DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

Aakash Parwani
May 6, 2017

- ## Introduction

Arhus is Denmark's second-greatest city and the cash related concentration of the Central Denmark Region. The city has a catchment zone of 1.2 million people inside a one-hour travel go and is especially connected with Copenhagen and Hamburg. The tenants of Aarhus live inside walking division of parks and recreational reaches, and inside a 15-minute bike ride of an immaculate coastline, and they advantage from close-by provisions of clean drinking water. In a nearby organized exertion with the business assemble and the city's various data foundations, the City of Aarhus will reduce the city's CO2 transmissions and make keen game plans and green advancement. Creative demonstrating wanders ensure exchange offers of home-created environment courses of action abroad, attract overall theory and fulfill the goal of being CO2-fair by 2030.

This document proposes a review of how atmosphere data examination and representation attempt can help in building framework for sharp urban ranges to stimulate the introduction of splendid city applications for atmosphere envisioning and takes note. In this document we will first talk about the contributions of the competitor article. Then, a new weather analysis and prediction method will come into picture to outperform the competitor's method.

- ## Contribution of Competitor's Article:

Average Daily Air Temperature's Long-Range Forecast Using Inductive Modeling (Zubov, 2013)

In this paper, long-range forecasting average daily air temperature using inductive method was proposed. Principle of high-impact weather events substantiates the different places' interaction by atmosphere, hydrosphere, landmass, biosphere, etc. Forecasting model reasoning's first stage is selection of three most data-related places using Pearson product-moment correlation coefficient, which has to be greater than 0.8 in absolute value. 66 datasets were acquired from NOAA Satellite and Information Service. Second stage is finding weighting coefficients of forecasting model and criterion "minimum of regularity plus maximum of conjunctions" by combinatorial algorithm. This concept is illustrated by Skopje Airport's forecasting model and criterion reasoning, which include datasets from Beijing (China), Ulaanbaatar (Mongolia), and Paphos Airport (Cyprus). Results (conjunctions' percent is 74.8 %, mean absolute error (MAE) is up to 5.7 °F, 166 days lead-time) showed an efficiency of proposed approach. Similar results were achieved for Kiev (MAE is up to 7.2 °F, 167 days lead-time) and Washington National Airport (MAE is up to 6.07 °F, 173 days lead-time). Web-site prototype www.weatherforecast.tk was developed using Ms Windows Azure public cloud computing technology. Proposed approach is characterized by high accuracy, final linear difference equations' simplicity, low computational complexity, and user-friendly interface, which is very suitable for meteorological services.

DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

Highlights

- This paper emphasizes importance of the average daily air temperature's long-range forecast. Further, daily values may be used as very good basis for week, month, and season forecasts.
- The air temperature has great influence on power service's load (Robinson, Peter J., 1997), and one's direct application is for estimating future fuel needs.
- For air temperature's long range forecasting 66 datasets of different countries were acquired from NOAA satellite and information service from 1 January 1973 to 20 April 2013.

Numerical Method

$$\frac{X_F[i]}{\max\{X_R[p]\}} = k_0 + \frac{k_1 X^*_{j_1}[i-d_1]}{\max\{X^*_{j_1}[p-d_1]\}} + \frac{k_2 X^*_{j_2}[i-d_2]}{\max\{X^*_{j_2}[p-d_2]\}} + \frac{k_3 X^*_{j_3}[i-d_3]}{\max\{X^*_{j_3}[p-d_3]\}},$$
$$p = 1,\ 2,\ \ldots,\ l,\quad j_1 \neq j_2 \neq j_3,$$

XF[i], XR[i] – prediction and true air temperature in the forecasted point; i – data position's number in time series, i = 1, 2, 3... 14521 (19 July 1973 – 20 April 2013);

l = 14045 – training sequence's length; k0, k1, K2, k3 – weighting Coefficients.

Xj1[I – d1] , Xj2*[I – d2] , Xj3*[I – d3]* - biased by lead-times *d1, d2, d3 and three days averaged true* air temperature time series for the appropriate places *j1, j2, j3 = 0,1,2, ..., 66 (number of the place from the above list;* plus additional parameters optionally; 0 means the same place in the right and left sides).

- ## Description of Your Contribution:

Contribution to process improvement

Focus of competitor process was only on time series analyses of air temperature and then predicts the same. But in order to make an accurate prediction, consideration of one environment variable is not enough.

To make the weather analyses process reliable, independent variables like **Dew point, Humidity & Air Pressure** are also considered in the new proposed method.

Contribution to reliability improvement

For reliability improvement, in proposed algorithm short range forecasting method is used that generally forecasts to go out to **72 hours or less**. That has **error less than 10% in the mean.**

Model improvement

A). Temperature - K0 + K1 * Dew Point + K2 * Humidity + K3 * Pressure.

The new model has linear structure. Where **K0, K1, K2, K3** are weighting coefficients of independent variable and **Temperature** is dependent variable.

B). For short range time series analyses, dataset is partitioned on monthly basis. And then forecasting is performed.

- ## Data Source and Content:

The weather data for the city of Aarhus in Denmark is public for analyses purpose, available at Weather Data . The dataset is a collection of weather observations from the city of Aarhus. Measurements are recorded from February 2014 – June 2014 and August 2014 – September 2014.

Date Time: - Date and time of weather observation. **Data structure: YYYY/MM/DD HH:MM:SS**

Dew Point (In Degree Celsius): - Dew point is the temperature at which airborne water vapor will amass to shape fluid dew. A higher dew point proposes there will be more steepness discernible all around. Dew point is every so often called ice minute that the temperature is underneath nippy. The estimation of dew indicates is connected mugginess.

Real-Time application, most builders appreciate that development can shape when warm, sodden air encounters a cool surface. The development is shocking, and makers need to keep up a key separation from it. There's an answer, be that as it may: According to building specialists, we can expect development issues in dividers by choosing a divider's temperature profile and playing out a dew-point tally. It's certainly profitable to know whether you're sheathing will be over the dew point or underneath the dew point in winter. When sheathing is underneath the dew point, it's most likely going to gather sogginess. Warm sheathing is better than anything cold sheathing.

Humidity (In Percentage): - Humidity is the measure of water vapor noticeable all around. Water vapor is the vaporous condition of water and is imperceptible. Humidity shows the probability of precipitation, dew, or haze. Higher dampness diminishes the viability of sweating in cooling the body by decreasing the rate of dissipation of dampness from the skin. There are three primary estimations of humidity: total, relative and particular.

Relative humidity is the most frequently encountered measurement of humidity because it is regularly used in weather forecasts. It's an important part of weather reports because it indicates the likelihood of precipitation, dew, or fog. Higher relative humidity also makes it feel hotter outside in the summer because it reduces the effectiveness of sweating to cool the body by preventing the evaporation of perspiration from the skin.

Humidity expects a key part in our step by step atmosphere. Without water vapor detectable all around, our atmosphere may take after the atmosphere on Mars. Would you have the capacity to imagine presence without fogs, rain, snow, thunder, or lightning?

So how does humidity impact us on a hot day? Individuals are sensitive to changes in humidity, in light of the way that our skin uses the air around us to discard clamminess as sweat. In case the relative humidity is high, the air is starting at now drenched with water vapor and our sweat won't vanish. Right when this happens, we feel more smoking than the genuine temperature.

Similarly, low humidity can make us feel cooler than the genuine temperature. This happens in light of the fact that the dry air sweats dissipate more rapidly than expected.

<u>Pressure (In mBar)</u>: - Atmospheric pressure, here and there additionally called barometric weight is the weight applied by the heaviness of air in the environment of Earth (or that of another planet). Much of the time climatic weight is nearly approximated by the hydrostatic weight created by the heaviness of air over the estimation point.

The modification in air pressure after some time has basic suspecting recommendations. As pressure brings down after some time, especially if it is snappy, that implies that a low-pressure system or front is moving closer. This cutting down pressure exhibits an extending likeliness of precipitation. If the air pressure rises basically or stays well superior to anything expected for a drawn out extend of time, which is a sign precipitation is more stunning.

<u>Temperature (In Degree Celsius)</u>: - Temperature is a level of hotness or coldness the can be measured utilizing a thermometer. It's likewise a measure of how quick the particles and atoms of a substance are moving. Temperature is measured in degrees on the Fahrenheit, Celsius, and Kelvin scales.

Guesses in light of temperature and precipitation are fundamental to cultivation, and in this way to dealers inside product markets. Temperatures figures are used by administration associations to gage ask for over coming days. On a standard commence, people use atmosphere evaluations to make sense of what to wear on a given day.

Power and gas associations rely on upon atmosphere assessments to speculate ask for which can be immovably affected by the atmosphere. They use the sum named the degree day to choose how strong of a use there will be for (warming degree day) or (cooling degree day). These sums rely on upon a consistently typical temperature of 65 °F (18 °C). Cooler temperatures drive warming degree days (one for each degree Fahrenheit), while more blazing temperatures oblige cooling degree days. In winter, genuine crisp atmosphere can achieve a surge well known as people turn up their warming. Correspondingly, in summer a surge looked for after can be associated with the extended use of air trim structures in a hot atmosphere. By retribution a surge looked for after, administration associations can purchase additional provisions of constrain or regular gas before the cost increases, or in a couple conditions, supplies are restricted utilizing brownouts and power blackouts.

<u>Wind Direction (In Degrees)</u>: - Wind direction is represented by the course from which it starts. For example, a northerly wind blows from the north toward the south. The wind bearing will vitally affect the ordinary atmosphere. You can frequently be given a twist course and you will have a completely brilliant considered how the atmosphere will change and what atmosphere can be typical with that wind heading.

The wind heading will critically affect the ordinary atmosphere. You can as often as possible be given a wind heading and you will have a truly savvy considered how the atmosphere will change and what atmosphere can be ordinary with that wind bearing.

The average wind bearing that a range has for a particular time is known as the general wind. Right, when the curve is from the regular bearing then the atmosphere is generally common. Exactly when the twist moves a long way from the all-encompassing course then it consistently demonstrates atypical or advancing atmosphere.

Wind direction changes routinely run with changes in the atmosphere. The wind streams cyclonically around low-weight systems. If the wind modifies course in a cyclonic way it oftentimes infers a low weight or front is affecting the figure zone. A wind moving from the south routinely infers more smoking air is moving ever closer contort from the north as often as possible suggests cooler air is moving closer. Right, when the wind changes surprisingly it could be a frontal segment or wind direction change made by tempest surge.

Wind Speed (In Kmph): - Wind speed, or wind stream speed, is a basic climatic amount. Wind speed is brought on via air moving from high weight to low weight, more often than not because of changes in temperature. Wind speed influences climate determining, airplane and oceanic operations, development undertakings, development and digestion system rate of many plant species, and endless different ramifications.

The wind speed will accept a key part in the surface temperature in conditions where there is a strong temperature change with stature in the farthest point layer. In particular, a strong temperature sneaks past rate in the bit of the farthest point layer nearest the surface. The earth is warmed and cooled beginning from the most punctual stage. The wind mixes this air at ground level with air higher overtops. In the midst of the day when winds are light and the skies are clear, warmth will work at the surface. The temperature for this circumstance will tend to be sultrier than if the wind speeds were more grounded. This is in light of the fact that more grounded winds will mix the warm air near the surface with cooler air overhead.

On a fresh night, the opposite is the circumstance. Light winds amid the night will allow cool air to work at the surface. In case winds are more grounded than expected then the surface temperature will be more smoking since the wind will mix more sizzling air high up with the shallow cool air working at the surface. The wind speed is moreover basic in choosing the rate at which warm move in climate conditions will happen.

Let us take a graphical look at important parameters of Aarhus city weather dataset

DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

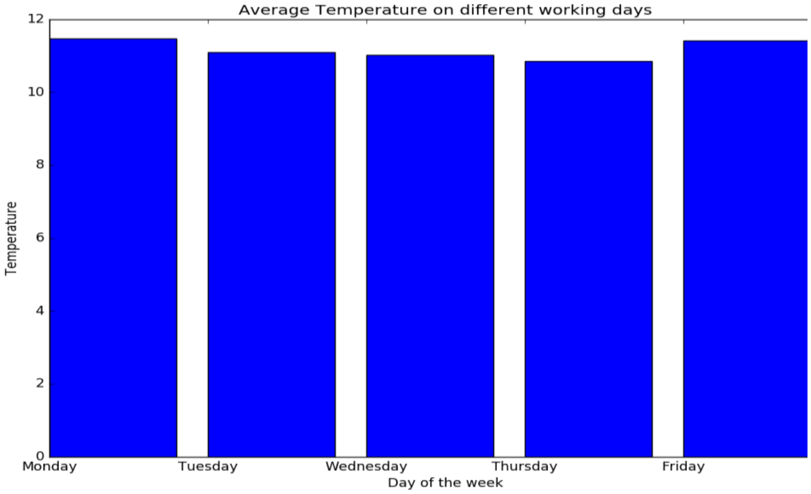


Fig 1: Average temperature on working days

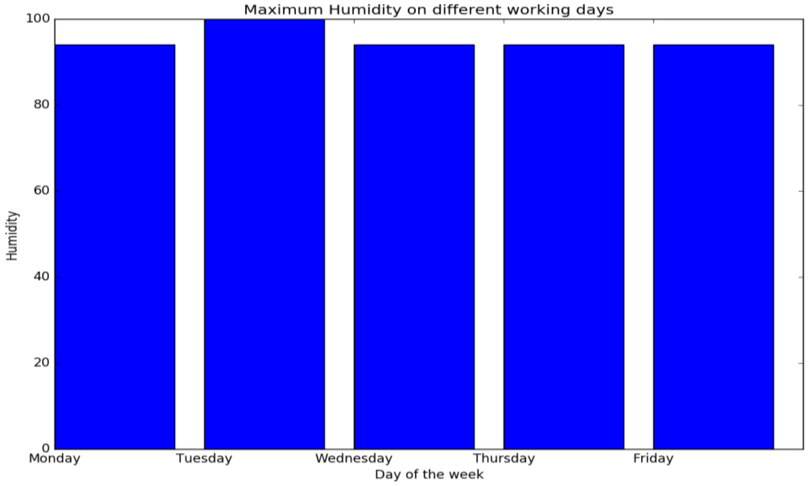


Fig 2: Maximum humidity on working days

- ## Your Method:

In this section we will understand the complete process and importance of all the steps used in method.

<u>DATA COLLECTION:</u> - Data collection is the route toward get-together and measuring data on elements of energy, in a set up efficient form that engages one to answer communicated inquire about inquiries, test hypotheses, and survey comes about. The data collection portion of research is essential to all fields of study including physical and humanistic systems, humanities, business, et cetera. While methodologies vary via prepare, the emphasis on ensuring accurate and reasonable collection proceeds as some time recently.

DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

The weather data for the city of Aarhus in Denmark is public for analyses purpose, available at Weather Data . The dataset is a collection of weather observations from the city of Aarhus. Measurements are recorded from February 2014 – June 2014 and August 2014 – September 2014. Weather data values will be analyzed on the basis of components like - Dew Point, Humidity, Pressure, Temperature, Wind Direction and Wind Speed.

For data analyses and exploration we will use **Apache Zeppelin** environment more specifically **Spark Module.**

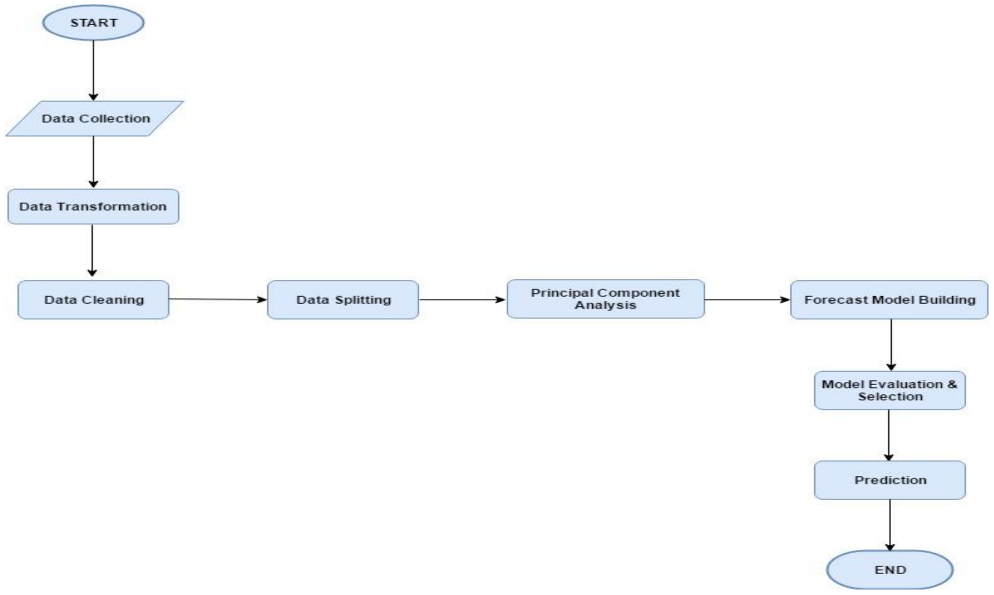


Fig 3: Algorithm

DATA TRANSFORMATION: - Data provided in online repository is in .Json file format. And in this project we will consider processing of only .Csv file format. So, we will perform transformation of. Json to .Csv file format.

Data transformation is the path toward changing over data from one association (e.g. a database record, XML file, or Excel sheet) to another. Since data often lives in different territories and designs over the attempt, data transformation is essential to ensure data from one application or database is conceivable to various applications and databases, a fundamental part for applications blend.

DATA CLEANING: - Data cleaning is the route toward recognizing and helping (or clearing) decline or mistaken records from a record set, table, or database and implies perceiving divided, misguided, off base or unessential parts of the data and after that supplanting, changing, or deleting the muddled or coarse data. Data cleaning may be performed keenly with data wrangling instruments, or as cluster get ready through scripting.

DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

<u>**DATA SPLITTING**</u>**: -** In our case Aarhus city data set will be divided in 60-40 ratio. In which **60%** will be **training part** and **40%** of data will go in **testing bucket.**

PRINCIPAL COMPONENT ANALYSIS: - Principal components analysis is a procedure for identifying a smaller number of uncorrelated variables, called "principal components", from a large set of data. The goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components. Principal components analysis is commonly used in the social sciences, market research, and other industries that use large data sets.

Principal components analysis is commonly used as one step in a series of analyses. You can use principal components analysis to reduce the number of variables and avoid multi co linearity, or when you have too many predictors relative to the number of observations.

MODEL BUILDING: **-** Model Building–choosing predictors–is one of those skills in statistics that is difficult to teach. It's hard to lay out the steps, because at each step, you have to evaluate the situation and make decisions on the next step.

If you're running purely predictive models, and the relationships among the variables aren't the focus, it's much easier. Go ahead and run a stepwise regression model. Let the data give you the best prediction.

But if the point is to answer a research question that describes relationships, you're going to have to get your hands dirty.

It's easy to say "use theory" or "test your research question" but that ignores a lot of practical issues. Like the fact that you may have 10 different variables that all measure the same theoretical construct and it's not clear which one to use. Or that you could, theoretically, make the case for all 40 demographic control variables. But when you put them all in together, all of their coefficients become non significant.

All the variables in Aarhus city dataset are of type continuous. We will build **linear model** for prediction of response variable. And we will also perform **time series forecasting,** because at the end we want to outperform competitor's statistical method. Let's take a look at some important steps while statistical model building.

Linear Model:
A). Perform "Kstest" to check whether environmental variables are normally distributed or not. If not then normalize the variables first.
B). Visualize the linear relationships in the dataset.
C). Perform "Pearson correlation" to get an idea of relationship between dependent and independent variable and it also informs about weighting coefficient of independent variables.
D). Finally, design the linear model with appropriate variables.

Time Series Forecast:
A). First convert "datetime" variable into index column.

B). Check stationarity of time series i.e. constant mean, constant variance & auto covariance that does not depend on time.

C). Eliminate trend and seasonality.

D). Perform time series forecasting.

PREDICTION: **-** After selection of best model. We will use that model on **testing data set** for prediction of **Air Temperature,** which is our response variable**.**

- ## Quantitative Results 1:

Results from linear model:

KSTest Results: Pvalue of all the variables is 0. Variables are normalized.

A). Dewpoint: KstestResult(statistic=0.3729231258446617, pvalue=0.0).

B). Humidity: KstestResult(statistic=1.0, pvalue=0.0).

C). Temperature: KstestResult(statistic=0.651, pvalue=0.0).

D). Pressure: KstestResult(statistic=1.0, pvalue=0.0).

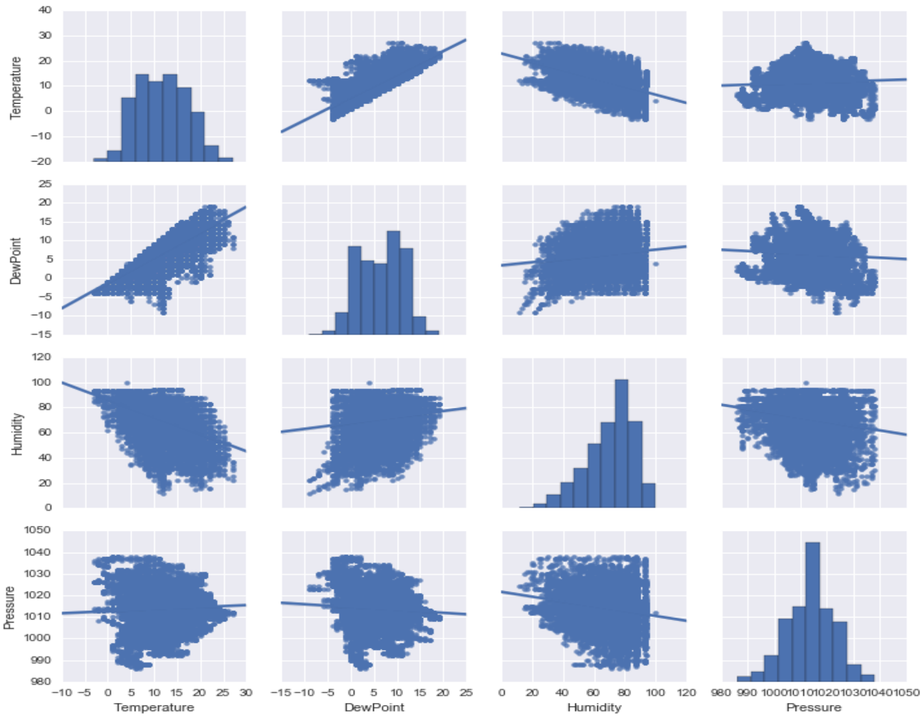
Pearson Correlation: Positive relation of **temperature** with **dewpoint & pressure**. Negative relation with **humidity** in the atmosphere.

A). Dewpoint: PearsonResult(coefficient= 0.78541330434839418, pvalue = 0.0).

B). Humidity: PearsonResult(coefficient= -0.47275732315191615, pvalue = 0.0).

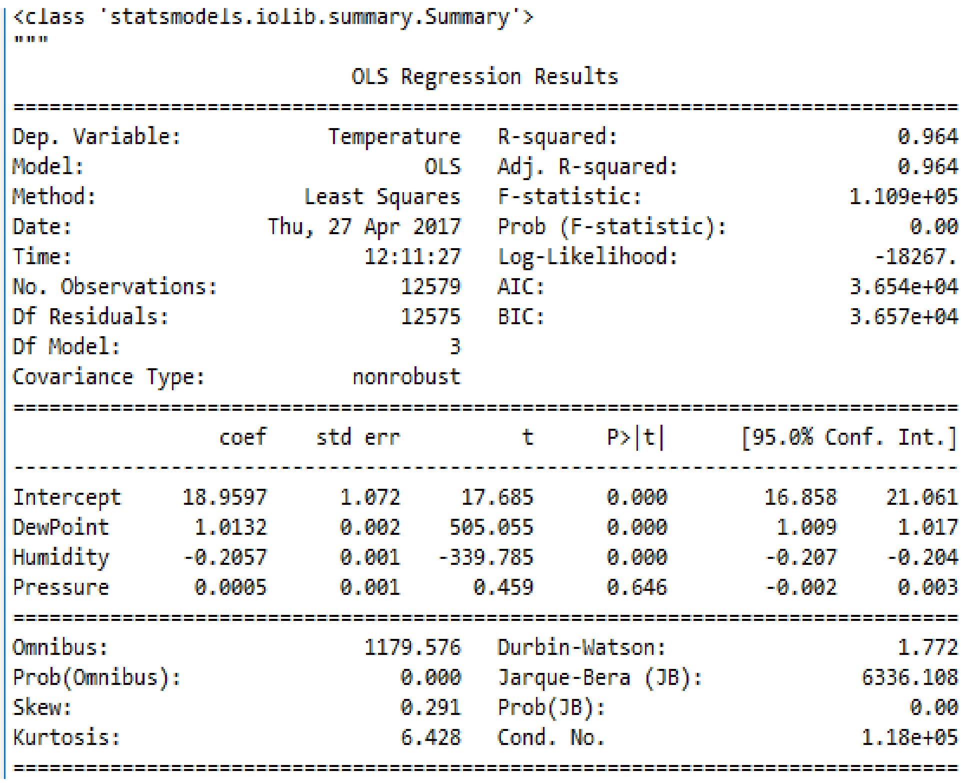
C). Pressure: PearsonResult(coefficient= 0.058368756214757185, pvalue = 0.0).

Correlation Plot:



DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

OLS Regression Results:

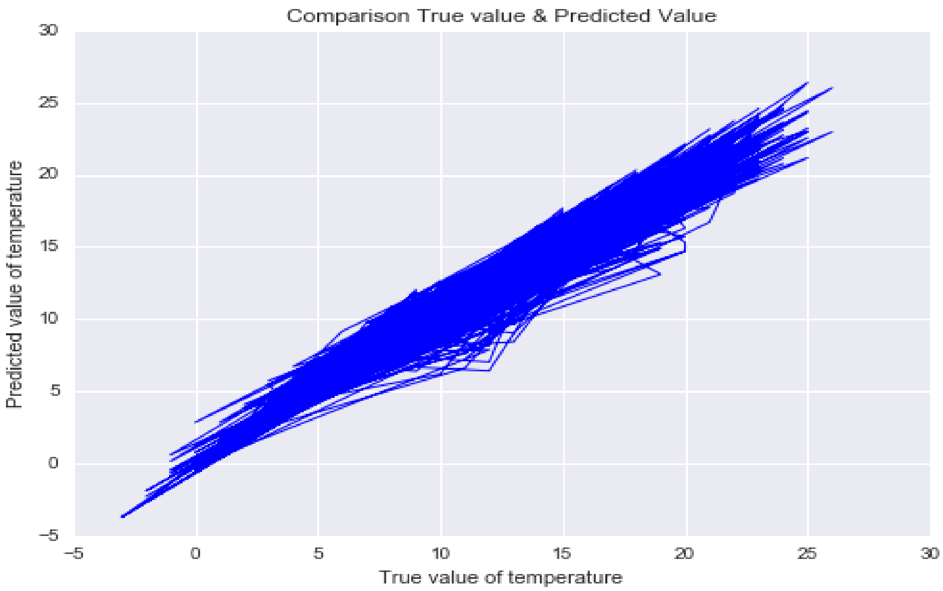


```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:            Temperature   R-squared:                       0.964
Model:                            OLS   Adj. R-squared:                  0.964
Method:                 Least Squares   F-statistic:                 1.109e+05
Date:                Thu, 27 Apr 2017   Prob (F-statistic):               0.00
Time:                        12:11:27   Log-Likelihood:                -18267.
No. Observations:               12579   AIC:                         3.654e+04
Df Residuals:                   12575   BIC:                         3.657e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     18.9597      1.072     17.685      0.000        16.858    21.061
DewPoint       1.0132      0.002    505.055      0.000         1.009     1.017
Humidity      -0.2057      0.001   -339.785      0.000        -0.207    -0.204
Pressure       0.0005      0.001      0.459      0.646        -0.002     0.003
==============================================================================
Omnibus:                     1179.576   Durbin-Watson:                   1.772
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6336.108
Skew:                           0.291   Prob(JB):                         0.00
Kurtosis:                       6.428   Cond. No.                     1.18e+05
==============================================================================
```

Comparison Predicted Temperature & True Value:

A). Variance Score: Variance score: 0.96.

B). Mean Squared Error: Mean squared error: 1.15.

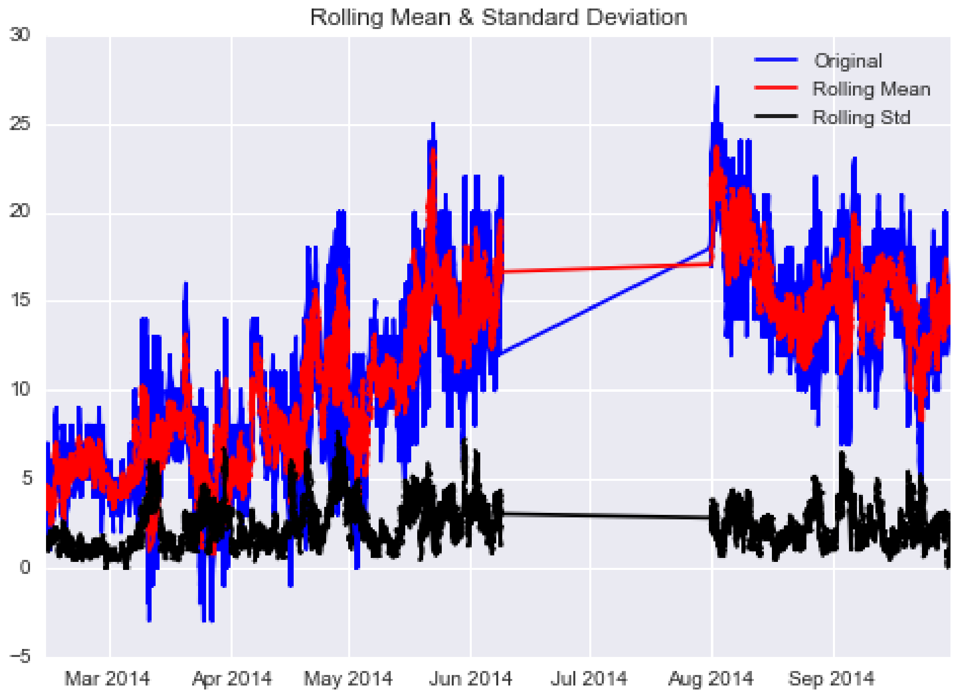


DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

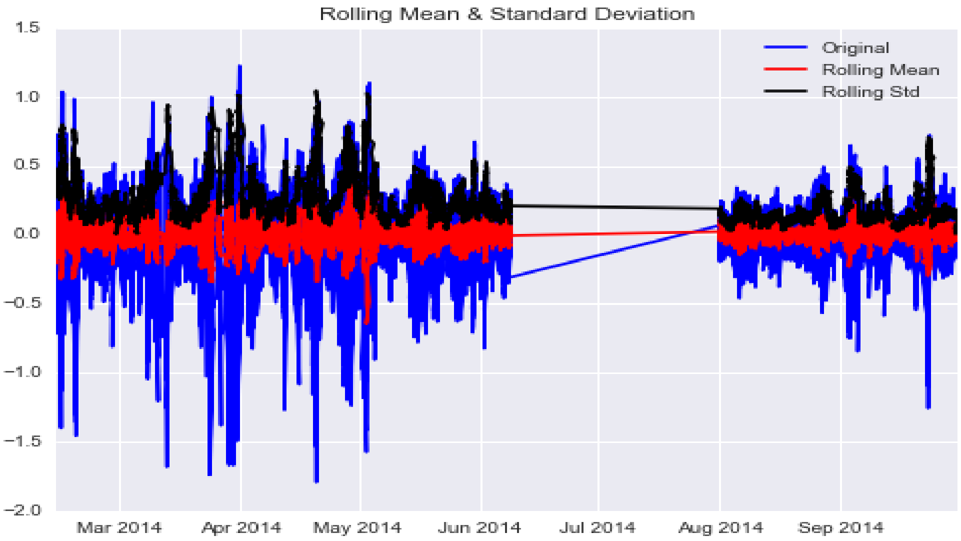
- **Quantitative Results 2:**

Results from Time Series model:

Stationarity test: From stationary test graph we can observe that standard deviation is stationary but mean is not stationary.

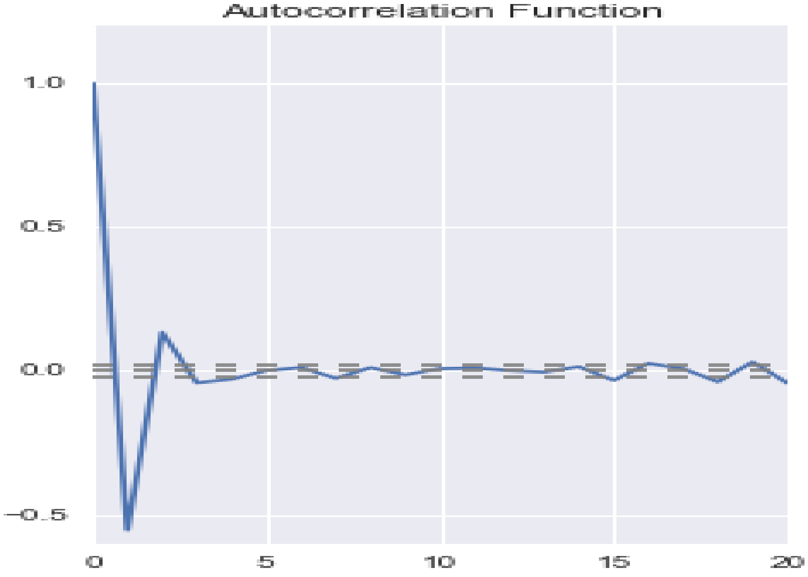


Make rolling mean stationary: From stationary view graph we can observe that mean is now stationary.

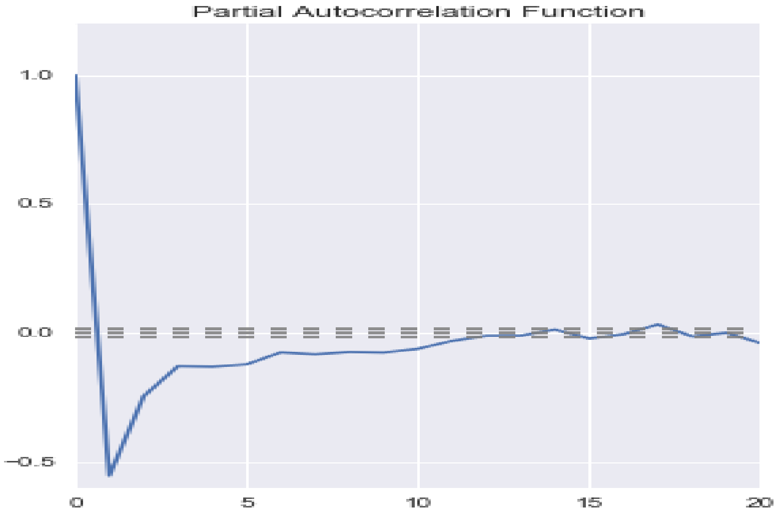


DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

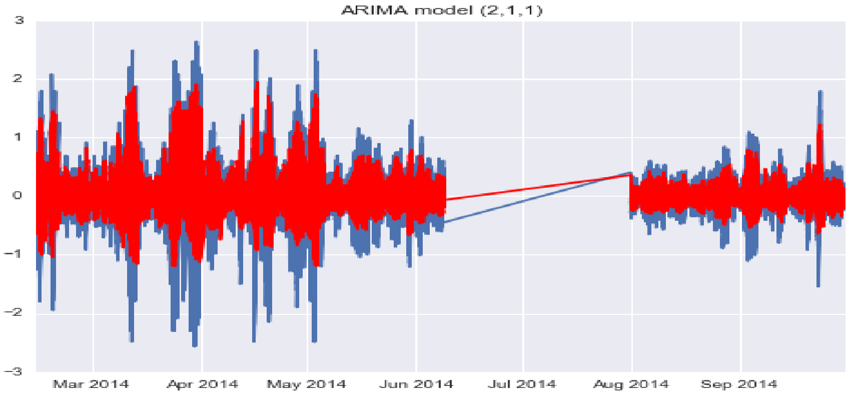
ACF Plot: From plot **we can notice value of q = 2.**



PACF Plot: From plot **we can notice value of p = 1.**



ARIMA Model:



DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

RMSE Score: Root mean square error: 20.01.

Prediction of time series model is not that good as we can see that RMSE value is too high than the Linear Model designed.

- ## Discussion: Comparison With Your Competitor:

Competitor Results	Your Results
The forecasting method used has **significant errors at long-range period (more than 20% in the mean).**	My algorithm has used short range forecasting method. That has **error less than 5% in the mean.**
Pearson product-moment correlation coefficient is greater than 0.8 and less than 0.85 in absolute value.	**Pearson product-moment correlation coefficient calculated using Zeppelin IDE is greater than 0.85 (chances for improvement).**
Recent research based on the inductive methods showed possibility of the long-range (half-year lead-time) forecast with **mean absolute error (MAE) up to 8 degree Celsius.**	**Mean absolute error (MAE) of my algorithm is 5.7 degree Celsius. And there are chances of improvement.**
The model designed is combination of air temperature time series analysis for different places. So the **accuracy depends on the number of places into consideration.**	Here, multi linear regression analysis is performed. That considers relations between variables like: Temperature, Humidity, DewPoint. That improves the **accuracy of model to 85 percent.**

- ## Performance on Big Data: Time Measurements:

Operation	Time Taken
Loading and cleaning the data.	90 Seconds.
Data Transformation and analysis.	60 Seconds.
Linear Model building & analysis.	180 Seconds. (approx)
Stationary test of data.	90 Seconds.
Make data Stationary.	50 Seconds.
Time series forecasting & analysis.	300 Seconds. (approx)

- ## Conclusion:

After the use of short-range forecast method there is significant improvement in the error rate of prediction. And consideration of relationship between different independent variables gave a good platform for prediction of **air temperature**.

DS 670 – Assignment 14 – Teleprompter Script
Weather Data Analysis of Arhus City in Denmark

A). Performance of linear model is much better than time series forecasting method proposed by competitor.

B). For now **dewpoint, humidity & pressure** are considered as independent variables.

C). The long range forecasting method has **significant errors at long-range period (more than 20% in the mean).**

D). Short range forecasting method is much reliable. That has **error less than 5% in the mean**.

E). You can’t consider just temperature, or relative humidity, or wind direction alone when trying to make a forecast. You have to understand each measurement and what it could mean to the larger weather picture. Only then can you put them together to make an accurate forecast.

F). Generated variables could be the future enhancement for this project.