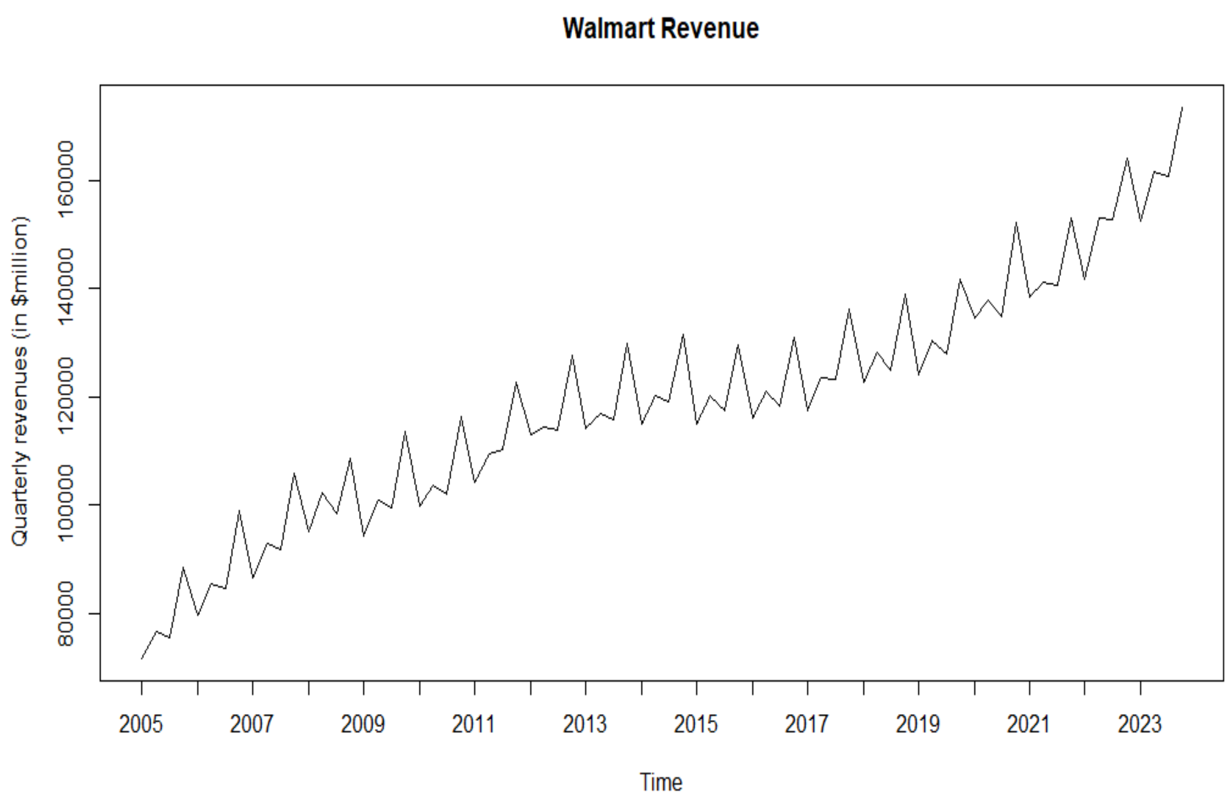
Case Study #2 - Regression Models

1. Plot the data and visualize time series components
   a. Create time series dataset.

```
revenue.ts <- ts(revenue.data$Revenue, start=c(2005, 1), end=c(2023, 4), frequency = 4)
revenue.ts
```

   b. Here is a plot of historical data for Walmart's revenue. In this time series for Walmart's revenue from 2005 to 2023 we can see that there is an upward trend of some sort since over time the values of revenue increase as the years pass. It is unclear if this upwards trend is linear or quadratic or other but once we develop some models, we can see which performs the best to provide further insight. Next, we can notice that there is additive seasonality in this time series. This is because there are regular intervals where revenue increases and decreases which occur at similar periods of times.



2. Apply five regression models using data partition.
   a. Develop a data partition with the validation partition of 16 periods and the rest for the training partition.

```
nValid <- 16
nTrain <- length(revenue.ts) - nValid

train.ts <- window(revenue.ts, start = c(2005, 1), end = c(2005, nTrain))
valid.ts <- window(revenue.ts, start = c(2005, nTrain + 1), end = c(2005, nTrain + nValid))
```

b. Model Building and Evaluation

i. Linear Trend

```
tslm(formula = train.ts ~ trend)

Residuals:
     Min       1Q   Median       3Q      Max 
-13713.4  -5045.9   -416.3   4058.6  15335.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 84527.92    1829.19   46.21   <2e-16 ***
trend         865.48      52.15   16.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6996 on 58 degrees of freedom
Multiple R-squared:  0.826,	Adjusted R-squared:  0.823 
F-statistic: 275.4 on 1 and 58 DF,  p-value: < 2.2e-16
```

Equation: 84527.92+865.48(trend)
Predictor: trend

Explanation: The model shows high statistical significance and is a decent fit for the data but not good enough for forecasting in the validation partition. Too smooth of predictions.

The trend explains a significant portion of the variability in train.ts, with an adjusted R-squared value of 0.823 and small p-value throughout.

```
> train.lin.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2020 137322.1 138187.6 139053.0 139918.5
2021 140784.0 141649.5 142515.0 143380.4
2022 144245.9 145111.4 145976.9 146842.3
2023 147707.8 148573.3 149438.8 150304.3
```

ii. Quadratic Trend

```
tslm(formula = train.ts ~ trend + I(trend^2))

Residuals:
   Min     1Q Median     3Q    Max 
 -8848  -4356  -1331   5045  12581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 76386.191   2442.424  31.275  < 2e-16 ***
trend        1653.387    184.749   8.949 1.87e-12 ***
I(trend^2)    -12.917      2.936  -4.400 4.80e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6097 on 57 degrees of freedom
Multiple R-squared:  0.8701,	Adjusted R-squared:  0.8656 
F-statistic:   191 on 2 and 57 DF,  p-value: < 2.2e-16
```

Equation:76386.191+1653.387(trend)−12.917(trend^2)
Predictors: trend, trend^2

Explanation: The model shows high statistical significance and seems to be a good fit for the data but when looking at the forecast this model provided, underestimating is prevalent so this would not be a good model to choose. The trend and squared trend together explain a significant portion of the variability in train.ts, with an adjusted R-squared value of 0.8656 and small p-values throughout.

```
> train.quad.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2020 129180.4 129245.0 129283.8 129296.8
2021 129284.0 129245.3 129180.8 129090.4
2022 128974.3 128832.2 128664.4 128470.7
2023 128251.2 128005.9 127734.7 127437.7
```

iii. Seasonality

```
tslm(formula = train.ts ~ season)

Residuals:
   Min     1Q Median     3Q    Max 
-33029  -9632   5943  10617  20676 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   104525       4072  25.667   <2e-16 ***
season2         5176       5759   0.899    0.373    
season3         3593       5759   0.624    0.535    
season4        16831       5759   2.922    0.005 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15770 on 56 degrees of freedom
Multiple R-squared:  0.1463,	Adjusted R-squared:  0.1006 
F-statistic: 3.199 on 3 and 56 DF,  p-value: 0.03017
```

Equation:104525+5176(season2)+3593(season3)+16831(season4)

Predictors: season2, season3, season4

Explanation: The model shows high statistical significance but is not a good fit for the data. Looking at the forecast, underestimating is prevalent and the entire model only captures seasonality as expected. P-values are alright but the adjusted R-squared value is very low showing that this model does not capture the variability in the data.

```
> train.seas.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2020 104525.0 109701.1 108117.7 121356.3
2021 104525.0 109701.1 108117.7 121356.3
2022 104525.0 109701.1 108117.7 121356.3
2023 104525.0 109701.1 108117.7 121356.3
```

iv. Linear Trend and Seasonality

```
tslm(formula = train.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-9267.6 -3135.2   307.5  3637.7  8485.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 79914.73    1455.18  54.917  < 2e-16 ***
trend         848.63      32.25  26.312  < 2e-16 ***
season2      4327.44    1576.86   2.744  0.00817 ** 
season3      1895.41    1577.85   1.201  0.23480    
season4     14285.38    1579.50   9.044  1.8e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4318 on 55 degrees of freedom
Multiple R-squared:  0.9372,	Adjusted R-squared:  0.9326 
F-statistic: 205.1 on 4 and 55 DF,  p-value: < 2.2e-16
```

Equation:79914.73+848.63(trend)+4327.44(season2)+1895.41(season3)+14285.38(season4)

Predictors: trend, season2, season3, season4

Explanation: The model shows high statistical significance and is a decent fit for the data. The forecast results show underestimation for half of the validation partition. The trend and seasonal dummies together explain a significant portion of the variability in train.ts, with an adjusted R-squared value of 0.9326.

```
> train.lin.seas.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2020 131681.2 136857.2 135273.8 148512.4
2021 135075.7 140251.8 138668.4 151907.0
2022 138470.2 143646.3 142062.9 155301.5
2023 141864.7 147040.8 145457.4 158696.0
```

v. Quadratic Trend and Seasonality

```
tslm(formula = train.ts ~ trend + I(trend^2) + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-4072.5 -1738.4    33.4  1486.5  5873.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71768.162   1072.465  66.919  < 2e-16 ***
trend        1638.260     71.710  22.846  < 2e-16 ***
I(trend^2)    -12.945      1.139 -11.362 6.15e-16 ***
season2      4301.547    864.246   4.977 6.95e-06 ***
season3      1869.517    864.788   2.162   0.0351 *  
season4     14285.376    865.688  16.502  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2366 on 54 degrees of freedom
Multiple R-squared:  0.9815,	Adjusted R-squared:  0.9798 
F-statistic:   572 on 5 and 54 DF,  p-value: < 2.2e-16
```

Equation:71768.162+1638.260(trend)−12.945(trend^2)+4301.547(season2)+1869.517(season3)+14285.376(season4)
Predictors: trend, trend^2, season2, season3, season4
Explanation: The model shows high statistical significance and is a good fit for the data but when looking at the forecast, the model underestimates a lot in the validation partition. High adjusted R-squared and small p-values don't seem to mean much here due to the forecast.

```
> train.quad.trend.seas.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2020 123534.6 127882.2 125470.3 137880.5
2021 123563.5 127807.5 125292.1 137598.7
2022 123178.1 127318.6 124699.6 136902.7
2023 122378.6 126415.5 123692.9 135792.4
```

c. After evaluating the forecasting accuracy of various regression models using MAPE and RMSE metrics, we've identified the three most accurate models. Topping the list is the model incorporating both a Linear Trend and Seasonality, with a MAPE of 4.058 and an RMSE of 8282.813. Following closely is the Linear Trend model, with a MAPE of 4.799 and an RMSE of 9805.521. Lastly, the Quadratic Trend and Seasonality model also shows strong performance, albeit less accurate than the previous two, with a MAPE of 13.957 and an RMSE of 23464.216. It's worth noting that these models, which account for both trend and seasonality, outperform the individual trend and seasonality models, underscoring the importance of incorporating both factors for more accurate forecasting

Model	MAPE	RMSE
Linear Trend	4.799	9805.521
Quadratic Trend	13.344	23801.483
Seasonality	25.558	39550.732
Linear trend and Seasonality	4.058	8282.813
Quadratic Trend and Seasonality	13.957	23464.216

Choices: Linear Trend and Seasonality, Linear Trend, Quadratic Trend and Seasonality.

3. Employ the entire data set to make time series forecast.
   a. Train two most accurate models with the entire dataset for future prediction.
      i. Linear Trend

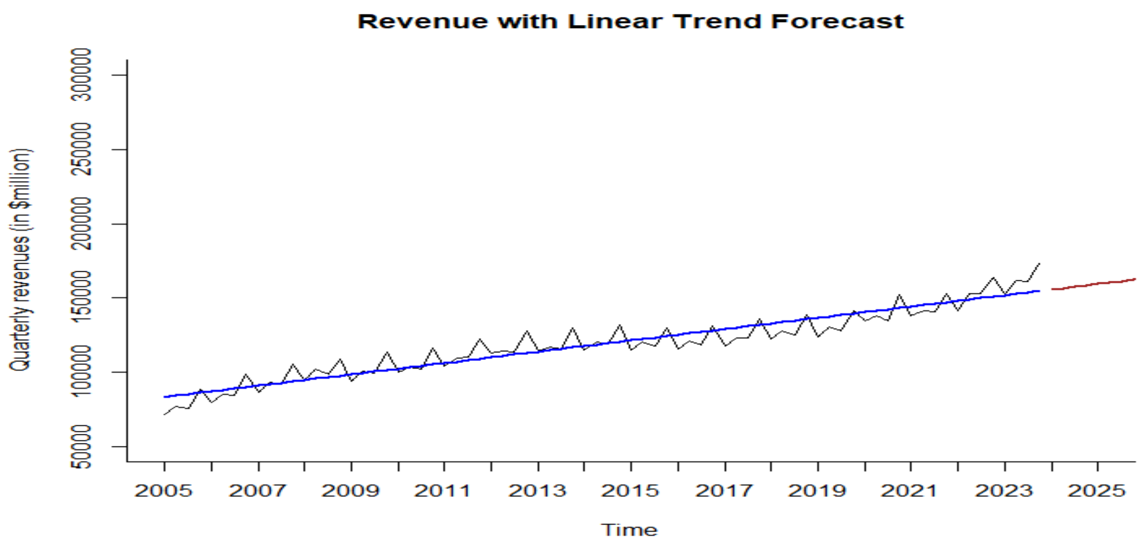
```
tslm(formula = revenue.ts ~ trend)

Residuals:
   Min     1Q Median     3Q    Max 
-12710  -5841     70   3462  18679 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 82414.01    1703.65   48.38   <2e-16 ***
trend         951.25      38.45   24.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7353 on 74 degrees of freedom
Multiple R-squared:  0.8922,    Adjusted R-squared:  0.8907 
F-statistic: 612.1 on 1 and 74 DF,  p-value: < 2.2e-16
```

Equation: 82414.01+951.25(trend)
Predictor: trend
Explanation: The model shows high statistical significance and is a decent fit for the data but not ideal as its forecast provides values that smooths revenue too much.
The trend explains a significant portion of the variability in revenue.ts, with an adjusted R-squared value of 0.8907 and small p-value throughout.




```
> tot.lin.trend.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2024 155660.2 156611.4 157562.7 158513.9
2025 159465.2 160416.4 161367.7 162318.9
2026 163270.2 164221.4 165172.7 166123.9
2027 167075.2 168026.4 168977.7 169928.9
```

ii. Linear Trend and Seasonality

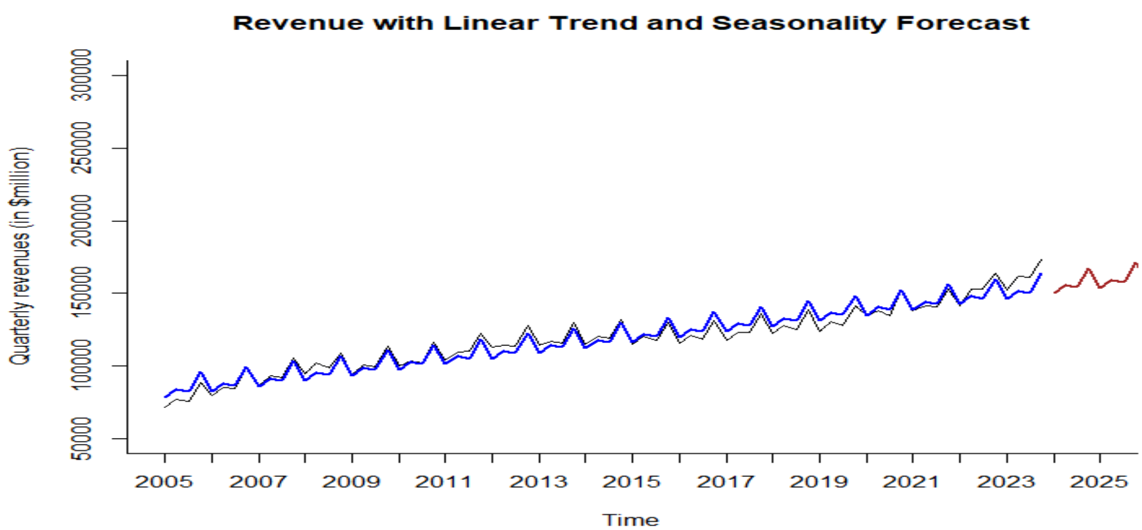
```
tslm(formula = revenue.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-7427.9 -4275.5   524.9  3108.0 10593.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 77548.36    1460.13  53.111  < 2e-16 ***
trend         940.62      25.46  36.945  < 2e-16 ***
season2      4539.38    1577.91   2.877   0.0053 ** 
season3      2115.49    1578.52   1.340   0.1845    
season4     14444.08    1579.55   9.144 1.27e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4863 on 71 degrees of freedom
Multiple R-squared:  0.9547,	Adjusted R-squared:  0.9522 
F-statistic: 374.4 on 4 and 71 DF,  p-value: < 2.2e-16
```

Equation:
77548.36+940.62(trend)+4539.38(season2)+2115.49(season3)+14444.08(season4)
Predictors: trend, season2, season3, season4
Explanation: The model shows high statistical significance and is a good fit for the data. Trend and seasons explains a significant portion of the variability in revenue.ts, with an adjusted R-squared value of 0.9522 and small p-value throughout except for season3 coefficient. Forecast is good since it captures the trend and seasonality of the previous years but may be underestimating.




```
> tot.lin.trend.seas.pred$mean
         Qtr1     Qtr2     Qtr3     Qtr4
2024 149976.4 155456.4 153973.1 167242.3
2025 153738.8 159218.8 157735.6 171004.8
2026 157501.3 162981.3 161498.1 174767.3
2027 161263.8 166743.8 165260.6 178529.8
```

b. See results table below. After evaluating the forecasting accuracy of the different models using MAPE and RMSE, we can determine the most accurate model for forecasting Walmart’s quarterly revenue in Q1-Q4 of 2024-2025. Among the models considered, the "Linear Trend and Seasonality" model emerges as the most accurate with a MAPE of 3.428 and an RMSE of 4700.120. It outperforms the other models, including the "Linear Trend" model with a MAPE of 5.027 and an RMSE of 7255.488, the "Naive" model with a MAPE of 6.928 and an RMSE of 9705.706, and the "Seasonal Naive" model with a MAPE of 4.081 and an RMSE of 5863.128. Thus, the "Linear Trend and Seasonality" model provides the most accurate forecasts for Walmart’s quarterly revenue during the specified period.

Model	MAPE	RMSE
Linear Trend	5.027	7255.488
Linear Trend and Seasonality	3.428	4700.120
Naive	6.928	9705.706
Seasonal Naive	4.081	5863.128