A.1) Study of Hadoop Installation.

a. Single Node

b. Multiple NodeApache Hadoop Installation

Open Terminal

 # Update the source list
aaditya@laptop:~$ **sudo apt-get update**

The OpenJDK project is the default version of Java
Install JDK

aaditya@laptop:~$ **sudo apt-get install default-jdk`**

aaditya@laptop:~$ **java -version**
java version "1.7.0_95"

Adding a dedicated Hadoop user aaditya@laptop:~$ **sudo addgroup hadoop** Adding group `hadoop' (GID 1002) ...
Done.

aaditya@laptop:~$ **sudo adduser --ingroup hadoop hduser (Press Enter)**

Is the information correct? [Y/n] Y

Enter new password for new hduser

aaditya@laptop:~$ **sudo adduser hduser sudo**
Adding user `hduser' to group `sudo' ... Adding user
hduser to group sudo Done.

Logout from current user and log in to hduser

Installing SSH

aaditya@laptop:~$**su hduser**

hduser@laptop:~$ **sudo apt-get install ssh**

Create and Setup SSH Certificates

hduser@laptop:~$**cd /home/hduser/**

hduser@laptop:~$ **ssh-keygen -t rsa -P ""**
Generating public/private rsa key pair.

Enter file in which to save the key (/home/hduser/.ssh/id_rsa): **(**press enter)**
Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa. Your public key has been saved in /home/hduser/.ssh/id_rsa.pub. The key fingerprint is:
50:6b:f3:fc:0f:32:bf:30:79:c2:41:71:26:cc:7d:e3 hduser@laptop The key's randomart image is:

```
+--[ RSA 2048] ---+
|        .oo.o    |
|       . .o=. o  |
|      . + . o . |
|       o =    E  |
|        S +      |
|         . +     |
|          O +    |
|           O o   |
|            o..  |
+-----------------+
```

hduser@laptop:$ **cat /home/hduser/.ssh/id_rsa.pub >> /home/hduser/.ssh/authorized_keys**

hduser@laptop:$ **ssh localhost**

hduser@laptop:$ **exit**

hduser@laptop:$ **cd /home/hduser**

Download Hadoop

hduser@laptop:~$

wget http://mirrors.sonic.net/apache/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz

Or from http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.2/hadoop-2.7.2.tar.gz hduser@laptop:~$

tar xvzf hadoop-2.8.3.tar.gz

hduser@laptop:~$ **cd hadoop-2.8.3**

hduser@laptop:~$ **sudo mkdir /usr/local/hadoop**

hduser@laptop:~/hadoop-2.7.2$ **sudo mv * /usr/local/hadoop**
hduser@laptop:~/hadoop-2.7.2$ **sudo chown -R hduser:hadoop /usr/local/hadoop**

hduser@laptop:~/hadoop-2.7.2$ **sudo apt-get install vim Check Java**
hduser@laptop:~$ **readlink -f /usr/bin/javac**
/usr/lib/jvm/java-8-openjdk-amd64/bin/javac *** for 64 bit

/usr/lib/jvm/java-8-openjdk-i386/bin/javac ***for 32 bit

Set-up the Configuration Files

hduser@laptop:~$ **vim ~/.bashrc**
#HADOOP VARIABLES START ****Append this at end of file******* export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64**
export HADOOP_INSTALL=/usr/local/hadoop export PATH=$PATH:$HADOOP_INSTALL/bin export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL export HADOOP_COMMON_HOME=$HADOOP_INSTALL export HADOOP_HDFS_HOME=$HADOOP_INSTALL export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib" #HADOOP VARIABLES END

Save the file and Exit (**esc:wq!** to save the file in vim) hduser@laptop:~$
source ~/.bashrc

hduser@laptop:~$ **vim /usr/local/hadoop/etc/hadoop/hadoop-env.sh**

(Change existing JAVA_HOME to **/usr/lib/jvm/java-7-openjdk-amd64 if machine is 64 Bit** or

If machine is 32 bit change it to /usr/lib/jvm/java-7-openjdk-i386) **export**

JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
Save and Exit

hduser@laptop:~$ **vim /usr/local/hadoop/etc/hadoop/yarn-site.xml**

Open the file and enter the following in between <configuration></configuration> tag and save :

```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
```

Create Temporary Directory to store App-data hduser@laptop:~$ **sudo**

mkdir -p /app/hadoop/tmp

hduser@laptop:~$ **sudo chown hduser:hadoop /app/hadoop/tmp**

hduser@laptop:~$ **vim /usr/local/hadoop/etc/hadoop/core-site.xml**

Open the file and enter the following in between the <configuration></configuration> tag:

```
<property>
<name>hadoop.tmp.dir</name>
<value>/app/hadoop/tmp</value>
<description>A base for other temporary directories.</description>
</property>
```

<property>
<name>fs.default.name</name>
<value>hdfs://localhost:54310</value></property>

hduser@laptop:~$ **cp /usr/local/hadoop/etc/hadoop/mapred-site.xml.template /usr/local/hadoop/etc/hadoop/mapred-site.xml**

hduser@laptop:~$ **vim /usr/local/hadoop/etc/hadoop/mapred-site.xml**

Add these properties in <configuration></configuration> tag:

<property>

<name>mapred.job.tracker</name>

<value>localhost:54311</value>

</property>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>
Save and Exit

hduser@laptop:~$ **sudo mkdir -p /usr/local/hadoop_store/hdfs/namenode**

hduser@laptop:~$ **sudo mkdir -p /usr/local/hadoop_store/hdfs/datanode**
hduser@laptop:~$ **sudo chown -R hduser:hadoop /usr/local/hadoop_store**

hduser@laptop:~$ **vim /usr/local/hadoop/etc/hadoop/hdfs-site.xml**
Open the file and add the following properties between the <configuration></configuration> tag:

<property>
<name>dfs.replication</name><value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/usr/local/hadoop_store/hdfs/namenode</value>
</property>

```
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop_store/hdfs/datanode</value>
</property>
```

hduser@laptop:~$ **hadoop namenode -format** **Starting Hadoop**

hduser@laptop$: **/usr/local/hadoop/sbin/start-all.sh** **We can check if it's really up and running:**

hduser@laptop$: **jps**

9026 NodeManager

7348 NameNode

9766 SecondaryNameNode

8887 ResourceManager

7507 DataNode

Hadoop UI visible in browser at localhost:50070

MR Job progress and history UI visible atlocalhost:8088 **To Stop Hadoop**

hduser@laptop$: **/usr/local/hadoop/sbin/stop-all.sh**

Done

Conclusion:

In this way single node Hadoop was installed & configured on Ubuntu for Big Data analytics

PartA: Assignment No2

Aim: Design a distributed application using MapReduce which processes log file of asystem.List out users who have logged for maximum period on the system.

Name of input file is access_log_short.csv

PARTA

1. Open Eclipse> File > New > Java Project >(Name it – MRProgramsDemo) > Next>Click on Libraries Tab>Click on Add External JARS tab

   jar FILE LOCATION
   /usr/lib/Hadoop ->select all jar files
   /usr/lib/Hadoop/client ->select all jar files

2. Right Click > New > Package (Name it - mrLogFile_demo > Finish.
3. Right Click on mrLogFile_demo Package > New > Class (Name it – UserLogDriver).

Add following code in that class

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;

public class UserLogDriver {
	public static void main(String[] args) {
		JobClient my_client = new JobClient();
		// Create a configuration object for the job
		JobConf job_conf = new JobConf(UserLogDriver.class);

		// Set a name of the Job
		job_conf.setJobName("MaxLoggedUsers");

		// Specify data type of output key and value
		job_conf.setOutputKeyClass(Text.class);
		job_conf.setOutputValueClass(IntWritable.class);

		// Specify names of Mapper and Reducer Class
		job_conf.setMapperClass(UserLogMapper.class);


		job_conf.setReducerClass(UserLogReducer .class);

		// Specify formats of the data type of Input and output
		job_conf.setInputFormat(TextInputFormat.class);
		job_conf.setOutputFormat(TextOutputFormat.class);

		// Set input and output directories using command line arguments,
		//arg[0] = name of input directory on HDFS, and arg[1] =  name of
output directory to be created to store the output file.

		FileInputFormat.setInputPaths(job_conf, new Path(args[0]));
		FileOutputFormat.setOutputPath(job_conf, new Path(args[1]));

		my_client.setConf(job_conf);
		try {
			// Run the job
			JobClient.runJob(job_conf);
		} catch (Exception e) {
			e.printStackTrace();
		}
	}
}
```

 # Save the file

4. Right Click on mrLogFile_demo Package > New > Class (Name it – `UserLogReducer`).

```
import java.io.IOException;
import java.util.*;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class UserLogReducer  extends MapReduceBase implements Reducer<Text,
IntWritable, Text, IntWritable> {

      public void reduce(Text t_key, Iterator<IntWritable> values,
OutputCollector<Text,IntWritable> output, Reporter reporter) throws IOException
{
            Text key = t_key;
            int frequencyForUser = 0;
            while (values.hasNext()) {
                  // replace type of value with the actual type of our value
                  IntWritable value = (IntWritable) values.next();
                  frequencyForUser += value.get();

            }
            output.collect(key, new IntWritable(frequencyForUser));
      }
}
```

 # Save the file

5. **Right Click on mrLogFile_demo Package > New > Class (Name it – `UserLogMapper`).**

Add following code in that class

```
package MRLogFile;
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class UserLogMapper extends MapReduceBase implements Mapper<LongWritable,
Text, Text, IntWritable> {
      private final static IntWritable one = new IntWritable(1);

      public void map(LongWritable key, Text value, OutputCollector<Text,
IntWritable> output, Reporter reporter) throws IOException {

            String valueString = value.toString();
            String[] SingleUserData = valueString.split("-");
            output.collect(new Text(SingleUserData[0]), one);
      }
}
```

Save the file

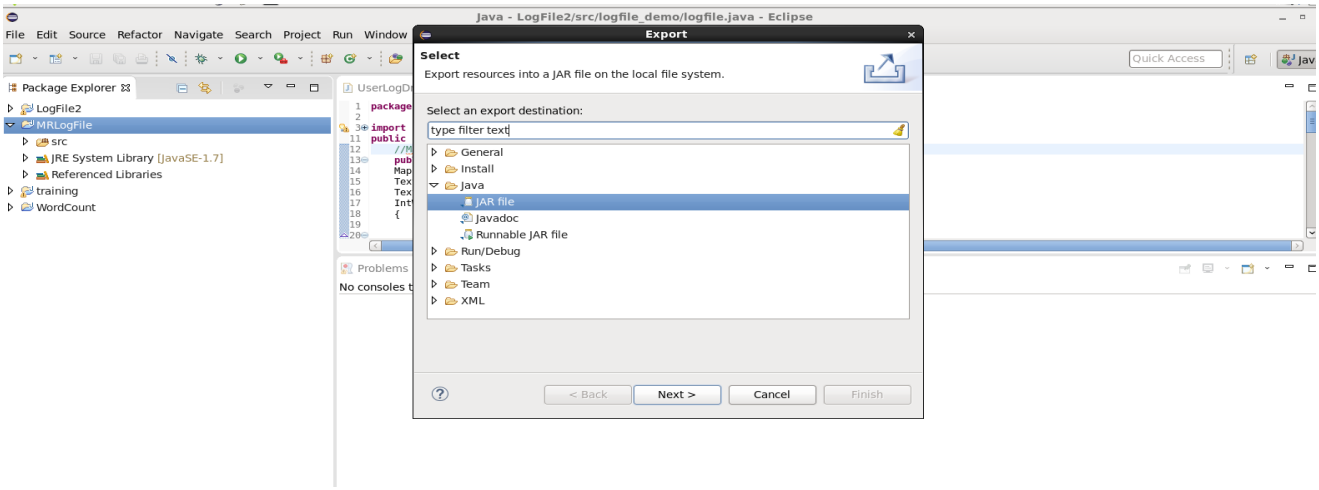
PART B

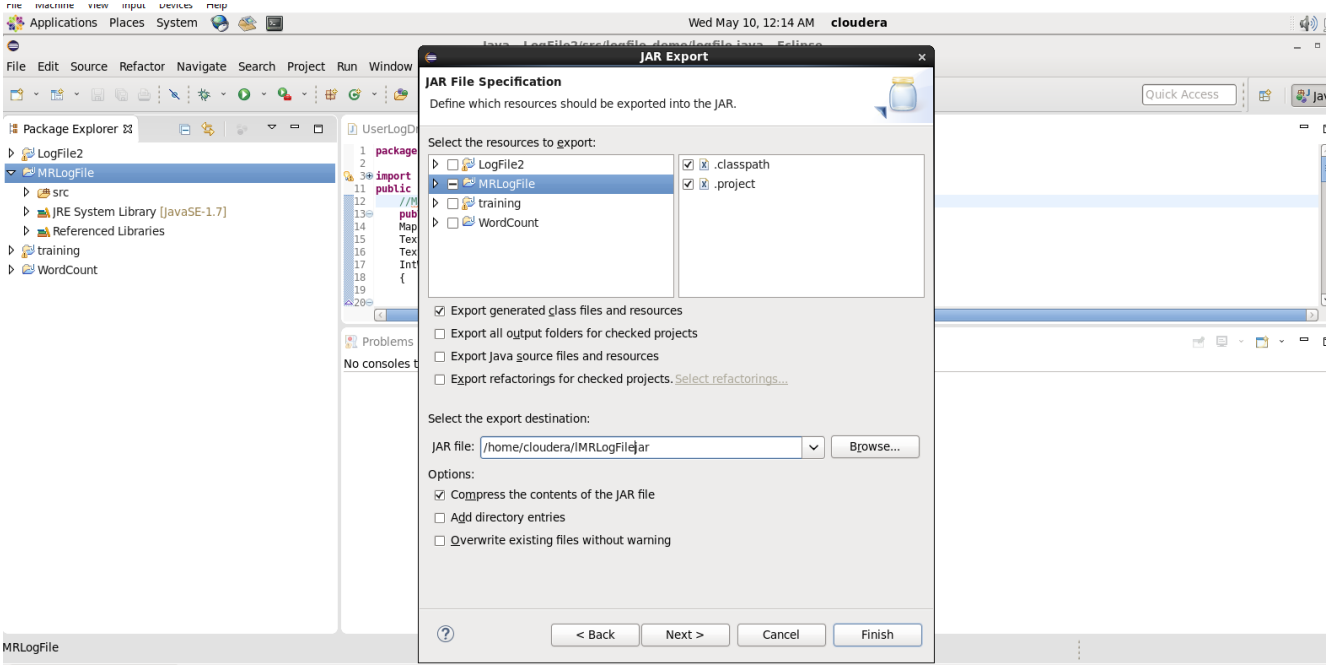
Create .jar file for your program execution :

Make a jar file

In eclipse Right click on MRLogFile Project > then select Export> Click on Java>JAR Files>Click on Next>then select export destination for JAR file as /home/Cloudera/MRlogFile.jar>Finish

*MRLogFile.jar file will get created in your /home/Cloudera/ folder





PART C:

 # Open terminal

#Check for present working Directory

```
[cloudera@quickstart ~]$ pwd
/home/cloudera
```

#Create inputfoder with name MRinputfolder1

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /MRinputfolder1
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 21 items
drwxr-xr-x   - cloudera supergroup          0 2023-05-10 00:22 /MRInputfolder
drwxr-xr-x   - cloudera supergroup          0 2023-05-10 00:29 /MRInputfolder1
drwxr-xr-x   - cloudera supergroup          0 2023-05-10 00:38
/MRoutputfolder1
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2023-05-10 00:02 /hbase
drwxr-xr-x   - cloudera supergroup          0 2023-05-06 01:27 /inputfolder
drwxr-xr-x   - cloudera supergroup          0 2023-05-07 23:02 /inputfolder1
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 01:45 /inputfolder5
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 03:10 /inputfolder8
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 03:13 /inputfolder9
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 03:31 /out10
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 03:38 /out11
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 03:50 /out14
drwxr-xr-x   - cloudera supergroup          0 2023-05-07 23:55 /out2
drwxr-xr-x   - cloudera supergroup          0 2023-05-08 03:22 /out9
drwxr-xr-x   - cloudera supergroup          0 2023-05-06 01:28 /outputfolder
drwxr-xr-x   - cloudera supergroup          0 2023-05-07 23:04 /outputfolder1
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs     supergroup          0 2023-05-05 23:26 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
```

```
[cloudera@quickstart ~]$ hdfs dfs -put
/home/cloudera/access_log_short.txt /MRInputfolder1

[cloudera@quickstart ~]$ hdfs dfs -cat
/MRInputfolder1/access_log_short.txt

[cloudera@quickstart ~]$ hadoop jar /home/cloudera/MRLogFile.jar
mrLogFile_demo.UserLogDriver /MRInputfolder1/access_log_short.txt
/MRoutputfolder1
```

```
23/05/10 00:38:06 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
23/05/10 00:38:06 INFO client.RMProxy: Connecting to ResourceManager at
/0.0.0.0:8032
```

```
23/05/10 00:38:07 WARN mapreduce.JobResourceUploader: Hadoop command-line
option parsing not performed. Implement the Tool interface and execute your
application with ToolRunner to remedy this.
23/05/10 00:38:07 INFO mapred.FileInputFormat: Total input paths to process :
1
23/05/10 00:38:07 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
      at java.lang.Object.wait(Native Method)
      at java.lang.Thread.join(Thread.java:1281)
      at java.lang.Thread.join(Thread.java:1355)
      at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputSt
ream.java:967)
      at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.j
ava:705)
      at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:8
94)
23/05/10 00:38:07 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
      at java.lang.Object.wait(Native Method)
      at java.lang.Thread.join(Thread.java:1281)
      at java.lang.Thread.join(Thread.java:1355)
      at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputSt
ream.java:967)
      at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.j
ava:705)
      at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:8
94)
23/05/10 00:38:07 INFO mapreduce.JobSubmitter: number of splits:2
23/05/10 00:38:08 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1683702103820_0001
23/05/10 00:38:08 INFO impl.YarnClientImpl: Submitted application
application_1683702103820_0001
23/05/10 00:38:08 INFO mapreduce.Job: The url to track the job:
http://quickstart.cloudera:8088/proxy/application_1683702103820_0001/
23/05/10 00:38:08 INFO mapreduce.Job: Running job: job_1683702103820_0001
23/05/10 00:38:19 INFO mapreduce.Job: Job job_1683702103820_0001 running in
uber mode : false
23/05/10 00:38:19 INFO mapreduce.Job:  map 0% reduce 0%
23/05/10 00:38:37 INFO mapreduce.Job:  map 100% reduce 0%
23/05/10 00:38:46 INFO mapreduce.Job:  map 100% reduce 100%
23/05/10 00:38:47 INFO mapreduce.Job: Job job_1683702103820_0001 completed
successfully
23/05/10 00:38:47 INFO mapreduce.Job: Counters: 49
      File System Counters
            FILE: Number of bytes read=26793
            FILE: Number of bytes written=484376
            FILE: Number of read operations=0
            FILE: Number of large read operations=0
            FILE: Number of write operations=0
            HDFS: Number of bytes read=147418
            HDFS: Number of bytes written=3838
            HDFS: Number of read operations=9
            HDFS: Number of large read operations=0
            HDFS: Number of write operations=2
      Job Counters
            Launched map tasks=2
            Launched reduce tasks=1
            Data-local map tasks=2
```

```
		Total time spent by all maps in occupied slots (ms)=28992
		Total time spent by all reduces in occupied slots (ms)=7394
		Total time spent by all map tasks (ms)=28992
		Total time spent by all reduce tasks (ms)=7394
		Total vcore-milliseconds taken by all map tasks=28992
		Total vcore-milliseconds taken by all reduce tasks=7394
		Total megabyte-milliseconds taken by all map tasks=29687808
		Total megabyte-milliseconds taken by all reduce tasks=7571456
	Map-Reduce Framework
		Map input records=1295
		Map output records=1295
		Map output bytes=24197
		Map output materialized bytes=26799
		Input split bytes=238
		Combine input records=0
		Combine output records=0
		Reduce input groups=227
		Reduce shuffle bytes=26799
		Reduce input records=1295
		Reduce output records=227
		Spilled Records=2590
		Shuffled Maps =2
		Failed Shuffles=0
		Merged Map outputs=2
		GC time elapsed (ms)=311
		CPU time spent (ms)=2690
		Physical memory (bytes) snapshot=556244992
		Virtual memory (bytes) snapshot=4519596032
		Total committed heap usage (bytes)=391979008
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
		IO_ERROR=0
		WRONG_LENGTH=0
		WRONG_MAP=0
		WRONG_REDUCE=0
	File Input Format Counters
		Bytes Read=147180
	File Output Format Counters
		Bytes Written=3838
[cloudera@quickstart ~]$ hdfs dfs -ls /MRoutputfolder1
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2023-05-10 00:38
/MRoutputfolder1/_SUCCESS
-rw-r--r--   1 cloudera supergroup       3838 2023-05-10 00:38
/MRoutputfolder1/part-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /MRoutputfolder1/part-00000
10.1.1.236  7
10.1.181.142     14
10.1.232.31      5
10.10.55.142     14
10.102.101.66    1
10.103.184.104   1
10.103.190.81    53
10.103.63.29     1
10.104.73.51     1
10.105.160.183   1
10.108.91.151    1
10.109.21.76     1
10.11.131.40     1
10.111.71.20     8
10.112.227.184   6
10.114.74.30     1
10.115.118.78    1
```

10.117.224.230 1
10.117.76.22 12
10.118.19.97 1
10.118.250.30 7
10.119.117.132 23
10.119.33.245 1
10.119.74.120 1
10.12.113.198 2
10.12.219.30 1
10.120.165.113 1
10.120.207.127 4
10.123.124.47 1
10.123.35.235 1
10.124.148.99 1
10.124.155.234 1
10.126.161.13 7
10.127.162.239 1
10.128.11.75 10
10.13.42.232 1
10.130.195.163 8
10.130.70.80 1
10.131.163.73 1
10.131.209.116 5
10.132.19.125 2
10.133.222.184 12
10.134.110.196 13
10.134.242.87 1
10.136.84.60 5
10.14.2.86 8
10.14.4.151 2
10.140.139.116 1
10.140.141.1 9
10.140.67.116 1
10.141.221.57 5
10.142.203.173 7
10.143.126.177 32
10.144.147.8 1
10.15.208.56 1
10.15.23.44 13
10.150.212.239 14
10.150.227.16 1
10.150.24.40 13
10.152.195.138 8
10.153.23.63 2
10.153.239.5 25
10.155.95.124 9
10.156.152.9 1
10.157.176.158 1
10.164.130.155 1
10.164.49.105 8
10.164.95.122 10
10.165.106.173 14
10.167.1.145 19
10.169.158.88 1
10.170.178.53 1
10.171.104.4 1
10.172.169.53 18
10.174.246.84 3
10.175.149.65 1
10.175.204.125 15
10.177.216.164 6
10.179.107.170 2
10.181.38.207 13
10.181.87.221 1

10.185.152.140	1
10.186.56.126	16
10.186.56.183	1
10.187.129.140	6
10.187.177.220	1
10.187.212.83	1
10.187.28.68	1
10.19.226.186	2
10.190.174.142	10
10.190.41.42	5
10.191.172.11	1
10.193.116.91	1
10.194.174.4	7
10.198.138.192	1
10.199.103.248	2
10.199.189.15	1
10.2.202.135	1
10.200.184.212	1
10.200.237.222	1
10.200.9.128	2
10.203.194.139	10
10.205.72.238	2
10.206.108.96	2
10.206.175.236	1
10.206.73.206	7
10.207.190.45	17
10.208.38.46	1
10.208.49.216	4
10.209.18.39	9
10.209.54.187	3
10.211.47.159	10
10.212.122.173	1
10.213.181.38	7
10.214.35.48	1
10.215.222.114	1
10.216.113.172	48
10.216.134.214	1
10.216.227.195	16
10.217.151.145	10
10.217.32.16	1
10.218.16.176	8
10.22.108.103	4
10.220.112.1	34
10.221.40.89	5
10.221.62.23	13
10.222.246.34	1
10.223.157.186	11
10.225.137.152	1
10.225.234.46	1
10.226.130.133	1
10.229.60.23	1
10.230.191.135	6
10.231.55.231	1
10.234.15.156	1
10.236.231.63	1
10.238.230.235	1
10.239.100.52	1
10.239.52.68	4
10.24.150.4	5
10.24.67.131	13
10.240.144.183	15
10.240.170.50	1
10.241.107.75	1
10.241.9.187	1

10.243.51.109 5
10.244.166.195 5
10.245.208.15 20
10.246.151.162 3
10.247.111.104 9
10.247.175.65 1
10.247.229.13 1
10.248.24.219 1
10.248.36.117 3
10.249.130.132 3
10.25.132.238 2
10.25.44.247 6
10.250.166.232 1
10.27.134.23 1
10.30.164.32 1
10.30.47.170 8
10.31.225.14 7
10.32.138.48 11
10.32.247.175 4
10.32.55.216 12
10.33.181.9 8
10.34.233.107 1
10.36.200.176 1
10.39.45.70 2
10.39.94.109 4
10.4.59.153 1
10.4.79.47 15
10.41.170.233 9
10.41.40.17 1
10.42.208.60 1
10.43.81.13 1
10.46.190.95 10
10.48.81.158 5
10.5.132.217 1
10.5.148.29 1
10.50.226.223 9
10.50.41.216 3
10.52.161.126 1
10.53.58.58 1
10.54.242.54 10
10.54.49.229 1
10.56.48.40 16
10.59.42.194 11
10.6.238.124 6
10.61.147.24 1
10.61.161.218 1
10.61.23.77 8
10.61.232.147 3
10.62.78.165 2
10.63.233.249 7
10.64.224.191 13
10.66.208.82 2
10.69.20.85 26
10.70.105.238 1
10.70.238.46 6
10.72.137.86 6
10.72.208.27 1
10.73.134.9 4
10.73.238.200 1
10.73.60.200 1
10.73.64.91 1
10.74.218.123 1
10.75.116.199 1
10.76.143.30 1

```
10.76.68.178      16
10.78.95.24       8
10.80.10.131      10
10.80.215.116     17
10.81.134.180     1
10.82.30.199      63
10.82.64.235      1
10.84.236.242     1
10.87.209.46      1
10.87.88.214      1
10.88.204.177     1
10.89.178.62      1
10.89.244.42      1
10.94.196.42      1
10.95.136.211     4
10.95.232.88      1
10.98.156.141     1
10.99.228.224     1
[cloudera@quickstart ~]$
```

OR

Goto Browser and enter
localhost:50070 and check the output in output directory.

A3 Write an application using HiveQL for flight information system which will include

a. Creating, Dropping, and altering Database tables.

b. Creating an external Hive table.

c. Load table with data, insert new values and field in the table, Join tables with Hive

d. Create index on Flight Information Table

e. Find the average departure delay per day in 2008.

a) Creating, Dropping, and altering Database tables

Using Hbase

#Create Table:

```
hbase(main):002:0> create 'flight','finfo','fsch'
0 row(s) in 4.6960 seconds
=> Hbase::Table - flight
```

#Table Created-list

```
hbase(main):003:0> list
TABLE
flight
table1
table2
3 row(s) in 0.0120 seconds
```

#Insert records in created table

```
hbase(main):004:0> put 'flight',1,'finfo:source','pune'
0 row(s) in 0.2480 seconds
hbase(main):008:0> put 'flight',1,'finfo:dest','mumbai'
0 row(s) in 0.0110 seconds
hbase(main):010:0> put 'flight',1,'fsch:at','10.25a.m.'
0 row(s) in 0.0060 seconds
```

```
hbase(main):011:0> put 'flight',1,'fsch:dt','11.25 a.m.'
0 row(s) in 0.0070 seconds
hbase(main):012:0> put 'flight',1,'fsch:delay','5min'
hbase(main):015:0> put 'flight',2,'finfo:source','pune'
0 row(s) in 0.0160 seconds
hbase(main):016:0> put 'flight',2,'finfo:dest','kolkata'
0 row(s) in 0.0070 seconds
hbase(main):017:0> put 'flight',2,'fsch:at','7.00a.m.'
0 row(s) in 0.0080 seconds

hbase(main):018:0> put 'flight',2,'fsch:dt','7.30a.m.'
0 row(s) in 0.0050 seconds
hbase(main):019:0> put 'flight',2,'fsch:delay','2 min'
0 row(s) in 0.0090 seconds
hbase(main):021:0> put 'flight',3,'finfo:source','mumbai'
0 row(s) in 0.0040 seconds
hbase(main):022:0> put 'flight',3,'finfo:dest','pune'
0 row(s) in 0.0070 seconds
hbase(main):023:0> put 'flight',3,'fsch:at','12.30p.m.'
0 row(s) in 0.0100 seconds
hbase(main):024:0> put 'flight',3,'fsch:dt','12.45p.m.'
0 row(s) in 0.0040 seconds
hbase(main):025:0> put 'flight',3,'fsch:delay','1 min'
0 row(s) in 0.0190 seconds
hbase(main):026:0> put 'flight',4,'finfo:source','mumbai'
0 row(s) in 0.0060 seconds
hbase(main):027:0> put 'flight',4,'finfo:dest','delhi'
0 row(s) in 0.0050 seconds
hbase(main):028:0> put 'flight',4,'fsch:at','2.00p.m.'
0 row(s) in 0.0080 seconds
hbase(main):029:0> put 'flight',4,'fsch:dt','2.45p.m.'
0 row(s) in 0.0040 seconds
hbase(main):030:0> put 'flight',4,'fsch:delay','10 min'
0 row(s) in 0.0140 seconds
```

#Display Records from Table ‘flight’

```
hbase(main):031:0>
scan 'flight' ROW
        COLUMN+CELL
 1                    column=finfo:dest, timestamp=1521312730758,
 value=mumbai
 1                    column=finfo:source, timestamp=1521312493881,
 value=pune
 1                    column=fsch:at, timestamp=1521312789417,
```

value=10.25a.m.
1 column=fsch:delay, timestamp=1521312850594, value=5min
1 column=fsch:dt, timestamp=1521312823256, value=11.25 a.m.
2 column=finfo:dest, timestamp=1521313135697, value=kolkata 2
column=finfo:source, timestamp=1521313092772, value=pune 2
column=fsch:at, timestamp=1521313166540, value=7.00a.m.
2 column=fsch:delay, timestamp=1521313229963, value=2 min
2 column=fsch:dt, timestamp=1521313202767, value=7.30a.m.
3 column=finfo:dest, timestamp=1521313310302, value=pune
3 column=finfo:source, timestamp=1521313290906, value=mumbai 3
column=fsch:at, timestamp=1521313333432, value=12.30p.m.
3 column=fsch:delay, timestamp=1521313379725, value=1 min
3 column=fsch:dt, timestamp=1521313353804, value=12.45p.m.
4 column=finfo:dest, timestamp=1521313419679, value=delhi
4 column=finfo:source, timestamp=1521313404831, value=mumbai 4
column=fsch:at, timestamp=1521313440328, value=2.00p.m.
4 column=fsch:delay, timestamp=1521313472389, value=10 min 4
column=fsch:dt, timestamp=1521313455226, value=2.45p.m. 4
row(s) in 0.0300 seconds
#Alter Table (add one more column family)
hbase(main):036:0> alter 'flight',NAME=>'revenue'
Updating all regions with the new schema...

0/1 regions
updated. 1/1
regions
updated.
Done.
0 row(s) in 3.7640 seconds
hbase(main):037:0> scan
'flight' ROW
COLUMN+CELL
1 column=finfo:dest, timestamp=1521312730758, value=mumbai 1
column=finfo:source, timestamp=1521312493881, value=pune 1
column=fsch:at, timestamp=1521312789417, value=10.25a.m. 1
column=fsch:delay, timestamp=1521312850594, value=5min
1 column=fsch:dt, timestamp=1521312823256, value=11.25 a.m.
2 column=finfo:dest, timestamp=1521313135697, value=kolkata 2
column=finfo:source, timestamp=1521313092772, value=pune 2
column=fsch:at, timestamp=1521313166540, value=7.00a.m.
2 column=fsch:delay, timestamp=1521313229963, value=2 min
2 column=fsch:dt, timestamp=1521313202767, value=7.30a.m.
3 column=finfo:dest, timestamp=1521313310302, value=pune
3 column=finfo:source, timestamp=1521313290906, value=mumbai 3
column=fsch:at, timestamp=1521313333432, value=12.30p.m.
3 column=fsch:delay, timestamp=1521313379725, value=1 min 3
column=fsch:dt, timestamp=1521313353804, value=12.45p.m.
4 column=finfo:dest, timestamp=1521313419679, value=delhi
4 column=finfo:source, timestamp=1521313404831, value=mumbai 4
column=fsch:at, timestamp=1521313440328, value=2.00p.m.
4 column=fsch:delay, timestamp=1521313472389, value=10 min 4
column=fsch:dt, timestamp=1521313455226, value=2.45p.m. 4
row(s) in 0.0290 seconds
#Insert records into added column family
hbase(main):038:0> put 'flight',4,'revenue:rs','45000' 0
row(s) in 0.0100 seconds
#Check the updates
hbase(main):039:0> scan
'flight' ROW
COLUMN+CELL
1 column=finfo:dest, timestamp=1521312730758, value=mumbai 1
column=finfo:source, timestamp=1521312493881, value=pune 1

column=fsch:at, timestamp=1521312789417, value=10.25a.m. 1
column=fsch:delay, timestamp=1521312850594, value=5min
1 column=fsch:dt, timestamp=1521312823256, value=11.25 a.m.
2 column=finfo:dest, timestamp=1521313135697, value=kolkata 2
column=finfo:source, timestamp=1521313092772, value=pune 2
column=fsch:at, timestamp=1521313166540, value=7.00a.m.
2 column=fsch:delay, timestamp=1521313229963, value=2 min
2 column=fsch:dt, timestamp=1521313202767, value=7.30a.m.
3 column=finfo:dest, timestamp=1521313310302, value=pune
3 column=finfo:source, timestamp=1521313290906, value=mumbai 3
column=fsch:at, timestamp=1521313333432, value=12.30p.m.
3 column=fsch:delay, timestamp=1521313379725, value=1 min
3 column=fsch:dt, timestamp=1521313353804, value=12.45p.m.
4 column=finfo:dest, timestamp=1521313419679, value=delhi
4 column=finfo:source, timestamp=1521313404831, value=mumbai 4
column=fsch:at, timestamp=1521313440328, value=2.00p.m.
4 column=fsch:delay, timestamp=1521313472389, value=10 min 4
column=fsch:dt, timestamp=1521313455226, value=2.45p.m. 4
column=revenue:rs, timestamp=1521314406914, value=45000 4
row(s) in 0.0340 seconds
#Delete Column family
hbase(main):040:0> alter 'flight',NAME=>'revenue',METHOD=>'delete'
Updating all regions with the new schema...
0/1 regions
updated. 1/1
regions
updated.
Done.
0 row(s) in 3.7880
seconds #changes
Reflected in Table
hbase(main):041:0> scan
'flight' ROW
COLUMN+CELL
1 column=finfo:dest, timestamp=1521312730758, value=mumbai 1
column=finfo:source, timestamp=1521312493881, value=pune 1
column=fsch:at, timestamp=1521312789417, value=10.25a.m. 1
column=fsch:delay, timestamp=1521312850594, value=5min
1 column=fsch:dt, timestamp=1521312823256, value=11.25 a.m.
2 column=finfo:dest, timestamp=1521313135697, value=kolkata 2
column=finfo:source, timestamp=1521313092772, value=pune 2
column=fsch:at, timestamp=1521313166540, value=7.00a.m.
2 column=fsch:delay, timestamp=1521313229963, value=2 min
2 column=fsch:dt, timestamp=1521313202767, value=7.30a.m.
3 column=finfo:dest, timestamp=1521313310302, value=pune
3 column=finfo:source, timestamp=1521313290906, value=mumbai 3
column=fsch:at, timestamp=1521313333432, value=12.30p.m.
3 column=fsch:delay, timestamp=1521313379725, value=1 min
3 column=fsch:dt, timestamp=1521313353804, value=12.45p.m.
4 column=finfo:dest, timestamp=1521313419679, value=delhi
4 column=finfo:source, timestamp=1521313404831, value=mumbai 4
column=fsch:at, timestamp=1521313440328, value=2.00p.m.
4 column=fsch:delay, timestamp=1521313472389, value=10 min
4 column=fsch:dt, timestamp=1521313455226, value=2.45p.m. 4
row(s) in 0.0280 seconds
#Drop Table
#Create Table for dropping
hbase(main):046:0* create
'tb1','cf' 0 row(s) in 2.3120
seconds
=>
Hbase::Table -

tb1
hbase(main):04
7:0> list TABLE
flight
table1
table 2
4 row(s) in 0.0070 seconds

=> ["flight", "table1", "table2", "tb1"]
#Drop Table
hbase(main):048:0> drop 'tb1'

ERROR: Table tb1 is enabled. Disable it first.
Here is some help for this command:
Drop the named table. Table must first be disabled:
hbase> drop 't1'

```
hbase>
drop
'ns1:t1'
#Disabl
e table
hbase(main):049:0>
disable 'tb1' 0 row(s) in
4.3480 seconds
hbase(main):050:0>
drop 'tb1'
0 row(s) in 2.3540
seconds
hbase(main):051:0
> list TABLE
flight
table1
table2
3 row(s) in 0.0170 seconds

=> ["flight", "table1",
"table2"] #Read data from
table for row key 1:
hbase(main):052:0> get
'flight',1 COLUMN CELL
finfo:dest timestamp=1521312730758,
value=mumbai finfo:source
timestamp=1521312493881, value=pune fsch:at
timestamp=1521312789417, value=10.25a.m.
fsch:delay timestamp=1521312850594, value=5min
fsch:dt timestamp=1521312823256, value=11.25
a.m.
5 row(s) in 0.0450 seconds
Read data for particular column from HBase table:
hbase(main):053:0> get
'flight','1',COLUMN=>'finfo:source' COLUMN CELL
finfo:source timestamp=1521312493881,
value=pune 1 row(s) in 0.0110 seconds
Read data for multiple columns in HBase Table:
hbase(main):054:0> get 'flight','1',COLUMN=>['finfo:source','finfo:dest']
COLUMN CELL
finfo:dest timestamp=1521312730758,
value=mumbai finfo:source
timestamp=1521312493881, value=pune 2 row(s)
in 0.0190 seconds
hbase(main):055:0> scan
'flight',COLUMNS=>'finfo:source' ROW
COLUMN+CELL
1 column=finfo:source, timestamp=1521312493881, value=pune
2 column=finfo:source, timestamp=1521313092772, value=pune
3 column=finfo:source, timestamp=1521313290906, value=mumbai

4 column=finfo:source, timestamp=1521313404831, value=mumbai
4 row(s) in 0.0320 seconds
```

b) Creating an external Hive table to connect to the HBase for **Customer Information Table Covers===>**

c) Load table with data, insert new values and field in the table, Join tables with Hive

```
# Create the external table emp using hive

hive>create external table empdata2 ( ename string, esal int)
row format delimited fields terminated by "," stored as textfile location
"/home/hduser/Desktop/empdata2";
hive>load data local inpath '/home/hduser/Desktop/empdb.txt' into table empdata2;
```

#Create External Table in hive referring to hbase table

create hbase table emphive first
hbase(main):003:0> create
'emphive', 'cf' 0 row(s) in 4.6260
seconds
#create hive external table
CREATE external TABLE hive_table_emp(id int, name string, esal string) STORED
BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,cf:name,cf:esal")
TBLPROPERTIES ("hbase.table.name" = "emphive");

load data into hive_table_emp

(Hive doesn't allow directly inserting data into external hive table)
#for that create one hive table(managed table in hive)
Managed table and External table in Hive. There are two types of tables in Hive ,one is Managed table and second is external table. the difference is , when you drop a table, if it is managed table hive deletes both data and meta data,if it is external table Hive only deletes metadata.
hive>create table empdbnew(eno int, ename string, esal int) row format delimited fields
terminated
by ',' stored as
textfile; #load
data in managed
table
hive>load data local inpath '/home/hduser/Desktop/empdbnew.txt' into table empdbnew; #Load
data in external table from managed table.
hive>INSERT INTO hive_table_emp select * from empdbnew;
hive> select * from hive_table_emp;
OK
deepali120000
1 mahesh 30000
2 mangesh 25000
3 ram 39000
4 brijesh 40000
5 john 300000
Time taken: 0.52 seconds, Fetched: 6 row(s)
#display records where salary is greater than
40000
hive> select * from hive_table_emp where
esal>40000; OK
1 deepali120000
6 john 300000
Time taken: 0.546 seconds, Fetched: 2 row(s)

#Check hbase for updates(The records are available in associated Hbase table)
hbase(main):008:0> scan 'emphive'
ROW COLUMN+CELL
1 column=cf:esal, timestamp=1522212425665, value=120000
1 column=cf:name, timestamp=1522212425665, value=deepali
2 column=cf:esal, timestamp=1522212425665, value=30000
2 column=cf:name, timestamp=1522212425665, value=mahesh
3 column=cf:esal, timestamp=1522212425665, value=25000
3 column=cf:name, timestamp=1522212425665, value=mangesh
4 column=cf:esal, timestamp=1522212425665, value=39000
4 column=cf:name, timestamp=1522212425665, value=ram
5 column=cf:esal, timestamp=1522212425665, value=40000
5 column=cf:name, timestamp=1522212425665, value=brijesh
6 column=cf:esal, timestamp=1522212425665, value=300000
6 column=cf:name, timestamp=1522212425665,
value=john 6 row(s) in 0.0700 seconds

Creating external table in Hive referring to Hbase #referring to flight table
created in Hbase

CREATE external TABLE hbase_flight_new(fno int, fsource string,fdest string,fsh_at
string,fsh_dt string,fsch_delay

string,delay int)

STORED BY
'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH
SERDEPROPERTIES ("hbase.columns.mapping" =
":key,finfo:source,finfo:dest,fsch:at,fsch:dt,fsch:delay,delay:d
l") TBLPROPERTIES ("hbase.table.name" = "flight");
hive> CREATE external TABLE hbase_flight_new(fno int, fsource string,fdest string,fsh_at
string,fsh_dt string,fsch_delay string,delay int)
> STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
> WITH SERDEPROPERTIES ("hbase.columns.mapping"
=":key,finfo:source,finfo:dest,fsch:at,fsch:dt,fsch:delay,delay:dl")
> TBLPROPERTIES ("hbase.table.name" = "flight");
OK
Time taken:
0.361 seconds
#table created in
hive hive>
show tables;
OK
abc
ddl_hive
emp
empdata
empdata1
empdata2
empdbnew
hbase_flight
hbase_flight
1
hbase_flight
_new
hbase_table
_1
hive_table_e
mp
Time taken: 0.036 seconds, Fetched: 12
row(s) # Display records from that table
hive> select * from
hbase_flight_new; OK
1 pune mumbai 10.25a.m. 11.25 a.m. 5min 10
2 pune kolkata7.00a.m. 7.30a.m. 2 min 4
3 mumbai pune 12.30p.m. 12.45p.m. 1 min 5
4 mumbai delhi 2.00p.m. 2.45p.m. 10
min 16 Time taken: 0.581 seconds,
Fetched: 4 row(s)

D) Create index on Flight information Table

#create index on
hbase_flight_new CREATE
INDEX hbasefltnew_index ON
TABLE hbase_flight_new
(delay)
AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'
WITH DEFERRED REBUILD;
SHOW INDEX ON
hbase_flight_new; #create index on
table hbase_flight_new
hive> CREATE INDEX hbasefltnew_index
> ON TABLE hbase_flight_new (delay)
> AS 'org.apache.hadoop.hive.ql.index.compact.CompactIndexHandler'

> WITH DEFERRED
REBUILD; OK
Time taken: 0.74 seconds

#show index on table
hbase_flight_new hive> SHOW
INDEX ON hbase_flight_new; OK
hbasefltnew_index hbase_flight_new delay
default__hbase_flight_new_hbasefltnew_index_____compact
Time taken: 0.104 seconds, Fetched: 1 row(s)

#join two tables in Hive

#create table B for join
hive> create table empinfo(empno int, empgrade string) row format delimited fields terminated
by
',' stored as
textfile;
#Load Data
into table
hive> load data local inpath '/home/hduser/Desktop/empinfo.txt' into table empinfo; Loading data
to table default.empinfo
OK
Time taken: 0.552
seconds #insert
data into the table
hive> load data local inpath '/home/hduser/Desktop/empinfo.txt' into table empinfo; #
Table A empdbnew
hive> select * from
empdbnew; OK
1 deepali120000
2 mahesh 30000
3 mangesh 25000
4 ram 39000
5 brijesh 40000
6 john 300000
Time taken: 0.258 seconds, Fetched: 6
row(s) # Table B empinfo
hive> select * from
empinfo; OK
1 Time taken: 0.207 seconds, Fetched: 6 row(s)
#Join two tables(empdbnew with empinfo on empno)
hive> SELECT eno, ename, empno, empgrade FROM empdbnew JOIN empinfo ON eno = empno;
#Join==> Result

hive> SELECT eno, ename, empno, empgrade
> FROM empdbnew JOIN empinfo ON eno = empno;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions.
Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases. Query ID
= hduser_20180328153258_bc345f46-a1f1-4589-ac5e-4c463834731a
Total jobs = 1
Launching
Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1 In
order to change the average load for a reducer (in bytes):
set
hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of
reducers: set hive.exec.reducers.max=<number>
In order to set a constant number of
reducers: set
mapreduce.job.reduces=<number>
Starting Job = job_1522208646737_0005, Tracking URL =

http://localhost:8088/proxy/application_1522208646737_0005/

Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1522208646737_0005
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1 2018-
03-28 15:33:09,615 Stage-1 map = 0%, reduce = 0%
hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job_1522208646737_0003, Tracking URL =
http://localhost:8088/proxy/application_1522208646737_0003/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1522208646737_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1 2018-
03-28 13:00:20,256 Stage-1 map = 0%, reduce = 0%

2018-03-28 13:00:28,747 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.68 sec

2018-03-28 13:00:35,101 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.26 sec
MapReduce Total cumulative CPU time: 6 seconds 260 msec
Ended Job =
job_1522208646737_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.26 sec HDFS Read: 9095 HDFS Write:
102 SUCCESS

Total MapReduce CPU Time Spent: 6 seconds 260
msec OK

Time taken: 31.866 seconds, Fetched: 1 row(s) hive>

e) Find the average departure delay per day in 2008.

#calculate average delay

hive> select sum(delay) from hbase_flight_new;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions.
Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases. Query ID
= hduser_20180328130004_47384e9a-7490-4dfb-809d-ae240507bfab
Total jobs = 1
Launching
Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1522208646737_0003, Tracking URL =
http://localhost:8088/proxy/application_1522208646737_0003/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1522208646737_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1 2018-
03-28 13:00:20,256 Stage-1 map = 0%, reduce = 0%
2018-03-28 13:00:28,747 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.68 sec
2018-03-28 13:00:35,101 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.26 sec
MapReduce Total cumulative CPU time: 6 seconds 260 msec
Ended Job =
job_1522208646737_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.26 sec HDFS Read: 9095 HDFS Write:
102 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 260
msec OK
Time taken: 31.866 seconds, Fetched: 1
row(s) hive>

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('adult_dataset.csv')
```

In [3]:
```
df
```

Out[3]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationshi
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-chil
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husban
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husban
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husban
4	18	?	103497	Some-college	10	Never-married	?	Own-chil
...	...	...	...	...	...	...	...	.
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wif
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husban
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarrie
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-chil
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wif

48842 rows × 15 columns

In [4]:
```
df.describe()
```

Out[4]:

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	48842.000000	4.884200e+04	48842.000000	48842.000000	48842.000000	48842.000000
mean	38.643585	1.896641e+05	10.078089	1079.067626	87.502314	40.422382
std	13.710510	1.056040e+05	2.570973	7452.019058	403.004552	12.391444
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.175505e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.781445e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.376420e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   age              48842 non-null  int64 
 1   workclass        48842 non-null  object
 2   fnlwgt           48842 non-null  int64 
 3   education        48842 non-null  object
 4   educational-num  48842 non-null  int64 
 5   marital-status   48842 non-null  object
 6   occupation       48842 non-null  object
 7   relationship     48842 non-null  object
 8   race             48842 non-null  object
 9   gender           48842 non-null  object
 10  capital-gain     48842 non-null  int64 
 11  capital-loss     48842 non-null  int64 
 12  hours-per-week   48842 non-null  int64 
 13  native-country   48842 non-null  object
 14  income           48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
```

In [6]:
```
# First subset: Like and Share
df_subset_1 = df[['workclass','education','capital-gain']]
df_subset_1
```

Out[6]:

	workclass	education	capital-gain
0	Private	11th	0
1	Private	HS-grad	0
2	Local-gov	Assoc-acdm	0
3	Private	Some-college	7688
4	?	Some-college	0
...	...	...	...
48837	Private	Assoc-acdm	0
48838	Private	HS-grad	0
48839	Private	HS-grad	0
48840	Private	HS-grad	0
48841	Self-emp-inc	HS-grad	15024

48842 rows × 3 columns

In [7]:
```
# second subset: Comment and Type
df_subset_2 = df[['race','native-country']]
df_subset_2
```

Out[7]:

	race	native-country
0	Black	United-States
1	White	United-States
2	White	United-States
3	Black	United-States
4	White	United-States
...	...	...
48837	White	United-States
48838	White	United-States
48839	White	United-States
48840	White	United-States
48841	White	United-States

48842 rows × 2 columns

In [8]:
```
merged_data  = pd.concat([df_subset_1,df_subset_2],axis=1)
merged_data
```

Out[8]:

	workclass	education	capital-gain	race	native-country
0	Private	11th	0	Black	United-States
1	Private	HS-grad	0	White	United-States
2	Local-gov	Assoc-acdm	0	White	United-States
3	Private	Some-college	7688	Black	United-States
4	?	Some-college	0	White	United-States
...	...	...	...	...	...
48837	Private	Assoc-acdm	0	White	United-States
48838	Private	HS-grad	0	White	United-States
48839	Private	HS-grad	0	White	United-States
48840	Private	HS-grad	0	White	United-States
48841	Self-emp-inc	HS-grad	15024	White	United-States

48842 rows × 5 columns

In [9]:
```
# Sorting merged_data in descending order wrt 'capital-gain'
merged_data.sort_values(by=['capital-gain'],ascending=False)
```

Out[9]:

	workclass	education	capital-gain	race	native-country
28936	Self-emp-inc	HS-grad	99999	White	?
18384	Self-emp-inc	Some-college	99999	White	United-States
34689	Self-emp-inc	Prof-school	99999	White	United-States
34744	Private	Prof-school	99999	White	United-States
48519	Private	Prof-school	99999	White	United-States
...	...	...	...	...	...
16967	?	7th-8th	0	White	United-States
16968	Local-gov	Bachelors	0	White	United-States
16969	Private	Masters	0	White	Dominican-Republic
16970	Private	HS-grad	0	White	United-States
24421	Private	HS-grad	0	White	United-States

48842 rows × 5 columns

In [10]:
```
# Method 1
merged_data.transpose()
```

Out[10]:

	0	1	2	3	4	5	6	7	8	
workclass	Private	Private	Local-gov	Private	?	Private	?	Self-emp-not-inc	Private	P
education	11th	HS-grad	Assoc-acdm	Some-college	Some-college	10th	HS-grad	Prof-school	Some-college	7t
capital-gain	0	0	0	7688	0	0	0	3103	0	
race	Black	White	White	Black	White	White	Black	White	White	
native-country	United-States	United-States	United-States	United-States	United-States	United-States	United-States	United-States	United-States	U

5 rows × 48842 columns

In [11]:
```
# Method 2
merged_data.T
```

Out[11]:

	0	1	2	3	4	5	6	7	8	
workclass	Private	Private	Local-gov	Private	?	Private	?	Self-emp-not-inc	Private	P
education	11th	HS-grad	Assoc-acdm	Some-college	Some-college	10th	HS-grad	Prof-school	Some-college	7t
capital-gain	0	0	0	7688	0	0	0	3103	0	
race	Black	White	White	Black	White	White	Black	White	White	
native-country	United-States	United-States	United-States	United-States	United-States	United-States	United-States	United-States	United-States	U

5 rows × 48842 columns

In [12]:
```
df
```

Out[12]:

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationshi
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-chil
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husban
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husban
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husban
4	18	?	103497	Some-college	10	Never-married	?	Own-chil
...	...	...	...	...	...	...	...	.
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wif
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husban
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarrie
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-chil
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wif

48842 rows × 15 columns

In [13]:
```
# Reshape
pd.melt(df, id_vars =['education'], value_vars =['capital-gain'])
```

Out[13]:

	education	variable	value
0	11th	capital-gain	0
1	HS-grad	capital-gain	0
2	Assoc-acdm	capital-gain	0
3	Some-college	capital-gain	7688
4	Some-college	capital-gain	0
...	...	...	...
48837	Assoc-acdm	capital-gain	0
48838	HS-grad	capital-gain	0
48839	HS-grad	capital-gain	0
48840	HS-grad	capital-gain	0
48841	HS-grad	capital-gain	15024

48842 rows × 3 columns

In [ ]:

In [1]:
```
import pandas as pd
```

In [3]:
```
df = pd.read_csv('dataset_Facebook.csv',delimiter=';')
```

In [4]:
```
df
```

Out[4]:

	Page total likes	Type	Category	Post Month	Post Weekday	Post Hour	Paid	Lifetime Post Total Reach	Lifetime Post Total Impressions	Lifet Enga U
0	139441	Photo	2	12	4	3	0.0	2752	5091	
1	139441	Status	2	12	3	10	0.0	10460	19057	1
2	139441	Photo	3	12	3	3	0.0	2413	4373	
3	139441	Photo	2	12	2	10	1.0	50128	87991	2
4	139441	Photo	2	12	2	3	0.0	7244	13594	
...	...	...	...	...	...	...	...	...	...	
495	85093	Photo	3	1	7	2	0.0	4684	7536	
496	81370	Photo	2	1	5	8	0.0	3480	6229	
497	81370	Photo	1	1	5	2	0.0	3778	7216	
498	81370	Photo	3	1	4	11	0.0	4156	7564	
499	81370	Photo	2	1	4	4	NaN	4188	7292	

500 rows × 19 columns

In [5]:
```
df.describe()
```

Out[5]:

	Page total likes	Category	Post Month	Post Weekday	Post Hour	Paid	Life Post R
count	500.000000	500.000000	500.000000	500.000000	500.000000	499.000000	500.0
mean	123194.176000	1.880000	7.038000	4.150000	7.840000	0.278557	13903.3
std	16272.813214	0.852675	3.307936	2.030701	4.368589	0.448739	22740.7
min	81370.000000	1.000000	1.000000	1.000000	1.000000	0.000000	238.0
25%	112676.000000	1.000000	4.000000	2.000000	3.000000	0.000000	3315.0
50%	129600.000000	2.000000	7.000000	4.000000	9.000000	0.000000	5281.0
75%	136393.000000	3.000000	10.000000	6.000000	11.000000	1.000000	13168.0
max	139441.000000	3.000000	12.000000	7.000000	23.000000	1.000000	180480.0

In [6]:
```
# First subset: Like and Share
df_subset_1 = df[['like','share']]
df_subset_1
```

Out[6]:

	like	share
0	79.0	17.0
1	130.0	29.0
2	66.0	14.0
3	1572.0	147.0
4	325.0	49.0
...	...	...
495	53.0	26.0
496	53.0	22.0
497	93.0	18.0
498	91.0	38.0
499	91.0	28.0

500 rows × 2 columns

In [7]:
```
# second subset: Comment and Type
df_subset_2 = df[['comment','Type']]
df_subset_2
```

Out[7]:

	comment	Type
0	4	Photo
1	5	Status
2	0	Photo
3	58	Photo
4	19	Photo
...	...	...
495	5	Photo
496	0	Photo
497	4	Photo
498	7	Photo
499	0	Photo

500 rows × 2 columns

In [8]:
```
merged_data  = pd.merge(df_subset_2, df_subset_1, left_on='comment', right_on= 'lik
merged_data
```

Out[8]:

	comment	Type	like	share
0	4	Photo	4.0	2.0
1	4	Photo	4.0	1.0
2	4	Photo	4.0	0.0
3	4	Photo	4.0	1.0
4	5	Status	5.0	2.0
...	...	...	...	...
1462	0	Photo	0.0	0.0
1463	0	Photo	0.0	0.0
1464	0	Photo	0.0	0.0
1465	0	Photo	0.0	0.0
1466	0	Photo	0.0	0.0

1467 rows × 4 columns

In [9]:
```
# Define a dictionary containing employee data
data1 = {
'key': ['K0', 'K1', 'K2', 'K3'],
'Name':['Jai', 'Princi', 'Gaurav', 'Anuj'],
'Age':[27, 24, 22, 32],}
# Define a dictionary containing employee data
data2 = {
'key': ['K0', 'K1', 'K2', 'K3'],
'Address':['Nagpur', 'Kanpur', 'Allahabad', 'Kannuaj'],
'Qualification':['Btech', 'B.A', 'Bcom', 'B.hons']}
# Convert the dictionary into DataFrame
data1 = pd.DataFrame(data1)
# Convert the dictionary into DataFrame
data2 = pd.DataFrame(data2)

# print(df, "\n\n", df1)
res = pd.merge(data1, data2, on='key')
res
```

Out[9]:

	key	Name	Age	Address	Qualification
0	K0	Jai	27	Nagpur	Btech
1	K1	Princi	24	Kanpur	B.A
2	K2	Gaurav	22	Allahabad	Bcom
3	K3	Anuj	32	Kannuaj	B.hons

In [10]:
```
# Sorting merged_data in descending order wrt 'Like'
merged_data.sort_values(by=['like'],ascending=False)
```

Out[10]:

	comment	Type	like	share
1351	146	Photo	146.0	9.0
1352	146	Photo	146.0	15.0
549	144	Photo	144.0	10.0
550	144	Photo	144.0	29.0
711	64	Photo	64.0	19.0
...	...	...	...	...
891	0	Photo	0.0	0.0
892	0	Photo	0.0	0.0
895	0	Photo	0.0	0.0
896	0	Photo	0.0	0.0
1466	0	Photo	0.0	0.0

1467 rows × 4 columns

In [11]:
```
# Method 1
merged_data.transpose()
```

Out[11]:

	0	1	2	3	4	5	6	7	8	9	...	145
comment	4	4	4	4	5	0	0	0	0	0	...	
Type	Photo	Photo	Photo	Photo	Status	Photo	Photo	Photo	Photo	Photo	...	Phot
like	4.0	4.0	4.0	4.0	5.0	0.0	0.0	0.0	0.0	0.0	...	7
share	2.0	1.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	...	2

4 rows × 1467 columns

In [12]:
```
# Method 2
merged_data.T
```

Out[12]:

	0	1	2	3	4	5	6	7	8	9	...	145
comment	4	4	4	4	5	0	0	0	0	0	...	
Type	Photo	Photo	Photo	Photo	Status	Photo	Photo	Photo	Photo	Photo	...	Pho
like	4.0	4.0	4.0	4.0	5.0	0.0	0.0	0.0	0.0	0.0	...	7
share	2.0	1.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	...	2

4 rows × 1467 columns

In [13]:
```
df
```

Out[13]:

	Page total likes	Type	Category	Post Month	Post Weekday	Post Hour	Paid	Lifetime Post Total Reach	Lifetime Post Total Impressions	Lifet Enga U
0	139441	Photo	2	12	4	3	0.0	2752	5091	
1	139441	Status	2	12	3	10	0.0	10460	19057	1
2	139441	Photo	3	12	3	3	0.0	2413	4373	
3	139441	Photo	2	12	2	10	1.0	50128	87991	2
4	139441	Photo	2	12	2	3	0.0	7244	13594	
...	...	...	...	...	...	...	...	...	...	
495	85093	Photo	3	1	7	2	0.0	4684	7536	
496	81370	Photo	2	1	5	8	0.0	3480	6229	
497	81370	Photo	1	1	5	2	0.0	3778	7216	
498	81370	Photo	3	1	4	11	0.0	4156	7564	
499	81370	Photo	2	1	4	4	NaN	4188	7292	

500 rows × 19 columns

In [14]:
```
df.Type.unique()
```

Out[14]:
```
array(['Photo', 'Status', 'Link', 'Video'], dtype=object)
```

In [15]:
```
# Reshape
# Comment is id_vars and Type is value_vars
pd.melt(df, id_vars =['Type'], value_vars =['comment'])
```

Out[15]:

	Type	variable	value
0	Photo	comment	4
1	Status	comment	5
2	Photo	comment	0
3	Photo	comment	58
4	Photo	comment	19
...	...	...	...
495	Photo	comment	5
496	Photo	comment	0
497	Photo	comment	4
498	Photo	comment	7
499	Photo	comment	0

500 rows × 3 columns

In [16]:
```
# Reshape
df_temp = pd.DataFrame({'foo': ['one', 'one', 'one', 'two', 'two', 'two'],
                        'bar': ['A', 'B', 'C', 'A', 'B', 'C'],
                        'baz': [1, 2, 3, 4, 5, 6],
                        'zoo': ['x', 'y', 'z', 'q', 'w', 't']})
df_temp
```

Out[16]:

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

In [17]:
```
df_temp.pivot(index='foo', columns='bar', values='baz')
```

Out[17]:

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

In [ ]:

In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
df = pd.read_csv('airquality_data.csv', encoding='cp1252', low_memory=False)
```

In [4]:
```
df.head()
```

Out[4]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm
0	150	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN
1	151	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN
2	152	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN
3	150	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN
4	151	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   stn_code                     291665 non-null  object 
 1   sampling_date                435739 non-null  object 
 2   state                        435742 non-null  object 
 3   location                     435739 non-null  object 
 4   agency                       286261 non-null  object 
 5   type                         430349 non-null  object 
 6   so2                          401096 non-null  float64
 7   no2                          419509 non-null  float64
 8   rspm                         395520 non-null  float64
 9   spm                          198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object 
 11  pm2_5                        9314 non-null    float64
 12  date                         435735 non-null  object 
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

In [6]:
```
df.columns
```

Out[6]:
```
Index(['stn_code', 'sampling_date', 'state', 'location', 'agency', 'type',
       'so2', 'no2', 'rspm', 'spm', 'location_monitoring_station', 'pm2_5',
       'date'],
      dtype='object')
```

In [7]:
```
# Change data type from float64 to float32 for Space Complexity
df['so2'] = df['so2'].astype('float32')
df['no2'] = df['no2'].astype('float32')
df['rspm'] = df['rspm'].astype('float32')
df['spm'] = df['spm'].astype('float32')
df['date'] = df['date'].astype('string')

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   stn_code                     291665 non-null  object 
 1   sampling_date                435739 non-null  object 
 2   state                        435742 non-null  object 
 3   location                     435739 non-null  object 
 4   agency                       286261 non-null  object 
 5   type                         430349 non-null  object 
 6   so2                          401096 non-null  float32
 7   no2                          419509 non-null  float32
 8   rspm                         395520 non-null  float32
 9   spm                          198355 non-null  float32
 10  location_monitoring_station  408251 non-null  object 
 11  pm2_5                        9314 non-null    float64
 12  date                         435735 non-null  string 
dtypes: float32(4), float64(1), object(7), string(1)
memory usage: 36.6+ MB
```

In [8]:
```
df=df.drop_duplicates()
```

In [9]:
```
df.isna().sum()
```

Out[9]:
```
stn_code                     144077
sampling_date                     3
state                             0
location                          3
agency                       149466
type                           5357
so2                           34632
no2                           16222
rspm                          40035
spm                          236908
location_monitoring_station   27303
pm2_5                        425754
date                              7
dtype: int64
```

In [10]:
```
percent_missing = df.isnull().sum() * 100 / len(df)
```

In [11]:
```
percent_missing.sort_values(ascending=False)
```

Out[11]:
```
pm2_5                            97.859185
spm                              54.453097
agency                           34.354630
stn_code                         33.115973
rspm                              9.202010
so2                               7.960135
location_monitoring_station       6.275571
no2                               3.728613
type                              1.231302
date                              0.001609
sampling_date                     0.000690
location                          0.000690
state                             0.000000
dtype: float64
```

In [12]:
```
df=df.drop(['stn_code', 'agency','sampling_date','location_monitoring_station','pm2
```

In [13]:
```
df.head()
```

Out[13]:

	state	location	type	so2	no2	rspm	spm	date
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	1990-02-01
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	NaN	NaN	1990-02-01
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	1990-02-01
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	1990-03-01
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	NaN	NaN	1990-03-01

In [14]:
```
df.columns
```

Out[14]:
```
Index(['state', 'location', 'type', 'so2', 'no2', 'rspm', 'spm', 'date'], dtype='o
bject')
```

In [15]:
```
col_var = ['state', 'location', 'type','date']
col_num = ['so2','no2','rspm','spm']
```

In [16]:
```
for col in df.columns:
    if df[col].dtype == 'object' or df[col].dtype == 'string':
        df[col] = df[col].fillna(df[col].mode()[0])
    else:
        df[col] = df[col].fillna(df[col].mean())
```

In [17]:
```
df.isna().sum()
```

Out[17]:
```
state       0
location    0
type        0
so2         0
no2         0
rspm        0
spm         0
date        0
dtype: int64
```

In [18]:
```
df
```

Out[18]:

	state	location	type	so2	no2	rspm	spm
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.800000	17.400000	108.871712	220.774796
1	Andhra Pradesh	Hyderabad	Industrial Area	3.100000	7.000000	108.871712	220.774796
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.200000	28.500000	108.871712	220.774796
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.300000	14.700000	108.871712	220.774796
4	Andhra Pradesh	Hyderabad	Industrial Area	4.700000	7.500000	108.871712	220.774796
...	...	...	...	...	...	...	...
435737	West Bengal	ULUBERIA	RIRUO	22.000000	50.000000	143.000000	220.774796
435738	West Bengal	ULUBERIA	RIRUO	20.000000	46.000000	171.000000	220.774796
435739	andaman-and-nicobar-islands	Guwahati	Residential, Rural and other Areas	10.830467	25.823299	108.871712	220.774796
435740	Lakshadweep	Guwahati	Residential, Rural and other Areas	10.830467	25.823299	108.871712	220.774796
435741	Tripura	Guwahati	Residential, Rural and other Areas	10.830467	25.823299	108.871712	220.774796

435068 rows × 8 columns

In [19]:
```
df.isna().sum()
```

Out[19]:
```
state       0
location    0
type        0
so2         0
no2         0
rspm        0
spm         0
date        0
dtype: int64
```

In [20]:
```
subSet1 = df[['state', 'type']]
subSet2 = df[['state','location']]
```

In [21]:
```
subSet1.head()
```

Out[21]:

	state	type
0	Andhra Pradesh	Residential, Rural and other Areas
1	Andhra Pradesh	Industrial Area
2	Andhra Pradesh	Residential, Rural and other Areas
3	Andhra Pradesh	Residential, Rural and other Areas
4	Andhra Pradesh	Industrial Area

In [22]:
```
subSet2.head()
```

Out[22]:

	state	location
0	Andhra Pradesh	Hyderabad
1	Andhra Pradesh	Hyderabad
2	Andhra Pradesh	Hyderabad
3	Andhra Pradesh	Hyderabad
4	Andhra Pradesh	Hyderabad

In [23]:
```
concatenated_df = pd.concat([subSet1, subSet2], axis=1)
```

In [24]:
```
concatenated_df
```

Out[24]:

	state	type	state	location
0	Andhra Pradesh	Residential, Rural and other Areas	Andhra Pradesh	Hyderabad
1	Andhra Pradesh	Industrial Area	Andhra Pradesh	Hyderabad
2	Andhra Pradesh	Residential, Rural and other Areas	Andhra Pradesh	Hyderabad
3	Andhra Pradesh	Residential, Rural and other Areas	Andhra Pradesh	Hyderabad
4	Andhra Pradesh	Industrial Area	Andhra Pradesh	Hyderabad
...	...	...	...	...
435737	West Bengal	RIRUO	West Bengal	ULUBERIA
435738	West Bengal	RIRUO	West Bengal	ULUBERIA
435739	andaman-and-nicobar-islands	Residential, Rural and other Areas	andaman-and-nicobar-islands	Guwahati
435740	Lakshadweep	Residential, Rural and other Areas	Lakshadweep	Guwahati
435741	Tripura	Residential, Rural and other Areas	Tripura	Guwahati

435068 rows × 4 columns

In [25]:
```
def remove_outliers(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    threshold = 1.5 * IQR
    outlier_mask = (column < Q1 - threshold) | (column > Q3 + threshold)
    return column[~outlier_mask]
```

In [26]:
```
df.columns
```

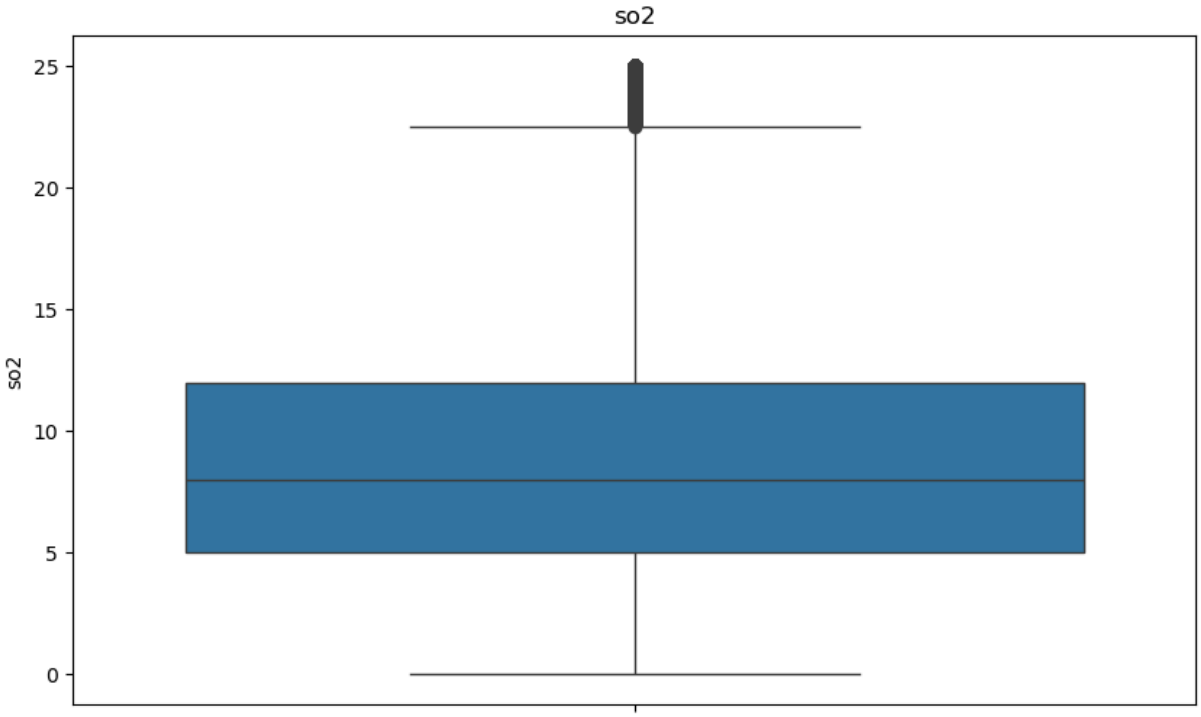
Out[26]:
```
Index(['state', 'location', 'type', 'so2', 'no2', 'rspm', 'spm', 'date'], dtype='o
bject')
```

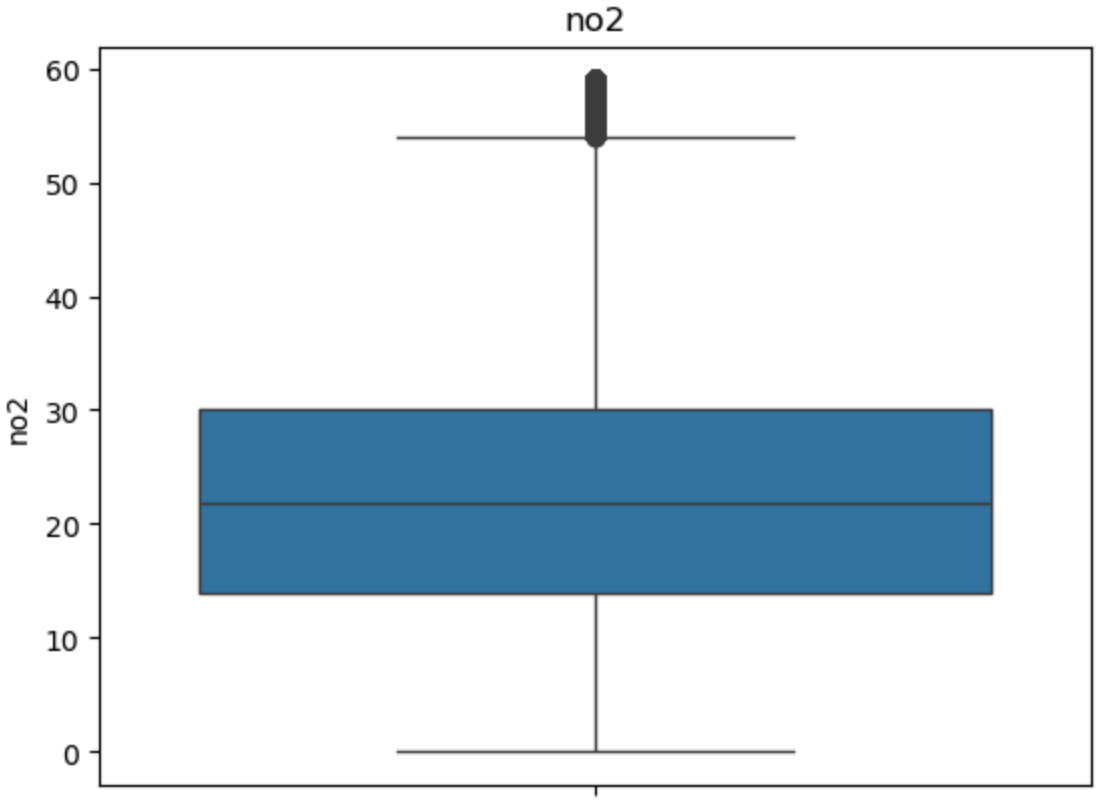
In [27]:
```
# Remove outliers for each column using a loop
col_name = ['so2', 'no2', 'rspm', 'spm']
for col in col_name:
    df[col] = remove_outliers(df[col])
```

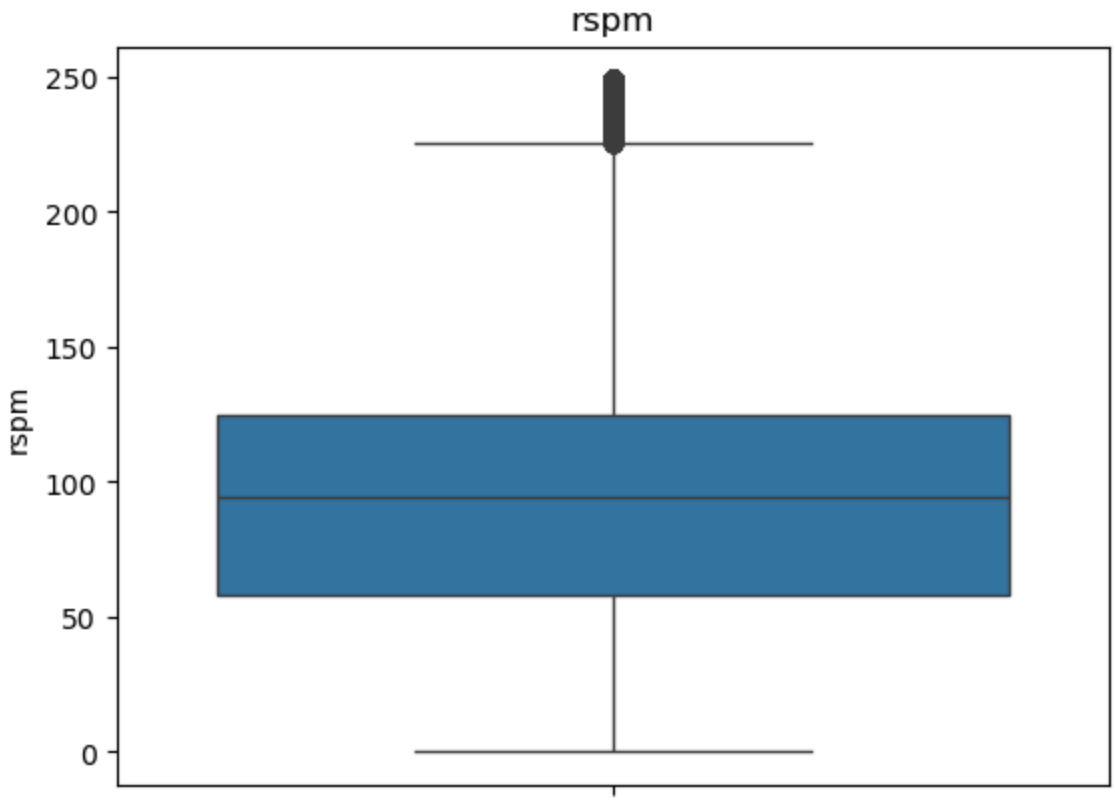
In [28]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

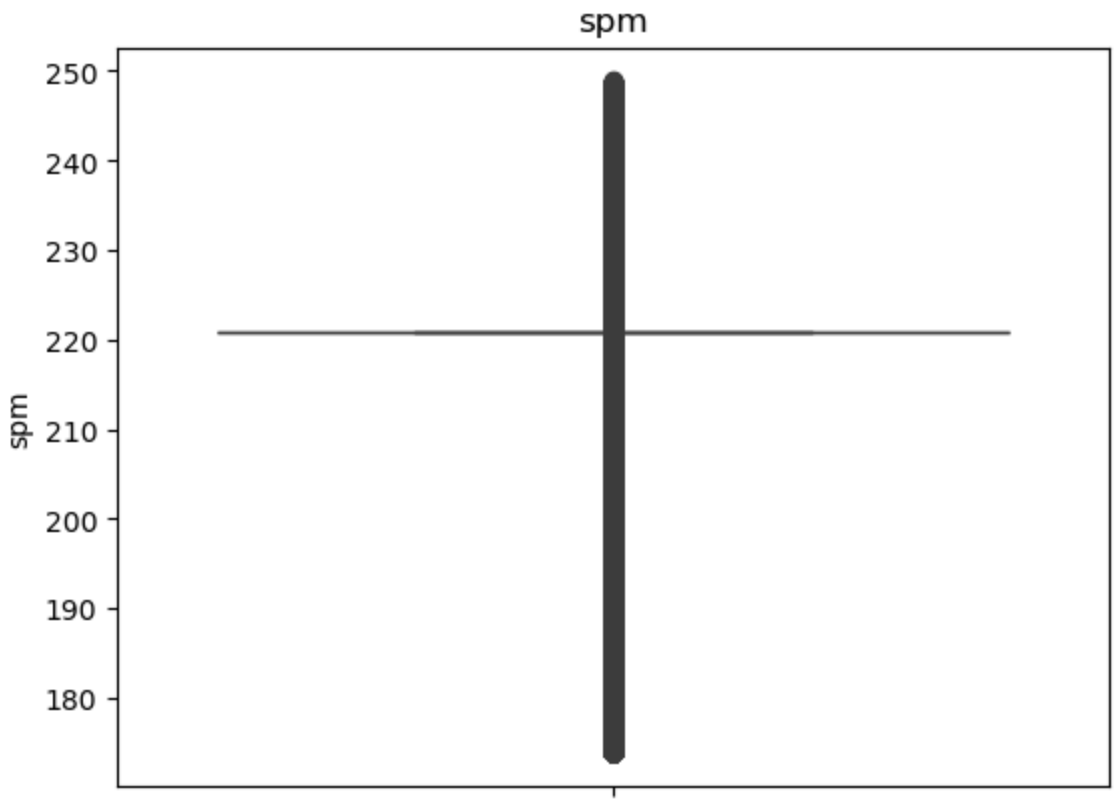
In [29]:
```
plt.figure(figsize=(10, 6))  # Adjust the figure size if needed

for col in col_name:
```

```
    sns.boxplot(data=df[col])
    plt.title(col)
    plt.show()
```









In [30]:
```
from sklearn.preprocessing import LabelEncoder

col_label= ['state','location','type']
# Initialize LabelEncoder

encoder = LabelEncoder()
```

```
# Iterate over columns
for col in df.columns:
        # Fit and transform the column
        df[col] = encoder.fit_transform(df[col])
```

In [31]:
```
df
```

Out[31]:

	state	location	type	so2	no2	rspm	spm	date
0	0	114	6	446	1489	2030	464	213
1	0	114	1	197	250	2030	464	213
2	0	114	6	790	3096	2030	464	213
3	0	114	6	823	1144	2030	464	214
4	0	114	1	427	301	2030	464	214
...	...	...	...	...	...	...	...	...
435737	35	282	3	2888	5307	2534	464	5059
435738	35	282	3	2809	5113	3098	464	5064
435739	36	100	6	1638	2696	2030	464	4779
435740	17	100	6	1638	2696	2030	464	4779
435741	31	100	6	1638	2696	2030	464	4779

435068 rows × 8 columns

In [ ]:
```

```

In [1]:
```
# import pandas library
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score,confusion_matrix
from sklearn.linear_model import LogisticRegression
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
# Reading csv file
df = pd.read_csv("Heart.csv")
df.head()
```

Out[2]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	out
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	

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

In [4]:
```
# Count ,min,max ,etc of each column
df.describe()
```

Out[4]:

	age	sex	cp	trtbps	chol	fbs	restecg
count	302.00000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.42053	0.682119	0.963576	131.602649	246.500000	0.149007	0.526490
std	9.04797	0.466426	1.032044	17.563394	51.753489	0.356686	0.526027
min	29.00000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	48.00000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	55.50000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000
75%	61.00000	1.000000	2.000000	140.000000	274.750000	0.000000	1.000000
max	77.00000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

In [5]:
```
# Information about each column data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 302 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       302 non-null    int64  
 1   sex       302 non-null    int64  
 2   cp        302 non-null    int64  
 3   trtbps    302 non-null    int64  
 4   chol      302 non-null    int64  
 5   fbs       302 non-null    int64  
 6   restecg   302 non-null    int64  
 7   thalachh  302 non-null    int64  
 8   exng      302 non-null    int64  
 9   oldpeak   302 non-null    float64
 10  slp       302 non-null    int64  
 11  caa       302 non-null    int64  
 12  thall     302 non-null    int64  
 13  output    302 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 35.4 KB
```

In [6]:
```
#Finding null values in each column
df.isna().sum()
```

Out[6]:
```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

In [7]:
```
df.head()
```

Out[7]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	out
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	

In [8]:
```
df.fbs.unique()
```

Out[8]:
```
array([1, 0], dtype=int64)
```

In [9]:
```
subSet1 = df[['age','cp','chol','thalachh']]
```

In [10]:
```
subSet2 = df[['exng','slp','output']]
```

In [11]:
```
merged_df = subSet1.merge(right=subSet2,how='cross')
merged_df.head()
```

Out[11]:

	age	cp	chol	thalachh	exng	slp	output
0	63	3	233	150	0	0	1
1	63	3	233	150	0	0	1
2	63	3	233	150	0	2	1
3	63	3	233	150	0	2	1
4	63	3	233	150	1	2	1

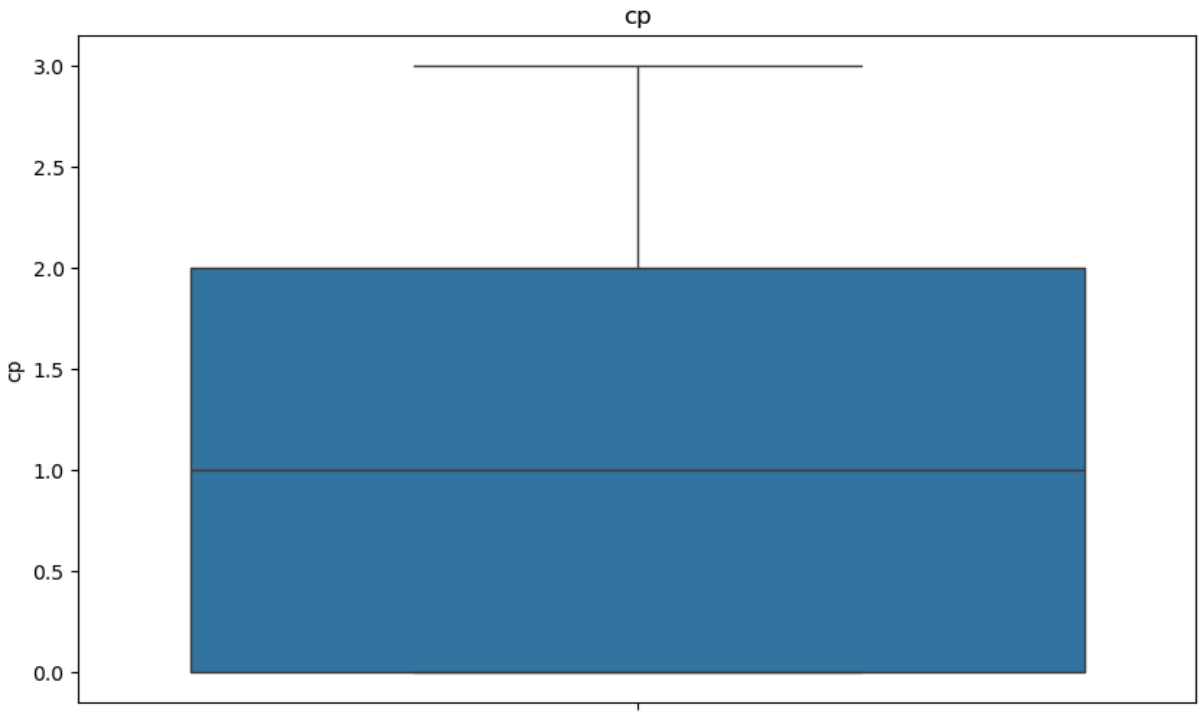
In [12]:
```
df.columns
```

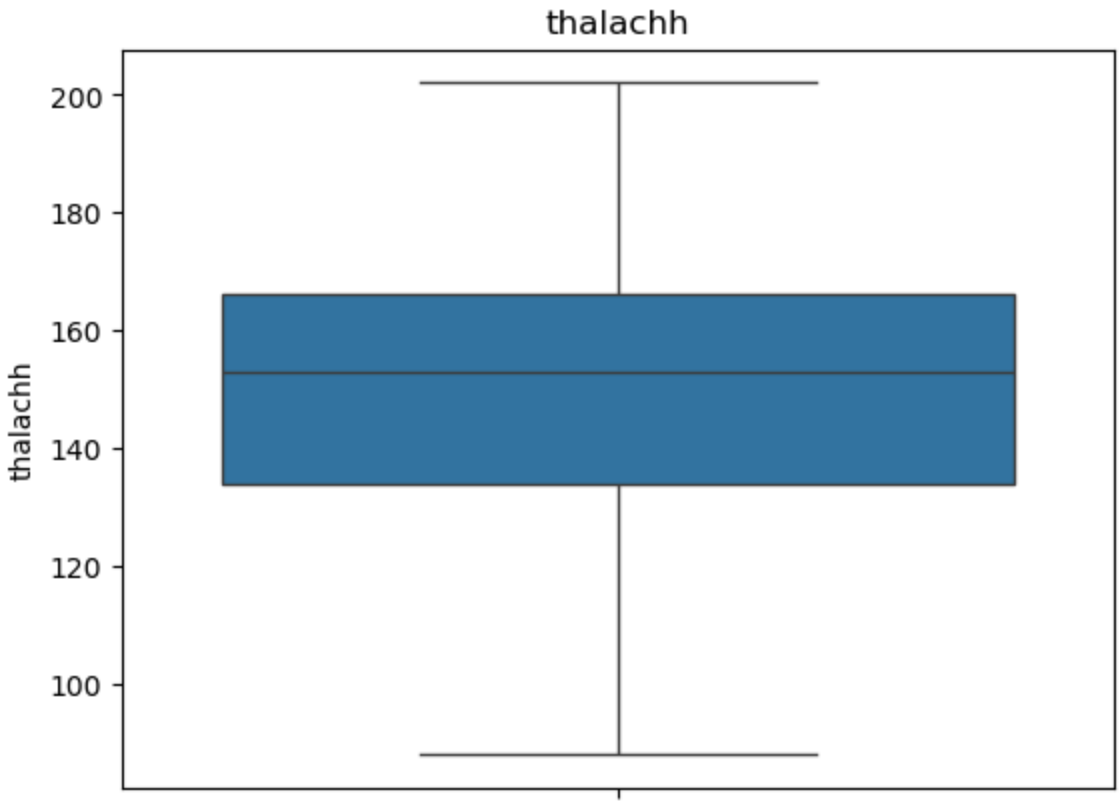
Out[12]:
```
Index(['age', 'sex', 'cp', 'trtbps', 'chol', 'fbs', 'restecg', 'thalachh',
       'exng', 'oldpeak', 'slp', 'caa', 'thall', 'output'],
      dtype='object')
```

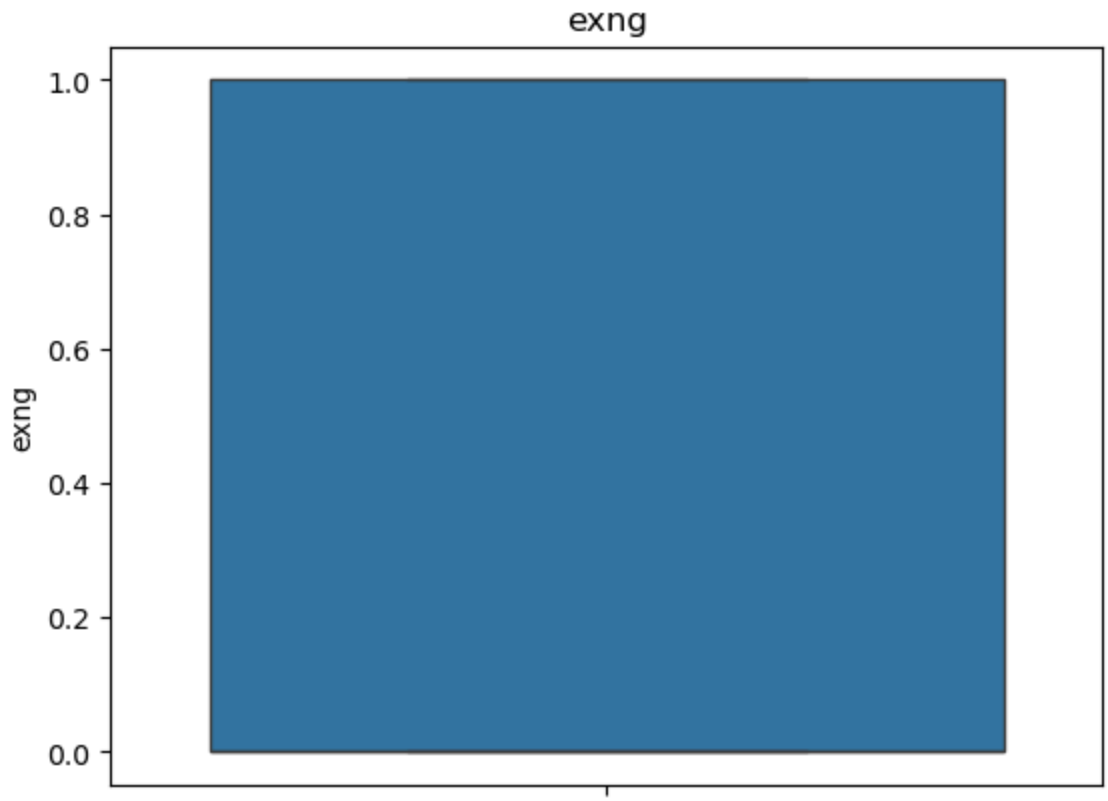
In [13]:
```
def remove_outliers(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    threshold = 1.5 * IQR
    outlier_mask = (column < Q1 - threshold) | (column > Q3 + threshold)
    return column[~outlier_mask]
```

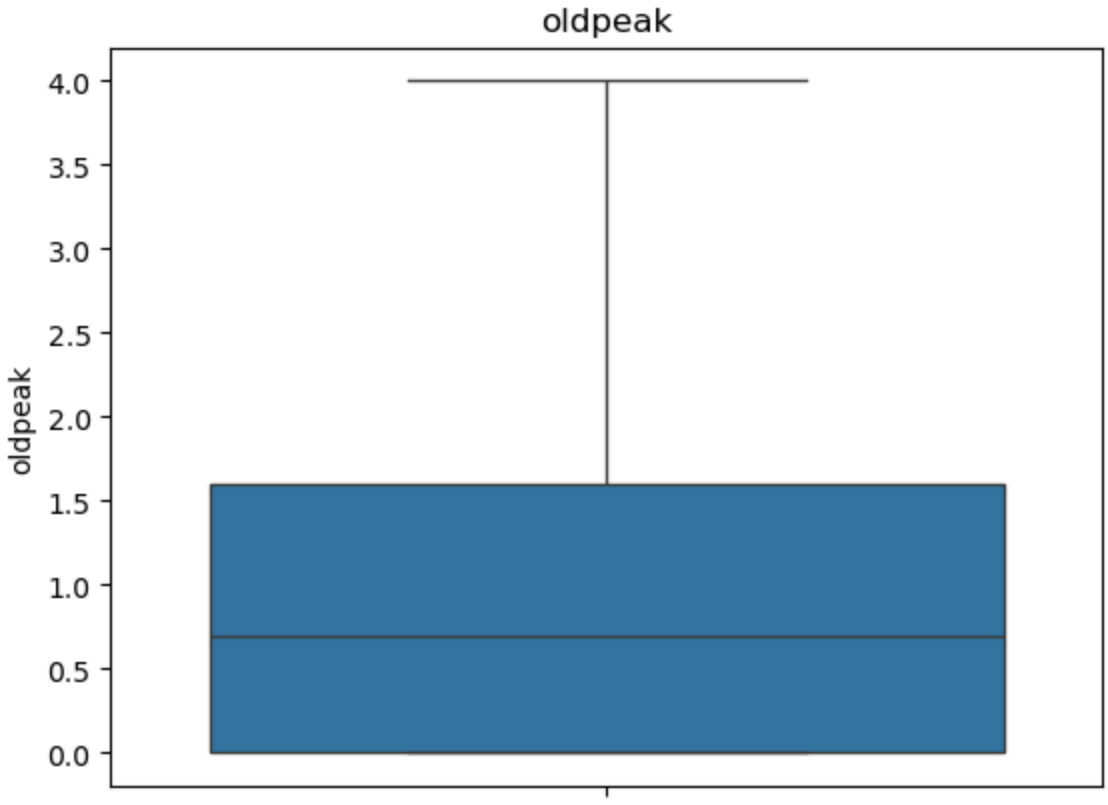
In [14]:
```
# Remove outliers for each column using a loop
col_name = ['cp','thalachh','exng','oldpeak','slp','caa']
for col in col_name:
    df[col] = remove_outliers(df[col])
```

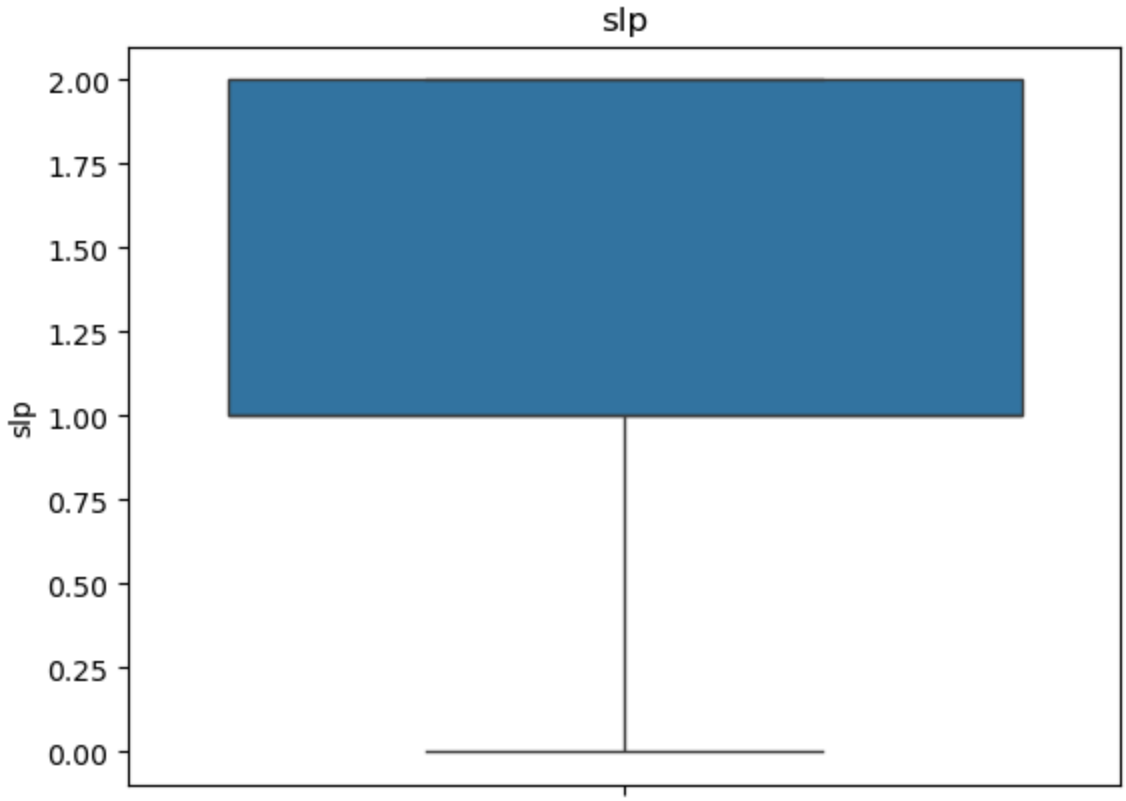
In [15]:
```
plt.figure(figsize=(10, 6))  # Adjust the figure size if needed

for col in col_name:
    sns.boxplot(data=df[col])
    plt.title(col)
    plt.show()
```

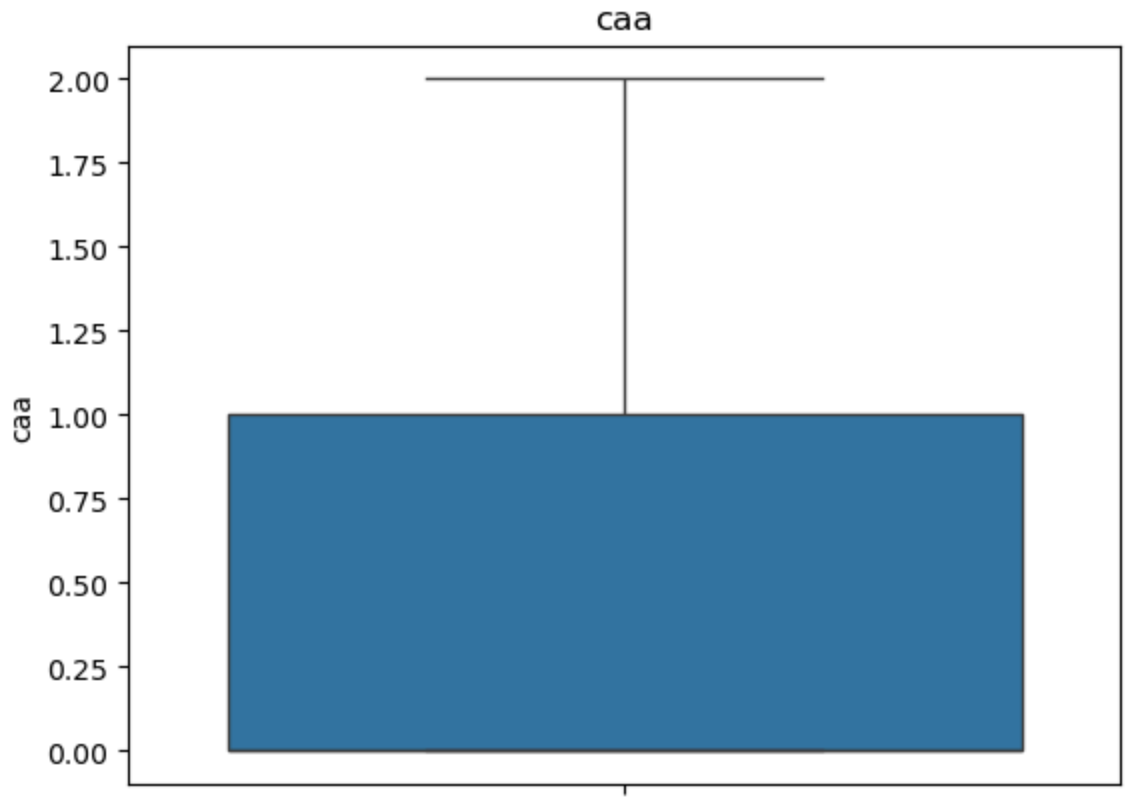





exng
exng
1.0
0.8
0.6
0.4
0.2
0.0


oldpeak
oldpeak
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0





In [16]:
```
df = df.dropna()
```

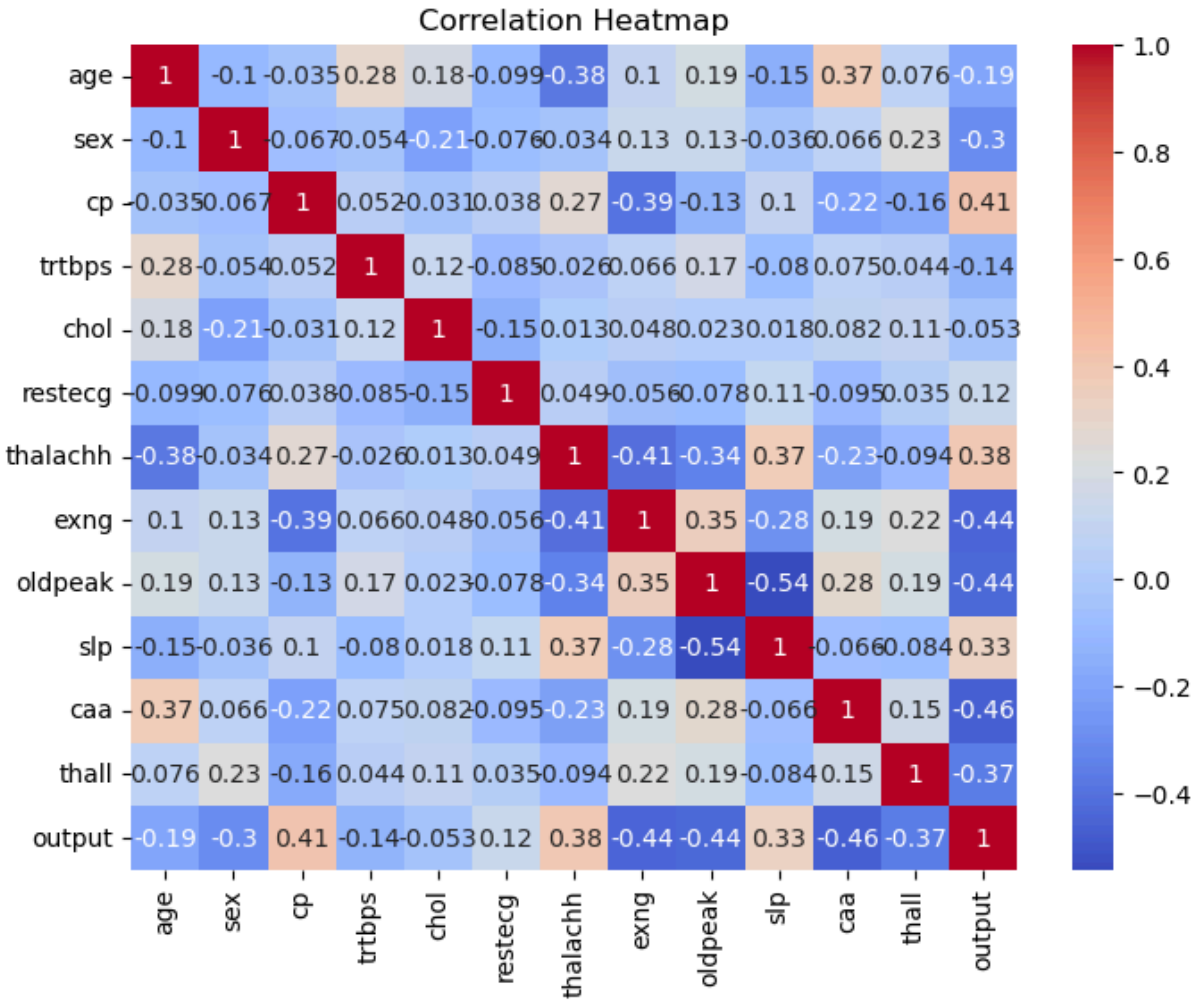
In [17]:
```
df.isna().sum()
```

Out[17]:
```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

In [18]:
```
df = df.drop('fbs',axis=1)
```

In [19]:
```
# Compute correlations between features and target
correlations = df.corr()['output'].drop('output')

# Print correlations
print("Correlation with the Target:")
print(correlations)
print()

# Plot correlation heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

```
Correlation with the Target:
age        -0.193798
sex        -0.303271
cp          0.410807
trtbps     -0.135238
chol       -0.052796
restecg     0.122071
thalachh    0.384609
exng       -0.444401
oldpeak    -0.437895
slp         0.329432
caa        -0.460816
thall      -0.366390
Name: output, dtype: float64
```



In [20]:
```
# df.isna().sum()
```

In [21]:
```
# splitting data using train test split
x = df[['cp','thalachh','exng','oldpeak','slp','caa']]
y = df.output
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)

x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

Out[21]:
```
((220, 6), (55, 6), (220,), (55,))
```

In [22]:
```
from sklearn.preprocessing import StandardScaler
```

In [23]:
```
scaler = StandardScaler()
```

In [24]:
```
x_train_scaled = scaler.fit_transform(x_train)
x_test_scaled = scaler.transform(x_test)
```

In [25]:
```
y_train= np.array(y_train).reshape(-1, 1)
y_test= np.array(y_test).reshape(-1, 1)
```

In [26]:
```
y_train.shape
```

Out[26]: (220, 1)

In [29]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Train the model with corrected label shape
model = LogisticRegression()
model.fit(x_train_scaled, y_train.ravel())

# Predict on the test set
y_pred = model.predict(x_test_scaled)

# Evaluate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.8363636363636363
```

In [30]:
```
#Classification model using Decision Tree
from sklearn.tree import DecisionTreeClassifier
tc=DecisionTreeClassifier(criterion='entropy')
tc.fit(x_train_scaled,y_train)
y_pred=tc.predict(x_test_scaled)

print("Training Accuracy Score :",accuracy_score(y_pred,y_test))
print("Training Confusion Matrix  :",confusion_matrix(y_pred,y_test))
```

```
Training Accuracy Score : 0.7818181818181819
Training Confusion Matrix  : [[21  6]
 [ 6 22]]
```

In [ ]:

1. Average Burned Area by Month

mapper.py

```
#!/usr/bin/env python
import sys

for line in sys.stdin:
    if line.startswith('X,Y,month'):
        continue
    fields = line.strip().split(',')
    month = fields[2]
    try:
        area = float(fields[-1])
    except ValueError:
        continue
    print(f"{month}\t{area}")
```

reducer.py

```
#!/usr/bin/env python
import sys
from collections import defaultdict

monthly_totals = defaultdict(list)

for line in sys.stdin:
    month, area = line.strip().split('\t')
    monthly_totals[month].append(float(area))

for month, areas in monthly_totals.items():
    avg_area = sum(areas) / len(areas)
    print(f"{month}\t{avg_area}")
```

Command to Run

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar \
-file mapper.py -mapper mapper.py \
-file reducer.py -reducer reducer.py \
-input /user/hadoop/forestfires.csv \
-output /user/hadoop/output/monthly_avg_area
```

Output

```
jan     0.45
feb     1.23
mar     12.34
apr     8.90
may     5.67
jun     7.89
jul     10.23
aug     45.68
sep     32.46
oct     6.78
nov     2.34
dec     3.46
```

2. Temperature-Area Correlation

Temp_mapper.py

```
#!/usr/bin/env python
import sys

for line in sys.stdin:
    if line.startswith('X,Y,month'):
        continue
    fields = line.strip().split(',')
    try:
        temp = float(fields[8])
        area = float(fields[-1])
    except ValueError:
        continue
    print(f"temp_area\t{temp}\t{area}")
```

Temp_reducer.py

```
#!/usr/bin/env python
import sys
import math

n = 0
sum_x = 0
```

```
sum_y = 0
sum_xy = 0
sum_x2 = 0
sum_y2 = 0

for line in sys.stdin:
    _, temp, area = line.strip().split('\t')
    temp = float(temp)
    area = float(area)

    n += 1
    sum_x += temp
    sum_y += area
    sum_xy += temp * area
    sum_x2 += temp ** 2
    sum_y2 += area ** 2

if n > 0:
    numerator = sum_xy - (sum_x * sum_y)/n
    denominator = math.sqrt((sum_x2 - (sum_x**2)/n) * (sum_y2 - (sum_y**2)/n))
    correlation = numerator / denominator if denominator != 0 else 0
    print(f"Correlation: {correlation:.4f}")
```

Output

```
Correlation: 0.7245
```

Hive Data Mining

1. Table Creation

```
CREATE EXTERNAL TABLE forest_fires (
    x INT, y INT, month STRING, day STRING,
    ffmc DOUBLE, dmc DOUBLE, dc DOUBLE, isi DOUBLE,
    temp DOUBLE, rh DOUBLE, wind DOUBLE, rain DOUBLE,
    area DOUBLE
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION '/user/hadoop/forestfires'
TBLPROPERTIES ("skip.header.line.count"="1");
```

2. Top Dangerous Months

```
SELECT month, AVG(area) as avg_area
FROM forest_fires
GROUP BY month
ORDER BY avg_area DESC
LIMIT 5;
```

Output

```
month  avg_area
aug    45.68
sep    32.46
jul    10.23
mar    12.34
apr    8.90
```

3. Risk Analysis

```
SELECT
   CASE
      WHEN temp > 30 AND wind > 20 AND rh < 30 THEN 'High Risk'
      WHEN temp > 25 AND wind > 15 AND rh < 40 THEN 'Medium Risk'
      ELSE 'Low Risk'
   END as risk_level,
   COUNT(*) as cases,
   AVG(area) as avg_area
FROM forest_fires
GROUP BY risk_level;
```

Output

```
risk_level  cases  avg_area
High Risk   45     78.90
Medium Risk 120    34.56
Low Risk    235    5.67
```

4. Day-wise Analysis

```
SELECT day, AVG(area) as avg_area
FROM forest_fires
GROUP BY day
ORDER BY avg_area DESC;
```

Output

day avg_area
fri 25.67
sun 22.45
sat 18.90
thu 15.34
mon 12.56
tue 10.23
wed 8.90

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv('AirQuality_visualization.csv',delimiter=';')
```

In [3]:
```
df
```

Out[3]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0
...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN

9471 rows × 17 columns

In [4]:
```
df = df.rename(columns={'T': 'Temperature'})
df = df.rename(columns={'RH': 'Relative Humidity'})
df = df.rename(columns={'AH': 'Absolute Humidity'})
df
```

Out[4]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0
...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN

9471 rows × 17 columns

In [5]:
```
df = df.drop(['Unnamed: 15', 'Unnamed: 16'],axis=1)
```

In [6]:
```
df
```

Out[6]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0
...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN

9471 rows × 15 columns

In [7]:
```
df['CO(GT)'] = df['CO(GT)'].str.replace(',', '.').astype(float)
df['C6H6(GT)'] = df['C6H6(GT)'].str.replace(',', '.').astype(float)
df['Temperature'] = df['Temperature'].str.replace(',', '.').astype(float)
df['Relative Humidity'] = df['Relative Humidity'].str.replace(',', '.').astype(floa
df['Absolute Humidity'] = df['Absolute Humidity'].str.replace(',', '.').astype(floa
df
```

Out[7]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0
...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN

9471 rows × 15 columns

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9471 entries, 0 to 9470
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Date               9357 non-null   object 
 1   Time               9357 non-null   object 
 2   CO(GT)             9357 non-null   float64
 3   PT08.S1(CO)        9357 non-null   float64
 4   NMHC(GT)           9357 non-null   float64
 5   C6H6(GT)           9357 non-null   float64
 6   PT08.S2(NMHC)      9357 non-null   float64
 7   NOx(GT)            9357 non-null   float64
 8   PT08.S3(NOx)       9357 non-null   float64
 9   NO2(GT)            9357 non-null   float64
 10  PT08.S4(NO2)       9357 non-null   float64
 11  PT08.S5(O3)        9357 non-null   float64
 12  Temperature        9357 non-null   float64
 13  Relative Humidity  9357 non-null   float64
 14  Absolute Humidity  9357 non-null   float64
dtypes: float64(13), object(2)
memory usage: 1.1+ MB
```

In [9]:
```
df.head()
```

Out[9]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NO
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0	
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0	
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0	
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0	
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0	

In [10]:
```
df=df.drop_duplicates()
```

In [11]:
```
df.isna().sum()
```

Out[11]:
```
Date                 1
Time                 1
CO(GT)               1
PT08.S1(CO)          1
NMHC(GT)             1
C6H6(GT)             1
PT08.S2(NMHC)        1
NOx(GT)              1
PT08.S3(NOx)         1
NO2(GT)              1
PT08.S4(NO2)         1
PT08.S5(O3)          1
Temperature          1
Relative Humidity    1
Absolute Humidity    1
dtype: int64
```

In [15]:
```
numeric_cols = df.select_dtypes(include='number').columns
df.loc[:, numeric_cols] = df[numeric_cols].fillna(df[numeric_cols].mean())
df = df.dropna()
```

In [16]:
```
df.isna().sum()
```

Out[16]:
```
Date                 0
Time                 0
CO(GT)               0
PT08.S1(CO)          0
NMHC(GT)             0
C6H6(GT)             0
PT08.S2(NMHC)        0
NOx(GT)              0
PT08.S3(NOx)         0
NO2(GT)              0
PT08.S4(NO2)         0
PT08.S5(O3)          0
Temperature          0
Relative Humidity    0
Absolute Humidity    0
dtype: int64
```

In [17]:
```
df
```

Out[17]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0
...	...	...	...	...	...	...	...
9352	04/04/2005	10.00.00	3.1	1314.0	-200.0	13.5	1101.0
9353	04/04/2005	11.00.00	2.4	1163.0	-200.0	11.4	1027.0
9354	04/04/2005	12.00.00	2.4	1142.0	-200.0	12.4	1063.0
9355	04/04/2005	13.00.00	2.1	1003.0	-200.0	9.5	961.0
9356	04/04/2005	14.00.00	2.2	1071.0	-200.0	11.9	1047.0

9357 rows × 15 columns

In [18]:
```
df['Absolute Humidity'] = df['Absolute Humidity'].multiply(100)
df
```

Out[18]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
0	10/03/2004	18.00.00	2.6	1360.0	150.0	11.9	1046.0
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.4	955.0
2	10/03/2004	20.00.00	2.2	1402.0	88.0	9.0	939.0
3	10/03/2004	21.00.00	2.2	1376.0	80.0	9.2	948.0
4	10/03/2004	22.00.00	1.6	1272.0	51.0	6.5	836.0
...	...	...	...	...	...	...	...
9352	04/04/2005	10.00.00	3.1	1314.0	-200.0	13.5	1101.0
9353	04/04/2005	11.00.00	2.4	1163.0	-200.0	11.4	1027.0
9354	04/04/2005	12.00.00	2.4	1142.0	-200.0	12.4	1063.0
9355	04/04/2005	13.00.00	2.1	1003.0	-200.0	9.5	961.0
9356	04/04/2005	14.00.00	2.2	1071.0	-200.0	11.9	1047.0

9357 rows × 15 columns

In [19]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [20]:
```
def remove_outliers(column):
    Q1 = column.quantile(0.25)
    Q3 = column.quantile(0.75)
    IQR = Q3 - Q1
    threshold = 1.5 * IQR
    outlier_mask = (column < Q1 - threshold) | (column > Q3 + threshold)
    return column[~outlier_mask]
```

In [21]:
```
df.columns
```

Out[21]:
```
Index(['Date', 'Time', 'CO(GT)', 'PT08.S1(CO)', 'NMHC(GT)', 'C6H6(GT)',
       'PT08.S2(NMHC)', 'NOx(GT)', 'PT08.S3(NOx)', 'NO2(GT)', 'PT08.S4(NO2)',
       'PT08.S5(O3)', 'Temperature', 'Relative Humidity', 'Absolute Humidity'],
      dtype='object')
```

In [22]:
```
# Remove outliers for each column using a loop
col_name = ['Temperature','Relative Humidity','Absolute Humidity','PT08.S4(NO2)','P
       'PT08.S2(NMHC)','PT08.S1(CO)']
for col in col_name:
    df[col] = remove_outliers(df[col])
```

In [24]:
```
df['Date'] = pd.to_datetime(df['Date'], dayfirst=True, errors='coerce')
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
```

In [25]:
```
df
```

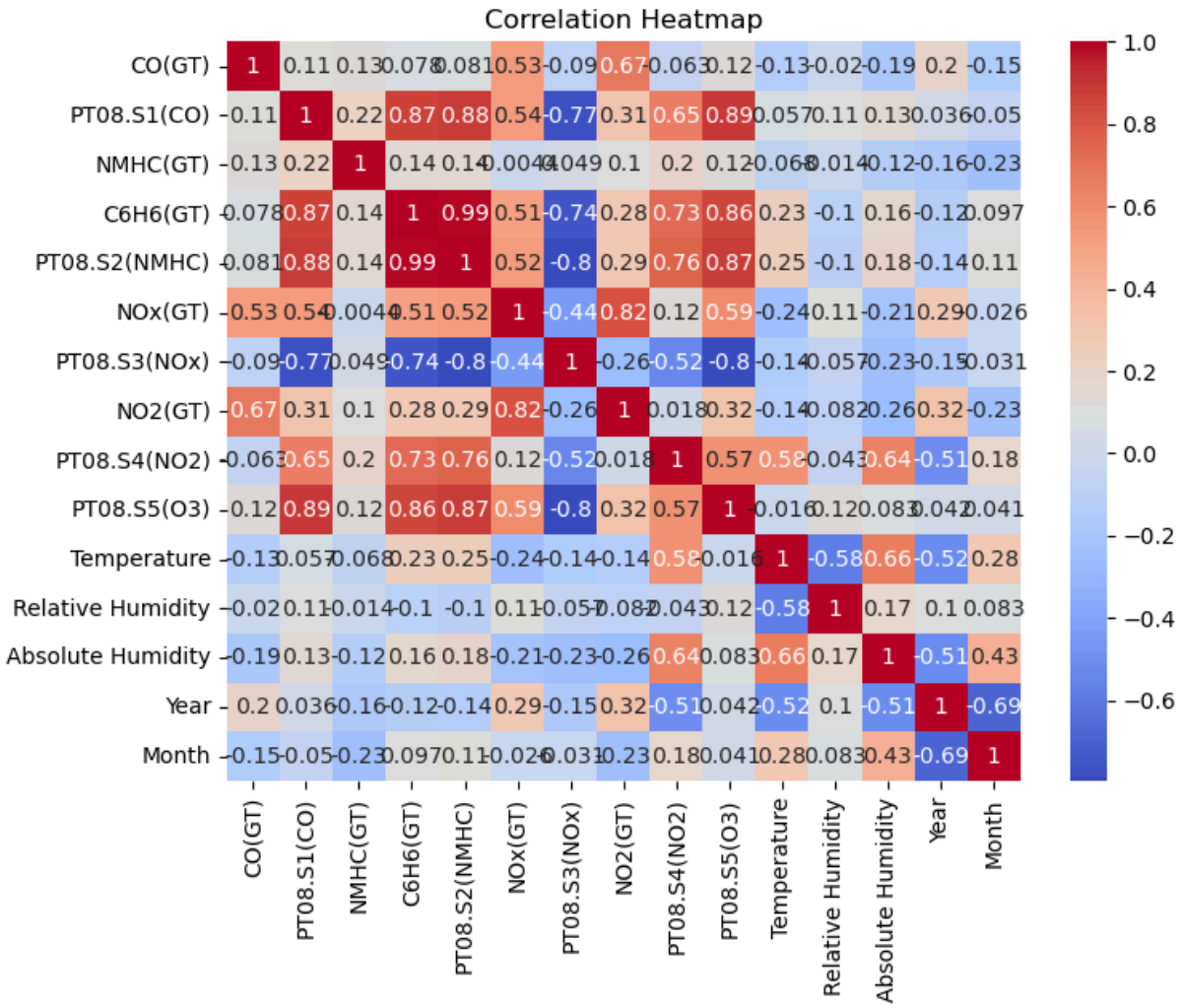
Out[25]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(
0	2004-03-10	18.00.00	2.6	1360.0	150.0	11.9	1046.0	1(
1	2004-03-10	19.00.00	2.0	1292.0	112.0	9.4	955.0	1(
2	2004-03-10	20.00.00	2.2	1402.0	88.0	9.0	939.0	1:
3	2004-03-10	21.00.00	2.2	1376.0	80.0	9.2	948.0	1'
4	2004-03-10	22.00.00	1.6	1272.0	51.0	6.5	836.0	1:
...	...	...	...	...	...	...	...	
9352	2005-04-04	10.00.00	3.1	1314.0	-200.0	13.5	1101.0	4'
9353	2005-04-04	11.00.00	2.4	1163.0	-200.0	11.4	1027.0	3:
9354	2005-04-04	12.00.00	2.4	1142.0	-200.0	12.4	1063.0	2'
9355	2005-04-04	13.00.00	2.1	1003.0	-200.0	9.5	961.0	2:
9356	2005-04-04	14.00.00	2.2	1071.0	-200.0	11.9	1047.0	2(

9357 rows × 17 columns

In [26]:
```
df['yearr']= df.Year.astype(str)
df['month']= df.Month.astype(str)
```

In [28]:
```
# Select only numeric columns
numeric_df = df.select_dtypes(include=['number'])

# Plot the correlation heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

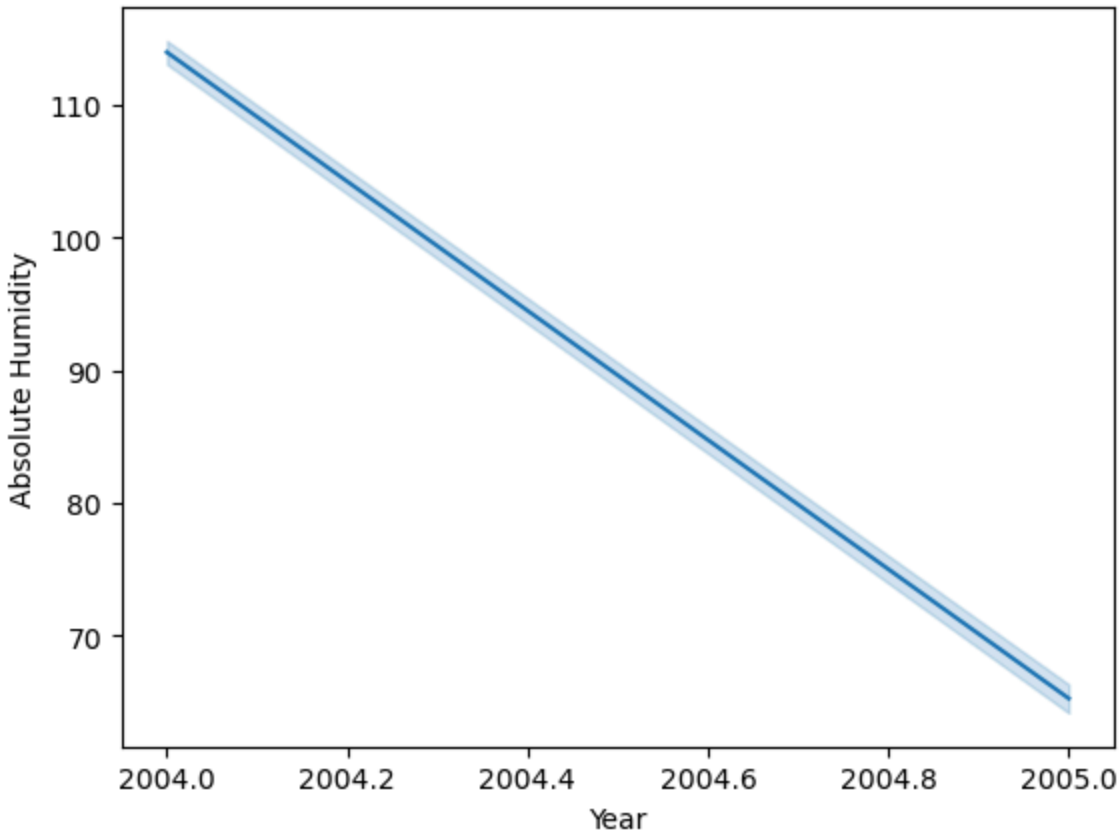


In [29]:
```
df.columns
```

Out[29]:
```
Index(['Date', 'Time', 'CO(GT)', 'PT08.S1(CO)', 'NMHC(GT)', 'C6H6(GT)',
       'PT08.S2(NMHC)', 'NOx(GT)', 'PT08.S3(NOx)', 'NO2(GT)', 'PT08.S4(NO2)',
       'PT08.S5(O3)', 'Temperature', 'Relative Humidity', 'Absolute Humidity',
       'Year', 'Month', 'yearr', 'month'],
      dtype='object')
```

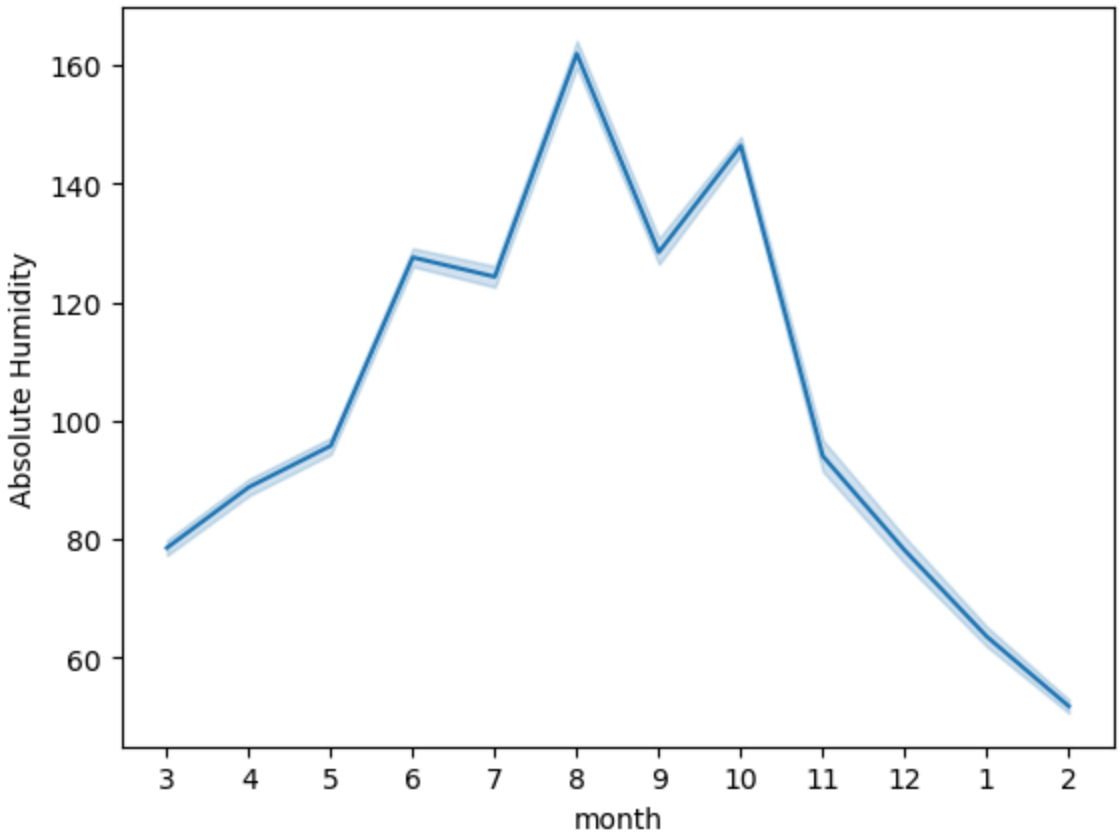
In [30]:
```
sns.lineplot(df,x="Year",y='Absolute Humidity')
```

Out[30]:
```
<Axes: xlabel='Year', ylabel='Absolute Humidity'>
```



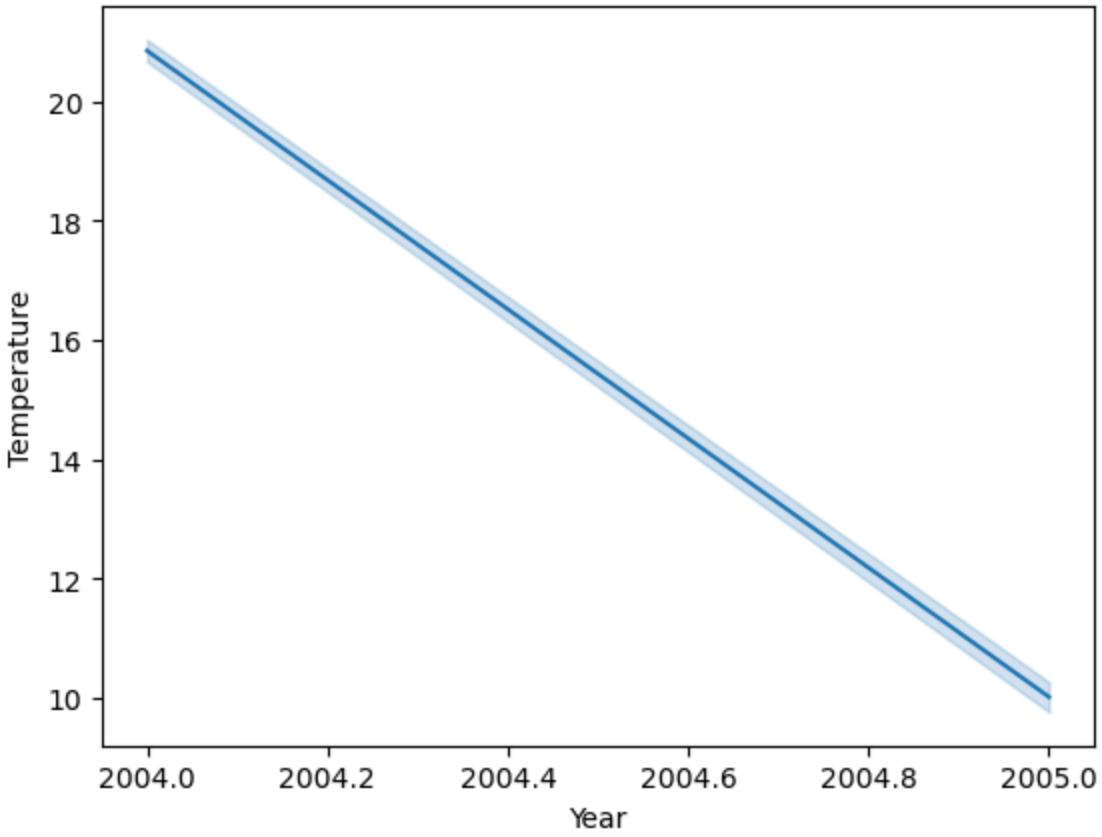
In [31]:
```
sns.lineplot(df,x="month",y='Absolute Humidity')
```

Out[31]: <Axes: xlabel='month', ylabel='Absolute Humidity'>



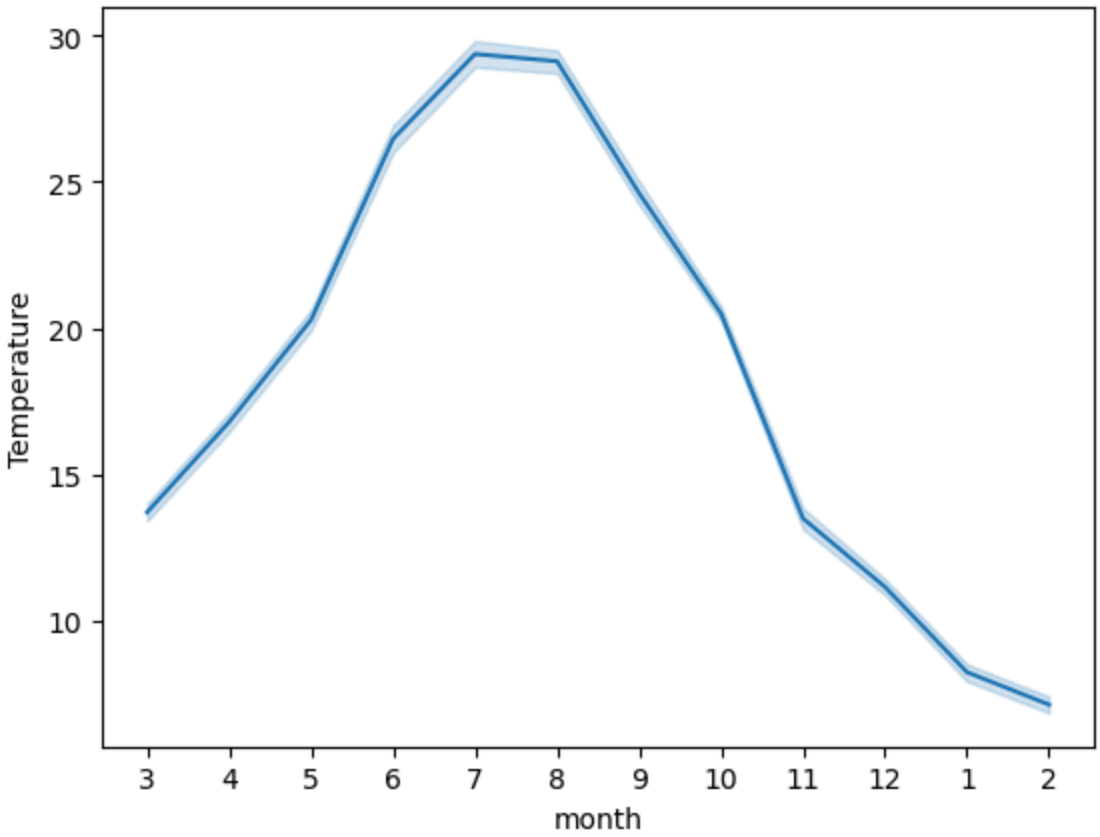
In [32]:
```
sns.lineplot(df,x="Year",y='Temperature')
```

Out[32]: <Axes: xlabel='Year', ylabel='Temperature'>



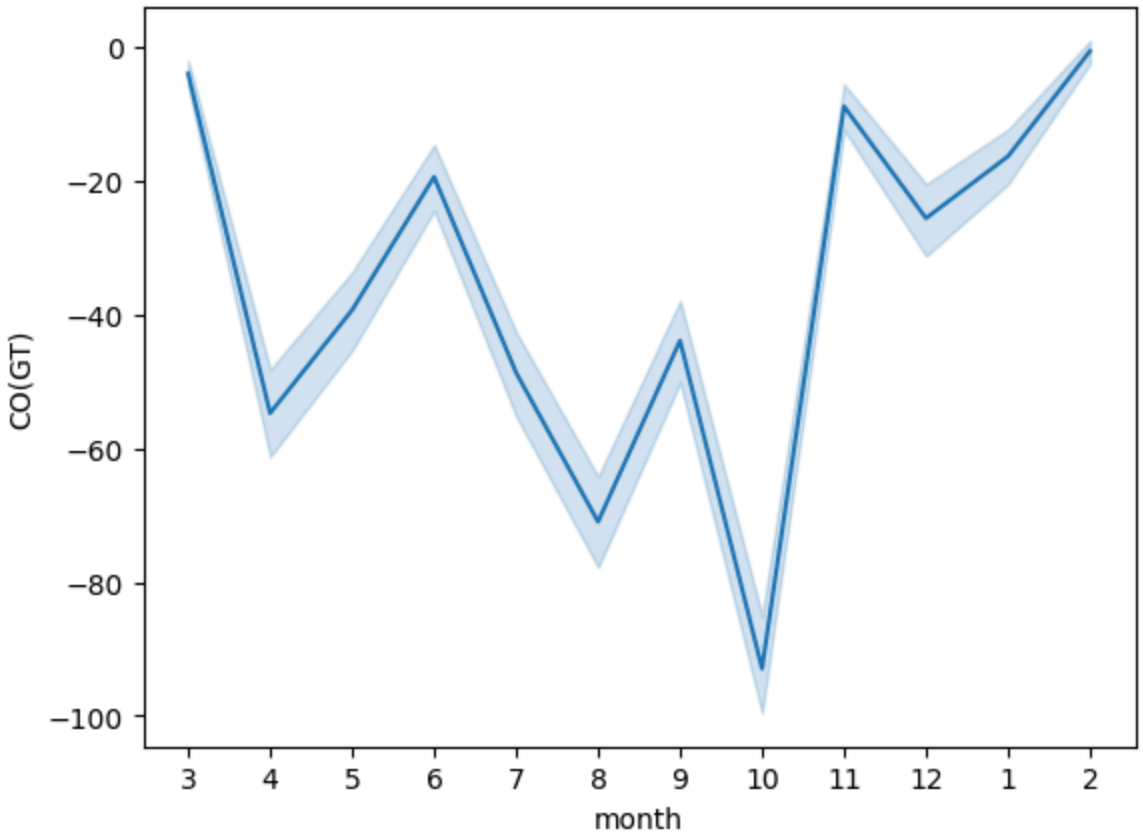
In [33]:
```
sns.lineplot(df,x="month",y='Temperature',)
```

Out[33]: <Axes: xlabel='month', ylabel='Temperature'>



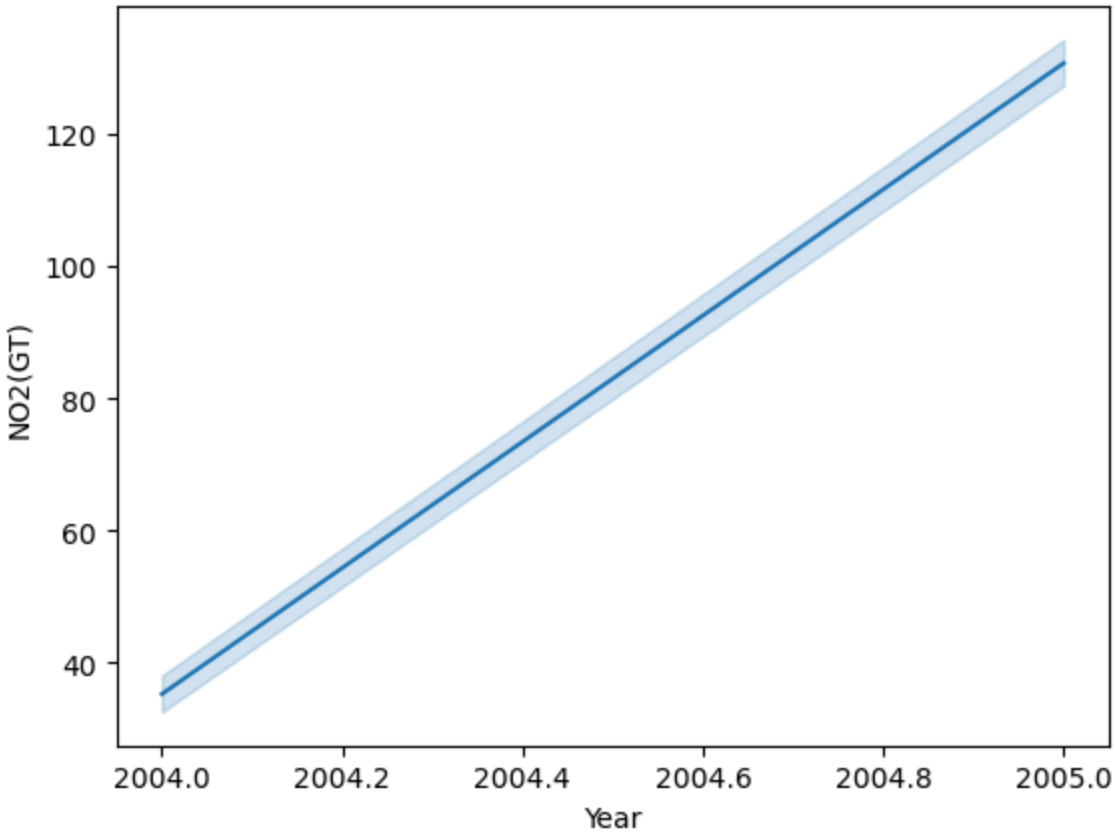
In [34]:
```
sns.lineplot(df,x="month",y='CO(GT)',)
```

Out[34]: <Axes: xlabel='month', ylabel='CO(GT)'>



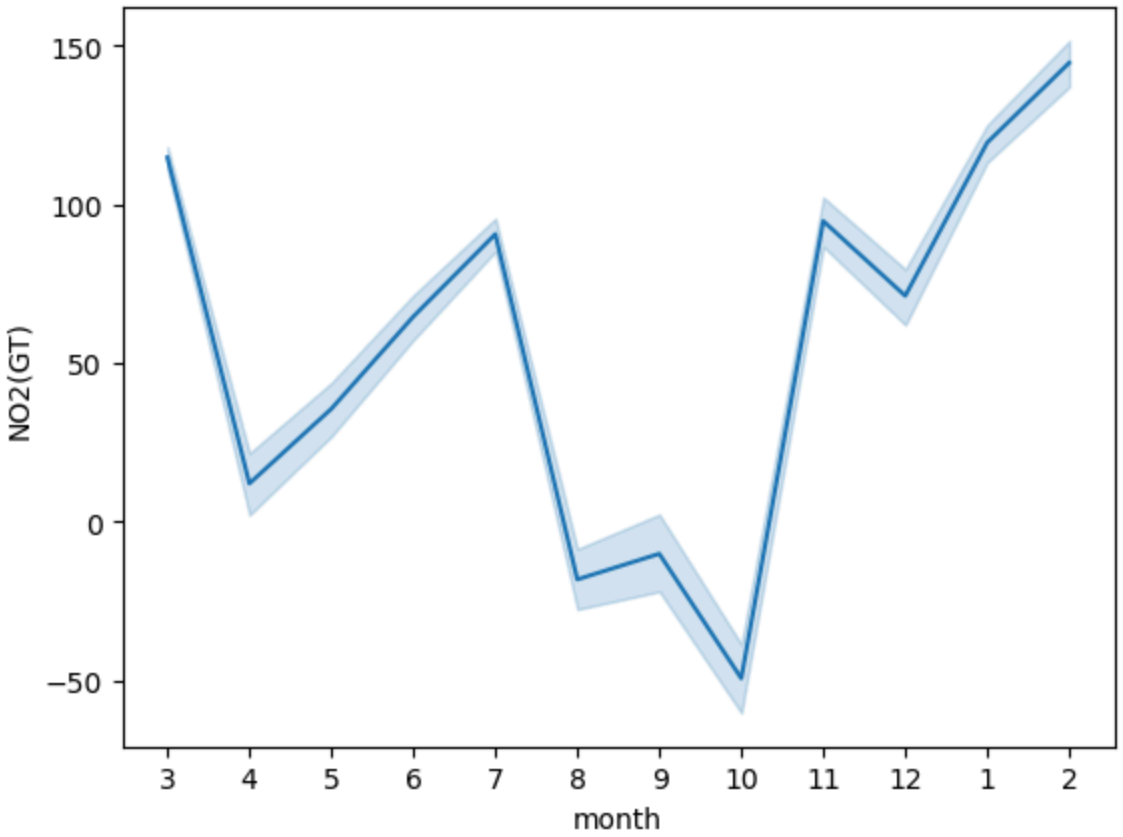
In [35]:
```
sns.lineplot(df,x="Year",y='NO2(GT)')
```

Out[35]: <Axes: xlabel='Year', ylabel='NO2(GT)'>



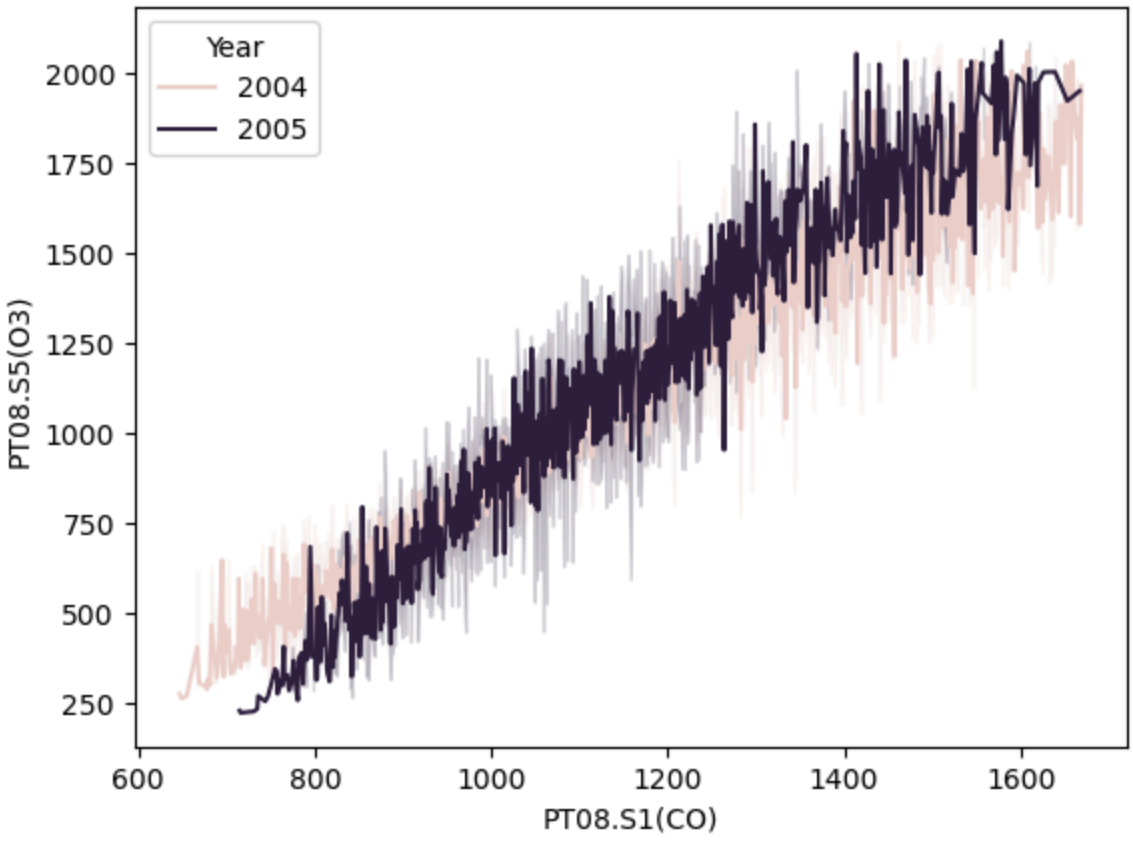
In [36]:
```
sns.lineplot(df,x="month",y='NO2(GT)')
```

Out[36]: <Axes: xlabel='month', ylabel='NO2(GT)'>



In [37]:
```
sns.lineplot(df,x='PT08.S1(CO)',y='PT08.S5(O3)',hue='Year')
```

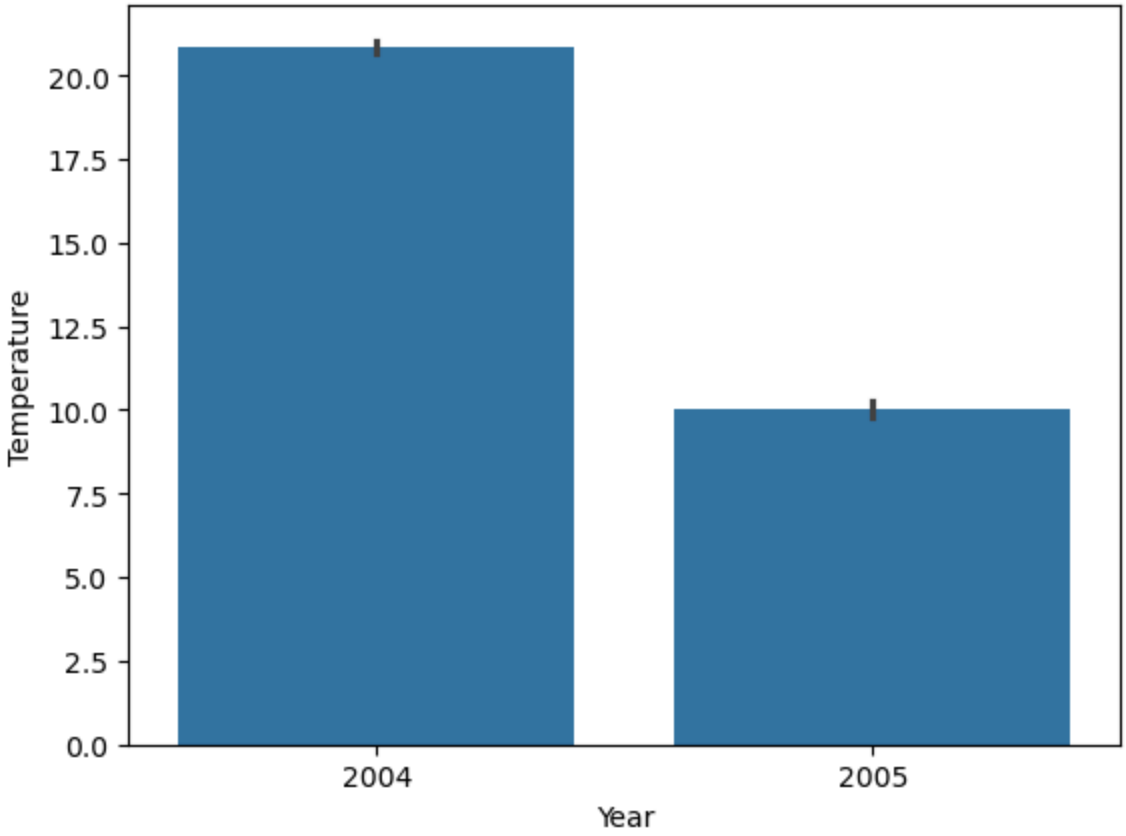
Out[37]: <Axes: xlabel='PT08.S1(CO)', ylabel='PT08.S5(O3)'>



In [38]:
```
# sns.lineplot(df,x='PT08.S1(CO)',y='PT08.S5(O3)',hue='month')
```

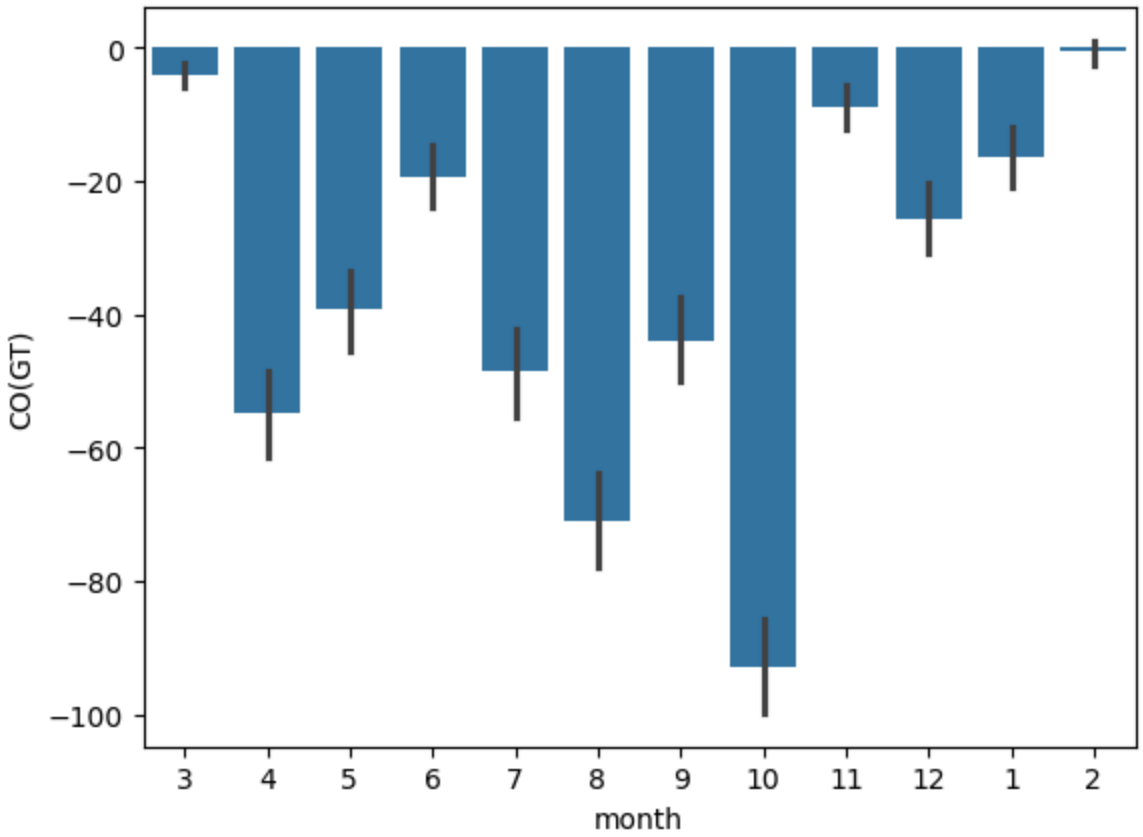
In [39]:
```
sns.barplot(df,x=df.Year,y=df.Temperature)
```

Out[39]:
```
<Axes: xlabel='Year', ylabel='Temperature'>
```



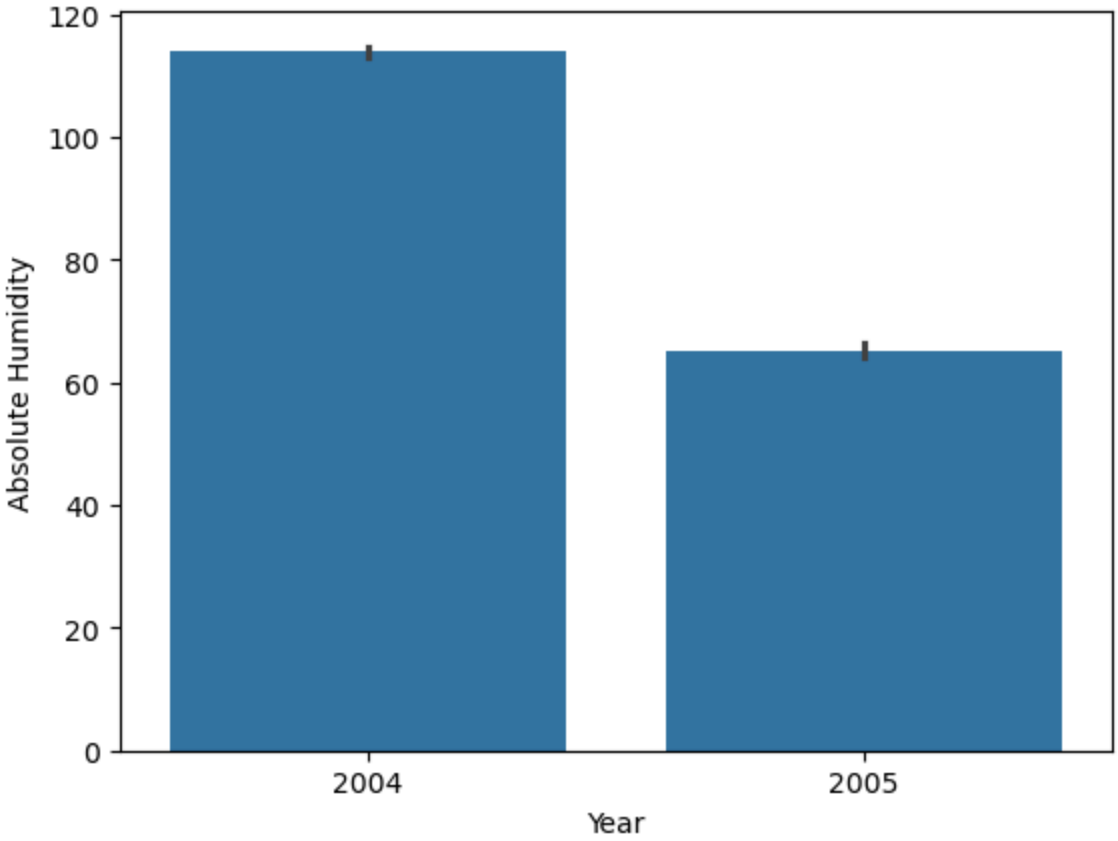
In [40]:
```
sns.barplot(df,x=df.month,y=df['CO(GT)'])
```

Out[40]:
```
<Axes: xlabel='month', ylabel='CO(GT)'>
```



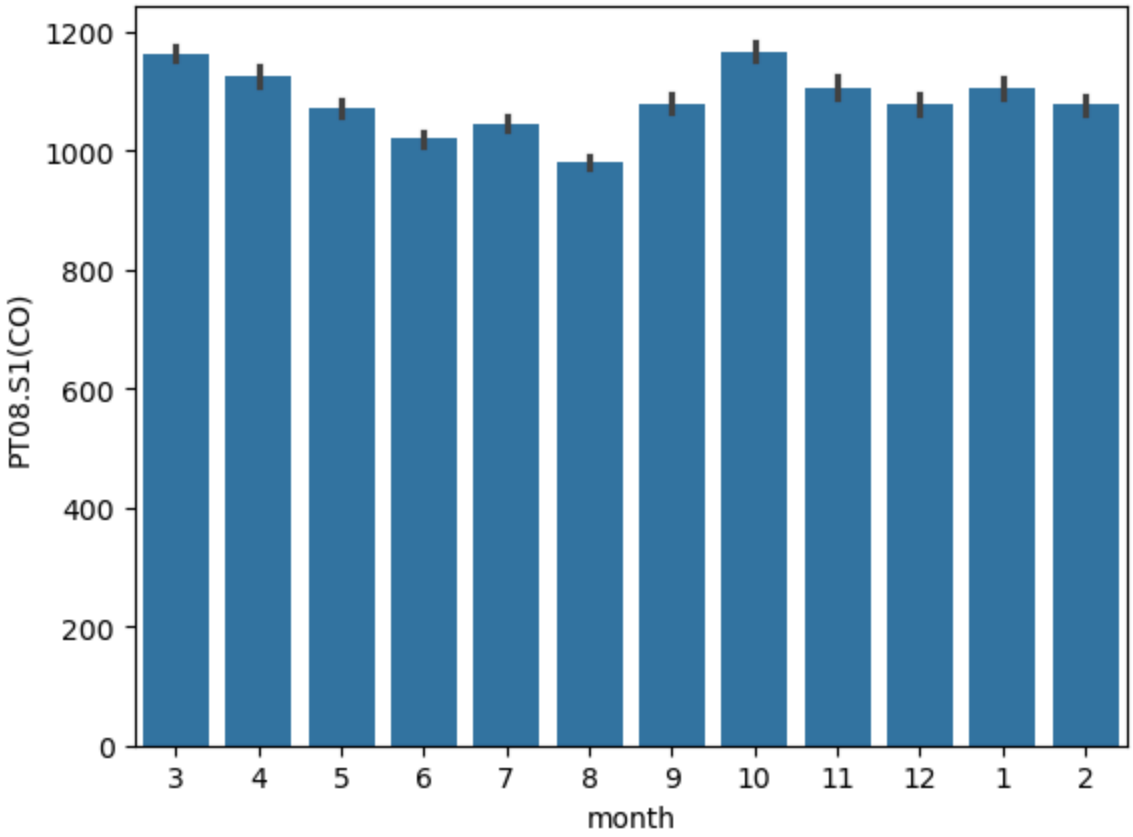
In [41]:
```
sns.barplot(df,x=df.Year,y='Absolute Humidity')
```

Out[41]: <Axes: xlabel='Year', ylabel='Absolute Humidity'>



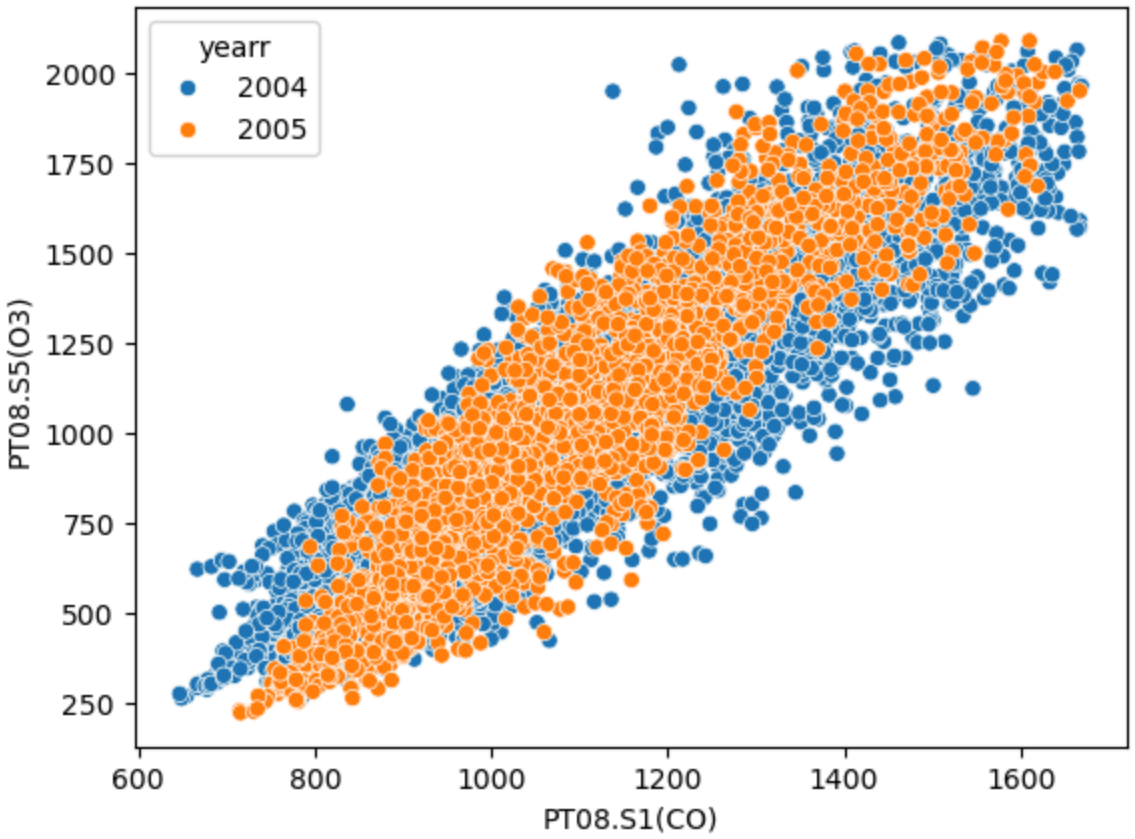
In [42]:
```
sns.barplot(df,x=df.month,y='PT08.S1(CO)')
```

Out[42]:
```
<Axes: xlabel='month', ylabel='PT08.S1(CO)'>
```



In [43]:
```
sns.scatterplot(df,x='PT08.S1(CO)',y='PT08.S5(O3)', hue='yearr')
```

Out[43]:
```
<Axes: xlabel='PT08.S1(CO)', ylabel='PT08.S5(O3)'>
```

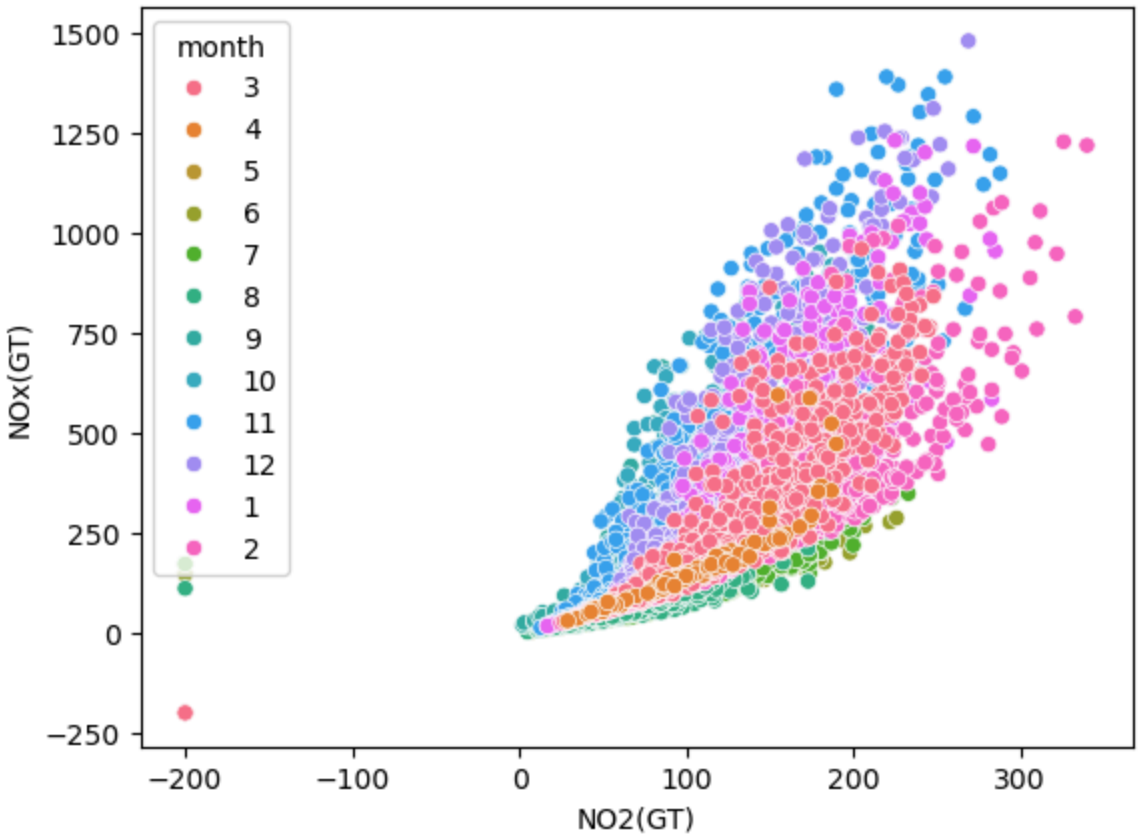


In [44]:
```
df.columns
```

Out[44]:
```
Index(['Date', 'Time', 'CO(GT)', 'PT08.S1(CO)', 'NMHC(GT)', 'C6H6(GT)',
       'PT08.S2(NMHC)', 'NOx(GT)', 'PT08.S3(NOx)', 'NO2(GT)', 'PT08.S4(NO2)',
       'PT08.S5(O3)', 'Temperature', 'Relative Humidity', 'Absolute Humidity',
       'Year', 'Month', 'yearr', 'month'],
      dtype='object')
```

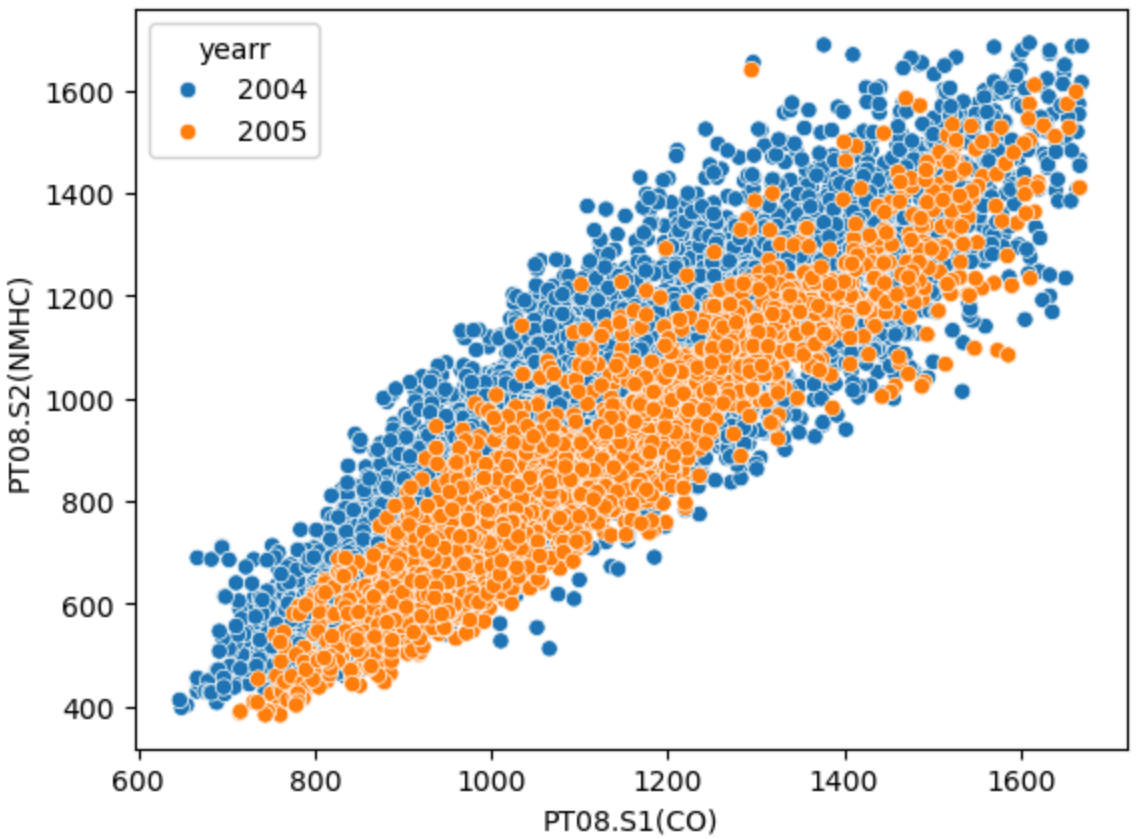
In [45]:
```
sns.scatterplot(df,x='NO2(GT)',y='NOx(GT)', hue='month')
```

Out[45]:
```
<Axes: xlabel='NO2(GT)', ylabel='NOx(GT)'>
```



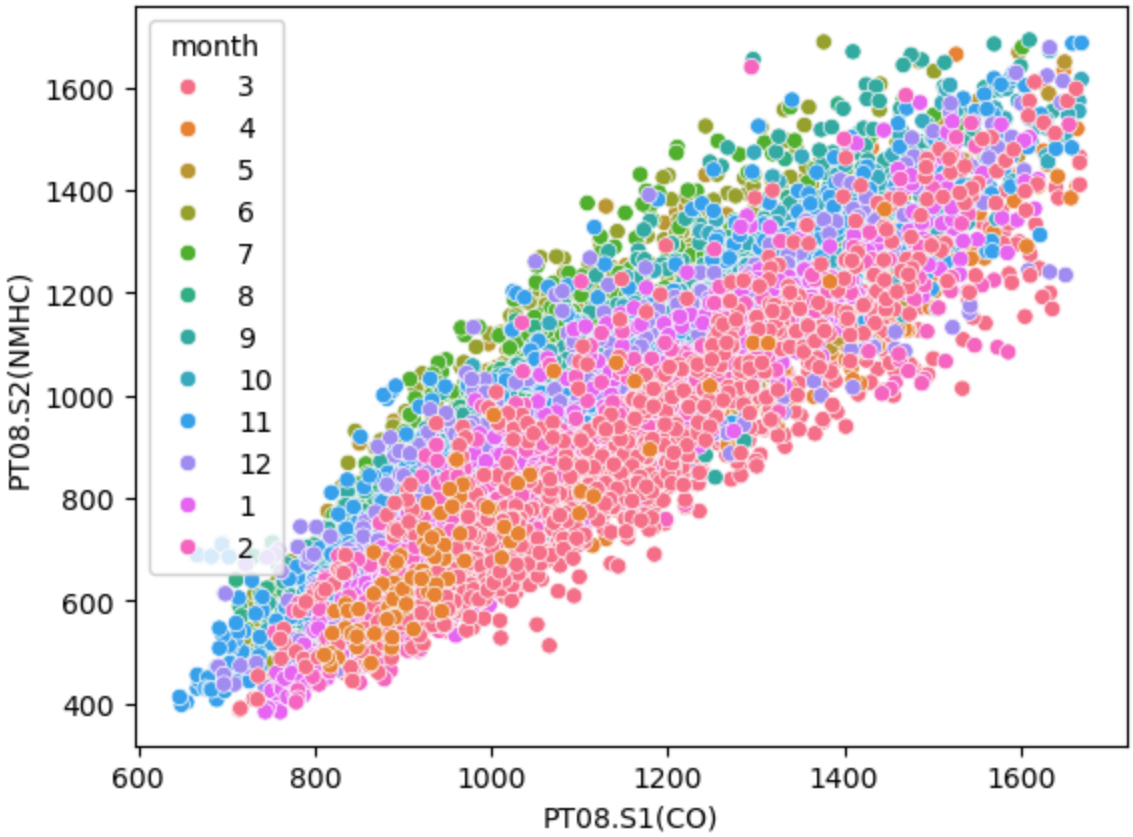
In [46]:
```
sns.scatterplot(df,y='PT08.S2(NMHC)',x='PT08.S1(CO)', hue='yearr')
```

Out[46]: <Axes: xlabel='PT08.S1(CO)', ylabel='PT08.S2(NMHC)'>



In [47]:
```
sns.scatterplot(df,y='PT08.S2(NMHC)',x='PT08.S1(CO)', hue='month')
```

Out[47]: <Axes: xlabel='PT08.S1(CO)', ylabel='PT08.S2(NMHC)'>



In [ ]:

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('heart.csv')
```

In [3]:
```
df
# 303 rows × 14 columns
```

Out[3]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	o
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	

303 rows × 14 columns

In [4]:
```
df = df.drop_duplicates()
```

In [5]:
```
df
```

Out[5]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	o
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	

302 rows × 14 columns

In [6]:
```
df.isna().sum()
# No null values, it's clean
```

Out[6]:
```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

In [7]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

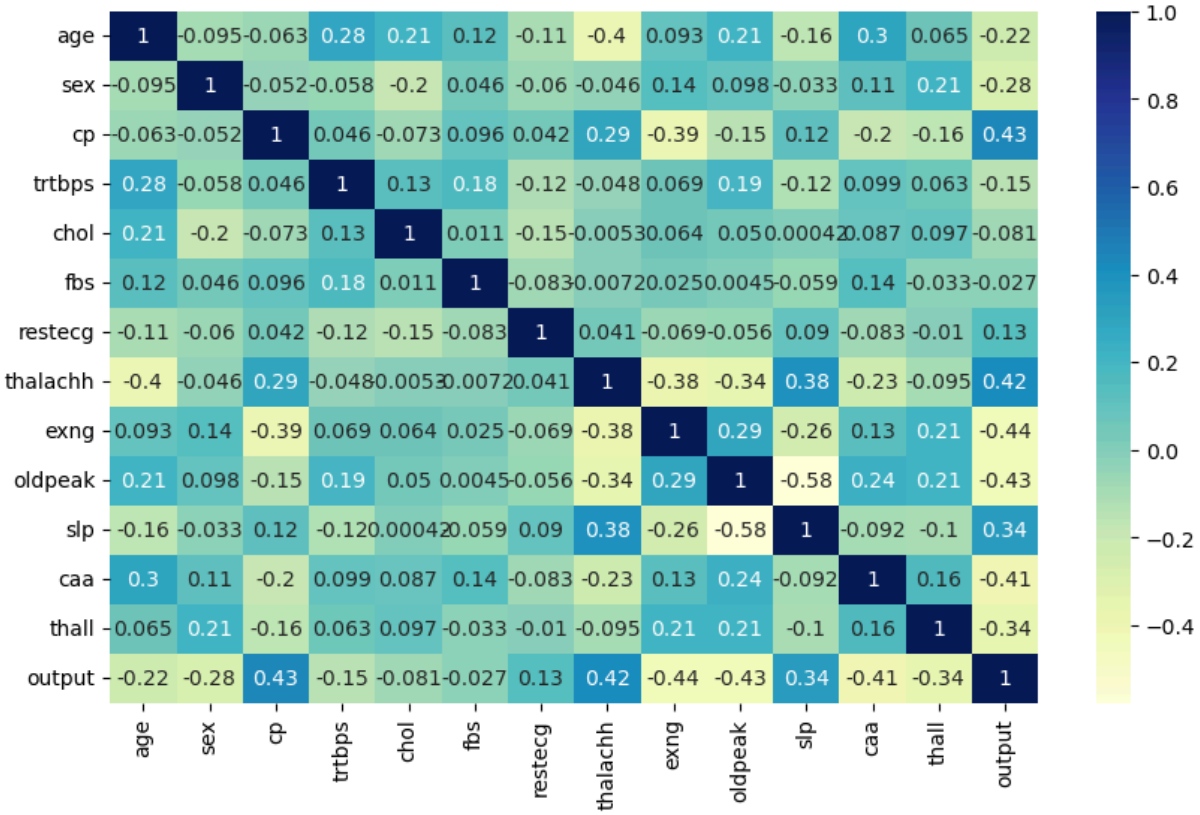
In [9]:
```
df.columns
```

Out[9]:
```
Index(['age', 'sex', 'cp', 'trtbps', 'chol', 'fbs', 'restecg', 'thalachh',
       'exng', 'oldpeak', 'slp', 'caa', 'thall', 'output'],
      dtype='object')
```

In [10]:
```
plt.figure(figsize=(10,6))
```

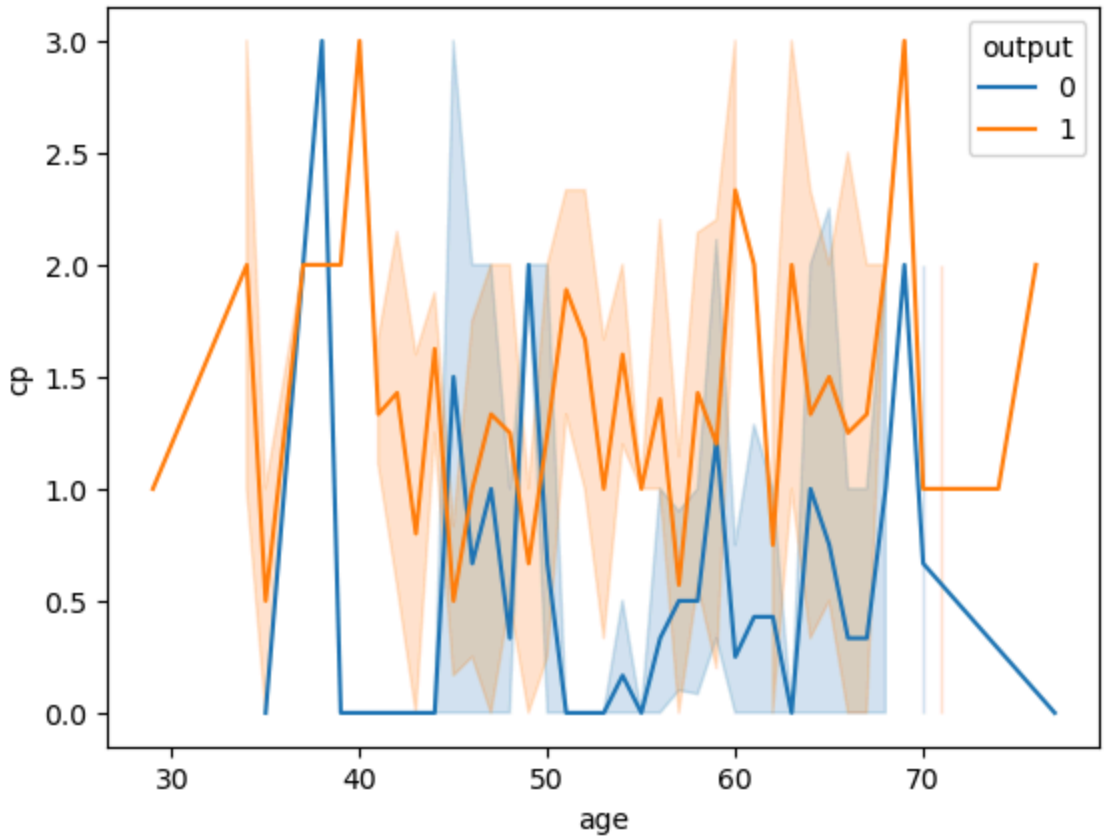
```
sns.heatmap(df.corr(),cmap = 'YlGnBu', annot = True)
```

Out[10]: <Axes: >



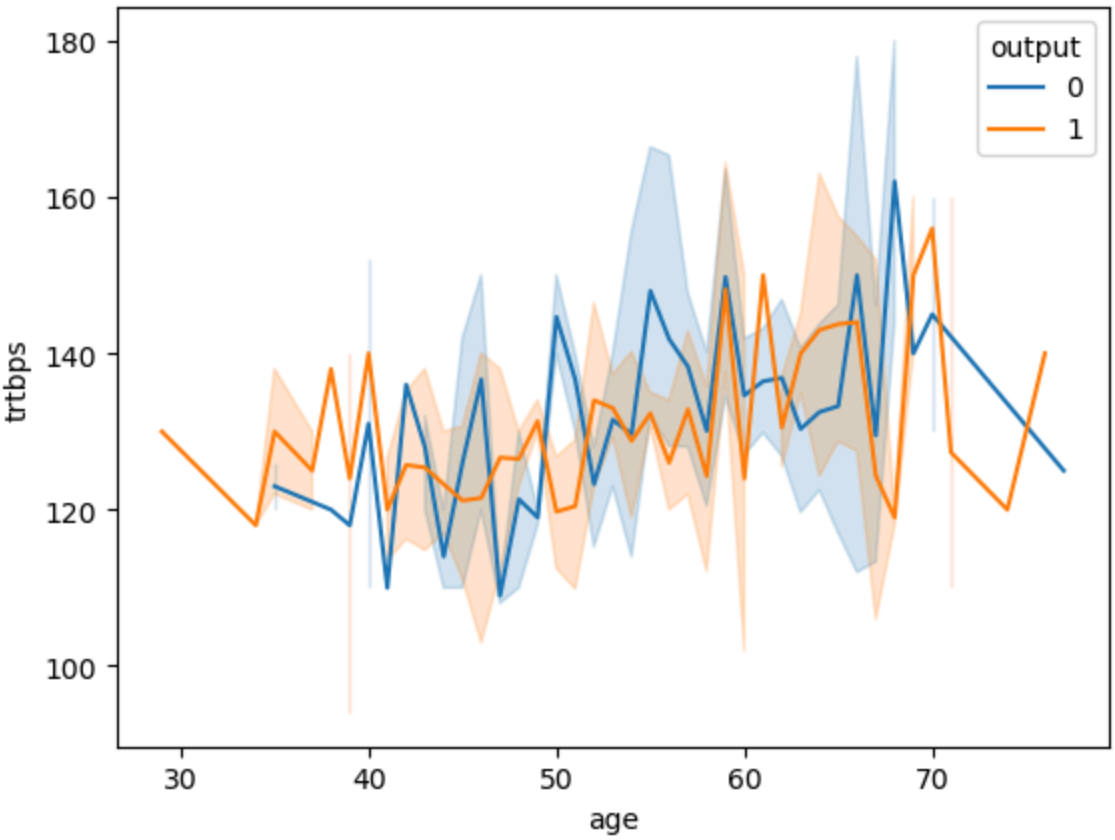
In [11]:
```
sns.lineplot(data=df,x=df.age,y=df.cp,hue='output')
```

Out[11]: <Axes: xlabel='age', ylabel='cp'>



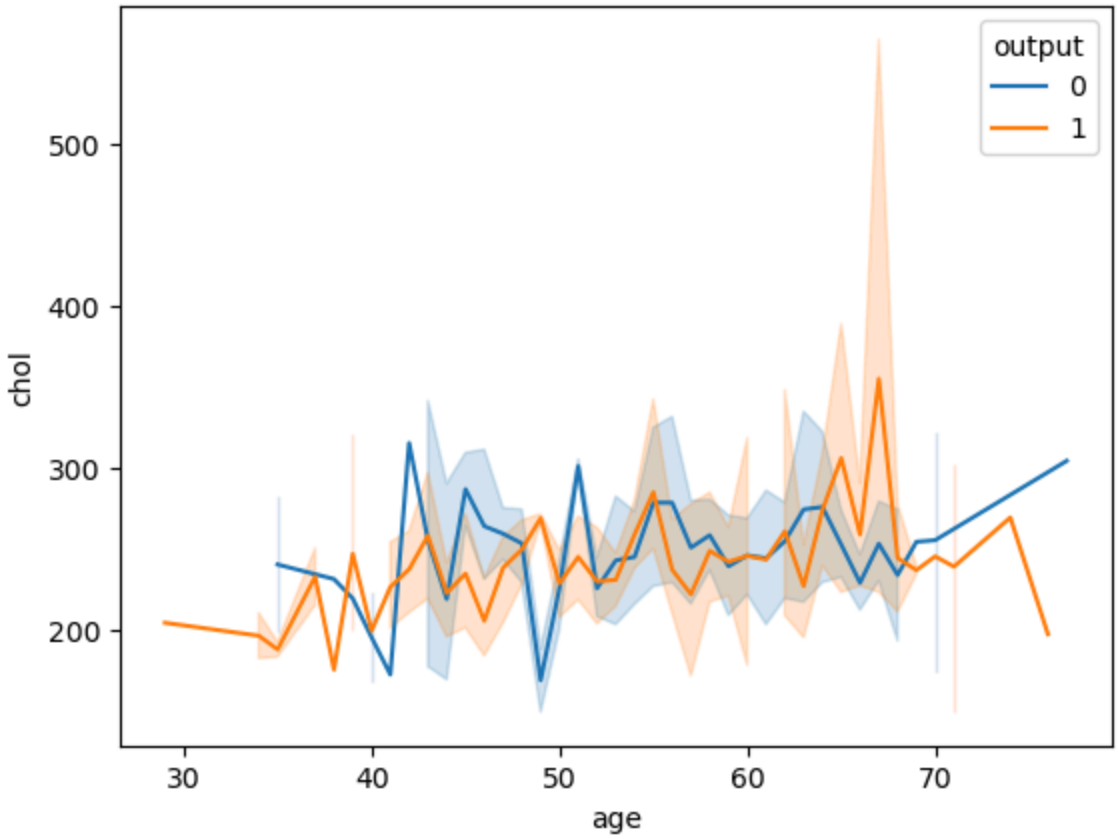
In [12]:
```
sns.lineplot(data=df,x=df.age,y=df.trtbps,hue='output')
```

Out[12]: <Axes: xlabel='age', ylabel='trtbps'>



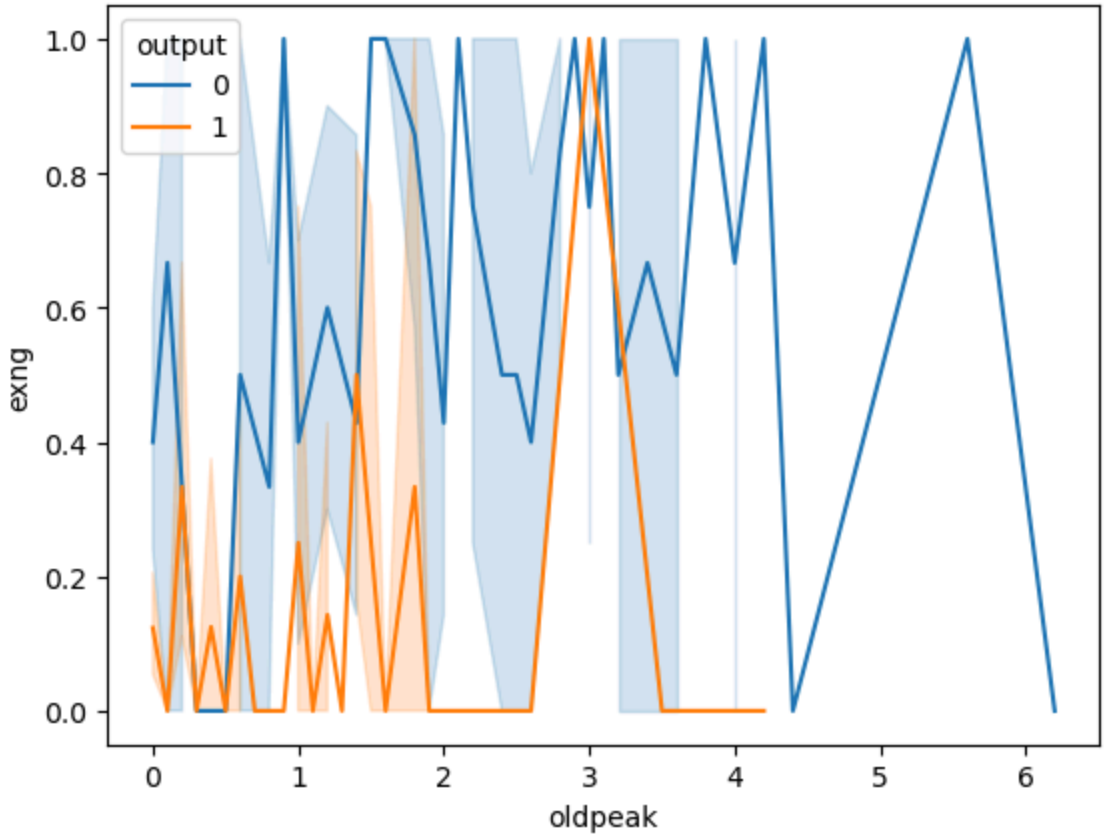
In [13]:
```
sns.lineplot(data=df,x=df.age,y=df.chol,hue='output')
```

Out[13]:
```
<Axes: xlabel='age', ylabel='chol'>
```



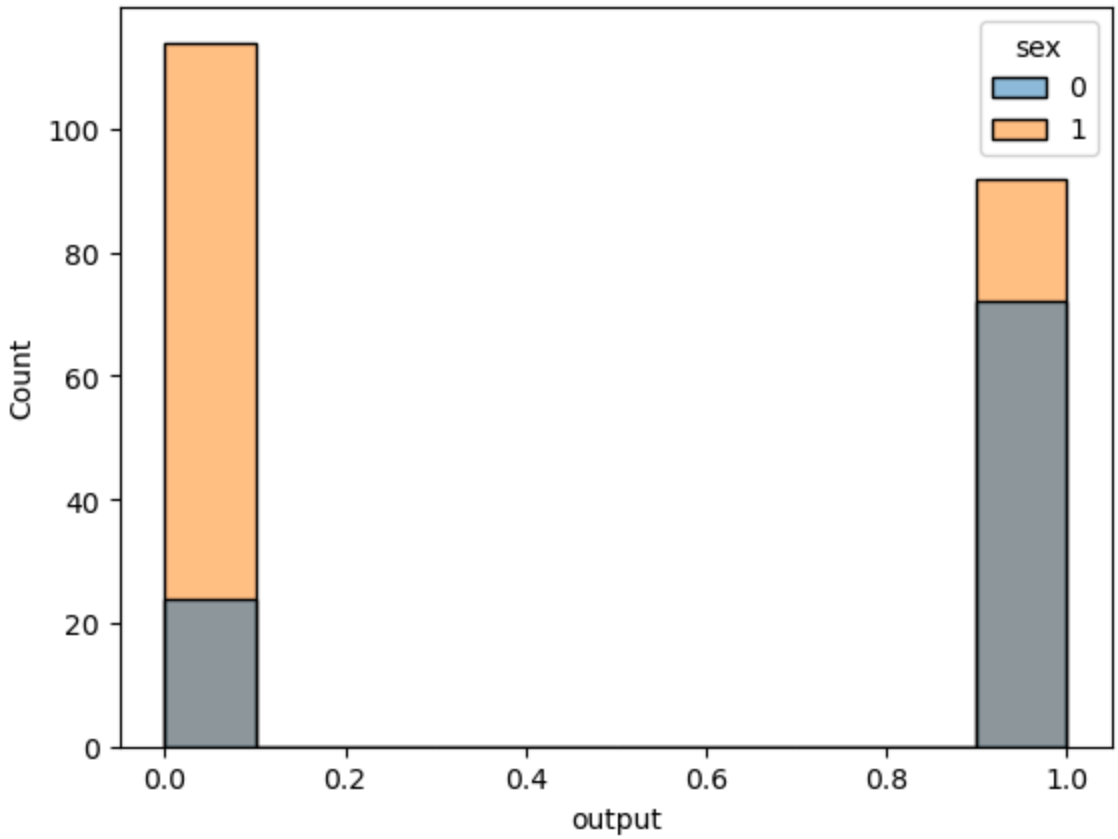
In [14]:
```
sns.lineplot(df,x=df.oldpeak,y=df.exng,hue='output')
```

Out[14]:
```
<Axes: xlabel='oldpeak', ylabel='exng'>
```



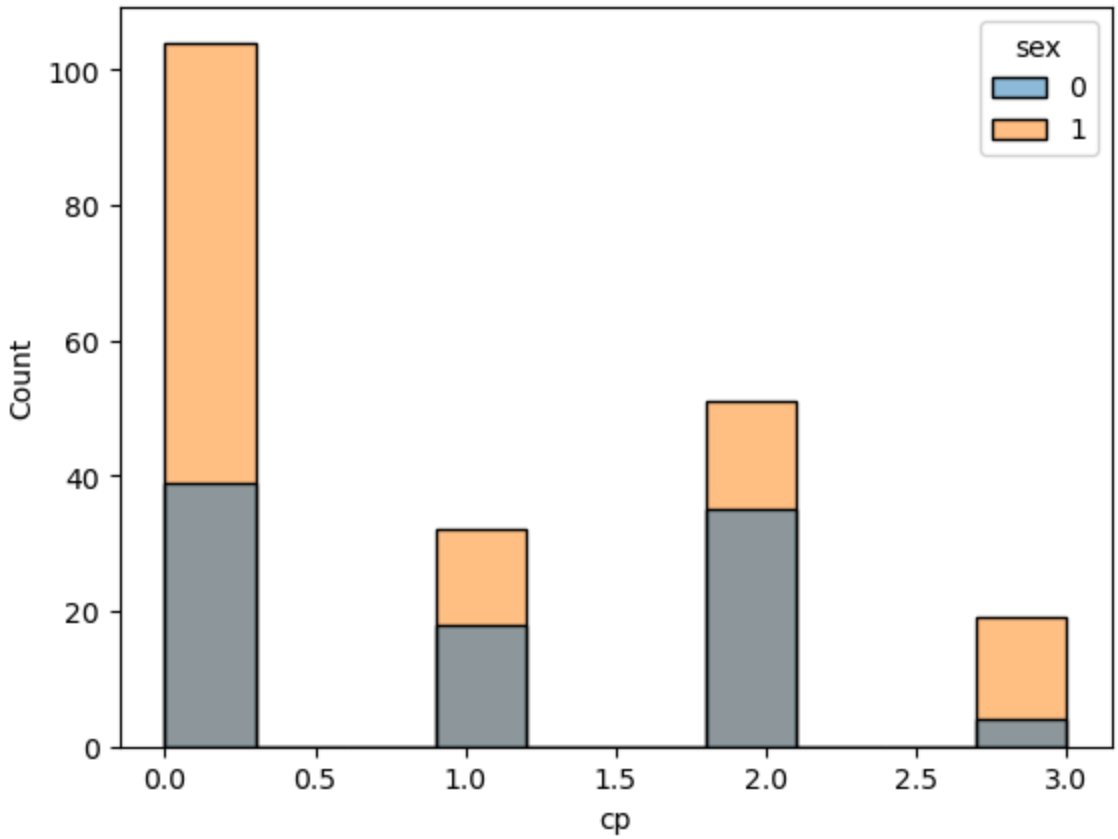
In [15]:
```
# Shows the Distribution of Heat Diseases with respect to male and female
sns.histplot(data=df,
             x=df.output,
             hue=df.sex)
```

Out[15]:
```
<Axes: xlabel='output', ylabel='Count'>
```



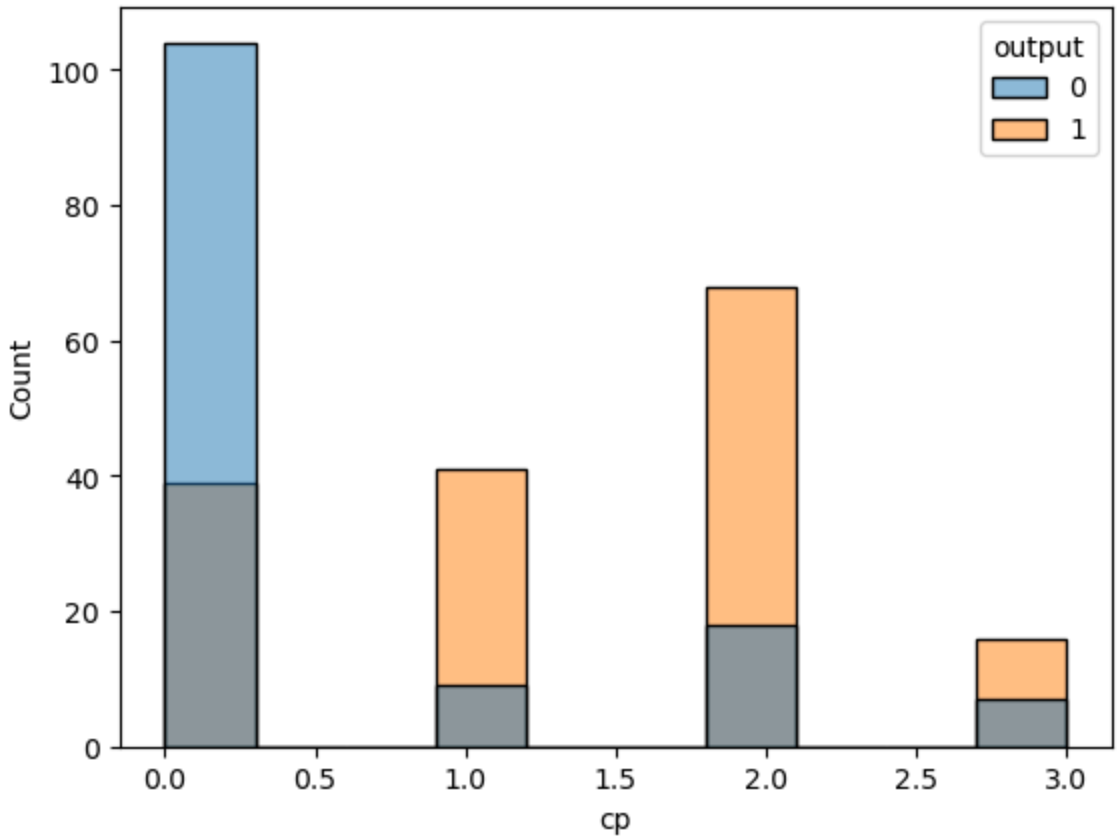
In [16]:
```
# Shows the Distribution of cp types with respect to male and female
sns.histplot(data=df,
             x=df.cp,
             hue=df.sex)
```

Out[16]:
```
<Axes: xlabel='cp', ylabel='Count'>
```



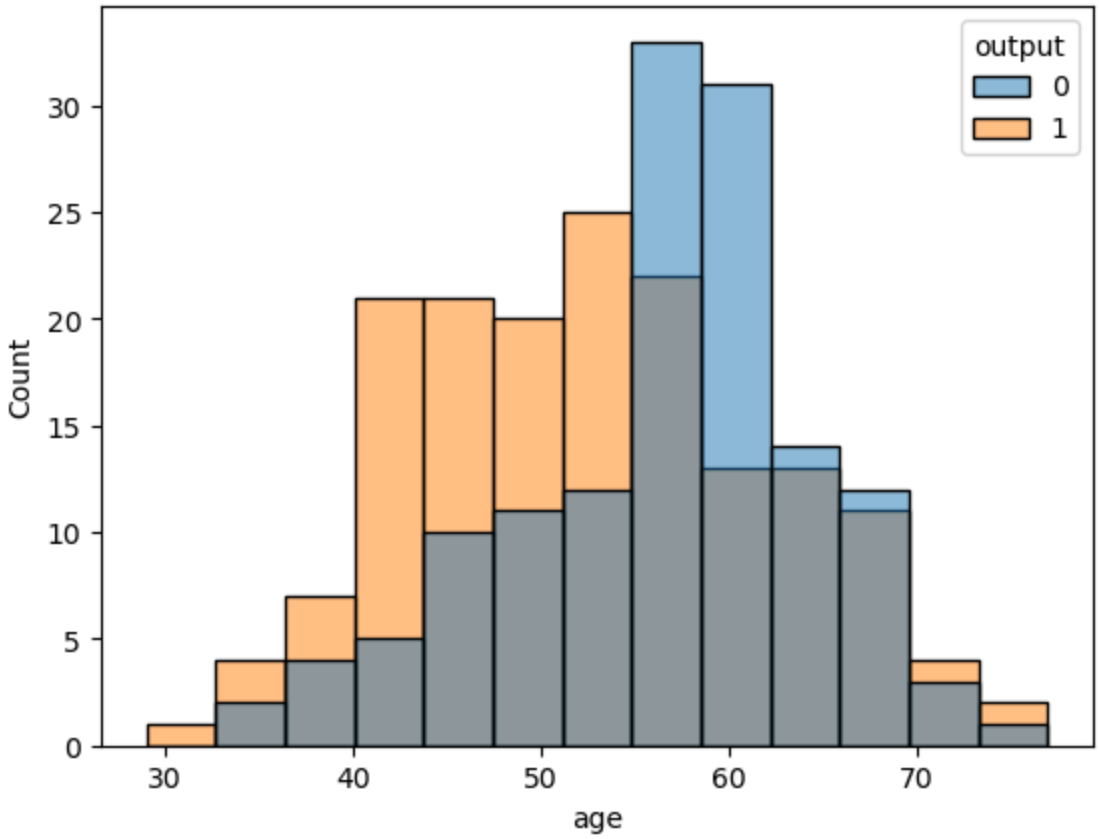
In [17]:
```
sns.histplot(data=df,x=df.cp, hue='output')
```

Out[17]: <Axes: xlabel='cp', ylabel='Count'>



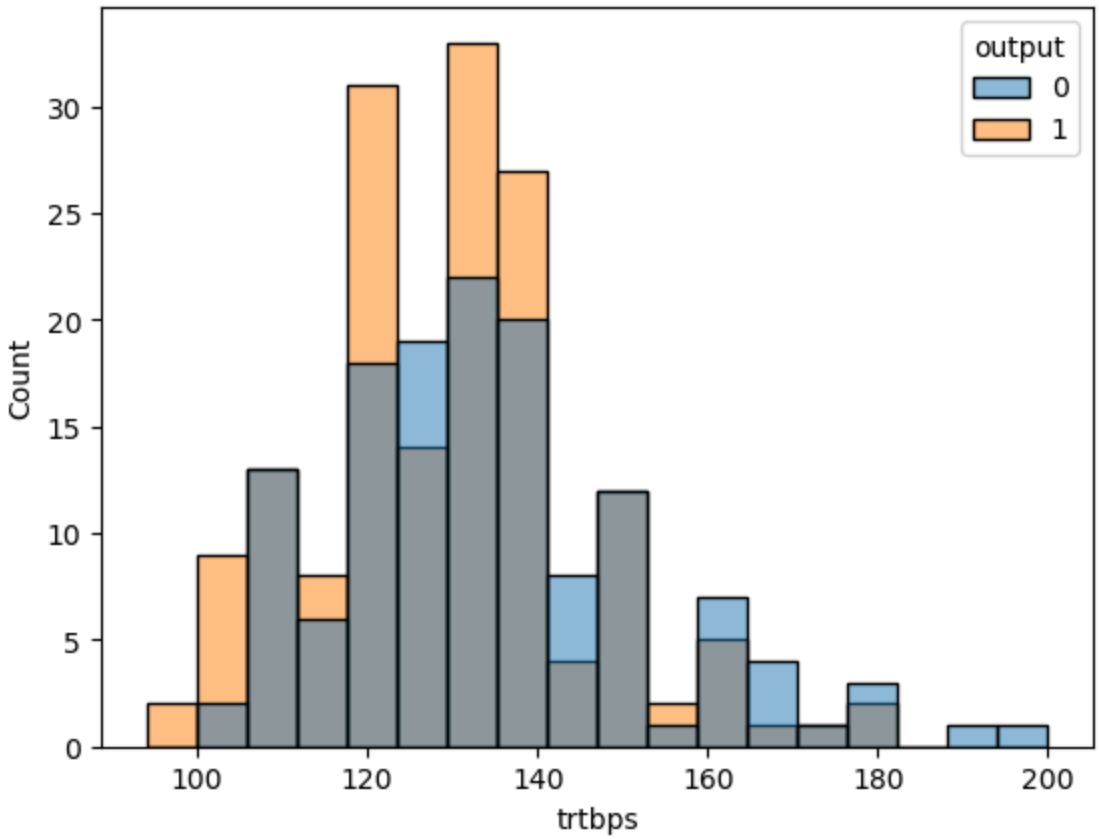
In [18]:
```
# Shows the Distribution of age w.r.t output
sns.histplot(data=df,x=df['age'], hue='output')
```

Out[18]:
```
<Axes: xlabel='age', ylabel='Count'>
```



In [19]:
```
sns.histplot(data=df,x=df.trtbps, hue='output')
```

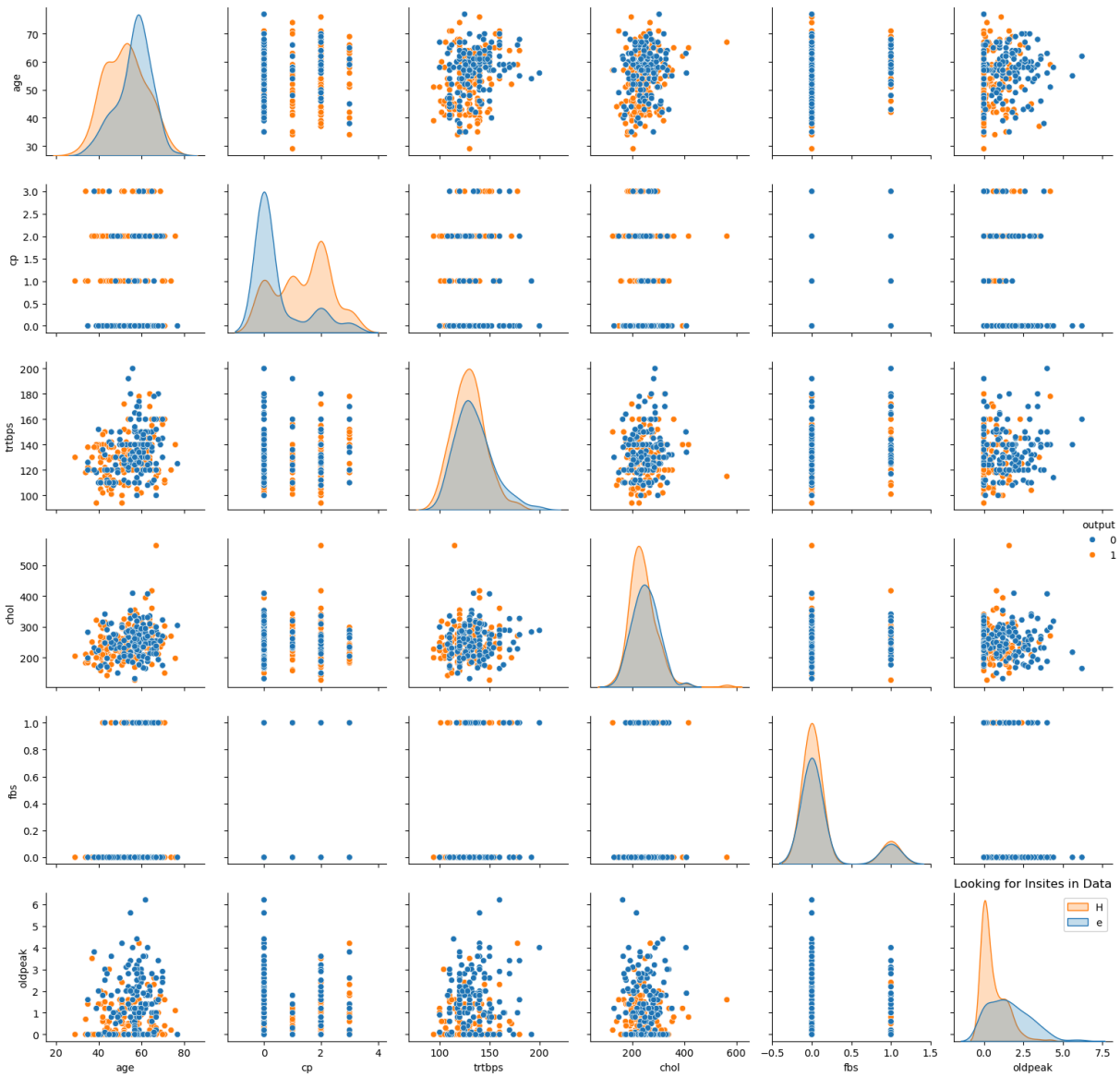
Out[19]:
```
<Axes: xlabel='trtbps', ylabel='Count'>
```



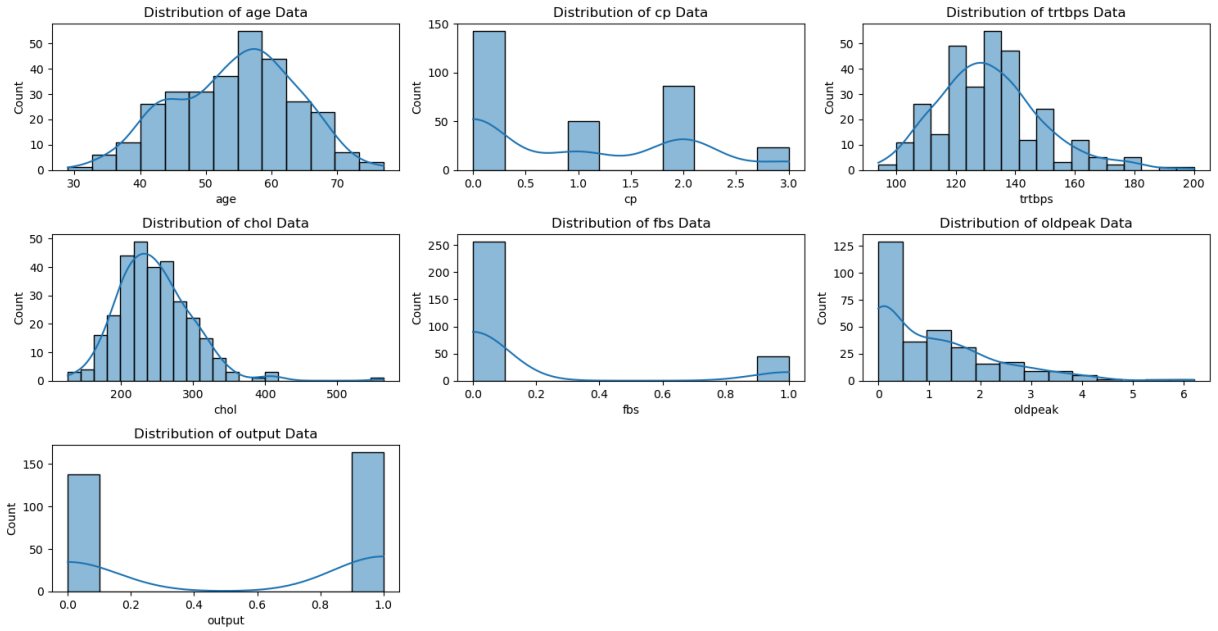
In [20]:
```
temp_df = df[['age','cp', 'trtbps','chol','fbs','oldpeak','output']]
plt.figure(figsize=(15,10))
sns.pairplot(temp_df,hue="output")
plt.title("Looking for Insites in Data")
plt.legend("HeartDisease")
plt.tight_layout()
plt.plot()
```

Out[20]: []

<Figure size 1500x1000 with 0 Axes>



In [21]:
```
plt.figure(figsize=(15,10))
for i,col in enumerate(temp_df.columns,1):
    plt.subplot(4,3,i)
    plt.title(f"Distribution of {col} Data")
    sns.histplot(df[col],kde=True)
    plt.tight_layout()
    plt.plot()
```


Distribution of age Data
Count
age
Distribution of cp Data
Count
cp
Distribution of trtbps Data
Count
trtbps
Distribution of chol Data
Count
chol
Distribution of fbs Data
Count
fbs
Distribution of oldpeak Data
Count
oldpeak
Distribution of output Data
Count
output

In [ ]:

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('tips_DV.csv')
```

In [3]:
```
df
```

Out[3]:

	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 × 7 columns

In [4]:
```
df.rename(columns = {'size':'peoples'}, inplace = True)
```

In [5]:
```
df.isna().sum()
```

Out[5]:
```
total_bill    0
tip           0
sex           0
smoker        0
day           0
time          0
peoples       0
dtype: int64
```

In [6]:
```
col_var = df[['peoples','sex', 'smoker', 'day', 'time']]

for i in col_var:
    print(i, df[i].unique())
```

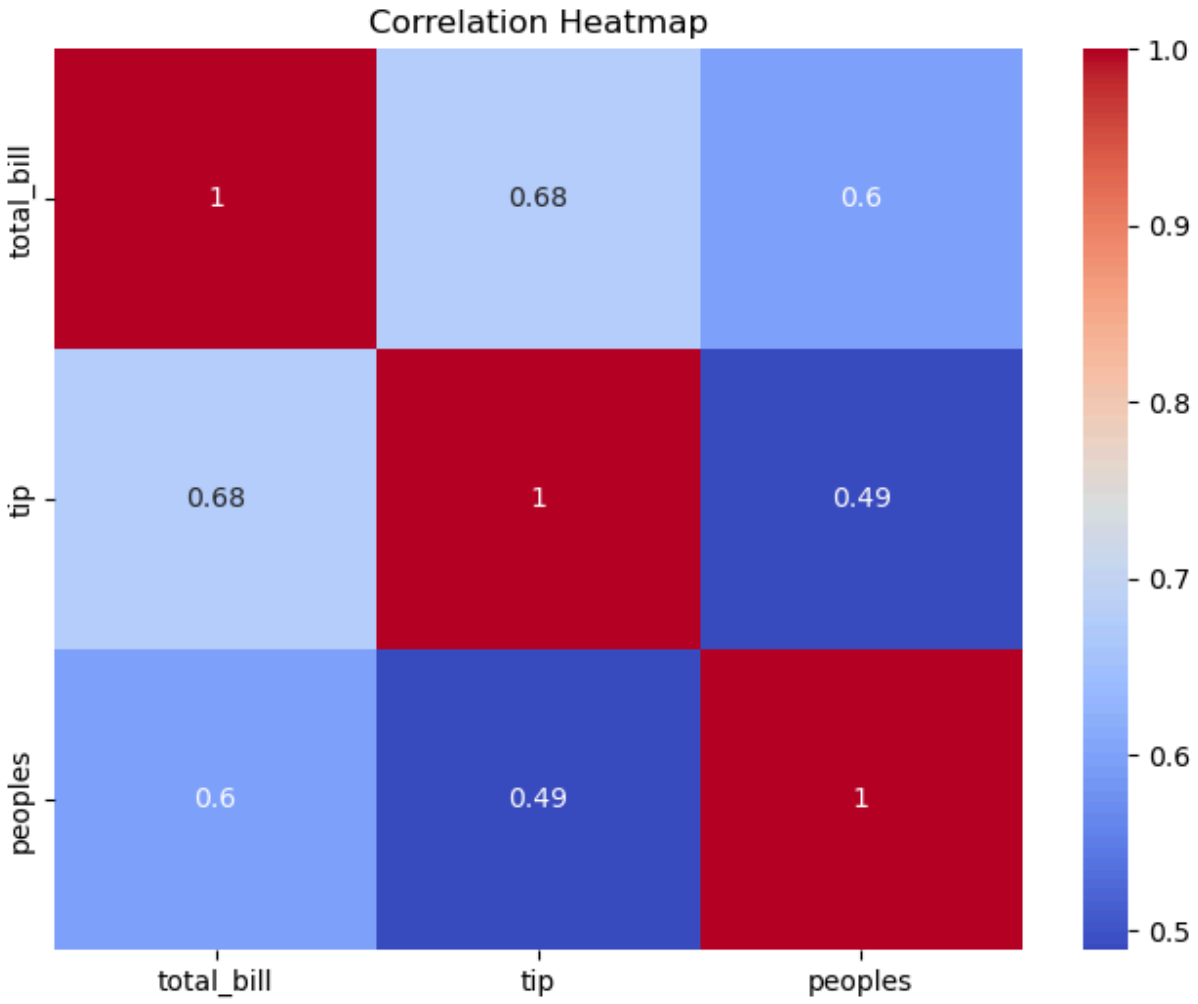
```
peoples [2 3 4 1 6 5]
sex ['Female' 'Male']
smoker ['No' 'Yes']
day ['Sun' 'Sat' 'Thur' 'Fri']
time ['Dinner' 'Lunch']
```

In [7]:
```
df.columns
```

Out[7]:
```
Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'peoples'], dtype='obj
ect')
```

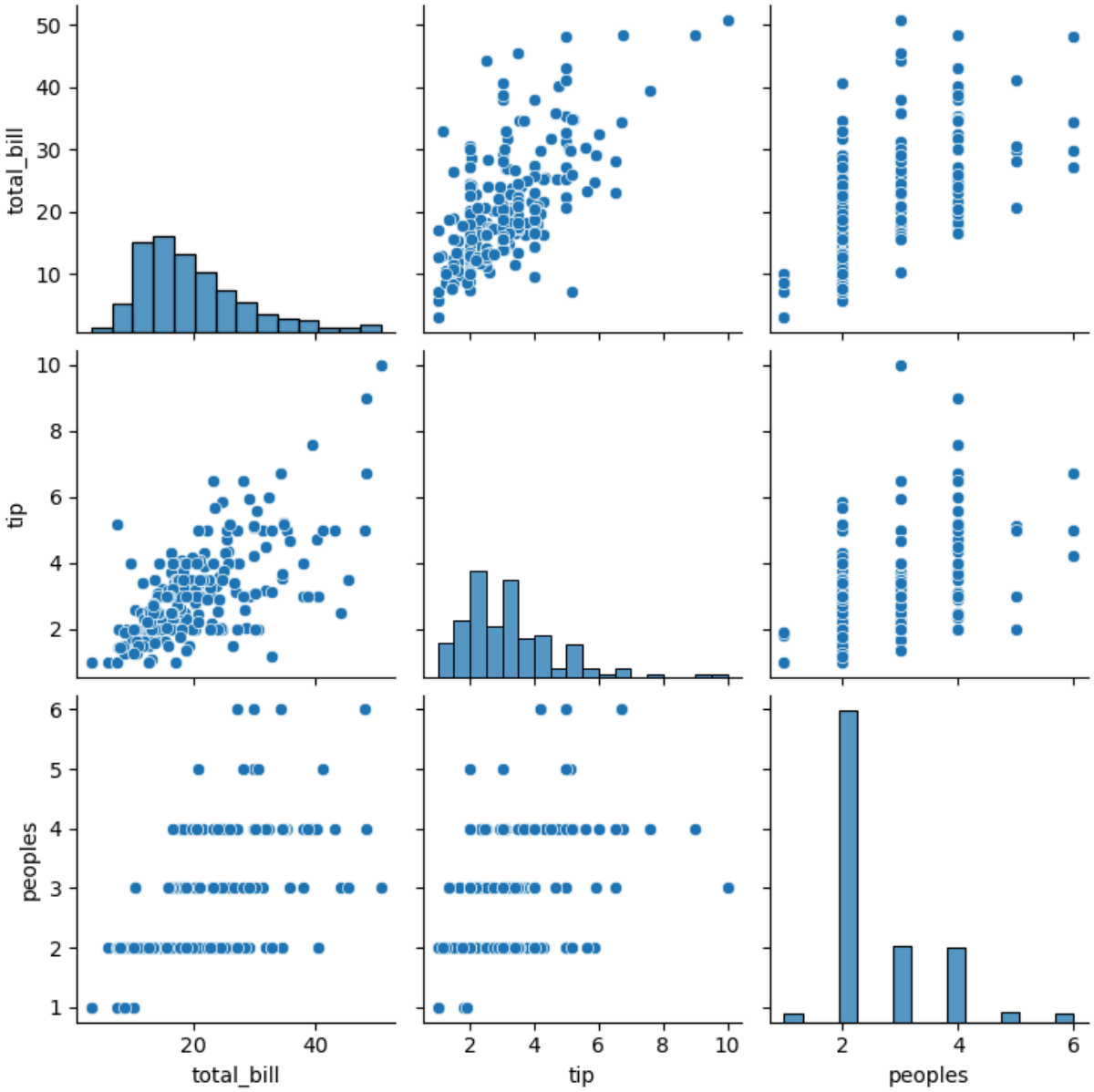
In [8]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [10]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Select only numeric columns from the DataFrame
numeric_df = df.select_dtypes(include='number')

# Now plot the heatmap safely
plt.figure(figsize=(8, 6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



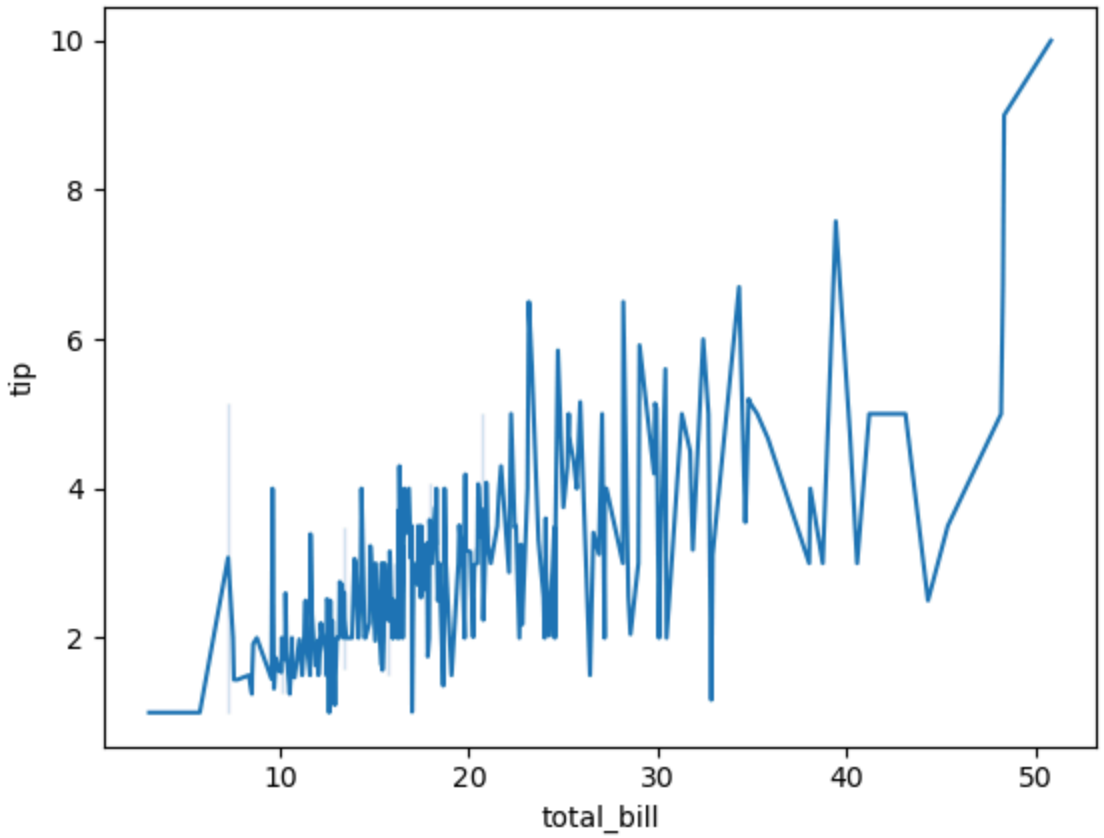
In [11]:
```
sns.pairplot(df)
```

Out[11]:
```
<seaborn.axisgrid.PairGrid at 0x274f4242570>
```



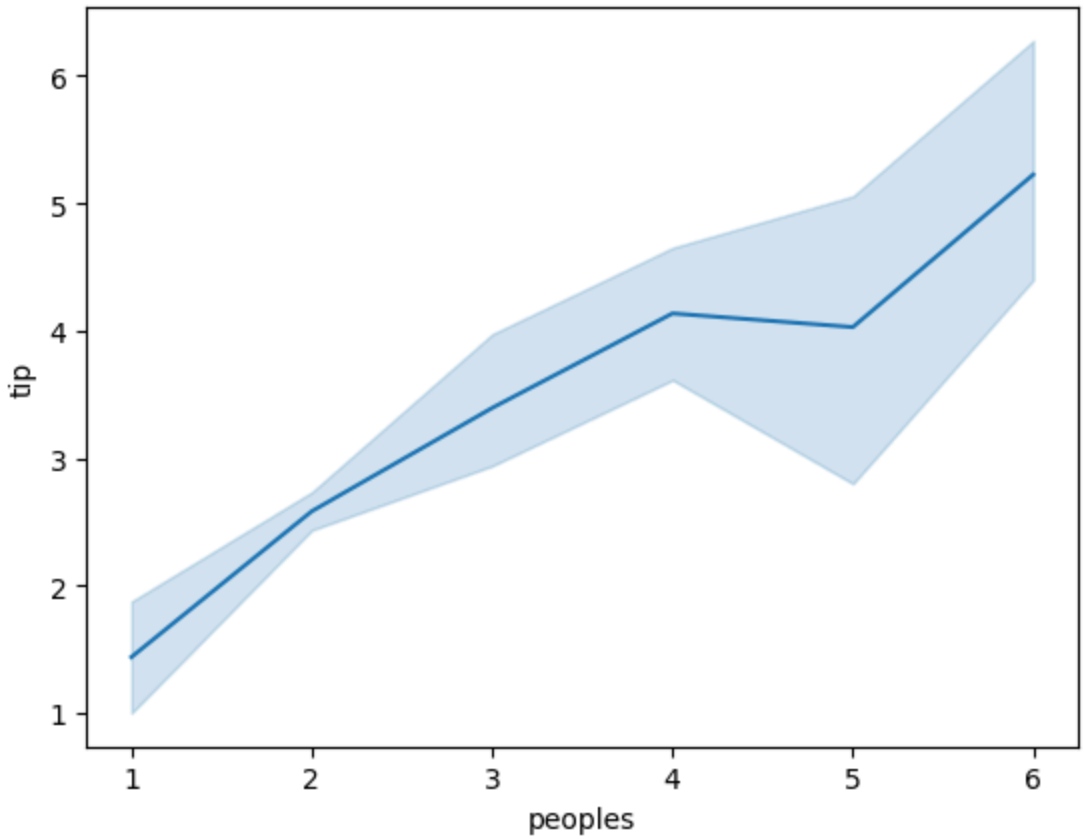
In [12]:
```
sns.lineplot(df,x='total_bill',y='tip')
# Tip increases as Totl_bill increases
```

Out[12]: <Axes: xlabel='total_bill', ylabel='tip'>



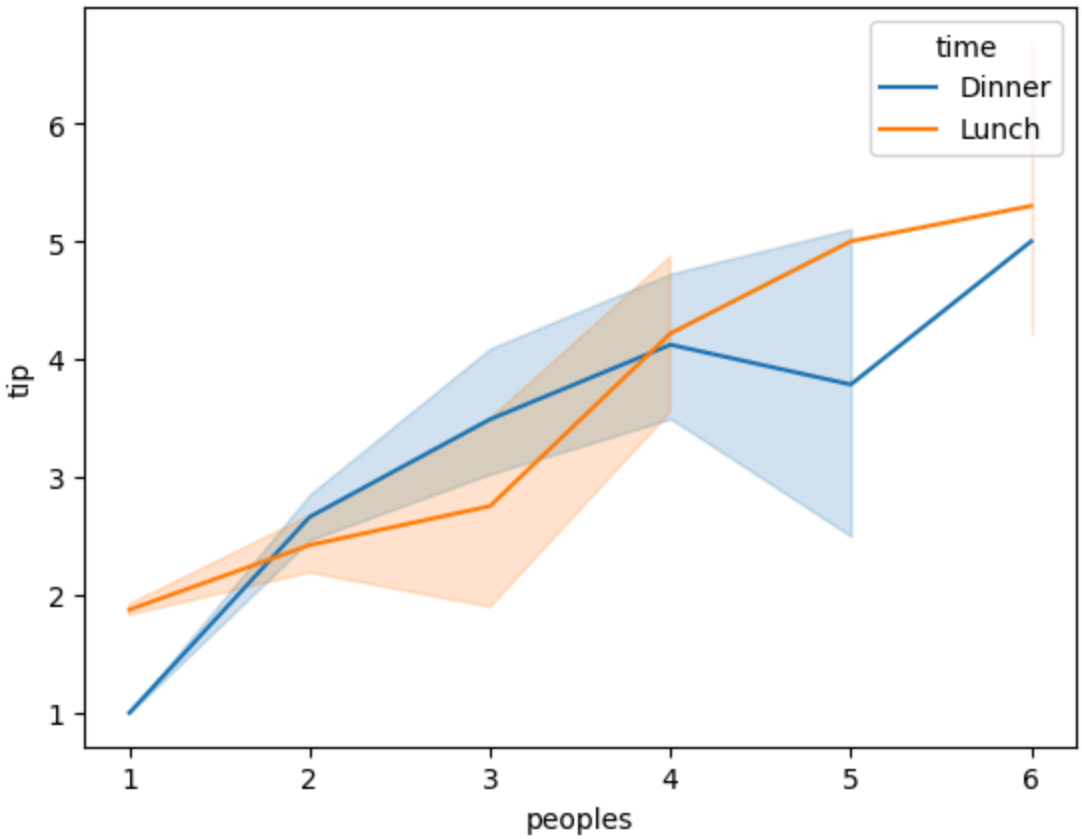
In [13]:
```
sns.lineplot(df,x='peoples',y='tip')
# Tip increases as no of people increases
```

Out[13]:
```
<Axes: xlabel='peoples', ylabel='tip'>
```



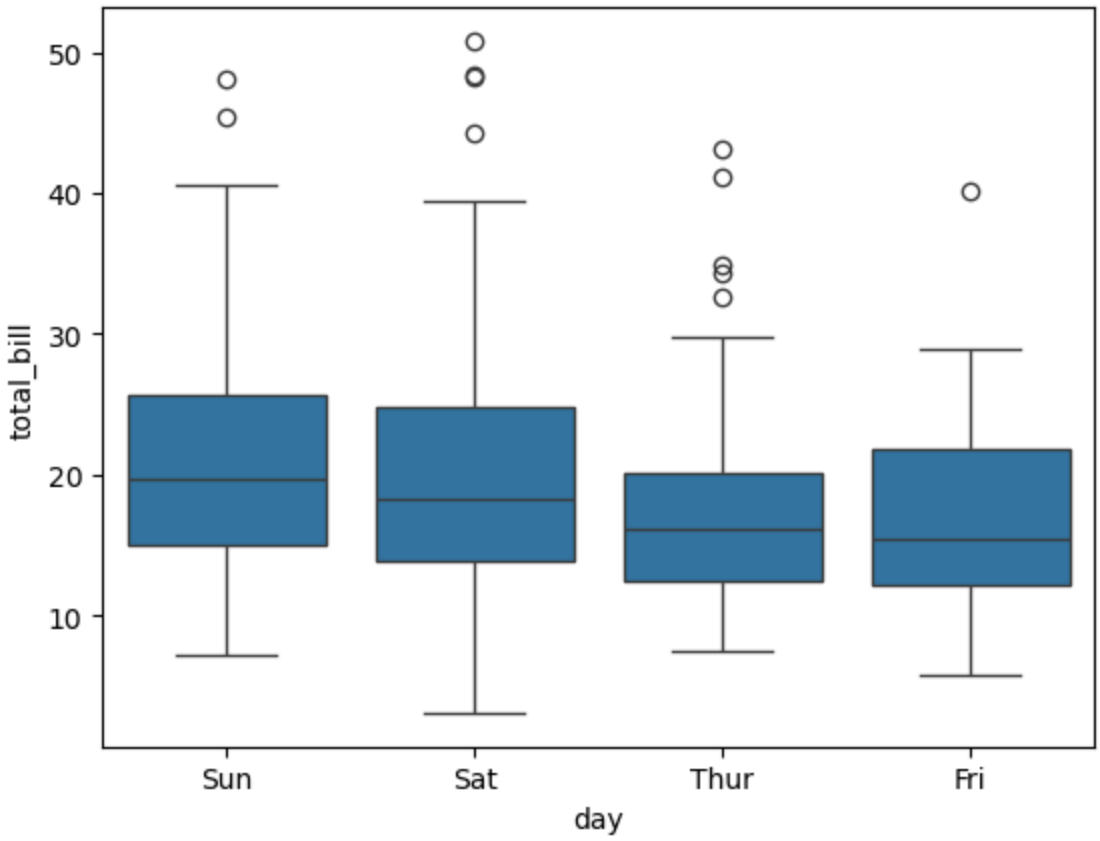
In [14]:
```
sns.lineplot(df,x='peoples',y='tip',hue='time')
# higher tip at lauch time as compared to Dinner
```

Out[14]: <Axes: xlabel='peoples', ylabel='tip'>



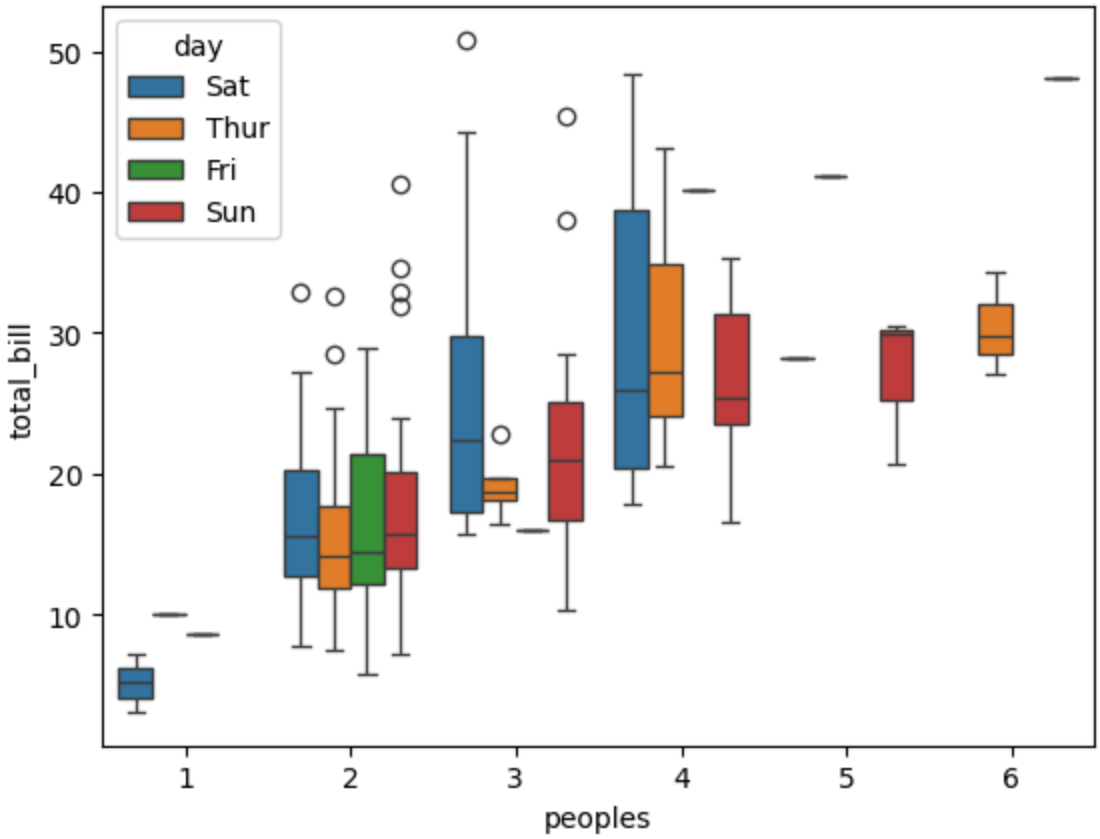
In [15]:
```
sns.boxplot(df,x='day',y='total_bill')
```

Out[15]:
```
<Axes: xlabel='day', ylabel='total_bill'>
```



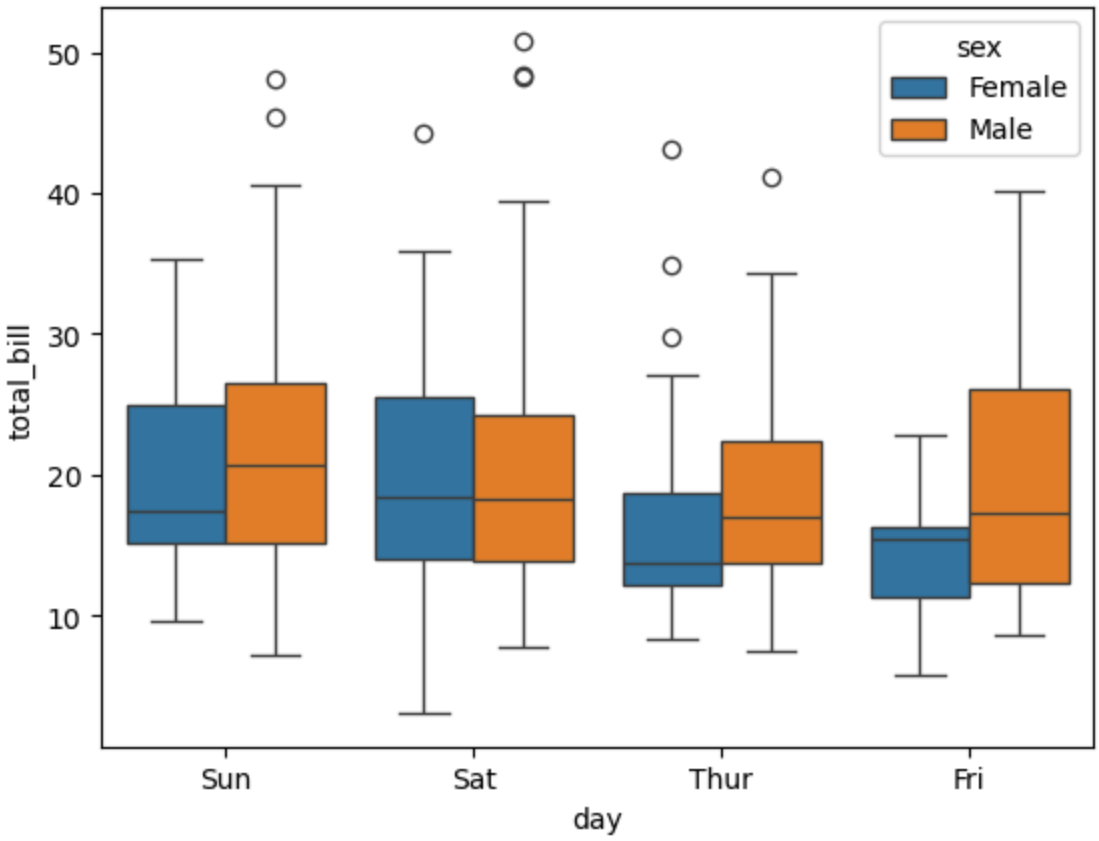
In [16]:
```
sns.boxplot(df,x='peoples',y='total_bill',hue='day')
```

Out[16]:
```
<Axes: xlabel='peoples', ylabel='total_bill'>
```



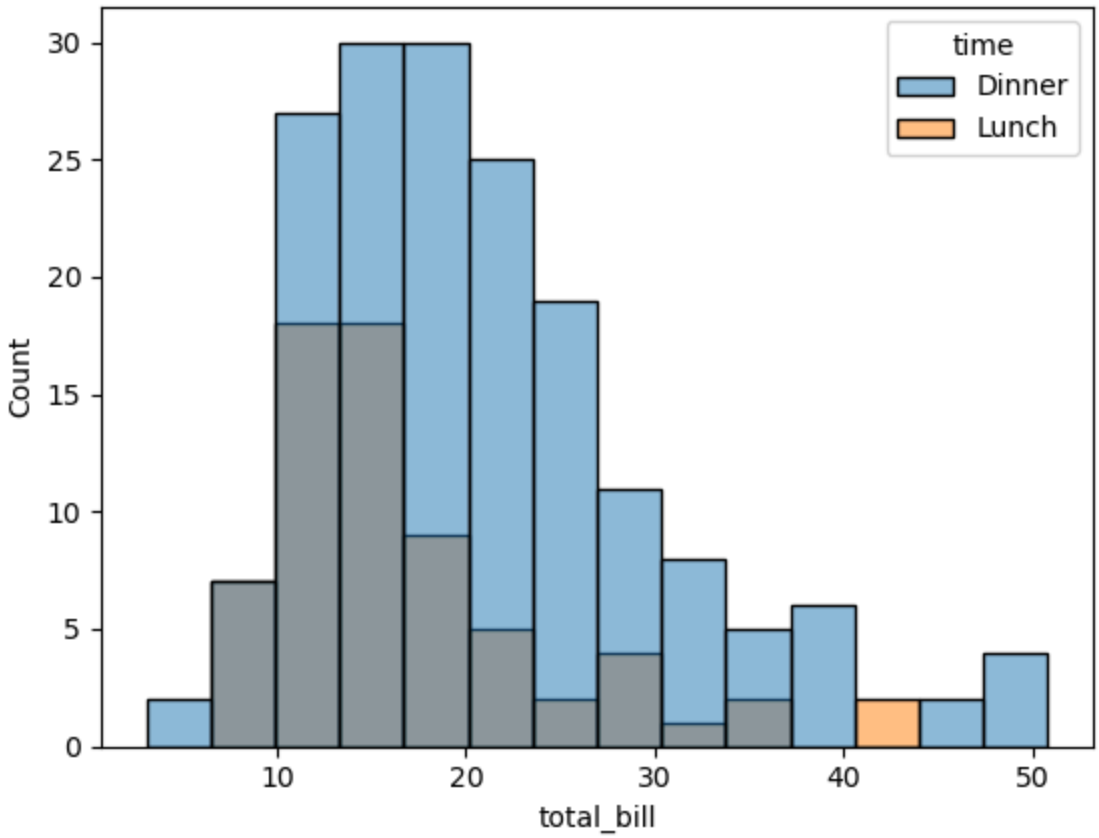
In [17]:
```
sns.boxplot(df,x='day',y='total_bill',hue='sex')
```

Out[17]:
```
<Axes: xlabel='day', ylabel='total_bill'>
```



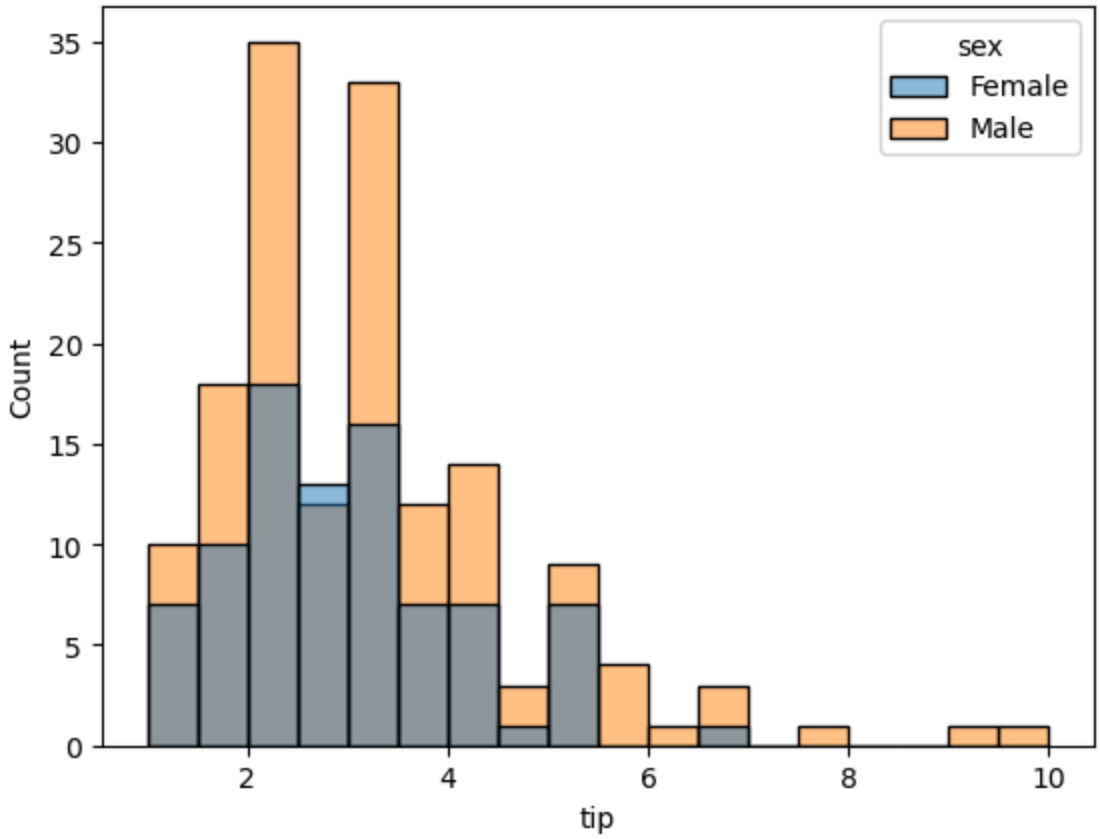
In [18]:
```
sns.histplot(df,x='total_bill',hue='time')
```

Out[18]:
```
<Axes: xlabel='total_bill', ylabel='Count'>
```



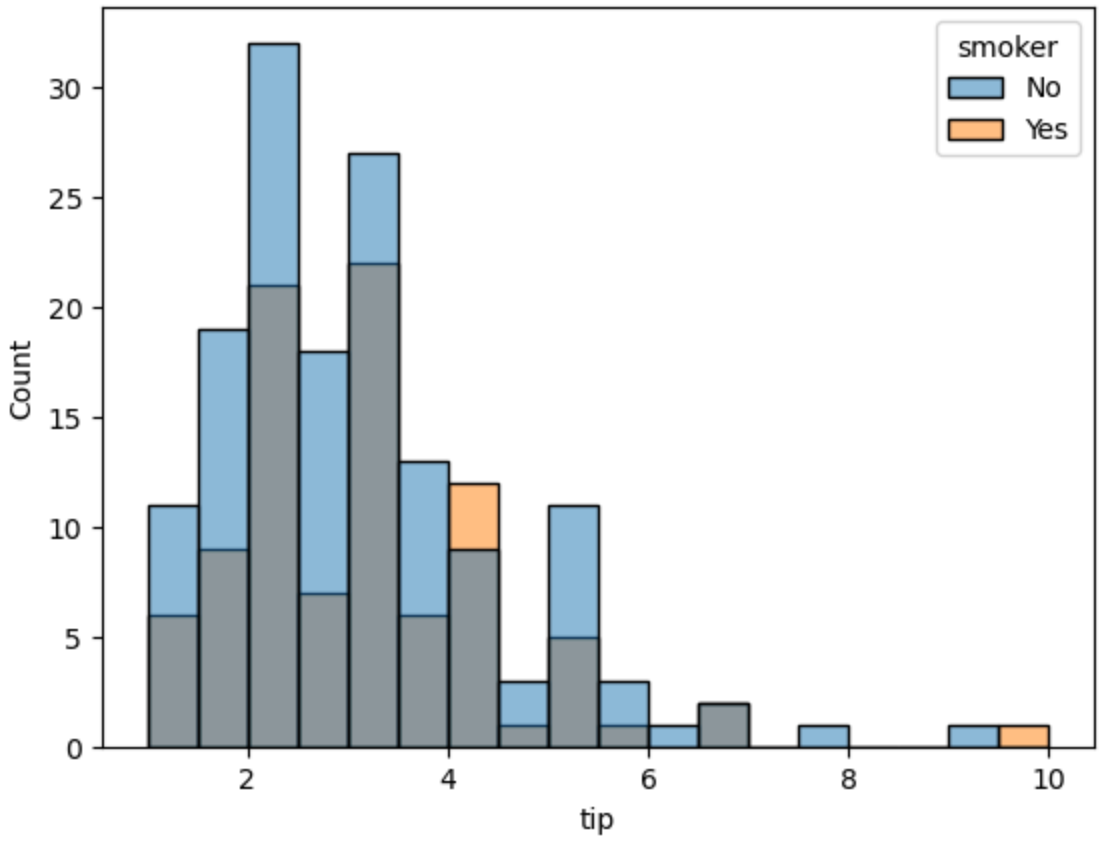
In [19]:
```
sns.histplot(df,x='tip',hue='sex')
```

Out[19]: <Axes: xlabel='tip', ylabel='Count'>



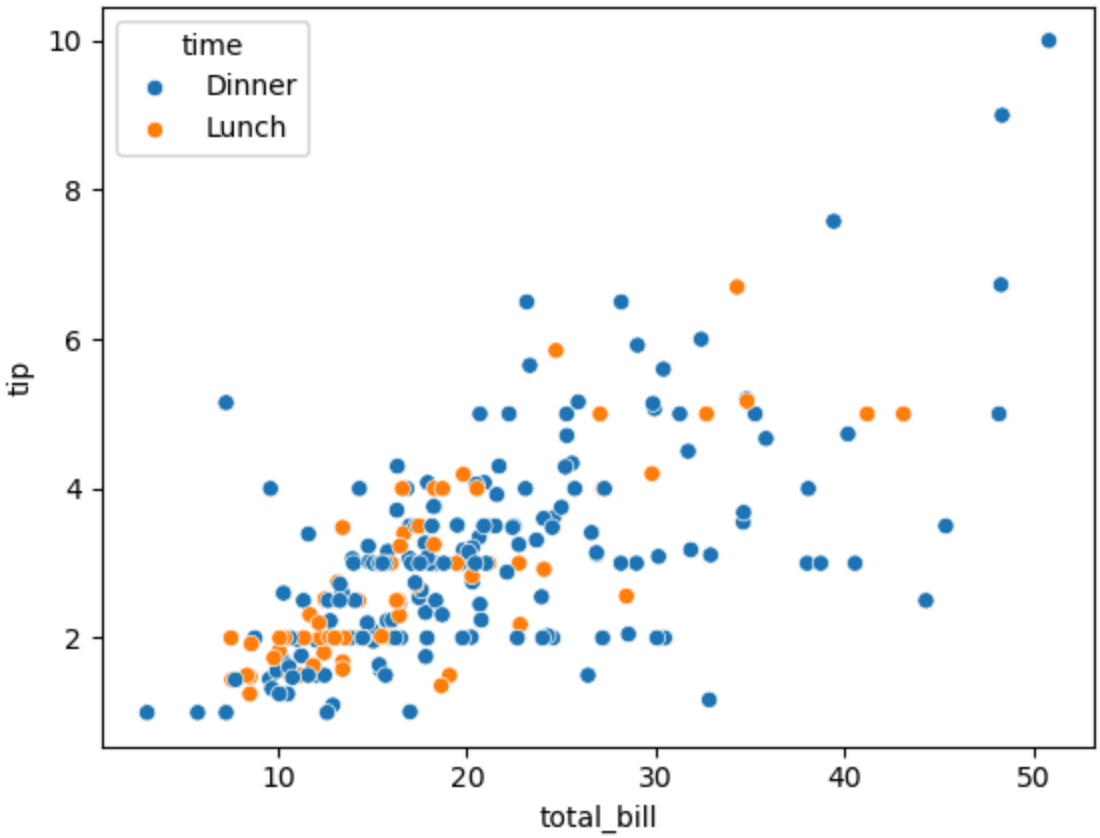
In [20]:
```
sns.histplot(df,x='tip',hue='smoker')
```

Out[20]: <Axes: xlabel='tip', ylabel='Count'>



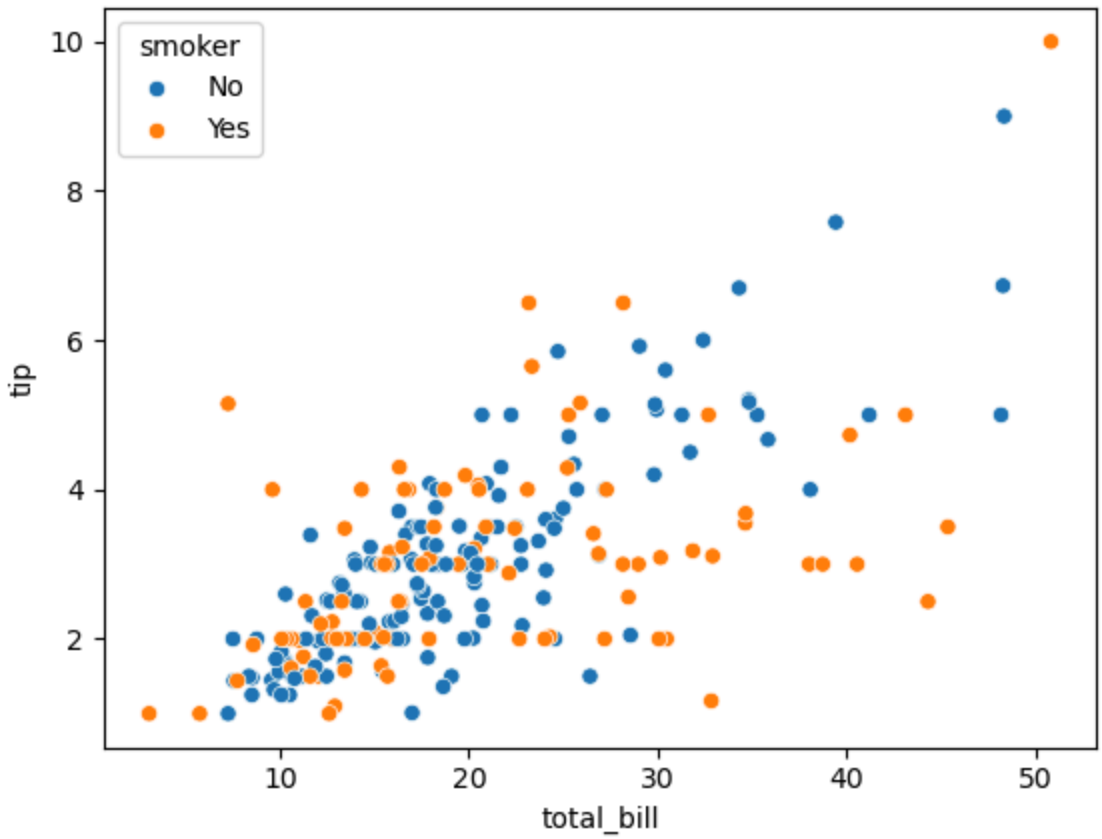
In [21]:
```
sns.scatterplot(df,x='total_bill',y='tip',hue='time')
```

Out[21]:
```
<Axes: xlabel='total_bill', ylabel='tip'>
```



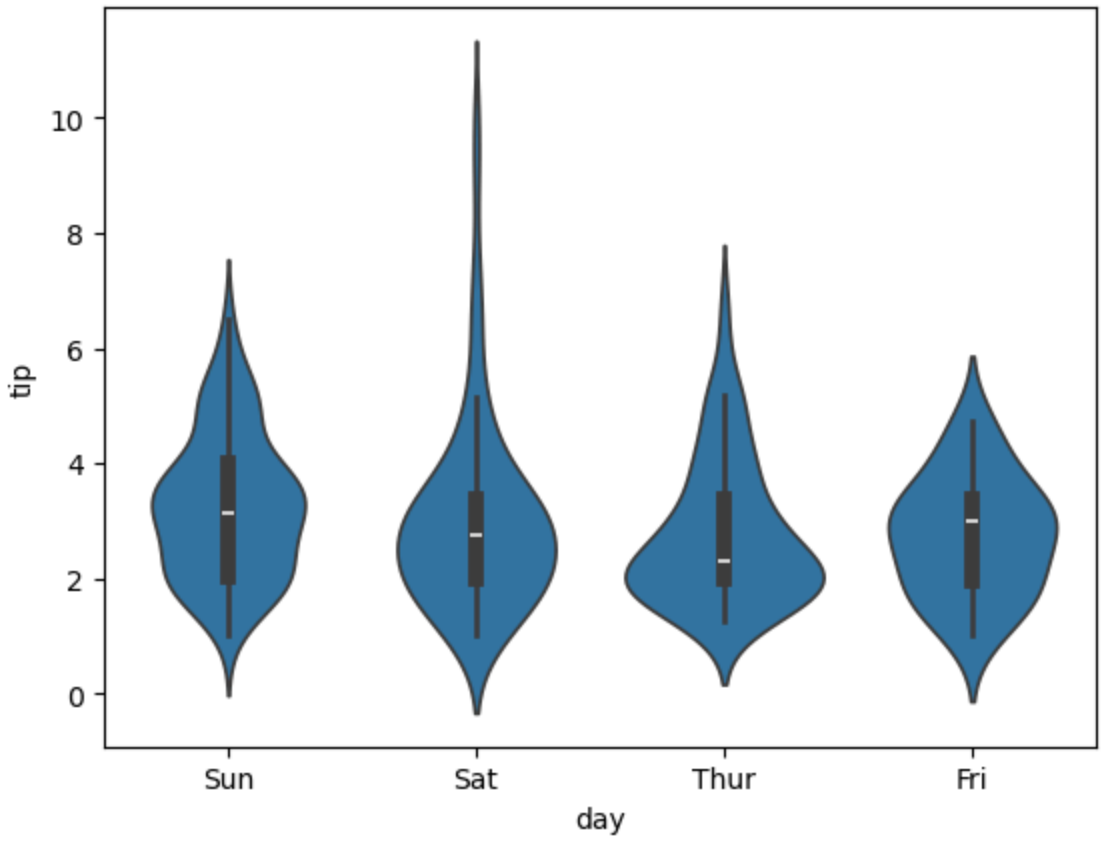
In [22]:
```
sns.scatterplot(df,x='total_bill',y='tip',hue='smoker')
```

Out[22]: <Axes: xlabel='total_bill', ylabel='tip'>



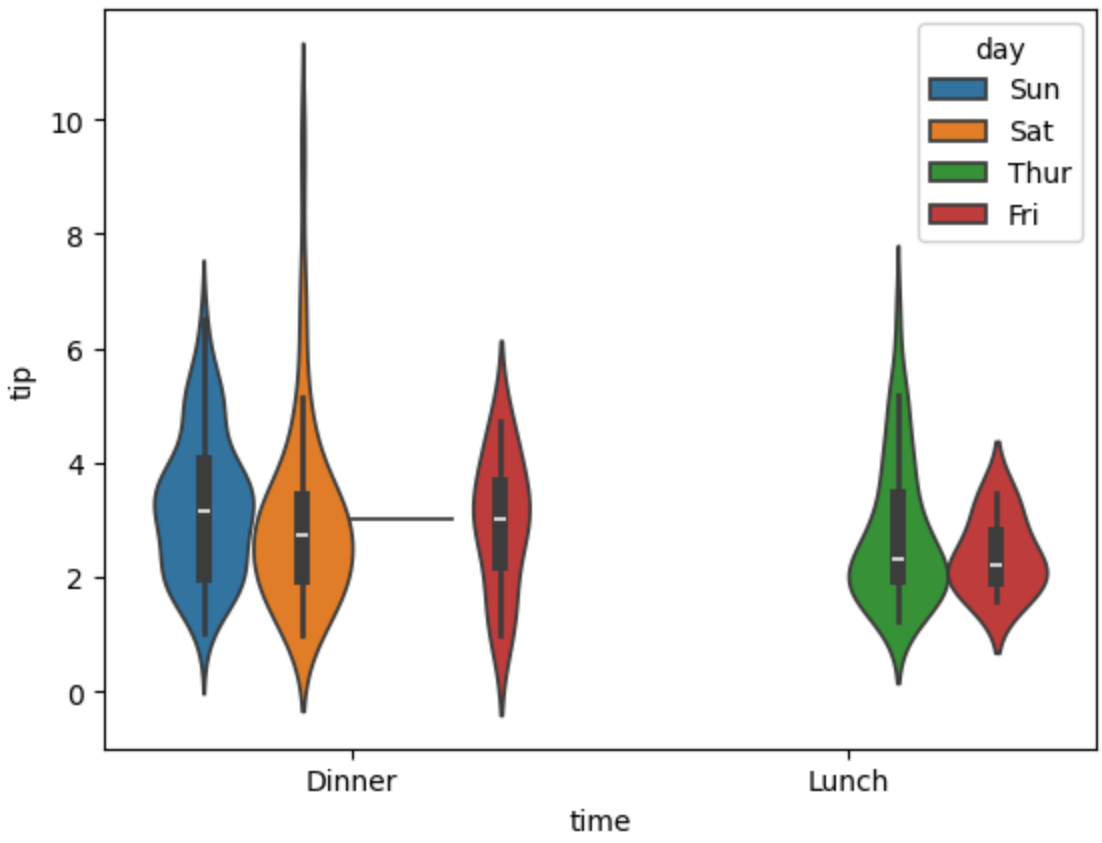
In [23]:
```
sns.violinplot(df,x ='day', y ='tip')
```

Out[23]: <Axes: xlabel='day', ylabel='tip'>



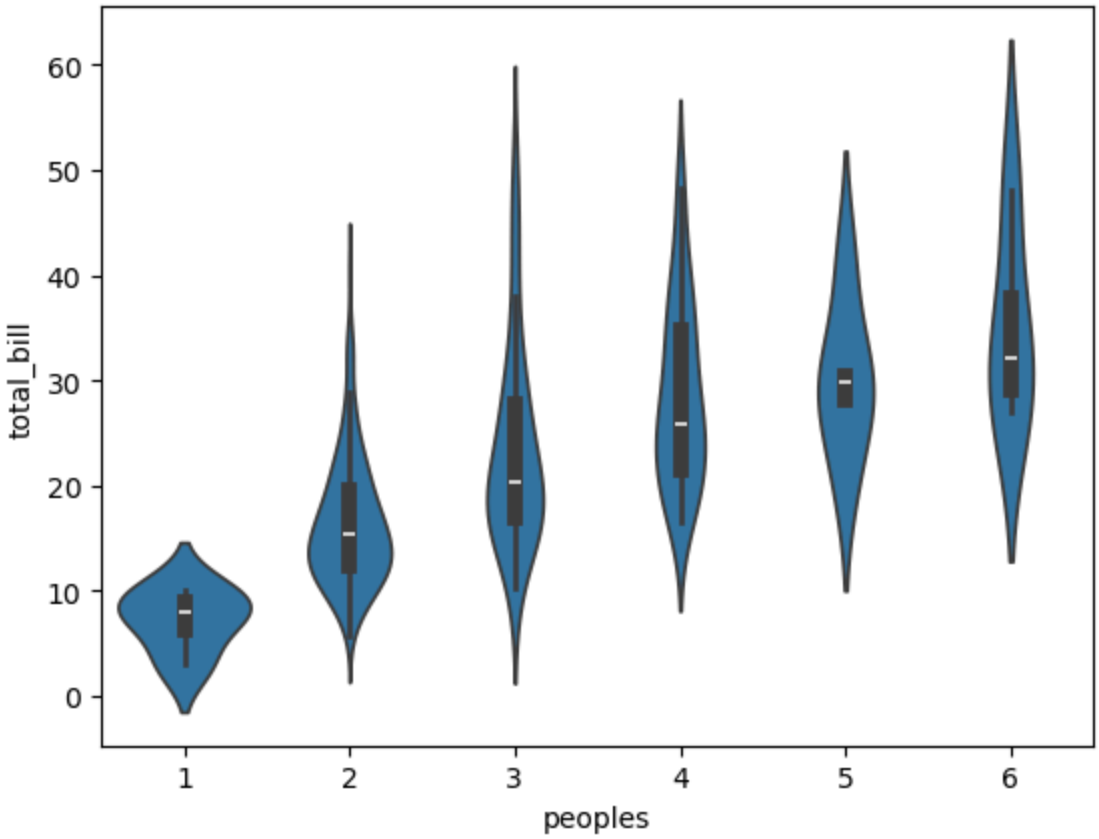
In [24]:
```
sns.violinplot(df,x ='time', y ='tip',hue='day')
```

Out[24]: <Axes: xlabel='time', ylabel='tip'>



In [25]:
```
sns.violinplot(df,x ='peoples', y ='total_bill')
```

Out[25]:
```
<Axes: xlabel='peoples', ylabel='total_bill'>
```



In [ ]:
```

```

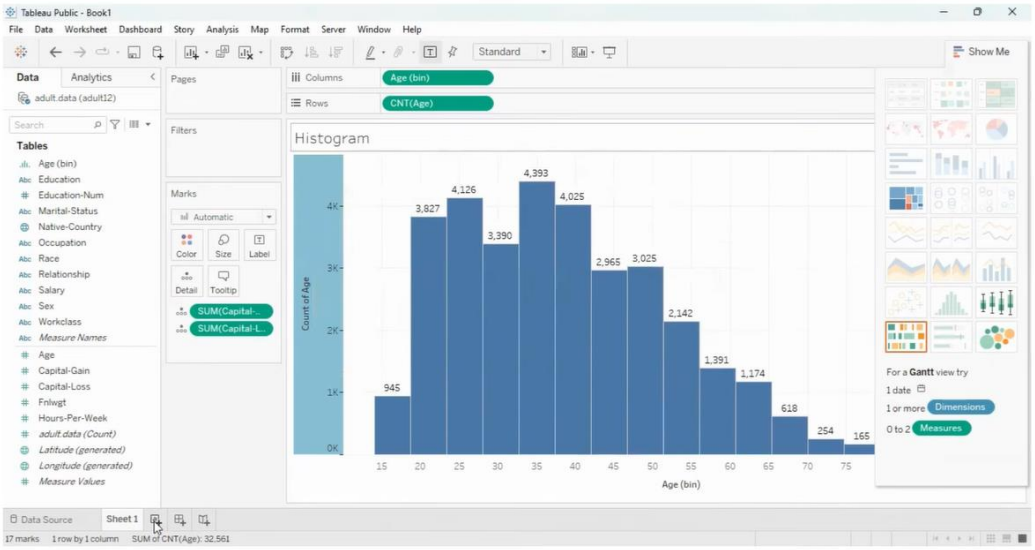


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
adult.data (adult12)
Search
Tables
Age (bin)
Education
Education-Num
Marital-Status
Native-Country
Occupation
Race
Relationship
Salary
Sex
Workclass
Measure Names
Age
Capital-Gain
Capital-Loss
Fnlwgt
Hours-Per-Week
adult.data (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Capital-...
SUM(Capital-L...
Columns
Age (bin)
Rows
CNT(Age)
Histogram
Count of Age
4K
3K
2K
1K
0K
945
3,827
4,126
3,390
4,393
4,025
2,965
3,025
2,142
1,391
1,174
618
254
165
15
20
25
30
35
40
45
50
55
60
65
70
75
Age (bin)
For a Gantt view try
1 date
1 or more Dimensions
0 to 2 Measures
Data Source
Sheet 1
17 marks
1 row by 1 column
SUM of CNT(Age): 32,561

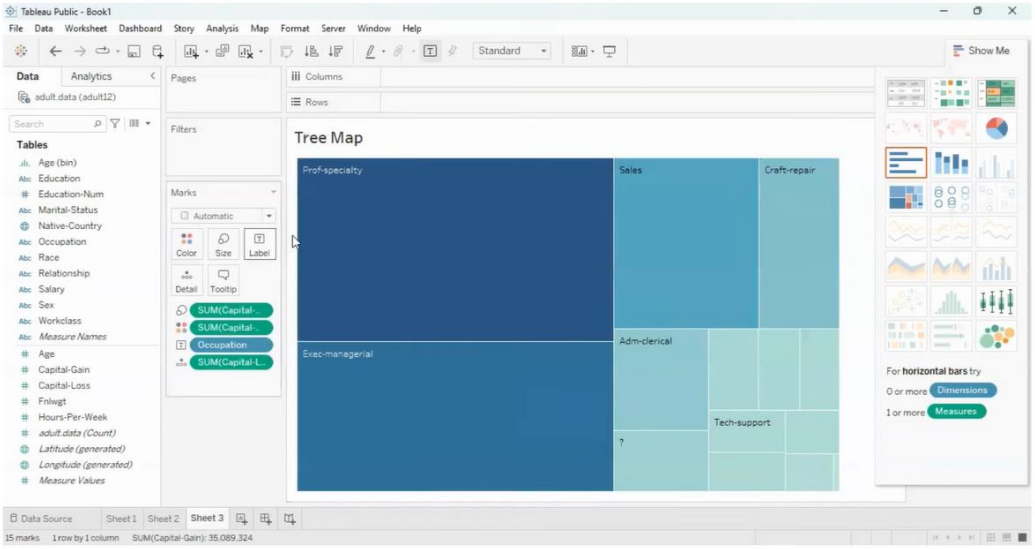


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
adult.data (adult12)
Search
Tables
Age (bin)
Education
Education-Num
Marital-Status
Native-Country
Occupation
Race
Relationship
Salary
Sex
Workclass
Measure Names
Age
Capital-Gain
Capital-Loss
Fnlwgt
Hours-Per-Week
adult.data (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Capital-...
SUM(Capital-...
Occupation
SUM(Capital-L...
Columns
Rows
Tree Map
Prof-specialty
Sales
Craft-repair
Exec-managerial
Adm-clerical
Tech-support
?
For horizontal bars try
0 or more Dimensions
1 or more Measures
Data Source
Sheet 1
Sheet 2
Sheet 3
15 marks
1 row by 1 column
SUM(Capital-Gain): 35,089,324

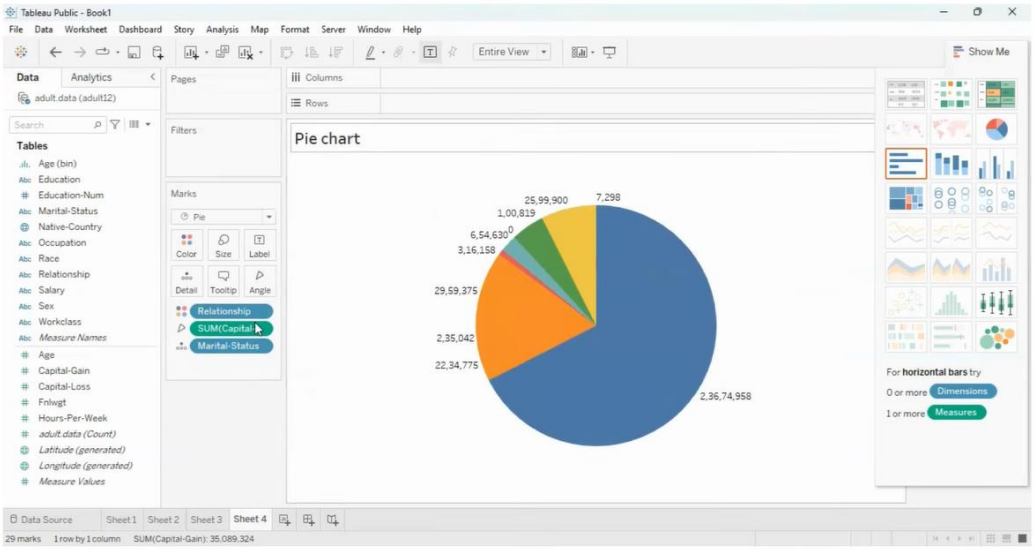


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Entire View
Show Me
Data
Analytics
adult.data (adult12)
Tables
Age (bin)
Education
Education-Num
Marital-Status
Native-Country
Occupation
Race
Relationship
Salary
Sex
Workclass
Measure Names
Age
Capital-Gain
Capital-Loss
Fnlwgt
Hours-Per-Week
adult.data (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Pie
Color
Size
Label
Detail
Tooltip
Angle
Relationship
SUM(Capital-
Marital-Status
Columns
Rows
Pie chart
25,99,900
7,298
1,00,819
6,54,630
3,16,158
29,59,375
2,35,042
22,34,775
2,36,74,958
For horizontal bars try
0 or more Dimensions
1 or more Measures
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
29 marks
1 row by 1 column
SUM(Capital-Gain): 35,089,324

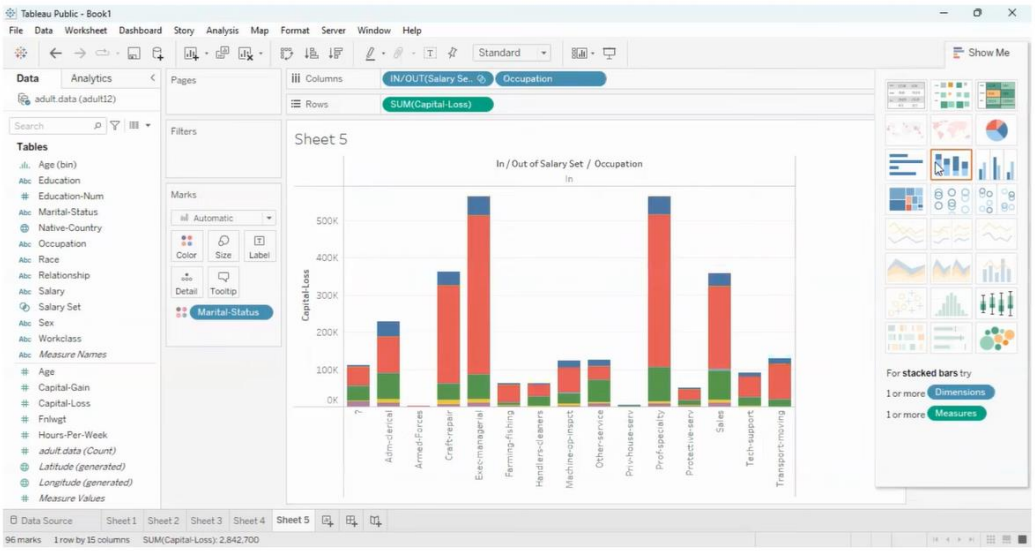


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
adult.data (adult12)
Tables
Age (bin)
Education
Education-Num
Marital-Status
Native-Country
Occupation
Race
Relationship
Salary
Salary Set
Sex
Workclass
Measure Names
Age
Capital-Gain
Capital-Loss
Fnlwgt
Hours-Per-Week
adult.data (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Marital-Status
Columns
IN/OUT(Salary Se..
Occupation
Rows
SUM(Capital-Loss)
Sheet 5
In / Out of Salary Set / Occupation
In
Capital-Loss
500K
400K
300K
200K
100K
0K
?
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
For stacked bars try
1 or more Dimensions
1 or more Measures
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
96 marks
1 row by 15 columns
SUM(Capital-Loss): 2,842,700

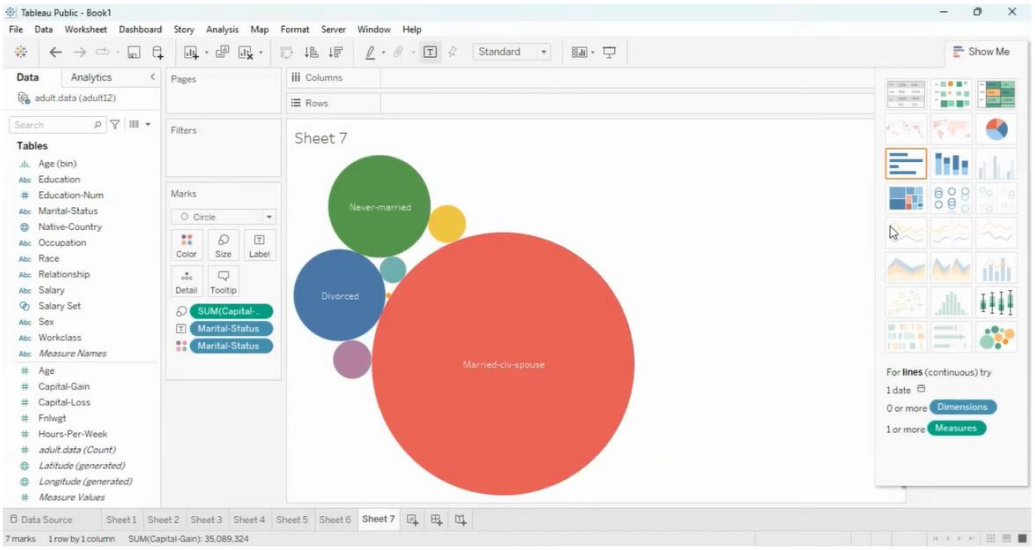


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
adult.data (adult12)
Search
Tables
Age (bin)
Education
Education-Num
Marital-Status
Native-Country
Occupation
Race
Relationship
Salary
Salary Set
Sex
Workclass
Measure Names
Age
Capital-Gain
Capital-Loss
Fnlwgt
Hours-Per-Week
adult.data (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Circle
Color
Size
Label
Detail
Tooltip
SUM(Capital-..
Marital-Status
Marital-Status
Columns
Rows
Sheet 7
Never-married
Divorced
Married-civ-spouse
For lines (continuous) try
1 date
0 or more Dimensions
1 or more Measures
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
Sheet 6
Sheet 7
7 marks
1 row by 1 column
SUM(Capital-Gain): 35,089,324

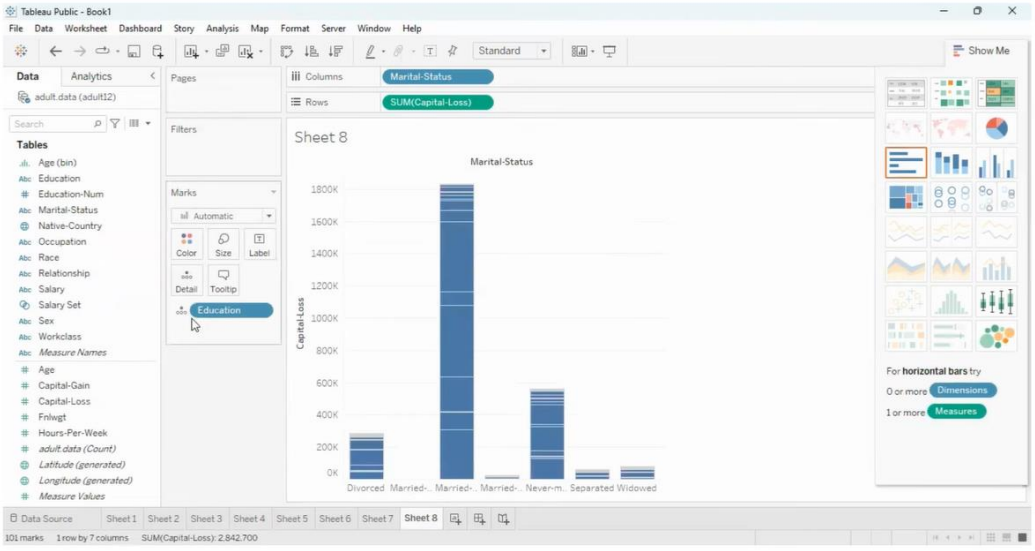


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
adult.data (adult12)
Search
Tables
Age (bin)
Education
Education-Num
Marital-Status
Native-Country
Occupation
Race
Relationship
Salary
Salary Set
Sex
Workclass
Measure Names
Age
Capital-Gain
Capital-Loss
Fnlwgt
Hours-Per-Week
adult.data (Count)
Latitude (generated)
Longitude (generated)
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Education
Columns
Marital-Status
Rows
SUM(Capital-Loss)
Sheet 8
Marital-Status
Capital-Loss
1800K
1600K
1400K
1200K
1000K
800K
600K
400K
200K
0K
Divorced
Married-..
Married-..
Married-..
Never-m..
Separated
Widowed
For horizontal bars try
0 or more Dimensions
1 or more Measures
Data Source
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
Sheet 6
Sheet 7
Sheet 8
101 marks
1 row by 7 columns
SUM(Capital-Loss): 2,842,700

In [58]:
```
import pandas as pd
import requests
from bs4 import BeautifulSoup
```

In [59]:
```
req = requests.get('https://books.toscrape.com/catalogue/page-1.html')
```

In [60]:
```
print(req)
```

```
<Response [200]>
```

In [61]:
```
soup = BeautifulSoup(req.text, 'html.parser')
```

In [62]:
```
print(soup.prettify())
```

```
<!DOCTYPE html>
<!--[if lt IE 7]>      <html lang="en-us" class="no-js lt-ie9 lt-ie8 lt-ie7"> <![end
if]-->
<!--[if IE 7]>         <html lang="en-us" class="no-js lt-ie9 lt-ie8"> <![endif]-->
<!--[if IE 8]>         <html lang="en-us" class="no-js lt-ie9"> <![endif]-->
<!--[if gt IE 8]><!-->
<html class="no-js" lang="en-us">
 <!--<![endif]-->
 <head>
  <title>
   All products | Books to Scrape - Sandbox
  </title>
  <meta content="text/html; charset=utf-8" http-equiv="content-type"/>
  <meta content="24th Jun 2016 09:30" name="created"/>
  <meta content="" name="description"/>
  <meta content="width=device-width" name="viewport"/>
  <meta content="NOARCHIVE,NOCACHE" name="robots"/>
  <!-- Le HTML5 shim, for IE6-8 support of HTML elements -->
  <!--[if lt IE 9]>
        <script src="//html5shim.googlecode.com/svn/trunk/html5.js"></script>
        <![endif]-->
  <link href="../static/oscar/favicon.ico" rel="shortcut icon"/>
  <link href="../static/oscar/css/styles.css" rel="stylesheet" type="text/css"/>
  <link href="../static/oscar/js/bootstrap-datetimepicker/bootstrap-datetimepicker.c
ss" rel="stylesheet"/>
  <link href="../static/oscar/css/datetimepicker.css" rel="stylesheet" type="text/cs
s"/>
 </head>
 <body class="default" id="default">
  <header class="header container-fluid">
   <div class="page_inner">
    <div class="row">
     <div class="col-sm-8 h1">
      <a href="../index.html">
       Books to Scrape
      </a>
      <small>
       We love being scraped!
      </small>
     </div>
    </div>
   </div>
  </header>
  <div class="container-fluid page">
   <div class="page_inner">
    <ul class="breadcrumb">
     <li>
      <a href="../index.html">
       Home
      </a>
     </li>
     <li class="active">
      All products
     </li>
    </ul>
    <div class="row">
```

```
<aside class="sidebar col-sm-4 col-md-3">
 <div id="promotions_left">
 </div>
 <div class="side_categories">
  <ul class="nav nav-list">
   <li>
    <a href="category/books_1/index.html">
     Books
    </a>
    <ul>
     <li>
      <a href="category/books/travel_2/index.html">
       Travel
      </a>
     </li>
     <li>
      <a href="category/books/mystery_3/index.html">
       Mystery
      </a>
     </li>
     <li>
      <a href="category/books/historical-fiction_4/index.html">
       Historical Fiction
      </a>
     </li>
     <li>
      <a href="category/books/sequential-art_5/index.html">
       Sequential Art
      </a>
     </li>
     <li>
      <a href="category/books/classics_6/index.html">
       Classics
      </a>
     </li>
     <li>
      <a href="category/books/philosophy_7/index.html">
       Philosophy
      </a>
     </li>
     <li>
      <a href="category/books/romance_8/index.html">
       Romance
      </a>
     </li>
     <li>
      <a href="category/books/womens-fiction_9/index.html">
       Womens Fiction
      </a>
     </li>
     <li>
      <a href="category/books/fiction_10/index.html">
       Fiction
      </a>
     </li>
     <li>
```

```
 <a href="category/books/childrens_11/index.html">
  Childrens
 </a>
</li>
<li>
 <a href="category/books/religion_12/index.html">
  Religion
 </a>
</li>
<li>
 <a href="category/books/nonfiction_13/index.html">
  Nonfiction
 </a>
</li>
<li>
 <a href="category/books/music_14/index.html">
  Music
 </a>
</li>
<li>
 <a href="category/books/default_15/index.html">
  Default
 </a>
</li>
<li>
 <a href="category/books/science-fiction_16/index.html">
  Science Fiction
 </a>
</li>
<li>
 <a href="category/books/sports-and-games_17/index.html">
  Sports and Games
 </a>
</li>
<li>
 <a href="category/books/add-a-comment_18/index.html">
  Add a comment
 </a>
</li>
<li>
 <a href="category/books/fantasy_19/index.html">
  Fantasy
 </a>
</li>
<li>
 <a href="category/books/new-adult_20/index.html">
  New Adult
 </a>
</li>
<li>
 <a href="category/books/young-adult_21/index.html">
  Young Adult
 </a>
</li>
<li>
 <a href="category/books/science_22/index.html">
```

```
  Science
 </a>
</li>
<li>
 <a href="category/books/poetry_23/index.html">
  Poetry
 </a>
</li>
<li>
 <a href="category/books/paranormal_24/index.html">
  Paranormal
 </a>
</li>
<li>
 <a href="category/books/art_25/index.html">
  Art
 </a>
</li>
<li>
 <a href="category/books/psychology_26/index.html">
  Psychology
 </a>
</li>
<li>
 <a href="category/books/autobiography_27/index.html">
  Autobiography
 </a>
</li>
<li>
 <a href="category/books/parenting_28/index.html">
  Parenting
 </a>
</li>
<li>
 <a href="category/books/adult-fiction_29/index.html">
  Adult Fiction
 </a>
</li>
<li>
 <a href="category/books/humor_30/index.html">
  Humor
 </a>
</li>
<li>
 <a href="category/books/horror_31/index.html">
  Horror
 </a>
</li>
<li>
 <a href="category/books/history_32/index.html">
  History
 </a>
</li>
<li>
 <a href="category/books/food-and-drink_33/index.html">
  Food and Drink
```

```
 </a>
</li>
<li>
 <a href="category/books/christian-fiction_34/index.html">
  Christian Fiction
 </a>
</li>
<li>
 <a href="category/books/business_35/index.html">
  Business
 </a>
</li>
<li>
 <a href="category/books/biography_36/index.html">
  Biography
 </a>
</li>
<li>
 <a href="category/books/thriller_37/index.html">
  Thriller
 </a>
</li>
<li>
 <a href="category/books/contemporary_38/index.html">
  Contemporary
 </a>
</li>
<li>
 <a href="category/books/spirituality_39/index.html">
  Spirituality
 </a>
</li>
<li>
 <a href="category/books/academic_40/index.html">
  Academic
 </a>
</li>
<li>
 <a href="category/books/self-help_41/index.html">
  Self Help
 </a>
</li>
<li>
 <a href="category/books/historical_42/index.html">
  Historical
 </a>
</li>
<li>
 <a href="category/books/christian_43/index.html">
  Christian
 </a>
</li>
<li>
 <a href="category/books/suspense_44/index.html">
  Suspense
 </a>
```

```
      </li>
      <li>
       <a href="category/books/short-stories_45/index.html">
        Short Stories
       </a>
      </li>
      <li>
       <a href="category/books/novels_46/index.html">
        Novels
       </a>
      </li>
      <li>
       <a href="category/books/health_47/index.html">
        Health
       </a>
      </li>
      <li>
       <a href="category/books/politics_48/index.html">
        Politics
       </a>
      </li>
      <li>
       <a href="category/books/cultural_49/index.html">
        Cultural
       </a>
      </li>
      <li>
       <a href="category/books/erotica_50/index.html">
        Erotica
       </a>
      </li>
      <li>
       <a href="category/books/crime_51/index.html">
        Crime
       </a>
      </li>
     </ul>
    </li>
   </ul>
  </div>
</aside>
<div class="col-sm-8 col-md-9">
 <div class="page-header action">
  <h1>
   All products
  </h1>
 </div>
 <div id="messages">
 </div>
 <div id="promotions">
 </div>
 <form class="form-horizontal" method="get">
  <div style="display:none">
  </div>
  <strong>
   1000
```

```
        </strong>
        results - showing
        <strong>
         1
        </strong>
        to
        <strong>
         20
        </strong>
        .
       </form>
       <section>
        <div class="alert alert-warning" role="alert">
         <strong>
          Warning!
         </strong>
         This is a demo website for web scraping purposes. Prices and ratings here we
re randomly assigned and have no real meaning.
        </div>
        <div>
         <ol class="row">
          <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
           <article class="product_pod">
            <div class="image_container">
             <a href="a-light-in-the-attic_1000/index.html">
              <img alt="A Light in the Attic" class="thumbnail" src="../media/cache/2
c/da/2cdad67c44b002e7ead0cc35693c0e8b.jpg"/>
             </a>
            </div>
            <p class="star-rating Three">
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
            </p>
            <h3>
             <a href="a-light-in-the-attic_1000/index.html" title="A Light in the Att
ic">
              A Light in the ...
             </a>
            </h3>
            <div class="product_price">
             <p class="price_color">
              Â£51.77
             </p>
             <p class="instock availability">
              <i class="icon-ok">
              </i>
              In stock
             </p>
```

```
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="tipping-the-velvet_999/index.html">
             <img alt="Tipping the Velvet" class="thumbnail" src="../media/cache/26/
0c/260c6ae16bce31c8f8c95daddd9f4a1c.jpg"/>
            </a>
           </div>
           <p class="star-rating One">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="tipping-the-velvet_999/index.html" title="Tipping the Velvet">
             Tipping the Velvet
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£53.74
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="soumission_998/index.html">
```

```
              <img alt="Soumission" class="thumbnail" src="../media/cache/3e/ef/3eef9
9c9d9adef34639f510662022830.jpg"/>
             </a>
            </div>
            <p class="star-rating One">
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
            </p>
            <h3>
             <a href="soumission_998/index.html" title="Soumission">
              Soumission
             </a>
            </h3>
            <div class="product_price">
             <p class="price_color">
              Â£50.10
             </p>
             <p class="instock availability">
              <i class="icon-ok">
              </i>
              In stock
             </p>
             <form>
              <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
               Add to basket
              </button>
             </form>
            </div>
           </article>
          </li>
          <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
           <article class="product_pod">
            <div class="image_container">
             <a href="sharp-objects_997/index.html">
              <img alt="Sharp Objects" class="thumbnail" src="../media/cache/32/51/32
51cf3a3412f53f339e42cac2134093.jpg"/>
             </a>
            </div>
            <p class="star-rating Four">
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
```

```
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="sharp-objects_997/index.html" title="Sharp Objects">
             Sharp Objects
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£47.82
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="sapiens-a-brief-history-of-humankind_996/index.html">
             <img alt="Sapiens: A Brief History of Humankind" class="thumbnail" src
="../media/cache/be/a5/bea5697f2534a2f86a3ef27b5a8c12a6.jpg"/>
            </a>
           </div>
           <p class="star-rating Five">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="sapiens-a-brief-history-of-humankind_996/index.html" title="Sap
iens: A Brief History of Humankind">
             Sapiens: A Brief History ...
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£54.23
            </p>
```

```
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="the-requiem-red_995/index.html">
             <img alt="The Requiem Red" class="thumbnail" src="../media/cache/68/33/
68339b4c9bc034267e1da611ab3b34f8.jpg"/>
            </a>
           </div>
           <p class="star-rating One">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="the-requiem-red_995/index.html" title="The Requiem Red">
             The Requiem Red
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£22.65
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
```

```
        </li>
        <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
         <article class="product_pod">
          <div class="image_container">
           <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.ht
ml">
            <img alt="The Dirty Little Secrets of Getting Your Dream Job" class="th
umbnail" src="../media/cache/92/27/92274a95b7c251fea59a2b8a78275ab4.jpg"/>
           </a>
          </div>
          <p class="star-rating Four">
           <i class="icon-star">
           </i>
           <i class="icon-star">
           </i>
           <i class="icon-star">
           </i>
           <i class="icon-star">
           </i>
           <i class="icon-star">
           </i>
          </p>
          <h3>
           <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.ht
ml" title="The Dirty Little Secrets of Getting Your Dream Job">
            The Dirty Little Secrets ...
           </a>
          </h3>
          <div class="product_price">
           <p class="price_color">
            Â£33.34
           </p>
           <p class="instock availability">
            <i class="icon-ok">
            </i>
            In stock
           </p>
           <form>
            <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
             Add to basket
            </button>
           </form>
          </div>
         </article>
        </li>
        <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
         <article class="product_pod">
          <div class="image_container">
           <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-femi
nist-victoria-woodhull_993/index.html">
            <img alt="The Coming Woman: A Novel Based on the Life of the Infamous F
eminist, Victoria Woodhull" class="thumbnail" src="../media/cache/3d/54/3d54940e57e6
62c4dd1f3ff00c78cc64.jpg"/>
           </a>
          </div>
```

```
            <p class="star-rating Three">
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
            </p>
            <h3>
             <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-femi
nist-victoria-woodhull_993/index.html" title="The Coming Woman: A Novel Based on the
Life of the Infamous Feminist, Victoria Woodhull">
              The Coming Woman: A ...
             </a>
            </h3>
            <div class="product_price">
             <p class="price_color">
              Â£17.93
             </p>
             <p class="instock availability">
              <i class="icon-ok">
              </i>
              In stock
             </p>
             <form>
              <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
               Add to basket
              </button>
             </form>
            </div>
           </article>
          </li>
          <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
           <article class="product_pod">
            <div class="image_container">
             <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-go
ld-at-the-1936-berlin-olympics_992/index.html">
              <img alt="The Boys in the Boat: Nine Americans and Their Epic Quest for
Gold at the 1936 Berlin Olympics" class="thumbnail" src="../media/cache/66/88/66883b
91f6804b2323c8369331cb7dd1.jpg"/>
             </a>
            </div>
            <p class="star-rating Four">
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
```

```
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-go
ld-at-the-1936-berlin-olympics_992/index.html" title="The Boys in the Boat: Nine Ame
ricans and Their Epic Quest for Gold at the 1936 Berlin Olympics">
             The Boys in the ...
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£22.60
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="the-black-maria_991/index.html">
             <img alt="The Black Maria" class="thumbnail" src="../media/cache/58/46/
5846057e28022268153beff6d352b06c.jpg"/>
            </a>
           </div>
           <p class="star-rating One">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="the-black-maria_991/index.html" title="The Black Maria">
             The Black Maria
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£52.15
```

```
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html">
             <img alt="Starving Hearts (Triangular Trade Trilogy, #1)" class="thumbn
ail" src="../media/cache/be/f4/bef44da28c98f905a3ebec0b87be8530.jpg"/>
            </a>
           </div>
           <p class="star-rating Two">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html" titl
e="Starving Hearts (Triangular Trade Trilogy, #1)">
             Starving Hearts (Triangular Trade ...
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£13.99
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
```

```
            </div>
           </article>
          </li>
          <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
           <article class="product_pod">
            <div class="image_container">
             <a href="shakespeares-sonnets_989/index.html">
              <img alt="Shakespeare's Sonnets" class="thumbnail" src="../media/cache/
10/48/1048f63d3b5061cd2f424d20b3f9b666.jpg"/>
             </a>
            </div>
            <p class="star-rating Four">
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
             <i class="icon-star">
             </i>
            </p>
            <h3>
             <a href="shakespeares-sonnets_989/index.html" title="Shakespeare's Sonne
ts">
              Shakespeare's Sonnets
             </a>
            </h3>
            <div class="product_price">
             <p class="price_color">
              Â£20.66
             </p>
             <p class="instock availability">
              <i class="icon-ok">
              </i>
              In stock
             </p>
             <form>
              <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
               Add to basket
              </button>
             </form>
            </div>
           </article>
          </li>
          <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
           <article class="product_pod">
            <div class="image_container">
             <a href="set-me-free_988/index.html">
              <img alt="Set Me Free" class="thumbnail" src="../media/cache/5b/88/5b88
c52633f53cacf162c15f4f823153.jpg"/>
             </a>
            </div>
            <p class="star-rating Five">
```

```
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
               </p>
               <h3>
                <a href="set-me-free_988/index.html" title="Set Me Free">
                 Set Me Free
                </a>
               </h3>
               <div class="product_price">
                <p class="price_color">
                 Â£17.46
                </p>
                <p class="instock availability">
                 <i class="icon-ok">
                 </i>
                 In stock
                </p>
                <form>
                 <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
                  Add to basket
                 </button>
                </form>
               </div>
              </article>
             </li>
             <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
              <article class="product_pod">
               <div class="image_container">
                <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.h
tml">
                 <img alt="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)" clas
s="thumbnail" src="../media/cache/94/b1/94b1b8b244bce9677c2f29ccc890d4d2.jpg"/>
                </a>
               </div>
               <p class="star-rating Five">
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
               </p>
               <h3>
```

```
            <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.h
tml" title="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)">
             Scott Pilgrim's Precious Little ...
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£52.29
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="rip-it-up-and-start-again_986/index.html">
             <img alt="Rip it Up and Start Again" class="thumbnail" src="../media/ca
che/81/c4/81c4a973364e17d01f217e1188253d5e.jpg"/>
            </a>
           </div>
           <p class="star-rating Five">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="rip-it-up-and-start-again_986/index.html" title="Rip it Up and
Start Again">
             Rip it Up and ...
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£35.02
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
```

```
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="our-band-could-be-your-life-scenes-from-the-american-indie-unde
rground-1981-1991_985/index.html">
             <img alt="Our Band Could Be Your Life: Scenes from the American Indie U
nderground, 1981-1991" class="thumbnail" src="../media/cache/54/60/54607fe8945897cdc
ced0044103b10b6.jpg"/>
            </a>
           </div>
           <p class="star-rating Three">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="our-band-could-be-your-life-scenes-from-the-american-indie-unde
rground-1981-1991_985/index.html" title="Our Band Could Be Your Life: Scenes from th
e American Indie Underground, 1981-1991">
             Our Band Could Be ...
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£57.25
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
```

```
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="olio_984/index.html">
             <img alt="Olio" class="thumbnail" src="../media/cache/55/33/553310a7162
dfbc2c6d19a84da0df9e1.jpg"/>
            </a>
           </div>
           <p class="star-rating One">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="olio_984/index.html" title="Olio">
             Olio
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£23.88
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.
html">
             <img alt="Mesaerion: The Best Science Fiction Stories 1800-1849" class
="thumbnail" src="../media/cache/09/a3/09a3aef48557576e1a85ba7efea8ecb7.jpg"/>
            </a>
           </div>
           <p class="star-rating One">
            <i class="icon-star">
```

```
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
               </p>
               <h3>
                <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.
html" title="Mesaerion: The Best Science Fiction Stories 1800-1849">
                 Mesaerion: The Best Science ...
                </a>
               </h3>
               <div class="product_price">
                <p class="price_color">
                 Â£37.59
                </p>
                <p class="instock availability">
                 <i class="icon-ok">
                 </i>
                 In stock
                </p>
                <form>
                 <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
                  Add to basket
                 </button>
                </form>
               </div>
              </article>
             </li>
             <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
              <article class="product_pod">
               <div class="image_container">
                <a href="libertarianism-for-beginners_982/index.html">
                 <img alt="Libertarianism for Beginners" class="thumbnail" src="../medi
a/cache/0b/bc/0bbcd0a6f4bcd81ccb1049a52736406e.jpg"/>
                </a>
               </div>
               <p class="star-rating Two">
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
                <i class="icon-star">
                </i>
               </p>
               <h3>
                <a href="libertarianism-for-beginners_982/index.html" title="Libertarian
```

```
ism for Beginners">
             Libertarianism for Beginners
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£51.33
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
         <li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
          <article class="product_pod">
           <div class="image_container">
            <a href="its-only-the-himalayas_981/index.html">
             <img alt="It's Only the Himalayas" class="thumbnail" src="../media/cach
e/27/a5/27a53d0bb95bdd88288eaf66c9230d7e.jpg"/>
            </a>
           </div>
           <p class="star-rating Two">
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
            <i class="icon-star">
            </i>
           </p>
           <h3>
            <a href="its-only-the-himalayas_981/index.html" title="It's Only the Him
alayas">
             It's Only the Himalayas
            </a>
           </h3>
           <div class="product_price">
            <p class="price_color">
             Â£45.17
            </p>
            <p class="instock availability">
             <i class="icon-ok">
             </i>
             In stock
```

```
            </p>
            <form>
             <button class="btn btn-primary btn-block" data-loading-text="Adding..."
type="submit">
              Add to basket
             </button>
            </form>
           </div>
          </article>
         </li>
        </ol>
        <div>
         <ul class="pager">
          <li class="current">
           Page 1 of 50
          </li>
          <li class="next">
           <a href="page-2.html">
            next
           </a>
          </li>
         </ul>
        </div>
       </div>
      </section>
     </div>
    </div>
    <!-- /row -->
   </div>
   <!-- /page_inner -->
  </div>
  <!-- /container-fluid -->
  <footer class="footer container-fluid">
  </footer>
  <!-- jQuery -->
  <script src="http://ajax.googleapis.com/ajax/libs/jquery/1.9.1/jquery.min.js">
  </script>
  <script>
   window.jQuery || document.write('<script src="../static/oscar/js/jquery/jquery-1.
9.1.min.js"><\/script>')
  </script>
  <!-- Twitter Bootstrap -->
  <script src="../static/oscar/js/bootstrap3/bootstrap.min.js" type="text/javascrip
t">
  </script>
  <!-- Oscar -->
  <script charset="utf-8" src="../static/oscar/js/oscar/ui.js" type="text/javascrip
t">
  </script>
  <script charset="utf-8" src="../static/oscar/js/bootstrap-datetimepicker/bootstrap
-datetimepicker.js" type="text/javascript">
  </script>
  <script charset="utf-8" src="../static/oscar/js/bootstrap-datetimepicker/locales/b
ootstrap-datetimepicker.all.js" type="text/javascript">
  </script>
  <script type="text/javascript">
```

```
            $(function() {



              oscar.init();

              oscar.search.init();

                      });
            </script>
            <!-- Version: N/A -->
           </body>
          </html>
```

In [63]:
```
soup.select('article')
```

Out[63]:
```
[<article class="product_pod">
 <div class="image_container">
 <a href="a-light-in-the-attic_1000/index.html"><img alt="A Light in the Attic" cl
ass="thumbnail" src="../media/cache/2c/da/2cdad67c44b002e7ead0cc35693c0e8b.jpg"/>
</a>
 </div>
 <p class="star-rating Three">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="a-light-in-the-attic_1000/index.html" title="A Light in the Attic">A
Light in the ...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£51.77</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="tipping-the-velvet_999/index.html"><img alt="Tipping the Velvet" class
="thumbnail" src="../media/cache/26/0c/260c6ae16bce31c8f8c95daddd9f4a1c.jpg"/></a>
 </div>
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="tipping-the-velvet_999/index.html" title="Tipping the Velvet">Tippin
g the Velvet</a></h3>
 <div class="product_price">
 <p class="price_color">Â£53.74</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
```

```
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="soumission_998/index.html"><img alt="Soumission" class="thumbnail" src
="../media/cache/3e/ef/3eef99c9d9adef34639f510662022830.jpg"/></a>
 </div>
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="soumission_998/index.html" title="Soumission">Soumission</a></h3>
 <div class="product_price">
 <p class="price_color">Â£50.10</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="sharp-objects_997/index.html"><img alt="Sharp Objects" class="thumbnail"
src="../media/cache/32/51/3251cf3a3412f53f339e42cac2134093.jpg"/></a>
 </div>
 <p class="star-rating Four">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="sharp-objects_997/index.html" title="Sharp Objects">Sharp Objects</a
></h3>
 <div class="product_price">
 <p class="price_color">Â£47.82</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
```

```
 <article class="product_pod">
 <div class="image_container">
 <a href="sapiens-a-brief-history-of-humankind_996/index.html"><img alt="Sapiens:
A Brief History of Humankind" class="thumbnail" src="../media/cache/be/a5/bea5697f
2534a2f86a3ef27b5a8c12a6.jpg"/></a>
 </div>
 <p class="star-rating Five">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="sapiens-a-brief-history-of-humankind_996/index.html" title="Sapiens:
A Brief History of Humankind">Sapiens: A Brief History ...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£54.23</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="the-requiem-red_995/index.html"><img alt="The Requiem Red" class="thumbn
ail" src="../media/cache/68/33/68339b4c9bc034267e1da611ab3b34f8.jpg"/></a>
 </div>
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="the-requiem-red_995/index.html" title="The Requiem Red">The Requiem
Red</a></h3>
 <div class="product_price">
 <p class="price_color">Â£22.65</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
```

```
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.html"><img
alt="The Dirty Little Secrets of Getting Your Dream Job" class="thumbnail" src
="../media/cache/92/27/92274a95b7c251fea59a2b8a78275ab4.jpg"/></a>
 </div>
 <p class="star-rating Four">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.html" t
itle="The Dirty Little Secrets of Getting Your Dream Job">The Dirty Little Secrets
...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£33.34</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-vict
oria-woodhull_993/index.html"><img alt="The Coming Woman: A Novel Based on the Lif
e of the Infamous Feminist, Victoria Woodhull" class="thumbnail" src="../media/cac
he/3d/54/3d54940e57e662c4dd1f3ff00c78cc64.jpg"/></a>
 </div>
 <p class="star-rating Three">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-
victoria-woodhull_993/index.html" title="The Coming Woman: A Novel Based on the Li
fe of the Infamous Feminist, Victoria Woodhull">The Coming Woman: A ...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£17.93</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
```

```
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the
-1936-berlin-olympics_992/index.html"><img alt="The Boys in the Boat: Nine America
ns and Their Epic Quest for Gold at the 1936 Berlin Olympics" class="thumbnail" sr
c="../media/cache/66/88/66883b91f6804b2323c8369331cb7dd1.jpg"/></a>
 </div>
 <p class="star-rating Four">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at
-the-1936-berlin-olympics_992/index.html" title="The Boys in the Boat: Nine Americ
ans and Their Epic Quest for Gold at the 1936 Berlin Olympics">The Boys in the ...
</a></h3>
 <div class="product_price">
 <p class="price_color">Â£22.60</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="the-black-maria_991/index.html"><img alt="The Black Maria" class="thumbn
ail" src="../media/cache/58/46/5846057e28022268153beff6d352b06c.jpg"/></a>
 </div>
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="the-black-maria_991/index.html" title="The Black Maria">The Black Ma
ria</a></h3>
 <div class="product_price">
 <p class="price_color">Â£52.15</p>
 <p class="instock availability">
 <i class="icon-ok"></i>
```

```
        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html"><img alt="Sta
rving Hearts (Triangular Trade Trilogy, #1)" class="thumbnail" src="../media/cach
e/be/f4/bef44da28c98f905a3ebec0b87be8530.jpg"/></a>
 </div>
 <p class="star-rating Two">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="starving-hearts-triangular-trade-trilogy-1_990/index.html" title="St
arving Hearts (Triangular Trade Trilogy, #1)">Starving Hearts (Triangular Trade
...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£13.99</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="shakespeares-sonnets_989/index.html"><img alt="Shakespeare's Sonnets" cl
ass="thumbnail" src="../media/cache/10/48/1048f63d3b5061cd2f424d20b3f9b666.jpg"/>
</a>
 </div>
 <p class="star-rating Four">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="shakespeares-sonnets_989/index.html" title="Shakespeare's Sonnets">S
hakespeare's Sonnets</a></h3>
 <div class="product_price">
```

```
 <p class="price_color">Â£20.66</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="set-me-free_988/index.html"><img alt="Set Me Free" class="thumbnail" src
="../media/cache/5b/88/5b88c52633f53cacf162c15f4f823153.jpg"/></a>
 </div>
 <p class="star-rating Five">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="set-me-free_988/index.html" title="Set Me Free">Set Me Free</a></h3>
 <div class="product_price">
 <p class="price_color">Â£17.46</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html"><img
alt="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)" class="thumbnail" sr
c="../media/cache/94/b1/94b1b8b244bce9677c2f29ccc890d4d2.jpg"/></a>
 </div>
 <p class="star-rating Five">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html"
title="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)">Scott Pilgrim's Pr
ecious Little ...</a></h3>
```

```
 <div class="product_price">
 <p class="price_color">Â£52.29</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="rip-it-up-and-start-again_986/index.html"><img alt="Rip it Up and Start
Again" class="thumbnail" src="../media/cache/81/c4/81c4a973364e17d01f217e1188253d5
e.jpg"/></a>
 </div>
 <p class="star-rating Five">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="rip-it-up-and-start-again_986/index.html" title="Rip it Up and Start
Again">Rip it Up and ...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£35.02</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="our-band-could-be-your-life-scenes-from-the-american-indie-underground-1
981-1991_985/index.html"><img alt="Our Band Could Be Your Life: Scenes from the Am
erican Indie Underground, 1981-1991" class="thumbnail" src="../media/cache/54/60/5
4607fe8945897cdcced0044103b10b6.jpg"/></a>
 </div>
 <p class="star-rating Three">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
```

```
 </p>
 <h3><a href="our-band-could-be-your-life-scenes-from-the-american-indie-undergrou
nd-1981-1991_985/index.html" title="Our Band Could Be Your Life: Scenes from the A
merican Indie Underground, 1981-1991">Our Band Could Be ...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£57.25</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="olio_984/index.html"><img alt="Olio" class="thumbnail" src="../media/cac
he/55/33/553310a7162dfbc2c6d19a84da0df9e1.jpg"/></a>
 </div>
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="olio_984/index.html" title="Olio">Olio</a></h3>
 <div class="product_price">
 <p class="price_color">Â£23.88</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html"><im
g alt="Mesaerion: The Best Science Fiction Stories 1800-1849" class="thumbnail" sr
c="../media/cache/09/a3/09a3aef48557576e1a85ba7efea8ecb7.jpg"/></a>
 </div>
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
```

```
 <i class="icon-star"></i>
 </p>
 <h3><a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html"
title="Mesaerion: The Best Science Fiction Stories 1800-1849">Mesaerion: The Best
Science ...</a></h3>
 <div class="product_price">
 <p class="price_color">Â£37.59</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="libertarianism-for-beginners_982/index.html"><img alt="Libertarianism fo
r Beginners" class="thumbnail" src="../media/cache/0b/bc/0bbcd0a6f4bcd81ccb1049a52
736406e.jpg"/></a>
 </div>
 <p class="star-rating Two">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="libertarianism-for-beginners_982/index.html" title="Libertarianism f
or Beginners">Libertarianism for Beginners</a></h3>
 <div class="product_price">
 <p class="price_color">Â£51.33</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>,
 <article class="product_pod">
 <div class="image_container">
 <a href="its-only-the-himalayas_981/index.html"><img alt="It's Only the Himalaya
s" class="thumbnail" src="../media/cache/27/a5/27a53d0bb95bdd88288eaf66c9230d7e.jp
g"/></a>
 </div>
 <p class="star-rating Two">
 <i class="icon-star"></i>
```

```
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="its-only-the-himalayas_981/index.html" title="It's Only the Himalaya
s">It's Only the Himalayas</a></h3>
 <div class="product_price">
 <p class="price_color">Â£45.17</p>
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>
 <form>
 <button class="btn btn-primary btn-block" data-loading-text="Adding..." type="sub
mit">Add to basket</button>
 </form>
 </div>
 </article>]
```

In [64]:
```
soup.select('article h3 a')
```

Out[64]:
```
[<a href="a-light-in-the-attic_1000/index.html" title="A Light in the Attic">A Lig
ht in the ...</a>,
 <a href="tipping-the-velvet_999/index.html" title="Tipping the Velvet">Tipping th
e Velvet</a>,
 <a href="soumission_998/index.html" title="Soumission">Soumission</a>,
 <a href="sharp-objects_997/index.html" title="Sharp Objects">Sharp Objects</a>,
 <a href="sapiens-a-brief-history-of-humankind_996/index.html" title="Sapiens: A B
rief History of Humankind">Sapiens: A Brief History ...</a>,
 <a href="the-requiem-red_995/index.html" title="The Requiem Red">The Requiem Red
</a>,
 <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.html" title
="The Dirty Little Secrets of Getting Your Dream Job">The Dirty Little Secrets ...
</a>,
 <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-vict
oria-woodhull_993/index.html" title="The Coming Woman: A Novel Based on the Life o
f the Infamous Feminist, Victoria Woodhull">The Coming Woman: A ...</a>,
 <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the
-1936-berlin-olympics_992/index.html" title="The Boys in the Boat: Nine Americans
and Their Epic Quest for Gold at the 1936 Berlin Olympics">The Boys in the ...</a
>,
 <a href="the-black-maria_991/index.html" title="The Black Maria">The Black Maria
</a>,
 <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html" title="Starvi
ng Hearts (Triangular Trade Trilogy, #1)">Starving Hearts (Triangular Trade ...</a
>,
 <a href="shakespeares-sonnets_989/index.html" title="Shakespeare's Sonnets">Shake
speare's Sonnets</a>,
 <a href="set-me-free_988/index.html" title="Set Me Free">Set Me Free</a>,
 <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html" titl
e="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)">Scott Pilgrim's Precio
us Little ...</a>,
 <a href="rip-it-up-and-start-again_986/index.html" title="Rip it Up and Start Aga
in">Rip it Up and ...</a>,
 <a href="our-band-could-be-your-life-scenes-from-the-american-indie-underground-1
981-1991_985/index.html" title="Our Band Could Be Your Life: Scenes from the Ameri
can Indie Underground, 1981-1991">Our Band Could Be ...</a>,
 <a href="olio_984/index.html" title="Olio">Olio</a>,
 <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html" tit
le="Mesaerion: The Best Science Fiction Stories 1800-1849">Mesaerion: The Best Sci
ence ...</a>,
 <a href="libertarianism-for-beginners_982/index.html" title="Libertarianism for B
eginners">Libertarianism for Beginners</a>,
 <a href="its-only-the-himalayas_981/index.html" title="It's Only the Himalayas">I
t's Only the Himalayas</a>]
```

In [65]:
```
len(soup.select('article h3 a'))
```

Out[65]:
```
20
```

In [66]:
```
soup.select('article h3 a')
```

Out[66]:
```
[<a href="a-light-in-the-attic_1000/index.html" title="A Light in the Attic">A Lig
ht in the ...</a>,
 <a href="tipping-the-velvet_999/index.html" title="Tipping the Velvet">Tipping th
e Velvet</a>,
 <a href="soumission_998/index.html" title="Soumission">Soumission</a>,
 <a href="sharp-objects_997/index.html" title="Sharp Objects">Sharp Objects</a>,
 <a href="sapiens-a-brief-history-of-humankind_996/index.html" title="Sapiens: A B
rief History of Humankind">Sapiens: A Brief History ...</a>,
 <a href="the-requiem-red_995/index.html" title="The Requiem Red">The Requiem Red
</a>,
 <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.html" title
="The Dirty Little Secrets of Getting Your Dream Job">The Dirty Little Secrets ...
</a>,
 <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-vict
oria-woodhull_993/index.html" title="The Coming Woman: A Novel Based on the Life o
f the Infamous Feminist, Victoria Woodhull">The Coming Woman: A ...</a>,
 <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the
-1936-berlin-olympics_992/index.html" title="The Boys in the Boat: Nine Americans
and Their Epic Quest for Gold at the 1936 Berlin Olympics">The Boys in the ...</a
>,
 <a href="the-black-maria_991/index.html" title="The Black Maria">The Black Maria
</a>,
 <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html" title="Starvi
ng Hearts (Triangular Trade Trilogy, #1)">Starving Hearts (Triangular Trade ...</a
>,
 <a href="shakespeares-sonnets_989/index.html" title="Shakespeare's Sonnets">Shake
speare's Sonnets</a>,
 <a href="set-me-free_988/index.html" title="Set Me Free">Set Me Free</a>,
 <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html" titl
e="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)">Scott Pilgrim's Precio
us Little ...</a>,
 <a href="rip-it-up-and-start-again_986/index.html" title="Rip it Up and Start Aga
in">Rip it Up and ...</a>,
 <a href="our-band-could-be-your-life-scenes-from-the-american-indie-underground-1
981-1991_985/index.html" title="Our Band Could Be Your Life: Scenes from the Ameri
can Indie Underground, 1981-1991">Our Band Could Be ...</a>,
 <a href="olio_984/index.html" title="Olio">Olio</a>,
 <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html" tit
le="Mesaerion: The Best Science Fiction Stories 1800-1849">Mesaerion: The Best Sci
ence ...</a>,
 <a href="libertarianism-for-beginners_982/index.html" title="Libertarianism for B
eginners">Libertarianism for Beginners</a>,
 <a href="its-only-the-himalayas_981/index.html" title="It's Only the Himalayas">I
t's Only the Himalayas</a>]
```

In [67]:
```
soup.select('article h3 a')[0]
```

Out[67]:
```
<a href="a-light-in-the-attic_1000/index.html" title="A Light in the Attic">A Ligh
t in the ...</a>
```

In [68]:
```
soup.select('article h3 a')[0]['title']
```

Out[68]:
```
'A Light in the Attic'
```

In [69]:
```
soup.select('article p')
```

Out[69]:
```
[<p class="star-rating Three">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>,
 <p class="price_color">Â£51.77</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>,
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>,
 <p class="price_color">Â£53.74</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>,
 <p class="star-rating One">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>,
 <p class="price_color">Â£50.10</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>,
 <p class="star-rating Four">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>,
 <p class="price_color">Â£47.82</p>,
 <p class="instock availability">
 <i class="icon-ok"></i>

        In stock

 </p>,
```

```
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£54.23</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£22.65</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£33.34</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£17.93</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
```

```
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£22.60</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£52.15</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£13.99</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£20.66</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
```

```
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£17.46</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£52.29</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£35.02</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£57.25</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
```

```
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£23.88</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£37.59</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£51.33</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>,
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>,
<p class="price_color">Â£45.17</p>,
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>]
```

In [70]:
```
soup.select('article p')[0]
```

Out[70]:
```
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
```

In [71]:
```
soup.select('article p')[0]['class']
```

Out[71]:
```
['star-rating', 'Three']
```

In [72]:
```
soup.select('article p')[1]
```

Out[72]:
```
<p class="price_color">Â£51.77</p>
```

In [73]:
```
soup.select('article p')[1]
```

Out[73]:
```
<p class="price_color">Â£51.77</p>
```

In [74]:
```
soup.select('article p')[1].string
```

Out[74]:
```
'Â£51.77'
```

In [75]:
```
soup.select('article p')[1].string.replace('Â£','')
```

Out[75]:
```
'51.77'
```

In [76]:
```
title_list = []
rating_list = []
price_list = []

for pg_num in range(1, 51):
    if req.status_code == 200:
        soup_pg = bs4.BeautifulSoup(req.text, 'html.parser')

        # Extract book containers
        product_pods = soup_pg.select('article.product_pod')

        for pod in product_pods:
            # Extract title
            title_element = pod.select_one('h3 a')
            if title_element:
                title_list.append(title_element['title'])

            # Extract rating
            rating_element = pod.select_one('p[class^="star-rating"]')
            if rating_element:
                rating_list.append(rating_element['class'][1])

            # Extract price
            price_element = pod.select_one('.price_color')
```

```
            if price_element:
                price_list.append(price_element.text.replace('£', '').strip())
    else:
        print(f"Failed to retrieve page {pg_num}: {req.status_code}")
```

In [77]:
```
pip install requests beautifulsoup4
```

```
Requirement already satisfied: requests in c:\users\91937\anaconda3\lib\site-package
s (2.32.2)
Requirement already satisfied: beautifulsoup4 in c:\users\91937\anaconda3\lib\site-p
ackages (4.12.3)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\91937\anaconda3
\lib\site-packages (from requests) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\91937\anaconda3\lib\site-pac
kages (from requests) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\91937\anaconda3\lib\si
te-packages (from requests) (2.2.2)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\91937\anaconda3\lib\si
te-packages (from requests) (2024.6.2)
Requirement already satisfied: soupsieve>1.2 in c:\users\91937\anaconda3\lib\site-pa
ckages (from beautifulsoup4) (2.5)
Note: you may need to restart the kernel to use updated packages.
```

In [78]:
```
# price_list
dic = dict(zip(title_list,zip(rating_list, price_list)))
df = pd.DataFrame.from_dict(dic, orient='index')
df
```

Out[78]:

	0	1
A Light in the Attic	Three	Â51.77
Tipping the Velvet	One	Â53.74
Soumission	One	Â50.10
Sharp Objects	Four	Â47.82
Sapiens: A Brief History of Humankind	Five	Â54.23
The Requiem Red	One	Â22.65
The Dirty Little Secrets of Getting Your Dream Job	Four	Â33.34
The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria Woodhull	Three	Â17.93
The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics	Four	Â22.60
The Black Maria	One	Â52.15
Starving Hearts (Triangular Trade Trilogy, #1)	Two	Â13.99
Shakespeare's Sonnets	Four	Â20.66
Set Me Free	Five	Â17.46
Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)	Five	Â52.29
Rip it Up and Start Again	Five	Â35.02
Our Band Could Be Your Life: Scenes from the American Indie Underground, 1981-1991	Three	Â57.25
Olio	One	Â23.88
Mesaerion: The Best Science Fiction Stories 1800-1849	One	Â37.59
Libertarianism for Beginners	Two	Â51.33
It's Only the Himalayas	Two	Â45.17

In [79]:
```
# we already know the url
soup.select('article h3 a')
```

Out[79]:
```
[<a href="a-light-in-the-attic_1000/index.html" title="A Light in the Attic">A Lig
ht in the ...</a>,
 <a href="tipping-the-velvet_999/index.html" title="Tipping the Velvet">Tipping th
e Velvet</a>,
 <a href="soumission_998/index.html" title="Soumission">Soumission</a>,
 <a href="sharp-objects_997/index.html" title="Sharp Objects">Sharp Objects</a>,
 <a href="sapiens-a-brief-history-of-humankind_996/index.html" title="Sapiens: A B
rief History of Humankind">Sapiens: A Brief History ...</a>,
 <a href="the-requiem-red_995/index.html" title="The Requiem Red">The Requiem Red
</a>,
 <a href="the-dirty-little-secrets-of-getting-your-dream-job_994/index.html" title
="The Dirty Little Secrets of Getting Your Dream Job">The Dirty Little Secrets ...
</a>,
 <a href="the-coming-woman-a-novel-based-on-the-life-of-the-infamous-feminist-vict
oria-woodhull_993/index.html" title="The Coming Woman: A Novel Based on the Life o
f the Infamous Feminist, Victoria Woodhull">The Coming Woman: A ...</a>,
 <a href="the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-gold-at-the
-1936-berlin-olympics_992/index.html" title="The Boys in the Boat: Nine Americans
and Their Epic Quest for Gold at the 1936 Berlin Olympics">The Boys in the ...</a
>,
 <a href="the-black-maria_991/index.html" title="The Black Maria">The Black Maria
</a>,
 <a href="starving-hearts-triangular-trade-trilogy-1_990/index.html" title="Starvi
ng Hearts (Triangular Trade Trilogy, #1)">Starving Hearts (Triangular Trade ...</a
>,
 <a href="shakespeares-sonnets_989/index.html" title="Shakespeare's Sonnets">Shake
speare's Sonnets</a>,
 <a href="set-me-free_988/index.html" title="Set Me Free">Set Me Free</a>,
 <a href="scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html" titl
e="Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)">Scott Pilgrim's Precio
us Little ...</a>,
 <a href="rip-it-up-and-start-again_986/index.html" title="Rip it Up and Start Aga
in">Rip it Up and ...</a>,
 <a href="our-band-could-be-your-life-scenes-from-the-american-indie-underground-1
981-1991_985/index.html" title="Our Band Could Be Your Life: Scenes from the Ameri
can Indie Underground, 1981-1991">Our Band Could Be ...</a>,
 <a href="olio_984/index.html" title="Olio">Olio</a>,
 <a href="mesaerion-the-best-science-fiction-stories-1800-1849_983/index.html" tit
le="Mesaerion: The Best Science Fiction Stories 1800-1849">Mesaerion: The Best Sci
ence ...</a>,
 <a href="libertarianism-for-beginners_982/index.html" title="Libertarianism for B
eginners">Libertarianism for Beginners</a>,
 <a href="its-only-the-himalayas_981/index.html" title="It's Only the Himalayas">I
t's Only the Himalayas</a>]
```

In [80]:
```
nreq= requests.get('https://books.toscrape.com/catalogue/in-her-wake_980/index.html
```

In [81]:
```
nsoup = bs4.BeautifulSoup(nreq.text)
nsoup
```

Out[81]:
```
<!DOCTYPE html>
<!--[if lt IE 7]>      <html lang="en-us" class="no-js lt-ie9 lt-ie8 lt-ie7"> <![e
ndif]--><!--[if IE 7]>         <html lang="en-us" class="no-js lt-ie9 lt-ie8"> <!
[endif]--><!--[if IE 8]>         <html lang="en-us" class="no-js lt-ie9"> <![endi
f]--><!--[if gt IE 8]><!--><html class="no-js" lang="en-us"> <!--<![endif]-->
<head>
<title>
    In Her Wake | Books to Scrape - Sandbox
</title>
<meta content="text/html; charset=utf-8" http-equiv="content-type"/>
<meta content="24th Jun 2016 09:30" name="created"/>
<meta content="
    A perfect life â¦ until she discovered it wasnâ€™t her own.A tragic family ev
ent reveals devastating news that rips apart Bellaâ€™s comfortable existence. Emba
rking on a personal journey to uncover the truth, she faces a series of traumatic
discoveries that take her to the ruggedly beautiful Cornish coast, where hidden tr
uths, past betrayals and a 25-year-old mystery threaten n A perfect life â¦ until
she discovered it wasnâ€™t her own.A tragic family event reveals devastating news
that rips apart Bellaâ€™s comfortable existence. Embarking on a personal journey t
o uncover the truth, she faces a series of traumatic discoveries that take her to
the ruggedly beautiful Cornish coast, where hidden truths, past betrayals and a 25
-year-old mystery threaten not just her identity, but also her life.Chilling, comp
lex and profoundly moving, In Her Wake is a gripping psychological thriller that q
uestions the nature of family â€” and reminds us that sometimes the most shocking
crimes are committed closest to home. ...more
" name="description"/>
<meta content="width=device-width" name="viewport"/>
<meta content="NOARCHIVE,NOCACHE" name="robots"/>
<!-- Le HTML5 shim, for IE6-8 support of HTML elements -->
<!--[if lt IE 9]>
        <script src="//html5shim.googlecode.com/svn/trunk/html5.js"></script>
        <![endif]-->
<link href="../../static/oscar/favicon.ico" rel="shortcut icon"/>
<link href="../../static/oscar/css/styles.css" rel="stylesheet" type="text/css"/>
<link href="../../static/oscar/js/bootstrap-datetimepicker/bootstrap-datetimepicke
r.css" rel="stylesheet"/>
<link href="../../static/oscar/css/datetimepicker.css" rel="stylesheet" type="tex
t/css"/>
</head>
<body class="default" id="default">
<header class="header container-fluid">
<div class="page_inner">
<div class="row">
<div class="col-sm-8 h1"><a href="../../index.html">Books to Scrape</a><small> We
love being scraped!</small>
</div>
</div>
</div>
</header>
<div class="container-fluid page">
<div class="page_inner">
<ul class="breadcrumb">
<li>
<a href="../../index.html">Home</a>
</li>
<li>
```

```
<a href="../category/books_1/index.html">Books</a>
</li>
<li>
<a href="../category/books/thriller_37/index.html">Thriller</a>
</li>
<li class="active">In Her Wake</li>
</ul>
<div id="messages">
</div>
<div class="content">
<div id="promotions">
</div>
<div id="content_inner">
<article class="product_page"><!-- Start of product page -->
<div class="row">
<div class="col-sm-6">
<div class="carousel" id="product_gallery">
<div class="thumbnail">
<div class="carousel-inner">
<div class="item active">
<img alt="In Her Wake" src="../../media/cache/27/92/2792ef951651ff1eae40a410cac41e
0f.jpg"/>
</div>
</div>
</div>
</div>
</div>
<div class="col-sm-6 product_main">
<h1>In Her Wake</h1>
<p class="price_color">Â£12.84</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock (19 available)

</p>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<!-- <small><a href="/catalogue/in-her-wake_980/reviews/">


                    0 customer reviews

        </a></small>
         -->


<!--
    <a id="write_review" href="/catalogue/in-her-wake_980/reviews/add/#addreview"
class="btn btn-success btn-sm">
        Write a review
    </a>
```

```
 --></p>
<hr/>
<div class="alert alert-warning" role="alert"><strong>Warning!</strong> This is a
demo website for web scraping purposes. Prices and ratings here were randomly assi
gned and have no real meaning.</div>
</div><!-- /col-sm-6 -->
</div><!-- /row -->
<div class="sub-header" id="product_description">
<h2>Product Description</h2>
</div>
<p>A perfect life â¦ until she discovered it wasnâ€™t her own.A tragic family eve
nt reveals devastating news that rips apart Bellaâ€™s comfortable existence. Embar
king on a personal journey to uncover the truth, she faces a series of traumatic d
iscoveries that take her to the ruggedly beautiful Cornish coast, where hidden tru
ths, past betrayals and a 25-year-old mystery threaten n A perfect life â¦ until
she discovered it wasnâ€™t her own.A tragic family event reveals devastating news
that rips apart Bellaâ€™s comfortable existence. Embarking on a personal journey t
o uncover the truth, she faces a series of traumatic discoveries that take her to
the ruggedly beautiful Cornish coast, where hidden truths, past betrayals and a 25
-year-old mystery threaten not just her identity, but also her life.Chilling, comp
lex and profoundly moving, In Her Wake is a gripping psychological thriller that q
uestions the nature of family â€ and reminds us that sometimes the most shocking
crimes are committed closest to home. ...more</p>
<div class="sub-header">
<h2>Product Information</h2>
</div>
<table class="table table-striped">
<tr>
<th>UPC</th><td>23356462d1320d61</td>
</tr>
<tr>
<th>Product Type</th><td>Books</td>
</tr>
<tr>
<th>Price (excl. tax)</th><td>Â£12.84</td>
</tr>
<tr>
<th>Price (incl. tax)</th><td>Â£12.84</td>
</tr>
<tr>
<th>Tax</th><td>Â£0.00</td>
</tr>
<tr>
<th>Availability</th>
<td>In stock (19 available)</td>
</tr>
<tr>
<th>Number of reviews</th>
<td>0</td>
</tr>
</table>
<section>
<div class="sub-header" id="reviews">
</div>
</section>
```

```
<div class="sub-header">
<h2>Products you recently viewed</h2>
</div>
<ul class="row">
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="../the-rise-of-theodore-roosevelt-theodore-roosevelt-1_276/index.html"><i
mg alt="The Rise of Theodore Roosevelt (Theodore Roosevelt #1)" class="thumbnail"
src="../../media/cache/ff/d4/ffd45d95f314555e20c923d3522adea7.jpg"/></a>
</div>
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="../the-rise-of-theodore-roosevelt-theodore-roosevelt-1_276/index.htm
l" title="The Rise of Theodore Roosevelt (Theodore Roosevelt #1)">The Rise of Theo
dore ...</a></h3>
<div class="product_price">
<p class="price_color">Â£42.57</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding..." type="subm
it">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="../benjamin-franklin-an-american-life_460/index.html"><img alt="Benjamin
Franklin: An American Life" class="thumbnail" src="../../media/cache/25/f8/25f869f
a75340fca0fc2a68e8a0412a1.jpg"/></a>
</div>
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="../benjamin-franklin-an-american-life_460/index.html" title="Benjamin
Franklin: An American Life">Benjamin Franklin: An American ...</a></h3>
<div class="product_price">
<p class="price_color">Â£48.19</p>
<p class="instock availability">
<i class="icon-ok"></i>
```

```
        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding..." type="subm
it">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="../the-faith-of-christopher-hitchens-the-restless-soul-of-the-worlds-most
-notorious-atheist_495/index.html"><img alt="The Faith of Christopher Hitchens: Th
e Restless Soul of the World's Most Notorious Atheist" class="thumbnail" src
="../../media/cache/cc/a4/cca4e6a4cd5c207e7ce7d992ff464c3b.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="../the-faith-of-christopher-hitchens-the-restless-soul-of-the-worlds-
most-notorious-atheist_495/index.html" title="The Faith of Christopher Hitchens: T
he Restless Soul of the World's Most Notorious Atheist">The Faith of Christopher
...</a></h3>
<div class="product_price">
<p class="price_color">Â£39.55</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding..." type="subm
it">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="../setting-the-world-on-fire-the-brief-astonishing-life-of-st-catherine-o
f-siena_603/index.html"><img alt="Setting the World on Fire: The Brief, Astonishin
g Life of St. Catherine of Siena" class="thumbnail" src="../../media/cache/8b/c4/8
bc43a6b42d0283ab4bf611f1b497126.jpg"/></a>
</div>
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
```

```
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="../setting-the-world-on-fire-the-brief-astonishing-life-of-st-catheri
ne-of-siena_603/index.html" title="Setting the World on Fire: The Brief, Astonishi
ng Life of St. Catherine of Siena">Setting the World on ...</a></h3>
<div class="product_price">
<p class="price_color">Â£21.15</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding..." type="subm
it">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="../louisa-the-extraordinary-life-of-mrs-adams_818/index.html"><img alt="L
ouisa: The Extraordinary Life of Mrs. Adams" class="thumbnail" src="../../media/ca
che/6f/d9/6fd92e5143cbd5bb8bcf034e5f007dde.jpg"/></a>
</div>
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="../louisa-the-extraordinary-life-of-mrs-adams_818/index.html" title
="Louisa: The Extraordinary Life of Mrs. Adams">Louisa: The Extraordinary Life ...
</a></h3>
<div class="product_price">
<p class="price_color">Â£16.85</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding..." type="subm
it">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
```

```
<div class="image_container">
<a href="../rework_212/index.html"><img alt="Rework" class="thumbnail" src="../../
media/cache/54/10/5410a58193e2373c04b3021ade78a82b.jpg"/></a>
</div>
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="../rework_212/index.html" title="Rework">Rework</a></h3>
<div class="product_price">
<p class="price_color">Â£44.88</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding..." type="subm
it">Add to basket</button>
</form>
</div>
</article>
</li>
</ul>
</article><!-- End of product page -->
</div>
</div>
</div>
</div>
<footer class="footer container-fluid">
</footer>
<!-- jQuery -->
<script src="http://ajax.googleapis.com/ajax/libs/jquery/1.9.1/jquery.min.js"></sc
ript>
<script>window.jQuery || document.write('<script src="../../static/oscar/js/jquer
y/jquery-1.9.1.min.js"><\/script>')</script>
<!-- Twitter Bootstrap -->
<script src="../../static/oscar/js/bootstrap3/bootstrap.min.js" type="text/javascr
ipt"></script>
<!-- Oscar -->
<script charset="utf-8" src="../../static/oscar/js/oscar/ui.js" type="text/javascr
ipt"></script>
<script charset="utf-8" src="../../static/oscar/js/bootstrap-datetimepicker/bootst
rap-datetimepicker.js" type="text/javascript"></script>
<script charset="utf-8" src="../../static/oscar/js/bootstrap-datetimepicker/locale
s/bootstrap-datetimepicker.all.js" type="text/javascript"></script>
<script type="text/javascript">
            $(function() {


    oscar.init();
```

```
            });
        </script>
<!-- Version: N/A -->
</body>
</html>
```

In [82]:
```
nsoup.select('tr')
```

Out[82]:
```
[<tr>
 <th>UPC</th><td>23356462d1320d61</td>
 </tr>,
 <tr>
 <th>Product Type</th><td>Books</td>
 </tr>,
 <tr>
 <th>Price (excl. tax)</th><td>Â£12.84</td>
 </tr>,
 <tr>
 <th>Price (incl. tax)</th><td>Â£12.84</td>
 </tr>,
 <tr>
 <th>Tax</th><td>Â£0.00</td>
 </tr>,
 <tr>
 <th>Availability</th>
 <td>In stock (19 available)</td>
 </tr>,
 <tr>
 <th>Number of reviews</th>
 <td>0</td>
 </tr>]
```

In [83]:
```
nsoup.select('tr th')
```

Out[83]:
```
[<th>UPC</th>,
 <th>Product Type</th>,
 <th>Price (excl. tax)</th>,
 <th>Price (incl. tax)</th>,
 <th>Tax</th>,
 <th>Availability</th>,
 <th>Number of reviews</th>]
```

In [84]:
```
nsoup.select('tr td')
```

Out[84]:
```
[<td>23356462d1320d61</td>,
 <td>Books</td>,
 <td>Â£12.84</td>,
 <td>Â£12.84</td>,
 <td>Â£0.00</td>,
 <td>In stock (19 available)</td>,
 <td>0</td>]
```

In [85]:
```
nsoup.select('tr td')[2].text
```

Out[85]:
```
'Â£12.84'
```

In [87]:
```
title_list = []
rating_list = []
price_list = []

for pg_num in range(1, 51):  # Loop through 50 pages
    url = f'https://books.toscrape.com/catalogue/page-{pg_num}.html'
    req = requests.get(url)
    soup = BeautifulSoup(req.text, 'html.parser')

    books = soup.select('article.product_pod')  # Select all book containers

    for book in books:
        # Title
        title = book.h3.a['title']
        title_list.append(title)

        # Rating
        rating = book.p['class'][1]
        rating_list.append(rating)

        # Price
        price = book.select_one('p.price_color').text.replace('Â', '').replace('£',
        price_list.append(price)
```

In [88]:
```
new_df
```

Out[88]:

	UPC	Product Type	Price (excl. tax)	Price (incl. tax)	Tax	Availability	Number of reviews
0	a897fe39b1053632	Books	Â£51.77	Â£51.77	Â£0.00	In stock (22 available)	0
1	90fa61229261140a	Books	Â£53.74	Â£53.74	Â£0.00	In stock (20 available)	0
2	6957f44c3847a760	Books	Â£50.10	Â£50.10	Â£0.00	In stock (20 available)	0
3	e00eb4fd7b871a48	Books	Â£47.82	Â£47.82	Â£0.00	In stock (20 available)	0
4	4165285e1663650f	Books	Â£54.23	Â£54.23	Â£0.00	In stock (20 available)	0
...	...	...	...	...	...	...	...
963	39592d9d72e717c4	Books	Â£47.11	Â£47.11	Â£0.00	In stock (1 available)	0
964	63e20a0f98218a87	Books	Â£58.75	Â£58.75	Â£0.00	In stock (1 available)	0
965	fd78f89878479f56	Books	Â£38.92	Â£38.92	Â£0.00	In stock (1 available)	0
966	35a60467893aa168	Books	Â£39.39	Â£39.39	Â£0.00	In stock (1 available)	0
967	4f19709e47883df5	Books	Â£25.89	Â£25.89	Â£0.00	In stock (1 available)	0

968 rows × 7 columns

In [91]:
```
# Make sure the lengths match to avoid error
min_length = min(len(new_df), len(title_list), len(rating_list))

# Slice the lists to match the number of rows in new_df
new_df["Rating"] = rating_list[:min_length]
new_df["Title"] = title_list[:min_length]

new_df.head()
```

Out[91]:

	UPC	Product Type	Price (excl. tax)	Price (incl. tax)	Tax	Availability	Number of reviews	Rating	
0	a897fe39b1053632	Books	Â£51.77	Â£51.77	Â£0.00	In stock (22 available)	0	Three	A
1	90fa61229261140a	Books	Â£53.74	Â£53.74	Â£0.00	In stock (20 available)	0	One	Tip
2	6957f44c3847a760	Books	Â£50.10	Â£50.10	Â£0.00	In stock (20 available)	0	One	Sou
3	e00eb4fd7b871a48	Books	Â£47.82	Â£47.82	Â£0.00	In stock (20 available)	0	Four	
4	4165285e1663650f	Books	Â£54.23	Â£54.23	Â£0.00	In stock (20 available)	0	Five	Sa H Hu

In [ ]: