

ML IN PRACTICE - TYPES OF MODELS

Marcus Birkenkrahe

February 18, 2023

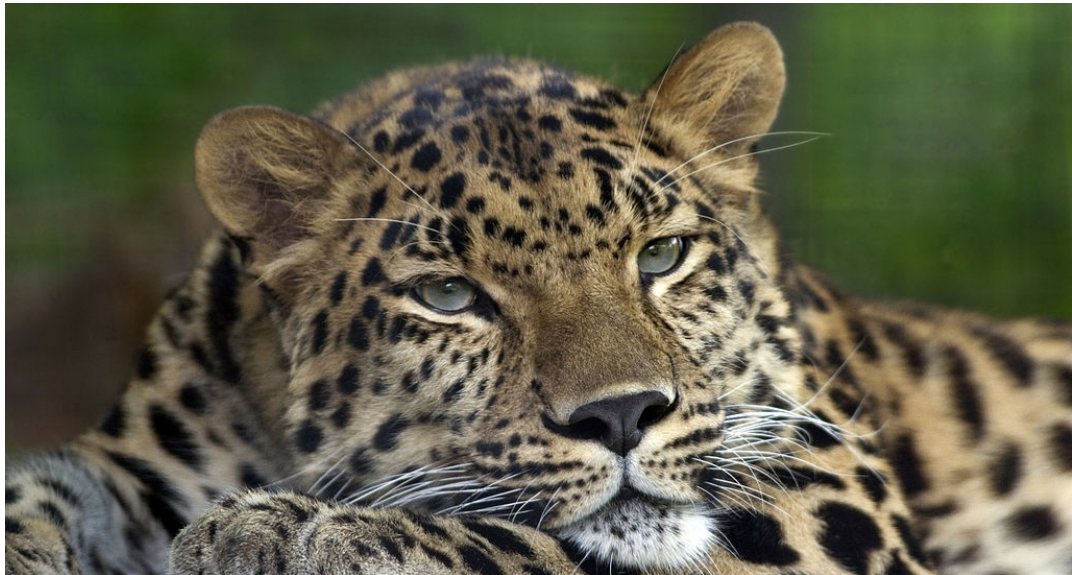


Figure 1: The big cat is ready to pounce

ML IN PRACTICE - PROCESS

1. **Data collection:** The data collection step involves gathering the learning material an algorithm will use to generate actionable knowledge. In most cases, the data will need to be combined into a single source, such as a text file, spreadsheet, or database.
2. **Data exploration and preparation:** The quality of any machine learning project is based largely on the quality of its input data. Thus,

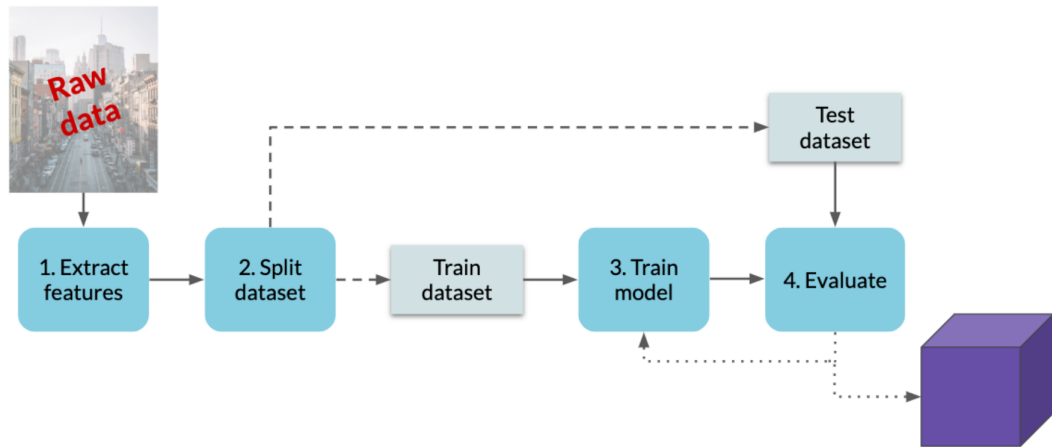


Figure 2: DataCamp "Understanding ML" course

it is important to learn more about the data and its nuances during a practice called data exploration. Additional work is required to prepare the data for the learning process. This involves fixing or cleaning so-called "messy" data, eliminating unnecessary data, and recoding the data to conform to the learner's expected inputs.

3. **Model training:** By the time the data has been prepared for analysis, you are likely to have a sense of what you are capable of learning from the data. The specific machine learning task chosen will inform the selection of an appropriate algorithm, and the algorithm will represent the data in the form of a model.
4. **Model evaluation:** Each machine learning model results in a biased solution to the learning problem, which means that it is important to evaluate how well the algorithm learned from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset, or you may need to develop measures of performance specific to the intended application.
5. **Model improvement:** If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the model's performance. Sometimes it may be necessary to switch to a different type of model altogether. You may need to supplement your data with additional data or perform additional preparatory work, as in step two of this process.

6. **Model deployment:** if the model appears to be performing well, it can be deployed for its intended task. As the case may be, you might utilize your model to provide score data for predictions (possibly in real time); for projections of financial data; to generate useful insight for marketing or research; or to automate tasks, such as mail delivery or flying aircraft. The successes and failures of the deployed model might even provide additional data to train your next-generation learner.

UNITS OF OBSERVATION VS. ANALYSIS

- A **unit of observation** is the smallest entity with measured properties of interest for a study.
- Examples: The unit of observation is in the form of persons, objects or things, transactions, time points, geographic regions, or measurements.
- Sometimes, units of observation are combined to form units, such as person-years, which denote cases where the same person is tracked over multiple years, and each person-year comprises a person's data for one year.
- When **visualizing data**, you always need to add the unit to the axis label information!
- The **unit of analysis** is the smallest unit from which inference is made. This is not always the same as the unit of observation.
- Example: data observed from people (= unit of observation) might be used to analyze trends across countries (= unit of analysis).
- Distinguish **examples**, instances of the unit of observation for which properties were recorded, and **features**, the recorded properties or attributes of examples that may be useful for ML

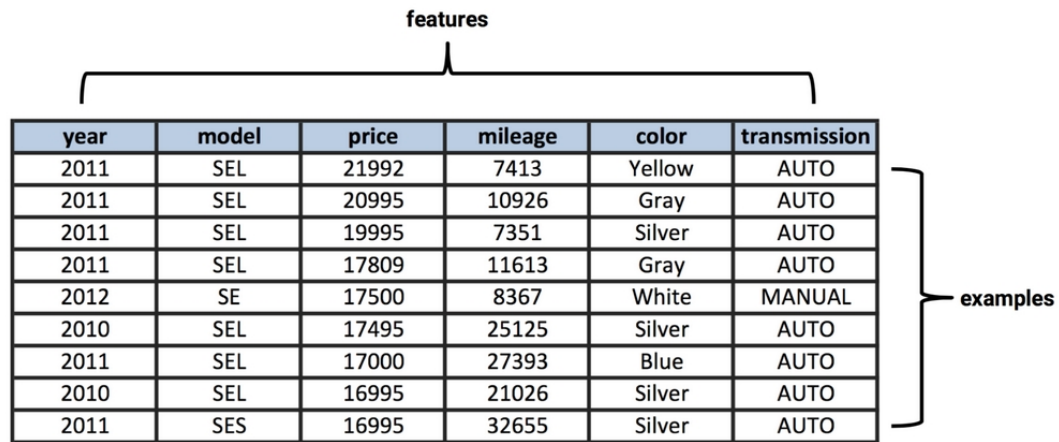
USE CASES: SPAM / CANCER DETECTION

- **Spam email identification:**
 1. Unit of observation: email messages
 2. Examples or instances: specific individual messages
 3. Features: words, punctuation marks used in the messages etc.

- **Cancer detection.**

1. Unit of observation: patients
2. Examples or instance: a random sample of cancer patients
3. Features: genomic markers from biopsied cells, and patient characteristics like weight, height, blood pressure etc.

STRUCTURED VS. UNSTRUCTURED DATA



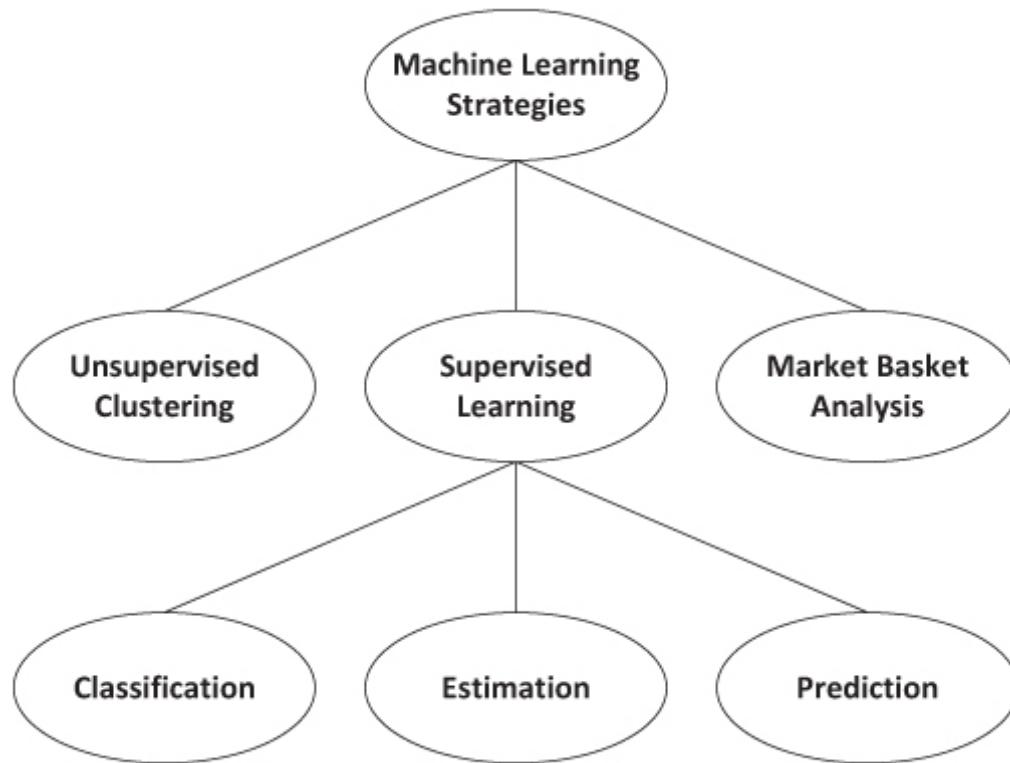
The diagram shows a table with 6 columns and 10 rows. A bracket above the columns is labeled "features". A bracket to the right of the rows is labeled "examples".

year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

Figure 3: "Examples" = vehicles for sale, "features" = car properties

- Humans can consume *unstructured* data - free-form text, pictures, sound, and they can handle cases with many or few features
- Computers required data to be *structured* - each example of the phenomenon has the same features, which are organized in data structures like tables or matrices or data frames
- In data tables, matrices or data frames, rows correspond to examples or records or observations of features, which correspond to columns
- Data entries can have different types: *numeric-discrete*, *numeric-continuous*, *categorical-nominal*, or *categorical-ordinal*
- Clarity about features, observations, and data types is crucial for selecting the best learning algorithm

TYPES OF ML ALGORITHMS



Machine learning algorithms are divided into categories according to their purpose. Understanding the categories of learning algorithms is an essential first step toward using data to drive the desired action.

PREDICTIVE MODELS-SUPERVISED LEARNING-CLASSIFICATION

- **Predictive models** involve prediction of one value using other values in the same dataset. The algorithm models the relationship between the target feature (predicted) and the other features (predictors).
- These models do not need to be forecasting models (for the future), they can also predict past events or work in real-time.

- The process of training a predictive model is called **supervised learning**. The "supervision" refers to the fact that the target values let the learner (the machine) know how well it's doing.
- Given a set of data, a **supervised learning algorithm** optimizes a **function** (the **model**) to find the combination of **feature** input values that result in the **target** output.
- **Classification** means predicting which category an example belongs to. The corresponding supervised ML algorithm is a **classifier**, e.g.
 1. An email message is spam
 2. A person has cancer
 3. A football team will win or lose
 4. An applicant will default on a loan
- The classification target feature is the **class**, which is divided into category values called **levels**, which may be nominal or ordinal
- The most widely used supervised learning algorithm for **numeric prediction**, especially forecasting, is **linear regression**
- Since discrete numbers can be converted to categories, the boundary between classification and numeric prediction models is blurry

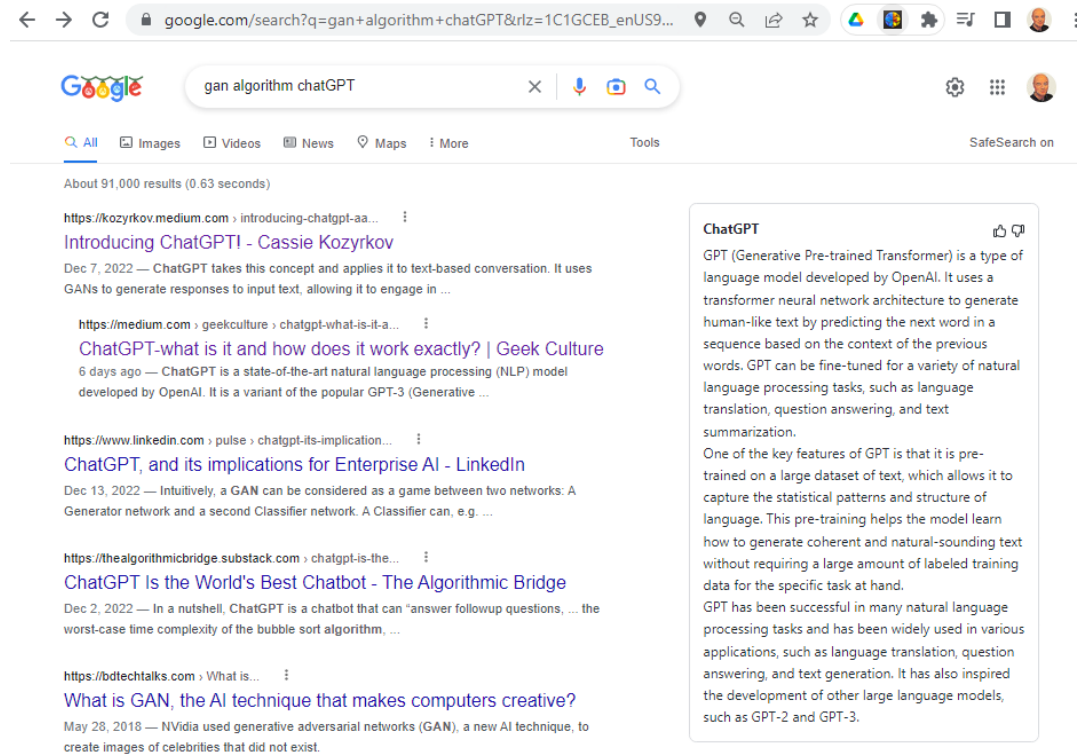
DESCRIPTIVE MODELS-UNSUPERVISED LEARNING-CLUSTERING

- **Descriptive models** are used to summarize data in new and interesting ways. No single feature is more important than any other.
- Because there is no target to be supervised, the process of training a descriptive model is called **unsupervised learning**.
- An example is **pattern discovery** in **data mining** to identify useful associations (correlations) within data.
- Application: **market basket analysis** of transactional purchase data in retail: if the retailer learns that swimming trunks are purchased at the same time as sunscreen, it could use this information when marketing both products, e.g. reposition them in the store, run a promotion etc.

- **Clustering** is descriptive modeling - it means dividing a dataset into homogenous groups. This can be used for **segmentation analysis** to identify groups of individuals with similar behavior or demographics, e.g. to create a "people like you have bought this item, too" type of promotion.

META-LEARNERS-ENSEMBLES-REINFORCEMENT LEARNING

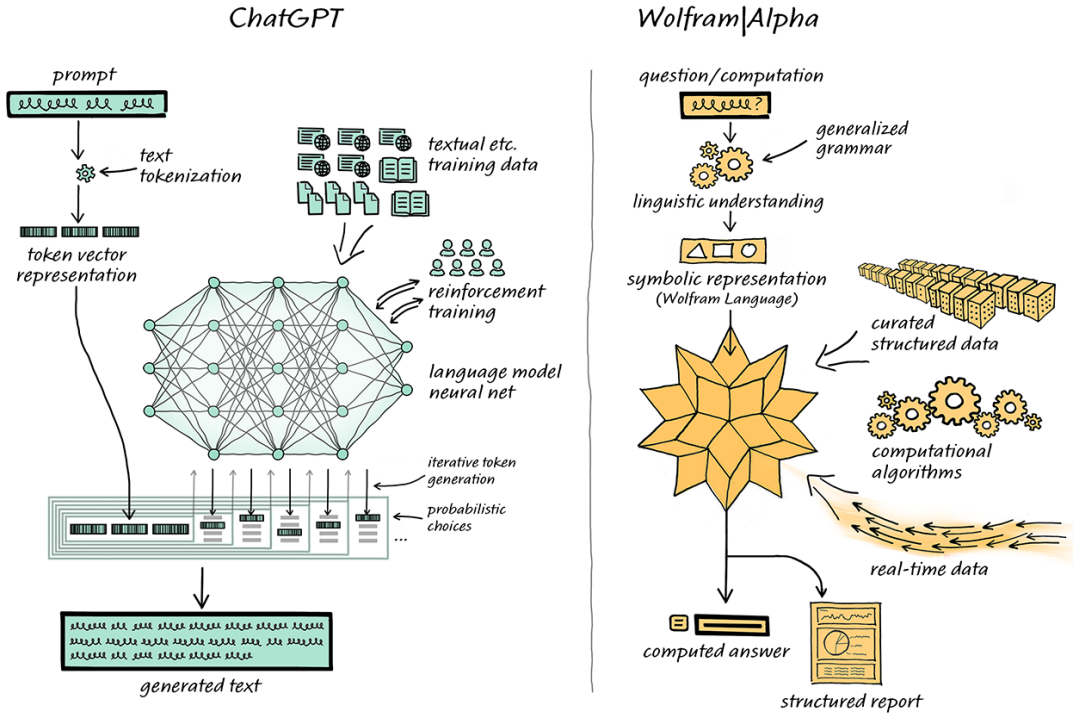
- "**Meta-learners**" are models that learn how to learn more effectively by using the result of past learning to inform additional learning
- **Ensembles** are algorithms that work in teams, and algorithms that evolve over time in a process called **reinforcement learning**
- **Adversarial learning** involves learning about a model's weaknesses in order to harden it against malicious attacks
- The popular **ChatGPT** model is a natural-language processing (NLP) variant of the GPT-3 (Generative Pretained Transformer 3) model, which was trained in massive amount of text data to generate human-like responses to a given input. #+captions; ChatGPT output next to Google.com output in browser



The image shows ChatGPT output via Google Chrome extension (right) next to "classic" Google search engine output (left)¹.

- Here is an excellent overview article (75p.) on ChatGPT's architecture by Stephen Wolfram (who created Wolfram language - not unlike R).
+captions; ChatGPT mechanics vs. Wolfram|Alpha mechanics

¹Meanwhile (Feb 8, 2023), Microsoft, the main sponsor of OpenAI, the developers of ChatGPT, have integrated the bot in their Bing search engine (Lardinois, 2023).



The image contrasts Wolfram language with ChatGPT from another article (Wolfram, 2023). By the way, I don't agree with the author's conclusion that "human language is somehow simpler and more 'law like' in its structure than we thought". I think the opposite is the case and ChatGPT shows exactly that. But the article still does its job in explaining the mechanics of neural nets and transformer technology.

ALGORITHMS

- List of Supervised Learning algorithms (Lantz, 2019):

NAME	TYPE	CH.
Naive Bayes	Classification	4
Decision trees	Classification	5
Linear regression	Numeric prediction	6
Regression trees	Numeric prediction	6
Model trees	Numeric prediction	6
Neural networks	Dual use	7
Support Vector Machines	Dual use	7

- List of Unsupervised Learning algorithms:

NAME	TYPE	CH.
Association rules	Pattern detection	8
k-means clustering	Clustering	9

- Meta-learning algorithms:

NAME	TYPE	CH.
Bagging	Dual use	11
Boosting	Dual use	11
Random forests	Dual use	11

ML WITH R - R PACKAGES

- R is free, open source software (FOSS) for statistical programming
- Many ML algorithms must be installed on top of base R as packages
- Both base R and packages can be obtained from CRAN, the Comprehensive R Archive Network (CRAN), at cran.r-project.org
- There is a separate *task view* for ML on CRAN

CRAN Task View: Machine Learning & Statistical Learning

Maintainer: Torsten Hothorn

Contact: [Torsten.Hothorn at R-project.org](mailto:Torsten.Hothorn@R-project.org)

Version: 2022-03-07

URL: <https://CRAN.R-project.org/view=MachineLearning>

Source: <https://github.com/cran-task-views/MachineLearning/>

Contributions: Suggestions and improvements for this task view are very welcome and can be made through issues or pull requests on GitHub or via e-mail to the maintainer address. For further details see the [Contributing guide](#).

Citation: Torsten Hothorn (2022). CRAN Task View: Machine Learning & Statistical Learning. Version 2022-03-07. URL <https://CRAN.R-project.org/view=MachineLearning>.

Installation: The packages from this task view can be installed automatically using the [ctv](#) package. For example, `ctv::install.views("MachineLearning", coreOnly = TRUE)` installs all the core packages or `ctv::update.views("MachineLearning")` installs all packages that are not yet installed and up-to-date. See the [CRAN Task View Initiative](#) for more details.

THE RWeka PACKAGE

- RWeka was developed by Hornik et al (2009). See here for more information on `weka` - you also need to have Java installed
- When installing the package with `install.packages`, required *dependencies* (other packages) will also be installed
- When installing, pick a mirror near you for greater download speed or (better) put "`https://cloud.r-project.org/`" into your `.Rprofile`
- The *default* location will be announced at the end of the install, or your system may ask you to specify a location (accept the default)
- You could also specify a location to install using the `lib` parameter:

```
> install.packages("RWeka", lib = "/path/to/library")
```

- To load the package, use the `library` function. To see it in the work environment, use `search()`, and to detach it from the current session, use `detach`:

```
library(RWeka)
search()
detach("package:RWeka", unload=TRUE)
search()
```

```
[1] ".GlobalEnv"      "package:RWeka"      "ESSR"
[4] "package:stats"    "package:graphics"   "package:grDevices"
[7] "package:utils"    "package:datasets"   "package:methods"
[10] "Autoloads"        "package:base"
[1] ".GlobalEnv"      "ESSR"                "package:stats"
[4] "package:graphics" "package:grDevices"   "package:utils"
[7] "package:datasets" "package:methods"     "Autoloads"
[10] "package:base"
```

THE RStudio IDE

- RStudio is an additional interface to R available at <https://www.rstudio.com>
- RStudio includes:

1. an integrated code editor
 2. an R command-line console
 3. a file browser
 4. code output, plot, graphics
 5. project and package management
 6. integration with source / version control tools
 7. database connection management
 8. compilation of R output to HTML, PDF, WORD
- RStudio Notebook formats allow for literate programming

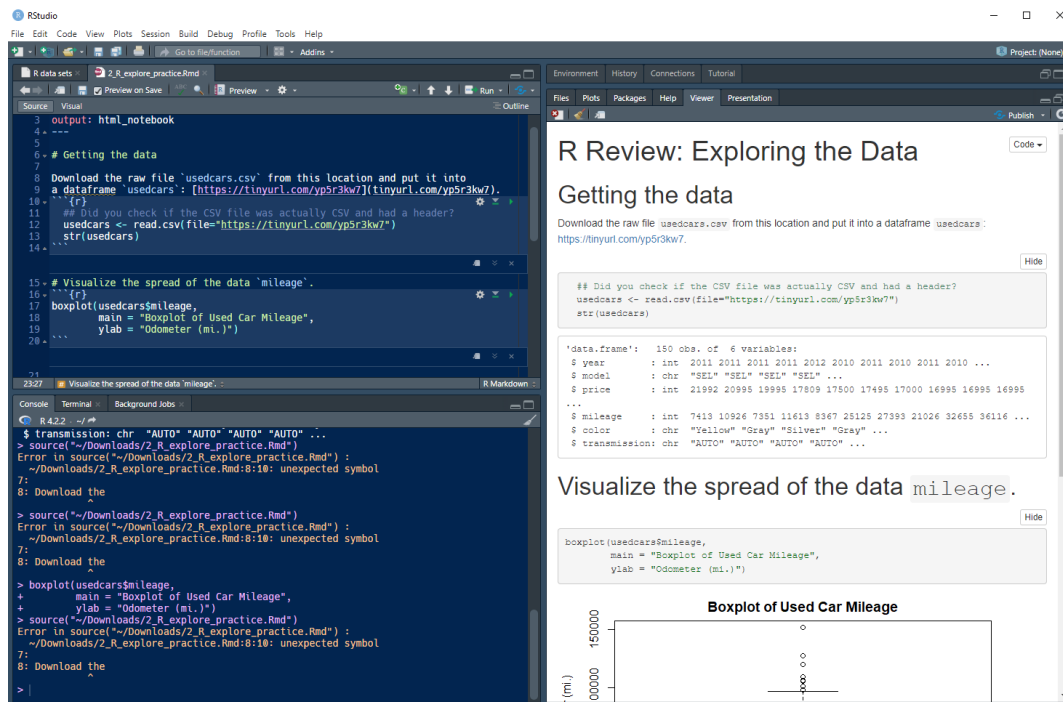


Figure 4: RStudio implementation of an R practice file

SUMMARY

- The ML model is used for prescriptive or descriptive purposes

- ML purposes can be: category classification, numeric prediction, pattern detection, and clustering
- Algorithms are chosen based on input data and learning task
- R supports ML through community-authored, FOSS packages that need to be installed and loaded

REFERENCES

- Anderson (2017). Twenty years on from Deep Blue vs Kasparov: how a chess match started the big data revolution. @theconversation.com.
- Hosseini, Z., Hytönen, K., & Kinnunen, J. (2022). Improving Online Content Quality Through Technological Pedagogical Content Design (TPCD). In S. Vachkova, & S. S. Chiang (Eds.), Education and City: Quality Education for Modern Cities, vol 3. European Proceedings of Educational Sciences (pp. 284-296). European Publisher. <https://doi.org/10.15405/epes.22043.25>
- Lantz (2019). Machine Learning with R. Packt.
- Lardinois (February 8, 2023). Hands-on with Bing's new ChatGPT-like features. Online: techcrunch.com.
- Roiger (2020). Just Enough R!. CRC Press.
- Serrano (2021). Grokking Machine Learning.