Credentials

1. Name : Harish Thangaraj
2. VIT Mail ID : harish.thangaraj2021@vitstudent.ac.in
3. Date of assignment: 08-09-2023

Task

1. Take car crashes dataset from seaborn library
2. load the dataset
3. Perform Data Visualization
4. Inference is must for each and every graph

In [19]:
```
#Importing necessary libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [3]:
```
print(sns.get_dataset_names())
```

```
['anagrams', 'anscombe', 'attention', 'brain_networks', 'car_crashes', 'diamonds', 'dots', 'dowjones', 'exercise', 'flights', 'fmri', 'geyser', 'glue', 'healthexp', 'iris', 'mpg', 'penguins', 'planets', 'seaice', 'taxis', 'tips', 'titanic']
```

In [5]:
```
#Loading dataset
data=sns.load_dataset('car_crashes')
data
```

Out[5]:

	total	speeding	alcohol	not_distracted	no_previous	ins_premium	ins_losses	abbrev
0	18.8	7.332	5.640	18.048	15.040	784.55	145.08	AL
1	18.1	7.421	4.525	16.290	17.014	1053.48	133.93	AK
2	18.6	6.510	5.208	15.624	17.856	899.47	110.35	AZ
3	22.4	4.032	5.824	21.056	21.280	827.34	142.39	AR
4	12.0	4.200	3.360	10.920	10.680	878.41	165.63	CA
5	13.6	5.032	3.808	10.744	12.920	835.50	139.91	CO
6	10.8	4.968	3.888	9.396	8.856	1068.73	167.02	CT
7	16.2	6.156	4.860	14.094	16.038	1137.87	151.48	DE
8	5.9	2.006	1.593	5.900	5.900	1273.89	136.05	DC
9	17.9	3.759	5.191	16.468	16.826	1160.13	144.18	FL
10	15.6	7.964	3.900	14.820	14.508	913.15	142.80	GA
11	17.5	9.450	7.175	14.350	15.225	861.18	120.92	HI
12	19.3	5.509	4.437	13.506	14.864	641.96	82.75	ID
13	12.8	4.608	4.352	12.032	12.288	803.11	139.15	IL
14	14.5	3.625	4.205	13.775	13.775	710.46	108.92	IN
15	15.7	2.669	3.925	15.229	13.659	649.06	114.47	IA
16	17.8	4.806	4.272	13.706	15.130	780.45	133.80	KS
17	21.4	4.066	4.922	16.692	16.264	872.51	137.13	KY
18	20.5	7.175	6.765	14.965	16.195	1281.55	194.78	LA
19	15.1	5.738	4.530	13.137	12.684	661.88	96.57	ME
20	12.5	4.250	4.000	8.875	12.375	1048.78	192.70	MD
21	8.2	1.886	2.870	7.134	6.560	1011.14	135.63	MA
22	14.1	3.384	3.948	13.395	10.857	1110.61	152.26	MI
23	9.6	2.208	2.784	8.448	8.448	777.18	133.35	MN
24	17.6	3.640	5.456	15.752	17.600	896.07	155.77	MS
25	16.1	6.923	5.474	14.812	13.524	790.32	144.45	MO
26	21.4	8.346	9.416	17.976	18.190	816.21	85.15	MT
27	14.9	3.427	5.811	13.857	13.410	732.28	114.82	NE
28	14.7	5.439	4.704	13.965	14.553	1029.87	138.71	NV
29	11.6	4.060	3.480	10.092	9.628	746.54	120.21	NH
30	11.2	3.696	3.696	9.632	8.736	1301.52	159.85	NJ
31	18.4	3.496	4.968	12.328	18.032	869.85	120.75	NM
32	12.3	3.936	3.567	10.824	9.840	1234.31	150.01	NY
33	16.8	6.552	5.208	15.792	13.608	708.24	127.82	NC
34	23.9	5.497	10.038	23.661	20.554	688.75	109.72	ND
35	14.1	3.948	4.794	13.959	11.562	697.73	133.52	OH
36	19.9	5.368	5.771	18.308	18.706	881.51	178.86	OK
37	12.8	4.224	3.328	8.576	11.520	804.71	104.61	OR
38	18.2	9.100	5.642	17.472	16.016	905.99	153.86	PA
39	11.1	3.774	4.218	10.212	8.769	1148.99	148.58	RI
40	23.9	9.082	9.799	22.944	19.359	858.97	118.29	SC
41	19.4	6.014	6.402	19.012	16.684	669.31	96.87	SD
42	19.5	4.095	5.655	15.990	15.795	767.91	155.57	TN
43	19.4	7.760	7.372	17.654	16.878	1004.75	156.83	TX
44	11.3	4.859	1.808	9.944	10.848	809.38	109.48	UT
45	13.6	4.080	4.080	13.056	12.920	716.20	109.61	VT
46	12.7	2.413	3.429	11.049	11.176	768.95	153.72	VA
47	10.0	4.452	3.488	8.692	9.116	890.03	111.62	WA
48	23.6	9.092	8.684	23.006	20.706	992.61	152.56	WV
49	13.8	4.968	4.554	5.382	11.592	670.31	106.62	WI
50	17.4	7.308	5.568	14.094	15.660	791.14	122.04	WY

Dataset details:

1. total -> Number of drivers involved in fatal collisions per billion miles.
2. speeding -> Percentage Of Drivers Involved In Fatal Collisions Who Were speeding.
3. alcohol -> Percentage Of Drivers Involved In Fatal Collisions Who Were Alcohol-Impaired.
4. not_distracted -> Percentage Of Drivers Involved In Fatal Collisions Who Were Not Distracted.
5. no_previous -> Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents.
6. ins_premium -> Car Insurance Premiums
7. ins_losses -> Losses incurred by insurance companies for collisions per insured driver.
8. abbrev -> USA states.

In [6]:
```
#Printing correlation matrix
corr=data.corr()
corr
```

Out[6]:

	total	speeding	alcohol	not_distracted	no_previous	ins_premium	ins_losses
total	1.000000	0.611548	0.852613	0.827560	0.956179	-0.139702	-0.175011
speeding	0.611548	1.000000	0.669719	0.588510	0.573976	-0.077675	-0.069328
alcohol	0.852613	0.669719	1.000000	0.732016	0.783520	-0.170612	-0.112547
not_distracted	0.827560	0.588510	0.732016	1.000000	0.747307	-0.174956	-0.075970
no_previous	0.956179	0.573976	0.783520	0.747307	1.000000	-0.156695	-0.090359
ins_premium	-0.139702	-0.077675	-0.170612	-0.174956	-0.156695	1.000000	0.623116
ins_losses	-0.175011	-0.069328	-0.112547	-0.075970	-0.090359	0.623116	1.000000

In [20]:
```
#Plotting heatmap for the above correlation
sns.heatmap(corr,annot=True)
```

Out[20]:

Inference:

1. Positive correlation represents variables are moving in the same direction
2. Negative correlation represents variables moving in opposite direction
3. Correlation = 1 represents perfect correlation
4. Darker shades represent higher correlation between the variables

In [46]:
```
#Insurance premium vs insurance losses using scatterplot
sns.scatterplot(x="ins_premium",y="ins_losses",data=data)
plt.show()
```


In [49]:
```
#Insurance premium vs insurance losses using lineplot
sns.lineplot(x="ins_premium",y="ins_losses",data=data,ci=None)
plt.show()
```


Inference:

1. From the above scatterplot and line, the insurance losses are increasing with respect to increase in insurance premium paid.

In [58]:
```
#Stats vs total using barplot
plt.subplots(figsize=(25,20))
sns.barplot(data=data,x="abbrev",y="total")
plt.show()
```


Inference:

1. Clearly AR,ND,SC,WV are states with the highest total car crashes

In [34]:
```
#Boxplot for speeding crashes
speeding=data.iloc[0:6,1:2]
plt.boxplot(speeding)
plt.show()
```


Inference:

1. Clearly, from the box plot the speeding data is symmetrical

In [57]:
```
#Univariate observation of alcohol percentage density using distplot
sns.distplot(data['alcohol'])
plt.show()
```

```
/Users/casanalez/opt/anaconda3/lib/python3.9/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```


Inference:

1. The density of alcohol related car crashes reaches a density peak between 4 and 6.

In [62]:
```
#Plotting a pie chart holding total crashes in respective states (first 6)
total=data.total.values.tolist()
abbrev=data.abbrev.values.tolist()
df=pd.DataFrame({"total":total[0:6],"abbrev":abbrev[0:6]})
plt.pie(df["total"], labels = df["abbrev"],autopct='%.1f%%',explode=[0.0,0.0,0.0,0.0,1.0,0.0,0.0])
plt.show()
```


Inference:

1. Clearly, AR state has the highest car crash percentage of 21.1% and the respective percentages are displayed in the pie plot above.

In [ ]:
```

```