*University of California, Santa Barbara*


PSTAT 174 Final Project

**Time Series Analysis:**

**Monthly Hotel Occupied Room Average**

---

Group Pi - Angela Ho, Amy Li, Darian Lam, Timmy Kubota, Omar Jones

June 6, 2018

## Contents

**1 Introduction ……..…. 2**

**2 Data Exploratory Analysis ……..…….…. 3**

**3 Data Transformation ……...……**

**3.1 Box-Cox Transformation 3**

**3.2 Differencing 5**

**4 Model Building and Selection ……...……..**

**4.1 Analysis 7**

**4.2 Model Selection 7**

**5 Model Checking and Diagnostic Plots …… 7**

**6 Forecasting ……... 10**

**7 Conclusion ……... 11**

**References ……... 13**

**Appendices ……..… 13**

## Abstract

The purpose of our project is to build a time series model that can accurately forecast the monthly average number of occupied rooms in hotels based on historical data. After transforming and differencing the original data, we looked at the ACF and PACF plots to determine possible SARIMA models and used Akaike information criterion and the Ljung–Box test to find the model with the best fit. We then performed diagnostics and used this model to forecast average occupied rooms for the next ten months. We found that the predicted values are comparable to actual values.

## 1 Introduction

Hotel room occupancy are the main statistic planners used when deciding how to best meet their customers' demands. For most people, travel depends heavily on the season and major holidays, so these numbers fluctuate throughout the year. This information is used to determine many factors such as room prices and staffing needs; therefore, it is crucial for hotels to be able to accurately predict the number of occupied rooms they will service in difference months.

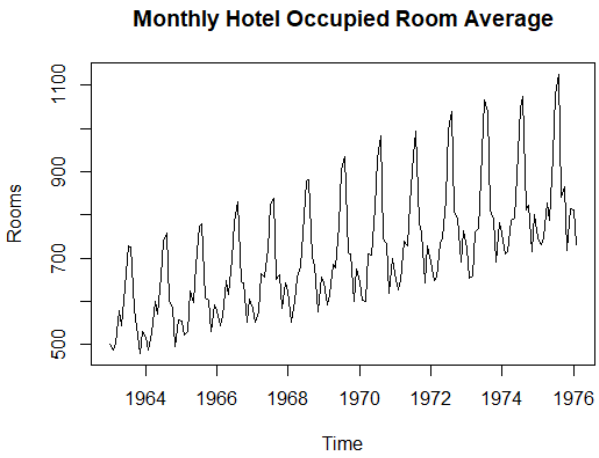
In our project, we look at data tracking the monthly hotel occupied room averages from January 1963 to December 1976, found on DataMarket. Although this is an older data set, we thought it would be interesting to look at the movements in number of occupied rooms during those years. We want to find a time series model that fits the data and will accurately predict future observations, so we use the first 158 values of the data and reserve the last 10 values to check if our model is accurate. In R, we perform a box-cox transformation to make variance constant and difference at lag 1 and lag 12 in order to detrend and deseasonalize the data. We

then look at the ACF and PACF plots to find possible SARIMA models and use AIC and Ljung-Box test to select the two best models and check their diagnostic plots. We found the best model to be *SARIMA* (3,1,3) x $(1,1,0)_{12}$. We use this model to forecast the next ten months. All forecasted values fall within a 95% confidence interval.

## 2 Exploratory Data Analysis

Our data contains 168 observations and spans the time between January 1963 and December 1976. We remove the last ten observations so that we can compare them to our final model's predicted values in order to check for accuracy.

The time series plot of the first 158 observations is on the right. The plot shows regular peaks and valleys, which indicate seasonality. This makes sense, as we would expect there to be more travelers during the summer season and major holidays. There is also increasing variance and an upward trend in occupied rooms in more recent years compared to earlier ones.




## 3 Data Transformation

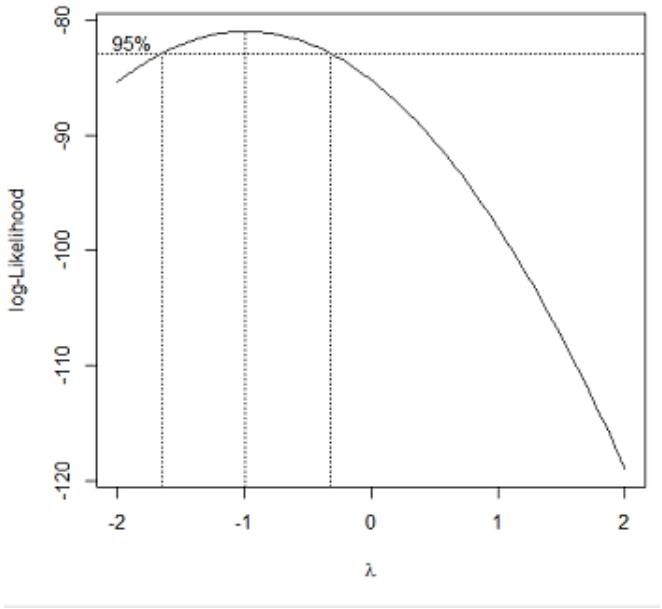
### 3.1 Box-Cox Transformation

We consider the Box-Cox Transformation as a means to stabilize the variance of our data.

$$Y = \frac{Y^{\lambda}-1}{\lambda} \quad \lambda \neq 0$$

$$Y = log\, Y \quad \lambda = 0$$

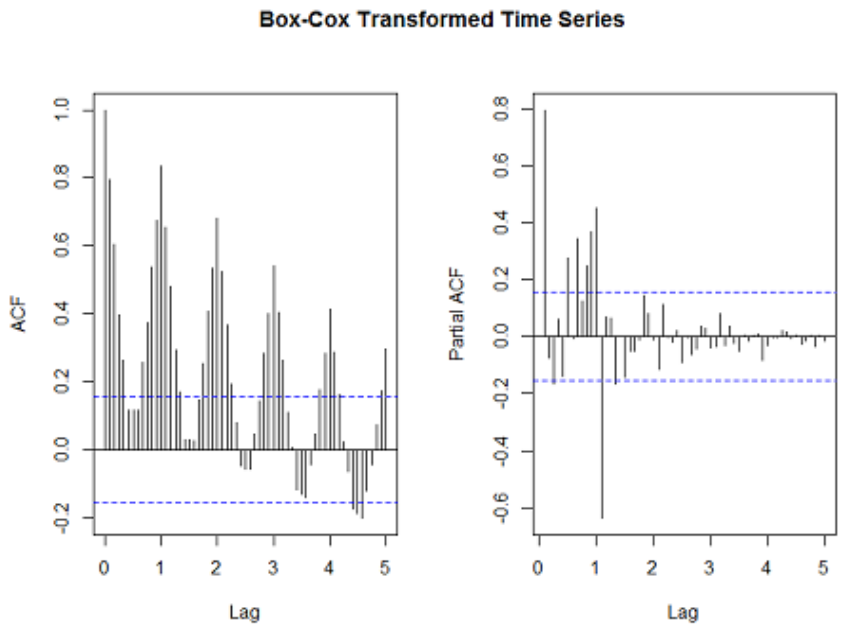
**Figure: Log-Likelihood and Box-Cox Transformation**




We want to obtain Y, the transformed data, according to the parameter, λ. Once λ is determined, the transformation can be performed based on the above equation. In the figure above, the log-likelihood is plotted vs possible λ-values. The optimal λ, which leads to the maximal log-likelihood, is around - 0.989. Since λ is not equal to 0, we use the first equation to transform the original time series.

**Figure (Right) : ACF and PACF of the transformed data, Y**.

The seasonal spikes in the ACF plot and the tailing-off of the PACF plots of the Box-Cox



transformed data indicate that the data is not stationary. Hence, it is necessary to de-trend the data and remove the seasonality. We will check that differencing is required by observing the plots of the data and compare the values of the variance before and after the differencing has been performed.

## 3.2 Differencing

We conduct differencing to remove the trend and seasonality. Define the lag d difference as:

$$\nabla \mathbf{Y}_t = Y_t - Y_{t-d}$$

where $Y_t$ is our data after performing the Box-Cox transformation, and *d* is the lag. First we perform a lag-1 difference to remove the trend in the data. The variance in the data decreases from $7.99 \times 10^{-8}$ to $3.06 \times 10^{-8}$. The plot of the data shows a linear trend, so we expect to perform a lag-1 difference only once. Indeed, differencing again at lag 1 increases the variance, so next we deal with the seasonality. From observing the plot and the recurring spikes in the ACF plot of the time series, we conclude that a differencing at lag-12 is needed to remove the seasonality. Hence, $d = 12$ and our data is now in the form $\nabla_{12} \nabla \mathbf{Y}_t$.

We can see from the ACF plot that the spikes oscillate below the significance level and the PACF plot tails off exponentially toward 0. In addition, we can compare the variance of the data at each step. After performing a lag-12 difference to remove the seasonality, the variance decreases from $3.06 \times 10^{-8}$ to $2.04 \times 10^{-9}$. These results indicate that our data is now stationary.

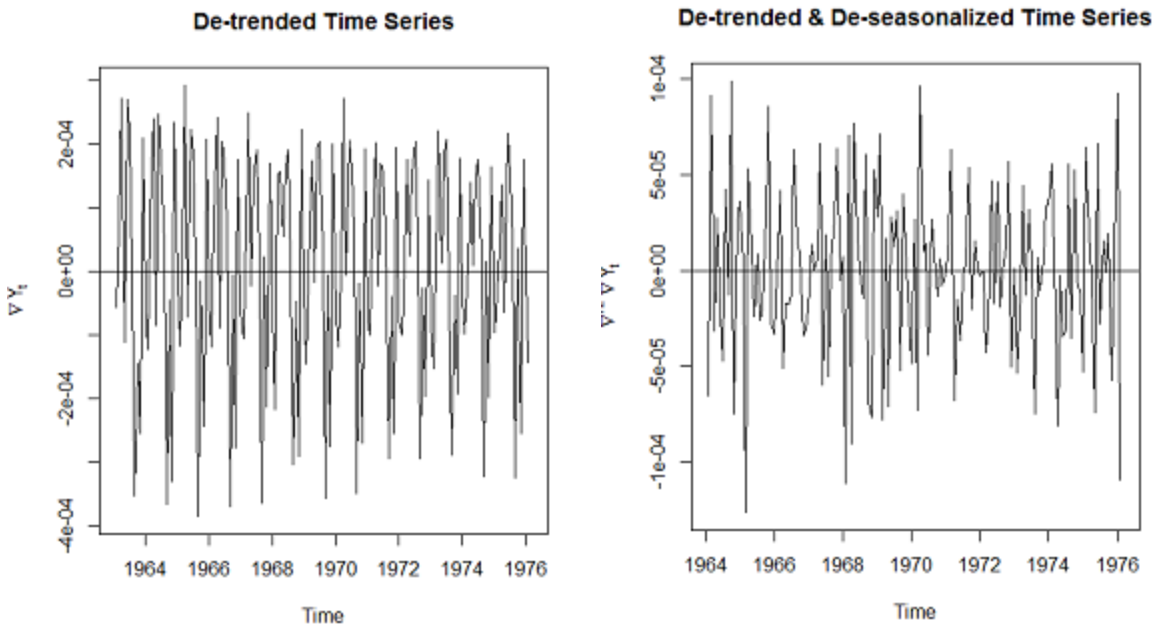



**Figure (Above) : On the left: The data after lag-1 difference; On the right: The data after lag-12 additional difference.**

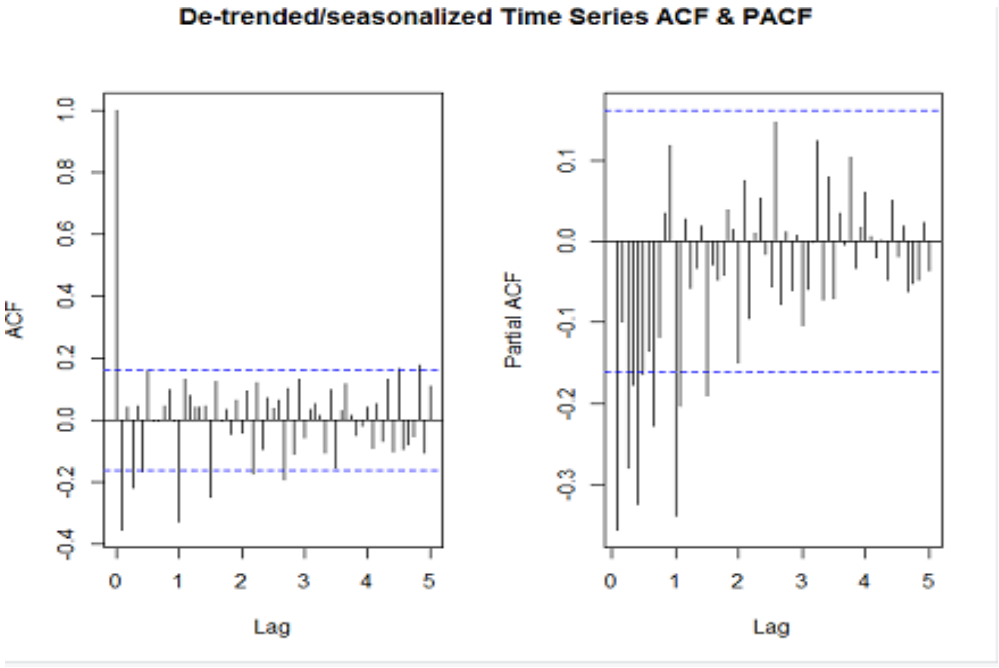



**Figure (Above) : ACF and PACF of $Y_t$ after it has been differenced to remove the linear trend and seasonality at d = 12.**

# 4 Model Building and Selection

## 4.1 Analysis

After deseasonalizing and detrending our time series data, we are given the graphs above. In these graphs, we are able to identify what our possible p, q, P and Q values. Looking at the ACF and PACF graphs, we are able to think of the following possible models. Due to the fact that it seems like for the seasonal values PACF cuts off after 1 and ACF tails off, we have P = 1 and Q=0. For the non-seasonal values, we observed p, q in a subset of 0, 1, 2, and 3, because it seems to tail off after 3. D and d are both left at 1 since we differenced it once to remove seasonality and once more to remove trend.

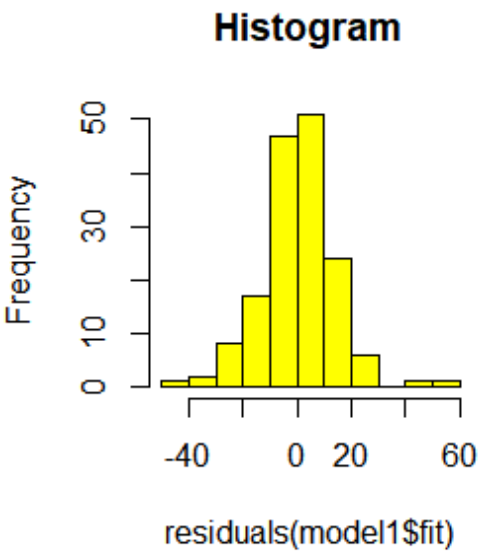
## 4.2 Model Selection

We then use AICc to figure out the tradeoff between the goodness of fit and model complexity for the best model. When accounting for AICc, we look for the smallest AIC out of our possible candidates. We created a loop that ran the different parameters p, q from 0 to 3 in a SARIMA model and extracted the AICc values. After extracting the AICc values we sorted it to find out the two smallest AICc models, thus giving us two possible SARIMA models of $(3,1,3) \times (1,1,0)_{12}$ and $(3,1,2) \times (1,1,0)_{12}$.

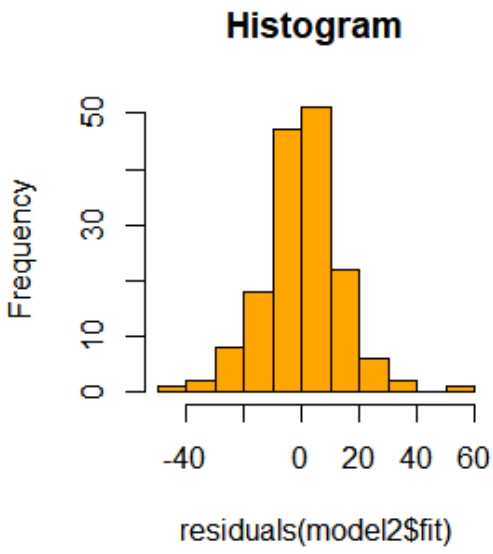
# 5 Model Checking & Diagnostic Plots

We chose a time series SARIMA model by observing its diagnostic plots. The assumptions are checked to determine whether these models are a good enough fit for us to continue onto the forecasting portion of the project. To begin the diagnostic checks, we observe the histogram of standardized residuals below. Both model 1 and model 2 histogram's frequency looks somewhat close to bell curves, although they are not perfect.

**Model 1 Histogram**

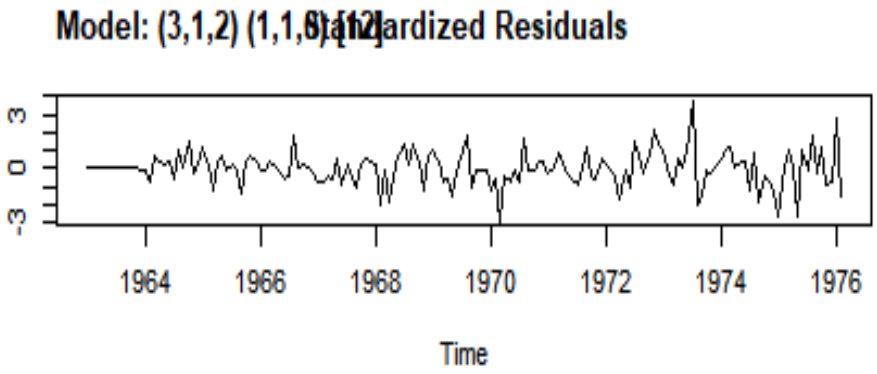


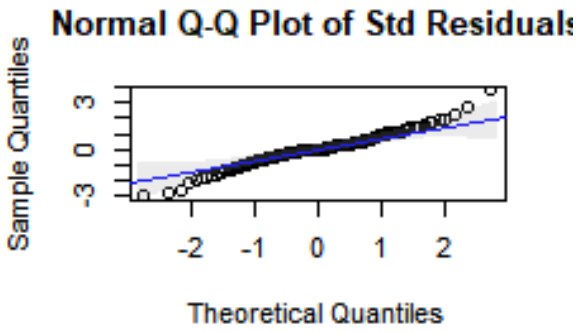

**Model 2 Histogram**



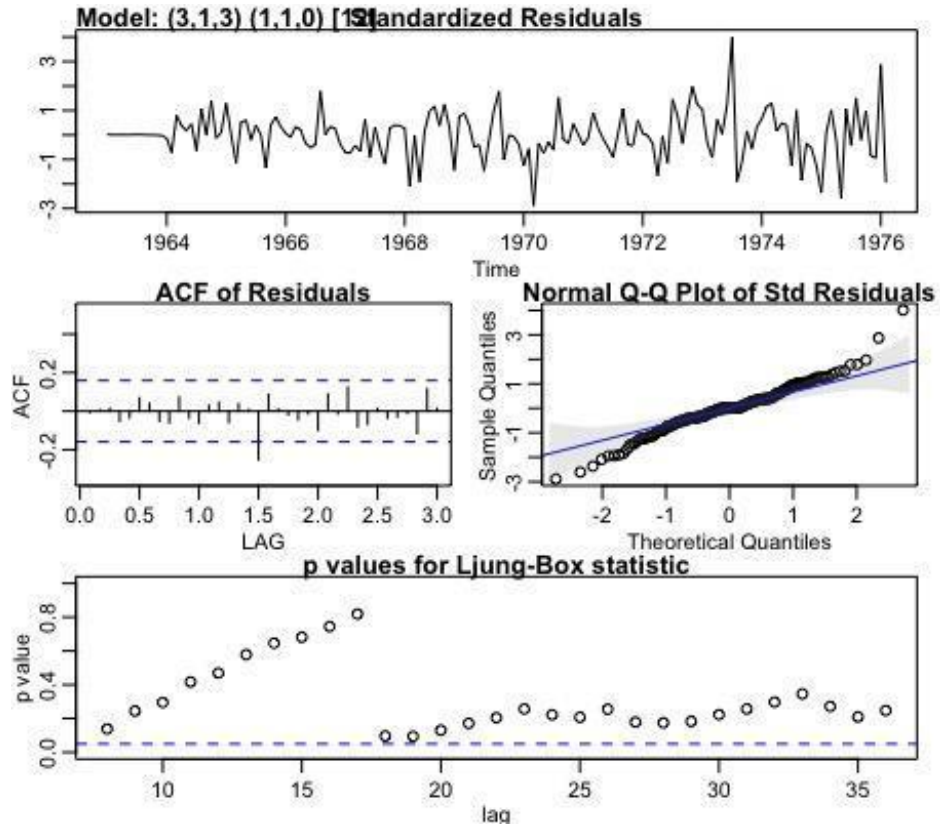

This is probably because our data set only consists of 168 points, whereas more data points will give a better approximation such as the histogram of 100,000 values of Gaussian White Noise studied in class (Notes 9). Therefore, we move on and check the next diagnostic plots for better accuracy.

**Model 2 Diagnostic Plots**






**Model 1 Diagnostic Plots :**




The standardized residuals plots for both models exhibit no seasonality or trends, indicating that its residuals are normal. In addition, all the residuals in the ACF plots reside within the 95% confidence interval, with the exception of a single point at lag at 1.5 in both models. This satisfies the assumption that the ACF values have constant errors.

The Normal QQ Plot is generally used to check the assumption that our variables are normally distributed. It consists of a scatter plot made up of two sets of quantiles: the sample quantile versus the theoretical quantile. Our plotted quantile values do not exactly lie within the confidence interval on a full straight line in both models. However, in both plots the values are close to the straight line, so we select model 1 with the smallest AIC value that passes most of the tests since its residuals are somewhat normally distributed. After fitting a SARIMA model 1 to our data, the Ljung-Box Test is also applied to check the model. The dashed line in our plot signifies the limit of the 0.05 p-value. We used the test by observing its p-values in the Ljung-Box Test Diagram above. The diagram shows that the model's p-values all appeared to be greater than 0.05, as desired. This means we should accept the null hypothesis that the time series model is a good fit to our data, and no autocorrelation exists in our residuals. For further diagnostics, we also checked the Box Pierce p-values for autocorrelation and large p-values.

**Box Pierce Test P-Values**

| | |
|---|---|
| Model 1 | 0.9706 |
| Model 2 | 0.8787 |

After performing multiple diagnostics tests, we can safely say that model 1 is a better fit for our data:

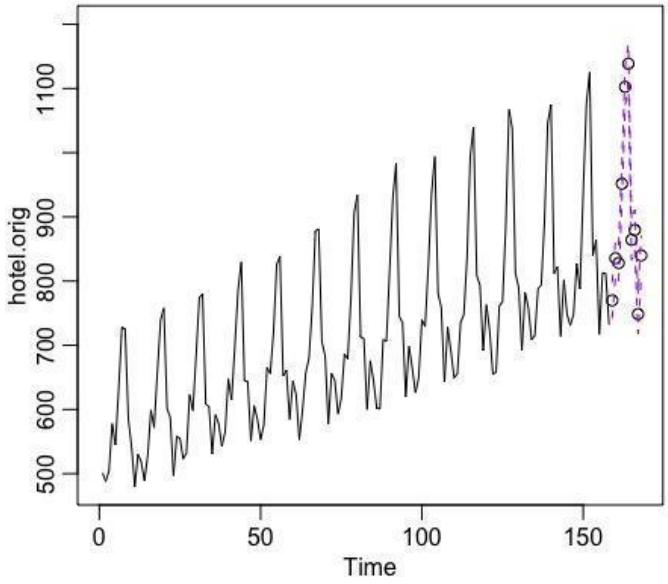
$$SARIMA\ (3,1,3)\ x\ (1,1,0)_{12}$$

$$(1 - .4886_{.2135}B - .3697_{.2208}B^2 + .4182_{.0821}B^3)\nabla_{12}\nabla X_t = (1 - 1.14233_{.2383}B + .0702_{.4690}B^2 + .3531_{.2507}B^3)\ Z_t$$

## 6 Forecasting

A goal of this project was to figure out a model which could accurately predict 10 future observations of our data. In the beginning we removed 10 data observations so we could use it to

forecast and compare our predicted values to the observed values. We are predicting the values of March of 1976 to December of 1976.




From the time series graph above, we can see the overall trend, and seasonality is present. From March to August, the data tends to increase as in the previous years, while from September to November, it decreases again. Also, in December the graph increases again, as expected. We can also see that each data point falls into a 95% confidence interval. Our predicted data points are pretty close to the actual values. For example, in May of 1976 our predicted is 828.19 compared to the actual value of 833. And in July of 1976 our predicted value is 1102.56 compared to the actual of 1110. Therefore, we are satisfied with our model since the model prediction values are fairly close to the observed values.

## 7 Conclusion

The goal of this project was to build a time series model that could explain monthly hotel occupied rooms between January of 1963 to February of 1976. Then, we need to use the model obtained to predict 10 future observations, the monthly hotel occupied rooms from March of

1976 to December of 1976. In this project, we found a seasonal component of our data of 12 months. We also saw that there was an upward trend in the data, meaning as the years progressed, the monthly hotel occupied room average also increased as well. Our final model was:

$$SARIMA\ (3,1,3)\ x\ (1,1,0)_{12}$$

With an algebraic form of:

$$(1 - .4886_{.2135}B - .3697_{.2208}B^2 + .4182_{.0821}B^3)\nabla_{12}\nabla X_t = (1 - 1.14233_{.2383}B + .0702_{.4690}B^2 + .3531_{.2507}B^3)\ Z_t$$

With this new model, we predicted the values of March 1976 to December 1976, based on the model selection and a 95% confidence interval. We determined and observed that all values fall within the interval, and are close to the actual values. Therefore, we are satisfied with this model and it is a good sign that our model is accurate.

## References

“Monthly Hotel Occupied Room Av. '63-'76 B.L.Bowerman Et Al.” *DataMarket*, datamarket.com/data/set/22no/monthly-hotel-occupied-room-av-63-76-blbowerman-et-al#!ds=22no&display=line.

## Appendices

## R Code

```
#Data
hotel.csv <- read.table("monthly-hotel-occupied-room-av-6.csv",sep=",", header=FALSE, skip=1)
#create time series object
hotel<-ts(hotel.csv[-c(159:168),2], frequency = 12, start=c(1963,1))
test<-ts(hotel.csv[c(159:168),2], frequency=12, start=c(1976,3))

#Time Series Plot
plot(hotel, main="Monthly Hotel Occupied Room Average", ylab="Rooms")
ts.plot(hotel, main="Monthly Hotel Occupied Room Average", ylab="Rooms")
#increasing trend
#seasonality
#Our variance is increasing

#box-cox, sqrt, and log transformations
library(MASS)
t = 1:length(hotel)
fit = lm(hotel ~ t)
bcTransform = boxcox(hotel ~ t,plotit = TRUE)
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
hotel.bc = (1/lambda)*(hotel^lambda-1)
op = par(mfrow = c(1,2))
ts.plot(hotel,main = "Original data",ylab = expression(X[t]))
ts.plot(hotel.bc,main = "Box-Cox tranformed data", ylab = expression(Y[t]))
#linear trend, need to difference

#box cox transformed data - ACF & PACF
op = par(mfrow = c(1,2))
acf(hotel.bc,lag.max = 60,main = "")
pacf(hotel.bc,lag.max = 60,main = "")
title("Box-Cox Transformed Time Series", line = -1, outer=TRUE)
par(op)

#difference at lag1 to remove trend component
op = par(mfrow = c(1,1))
y1 = diff(hotel.bc, 1)
plot(y1,main = "De-trended Time Series",ylab = expression(nabla~Y[t]))
abline(h = 0)
par(op)
```

```
#difference at lag12 to remove seasonal component
op = par(mfrow = c(1,1))
y12 = diff(y1, 12)
ts.plot(y12,main = "De-trended & De-seasonalized Time Series",ylab =
expression(nabla^{12}~nabla~Y[t]))
abline(h = 0)
par(op)

var(y1)
var(y12)
var(hotel.bc)
ts.plot(y1)
ts.plot(y12)

#de-trended & de-seasonalized ACF & PACF
op = par(mfrow = c(1,2))
acf(y12,lag.max = 60,main = "")
pacf(y12,lag.max = 60,main = "")
title("De-trended/seasonalized Time Series ACF & PACF",line = -1, outer=TRUE)
par(op)
#P=0, and Q=2 because PACF cuts off at 2 and ACF cuts off at 0

library(sarima)
library(astsa)
Fit = list()
AICc = matrix(nrow =16 , ncol =6)
colnames ( AICc ) =c("p" ,"q" ,"P" ,"Q" ,"AICc","LjungBox")
i =0
for (P in c(1) ) {
  for (Q in c(0)) {
    for (p in c(0, 1, 2, 3)){
      for (q in c (0, 1, 2, 3)){
        Fit [[i+1]]= sarima (hotel.bc, p ,1 ,q , P ,1 ,Q ,12 , Model = TRUE , details = FALSE ) $fit
        plot.new()
        AICc [i+1,1]= p
        AICc [i+1,2]= q
        AICc [i+1,3]= P
        AICc [i+1,4]= Q
        AICc [i+1,5]= sarima (hotel.bc,p,1,q ,P ,1 ,Q ,12 , Model = TRUE , details = FALSE ) $AICc
        plot.new()
        AICc [i+1,6]= Box.test(resid(Fit[[i+1]]), type = c("Ljung-Box"), lag = 12)$p.value
        i=i+1
      }
    }
  }
}
AICc <- data.frame ( AICc )
AICc
AICc.sorted<-AICc[order(AICc$AICc),] #sorted by AIC in increasing order
AICc.sorted

#model1 diagnostics plots for residuals
sarima(hotel, 3,1,3,1,1,0,12)
```

```
#model2 diagnostics plots for residuals
sarima(hotel, 3,1,2,1,1,0,12)

model1<-sarima(hotel, 3,1,3,1,1,0,12)
model2<-sarima(hotel, 3,1,2,1,1,0,12)
hist(residuals(model1$fit), main = 'Histogram', col = 'yellow')
hist(residuals(model2$fit), main = 'Histogram', col = ' orange')
Box.test(residuals(model1$fit), lag=12, type='Ljung-Box')
Box.test(residuals(model2$fit), lag=12, type='Ljung-Box')
#our best model with the smallest AICc value is p=3, q=3, P=1, Q=0
#model here is our final new model
model<-sarima(hotel, 3,1,3,1,1,0,12)
#now we forecast and predict our next 10 data points
op=par(mfrow=c(1,1))
mypred<-predict(model$fit, n.ahead=10)
hotel.orig<-ts(hotel.csv)
ts.plot(hotel.orig, xlim=c(0, 168), ylim=c(min(hotel),1200))
points(159:168, mypred$pred)
lines(159:168, mypred$pred+1.96*mypred$se,lty=2, col = 'purple')
lines(159:168, mypred$pred-1.96*mypred$se,lty=2, col = 'purple')
```