In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df=pd.read_csv(r"C:\Users\karan\OneDrive\Desktop\MyWorld\data project\australiaW
```

In [3]:
```
df.columns
```

Out[3]:
```
Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation',
       'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm',
       'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',
       'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am',
       'Temp3pm', 'RainToday', 'RainTomorrow'],
      dtype='object')
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   Date           145460 non-null  object 
 1   Location       145460 non-null  object 
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object 
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object 
 10  WindDir3pm     141232 non-null  object 
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object 
 22  RainTomorrow   142193 non-null  object 
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

In [5]:
```
df.shape
```

Out[5]:
```
(145460, 23)
```

In [6]:
```
df.describe
```

Out[6]:
```
<bound method NDFrame.describe of              Date Location  MinTemp  MaxTemp
Rainfall  Evaporation  \
0       2008-12-01   Albury     13.4     22.9       0.6          NaN
1       2008-12-02   Albury      7.4     25.1       0.0          NaN
2       2008-12-03   Albury     12.9     25.7       0.0          NaN
3       2008-12-04   Albury      9.2     28.0       0.0          NaN
4       2008-12-05   Albury     17.5     32.3       1.0          NaN
...            ...      ...      ...      ...       ...          ...
145455  2017-06-21    Uluru      2.8     23.4       0.0          NaN
145456  2017-06-22    Uluru      3.6     25.3       0.0          NaN
145457  2017-06-23    Uluru      5.4     26.9       0.0          NaN
145458  2017-06-24    Uluru      7.8     27.0       0.0          NaN
145459  2017-06-25    Uluru     14.9      NaN       0.0          NaN

        Sunshine WindGustDir  WindGustSpeed WindDir9am  ... Humidity9am  \
0            NaN           W           44.0          W  ...        71.0
1            NaN         WNW           44.0        NNW  ...        44.0
2            NaN         WSW           46.0          W  ...        38.0
3            NaN          NE           24.0         SE  ...        45.0
4            NaN           W           41.0        ENE  ...        82.0
...          ...         ...            ...        ...  ...         ...
145455       NaN           E           31.0         SE  ...        51.0
145456       NaN         NNW           22.0         SE  ...        56.0
145457       NaN           N           37.0         SE  ...        53.0
145458       NaN          SE           28.0        SSE  ...        51.0
145459       NaN         NaN            NaN        ESE  ...        62.0

        Humidity3pm  Pressure9am  Pressure3pm  Cloud9am  Cloud3pm  Temp9am  \
0              22.0       1007.7       1007.1       8.0       NaN     16.9
1              25.0       1010.6       1007.8       NaN       NaN     17.2
2              30.0       1007.6       1008.7       NaN       2.0     21.0
3              16.0       1017.6       1012.8       NaN       NaN     18.1
4              33.0       1010.8       1006.0       7.0       8.0     17.8
...             ...          ...          ...       ...       ...      ...
145455         24.0       1024.6       1020.3       NaN       NaN     10.1
145456         21.0       1023.5       1019.1       NaN       NaN     10.9
145457         24.0       1021.0       1016.8       NaN       NaN     12.5
145458         24.0       1019.4       1016.5       3.0       2.0     15.1
145459         36.0       1020.2       1017.9       8.0       8.0     15.0

        Temp3pm  RainToday  RainTomorrow
0          21.8         No            No
1          24.3         No            No
2          23.2         No            No
3          26.5         No            No
4          29.7         No            No
...         ...        ...           ...
145455     22.4         No            No
145456     24.5         No            No
145457     26.1         No            No
145458     26.0         No            No
145459     20.9         No           NaN

[145460 rows x 23 columns]>
```

In [7]:
```
df['Date']=pd.to_datetime(df['Date'])
```

In [8]:
```
df.head()
```

Out[8]:

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W

5 rows × 23 columns

In [9]:
```
df.duplicated().sum()
```

Out[9]:
```
0
```

In [10]:
```
df.isnull()
```

Out[10]:

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGu
0	False	False	False	False	False	True	True	
1	False	False	False	False	False	True	True	
2	False	False	False	False	False	True	True	
3	False	False	False	False	False	True	True	
4	False	False	False	False	False	True	True	
...	...	...	...	...	...	...	...	
145455	False	False	False	False	False	True	True	
145456	False	False	False	False	False	True	True	
145457	False	False	False	False	False	True	True	
145458	False	False	False	False	False	True	True	
145459	False	False	False	True	False	True	True	

145460 rows × 23 columns

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
Date                0
Location            0
MinTemp          1485
MaxTemp          1261
Rainfall         3261
Evaporation     62790
Sunshine        69835
WindGustDir     10326
WindGustSpeed   10263
WindDir9am      10566
WindDir3pm       4228
WindSpeed9am     1767
WindSpeed3pm     3062
Humidity9am      2654
Humidity3pm      4507
Pressure9am     15065
Pressure3pm     15028
Cloud9am        55888
Cloud3pm        59358
Temp9am          1767
Temp3pm          3609
RainToday        3261
RainTomorrow     3267
dtype: int64
```

In [12]:
```
missing_percent=((df.isnull().sum())/len(df))*100
print(missing_percent)
```

```
Date             0.000000
Location         0.000000
MinTemp          1.020899
MaxTemp          0.866905
Rainfall         2.241853
Evaporation     43.166506
Sunshine        48.009762
WindGustDir      7.098859
WindGustSpeed    7.055548
WindDir9am       7.263853
WindDir3pm       2.906641
WindSpeed9am     1.214767
WindSpeed3pm     2.105046
Humidity9am      1.824557
Humidity3pm      3.098446
Pressure9am     10.356799
Pressure3pm     10.331363
Cloud9am        38.421559
Cloud3pm        40.807095
Temp9am          1.214767
Temp3pm          2.481094
RainToday        2.241853
RainTomorrow     2.245978
dtype: float64
```

In [13]:
```
df=df.dropna(subset=['RainTomorrow'])
```

In [14]:
```
df['RainTomorrow'].isnull().sum()
```

Out[14]:
```
0
```

In [15]:
```
df=df.ffill()
```

In [16]:
```
df.drop(['Date','Evaporation','Sunshine','WindGustDir','WindDir9am','WindDir3pm'
```

In [17]:
```
df.dtypes
```

Out[17]:
```
Location          object
MinTemp          float64
MaxTemp          float64
Rainfall         float64
WindGustSpeed    float64
WindSpeed9am     float64
WindSpeed3pm     float64
Humidity9am      float64
Humidity3pm      float64
Pressure9am      float64
Pressure3pm      float64
Cloud9am         float64
Cloud3pm         float64
Temp9am          float64
Temp3pm          float64
RainToday         object
RainTomorrow      object
dtype: object
```

In [18]:
```
df['RainToday']=df["RainToday"].map({'No':0,'Yes':1})
df['RainTomowrrow']=df['RainTomorrow'].map({'No':0,'Yes':1})
```

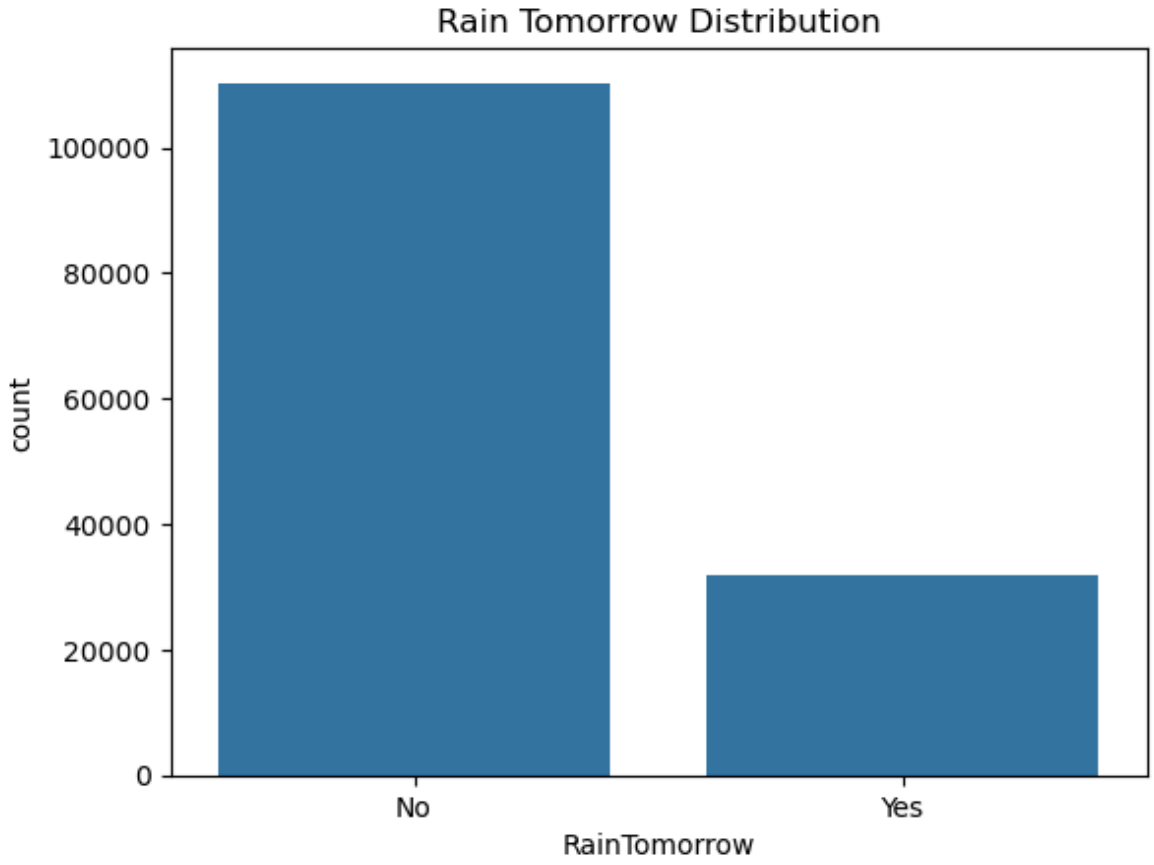
In [19]:
```
print(f"Missing Values :{df.isnull().sum()}")
print(f"Shape:{df.shape}")
df.head()
```

```
Missing Values :Location         0
MinTemp          0
MaxTemp          0
Rainfall         0
WindGustSpeed    0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         2
Temp9am          0
Temp3pm          0
RainToday        0
RainTomorrow     0
RainTomowrrow    0
dtype: int64
Shape:(142193, 18)
```

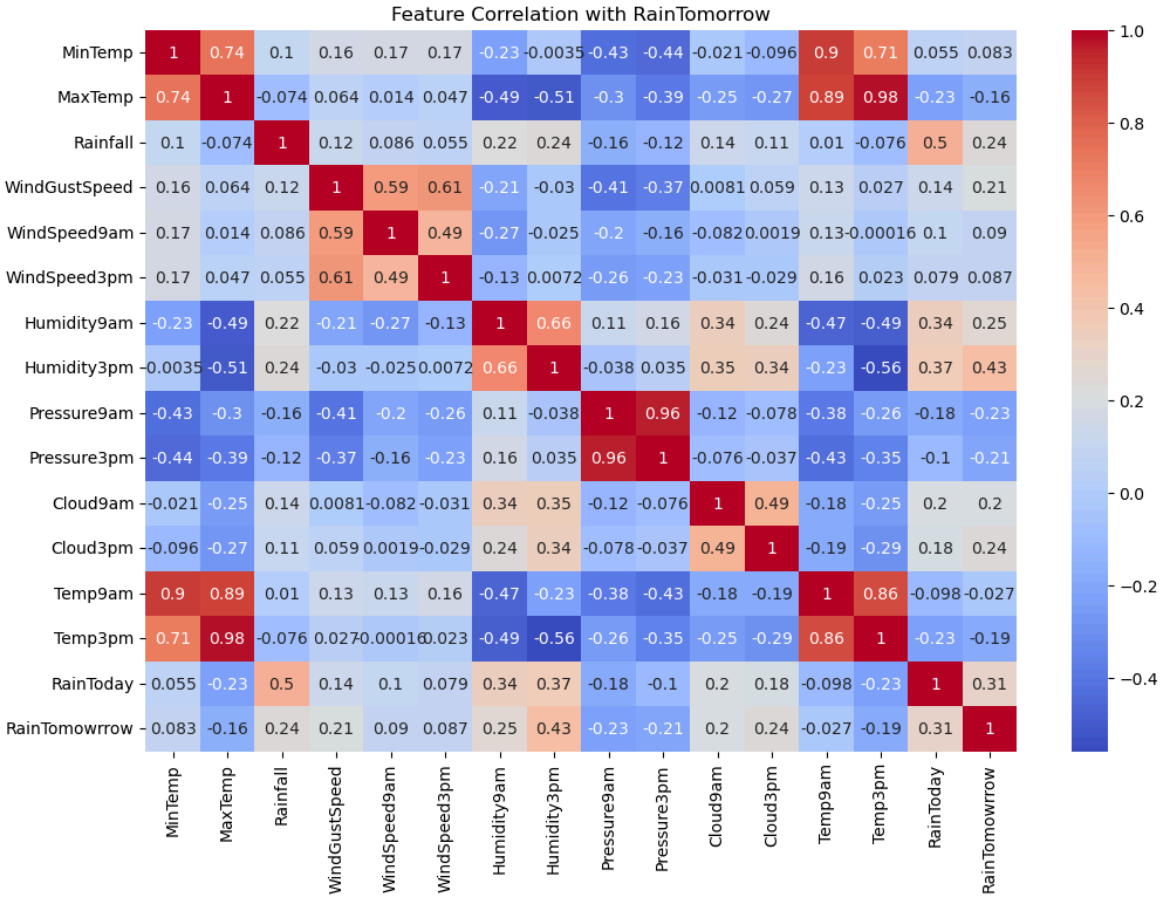
Out[19]:

	Location	MinTemp	MaxTemp	Rainfall	WindGustSpeed	WindSpeed9am	WindSpee
0	Albury	13.4	22.9	0.6	44.0	20.0	
1	Albury	7.4	25.1	0.0	44.0	4.0	
2	Albury	12.9	25.7	0.0	46.0	19.0	
3	Albury	9.2	28.0	0.0	24.0	11.0	
4	Albury	17.5	32.3	1.0	41.0	7.0	

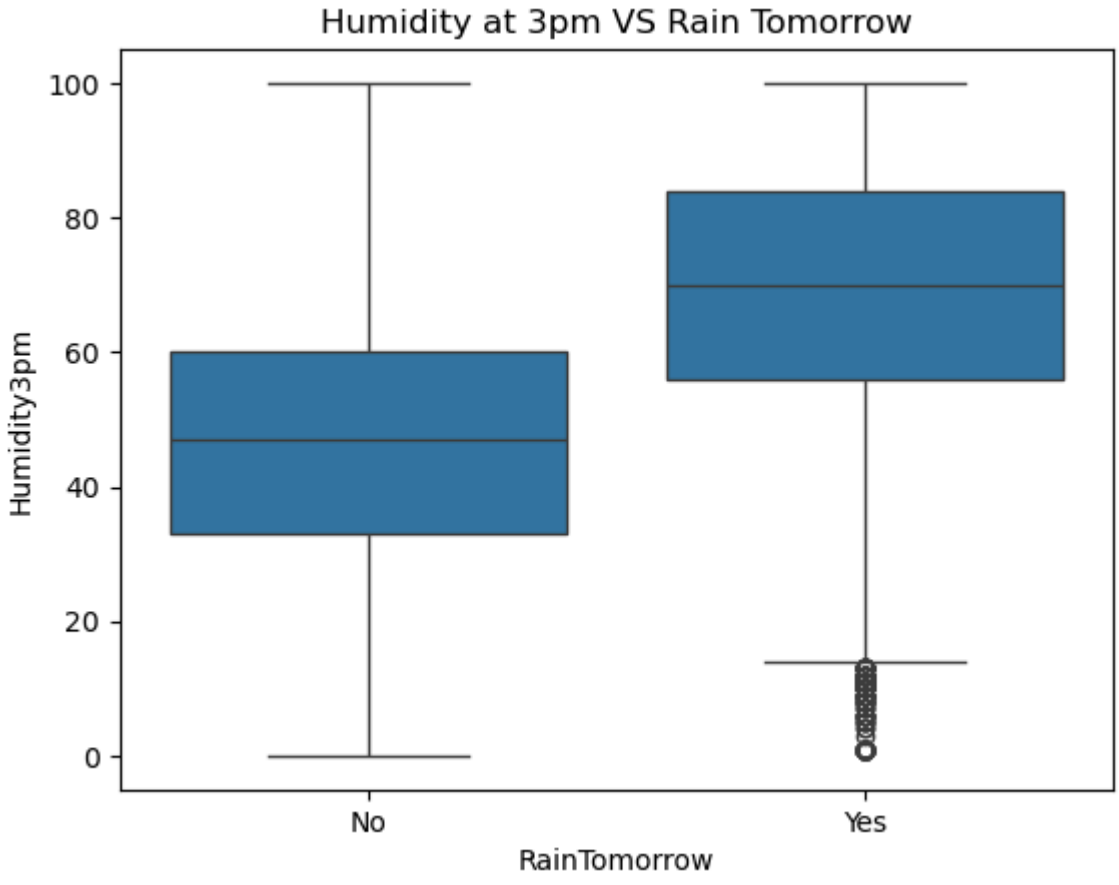
In [20]:
```
sns.countplot(x='RainTomorrow',data=df)
plt.title("Rain Tomorrow Distribution")
plt.show()
```



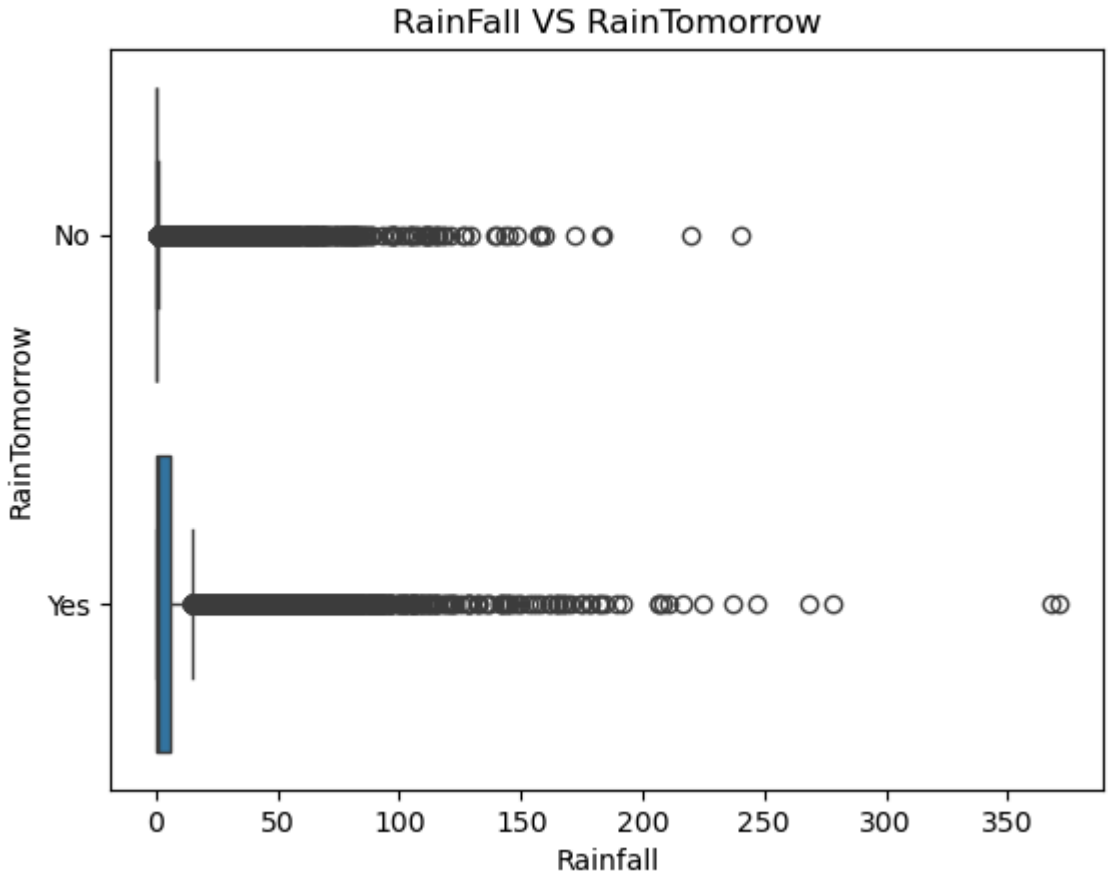
In [21]:
```
numeric_df = df.select_dtypes(include='number')
plt.figure(figsize=(12,8))
sns.heatmap(numeric_df.corr(),annot=True,cmap="coolwarm")
plt.title("Feature Correlation with RainTomorrow")
plt.show()
```



In [22]:
```
sns.boxplot(x='RainTomorrow',y='Humidity3pm',data=df)
plt.title("Humidity at 3pm VS Rain Tomorrow")
plt.show()
```

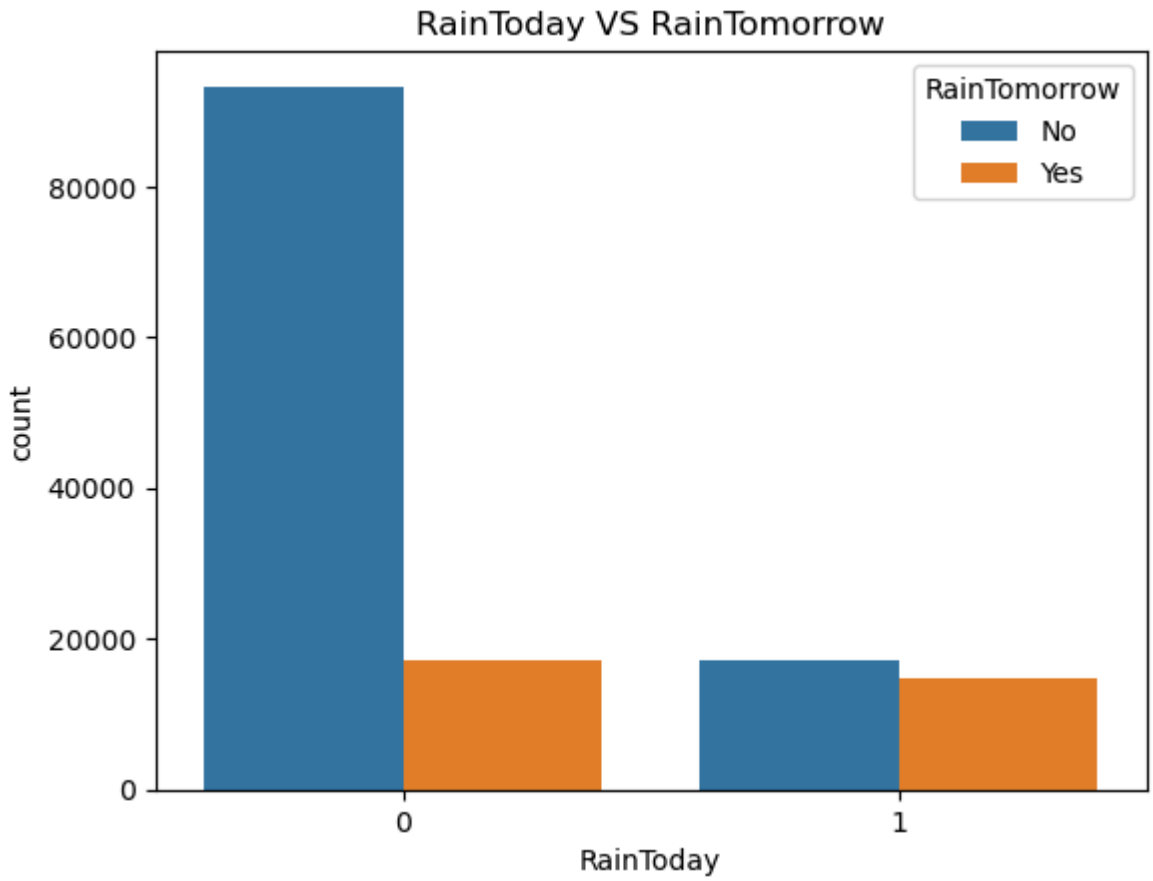


In [23]:
```
sns.boxplot(x='Rainfall',y='RainTomorrow', data=df)
plt.title("RainFall VS RainTomorrow")
plt.show()
```



In [24]:
```
sns.countplot(x='RainToday',hue='RainTomorrow',data=df)
plt.title("RainToday VS RainTomorrow")
plt.show
```

Out[24]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



In [25]:
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score,classification_report
```

In [28]:
```
selected_features=['MinTemp','MaxTemp','Rainfall','Humidity3pm','Pressure9am','R
X=df[selected_features]
y =df['RainTomorrow']
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=
```

In [30]:
```
print("Train Shape:",X_train)
print("Test Shape:",X_test)
```

```
Train Shape:         MinTemp  MaxTemp  Rainfall  Humidity3pm  Pressure9am  RainTo
day
18401       15.1     23.9       0.0         68.0       1001.9           0
127797       9.7     14.2       7.6         56.0       1008.2           1
40012       13.2     25.4       0.0         63.0       1025.2           0
130914       7.6     14.8       0.0         45.0       1004.6           0
41742       12.9     22.2       0.0         52.0       1023.0           0
...          ...      ...       ...          ...          ...         ...
112920       9.0     18.9      12.6          1.0       1017.5           1
122810      13.1     19.9       3.0         55.0       1018.1           1
106280      10.8     25.5       0.0         21.0       1015.7           0
135107      11.3     21.4       0.0         79.0       1014.5           0
124925      10.3     28.1       0.0         28.0       1028.6           0

[113754 rows x 6 columns]
Test Shape:         MinTemp  MaxTemp  Rainfall  Humidity3pm  Pressure9am  RainTod
ay
57760        7.1     13.0       8.8         98.0       1001.7           1
127128      13.2     18.3       0.0         73.0       1027.6           0
119994       9.2     22.7       0.0         25.0       1030.1           0
7088        15.3     26.1       0.0         40.0       1013.2           0
62992       11.9     31.8       0.0         25.0       1006.7           0
...          ...      ...       ...          ...          ...         ...
59458        2.7      9.7       0.0         95.0       1017.9           0
73138        5.0     22.9       3.4         32.0       1019.9           1
35876       11.2     26.1       0.0         38.0       1028.3           0
59558        8.3     20.8       0.0         36.0       1015.1           0
44133        8.9     18.6       0.0         52.0       1023.5           0

[28439 rows x 6 columns]
```

In [32]:
```
from sklearn.preprocessing import StandardScaler
```

In [35]:
```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

In [37]:
```
model=RandomForestClassifier(n_estimators=100,random_state=42)
model.fit(X_train,y_train)
```

Out[37]:
```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

In [38]:
```
y_pred=model.predict(X_test_scaled)
```

```
C:\Users\karan\anaconda3\Lib\site-packages\sklearn\base.py:493: UserWarning: X do
es not have valid feature names, but RandomForestClassifier was fitted with featu
re names
  warnings.warn(
```

In [40]:
```
accuracy= accuracy_score(y_test,y_pred)
print("Accuracy:",accuracy)
```

```
Accuracy: 0.22205422131579872
```

In [41]:
```
print(df['RainTomorrow'].value_counts(normalize=True) * 100)
```

```
RainTomorrow
No     77.581878
Yes    22.418122
Name: proportion, dtype: float64
```

In [42]:
```
unique,counts=np.unique(y_pred,return_counts=True)
print(dict(zip(unique,counts)))
```

```
{'No': 386, 'Yes': 28053}
```

In [ ]:
```

```