University ranks

Backé Julian, Salzer Tobias, Singh Ajayvir
Group 3

Agenda / Questions

- How do university rankings change over time?
- Which characteristics of universities contribute most to good rankings, or to large changes in the ranking position?
- How do these characteristics correlate with characteristics of cities or countries in which the university is located?
- Are there predictors for increases or decreases in the rankings?

Workflow

1. **Data acquisition and preprocessing**
2. **Data Exploration – with results**
3. **Data Modelling – with results**

Data acquisition and loading

- **Center for World University Rankings** (CWUR) → Main data source
- **Academic Ranking of World Ranking Universities** → by ShanghaiRanking, contains rankings from 2005
- **World University Rankings** → by Times Higher Education, contains inherent characterstics of universities such as number of students, ratio, ...
- **Public Expenditure** → by National Center of for Education Statistics, for correlation with characteristics of country

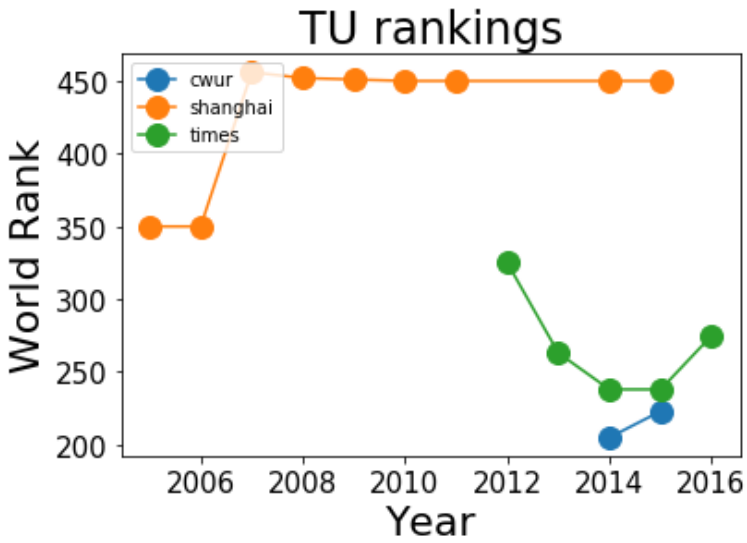
Data acquisition and loading

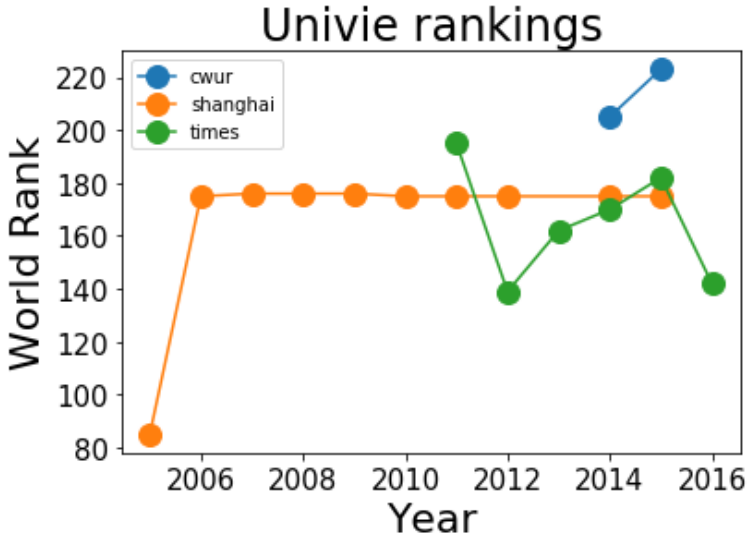
- **Human Development Index** → by United Nations Development Programme, for correlation with HDI
- **Countries By Region** → by US Government, for data analysis of rankings by region
- **Corruption Perception Index** → by Transparency International, for correlation with corruption

Challenges

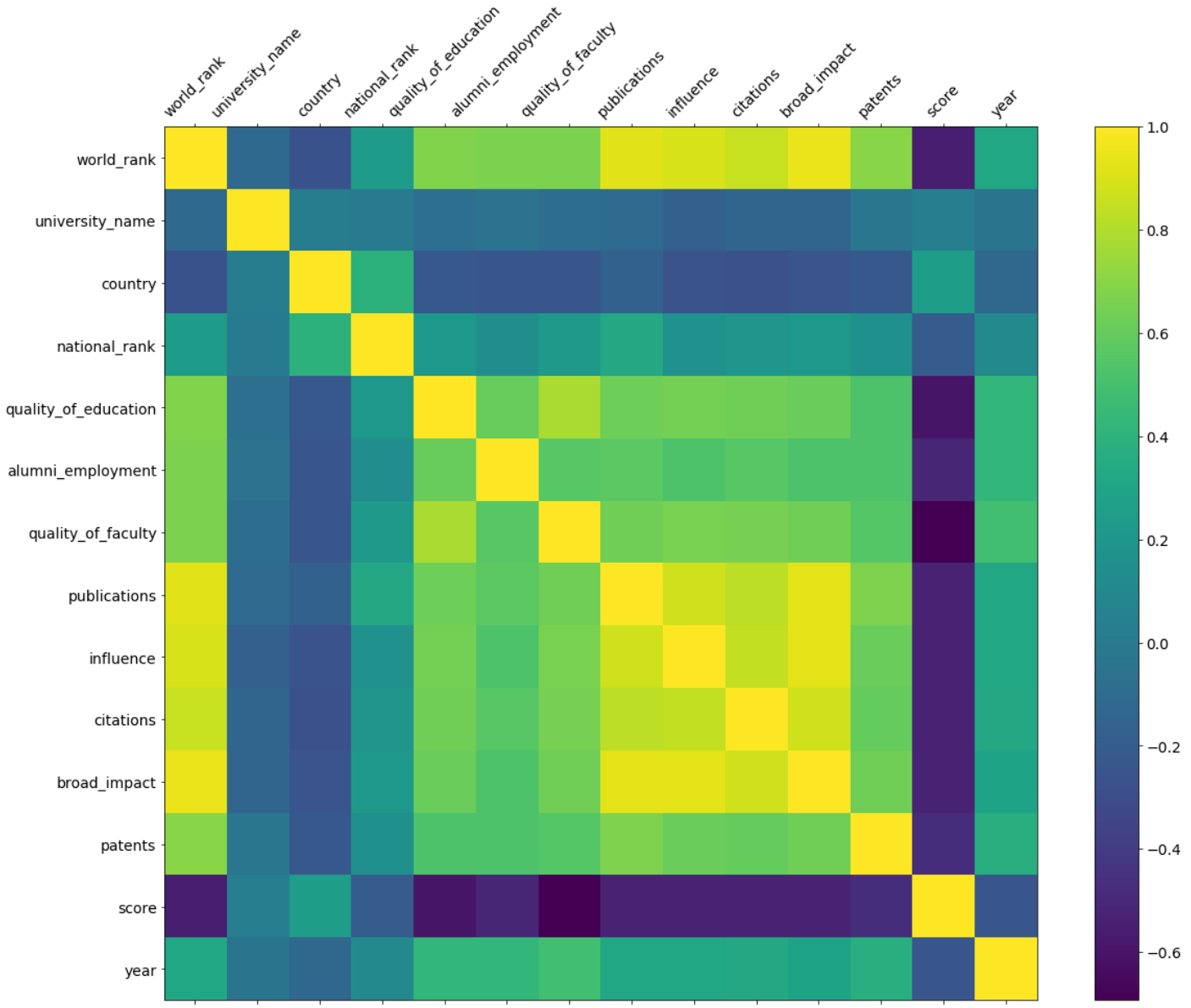
- Rankings throughout the surveys very different - performance of ML-algorithms dependent on chosen survey.
- University names and country names are not standardized - makes it difficult to merge the datasets.
- In general, very messy data: missing values stored in many forms (NAN, '?', -,...), a lot of typos, random symbols attached to numbers,...

Data Exploration - Univie vs TU Vienna

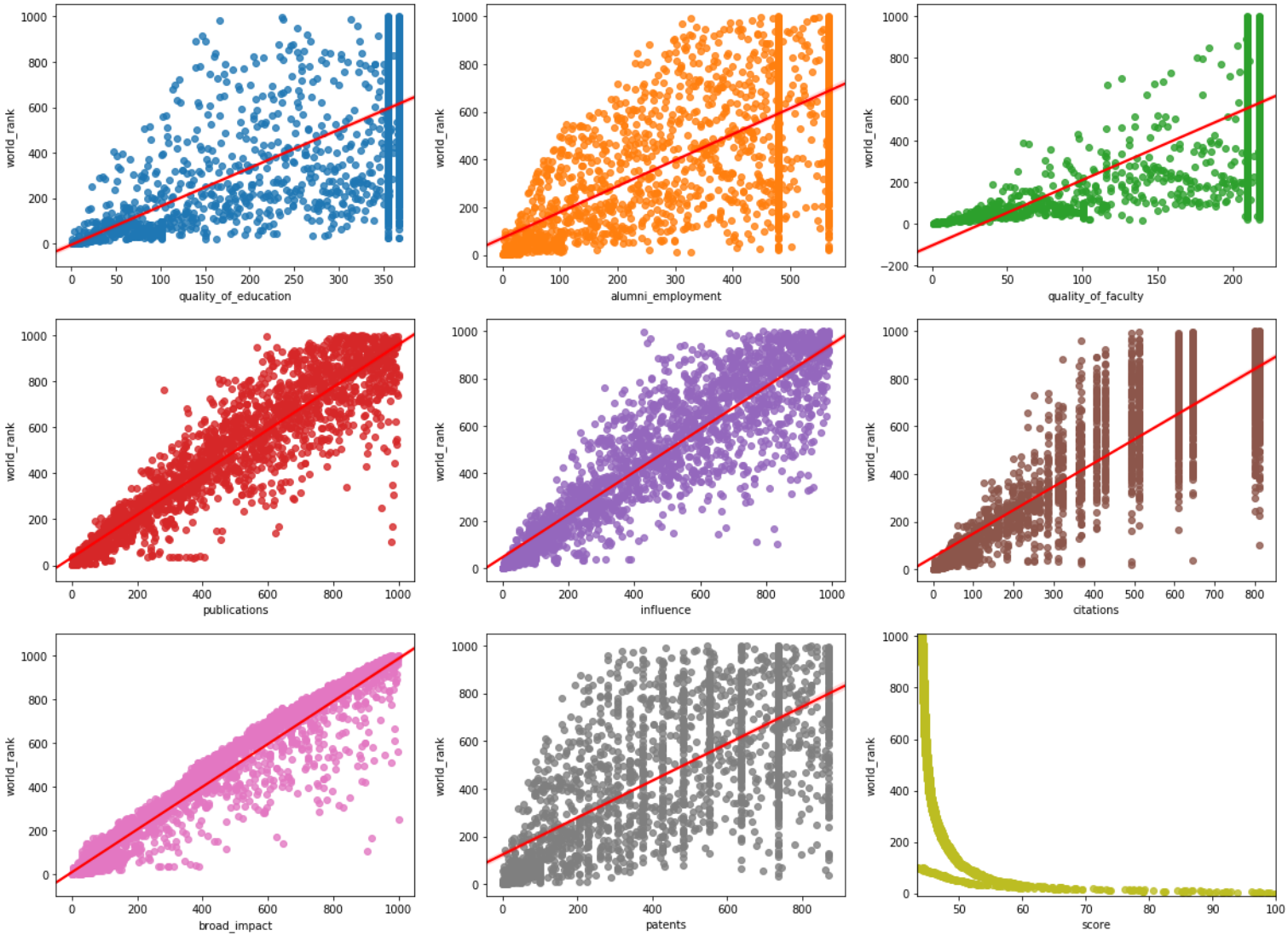




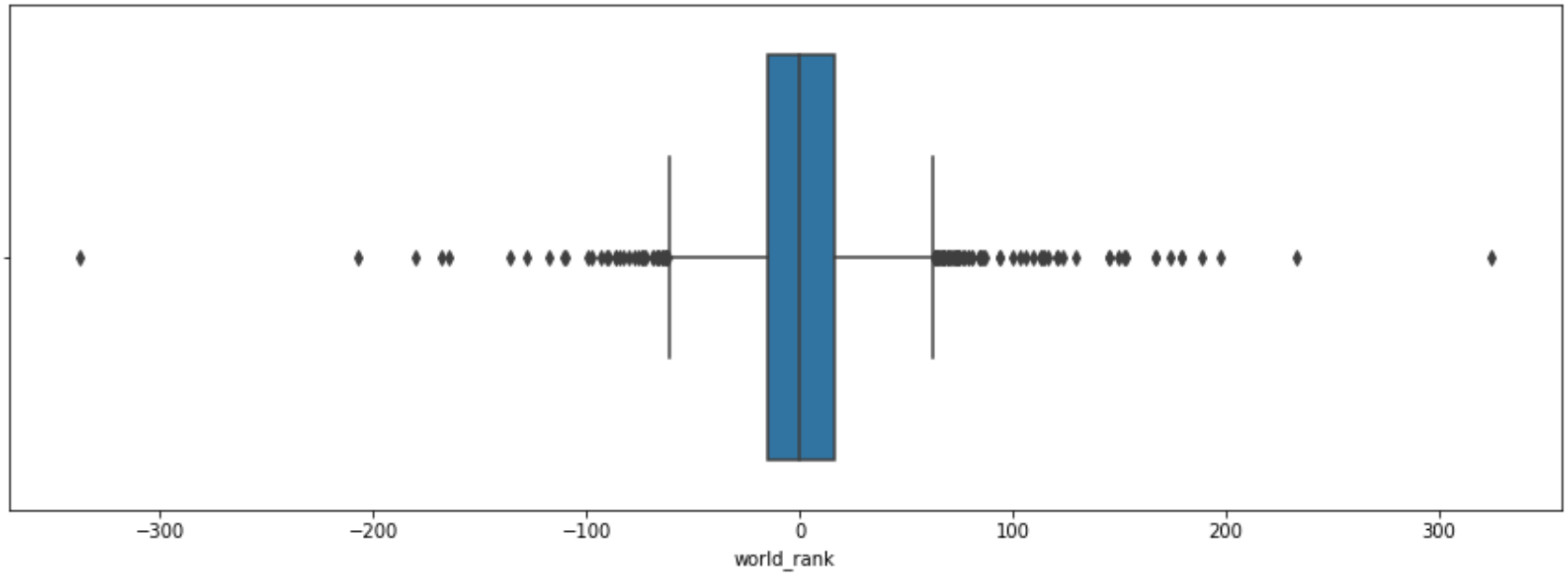
Data exploration - Heatmap for overview



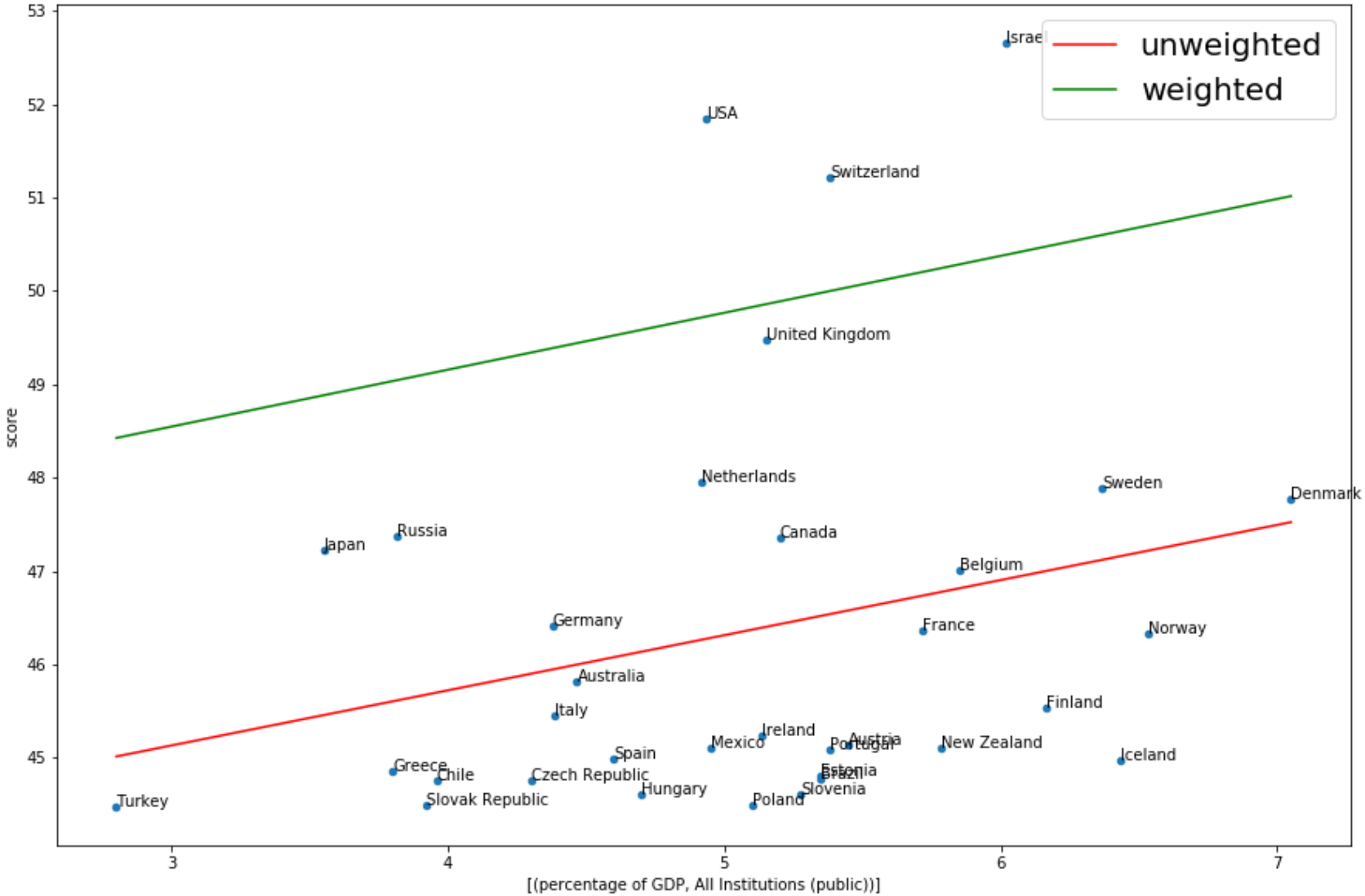
Data exploration - World rank vs other characteristics



Data exploration - Ranking deviation over years (2012-2015)



Data exploration - Expenditure for education



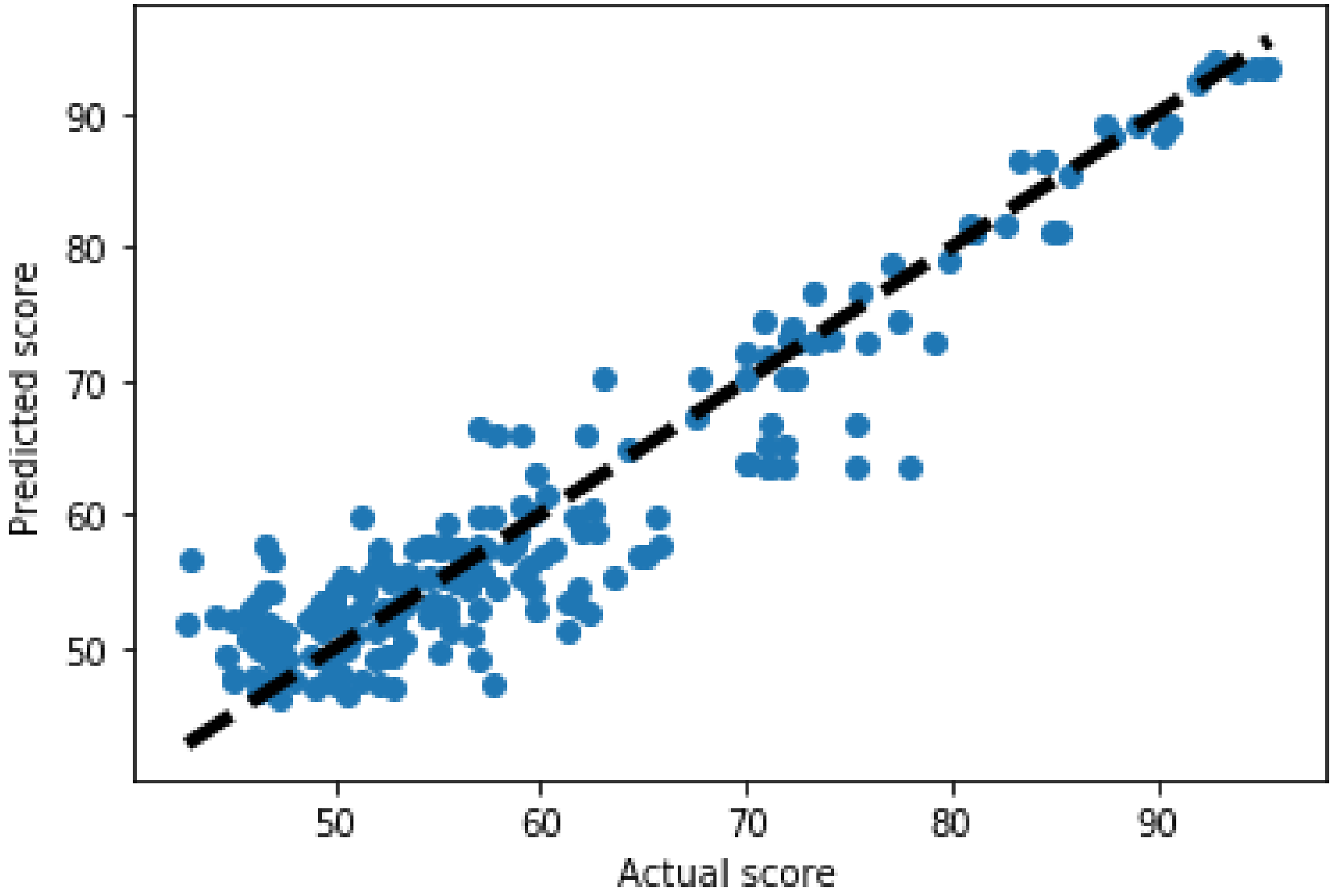
Data exploration - Other results

independent variable	dependent variable	impact
expenditures for education (all institutions)	mean score of country	YES
expenditures for education (all institutions)	max. score of country	NO
expenditures for education (higher institutions)	mean score of country	YES
expenditures for education (higher institutions)	max. score of country	NO
number of universities	mean score of country	SLIGHT
number of inhabitants	mean score of country	NO
univerisites per inhabitant	mean score of country	YES
HDI	mean score of country	YES
corruption	mean score of country	NO

Predicting times-score

1. **Setup:**
    1. only top 200 universities from times-survey considered
    2. predict times-score using number of students, student-staff-ratio, percentage of international students and percentage of female students.
    3. grid-search over different ML-algorithm and parameters
2. best ML-algorithm: random forest
3. Mean squared error: 3.4

Actual scores vs predicted scores



Final thoughts

- Working with messy data can be pretty tedious
- Preprocessing takes up a big chunk of human and computational effort
- Preprocessing and data cleaning essential for meaningful results
- Acquisition of suitable data is difficult

Thank you for your attention!

Any Questions?