**1, Introduction**

**1,1 Intro**

The Hungarian capital, Budapest's growing, nearly 2 million population require an extensive public transport network. This paper aim to analyse the operation of Budapest public transport by exploring the openly accessible GTFS data format. The analysis will further investigate the public bus routes as a key part of the urban transportation mix. Discovering key similarities among bus routes will help identify clusters of bus services with different characteristics across the city. Better understanding of the bus network will lead to more sufficient, data-led city planning such as reduce congestions, create more efficient schedule, discover bus lane need, or improve air quality.

Budapest and most cities transportation need significantly differ based on whether it’s a weekday or weekend. Therefore, paper will differentiate between a regular weekday’s and weekend’s transportation schedule to provide comparison between the two contrasting operation mode.

**1,2 Technical requirements**

For the analysis, the following Python libraries will be used. The aim was to use standard libraries to showcase how regular Pandas and Geopandas can be sufficient to work with GTFS format. For the clustering, Sklearn machine learning library is imported. To create the data visualisations Seaborn, Contextily and Folium packages are utilised.

**Literature Review**

Public transportation is a key component of urban development and planning. In most major cities, significant public spending is allocated towards enhancing transportation networks, as they impact the socio-economic development of areas, through providing (or limiting) access to education, employment, healthcare and leisure (Deka, D., 2004). Access to various forms of transportation has long been seen as a source of economic inequality – owning a horse in the middle ages meant you could get further than simply on foot. Today, numerous studies suggest that lower income households are drawn towards the more impoverished, inner-city areas due to cheaper access to transportation than in the suburbs (3&4).

In addition, transportation networks can have substantial environmental impacts. As the global rate of automobiles per capita continues to rise, countries and particularly large cities can expect a continued rise in traffic congestions and related pollution (via emissions) unless significant action is taken (5). As such, enhancements opportunities in urban transportation planning are highly sought-after, particularly in the developed world, where countries have means for funding such projects.

There are various modes of transport that define a particular cities urban transportation structure. In most cities cars and public busses tend to be the primary form of transportation, however many cities have popular alternatives, such as the underground in London (6), or cycling network in Amsterdam (7). For Budapest in particular, the public transportation mix is fairly complex, with numerous metro lines, buses, trams, trolleybuses and overground systems collectively operating to sustain the population’s transportation needs. The following essay will focus on the 229 bus services with 710 routes across Budapest and their characteristics.

This analysis will only use GTFS feed data. GTFS is becoming more and more accessible across the world. The larges data store can be found on transitfeeds.com, with currently 677 location’s GTFS feed provided by 1327 provider. Analysing and understanding these feeds lead to better understand the transit of cities and can provide improvement opportunities

**Research Question:**

R1: What are the key components of Budapest’s public transportation mix?

R2: How can we categories Budapest public bus route system based on key variables?

**3. Methodology**

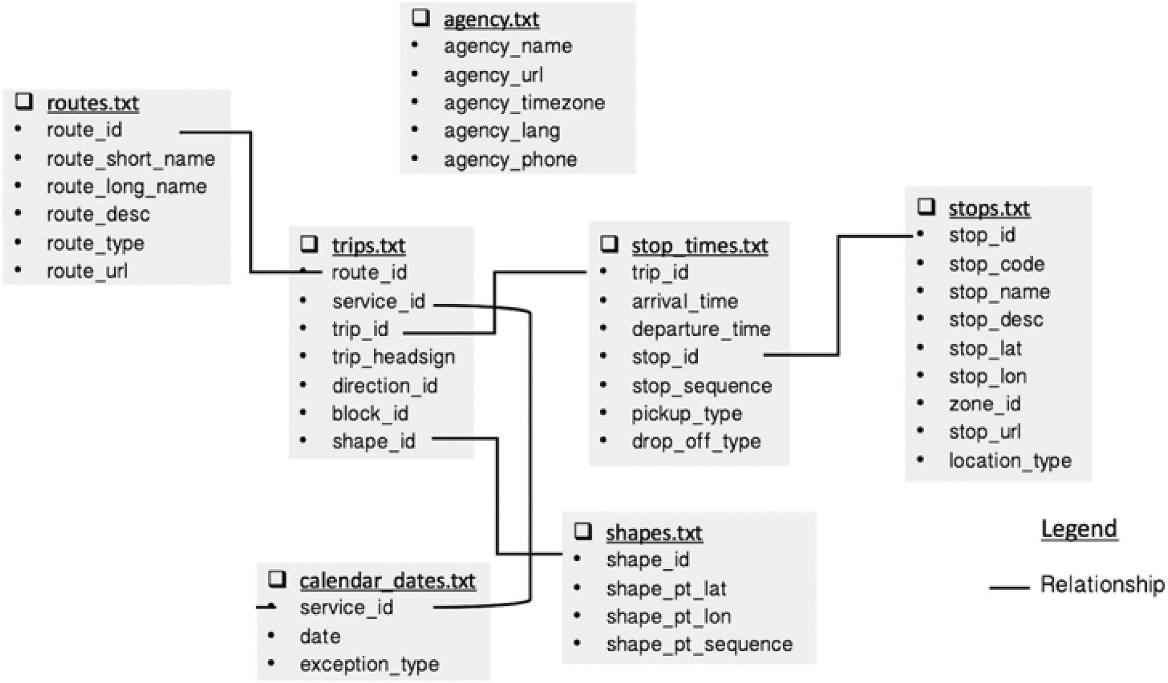
The accessed GTFS data contains an entire month (November 2020) of public transport schedule. This dataset was selected for the purpose of excluding Covid-19 related anomalies in the transportation schedule. According to the operator company the schedule is defined monthly, therefore generally the same day of the weeks has essentially the same routes and timetable planned for. For memory efficiency reasons, the busiest weekend (Saturday or Sunday) and the busiest weekday (Monday, Tuesday, Wednesday, Thursday, Friday) in the month will be selected as samples for the analysis and clustering.

This workpaper will provided step-by-step instructions from reading the raw data, cleaning and preparing it to the explanatory analysis and bus route clustering. The explanatory analysis will be aimed to showcase the information that can be gathered from the rich GTFS file format. Apart from highlight the general usefulness of the GTFS public transport data, we wish to gain more understanding of Budapest public transportation system. The workpaper also written with the aim of reusability whether using another month of Budapest GTFS data or use a different cities GTFS dataset.

This analysis will be using GTFS (General Transit Feed Specification) Static and not the real-time extensions of GTFS. GTFS contains multiple txt files which has complex relational relationships. In the case of BKK it contains the following files:

* **agency.txt** – It contains information about the service agency (BKK) of the feed, and time zone of the city where the transit operates and contact information.
* **stops.txt** –It contains the transit station/stop names, stop ID, and geo-locations (latitude, longitude).
* **trips.txt** –It contains the directions headed of each vehicle movement. Each trip has a service ID which specifies the days it operates on.
* **route.txt** –It contains the information on how trips are grouped into single services. Vehicle type of transportation, name/number of the route,
* **stop\_times.txt** –It contains the information on the arrival and departure time of each transit station/stop, as well as station sequence and corresponding trip IDs. This is the largest file in the folder
* **calendar\_dates.txt** –It contains the service id and their related dates
* **shapes.txt** –It contains geographical polylines representing the routes that a transit vehicle takes. (source GTFS-Viz paper)

The relationship among the tables is illustrated below.



(Visualizing public transit system operation with GTFS)

In order to execute the clustering of bus routes, K-means classification method will be utilised. K-means is an unsupervised machine learning algorithm, therefore by definition it seeks to uncover hidden characteristics of the data feed into it. (Allahyari et al., 2017). No training dataset required, and the algorithm can be used on numerical and categorical data attributes. The following numerical attribute will be calculated in order to create the clusters:

1. Average speed of the route/shape – defined in km/h
2. Stop Density of route/shape - stops per kilometre across the route
3. Trip activity level - number of trips have been taken on the route in a day

For defining the K – number of clusters for our classifications, we’ll be using the Elbow Method. Elbow Method is a popular test to conduct to identify the appropriate number of clusters that the data requires.

4, Data Processing

4,1 Data Gathering

4,2 Data Cleaning, Preparation

5, Analysis

5,1 Descriptive Stats

5,1,1 Vehicle types

5,1,2 Traffic of stops

5,1,3 Visualise a daily changes of the public transport traffic

TALK ABOUT EACH CLUSTER’S CHARACTERISTICS < WHY IS IT DIFFERENT WHY IT’S IMPORTANT TO DIFFERENCIATE

6. Discussion and Conclusion

References

Kadir, R.A., Shima, Y., Sulaiman, R. and Ali, F., 2018, March. Clustering of public transport operation using k-means. In *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)* (pp. 427-432). IEEE.

Deka, D., 2004. Social and environmental justice issues in urban transportation.

(3) S. LeRoy, J. Sonstelie, Paradise lost and regained: Transportation innovation, income and residential location, Jour-nal of Urban Economics 13 (1983) 67–89.

(4) Glaeser, E.L., Kahn, M.E. and Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *Journal of urban Economics*, *63*(1), pp.1-24.

(5) May, A.D., 2013. Urban transport and sustainability: The key challenges. *International journal of sustainable transportation*, *7*(3), pp.170-185.

(6) Guo, Z. and Wilson, N.H., 2011. Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. *Transportation Research Part A: Policy and Practice*, *45*(2), pp.91-104.

(7) Buehler, R. and Pucher, J., 2009. Cycling to sustainability in Amsterdam.