Universität Stuttgart

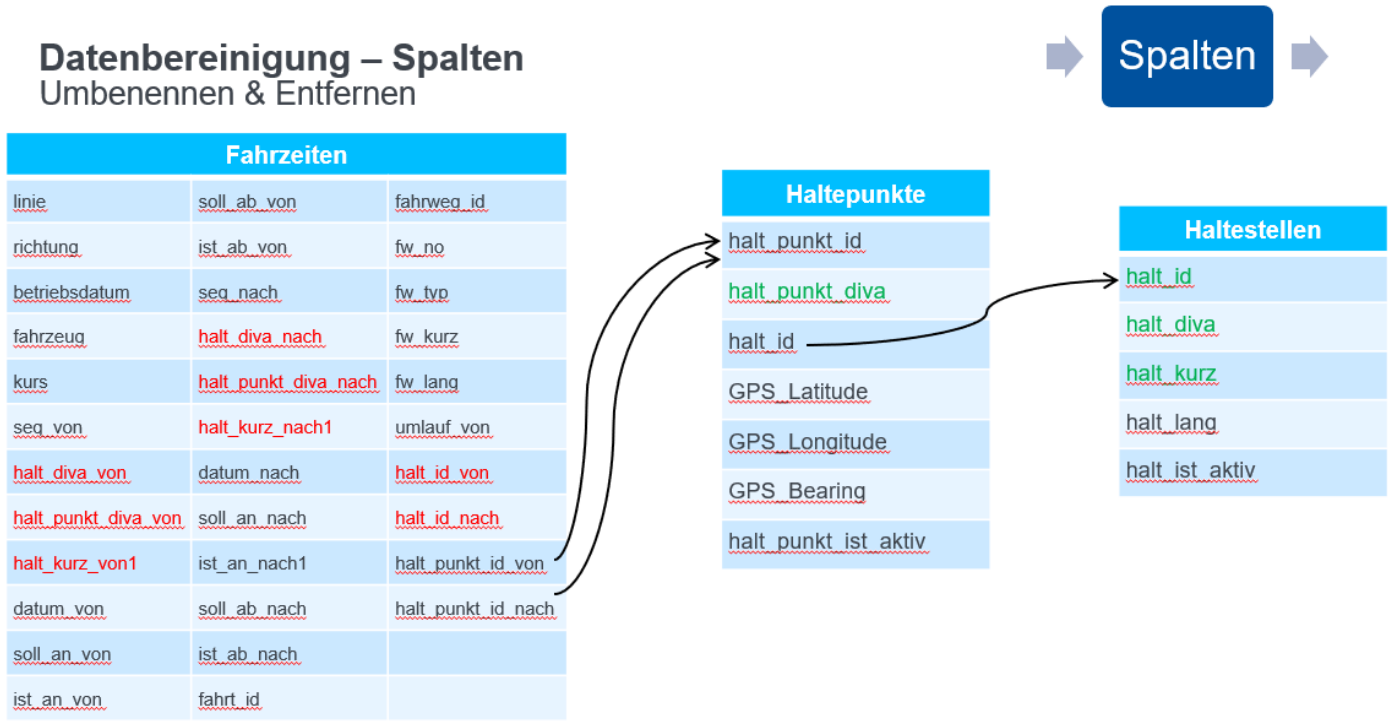
**Projekt Data Science**
**Analyse von Mobilitätsdaten**

# Hypothesen

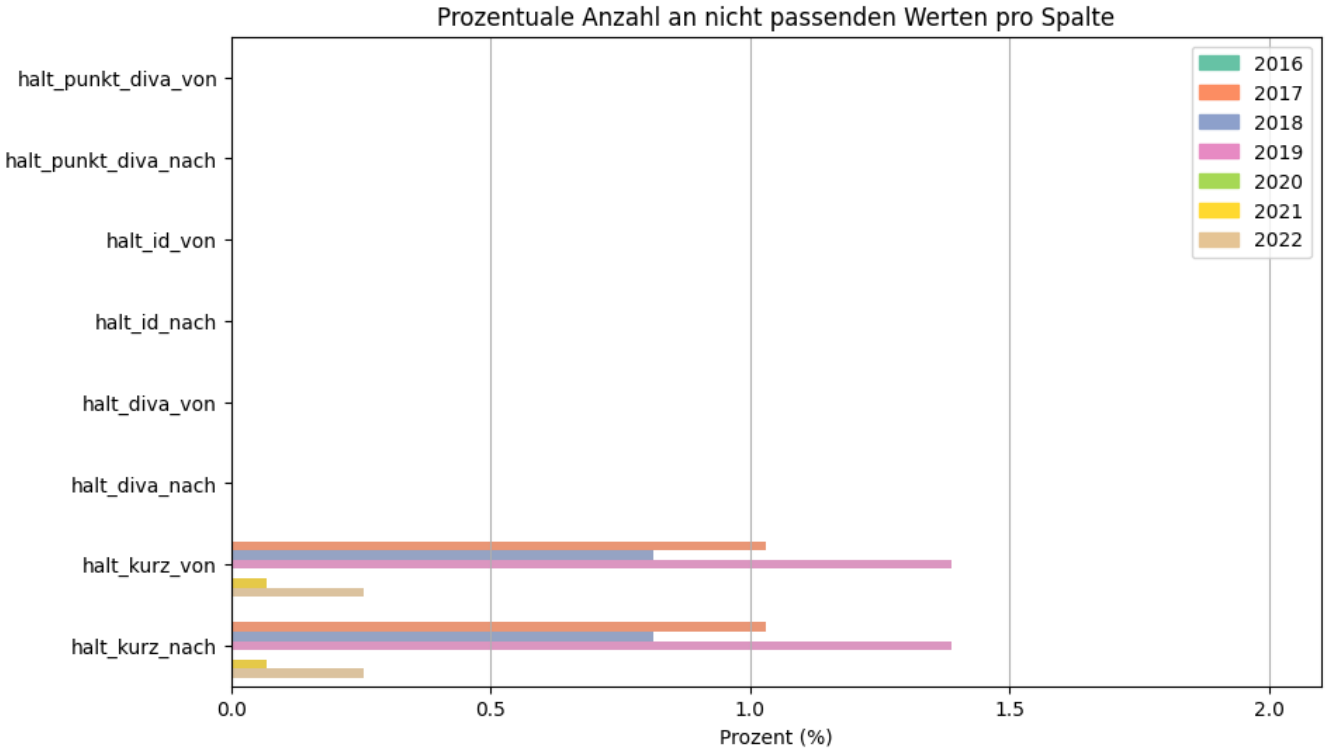
Wintersemester 2023/24

Gruppe 02 – Ozan Tastekin & Tony Klasan

# Letztes Mal bei…
Datenbereinigung



# Letztes Mal bei…
## Datenbereinigung



# Hypothese 01/08 - Ozan

Schwierigkeit: ★★★★☆

**Eine erhöhte Passagierfrequenz an einer Haltestelle führt zu einer Zunahme von Verspätungen.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022
- Passagierfrequenzen in Zürich[2]
  - Ein- und Aussteiger Zählungen an Bahnhöfen in Zürich
  - 2014-2022

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

[2] https://data.stadt-zuerich.ch/dataset/vbz_fahrgastzahlen_ogd

## Annahmekriterium

Korrelationsanalyse z.B. Pearson-Korrelation (oder Spearman-Korrelation).

Wert zwischen -1 und +1 => Annahmewert: $\geq 0{,}6$

## Erklärung

- -1 = negative, 0 = keine, +1 = positive Korrelation
- 0,5-0,7 = hoher Zusammenhang

Um Hypothese anzunehmen, wurde 0,6 gewählt, was einen hohen Zusammenhang aufweist

# Hypothese 02/08 - Tony

Schwierigkeit: ★★★☆☆

**Fahrplanänderungen beeinflussen die Verspätungen positiv.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022
- Fahrpläne vom Züricher Netz
  - 2016-2022

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

## Annahmekriterium

Statistischer Test z.B. ANOVA (Kruskal-Wallis-Test)

Wert zwischen 0 und 1 => Annahmewert $\geq 0{,}05$

## Erklärung

- $p < 0{,}01$ – starke Evidenz gegen die Nullhypothese
- $p > 0{,}10$ – keine Evidenz gegen die Nullhypothese

Die Hypothese wird angenommen wenn der p-Wert mehr als 0,05 ist, was eine starke Evidenz aufweist

# Hypothese 03/08 - Ozan

Schwierigkeit: ★★★★☆

**Die Verspätung von Bahnen korreliert mit den Wetterbedingungen.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022
- Historische Wetterdaten in Zürich[2]
  - 3 Standorte der Messung in Zürich
  - Temperatur, Regendauer, Luftdruck, Luftfeuchtigkeit, etc.
  - 2000-2023

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

[2] https://data.stadt-zuerich.ch/dataset/ugz_meteodaten_tagesmittelwerte

## Annahmekriterium

Korrelationsanalyse z.B. Pearson-Korrelation (oder Spearman-Korrelation).

Wert zwischen -1 und +1 => Annahmewert: $\geq 0{,}6$ oder $\leq -0{,}6$

## Erklärung

- -1 = negative, 0 = keine, +1 = positive Korrelation
- 0,5-0,7 = hoher Zusammenhang

Um Hypothese anzunehmen, wurde 0,6 oder -0,6 gewählt, was einen hohen Zusammenhang in positive bzw. negative Richtung aufweist

# Hypothese 04/08 - Tony

Schwierigkeit: ★★☆☆☆

**Während Rush Hours gibt es mehr Verspätungen.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

## Annahmekriterium

Statistischer Test z.B. T-Test (Mann-Whitney-U-Test)

Wert zwischen 0 und 1 => Annahmewert ≥ 0,05

## Erklärung

- $p < 0,01$ – starke Evidenz gegen die Nullhypothese
- $p > 0,10$ – keine Evidenz gegen die Nullhypothese

Die Hypothese wird angenommen wenn der p-Wert mehr als 0,05 ist, was eine starke Evidenz aufweist

# Hypothese 05/08 - Ozan

Schwierigkeit: ★★☆☆☆

**Die Pünktlichkeit von Trams ist im Vergleich zu Bussen signifikant höher.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022

## Annahmekriterium

Statistischer Test z.B. T-Test Analyse für Signifikanzniveau α

Wert zwischen 0 und 1 => Annahmewert: ≤ 0,05

## Erklärung

In der Forschung wird oft der Wert 0,05 für α verwendet, was auch hier der Wert für die Annahme der Hypothese sein wird

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

# Hypothese 06/08 - Tony

Schwierigkeit: ★★☆☆☆

**Die Richtung einer Linie beeinflusst die Verspätungen der Linie negativ.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

## Annahmekriterium

Regressionsanalyse

Wert zwischen 0 und 1 => Annahmewert ≥ 0,05

## Erklärung

- $p < 0{,}01$ – starke Evidenz gegen die Nullhypothese
- $p > 0{,}10$ – keine Evidenz gegen die Nullhypothese

Die Hypothese wird angenommen wenn der p-Wert mehr als 0,05 ist, was eine starke Evidenz aufweist

# Hypothese 07/08 - Ozan

Schwierigkeit: ★★★★☆

**Das Wetter beeinflusst, wie viele Passagiere mit dem öffentlichen Verkehr fahren in [Zürich/Paris] signifikant.**

## Welche Daten werden benötigt?

- Historische Wetterdaten in Zürich[1]
  - 3 Standorte der Messung in Zürich
  - Temperatur, Regendauer, Luftdruck, Luftfeuchtigkeit, etc.
  - 2000-2023
- Passagierfrequenzen in Zürich[2]
  - Ein- und Aussteiger Zählungen an Bahnhöfen in Zürich
  - 2014-2022

[1] https://data.stadt-zuerich.ch/dataset/ugz_meteodaten_tagesmittelwerte

[2] https://data.stadt-zuerich.ch/dataset/vbz_fahrgastzahlen_ogd

## Annahmekriterium

Korrelationsanalyse z.B. Pearson-Korrelation (oder Spearman-Korrelation).

Wert zwischen -1 und +1 => Annahmewert: $\geq 0{,}6$ oder $\leq -0{,}6$

## Erklärung

- -1 = negative, 0 = keine, +1 = positive Korrelation
- 0,5-0,7 = hoher Zusammenhang

Um Hypothese anzunehmen, wurde 0,6 oder -0,6 gewählt, was einen hohen Zusammenhang in positive bzw. negative Richtung aufweist

# Hypothese 08/08 - Tony

Schwierigkeit: ★★★★☆

**Der öffentliche Verkehr in Zürich hat weniger Verspätungen als der in Brüssel.**

## Welche Daten werden benötigt?

- Fahrzeiten SOLL und IST Vergleich in Zürich[1]
  - Basisdatensatz
  - 2016-2022
- Fahrzeiten SOLL und IST Vergleich in Belgien[2]
  - Monatliche Verspätungen in ganz Belgien

## Annahmekriterium

Statistischer Test z.B. T-Test (Mann-Whitney-U-Test)

Wert zwischen 0 und 1 => Annahmewert ≥ 0,05

## Erklärung

- $p < 0,01$ – starke Evidenz gegen die Nullhypothese
- $p > 0,10$ – keine Evidenz gegen die Nullhypothese

Die Hypothese wird angenommen wenn der p-Wert mehr als 0,05 ist, was eine starke Evidenz aufweist

[1] https://data.europa.eu/data/datasets/878a98b8-4973-4d76-858e-eddd88652d9f-stadt-zurich

[2] https://data.europa.eu/data/datasets/https-opendata-infrabel-be-explore-dataset-stiptheid-per-type-trein-en-per-moment-?locale=de
https://data.europa.eu/data/datasets/https-opendata-infrabel-be-explore-dataset-data_punctualite_typedetrain-?locale=de

# Weitere Ideen/Probleme

- Ob ein Gesetz/Änderung etwas an den Verspätungen geändert hat
  - Problem: Nichts relevantes gefunden auf Züricher Nachrichten Seite von 2017-2022
- Bei Haltestellen, wo es in der Nähe alternative Verkehrsmittel (Leihrad, Sharepoints, etc.) steigen weniger Menschen ein/mehr Menschen aus.
  - Problem: Keinen passenden Datensatz gefunden dafür

- Größtes Problem:
  - Hypothesen, die den Basisdatensatz beinhalten müssen, gehen immer nur um Verspätungen
  - Basisdatensatz:
    - Fahrzeiten SOLL und IST in Zürich
      - Problem: Informationsgehalt ist eigentlich nur: Verspätungszeiten
    - Passagierfrequenz in Fernzugbahnhöfen in der Schweiz
      - Problem: Nur auf Jahresebene (einen) Wert, nur 2018 und 2022
      - => Entweder zu wenige Jahre oder zu grobe Zeitebene

# Weitere gefundene (brauchbare) Daten

- **Zürich**
  - Standorte von Leihfahrrädern (ZüriVelo) (Zu der Hypothese von vorher)
    - => Problem: Daten sind ganz aktuell, d.h. keine Information, welche Leihfahrradstände es seit wann gibt
- **Belgien**
  - Verteilung in % der Verantwortlichen Parteien bei Verspätungen (bestimmte Bahngesellschaft, dritte, etc.)
  - Gründe, weswegen es in einem Jahr ≥ 1.000 Minuten Verspätungen gab
- **Frankreich**
  - Passagierfrequenzen in Fernverkehrsbahnhöfen in Frankreich
    - Auch auf Jahresebene, wie Basisdatensatz, aber von 2015-2022
  - Monatliche Verspätungen/Ausfälle pro Region in Frankreich (2013-2023)
  - Befragungen in 2010-2017 => Verteilung in % für…
    - Reisegründe für Reisende und nicht Reisende
    - Kunden nach Alter
    - Kunden nach Häufigkeit der Verwendung vom öffentlichen Verkehr

# Weitere gefundene (brauchbare) Daten

- **Spanien**
  - Passagierfrequenzen in verschiedensten Städten in Spanien (Bilbao, Barcelona, Sevilla, Valencia, etc.)
    - Alle 30 Minuten => Einsteiger und Aussteiger an einer Haltestelle
    - Nur 2018
  - Bahnhofsdaten mit GPS

# Zeitaufwand
## Hypothesenaufstellung

- Ozan: ~74 Stunden + ~ 20 Stunden von vorherigen Meilensteinen, die ausgelassen wurden
  - Feedback bearbeitet
  - Daten gesucht
  - Hypothesen erstellt
  - Präsentationsfolien erstellt

- Tony: ~32 Stunden
  - Hypothesen erstellt

- Hypothesenaufstellung: ~106 Stunden + ~ 20 Stunden
- Insgesamt: Ozan: ~232 Stunden Tony: ~94 Stunden