Navigating Zurich: A Comprehensive Analysis of Urban Traffic Dynamics

Project Proposal for the Semester Work in the Course "Introduction to Data Science"

Luca Zhao
Xiaohan Zhu
Weijia Zhong
Pascal Sager

Departement of Astrophysics
Faculty of Science
University of Zurich

March 28, 2024

A project proposal submitted in the course Introduction to Data Science ESC 403 at the University of Zurich.

Proposal

1.1 | Introduction

The Canton of Zurich frequently releases public transportation data, often in the form of CSV files (Stadt Zürich Open Data, 2024b). However, the raw nature of these datasets poses challenges for non-technical persons in comprehending and utilizing the information effectively. Fortunately, modern data analytic techniques (Mathar et al., 2020) offer solutions to aggregate, interpret, and visualize these datasets, enabling a comprehensive understanding of various aspects of public transportation. Moreover, leveraging machine learning methods allows predictive analysis (Kelleher et al., 2020), providing insights into future transportation patterns based on historical data statistics.

In the upcoming semester project, we aim to harness these advanced tools to conduct a thorough analysis of public transportation in Zurich. Specifically, our focus will be on examining transportation frequencies over time and assessing the trends of transportation in recent years. By doing so, we seek to uncover insights into past, current, and future transportation dynamics in the city of Zurich.

1.2 | Research Questions

Our analysis is structured around the following fundamental research questions:

1. The development of Zurich's public transportation system over time, assessed through metrics such as spatial coverage and yearly passenger volume.

2. The utilization intensity of Zurich's public transportation infrastructure, quantified in passengers per kilometer and compared to other major cities in the world (e.g., based on data sourced from OECD.org (2023)).

3. The spatiotemporal distribution of passengers as a proportion of maximum vehicle capacity per transit route, visualized geospatially within Zurich's city map.

4. Analysis of the interplay between diverse factors, including spatial location, weekday versus weekend patterns, and academic calendar.

5. Prediction of seat availability on public transit (through predictive analysis); i.e., evaluating the feasibility of forecasting the number of available seats on buses departing from a specific stop at a given time of the day, leveraging historical data.

We will answer these questions using the data described in Section 1.4 and the analytic techniques outlined in Section 1.5.

1.3 | Limitations

Examining past and present data yields multifaceted insights into the present state of Zurich's public transportation system. Nonetheless, we point out the inherent dynamics of this domain, subject to diverse influences such as evolving traffic schedules, political decisions, societal behaviors, demographic shifts, etc. Consequently, while predictions of the future may indicate trends, they inherently carry a retrospective bias and should be interpreted cautiously. Changing schedules and political decisions are especially unpredictable, making long-term precise predictions non-realistic. However, retrospective data analysis might offer insight into the confidence of such predictions.

1.4 | Data

To address our research inquiries, we utilized the "VBZ passenger numbers dataset"[1] sourced from Stadt Zürich Open Data (2024b). This dataset encompasses computed yearly averages for passengers embarking and disembarking at stops and for occupancy within vehicles operated by the Zurich Transport Authority (VBZ). The frequency measurement is conducted within vehicles employing a counting apparatus, with approximately 20% of VBZ vehicles being equipped accordingly, while the missing data are interpolated. The data were captured between 2014 and 2023 and contain 1,048,575 entries. We provide an overview of the measured features in Appendix A.

1.5 | Data Analysis Methods

Exploratory Data Analysis *(research questions 1, 2, 3, 4)*: Descriptive statistics are utilized to understand the characteristics of the dataset. We will compute summary statistics such as mean, median, standard deviation, and percentiles for attributes such as passengers boarding and alighting, occupancy, and trip distances. These computations offer insights into the data's patterns, variability, and distribution. Accompanying visualizations such as boxplots are utilized to provide an overview of these statistics. Additionally, outliers and leverage points are identified through statistical metrics and analyzed for their origin.

Temporal Analysis *(research questions 1, 5)*: Temporal analysis will allow us to uncover trends and patterns in public transportation over time. We will examine yearly trends from 2014 to 2023, aiming to identify long-term trends in public transportation, potentially including the impact of the pandemic. Additionally, we will analyze passenger numbers and occupancy rates for weekdays versus weekends to identify any notable differences in demand patterns. Furthermore, we will examine

[1] The dataset is publicly accessible at `https://data.stadt-zuerich.ch/dataset/vbz_fahrgastzahlen_ogd`.

how public transportation usage fluctuates throughout the day by analyzing passenger numbers and occupancy rates during different time intervals (e.g., morning rush hour, midday, and evening rush hour) to identify peak periods.

Predictive Modeling *(research question 5)*: We will utilize time series techniques such as ARIMA (Autoregressive Integrated Moving Average) and regression models such as linear regression or random forest regression to forecast passenger numbers and occupancy rates for future periods. Based on historical data, we estimate the confidence interval of our predictions and analyze the influence of different predictor variables. We contrast the importance of the predictor variable with previously defined assumptions, such as that the time of day has a high influence on passenger volume.

Data Visualization *(research question 3)*: Visualizing the data will enable us to identify patterns and trends more intuitively, facilitating more profound insights into Zurich's public transportation. The visualization of spatiotemporal passenger distribution will be achieved by mapping the geographical coordinates of stops provided by VBZ (Stadt Zürich Open Data (2024a)) to a city map. This offers a graphical representation illustrating passenger flows in relation to temporal variations throughout the day. Additionally, other interactive visualization tools may be utilized to provide a dynamic and exploratory view of the data, enhancing understanding of the analysis results.

1.6 | Adaptability and Outlook

Our research on Zurich's public transportation aims to provide insights by addressing our research questions. Consequently, we remain flexible in adjusting our data and analytical methodologies based on interim findings. For instance, the City of Zurich provides diverse datasets (Stadt Zürich Open Data, 2024b), encompassing motorized traffic, pedestrian movements, and public transportation schedules. Integrating these supplementary datasets into our analysis could build a more affluent data foundation, enabling more comprehensive responses to our research queries. Likewise, our outlined data analysis approaches serve as an initial pipeline and could be extended with alternative techniques for enhanced insights.

Appendix: Data Features

Table A.1: The features contained in the "VBZ passenger numbers dataset". The dataset consists of 5 separate CSV files that share a common ID called a foreign key. For each data attribute, we provide the attribute name, the technical definition (the feature name used in the dataset), and a corresponding definition.

Attribute	**Description**
Tag type ID (technical: Tagtyp_Id)	Foreign key to the TAGTYPE table
Line type ID (technical: Linien_Id)	Foreign key to the LINES table. The ID may vary from year to year.
Line name (technical: Linienname)	Line number used by VBZ (sometimes not identical to published line).
Scheduled trip ID (technical: Plan_Fahrt_Id)	ID of the scheduled trip.
Direction (technical: Richtung)	Direction of a trip. Values: 1 or 2.
Stop sequence (technical: Sequenz)	Sequence of stops for a trip. Values: 1, 2, 3 - Number of stops.
Stop ID (technical: Haltestellen_Id)	Foreign key to the STOPS table.
Following stop ID (technical: Nach_Hst_Id)	ID of the following stop. This field is empty for the last stop of a line.
Departure times (technical: FZ_AB)	Departure time as HH:MM.
Measurement count (technical: Anzahl_Messungen)	Number of measurements considered for this trip.
Passengers boarding (technical: Einsteiger)	Mean value of passengers boarding at the stop from the considered measurements. If the value for "Measurement count" = 0, the value is estimated.
Passengers alighting (technical: Aussteiger)	Mean value of passengers alighting at the stop from the considered measurements. If the value for "Measurement count" = 0, the value is estimated.

Table A.1: The features contained in the "VBZ passenger numbers dataset" (continued).

Attribute	Description
Occupancy (technical: Besetzung)	Mean value of occupancy from "Stop ID" to "Following stop ID" from the considered measurements. If "Measurement count" = 0, the value is estimated.
Trip distance (technical: Distanz)	Distance of the stop section from "Stop ID" to "Following stop ID" in meters.
DTV Days count (technical: Tage_DTV)	Days count for extrapolating to daily traffic volume (DTV). If the value is 0, counts have been taken but deemed non-representative for extrapolating to DTV.
DWV Days count (technical: Tage_DWV)	Days count for extrapolating to weekly traffic volume (DWV). If the value is 0, counts have been taken but deemed non-representative for extrapolating to DWV.
Saturday Traffic Days count (technical: Tage_SA)	Days count for extrapolating Saturday traffic.
Sunday Traffic Days count (technical: Tage_SO)	Days count for extrapolating Sunday and holiday traffic.
Night network (technical: Nachtnetz)	Numbers belong to the night network Fri/Sat or Sat/Sun night.
Saturday Night Days count (technical: Tage_SA_N)	Days count for extrapolating the night network on Friday to Saturday night.
Sunday Night Days count (technical: Tage_SO_N)	Days count for extrapolating the night network on Saturday to Sunday night.
Section ID (technical: ID_Abschnitt)	Calculated key for the stop section. Calculation: ("Stop ID" * 10000) + "Following stop ID". If "Following stop ID" is empty, it's calculated with 0.
Tag type name (technical: Tagtypname)	Name of the used tag type.
Tag type description (technical: Bemerkung)	Description of the tag type.
Transport system (technical: VSYS)	Transport system: T: Tram; TR: Trolleybus; B: Bus; SB: Cable car; FB: Forchbahn; N: Night bus.
Line name for passenger information (technical: Liniennname_Fahrgastauskunft)	Line name used in the timetable information.

Table A.1: The features contained in the "VBZ passenger numbers dataset" (continued).

Attribute	**Description**
VBZ internal stop number (technical: Haltestellennummer)	VBZ internally used stop number.
VBZ internal stop short name (technical: Haltestellenkurzname)	VBZ internally used stop abbreviation.
Stop name (technical: Haltestellenlangname)	Stop name. The number of characters available is limited, so some strange names may occur.
Seats (technical: SITZPLAETZE)	Number of seats excluding the driver's seat.
Capacity 1 person/m² (technical: KAP_1m2)	Capacity when all seats are occupied plus one person per m² of standing space.
Capacity 2 persons/m² (technical: KAP_2m2)	Capacity when all seats are occupied plus two persons per m² of standing space.
Capacity 3 persons/m² (technical: KAP_3m2)	Capacity when all seats are occupied plus three persons per m² of standing space.
Capacity 4 persons/m² (technical: KAP_4m2)	Capacity when all seats are occupied plus four persons per m² of standing space.

References

John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. The MIT Press, 2020. ISBN 978-0-262-04469-1.

Rudolf Mathar, Gholamreza Alirezaei, Emilio Rafael Balda Cañizares, and Arash Behboodi. *Fundamentals of data analytics: with a view to machine learning*. Springer, 2020. ISBN 978-3-030-56830-6.

OECD.org. Passenger transport, 2023. URL `https://data.oecd.org/transport/passenger-transport.htm`. Accessed on March 25, 2024.

Stadt Zürich Open Data. Fahrzeiten der vbz im soll-ist-vergleich, 2024a. URL `https://data.stadt-zuerich.ch/dataset/vbz_fahrzeiten_ogd`. Accessed on March 25, 2024.

Stadt Zürich Open Data. Fahrgastzahlen vbz, 2024b. URL `https://data.stadt-zuerich.ch`. Accessed on March 25, 2024.