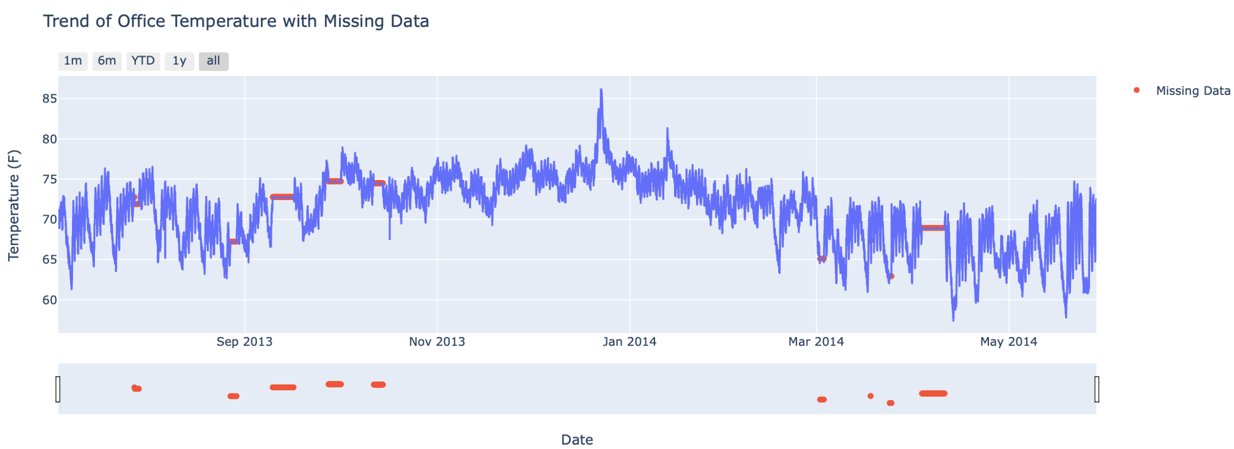
Time Series Anomaly Detection: Temperature Data

Analysis Summary

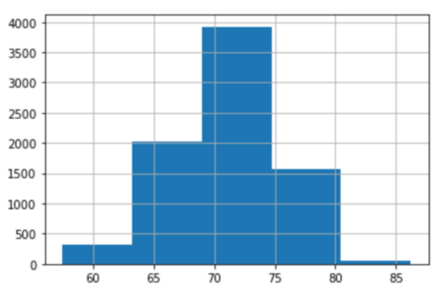
1. Explore and Feature Engineering:

- The dataset contains two columns: timestamp and the temperature values.
- The timestamps are at an interval of an hour from the start date 2013-07-04 to 2014-05-28.
- There were no Null values in the dataset but few hours missing, so the hours were added into the dataset and empty values forward filled. (261 rows)

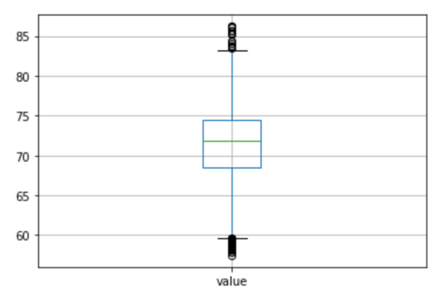


Red points showing missing data was forward filled.

- Understanding the distribution of the temperature values.



Almost Normal Distribution



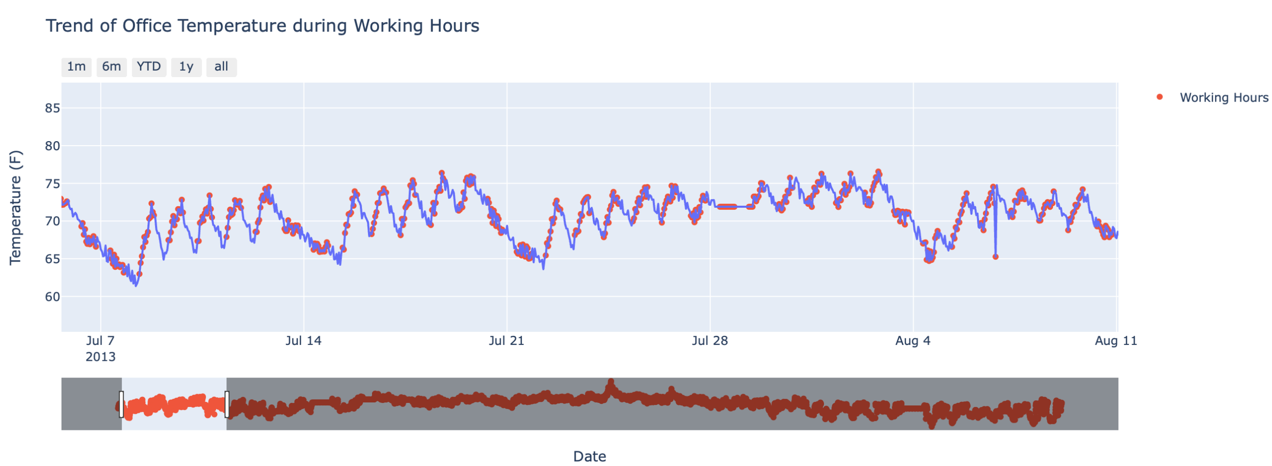
Box Plot showing extreme value

- New features extracted from the datetime and temperature value column:
  - Hour, day, month, weekday, quarter.
  - Weekend column if the day of week is Sunday or Saturday.
  - Working Hours (assuming working hours from 8am to 8pm)
  - Temperature lag (24-hours lag variable)
  - Change in lag (difference between current and 24-hour lag temperature.)

- Examining weekend temperature trends
  - On most weekends there is a continuous drop in temperature as shown in the graph below.
  - Possible inference: the office is in a cold region where the temperature drops continuously if heating is not on.

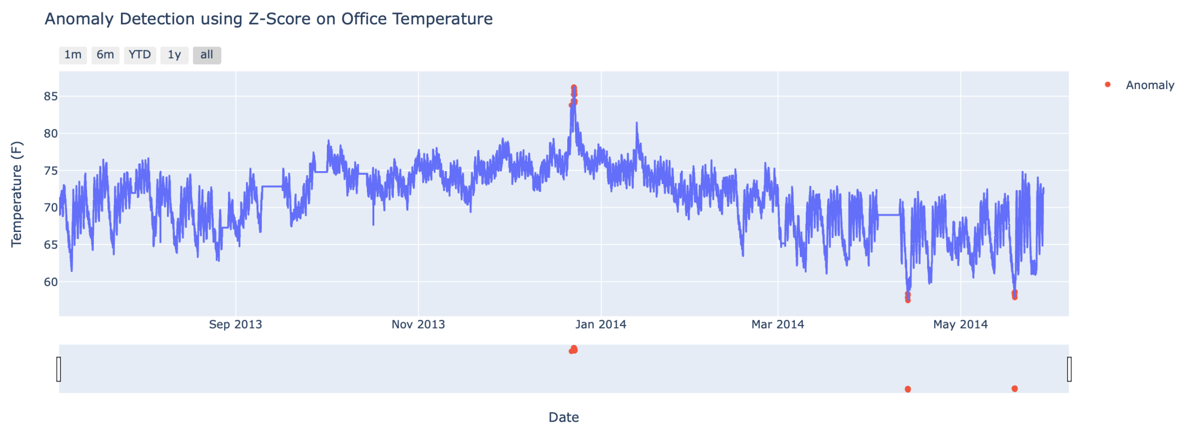
Red points showing weekend data.

- Examining working hours temperature trends
  - On most working hours the temperature keeps on rising as shown in the graph below.
  - Possible inference: the office is in a cold region where the temperature is maintained by heating during working hours.



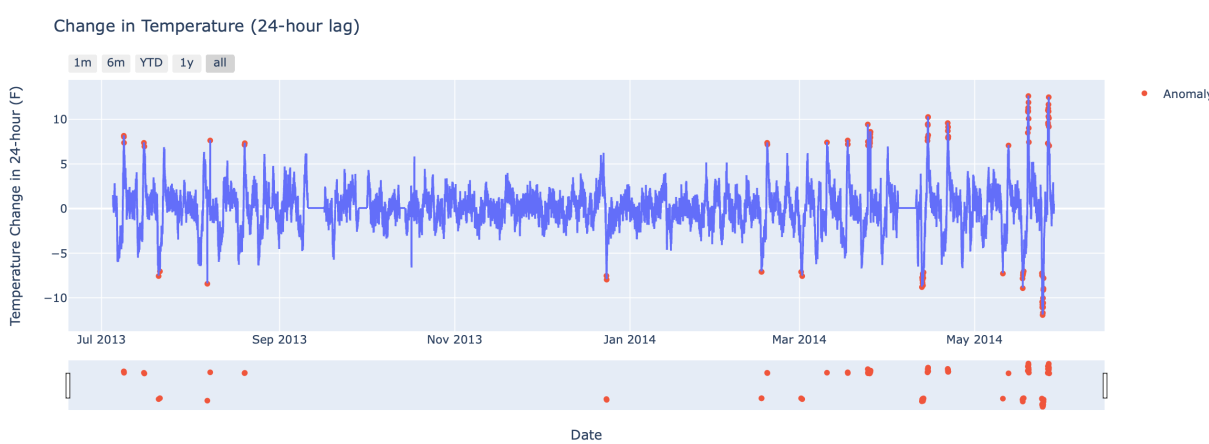
Red points showing working hours.

2. **Anomaly Detection using Statistical Method (z-score)**
    - The z-score measures how far a data point is away from the mean as a signed multiple of standard deviation. Large absolute values of z-score suggest an anomaly.
    - Z-score method of anomaly detection can be used for real world anomaly detection in production using the sliding window technique.
    - Easy of infer and not computation heavy.
    - In-case the extreme values are too skewed (mean and median differ significantly), a technique called *<u>modified z-score</u>* can be used that utilizes the median. In this dataset the median and mean are almost similar.
    - Z-Score on Temperature values:
    - This method helps us to identify the extreme point from temperature values as anomalies. (22 anomalous points)



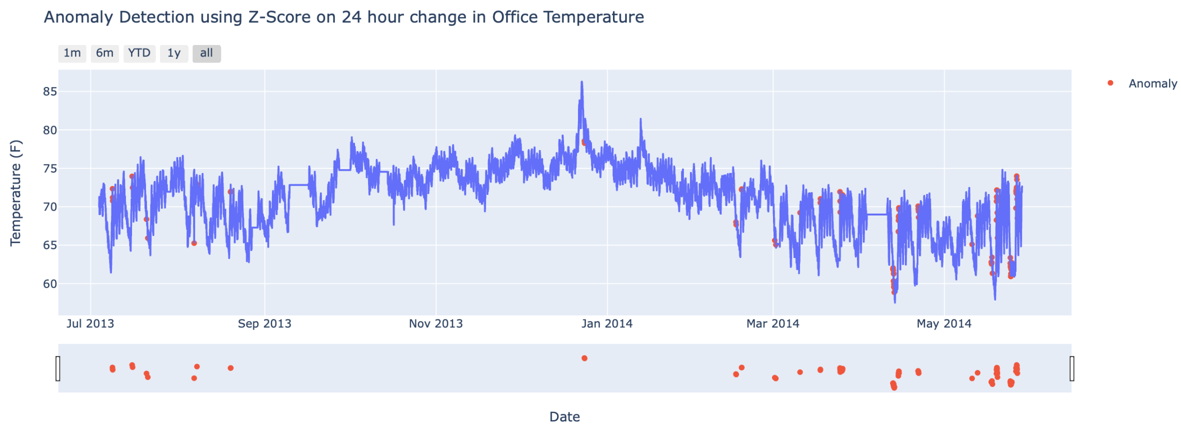
Red points showing anomalies.

- Z-Score on lag difference values:
- The graph below shows the lag difference value (temperature – temperature 24 hours back) and the anomalies detected.



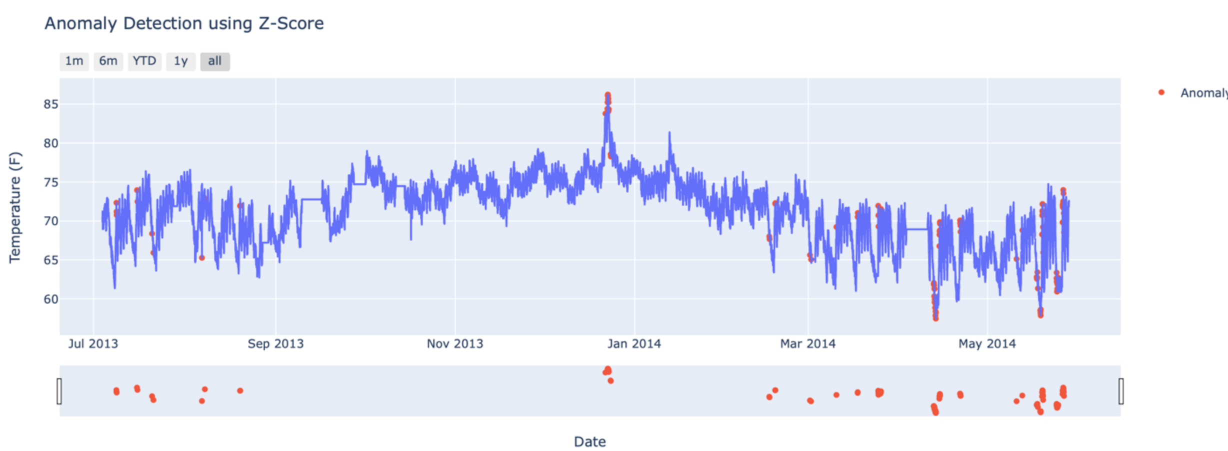
Red points showing anomalies.

- This method helps us to identify the local anomalous points with respect to 24-hour time period. (101 anomalous points)



Red points showing anomalies.

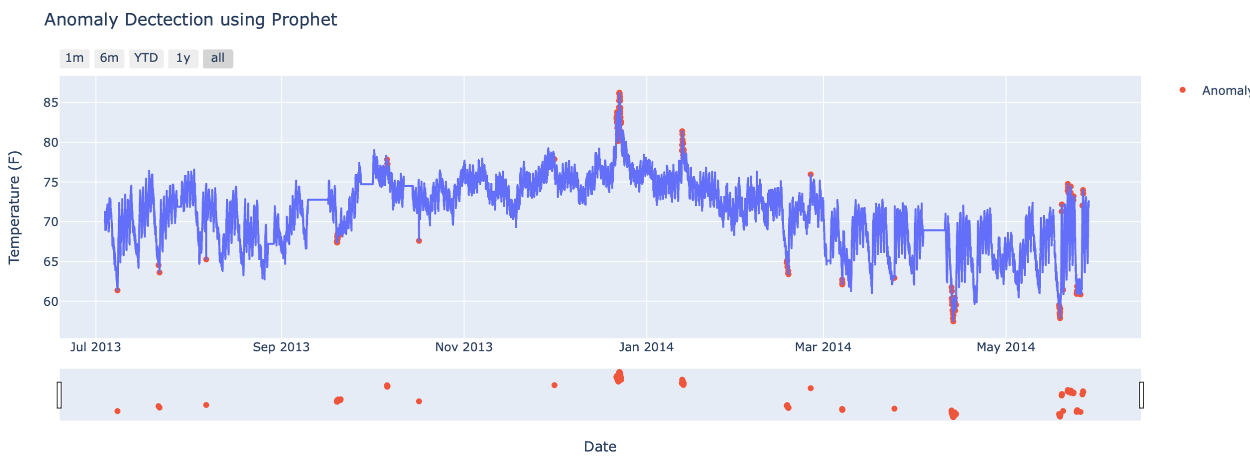
- Combined Z-Score based model:
- 123 anomalous points detected from 7888 rows (~1.5%)



Red points showing anomalies.

3. Anomaly Detection using FBProphet

- Prophet library helps with forecasting time series data that maybe non-linear by considering the seasonality. It recognizes the trends over the given dataset and is easy to tune.
- Prophet is used to forecast and can be extended to find anomalies. When the original values cross the forecasted lower or upper range, these points can be termed as anomalies.
- The plots build by the library also helps in understanding the daily, weekly, yearly trends.
- Anomaly Graph using the library
- 150 anomalous points recognized from the 7888 rows of data. (~1.9%)



Red points showing anomalies.

Possible Alternate Methods:

1. Use of ML based Anomaly detection like DBSCAN and Isolation Forest.
   - Distance Based Spatial Clustering and Isolation Forest is highly used for anomaly detection projects across industry.
   - As there were no other feature available apart from the Temperature itself, the decision to use Prophet seems better. Prophet considers the seasonality (daily, weekly, yearly) that might be non-linear to predict the values on time series.
   - After the feature engineering of extracting weekend, working hours and few others it can however be used to build a DBSCAN or Isolation Forest model.
2. Auto Regressive (AR) Integrated Moving Average (MA) based Anomaly Detection.
   - ARIMA can be used on stationary data to analyze and predict time series data which leverages the lagged features and moving averages.
   - Though in this case the Temperature data is stationary, it is a more complex procedure to tune ARIMA models and then define the error range for anomaly.
3. Clustering based Anomaly Detection.
   - Finding the right K value becomes a challenge in real time. The elbow curve can be used but the K value cannot be static as the incoming data might have a different optimal K value.