**Predicting waiting times at Disney World**

* Need to decide if do for California or Florida
* Need to decide what attraction to do it for
  + Animal Kingdom, Epcot?

The seven parks in Florida: <https://www.wdwinfo.com/maps/index.htm>

Magic Kingdom Park:

* Most popular part of Disney World, original theme park
* 57’000 visitors per day (vs. EPCOT 34’000) (<https://www.magicalguides.com/how-many-people-go-to-disney-world/> )
  + Most popular ride within Magic Kingdom: **Haunted House**
    - Have data on it from 2015-2021
      * A lot of irregular time stamps (sometimes even 10 min or less)
    - Have predicted waittime already (<https://touringplans.com/magic-kingdom/attractions/haunted-mansion/wait-times> )
    - a lot of featues in Metadata

Mein Favorit für die Attraktion:

Haunted House im Magic Kingdom

Das Magic Kingdom ist mit 57’000 Besuchern pro Tag der meist besuchte Park

Wir haben viele features zum Magic Kingdom in den Metadaten

Das Haunted House ist der beliebteste ride im Magic Kingdom

Wir haben Wartezeiten dazu von 2015-2021

Ausserdem können wir unsere prediction mit den predictions auf dieser Website vergleichen (<https://touringplans.com/magic-kingdom/attractions/haunted-mansion/wait-times> )

Metadata:

Shape: (2079, 181)

Columns:

* Date, day of weekn month, year, season, time of sunset etc.
* Price: WDW\_TICKET\_SEASON
* Holidays:
* Temperature, Weather, Precipitation
* Special events (parades, fireworks etc.)
* Schools in session, within distance to Florida/California
* Opening hours (yesterday, today, tomorrow) and Closing hours
* Days to next event or since last event
* Capacity lost on park day

Have data from: 2015 -2021

Actual number of visitors do we have? No, don’t need them. Have wait time data in individual data for the attractions.

What we want to predict: a number -> regression

* OLS (Ordinary Leas Squares) used as statistical model
  + MSE loss function
  + Compute MAE and R^2
* PCA as additional model
* Do SVD if have a lot of data to reduce data
* Clustering

To do:

* Decide on Florida / California and on what attractions we want to do
* Metadata cleaning: only keep necessary columns
* Merge metadata and dataset for chosen attraction
* Start modelling

Questions:

* Data from six years enough?
  + Alte Jahre vielleicht weglassen -> Data set variieren
* Which model to do? Start with OLS?
  + Can give negative values
  + Yes, start with OLS
  + Special OLS models for waiting times
* SVD supposed to use to reduce data?
* Time series
  + Ja als Vergleich zu OLS
  + Als \* noch zu cross validation tun
* Random forest, booster trees anschauen
* Alte Jahre vielleicht weglassen -> Data set variieren
* Use case ist wichtig!
  + Time series hilft

Corona zahlen weg lassen? Waren komplett geschlossen? Grenzen?