|  |  |  |
| --- | --- | --- |
| Predictive Analytics | | January 17  2018 |
| Implementing a Time Series Modelling algorithm on a Chocolate Production data. Checking and correcting seasonality of the sale, stats for highest and lowest production, predicting production of an upcoming month. | Chocolate Production Data - Monthly | |

*Team Members ( Group 6)–*

*Namit Gaur*

*Praveen Garlapati*

*Lohit Khanna*

*Akshaya Bharadwaj*

*Kumar Madhav*

*Harsh Vardhan*

# Time Series Model Using on US Chocolate Production Dataset

**Abstract**:

The objective of this analysis and modeling is to review basic time series theories and do some basic exploration on the dataset as well.

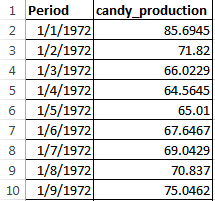
We will be following an ARIMA modeling procedure of the dataset as follows:

1. Exploratory Data Analysis
2. Decompostion of Data
3. Test the stationarity
4. Model fitting using an algorithm
5. Calculate forecasts

**Data Source:**

We have used the US Chocolate Production Dataset from kaggle. This dataset tracks industrial production every month from January 1972 to August 2017. It contains two variables:

1. observation\_date
2. chocolate\_production



**Setup:**

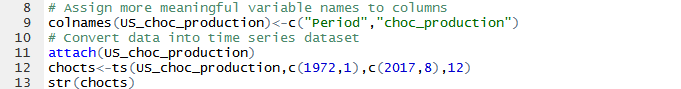
**Loading packages**



**Loading data**





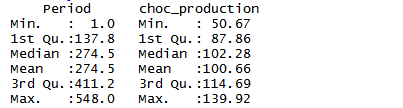




**Exploratory Data Analysis:**

**Summary of Data**





**Check missing values**





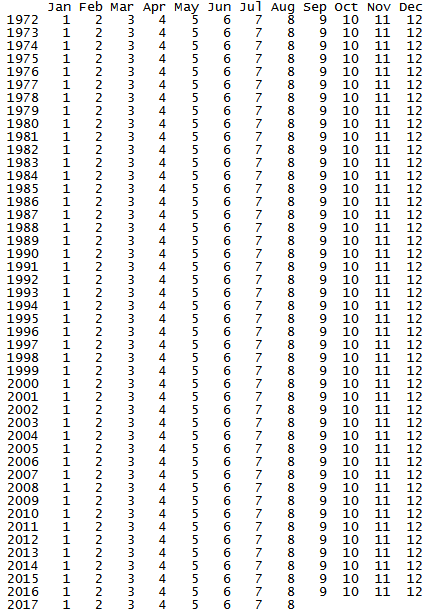
**Check frequency of time series data**





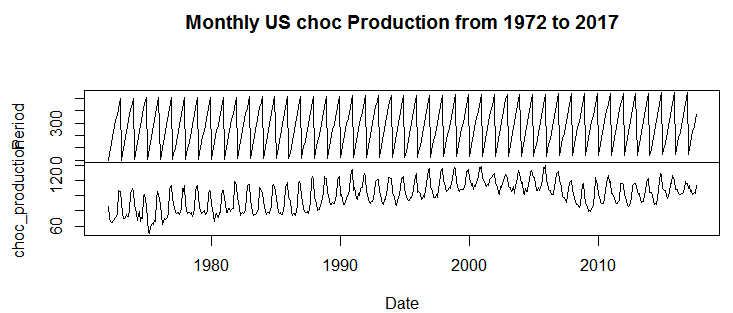
**Check the cycle of time series data**





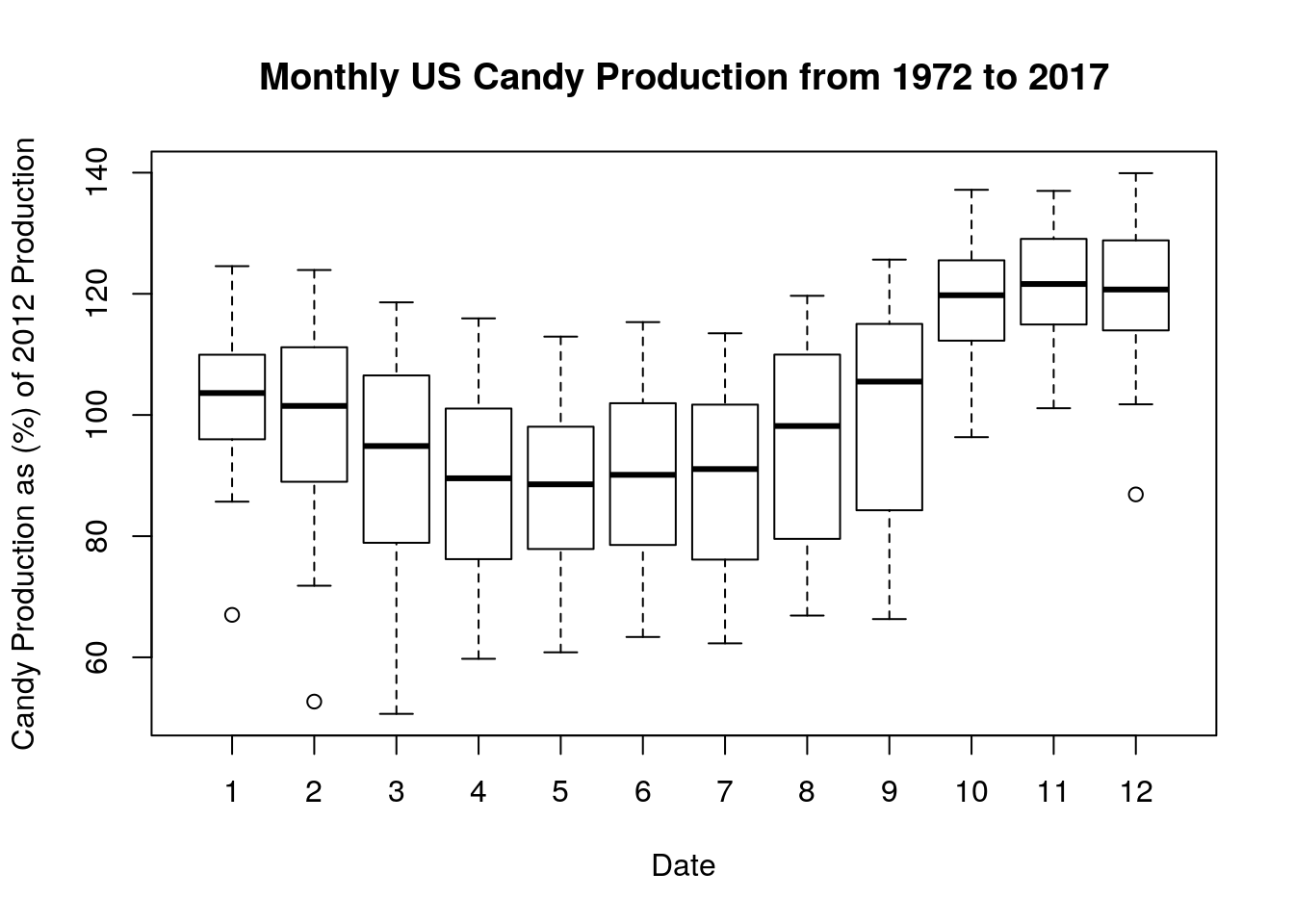
**Plot the raw data using the base plot function**





**Boxplot function to see any seasonal effects**





Some inferences can be made from these plots:

1. The chocolate production numbers increase over time with each year from 1972 to early 2000’s which may be indicative of an increasing linear trend. However, production seems to slow down during mid to late 2000’s, while the trend during the present decade seems to be pretty stagnant.
2. Boxplot reveals that chocolate production is higher in months 10 to 12, having higher means and lower variances than the other months, indicating seasonality with an apparent cycle of 12 months.
3. Chocolate production dataset appears to be multiplicative time series as the production numbers increase, it appears so does the pattern of seasonality. There do not appear to be significant amount of outliers and there are no missing values. Therefore no further data cleaning is required.

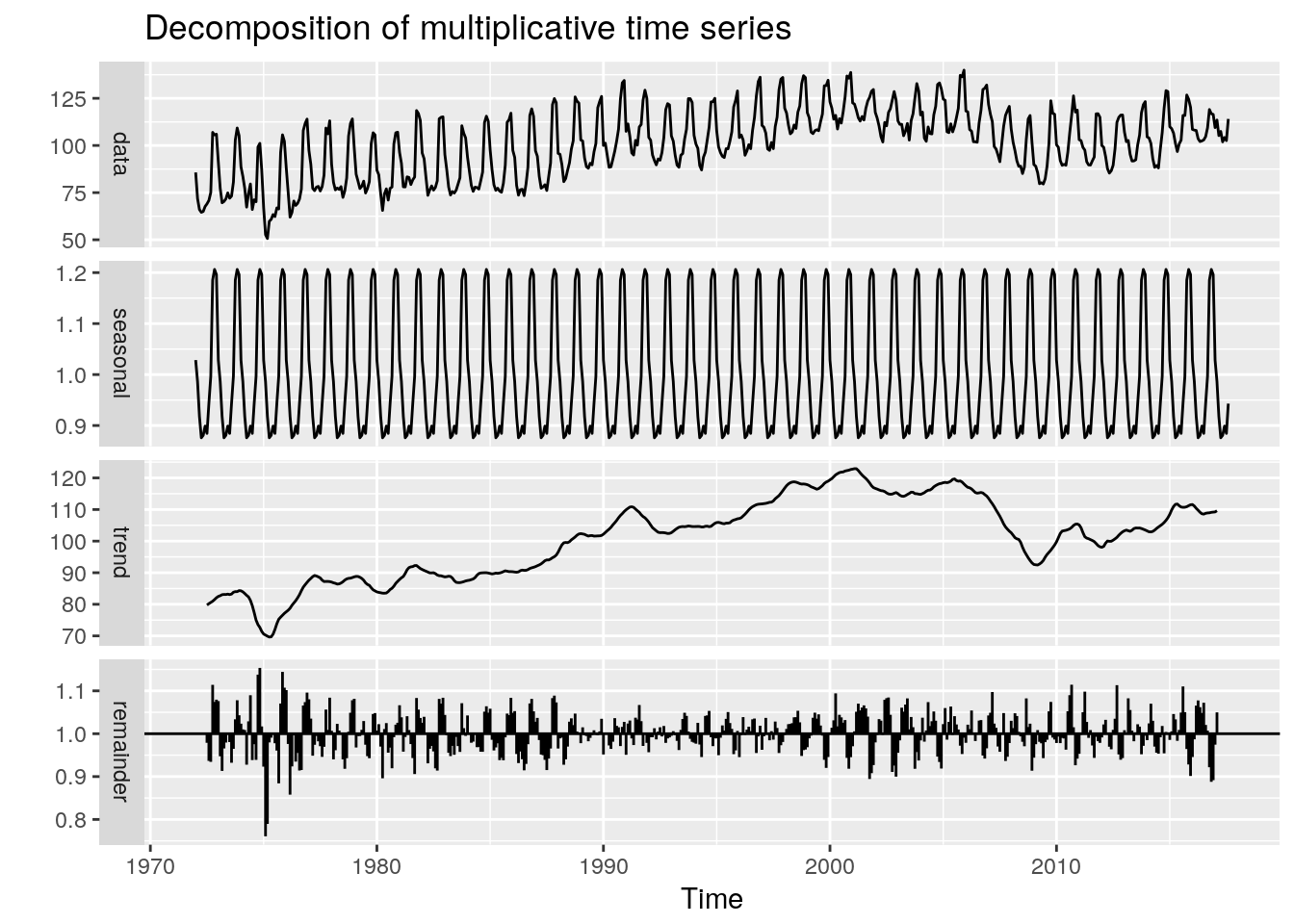
# TIME SERIES DECOMPOSITION:

The multiplicative model is:

Y[t]=T[t] x S[t] x e[t]

where Y(t) is the chocolate production at time t, T(t) is the trend component at time t, S(t) is the seasonal component at time t, e(t) is the random error component at time t.





In these decomposed plots we can again see the trend and seasonality as inferred previously, but we can also observe the estimation of the random component depicted under the remainder.

**Testing stationarity of the Time Series:**

To fit the arima series models, time series is required to be stationary. Below two methods have been used to test the stationarity.

## Augmented Dickey-Fuller

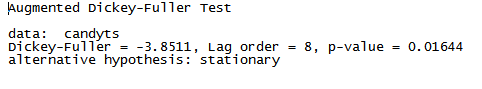
To test the stationarity of the time series, run the Augmented Dickey-Fuller Test using the adf.test() function.

Setting the hypothesis test:

H0 : that the time series is non stationary

HA : that the time series is stationary



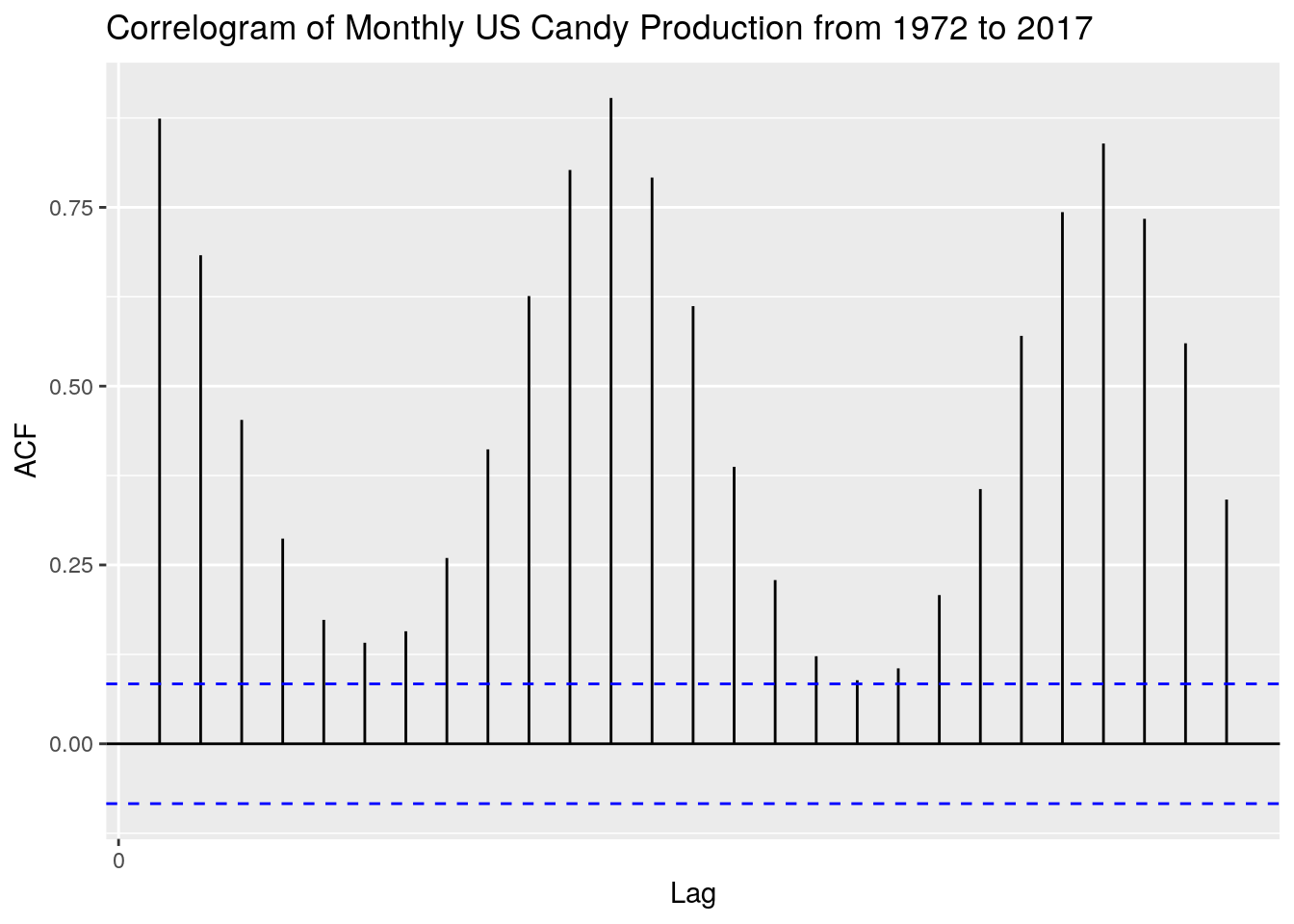


Since the p-value is less than 5%, we have a strong evidence against the null hypothesis, so we reject the null hypothesis. As per the test results above, the p-value is 0.016 which is <0.05 therefore we reject the null in favour of the alternative hypothesis that the time series is stationary.

1. Autocorrelation

This function plots the correlation between a series and its lags (ie previous observations) with a 95% confidence interval in blue. If the autocorrelation crosses the dashed blue line, it means that specific lag is significantly correlated with current series.

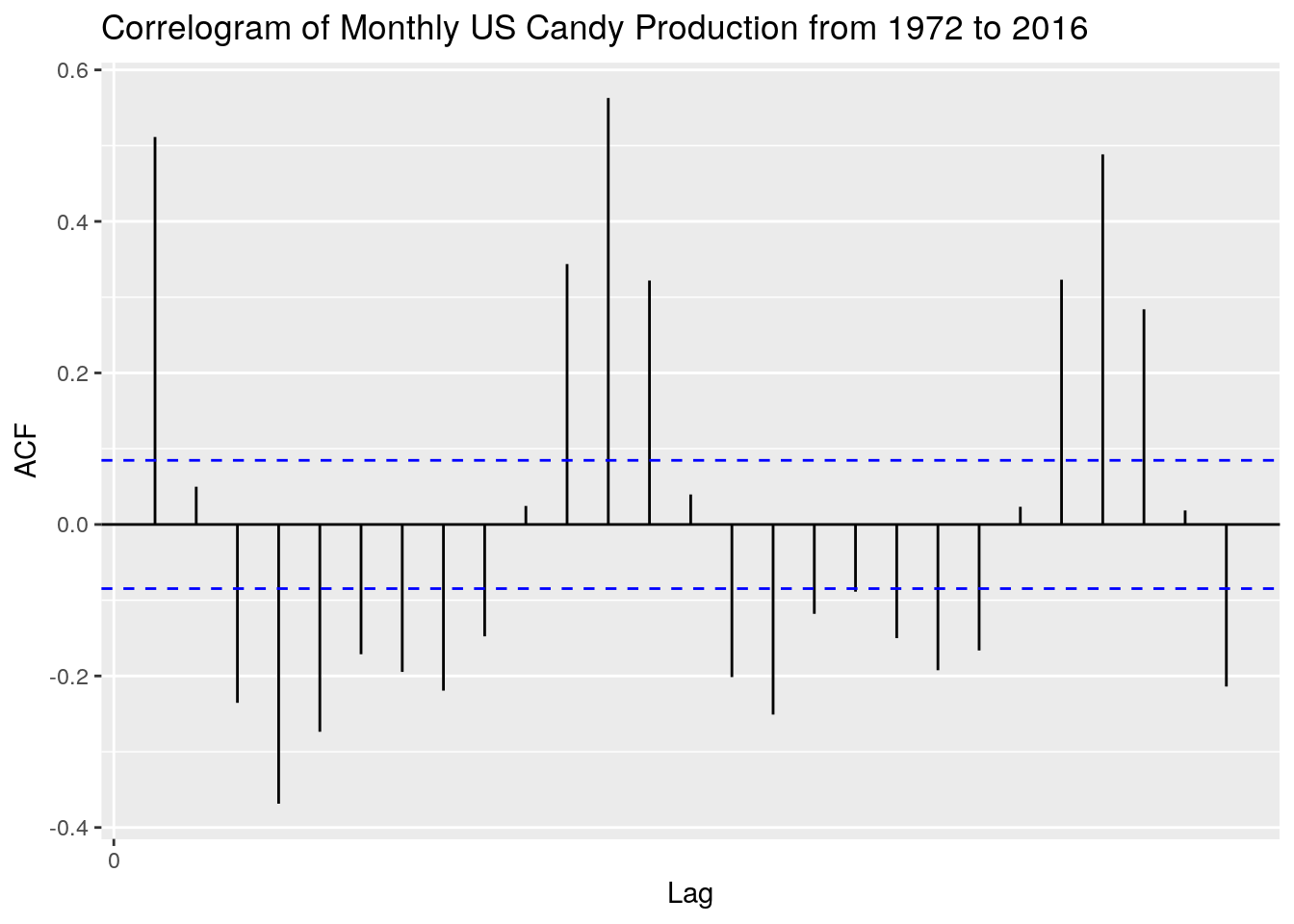




The maximum at lag 1 or 12 months, indicates a positive relationship with the 12 month cycle.

**Autoplot the random time series which exclude the NA values**



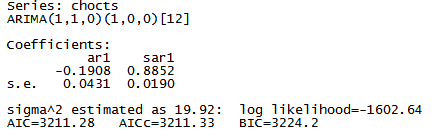


It can be seen the acf of the residuals is centered around 0

**Fit a Time Series model:**

**ARIMA MODEL**

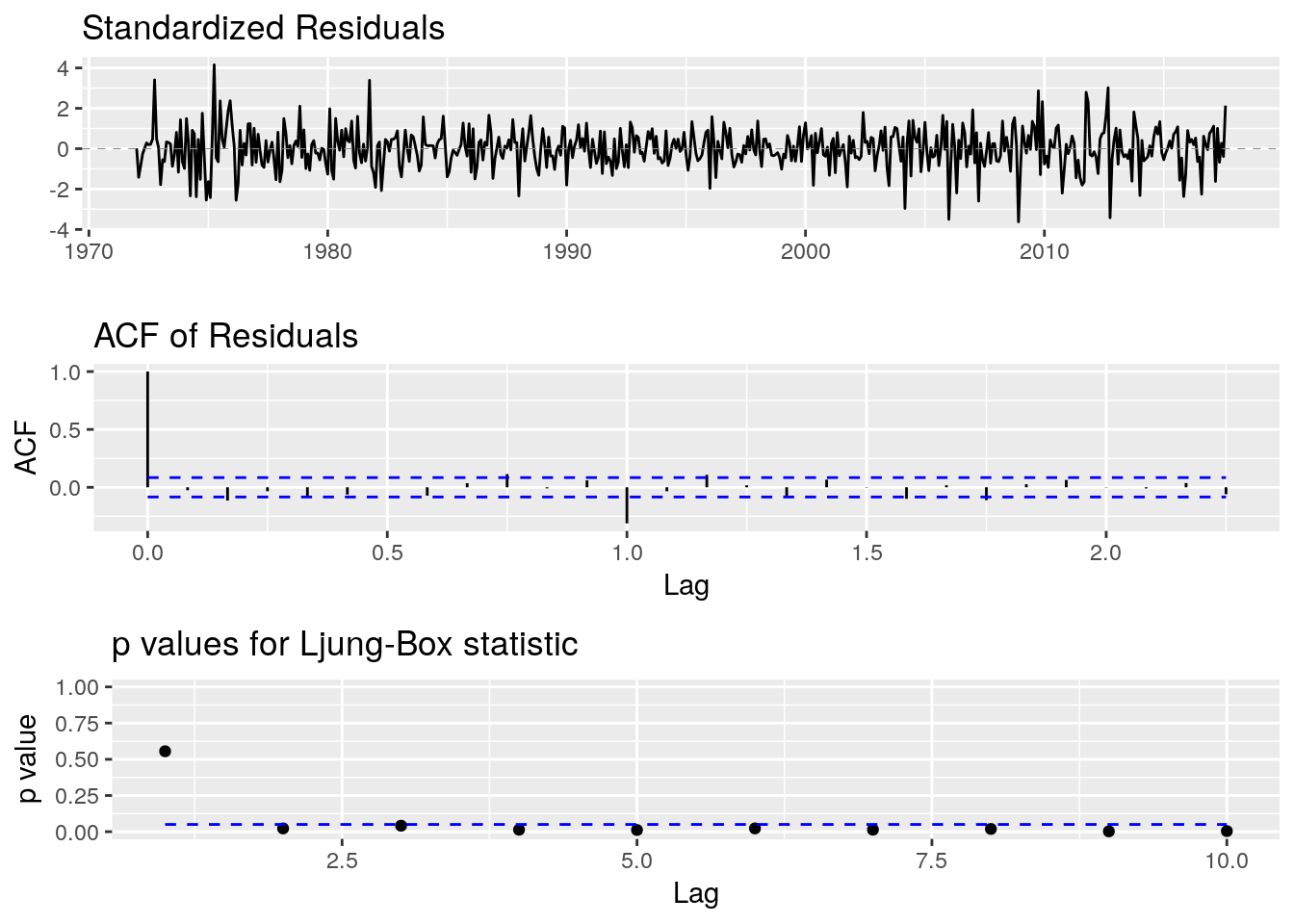




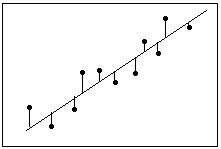
The ARIMA(1,1,0)(1,0,0)[12] model parameters are lag 1 differencing (d), and an autoregressive term of first lag (p). Then the seasonal model has an autoregressive term of first lag (D) at model period 12 units, in this case months.

The ggtsdiag function from ggfortify R package performs model diagnostics of the residuals and the acf will include an autocovariance plot.





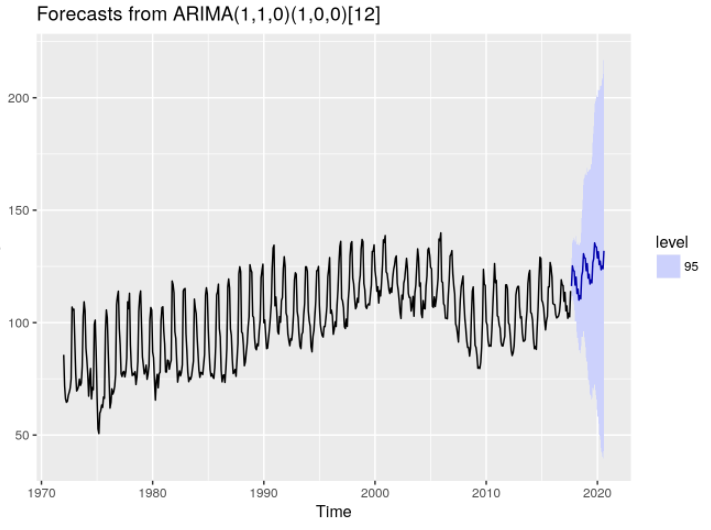
The residual plots appear to be centered around 0 as noise, with no pattern. The arima model is a fairly good fit.



# CALCULATE FORECASTS

# Finally plotting a forecast of the time series using the forecast function





# Conclusion and Recommendations

# There is no pattern in residual plot and using residual plots, we can assess whether the observed error (residuals) is consistent with stochastic error. The residuals is neither systematically high nor low. So, the residuals is centered on zero throughout the range of fitted values. In other words, the model is correct on average for all fitted values. Further, in the OLS context, random errors are assumed to produce residuals that are normally distributed. Therefore, the residuals should fall in a symmetrical pattern and have a constant spread throughout the range.

# As we considered seasonal ARIMA model which first checks their basic requirements and has performed forecasting. Forecasts from the model for the next three years has been made and the forecasts follow the recent trend in the data.

# The time series prediction has been one with 95% confidence interval.

# The chocolate production will grow marginally in coming next three years.