

# Data Engineering 1: SQL and Different Shapes of Data

## Term Project 2

Artem Timonov, Işıl Oral, Saad Joiya, Vasilisa Vashchenko

Central European University

December 4, 2024

# Introduction

- ▶ Machine learning algorithm to predict happiness score of the countries
- ▶ Data sources enriched by APIs
- ▶ Knime, low-code data analytics tool, as the central platform

# Data & ETL

## Data Sources:

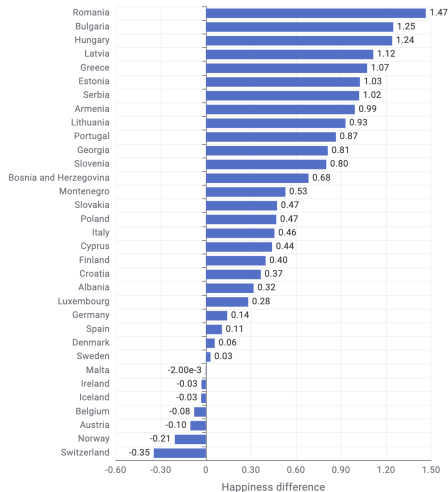
- ▶ Yearly data of 'World Happiness Report' from 2015 to 2023 retrieved from Kaggle
- ▶ World Bank data
- ▶ Eurostat APIs
  1. Standardised house price-to-income ratio – annual data
  2. Pollution, grime or other environmental problems
  3. Population on 1 January
  4. Mobile broadband internet traffic (within the country)

## ETL Processes

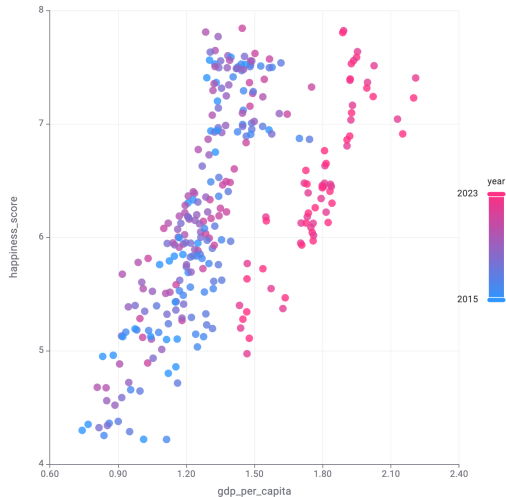
- ▶ using string replacer in Knime to ensure data compatibility across data sets
- ▶ Python script and HTML query for integration of API
- ▶ normalizing variables
- ▶ handling missing data

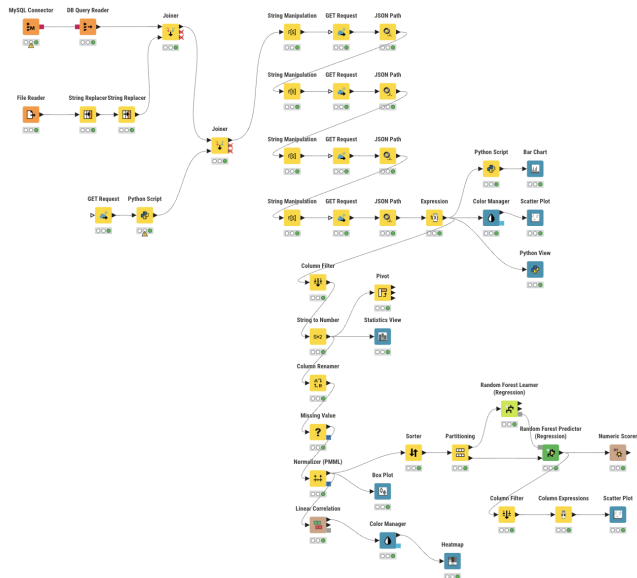
# Descriptive Analysis

Happiness level change from 2015 to 2023



Scatter Plot





to the Knime extensions

to the normalized variables

to the heatmap

# Algorithm

RowID	Prediction (happiness_score) <i>Number (double)</i>
R^2	0.729
mean absolute error	0.296
mean squared error	0.131
root mean squared error	0.362
mean signed difference	0.015
mean absolute percentage error	0.047
adjusted R^2	0.729

Figure: Statistics of predicted outcome

Scatter Plot: Prediction error by target value

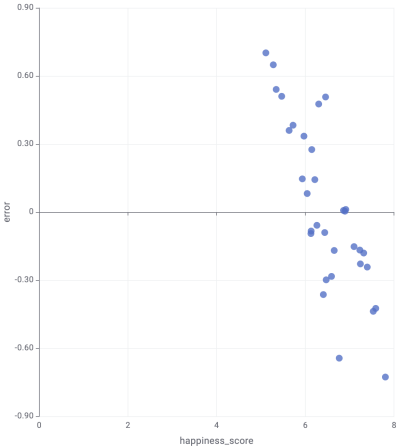


Figure: Prediction errors

# Conclusion

- ▶ Diverse data sources: World Happiness Report 2023, World Bank, Eurostat APIs
- ▶ Data processing with AWS and careful ETL process in Knime
- ▶ "Random Forest Learner" as machine learning algorithm to predict last year of data
- ▶ Fitness of the model comparison to actual data

# After ETL Process

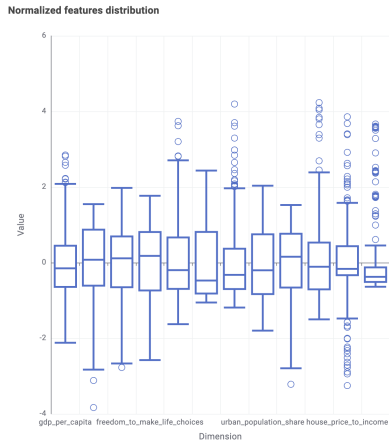


Figure: Normalized features



# References

- ▶ Databank. World Bank Group DataBank. (n.d.). <https://databank.worldbank.org/>
- ▶ Eurostat. Standardised house price-to-income ratio – annual data [tipsho60]. Available at: <https://ec.europa.eu/eurostat/databrowser/bookmark/3e391ba8-1b5a-4adf-a32f-6a9c02e21b73?lang=en>
- ▶ Eurostat. Pollution, grime or other environmental problems [ilc\_mddw02]. [https://doi.org/10.2908/ILC\\_MDDW02](https://doi.org/10.2908/ILC_MDDW02).
- ▶ Eurostat. Population on 1 January [tps00001]. <https://doi.org/10.2908/TPS00001>
- ▶ Eurostat. Mobile broadband internet traffic (within the country) [isoc\_tmi\$defaultview]. Available at: <https://ec.europa.eu/eurostat/databrowser/bookmark/c815e093-d12b-4bd6-b2dc-dbcd3e970bba?lang=en>
- ▶ Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., Wang, S. (2023). Statistical Appendix for “World happiness, trust and social connections in times of crisis,” Chapter 2 of World Happiness Report 2023. In World Happiness Report 2023 (11th ed.). (Eds.). Sustainable Development Solutions Network.
- ▶ Islam, S. (2023, September 9). World happiness report (till 2023). Kaggle. <https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors>
- ▶ Wikimedia Foundation. (2024, October 26). List of ISO 3166 country codes. Wikipedia. [https://en.wikipedia.org/wiki/List\\_of\\_ISO\\_3166\\_country\\_codes](https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes)

# Knime Extensions

- ▶ KNIME Database Extension
- ▶ KNIME Python Script Extension

Back

# Appendix

Heatmap

