SPRINT 1

Team ID	PNT2022TMID30419
Project Name	Exploratory Analysis Of Rainfall Data In India For Agriculture

DATA READ AND PREPROCESSING

```
df = pd.read_csv(r"C:/Users/NIVEDITHA/Downloads/rainfall.csv")

df = df.fillna(df.mean())

df.info()
```

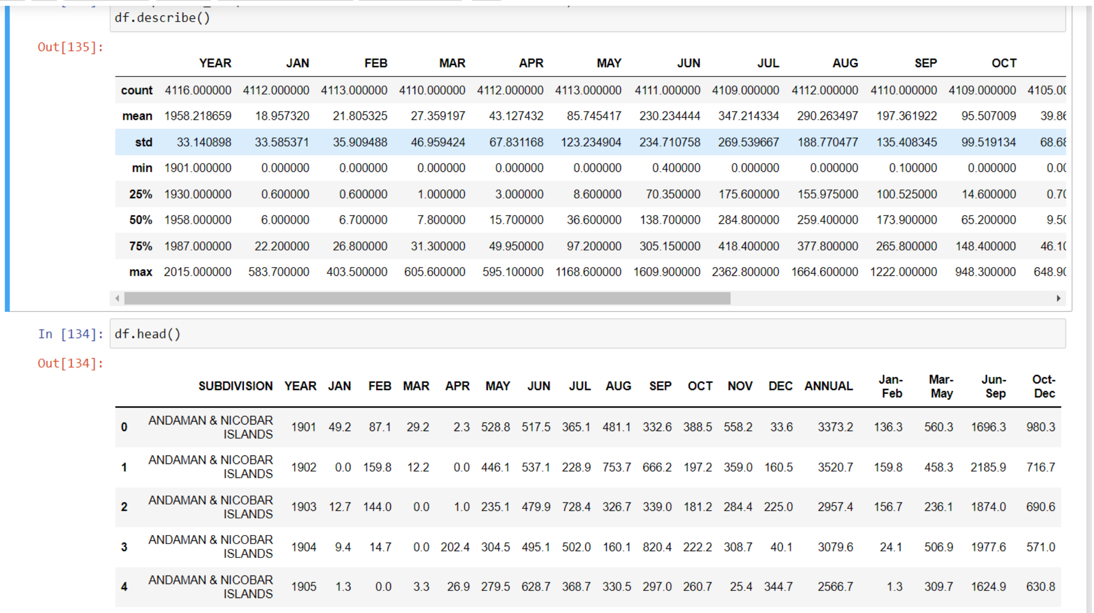
OUTPUT:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4116 entries, 0 to 4115
Data columns (total 19 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   SUBDIVISION  4116 non-null   object 
 1   YEAR         4116 non-null   int64  
 2   JAN          4116 non-null   float64
 3   FEB          4116 non-null   float64
 4   MAR          4116 non-null   float64
 5   APR          4116 non-null   float64
 6   MAY          4116 non-null   float64
 7   JUN          4116 non-null   float64
 8   JUL          4116 non-null   float64
 9   AUG          4116 non-null   float64
 10  SEP          4116 non-null   float64
 11  OCT          4116 non-null   float64
 12  NOV          4116 non-null   float64
 13  DEC          4116 non-null   float64
 14  ANNUAL       4116 non-null   float64
 15  Jan-Feb      4116 non-null   float64
 16  Mar-May      4116 non-null   float64
 17  Jun-Sep      4116 non-null   float64
 18  Oct-Dec      4116 non-null   float64
dtypes: float64(17), int64(1), object(1)
memory usage: 611.1+ KB
```

df.head()

df.describe()

OUTPUT:



```
df.describe()
```

Out[135]:

	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	
count	4116.000000	4112.000000	4113.000000	4110.000000	4112.000000	4113.000000	4111.000000	4109.000000	4112.000000	4110.000000	4109.000000	4105.0
mean	1958.218659	18.957320	21.805325	27.359197	43.127432	85.745417	230.234444	347.214334	290.263497	197.361922	95.507009	39.8
std	33.140898	33.585371	35.909488	46.959424	67.831168	123.234904	234.710758	269.539667	188.770477	135.408345	99.519134	68.6
min	1901.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.400000	0.000000	0.000000	0.100000	0.000000	0.0
25%	1930.000000	0.600000	0.600000	1.000000	3.000000	8.600000	70.350000	175.600000	155.975000	100.525000	14.600000	0.7
50%	1958.000000	6.000000	6.700000	7.800000	15.700000	36.600000	138.700000	284.800000	259.400000	173.900000	65.200000	9.5
75%	1987.000000	22.200000	26.800000	31.300000	49.950000	97.200000	305.150000	418.400000	377.800000	265.800000	148.400000	46.1
max	2015.000000	583.700000	403.500000	605.600000	595.100000	1168.600000	1609.900000	2362.800000	1664.600000	1222.000000	948.300000	648.9

In [134]:
```
df.head()
```

Out[134]:

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
1	ANDAMAN & NICOBAR ISLANDS	1902	0.0	159.8	12.2	0.0	446.1	537.1	228.9	753.7	666.2	197.2	359.0	160.5	3520.7	159.8	458.3	2185.9	716.7
2	ANDAMAN & NICOBAR ISLANDS	1903	12.7	144.0	0.0	1.0	235.1	479.9	728.4	326.7	339.0	181.2	284.4	225.0	2957.4	156.7	236.1	1874.0	690.6
3	ANDAMAN & NICOBAR ISLANDS	1904	9.4	14.7	0.0	202.4	304.5	495.1	502.0	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571.0
4	ANDAMAN & NICOBAR ISLANDS	1905	1.3	0.0	3.3	26.9	279.5	628.7	368.7	330.5	297.0	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8

PREPROCESSING THE DATASET:

TO CHECK FOR NULL VALUES AND FILLING THEM:

df.isnull().sum()

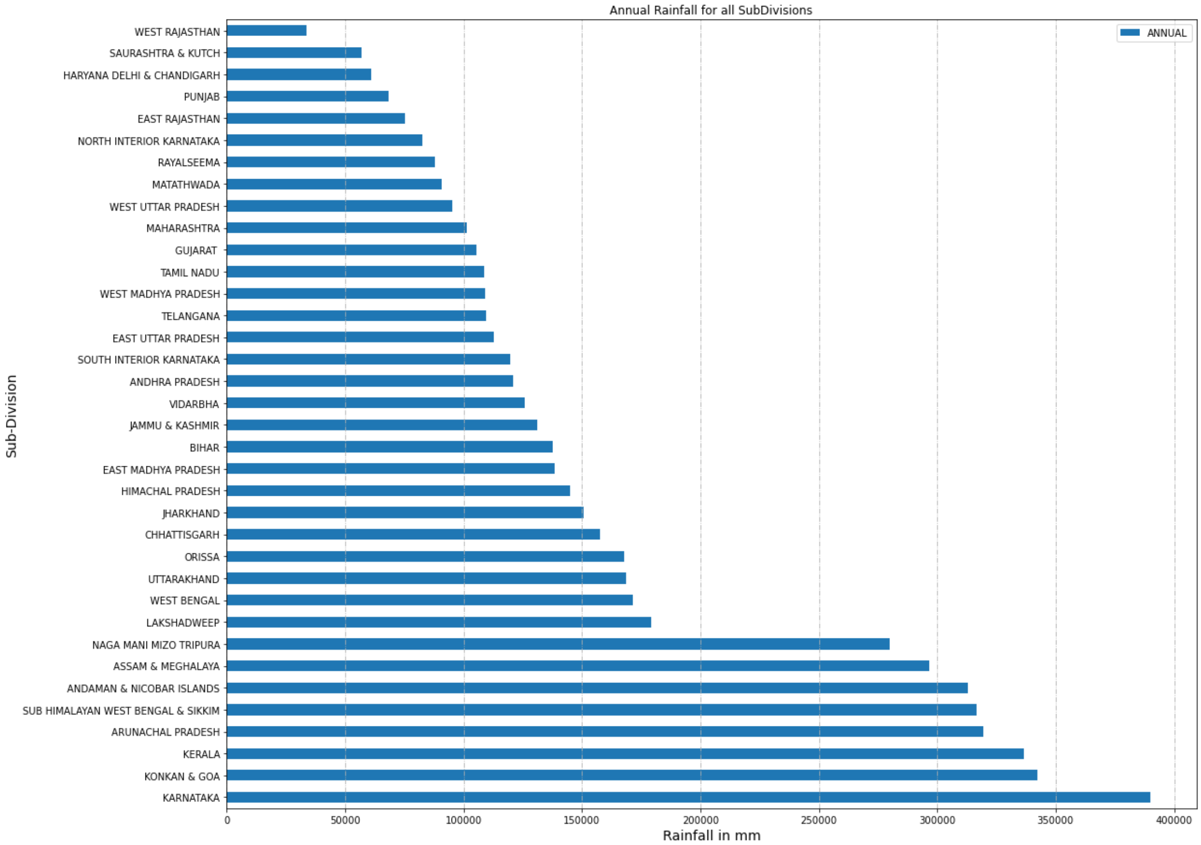
```
SUBDIVISION      0
YEAR             0
JAN              4
FEB              3
MAR              6
APR              4
MAY              3
JUN              5
JUL              7
AUG              4
SEP              6
OCT              7
NOV             11
DEC             10
ANNUAL          26
Jan-Feb          6
Mar-May          9
Jun-Sep         10
Oct-Dec         13
dtype: int64
```

df=df.fillna(df.mean(numeric_only=True).round(1))

DATA VISUALIZATION

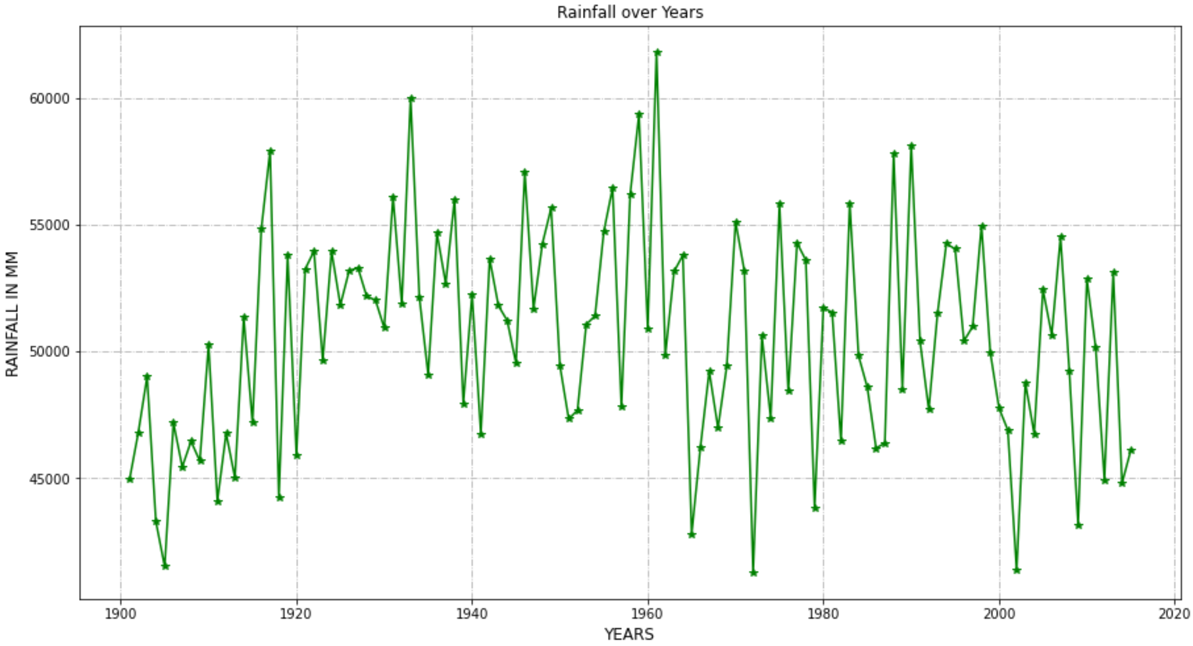
1)

```
df[["SUBDIVISION","ANNUAL"]].groupby("SUBDIVISION").sum().sort_values(by='ANNUAL',ascending=False).plot(kind='barh',stacked=True,figsize=(18,15))

plt.xlabel("Rainfall in mm",size=14)

plt.ylabel("Sub-Division",size=14)

plt.title("Annual Rainfall for all SubDivisions")

plt.grid(axis="x",linestyle="-.")

plt.show()
```



2)

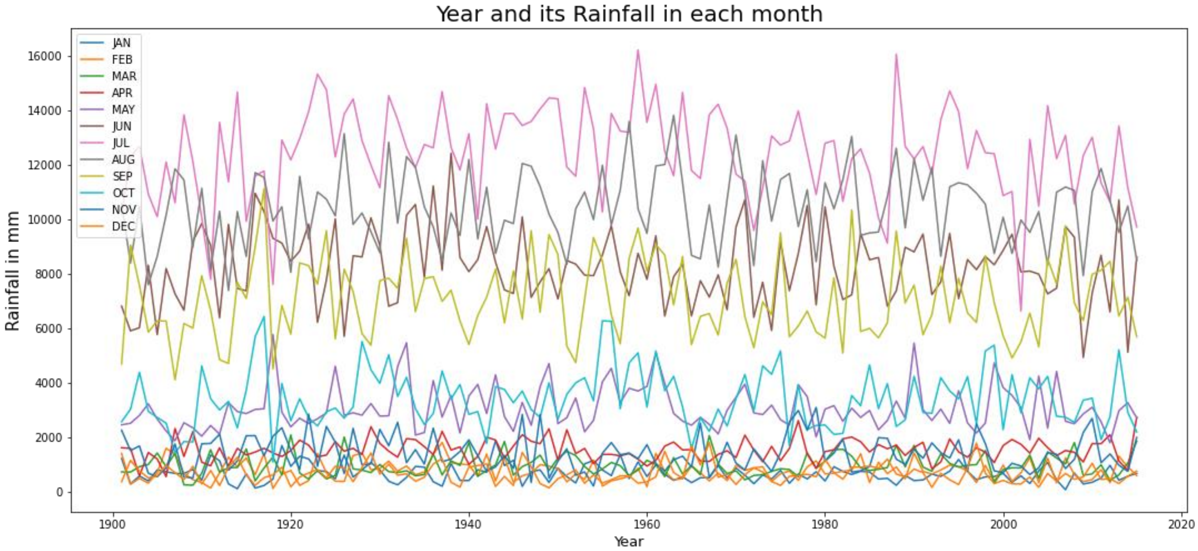
```
plt.figure(figsize=(15,8))

df.groupby("YEAR").sum()['ANNUAL'].plot(kind="line",color="g",marker="*")

plt.xlabel("YEARS",size=12)

plt.ylabel("RAINFALL IN MM",size=12)

plt.grid(axis="both",linestyle="-.")
plt.title("Rainfall over Years")

plt.show()
```



3)

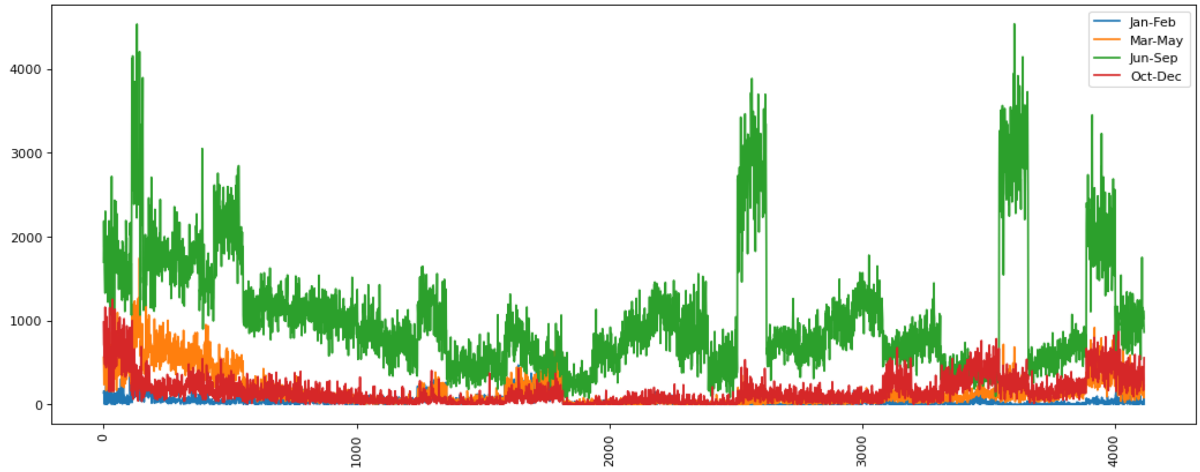
```
df[['YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL','AUG', 'SEP',
    'OCT', 'NOV', 'DEC']].groupby("YEAR").sum().plot(kind="line",figsize=(18,8))

plt.xlabel("Year",size=13)

plt.ylabel("Rainfall in mm",size=15)

plt.title("Year and its Rainfall in each month",size=20)
```

```
lt.show()
```



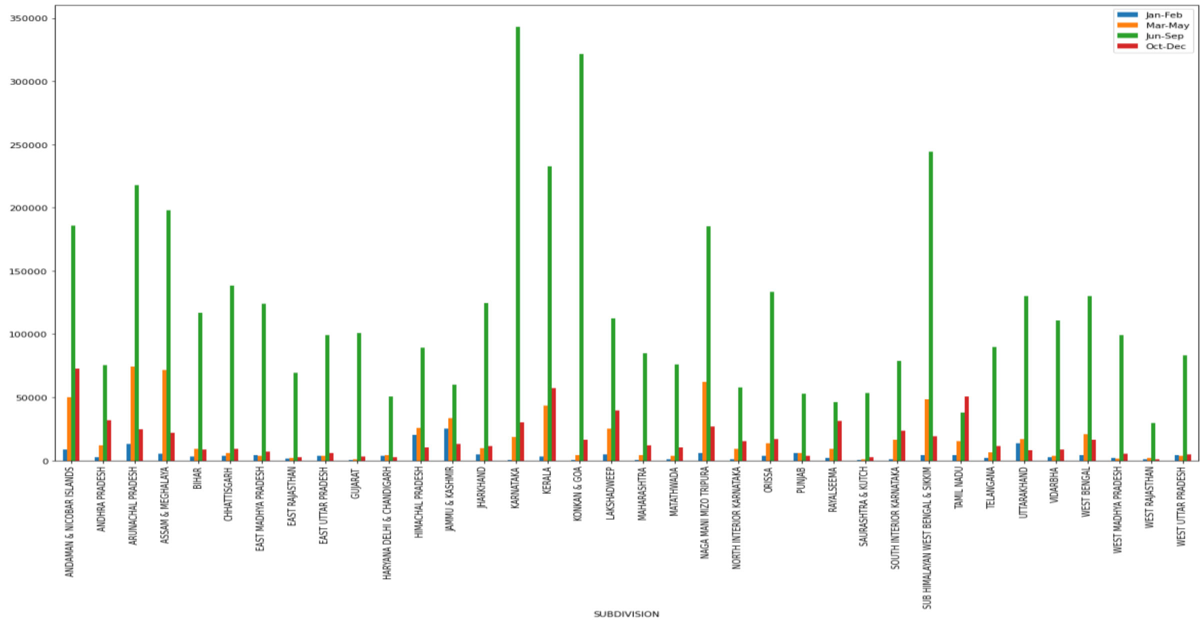
4)

```
plt.figure(figsize=(16,6),dpi=80)
plt.xticks(rotation=90)
plt.plot(df['Jan-Feb'],label='Jan-Feb')
plt.plot(df['Mar-May'],label='Mar-May')
plt.plot(df['Jun-Sep'],label='Jun-Sep')
plt.plot(df['Oct-Dec'],label='Oct-Dec')
plt.legend(loc='best')
```



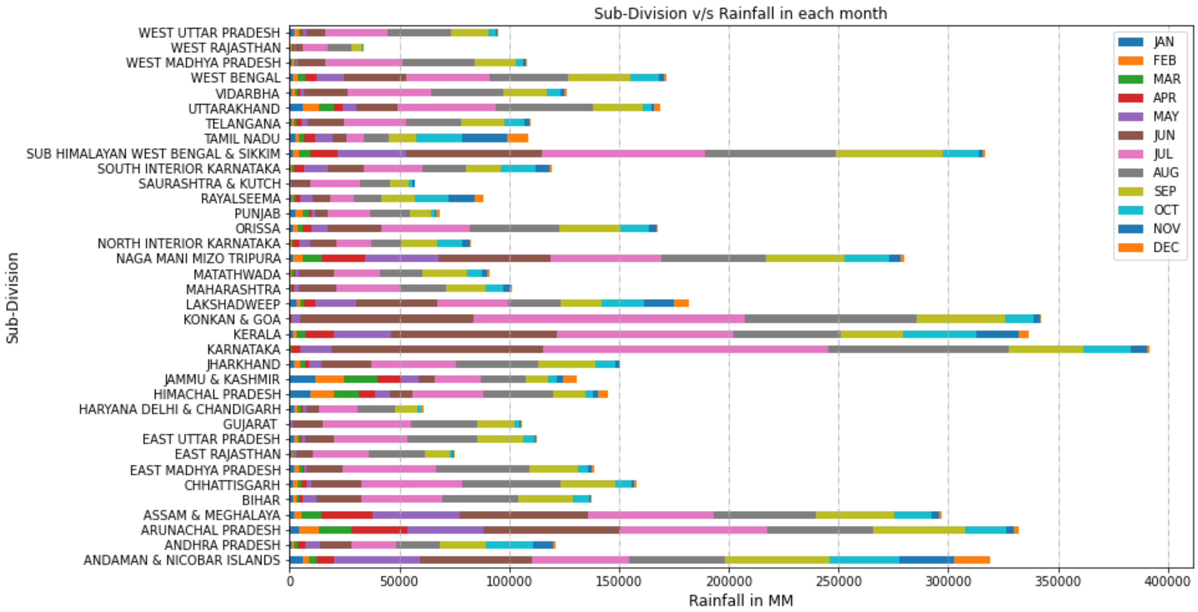
5)

```
ax = df[['SUBDIVISION', 'Jan-Feb', 'Mar-May','Jun-Sep', 'Oct-Dec']].groupby("SUBDIVISION").sum().plot.bar(stacked=False,figsize=(20,12))
```



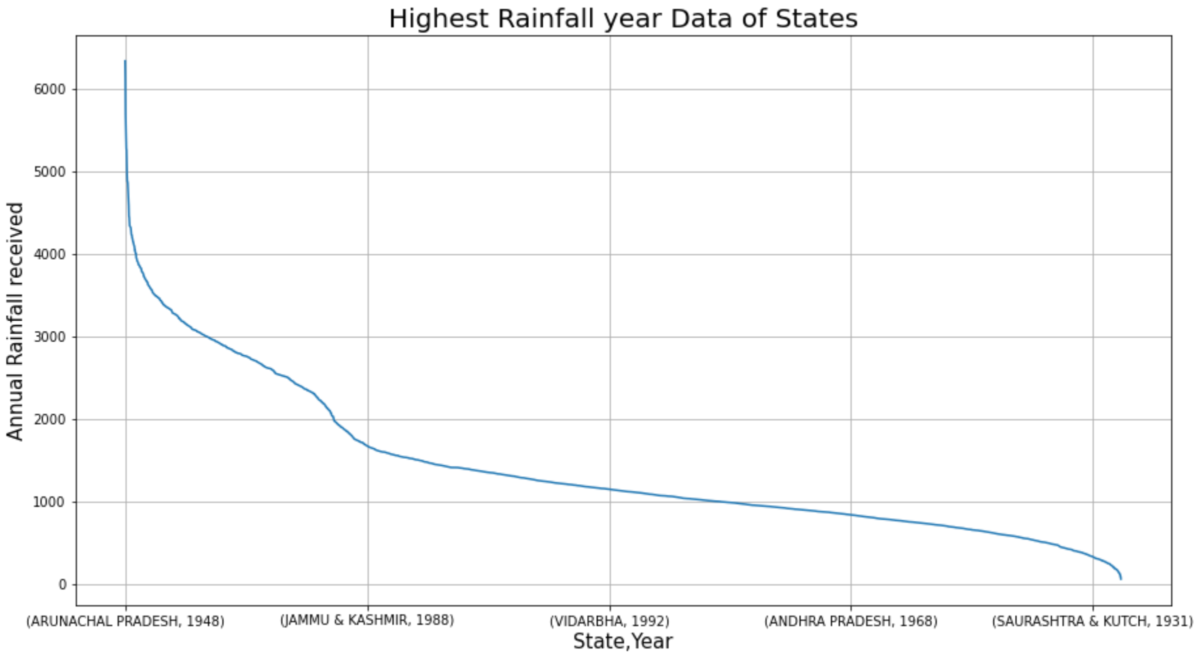
6)

```
df[['SUBDIVISION', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL',
                                        'AUG', 'SEP', 'OCT', 'NOV', 'DEC']].groupby("SUBDIVISION").sum().plot(kind="barh",stacked=True,figsize=(13,8))
plt.title("Sub-Division v/s Rainfall in each month")
plt.xlabel("Rainfall in MM",size=12)
plt.ylabel("Sub-Division",size=12)
plt.grid(axis="x",linestyle="-.")
plt.show()
```



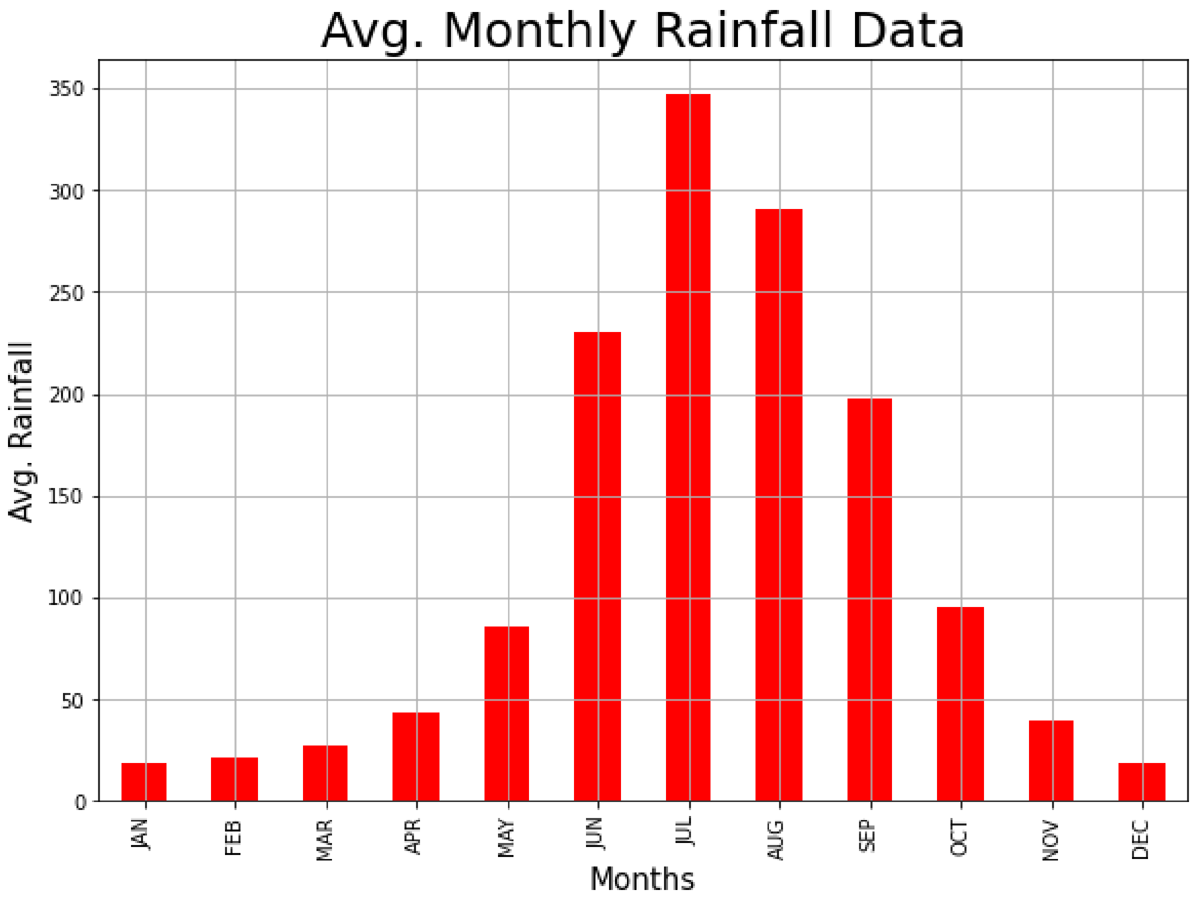
7)

```
#Highest rainfall receiving regions


plt.figure(figsize=(15,8))

df.groupby(['SUBDIVISION','YEAR'])['ANNUAL'].sum().sort_values(ascending=False).plot()

plt.grid()

plt.xlabel("State,Year",fontsize=15)

plt.ylabel("Annual Rainfall received",fontsize=15)

plt.title('Highest Rainfall year Data of States',fontsize=20)
```



8)

```
#Month with highest rainfall

plt.figure(figsize=(10,7))
df[['JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL', 'AUG',
      'SEP', 'OCT', 'NOV', 'DEC']].mean().plot(kind= 'bar', color='red')
plt.xlabel('Months',fontsize=15)
plt.ylabel('Avg. Rainfall',fontsize=15)
plt.title('Avg. Monthly Rainfall Data',fontsize=25)
plt.grid()
plt.show()
```



These are the visualizations that we have made for our analysis of rainfall in India. These predictions show the highest rainfall region, average rainfall, the rainfall in each month with respect to the region, etc.