Issues:

1. Features selection (after 3rd iteration, all p-values are zero, so dead end)
2. Addressing negative skew in feature: The transformation flipped skew from -ve to +ve, so not using transformation method.

1. Features selection (using p-value):

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn import metrics
import warnings
```

1st iteration: Drop least important feature (p-value of 'Temp9am': 0.995)

```
x = sm.add_constant(rain_imp)
model = sm.OLS(target,x)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.335
Model:                            OLS   Adj. R-squared:                  0.335
Method:                 Least Squares   F-statistic:                     4181.
Date:                Wed, 30 Nov 2022   Prob (F-statistic):               0.00
Time:                        07:15:37   Log-Likelihood:                -47293.
No. Observations:              140787   AIC:                         9.462e+04
Df Residuals:                  140769   BIC:                         9.480e+04
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.5272      0.181     47.137      0.000       8.173       8.882
MinTemp       -0.0060      0.000    -13.201      0.000      -0.007      -0.005
MaxTemp        0.0087      0.001      9.812      0.000       0.007       0.010
Rainfall       0.0016      0.000     12.366      0.000       0.001       0.002
Evaporation    0.0019      0.000      5.051      0.000       0.001       0.003
Sunshine      -0.0275      0.001    -44.349      0.000      -0.029      -0.026
WindGustSpeed  0.0073      0.000     62.126      0.000       0.007       0.007
WindSpeed9am  -0.0007      0.000     -5.234      0.000      -0.001      -0.000
WindSpeed3pm  -0.0043      0.000    -27.785      0.000      -0.005      -0.004
Humidity9am   -0.0006   9.97e-05     -5.775      0.000      -0.001      -0.000
Humidity3pm    0.0075      0.000     65.814      0.000       0.007       0.008
Pressure9am    0.0180      0.001     29.731      0.000       0.017       0.019
Pressure3pm   -0.0266      0.001    -44.061      0.000      -0.028      -0.025
Cloud9am      -0.0077      0.001    -12.650      0.000      -0.009      -0.006
Cloud3pm       0.0036      0.001      5.349      0.000       0.002       0.005
Temp9am    -3.939e-06      0.001     -0.006      0.995      -0.001       0.001
Temp3pm       -0.0018      0.001     -1.865      0.062      -0.004    9.17e-05
RainToday      0.1077      0.003     38.549      0.000       0.102       0.113
==============================================================================
Omnibus:                    12905.385   Durbin-Watson:                   1.904
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            16747.764
Skew:                           0.815   Prob(JB):                         0.00
Kurtosis:                       3.444   Cond. No.                     2.89e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.89e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
x.drop('Temp9am',axis=1, inplace=True)
```

2nd iteration: Drop least important feature (p-value of 'Temp3pm': 0.056)

```
model = sm.OLS(target,x)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.335
Model:                            OLS   Adj. R-squared:                  0.335
Method:                 Least Squares   F-statistic:                     4442.
Date:                Wed, 30 Nov 2022   Prob (F-statistic):               0.00
Time:                        07:15:39   Log-Likelihood:                -47293.
No. Observations:              140787   AIC:                         9.462e+04
Df Residuals:                  140770   BIC:                         9.479e+04
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.5271      0.181     47.221      0.000       8.173       8.881
MinTemp       -0.0060      0.000    -17.095      0.000      -0.007      -0.005
MaxTemp        0.0087      0.001      9.964      0.000       0.007       0.010
Rainfall       0.0016      0.000     12.370      0.000       0.001       0.002
Evaporation    0.0019      0.000      5.078      0.000       0.001       0.003
Sunshine      -0.0275      0.001    -44.510      0.000      -0.029      -0.026
WindGustSpeed  0.0073      0.000     62.348      0.000       0.007       0.007
WindSpeed9am  -0.0007      0.000     -5.256      0.000      -0.001      -0.000
WindSpeed3pm  -0.0043      0.000    -28.223      0.000      -0.005      -0.004
Humidity9am   -0.0006   8.03e-05     -7.170      0.000      -0.001      -0.000
Humidity3pm    0.0075   9.97e-05     75.150      0.000       0.007       0.008
Pressure9am    0.0180      0.001     29.961      0.000       0.017       0.019
Pressure3pm   -0.0266      0.001    -44.318      0.000      -0.028      -0.025
Cloud9am      -0.0077      0.001    -12.706      0.000      -0.009      -0.006
Cloud3pm       0.0036      0.001      5.364      0.000       0.002       0.005
Temp3pm       -0.0018      0.001     -1.910      0.056      -0.004    4.71e-05
RainToday      0.1077      0.003     38.555      0.000       0.102       0.113
==============================================================================
Omnibus:                    12905.360   Durbin-Watson:                   1.904
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            16747.722
Skew:                           0.815   Prob(JB):                         0.00
Kurtosis:                       3.444   Cond. No.                     2.88e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.88e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
x.drop('Temp3pm',axis=1, inplace=True)
```

3rd Iteration: Now, all the p-values are resulting in zero, which is a dead end. So not pursuing this method for features selection.

```
model = sm.OLS(target,x)
results = model.fit()
print(results.summary())
```

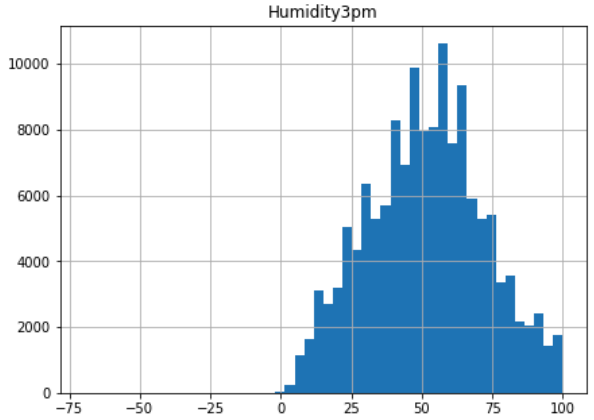
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.335
Model:                            OLS   Adj. R-squared:                  0.335
Method:                 Least Squares   F-statistic:                     4738.
Date:                Wed, 30 Nov 2022   Prob (F-statistic):               0.00
Time:                        07:15:41   Log-Likelihood:                -47295.
No. Observations:              140787   AIC:                         9.462e+04
Df Residuals:                  140771   BIC:                         9.478e+04
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.5389      0.180     47.313      0.000       8.185       8.893
MinTemp       -0.0063      0.000    -18.809      0.000      -0.007      -0.006
MaxTemp        0.0072      0.000     19.576      0.000       0.006       0.008
Rainfall       0.0015      0.000     12.272      0.000       0.001       0.002
Evaporation    0.0020      0.000      5.253      0.000       0.001       0.003
Sunshine      -0.0275      0.001    -44.506      0.000      -0.029      -0.026
WindGustSpeed  0.0073      0.000     63.179      0.000       0.007       0.008
WindSpeed9am  -0.0008      0.000     -5.374      0.000      -0.001      -0.000
WindSpeed3pm  -0.0043      0.000    -28.230      0.000      -0.005      -0.004
Humidity9am   -0.0006   7.79e-05     -7.872      0.000      -0.001      -0.000
Humidity3pm    0.0076    8.7e-05     87.181      0.000       0.007       0.008
Pressure9am    0.0177      0.001     30.441      0.000       0.017       0.019
Pressure3pm   -0.0264      0.001    -45.122      0.000      -0.027      -0.025
Cloud9am      -0.0077      0.001    -12.739      0.000      -0.009      -0.006
Cloud3pm       0.0037      0.001      5.578      0.000       0.002       0.005
RainToday      0.1075      0.003     38.510      0.000       0.102       0.113
==============================================================================
Omnibus:                    12892.160   Durbin-Watson:                   1.904
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            16725.934
Skew:                           0.815   Prob(JB):                         0.00
Kurtosis:                       3.443   Cond. No.                     2.88e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.88e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

2. Data transformation for positive/negative skew:

Column: 'Humidity3pm' (negative skew)

```
[156] rain_imp_feat.hist(column=['Humidity3pm'],bins=50, figsize=(7,5))
      print('Min:',round((rain_imp_feat['Humidity3pm']).min(),3))
      print('Max:',round((rain_imp_feat['Humidity3pm']).max(),3))
```

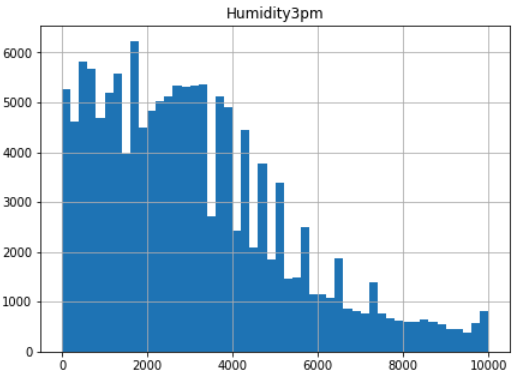
```
Min: -69.825
Max: 100.0
```



Operation: Square transformation

```
[160] # # Square transformation to remove negative skewness
      for i in tqdm(range(len(rain_imp_feat))):
        rain_imp_feat.iloc[i]['Humidity3pm'] =  (rain_imp_feat.iloc[i]['Humidity3pm']) ** 2

      rain_imp_feat.hist(column=['Humidity3pm'],bins=50, figsize=(7,5))
      print('Min:',round((rain_imp_feat['Humidity3pm']).min(),3))
      print('Max:',round((rain_imp_feat['Humidity3pm']).max(),3))
      # rain_imp_feat.head(3)
```

```
100%|██████████| 140787/140787 [00:19<00:00, 7097.05it/s]
Min: 0.0
Max: 10000.0
```



Output of transformation: we have got a positive skew now. So, we are **skipping** this transformation.

```
[177] # After square transformation on variable 'Humidity3pm' our distribution changed from -ve skew to +ve
      #      I have documented the work(-ve skew transformation) in doc file, please refer for detail (file name: Issues in data pre-processing.docx)
      print('We are skipping this transformation because it flipped from -ve skew to +ve skew')
      print('See the documentation (file name): `Issues in data pre-processing.docx`')
```

```
We are skipping this transformation because it flipped from -ve skew to +ve skew
See the documentation (file name): `Issues in data pre-processing.docx`
```