Organising Your Work

Data Science Workflows

Dr Zak Varty

Week 1: What are we trying to do?

- Data science is a collaborative discipline
- Be a good collaborator, to others and to your future self
- This week will show one framework to help you with that task
- Like flossing not difficult but requires discipline
- We will take an opinionated and R focused approach, ideas transfer to other settings.

One Project = One Directory

- Sounds easy, but in practice it is not.
- Requires prospective project scoping.
- Entropy is ever increasing.

1 project = 1 directory

Properties of a well organised project

Ideally, we would like to organise our projects so that they are:

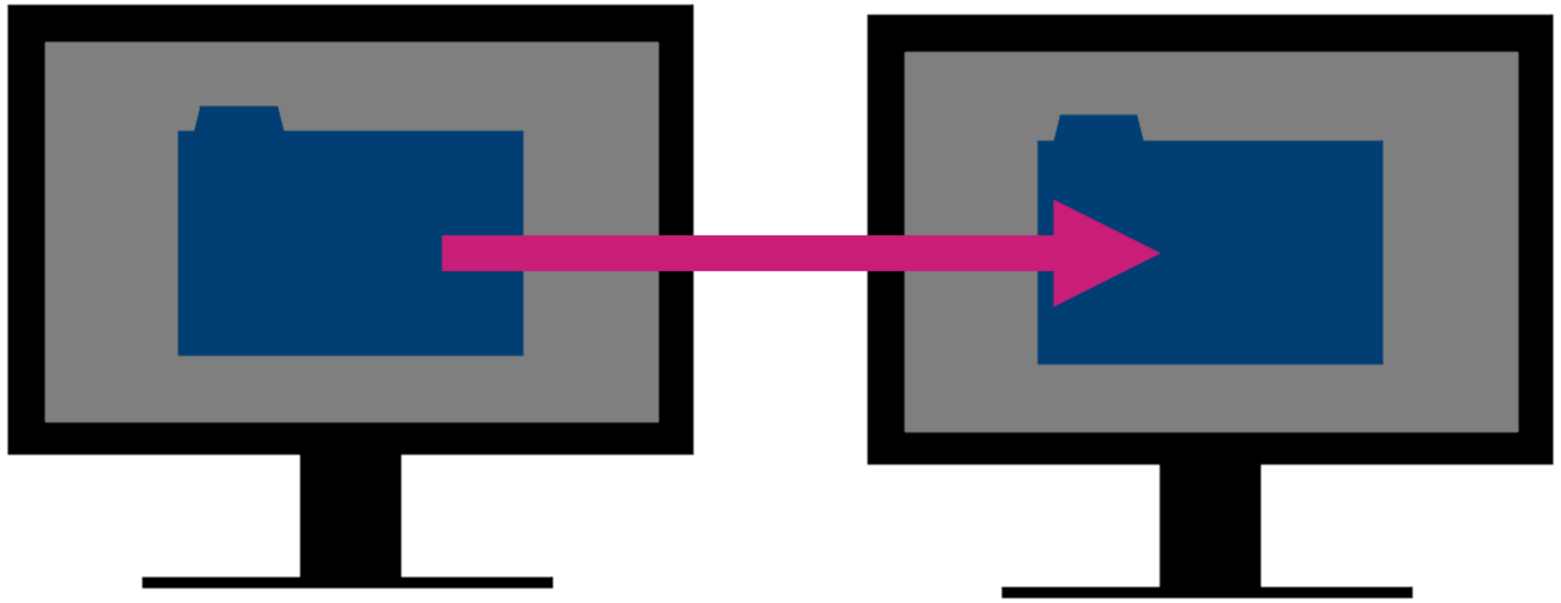
- Portable
- Version control friendly
- Reproducible
- Integrated Development Environment friendly

Portability

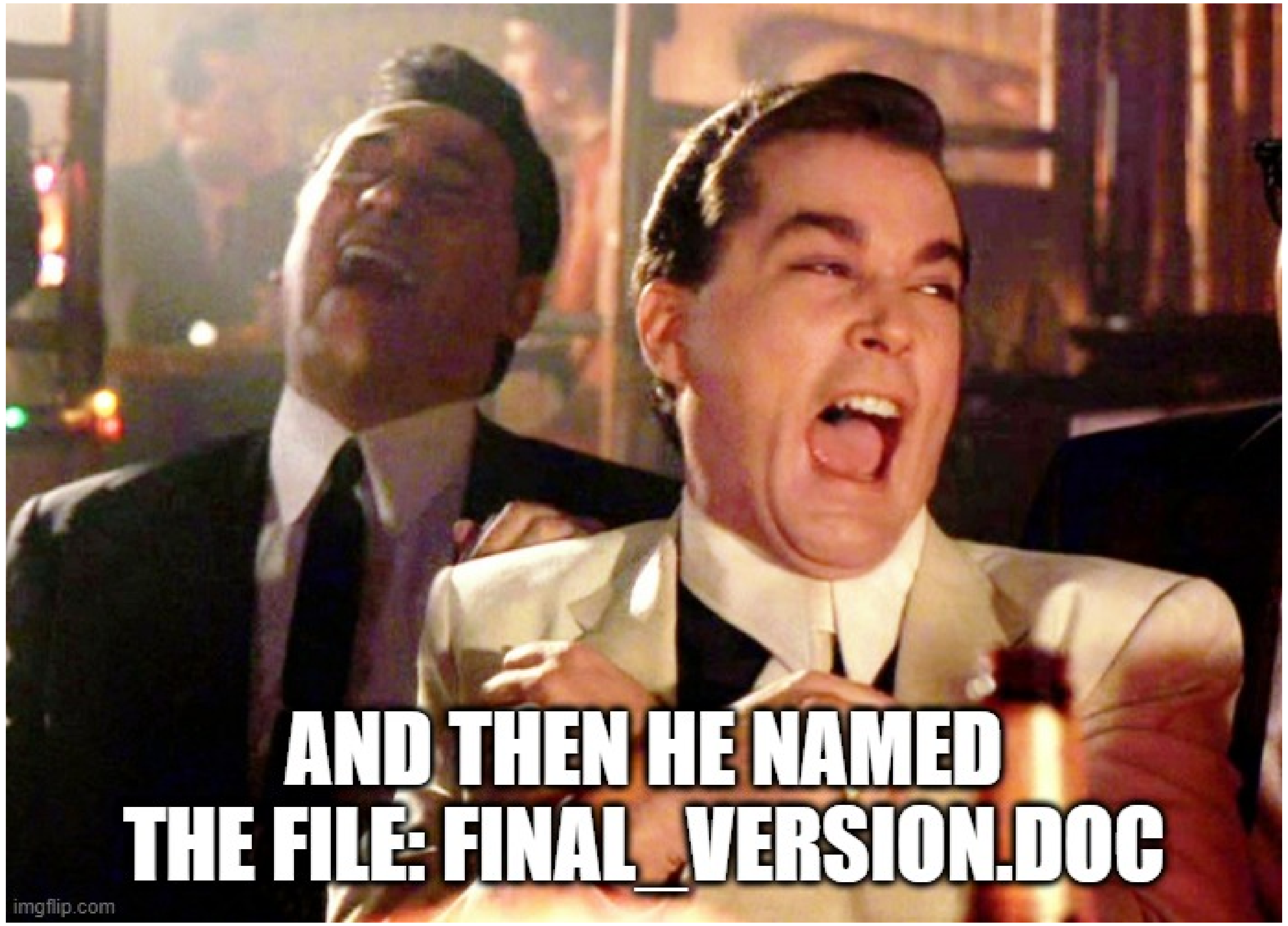
Is your work all in one place or scattered?

Can it be moved to a new location without breaking?

What do we mean by a new location, exactly?



Version Control Friendly



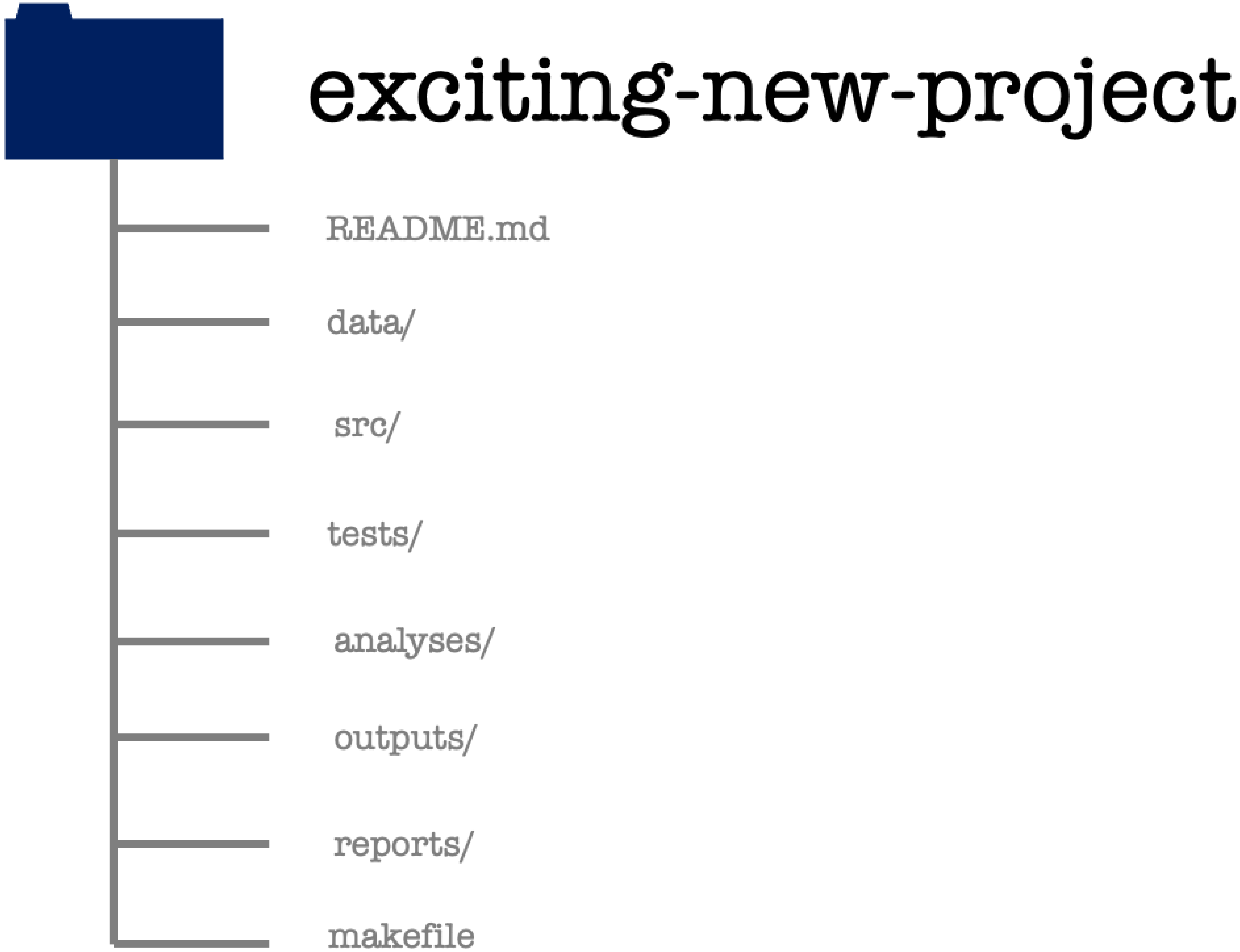
Reproducible

> A study is **reproducible** if you can take the original data and the computer code used to analyze the data and recreate all of the numerical findings from the study.
>
> Broman et al. (2017) "Recommendations to Funding Agencies for Supporting Reproducible Research"

Integrated Development Environment Friendly

- Possible to code and manage projects entirely in notepad or at the command line.
- Puts a lot of strain on you, both your fingers and your brain
- Integrated development environments such as RStudio, PyCharm and VisualStudio aim to reduce this burden
  - Autocomplete / shortcodes
  - Environment panes
  - Templating

exciting-new-project



exciting-new-project

README.md

data/

src/

tests/

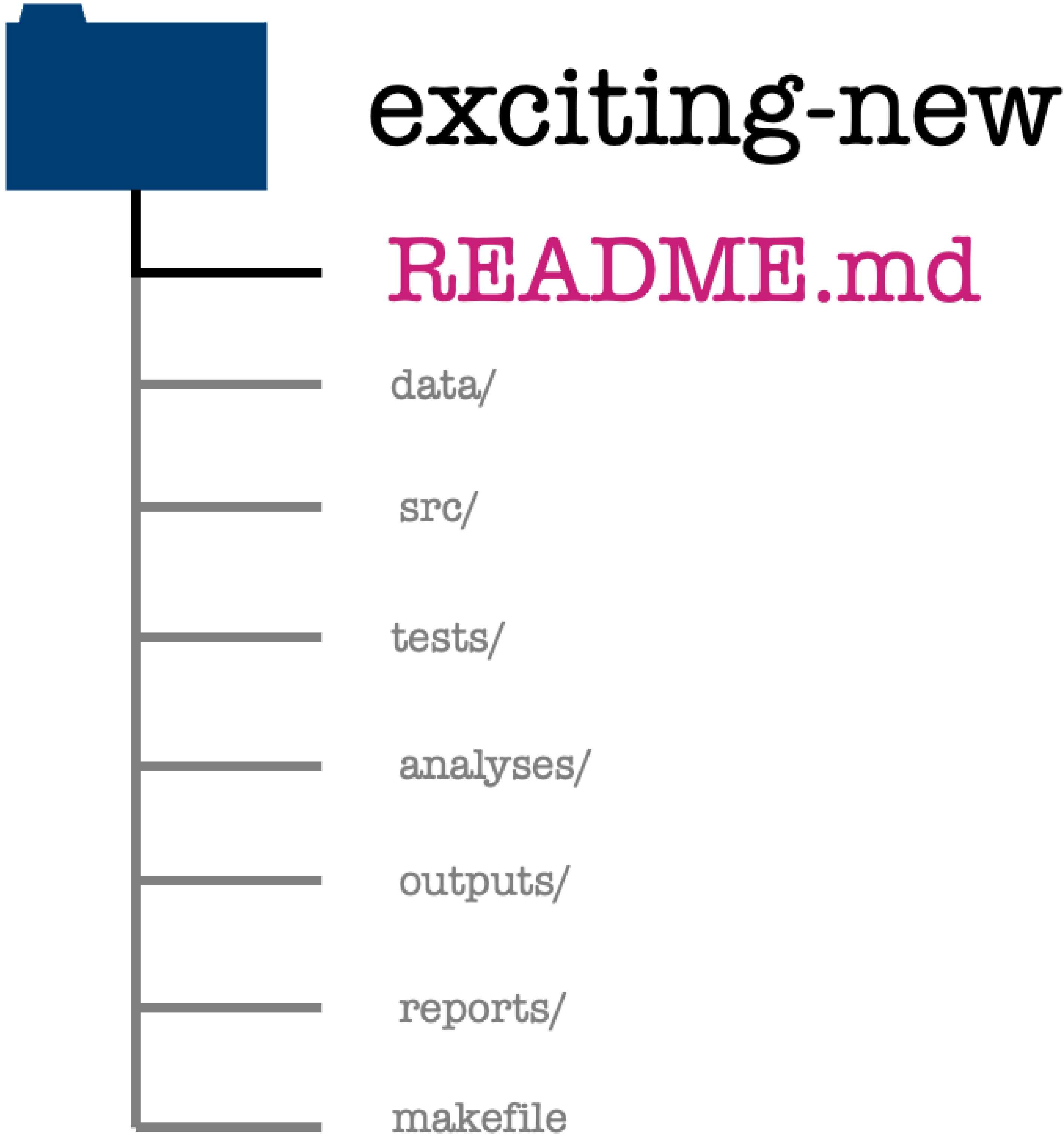
analyses/

outputs/

reports/

makefile

exciting-new-project



README.md

data/

src/

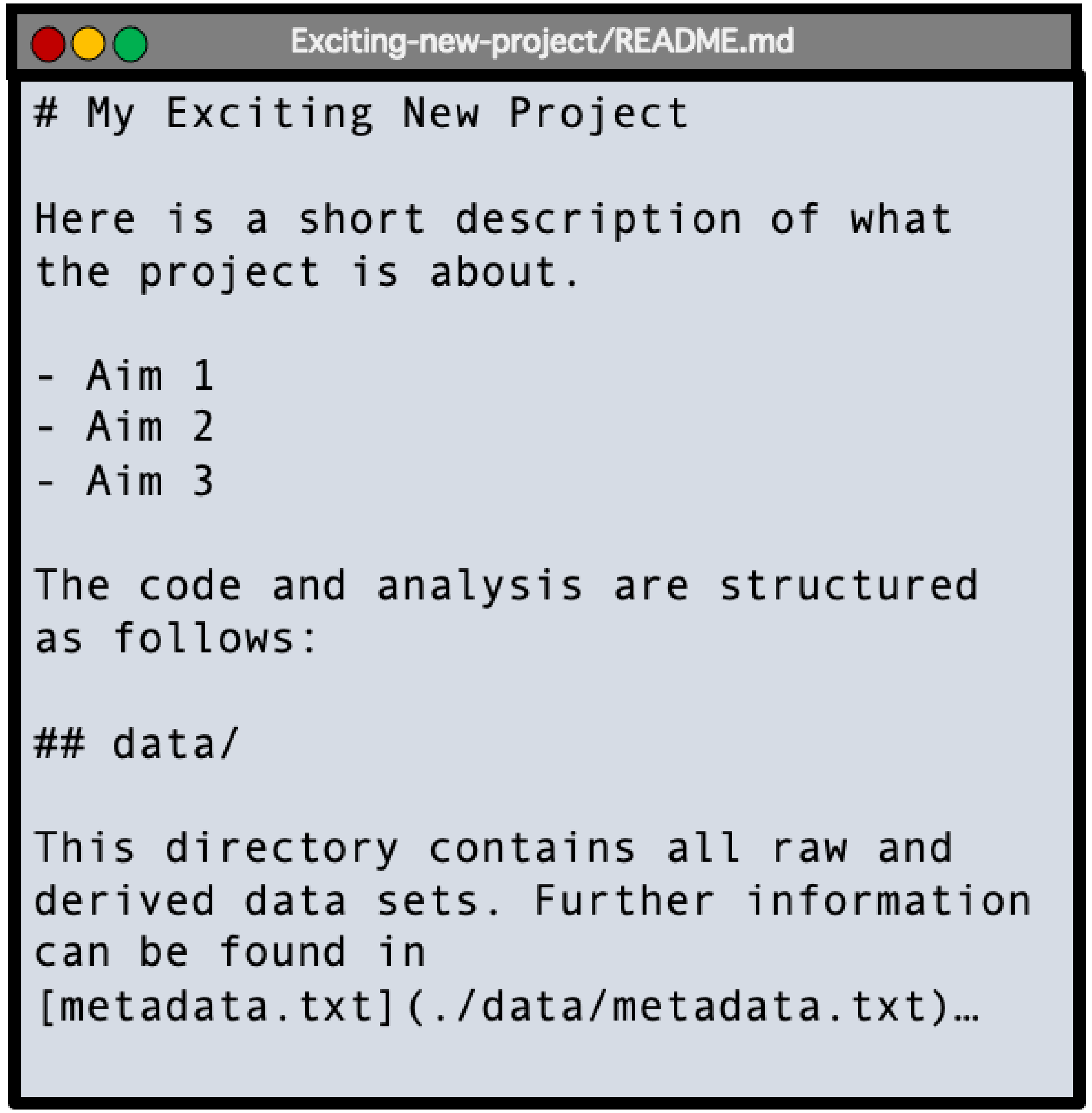
tests/

analyses/

outputs/

reports/

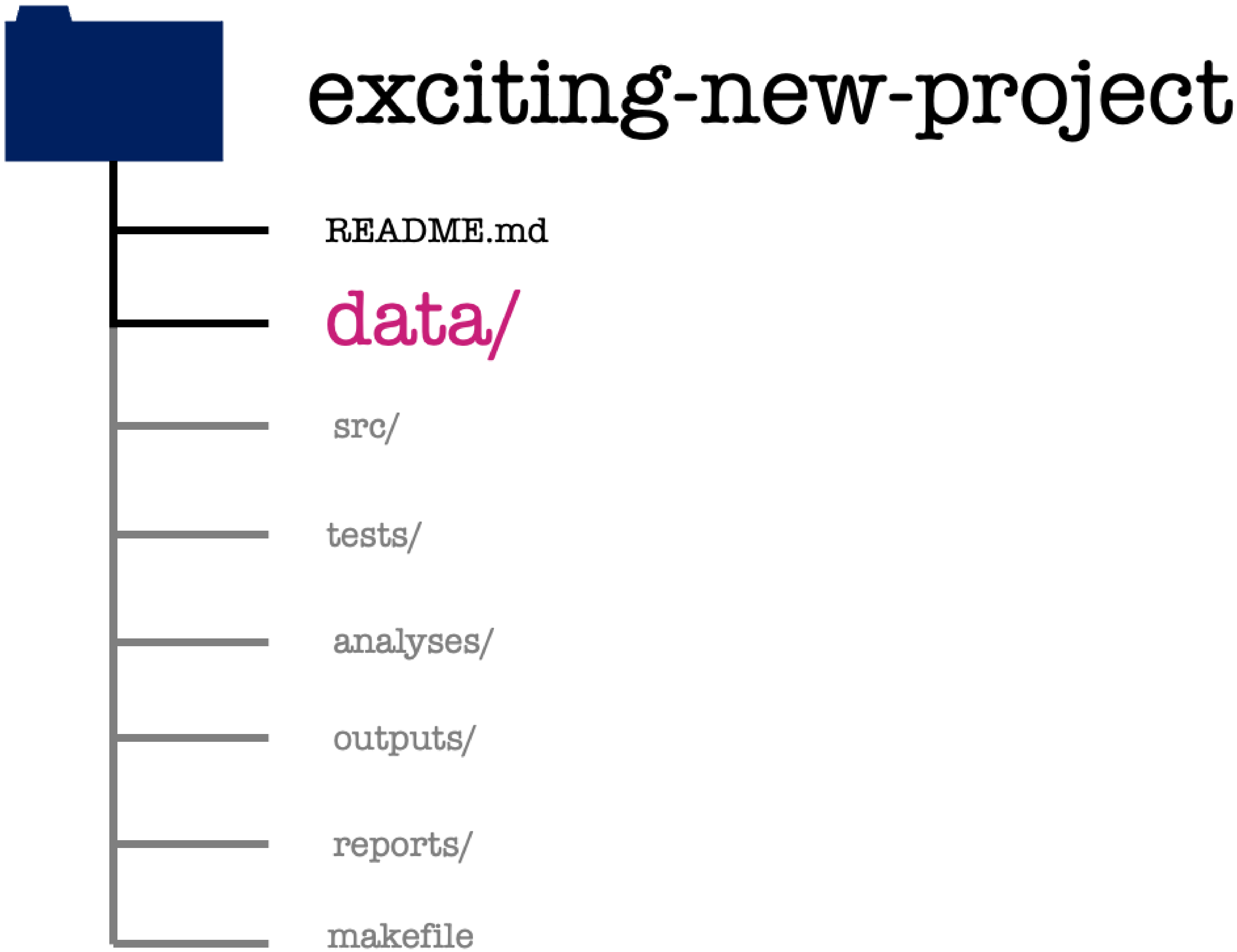
makefile



```
# My Exciting New Project

Here is a short description of what
the project is about.

- Aim 1
- Aim 2
- Aim 3

The code and analysis are structured
as follows:

## data/

This directory contains all raw and
derived data sets. Further information
can be found in
[metadata.txt](./data/metadata.txt)…
```

Inside the README

- Name
- Project Status
- Description
- Installation
- Usage / Examples
- Support
- Contributing



exciting-new-project

README.md

data/

src/

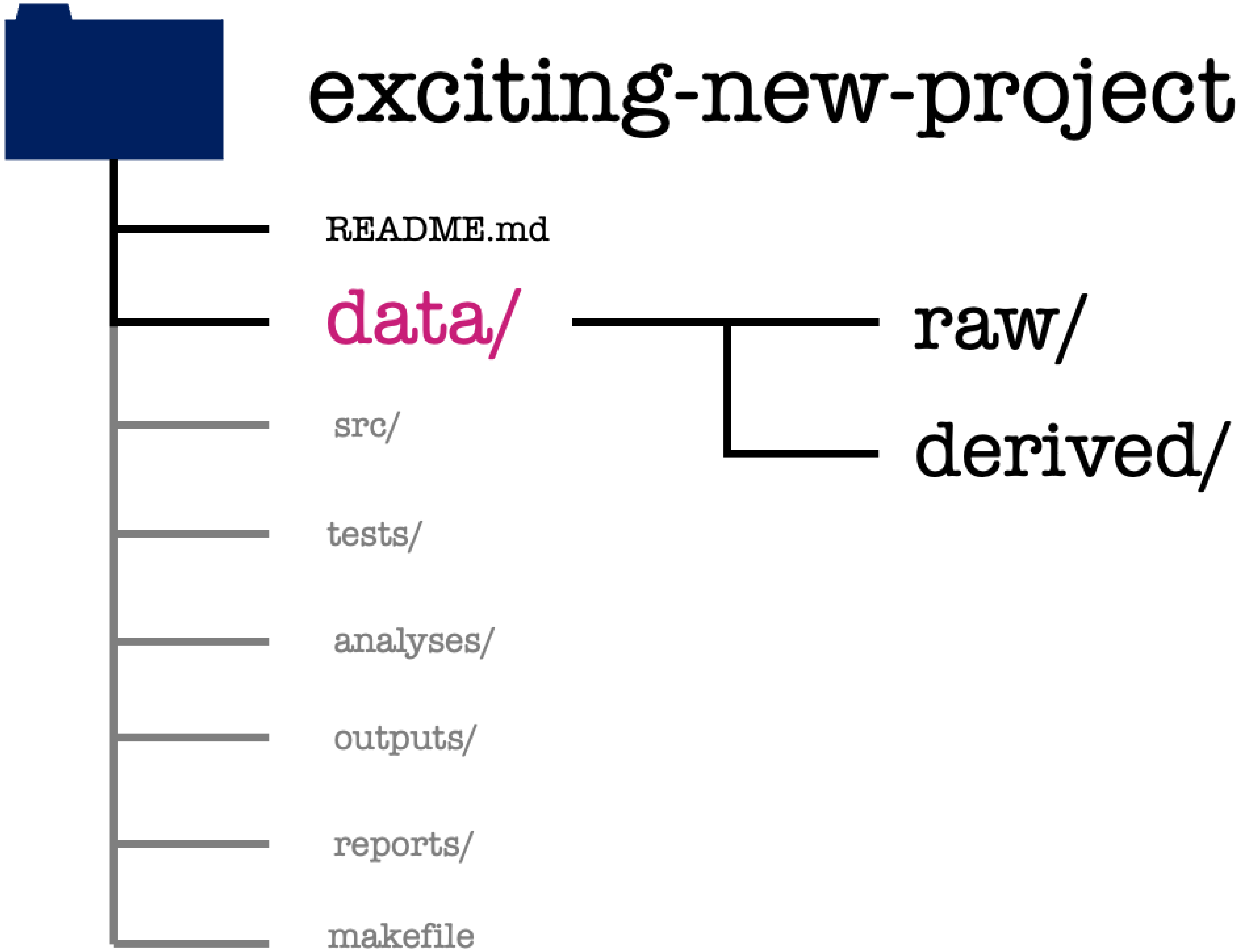
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

derived/

src/

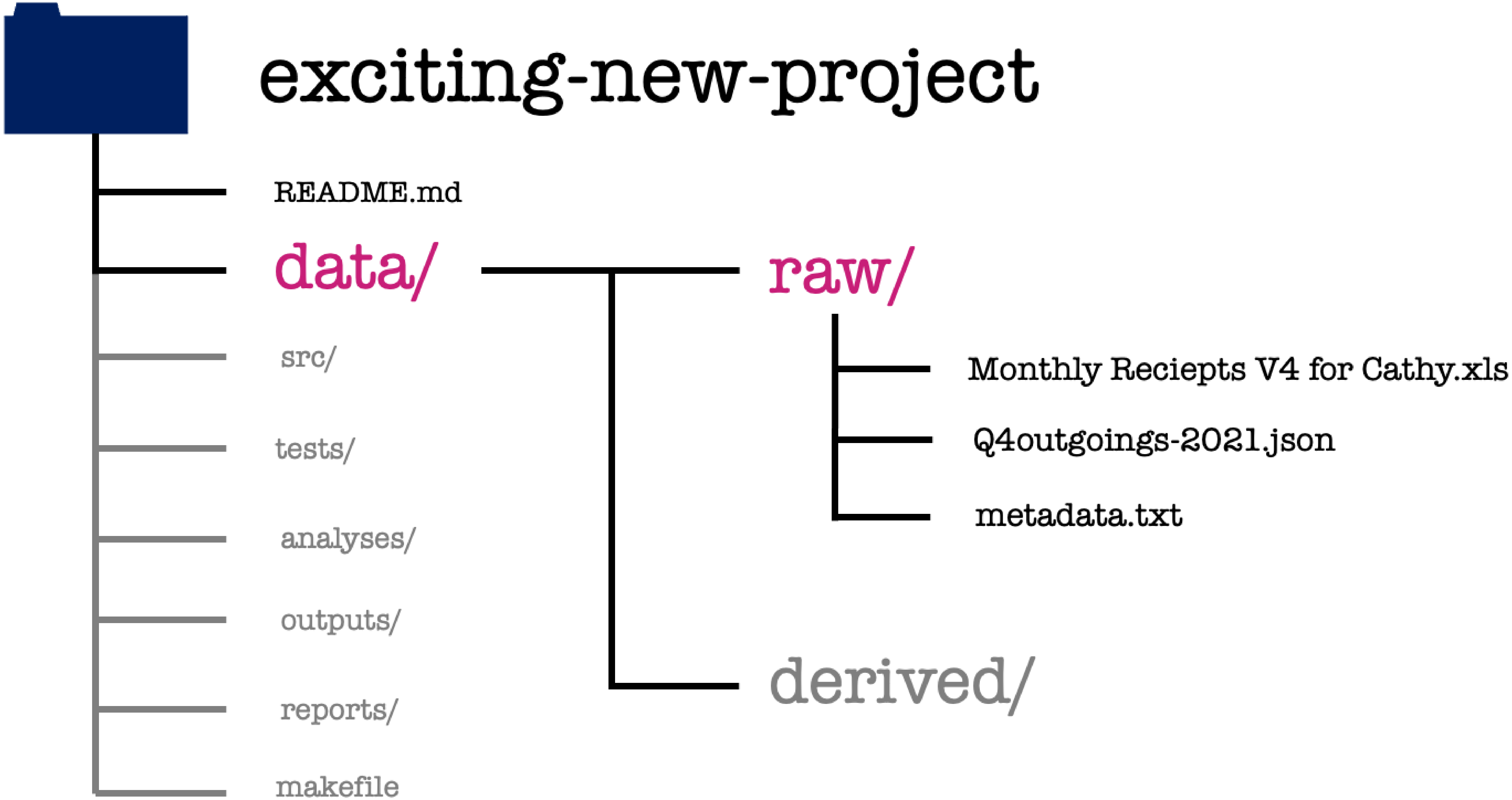
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

Monthly Reciepts V4 for Cathy.xls

Q4outgoings-2021.json

metadata.txt

derived/

src/

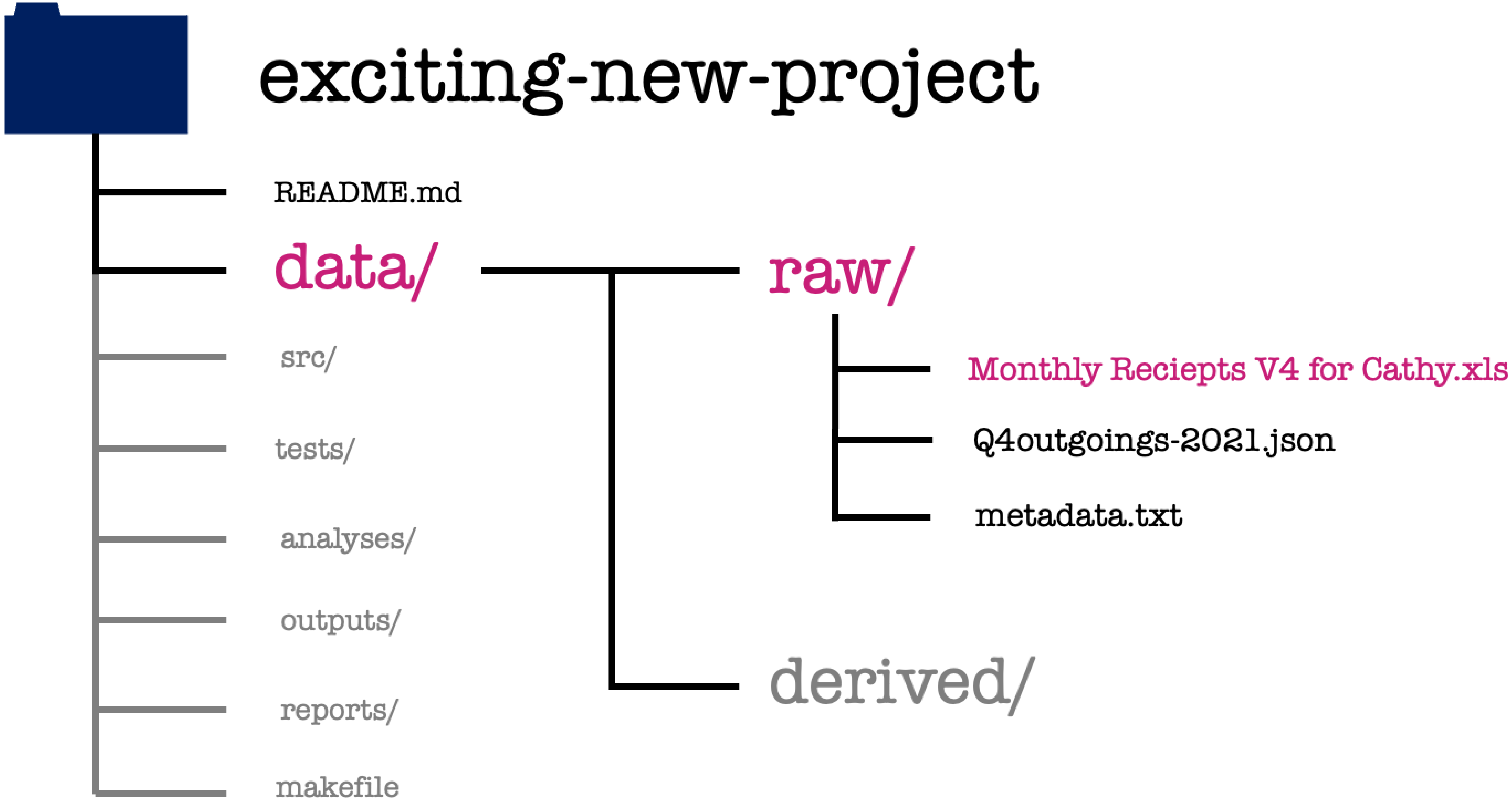
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

Monthly Reciepts V4 for Cathy.xls

Q4outgoings-2021.json

metadata.txt

derived/

src/

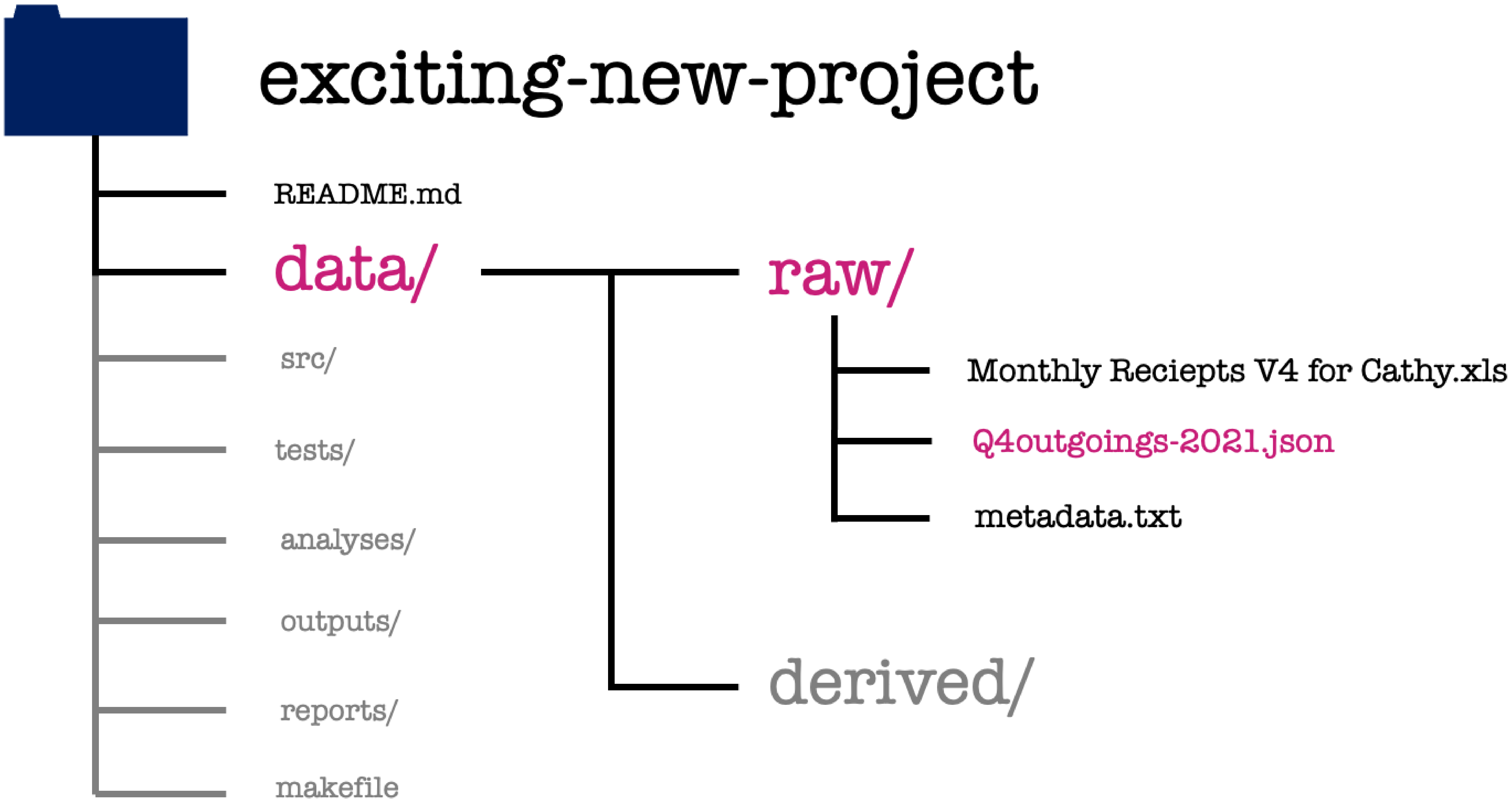
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

Monthly Reciepts V4 for Cathy.xls

Q4outgoings-2021.json

metadata.txt

derived/

src/

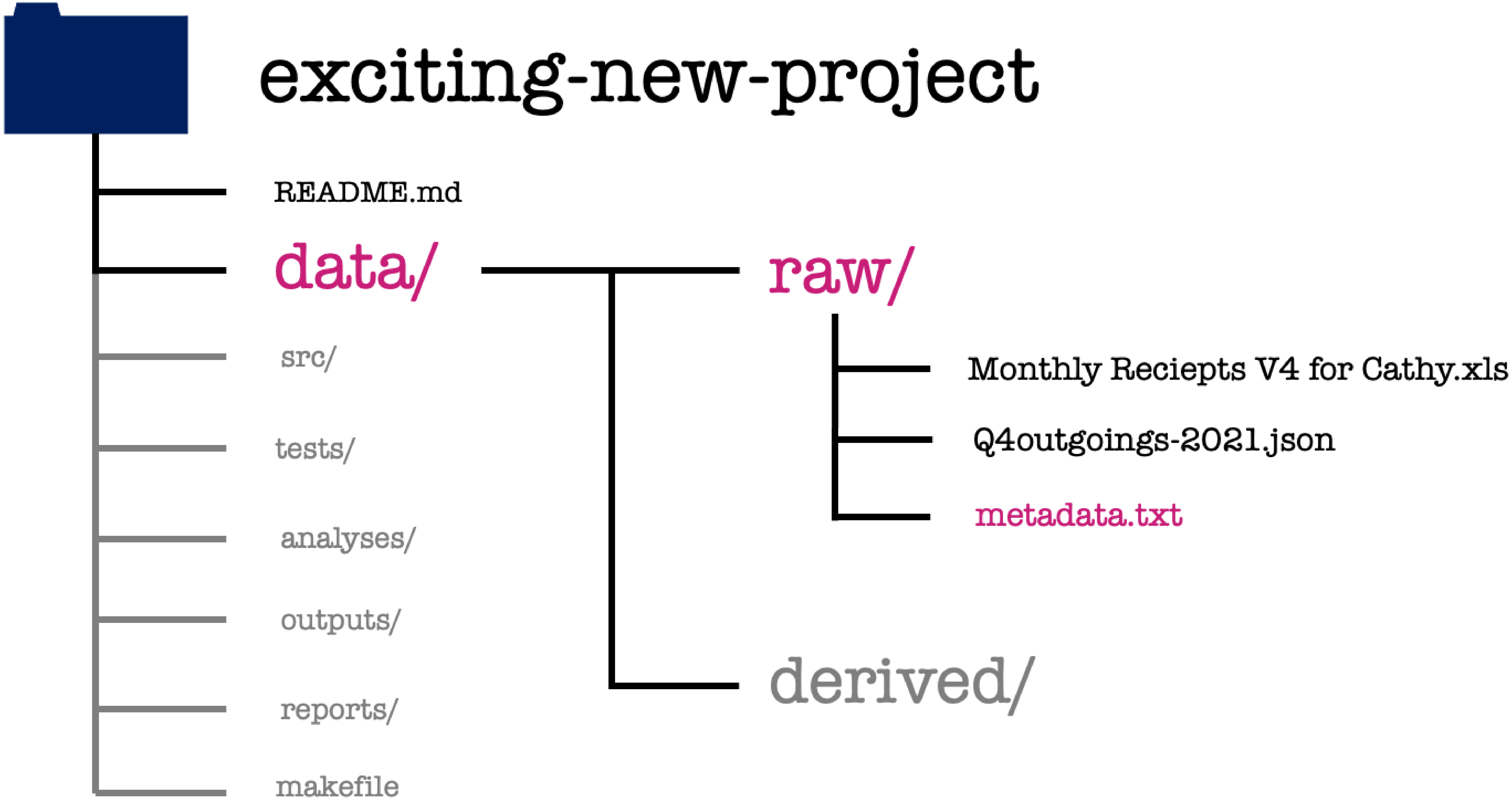
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

Monthly Reciepts V4 for Cathy.xls

Q4outgoings-2021.json

metadata.txt

derived/

src/

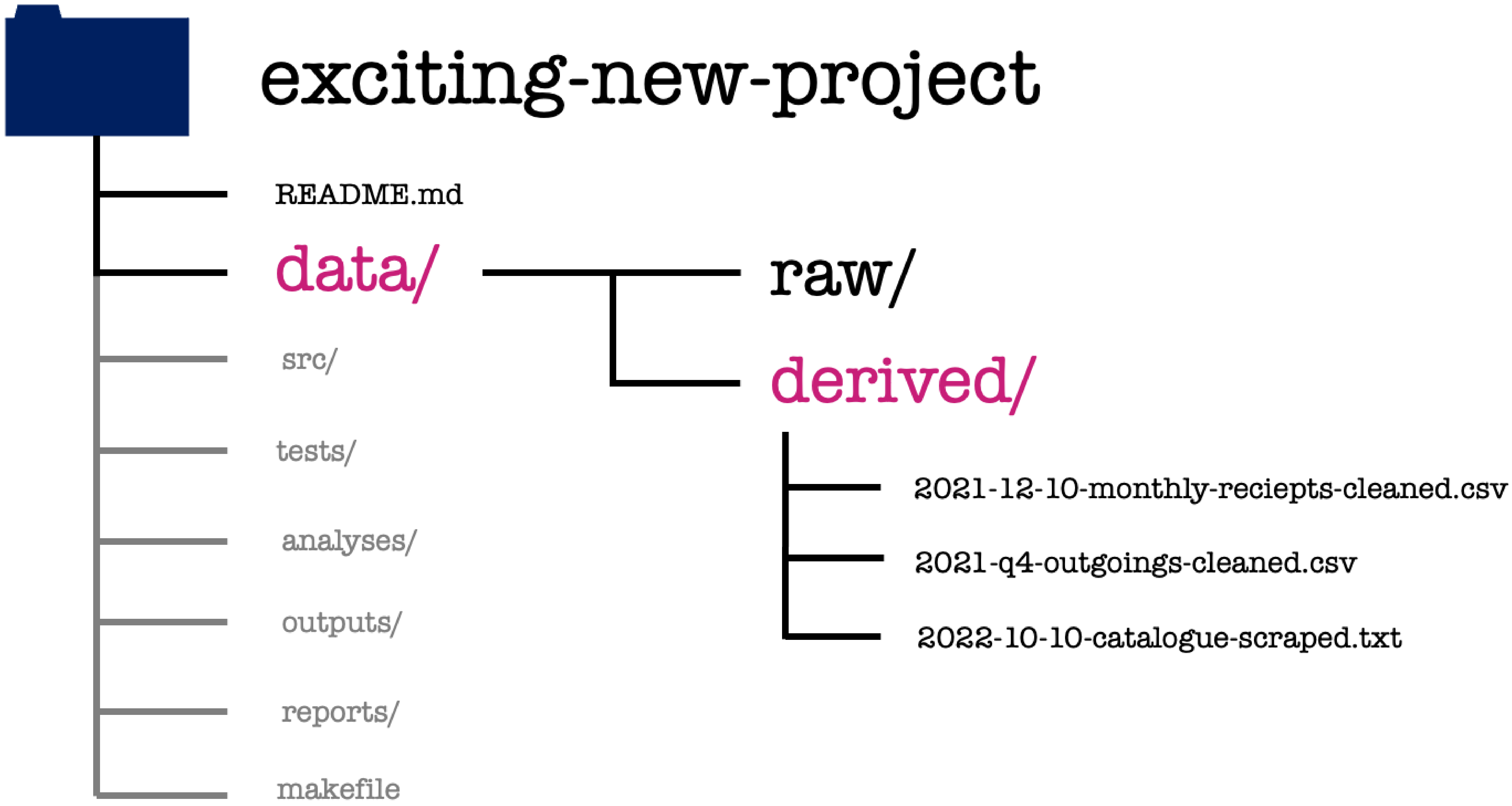
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

derived/

2021-12-10-monthly-reciepts-cleaned.csv

2021-q4-outgoings-cleaned.csv

2022-10-10-catalogue-scraped.txt

src/

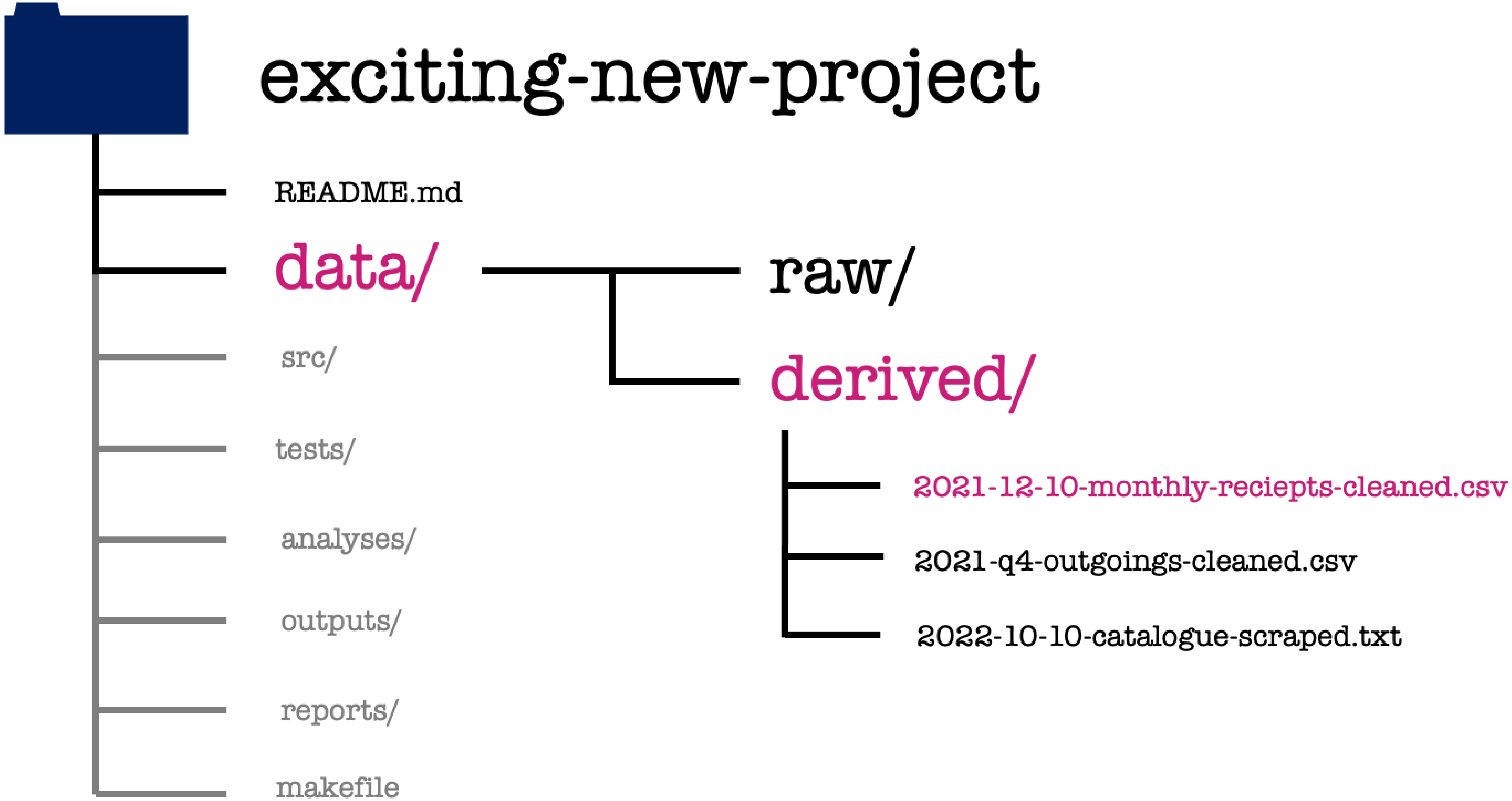
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

raw/

derived/

2021-12-10-monthly-reciepts-cleaned.csv

2021-q4-outgoings-cleaned.csv

2022-10-10-catalogue-scraped.txt

src/

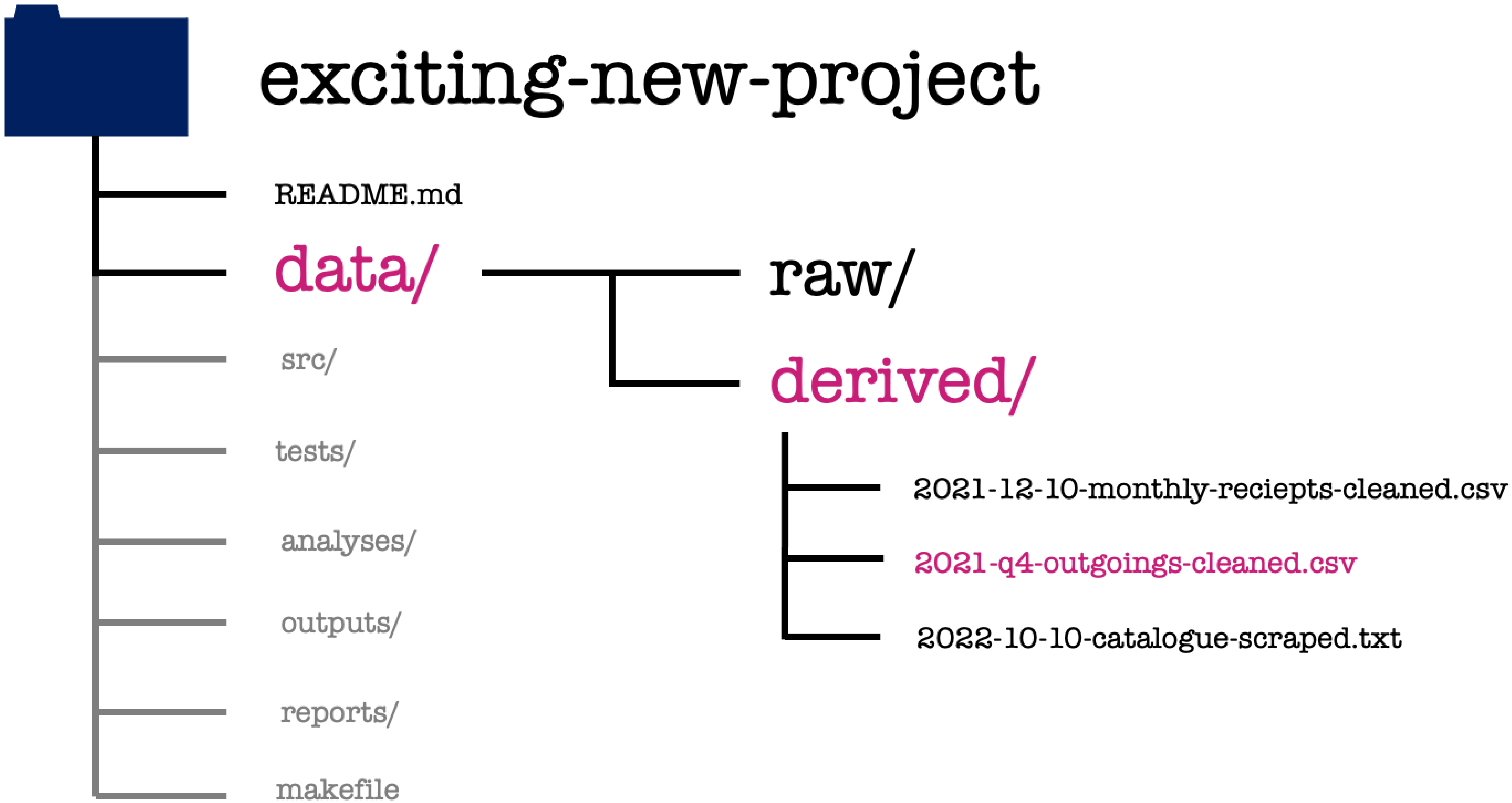
tests/

analyses/

outputs/

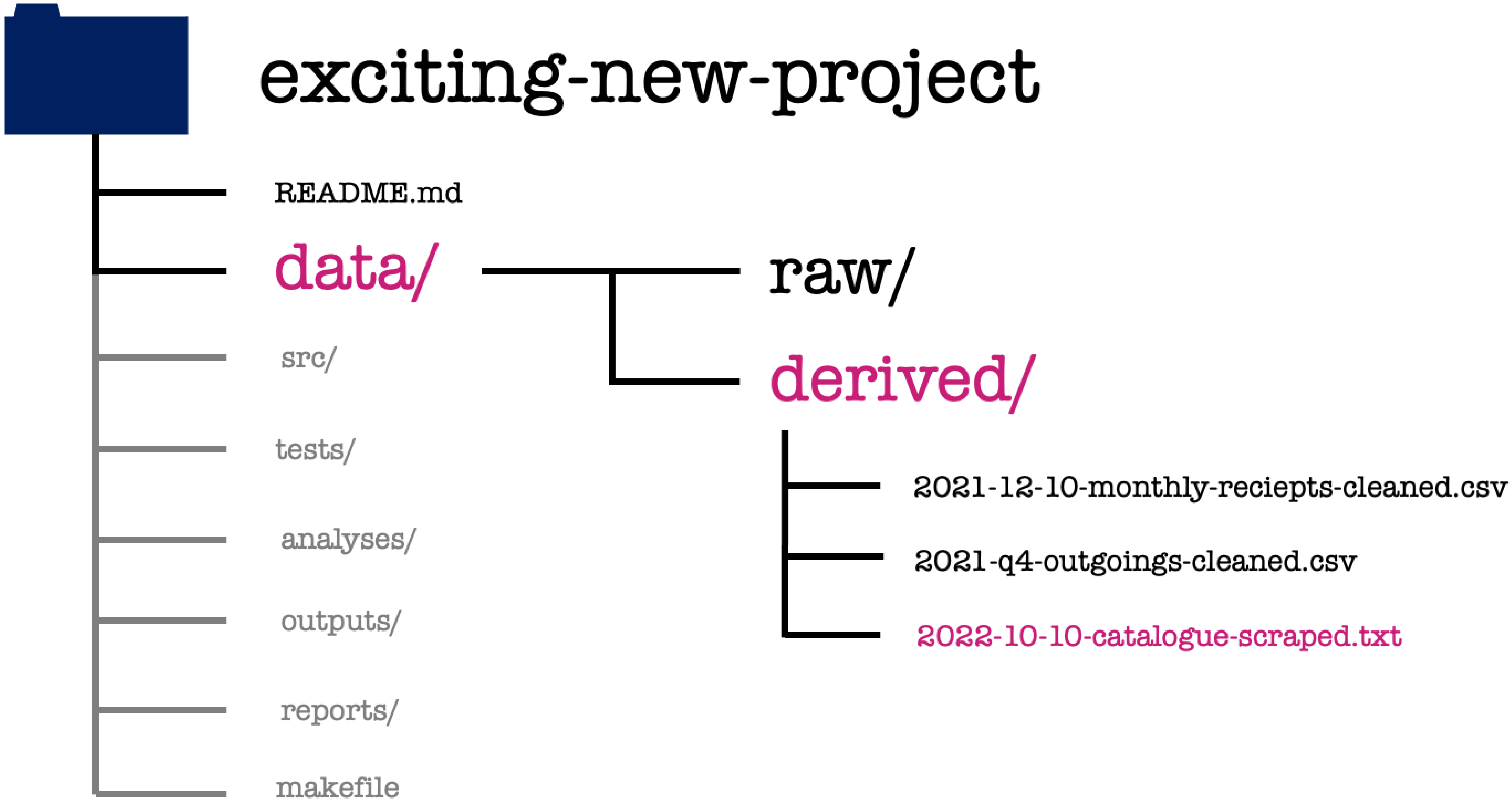
reports/

makefile



exciting-new-project

- README.md
- data/
  - raw/
  - derived/
    - 2021-12-10-monthly-reciepts-cleaned.csv
    - 2021-q4-outgoings-cleaned.csv
    - 2022-10-10-catalogue-scraped.txt
- src/
- tests/
- analyses/
- outputs/
- reports/
- makefile



exciting-new-project

README.md

data/

raw/

derived/

2021-12-10-monthly-reciepts-cleaned.csv

2021-q4-outgoings-cleaned.csv

2022-10-10-catalogue-scraped.txt

src/

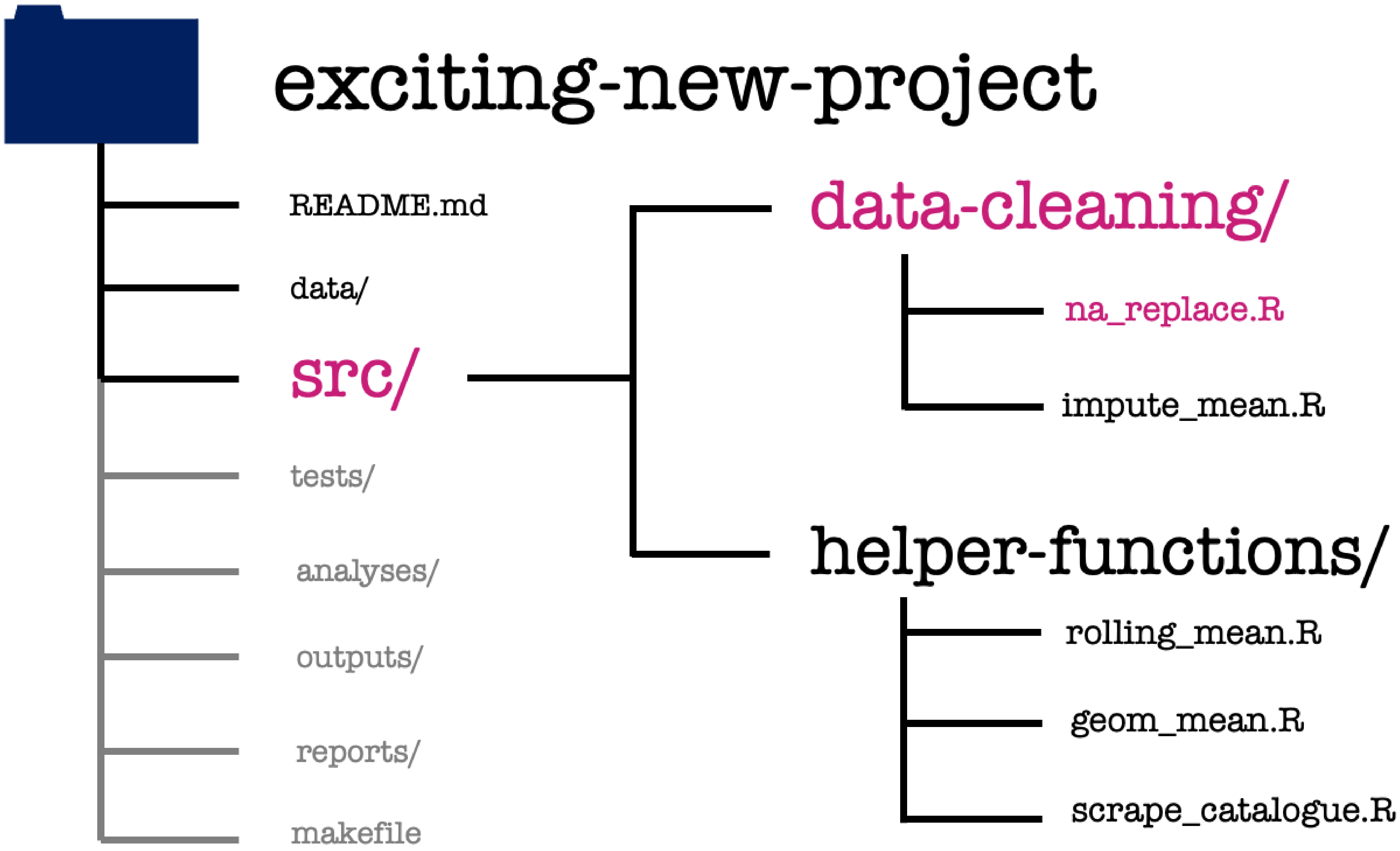
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

src/

data-cleaning/

na_replace.R

impute_mean.R

helper-functions/

rolling_mean.R

geom_mean.R

scrape_catalogue.R

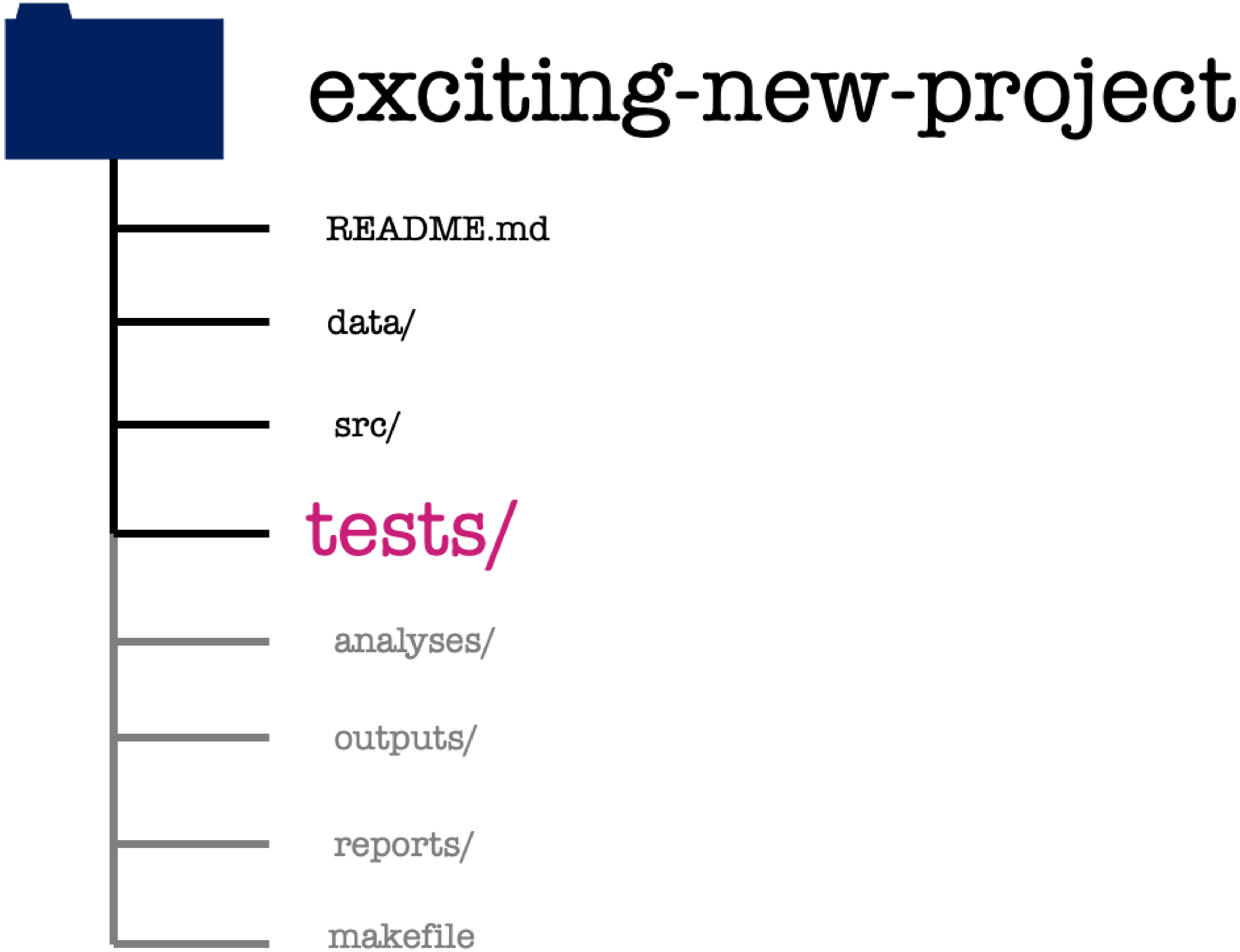
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

src/

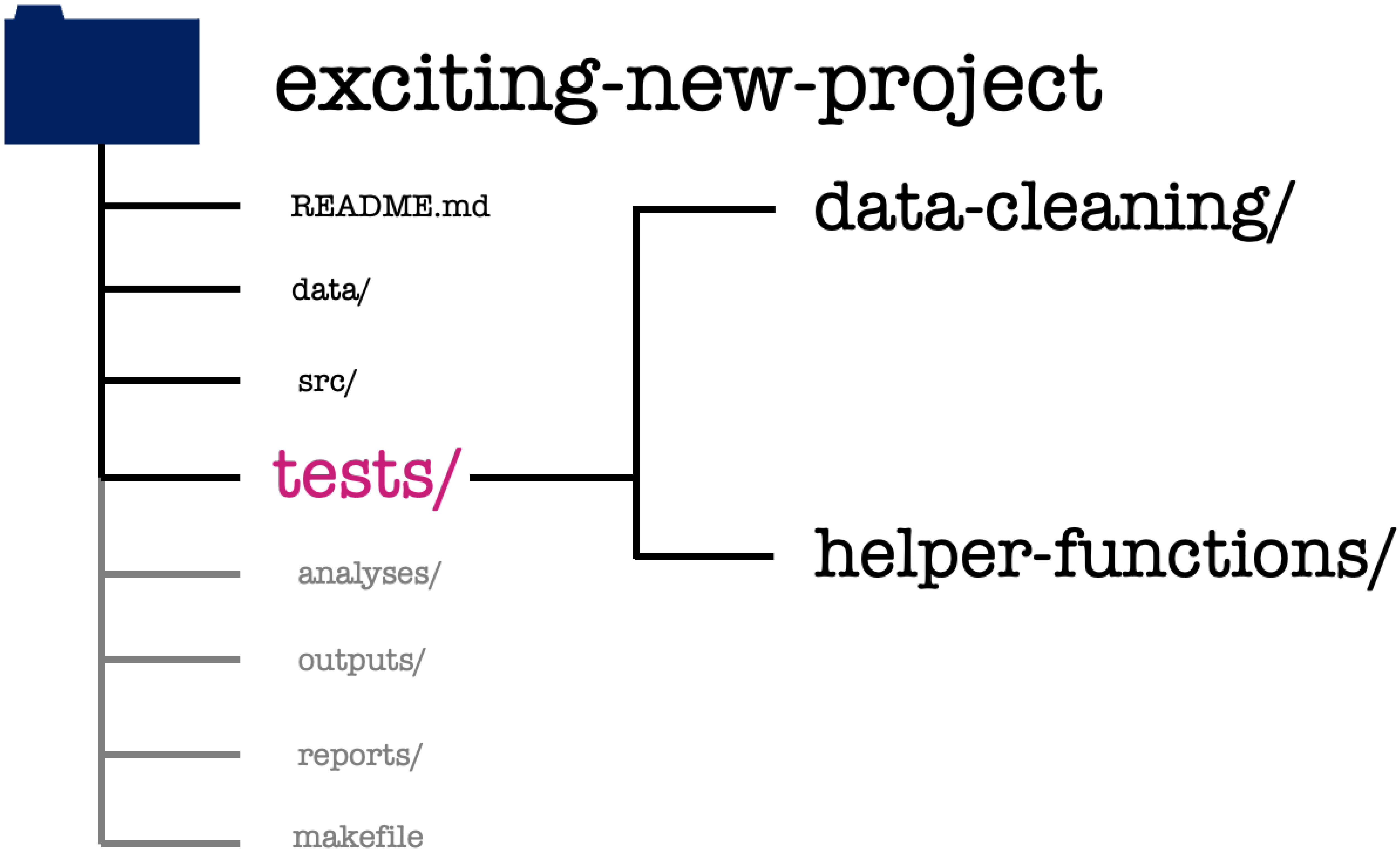
tests/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

src/

tests/

data-cleaning/

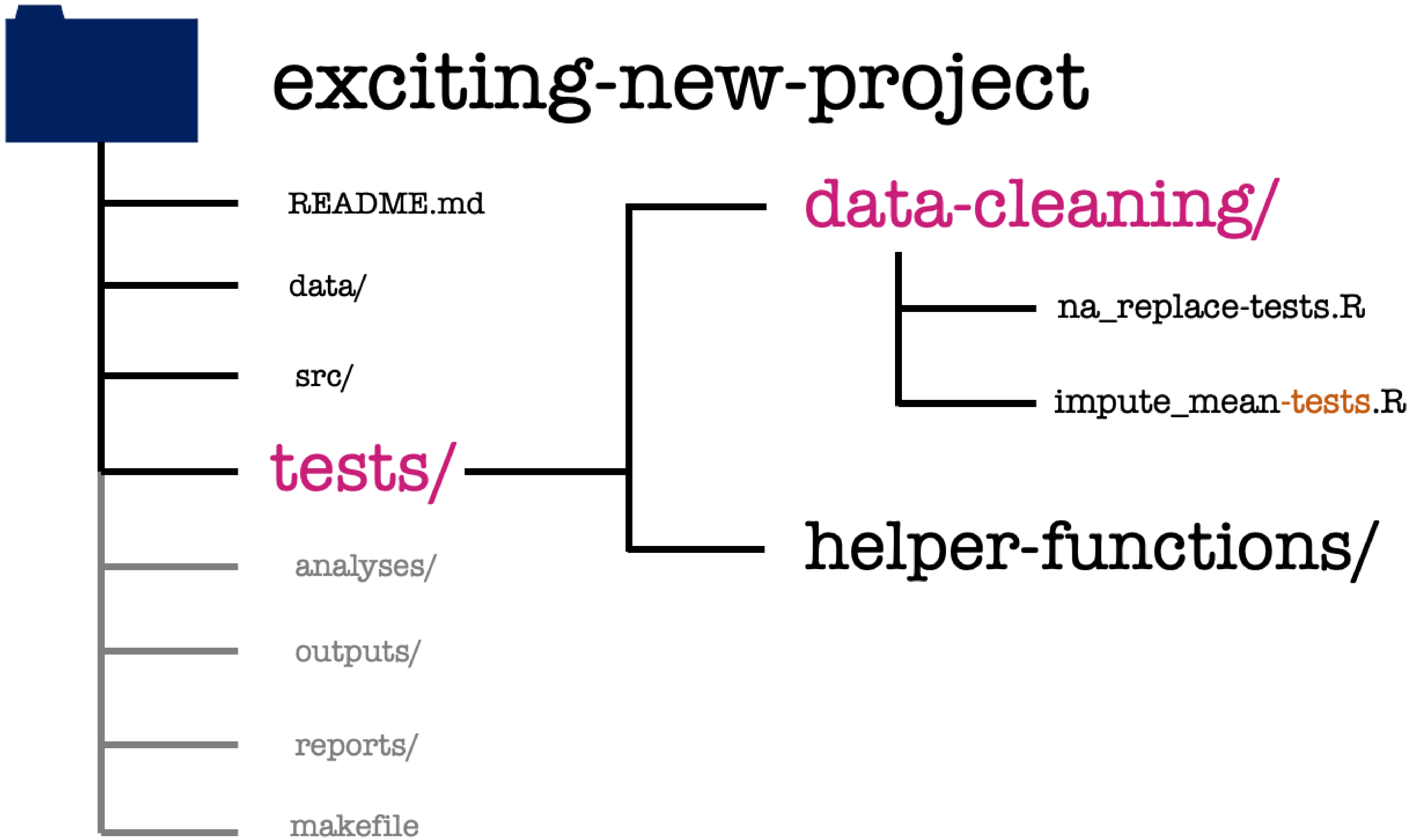
helper-functions/

analyses/

outputs/

reports/

makefile



exciting-new-project

README.md

data/

src/

tests/

data-cleaning/

na_replace-tests.R

impute_mean-tests.R

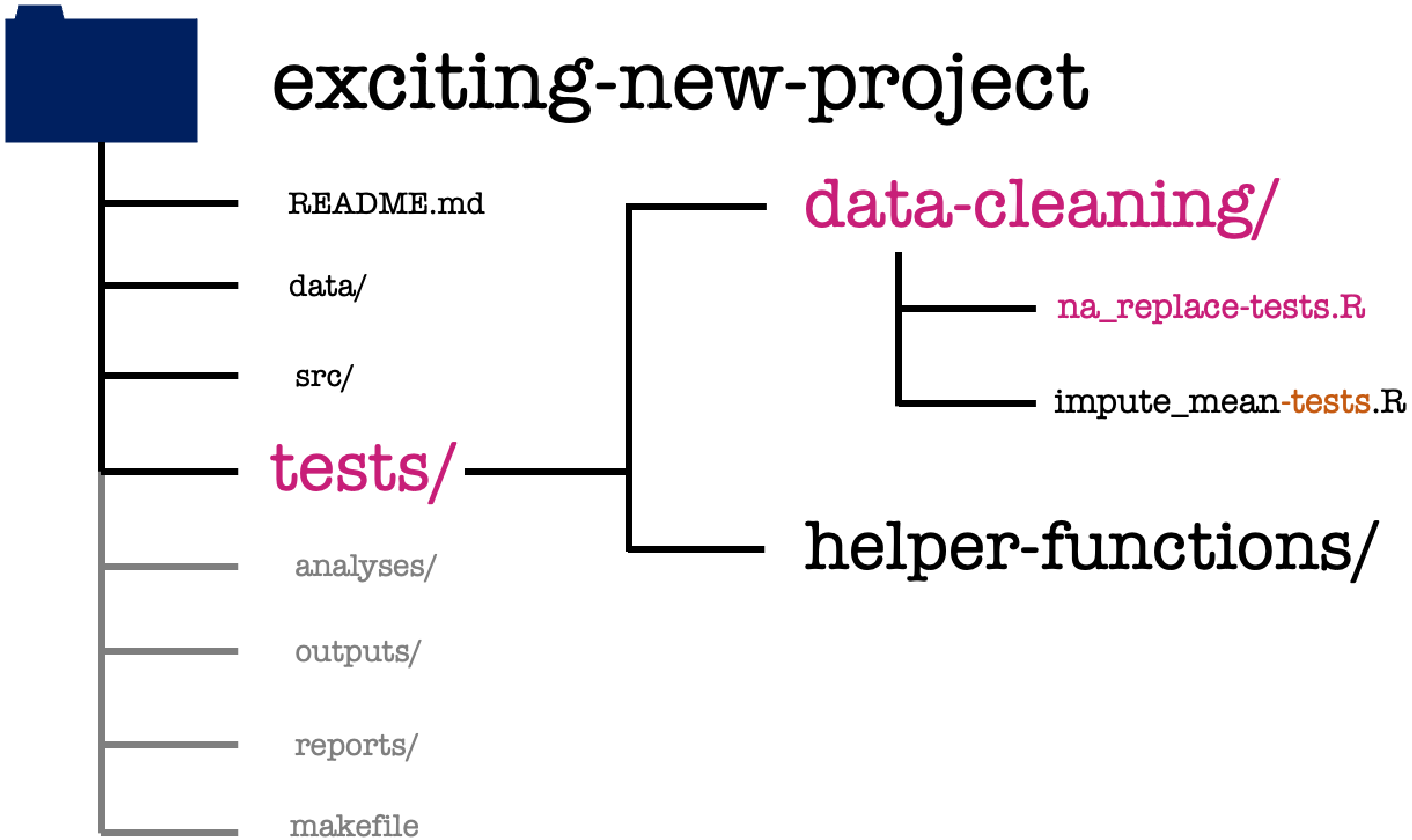
helper-functions/

analyses/

outputs/

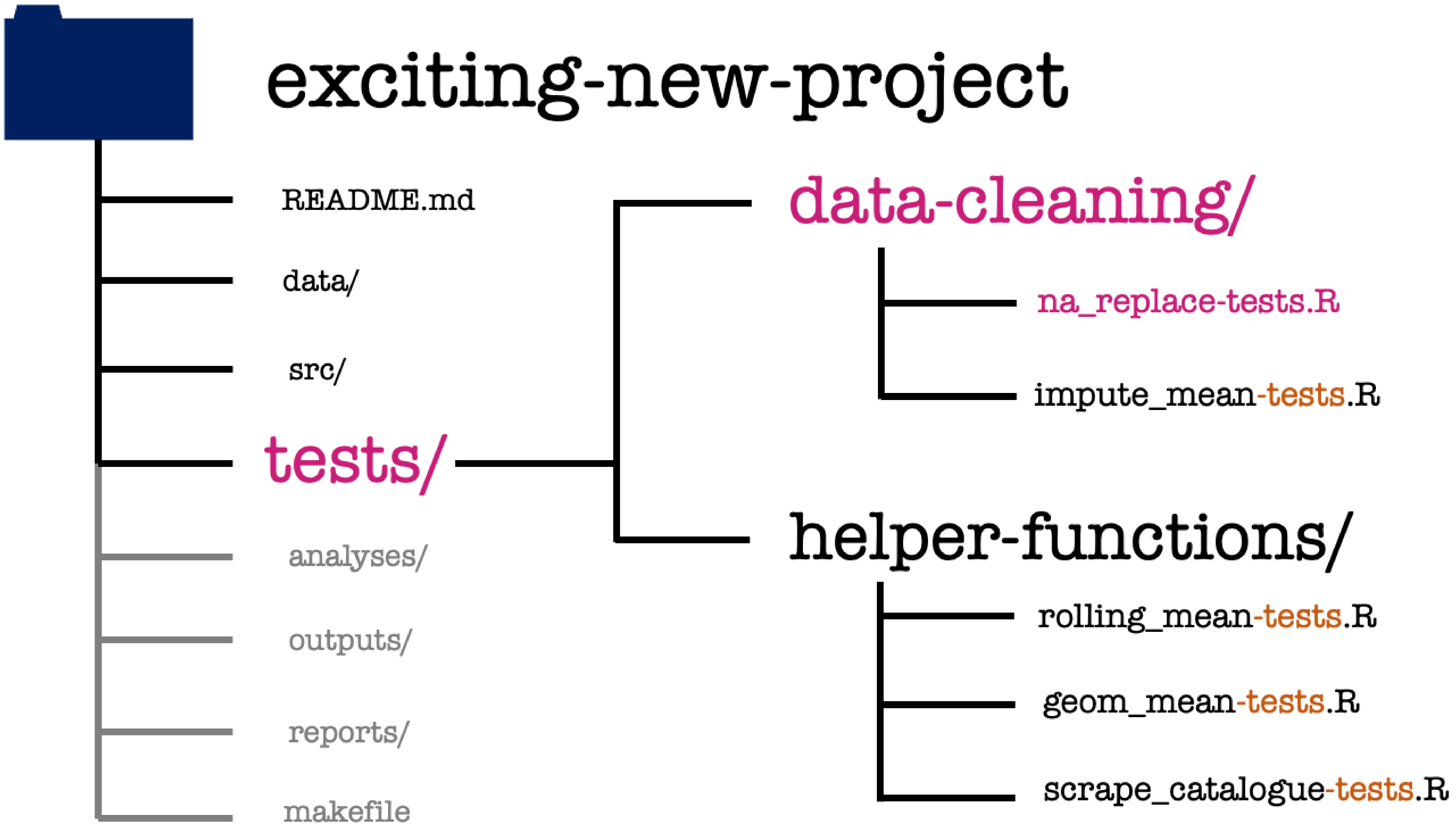
reports/

makefile



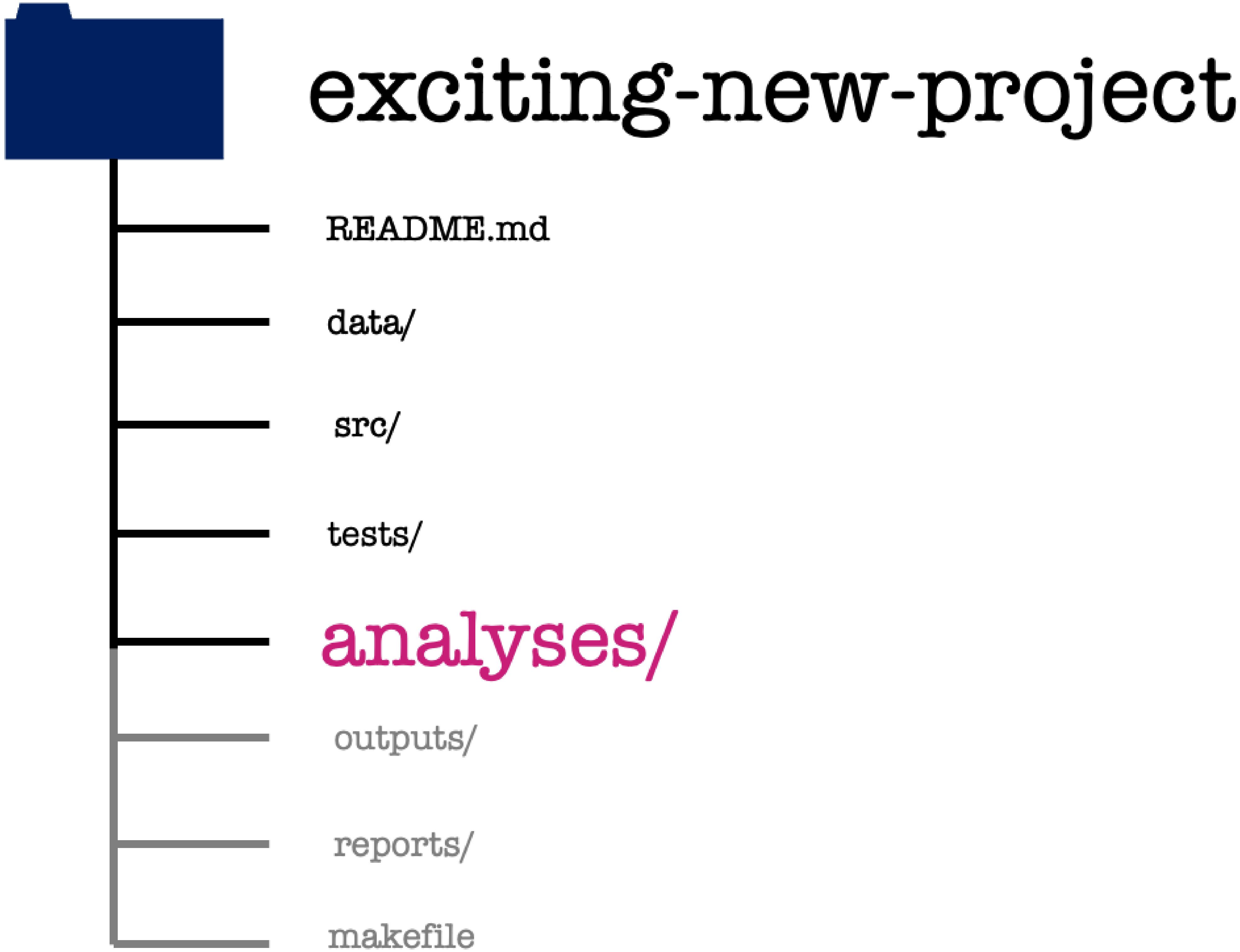
exciting-new-project

- README.md
- data/
- src/
- tests/
  - data-cleaning/
    - na_replace-tests.R
    - impute_mean-tests.R
  - helper-functions/
- analyses/
- outputs/
- reports/
- makefile



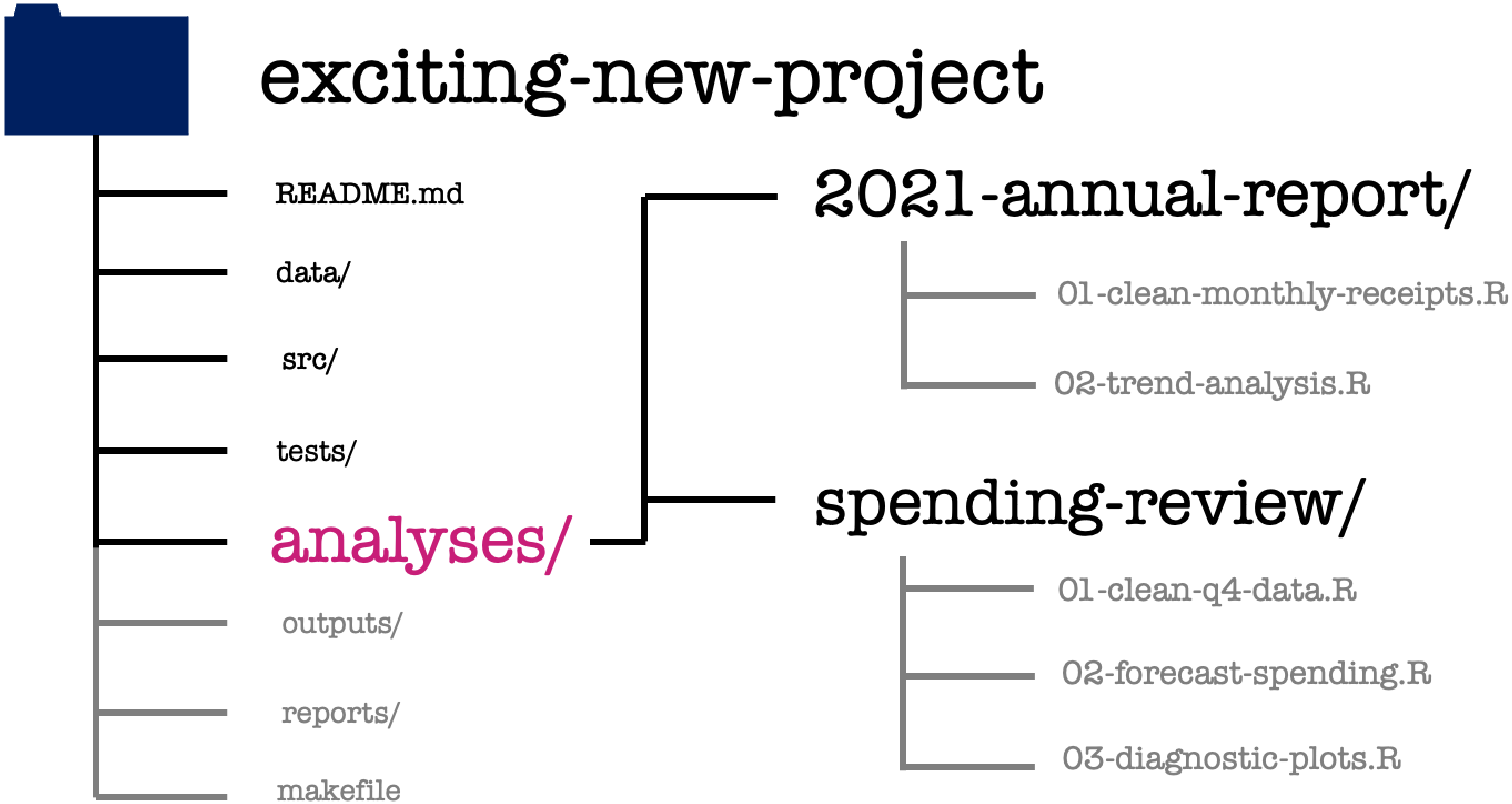
exciting-new-project

- README.md
- data/
- src/
- tests/
  - data-cleaning/
    - na_replace-tests.R
    - impute_mean-tests.R
  - helper-functions/
    - rolling_mean-tests.R
    - geom_mean-tests.R
    - scrape_catalogue-tests.R
- analyses/
- outputs/
- reports/
- makefile



exciting-new-project

- README.md
- data/
- src/
- tests/
- analyses/
- outputs/
- reports/
- makefile



exciting-new-project

README.md

data/

src/

tests/

analyses/

outputs/

reports/

makefile

2021-annual-report/

01-clean-monthly-receipts.R

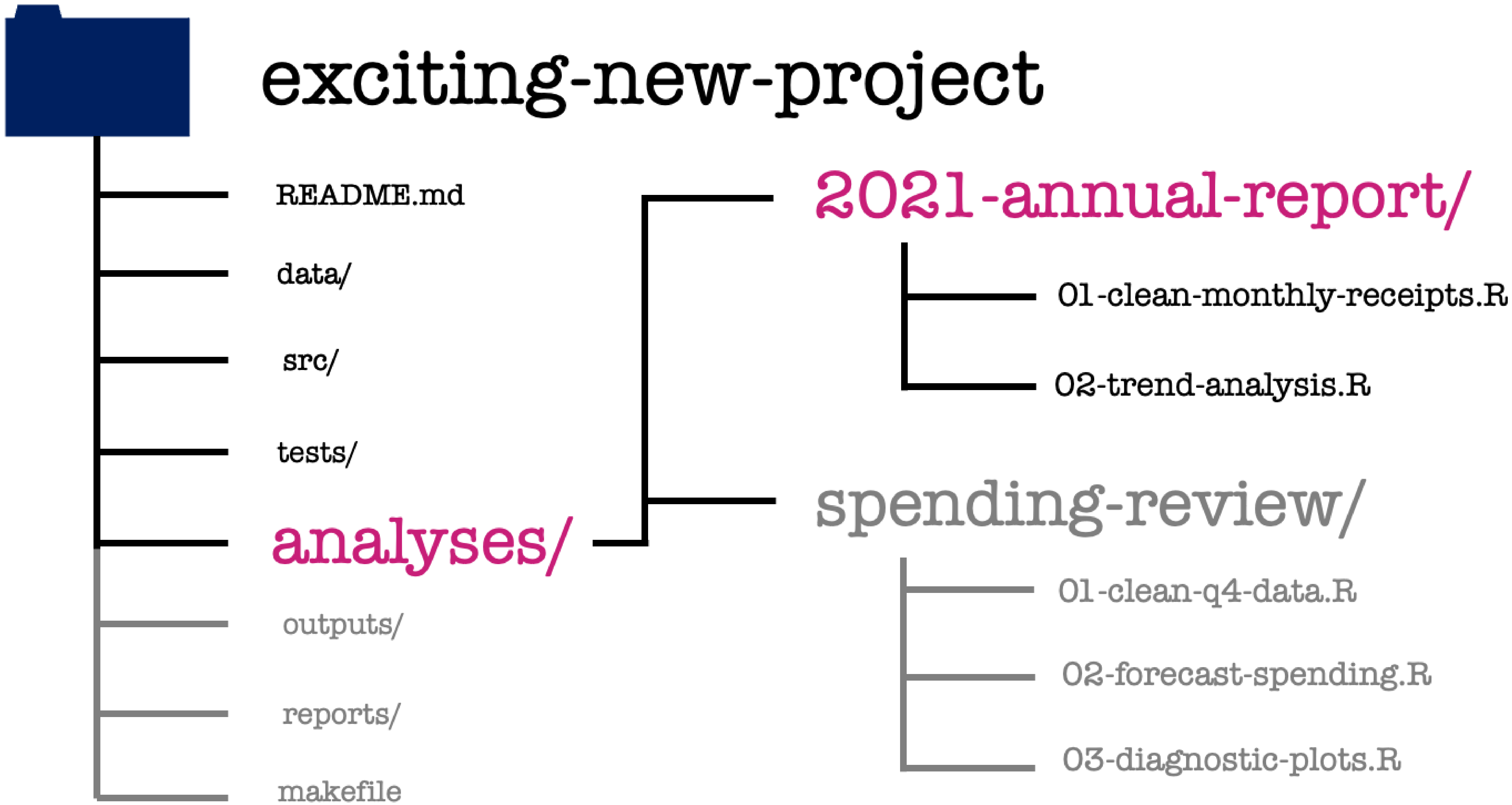
02-trend-analysis.R

spending-review/

01-clean-q4-data.R

02-forecast-spending.R

03-diagnostic-plots.R



exciting-new-project

README.md

data/

src/

tests/

analyses/

outputs/

reports/

makefile

2021-annual-report/

01-clean-monthly-receipts.R

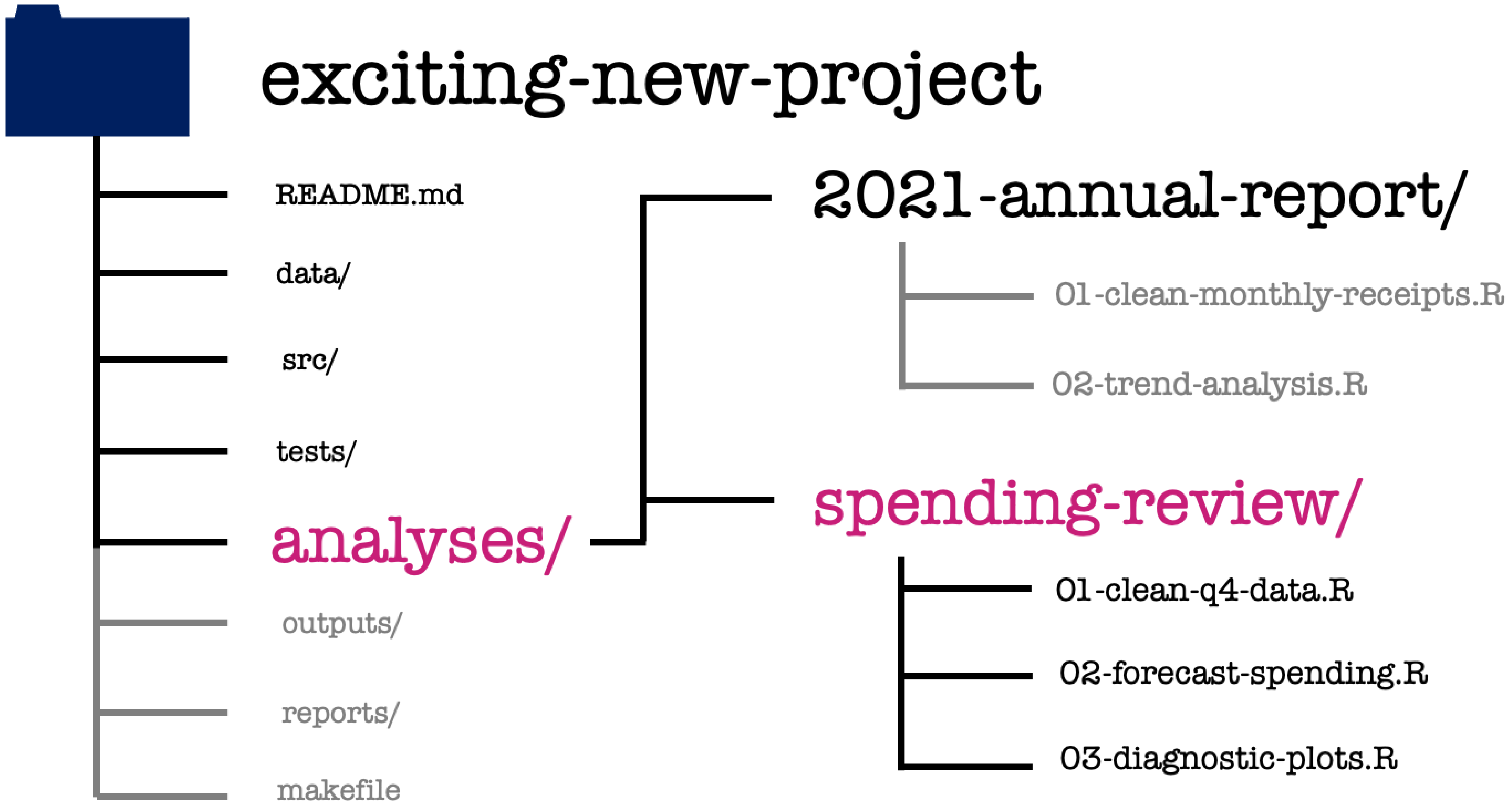
02-trend-analysis.R

spending-review/

01-clean-q4-data.R

02-forecast-spending.R

03-diagnostic-plots.R



exciting-new-project

README.md

data/

src/

tests/

analyses/

2021-annual-report/

01-clean-monthly-receipts.R

02-trend-analysis.R

spending-review/

01-clean-q4-data.R

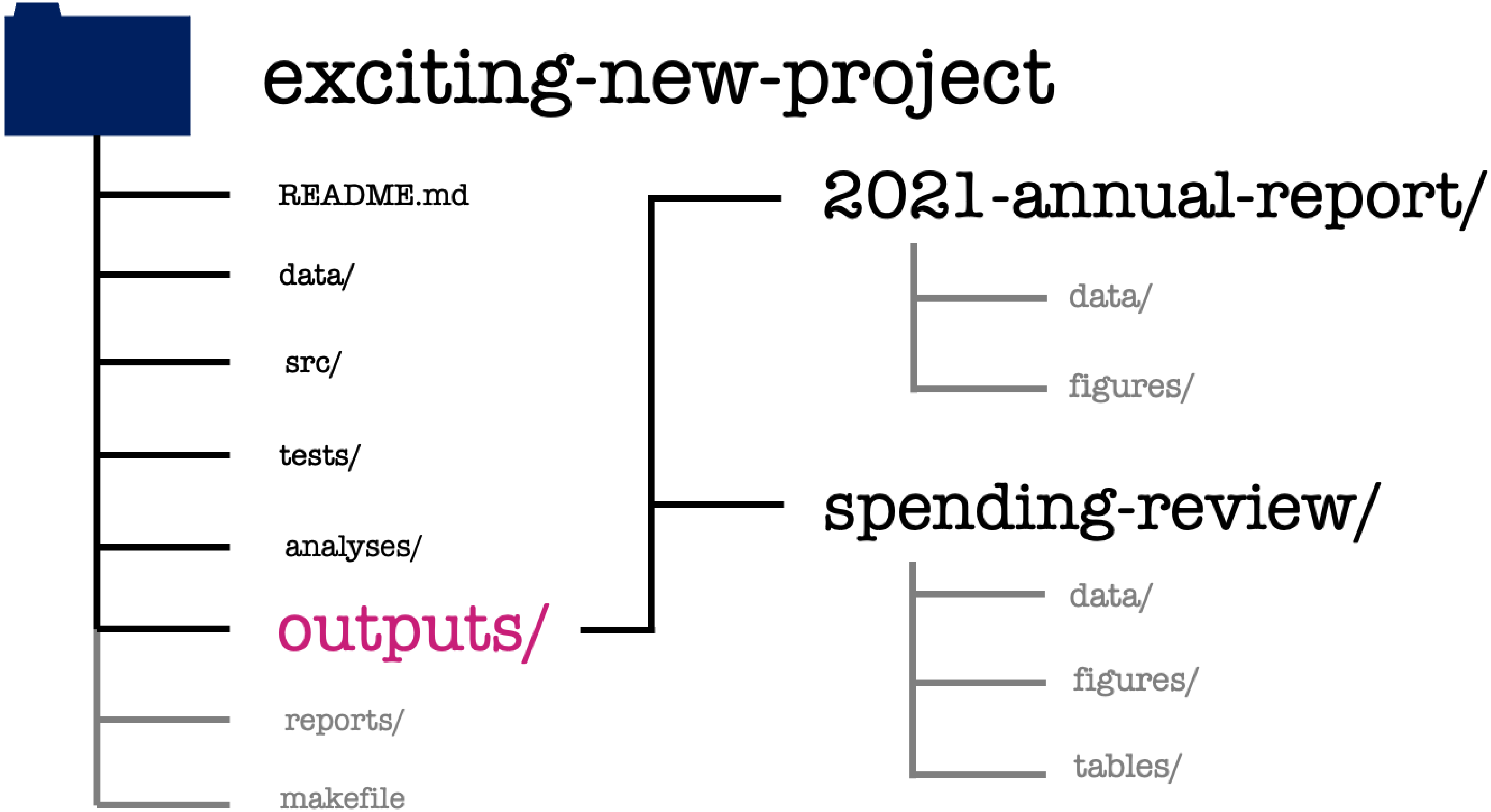
02-forecast-spending.R

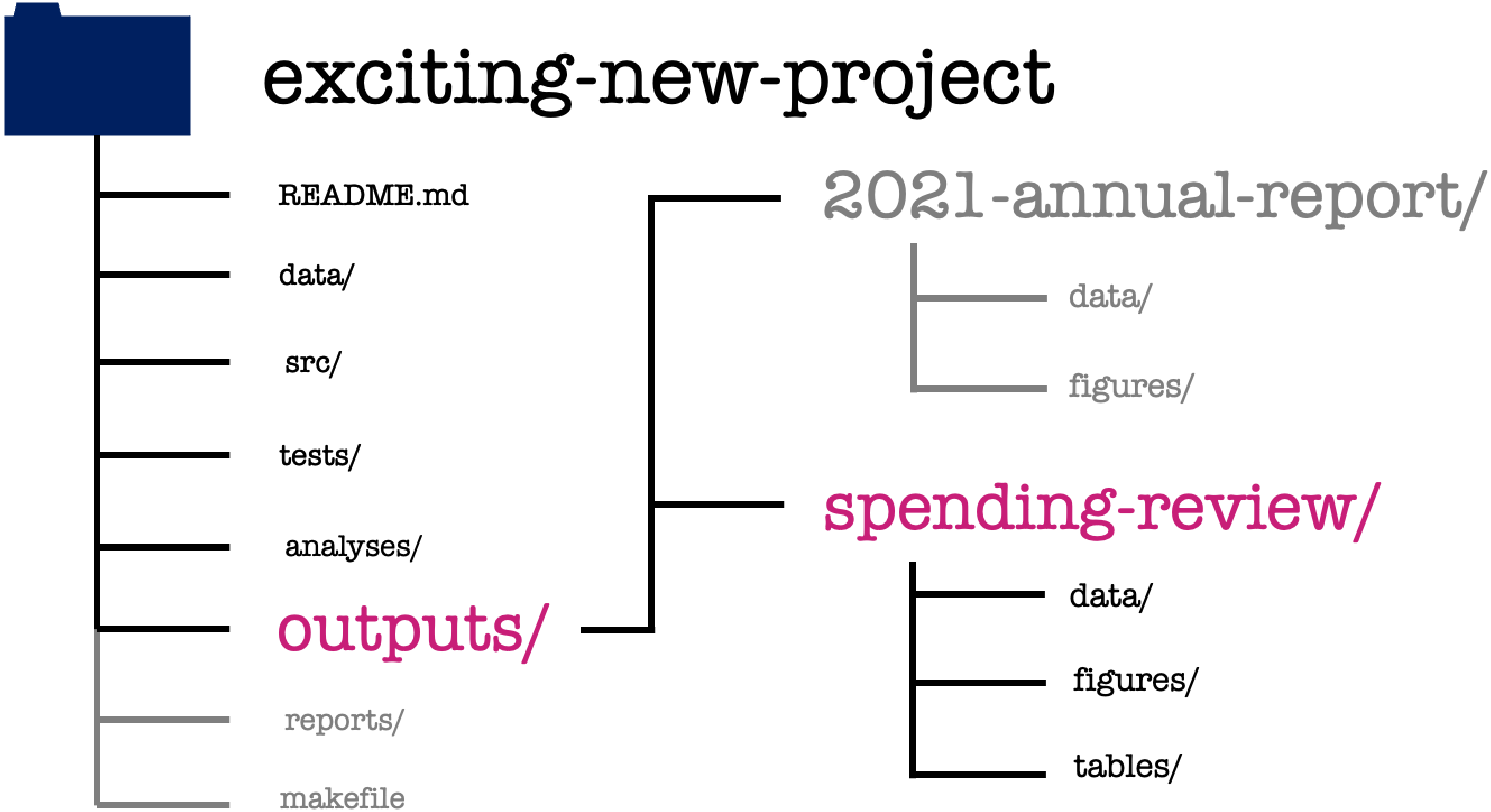
03-diagnostic-plots.R

outputs/

reports/

makefile


exciting-new-project
README.md
data/
src/
tests/
analyses/
outputs/
reports/
makefile
2021-annual-report/
data/
figures/
spending-review/
data/
figures/
tables/



exciting-new-project

README.md

data/

src/

tests/

analyses/

outputs/

reports/

makefile

2021-annual-report/

data/

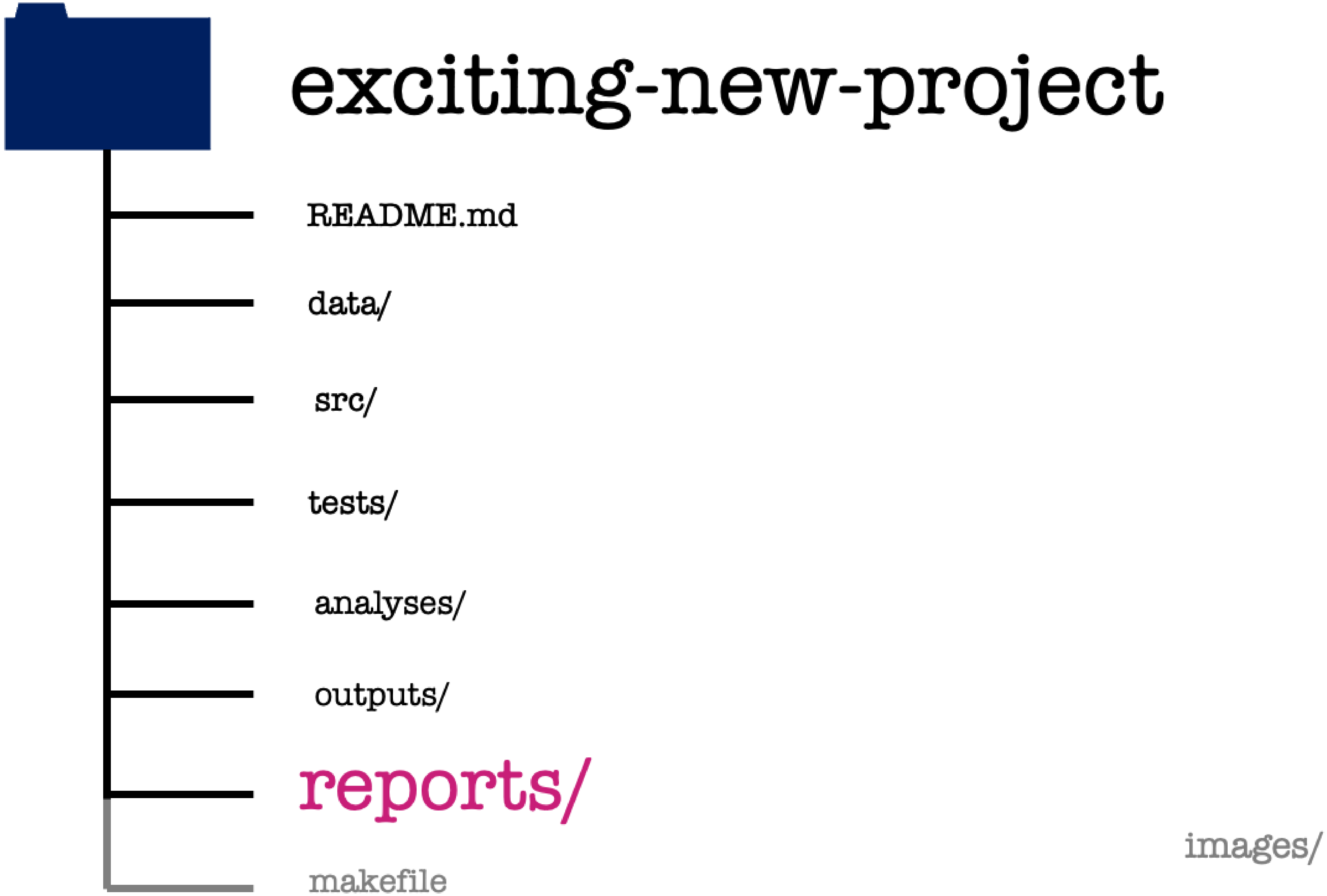
figures/

spending-review/

data/

figures/

tables/



exciting-new-project

README.md

data/

src/

tests/

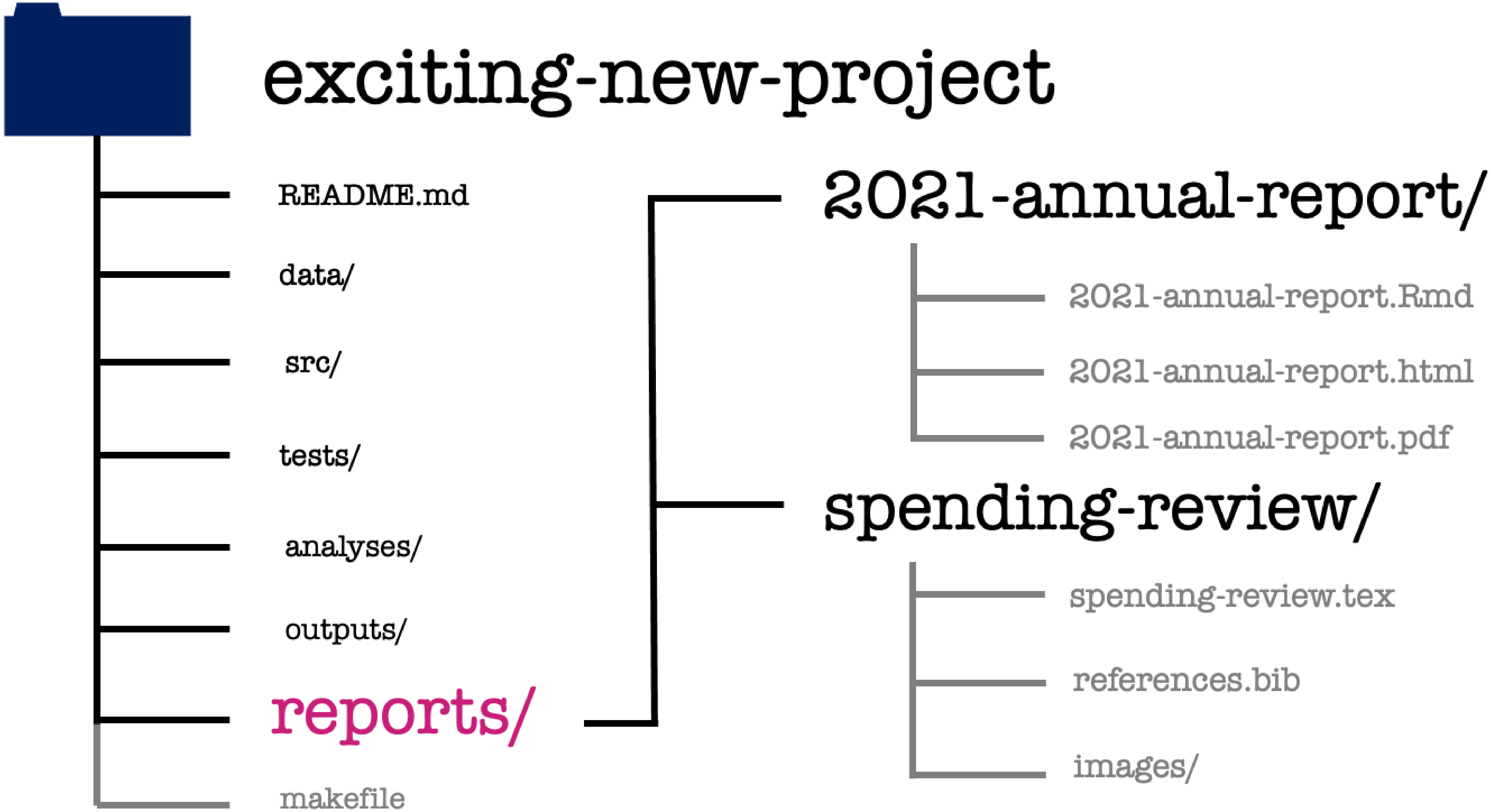
analyses/

outputs/

reports/

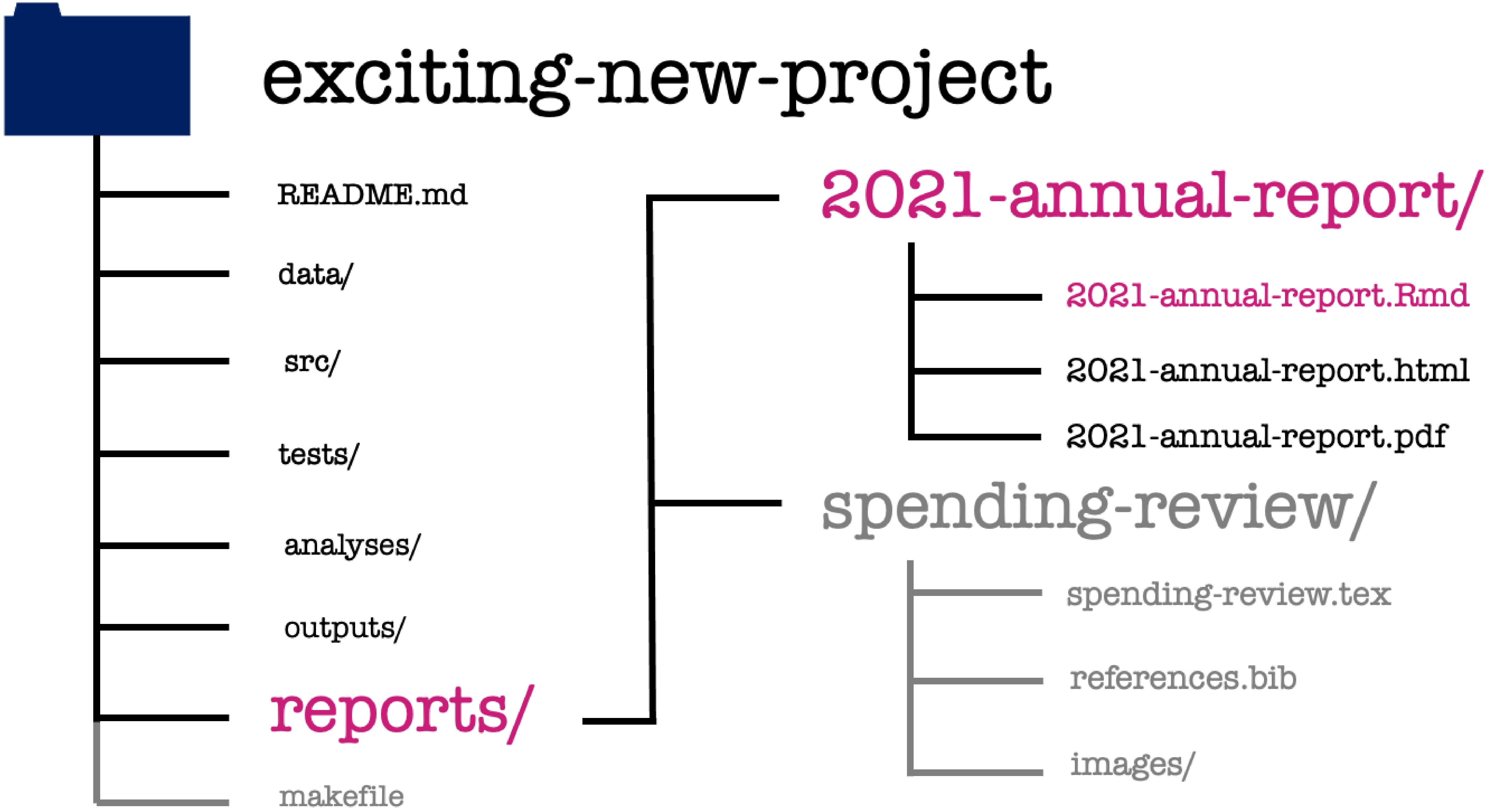
images/

makefile



exciting-new-project

- README.md
- data/
- src/
- tests/
- analyses/
- outputs/
- reports/
    - 2021-annual-report/
        - 2021-annual-report.Rmd
        - 2021-annual-report.html
        - 2021-annual-report.pdf
    - spending-review/
        - spending-review.tex
        - references.bib
        - images/
- makefile



exciting-new-project

README.md

data/

src/

tests/

analyses/

outputs/

reports/

makefile

2021-annual-report/

2021-annual-report.Rmd

2021-annual-report.html

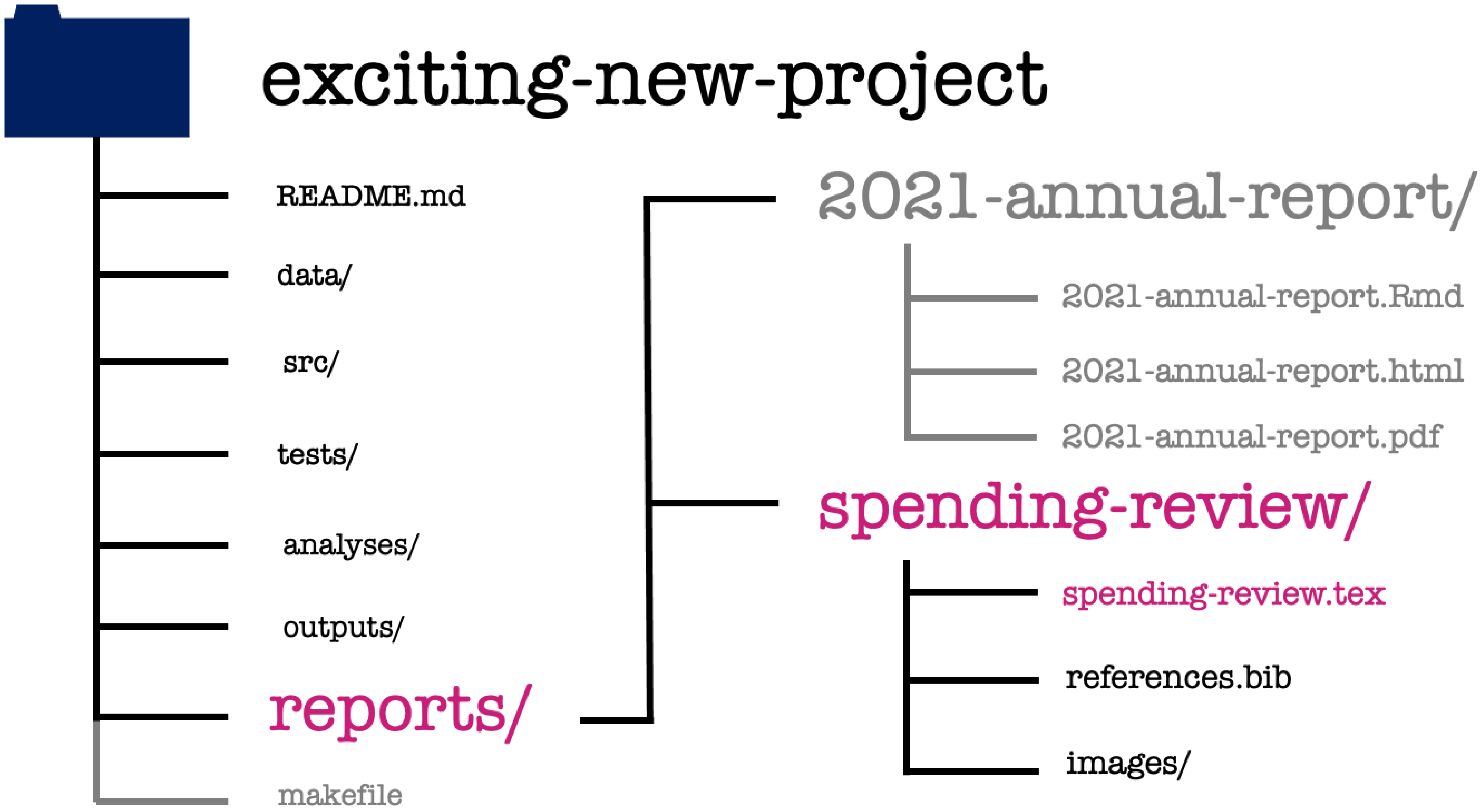
2021-annual-report.pdf

spending-review/

spending-review.tex

references.bib

images/



exciting-new-project

README.md

data/

src/

tests/

analyses/

outputs/

reports/

makefile

2021-annual-report/

2021-annual-report.Rmd

2021-annual-report.html

2021-annual-report.pdf

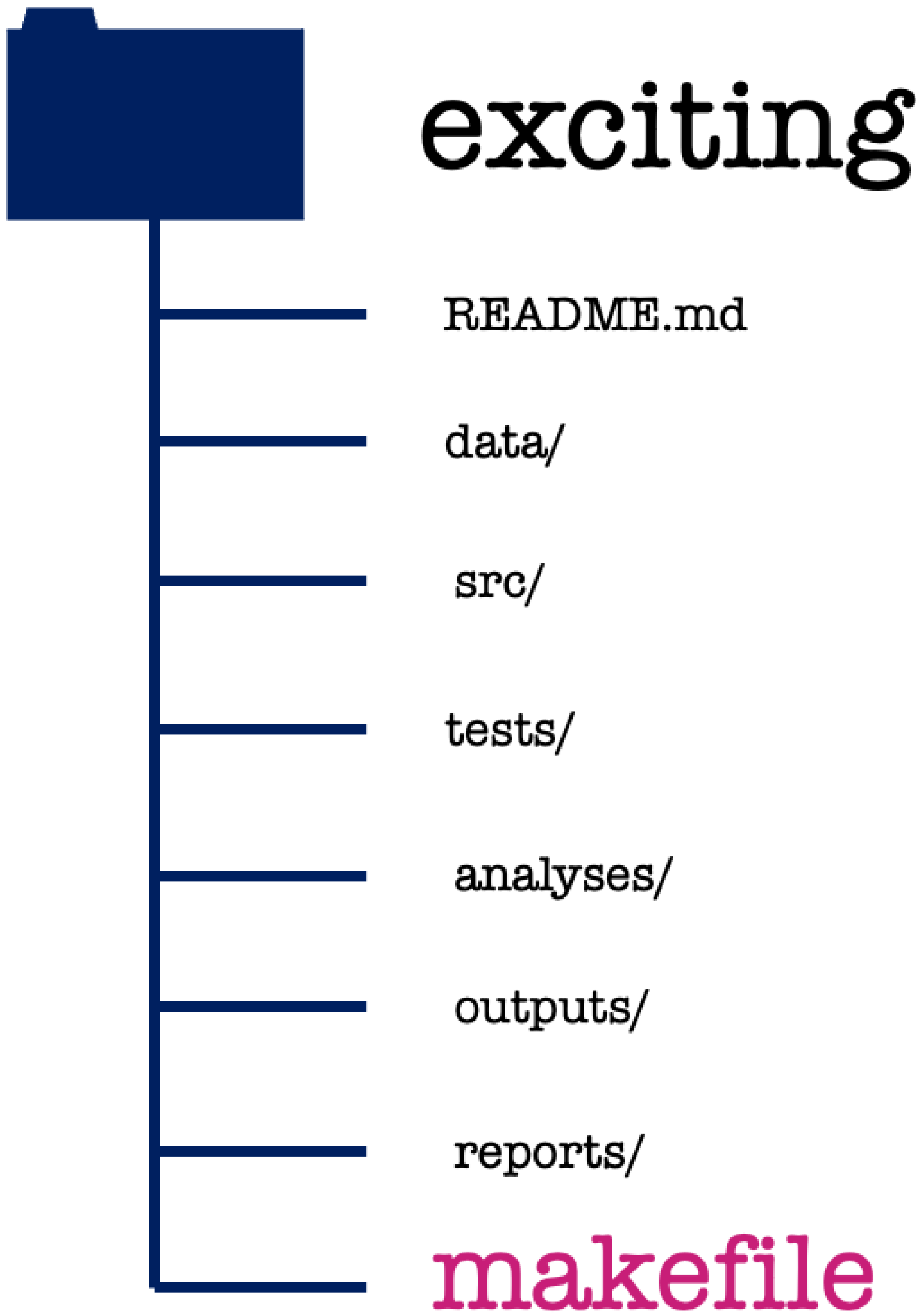
spending-review/

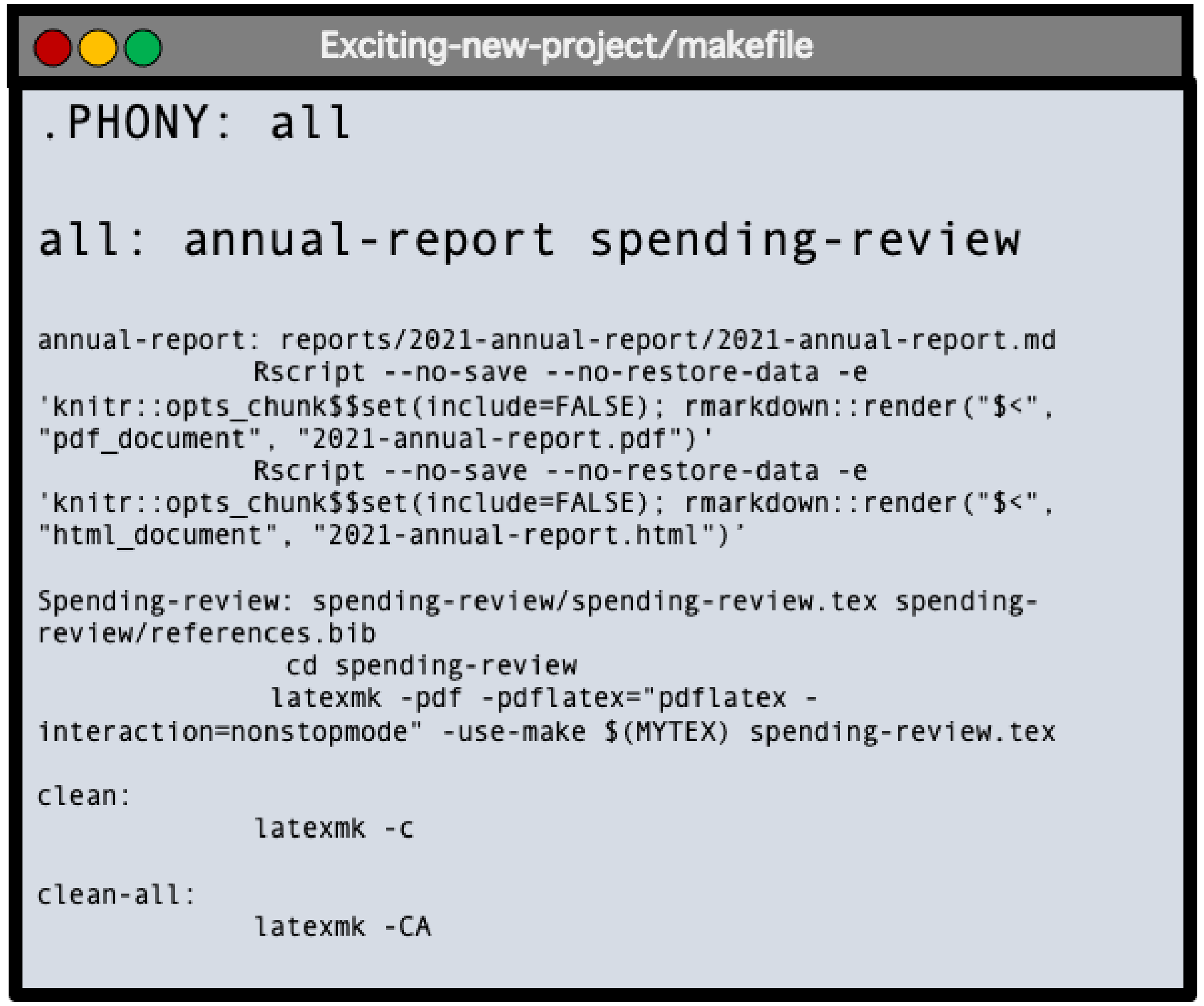
spending-review.tex

references.bib

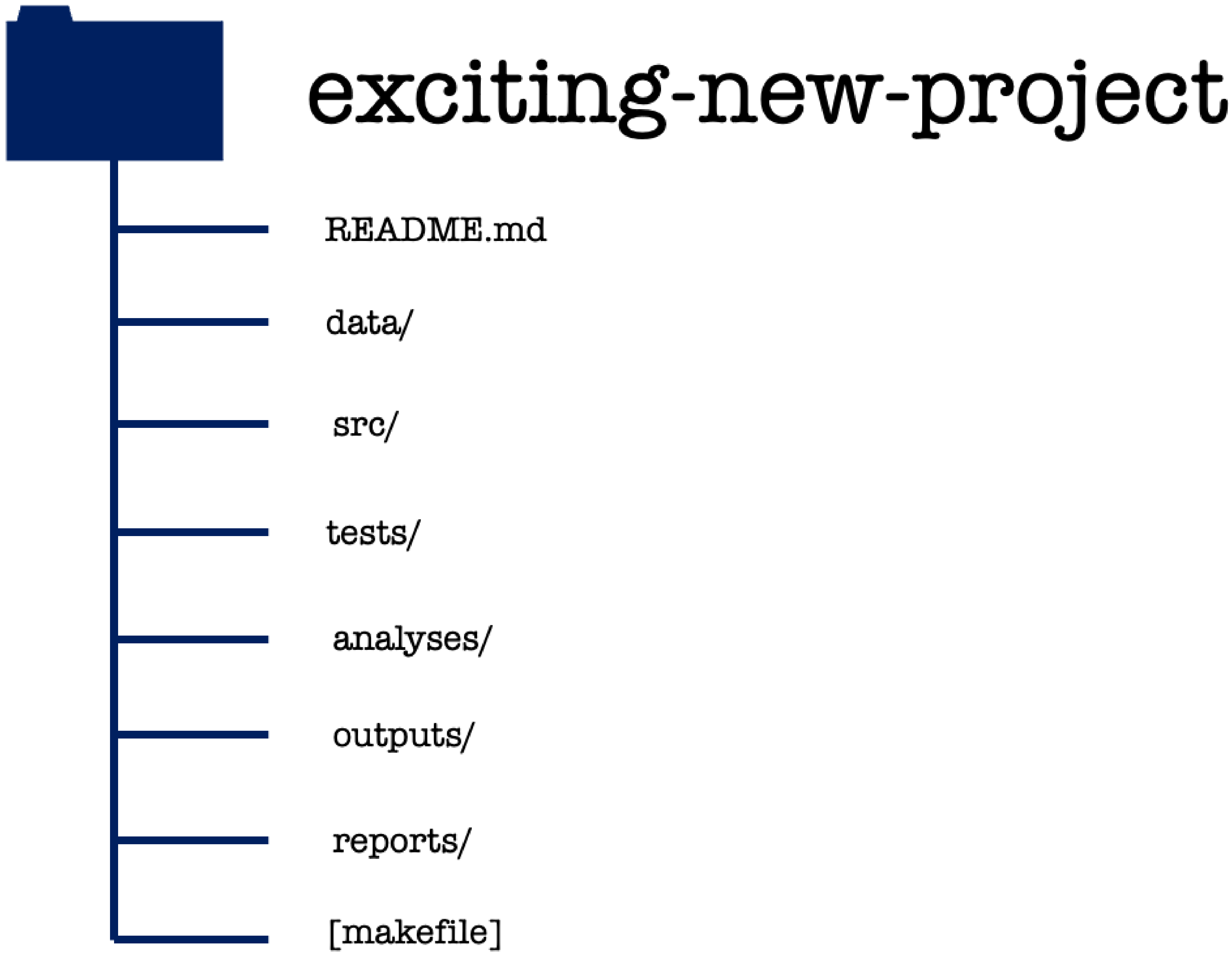
images/

exciting-new-project





```
.PHONY: all

all: annual-report spending-review

annual-report: reports/2021-annual-report/2021-annual-report.md
              Rscript --no-save --no-restore-data -e
'knitr::opts_chunk$$set(include=FALSE); rmarkdown::render("$<",
"pdf_document", "2021-annual-report.pdf")'
              Rscript --no-save --no-restore-data -e
'knitr::opts_chunk$$set(include=FALSE); rmarkdown::render("$<",
"html_document", "2021-annual-report.html")'

Spending-review: spending-review/spending-review.tex spending-
review/references.bib
                cd spending-review
               latexmk -pdf -pdflatex="pdflatex -
interaction=nonstopmode" -use-make $(MYTEX) spending-review.tex

clean:
              latexmk -c

clean-all:
              latexmk -CA
```



exciting-new-project

README.md

data/

src/

tests/

analyses/

outputs/

reports/

[makefile]

Summary

- Introduced a standardised project structure;

- Good starting point for most data science projects;

- Exceptions are **apps** and **packages**.

References

Broman, Karl, Mine Cetinkaya-Rundel, Amy Nussbaum, Christopher Paciorek, Roger Peng, Daniel Turek, and Hadley Wickham. 2017. “Recommendations to Funding Agencies for Supporting Reproducible Research.” In **American Statistical Association**, 2:1–4.