Homework 2 - Linear Regression

Data set - NOAA GSOD

National Oceanic and Atmospheric Administration, NOAA, publishes Global Surface Summary of the Day, GSOD.[1] “GSOD is derived from The Integrated Surface Hourly (ISH) dataset. The ISH dataset includes global data obtained from the USAF Climatology Center, located in the Federal Climate Complex with NCDC. The latest daily summary data are normally available 1-2 days after the date-time of the observations used in the daily summaries.[1]” This dataset can be found at https://www.kaggle.com/noaa/gsod/data. The data included in the set starts in 1929 and continues through the current day. This dataset includes date, temperature, dew point, sea level pressure, and station number. Date is the date the samples were taken on including year, month, and day. Temperature is the mean temperature of the day in degrees Fahrenheit and includes the number of samples used to create the mean. Dew Point is the mean dew point or temperature at which water vapor will condense out the air, for the day in degrees Fahrenheit and includes the number of samples used to create the mean. Sea level pressure is the mean pressure for the day in millibars to the tenth and includes the number of samples used to create the mean. Station number is the recording station that is doing the monitoring this data includes station name, country, latitude, longitude, and elevation.

I took a large sample of the data for a recording station from January 1929 through August 2001. The data uses 9999, 999.9, and 9999.9 as null values so those needed to be accounted for in the data.

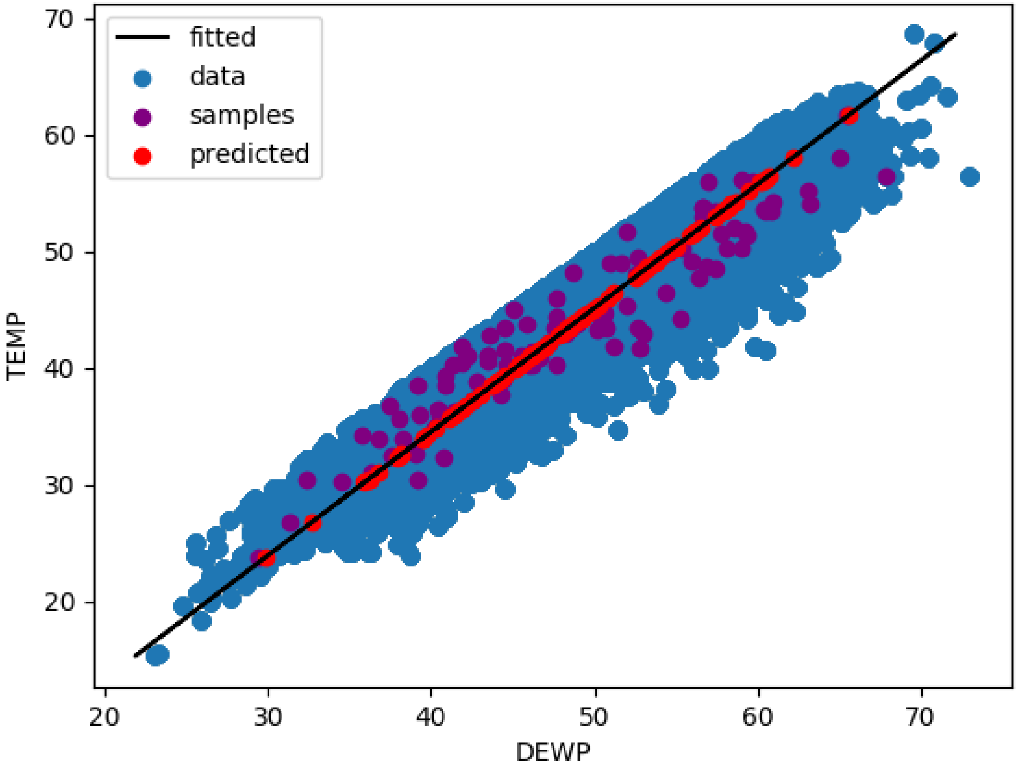
Part A

Variables

I chose do my analysis on the relationship between Temperature and Dew Point. In particular I used Dew Point as my Explanatory Variable and Temperature as my Outcome Variable.

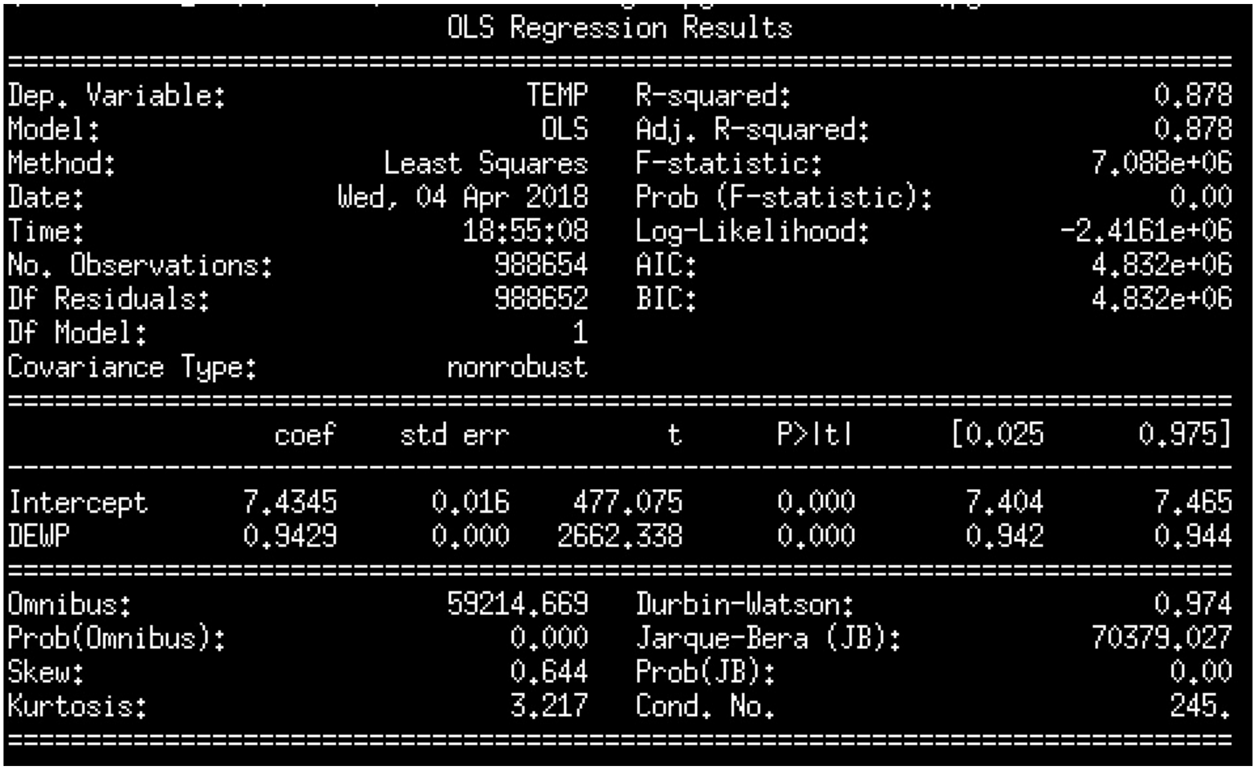
I randomly picked a recording station and used data from between January 1929 through August 2001. This dataset has 98,8654 rows and 16 columns.

Scatterplot



The linear model doesn't seem to fit the data very well. The scatter plot shows that the data is very spread out and varies widely, making it fairly obvious that there is a relationship between the variables, but it isn't clear that the relationship is a single linear relationship.

Linear Regression



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   TEMP   R-squared:                       0.878
Model:                            OLS   Adj. R-squared:                  0.878
Method:                 Least Squares   F-statistic:                 7.088e+06
Date:                Wed, 04 Apr 2018   Prob (F-statistic):               0.00
Time:                        18:55:08   Log-Likelihood:            -2.4161e+06
No. Observations:              988654   AIC:                         4.832e+06
Df Residuals:                  988652   BIC:                         4.832e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.4345      0.016    477.075      0.000       7.404       7.465
DEWP           0.9429      0.000   2662.338      0.000       0.942       0.944
==============================================================================
Omnibus:                    59214.669   Durbin-Watson:                   0.974
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            70379.027
Skew:                           0.644   Prob(JB):                         0.00
Kurtosis:                       3.217   Cond. No.                         245.
==============================================================================
```

Predicted Values

As can be seen from the sample of the sample set Using the Dew Point to predict the Temperature is not the best data point to use to predict the temperature.

TEMP	DEWP	Predicted TEMP
44.6	37.1	42.4174124
37.7	35.3	40.720113
44	38.5	43.7375343
62.7	57.6	61.7477676
43.3	42.4	47.4150165
60.8	49.6	54.2042144
46	42.8	47.7921941
52.5	50.1	54.6756864
53.1	48.9	53.5441535

50.2	49.3	53.9213311
53.6	51.1	55.6186306
48.8	44.5	49.3951992
36.4	35.2	40.6258185
50.9	46.2	50.9982042
43.5	40.1	45.2462449
38.2	31.2	36.8540419
59.8	58.8	62.8793006
51.1	46.9	51.6582651
60.4	56.4	60.6162346
42.3	39	44.2090063
50	43	47.9807829
41.5	35.8	41.191585
53.3	49.7	54.2985088
54.1	45.5	50.3381433
38.4	36.2	41.5687627
53.5	45.8	50.6210266
56.4	43.5	48.452255
63.8	53.6	57.975991
56.8	50.2	54.7699809
35.3	30.2	35.9110978

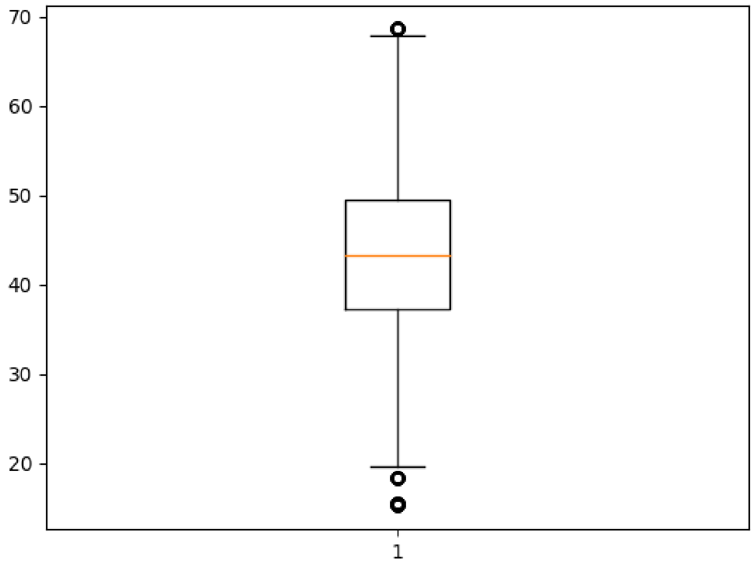
Root Mean Square

The Root Mean Square of the predicted to actual values is 3.251303848483339 which is a fairly large number indicating that the dew point is not a good predictor for temperature.

Part B

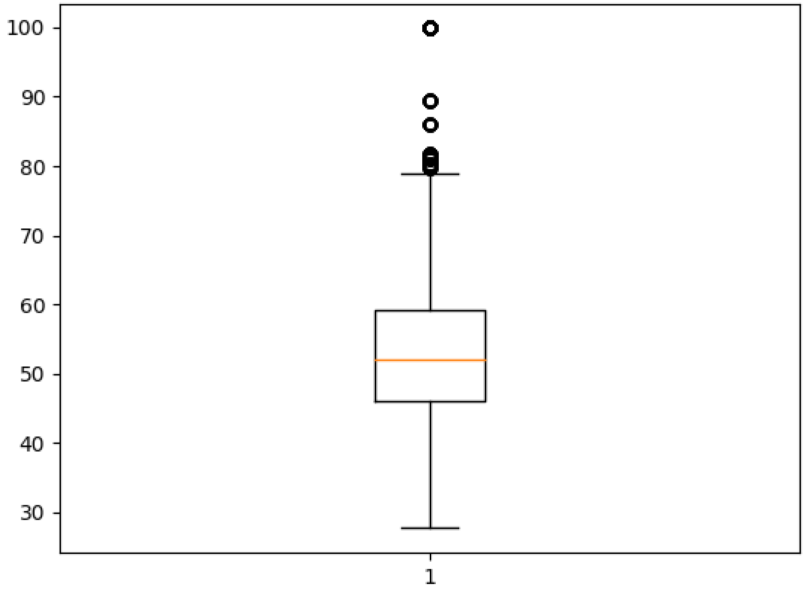
1. Box Plots

1.1. Dew Point



-
- This plot of Dew Point shows that most of the values for dew point are within the standard deviation.

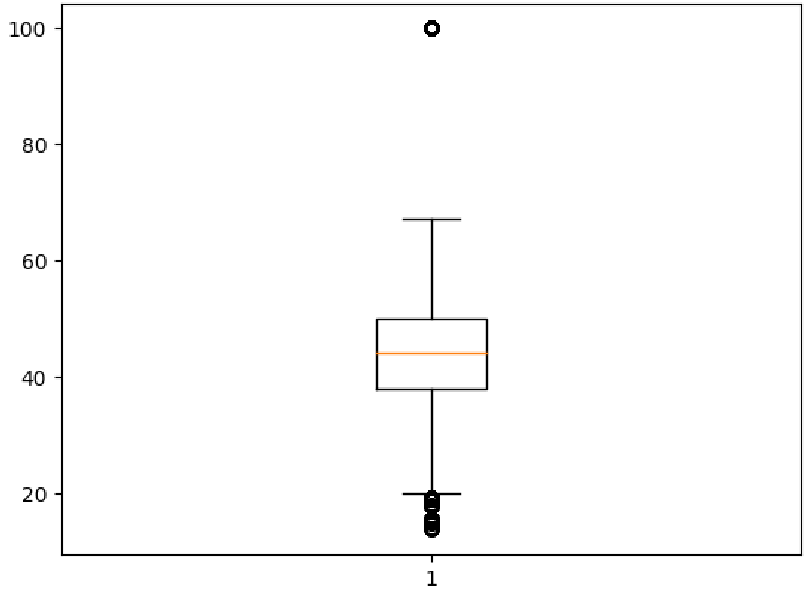
1.2. Max Temperature



-

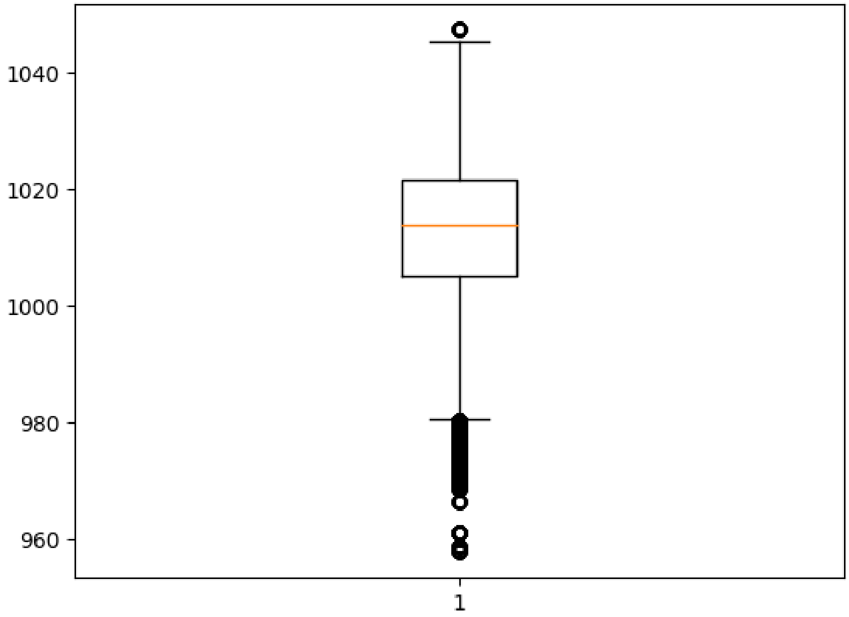
- This plot of maximum temperature shows that the maximum temperature for any given day tends to be within 1 standard deviation of 52° or so, with a few really hot days.

1.3. Minimum Temperature



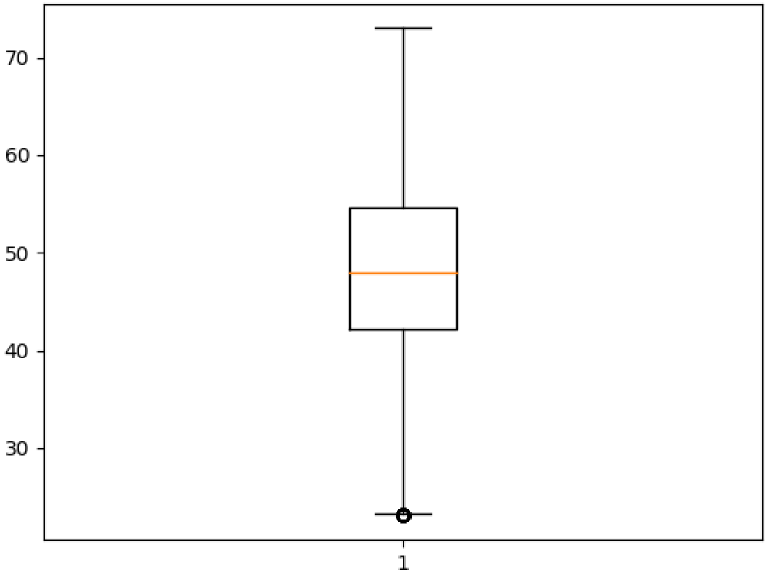
-
- This chart of minimum temperature shows that most of the time the minimum temperature hangs near 45°, with a few below 20°.

1.4. Sea Level Pressure



-
- Sea level pressure tends to mostly be around 1028 millibars, but a not insignificant number lie far below the 1st quantile.

1.5. Average Daily Temperature

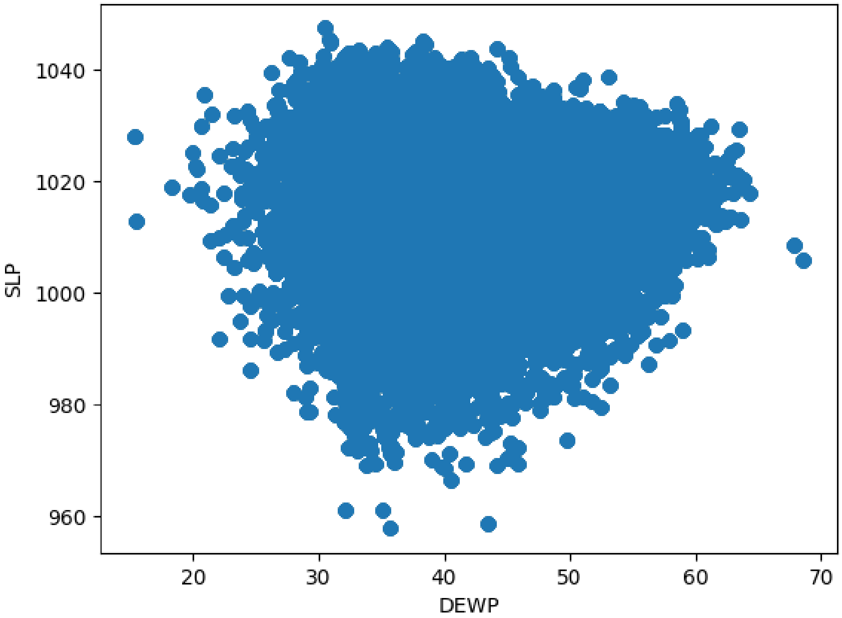


-
- The average daily temperature is most often between 40 and 55°, but has been known to get as high as 70° and as low as 25°. There are relatively few outliers. With a fairly cold average temperature

this is not someplace that I would want to spend a great deal of time.

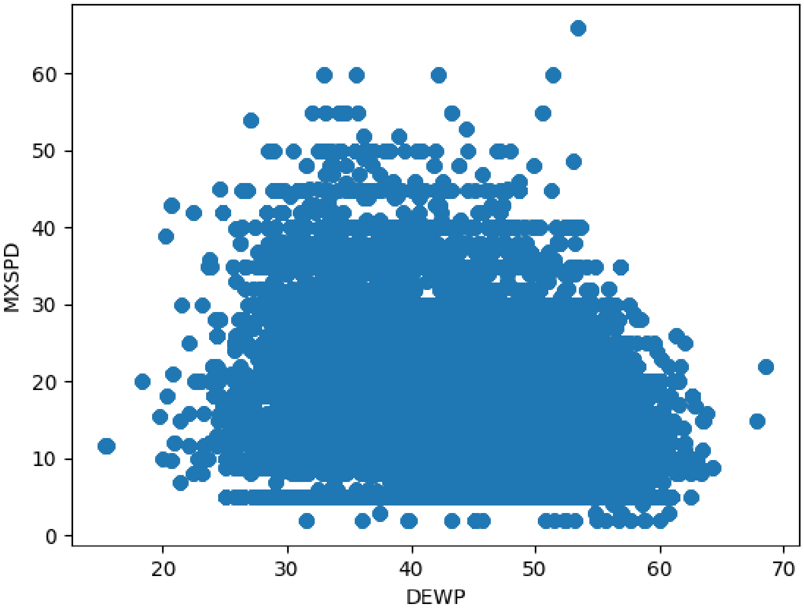
2. Scatter Plots

2.1. Dew Point to Sea Level Pressure



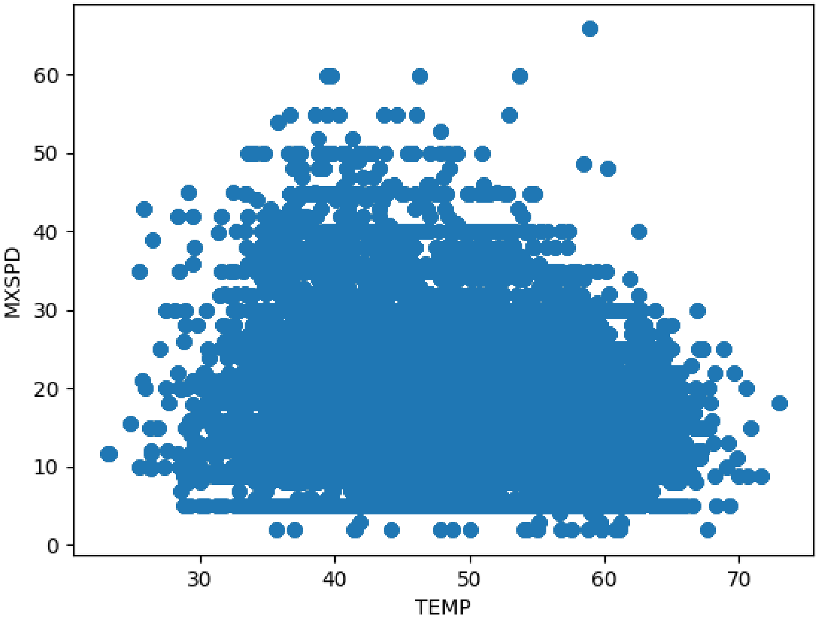
-
- This plot tends to make me think that Dew Point and Sea Level pressure have a relationship via another variable or two, the data seems to make a heart shape, indicating a complex relationship.

2.2. Dew Point to Max Wind Speed



-
- This plot seems to indicate that there is another variable that affects the relationship between dew point and max wind speed.

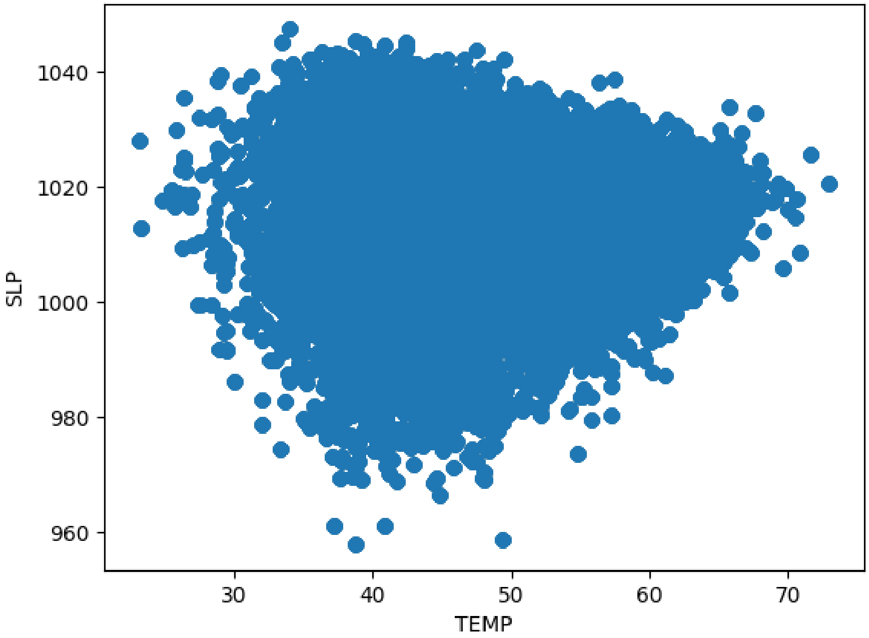
2.3. Temperature to Max Wind Speed



-
- This plot seems to indicate that there is another variable that affects the relationship between temperature and max wind speed.

Viewing this plot next to the previous I am inclined to believe that the other variable in the relationship is dew point.

2.4. Temperature to Sea Level Pressure



-
- This plot tends to make me think that Temperature and Sea Level pressure have a relationship via another variable or two, the data seems to make a heart shape, indicating a complex relationship. Looking at the Dew Point and Sea Level Pressure plot with this one the conclusion might be drawn to see if dew point is part of the relationship.

References

[1] National Oceanic and Atmospheric Administration. "NOAA GSOD." Kaggle. March 13, 2018. Accessed March 28, 2018.