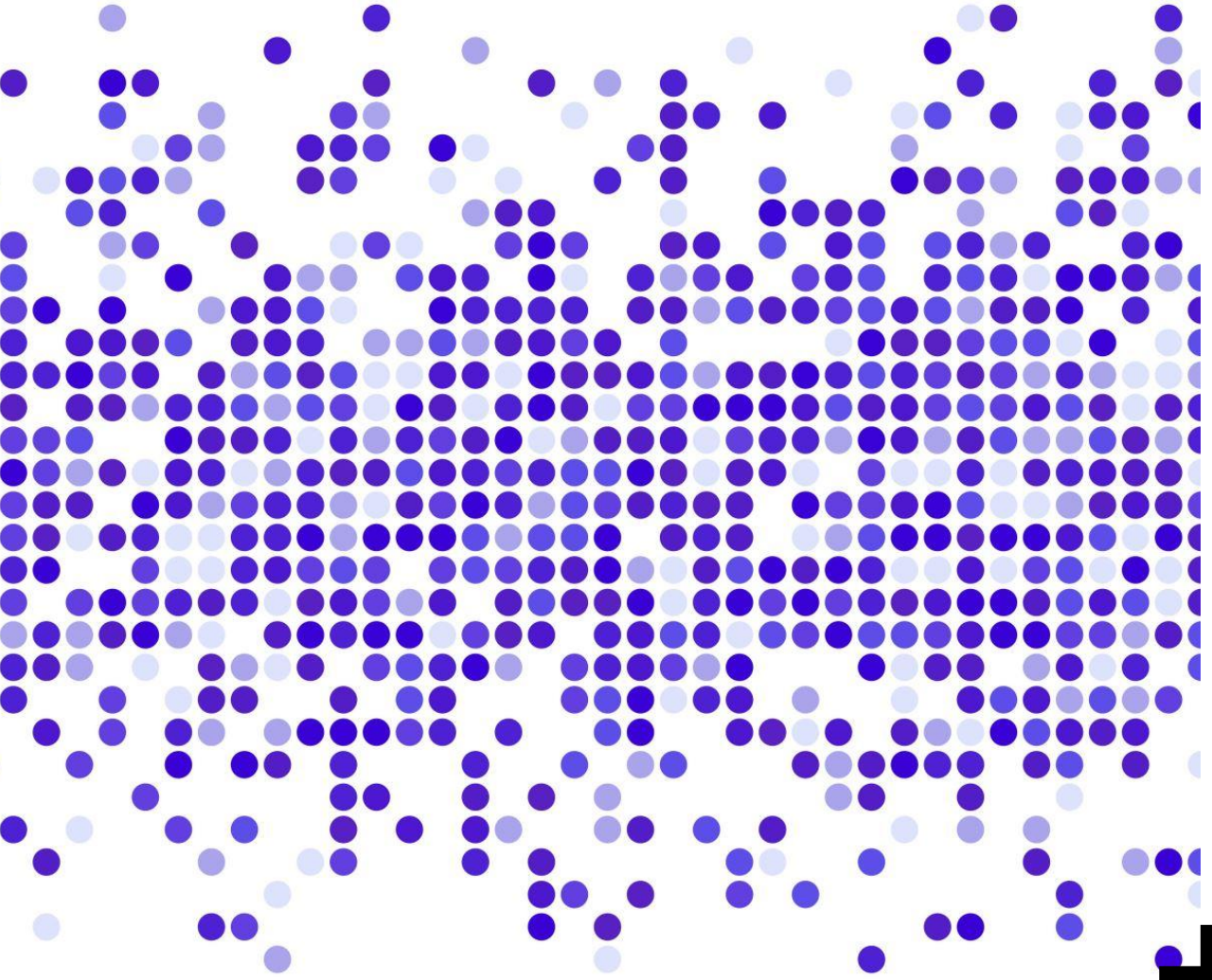


Predicting sales prices for US houses

Team 1

Bart Lesschen

Bhinawa Putra Raja

Martijn Ma

Sebastian Piest

BI meets Data Science

– Learn PowerBI
– Explore use of R Studio (and similar)
– Experiment with Power BI + R
– Create a working ML pipeline
– Sharpen/update data science skills
– Have fun ☺

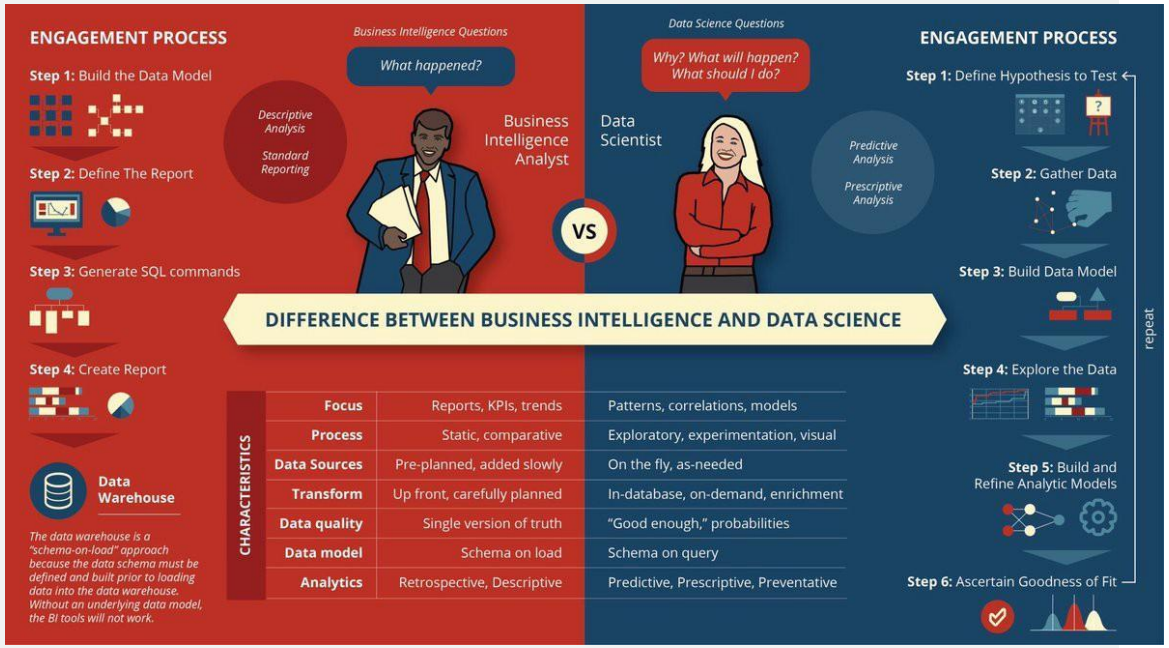


Image source:
https://caiomsouza.medium.com/difference-between-bi-business-intelligence-and-data-science-1a9c7628bbdb (23-09-2022)

Starting with some science before diving into the data (swamp)

- Variety of approaches:
  - Machine learning (Komagome-Towne, 2016; Ravikumar, 2017; Phan, 2018; Truong et al., 2020)
  - Deep learning (Wang et al., 2021)
  - Lineair- and Logistic regression (He, He, 2021), Fuzzy (Sarip, 2016) and Multiple regression (Zhang, 2021)
  - Time series (Wang, Juntao, et al. 2018)
  - AutoML (Li et al., 2020)
  - Particle swarm optimization (Zhou, 2021)
- Different sources:
  - Mainly based on historical data
  - Economic parameters (Li and Chu, 2017)
  - News data (Kirkeby and Larsen, 2021)

Clearly no “silver bullet” solution or generalized set of features

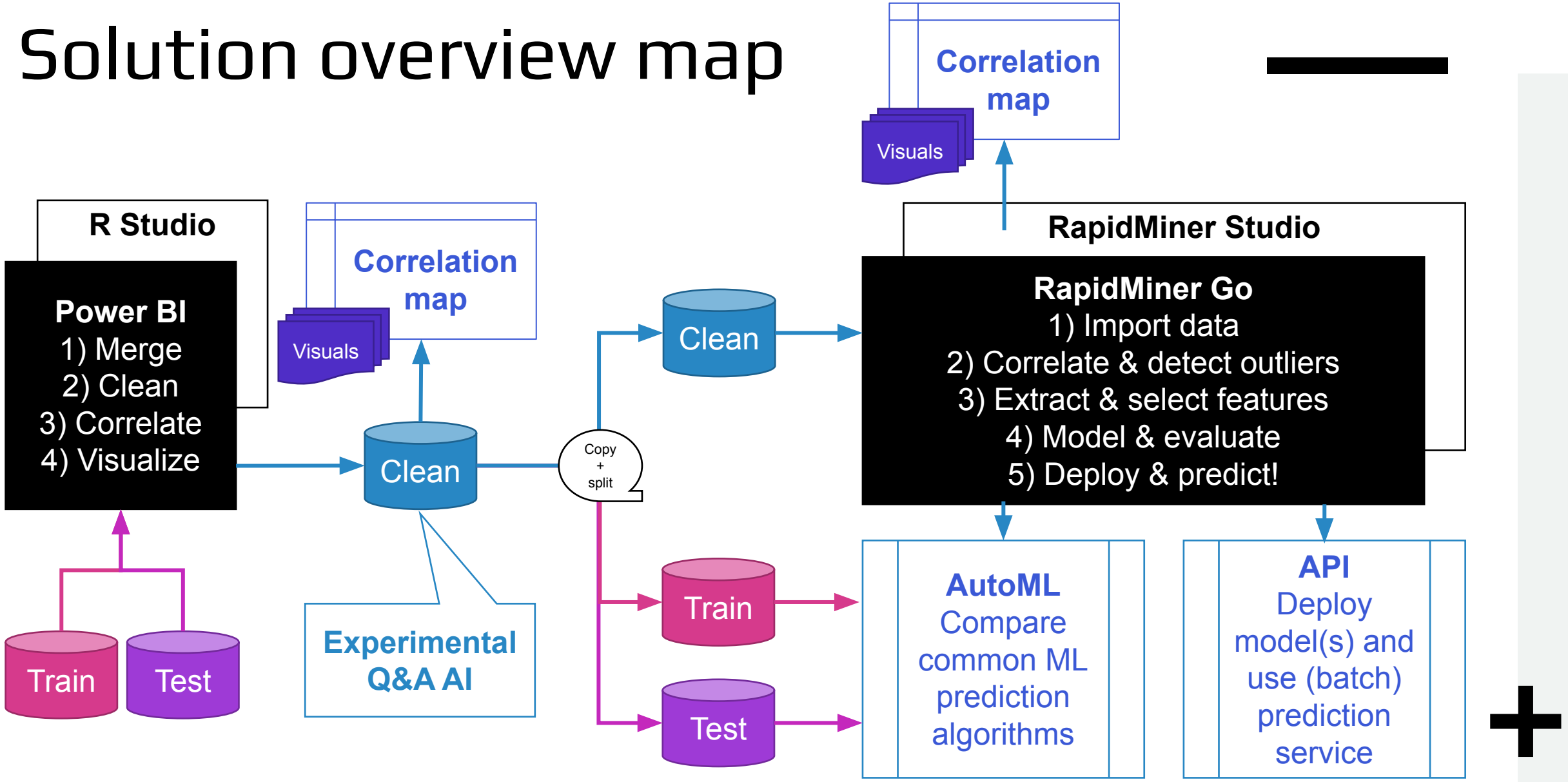
Questions for this week:

- How to test historical, spatial and temporal influences?
- Which risk of bias is in the data?

Questions for later/others:

- Which indices correlate with housing prices?
- What influence does public opinion and sentiment have?

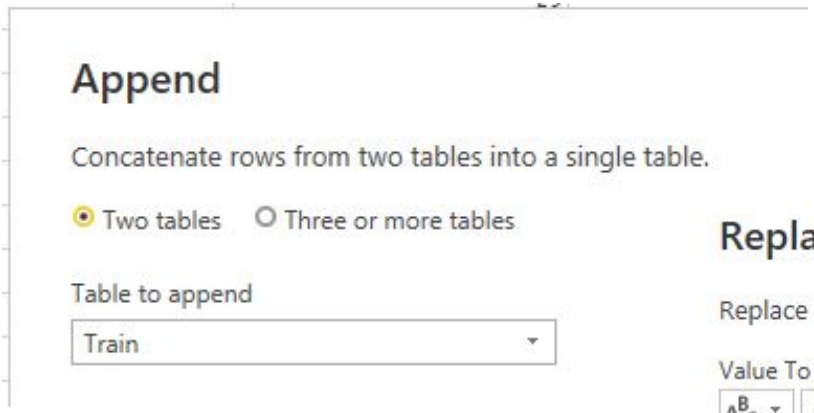
Solution overview map

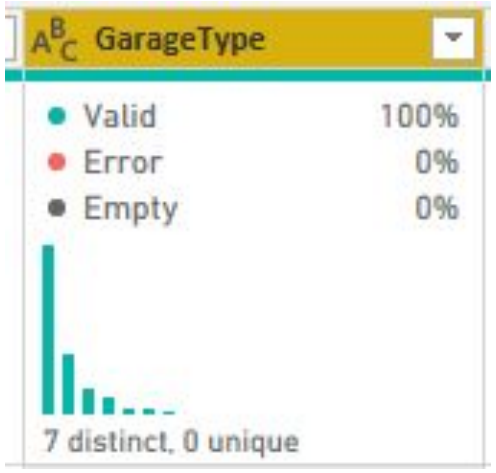


Use of PowerBI

Data crunching

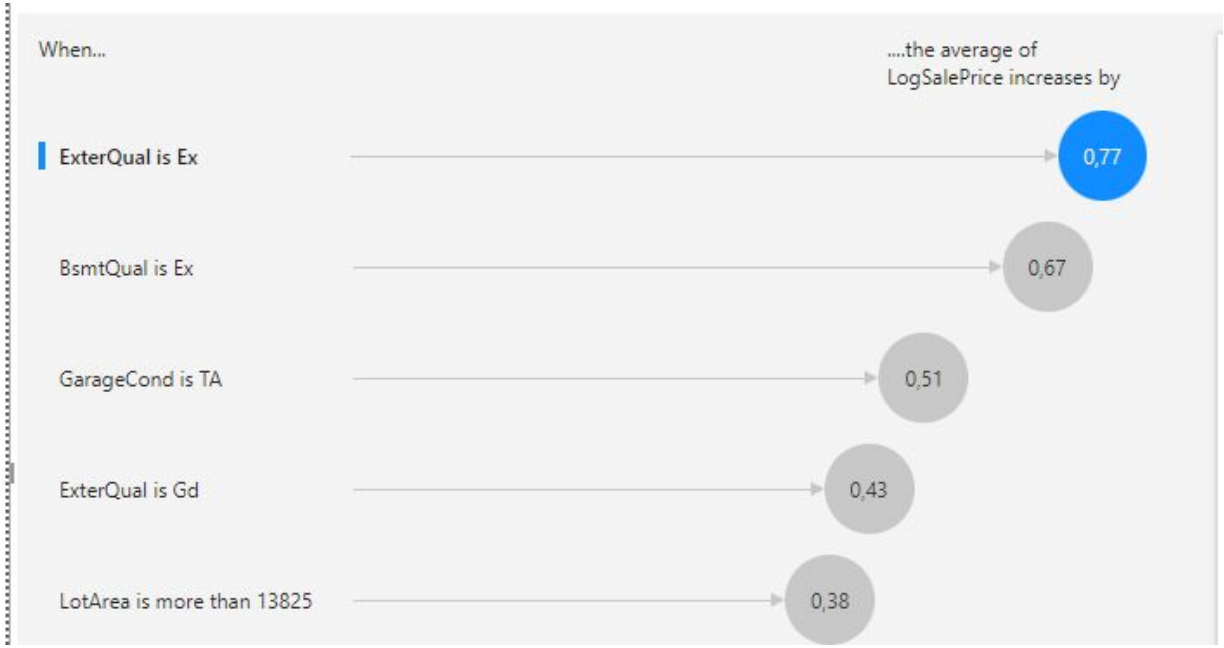
- Dataset merge
- Data cleaning
- Data exploration/visualizations
- Correlation map
- Influencing factors
- Experimental AI: Q&A
- Experimental: integrate R script





Use of PowerBI

Correlations and AI-based Q&A

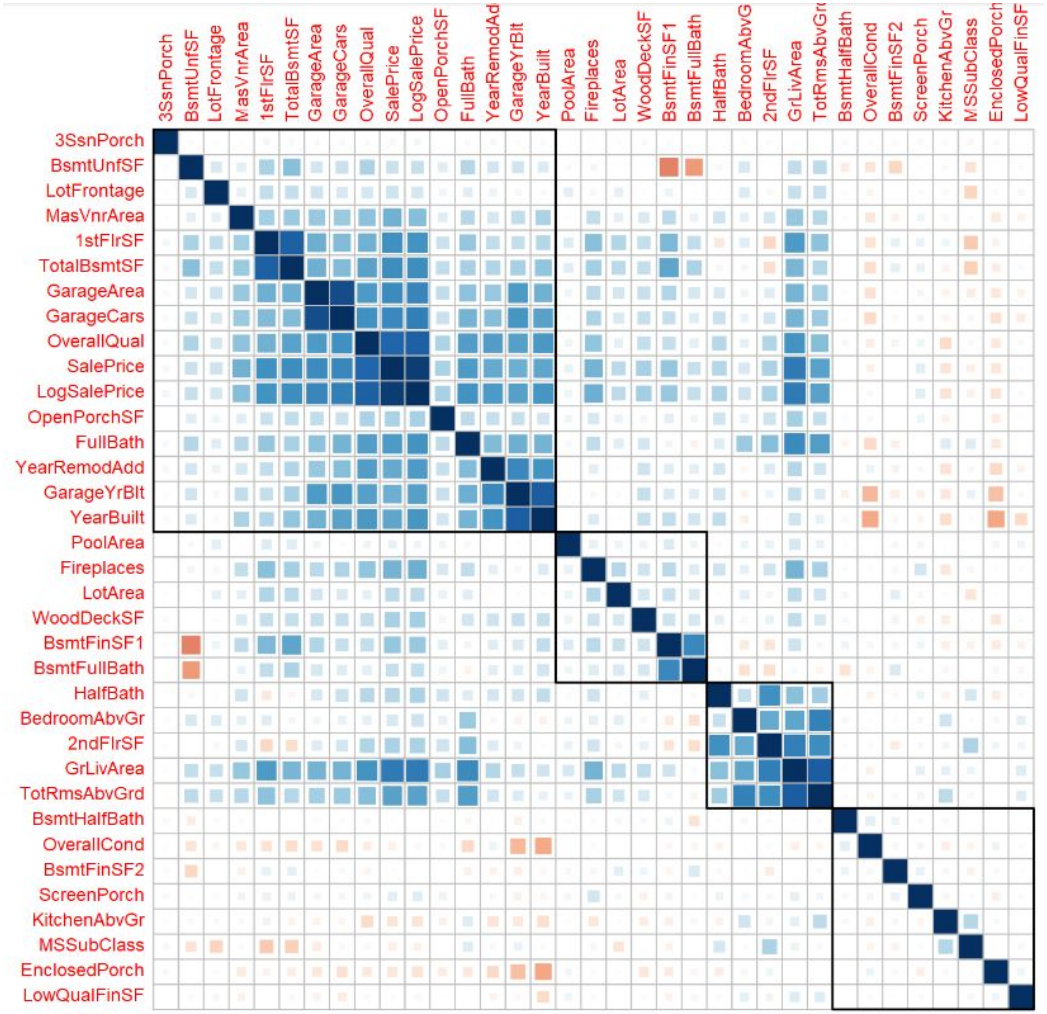


Use of R in Power BI

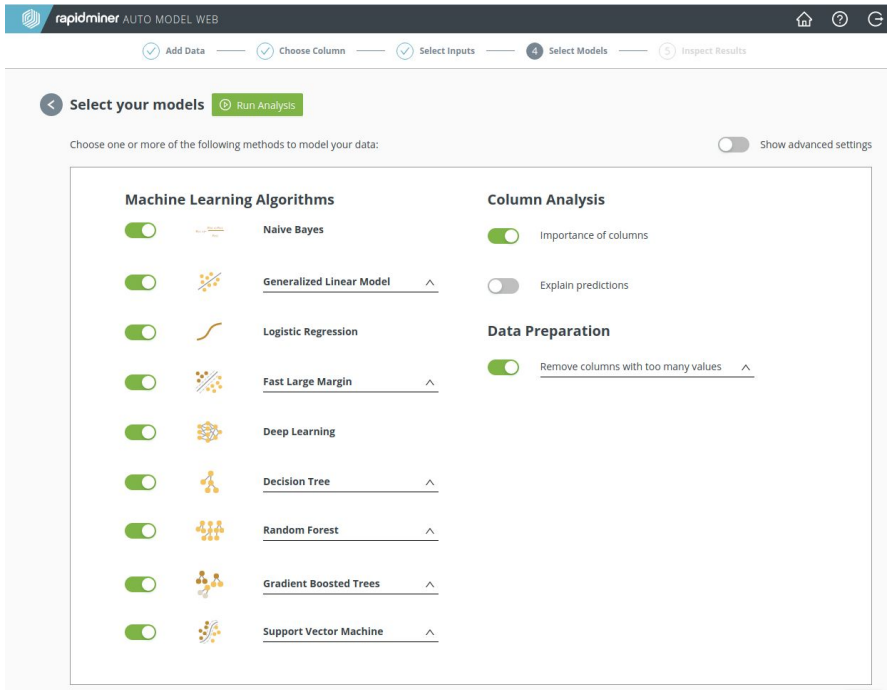
- Data visualization (library: corrplot)
- R script editor in PowerBI
- Explored functionalities in R
- Continued in RapidMiner

R script editor

```
1 # The following code to create a datafram
2
3 # dataset <- data.frame(LogSalePrice)
4 # dataset <- unique(dataset)
```

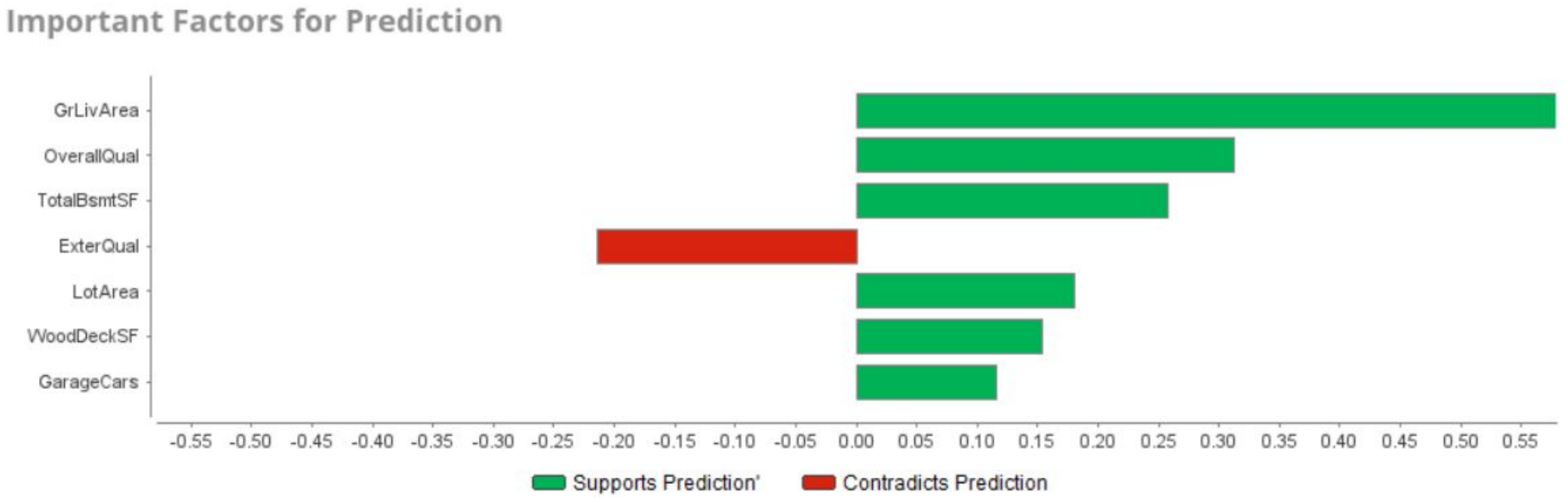


Use of RapidMiner



- Data preparation
- Statistics/visualizations
- Outlier detection
- Modeling with AutoML (6 models + correlation map)
- Model evaluation and selection: Gradient Boosted Trees
- Model deployment
- Scoring

Predicting future sales prices



Model	Relative Error	Standard Deviation	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Generalized Linear Model	0,70%	0,08%	1369,0	42,2	6,9
Deep Learning	0,77%	0,07%	5162,0	923,9	133,2
Decision Tree	1,40%	0,09%	1331,0	11,8	12,1
Random Forest	0,86%	0,12%	21752,0	204,8	192,0
Gradient Boosted Trees	0,69%	0,04%	32132,0	341,9	57,1
Support Vector Machine	138,15%	0,34%	27709,0	3815,2	207,6

Concluding remarks

- Nice combination of BI & Data Science
- Quite a lot features in PowerBI that go beyond BI
- RapidMiner really speeds up process with TurboPrep and AutoML - and also offers a variety of statistics / visualizations
- Working ML pipeline in 1 week based on Gradient Boosted Trees and iterative development

Didn’t go in-depth with:

- R scripting
- (Hyper)parameter tuning

PowerBI correlations limited to numeric values

We had fun and got predictions :)

References used in this presentation

He, K., & He, C. (2021, November). Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data. In 2021 IEEE International Conference on Computing (ICOCO) (pp. 241-246). IEEE.

Kirkeby, S. J., & Larsen, V. H. (2021). House price prediction using daily news data (No. 5/2021). Staff Memo.

Komagome-Towne, A. (2016). Models and visualizations for housing price prediction. Faculty of California State Polytechnic University, Pomona.

Li, R. Y. M., Chau, K. W., Li, H. C. Y., Zeng, F., Tang, B., & Ding, M. (2020, July). Remote sensing, heat island effect and housing price prediction via AutoML. In International Conference on Applied Human Factors and Ergonomics (pp. 113-118). Springer, Cham.

Li, L., & Chu, K. H. (2017, May). Prediction of real estate price variation based on economic parameters. In 2017 International Conference on Applied System Innovation (ICASI) (pp. 87-90). IEEE.

Phan, T. D. (2018, December). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In 2018 International conference on machine learning and data engineering (iCMLDE) (pp. 35-42). IEEE.

Ravikumar, A. S. (2017). Real estate price prediction using machine learning (Doctoral dissertation, Dublin, National College of Ireland).

Sarip, A. G., Hafez, M. B., & Daud, M. N. (2016). Application of fuzzy regression model for real estate price prediction. Malaysian Journal of Computer Science, 29(1), 15-27.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442.

Wang, P. Y., Chen, C. T., Su, J. W., Wang, T. Y., & Huang, S. H. (2021). Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE Access, 9, 55244-55259.

Wang, J., Yam, W. K., Fong, K. L., Cheong, S. A., & Wong, K. Y. (2018, December). Gaussian process kernels for noisy time series: Application to housing price prediction. In *International Conference on Neural Information Processing* (pp. 78-89). Springer, Cham.

Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. Scientific Programming, 2021.

Zhou, C. (2021, March). House price prediction using polynomial regression with Particle Swarm Optimization. In Journal of Physics: Conference Series (Vol. 1802, No. 3, p. 032034). IOP Publishing.