Task - Develop a code that can clean this dataset and extract Technical (Hard) skills?

Solution: Starting with reading the dataset and dropping all the null values from dataset. Then I drop all the duplicates from our dataset, which almost reduce the number by half in our dataset

```
In [1]: import pandas as pd
        df = pd.read_csv("Raw_Skills_Dataset.csv")
        df.shape
```

```
Out[1]: (34116, 1)
```

```
In [2]: df.head()
```

Out[2]:

	RAW DATA
0	What ifs
1	seniority
2	familiarity
3	functionalities
4	Lambdas

```
In [3]: df.dropna(inplace = True)
        df.drop_duplicates(inplace = True)
```

```
In [4]: df.shape
```

```
Out[4]: (15677, 1)
```

Then I used some regex to find all the required data, and replace all the numerical data with empty sting from our dataset.

```
In [5]: y = df['RAW DATA'].str.findall(r'(^[A-Z].*)')
```

```
In [6]: y.replace(to_replace ="\d+",
                  value ="")
```

```
Out[6]: 0                                                [What ifs]
        1                                                        []
        2                                                        []
        3                                                        []
        4                                                 [Lambdas]
                                      ...
        34106                              [Leadership qualities]
        34108                                          [Actalent]
        34109                                                    []
        34110    [Self-motivated, enthusiastic and strong drive]
        34111                                                    []
        Name: RAW DATA, Length: 15677, dtype: object
```

After that I did the basic things, like removing empty list from our dataset and some simple cleaning tasks. And convert it to pandas data frame

```
In [66]: d=[]

         for i in az:
             d.append(i[2:-2])

         print(d)
```

```
['Java Streams', 'Object Oriented analysis', 'Relational Databases', 'SQL', 'ORM', 'JPA2', 'Hibernate', 'MyBatis', 'Git.. Fam
iliarity', 'Maven', 'Gradle.. Familiarity', 'Familiarity', 'Demonstrable experience', 'Bachelor or Master degree', 'STEM majo
rs', 'Strong algorithms', 'Python', 'Scala programming experience', 'Exceptional proficiency', 'MySQL', 'Oracle', 'NoSQL Stor
es', 'Clickhouse', 'Distributed Processing Engines', 'Apache Flink', 'Celery', 'Distributed Queues', 'AWS Kinesis', 'GCP PusS
ub', 'GCP', 'Azure', 'AWS', 'Excellent Unix/Linux experience and programming experience', 'IPC mechanisms', 'TCP/IP', 'Kafk
a', 'API gateways', 'CAP Theorem', 'RAFT', 'Paxos', 'CI/CD', 'Github', 'Makefiles', 'Microservices', 'GCPNode.js', 'REST APIs
servicesExperience writing tools', 'Ruby', 'PerlExperience', 'Bazel', 'HTTP 2/websocket', 'Redis', 'Web UIExperience', 'JavaS
cript', 'HTML', 'CSS', 'Angular', 'Meteor', 'Ember', 'Knockout', 'UI Component libraries', 'Vue', 'Bootstrap', 'ES6', 'TypeSc
ript', 'Dependency injection patterns', 'Async Module Definition', 'AMD, requireJS', 'Stylesheet languages', 'SASS', 'Redux',
'Reactive Programming', 'RxJS', 'JavaStrong programming skills', 'Guice knowledge', 'Protobuf knowledge', 'APIs', 'Web UI:Dee
p understanding', 'JSON', 'JavaScript Design', 'Architectural Patterns', 'Proficiency', 'Spring/Guice', 'JPA/Hibernate', 'Jav
aExperience', 'Containers knowledge', 'Multi-threaded concepts', 'NoSQL', 'MongoDB', 'Cassandra', 'Java Application Servers',
'Oracle WebLogic', 'JBoss', 'Tomcat', 'Jasmine', 'Karma', 'Protractor', 'Junit', 'Cross-browser techniques', 'Chrome Dev Tool
s', 'Responsive design implementation', 'Browser performance auditing', 'Scrum / Kanban', 'Guice', 'GRPC', 'Excellent analyti
cal and problem-solving skills', 'Proven-deep-expertise', 'Python programming', 'Django', 'Flask', 'Unix / Linux command lin
e', 'Object-oriented concepts', 'UI frameworks', 'Hadoop / Hive / Presto', 'ML/AI frameworks', 'TensorFlow', 'H20', 'Mesos',
'Kubernetes', 'Application benchmarking', 'Building user interface applications', 'UI Applications', 'QML', 'Qt/C++', 'RasP
i', 'BeagleBone', 'IP-based network protocols', 'Myntra's new product platforms', 'Architects', 'DevOps engineers', 'Myntra',
'Java/GO', 'Developing user-facing web applications', 'Bachelor', 'Backend services', 'React JS', 'JS/TypeScript', 'Webpack b
```

```
In [67]: df = pd.DataFrame(d,columns = ['Technical Skills'])
```

Then I convert our panda's data frame to a CSV file as required in our assignment

```
In [67]: df = pd.DataFrame(d,columns = ['Technical Skills'])
```

```
In [68]: df
```

Out[68]:

	Technical Skills
0	Java Streams
1	Object Oriented analysis
2	Relational Databases
3	SQL
4	ORM
5	JPA2
6	Hibernate
7	MyBatis
8	Git.. Familiarity
9	Maven
10	Gradle.. Familiarity
11	Familiarity

```
In [69]: df.to_csv("output.csv",index = False)
```

```
In [70]: #to check
         xc = pd.read_csv("output.csv")
```

Final to check our CSV file

```
In [70]: #to check
         xc = pd.read_csv("output.csv")
```

```
In [71]: xc.head()
```

Out[71]:

	Technical Skills
0	Java Streams
1	Object Oriented analysis
2	Relational Databases
3	SQL
4	ORM

```
In [72]: xc.shape
```

```
Out[72]: (6540, 1)
```