

Mini-Lesson 2:

Reproducible Example

(Reprex)

Richard Chen

Intro

Communicating clearly is an extremely valuable skill in all aspects of life.

Coding is no exception. Use **reprex**.

What is Reproducible Example (Reprex)?

A **reproducible example (reprex)** is a simplified piece of code designed to help others understand and troubleshoot an issue quickly. Reprex involves:

- code that **actually runs**
- code that **I don't have to run**
- code that **I can easily run**

How to create reprex

1. Isolate the part of your code where the issue happens.
2. Remove unrelated parts to keep things simple.
3. Use dummy data, instead of using real files or large datasets.
4. Make sure all variables and data are defined.

Know where the error is!

The screenshot shows a dark-themed terminal window in VSCode. At the top, there are tabs: PROBLEMS, OUTPUT, DEBUG CONSOLE, TERMINAL (which is underlined), and PORTS. The terminal content is as follows:

```
paid_fraction = paid_fraction.sort_values(ascending=False).reset_index()
print(paid_fraction.head())
paid_fraction.columns = ['vehicle_make', 'fraction_paid']

-----
AttributeError                                     Traceback (most recent call last)
Cell In[7], line 1
----> 1 paid_fraction = paid_fraction.sort_values(ascending=False).reset_index()
      2 print(paid_fraction.head())
      3 paid_fraction.columns = ['vehicle_make', 'fraction_paid']

AttributeError: 'int' object has no attribute 'sort_values'
```

At the bottom left of the terminal area, it says "Quarto: 1.5.57".

Your terminal or output panel has all the information.

Can't see your terminal? Go to Terminal > New Terminal in VSCode menu bar.

Bad example

```
1 paid_fraction = paid_fraction.sort_values(ascending=False).reset_index()  
2 print(paid_fraction.head())  
3 paid_fraction.columns = [ 'vehicle_make', 'fraction_paid' ]
```

- `paid_fraction` is not defined (we don't know what it contains)
- `print()` is not necessary
- Not clear where (and what) the error is

Good example (using reprex)

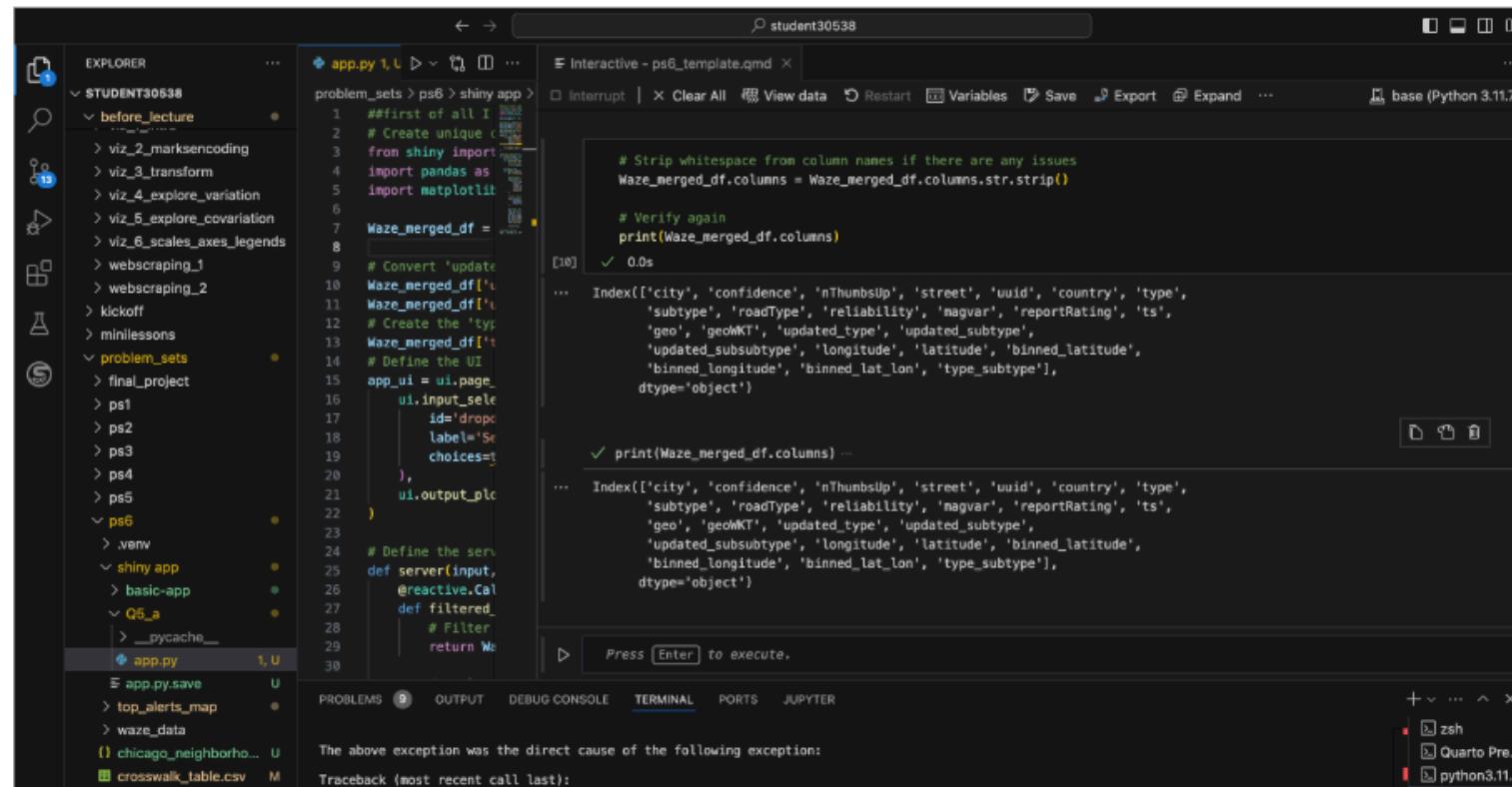
```
1 import pandas as pd
2 df = pd.DataFrame({ "vehicle_make": ["LEXU", "FORD"],
3                     "ticket_queue": ["Paid", "Paid"]})
4
5 ticket_freq = df['vehicle_make'].value_counts()
6 paid_tickets = df[df['ticket_queue'] == 'Paid'].groupby('vehicle_make').size()
7 paid_fraction = dict(paid_tickets / ticket_freq)
8
9 # The following line produces error:
10 # AttributeError: 'dict' object has no attribute 'sort_values'
11 paid_fraction = paid_fraction.sort_values(ascending=False).reset_index()
```

- all variables and data are defined
- focus on where the bug is
- contain only necessary things
- **BUT, doesn't mean you copy the entire code!**

Bad Example (from previous class)

App one is literally not working for me on the most stupid thing, it keeps failing to recognize the "updated-type" and "updated_subtype" columns, even though I printed both of them and they exist, failing to create a dropdown menu.

I have tried everything: changing to strings, creating a list, even GPT does not know what is wrong!



The screenshot shows a Jupyter Notebook interface with the following details:

- EXPLORER** sidebar: Shows a file tree with a folder named `STUDENT30538` containing subfolders like `before_lecture`, `problem_sets`, and `ps6`. Inside `ps6`, there are files for `basic-app`, `Q5_a`, and `pycache`.
- Interactive - ps6_template.qmd**: The main notebook cell contains the following Python code:

```
# Strip whitespace from column names if there are any issues
Waze_merged_df.columns = Waze_merged_df.columns.str.strip()

# Verify again
print(Waze_merged_df.columns)

Index(['city', 'confidence', 'nThumbsUp', 'street', 'uuid', 'country', 'type',
       'subtype', 'roadType', 'reliability', 'magvar', 'reportRating', 'ts',
       'geo', 'geomKT', 'updated_type', 'updated_subtype',
       'updated_subsubtype', 'longitude', 'latitude', 'binned_latitude',
       'binned_longitude', 'binned_lat_lon', 'type_subtype'],
      dtype='object')

print(Waze_merged_df.columns)

Index(['city', 'confidence', 'nThumbsUp', 'street', 'uuid', 'country', 'type',
       'subtype', 'roadType', 'reliability', 'magvar', 'reportRating', 'ts',
       'geo', 'geomKT', 'updated_type', 'updated_subtype',
       'updated_subsubtype', 'longitude', 'latitude', 'binned_latitude',
       'binned_longitude', 'binned_lat_lon', 'type_subtype'],
      dtype='object')
```
- TERMINAL**: Shows the command `base (Python 3.11.7)`.
- PROBLEMS**: Shows 9 problems.
- OUTPUT**: Shows the output of the code execution.
- DEBUG CONSOLE**: Shows the output of the code execution.
- PORTS**: Shows port information.
- JUPYTER**: Shows the status of the Jupyter server.
- Bottom Status Bar**: Shows the message "The above exception was the direct cause of the following exception:" and "Traceback (most recent call last):".

Good Example

```
1 import pandas as pd
2 df = pd.DataFrame({"id": [1, 2, 3], "score": [10, 20, 30]})
3 # This line will create an error ValueError: Cannot set a DataFrame with mu
4 # df["flag"] = df[df["score"] > 15]
```

- all variables and data are defined
- focus on where the bug is
- contain only necessary things

Thank you!