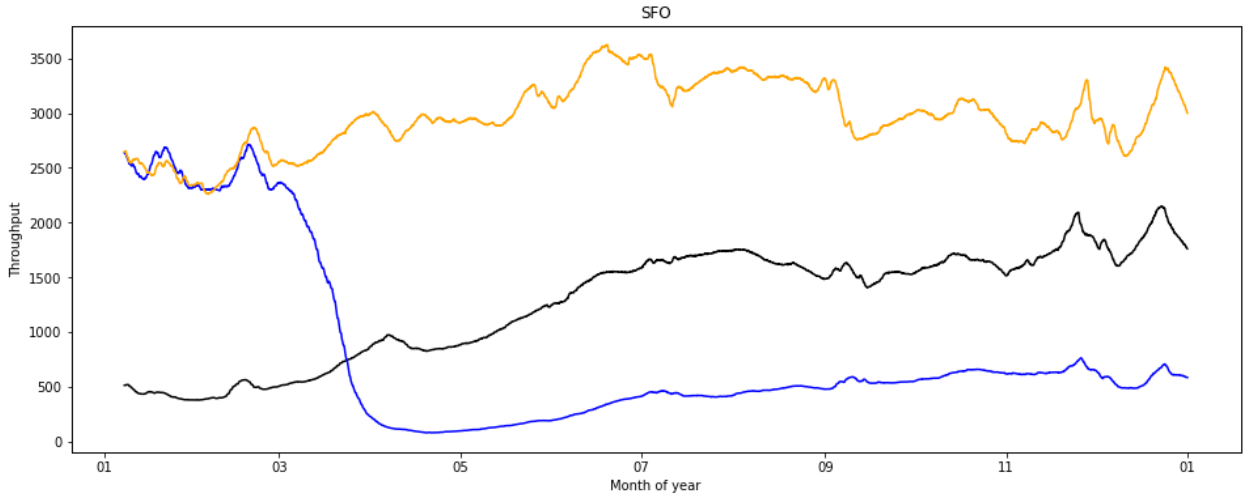
The zoomed in time series above allow us to visualize our missing values better and we can further verify that missingness occurs in the evening and early morning hours and from experience in small airports such as those in Montana (BZN and MCO) as well as researching flight departures, we can gather that fewer to no outbound flights occur around midnight until 4 am and thus it is most likely that no one passed through TSA checkpoints for those times where values are missing. I shall then fill in the missing values with zero and visualize the data again.

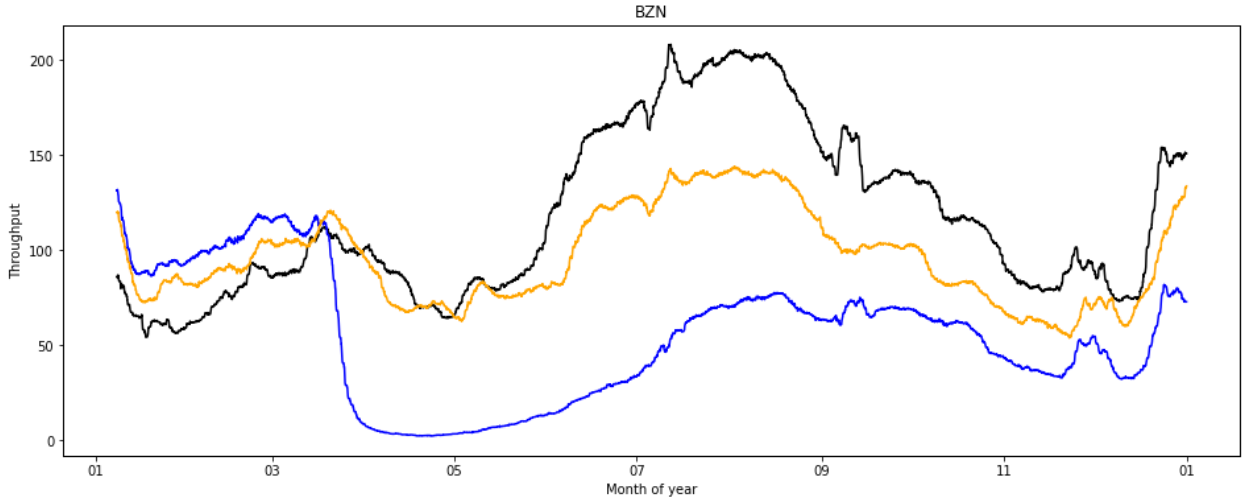


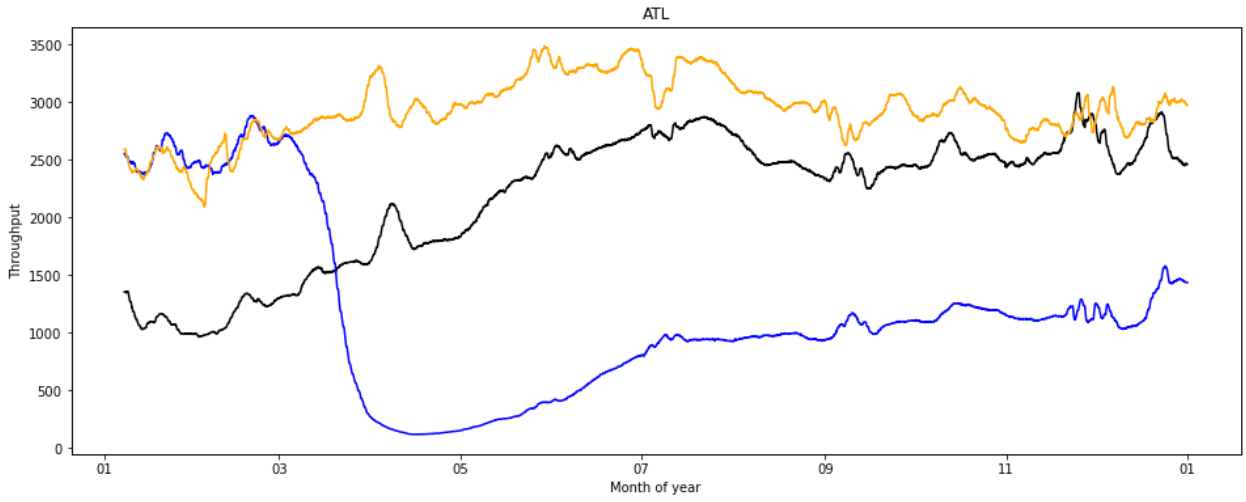
I will also keep a version of the dataframe that includes all null values for potential future analysis where we may include missingness as a feature and allow the model to determine what the best possible value might be.

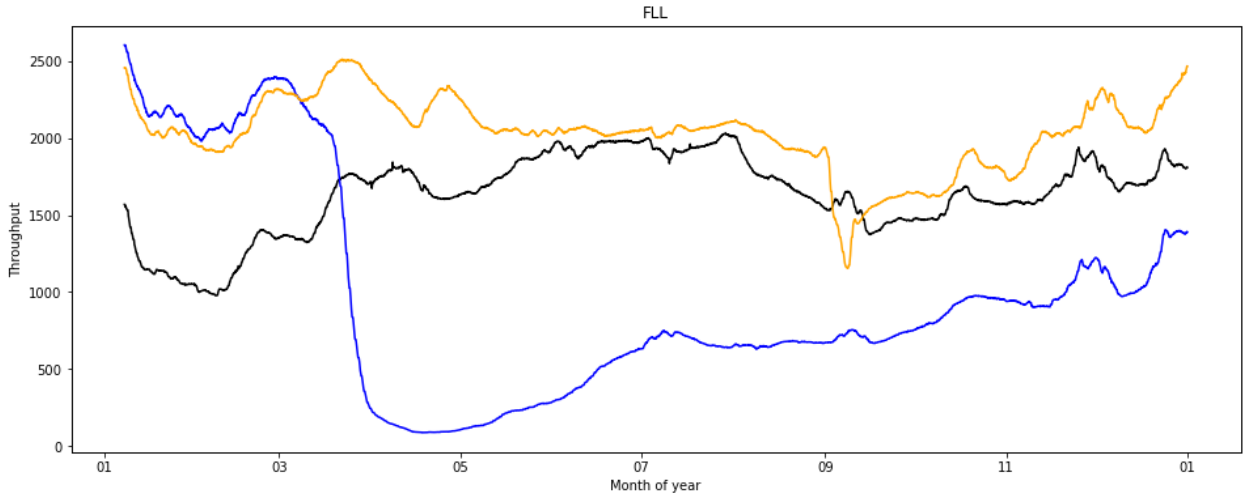
Yearly Trends:

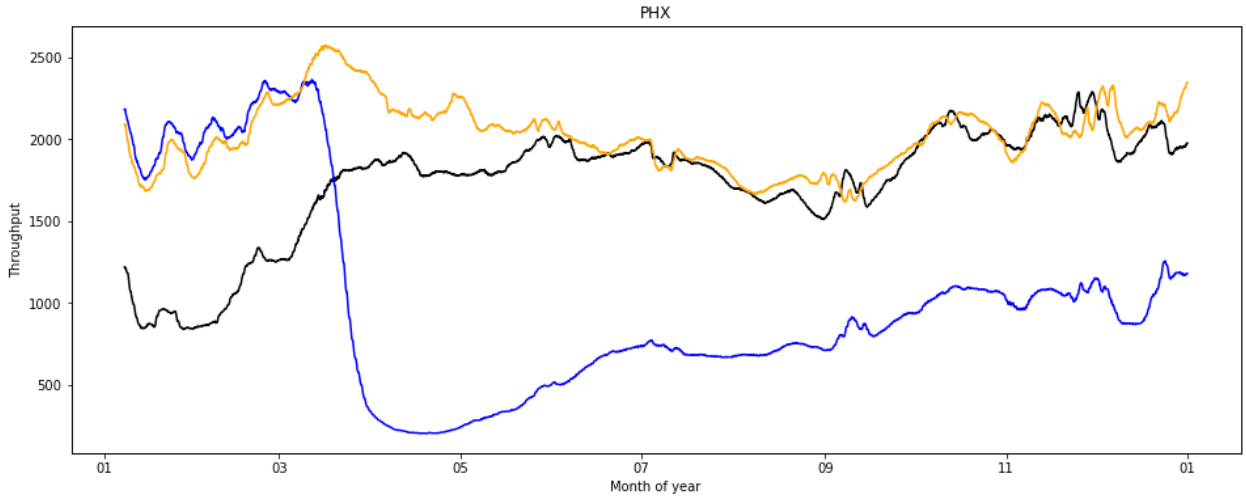



SFO
Throughput
0
500
1000
1500
2000
2500
3000
3500
01
03
05
07
09
11
01
Month of year


BZN
Throughput
0
50
100
150
200
01
03
05
07
09
11
01
Month of year


ATL
Throughput
0
500
1000
1500
2000
2500
3000
3500
01
03
05
07
09
11
01
Month of year

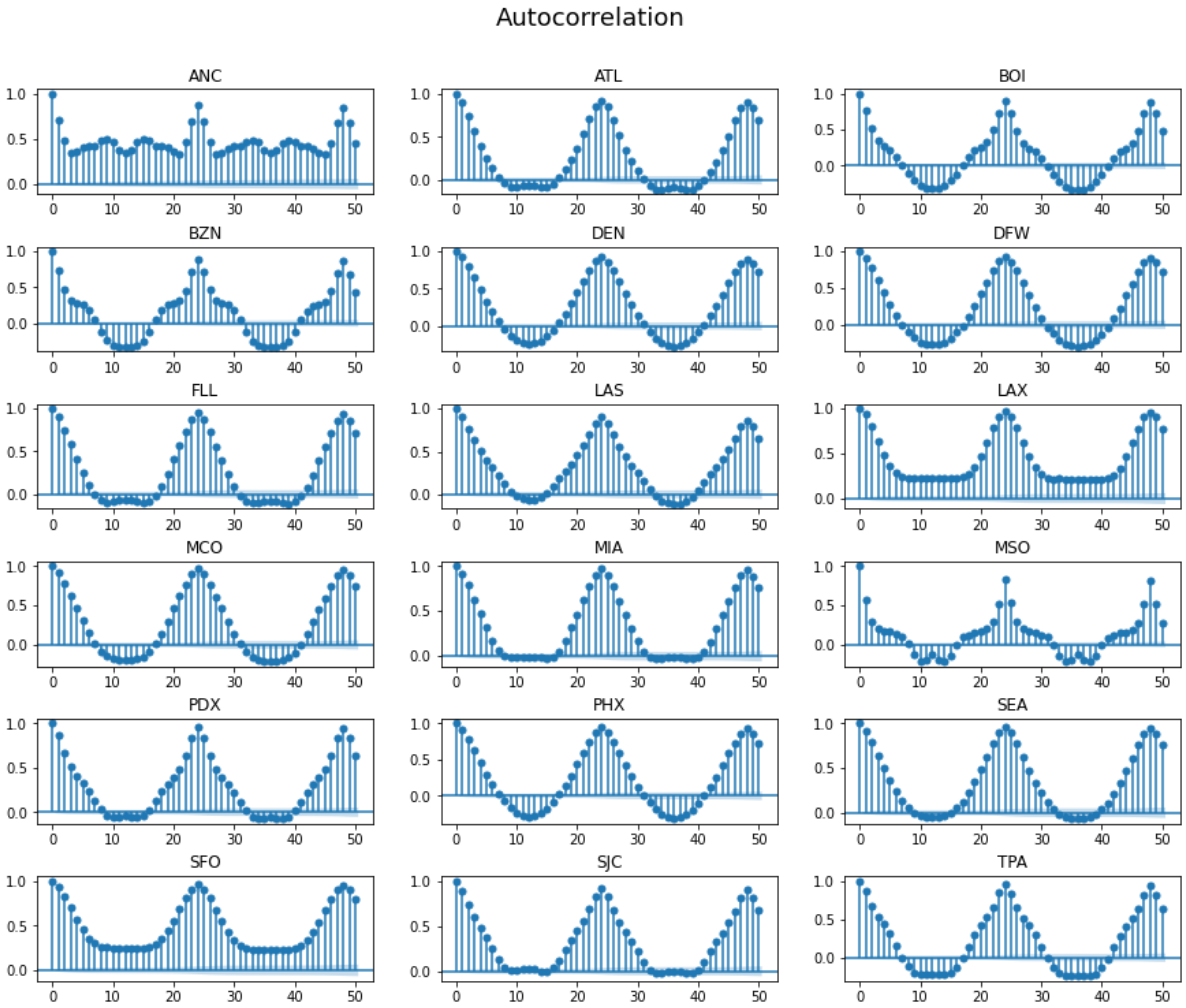

FLL
Throughput
2500
2000
1500
1000
500
0
01
03
05
07
09
11
01
Month of year

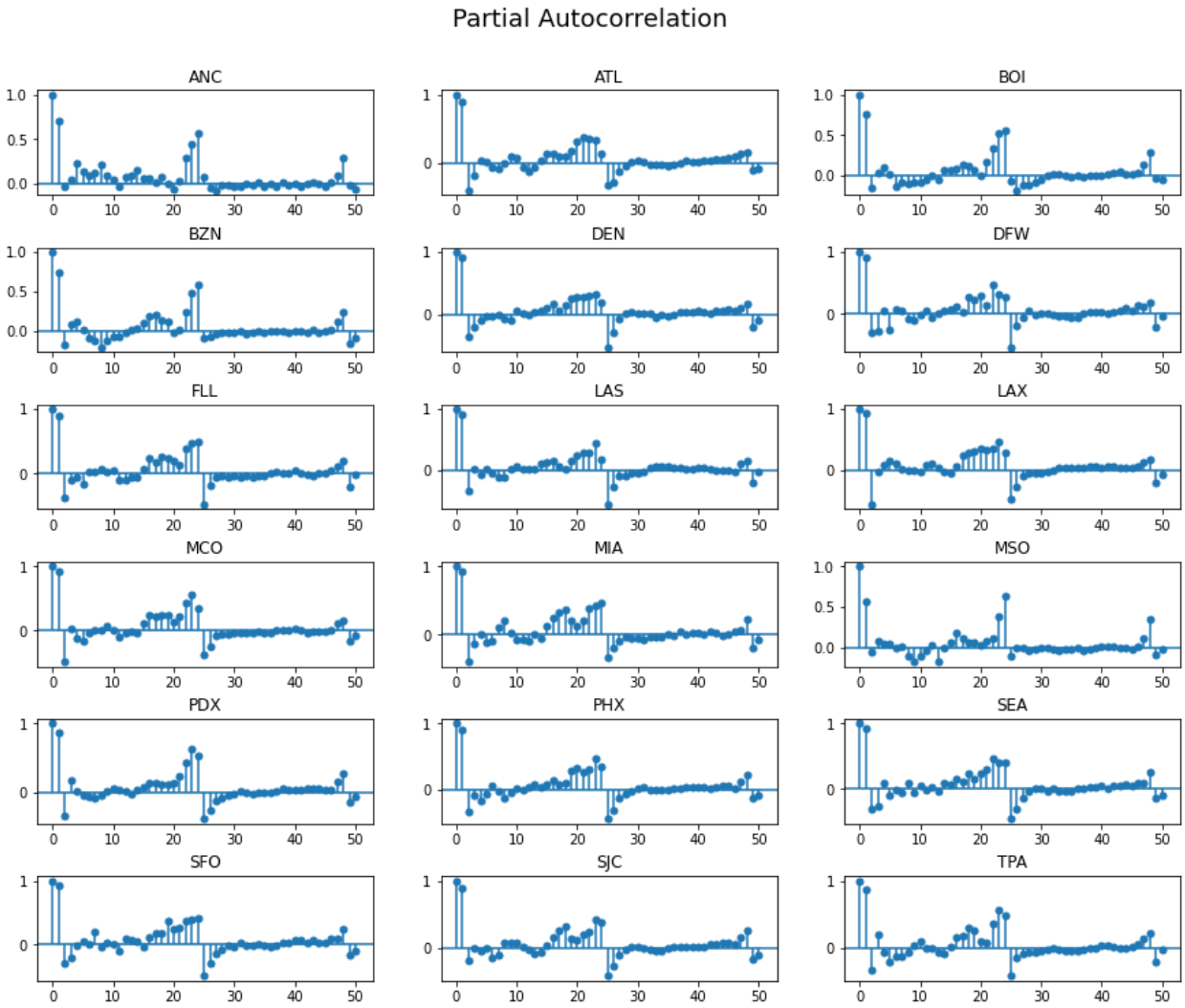

PHX
Throughput
2500
2000
1500
1000
500
01
03
05
07
09
11
01
Month of year

Preprocessing and training

Seasonality: Strong seasonality with period 24

- Autocorrelation plots
- Partial Autocorrelation plots



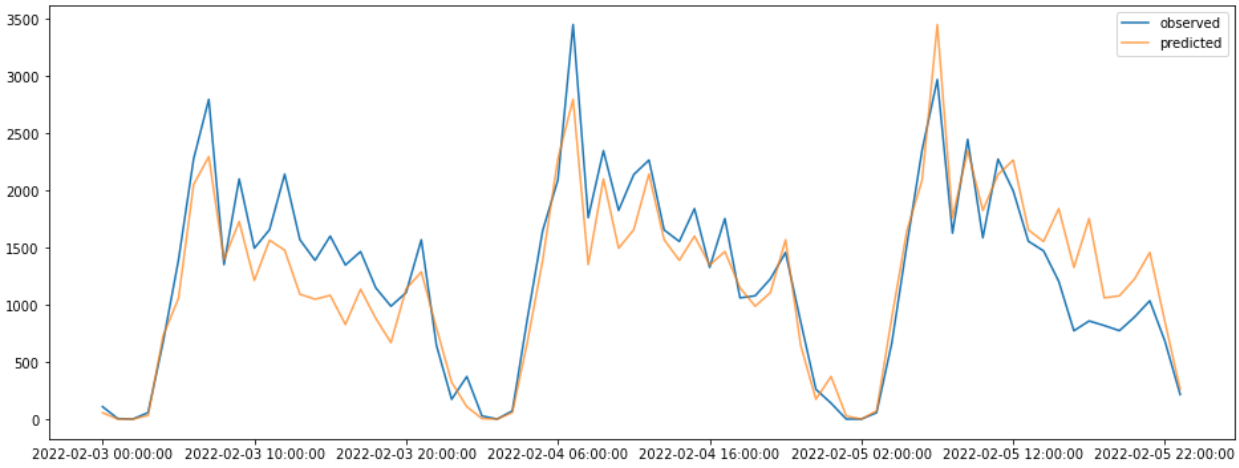


Stationarity:

- Use Dickey Fuller to test for trend stationarity
- Time series plot
- Autocorrelation plot

Baseline: Use yesterday's values

Mean error: 765 people



Modeling