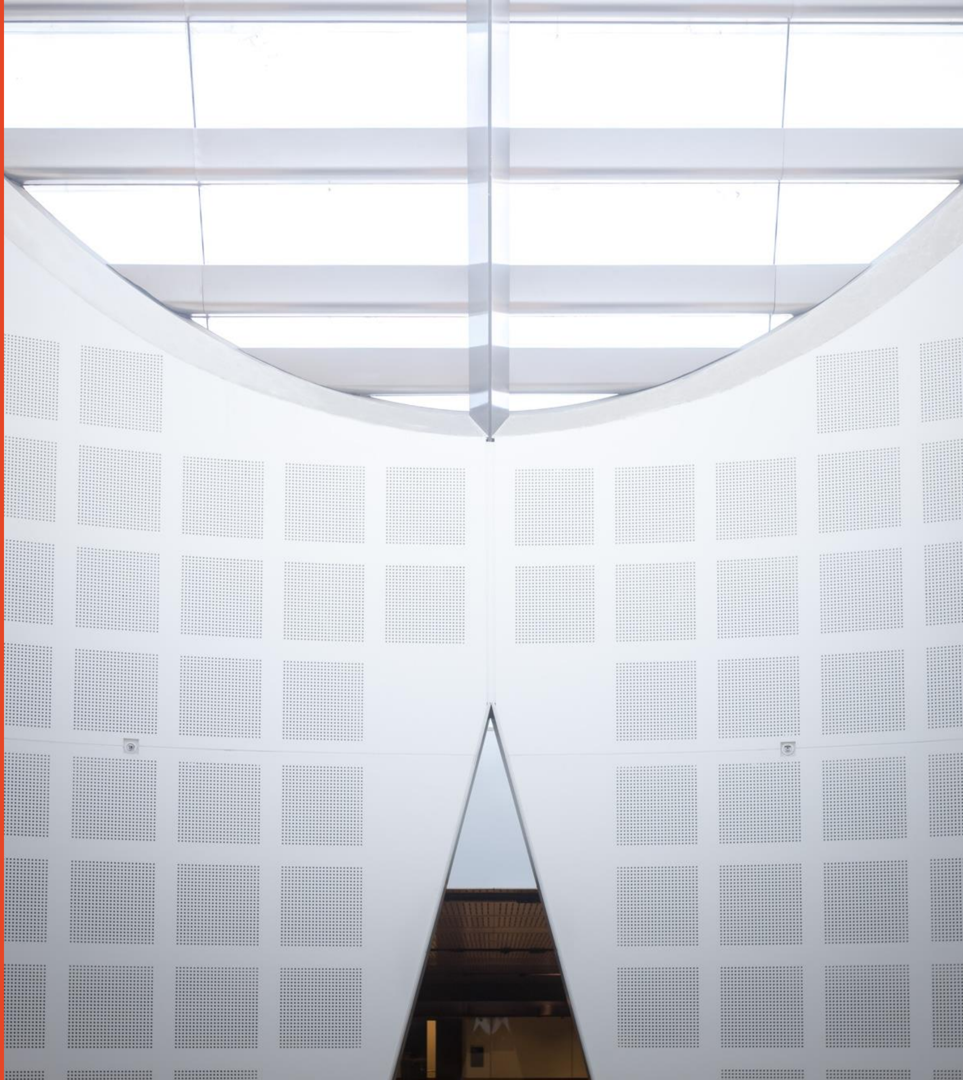
OCMP5310: Principles of Data Science

Week 1 Live Session

Presented by

Daniela Rivas

THE UNIVERSITY OF SYDNEY



WELCOME!

Daniela Rivas

- Bachelor in Engineering.
- PhD in Computational Biophysics.
- Postdoctoral Research Fellow at the Daffodil Centre.
  - Cervical Cancer and HPV modelling.
- Casual Academic at the University of Sydney.

LIVE SESSIONS

Live Sessions Details

- 6 Weeks.
- Every **Monday** at **6:30 PM.**
- First Live Session: Monday 17th of April.
- Last Live Session: Monday 22nd of May.
- 90 minutes.

Live Sessions Expectations

Before the session:

- Watch the lecture videos.
- Work on the practical exercises.

During the session:

- Reflect on the main topics of the week.
- Run some additional group exercises.
- Give feedback on the exercises and the assessment tasks.

ASSESSMENTS

Assessments

- The official syllabus is the authoritative source of assessment information.
  - https://www.sydney.edu.au/units/OCMP5310/2023-S1CRB-OL-OP
- 10%: Weekly Review Quizzes.
- 10%: Project Stage 1 (Week 3, 7 May 2023, 23:59*).
- 10%: Project Stage 2 (Week 5, 16 May 2023, 23:59*).
- 5%: Project Oral Presentation (Week 6, 22 May 2023, during Live Session).
- 15%: Project Stage 3 (Week 6, 28 May 2023, 23:59*).
- 50%: Final exam (Week 8, 5 June 2023, 18:30*).

*Sydney time

Project

This is a small data science project that will be developed incrementally through the course, and which focuses on **understanding the data science process** from data ingestion and cleaning, over understanding data and gathering descriptive statistics, to building a (simple) predictive model. There are three main submissions, as well as a short oral presentation that is designed to train presentation skills.

Project

- **Stage 0:** Project Proposal.
- **Stage 1:** Data Acquisition and Cleaning, and Problem definition.
- **Project Stage 2:** Data Summarisation and Analysis.
- **Project Stage 3:** Predictive Model and Evaluation.

- More details here.

WEEK 1: SESSION ACTIVITY

THE UNIVERSITY OF
SYDNEY

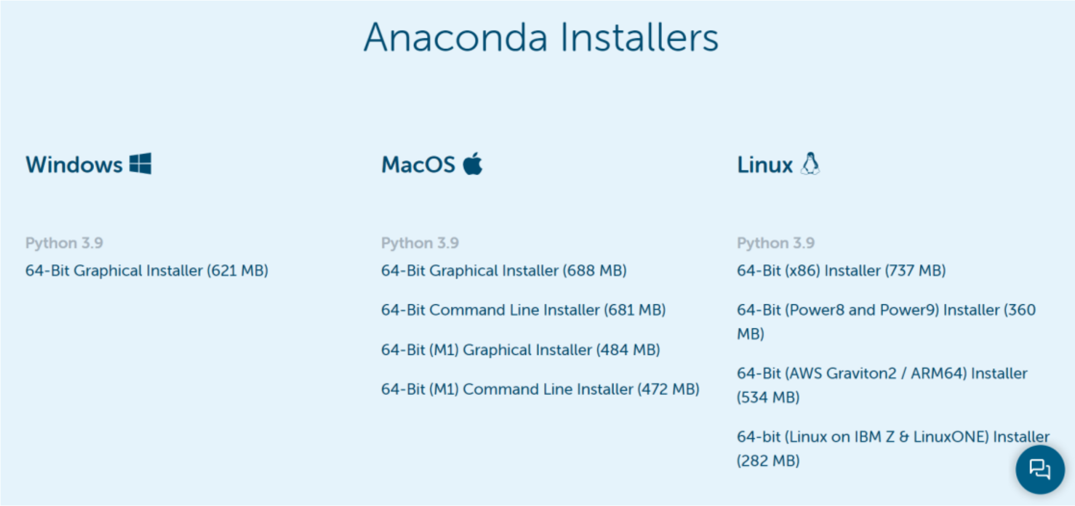
Week 1: Session Activity

- Installing Python and PostgreSQL.
- Exploratory Data Analysis with Spreadsheets/Python (Calculating Descriptive Statistics, plotting).
- Descriptive Statistics and Data visualization with Python.

Installing Python and PostgreSQL

Installing Anaconda

- Go to https://www.anaconda.com/products/distribution
- Choose the right installer for your Operating System and download it.



Anaconda Installers

Windows

Python 3.9

64-Bit Graphical Installer (621 MB)

MacOS

Python 3.9

64-Bit Graphical Installer (688 MB)

64-Bit Command Line Installer (681 MB)

64-Bit (M1) Graphical Installer (484 MB)

64-Bit (M1) Command Line Installer (472 MB)

Linux

Python 3.9

64-Bit (x86) Installer (737 MB)

64-Bit (Power8 and Power9) Installer (360 MB)

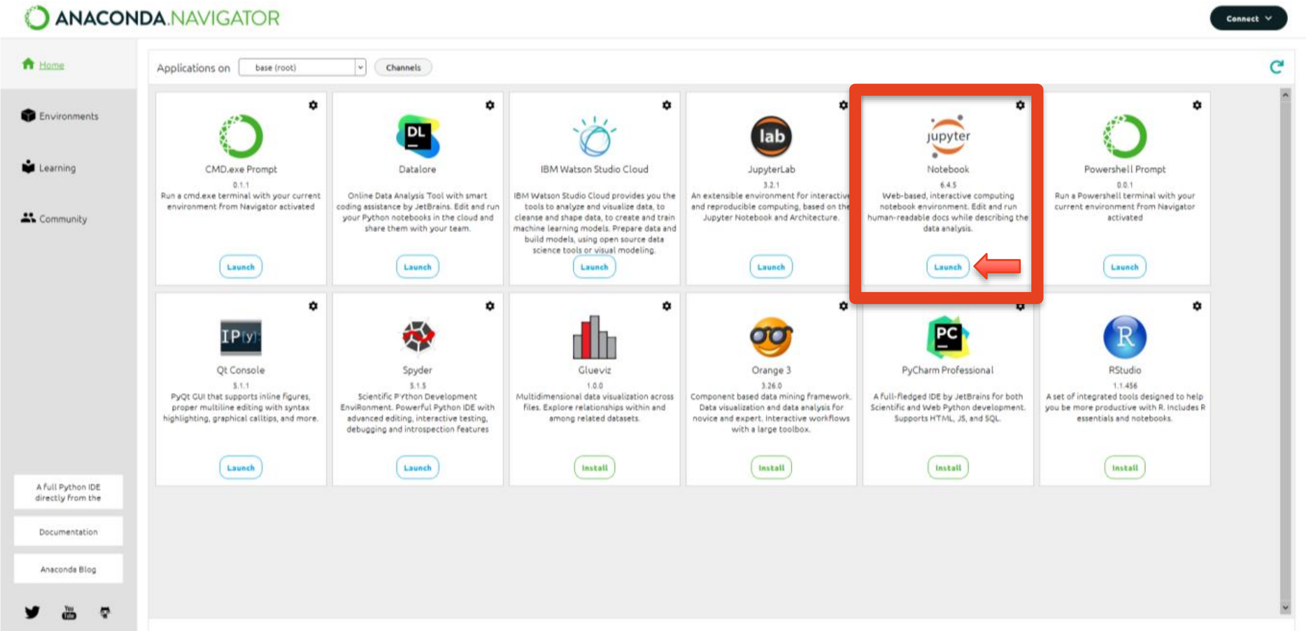
64-Bit (AWS Graviton2 / ARM64) Installer (534 MB)

64-bit (Linux on IBM Z & LinuxONE) Installer (282 MB)

- Open the installer and follow the instructions to install.

Opening Jupyter Notebook

– Open the Anaconda Navigator and click on “Launch” under Jupyter Notebook from the options available.



CMD.exe Prompt
0.1.1
Run a cmd.exe terminal with your current environment from Navigator activated
Launch

Datalore
Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team.
Launch

IBM Watson Studio Cloud
IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.
Launch

JupyterLab
3.2.1
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch

Notebook
6.4.5
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch

Powershell Prompt
0.0.1
Run a Powershell terminal with your current environment from Navigator activated
Launch

Qt Console
5.1.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch

Spyder
5.1.5
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch

Glueviz
1.0.0
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Install

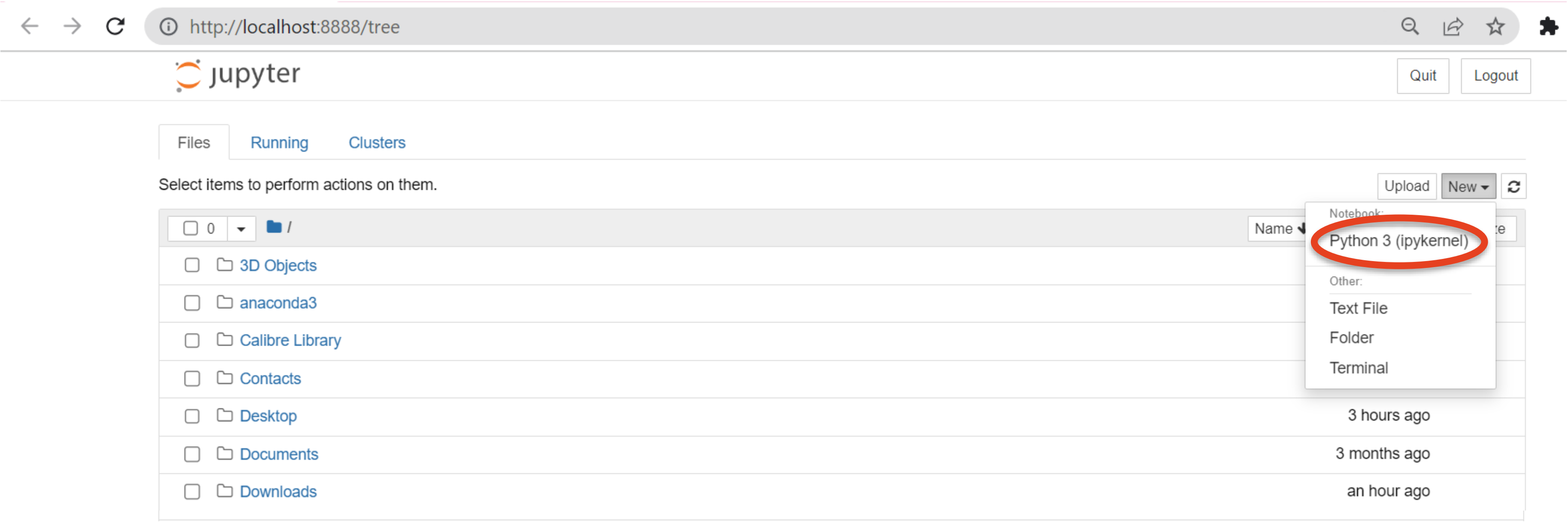
Orange 3
3.26.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.
Install

PyCharm Professional
A full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL.
Install

RStudio
1.1.456
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.
Install

Opening a New Jupyter Notebook

– In the upper right select [New] –> [Python 3] to open a new Python 3 notebook.

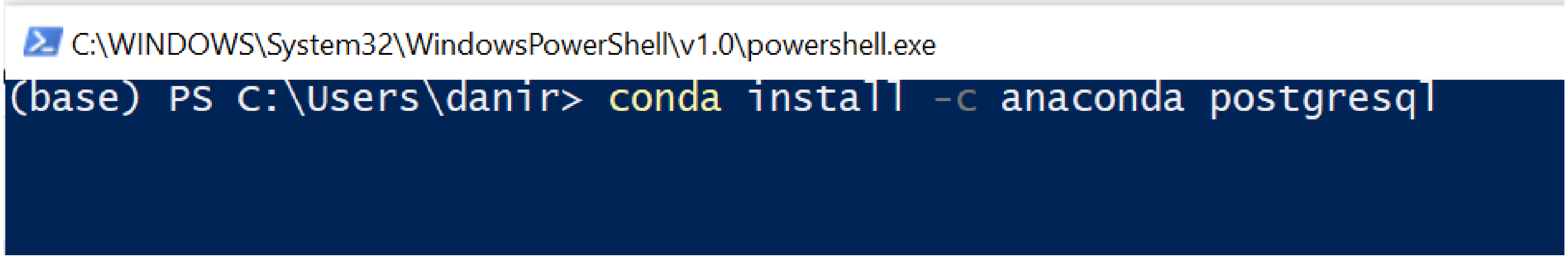


Select items to perform actions on them.

3D Objects	
anaconda3	
Calibre Library	
Contacts	
Desktop	3 hours ago
Documents	3 months ago
Downloads	an hour ago

Installing PostgreSQL on Anaconda

- Open your computer terminal or the Anaconda Powershell Prompt and type:
  - `conda install -c anaconda postgresql`

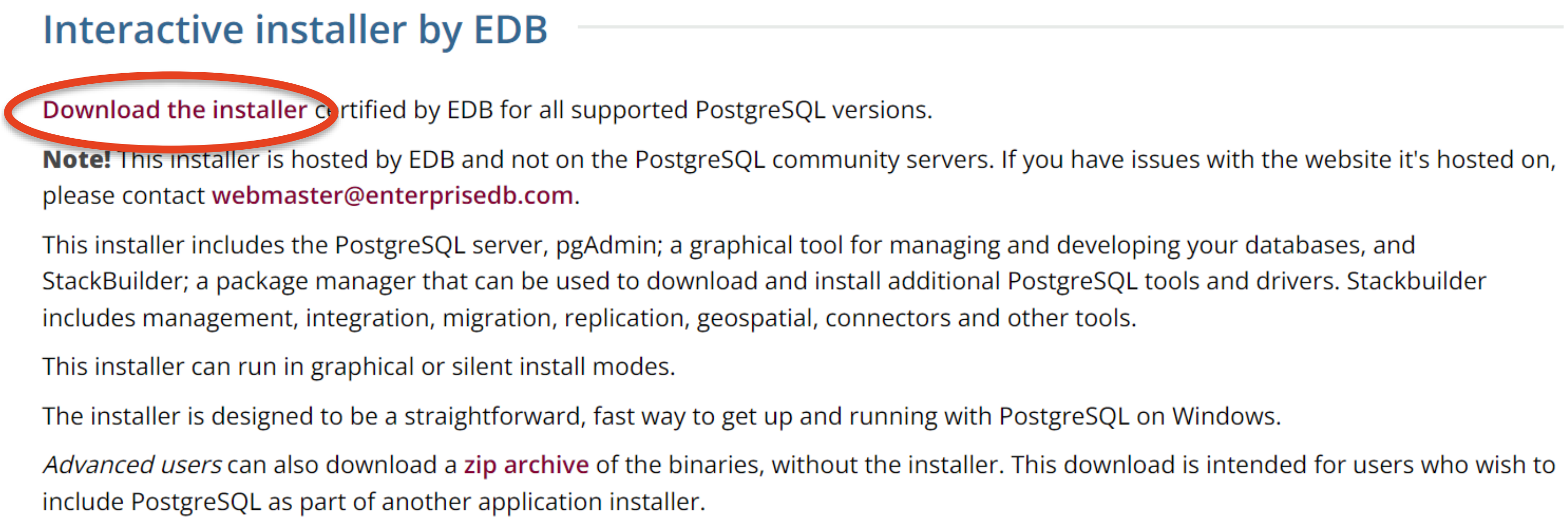


C:\WINDOWS\System32\WindowsPowerShell\v1.0\powershell.exe

```
(base) PS C:\Users\danir> conda install -c anaconda postgresql
```

Installing PostgreSQL on your computer

- Go to https://www.postgresql.org/download/
- Choose the right option for your Operating System and click on **download the installer**.



Interactive installer by EDB

Download the installer certified by EDB for all supported PostgreSQL versions.

Note! This installer is hosted by EDB and not on the PostgreSQL community servers. If you have issues with the website it's hosted on, please contact webmaster@enterprisedb.com.

This installer includes the PostgreSQL server, pgAdmin; a graphical tool for managing and developing your databases, and StackBuilder; a package manager that can be used to download and install additional PostgreSQL tools and drivers. Stackbuilder includes management, integration, migration, replication, geospatial, connectors and other tools.

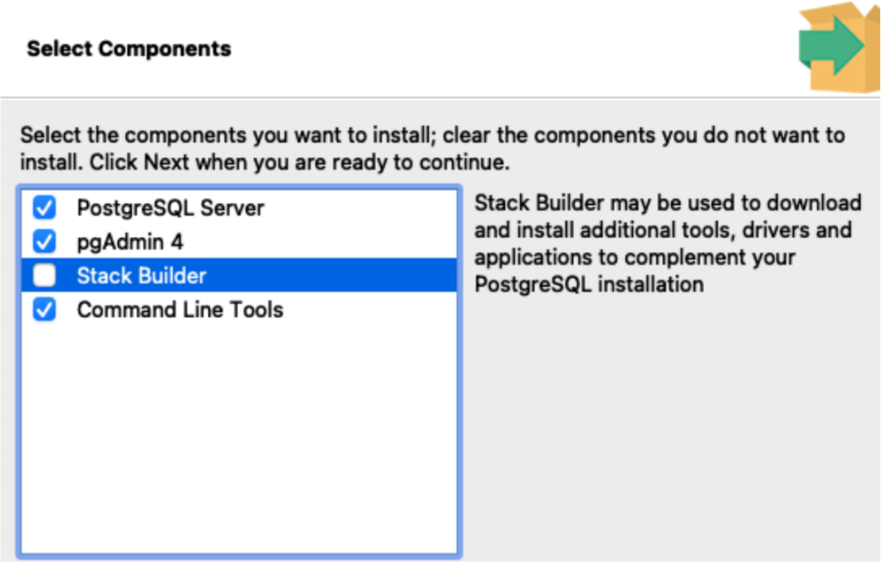
This installer can run in graphical or silent install modes.

The installer is designed to be a straightforward, fast way to get up and running with PostgreSQL on Windows.

Advanced users can also download a zip archive of the binaries, without the installer. This download is intended for users who wish to include PostgreSQL as part of another application installer.

Installing PostgreSQL on your computer

- Select the latest PostgreSQL version available for your OS and download it.
- Open the installer and follow the instructions to install.
  - In "Select Components", select:
    - The PostgreSQL server.
    - pgAdmin.
    - Command Line Tools.
  - In "Password"
    - Type a password.
    - **Remember this password!**



Select Components

Select the components you want to install; clear the components you do not want to install. Click Next when you are ready to continue.

☑ PostgreSQL Server
☑ pgAdmin 4
☐ Stack Builder
☑ Command Line Tools

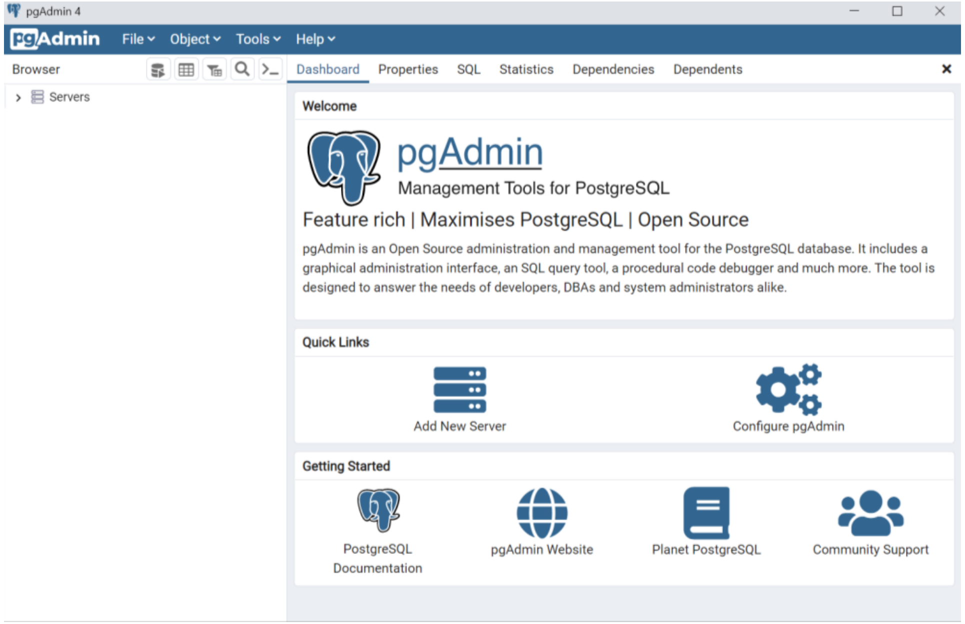
Stack Builder may be used to download and install additional tools, drivers and applications to complement your PostgreSQL installation

Launching PostgreSQL

- Open pgAdmin4.

pgAdmin 4

- Type in the **password** you used during the installation.
- After entering your password, you will see the pgAdmin homepage.

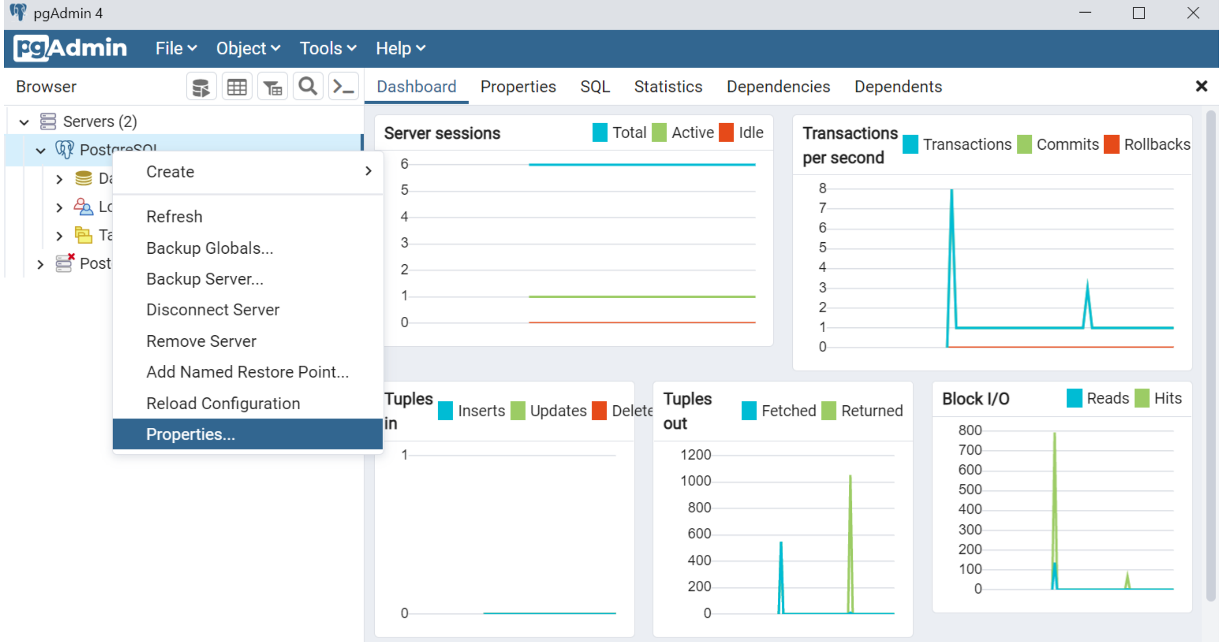


Feature rich | Maximises PostgreSQL | Open Source

pgAdmin is an Open Source administration and management tool for the PostgreSQL database. It includes a graphical administration interface, an SQL query tool, a procedural code debugger and much more. The tool is designed to answer the needs of developers, DBAs and system administrators alike.

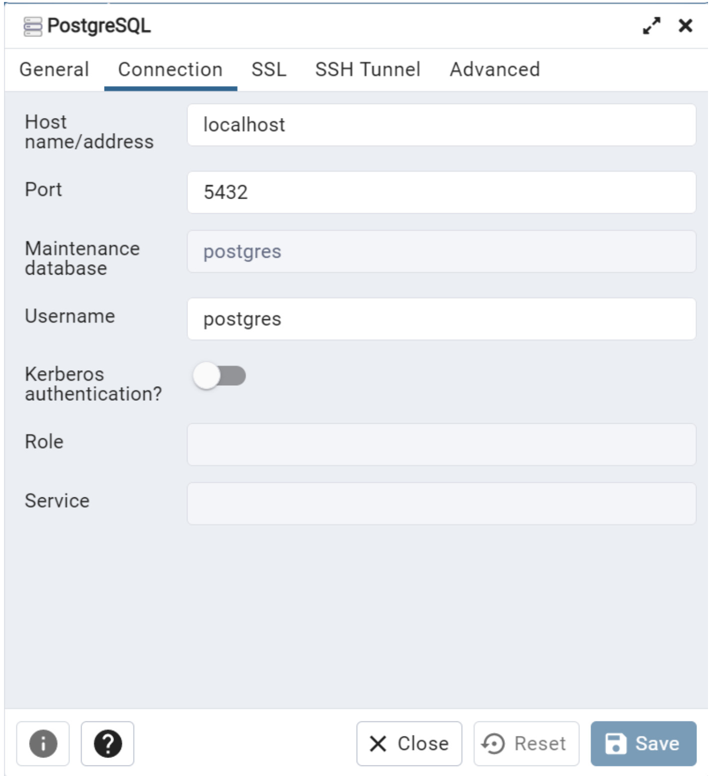
Launching PostgreSQL

– Find PostgreSQL Server in the left tree, right-click 'PostgresSQL', and select 'Properties'.



Launching PostgreSQL

- Go to ‘Connection’ and select:
  - host name: localhost
  - port: 5432
  - username: postgres



Host name/address: localhost

Port: 5432

Maintenance database: postgres

Username: postgres

Kerberos authentication?

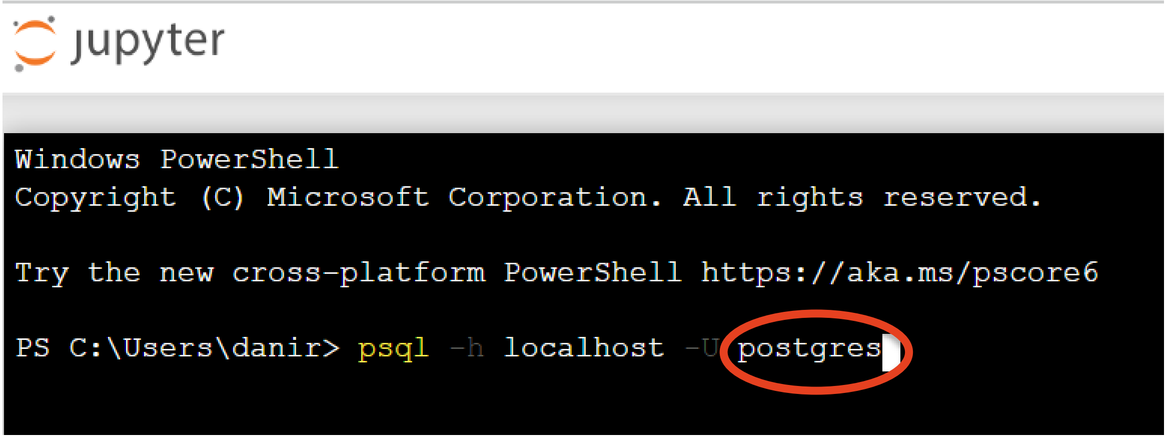
Role:

Service:

Connecting to the server

Option 1

- Open Jupyter Notebook, and select New > Terminal.
- Once a new terminal is opened, run the following command:
  - **`psql -h localhost -U <username>`**



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\danir> psql -h localhost -U postgres
```

Connecting to the server

Option 2

- Only for the **first time you connect**, in your Jupyter Notebook, type:
  - **`!pip install psycopg2`**
- To connect to the database server, copy the code on the right to your Jupyter Notebook (Do not forget to **change the password** to your own password)

```
import psycopg2
def pgconnect():
      """ Connect to the PostgreSQL database server """
      conn = None
      try:
             # connect to the PostgreSQL server
             print('Connecting to the PostgreSQL database...’)
             conn = psycopg2.connect(host = 'localhost’,
                                 database = 'postgres’,
                                 user = 'postgres’,
                                 password = 'abcd1234’)
             print("connected")

      except Exception as e:
             print("unable to connect to the database")
             print(e)
      return conn
conn = pgconnect()
```

Data exploration with Python

THE UNIVERSITY OF
SYDNEY

Activity

- In Canvas, go to:
  - Exercise: Data Acquisition and Data Cleaning with Python.
- Download Jupyter Notebook:
  - data_exploration_with_Python.ipynb
- Download WFH survey responses:
  - WFH-Survey-Responses-NSW.csv

Project Stage 0

Project Stage 0

- Identify and explore possible problems and datasets.
- Select project dataset, define problem, submit a project proposal before April 23rd.
- For those without own proposal, a project will be chosen and assigned.

Where to find a dataset/problem

- WHO - World health organisation
- The World Bank
- International Labour Organization
- United Nations
- OECD
- Earth Data (NASA)
- UNICEF
- Australian Government
- Australian Bureau of Statistics
- Research Data Australia
- UCI Machine Learning Repository
- Kaggle Datasets
- AIHW Data

QUESTIONS?