David Leifer

01/28/2019

Surface Weather Stations & Univariate Analysis

Geography 531

Link to code: <https://github.com/DavidLeifer/GEOG431531/blob/master/assignment1/Assignment1.ipynb>

Introduction

Climatic and weather data is routinely downloaded, scrapped, and processed as a source of information to provide insights into spatial patterns of climatology. Although these data are the gold standard of climatology, there are many processing issues that need to be addressed before conclusions can be drawn. We needed to decipher to what extent temperature and dew point values have changed over the past 70 years in Chicago. To do this, we started by downloading Automated Surface Observing Systems (ASOS) weather data from January 1st, 1948 until December 31st, 2018 from the Iowa Environmental Mesonet database for the station located at Chicago O’Hare. The data was then cleaned using data science techniques and the scripting language Python in the Jupyter Notebook environment. Plots were drawn using built in libraries and analysis was conducted. We found a positive trend in the data indicating a slight warming pattern for this time period. Further work is required to expand the data from one dimension into higher dimensionality and include gridded coverage along a timeline.

Methodology

Once the data was downloaded, the Python package Pandas was used to import the data and visualize the data in tabular form. The columns for station, day, max temperature, min temperature, max due point, and minimum due point were identified. The most insidious of values in a weather dataset are erroneous data values, which take the form of -99 or -999. It was thus necessary to remove these from our dataset and replace them with a moniker such as “NaN” or “none” using panda’s built in function “replace”. It was then required to count the amount of none values in our dataset. This number was calculated by combining the .isna() and .sum() built in function. This number was 1841. It was also required that the average temperature and dew point between max and min columns be calculated, which occurred using the .mean() function and creating a new column indicator in Pandas. Since we are budding scientists and our data was in Fahrenheit, we converted all six of our columns into Celsius using a custom Python function.

It was established that we can only use data for years that had enough non-missing data in the column. The threshold for this was created at 10%. If a year’s worth of days had less than 10% worth of data, the year was dropped completely from the data. This was accomplished using the .groupby() function of Pandas for each variable and then merged together along the year column. For these years, we then calculated a mean for each of the six Celsius columns. We also decided to look at changes in extreme temperatures and dew points. This was completed using the 90th and 10th percentile quantiles for all the days from 1948 until 2018. We were required to find the days that exceeded the 90th percentile and fell below the 10th percentile.

Once the data was successfully processed and cleaned, it was time to start analyzing the data. Our primary research question was to what extent has the temperature and dewpoint changed over the past 70 years near Chicago. To accomplish this, we needed to decide on a univariate statistical test to apply to our newly cleaned data, each of which have their own positives and negatives. The first univariate statistical test examined in class was the running mean or median. This is an easy to implement test that calculates the mean of data points in successive, temporal order. The next test examined was the ordinary least squares linear regression (OLS). This test is the dependence of a single variable on an independent variable, in our case time. This measures the difference between a linear regression and the data provided, by minimizing the sum of squares error. It is often included with t-testing to discover the significance of the trend. The third method is the Theil-Sen trend estimator. This estimation is similar to OLS, however it estimates the regression using a median of the slope (rise over run). It is a non-parametric test, thus it does not assume normal distribution of the data. The final univariate test examined was the difference in time, which situates slices of the mean, median, and tails into time periods. We will use OLS to summarize our data because it is a simple yet powerful way to encapsulate data.

Results

There are multiple ways to calculate OLS, from using software packages like SPSS and ArcMap to rolling your own statistical test in a scripting language like Python or MATLAB. For the sake of brevity, we used a built-in Python package called statsmodels.api. The only requirements were to set the x and y variable as column subsets of our Pandas dataframe. Then with three simple lines, an OLS regression results table was produced for max temperature mean (*Table 1*).

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | max\_temp\_c\_mean | **R-squared:** | 0.997 |
| **Model:** | OLS | **Adj. R-squared:** | 0.997 |
| **Method:** | Least Squares | **F-statistic:** | 2.181e+04 |
| **Date:** | Mon, 28 Jan 2019 | **Prob (F-statistic):** | 4.52e-89 |
| **Time:** | 14:26:56 | **Log-Likelihood:** | -87.938 |
| **No. Observations:** | 71 | **AIC:** | 177.9 |
| **Df Residuals:** | 70 | **BIC:** | 180.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0074 | 5.03e-05 | 147.683 | 0.000 | 0.007 | 0.008 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.361 | **Durbin-Watson:** | 1.890 |
| **Prob(Omnibus):** | 0.307 | **Jarque-Bera (JB):** | 1.611 |
| **Skew:** | 0.305 | **Prob(JB):** | 0.447 |
| **Kurtosis:** | 3.415 | **Cond. No.** | 1.00 |

*Table 1:* The OLS output from our code for max temperature mean.

As is made apparent in the table, we have a very large t-statistic, indicating that we should reject the null hypothesis. Using the Python package matplotlib, we were then able to plot the data with a linear trend line and confidence intervals (*Figure 1*.)

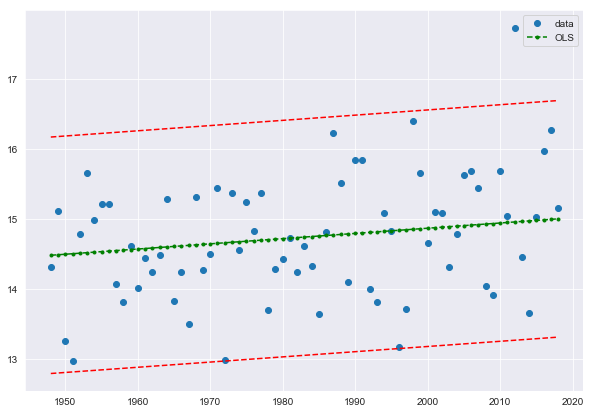


Figure 1: The trendline for max temperature appears positive.

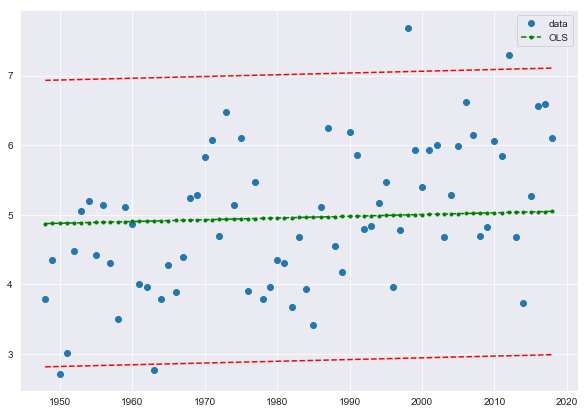
The same method was applied to min temperature mean (*Table 2*) and plotted (*Figure 2*). The min temperature again had a large t value, indicating to reject the null hypothesis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | min\_temp\_c\_mean | **R-squared:** | 0.960 |
| **Model:** | OLS | **Adj. R-squared:** | 0.959 |
| **Method:** | Least Squares | **F-statistic:** | 1664. |
| **Date:** | Tue, 29 Jan 2019 | **Prob (F-statistic):** | 1.57e-50 |
| **Time:** | 13:38:29 | **Log-Likelihood:** | -101.98 |
| **No. Observations:** | 71 | **AIC:** | 206.0 |
| **Df Residuals:** | 70 | **BIC:** | 208.2 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0025 | 6.13e-05 | 40.794 | 0.000 | 0.002 | 0.003 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.575 | **Durbin-Watson:** | 1.151 |
| **Prob(Omnibus):** | 0.750 | **Jarque-Bera (JB):** | 0.709 |
| **Skew:** | 0.178 | **Prob(JB):** | 0.702 |
| **Kurtosis:** | 2.664 | **Cond. No.** | 1.00 |

*Table 2*: The OLS output from our code for min temperature



*Figure 2*: The result for the min temperature is slightly positive.

The method was applied to max dewpoint mean (*Table 3*) and plotted (*Figure 3*). The max dewpoint again had a large t value, indicating to reject the null hypothesis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | max\_dewpoint\_c\_mean | **R-squared:** | 0.988 |
| **Model:** | OLS | **Adj. R-squared:** | 0.988 |
| **Method:** | Least Squares | **F-statistic:** | 5999. |
| **Date:** | Tue, 29 Jan 2019 | **Prob (F-statistic):** | 1.41e-69 |
| **Time:** | 13:51:08 | **Log-Likelihood:** | -84.912 |
| **No. Observations:** | 71 | **AIC:** | 171.8 |
| **Df Residuals:** | 70 | **BIC:** | 174.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0037 | 4.82e-05 | 77.455 | 0.000 | 0.004 | 0.004 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.958 | **Durbin-Watson:** | 1.480 |
| **Prob(Omnibus):** | 0.619 | **Jarque-Bera (JB):** | 0.396 |
| **Skew:** | -0.029 | **Prob(JB):** | 0.820 |
| **Kurtosis:** | 3.361 | **Cond. No.** | 1.00 |

*Table 3*: The OLS output table from max dew point mean.



*Figure 3*: The dew point max is slightly positive.

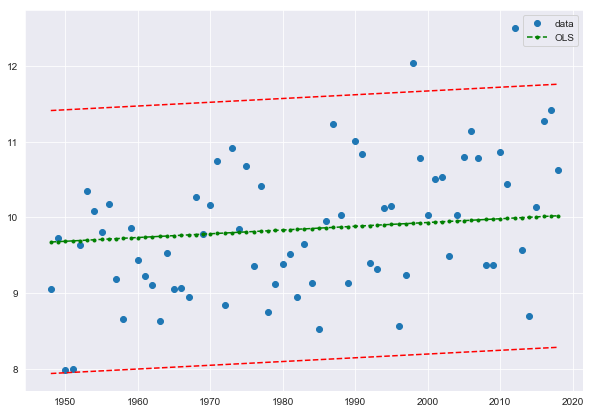
When the same methodology was applied to the min dewpoint mean, no results were made available. When plotted, the graph lacked the OLS trendline and confidence intervals. The average temperature mean however did create the correct table (*Table 4*) and plot (*Figure 4*).

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | avg\_temp\_c\_mean | **R-squared:** | 0.992 |
| **Model:** | OLS | **Adj. R-squared:** | 0.992 |
| **Method:** | Least Squares | **F-statistic:** | 9177. |
| **Date:** | Tue, 29 Jan 2019 | **Prob (F-statistic):** | 5.61e-76 |
| **Time:** | 13:57:38 | **Log-Likelihood:** | -90.059 |
| **No. Observations:** | 71 | **AIC:** | 182.1 |
| **Df Residuals:** | 70 | **BIC:** | 184.4 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0050 | 5.18e-05 | 95.794 | 0.000 | 0.005 | 0.005 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.072 | **Durbin-Watson:** | 1.536 |
| **Prob(Omnibus):** | 0.355 | **Jarque-Bera (JB):** | 1.750 |
| **Skew:** | 0.384 | **Prob(JB):** | 0.417 |
| **Kurtosis:** | 2.983 | **Cond. No.** | 1.00 |

*Table 4*: The OLS output table from average temperature mean.



*Figure 4*: The average temperature mean is again slightly positive.

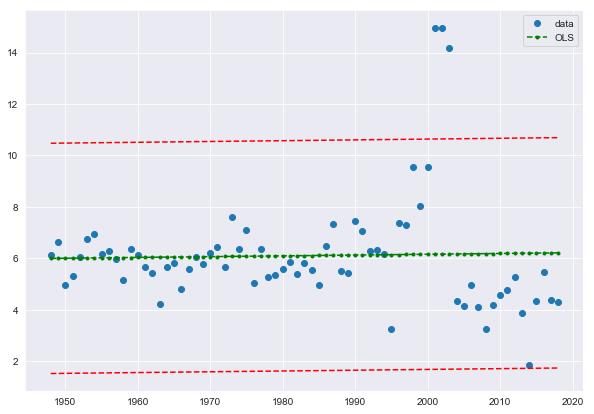
The table (*Table 5*) and plot (*Figure 5*) were again constructed for average dew point mean.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | avg\_dp\_c\_mean | **R-squared:** | 0.884 |
| **Model:** | OLS | **Adj. R-squared:** | 0.882 |
| **Method:** | Least Squares | **F-statistic:** | 533.5 |
| **Date:** | Tue, 29 Jan 2019 | **Prob (F-statistic):** | 1.82e-34 |
| **Time:** | 14:00:59 | **Log-Likelihood:** | -157.13 |
| **No. Observations:** | 71 | **AIC:** | 316.3 |
| **Df Residuals:** | 70 | **BIC:** | 318.5 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0031 | 0.000 | 23.097 | 0.000 | 0.003 | 0.003 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 52.487 | **Durbin-Watson:** | 0.632 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 208.083 |
| **Skew:** | 2.283 | **Prob(JB):** | 6.54e-46 |
| **Kurtosis:** | 10.035 | **Cond. No.** | 1.00 |

*Table 5*: The OLS output table from average dew point mean.



*Figure 5*: The average dew point mean is again slightly positive.

The final step was to take the tables containing the extreme values lower than the threshold of 10th percentile and higher than 90th percentile and again plot these values and run OLS on them. The first OLS table was for max temperature means under the 10th percentile threshold, then averaged by year (*Table 6* and *Figure 6*). The same method was applied to max temperature values above the 90th percentile threshold (*Table 7* and *Figure 7*). The methodology was again applied to min temperature values for upper and lower percentiles (*Tables 8 and 9* and *Figures 8 and 9*). And again, for max (*Tables 10 and 11* and *Figures 10 and 11*). We had issues plotting the min dew point.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | max\_temp\_c\_mean | **R-squared:** | 0.997 |
| **Model:** | OLS | **Adj. R-squared:** | 0.997 |
| **Method:** | Least Squares | **F-statistic:** | 2.181e+04 |
| **Date:** | Wed, 30 Jan 2019 | **Prob (F-statistic):** | 4.52e-89 |
| **Time:** | 13:01:35 | **Log-Likelihood:** | -87.938 |
| **No. Observations:** | 71 | **AIC:** | 177.9 |
| **Df Residuals:** | 70 | **BIC:** | 180.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0074 | 5.03e-05 | 147.683 | 0.000 | 0.007 | 0.008 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.361 | **Durbin-Watson:** | 1.890 |
| **Prob(Omnibus):** | 0.307 | **Jarque-Bera (JB):** | 1.611 |
| **Skew:** | 0.305 | **Prob(JB):** | 0.447 |
| **Kurtosis:** | 3.415 | **Cond. No.** | 1.00 |

*Table 6*: A large t-statistic for this value.



*Figure 6*: The extremes for each year for max temperature seem to be increasing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | max\_temp\_c\_mean | **R-squared:** | 0.997 |
| **Model:** | OLS | **Adj. R-squared:** | 0.997 |
| **Method:** | Least Squares | **F-statistic:** | 2.181e+04 |
| **Date:** | Wed, 30 Jan 2019 | **Prob (F-statistic):** | 4.52e-89 |
| **Time:** | 13:14:55 | **Log-Likelihood:** | -87.938 |
| **No. Observations:** | 71 | **AIC:** | 177.9 |
| **Df Residuals:** | 70 | **BIC:** | 180.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0074 | 5.03e-05 | 147.683 | 0.000 | 0.007 | 0.008 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.361 | **Durbin-Watson:** | 1.890 |
| **Prob(Omnibus):** | 0.307 | **Jarque-Bera (JB):** | 1.611 |
| **Skew:** | 0.305 | **Prob(JB):** | 0.447 |
| **Kurtosis:** | 3.415 | **Cond. No.** | 1.00 |

*Table 7*: The values for max temperature over the 90th percentile.



*Figure 7*: The plot of values for max temperature over the 90th percentile.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | min\_temp\_c\_mean | **R-squared:** | 0.960 |
| **Model:** | OLS | **Adj. R-squared:** | 0.959 |
| **Method:** | Least Squares | **F-statistic:** | 1664. |
| **Date:** | Wed, 30 Jan 2019 | **Prob (F-statistic):** | 1.57e-50 |
| **Time:** | 13:22:03 | **Log-Likelihood:** | -101.98 |
| **No. Observations:** | 71 | **AIC:** | 206.0 |
| **Df Residuals:** | 70 | **BIC:** | 208.2 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0025 | 6.13e-05 | 40.794 | 0.000 | 0.002 | 0.003 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.575 | **Durbin-Watson:** | 1.151 |
| **Prob(Omnibus):** | 0.750 | **Jarque-Bera (JB):** | 0.709 |
| **Skew:** | 0.178 | **Prob(JB):** | 0.702 |
| **Kurtosis:** | 2.664 | **Cond. No.** | 1.00 |

*Table 8*: The values for min temperature under the 10th percentile.



*Figure 8*: The plot for min temperature under the 10th percentile.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | min\_temp\_c\_mean | **R-squared:** | 0.960 |
| **Model:** | OLS | **Adj. R-squared:** | 0.959 |
| **Method:** | Least Squares | **F-statistic:** | 1664. |
| **Date:** | Wed, 30 Jan 2019 | **Prob (F-statistic):** | 1.57e-50 |
| **Time:** | 13:24:28 | **Log-Likelihood:** | -101.98 |
| **No. Observations:** | 71 | **AIC:** | 206.0 |
| **Df Residuals:** | 70 | **BIC:** | 208.2 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0025 | 6.13e-05 | 40.794 | 0.000 | 0.002 | 0.003 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.575 | **Durbin-Watson:** | 1.151 |
| **Prob(Omnibus):** | 0.750 | **Jarque-Bera (JB):** | 0.709 |
| **Skew:** | 0.178 | **Prob(JB):** | 0.702 |
| **Kurtosis:** | 2.664 | **Cond. No.** | 1.00 |

*Table 9*: The values for min temperature over the 90th percentile.



*Figure 9*: The plot of values for min temperature over the 90th percentile.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | max\_dewpoint\_c\_mean | **R-squared:** | 0.988 |
| **Model:** | OLS | **Adj. R-squared:** | 0.988 |
| **Method:** | Least Squares | **F-statistic:** | 5999. |
| **Date:** | Wed, 30 Jan 2019 | **Prob (F-statistic):** | 1.41e-69 |
| **Time:** | 13:28:57 | **Log-Likelihood:** | -84.912 |
| **No. Observations:** | 71 | **AIC:** | 171.8 |
| **Df Residuals:** | 70 | **BIC:** | 174.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0037 | 4.82e-05 | 77.455 | 0.000 | 0.004 | 0.004 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.958 | **Durbin-Watson:** | 1.480 |
| **Prob(Omnibus):** | 0.619 | **Jarque-Bera (JB):** | 0.396 |
| **Skew:** | -0.029 | **Prob(JB):** | 0.820 |
| **Kurtosis:** | 3.361 | **Cond. No.** | 1.00 |

*Table 10*: The values of max dew point under the 10th percentile.



*Figure 10*: The plot of values for max dew point under the 10th percentile.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | max\_dewpoint\_c\_mean | **R-squared:** | 0.988 |
| **Model:** | OLS | **Adj. R-squared:** | 0.988 |
| **Method:** | Least Squares | **F-statistic:** | 5999. |
| **Date:** | Wed, 30 Jan 2019 | **Prob (F-statistic):** | 1.41e-69 |
| **Time:** | 13:32:03 | **Log-Likelihood:** | -84.912 |
| **No. Observations:** | 71 | **AIC:** | 171.8 |
| **Df Residuals:** | 70 | **BIC:** | 174.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **year** | 0.0037 | 4.82e-05 | 77.455 | 0.000 | 0.004 | 0.004 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.958 | **Durbin-Watson:** | 1.480 |
| **Prob(Omnibus):** | 0.619 | **Jarque-Bera (JB):** | 0.396 |
| **Skew:** | -0.029 | **Prob(JB):** | 0.820 |
| **Kurtosis:** | 3.361 | **Cond. No.** | 1.00 |

*Table 11*: The values for max dew point over the 90th percentile.



*Figure 11*: The plot of values for max dew point over the 90th percentile.

Conclusions

From these exercises, we were able to successfully clean data as a real meteorologist or climatologist would. We discovered an overall positive trend for maximum temperature, minimum temperature, and maximum dew point. Maximum temperature was more pronounced as a positive trendline and we had difficulty plotting the minimum dew point. For extreme weather, we also discovered a positive trendline for values that exceeded the 90th percentile and were less than the 10th percentile for maximum temperature, minimum temperature, and maximum dew point. We again had difficulty plotting the minimum dew point extreme values. This evidence supports the hypothesis that the Chicago O’Hare station is warming slightly over the past 70 years.