Data Science 2

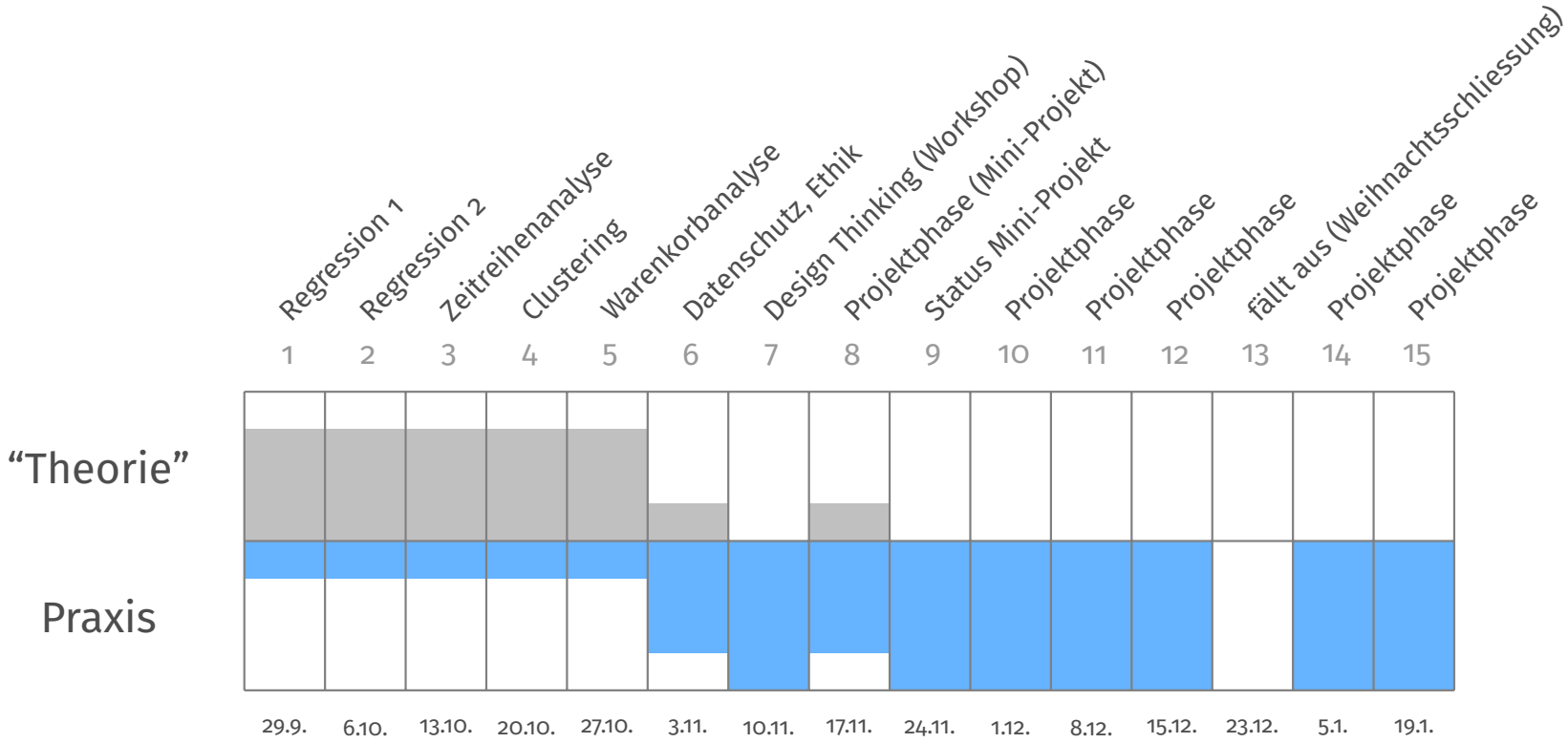
Projektphase

Prof. Dr. Christian Bockermann

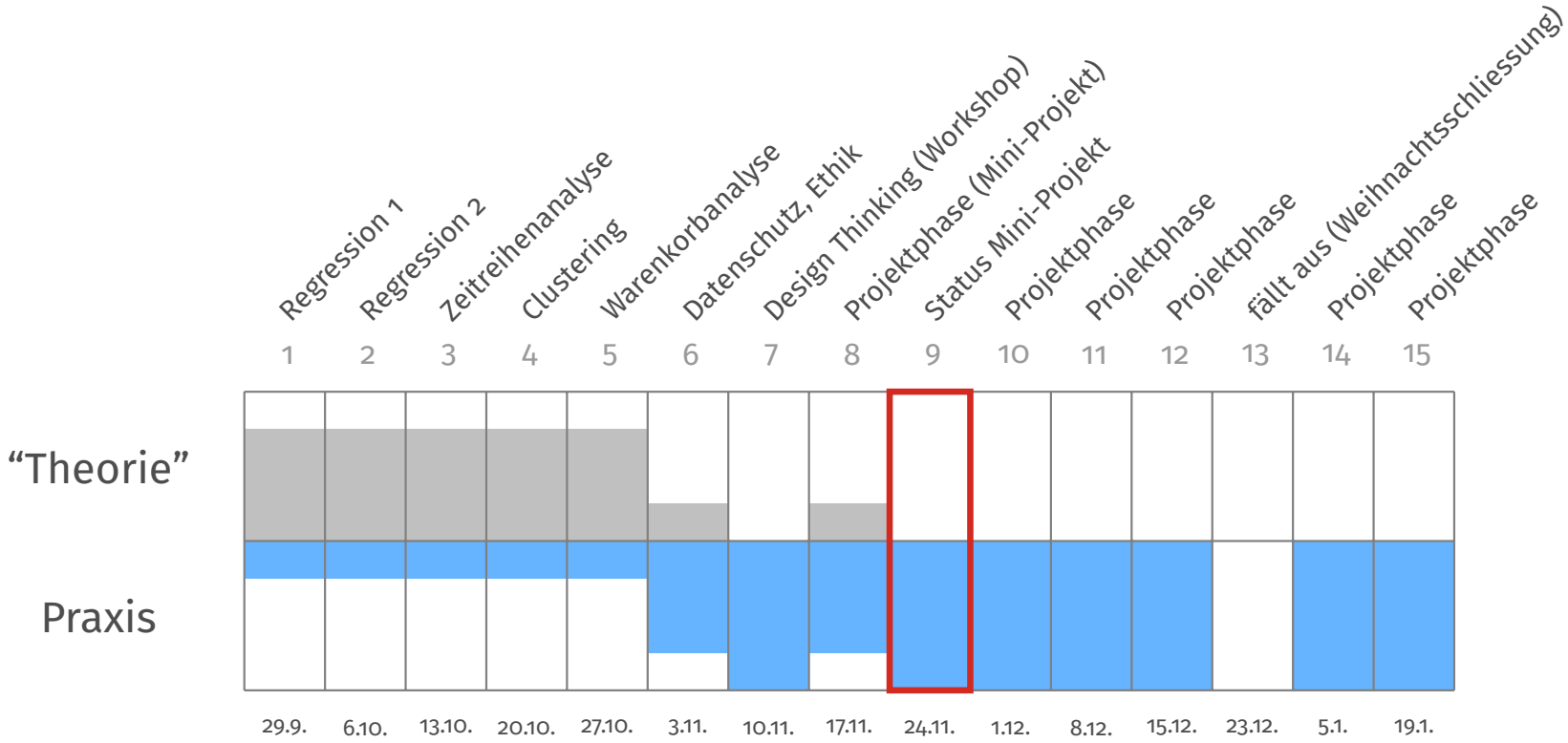
Hochschule Bochum

Wintersemester 2022 / 2023

Zeitplan


Regression 1
Regression 2
Zeitreihenanalyse
Clustering
Warenkorbanalyse
Datenschutz, Ethik
Design Thinking (Workshop)
Projektphase (Mini-Projekt)
Status Mini-Projekt
Projektphase
Projektphase
Projektphase
fällt aus (Weihnachtsschliessung)
Projektphase
Projektphase
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
“Theorie”
Praxis
29.9.
6.10.
13.10.
20.10.
27.10.
3.11.
10.11.
17.11.
24.11.
1.12.
8.12.
15.12.
23.12.
5.1.
19.1.

Zeitplan


Regression 1
Regression 2
Zeitreihenanalyse
Clustering
Warenkorbanalyse
Datenschutz, Ethik
Design Thinking (Workshop)
Projektphase (Mini-Projekt)
Status Mini-Projekt
Projektphase
Projektphase
Projektphase
fällt aus (Weihnachtsschliessung)
Projektphase
Projektphase
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
“Theorie”
Praxis
29.9.
6.10.
13.10.
20.10.
27.10.
3.11.
10.11.
17.11.
24.11.
1.12.
8.12.
15.12.
23.12.
5.1.
19.1.

Agenda

Heute:

- Präsentation der Titanic Challenge
- Feedback-Runde - Was lief gut? Was nicht?
- Vorstellung möglicher Datensätze

Danach:

- Themenauswahl
- Gruppenfindung

Projektphase – Teil 2

Was ist das Ziel?

- Eigenständig Datenanalyse “erarbeiten”
- Wirtschaftliche Fragestellung überlegen
- Wissenschaftliche Fragestellung ableiten, Lernaufgabe(n) formulieren/anpassen
- Datensatz explorieren
- Daten analysieren und eigene Analyse bewerten

Organisation

- selbstständiges Arbeiten in Kleingruppen (3-4)
- möglichst WiInf'ler und BWL/VWLer gemischt
- Abgabe als Jupyter-Notebook/Blog-Eintrag pro Person
- Präsentation als Gruppe

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil
- Problem: Individuelle Prüfungsleistung?

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil
- Problem: Individuelle Prüfungsleistung?

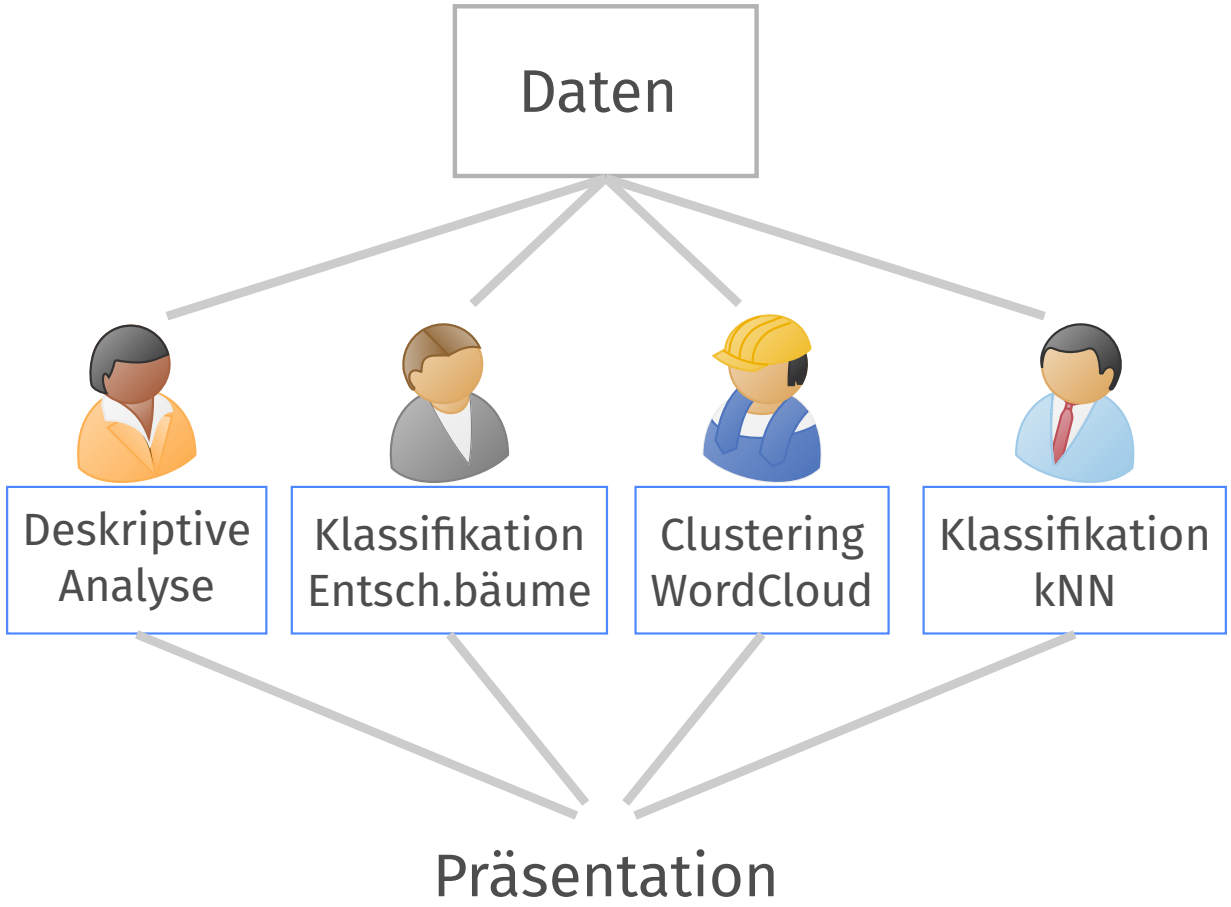
Modus B

- Freie Wahl des Datensatzes
- Je Gruppe unterschiedliche Fragen/Aspekte
- Jeder Teilnehmer bearbeitet u.a. eigenständigen Aspekt
- Gemeinsame Präsentation der Analyse des Datensatzes

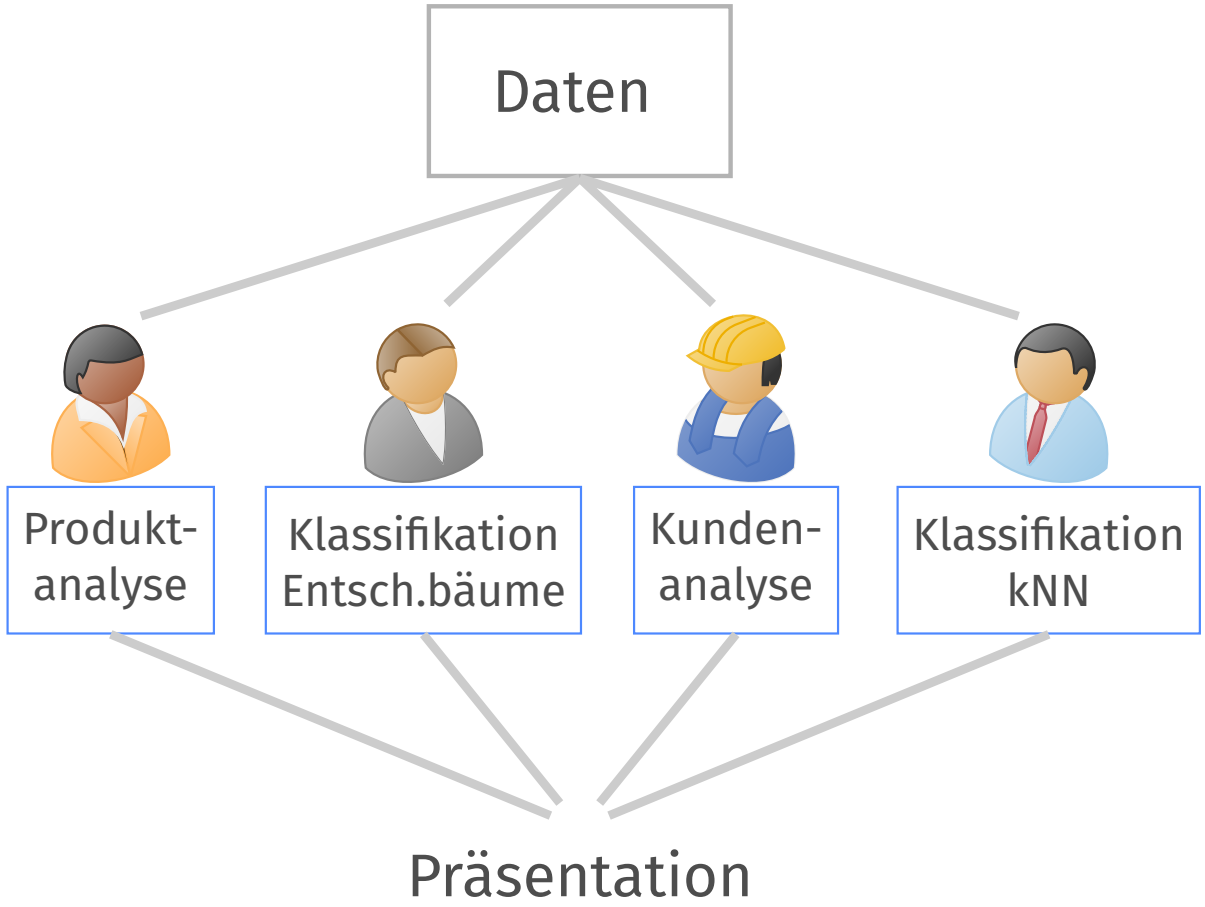
Abschlussprojekte

- Daten-/Themenvergabe heute
- Mögliche Themen/Datensätze stelle ich gleich vor
- Am Ende 1 Präsentation je Gruppe
- Jeder Teilnehmer lädt Notebook mit Analyse (ausführlich!) **bis TBD 23:59 Uhr** in Moodle-Kurs hoch
- Präsentationen am **TBD**

Abschlussprojekte



Abschlussprojekte



Abschlusspräsentation

- Termin: **TBD**

Abschlusspräsentation

- Termin: **TBD**

Bewertung

- Verschiedene Aspekte je Teilnehmer
- Schlüssige Analyse wichtig
- Ordentliches Notebook (Visualisierungen!)
- Auch erfolglose Modelle (mit Begründung!) gut

- Bewertet wird Präsentation + Hausarbeit
- “Leitfaden” für Projektphase 2:

https://datascience.hs-bochum.de/vorlesungen/ws2022_2023/datascience2

Mögliche Datensätze

- California House Pricing, House Prices Advances

  `https://www.kaggle.com/c/california-house-prices/overview`

  `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`

- Predict Future Sales for Store/Product

  `https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview`

  `https://www.kaggle.com/c/instacart-market-basket-analysis/data`

- Natural Language Processing with Disaster Tweets

  `https://www.kaggle.com/c/nlp-getting-started`

- Inside AirBnB

  `http://insideairbnb.com/get-the-data.html`

- RKI Covid19

  `https:`

  `//npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0`

Mögliche Datensätze

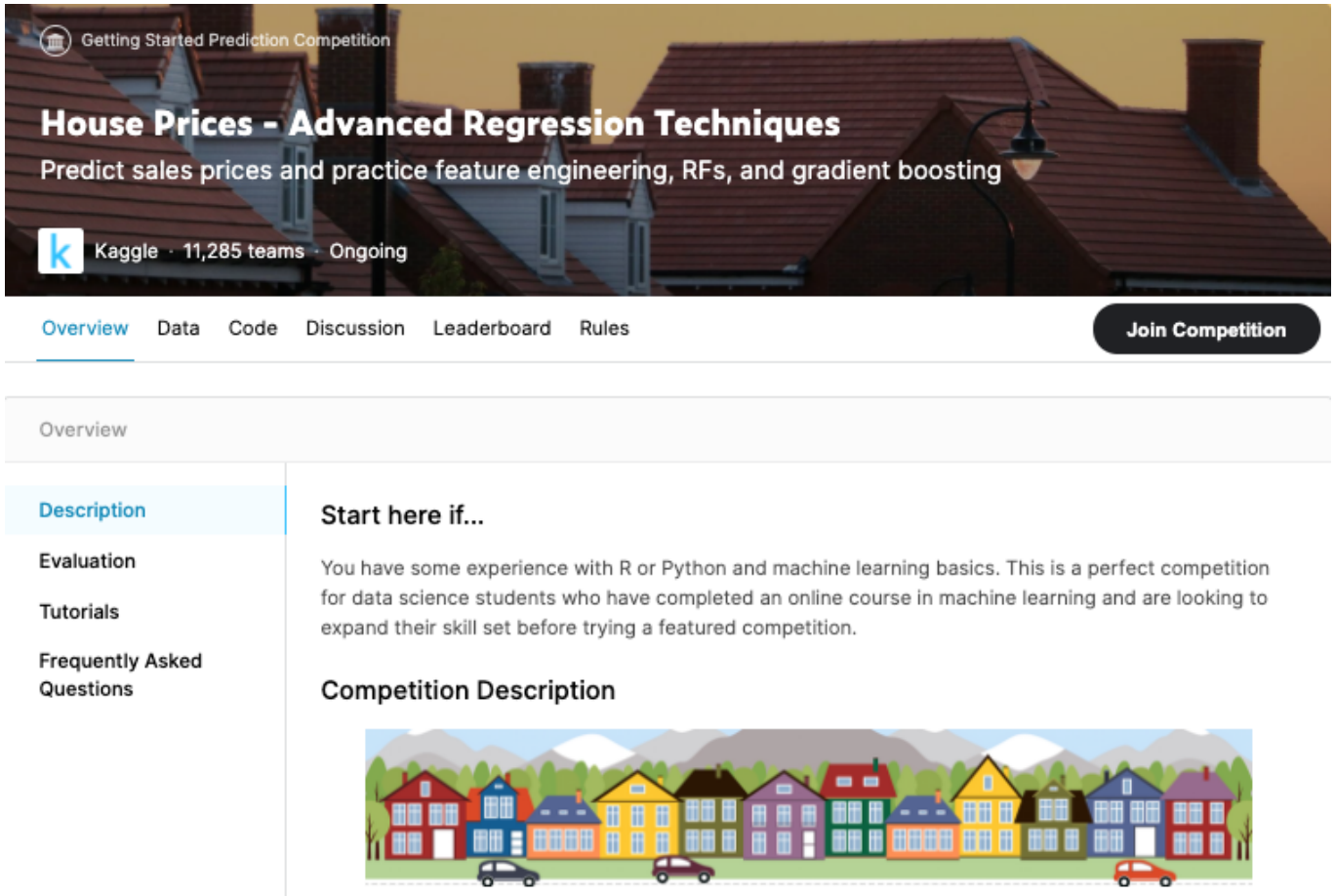
- Sprint-Preise bei Tankerkönig

  `https://tankerkoenig.de`

- Immobilienpreise von Immoscout 24

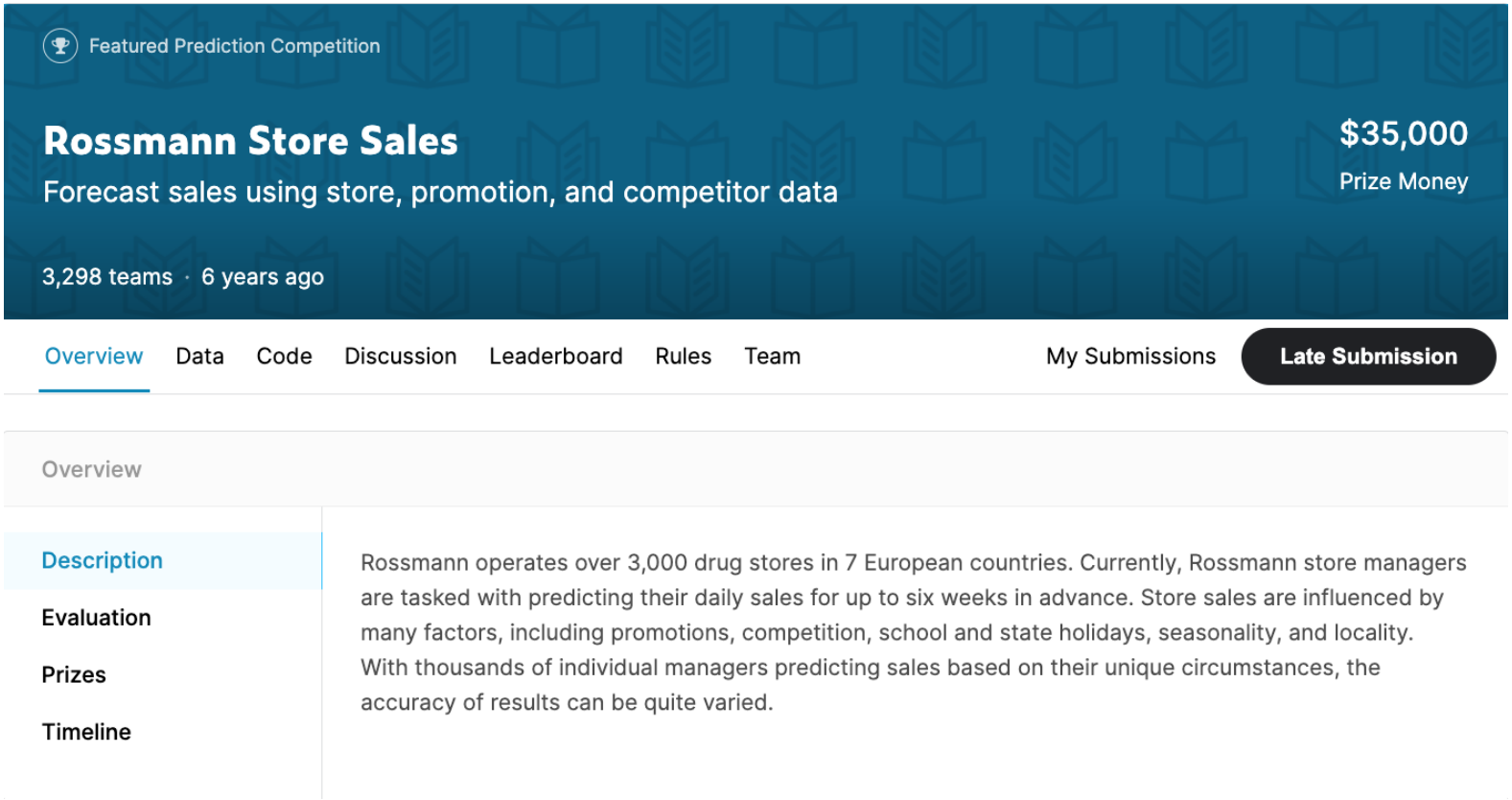
  `https://www.rwi-essen.de/forschung-beratung/weitere/forschungsdatenzentrum-ruhr/datenangebot/rwi-geo-red-real-estate-data`

Vorhersage von Immobilienpreisen



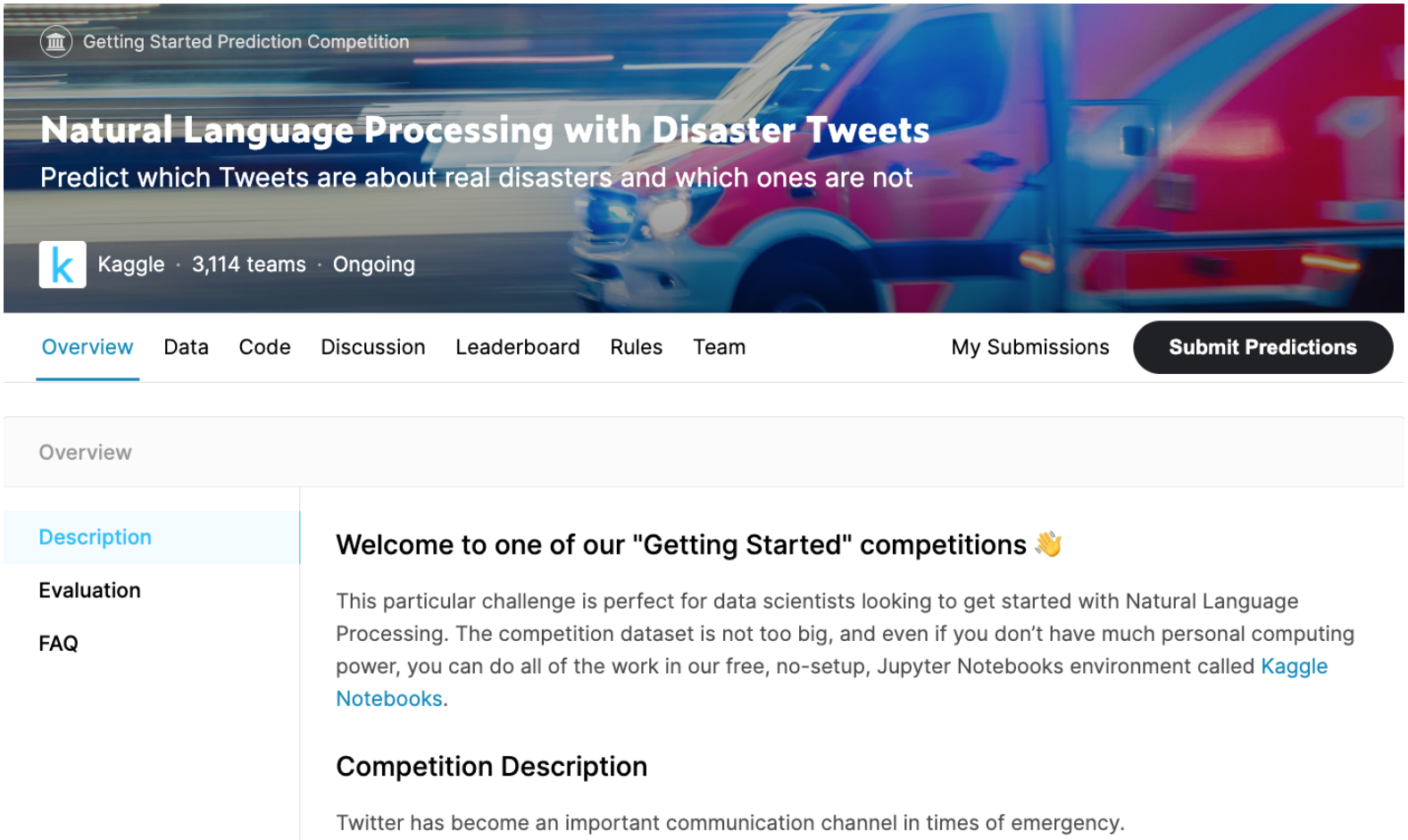
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Vorhersage von Verkäufen



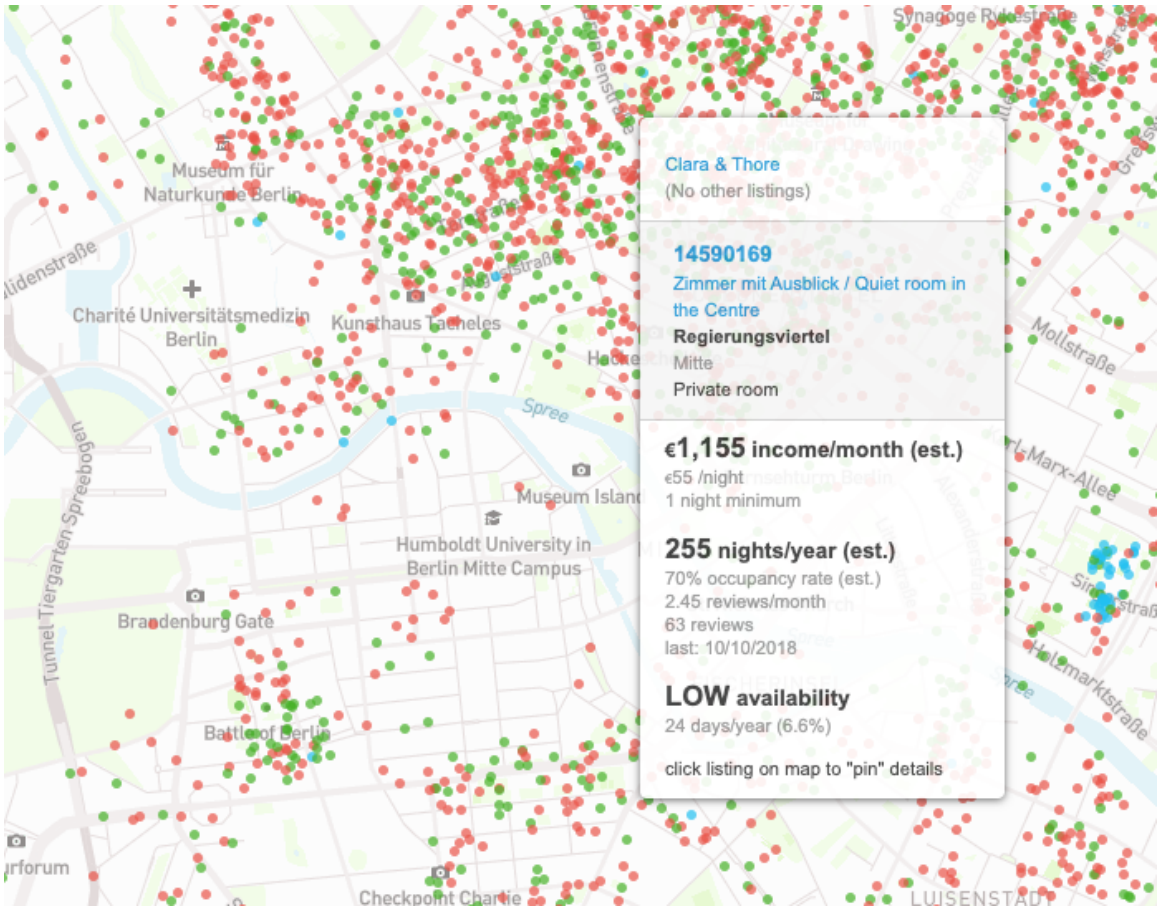
https://www.kaggle.com/c/rossmann-store-sales/overview

Meint der Tweet eine Katastrophe? Ja/Nein



https://www.kaggle.com/c/nlp-getting-started

Analyse von AirBnB-Daten (z.B. Berlin)



http://insideairbnb.com/get-the-data.html

Mögliche Fragestellungen

- Welche Gegenden von z.B. Berlin sind teuer/günstig?
- Welche Eigenschaften von Wohnungen führen zu hohen Mietpreisen?
- Wie gut läßt sich der *est. income* vorhersagen?
- Wo befinden Sie die Hotspots *professioneller* Vermieter? (Hosts mit mehr als 2 oder 3 Angeboten)

Mögliche Fragestellungen

- Welche Gegenden von z.B. Berlin sind teuer/günstig?
- Welche Eigenschaften von Wohnungen führen zu hohen Mietpreisen?
- Wie gut läßt sich der *est. income* vorhersagen?
- Wo befinden Sie die Hotspots *professioneller* Vermieter? (Hosts mit mehr als 2 oder 3 Angeboten)

Advanced

- Wie unterscheiden sich Preise/Angebote von Stadt A und Stadt B?

Analyse der Spritpreise

- Veröffentlichungspflicht der Tankstellen
- Spritpreise seit 2014

Spritpreise für alle!

Analyse der Spritpreise

- Steigen/sinken die Preise in allen Regionen gleich?
- Wie verläuft der Spritpreis zum Ölpreis?
- Gibt es Marken, die stärker an den Ölpreis gebunden sind?
- Hat sich das Verhalten seit Kriegsbeginn (Ukraine) geändert?

Analyse von Daten von ImmoScout24

- Leibniz-Institut für Wirtschaftsforschung
- Forschungsdatenzentrum Ruhr (FDZ)

`https://www.rwi-essen.de/forschung-beratung/weitere/forschungsdatenzentrum-ruhr/datenangebot/rwi-geo-red-real-estate-data`

Wie geht’s weiter?

Zeitplan Projektphase – 2022

1.12.2022, 9 Uhr – Visualisierung

- Status: Bericht aus den Gruppen
- Vortrag: Visualisierung

8.12.2022, 9 Uhr – Big Data

- Status: Bericht aus den Gruppen
- Vortrag: Big Data

15.12.2022, 9 Uhr – No-Code Datenanalyse

- Status: Bericht aus den Gruppen
- Vorstellung: No-Code Datenanalyse

22.12.2022, 9 Uhr – TBD?

- Status: Bericht aus den Gruppen