Comparison of target and actual travel times

As part of Open Government Data (OGD), the City of Zurich provides data records from public administrations for the general public in digital form. This task is now about analyzing deviations from the target departure time "soll" in the public transportation network ("VBZ"). The actual departure time is called "ist". The data, all in csv format, can be found at: https://data.stadt-zuerich.ch/dataset/vbz_fahrzeiten_ogd

You need the following four data files to process the tasks:

- fahrzeiten_soll_ist_20191020_20191026.csv
- fahrzeiten_soll_ist_20191027_20191102.csv
- fahrzeiten_soll_ist_20191103_20191109.csv
- fahrzeiten_soll_ist_20191110_20191116.csv

as well as the following tables with stops:

- Haltestelle.csv (name of the stop points)
- Haltepunkt.csv (GPS location of the stop points)

The link between the files is illustrated on the website in the comments section with a database schema and the explanations for the individual variables can be found in the attributes section further down on the website. (Github: https://github.com/opendatazurich)

Make sure that the illustrations are created in such a way that someone with mediocre statistics knowledge can interpret them correctly without any context knowledge.

Use the packages **ggplot2** and **ggmap** and **tmaptools** to generate the images. For reading in and preprocessing the data you can e.g. use the **readr**, **dplyr**, **stringr**, and **tidyr** packages.

a) *Reading in and filtering the data:* Read the four data files with the travel times into Rstudio and filter them according to the following criteria:

   • We only want to examine trips on line 7 (in both directions).

   • We are only interested in trips from Monday to Friday ("Montag", "Freitag"). Tip: Convert the operating days to timestamps with **as.POSIXct()** and use the built-in function **weekdays()**.

   • There are different routes, so-called "Fahrtwege"; these are differentiated according to the direction of travel. We want to exclude depot journeys, detours etc., therefore filter the data so that only the most frequent route remains in direction 1 and 2 in the data record.

b) Draw the stops of line 7 in different colors per direction on a map of the city of Zurich. Label the stops on the map with their names, but without the characters 'Zürich, ' or 'Zch, '.

The R functions **merge()** (built-in) or **left_join()** (from dplyr) and **unique()** could be helpful, but do not have to be used.

c) Visualize the distribution of departure delays (actual from - target from) ("Ist ab - Soll ab") for each stop for both directions of travel. Make sure that these distributions are plotted according to the station stop order. What did you notice?

d) The Billoweg stop is one of the last in direction 1 or one of the first in direction 2. One could therefore assume that the departure delays at this stop in direction 1 are systematically greater than in direction 2. Filter the data again:

- consider only the Billoweg stop.
- create a new variable that specifies the hour. You can divide and round off the target departure times given in seconds by 3600.
- Only consider trips between 9:00 and 15:00.
- Take only one trip per hour in each direction. You can do this with **dplyr**, for example, by using **summarize()** with **sample()** or **first()**.

Run an appropriate test or calculate a confidence interval. Justify the choice of method.

e) In d) we only looked at one trip per hour, which of course means using only part of the available data. Why does that make sense? What would be problematic when using your chosen method if you use all trips?

English – German (for the csv files):

Target: Soll
Actual: Ist
Line: Linie
Direction: Richtung
Monday: Montag
Friday: Freitag
Date: Datum
Stop: Haltestelle
Location of the stop: Haltepunkt
Name of the stop: Halt_lang
Short name of the stop: halt_kurz