**CAPSTONE**

**PROJECT REPORT**



TOPIC: Weather Forecasting WW2

By: Neha Bhandari

Master’s in Information Technology and Analytics Rutgers University, Newark, NJ, USA neha.bhandari@rutgers.edu

RUID:183008806

**TABLE OF CONTENTS**

**TOPIC PAGE NO.**

I. Introduction …………………………………. 3

II. Overview of Analysis ………………………. 5

III. Dataset Description ………………………… 5

IV. Setting up environment and Data Import … 7

V. Data Cleaning ………………………………. 9

VI. Data Analysis and Visualization …………… 14

VII. Time Series Analysis & Modelling………….. 27

VIII. Results and Discussions ………………… 49

IX. Conclusion …………………………………… 50

**1.Introduction**

War does not affect weather. Weather does affect war. To know in advance the climate of a war zone; to have as accurate knowledge as possible of the probability of occurrence of severe cold; of sudden thaws; of heavy rains; of great heat; of high winds--all this is a very essential element in planning a campaign or in organizing a single engagement.

Weather has long played a vital role in human history. Kublai Khan’s attempted conquest of Japan was foiled when his invasion fleet was destroyed by a typhoon. Napoleon’s Grand Armee perished during his ill-fated Russian campaign, laid low by the sweltering heat of summer and the frigid cold of winter.

However, the weather became even more important during the 20th century thanks to the invention of the airplane, tank, and modern ship.

Bombers and other aircraft can be grounded by bad weather or their targets obscured by fog or clouds. Land offensives depend on accurate predictions of the weather, and at sea, convoys bearing vital supplies also need reliable forecasts to deliver their cargoes.

Seasons can affect military operations. Battles are sometimes planned according to the seasons, so they know when it is the best time to attack. Day is only decided when the weather was predicted calm enough for amphibious landings.





**World War 2**

World War II was the biggest and deadliest war in history, involving more than 30 countries

In the WORLD WAR 2, weather played an important role for the countries to decide when to

attack.

They had Weather Station locations near the target cities which usually was a cluster of

huts that was lavish by Arctic standards. The people were a mixture of technicians and

soldiers—with the facilities of radio transmission.

They used to broadcast weather data but in plain transmissions without code. And that gave

the countries information to decide their plan of attack.

**2.Overview of analysis**

The goal of this analysis serves two purposes:

* The first one is to understand the World war 2 scenario- the countries who attacked, the countries which got attacked, the weather station locations, what was the weather during that time and then establish a relationship between these variables.
* The second is to use forecast techniques to predict the future weather conditions.

To start the analysis, we will first clean the data. For cleaning the data, we have to take care of the missing values, outliers, unused features.

Cleaning the data helps us to better understand and visualize the data.

Once the cleaning is done, we visualize the data to understand the data more.

We visualize the top attacker countries, top target countries, the paths of attack, what are the weather locations etc.

The countries who attacked had weather stations located in different places. A weather station is a facility, either on land or sea, with instruments and equipment for measuring atmospheric conditions to provide information for weather forecasts and to study the weather and climate.

This information is used by them to plan their attacks.

The weather information provided by these stations is forecasted to predict the future weather conditions.

**3.Dataset Description**

The dataset contains information on Bombing Operations about the attacker and target countries, their mission dates, the aircrafts used etc., the weather stations locations and the weather conditions recorded on each day at various weather stations around the world.

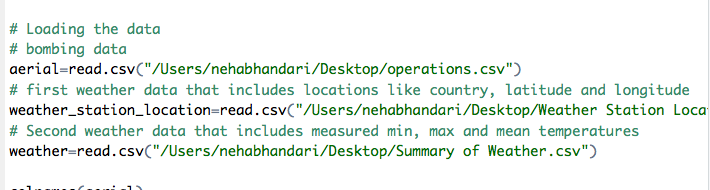
* + Aerial Bombing Operations in WW2
    - This data includes bombing operations. For example, USA who use Ponte olivo airfield bomb Germany (Berlin) with A36 air craft in 1945.
  + Weather Conditions in WW2
    - The weather conditions during ww2. For example, according to George Town weather station, average temperature is 23.88 in 1/7/1942.
    - This data set has 2 subsets in it. First one includes weather station locations like country, latitude and longitude.
    - Second one includes measured min, max and mean temperatures from weather stations.

**Below are the features that I will be using for my analysis:**

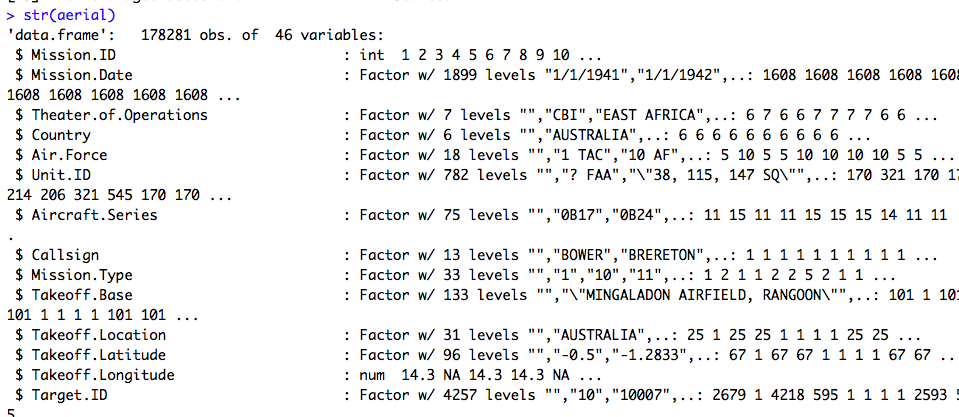
* **Aerial bombing Data description:**
  + Mission Date: Date of mission
  + Theater of Operations: Region in which active military operations are in progress; "the army was in the field awaiting action"; Example: "he served in the Vietnam theater for three years"
  + Country: Country that makes mission or operation like USA
  + Air Force: Name or id of air force unity like 5AF
  + Aircraft Series: Model or type of aircraft like B24
  + Call sign: Before bomb attack, message, code, announcement, or tune that is broadcast by radio.
  + Takeoff Base: Takeoff airport name like Ponte Olivo Airfield
  + Takeoff Location: Takeoff region Sicily
  + Takeoff Latitude: Latitude of takeoff region
  + Takeoff Longitude: Longitude of takeoff region
  + Target Country: Target country like Germany
  + Target City: Target city like Berlin
  + Target Type: Type of target like city area
  + Target Industry: Target industry like town or urban
  + Target Priority: Target priority like 1 (most)
  + Target Latitude: Latitude of target
  + Target Longitude: Longitude of target
* **Weather Condition data description:**
  + Weather station location:
    - WBAN: Weather station number
    - NAME: weather station name
    - STATE/COUNTRY ID: acronym of countries
    - Latitude: Latitude of weather station
    - Longitude: Longitude of weather station
  + Weather:
    - STA: weather station number (WBAN)
    - Date: Date of temperature measurement
    - MeanTemp: Mean temperature

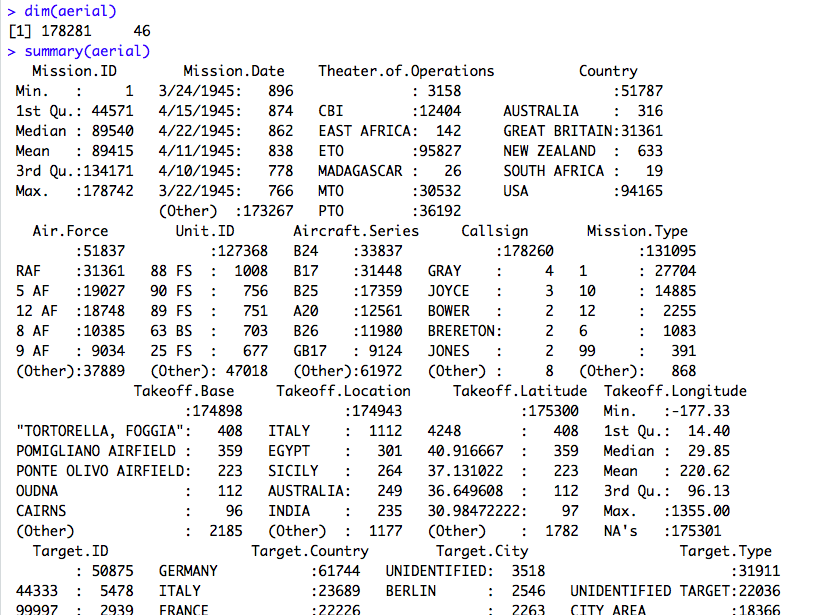
**4.Setting up the environment and importing the data**

****

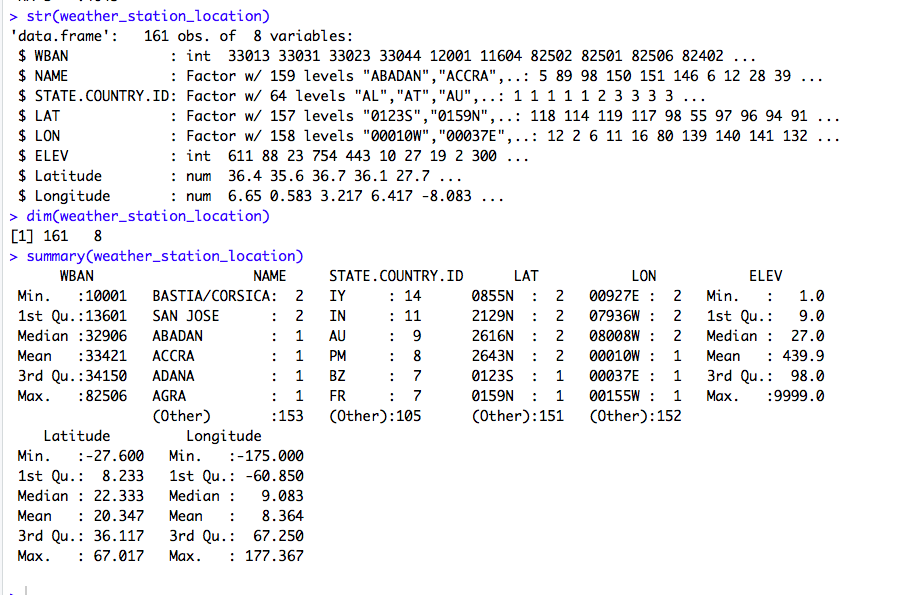


**AERIAL DATASET**

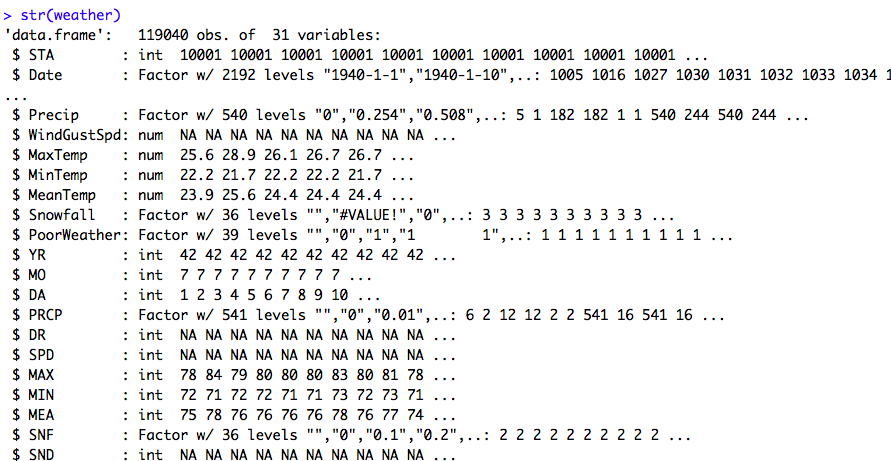




**WEATHER STATION LOCATION**



**WEATHER**



**5. Data Cleaning**

After generating various descriptions of the data, we can see that there are 178281 observations 46 columns in aerial data set ,161\*8 in weather location and 119040\*31 in weather data set.

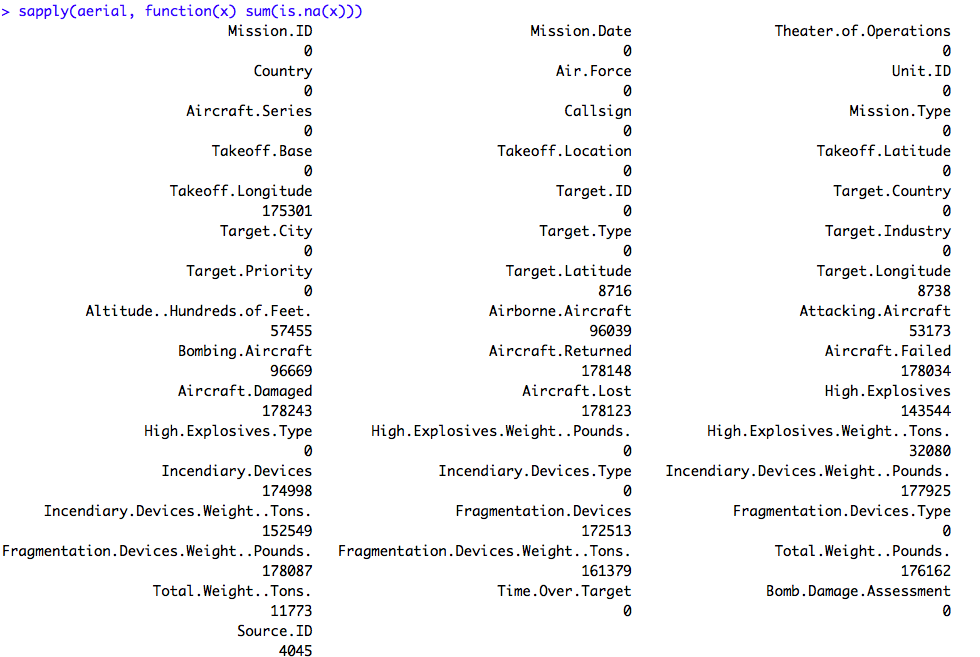
There are a lot of columns which are not useful for my analysis and I am going to drop them. In addition, I noticed that there are some extreme values for the longitude and latitude column which can be considered as outliers, so I will remove them also.

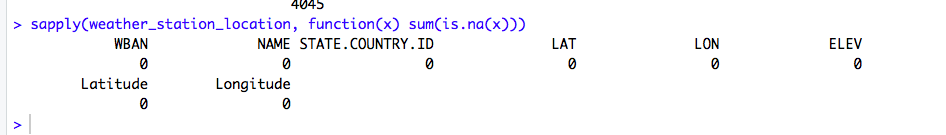
Of the remaining observations, there are some missing values which will be omitted.

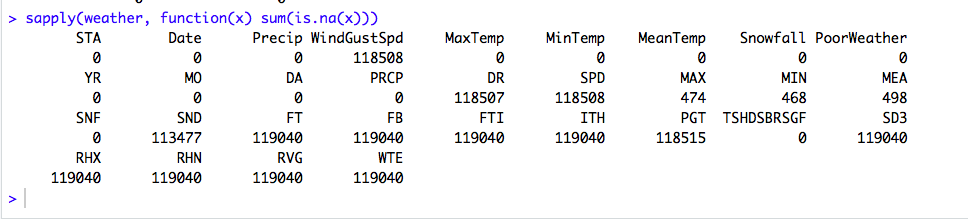
It does not only remove the uncertainty, but it also makes the visualization process easier.

* Drop countries that are NaN
* Drop if target longitude is NaN
* Drop if takeoff longitude is NaN
* Drop unused features

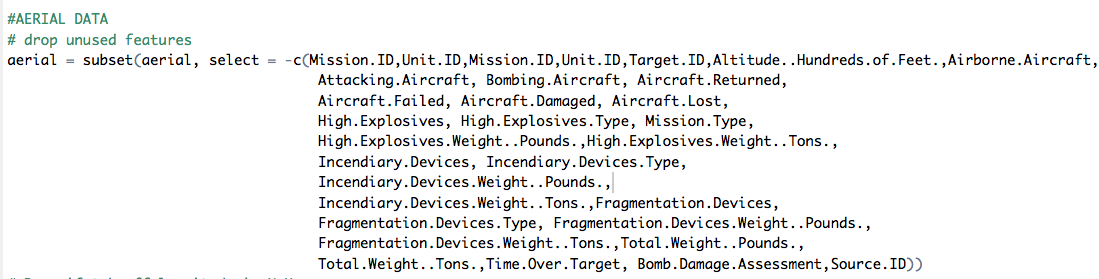
**BEFORE CLEANING THE DATA LOOKS LIKE THIS:**

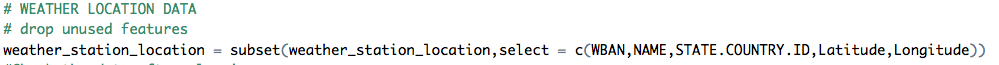


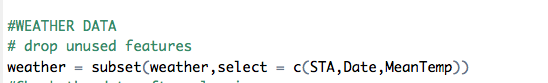




**Keeping only the necessary features**



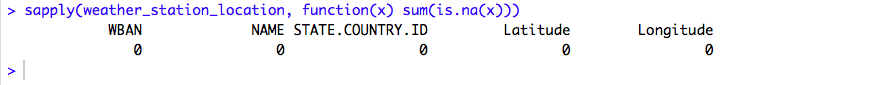


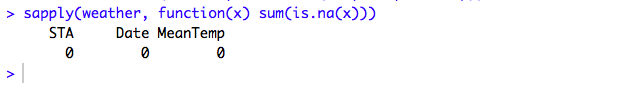


**This is how the data looks like after feature removal:**

**Most of the missing values are taken care of:**



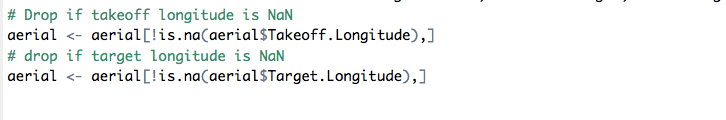




**Dropping the remaining rows with Nan values:**

The datasets weather station location and weather don’t have any Na values left. However, there are still some in the aerial dataset. We will now take care of that:

There are columns Target Longitude and Takeoff longitude with Na values and since they are used in the further analysis, it is important to drop those with no values.



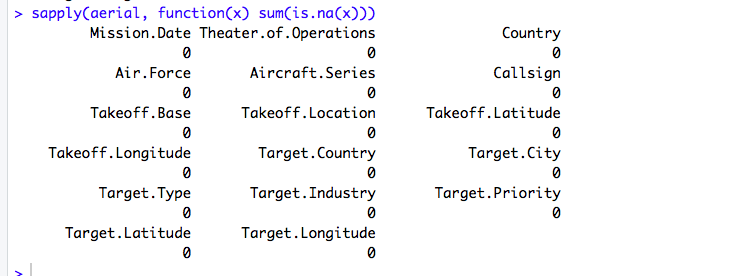
**Taking care of the Outliers:**

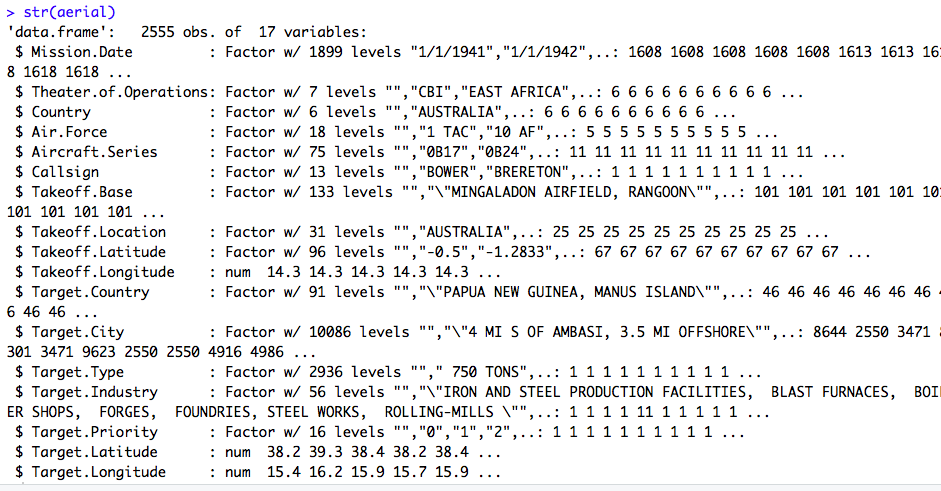
After checking the data, we see that Takeoff latitude and longitude have some extreme values:

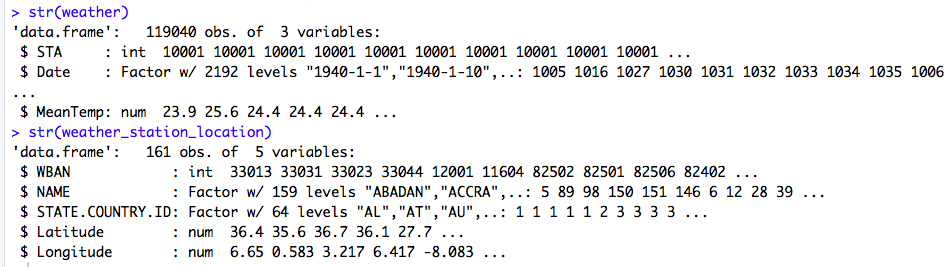
We will drop those rows as we don’t want our analysis to be affected by the outliers:



**After all the cleaning is done:**





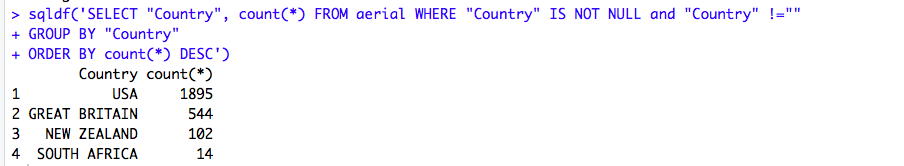


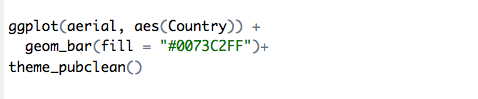
**6.Data Visualization**

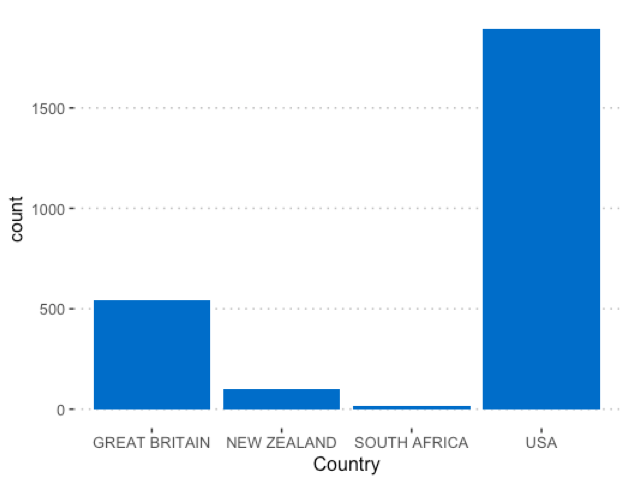
**I will start with basics of visualization to understand our data:**

* How many countries are there which attack
* What are the top target countries
* What are Top 10 aircraft series
* What are the Takeoff base locations (for the Attack countries)
* Target locations
* Bombing paths for the countries
* Theater of Operations
* Weather station location

**The countries which attack most of the times:**

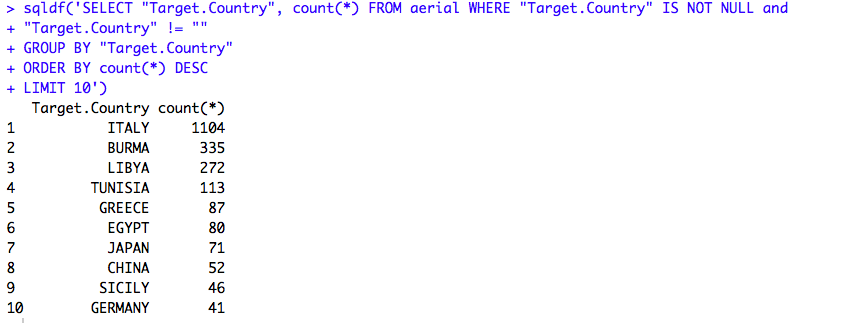


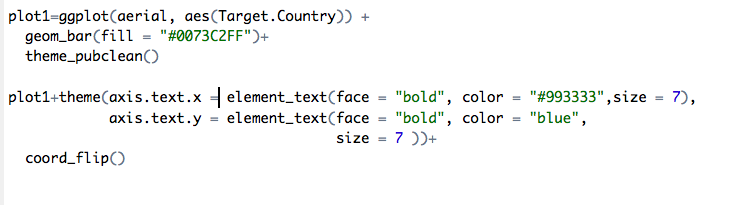


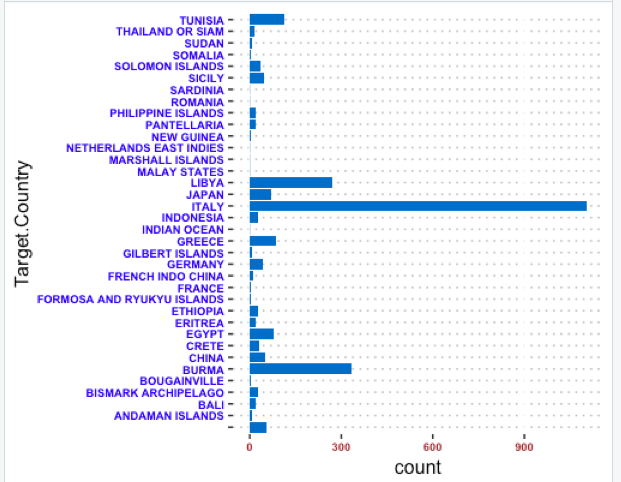


Here we can see that USA does most of the attacks followed by Great Britain.

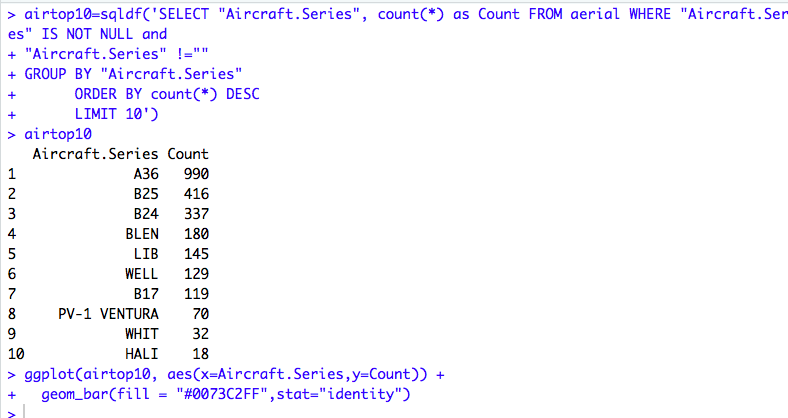
**Next, we look at the top target countries (countries which get attacked most)**

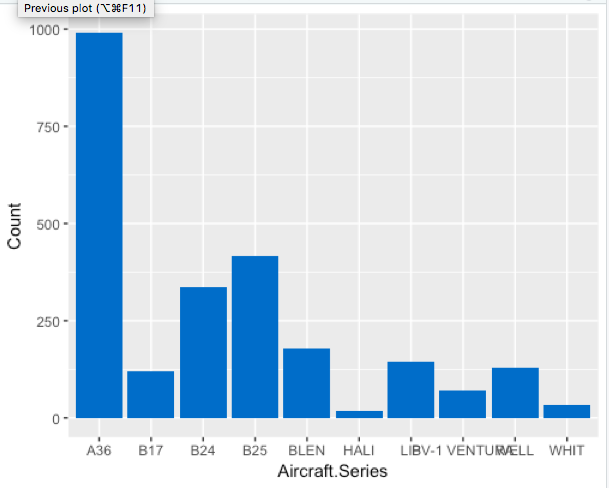






**Now let’s see which were the aircrafts used for most of the attacks:**





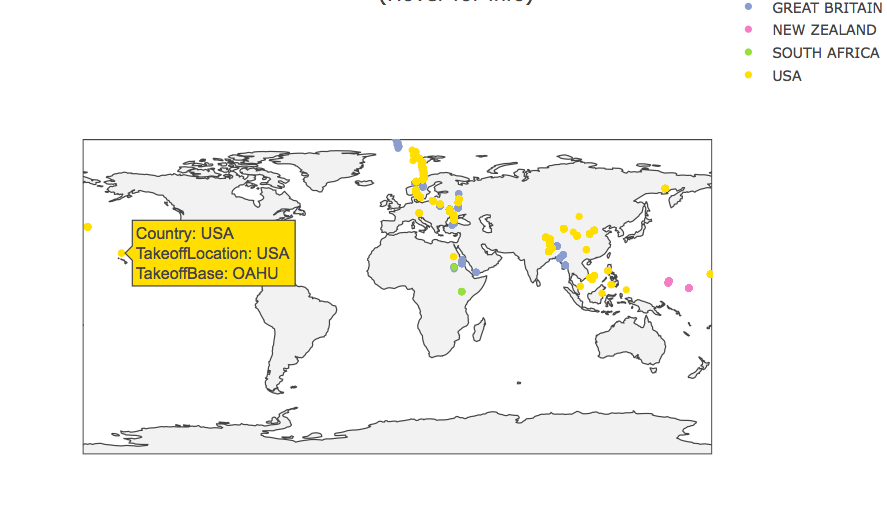
We can see that A36 was used most of the times for the attacks.

**Now let’s visualize take off bases and location of countries who attack**

****

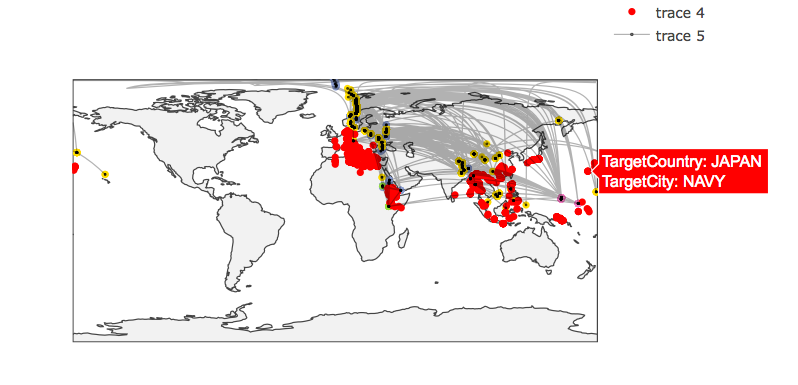
**In plot below, yellow color draws the attention, it is USA and blue color is Great Britain**

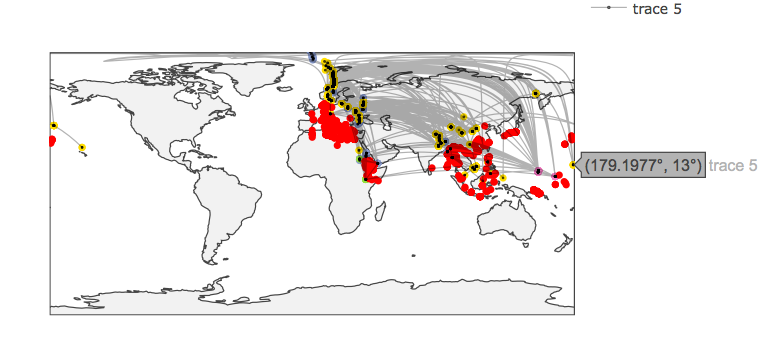
****



**Now let’s visualize bombing paths: which country from which take off base locations bomb which countries and cities**



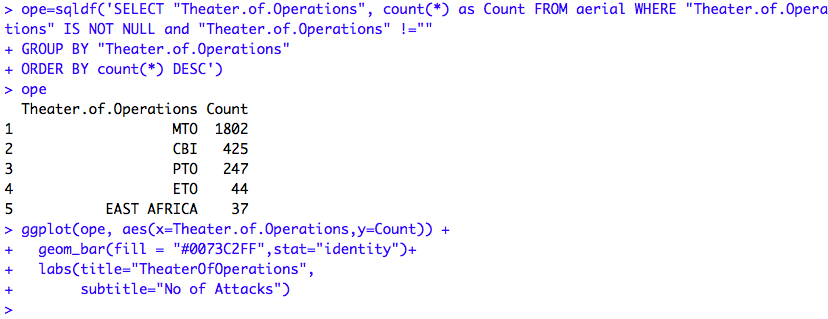


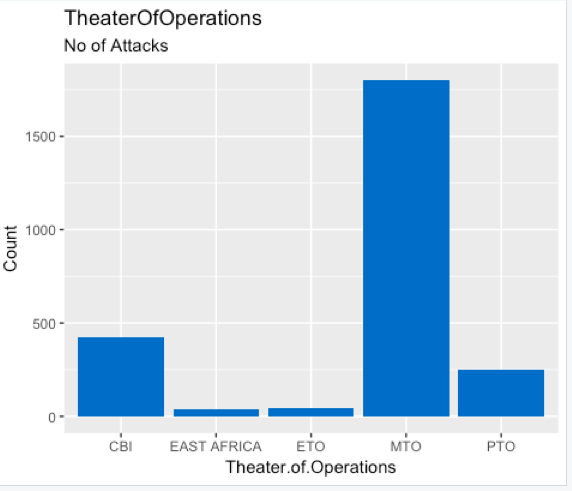


The grey lines are bombing paths from the countries who attack to the target countries. It shows the path connecting the takeoff longitude and takeoff latitude to the target longitude and target latitude.

We can see from the above graphs that most of the attacks happened in the Mediterranean Area

**Let’s visualize The Theater of Operations**





Most of the bombing attack is done in Mediterranean theater of operations.

**Different Theater of Operations:**

* ETO: European Theater of Operations
* PTO: Pacific Theater of Operations
* MTO: Mediterranean Theater of Operations
* CBI: China-Burma-India Theater of Operations
* EAST AFRICA: East Africa Theater of Operations

**We have other data sets which talk about the weather conditions during world war 2.**

**Now we will visualize the weather station locations like country, latitude and longitude**

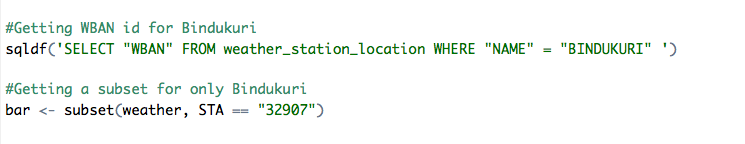


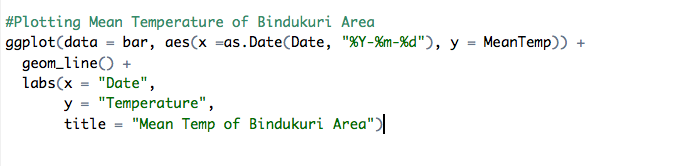


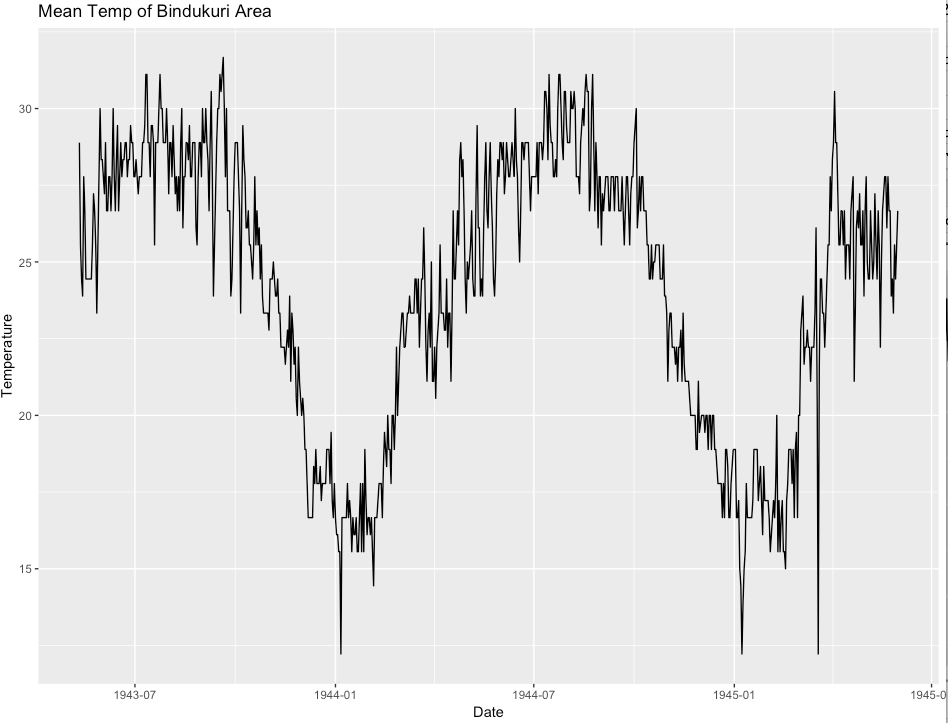
**For my analysis and prediction, I will focus on the USA & BURMA war.**

* In this war USA bomb BURMA (KATHA city) from 1942 to 1945.
* The closest weather station to this war is BINDUKURI and it has temperature record from 1943 to 1945.

**Let’s visualize the mean temperature of this station:**



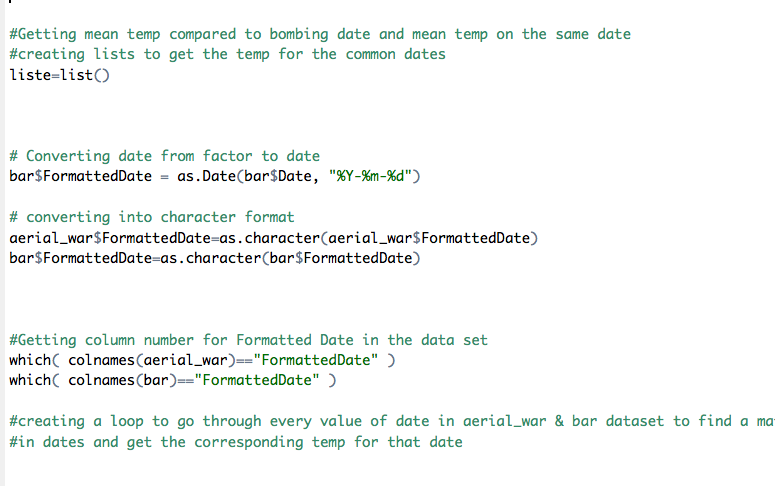


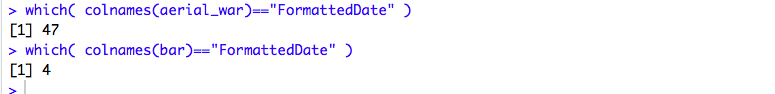


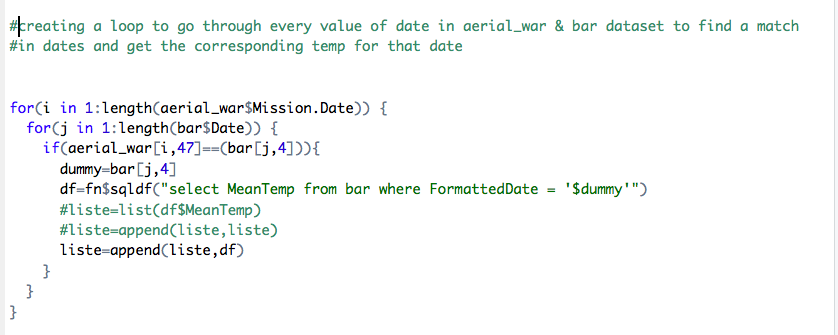
* As seen, we have temperature measurement from 1943 to 1945.
* Temperature oscillates between 12 and 32 degrees.
* Temperature of winter months is colder than temperature of summer months

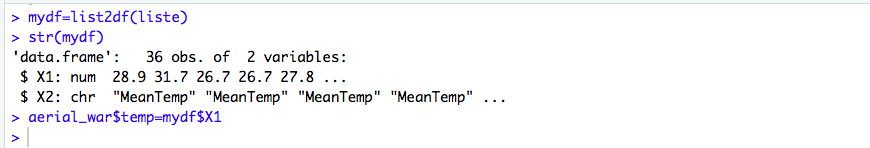
**Now let’s understand how the mean temperature of this Bindukuri area compares with the bombing temperature (USA) on the same dates.**

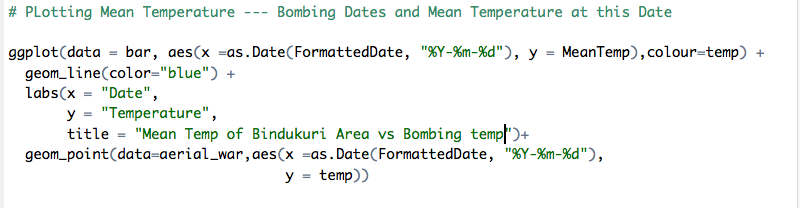


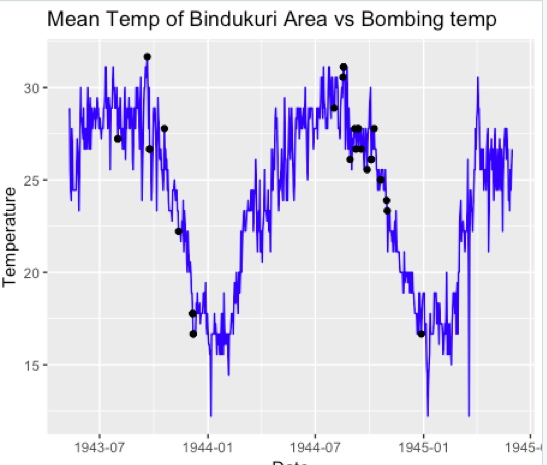












* Blue lines are mean temperature that is measured in Bindukuri.
* Black markers are bombing dates and bombing date temperature.
* As it can be seen from plot, USA bomb at high temperatures:
* The question is that can we predict future weather and according to this prediction can we know whether bombing will be done or not.
* In order to answer this question lets first start with time series prediction

**7.Time Series Analysis**

Time-series analysis is a basic concept within the field of statistical learning that allows the user to find meaningful information in data collected over time.

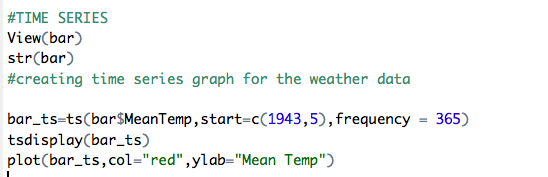
To demonstrate the power of this technique, we applied it to our dataset to forecast the future weather mean temperatures and predict when bombing operations can be done.

We will first decompose the time series to check for trend, seasonality and level in the data.

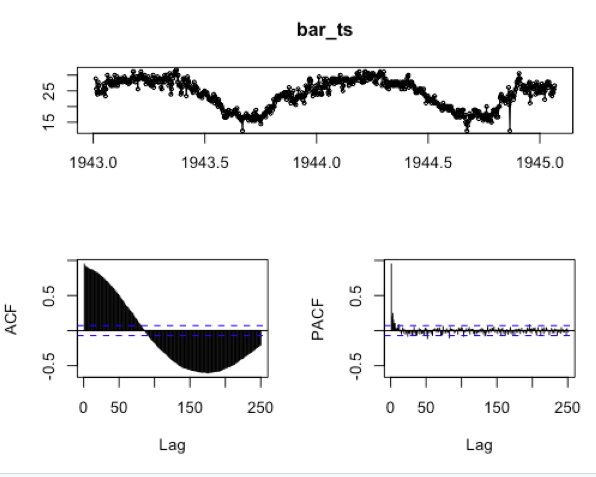
I will use different forecasting techniques such as ARIMA, SMA, Holt Winters etc. to predict the values and finally compare their accuracies.

Visualizing our time-series data enables us to make inferences about important components, such as trend, seasonality, heteroscedasticity, and stationarity. Here is a quick summary of each:

* **Trend**: We say that a dataset has a trend when it has either a *long-term increase* or *decrease*.
* **Seasonality**: We say that a dataset has [seasonality](https://www.datascience.com/blog/seasonality-what-it-means-and-how-you-can-leverage-it/) when it has patterns that repeat over known, fixed periods of time (e.g. monthly, quarterly, yearly).
* **Heteroscedasticity**: We say that a data is heteroskedastic when its variability is not constant (i.e., its variance increases or decreases as a function of the explanatory variable).
* **Stationarity**: A stochastic process is called *stationary* if the mean and variance are constant (i.e., their joint distribution does not change over time).





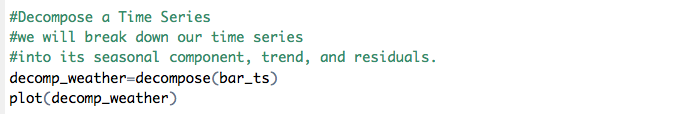


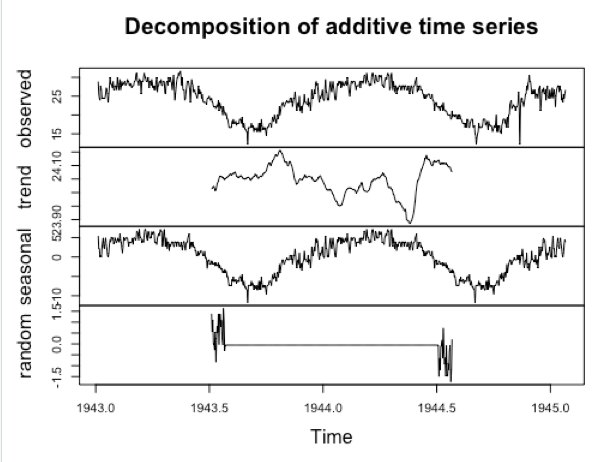
We can easily see that our time series has instances of both positive and negative trend. Overall, it is very volatile.

**Notice**: Here in the plot 1943.0 signify May’1943 as our data starts from that date.

**Decomposition of Time Series**

Beyond understanding the trend of your time series, you want to further understand the anatomy of your data. For this reason, we will break down our time series into its seasonal component, trend, and residuals.



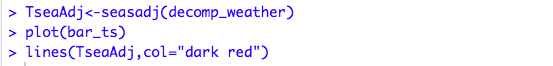


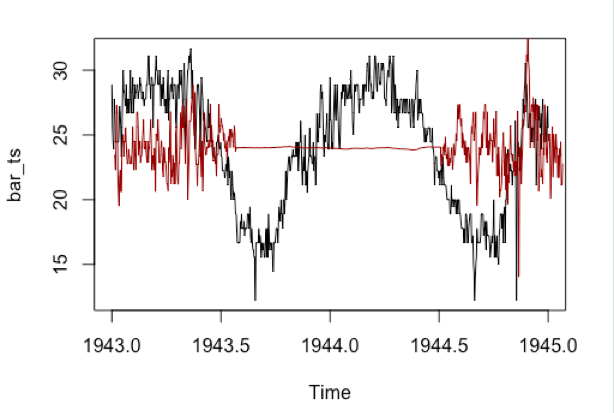
Also, we can see from plot above, our time series has seasonal variation. In summer, mean temperature is higher and in winter mean temperature is lower for each year.

**De-seasonalize a time series:**

De-seasonalizing throws insight about the seasonal pattern in the time series and helps to model the data without the seasonal effects.

If we have a seasonal time series that can be described using an additive model, we can seasonally adjust the time series by estimating the seasonal component and subtracting the estimated seasonal component from the original time series.



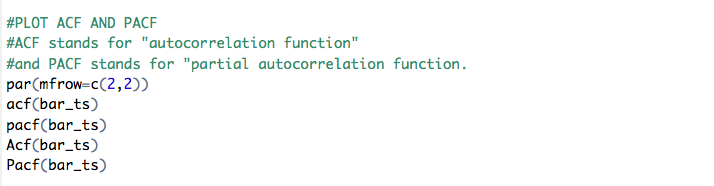


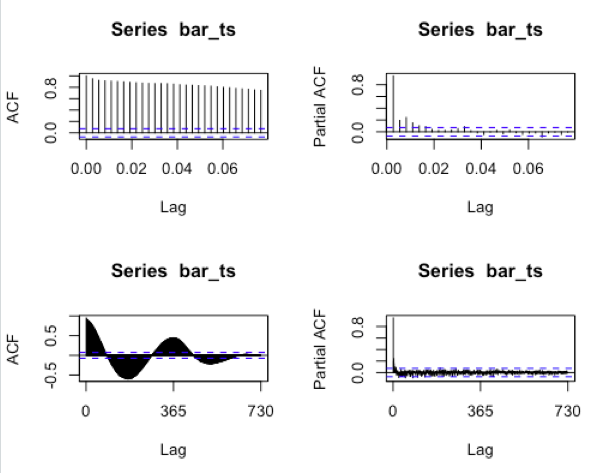
**ACF & PACF**

**Autocorrelation** is the correlation of a Time Series with lags of itself. This is a significant metric because,

* It shows if the previous states (lagged observations) of the time series has an influence on the current state. In the autocorrelation chart, if the autocorrelation crosses the dashed blue line, it means that specific lag is significantly correlated with current series.
* It is used commonly to determine if the time series is stationary or not. A stationary time series will have the autocorrelation fall to zero fairly quickly but for a non-stationary series it drops gradually.

**Partial Autocorrelation** is the correlation of the time series with a lag of itself, with the linear dependence of all the lags between them removed.



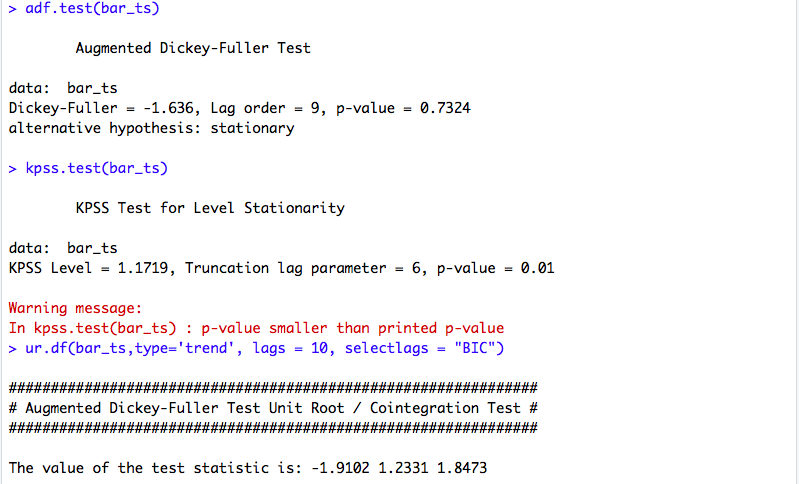


**Test for Stationarity**

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary through the use of mathematical transformations. A stationarized series is relatively easy to predict: you simply predict that its statistical properties will be the same in the future as they have been in the past!

We will utilize a few statistical methods to test for stationarity. We must be wary of our model having a unit root; this will lead to non-stationary processes. We will utilize the Augmented Dickey-Fuller Test for stationarity. The null hypothesis states that large p values indicate non-stationarity and smaller p values indicate stationarity. (We will be using 0.05 as our alpha value.)

We all use KPSS test to check for stationarity. Using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test we will test the null hypothesis of trend stationarity (a low *p*-value will indicate a signal that is not trend stationary, has a unit root):



We can see our p value for the ADF test is relatively high and p value of KPP is quite low. So, our time series is not stationary.

**Make a Time Series Stationary?**

A time series is said to be stationary if it holds the following conditions true.

* The mean value of time-series is constant over time, which implies, the trend component is nullified.
* The variance does not increase over time.
* Seasonality effect is minimal.

This means it is devoid of trend or seasonal patterns, which makes it looks like a random *white noise* irrespective of the observed time interval.

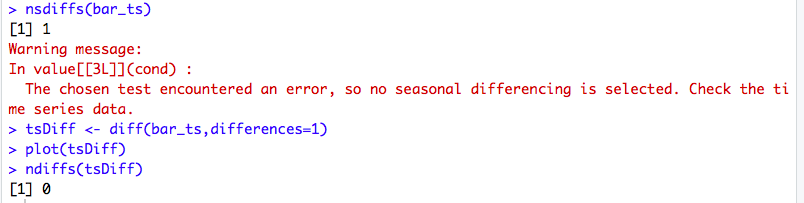
**Differencing method**: Differencing a time series means, to subtract each data point in the series from its successor. It is commonly used to make a time series stationary. For most time series patterns, 1 or 2 differencing is necessary to make it a stationary series.

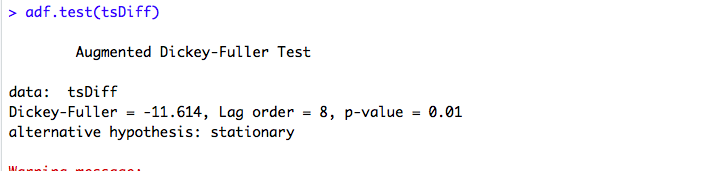
A way to make a time series stationary is to find the difference across its consecutive values. (take difference between time series and shifted time series)

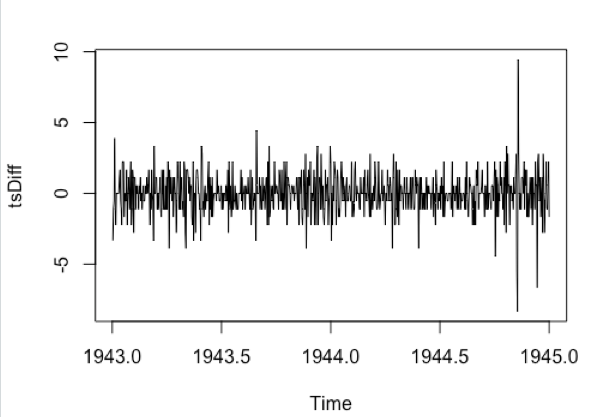
This helps stabilize the mean, thereby making the time-series object stationary.

Since our time series appears to be seasonal, a better approach is to difference with respective season’s data points to remove seasonal effect.

After that, if needed, difference it again with successive data points.

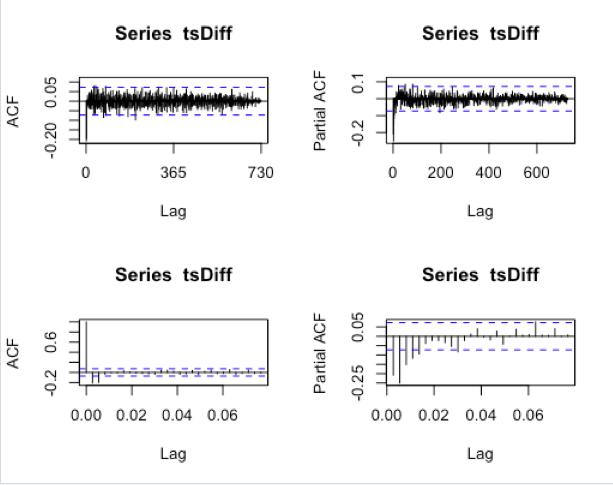






**ACF/PACF PLOTS AFTER MAKING IT STATIONARY**





MODELS FOR FORECASTING

SIMPLE MOVING AVERAGES

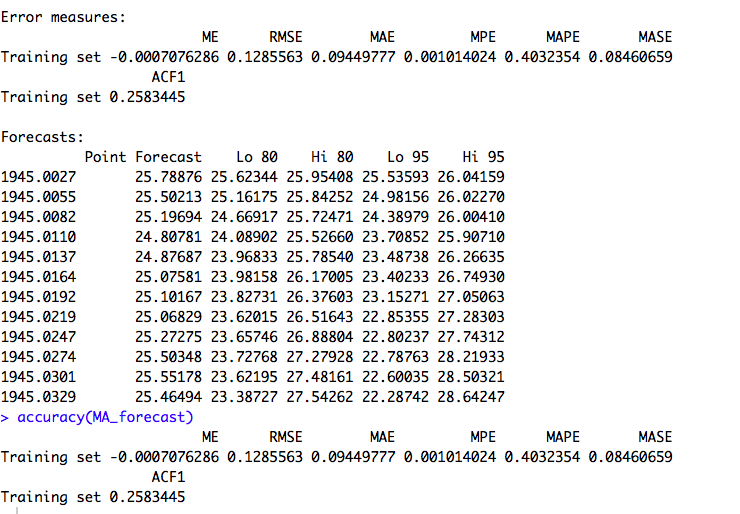
Smoothing methods are a family of forecasting methods that average values over multiple periods in order to reduce the noise and uncover patterns in the data.

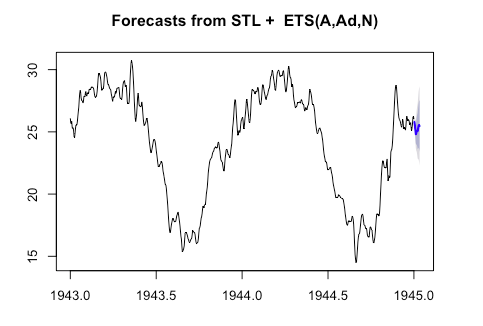
Simple Moving Average is a method of time series smoothing and is actually a very basic forecasting technique. It does not need estimation of parameters, but rather is based on order selection.

The moving average smoother averages the nearest order periods of each observation. As neighboring observations of a time series are likely to be similar in value, averaging eliminates some of the randomness in the data, leaving a smooth trend-cycle component.

Moving averages is a smoothing approach that averages values from a window of consecutive time periods, thereby generating a series of averages.





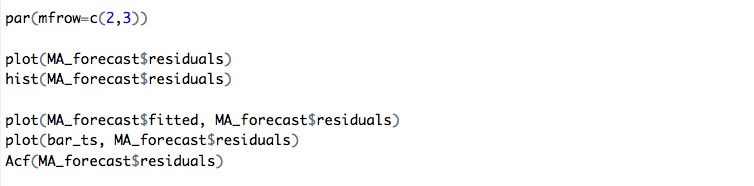


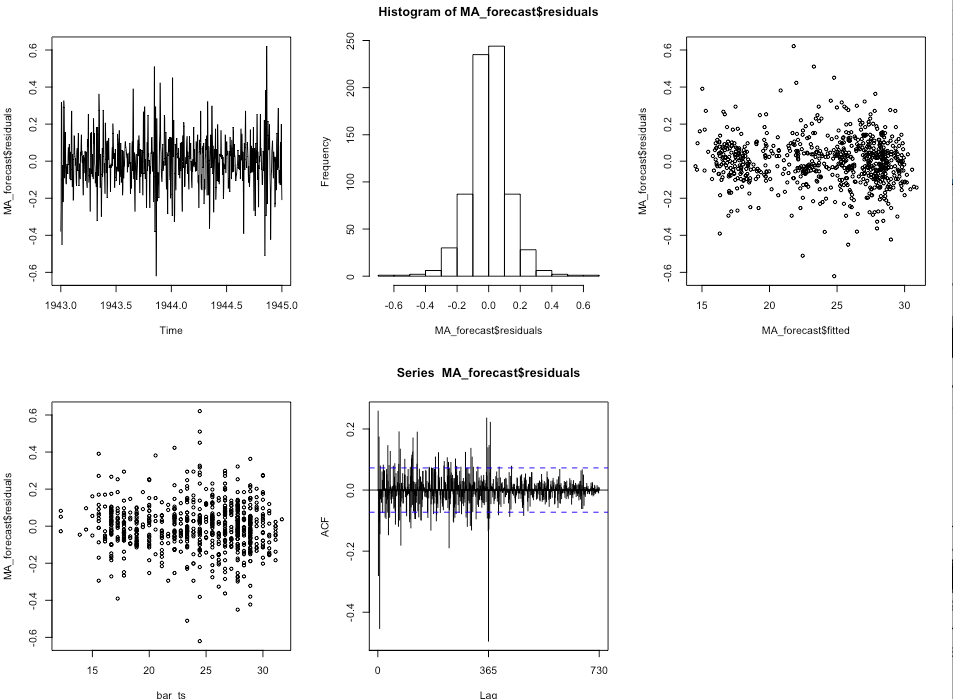
MA assigns equal weighting to all values. The SMA may rely too heavily on outdated data since it treats the 10th or 200th day's impact just as much as the first or second.

**Our next step is to run residual diagnostics to ensure our residuals are white noise under our initial assumptions:**

Residuals are useful in checking whether a model has adequately captured the information in the data. A good forecasting method will yield residuals with the following properties: The residuals are uncorrelated.

If the residuals have a mean other than zero, then the forecasts are biased.



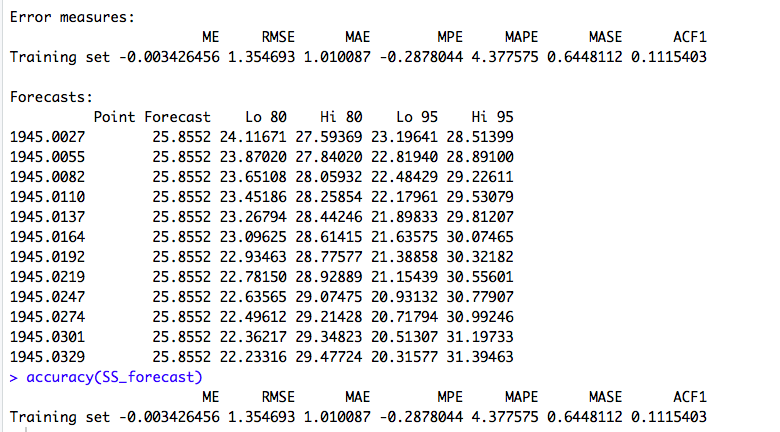


SIMPLE EXPONENTIAL SMOOTHING

**Exponential smoothing** is a technique for smoothing time series data using the exponential window function. Whereas in the SMA the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time. It is an easily learned and easily applied procedure for making some determination based on prior assumptions by the user, such as seasonality.

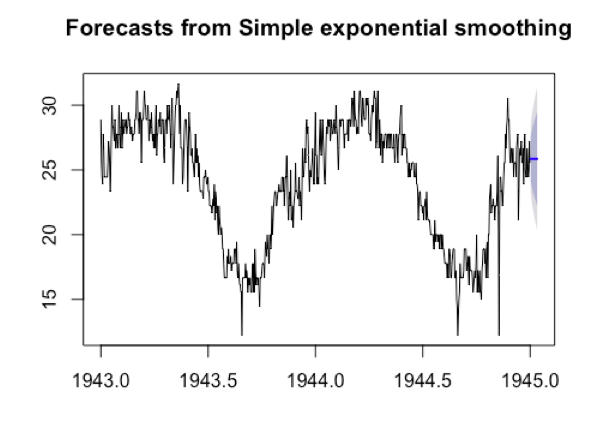
Where naïve  places 100% weight on most recent observation and [moving averages](http://uc-r.github.io/ts_moving_averages) place equal weight on *k* values, exponential smoothing allows for weighted averages where greater weight can be placed on recent observations and lesser weight on older observations.





The simple exponential smoothing method provides a way of estimating the level at the current time point. Smoothing is controlled by the parameter alpha (smoothing constant); for the estimate of the level at the current time point. The value of alpha lies between 0 and 1. Values of alpha that are close to 0 mean that little weight is placed on the most recent observations when making forecasts of future values.

For our data, α = 0.5, puts a weight of .5 on the most recent observation and a weight of 1 − .5 = .5 on the most recent forecast. With a relatively small value of α, the smoothing will be relatively more extensive.  With a relatively large value of α, the smoothing is relatively less extensive as more weight will be put on the observed value.

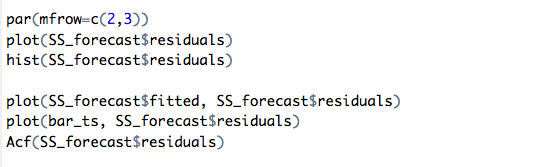


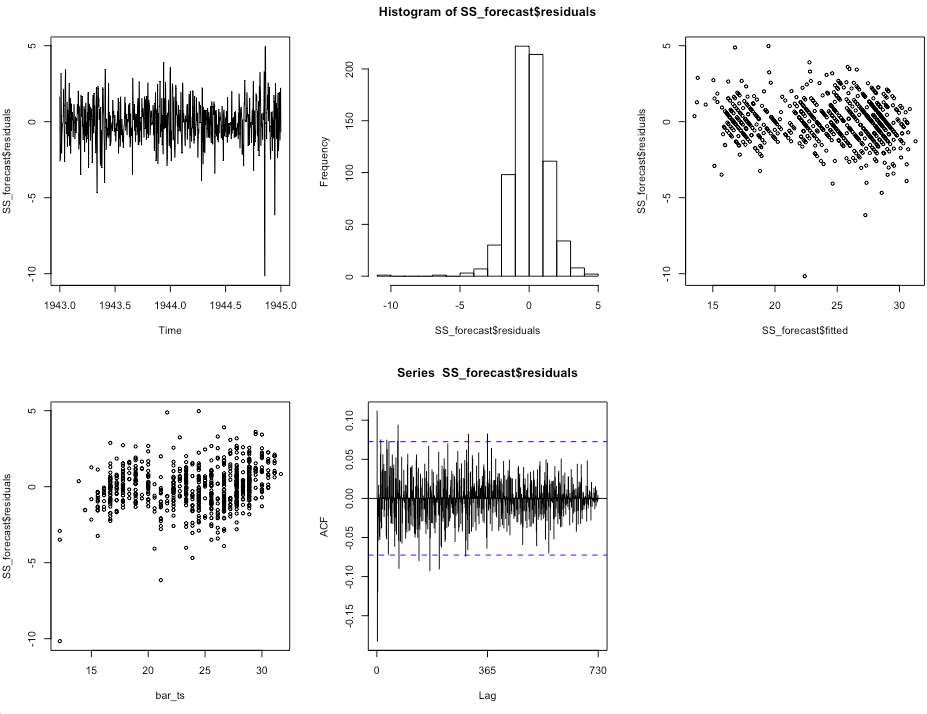
The key point to remember is that SES is suitable for data with no trend or seasonal pattern

because the algorithm gives more weight to the most recent observation; therefore, recent changes in the data will have a bigger impact on forecasted values.

We see that our forecast projects a flat lined estimate into the future, which does not capture the seasonality or trend in the data. This is why SES should not be used on data with a trend or seasonal component.

Running residual diagnostics:





Residuals are randomly distributed across the data. The residuals plot is random and have high variation over the years.

It shows that the residuals fluctuate across the data therefore variance cannot be treated as constant since it is random.

HOLT WINTERS

To make predictions using data with a trend and seasonality, we turn to the Holt-Winters Seasonal Method.

The Additive model is used because the seasonal trend is of the same magnitude throughout the data set.

Triple Exponential Smoothing is an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series.

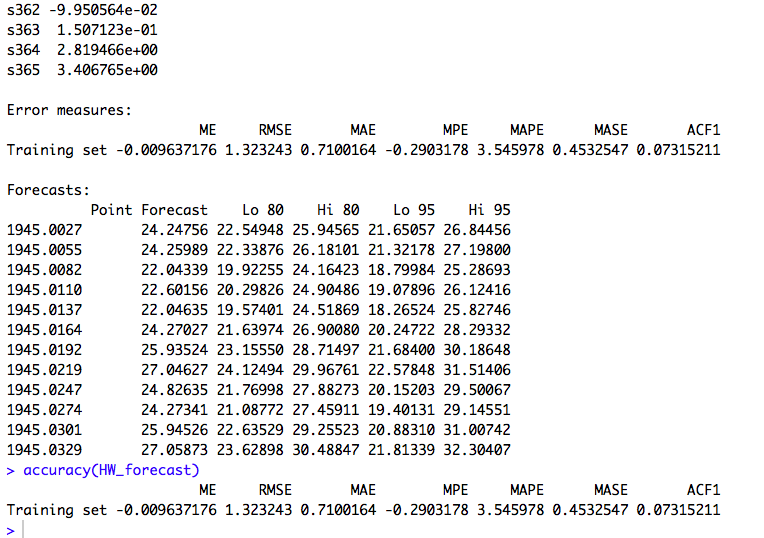
In addition to the alpha and beta smoothing factors, a new parameter is added called gamma (g) that controls the influence on the seasonal component.

Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality). Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict “typical” values for the present and future.

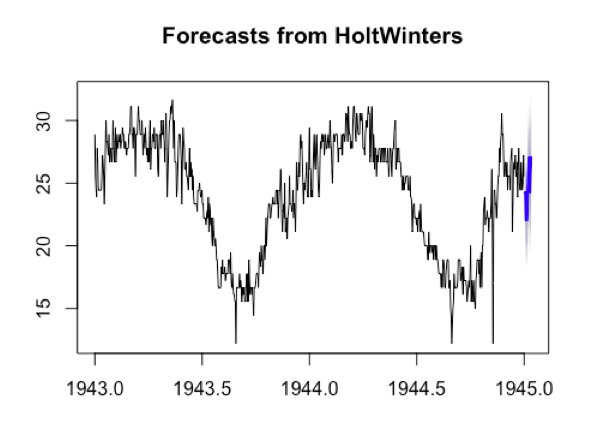
The three aspects of the time series behavior—value, trend, and seasonality—are expressed as three types of exponential smoothing, so Holt-Winters is called triple exponential smoothing.

Smoothing is controlled by three parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point. The parameters alpha, beta and gamma all have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.

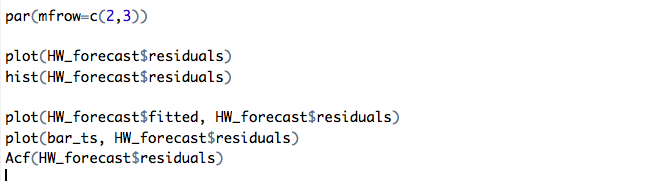


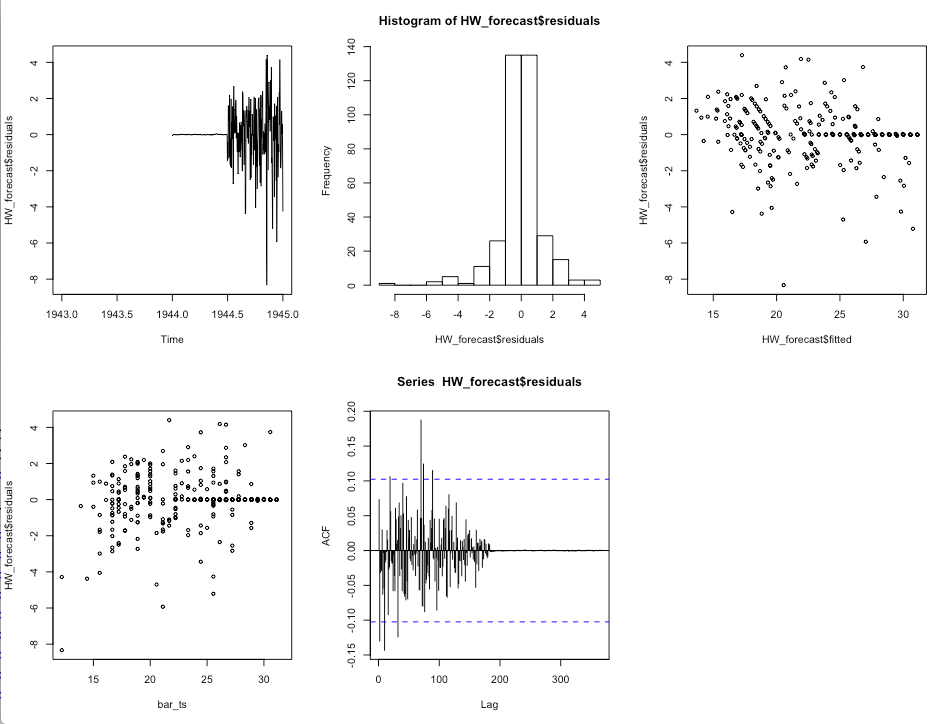


The value of alpha (0.52) is in between, indicating that the estimate of the level at the current time point is based upon both recent observations and some observations in the more distant past. The value of beta is 0.00, indicating that the estimate of the slope b of the trend component is not updated over the time series, and instead is set equal to its initial value. This makes good intuitive sense, as the level changes quite a bit over the time series, but the slope b of the trend component remains roughly the same. In contrast, the value of gamma (1) is high, indicating that the estimate of the seasonal component at the current time point is just based upon very recent observations.



Checking the residuals:





A good forecasting method will have residuals scattered and uncorrelated.

Exponential smoothing methods are useful for making forecasts and make no assumptions about the correlations between successive values of the time series. However, if we want to make prediction intervals for forecasts made using exponential smoothing methods, the prediction intervals require that the forecast errors are uncorrelated and are normally distributed with mean zero and constant variance.

Autoregressive Integrated Moving Average (ARIMA) models include an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.

ARIMA

ARIMA stands for auto-regressive integrated moving average and is specified by these three order parameters: *(p, d, q)*. The process of fitting an ARIMA model is sometimes referred to as the Box-Jenkins method.

ARIMA models work on the following assumptions –

* The data series is stationary, which means that the mean and variance should not vary with time. A series can be made stationary by using log transformation or differencing the series.
* The data provided as input must be a univariate series, since arima uses the past values to predict the future values.

ARIMA has three components – AR (autoregressive term), I (differencing term) and MA (moving average term). Let us understand each of these components –

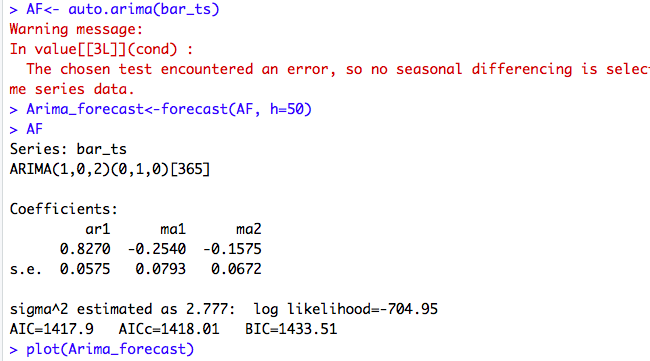
* AR term refers to the past values used for forecasting the next value. The AR term is defined by the parameter ‘p’ in arima. The value of ‘p’ is determined using the PACF plot.
* MA term is used to defines number of past forecast errors used to predict the future values. The parameter ‘q’ in arima represents the MA term. ACF plot is used to identify the correct ‘q’ value.
* Order of differencing specifies the number of times the differencing operation is performed on series to make it stationary. Test like ADF and KPSS can be used to determine whether the series is stationary and help in identifying the d value

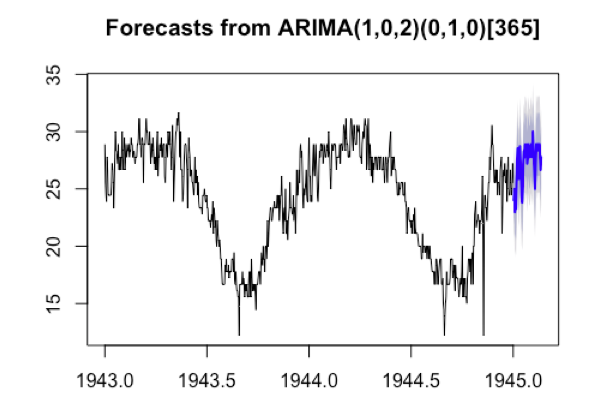
ARIMA models can be also specified through a seasonal structure. In this case, the model is specified by two sets of order parameters: (p, d, q) as described above and parameters describing the seasonal component of m periods.

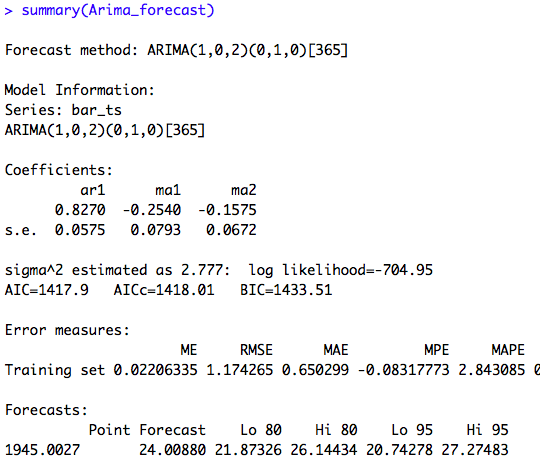
So a seasonal ARIMA model is classified as an ARIMA(p,d,q)x(P,D,Q) model, where P=number of seasonal autoregressive (SAR) terms, D=number of seasonal differences, Q=number of seasonal moving average (SMA) terms

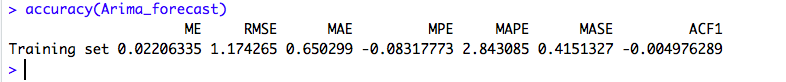
Although ARIMA is a very powerful model for forecasting time series data, the data preparation and parameter tuning processes end up being really time consuming. Before implementing ARIMA, we need to make the series stationary, and determine the values of p and q using the plots we discussed above. Auto ARIMA makes this task really simple for us as it eliminates these steps.

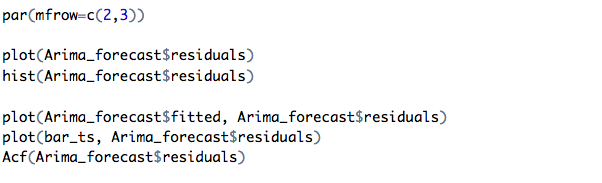
Auto ARIMA takes into account the AIC and BIC values generated to determine the best combination of parameters. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values are estimators to compare models. The lower these values, the better is the model.

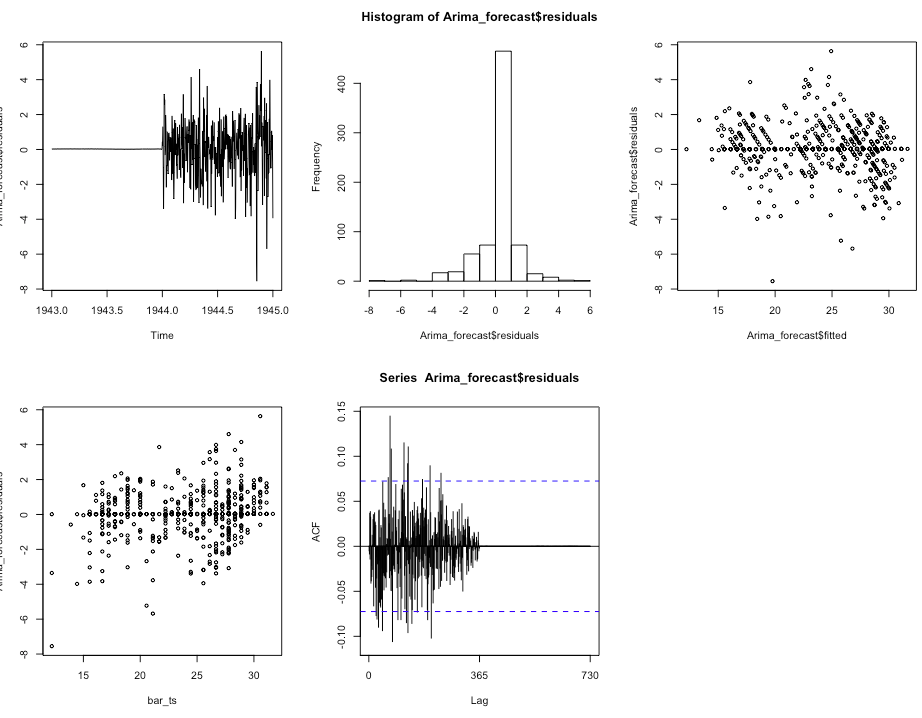








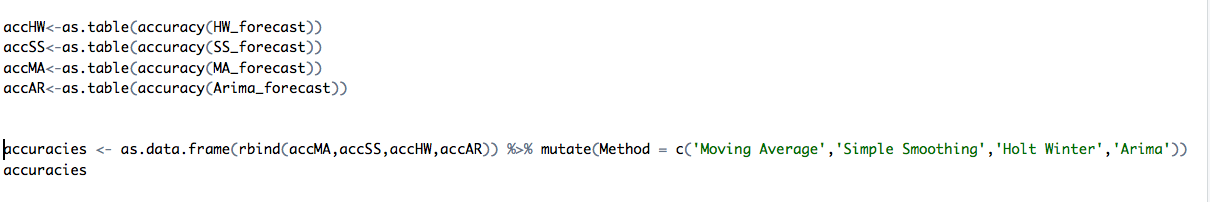


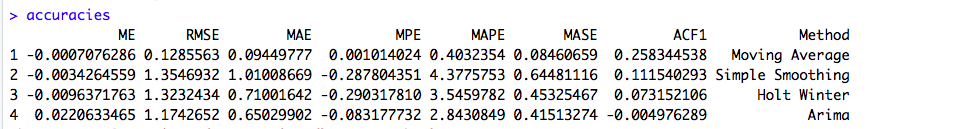


The forecast errors are normally distributed with mean zero and constant variance.

**8.Results and Discussion**

Summarizing all the models based on their accuracies:





A forecast “error” is the difference between an observed value and its forecast.

The simplest measure of forecast accuracy is called [Mean Absolute Error (MAE)](http://en.wikipedia.org/wiki/Mean_absolute_error). MAE is simply, as the name suggests, the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value.

[Mean Absolute Percentage Error (MAPE)](http://en.wikipedia.org/wiki/Mean_absolute_percentage_error" \o "mean absolute percentage error" \t "_blank) allows us to compare forecasts of different series in different scales.

To adjust for large rare errors, we calculate the [Root Mean Square Error (RMSE)](http://www.math-interactive.com/products/calgraph/help/fit_curve_to_data/root_mean_squared_error.htm)

The forecasting method we used to find the best model receives the lowest MAE, MAPE & RMSE.

**Simple Exponential Smoothing** This model smoothens the data using the exponential window function and is used to assign exponentially decreasing weights over time. More useful when recent observations need to be given more weightage than past observations.

**Holt Winter** is a good method of forecasting as it covers all the components.

**Moving average** depends on the order, with the increase in order the curve is smooth but for lower orders it is not recommendable.

**ARIMA** a powerful tool for accurate short-range forecast. Models are quite flexible and can represent a wide range of characteristics of time series occurring in practices, but a large amount of data is required. In our case, it works well given the huge range of data. Model has to be periodically completely refitted or new model developed.

**9. Conclusion**

The most common evaluation metrics for forecasting are RMSE, which we have used on regression problems; MAPE, as it is scale-independent and represents the ratio of error to actual values as a percent; and MASE, which indicates how well the forecast performs compared to a naïve average forecast.

Surprisingly, Moving Average performed better on the data set than the ARIMA model. This could be a result of overfitting. Also, for our weather data the values are not changing rapidly, and we have them for longer time frames(daily)

However, as compared to other models remaining ARIMA performed better and is better at predicting the future values.

We see that our data set has an increasing /decreasing trend, which can be said to be very unpredictable.

However, there is seasonality where the area shows high temperature in summers and lower in winters. The reason is obvious as that is how the seasons are perceived.

Since the seasonality factor plays an important role in the data, the arima model performed better to forecast the future values.