Final Capstone Project (CKME 136)

# Introduction

Weather is always in the news and makes headlines because it determines a person’s day outdoors. The dataset I have chosen represents the weather of New Delhi, the capital of India, on a daily basis from 1996-2017. Using classification, regression we can predict the weather patterns in the coming years in New Delhi. Are the weather conditions interrelated? What will be the highest and lowest temperature point in the coming years? What will be the weather in the future (beyond 2017)? Tools such as R, multivariate regression, correlations shall help and help in identifying patterns, trends, insights, analyses and predictions.

# Literature Review

Article 1

Based on the study conducted from the article, Trend analysis for rainfall in Delhi and Mumbai, India, in the year 1951-2004, it was found the precipitation level in New Delhi had an insignificant decrease in the monsoon season (July - October) while in the winter time and pre-monsoon season there was an increase in perception. However, based on statistical analysis, the authors found out that daily rainfall rate per year had significantly decreased followed by a decrease amount of rainfall on a average maximum rainfall in a day.

Article 2

In this article, it was found that there is a relationship between ozone concentration levels and humidity. This relationship focuses on Pune (city in India) and New Delhi, but I will focus on just New Delhi. The authors found out that between 1990 – 1999, in New Delhi the ozone concentration levels in New Delhi reach a level of 0-5 ppb in all seasons which indicates a low level of concentration due to the high humidity levels. Abhishek Pathak CKME 136 October 10th, 2017 Regression analyses were conducted of ozone concentration with maximum temperature, relative humidity and air pollution to test this relationship. Based on the results it can be concluded that when humidity increases the removal of ozone become effective thus decreasing the rate of the creation of ozone concentration.

**Bibliography**

Ali, K., Inamdar, S. R., Beig, G., Ghude, S., & Peshin, S. (2012). Surface ozone scenario at pune and delhi during the decade of 1990s. Journal of Earth System Science, 121(2), 373-383. doi:http://dx.doi.org.ezproxy.lib.ryerson.ca/10.1007/s12040-012-0170-1 D

Rana, A., Uvo, C. B., Bengsston, L., & Sarthi, P. P. (2011). Trend analysis for rainfall in Delhi and Mumbai, India. Climate Dynamics, 38(1-2), 45-56. Doi:10.1007/s00382-001-1083-4

# Dataset

The dataset that is being used is the weather in New Delhi, on a daily basis from 1996 – 2017 (source: https://www.kaggle.com/mahirkukreja/delhi-weather-data/version/2). In this dataset there are 20 attributes.

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute** | **Description** |
| 1 | Date & Time | This indicates the date and time of the weather |
| 2 | Condition | What time of weather condition it is (eg: overcast, haze, cloudy) |
| 3 | Dew point | Temperature based on pressure and humidity which causes water droplets to form |
| 4 | Fog | Percentage of fog occurring during the day |
| 5 | Hail | Percentage of hail occurring during the day |
| 6 | Heat Index | Temperature that factors in the humidity plus the actual air temperature |
| 7 | Humidity | Percentage (chance) of how hot it will be during the day |
| 8 | Precipitation | Percentage (chance) of rain during the day |
| 9 | Pressure | Air pressure in (kPa) during the day |
| 10 | Rain | The amount of rain (mm) in a day |
| 11 | Snow | The amount of snow (cm) in a day |
| 12 | Temperature | The temperature during the day |
| 13 | Thunder | Percentage (chance) of thunderstorms occurring during the day |
| 14 | Tornado | Percentage (chance) of tornado ocuuring during the day |
| 15 | Visibility | Measurement in (km) of how far can one clearly see |
| 16 | Wind direction speed | Wind speed with wind gust |
| 17 | Wind direction | The direction of the wind |
| 18 | Wind gust | Increase speed (km) of the wind |
| 19 | Wind chill | Lowering of air temperature caused by the wind |
| 20 | Wind speed | Actual speed of the wind without wind gust |

* Out of the attributes listed above, I do not plan on using attributes # 6, 8,17, 18,19.
* After checking the class of this dataset in R, it was determined that the class was a data frame.

# Approach

# Step 1: Cleaning the dataset

Cleaning the dataset is done to help map raw data into a format where the data is presented without any inaccuracies and biases. This can contain various ways such as finding if there is any extra data which is not required; ensuring that computations can be carried out easily, the fields/attributes have the appropriate classes. The most important aspect of cleaning the data is to organize the data in such a way that it can be fitted into a consumable structure which can be used for data analysis, aggregation and visualization. In this dataset, there are quite a few blanks and NAs which need to be removed thereby restricting me to carry any analyses. Furthermore, certain visualizations cannot be presented due to the NAs or blanks present. Having blanks or NAs also add bias to the dataset and must be removed in a systematical manner as to not be detrimental to the existing data. Further, to compute calculations and for visualizing them in certain graphs – for instance, the class for these attributes would have to be changed from its default class numerical to factor (especially for the continuous variables). It must also be noted that cleaning also involves ensuring that column headings are clear, so that it is easier to understand when reading the data thus making the data more presentable.

# Step 2: Relationships/Trends, Analysis and Predictions

After the data has been cleaned, identifying/finding trends and relationships among the attributes is important. Some examples which could give more insight into the dataset include: • Condition and dew point (2 attributes);

• Humidity and date & time (2 attributes);

• Snow and temperature and date & time (3 attributes);

• Condition and temperature and date & time (3 attributes);

• Rain and thunder and temperature and date & time (4 attributes);

These and many other relationships can be further analyzed through correlations and correlations matrices. This would allow to provide more data insight and assist in predicting the various factors relating to the temperature in the coming years. For this dataset, the plan is to use multivariate linear regression. As majority of the data is numerical in nature, performing regression testing shall help provide better insight. Using multi-variate linear regression shall assist in in giving a better insight into the relationships of the fields/attributes. This shall allow us to identify the manner in which the various relationships are shaped – in terms of the shape of the graph (normal curve or any skews), whether it has a positive or negative slope or whether is totally horizontal or vertical, i.e. no relationship. In the case of the multivariate linear regression, temperature shall be the dependent variable while attributes such as condition, humidity, dew point can be the independent variables. This shall help establish the relationships among such variables and give better data insights. There are also some variables or attributes that are binary in nature, i.e. either zeros or ones. Some of these attributes are fog, hail, rain, snow, etc. Regarding predicting the data, on how the temperature will be in the coming years based on the various attributes, a training dataset and testing dataset will have to be chosen. Using the predict and forecast function, for time series analysis, shall assist in predicting the data in the coming years. Also, cross validation techniques can be used to further validate the data.

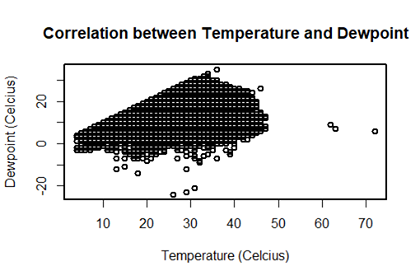
# Step 3: Visualizations

As mentioned in Step 2, after performing the various analyses, creating visualizations in this dataset shall help in portraying as well as understanding the trends and relationships of the fields/attributes. Some of the visualizations can be but not limited to are line graphs, box plots, bar plots, etc. For multivariate linear regression, visualizations shall help in showing the various points on the graph and how each attribute affects the different points. Regarding the time series analyses, cross validation graphs can also be shown to indicate the trends and confirm the validity of the predictions.

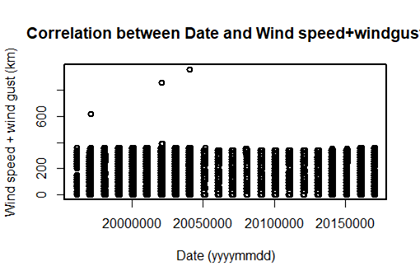
# Results

In this dataset, it is important to note that all the analyses has been carried out in R. Further, understanding the trends, relationships among the variables have been ascertained. From the analyses performed, the results indicate that temperature is mainly independent, i.e. does not work in tandem with other attributes.

After carrying out a correlation among various sets of attributes, I found that the correlation between temperature and dewpoint were the highest among all attributes with a correlation of 0.50 which indicates this to be a medium correlation. This is not particularly a strong correlation but it does indicate that almost half of the time, when temperature increases, dewpoint increases and vice versa. It reverts back to point that temperature seems to be mainly independent of the other attributes within this dataset.



An example, of a weak correlation, is between date and wind direction where the correlation resulted in 0.17. The main reason some of the correlations are weak in nature are because these attributes are portrayed as 0’s and 1’s, i.e. 0’s indicating that the event will not occur and 1’s indicating the event will occur, thus making it hard to correlation amongst variables. Below is a depiction of a weak correlation:



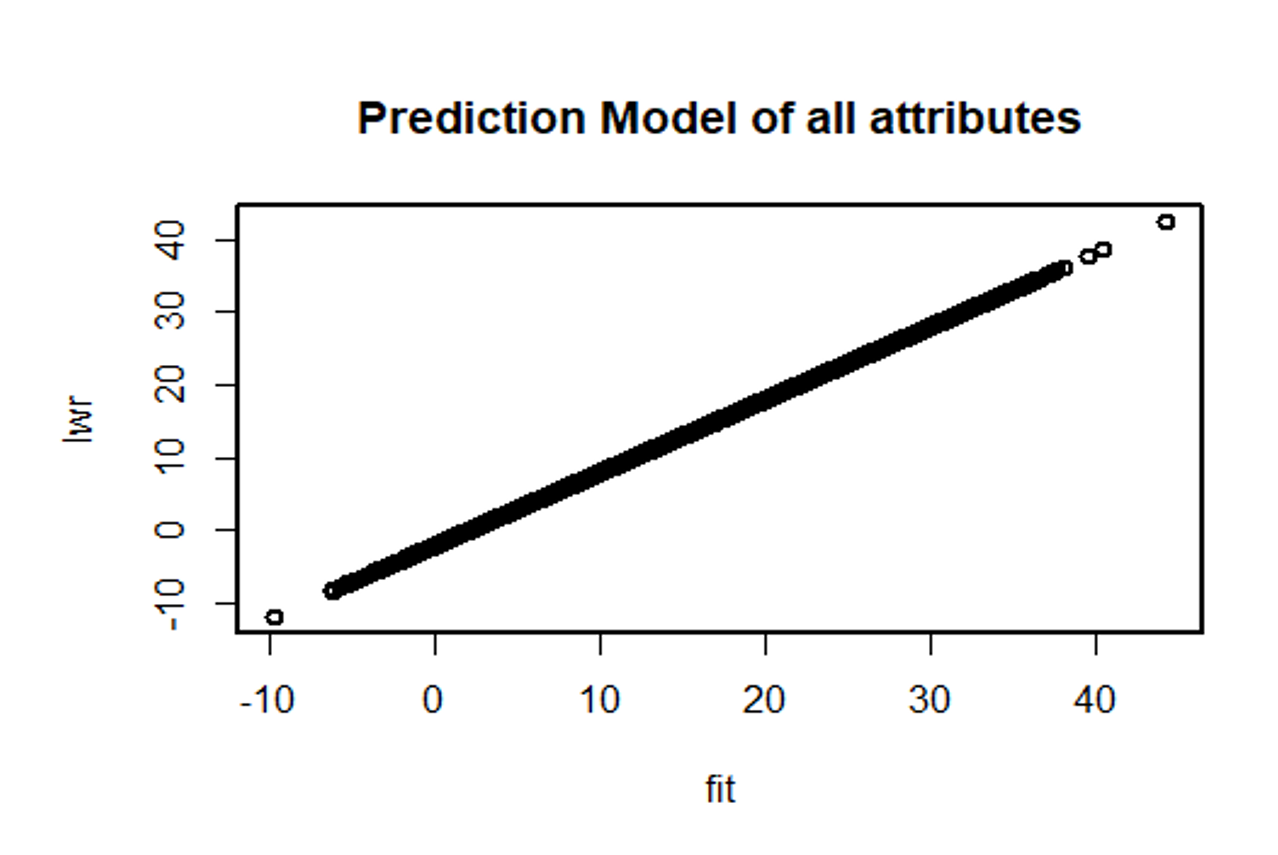
In terms of applying a machine learning technique and creating models, Linear Regression has been used. As almost 90% of the dataset was continuous in nature, and the purpose was to predict the temperature in the coming years for New Delhi, I decided to use multivariate linear regression.

Initially I had created multiple models with two, three attributes. It turned out that the multiple R squared for those models was low which means they are ineffective. Using just two – three attributes in a linear regression model causes bias which would indicate that the models are portraying inaccurate results.

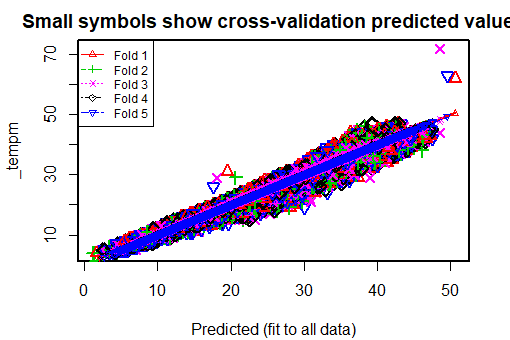
After giving a lot of thought, I decided to make one model (i.e. that consisted of all the attributes. But before creating the model, I had to separate the date and time column to avoid encountering any errors and plus to get a much more accurate prediction.

Following that, I created and predicted the model and it turned out the model had multiple R squared of 0.96 which indicates that the model explains 96% of the variance in the dependent variable.

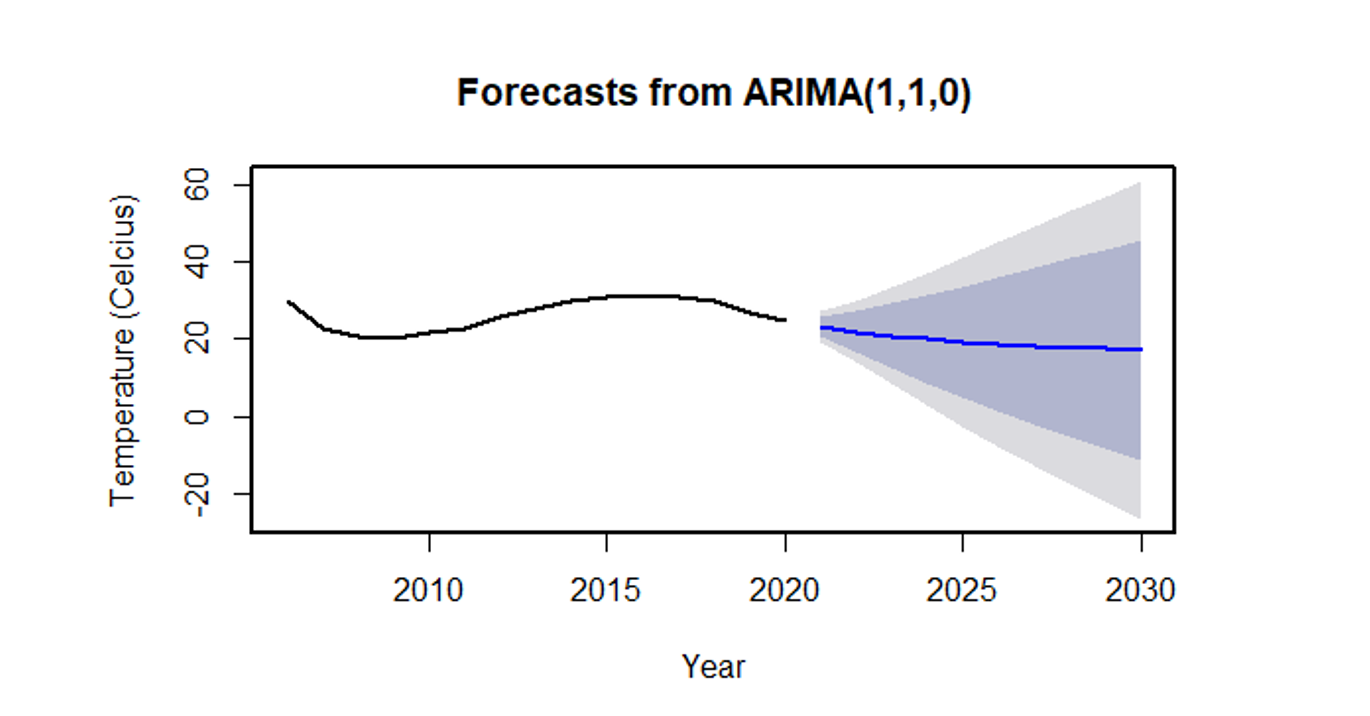
Below is a graph which highlights the linearity of the various attributes used, keeping temperature as the dependent variable.



To further verify the accuracy testing and training models, I used the cross-validation function. In R, with a package known as’ DAAG’, I installed the package which assisted me in the cross-validation technique known as k-fold cross validation. According to the graph below, as the x-axis increases, y-axis also increases, thereby showing that the cross-validation model matches the prediction model i.e. having a reasonable accuracy level. In addition, the 5 folds are all one above another, clearly indicating the accuracy of this prediction.



Having a decently reasonable r squared value for my linear regression and the cross validation seems to matching the trend / path of the graphs, I felt confident in wanting to predict the future temperatures for New Delhi. To predict the weather for New Delhi in the future I used a similar function to predict, i.e. the forecast function. The forecast function is a subset of the prediction function and it focuses mainly on time-series. Based on the graph below, the light grey area in the graph is the 95% interval and purplish-blue area is the 80% interval consisting of the predicted values.

To further understand the graph below in a thorough manner, from mid-2009 the temperature slightly decreases but increases from 2011 onwards. As around mid 2018 approaches, the temperature seems to be slightly decreasing again and keeps decreasing very marginally and even plateaus after the year 2020.

# Conclusions

Hence, in conclusion, the temperature in New Delhi beyond 2018 decreases and then after the year 2020 it plateaus / decreases marginally. In regards to other factors affecting the temperature, the only factor that had the most effect was Dewpoint (with a correlation of 0.5). Although there can be others that affect the temperature, within this dataset, after trying to identify one with a strong relationship was unfortunately not found. Using techniques such as linear regression, cross validation and then the forecast function helped in identifying the level of accuracy rate, especially for the immediate years (i.e. from 2020 to 2023) as most of the predicted values are in the narrow purplish blue area along with the narrow grayish area (as per the graph above).

This report, along with the R code is on Github: <https://github.com/abhiryerson/ryerson-ckme136>