EXPERIMENT NO- 3

AIM: To implement Data Cleaning and Storage.

RESOURCES REQUIRED: Windows/MAC/Linux O.S, Compatible version of Python.

THEORY:

What is data cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to map out a framework for your organization.

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with.

Step 4: Handle missing data

There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

- As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
- As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
- As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Validate and QA

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation

CONCLUSION: Hence, we have successfully studied Data Cleaning and Storage.

In [72]:

```
import pandas as pd
import numpy as np
from tinydb import TinyDB
```

In [73]:

```
post_db, comment_db = TinyDB('post.json'), TinyDB('comment.json')
```

In [74]:

```
post_list, comment_list = post_db.all(), comment_db.all()
```

In [75]:

```
post_list[:5]
```

```
Out[75]: [{'id': '12glkw4',
            'author': 'TheBodyPolitic1',
            'author_id': 'von3w6y2',
            'total_comments': 319,
            'upvote': 591,
            'post_type': 'top week',
            'title': "Why didn't Python become popular until long after its creati
          on?",
            'body': "Python was invented in 1994, two years before Java.\n\nGiven
          it's age, why didn't Python become popular or even widely known about, u
          ntil much later?",
            'downvote': 37.72340425531918,
            'url': 'https://www.reddit.com/r/Python/comments/12glkw4/why_didnt_pyt
          hon_become_popular_until_long_after/',
            'created_on': 1681051909.0,
            'subreddit': 'python',
            'subreddit_id': 't5_2qh0y'},
           {'id': '12hj9oc',
            'author': 'MetonymyQT',
            'author_id': 'shnqm',
            'total_comments': 80,
            'upvote': 527,
            'post_type': 'top week',
            'title': 'Free course: Build a modern API with FastAPI and Python',
            'body': "Hello everyone! \n\nI've posted this course 4 months ago on t
          his sub Reddit and it was well received. I want to do another giveawa
          y,\n\nAll 3 coupons expire in 4 days and allow for a maximum 1k per coup
          on redeems.\n\n[https://www.udemy.com/course/build-a-movie-tracking-api-
          with-fastapi-and-python/?couponCode=90707F6B0050F6D60303](https://www.ud
          emy.com/course/build-a-movie-tracking-api-with-fastapi-and-python/?coupo
          nCode=90707F6B0050F6D60303)\n\n[https://www.udemy.com/course/build-a-mov
          ie-tracking-api-with-fastapi-and-python/?couponCode=F0744D2CC6E3E1C6E62
          2](https://www.udemy.com/course/build-a-movie-tracking-api-with-fastapi-
          and-python/?couponCode=F0744D2CC6E3E1C6E622)\n\n[https://www.udemy.com/c
          ourse/build-a-movie-tracking-api-with-fastapi-and-python/?couponCode=A62
          0331B2F48333F76D7](https://www.udemy.com/course/build-a-movie-tracking-a
          pi-with-fastapi-and-python/?couponCode=A620331B2F48333F76D7)\n\nI know t
          he course is not top notch and can be improved a lot but honestly I hope
          you like it as it is. I've set the lowest price I could set for it on Ud
          emy and I'm just grateful that it helped cover my blog hosting fees over
          the last 3 years.\n\nThank you!",
            'downvote': 21.958333333333353,
            'url': 'https://www.reddit.com/r/Python/comments/12hj9oc/free_course_b
          uild_a_modern_api_with_fastapi_and/',
            'created_on': 1681134311.0,
            'subreddit': 'python',
            'subreddit_id': 't5_2qh0y'},
           {'id': '12egsoz',
            'author': '2broke2code',
            'author_id': 'rs4dqilj',
            'total_comments': 105,
            'upvote': 467,
            'post_type': 'top week',
            'title': "I trained a RoastBot on >120,000 faces and >0.5 million comm
          ents and it's a menace 😈.",
            'body': "It uses facial recognition to fetch roasts for users from the
          r/RoastMe subreddit.\n\nTry it out [here](https://subroast.me)\n\n**App:
          ** [https://subroast.me](https://subroast.me)\n\n**Code:**  [nizarhaide
          r/RoastMe (github.com)](https://github.com/nizarhaider/RoastMe)\n\n# Tec
          h Stack\n\n**Front End:** Bootstrap5 + Vanilla JS\n\n**Back End:** Flask
          chipped away at PHP, but clearly that never overtook it.\n\nPython was k
          ind of "just around."  Plone was its only killer app, but it kept it ali
          ve.  CMS was a big enterprise need that didn\'t have an elegant solutio
          n, and so people were trying everything.  Google was using it for a lot
          of scientific research and from there, it crept into the research world,
          because the syntax is easy, where it got a stranglehold.  (Just five yea
          rs ago, I was working at a bioresearch lab, and I can\'t tell you how mu
          ch bad Python 2.7 was still around.). Being in research, it serendipitou
          sly positioned itself into the math and AI boom that we\'re in today, an
          d it\'s everywhere.\n\nSo basically, Plone and Google kept it alive long
          enough for dumb luck to take over and show people how good it was.  Othe
          rwise, I think it\'d just be another niche academic language.',
            'created_on': 1681079618.0,
            'upvotes': 3,
            'author_id': '380he',
            'author_name': 'snapetom'}]
```

In [77]:
```
post_df = pd.DataFrame(post_list).set_index("id")
post_df['downvote'] = post_df['downvote'].astype('int')
comment_df = pd.DataFrame(comment_list).set_index('comment_id')
```

In [78]:
```
post_df.head(5)
```

Out[78]:

	author	author_id	total_comments	upvote	post_type	ti
id						
12glkw4	TheBodyPolitic1	von3w6y2	319	591	top week	Why did Python becor popular ur long a
12hj9oc	MetonymyQT	shnqm	80	527	top week	Free cours Build a mode API w FastAPI a
12egsoz	2broke2code	rs4dqilj	105	467	top week	I trainec RoastBot >120,000 fac and >0
12ffsif	midnitte	3gad9	64	378	top week	EP 684: A P Interpreter G Accept
12fzdu2	aeluro1	88efmodhn	12	370	top week	Comprehensi Reddit Sav Pos Downloade

In [79]:
```
comment_df.head()
```

Out[79]:

	post_id	parent_id	body	created_on	upvotes	author_id	autho
comment_id							
jfkvwx0	12glkw4	t3_12glkw4	Hardware wasn't ready for Python in that time	1.681054e+09	5	22vat21u	D
jfnha98	12glkw4	t3_12glkw4	Because Python was developed with the conceit ...	1.681095e+09	2	4wtjvsh6	Fred\
jflnf2e	12glkw4	t3_12glkw4	Perl was *the* scripting language in the early...	1.681065e+09	2	4i9hp	to
jflbch7	12glkw4	t3_12glkw4	I was a web developer between 2000 and 2010, a...	1.681060e+09	2	8hi6986p	Dre
jfmk972	12glkw4	t1_jflbch7	Adding on to this from my POV, early web in th...	1.681080e+09	3	380he	sr

In [80]:

```
from datetime import datetime

post_df['created_on'] = post_df['created_on'].apply(
    lambda unix_time: datetime.utcfromtimestamp(unix_time).strftime('%Y-%
)


comment_df['created_on'] = comment_df['created_on'].apply(
    lambda unix_time: datetime.utcfromtimestamp(unix_time).strftime('%Y-%
)
```

In [81]:

```
post_df['created_on'], comment_df['created_on']
```

```
Out[81]: (id
          12glkw4    2023-04-09 14:51:49
          12hj9oc    2023-04-10 13:45:11
          12egsoz    2023-04-07 10:36:18
          12ffsif    2023-04-08 08:26:37
          12fzdu2    2023-04-08 21:43:41
          12f3glm    2023-04-07 23:24:08
          12ha6mc    2023-04-10 06:58:33
          12dfdq1    2023-04-06 10:11:05
          12cgplg    2023-04-05 11:10:38
          12bl2uj    2023-04-04 14:37:50
          Name: created_on, dtype: object,
          comment_id
          jfkvwx0    2023-04-09 15:27:18
          jfnha98    2023-04-10 02:55:16
          jflnf2e    2023-04-09 18:37:04
          jflbch7    2023-04-09 17:13:45
          jfmk972    2023-04-09 22:33:38
                            ...
          jeyxtqh    2023-04-04 21:19:50
          jez63w8    2023-04-04 22:17:42
          jeywttj    2023-04-04 21:13:04
          jf0v8f8    2023-04-05 07:15:14
          jf00txb    2023-04-05 02:06:08
          Name: created_on, Length: 325, dtype: object)
```

```
In [82]: post_df.to_csv('post.csv')
         comment_df.to_csv('comment.csv')
```