Deep Neural Network for Weather Time Series Forecasting – Milestone Report

Contents

1. Introduction 2
2. Time Series Analysis 2
3. Weather Forecast 2
4. Historical Hourly Weather Dataset 2
5. Data Wrangling 3
6. Exploratory Data Analysis 3
   a. Outliers 3
   b. Resampling and Converting Frequencies 3
   c. Rolling windows 4
   d. Deterministic vs Non deterministic Time series 4
   e. Time Series Histogram and Density Plot 5
   f. Seasonal Decomposition 5
   g. Lag Plot 6
   h. Auto Correlation Factor Plot 7
   i. Auto Correlation Factor Plot 8
   j. Intuition of ACF and PACF 8
   k. MultiVariate Analysis 9
   l. Weather Statistics of different cities 9
      1. Hottest and Coldest cities 9
      2. Minimum Humid of all Cities 11
      3. Maximum Wind Speeds 11
7. References: 11

1. Introduction

Forecasting future values of a time series plays an important role in nearly all fields of science and engineering, such as economics, finance, business intelligence and industrial applications, also in real world applications such as speech recognition, real time sign language translation, finance markets, weather forecast etc. Deep Learning algorithms are known to perform best when there is a massive dataset available for learning. Not every time series problem has massive dataset available. In that case advanced ML algorithms are available for time series applications. Because of the nature of time series analysis problem where two values of the same feature in two different time steps are considered as different features, the data size available for processing becomes larger. It is hard to decide which algorithms will perform better for a medium size dataset. The recommendation made in this project can be useful for anyone looking for best ML/DL algorithms for medium size time series problem dataset. I used Historical hourly Weather Data from Kaggle website. The dataset contains ~5 years of high temporal resolution (hourly measurements) data of various weather attributes, such as temperature, humidity, air pressure, etc. There is non-temporal data such as longitude and latitude of cities, which is also used in forecasting. Both Univariate and Multivariate time series analysis is done to collect the data for the recommendation.

2. Time Series Analysis

A time series is a sequence of numeric data observations collected over a period of time at regular intervals. The temporal dependency in time series cause two otherwise identical points of time to belong to different classes or predict different behavior. This characteristics generally increases the difficulty of analyzing them.

3. Weather Forecast

Weather forecast is among the most popular forecast problems. Weather forecast includes forecasting temperature, pressure, humidity, wind direction and wind speed. Unlike other time series datasets, weather data has unique features. There is season-to-season, year-to-year variability in the trends of weather data. The temperatures and Pressures are correlated. Wind speed and direction have similar attributes and changing patterns. So all these features could be separately predicted using univariate time series analysis techniques or could be used jointly to predict using multivariate time series techniques.

4. Historical Hourly Weather Dataset

The dataset contains ~5 years of high temporal resolution i.e. hourly measurements of six weather attributes: temperature, humidity, air pressure, wind direction, wind speed and general weather description of 30 US and Canadian cities and 6 Israeli cities. The dataset contains separate file for each weather attribute. Each file contains 36 cities as columns.

Each weather attribute has it’s own file and is organized such that the rows are the time axis (it's the same time axis for all files), and the columns are the different cities (it's the same city ordering for all files as well). Additionally, for each city we also have the country, latitude and longitude information in a separate file.

5. Data Wrangling

The following Data Wrangling steps are performed:

- There are missing values in the datasets. For several cities first full months' data is missing. Deleted those time steps as filling those values could cause the distraction in the training of the models.
- There are other values missing which were forward fill and then backward fill to make sure all the entries are filled.
- Wind Speed and Wind Direction have a value of 0 for several samples. Replaced them with a small value of 0.001 to avoid divide by zero error later during the evaluation of MAPE (mean absolute percent error) performance metric.
- Temperature, Pressure and Wind Speed are converted into popular units of Degree Fahrenite, inches of Mercury and miles per hour respectively.
- Cleaned Data is stored in separate csv files for later univariate and multivariate ML and DL analysis.

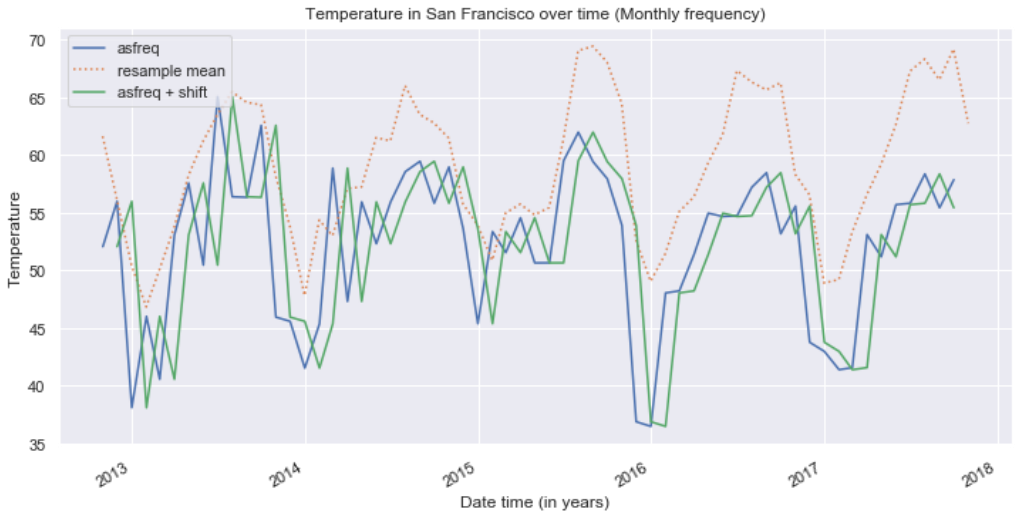
6. Exploratory Data Analysis

a. Outliers

The Box Plot shows that there are some outliers according to the statistics. We will not try to change anything for these outliers since they represent extreme weather conditions which happen in real world time to time. We want to model the time series using these outliers and see if the model can find a pattern and predict the extreme weather for future.

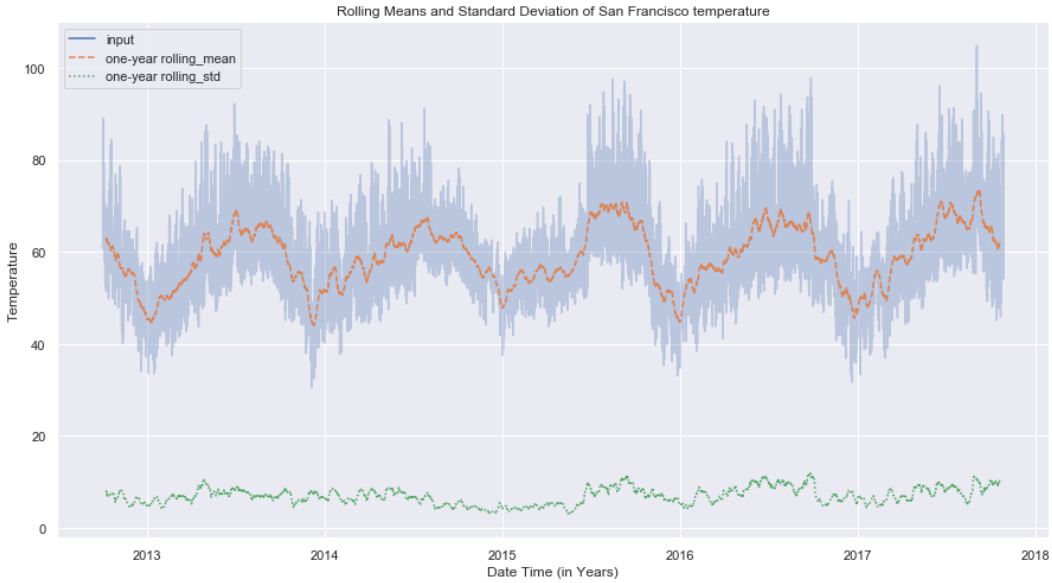
b. Resampling and Converting Frequencies

The above plot shows the monthly temperature of San Francisco and second graph show the temperature at a lag of one month. The graph consists of straight lines which indicates Temperature is a Linear Time Series so gradient is constant between two local minima and maxima values of the graph. The gradients changes between positive and negative value over the period of time. Negative gradient indicates negative correlation and positive gradient indicates positive relationship between time and values of the Temperature.



c. Rolling windows

Rolling statistics are time series-specific operation implemented by Pandas. These can be accomplished via the rolling() attribute of Series and DataFrame objects, which returns a view similar to what we saw with the groupby operation such as Aggregation and Grouping. This rolling view makes available a number of aggregation operations by default. For example, here is the one-year centered rolling mean and standard deviation of the San Francisco Temperature: The visuals of the plot show that rolling mean and rolling standard deviation are constant.



d. Deterministic vs Non deterministic Time series

Time series can be deterministic or non-deterministic in nature. Deterministic time series always behave in an expected manner where as non-deterministic time series is stochastic or random in nature. The following measure indicate is a time series is deterministic or not.

Covariance Stationary - If a time series mean, variance and covariance with past and future values do not change over time then the model is known to be covariance stationary. Time series needs to meet following three criteria to be stationary:

1. Constant Mean - Mean or expected value of a time series over successive time periods needs to be constant for a time series to be considered covariance stationary. This implies that the expected value should not be time dependent.

2. Constant Variance - Variance or standard deviation of a time series needs to be constant over time and should not be dependent on time. This is the second criteria for a time series to be covariance stationary.

3. Constant Covariance - If a covariance is not constant in a time series then the time series exhibits randomness. Additionally, time series distribution changes without any obvious pattern. This indicates that the time series time points have changing correlation.

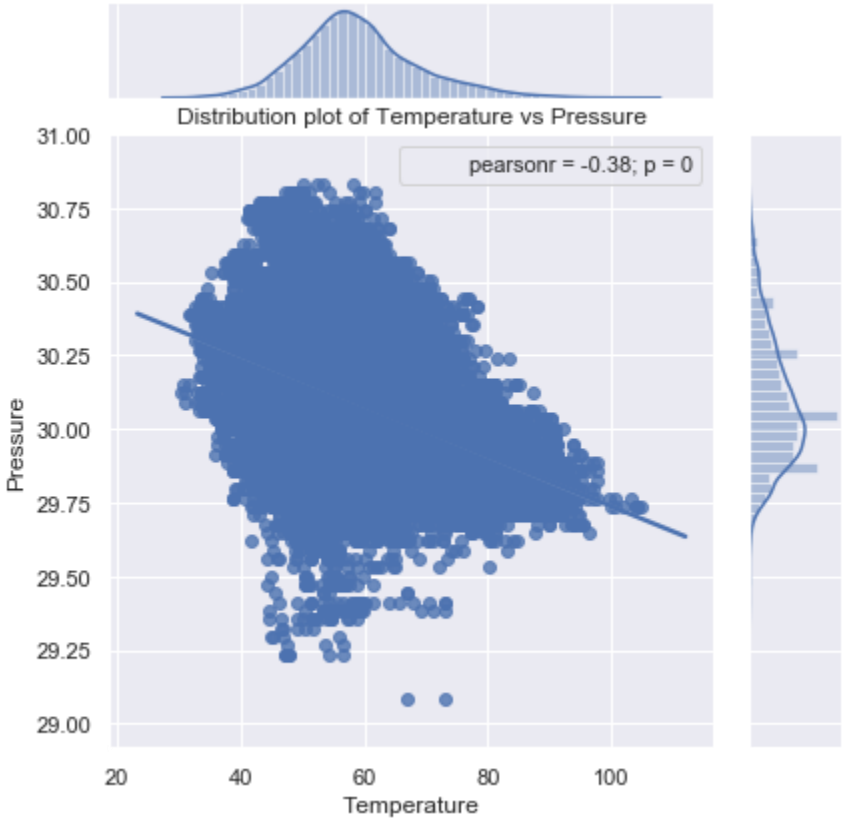
The San Francisco Temperature Dataset is divided into 2 sets and mean, standard deviation and co-variance is calculated for each set. The result indicated that mean and

standard deviation are very similar in both sets of datasets. So the time series is Deterministic and Stationary.

Also Augmented Dickey-Fuller Test from statsmodel library indicated that the temperature is a stationary time series.

e. Time Series Histogram and Density Plot

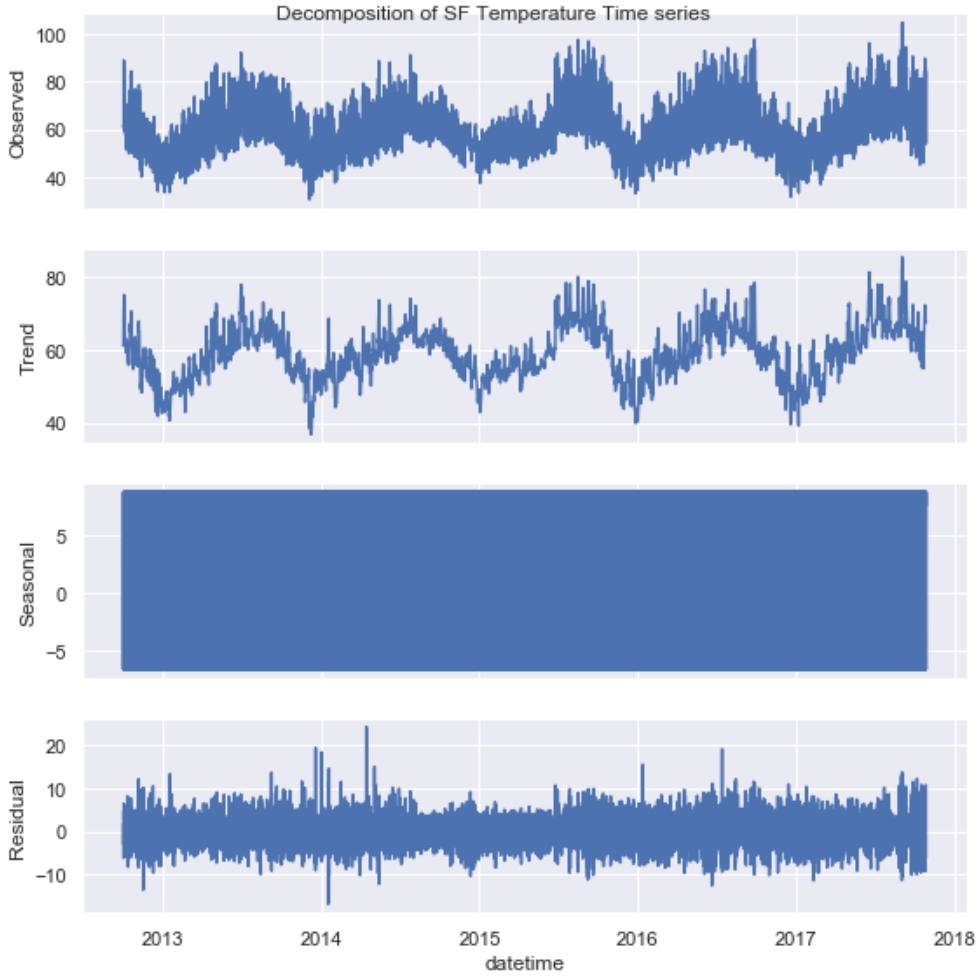
The joint plot below shows that Temperature distribution is Normal distribution with little skewness to the right. Which could indicate outliers in higher temperatures. Pressure distribution is also normal distribution with little skewness to the left. Pearson coefficient is -0.38 which means they are inversely co-related.



f. Seasonal Decomposition

Using time-series decomposition makes it easier to quickly identify a changing mean or variation in the data. The plots below clearly show that there is yearly (long term) trend. Also there is constant seasonality. The residuals seem random with zero mean which makes it a white noise. These can be used to understand the structure of our time-series. The intuition behind time-series decomposition is important, as many forecasting methods build

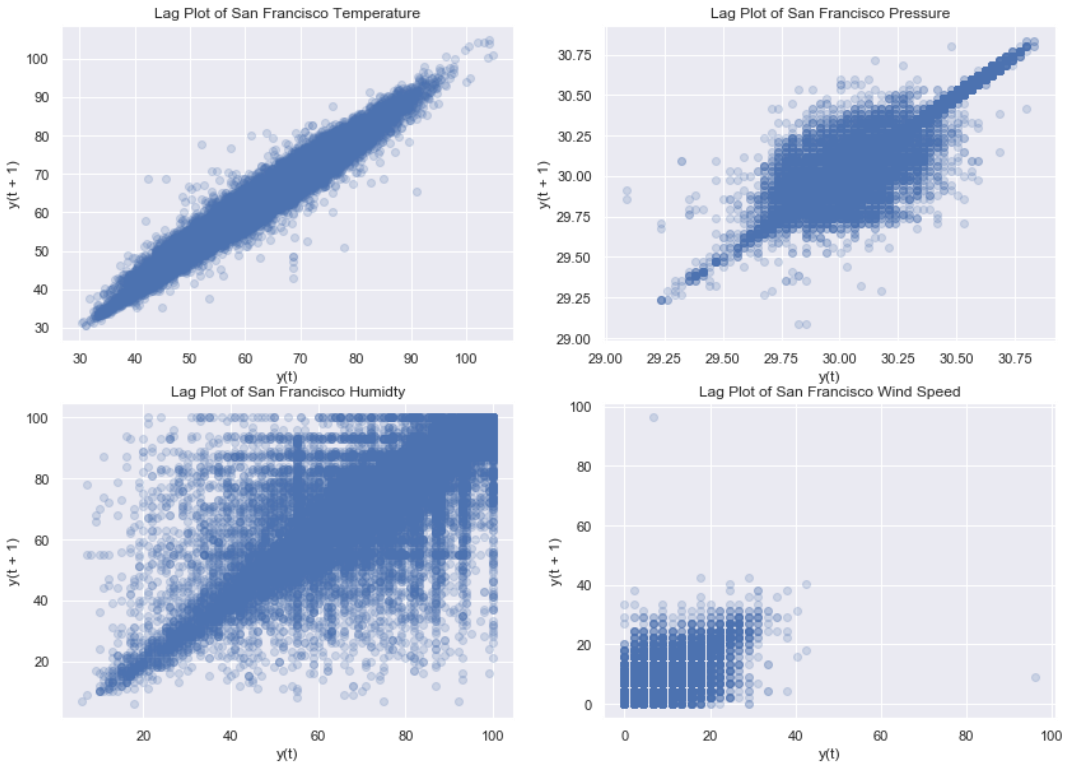
upon this concept of structured decomposition to produce forecasts.



g. Lag Plot

A lag plot is a scatter plot for a time series and the same data lagged. With such a plot, we can check whether there is a possible correlation between current value and the lagged value. Following observations can be made from the plot below.

1. A linear shape shows a relatively strong positive correlation between observations and their lag1 values.
2. Also it suggests that an autoregressive model is probably a better choice.
3. Outliers are easily discernible on a lag plot. The plot shows that there are several outliers.

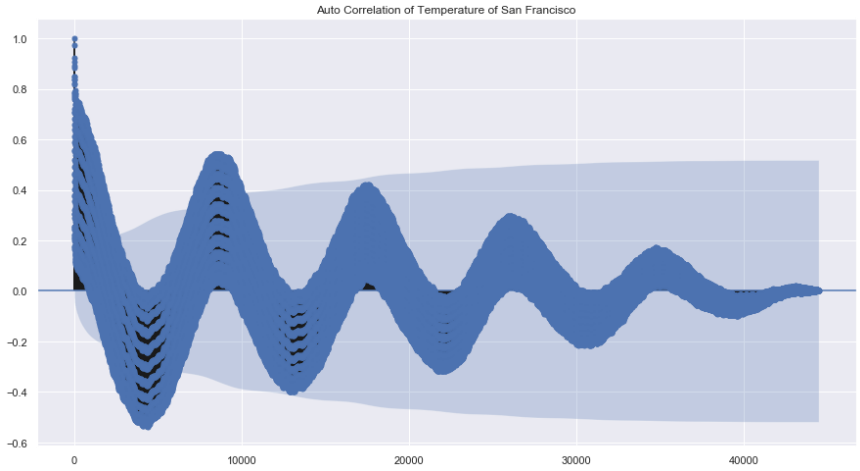


h. Auto Correlation Factor Plot

We can calculate the correlation for time series observations with observations with previous time steps, called lags. Because the correlation of the time series observations is calculated with values of the same series at previous times, this is called a serial correlation, or an autocorrelation.

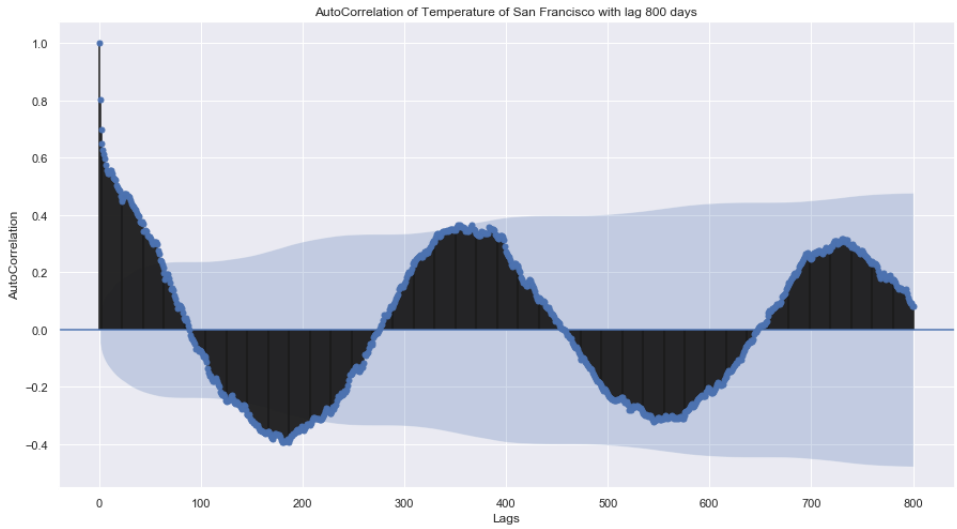
ACF Plot or Auto Correlation Factor Plot is generally used in analyzing the raw data for the purpose of fitting the Time Series Forecasting Models. ACF is used in tandem with PACF (Partial Auto Correlation Factor) to identify which Time series forecasting model to be used.

The plot below shows the lag value along the x-axis and the correlation on the y-axis between -1 and 1. Confidence intervals are drawn as a shaded cone. By default, this is set to a 95% confidence interval, suggesting that correlation values outside of this cone are very likely a correlation and not a statistical fluke. The below plot shows the with a smaller lag value. All the points except are first point are outside the confidence interval. So they are statistically significant.



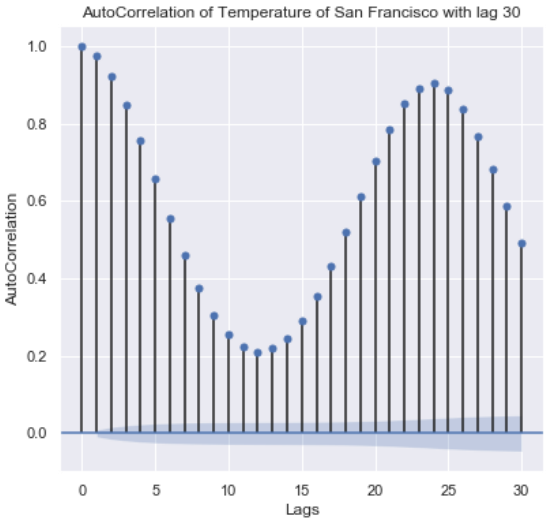
i. Auto Correlation Factor Plot

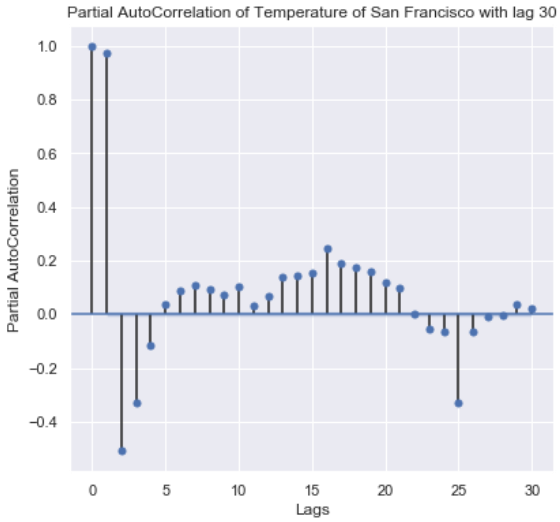
A partial autocorrelation is a summary of the relationship between observations in a time series with observations at prior time steps with the relationships of intervening observations removed. The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags. In the plot below all the points lie outside the confidence interval. So all the lags are statically significant.



j. Intuition of ACF and PACF

The ACF for the AR(k) time series is strong to a lag of k and the inertia of that relationship carry on to subsequent lag values, trailing off at some point as the effect is weakened. We know that the PACF only describes the direct relationship between an observation and its lag. This would suggest that there is no correlation for lag values beyond k. In the ACF plot below, the first correlation value is strong and then it slowly becomes week in subsequent lags. In PACF plot, the correlation value become negative after 2nd value. This indicates that AR model will work on this time series with a value p = 2 for ARIMA model.





k. MultiVariate Analysis

Multivariate time series has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. This dependency is used for forecasting future values.

Johansen Cointegration Test - Cointegration is a statistical property of a collection (X1, X2, ..., Xk) of time series variables. First, all of the series must be integrated of order d (see Order of integration). Next, if a linear combination of this collection is integrated of order less than d, then the collection is said to be co-integrated.

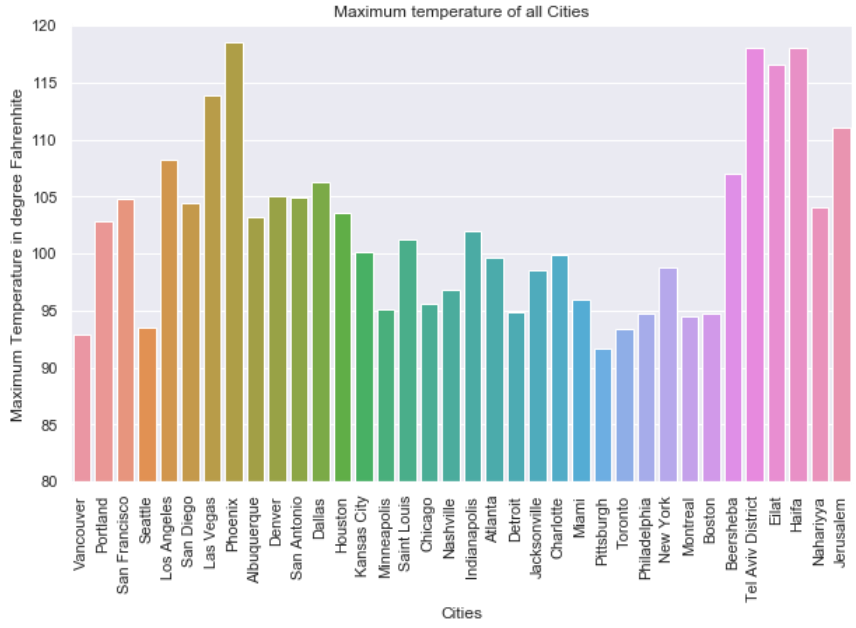
In the Johansen test, we check whether lambda has a zero eigenvalue. When all the eigenvalues are zero, that would mean that the series are not cointegrated, whereas when some of the eigenvalues contain negative values, it would imply that a linear combination of the time series can be created, which would result in stationarity.

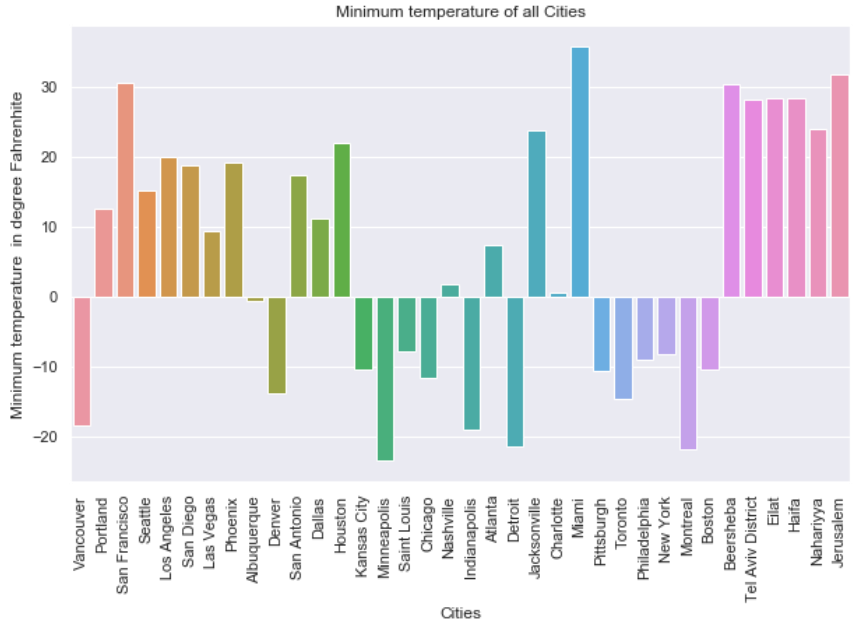
The results of the Johansen Cointegration Test show that 4 vectors of 4 weather attributes are co-integrated.

l. Weather Statistics of different cities

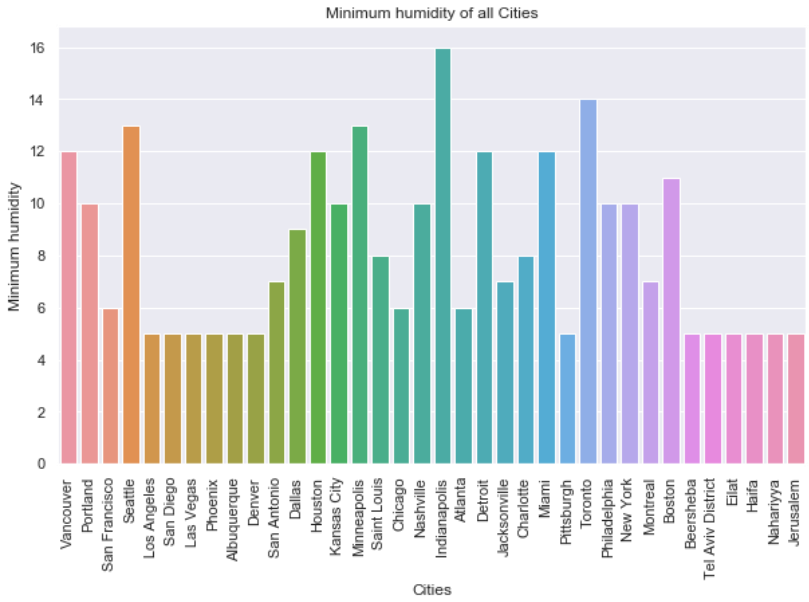
1. Hottest and Coldest cities

Plot below show that Phoenix is the hottest city of all the cities in the dataset and Minneapolis is the coldest city.

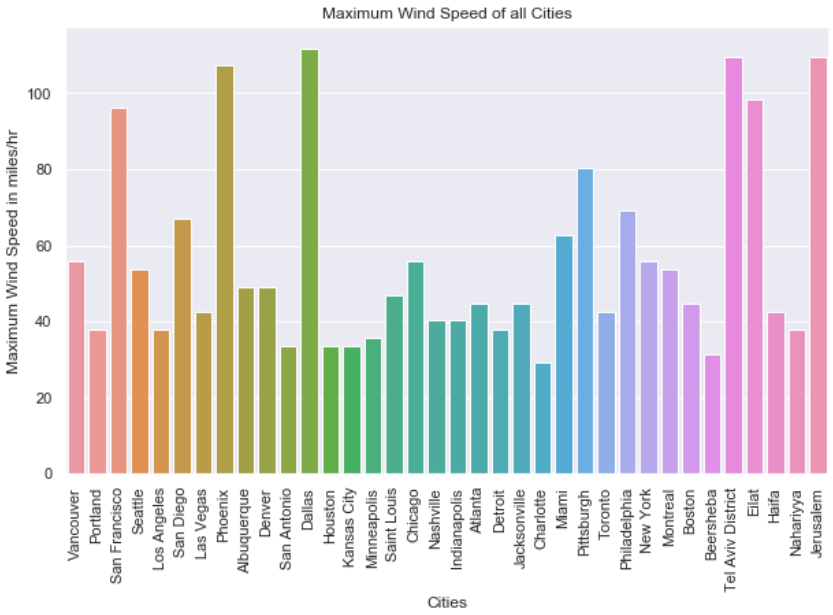

Maximum temperature of all Cities
Maximum Temperature in degree Fahrenhite
120
115
110
105
100
95
90
85
80
Vancouver
Portland
San Francisco
Seattle
Los Angeles
San Diego
Las Vegas
Phoenix
Albuquerque
Denver
San Antonio
Dallas
Houston
Kansas City
Minneapolis
Saint Louis
Chicago
Nashville
Indianapolis
Atlanta
Detroit
Jacksonville
Charlotte
Miami
Pittsburgh
Toronto
Philadelphia
New York
Montreal
Boston
Beersheba
Tel Aviv District
Eilat
Haifa
Nahariyya
Jerusalem
Cities


Minimum temperature of all Cities
Minimum temperature in degree Fahrenhite
30
20
10
0
−10
−20
Vancouver
Portland
San Francisco
Seattle
Los Angeles
San Diego
Las Vegas
Phoenix
Albuquerque
Denver
San Antonio
Dallas
Houston
Kansas City
Minneapolis
Saint Louis
Chicago
Nashville
Indianapolis
Atlanta
Detroit
Jacksonville
Charlotte
Miami
Pittsburgh
Toronto
Philadelphia
New York
Montreal
Boston
Beersheba
Tel Aviv District
Eilat
Haifa
Nahariyya
Jerusalem
Cities

2. Minimum Humid of all Cities



3. Maximum Wind Speeds



7. References:

i. Link to data set: https://www.kaggle.com/selfishgene/historical-hourly-weather-data