Midterm Two Project

Erica Wong (SID:24302634) and Bryana Gutierrez (SID:24504003)

April 18, 2017

In this project we received data and were tasked with coming up with the best predictions for the next 104 data points. Since our data are all time series, we took the time series approach. For the purposes of this report, we focused on the first data set Q1. Our first step was to load and format the data so that we may work with it. Then we checked to make sure that the data didn't have any missing values. Q1 did not contain any missing values so then we proceeded to plot the data.

From the plot of the original data, Figure 1, we can see that the variance of the data increases with time. Also, given the slight upward trend, we considered that the data might need to be differenced. To deal with the increase in variance we took the log of the data. Since there are negative values in the data, we added a constant of 1.5 which would later be subtracted when we calculated the predictions. The log transformed data looks as follows:

This helped take care of the variance that increases with time. Now it almost looks like the data decreases in variance with time. There is still a slight upward trend in the data, to take care of this we differenced the data. This is how the differenced data looked like.

This looked fairly stationary. So we moved on to look at what the ACF and PACF could tell us about the remaining data.

We can see that there is some strong AR presence from the large PACF values and also some MA presence from some high values in the ACF plot. The ACF plot would imply that the model may be MA(2) because the lags die off after those points. The PACF plot shows that our model may also have some AR(4) component because after those lags the lags become insignificant. To help give us an initial guess at the ARIMA distribution, we used auto.arima. Since an ARIMA model takes differencing into consideration, we used the data prior to differencing for the auto.arima function. R gave us the following output:

```r
#running auto.arima to get a better idea of what the model is
auto.arima(ts2)
```

```
## Series: ts2
## ARIMA(4,1,2)
##
## Coefficients:
##           ar1      ar2      ar3      ar4     ma1     ma2
##       -0.4732  -1.0731  -0.2677  -0.1538  0.2798  0.9215
## s.e.   0.0778   0.0626   0.0477   0.0471  0.0685  0.0391
##
## sigma^2 estimated as 0.054:  log likelihood=23.99
## AIC=-33.98   AICc=-33.76   BIC=-4.15
```

We used this information to help give us a better idea about what our model is and then moved on to conduct some model selection.

One method of model selection is Cross Validation. First we wrote a function that will run the cross validation process for us. The output of the function will be the vector of mean-squared errors (MSE) from the different cross validation groupings from the same model. To compare the MSEs from each of the groups, we took the average MSEs from each of the models we tested.

Before actually conducting cross validation, we first had to come up with some models. We knew that in order for our data to look stationary, we must take 1 difference at one point. Additionally, from the PACF and ACF plot, we can see that there are AR(4) and MA(2) components in our model. So, we started off by

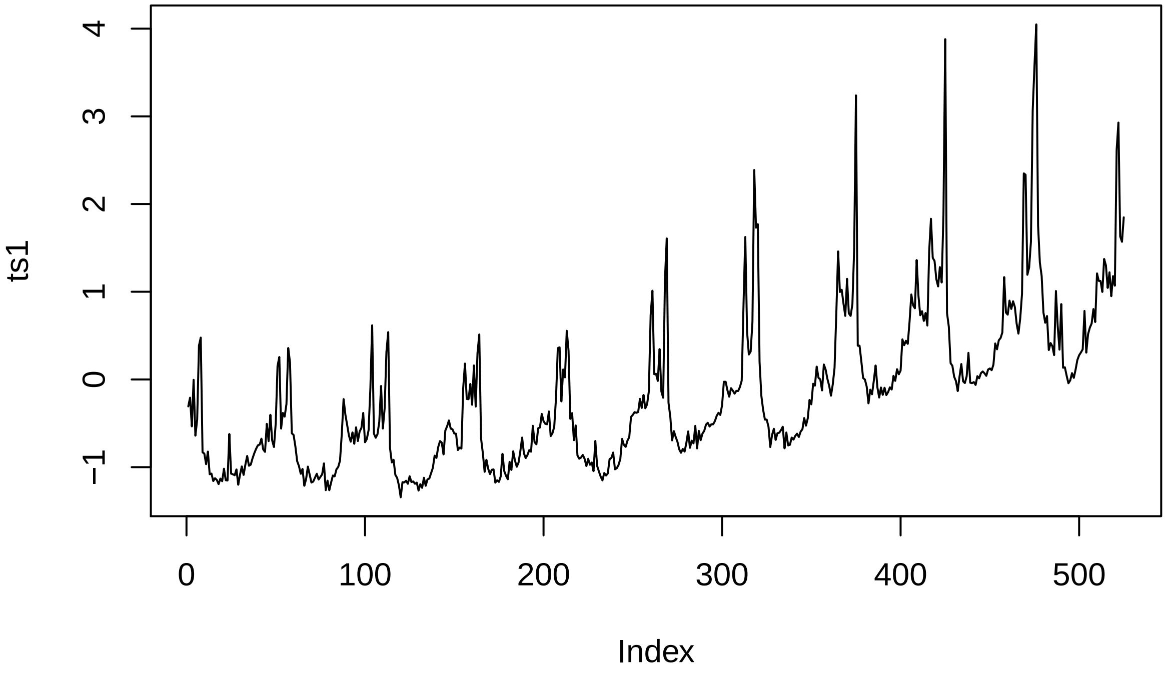

Figure 1: Plot of Original Data

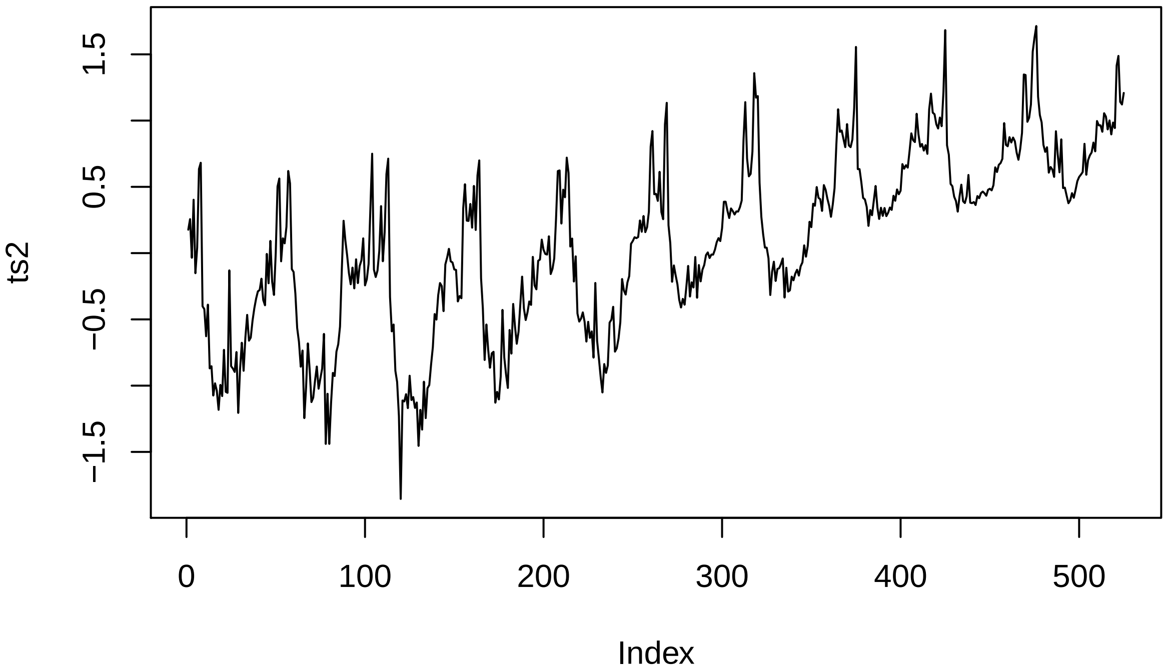

Figure 2: Plot of Log Data Shifted Up 1.5

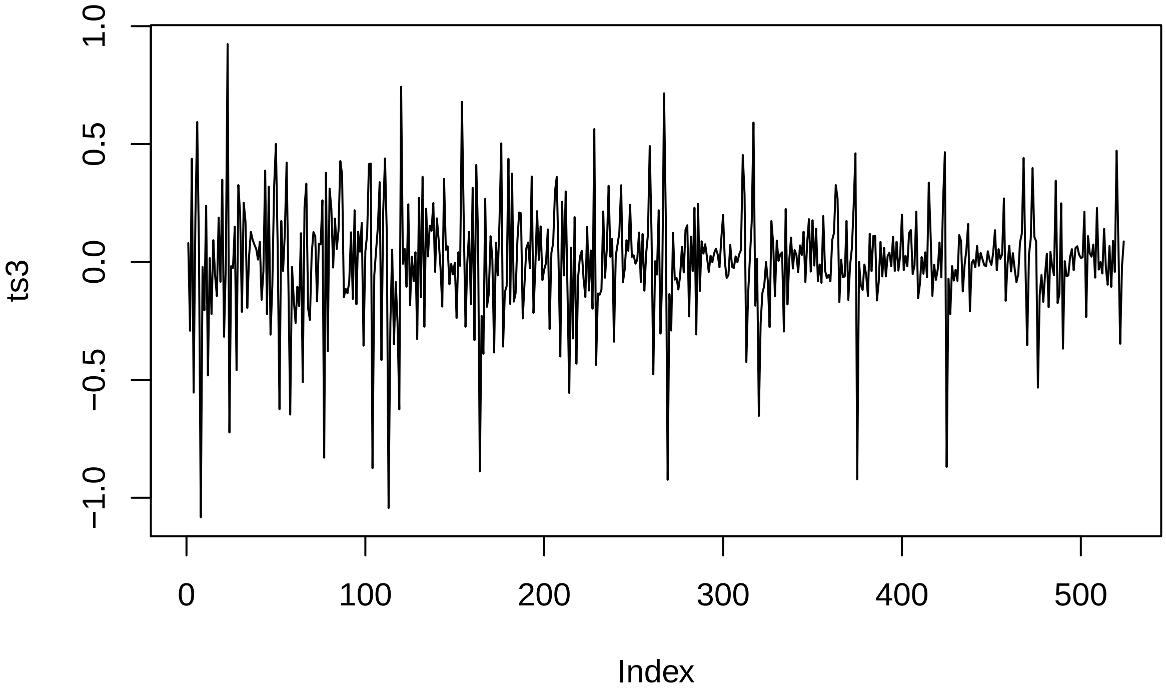

Figure 3: Plot of the Differenced Log Data

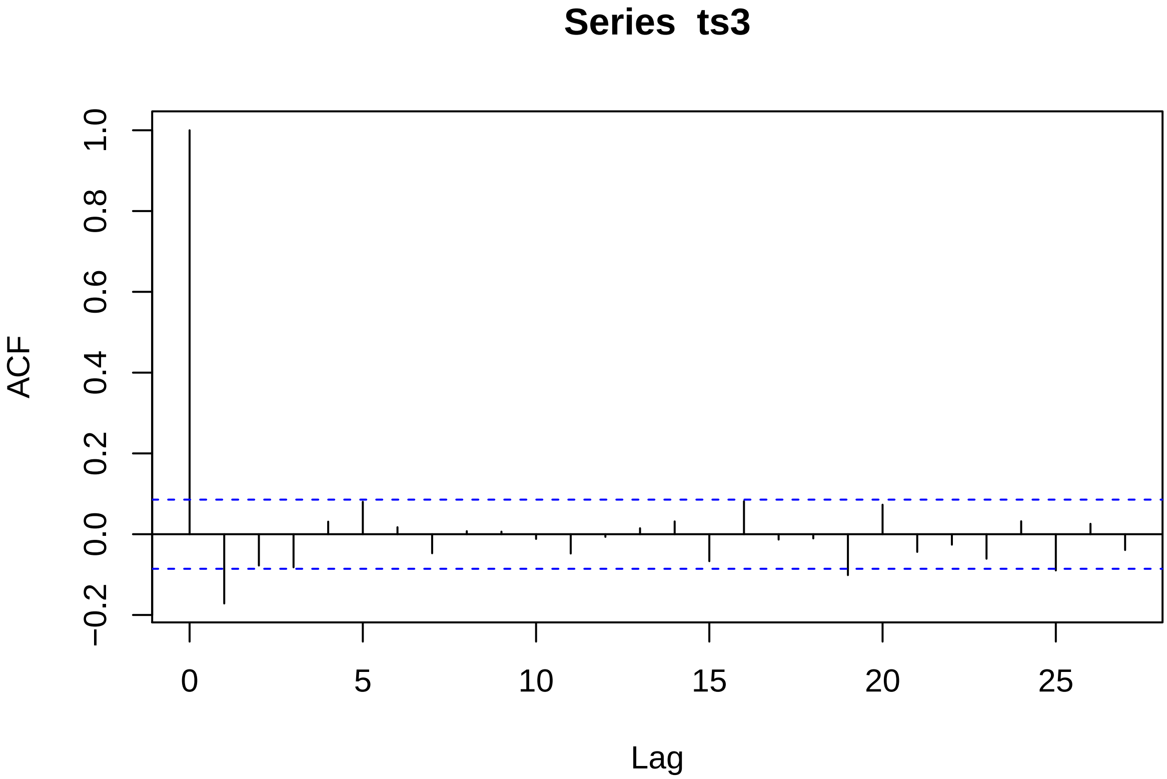

Figure 4: ACF of Log Transformed Data

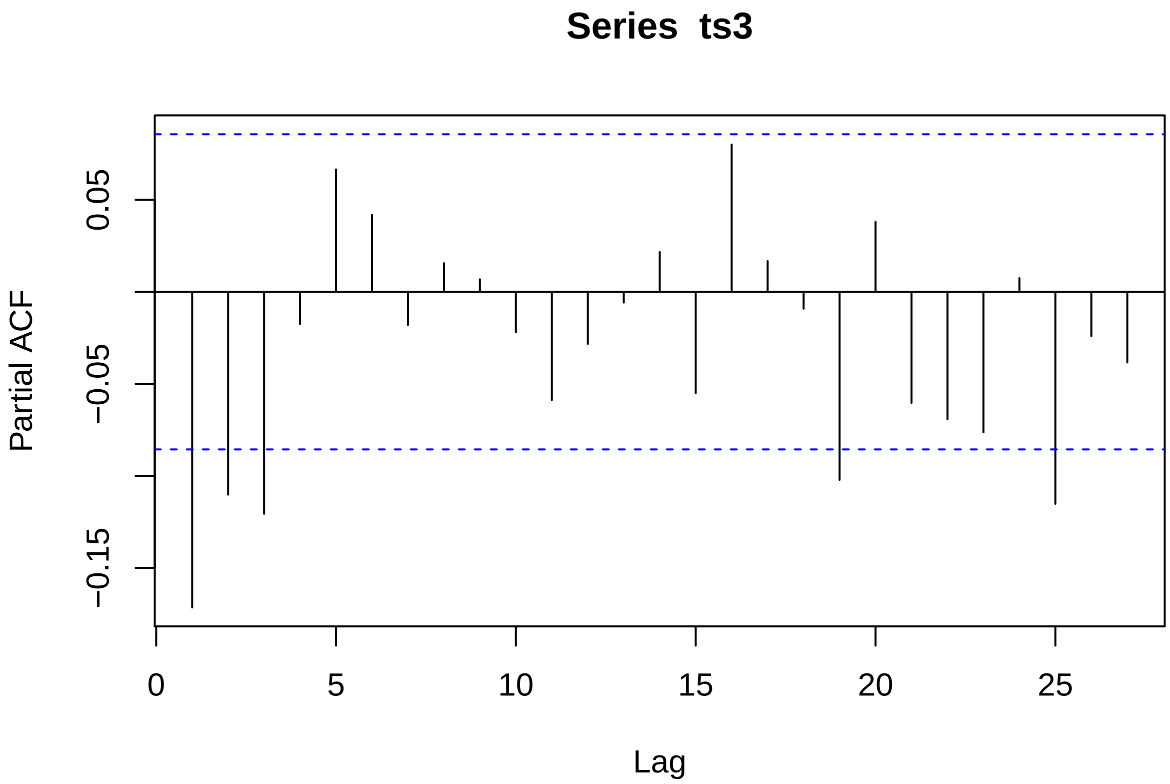

Figure 5: PACF of Log Transformed Data

thinking we should test an ARIMA(4,1,2) model. This model also happen to match what auto.arima() gave us. However, we also noticed from the original data (Figure 1) that there is some seasonality that exists. Since, this is data that is collected every week over years, we believed that the seasonality may be lag 52 because there are 52 weeks in a year. So, that is how we created our original model. We then conducted cross validation with models that were similar to what our original prediction was.

After coming up with various models and running cross validation, the best model in our case was from model 1 and the second best was model 6. In table 1, we can see a comparison of the MSEs.

	Average MSE
Model 1	2.5975
Model 2	2.8855
Model 3	4.2923
Model 4	2.7160
Model 5	2.6129
Model 6	2.6119

Table 1: MSE Table

After conducting cross validation, we still wanted to which model other model selection methods would tell us to pick. So, we conducted AIC and BIC. We knew that for AIC and BIC model selection, we wanted to pick the model with the lowest value. Tables 2 and 3 give us the values of AIC and BIC respectively. The best models for both are model 6, followed by model 4, and then model 1. However, when generating model 4, we did get some errors, so we believed that it would be best to eliminate that model from being considered.

From all three of our model selection processes, we get model 6 and model 1 as the best. So to help make our final decision, we decided to graph both of them to see which one we liked better or believed fit in our model better.

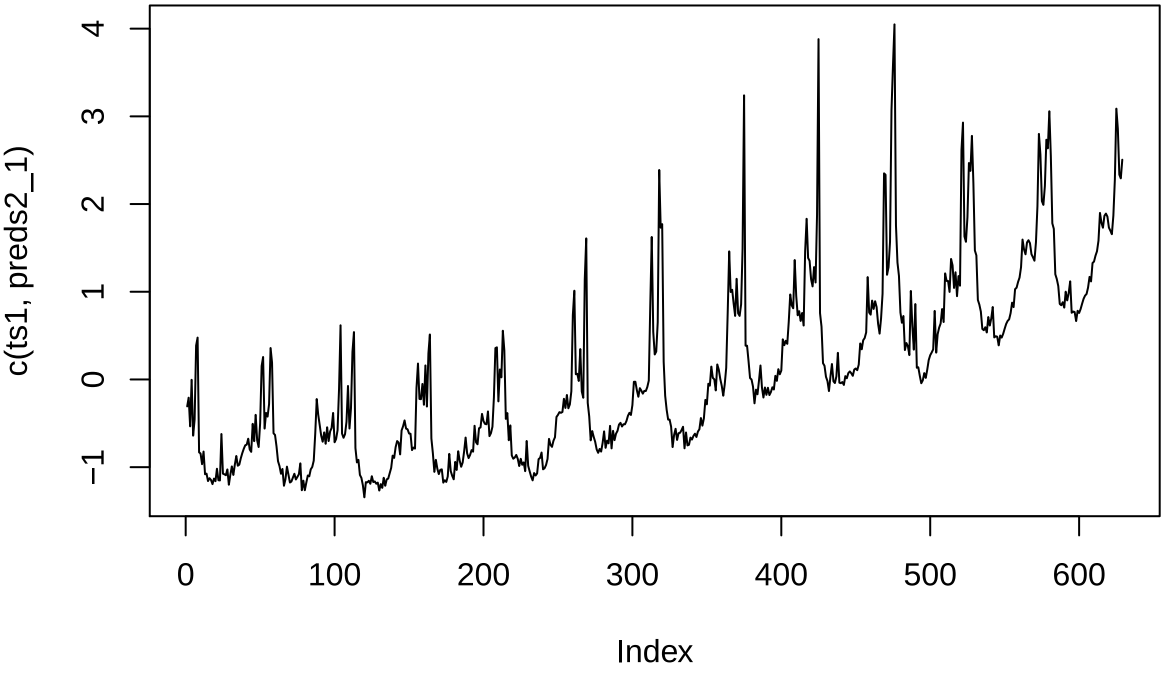

Figure 6: Model 1 Plot

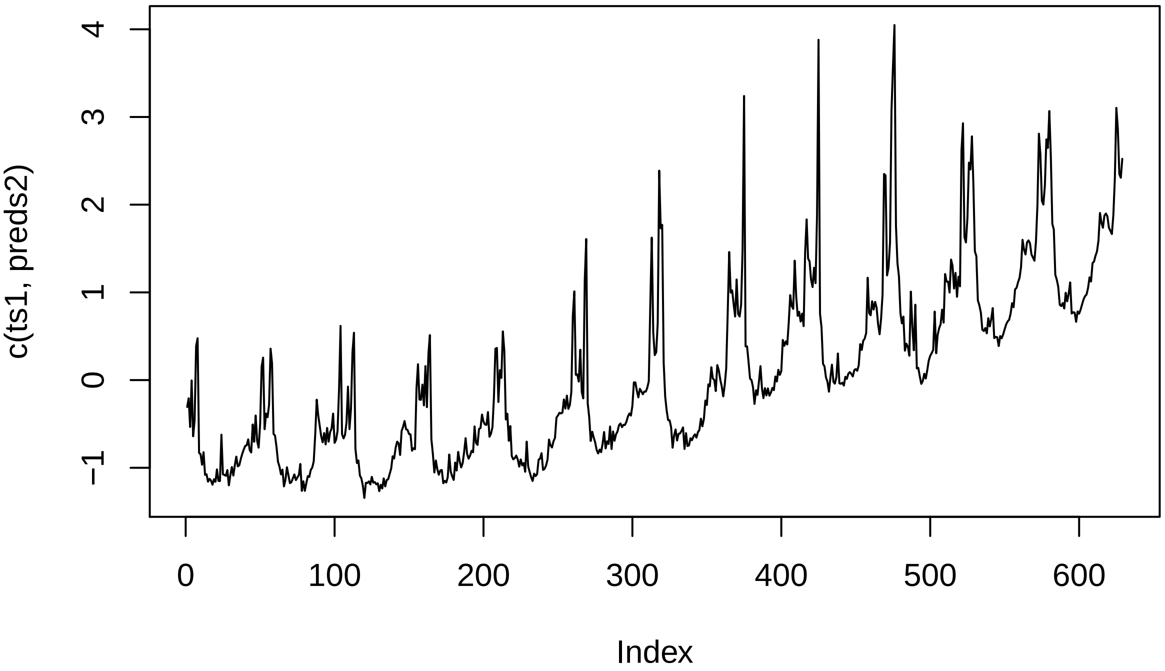

Figure 7: Model 6 Plot

	AIC Value
Model 1	-115.2144
Model 2	-107.8405
Model 3	-60.0934
Model 4	-121.1718
Model 5	-116.2298
Model 6	-117.2269

Table 2: AIC Table

	BIC Value
Model 1	-76.8609
Model 2	-69.5042
Model 3	-26.0015
Model 4	-78.5569
Model 5	-73.6149
Model 6	-83.1349

Table 3: BIC Table

The plots only vary slightly, so we decided to go with model 6 because overall we believed that it performed better in model selection, ranking consistently in the top two best models. So, the final model that we went with was arima(ts2, order = c(4,1,1), seasonal = list(order = c(1, 0, 1), period = 52)).

Apendix

Q1 Code

```r
#Loading the Data:
data1 <- read.csv("q1_train.csv", as.is = TRUE)
plot(data1$activity, type = 'l')
data1_df <- data.frame(ts(data1))
data1_df$Date <- 1:nrow(data1)

#Cleaning the Data:
ts1 <- data1$activity

#Taking the log to get rid of the trend
plot(log(ts1+ 1.5), type = 'l')
ts2 <- log(ts1+ 1.5)

#running auto.arima to get a better idea of what the model is
auto.arima(ts2)

#Doing Different Transformations to get a better idea of the wanted model
#took difference twice so data looked reasonably stationary
ts3 <- diff(ts2)
plot(ts3, type = "l")

ts4 <- diff(ts3)
plot(ts4, type = "l")
```

```r
#plotted acf
acf(ts3)
pacf(ts3)

acf(ts4) #MA(1) leading
pacf(ts4)

#Cross-Validation as a Function:
CV <- function(test_order, test_seasonality, test_period){
  mse <- NULL
  leng <- length(ts2)
  for (i in 4:1){
    train <- ts2[1:(leng - i*105)]
    test <- ts2[(leng - i*105 + 1):(leng - (i-1)*105)]
    mod_train <- arima(ts2, order = test_order, seasonal = list(order = test_seasonality,
                                                                period= test_period))
    forcast <- predict(mod_train, n.ahead = 105)
    mse[i] <- mean((exp(forcast$pred) - exp(test))^2)
  }
  return(mse)
}

#Running Cross-Validation on different models: The best model we found via auto.arima was ARIMA(4,1,2)
#but from our graph we can see that there is some seasonality, so we wanted to test that
mse1 <- CV(c(4,1,2),c(1, 0, 1),52)
mse2 <- CV(c(4,2,2),c(1, 0, 1),52)
mse3 <- CV(c(4,1,2),c(0, 0, 1),52)
mse4 <- CV(c(5,1,2),c(1, 0, 1),52)
mse5 <- CV(c(4,1,3),c(1, 0, 1),52)
mse6 <- CV(c(4,1,1),c(1, 0, 1),52)

sum(mse1)/4
sum(mse2)/4
sum(mse3)/4
sum(mse4)/4
sum(mse5)/4
sum(mse6)/4

#the smallest MSE is mse1, so the model choosen is CV(c(4,1,2),c(1, 0, 1),52), models 4 and 5
#produced errors which may imply that it is not the best idea to use them

#Let us check AIC and BIC to see what model is chosen
mod_test1 <- arima(ts2, order = c(4,1,2),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test2 <- arima(ts2, order = c(4,2,2),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test3 <- arima(ts2, order = c(4,1,2),
                   seasonal = list(order = c(0, 0, 1), period = 52))
mod_test4 <- arima(ts2, order = c(5,1,2),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test5 <- arima(ts2, order = c(4,1,3),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test6 <- arima(ts2, order = c(4,1,1),
                    seasonal = list(order = c(1, 0, 1), period = 52))

AIC(mod_test1)
AIC(mod_test2)
AIC(mod_test3)
AIC(mod_test4)
AIC(mod_test5)
AIC(mod_test6)

BIC(mod_test1)
BIC(mod_test2)
BIC(mod_test3)
BIC(mod_test4)
BIC(mod_test5)
BIC(mod_test6)

#Both AIC and BIC pick mod_test6 as the best model, which is
#arima(ts2, order = c(4,1,1), seasonal = list(order = c(1, 0, 1), period = 52))

#forcasting
preds <- predict(mod_test6, n.ahead = 104)
preds2 <- exp(preds$pred) - 1.5
plot(c(ts1, preds2), type = "l")

preds_1 <- predict(mod_test1, n.ahead = 104)
preds2_1 <- exp(preds_1$pred) - 1.5
plot(c(ts1, preds2_1), type = "l")

#there was not a difference between the graphs, but we ended up choosing model 1 because we found that
#the AIC and BIC were fairly close to that of model 6

write.table(preds2, sep = ",", col.names = FALSE,
            row.names = FALSE, file = "Q1_Bryana_Gutierrez_24504003.txt")
```

Q2 Code

```
#Loading the Data:
data2 <- read.csv("q2_train.csv", as.is = TRUE)
plot(data2$activity, type = 'l')
data2_df <- data.frame(ts(data2))
data2_df$Date <- 1:nrow(data2)

#Cleaning the Data:
ts2 <- data2$activity
min(ts2) # the minimum is about -1.5 and since I want to be able to take the log to get
#rid of the trend, I am going to add in 1.5 to shift the data upward and make it postive

#Taking the log to get rid of the trend
plot(log(ts2+ 1.5), type = 'l')
ts2_adj <- log(ts2+ 1.5)

#using auto.arima to see if there is some good fit
```

```
auto.arima(ts2_adj)
which.min(acf(ts2_adj, lag.max = 100)$acf[-1])

#Cross-Validation as a Function:
CV <- function(test_order, test_seasonality, test_period){
  mse <- NULL
  leng <- length(ts2_adj)
  for (i in 4:1){
    train <- ts2_adj[1:(leng - i*105)]
    test <- ts2_adj[(leng - i*105 + 1):(leng - (i-1)*105)]
    mod_train <- arima(ts2_adj, order = test_order, seasonal = list(order = test_seasonality,
                                                                    period= test_period))
    forcast <- predict(mod_train, n.ahead = 105)
    mse[i] <- mean((exp(forcast$pred) - exp(test))^2)
  }
  return(mse)
}

mse1 <- CV(c(2,1,1),c(1, 0, 1),52)
mse2 <- CV(c(2,1,0),c(1, 0, 1),52)
mse3 <- CV(c(2,1,2),c(0, 0, 0),NA)
mse4 <- CV(c(2,1,1),c(0, 0, 0),NA)
mse5 <- CV(c(3,1,2),c(0, 0, 0),NA)
mse6 <- CV(c(1,1,2),c(1, 0, 1),52)

sum(mse1)/4
sum(mse2)/4
sum(mse3)/4
sum(mse4)/4
sum(mse5)/4
sum(mse6)/4

mod_test1 <- arima(ts2_adj, order = c(2,1,1),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test2 <- arima(ts2_adj, order = c(2,1,0),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test3 <- arima(ts2_adj, order = c(2,1,2))
mod_test4 <- arima(ts2_adj, order = c(2,1,1))
mod_test5 <- arima(ts2_adj, order = c(3,1,2))
mod_test6 <- arima(ts2_adj, order = c(1,1,2),
                   seasonal = list(order = c(1, 0, 1), period = 52))

AIC(mod_test1)
AIC(mod_test2)
AIC(mod_test3)
AIC(mod_test4)
AIC(mod_test5)
AIC(mod_test6)

BIC(mod_test1)
BIC(mod_test2)
BIC(mod_test3)
BIC(mod_test4)
BIC(mod_test5)
BIC(mod_test6)

#forcasting both of these models look the same, i think the second one looks slightly better
preds_1 <- predict(mod_test1, n.ahead = 104)
preds2_1 <- exp(preds_1$pred) - 1.5
plot(c(ts2, preds2_1), type = "l")

write.table(preds2, sep = ",", col.names = FALSE,
            row.names = FALSE, file = "Q2_Bryana_Gutierrez_24504003.txt")
```

Q3 Code

```
#Loading the Data:
data3 <- read.csv("q3_train.csv", as.is = TRUE)
plot(data3$activity, type = 'l')
data3_df <- data.frame(ts(data3))
data3_df$Date <- 1:nrow(data3)

#Cleaning the Data:
ts3 <- data3$activity
min(ts3) # the minimum is about -1.5 and since I want to be able to take the log to get
#rid of the trend, I am going to add in 1.5 to shift the data upward and make it postive

#Taking the sqrt to get rid of the trend, it is slightly parabolic which is why I chose sqrt
plot(sqrt(ts3+ 1.5), type = 'l')
ts3_adj <- sqrt(ts3+ 1.5)

#using auto.arima to see if there is some good fit
auto.arima(ts3_adj)
which.max(acf(ts3_adj, lag.max = 100)$acf[-1])

#Cross-Validation as a Function:
CV <- function(test_order, test_seasonality, test_period){
  mse <- NULL
  leng <- length(ts3_adj)
  for (i in 4:1){
    train <- ts3_adj[1:(leng - i*105)]
    test <- ts3_adj[(leng - i*105 + 1):(leng - (i-1)*105)]
    mod_train <- arima(ts3_adj, order = test_order, seasonal = list(order = test_seasonality,
                                                                    period= test_period))
    forcast <- predict(mod_train, n.ahead = 105)
    mse[i] <- mean((exp(forcast$pred) - exp(test))^2)
  }
  return(mse)
}

mse1 <- CV(c(3,1,3),c(1, 0, 1),52)
#mse2 <- CV(c(3,0,3),c(1, 0, 1),52)
mse3 <- CV(c(3,1,3),c(0, 0, 0),NA)
mse4 <- CV(c(2,1,1),c(0, 0, 0),NA)
mse5 <- CV(c(3,1,2),c(0, 0, 0),NA)
```

```r
mse6 <- CV(c(1,1,2),c(1, 0, 1),52)

sum(mse1)/4
#sum(mse2)/4
sum(mse3)/4
sum(mse4)/4
sum(mse5)/4
sum(mse6)/4

mod_test1 <- arima(ts3_adj, order = c(3,1,3),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test2 <- arima(ts2_adj, order = c(3,1,3),
                   seasonal = list(order = c(1, 0, 2), period = 52))
mod_test3 <- arima(ts3_adj, order = c(3,1,3),
                   seasonal = list(order = c(0, 0, 1), period = 52))
mod_test4 <- arima(ts3_adj, order = c(2,1,1),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test5 <- arima(ts3_adj, order = c(3,1,2))
mod_test6 <- arima(ts3_adj, order = c(1,1,2),
                   seasonal = list(order = c(1, 0, 1), period = 52))

AIC(mod_test1)
AIC(mod_test2)
AIC(mod_test3)
AIC(mod_test4)
AIC(mod_test5)
AIC(mod_test6)

BIC(mod_test1)
BIC(mod_test2)
BIC(mod_test3)
BIC(mod_test4)
BIC(mod_test5)
BIC(mod_test6)

#forcasting both of these models look the same, i think the second one looks slightly better
preds <- predict(mod_test6, n.ahead = 104)
preds2 <- (preds$pred)^2 - 1.5
plot(c(ts3, preds2), type = "l")

write.table(preds2, sep = ",", col.names = FALSE,
            row.names = FALSE, file = "Q3_Bryana_Gutierrez_24504003.txt")
```

Q4 Code

```r
#Loading the Data:
data4 <- read.csv("q4_train.csv", as.is = TRUE)
plot(data4$activity, type = 'l')
data4_df <- data.frame(ts(data4))
data4_df$Date <- 1:nrow(data4)

#Cleaning up data
```

```r
ts_4 <- data4_df$activity

#data appears to be made stationary by differencing
plot(diff(ts_4), type = "l")

#also helps to apply a log transformation to the data
log_ts4 <- log(ts_4 + 1.5)

#testing out the log data
auto.arima(log_ts4)

#plotting acf and pacf to see what kind of model the resulting differenced
#data follows
acf(diff(ts_4), lag.max = 100)
pacf(diff(ts_4), lag.max = 100)

acf(diff(log_ts4), lag.max = 100)
pacf((log_ts4), lag.max = 100)


#MA(1), AR(2) with the diffferenced data
#3,1,4

auto.arima(ts_4)
auto.arima(log_ts4)

#Cross validation function
CV4 <- function(test_order, test_seasonality, test_period){
  mse <- NULL
  leng <- length(log_ts4)
  for (i in 4:1){
    train <- log_ts4[1:(leng - i*105)]
    test <- log_ts4[(leng - i*105 + 1):(leng - (i-1)*105)]
    mod_train <- arima(log_ts4, order = test_order, seasonal = list(order = test_seasonality,
                                                                    period= test_period))
    forcast <- predict(mod_train, n.ahead = 105)
    mse[i] <- mean((exp(forcast$pred) - exp(test))^2)
  }
  return(mse)
}

#Running Cross-Validation on different models
mse1 <- CV4(c(3,1,1),c(1, 0, 1),52)
mse2 <- CV4(c(3,1,4),c(1, 0, 1),52)
mse3 <- CV4(c(3,1,4),c(0, 0, 1),52)
mse4 <- CV4(c(3,1,1),c(0, 0, 1),52)
mse5 <- CV4(c(0,1,1),c(1, 0, 1),52)
mse6 <- CV4(c(0,1,1),c(1, 0, 1),52)


sum(mse1)/4
sum(mse2)/4
sum(mse3)/4
sum(mse4)/4
sum(mse5)/4
sum(mse6)/4


#Let us check AIC and BIC to see what model is chosen
mod_test1 <- arima(log_ts4, order = c(3,1,1),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test2 <- arima(log_ts4, order = c(3,1,4),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test3 <- arima(log_ts4, order = c(3,1,4),
                   seasonal = list(order = c(1, 1, 1), period = 52))
mod_test4 <- arima(log_ts4, order = c(3,1,1),
                   seasonal = list(order = c(0, 0, 1), period = 52))
mod_test5 <- arima(log_ts4, order = c(0,1,1),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test6 <- arima(log_ts4, order = c(0,1,1),
                   seasonal = list(order = c(0, 0, 1), period = 52))


AIC(mod_test1)
AIC(mod_test2)
AIC(mod_test3)
AIC(mod_test4)
AIC(mod_test5)
AIC(mod_test6)


BIC(mod_test1)
BIC(mod_test2)
BIC(mod_test3)
BIC(mod_test4)
BIC(mod_test5)
BIC(mod_test6)

#The second model is the best

preds <- predict(mod_test2, n.ahead = 104)
preds2 <- exp(preds$pred) - 1.5
plot(c(ts_4, preds2), type = "l")


write.table(preds2, sep = ",", col.names = FALSE,
            row.names = FALSE, file = "Q4_Bryana_Gutierrez_24504003.txt")
```

Q5 Code

```
#Loading the Data:
data5 <- read.csv("q5_train.csv", as.is = TRUE)
plot(data5$activity, type = 'l')
data5_df <- data.frame(ts(data5))
data5_df$Date <- 1:nrow(data5)
```

```
#Writing data in much more accessible way
ts_5 <- data5_df$activity

#EDA
plot(log(log(ts_5 + 1.5)+1.5), type = "l")
acf(diff(ts_5), lag.max = 100)
pacf(diff(ts_5))

#transforing the data
log_ts5 <- log(ts_5 + 1.5)

#auto arima functions
auto.arima(ts_5)
#3,1,5

auto.arima(log(ts_5 +1.5))
#2,1,1

CV5 <- function(test_order, test_seasonality, test_period){
  mse <- NULL
  leng <- length(log_ts5)
  for (i in 4:1){
    train <- log_ts5[1:(leng - i*105)]
    test <- log_ts5[(leng - i*105 + 1):(leng - (i-1)*105)]
    mod_train <- arima(log_ts5, order = test_order, seasonal = list(order = test_seasonality,
                                                                    period= test_period))
    forcast <- predict(mod_train, n.ahead = 105)
    mse[i] <- mean((exp(forcast$pred) - exp(test))^2)
  }
  return(mse)
}

#Running Cross-Validation on different models
mse1 <- CV4(c(3,1,5),c(1, 0, 1),52)
mse2 <- CV4(c(3,1,5),c(1, 1, 1),52)
mse3 <- CV4(c(2,1,1),c(1, 0, 1),52)
mse4 <- CV4(c(2,1,1),c(1, 1, 1),52)
mse5 <- CV4(c(0,1,1),c(1, 0, 1),52)
mse6 <- CV4(c(0,1,1),c(1, 1, 1),52)


sum(mse1)/4
sum(mse2)/4
sum(mse3)/4
sum(mse4)/4
sum(mse5)/4
sum(mse6)/4

#Let us check AIC and BIC to see what model is chosen
mod_test1 <- arima(log_ts5, order = c(3,1,5),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test2 <- arima(log_ts5, order = c(3,1,5),
                   seasonal = list(order = c(1, 1, 1), period = 52))
```

```
mod_test3 <- arima(log_ts5, order = c(2,1,1),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test4 <- arima(log_ts5, order = c(2,1,1),
                   seasonal = list(order = c(1, 1, 1), period = 52))
mod_test5 <- arima(log_ts5, order = c(0,1,1),
                   seasonal = list(order = c(1, 0, 1), period = 52))
mod_test6 <- arima(log_ts5, order = c(0,1,1),
                   seasonal = list(order = c(1, 1, 1), period = 52))


AIC(mod_test1)
AIC(mod_test2)
AIC(mod_test3)
AIC(mod_test4)
AIC(mod_test5)
AIC(mod_test6)


BIC(mod_test1)
BIC(mod_test2)
BIC(mod_test3)
BIC(mod_test4)
BIC(mod_test5)
BIC(mod_test6)

#The second model is the best

preds <- predict(mod_test1, n.ahead = 104)
preds2 <- exp(preds$pred) - 1.5
plot(c(ts_5, preds2), type = "l")

write.table(preds2, sep = ",", col.names = FALSE,
            row.names = FALSE, file = "Q5_Bryana_Gutierrez_24504003.txt")
```