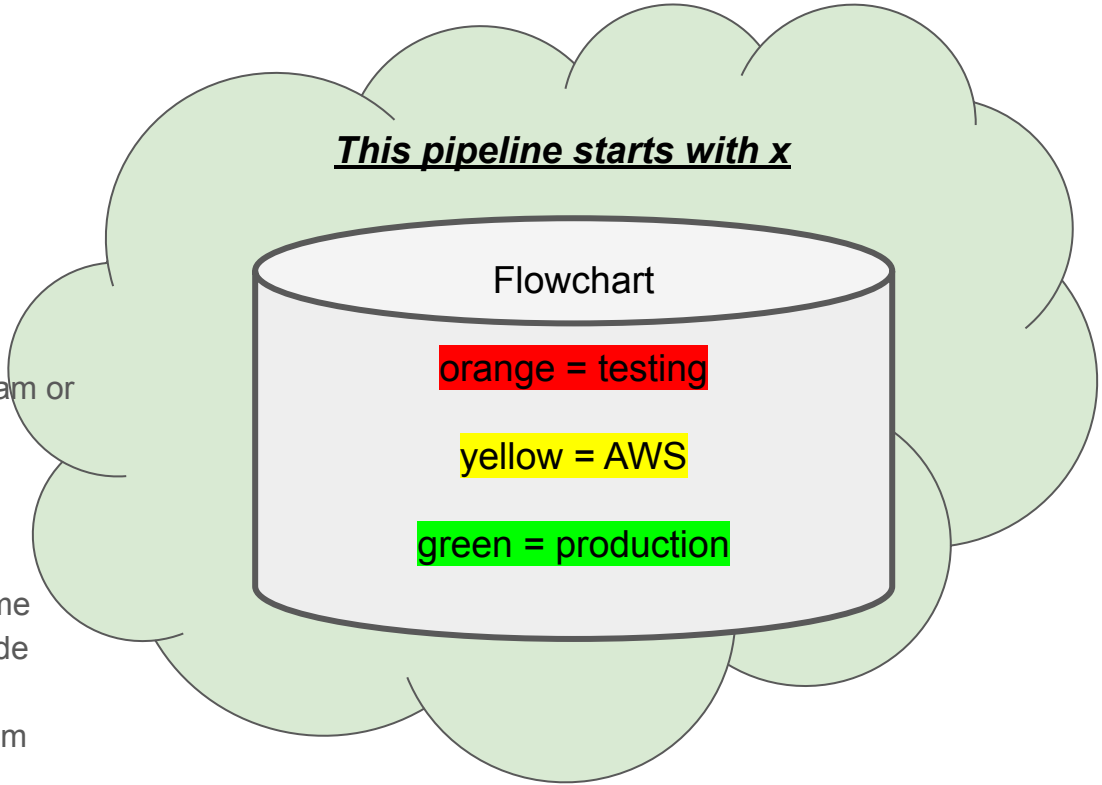
Data /**P**rocessing /**C**leaning /**I**mputation Pipeline

on Real-Time data

Sana Wajid
12/6/19

0P. Step

P = Processing
C = Cleaning
I = Imputation



Assumptions:

- For testing using csv files
- A pipeline can move it's starting point upstream or downstream.

Checks:

1. Listed in order of least to most processing time
2. Numbers will refer to function numbers in code comment or headers in Jupyter notebook
    a. e.g. 1P-2: File contains minimum number of headers

Libraries:

Libraries used in {pandas, scikit-learn} ∈ Python, sed/awk, unix

2P. Raw data conversion to pandas

Assumptions:

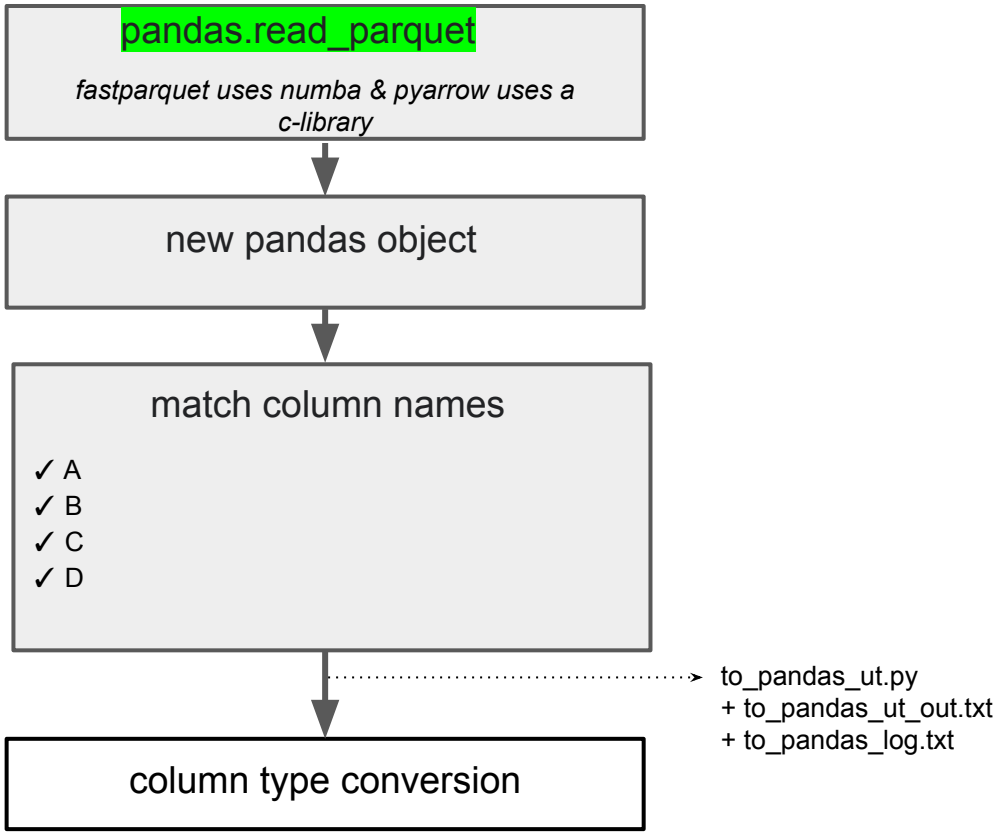
- File contains minimum number of headers
- File contains minimum number of rows
- File contains correct headers
- Column names don’t contain spaces

Functions:

1. `to_pandas`

Libraries:

`read_parquet` ∈ pandas ∈ Python



3C. datetime column type conversion

Assumptions:

- time column is formatted as:

YEAR-MONTH-DAY

space

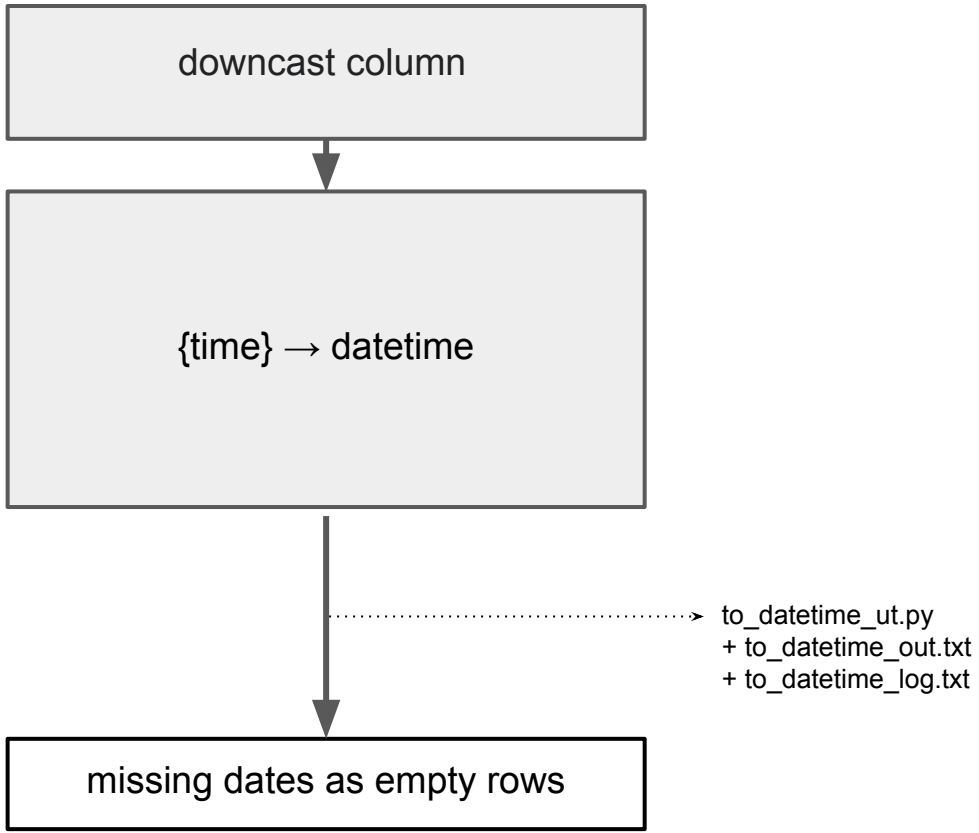
HOUR:MINUTE:SECOND

Functions:

1. `column_to_datetime`

Libraries:

datetime ∈ Python



7I. Impute strategy, simple: scikit-learn

Assumptions:

- fill value is defined per building and per x
- every $t_0 : t_{end}$ has a row

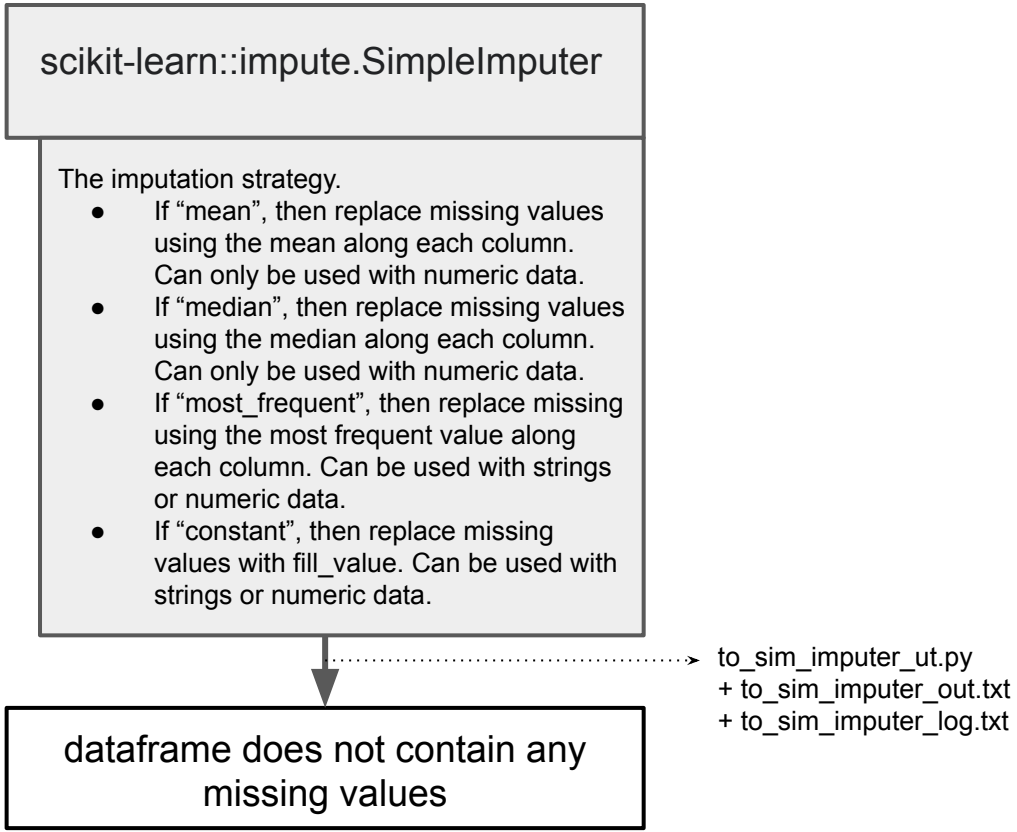
Functions:

accessor (get) functions return dataframe

1. `get_flanking_cluster` (dataframe obj of pandas type)

Libraries:

scikit-learn ∈ Python



7I. Impute strategy, simple: scikit-learn

Assumptions:

- fill value is defined per building and per x
- every $t_0 : t_{end}$ has a row

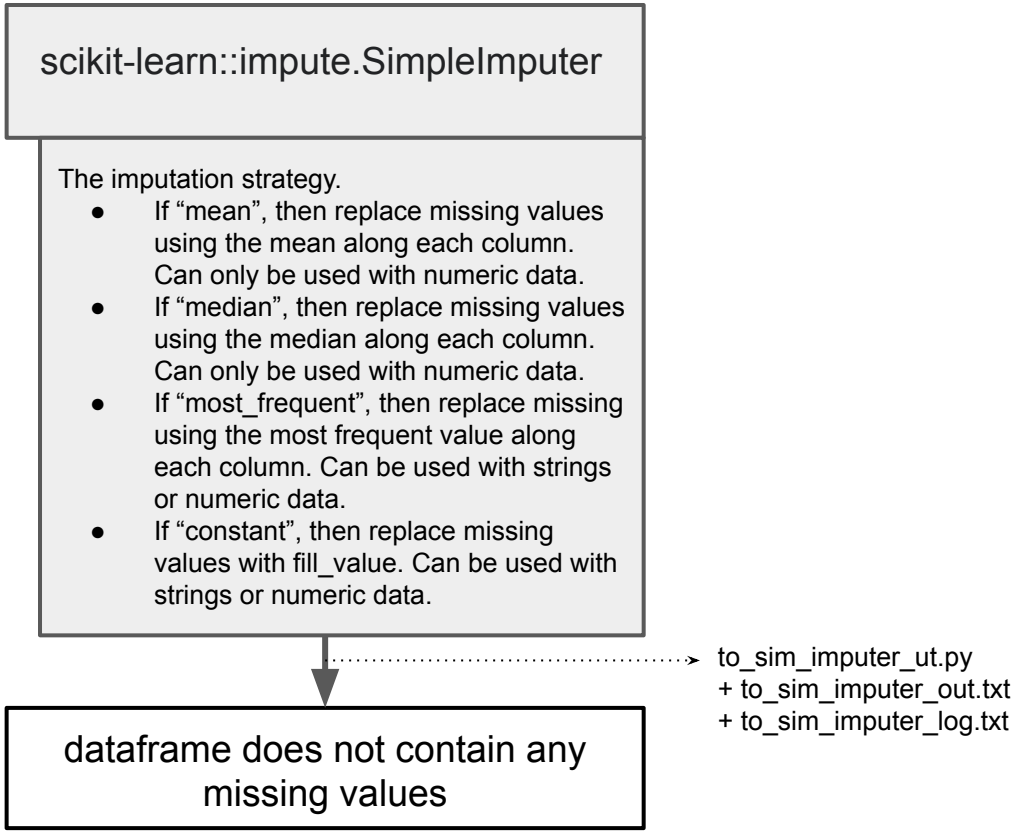
Functions:

mutators return True or False

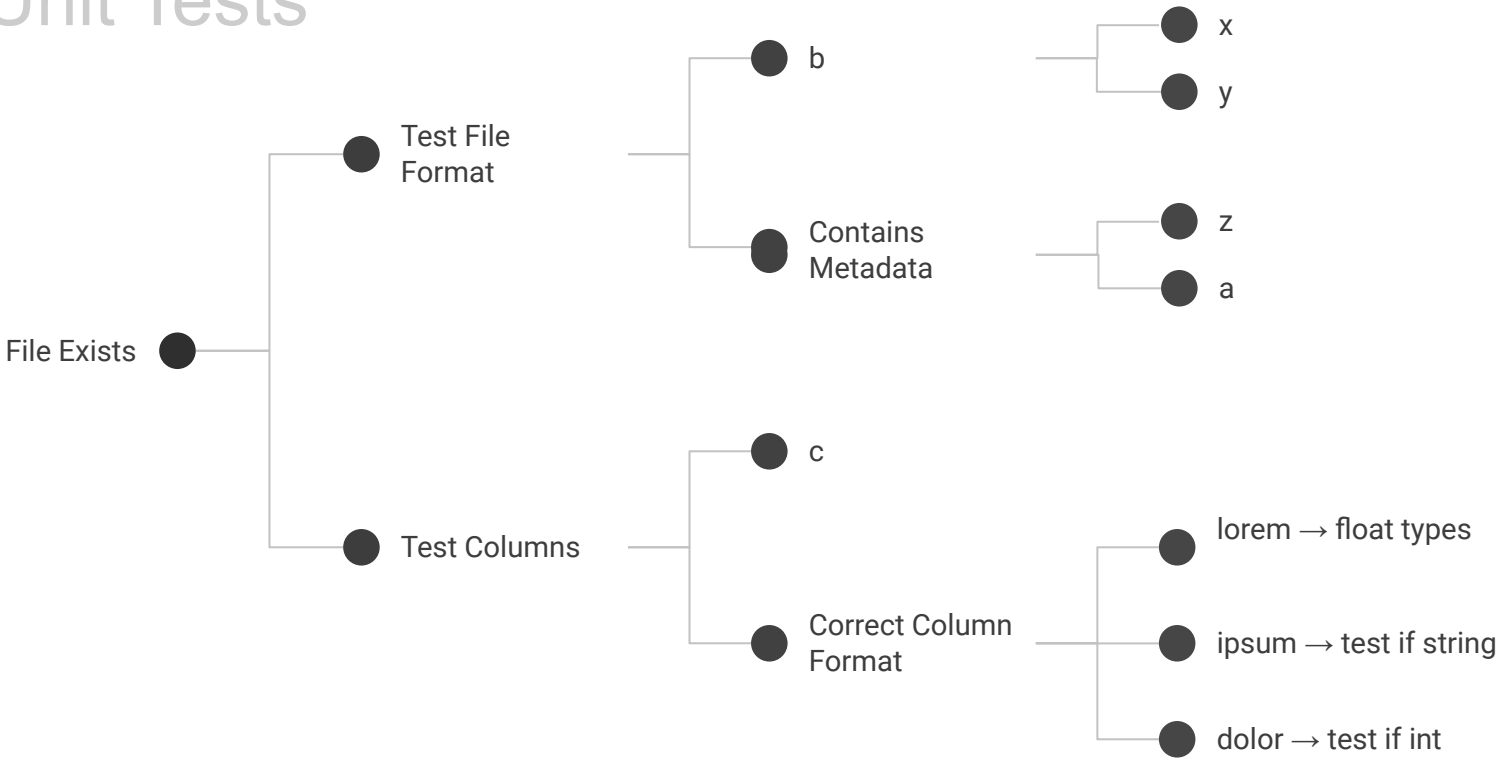
```
2.  impute_cluster_by_mean
3.  impute_cluster_by_median
4.  impute_cluster_by_most_freq
5.  impute_cluster_by_constant
```

Libraries:

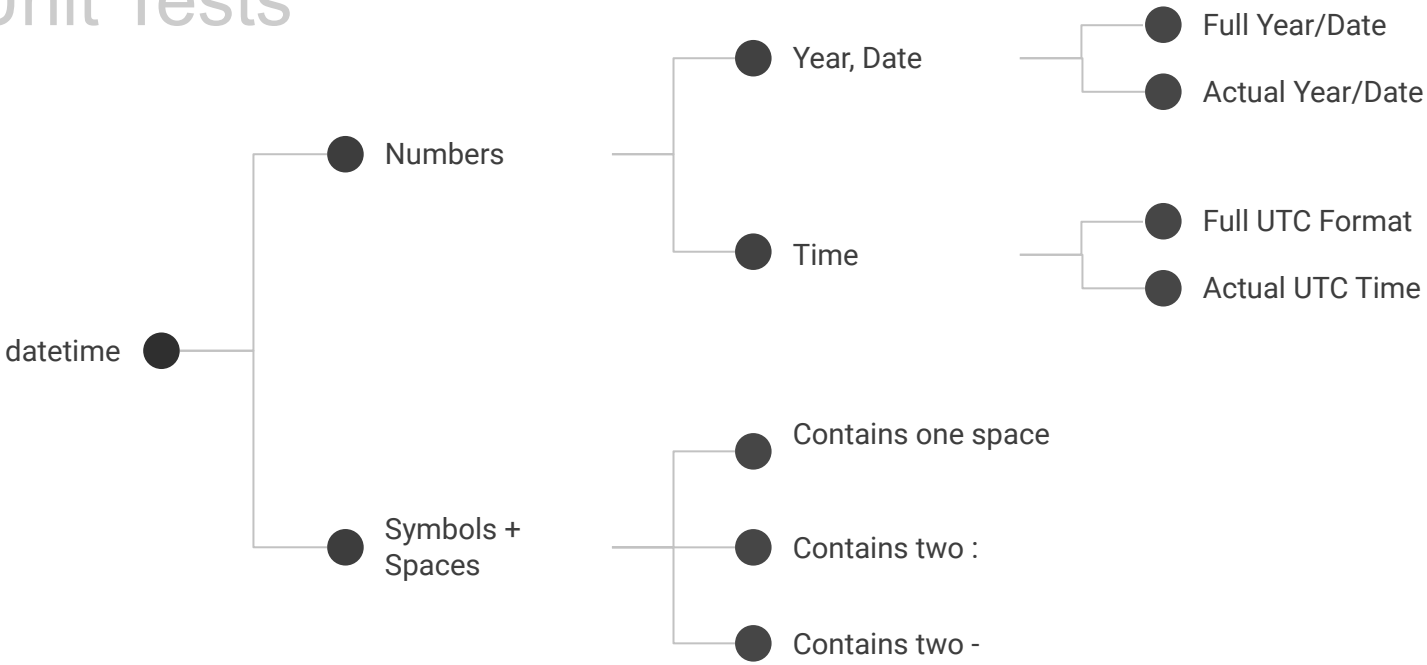
scikit-learn ∈ Python



Unit Tests


File Exists
Test File Format
b
x
y
Contains Metadata
z
a
Test Columns
c
Correct Column Format
lorem → float types
ipsum → test if string
dolor → test if int

Unit Tests


datetime
Numbers
Year, Date
Full Year/Date
Actual Year/Date
Time
Full UTC Format
Actual UTC Time
Symbols + Spaces
Contains one space
Contains two :
Contains two -

References

etc, more will be added