The Tidyverse & the Data Science Workflow

Marketing Research and Analytics

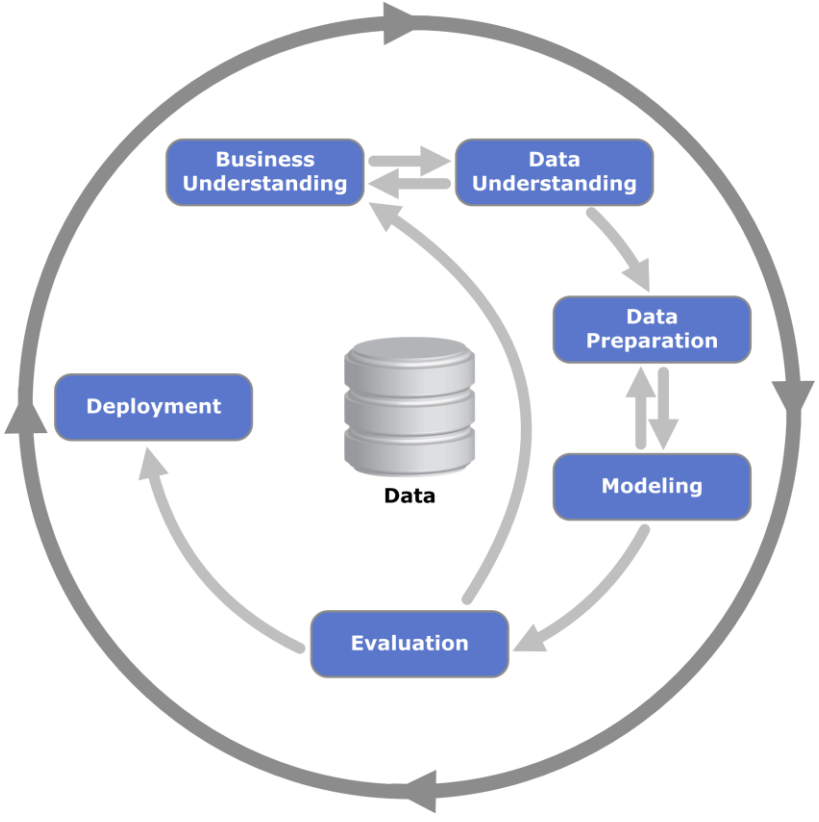
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Andreas MILD

Overview

- CRISP-DM data science workflow
- The Tidyverse for R

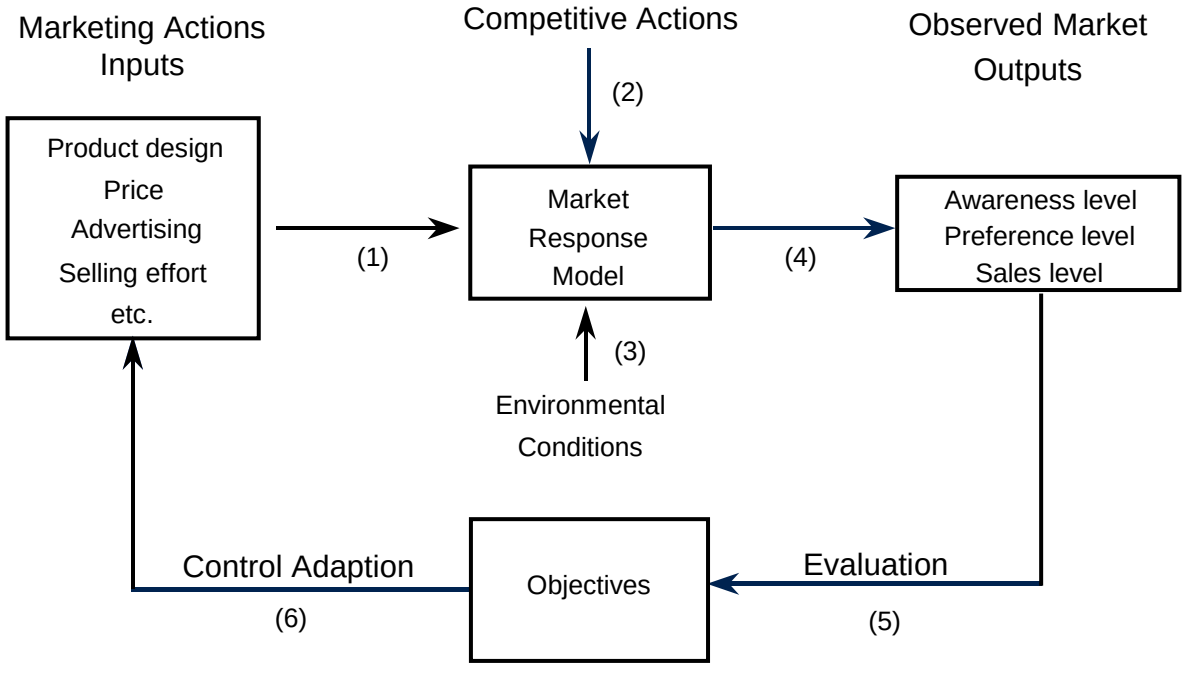
Cross-industry standard process for data mining (CRISP-DM)



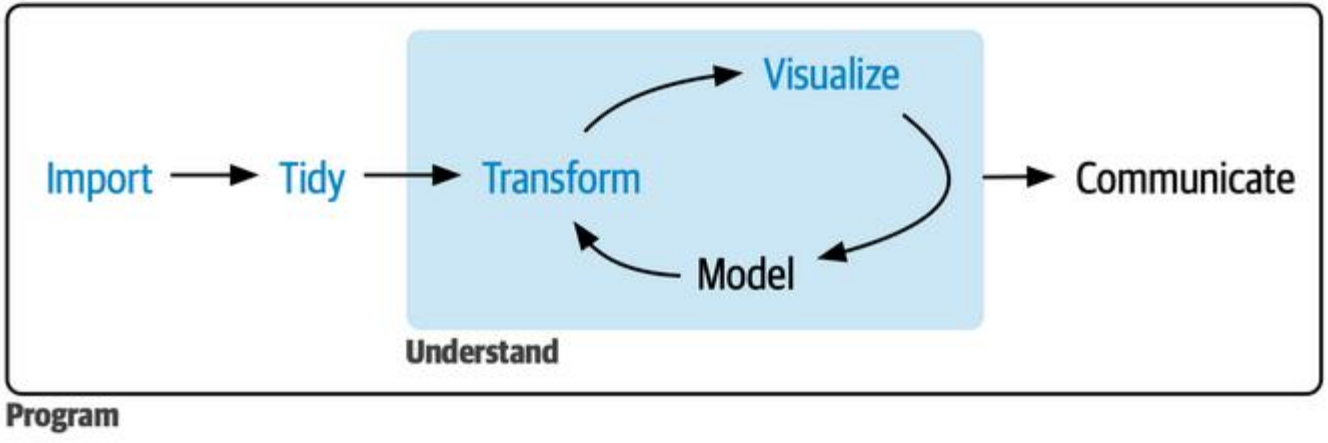
1. **Business understanding –** What is the business problem/need?
2. **Data understanding –** What data do we have/need? Sources? Structure?
3. **Data preparation –** Organize the data for modeling
4. **Modeling –** What modeling techniques should we used?
5. **Evaluation –** Which model best meets the business objectives
6. **Deployment –** How do stakeholders access the results?

Business understanding

Example Structure of a marketing-model



Tools needed for data science



Tidyverse

- Tidyverse is a collection of R packages
- Consistent Syntax and Philosophy
- Data Manipulation
  - **dplyr** provides powerful functions for data manipulation
- Data Tidying
  - **tidyR** helps in reshaping and tidying data
- Visualization
  - **ggplot2** is a highly flexible and powerful tool for creating visualizations
- Integration and Ecosystem
- Community and Documentation
- Reproducibility
  - The code written using Tidyverse is generally more readable and reproducible.

Some notes for the practical application

- **Pipes %>%**

Pipes let you compose a sequence of function calls in a more readable way. The following two lines do the same.

Standard functional form in R using nested functions: print(head(iris))
Using pipes as a sequence of operations: iris %>% head() %>% print()

The pipe supplies the result of the previous function as the first argument to the next function.

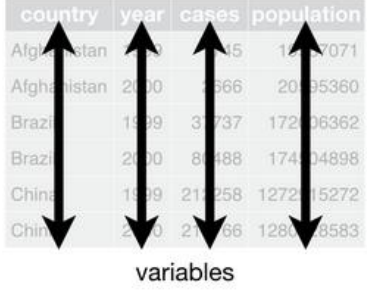
Tidy data

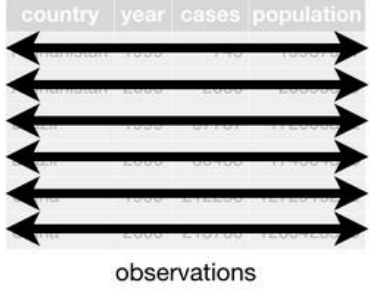
> "Happy families are all alike; every unhappy family is unhappy in its own way."
> — Leo Tolstoy

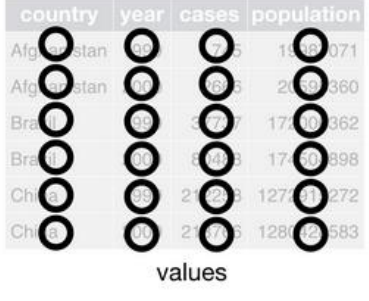
> "Tidy datasets are all alike, but every messy dataset is messy in its own way."
> — Hadley Wickham

What makes a dataset tidy?

1. Each variable is a column; each column is a variable.
2. Each observation is a row; each row is an observation.
3. Each value is a cell; each cell is a single value.



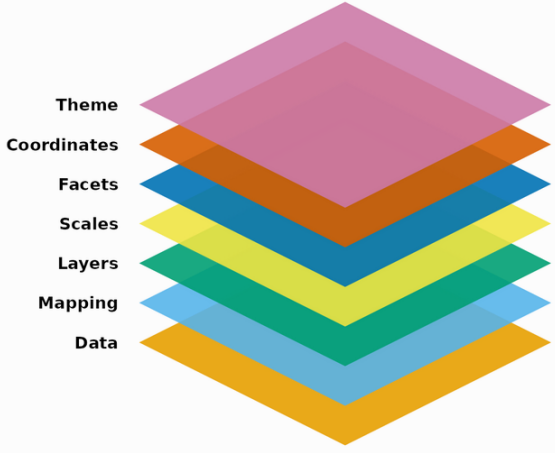




ggplot2

- https://ggplot2.tidyverse.org/articles/ggplot2.html

For structure, we go over the 7 composable parts that come together as a set of instructions on how to draw a chart.



Out of these components, ggplot2 needs at least the following three to produce a chart: data, a mapping, and a layer. The scales, facets, coordinates, and themes have sensible defaults that take away a lot of finicky work.