MineThatData Forecasting Challenge
Time Series Analysis
Zlatka Staykova, PhD and Bill Bass, lesser degrees than a PhD
SabinoDB

INTRODUCTION:

We analyzed 80 months of Total Sales from an anonymous retailer to forecast the ensuing three months of sales. For the extra credit competition, we then analyzed and forecast each of Total Sales' 15 component variables.

We tested both time series analysis and regression splines on the data using R. Because the data 1) showed a seasonal component, 2) the extra credit competition required the forecast months to incorporate previous month forecasts and 3) the final knot on the regression spline was at month 61 thereby eliminating prior months' information from the forecast, we landed on time series as the analysis tool of choice.

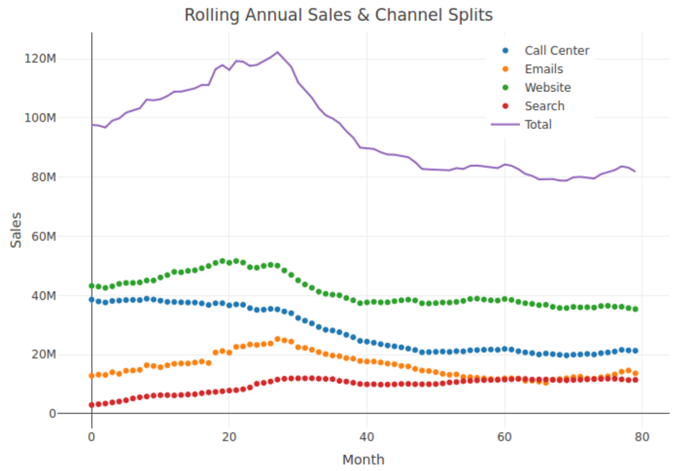
However, we believe that a shift in email strategy in the last few months of the data might be overly skewing the time series forecast – the autoregressive and moving average components of the time series model concentrate on these problematic months. We have included the regression spline model in this paper which accounts for, but minimizes the impact of, the problematic months. We then applied a "wisdom of the crowds" approach by averaging the forecast of the time series model and the splines model for our point forecasts. If we are allowed more than one submission, we would submit the average, the time series, and the splines in that order.

RESULTS

Data Exploration

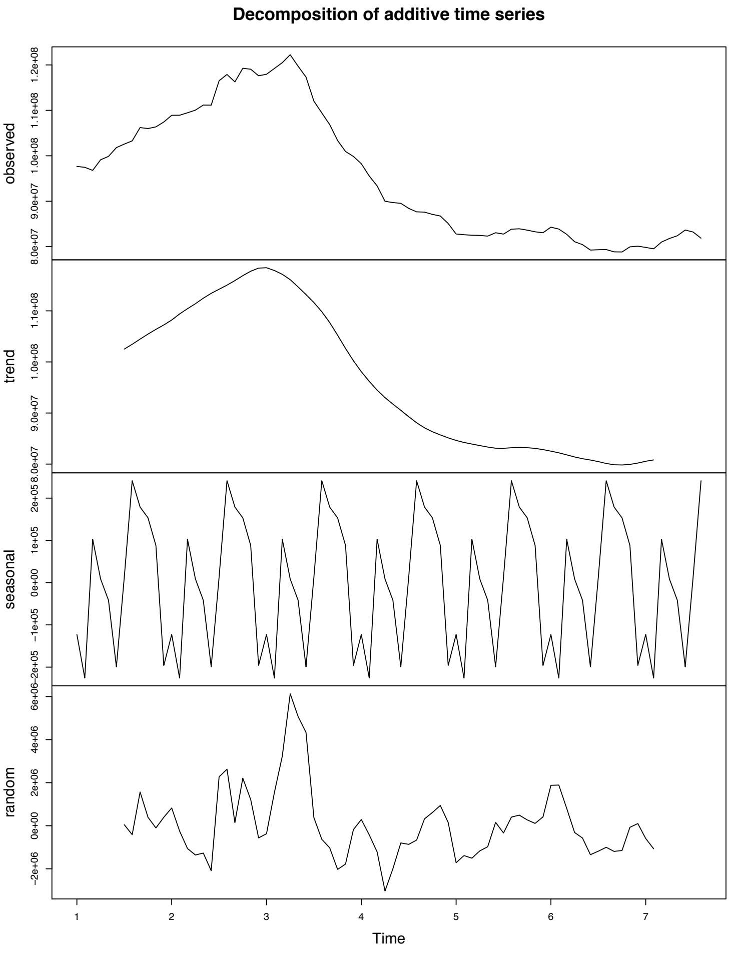
The dataset contains 80 observations (82 including two repeats of month 41) representing monthly rolling 12 month Total Sales and 15 component variables for each observation.

There are no missing variables in the dataset, however there are three identical entries for month 41. We removed the two extra observations.



The adjacent chart shows the trailing 12 month sales for each of the 80 months and the channel components that comprise the total. We noted that the bump around month 78 resulted solely from the email channel. It appears by month 80 the company had reversed whatever strategy had driven email sales higher.

This represented a conundrum. Call center, website, and search strategies appear consistent and forecasting three months into the future based on past behavior is a reasonable assumption. The email strategy appears to be in flux. Harder to forecast without knowing whether the email strategy has reverted to pre-month 78 status, or if the decline in email sales is the result of a new, failed strategy and the decline will continue further.



There is no obvious seasonality in the total sales chart above, but we ran the following decomposition in R that shows a clear seasonal component.

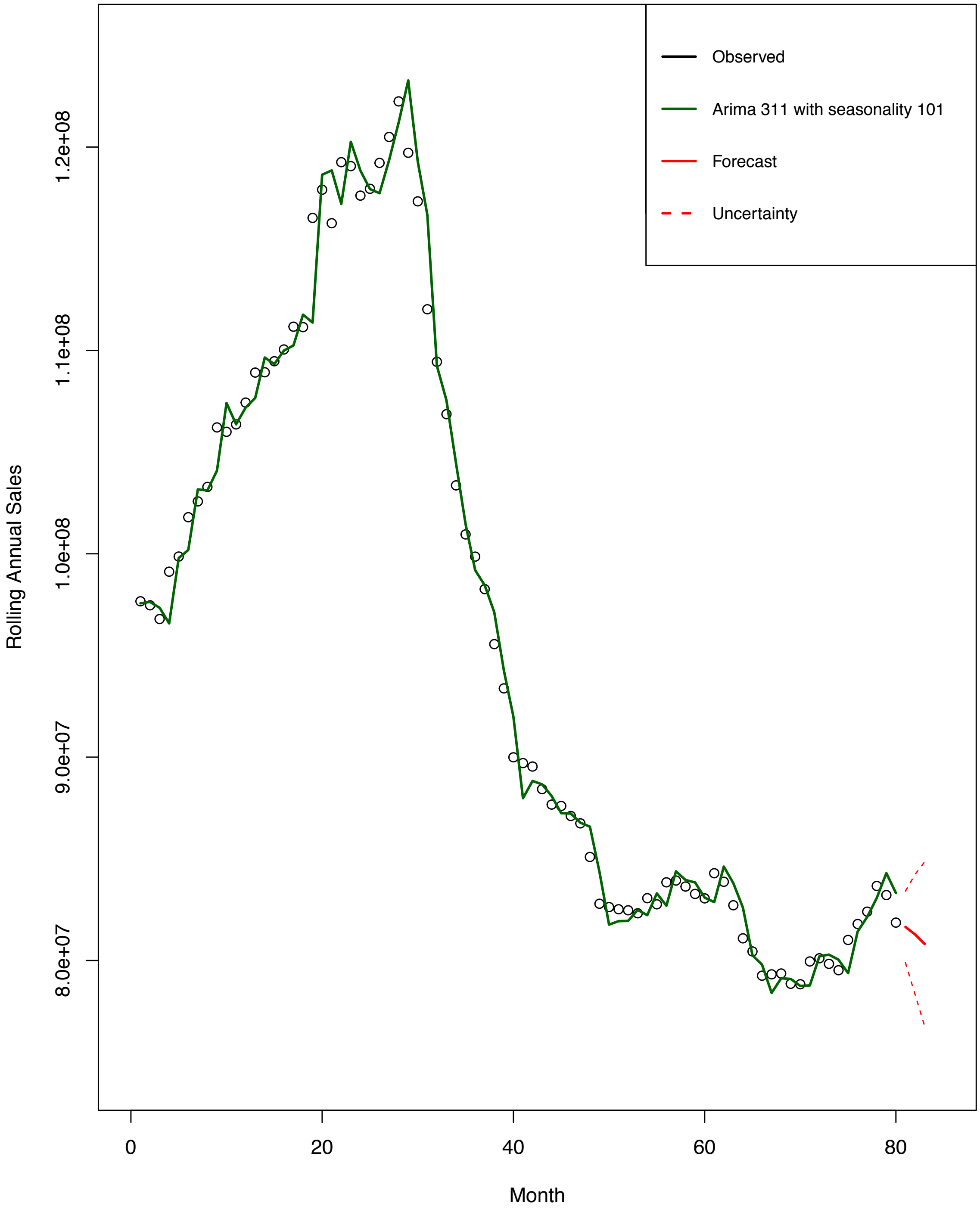
This chart shows the observed Total Sales in the top box. The y-axis displays time in years rather than months. The third box breaks out the seasonal component of sales that recurs on an annual basis. In this case, there is a clear recurring seasonal effect. The second box labeled trend shows the remaining sales trend after removing the seasonal effects. Looked good the first few years. Not so good since year three, although stabilized in the last couple of years. The final box shows the remaining random fluctuations in sales after accounting for trend and seasonality.

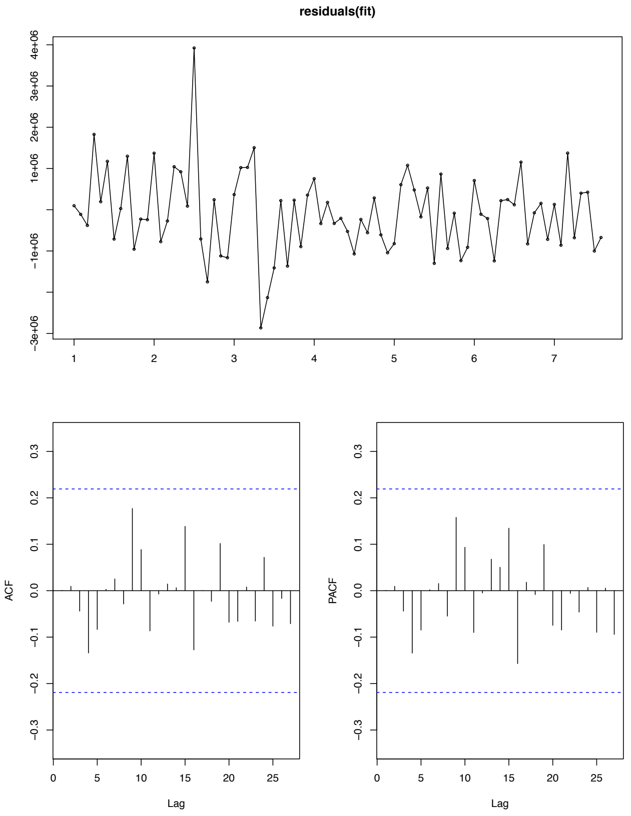
Data Preparation

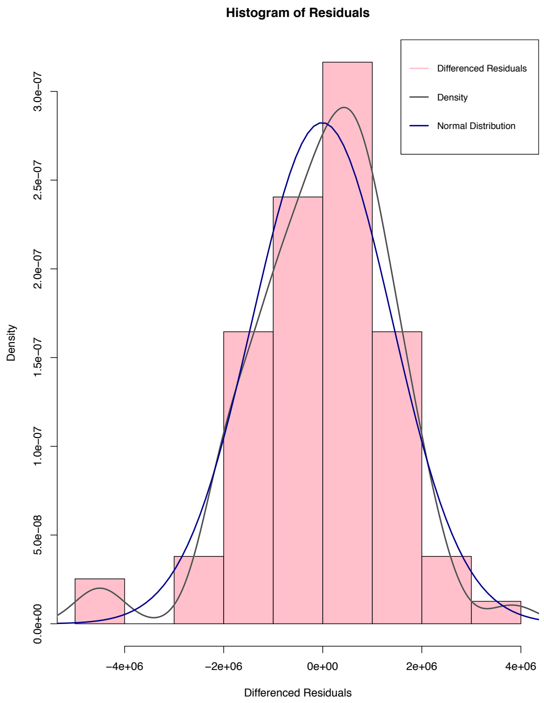
None. As detailed below, the time series data is not stationary and requires differencing but this is accomplished through the ARIMA model rather than by transforming the raw data.

The Model

We conducted ADF and KSS tests which indicated the data were not stationary (time series analysis assumes stationary data). We then ran NDIFFS and NSDIFFS which showed one degree of differencing resolved the stationary problem and no seasonal differencing was required. Auto.arima confirmed this when it generated an optimum model of ARIMA(3,1,1)(1,0,1) – the second number in each bracket shows the level of differencing. The first bracket shows the non-seasonal portion of the model with 3 autoregressive and 1 moving average factors. The second bracket shows the seasonal portion with 1 autoregressive and 1 moving average factors. This model produced the following graph with the observations as black circles, the fitted model as the green line and the red line as the three-month forecast.



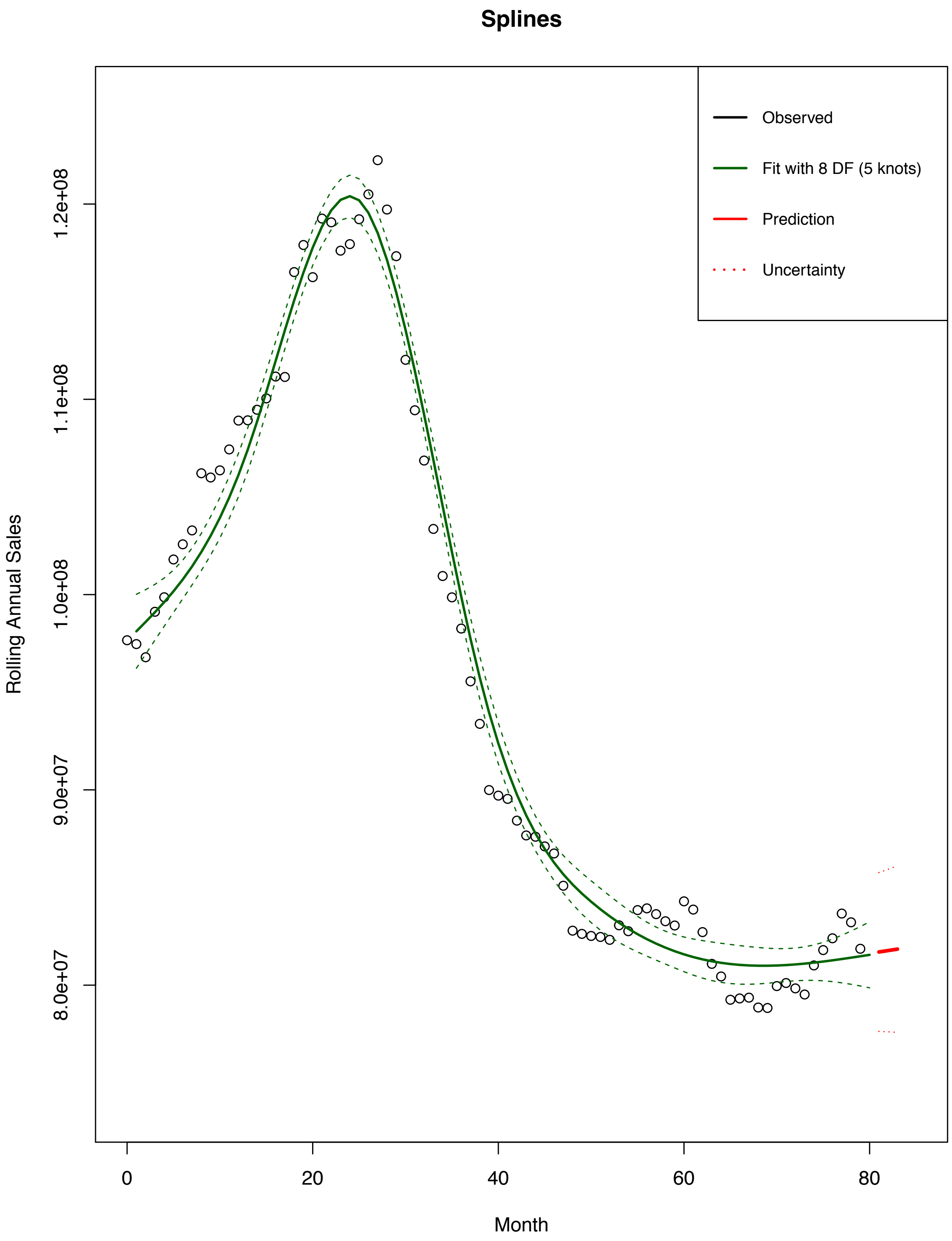




The two charts above show the distribution of the model's residuals. The graph on the left shows randomly distributed residuals with the ACF and PACF charts showing all variations within the 95% bands. The histogram on the right shows the residuals follow a roughly normal distribution.

For the extra credit portion of the contest, we applied this same time series methodology to each of the 15 remaining component variables.

However, for Total Sales, we believe that the time series model could be overly penalizing the forecast because of the last three-month performance trend that is impacted by the email channel. We tested a smoothing spline with leave one out cross-validation to determine the number of knots but the resulting 30+ knot model appeared to over fit the data. We settled on a 5-knot natural spline as the best balance of fit and potential predictive value. The following chart shows the results of this model:


Splines
Observed
Fit with 8 DF (5 knots)
Prediction
Uncertainty
Rolling Annual Sales
1.2e+08
1.1e+08
1.0e+08
9.0e+07
8.0e+07
0
20
40
60
80
Month

CONCLUSION

We suspect splines might be overly optimistic in the short term, but probably trend right in the longer term. Not clear from the contest rules if we can submit multiple entries, but if so, we would submit the average of the point forecasts between time series and splines as our first-choice entry, followed by time series as our second choice, with splines as the third choice for a short term, three-month forecast.

If we can only submit one forecast, we choose the average of time series and splines.

For the extra credit, we are only using the time series forecasts because of the requirement to base each month's forecast on the previous predicted months.

Time Series		
Month		Rolling_12Month_Sales
	80	81,651,702
	81	81,281,656
	82	80,814,374
Splines		
	80	81,698,973
	81	81,772,261
	82	81,845,550
Average		
	80	81,675,338
	81	81,526,959
	82	81,329,962

CODE

TIME SERIES:

```
library(timeSeries)
library(forecast)

root.Dir <- as.character('/home/zlatusha/Projects/SabinoDB/Mine That Data/toCompete')
setwd(root.Dir)
```

```
input.data <- read.csv(paste0(root.Dir, '/Data/MineThatData_ForecastChallenge_20170914.csv'))

summary(input.data)
names(input.data)

rolling12_ts <- ts(input.data[,2], frequency = 12)
rts_decompose <- decompose(rolling12_ts)

plot(rts_decompose)

fit <- auto.arima(rolling12_ts, seasonal = TRUE)
fit
tsdisplay(residuals(fit))

fit_manual <- Arima(input.data[,2], order = c(3,1,1), seasonal = c(1,0,1))
tsdisplay(residuals(fit_manual))

fore_fit <- forecast(fit_manual, h = 3)

plot(input.data[,2],
    xlim = c(0,85),
    xlab = 'Month',
    ylab = 'Rolling Annual Sales',
    ylim = c(74645373, 125000000)
)

lines(fore_fit$fitted, col = 'dark green', lwd = 2)
lines(rownames(as.data.frame(fore_fit)),
     as.data.frame(fore_fit)[,1],
     lwd = 2, col = 'red')

lines(rownames(as.data.frame(fore_fit)),
     as.data.frame(fore_fit)[,2], lty = 'dashed', col = 'red')

lines(rownames(as.data.frame(fore_fit)),
     as.data.frame(fore_fit)[,3], lty = 'dashed', col = 'red')


legend ("topright", legend = c('Observed', "Arima 311 with seasonality 101", 'Forecast',
'Uncertainty') ,
       col = c('black', 'dark green', 'red','red'), lty = c(1,1,1,2), lwd =2, cex =.8)
```

```
###
hist(diff(residuals(fit)),
    prob=T, col = 'pink',
    xlab = 'Differenced Residuals',
    main = 'Histogram of Residuals')

lines(density(diff(residuals(fit))), lwd=2, col = grey(0.3))

mu<-mean(diff(residuals(fit)))
sigma<-sd(diff(residuals(fit)))
x <- seq(-6000000,6000000,length=100)
y <- dnorm(x,mu,sigma)
lines(x,y,lwd=2,col="dark blue")

legend ("topright", legend = c('Differenced Residuals', "Density", 'Normal Distribution') ,
      col = c('pink', grey(0.3), 'dark blue'), lwd =2, cex =.8)


## Individual Contribution
result <- data.frame(Month=c(80:82))
result <- cbind(result, as.data.frame(fore_fit)[,1])
colnames(result)[2] <- colnames(input.data)[2]

for (i in 3:ncol(input.data)) {
  this.ts <- ts(input.data[,i], frequency = 12)
  dThis.ts <- decompose(this.ts)
  #plot(dThis.ts)

  au <- auto.arima(this.ts, seasonal = TRUE)
  this.arima <- Arima(input.data[,i], order = arimaorder(au)[1:3], seasonal = arimaorder(au)[4:6])

  this.forecast <- forecast(this.arima, h=3)

  result <- cbind(result, as.data.frame(this.forecast)[,1])
  colnames(result)[i] <- colnames(input.data)[i]
}

write.csv(result, 'result.csv')
```

SPLINES:

```
library(splines)
library(ISLR)
```

```
## PLACE THE PATH OF YOUR WORKING DIRECTORY HERE!!!!
root.Dir <- as.character('/home/zlatusha/Projects/SabinoDB/Mine That Data/toCompete')
setwd(root.Dir)

input.data <- read.csv(paste0(root.Dir, '/Data/MineThatData_ForecastChallenge_20170914.csv'))

summary(input.data)
names(input.data)

attr(bs(input.data[,1], df = 8), 'knots')

fit =lm(Rolling_12Month_Sales ~ ns(Month , knots =c(12, 24, 37, 49, 61)), data = input.data)
pred <- predict(fit, se=T, type='response')

plot(input.data[,1], input.data[,2], col = 'black',
    xlim = c(0, 85), ylim = c(74000000,125000000),
    xlab = 'Month',
    ylab = 'Rolling Annual Sales')
title('Splines')
lines(pred$fit, lwd = 2, col = 'dark green')
lines(pred$fit+2*pred$se, lty = 'dashed', col = 'dark green')
lines(pred$fit-2*pred$se, lty = 'dashed', col = 'dark green')

pred.result <- predict(fit, se = T, newdata = data.frame(Month=c(81:83)), interval = "prediction",
type = 'response')

lines(c(81:83), as.data.frame(pred.result$fit[,1])[,1], lwd = 3, col = 'red')

lines(c(81:83), as.data.frame(pred.result$fit[,2])[,1], lty = 3, lwd = 1, col = 'red')
lines(c(81:83), as.data.frame(pred.result$fit[,3])[,1], lty = 3, lwd = 1, col = 'red')

legend ("topright", legend = c('Observed', "Fit with 8 DF (5 knots)", 'Prediction', 'Uncertainty') ,
      col = c('black', 'dark green', 'red','red'), lty = c(1,1,1,3), lwd =2, cex =.8)
```