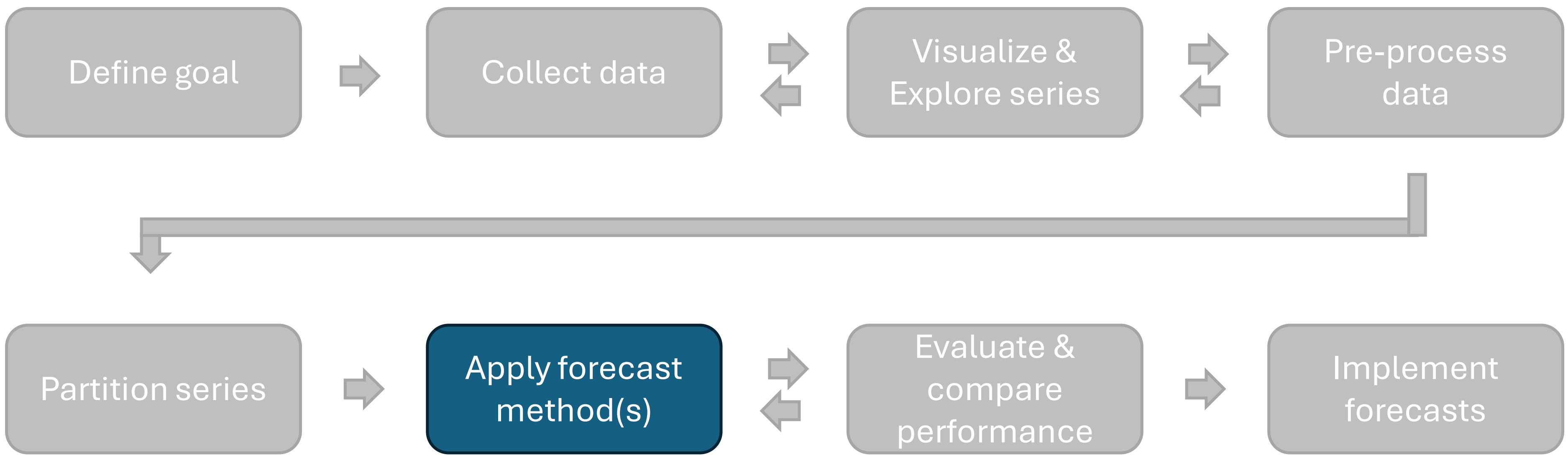
Time series Forecasting

Part 3: Linear Regression

The forecasting process



Components of a Time Series - Recap

1. **Level** (always present)
2. **Trend**: steady increase/decrease over time.
3. **Seasonality**: pattern that repeats itself every season
4. Random **noise** (always present)

Additive:

Y_t = Level + Trend + Seasonality + Noise

Multiplicative:

Y_t = Level x Trend x Seasonality x Noise

Regression-Based Methods

<u>The Idea:</u>
Using suitable predictors to capture trend and/or seasonality

<u>Advantages:</u>
Simple, popular

<u>Disadvantages:</u>
Model can become too complicated very quickly

<u>Uses:</u>
Examine trend and/or seasonality

<u>Key concept:</u>
Using suitable predictors

Models that we will examine

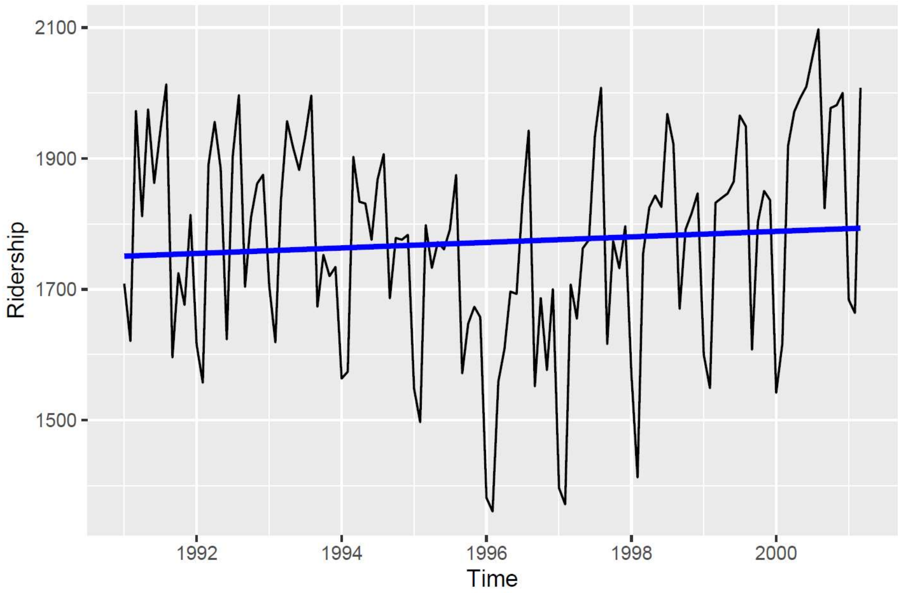
- Models with Trend
  - Linear Trend
  - Exponential Trend
  - Polynomial Trend
- Models with Seasonality
- Models with trend & Seasonality

Models with linear trend

- Values of the series increase or decrease linearly with time

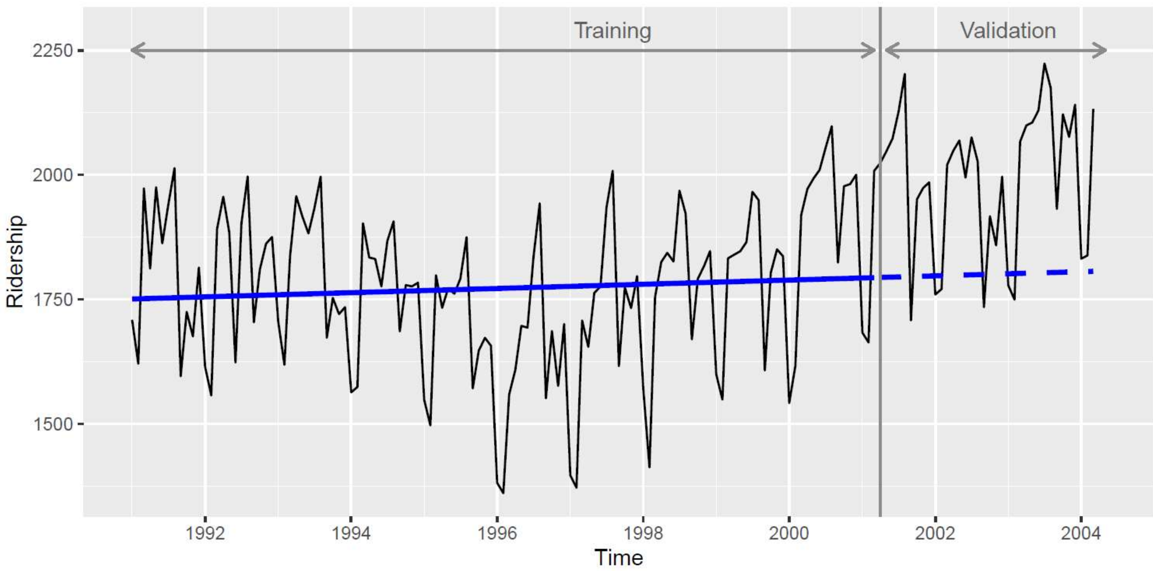
$$y_t = \beta_0 + \beta_1 t + e_t$$

```
train.ridership <- ridership |> filter_index(~ "2001 Mar")
valid.ridership <- ridership |> filter_index("2001 Apr" ~ .)

train.lm <- train.ridership |> model( TSLM( Ridership ~ trend() ) )

train.ridership |>
        autoplot(Ridership) +
        autolayer(fitted.values(train.lm), colour = "blue1", size = 1.2) +
        scale_x_yearmonth(date_breaks = "2 years", date_labels = "%Y") +
        labs(x = "Time")
```



Linear trend fitted to the Amtrak ridership data.

```
# Generate forecasts for the validation period

fc.lm <- train.lm |> forecast(h = 36)

train.ridership |>
        autoplot(Ridership) +
        autolayer(fitted.values(train.lm), colour = "blue1", size = 1.2) +
        geom_line(aes(y = .mean), data = fc.lm, colour = "blue1", linetype = "dashed", size = 1.2) +
        scale_x_yearmonth(date_breaks = "2 years", date_labels = "%Y") +
        labs(x = "Time")
```



Linear trend fitted to the Amtrak ridership training data and forecasts for the validation period.

```
# Model's summary output

> report(train.lm)

Series: Ridership
Model: TSLM

Residuals:    Min        1Q   Median       3Q      Max
          -411.29   -114.02    16.06   129.28   306.35

Coefficients:
                 Estimate    Std. Error    t value     Pr(>|t|)
(Intercept)     1750.3595    29.0729       60.206     <2e-16 ***
trend()            0.3514     0.4069        0.864     0.39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 160.2 on 121 degrees of freedom
Multiple R-squared: 0.006125,      Adjusted R-squared: -0.002089
F-statistic: 0.7456 on 1 and 121 DF, p-value: 0.38957
```

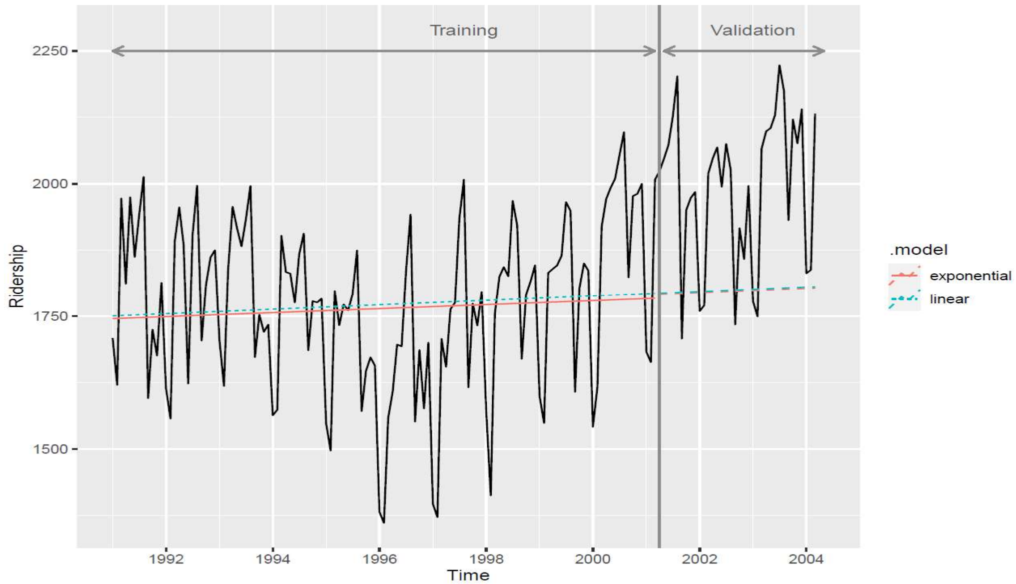
An insignificant coefficient does not mean no trend in the data. There may be a trend in the data once we control for seasonality. Examine the plot!

Models with Exponential trend

- An exponential trend implies a multiplicative increase/decrease of the series over time ($y_t = c\ e^{\beta_1 t + e_t}$).
- It reflects a percentage increase/decrease.
- To fit exponential trend, simply replace the output variable y with $\ln(y)$ and fit a linear regression:

$$\ln(y_t) = \beta_0 + \beta_1 t + e_t$$

```
train.lm <- train.ridership |>
                    model(    linear = TSLM( Ridership ~ trend() ),
                              exponential = TSLM( log(Ridership) ~ trend() )
                         )
fc_trends <- train.lm |> forecast(h = 36)

ridership |>  autoplot(Ridership) +
          geom_line(data = fitted(train.lm), aes(y = .fitted, colour = .model, linetype = .model)) +
          autolayer(fc_trends, level = NULL, linetype = "dashed") +
          labs(x = "Time") +
          scale_x_yearmonth(date_breaks = "2 years", date_labels = "%Y")
```



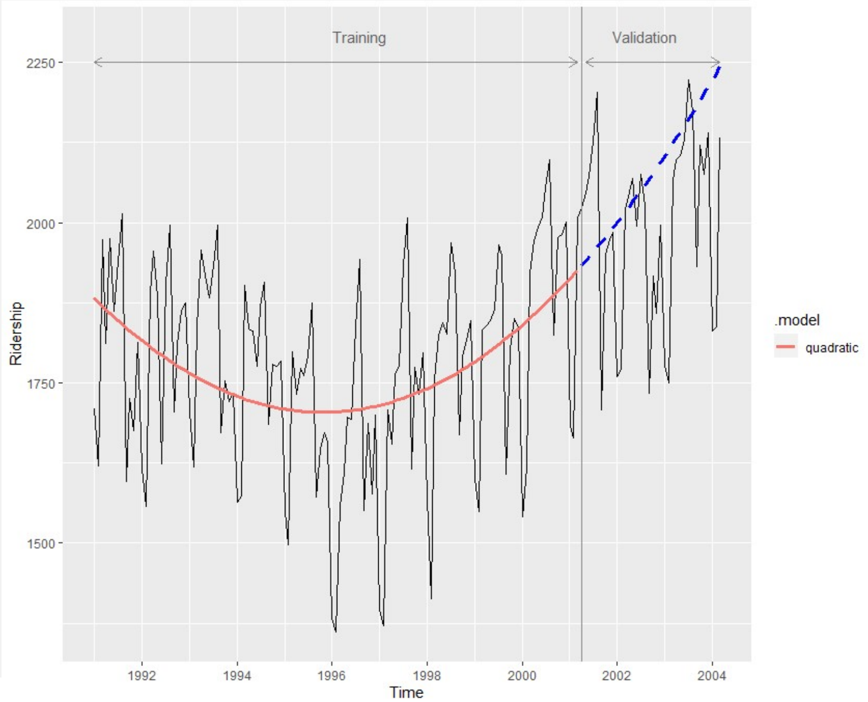
Linear and exponential trends fitted to the Amtrak ridership training data and forecasts for the validation period.

Models with polynomial trend

$$y_t = \beta_0 + \beta_1 t + \beta_t^2 + \beta_t^3 + \cdots + e_t$$

Linear Regression

```
# quadratic (U-shape) trend

> train.lm.poly.trend <- train.ridership |>  model(TSLM(Ridership ~ trend() + I(trend()^2)))

> report(train.lm.poly.trend)
Series: Ridership
Model: TSLM

Residuals:
     Min        1Q    Median       3Q       Max
-344.79    -101.86     40.89    98.54    279.81

Coefficients:
                  Estimate    Std. Error   t value     Pr(>|t|)
(Intercept)   1888.88401      40.91521     46.166     < 2e-16 ***
trend()          -6.29780      1.52327     -4.134     6.63e-05 ***
I(trend()^2)      0.05362      0.01190      4.506     1.55e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148.8 on 120 degrees of freedom
Multiple R-squared: 0.1499, Adjusted R-squared: 0.1358
F-statistic: 10.58 on 2 and 120 DF, p-value: 5.8437e-05
```



Hands-On # 3.1

- Recreate the plot

Models that we will examine

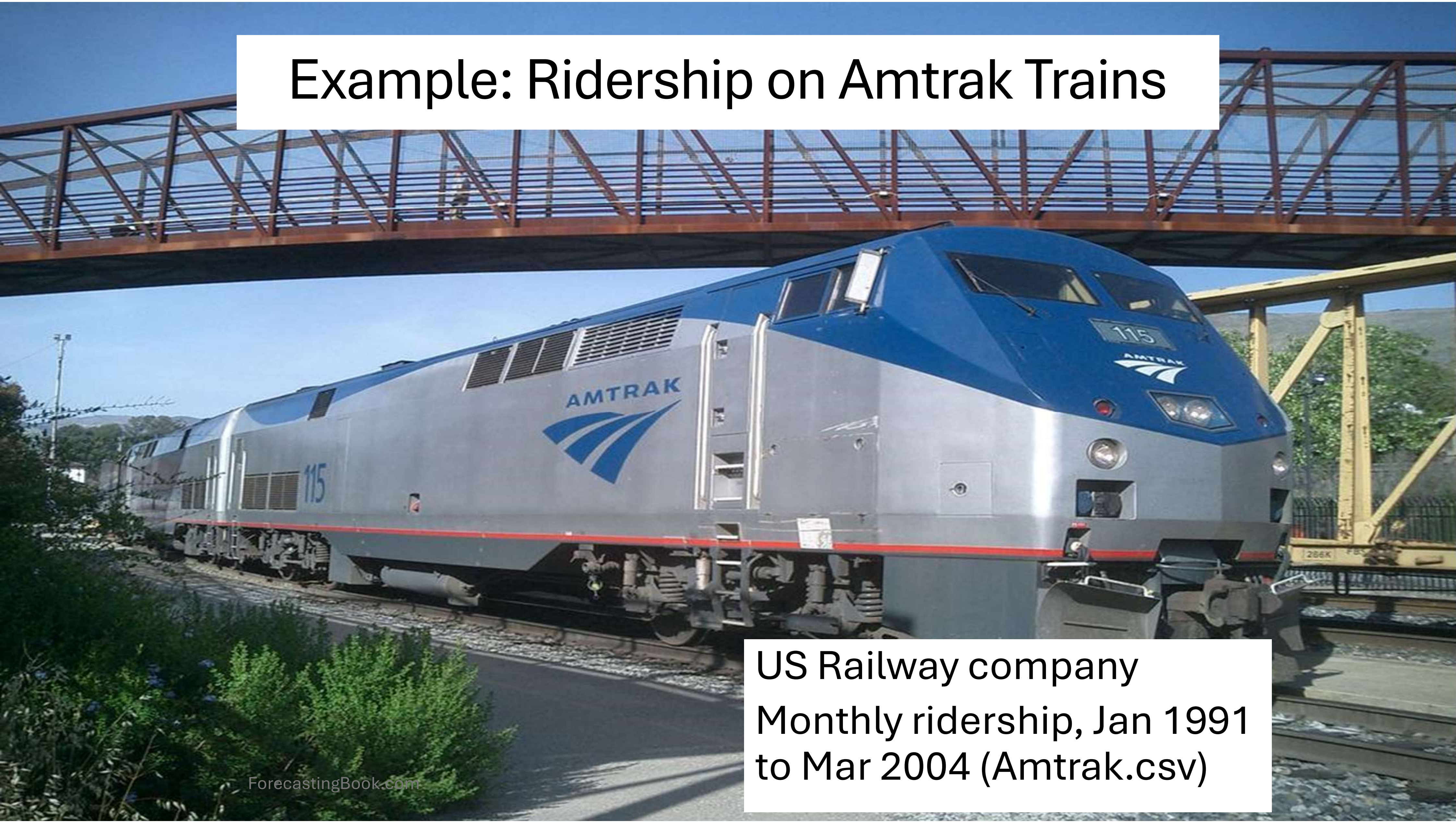
- Models with Trend
  - Linear Trend
  - Exponential Trend
  - Polynomial Trend
- Models with Seasonality
- Models with trend & Seasonality

Models with seasonality

- Observations that fall in some seasons have consistently higher or lower values than those in others.
- Amtrak ridership monthly time series, for example, exhibits strong monthly seasonality (pick in summer months).
- To capture seasonality, we can create dummy variables

  - For *m* seasons, we create *m-1* dummy variables
  - Each take on the value *1* if the record falls in that particular season and *0* otherwise.
  - The *m*-th season does not require a dummy, since it is identified when all the *m-1* dummies take on zero values.

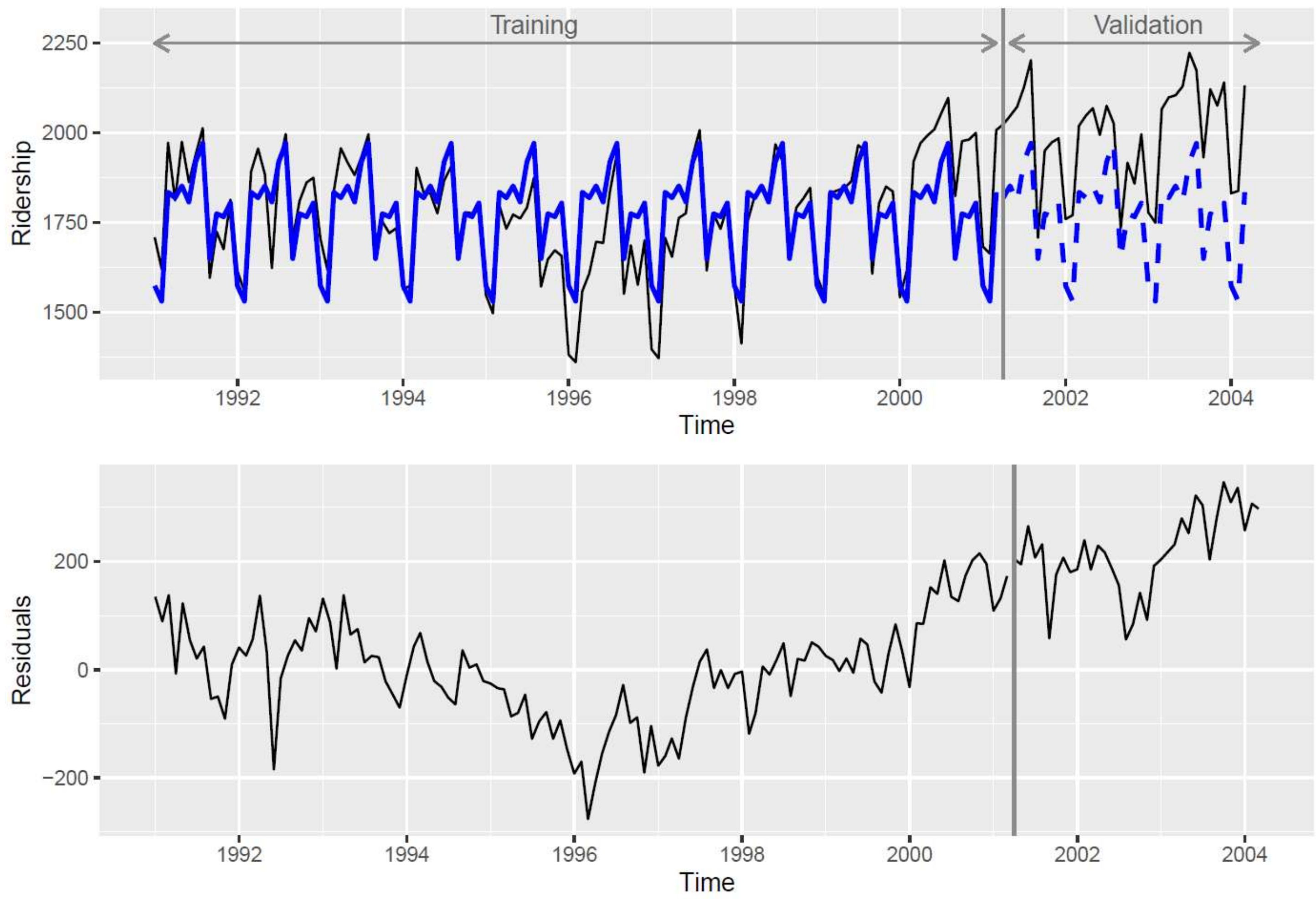
Example: Ridership on Amtrak Trains



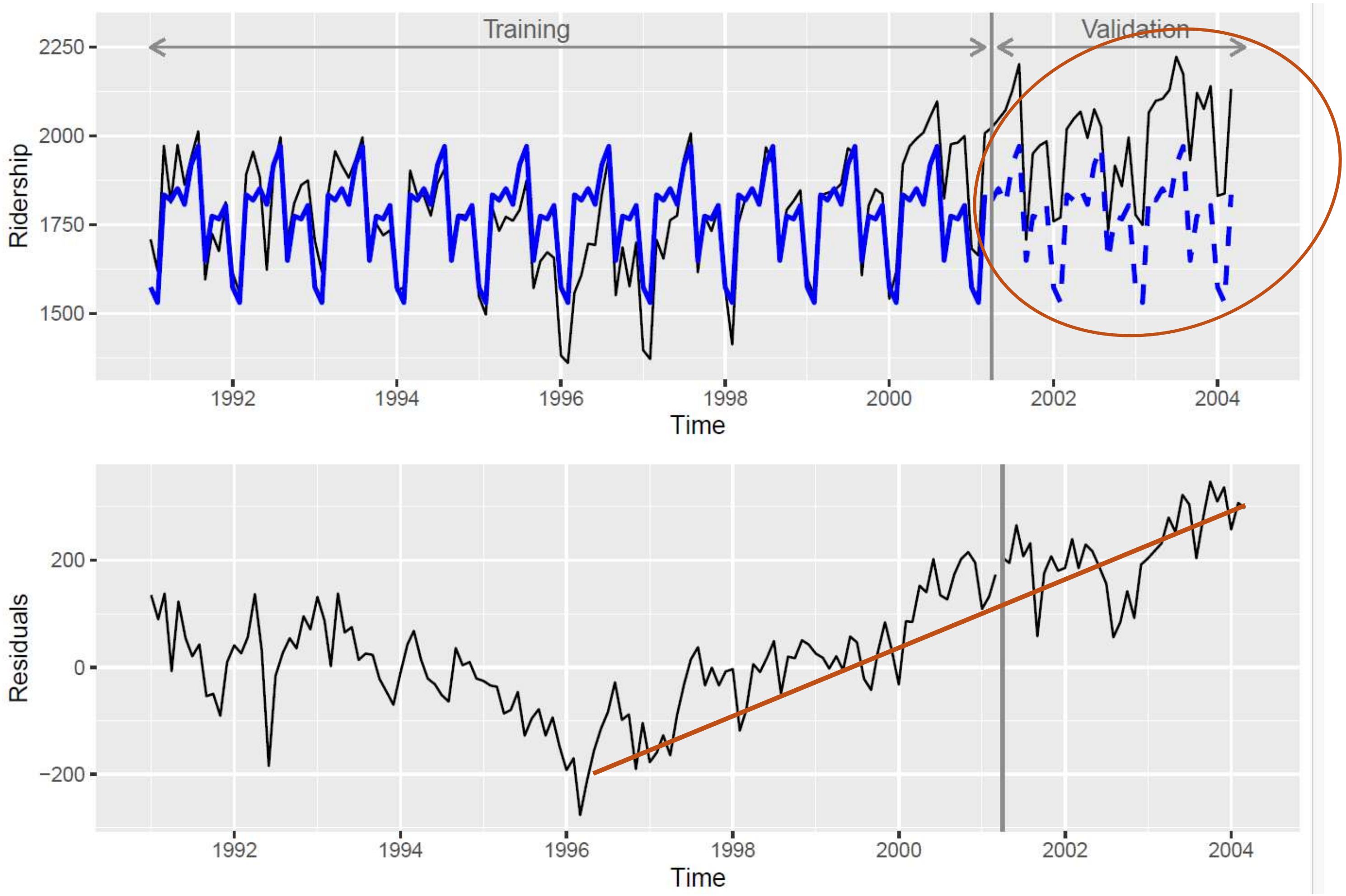
US Railway company

Monthly ridership, Jan 1991 to Mar 2004 (Amtrak.csv)

```
> train.lm.season <- train.ridership |> model(TSLM(Ridership ~ season()))
> report(train.lm.season)
Series: Ridership
Model: TSLM

Residuals:
     Min       1Q   Median       3Q      Max
-276.165  -52.934    5.868   54.544  215.081

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1573.97      30.58  51.475  < 2e-16 ***
season()year2    -42.93      43.24  -0.993   0.3230
season()year3    260.77      43.24   6.030 2.19e-08 ***
season()year4    245.09      44.31   5.531 2.14e-07 ***
season()year5    278.22      44.31   6.279 6.81e-09 ***
season()year6    233.46      44.31   5.269 6.82e-07 ***
season()year7    345.33      44.31   7.793 3.79e-12 ***
season()year8    396.66      44.31   8.952 9.19e-15 ***
season()year9     75.76      44.31   1.710   0.0901 .
season()year10   200.61      44.31   4.527 1.51e-05 ***
season()year11   192.36      44.31   4.341 3.14e-05 ***
season()year12   230.42      44.31   5.200 9.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.4 on 111 degrees of freedom
Multiple R-squared: 0.6348,	Adjusted R-squared: 0.5986
F-statistic: 17.54 on 11 and 111 DF, p-value: < 2.22e-16
```



A regression model with additive seasonality fitted to the Amtrak ridership data in the top panel (model is a blue thick line in training and dashed line in validation). The model's residuals are in the bottom panel.

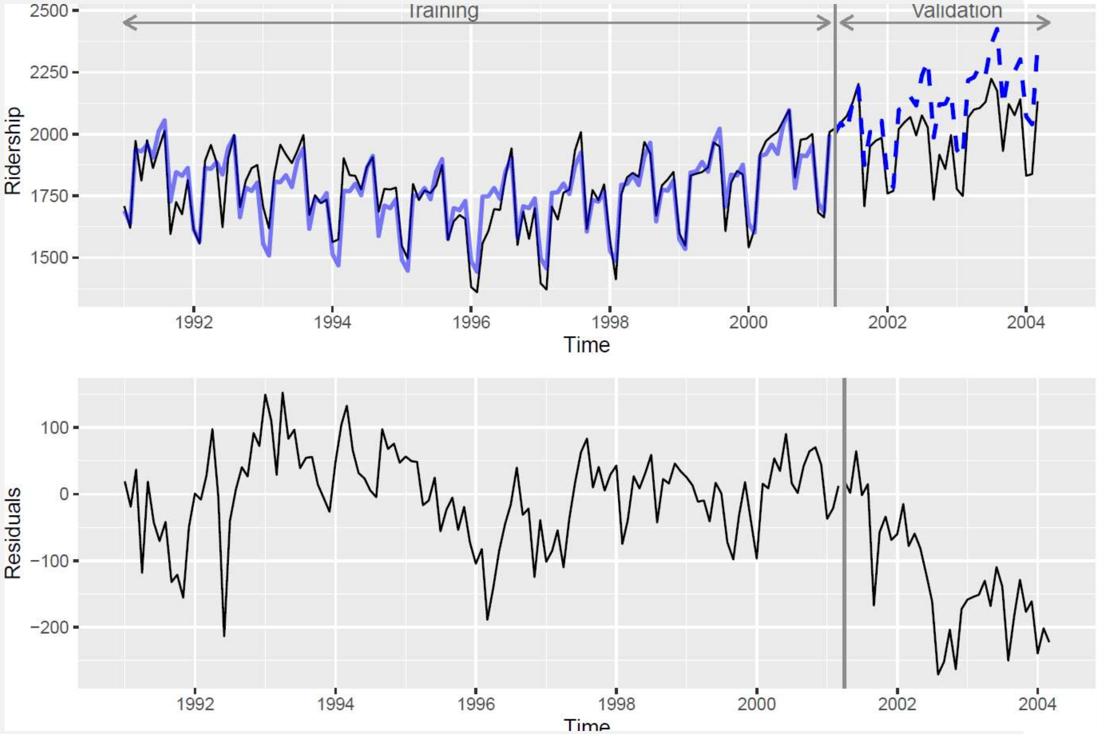


This can be fixed by adding a linear trend

Models with trend and seasonality

- Models that can capture both trend and seasonality
- Amtrak ridership series, for example, exhibits a quadratic trend and monthly seasonality (pick in summer months).
- To capture tend and seasonality, we will use 11 dummy variables, t and t^2 for trend.

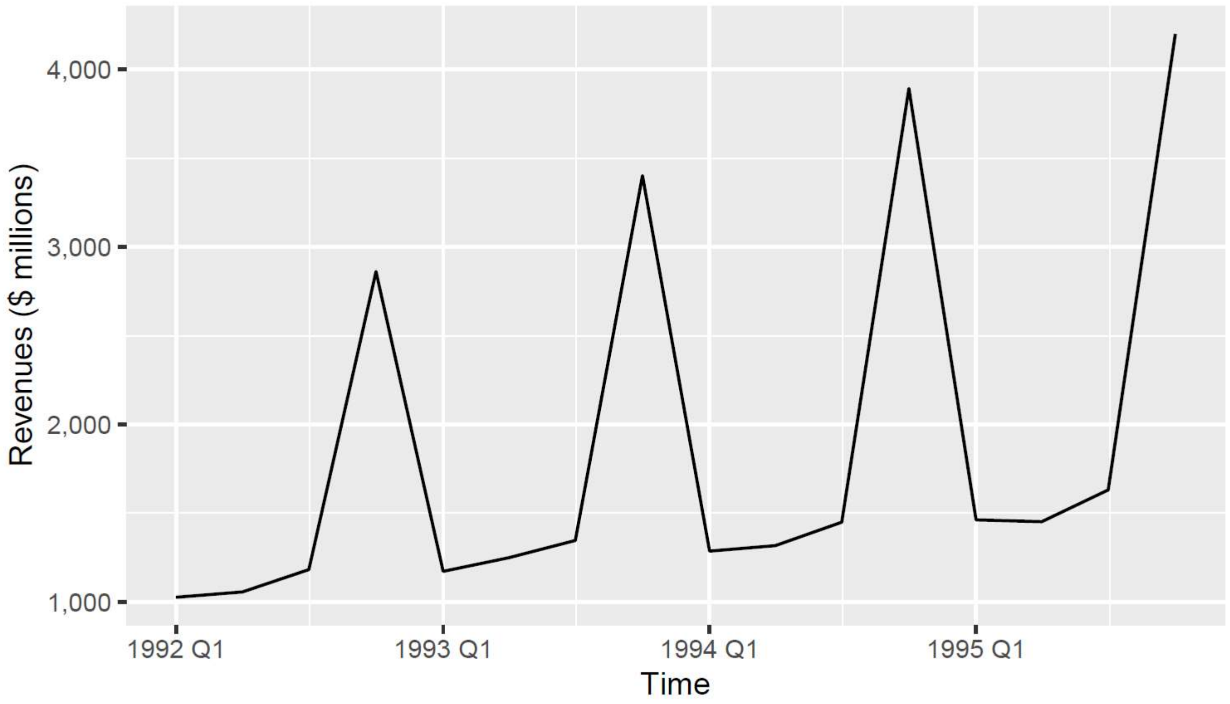
```
> train.lm.trend.season <- train.ridership |>
                                        model(TSLM(Ridership ~ trend() + I(trend()^2) + season()))

> report(train.lm.trend.season)
Series: Ridership
Model: TSLM
Residuals:
      Min       1Q    Median      3Q      Max
 -213.775  -39.363     9.711  42.422  152.187

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    1.697e+03   2.768e+01   61.318  < 2e-16 ***
trend()       -7.156e+00   7.293e-01   -9.812  < 2e-16 ***
I(trend()^2)   6.074e-02   5.698e-03   10.660  < 2e-16 ***
season()year2 -4.325e+01   3.024e+01   -1.430  0.15556
season()year3  2.600e+02   3.024e+01    8.598  6.60e-14 ***
season()year4  2.606e+02   3.102e+01    8.401  1.83e-13 ***
season()year5  2.938e+02   3.102e+01    9.471  6.89e-16 ***
season()year6  2.490e+02   3.102e+01    8.026  1.26e-12 ***
season()year7  3.606e+02   3.102e+01   11.626  < 2e-16 ***
season()year8  4.117e+02   3.102e+01   13.270  < 2e-16 ***
season()year9  9.032e+01   3.102e+01    2.911  0.00437 **
season()year10 2.146e+02   3.102e+01    6.917  3.29e-10 ***
season()year11 2.057e+02   3.103e+01    6.629  1.34e-09 ***
season()year12 2.429e+02   3.103e+01    7.829  3.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.92 on 109 degrees of freedom
Multiple R-squared: 0.8246,   Adjusted R-squared: 0.8037
F-statistic: 39.42 on 13 and 109 DF, p-value: < 2.22e-16
```



Hands-On # 3.2

- Modelling Toys "R" Us Revenues.
- The Figure on the right is a time plot of the quarterly revenues of Toys "R" Us between 1992 and 1995.
- The data is available in the *ToysRUsRevenues.csv* file.

1. Fit a regression model with a linear trend and seasonality. Use the entire series (excluding the last two quarters) as the training period.



Hands-On # 3.2

2. A partial regression model output is shown in the Table on the left (where *season()year2* is the Quarter 2 dummy).
   Use this output to answer the following questions:

i. Mention two statistics (and their values) that measure how well this model fits the training period.

ii. Mention two statistics (and their values) that measure the predictive accuracy of this model.

```
Series: Revenues
Model: TSLM

Residuals:
Min      1Q  Median      3Q     Max
-335.90  -54.29   18.50   63.80  319.24

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     906.75     115.35   7.861 2.55e-05 ***
trend()          47.11      11.26   4.185  0.00236 **
season()year2   -15.11     119.66  -0.126  0.90231
season()year3    89.17     128.67   0.693  0.50582
season()year4  2101.73     129.17  16.272 5.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 168.5 on 9 degrees of freedom
Multiple R-squared: 0.9774,  Adjusted R-squared: 0.9673
F-statistic: 97.18 on 4 and 9 DF, p-value: 2.1289e-07

                   ME      RMSE        MAE       MPE      MAPE      MASE
Training set        0  135.0795   92.53061 0.1614994  5.006914  0.434212
Test set     183.1429  313.6820  254.66667 3.0193814  7.404655  1.195057
```

Hands-On # 3.2

2. A partial regression model output is shown in the Table on the left (where *season()year2* is the Quarter 2 dummy).
   Use this output to answer the following questions:

iii. After adjusting for trend, what is the average difference between sales in Q3 and sales in Q1?

iv. After adjusting for seasonality, which quarter (Q1, Q2, Q3 or Q4) has the highest average sales?

```
Series: Revenues
Model: TSLM

Residuals:
Min      1Q  Median      3Q     Max
-335.90  -54.29   18.50   63.80  319.24

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      906.75     115.35   7.861 2.55e-05 ***
trend()           47.11      11.26   4.185  0.00236 **
season()year2    -15.11     119.66  -0.126  0.90231
season()year3     89.17     128.67   0.693  0.50582
season()year4   2101.73     129.17  16.272 5.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 168.5 on 9 degrees of freedom
Multiple R-squared: 0.9774,  Adjusted R-squared: 0.9673
F-statistic: 97.18 on 4 and 9 DF, p-value: 2.1289e-07
```

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	0	135.0795	92.53061	0.1614994	5.006914	0.434212
Test set	183.1429	313.6820	254.66667	3.0193814	7.404655	1.195057