Data Visualization
DS 116 / CS 343

Final Project

Shanghai Transportation Analysis

Levon Gevorgyan
Albert Hakobyan

American University of Armenia
May 1, 2025

Contents

Overview

In the scope of the (DS116/CS343) Data Visualization course, Shanghai transportation analysis was conducted by students, Levon Gevorgyan and Albert Hakobyan, as a course project.

This analysis studied three main transportation categories: ***bike, taxi*** and ***bus.*** This analysis gives valuable insights into global and local transportation issues, and clear observations and solutions are provided through visualizations. Transportation can pose significant challenges to its residents and tourists. Shanghai is no exception, as tourism is very popular *(6,705.900 Person-Time in 2024)* and the population is dense *(around 24.87 million inhabitants in 2023).* Transportation directly influences people's ability to take an easy ride, access different public spaces, and ensure a better time usage. So, it is very crucial for these three components—bike, taxi, and bus—to be optimized to meet simple human needs.

This paper includes various distinct findings, and it can be helpful in many study areas, such as helping the government design better policies that will improve traffic management, reduce overcrowding, and encourage sustainable transportation. This analysis can further be used to develop some infrastructure projects that meet the demand of urban residents. The authorities who take care of public transportation can use this paper to improve route planning, optimize bus schedules, and improve traffic efficiency. Additionally, this report can help NGOs and environmental groups find more eco-friendly alternatives, such as cycling programs or replacing engines with electric buses, contributing to Shanghai's broader sustainability goals.

Literature Review

Shanghai's transportation patterns were thoroughly analyzed to understand the existing data, available sources, and the bounds of this research, using several reports that included insights about global bus ridership trends and taxi usage in big metropolitan areas. Such sources were the International Transport Forum (ITF) and the World Bank, which provide essential features' analysis, such as economic growth, population density, and infrastructure development, that influence transportation patterns.

One of the valuable sources for the analysis of this paper was provided by the ***Shanghai Municipal Transportation Commission,*** which gives us insights about the transportation network and vehicle emissions statistics. This report, followed by thorough analysis and investigations, can shed light on key problems that Shanghai faces today. To ensure better interpretability of Shanghai's official transportation report and the city's public bus-sharing system, appropriate analytical tools and visualization were used, which you can get familiar with in the section on *Analysis and Visualizations*.

Moreover, a study by the ***China Academy of Transportation Sciences*** examined various types of transportation integrity in Shanghai, offering a valu-

able understanding of how bikes, taxis, and buses interact with the city's transportation ecosystem.

**All of the mentioned papers can be found in the References.*

The Data

The data used in the project was obtained from the ***Shanghai International Open Data Platform,*** which is a public, certified repository that provides access to different datasets published by the government's transportation bodies. This specific data contains information about transportation activity, including bike, taxi, and bus statistics, and is only open for research purposes. The data allows users to explore real-time and historical transportation patterns across different regions of Shanghai.

The dataset was given in tabular format, and it was mostly clean, containing no missing or duplicate values. However, during the preprocessing some irrelevant records such as incomplete trajectories, outliers, and zero-activity segments were captured, which contained no insightful information for analytical purposes, that is why they were filtered out and stored in a new datafile named `data_sparsity.csv` which you can access *here* and see those `trash` records.

To analyze and visualize the data, two software environments—**Python** and **R** were used to get powerful information, statistical analysis, and visualizations.

Analysis and Visualizations

Bike

The analysis begins with an examination of bike traffic in Shanghai, using **Python** and libraries such as *pandas, matplotlib, seaborn, numpy, geopandas,* and *folium.* The dataset reveals spatial and temporal trends in bike usage, including traffic concentrations across districts, peak usage times, and rider behavior throughout the day. This approach provides insights into commuting patterns and the efficiency of the public bike-sharing system.

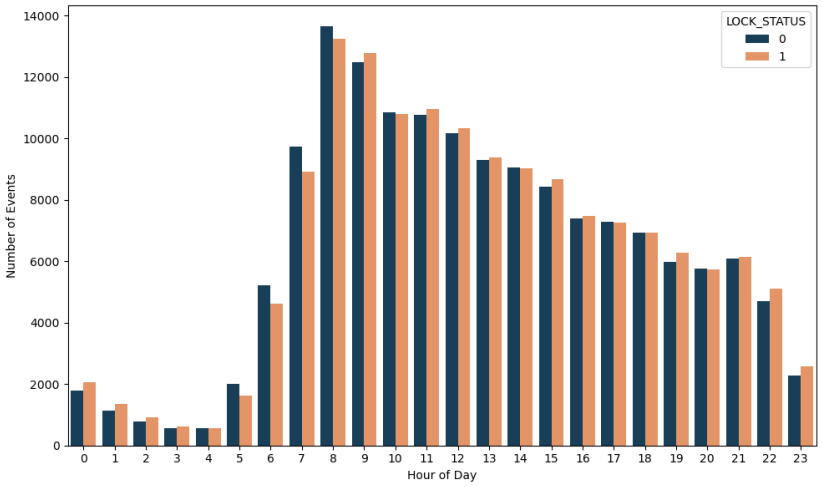


Figure 1: Bike Status Analysis in Shanghai

The temporal distribution of bike usage *(Figure 1)* gives us insights about daily patterns in Shanghai. Activity peaks at 8 AM, with both lock and unlock having their highest peak; this might suggest a busy morning rush. Usage drops between 10 AM and 3 PM, which might suggest midday errands. From 4 PM, usage again rises as people begin the evening commute, continuing till 10 PM to 11 PM, then the activity drops. Late-night use is the minimum between 12 AM and 5 AM.

The next visualization is an interactive heatmap of bike unlock locations, which was created using the *folium* library and saved into an HTML file, which you can access *here*. Each point on this map shows a GPS coordinate where a bike trip began. High concentration of unlocked bikes are shown in brighter regions (yellow to white) suggesting areas of high demand or frequent activity such as business locations, transport hubs, or highly populated neighbourhoods. In contrast, darker regions (blue to green) mean lower usage, typically corresponding to more residential or less accessible areas.

This spatial visualization gives valuable insights about bike-sharing services; when they are the most actively used, and helps to identify and inform potential decisions about infrastructure improvements or bike redistribution strategies.

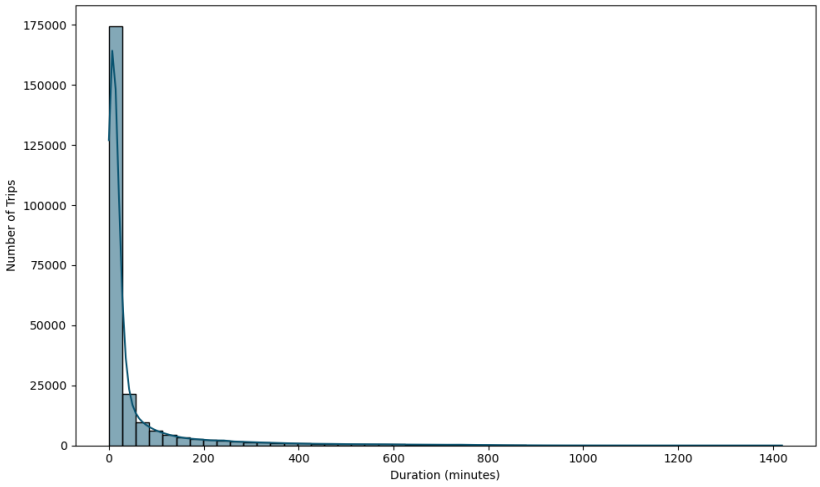


Figure 2: Estimated Bike Trip Duration Distribution

The next visualization *(Figure 2)* is a density, where Bike trip durations were calculated by the difference between consecutive unlock and lock events, with the dataset sorted by `BIKE_ID` and `DATA_TIME` to ensure accuracy. Only positive durations were retained. Most trips are short, with a peak around 5–10 minutes. Some trips are longer, up to 1400 minutes, and may indicate errors in docking. The majority are short trips. These insights may suggest some strategies for pricing campaigns, like offering lower rates for shorter trips.

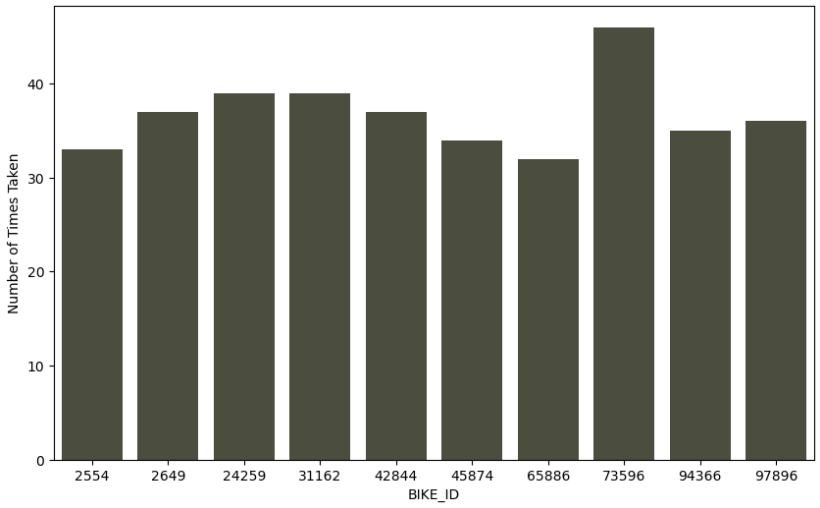


Figure 3: Top 10 Most Used Bikes in Shanghai

To identify top bikes by their usage, bike usage frequency *(Figure 3)* was analyzed by counting the unlock events (`LOCK_STATUS = 1`). The barplot graph shows the top 10 most-used bikes, with their frequency visualized on the y-axis. Bike `73596` showed high demand, suggesting that it is either located in a high-traffic area or easily accessible. One can see here that bikes with high usage are often in optimal locations or well-maintained. In contrast, those with lower usage may be underutilized, meaning a need for rebalancing to improve availability. High usage bikes like `73596` need more frequent maintenance due to wear and tear.

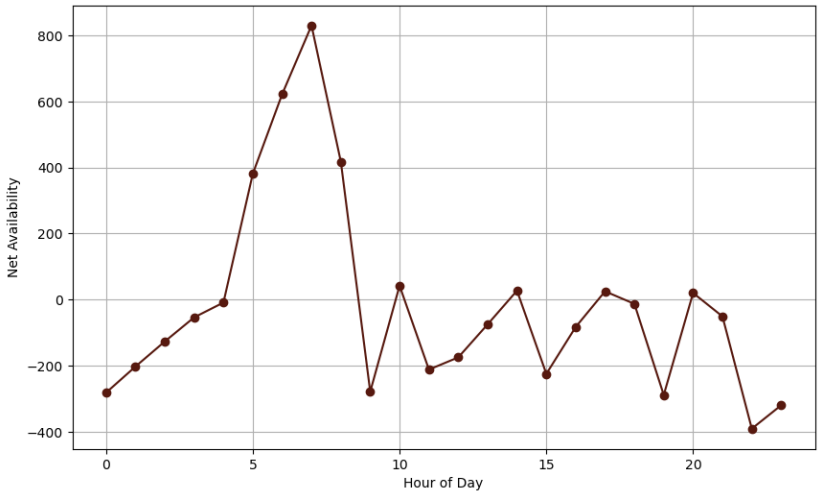


Figure 4: Net Bike Availability by Hour

Hourly net availability *(Figure 4)* of bikes was calculated by comparing unlock and lock events each hour.

- *Morning Peak (6 AM – 8 AM):* A significant increase in unlocking activity starts around 8 AM, peaking at +820, suggesting a rush for commuting.
- *Sharp Drop Around 9 AM:* A huge drop suggests that bikes are being returned after the morning rush, likely at workplaces or schools.
- *Evening Fluctuations (3 PM – 9 PM):* Net availability stays near zero during these hours, indicating a balance between pickups and returns during the evening commute.
- *Late Night (10 PM – 4 AM):* Net availability is negative or near zero, reflecting minimal bike usage.

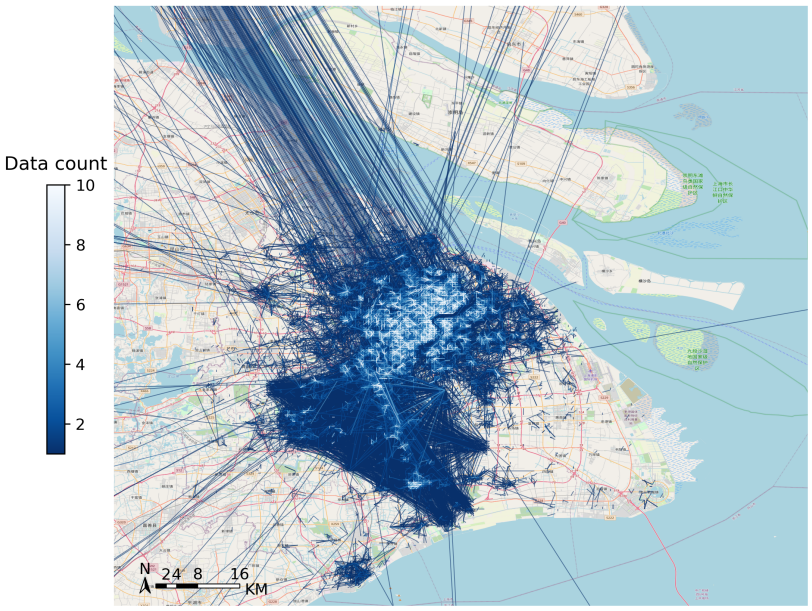


Figure 5: Origin-Destination (OD) Graph

The origin-destination (OD) graph *(Figure 5)* created using the function `tbd.bikedata_to_od`, was designed to visualize the flow of bike trips in a specific area of Shanghai. This analysis focused on the central part of the city, bounded by these coordinates: `120.85`, `30.67` and `122.24`, `31.87`, and was divided into a 500-meter grid for detailed mapping. The graph gives insights about the bike trips from their origin (unlock) to destination (lock) locations. The color represents the frequency of trips within grid points and shows higher traffic in brighter regions and lower traffic in darker areas. Results are displayed as a heatmap, where the scale of color ranges from light blue (high counts) to dark blue (low counts), providing a clear understanding of where the bike usage was concentrated. The heatmap visualizes high-demand areas and offers knowledge that can be used in making decisions about bike placement and rebalancing across the city.

Conclusion of *Bike* Analysis

The bike usage analysis in Shanghai was valuable in understanding one of the criteria of Shanghai's transportation system - the bike and its spatial, temporal, and operational patterns. It assists in designing a more efficient bike distribution and maintenance. This analysis highlights that short trips dominate the bike-sharing system and suggests opportunities for dynamic pricing and targeted consumers. Also, some problems concerning the redistribution of bikes in low-

demand areas were suggested.

Additionally, the spatial distribution gave insights into the most optimal location for bike traffic to be offered, where it has the maximum impact.

Taxi

The taxi analysis focused on the key patterns about demand, operational efficiency, and supply across Shanghai. It is obvious that taxis play a significant role in urban transportation, providing insights into peak demand times, potential gaps in a system, and commuting behaviors.

The analysis was conducted in **R** software, using libraries *dplyr, ggplot2, lubridate,* and *leaflet.* Additionally, the *sf* package was installed to do spatial data processing, and via the *ggmap* package, taxi trip data were overlaid on city maps.

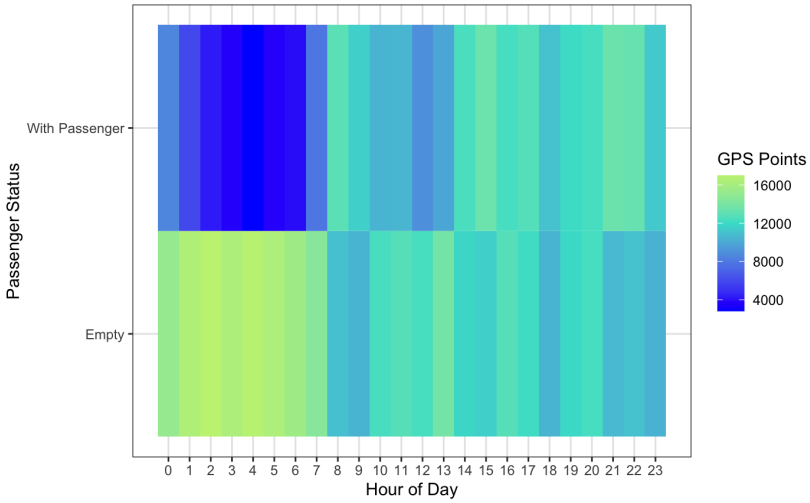


Figure 6: Day Activity Patterns of Taxi

The temporal heatmap shown in *Figure 6* was developed to visualize taxi activity patterns based on the frequency of GPS points by hour of the day and passenger status.

The x-axis shows hours (0-23), and the y-axis shows whether the taxi is empty or has a passenger. The color gradients visualize GPS points' frequency, with blue indicating low activity and green-yellow indicating high activity.

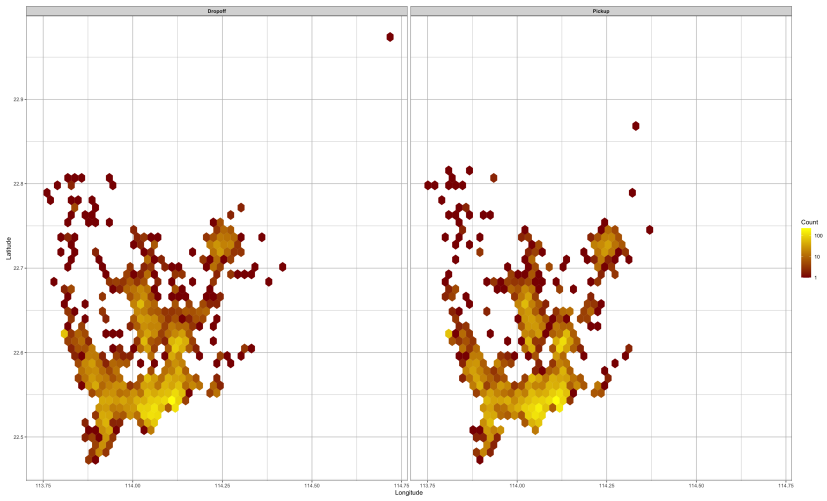


Figure 7: Spatial Distribution of Taxi Pickups and Drop-offs

Taxi pickup and drop-off transitions underwent spatial analysis *(Figure 7)* using hexbin density plots to visualize the passenger transitions' distributions. The data is filtered to identify pickup and drop-off events dependent on the status that changes in the taxi's `OpenStatus`.

The transitions were mapped as hexagonal visualizations with separate pickup and drop-off panels. The color intensity emphasizes the density of the events, with yellow color showing lower concentration and red color showing higher concentrations. This helps identify high-traffic areas for taxi pickup and drop-offs in Shanghai.

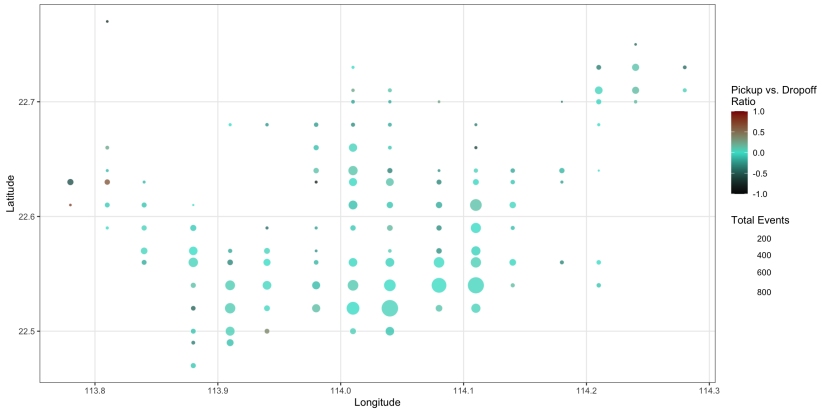


Figure 8: Comparative Analysis of Taxi Pickup and Drop-off Patterns

The analysis shown in *Figure 7* was then enhanced using data binning

into longitude and latitude intervals *(Figure 8)*. The binned data was used to visualize the ratio of pickups to drop-offs in each area. The color scale shows the ratio where blue means that the area has more drop-off events, red indicates that the area has more pickups, and the turquoise circles mean the balance between these two events.

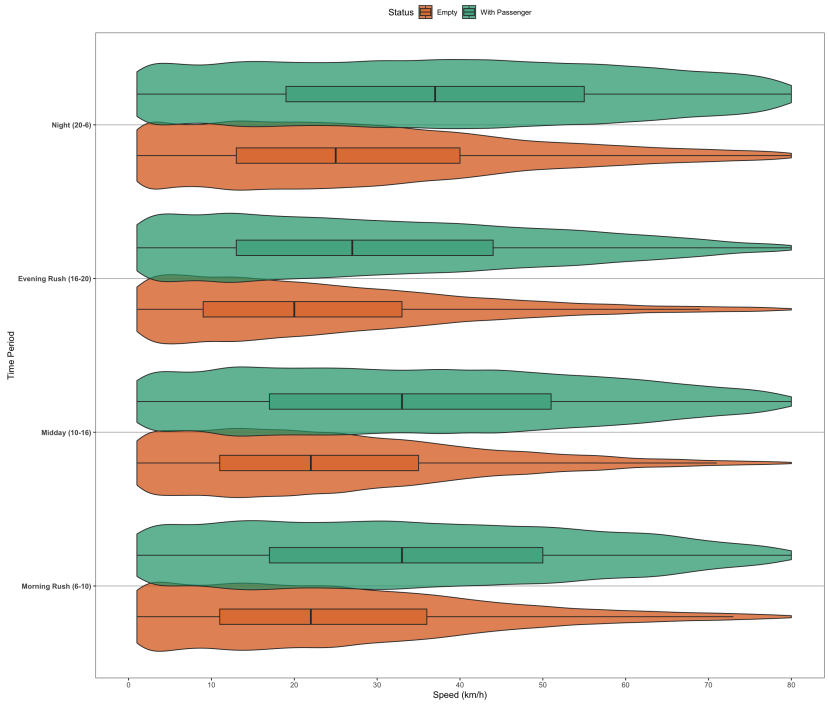
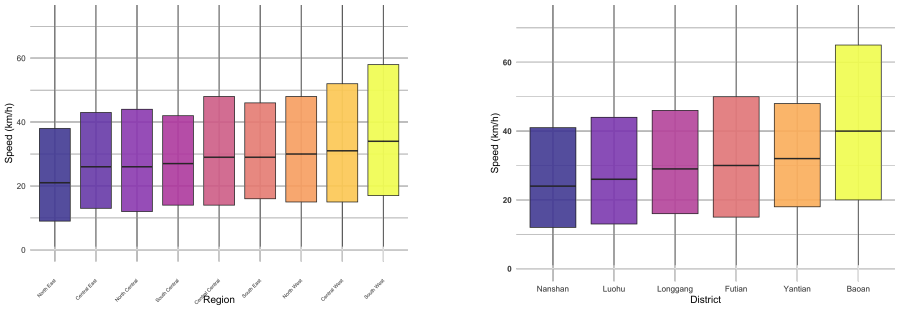


Figure 9: Taxi Speed Distributions by Time of Day

The speed of a transportation vehicle is very crucial, specifically in urban areas, as it can directly lead to accidents or to optimized time. For that purpose, taxi's speed distribution was analyzed *(Figure 9)* by time period, which revealed speed variations in taxi during the day based on the status of the passenger. The dataset was divided into four subgroups: `Morning Rush` (6-10), `Midday` (10-16), `Evening Rush` (16-20), and `Night` (20-6).

Violin Plot was used with an embedded box plot for better visualization of taxis' speed distribution with and without passengers during the aforementioned time periods.

This visualization gives insights into the importance of factors such as traffic conditions and passenger status influence in a daily hours frame.



(a) Shanghai's Urban Regions

(b) Shenzhen's Districts

Figure 10: Comparison of Taxi Speeds Across Urban Regions

Another analysis was of the taxi speed distribution in different urban regions in Shanghai. This was done by dividing the city into three latitudinal and three longitudinal areas. The plot *(Figure 10.(a))* provides insights into taxi speed distribution based on region, giving information about speed variations. To ensure an even spatial distribution of data, the quantiles of latitude and longitude metrics were used to determine regions during this analysis.

The outcome plot shows taxi speed variations, with a cap in the 95th percentile, which provides a clear view of typical speeds by excluding extreme outliers. This can help users to understand where traffic congestion or other factors can affect the performance of taxis as a separate transportation service.

The analysis covered in *Figure 10.(b)* continue the idea of taxi speed distribution and focus on speed in different districts in Shenzhen, which is the third most populous city by urban population in China after Shanghai and Beijing, and as some additional data was found about Shenzhen as well, this can assist in camparisons between Shaghai's and Shenzhen's taxi analysis. Geographic coordinates were used to perform a spatial join to assign each taxi trip to a unique area. The same idea of a boxplot as in *(Figure 10.(a))* was used here.

The graphs' insight includes the variability of taxi speed across different districts, with the quantile cap at the 95th percentile.

It can be observed which district has faster or slower taxi speeds, providing valuable information for urban planning and taxi transportation efficiency in both cities.

Conclusion of *Taxi* Analysis

This analysis of taxi activity in urban areas reveals important spatial and temporal patterns. By examining speed distributions across districts, it was observed that certain regions experience higher or lower speeds, suggesting varying levels of congestion. Temporal patterns highlight peak demand periods, which are essential for improving fleet management. The findings indicate that

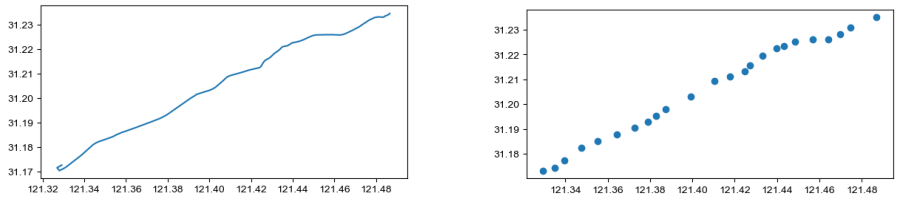
optimizing taxi operations based on district and time could enhance efficiency, reduce congestion, and improve service quality in urban transportation systems.

Bus

Python Analysis of *Bus*

Shanghai's public transportation system is one of the world's most extensive and efficient ones, and the bus itself plays a crucial role in connecting urban and suburban areas. The bus system complements the metro and covers routes and neighbourhoods that are not directly accessible by rail, offering better flexibility and frequent use of buses across the city.

To visualize and analyze the bus network, both software programs—**Python** and **R** were used. To enable handling spatial data analysis, and plotting geospatial data, libraries like *transbigdata, pandas, geopandas, matplotlib, cartopy, seaborn,* and *scipy* were used in Python.



(a) Bus-Line Plot (b) Bus-Stop Plot

Figure 11: Bus Traffic

The two visualizations focused on route 71 (`Yanan East Road`–`Bund`–`Shenkun Road Hub`), which is a major east-west route in Shanghai. The first plot *(Figure 11.(a))* visualizes the bus stops' distributions along this route, from southwest to northeast, aligning with the actual path from the `Bund` toward `Shenkun Road Hub`. The graph shows that the stops are evenly spaced, suggesting that the well-planned route is consistent with rapid transit principles.

The second plot *(Figure 11.(b))* shows the distinct bus stop locations as separate points along the route. The bus-line and bus-stop plots are a great visual of the spatial distribution of the bus traffic and give insights about the areas with higher stop density.

	stopname	arrivetime	leavetime	VehicleId
0	延安东路外滩	2019-01-17 07:16:40	2019-01-17 07:32:15	沪D-R0725
1	延安东路外滩	2019-01-17 09:51:22	2019-01-17 10:10:50	沪D-R0725
2	延安东路外滩	2019-01-17 12:54:13	2019-01-17 13:14:42	沪D-R0725
3	延安东路外滩	2019-01-17 15:37:00	2019-01-17 15:43:24	沪D-R0725
4	延安东路外滩	2019-01-17 18:34:27	2019-01-17 18:53:18	沪D-R0725
...	...	...	...	...
8516	吴宝路	2019-01-17 10:23:30	2019-01-17 10:54:31	沪D-T9651
8517	吴宝路	2019-01-17 12:57:54	2019-01-17 13:48:38	沪D-T9651
8518	吴宝路	2019-01-17 15:50:34	2019-01-17 15:55:34	沪D-T9651
8519	吴宝路	2019-01-17 16:02:53	2019-01-17 16:36:17	沪D-T9651
8520	吴宝路	2019-01-17 19:22:11	2019-01-17 22:44:50	沪D-T9651

Figure 12: Arrival Information

The descriptive analysis *(Figure 12)* of bus arrival and departure times from GPS data uses the *transbigdata* library in Python, and deriving the longitude and latitude coordinates from the `Strlatlon` column, the `busgps_arriveinfo` function was applied to match GPS records with bus stops along route 71. The output dataset provides detailed information about stop names, arrival and departure times, and corresponding vehicle identifiers.

Some potential outcomes can be observed from this tabular dataset, as for instance, the bus with `ID Hu D-R0725` had multiple stops during the day, one of which was at `Yanan East Road (Bund)` by a few up to 20 minutes, suggesting potential passenger boarding volumes and peak hour congestion.

This gives deeper insights about operational patterns of the bus traffic, and by analyzing the dwell times (the difference between arrival and departure), it will be possible to identify stops with potential delays, evaluate routes' efficiency, and modify the schedule to satisfy consumer needs.

The duplicated records suggest looped service across the day, which enables the study of temporal trends, such as increased congestion during the morning and evening rush hours.

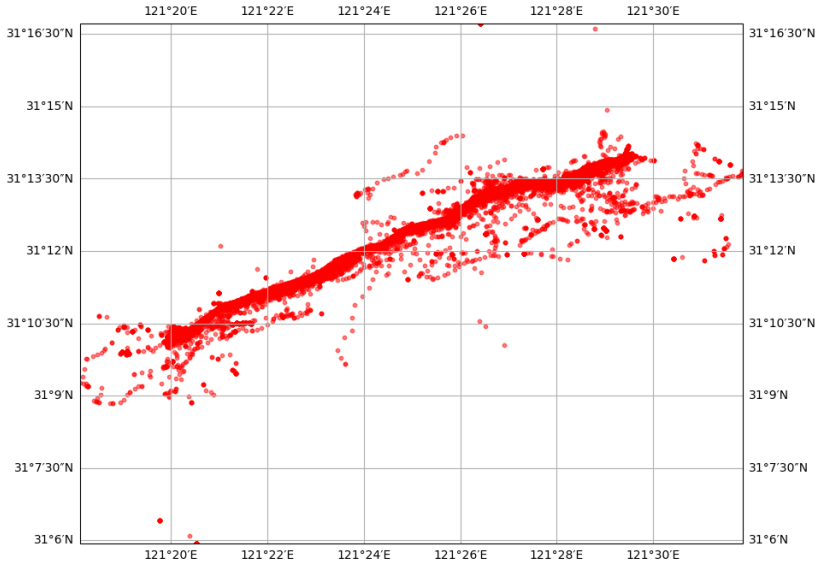


Figure 13: Bus Locations

A scatter plot of bus GPS locations *(Figure 13)* was generated using *matplotlib* and *cartopy*, with the latter providing geographic context through its map projection and optional features like coastlines and gridlines. The plot visualizes bus positions along a clearly defined linear corridor, suggesting a fixed-route service operating predominantly along an east-west axis. A higher concentration of points in the central segment indicates either increased bus frequency, congestion, or a central hub. The visualization is useful for understanding route structure, identifying operational hotspots, and can serve as a foundation for further spatiotemporal analysis of public transit behavior.

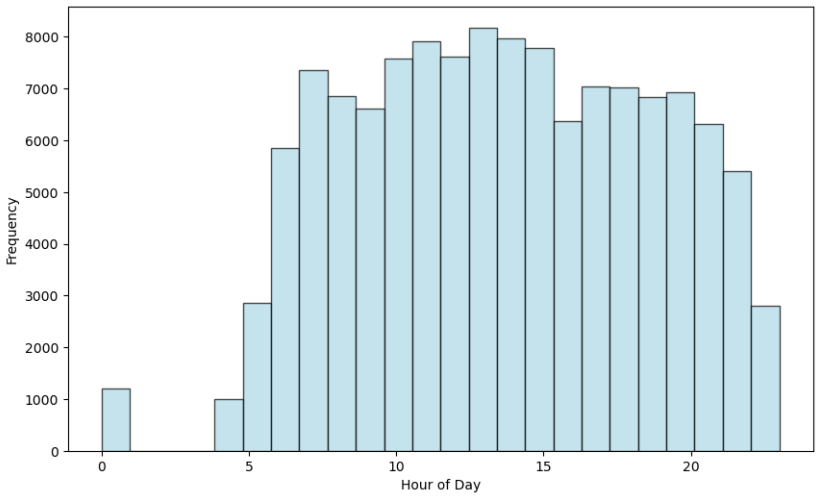


Figure 14: Distribution of Bus Arrivals and Departures by Hour

Bus arrivals and departures during different hours of a day are illustrated in the histogram shown in *Figure 14*. It is clearly observed that a peak point appears to be between 7 AM and 3 PM, with the highest frequency around 1 PM. This clearly suggests that intensive usage of buses occurs during standard working hours and school time. In comparison, the activity is minimal between midnight and 4 AM. This pattern looks very similar to the Hourly Net Availability shown in *Figure 4*, where bike traffic was observed. This can mean that these two components have very similar patterns, suggesting that the bike is as congested as a bus.

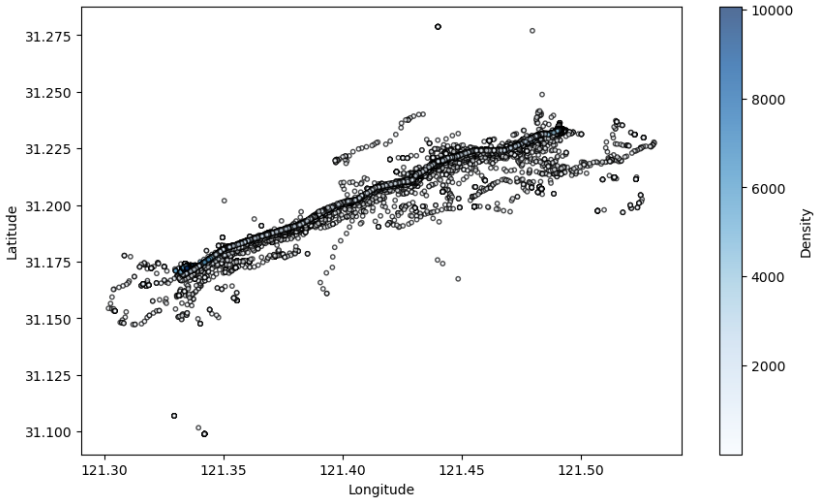


Figure 15: Density of Bus Stops

The spatial location of bus stops was observed using a 2D kernel density estimate *(Figure 15)*. The distribution shows that the bus stops are highly concentrated in the linear corridor, as in the example of Bus-Locations' graph shown in *Figure 13*, meaning a major road or transit line that is a dominant transportation axis. Darker blue shades indicate areas of higher density and the locations that have the most frequent stop hubs. This can help assess the infrastructure planning of the bus service and detect its gaps.

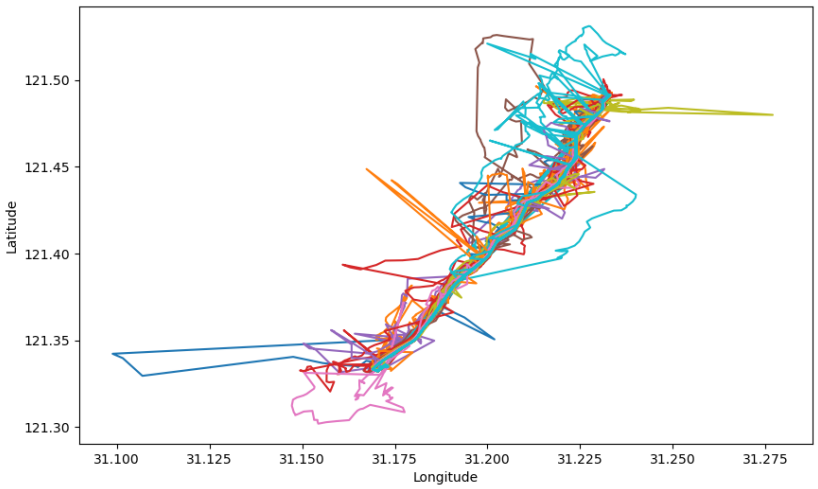


Figure 16: Bus Routes and Stops

The GPS stop locations based visualization *(Figure 16)* shows the paths of individual bus routes, where each line represents a different vehicle's trajectory through the city. Overlapping and crossing routes suggest a public transit network that is interconnected and dense, mainly concentrated on the central corridor.

This map is valuable in extending the bus network and understanding its structure. It helps to analyze some key values such as high-traffic zones and areas with potentially redundant and insufficient coverage, as well as to identify transit corridors, assess route planning, improve connectivity, and improve service efficiency.

R Analysis of *Bus*

Then moving on to the analysis in **R,** libraries such as *ggplot2, dplyr, sf, raster, ggpubr, maps, rnaturalearth, classInt,* and *ggmap* were used for geospatial data processing and visualization. These packages enabled the creation of detailed route plots and facilitated the integration of spatial datasets for further analysis.

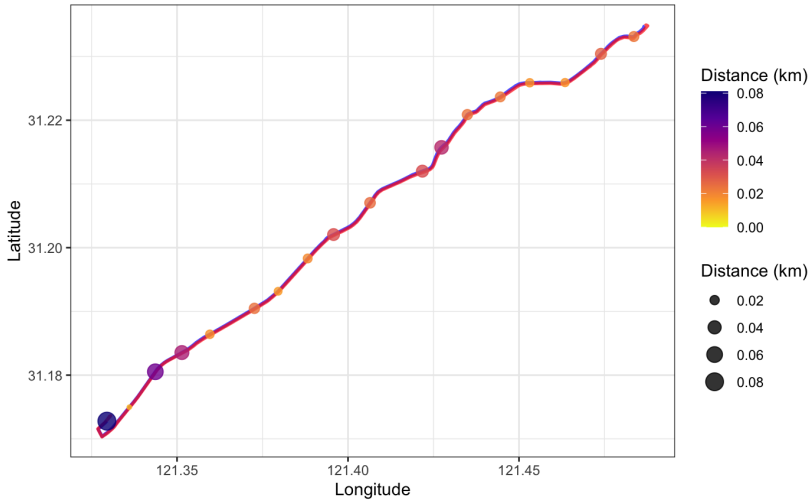


Figure 17: Bus Route Segment Divergence

The divergence of `Bus Route 71` from the `Yanan Road` path was referenced using geographic coordinates, visualized in *Figure 17.* Segments point represents segments of the route, the size and color indicate the divergence magnitude in kilometers. The points that are larger and darker show greater distances from the reference path, which are mostly distributed on the route's southwestern part. The points become smaller and lighter as the route continues northeast, indicating closer alignment with the reference. This variation in route adherence may be affected by infrastructural differences or route planning decisions.

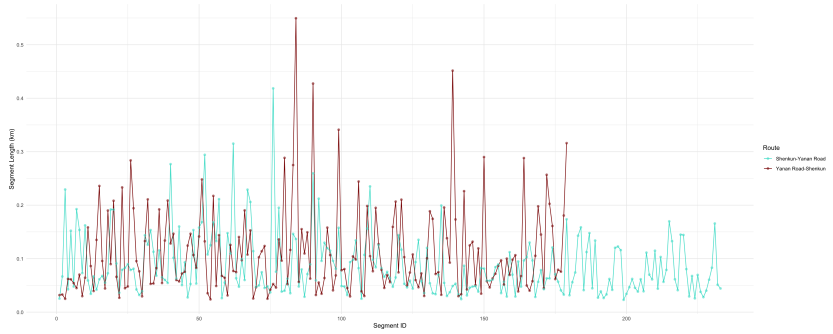


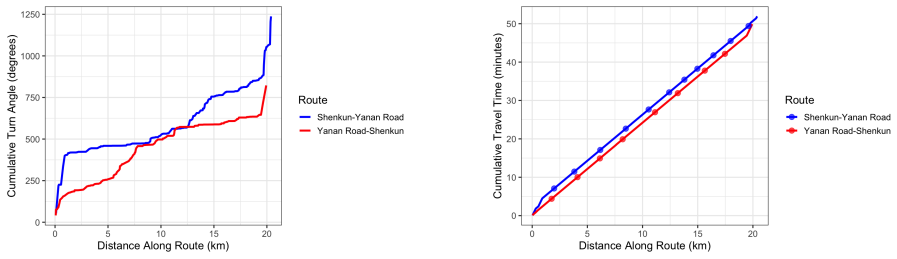
Figure 18: Segment Length Analysis

`Shenkun`–`Yanan Road` and `Yanan Road`–`Shenkun` routes' segment length analysis is depicted in *Figure 18*, where on the x-axis are the points, each corresponding to a unique segment ID along the route, and the y-axis demonstrates the segment length in kilometers. As analyzed, the segment lengths vary significantly and asymmetrically between the two directions.

The brown line is the `Yanan Road`–`Shenkun` route, which displays a greater variability, including sharper and more frequent peaks where some of them exceed 0.5 km. In comparison `Shenkun`–`Yanan Road` (cyan line) has shorter and more consistent segments, rarely reaching above 0.3 km.

To identify route-specific segmentation differences, this graph can be used to serve that purpose, as well as GPS route optimization and traffic modelling.

Also, some issues might be analyzed concerning how segments are identified for each direction, and improvements in routing algorithms might be suggested.



(a) Cumulative Turn Analysis

(b) Estimated Travel Time

Figure 19: Bus Ride Analysis

Cumulative churn angle *(Figure 19.(a))* and estimated travel time *(Figure 19.(b))* along `Bus Route 71` were analyzed, and a strong relationship between operational performance and route complexity was found.

The `Shenkun`–`Yanan Road` direction consistently showed a higher cumulative turn angle, exceeding 1200°, compared to the opposite direction, under 850°.

This suggests that there are sections with sharper and more frequent turns, which likely cause congestion and more delays. The direction also shows a longer estimated travel time, more than 52 minutes, compared to the `Yanan–Shenkun` direction, which has a lower duration of 51 minutes.

Although the difference is not huge in absolute terms, the time gap is still consistent through the route. This means that not only is physical complexity added, but overall speed is also impacted by the increase of turning, supporting the idea that route geometry can influence service time significantly, mostly by rapid directional changes. These insights are valuable for urban mobility planning, transit reliability improvement, and route optimization.

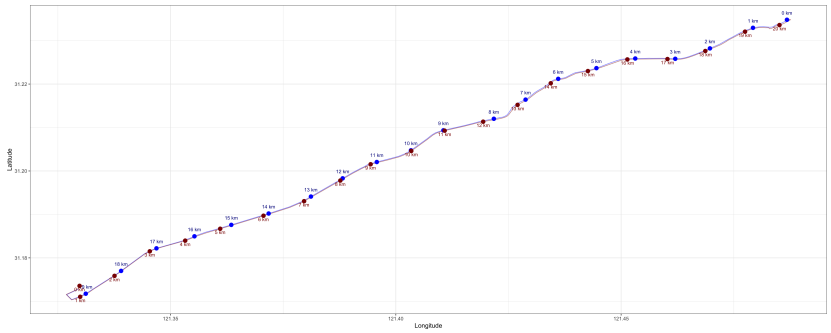


Figure 20: Key Points Along Bus Route 71

`Bus Route 71` was analyzed using key points spaced at 1 km intervals along both directions for a consistent spatial comparison *(Figure 20)*. The cumulative distance across longitude and latitude was plotted, offering a clear geospatial view of how the route evolves.

This visualization gives better transport planning and route optimization strategies by helping to identify deviations, overlaps, and directional symmetry.

Additionally, it shows distance metrics analysis, helping in tasks like travel time modelling or evaluating stop placement. The `sample_interval_km` allows users, especially within a Shiny application—which you can find in *Interactive Visualizations* or click *here*)—adjust and explore the route at various levels, which can help with strategic planning and operational management.

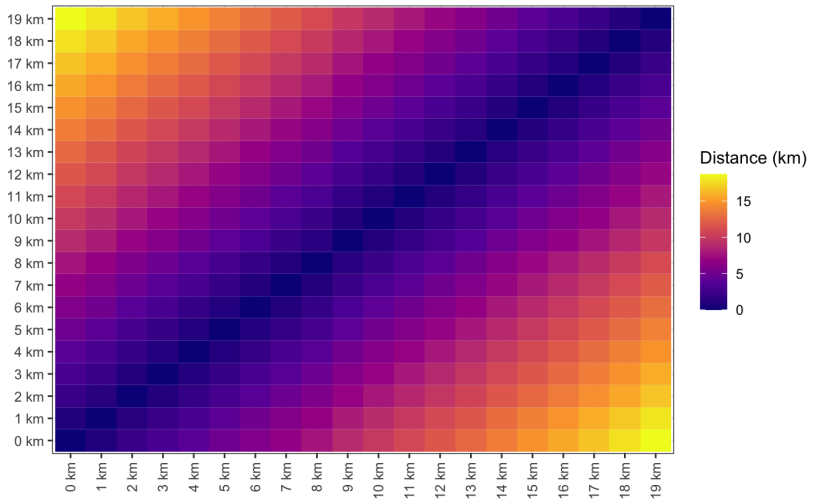


Figure 21: Distance Matrix Between Key Points

This distance matrix *(Figure 21)* was created by calculating the pairwise distances in kilometers between key points sampled every 1 km along `Bus Route 71`. Each axis represents these sequential key points, and the fill color reflects the spatial separation between them, with a plasma color scale emphasizing variation.

From the matrix, patterns of linearity and irregularity of the route can be observed, as diagonal symmetry suggests spacing between the points to be consistent. In contrast, visible abrupt shifts or bands in color intensity may emphasize loops, detours, or segments of the route that converge or diverge. Areas that have lower distances across multiple adjacent cells suggest overlapping or tightly packed path segments, while high values indicate stretches with higher geographic spread.

This matrix helps in understanding the spatial structure of the route, identifying redundant or circuitous segments, assessing route efficiency, and informing optimization strategies.

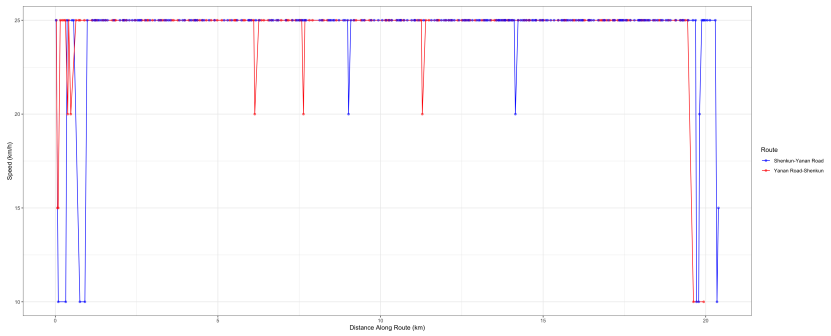


Figure 22: Estimated Speeds Along Bus Route 71

Speeds along two directions of `Bus Route 71` are plotted against cumulative distance, *(Figure 22)* based on interpolated travel time segments. The route from `Shenkun` to `Yanan` is shown in blue, and the reverse in red. Points reflect local segment speeds, with lines connecting them for clarity.

Most segments indicate speeds close to the estimated maximum speed (≈25 km/h), suggesting smooth flow for most routes. However, some segments drop significantly below this, in some cases from 10-15 km/h; these are mostly the ones that are near the route's ends. These may mean some signal delays, high congestion areas, or intersections.

This graph suggests that both directions experience similar delay patterns, but not always at the same segments, which means there are directional asymmetries in traffic behavior or signal phasing.

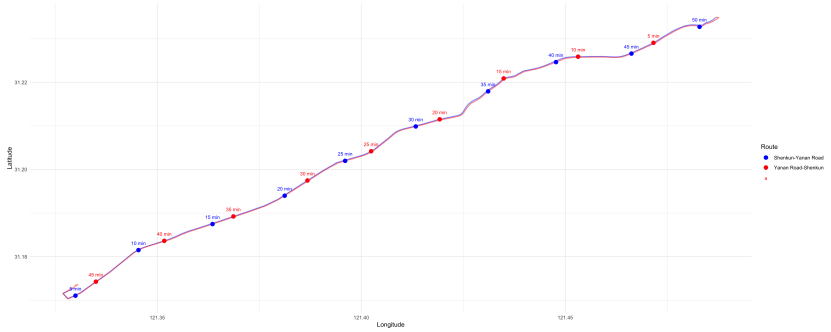


Figure 23: Bus Route 71 with Time Markers

Estimated travel times *(Figure 23)* are mapped along both directions of Route 71 using geographic coordinates. Time markers, labeled in 5-minute intervals, are calculated based on segment length and modeled speeds. Blue and red points represent the `Shenkun`–`Yanan` and `Yanan`–`Shenkun` directions respectively.

Both directions have very similar paths, which confirms the route symmetry; however, the positions of time makers suggest that travel times are not uniform across segments. Several segments do not contribute to overall time proportionally, particularly in `Yanan`–`Shenkun` direction, highlighting directional inefficiencies.

This spatial-temporal visualization assists in pinpointing where the delays accumulate and can help optimize infrastructures.

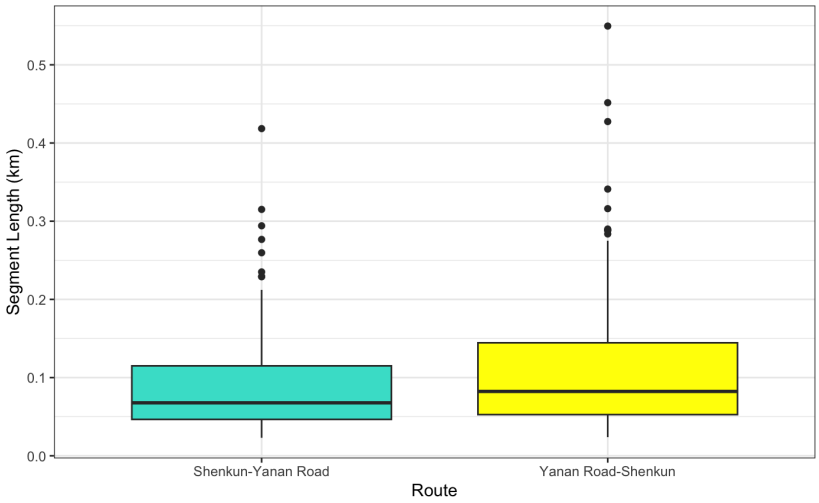


Figure 24: Distribution of Segment Lengths by Route

The boxplot *(Figure 24)* shows the distribution of segment lengths for the two directions of `Route 71`:

- `Shenkun`–`Yanan Road` (turquoise) has a slightly more compact distribution with a lower median and shorter interquartile range (IQR), indicating more uniform segment lengths.
- `Yanan`–`Shenkun Road` (yellow) shows higher variability with a larger median and is more spread in the upper quartile. More outliers in this direction suggest that the segmentation here is inconsistent. This asymmetry in segment lengths can mean differences in how the route is operationally and physically segmented, for example, signal timing, differing stop spacing, or geometry, and can influence vehicle scheduling strategies or travel time estimation.

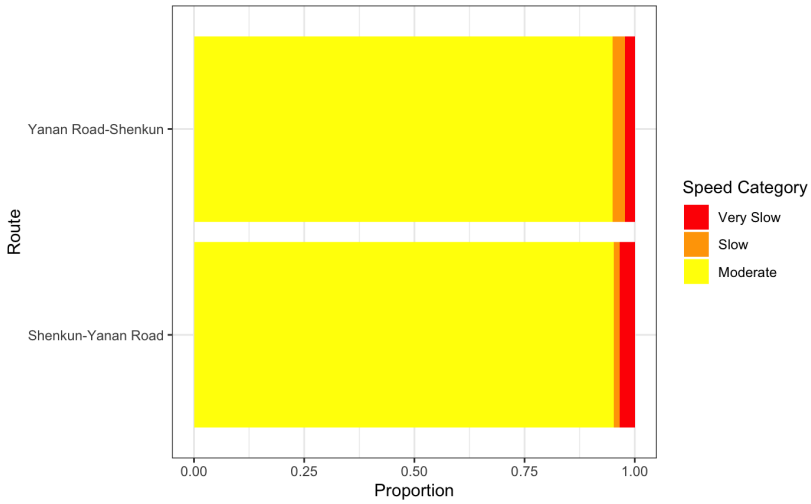


Figure 25: Proportion of Speed Categories by Route

The horizontal stacked bar chart *(Figure 25)* shows the speed categories' proportional distribution along two directions of `Route 71`:

- *Moderate speeds* (yellow) significantly dominate both directions, meaning traffic flow is relatively consistent.
- *Slow* (orange) and *Very Slow* (red) segments compose a small proportion. However, they appear more in the `Yanan`–`Shenkun` direction. This means that this direction has occasional congestion or delays.
- *Normal* (green) segment is not visible on the graph because no segment has achieved the highest speed.

Despite a larger variation in segment lengths, overall speed remains the same for most of the route. However, a slight increase in *Very Slow* in the reverse direction can suggest possible operational issues like passenger boarding delays, signal timing, or congestion patterns.

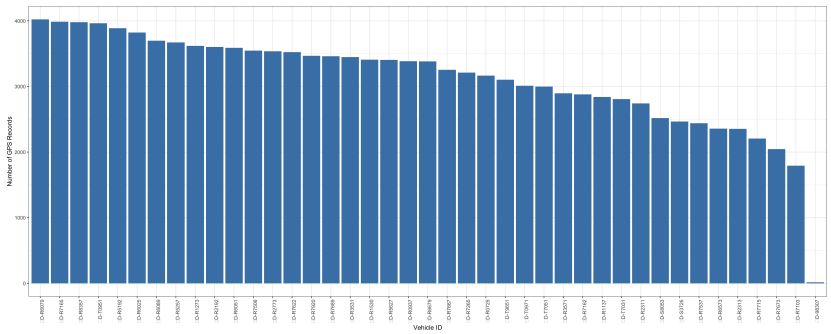


Figure 26: Vehicle Activity Distribution

The bar chart *(Figure 26)* shows the distribution of GPS records across vehicles, emphasizing how active each was during the data collection period. Most vehicles show a high volume of GPS points, centered around 3,000-4,000 records, meaning frequent activity and consistent tracking.

Moving toward the right side of the chart, with a few vehicles showing significantly lower counts, the number of records per vehicle declines. One of the vehicles has almost no recorded data, which might mean a multifunctioning GPS device that was inactive during the study period or a newly introduced one.

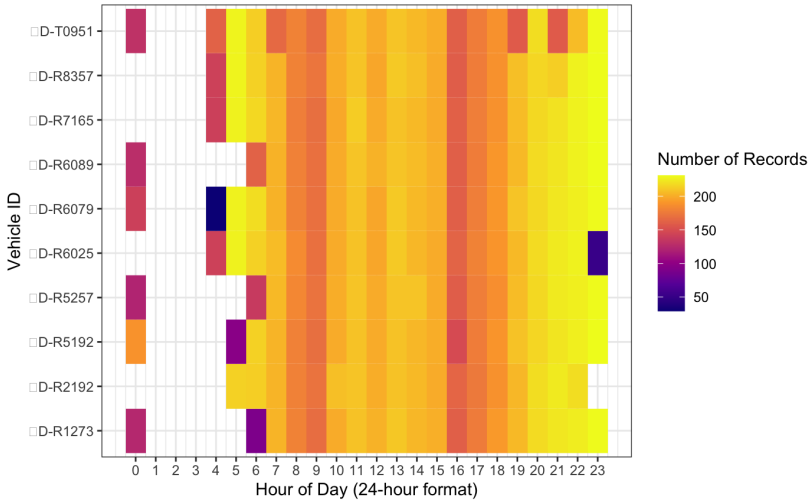


Figure 27: Vehicle Activity by Hour of Day

The next graph is a heatmap *(Figure 27)* that shows the ten most active vehicles' hourly activity based on their GPS records. As observed, the activity peaks between 6 and 10 AM, as these are the typical working hours for people

and public transport.

The early morning and late afternoon periods appear to be the busiest periods as these are the rush hours, with consistent activity during the day. Some of the vehicles have very few records in the late-night-very early morning hours from midnight to 5 AM, suggesting that it is a downtime or off-duty period, as most people are asleep. However, a few vehicles show activity spikes late at night.

This is a helpful graph for understanding operational patterns, identifying anomalies like active vehicles during unexpected hours, or optimizing fleet schedules.

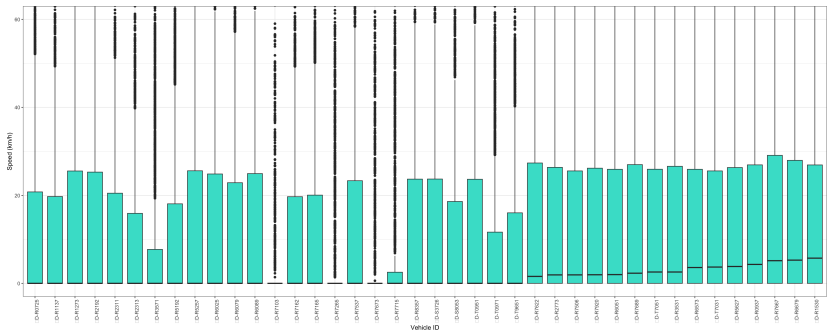


Figure 28: Vehicle Activity by Hour of Day

The boxplot in *Figure 28* shows the speed distribution of various vehicles, allowing for comparison of how fast each is. Each box represents the interquartile range (IQR) of speeds, whereas the lines below and above show the variability outside the lower and upper quartiles.

Some vehicles have extremely low median speeds, which can mean short routes, data anomalies, or frequent stops. Such vehicles are `D-R2313` and `D-R71745`. Other vehicles maintain higher median speeds with consistent distributions, such as `D-R1137`, `D-R2192`, and `D-R6025`.

Outliers above 60 km/h are excluded from the graph, but are still represented in the raw data, suggesting occasional high-speed segments or possible GPS jitter. This can help identify operational inconsistencies, pinpoint potential mechanical or routing issues, or verify expected behaviors, such as slower speed for inner-city vehicles.

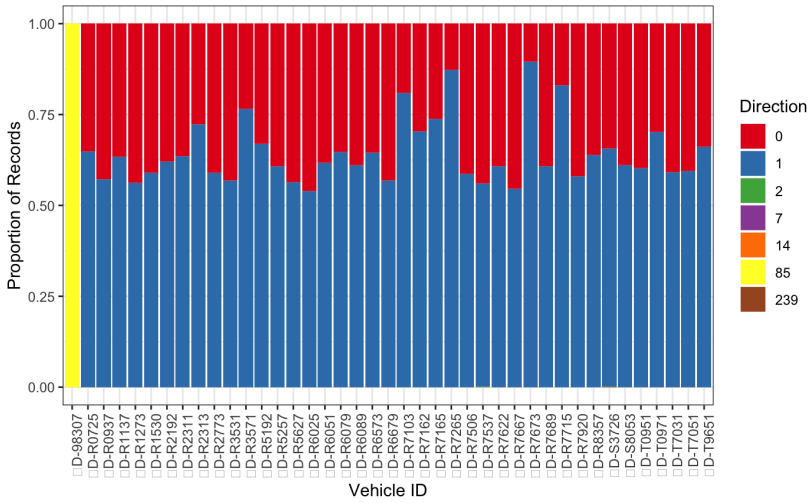


Figure 29: Direction Distribution by Vehicle

The stacked bar *(Figure 29)* chart shows how frequently each vehicle serves various directions, using the `ToDir` field. For most vehicles, the data is heavily concentrated in just two directions—0 (red) and 1 (blue)—which likely represent outbound and inbound routes, respectively. Vehicles in both directions are evenly split between them, whereas others strongly skew toward one direction.

Additionally, a few vehicles belong to uncommon directions, such as 7, 14, 85, and even 239; such vehicles are `D-98307` or `D-R1530`. These can be data entry issues, temporary routes, or anomalies. Their presence suggests potential data quality issues or operation complexity, such as multi-route service.

This graph can help detect one-way usage, route errors, or confirm expected service patterns such as round trips.

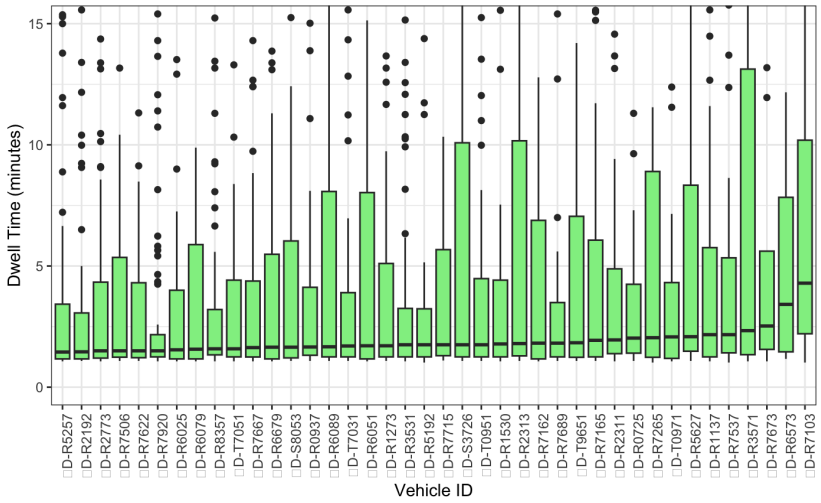


Figure 30: Vehicle Dwell Time Analysis

The next plot *(Figure 30)* reveals how long each bus remains stationary at specific locations, which helps identify patterns in dwell behavior.

Delays due to high passenger activity or traffic congestion can be suggested by longer dwell times. Frequent stops with extended durations can represent high-demand or unofficial locations, suggesting a need to add or review bus stops. This can also help to enhance overall service reliability, optimize schedules, and improve the efficiency of the route.

**You will get a free snack if you screenshot this and send it to Levon Gevorgyan via SLACK.*

Conclusions

Shanghai's transportation system was explored and analyzed in this paper using a statistical background and appropriate visualizations. Three main criteria—***bike, taxi,*** and ***bus***—were explored using softwares, **Python,** and **R**.

Bike analysis showed peak usage during morning and evening hours, minimal activity overnight, and distinct commuting behaviors. The commuter use rather than leisure was identified based on most of the trips being short. Trip duration analysis suggested system inefficiencies or outliers, whereas heatmaps and spatial analysis illustrated the use of taxis in central districts. These findings can assist in constructing pricing strategies, infrastructure planning, and better redistributions of bikes.

Taxi analysis used spatial and temporal trends, exploring speed variations and demand across different hours and regions. Taxi analysis was compared to bus and bike analysis to identify similar patterns and problems that can help to develop fleet management and service responsiveness.

Bus analysis were done particularly for `Route 71` and showed spatial visualizations about operational dynamics, stop spacing, direction symmetry, and route structure. Some aspects maintained transportation problems, such as the time, speed, and complexity of directions.

Overall, these analysis gave insights about urban mobility in a big city like Shanghai, with the spatiotemporal analysis suggesting a need to explore and analyze the transportation system.

Appendix

Interactive Visualizations

Shanghai's transportation analysis was thoroughly covered during this paper, with each criterion having a detailed explanation and further goal identification. Still, to visually look at all the graphs discussed during this paper, as well as in the appropriate **Python** and **R** code files, an interactive dashboard was designed and developed. The dashboard provides dynamic and accessible insights for public transportation authorities, government policy makers, and NGOs focused on urban traffic optimization.

The dashboard allows users to explore transportation details, and observe each of the criteria—***bike, taxi,*** and ***bus,*** filtering the time, location, and other metrics like speed, trip duration, and traffic distribution. For this perspective, you can access the dashboard *here*.

References

China Academy of Transportation Sciences. (n.d.). Transportation system integration studies in Shanghai.

Forum, I. T. (2023). Transit-oriented development and accessibility. In International Transport Forum Policy Papers. `https://doi.org/10.1787/41b95623-en`

Shanghai transportation system. (n.d.). ITF. `https://www.itf-oecd.org/search/site/shanghai`

Shanghai International Open Data Platform. (n.d.). Public datasets on transportation in Shanghai. `https://data.sh.gov.cn/`

Wikipedia contributors. (2025, April 19). Shanghai. Wikipedia. `https://en.wikipedia.org/wiki/Shanghai`

WorldBank. (n.d.). World Bank Group - International Development, Poverty and Sustainability. `https://www.worldbank.org/ext/en/home`

GitHub repository of this paper's analysis: `https://github.com/levongevorgian/Data-Visualization/tree/main/Project`