**TIME SERIES ANALYSIS AND AIR PASSENGER FORECAST**

Submitted as a part of

**DATA SCIENCE AND BIG DATA ANALYTICS**

Course Requirement

15IT423E

By-

**ASHUTOSH UPADHYAY - RA1711003010663**

**J.SRIYASH - RA1711003011257**

**ISHAN JAIN - RA1711003010767**

Under the supervision of

**Ms. B. GRACELIN SHEENA**

**Teaching Staff**

(102356)

Department of Computer Science and Engineering SRM Institute of Science and Technology Kattankulathur, Chennai.



**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**FACULTY OF ENGINEERING AND TECHNOLOGY DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**DECLARATION**

We **ASHUTOSH UPADHYAY- RA1711003010663, J.SRIYASH- RA1711003011257, ISHAN JAIN- RA1711003010767** studying in III year, V semester, B.Tech in Computer Science and Engineering at SRM Institute of Science and Technology, Kattankulathur, Chennai, hereby declare that this Mini project is an original work of mine and we have not verbatim copied / duplicated any material from sources like internet or from print media, excepting some vital company information / statistics and data that is provided by the Technical organizations itself.

Signature of the Student :

Date:

Place: SRM IST

**ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend my sincere thanks to all of them.

We are highly indebted to my Data Science and Big Data Analytics instructor

Mrs B.Gracelin Sheena, for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express my gratitude towards the HOD of Computer Science Department Dr. B. Amutha for her kind co-operation and encouragement which help us in completion of this project.

My thanks and appreciations also go to my peers in developing the project and people who have willingly helped me out with their abilities.

**ABSTRACT**

**Time series analysis** and **forecasting** future values has been a major research focus since years ago. In this data finds it significance in many applications such as business, stock market and exchange, weather, cost and usage of products such as fuels, electricity, etc. and in any kind of place that has specific seasonal or trendy changes with time. It provides the organization with useful information that is necessary for making important decisions. This survey covers the overall forecasting models, the algorithms used within the model and the other optimization techniques used for better performance and accuracy. The most widely employed approach is based on the class of models known as Autoregressive Integrated Moving Average (ARIMA) models. This techniques include multiunit time-series design, multivariate time-series analysis, inclusion of covariates, and analysis of patterns of intra-individual differences across time. In this we are working on the time series analysis of Air Passengers.

**LIST OF FIGURES / GRAPHS**

**S no. Graph name**

1 Stationary & non-stationary

2 Acf graph

3 Pacf graph

4 Box plot

5 Trend graph

6 Seasonal graph

7 Randomness graph

8 Model Residual graph

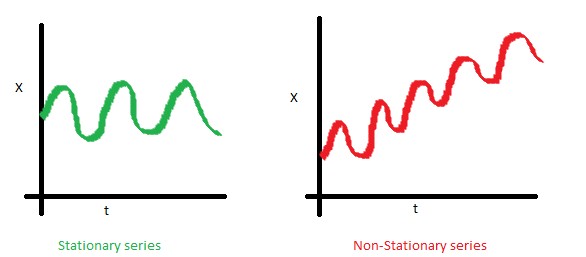
9 Forecast graph

**INTRODUCTION**

A time series is a sequence of data being recorded at specific time intervals. These data points are analysed to forecast a future. It is time dependent.

Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

A time series is a series of data points indexed (or listed or graphed) in time order Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.



**Stationary and non-stationary graphs**

Methods for time series analysis may be divided into two classes: frequency-domain methods and time-domain methods .In the time domain, correlation and analysis can be made in a filter- like manner using scaled correlation, thereby mitigating the need to operate in the frequency domain. Additionally, time series analysis techniques may be divided into parametric and non- parametric methods. The parametric approaches assume that the underlying stationary stochastic process has a certain structure which can be described using a small number of parameters (for example, using an autoregressive or moving average model). In these approaches, the task is to estimate the parameters of the model that describes the stochastic process. By contrast, non- parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure. Methods of time series analysis may also be divided into linear and non-linear, and univariate and multivariate.

**REVIEW OF LITERATURE**

The main objective is to predict the airline tickets sales of 1961 using time series analysis. In this we will be analysing the 10 years air-ticket sales data of airline industry from 1949-

1960.The behaviour of time series is the identification of the time series components like trend, seasonality to describe the behaviour. Time series analysis forecasts the values depending on its actual and past values. The authors find evidence that generalized autoregressive conditional heteroscedasticity models outperform their autoregressive– moving-average counterparts and that the consideration of explanatory variables improves forecasts. Various statistics are presented to characterize the empirical literature on electricity spot price modelling, and the forecasting performance of several model types and modifications is analysed. Modern time series forecasting methods are essentially rooted in the idea that the past tells us something about the future. Of course, the question of how exactly we are to go about interpreting the information encoded in past events, and furthermore, how we are to extrapolate future events based on this information, constitute the main subject matter of time series analysis. Typically, the approach to forecasting time series is to first specify a model, although this need not be so. This model is a statistical formulation of the dynamic relationships between that which we observe (i.e. the so called information set), and those variables we believe are related to that which we observe. The key issue of this study is to offer a comparison between different model types and modelling conditions regarding their forecasting performance, which is referred to as a quasi-meta- analysis, i.e. the analysis of analyses to achieve more general findings independent of the circumstances of single studies. The predicted graph can be changed accordingly for the number of years we actually want to predict and get the loss or profit in the future as output. During this period, over one third of all papers published in these journals concerned time series forecasting. We also review highly influential works on time series forecasting that have been published elsewhere during this period. Enormous progress has been made in many areas, but we find that there are a large number of topics in need of further development. This is what has been done before, the additional thing we have added is to find the accuracy (validate) of the output graph for the predicted years.

**RESEARCH METHODOLOGY**

**ARIMA**

ARIMA stands for autoregressive integrated moving average. This method is also known as the

Box-Jenkins method.

**Identification of ARIMA parameters:**

**Autoregressive component:** AR stands for autoregressive. Autoregressive paratmeter is denoted by p. When p =0, it means that there is no auto-correlation in the series. When p=1, it means that the series auto-correlation is till one lag.

**Integrated:** In ARIMA time series analysis, integrated is denoted by d. Integration is the inverse of differencing. When d=0, it means the series is stationary and we do not need to take the difference of it. When d=1, it means that the series is not stationary and to make it stationary, we need to take the first difference. When d=2, it means that the series has been differenced twice. Usually, more than two time difference is not reliable.

**Moving average component:** MA stands for moving the average, which is denoted by q. In ARIMA, moving average q=1 means that it is an error term and there is auto-correlation with one lag.

It is specified by three parameters : p, d, q

P=number of auto regressive terms

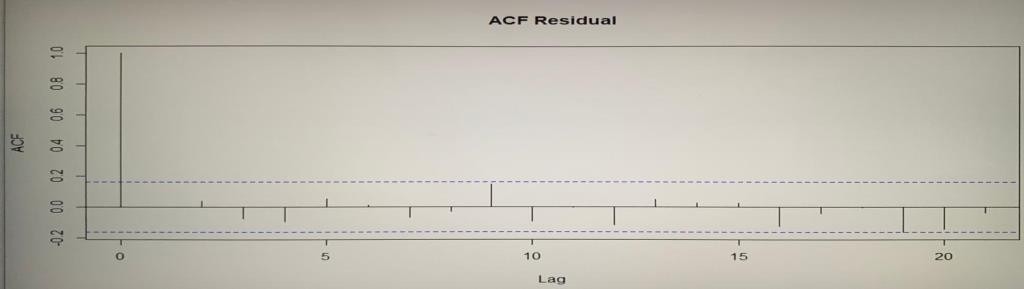
D=how many non-seasonal differences required to achieve stationarity

Q=number of lagged forecast errors in the prediction equation.

In terms of regressive model, auto regressive components refer to prior values of current values. Level of differencing would mean to take the current value and subtract the prior value from it. ARIMA model works on the assumption of stationarity.

**Acf – Auto correlation function**

Auto correlation is the similarity between values of a same variable across observations. It tells you how correlated points are with each other, based on how many times steps they are separated by. It is used to determine how past and future values are related. Its values range from -1 to +1.

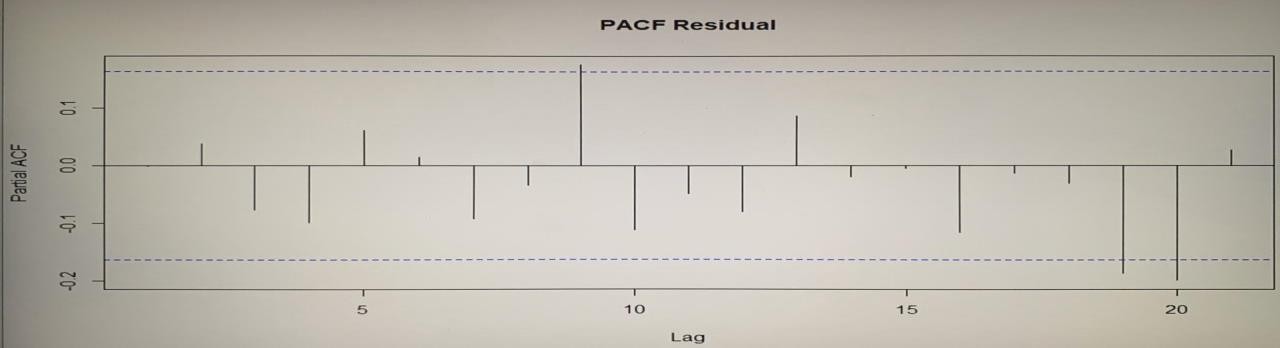


Grapg 1.2 : acf graph

**Pacf – Partial Auto Correlation function**

Partial auto correlation is the degree of association between two variables while adjusting the effect of one or more additional variables. Pacf (partial auto correlation function) gives partial correlation of time series with its own tagged values. Its values can range from -1 to

+1.



pacf graph

**LJUNG-BOX TEST**

The **Ljung–Box test** is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the overall randomness based on a number of lags, and is therefore a portmanteau test.

**DATA ANALYSIS**

**Data collection**

The data scientist identifies and gathers data resources—structured, unstructured and semi- structured—that are relevant to the problem domain. On encountering gaps in data collection,

the data scientist might need to revise the data requirements and collect more data.Here we have included AIR PASSENGERS data set which consists 10 years (1949-1960) of air tickets sales.

**Data Set Table**

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

1949 112 118 132 129 121 135 148 148 136 119 104 118

1950 115 126 141 135 125 149 170 170 158 133 114 140

1951 145 150 178 163 172 178 199 199 184 162 146 166

1952 171 180 193 181 183 218 230 242 209 191 172 194

1953 196 196 236 235 229 243 264 272 237 211 180 201

1954 204 188 235 227 234 264 302 293 259 229 203 229

1955 242 233 267 269 270 315 364 347 312 274 237 278

1956 284 277 317 313 318 374 413 405 355 306 271 306

1957 315 301 356 348 355 422 465 467 404 347 305 336

1958 340 318 362 348 363 435 491 505 404 359 310 337

1959 360 342 406 396 420 472 548 559 463 407 362 405

1960 417 391 419 461 472 535 622 606 508 461 390 432

**Data understanding**

Descriptive statistics and visualization techniques can help a data scientist understand data content, assess data quality and discover initial insights into the data. A revisiting of the previous step, data collection, might be necessary to close gaps in understanding. Here we check for any null values in the data set and visualize the data through various plots.

**Cycle Period Table**

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

1949 1 2 3 4 5 6 7 8 9 10 11 12

1950 1 2 3 4 5 6 7 8 9 10 11 12

1951 1 2 3 4 5 6 7 8 9 10 11 12

1952 1 2 3 4 5 6 7 8 9 10 11 12

1953 1 2 3 4 5 6 7 8 9 10 11 12

1954 1 2 3 4 5 6 7 8 9 10 11 12

1955 1 2 3 4 5 6 7 8 9 10 11 12

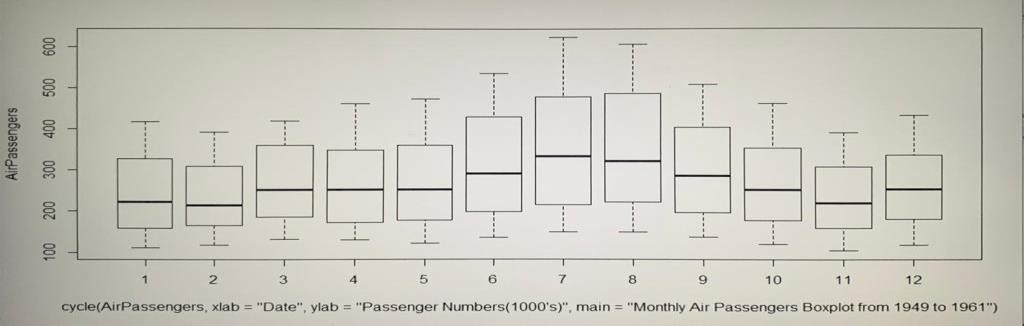
1956 1 2 3 4 5 6 7 8 9 10 11 12

1957 1 2 3 4 5 6 7 8 9 10 11 12

1958 1 2 3 4 5 6 7 8 9 10 11 12

1959 1 2 3 4 5 6 7 8 9 10 11 12

1960 1 2 3 4 5 6 7 8 9 10 11 12



box plot

**Data preparation**

We will decompose the time series for estimates of trend, seasonal, and random components using moving average method. The data preparation stage comprises all activities used to construct the data set that will be used in the modelling stage. These include data cleaning, combining data from multiple sources and transforming data into more useful variables. Moreover, feature engineering and text analytics may be used to derive new structured variables, enriching the set of predictors and enhancing the model’s accuracy.

Air passengers is the data set that we use in this model.it consists of 10 years air ticket data sales data of airline industry from 1949-1960. Forecast and tseries are the packages we use in this model to predict the output. We will decompose the time series for estimates of trend, seasonal, and random components using moving average method.

The multiplicative model is:

Y[t] =T[t] ∗S[t] ∗e[t] Y[t] =T[t] ∗S[t] ∗e[t] Where

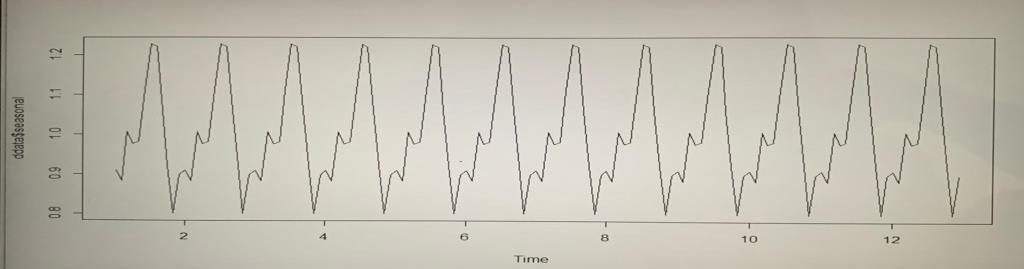
Y (t) is the number of passengers at time t

T (t) is the trend component at time t,



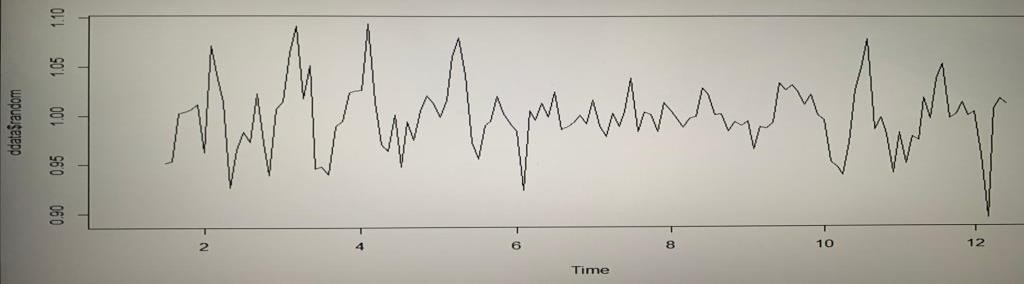
trend graph

S (t) is the seasonal component at time t,



seasonal graph

E (t) is the random error component at time t.



randomness graph

**1. Test stationarity of the time series (ADF)**

In order to test the stationarity of the time series, let’s run the Augmented Dickey-Fuller Test using the adf test function from the tseries R package.

First set the hypothesis test:

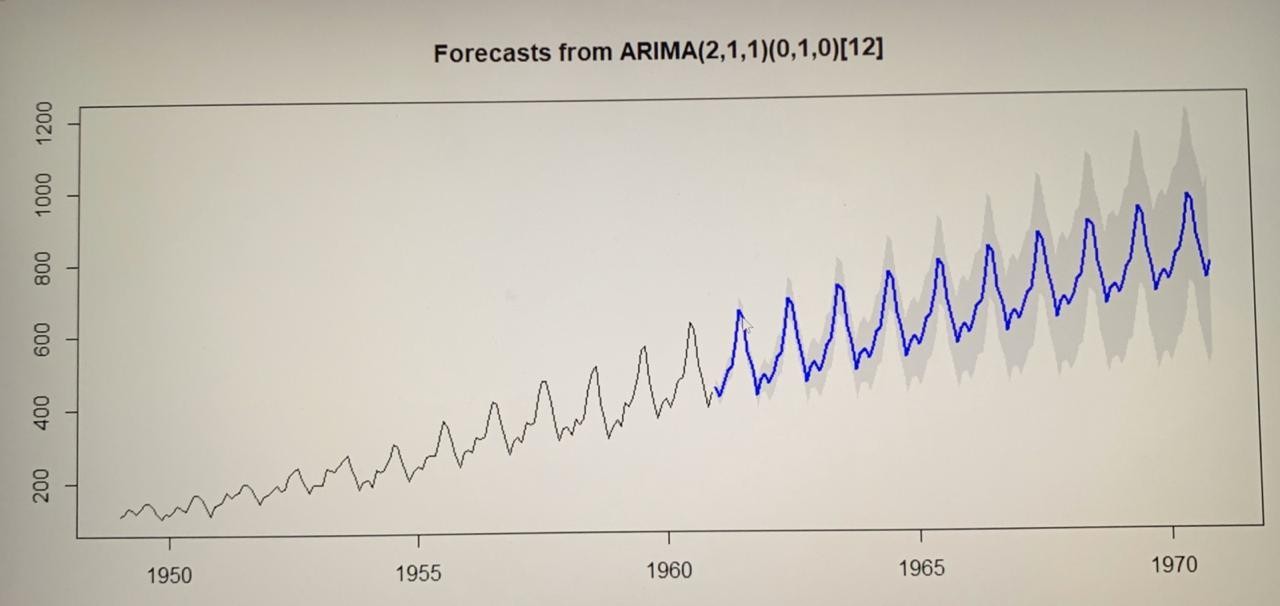
The null hypothesis H0 : that the time series is non stationary

The alternative hypothesis Ha : that the time series is stationary

**INTERPRETATION AND FINDINGS**

Here we find the output for 10 years. These 10 years can be varied accordingly. Here the data set is from 1949-1960. The blue curve represents the trend in the next 10 years (1961-1970), and the shaded part of the curve represents the maximum and minimum value of variation.

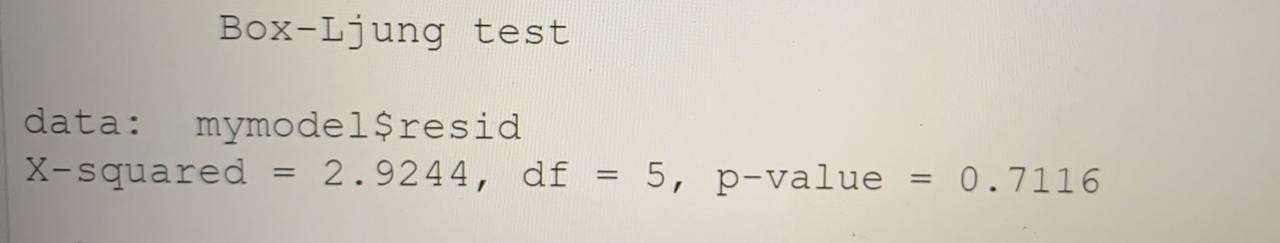
Best model: ARIMA(0,1,1)(2,1,0)[12] Series: AirPassengers ARIMA(0,1,1)(2,1,0)[12]



**Validating the forecast**:

Here we find the accuracy of the predicted value for the given input period.

The **Ljung–Box test** is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the overall randomness based on a number of lags, and is therefore a portmanteau test.



We can arbitrarily select the lag value i.e., 5,10,15 ,etc and the p-values are insignificant which indicates that the model is free from auto-correlation

**Code:**

install.packages('forecast') library(forecast) data("AirPassengers") class(AirPassengers) start(AirPassengers) end(AirPassengers) frequency(AirPassengers) sum(is.na(AirPassengers)) summary(AirPassengers) AirPassengers

tsdata<-ts(AirPassengers,frequency = 12) ddata<-decompose(tsdata,"multiplicative") plot(ddata)

plot(ddata$trend) plot(ddata$random) plot(AirPassengers)

abline(reg=lm(AirPassengers~time(AirPassengers)))

cycle(AirPassengers)

boxplot(AirPassengers~cycle(AirPassengers, xlab="Date",ylab= "Passenger Numbers(1000's)"

,main="Monthly Air Passengers Boxplot from 1949 to 1961"))

plot(AirPassengers)

plot(AirPassengers,xlab="Date", ylab = "Passenger numbers (1000's)",main="Air Passenger numbers from 1949 to 1961")

mymodel<-auto.arima(AirPassengers)

mymodel

auto.arima(AirPassengers,ic="aic" ,trace = TRUE)

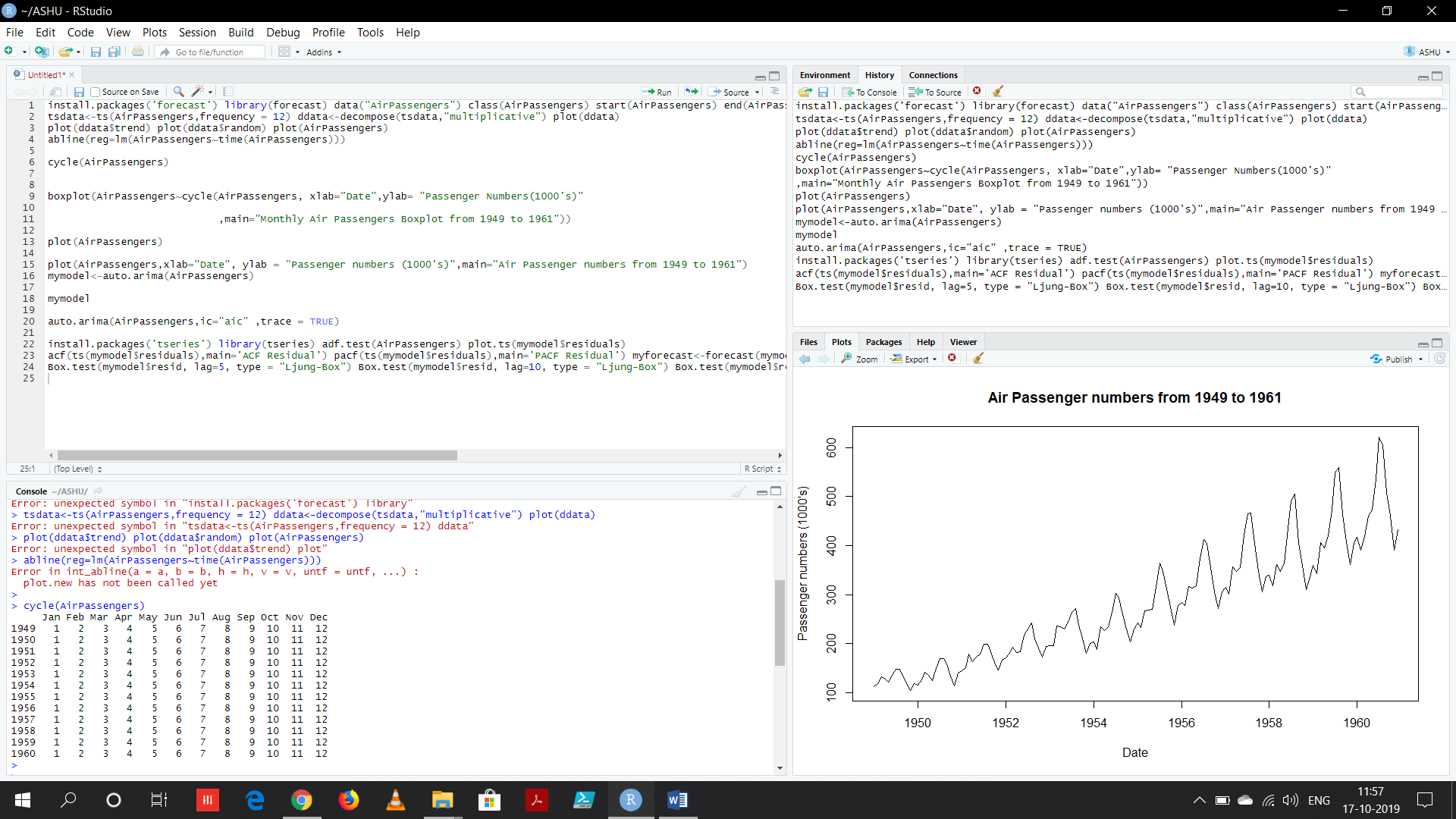
install.packages('tseries') library(tseries) adf.test(AirPassengers) plot.ts(mymodel$residuals)

acf(ts(mymodel$residuals),main='ACF Residual') pacf(ts(mymodel$residuals),main='PACF Residual') myforecast<-forecast(mymodel , level=c(95),h=10\*12) plot(myforecast)

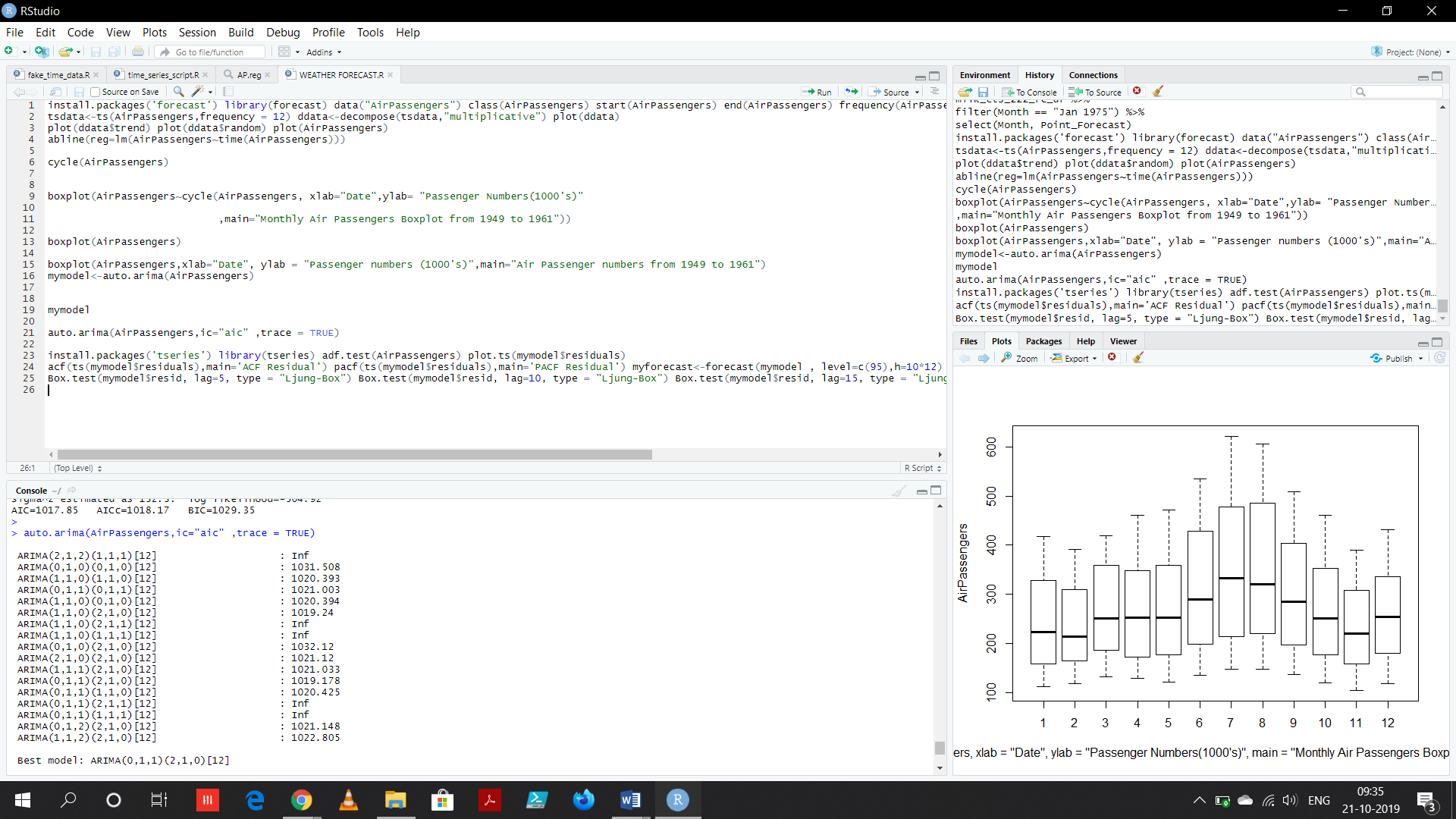
Box.test(mymodel$resid, lag=5, type = "Ljung-Box") Box.test(mymodel$resid, lag=10, type = "Ljung-Box") Box.test(mymodel$resid, lag=15, type = "Ljung-Box")

**Output:**

Analysis for further 10 years.



OUTPUT



**SUGGESTIONS AND CONCLUSION**

The major aim of forecasting time series data is to be able to determine and get a clue or understanding on how and for how long the observation will continue to future. As we have seen that, we can predict or forecast the variations. We can also include a feedback system from the customers ,in which we can segregate the negative and positive comments using some analysis algorithm. On which the organization can work on improving their drawbacks to increase the sales of air tickets. And they can plan the increasing occupancy and schedules for the upcoming time period for a better hospitality for the customers.

Time series analysis is very important because to each and every organization to succeed then it has to understand itself regarding performance, achievements, behaviorally, etc. We can conclude from the arima output , that our model using parameters (2,1,1) has been shown to adequately fit the data

**BIBLIOGRAPHY**

1. Introductory Time Series with R.- Paul Cowpertwait

2. Time Series Analysis and Its Applications: With R Examples. -Robert H. Shumway and

Stoffer

3. [www.statisticssolutions.com/time-series-analysis/](http://www.statisticssolutions.com/time-series-analysis/)

4. rstudio-pubs- static.s3.amazonaws.com/311446\_08b00d63cc794e158b1f4763eb70d43a.html

5. Wikipedia for reference