Using Time Series Analysis to Identify Climate Trends and Potential Impacts to Local Bird Species in Delhi, India

Christopher Robinson, Hanmaro Song, Nima Amin Taghavi

University of San Diego

ADS-506-02-FA21 - Applied Time Series Analysis

Abstract

Global warming is a major concern, and as time progresses global warming leads to climate change. Climate change is thought to be the cause of many challenges, such as rising sea levels and drought, affecting ecosystems around the globe. Although humans can likely adapt to changing environments, plants and animals may not be able to adjust as easily. In this paper weather data from 1973 to 2021 for Delhi, India was utilized to explore potential patterns and trends for temperature, humidity, pressure, and wind to attempt to identify changes in climate over time. Local bird populations in Delhi, India are already in decline and climate change has been suspected of being a primary cause. According to Padmaparna Ghosh, "Indian bird experts agree that it is crucial to start integrating the impacts of climate change on birds in India" (2018). This project attempts to determine the validity of these claims. ARIMA models were created for variable of importance in order to forecast values based on past data. In addition, machine learning model(s) were created in an attempt to forecast future climate trends in order to predict potential challenges to local bird populations. While seasonal weather patterns in Delhi have been consistent, overall temperatures have been steadily increasing since the 1970's. The data suggests a slight upward trend in overall annual temperature in Delhi. While a small trend over decades may seem insignificant, according to the climate reality project, "one degree can have a huge effect on a global scale in the natural world" (2021). Research has shown that Indian bird species are struggling; However, the time series data analysis done for this project proved inconclusive and could not prove with any certainty that climate change may be a primary factor. India is a rapidly developing country and Delhi is one of the country's largest urban centers. Due to this fact, there may be other factors which are having a greater impact on local bird species than climate change, such as changing land use, direct pollution, and competition for resources.

Background

Global climate change has become an increasing concern and is thought to be responsible for changes in local climates which have negative impacts on local flora and fauna. This paper focusses on local climate conditions in Delhi, India and what effects, if any, they could have on local bird populations. As the population in Delhi, India grows each year, so does traffic congestion and industrial processes which are leading contributors to air pollution. In this project, we will evaluate climate data from Delhi, India to evaluate patterns in climate change that could pose risks to local bird populations. The conditions we are focusing on are temperature, humidity, wind, and barometric pressure, all of which are used to indicate warming and cooling trends.

Literature Review

In preparation for the analysis, several peer reviewed sources were explored to investigate what research has been done on the subject and identify any information which may be helpful in our work. As stated previously, this paper focuses on the effects of climate change in Delhi, India and its potential impacts on local bird species. Birds are especially vulnerable to changes in climate and air pollution and it is important to look at how climate change is affecting local species in order to present an overall assessment on the effects global climate change may have on bird populations worldwide. In the paper, “Long-term time series of ornithological data”, Moller and Hochachka explore the diversity of data sources that can be used to study the long-term effects of climate change on birds. The paper discusses the use of local studies, such as this one, to build a robust database of climate and avian data, including nest record schemes, population studies, bird surveys, bird atlases, bird ringing and observation depositories, and museum collections. Unfortunately, coordination on this effort has been inconsistent.

According to Moller and Hochachka, “Numerous data allow studies of the effects of climate change on birds. There have been few attempts to coordinate databases” (2019). As more studies are made available perhaps a better understanding of the implications of climate change will become more widespread and mutual interest will facilitate more cooperation between groups.

There have been many research studies conducted about how the change in climate temperature could lead to unforeseen events such as erosion and accretion in land dynamics as well as the alteration in species diversity and productivity. One past study, “Climate change impacts on Indian Sunderbans: a time series analysis” discusses how climate change is affected surrounding lands. The increase of downstream salinity due obstruction in upstream has led to decrease in transparency of water causing decrease in phytoplankton and fish, density and diversity in the central sector of Indian Sunderbans. Based on this study, we could easily come to the conclusion that not only does temperature effect the environment and non-living things but also living organisms such as plankton, fish, and birds.

In addition to literature focused solely on environmental factors, it is important to look at past research and methodologies used to quantify climate change in general. In the paper “Generating a Set of Temperature Time Series Representative of Recent Past and Near Future Climate”, the author discusses a methodology for building a large sample of temperature specific indicators based on the decomposition of the time series into deterministic parts, such as seasonality and trends. While this study focused on the interrelation between climate and energy systems, the methodology for identifying trends in temperature was of specific interest and this information could also be useful in other areas or study as they relate to climate change.

Climate change is a difficult area of study as there is yet to be a consensus on the underlying causes. Some say climate change is a naturally occurring process, which undoubtedly to some extent it is. Even if climate change does occur naturally, it does not negate the fact that human intervention may be contributing to unnatural levels of change. Data is the driving force behind the study of climate change, and climate science is a field focused on studying large-scale changes. In contrast, according to James Faghmous and Vipin Kumar, "some consider shorter (weather) time scales such as days or weeks to also be part of studying climate" (2014). This paper will not be looking at shorter durations in the data, but it is worth noting that some environmental interactions may only last hours or days, such as the influence of ocean surface temperatures on hurricanes. Other interactions might occur over several years, such as rising sea levels or desertification, which would be of greater concern for a study such as this.

Data

The data was collected through "https://weather.visualcrossing.com" using json to extract a sample dataset containing 17,866 records from 1973 through 2021. The Json code is contained in appendix A and extracts the basic fields used in the analysis such as "datetime", "tempmax", "tempmin", "temp", "humidity", "precip", "windspeed", "pressure", "visibility", "conditions", and "cloudcover". Addition fields were created for the analysis including "ID", "season", "month", "year", "decade", "conditions_scaler", "season_scaler", and "month_numeric". The dataset fields and types can be seen in figure 1.1.

Figure 1.1 – Dataset Data Dictionary

Field Name	Field Type	Description
ID	Numeric	Record ID
datetime	Date	Recorded data
tempmax	Decimal	Maximum temperature
tempmin	Decimal	Minimum temperature
temp	Decimal	Average temperature
humidity	Decimal	Average humidity
precip	Decimal	Total percipitation
windspeed	Decimal	Average windspeed
pressure	Decimal	Average pressure
visibility	Decimal	Visibility (Miles)
conditions	Text	Weather conditions
conditions_scaler	Numeric	Weather conditions (Numberic)
cloudcover	Numeric	Cloud cover (Percent)
season	Text	Season
season_scaler	Numeric	Season (Numeric)
month	Text	Month
year	Text	Year
decade	Text	Decade
month_numeric	Numeric	Month (Numeric)

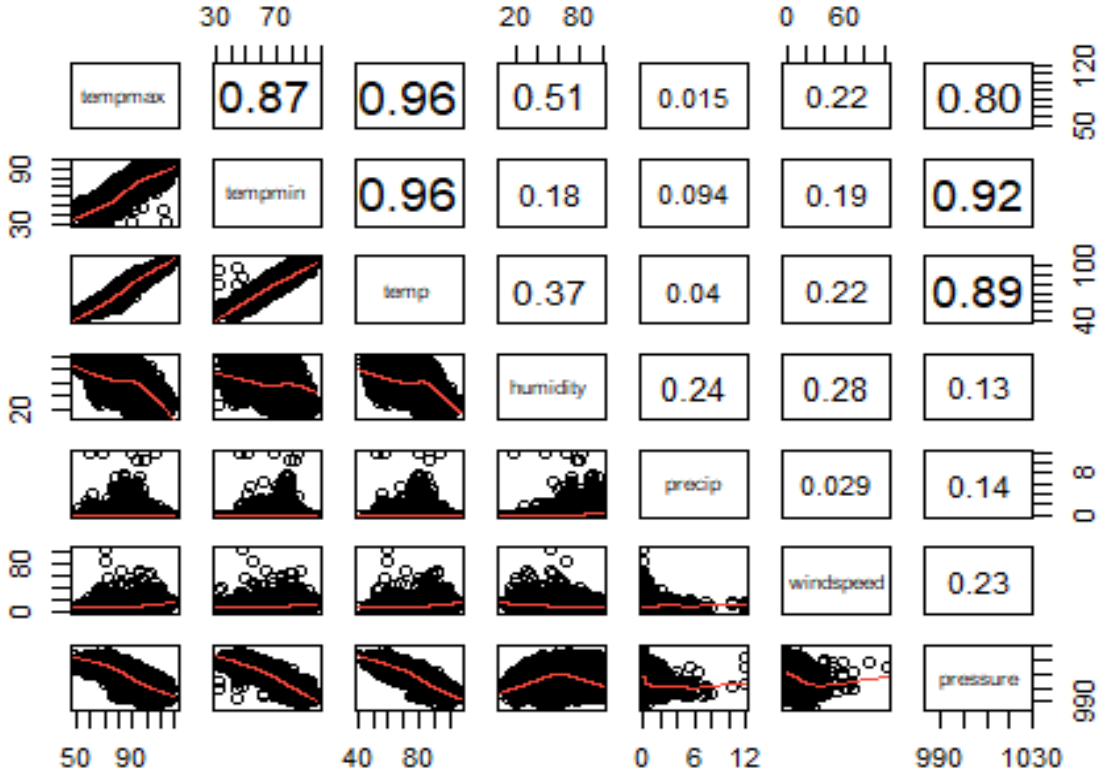
The dataset was relatively clean with very few null values. The few nulls that were present were interpolated using a 6-day average. For example, if a "temp" value is missing, the data from 3 days prior and 3 days after will be added together and then divided by 6 to find the average "temp". Identifying correlated features can help eliminate redundant information in order to simplify the model.

Figure 1.2 – Null Values (%)

tempmin	0.04%
tempmax	0.04%
temp	0.04%
humidity	0.04%
pressure	0.07%
windspeed	0.04%

Figure 1.3 shows correlation between each pair of features. Based on the table, it suggests that there is strong positive correlation between the temperature columns (min, max, mean) and pressure (all above 0.8). The next largest correlation is between humidity and temperature. Precipitation and windspeed has slightly higher correlation than neutral.

Figure 1.3 – Correlation Table



For this analysis, temperature produced consistent results between all three fields (min, max, avg), so maximum temperature was chosen to represent the temperature variable.

Figure 1.4 – Average Temperature

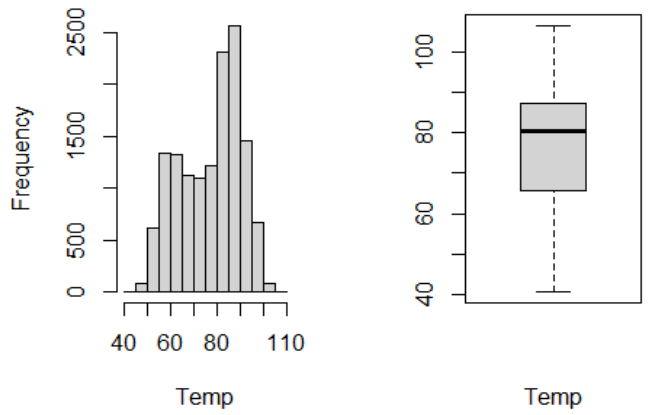
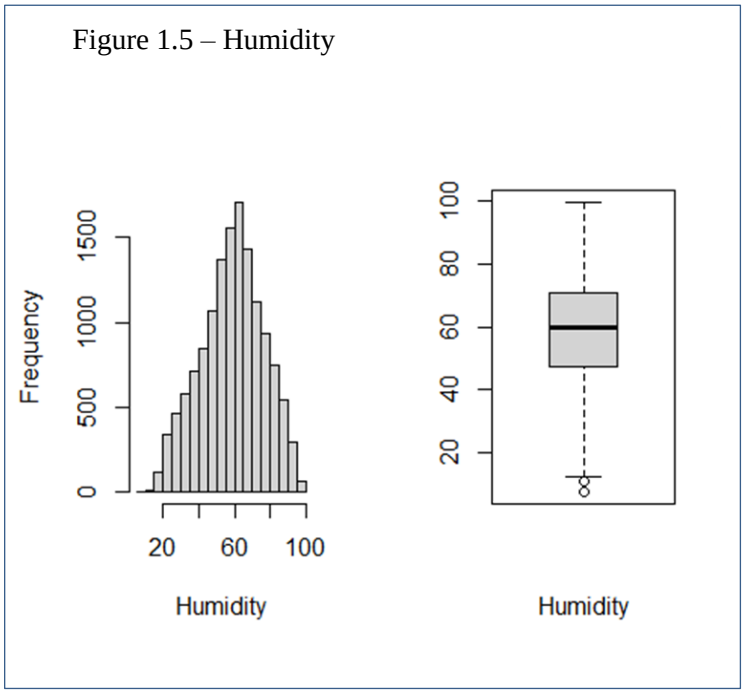


Figure 1.4 shows both the histogram and boxplot of average temperature which is bi-modal with a spike in at around 60 degrees and again at 90 degrees. The mean was around 80 degrees and there were no outliers in the dataset.

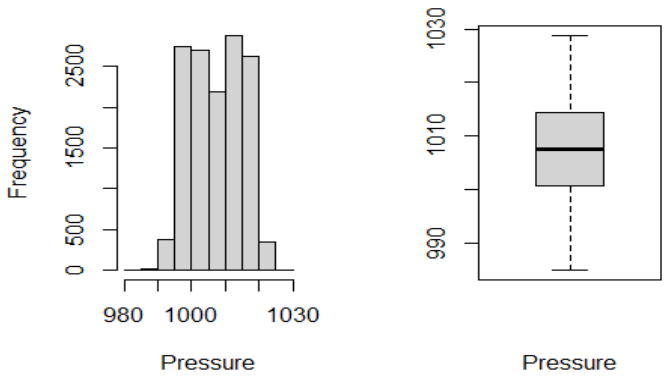
Based on the figure 1.5, the humidity shows data points quite normally distributed without any skewness. It has the mean value at around 60%. There were some outliers towards the bottom of the scale, which were likely from "zero" readings.

Figure 1.5 – Humidity



The pressure variable was bimodal with an initial spike at approximately 1000kPa and another at 1025kPa. Figure 1.6 shows the median pressure measurement was around 1008kPa. Otherwise, the data was very well distributed with no outliers.

Figure 1.6 – Pressure



The windspeed variable was highly skewed right with a median around 20mph. Figure 1.7 indicates there were some outliers with values extending to over 100mph. Likely these outliers were due to storm conditions as opposed to inaccurate measurement so they were not altered or removed.

Figure 1.7 – Windspeed

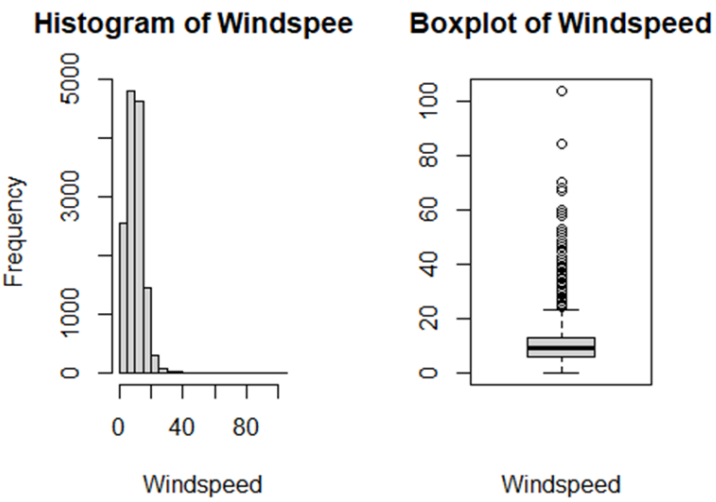
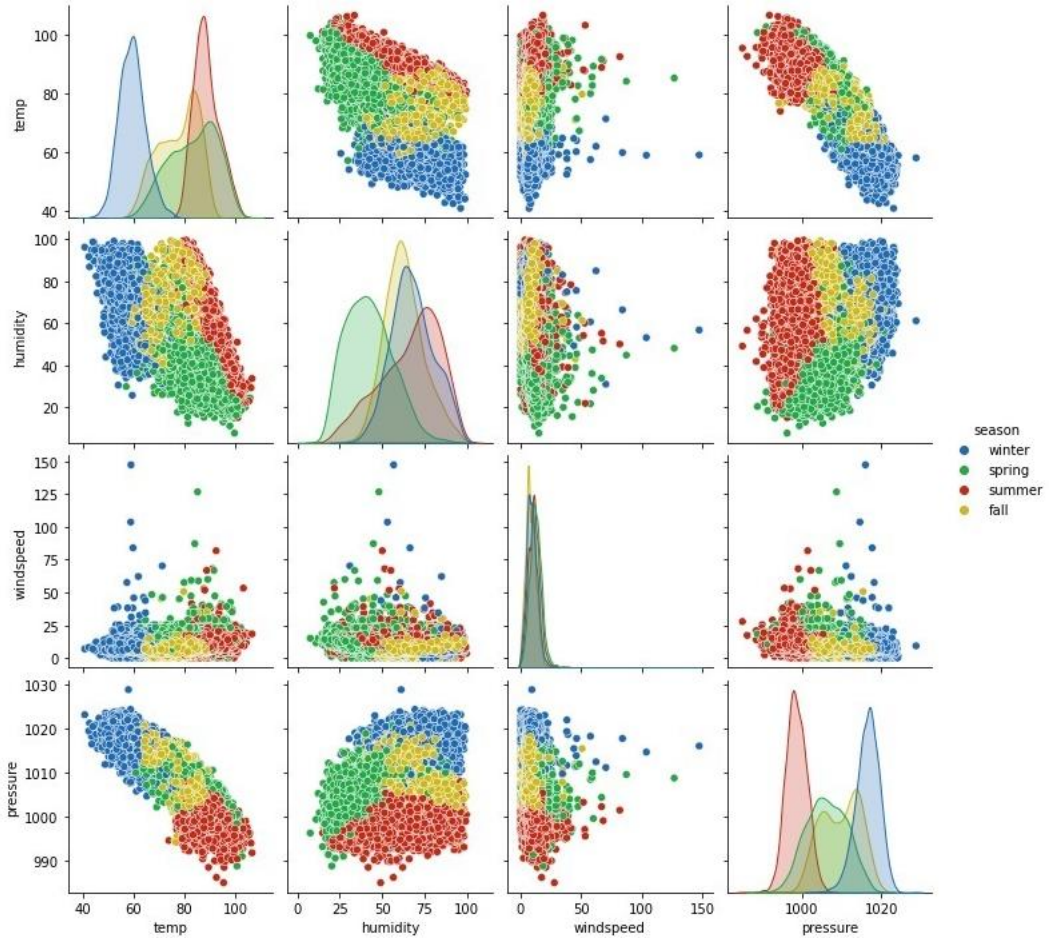


Figure 1.8 indicates the seasonality of the data. The values cluster together into four very distinct groups by season (fall, winter, spring, summer).

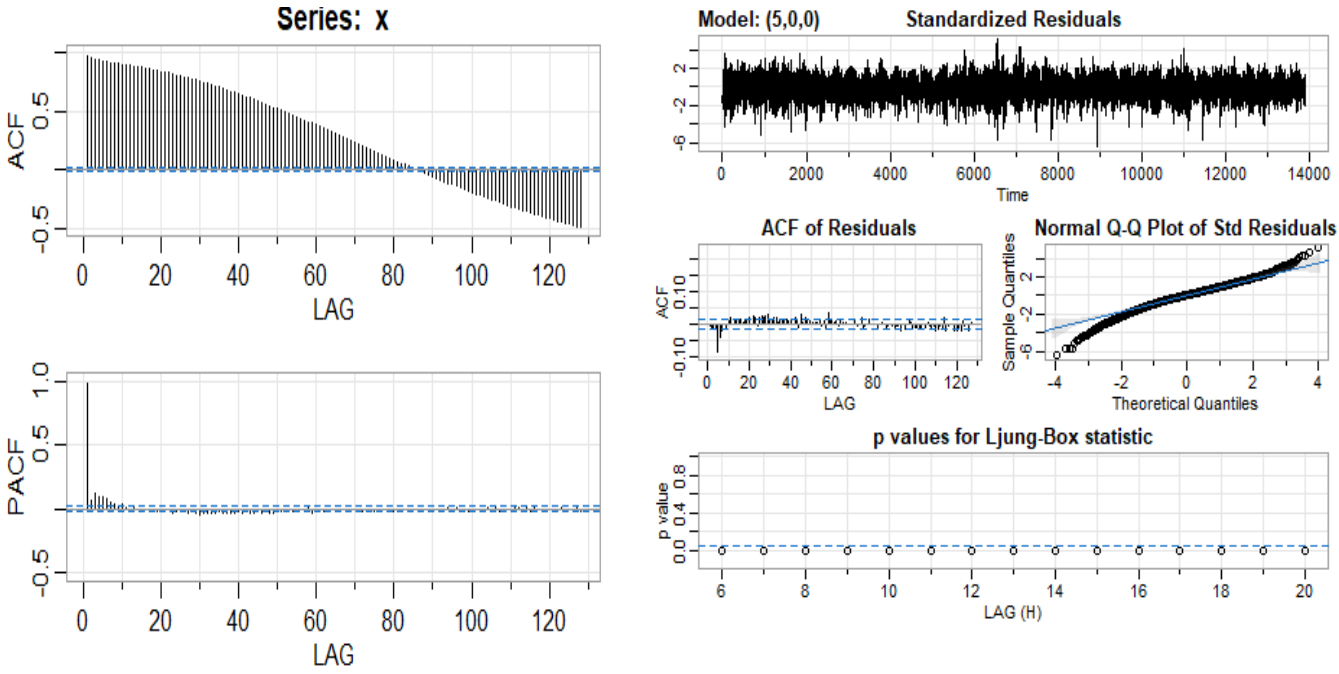
Figure 1.8 – Seasonality



Analysis

The dataset variables were analyzed individually to see what traits each possessed. The results could then be reviewed together to see if groups exhibited similar trends and conclusions made based on known interactions as they pertain to weather. For example, temperature trending upwards with pressure also trending upward and humidity trending downward would suggest an overall drying trend. Likewise, a pressure trending downward with a humidity trending upwards would indicate a trend towards wetter conditions. A change in these combined trends could indicate an overall trend towards a climate changing in a particular direction towards wetter or dryer. Figure 2.1 shows the time series analysis for temperature.

Figure 2.1 – Temperature Time Series



The data was not stationary so the distance parameter in the ARIMA model was set to 2 based on the auto ARIMA output. The ARIMA output ACF plot decays very slowly over time and the PACF plot cuts off after lag = 1 indicating SARIMA (p = 5 , d = 0 , q = 1 , P = 0 , D = 1 , Q = 0). Temperature looks to have an 85-day cycle which would approximately correspond the four seasons (fall, winter, spring, summer). The ACF of Residuals from the right plot does not show clear patterns in them and all the data points lie before the p-value of 0.05 which suggests we reject the null hypothesis. Other variables such as humidity and pressure had similar model output to temperature, which would make sense as there is some correlation between the variables. Windspeed showed a different output altogether and did not experience a sharp cutoff after lag 1. Windspeed displayed a sinusoidal seasonality pattern. Figures 2.2 through 2.4 show model output for humidity, windspeed, and pressure. The R code for all modeling can be found in appendix B.

Figure 2.2 – Humidity Time Series

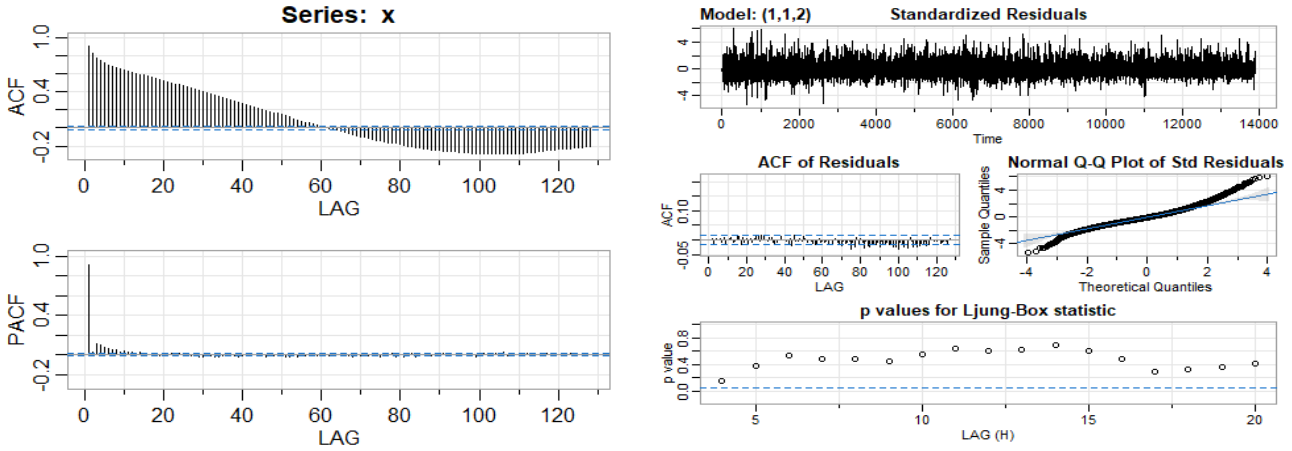


Figure 2.3 – Pressure Time Series

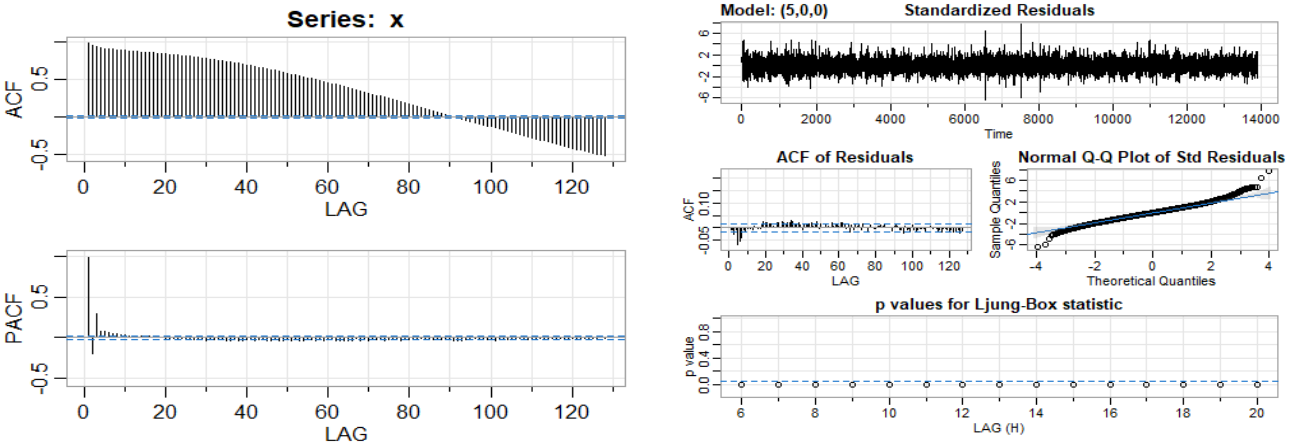


Figure 2.4 – Windspeed Time Series

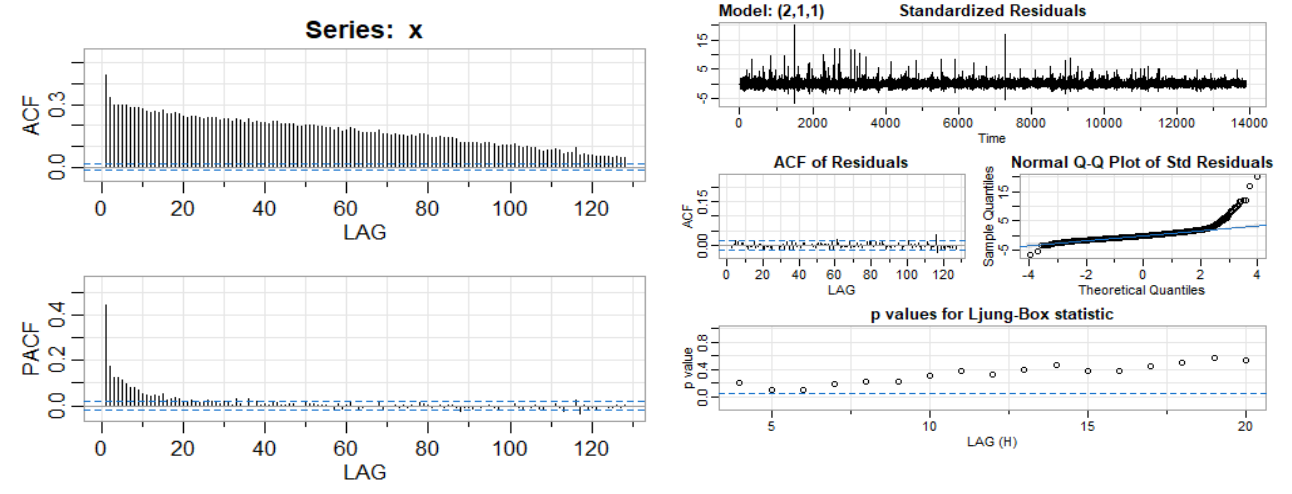
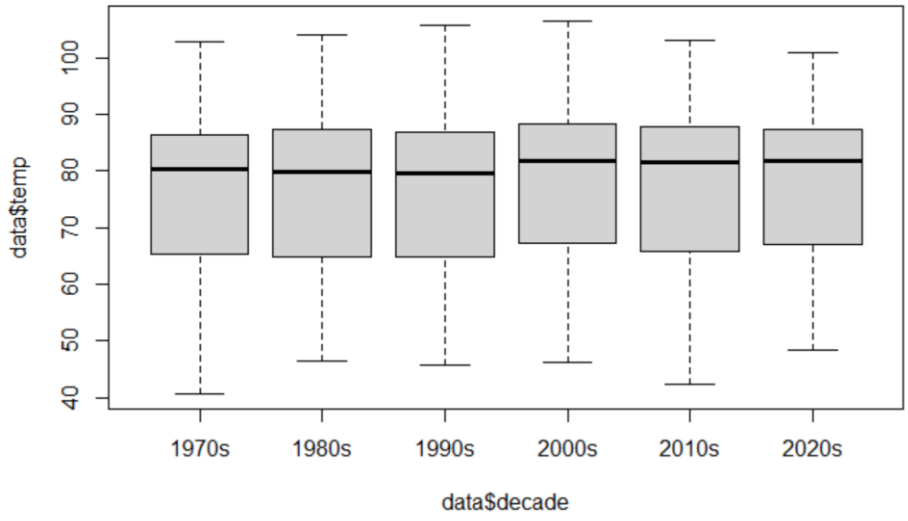


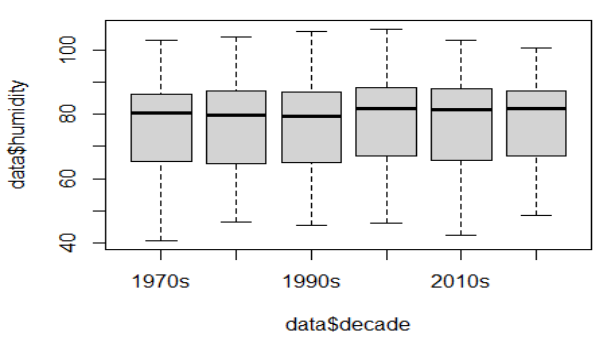
Figure 6.1 shows the boxplots for temperature by decade. The median temperature between decades is similar with an increase between 1990 and 2000. Overall, there is a slight temperature increase across decades. Even a slight upward trend in temperature could have a dramatic effect on climate.

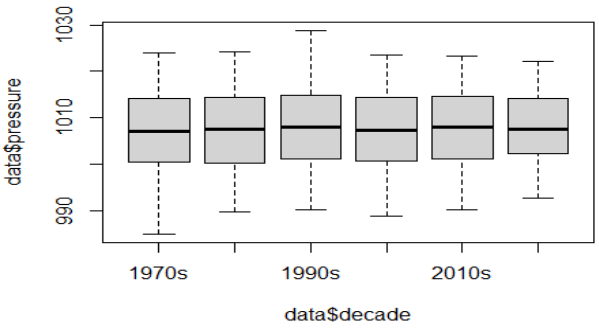
Figure 3.1 – Temperature by Decade



Similar to above, figure 3.2 shows how the additional variables humidity and pressure. Both had very similar median values and interval ranges, as well as overall trend.

Figure 3.2 – Humidity and Pressure by Decade





While only a slight increasing trend in temperature was noticed during the analysis, there was a definite increase in the high temperatures over the five decades included in the time series. Figure 4.1 indicates an overall increase in high temperatures which appear to level off after 2000. This same trend was also apparent in the average temperature, but not as extreme. There was also an increase in outliers which indicates more variability in the high temperature values. The leveling off after the turn of the century could be due to factors such as increased environmental regulation as global warming and climate change become more of a concern for governments around the world. The trend could also be cause by other factors such as a reduction in in economic or population growth.

Figure 4.1

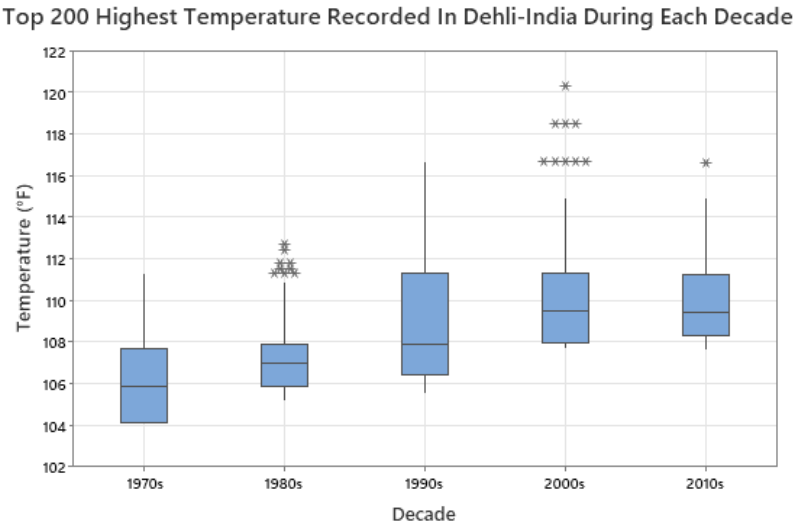
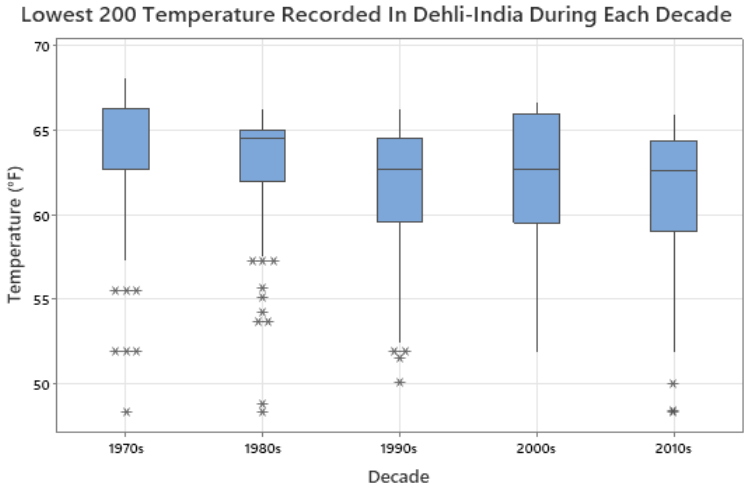


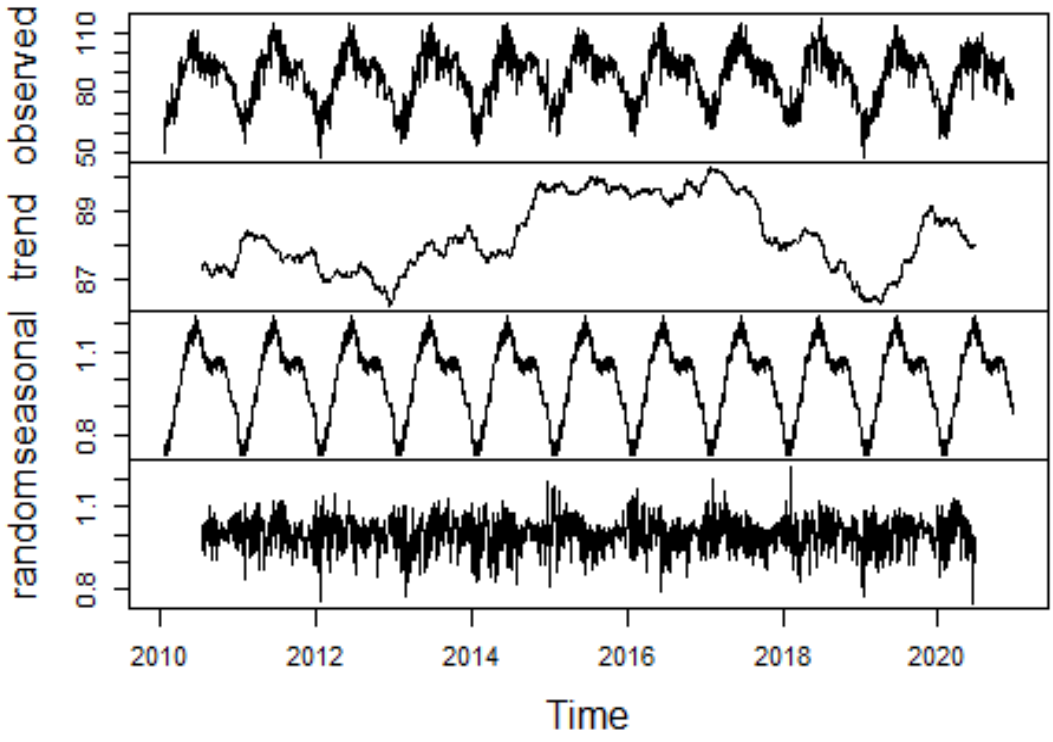
Figure 4.2



There was also a noticeable downward trend in low temperatures across the decades included in the time series. Similar to the high temperatures, there was an increase in outliers, indicating more variability in the data prior to 2000. This trend also levels off around the turn of the century consistent with the high temperatures.

Using the temperature variable, three different plots were created (figure 4.3).

Figure 4.3 – Decomposition of Multiplicative Time Series



The first shows the observed values over time. The data shows a definite seasonality and indicates a slight upward trend. The second plot shows the trend which can change from year to year. While the overall trend from 2010 to 2021 is upward the time segments in between may exhibit a downward trend. This information is useful to study how the data trends from year to year. The third plot shows seasonality and closely mimics the observed values. The last plot depicts the randomness of the data and appears to exhibit the same properties as white noise. By comparing the last plot to white noise, we can determine if the temperature plotted over time is simply randomly plotted data points.

Figure 5.1 shows the importance value for each variable used to train the random forest and gradient boost algorithms. The table indicates that pressure has the most impact when determining and predicting temperature. The “month_numeric” and “season_scaler” were also important variables based on the output.

Figure 5.1

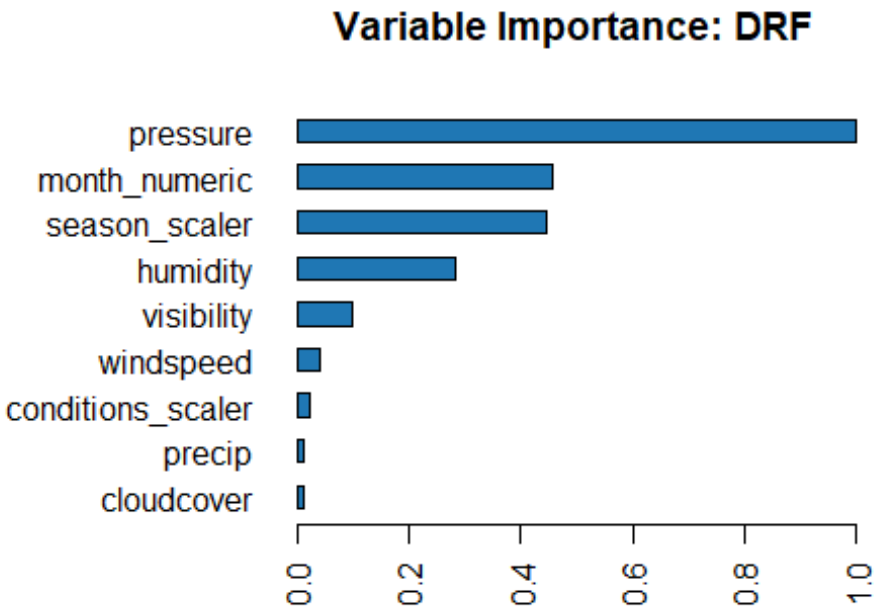


Figure 6.1 shows the two algorithms, random forest and gradient boost, that were created to test if an accurate forecast could be created. The first algorithm was meant to be robust, while the second algorithm was meant to expand on the first. The model results were promising and appear to capture the overall trend.

Figure 6.1

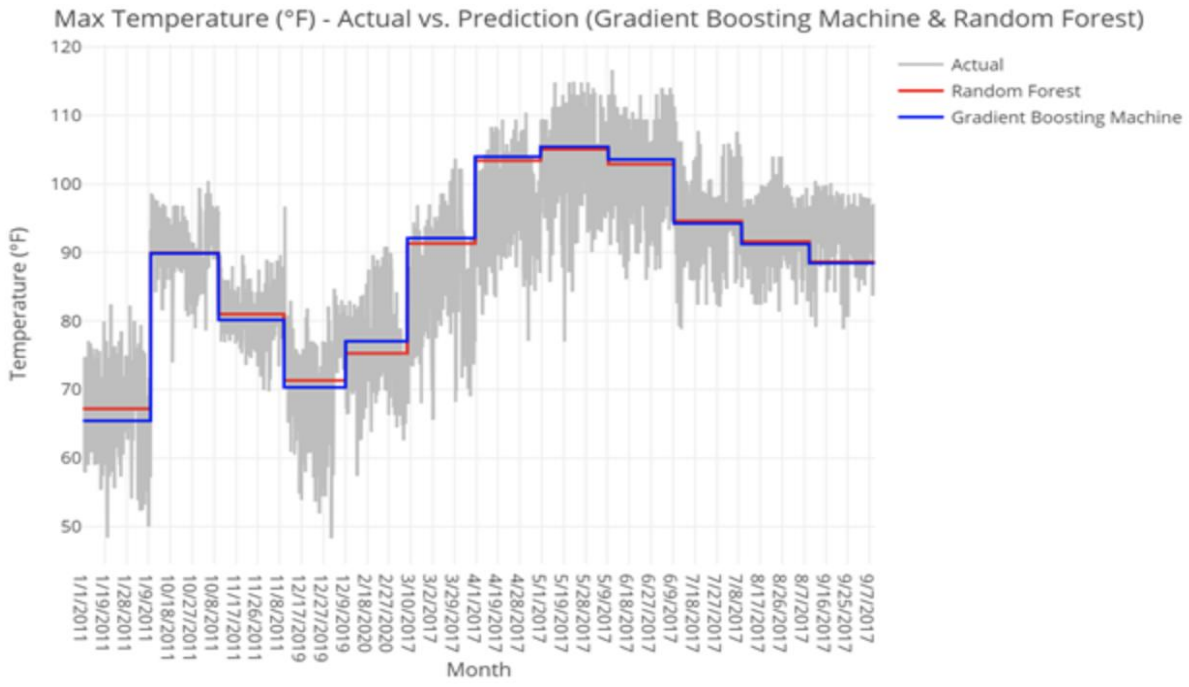
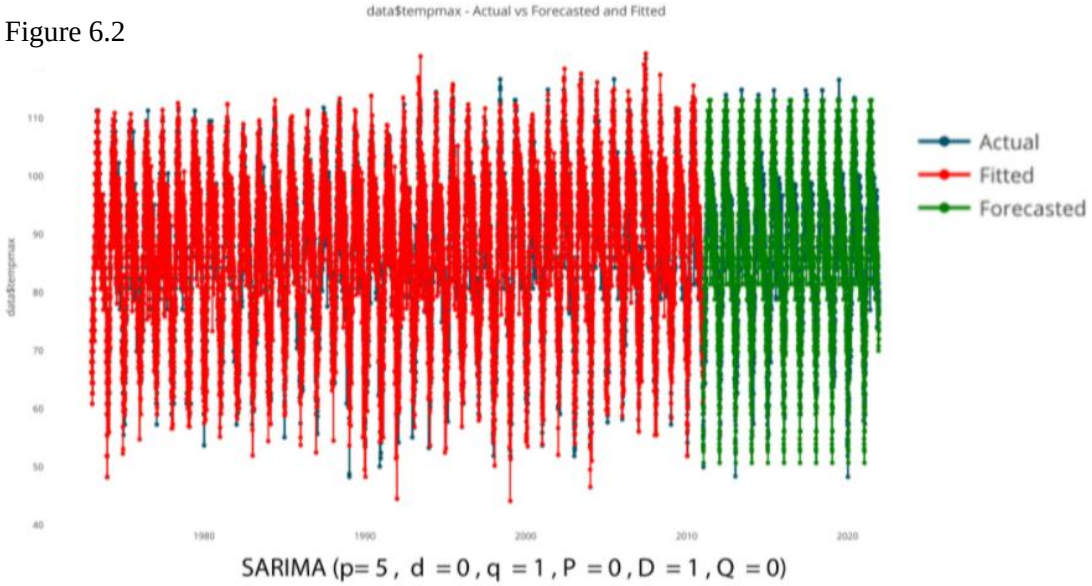


Figure 6.2 shows the trained and forecasted temperature over time. The overall temperature shows a slight upward trend in the actual data as well as the forecasted data.

Figure 6.2




```
Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE     MASE         ACF1
Training set 0.01179757 5.099908 3.767222 -0.1716369 4.444871 0.660315 0.0002324551
```

Conclusion

The data suggests that seasonal weather patterns in Delhi have been consistent over the past few decades, but overall temperatures have shown to be slightly increasing over the entire study period. There was a more noticeable jump in average temperatures between the 1990's and 2000's which leveled off potentially due to greater environmental awareness and regulation towards the end of the last century. While this small trend may seem insignificant, it could still have a huge impact on local wildlife. Research has shown that Indian bird species are in fact struggling. Unfortunately, the time series data analysis done for this project proved inconclusive and could not prove with any certainty that climate change may be a primary factor in declining bird populations. There were no changes in climate identified that were significant enough to show a direct relationship to declining bird populations. India is a rapidly developing country and due to this fact, there may be other factors which are having a greater impact on local bird species than climate change, such as changing land use, direct pollution, and diminishing resources. Hopefully, as interest grows around the problems facing Indian wildlife, more research will be devoted to understanding the primary challenges affecting local bird populations in order to put in place better mitigation strategies to lessen the impact of a growing population.

References

Parey, S. (2019, June 28). *Generating a Set of Temperature Time Series Representative of Recent Past and Near Future Climate*. Frontiers in Environmental Science. 7(1), 99. https://doi.org/10.3389/fenvs.2019.00099

Moller, A., Hochachka, W. (2019). *Long-term time series of ornithological data. Effects of Climate Change on Birds.* Oxford University Press. 37-43. DOI: 10.1093/OSO/9780198824268.003.0004

Atanu, R., Susmita, D., Kakoli, B., Abhijit M. (2012, February 28). *Climate change impacts on Indian Sunderbans: a time series analysis (1924–2008).* Biodiversity and Conservation. 21, 1289–1307. https://doi.org/10.1007/s10531-012-0260-z

Faghmous, J., Kumar, V. (2014). *A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science*. Big data, 2(3), 155–163. https://doi.org/10.1089/big.2014.0026

Ghosh, P. (2018, July 4). *Climate change, disturbance to habitat are affecting the breeding patterns of Indian birds.* Scroll.in – Wildlife Conservation. https://scroll.in/article/884997/climate-change-disturbance-to-habitat-are-affecting-the-breeding-patterns-of-indian-birds

The Climate Reality Project. (2016, April 11). *What's the big deal if the planet is a few degrees warmer?* The Climate Reality Project. https://www.climaterealityproject.org/blog/what-big-deal-planet-few-degrees-warmer

Visual Crossing Weather. (2021). https://weather.visualcrossing.com [unpublished raw data].

Appendix A

JSON code used for data extraction from https://weather.visualcrossing.com

```
from datetime import date
from pandas.tseries import offsets
import datetime, pandas as pd, json, requests, ast

def get_weather_info(y_from,y_to):
    for y in range(y_from,y_to+1):
        d_from = pd.to_datetime(f'{y}-1-1').strftime("%Y-%m-%d")
        d_to = (pd.to_datetime(f'{y+1}-1-1') +
               datetime.timedelta(days = -1)).strftime("%Y-%m-%d")
        api_key = 'XXXXXXXXXXXXXXXXXXX'
        URL_full = f"https://weather.visualcrossing.com/VisualCrossingWebServices/"\
        "rest/services/timeline/delhi%2C%20india/{d_from}/{d_to}?unitGroup=us&key={api_key}"

        try:
            response = requests.get(URL_full)
            JSONCode = json.loads(response.text)
            with open(f'{y}.json', 'w', encoding='utf-8') as f:
                json.dump(JSONCode, f, ensure_ascii=False, indent=4)
        pd.DataFrame()
        except:pass

def combine_json_to_df(y_from,y_to):
    df = pd.DataFrame()
    for y in range(y_from,y_to+1):
        print(y)
        with open(f'{y}.json') as json_file:
            JSONCode = json.load(json_file)

        WeatherInfo = str(JSONCode.get('days'))
        BreakDown = ast.literal_eval(WeatherInfo)

        for row in range(len(BreakDown)):
            last_row = len(df) + 1
            df.loc[last_row,'datetime'] = BreakDown[row].get('datetime')
            df.loc[last_row,'tempmax'] = BreakDown[row].get('tempmax')
            df.loc[last_row,'tempmin'] = BreakDown[row].get('tempmin')
            df.loc[last_row,'temp'] = BreakDown[row].get('temp')
            df.loc[last_row,'humidity'] = BreakDown[row].get('humidity')
            df.loc[last_row,'precip'] = BreakDown[row].get('precip')
            df.loc[last_row,'windspeed'] = BreakDown[row].get('windspeed')
            df.loc[last_row,'pressure'] = BreakDown[row].get('pressure')
            df.loc[last_row,'visibility'] = BreakDown[row].get('visibility')
            df.loc[last_row,'conditions'] = BreakDown[row].get('conditions')
            df.loc[last_row,'cloudcover'] = BreakDown[row].get('cloudcover')

    df.to_csv('all.csv')


y_from, y_to = 1973, 2021

get_weather_info(y_from,y_to)
combine_json_to_df(y_from,y_to)
```

Appendix B
ADS 506 Final Project Code

```
library(astsa)
library(dplyr)
library(lubridate)
library(zoo)
library(forecast)
library(ggplot2)
library(reshape)
library(TSstudio)
library(h2o)
library(plotly)
```

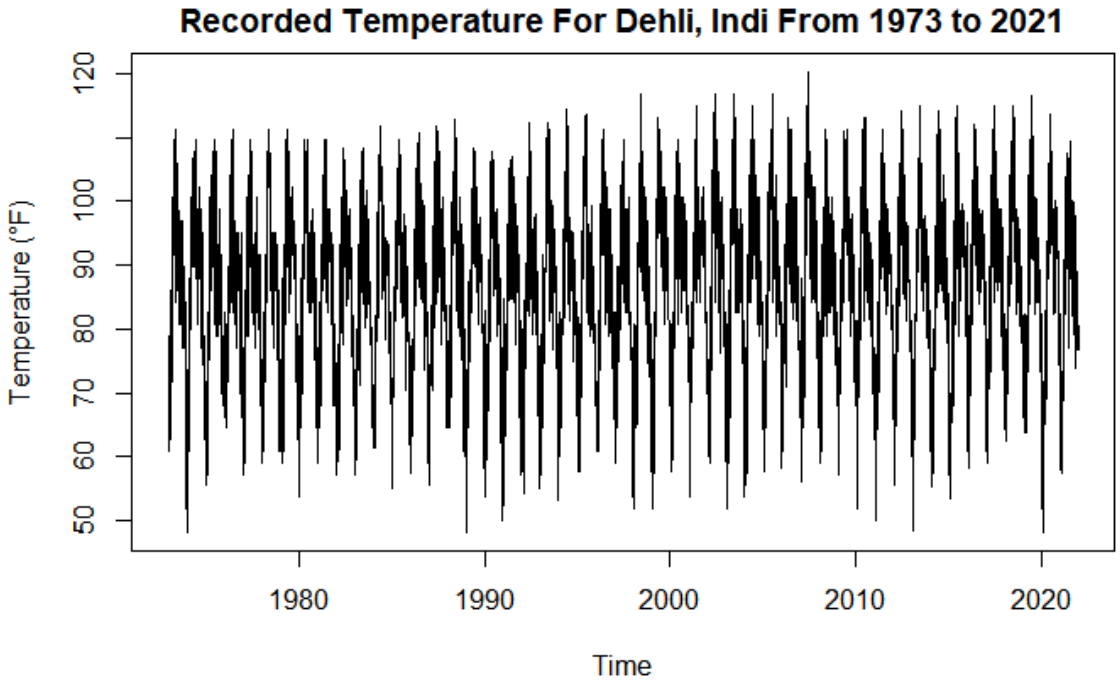
```
# Import Delhi_Weather_Data.csv
data = read.csv("F:/School/ADS/506/Final Project/dehli_weather_info.csv")
```

```
# Checking the percentage of NA values
length(which(is.na(data$tempmin))) / dim(data)[1]
length(which(is.na(data$tempmax))) / dim(data)[1]
length(which(is.na(data$temp))) / dim(data)[1]
length(which(is.na(data$humidity))) / dim(data)[1]
length(which(is.na(data$pressure))) / dim(data)[1]
length(which(is.na(data$windspeed))) / dim(data)[1]
```

```
[1] 0.0003918057
[1] 0.0003918057
[1] 0.0003918057
[1] 0.0003918057
[1] 0.0007276391
[1] 0.0003918057
```

```
# Replace NA's
# The missing values will be replaced with the average during a certain week. For
example, if a "temp" value is missing,
# the data from 3 days prior and 3 days after will be added together and then
divided by 6 to find the average "temp".

NA_replace <- function(df) {
  for(j in 1:ncol(df)){
    for(i in 1:nrow(df)){
      if(is.na(df[i,j]) == TRUE && i > 3){
        avg <- sum(df[(i-3):(i+3),j], na.rm = TRUE) / 6
        df[i,j] <- avg
      }
    }
  }
  return(df)
}
```

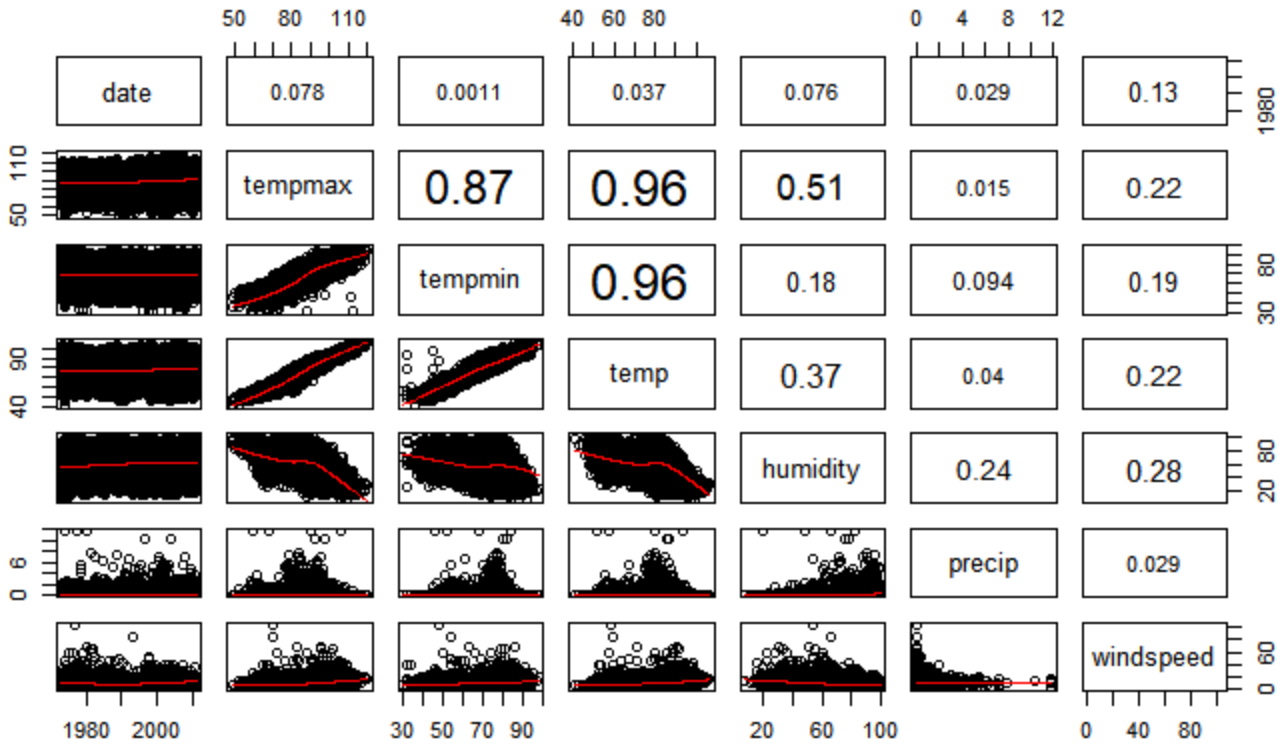
```
data = NA_replace(data)

colnames(data)[2] = 'date'

data$date = strptime(data$date, "%m/%d/%Y")
data$date = format(data$date, "%Y-%m-%d")
data$date = as.Date(data$date)


data$tempmax = ts(data = data$tempmax,
                  frequency = 365,
                  start = c(1973, yday(head(data$date,1))))

plot(data$tempmax,
     main = 'Recorded Temperature For Dehli, Indi From 1973 to 2021',
     ylab = 'Temperature (°F)',
     xlab = 'Time')
```




```
# for use later in ML model
train = subset(data, data$year <= '2010')
test = subset(data, data$year > '2010' & data$year <= '2021' )
```

```
# Checking the percentage of NA values
length(which(is.na(data$tempmin))) / dim(data)[1]
length(which(is.na(data$tempmax))) / dim(data)[1]
length(which(is.na(data$temp))) / dim(data)[1]
length(which(is.na(data$humidity))) / dim(data)[1]
length(which(is.na(data$pressure))) / dim(data)[1]
length(which(is.na(data$windspeed))) / dim(data)[1]
```

```r
# **Correlation between each feature**


  panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
    usr <- par("usr")
    on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    Cor <- abs(cor(x, y))
    txt <- paste0(prefix, format(c(Cor, 0.123456789), digits = digits)[1])
    if(missing(cex.cor)) {
        cex.cor <- 0.4 / strwidth(txt)
    }
    text(0.5, 0.5, txt,
         cex = 1 + cex.cor * Cor) # This will show correlation value and change the
size based on the value
}

# Plot correlation
pairs(train[, c(2:8)],
      upper.panel = panel.cor,
      lower.panel = panel.smooth)
```

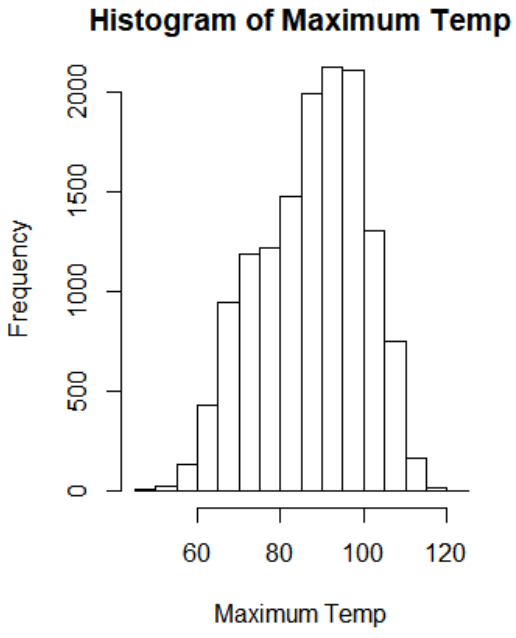


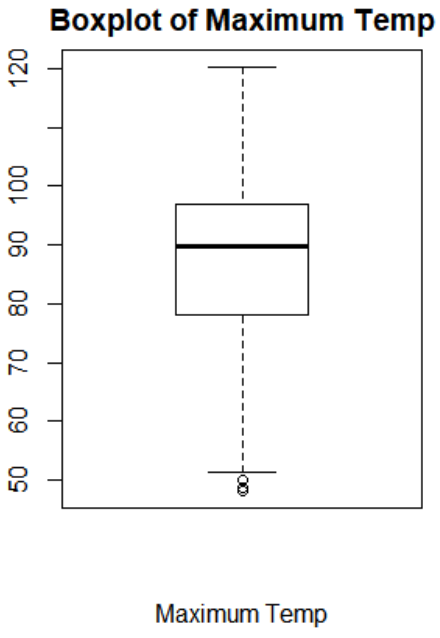
There are strong positive correlation between temp columns (min, max, mean) and pressure. The next big correlation is between humidity and temp. Precipitation and windspeed has slightly higher correlation than neutral.

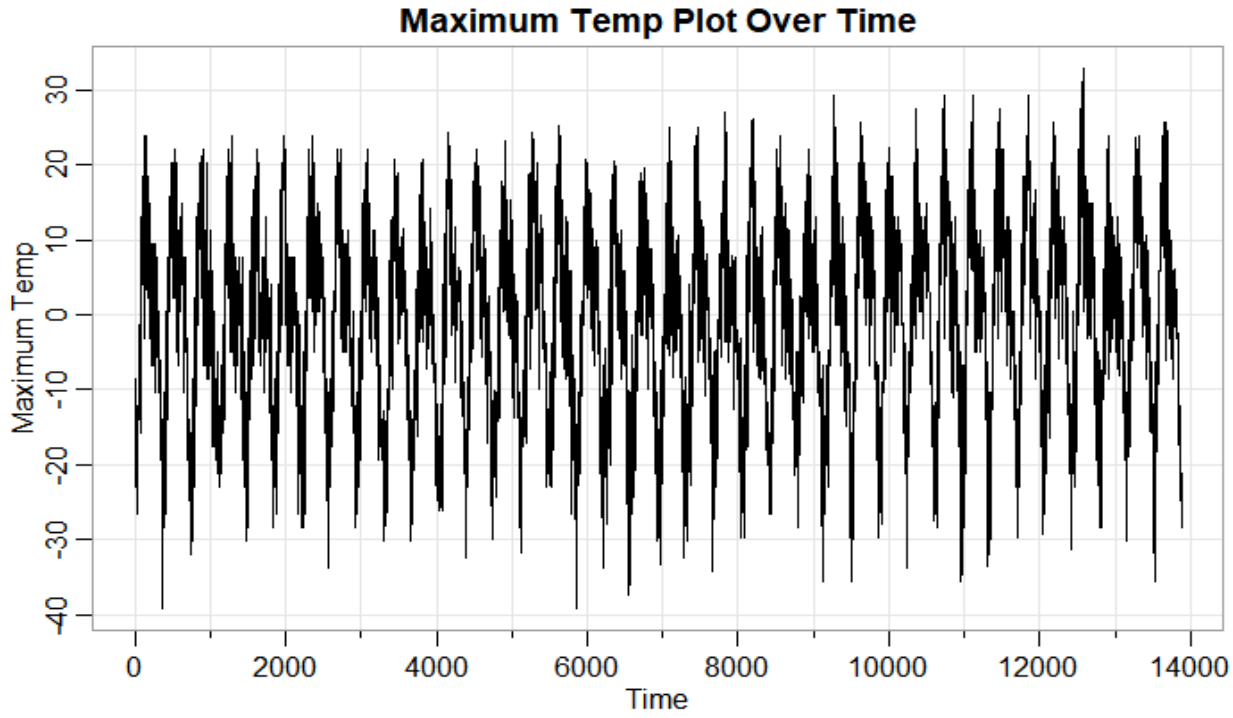
```r
visualize_feature = function(x, name) {

  # Show Histogram of feature before subtracting mean value

  par(mfrow=c(1,2))
  hist(x, main=paste('Histogram of', name), xlab=name)
  boxplot(x, main=paste('Boxplot of', name), xlab=name)

  x = x - mean(x)
  summary(x)

  par(mfrow=c(1,1))
  tsplot(x, main=paste(name, 'Plot Over Time'), ylab=name)

  acf2(x)

  best_param = auto.arima(x)
  print(best_param)

  x
}
```

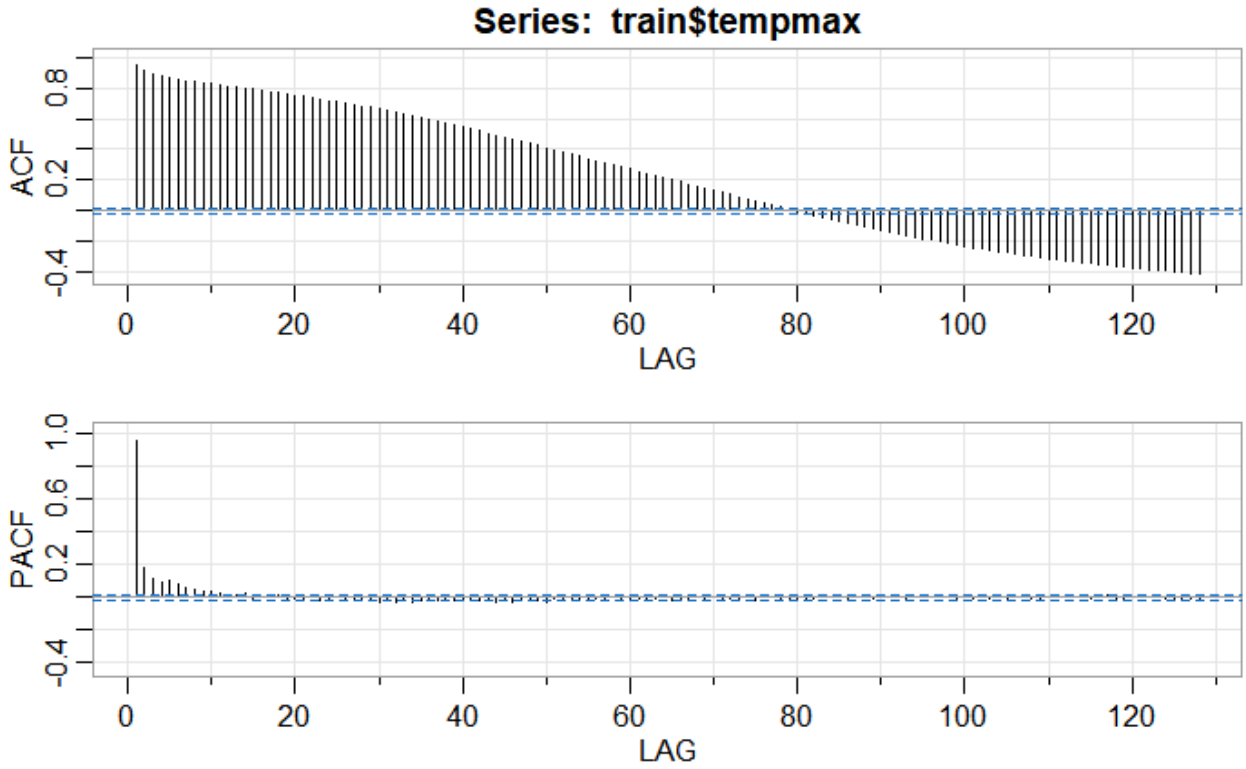
```r
# Maximum temperature

tempmax = visualize_feature(train$tempmax, 'Maximum Temp')
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-39.117	-9.217	2.283	0.000	9.483	32.883



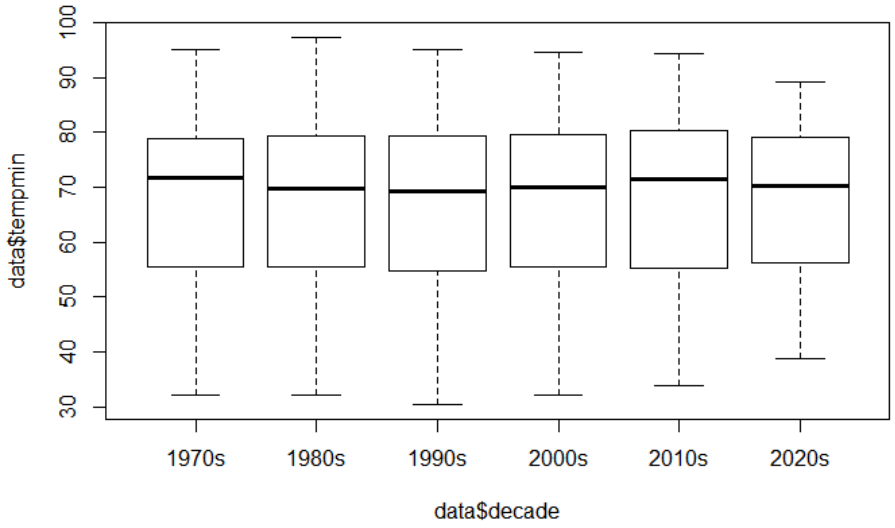


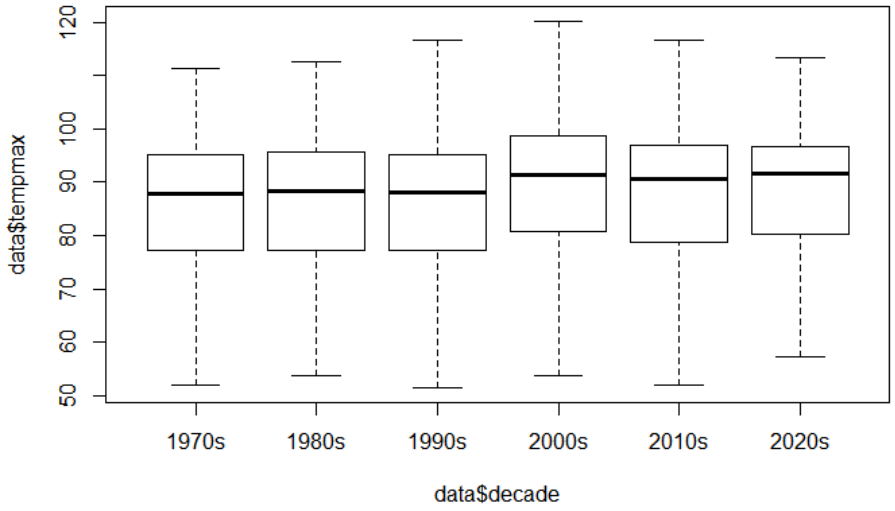


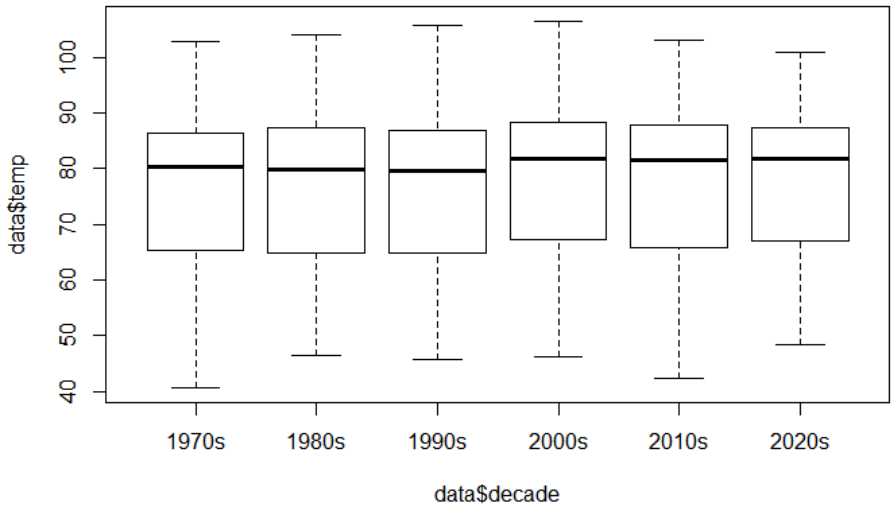



```
Series: y_train
ARIMA(5,0,1)(0,1,0)[365]

Coefficients:
         ar1      ar2      ar3      ar4     ar5      ma1
      1.3094  -0.3890  -0.0142  -0.0156  0.0307  -0.6460
s.e.  0.0984   0.0668   0.0151   0.0143  0.0112   0.0983

sigma^2 estimated as 26.72:  log likelihood=-41373.25
AIC=82760.49   AICc=82760.5   BIC=82813.08

Training set error measures:
                     ME     RMSE      MAE        MPE     MAPE     MASE          ACF1
Training set 0.01179757 5.099908 3.767222 -0.1716369 4.444871 0.660315 0.0002324551
```

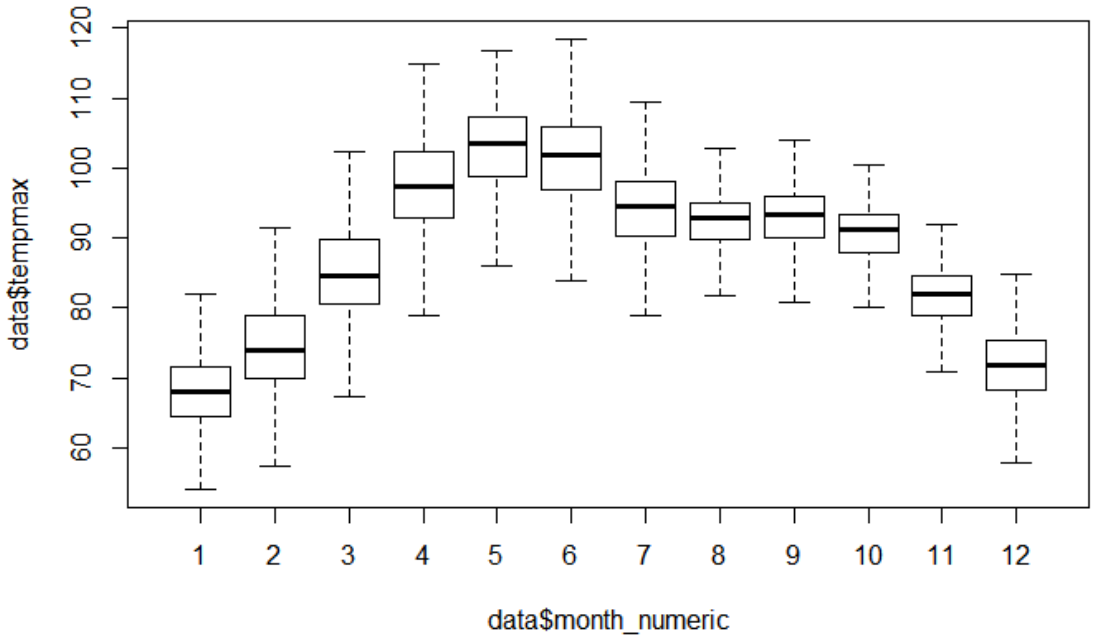
```r
boxplot(data$tempmin ~ data$decade, outline = FALSE)

boxplot(data$tempmax ~ data$decade, outline = FALSE)

boxplot(data$temp ~ data$decade, outline = FALSE)
```

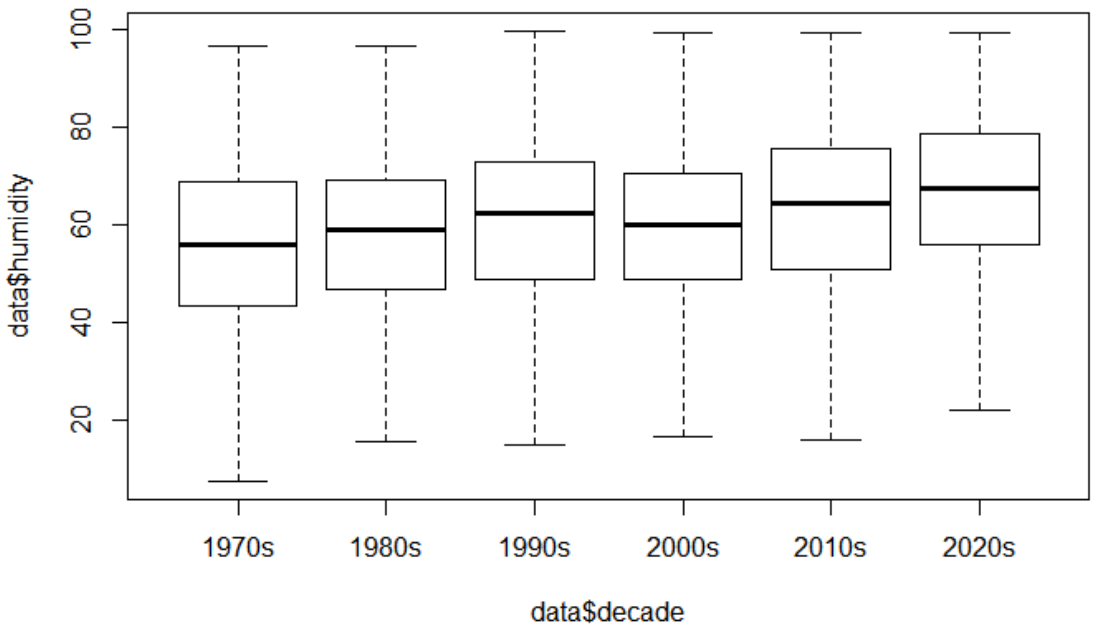


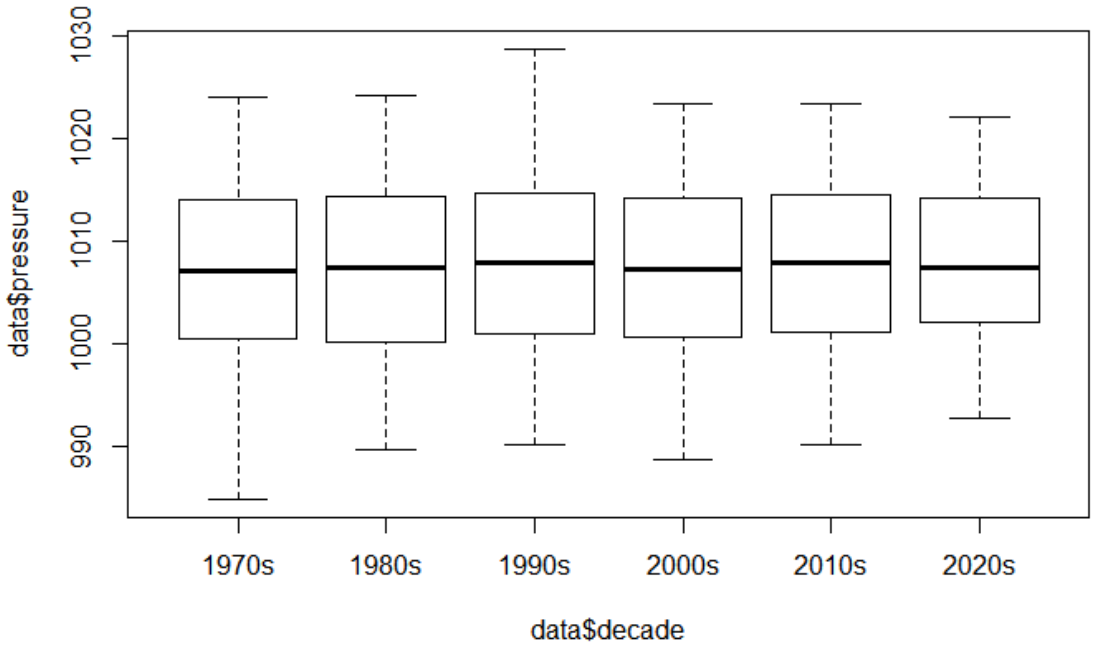


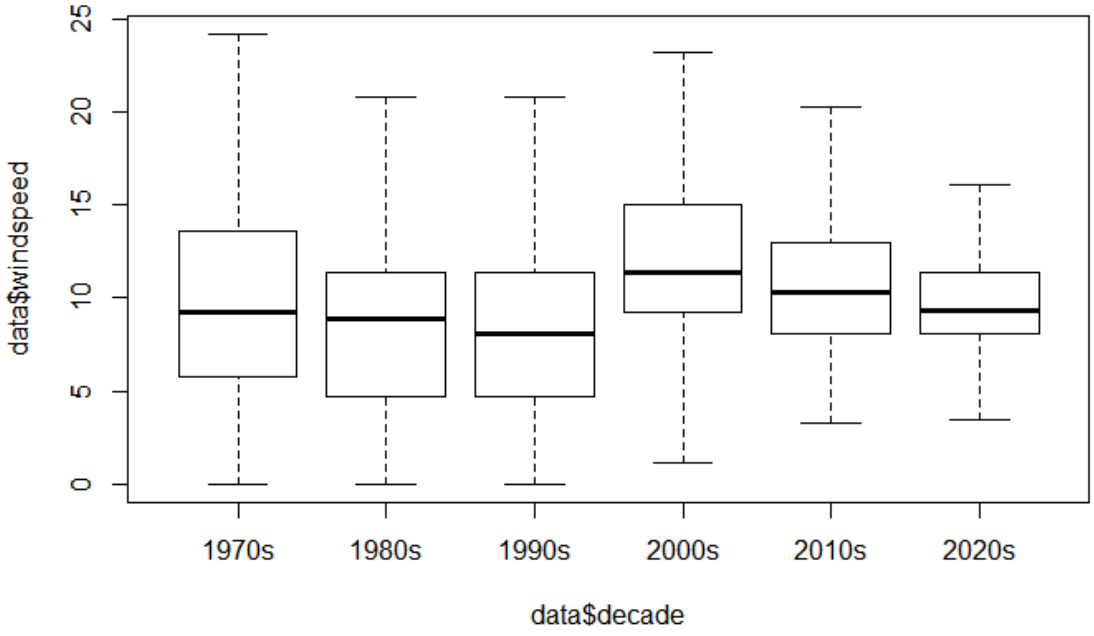


```r
boxplot(data$tempmax ~ data$month_numeric, outline = FALSE)

boxplot(data$humidity ~ data$decade, outline = FALSE)

boxplot(data$pressure ~ data$decade, outline = FALSE)

boxplot(data$windspeed ~ data$decade, outline = FALSE)
```

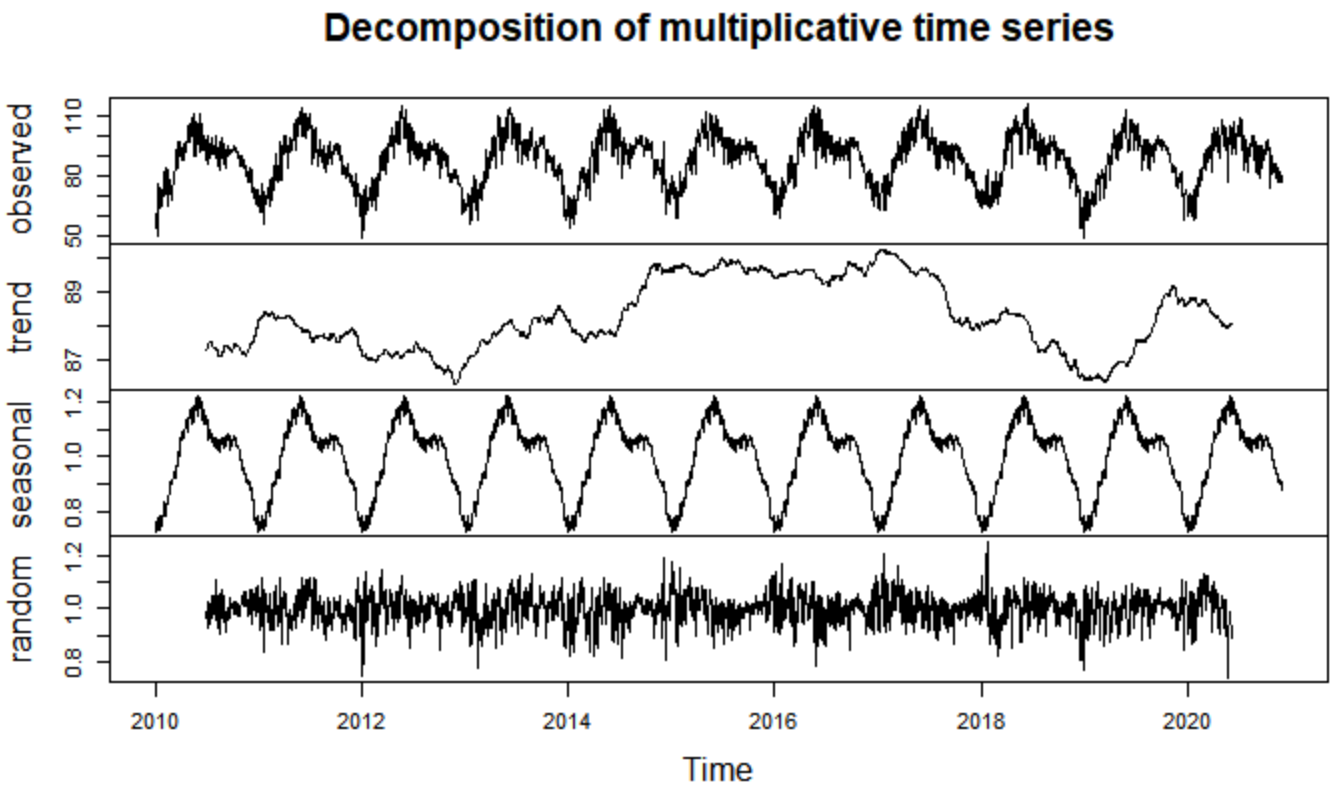





data$pressure
990
1000
1010
1020
1030
1970s
1980s
1990s
2000s
2010s
2020s
data$decade


data$windspeed
0
5
10
15
20
25
1970s
1980s
1990s
2000s
2010s
2020s
data$decade

```r
tempmax_ts = ts(data = test$tempmax,
                frequency = 365,
                start = c(2010, yday(head(test$date,1))))

# Show Classical Seasonal Decomposition by Moving Averages
decomposition = decompose(tempmax_ts, type = 'multiplicative')

plot(decomposition)
```

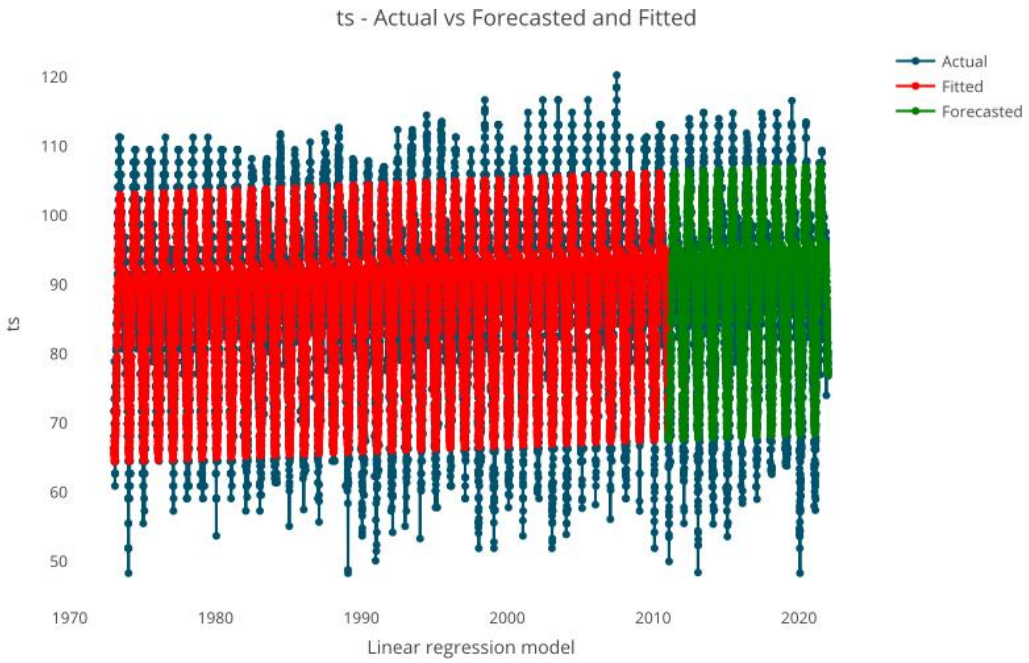


```r
# Moldeing Setup
h = nrow(test)

ts = na.locf(data$tempmax)
ts_par = ts_split(ts, sample.out = h)
y_train = ts_par$train
y_test = ts_par$test

# Training Model
md_tslm = tslm(y_train ~ season + trend)
hw_model = HoltWinters(y_train, seasonal = 'multiplicative')
arima_model = auto.arima(y_train)
```

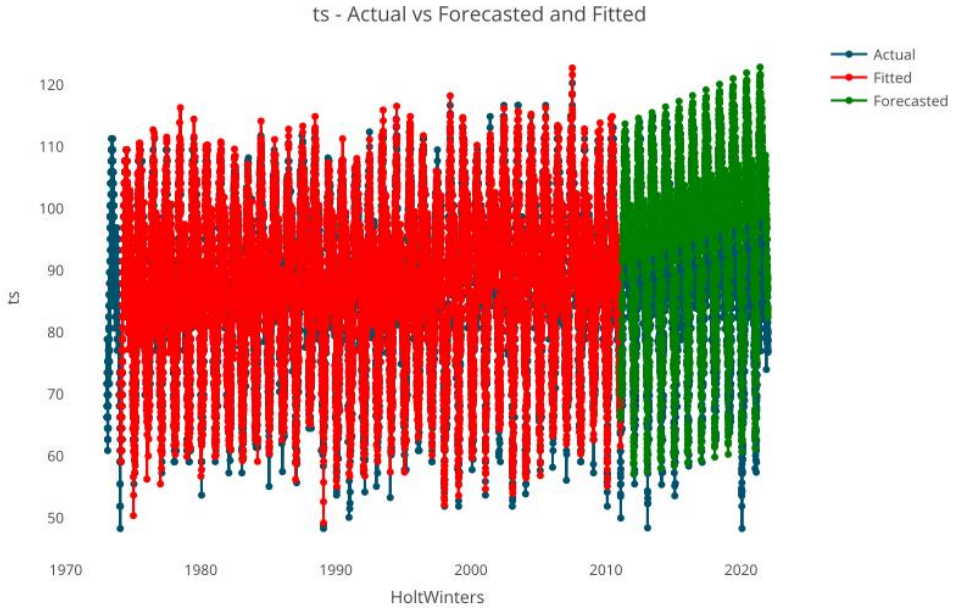
```r
# Forcast Portion
fc_tslm = forecast(md_tslm, h = h)
fc_hw = forecast(hw_model, h = h)
fc_arima = forecast(arima_model, h = h)
```

```r
# Plot
test_forecast(actual = ts,forecast.obj = fc_tslm, test = y_test)
test_forecast(actual = ts,forecast.obj = fc_hw,   test = y_test)
test_forecast(actual = ts,forecast.obj = fc_arima,test = y_test)
```

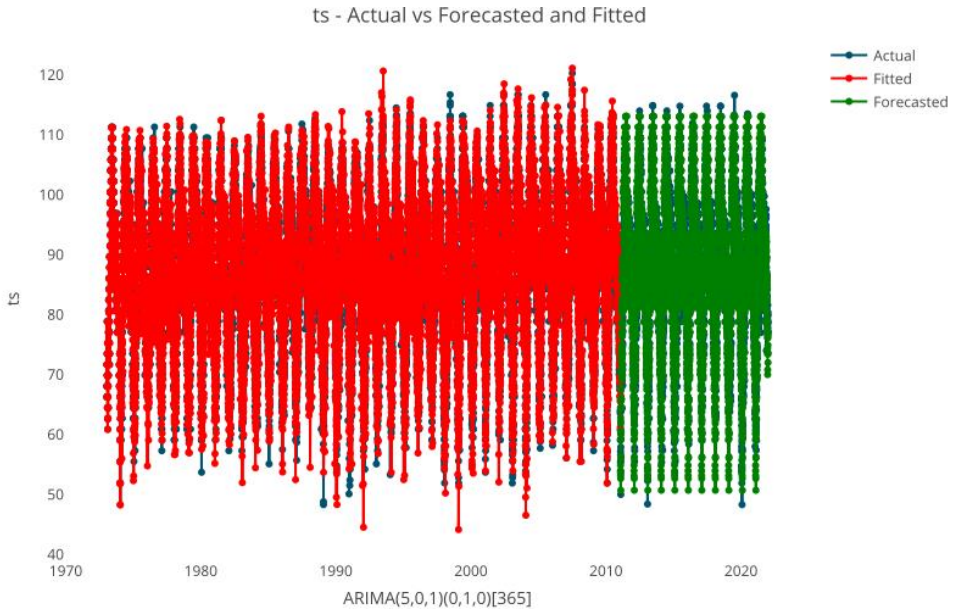
Plot 1
Linear Model



Plot 2
Holt Winters



Plot 3
SARIMA (5 , 0 , 1 , 0 , 1 , 0)



```r
# Checking Performance Of Each Model
accuracy(fc_tslm, y_test)
accuracy(fc_hw, y_test)
accuracy(fc_arima, y_test)
```

```
                      ME     RMSE      MAE        MPE     MAPE      MASE      ACF1 Theil's U
Training set -1.001005e-16 5.244222 4.051048 -0.3883931 4.777445 0.7100637 0.7199728        NA
Test set     -1.352905e+00 5.420191 4.124090 -2.0379698 4.967317 0.7228663 0.7220514  1.468115
```

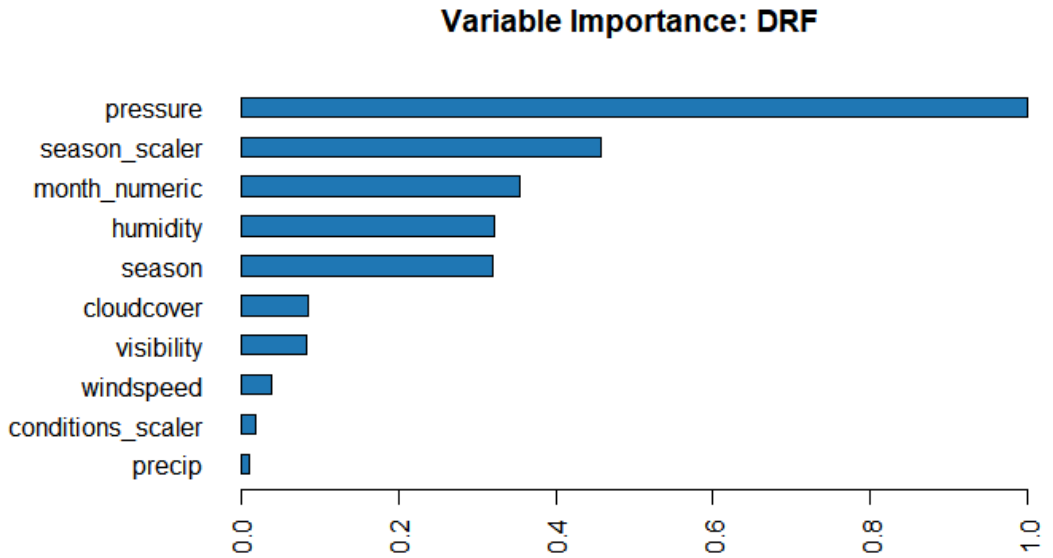
```
                    ME      RMSE      MAE        MPE     MAPE      MASE      ACF1 Theil's U
Training set -0.0276418  4.350542 3.245102 -0.1937894 3.849248 0.5687983 0.1625329        NA
Test set     -7.1987998 10.282641 8.248469 -8.5142413 9.768979 1.4457835 0.8363367  2.684651
```

```
                     ME      RMSE      MAE        MPE     MAPE     MASE          ACF1 Theil's U
Training set  0.01179757  5.099908 3.767222 -0.1716369 4.444871 0.660315 0.0002324551        NA
Test set     -0.34436945 10.171569 8.291481 -0.9856161 9.875336 1.453323 0.8580748692  2.710774
```

```
h2o.init(max_mem_size = "16G")


train_h = as.h2o(train)
test_h <- as.h2o(test)


x <-
c('humidity','precip','windspeed','pressure','visibility','cloudcover','conditions_s
caler','season','season_scaler', 'month_numeric')
y <- "tempmax"

rf_md <- h2o.randomForest(training_frame = train_h,
                          nfolds = 5,
                          x = x,
                          y = y,
                          ntrees = 100,
                          stopping_rounds = 10,
                          stopping_metric = "RMSE",
                          score_each_iteration = TRUE,
                          stopping_tolerance = 0.0001,
                          seed = 1234)

h2o.varimp_plot(rf_md)
tree_score <- rf_md@model$scoring_history$training_rmse
plot_ly(x = seq_along(tree_score), y = tree_score,
        type = "scatter", mode = "line") %>%
  layout(title = "The Trained Model Score History",
         yaxis = list(title = "RMSE"),
         xaxis = list(title = "Num. of Trees"))

x = c('month_numeric','year','season_scaler')
y = "tempmax"
```

```
rf_md = h2o.randomForest(training_frame = train_h,
                         nfolds = 5,
                         x = x,
                         y = y,
                         ntrees = 100,
                         stopping_rounds = 10,
                         stopping_metric = "RMSE",
                         score_each_iteration = TRUE,
                         stopping_tolerance = 0.0001,
                         seed = 1234)

test_h$pred_rf = h2o.predict(rf_md, test_h)

test_1 <- as.data.frame(test_h)
mape_rf <- mean(abs(test_1$tempmax - test_1$pred_rf) / test_1$tempmax)
mape_rf
```

```r
gbm_md <- h2o.gbm(
  training_frame = train_h,
  nfolds = 5,
  x = x,
  y = y,
  max_depth = 20,
  distribution = "gaussian",
  ntrees = 500,
  learn_rate = 0.1,
  score_each_iteration = TRUE
)

test_h$pred_gbm <- h2o.predict(gbm_md, test_h)
test_1 <- as.data.frame(test_h)
mape_gbm <- mean(abs(test_1$tempmax - test_1$pred_gbm) / test_1$tempmax)
mape_gbm

mape_hw <- mean(abs(test_1$tempmax - fc_hw$mean) / test_1$tempmax)
mape_hw
```



```
plot_ly(data = test_1) %>%
  add_lines(x = ~ test_1$date, y = ~ test_1$tempmax,
  name = "Actual",color = I("gray")) %>%
  add_lines(x = ~ test_1$date, y = ~ test_1$pred_rf,
  name = "Random Forest", color = I("red")) %>%
  add_lines(x = ~ test_1$date, y = ~ test_1$pred_gbm,
  name = "Gradient Boosting Machine",color = I("blue")) %>%
  layout(title = "Max Temperature (°F) - Actual vs. Prediction (Gradient Boosting
  Machine & Random Forest)", yaxis = list(title = "Temperature (°F)"), xaxis =
  list(title = "Month"))
```