# Validating R DataFrames with Pandera via `reticulate`

Niels Bantilan

2023-03-30

# Pandera in Quarto

```r
library(reticulate)
library(dplyr)

use_condaenv("pandera-nyhackr")
```

# Define a Pandera Schema

```python
import numpy as np
import pandas as pd
import pandera as pa
from pandera.typing import Series


class Schema(pa.DataFrameModel):
    item: Series[str] = pa.Field(isin=["apple", "orange"],
    price: Series[float] = pa.Field(gt=0, coerce=True, nul


python_data = pd.DataFrame.from_records([
    {"item": "orange", "price": 0.75},
    {"item": "orange", "price": np.nan},
])
```

# Define some R data

```r
r_data <- data.frame(
    item = c("apple", "orange", "orange"),
    price = c(0.5, 0.75, NaN)
)
```

# Validate Python and R Data

```r
# validate the python data
print(py$Schema$validate(py$python_data))

# validate an R dataframe
print(py$Schema$validate(r_data))
```

# Synthesize Test Data with Pandera

```r
print(py$Schema$example(size = as.integer(5)))
```