Class 6a: Time Series

Business Forecasting

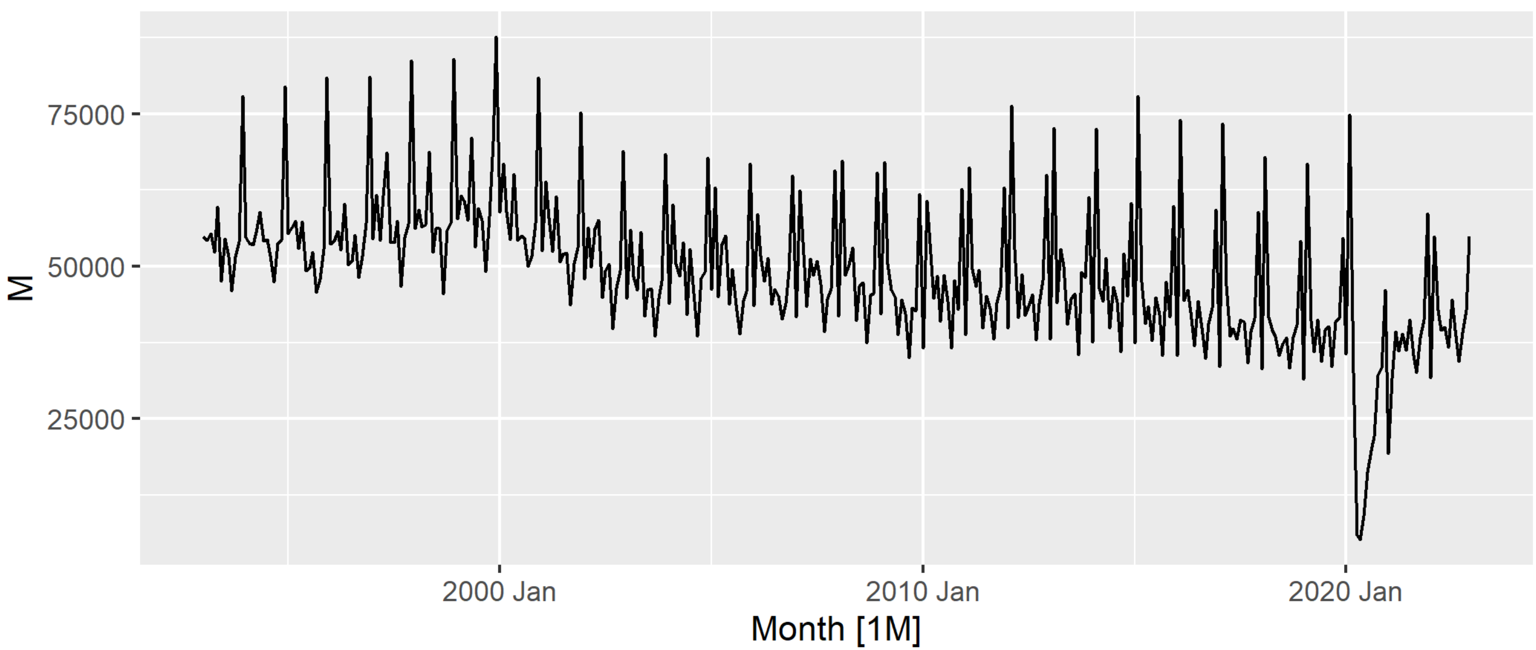
Roadmap:

1. Components of time series
2. Patterns of correlation in time series
3. Simple forecasting methods
4. Evaluating forecasts
5. Time series decomposition
6. Forecasting with time series decomposition

Example

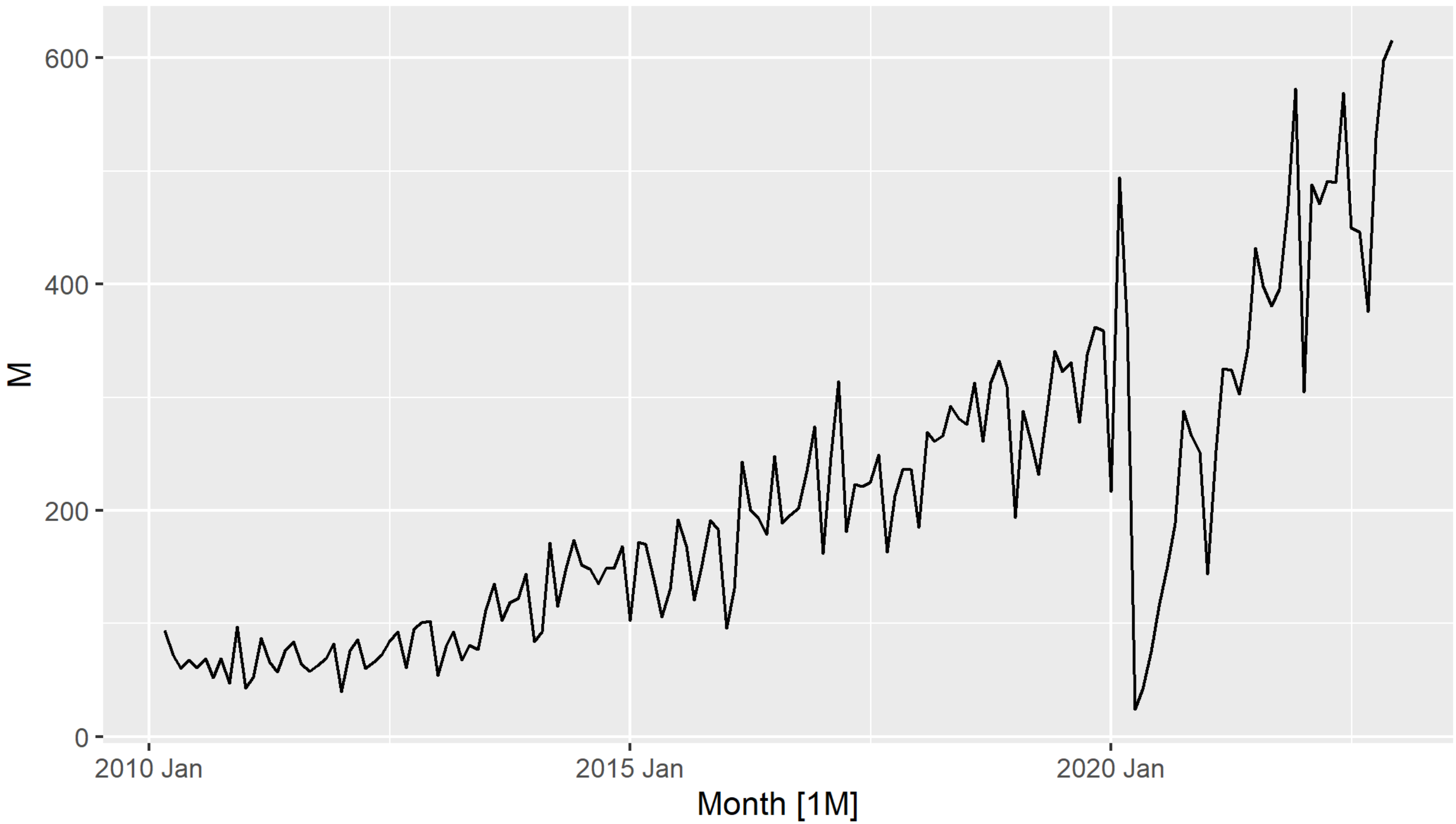
- Suppose you wonder if you should go into the wedding business.
- You need to predict whether there is potential for work
- So you look at evolution in the number of weddings across years

Heterosexual Marriages in Mexico



What patterns can you identify in that time series?

Same Sex Marriages in Mexico

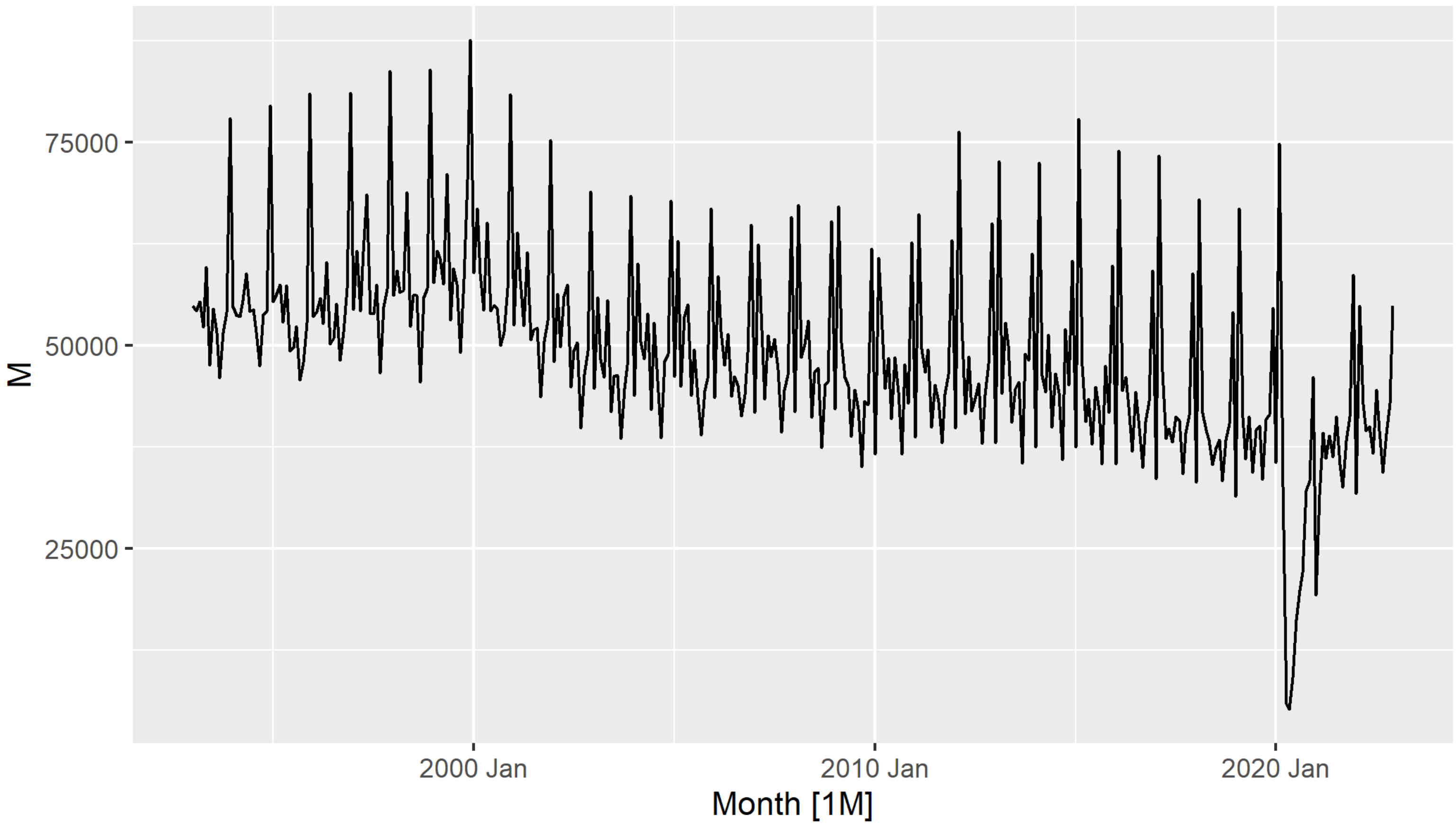


Going into gay marriage business is probably a better idea!

Components

1. **Trend** - long term change in the level of data, positive or negative.
   - If flat, we call the data stationary
   - Formally, the mean, variance, and autocorrelation does not depend on time

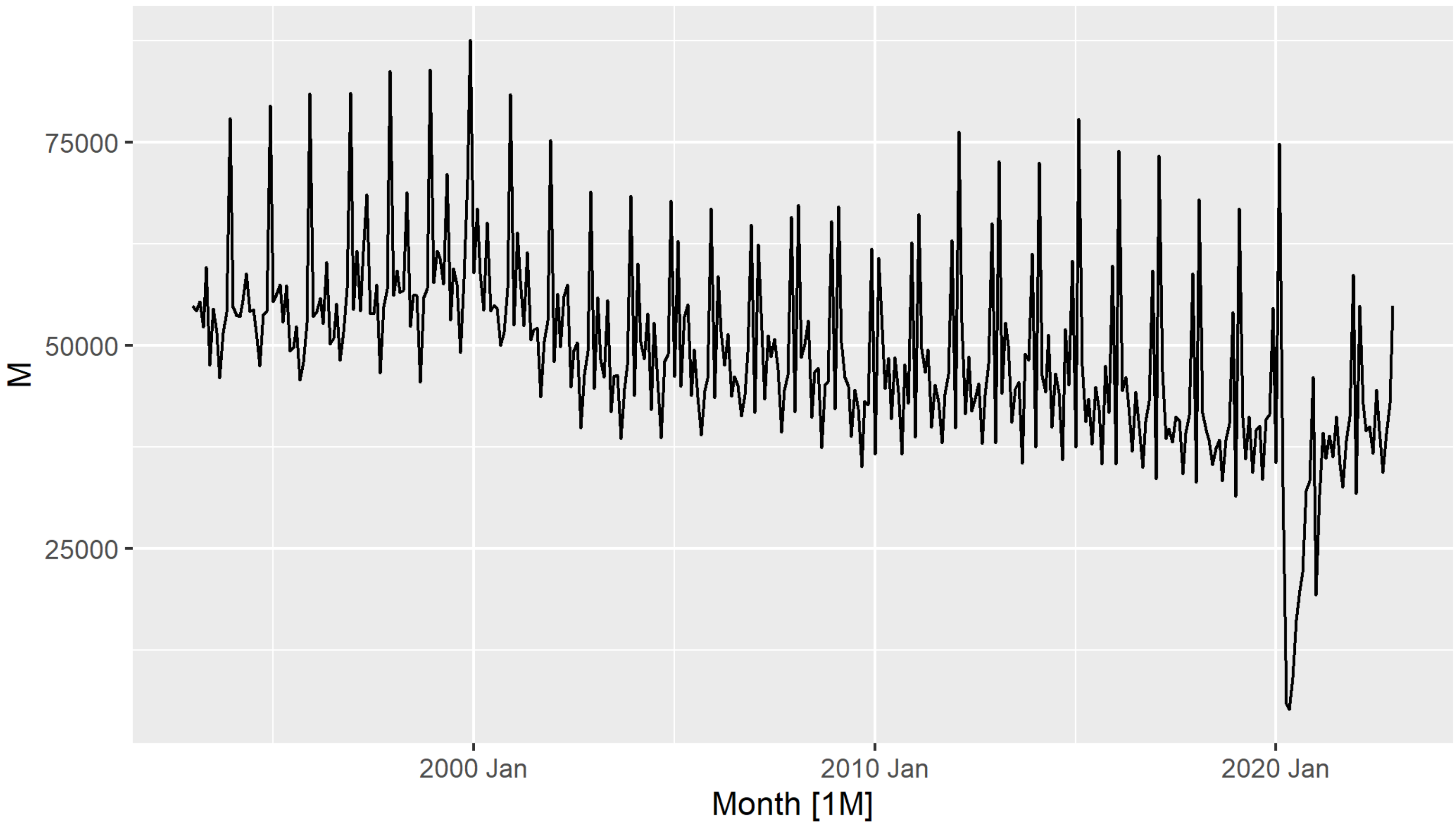
Heterosexual Marriages in Mexico



Components

1. **Trend** - long term change in the level of data, positive or negative.
    - If flat, we call the data stationary
    - Formally, the mean, variance, and autocorrelation does not depend on time
2. **Seasonal pattern**: Variation in level that repeats at the same time each period
    - If there is seasonality, data is not stationary

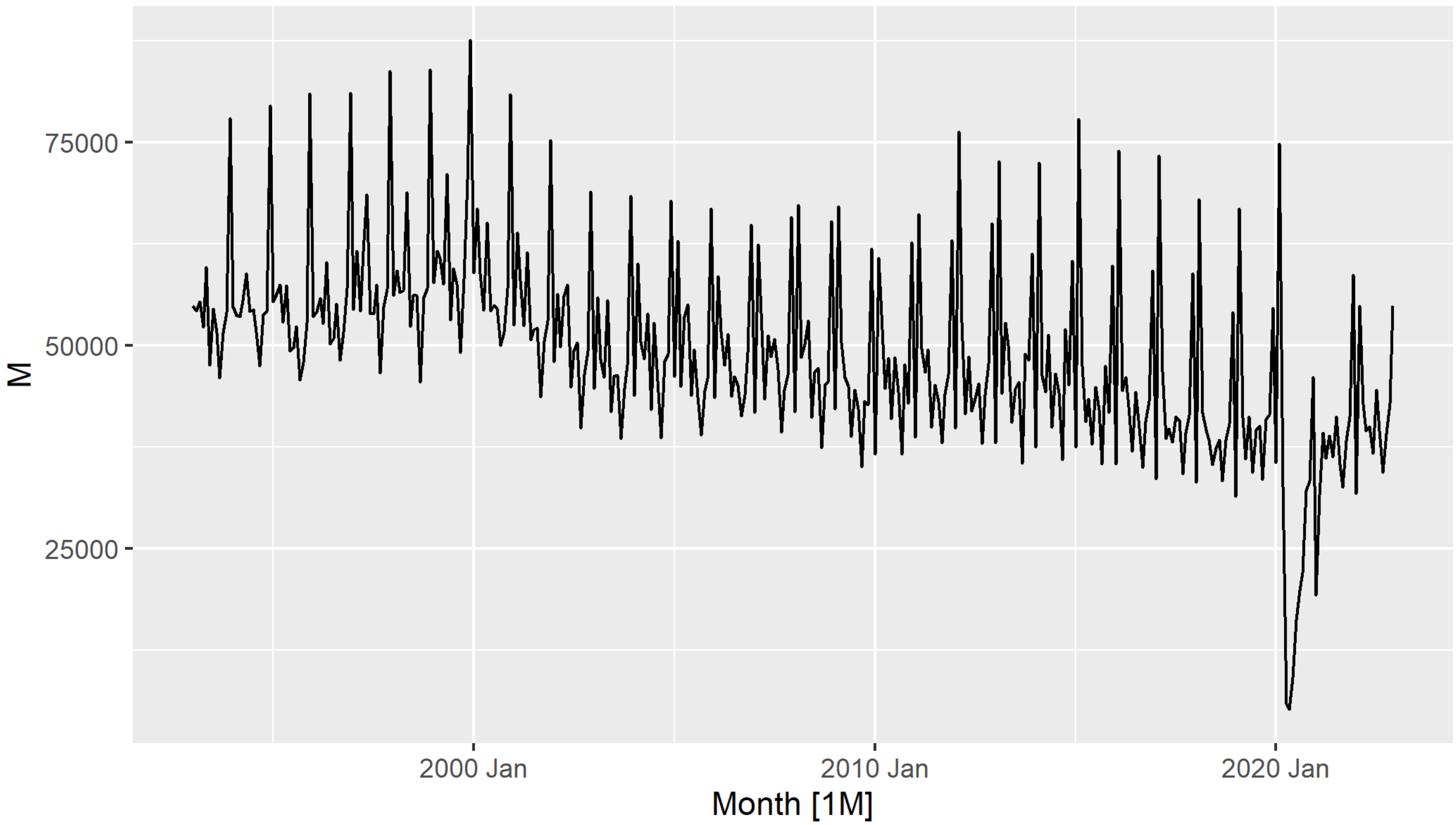
Heterosexual Marriages in Mexico



Components

1. **Trend**: Long term change in the level of data, positive or negative.
    - If flat, we call the data stationary
    - Formally, the mean, variance, and autocorrelation does not depend on time
2. **Seasonal pattern**: Variation in level that repeats at the same time each period
    - If there is seasonality, data is not stationary
3. **Cyclical pattern**: Wavelike upward and downward movements along the trend. Not always the same length, not always the same time of year
    - Different from seasonality which always happens at the same time and has same length
    - Often related to business cycles

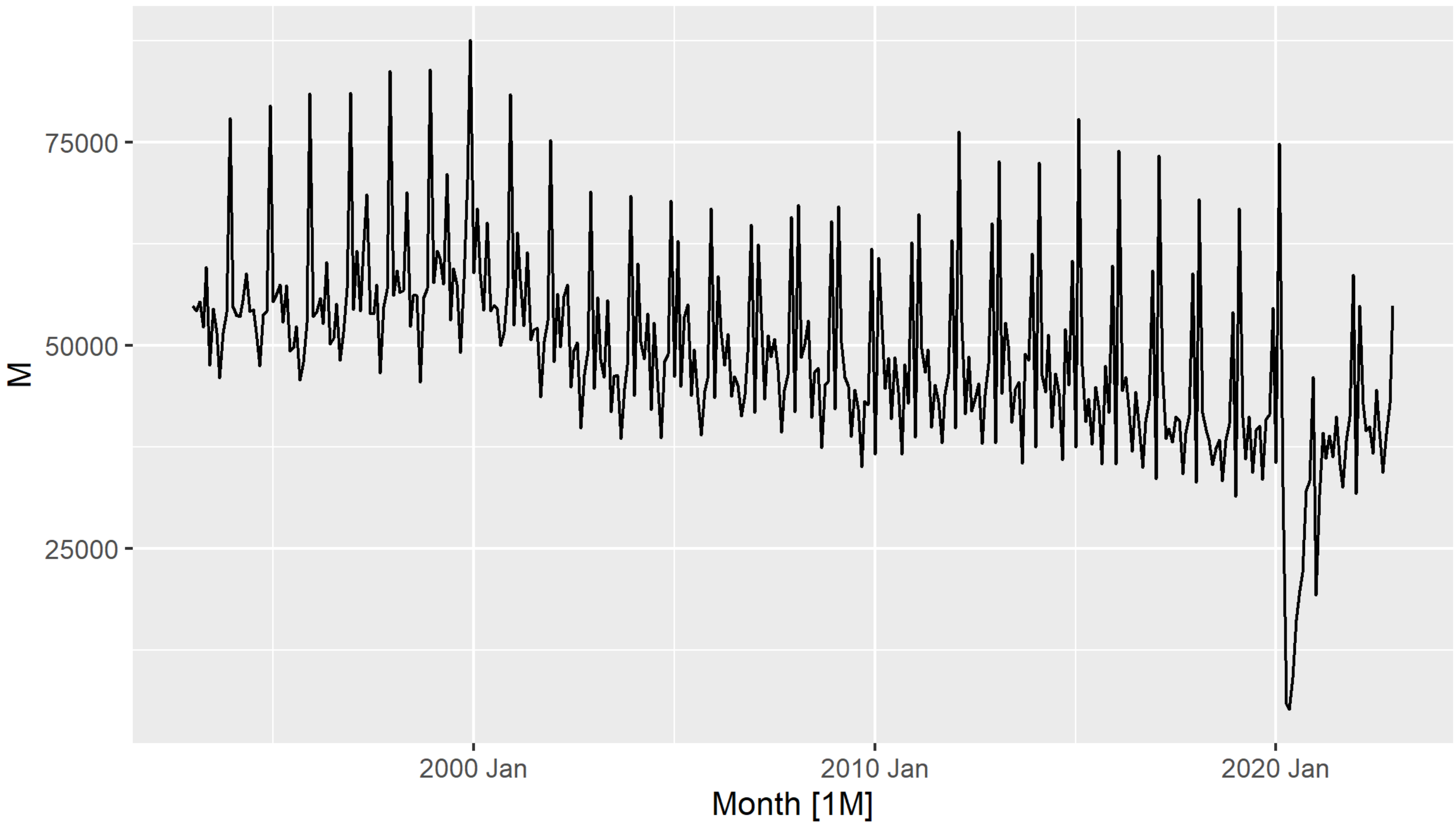
Heterosexual Marriages in Mexico



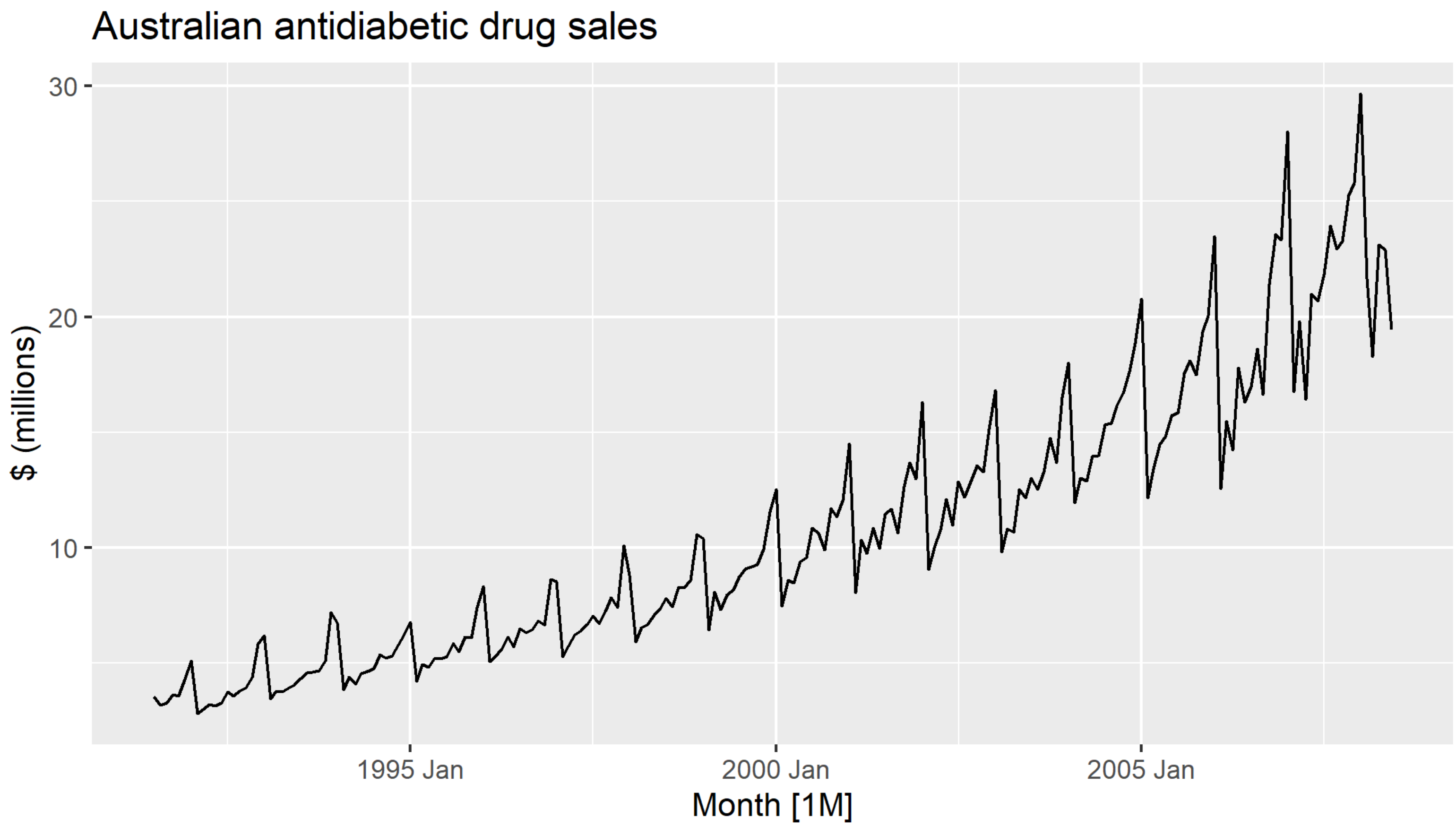
Components

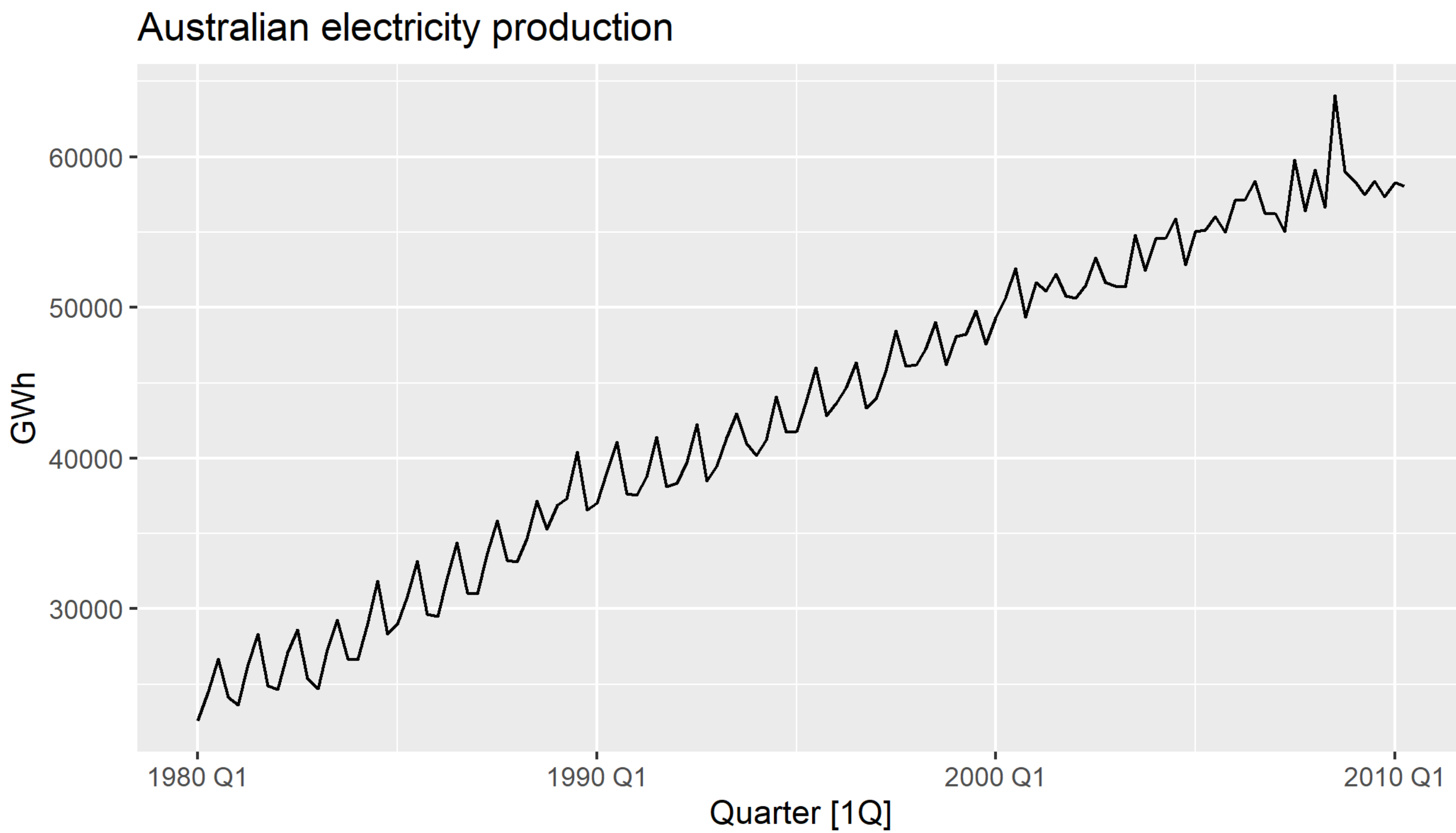
1. **Trend**: Long term change in the level of data, positive or negative.
    - If flat, we call the data stationary
    - Formally, the mean, variance, and autocorrelation does not depend on time
2. **Seasonal pattern**: Variation in level that repeats at the same time each period
    - If there is seasonality, data is not stationary
3. **Cyclical pattern**: Wavelike upward and downward movements along the trend. Not always the same length, not always the same time of year
    - Different from seasonality which always happens at the same time and has same length
    - Often related to business cycles
4. **Random components**: Can't be attributed to other parts of the model. The most difficult to predict

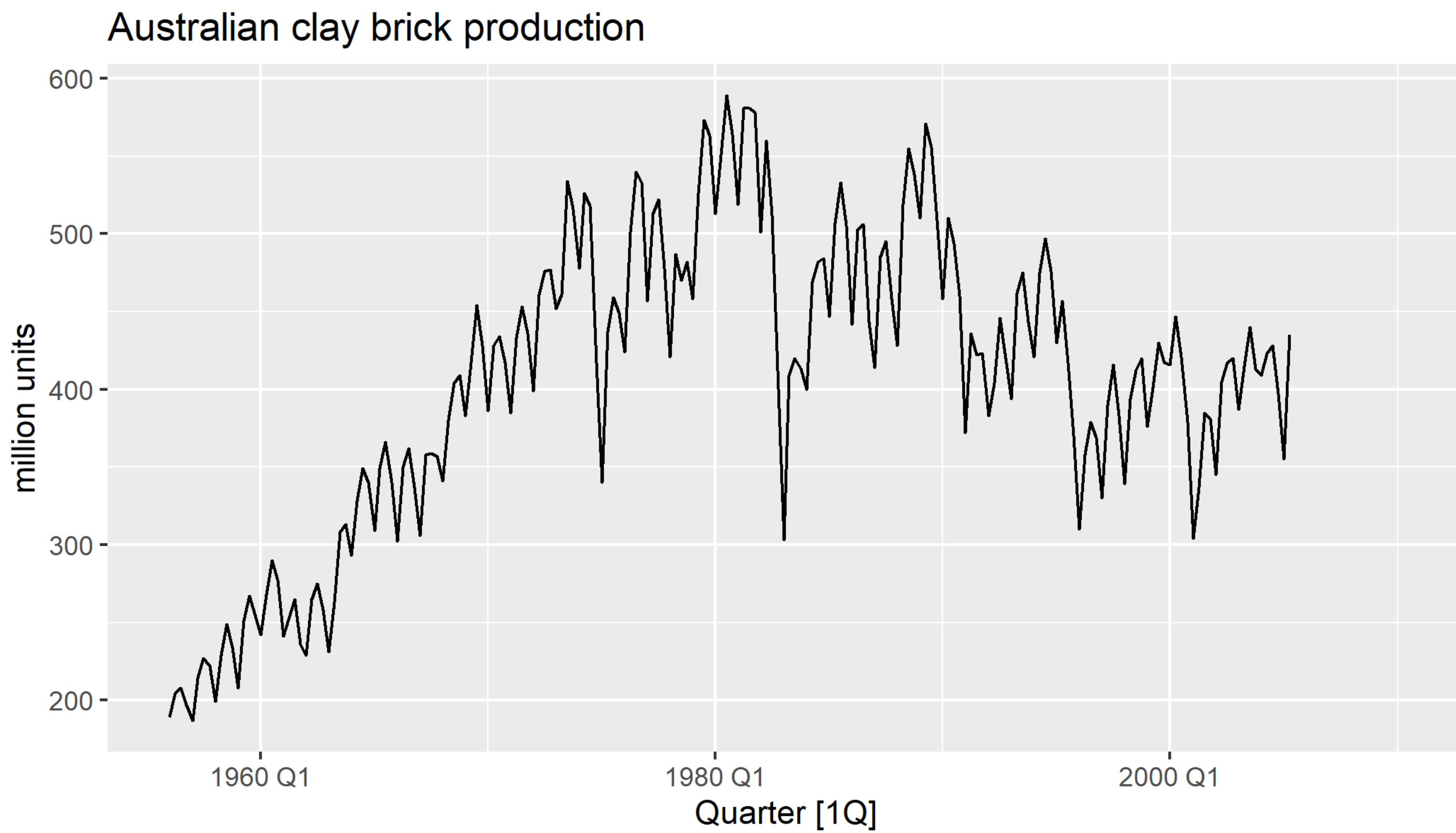
Heterosexual Marriages in Mexico

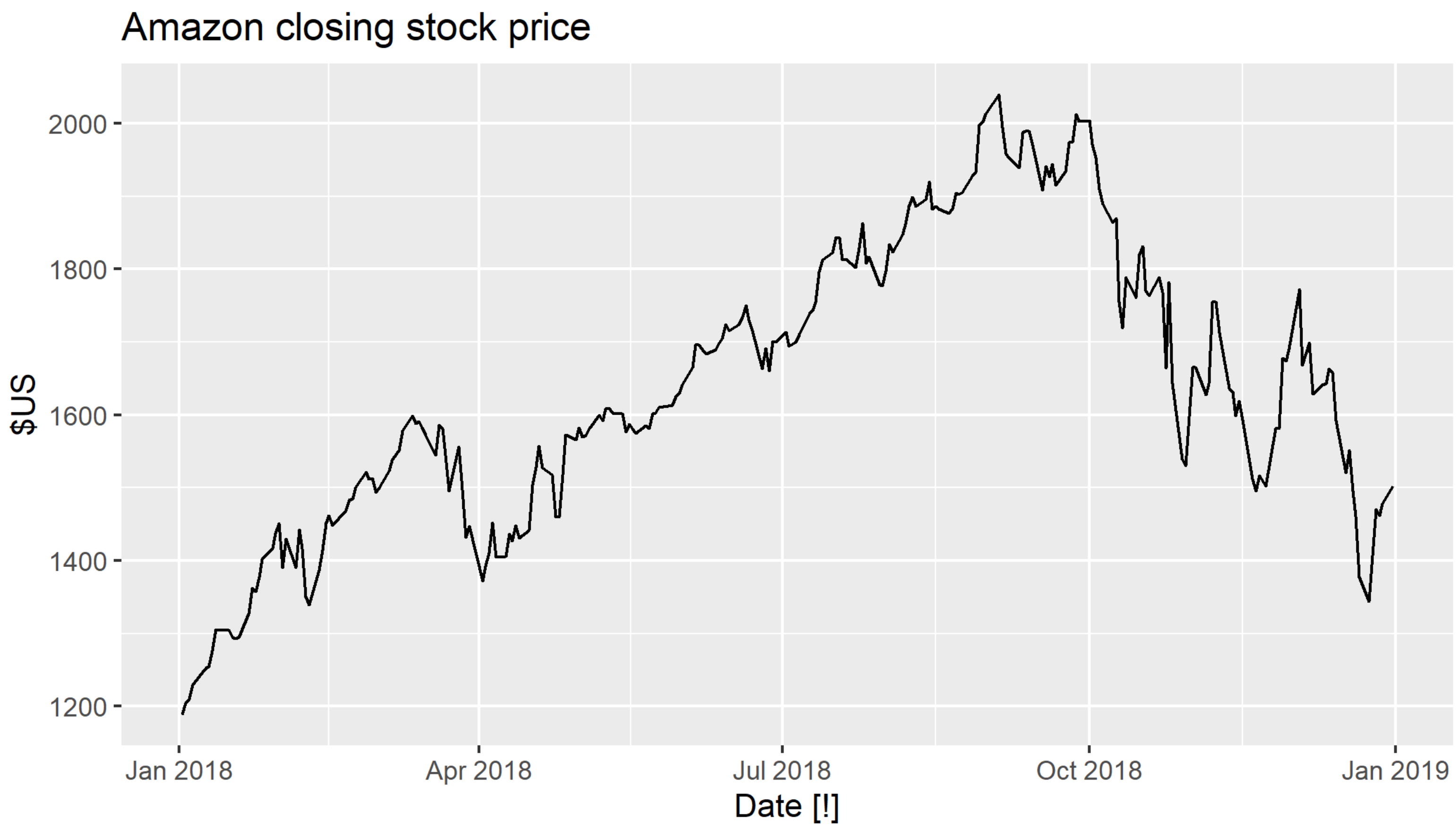


Some other examples




Australian electricity production
GWh
60000
50000
40000
30000
1980 Q1
1990 Q1
2000 Q1
2010 Q1
Quarter [1Q]


Australian clay brick production
600
500
400
300
200
million units
1960 Q1
1980 Q1
2000 Q1
Quarter [1Q]


Amazon closing stock price
$US
2000
1800
1600
1400
1200
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
Date [!]

Autocorrelation

- Can past values predict future values?
- Yes, if they are correlated
- We will measure **Autocorrelation**:
  - Are values in previous period correlated with values in the next period?
  - So between y_t and y_{t-1},\$yt\$ *and* \`(*y*{t-2})\` etc

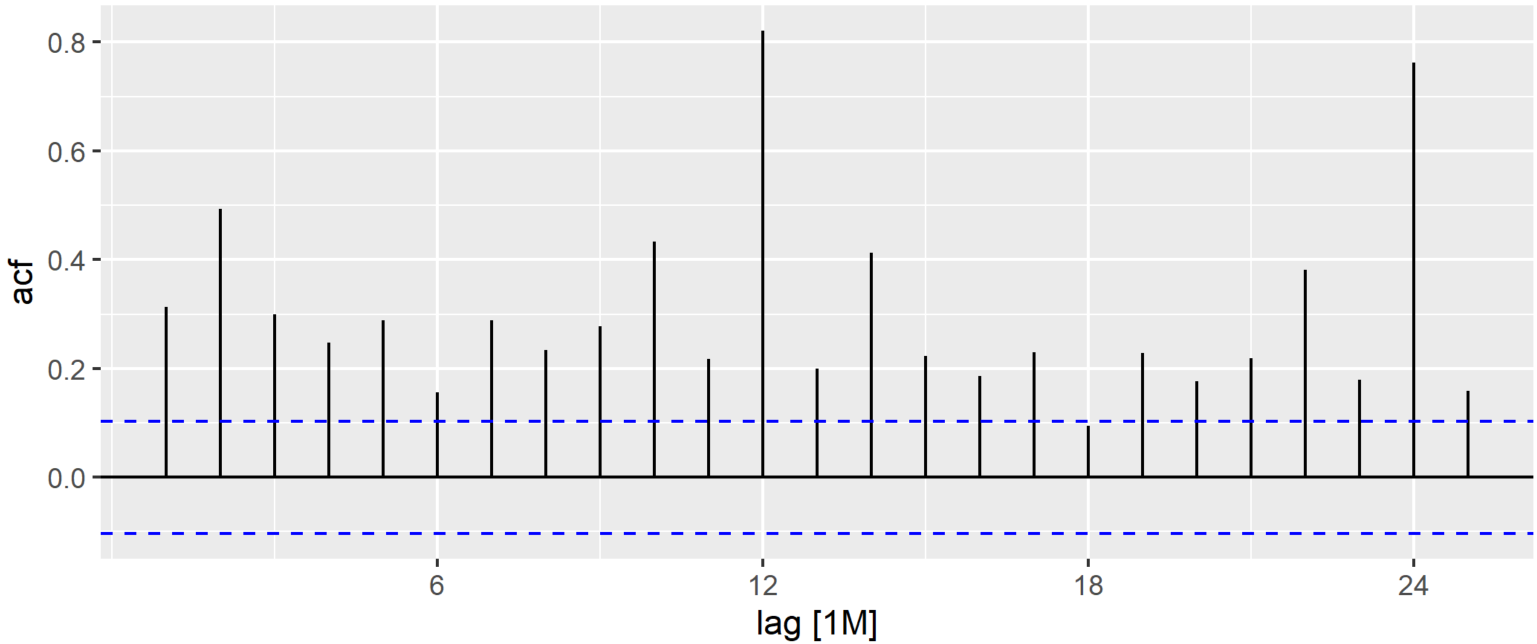
$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

```
## # A tsibble: 360 x 5 [1M]
##       Month     M Lag1_M Lag2_M Lag3_M
##       <mth> <dbl>  <dbl>  <dbl>  <dbl>
##  1 1993 Jan 54850     NA     NA     NA
##  2 1993 Feb 54271  54850     NA     NA
##  3 1993 Mar 55350  54271  54850     NA
##  4 1993 Apr 52268  55350  54271  54850
##  5 1993 May 59671  52268  55350  54271
##  6 1993 Jun 47557  59671  52268  55350
##  7 1993 Jul 54503  47557  59671  52268
##  8 1993 Aug 51534  54503  47557  59671
##  9 1993 Sep 46000  51534  54503  47557
## 10 1993 Oct 51590  46000  51534  54503
## # ℹ 350
```

We can calculate the values for marriage data:

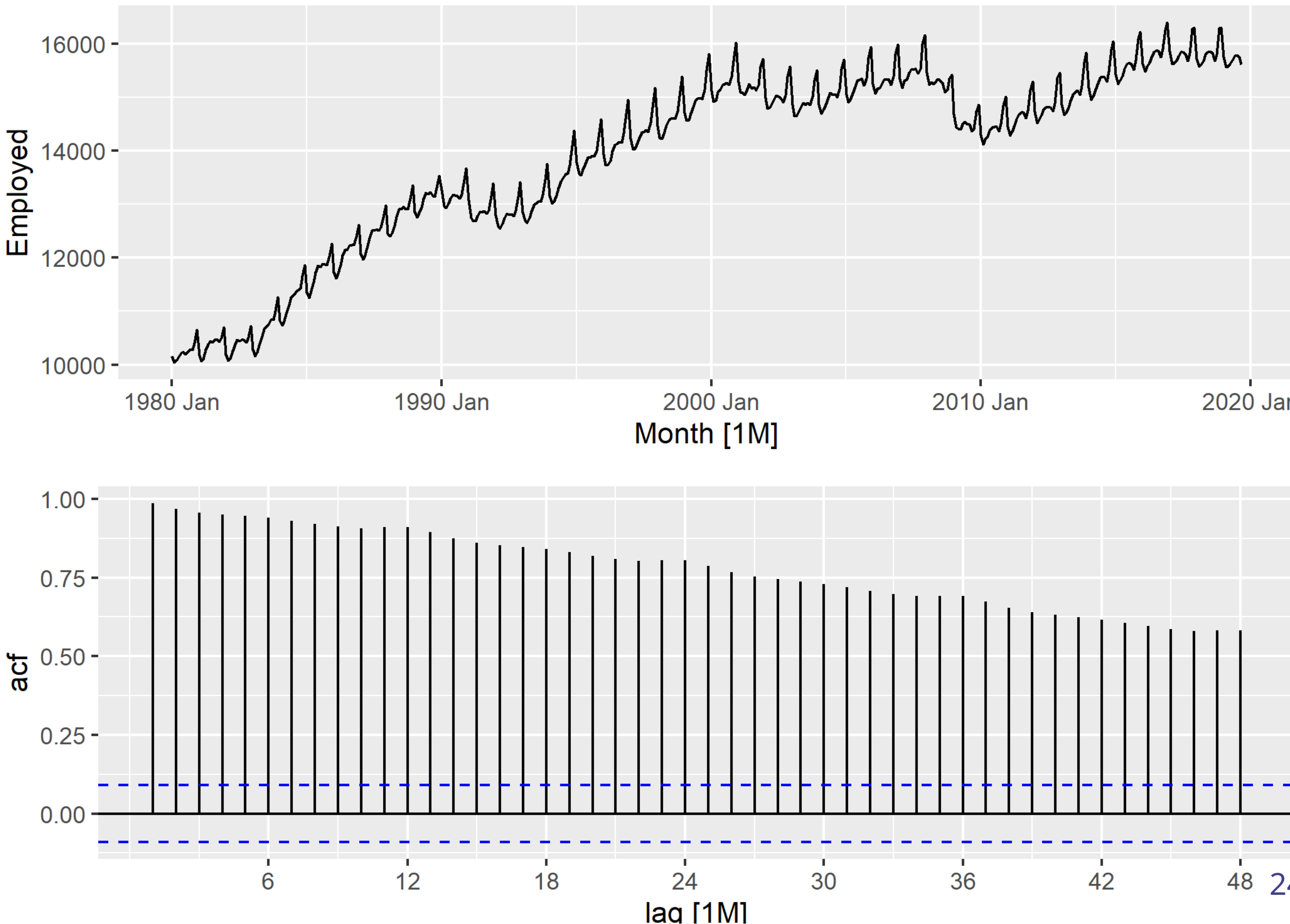
lag	1.0000000	2.0000000	3.0000000	4.0000000	5.0000000	6.0000000
acf	0.3126539	0.4934558	0.2992763	0.2474031	0.2879573	0.1557756

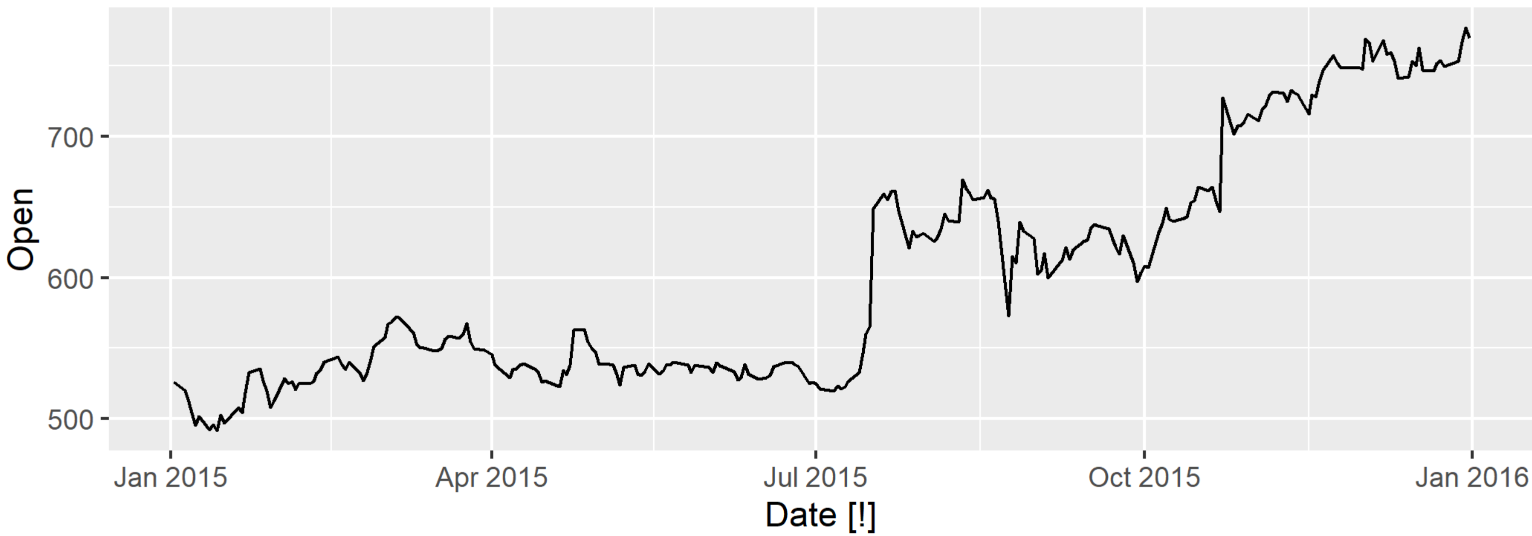
And plot the Autocorrelation Function (ACF) on a correlogram:

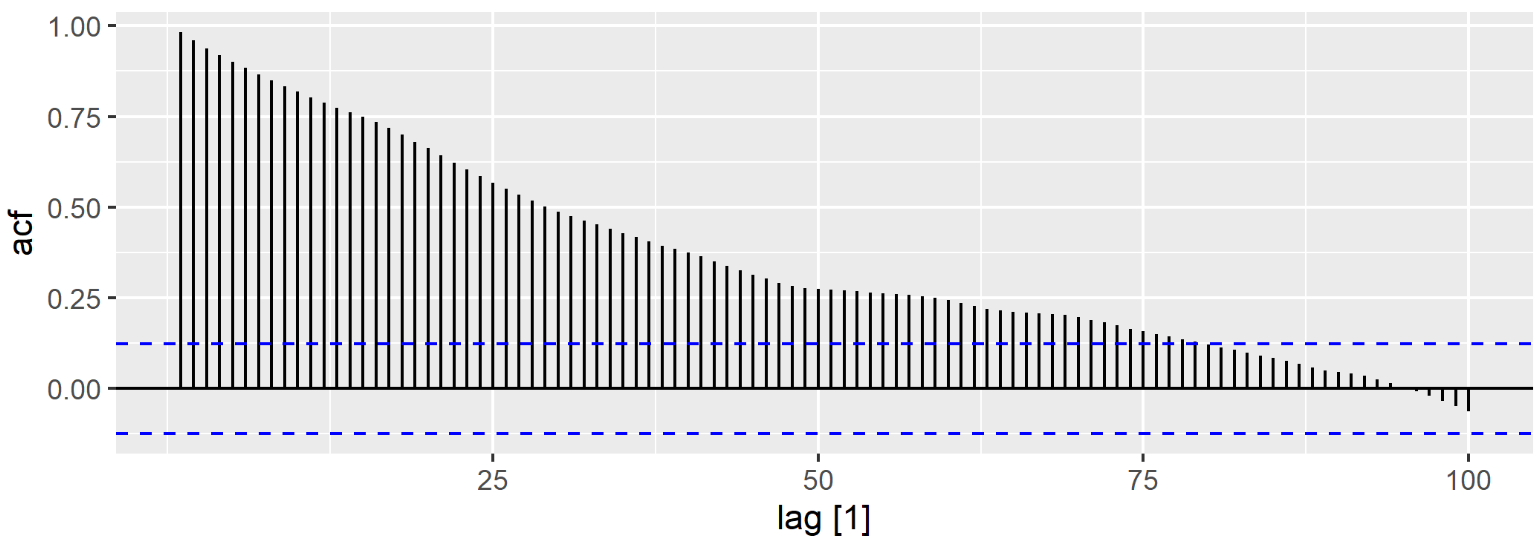


- Why high values at 12 and 24 lag?

Some other examples:



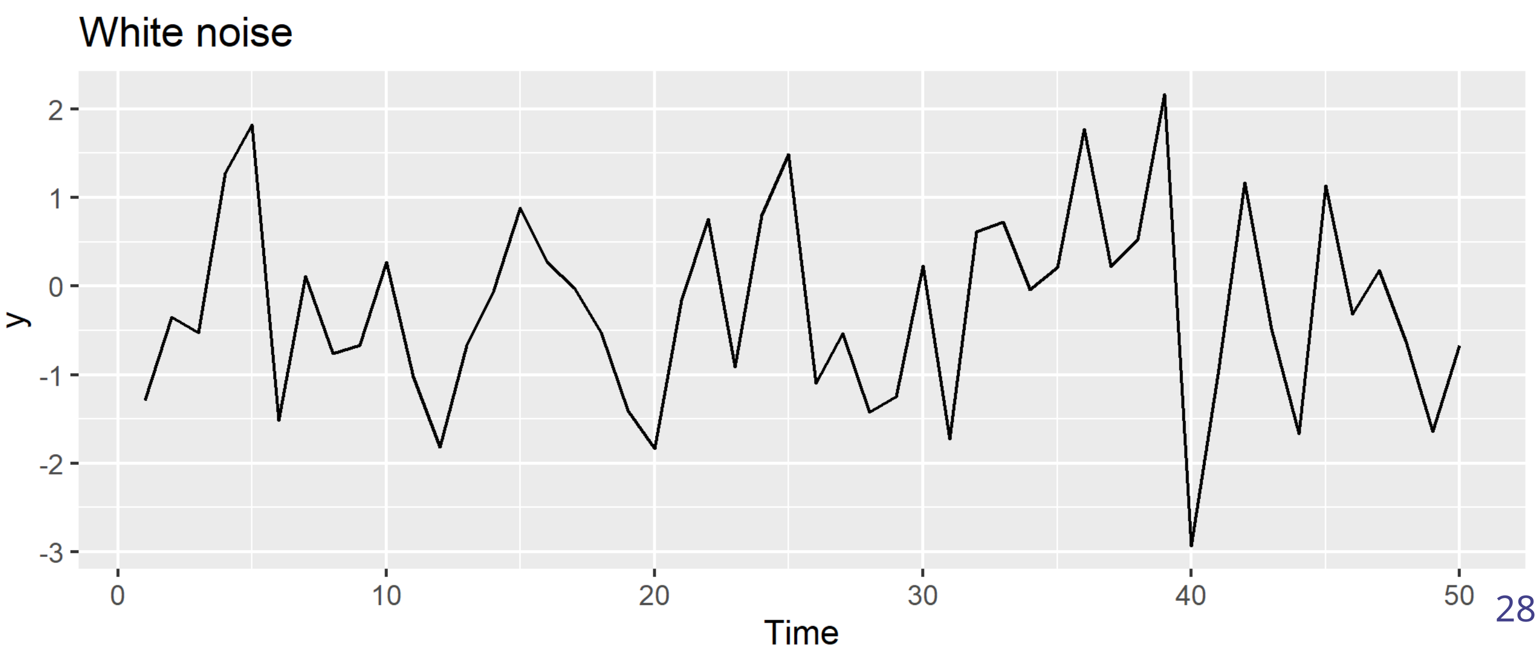




- Shock persists for a long time
- If stationary, shocks should not persist, autocorrelation should decay quickly

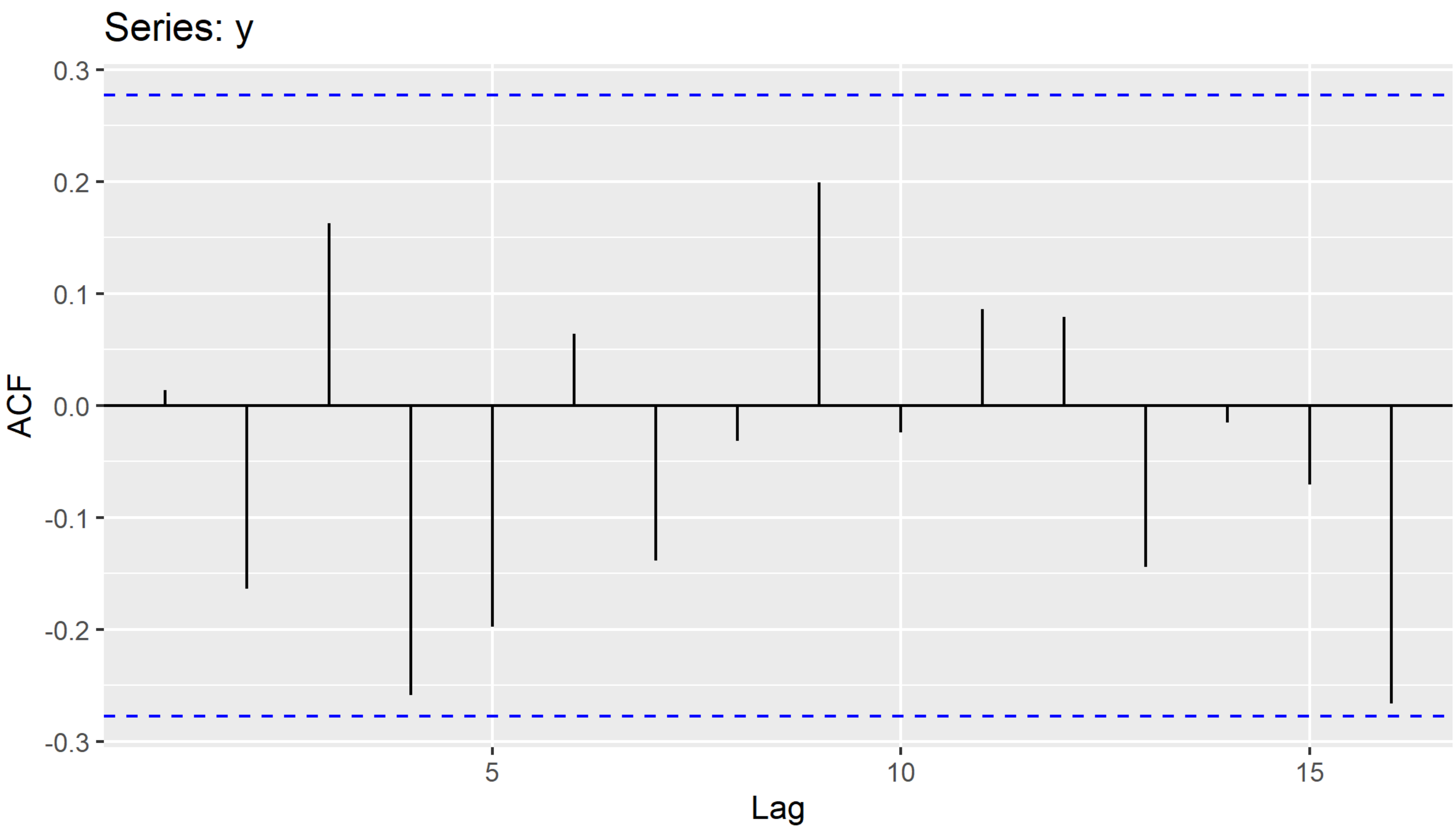
Autocorrelation

- How do we know that the correlation is significant and not just sampling randomness?
- Test:
  - $H_0 : \rho_k = 0$ or data is white noise
  - $H_A : \rho_k \neq 0$
- What is **White Noise**?



White Noise

Autocorrelation of white noise



Test

- Intuitively:
  1. We will calculate test statistic
  2. Figure out how likely to obtain such value if data was White Noise
  - If test statistic is big, it's unlikely to come from White Noise, so we reject null

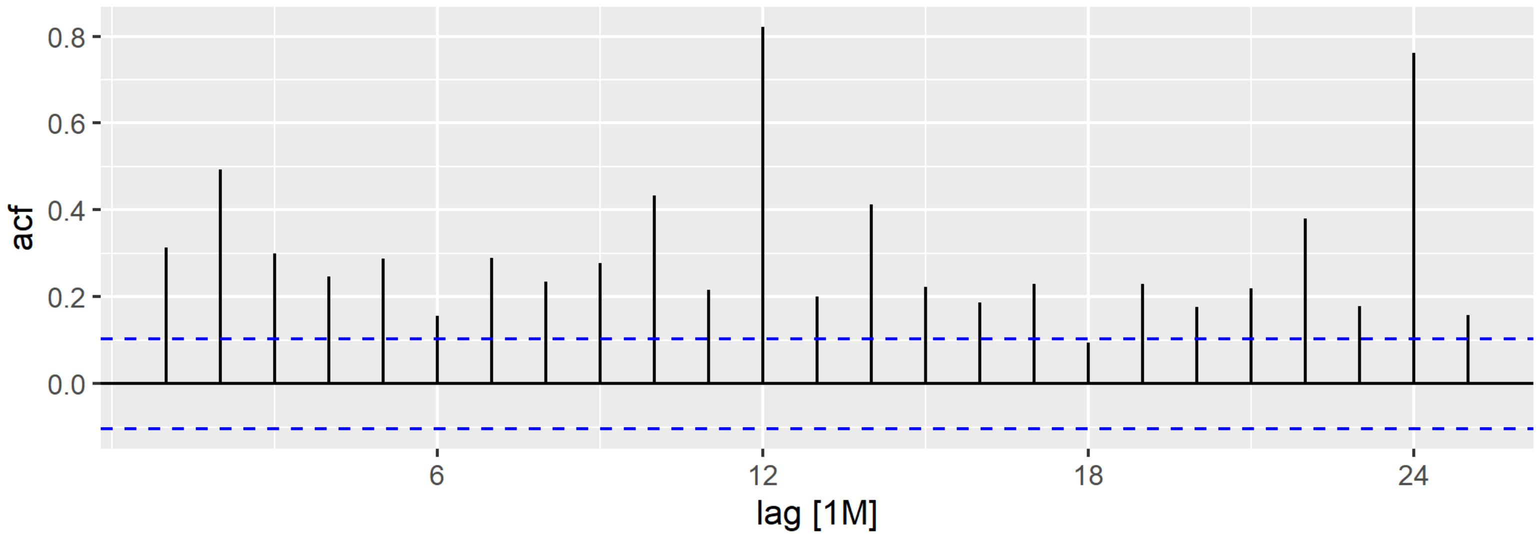
$$t_{test} = \frac{\hat{\rho}_k - 0}{1/\sqrt{n-k}}$$

- Compare it to t distribution with t_{n-k} degrees of freedom
- Rule of thumb for larger datasets: reject at 95% if:

$$|\hat{\rho}_k| > \frac{2}{\sqrt{n}}$$

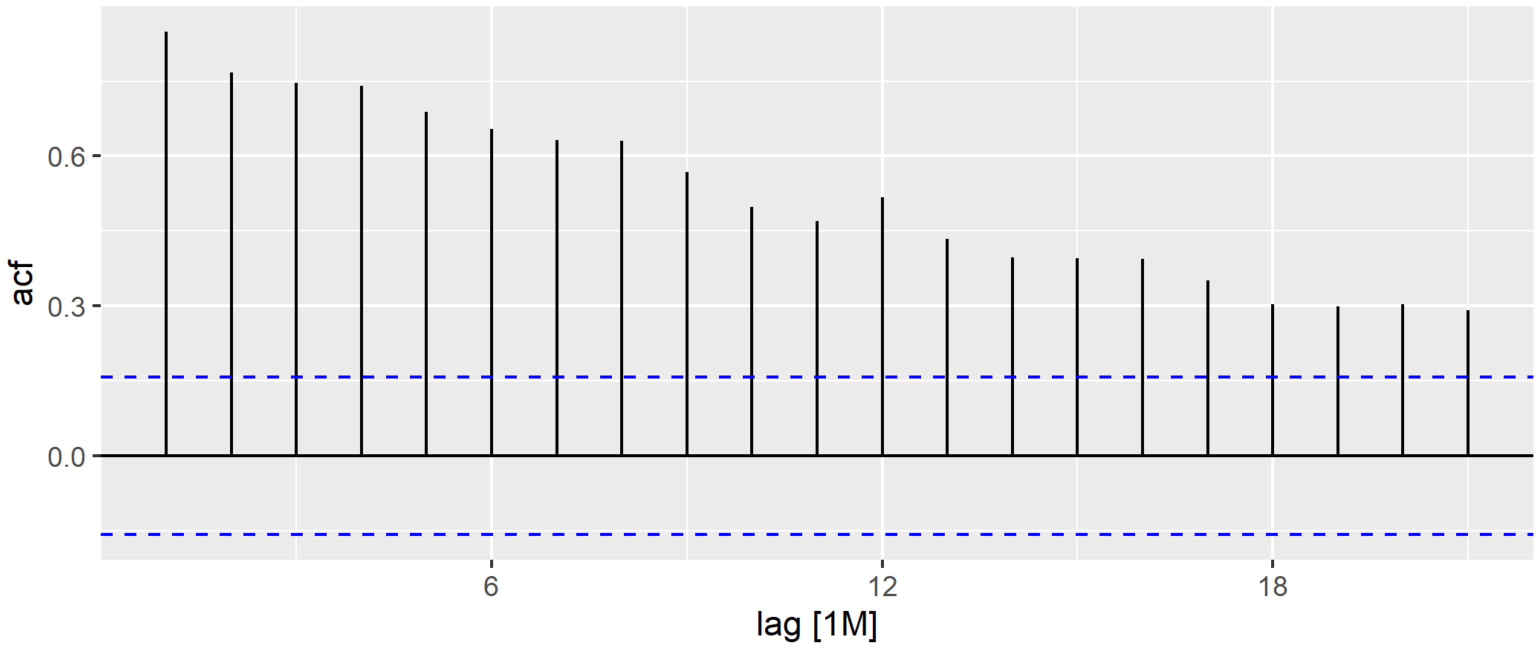
Confidence bands

- We can compute confidence bands such that if $\hat{\rho}_k$ is within these bands, it's not significant.
- In our data on straight marriage, n=360
- If data is white noise, autocorrelations should not cross 0.1054



- The more observation you have, the better you are at detecting autocorrelation

Gay marriages



- Is there a way to transform the data, so it's stationary?

First differencing

- Take the first differences

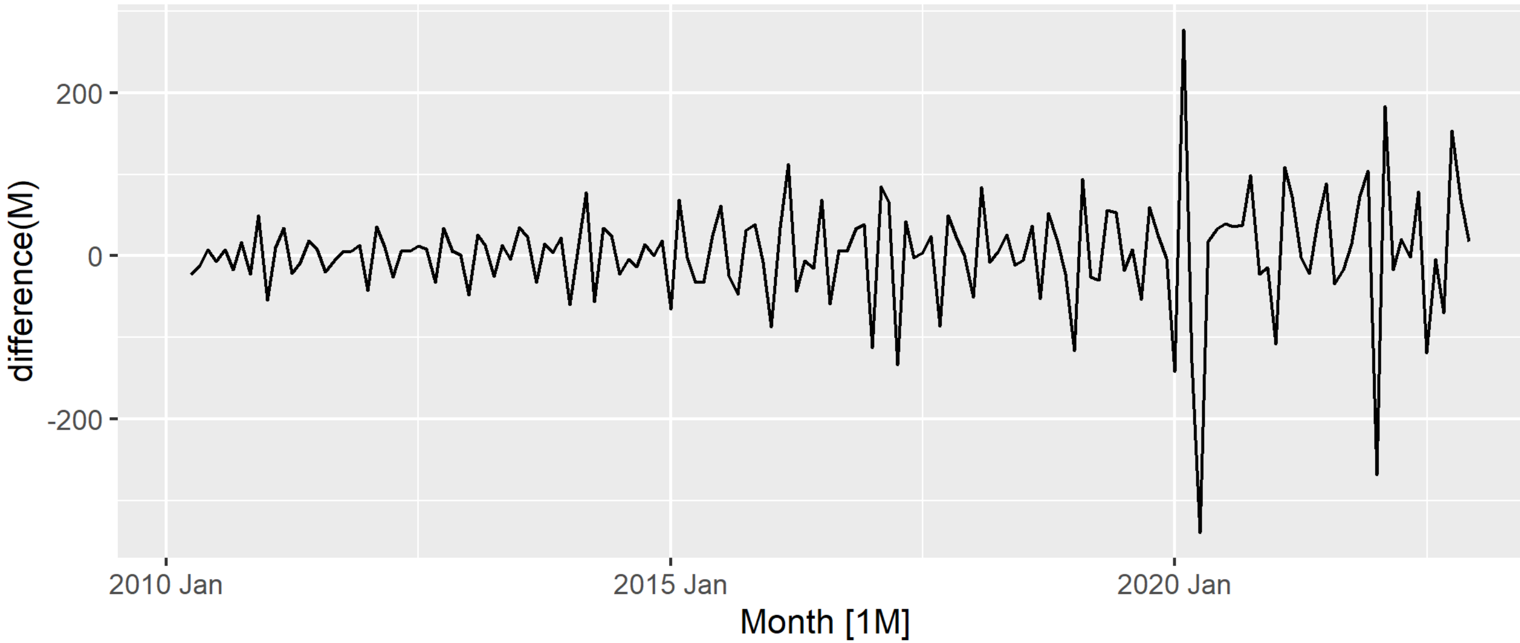
$$\Delta y_t = y_t - y_{t-1}$$

- First differences approximate how much data growth in each period
- If trend is linear, this variable should have more or less constant mean

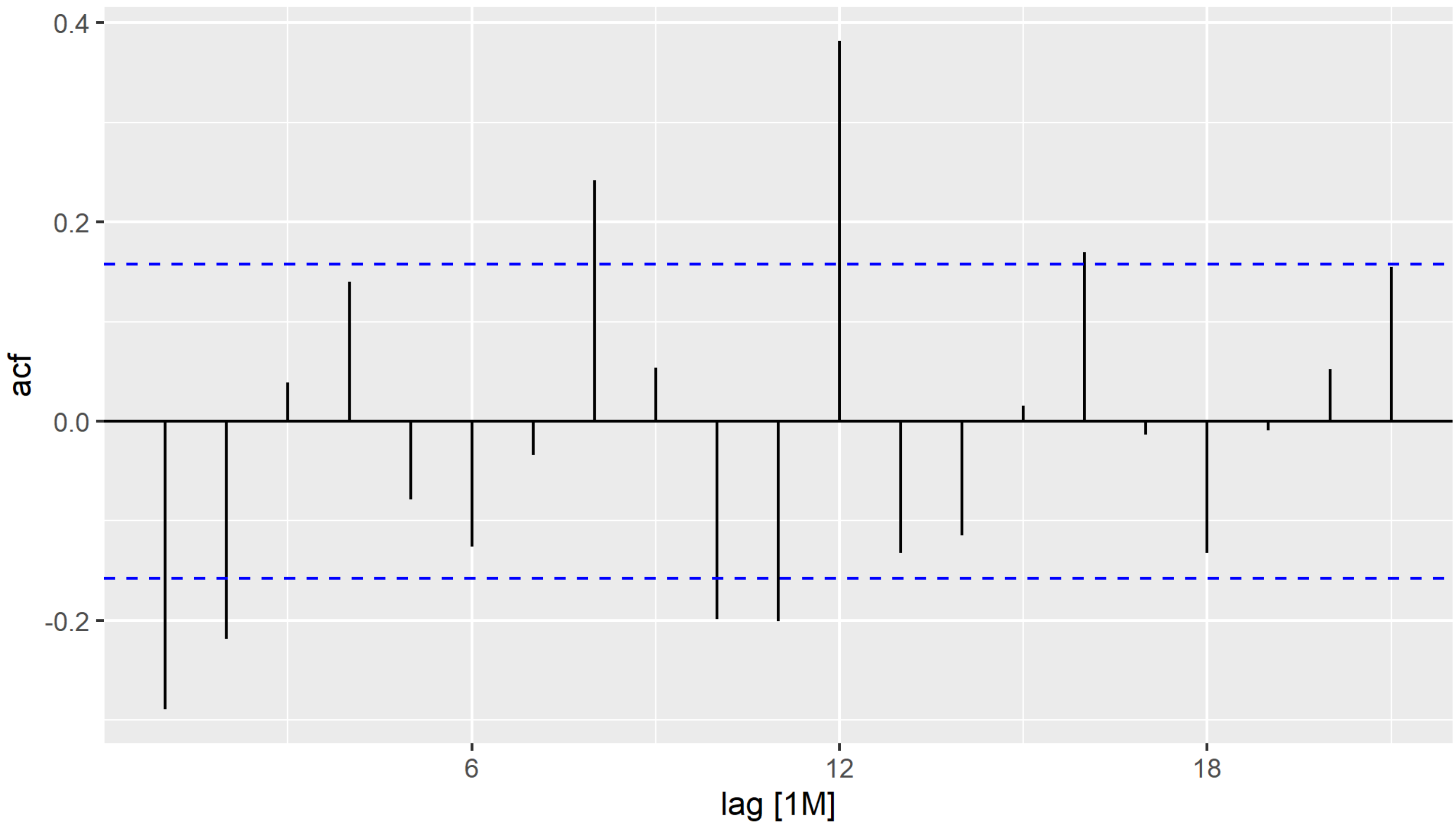
First differencing

```
## # A tsibble: 154 x 3 [1M]
##       Month     M Diff_M
##       <mth> <int>  <int>
##  1 2010 Mar    94     NA
##  2 2010 Apr    72    -22
##  3 2010 May    60    -12
##  4 2010 Jun    68      8
##  5 2010 Jul    61     -7
##  6 2010 Aug    69      8
##  7 2010 Sep    52    -17
##  8 2010 Oct    69     17
##  9 2010 Nov    47    -22
## 10 2010 Dec    97     50
## # ℹ 144 more rows
```

Is transform data stationary?



- Does it have constant mean?
- What about constant variance?
- What about autocorrelation?

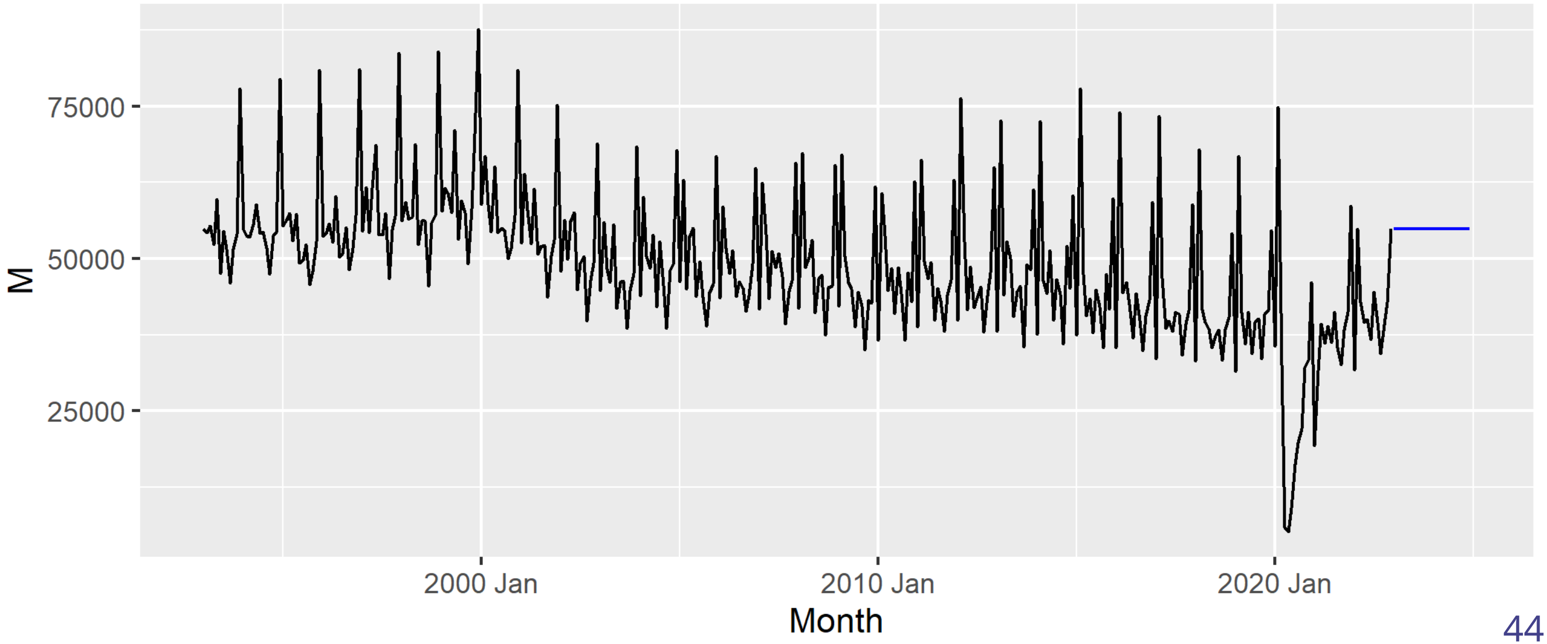

0.4
0.2
0.0
-0.2
acf
6
12
18
lag [1M]

Simple forecasting methods

Naive Model

The simplest way to forecast is to assume that it will be the same as previous period

- One step forecast: $\hat{y}_{T+1|T} = y_T$
- h-step forecast: $\hat{y}_{T+h|T} = y_T$



Simple forecasting methods

What is the confidence interval for such prediction?

- We need to know the variance of the forecast error
- What is **Forecast Error**?

$$e_t = y_{T+h} - \hat{y}_{T+h|T}$$

- It's the difference between what we forecasted and what actually happened onece we observe this datapoint

- Also known as out-of-sample error
- We only used observations up to point T when estimating this model!
- Different from **Fitted Residuals**!

$$u_t = y_t - \hat{y}_t$$

These are fitted residuals for observations that we used in estimation.

Simple forecasting methods

- In the simplest model, and one step ahead, residuals and forecast errors are similar.
- So we can approximate the standard deviation of e_t with standard deviation of u_t in this naive model.
- Let σ_h be the h-step forecast error.
- We will assume:

$$\sigma_1 = \sigma_u$$

  so the standard deviation of the one step ahead forecast is the same as the standard deviation of the residuals

- This gives us the following confidence interval for one step ahead error:

$$CI_{95} = \hat{y}_{T+1|T} \pm 1.96\hat{\sigma}_u$$

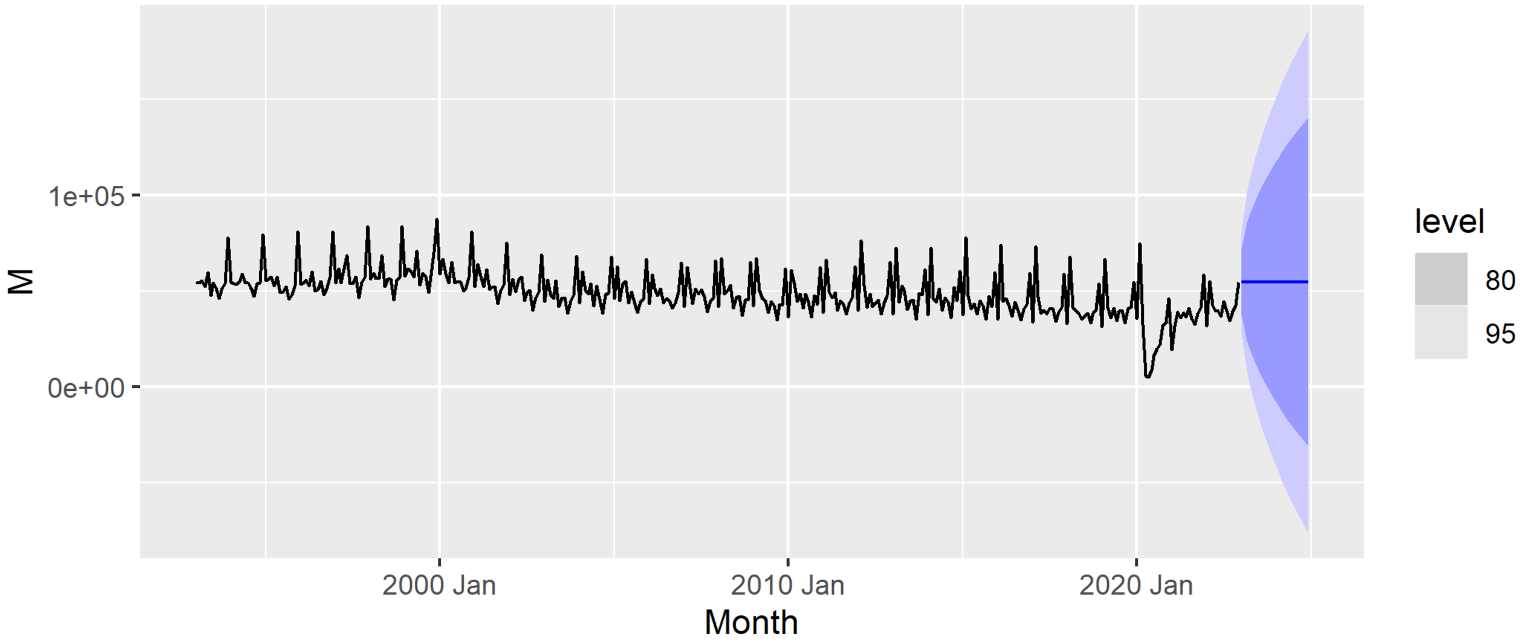
For longer horizon, forecast error in naive forecast is:

- Let $\sigma_h = \sigma_u\sqrt{h}$ be the sd of h-step forecast error, and

Simple forecasting methods

```
## [1] 13683.56

##     Point Forecast      Lo 80    Hi 80    Lo 95    Hi 95
## 361          54887 37375.25 72398.75 28105.09 81668.91
```



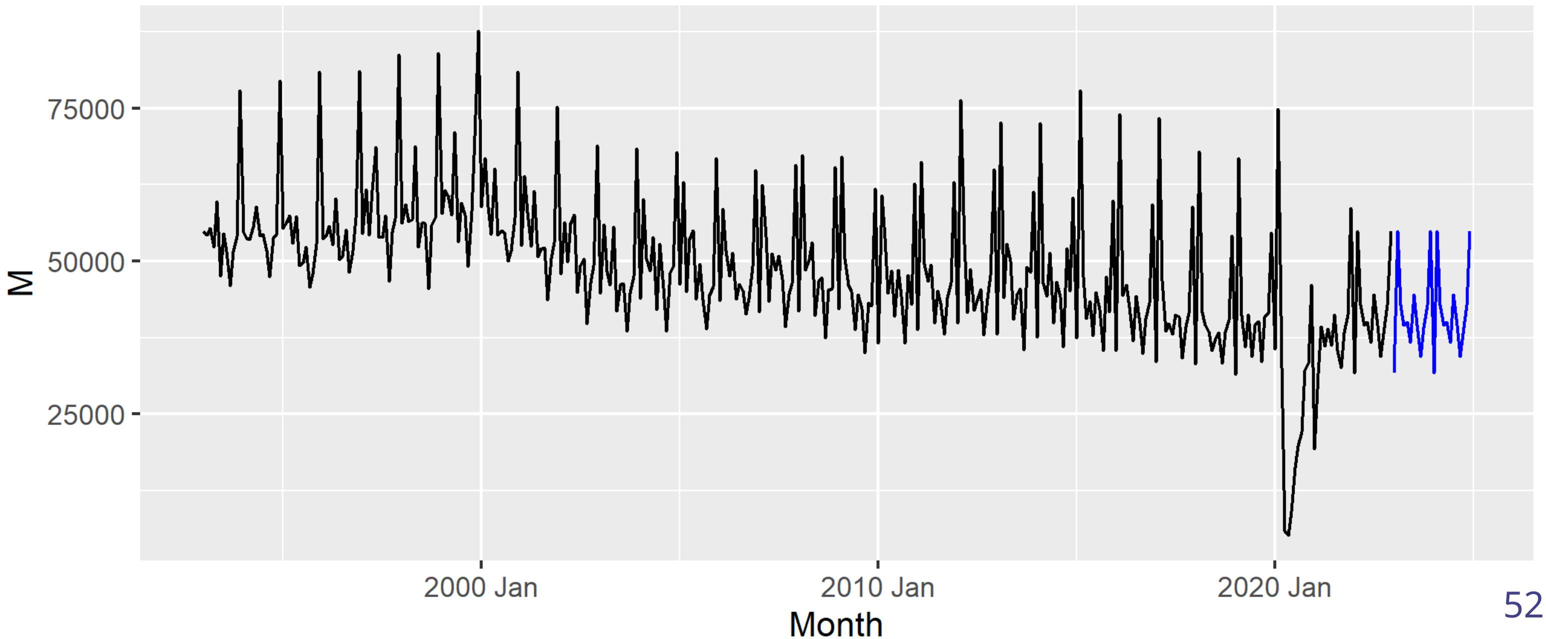
Simple forecasting methods

Seasonal naive

We can make it slightly more elaborate by assuming it's the same value as in the last same season:

$\hat{y}_{T+1|T} = y_{m(T+1)}$

- m(T) is the last time period with the same season as T+1



Simple forecasting methods

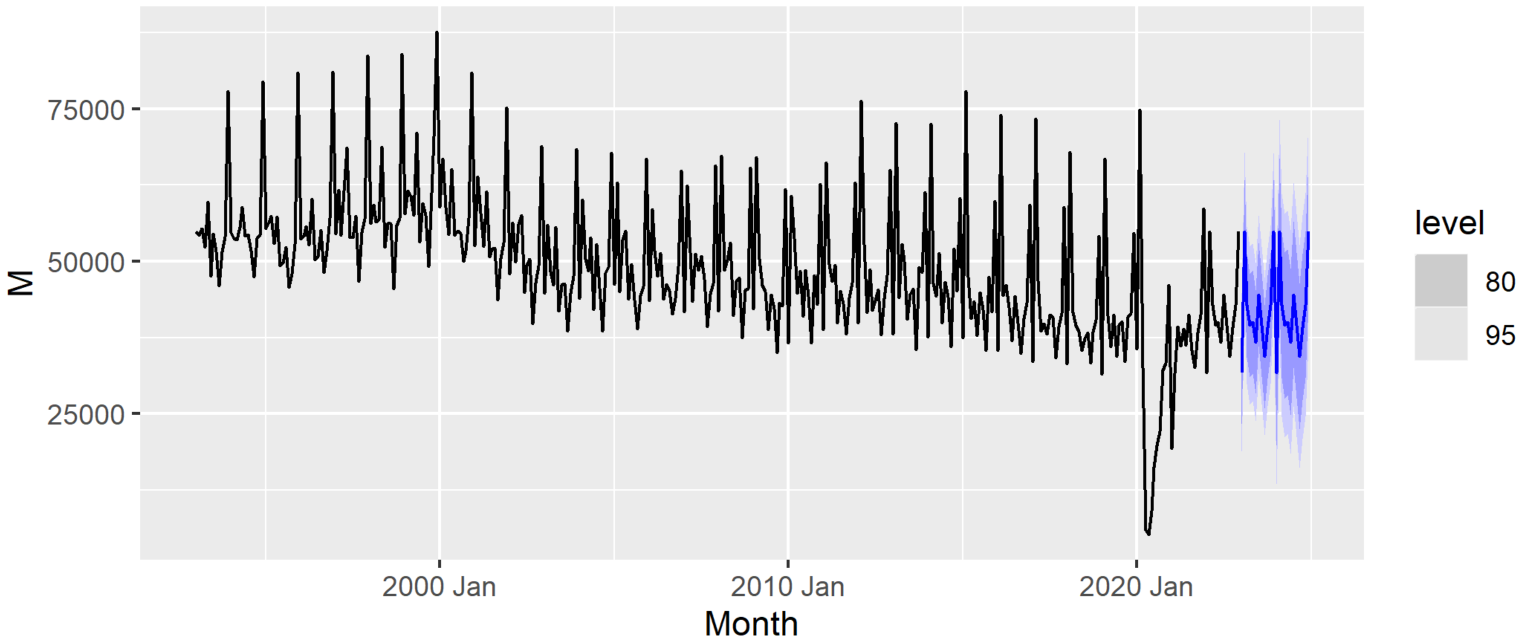
- At one step ahead, the confidence interval is the same:

$$CI_{95} = \hat{y}_{T+1|T} \pm 1.96\hat{\sigma}_u$$

- For longer horizon, forecast error is slightly different:
- Let $\sigma_h = \sigma_u\sqrt{h}$ be the h-step forecast error sd
- Let k be the number of seasonal cycles in the forecast prior to forecast time
  - If it's the first January since time T, k+1=1
  - If it's the second January since time T, k+1=2

$$CI_{95} = \hat{y}_{T+h|T} \pm 1.96\hat{\sigma}_u\sqrt{k+1}$$

Simple forecasting methods



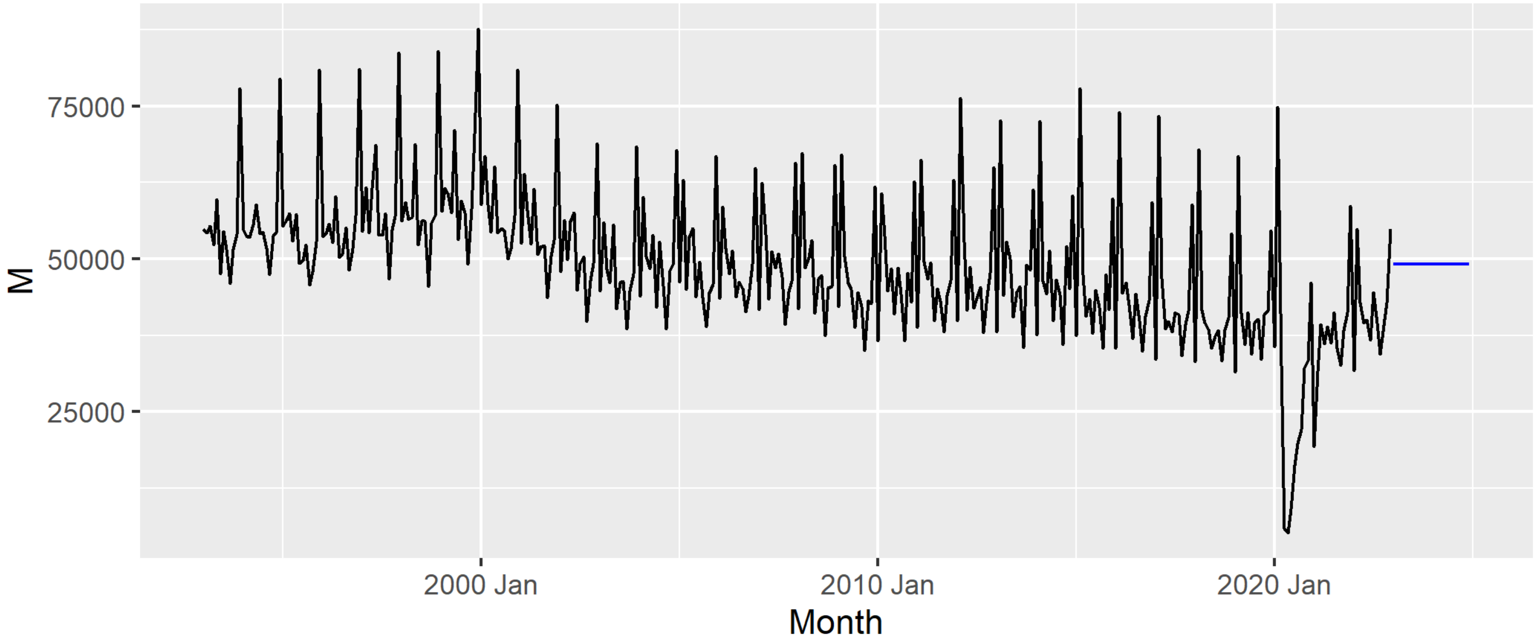
- Why the interval is smaller than in the previous case?
- The forecast errors are smaller
- So the standard deviation of errors is smaller!

Simple forecasting methods

Simple Average

We can also just take an average of the time series and make it our prediction:

$$\hat{y}_{T+1|T} = \bar{y}_T = \frac{\sum_{t \leq T} y_t}{T}$$



Simple forecasting methods

- At one step ahead, the confidence interval is the same:

$$CI_{95} = \hat{y}_{T+1|T} \pm 1.96\hat{\sigma}_u$$

- For longer horizon, forecast error is slightly different:
- Let $\sigma_h = \sigma_u\sqrt{h + \frac{1}{T}}$ be the h-step forecast error sd

$$CI_{95} = \hat{y}_{T+h|T} \pm 1.96\hat{\sigma}_u\sqrt{h + \frac{1}{T}}$$

- Generally,aAverage value acros 20 years is not a good prediction

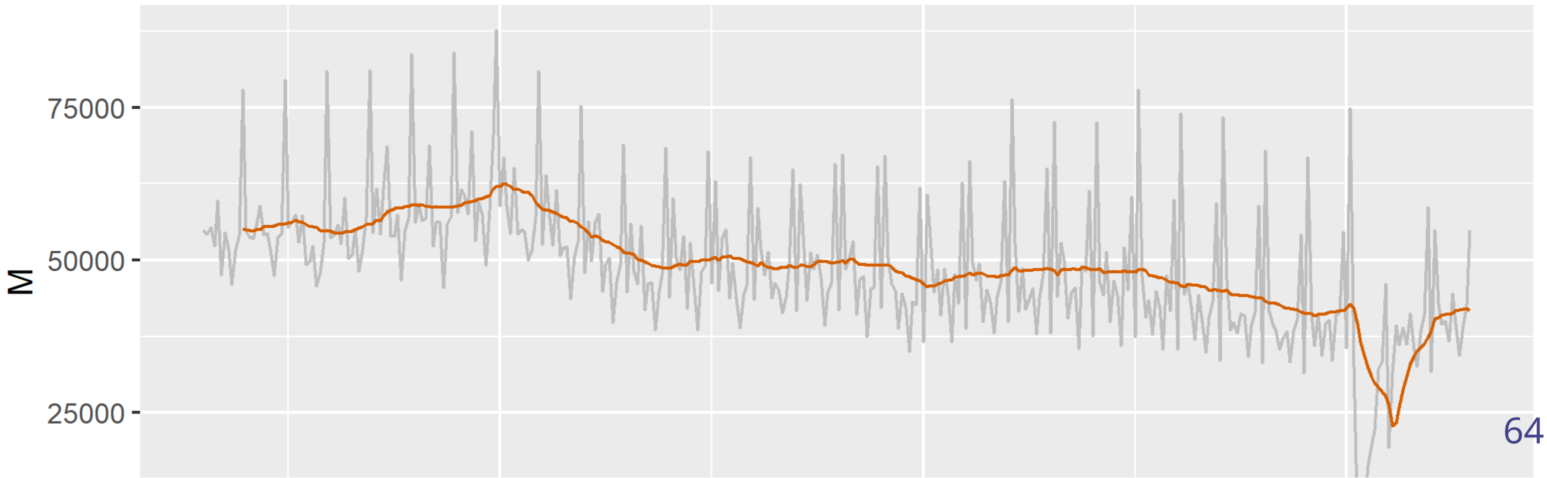
Moving average

Consider an average of the last k observations:

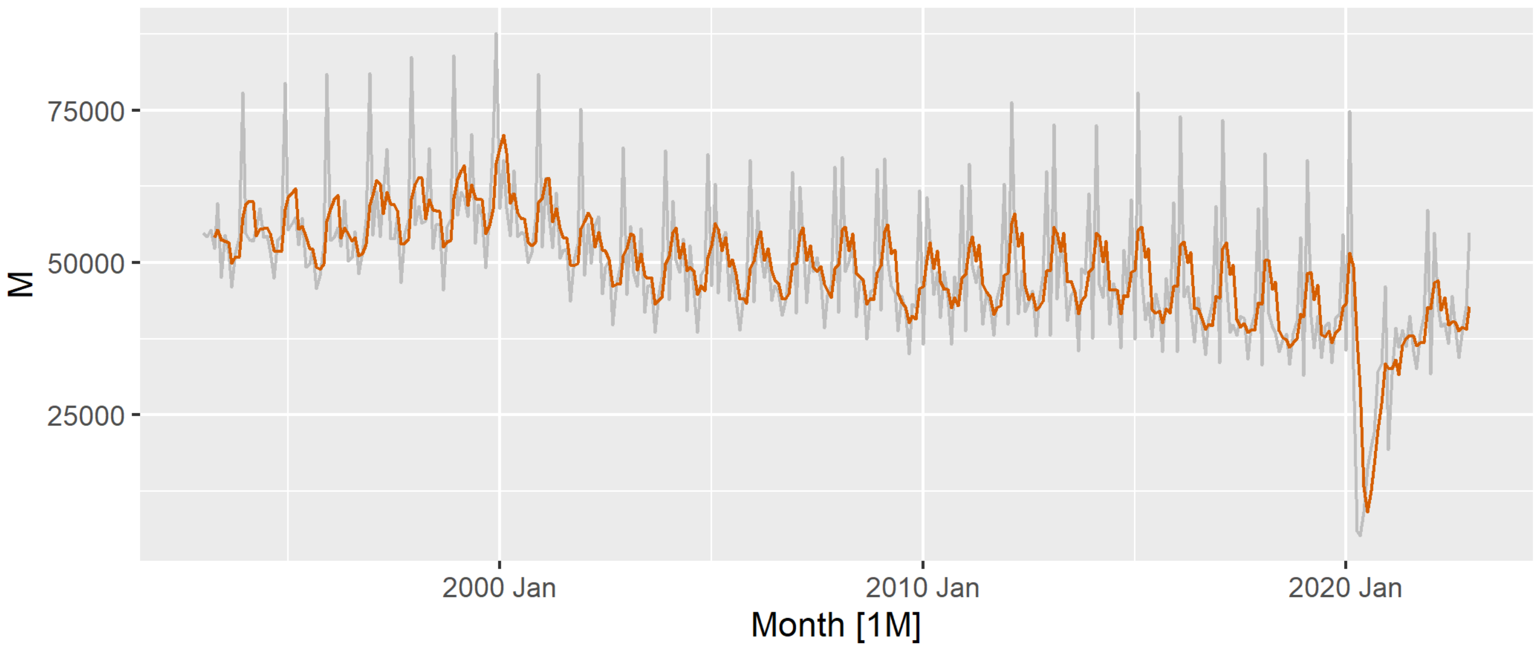
$$MA(k)_t : \frac{\sum_{j=1}^{k} y_{t+1-j}}{k} = \frac{y_t + y_{t-1} .. + y_{t+1-k}}{k}$$

- How many? Usually equal to number of seasons, so the seasonal variation smoothed out
- As we will see later, this is more useful in identifying trend and cycle components

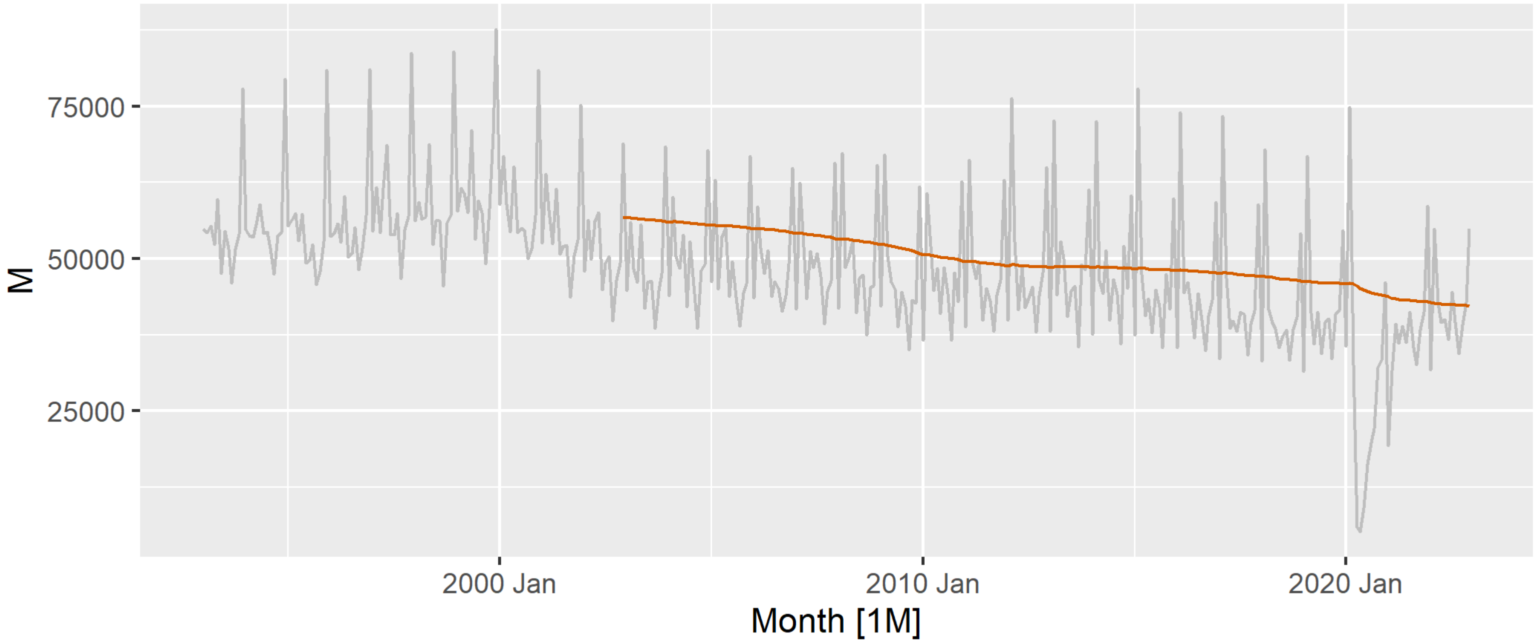
12 months



4 Months



3 years



Evaluating forecasts

Which of the forecasts was the best?

- There is couple of ways to evaluate the forecast accuracy
- They all have advantages and disadvantages
- General idea: how close the forecast was to the observed value
- You always use OUT-OF-SAMPLE errors, not fitted residuals

Mean Error

$$ME = \frac{\sum_{t=1}^{T-h}(y_{t+h} - \hat{y}_{t+h|t})}{T-h}$$

- This is the average of forecast error
- Can tell us which direction is the bias
- You can test for the existence of bias with a usual t test:
  - $H_0 : E(e_t) = 0$
  - $H_A : E(e_t) \neq 0$ (or inequality)
- Test statistic and the null distribution:

$$T_{test} = \frac{\bar{e} - 0}{\frac{\hat{\sigma}_e}{\sqrt{n}}} \sim t_{n-1}$$

- Positive and negative values can add up to 0
- So even if errors are large, but symmetric, this measure will be close to 0

Mean Error:

```
## # A tibble: 3 × 11
##   .model          SS    .type       ME   RMSE    MAE   MPE  MAPE  MASE RMSSE
##   <chr>           <lgl> <chr>    <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mean            FALSE Test   -4822. 12980. 12050. -17.1  29.0  4.19  3.51
## 2 Naïve           FALSE Test  -13170. 17851. 13170. -35.3  35.3  4.58  4.83
## 3 Seasonal naïve FALSE Test   -5940.  5951.  5940. -12.8  12.8  2.06  1.61
```

- If the error is negative, we overestimate!

Mean Absolute Error

$$MAE = \frac{\sum_{t=1}^{T-h} \left| y_{t+h} - \hat{y}_{t+h|t} \right|}{T-h}$$

- Similar, but we take absolute value of errors. So they don't cancel out!
- This measure is **always** positive
- But we can't say whether we underpredict or overpredict

```
## # A tibble: 3 × 11
##   .model         SS    .type       ME   RMSE    MAE   MPE  MAPE  MASE RMSSE
##   <chr>          <lgl> <chr>    <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mean           FALSE Test    -4822. 12980. 12050. -17.1  29.0  4.19  3.51
## 2 Naïve          FALSE Test   -13170. 17851. 13170. -35.3  35.3  4.58  4.83
## 3 Seasonal naïve FALSE Test    -5940.  5951.  5940. -12.8  12.8  2.06  1.61
```

- Now clearly seasonal is the best

Mean Percentage Error

$$MPE = \frac{\sum_{t=1}^{T-h}(y_{t+h} - \hat{y}_{t+h|t})/y_{t+h}}{T-h}$$

- Answers the question:
  - on average, my forecast is x% wrong
  - It's unitless, so I can compare forecasts of different measures
  - EG: comparing forecast of inflation vs exports
- But again, negative and positive can cancel out...
- So average forecast is again performing well!

```
## # A tibble: 3 × 11
##   .model          SS    .type      ME   RMSE    MAE   MPE  MAPE  MASE RMSSE
##   <chr>           <lgl> <chr>   <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mean            FALSE Test    -4822. 12980. 12050. -17.1  29.0  4.19  3.51
## 2 Naïve           FALSE Test   -13170. 17851. 13170. -35.3  35.3  4.58  4.83
## 3 Seasonal naïve  FALSE Test    -5940.  5951.  5940. -12.8  12.8  2.06  1.61
```

Mean Absolute Percentage Error

$$MAPE = \frac{\sum_{t=1}^{T-h} \left| y_{t+h} - \hat{y}_{t+h|t} \right| / y_{t+h}}{T-h}$$

- Similar as before, but we take the absolute value

```
## # A tibble: 3 × 11
##   .model          SS    .type      ME   RMSE    MAE   MPE  MAPE  MASE RMSSE
##   <chr>           <lgl> <chr>   <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mean            FALSE Test   -4822. 12980. 12050. -17.1  29.0  4.19  3.51
## 2 Naïve           FALSE Test  -13170. 17851. 13170. -35.3  35.3  4.58  4.83
## 3 Seasonal naïve  FALSE Test   -5940.  5951.  5940. -12.8  12.8  2.06  1.61
```

Squared Errors

- Mean Squared Errors

$$MSE = \frac{\sum(A_t - F_t)^2}{n}$$

- Root Mean Squared Errors

$$RMSE = \sqrt{\frac{\sum(A_t - F_t)^2}{n}}$$

- If we take squre instead of absolute value, we penalize more big deviations

- Then we need to take square root to get the right units back

```
## # A tibble: 3 × 11
##   .model         SS    .type      ME   RMSE    MAE   MPE  MAPE  MASE RMSSE
##   <chr>          <lgl> <chr>   <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mean           FALSE Test   -4822. 12980. 12050. -17.1  29.0  4.19  3.51
## 2 Naïve          FALSE Test  -13170. 17851. 13170. -35.3  35.3  4.58  4.83
## 3 Seasonal naïve FALSE Test   -5940.  5951.  5940. -12.8  12.8  2.06  1.61
```

Time series decomposition

- Helps in analyzing the patterns in the time series data
- Sometimes used for forecasting

Multiplicative Decomposition: Assume time series is a product of 4 elements:

$$y_t = S_t T_t C_c R_t$$

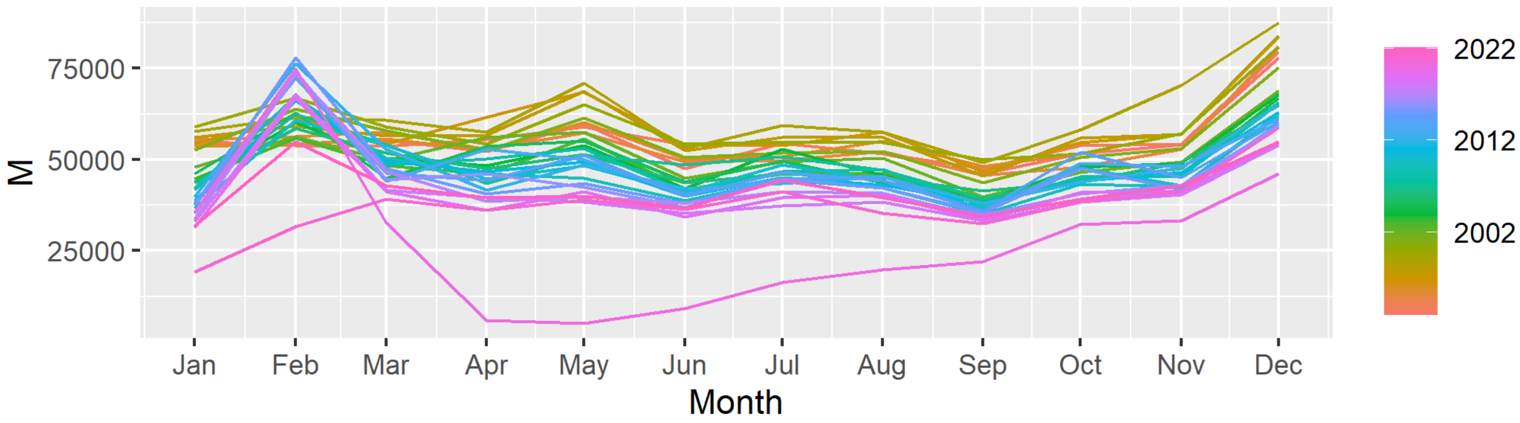
Goal: Identify the elements of the time series:

1. Seasonality S_t
2. Trend T_t
3. Cycles C_t
4. Irregular/Reminder R_t

- Two notes:
  - We will often ignore irregular components
  - Some methods don't distinguish between Trend and Cycles

Seasonality

- How would you identify which variations are due to seasonality?



- **Idea**:
    1. Eliminate seasonal variation
    2. Compare the actual series to the one without seasonal variation
    3. The difference is due to seasonality!
- How to eliminate seasonal variation?
- We will use (Centred) Moving Averages for smoothing.

Seasonality

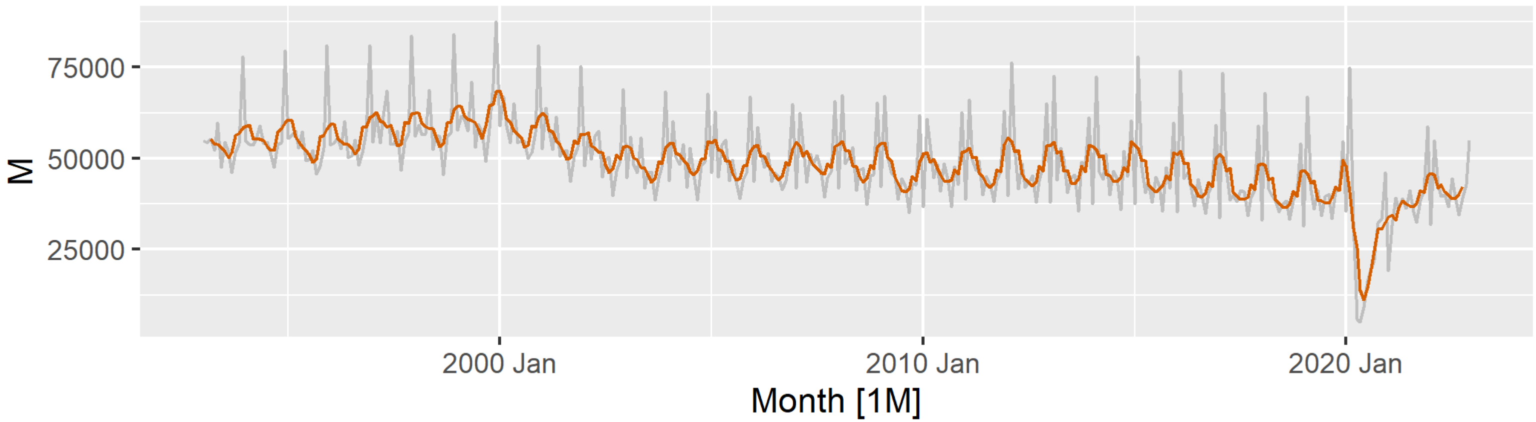
Moving averages for smoothing

- Why moving average smoothes out seasonal variation?
- It averages out variation over some period of time
- In some seasons we have more weddings, in some seasons we have less wedding. On average these positive and negative seasonalities will average out.
- Over which period should we take average?

Seasonality

- Suppose I take average over 5 months.
- Note that this time the period in focus is at the center
- I look at y_t, two observations before it and two observations after it!
- So the closest observations to y_t

$$MA(5)_t : \frac{\sum_{j=-2}^{2} y_{t+j}}{5} = \frac{y_{t+2} + y_{t+1} + y_t + y_{t-1} + y_{t-2}}{5}$$

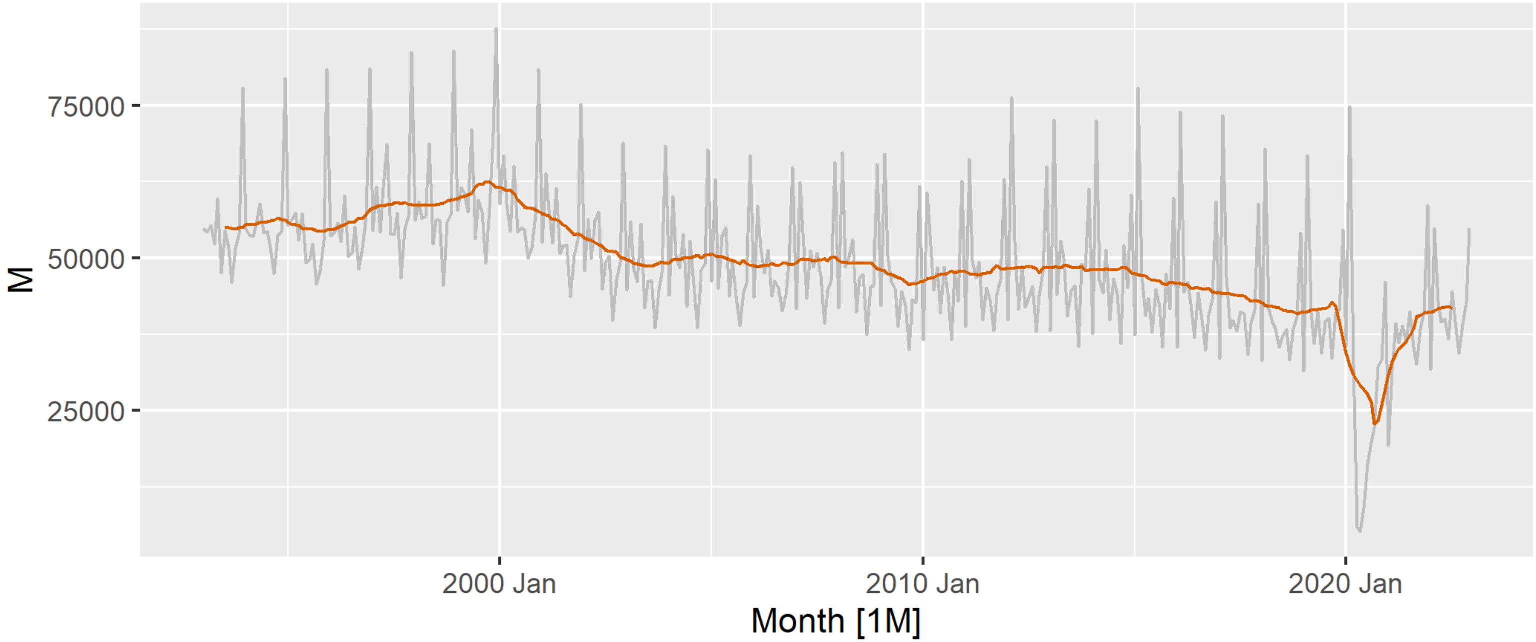


- If I take only 5, I don't average over all seasons!
- Sometimes I capture more high seasons but not low seasons, so seasonality persists

Seasonality

- Suppose I take average over 12 seasons
- Now each average is over all months (seasons)
- We eliminated seasonal variation

$$MA(12)_t : \frac{\sum_{j=-6}^{5} y_{t+j}}{12} = \frac{y_{t+5} + y_{t+4} .. + y_t + .. + y_{t-5} + y_{t-6}}{12}$$



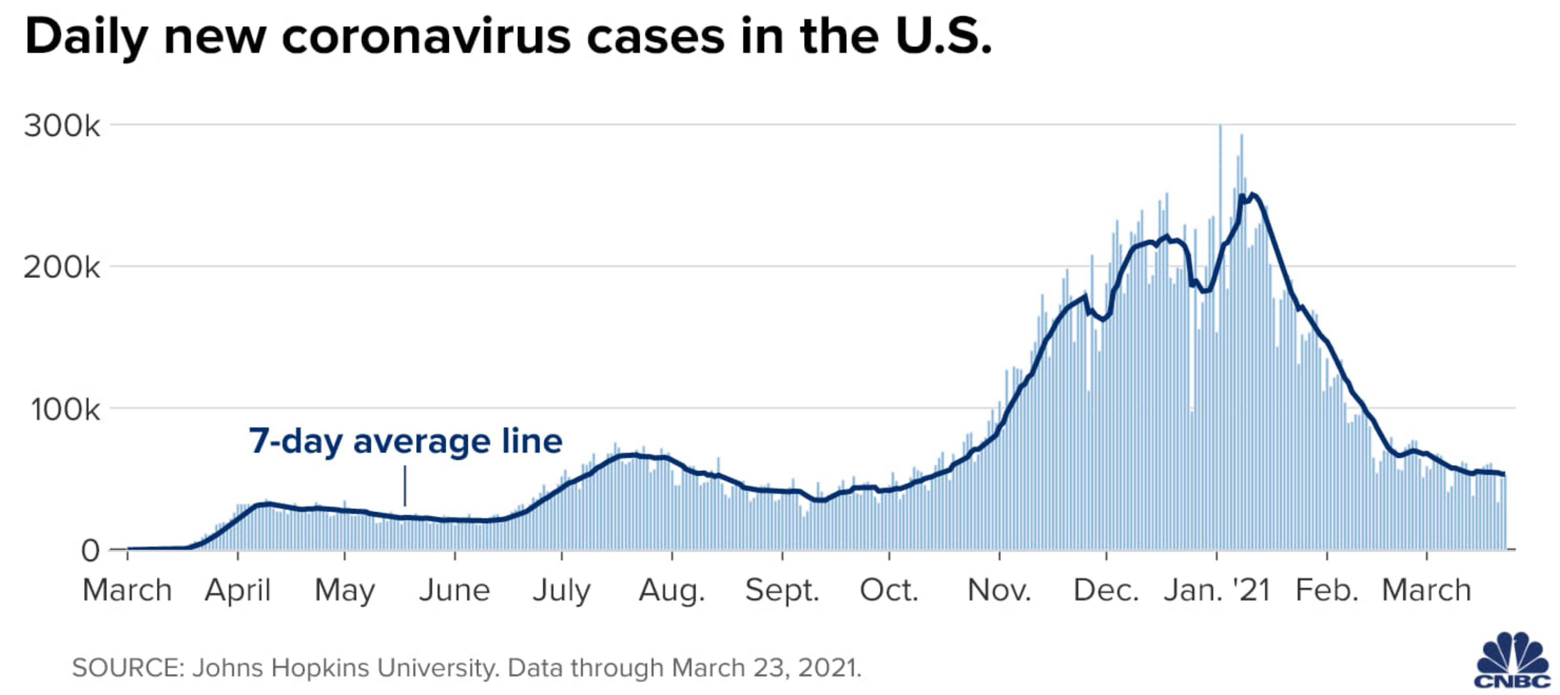
Seasonality

- Mathematical caveat
  - Since the number of periods is even (12), our main observations is not really at the center
  - We can have 5 obs before and 6 after
  - Or 6 obs bevore and 5 after **Centered** Moving Average
- Or we can have both!
- Calculate moving average both ways and then take the average of the two.

$$CMA(12)_t = \left(\frac{\sum_{j=-6}^{5} y_{t+j}}{12} + \frac{\sum_{j=-5}^{6} y_{t+j}}{12}\right)/2$$

- Note that we lose some data at the end and at the beginning.

- What was the "seasonality" in terms of Covid?



- Less testing on weekends
- Seasonality was by the day of the week
- So we take 7 days average to smooth it out

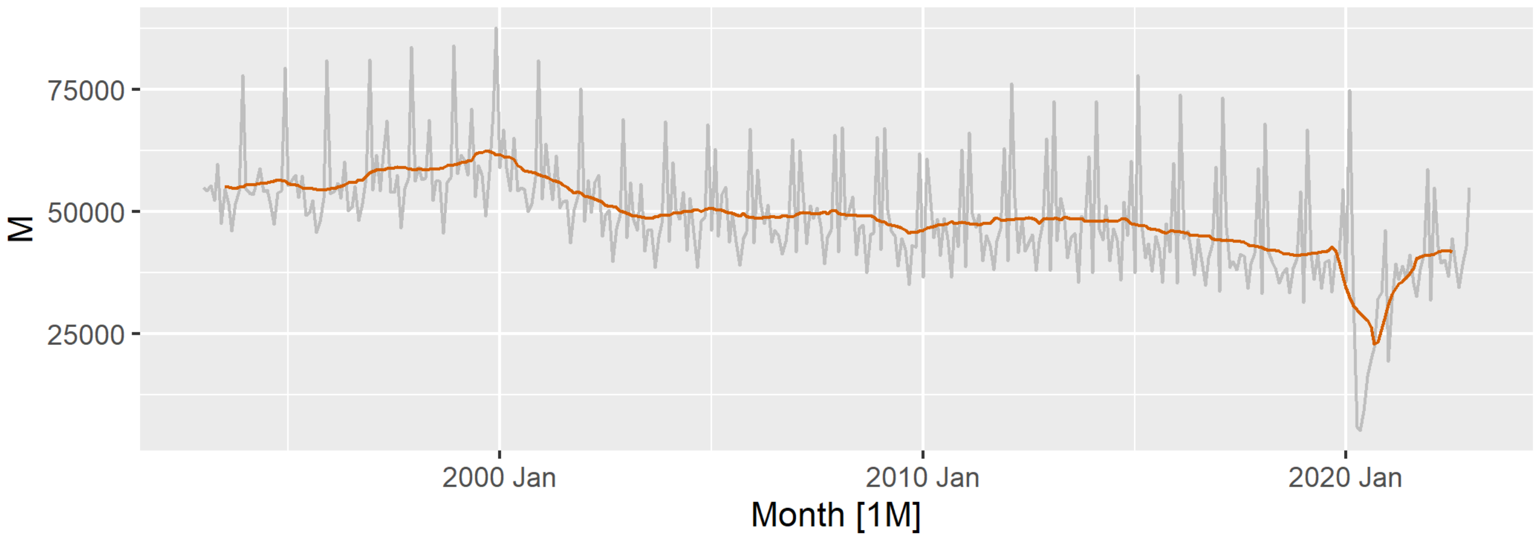
Seasonality

- We achieved first step, we have a series without seasonal variation
- So how we identify which parts are due to seasonality?

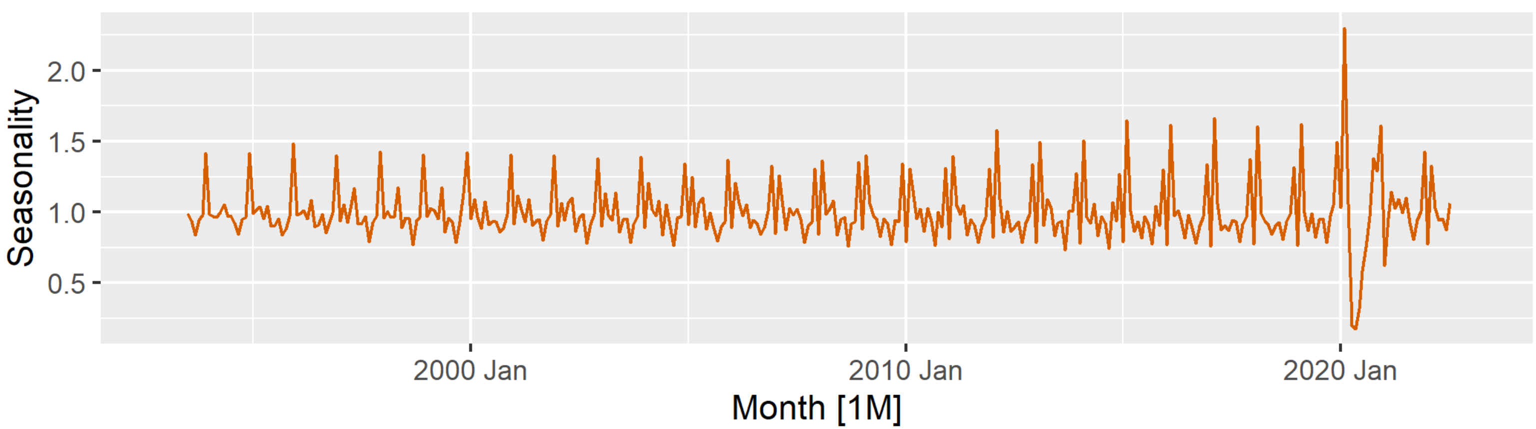
Seasonal indices

- Compare actual data to data without seasons
- January 2010 seasonal factor would be

$$SF_{January,2010} = Y_{January,2010}/CMA_{January,2010}$$


75000
50000
25000
M
2000 Jan
2010 Jan
2020 Jan
Month [1M]

$$SF_t = Y_t / CMA_t$$


2.0
1.5
1.0
0.5
Seasonality
2000 Jan
2010 Jan
2020 Jan
Month [1M]

Seasonality

Seasonal indices

- We assume seasonal indices are the same across the time, so we just take the average of all of them for each season:

$$SF_{January} = \sum_{year} Y_{January,year} / CMA_{January,year}$$

```
##    Month Seasonal_index
## 1      1      0.8646459
## 2      2      1.3184852
## 3      3      1.0124035
## 4      4      0.9360894
## 5      5      1.0167274
## 6      6      0.8573390
## 7      7      0.9494380
## 8      8      0.9299655
## 9      9      0.8000362
## 10    10      0.9524714
## 11    11      0.9844850
## 12    12      1.3779134
```

- In January, we have 13.5% less weddings than yearly average
- In December, we have on average 38% weddings than yearly avearge
- in June, we have 14.3% less weddings than yearly average

Seasonality

- They should average to 1
  - Because they represent how much they deviate from average in a given season
- (or in other words) They should add up to the number of seasons!

$$\sum_{s=1}^{S} SF_s = S$$

- If you don't know one index, you can identify it from the sum

Trend

- We isolated seasonality
- Now that our time series is not contaminated by seasonal variation, we can identify the trend

Assumption: Trend is linear

- We are trying to find a line that best approximate the *deseasoned data*
- That's what a linear regression do!
- My outcome is the deseasoned time series values
- My predictor is time

$$CMA_t = a + bt + e_t$$

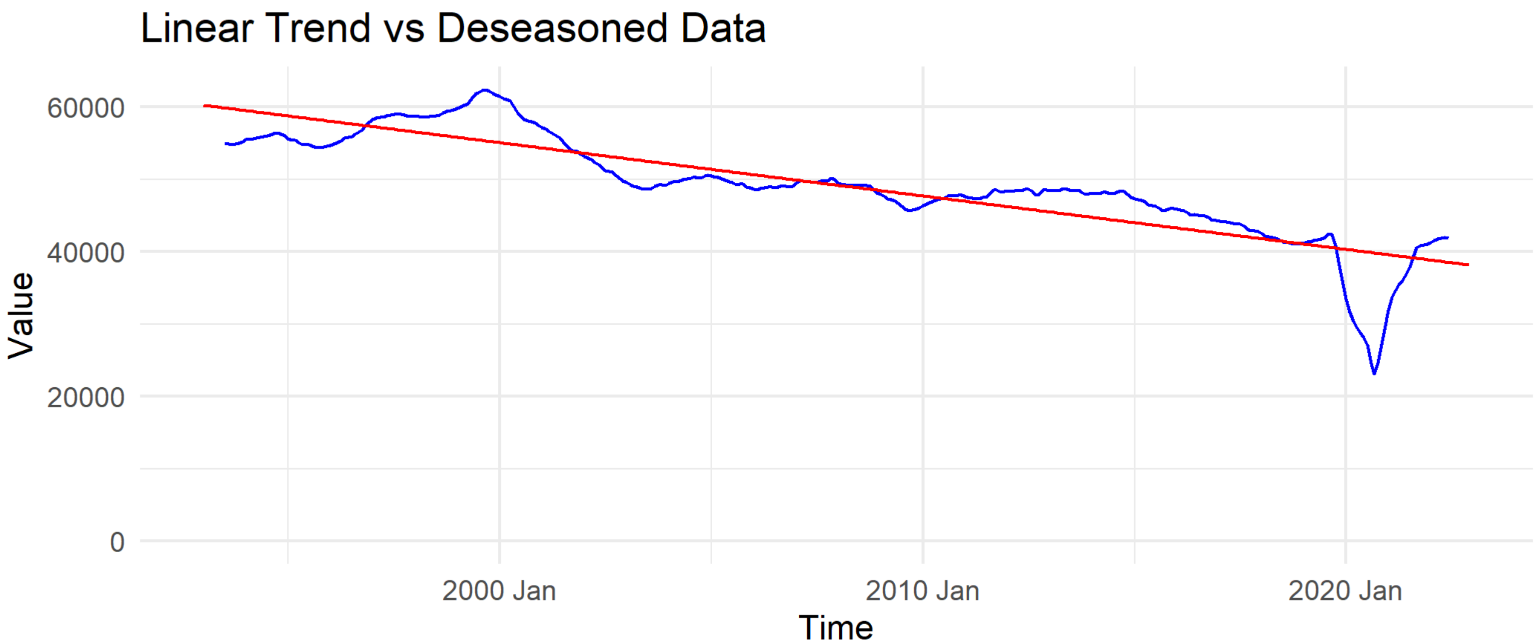
- We find a and b by OLS

-- Our predicted trend at time t is:

$$T_t = \hat{a} + \hat{b}t$$

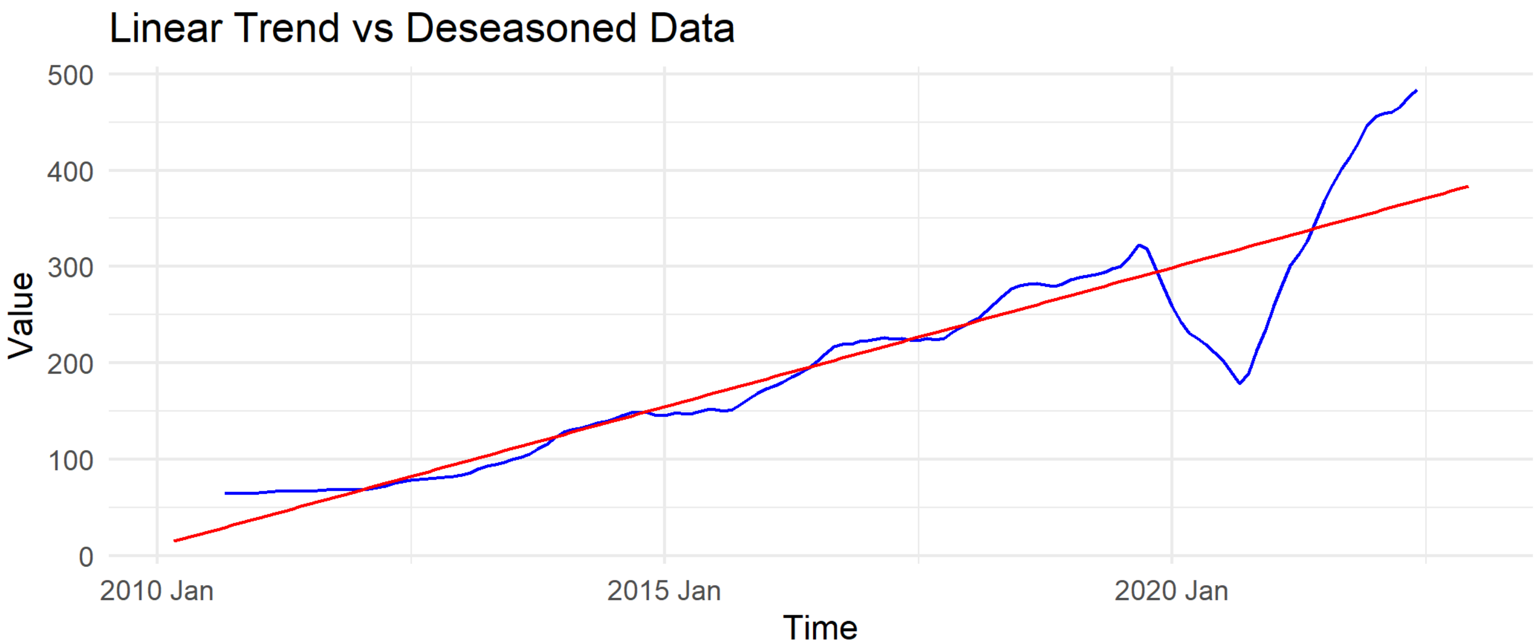
Straight marriages

```
##
## Call:
## lm(formula = trend ~ Time, data = a)
##
## Coefficients:
## (Intercept)         Time
##    60304.43       -61.46
```



Same-sex marriages

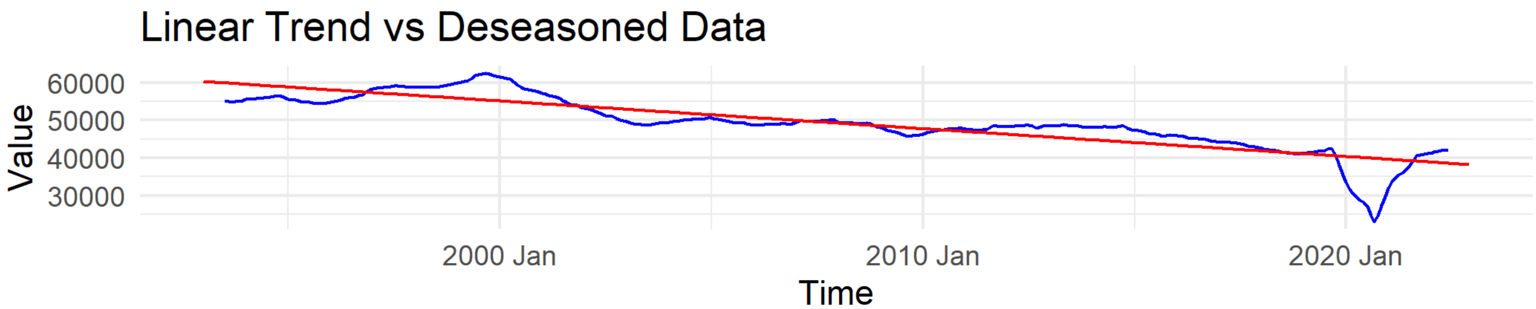
```
##
## Call:
## lm(formula = trend ~ Time, data = a)
##
## Coefficients:
## (Intercept)         Time
##      12.650        2.407
```

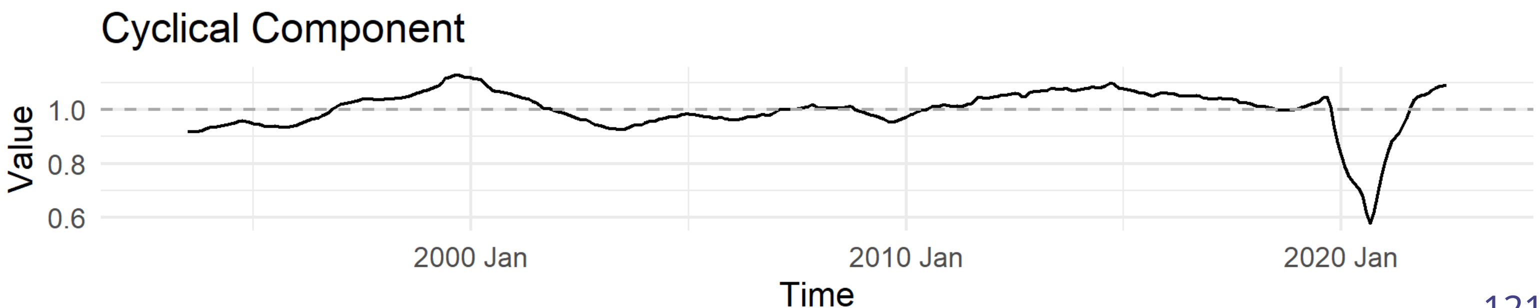


Cyclical element

- Cyclical element is the upward and downward movements around the trend in the deseasoned data
- Divide the centered moving average (deseasoned time series) by the trend value

$$C_t = \frac{CMA_t}{T_t}$$





Multiplicative Decomposition

- What about the irregular component?
- We will assume it's one, unless someone tells us there will be some shock

- Once we identified all the elements, we can make predictions for the original variable using the model:

$$y_t = S_t T_t C_c R_t$$

Prediction

What will be the marriage rate in January 2023 (T+1)?

- What is my S_{T+1}
  - S_{T+1} for January is: 0.865
- What is my T_{T+1}?
  - Formula: $60304.43 - 61.46 * 361 = 38117.37$
  - January 1993 is t=1, February 1993 is t=2 ... January 2023 is t=361
- What is my C_{T+1}?
  - Hardest to predict
  - Assume it's the same as last available one: $C_{T-6} = 1.0876$
- What is my R_{T+1}?
  - We don't expect anything crazy to happen so $R_{T+1} = 1$
  - Putting it all together:

$$\hat{y}_{T+1} = S_{T+1} T_{T+1} C_{T+1} R_{T+1} = 35859.83$$

Prediction

Confidence Interval

Step 1 Find interval bands from trend regression

- Just use the standard formula

$$\hat{T}_{T+1} \pm t_{\alpha/2,n-2}\sqrt{\hat{\sigma_T}^2(1+\frac{1}{T}+\frac{(T+1-\bar{t})^2}{\sum_{t=1}^{T}(t-\bar{t})^2})}$$

- T is the number of the last observation
- T_{T+1} is the trend prediction
- $\hat{\sigma}_T = \frac{SSE}{T-2}$ is the st.dev of residuals from the linear regression.

```
##        fit      lwr      upr
## 1 38119.08 31187.72 45050.44
```

Step 2 Mulitply these bands by the seasonal and cyclical component

$$CI_{95} = (31187 * \underbrace{1.0876}_{C_{T+1}} * \underbrace{0.865}_{S_{T+1}}, 45050 * \underbrace{1.0876}_{C_{T+1}} * \underbrace{0.865}_{S_{T+1}}) = (29339.92, 42381.87)$$