ISA 444: Business Forecasting

04 - Basic Tools for Time Series Analysis

Fadel M. Megahed

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Click here to schedule an appointment

Spring 2021

Outline

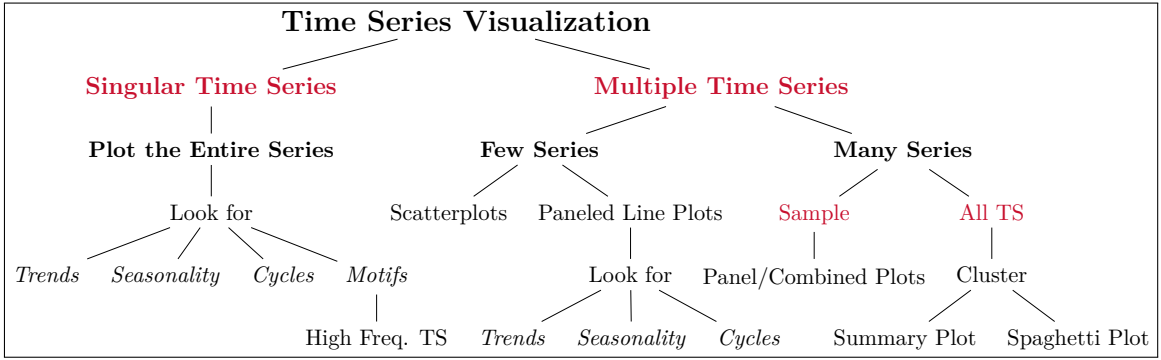
What we Covered Last Class

Main Learning Outcomes

- ☑ **Explain different goals for visualizing time series data**
- ☑ **Identify an appropriate chart for a specific time series data visualization goal**
- ☑ **Use software to construct charts of interest**

For the last two bullets, we only described the univariate case.

A Structured Approach for Time Series Visualization



A Potential Framework for Time Series Visualization.[1]

[1] This is my best attempt to improve on the general advice provided in the previous slide. Many of the suggestions, presented in this flow chart, stem from my past and current research/consulting collaborations. They are by no means a comprehensive list of everything that you can do.

Learning Objectives for Today's Class

Main Learning Outcomes

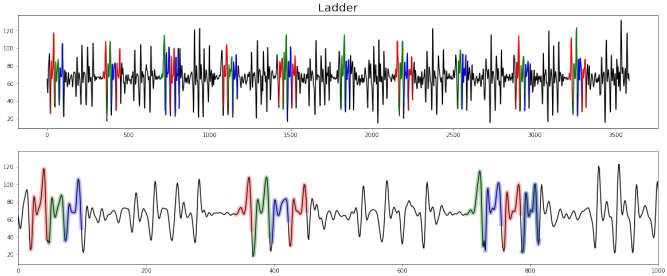
- **Use numerical summaries to describe a time series.**
- **Apply transformations to a time series.**

Outline

Outline

Looking for Motifs based on Wearable Sensors Data

Motifs allow us to cluster subsequences of a time series. It is a popular (unsupervised) learning approach, where patterns are automatically detected in time series.



Three dominant motifs discovered in an electrical utility application.[2]

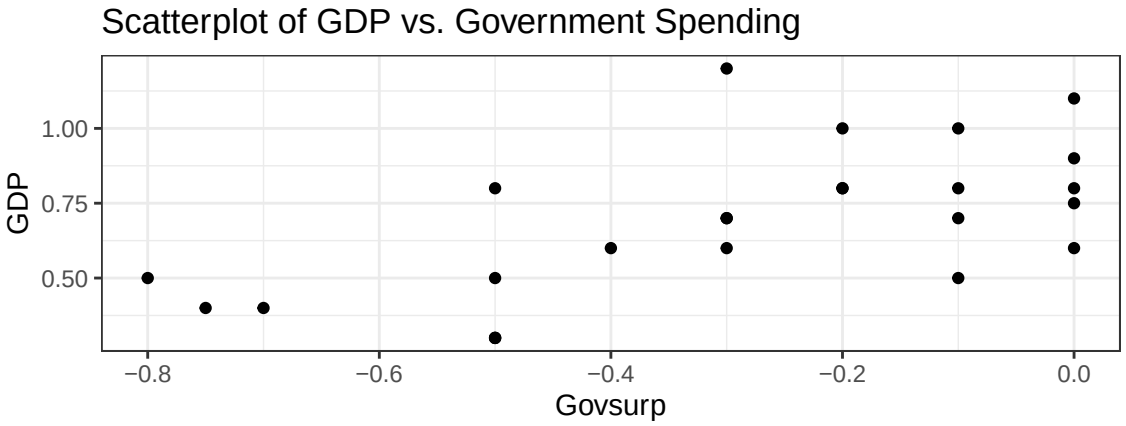
[2] Joint Work with GE Research and the University at Buffalo. Recall that the sensors had a frequency of 60Hz per Slide 25 in Lecture 02 Notes.

Outline

Scatterplots [1]

- **Scatterplots** are frequently used to visualize the correlation between two continuous variables.
- In this Example, we will be using the German_Forecast Data. The file can be downloaded to your working directory, using the `download.file()` from base R.
- Note that the data is an xlsx file, which would require us to use the `read_excel()` from the readxl package.
- We will remake the plot of GDP vs Govsurp (Figure 2.4 in our textbook) using R. As noted in the chapter, the figure was created using Minitab for the book.
- The plot using the `ggplot()` is shown in the next slide. We will recreate it in class.

Scatterplots [2]



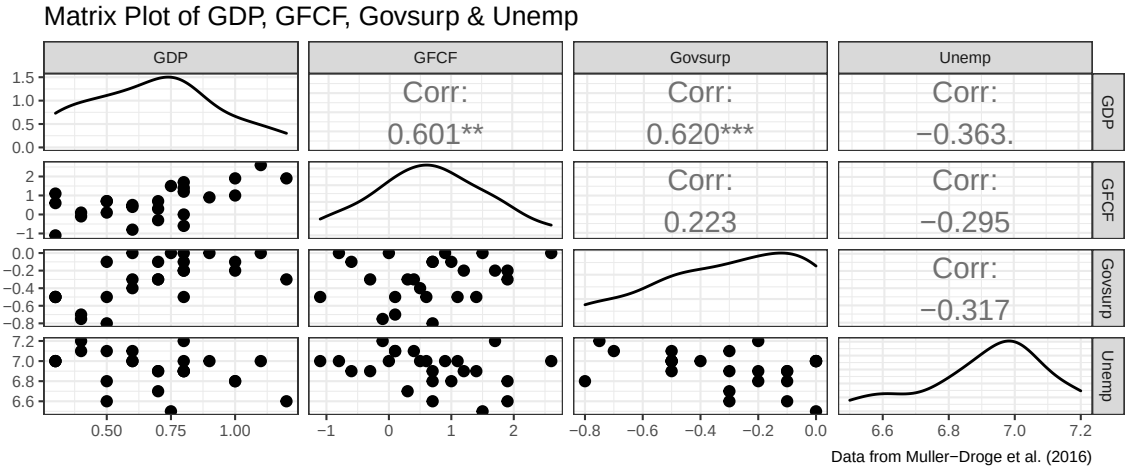
Data from Muller−Droge et al. (2016)

Scatterplot Matrix / Generalized Pairs Plots [1]

- Extending scatterplots for when we have more than two variables.[3]
- Can be easily created in R using the `ggpairs()` from the GGally package.

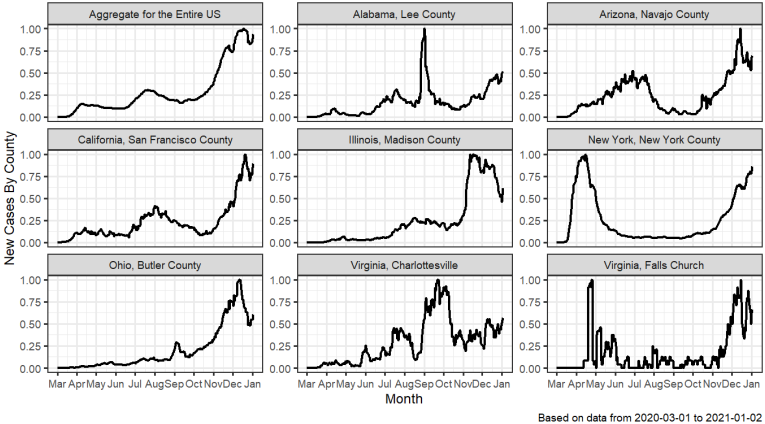
[3] John W Emerson, Walton A Green, Barret Schloerke, Jason Crowley, Dianne Cook, Heike Hofmann, Hadley Wickham. The Generalized Pairs Plot. Journal of Computational and Graphical Statistics, vol. 22, no. 1, pp. 79–91, 2012. Click here to access paper.

Scatterplot Matrix / Generalized Pairs Plots [2]



Data from Muller−Droge et al. (2016)

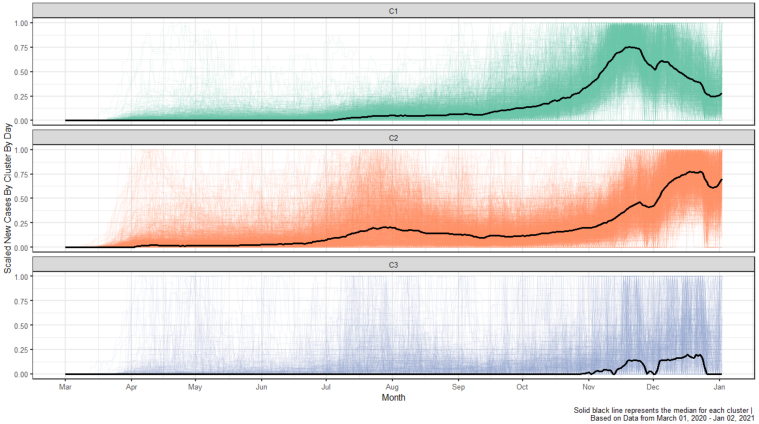
Panel Plots



New COVID-19 Cases in the United States.[4]

[4] Joint Work with Saint Louis University.

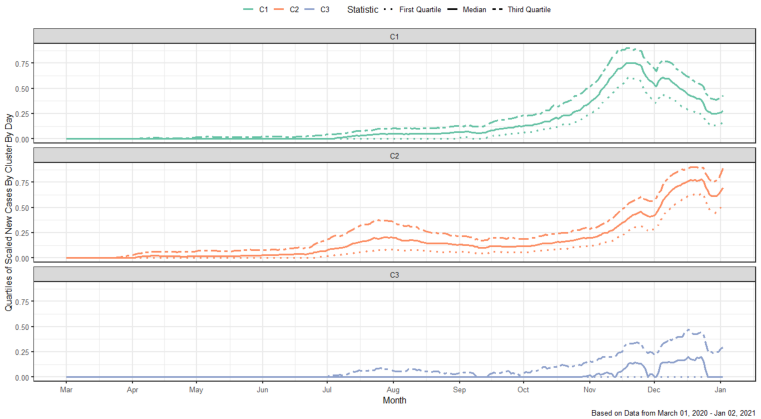
Clustering of COVID-19 New Cases: Spaghetti Plot



Spaghetti Plot of the 3 Major Clusters of COVID-19 Time-Series Profile Shapes.[5]

[5] Joint Work with Saint Louis University.

Clustering of COVID-19 New Cases: Summary Plot



Summary Plot of the 3 Major Clusters of COVID-19 Time-Series Profile Shapes.[6]

[6] Joint Work with Saint Louis University.

Outline

Measures of Average

Mean: Given a set of n values $Y_1, Y_2, \ldots, Y_n$, the arithmetic mean can be computed as:

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \frac{1}{n} \sum_{i=1}^{i=n} Y_i. \tag{1}$$

Order Statistics: Given a set of n values $Y_1, Y_2, \ldots, Y_n$, we place them in an ascending order to define the order statistics, written as $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$.

Median:

- If n is odd, $n = 2m + 1$ and the median is $Y_{(m+1)}$.
- If n is even, $n = 2m$ and the median is the average of the two middle numbers, i.e., $\frac{1}{2}[Y_{(m)} + Y_{(m+1)}]$.

Measures of Variation

The **range** denotes the difference between the largest and smallest value in a sample:

$$\text{Range} = Y_{(n)} - Y_{(1)}. \tag{2}$$

The **deviation** is defined as the difference between a given observation Y_i and the mean $\bar{Y}$.

The **mean absolute deviation (MAD)** is the average deviations about the mean, irrespective of their sign:

$$\text{MAD} = \frac{\sum_{i=1}^{i=n} |d_i|}{n}. \tag{3}$$

The **variance** is the average of the squared deviations around the mean:

$$S^2 = \frac{\sum_{i=1}^{i=n} d_i^2}{n-1}. \tag{4}$$

A Comment on the mad() Function in R

- The `mad()` in R is used for computing the median absolute deviation and **Not** the mean absolute deviation. This can be easily checked using `?mad()` in your R console.
- Thus, we will have to create our custom R function, `MAD()`, which we will define as follows:

```
MAD = function(x){
  return( mean( abs(x-mean(x)) ) )
  }
```

- Now, let us make sure that this formula works as expected by testing it on the vector `x = c(1, 2, 3)` and comparing it with manually computing the MAD.

Applications of Measures of Average/Variance: $GME

- Let us examine the stock prices for GameStop from September 01, 2020 up to January 31, 2021.
- Let us compute the aforementioned measures, on the adjusted closing price, using the following two approaches: (a) averages **across** all months, and (b) averages **by/within** month. The printout for those two methods are shown in the tables below.

meanACP	medianACP	madACP	varACP	sdACP
24.73	13.38	20.91	2501.19	50.01

month	meanACP	medianACP	madACP	varACP	sdACP
Sep 2020	8.51	8.68	1.23	1.98	1.41
Oct 2020	12.17	12.13	1.54	3.54	1.88
Nov 2020	12.42	11.69	1.32	2.90	1.70
Dec 2020	16.58	16.24	2.11	6.56	2.56
Jan 2021	79.62	39.12	73.22	10365.20	101.81

Outline

The Pearson Correlation Coefficient

- **Correlation:** measures the strength of the **linear relationship** between two quantitative variables.
- It can be computed using the `cor()` from base R. Mathematically speaking, the pearson correlation coefficient, r, can be computed as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5}$$

- Do **not** use the Pearson Correlation coefficient if both variables are not quantitative. Instead, refer to the `mixed.cor()` from the psch package to compute the correlations for mixtures of continuous, polytomous, and/or dichotomous variables.
- You should supplement any descriptive summaries with visualizations to ensure that you are able to interpret the computations correctly.

A Synthetic Example: The Anscombe Dataset [1]

In a seminal paper, Anscombe stated:[7] *Few of us escape being indoctrinated with these notions*

- *numerical calculations are exact, but graphs are rough;*
- *for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;*
- *performing intricate calculations is virtuous, whereas actually looking at the data is cheating.*

He proceeded by stating that *a computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.*

Now, let us consider his four datasets, each consisting of eleven (x,y) pairs.

[7] Anscombe, Francis J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. (Click here to access the full paper).

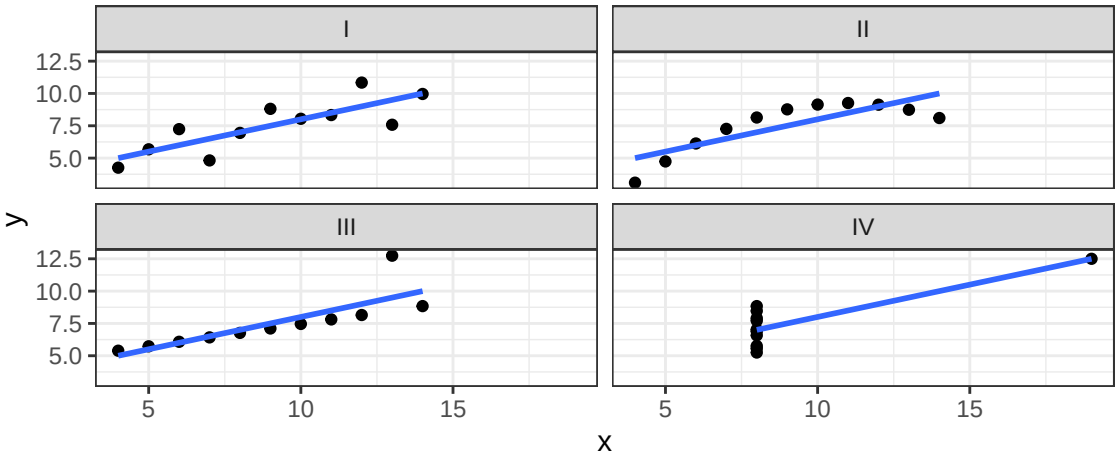
A Synthetic Example: The Anscombe Dataset [2]

x1	x2	x3	x4	y1	y2	y3	y4
10.00	10.00	10.00	8.00	8.04	9.14	7.46	6.58
8.00	8.00	8.00	8.00	6.95	8.14	6.77	5.76
13.00	13.00	13.00	8.00	7.58	8.74	12.74	7.71
9.00	9.00	9.00	8.00	8.81	8.77	7.11	8.84
11.00	11.00	11.00	8.00	8.33	9.26	7.81	8.47
14.00	14.00	14.00	8.00	9.96	8.10	8.84	7.04
6.00	6.00	6.00	8.00	7.24	6.13	6.08	5.25
4.00	4.00	4.00	19.00	4.26	3.10	5.39	12.50
12.00	12.00	12.00	8.00	10.84	9.13	8.15	5.56
7.00	7.00	7.00	8.00	4.82	7.26	6.42	7.91
5.00	5.00	5.00	8.00	5.68	4.74	5.73	6.89

A Synthetic Example: The Anscombe Dataset [3]

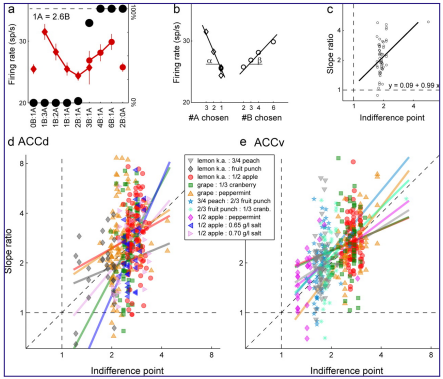
set	x.mean	x.sd	y.mean	y.sd	corr
I	9.00	3.32	7.50	2.03	0.82
II	9.00	3.32	7.50	2.03	0.82
III	9.00	3.32	7.50	2.03	0.82
IV	9.00	3.32	7.50	2.03	0.82

A Synthetic Example: The Anscombe Dataset [4]



Anscombe-Like Mistakes in Research and Practice

In my estimation, Figure 8c represents an example where regression should not have been performed[8].



[8] Cai, Xinying, and Camillo Padoa-Schioppa. 2012. "Neuronal Encoding of Subjective Value in Dorsal and Ventral Anterior Cingulate Cortex." *Journal of Neuroscience* 32(11):3791–3808.

Outline

First Differences

The change in the time series from one period to the next is known as the (first) difference. It can be computed as follows:

$$DY_t = Y_t - Y_{t-1} \tag{6}$$

symbol	date	adjusted	Yt-1	DYt
^DJI	2021-01-25	30960.00		
^DJI	2021-01-26	30937.04	30960.00	-22.96
^DJI	2021-01-27	30303.17	30937.04	-633.87
^DJI	2021-01-28	30603.36	30303.17	300.19
^DJI	2021-01-29	29982.62	30603.36	-620.74

Differences can be computed in one step using `diff()` from base R as follows.

symbol	date	adjusted	DYt
^DJI	2021-01-25	30960.00	
^DJI	2021-01-26	30937.04	-22.96
^DJI	2021-01-27	30303.17	-633.87
^DJI	2021-01-28	30603.36	300.19
^DJI	2021-01-29	29982.62	-620.74

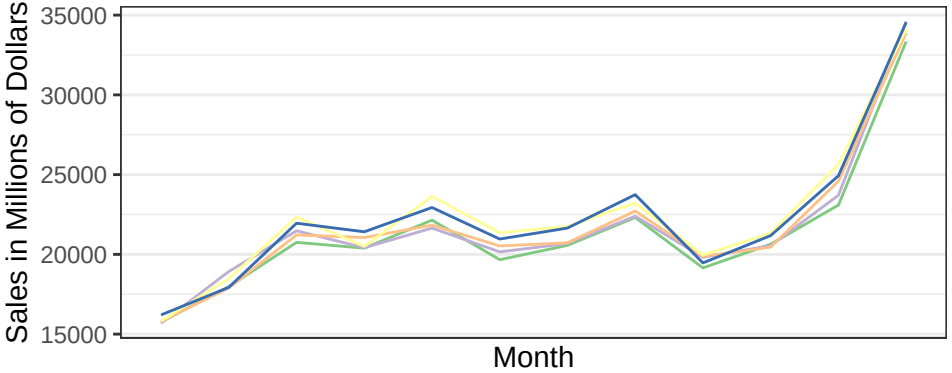
Differencing for Seasonal Data [1]

Recall the plot of the Advance Retail Sales: Clothing and Clothing Accessory Stores (RSCCASN) FRED dataset.

```
pacman::p_load(tidyquant)
retailSales = tq_get("RSCCASN", get = "economic.data",
                     from = "2015-01-01", to = "2019-12-31")
retailSales$month = month(retailSales$date)
retailSales$year = year(retailSales$date) %>% factor()
retailSales %>%
  ggplot(aes(x = month, y = price, color = year)) + geom_line() +
  labs(title = "How would you difference based on that plot?",
       x = "Month", y = "Sales in Millions of Dollars",
       caption = "Data from FRED, extracted using the tidyquant package") +
  scale_x_discrete() + scale_color_brewer(type = "qual") + theme_bw()
```

Differencing for Seasonal Data [2]

How would you difference based on that plot?



Data from FRED, extracted using the tidyquant package

Differencing with Seasonal Data [3]

The table below is the result of Approach #1 in R

date	price	Yt-m	DYt-m
2015-01-01	15764		
2015-02-01	17980		
2015-03-01	20752		
2015-04-01	20389		
2015-05-01	22145		
2015-06-01	19667		
2015-07-01	20564		
2015-08-01	22314		
2015-09-01	19151		
2015-10-01	20637		
2015-11-01	23090		
2015-12-01	33345		
2016-01-01	15685	15764	-79
2016-02-01	18926	17980	946
2016-03-01	21477	20752	725
2016-04-01	20412	20389	23
2016-05-01	21648	22145	-497
2016-06-01	20157	19667	490

Differencing with Seasonal Data [4]

The table below is the result of Approach #2 in R

date	price	DYt-m
2015-01-01	15764	
2015-02-01	17980	
2015-03-01	20752	
2015-04-01	20389	
2015-05-01	22145	
2015-06-01	19667	
2015-07-01	20564	
2015-08-01	22314	
2015-09-01	19151	
2015-10-01	20637	
2015-11-01	23090	
2015-12-01	33345	
2016-01-01	15685	-79
2016-02-01	18926	946
2016-03-01	21477	725
2016-04-01	20412	23
2016-05-01	21648	-497
2016-06-01	20157	490

Growth Rates: The Formulation

In the absence of seasonality, the growth rate for a time series is given by

$$GY_t = 100\frac{Y_t - Y_{t-1}}{Y_{t-1}} \tag{7}$$

In the presence of seasonality (with period $= m$), the growth rate for a time series is given by

$$GY_t = 100\frac{Y_t - Y_{t-m}}{Y_{t-m}} \tag{8}$$

Growth Rates in Practice – a Non-Graded Class Activity

- **Question 1:** Let us say that an investor purchased 10 stocks of \$GME, on 2021-01-29, at \$325/stock. The next trading day, 2021-02-01, the \$GME stock closed at \$225. Compute the growth rate in their portfolio worth (assuming it only has the GME stock) over this time period.
- Let us say that the growth rate, $GY_t = -g$. Now let us assume that the \$GME stock went up by g (i.e., if it went down 10%, it increased by 10% over the next trading day). What is the value of the investor's portfolio by stock market closing on 2021-02-02?
  - Provide the answer to both computational questions on Canvas.

The Log Transform [1]

The log transformation can be computed as follows:

$$L_t = \ln(Y_t) \tag{9}$$

Note that the `log()` in R takes the natural logarithm as its default base, i.e., would transform a variable/statistic based on the above equation.

The reverse transformation using the exponential function is:

$$e^{L_t} = e^{\ln(Y_t)} = Y_t \tag{10}$$

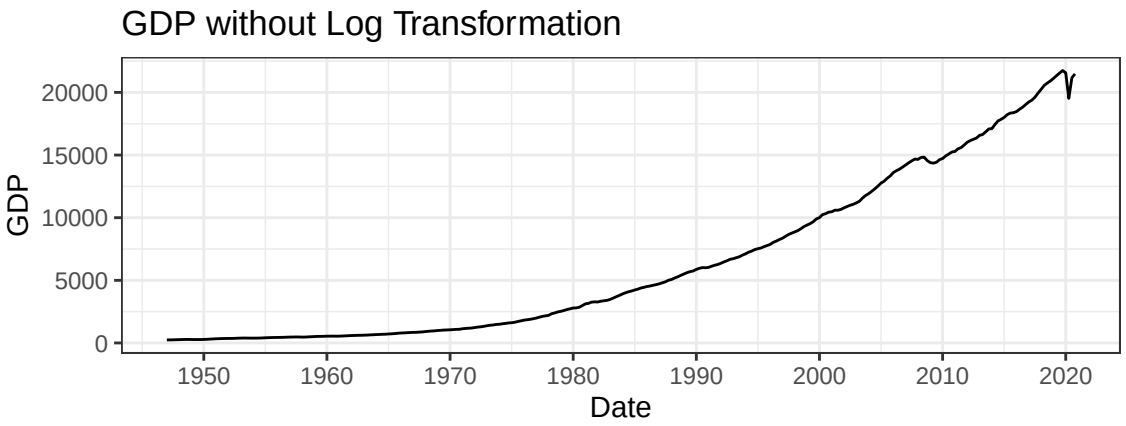
The first difference in logarithms represents the logarithm of the ratio:

$$L_t = \ln\left(\frac{Y_t}{Y_{t-1}}\right) = \ln(Y_t) - \ln(Y_{t-1}) \tag{11}$$

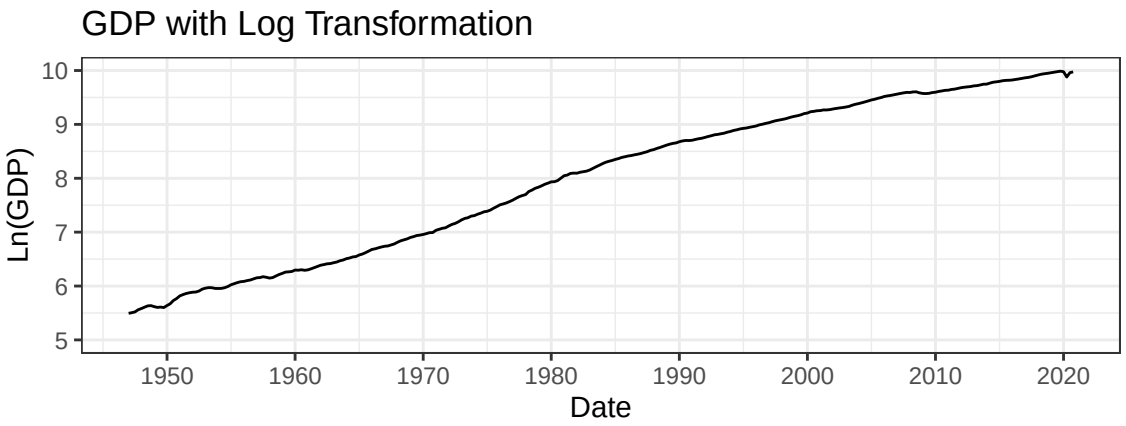
The Log Transform [2]

- The primary purpose of the log transform is to **convert exponential growth into linear growth.**
- The transform often has the **secondary purpose of balancing the variance.**
- Difference in logs and growth rate transformations produce similar results and interpretations.

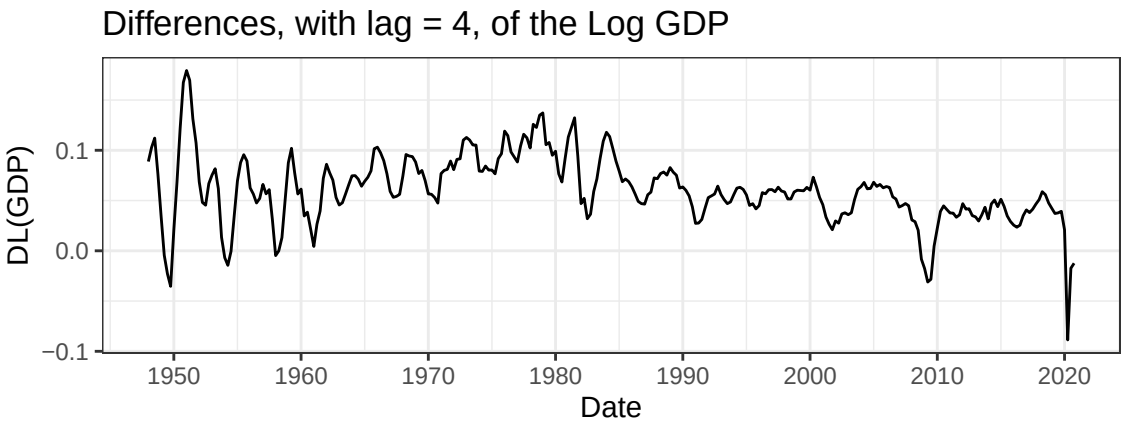
Plots with and without the Log Transformation [1]



Plots with and without the Log Transformation [2]



Plots with and without the Log Transformation [3]



Outline

Summary of Main Points

Main Learning Outcomes

- **Use numerical summaries to describe a time series.**
- **Apply transformations to a time series.**

Things to Do

- **Recommended:** Thoroughly read Chapter 2.1-2.6 of our book.
- Go through the slides, examples and make sure you have a good understanding of what we have covered.
- **Highly Recommended:** Go through the Week 02 Self-Paced Study Guide.
- **Required:** Complete the graded assignment.

Graded Assignment 04: Evaluating your Understanding

Please go to Canvas (click here) and answer the four questions. **Due February 08, 2021 [11:40 AM, Ohio local time].**

What/Why/Prep? The purpose of this assignment is to evaluate your understanding and retention of the material covered up to the end of Class 04. To reinforce your understanding of the covered material, I also suggest reading up to and including Chapter 2.6 of the book.

General Guidelines:

- Individual assignment.
- This is **NOT** a timed assignment.
- Proctorio is NOT required for this assignment.
- You will need to have R installed (or accessible through the Remote Desktop)

ISA 444: Business Forecasting

04 - Basic Tools for Time Series Analysis

Fadel M. Megahed

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University
Email: fmegahed@miamioh.edu
Office Hours: Click here to schedule an appointment

Spring 2021