**Tornado Prediction and Impact Analysis**

Table of Contents

[1. Introduction 2](#_Toc500275175)

[2. Data 2](#_Toc500275176)

[2. Problems to be Solved 3](#_Toc500275177)

[3. Data Processing 3](#_Toc500275178)

[5. Methods and Process 4](#_Toc500275179)

[6. Evaluations and Results 39](#_Toc500275180)

[6.1. Evaluation Methods 39](#_Toc500275181)

[6.2. Results and Findings 49](#_Toc500275182)

[7. Conclusions and Future Work 51](#_Toc500275183)

[7.1. Conclusions 51](#_Toc500275184)

[7.2. Limitations 51](#_Toc500275185)

[7.3. Potential Improvements or Future Work 51](#_Toc500275186)

# **Introduction**

Natural disasters around the world have always had a serious impact on the people and property. The destruction varies largely from property damage to even many more deaths of people. United States is prone to such serious natural disasters for more than 80 years from the past and is still suffering because of such natural disasters. At present, we have intelligent disaster mitigation plans in place which helps us to plan and manage the resources efficiently in such a way that the loss incurred by any disaster is reduced.

Tornado is one such natural disaster and has had severe bad impacts on the lives of people and on the property and agriculture. Tornado is the most violent of all atmospheric that has had severe impacts on people and property. Around 1200 tornadoes hit the US yearly. Hence, National Oceanic and Atmospheric Administration’s National Weather Service Storm Prediction Center generates severe reports and maintains it in its Geographic Information System(GIS) database.

In this project we collected data from the GIS report on tornadoes and examined the occurrence of tornadoes over 60-65 years in the past, apply the different analytical methodologies and deduce the relationship between several factors that would aid in predicting the severity of the tornadoes, property loss and crop loss caused by any tornado in future. This study would not only help to predict the tornadoes severity and loss incurred but also determines if the application of science and technology in disaster management has really reduced the impact of such disasters on people, property and agriculture.

# **2. Data**

The data set belongs to Database of tornado activity from 1950 to 2016 created by NOAA’s National Weather Service Storm Prediction Center (data set available at <http://www.spc.noaa.gov/gis/svrgis/>) to enhance understanding of where tornados happen, indicators of damage, and weather conditions associated with tornados.

Metadata available at <http://www.spc.noaa.gov/wcm/data/SPC_severe_database_description.pdf>

Reference: <https://www.kaggle.com/jtennis/spctornado>

The data contains 62208 records.

Attributes:

|  |  |  |
| --- | --- | --- |
| **No** | **Attribute** | **Description** |
| 1 | Om | Tornado number |
| 2 | Yr | Year when respective Tornado occurred |
| 3 | Mo | Month of the year when respective Tornado occurred |
| 4 | Dy | Day of the year when respective Tornado occurred |
| 5 | Date | Exact date of tornado |
| 6 | Time | Exact time of tornado |
| 7 | Tz | Timezone of the region where tornado occurred |
| 8 | St | State where tornado occurred |
| 9 | Stf | State FIPS number |
| 10 | Stn | Monitoring station of tornado |
| 11 | Mag | Magnitude of tornado |
| 12 | Inj | Injuries because of tornado |
| 13 | Fat | Fatalities because of tornado |
| 14 | Loss | Estimated Property loss because of tornado in millions of dollars |
| 15 | Closs | Estimated Crop loss because of tornado in millions of dollars |
| 16 | Slat | Tornado starting latitude in degrees |
| 17 | Slon | Tornado starting longitude in degrees |
| 18 | Elat | Tornado ending latitude in degrees |
| 19 | Elon | Tornado ending longitude in degrees |
| 20 | Len | Length of tornado in miles |
| 21 | Wid | Width of tornado in yards |
| 22 | Fc | Altered or unaltered f-scale rating |

# **Problems to be Solved**

Below are the list of problems that would be researched in this project using the above mentioned dataset

1. Is the distance travelled by the tornadoes of magnitude 3 and 4 are the same?
2. Is the number of injured people affected by the states Alabama and Texas are the same?
3. Is the average number of fatalities caused by the tornadoes is 4?
4. Is the number of states affected by tornadoes of magnitude 0 to 3 , is 3?
5. Is the property loss incurred by different regions across the US the same or not?
6. Is the no. of injuries due to tornadoes of different magnitudes are the same or not?
7. Is it possible to deduce magnitude category from length, width and distance travelled by a tornado?
8. Is it possible to deduce the property loss category from magnitude category of tornado, region of occurrence and no. of states affected by a tornado?
9. Is it possible to make a prediction on the length of the tornadoes that would occur in future date?

# **Data Processing**

The dataset available in the Kaggle site for Tornadoes is not a clean data that can be directly used for analysis and for solving the different problems posted in the section 2 of this document. We had to apply a lot of pre-processing methods to make the data suitable for the analysis and for solving the problems and we have provided the list below

1. Calculated distance travelled by the tornado from start and end latitude-longitude pairs in the data set using distHaversine() function in geosphere package of R.

distHaversine give the shortest distance between latitude-longitude pairs considering spherical feature of earth and ignoring ellipsoidal feature.

1. Calculated average length and width by month for years 1950-2016 using aggregate () function in R.
2. Computed yearly average of length and width, and filled it for the months for which data were missing.
3. Property loss values follow different format over years. Manipulated it to be of the similar format without altering information for ease of analysis.

Prior to 1996, property loss data was categorical ranging from 1 to 9 (1 : <50$ , 2: 50-500$, 3: 500-5000$,…..9:>5000,000,000$.

1. Grouped the States into five geographic Regions, MidWest, NorthEast, SouthEast, SouthWest and West for ease of data analysis.
2. Categorical data in numbers replaced by corresponding category labels, eg: Property loss categories (1950-1995), magnitude scale(0-5), etc.

# **Methods and Process**

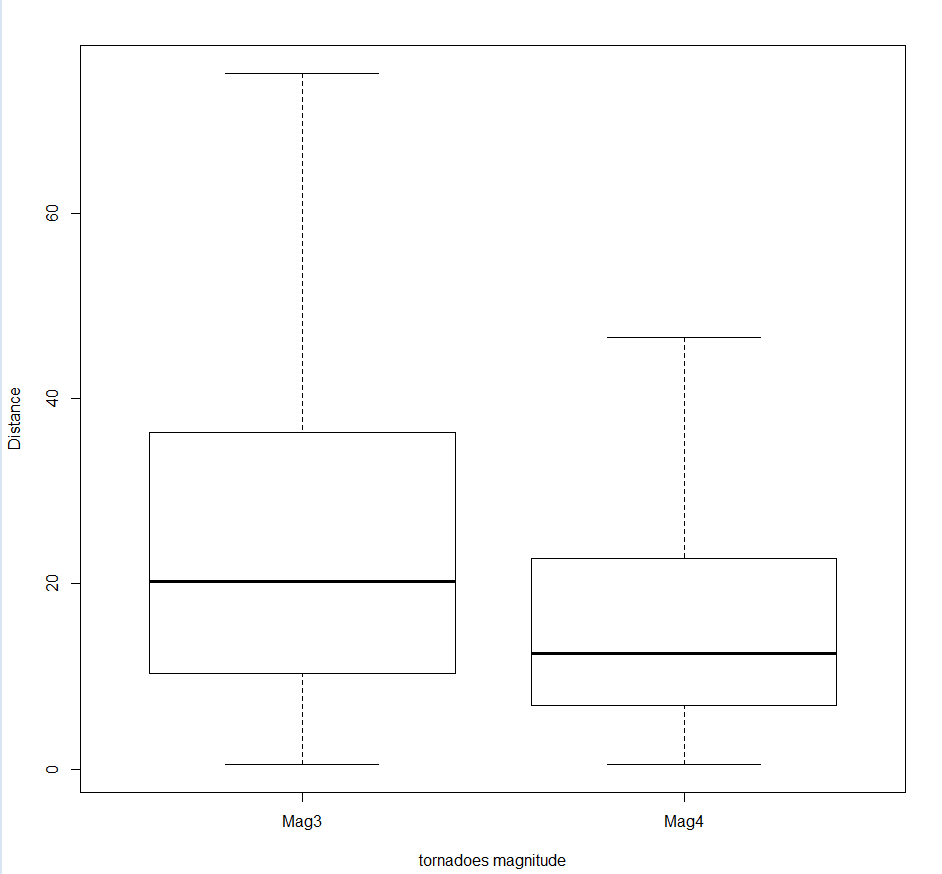
1. **Hypothesis #1:**

***H0: Distance travelled by tornadoes of magnitude 3 and 4 are the same***

***Ha: Distance travelled by tornadoes of magnitude 3 is more than that of 4***

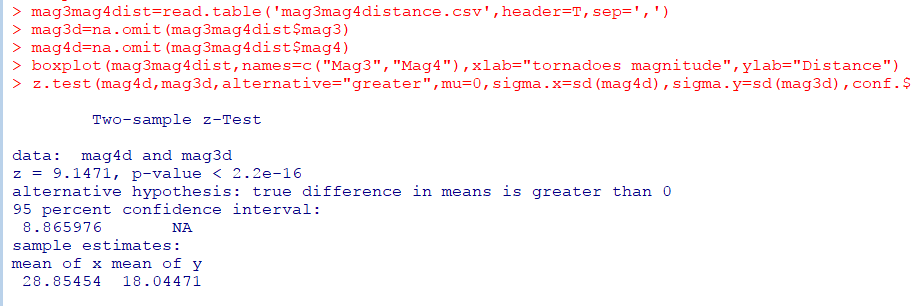
*Data used: mag3mag4distance.csv*

Box plot showing the distribution of the distance data for magnitude 3 and 4



It is clear from the box plot that magnitude 3 has a lengthier distance travelled than mag4. But we use hypothesis testing to confirm this.

Here, the two samples we used are two independent samples and are of size (2202/655) more than 30. Hence, we perform two sample one tailed hypothesis testing



From the screenshot, we could see that the p-value is less than 0.05. Hence at 95% confidence level, we have no enough evidence to accept the null hypothesis. Hence, **at 95% confidence level,** **we conclude that the tornadoes of magnitude 3 travel a longer distance compared to the tornadoes of magnitude 4**.

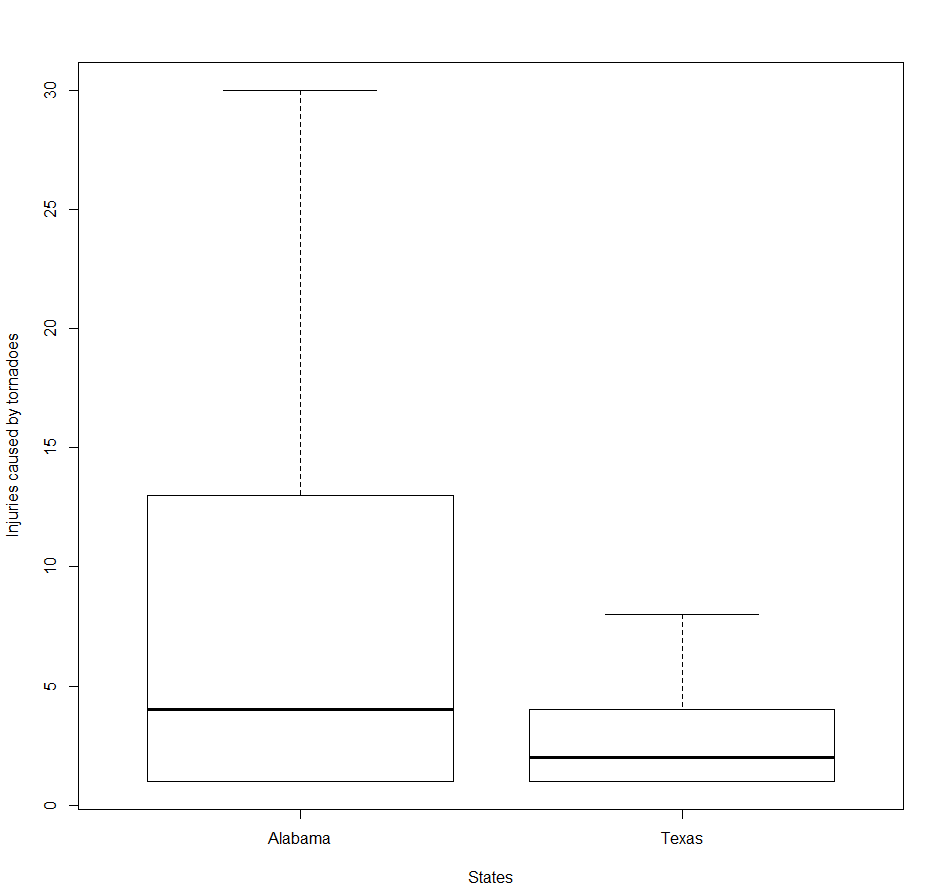
1. **Hypotheis#2:**

***H0: Alabama and Texas have the same number of injuries list caused by tornadoes***

***Ha: Alabama and Texas injuries list is not the same***

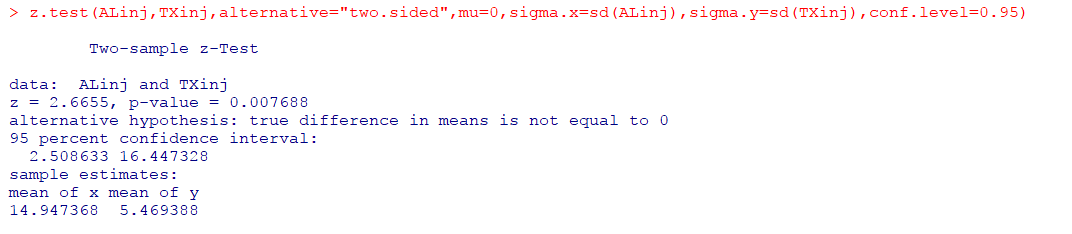
*Data used: ALTXinjuries.csv*

Box plot showing the distribution of the injury data of Alabama and Texas



It is clear from the box plot that state Alabama has a higher average of injury than that of state Texas. But we use hypothesis testing to confirm this.

Here, the two samples we used are two independent samples and are of size (58,50) more than 30. Hence, we perform two sample two tailed hypothesis testing



From the above z test screenshot, we could see that p-value is less than 0.05. Hence at 95% confidence level, we do not have enough evidence to accept the null hypothesis. **Hence at 95% confidence level, we conclude that the number of people injured by tornadoes in states Texas and Alabama are not the same and that they are different.**

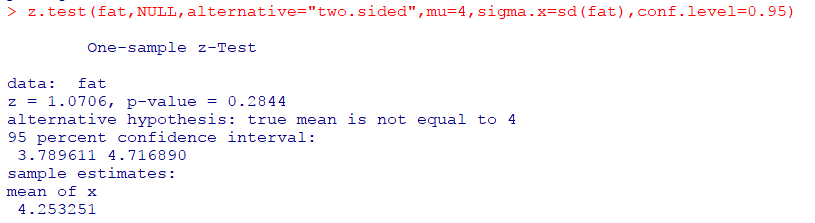
1. **Hypothesis#3:**

***H0: Average number of fatalities caused by tornadoes is 4***

***Ha: Average number of fatalities caused by tornadoes is not 4***

*Data used: fatalities.csv*

Since the sample size of fatalities data (1616) > 30, we use z-test to perform one sample two tailed hypothesis testing to confirm if our assumption is right or not



From the above z test screenshot, we could see that the p-value > 0.05. Hence at 95% confidence level, we have enough evidence to accept the null hypothesis. Hence, **at 95% confidence level, we conclude that the average number of fatalities caused by tornadoes is 4.**

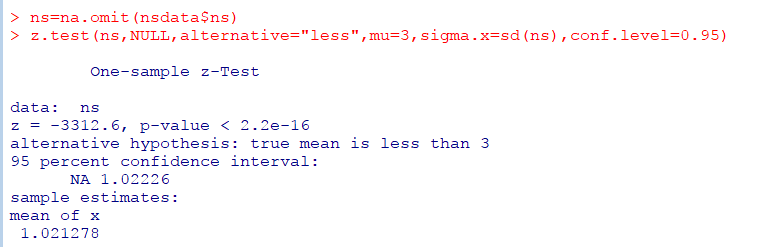
1. **Hypothesis#4:**

**H0: No of states affected by tornado of magnitudes 0 to 3 is 3**

**Ha: No of states affected by tornado of magnitude less than 3 is less than 3**

*Data used: noofstates.csv*

Since the sample size of no of states data (61380) > 30, we use z-test to perform one sample one tailed hypothesis testing to confirm if our assumption is right or not.



Here, the z-stat falls within the rejection region and the p-value is also less than 0.05. Hence, we do not have enough evidence to accept the null hypothesis at 95% confidence level which in turn favors the alternate hypothesis. Hence, **at 95% confidence level, it can be concluded that the no of states affected by tornado of magnitudes 0 to 3 are less than 3.**

1. **Comparison of property loss incurred by different regions – ANOVA**

Test data set*: ploss19962016.csv*

Sample size: 12406 records

Attributes: Region, property loss in dollars

Response variable: Property loss

Region groups: Midwest, NorthEast, SouthEast, SouthWest and West

Assumptions:

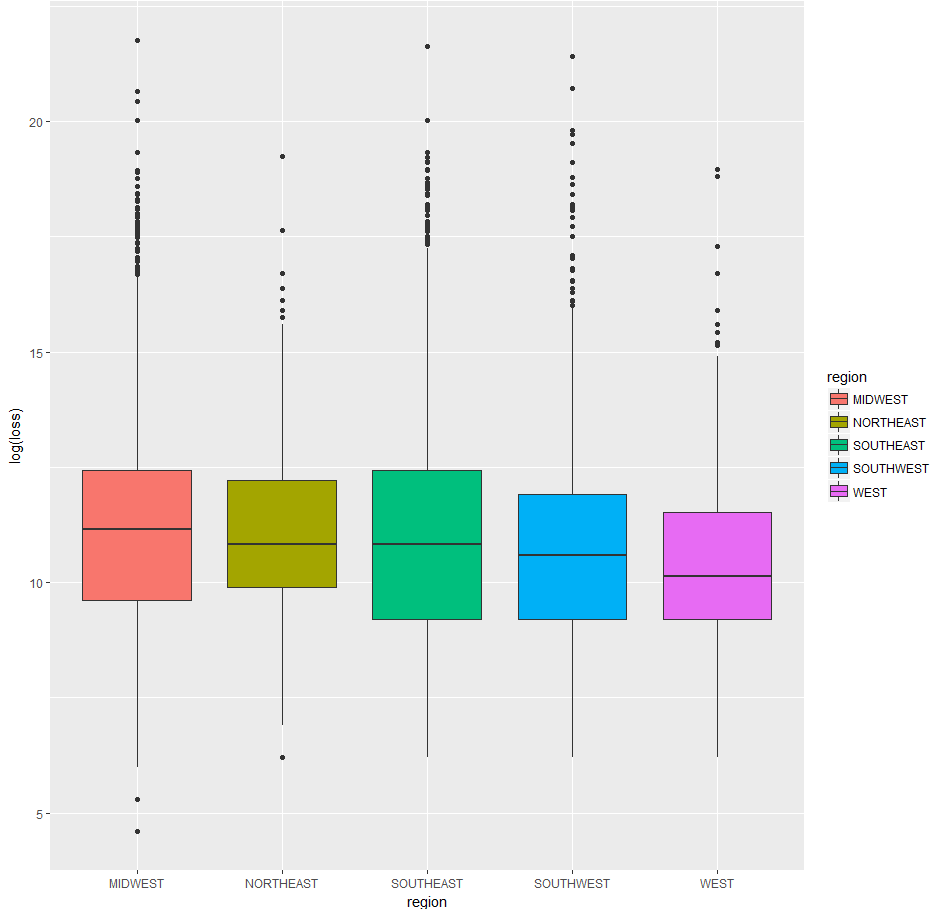
1. samples of losses incurred for each region is independent
2. populations from which the samples are taken is normal with unknown average loss
3. populations have the same standard deviation.

The ANOVA F-test tests the hypotheses:

*H0: Average property loss for all the regions are the same.*

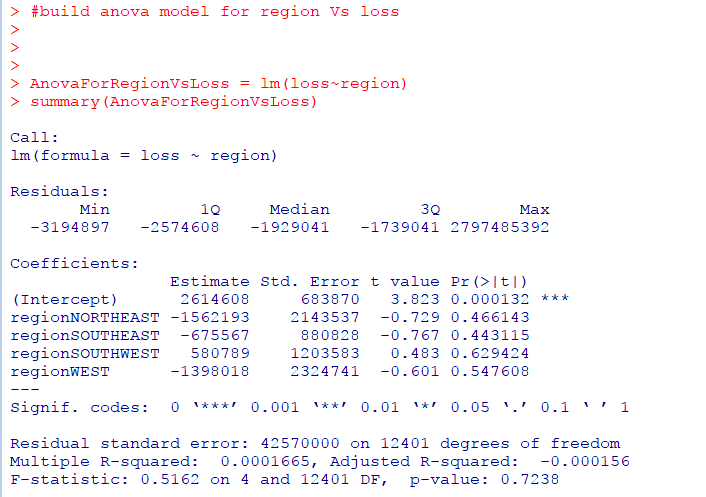
*Ha: Not all the averages are equal*

Box Plot:



The group means appears to be almost similar, with slight difference in in-group variation, from the analysis of boxplot.

ANOVA model:



F-statistic is 0.5 with p=0.72 (>0.05). This indicates our null hypothesis is true, that the average loss incurred by different regions are the same.

T-test on individual parameters also suggest the difference in averages are not significant for all the regions as all p-values > 0.05.

The model needs to be evaluated for model assumptions by residual analysis before concluding (Section 6.1)

1. **Comparison of injuries due to tornadoes of different magnitude – ANOVA**

Test data set*: mag injuries.csv*

Sample size: 7719 records

Attributes: Magnitude (0-5 scale), no. of people injured

Response variable: Magnitude

Magnitude groups: mag0, mag1, mag2, mag3, mag4, mag5.

Assumptions:

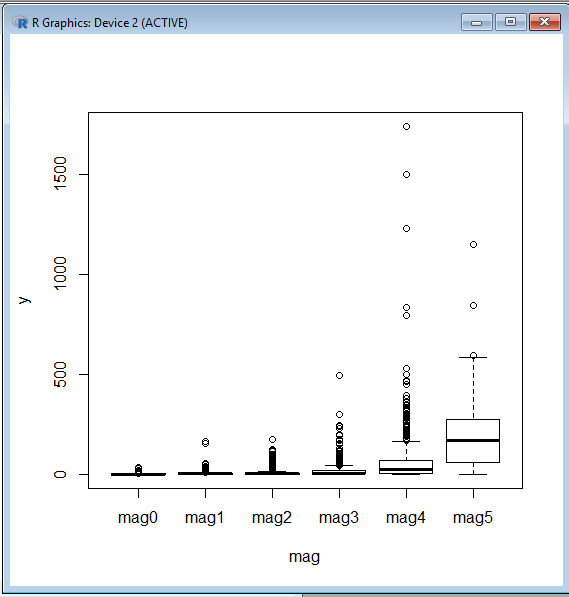
1. samples of injuries due to each tornado category is independent
2. populations from which the samples are taken is normal with unknown average no. of injuries
3. populations have the same standard deviation.

The ANOVA F-test tests the hypotheses:

*H0: Average no. of injuries due to all magnitude categories are the same.*

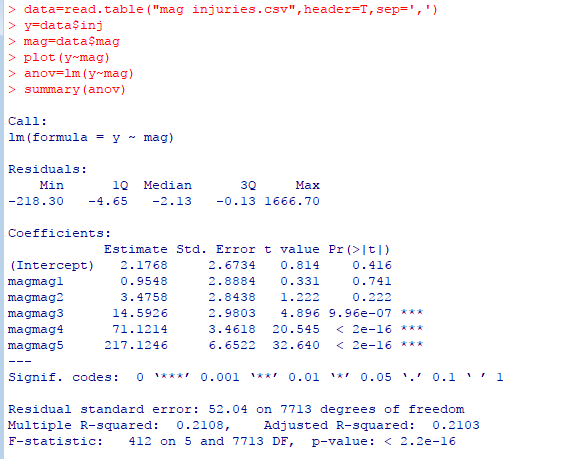
*Ha: Not all the averages are equal*

Box Plot:



The group means appears to be almost similar except for mag4 and mag5, with slight difference in in-group variation, from the analysis of boxplot.

ANOVA model:



F-statistic is 412 and p-value<0.05, hence we can reject null hypothesis.

Also, t-test for individual parameters indicate that the difference in means is significant only for mag3, mag4 ang mag5 with respect to mag0 as p-values<0.05. The difference in means is insignificant for mag1 and mag2 with respect to mag0 as p-values>0.05.

The model needs to be evaluated for model assumptions by residual analysis before concluding (Section 6.1)

1. **Group Magnitude in terms of Length, width and distance travelled by a tornado – KNN Classification**

Test data set: *mag categ prediction.csv*

Sample size: 11797

Attributes: length, width, distance, magnitude

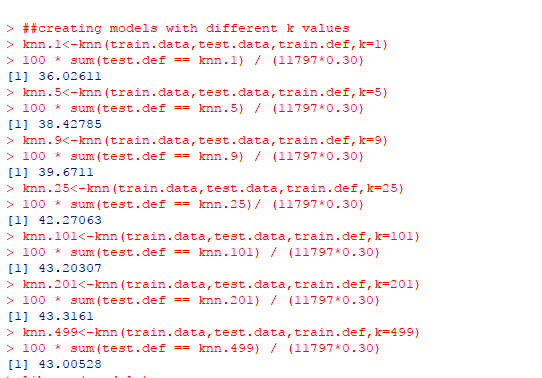
Label: Magnitude

Features: Length, width, distance

Data split: 70/30 split by Hold out Evaluation as the sample size is large. (note: Attempted 80/20, 75/25 splits, with different k-values, but result was better with 70/30. Hence, documenting for 70/30 split alone)

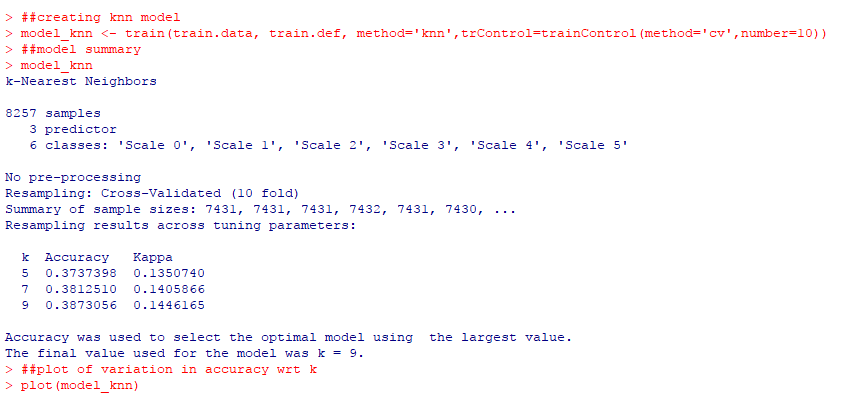
Data preparation: Convert dependent variable magnitude to factor. Normalize numeric variables length, width and distance.

Method: Build K-Nearest Neighbour model for k=1, 5,9,25,101,201,499 and calculate the accuracy.



The accuracy levels are less for almost all models with k=201 being the better model.

We also tried model by training data by knn and resampling by cross validation.



The method took k=9 as the best model, but accuracy is only 0.38

Model evaluation for accuracy needs to be carried out using crosstable/confusion matrix (section 6.1)

1. **Group Property Loss category in terms of Magnitude category of tornado, region of occurrence and no. of states affected – Naïve Bayes Classification**

Test data set: *property loss prediction.csv*

Sample size: 9705

Attributes: Property loss, magnitude, region, no. of states

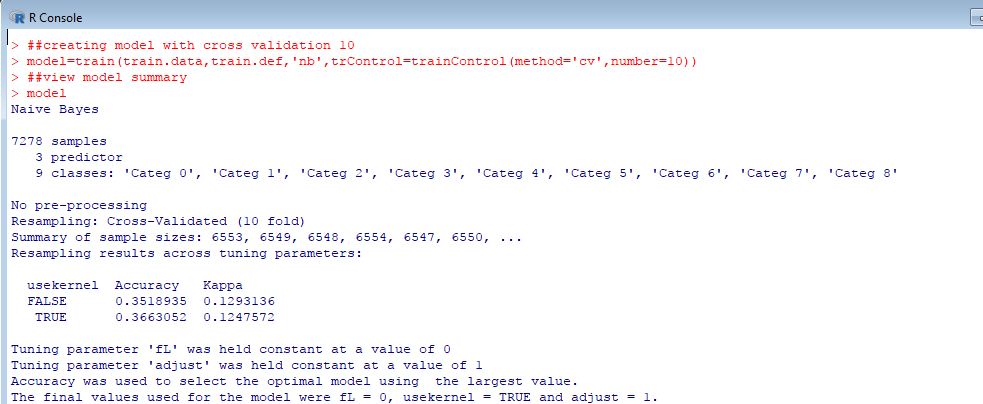
Label: Property loss

Features: magnitude, region, no. of states

Data split: 10-folds cross validation as the sample size is less

Data preparation: Convert dependent variable magnitude to factor. Normalize numeric variables length, width and distance.

Method: We use Naïve Bayes since the features are nominal.



The accuracy of the model obtained is less 0.35.

We need to further evaluate the model using predictions to conclude (Sec 6.1)

1. **Time series analysis on Tornadoes Length data:**

Tornadoes occurred over the years have had different lengths throughout. Time series analysis is an attempt to learn the distribution of length of tornadoes over the years. Since the daily data collected were not continuous, we took the average length data per month and used the new monthly data for the time series analysis. There were data missing for 3 months in three different years. So, we arrived at the average length of the year and filled this average for the respective months which missed the length data

*Data used: torn-length-monthly-data.csv*

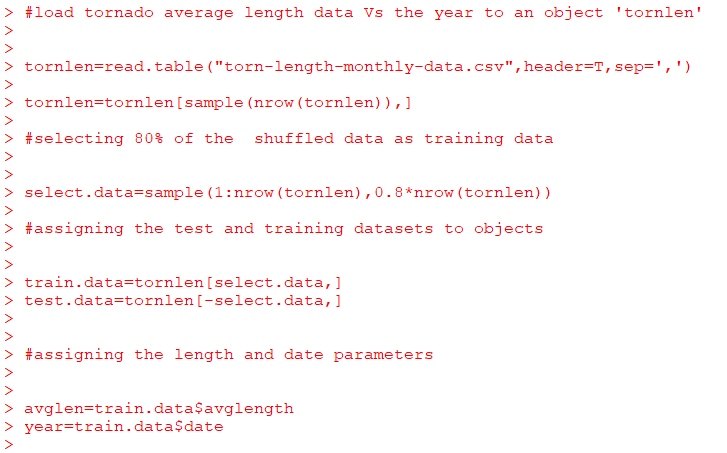
Data spans from 1950 to 2016 covering 67 years and 67X12 = 804 months.

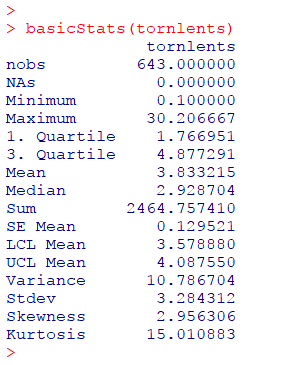
Process followed for Time Series:

* + Loading time series data to an object
  + Split the data into training and testing dataset
  + Creation of Time Series Object for the training dataset
  + Preliminary Data Analysis to test assumptions
  + Decide if transformation or differencing is required
  + Identify p and q values for building time series models – AR/MA/ARMA
  + Build models
  + Perform residual analysis of every model
  + Evaluate the models based on RMSE
  + Choose the best time series model
  + Plot the predictions and forecasts graph with predicted values

Every step followed is as detailed below with appropriate screenshots

**Loading time series data and split-up**



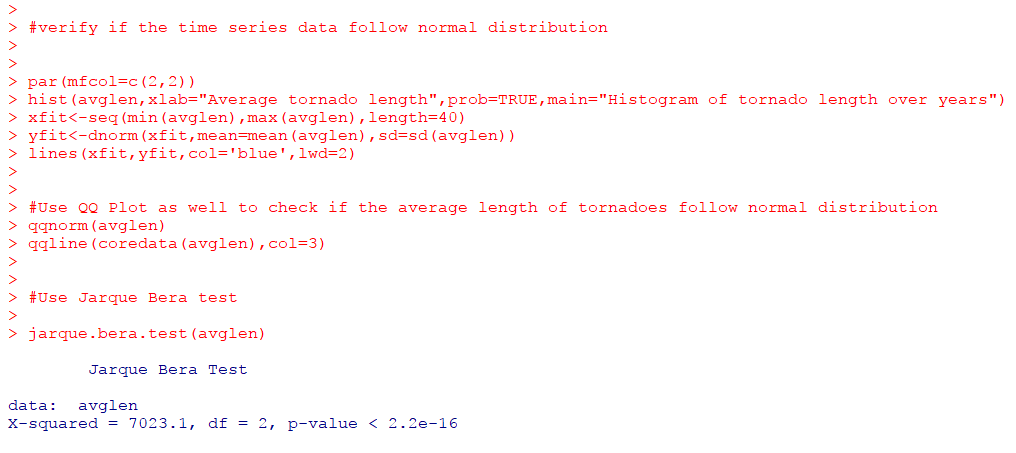


As you can see here, we have split the data into test and training and created the time series object

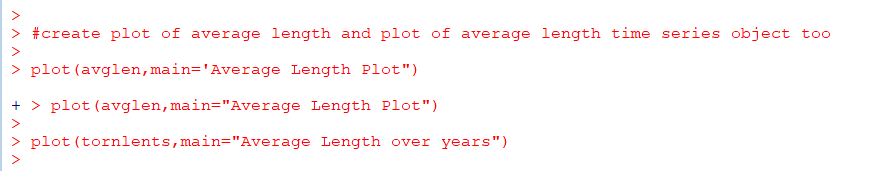
**Preliminary data analysis to test assumptions of time series:**

Normal Distribution: Most important assumption for time series which the time series data is expected to meet.

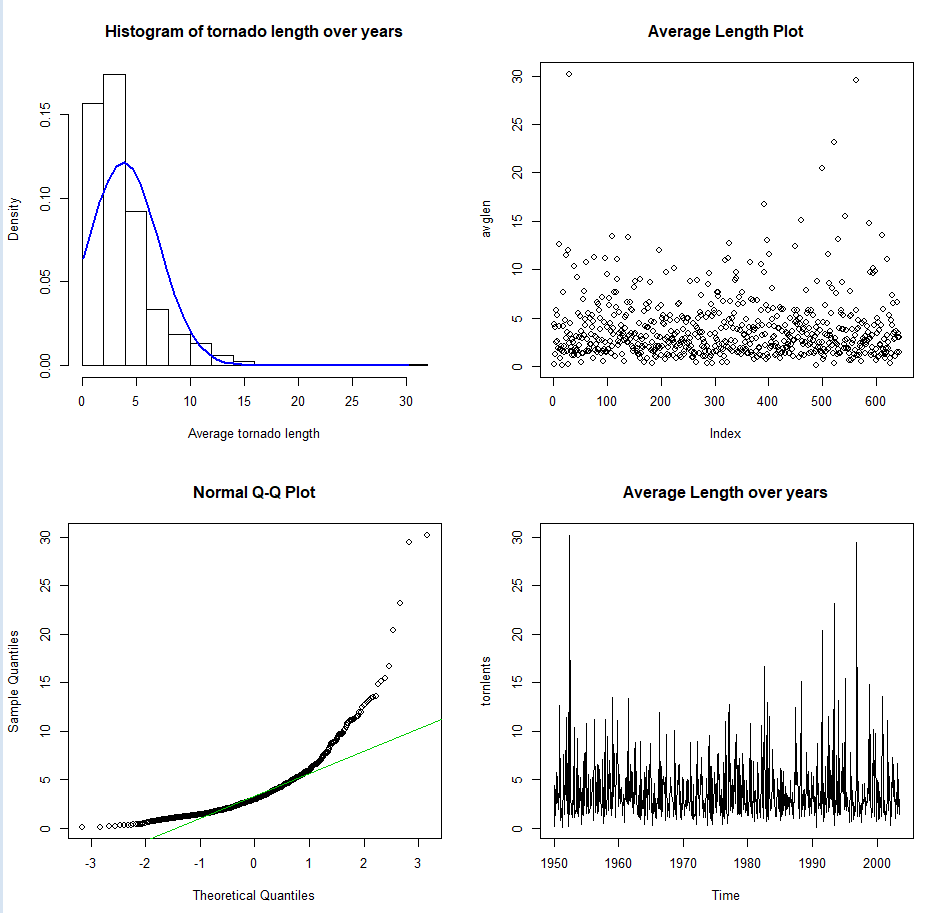
Tornadoes length data meets this requirement?

To test this, we plotted the histogram, QQ plot and Jarque Bera test as shown below 

We created the time and time series plots too



Resulting graphs and plots of Normality tests and time plots

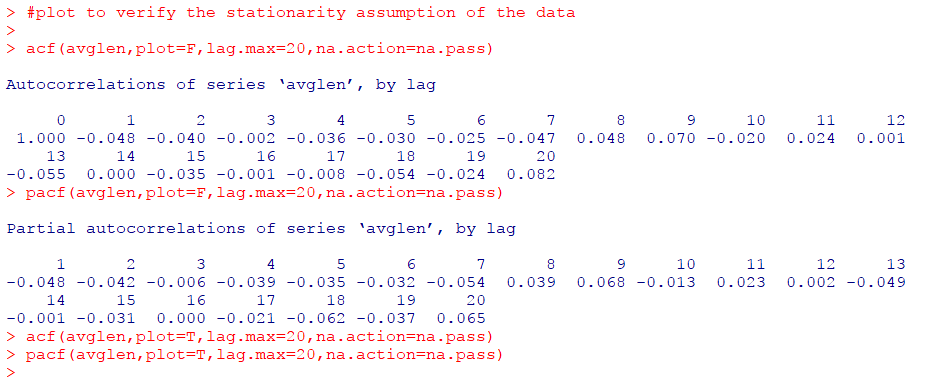


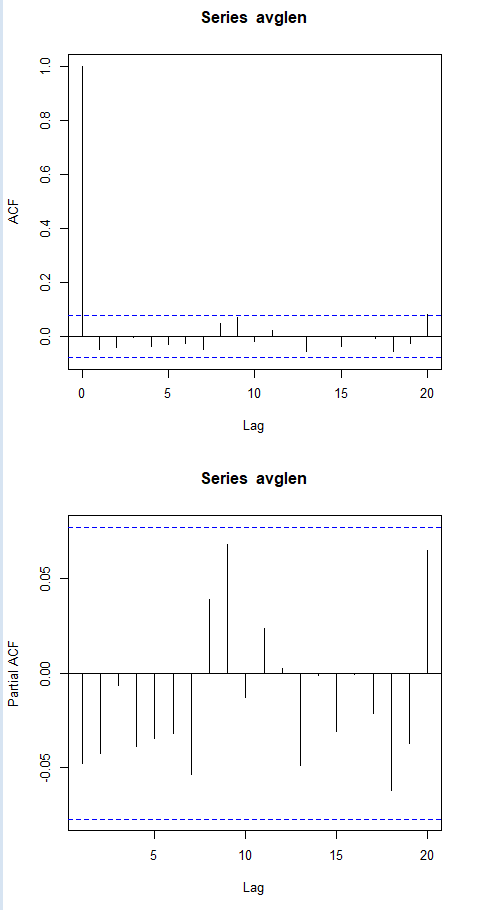
It is clear from the above histogram, QQ Plot and Jarque Bera Test that the tornadoes length data **doesn’t follow normal distribution**

**Stationarity:**

Stationarity refers to the constant variance and constant mean of the tornadoes average length throughout the years which is one of the critical assumptions that the time series data is expected to meet.

Let’s see if our data meets this stationarity requirement. We plot the ACF plot to verify this.





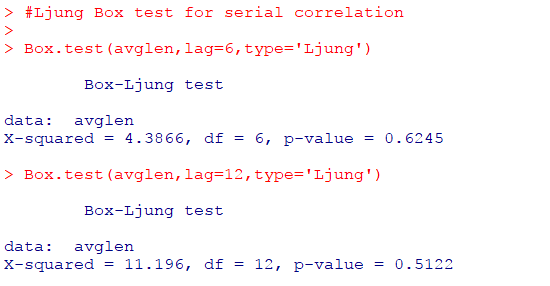
There is no single non-zero auto-correlation value present and the acf values decay slowly in the ACF plot and hence this confirms that the tornadoes length data is not stationary

**Serial Correlation:**

In time series, the data should be correlated to the past data and only then the prediction and model building can be done. This aspect of correlation between the current and past data is referred to as serial correlation.

Tornadoes length data meets this requirement?

To answer this, we need to perform the Ljung Box Test on the tornadoes length data as shown below



The hypothesis statements considered in this Ljung Box test are as follows:

*H0: Time series data has serial correlation and are not white noise*

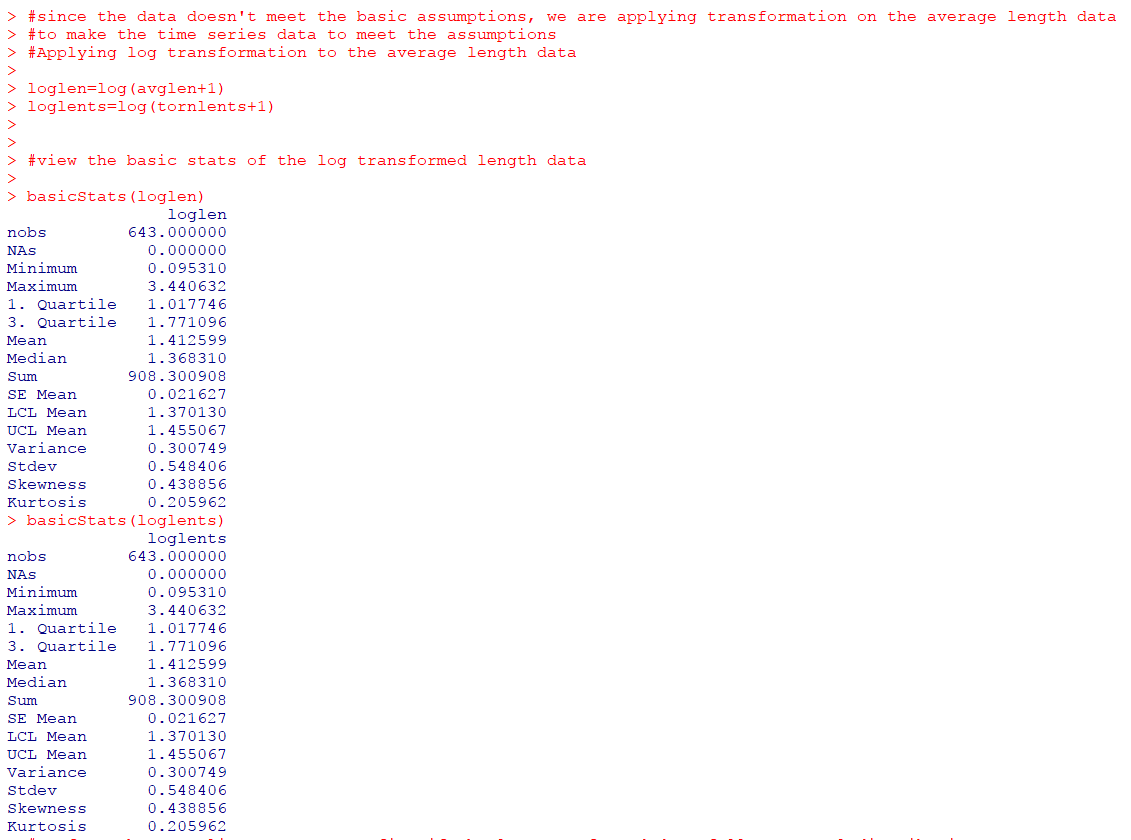
*Ha: Time series data has no serial correlation and are white noise*

From the above test result, we could see that the p-value of this test at both lags 6 & 12 are more than 0.05 favoring the null hypothesis. Hence, this indicates **that the tornadoes length data is not serial correlated**

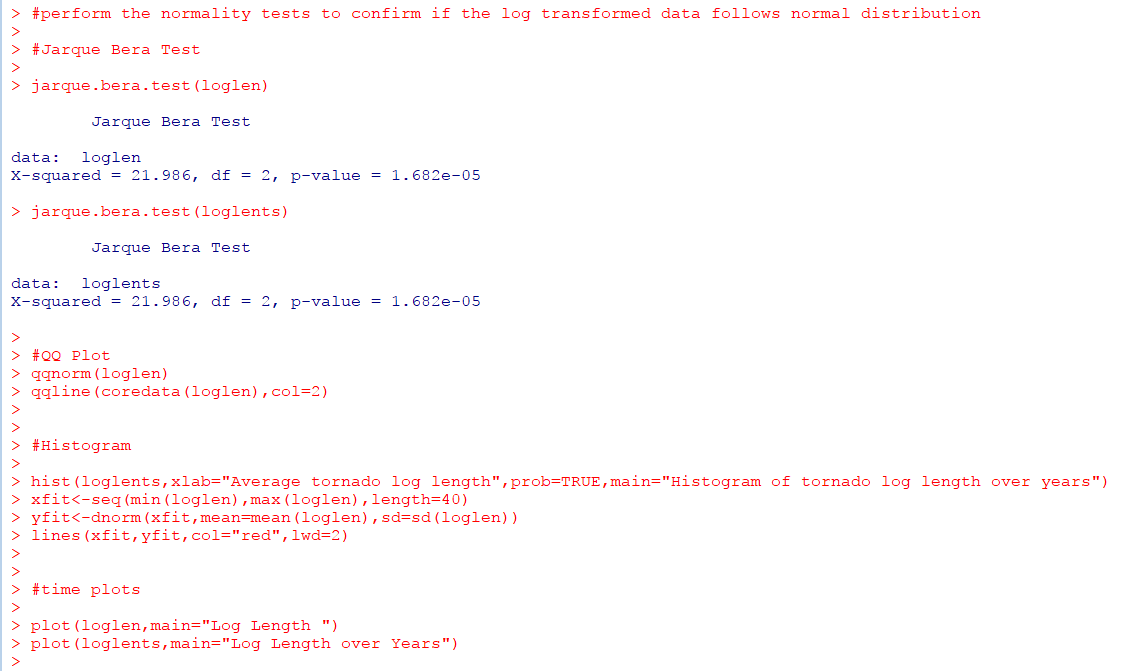
**Decide if transformation or differencing is required**

It is obvious from the above screenshots that the data doesn’t meet the requirements of time series data. So, we applied log transformation on the tornado length data since log transformation normalizes the data. We need to perform the preliminary analysis on this log transformed data as shown below.

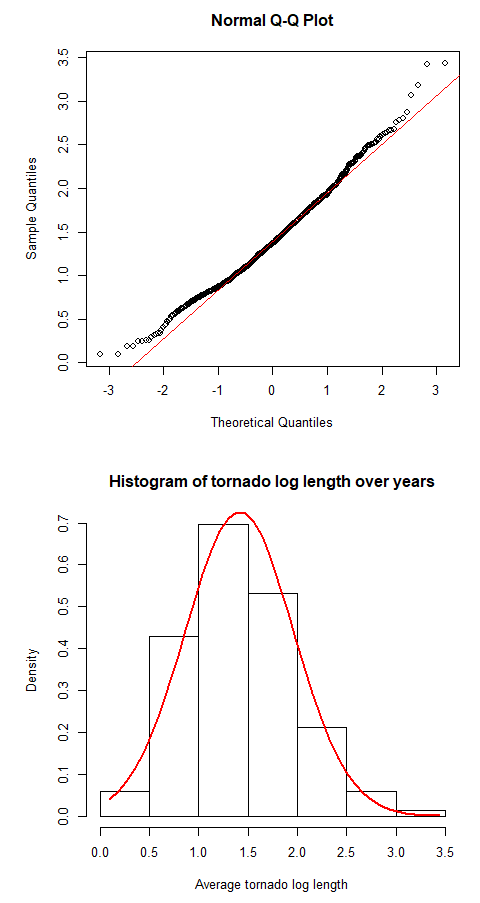
**Log transformed data:**



**Normality tests and time plots on log transformed data:**



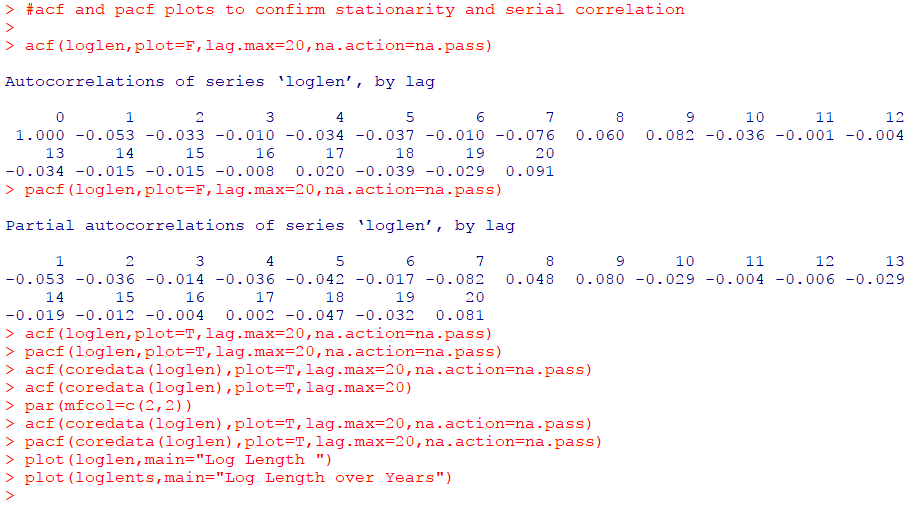
Log transformed data normality test results:

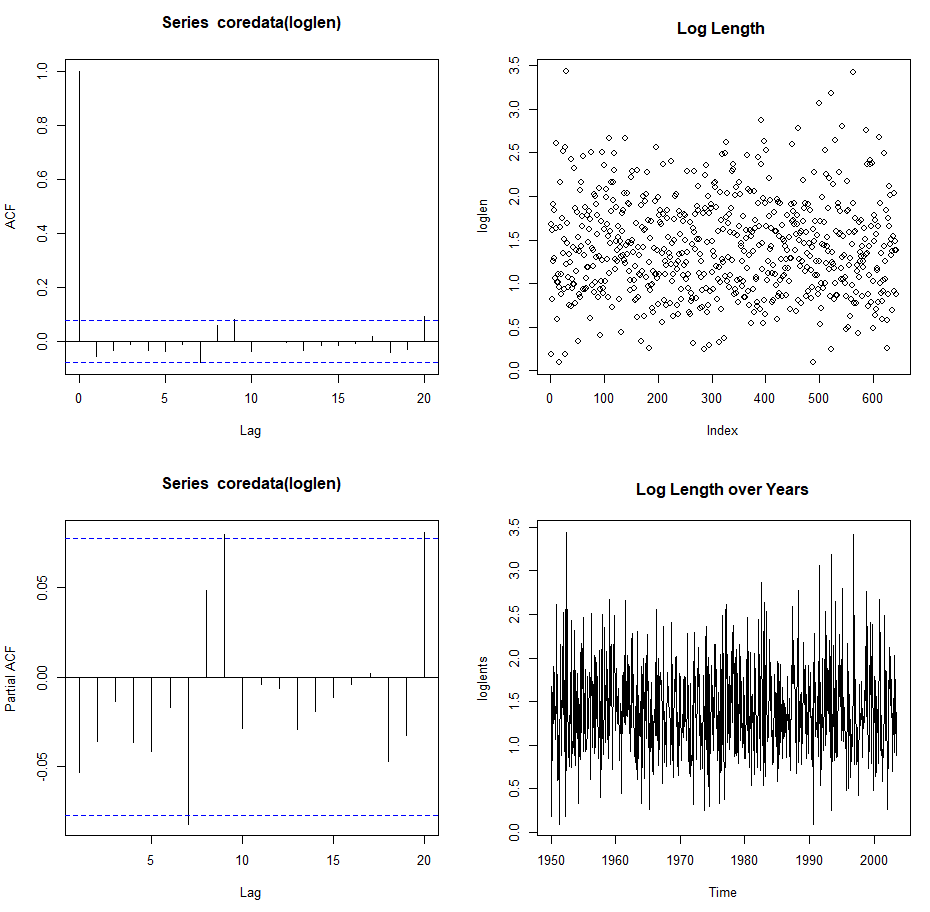


**It is now clear from the above screenshots that the log transformed data follows normal distribution. The skewness and kurtosis values in the Basic Stats also prove this.**

**Stationarity tests:**

We plotted the ACF and PACF plots to confirm the stationarity as given below



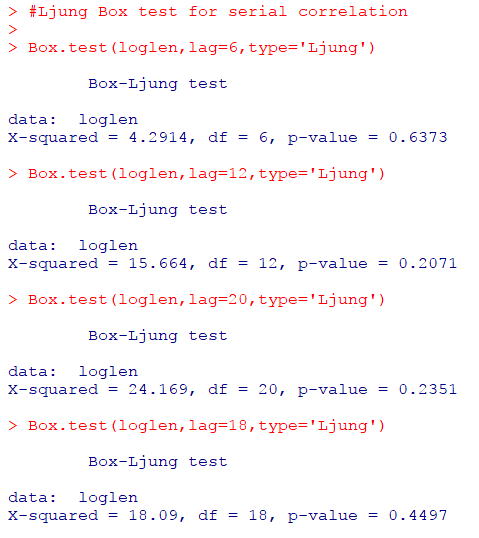


There is one non-zero autocorrelation value present in the ACF plot indicating serial correlation but it is not stationary.

We need to confirm this with the Ljung Box test as well

**Serial correlation:**

We performed Ljung Box test on the log transformed data to test serial correlation as shown below



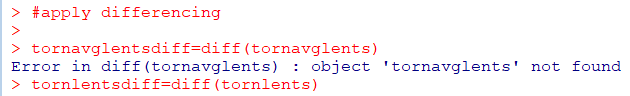
As stated earlier, the null hypothesis in Ljung Box test is ‘there is no serial correlation and the time series data is white noise’. Here again, the p-value is > 0.05 favoring the null hypothesis and hence the log transformed data is not serially correlated.

Hence we can conclude that , though **the log transformed data is normally distributed, it is not stationary and serial correlated and cannot be used for time series analysis**.

We no need to apply differencing on the length data to see if they make the length data meet the time series assumptions.

**First Differencing:**

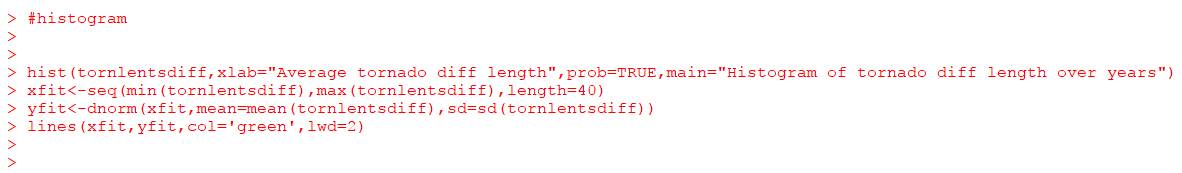
Differencing is applied on the time series object of tornadoes length data as shown below.

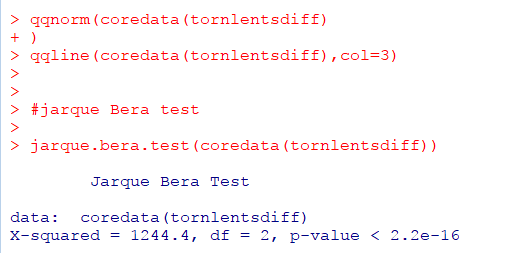


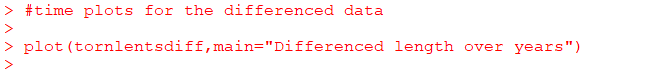
The preliminary analysis process should be now followed for this differenced time series object as well.

**Normality tests:**

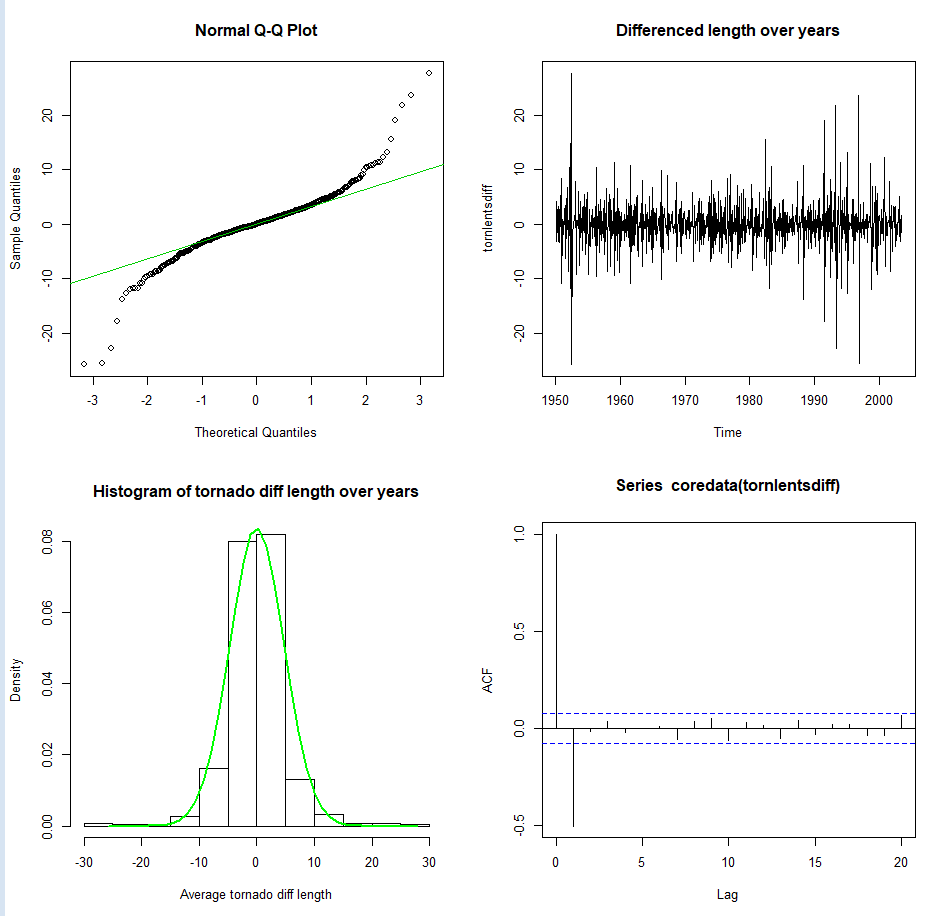
We plotted the histogram, QQ plot and performed Jarque Bera tests on the first differencing applied time series object as shown below:







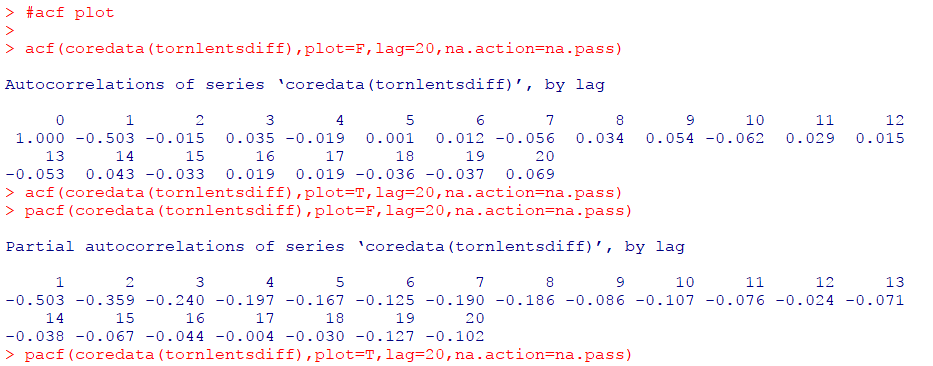
**Normality test results:**

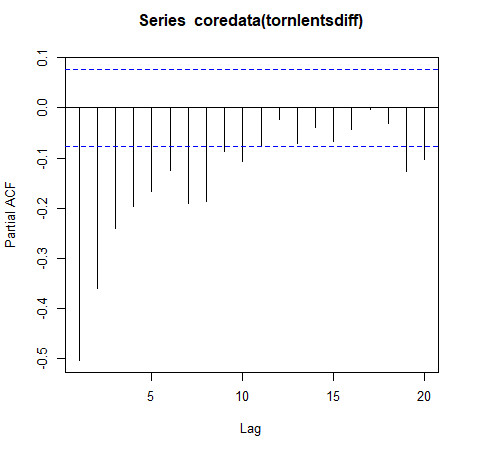


We could see from the normality plots and test results that the **differencing applied length data now follows normal distribution meeting the first assumption required for time series analysis**

**Stationarity:**

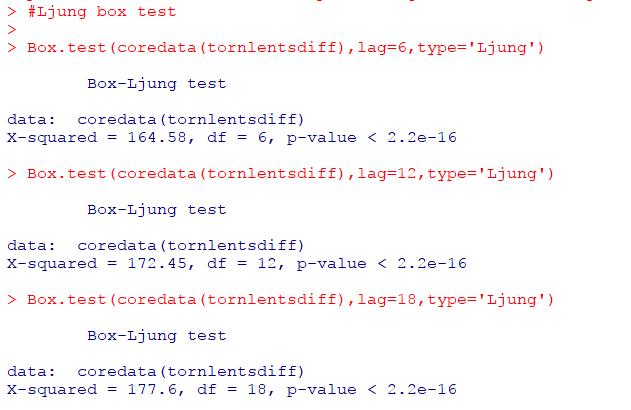
**The ACF plot given above shows that there is quick decay of the acf values and there is atleast one non-zero auto-correlation value confirming that the differencing applied length data meets the stationarity assumption of time series analysis**. The code used and the pacf plot of the differenced data is as given below.





**Serial correlation:**

We performed the Ljung box test on the first differencing applied tornadoes length data as shown below



Here, the p-value is less than 0.05 and hence we reject the null hypothesis ‘no serial correlation exists in the time series data and it is white noise . Hence, we can conclude that the differencing applied time series data is serial correlated.

Hence, **the first differencing applied tornadoes length data now satisfies all the three assumptions – normal distribution, stationarity and serial correlation – of the time series modelling**.

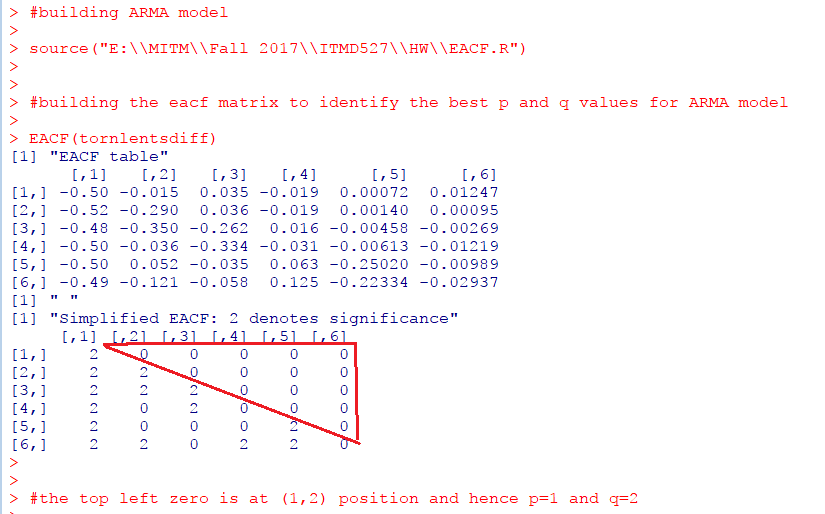
The next step in time series is to identify the p and q values from the PACF and ACF plots respectively to build models

**Identify p and q values for building time series models – AR/MA/ARMA:**

From the ACF plot, we could see that after lag1 the successive lags are all within the zero bounds of the ACF plot. The ACF plot cuts off at q. Hence the **q value is identified to be 1**.

From the PACF plot, we could see that after lag10, the successive lags are all within the zero bounds of the PACF plot. The PACF plot cuts off at lag p. Hence the **p value is identified to be 10.**

We can also identify p and q values for building ARMA models automatically using the EACF plot as shown below:



**Build Models:**

We will now build Auto-Regressive (AR(p)), Moving Average(MA(q)), ARMA(p,q) models with the order values ( p & q) we identified from the ACF,PACF and EACF plots respectively.

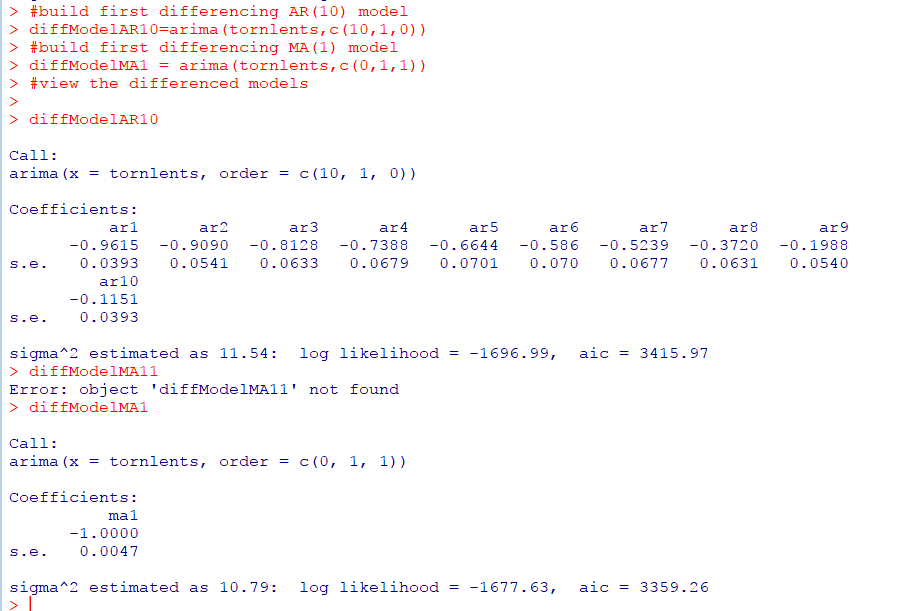
We will build all the models using ARIMA () function available in R.

**Note:** since we cannot use the model built using ARMA () function for prediction purposes, we will build the ARMA model also using the ARIMA () function in R.

Since p=10 from PACF plot and q = 1 from ACF plot, we will be building AR (10) and MA (1) models as shown below:

**AR (10) and MA (1) models:**

Since we have applied differencing, the order of AR (10) model will be (10,1,0) and MA(1) will be (0,1,1) as shown below



Once the models are built, we should analyze the residuals of every model that is built to check if the residuals meet the below requirements:

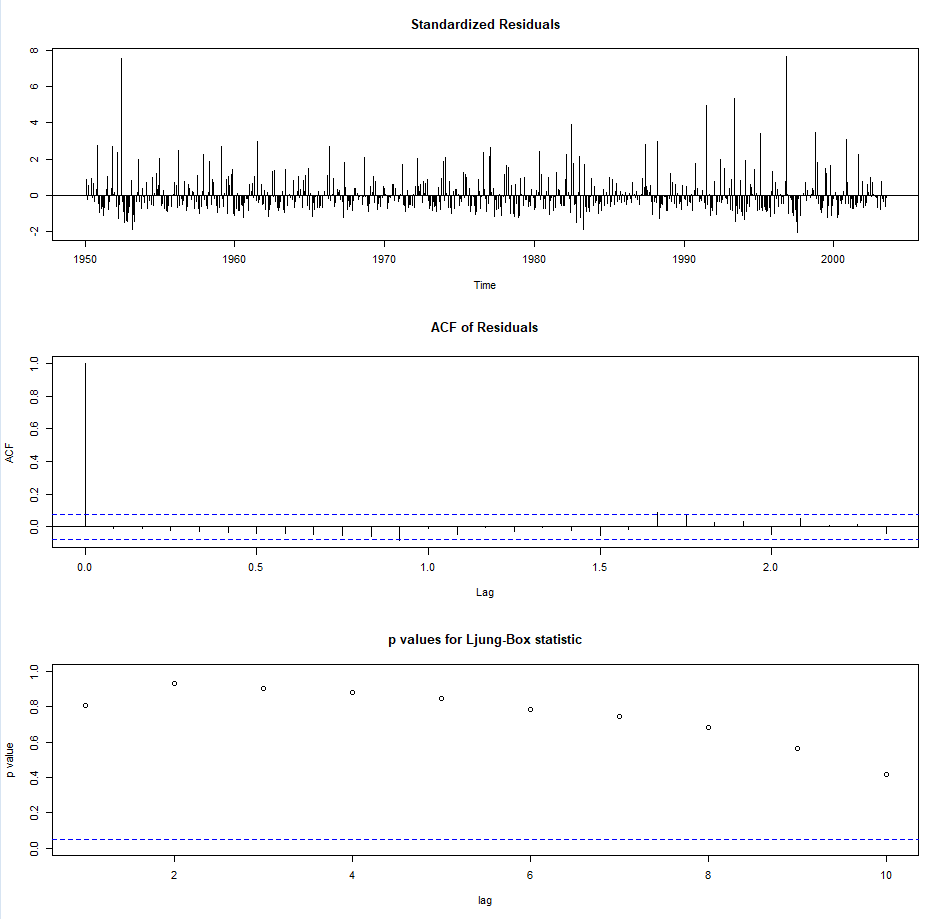
Residuals must follow normal distribution

Residuals must be white noise.

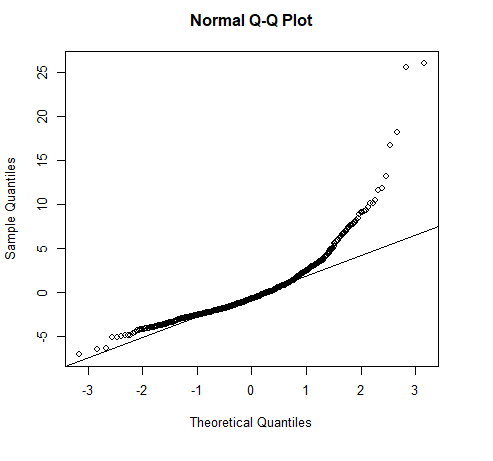
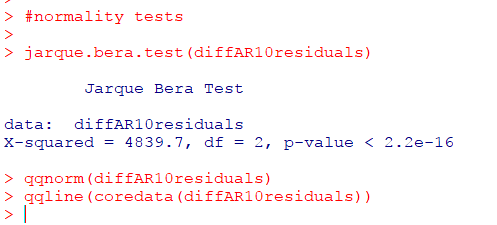
We test all these assumptions of residuals using the residual plots, QQ Plot, jarque Bera test and Ljung Box test.

These are carried out as shown below.

**Residual plots of AR (10) model:**

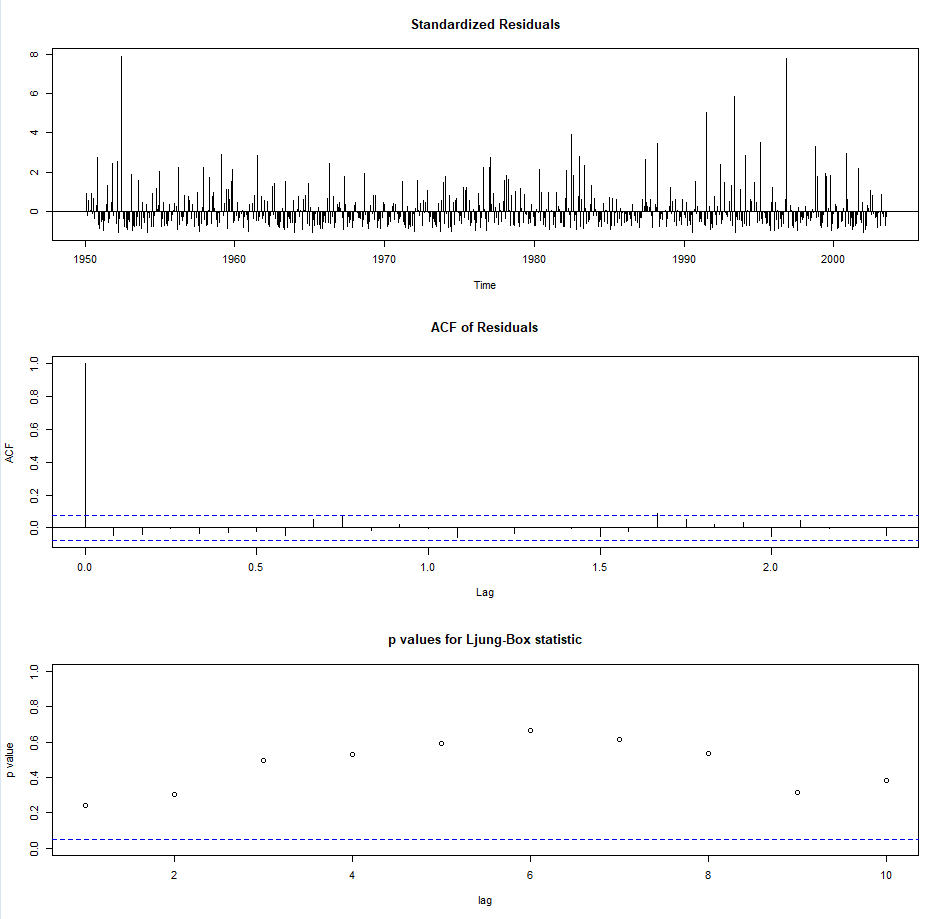


**Normality tests of AR (10) model residuals:**

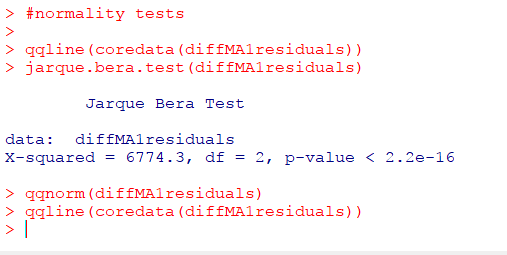


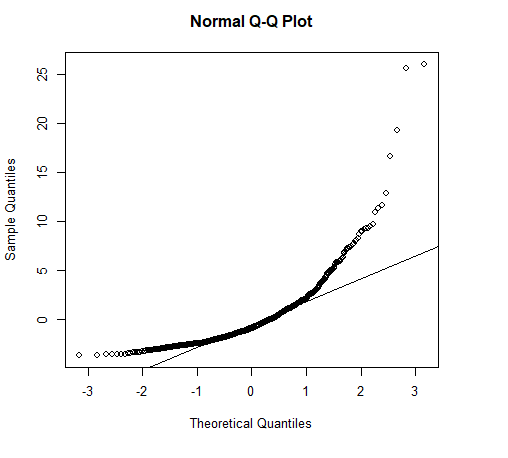
From the above plots and the p-value < 0.05 in the jarque bera test, **we could see that the residuals of AR (10) model follows normal distribution.**

**Residual plots of MA (1) model:**



**Normality tests of MA (1) model**

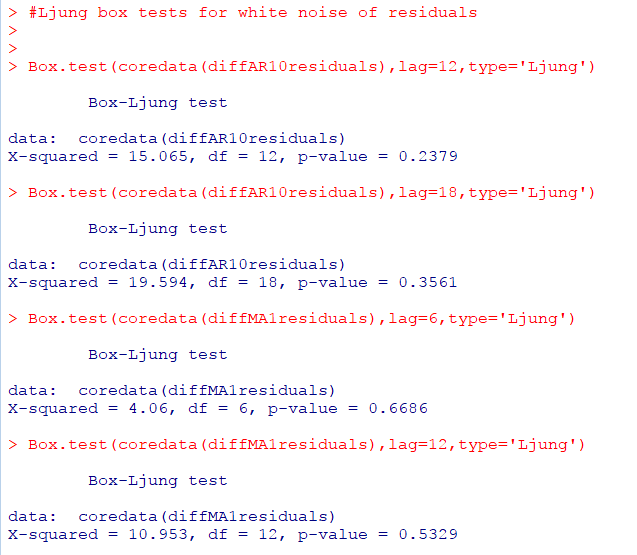




From the above plots and p-value < 0.05 in Jarque Bera test, we could see that **the residuals of MA (1) model also follow normal distribution.**

We will now see if the residuals are white noise or not for these two models.

**Ljung Box test for residuals white noise:**



The hypothesis statements in Ljung Box test for residuals are as given below

*H0: Residuals are not white noise*

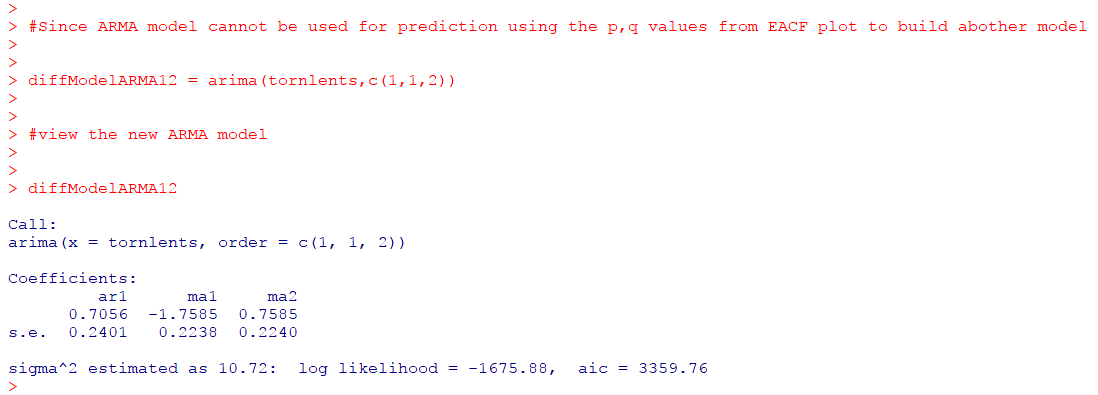
*Ha: Residuals are white noise*

As shown in the above plot, the p-values are more than 0.05 at both lags for both the models.

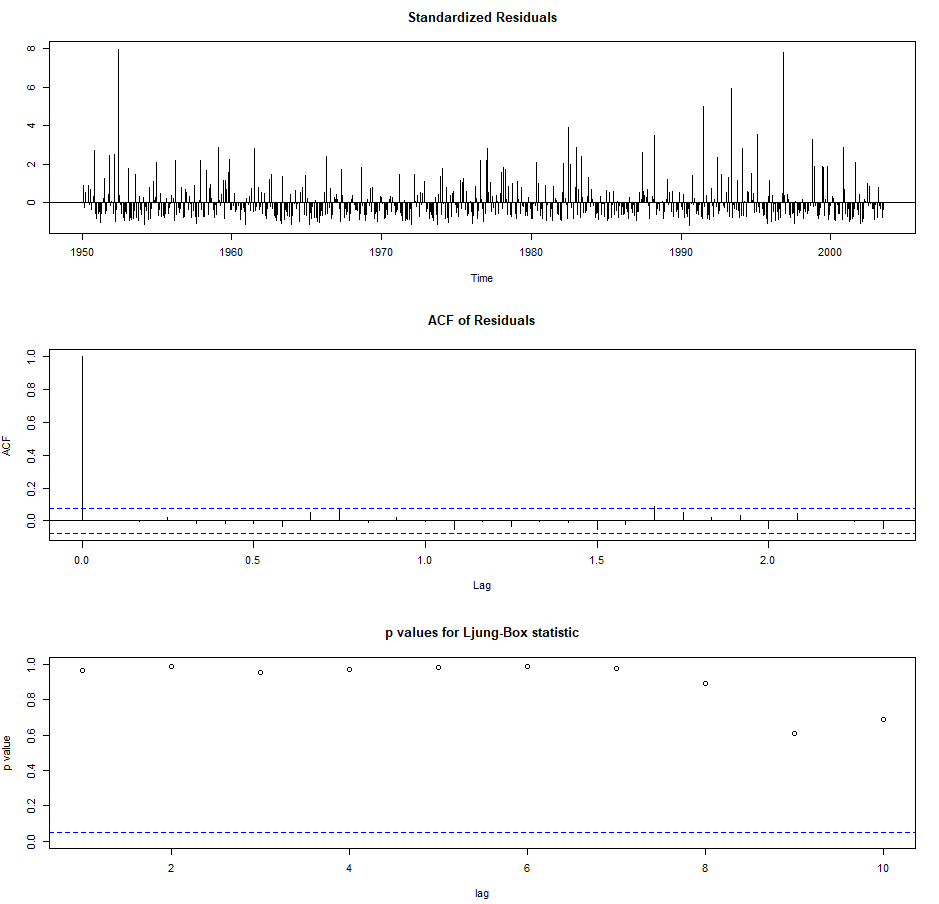
Hence, we do not have enough evidence to accept the above null hypothesis. Hence, we can conclude that **the residuals of both models are white noise**.

**ARMA (1,2) model**

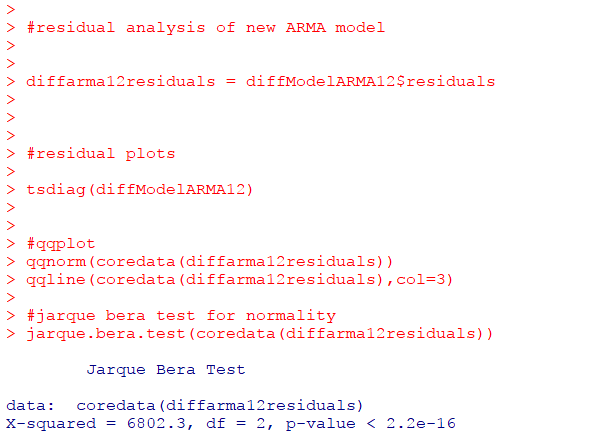
From EACF plot, we got the p and q values to 1 and 2. We will now build ARMA (1,2) model as shown below:

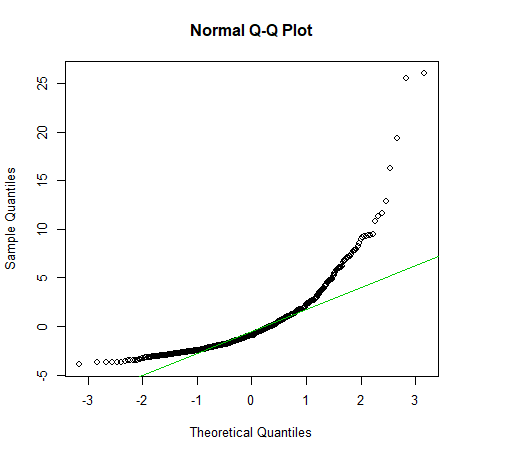


**Residual plots of ARMA (1,2) model**

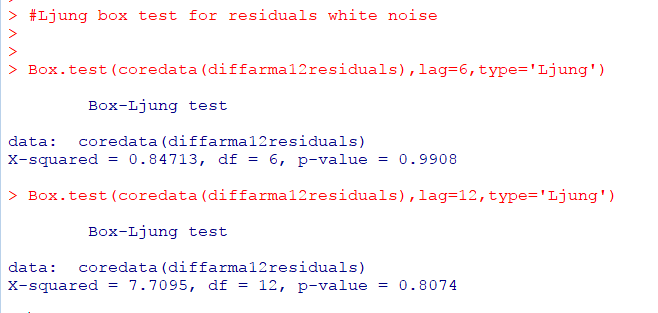


**Normality tests of ARMA (1,2) residuals**





**Ljung Box test for ARMA (1,2) model residuals white noise**



It is clear from the QQ plot that most of the residuals are along the QQ line and the p-value in Jarque Bera test is < 0.05. This indicates that **the residuals of ARMA (1,2) model follow normal distribution.**

Also, from the Box test, the p-values are more than 0.05 at both lags. Hence **the residuals of ARMA (1,2) model are white noise.**

A detailed summary of all the 3 models built with the training data is as given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Criteria | AR (10) | MA (1) | ARMA (1,2) |
| AIC | 3415.97 | 3359.26 | 3359.76 |
| Residuals white noise? | yes | Yes | Yes |
| Residuals normally distributed | Yes | Yes | Yes |
| Qualified Model? | Yes | Yes | Yes |

# **6. Evaluations and Results**

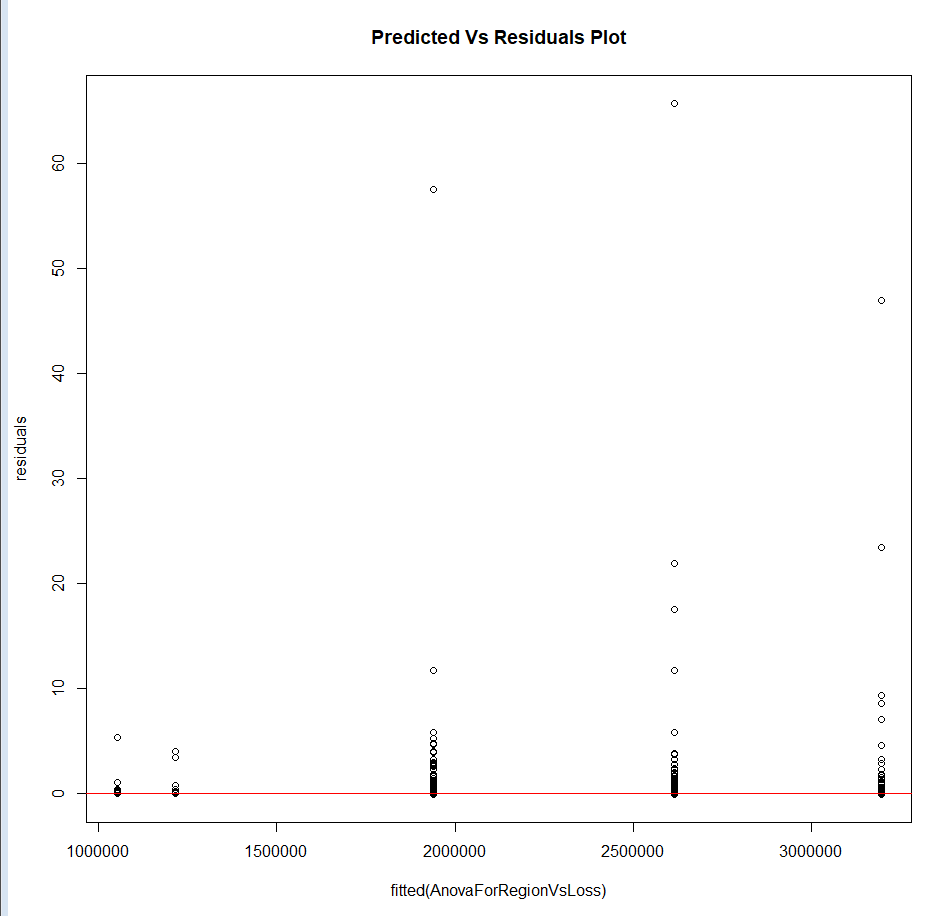
## 6.1. Evaluation Methods

**ANOVA model evaluation (Comparison of property loss incurred by different regions)**

Residual Analysis:

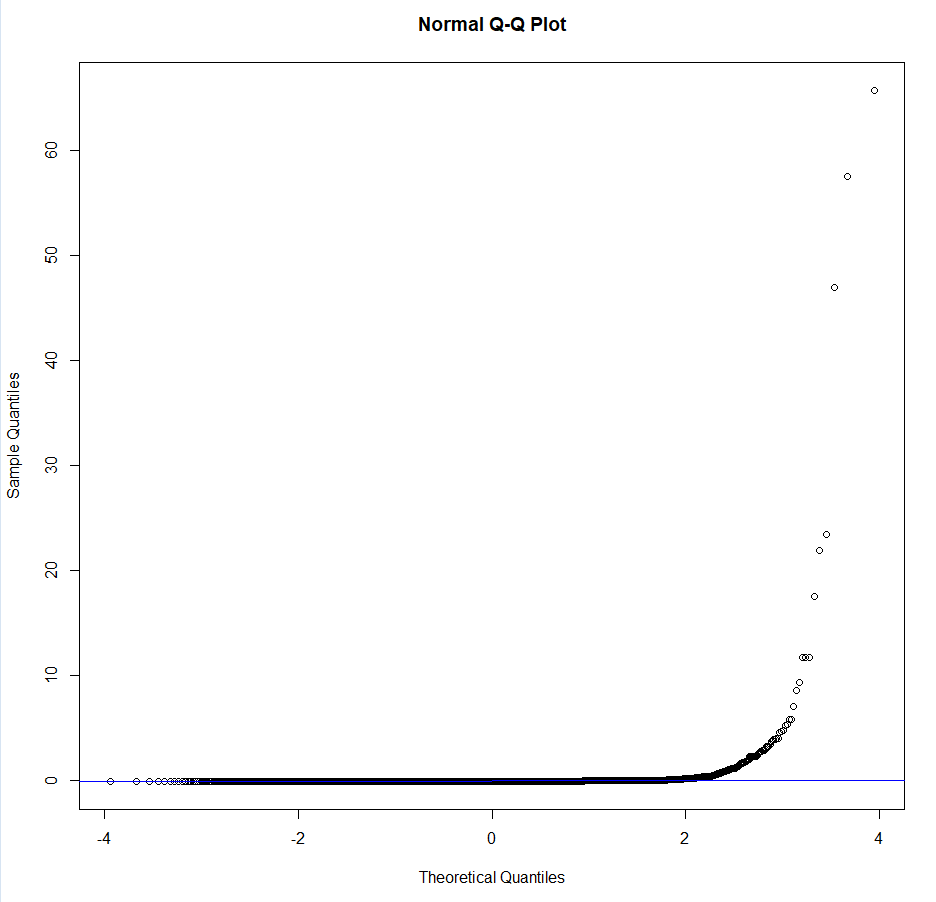
1. Residual plot – is the within group variability constant?
2. Normal Probability plot – does it follow normal distribution?

Predicted vs Residual plot:



Here, the within- group variability is not constant.

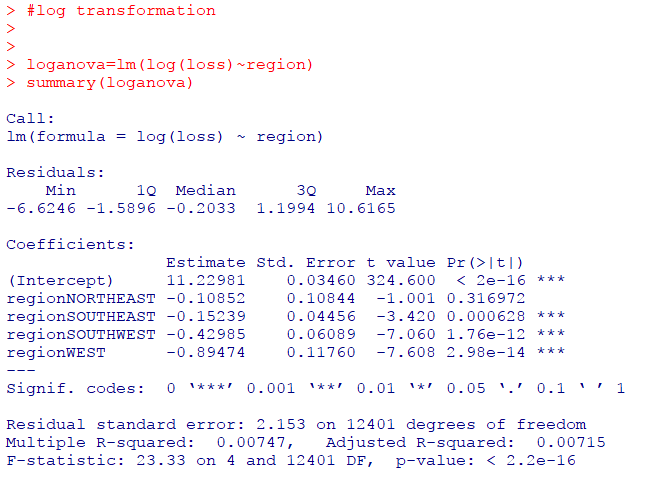
Normal Probability plot:



The residuals do not follow normal distribution.

Hence, our model is not sufficient to conclude on the ANOVA test hypothesis.

Hence, we apply build ANOVA model with log transformed loss variable.

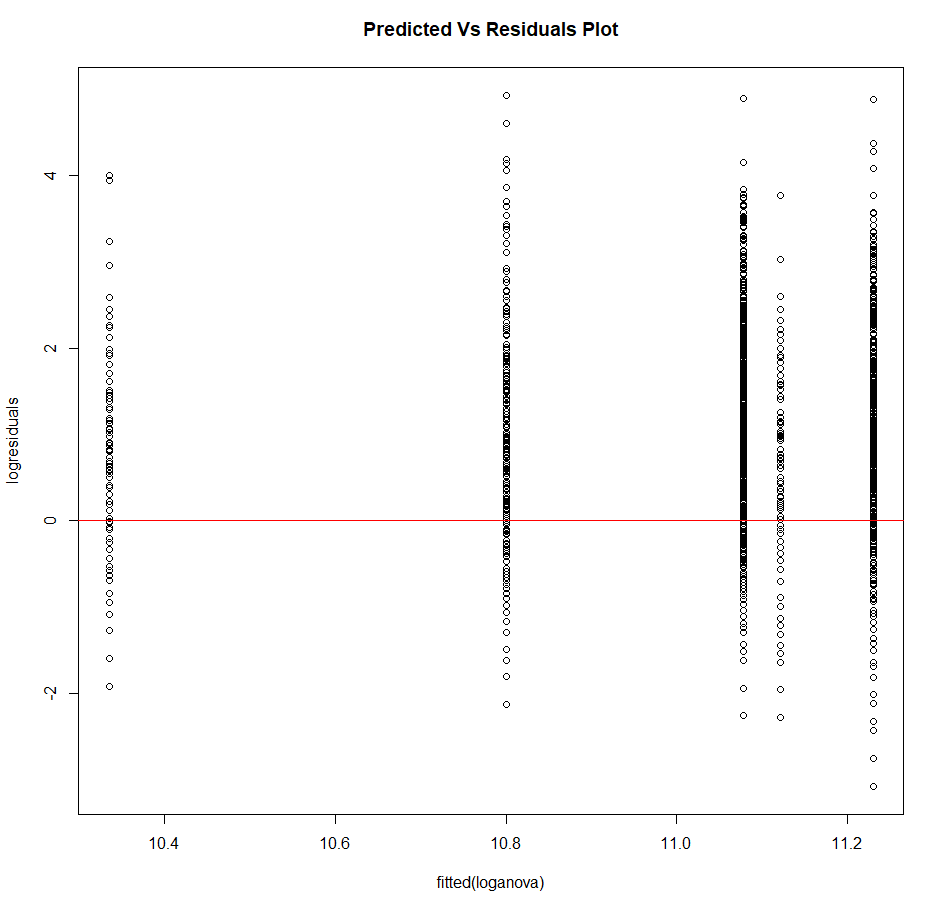


F-statistic is 23.33 (which is greater than the previous model) with p-value < 0.05. This indicates that we can reject null hypothesis at 95% confidence interval.

Also, t-test on individual parameters suggest that the difference in means in not significant for NorthEast region with respect to MidWest as p-value>0.05. But, it is significant for SouthEast, SouthWest and West with respect to MidWest as p-values>0.05.

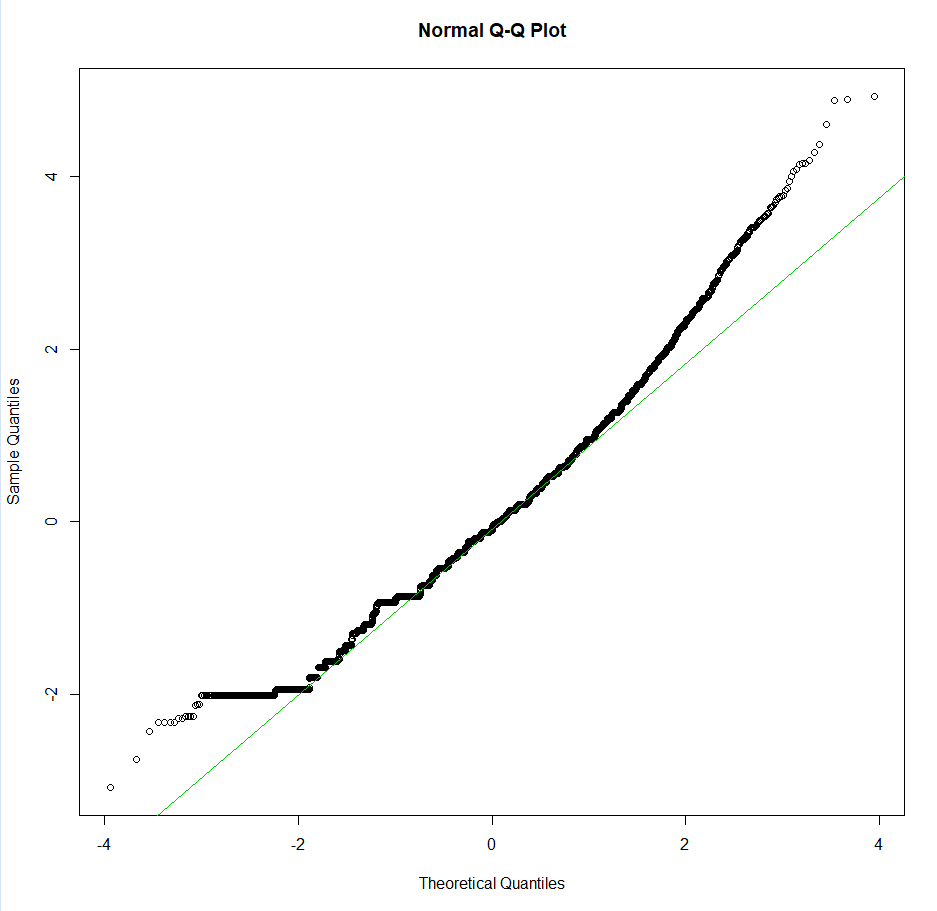
Model evaluation for the log transformed ANOVA model:

Residual analysis:



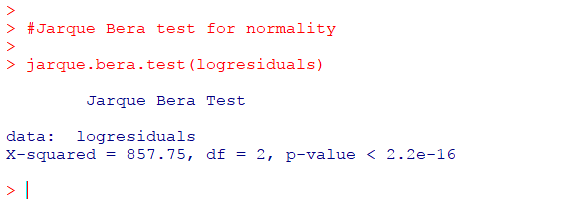
The within group variability is now constant.

Normal Probability plot



The residuals follow normal distribution.

To confirm we do Jarque-Bera test for Normality



Since, p-value<0.05, we can confirm the residuals follow normal distribution.

Thus, the log transformed model confirms our assumptions on normality and constant variance.

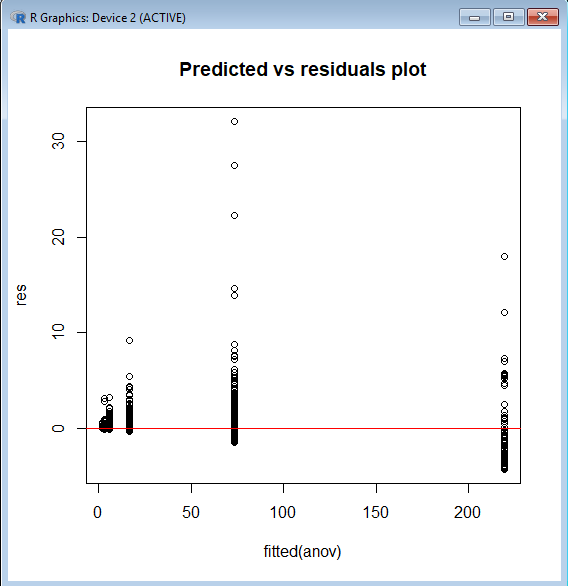
Hence, the log transformed model is the best fitted ANOVA model with which we conclude that we can reject null hypothesis, and the difference in mean loss in some regions are significant.

**ANOVA model evaluation (Comparison of injuries due to tornadoes of different magnitudes)**

Residual Analysis:

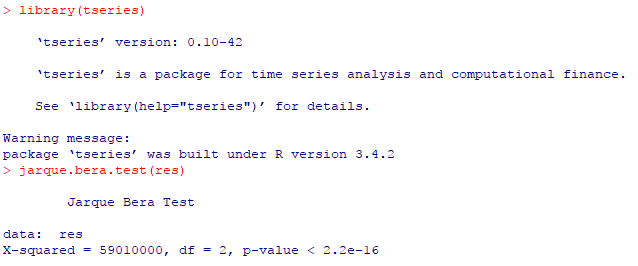
1. Residual plot – is the within group variability constant?
2. Normal Probability plot – does it follow normal distribution?

Predicted vs Residual plot:



The within group variance is almost constant.

Test for normality:

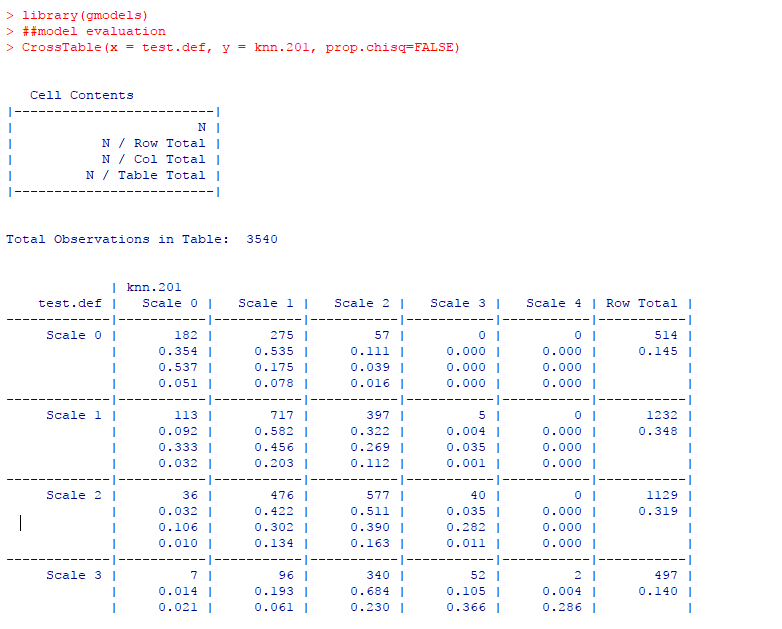


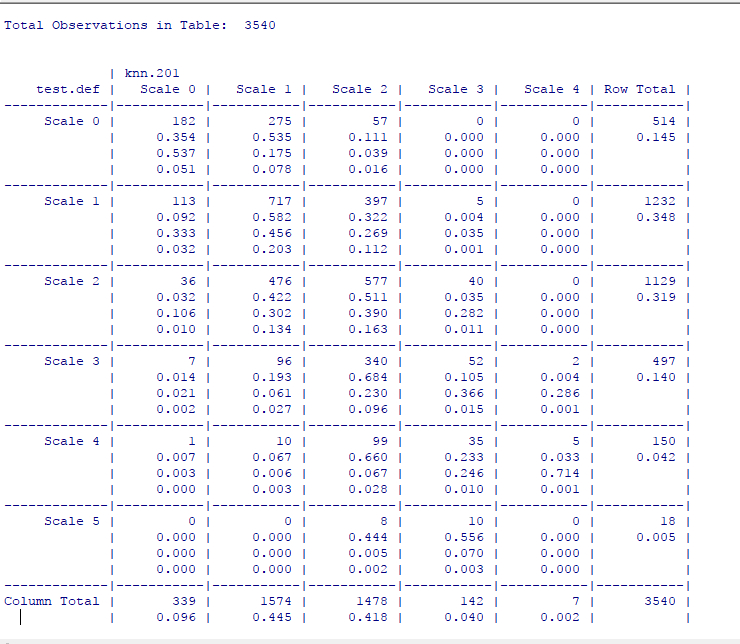
Since p<0.05 for Jarque Bera test on residuals, we say the residuals follow normal distribution.

Hence, our model is sufficient to conclude on the ANOVA test hypothesis and we reject null hypothesis, i.e., the difference in injuries is significant for some mag categories.

**KNN Classification Model Evaluation (Group Magnitude in terms of Length, width and distance travelled by a tornado)**

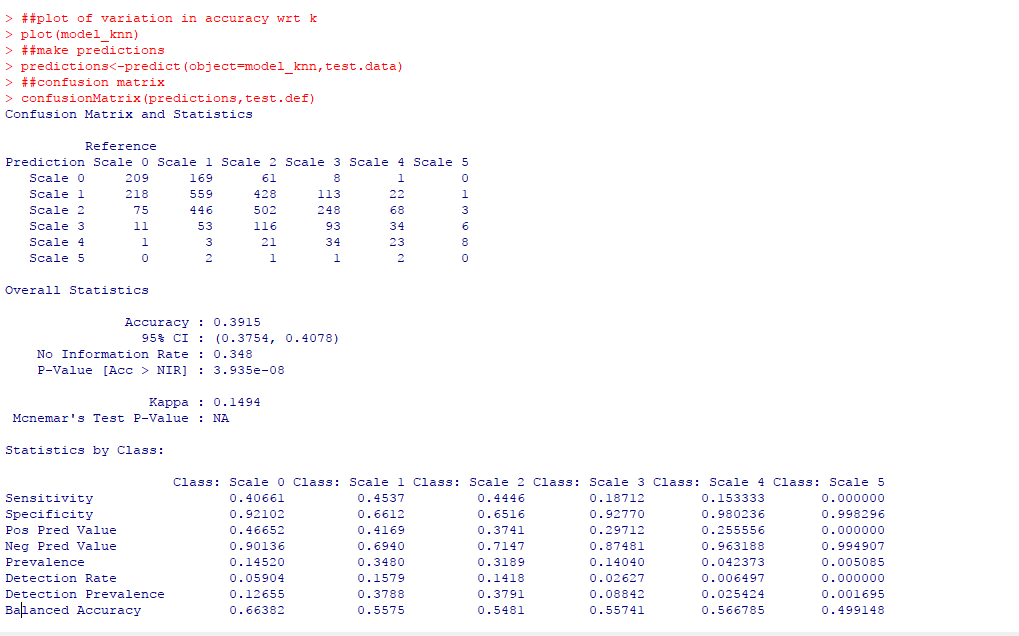
We evaluate the kNN model, knn.201 obtained for k=201 which had better accuracy.





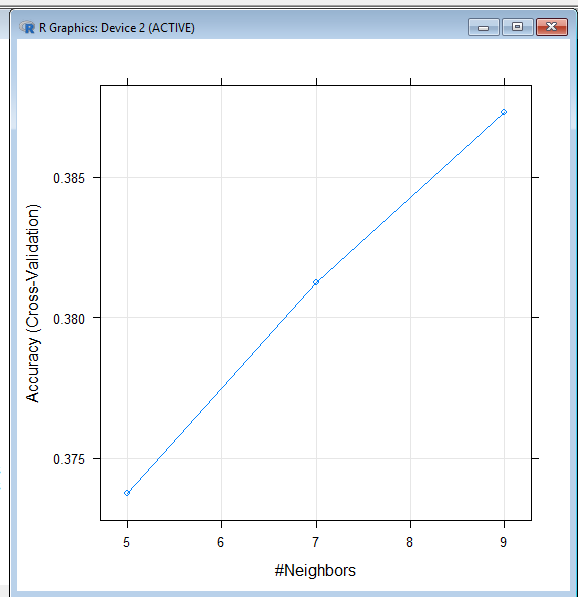
As per above cross table, the prediction results on test data does not seem good, as for example 182 records of magnitude 0 were predicted correctly as mag0, 275 which were actually of mag0 was predicted incorrectly as Scale 1, which is not preferred.

Also, on evaluating data obtained by training and resampling by cross validation, the overall accuracy is less 0.39.



As per the confusion matrix, the accuracy of predicted data is less.

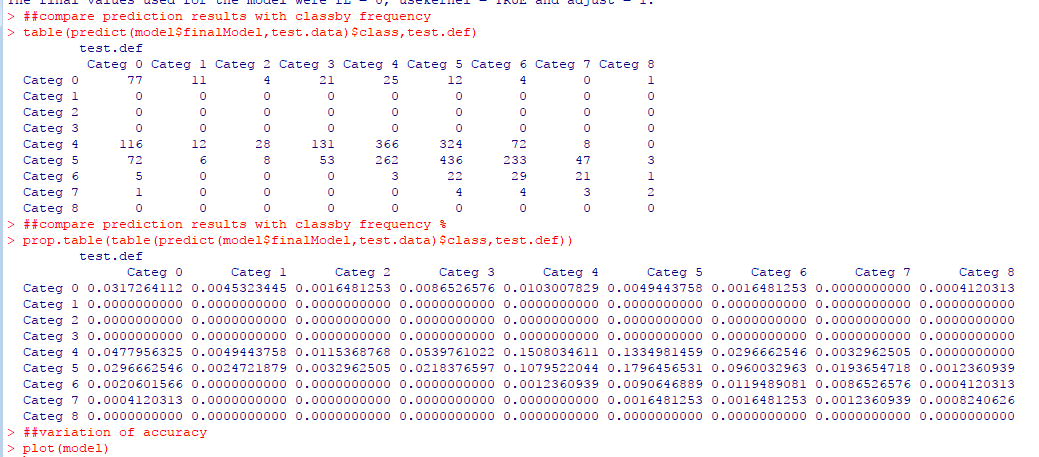
We plot the variation of accuracy with respect to k-value. We found, accuracy is of range 0.32 to 0.39.



Hence, we conclude that we cannot group magnitude category based on length, width and distance of tornado.

**Naïve Bayes Classification Model Evaluation (Group Property Loss category in terms of Magnitude category of tornado, region of occurrence and no. of states affected)**

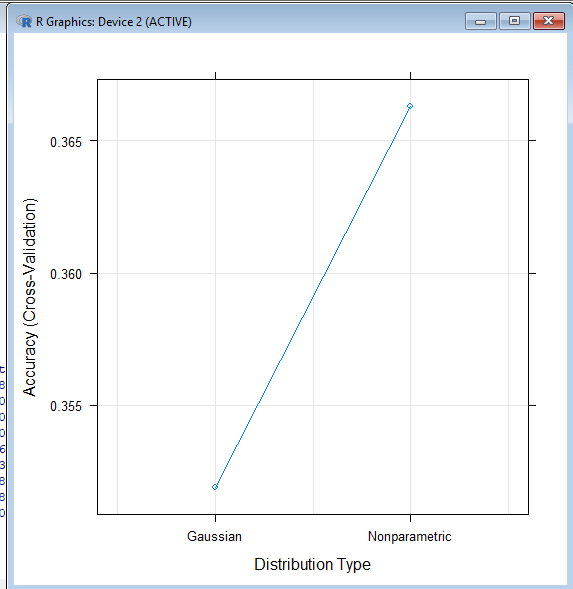
We predict on test data set, to evaluate the model.



The first prediction table shows the actual no. of predicted values and the second shows the % of data predicted on various categories.

The prediction results are not good, as for example, for loss category 0, 77 samples were predicted correctly as category 0, but 11 which are actually category 1 were predicted wrongly to be of category 1.

We also evaluated the accuracy variation by model plot.



The accuracy seems to be varying between 0.33 and 0.37. Hence, **we conclude we cannot accurately predict loss category based on magnitude, region and no. of states affected.**

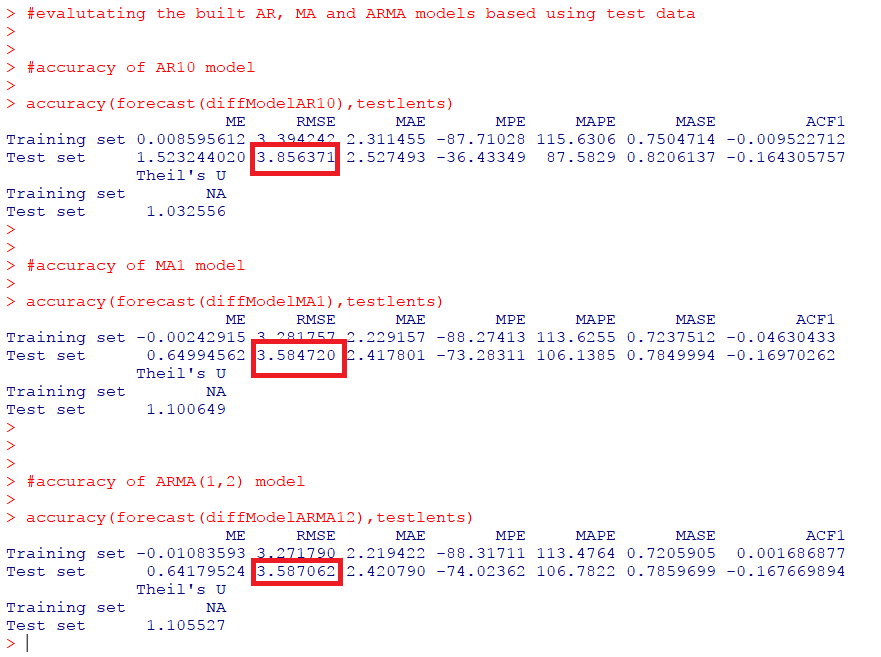
**Time series modelling of Tornadoes length data**

**Evaluation method used for splitting data for time series**

We used the Hold-out evaluation method to split the data into training and testing data

**Evaluation method applied for choosing the best model**

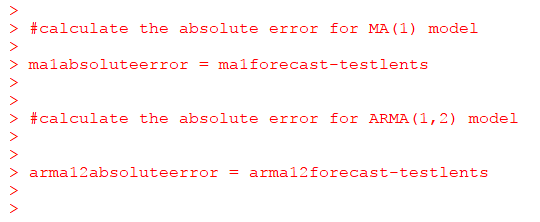
We used the accuracy () function in R to apply the models built using training data on the test data to arrive at the RMSE values of all the models that we built. It is this RMSE value that we use as the criteria to choose the best model as shown below



We could see that the RMSE values of MA (1) model and ARMA (1,2) model are the lowest compared to that of AR (10) model

But MA (1) and ARMA (1,2) model have very close RMSE values and hence we have a doubt if both these RMSEs are the same or not. So, to confirm this, we will perform two paired sample two-tailed hypothesis testing on absolute errors and make a conclusion on the best time series model for prediction.

Calculating absolute errors for MA (1) model and ARMA (1,2) models as shown below



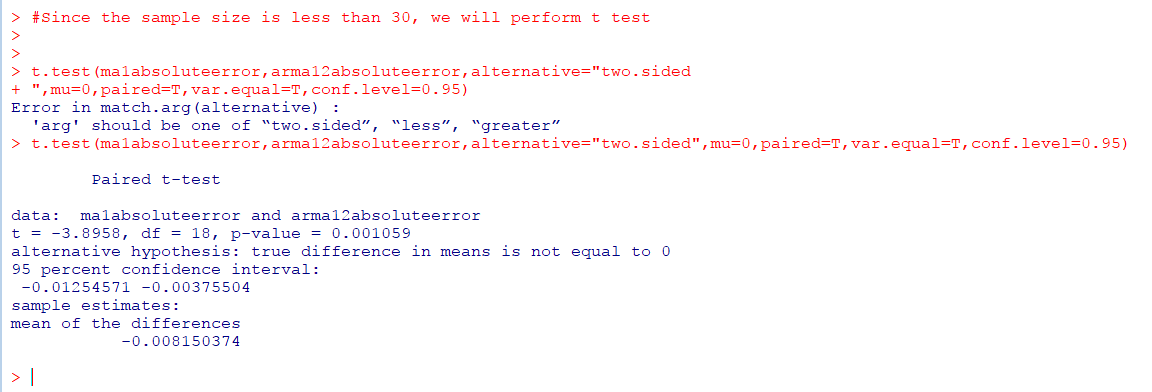
Since both the models are built on the same data and the mean errors are also for the same data, we will carry out two sample paired hypothesis.

Next step is to perform the two paired samples two tailed hypothesis testing. Since the sample size here is less than 30, we perform t-test.

The Hypothesis statements are:

*H0: Errors of models MA (1) and ARMA (1,2) are the same*

*Ha: Errors of models MA (1) and ARMA (1,2) are not the same*



From the above hypothesis testing screenshot, we see that the p-value is less than 0.05. Hence, we do not have enough evidence to accept the null hypothesis at 95% confidence level. Hence, we can conclude that the errors of models MA (1) and ARMA (1,2) are not the same.

Hence, we confirm that **model MA (1) is the best model for tornado length prediction for different months in future**

## 6.2. Results and Findings

* **By Hypothesis Testing**
* **Average distance travelled by tornadoes of magnitude 3 is more than that of magnitude 4**
* **No of injured people affected by tornadoes in states Alabama and Texas are not the same**
* **Average number of fatalities caused by tornadoes is 4**
* **Number of states affected by tornado of magnitude 0 to 3 is less than 3**
* **Comparison of Average loss incurred across different regions by log transformed ANOVA model**

The average loss incurred across various regions are not the same. The difference is significant for SouthEast, SouthWest and West with respect to MidWest, and insignificant for NorthEast with respect to MidWest.

* **Comparison of injuries due to tornadoes of different magnitudes by ANOVA model**

The average no. of injuries due to tornadoes of different magnitude categories are not the same. The difference is significant for magnitude categories 3,4 and 5 with respect to scale 0. But, it is insignificant for magnitude categories 1 and 2 with respect to 0.

* **Group Magnitude in terms of Length, width and distance travelled by a tornado by KNN classification**

We cannot deduce magnitude category of tornado accurately based on length, width and distance.

* **Group Property Loss category in terms of Magnitude category of tornado, region of occurrence and no. of states affected by Naïve Bayes Classification**

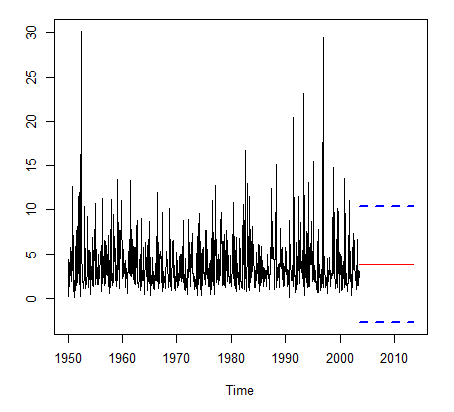
We cannot deduce property loss category accurately based on magnitude, region and no. of states affected.

* **Prediction of tornado average length for future years:**

On evaluating the different models based on the RMSE value, we conclude that **the MA (1) model is the best model in predicting the tornado length data for the coming years.**

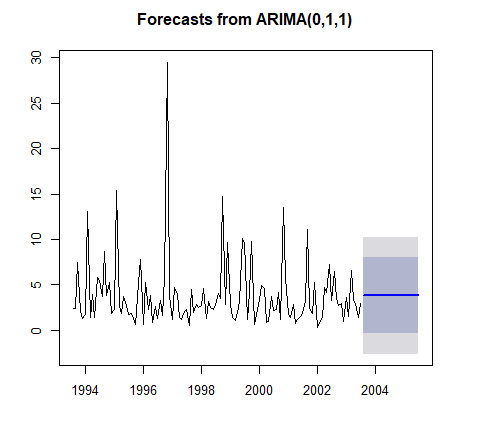
The predictions graph and the forecast graph for the next few years are as given below

**Prediction Graph:**



The lines in blue represent the prediction interval and the one in red indicates the predicted tornado length for the years from 2004.

**Forecast Graph**



The line in blue represents the forecasted tornado length for the years after 2004, the dark gray portion represent the 80% confidence interval and the light gray portion represents the 95% confidence interval.

# **7. Conclusions and Future Work**

## 7.1. Conclusions

We were able to predict, and forecast tornadoes based on the time series model. But, we could not find any effective relationship between the tornado and attributes defining a tornado, that could help predict the magnitude or the impact on people and property.

## 7.2. Limitations

The data provides a huge amount of information regarding the characteristics of the tornado. But, it has not taken into consideration, the geographic conditions of the region where it occurred. This could have helped identify the factors contributing to the tornado and its intensity.

Also, the data being vast, a deeper knowledge on the various analytics techniques was required, along with better understanding of the domain of the dataset.

## 7.3. Potential Improvements or Future Work

There is scope for various analytical models on the data set, especially time based. And, with attributes pertaining to the geographic conditions at the region of occurrence during the time could help identify the factors, and hence predict the intensity and extend of loss on people and property.