Data Science - A practical Approach

Lorenz Feyen

Oct 08, 2021

CONTENTS

pdf version can be found here.

Part I

1. Introduction

CHAPTER

ONE

INTRODUCTION

this is an introduction

1.1 Structured vs Unstructured

When performing data preparation an important aspect is to consider with the type of data we are working with. In general there are 2 types of data, but you could consider a third.

1.1.1 Structured data

Structured data is data that adheres to a pre-defined data model and is therefore straightforward to analyze. This data model is the description of our data, each record has to be conform to the model. A table in a spreadsheet is a good example of the concept of structured data however often no data types are enforced, meaning a column can contain e.g. both numbers and text. Later we will see that a mixture of data types is often problematic therefor the need of a data model.

1.1.2 Unstructured data

In contrast to structured data, there is no apparent data model but this does not mean the data is unusable or cluttered. Usually it means either no data model has yet been applied or we are dealing with data that is difficult to confine in a model. A great example of this would be images, or more general (binary) files. These obviously are hard to sort yet often data structures also contain metadata from these files, with data describing things as when the file was uploaded, what is shown in the file, … In turn the metadata can be structured and a data model can be related to the unstructured data.

1.1.3 Semi-structured data

As an intermediate option, we have what is called semi-structured data. The reasoning behind this is that the concept of tables is not always applicable, in some occasions e.g. data lakes there is no complex structure present compared to a database. In a data lake files are stored similar to the folder structure in your computer, with no fancy infrastructure behind it, thus reducing operation costs. This implies that a data model can not be enforced and the data is stored in generic files.

1.2 Data Structures

There are several structures in which data can be stored and accessed, here we cover the 3 most important.

1.2.1 Data Lake

As mentioned earlier a data lake would be the most cost efficient method as it relies on the least infrastructure and can be serverless. The concept behind a data lake is straight-forward, the data is stored in simple files with a specific notation e.g. parquet, csv, xml,… What is important when designing a data lake would be partitioning, this can be achieved by using subfolders and saving parts of the data in different files. To make this more tangible, take a look at this symbolic example I provided. Instead of putting all data in one csv file, subfolder divide the data in Country, City and then the year. We could even further partition yet the data is here in daily frequency so that would create many small partitions. The difficulty for a data lake lies in the method of interacting, when adding new data one has to adhere to a agreed upon data model that is not enforced, meaning you could create incorrect data which then need to be cleaned. On the other hand efficiency of you data lake depends on good partitioning, as the order of divisioning of your folders. We could have also divided first on year and then on country and city. As a data scientist seeing the data lake might not be as common, as this is rather an engineering task, however using the concepts of a data lake in experimental projects can make a big difference.

1.2.2 Database

Another interesting data structure is the database, widely used for exceptional speeds and ease of use, yet costly in storage. Numerous implementations of servers using the SQL language are developed over the years with each their own dialect and advantages. The important take home message here is that you can easily perform queries on the database that pre-handles the data to retrieve the information you need. these operations include filtering, grouping categories, joining tables, ordering and much more, as SQL is a complete language on its own. As a data scientist these databases are much more common, so SQL is a good asset to learn!

1.2.3 Data Warehouse

A next step towards data analysis is the data warehouse, where a database is composed of the most pragmatic method of storing your data a data warehouse consist of multiple views on your data. Based upon the data of a dataset the data warehouse transforms this data into a new format that displays the data in a new way. Let me illustrate with with a simple example, we have a database with a table that contains the rentals of books from multiple libraries. This table has a few columns: a timestamp, the library, the action (rent, return, …), the client_id and the book_id. If you would want to know if a book is available this database is perfect for your needs as you just have to find the last event for that book and if its a return the book is (or should be) there. Now image we would want to know how many books are being rented per month this database is insufficient, yet our data warehouse might contain such a view! It is up to the data engineer/scientist to create a computation that displays the amount of books rented per month. If they also would like to subdivided it per category of books, you would need to incorporate another table of the database where information of the books is stored. More on these operations of a data warehouse will be seen in the data preprocessing chapter. One last remark about data warehousing, it is important to optimize between memory and computation. Tables in our data warehouse compared to database can be computed in place reducing memory costs yet increasing computation costs. If a visualization tool often queries a table in your warehouse it is favorable to create it as a table in your database.

1.3 OLTP and OLAP

From the previous section you might have deduced that a database and Data Warehouse serve 2 different purposes. These are denoted as OnLine Transaction Processing and OnLine Analytical Processing, as the names suggest these are used for transactional and analytical processes.

1.3.1 OLTP

For this method the database structure is optimal, let us review the example where we have libraries renting out books. Renting out a book would send a message to our OLTP system creating a new record stating that specific book is at this moment rented out from our library. OLTP handles day-to-day operational data that can be both written and read from our database.

1.3.2 OLAP

In the case we would like to analyse data from the libraries we would use the OLAP method, creating multi-dimensional views from our transactional data. Our dimensions would be the date (aggregated per month), the library and the category of book, the chapter of data preprocessing will use these operations practically. I could write a whole chapter on OLAP operations however they are well described in this wikipedia page.

Part II

2. Data Preparation

CHAPTER

TWO

INTRODUCTION

When performing data science, we often do not elaborate about the preparation that went into the dataset. It is considered tedious and irrelevant to the story of the analysis, however it is often the most important part of data analysis. Data Preparation is the metaphorical foundation of your construction, if you fail to prepare data, you prepare to fail your analysis.

> Good data beats a fancy algorithm

If you would perform an analysis and insert unprepared data, you will mostly be disappointed with the result.

2.1 why Data Preparation?

Aside from metaphors let us make the reasoning behind this step more tangile, to explain the relevance of this step, we partitioned the answer into a few key points.

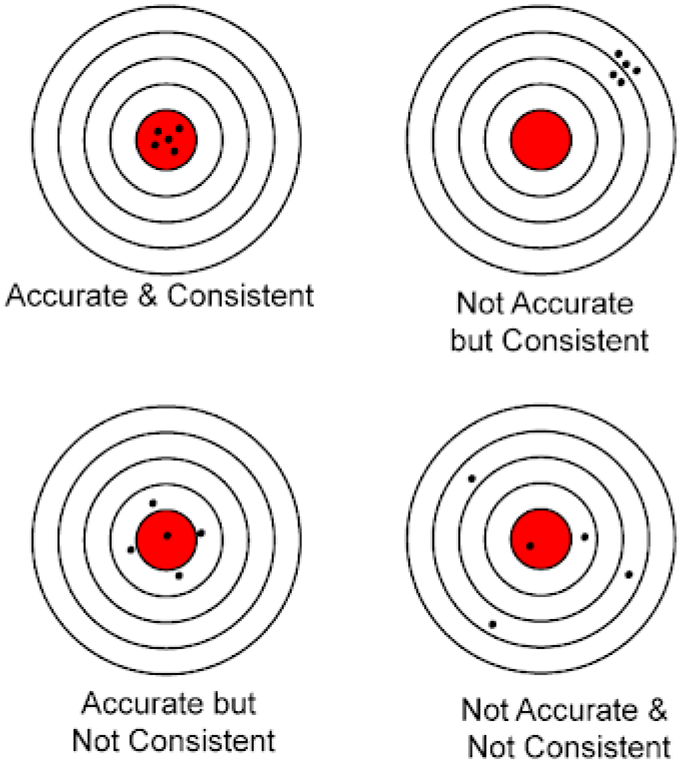
2.1.1 Accuracy

There is no excuse for incorrect data and accuracy is the most important attribute. Let us assume that we have a dataset where for some reason the result are not accurate. This would led us to an analysis where we conclude a result that contains a bias. An example would be a dataset of sold cars, where the listed price is that of the stock car without options. Options are not incorporated in the price and we are perhaps training an algorithm that predicts the stock price. If you as a data scientist fail to report/correct this, your predictions are making sense, but always underestimate!

2.1.2 Consistency

They usually say something such as 'consistency is key' and with data preparation that is likewise true. A dataset where we do not have consistent results will never converge towards a particular answer. Note however that it might not be a problem of consistency but rather you are missing crucial information. If we would have a dataset where local temperatures are logged, we would like to see a consistency each 24 hours. However we do see there are day to day fluctuations, so perhaps we need to keep track of cloud and rain data to make the dataset more complete. We could then see that the results are more consistent yet the possibility of outliers is still present. Equally possible would be that our temperature sensor is not sensitive enough or has large fluctuations in readings, it is the task of the data scientist to figure this out.

To get a visual about accuracy and consistency this picture might help:



2.1.3 Completeness

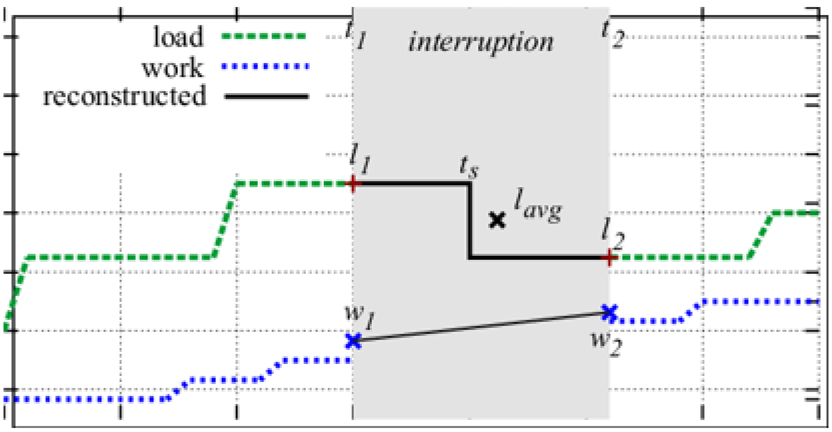
As hinted in the previous point, completeness is something you have to be aware of. Having 'complete' data is crucial for you narrative to give a correct answer, as you might otherwise lose detail. Note that you never will know if your data is complete as there might always be more data to mine. Yet you have to make a consideration between collecting more data and the effort required. This collecting can happen in multiple methods, as an example we use a survey where we asked several people 10 different questions, we could:

- gather new data, here our data grows 'longer' by asking the 10 question to more people. It might be that our sample of people were only students at a campus, so our data was not complete.
- gather new feature, by asking more questions to the same people (in case we could still find them). By doing this we get a better understanding of their opinion, again making our data more complete.
- fill missing values, by imputing the abstained questions with answers of similar records. When someone answered they did not want to answer we could figure out what they would have answered by looking at what persons answered that reply in a similar way.

2.1.4 Timeliness

For some datasets we are dealing with data that is time related. It can happen that data at specific timepoints is missing or delayed, resulting in a failure to use machine learning algorithms. A well-organised data pipeline utilises techniques of data preparation to circumvent these outages, usually this would be to retain the last successful datapoint. However in hindsight we could use more complex strategies to fill in these gaps or correct datetimes in our dataset,

In this example the data stream is interrupted and data preparation is there to handle these outages before we can perform analysis.



2.1.5 Believability

You could collect the most intricate dataset possible, but if the narrative that you are conducting contradicts itself, you will end up nowhere. During the process of data analytics it is important to apply a critical mind to what your dataset is telling you. Obviously this is not a reason to mask or mold the data so it agrees with your opinion. Rather you should be wary when conflicts happen and act accordingly, unfortunately it is impossible to write a generic tactic for this. As a data scientist your experience of the underlying subject should help create understanding of the topic, remember, gathering information from experts in the field is crucial here!

2.1.6 Interpretability

Another problem that might arise when you are diving deep into the data might be that you have created something no human could ever interpret. The Machine Learning algorithms outputs plausible and believable results, but it is impossible to understand the reasoning behind. For some this is perfectly acceptible, for some this is undesirable. It is your task as a data scientist to cater the wishes of the product operator and if they desire understanding as they would like to learn from the data driven process you need to unfold the process. Usually this comes down to which data transformations are used as some do produce an output that only makes mathematical sense.

2.1.7 In conclusion

There are multiple ways to deteriorate the quality of your data and raw formats of data often contain multiple. Before we can do anything with it these problems need to be resolved, if you fail to do so, the final output fails too.

2.2 Further reading

Towards Data Science

CHAPTER

THREE

MISSING DATA

In this notebook we will look at a few datasets where values from columns are missing. It is crucial for data science and machine learning to have a dataset where no values are missing as algorithms are usually not able to handle data with information missing.

For python, we will be using the pandas library to handle our dataset.

```
import pandas as pd
```

3.1 Kamyr digester

The first dataset we will be looking at is taken from a physical device equiped with numerous sensors, each timepoint (1 hour) these sensors are read out and the data is collected. Let's have a look at the general structure

```
kamyr_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-
↪practical-approach/main/src/c2_data_preparation/data/kamyr-digester.csv')
kamyr_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
```

(continues on next page)

(continued from previous page)

```
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]
```

Interesting, there seem to be 22 sensor values and 1 timestamp for each record. As mechanical devices are prone to noise and dropouts of sensors we would be foolish to assume no missing values are present.

```
kamyr_df.isna().sum().divide(len(kamyr_df)).round(4)*100
```

```
Observation          0.00
Y-Kappa              0.00
ChipRate             1.33
BF-CMratio           4.65
BlowFlow             4.32
ChipLevel4           0.33
T-upperExt-2         0.33
T-lowerExt-2         0.33
UCZAA                7.97
WhiteFlow-4          0.33
AAWhiteSt-4         46.84
AA-Wood-4            0.33
ChipMoisture-4       0.33
SteamFlow-4          0.33
Lower-HeatT-3        0.33
Upper-HeatT-3        0.33
ChipMass-4           0.33
WeakLiquorF          0.33
BlackFlow-2          0.33
WeakWashF            0.33
SteamHeatF-3         0.33
T-Top-Chips-4        0.33
SulphidityL-4       46.84
dtype: float64
```

As expected, the datapoint 'AAWhiteSt-4' even has 46% of data missing! It seems we only have 300 datapoints and presumably these missing values occur in different records our dataset will be decimated if we just drop all rows with missing values.

```
kamyr_df.shape
```

```
(301, 23)
```

```
kamyr_df.dropna().shape
```

```
(131, 23)
```

As we drop all rows with missing values, we are left with only 131 records. Whilst this might be good enough for some purposes, there are more viable options.

Perhaps we can first remove the column with the most missing values and then drop all remaining

```
kamyr_df.drop(columns=['AAWhiteSt-4 ','SulphidityL-4 ']).dropna().shape
```

```
(263, 21)
```

Significantly better, although we lost the information of 2 sensors we now have a complete dataset with 263 records. For purposes where those 2 sensors are irrelevant this is a viable option, keep in mind that this dataset is still 100% truthful, as we have not imputed any values.

Another option, where we retain all our records would be using the timely nature of our dataset, each record is a measurement with an interval of 1 hour. I have no knowledge of this dataset but one might make the assumption that the interval of 1 hour is taken as the state of the machine does not alter much in 1 hour. Therefore we could do what is called a forward fill, where we fill in the missing values with the same value of the sensor for the previous measurement.

This would solve nearly all nan values as there might be a problem where the first value is missing. This is shown below.

```
kamyr_df.fillna(method='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
```

Although our dataset is not fully the truth, we can see that little to no changes occur in the sensor and using a forward fill is arguably the most suitable option.

3.2 Travel times

Another dataset from the same source contains a collection of recorded travel times and specific information about the travel itself as e.g.: the day of the week, where they were going, …

```
travel_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-
↪practical-approach/main/src/c2_data_preparation/data/travel-times.csv')
travel_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

     AvgMovingSpeed FuelEconomy  TotalTime  MovingTime Take407All Comments
```

(continues on next page)

(continued from previous page)

```
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]
```

we have a total of 205 records and we can already see that the FuelEconomy column seems pretty bad, let's quantify that.

```
travel_df.isna().sum().divide(len(travel_df)).round(4)*100
```

```
Date                  0.00
StartTime             0.00
DayOfWeek             0.00
GoingTo               0.00
Distance              0.00
MaxSpeed              0.00
AvgSpeed              0.00
AvgMovingSpeed        0.00
FuelEconomy           8.29
TotalTime             0.00
MovingTime            0.00
Take407All            0.00
Comments             88.29
dtype: float64
```

In the end, it doesn't seem that bad, but there are comments and nearly none of them are filled in. Which in perspective is understandable. Let's see what the comments look like

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

(continues on next page)

(continued from previous page)

```
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
Name: Comments, dtype: object
```

As you would expect, these comments are text based. Now imagine we would like to run some Natural Language Processing (NLP) on these, it would be a pain to perform string operations on it when it is riddled with missing values.

Here a simple example where we select all records containing the word 'rain', with no avail.

```
travel_df[travel_df.Comments.str.lower().str.contains('rain')]
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
/tmp/ipykernel_6376/1298831137.py in <module>
----> 1 travel_df[travel_df.Comments.str.lower().str.contains('rain')]

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪frame.py in __getitem__(self, key)
   3446 
   3447         # Do we have a (boolean) 1d indexer?
-> 3448         if com.is_bool_indexer(key):
   3449             return self._getitem_bool_array(key)
   3450 

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪common.py in is_bool_indexer(key)
    137                 # Don't raise on e.g. ["A", "B", np.nan], see
    138                 #  test_loc_getitem_list_of_labels_categoricalindex_with_
↪na
--> 139                 raise ValueError(na_msg)
    140             return False
    141         return True

ValueError: Cannot mask with non-boolean array containing NA / NaN values
```

The last line of the python error traceback gives us the reason it failed, because there were NaN values present.

Luckily the string variable has more or less it's on 'null' value, being an empty string, this way these operations are still possible, most of the comments will just contain nothing.

```
travel_df.Comments = travel_df.Comments.fillna('')
```

```
travel_df[travel_df.Comments.str.lower().str.contains('rain')]
```

```
           Date StartTime  DayOfWeek GoingTo  Distance  MaxSpeed  AvgSpeed  \
39   11/29/2011     07:23    Tuesday     GSK     51.74     112.2      55.3
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
110   9/27/2011     07:36    Tuesday     GSK     50.65     128.1      86.3
172    8/9/2011     08:15    Tuesday     GSK     49.08     134.8      60.5
```

(continues on next page)

(continued from previous page)

```
     AvgMovingSpeed FuelEconomy  TotalTime  MovingTime Take407All  \
39             61.0         NaN       56.2        50.9         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
110            88.6        8.31       35.2        34.3        Yes
172            67.2        8.54       48.7        43.8         No

                  Comments
39              Heavy rain
60                   Rainy
78        Rain, rain, rain
91        Rain, rain, rain
110                Raining
172  Medium amount of rain
```

Fixed! now we can use the comments for analysis.

We still have to fix the FuelEconomy, let us take a look at the non NaN values

```
travel_df[~travel_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]
```

It seems that aside NaN values there are also other intruders, a quick check on the data type (Dtype) reveils it is not recognised as a number!

```
travel_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     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
```

The column is noted as an object or string type, meaning that these numbers are given as '9.24' instead of 9.24 and numerical operations are not possible. We can cast them to numeric but have to warn pandas to coerce errors, meaning errors will be converted to NaN values. Later we'll handle the NaN's.

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

Wonderful, now the column is numerical and we can see 2 more missing values have popped up! We could easily drop these 19 records and have a complete dataset.

```
travel_df.dropna()
```

```
           Date StartTime  DayOfWeek GoingTo  Distance  MaxSpeed  AvgSpeed  \
8    12/23/2011     08:01     Friday     GSK     52.91     130.3      80.9
```

(continues on next page)

(continued from previous page)

```
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

     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]
```

However im leaving them as an excercise for you to apply a technique we will see in the next part

3.3 Material properties

Another dataset from the same source contains the material properties from 30 samples, this time there is not timestamp as the samples are not related in time with each other.

```
material_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-
↪practical-approach/main/src/c2_data_preparation/data/raw-material-properties.csv')
material_df
```

```
    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
5   X24055  0.762   2.81   6.36      38.1     13.28      3.89
6   X24905  0.552   2.34   5.03      41.3     16.71      3.86
7   X25917  0.501   2.17   5.09       NaN       NaN       NaN
8   X27871  0.619   2.11   5.13       NaN       NaN       NaN
9   X28690  0.610   2.10   4.18      35.0     12.15      3.86
10  X31385  0.532   2.09   4.93       NaN       NaN       NaN
11  X31813  0.738   2.29   5.47       NaN       NaN       NaN
12  X32807  0.779   2.62   5.59       NaN       NaN       NaN
13  X33943  0.537   2.23   5.41      35.2     11.34      3.99
14  X35035  0.702   2.05   5.10      34.2     10.54      4.02
15  X39223  0.768   2.51   5.09      34.9     12.55      3.90
```

(continues on next page)

(continued from previous page)

```
16  X40503  0.714   2.56   6.03       35.6      12.20       4.02
17  X41400  0.621   2.42   5.10       38.7      14.27       3.98
18  X42988  0.726   2.11   4.69       37.1      13.14       3.98
19  X44749  0.698   2.36   5.40       36.6      12.16       4.01
20  X45295    NaN    NaN    NaN       38.1      13.34       3.89
21  X46965  0.759   2.47   4.83       38.7      14.83       3.89
22  X49666  0.535   2.13   5.23        NaN        NaN        NaN
23  X50678  0.716   2.29   5.45       37.3      13.70       3.92
24  X52894  0.635   2.08   4.94        NaN        NaN        NaN
25  X53925  0.598   2.12   4.69       37.9      13.45       3.78
26  X54254  0.700   2.47   5.22       38.8      14.72       3.92
27  X54272  0.957   2.96   7.37       36.2      13.38       4.20
28  X54394  0.759   2.66   5.36       35.2      12.19       3.98
29  X55408  0.661   2.10   4.27        NaN        NaN        NaN
30  X56952  0.646   2.38   4.51       40.1      15.68       3.86
31  X57095  0.662   2.34   4.71       35.0      12.37       3.90
32  X57128  0.749   2.43   5.16       37.3      13.04       3.92
33  X61870  0.598   2.21   4.90        NaN        NaN        NaN
34  X61888  0.619   2.59   5.81        NaN        NaN        NaN
35  X72736  0.693   2.05   5.02       39.6      15.55       3.94
```

let us quantify the amount of missing data

```
material_df.isna().sum().divide(len(material_df)).round(4)*100
```

```
Sample        0.00
size1         2.78
size2         2.78
size3         2.78
density1     27.78
density2     27.78
density3     27.78
dtype: float64
```

Unfortunately that is a lot of missing data, covered in all records, dropping here seems almost impossible if we want to keep a healthy amount of records.

Here it would be wise to go for a more elaborate method of imputation, I opted for the K-nearest neighbours method, which looks at the K most similar records in the dataset to make an educated guess on what the missing value could be, this because we can assume that records with similar data are also similar over all the properties (columns).

Im using the sklearn library for this, which has more imputation techniques such as MICE. More info can be found here

```
from sklearn.impute import KNNImputer
```

im creating an imputer object and specify that i want to use the 5 most similar records and weigh them by distance from the to imputed record, meaning closer neighbours are more important.

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

As the imputer only takes numerical values I had to do some pandas magic and drop the first column, which I then added again. The result is a fully filled dataset, you can recognise the new values as they are not rounded.

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

This concludes the part of missing values, perhaps you can try yourself and impute the missing values for the FuelEconomy using the SimpleImputer or even the IterativeImputer.

CHAPTER

FOUR

CONCATENATION AND DEDUPLICATION

In this notebook we are going to investigate the concepts of stitching data files (concatenation) and verifying the integrity of our data concercing duplicates

4.1 Concatenation

When dealing with large amounts of data, fractioning is often the only solution. Not only does this tidy up your data space, but it also benefits computation. Aside from that, appending new data to your data lake is independent of the historical data. However if you want to perform historical analysis this means you will need to perform additional operations.

In this notebook we have a setup of a very small data lake containing daily minimal temperatures. If you would look closely in the url you would see the following structure.

> data/temperature/australia/melbourne/1981.csv

This is a straight-forward but perfect example on how fragmentation works, in our data lake we have: temperatures data fractioned by country, city and year. As we are working with daily temperatures further fractioning would not be interesting, but you could fraction e.g. per month.

In the cells below, we read our both 1981 and 1982 data and concatenate them using python.

```
import pandas as pd
```

```
melbourne_1981_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-
↪science-practical-approach/main/src/c2_data_preparation/data/temperatures/australia/
↪melbourne/1981.csv')
```

```
melbourne_1982_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-
↪science-practical-approach/main/src/c2_data_preparation/data/temperatures/australia/
↪melbourne/1982.csv')
```

```
df = pd.concat(
    [
        melbourne_1981_df,
        melbourne_1982_df,
    ]
)
```

```
df
```

```
            Date  Temp
0     1981-01-01  20.7
1     1981-01-02  17.9
2     1981-01-03  18.8
3     1981-01-04  14.6
4     1981-01-05  15.8
..           ...   ...
360   1982-12-27  15.3
361   1982-12-28  16.3
362   1982-12-29  15.8
363   1982-12-30  17.7
364   1982-12-31  16.3

[730 rows x 2 columns]
```

And there you have it! we now have a dataframe containing both data from 1981 as 1982. Can you figure out what I calculated in the next cell? Do you think there might be a more 'clean' solution?

```
df[df.Date.str[5:7]== '01'].Temp.mean()
```

```
17.140322580645158
```

As an exercise I would ask you now to create a small python script that given a begin and end year (between 1981 and 1990) can automatically concatenate all the necessary data

```
for i in range(1982,1987):
    print(i)
```

```
1982
1983
1984
1985
1986
```

4.2 Deduplication

Another important aspect of data cleaning is the removal of duplicates. Here we fragment of a dataset from activity on a popular games platform. We can see which user has either bought or played specific games and how often. Unfortunately for some reason, entries might have duplicates which we have to deal with as otherwise users might have e.g. bought a game twice.

```
df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-practical-
↪approach/main/src/c2_data_preparation/data/steam.csv')
df
```

```
        user_id                                     game    action  freq
0      11373749                Sid Meier's Civilization IV  purchase   1.0
1      11373749                Sid Meier's Civilization IV      play   0.1
2      11373749                Sid Meier's Civilization IV  purchase   1.0
3      11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
4      11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
...         ...                                          ...       ...   ...
1834  112845094                                     Arma 2  purchase   1.0
```

(continues on next page)

(continued from previous page)

```
1835  112845094                  Grand Theft Auto San Andreas  purchase   1.0
1836  112845094                    Grand Theft Auto Vice City  purchase   1.0
1837  112845094                    Grand Theft Auto Vice City  purchase   1.0
1838  112845094                          Grand Theft Auto III  purchase   1.0

[1839 rows x 4 columns]
```

We have a dataframe with 1839 interactions, you can see that the freq either notes the amount they bought (which always 1 as there is not use in buying it more) or the amount in hours they played.

Let us straightforward ask pandas to remove all rows that have an exact duplicate

```
df.drop_duplicates()
```

```
        user_id                                          game    action  freq
0      11373749                   Sid Meier's Civilization IV  purchase   1.0
1      11373749                   Sid Meier's Civilization IV      play   0.1
3      11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
5      11373749          Sid Meier's Civilization IV Warlords  purchase   1.0
7      56038151                       Tom Clancy's H.A.W.X. 2  purchase   1.0
...         ...                                           ...       ...   ...
1831  112845094                  Grand Theft Auto San Andreas  purchase   1.0
1832  112845094                  Grand Theft Auto San Andreas      play   0.2
1833  112845094                          Grand Theft Auto III  purchase   1.0
1834  112845094                                        Arma 2  purchase   1.0
1836  112845094                    Grand Theft Auto Vice City  purchase   1.0

[1132 rows x 4 columns]
```

Alright! this seemed to have dropped 707 rows from our dataset, but we would like to know more about those. Let's ask which rows the algorithm has dropped:

```
df[df.duplicated()]
```

```
        user_id                                          game    action  freq
2      11373749                   Sid Meier's Civilization IV  purchase   1.0
4      11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
6      11373749          Sid Meier's Civilization IV Warlords  purchase   1.0
10     56038151                  Grand Theft Auto San Andreas  purchase   1.0
12     56038151                    Grand Theft Auto Vice City  purchase   1.0
...         ...                                           ...       ...   ...
1827   39146470          Sid Meier's Civilization IV Warlords  purchase   1.0
1830   48666962                                      Crysis 2  purchase   1.0
1835  112845094                  Grand Theft Auto San Andreas  purchase   1.0
1837  112845094                    Grand Theft Auto Vice City  purchase   1.0
1838  112845094                          Grand Theft Auto III  purchase   1.0

[707 rows x 4 columns]
```

Here we can see the duplicates, no particular pattern seems to be present, we could just for curiosity count the games that are duplicated

```
df[df.duplicated()].game.value_counts()
```

```
Grand Theft Auto San Andreas                    172
Grand Theft Auto Vice City                      103
```

(continues on next page)

(continued from previous page)

```
Sid Meier's Civilization IV                        98
Grand Theft Auto III                               90
Sid Meier's Civilization IV Beyond the Sword       80
Sid Meier's Civilization IV Warlords               79
Sid Meier's Civilization IV Colonization           75
Crysis 2                                            7
Arma 2                                              1
Tom Clancy's H.A.W.X. 2                             1
TERA                                                1
Name: game, dtype: int64
```

It seems there are some games which are very prone to being duplicated, at this point we could go and ask the IT department why these games are acting weird.

Another thing im interested about is the perspective of a single gamer, here we took a single user_id and printed all his games

```
df[df.user_id == 11373749]
```

```
    user_id                                          game    action  freq
0  11373749                   Sid Meier's Civilization IV  purchase   1.0
1  11373749                   Sid Meier's Civilization IV      play   0.1
2  11373749                   Sid Meier's Civilization IV  purchase   1.0
3  11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
4  11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
5  11373749          Sid Meier's Civilization IV Warlords  purchase   1.0
6  11373749          Sid Meier's Civilization IV Warlords  purchase   1.0
```

Ah, you can see all of his three games are somehow duplicated in purchase, also it seems he only played one of them for only 0.1 hours. Looks like he fell to the bait of a tempting summer sale but didn't realise he had no time to actually play it.

Another thing I would like to mention here is that this dataset would make a fine recommender system as it contains user ids and hours played. Add game metadata (description) and reviews to the mix and your data preparation is done!

We can remove all duplicates now by overwriting our dataframe

```
df = df.drop_duplicates()
```

One thing still bothers me, as hours played can change over time it might be that different snapshots have produced different values, therefore more duplicates might be present with different hours_played.

Time to investigate this by using a subset of columns in the drop_duplicates algorithm

```
df.drop_duplicates(subset=['user_id', 'game', 'action'])
```

```
        user_id                                          game    action  freq
0      11373749                   Sid Meier's Civilization IV  purchase   1.0
1      11373749                   Sid Meier's Civilization IV      play   0.1
3      11373749  Sid Meier's Civilization IV Beyond the Sword  purchase   1.0
5      11373749          Sid Meier's Civilization IV Warlords  purchase   1.0
7      56038151                       Tom Clancy's H.A.W.X. 2  purchase   1.0
...         ...                                           ...       ...   ...
1831  112845094                      Grand Theft Auto San Andreas  purchase   1.0
1832  112845094                      Grand Theft Auto San Andreas      play   0.2
1833  112845094                          Grand Theft Auto III  purchase   1.0
1834  112845094                                        Arma 2  purchase   1.0
```

(continues on next page)

(continued from previous page)

```
1836  112845094                    Grand Theft Auto Vice City  purchase   1.0

[1120 rows x 4 columns]
```

Seems we have shaved off another 12 records, so our intuition was right, again lets see which the duplicates are:

```
df[df.duplicated(subset=['user_id', 'game', 'action'])]
```

```
        user_id                                       game action  freq
118   118664413                  Grand Theft Auto San Andreas   play   0.2
458    50769696                  Grand Theft Auto San Andreas   play   3.1
521    71411882                          Grand Theft Auto III   play   0.2
607    33865373                   Sid Meier's Civilization IV   play   2.0
898    71510748                  Grand Theft Auto San Andreas   play   0.2
908    28472068                    Grand Theft Auto Vice City   play   0.4
910    28472068                  Grand Theft Auto San Andreas   play   0.2
912    28472068                          Grand Theft Auto III   play   0.1
1506   59925638                       Tom Clancy's H.A.W.X. 2   play   0.3
1553  148362155                  Grand Theft Auto San Andreas   play  12.5
1709  176261926  Sid Meier's Civilization IV Beyond the Sword   play   0.4
1711  176261926                   Sid Meier's Civilization IV   play   0.2
```

As expected the duplicates are all in the 'play' action, to complete our view we extract the data of a single user

```
df[df.user_id==118664413]
```

```
       user_id                          game    action  freq
115  118664413  Grand Theft Auto San Andreas  purchase   1.0
116  118664413  Grand Theft Auto San Andreas      play   1.9
118  118664413  Grand Theft Auto San Andreas      play   0.2
```

It looks like we have a problem now, we know these are duplicates and should be removed, but which one? Personally I would argue here that we keep the highest value, as it is impossible to 'unplay' hours on the game. I will leave this as an exercise for you, but the solution is pretty tricky so i'll give a hint:

The algorithm always keeps the first record in case of duplicates, so you could sort the rows making sure the higher value is always encountered first, good luck!

CHAPTER

FIVE

OUTLIERS AND VALIDITY

When preparing data we have to be cautious with the accuracy of our set. Outliers and invalid data points are difficult to detect but should be handled with caution.

we start out by importing our most important library.

```
import pandas as pd
```

5.1 Silicon wafer thickness

Our first dataset contains information about the production of silicon wafers, each wafers thickness is measure on 9 different spots. More information on the dataset can be found here.

```
wafer_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-
↪practical-approach/main/src/c2_data_preparation/data/silicon-wafer-thickness.csv')
wafer_df.head()
```

```
      G1     G2     G3     G4     G5     G6     G7     G8     G9
0  0.175  0.188 -0.159  0.095  0.374 -0.238 -0.800  0.158 -0.211
1  0.102  0.075  0.141  0.180  0.138 -0.057 -0.075  0.072  0.072
2  0.607  0.711  0.879  0.765  0.592  0.187  0.431  0.345  0.187
3  0.774  0.823  0.619  0.370  0.725  0.439 -0.025 -0.259  0.496
4  0.504  0.644  0.845  0.681  0.502  0.151  0.404  0.296  0.260
```

we would like to investigate the distribution of measurements here, as we are early in this course using visualisation techniques would be too soon. This does not mean we can't use simple mathematics, introducing the InterQuartile Range. A reason for using IQR over standard deviation is that with IQR we do not assume a normal distribution. The IQR calculates the range between the bottom 'quart' or 25% and the top 25%, giving us an indication of the spread of our results, we calculate this IQR for each of the 9 measurements independently. For more info about IQR you can visit wikipedia.

```
iqr = wafer_df.quantile(0.75)-wafer_df.quantile(0.25)
iqr
```

```
G1    0.54425
G2    0.61000
G3    0.54075
G4    0.52475
G5    0.61175
G6    0.86750
G7    0.76175
```

(continues on next page)

(continued from previous page)

```
G8    0.87225
G9    0.86300
dtype: float64
```

you can see that the IQR spread for each measurement lays between 0.5 and 1 unit indicating that the 9 measurements of the wafer have a similar spread. With these IQR's we could calculate for each point relative to the spread of the measurement how far it is from the median.

```
relative_spread_df = (wafer_df-wafer_df.median())/iqr
relative_spread_df.head()
```

```
         G1        G2        G3        G4        G5        G6        G7  \
0 -0.011024 -0.077869 -0.819233 -0.367794  0.176543 -0.352738 -1.029865
1 -0.145154 -0.263115 -0.264448 -0.205812 -0.209236 -0.144092 -0.078110
2  0.782729  0.779508  1.100324  0.909004  0.532897  0.137176  0.586150
3  1.089573  0.963115  0.619510  0.156265  0.750306  0.427666 -0.012471
4  0.593477  0.669672  1.037448  0.748928  0.385779  0.095677  0.550706

         G8        G9
0 -0.130696 -0.254925
1 -0.229292  0.073001
2  0.083692  0.206257
3 -0.608770  0.564311
4  0.027515  0.290846
```

You can now see that some points are close to the median, whilst others are much higher, both positive as negative. By defining a threshold, we quantify what deviation has to be there to flag a reading as an outlier. The high outliers are seperated, note that only a single measurement of the 9 can trigger and render the total measurement as an outlier. Yet judging from the setup where we would want to find wafers with varying thickness that approach is desirable.

```
relative_spread_df[(relative_spread_df>2).any(axis='columns')]
```

```
            G1         G2         G3         G4         G5        G6  \
8     2.232430   2.009016   1.956542   1.589328   1.843890  1.544669
38   12.891135  12.827049  12.832178  13.913292  11.429506  9.500865
39    3.691318   3.981148   3.774387   4.081944   3.248059  3.729107
61    2.010106   2.153279   1.987980   1.863745   1.858602  1.274928
110   3.678457   2.841803   3.204808   3.180562   2.669391  0.518732
112   2.361047   2.086066   2.363384   2.107670   1.925623  1.238040
117   1.475425   1.043443   2.154415   2.582182   0.653862  1.823631
120   1.791456   1.484426   2.583449   1.440686   2.085819  0.990202
121   1.791456   1.484426   2.583449   1.440686   2.085819  0.990202
152   2.610932   2.102459   2.387425   2.549786   2.169187  1.730259
154  -0.529169  -0.538525  -0.404993  -0.331586  -0.552513  4.565994

            G7        G8        G9
8     1.233344  0.419604  1.582851
38   10.305875  9.927200  9.055620
39    3.304890  3.846374  3.149479
61    1.237283  0.825451  0.955968
110   0.700361  0.176555  0.727694
112   1.766328  0.890800  1.377752
117   1.581227  0.857552  1.188876
120   1.782081  1.034107  1.822711
121   1.782081  1.034107  1.822711
```

(continues on next page)

(continued from previous page)

```
152   2.241549  1.713958  1.592121
154  -0.051854 -0.382918 -0.536501
```

seems we have a few high outliers, you can clearly see the measurements are mostly all across the board high, but in some cases (e.g. id 154) only one measurement was an outlier. We can do the same for the low outliers.

```
relative_spread_df[(relative_spread_df<-2).any(axis='columns')]
```

```
           G1        G2        G3        G4        G5        G6        G7  \
54  -1.550758 -1.525410 -1.843736 -2.082897 -1.659174 -1.203458 -1.184772
56  -1.732660 -1.510656 -2.121128 -2.122916 -1.781774 -1.521614 -1.909419
59  -1.971520 -1.310656 -2.328248 -1.175798 -2.067838 -0.915274 -1.783394
64  -1.234727 -1.361475 -0.736015 -1.055741 -2.224765 -0.839193 -0.679357
65  -2.226918 -1.194262 -2.117429 -2.161029 -2.043318 -0.190202 -1.004923
102 -2.484153 -2.330328 -1.568192 -2.808957 -1.945239 -1.340634 -0.846078

           G8        G9
54  -1.650903 -1.245655
56  -1.782746 -1.159907
59  -1.304672 -1.514484
64  -0.865578 -0.663963
65  -0.270565 -0.794902
102 -1.691029 -0.887601
```

For a simple mathematical equation these result look promising, yet it can always be more sophisticated. Not going to deep into the subject we could perform some Machine Learning, using a unsupervised method. Here we use the sklearn library which contains the Isolation forest algorithm. More info about the algorithm here.

```
from sklearn.ensemble import IsolationForest
```

We first create the classifier and train (fit) it with the generic wafer data. Then for each record of the wafer data we make a prediction, if it thinks its an outlier, we keep them

```
clf = IsolationForest(random_state=0).fit(wafer_df)
wafer_df[clf.predict(wafer_df)==-1]
```

```
        G1     G2     G3     G4     G5     G6     G7     G8     G9
8    1.396  1.461  1.342  1.122  1.394  1.408  0.924  0.638  1.375
20  -0.558 -0.705 -0.526 -0.412 -0.753 -0.998 -0.270  0.598 -1.416
38   7.197  8.060  7.223  7.589  7.258  8.310  7.835  8.931  7.824
39   2.190  2.664  2.325  2.430  2.253  3.303  2.502  3.627  2.727
54  -0.663 -0.695 -0.713 -0.805 -0.749 -0.976 -0.918 -1.168 -1.066
56  -0.762 -0.686 -0.863 -0.826 -0.824 -1.252 -1.470 -1.283 -0.992
59  -0.892 -0.564 -0.975 -0.329 -0.999 -0.726 -1.374 -0.866 -1.298
61   1.275  1.549  1.359  1.266  1.403  1.174  0.927  0.992  0.834
65  -1.031 -0.493 -0.861 -0.846 -0.984 -0.097 -0.781  0.036 -0.677
102 -1.171 -1.186 -0.564 -1.186 -0.924 -1.095 -0.660 -1.203 -0.757
106 -0.659 -0.451 -0.692 -0.708 -0.595 -0.726 -1.031 -0.877 -1.080
110  2.183  1.969  2.017  1.957  1.899  0.518  0.518  0.426  0.637
112  1.466  1.508  1.562  1.394  1.444  1.142  1.330  1.049  1.198
117  0.984  0.872  1.449  1.643  0.666  1.650  1.189  1.020  1.035
120  1.156  1.141  1.681  1.044  1.542  0.927  1.342  1.174  1.582
121  1.156  1.141  1.681  1.044  1.542  0.927  1.342  1.174  1.582
152  1.602  1.518  1.575  1.626  1.593  1.569  1.692  1.767  1.383
```

Comparing the results with our IQR approach we see a lot of similarities, here the id 154 record did not show up as we already realised this was perhaps not a strong enough outlier. You could enhance our IQR technique by checking the amount of measurements that are above the threshold and respond accordingly, I will leave you a little hint.

```
(relative_spread_df>2).sum()
```

```
G1     7
G2     7
G3     8
G4     6
G5     6
G6     3
G7     3
G8     2
G9     2
dtype: int64
```

5.2 Distillation column

As an exercise you can try the same technique to this dataset and see what you would find, good luck! Be mindful that you do not incorporate the date as a variable in your outlier algorithm.

```
distil_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-
↪practical-approach/main/src/c2_data_preparation/data/distillation-tower.csv')
distil_df
```

```
           Date     Temp1    FlowC1     Temp2    TempC1     Temp3    TempC2  \
0    2000-08-21  139.9857  432.0636  377.8119  100.2204  492.1353  490.1459
1    2000-08-23  131.0470  487.4029  371.3060  100.2297  482.2100  480.3128
2    2000-08-26  118.2666  437.3516  378.4483  100.3084  488.7266  487.0040
3    2000-08-29  118.1769  481.8314  378.0028   95.5766  493.1481  491.1137
4    2000-08-30  120.7891  412.6471  377.8871   92.9052  490.2486  488.6641
..          ...       ...       ...       ...       ...       ...       ...
248  2003-01-26  130.8138  212.6385  341.5964  121.4354  468.3401  467.0299
249  2003-01-28  128.9673  225.1412  349.8965  118.8604  479.7665  478.4652
250  2003-01-31  130.5328  223.5965  345.9366  120.4027  474.5378  473.1145
251  2003-02-03  128.5248  213.5613  343.4950  119.6989  469.3802  467.9954
252  2003-02-04  131.0491  217.4117  346.1960  119.0825  474.6599  473.0381

       TempC3     Temp4  PressureC1  ...     Temp10  FlowC3   FlowC4   Temp11  \
0    180.5578  187.4331    215.0627  ...   513.9653  8.6279  10.5988  30.8983
1    172.6575  179.5089    205.0999  ...   504.5145  8.7662  10.7560  31.9099
2    165.9400  172.9262    205.0304  ...   508.9997  8.5319  10.5737  29.9165
3    167.2085  174.2338    205.2561  ...   514.1794  8.6260  10.6695  30.6229
4    167.0326  173.9681    205.0883  ...   511.0948  8.5939  10.4922  29.4977
..        ...       ...         ...  ...        ...     ...      ...      ...
248  174.7639  180.7649    229.7393  ...   479.0290  5.5590   6.4470  16.4131
249  176.2176  182.3646    230.5049  ...   491.2362  5.6342   6.4360  17.2385
250  176.3310  182.2578    230.6638  ...   485.8786  5.4810   6.3575  16.9866
251  174.6435  180.5093    230.5226  ...   480.2879  5.4727   6.4175  16.6778
252  177.1088  183.1810    225.6420  ...   486.0253  5.4597   6.3291  16.8766

       Temp12  InvTemp1  InvTemp2  InvTemp3  InvPressure1  VapourPressure
0    489.9900    2.0409    2.6468    2.1681        4.3524         32.5026
1    480.2888    2.0821    2.6932    2.2207        4.5497         34.8598
```

(continues on next page)

(continued from previous page)

```
2    486.6190    2.0550    2.6424    2.1796         4.5511         32.1666
3    491.1304    2.0361    2.6455    2.1620         4.5464         30.4064
4    487.6475    2.0507    2.6463    2.1704         4.5499         30.9238
..        ...       ...       ...       ...            ...             ...
248  466.3347    2.1444    2.9274    2.2127         4.0911         38.8507
249  477.8816    2.0926    2.8580    2.1620         4.0783         34.2653
250  472.3176    2.1172    2.8907    2.1855         4.0756         36.5717
251  467.0001    2.1413    2.9113    2.2090         4.0780         38.1054
252  472.2701    2.1174    2.8885    2.1844         4.1608         35.6298

[253 rows x 28 columns]
```

CHAPTER

SIX

STRING OPERATIONS

CHAPTER

SEVEN

DATETIME OPERATIONS

When our dataset contains time-related data, datetime operations are a great asset to our data science toolkit. For this exercise we obtain a public covid dataset containing A LOT of information on infection cases, deaths, tests and vaccinations.

Let's start by importing the data, as the dataset is about 60MB at the time of writing, this might take some time. Perhaps you could think of a method to make this more efficient, do we always need all of the data?

More info about the data can be found here

```
import pandas as pd
```

```
covid_df = pd.read_csv('https://raw.githubusercontent.com/owid/covid-19-data/master/
↪public/data/owid-covid-data.csv', on_bad_lines='skip')
covid_df.head()
```

```
  iso_code continent     location        date  total_cases  new_cases  \
0      AFG      Asia  Afghanistan  2020-02-24          5.0        5.0
1      AFG      Asia  Afghanistan  2020-02-25          5.0        0.0
2      AFG      Asia  Afghanistan  2020-02-26          5.0        0.0
3      AFG      Asia  Afghanistan  2020-02-27          5.0        0.0
4      AFG      Asia  Afghanistan  2020-02-28          5.0        0.0

   new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  ...  \
0                 NaN           NaN         NaN                  NaN  ...
1                 NaN           NaN         NaN                  NaN  ...
2                 NaN           NaN         NaN                  NaN  ...
3                 NaN           NaN         NaN                  NaN  ...
4                 NaN           NaN         NaN                  NaN  ...

   female_smokers  male_smokers  handwashing_facilities  \
0             NaN           NaN                  37.746
1             NaN           NaN                  37.746
2             NaN           NaN                  37.746
3             NaN           NaN                  37.746
4             NaN           NaN                  37.746

   hospital_beds_per_thousand  life_expectancy  human_development_index  \
0                         0.5            64.83                    0.511
1                         0.5            64.83                    0.511
2                         0.5            64.83                    0.511
3                         0.5            64.83                    0.511
4                         0.5            64.83                    0.511

   excess_mortality_cumulative_absolute  excess_mortality_cumulative  \
```

(continues on next page)

(continued from previous page)

```
0                                    NaN                            NaN
1                                    NaN                            NaN
2                                    NaN                            NaN
3                                    NaN                            NaN
4                                    NaN                            NaN

   excess_mortality  excess_mortality_cumulative_per_million
0               NaN                                      NaN
1               NaN                                      NaN
2               NaN                                      NaN
3               NaN                                      NaN
4               NaN                                      NaN

[5 rows x 65 columns]
```

As mentioned a lot of information is present here, about 65 columns. yet for this exercise my main objective is the 'date' column. If we would print out the data types using the info method, we can see that the date is recognized as an 'object' stating that it is an ordinary string, not a datetime.

```
covid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121744 entries, 0 to 121743
Data columns (total 65 columns):
 #   Column                                   Non-Null Count   Dtype
---  ------                                   --------------   -----
 0   iso_code                                 121744 non-null  object
 1   continent                                116202 non-null  object
 2   location                                 121744 non-null  object
 3   date                                     121744 non-null  object
 4   total_cases                              115518 non-null  float64
 5   new_cases                                115515 non-null  float64
 6   new_cases_smoothed                       114500 non-null  float64
 7   total_deaths                             104708 non-null  float64
 8   new_deaths                               104863 non-null  float64
 9   new_deaths_smoothed                      114500 non-null  float64
 10  total_cases_per_million                  114910 non-null  float64
 11  new_cases_per_million                    114907 non-null  float64
 12  new_cases_smoothed_per_million           113897 non-null  float64
 13  total_deaths_per_million                 104113 non-null  float64
 14  new_deaths_per_million                   104268 non-null  float64
 15  new_deaths_smoothed_per_million          113897 non-null  float64
 16  reproduction_rate                        98318 non-null   float64
 17  icu_patients                             14443 non-null   float64
 18  icu_patients_per_million                 14443 non-null   float64
 19  hosp_patients                            16504 non-null   float64
 20  hosp_patients_per_million                16504 non-null   float64
 21  weekly_icu_admissions                    1268 non-null    float64
 22  weekly_icu_admissions_per_million        1268 non-null    float64
 23  weekly_hosp_admissions                   2088 non-null    float64
 24  weekly_hosp_admissions_per_million       2088 non-null    float64
 25  new_tests                                52248 non-null   float64
 26  total_tests                              52352 non-null   float64
 27  total_tests_per_thousand                 52352 non-null   float64
 28  new_tests_per_thousand                   52248 non-null   float64
 29  new_tests_smoothed                       62816 non-null   float64
```

(continues on next page)

(continued from previous page)

```
 30  new_tests_smoothed_per_thousand          62816 non-null   float64
 31  positive_rate                            58959 non-null   float64
 32  tests_per_case                           58319 non-null   float64
 33  tests_units                              64746 non-null   object
 34  total_vaccinations                       28115 non-null   float64
 35  people_vaccinated                        26746 non-null   float64
 36  people_fully_vaccinated                  23714 non-null   float64
 37  total_boosters                           3057 non-null    float64
 38  new_vaccinations                         23298 non-null   float64
 39  new_vaccinations_smoothed                50221 non-null   float64
 40  total_vaccinations_per_hundred           28115 non-null   float64
 41  people_vaccinated_per_hundred            26746 non-null   float64
 42  people_fully_vaccinated_per_hundred      23714 non-null   float64
 43  total_boosters_per_hundred               3057 non-null    float64
 44  new_vaccinations_smoothed_per_million    50221 non-null   float64
 45  stringency_index                         101767 non-null  float64
 46  population                               120880 non-null  float64
 47  population_density                       112501 non-null  float64
 48  median_age                               107423 non-null  float64
 49  aged_65_older                            106229 non-null  float64
 50  aged_70_older                            106834 non-null  float64
 51  gdp_per_capita                           108055 non-null  float64
 52  extreme_poverty                          72482 non-null   float64
 53  cardiovasc_death_rate                    107695 non-null  float64
 54  diabetes_prevalence                      111063 non-null  float64
 55  female_smokers                           84078 non-null   float64
 56  male_smokers                             82858 non-null   float64
 57  handwashing_facilities                   54111 non-null   float64
 58  hospital_beds_per_thousand               97911 non-null   float64
 59  life_expectancy                          115458 non-null  float64
 60  human_development_index                  107790 non-null  float64
 61  excess_mortality_cumulative_absolute     4317 non-null    float64
 62  excess_mortality_cumulative              4317 non-null    float64
 63  excess_mortality                         4317 non-null    float64
 64  excess_mortality_cumulative_per_million  4317 non-null    float64
dtypes: float64(60), object(5)
memory usage: 60.4+ MB
```

We would like to change that, as we can only perform datetime operations if pandas recognises the datetime format used. Good for us, pandas has a method to automatically infer the date format, we do that now.

```
covid_df.date = pd.to_datetime(covid_df.date)
covid_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121744 entries, 0 to 121743
Data columns (total 65 columns):
 #   Column                                   Non-Null Count   Dtype
---  ------                                   --------------   -----
 0   iso_code                                 121744 non-null  object
 1   continent                                116202 non-null  object
 2   location                                 121744 non-null  object
 3   date                                     121744 non-null  datetime64[ns]
 4   total_cases                              115518 non-null  float64
 5   new_cases                                115515 non-null  float64
 6   new_cases_smoothed                       114500 non-null  float64
```

(continues on next page)

(continued from previous page)

```
 7   total_deaths                            104708 non-null  float64
 8   new_deaths                              104863 non-null  float64
 9   new_deaths_smoothed                     114500 non-null  float64
 10  total_cases_per_million                 114910 non-null  float64
 11  new_cases_per_million                   114907 non-null  float64
 12  new_cases_smoothed_per_million          113897 non-null  float64
 13  total_deaths_per_million                104113 non-null  float64
 14  new_deaths_per_million                  104268 non-null  float64
 15  new_deaths_smoothed_per_million         113897 non-null  float64
 16  reproduction_rate                       98318 non-null   float64
 17  icu_patients                            14443 non-null   float64
 18  icu_patients_per_million                14443 non-null   float64
 19  hosp_patients                           16504 non-null   float64
 20  hosp_patients_per_million               16504 non-null   float64
 21  weekly_icu_admissions                   1268 non-null    float64
 22  weekly_icu_admissions_per_million       1268 non-null    float64
 23  weekly_hosp_admissions                  2088 non-null    float64
 24  weekly_hosp_admissions_per_million      2088 non-null    float64
 25  new_tests                               52248 non-null   float64
 26  total_tests                             52352 non-null   float64
 27  total_tests_per_thousand                52352 non-null   float64
 28  new_tests_per_thousand                  52248 non-null   float64
 29  new_tests_smoothed                      62816 non-null   float64
 30  new_tests_smoothed_per_thousand         62816 non-null   float64
 31  positive_rate                           58959 non-null   float64
 32  tests_per_case                          58319 non-null   float64
 33  tests_units                             64746 non-null   object
 34  total_vaccinations                      28115 non-null   float64
 35  people_vaccinated                       26746 non-null   float64
 36  people_fully_vaccinated                 23714 non-null   float64
 37  total_boosters                          3057 non-null    float64
 38  new_vaccinations                        23298 non-null   float64
 39  new_vaccinations_smoothed               50221 non-null   float64
 40  total_vaccinations_per_hundred          28115 non-null   float64
 41  people_vaccinated_per_hundred           26746 non-null   float64
 42  people_fully_vaccinated_per_hundred     23714 non-null   float64
 43  total_boosters_per_hundred              3057 non-null    float64
 44  new_vaccinations_smoothed_per_million   50221 non-null   float64
 45  stringency_index                        101767 non-null  float64
 46  population                              120880 non-null  float64
 47  population_density                      112501 non-null  float64
 48  median_age                              107423 non-null  float64
 49  aged_65_older                           106229 non-null  float64
 50  aged_70_older                           106834 non-null  float64
 51  gdp_per_capita                          108055 non-null  float64
 52  extreme_poverty                         72482 non-null   float64
 53  cardiovasc_death_rate                   107695 non-null  float64
 54  diabetes_prevalence                     111063 non-null  float64
 55  female_smokers                          84078 non-null   float64
 56  male_smokers                            82858 non-null   float64
 57  handwashing_facilities                  54111 non-null   float64
 58  hospital_beds_per_thousand              97911 non-null   float64
 59  life_expectancy                         115458 non-null  float64
 60  human_development_index                 107790 non-null  float64
 61  excess_mortality_cumulative_absolute    4317 non-null    float64
 62  excess_mortality_cumulative             4317 non-null    float64
 63  excess_mortality                        4317 non-null    float64
```

(continues on next page)

(continued from previous page)

```
 64  excess_mortality_cumulative_per_million  4317 non-null    float64
dtypes: datetime64[ns](1), float64(60), object(4)
memory usage: 60.4+ MB
```

now we are ready to perform datetime operations, however we can see that dates are appearing multiple times, this because we have records for multiple countries. I live in Belgium, so decided to isolate that subsection of the data. If they had used a data lake and partitioned into countries, reading out the data would have been much more efficient, but efficiency is not something I would expect from government as a Belgian.

```
covid_belgium_df = covid_df[covid_df.location=='Belgium'].set_index('date')
covid_belgium_df.head()
```

```
           iso_code continent location  total_cases  new_cases  \
date
2020-02-04      BEL    Europe  Belgium          1.0        1.0
2020-02-05      BEL    Europe  Belgium          1.0        0.0
2020-02-06      BEL    Europe  Belgium          1.0        0.0
2020-02-07      BEL    Europe  Belgium          1.0        0.0
2020-02-08      BEL    Europe  Belgium          1.0        0.0

            new_cases_smoothed  total_deaths  new_deaths  new_deaths_smoothed  \
date
2020-02-04                 NaN           NaN         NaN                  NaN
2020-02-05                 NaN           NaN         NaN                  NaN
2020-02-06                 NaN           NaN         NaN                  NaN
2020-02-07                 NaN           NaN         NaN                  NaN
2020-02-08                 NaN           NaN         NaN                  NaN

            total_cases_per_million  ...  female_smokers  male_smokers  \
date                                 ...
2020-02-04                    0.086  ...            25.1          31.4
2020-02-05                    0.086  ...            25.1          31.4
2020-02-06                    0.086  ...            25.1          31.4
2020-02-07                    0.086  ...            25.1          31.4
2020-02-08                    0.086  ...            25.1          31.4

            handwashing_facilities  hospital_beds_per_thousand  \
date
2020-02-04                     NaN                        5.64
2020-02-05                     NaN                        5.64
2020-02-06                     NaN                        5.64
2020-02-07                     NaN                        5.64
2020-02-08                     NaN                        5.64

            life_expectancy  human_development_index  \
date
2020-02-04            81.63                    0.931
2020-02-05            81.63                    0.931
2020-02-06            81.63                    0.931
2020-02-07            81.63                    0.931
2020-02-08            81.63                    0.931

            excess_mortality_cumulative_absolute  excess_mortality_cumulative  \
date
2020-02-04                                   NaN                          NaN
2020-02-05                                   NaN                          NaN
```

(continues on next page)

(continued from previous page)

```
2020-02-06                                    NaN                          NaN
2020-02-07                                    NaN                          NaN
2020-02-08                                    NaN                          NaN

            excess_mortality  excess_mortality_cumulative_per_million
date
2020-02-04               NaN                                      NaN
2020-02-05               NaN                                      NaN
2020-02-06               NaN                                      NaN
2020-02-07               NaN                                      NaN
2020-02-08               NaN                                      NaN

[5 rows x 64 columns]
```

Now that we have our dataset containing only Belgium I would like to emphasize another aspect, for features such as population density we would not expect a 'head count' to differ each day, and as we can see this number is steady over the whole line (results may vary for those who execute this in the future).

```
covid_belgium_df.population.value_counts()
```

```
11632334.0    611
Name: population, dtype: int64
```

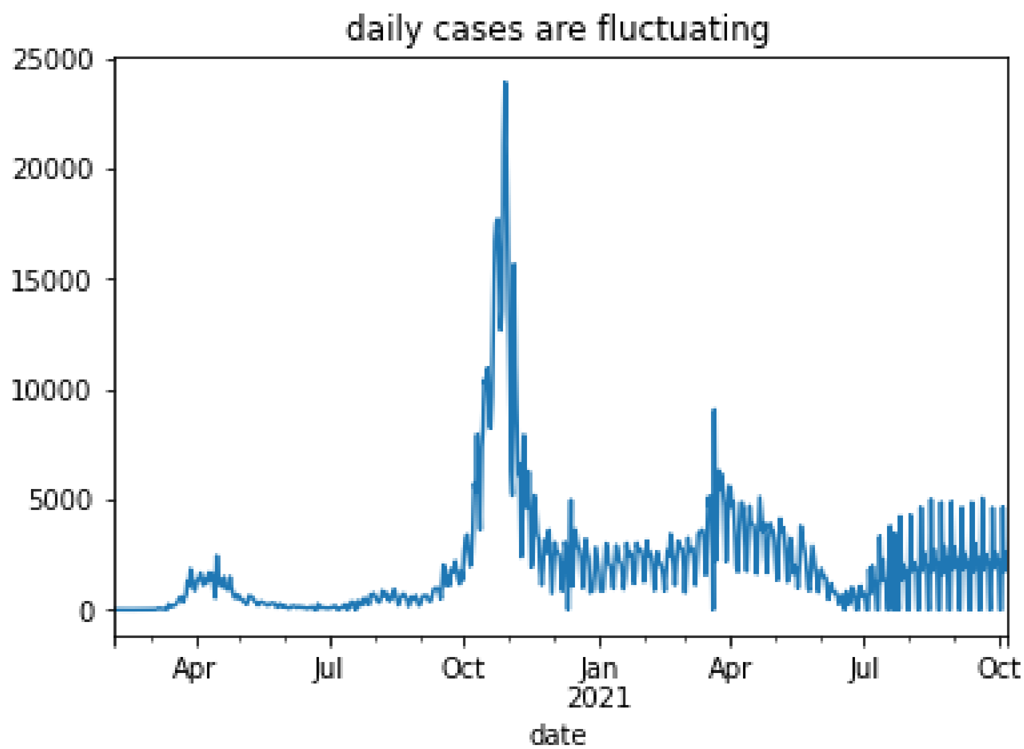
we only have a single value (in my case 11.6M) that is repeated over the whole dataset, would this look optimal to you? How would you perhaps approach this to improve data management? If you would like to go hands-on I left you a blank cell to experiment.

Optimalizations aside, we can not do that which we came for! Datetime operations, the first thing that I have in mind is that due to weekends, the cases might fluctuate a lot per day, so it is not optimal to view it on a daily basis.

First we create a simple line plot with the raw daily cases, then we perform a weekly sum to create a more smooth version of the new cases.

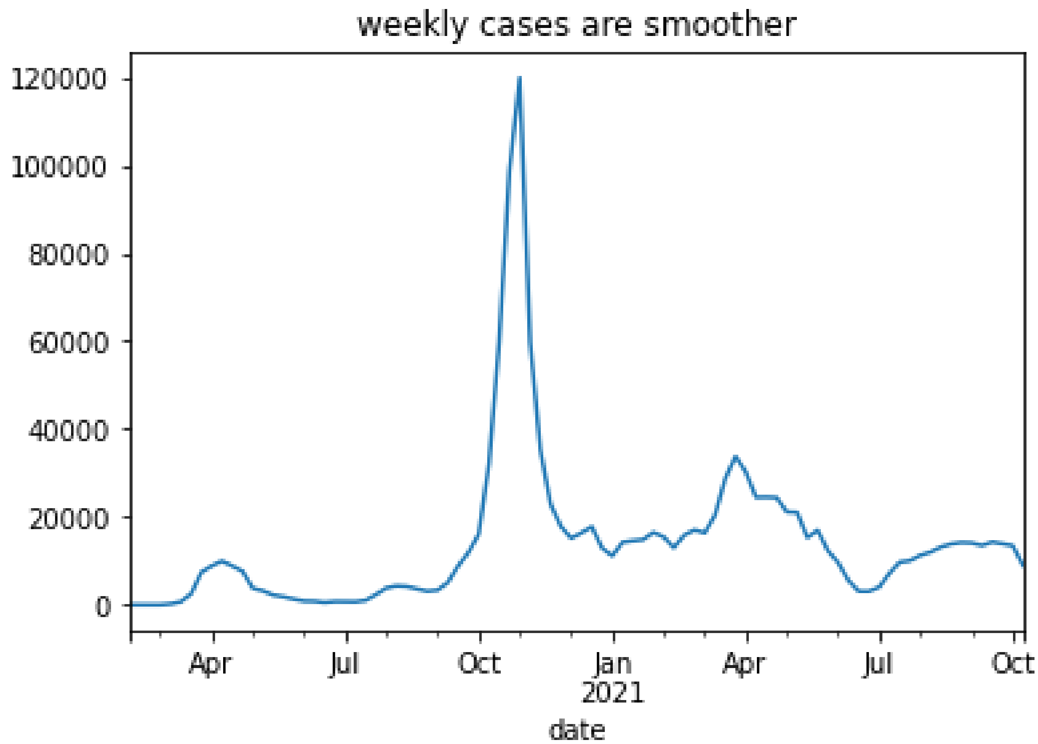
```
covid_belgium_df['new_cases'].plot(title='daily cases are fluctuating')
```

```
<AxesSubplot:title={'center':'daily cases are fluctuating'}, xlabel='date'>
```




```
weekly_cases_df = covid_belgium_df['new_cases'].resample('W').sum()
weekly_cases_df.plot(title='weekly cases are smoother')
```

```
<AxesSubplot:title={'center':'weekly cases are smoother'}, xlabel='date'>
```



That looks great! Those who inspected carefully saw that the x-axis was correclty identified as datetimes and that the y-axis for weekly sums have a much higher range.

In a next example we would like to have the relative changes from week to week, this can be done using the shift operation.

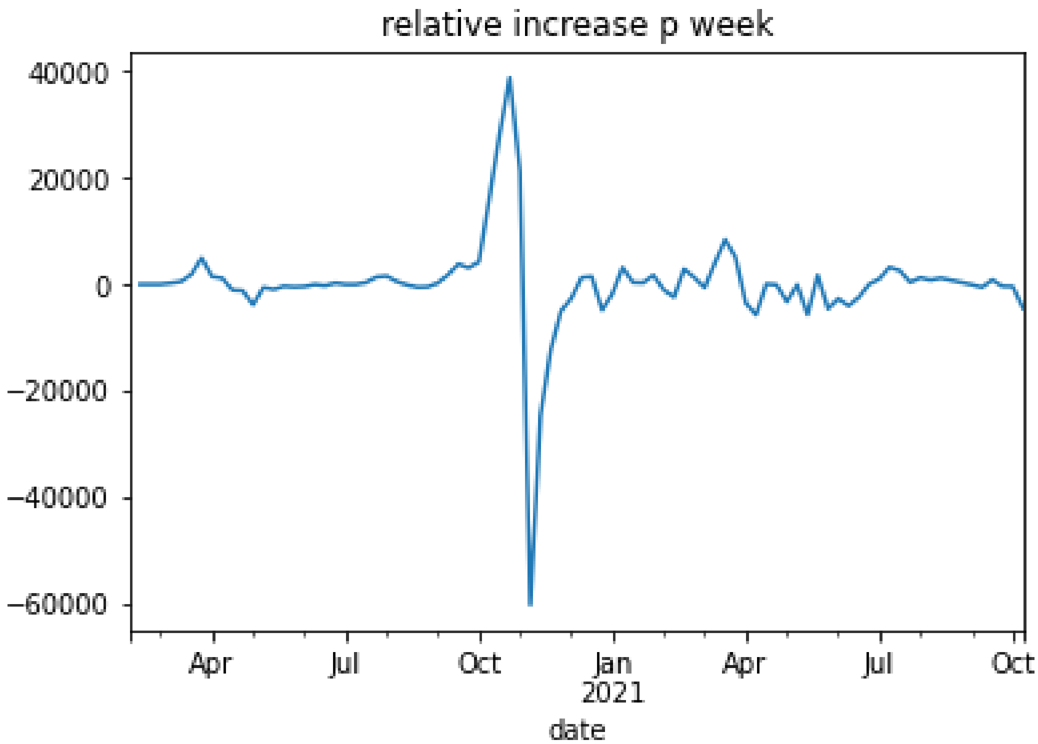
```
weekly_cases_df.shift(1)
```

```
date
2020-02-09        NaN
2020-02-16        1.0
2020-02-23        0.0
2020-03-01        0.0
2020-03-08        1.0
               ...
2021-09-12    14099.0
2021-09-19    13508.0
2021-09-26    14298.0
2021-10-03    13909.0
2021-10-10    13474.0
Freq: W-SUN, Name: new_cases, Length: 88, dtype: float64
```

This method shifted our data by 1 week forwards, this way we can subtract these results from our original data creating a relative increase (this_week_cases - last_week_cases).

```
(weekly_cases_df-weekly_cases_df.shift(1)).plot(title='relative increase p week')
```

```
<AxesSubplot:title={'center':'relative increase p week'}, xlabel='date'>
```



Another powerfull asset of datetimes is that we can utilize the concepts of days, weeks, months and years. In Belgium they speak about a phenomenon called 'the weekend effect' where a lot of reports are delayed and therefore Sundays have less cases whereas Mondays have more.

Do we see that in our data? let us seperate the Sundays and Mondays and take a mean!

```
print('mean deaths on Monday')
covid_belgium_df.loc[covid_belgium_df.index.dayofweek==0,"new_deaths"].mean()
```

```
mean deaths on Monday
```

```
39.02439024390244
```

```
print('mean deaths on Sunday')
covid_belgium_df.loc[covid_belgium_df.index.dayofweek==6,"new_deaths"].mean()
```

```
mean deaths on Sunday
```

```
36.646341463414636
```

It seems indeed that more people are reported to pass away no a Monday than on a Sunday, it would be optimal to verify this with statistics, but for now we keep it simple.

As a last example I would like to use slicing of our dataset to demonstrate we can also take a subset of our data and handle this, here we took the months of dec2020-jan2021 for belgium and calculated the total deaths.

```
covid_belgium_df.loc['2020-12-01':'2021-01-31'].new_deaths.sum()
```

```
4447.0
```

Now let's compare this to our neighbours, the Netherlands and France, we do exactly the same operations by selecting exaclty the same time window.

```
covid_netherlands_df = covid_df[covid_df.location=='Netherlands'].set_index('date')
covid_netherlands_df.loc['2020-12-01':'2021-01-31'].new_deaths.sum()
```

```
4655.0
```

```
covid_france_df = covid_df[covid_df.location=='France'].set_index('date')
covid_france_df.loc['2020-12-01':'2021-01-31'].new_deaths.sum()
```

```
23382.0
```

You can see that Belgium has the lowest of total deaths in that time interval, so you could assume we performed the best! However this approach is a bit simplified as there are not as many Belgians as French and Dutch. Could you perhaps think if an improvement to create a better understanding?

CHAPTER

EIGHT

CATEGORICAL ENCODING

Often we deal with categorical data and this kind of data is something computer algorithms are not able to understand. On the other hand long categorical features might take up unnecessary memory in our dataset, so converting to a categorical feature is optimal.

```
import pandas as pd
```

8.1 Raw Material Charaterization

In this dataset, we have a few numerical features describing characteristics of our material, next to that we also have an Outcome feature describing the state of our material in a category.

Let's have a look at the data

```
raw_material_df = pd.read_csv('./data/raw-material-characterization.csv')
raw_material_df.head()
```

```
  Lot number   Outcome  Size5  Size10  Size15    TGA   DSC   TMA
0       B370  Adequate   13.8     9.2    41.2  787.3  18.0  65.0
1       B880  Adequate   11.2     5.8    27.6  772.2  17.7  68.8
2       B452  Adequate    9.9     5.8    28.3  602.3  18.3  50.7
3       B287  Adequate   10.4     4.0    24.7  677.9  17.7  56.5
4       B576  Adequate   12.3     9.3    22.0  593.5  19.5  52.0
```

So we can see that the outcome is indeed a text field with a human interpretable value. The different values are:

```
raw_material_df.Outcome.unique()
```

```
array(['Adequate', 'Poor'], dtype=object)
```

Image that we would like to get all records where the Outcome is less than adequate, using strings this is not possible as the computer does not understand relations of Adequate and Poor when they are denoted as text.

```
raw_material_df[raw_material_df.Outcome<'Adequate']
```

```
Empty DataFrame
Columns: [Lot number, Outcome, Size5, Size10, Size15, TGA, DSC, TMA]
Index: []
```

To overcome this we can change the type of the feature from 'object' (string) to 'category' let us look at the data types of our data

```
raw_material_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24 entries, 0 to 23
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Lot number  24 non-null     object
 1   Outcome     24 non-null     object
 2   Size5       24 non-null     float64
 3   Size10      24 non-null     float64
 4   Size15      24 non-null     float64
 5   TGA         24 non-null     float64
 6   DSC         24 non-null     float64
 7   TMA         24 non-null     float64
dtypes: float64(6), object(2)
memory usage: 1.6+ KB
```

Now we can change that of Outcome to category using the astype method

```
raw_material_df.Outcome = raw_material_df.Outcome.astype('category')
raw_material_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24 entries, 0 to 23
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Lot number  24 non-null     object
 1   Outcome     24 non-null     category
 2   Size5       24 non-null     float64
 3   Size10      24 non-null     float64
 4   Size15      24 non-null     float64
 5   TGA         24 non-null     float64
 6   DSC         24 non-null     float64
 7   TMA         24 non-null     float64
dtypes: category(1), float64(6), object(1)
memory usage: 1.6+ KB
```

Our feature might be of categorical nature now, however we still have to define it is an ordinal category and has an order.

```
raw_material_df.Outcome = raw_material_df.Outcome.cat.as_ordered().cat.reorder_
↪categories(['Poor', 'Adequate'])
```

If we retry to effort to only take the records where the Outcome is less than Adequate, we now get an outcome! Since we only have 2 categories this is a bit unfortunate, but you should get the idea behind it.

```
raw_material_df[raw_material_df.Outcome<'Adequate']
```

```
   Lot number Outcome  Size5  Size10  Size15    TGA   DSC   TMA
5        B914    Poor   13.7     7.8    27.0  597.9  18.1  49.8
6        B404    Poor   15.5    10.7    34.3  668.5  19.6  55.7
7        B694    Poor   15.4    10.7    35.9  602.8  19.2  53.6
8        B875    Poor   14.9    11.3    41.0  614.6  18.5  50.0
10       B517    Poor   16.1    11.6    39.2  682.8  17.5  56.4
```

(continues on next page)

(continued from previous page)

```
13        B430    Poor    12.9      9.7     36.3   642.4  19.1  55.0
21        B745    Poor    10.2      5.8     24.7   575.9  18.5  46.2
```

Let's take this a step further, since computer algorithms still have no idea what the numerical relation is between Adequate and Poor, we could use a Label Encoder for that.

```
from sklearn.preprocessing import LabelEncoder
```

the label encoder is inputted with the Outcome feature and recognises 2 types, it chooses a numerical value for each while fitting.

```
le = LabelEncoder()
le.fit(raw_material_df.Outcome)
```

```
LabelEncoder()
```

After fitting we can use this encoder to transform our dataset!

```
raw_material_df['outcome_label'] = le.transform(raw_material_df.Outcome)
raw_material_df.head()
```

```
  Lot number   Outcome  Size5  Size10  Size15    TGA   DSC   TMA  \
0       B370  Adequate   13.8     9.2    41.2  787.3  18.0  65.0
1       B880  Adequate   11.2     5.8    27.6  772.2  17.7  68.8
2       B452  Adequate    9.9     5.8    28.3  602.3  18.3  50.7
3       B287  Adequate   10.4     4.0    24.7  677.9  17.7  56.5
4       B576  Adequate   12.3     9.3    22.0  593.5  19.5  52.0

   outcome_label
0              0
1              0
2              0
3              0
4              0
```

It seems something unfortunate has happened, the encoder gave the Adequate an outcome label of 0, which is lower than the label of Poor (1), this might be bad if we would like to give a score as our outcome.

There is in pandas another method of mapping a label to a category albeit less automated, as you would have to know the categories in your feature.

```
raw_material_df.outcome_label = raw_material_df.Outcome.map({'Poor': 0, 'Adequate':1})
raw_material_df.head()
```

```
  Lot number   Outcome  Size5  Size10  Size15    TGA   DSC   TMA outcome_label
0       B370  Adequate   13.8     9.2    41.2  787.3  18.0  65.0             1
1       B880  Adequate   11.2     5.8    27.6  772.2  17.7  68.8             1
2       B452  Adequate    9.9     5.8    28.3  602.3  18.3  50.7             1
3       B287  Adequate   10.4     4.0    24.7  677.9  17.7  56.5             1
4       B576  Adequate   12.3     9.3    22.0  593.5  19.5  52.0             1
```

Yes! This did the trick, now we can use that outcome label to predict an outcome for future samples.

CHAPTER NINE

RESTAURANT TIPS

Now we are going to look at a dataset of tips, here a restaurant tracked the table bills and tips for a few days in the week whilst also noting the gender, smoking habit and time of day. This led to a small yet very interesting dataset, let's have a look!

```
tips_df = pd.read_csv('https://raw.githubusercontent.com/mwaskom/seaborn-data/master/
↪tips.csv')
tips_df
```

```
     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
..          ...   ...     ...    ...   ...     ...   ...
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

[244 rows x 7 columns]
```

We can see here that we have a lot of categorical variables: gender, smoker, the day and the time. In later sections we will see how we can aggregate on these categorical variables. Now however we would like to process them for a machine learning exercise, where we need numbers not text. For the features smoker and day, you could argue there is a clear numbering between them, smoking is 1 if the person was smoking and e.g. Sun relates to 7 as it is the seventh day of the week.

But for the gender this is different, we can't really say that women are 1 and Men are 0 or vice versa (although in this binary case it might work). The same theory applies for time, if we would say that breakfast, lunch and dinner equal to 0, 1 and 2 this would give our algorithm a bad impression as it would think dinner is twice lunch…

We use One Hot Encoding for this, the idea is that for each value of the feature we create a new column.

```
from sklearn.preprocessing import OneHotEncoder
```

First we create our encoder, then we give it the day column to learn and see which values of categories there are.

```
ohe = OneHotEncoder()
ohe.fit(tips_df[['day']])
```

```
OneHotEncoder()
```

Now we can perform an encoding, here we insert the day column and it returns a matrix with 4 columns corresponding to the 4 days in our feature.

```
ohe.transform(tips_df[['day']]).todense()
```

```
matrix([[0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
        [0., 0., 1., 0.],
```

(continues on next page)

(continued from previous page)

```
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
```

(continues on next page)

(continued from previous page)

```
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
```

(continues on next page)

(continued from previous page)

```
       [0., 0., 1., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 1., 0.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 0., 0., 1.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [0., 1., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
```

(continues on next page)

(continued from previous page)

```
        [1., 0., 0., 0.],
        [1., 0., 0., 0.],
        [1., 0., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 1., 0., 0.],
        [0., 0., 0., 1.]])
```

As this is a rather mathematical approach for this simple problem I prefer to use the pandas approach for this, which is the get_dummies method. The outcome is much more pleasing yet completely the same.

```
pd.get_dummies(tips_df.day)
```

```
     Fri  Sat  Sun  Thur
0      0    0    1     0
1      0    0    1     0
2      0    0    1     0
3      0    0    1     0
4      0    0    1     0
..   ...  ...  ...   ...
239    0    1    0     0
240    0    1    0     0
241    0    1    0     0
242    0    1    0     0
243    0    0    0     1

[244 rows x 4 columns]
```

As an exercise you could create a script that given a specific feature (e.g. day):

- extracts that feature
- creates dummies
- concattenates it to the dataframe

Good luck!

CHAPTER

TEN

SCALING AND NORMALIZATION

In this notebook we are going to look into 2 rather mathematical concepts, Scaling and Normalization. The idea is that we can scale the values and shape the distribution of feature in our dataset.

As an example we take a dataset containing samples from a low density polyethylene production process, containing several numerical features such as temperatures, Forces, Pressure,…

The idea is that by using Scaling and normalization, the 'range of motion' for these sensors is equal and we can compare the fluxtuations not only inbetween records, but also inbetween sensors.

```
import pandas as pd
```

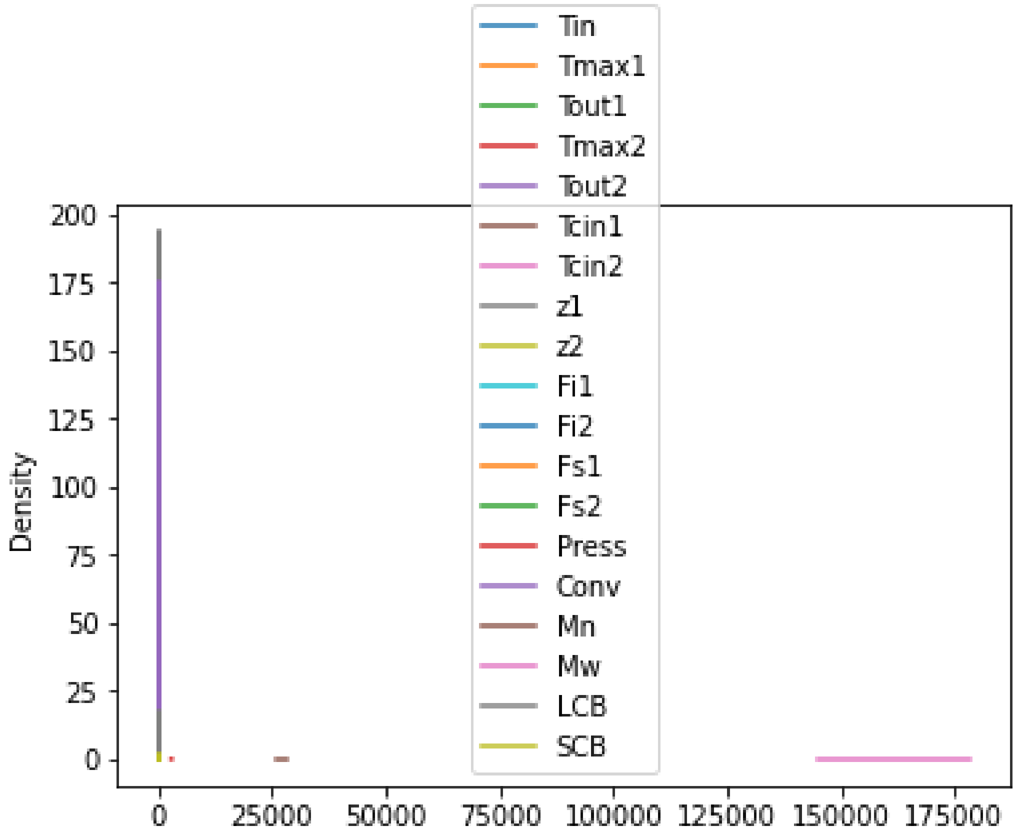
```
ldpe_df = pd.read_csv('https://openmv.net/file/LDPE.csv').drop(columns=['Unnamed: 0'])
ldpe_df.head()
```

```
      Tin   Tmax1   Tout1   Tmax2   Tout2   Tcin1   Tcin2     z1     z2  \
0  208.17  296.35  233.81  283.41  239.05  117.14  117.20  0.029  0.581
1  207.26  298.26  230.88  287.55  242.55  116.39  117.23  0.028  0.574
2  205.30  296.57  235.38  284.35  245.19  117.33  118.42  0.031  0.578
3  209.29  294.11  225.61  283.31  242.04  116.15  117.94  0.030  0.581
4  206.76  295.13  230.26  283.74  244.92  116.75  118.49  0.030  0.579

      Fi1     Fi2     Fs1     Fs2  Press    Conv     Mn      Mw    LCB    SCB
0  0.4507  0.4518  666.42  248.95   3021  0.1322  27379  160326  0.781  26.11
1  0.4765  0.5091  658.61  246.36   3033  0.1365  27043  165044  0.819  26.29
2  0.4744  0.4505  666.51  244.65   3004  0.1335  27344  165621  0.801  26.13
3  0.4429  0.4516  667.31  242.28   2980  0.1300  27502  160497  0.778  25.92
4  0.4394  0.4414  670.83  244.31   2997  0.1316  27518  165713  0.786  26.02
```

We can see that our features clearly have different ranges, but lets try to visualise these ranges using a density plot

```
ldpe_df.plot(kind='density')
```

```
<AxesSubplot:ylabel='Density'>
```

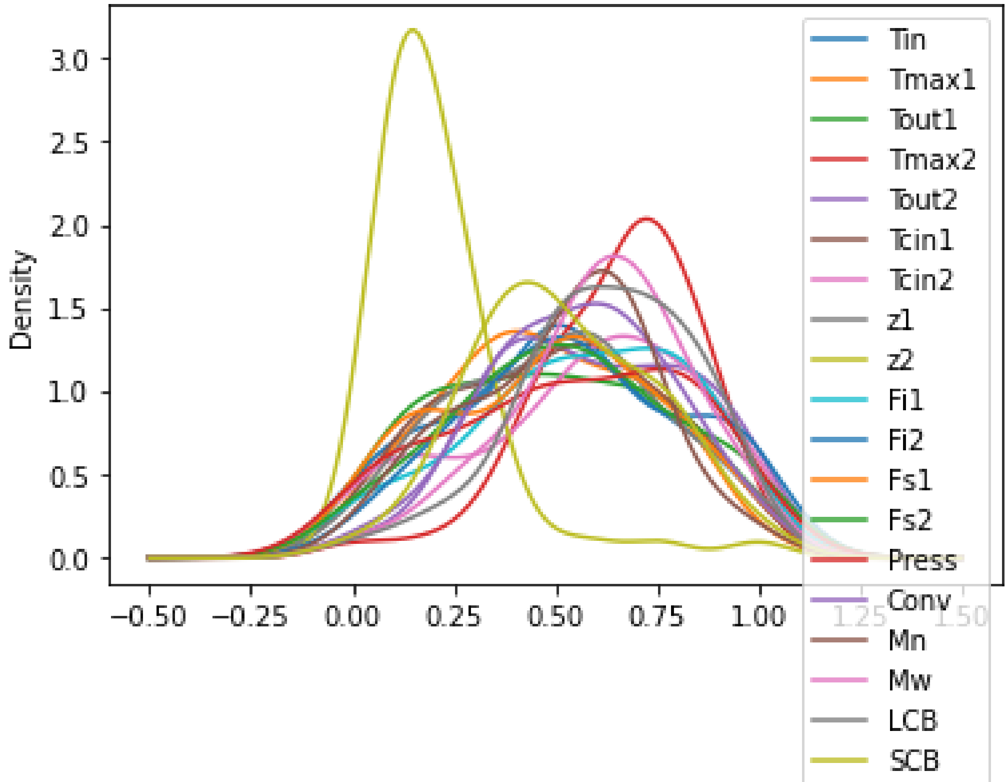


Ouch, this is clearly not working! Because the 'Mw' feature is in the range of 150k-175k our plot is so dilluted the rest are pinned to 0. We can use the sklearn library to perform a min max scaling, this scaling will shift the distribution of each feature between 0 and 1, but that can also be adjusted.

```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler()
scaler.fit(ldpe_df)
pd.DataFrame(scaler.transform(ldpe_df), columns=ldpe_df.columns).plot(kind='density')
```

```
<AxesSubplot:ylabel='Density'>
```



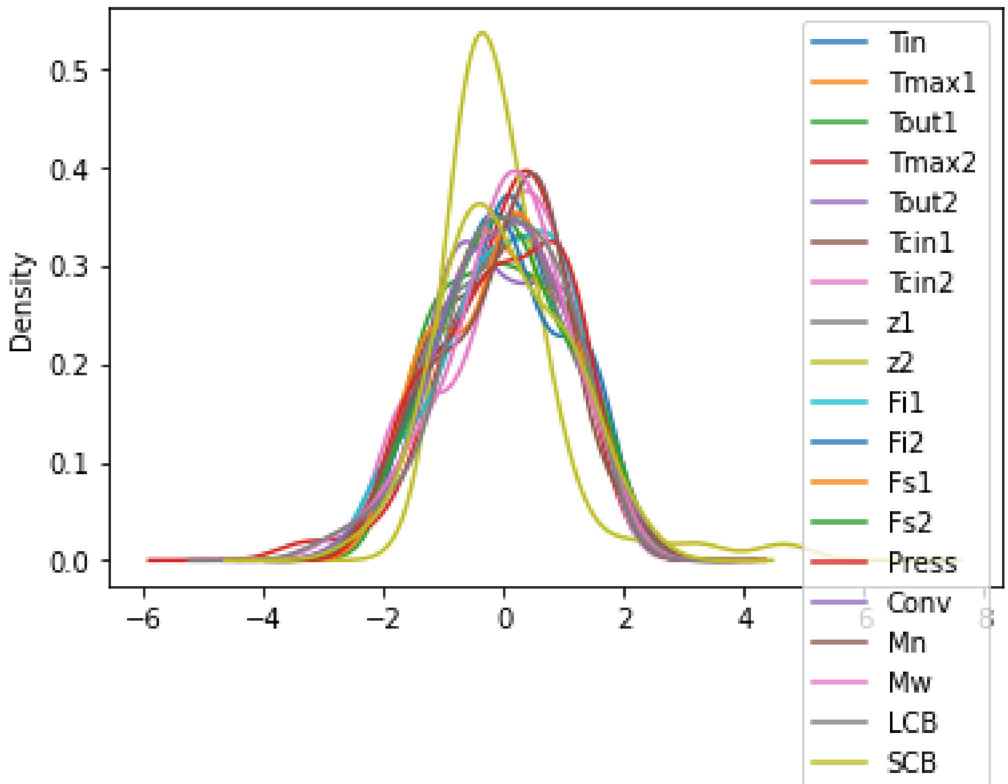
That makes a lot more sense, you can now see all of the distribution at once. Also there seems to be one (yellow) feature that has some outliers perhaps something weird is going on there…

Taking it a step further we could also alter the distributions by using a standard scaler instead of a min max scaler, redistributing the values mathematically into a normal distribution.

```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
scaler.fit(ldpe_df)
pd.DataFrame(scaler.transform(ldpe_df), columns=ldpe_df.columns).plot(kind='density')
```

```
<AxesSubplot:ylabel='Density'>
```



You can see it had some trouble fitting our special feature into the normal distribution but it did work out in the end. With this we are ready to perform machine learning algorithms on this data, but first why not try and figure out where those outliers are I mentioned earlier?

CHAPTER

ELEVEN

BINNING AND RANKING

When dealing with numerical data the trouble can sometimes be that numbers can have a wide variety.

Here we apply 2 methods to deal with that, binning and ranking. With binning we change the numerical feature into a categorical/ordinal feature. Ranking is used when our numerical feature contains a non normal distribution that fails to be normalized.

For this example we use a food consumption dataset, where european countries are listed and the relative percentage of each country is given that consumes the type of food, e.g. a value of 67 means that 67% of that country eats that type of food.

```
import pandas as pd
pd.set_option('display.max_columns', None)
```

```
food_df = pd.read_csv('https://openmv.net/file/food-consumption.csv')
food_df
```

```
        Country  Real coffee  Instant coffee  Tea  Sweetener  Biscuits  \
0       Germany           90              49   88       19.0      57.0
1         Italy           82              10   60        2.0      55.0
2        France           88              42   63        4.0      76.0
3       Holland           96              62   98       32.0      62.0
4       Belgium           94              38   48       11.0      74.0
5    Luxembourg           97              61   86       28.0      79.0
6       England           27              86   99       22.0      91.0
7      Portugal           72              26   77        2.0      22.0
8       Austria           55              31   61       15.0      29.0
9   Switzerland           73              72   85       25.0      31.0
10       Sweden           97              13   93       31.0       NaN
11      Denmark           96              17   92       35.0      66.0
12       Norway           92              17   83       13.0      62.0
13      Finland           98              12   84       20.0      64.0
14        Spain           70              40   40        NaN      62.0
15      Ireland           30              52   99       11.0      80.0

    Powder soup  Tin soup  Potatoes  Frozen fish  Frozen veggies  Apples  \
0            51        19        21           27              21      81
1            41         3         2            4               2      67
2            53        11        23           11               5      87
3            67        43         7           14              14      83
4            37        23         9           13              12      76
5            73        12         7           26              23      85
6            55        76        17           20              24      76
7            34         1         5           20               3      22
```

(continues on next page)

(continued from previous page)

```
8           33        1        5          15             11      49
9           69       10       17          19             15      79
10          43       43       39          54             45      56
11          32       17       11          51             42      81
12          51        4       17          30             15      61
13          27       10        8          18             12      50
14          43        2       14          23              7      59
15          75       18        2           5              3      57

    Oranges  Tinned fruit  Jam  Garlic  Butter  Margarine  Olive oil  Yoghurt  \
0        75            44   71      22      91         85         74     30.0
1        71             9   46      80      66         24         94      5.0
2        84            40   45      88      94         47         36     57.0
3        89            61   81      15      31         97         13     53.0
4        76            42   57      29      84         80         83     20.0
5        94            83   20      91      94         94         84     31.0
6        68            89   91      11      95         94         57     11.0
7        51             8   16      89      65         78         92      6.0
8        42            14   41      51      51         72         28     13.0
9        70            46   61      64      82         48         61     48.0
10       78            53   75       9      68         32         48      2.0
11       72            50   64      11      92         91         30     11.0
12       72            34   51      11      63         94         28      2.0
13       57            22   37      15      96         94         17      NaN
14       77            30   38      86      44         51         91     16.0
15       52            46   89       5      97         25         31      3.0

    Crisp bread
0            26
1            18
2             3
3            15
4             5
5            24
6            28
7             9
8            11
9            30
10           93
11           34
12           62
13           64
14           13
15            9
```

Here you could do some data validity, where we check if all values are between 0 and 100, or we check for missing values. I will leave that up to you

11.1 Binning

the first thing we want to do is seperate the countries based on their coffee consumption, instead of creating arbitrary values we can perform a quantitative cut. This means we create a number of equally sized groups using the qcut function, we give them the labels low, medium and high.

```
food_df['bin_coffee'] = pd.qcut(food_df['Real coffee'], q=3, labels=['low', 'medium',
↪'high'])
food_df
```

```
       Country  Real coffee  Instant coffee  Tea  Sweetener  Biscuits  \
0      Germany           90              49   88       19.0      57.0
1        Italy           82              10   60        2.0      55.0
2       France           88              42   63        4.0      76.0
3      Holland           96              62   98       32.0      62.0
4      Belgium           94              38   48       11.0      74.0
5   Luxembourg           97              61   86       28.0      79.0
6      England           27              86   99       22.0      91.0
7     Portugal           72              26   77        2.0      22.0
8      Austria           55              31   61       15.0      29.0
9  Switzerland           73              72   85       25.0      31.0
10      Sweden           97              13   93       31.0       NaN
11     Denmark           96              17   92       35.0      66.0
12      Norway           92              17   83       13.0      62.0
13     Finland           98              12   84       20.0      64.0
14       Spain           70              40   40        NaN      62.0
15     Ireland           30              52   99       11.0      80.0

   Powder soup  Tin soup  Potatoes  Frozen fish  Frozen veggies  Apples  \
0           51        19        21           27              21      81
1           41         3         2            4               2      67
2           53        11        23           11               5      87
3           67        43         7           14              14      83
4           37        23         9           13              12      76
5           73        12         7           26              23      85
6           55        76        17           20              24      76
7           34         1         5           20               3      22
8           33         1         5           15              11      49
9           69        10        17           19              15      79
10          43        43        39           54              45      56
11          32        17        11           51              42      81
12          51         4        17           30              15      61
13          27        10         8           18              12      50
14          43         2        14           23               7      59
15          75        18         2            5               3      57

   Oranges  Tinned fruit  Jam  Garlic  Butter  Margarine  Olive oil  Yoghurt  \
0       75            44   71      22      91         85         74     30.0
1       71             9   46      80      66         24         94      5.0
2       84            40   45      88      94         47         36     57.0
3       89            61   81      15      31         97         13     53.0
4       76            42   57      29      84         80         83     20.0
5       94            83   20      91      94         94         84     31.0
6       68            89   91      11      95         94         57     11.0
7       51             8   16      89      65         78         92      6.0
8       42            14   41      51      51         72         28     13.0
9       70            46   61      64      82         48         61     48.0
```

(continues on next page)

(continued from previous page)

```
10       78            53  75       9      68         32         48      2.0
11       72            50  64      11      92         91         30     11.0
12       72            34  51      11      63         94         28      2.0
13       57            22  37      15      96         94         17      NaN
14       77            30  38      86      44         51         91     16.0
15       52            46  89       5      97         25         31      3.0

    Crisp bread bin_coffee
0            26     medium
1            18     medium
2             3     medium
3            15       high
4             5     medium
5            24       high
6            28        low
7             9        low
8            11        low
9            30        low
10           93       high
11           34       high
12           62     medium
13           64       high
14           13        low
15            9        low
```

a new column has appeared at the end of our dataframe, containing the labels of our binning, countries with low coffee consumption are put in the low category and vice versa. Now we can seperate the countries with low coffee consumption from the rest

```
food_df[food_df.bin_coffee == 'low']
```

```
        Country  Real coffee  Instant coffee  Tea  Sweetener  Biscuits  \
6       England           27              86   99       22.0      91.0
7      Portugal           72              26   77        2.0      22.0
8       Austria           55              31   61       15.0      29.0
9   Switzerland           73              72   85       25.0      31.0
14        Spain           70              40   40        NaN      62.0
15      Ireland           30              52   99       11.0      80.0

    Powder soup  Tin soup  Potatoes  Frozen fish  Frozen veggies  Apples  \
6            55        76        17           20              24      76
7            34         1         5           20               3      22
8            33         1         5           15              11      49
9            69        10        17           19              15      79
14           43         2        14           23               7      59
15           75        18         2            5               3      57

    Oranges  Tinned fruit  Jam  Garlic  Butter  Margarine  Olive oil  Yoghurt  \
6        68            89   91      11      95         94         57     11.0
7        51             8   16      89      65         78         92      6.0
8        42            14   41      51      51         72         28     13.0
9        70            46   61      64      82         48         61     48.0
14       77            30   38      86      44         51         91     16.0
15       52            46   89       5      97         25         31      3.0

    Crisp bread bin_coffee
```

(continues on next page)

(continued from previous page)

```
6             28        low
7              9        low
8             11        low
9             30        low
14            13        low
15             9        low
```

You can already see the England/Ireland stereotype here, note that those are the only 2 with really low coffee consumption, the others are only in this low binning because we requested equally spaced bins in our qcut function. using the cut function would result in a different outcome. Perhaps you could try that out?

I tried to think of some metric to quantify the status of coffee drinkers, since we also have the instant coffee consumption we could create a metric where we subtract the amount of instant coffe drinkers from the amount of real coffee drinkers. This way we can measure that difference between them, I already went ahead and made equal quantity bins for them with labels low, medium and high 'quality coffee'.

```
food_df['bin_qual_coffee'] = pd.qcut(food_df['Real coffee'] - food_df['Instant coffee
↪'], q=3, labels=['low', 'medium', 'high'])
```

```
food_df[food_df.bin_qual_coffee=='high']
```

```
    Country  Real coffee  Instant coffee  Tea  Sweetener  Biscuits  \
1     Italy           82              10   60        2.0      55.0
10   Sweden           97              13   93       31.0       NaN
11  Denmark           96              17   92       35.0      66.0
12   Norway           92              17   83       13.0      62.0
13  Finland           98              12   84       20.0      64.0

    Powder soup  Tin soup  Potatoes  Frozen fish  Frozen veggies  Apples  \
1            41         3         2            4               2      67
10           43        43        39           54              45      56
11           32        17        11           51              42      81
12           51         4        17           30              15      61
13           27        10         8           18              12      50

    Oranges  Tinned fruit  Jam  Garlic  Butter  Margarine  Olive oil  Yoghurt  \
1        71             9   46      80      66         24         94      5.0
10       78            53   75       9      68         32         48      2.0
11       72            50   64      11      92         91         30     11.0
12       72            34   51      11      63         94         28      2.0
13       57            22   37      15      96         94         17      NaN

    Crisp bread bin_coffee bin_qual_coffee
1            18     medium            high
10           93       high            high
11           34       high            high
12           62     medium            high
13           64       high            high
```

Aha! you can see here which countries prefer the real coffee over the instant version. It seems the scandinavian countries together with obviously Italy are the true Caffeine connoisseur of Europe. Another intersting thing we can do now is take the mean for each product for both group high and low and take the difference for high - low. We can see the result below

```
food_df[food_df.bin_qual_coffee=='high'].mean()-food_df[food_df.bin_qual_coffee=='low
↪'].mean()
```

```
/tmp/ipykernel_16521/3908782487.py:1: FutureWarning: Dropping of nuisance columns in␣
↪DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version␣
↪this will raise TypeError.  Select only valid columns before calling the reduction.
  food_df[food_df.bin_qual_coffee=='high'].mean()-food_df[food_df.bin_qual_coffee==
↪'low'].mean()
```

```
Real coffee       34.500000
Instant coffee   -43.366667
Tea                2.066667
Sweetener         -0.800000
Biscuits           2.583333
Powder soup      -18.200000
Tin soup          -9.600000
Potatoes           5.066667
Frozen fish       15.400000
Frozen veggies    10.866667
Apples            -4.166667
Oranges            3.666667
Tinned fruit     -14.066667
Jam              -12.233333
Garlic           -13.466667
Butter            10.333333
Margarine          2.500000
Olive oil         -3.433333
Yoghurt          -19.000000
Crisp bread       36.533333
dtype: float64
```

It seems a preference for quality coffee also pairs with crisp bread, who knew? Do you think scaling/normalization might be interesting here? why (not)?

11.2 Ranking

In case normalization fails us and we are for some reason not able to get a normal distribution out of a feature, we can still resort to ranking. Note that non linear machine learning techniques often use a ranking functionality under the hood, therefore this technique is often not required, yet for educational purposes we are going to use it here anyway. Let's see how the distribution for Real coffee consumption looks like.

```
food_df.sort_values('Real coffee')
```

```
        Country  Real coffee  Instant coffee  Tea  Sweetener  Biscuits  \
6       England           27              86   99       22.0      91.0
15      Ireland           30              52   99       11.0      80.0
8       Austria           55              31   61       15.0      29.0
14        Spain           70              40   40        NaN      62.0
7      Portugal           72              26   77        2.0      22.0
9   Switzerland           73              72   85       25.0      31.0
1         Italy           82              10   60        2.0      55.0
2        France           88              42   63        4.0      76.0
0       Germany           90              49   88       19.0      57.0
12       Norway           92              17   83       13.0      62.0
4       Belgium           94              38   48       11.0      74.0
3       Holland           96              62   98       32.0      62.0
11      Denmark           96              17   92       35.0      66.0
```

(continues on next page)

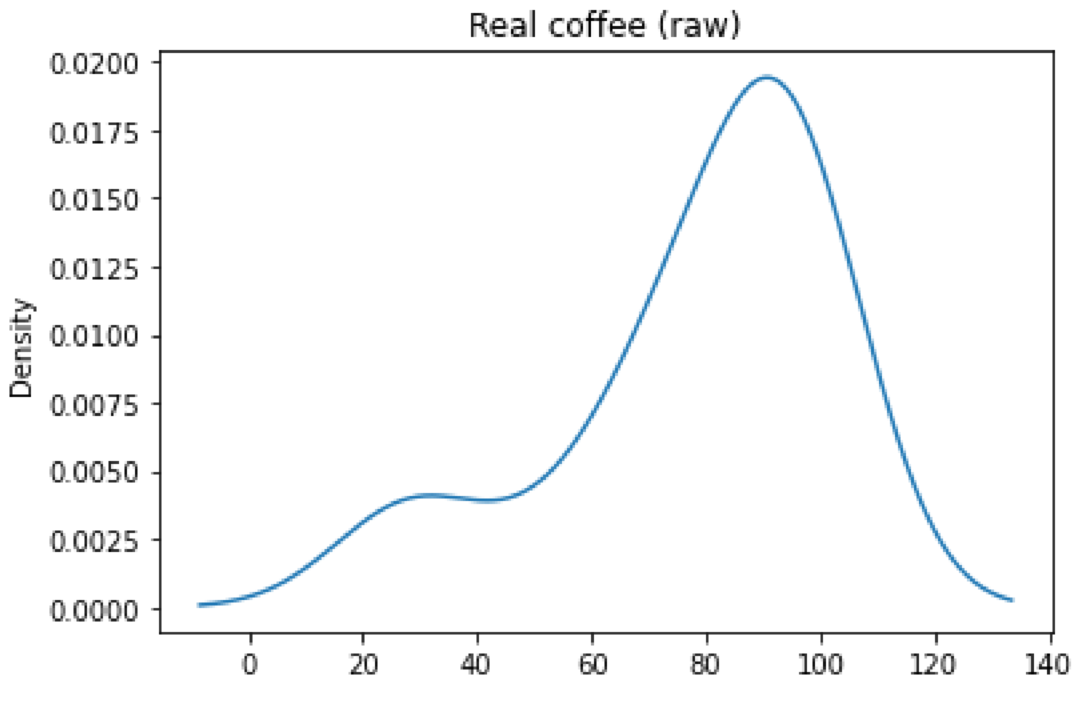
(continued from previous page)

```
5    Luxembourg            97              61   86       28.0      79.0
10       Sweden            97              13   93       31.0       NaN
13      Finland            98              12   84       20.0      64.0

    Powder soup  Tin soup  Potatoes  Frozen fish  Frozen veggies  Apples  \
6            55        76        17           20              24      76
15           75        18         2            5               3      57
8            33         1         5           15              11      49
14           43         2        14           23               7      59
7            34         1         5           20               3      22
9            69        10        17           19              15      79
1            41         3         2            4               2      67
2            53        11        23           11               5      87
0            51        19        21           27              21      81
12           51         4        17           30              15      61
4            37        23         9           13              12      76
3            67        43         7           14              14      83
11           32        17        11           51              42      81
5            73        12         7           26              23      85
10           43        43        39           54              45      56
13           27        10         8           18              12      50

    Oranges  Tinned fruit  Jam  Garlic  Butter  Margarine  Olive oil  Yoghurt  \
6        68            89   91      11      95         94         57     11.0
15       52            46   89       5      97         25         31      3.0
8        42            14   41      51      51         72         28     13.0
14       77            30   38      86      44         51         91     16.0
7        51             8   16      89      65         78         92      6.0
9        70            46   61      64      82         48         61     48.0
1        71             9   46      80      66         24         94      5.0
2        84            40   45      88      94         47         36     57.0
0        75            44   71      22      91         85         74     30.0
12       72            34   51      11      63         94         28      2.0
4        76            42   57      29      84         80         83     20.0
3        89            61   81      15      31         97         13     53.0
11       72            50   64      11      92         91         30     11.0
5        94            83   20      91      94         94         84     31.0
10       78            53   75       9      68         32         48      2.0
13       57            22   37      15      96         94         17      NaN

    Crisp bread bin_coffee bin_qual_coffee
6            28        low             low
15            9        low             low
8            11        low             low
14           13        low             low
7             9        low          medium
9            30        low             low
1            18     medium            high
2             3     medium          medium
0            26     medium          medium
12           62     medium            high
4             5     medium          medium
3            15       high             low
11           34       high            high
5            24       high          medium
10           93       high            high
13           64       high            high
```

Ah yes, about half of the values are 90 or higher, not really optimal as the range is between 0 and 100! We can also view this in a visual way using a density plot.

```
food_df['Real coffee'].plot(kind='density', title='Real coffee (raw)')
```

```
<AxesSubplot:title={'center':'Real coffee (raw)'}, ylabel='Density'>
```



For larger datasets this would be more useful as we cannot see our whole dataset, it is clear we have to do something about this, now imagine we can not use regular normalization techniques. The rank method now comes in handy!

```
food_df['rank_coffee'] = food_df['Real coffee'].rank()
food_df
```

```
        Country  Real coffee  Instant coffee  Tea  Sweetener  Biscuits  \
0       Germany           90              49   88       19.0      57.0
1         Italy           82              10   60        2.0      55.0
2        France           88              42   63        4.0      76.0
3       Holland           96              62   98       32.0      62.0
4       Belgium           94              38   48       11.0      74.0
5    Luxembourg           97              61   86       28.0      79.0
6       England           27              86   99       22.0      91.0
7      Portugal           72              26   77        2.0      22.0
8       Austria           55              31   61       15.0      29.0
9   Switzerland           73              72   85       25.0      31.0
10       Sweden           97              13   93       31.0       NaN
11      Denmark           96              17   92       35.0      66.0
12       Norway           92              17   83       13.0      62.0
13      Finland           98              12   84       20.0      64.0
14        Spain           70              40   40        NaN      62.0
15      Ireland           30              52   99       11.0      80.0

    Powder soup  Tin soup  Potatoes  Frozen fish  Frozen veggies  Apples  \
```

(continues on next page)

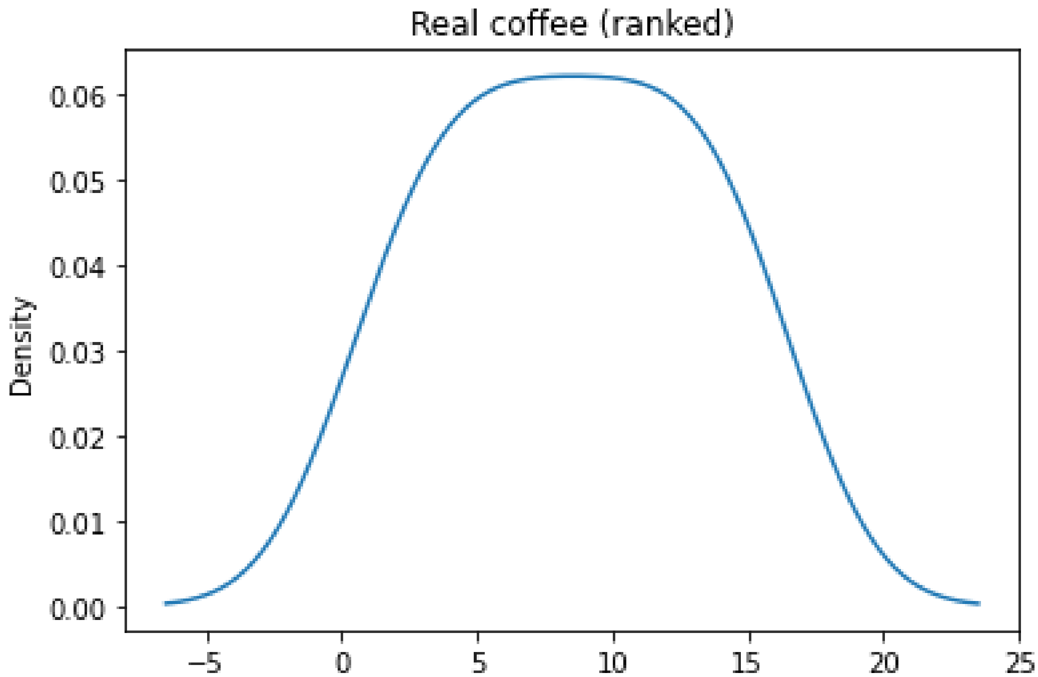
(continued from previous page)

```
0            51        19        21          27              21      81
1            41         3         2           4               2      67
2            53        11        23          11               5      87
3            67        43         7          14              14      83
4            37        23         9          13              12      76
5            73        12         7          26              23      85
6            55        76        17          20              24      76
7            34         1         5          20               3      22
8            33         1         5          15              11      49
9            69        10        17          19              15      79
10           43        43        39          54              45      56
11           32        17        11          51              42      81
12           51         4        17          30              15      61
13           27        10         8          18              12      50
14           43         2        14          23               7      59
15           75        18         2           5               3      57

    Oranges  Tinned fruit  Jam  Garlic  Butter  Margarine  Olive oil  Yoghurt  \
0        75            44   71      22      91         85         74     30.0
1        71             9   46      80      66         24         94      5.0
2        84            40   45      88      94         47         36     57.0
3        89            61   81      15      31         97         13     53.0
4        76            42   57      29      84         80         83     20.0
5        94            83   20      91      94         94         84     31.0
6        68            89   91      11      95         94         57     11.0
7        51             8   16      89      65         78         92      6.0
8        42            14   41      51      51         72         28     13.0
9        70            46   61      64      82         48         61     48.0
10       78            53   75       9      68         32         48      2.0
11       72            50   64      11      92         91         30     11.0
12       72            34   51      11      63         94         28      2.0
13       57            22   37      15      96         94         17      NaN
14       77            30   38      86      44         51         91     16.0
15       52            46   89       5      97         25         31      3.0

    Crisp bread bin_coffee bin_qual_coffee  rank_coffee
0            26     medium          medium          9.0
1            18     medium            high          7.0
2             3     medium          medium          8.0
3            15       high             low         12.5
4             5     medium          medium         11.0
5            24       high          medium         14.5
6            28        low             low          1.0
7             9        low          medium          5.0
8            11        low             low          3.0
9            30        low             low          6.0
10           93       high            high         14.5
11           34       high            high         12.5
12           62     medium            high         10.0
13           64       high            high         16.0
14           13        low             low          4.0
15            9        low             low          2.0
```

At the end of our data a new column was appended, containing the ranking of each country with the lowest being 1 and the highest equal to the amount of countries. When we visualise this distribution we get a uniform distribution, not normal but still better than before!

```
food_df['rank_coffee'].plot(kind='density', title='Real coffee (ranked)')
```

```
<AxesSubplot:title={'center':'Real coffee (ranked)'}, ylabel='Density'>
```



CHAPTER TWELVE

SOME PRACTICE

Now that you have learned techniques in data preparation, why don't you put them to use in this wonderfully horrifying dataset. Good luck!

```
import os
import json

import pandas as pd
```

```
kaggle_dir = os.path.expanduser("~/.kaggle")
if not os.path.exists(kaggle_dir):
    os.mkdir(kaggle_dir)

with open(f'{kaggle_dir}/kaggle.json', 'w') as f:
    json.dump(
        {
            "username":"lorenzf",
            "key":"7a44a9e99b27e796177d793a3d85b8cf"
        }
        , f)
```

```
import kaggle
kaggle.api.dataset_download_files(dataset='PromptCloudHQ/us-jobs-on-monstercom', path=
↪'./data', unzip=True)
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
/tmp/ipykernel_25600/39646943.py in <module>
----> 1 import kaggle
      2 kaggle.api.dataset_download_files(dataset='PromptCloudHQ/us-jobs-on-monstercom
↪', path='./data', unzip=True)

ModuleNotFoundError: No module named 'kaggle'
```

```
df = pd.read_csv('./data/monster_com-job_sample.csv')
```

```
df.head()
```

```
                    country country_code date_added has_expired  \
0  United States of America           US        NaN          No
1  United States of America           US        NaN          No
2  United States of America           US        NaN          No
```

(continues on next page)

(continued from previous page)

```
3  United States of America           US        NaN          No
4  United States of America           US        NaN          No

          job_board                                    job_description  \
0  jobs.monster.com  TeamSoft is seeing an IT Support Specialist to...
1  jobs.monster.com  The Wisconsin State Journal is seeking a flexi...
2  jobs.monster.com  Report this job About the Job DePuy Synthes Co...
3  jobs.monster.com  Why Join Altec? If you're considering a career...
4  jobs.monster.com  Position ID#  76162 # Positions  1 State  CT C...

                                           job_title             job_type  \
0            IT Support Technician Job in Madison   Full Time Employee
1         Business Reporter/Editor Job in Madison            Full Time
2  Johnson & Johnson Family of Companies Job Appl...  Full Time, Employee
3                   Engineer - Quality Job in Dixon            Full Time
4      Shift Supervisor - Part-Time Job in Camphill   Full Time Employee

                                            location  \
0                                  Madison, WI 53702
1                                  Madison, WI 53708
2  DePuy Synthes Companies is a member of Johnson...
3                                          Dixon, CA
4                                       Camphill, PA

                     organization  \
0                             NaN
1         Printing and Publishing
2  Personal and Household Services
3                 Altec Industries
4                          Retail

                                            page_url salary  \
0  http://jobview.monster.com/it-support-technici...    NaN
1  http://jobview.monster.com/business-reporter-e...    NaN
2  http://jobview.monster.com/senior-training-lea...    NaN
3  http://jobview.monster.com/engineer-quality-jo...    NaN
4  http://jobview.monster.com/shift-supervisor-pa...    NaN

                      sector                           uniq_id
0    IT/Software Development  11d599f229a80023d2f40e7c52cd941e
1                        NaN  e4cbb126dabf22159aff90223243ff2a
2                        NaN  839106b353877fa3d896ffb9c1fe01c0
3   Experienced (Non-Manager)  58435fcab804439efdcaa7ecca0fd783
4  Project/Program Management  64d0272dc8496abfd9523a8df63c184c
```

Need some inspiration? perhaps this might help!

Part III

3. Data Preprocessing

CHAPTER

THIRTEEN

DATA PREPROCESSING

this is an introduction

CHAPTER FOURTEEN

INDEXING AND SLICING

In

```
import pandas as pd
```

```
min_temp_df = pd.read_csv('https://raw.githubusercontent.com/LorenzF/data-science-
↪practical-approach/main/src/c2_data_preparation/data/temperatures/australia/
↪melbourne/1981.csv')
min_temp_df
```

```
           Date  Temp
0    1981-01-01  20.7
1    1981-01-02  17.9
2    1981-01-03  18.8
3    1981-01-04  14.6
4    1981-01-05  15.8
..          ...   ...
360  1981-12-27  15.5
361  1981-12-28  13.3
362  1981-12-29  15.6
363  1981-12-30  15.2
364  1981-12-31  17.4

[365 rows x 2 columns]
```

```
min_temp_df.Date = pd.to_datetime(min_temp_df.Date)
```

```
min_temp_df = min_temp_df.set_index('Date')
```

```
min_temp_df.loc['1981-06-01':'1981-06-30']
```

```
            Temp
Date
1981-06-01  11.6
1981-06-02  10.6
1981-06-03   9.8
1981-06-04  11.2
1981-06-05   5.7
1981-06-06   7.1
1981-06-07   2.5
1981-06-08   3.5
1981-06-09   4.6
```

(continues on next page)

(continued from previous page)

```
1981-06-10  11.0
1981-06-11   5.7
1981-06-12   7.7
1981-06-13  10.4
1981-06-14  11.4
1981-06-15   9.2
1981-06-16   6.1
1981-06-17   2.7
1981-06-18   4.3
1981-06-19   6.3
1981-06-20   3.8
1981-06-21   4.4
1981-06-22   7.1
1981-06-23   4.8
1981-06-24   5.8
1981-06-25   6.2
1981-06-26   7.3
1981-06-27   9.2
1981-06-28  10.2
1981-06-29   9.5
1981-06-30   9.5
```

```
min_temp_df.loc['1989-06-01':'1989-06-30'].mean()
```

```
Temp   NaN
dtype: float64
```

```
min_temp_df.resample('MS').mean()
```

```
                 Temp
Date
1981-01-01  17.712903
1981-02-01  17.678571
1981-03-01  13.500000
1981-04-01  12.356667
1981-05-01   9.490323
1981-06-01   7.306667
1981-07-01   7.577419
1981-08-01   7.238710
1981-09-01  10.143333
1981-10-01  10.087097
1981-11-01  11.890000
1981-12-01  13.680645
```

```
import seaborn as sns
```

```
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
```

(continues on next page)

(continued from previous page)

```
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_25625/2537502835.py:1: PerformanceWarning: indexing past lexsort depth␣
↪may impact performance.
  tip_index_df.loc[('Thur','Lunch')].tip.mean()
```

```
2.767704918032786
```

```
pd.pivot_table(tip_df, values='total_bill', index='day', columns='time', aggfunc=
↪'median')
```

```
time  Lunch  Dinner
day
Thur  16.00  18.780
Fri   13.42  18.665
Sat     NaN  18.240
Sun     NaN  19.630
```

```
tip_df.set_index(['sex', 'time','smoker']).loc[('Male', 'Dinner','Yes')]['tip'].mean()
```

```
/tmp/ipykernel_25625/3467525553.py:1: PerformanceWarning: indexing past lexsort depth␣
↪may impact performance.
  tip_df.set_index(['sex', 'time','smoker']).loc[('Male', 'Dinner','Yes')]['tip'].
↪mean()
```

```
3.123191489361702
```

CHAPTER

FIFTEEN

MERGE

https://www.kaggle.com/uciml/restaurant-data-with-consumer-ratings

```
import pandas as pd
```

```
rating_df = pd.read_csv('./data/cuisine/rating_final.csv')
```

```
rating_df
```

```
     userID  placeID  rating  food_rating  service_rating
0     U1077   135085       2            2               2
1     U1077   135038       2            2               1
2     U1077   132825       2            2               2
3     U1077   135060       1            2               2
4     U1068   135104       1            1               2
...     ...      ...     ...          ...             ...
1156  U1043   132630       1            1               1
1157  U1011   132715       1            1               0
1158  U1068   132733       1            1               0
1159  U1068   132594       1            1               1
1160  U1068   132660       0            0               0

[1161 rows x 5 columns]
```

```
cuisine_df = pd.read_csv('./data/cuisine/chefmozcuisine.csv')
cuisine_df
```

```
     placeID        Rcuisine
0     135110         Spanish
1     135109         Italian
2     135107  Latin_American
3     135106         Mexican
4     135105       Fast_Food
..       ...             ...
911   132005         Seafood
912   132004         Seafood
913   132003   International
914   132002         Seafood
915   132001    Dutch-Belgian

[916 rows x 2 columns]
```

```
merged_df = pd.merge(rating_df, cuisine_df, on='placeID', how='inner')
merged_df
```

```
     userID  placeID  rating  food_rating  service_rating   Rcuisine
0     U1077   135085       2            2               2  Fast_Food
1     U1108   135085       1            2               1  Fast_Food
2     U1081   135085       1            2               1  Fast_Food
3     U1056   135085       2            2               2  Fast_Food
4     U1134   135085       2            1               2  Fast_Food
...     ...      ...     ...          ...             ...        ...
1038  U1061   132958       2            2               2   American
1039  U1025   132958       1            0               0   American
1040  U1097   132958       2            1               1   American
1041  U1096   132958       1            2               2   American
1042  U1136   132958       2            2               2   American

[1043 rows x 6 columns]
```

```
merged_df[merged_df.Rcuisine.isna()]
```

```
Empty DataFrame
Columns: [userID, placeID, rating, food_rating, service_rating, Rcuisine]
Index: []
```

```
merged_df.Rcuisine.unique()
```

```
array(['Fast_Food', 'Mexican', 'Seafood', 'Pizzeria', 'Regional',
       'Armenian', 'Cafeteria', 'American', 'Bar', 'Bar_Pub_Brewery',
       'Italian', 'Japanese', 'Contemporary', 'Burgers', 'Mediterranean',
       'International', 'Vietnamese', 'Family', 'Chinese',
       'Cafe-Coffee_Shop', 'Game', 'Bakery', 'Breakfast-Brunch'],
      dtype=object)
```

```
merged_df[merged_df.Rcuisine=='Bar_Pub_Brewery'][['rating', 'food_rating', 'service_
↪rating']].mean()
```

```
rating            1.305085
food_rating       1.169492
service_rating    1.203390
dtype: float64
```

```
merged_df[merged_df.Rcuisine=='Bar'][['rating', 'food_rating', 'service_rating']].
↪mean()
```

```
rating            1.200000
food_rating       1.135714
service_rating    1.085714
dtype: float64
```

```
merged_df[merged_df.Rcuisine=='Cafeteria'][['rating', 'food_rating', 'service_rating
↪']].mean()
```

```
rating            1.205882
food_rating       1.127451
service_rating    1.078431
dtype: float64
```

```
merged_df[merged_df.Rcuisine=='Cafe-Coffee_Shop'][['rating', 'food_rating', 'service_
↪rating']].mean()
```

```
rating            1.583333
food_rating       1.333333
service_rating    1.416667
dtype: float64
```

```
user_payment_df = pd.read_csv('./data/cuisine/userpayment.csv')
```

```
payment_df = pd.merge(rating_df, user_payment_df, how='inner')
```

```
payment_df.Upayment.unique()
```

```
array(['VISA', 'cash', 'bank_debit_cards', 'MasterCard-Eurocard',
       'American_Express'], dtype=object)
```

```
payment_df[payment_df.Upayment=='cash'][['rating', 'food_rating', 'service_rating']].
↪mean()
```

```
rating            1.182815
food_rating       1.200183
service_rating    1.080439
dtype: float64
```

```
payment_df[payment_df.Upayment=='bank_debit_cards'][['rating', 'food_rating',
↪'service_rating']].mean()
```

```
rating            1.437908
food_rating       1.562092
service_rating    1.398693
dtype: float64
```

CHAPTER

SIXTEEN

GROUPBY

```
import pandas as pd
```

```
rating_df = pd.read_csv('./data/cuisine/rating_final.csv')
```

```
grouped_rating_df = rating_df.groupby('placeID').mean().sort_values('rating')
grouped_rating_df
```

```
           rating  food_rating  service_rating
placeID
132654   0.250000         0.25        0.250000
135040   0.250000         0.25        0.250000
132560   0.500000         1.00        0.250000
132663   0.500000         0.50        0.666667
135069   0.500000         0.50        0.750000
...           ...          ...             ...
132755   1.800000         2.00        1.600000
132922   1.833333         1.50        1.833333
134986   2.000000         2.00        2.000000
135034   2.000000         2.00        1.600000
132955   2.000000         1.80        1.800000

[130 rows x 3 columns]
```

```
geo_df = pd.read_csv('./data/cuisine/geoplaces2.csv').set_index('placeID')
```

```
merged_rating_df = pd.merge(grouped_rating_df, geo_df, left_index=True, right_
↪index=True)
merged_rating_df
```

```
           rating  food_rating  service_rating   latitude   longitude  \
placeID
132654   0.250000         0.25        0.250000  23.735523  -99.129588
135040   0.250000         0.25        0.250000  22.135617 -100.969709
132560   0.500000         1.00        0.250000  23.752304  -99.166913
132663   0.500000         0.50        0.666667  23.752511  -99.166954
135069   0.500000         0.50        0.750000  22.140129 -100.944872
...           ...          ...             ...        ...         ...
132755   1.800000         2.00        1.600000  22.153324 -101.019546
132922   1.833333         1.50        1.833333  22.151135 -100.982311
134986   2.000000         2.00        2.000000  18.928798  -99.239513
135034   2.000000         2.00        1.600000  22.140517 -101.021422
```

(continues on next page)

(continued from previous page)

```
132955   2.000000         1.80        1.800000  22.147622 -101.010275

                                   the_geom_meter  \
placeID
132654   0101000020957F000040E8F628488557C18224E8B94845...
135040   0101000020957F00001B552189B84A58C15A2AAEFD2CA2...
132560   0101000020957F0000FC60BDA8E88157C1B2C357D6DA4E...
132663   0101000020957F0000FDF8D26EE08157C1FEDB6A1FDB4E...
135069   0101000020957F000038E5D546B74A58C18FD29AD0D29A...
...                                                 ...
132755   0101000020957F000026CADE45A14658C1F011EBCA55AF...
132922   0101000020957F000060A98A38FF4758C146718E41D9A4...
134986   0101000020957F00002A0D05E2D96D5AC1AB058CB1EC56...
135034   0101000020957F000026D92BB4894858C161A7552DA2B0...
132955   0101000020957F000068BE7C87C24758C1920A360A08AD...

                                 name  \
placeID
132654   Carnitas Mata  Calle 16 de Septiembre
135040                Restaurant los Compadres
132560                      puesto de gorditas
132663                               tacos abi
135069                 Abondance Restaurante Bar
...                                        ...
132755                     La Estrella de Dimas
132922                     cafe punta del cielo
134986                 Restaurant Las Mananitas
135034                Michiko Restaurant Japones
132955                                emilianos

                                        address             city  \
placeID
132654                          16 de Septiembre        victoria
135040              Camino a Simon Diaz 155 Centro  San Luis Potosi
132560                        frente al tecnologico         victoria
132663                                        ?         victoria
135069                 Industrias 908 Valle Dorado  San Luis Potosi
...                                             ...              ...
132755                         Av. de los Pintores  San Luis Potosi
132922                                        ?                ?
134986                         Ricardo Linares 107       Cuernavaca
135034   Cordillera de Los Alpes 160 Lomas 2 Seccion  San Luis Potosi
132955                         venustiano carranza   san luis potos

             state  ...            alcohol smoking_area dress_code  \
placeID             ...
132654   tamaulipas  ...  No_Alcohol_Served         none   informal
135040          SLP  ...          Wine-Beer         none   informal
132560   tamaulipas  ...  No_Alcohol_Served    permitted   informal
132663   tamaulipas  ...  No_Alcohol_Served         none   informal
135069          SLP  ...          Wine-Beer         none   informal
...             ...  ...                ...          ...        ...
132755       S.L.P.  ...  No_Alcohol_Served         none   informal
132922            ?  ...  No_Alcohol_Served    permitted     formal
134986      Morelos  ...          Wine-Beer         none     formal
135034          SLP  ...  No_Alcohol_Served         none   informal
132955       mexico  ...          Wine-Beer         none   informal
```

(continues on next page)

(continued from previous page)

```
           accessibility   price                  url Rambience franchise  \
placeID
132654        completely     low                    ?  familiar         f
135040   no_accessibility    high                    ?  familiar         f
132560   no_accessibility     low                    ?  familiar         f
132663        completely     low                    ?  familiar         f
135069   no_accessibility     low                    ?  familiar         f
...                  ...     ...                  ...       ...       ...
132755         partially  medium                    ?  familiar         f
132922        completely  medium                    ?  familiar         f
134986   no_accessibility    high  lasmananitas.com.mx  familiar         f
135034   no_accessibility  medium                    ?  familiar         f
132955        completely     low                    ?  familiar         t

          area other_services
placeID
132654   closed           none
135040   closed           none
132560     open           none
132663   closed           none
135069   closed           none
...         ...            ...
132755   closed        variety
132922   closed           none
134986   closed           none
135034   closed           none
132955   closed        variety

[130 rows x 23 columns]
```

```
merged_rating_df.columns
```

```
Index(['rating', 'food_rating', 'service_rating', 'latitude', 'longitude',
       'the_geom_meter', 'name', 'address', 'city', 'state', 'country', 'fax',
       'zip', 'alcohol', 'smoking_area', 'dress_code', 'accessibility',
       'price', 'url', 'Rambience', 'franchise', 'area', 'other_services'],
      dtype='object')
```

```
merged_rating_df.groupby('country')[['rating', 'food_rating', 'service_rating']].
↪mean()
```

```
           rating  food_rating  service_rating
country
?        1.166045     1.232946        1.069169
Mexico   1.200977     1.229093        1.118162
mexico   1.062660     1.069006        0.900064
```

```
can you fix this string problem?
```

```
Object `problem` not found.
```

```
merged_rating_df.groupby('alcohol')[['rating', 'food_rating', 'service_rating']].
↪mean()
```

```
                     rating  food_rating  service_rating
alcohol
Full_Bar           1.287124     1.218315        1.170311
No_Alcohol_Served  1.148075     1.194730        1.042417
Wine-Beer          1.231887     1.261840        1.174437
```

```
merged_rating_df.groupby('accessibility')[['rating', 'food_rating', 'service_rating
↪']].mean()
```

```
                    rating  food_rating  service_rating
accessibility
completely        1.132494     1.203597        1.049709
no_accessibility  1.196189     1.206242        1.091278
partially         1.275356     1.330294        1.219991
```

```
merged_rating_df.accessibility.value_counts()
```

```
no_accessibility    76
completely          45
partially            9
Name: accessibility, dtype: int64
```

```
merged_rating_df.groupby('price')[['rating', 'food_rating', 'service_rating']].mean()
```

```
          rating  food_rating  service_rating
price
high    1.258106     1.253816        1.174754
low     1.063059     1.135805        0.935632
medium  1.234342     1.255871        1.161361
```

```
can you solve the mean-mean problem?
```

```
Object `problem` not found.
```

```
in the merge example we added the cuisine type, could you perform a groupby analysis␣
↪on this?
```

```
Object `this` not found.
```

CHAPTER

SEVENTEEN

PIVOT

```
import pandas as pd
```

```
rating_df = pd.read_csv('./data/cuisine/rating_final.csv')
```

```
rating_df
```

```
     userID  placeID  rating  food_rating  service_rating
0     U1077   135085       2            2               2
1     U1077   135038       2            2               1
2     U1077   132825       2            2               2
3     U1077   135060       1            2               2
4     U1068   135104       1            1               2
...     ...      ...     ...          ...             ...
1156  U1043   132630       1            1               1
1157  U1011   132715       1            1               0
1158  U1068   132733       1            1               0
1159  U1068   132594       1            1               1
1160  U1068   132660       0            0               0

[1161 rows x 5 columns]
```

```
geo_df = pd.read_csv('./data/cuisine/geoplaces2.csv').set_index('placeID')
```

```
merged_rating_df = pd.merge(rating_df, geo_df, on='placeID')
merged_rating_df
```

```
     userID  placeID  rating  food_rating  service_rating   latitude  \
0     U1077   135085       2            2               2  22.150802
1     U1108   135085       1            2               1  22.150802
2     U1081   135085       1            2               1  22.150802
3     U1056   135085       2            2               2  22.150802
4     U1134   135085       2            1               2  22.150802
...     ...      ...     ...          ...             ...        ...
1156  U1061   132958       2            2               2  22.144979
1157  U1025   132958       1            0               0  22.144979
1158  U1097   132958       2            1               1  22.144979
1159  U1096   132958       1            2               2  22.144979
1160  U1136   132958       2            2               2  22.144979

       longitude                                      the_geom_meter  \
0    -100.982680  0101000020957F00009F823DA6094858C18A2D4D37F9A4...
```

(continues on next page)

(continued from previous page)

```
1    -100.982680  0101000020957F00009F823DA6094858C18A2D4D37F9A4...
2    -100.982680  0101000020957F00009F823DA6094858C18A2D4D37F9A4...
3    -100.982680  0101000020957F00009F823DA6094858C18A2D4D37F9A4...
4    -100.982680  0101000020957F00009F823DA6094858C18A2D4D37F9A4...
...         ...                                                ...
1156 -101.005683  0101000020957F000049095EB34A4858C15CB4BD1EE1AB...
1157 -101.005683  0101000020957F000049095EB34A4858C15CB4BD1EE1AB...
1158 -101.005683  0101000020957F000049095EB34A4858C15CB4BD1EE1AB...
1159 -101.005683  0101000020957F000049095EB34A4858C15CB4BD1EE1AB...
1160 -101.005683  0101000020957F000049095EB34A4858C15CB4BD1EE1AB...

                        name                         address  ...  \
0     Tortas Locas Hipocampo  Venustiano Carranza 719 Centro  ...
1     Tortas Locas Hipocampo  Venustiano Carranza 719 Centro  ...
2     Tortas Locas Hipocampo  Venustiano Carranza 719 Centro  ...
3     Tortas Locas Hipocampo  Venustiano Carranza 719 Centro  ...
4     Tortas Locas Hipocampo  Venustiano Carranza 719 Centro  ...
...                      ...                             ...  ...
1156      tacos los volcanes          avenida hivno nacional  ...
1157      tacos los volcanes          avenida hivno nacional  ...
1158      tacos los volcanes          avenida hivno nacional  ...
1159      tacos los volcanes          avenida hivno nacional  ...
1160      tacos los volcanes          avenida hivno nacional  ...

                alcohol   smoking_area dress_code     accessibility   price  \
0     No_Alcohol_Served  not permitted   informal  no_accessibility  medium
1     No_Alcohol_Served  not permitted   informal  no_accessibility  medium
2     No_Alcohol_Served  not permitted   informal  no_accessibility  medium
3     No_Alcohol_Served  not permitted   informal  no_accessibility  medium
4     No_Alcohol_Served  not permitted   informal  no_accessibility  medium
...                 ...            ...        ...               ...     ...
1156  No_Alcohol_Served           none   informal        completely     low
1157  No_Alcohol_Served           none   informal        completely     low
1158  No_Alcohol_Served           none   informal        completely     low
1159  No_Alcohol_Served           none   informal        completely     low
1160  No_Alcohol_Served           none   informal        completely     low

     url Rambience franchise    area other_services
0      ?  familiar         f  closed           none
1      ?  familiar         f  closed           none
2      ?  familiar         f  closed           none
3      ?  familiar         f  closed           none
4      ?  familiar         f  closed           none
...   ..       ...       ...     ...            ...
1156   ?     quiet         t  closed           none
1157   ?     quiet         t  closed           none
1158   ?     quiet         t  closed           none
1159   ?     quiet         t  closed           none
1160   ?     quiet         t  closed           none

[1161 rows x 25 columns]
```

```
merged_rating_df.pivot(index='alcohol', columns='smoking_area', values='rating')
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
```

(continues on next page)

(continued from previous page)

```
/tmp/ipykernel_5491/1397663245.py in <module>
----> 1 merged_rating_df.pivot(index='alcohol', columns='smoking_area', values='rating
↪')

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪frame.py in pivot(self, index, columns, values)
   7791         from pandas.core.reshape.pivot import pivot
   7792
-> 7793         return pivot(self, index=index, columns=columns, values=values)
   7794
   7795     _shared_docs[

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪reshape/pivot.py in pivot(data, index, columns, values)
    515         else:
    516             indexed = data._constructor_sliced(data[values]._values,␣
↪index=multiindex)
--> 517     return indexed.unstack(columns_listlike)
    518
    519

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪series.py in unstack(self, level, fill_value)
   4079         from pandas.core.reshape.reshape import unstack
   4080
-> 4081         return unstack(self, level, fill_value)
   4082
   4083     # ----------------------------------------------------------------------

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪reshape/reshape.py in unstack(obj, level, fill_value)
    458         if is_1d_only_ea_dtype(obj.dtype):
    459             return _unstack_extension_series(obj, level, fill_value)
--> 460         unstacker = _Unstacker(
    461             obj.index, level=level, constructor=obj._constructor_expanddim
    462         )

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪reshape/reshape.py in __init__(self, index, level, constructor)
    131             raise ValueError("Unstacked DataFrame is too big, causing int32␣
↪overflow")
    132
--> 133         self._make_selectors()
    134
    135     @cache_readonly

~/git/data-science-practical-approach/venv/lib/python3.8/site-packages/pandas/core/
↪reshape/reshape.py in _make_selectors(self)
    183
    184         if mask.sum() < len(self.index):
--> 185             raise ValueError("Index contains duplicate entries, cannot reshape
↪")
    186
    187         self.group_index = comp_index

ValueError: Index contains duplicate entries, cannot reshape
```

```
merged_rating_df.groupby(['alcohol','smoking_area']).mean().reset_index().pivot(index=
↪'alcohol', columns='smoking_area', values='rating')
```

```
smoking_area            none  not permitted  only at bar  permitted   section
alcohol
Full_Bar            1.305556       0.857143          NaN   1.500000  1.272727
No_Alcohol_Served   1.186788       1.124402          NaN   1.114286  1.265823
Wine-Beer           1.217391       1.000000     1.368421   1.300000  1.275000
```

```
merged_rating_df.groupby(['alcohol','smoking_area']).count().reset_index().
↪pivot(index='alcohol', columns='smoking_area', values='rating')
```

```
smoking_area         none  not permitted  only at bar  permitted  section
alcohol
Full_Bar             36.0            7.0          NaN        4.0     33.0
No_Alcohol_Served   439.0          209.0          NaN       35.0     79.0
Wine-Beer           161.0            9.0         19.0       10.0    120.0
```

```
geo_df.columns
```

```
Index(['latitude', 'longitude', 'the_geom_meter', 'name', 'address', 'city',
       'state', 'country', 'fax', 'zip', 'alcohol', 'smoking_area',
       'dress_code', 'accessibility', 'price', 'url', 'Rambience', 'franchise',
       'area', 'other_services'],
      dtype='object')
```

```
merged_rating_df.groupby(['city','smoking_area']).mean().reset_index().pivot(index=
↪'city', columns='smoking_area', values='rating')
```

```
smoking_area            none  not permitted  only at bar  permitted   section
city
?                   1.052632       1.000000          NaN   1.833333  1.300000
Cd Victoria         0.625000            NaN          NaN        NaN       NaN
Cd. Victoria             NaN       0.750000          NaN        NaN       NaN
Ciudad Victoria          NaN       1.300000          NaN   0.750000       NaN
Cuernavaca          1.411765       1.157895     1.555556   1.500000  1.000000
Jiutepec            1.461538            NaN          NaN   1.166667       NaN
San Luis Potosi     1.211494       1.168317     1.200000        NaN  1.273743
Soledad             1.058824            NaN          NaN        NaN       NaN
cuernavaca               NaN       1.600000          NaN        NaN       NaN
s.l.p                    NaN       0.916667          NaN        NaN       NaN
s.l.p.              1.281250            NaN          NaN        NaN       NaN
san luis potos      2.000000            NaN          NaN        NaN       NaN
san luis potosi     1.272727       1.000000          NaN        NaN  1.500000
san luis potosi          NaN            NaN          NaN   1.400000       NaN
slp                 1.166667            NaN          NaN        NaN       NaN
victoria            1.000000       0.600000          NaN   0.937500       NaN
victoria            0.750000            NaN          NaN        NaN       NaN
```

Part IV

4. Data Exploration

CHAPTER EIGHTEEN

DATA EXPLORATION

this is an introduction

Part V

5. Data Visualisation

CHAPTER

NINETEEN

DATA VISUALISATION

this is an introduction

Part VI

6. Machine Learning

CHAPTER TWENTY

MACHINE LEARNING

this is an introduction