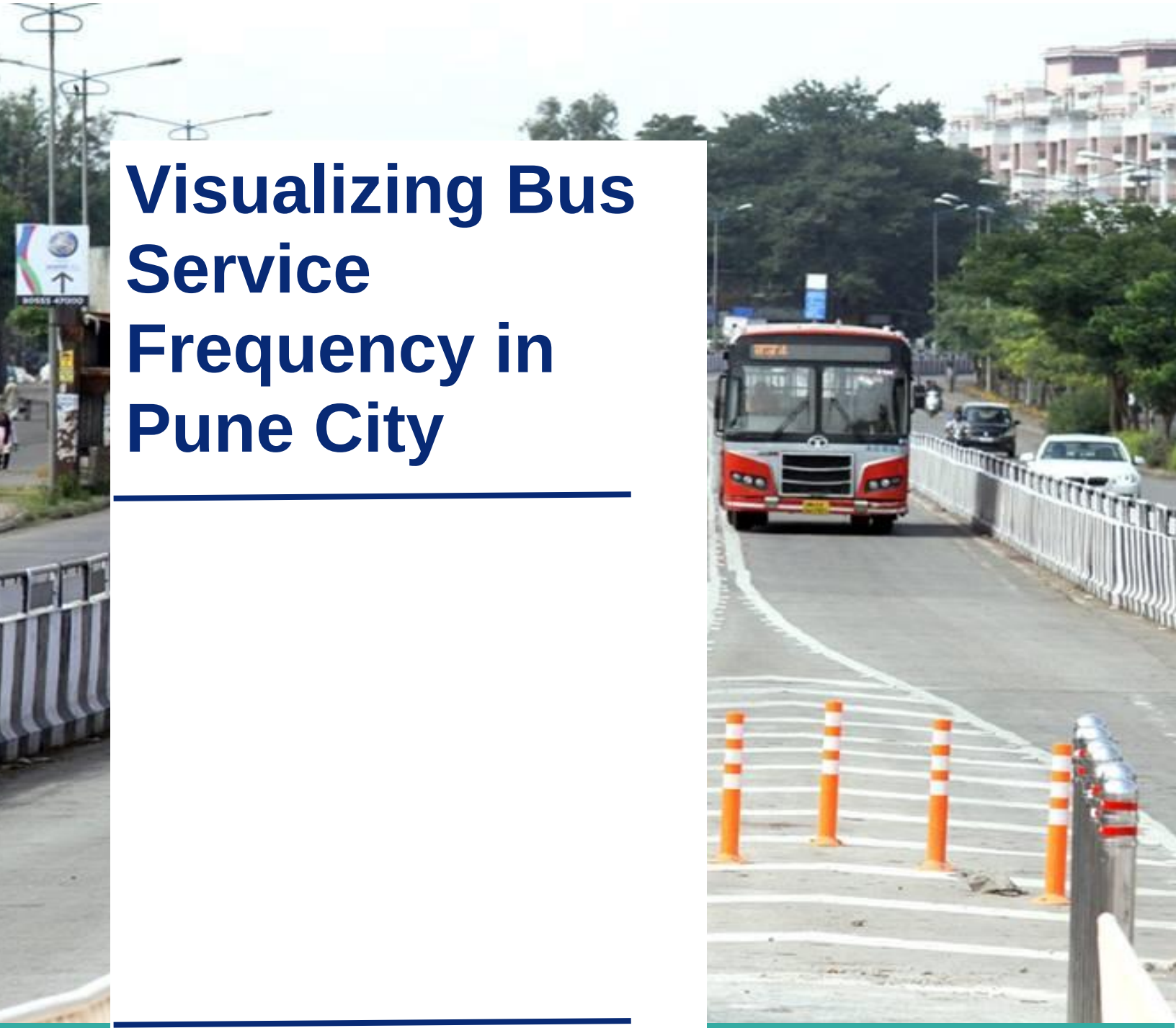

Visualizing Bus Service Frequency in Pune City

15/04/2021

Authored by:

Tushar Jadhav [2020MT93213]

Ajay Patil [2020MT93265]

Contents

Introduction

Background

Pune Mahanagar Parivahan Mahamandal Ltd (PMPML) is the public transport bus service provider in Pune and Pimpri Chinchwad Corporation limits. As per latest data PMPML has approx. 1500 buses maintained at several depots, out of which approx. 1300 buses are operational. This fleet serves around one million one hundred thousand passengers each day! The frequencies on individual routes vary greatly according to demand. Certain long-distance routes are serviced only a few times per day.
Now there are many private transport options like Ola, Uber available at fingertips. But if there is a better public transport network and frequency, then one may consider going by public transport as well.

Problem

Here we will try to visualize the bus frequency of some of the busiest bus-stops in Pune city. We've two main resources, first is map of bus routes and second is a timetable of buses on the route. Let's talk about the first one, map of bus routes. It gives us information about what route bus will follow in a geographic context, but it won't give any information about frequency of buses for a particular bus stop. Let's talk about the second one, bus timetable. It gives us information about bus frequency at a particular bus stop but lacking geographic context. How about we combine this two information and visualize them!

Apart from analyzing bus frequency we will also try to identify or cluster the bus stops according their locality. For example, we will try to find out whether the particular bus stop is located in the commercial / residential zone, developed / under-developed area etc. This will help us to analyze commute patterns and justify the number of bus stops for particular neighborhoods.

Interest

Bus transit visualization will give the frequency of buses on particular routes for a particular time period. This can help us to analyse service efficacy, transit performance, traffic management, route spacing, geographical coverage / network of service, cost efficiency and service effectiveness. This will help city planners to propose the bus stops and route network. Visualization can also help planners to explore new modes of public transport such as metro, trains as a replacement to fulfil the demand of the growing population of the city.

Lack of foresight can waste lot of people's money (news), city planning with the help of modern methods can prevent our economic waste !

Data

Data Sources

The PMPML Data

Bus Timetable and Bus Stop Co-ordinates

Pune Municipal Corporation (PMC) is providing open data sets for citizens of Pune as part of Open Governance, Open Data initiative of Government of India. Anyone can access this non-sensitive data relating to the city by just simple email registration and help PMC to develop solutions to Pune's problems.

Below are the details of dataset we are going access:

URL : PMC Open Data Store
Dataset title : PMPML Bus Routes - July 2019
Tag/ category : PMPML
File type : .zip
Size : 8.81 MB

One needs to access the above link using a personal account login and search the dataset with mentioned title and category. The dataset can be downloaded and unzipped to access text files. There are total 9 text files out of which we are going to use below 5 files for analysis.

File details:

1. stops.txt

Contains list of bus stops along with their location coordinates and stop ID.

2. shapes.txt

Contains list of unique route shapes in the form of coordinates in the sequential order.

3. calendar.txt

Contains trip service IDs that are executed for the period 2nd Aug 2019 to 31st Oct 2019.

4. trips.txt

Contains unique trip IDs and route shape ID, service ID required for trip execution.

5. stop_times.txt

Contains trip ID wise departure and arrival time at each bus stop of the trip in the sequential manner.

The Foursquare Data

Bus Stop Surrounding Venues Information

To categorize bus stops according to their type of neighbourhood we will need a database with venues data. We are going to use Foursquare Places API that will give us the latest information about our bus stop surroundings.

For Foursquare query we will need following information:

1. Bus stop location (can be taken from 'stop.txt' mentioned above)
2. Search code of main venue categories to search for.
   We can find search codes in below Foursquare documentation.
   We'll query 10 main categories (refer Table 1Table 1) with their search codes for every bus stop.
3. Search radius (we will search within 1km radius from bus stop location)

Sr. No.	Venue Category	Examples
1	Arts & Entertainment	Movies theatres, museum, sports stadium, theme park, zoo, etc.
2	College & University	University, college, school, grounds, etc.
3	Event	Street sale, festival place, main market, etc.
4	Food	Restaurant, bakery, coffee shop, cafe, etc.
5	Nightlife Spot	Bar, lounge, pub, etc.
6	Outdoors & Recreation	Botanical garden, gym, pool, track, hill, farm, forest, lake, river, etc.
7	Professional & Other Places	Business center, distribution center, factory, government building, hospitals, camps, etc.
8	Residence	Home, bungalow, apartments, residential building, etc.
9	Shop & Service	ATM, bank, shops, petrol pumps, salons, etc.
10	Travel & Transport	bus/railway/metro station, airport, hotels, tunnels, roads, etc.

Table 1 Foursquare venue categories

Sample of Foursquare data for few bus-stops (refer Table 2):

Bus_stop_id	32769	32770	32771	32772	32773
Arts & Entertainment	1	1	3	5	5
College & University	1	5	1	14	14
Event	1	0	0	0	0
Food	26	8	5	46	45
Nightlife Spot	0	3	2	5	5
Outdoors & Recreation	3	5	5	5	5
Professional & Other Places	3	4	2	21	23
Residence	4	3	6	5	5
Shop & Service	3	23	9	27	26
Travel & Transport	3	2	3	22	22

Table 2 Sample Foursquare data for some bus stop ids

Initial feature selection, Data cleaning and validation

The PMPML Data

Before processing bus data downloaded from PMC website, we checked each file and data associated with each feature. Based on sufficient data availability and its use for our analysis, we decided to keep or discard the initial features. (Refer Table 3 for more details)

Sr. No.	Resource File Name	Feature Name	Kept / Dropped	Reason
1	stops.txt	stop_id	Kept	Can be used to visualize bus stops on map
2		stop_name		
3		stop_lat		
4		stop_lon		
5		zone_id	Dropped	Insufficient data
6		stop_url		
7		location_type		
8		parent_station		
9		stop_timezone		
10		wheelchair_boarding		
11		stop_desc		
12		stop_code		
13	shapes.txt	shape_id	Kept	Can be used to calculate trip/ route distance and can be dropped after that.
14		shape_pt_lat		
15		shape_pt_lon		
16		shape_pt_sequence		
17		shape_dist_traveled	Dropped	Insufficient data
18	trips.txt	route_id	Kept	Can be used to identify trip service day and route to be used Cont..
19		service_id		
20		trip_id		
21		trip_headsign		
22		direction_id		
23		shape_id		

Sr. No.	Resource File Name	Feature Name	Kept / Dropped	Reason
24	trips.txt	duty	Dropped	This data is not required for analysis
25		duty_sequence_number		
26		run_sequence_number		
27		trip_short_name	Dropped	Insufficient data
28		block_id		
29		wheelchair_accessible		
30		bikes_allowed		
31	stop_times.txt	trip_id	Kept	Can be used to filter arriving bus/es at particular bus stop for particular time.
32		arrival_time		
33		departure_time		
34		stop_id		
35		stop_sequence		
36		stop_headsign	Dropped	Insufficient data
37		pickup_type		
38		drop_off_type		
39		shape_dist_traveled		
40		timepoint		

Table 3 Initial feature selection of PMPML data

I found a few problems with “stop_times.txt” which contains trip ID wise departure and arrival time at each bus stop of the trip in the sequential manner.

First, arrival and departure times were not in correct 24hr format.
E.g. For arrival time “23:59:40”, departure time is “24:01:21”, such values are observed when the trip begins late night and ends early in the morning. Ideally here departure time should be “+1 Day 00:01:21”. However, we can use this problem itself as filter to correct the time information as per our requirement.

The second problem with dataset is the difference between arrival and departure time is not realistic. For some bus stops difference is greater than 5 hrs. To solve this issue, we can use 75% occurrence value to replace time differences greater than 3 mins.

The third problem is also similar to second where the difference between the arrival times of two consecutive bus stops is greater than 5 hrs., so similar solution can be applied to this issue as well. We can use 30 mins as lowest time difference for replacement considering there are some long-distance trips.
Apart from the above issues, there were a few entries that need individual attention to correct arrival and departure times.

Now we can derive some features from available features, refer Table 4:

Sr. No.	Derived-Features	Derived from features	Method
1	Trip distance	Shape point latitude and longitude	Every trip or trip ID has associated route ID. Every route ID has associated shape ID. Shape ID has number of shape points. Shape point is nothing but location on map with latitude and longitude information. When we connect these shape points a route is formed on the map. Now, if we calculate distance between consecutive shape points and sum them up then we'll get distance between 1st and last shape point which is nothing but route distance or trip distance. We can now drop kept features from 'shapes.txt'.
2	Trip begin and end time, Trip Duration	bus stop arrival and departure time, bus stop sequence number	For every trip bus follows a sequence of bus stops to reach destination. We can use arrival time at first bus stop as trip begin time and arrival time at last bus stop as trip end time. Trip duration is nothing but difference between trip end and begin time.

Table 4 Derived features for PMPML data

After preparing and combining the data from all the text files, there were 21803 unique trips with 19 features in the data (refer Table 5).

Sr. No.	Resource / File Name	Feature Name
1	stops.txt	stop_id
2		stop_name
3		stop_lat
4		stop_lon
5	trips.txt	route_id
6		service_id
7		trip_id
8		trip_headsign
9		direction_id
10		shape_id
11	stop_times.txt	trip_id
12		arrival_time
13		departure_time
14		stop_id
15		stop_sequence
16	Derived features	trip_distance
17		trip_duration
18		trip_bgn_time
19		trip_end_time

Table 5 Final features

Methodology

For clustering the data on the basis of stop location/venue details

We've PMPML's bus stop data and Foursquare venues data of each bus stop surroundings. As a first step we'll normalize the data set of Foursquare data and cluster them using the KNN algorithm. We'll find out the best suitable value of number of clusters using fixed random state to make randomness deterministic. Then we'll train the model with the best value of clusters and merge the generated cluster labels with our PMPML's bus-stop dataframe. For further analysis on how clusters are formed we can plot the 75% values of each venue category for each cluster.

For Visualizing the bus frequency

There are 349 bus-stops in Pune city. We've to limit the number of bus-stops for visualization because visualizing all of them at a time may not look useful. So, we'll find out busiest bus-stops on the basis of trip count. Here trip count is nothing but the summation of trips arriving and departing from bus stop. We'll normalize the trip count data for each bus stop and apply the KNN algorithm to get clusters. We can then check how clustering is performed using bar plot, where we can select top busy bus-stops. Once we limit the of bus-stops we'll filter our dataframe with these bus stops depending on trip begin stop id or trip end stop id.
Trips are executed depending on service ID. Service ID is selected depending on day of the week. So, for each day of the week we will have a list of service IDs to run. We'll select duration as Friday 02/08/2019 00:00:00 hrs to 03/08/2019 02:00:00 hrs for our analysis. In-short, we're going to visualize the bus-frequency on Friday.

Now we've our dataframe filtered with busiest bus stops and service IDs. All we've to do is repeatedly filter dataframe for every second for mentioned duration and map bus location for every trip at given time. Also, as we map the buses, we've to save them into the image format. But that will create an image for every second and that won't be easy to handle. On an average for every trip, it takes approx. 2 mins to reach its next bus stop. So, we'll change our filter frequency from 1 Sec to 2mins, that will show considerable movement of buses on the map/image.
Once we've all the map images, we can combine them to form GIF for visualization.

Results & Discussions

Clustering the data on the basis of stop location/venue details

After clustering the bus stops based on Foursquare data, we can get the following map. (refer Figure 1)

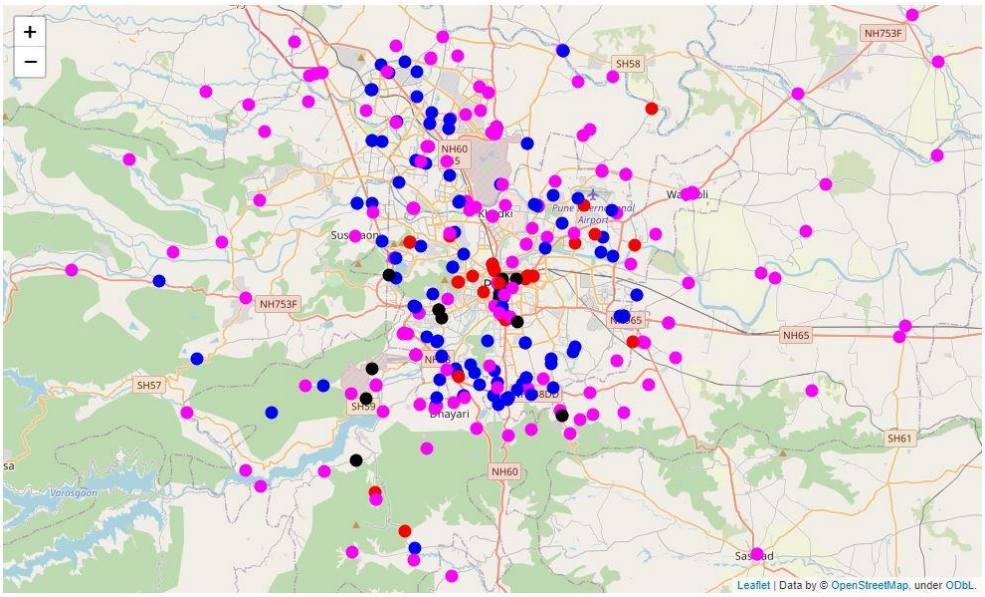

Figure 1 Map: clustering bus stops on the basis of Foursquare venue details

Let's try to understand clustering from the below line plot. (refer Figure 2)

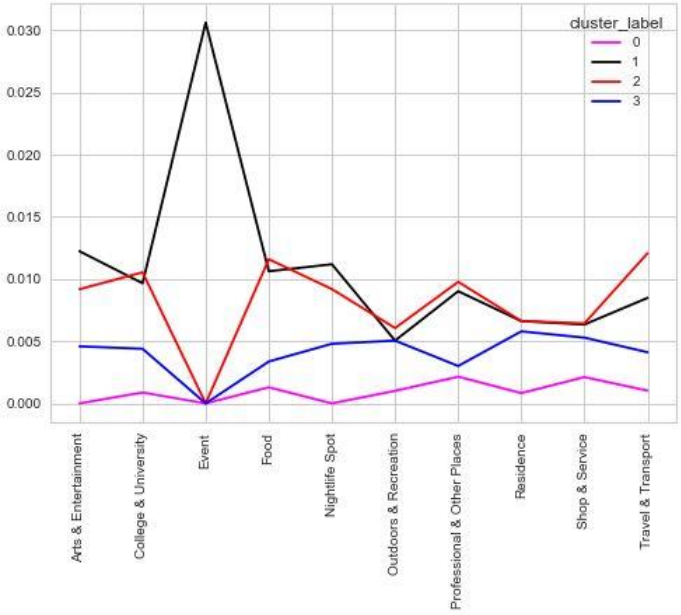

Figure 2 Line plot: Understanding the bus stop clusters

What are cluster 1 & 2?

There are two significant differences between these two clusters (Number of Events and travel/transport venues). Cluster 2 contains the highest amount of transport venues compared to any of the clusters. Cluster 1 is very similar to cluster 2 with the only significant difference in the number of Events venues.

What are cluster 0 & 3?

Cluster 0 has a smaller number of venues compared to any of the other clusters.
Cluster 3 has more venues than cluster 0 but lesser than cluster 1 or 2.

So, from the above observation and plot, we can say that:

- Cluster 2 is a developed part of the city and contains bus stops that are close to transport venues such as railway / bus / metro stations or airports etc.
- Cluster 1 is also a developed part of the city with the highest amount of Event venues.
- Cluster 0 is remote / underdeveloped part of the city and
- Cluster 3 is a developing part of the city.
- From the map, we can say that, although clusters 1,2, and 3 have lesser bus stops, they are densely located. This suggests that bus-stops have very good connectivity in the central / old part of the city.

We also plotted the bar chart of the number of bus stops for each cluster. (refer Figure 3)

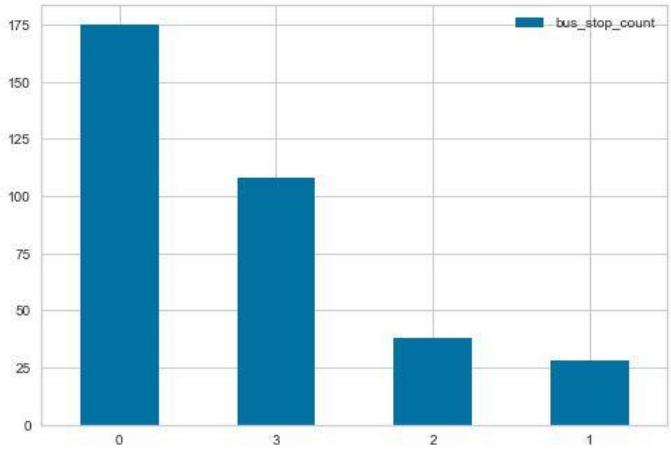

Figure 3 Bar plot: Understanding the bus stop clusters

From the bar chart, we can say that bus service has a great network in for remote parts of the city (cluster 0), then it also covers travel / transport venues with a comparatively (compared to cluster 1) higher number of bus stops (cluster 2).

Visualizing the bus frequency

We can also plot bus frequency for different times of the day. (refer Figure 4)

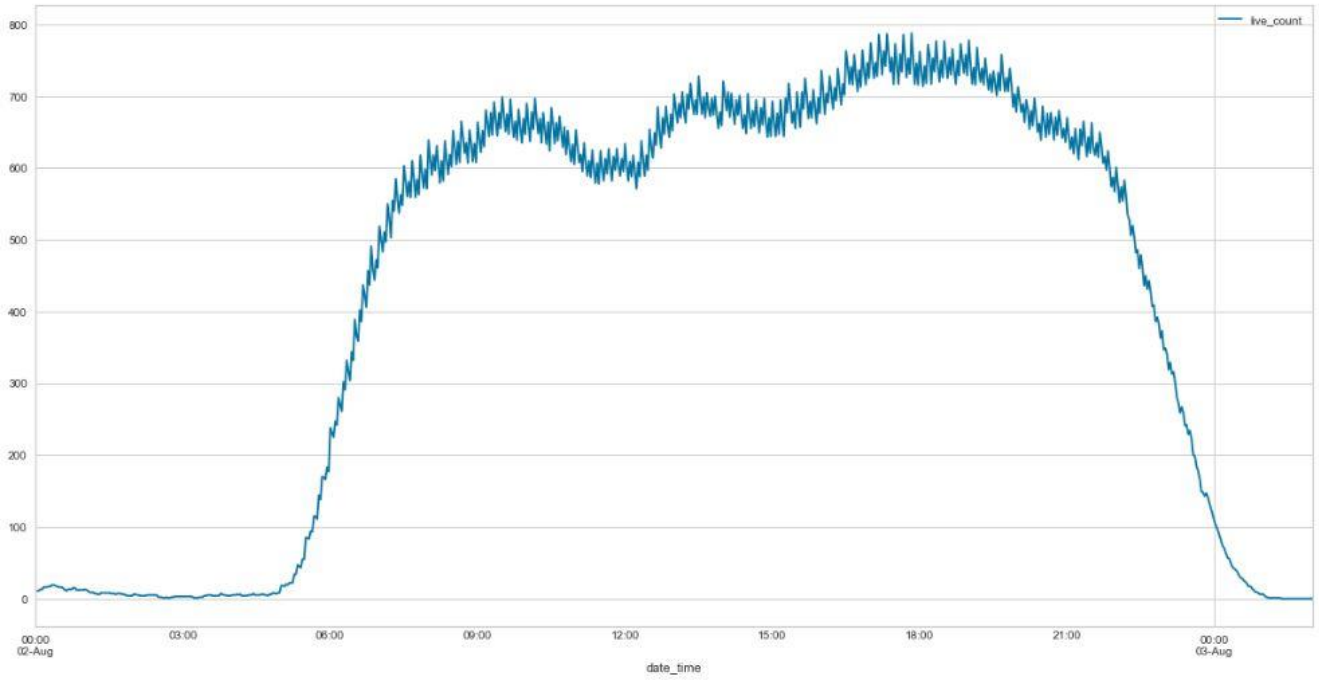

Figure 4 Bus service frequency during whole day

From the plot, we can say that bus frequency rapidly increases around 6 am and touches the morning peak of approx. 660 buses in motion at around 9:30 am. We can call it 'morning rush'.

After the noon bus frequency again increases close to 700 buses in motion and remains there up till 05:00 pm.

Between 05:00 pm to 07:30 pm bus frequency is close to 750 buses in motion, we can call it as 'evening rush'.

From 07:30 pm onwards frequency declines gradually till 10:00 pm, and after that, it takes a plunge.

Now, we've images/frames that can be used to make GIF. (refer Figure 5)
There are approx. 819 images/frames (size: 339 MB), since combining these images to form a GIF will require a lot of memory, I've used 'PhotoScape X' software (for Windows 10) to make final GIF.
Software settings are:

- Frame rate: 20 frames/sec
- Frame size: 500 x 251 px
- Loop: Infinite

Figure 5 Screen grab of bus service frequency visualization for whole day

From the GIF visualization we can conclude on the following points:

- There are approx. 7 roads that are used by bus trips to connect / pass through the central part of the city.
- Bus frequency is fairly similar on every bus stop on these 7 main roads of the city for any given period.
- We can observe comparatively high traffic in the central part of the city from 7:30 am to 10:30 pm

Conclusion

Location data is continuously growing so we have to find new ways to visualize this information and it must be clearer and more intuitive. In this project, we tried this new approach to visualize the bus frequency in Pune city. We can study individual bus stops and their pattern for more details. However, this data will be more effective if we can analyse it with community mobility data. It can give more insights into how people move during a different time and in different parts of the city. We can have a comparison with current transit options and improve or if required implement new transit modes.

References

1. PMC Open Data Store:
   http://opendata.punecorporation.org/Citizen/CitizenDatasets/Index

2. Foursquare documentation:
   https://developer.foursquare.com/docs/build-with-foursquare/categories/

3. GIF software, PhotoScape X
   http://x.photoscape.org/

4. Final GIF of Bus service frequency in Pune city
   https://github.com/2020mt93213/Pune_BusRoutes/blob/main/00_Ntbk_Resources/01_DataAnalysis/02_Final_GIF/BusesInMotion_20190802.gif

5. Github repository
   https://github.com/2020mt93213/Pune_BusRoutes