```
def avgrainfall(df):
    # to get avg rainfall:
    df.groupby()

    # but
    # df = df.groupby('state')['rainfall'].mean()# -> Here state would then
become the index of your dataframe and since in output we required state we
need to reset the index

    #     state    rainfall
    #     mh        100
    #     tn        200

    # and in output youll get

    #     state    rainfall
    #     0        100
    #     1        200


    # # TO avoid
    df = df.groupby('state')['rainfall'].mean().reset_index()

    #          state    rainfall
    #     0    mh         100
    #     1    tn         200

    # now
    df.rename(columns={0:'state',1:'rainfall'},inplace=True)

    return df[df['rainfall']>150]   ---> #since we only need rows where
rainfall>150
```

```
Questions asked - Viraj

Basic Introduction

1. Project
2. Tell me about youself

Technical - SQL
1.Talk about DDL,DML,DCL,TCL,DQL
2. Difference About truncate delete drop
3. execution sequence of sql commands and (why select should not be the
first in your order)
4. difference between group by and having
```

```
5. what is normalization and denormalization and why they are needed
6. Joins
7. window functions
8. what is cardinality in terms of database

9. following table:
|name|
------
|jack|
|jack|
|jack|
|Ryan|
|Ryan|
|Ryan|

how will you get output like this :

jack
ryan
jack
ryan
jack
ryan



10. following tables:-

tab1       tab2
----       -----
id1        id2
----       -----
1           1
1           2
1           1
null        null

what will be the output for this of : inner join,left join,right join, full
outer join,cross join ?

11. find name,salary of second highest earner. show all possible ways you
can do this and which is most optimum
```

```
technical big data/python/pandas/linux/cloud

1. what is airflow. how to use airflow. can you write a code for it?
2. what are operators in airflow ? types of operators
3. diff between hadoop1.x and 2.x
4. what is yarn ?
```

```
5. explain spark rdd and dataframes. can you write a code in any of the two
and which one will you choose and why ?
6. what is amazon s3
7. what is route 53 (Project related)
8. which AMI have you used in project in ec2?
9. which instance type you have used for project?
10. what is a serverless functionality ?
11. what are facts and dimensions. Give real life examples
12. what is slowly changing dimensions and explain it
13. what is slicing and deslicing ?
14. how to handle missing values in df?
15. in table :

name | age |
-----------
ab     25
cd     null
xz     30
qv     null
ty     null


get the output this way using missing value imputation (no mode,median
etc):

name | age |
-----------
ab     25
cd     25
xz     30
qv     30
ty     30

16. l1 = [1,1,1,2,2,3,4,5,6,7]
show me all possible ways for getting numbers and their counts and identify
which is the most optimal out of all ? use one that is not brute force.


17. linux command to get no of lines
18. difference between cat and vi
19. what is git?
20. write git commands, pull,push,commit
21. cat >, cat >> difference
22. grep command and all its flavors
23. how to enter and exit insertion mode and also the commands in your vi
editor
24. how to safely exit vi editor
25. differences between list, tuple,set,dictionary
26. when would you use which collection ?
27. OLAP vs OLTP
28. give real life example of above question
29. schemas (star snowflake etc)
30. airflow architecture
```

```
31. aws and its components
32. some pandas functions
```