

# TSMC stock price prediction

Ashish Das

12/11/2021

## Set up environment

```
library("tidyverse")
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr    0.3.4
## v tibble  3.1.6      v dplyr    1.0.7
## v tidyr   1.1.4      v stringr  1.4.0
## v readr   2.0.1      v forcats  0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("tidymodels")
```

```
## Warning: package 'tidymodels' was built under R version 4.1.2
```

```
## Registered S3 method overwritten by 'tune':
##   method                from
##   required_pkgs.model_spec parsnip
```

```
## -- Attaching packages ----- tidymodels 0.1.4 --
```

```
## v broom        0.7.9      v rsample      0.1.1
## v dials        0.0.10     v tune         0.1.6
## v infer        1.0.0      v workflows    0.2.4
## v modeldata    0.1.1      v workflowsets 0.1.0
## v parsnip      0.1.7      v yardstick    0.0.9
## v recipes      0.1.17
```

```
## Warning: package 'dials' was built under R version 4.1.2
```

```
## Warning: package 'infer' was built under R version 4.1.2
```

```
## Warning: package 'modeldata' was built under R version 4.1.2
```

```
## Warning: package 'parsnip' was built under R version 4.1.2
```

```
## Warning: package 'recipes' was built under R version 4.1.1
```

```
## Warning: package 'rsample' was built under R version 4.1.2
```

```
## Warning: package 'tune' was built under R version 4.1.2
```

```
## Warning: package 'workflows' was built under R version 4.1.2
```

```
## Warning: package 'workflowsets' was built under R version 4.1.2
```

```
## Warning: package 'yardstick' was built under R version 4.1.2
```

```
## -- Conflicts ----- tidymodels_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x yardstick::spec() masks readr::spec()  
## x recipes::step() masks stats::step()  
## * Use tidymodels_prefer() to resolve common conflicts.
```

```
library("timetk")
```

```
## Warning: package 'timetk' was built under R version 4.1.2
```

```
library("patchwork")
```

```
## Warning: package 'patchwork' was built under R version 4.1.2
```

```
library("lubridate")
```

```
## Warning: package 'lubridate' was built under R version 4.1.2
```

```
##  
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':  
##  
##    date, intersect, setdiff, union
```

```
library("modeltime")
```

```
## Warning: package 'modeltime' was built under R version 4.1.2
```

```
# Read the dataset, convert the date and create new columns  
df_tsm <-  
  read_csv("TSM.csv") %>%  
  rename("Adj.Close" = "Adj Close") %>%  
  mutate(date = ymd(Date),  
         trend = row_number(),  
         year = as.factor(year(Date)),  
         quarter = as.factor(quarter(Date)),  
         month = as.factor(month(Date))) %>%  
  filter(date >= '2010-01-01')
```

```
## Rows: 5968 Columns: 7
```

```
## -- Column specification -----  
## Delimiter: ","  
## dbl  (6): Open, High, Low, Close, Adj Close, Volume  
## date (1): Date
```

```
##  
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
str(df_tsm)
```

```
## spec_tbl_df [2,891 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ Date      : Date[1:2891], format: "2010-01-04" "2010-01-05" ...
## $ Open      : num [1:2891] 11.5 11.6 11.6 11.4 11.1 ...
## $ High      : num [1:2891] 11.7 11.7 11.6 11.4 11.2 ...
## $ Low       : num [1:2891] 11.5 11.5 11.4 11.1 11 ...
## $ Close     : num [1:2891] 11.6 11.5 11.5 11.1 11.1 ...
## $ Adj.Close: num [1:2891] 7.83 7.8 7.77 7.51 7.51 ...
## $ Volume    : num [1:2891] 8096400 14375900 13608400 27346600 16895300 ...
## $ date      : Date[1:2891], format: "2010-01-04" "2010-01-05" ...
## $ trend     : int [1:2891] 3078 3079 3080 3081 3082 3083 3084 3085 3086 3087 ...
## $ year      : Factor w/ 25 levels "1997","1998",...: 14 14 14 14 14 14 14 14 14 14 ...
## $ quarter   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
## $ month     : Factor w/ 12 levels "1","2","3","4",...: 1 1 1 1 1 1 1 1 1 1 ...
## - attr(*, "spec")=
## .. cols(
## ..   Date = col_date(format = ""),
## ..   Open = col_double(),
## ..   High = col_double(),
## ..   Low = col_double(),
## ..   Close = col_double(),
## ..   `Adj Close` = col_double(),
## ..   Volume = col_double()
## .. )
## - attr(*, "problems")=<externalptr>
```

## Data preparing

```
# Check is any missing value in date and adj.close
is.nan(df_tsm$Adj.Close)
```

5/33

file:///E:/sem5/cse3505/Project/tsmc-stock-price-prediction.html

file:///E:/sem5/cse3505/Project/tsmc-stock-price-prediction.html

file:///E:/sem5/cse3505/Project/tsmc-stock-price-prediction.html



```
## [2677] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2689] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2701] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2713] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2725] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2737] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2749] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2761] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2773] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2785] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2797] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2809] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2821] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2833] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2845] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2857] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2869] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2881] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
is.finite(df_tsm$Adj.Close)
```

10/33

11/33

12/33

```
is.nan(df_tsm$date)
```

14/33

file:///E:/sem5/cse3505/Project/tsmc-stock-price-prediction.html

file:///E:/sem5/cse3505/Project/tsmc-stock-price-prediction.html



file:///E:/sem5/cse3505/Project/tsmc-stock-price-prediction.html

```
## [2677] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2689] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2701] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2713] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2725] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2737] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2749] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2761] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2773] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2785] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2797] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2809] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2821] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2833] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2845] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2857] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2869] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [2881] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
is.finite(df_tsm$date)
```

19/33

20/33

21/33

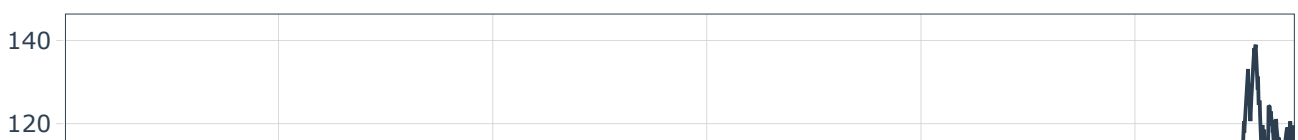
```
## [2339] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2353] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2367] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2381] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2395] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2409] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2423] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2437] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2451] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2465] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2479] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2493] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2507] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2521] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2535] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2549] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2563] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2577] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2591] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2605] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2619] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2633] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2647] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2661] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2675] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2689] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2703] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2717] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2731] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2745] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2759] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2773] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2787] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2801] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2815] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2829] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2843] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2857] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2871] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [2885] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

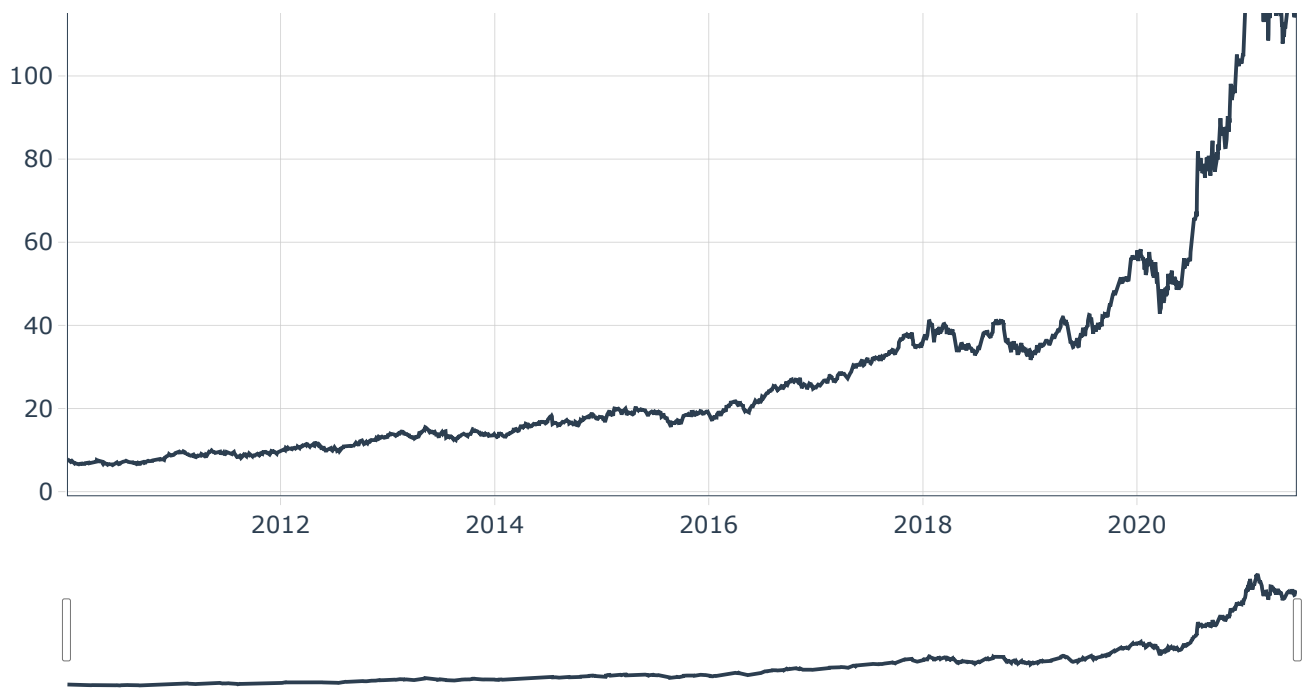
- There is no missing value in Adj.Close and date.

## EDA

```
# Plot the TSMC stock price over time
df_tsm %>%
  plot_time_series(date, Adj.Close, .smooth=FALSE, .plotly_slider = T)
```

### Time Series Plot





- As we can see stock price of TSMC is not stationary and it grew dramatically since 2020.

```
# Plot the TSMC stock price starting from last year
```

```
df_tsm %>%  
  filter(date>='2020-01-01') %>%  
  plot_time_series(date, Adj.Close, .smooth=FALSE)
```

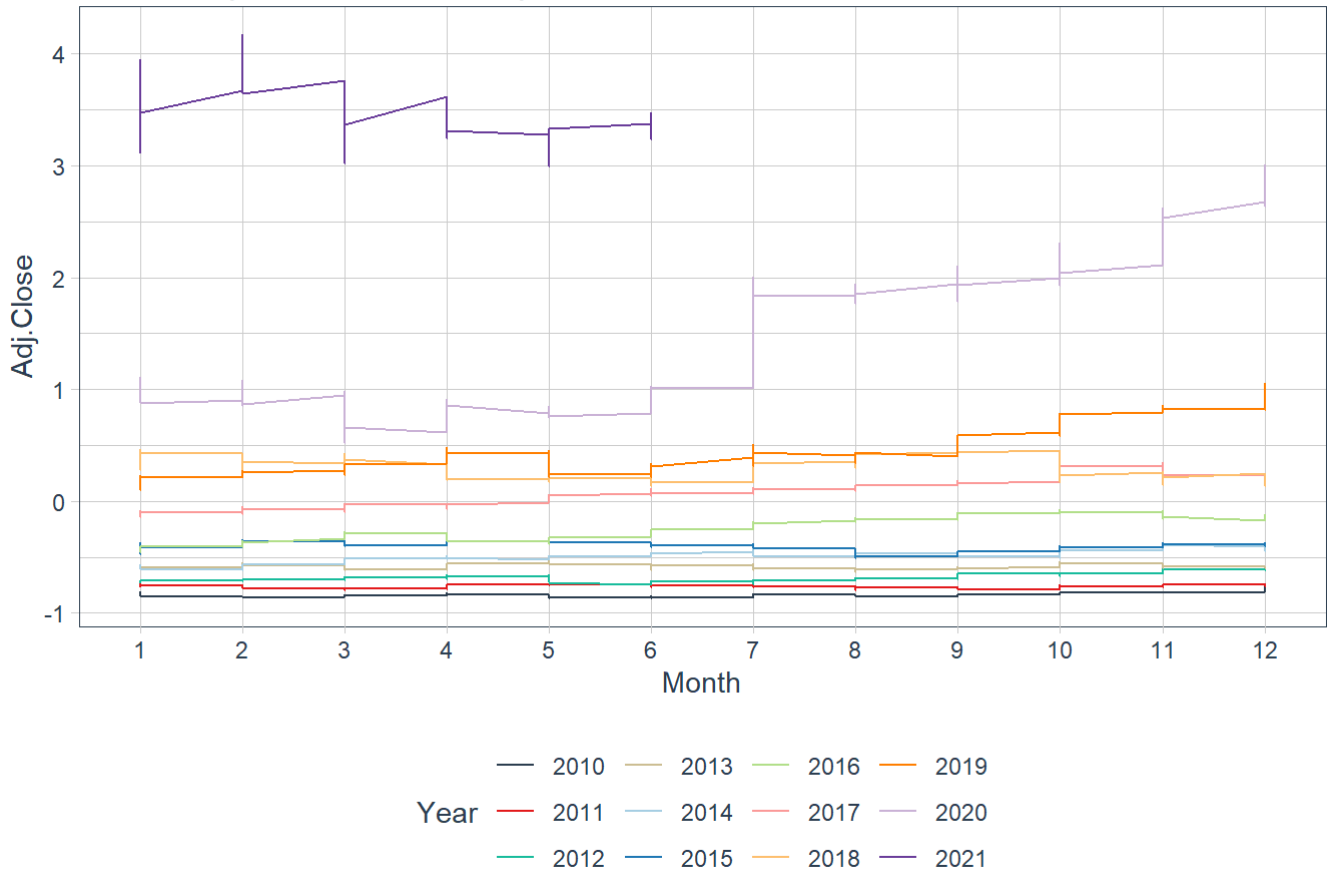
### Time Series Plot



```
# Create a plot of individual years
```

```
df_tsm %>%
  plot_time_series(month, scale(Adj.Close, center = TRUE, scale = TRUE),
    .smooth= FALSE,
    .color_var = lubridate::year(date),
    .interactive = FALSE,
    .color_lab = "Year",
    .title='Seasonal plot for TSM stock price since 2010',
    .y_lab='Adj.Close',
    .x_lab='Month')
```

Seasonal plot for TSM stock price since 2010



- From this plot, we can also see since 2020, prices are outliers.

```
# Create a STL decomposition plot for the full data
```

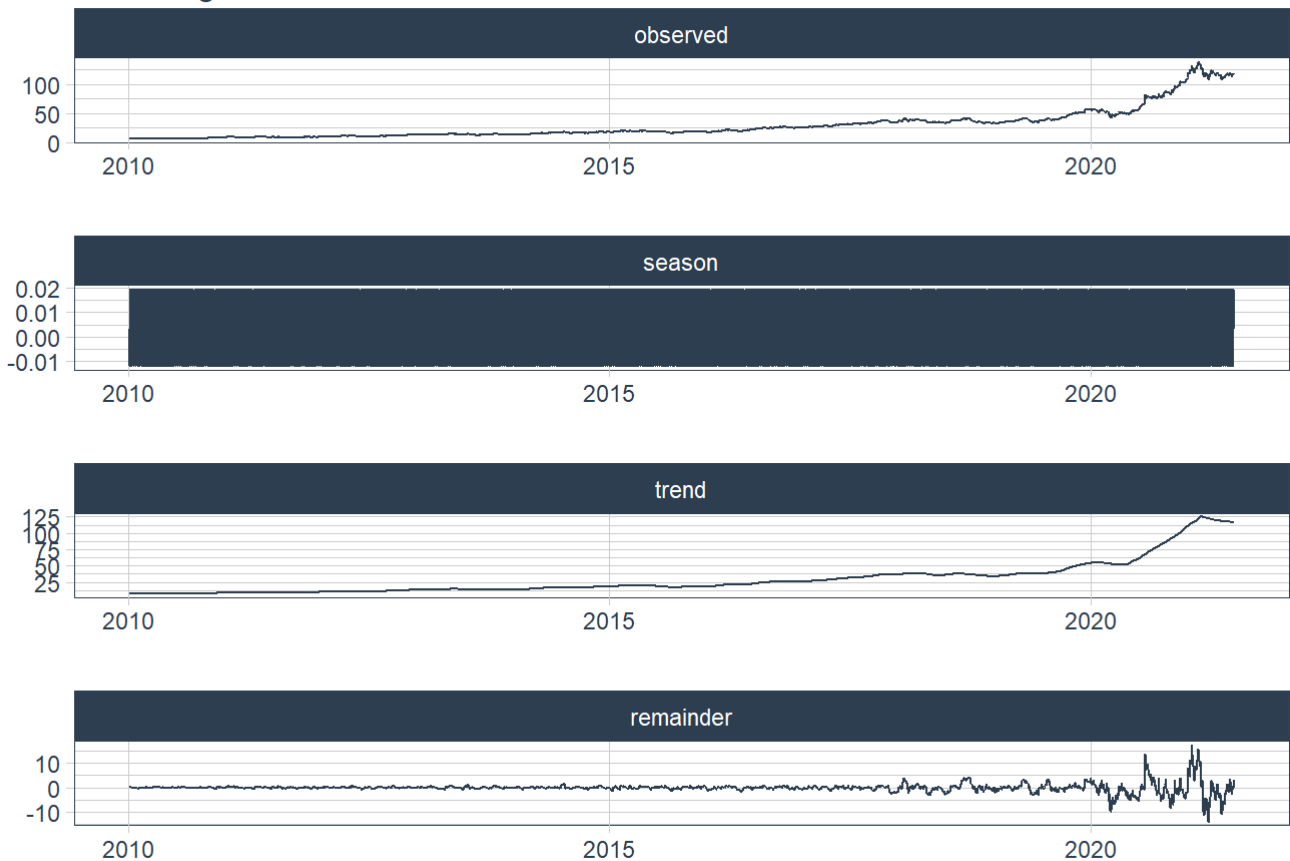
```
df_tsm %>%
  plot_stl_diagnostics(
    date, Adj.Close,
    # Set features to return, desired frequency and trend
    .feature_set = c("observed", "season", "trend", "remainder"),
    .interactive = FALSE)
```

```
## frequency = 5 observations per 1 week
```

```
## trend = 64 observations per 3 months
```



## STL Diagnostics



- From this STL decomposable plot, we can see a clear growth trend. Since 2020, remainder fluctuates in a larger scale.

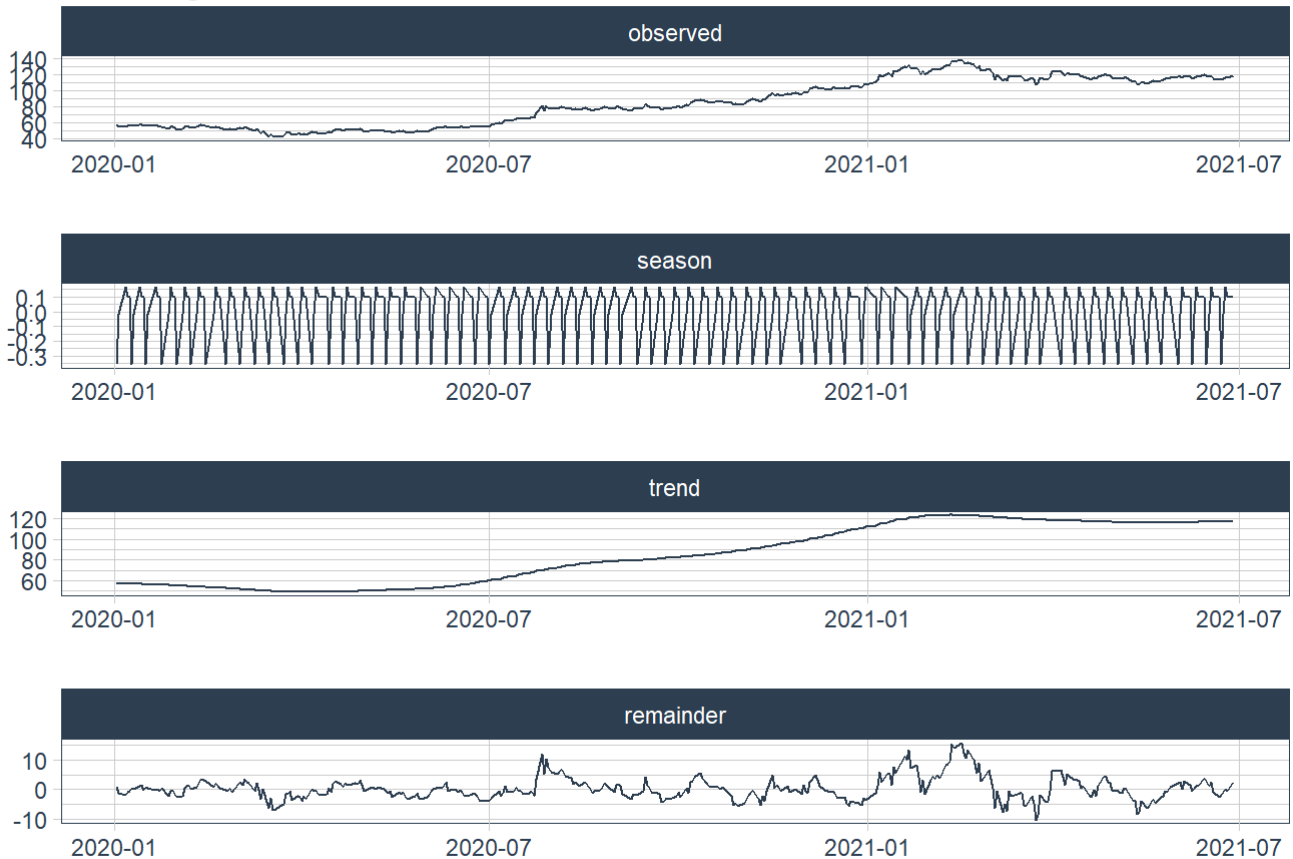
*# Create a STL decomposition plot for a subset of the data since 2020*

```
df_tsm %>%
  filter(date>='2020-01-01') %>%
  plot_stl_diagnostics(
    date, Adj.Close,
    # Set features to return, desired frequency and trend
    .feature_set = c("observed", "season", "trend", "remainder"),
    .interactive = FALSE)
```

## frequency = 5 observations per 1 week

## trend = 64 observations per 3 months

## STL Diagnostics



- If we get a closer look, it seems like TSMC stock price show “week” patterns in season subplot.

```
df_decompose_table <-
  df_tsm %>%
  tk_stl_diagnostics(date, Adj.Close, .frequency = "auto", .trend = "auto")
```

```
## frequency = 5 observations per 1 week
```

```
## trend = 64 observations per 3 months
```

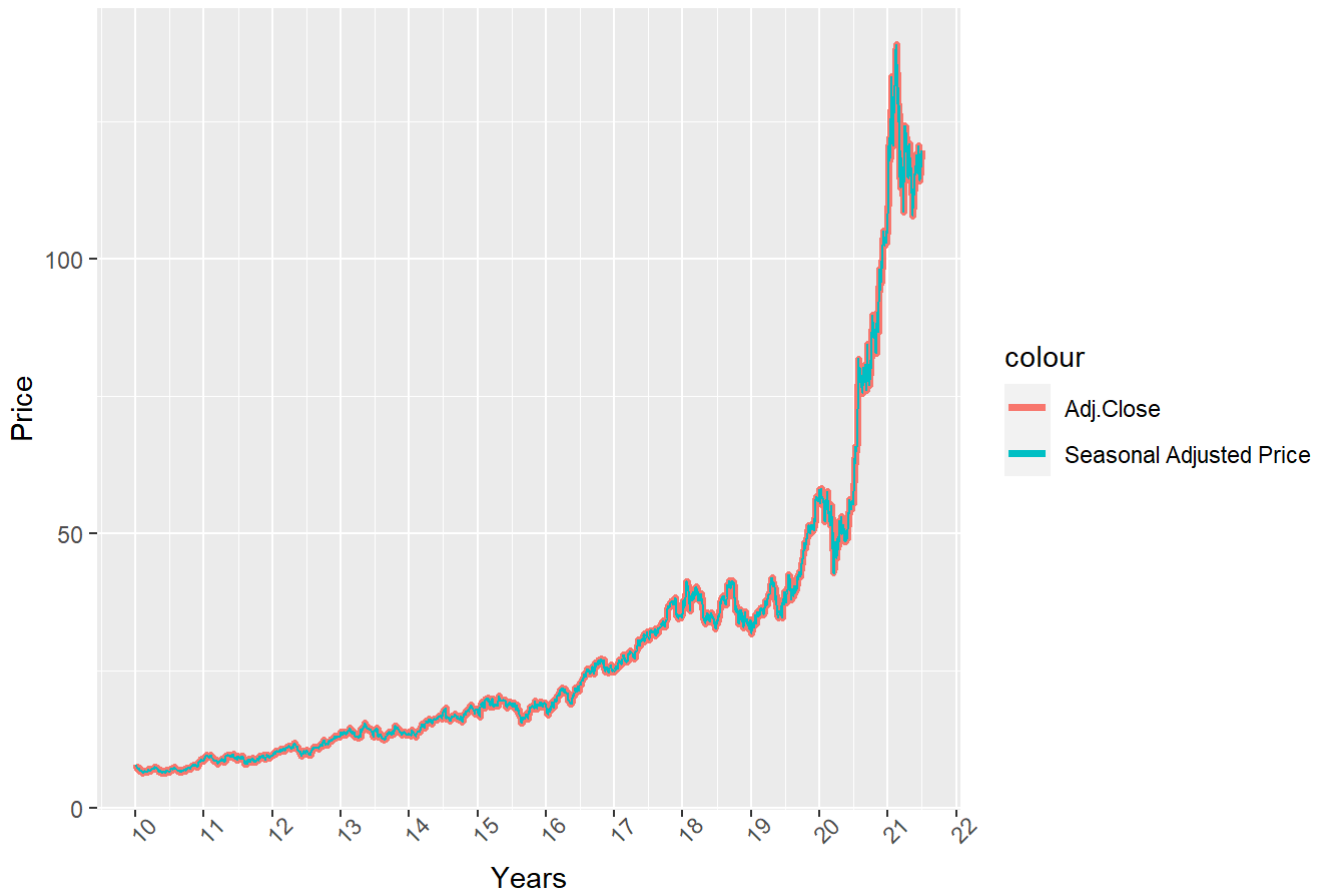
```
df_decompose_table
```

```
## # A tibble: 2,891 x 6
##   date      observed  season trend remainder seasadj
##   <date>      <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1 2010-01-04    7.83  0.00322  7.38    0.448    7.83
## 2 2010-01-05    7.80  0.00138  7.36    0.438    7.80
## 3 2010-01-06    7.77 -0.0122   7.34    0.446    7.78
## 4 2010-01-07    7.51 -0.0119   7.31    0.211    7.53
## 5 2010-01-08    7.51  0.0195   7.29    0.195    7.49
## 6 2010-01-11    7.47  0.00322  7.27    0.193    7.46
## 7 2010-01-12    7.24  0.00138  7.25   -0.0127   7.23
## 8 2010-01-13    7.43 -0.0122   7.23    0.212    7.44
## 9 2010-01-14    7.35 -0.0119   7.21    0.157    7.36
## 10 2010-01-15    7.17  0.0195   7.19   -0.0369   7.15
## # ... with 2,881 more rows
```

```
plot_seasadj <-
  df_decompose_table %>%
  ggplot(aes(x = date)) +
  geom_line(aes(y = observed, color = "Adj.Close"), size = 1.3) +
  geom_line(aes(y = seasadj, color = "Seasonal Adjusted Price")) +
  xlab("Years") + ylab("Price") +
  ggtitle("TSMC stock price over time") +
  scale_x_date(date_breaks = "years" , date_labels = "%y") +
  theme(axis.text.x = element_text(angle = 45))
```

```
plot_seasadj
```

TSMC stock price over time



- I try to use STL method to decompose Adj.Close and plot with original stock price. They are almost matched.

## Data Analyzing

```
# Build an ARIMA model
```

```
arima_auto <-
  arima_reg() %>%
  set_engine("auto_arima") %>%
  fit(Adj.Close ~ date, data = df_tsm)
```

```
## frequency = 5 observations per 1 week
```

```
arima_auto
```

```
## parsnip model object
##
## Fit time: 2.6s
## Series: outcome
## ARIMA(4,2,0)(2,0,2)[5]
##
## Coefficients:
##          ar1      ar2      ar3      ar4      sar1      sar2      sma1      sma2
##       -1.0815  -1.0205  -0.9683  -0.8892  -0.5126  -0.1177  -0.3759  -0.4194
## s.e.   0.0107   0.0124   0.0156   0.0179   0.1099   0.0221   0.1120   0.0993
##
## sigma^2 estimated as 0.7403: log likelihood=-3663.05
## AIC=7344.1   AICc=7344.16   BIC=7397.82
```

```
# Plot the residuals
```

```
models_table <- modeltime_table(
  arima_auto
)

#models_table %>%
#  modeltime_calibrate(new_data = df_tsm) %>%
#  modeltime_residuals() %>%
#  plot_modeltime_residuals(.interactive = FALSE)
```

- In the residual plot shows increasong in residual scale since 2020.

```
# Unit root test: ADF test
```

```
df_tsm %>%
  select(Adj.Close) %>%
  ts(start = c(2010, 1), end = c(2021, 5), frequency = 365) %>%
  tseries::adf.test()
```

```
##
## Augmented Dickey-Fuller Test
##
## data: .
## Dickey-Fuller = -2.5657, Lag order = 15, p-value = 0.3388
## alternative hypothesis: stationary
```

- In ADF test, p-value = 0.3388. This indicates there is unit root and TSMC stock price is “NOT” stationary.

```
# Run ADF test again after first order differencing
```

```
df_tsm %>%
  mutate(diffprices = diff_vec(Adj.Close)) %>%
  select(diffprices) %>%
  drop_na() %>%
  ts(start = c(2010, 1), end = c(2021, 5), frequency = 365) %>%
  tseries::adf.test()
```

```
## diff_vec(): Initial values: 7.831283
```

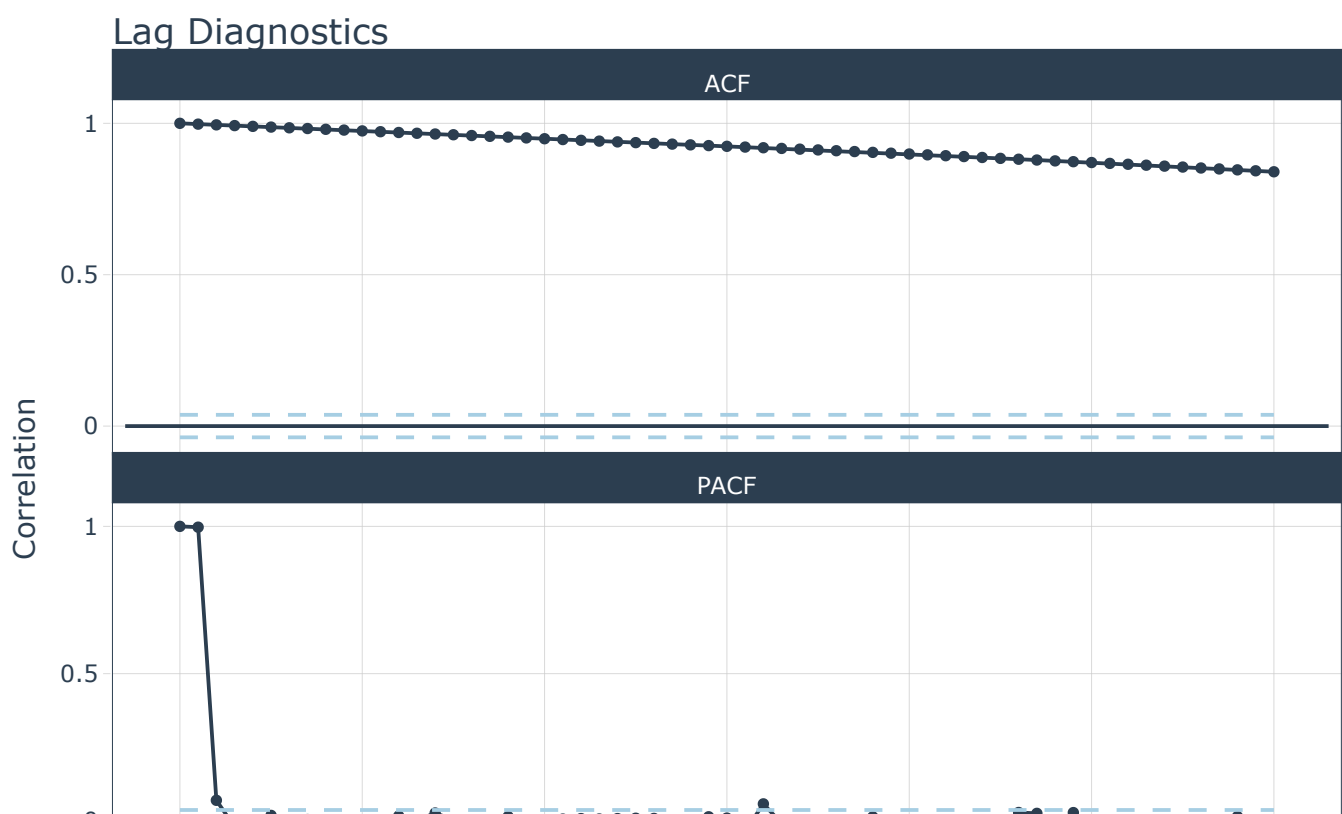
```
## Warning in tseries::adf.test(.): p-value smaller than printed p-value
```

```
##
## Augmented Dickey-Fuller Test
##
## data: .
## Dickey-Fuller = -14.932, Lag order = 15, p-value = 0.01
## alternative hypothesis: stationary
```

- After first order difference, p-value = 0.01. This indicates there is “Not” unit root and TSMC stock price after first order difference is stationary.

```
# Create two PACF plots for before vs. after difference
# Before any difference
```

```
df_tsm %>%
  plot_acf_diagnostics(date, Adj.Close,
    .lags = 60,
    .show_white_noise_bars = TRUE)
```



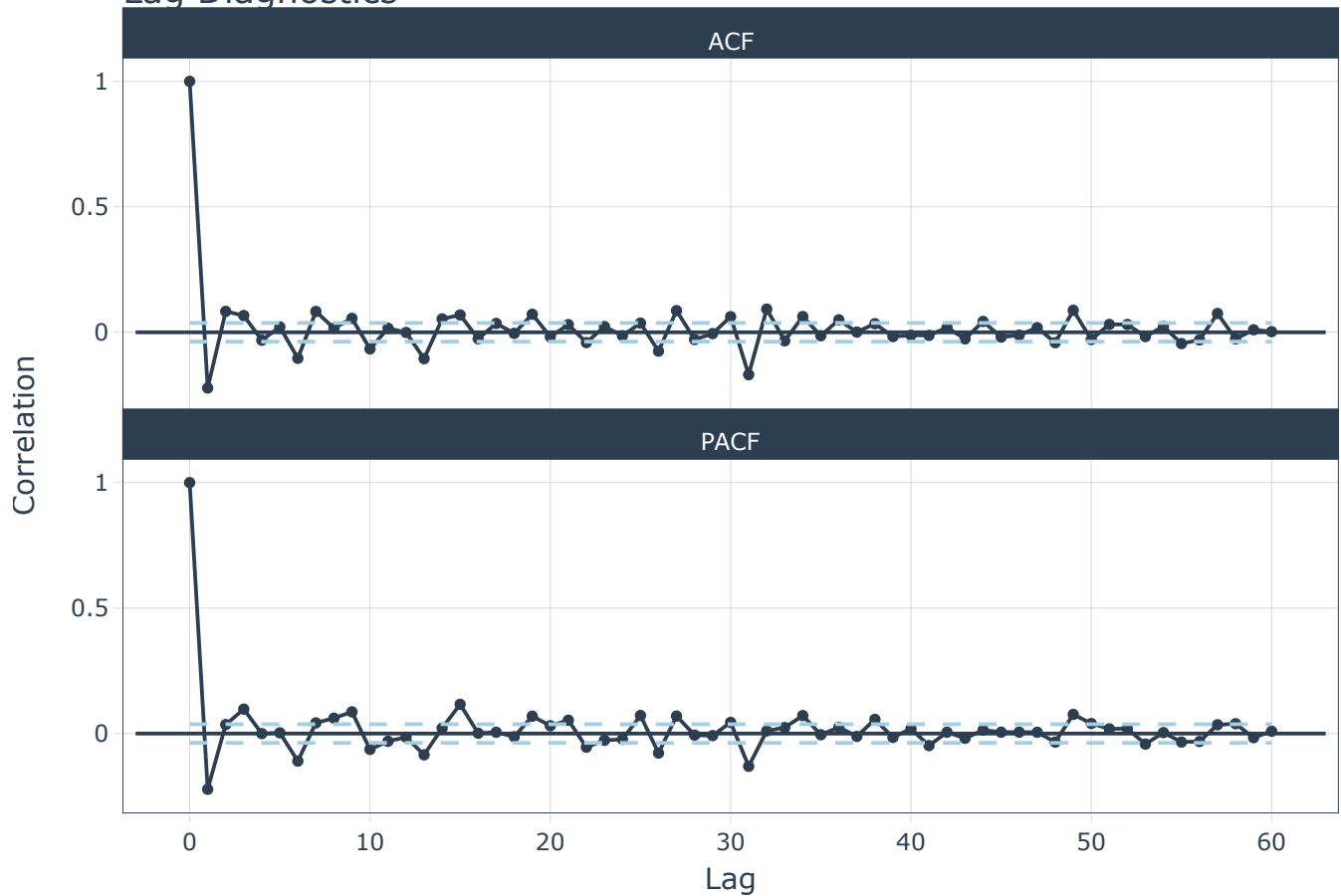


# After 1st order ordinary difference

```
df_tsm%>%
  plot_acf_diagnostics(date, diff_vec(df_tsm$Adj.Close, difference = 1),
    .lags = 60,
    .show_white_noise_bars =TRUE)
```

## diff\_vec(): Initial values: 7.831283

## Lag Diagnostics



- In these before - after comparison plots, we can see that after first order difference both ACF and PACF are with margin.

# Split the data

```
df_train <- df_tsm %>% filter(date < '2020-01-01')
df_test <- df_tsm %>% filter(date >= '2020-01-01')
```

```
# Auto ARIMA: gird search q, d, p
auto_arima <-
  arima_reg(seasonal_period = "auto") %>%
  set_engine(engine = "auto_arima") %>%
  fit(Adj.Close ~ date, data = df_train)
```

```
## frequency = 5 observations per 1 week
```

```
auto_arima
```

```
## parsnip model object
##
## Fit time: 3.1s
## Series: outcome
## ARIMA(3,2,0)(0,0,1)[5]
##
## Coefficients:
##          ar1      ar2      ar3      sma1
##      -0.7811 -0.5237 -0.2255 -0.0315
## s.e.   0.0194  0.0226  0.0195  0.0204
##
## sigma^2 estimated as 0.1586: log likelihood=-1250.78
## AIC=2511.56  AICc=2511.59  BIC=2540.71
```

```
# Auto ARIMA adding other variable
auto_arima_volume <-
  arima_reg(seasonal_period = "auto") %>%
  set_engine(engine = "auto_arima") %>%
  fit(Adj.Close ~ date + Volume, data = df_train)
```

```
## frequency = 5 observations per 1 week
```

```
auto_arima_volume
```

```
## parsnip model object
##
## Fit time: 1.6s
## Series: outcome
## Regression with ARIMA(1,1,3)(0,0,1)[5] errors
##
## Coefficients:
##          ar1      ma1      ma2      ma3      sma1      drift      volume
##      -0.8206  0.7930 -0.0321  0.0340 -0.0499  0.0192         0
## s.e.   0.0664  0.0689  0.0255  0.0218  0.0210  0.0000         0
##
## sigma^2 estimated as 0.1261: log likelihood=-960.99
## AIC=1937.98  AICc=1938.04  BIC=1984.62
```

- Both AIC and BIC are smaller in the second model. With this standard, ARIMA model adding volume is a better model in this case.

```
# Model table

models_tbl <- modeltime_table(
  auto_arima,
  auto_arima_volume
)

calibration_tbl <- models_tbl %>%
  modeltime_calibrate(new_data = df_test)

calibration_tbl %>%
  modeltime_accuracy()%>%
  table_modeltime_accuracy(.interactive = FALSE)
```

## Accuracy Table

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	ARIMA(3,2,0)(0,0,1)[5]	Test	32.19	31.37	19.18	40.14	41.46	0.87
2	REGRESSION WITH ARIMA(1,1,3)(0,0,1)[5] ERRORS	Test	28.91	28.64	17.22	35.09	36.90	0.87

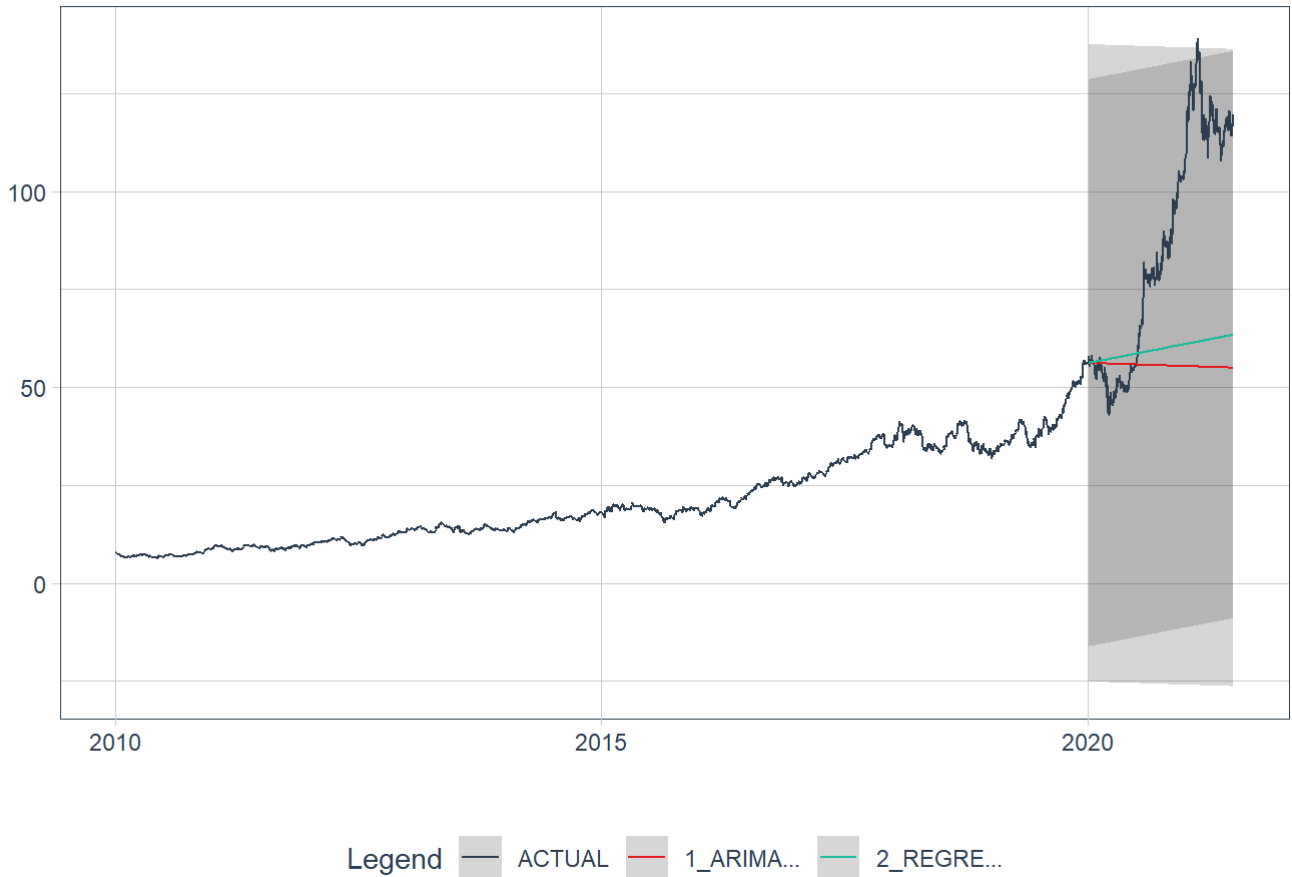
- From the matrix table, we can also see the second model has smaller mae and rmse. With this standard, ARIMA model adding volume is a better model in this case.

```
calibration_tbl %>%
  modeltime_forecast(
    new_data = df_test,
    actual_data = df_tsm
  ) %>%
  plot_modeltime_forecast(
    .legend_max_width = 10,
    .interactive = FALSE
  )
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



Forecast Plot



- Because we know from above analytics that since 2020 TSMC stock prices were growing too fast, and I considered them as outliers. So in my prediction, the second model predicts TSMC stock price should go up but not as fast as reality did.