1) Loading the dataset

```
import pandas as pd

# Load the dataset
df = pd.read_csv('rain.csv')
```

2) EDA

```
# Dimensions of the dataframe
df.shape
```

```
# Datatypes of all the attributes
df.dtypes
```

```
#first five rows of the dataframe
df.head()
```

```
# basic stats
df.describe()
```

```
# Summary of dataframe
df.info()
```

3) Handling missing values

```
# Check missing values in each attributes
print(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
```

```
# Mean imputation to fill missing values
for column in df.columns:
    if df[column].dtype == 'object':
        df[column].fillna(df[column].mode()[0], inplace=True)
    else:
        df[column].fillna(df[column].mean(), inplace=True)

# Try using this also
# df = df.fillna(df.select_dtypes(include='number').mean())


# After imputing missing values
print(df.isnull().sum())
```

```
SUBDIVISION    0
YEAR           0
JAN            0
FEB            0
```

```
MAR            0
APR            0
MAY            0
JUN            0
JUL            0
AUG            0
SEP            0
OCT            0
NOV            0
DEC            0
ANNUAL         0
Jan-Feb        0
Mar-May        0
Jun-Sep        0
Oct-Dec        0
dtype: int64
```

4) Standardization

Standardization transforms the data to have a mean of 0 and a standard deviation of 1

```
from sklearn.preprocessing import StandardScaler

# Select columns for standardization (excluding 'SUBDIVISION' and 'YEAR')
rainfall_columns = df.columns[2:]

# Apply standardization
scaler = StandardScaler()
df[rainfall_columns] = scaler.fit_transform(df[rainfall_columns])
df
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	ANDAMAN & NICOBAR ISLANDS	1901	0.901019	1.819197	0.039233	-0.602264	3.596952	1.224806	0.066421	1.011559	0.999593	2.946952
1	ANDAMAN & NICOBAR ISLANDS	1902	-0.564795	3.844716	-0.323090	-0.636192	2.925549	1.308374	-0.439377	2.456519	3.465350	1.022838
2	ANDAMAN & NICOBAR ISLANDS	1903	-0.186424	3.404507	-0.583110	-0.621441	1.212540	1.064492	1.415586	0.193137	1.046898	0.861909
3	ANDAMAN & NICOBAR ISLANDS	1904	-0.284741	-0.197964	-0.583110	2.349501	1.775966	1.129300	0.574818	-0.689952	4.605097	1.274291
4	ANDAMAN & NICOBAR ISLANDS	1905	-0.526064	-0.607525	-0.512776	-0.239378	1.573002	1.698926	0.079790	0.213280	0.736461	1.661527
...	...	...	...	...	...	...	...	...	...	...	...	...
4111	LAKSHADWEEP	2011	-0.412851	-0.529514	-0.517039	0.630957	0.174180	-0.326744	0.011088	-0.192220	0.427502	0.220202
4112	LAKSHADWEEP	2012	0.007230	-0.604739	-0.549009	0.496719	-0.524014	0.412577	-0.429721	0.482023	-0.129806	0.506858
4113	LAKSHADWEEP	2013	0.215781	0.350904	0.216132	-0.558009	0.020739	0.835533	-0.188706	-0.720166	-0.128328	-0.228389
4114	LAKSHADWEEP	2014	1.020191	-0.158958	-0.489332	-0.416395	-0.230123	0.059118	-0.858275	0.932049	-0.481635	0.741211
4115	LAKSHADWEEP	2015	-0.499250	-0.593595	-0.504251	0.648659	0.384450	0.282961	-0.333167	-0.762571	-0.273199	0.702991

4116 rows × 19 columns

Next steps: Generate code with `df` View recommended plots New interactive sheet

5) Normalization

Normalization rescales the data to fit within a specific range, typically [0, 1]

```
from sklearn.preprocessing import MinMaxScaler

# Apply normalization (to range [0,1])
normalizer = MinMaxScaler()
df[rainfall_columns] = normalizer.fit_transform(df[rainfall_columns])
df
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	N
0	ANDAMAN & NICOBAR ISLANDS	1901	0.084290	0.215861	0.048217	0.003865	0.452507	0.321280	0.154520	0.289018	0.272117	0.409680	0.860
1	ANDAMAN & NICOBAR ISLANDS	1902	0.000000	0.396035	0.020145	0.000000	0.381739	0.333458	0.096877	0.452781	0.545135	0.207951	0.553
2	ANDAMAN & NICOBAR ISLANDS	1903	0.021758	0.356877	0.000000	0.001680	0.201181	0.297919	0.308278	0.196263	0.277355	0.191079	0.438
3	ANDAMAN & NICOBAR ISLANDS	1904	0.016104	0.036431	0.000000	0.340111	0.260568	0.307363	0.212460	0.096179	0.671332	0.234314	0.475
4	ANDAMAN & NICOBAR ISLANDS	1905	0.002227	0.000000	0.005449	0.045202	0.239175	0.390370	0.156044	0.198546	0.242982	0.274913	0.039
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4111	LAKSHADWEEP	2011	0.008737	0.006939	0.005119	0.144345	0.091734	0.095185	0.148214	0.152589	0.208773	0.123800	0.284
4112	LAKSHADWEEP	2012	0.032894	0.000248	0.002642	0.129054	0.018141	0.202920	0.097977	0.229004	0.147066	0.153854	0.019
4113	LAKSHADWEEP	2013	0.044886	0.085254	0.061922	0.008906	0.075560	0.264554	0.125444	0.092755	0.147230	0.076769	0.120
4114	LAKSHADWEEP	2014	0.091143	0.039901	0.007266	0.025038	0.049119	0.151413	0.049137	0.280007	0.108110	0.178425	0.090
4115	LAKSHADWEEP	2015	0.003769	0.001239	0.006110	0.146362	0.113897	0.184032	0.108981	0.087949	0.131189	0.174417	0.355

4116 rows × 19 columns

Next steps: Generate code with df | View recommended plots | New interactive sheet

6) Log Transformation

Log transformation is used to stabilize variance and make the data more normally distributed, especially for skewed data.

```
import numpy as np

# Log transformation (adding 1 to avoid log(0))
df[rainfall_columns] = np.log1p(df[rainfall_columns])
df
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	N
0	ANDAMAN & NICOBAR ISLANDS	1901	0.080925	0.195453	0.047090	0.003857	0.373291	0.278601	0.143685	0.253881	0.240683	0.343363	0.620
1	ANDAMAN & NICOBAR ISLANDS	1902	0.000000	0.333636	0.019945	0.000000	0.323343	0.287775	0.092467	0.373480	0.435111	0.188926	0.440
2	ANDAMAN & NICOBAR ISLANDS	1903	0.021524	0.305186	0.000000	0.001679	0.183305	0.260762	0.268712	0.179203	0.244791	0.174859	0.363
3	ANDAMAN & NICOBAR ISLANDS	1904	0.015976	0.035783	0.000000	0.292752	0.231563	0.268012	0.192651	0.091831	0.513621	0.210515	0.389
4	ANDAMAN & NICOBAR ISLANDS	1905	0.002225	0.000000	0.005434	0.044211	0.214446	0.329570	0.145004	0.181109	0.217514	0.242878	0.038
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4111	LAKSHADWEEP	2011	0.008699	0.006915	0.005106	0.134833	0.087767	0.090923	0.138208	0.142011	0.189606	0.116716	0.249
4112	LAKSHADWEEP	2012	0.032364	0.000248	0.002639	0.121380	0.017979	0.184752	0.093469	0.206204	0.137207	0.143108	0.018
4113	LAKSHADWEEP	2013	0.043908	0.081814	0.060081	0.008867	0.072842	0.234720	0.118178	0.088702	0.137350	0.073965	0.113
4114	LAKSHADWEEP	2014	0.087226	0.039125	0.007239	0.024729	0.047950	0.140990	0.047968	0.246866	0.102656	0.164178	0.087
4115	LAKSHADWEEP	2015	0.003762	0.001238	0.006091	0.136593	0.107865	0.168926	0.103441	0.084294	0.123269	0.160772	0.304

4116 rows × 19 columns

Next steps: Generate code with df | View recommended plots | New interactive sheet

7) Aggregation

Aggregation is a way to group data and compute aggregate functions, such as the mean, sum, or count.

```
# Aggregating the data by 'SUBDIVISION' and 'YEAR' (calculating the mean for each group)
rain_aggregated = df.groupby(['SUBDIVISION','YEAR']).mean().reset_index()
rain_aggregated
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	ANDAMAN & NICOBAR ISLANDS	1958.918182	0.081125	0.063535	0.048527	0.109329	0.260364	0.252121	0.154646	0.212222	0.298894	0.260822
1	ARUNACHAL PRADESH	1965.824742	0.076052	0.197791	0.216548	0.358575	0.261455	0.327268	0.250516	0.253411	0.295980	0.177998
2	ASSAM & MEGHALAYA	1958.000000	0.028340	0.073914	0.120241	0.289457	0.254042	0.273896	0.189339	0.216423	0.224950	0.146853
3	BIHAR	1958.000000	0.022379	0.034425	0.016405	0.027704	0.044164	0.101725	0.128024	0.164655	0.162187	0.062871
4	CHHATTISGARH	1958.000000	0.023602	0.045241	0.024378	0.027424	0.017706	0.114735	0.155327	0.209566	0.162603	0.063817
5	COASTAL ANDHRA PRADESH	1958.000000	0.012556	0.030473	0.021104	0.043392	0.050960	0.073396	0.070765	0.100019	0.137581	0.174858
6	COASTAL KARNATAKA	1958.000000	0.003538	0.003690	0.010120	0.049928	0.095734	0.417266	0.386478	0.351872	0.214672	0.175294
7	EAST MADHYA PRADESH	1958.000000	0.032051	0.044148	0.021892	0.011863	0.007852	0.082767	0.145193	0.199047	0.145075	0.040146
8	EAST RAJASTHAN	1958.000000	0.010848	0.013175	0.007323	0.005222	0.008316	0.038072	0.089795	0.121976	0.075682	0.014711
9	EAST UTTAR PRADESH	1958.000000	0.026771	0.037779	0.014420	0.010638	0.014551	0.065608	0.115271	0.152338	0.138935	0.043072
10	GANGETIC WEST BENGAL	1958.000000	0.021053	0.052658	0.045808	0.071534	0.087450	0.141272	0.128813	0.170660	0.181386	0.112708
11	GUJARAT REGION	1958.000000	0.003024	0.002924	0.001984	0.001853	0.004891	0.071276	0.136154	0.142178	0.111589	0.020865
12	HARYANA DELHI & CHANDIGARH	1958.000000	0.028199	0.041343	0.020842	0.012532	0.012272	0.029316	0.061183	0.085834	0.068190	0.013139

13	HIMACHAL PRADESH	1958.000000	0.131818	0.197800	0.150349	0.098320	0.048195	0.054489	0.111460	0.151170	0.099240	0.031321
14	JAMMU & KASHMIR	1958.000000	0.156799	0.243828	0.191806	0.143844	0.055690	0.038702	0.073428	0.102475	0.068954	0.034801
15	JHARKHAND	1958.000000	0.029189	0.056532	0.029482	0.031693	0.040243	0.112775	0.132780	0.177909	0.169367	0.079106
16	KERALA	1958.000000	0.020439	0.036973	0.057979	0.168455	0.174349	0.337536	0.257149	0.223239	0.179698	0.267340
17	KONKAN & GOA	1958.000000	0.002139	0.001342	0.002238	0.007034	0.027200	0.352248	0.371181	0.338165	0.246516	0.109686
18	LAKSHADWEEP	1958.350877	0.044071	0.037026	0.023241	0.070358	0.127107	0.182835	0.112420	0.117033	0.123690	0.157740
19	MADHYA MAHARASHTRA	1958.000000	0.005158	0.003606	0.005868	0.015136	0.019272	0.086930	0.099790	0.104402	0.119903	0.070356
20	MATATHWADA	1958.000000	0.008376	0.010736	0.011465	0.012548	0.013138	0.080903	0.073248	0.094293	0.134263	0.058772
21	NAGA MANI MIZO TRIPURA	1958.000000	0.023374	0.083811	0.116309	0.247132	0.219393	0.242469	0.169491	0.219172	0.227632	0.166873
22	NORTH INTERIOR KARNATAKA	1958.000000	0.005094	0.007741	0.011495	0.039711	0.039226	0.060430	0.056778	0.068863	0.109615	0.094787
23	ORISSA	1958.000000	0.020572	0.046417	0.033700	0.055165	0.053544	0.122003	0.138052	0.192706	0.179152	0.110698
24	PUNJAB	1958.000000	0.041697	0.062844	0.037675	0.020692	0.011944	0.028022	0.068716	0.089973	0.066507	0.013768
25	RAYALSEEMA	1958.000000	0.016282	0.013602	0.013013	0.032352	0.041836	0.039001	0.039690	0.062098	0.101588	0.131430
26	SAURASHTRA & KUTCH	1958.000000	0.001942	0.003942	0.002097	0.001937	0.003888	0.044255	0.078223	0.067461	0.058265	0.014734
27	SOUTH INTERIOR KARNATAKA	1958.000000	0.004972	0.010101	0.015301	0.068056	0.075410	0.083647	0.093043	0.099071	0.105739	0.135382
28	SUB HIMALAYAN WEST BENGAL & SIKKIM	1958.000000	0.023456	0.054374	0.067703	0.167538	0.206113	0.286236	0.240339	0.269507	0.293881	0.137895
29	TAMIL NADU	1958.000000	0.038686	0.031721	0.031059	0.071970	0.057792	0.031525	0.029656	0.055788	0.086902	0.174915

30	TELANGANA	1958.000000	0.012850	0.023090	0.020180	0.029777	0.021292	0.083860	0.098959	0.120454	0.132696	0.073742
31	UTTARAKHAND	1958.000000	0.086200	0.141759	0.088400	0.056702	0.045841	0.094843	0.152387	0.205348	0.146105	0.038858
32	VIDARBHA	1958.000000	0.017625	0.028452	0.019109	0.015510	0.009762	0.101267	0.129937	0.157569	0.132458	0.052457
33	WEST MADHYA PRADESH	1958.000000	0.015534	0.015605	0.008403	0.003967	0.006493	0.066304	0.120007	0.158390	0.121497	0.028607
34	WEST RAJASTHAN	1958.000000	0.005656	0.011967	0.006482	0.005936	0.008007	0.017297	0.039263	0.054563	0.031939	0.005320
35	WEST UTTAR PRADESH	1958.000000	0.029480	0.042324	0.018491	0.010321	0.010429	0.046322	0.098721	0.139360	0.110944	0.028920

Next

Generate code with `rain_aggregated` View recommended plots New interactive sheet

8) Discretization

Discretization involves converting continuous variables into discrete categories. For example, we can categorize the ANNUAL rainfall into "low", "medium", and "high" bins.

```
# Discretizing the 'ANNUAL' rainfall into three categories: low, medium, and high
df['rainfall_category'] = pd.cut(df['ANNUAL'], bins=[-np.inf, 0.33, 0.66, np.inf],
                                 labels=["low", "medium", "high"])
df
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	ANDAMAN & NICOBAR ISLANDS	1958.918182	0.081125	0.063535	0.048527	0.109329	0.260364	0.252121	0.154646	0.212222	0.298894	0.260822
1	ARUNACHAL PRADESH	1965.824742	0.076052	0.197791	0.216548	0.358575	0.261455	0.327268	0.250516	0.253411	0.295980	0.177998
2	ASSAM & MEGHALAYA	1958.000000	0.028340	0.073914	0.120241	0.289457	0.254042	0.273896	0.189339	0.216423	0.224950	0.146853
3	BIHAR	1958.000000	0.022379	0.034425	0.016405	0.027704	0.044164	0.101725	0.128024	0.164655	0.162187	0.062871
4	CHHATTISGARH	1958.000000	0.023602	0.045241	0.024378	0.027424	0.017706	0.114735	0.155327	0.209566	0.162603	0.063817
5	COASTAL ANDHRA PRADESH	1958.000000	0.012556	0.030473	0.021104	0.043392	0.050960	0.073396	0.070765	0.100019	0.137581	0.174858
6	COASTAL KARNATAKA	1958.000000	0.003538	0.003690	0.010120	0.049928	0.095734	0.417266	0.386478	0.351872	0.214672	0.175294
7	EAST MADHYA PRADESH	1958.000000	0.032051	0.044148	0.021892	0.011863	0.007852	0.082767	0.145193	0.199047	0.145075	0.040146
8	EAST RAJASTHAN	1958.000000	0.010848	0.013175	0.007323	0.005222	0.008316	0.038072	0.089795	0.121976	0.075682	0.014711
9	EAST UTTAR PRADESH	1958.000000	0.026771	0.037779	0.014420	0.010638	0.014551	0.065608	0.115271	0.152338	0.138935	0.043072
10	GANGETIC WEST BENGAL	1958.000000	0.021053	0.052658	0.045808	0.071534	0.087450	0.141272	0.128813	0.170660	0.181386	0.112708
11	GUJARAT REGION	1958.000000	0.003024	0.002924	0.001984	0.001853	0.004891	0.071276	0.136154	0.142178	0.111589	0.020865
12	HARYANA DELHI & CHANDIGARH	1958.000000	0.028199	0.041343	0.020842	0.012532	0.012272	0.029316	0.061183	0.085834	0.068190	0.013139

13	HIMACHAL PRADESH	1958.000000	0.131818	0.197800	0.150349	0.098320	0.048195	0.054489	0.111460	0.151170	0.099240	0.031321
14	JAMMU & KASHMIR	1958.000000	0.156799	0.243828	0.191806	0.143844	0.055690	0.038702	0.073428	0.102475	0.068954	0.034801
15	JHARKHAND	1958.000000	0.029189	0.056532	0.029482	0.031693	0.040243	0.112775	0.132780	0.177909	0.169367	0.079106
16	KERALA	1958.000000	0.020439	0.036973	0.057979	0.168455	0.174349	0.337536	0.257149	0.223239	0.179698	0.267340
17	KONKAN & GOA	1958.000000	0.002139	0.001342	0.002238	0.007034	0.027200	0.352248	0.371181	0.338165	0.246516	0.109686
18	LAKSHADWEEP	1958.350877	0.044071	0.037026	0.023241	0.070358	0.127107	0.182835	0.112420	0.117033	0.123690	0.157740
19	MADHYA MAHARASHTRA	1958.000000	0.005158	0.003606	0.005868	0.015136	0.019272	0.086930	0.099790	0.104402	0.119903	0.070356
20	MATATHWADA	1958.000000	0.008376	0.010736	0.011465	0.012548	0.013138	0.080903	0.073248	0.094293	0.134263	0.058772
21	NAGA MANI MIZO TRIPURA	1958.000000	0.023374	0.083811	0.116309	0.247132	0.219393	0.242469	0.169491	0.219172	0.227632	0.166873
22	NORTH INTERIOR KARNATAKA	1958.000000	0.005094	0.007741	0.011495	0.039711	0.039226	0.060430	0.056778	0.068863	0.109615	0.094787
23	ORISSA	1958.000000	0.020572	0.046417	0.033700	0.055165	0.053544	0.122003	0.138052	0.192706	0.179152	0.110698
24	PUNJAB	1958.000000	0.041697	0.062844	0.037675	0.020692	0.011944	0.028022	0.068716	0.089973	0.066507	0.013768
25	RAYALSEEMA	1958.000000	0.016282	0.013602	0.013013	0.032352	0.041836	0.039001	0.039690	0.062098	0.101588	0.131430
26	SAURASHTRA & KUTCH	1958.000000	0.001942	0.003942	0.002097	0.001937	0.003888	0.044255	0.078223	0.067461	0.058265	0.014734
27	SOUTH INTERIOR KARNATAKA	1958.000000	0.004972	0.010101	0.015301	0.068056	0.075410	0.083647	0.093043	0.099071	0.105739	0.135382
28	SUB HIMALAYAN WEST BENGAL & SIKKIM	1958.000000	0.023456	0.054374	0.067703	0.167538	0.206113	0.286236	0.240339	0.269507	0.293881	0.137895
29	TAMIL NADU	1958.000000	0.038686	0.031721	0.031059	0.071970	0.057792	0.031525	0.029656	0.055788	0.086902	0.174915

30	TELANGANA	1958.000000	0.012850	0.023090	0.020180	0.029777	0.021292	0.083860	0.098959	0.120454	0.132696	0.073742
31	UTTARAKHAND	1958.000000	0.086200	0.141759	0.088400	0.056702	0.045841	0.094843	0.152387	0.205348	0.146105	0.038858
32	VIDARBHA	1958.000000	0.017625	0.028452	0.019109	0.015510	0.009762	0.101267	0.129937	0.157569	0.132458	0.052457
33	WEST MADHYA PRADESH	1958.000000	0.015534	0.015605	0.008403	0.003967	0.006493	0.066304	0.120007	0.158390	0.121497	0.028607
34	WEST RAJASTHAN	1958.000000	0.005656	0.011967	0.006482	0.005936	0.008007	0.017297	0.039263	0.054563	0.031939	0.005320
35	WEST UTTAR PRADESH	1958.000000	0.029480	0.042324	0.018491	0.010321	0.010429	0.046322	0.098721	0.139360	0.110944	0.028920