FH Salzburg
MultiMediaTechnology

Detecting and Visualizing Delay Hotspots in Vienna's Public Transport Network

Bachelor Thesis

Author: Martin Sonnberger
Advisor: Andreas Bilke, MSc
Repository: https://gitlab.mediacube.at/fhs45907/bachelor-thesis

Salzburg, Austria, May 2, 2023

Affidavit

I herewith declare on oath that I wrote the present thesis without the help of third persons and without using any other sources and means listed herein; I further declare that I observed the guidelines for scientific work in the quotation of all unprinted sources, printed literature and phrases and concepts taken either word for word or according to meaning from the Internet and that I referenced all sources accordingly.

This thesis has not been submitted as an exam paper of identical or similar form, either in Austria or abroad and corresponds to the paper graded by the assessors.

Date *Signature*

First Name *Last Name*

Kurzfassung

Um die Fahrgastzahlen und die Kundenzufriedenheit in öffentlichen Verkehrsnetzen zu erhöhen, müssen Verkehrsbetriebe verschiedene Maßnahmen ergreifen, um eine attraktive Alternative zu anderen Verkehrsmitteln zu bieten. Neben dem Angebot eines umfangreichen Netzes und der Verbesserung der Infrastruktur ist eine dieser Maßnahmen die Verringerung von Verspätungen und die Bereitstellung zuverlässiger Verbindungen. Um Verspätungen effizient zu reduzieren, ist es wünschenswert, Hotspots im Netz zu erkennen, an denen die meisten Verspätungen auftreten. Ziel dieser Arbeit ist es, Abfahrtsdaten zu analysieren, um potenzielle Hotspots zu finden und zu untersuchen, ob sich diese Hotspots an Umsteigeknoten und Endstationen befinden. Zu diesem Zweck wurden über einen Zeitraum von 30 Tagen Abfahrtsdaten im Wiener Nahverkehrsnetz erhoben. Nach einem Überblick über frühere Untersuchungen in anderen Städten werden die Ergebnisse der Analyse und mögliche Gründe für die Ergebnisse in dieser Arbeit diskutiert. Zusätzlich werden dem Leser Visualisierungen der gesammelten Verspätungsdaten präsentiert, um die Ergebnisse in einer leicht verständlichen Form darzustellen.

Abstract

In order to increase passenger numbers and customer satisfaction in public transport networks, transit authorities have to take various measures to provide an attractive alternative to other forms of transport. Besides offering an extensive network and improving infrastructure, one of those measures is reducing delays and providing reliable services. To efficiently reduce delays, it is desirable to detect hotspots in the network where delays occur the most. This thesis aims to analyze departure data to find potential hotspots and answer whether those hotspots are at transfer hubs and terminal stations. To do so, departure data in Vienna's public transport network was collected over 30 days. After reviewing previous research conducted in other cities, the results of the analysis and possible reasons for the results will be discussed in this thesis. Additionally, visualizations of the collected delay data will be presented to the reader in order to show the results in a more easy-to-understand way.

Table of Contents

List of Figures

Listings

List of Tables

Abbreviations and Acronyms

API Application Programming Interface

CSV comma-separated values

FPTF Friendly Public Transport Format

GTFS General Transit Feed Specification

HAFAS HaCon-Fahrplan-Auskunfts-System

JSON JavaScript Object Notation

RDBMS relational database management system

SQL Structured Query Language

VOR Verkehrsverbund Ost-Region

1 Introduction

Good public transportation is one of the most important factors for quality of life in any big city. A study by the European Union found that after frequency, reliability is the second biggest contributing factor to passenger satisfaction with public transport (European Commission 2020, 66). With the European Union's commitment to achieving 55% less greenhouse gas emission in comparison to 1990, it is of high importance to increase both usage of and satisfaction with public transport. In Vienna, Austria both of these numbers are already at very high levels with 55% of people using the public transport network on a typical day and 95% average passenger satisfaction, the best value among capital cities in the EU (European Commission 2020, 61). To be able to reduce delays and thus increase reliability, one has to first identify areas where delays occur most often and to the largest extent. After identifying these areas, possible reasons for occurring delays have to be explored, considering both internal and external reasons (Van Oort et al. 2015, 372). Only after both of these steps have been completed, effective measures can be researched and implemented.

The aim of this thesis is to identify said problem areas by collecting and analyzing departure data from the transport agency that operates metros, trams and buses in Vienna. Since raw numbers alone often do not provide an intuitive understanding of large data sets, different visualizations will be presented that aim to give the reader an easy-to-understand overview of potential delay hotspots. In particular, the hypothesis that delays occur more often at transfer hubs and terminal stations will be investigated. At transfer hubs, passenger volume is the highest which can lead to increased dwell times and thus delays. At terminal stations on the other hand, delays that occurred at the beginning of a trip can accumulate and therefore form delay hotspots.

The thesis is structured as follows. Section 2 presents previous research and different formats of public transport data that are available in Vienna. Section 3 first explains some relevant information about the network itself and continues to describe the data collection process in detail. For this process, a custom script was written that persisted the necessary data over a duration of 30 days. Next, the results of the conducted data analysis are presented, highlighting different aspects in varying levels of detail, starting with network-wide analysis and continuing with a closer look at single lines and stations. Additionally, visualizations that provide new views on the data set are presented, showing all stations on a map and coloring them by their punctuality rate in addition to plotting the punctuality rate by time of day. The results and possible reasons causing those results are discussed in section 4, mainly focusing on reasons like increased traffic or scheduling effects. Finally, this thesis concludes by summarizing the results found in previous sections. Furthermore, an outlook on possible further research and applications of public transport data is provided.

2 Related Work

2.1 The Hague

Van Oort et al. (2015) analyzed the bus and tram network of The Hague, Netherlands for their research of developing a tool that provides customers with predictions regarding the punctuality of their trips. First, they explained various reasons for delays, dividing them into *internal* and *external* reasons. Internal meaning that the problem originates from the network operation itself, for example, schedule quality, infrastructure design or driver behavior. External reasons on the other hand include things that the network operator has no influence over, such as traveler behavior, other traffic or simply bad weather conditions (Van Oort et al. 2015, 372–374).

After collecting data over several months using a newly available open data API provided by the Dutch government, the authors analyzed and visualized that data in varying degrees of detail. For the first visualization, as seen in fig. 1, they plotted all recorded trips of a single transit line on a line diagram, with stops on the x-axis and delay in minutes on the y-axis. This diagram shows good general punctuality and indicates that there are multiple stations where vehicles arrive early and wait for their scheduled departure. In a second line diagram, the authors instead showed the interval in minutes on the y-axis. Besides absolute delay, this is an important measure as a transit line with steady delays, yet high-frequency intervals are still desirable for passengers. Both diagrams together show that in the last third of the route, delays accumulate the most, leading to both higher absolute delays and increased intervals. (Van Oort et al. 2015, 378).

In fig. 2, one can see a network-wide visualization, where each station is colored in a gradient from green to red, representing the average delay recorded at that station. It shows that overall there are only a few problem areas which are mostly series of stops where delays may accumulate. Interestingly, there are stations with a negative average delay, meaning an earlier-than-planned arrival. This effect can in certain cases be an expected result of including a leeway in schedules for important transfer hubs. If not intentional, such early arrivals can simply be solved by updating scheduled departure data (Van Oort et al. 2015, 379–380).

After this static analysis of historical data, the authors fitted the data to a normal distribution in order to provide passengers with so-called *enriched travel advice*, where predictions based on said distribution are added to trip planning results. This information includes probabilities of delays, early departures or transfer feasibility. Additionally, they developed a tool for a more scientific analysis of collected data. This software was used on another data set from the transit network of Utrecht and was able to detect multiple bottlenecks in the network, for example, long dwell time on bus routes (Van Oort et al. 2015, 380–387).

2.2 Portland

In 1993, Strathman and Hopper (1993) presented a baseline study of the Portland metropolitan area transit network as part of an operations control plan implemented by the city's transit agency *Tri-Met*. This plan aimed to improve the growing challenges the network was facing

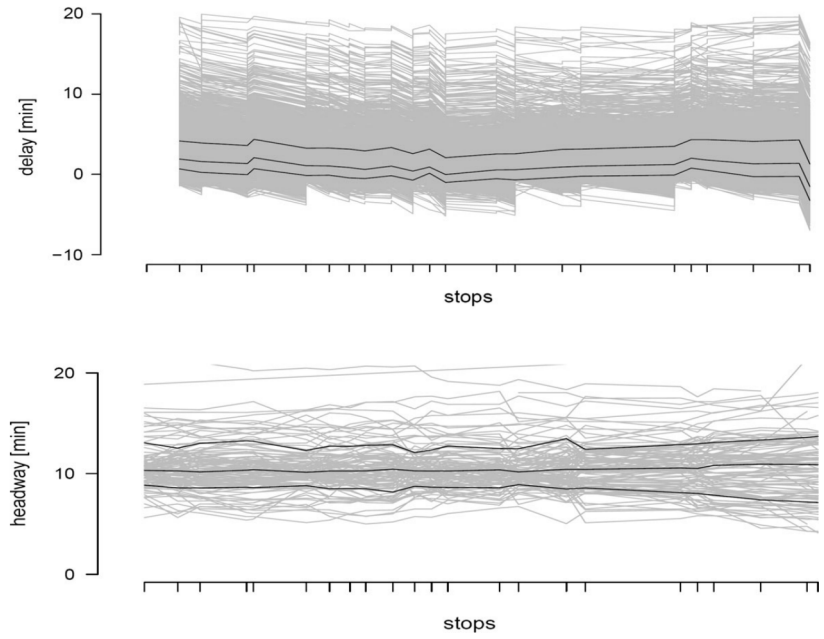

Figure 1: Vehicle delays (upper) and headways (lower) along a single route (Van Oort et al. 2015, 378)

regarding service reliability. Like Van Oort et al. (2015), the authors state that for short intervals, passengers do not care about vehicles running exactly on time, but rather whether the service runs regularly and reliably. In addition to grouping causes for delays into internal and external ones, the authors divide possible countermeasures into short-term and long-term actions. Short-term actions are ones to counter unanticipated delays such as holding vehicles at a stop in case of early arrival or adding additional vehicles to a route. Long-term actions on the other hand are helpful for systemic delays which are not caused by a single unexpected event. Examples include improving driver behavior through additional training or a reward system for good on-time performance. Additionally, the schedule itself can be the root cause and a redesign with longer run times or longer layover times at the end of lines may be the solution for consistent delays. The authors also state that a more complex network with longer routes generally results in worse on-time performance than one with short routes and only a few stops. Also, lines where the highest loads occur at the first few stations are more likely to show higher delay rates. As a result, public transport providers have to balance satisfying the demands of passengers and thus network complexity with on-time performance, taking various factors such as efficient use of resources into consideration. (Strathman and Hopper 1993, 93–94).

The study was conducted by manually recording arrival times on 200 trips on 59 different bus lines in the Portland metropolitan area, resulting in a total of 1552 recorded bus arrival times. Of those, 1360 or 87.6% were punctual, with an arrival considered punctual if it falls within a range of being one minute early to five minutes late. The authors found that driver experience has a measurable effect on on-time performance, with part-time drivers being more prone to

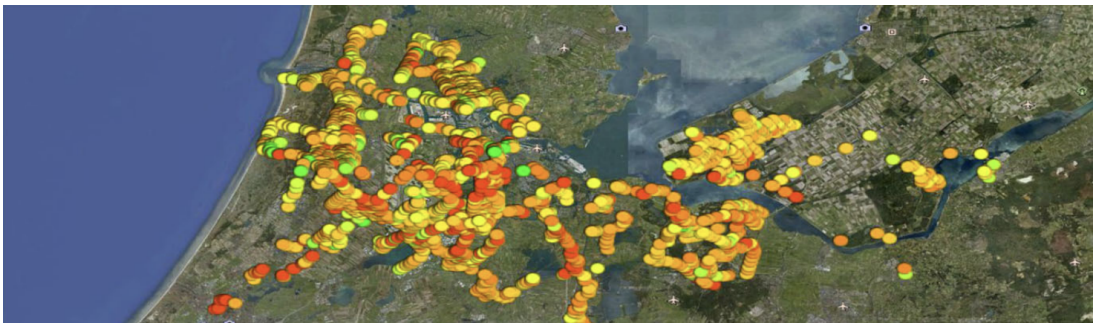

Figure 2: Average delay per stop (green early, yellow on time, red late) (Van Oort et al. 2015, 379)

late arrivals than their full-time colleagues. Furthermore, they confirmed that early arrivals and delays occur more often during afternoon peak hours. Interestingly, they did not observe this effect for morning peak hours. In fact, arrivals were more likely to be early than late if they did not follow the schedule, indicating that planners overcompensated for morning rush hour traffic by increasing running more than necessary (Strathman and Hopper 1993, 97–99).

2.3 New York City

A team of researchers from the Data, Research, Development (DRD) department of the New York City Transit Authority developed a tool for visualizing problem areas in New York's subway system. It detects areas where trains run slower than normal or show longer than normal intervals between trains. In addition to highlighting the results with a severity level, an interactive map display the locations of trains in real-time. This interactive map is used by the communications team in order to inform customers via social media channels about those problems occurring in the network. While there were similar tools already available for the New York transit network, these tools had significant limitations such as only allowing to view one direction of one line at a time. This required the monitoring staff to switch between the lines constantly to keep an overview. (Caspari et al. 2021, 1–3)

To gather the required train location data, the authors used a combination of real-time station arrival feeds and data provided by the signaling system. To compare running times between two stations or signaling segments, they relied on manual recordings by the transit authority rather than schedule data. They found that actual running times can deviate from scheduled ones and thus using schedules alone would result in less accurate calculations (Caspari et al. 2021, 4). Both large intervals and slow trains are calculated and displayed on multiple levels of detail. Longer-than-normal intervals are first calculated for stops or track segments, then aggregated into bigger areas of three or more stops if applicable. For slow trains it is done similarly, ranging from single stuck trains to entire areas consisting of multiple stations where trains run significantly slower than expected. All results are then categorized into *moderate*, *severe* and *very severe* problems, aiming to provide passengers with consistent information and suggested actions depending on the severity level (Caspari et al. 2021, 6–7).

The interactive map that Caspari et al. (2021) developed needed to cater to various needs, providing both an overview to detect emerging problem areas and the ability to see detailed information about a specific problem. The solution was a zoomable map, in addition to a text-based sidebar and tooltips for stations and line segments. Additional features include different available themes for severity levels, dark and light themes for the entire app and a text-only mode that displays the data in a tabular layout instead of the interactive map (Caspari et al. 2021, 8–9). The tool was developed for more than one year and is in successful use since then. Being entirely web-based, it even allowed for easy access during the COVID-19 pandemic for the communications personnel (Caspari et al. 2021, 11).

2.4 Stochastic models

Various authors have studied the fitting of delay data to statistical distribution models in order to be able to make predictions over possible future delays. Yuan (2008) found that while other distributions may provide more flexibility and adjustable parameters, a negative exponential distribution is most of the time a valid and easy-to-understand distribution to model both arrival and departure delays (Yuan 2008, 173–174). Depending on the data set and the exact delay variable being measured, other distributions might fit better. For example, Goverde, Corman, and D'Ariano (2013, 90) fitted their data of trains arriving in Utrecht to a Weibull distribution, which is a distribution that receives multiple input parameters and—depending on one of those parameters—can result in an exponential distribution too (Hallinan Jr 1993, 88). Yuan (2006, 61) confirmed that the Weibull distribution most of the time offers the best fit. In some cases, however, other distributions such as beta, gamma or log-normal distributions model departure delays the best (Yuan 2006, 61).

2.5 Data availability

2.5.1 HaCon-Fahrplan-Auskunfts-System (HAFAS)

In Germany, Austria and Switzerland but also other European countries, HAFAS, short for *HaCon-Fahrplan-Auskunfts-System* (HaCon Timetable-Information-System), is a popular system for retrieving timetable information as well as intermodal journey planning. It is developed by a subsidiary of Siemens and available since the late 1980s, and with that one of the first tools on the market, promising an algorithm that performs route calculation in less than six seconds. Each public transport company gets its own HAFAS deployment, which all share common terminology and endpoints, but can however use custom configurations and feature availability. While the endpoints are usually not openly documented by the respective companies, they do not restrict access using API tokens or similar measures and are therefore easily accessible (Computerwoche 1988; Redmann 2023b).

For this thesis, the HAFAS endpoint by Verkehrsverbund Ost-Region (VOR), the transport authority of the eastern regions of Austria, was used for collecting the necessary data. While the

API itself returns data in the HAFAS Raw Data Format, the JavaScript library *hafas-client* converts this into the much easier-to-work-with Friendly Public Transport Format (FPTF), which is inspired by the widely used GTFS format and uses JavaScript Object Notation (JSON) as its serialization format (Redmann 2023a).

The library provides a multitude of endpoints, both for station and trip data. Aside from fetching information about a stop or station, the `nearby` endpoint for example can retrieve all locations that are within a given walking distance in minutes, which is especially useful for building consumer-facing public transport applications. It also includes endpoints for journey planning between two given locations. Similar to `nearby`, the `reachableFrom` endpoint returns all stations that are reachable from another station, but instead of walking distance public transport journeys are calculated. Another endpoint that provides opportunities for interesting applications is `radar`, which returns the locations of all vehicles within a given radius. While location data is not available for all vehicles, this can certainly be used as a foundation for appealing visualizations. The two most important endpoints for this thesis however are `arrivals` and `departures`, which provide accurate real-time data for one or more given station(s). In addition to planned and actual arrival/departure times, they also calculate the difference, providing the consumer with a ready-to-use delay value in seconds (Redmann 2023a, 2023c).

2.5.2 General Transit Feed Specification (GTFS)

GTFS is a format specification initially developed by Google which later—after its widespread usage in many non-Google systems—has been renamed from *Google Transit Feed Specification* to *General Transit Feed Specification*. It allows public transport agencies to offer their data in a unified format, understood by various consumer applications, ranging from visualization tools to trip-planning mobile applications. The core of GTFS is a static feed of text files that contain the whole schedule and additional information about the network. The GTFS reference document specifies five required files: `agency.txt`, `stops.txt`, `routes.txt`, `trips.txt` and `stop_times.txt`. One can see that the structure of the data is similar to HAFAS. Routes are a group of trips that are offered to passengers as a line or service. Arrival and departure data for those trips are then collected in `stop_times.txt` (MobilityData 2022b).

In contrast to HAFAS however, these static files do not provide any historical or real-time data, but instead the planned schedule at the time of publishing. For real-time data, there exists a special extension to GTFS called GTFS Realtime. As a data format, GTFS Realtime uses Protocol Buffers, a language-agnostic serialization format developed by Google. As real-time updates naturally come with a significantly higher level of detail than static text files, the provided data structures are entirely different. Instead of static text files, messages such as `TripUpdate`, `VehiclePosition` or `StopTimeEvent` are exchanged between the feed provider and connected clients (MobilityData 2022a).

Since 2017, the city of Vienna provides GTFS schedule data as part of the *Open Government Data* initiative. With its real-time extension, GTFS would have been able to provide all the necessary data for this thesis, especially planned and actual departure data. However, as the

hafas-client library uses the JSON-based FPTF format instead of more complicated Protocol Buffers, HAFAS was chosen as a primary data source instead of GTFS.

3 Implementation

In this section, after introducing Vienna's public transport network, the data collection process is described in detail. Next, the results of the delay analysis are presented, showing results both on a network-wide level and in more detail for single lines or stations. Finally, visualizations that try to make the large data set easier to understand are presented. All information about the network, especially locations of stations and lines was gathered from the network map provided by the network operator Wiener Linien (Wiener Linien 2023).

3.1 Vienna's public transport network

The public transport network in the city of Vienna is very popular and well-received among the city's population. A case study by Haslauer et al. (2015, 917, 921) showed that 75 percent of the population uses the network multiple times per week and that the average satisfaction lies at 1.57 when rated on a scale from 1 to 5, with 1 being the best score (very satisfied). One factor of the high popularity is the comparatively cheap access to the system with a yearly ticket costing 365 euros, resulting in 852,300 of those tickets sold in 2019 (Wiener Linien 2020).

The network is owned and operated by Wiener Linien, a subsidiary of Wiener Stadtwerke Holding, which in turn is fully owned by the city of Vienna. The network consists of 5 metro lines, 28 tram lines and 131 bus lines and has a total length of 1169 kilometers (Wiener Linien 2020). Additionally, there is a suburban rail line connecting the city center with the town of Baden, operated by another subsidiary, Wiener Lokalbahnen. The line contributes 10.3 million to the total 606.1 million passengers transported by both companies in 2021 (Wiener Stadtwerke Group 2022, 24). Furthermore, Austria's federal railway operator ÖBB runs 10 suburban train lines from Vienna to and from neighboring towns in Lower Austria, adding another 89 million yearly passengers to Vienna's transit network (ÖBB-Personenverkehr AG 2023b).

3.2 Data collection

The necessary data for detecting potential delay hotspots were collected over a duration of one month, specifically from February 18, 2023 to March 20, 2023. The collection was done by automated scripts using the *hafas-client* JavaScript library available on GitHub.[1] It can be used to query HAFAS API endpoints from various public transport companies. In the case of Vienna, the relevant company is VOR, the transport authority for the eastern region of Austria, which includes the federal states of Vienna, Lower Austria and Burgenland. The following subsections

1. https://github.com/public-transport/hafas-client (Accessed April 30, 2023)

try to explain in more detail how both station and departure data were collected and persisted for further analysis.

3.2.1 Stations

Departure and delay data can be queried using the `departures` method available in *hafas-client*. This method expects a station object or identifier as its argument, thus a list of such station identifiers had to be created first. Since there exists no method to get a collection of all stations the network contains, the `nearby` method was used to get a list of stations within a radius (measured in meters of walking distance) from a specified coordinate pair. An example of such a query can be seen in listing 1; a `station` object contained in the response is shown in listing 2.

```
import { createClient } from "hafas-client";
import { profile } from "hafas-client/p/vor/index.js";

const hafas_client = createClient(profile, "hafas-ba");

const center = {
    type: "location",
    latitude: 48.2084,
    longitude: 16.3778,
};

const locations = await hafas_client.nearby(center, {
  products: {
    tram: true,
    "u-bahn": true,
    "city-bus": true,
  },
  subStops: false,
  entrances: false,
  linesOfStops: true,
  results: 5000,
  distance: 100_000,
});
```

Listing 1: Retrieving nearby stations from given coordinates

```
{
  "type": "station",
  "id": "490132000",
  "name": "Wien Stephansplatz",
  "location": {
    "type": "location",
    "id": "490132000",
    "latitude": 48.208133,
    "longitude": 16.371631
  },
  "products": {
    "train-and-s-bahn": false,
```

```
    "u-bahn": true,
    "tram": false,
    "city-bus": true,
  },
  "isMeta": true
}
```

Listing 2: `station` object returned by the HAFAS API

The idea was to choose a point in the city center and get all available stations by using a maximum walking distance of 20,000 meters, far exceeding the theoretical maximum distance to the city borders following the fact that Vienna has a maximum north-south and east-west extension of 22.8 and 29.4 kilometers respectively (Stadt Wien 2022, 14). It was found though that the number of stations did not increase with a maximum walking distance higher than 7200 meters, capping out at 997 found stations. In order to retrieve the rest of the stations, four additional coordinate pairs were selected which cover all missing regions, especially ones near the city border. Finally, duplicated results and results that did not include the prefix "Wien" were removed, as those are situated in the neighboring state of Lower Austria which is not covered by this thesis. Using this described method, a total of 1756 stations were collected and saved to an SQLite database. From there, they can be used for the next step, querying departure and delay data from these stations.

3.2.2 Departures

Collecting the desired delay data was achieved by using the `departures` method of the *hafas-client* library. The method receives a list of stations and returns departure objects for those stations, which can be seen in listing 3 and listing 4. These objects contain properties for planned and actual departure times and with that the resulting delay, if applicable.

```
import { createClient } from "hafas-client";
import { profile } from "hafas-client/p/vor/index.js";

const hafas_client = createClient(profile, "hafas-ba");
const station_id = "490132000"; // Wien Stephansplatz

const { departures } = await hafas_client.departures(station_id, {
  duration: 50,
  subStops: false,
  entrances: false,
  results: 60,
});
```

Listing 3: Retrieving departures from a given station ID

```
  {
    "tripId": "2|#VN#1#ST#1680820733#PI#0#ZI#66304#TA#1#DA...",
    "stop": {...},
    "when": "2023-04-07T09:33:00+02:00",
    "plannedWhen": "2023-04-07T09:32:00+02:00",
```

```
    "delay": 60,
    "platform": "2",
    "plannedPlatform": "2",
    "prognosisType": "prognosed",
    "direction": "Wien Alaudagasse",
    "provenance": null,
    "line": {
      "type": "line",
      "id": "vor-21-u1-j23-3",
      "fahrtNr": "586",
      "name": "U1",
      "public": true,
      "adminCode": "v04WL_",
      "mode": "train",
      "product": "u-bahn",
      "operator": {
        "type": "operator",
        "id": "wiener-linien",
        "name": "Wiener Linien"
      }
    },
    "remarks": [
      { "type": "hint", "code": "LF", "text": "Niederflurfahrzeug" }
    ],
    "origin": null,
    "destination": {
      "type": "stop",
      "id": "490001409",
      "name": "Wien Alaudagasse",
      "location": {...},
      "products": {...},
      "station": {...}
    },
    "currentTripPosition": {
      "type": "location",
      "latitude": 48.213994,
      "longitude": 16.383191
    }
  }
```

Listing 4: A `departure` object returned by the HAFAS API

Since the HAFAS API did not reliably work when fetching departures for all 1756 stations, an alternative strategy was chosen. Instead, only one station at a time was queried in a one-second interval, specifying a time span of 50 minutes of upcoming departures included in the result. After every station was visited, the cycle starts from the beginning, which results in each station being visited approximately every 30 minutes. A unique ID consisting of the identifiers of both the trip and station was assigned to each departure result in order to detect results that were already included in previous cycles. In those cases of duplicates, the newer result was selected in order to ensure the most current and accurate data was saved for each departure. The source

code for this process can be seen in listing 5. This script was then executed with `nohup` so that it can continuously run in the background on a Linux server during the collection period.

```
import { hafas_client } from "./client.js";
import { db } from "./db.js";
import { sleep } from "./utils.js";

const station_ids = await db.selectFrom("stations").select("id").execute();
let i = 0;

while (i < station_ids.length) {
  const { id: station_id } = station_ids[i];
  const { departures } = await hafas_client.departures(station_id, {
    duration: 50,
    subStops: false,
    entrances: false,
    results: 60,
  });
  let inserted = 0;

  for (const dep of departures) {
    const new_dep = {
      id: "",
      direction: dep.direction,
      delay: dep.delay,
      when: dep.when,
      planned_when: dep.plannedWhen,
      station_id: dep.stop?.station?.id ?? dep.stop?.id,
      station_name: dep.stop?.station?.name ?? dep.stop?.name,
      line_id: dep.line?.id,
      line_name: dep.line?.name,
      product: dep.line?.product,
    };
    new_dep.id = `${new_dep.station_id}_${dep.tripId}`;
    const result = await db
      .insertInto("departures")
      .values(new_dep)
      .onConflict((oc) => oc.column("id").doUpdateSet(new_dep))
      .execute();
    inserted += +result[0].numInsertedOrUpdatedRows ?? 0;
  }

  await sleep(1000);
  i++;

  if (i === stations.length) {
    i = 0;
  }
}
```

Listing 5: A `departure` object returned by the HAFAS API

3.2.3 Persisting the collected data

For persisting the collected data, SQLite, a self-contained relational database management system (RDBMS) was used together with the TypeScript based *kysely* library for building Structured Query Language (SQL) queries. Using a RDBMS with a query builder has the advantage of providing a great developer experience, with type-safe methods for querying the database and features like insertion conflict handling, which was especially important for assuring no duplicate stations and departures were saved.

While a RDBMS like SQLite is a good tool for data collection, the resulting departures table was converted into a comma-separated values (CSV) file for further persistence and analysis. One reason for that is the substantially smaller file size of 4.6 GB, compared to the SQLite file's size of 8.2 GB, resulting in a reduction of around 44%. The collected data set contains a total of 13,720,298 recorded departures with an average of 457,347 departures per day.

3.3 Delay analysis

3.3.1 Filtering irrelevant data

Since the main focus of this thesis is to analyze the public transport network of Vienna, long-distance train services were stripped of the resulting data set. Since the API endpoint does not differentiate between long-distance trains and suburban regional trains, the latter also were removed in this step. This means that in the following analysis, only public transport products by Wiener Linien, which includes metros, trams and buses, are considered. Long-distance trains have widely different characteristics in comparison to high-frequent inner-city transport methods such as trams or buses. These differences were quickly visible when conducting the first screening of the collected data. The most delayed trips were all Railjet (Austrian high-speed train service) services from neighboring countries, with delays in the multiple-hour range. While an interesting observation, these types of delays are not the intended research area of this thesis, which instead lays its focus on those inner city lines mentioned above. Additionally, the hybrid tram/train line *Badner Bahn* also had to be excluded from further analysis since the collected data did only include scheduled timetable departures and not real-time departure times. This means that for this particular line, no delay data could be collected.

3.3.2 Tools and technologies

As mentioned in section 3.2.3, the resulting SQLite file was converted into a CSV file for easier organization and handling. The following results were all gathered by utilizing the widely used data analysis library *pandas* for Python. Visualizations on the resulting data frames were then produced by *plotly*, another Python library commonly used for visualization and charting tasks. The most common figures such as line charts, bar charts, box plots or scatter plots are all available through the *plotly-express* subpackage, which was used for all following visualizations. For figures including a map, the underlying map data is provided by *Mapbox*.

count	13,720,398
mean	14.68
std	115.62
min	-3600
25%	0
50%	0
75%	0
max	10,920

Table 1: Basic summary of `delay` column

Delay threshold [min]	Punctuality rate
0	0.8791
1	0.9417
2	0.9704
3	0.9837
4	0.9895
5	0.9925

Table 2: Punctuality rates depending on minimum delay

3.3.3 Basic summary

To gather an initial overview of the collected data, we first want to highlight a few general characteristics of the data set. Table 1 shows the basics statistical data points of the `delay` measurements. One can see that the mean delay was 14.68 seconds, with a minimum of -3600, meaning a one-hour early departure, and a maximum delay of 10,920 seconds, or approximately 3 hours. The high standard deviation of 115.62 seconds indicates that the data points are highly spread out around the mean. It is therefore unsurprising that the 25^{th}, 50^{th} and 75^{th} percentile are all 0 since actual punctuality rates of public transport networks mostly lay between 90 and 99 percent, for example with national rail services in Vienna having punctuality rates of around 96% (ÖBB-Personenverkehr AG 2023a). In fact, table 2 shows that 12% of the recorded departures had a delay greater than zero. However, the definition of punctuality varies from company to company and rates are usually calculated with a specified maximum delay threshold of one to five minutes (Chen et al. 2009, 723). With a maximum delay of one minute, the punctuality rate falls at 94.17%, while a maximum of five minutes results in 99.25% punctuality. For all further punctuality rate calculations, we assume a departure is punctual if it has a maximum delay of three minutes or less.

Line	Product	Punctuality rate	Mean delay [s]
ZF	Bus	1.0	0.00
73A	Bus	1.0	0.00
34A	Bus	1.0	0.00
U6E	Metro/bus	1.0	0.00
N29	Bus	1.0	0.97
N66	Bus	1.0	1.54
2A	Bus	1.0	1.69
N46	Bus	1.0	5.99
N62	Bus	1.0	22.17
54B	Bus	0.9998	4.56

Table 3: Top ten most punctual lines

Line	Product	Punctuality rate	Mean delay [s]
42A	Bus	0.9030	149.01
10A	Bus	0.9245	51.89
63A	Bus	0.9505	30.05
95A	Bus	0.9527	39.14
71	Tram	0.9551	24.87

Table 4: Top five least punctual lines

3.3.4 Most and least delayed lines

When grouping all departures by line and sorting by punctuality rate, it shows that the most punctual lines are all bus lines, as shown in table 3. Also, the nine most punctual lines all have perfect punctuality rates of 100%, the first line with a rate less than 100% is bus 54B with 99.98%, landing in tenth place. Six of those lines with 100% punctual departures even have a mean delay of 6 seconds or less, meaning only very few or no delays were recorded. Line N62 is the first one with a higher mean delay of 22.17 seconds, but still a perfect punctuality rate of 100%. When looking at lines with the lowest punctuality rate in table 4, one can see that again buses dominate and only one tram line appears as the fifth least punctual line, namely tram 71 with a punctuality rate of 95.51% and a mean delay of 24.87 seconds. Line 42A has the lowest punctuality rate and also by far the highest mean delay with 149.01 seconds, while the other three bus lines show mean delays between 30.05 and 51.89 seconds.

Because of the dominance of bus lines, a further differentiation between bus, tram and metro lines may be of interest. Table 5 shows the city's five metro lines ordered by their punctuality rate. Line U6E is a special case as it is categorized as a metro line while it actually operated as a bus and acted as a replacement for a few stops on line U6 during one weekend of construction work. All other, actual metro lines show quite high punctuality rates too, with U2 being the least punctual line which still had 98.65% punctual departures. Line U6 is the most punctual

Line	Punctuality rate	Mean delay [s]
U6E	1.0	0.00
U6	0.9950	10.14
U1	0.9944	7.89
U3	0.9926	12.11
U4	0.9882	14.95
U2	0.9865	18.52

Table 5: Metro lines by punctuality rate

Line	Punctuality rate	Mean delay [s]
52	0.9977	4.22
U2Z	0.9955	-32.24
33	0.9954	-15.99
37	0.9929	10.53
30	0.9928	-2.41

Table 6: Top five most punctual tram lines

line, showing a punctuality rate of 99.5%. Mean delays generally increase for lower punctuality rates, though remain at a low level with values between 7.89 and 18.52 seconds.

Table 6 and table 7 show only tram lines, with line 52 showing the highest punctuality rate of 99.77% and line 71 the lowest with 95.51%. Similar to U6E, U2Z is a replacement line during construction work. However, in daily operation, it is treated like a regular tram line, as it is planned to be in service for more than two years (Stadt Wien 2021). For the most punctual trams, mean delays show very low values, with three out of the top five having negative values ranging as low as -32.24 seconds for line U2Z. The least punctual trams lines have mean delays of around 30 seconds, only line 1 has a lower value of 10.88 seconds.

Line	Punctuality rate	Mean delay [s]
71	0.9551	24.87
2	0.9554	26.62
1	0.9678	10.88
D	0.9727	32.49
O	0.9737	23.80

Table 7: Top five least punctual tram lines

Station	Products	Punctuality rate	Mean delay [s]
Blaasstraße	Bus	0.8093	115.53
Schafbergbad	Bus	0.8870	178.62
Schafberg/Werfelstraße	Bus	0.8957	164.45
Twarochgasse	Bus	0.8961	164.66
Minciostraße	Bus	0.8971	83.30
Ulanenweg	Bus	0.8975	62.63
Kieslerweg	Bus	0.8990	67.12
KGV Alsrückenweg	Bus	0.8991	155.32
Korbweidenweg	Bus	0.9000	70.19
Krenngasse	Bus	0.9005	153.04

Table 8: Top ten least punctual stations

3.3.5 Most and least delayed stations

Following the fact that entire lines showed punctuality rates of 100%, all stations that serve these lines (but not others), also have a punctuality rate of 100%. More interesting are the stations with the lowest rates of punctual departures, shown in table 8. They are all bus stops again, interestingly however the very least punctual stop, *Blaasstraße*, has a difference in punctuality rate to the next one (i.e. second least punctual) of almost eight percentage points while the maximum difference of all other neighboring punctuality rates in fact lies at 0.87 percentage points. Also, while mean delay values are generally high for the least punctual stations, they do not strictly increase with lower punctuality rates. *Krenngasse*, for example, has the fifth-highest mean delay but ranks tenth for punctuality rate. Figure 3 shows box plots of the stations' delays with stations grouped by their available products. In these plots, *Blaasstraße* is clearly visible on the very left of the plot for bus stations and overall, indicating it is an outlier.

Again, since most and least delayed stations are all bus stations and to get more results about highly frequented stations that are relevant to more passengers, table 9 shows only the highest-ranked stations that serve at least one metro line, while table 10 shows the five least punctual metro stations in the network. Of those stations, *Neue Donau* is the most punctual with a punctuality rate of 99.9%, while *Aspern Nord* shows the most delays with its punctuality rate being 97.11%. The five most punctual stations show low mean delays between 2.73 and 10.71 seconds while the least punctual metro stations still have mean delays of around 20 seconds, only five seconds higher than the average of the whole data set. Three out of the five stations with the lowest punctuality rates serve metro, tram and bus lines, while all of the five most punctual stations have metro and bus services only.

3.3.6 Tram line 71

In previous sections, all lines or stations of the network were considered. In this section, we want to analyze and focus on one specific line. For this, tram line 71 was chosen for three

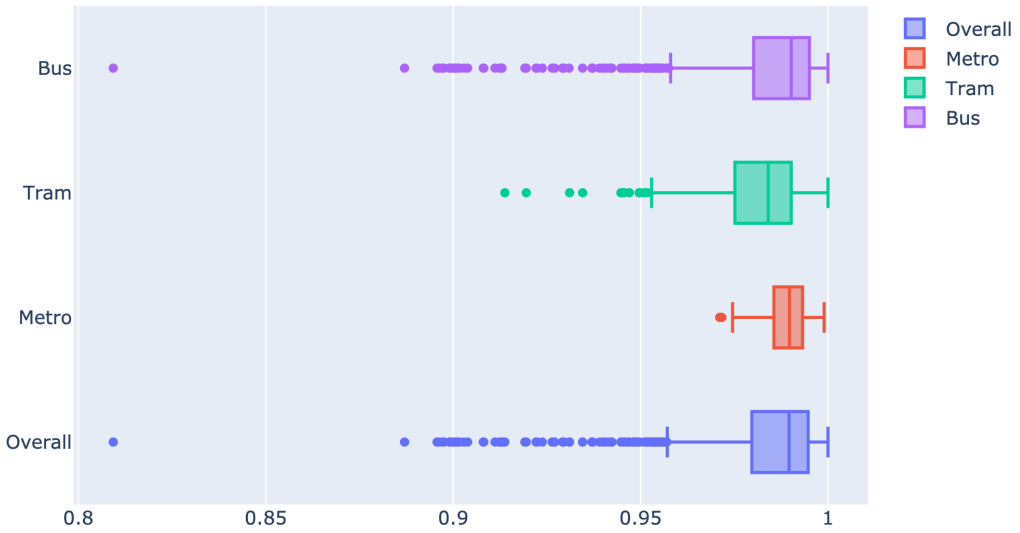

Figure 3: Box plots of delay rates grouped by station type

reasons: Firstly, it is a highly frequented line with intervals of 7-8 minutes on weekdays and therefore much departure data was collected and is available for analysis. Secondly, it is a radial line, starting in the city center and terminating in an outer district. This type of line represents the majority of tram lines in the network. Finally, it is the tram line with the lowest punctuality rate, promising interesting results for delay measures along the line.

Table 11 shows all 32 stations on line 71 sorted by their respective punctuality rate. Additionally, the mean delay was calculated for each station. The most punctual station is *Kasiereebersdorf, Zinnergasse* which shows a punctuality rate of 99.59% and a mean delay of -25.15 seconds. All other stations have positive mean delays with values ranging between 18.35 and 33.63 seconds. The least punctual station, *Valiergasse*, has a punctuality rate of 92.30% and

Station	Products	Punctuality rate	Mean delay [s]
Neue Donau	Metro, bus	0.9990	2.73
Siebenhirten	Metro, bus	0.9988	5.17
Hütteldorf	Metro, bus	0.9987	3.47
Seestadt	Metro, bus	0.9984	10.71
Nestroyplatz	Metro, bus	0.9966	6.28

Table 9: Top five most punctual stations serving metro lines

Station	Products	Punctuality rate	Mean delay [s]
Aspern Nord	Metro, bus	0.9711	21.36
Taborstraße	Metro, tram, bus	0.9717	21.40
Johnstraße	Metro, tram, bus	0.9746	23.66
Meidling Hauptstraße	Metro, bus	0.9785	17.75
Schwedenplatz	Metro, tram, bus	0.9789	15.21

Table 10: Top five least punctual stations serving metro lines

said mean delay of 33.63 seconds. Additionally, one can see that the mean delay generally increases with a lower ranking and thus lower punctuality rate. However, certain stations do not follow this pattern and show a higher mean delay than the next station down the list. For example, *Ring/Volkstheater U* has with 20.42 seconds a notably higher mean delay value than the next less punctual station, *Am Heumarkt*, which has a mean delay of 18.71 seconds. The same behavior can be observed for *Oper/Karlsplatz U* and its neighbor in the table *Unteres Belvedere*.

3.3.7 Station *Karlsplatz*

Like section 3.3.6 did for one line, this section focuses on the analysis of one station. For this, the station *Karlsplatz* was chosen, as it is one of the main transfer hubs in the network. It serves two metro lines, six tram lines and three bus lines. When construction finishes in the fall of 2023, the station's third metro line U2 will continue its service to *Karlsplatz*. In the meantime, tram U2Z serves as a replacement line, with one of its terminal stations located at *Karlsplatz*.

In table 12, departures at *Karlsplatz* are grouped by lines and again sorted by their respective punctuality rate. Line 2A, one of the inner city buses has the best punctuality rate with 99.70% punctual departures, together with a mean delay of 3.75 seconds. Next is U2Z, the replacement tram line for U2, showing a punctuality rate of 99.65% and, notably, a mean delay of -16.20 seconds. Metro lines U1 and U4 show considerably lower mean delays than tram lines, with U1 having a rather low value of just 8.29 seconds and a punctuality rate of 99.48%. Interestingly, tram lines 71, D, 62 and 2 all have similar mean delays of approximately 30 seconds. Line 1 however has a mean delay of only 0.76 seconds, even though its punctuality rate of 97.24% is similar to the other mentioned lines. The line with the least number of punctual departures is tram line 2 with its punctuality rate of 95.67% and mean delay of 29.59 seconds.

3.4 Visualizations

In section 3.3 we explored the available data and made interesting observations by looking at the numbers themselves and comparing them in a simple tabular form or with basic box plot diagrams. While this provided good insights, we as humans tend to comprehend visualizations better than raw numbers, especially for huge data sets with millions of entries such as the

Station	Punctuality rate	Mean delay [s]
Kaiserebersdorf, Zinnergasse	0.9959	-25.15
Rathausplatz/Burgtheater	0.9796	18.35
Parlament	0.9788	18.48
Burgring	0.9785	19.35
Börse	0.9784	18.83
Ring/Volkstheater U	0.9769	20.42
Am Heumarkt	0.9757	18.71
Schottentor	0.9747	22.08
Schwarzenbergplatz	0.9736	20.88
Oper/Karlsplatz U	0.9734	23.77
Unteres Belvedere	0.9699	21.18
Rennweg	0.9697	22.10
Kleistgasse	0.9679	20.34
Oberzellergasse	0.9666	20.40
St. Marx	0.9654	22.72
Litfaßstraße	0.9566	25.04
Molitorgasse	0.9554	25.76
Zippererstraße	0.9545	25.12
Hauffgasse	0.9526	26.26
Enkplatz	0.9459	29.50
Simmering	0.9444	30.23
Braunhubergasse	0.9441	29.26
Fickeysstraße	0.9424	29.95
Weißenböckstraße	0.9414	29.48
Zentralfriedhof 1.Tor	0.9393	29.49
Zentralfriedhof 2.Tor	0.9369	30.57
Zentralfriedhof 3.Tor	0.9337	31.50
Pantucekgasse/Widholzgasse	0.9334	30.75
Zentralfriedhof 4.Tor	0.9329	30.89
Leberberg	0.9322	30.87
Svetelskystraße	0.9288	31.78
Valiergasse	0.9230	33.63

Table 11: Stations on line 71 sorted by punctuality rate

Line	Product	Punctuality rate	Mean delay [s]
2A	Bus	0.9970	3.75
U2Z	Tram	0.9965	-16.20
U1	Metro	0.9948	8.29
4A	Bus	0.9889	12.69
59A	Bus	0.9854	18.83
U4	Metro	0.9811	17.84
71	Tram	0.9734	23.77
1	Tram	0.9724	0.76
D	Tram	0.9721	37.09
62	Tram	0.9702	30.18
2	Tram	0.9567	29.59

Table 12: Lines at *Karlsplatz* sorted by punctuality rate

one on hand (Ware 2019, 2). Therefore, the following section presents and discussed some visualizations which aim to show the collected delay data in a way that is easy to understand while still providing the reader with valuable information. Specifically, the first visualizations show colored dots on a map of Vienna, with the colors representing the delay rates of those stations. Next, we present line graphs showing the variation in punctuality rate and number of departures throughout the day, separated by weekdays and weekends.

3.4.1 Stations colored by punctuality rate

For the first visualization, depicted in fig. 4, all stations of the network were marked with a dot at their respective location on a map of Vienna. Then, stations were colored based on their individual punctuality rate, ranging from green for a rate of 100% to red representing a rate of 88.7%. Since the station *Blaasstraße* was already detected as an outlier in section 3.3, it is not included in this map as it would shift all other stations relatively far up on the color scale, resulting in an all-green map. At first glance, one can see a very good overall state, with a vast majority of stations colored light to dark green. Besides a few red spots, there is a noticeable sequence of orange and red stations, located in the northwest of the city center which all belong to bus line 42A. Besides these more delayed areas, the darkest green and thus most punctual sequences of stations can be seen on the edges of the network, most notably in the west and south of the city, in addition to the area east of the Danube River.

Since the previous map with all bus, tram and metro stations is rather dense, fig. 5 shows only tram and metro stations. On there, one can see two red dots, which are the stations *Marsanogasse* located north of the city center and *Gredlerstraße* which is right in the center. Besides these two stations, the most delayed tram lines are visible on the map because of their more yellow-colored stations. For example, line 71 goes from the center to the southeast and line 2 starts north of the city center and continues west. The most punctual regions this time

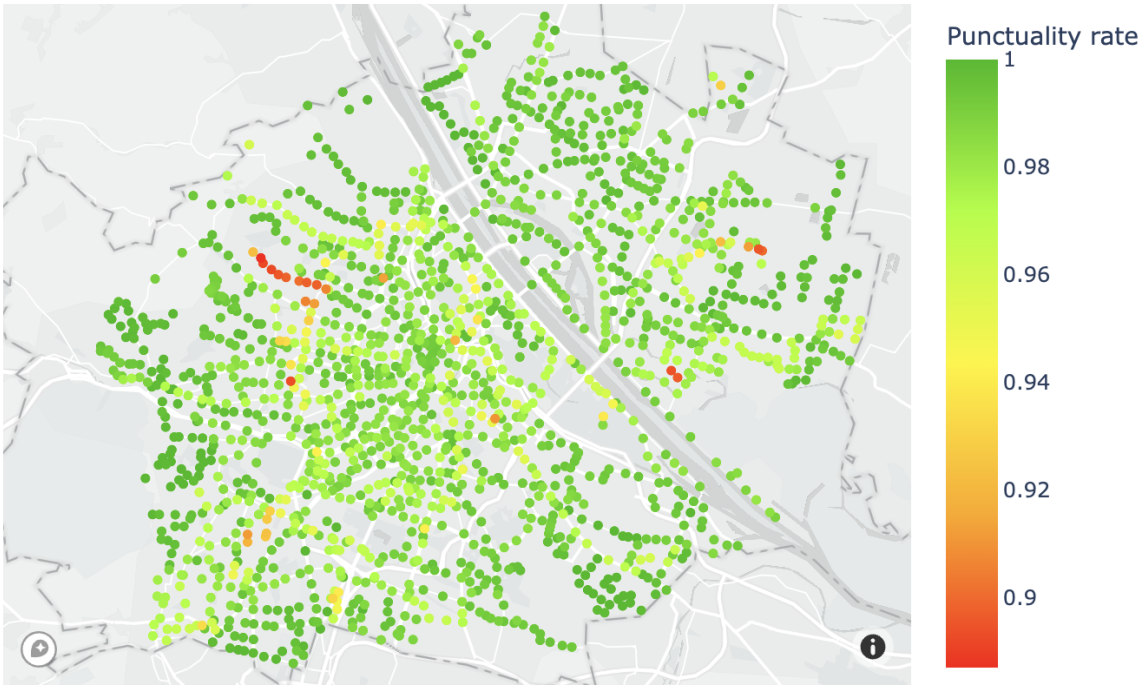

Figure 4: All stations colored by punctuality rate

are located in the west, which follows the fact that there are significantly more tram and metro stations in this area.

Figure 6 goes one step further and shows metro stations only, which allows an even more nuanced analysis. The map becomes once again significantly less densely dotted, with only 96 metro stations remaining. Almost all stations are green, confirming results from section 3.3 that metro lines generally have high punctuality rates. The two exceptions are the stations *Hausfeldstraße* and *Aspern Nord* located on the far-east, which are both part of metro line U2. Interestingly, the next station after the red-colored *Aspern Nord* shows a dark shade of green and therefore a high punctuality rate.

3.4.2 Punctuality rate by time of day

Until now, we have always analyzed all departures during the entire collection period of 30 days. However, another interesting aspect is the correlation of punctuality to the time of day. During rush hours in the morning and evening, higher delay rates are expected. In order to see this effect, all departures were grouped by the hour they were scheduled, ranging from 0 to 23. Furthermore, an additional separation between weekdays and weekends was done in order to capture the effect of rush hour traffic better. Figure 7 shows departures from Monday to Friday while fig. 8 shows departures that occurred on Saturdays and Sundays.

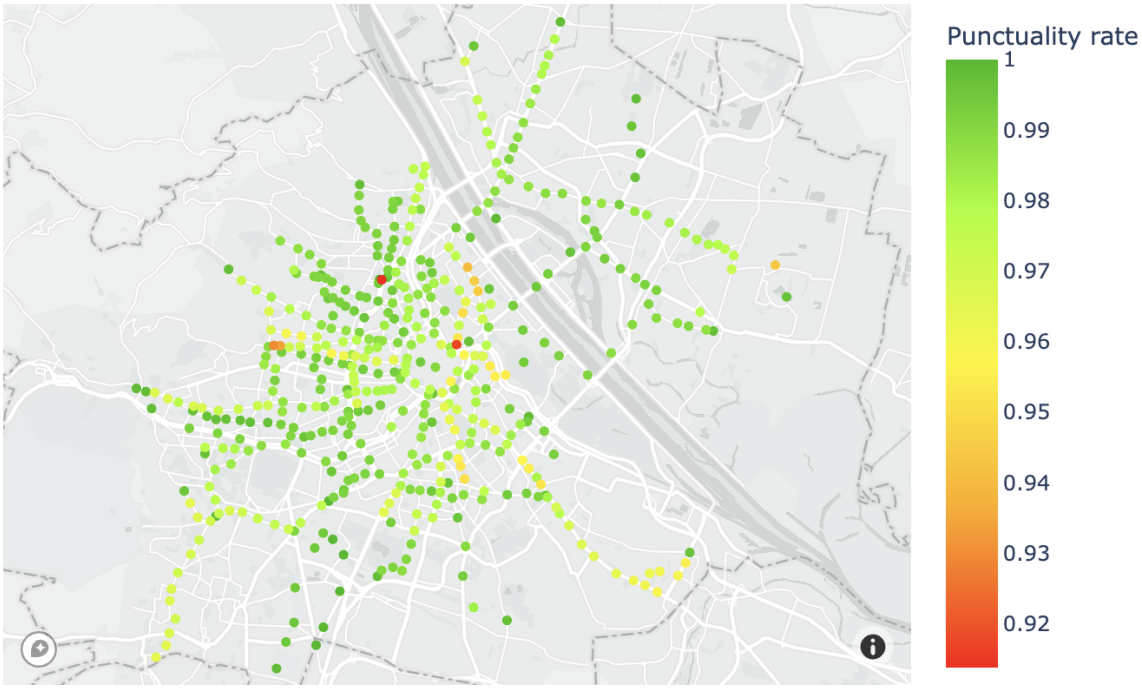

Figure 5: Metro and tram stations colored by punctuality rate

For weekdays, the punctuality rate drops significantly at 8am and 4pm. During this timeframe, the number of departures also rises to the highest points, peaking at 7am and 4pm with around 750,000 and 670,000 departures per hour respectively. While the amount of departures stays relatively high during midday, the punctuality rate recovers significantly. From 2am to 3am in the night, punctuality even reaches 100%. However, on weekdays there are only limited services with only a few bus lines running, shown by the number of departures which drop to around 8000 per hour. On weekends, the image differs a lot and shows that the number of departures rises during the morning and reaches its highest point at noon, going back down at 3pm. Further, those peaks are notably lower than on weekdays, reaching less than a third of departures with around 220,000. The punctuality rate behaves very differently too, with its lowest values being at 2am and 3pm, this time reaching its maximum at 7am.

4 Discussion

In this section, we will explore and discuss possible reasons for the results found in section 3.3 and section 3.4.

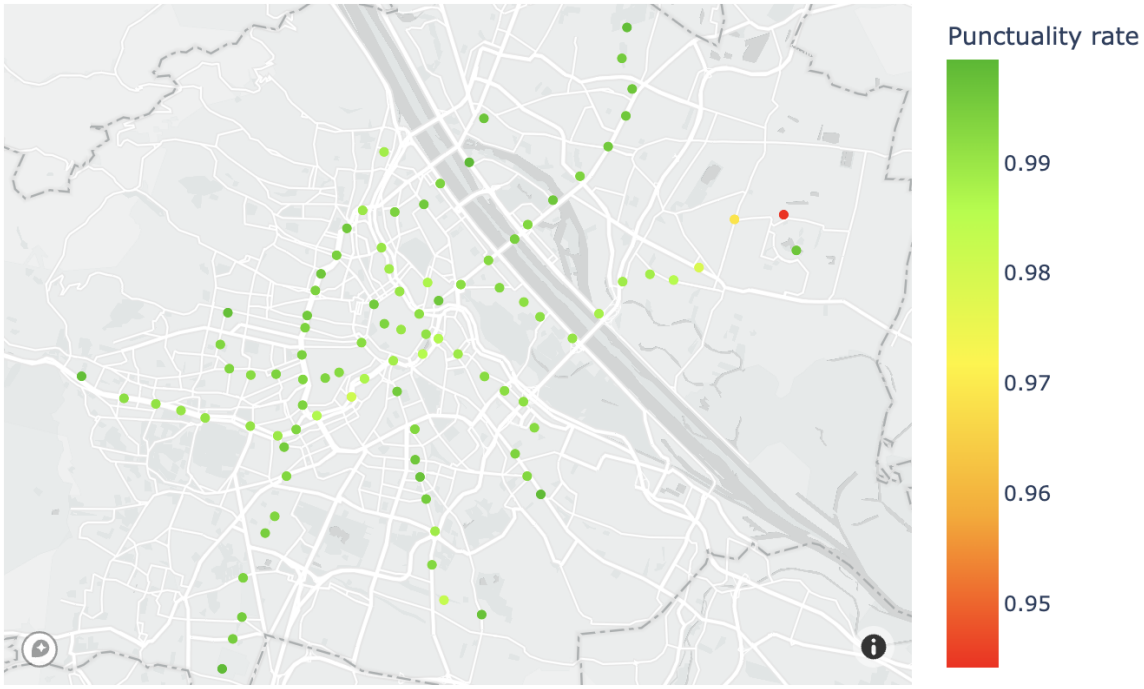

Figure 6: Metro stations colored by punctuality rate

4.1 Line and station overview

In section 3.3, we at first showed the most and least delayed lines in the network. While many lines had a perfect punctuality rate of 100%, most of those lines are either really short or have special routes where they face very little traffic. Line 2A for example is a line that has 10 stops within the city center where traffic is rather low due to the higher number of pedestrian areas. Line ZF is a special line that travels around the Central Cemetery and thus is not affected by any traffic at all. Lines in table 3 that start with the letter "N" are night lines, which again face very little traffic during the night. Because of those reasons, it is not surprising that these lines have punctuality rates of 100%. Lines 34A and 73A are also comparatively short lines, having 19 and 16 stations respectively. Additionally, many trips on those routes do not run all the way to the last station but rather end at an earlier one, resulting in trips with 10–15 stations, significantly less than the average bus route. The lines with the lowest punctuality rates all have relatively similar mean delays, ranging from 24.87 to 51.89 seconds. The least punctual line 42A, however, has a significantly higher mean delay of 149.01 seconds. Since this observation is also visible in one of the visualizations, it will be further discussed in section 4.3.

When looking only at metro lines, it became visible that all five lines have very high punctuality rates with the lowest value at 98.65%. Since metro lines are entirely separated from both external traffic and each other (i.e. the individual lines do not share any tracks), those high percentages are within expectations. Excluding U6E, which is a special replacement line during

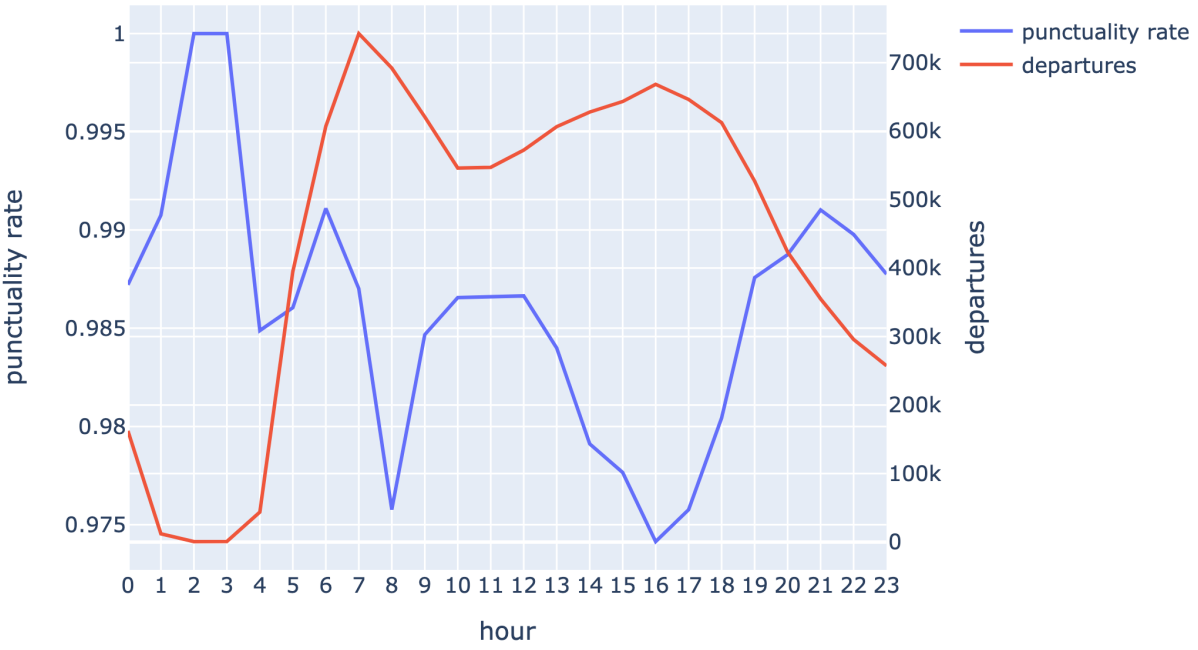

Figure 7: Punctuality rate and number of departures by time of day on weekdays

construction work, all metro lines showed similar mean delays of around 10 seconds, which indicates that there are no outliers, at least when looking at the lines as a whole.

Tram lines on the other hand showed more spread-out results, which seems obvious when considering that traffic can have great effects on how smoothly a tram can operate through narrow city streets. Interestingly, the data shows that the most punctual lines are lines with fewer stops than the least punctual lines which are all relatively long lines. The five most punctual tram lines have an average of 15.8 stops, while on the other hand, the five least punctual lines have an average of 31.8 stops, offering twice as many opportunities for delayed departures and therefore being a plausible cause for those higher delay rates. An interesting observation is that three out of the five most punctual tram lines have negative mean delays. This effect is likely caused by the terminal stations, where trams have scheduled dwell times of 5–10 minutes so that the next trip can start punctually and drivers have the opportunity for a small break. Even though the negative values only occur at the terminal stations, the mean delay of the entire line is also negative. This again is a consequence of the short length of those lines. Since this especially becomes visible when looking at each stop of one line, it will be further discussed in the next section.

Next, all departures were grouped by stations and their punctuality rates were aggregated. The results showed that there is one station with significantly more delays than others. While the station in question, *Blaasstraße*, is part of the second most delayed line 10A, at first sight, it is not affected by any special circumstances which could explain the drastically lower punctuality rate. This single outlier is also visible when looking at the box plots in fig. 3. Even when comparing the station to its direct neighbors on the line, namely *Dänenstraße* and *Hardtgasse*,

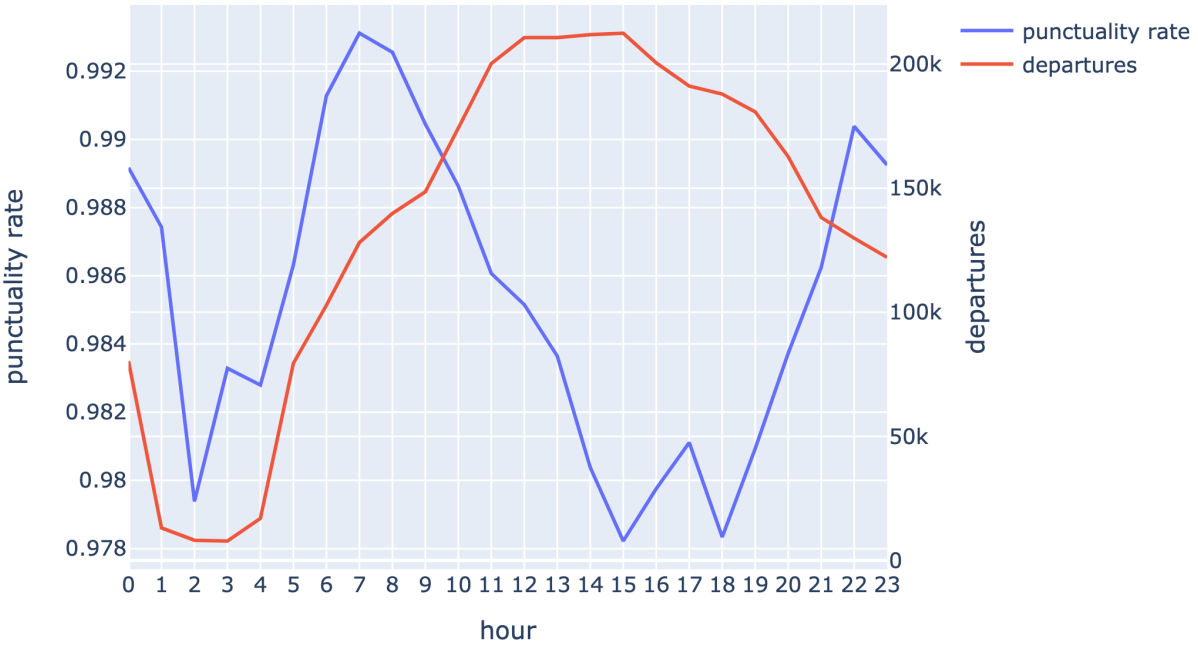

Figure 8: Punctuality rate and number of departures by time of day on weekends

the anomaly cannot be explained since those two stations show delay distributions in line with the overall trend. *Blaasstraße* however even has a median delay of 60 seconds, whereas one would expect the median as well as the 75th percentile to be 0, based on the rest of the data. To further confirm that there is no external reason for these high numbers, a local inspection and test rides on line 10A were carried out. Since this did not result in any new findings, a scheduling error or another non-obvious reason seems likely.

Looking only at stations that serve metro lines, one interesting observation is that out of the five stations with the most delays, shown in table 10, three have metro, tram and bus services and thus lots of transfers occurring. The five most punctual stations on the other hand, shown in table 9, all only serve exactly one metro line and buses, but no trams, which makes them less susceptible to delays happening due to large crowds interchanging between lines. Overall though, all of those stations have high punctuality rates, with the difference between the most and least delayed stations only being 2.8 percentage points, which the box plot in fig. 3 clearly demonstrates as well.

4.2 Tram line 71 and station *Karlsplatz*

After analyzing the collected data on a more general level by looking at all lines or stations at once, section 3.3.6 focused on one particular line, namely tram line 71. When looking at both punctuality rates and mean delays, one can observe that for lower punctuality rates, mean delay increases. However, for some stations such as *Ring/Volkstheater U* or *Oper/Karlsplatz U* this pattern does not hold, meaning that these stations have a higher mean delay than the respective

next ones in the ordered list. A plausible explanation for this is that both these stations are important interchange stations for metro lines (indicated by the letter "U" in the name). At such interchanges, it is more likely that delays occur due to big crowds entering and exiting the vehicles and thus increasing dwell time. The fact that these inflated numbers do not necessarily carry on to the next stations indicates that the network operator Wiener Linien has included this fact in their schedule planning. Regarding punctuality rates, it becomes obvious that delays gathered in the city center do accumulate to the end of the line. The most punctual stations at the top of table 11 are all within the first few stops in the center while stations with low punctuality rates are at the end of line 71. One eye-catching exception is the last station but very first entry, *Kaiserebersdorf, Zinnergasse*. While all other stations have relatively normal mean delays of 18 to 30 seconds, this station has a mean delay of -25.15 seconds. Trams in Vienna usually have a 5–10-minute stay at the terminal stations, enabling a short break for the driver and, as importantly, making up for accumulated delays and ensuring a punctual departure on the trip back. This explains the excellent punctuality rate for this particular station and likely other terminal stops as well. Since early departures are generally not desired though, there have to be more causes for this effect of negative departure delays. One of those possible causes could be that consistent intervals are preferred to perfectly following the schedule, resulting in drivers departing early for their next trip instead of waiting, which itself could cause an accumulation of waiting vehicles at the terminal station.

In addition to one single line, one station was analyzed in a similar manner, shown in table 12. The station of choice was *Karlsplatz*, as it is one of the busiest stations with metro, trams and bus lines stopping there. Metro lines again showed the highest punctuality rates and lowest mean delays. This again is most likely an effect of metros being entirely separated from other traffic, unlike trams and buses. The negative mean delay of -16.2 seconds for U2Z shows the same effect as the terminal stop of line 71 mentioned above. U2Z terminates at *Karlsplatz*, meaning that it has enough leeway scheduled to make up for possible delays. In general, though, trams show worse punctuality rates and mean delays than buses at *Karlsplatz*. While tram lines 71, D, 62 and 2 show consistent mean delays of around 30 seconds, line 1 has a significantly lower value with just 0.76 seconds. One possible explanation for this is that schedule planning is slightly more accurate for this line compared to the others mentioned.

4.3 Visualizations

In section 3.4, two types of visualizations were presented which allow for an intuitive understanding of the large amounts of data at hand. The first one showed a map of Vienna with all stations colored on a gradient from red to green, respective to their punctuality rate. A sequence of bus stops on line 42A was visibly more yellow than the rest of the network. Line 42A is a rather short line with twelve stops, none of which are located within the inner districts, where more traffic and passenger volume would be expected. While previous delays do accumulate towards the end of a line as seen in previous sections, it is interesting to see the effect this strongly at this line and leaves the question if there exist other reasons for this behavior.

The same visualization but with only metro and tram stations shown discovered that the stations *Gredlerstraße* and *Marsonogasse* were the two most delayed tram stations in the network. While the former is a regular station part of line 2 (one of the most delayed trams), the latter is a special station located at one of the tram depots, where vehicles only stop at the very beginning and end of service. In fig. 6, where only metro stations are shown, it became again apparent that delays that accumulate through the trip of a metro line are accounted for at the terminal stop by calculating time reserves accordingly. This pattern is visible at most terminal stops, where their immediate neighbors show a considerably lighter shade of green or even reaching to the end of the color scale like for the U2 station *Aspern Nord*, located in the far east of Vienna.

For the next visualization, all departures were grouped by the hour of the day and then plotted onto a line graph with both the mean delay and the number of departures visible. In order to see the potential effect of rush hour traffic more clearly, the departures were split into weekdays and weekends and visualized separately. This effect was indeed clearly visible, especially on weekdays when the punctuality rate was significantly lower at 8am and 4pm. On Saturdays and Sundays however, the image was very different, not showing any rush hours but with minimums at 2am and 3pm.

5 Conclusion

This section will conclude and summarize the results by highlighting key takeaways and putting them into the context of the scope and goals of this thesis. Finally, aspects that were not in the scope of this thesis and are subject to further research will be mentioned.

The aim of this thesis was to analyze the public transport network of Vienna for on-time performance and detect potential delay hotspots where punctuality rates are particularly low. Public transport is one of the key factors of a functioning livable city and good on-time performance is of high importance in order to ensure high usage and passenger satisfaction. In order to conduct this analysis, a custom data collection script was written which persisted 13.7 million departure records over a period of 30 days. Next, the collected data was cleaned up and filtered for incomplete or irrelevant results. For the following analysis, the data was either further filtered in order to highlight certain aspects, or it was aggregated and summarized as a whole.

Evaluating the data showed that the network offers a very good general on-time performance. While metro lines usually had very few delays due to their separation from other traffic, bus lines had both the highest and lowest punctuality rates, depending on their specific route. The initial assumption was that delays occur more often at busy interchange stations due to lots of passengers getting on and off the vehicles. This assumption did not hold in the general case, as the lowest punctuality rates were found at regular bus stations. On certain lines, however, this effect did become visible, like on tram line 71, which was analyzed in more detail in section 3.3.6. Other expected delay hotspots were terminal stations since those provide the opportunity for previously collected delays to accumulate. An interesting observation regarding this was that the terminal stops themselves surprisingly showed the best punctuality rates and even negative mean delays. However, the last 3–5 stops before them had in fact relatively low

punctuality rates, suggesting that previous delays do accumulate but are however accounted for by including a leeway in the schedule. Lastly, there were some stations and even parts of whole lines that showed substantially lower punctuality rates than others. In contrast to the aforementioned cases, these however did not show any apparent systemic reason for the bad on-time performances.

In addition to the analysis done in section 3.3, visualizations of the delay data were presented in section 3.4. The aim of those visualizations was to provide an intuitive understanding of the data. The first visualization was a city map with stations plotted at their locations. Each station was colored on a gradient ranging from green to red which represents its punctuality rate. With this, the already known results from section 3.3 were visualized as well as new delay hotspots discovered. Next, a line graph was presented which showed the number of departures and the punctuality rate per hour of the day, separated for weekdays and weekends. This showed the expected result that on-time performance drops significantly for morning and afternoon rush hour traffic during weekdays.

The presented statistics and graphics showed that there is definitely room for improvement in Vienna's public transport network, albeit the overall results suggest that there are already various measures put into practice to avoid long delays and with that unsatisfied passengers.

5.1 Future Work

There are many opportunities for further analysis of departure and in particular delay data, both in other cities and in Vienna. While the scope of this thesis was limited to departure delays, arrival delays could be analyzed and used in calculations combining arrival and departure times. For example, dwell times—another important measure in public transport analysis—could be calculated and analyzed. With dwell times, stations that actually cause delays could be detected better than when using only departure delays since one can see whether a delay was just carried on from the previous station (indicated by a normal, short dwell time) or a delay emerged at the station in question (indicated by a higher than expected dwell time). Of course, delays can also occur between stations, where this method would not be viable.

As mentioned in previous sections, another important indicator of on-time performance is the intervals between departures. When there are a lot of delays but consistent intervals, passengers are still satisfied if intervals are sufficiently short (Van Oort et al. 2015, 378). It could be studied how intervals differ from their schedule in order to detect possible delay hotspots. In combination with passenger surveys which could determine acceptable interval deviations, identified hotspots can be narrowed down to a few high-priority locations, where shorter intervals would result in the highest increase in customer satisfaction. Additionally, passenger volume should be taken into consideration. After all, very short intervals are only economically viable if they meet passenger demand.

In addition to the visualizations presented in this thesis, interactive visualizations could be explored in order to provide the reader with even more relevant data and a more engaging experience than a static graph or image. With location data, a live map showing the position of vehicles in real-time would be an interesting interactive application. Even without location data,

one could in theory calculate vehicle locations by multiplying the expected minutes or seconds until arrival by the average speed of the vehicle. Naturally, this method can only provide estimations, combined with smooth interpolation and animations this could however result in a useful, visually appealing visualization.

In order to improve the network and the situation for passengers, it is important that after the on-time performance has been studied, concrete measures are implemented that try to resolve identified problems. This can be done by both the transit company which can for example improve vehicles, driver behavior or station infrastructure and also by the city itself and its traffic planning, for example by reducing car lanes in order to avoid delays caused by traffic blocking buses or trams.

References

Caspari, Adam, Daniel Wood, Angel Campbell, Darian Jefferson, Tuan Huynh, and Alla Reddy. 2021. "Using Real-Time Data to Detect Delays and Improve Customer Communications at New York City Transit." *Transportation Research Record* 2675 (7): 45–57. https://doi.org/10.1177/0361198121994115.

Chen, Xumei, Lei Yu, Yushi Zhang, and Jifu Guo. 2009. "Analyzing urban bus service reliability at the stop, route, and network levels." *Transportation research part A: policy and practice* 43 (8): 722–734. https://doi.org/10.1016/j.tra.2009.07.006.

Computerwoche. 1988. "Neue Lösung auf der Basis von Tandem-Rechnern und Siemens-PCs: Bahn will offenen Rechner-Verbund schaffen" [in German]. *Computerwoche* 1988 (46). Accessed April 30, 2023. https://www.computerwoche.de/a/bahn-will-offenen-rechner-verbund-schaffen,1157261.

European Commission. 2020. *Report on the quality of life in European cities, 2020.* https://doi.org/10.2776/600407.

Goverde, Rob MP, Francesco Corman, and Andrea D'Ariano. 2013. "Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions." *Journal of rail transport planning & management* 3 (3): 78–94. https://doi.org/10.1016/j.jrtpm.2013.12.001.

Hallinan Jr, Arthur J. 1993. "A review of the Weibull distribution." *Journal of Quality Technology* 25 (2): 85–93. https://doi.org/10.1080/00224065.1993.11979431.

Haslauer, Eva, Elizabeth C Delmelle, Alexander Keul, Thomas Blaschke, and Thomas Prinz. 2015. "Comparing subjective and objective quality of life criteria: A case study of green space and public transport in Vienna, Austria." *Social Indicators Research* 124:911–927. https://doi.org/10.1007/s11205-014-0810-8.

MobilityData. 2022a. *GTFS Realtime Reference.* Accessed April 29, 2023. https://gtfs.org/realtime/reference/.

———. 2022b. *GTFS Schedule Reference.* Accessed April 29, 2023. https://gtfs.org/schedule/reference/.

Redmann, Jannis. 2023a. *friendly public transport format.* Accessed April 30, 2023. https://github.com/public-transport/friendly-public-transport-format/blob/3bd36faa721e85d9f5ca58fb0f38cdbedb87bbca/spec/readme.md.

———. 2023b. *hafas-client: A client for the "mobile APIs" of HAFAS public transport management systems.* Accessed April 30, 2023. https://github.com/public-transport/hafas-client#readme.

———. 2023c. *hafas-client API.* Accessed April 30, 2023. https://github.com/public-transport/hafas-client/blob/master/docs/api.md.

Stadt Wien. 2022. *Statistisches Jahrbuch der Stadt Wien 2022: Wien in Zahlen* [in German]. 338. Magistrat der Stadt Wien Wirtschaft, Arbeit und Statistik. Accessed April 29, 2023. https://www.digital.wienbibliothek.at/wbrup/download/pdf/4353659.

———. 2021. *U2-Teilsperre von Karlsplatz bis Rathaus bis Herbst 2023* [in German]. Accessed April 29, 2023. https://www.wien.gv.at/verkehr/oeffentlich/u2-sperre-karlsplatz-rathaus.html.

Strathman, James G, and Janet R Hopper. 1993. "Empirical analysis of bus transit on-time performance." *Transportation Research Part A: Policy and Practice* 27 (2): 93–100. https://doi.org/10.1016/0965-8564(93)90065-s.

Van Oort, Niels, Daniel Sparing, Ties Brands, and Rob MP Goverde. 2015. "Data driven improvements in public transport: the Dutch example." *Public transport* 7:369–389. https://doi.org/10.1007/s12469-015-0114-7.

Ware, Colin. 2019. *Information visualization: perception for design.* 4th ed. Morgan Kaufmann. ISBN: 9780128128756.

Wiener Linien. 2023. *Gesamtnetzplan.* Accessed April 29, 2023. https://www.wienerlinien.at/documents/843721/4763236/Tagnetz_2023-02_v1_Website.pdf.

———. 2020. *Zahlen und Fakten: Betriebsangaben 2019.* Accessed April 29, 2023. https://www.wienerlinien.at/media/files/2020/wl_betriebsangaben_2019_deutsch_358274.pdf.

Wiener Stadtwerke Group. 2022. *Working for the city: Financial Report 2021.* Accessed April 29, 2023. https://www.wienerstadtwerke.at/documents/238130/1547172/WStW_Financial_Report_2021.pdf.

Yuan, Jianxin. 2008. "Statistical Analysis of Train Delays." In *Railway Timetable & Traffic: Analysis, Modelling, Simulation*, edited by Ingo Arne Hansen and Jörn Pachl. Eurail Press. ISBN: 9783777103716.

———. 2006. "Stochastic modelling of train delays and delay propagation in stations." PhD diss., TU Delft. Accessed April 29, 2023. https://repository.tudelft.nl/islandora/object/uuid:caa72522-26b1-4088-afc0-59c6e5c346f6.

ÖBB-Personenverkehr AG. 2023a. *Pünktlichkeitswerte Wien* [in German]. Accessed April 29, 2023. https://www.oebb.at/de/rechtliches/puenktlichkeit/wien.

———. 2023b. *S-Bahn Wien, Niederösterreich und Burgenland* [in German]. Accessed April 29, 2023. https://www.oebb.at/de/regionale-angebote/wien/s-bahn-wien.

Appendices

A git-Repository

According to the respective guidelines.
The repository must be uploaded to the MMT/HCI git server gitlab.mediacube.at
https://gitlab.mediacube.at/fhs45907/bachelor-thesis

This work has the following word count (counted by texcount):

```
File: body.tex
Encoding: utf8
Sum count: 9259
Words in text: 9021
Words in headers: 80
Words outside text (captions, etc.): 158
Number of headers: 32
Number of floats/tables/figures: 20
Number of math inlines: 0
Number of math displayed: 0
Subcounts:
  text+headers+captions (#headers/#floats/#inlines/#displayed)
  1+0+0 (0/0/0/0) _top_
  479+1+0 (1/0/0/0) Section: Introduction
  0+2+0 (1/0/0/0) Section: Related Work
  438+2+23 (1/2/0/0) Subsection: The Hague
  414+1+0 (1/0/0/0) Subsection: Portland
  421+3+0 (1/0/0/0) Subsection: New York City
  149+2+0 (1/0/0/0) Subsection: Stochastic models
  672+4+0 (3/0/0/0) Subsection: Data availability
  82+1+0 (1/0/0/0) Section: Implementation
  215+4+0 (1/0/0/0) Subsection: Vienna's public transport network
  743+8+8 (4/0/0/0) Subsection: Data collection
  1794+25+83 (8/13/0/0) Subsection: Delay analysis
  833+12+44 (3/5/0/0) Subsection: Visualizations
  20+1+0 (1/0/0/0) Section: Discussion
  841+4+0 (1/0/0/0) Subsection: Line and station overview
  545+6+0 (1/0/0/0) Subsection: Tram line 71 and station \textit{Karlspla
  388+1+0 (1/0/0/0) Subsection: Visualizations
  558+1+0 (1/0/0/0) Section: Conclusion
  428+2+0 (1/0/0/0) Subsection: Future Work
```