Data analysis of Weather Dataset with python.

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

In [12]:
```
data = pd.read_csv(r'C:\Users\harka\Datasets\weatherHistory.csv')
data.head()
```

Out[12]:

	Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibili (km
0	2006-04-01 00:00:00.000 +0200	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251.0	15.826
1	2006-04-01 01:00:00.000 +0200	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259.0	15.826
2	2006-04-01 02:00:00.000 +0200	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204.0	14.956
3	2006-04-01 03:00:00.000 +0200	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269.0	15.826
4	2006-04-01 04:00:00.000 +0200	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259.0	15.826

In [25]:
```
data.isnull().sum()
```

Out[25]:
```
Summary                       0
Precip Type                 517
Temperature (C)               0
Apparent Temperature (C)      0
Humidity                      0
Wind Speed (km/h)             0
Wind Bearing (degrees)        0
Visibility (km)               0
Loud Cover                    0
Pressure (millibars)          0
Daily Summary                 0
dtype: int64
```

In [16]:
```
data.dtypes
```

Out[16]:
```
Formatted Date               object
Summary                      object
Precip Type                  object
Temperature (C)             float64
Apparent Temperature (C)    float64
Humidity                    float64
Wind Speed (km/h)           float64
```

```
Wind Bearing (degrees)      float64
Visibility (km)             float64
Loud Cover                  float64
Pressure (millibars)        float64
Daily Summary                object
dtype: object
```

In [17]:
```
data.columns
```

Out[17]:
```
Index(['Formatted Date', 'Summary', 'Precip Type', 'Temperature (C)',
       'Apparent Temperature (C)', 'Humidity', 'Wind Speed (km/h)',
       'Wind Bearing (degrees)', 'Visibility (km)', 'Loud Cover',
       'Pressure (millibars)', 'Daily Summary'],
      dtype='object')
```

In [18]:
```
#Changing tha data type of 'Formatted Date' from to datetime
data['Formatted Date'] = pd.to_datetime(data['Formatted Date'], utc=True)
data['Formatted Date']
```

Out[18]:
```
0       2006-03-31 22:00:00+00:00
1       2006-03-31 23:00:00+00:00
2       2006-04-01 00:00:00+00:00
3       2006-04-01 01:00:00+00:00
4       2006-04-01 02:00:00+00:00
                   ...
96448   2016-09-09 17:00:00+00:00
96449   2016-09-09 18:00:00+00:00
96450   2016-09-09 19:00:00+00:00
96451   2016-09-09 20:00:00+00:00
96452   2016-09-09 21:00:00+00:00
Name: Formatted Date, Length: 96453, dtype: datetime64[ns, UTC]
```

In [20]:
```
data.describe()
```

Out[20]:

	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibility (km)	Loud Cover
count	96453.000000	96453.000000	96453.000000	96453.000000	96453.000000	96453.000000	96453.0
mean	11.932678	10.855029	0.734899	10.810640	187.509232	10.347325	0.0
std	9.551546	10.696847	0.195473	6.913571	107.383428	4.192123	0.0
min	-21.822222	-27.716667	0.000000	0.000000	0.000000	0.000000	0.0
25%	4.688889	2.311111	0.600000	5.828200	116.000000	8.339800	0.0
50%	12.000000	12.000000	0.780000	9.965900	180.000000	10.046400	0.0
75%	18.838889	18.838889	0.890000	14.135800	290.000000	14.812000	0.0
max	39.905556	39.344444	1.000000	63.852600	359.000000	16.100000	0.0

In [21]:
```
#setting the 'Formatted Date' as index
data = data.set_index('Formatted Date')
data.head()
```

Out[21]:

Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibili (kr

Formatted Date	Summary	Precip Type	Temperature (C)	Apparent Temperature (C)	Humidity	Wind Speed (km/h)	Wind Bearing (degrees)	Visibili (kr
2006-03-31 22:00:00+00:00	Partly Cloudy	rain	9.472222	7.388889	0.89	14.1197	251.0	15.82
2006-03-31 23:00:00+00:00	Partly Cloudy	rain	9.355556	7.227778	0.86	14.2646	259.0	15.82
2006-04-01 00:00:00+00:00	Mostly Cloudy	rain	9.377778	9.377778	0.89	3.9284	204.0	14.95
2006-04-01 01:00:00+00:00	Partly Cloudy	rain	8.288889	5.944444	0.83	14.1036	269.0	15.82
2006-04-01 02:00:00+00:00	Mostly Cloudy	rain	8.755556	6.977778	0.83	11.0446	259.0	15.82

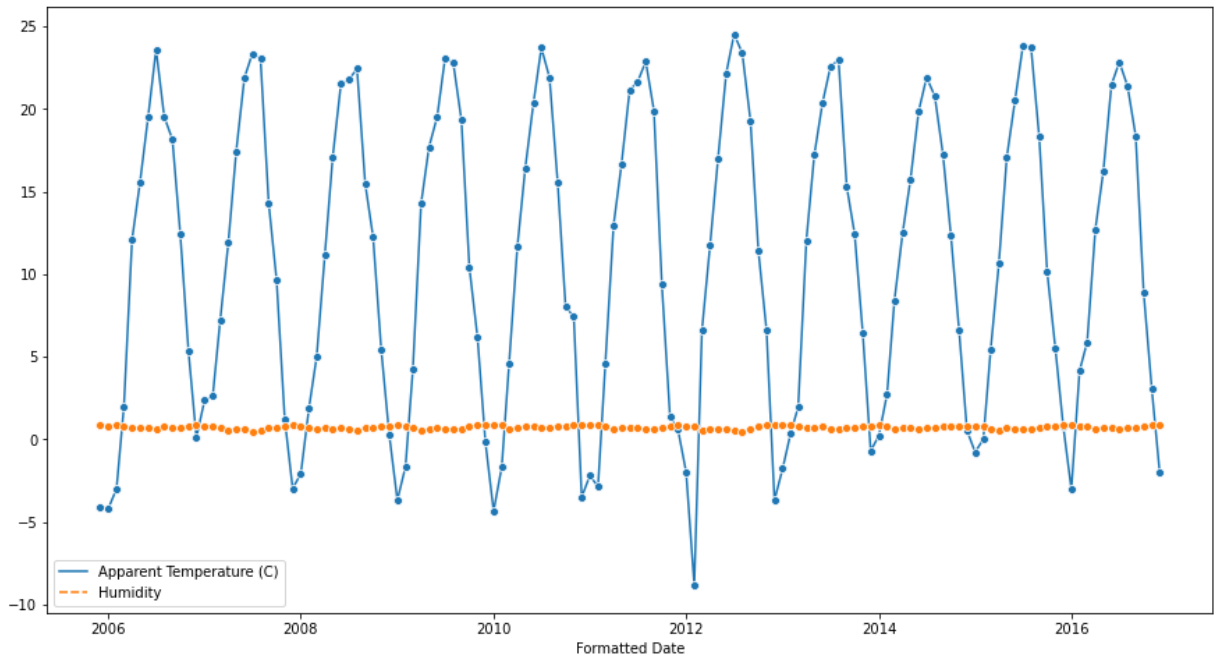
In [35]:
```
#Coverting the hourly data to monthly data and keeping only the 'Apparent Temperatur
data_columns = ['Apparent Temperature (C)', 'Humidity']data_monthly = data[data_colu
data_monthly.tail()
```

Out[35]:

Formatted Date	Apparent Temperature (C)	Humidity
2016-08-01 00:00:00+00:00	21.383094	0.674046
2016-09-01 00:00:00+00:00	18.355833	0.688833
2016-10-01 00:00:00+00:00	8.923947	0.799906
2016-11-01 00:00:00+00:00	3.048627	0.848472
2016-12-01 00:00:00+00:00	-2.017272	0.887981

In [ ]:
```
data_montly['Apparent Temperature (C)']
```

In [33]:
```
plt.figure(figsize=(15,8))
sns.lineplot(data=data_monthly, marker='o')
plt.show()
```



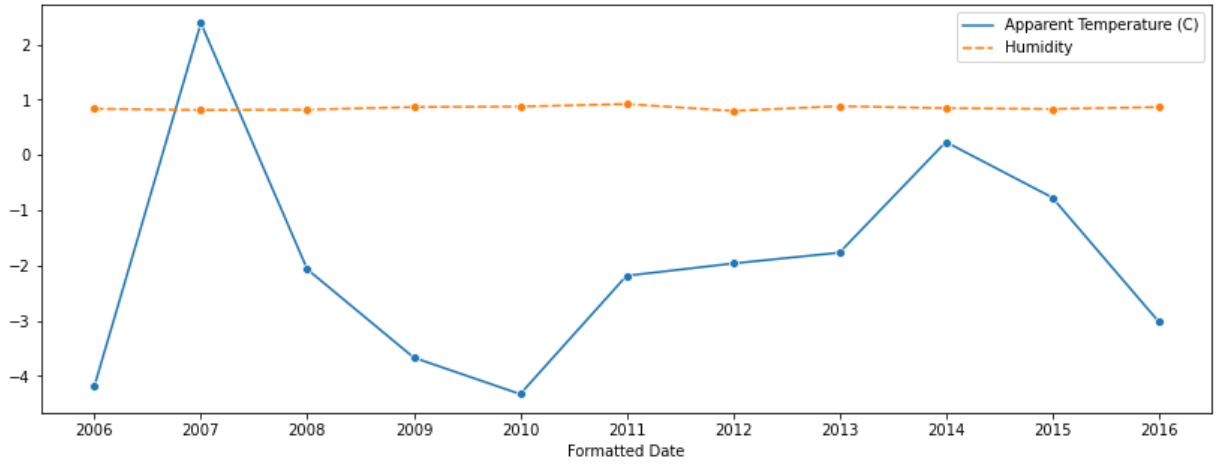
Obsevation: From the graph we we can the humidity 10 years is constant and apparent temperature is also same for all years except for 2012. In 2012 apparent temperature falls down a little bit.

Now Visualizing Graphically for each month.

In [37]:
```
#monthly analysis of month January
df1 = data_monthly[data_monthly.index.month==1]
print(df1)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-01-01 00:00:00+00:00                 -4.173708  0.834610
2007-01-01 00:00:00+00:00                  2.387015  0.813495
2008-01-01 00:00:00+00:00                 -2.069907  0.819476
2009-01-01 00:00:00+00:00                 -3.669937  0.867621
2010-01-01 00:00:00+00:00                 -4.329062  0.875914
2011-01-01 00:00:00+00:00                 -2.186813  0.922030
2012-01-01 00:00:00+00:00                 -1.965211  0.797581
2013-01-01 00:00:00+00:00                 -1.768578  0.883105
2014-01-01 00:00:00+00:00                  0.234536  0.846169
2015-01-01 00:00:00+00:00                 -0.770124  0.831519
2016-01-01 00:00:00+00:00                 -3.014576  0.866156
```

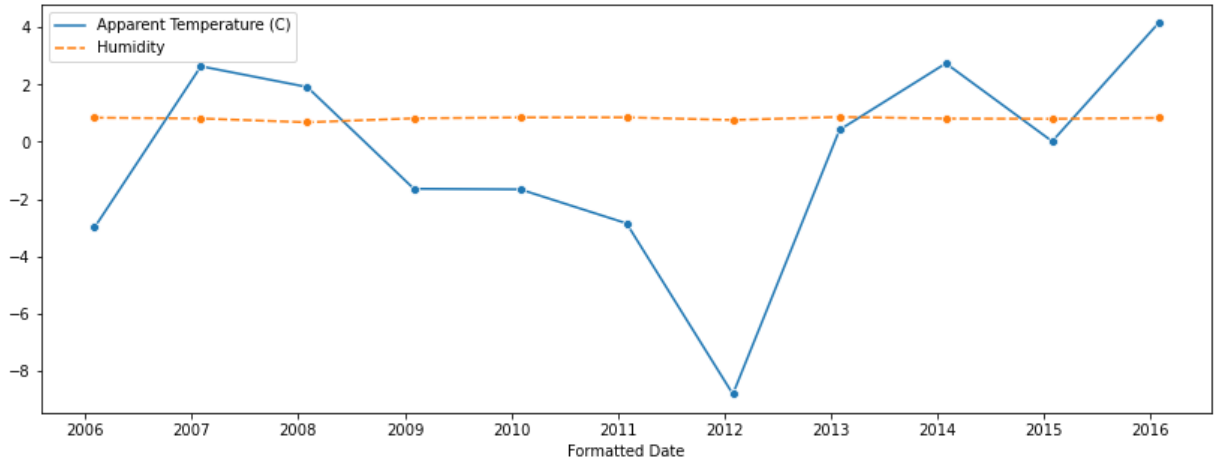
In [46]:
```
plt.figure(figsize=(14,5))
sns.lineplot(data=df1, marker='o')
plt.show()
```



In [41]:
```
#for feb
df2 = data_monthly[data_monthly.index.month==2]
print(df2)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-02-01 00:00:00+00:00                 -2.990716  0.843467
2007-02-01 00:00:00+00:00                  2.639848  0.815015
2008-02-01 00:00:00+00:00                  1.915597  0.682615
2009-02-01 00:00:00+00:00                 -1.641237  0.821161
2010-02-01 00:00:00+00:00                 -1.662045  0.851682
2011-02-01 00:00:00+00:00                 -2.849471  0.854137
2012-02-01 00:00:00+00:00                 -8.817241  0.762859
2013-02-01 00:00:00+00:00                  0.418171  0.869345
2014-02-01 00:00:00+00:00                  2.742998  0.812530
2015-02-01 00:00:00+00:00                  0.017006  0.803452
2016-02-01 00:00:00+00:00                  4.150782  0.836853
```

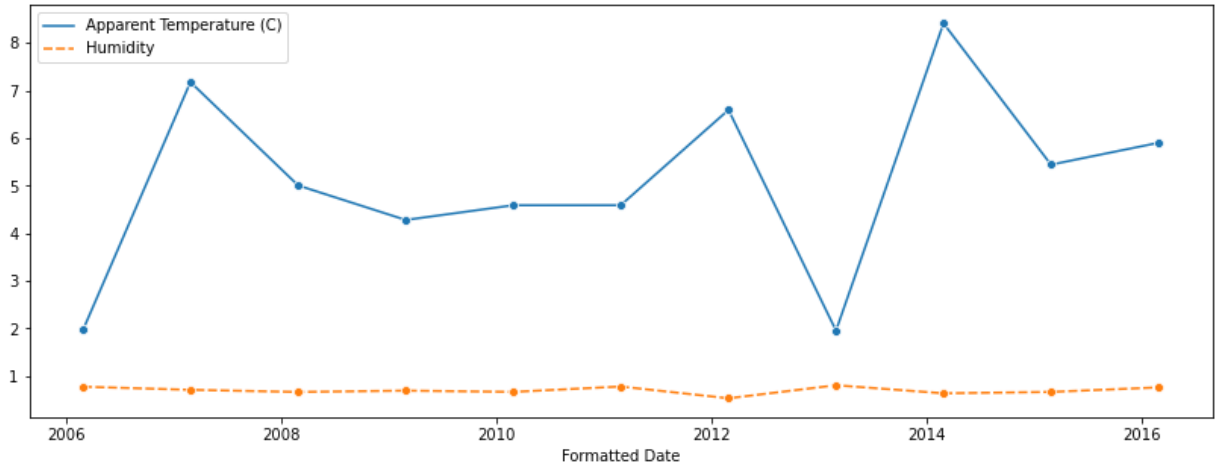
In [48]:
```
plt.figure(figsize=(14,5))
sns.lineplot(data=df2, marker='o')
plt.show()
```



In [50]:
```
#for march
df3 = data_monthly[data_monthly.index.month==3]
print(df3)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-03-01 00:00:00+00:00                  1.969780  0.778737
2007-03-01 00:00:00+00:00                  7.174619  0.713884
2008-03-01 00:00:00+00:00                  5.004353  0.668468
2009-03-01 00:00:00+00:00                  4.280585  0.696680
2010-03-01 00:00:00+00:00                  4.589038  0.670161
2011-03-01 00:00:00+00:00                  4.589785  0.782970
2012-03-01 00:00:00+00:00                  6.591502  0.535941
2013-03-01 00:00:00+00:00                  1.957445  0.809946
2014-03-01 00:00:00+00:00                  8.408303  0.640403
2015-03-01 00:00:00+00:00                  5.441592  0.669476
2016-03-01 00:00:00+00:00                  5.901404  0.764677
```

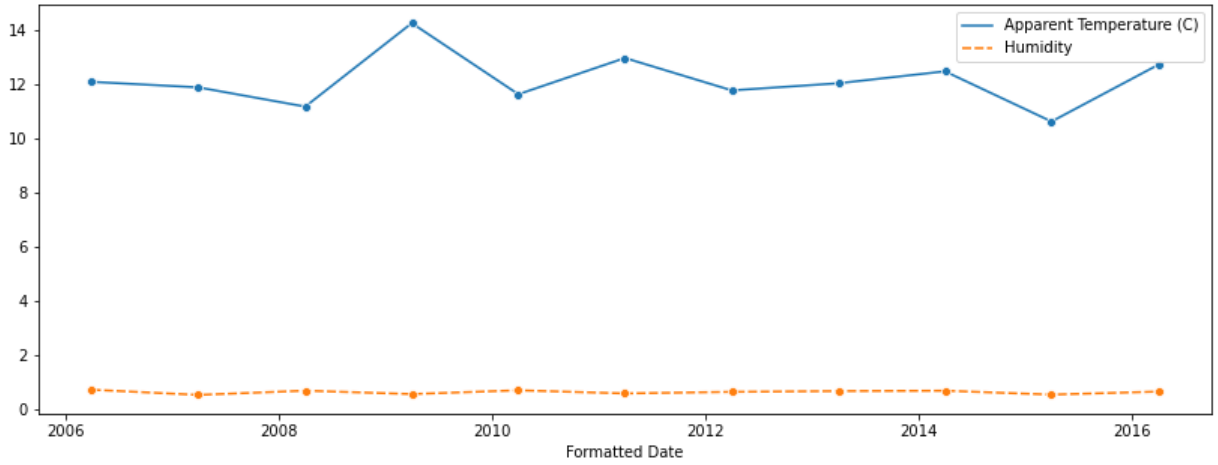
In [51]:
```
plt.figure(figsize=(14,5))
sns.lineplot(data=df3, marker='o')
plt.show()
```



In [59]:
```
#april
df4 = data_monthly[data_monthly.index.month==4]
print(df4)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-04-01 00:00:00+00:00                 12.098827  0.728625
2007-04-01 00:00:00+00:00                 11.894421  0.536361
2008-04-01 00:00:00+00:00                 11.183688  0.693194
2009-04-01 00:00:00+00:00                 14.267076  0.567847
2010-04-01 00:00:00+00:00                 11.639406  0.706875
2011-04-01 00:00:00+00:00                 12.978997  0.591625
2012-04-01 00:00:00+00:00                 11.782770  0.650222
2013-04-01 00:00:00+00:00                 12.045563  0.677667
2014-04-01 00:00:00+00:00                 12.486181  0.691403
2015-04-01 00:00:00+00:00                 10.632801  0.547764
2016-04-01 00:00:00+00:00                 12.731427  0.659972
```

In [60]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df4, marker='o')
plt.show()
```

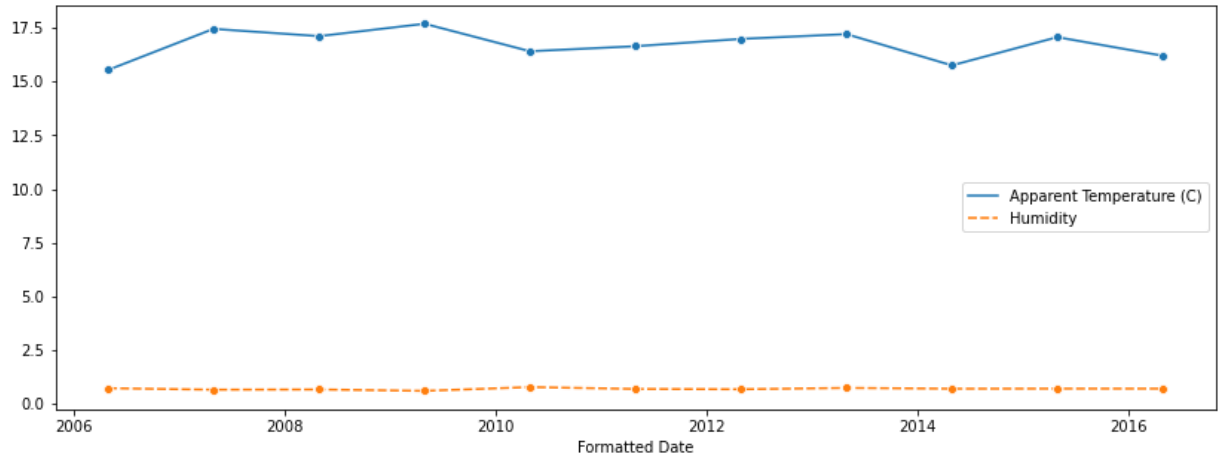


In [61]:
```
#may
df5 = data_monthly[data_monthly.index.month==5]
print(df5)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-05-01 00:00:00+00:00                 15.539479  0.721801
2007-05-01 00:00:00+00:00                 17.453136  0.653253
2008-05-01 00:00:00+00:00                 17.113583  0.663132
2009-05-01 00:00:00+00:00                 17.691256  0.597151
2010-05-01 00:00:00+00:00                 16.409879  0.773091
2011-05-01 00:00:00+00:00                 16.644922  0.688038
2012-05-01 00:00:00+00:00                 16.985596  0.672863
```

```
2013-05-01 00:00:00+00:00                17.208976  0.735309
2014-05-01 00:00:00+00:00                15.752218  0.698602
2015-05-01 00:00:00+00:00                17.067660  0.702742
2016-05-01 00:00:00+00:00                16.199216  0.702164
```

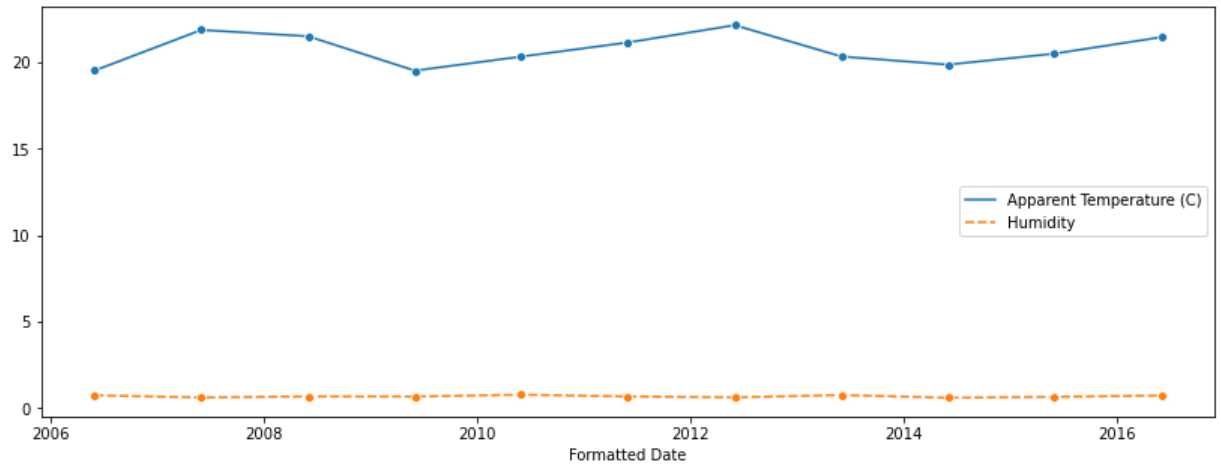
In [62]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df5, marker='o')
plt.show()
```



In [63]:
```
#june
df6 = data_monthly[data_monthly.index.month==6]
print(df6)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-06-01 00:00:00+00:00                 19.535965  0.747125
2007-06-01 00:00:00+00:00                 21.883102  0.616486
2008-06-01 00:00:00+00:00                 21.513750  0.679861
2009-06-01 00:00:00+00:00                 19.526790  0.675944
2010-06-01 00:00:00+00:00                 20.340571  0.778347
2011-06-01 00:00:00+00:00                 21.157114  0.677611
2012-06-01 00:00:00+00:00                 22.157130  0.622306
2013-06-01 00:00:00+00:00                 20.345664  0.761847
2014-06-01 00:00:00+00:00                 19.874306  0.602403
2015-06-01 00:00:00+00:00                 20.511782  0.655208
2016-06-01 00:00:00+00:00                 21.463387  0.733458
```

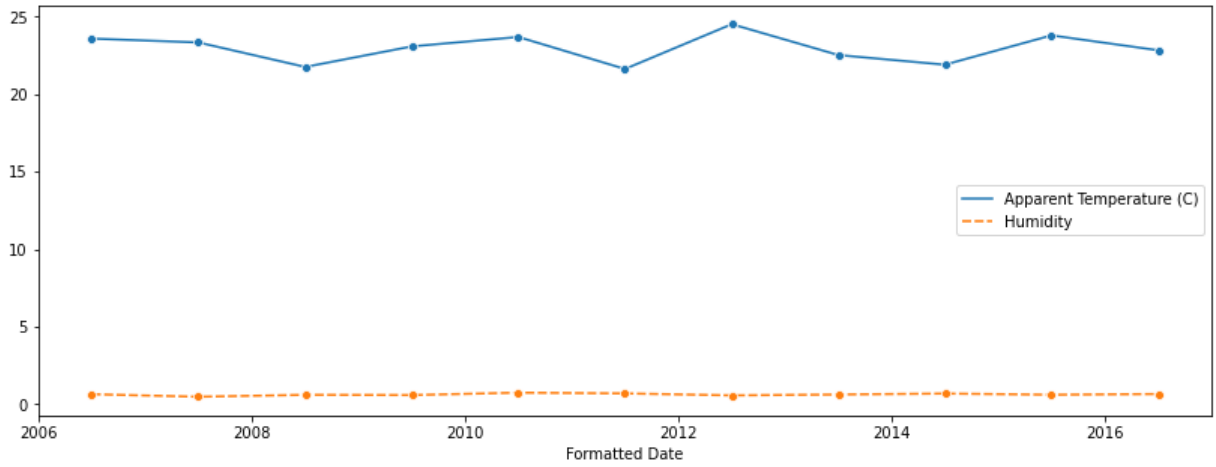
In [64]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df6, marker='o')
plt.show()
```



In [65]:
```
#july
df7 = data_monthly[data_monthly.index.month==7]
print(df7)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-07-01 00:00:00+00:00                 23.595348  0.656304
2007-07-01 00:00:00+00:00                 23.348081  0.491250
2008-07-01 00:00:00+00:00                 21.765562  0.616022
2009-07-01 00:00:00+00:00                 23.091614  0.600215
2010-07-01 00:00:00+00:00                 23.699447  0.755323
2011-07-01 00:00:00+00:00                 21.634984  0.707500
2012-07-01 00:00:00+00:00                 24.525343  0.580860
2013-07-01 00:00:00+00:00                 22.533669  0.636586
2014-07-01 00:00:00+00:00                 21.911598  0.699393
2015-07-01 00:00:00+00:00                 23.803487  0.622984
2016-07-01 00:00:00+00:00                 22.840226  0.669328
```

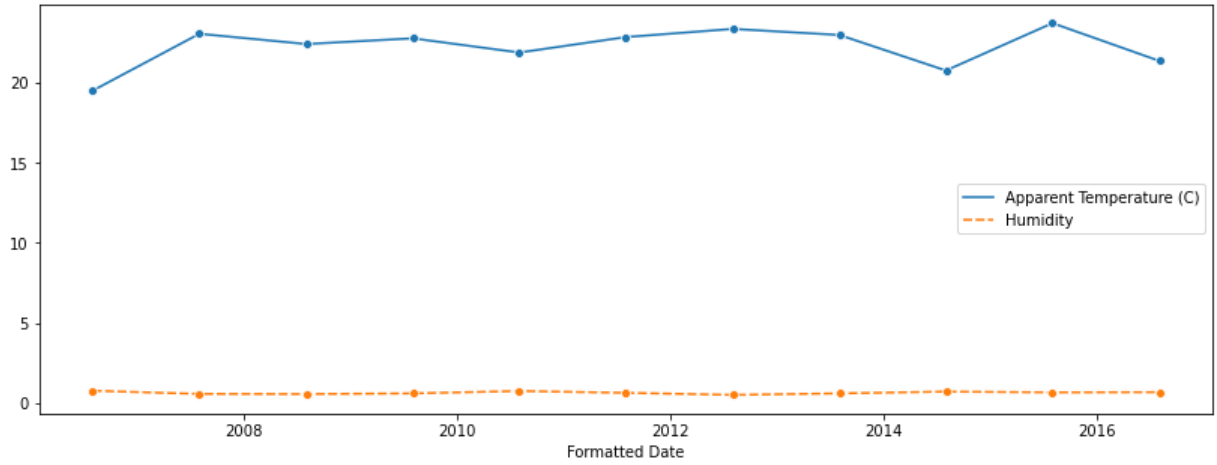
In [66]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df7, marker='o')
plt.show()
```



In [67]:
```
#august
df8 = data_monthly[data_monthly.index.month==8]
print(df8)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-08-01 00:00:00+00:00                 19.528241  0.760753
2007-08-01 00:00:00+00:00                 23.079689  0.562876
2008-08-01 00:00:00+00:00                 22.438852  0.551895
2009-08-01 00:00:00+00:00                 22.794205  0.597231
2010-08-01 00:00:00+00:00                 21.906713  0.742786
2011-08-01 00:00:00+00:00                 22.874126  0.631263
2012-08-01 00:00:00+00:00                 23.384334  0.500081
2013-08-01 00:00:00+00:00                 23.005249  0.596263
2014-08-01 00:00:00+00:00                 20.781870  0.707809
2015-08-01 00:00:00+00:00                 23.745766  0.659825
2016-08-01 00:00:00+00:00                 21.383094  0.674046
```

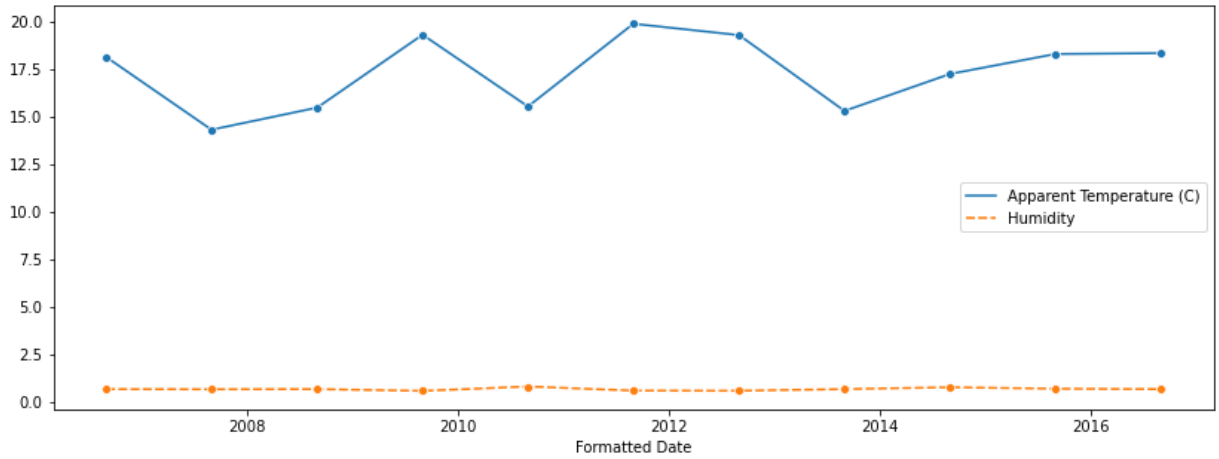
In [68]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df8, marker='o')
plt.show()
```



In [69]:
```
#september
df9 = data_monthly[data_monthly.index.month==9]
print(df9)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-09-01 00:00:00+00:00                 18.155571  0.689444
2007-09-01 00:00:00+00:00                 14.328457  0.682708
2008-09-01 00:00:00+00:00                 15.489606  0.690722
2009-09-01 00:00:00+00:00                 19.322353  0.596764
2010-09-01 00:00:00+00:00                 15.549414  0.826806
2011-09-01 00:00:00+00:00                 19.899900  0.611375
2012-09-01 00:00:00+00:00                 19.302948  0.603319
2013-09-01 00:00:00+00:00                 15.317477  0.691986
2014-09-01 00:00:00+00:00                 17.258387  0.785944
2015-09-01 00:00:00+00:00                 18.308472  0.712889
2016-09-01 00:00:00+00:00                 18.355833  0.688833
```

In [70]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df9, marker='o')
plt.show()
```

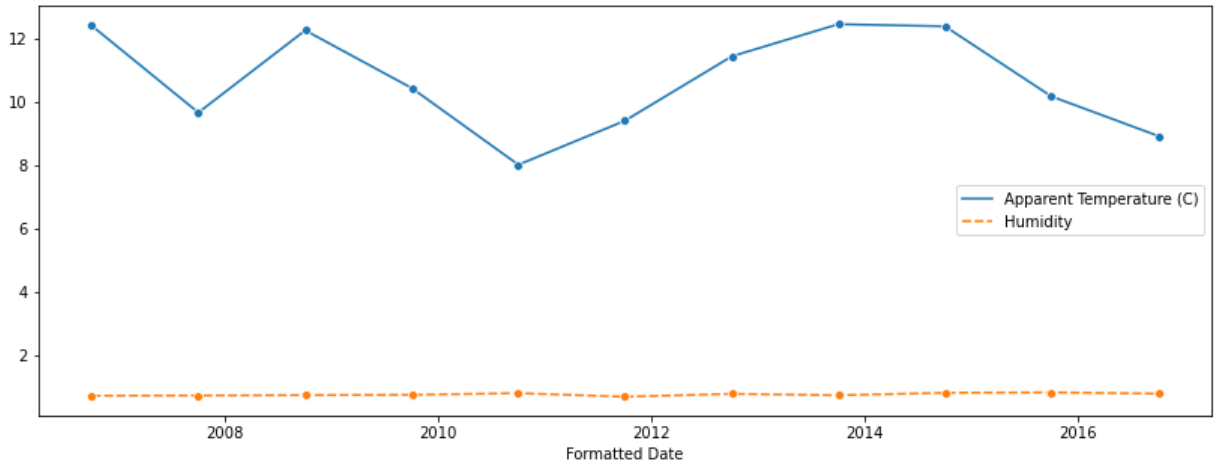


In [71]:
```
#october
df10 = data_monthly[data_monthly.index.month==10]
print(df10)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-10-01 00:00:00+00:00                 12.398678  0.733642
2007-10-01 00:00:00+00:00                  9.662612  0.740954
2008-10-01 00:00:00+00:00                 12.253390  0.753911
2009-10-01 00:00:00+00:00                 10.433535  0.763468
2010-10-01 00:00:00+00:00                  8.017145  0.815538
2011-10-01 00:00:00+00:00                  9.405167  0.701747
2012-10-01 00:00:00+00:00                 11.435581  0.794315
```

```
2013-10-01 00:00:00+00:00                12.449134  0.748750
2014-10-01 00:00:00+00:00                12.381803  0.826116
2015-10-01 00:00:00+00:00                10.170408  0.840524
2016-10-01 00:00:00+00:00                 8.923947  0.799906
```

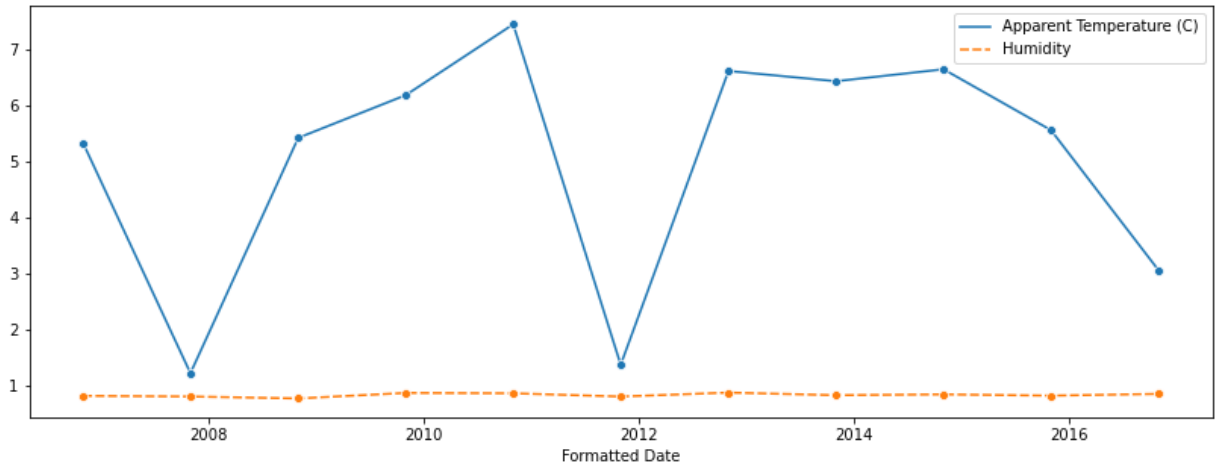
In [72]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df10, marker='o')
plt.show()
```



In [73]:
```
#november
df11 = data_monthly[data_monthly.index.month==11]
print(df11)
```

```
                           Apparent Temperature (C)  Humidity
Formatted Date
2006-11-01 00:00:00+00:00                  5.328310  0.812722
2007-11-01 00:00:00+00:00                  1.218225  0.801444
2008-11-01 00:00:00+00:00                  5.415039  0.766972
2009-11-01 00:00:00+00:00                  6.177222  0.865292
2010-11-01 00:00:00+00:00                  7.440934  0.858722
2011-11-01 00:00:00+00:00                  1.368519  0.800528
2012-11-01 00:00:00+00:00                  6.608133  0.871389
2013-11-01 00:00:00+00:00                  6.425664  0.824792
2014-11-01 00:00:00+00:00                  6.639097  0.839736
2015-11-01 00:00:00+00:00                  5.553040  0.817014
2016-11-01 00:00:00+00:00                  3.048627  0.848472
```

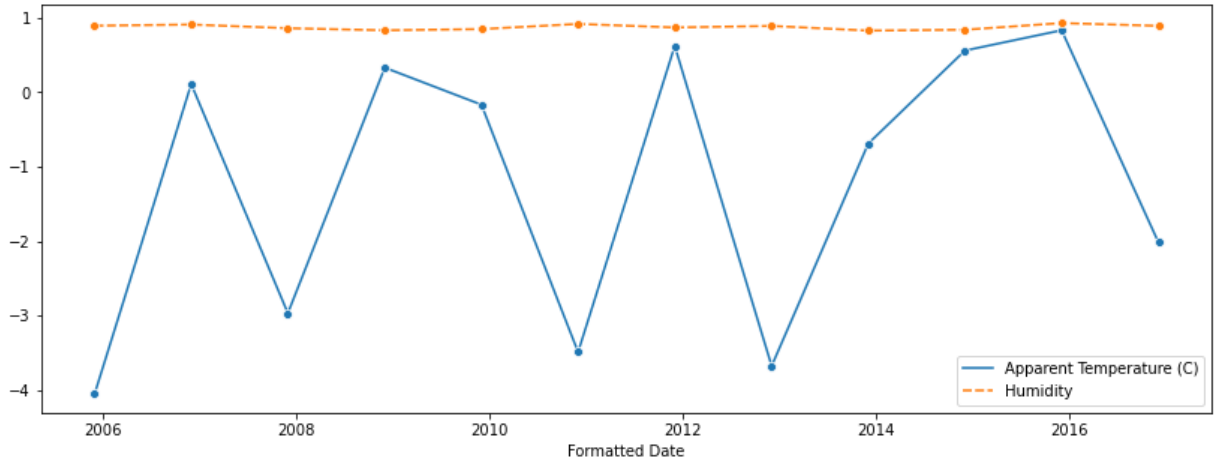
In [74]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df11, marker='o')
plt.show()
```



In [76]:
```
#december
df12 = data_monthly[data_monthly.index.month==12]
print(df12)
```

Formatted Date	Apparent Temperature (C)	Humidity
2005-12-01 00:00:00+00:00	-4.050000	0.890000
2006-12-01 00:00:00+00:00	0.107310	0.905376
2007-12-01 00:00:00+00:00	-2.964897	0.856250
2008-12-01 00:00:00+00:00	0.327389	0.828226
2009-12-01 00:00:00+00:00	-0.169086	0.844637
2010-12-01 00:00:00+00:00	-3.485947	0.913602
2011-12-01 00:00:00+00:00	0.618093	0.866223
2012-12-01 00:00:00+00:00	-3.672909	0.886801
2013-12-01 00:00:00+00:00	-0.690054	0.823965
2014-12-01 00:00:00+00:00	0.556586	0.835927
2015-12-01 00:00:00+00:00	0.828644	0.925390
2016-12-01 00:00:00+00:00	-2.017272	0.887981

In [77]:
```
plt.figure(figsize=(14, 5))
sns.lineplot(data=df12, marker='o')
plt.show()
```



Observation: We can see in the above graphs that humidity was constant in each month but temperature varied little bit.