Data Science - A practical Approach

Lorenz Feyen

Sep 17, 2021

CONTENTS

this is a foreword

pdf version can be found here here.

Part I

1. Introduction

CHAPTER

ONE

INTRODUCTION

this is an introduction

Part II

2. Data Preparation

CHAPTER

TWO

DATA PREPARATION

this is an introduction

CHAPTER

THREE

INDEXING AND SLICING

```
!pip install yfinance
import yfinance as yf
import pandas as pd
```

```
Collecting yfinance
  Using cached yfinance-0.1.63-py2.py3-none-any.whl
Requirement already satisfied: numpy>=1.15 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from yfinance) (1.21.2)
```

```
Collecting lxml>=4.5.1
```

```
  Using cached lxml-4.6.3-cp38-cp38-manylinux2014_x86_64.whl (6.8 MB)
Requirement already satisfied: requests>=2.20 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from yfinance) (2.26.0)
```

```
Collecting pandas>=0.24
```

```
  Using cached pandas-1.3.3-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl␣
 ↪(11.5 MB)
```

```
Collecting multitasking>=0.0.7
  Using cached multitasking-0.0.9-py3-none-any.whl
Requirement already satisfied: python-dateutil>=2.7.3 in /home/lorenzf/git/data-
 ↪science-practical-approach/venv/lib/python3.8/site-packages (from pandas>=0.24->
 ↪yfinance) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from pandas>=0.24->yfinance)␣
 ↪(2021.1)
Requirement already satisfied: six>=1.5 in /home/lorenzf/git/data-science-practical-
 ↪approach/venv/lib/python3.8/site-packages (from python-dateutil>=2.7.3->pandas>=0.
 ↪24->yfinance) (1.16.0)
```

```
Requirement already satisfied: charset-normalizer~=2.0.0 in /home/lorenzf/git/data-
 ↪science-practical-approach/venv/lib/python3.8/site-packages (from requests>=2.20->
 ↪yfinance) (2.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from requests>=2.20->yfinance)␣
 ↪(2021.5.30)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /home/lorenzf/git/data-
 ↪science-practical-approach/venv/lib/python3.8/site-packages (from requests>=2.20->
 ↪yfinance) (1.26.6)
```

(continues on next page)

(continued from previous page)

```
Requirement already satisfied: idna<4,>=2.5 in /home/lorenzf/git/data-science-
↪practical-approach/venv/lib/python3.8/site-packages (from requests>=2.20->yfinance)␣
↪(3.2)
```

```
Installing collected packages: pandas, multitasking, lxml, yfinance
```

```
Successfully installed lxml-4.6.3 multitasking-0.0.9 pandas-1.3.3 yfinance-0.1.63
```

```
WARNING: You are using pip version 21.1.2; however, version 21.2.4 is available.
You should consider upgrading via the '/home/lorenzf/git/data-science-practical-
↪approach/venv/bin/python -m pip install --upgrade pip' command.
```

```
pd.Timestamp.now()-pd.Timedelta(days=100)
```

```
Timestamp('2021-06-09 23:05:46.374369')
```

```
df = yf.download('TSLA', '2020-01-01', '2021-01-01')
```

```
[**********************100%***********************]  1 of 1 completed
```

```

```

```
df
```

```
                 Open        High         Low       Close   Adj Close  \
Date
2019-12-31   81.000000   84.258003   80.416000   83.666000   83.666000
2020-01-02   84.900002   86.139999   84.342003   86.052002   86.052002
2020-01-03   88.099998   90.800003   87.384003   88.601997   88.601997
2020-01-06   88.094002   90.311996   88.000000   90.307999   90.307999
2020-01-07   92.279999   94.325996   90.671997   93.811996   93.811996
...                ...         ...         ...         ...         ...
2020-12-24  642.989990  666.090027  641.000000  661.770020  661.770020
2020-12-28  674.510010  681.400024  660.799988  663.690002  663.690002
2020-12-29  661.000000  669.900024  655.000000  665.989990  665.989990
2020-12-30  672.000000  696.599976  668.359985  694.780029  694.780029
2020-12-31  699.989990  718.719971  691.119995  705.669983  705.669983

              Volume
Date
2019-12-31  51428500
2020-01-02  47660500
2020-01-03  88892500
2020-01-06  50665000
2020-01-07  89410500
...              ...
2020-12-24  22865600
2020-12-28  32278600
2020-12-29  22910800
2020-12-30  42846000
2020-12-31  49649900

[254 rows x 6 columns]
```

```
df.set_index('Open')
```

```
                 High         Low       Close   Adj Close    Volume
Open
81.000000    84.258003   80.416000   83.666000   83.666000  51428500
84.900002    86.139999   84.342003   86.052002   86.052002  47660500
88.099998    90.800003   87.384003   88.601997   88.601997  88892500
88.094002    90.311996   88.000000   90.307999   90.307999  50665000
92.279999    94.325996   90.671997   93.811996   93.811996  89410500
...                ...         ...         ...         ...       ...
642.989990  666.090027  641.000000  661.770020  661.770020  22865600
674.510010  681.400024  660.799988  663.690002  663.690002  32278600
661.000000  669.900024  655.000000  665.989990  665.989990  22910800
672.000000  696.599976  668.359985  694.780029  694.780029  42846000
699.989990  718.719971  691.119995  705.669983  705.669983  49649900

[254 rows x 5 columns]
```

```
df.loc['2020-06-01':'2020-06-30']
```

```
                 Open        High         Low       Close   Adj Close  \
Date
2020-06-01  171.600006  179.800003  170.820007  179.619995  179.619995
2020-06-02  178.940002  181.731995  174.199997  176.311996  176.311996
2020-06-03  177.623993  179.587997  176.020004  176.591995  176.591995
2020-06-04  177.975998  179.149994  171.688004  172.876007  172.876007
2020-06-05  175.567993  177.304001  173.240005  177.132004  177.132004
2020-06-08  183.800003  190.000000  181.832001  189.983994  189.983994
2020-06-09  188.001999  190.888000  184.785995  188.134003  188.134003
2020-06-10  198.376007  205.496002  196.500000  205.009995  205.009995
2020-06-11  198.039993  203.792007  194.399994  194.567993  194.567993
2020-06-12  196.000000  197.595993  182.520004  187.056000  187.056000
2020-06-15  183.557999  199.768005  181.699997  198.179993  198.179993
2020-06-16  202.369995  202.576004  192.477997  196.425995  196.425995
2020-06-17  197.542007  201.000000  196.514008  198.358002  198.358002
2020-06-18  200.600006  203.839996  198.893997  200.792007  200.792007
2020-06-19  202.556000  203.194000  198.268005  200.179993  200.179993
2020-06-22  199.990005  201.776001  198.003998  198.863998  198.863998
2020-06-23  199.776001  202.399994  198.802002  200.356003  200.356003
2020-06-24  198.822006  200.175995  190.628006  192.169998  192.169998
2020-06-25  190.854004  197.195999  187.429993  197.195999  197.195999
2020-06-26  198.955994  199.000000  190.973999  191.947998  191.947998
2020-06-29  193.802002  202.000000  189.703995  201.869995  201.869995
2020-06-30  201.300003  217.537994  200.746002  215.962006  215.962006

              Volume
Date
2020-06-01  74697500
2020-06-02  67828000
2020-06-03  39747500
2020-06-04  44438500
2020-06-05  39059500
2020-06-08  70873500
2020-06-09  56941000
2020-06-10  92817000
2020-06-11  79582500
```

(continues on next page)

(continued from previous page)

```
2020-06-12  83817000
2020-06-15  78486000
2020-06-16  70255500
2020-06-17  49454000
2020-06-18  48759500
2020-06-19  43398500
2020-06-22  31812000
2020-06-23  31826500
2020-06-24  54798000
2020-06-25  46272500
2020-06-26  44274500
2020-06-29  45132000
2020-06-30  84592500
```

```
df.loc['2020-05-01':'2020-05-31'].Volume.sum()
```

```
1363518000
```

```
!pip install seaborn
import seaborn as sns
```

```
Collecting seaborn
  Using cached seaborn-0.11.2-py3-none-any.whl (292 kB)
Requirement already satisfied: pandas>=0.23 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from seaborn) (1.3.3)
Requirement already satisfied: matplotlib>=2.2 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from seaborn) (3.4.3)
Requirement already satisfied: numpy>=1.15 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from seaborn) (1.21.2)
```

```
Collecting scipy>=1.0
```

```
  Using cached scipy-1.7.1-cp38-cp38-manylinux_2_5_x86_64.manylinux1_x86_64.whl (28.4␣
 ↪MB)
```

```
Requirement already satisfied: pyparsing>=2.2.1 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)␣
 ↪(2.4.7)
Requirement already satisfied: pillow>=6.2.0 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)␣
 ↪(8.3.1)
Requirement already satisfied: python-dateutil>=2.7 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)␣
 ↪(2.8.2)
Requirement already satisfied: cycler>=0.10 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)␣
 ↪(0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from matplotlib>=2.2->seaborn)␣
 ↪(1.3.1)
Requirement already satisfied: six in /home/lorenzf/git/data-science-practical-
 ↪approach/venv/lib/python3.8/site-packages (from cycler>=0.10->matplotlib>=2.2->
 ↪seaborn) (1.16.0)
Requirement already satisfied: pytz>=2017.3 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from pandas>=0.23->seaborn)␣
 ↪(2021.1)
```

(continues on next page)

(continued from previous page)

```
Installing collected packages: scipy, seaborn
```

```
Successfully installed scipy-1.7.1 seaborn-0.11.2
WARNING: You are using pip version 21.1.2; however, version 21.2.4 is available.
You should consider upgrading via the '/home/lorenzf/git/data-science-practical-
 ↪approach/venv/bin/python -m pip install --upgrade pip' command.
```

```
tip_df = sns.load_dataset('tips')
tip_df.head()
```

```
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
2       21.01  3.50    Male     No  Sun  Dinner     3
3       23.68  3.31    Male     No  Sun  Dinner     2
4       24.59  3.61  Female     No  Sun  Dinner     4
```

```
tip_index_df = tip_df.set_index('day')
```

```
tip_index_df.loc['Sun']
```

```
     total_bill   tip     sex smoker    time  size
day
Sun       16.99  1.01  Female     No  Dinner     2
Sun       10.34  1.66    Male     No  Dinner     3
Sun       21.01  3.50    Male     No  Dinner     3
Sun       23.68  3.31    Male     No  Dinner     2
Sun       24.59  3.61  Female     No  Dinner     4
..          ...   ...     ...    ...     ...   ...
Sun       20.90  3.50  Female    Yes  Dinner     3
Sun       30.46  2.00    Male    Yes  Dinner     5
Sun       18.15  3.50  Female    Yes  Dinner     3
Sun       23.10  4.00    Male    Yes  Dinner     3
Sun       15.69  1.50    Male    Yes  Dinner     2

[76 rows x 6 columns]
```

```
tip_index_df = tip_df.set_index(['day','time'])
```

```
tip_index_df.loc[('Thur','Lunch')].tip.mean()
```

```
/tmp/ipykernel_13322/2537502835.py:1: PerformanceWarning: indexing past lexsort depth␣
 ↪may impact performance.
  tip_index_df.loc[('Thur','Lunch')].tip.mean()
```

```
2.767704918032786
```

CHAPTER

FOUR

MISSING DATA

this is a notebook about missing data

```
variable = 'test'
```

```
import pandas as pd
```

```
df = pd.read_csv('https://openmv.net/file/kamyr-digester.csv')
df.head()
```

```
  Observation  Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4   \
0    31-00:00    23.10    16.520     121.717  1177.607      169.805
1    31-01:00    27.60    16.810      79.022  1328.360      341.327
2    31-02:00    23.19    16.709      79.562  1329.407      239.161
3    31-03:00    23.60    16.478      81.011  1334.877      213.527
4    31-04:00    22.90    15.618      93.244  1334.168      243.131

   T-upperExt-2   T-lowerExt-2    UCZAA  WhiteFlow-4   ...  SteamFlow-4   \
0        358.282        329.545  1.443       599.253  ...        67.122
1        351.050        329.067  1.549       537.201  ...        60.012
2        350.022        329.260  1.600       549.611  ...        61.304
3        350.938        331.142  1.604       623.362  ...        68.496
4        351.640        332.709    NaN       638.672  ...        70.022

   Lower-HeatT-3  Upper-HeatT-3   ChipMass-4   WeakLiquorF   BlackFlow-2   \
0        329.432        303.099      175.964      1127.197      1319.039
1        330.823        304.879      163.202       665.975      1297.317
2        329.140        303.383      164.013       677.534      1327.072
3        328.875        302.254      181.487       767.853      1324.461
4        328.352        300.954      183.929       888.448      1343.424

   WeakWashF   SteamHeatF-3   T-Top-Chips-4   SulphidityL-4
0     257.325         54.612         252.077             NaN
1     241.182         46.603         251.406           29.11
2     237.272         51.795         251.335             NaN
3     239.478         54.846         250.312           29.02
4     215.372         54.186         249.916           29.01

[5 rows x 23 columns]
```

```
df.isna().sum()
```

```
Observation           0
Y-Kappa               0
ChipRate              4
BF-CMratio           14
BlowFlow             13
ChipLevel4            1
T-upperExt-2          1
T-lowerExt-2          1
UCZAA                24
WhiteFlow-4           1
AAWhiteSt-4         141
AA-Wood-4             1
ChipMoisture-4        1
SteamFlow-4           1
Lower-HeatT-3         1
Upper-HeatT-3         1
ChipMass-4            1
WeakLiquorF           1
BlackFlow-2           1
WeakWashF             1
SteamHeatF-3          1
T-Top-Chips-4         1
SulphidityL-4       141
dtype: int64
```

```
df.ffill()['SulphidityL-4 ']
```

```
0        NaN
1      29.11
2      29.11
3      29.02
4      29.01
       ...
296    30.43
297    30.29
298    30.47
299    30.47
300    30.46
Name: SulphidityL-4 , Length: 301, dtype: float64
```

```
df = pd.read_csv('https://openmv.net/file/travel-times.csv')
df
```

```
          Date StartTime  DayOfWeek GoingTo  Distance  MaxSpeed  AvgSpeed  \
0     1/6/2012     16:37     Friday    Home     51.29     127.4      78.3
1     1/6/2012     08:20     Friday     GSK     51.63     130.3      81.8
2     1/4/2012     16:17  Wednesday    Home     51.27     127.4      82.0
3     1/4/2012     07:53  Wednesday     GSK     49.17     132.3      74.2
4     1/3/2012     18:57    Tuesday    Home     51.15     136.2      83.4
..         ...       ...        ...     ...       ...       ...       ...
200  7/18/2011     08:09     Monday     GSK     54.52     125.6      49.9
201  7/14/2011     08:03   Thursday     GSK     50.90     123.7      76.2
202  7/13/2011     17:08  Wednesday    Home     51.96     132.6      57.5
203  7/12/2011     17:51    Tuesday    Home     53.28     125.8      61.6
204  7/11/2011     16:56     Monday    Home     51.73     125.0      62.8
```

(continues on next page)

(continued from previous page)

```
     AvgMovingSpeed FuelEconomy  TotalTime  MovingTime Take407All Comments
0              84.8         NaN       39.3        36.3         No      NaN
1              88.9         NaN       37.9        34.9         No      NaN
2              85.8         NaN       37.5        35.9         No      NaN
3              82.9         NaN       39.8        35.6         No      NaN
4              88.1         NaN       36.8        34.8         No      NaN
..              ...         ...        ...         ...        ...      ...
200            82.4        7.89       65.5        39.7         No      NaN
201            95.1        7.89       40.1        32.1        Yes      NaN
202            76.7         NaN       54.2        40.6        Yes      NaN
203            87.6         NaN       51.9        36.5        Yes      NaN
204            92.5         NaN       49.5        33.6        Yes      NaN

[205 rows x 13 columns]
```

```
df.isna().sum()
```

```
Date                0
StartTime           0
DayOfWeek           0
GoingTo             0
Distance            0
MaxSpeed            0
AvgSpeed            0
AvgMovingSpeed      0
FuelEconomy        17
TotalTime           0
MovingTime          0
Take407All          0
Comments          181
dtype: int64
```

```
df[~df.Comments.isna()]
```

```
           Date StartTime  DayOfWeek GoingTo  Distance  MaxSpeed  AvgSpeed  \
15   12/19/2011     07:34     Monday     GSK     52.00     137.8      76.5
39   11/29/2011     07:23    Tuesday     GSK     51.74     112.2      55.3
49   11/21/2011     07:24     Monday     GSK     52.25     127.3      38.1
50   11/17/2011     16:16   Thursday    Home     51.16     127.6      72.4
52   11/16/2011     16:13  Wednesday    Home     51.12     125.1      65.0
54   11/15/2011     17:36    Tuesday    Home     51.06     122.8      61.4
60    11/9/2011     16:15  Wednesday    Home     51.28     121.4      65.9
78   10/25/2011     17:24    Tuesday    Home     52.87     123.5      65.1
91   10/12/2011     17:47  Wednesday    Home     51.40     114.4      59.7
92   10/12/2011     08:28  Wednesday     GSK     50.58     128.4      59.5
110   9/27/2011     07:36    Tuesday     GSK     50.65     128.1      86.3
132    9/7/2011     07:57  Wednesday     GSK     49.08     125.1      56.5
133    9/6/2011     16:27    Tuesday    Home     52.88     131.6      95.4
150   8/24/2011     07:59  Wednesday     GSK     49.07     127.1      58.5
156   8/19/2011     07:05     Friday     GSK     49.18     123.0      72.0
158   8/18/2011     08:11   Thursday     GSK     52.26     137.7      51.2
165   8/12/2011     17:25     Friday    Home     55.57     127.7      69.6
166   8/12/2011     08:05     Friday     GSK     49.02     128.4      76.7
172    8/9/2011     08:15    Tuesday     GSK     49.08     134.8      60.5
174    8/8/2011     08:07     Monday     GSK     49.25     126.3      68.5
```

(continues on next page)

(continued from previous page)

```
182    8/2/2011     07:38    Tuesday     GSK     53.48     124.9      68.8
184   7/29/2011     08:22     Friday     GSK     49.07     121.1      73.2
187   7/27/2011     17:24  Wednesday    Home     50.98     124.9      68.3
189   7/26/2011     17:15    Tuesday    Home     51.28     122.1      43.7

     AvgMovingSpeed FuelEconomy  TotalTime  MovingTime Take407All  \
15             87.8        8.89       40.8        35.5          No
39             61.0         NaN       56.2        50.9          No
49             50.3       10.05       82.3        62.4          No
50             77.4       10.05       42.4        39.6          No
52             73.1        9.53       47.2        41.9          No
54             70.9        9.53       49.9        43.2          No
60             71.8        9.35       46.7        42.1          No
78             72.4        8.97       48.7        43.8          No
91             65.8        8.75       51.7        46.9          No
92             67.3        8.75       51.0        45.1         Yes
110            88.6        8.31       35.2        34.3         Yes
132            66.5         8.5       52.1        44.3          No
133            98.3         8.5       33.3        32.3         Yes
150            71.5        8.54       50.3        41.1          No
156            81.4        8.37       41.0        36.3          No
158            64.1        8.37       61.2        48.9          No
165            77.1        8.54       47.9        43.2          No
166            82.9        8.54       38.4        35.5          No
172            67.2        8.54       48.7        43.8          No
174            78.2        8.54       43.1        37.8          No
182            78.8        8.48       46.7        40.7          No
184            77.7        8.45       40.2        37.9          No
187            71.9        8.45       44.8        42.6          No
189            51.5        8.45       70.5        59.8          No

                                          Comments
15                                Put snow tires on
39                                      Heavy rain
49                                Huge traffic backup
50    Pumped tires up: check fuel economy improved?
52                                Backed up at Bronte
54                                Backed up at Bronte
60                                             Rainy
78                                Rain, rain, rain
91                                Rain, rain, rain
92       Accident: backup from Hamilton to 407 ramp
110                                         Raining
132                          Back to school traffic?
133              Took 407 all the way (to McMaster)
150                           Heavy volume on Derry
156                        Start early to run a batch
158  Accident at 403/highway 6; detour along Dundas
165                                    Detour taken
166                                 Must be Friday
172                           Medium amount of rain
174                                       New tires
182                           Turn around on Derry
184                                   Empty roads
187                           Police slowdown on 403
189                         Accident blocked 407 exit
```

```
df.loc[df.Comments.isna(),'Comments'] = ''
```

```
df.Comments
```

```
0
1
2
3
4
      ..
200
201
202
203
204
Name: Comments, Length: 205, dtype: object
```

```
df[~df.FuelEconomy.isna()]
```

```
           Date StartTime  DayOfWeek GoingTo  Distance  MaxSpeed  AvgSpeed  \
6      1/2/2012     17:31     Monday    Home     51.37     123.2      82.9
7      1/2/2012     07:34     Monday     GSK     49.01     128.3      77.5
8    12/23/2011     08:01     Friday     GSK     52.91     130.3      80.9
9    12/22/2011     17:19   Thursday    Home     51.17     122.3      70.6
10   12/22/2011     08:16   Thursday     GSK     49.15     129.4      74.0
..          ...       ...        ...     ...       ...       ...       ...
197   7/20/2011     08:24  Wednesday     GSK     48.50     125.8      75.7
198   7/19/2011     17:17    Tuesday    Home     51.16     126.7      92.2
199   7/19/2011     08:11    Tuesday     GSK     50.96     124.3      82.3
200   7/18/2011     08:09     Monday     GSK     54.52     125.6      49.9
201   7/14/2011     08:03   Thursday     GSK     50.90     123.7      76.2

     AvgMovingSpeed FuelEconomy  TotalTime  MovingTime Take407All Comments
6              87.3           -       37.2        35.3         No
7              85.9           -       37.9        34.3         No
8              88.3        8.89       39.3        36.0         No
9              78.1        8.89       43.5        39.3         No
10             81.4        8.89       39.8        36.2         No
..              ...         ...        ...         ...        ...      ...
197            87.3        7.89       38.5        33.3        Yes
198           102.6        7.89       33.3        29.9        Yes
199            96.4        7.89       37.2        31.7        Yes
200            82.4        7.89       65.5        39.7         No
201            95.1        7.89       40.1        32.1        Yes

[188 rows x 13 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            205 non-null    object
```

(continues on next page)

(continued from previous page)

```
 1   StartTime       205 non-null    object
 2   DayOfWeek       205 non-null    object
 3   GoingTo         205 non-null    object
 4   Distance        205 non-null    float64
 5   MaxSpeed        205 non-null    float64
 6   AvgSpeed        205 non-null    float64
 7   AvgMovingSpeed  205 non-null    float64
 8   FuelEconomy     188 non-null    object
 9   TotalTime       205 non-null    float64
 10  MovingTime      205 non-null    float64
 11  Take407All      205 non-null    object
 12  Comments        205 non-null    object
dtypes: float64(6), object(7)
memory usage: 20.9+ KB
```

```
df.FuelEconomy = pd.to_numeric(df.FuelEconomy, errors='coerce')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            205 non-null    object
 1   StartTime       205 non-null    object
 2   DayOfWeek       205 non-null    object
 3   GoingTo         205 non-null    object
 4   Distance        205 non-null    float64
 5   MaxSpeed        205 non-null    float64
 6   AvgSpeed        205 non-null    float64
 7   AvgMovingSpeed  205 non-null    float64
 8   FuelEconomy     186 non-null    float64
 9   TotalTime       205 non-null    float64
 10  MovingTime      205 non-null    float64
 11  Take407All      205 non-null    object
 12  Comments        205 non-null    object
dtypes: float64(7), object(6)
memory usage: 20.9+ KB
```

```
df[~df.FuelEconomy.isna()]
```

```
           Date StartTime  DayOfWeek GoingTo  Distance  MaxSpeed  AvgSpeed  \
8    12/23/2011     08:01     Friday     GSK     52.91     130.3      80.9
9    12/22/2011     17:19   Thursday    Home     51.17     122.3      70.6
10   12/22/2011     08:16   Thursday     GSK     49.15     129.4      74.0
11   12/21/2011     07:45  Wednesday     GSK     51.77     124.8      71.7
12   12/20/2011     16:05    Tuesday    Home     51.45     130.1      75.2
..          ...       ...        ...     ...       ...       ...       ...
197   7/20/2011     08:24  Wednesday     GSK     48.50     125.8      75.7
198   7/19/2011     17:17    Tuesday    Home     51.16     126.7      92.2
199   7/19/2011     08:11    Tuesday     GSK     50.96     124.3      82.3
200   7/18/2011     08:09     Monday     GSK     54.52     125.6      49.9
201   7/14/2011     08:03   Thursday     GSK     50.90     123.7      76.2
```

(continues on next page)

(continued from previous page)

```
     AvgMovingSpeed  FuelEconomy  TotalTime  MovingTime Take407All Comments
8              88.3         8.89       39.3        36.0         No
9              78.1         8.89       43.5        39.3         No
10             81.4         8.89       39.8        36.2         No
11             78.9         8.89       43.3        39.4         No
12             82.7         8.89       41.1        37.3         No
..              ...          ...        ...         ...        ...      ...
197            87.3         7.89       38.5        33.3        Yes
198           102.6         7.89       33.3        29.9        Yes
199            96.4         7.89       37.2        31.7        Yes
200            82.4         7.89       65.5        39.7         No
201            95.1         7.89       40.1        32.1        Yes

[186 rows x 13 columns]
```

```
df = pd.read_csv('http://openmv.net/file/raw-material-properties.csv')
df.head()
```

```
   Sample  size1  size2  size3  density1  density2  density3
0  X12558  0.696   2.69   6.38      41.8     17.18      3.90
1  X14728  0.636   2.30   5.14      38.1     12.73      3.89
2  X15468  0.841   2.85   5.20      37.6     13.58      3.98
3  X21364  0.609   2.13   4.62      34.2     11.12      4.02
4  X23671  0.684   2.16   4.87      36.4     12.24      3.92
```

```
!pip install sklearn
from sklearn.impute import KNNImputer
```

```
Collecting sklearn
  Using cached sklearn-0.0-py2.py3-none-any.whl
```

```
Collecting scikit-learn
```

```
  Using cached scikit_learn-0.24.2-cp38-cp38-manylinux2010_x86_64.whl (24.9 MB)
```

```
Collecting threadpoolctl>=2.0.0
  Using cached threadpoolctl-2.2.0-py3-none-any.whl (12 kB)
Requirement already satisfied: scipy>=0.19.1 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from scikit-learn->sklearn) (1.
 ↪7.1)
```

```
Collecting joblib>=0.11
```

```
  Using cached joblib-1.0.1-py3-none-any.whl (303 kB)
Requirement already satisfied: numpy>=1.13.3 in /home/lorenzf/git/data-science-
 ↪practical-approach/venv/lib/python3.8/site-packages (from scikit-learn->sklearn) (1.
 ↪21.2)
```

```
Installing collected packages: threadpoolctl, joblib, scikit-learn, sklearn
```

```
Successfully installed joblib-1.0.1 scikit-learn-0.24.2 sklearn-0.0 threadpoolctl-2.2.
 ↪0
WARNING: You are using pip version 21.1.2; however, version 21.2.4 is available.
```

(continues on next page)

(continued from previous page)

```
You should consider upgrading via the '/home/lorenzf/git/data-science-practical-
↪approach/venv/bin/python -m pip install --upgrade pip' command.
```

```
imputer = KNNImputer(n_neighbors=5, weights="distance")
```

```
pd.DataFrame(
    imputer.fit_transform(df.drop(columns=['Sample'])),
    columns=df.columns.drop('Sample')
)
```

	size1	size2	size3	density1	density2	density3
0	0.696000	2.690000	6.380000	41.800000	17.180000	3.900000
1	0.636000	2.300000	5.140000	38.100000	12.730000	3.890000
2	0.841000	2.850000	5.200000	37.600000	13.580000	3.980000
3	0.609000	2.130000	4.620000	34.200000	11.120000	4.020000
4	0.684000	2.160000	4.870000	36.400000	12.240000	3.920000
5	0.762000	2.810000	6.360000	38.100000	13.280000	3.890000
6	0.552000	2.340000	5.030000	41.300000	16.710000	3.860000
7	0.501000	2.170000	5.090000	38.495282	14.029399	3.931180
8	0.619000	2.110000	5.130000	37.405275	13.157346	3.943667
9	0.610000	2.100000	4.180000	35.000000	12.150000	3.860000
10	0.532000	2.090000	4.930000	37.811132	13.646072	3.908364
11	0.738000	2.290000	5.470000	37.088833	13.255412	3.941654
12	0.779000	2.620000	5.590000	36.540567	12.889902	3.970973
13	0.537000	2.230000	5.410000	35.200000	11.340000	3.990000
14	0.702000	2.050000	5.100000	34.200000	10.540000	4.020000
15	0.768000	2.510000	5.090000	34.900000	12.550000	3.900000
16	0.714000	2.560000	6.030000	35.600000	12.200000	4.020000
17	0.621000	2.420000	5.100000	38.700000	14.270000	3.980000
18	0.726000	2.110000	4.690000	37.100000	13.140000	3.980000
19	0.698000	2.360000	5.400000	36.600000	12.160000	4.010000
20	0.733097	2.653959	5.881504	38.100000	13.340000	3.890000
21	0.759000	2.470000	4.830000	38.700000	14.830000	3.890000
22	0.535000	2.130000	5.230000	37.391815	13.089536	3.944335
23	0.716000	2.290000	5.450000	37.300000	13.700000	3.920000
24	0.635000	2.080000	4.940000	37.254724	13.206262	3.933904
25	0.598000	2.120000	4.690000	37.900000	13.450000	3.780000
26	0.700000	2.470000	5.220000	38.800000	14.720000	3.920000
27	0.957000	2.960000	7.370000	36.200000	13.380000	4.200000
28	0.759000	2.660000	5.360000	35.200000	12.190000	3.980000
29	0.661000	2.100000	4.270000	36.172345	12.755632	3.887375
30	0.646000	2.380000	4.510000	40.100000	15.680000	3.860000
31	0.662000	2.340000	4.710000	35.000000	12.370000	3.900000
32	0.749000	2.430000	5.160000	37.300000	13.040000	3.920000
33	0.598000	2.210000	4.900000	37.865882	13.826029	3.887021
34	0.619000	2.590000	5.810000	35.932339	12.318210	3.989911
35	0.693000	2.050000	5.020000	39.600000	15.550000	3.940000

CHAPTER

FIVE

CONCATENATION AND DEDUPLICATION

https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2020-01.csv

```
import pandas as pd
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
/tmp/ipykernel_13304/4080736814.py in <module>
----> 1 import pandas as pd

ModuleNotFoundError: No module named 'pandas'
```

```
df = pd.read_csv('https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2020-01.
↪csv')
```

```
/home/lorenzf/.local/lib/python3.8/site-packages/IPython/core/interactiveshell.
↪py:3441: DtypeWarning: Columns (6) have mixed types.Specify dtype option on import␣
↪or set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

```
df
```

```
         VendorID tpep_pickup_datetime tpep_dropoff_datetime  passenger_count  \
0             1.0  2020-01-01 00:28:15   2020-01-01 00:33:03              1.0
1             1.0  2020-01-01 00:35:39   2020-01-01 00:43:04              1.0
2             1.0  2020-01-01 00:47:41   2020-01-01 00:53:52              1.0
3             1.0  2020-01-01 00:55:23   2020-01-01 01:00:14              1.0
4             2.0  2020-01-01 00:01:58   2020-01-01 00:04:16              1.0
...           ...                  ...                   ...              ...
6405003       NaN  2020-01-31 22:51:00   2020-01-31 23:22:00              NaN
6405004       NaN  2020-01-31 22:10:00   2020-01-31 23:26:00              NaN
6405005       NaN  2020-01-31 22:50:07   2020-01-31 23:17:57              NaN
6405006       NaN  2020-01-31 22:25:53   2020-01-31 22:48:32              NaN
6405007       NaN  2020-01-31 22:44:00   2020-01-31 23:06:00              NaN

         trip_distance  RatecodeID store_and_fwd_flag  PULocationID  \
0                 1.20         1.0                  N           238
1                 1.20         1.0                  N           239
2                 0.60         1.0                  N           238
3                 0.80         1.0                  N           238
4                 0.00         1.0                  N           193
...                ...         ...                ...           ...
6405003           3.24         NaN                NaN           237
```

(continues on next page)

(continued from previous page)

```
6405004          22.13         NaN                NaN           259
6405005          10.51         NaN                NaN           137
6405006           5.49         NaN                NaN            50
6405007          11.60         NaN                NaN           179

         DOLocationID  payment_type  fare_amount  extra  mta_tax  tip_amount  \
0                 239           1.0         6.00   3.00      0.5        1.47
1                 238           1.0         7.00   3.00      0.5        1.50
2                 238           1.0         6.00   3.00      0.5        1.00
3                 151           1.0         5.50   0.50      0.5        1.36
4                 193           2.0         3.50   0.50      0.5        0.00
...               ...           ...          ...    ...      ...         ...
6405003           234           NaN        17.59   2.75      0.5        0.00
6405004            45           NaN        46.67   2.75      0.5        0.00
6405005           169           NaN        48.85   2.75      0.0        0.00
6405006            42           NaN        27.17   2.75      0.0        0.00
6405007           205           NaN        54.56   2.75      0.5        0.00

         tolls_amount  improvement_surcharge  total_amount  \
0                0.00                    0.3         11.27
1                0.00                    0.3         12.30
2                0.00                    0.3         10.80
3                0.00                    0.3          8.16
4                0.00                    0.3          4.80
...               ...                    ...           ...
6405003          0.00                    0.3         21.14
6405004         12.24                    0.3         62.46
6405005          0.00                    0.3         51.90
6405006          0.00                    0.3         30.22
6405007          0.00                    0.3         58.11

         congestion_surcharge
0                         2.5
1                         2.5
2                         2.5
3                         0.0
4                         0.0
...                       ...
6405003                   0.0
6405004                   0.0
6405005                   0.0
6405006                   0.0
6405007                   0.0

[6405008 rows x 18 columns]
```

Part III

3. Data Preprocessing

CHAPTER

SIX

DATA PREPROCESSING

this is an introduction

Part IV

4. Data Exploration

CHAPTER

SEVEN

DATA EXPLORATION

this is an introduction

Part V

5. Data Visualisation

CHAPTER

EIGHT

DATA VISUALISATION

this is an introduction

Part VI

6. Machine Learning

CHAPTER

NINE

MACHINE LEARNING

this is an introduction