Python + SQL

Dr. Andrea Villanes

A few things you learned in Fall 1…

- What is SQL
- How to use SQL to query data in SQLite
- How to use SQL in SAS
- Python
- Pandas

How can we use SQL in Python?

Let’s talk about Pandas…

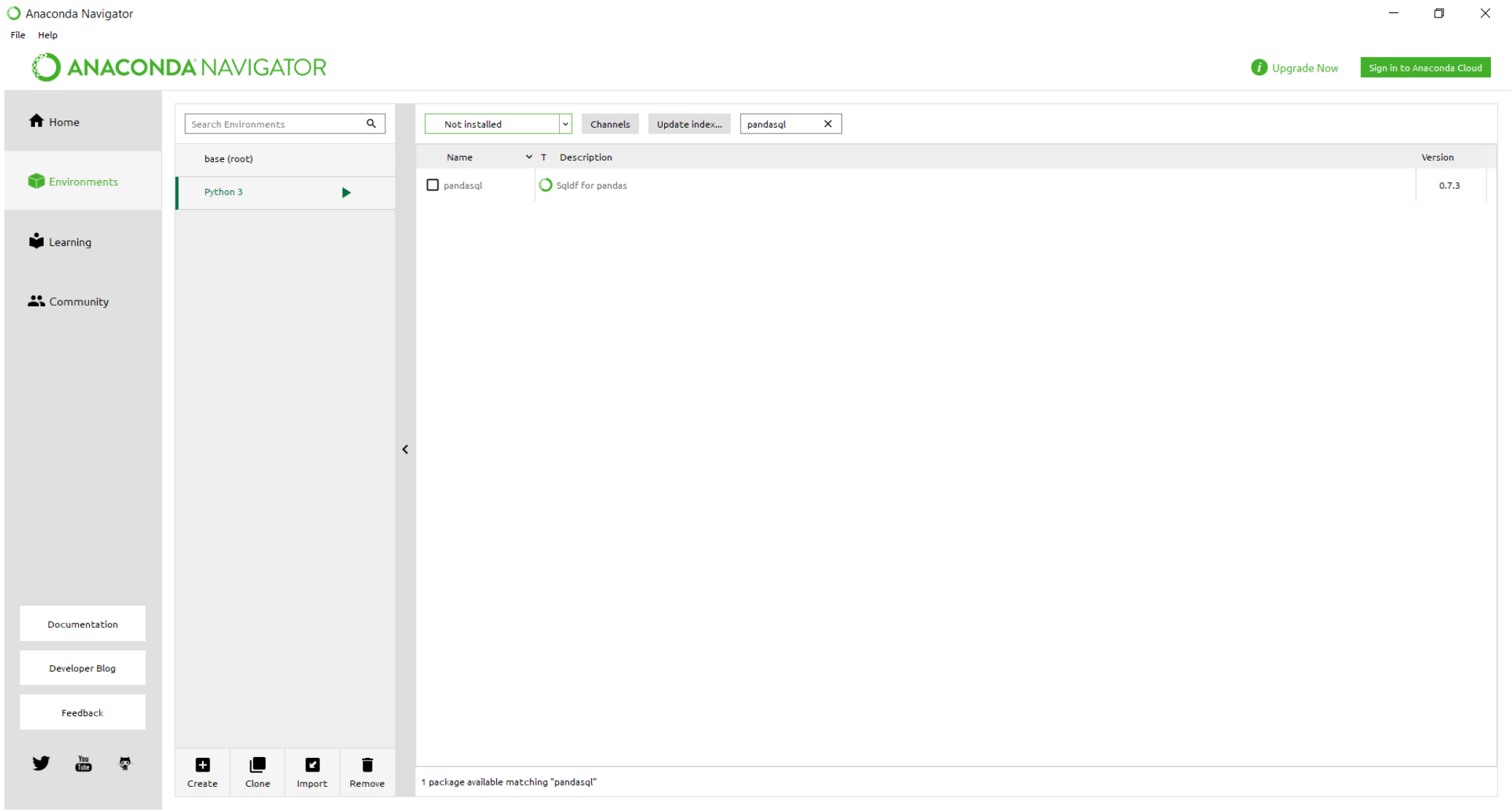
- Pandas is one of the most preferred and widely used tools in data wrangling if not the most used one
- Built in SQL support
- Uses underlying SQLite
- You can treat Pandas dataframes as if they were tables
- Documentation: https://pypi.org/project/pysqldf/
- “Do I have to use SQL Pandas?” → https://pandas.pydata.org/docs/getting_started/comparison/comparison_with_sql.html -- Check the R, SAS links as well

Install pandasql

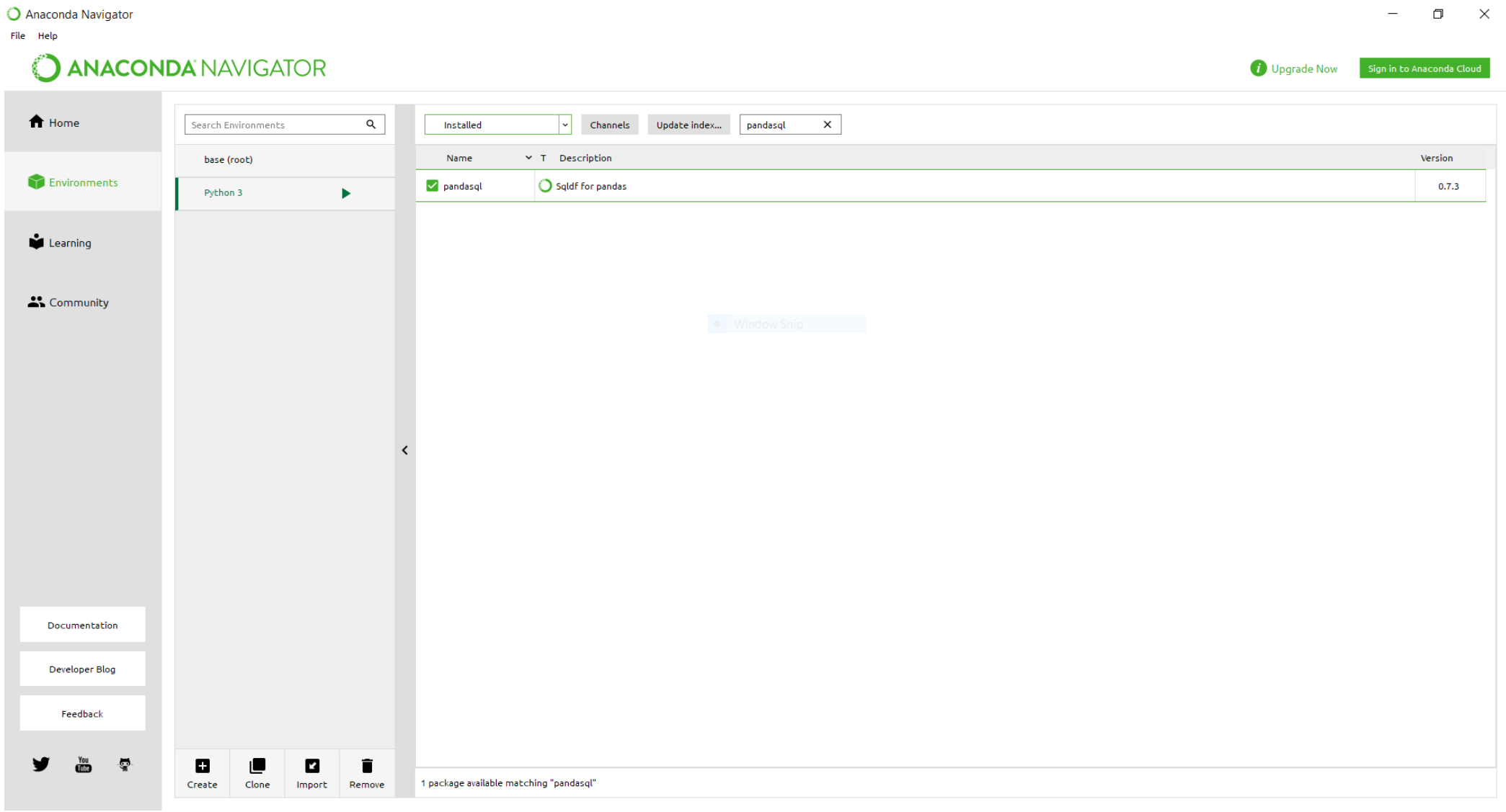
In [1]:
```
import pandas as pd
from pandasql import sqldf
from pandasql import load_births

births = load_births()

print(sqldf("SELECT * FROM births where births > 250000 limit 5;", locals()))
```

```
---------------------------------------------------------------------------
ImportError                               Traceback (most recent call last)
<ipython-input-1-b72c4429f14a> in <module>()
      1 import pandas as pd
----> 2 from pandasql import sqldf
      3 from pandasql import load_births
      4
      5 births = load_births()

ImportError: No module named 'pandasql'
```

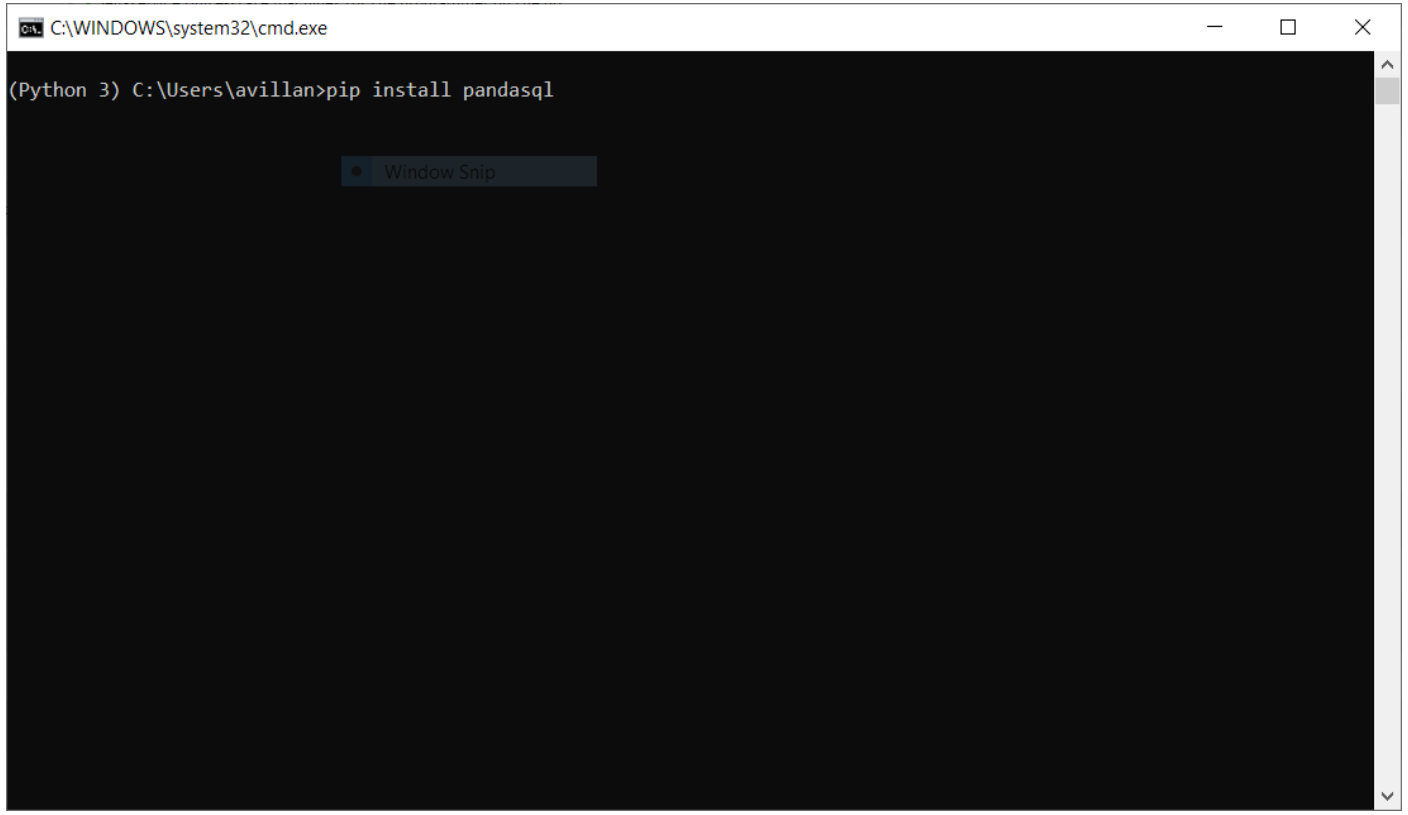
Let’s install the library we will need



Once it’s installed…



Another option: pip install pandasql



If you have Jupiter Notebook opened, closed everything and start again after pip install finishes

Resources we will use:

- Go to: https://github.com/andreavillanes/Python-SQL
- You will see 5 files:
  - Pandas + SQL.ipynb → I will use this file during class
  - Pandas + SQL - Exercise.ipynb → you will complete this Jupyter Notebook, and upload it to Moodle
  - AWProduct.csv & AWProductSubcategory.csv → the two files you need to complete the Exercise
  - Pandas: comparison with SQL