ECON 144 Project 1

BO LIN (304542813) MANYEE YU (304339557) ZIWEI ZHU (504614045)

April 27, 2017

I. Introduction

Due to the recent incident that took place on the United Airline, which has caused much discussion, our group has decided to look into the revenue of the company, and would like to project the future trend. In addition, we are also interested in learning the difference between the predicted value for United Airline's future quarterly revenue, especially the second quarter of 2017, and the actual value that they will be announcing. Such difference can then be used in assessing how the incident affects the company's revenue.

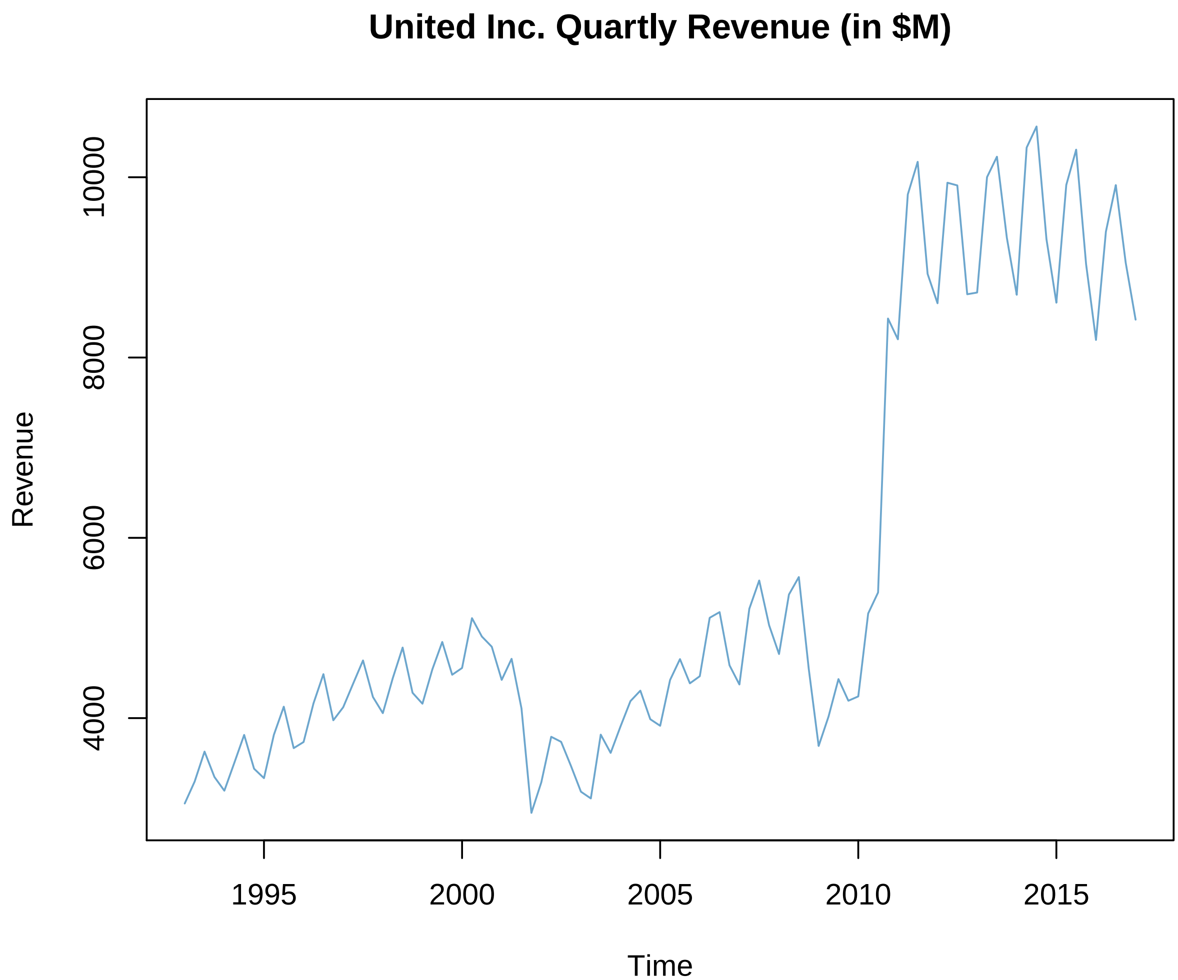
The data that we are currently using is the United Continental Holdings Inc quarterly revenue (hereafter referred as United Inc., in the future content) starting from the First quarter in the year 1993 to the first quarter in the year 2017.

Though there was no file that combined all the necessary data together, we were able to obtain all the data from their published annual financial report, and we manually entered the corresponding numbers into the excel file. One of the team members performed the entries, and the other team member double confirmed that the numbers are correct.

II. Results

1. Modeling and Forecasting

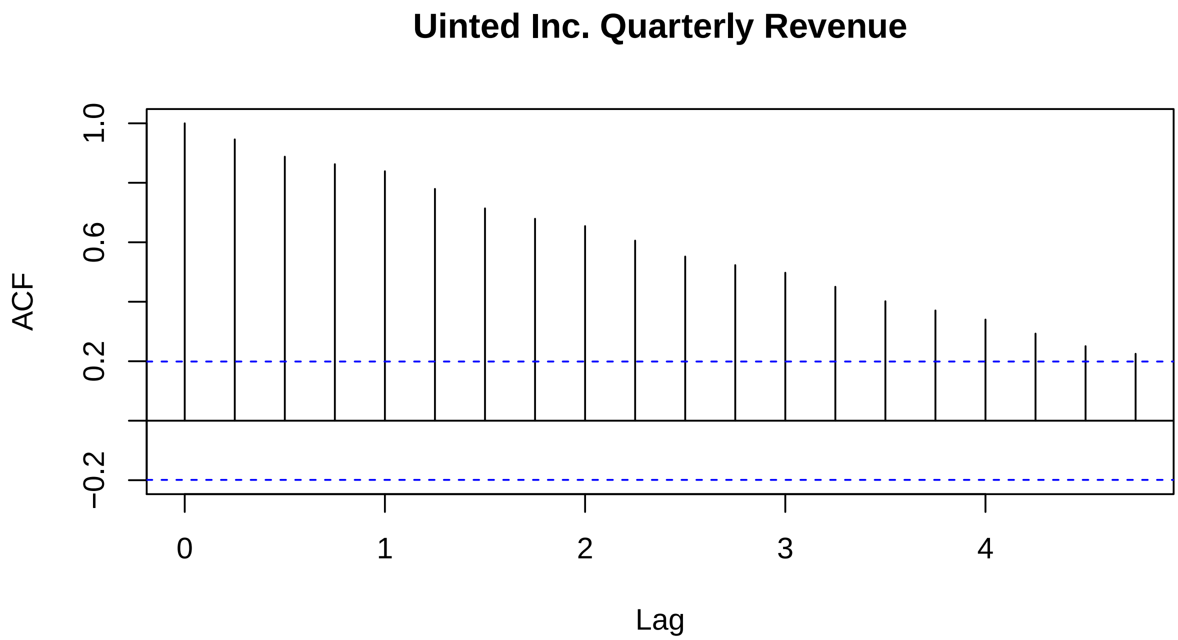
(a) Plot Time-series data of United Inc. Quartly Revenue

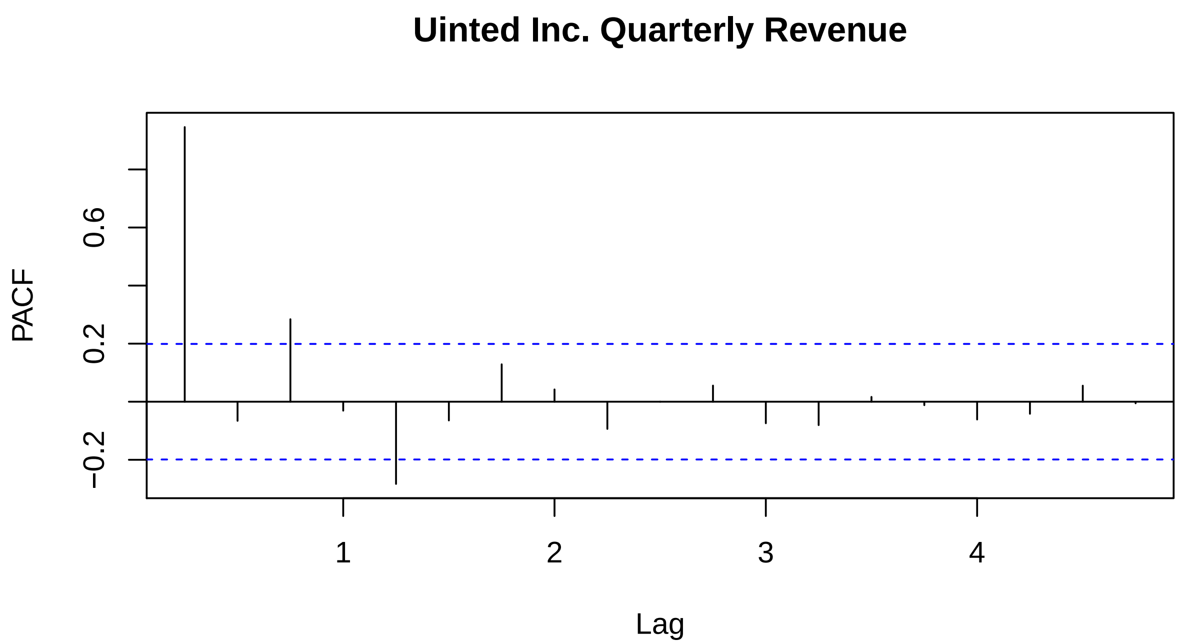


(b) Determination of Covariance Stationary

Our plot partially suggested that the data are covariance stationary. Let's take a closer look at the plot. From the year 1993 to approximately 2010, the data there does not seem to be covariance stationary as they fluctuate not around the sample mean. However, from the year 2010 to 2017,the plot actually has suggested that our data is covariance stationary. In fact, the data there has fluctuated around the sample mean and the variance is relatively stable during this time interval.

(c) Analyze Time Dependency by Utilizing ACF & PACF

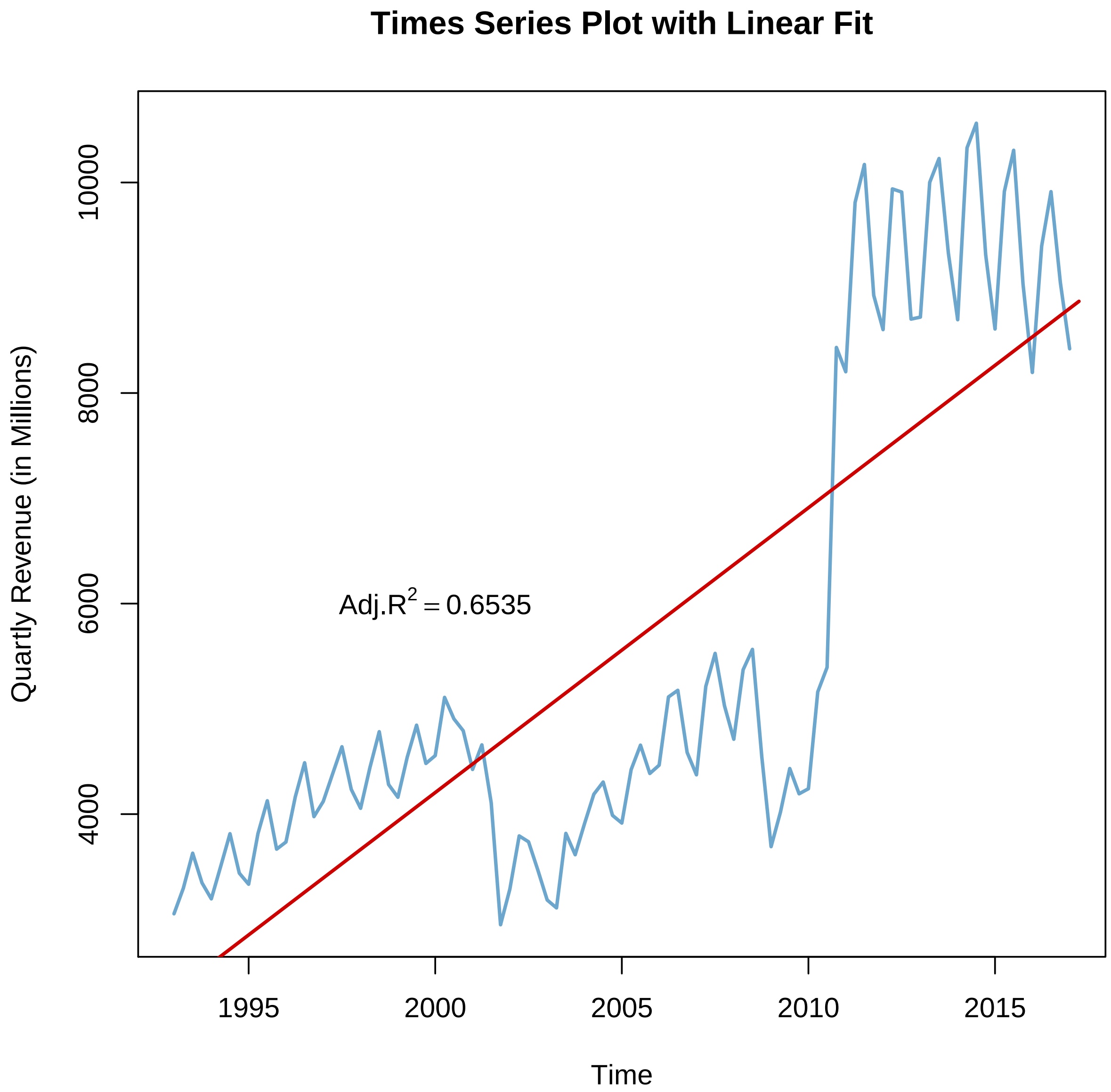




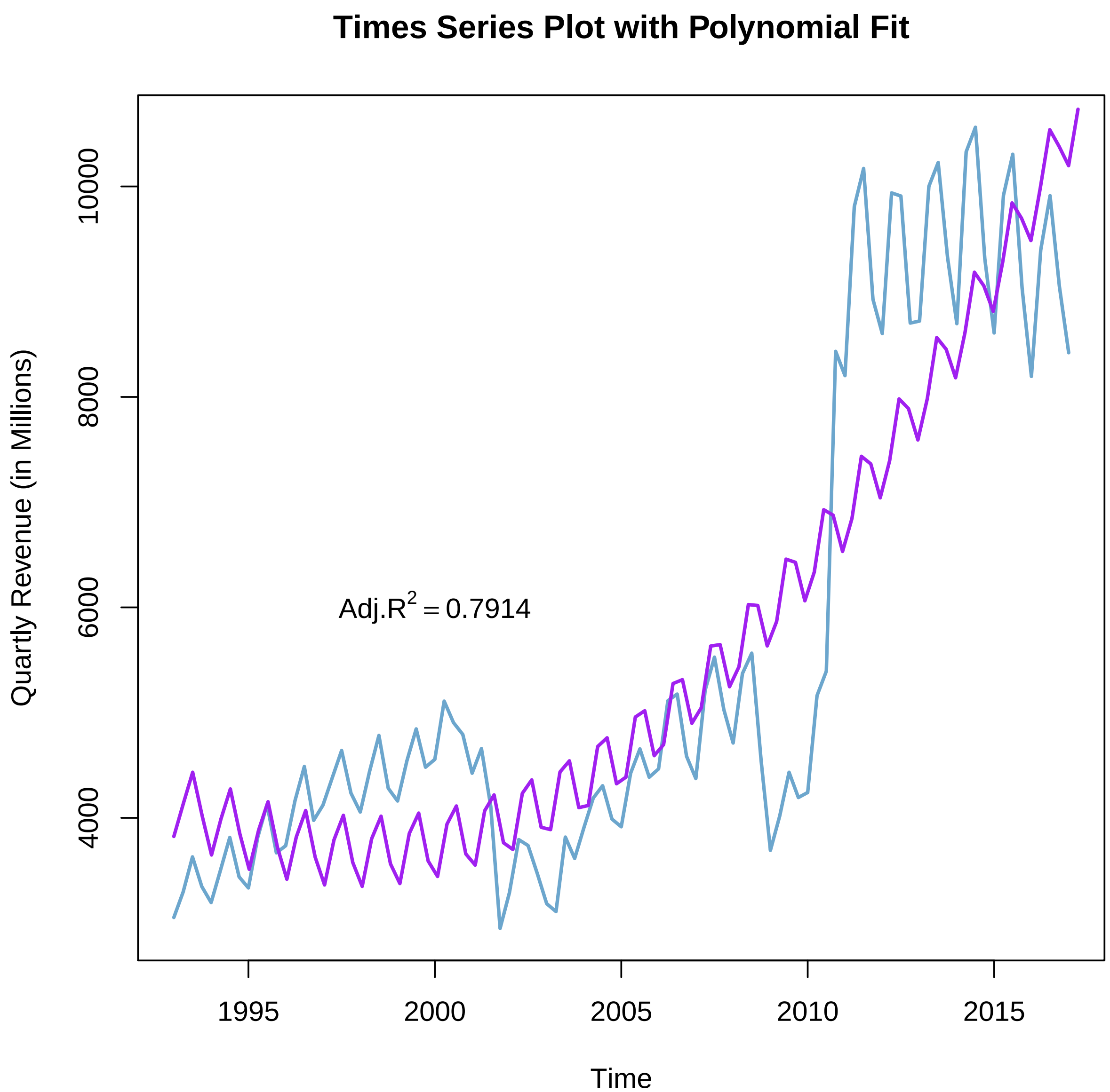
From the ACF nad PACF results produced by R, the time series is determined to have strong time dependency.

(d) Fit Models to the Time-series

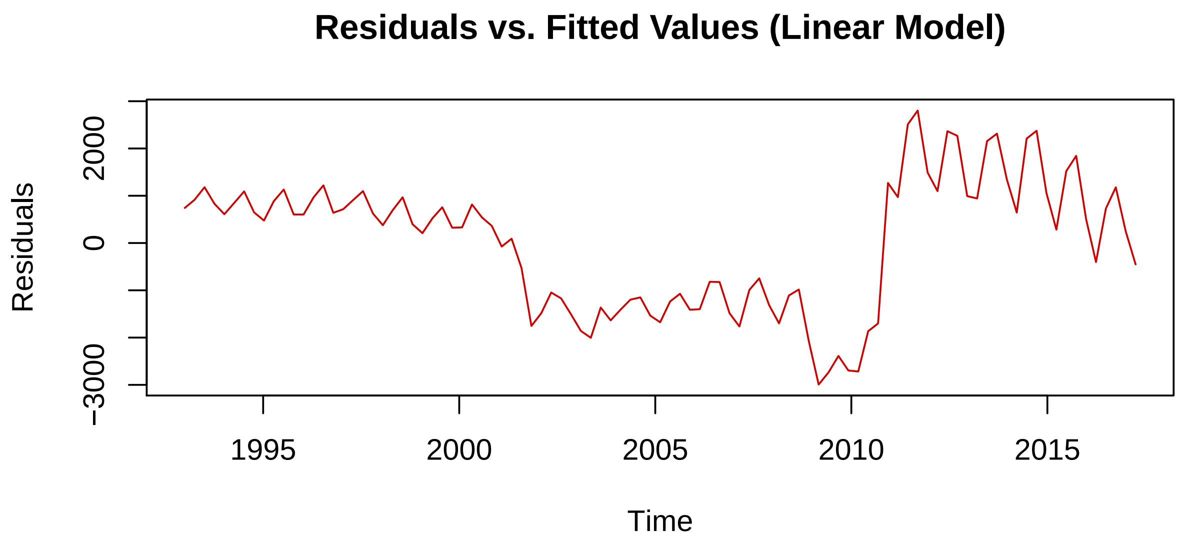
* Linear Model

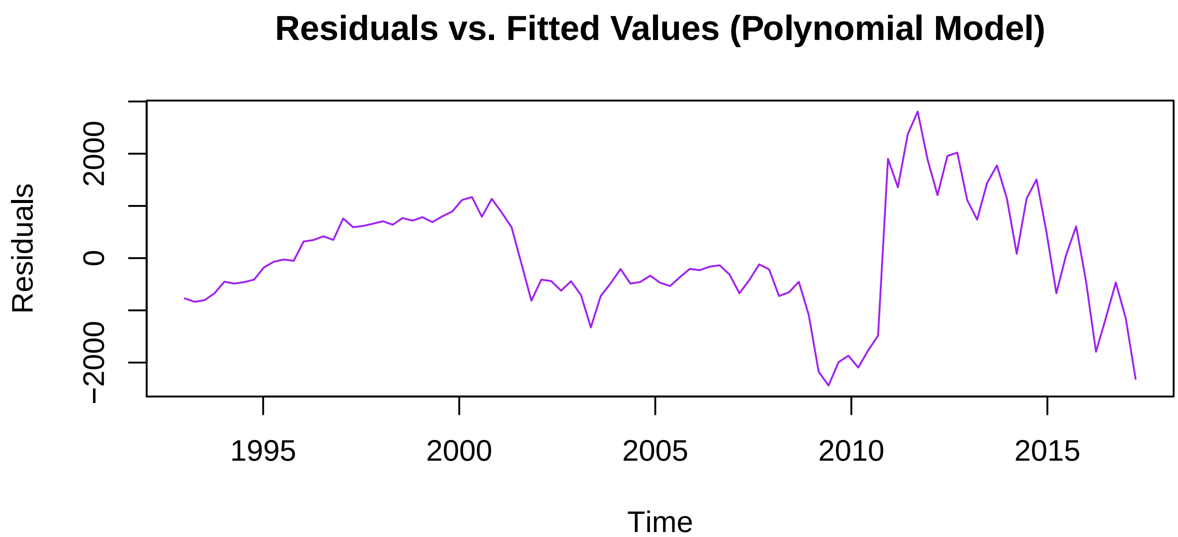


*** Polynomial**


Times Series Plot with Polynomial Fit
Quartly Revenue (in Millions)
4000
6000
8000
10000
Adj.R^2 = 0.7914
1995
2000
2005
2010
2015
Time

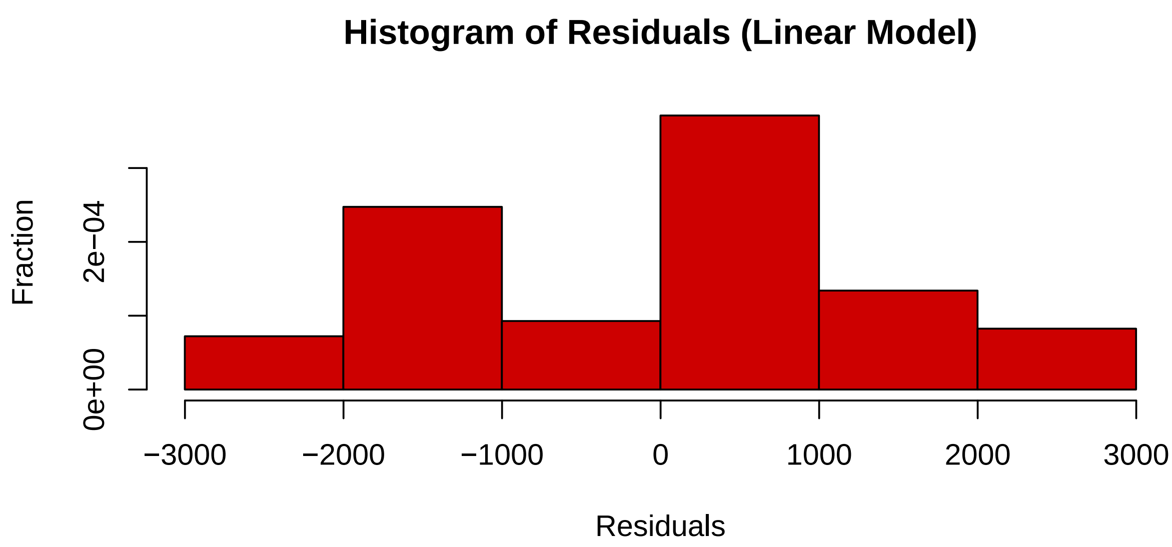
(e) Discussion on Residuals part.1

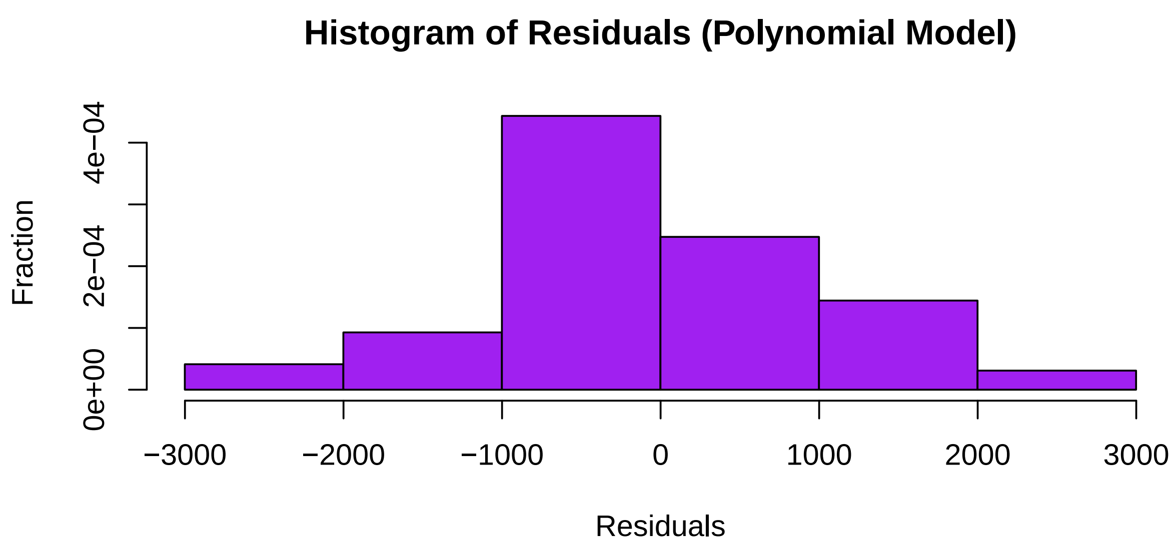




The residuals under each model behave similarly, both plots consist clear trend and seasonality. Thus, it indicates that non of the two models is a good fit to forecast this time series.

(f) Discussion on Residuals part.2





Unlike the Residuals vs. Fitted plots, histogram of residuals under each model look quite differently, residuals under the polynomial model(m2) have a more centralized distribution than those under the linear model(m1), which at some degree, implies that the is m2 a better fit compared to m1.

(g) Recognize Statistical Significance of Models

* Linear Model

```
##
## Call:
## lm(formula = UA_ts ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2993.5 -1236.6   398.0   964.3  2803.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -536805.59   40195.52  -13.36   <2e-16 ***
## t              270.51      20.05   13.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1396 on 95 degrees of freedom
## Multiple R-squared:  0.6572, Adjusted R-squared:  0.6535
## F-statistic: 182.1 on 1 and 95 DF,  p-value: < 2.2e-16
```

* Polynomial Model

```
##
## Call:
## lm(formula = UA_ts ~ t + I(t^2) + I(cos(2 * pi * t)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2440.3  -623.7  -177.3   737.1  2807.8
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.656e+07  9.893e+06   7.739 1.18e-11 ***
## t                   -7.663e+04  9.867e+03  -7.766 1.04e-11 ***
## I(t^2)               1.918e+01  2.461e+00   7.793 9.14e-12 ***
## I(cos(2 * pi * t)) -3.519e+02  1.557e+02  -2.261   0.0261 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1084 on 93 degrees of freedom
## Multiple R-squared:  0.7979, Adjusted R-squared:  0.7914
## F-statistic: 122.4 on 3 and 93 DF,  p-value: < 2.2e-16
```

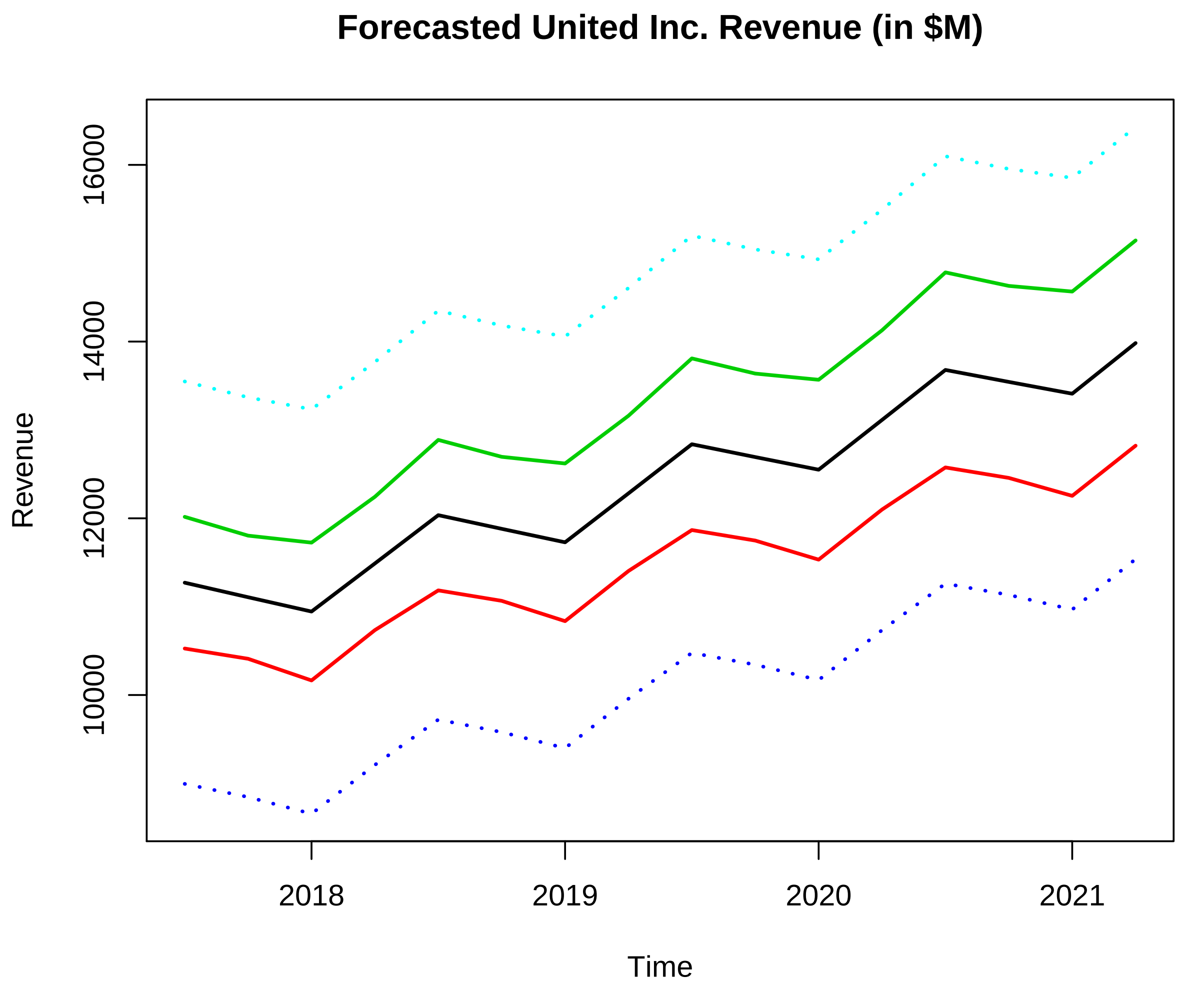
The linear model(m1) has an adjusted R-squared value of 0.6535; the Polynomial model(m2) has an adjusted R-squared value of 0.7914. The coefficients of each predictive variable under each model are all statistically significant. In addition, the F-statisitc and p-value of each model further consolidate that both regression fits are statistical significant models.

(h) Select Proper Model

	df	AIC	BIC
Linear	3	1684.137	1691.862
Polynomial	5	1636.866	1649.740

Both AIC and BIC test slightly favor in polynomial model(m2). When analyzing in accordance with regression summaries, model m2 is much more superior than m1. Thus, m2 is then chosen to predict future revenue.

(i) Forecast to 2021 Q1 Using m2



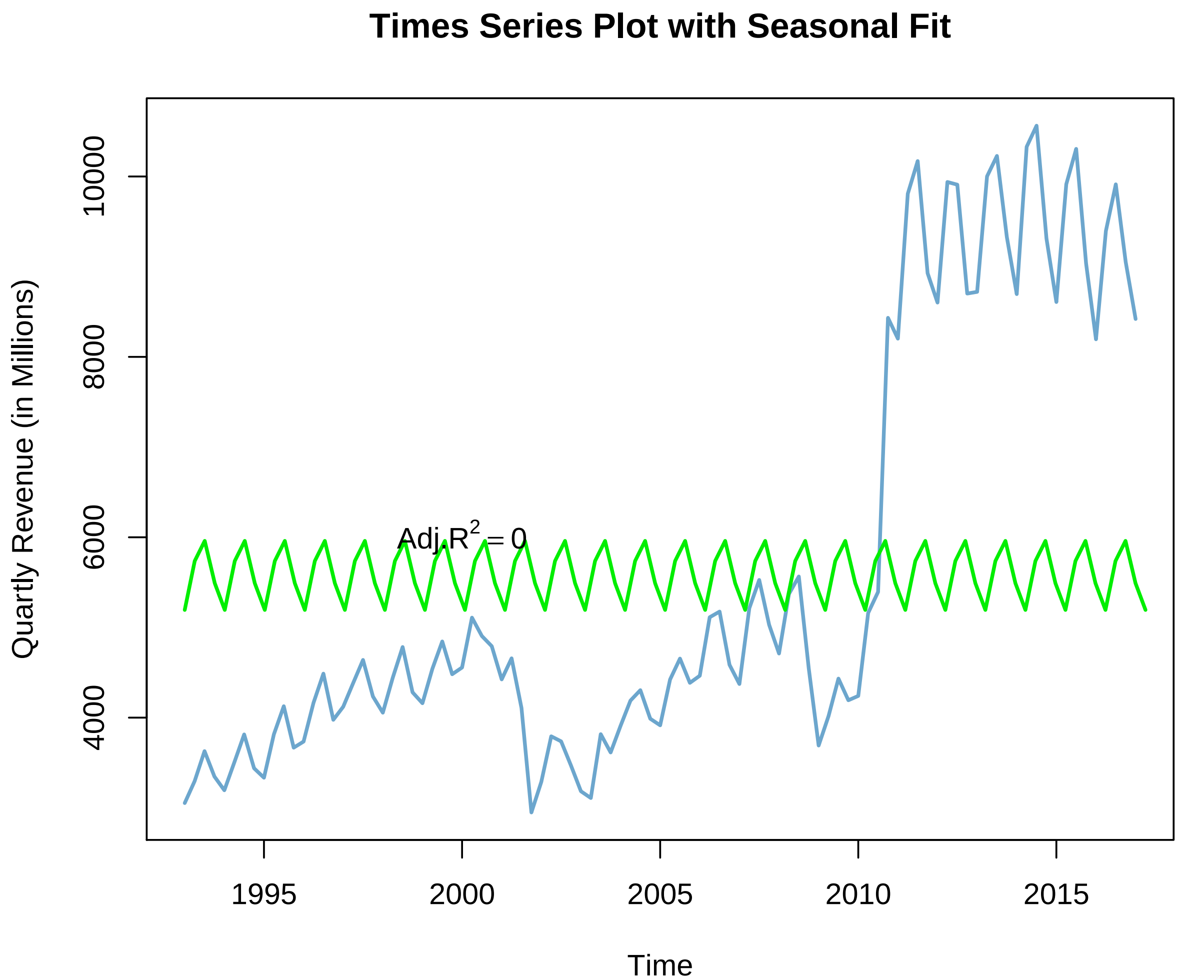

```
##        1        2        3        4        5        6        7        8 
## 11271.38 11106.91 10944.84 11489.04 12035.65 11880.77 11728.28 12282.07 
##        9       10       11       12       13       14       15       16 
## 12838.27 12692.97 12550.07 13113.46 13679.24 13543.53 13410.22 13983.19
```

In the above plot of forecasted revenue, the black line represents the forecasted revenue from 2017Q2 to 2021Q1; the region bounded between solid green line and red line represent the 95% confidence interval, and the region bounded between light-blue dashed-line and dark-blue dashed-line is the 95% prediction interval.

2. Modeling and Forecasting

(a) Construct New Model with Seasonal Dummies

* Plot

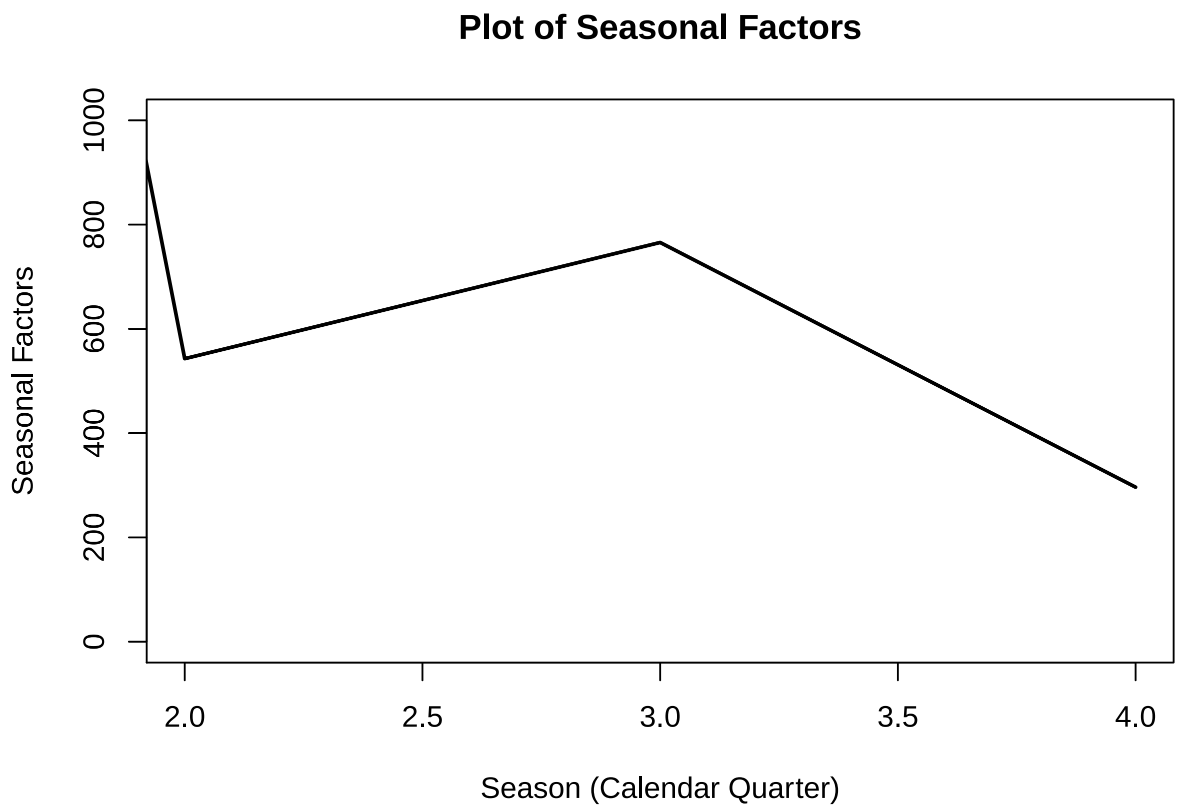


*** Regression Summary**

```
##
## Call:
## tslm(formula = UA_ts ~ season)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2628  -1573  -1073   3001   4603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5194.2      478.5  10.855   <2e-16 ***
## season2        542.8      683.7   0.794    0.429
## season3        765.7      683.7   1.120    0.266
## season4        296.3      683.7   0.433    0.666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2393 on 93 degrees of freedom
## Multiple R-squared:  0.01473,    Adjusted R-squared:  -0.01706
## F-statistic: 0.4633 on 3 and 93 DF,  p-value: 0.7086
```

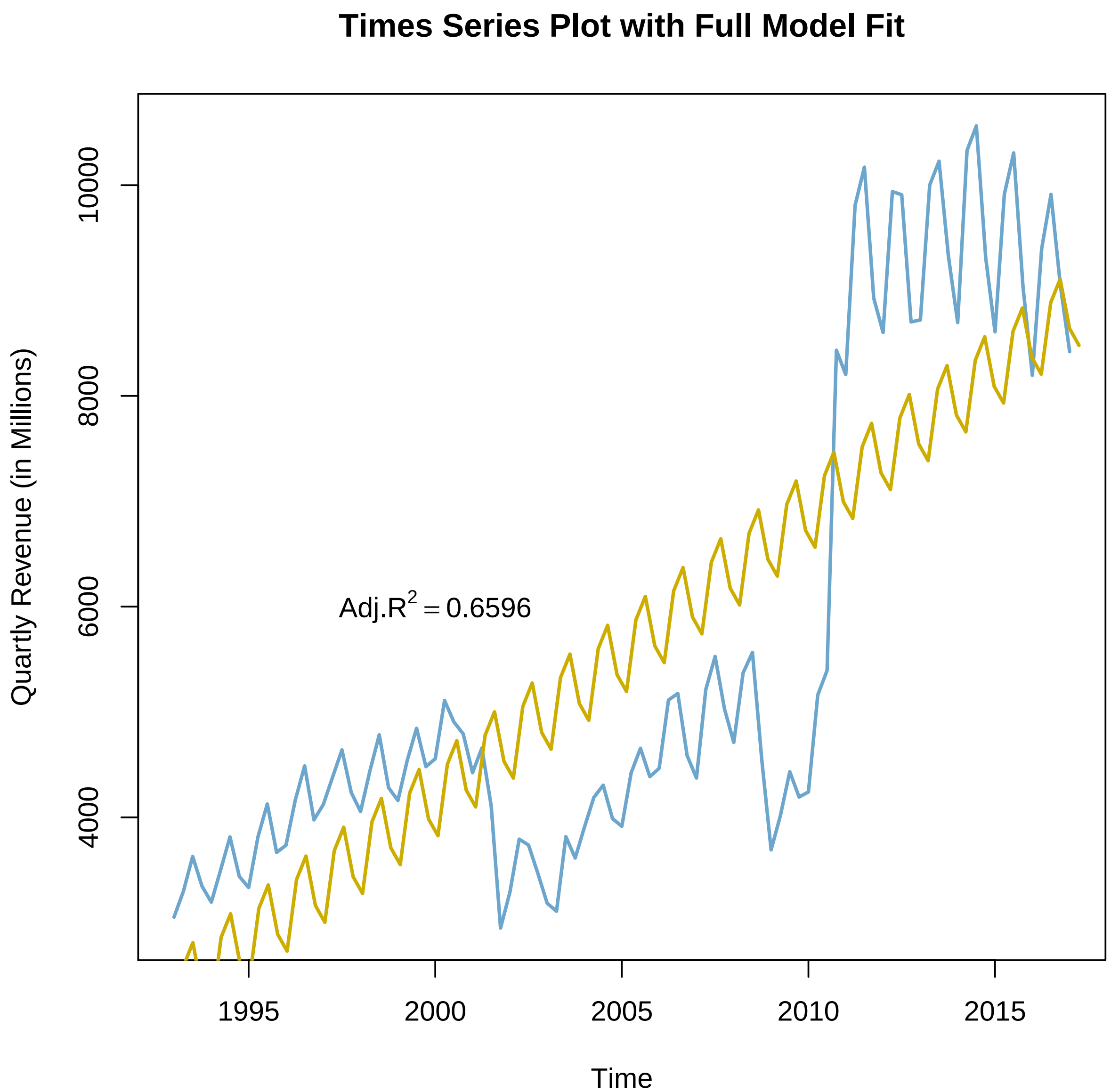
As the regression summary shows, this seasonal model has an adjusted R-squared value close to 0, and non of the seasonal variable coefficient is statistically significant. In summary, this model fits in the time series extremely poorly, and further adjustments are necessary before implementing this model for forecasting future revenue.

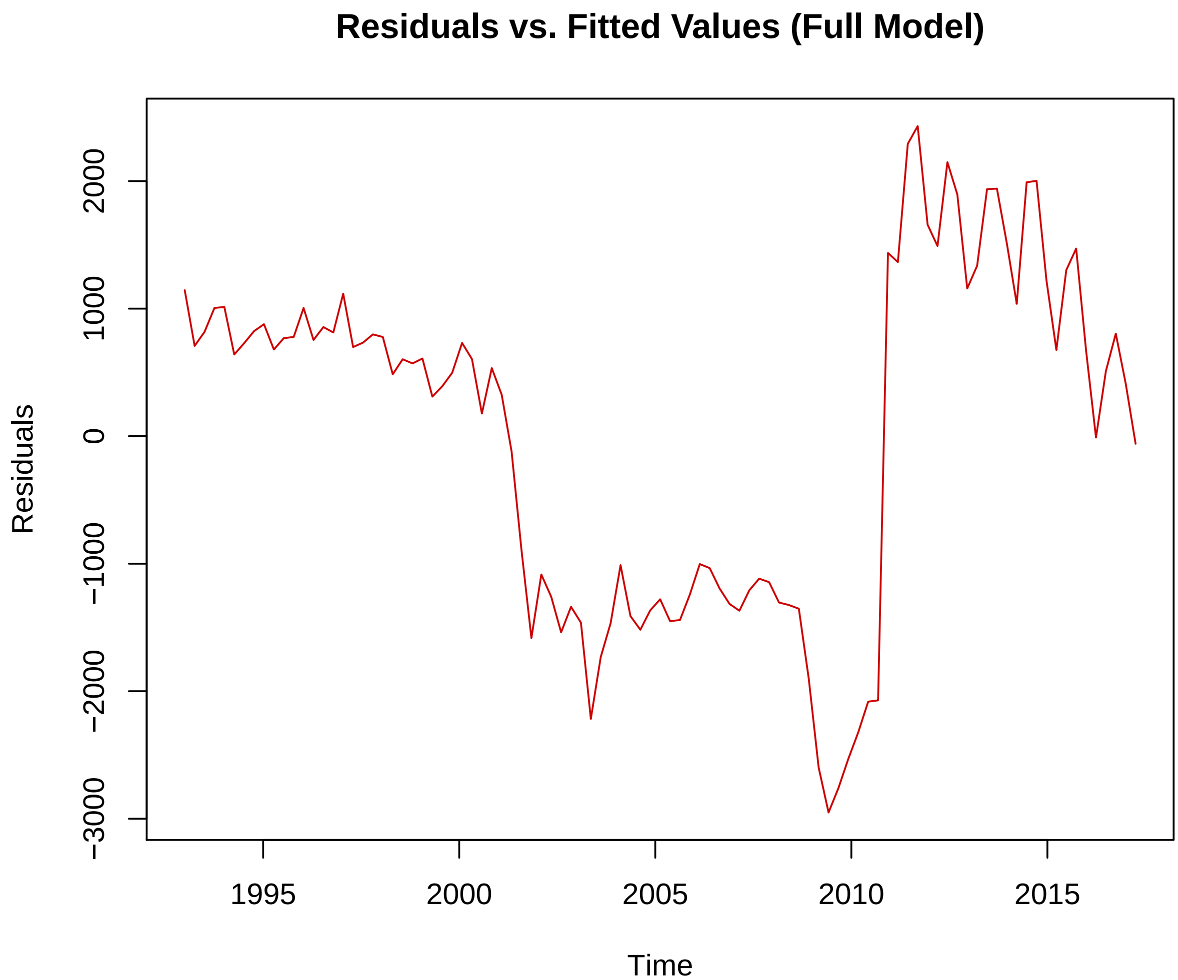
(b) Interpretation of Estimated Seasonal Factor



The Seaosnal Factors Plot provides a visual comparison between coefficient of each seasonal variable except the first variable since it is the intercept point. It appears that third quarters have the largest impact on generating revenue, and fourth quarters have the least impact.

(c) Building and Analyzing the Full Model w/ Trend





Under the full model, the behavior of residuals still preserves some trend and cycles, which indicates that the trend embedded in this model is not perfect, and there remains some degree of seasonality unadjusted by the 'season' component of the model. The inaccurate fit might be casued by the large scale of jump in year 2010 due to a change in United Inc.'s business structure.

(d) Summary and Error Metrics for the Full Model

*** Regression Summary**

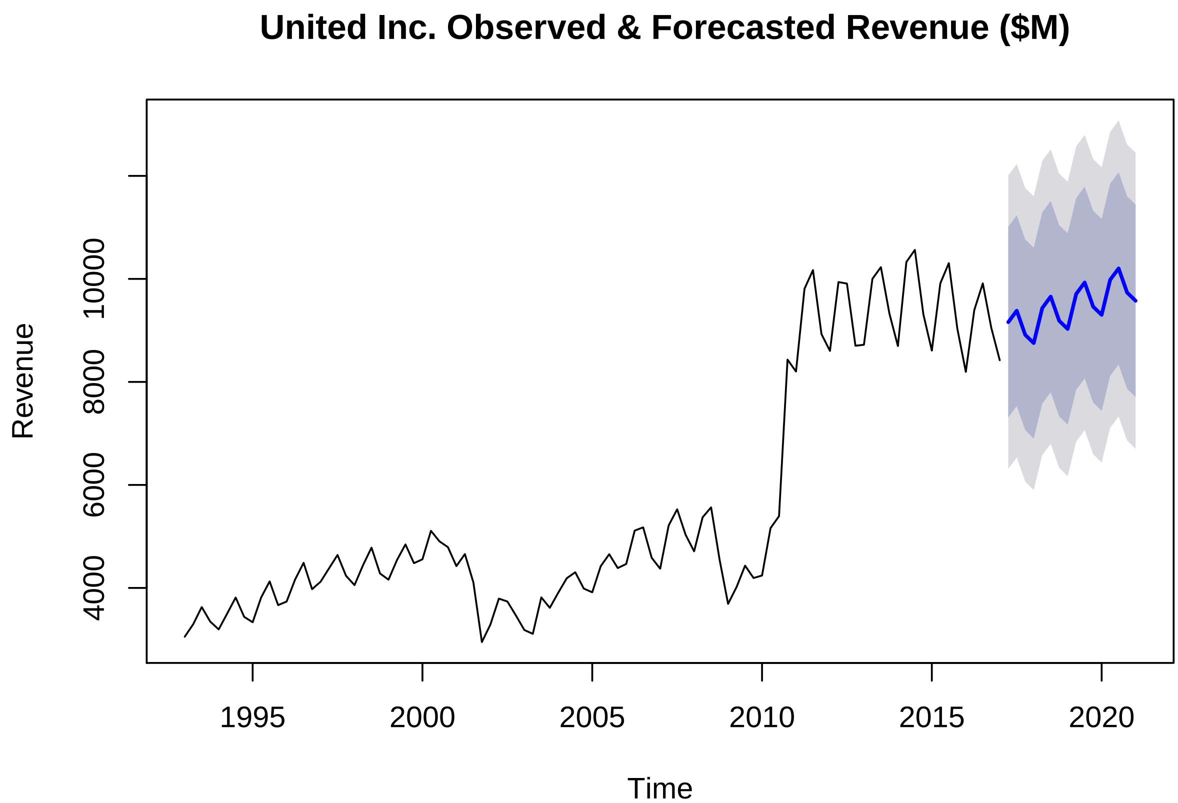
```
##
## Call:
## tslm(formula = UA_ts ~ trend + season)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2951.1 -1304.6   533.6   877.9  2431.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1840.025    370.353   4.968 3.11e-06 ***
## trend         68.452      5.021  13.634  < 2e-16 ***
## season2      611.250    395.586   1.545    0.126
## season3      765.715    395.554   1.936    0.056 .
## season4      227.888    395.586   0.576    0.566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1384 on 92 degrees of freedom
## Multiple R-squared:  0.6738, Adjusted R-squared:  0.6596
## F-statistic: 47.51 on 4 and 92 DF,  p-value: < 2.2e-16
```

*** MSE Error Matric**

```
## [1] 1348.006
```

The MSE error matric is calculated at 1348, which is a decent value considering the scale of the data. Although the full model outputs an adjusted R-squared value of 0.6596, the regression summary shows that non of the seasonal variable is statistically significant, which might be explained by the fact that the seasonal fluctuation has minor effect compared to the overall trend, yet this model is so far the best model to represent both trend and seasonality of this time series.

(e) Forecast to 2021 Q1 Using Full Model




```
##         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2017 Q2       9159.545 7307.441 11011.65 6309.857 12009.23
## 2017 Q3       9382.462 7530.358 11234.57 6532.773 12232.15
## 2017 Q4       8913.087 7060.983 10765.19 6063.398 11762.78
## 2018 Q1       8753.650 6900.669 10606.63 5902.613 11604.69
## 2018 Q2       9433.352 7576.537 11290.17 6576.415 12290.29
## 2018 Q3       9656.269 7799.453 11513.08 6799.332 12513.21
## 2018 Q4       9186.894 7330.078 11043.71 6329.957 12043.83
## 2019 Q1       9027.457 7169.587 10885.33 6168.896 11886.02
## 2019 Q2       9707.159 7845.283 11569.03 6842.436 12571.88
## 2019 Q3       9930.076 8068.200 11791.95 7065.353 12794.80
## 2019 Q4       9460.701 7598.825 11322.58 6595.978 12325.42
## 2020 Q1       9301.264 7438.156 11164.37 6434.644 12167.88
## 2020 Q2       9980.966 8113.684 11848.25 7107.924 12854.01
## 2020 Q3      10203.883 8336.601 12071.16 7330.841 13076.92
## 2020 Q4       9734.508 7867.226 11601.79 6861.466 12607.55
## 2021 Q1       9575.071 7706.380 11443.76 6699.861 12450.28
```

The forecasted revenue along with error bands under the full model is shown above.

III. Conclusions and Future Work

The final forecast model is currently the best one that we can produce under the guideline of this project. It has captured both the seasonality and trend in the model. However, this model is not perfect regarding to the dataset that we are currently modeling on. The future expected revenue could be under-estimated for short-term prediction and over-estimated for long-term predicition due to the fact that the model automatically compensates the huge jump in year 2010 by increasing trend slop and suppressing seasonal effect, which at some degree distorts future forecast.

The jump in revenue is caused by the $3 billion merger between the United Airline and Continental Airline that created the world's biggest airline (Mouawad and Merced, "United and Continental Said to Agree to Merge"). Since we do not expect another structural change in the close future, therefore, if in the future one would like to obtain more accurate predictions, he/she could do so by utilizing the data after the merger in 2010, in order words, in accordance with the pattern after 2010, to produce a more ideal result for short-term prediction.

In addition, as we mentioned in the introduction paragraph, one of the purposes of this project is to be able to see the differences between our expected revenue and the actual revenue that will be announced. According to the CNN Money, in the morning of the day after the incident, the United Airline's stock was off about 4%, knocking off close to $1 billion off the company's market value. In the same afternoon, the stock was eventually recovered from the worst losses after the CEO's apologization, but its market value was still off by $250 million (Kottasova, "United loses $250 million of its market value"). Therefore, in the short-run, we could expect a large difference between our predicted value and the actual revenue that will be reported, which has also illustrated the point that no matter how great the predicting model is, we still could not expect the unexpected.

IV. References

The following is the source of our data:

- http://ir.united.com/financial-performance/sec-filings.

As we have mentioned in the introduction part, we have constructed our data manually by pulling out the quarterly revenue number from the Unite Airline Financial Report. In addition, the reason why our data could only go back to 1993 is that we were not able to locate the data prior to that. One of the team member has manually entered the number into the excel file, and the other team members have verified and confirmed the numbers.

The following are the citations that we have utilized in the conclusion part

- Mouawad, Jad, and Michael J. De La Merced. "United and Continental Said to Agree to Merge." The New York Times. The New York Times, 02 May 2010. Web. 26 Apr. 2017.
- "United Airlines shares drop after man dragged off flight." CNNMoney. Cable News Network, n.d. Web. 26 Apr. 2017.

V. R Source Code

I

```r
#Read Data
UA <- read.xls("Data/UA-Revenue.xlsx")

#Pull out Revenue & Assign time series class
UA_ts <- ts(UA$Revenue...M., start = 1993, frequency = 4)
```

II

1(a)

```r
#Plot time series
plot(UA_ts, col = 'skyblue3', main = "United Inc. Quartly Revenue (in $M)",
     xlab = "Time", ylab = "Revenue")
```

1(c)

```r
#Plot ACF and PACF
par(mfrow = c(2,1))
acf(UA_ts, main = "Uinted Inc. Quarterly Revenue")
pacf(UA_ts, main = "Uinted Inc. Quarterly Revenue", ylab = "PACF")
```

1(d)

```r
#Introduce time sequence vector
t <- seq(1993, 2017.25, length = length(UA_ts))
```

```r
#Linear Fit
m1=lm(UA_ts ~ t)

#Plot time series overlay regression line
plot(UA_ts, main = "Times Series Plot with Linear Fit",
     xlab = "Time", ylab="Quartly Revenue (in Millions)",
     lwd=2, col='skyblue3')
lines(t,m1$fit,col="red3",lwd=2)
text(2000,6000,expression(Adj.R^2==0.6535))
```

```r
#Quadratic with Periodic fit
m2 <- lm(UA_ts ~ t + I(t^2) + I(cos(2*pi*t)))

#Plot time series overlay regression line
plot(UA_ts, main = "Times Series Plot with Polynomial Fit",
     xlab = "Time", ylab="Quartly Revenue (in Millions)",
     lwd=2, col='skyblue3')
lines(t,m2$fit,col="purple",lwd=2)
text(2000,6000,expression(Adj.R^2==0.7914))
```

1(e)

```r
par(mfrow=c(2,1))
#Plot Residual v.s Fitted for m1
plot(t,m1$res, col = "red3", type='l',xlab="Time", ylab="Residuals",
     main="Residuals vs. Fitted Values (Linear Model)")

#Plot Residual v.s Fitted for m2
plot(t,m2$res, col = "purple", type='l', xlab="Time",ylab="Residuals",
     main="Residuals vs. Fitted Values (Polynomial Model)")
```

1(f)

```r
par(mfrow=c(2,1))
#Plot Histogram of Residuals for m1
truehist(m1$res,col="red3", xlab="Residuals", ylab="Fraction",
         main="Histogram of Residuals (Linear Model)")

#Plot Histogram of Residuals for m2
truehist(m2$res,col="purple", xlab="Residuals", ylab="Fraction",
         main="Histogram of Residuals (Polynomial Model)")
```

1(g)

```r
summary(m1)
summary(m2)
```

1(h)

```
#Compute AIC & BIC
AIC <- AIC(m1,m2)
BIC <- BIC(m1,m2)

#Create AIC & BIC table
AIC_BIC <- cbind(AIC, BIC = BIC$BIC)
rownames(AIC_BIC) <- c("Linear", "Polynomial")
kable(AIC_BIC)
```

1(i)

```
#Plot forecast result
tn <- data.frame(t = seq(2017.5, 2021.25, length = 16))
pred <- predict(m2, tn, se.fit = TRUE)
pred.plim <- predict(m2, tn, level = 0.95, interval = "prediction")
pred.clim <- predict(m2, tn, level = 0.95, interval = "confidence")
matplot(tn$t, cbind(pred.clim, pred.plim[, -1]),
        lty=c(1,1,1,3,3), type="l", lwd=2,
        main = "Forecasted United Inc. Revenue (in $M)",
        ylab="Revenue", xlab="Time")

#Show predict value
pred$fit
```

2(a)

```
#Fit seasonal dummies for quarterly effects
fit1 <- tslm(UA_ts ~ season)

#Plot time series overlay fitted line
plot(UA_ts, main = "Times Series Plot with Seasonal Fit",
     xlab = "Time", ylab="Quartly Revenue (in Millions)",
     lwd=2, col='skyblue3')
lines(t,fit1$fit,col="green2",lwd=2)
text(2000,6000,expression(Adj.R^2==0.00))
```

```
#Diagnostic test for the model
summary(fit1)
```

2(b)

```
#Seasonal Factors Plot
plot(fit1$coef,type='l', lwd=2, xlim = c(2,4), ylim = c(0,1000),
     ylab='Seasonal Factors', xlab="Season (Calendar Quarter)",
     main="Plot of Seasonal Factors")
```

2(c)

```
#Fit the full model with trend and seasonal dummies
fit2 <- tslm(UA_ts ~ trend + season )

#Plot time series overlay fitted line
plot(UA_ts, main = "Times Series Plot with Full Model Fit",
     xlab = "Time", ylab="Quartly Revenue (in Millions)",
     lwd=2, col='skyblue3')
lines(t,fit2$fit,col="gold3",lwd=2)
text(2000,6000,expression(Adj.R^2==0.6596))
```

```
#Plot Residual v.s Fitted for the full model
plot(t,fit2$res, col = "red3", type='l',
     xlab="Time", ylab="Residuals", main="Residuals vs. Fitted Values (Full Model)")
```

2(d)

```
#Regression summary of the full model
summary(fit2)

#Mse of the full model
rmse(UA_ts, fit2$fitted.values)
```

2(e)

```
#Plot forecast result under the full model
pred.f <- forecast.lm(fit2, tn)

plot(pred.f, main = " United Inc. Observed & Forecasted Revenue ($M)",
     xlab = "Time", ylab = "Revenue")

#Show Predicted value
pred.f
```