WEATHER DATA ANALYSIS

```
from google.colab import files
df=files.upload
import numpy as np
import pandas as pd
df=pd.read_csv('/content/weather_Nexus_phase1.csv')
```

```
df.shape
```

```
(366, 22)
```

```
df.head()
```

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am
0	8.0	24.3	0.0	3.4	6.3	NW	30.0	SW	NW	6.0
1	14.0	26.9	3.6	4.4	9.7	ENE	39.0	E	W	4.0
2	13.7	23.4	3.6	5.8	3.3	NW	85.0	N	NNE	6.0
3	13.3	15.5	39.8	7.2	9.1	NW	54.0	WNW	W	30.0
4	7.6	16.1	2.8	5.6	10.6	SSE	50.0	SSE	ESE	20.0

5 rows × 22 columns

```
df.tail()
```

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	Winc
361	9.0	30.7	0.0	7.6	12.1	NNW	76.0	
362	7.1	28.4	0.0	11.6	12.7	N	48.0	
363	12.5	19.9	0.0	8.4	5.3	ESE	43.0	
364	12.5	26.9	0.0	5.0	7.1	NW	46.0	
365	12.3	30.2	0.0	6.0	12.6	NW	78.0	

5 rows × 22 columns

```
df.dtypes
```

```
MinTemp          float64
MaxTemp          float64
Rainfall         float64
Evaporation      float64
Sunshine         float64
WindGustDir       object
WindGustSpeed    float64
WindDir9am        object
WindDir3pm        object
WindSpeed9am     float64
WindSpeed3pm       int64
Humidity9am        int64
Humidity3pm        int64
Pressure9am      float64
Pressure3pm      float64
Cloud9am           int64
Cloud3pm           int64
Temp9am          float64
Temp3pm          float64
RainToday         object
RISK_MM          float64
RainTomorrow      object
dtype: object
```

```
df.isnull().sum()
```

```
MinTemp           0
MaxTemp           0
Rainfall          0
Evaporation       0
Sunshine          3
WindGustDir       3
WindGustSpeed     2
WindDir9am       31
WindDir3pm        1
WindSpeed9am      7
WindSpeed3pm      0
Humidity9am       0
Humidity3pm       0
Pressure9am       0
Pressure3pm       0
```

```
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
RainToday        0
RISK_MM          0
RainTomorrow     0
dtype: int64
```

```
df.nunique()
```

```
MinTemp          180
MaxTemp          187
Rainfall          47
Evaporation       55
Sunshine         114
WindGustDir       16
WindGustSpeed     35
WindDir9am        16
WindDir3pm        16
WindSpeed9am      22
WindSpeed3pm      26
Humidity9am       60
Humidity3pm       74
Pressure9am      190
Pressure3pm      193
Cloud9am           9
Cloud3pm           9
Temp9am          178
Temp3pm          200
RainToday          2
RISK_MM           47
RainTomorrow       2
dtype: int64
```

```
df=df.drop_duplicates()
```

```
df.describe()
```

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSp
count	366.000000	366.000000	366.000000	366.000000	363.000000	364.000000	359.
mean	7.265574	20.550273	1.428415	4.521858	7.909366	39.840659	9.
std	6.025800	6.690516	4.225800	2.669383	3.481517	13.059807	7.
min	-5.300000	7.600000	0.000000	0.200000	0.000000	13.000000	0.
25%	2.300000	15.025000	0.000000	2.200000	5.950000	31.000000	6.
50%	7.450000	19.650000	0.000000	4.200000	8.600000	39.000000	7.
75%	12.500000	25.500000	0.200000	6.400000	10.500000	46.000000	13.
max	20.900000	35.800000	39.800000	13.800000	13.600000	98.000000	41.

```
df.columns
```

```
Index(['MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine',
       'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm',
       'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',
       'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am',
       'Temp3pm', 'RainToday', 'RISK_MM', 'RainTomorrow'],
      dtype='object')
```

```
df=df.fillna('N/A')
```

```
df.isnull().sum()
```

```
MinTemp          0
MaxTemp          0
Rainfall         0
Evaporation      0
Sunshine         0
WindGustDir      0
WindGustSpeed    0
WindDir9am       0
WindDir3pm       0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
RainToday        0
```

```
RISK_MM          0
RainTomorrow     0
dtype: int64
```

To find outlier thresholds

```
# Outlier using IQR
Q1=df['Humidity9am'].quantile(0.25)
print(Q1)
Q3=df['Humidity9am'].quantile(0.75)
print(Q3)
```

```
64.0
81.0
```

```
IQR=Q3-Q1
IQR
```

```
17.0
```

```
# To find outlier thresolds
lower_bound = Q1-1.5*IQR
upper_bound = Q3+1.5*IQR
print(lower_bound)
print(upper_bound)
```

```
38.5
106.5
```

```
upper_array = np.where(df['Humidity9am']>=upper_bound)[0]
lower_array = np.where(df['Humidity9am']<=lower_bound)[0]
print(upper_array)
print(lower_array)
```

```
[]
[332 361]
```

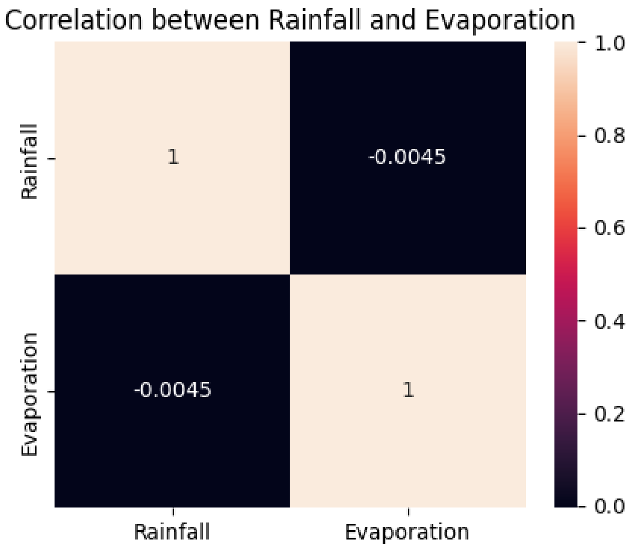
```
#To Remove outliers
df.drop(index=upper_array, inplace=True)
df.drop(index=lower_array, inplace=True)
print("New Shape of the test data: ", df.shape)
```

```
New Shape of the test data:  (364, 22)
```

```
# To find out correlation
corr=df[['Rainfall','Evaporation']].corr()
corr
```

	Rainfall	Evaporation
Rainfall	1.000000	-0.004548
Evaporation	-0.004548	1.000000

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(5, 4))
sns.heatmap(corr, annot=True)
plt.title('Correlation between Rainfall and Evaporation')
plt.show()
```

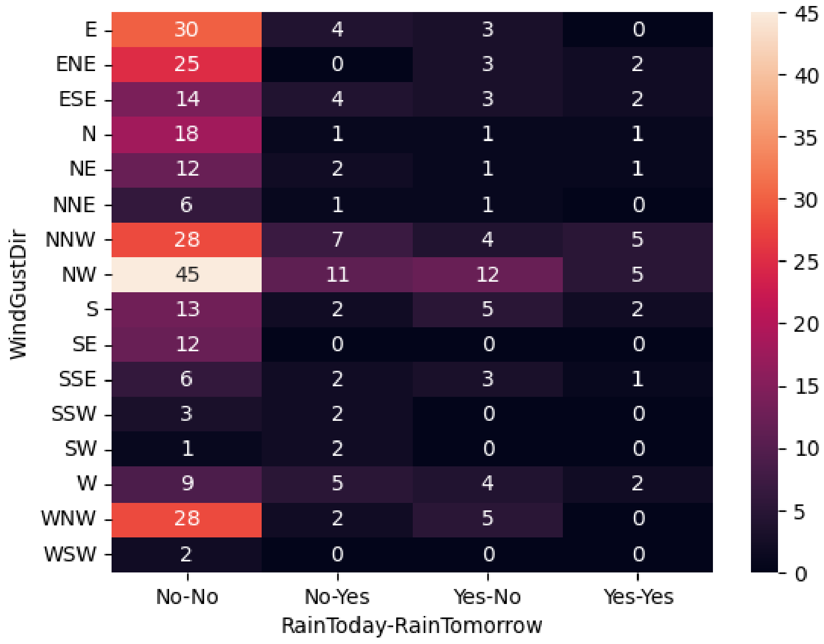


```
# To find out Regression
```

```
cross_tab=pd.crosstab(index=df['WindGustDir'],columns=[df['RainToday'],df['RainTomorrow']])
cross_tab
```

RainToday	No		Yes	
RainTomorrow	No	Yes	No	Yes
WindGustDir				
E	30	4	3	0
ENE	25	0	3	2
ESE	14	4	3	2
N	18	1	1	1
NE	12	2	1	1
NNE	6	1	1	0
NNW	28	7	4	5
NW	45	11	12	5
S	13	2	5	2
SE	12	0	0	0
SSE	6	2	3	1
SSW	3	2	0	0
SW	1	2	0	0
W	9	5	4	2
WNW	28	2	5	0
WSW	2	0	0	0

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.heatmap(cross_tab,annot=True,fmt='d')
plt.show()
```


WindGustDir
E
ENE
ESE
N
NE
NNE
NNW
NW
S
SE
SSE
SSW
SW
W
WNW
WSW
No-No
No-Yes
Yes-No
Yes-Yes
RainToday-RainTomorrow