Data Collection

- Data of Passenger Volume by Train Stations obtained from DataMall, an LTA Open Data initiative API.
- 3 months of tap in and tap out passenger volume by weekdays and weekends for each individual train station

1	YEAR_MONTH	DAY_TYPE	TIME_PER_HOUR	PT_TYPE	PT_CODE	TOTAL_TAP_IN_VOLUME	TOTAL_TAP_OUT_VOLUME
2	2021-07	WEEKDAY	22	TRAIN	NS28	730	137
3	2021-07	WEEKENDS/HOLIDAY	22	TRAIN	NS28	387	83
4	2021-07	WEEKDAY	0	TRAIN	DT10	10	55

1st three rows of the Passenger Volume by Train Stations for the July 2021.

Making sense of the data:

- TIME_PER_HOUR: Hour of the day. E.g. 22 = 2200hrs to 2259hrs.
- On a typical WEEKEND/HOLIDAY of July 2021, from 2200hrs to 2259hrs, at Train Station NS28, the passenger volume of tap in and tap out are 387 and 83 respectively.

Data Collection

Objective:

- Obtain suitable passenger arrival values for our Simulation model timeseries exponential distribution.

Considerations:

- The data only has a single value for passenger arrivals and exits at each station.
- Each station has 2 observations - weekday and weekend.
- Maximum Likelihood Estimator for Poisson arrivals = sample mean of observed hourly arrivals.
- Data collected spans across 3 months - "2021-07", "2021-08", "2021-09". **Note** the LTA API only provides the arrival volume data from the three most recent months.

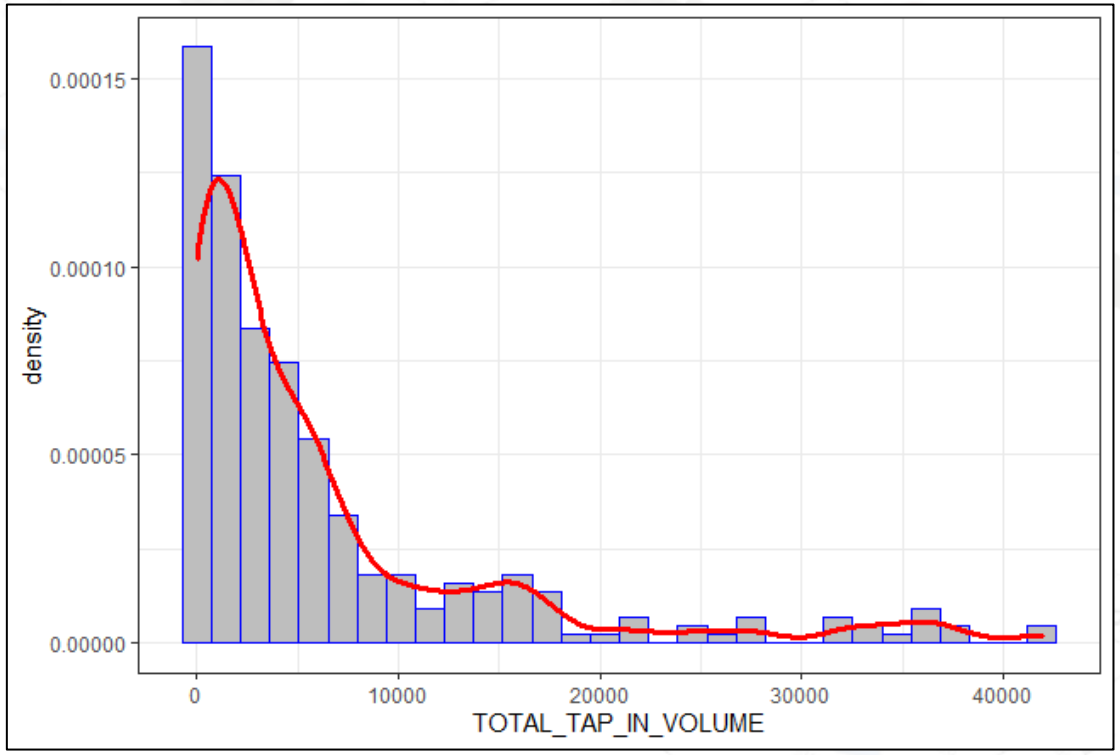
Assumptions:

- Peak hours are 0600hrs to 0900hrs (morning), 1700hrs to 1900hrs (evening).
- Passenger arrivals at stations connected to more than 1 line are assumed to be spilt evenly between all connected lines. E.g. Passenger arrivals at CE1/DT16 are spilt into 2 evenly.

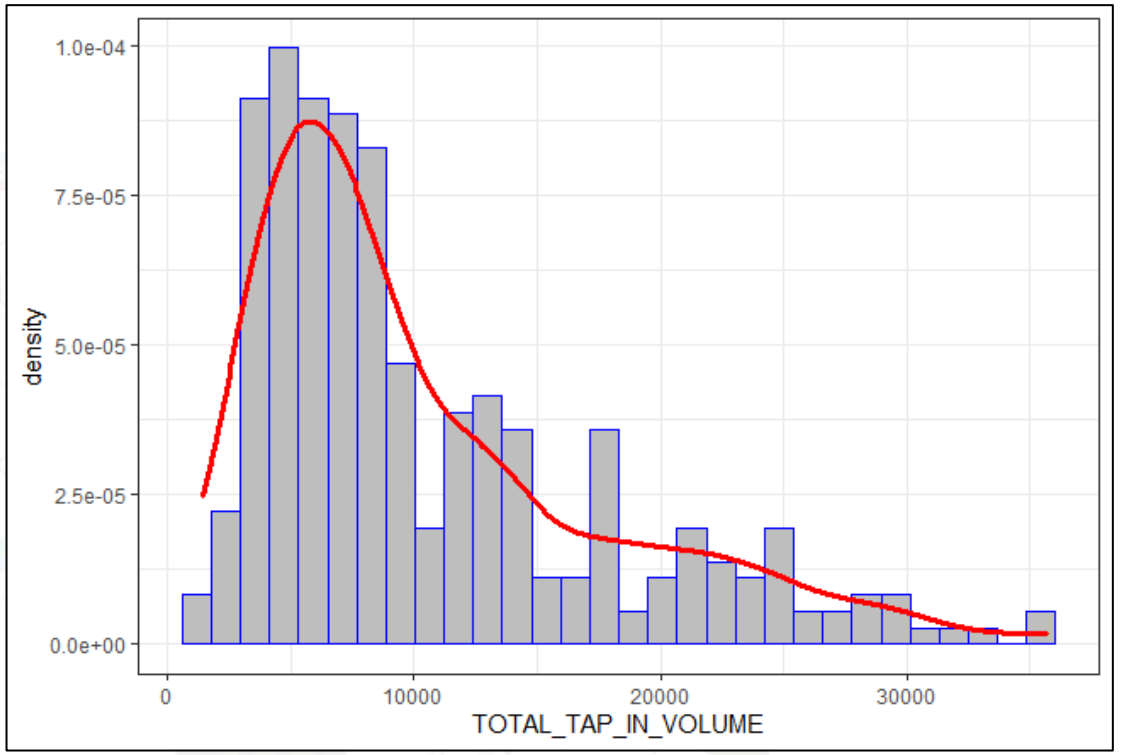
Data Visualization

- Combined all three csv datasets into a single large dataset.
- Split the single large data set into 6 categories:
    1. Weekday Morning Peak;
    2. Weekday Evening Peak;
    3. Weekday Non-Peak;
    4. Weekend Morning Peak;
    5. Weekend Evening Peak;
    6. Weekend Non-Peak.
- Plot Histogram with the density line over it for all 6 categories.

Data Visualization

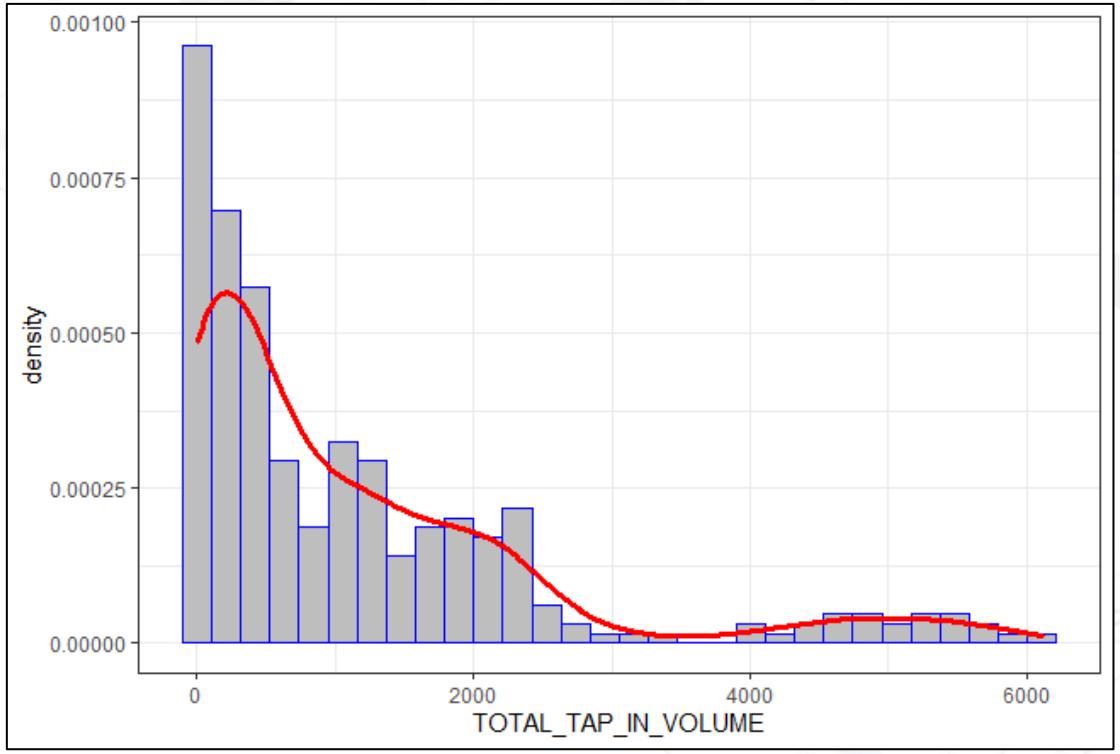


*Passenger Tap-In Volume for **Weekday Morning Peak***

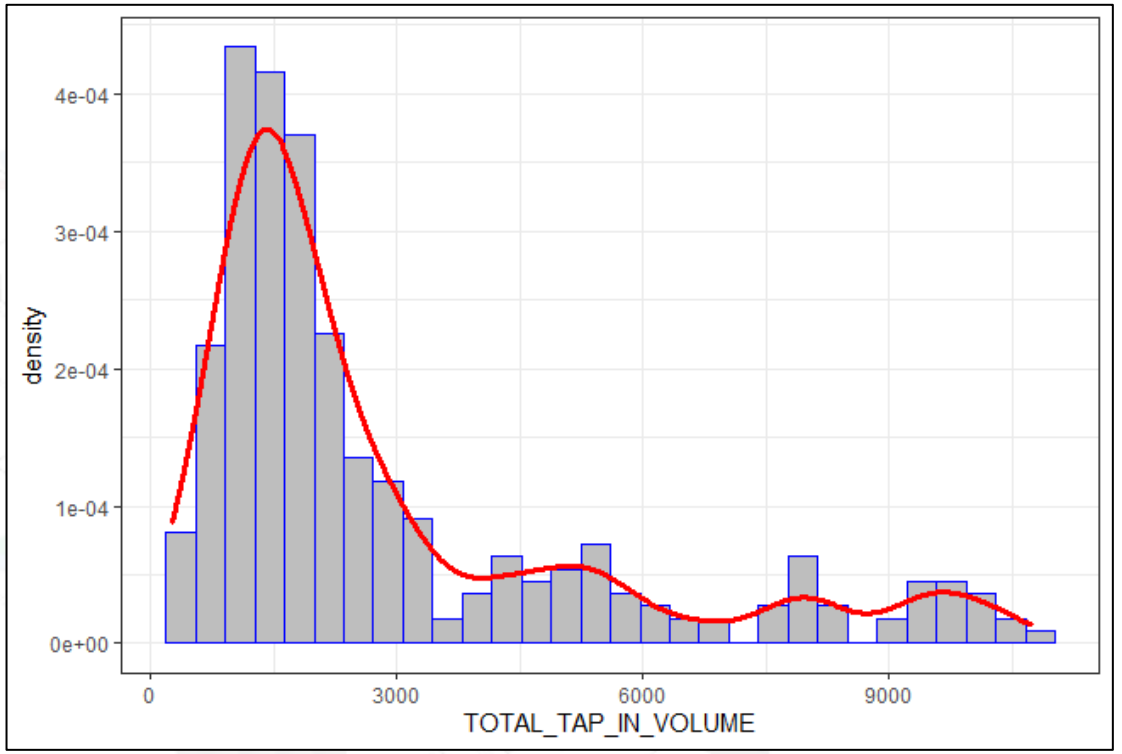


*Passenger Tap-In Volume for **Weekday Evening Peak***

Data Visualization

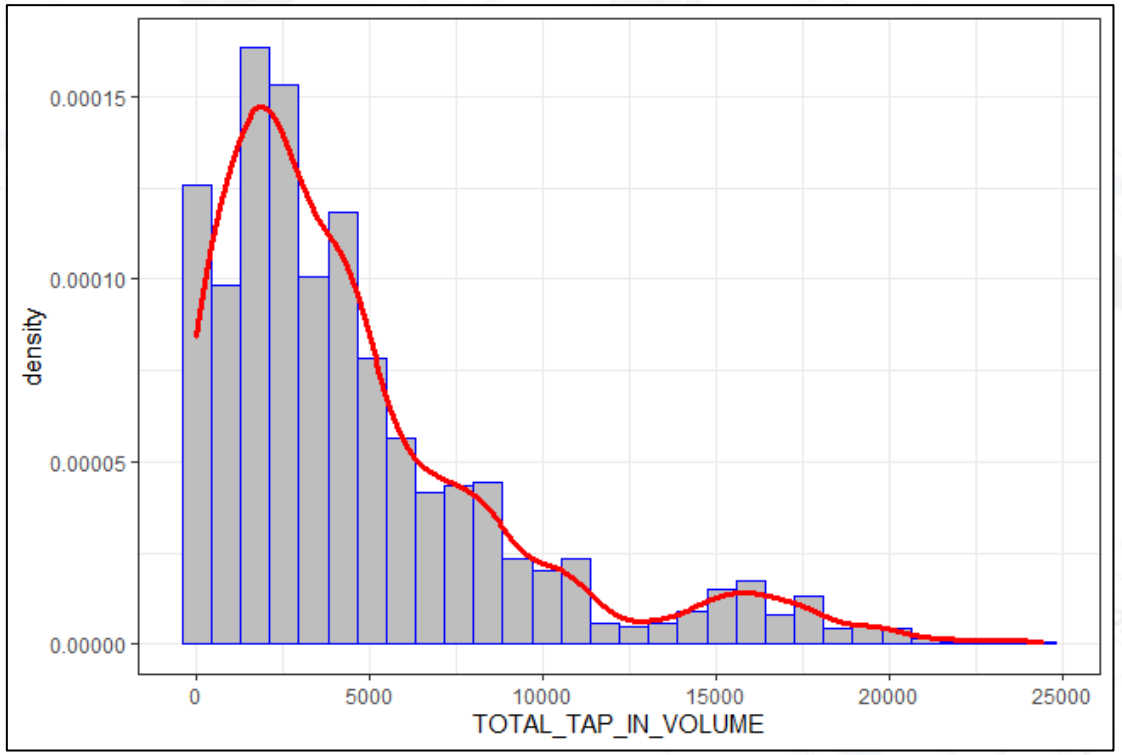


*Passenger Tap-In Volume for **Weekend Morning Peak***

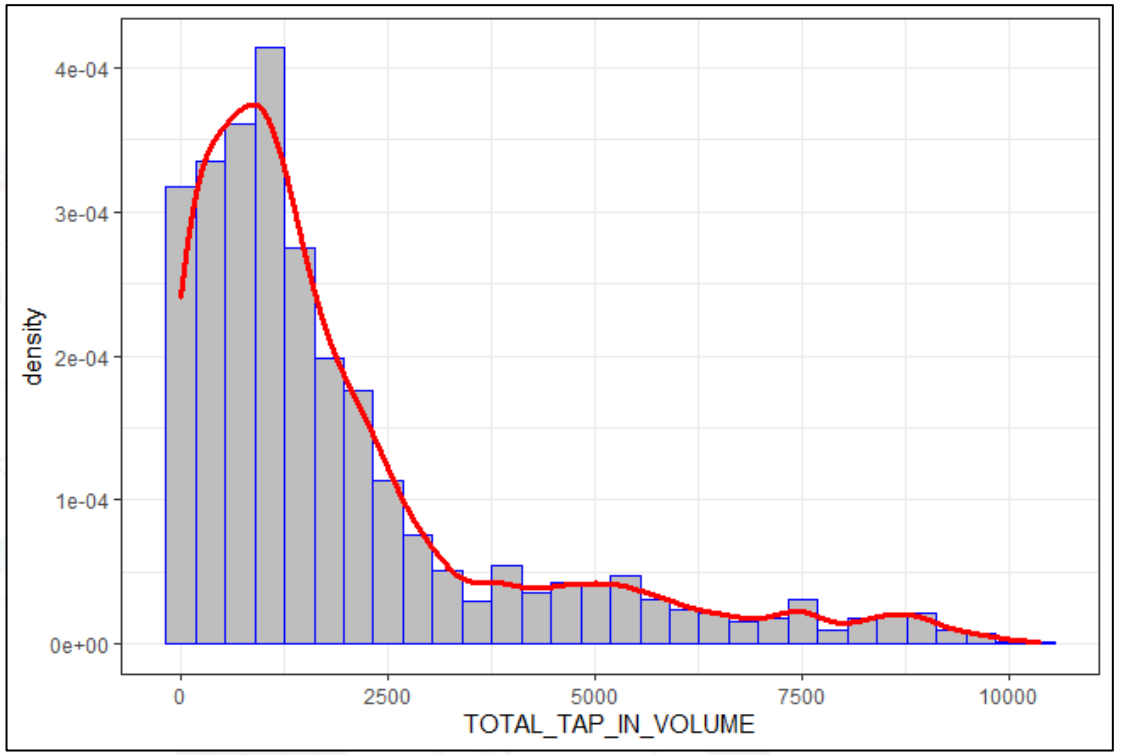


*Passenger Tap-In Volume for **Weekend Evening Peak***

Data Visualization



*Passenger Tap-In Volume for **Weekday Non-Peak***



*Passenger Tap-In Volume for **Weekend Non-Peak***

1st Statistical Test

Objective:

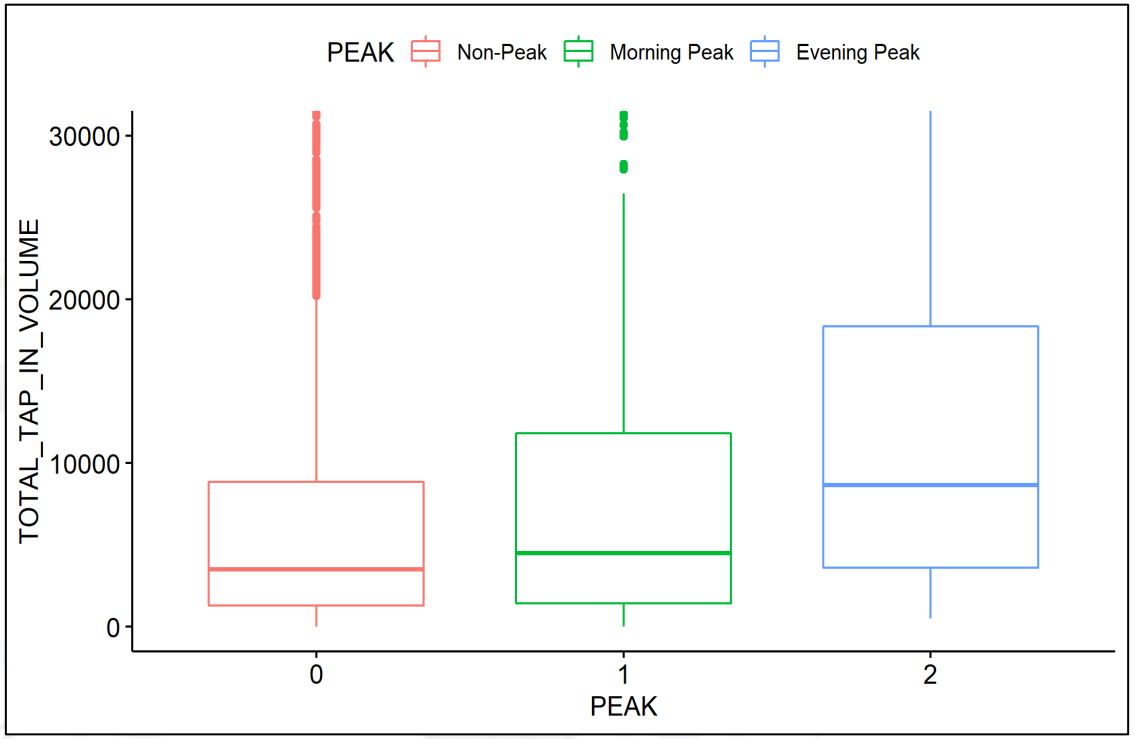
- To determine if the morning, evening peak hours and non-peak hours have statistically significant different number of passenger arrivals.
- To validate our assumption for splitting the day into 3 different categories: morning, evening peak hours and non-peak hours when building our timeseries exponential passenger arrival distribution in JaamSim.

How is it done:

- Kruskal-Wallis test used instead of ANOVA as responses are clearly not normally distributed.
- Kruskal-Wallis test conducted on weekday and weekend passenger arrival for each of the 3 months, comparing the mean arrivals of morning peak hours, evening peak hours and non-peak hours.
- Hence 6 Kruskal-Wallis tests were done:
    i. 3 for Weekday: July, August, September 2021;
    ii. 3 for Weekend: July, August, September 2021.

1st Statistical Test - Results

- P-values for all 6 tests obtained were all smaller than 0.05.
- Conclude at 95% confidence level that the mean arrivals per hour are statistically different between morning, evening peak hours and non-peak hours for every month weekdays and weekends.
- Hence it is valid to split a given weekday or weekend into 3 timeframes: morning, evening peak hours and non-peak hours.
- On the right is a boxplot (vertically cropped to see differences) to visualize just 1 of the 6 tests.



(Zoomed-in) Box-plot of the Passenger tap-in volume during ***weekday*** *for* ***July 2021****.*

2nd Statistical Test

Objective:

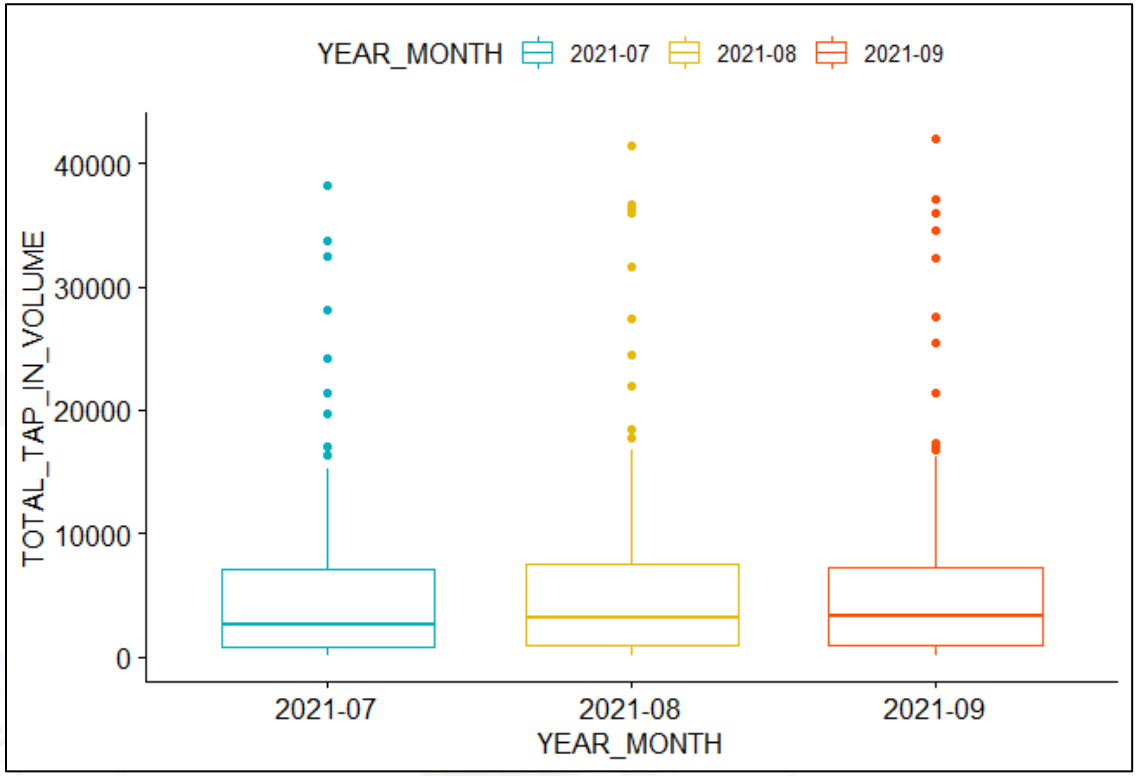
- To determine if the number of passenger arrival comes from the same distribution across the 3 months.
- To validate our assumption that the distribution of the number of passenger arrivals come from the same distribution.

How is it done:

- Kruskal-Wallis test conducted on each of the 6 categories, comparing the mean arrivals of each month.
- Hence another 6 Kruskal-Wallis tests were done:
    i. 3 for Weekday: Morning Peak hour, Evening Peak hour, Non-peak hour;
    ii. 3 for Weekend: Morning Peak hour, Evening Peak hour, Non-peak hour.

2nd Statistical Test - Results

- P-values for all 6 tests obtained were all greater than 0.05.
- Conclude at 95% confidence level that the mean arrivals per hour are **NOT** statistically different between each month.
- Hence our assumption that the distribution of passenger arrivals come from the same distribution is valid.
- On the right is a boxplot to visualize just 1 of the 6 tests.



*Box-plot Passenger tap-in volume for **weekday morning peak hours.***

Data Manipulation

Process:

1. Identify which observations are from morning peak hours, evening peak hours, and non-peak hours;
2. Identify which stations are connectors to other lines and spilt the arrivals accordingly;
3. Calculate the mean number of hourly arrivals (tap in) for the 6 identified groups.

DAY_TYPE	TIME_PER_HOUR	PT_TYPE	PT_CODE	TOTAL_TAP_IN_VOLUME	TOTAL_TAP_OUT_VOLUME
WEEKDAY	22	TRAIN	NS28	730	137
WEEKENDS/HOLIDAY	22	TRAIN	NS28	387	83
WEEKDAY	0	TRAIN	DT10	10	55
WEEKENDS/HOLIDAY	0	TRAIN	DT10	3	28
WEEKDAY	10	TRAIN	EW16/NE3	14824	21782
WEEKENDS/HOLIDAY	10	TRAIN	EW16/NE3	4243	5240
WEEKENDS/HOLIDAY	9	TRAIN	CC14	2172	1053
⋮					
WEEKENDS/HOLIDAY	15	TRAIN	CC25	2754	1672
WEEKENDS/HOLIDAY	12	TRAIN	CE1/DT16	4133	8609
WEEKDAY	12	TRAIN	CE1/DT16	6343	11666
WEEKDAY	10	TRAIN	SW6	2492	1350

For example, calculating mean for **Weekday Non-Peak** for DT line

Results: Simulation Parameters

- Average number of passenger **arrivals per hour:**

	Morning Peak	Evening Peak	Non-Peak
Weekday	5100	7980	3480
Weekend	840	1980	1380

- From the above values, we can then sample and create timeseries data such that we can run 1 day (finite cycle) of simulation using the non-stat Exponential Distribution in JaamSim. More details in the Model Documentation slides.

Limitations

Regarding the data distribution:

- Although we claim that the passenger arrivals follow a Poisson distribution, the sample mean is observed to exceed the sample variance (overdispersion).
- Since the data is over dispersed, a Negative Binomial distribution might be more appropriate in providing an estimate for the passenger arrival rate.

Regarding the LTA API:

- Does not provide the passenger tap-in volume across the whole day itself. Latest addition to the API only provides passenger volume across the day in categorical form (Low/Middle/High) which is not beneficial for our simulation modeling. Similar issue when using Google Popular Times view.
- Insufficient points for each station (currently 1 per month, and we only have 3 months). Hence, we can only validate and assume all the stations have the same passenger interarrival times.