

Save your tears for  
the data: A touch of  
Docker in a Data  
Scientist's workflow

# Save your tears for the data: A touch of Docker in a Data Scientist's workflow

---



Eric Wanjau, Data Scientist - LIDA

[github.com/R-icntay/res\\_comp\\_leeds\\_2022](https://github.com/R-icntay/res_comp_leeds_2022)

@ericntay at #ResCompLeedsCon2022

01:08:52

People

Chat

Reactions

More

Camera

Mic

Share

Leave

⚠ Recording and transcription have started. Let everyone know they're being recorded and transcribed. [Privacy policy](#)

Dismiss

# Fairy tale ending: Urban Transport Modelling for Sustainable Well-Being in Hanoi (UTM-Hanoi)

Alexis Comber

[urban-analytics.github.io/UTM-Hanoi/intro.html](https://urban-analytics.github.io/UTM-Hanoi/intro.html)

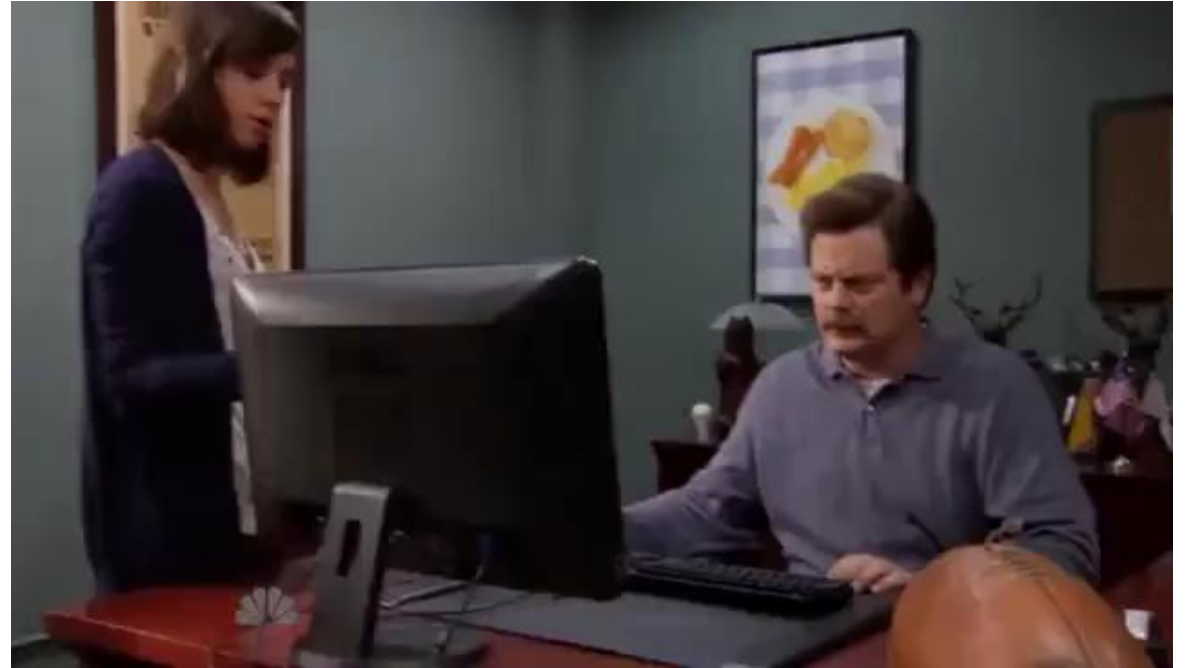
Nicolas Malleon

Kristina Bratkova

10°C Light rain ENG

12:40 PM  
4/4/2022

But somewhere  
along the line,  
it felt like ...





I am running on xxx OS



The model could make the dashboard lag



Reproducibility



Isolation



Portability



# To the rescue: Dev containers - Docker containers configured to provide a full-featured development environment

- Tailored environment with all the tools and runtimes e.g R, Python, Julia
- IDE of choice e.g VS Code, RStudio
- Portable i.e GitHub Codespaces, local machine, container hosts e.g Azure, AWS, GCP
- Same experience for every user





Let's set one  
up: R +  
Python +  
VS Code +  
RStudio

# Container Config: Dockerfile

```
# R version: 4, 4.1, 4.0  
ARG VARIANT="4.2"  
FROM rocker/tidyverse:${VARIANT}
```

**FROM** command describes a base environment, so we don't need to start from scratch

```
# Install Python and some packages  
RUN apt-get -y install \  
    python3-pip \  
&& pip --disable-pip-version-check --no-cache-dir install pandas \  
&& pip --disable-pip-version-check --no-cache-dir install seaborn \
```

```
# Install R packages  
&& install2.r --error --skipinstalled --ncpus -1 \  
    palmerpenguins \  
    tidymodels \  
    languageserver \
```

**RUN** changes the base environment  
e.g. installing additional packages



# Customizing dev environment: devcontainer.json -- VS Code

```
"name": "res_comp_leeds22",  
  "build": {  
    "dockerfile": "Dockerfile",  
    "args": { "VARIANT": "4.2" }  
  },
```

Specifies the name of dev container and references the Dockerfile we defined earlier

```
"settings": {  
  "r.rpath.linux": "/usr/local/lib/R/bin/R",  
},  
  
"extensions": [  
  // R extensions  
  "ikuyadeu.r",  
  "rdebugger.r-debugger",  
  "reitorsupport.r",  
  
  // Add Jupyter and Python vs code extensions  
  "ms-toolsai.jupyter",  
  "ms-python.python"  
],
```

Standard elements everyone working in the dev env would need

But, what if I really  
really really love the  
RStudio IDE?

# Customizing dev environment: devcontainer.json -- RStudio

```
"forwardPorts": [8787],  
"portsAttributes": {  
  "8787": {  
    "label": "RStudio",  
    "requireLocalPort": true,  
    "onAutoForward": "ignore"  
  }  
},
```

**forwardPorts:** makes a list of ports inside the container available locally.

**portsAttributes:** specifies the properties of the ports

```
"onCreateCommand": "sudo rstudio-server start"
```

**onCreateCommand:** finalizes setup by specifying what to execute immediately after container is created





# Python Demo: VS Code / Docker



# R Demo: Rstudio/ GithubCodespaces



# C, C++, Julia, Java and friends?

<https://github.com/microsoft/vscode-dev-containers>





# Wrapping up:

Reproducible and customizable development environments

Isolated environments that won't change pre-existing setups

Portable: MacOS, Linux, or Windows

Great for workshops!

Tears of joy ...

# Thank you!

Slides, links, resources and Dev Container:

[github.com/R-icntay/res\\_comp\\_leeds\\_2022](https://github.com/R-icntay/res_comp_leeds_2022)



Eric Wanjau, Data Scientist - LIDA

e.wanjau@leeds.ac.uk

@ericntay