Clémence André

Hans on Data science tools

ExerciseDBT

Analytics Engineering Exercise with DBT, PostgreSQL, and Metabase

This comprehensive exercise guides you through setting up a data pipeline with PostgreSQL, DBT, and Metabase. You'll start by importing data, then move on to modeling the data with DBT and visualizing insights with Metabase. Refere to dbt documentation to understand the different commands : https://docs.getdbt.com/docs/build/projects.

1. Setting Up the DBT Project

Install Dependencies

Create a Virtual Environment

```
C:\Users\cleme>python -m venv dbt-env

C:\Users\cleme>dbt-env\Scripts\activate
```

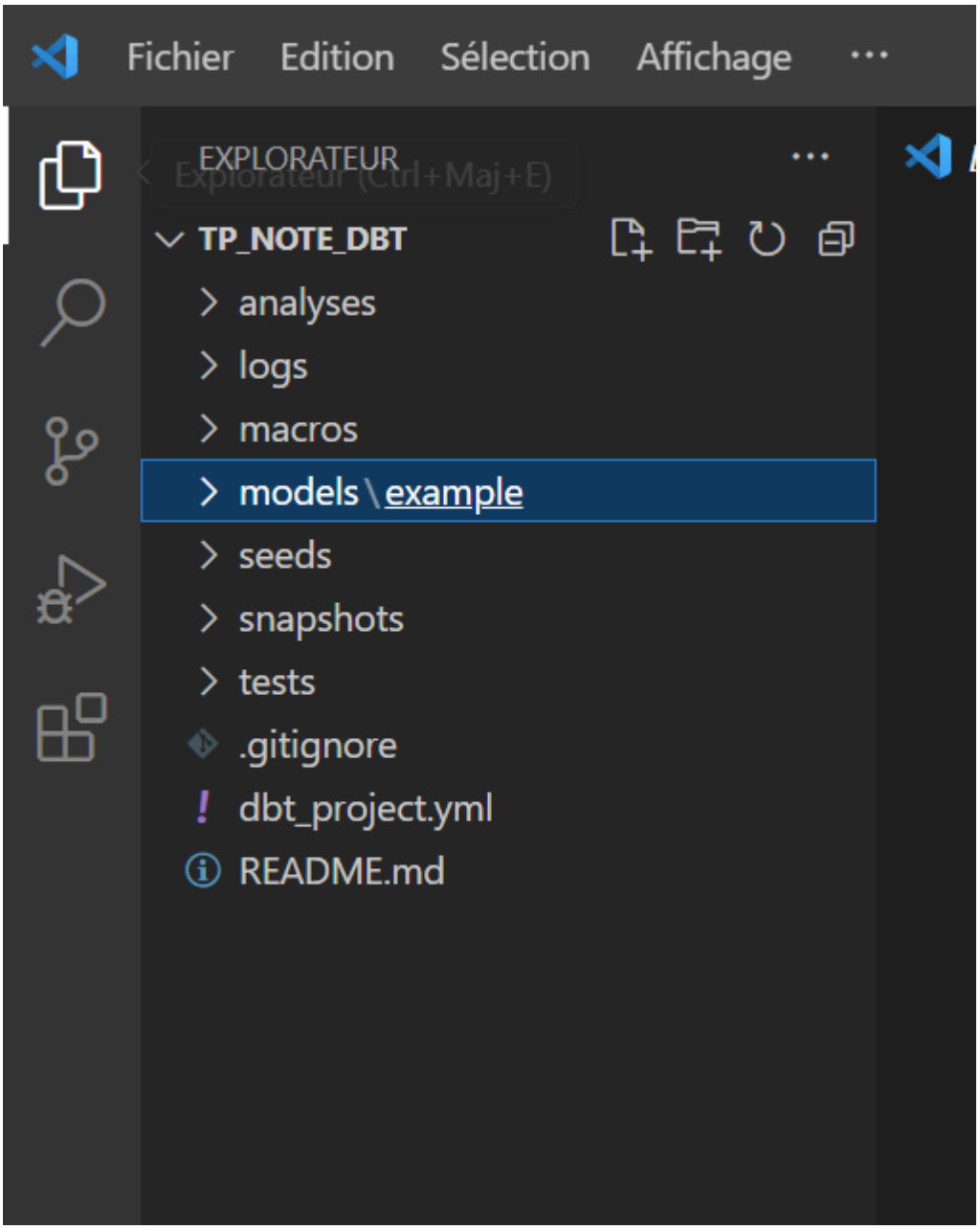
```
(dbt-env) C:\Users\cleme>pip install dbt-core dbt-postgres
Collecting dbt-core
  Downloading dbt_core-1.7.10-py3-none-any.whl (1.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.0/1.0 MB 9.3 MB/s eta 0:00:00
Collecting dbt-postgres
  Downloading dbt_postgres-1.7.10-py3-none-any.whl (28 kB)
Collecting jsonschema>=3.0
  Downloading jsonschema-4.21.1-py3-none-any.whl (85 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 85.5/85.5 KB ? eta 0:00:00
Collecting typing-extensions>=3.7.4
  Downloading typing_extensions-4.10.0-py3-none-any.whl (33 kB)
Collecting idna<4,>=2.5
  Downloading idna-3.6-py3-none-any.whl (61 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.6/61.6 KB 3.4 MB/s eta 0:00:00
Collecting minimal-snowplow-tracker~=0.0.2
  Using cached minimal_snowplow_tracker-0.0.2-py3-none-any.whl
Collecting mashumaro[msgpack]~=3.9
  Downloading mashumaro-3.12-py3-none-any.whl (89 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 89.9/89.9 KB 5.0 MB/s eta 0:00:00
Collecting protobuf<5,>=4.0.0
  Downloading protobuf-4.25.3-cp39-cp39-win_amd64.whl (413 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 413.4/413.4 KB 25.2 MB/s eta 0:00:00
Collecting colorama<0.5,>=0.3.9
  Downloading colorama-0.4.6-py2.py3-none-any.whl (25 kB)
```

Configure dbt profile

```
(dbt-env) C:\Users\cleme>dbt init
13:48:14  Running with dbt=1.7.10
Enter a name for your project (letters, digits, underscore): TP_note_DBT
13:48:18
Your new dbt project "TP_note_DBT" was created!

For more information on how to configure the profiles.yml file,
please consult the dbt documentation here:

  https://docs.getdbt.com/docs/configure-your-profile

One more thing:

Need help? Don't hesitate to reach out to us via GitHub issues or on Slack:

  https://community.getdbt.com/

Happy modeling!

13:48:18  Setting up your profile.
Which database would you like to use?
[1] postgres

(Don't see the one you want? https://docs.getdbt.com/docs/available-adapters)

Enter a number: 1
host (hostname for the instance): 172.17.144.1
port [5432]: 5432
user (dev username):
```

```
(dbt-env) C:\Users\cleme\TP_note_DBT>dbt debug
14:12:47  Running with dbt=1.7.10
14:12:47  dbt version: 1.7.10
14:12:47  python version: 3.9.13
14:12:47  python path: C:\Users\cleme\dbt-env\Scripts\python.exe
14:12:47  os info: Windows-10-10.0.22631-SP0
14:12:48  Using profiles dir at C:\Users\cleme\.dbt
14:12:48  Using profiles.yml file at C:\Users\cleme\.dbt\profiles.yml
14:12:48  Using dbt_project.yml file at C:\Users\cleme\TP_note_DBT\dbt_project.yml
14:12:48  adapter type: postgres
14:12:48  adapter version: 1.7.10
14:12:48  Configuration:
14:12:48    profiles.yml file [OK found and valid]
14:12:48    dbt_project.yml file [OK found and valid]
14:12:48  Required dependencies:
14:12:48   - git [OK found]
```

```
14:12:48  Connection:
14:12:48    host: 172.17.144.1
14:12:48    port: 5432
14:12:48    user: postgres
14:12:48    database: postgres
14:12:48    schema: public
14:12:48    connect_timeout: 10
14:12:48    role: None
14:12:48    search_path: None
14:12:48    keepalives_idle: 0
14:12:48    sslmode: None
14:12:48    sslcert: None
14:12:48    sslkey: None
14:12:48    sslrootcert: None
14:12:48    application_name: dbt
14:12:48    retries: 1
14:12:48  Registered adapter: postgres=1.7.10
14:12:48    Connection test: [OK connection ok]

14:12:48  All checks passed!

(dbt-env) C:\Users\cleme\TP_note_DBT>
```


Fichier
Edition
Sélection
Affichage
EXPLORATEUR
Explorateur (Ctrl+Maj+E)
TP_NOTE_DBT
analyses
logs
macros
models \ example
seeds
snapshots
tests
.gitignore
dbt_project.yml
README.md

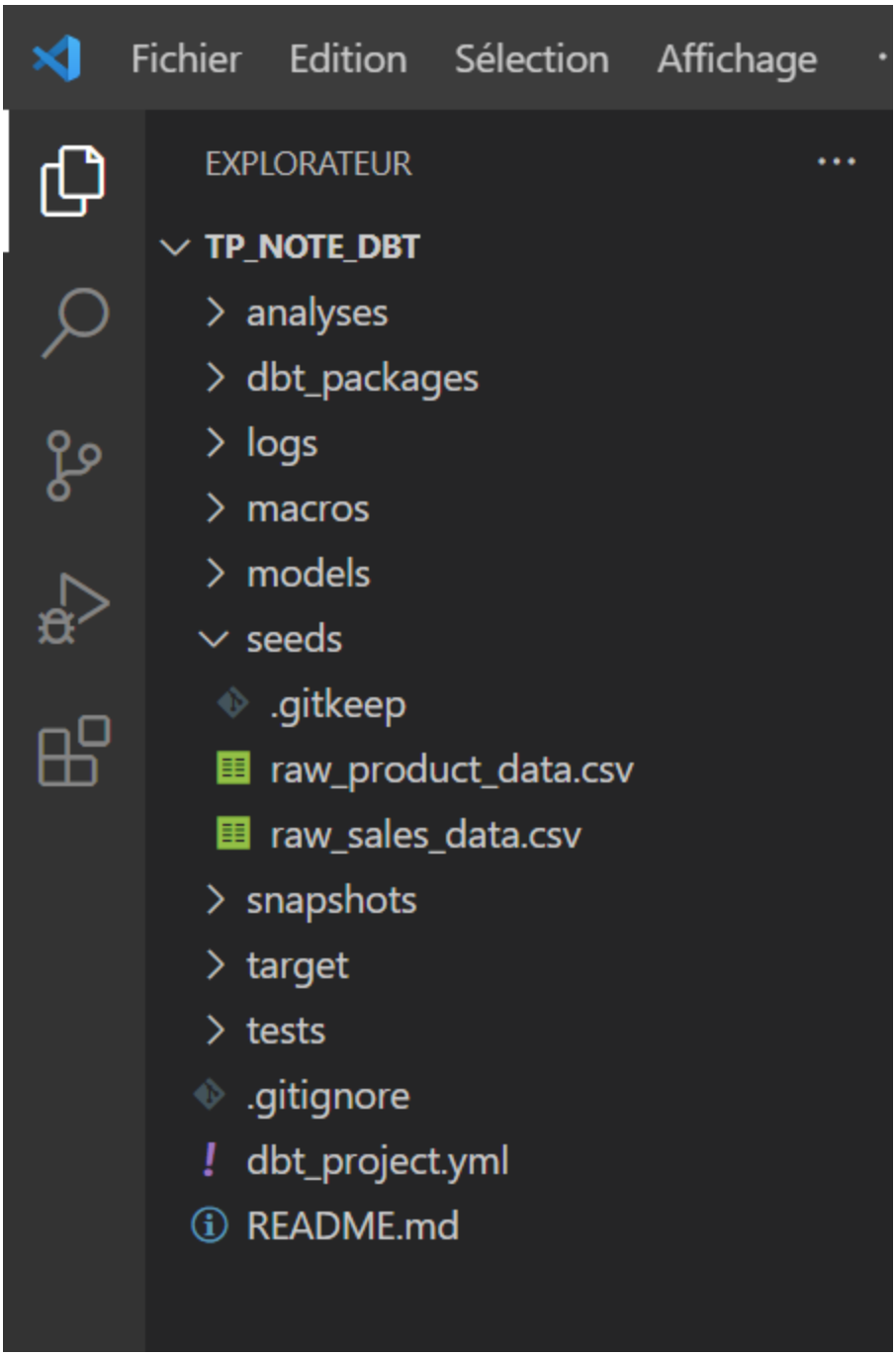
1. Preparing the Raw data

For simplicity, we'll use raw data from csv files. In the real life, data would have been a raw table.

> Ensure you've downloaded the two csv files and store them under `seeds` directory in your projec

Simply run the following command : `dbt seed`. It should create two views :

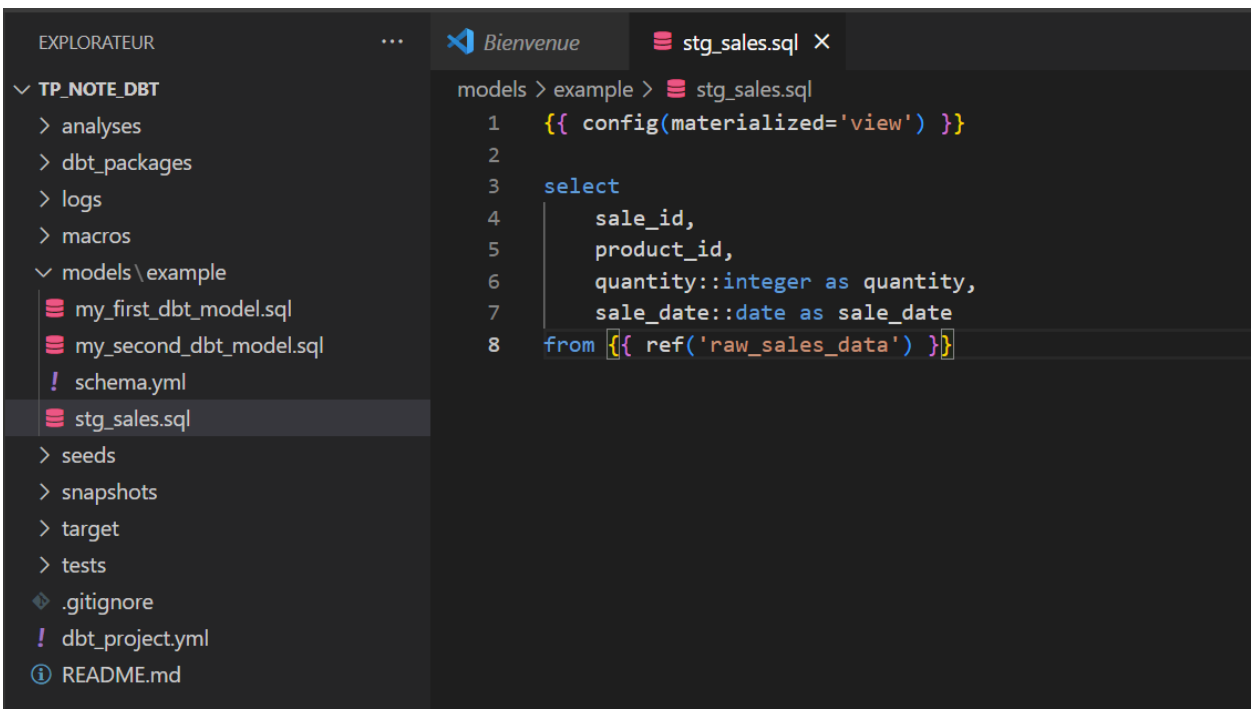
- `raw_product_data`
- `raw_sales_data`


Fichier
Edition
Sélection
Affichage
EXPLORATEUR
TP_NOTE_DBT
analyses
dbt_packages
logs
macros
models
seeds
.gitkeep
raw_product_data.csv
raw_sales_data.csv
snapshots
target
tests
.gitignore
dbt_project.yml
README.md

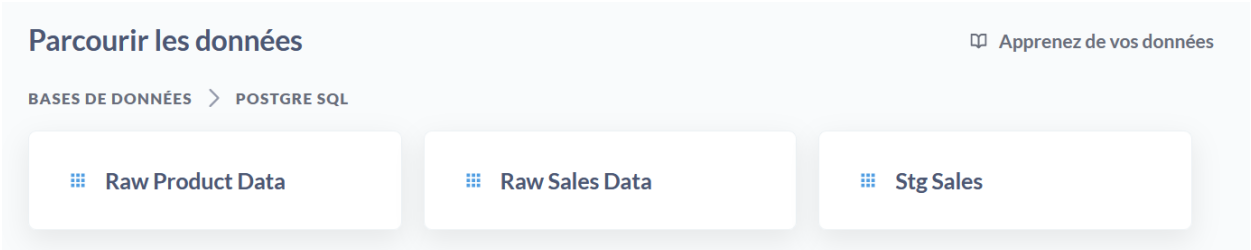
```
(dbt-env) PS C:\Users\cleme\TP_note_DBT> dbt seed
14:23:16  Running with dbt=1.7.10
14:23:17  Registered adapter: postgres=1.7.10
14:23:17  Unable to do partial parsing because saved manifest not found. Starting full parse.
14:23:18  Found 2 models, 2 seeds, 4 tests, 0 sources, 0 exposures, 0 metrics, 401 macros, 0 groups, 0 semantic models
14:23:18
14:23:18  Concurrency: 1 threads (target='dev')
14:23:18
14:23:18  1 of 2 START seed file public.raw_product_data ................................ [RUN]
14:23:18  1 of 2 OK loaded seed file public.raw_product_data ............................ [INSERT 4 in 0.15s]
14:23:18  2 of 2 START seed file public.raw_sales_data ................................... [RUN]
14:23:19  2 of 2 OK loaded seed file public.raw_sales_data ............................... [INSERT 500 in 0.86s]
14:23:19
14:23:19  Finished running 2 seeds in 0 hours 0 minutes and 1.21 seconds (1.21s).
14:23:19
14:23:19  Completed successfully
14:23:19
14:23:19  Done. PASS=2 WARN=0 ERROR=0 SKIP=0 TOTAL=2
```

3. Creating Staging Models

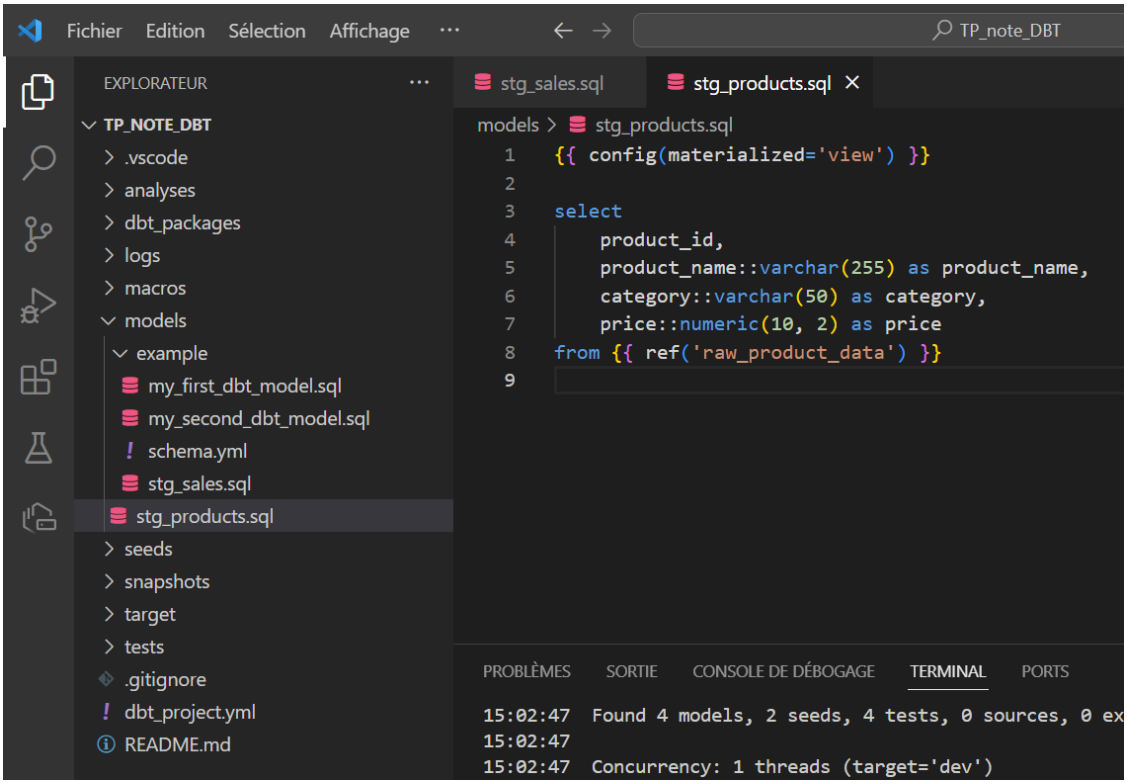
3.1 Staging Sales Data (stg_sales.sql)

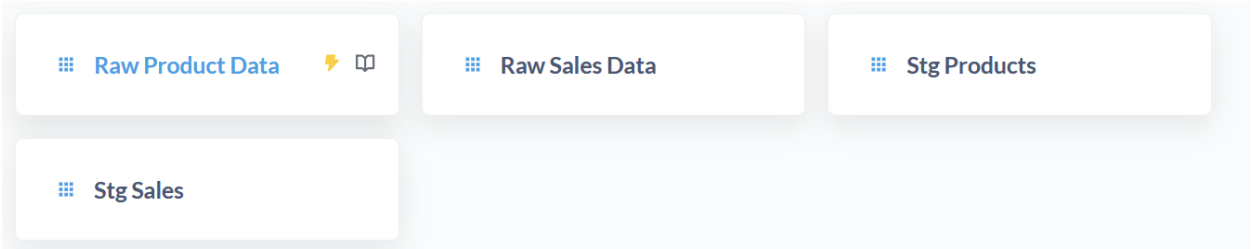


```
(dbt-env) PS C:\Users\cleme\TP_note_DBT> dbt run --select stg_sales
14:42:45  Running with dbt=1.7.10
14:42:45  Registered adapter: postgres=1.7.10
14:42:46  Found 3 models, 2 seeds, 4 tests, 0 sources, 0 exposures, 0 metrics, 401 macros, 0 groups, 0 semantic models
14:42:46
14:42:46  Concurrency: 1 threads (target='dev')
14:42:46
14:42:46  1 of 1 START sql view model public.stg_sales ..................................... [RUN]
14:42:46  1 of 1 OK created sql view model public.stg_sales ............................... [CREATE VIEW in 0.22s]
14:42:46
14:42:46  Finished running 1 view model in 0 hours 0 minutes and 0.51 seconds (0.51s).
14:42:46
14:42:46  Completed successfully
14:42:46
14:42:46  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
(dbt-env) PS C:\Users\cleme\TP_note_DBT>
```



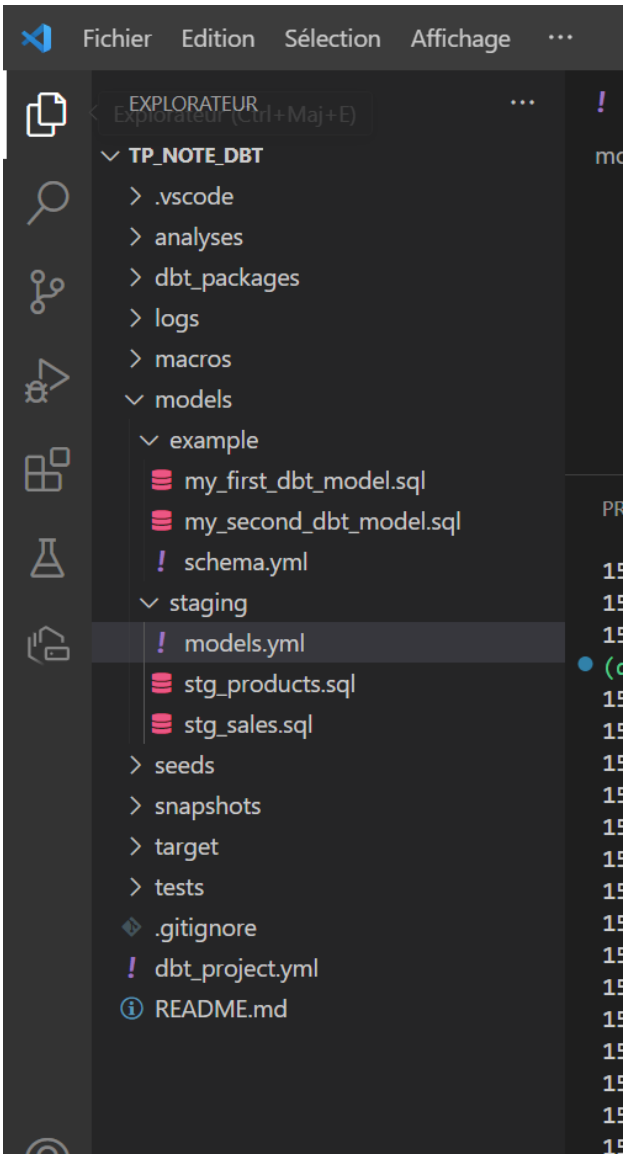
3.2 Staging Product Data (stg_products.sql)

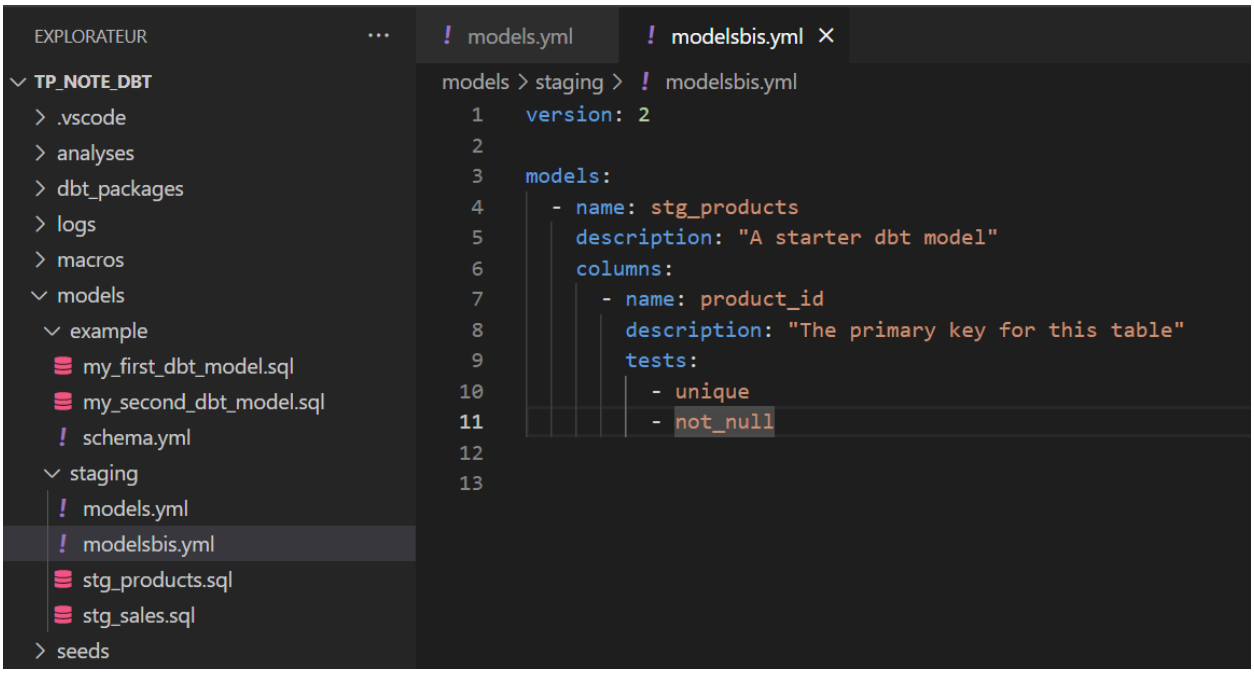




3.3 Testing Staging Models

```
(dbt-env) PS C:\Users\cleme\TP_note_DBT> dbt test --select stg_sales
15:06:34  Running with dbt=1.7.10
15:06:35  Registered adapter: postgres=1.7.10
15:06:36  Found 4 models, 2 seeds, 6 tests, 0 sources, 0 exposures, 0 metrics, 401 macros, 0 groups, 0 semantic models
15:06:36
15:06:36  Concurrency: 1 threads (target='dev')
15:06:36
15:06:36  1 of 2 START test not_null_stg_sales_sale_id ................................ [RUN]
15:06:36  1 of 2 PASS not_null_stg_sales_sale_id ...................................... [PASS in 0.16s]
15:06:36  2 of 2 START test unique_stg_sales_sale_id .................................. [RUN]
15:06:36  2 of 2 PASS unique_stg_sales_sale_id ........................................ [PASS in 0.11s]
15:06:36
15:06:36  Finished running 2 tests in 0 hours 0 minutes and 0.61 seconds (0.61s).
15:06:36
15:06:36  Completed successfully
15:06:36
15:06:36  Done. PASS=2 WARN=0 ERROR=0 SKIP=0 TOTAL=2
(dbt-env) PS C:\Users\cleme\TP_note_DBT>
```


Fichier
Edition
Sélection
Affichage
EXPLORATEUR
Explorateur (Ctrl+Maj+E)
TP_NOTE_DBT
.vscode
analyses
dbt_packages
logs
macros
models
example
my_first_dbt_model.sql
my_second_dbt_model.sql
schema.yml
staging
models.yml
stg_products.sql
stg_sales.sql
seeds
snapshots
target
tests
.gitignore
dbt_project.yml
README.md




```
(dbt-env) PS C:\Users\cleme\TP_note_DBT> dbt test --select stg_products
15:09:53  Running with dbt=1.7.10
15:09:54  Registered adapter: postgres=1.7.10
15:09:54  Found 4 models, 2 seeds, 8 tests, 0 sources, 0 exposures, 0 metrics, 401 macros, 0 groups, 0 semantic models
15:09:54
15:09:55  Concurrency: 1 threads (target='dev')
15:09:55
15:09:55  1 of 2 START test not_null_stg_products_product_id ............................ [RUN]
15:09:55  1 of 2 PASS not_null_stg_products_product_id ................................... [PASS in 0.14s]
15:09:55  2 of 2 START test unique_stg_products_product_id .............................. [RUN]
15:09:55  2 of 2 PASS unique_stg_products_product_id ..................................... [PASS in 0.10s]
15:09:55
15:09:55  Finished running 2 tests in 0 hours 0 minutes and 0.51 seconds (0.51s).
15:09:55
15:09:55  Completed successfully
15:09:55
15:09:55  Done. PASS=2 WARN=0 ERROR=0 SKIP=0 TOTAL=2
(dbt-env) PS C:\Users\cleme\TP_note_DBT>
```

4. Marts Models

Create models to perform aggregations and enrich the data.

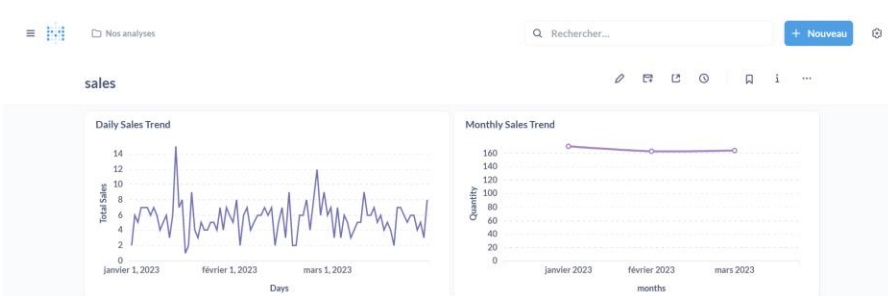
4.2 Example

models > marts > daily_sales_revenue_by_category.sql

```sql
{{ config(materialized='table') }}

select
    s.sale_date,
    p.category,
    sum(s.quantity * p.price) as daily_revenue
from {{ ref('stg_sales') }} s
join {{ ref('stg_products') }} p
    on s.product_id = p.product_id
group by s.sale_date, p.category
```

```
(dbt-env) PS C:\Users\cleme\TP_note_DBT> dbt run --models daily_sales_revenue_by_category
>>
15:33:33  Running with dbt=1.7.10
15:33:33  Registered adapter: postgres=1.7.10
15:33:34  Found 6 models, 2 seeds, 8 tests, 0 sources, 0 exposures, 0 metrics, 401 macros, 0 groups, 0 semantic models
15:33:34
15:33:34  Concurrency: 1 threads (target='dev')
15:33:34
15:33:34  1 of 1 START sql table model public.daily_sales_revenue_by_category ............ [RUN]
15:33:34  1 of 1 OK created sql table model public.daily_sales_revenue_by_category ....... [SELECT 169 in 0.23s]
15:33:34
15:33:34  Finished running 1 table model in 0 hours 0 minutes and 0.47 seconds (0.47s).
15:33:34
15:33:34  Completed successfully
15:33:34
15:33:34  Done. PASS=1 WARN=0 ERROR=0 SKIP=0 TOTAL=1
(dbt-env) PS C:\Users\cleme\TP_note_DBT>
```

5. Dashboard Creation with Metabase




Sales by Category
Gadgets 49.7%
Widgets 50.3%
84,500
TOTAL


Category Performance Over Time
Gadgets
Widgets
Daily Revenue
600
500
400
300
200
100
0
janvier 1, 2023
février 1, 2023
mars 1, 2023
Sale Date