Final Project

Natalie Pham

12/9/2020

1. Background

The data set below was found on Kaggle, seperately by year. The data set is about flight information of different airline carriers domestically in the US. The flight information includes flight date, airline carriers, origin airports, destination airports, arrival delay, departure delay, weather delay, cancelled-code, etc. The original data set was way before 2015 to June 2020. However, I choose the 5 most recent years to do forecast arrival delays because I think it will yield more accuracy to predict arrival delays for the next 3 years after 2020. Another reason I choose the most recent 5 years is because beginning of 2020 is the pandemic, resulted in airlines' activities have been drastically dropped. Using recent data will not overforecast after June 2020.

2. EDA

```
d15 <- read.csv("2015.csv")
d16 <- read.csv("2016.csv")
d17 <- read.csv("2017.csv")
d18 <- read.csv("2018.csv")
d19 <- read.csv("2019.csv")
d20 <- read.csv("20.csv")
```

```
summary(d15)
```

```
##        FL_DATE          OP_CARRIER      OP_CARRIER_FL_NUM     ORIGIN
##  2015-11-29:  17574   WN     :1261855   Min.   :   1      ATL    : 379424
##  2015-08-07:  17517   DL     : 875881   1st Qu.: 730      ORD    : 313536
##  2015-06-26:  17474   AA     : 725984   Median :1690      DFW    : 260595
##  2015-07-24:  17474   OO     : 588353   Mean   :2173      DEN    : 214191
##  2015-07-10:  17471   EV     : 571977   3rd Qu.:3230      LAX    : 212401
##  2015-07-17:  17469   UA     : 515723   Max.   :9855      SFO    : 162178
##  (Other)   :5714100   (Other):1279306                     (Other):4276754
##       DEST          CRS_DEP_TIME     DEP_TIME       DEP_DELAY
##  ATL    : 379498   Min.   :   1   Min.   :   1    Min.   : -82.00
##  ORD    : 313514   1st Qu.: 917   1st Qu.: 921    1st Qu.:  -5.00
##  DFW    : 260615   Median :1325   Median :1330    Median :  -2.00
##  DEN    : 214135   Mean   :1330   Mean   :1335    Mean   :   9.37
##  LAX    : 212435   3rd Qu.:1730   3rd Qu.:1740    3rd Qu.:   7.00
##  SFO    : 162136   Max.   :2359   Max.   :2400    Max.   :1988.00
##  (Other):4276746                  NA's   :86153   NA's   :86153
##     TAXI_OUT        WHEELS_OFF      WHEELS_ON        TAXI_IN
##  Min.   :  1.00   Min.   :   1    Min.   :   1    Min.   :  1.00
```

```
##  1st Qu.: 11.00   1st Qu.: 935    1st Qu.:1054    1st Qu.:  4.00
##  Median : 14.00   Median :1343    Median :1509    Median :  6.00
##  Mean   : 16.07   Mean   :1357    Mean   :1471    Mean   :  7.43
##  3rd Qu.: 19.00   3rd Qu.:1754    3rd Qu.:1911    3rd Qu.:  9.00
##  Max.   :225.00   Max.   :2400    Max.   :2400    Max.   :248.00
##  NA's   :89047    NA's   :89047   NA's   :92513   NA's   :92513
##   CRS_ARR_TIME     ARR_TIME       ARR_DELAY         CANCELLED
##  Min.   :   1   Min.   :   1    Min.   : -87.00   Min.   :0.00000
##  1st Qu.:1110   1st Qu.:1059    1st Qu.: -13.00   1st Qu.:0.00000
##  Median :1520   Median :1512    Median :  -5.00   Median :0.00000
##  Mean   :1494   Mean   :1476    Mean   :   4.41   Mean   :0.01545
##  3rd Qu.:1918   3rd Qu.:1917    3rd Qu.:   8.00   3rd Qu.:0.00000
##  Max.   :2400   Max.   :2400    Max.   :1971.00   Max.   :1.00000
##                 NA's   :92513   NA's   :105071
##  CANCELLATION_CODE    DIVERTED       CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
##   :5729195         Min.   :0.00000   Min.   : 18.0    Min.   : 14
##  A:  25262         1st Qu.:0.00000   1st Qu.: 85.0    1st Qu.: 82
##  B:  48851         Median :0.00000   Median :123.0    Median :118
##  C:  15749         Mean   :0.00261   Mean   :141.7    Mean   :137
##  D:     22         3rd Qu.:0.00000   3rd Qu.:173.0    3rd Qu.:168
##                    Max.   :1.00000   Max.   :718.0    Max.   :766
##                                      NA's   :6        NA's   :105071
##     AIR_TIME         DISTANCE      CARRIER_DELAY     WEATHER_DELAY
##  Min.   :  7.0    Min.   :  21.0   Min.   :   0      Min.   :   0
##  1st Qu.: 60.0    1st Qu.: 373.0   1st Qu.:   0      1st Qu.:   0
##  Median : 94.0    Median : 647.0   Median :   2      Median :   0
##  Mean   :113.5    Mean   : 822.4   Mean   :  19      Mean   :   3
##  3rd Qu.:144.0    3rd Qu.:1062.0   3rd Qu.:  19      3rd Qu.:   0
##  Max.   :690.0    Max.   :4983.0   Max.   :1971      Max.   :1211
##  NA's   :105071                    NA's   :4755640   NA's   :4755640
##    NAS_DELAY       SECURITY_DELAY    LATE_AIRCRAFT_DELAY Unnamed..27
##  Min.   :   0      Min.   :  0       Min.   :   0        Mode:logical
##  1st Qu.:   0      1st Qu.:  0       1st Qu.:   0        NA's:5819079
##  Median :   2      Median :  0       Median :   3
##  Mean   :  13      Mean   :  0       Mean   :  23
##  3rd Qu.:  18      3rd Qu.:  0       3rd Qu.:  29
##  Max.   :1134      Max.   :573       Max.   :1331
##  NA's   :4755640   NA's   :4755640   NA's   :4755640
```

```
summary(d16)
```

```
##        FL_DATE          OP_CARRIER      OP_CARRIER_FL_NUM     ORIGIN
##  2016-11-27:  17065   WN     :1299444   Min.   :   1      ATL    : 384375
##  2016-07-01:  17050   DL     : 922746   1st Qu.: 711      ORD    : 244082
##  2016-08-05:  17007   AA     : 914495   Median :1639      DEN    : 226136
##  2016-07-29:  17004   OO     : 605933   Mean   :2079      LAX    : 212983
##  2016-07-15:  16998   UA     : 545067   3rd Qu.:2855      DFW    : 196049
##  2016-07-22:  16992   EV     : 490990   Max.   :8402      SFO    : 172358
##  (Other)   :5515542   (Other): 838983                     (Other):4181675
##       DEST          CRS_DEP_TIME     DEP_TIME       DEP_DELAY
##  ATL    : 384252   Min.   :   1   Min.   :   1    Min.   :-204.00
##  ORD    : 243889   1st Qu.: 915   1st Qu.: 917    1st Qu.:  -5.00
##  DEN    : 226225   Median :1325   Median :1328    Median :  -2.00
##  LAX    : 212972   Mean   :1331   Mean   :1334    Mean   :   8.94
```

```
##  DFW    : 196031   3rd Qu.:1735   3rd Qu.:1742    3rd Qu.:   6.00
##  SFO    : 172398   Max.   :2359   Max.   :2400    Max.   :2149.00
##  (Other):4181891                  NA's   :63456   NA's   :63456
##     TAXI_OUT        WHEELS_OFF      WHEELS_ON        TAXI_IN
##  Min.   :  1.00   Min.   :   1    Min.   :   1    Min.   :  1.00
##  1st Qu.: 11.00   1st Qu.: 931    1st Qu.:1049    1st Qu.:  4.00
##  Median : 14.00   Median :1340    Median :1507    Median :  6.00
##  Mean   : 16.19   Mean   :1355    Mean   :1467    Mean   :  7.45
##  3rd Qu.: 19.00   3rd Qu.:1756    3rd Qu.:1914    3rd Qu.:  9.00
##  Max.   :186.00   Max.   :2400    Max.   :2400    Max.   :250.00
##  NA's   :65418    NA's   :65418   NA's   :67844   NA's   :67844
##   CRS_ARR_TIME     ARR_TIME       ARR_DELAY         CANCELLED
##  Min.   :   1   Min.   :   1    Min.   :-152.00   Min.   :0.00000
##  1st Qu.:1105   1st Qu.:1052    1st Qu.: -14.00   1st Qu.:0.00000
##  Median :1520   Median :1511    Median :  -6.00   Median :0.00000
##  Mean   :1491   Mean   :1472    Mean   :   3.52   Mean   :0.01172
##  3rd Qu.:1920   3rd Qu.:1918    3rd Qu.:   6.00   3rd Qu.:0.00000
##  Max.   :2400   Max.   :2400    Max.   :2142.00   Max.   :1.00000
##                 NA's   :67844   NA's   :79513
##  CANCELLATION_CODE    DIVERTED       CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
##   :5551797         Min.   :0.00000   Min.   :  5.0    Min.   : 14.0
##  A:  20279         1st Qu.:0.00000   1st Qu.: 88.0    1st Qu.: 84.0
##  B:  34465         Median :0.00000   Median :126.0    Median :121.0
##  C:  11091         Mean   :0.00243   Mean   :145.4    Mean   :140.2
##  D:     26         3rd Qu.:0.00000   3rd Qu.:178.0    3rd Qu.:173.0
##                    Max.   :1.00000   Max.   :705.0    Max.   :778.0
##                                      NA's   :6        NA's   :79513
##     AIR_TIME        DISTANCE      CARRIER_DELAY     WEATHER_DELAY
##  Min.   :  4.0   Min.   :  25.0   Min.   :   0      Min.   :   0
##  1st Qu.: 62.0   1st Qu.: 391.0   1st Qu.:   0      1st Qu.:   0
##  Median : 97.0   Median : 678.0   Median :   1      Median :   0
##  Mean   :116.5   Mean   : 850.1   Mean   :  20      Mean   :   3
##  3rd Qu.:148.0   3rd Qu.:1091.0   3rd Qu.:  18      3rd Qu.:   0
##  Max.   :723.0   Max.   :4983.0   Max.   :2142      Max.   :1157
##  NA's   :79513                    NA's   :4653419   NA's   :4653419
##    NAS_DELAY       SECURITY_DELAY    LATE_AIRCRAFT_DELAY Unnamed..27
##  Min.   :   0      Min.   :  0       Min.   :   0        Mode:logical
##  1st Qu.:   0      1st Qu.:  0       1st Qu.:   0        NA's:5617658
##  Median :   2      Median :  0       Median :   2
##  Mean   :  15      Mean   :  0       Mean   :  24
##  3rd Qu.:  19      3rd Qu.:  0       3rd Qu.:  30
##  Max.   :1446      Max.   :474       Max.   :1484
##  NA's   :4653419   NA's   :4653419   NA's   :4653419
```

```
summary(d17)
```

```
##        FL_DATE          OP_CARRIER      OP_CARRIER_FL_NUM     ORIGIN
##  2017-07-14:  17284   WN     :1329444   Min.   :   1      ATL    : 364655
##  2017-06-30:  17276   DL     : 923560   1st Qu.: 736      ORD    : 266460
##  2017-08-04:  17272   AA     : 896348   Median :1679      DEN    : 223165
##  2017-07-28:  17265   OO     : 706527   Mean   :2143      LAX    : 214297
##  2017-07-21:  17262   UA     : 584481   3rd Qu.:3064      DFW    : 181208
##  2017-07-17:  17260   EV     : 339541   Max.   :8402      SFO    : 174631
##  (Other)   :5571002   (Other): 894720                     (Other):4250205
```

```
##       DEST          CRS_DEP_TIME     DEP_TIME       DEP_DELAY
##  ATL    : 364596   Min.   :   1   Min.   :   1    Min.   :-234.00
##  ORD    : 266377   1st Qu.: 912   1st Qu.: 914    1st Qu.:  -5.00
##  DEN    : 223234   Median :1323   Median :1327    Median :  -2.00
##  LAX    : 214312   Mean   :1330   Mean   :1334    Mean   :   9.73
##  DFW    : 181208   3rd Qu.:1735   3rd Qu.:1743    3rd Qu.:   6.00
##  SFO    : 174674   Max.   :2359   Max.   :2400    Max.   :2755.00
##  (Other):4250220                   NA's   :80308   NA's   :80343
##     TAXI_OUT        WHEELS_OFF      WHEELS_ON        TAXI_IN
##  Min.   :  0.00   Min.   :   1    Min.   :   1    Min.   :  0.00
##  1st Qu.: 11.00   1st Qu.: 930    1st Qu.:1046    1st Qu.:  4.00
##  Median : 14.00   Median :1340    Median :1506    Median :  6.00
##  Mean   : 16.78   Mean   :1356    Mean   :1465    Mean   :  7.51
##  3rd Qu.: 20.00   3rd Qu.:1758    3rd Qu.:1913    3rd Qu.:  9.00
##  Max.   :183.00   Max.   :2400    Max.   :2400    Max.   :414.00
##  NA's   :82145    NA's   :82141   NA's   :84674   NA's   :84674
##   CRS_ARR_TIME     ARR_TIME       ARR_DELAY         CANCELLED
##  Min.   :   1   Min.   :   1    Min.   :-238.00   Min.   :0.00000
##  1st Qu.:1103   1st Qu.:1050    1st Qu.: -15.00   1st Qu.:0.00000
##  Median :1520   Median :1510    Median :  -6.00   Median :0.00000
##  Mean   :1489   Mean   :1469    Mean   :   4.33   Mean   :0.01457
##  3rd Qu.:1920   3rd Qu.:1918    3rd Qu.:   7.00   3rd Qu.:0.00000
##  Max.   :2359   Max.   :2400    Max.   :2189.00   Max.   :1.00000
##                 NA's   :84674   NA's   :95211
##  CANCELLATION_CODE    DIVERTED        CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
##   :5591928         Min.   :0.000000   Min.   :  1      Min.   : 15.0
##  A:  18602         1st Qu.:0.000000   1st Qu.: 90      1st Qu.: 85.0
##  B:  48459         Median :0.000000   Median :128      Median :123.0
##  C:  15313         Mean   :0.002208   Mean   :147      Mean   :141.8
##  D:    319         3rd Qu.:0.000000   3rd Qu.:180      3rd Qu.:175.0
##                    Max.   :1.000000   Max.   :718      Max.   :784.0
##                                       NA's   :7        NA's   :95211
##     AIR_TIME        DISTANCE      CARRIER_DELAY     WEATHER_DELAY
##  Min.   :  7.0   Min.   :  31.0   Min.   :   0      Min.   :   0
##  1st Qu.: 62.0   1st Qu.: 391.0   1st Qu.:   0      1st Qu.:   0
##  Median : 98.0   Median : 680.0   Median :   1      Median :   0
##  Mean   :117.5   Mean   : 856.7   Mean   :  20      Mean   :   3
##  3rd Qu.:149.0   3rd Qu.:1097.0   3rd Qu.:  17      3rd Qu.:   0
##  Max.   :712.0   Max.   :4983.0   Max.   :1934      Max.   :1934
##  NA's   :95211                    NA's   :4645148   NA's   :4645148
##    NAS_DELAY       SECURITY_DELAY    LATE_AIRCRAFT_DELAY Unnamed..27
##  Min.   :   0      Min.   :  0       Min.   :   0        Mode:logical
##  1st Qu.:   0      1st Qu.:  0       1st Qu.:   0        NA's:5674621
##  Median :   2      Median :  0       Median :   4
##  Mean   :  16      Mean   :  0       Mean   :  25
##  3rd Qu.:  19      3rd Qu.:  0       3rd Qu.:  31
##  Max.   :1605      Max.   :827       Max.   :1756
##  NA's   :4645148   NA's   :4645148   NA's   :4645148
```

```
summary(d18)
```

```
##        FL_DATE          OP_CARRIER      OP_CARRIER_FL_NUM     ORIGIN
##  2018-11-25:  22160   WN     :1352552   Min.   :   1      ATL    : 390046
##  2018-07-13:  22022   DL     : 949283   1st Qu.:1029      ORD    : 332953
```

```
##  2018-07-20:  22002   AA     : 916818   Median :2131      DFW    : 279298
##  2018-07-27:  21997   OO     : 774137   Mean   :2608      DEN    : 235989
##  2018-08-03:  21990   UA     : 621565   3rd Qu.:4074      CLT    : 233317
##  2018-07-12:  21966   YX     : 316090   Max.   :7909      LAX    : 221486
##  (Other)   :7081309   (Other):2283001                     (Other):5520357
##       DEST          CRS_DEP_TIME     DEP_TIME        DEP_DELAY
##  ATL    : 390079   Min.   :   1   Min.   :   1     Min.   :-122.00
##  ORD    : 332942   1st Qu.: 915   1st Qu.: 916     1st Qu.:  -5.00
##  DFW    : 279272   Median :1320   Median :1326     Median :  -2.00
##  DEN    : 236020   Mean   :1330   Mean   :1334     Mean   :   9.97
##  CLT    : 233309   3rd Qu.:1735   3rd Qu.:1744     3rd Qu.:   7.00
##  LAX    : 221516   Max.   :2359   Max.   :2400     Max.   :2710.00
##  (Other):5520308                  NA's   :112317   NA's   :117234
##     TAXI_OUT        WHEELS_OFF       WHEELS_ON         TAXI_IN
##  Min.   :  1.00   Min.   :   1     Min.   :   1     Min.   :  1.0
##  1st Qu.: 11.00   1st Qu.: 932     1st Qu.:1044     1st Qu.:  4.0
##  Median : 15.00   Median :1340     Median :1502     Median :  6.0
##  Mean   : 17.41   Mean   :1358     Mean   :1462     Mean   :  7.6
##  3rd Qu.: 20.00   3rd Qu.:1759     3rd Qu.:1911     3rd Qu.:  9.0
##  Max.   :196.00   Max.   :2400     Max.   :2400     Max.   :259.0
##  NA's   :115830   NA's   :115829   NA's   :119246   NA's   :119246
##   CRS_ARR_TIME     ARR_TIME        ARR_DELAY         CANCELLED
##  Min.   :   1   Min.   :   1     Min.   :-120.00   Min.   :0.00000
##  1st Qu.:1100   1st Qu.:1049     1st Qu.: -14.00   1st Qu.:0.00000
##  Median :1515   Median :1506     Median :  -6.00   Median :0.00000
##  Mean   :1486   Mean   :1467     Mean   :   5.05   Mean   :0.01616
##  3rd Qu.:1919   3rd Qu.:1916     3rd Qu.:   8.00   3rd Qu.:0.00000
##  Max.   :2400   Max.   :2400     Max.   :2692.00   Max.   :1.00000
##                 NA's   :119245   NA's   :137040
##  CANCELLATION_CODE    DIVERTED        CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
##   :7096862         Min.   :0.000000   Min.   :-99.0    Min.   : 14.0
##  A:  29484         1st Qu.:0.000000   1st Qu.: 88.0    1st Qu.: 83.0
##  B:  61984         Median :0.000000   Median :122.0    Median :118.0
##  C:  25072         Mean   :0.002476   Mean   :141.1    Mean   :136.5
##  D:     44         3rd Qu.:0.000000   3rd Qu.:171.0    3rd Qu.:167.0
##                    Max.   :1.000000   Max.   :704.0    Max.   :757.0
##                                       NA's   :10       NA's   :134442
##     AIR_TIME         DISTANCE    CARRIER_DELAY     WEATHER_DELAY
##  Min.   :  7.0    Min.   :  31   Min.   :   0      Min.   :   0
##  1st Qu.: 60.0    1st Qu.: 363   1st Qu.:   0      1st Qu.:   0
##  Median : 92.0    Median : 632   Median :   0      Median :   0
##  Mean   :111.5    Mean   : 800   Mean   :  19      Mean   :   4
##  3rd Qu.:141.0    3rd Qu.:1034   3rd Qu.:  17      3rd Qu.:   0
##  Max.   :696.0    Max.   :4983   Max.   :2109      Max.   :2692
##  NA's   :134442                  NA's   :5860736   NA's   :5860736
##    NAS_DELAY       SECURITY_DELAY    LATE_AIRCRAFT_DELAY Unnamed..27
##  Min.   :   0      Min.   :  0       Min.   :   0        Mode:logical
##  1st Qu.:   0      1st Qu.:  0       1st Qu.:   0        NA's:7213446
##  Median :   3      Median :  0       Median :   3
##  Mean   :  16      Mean   :  0       Mean   :  26
##  3rd Qu.:  20      3rd Qu.:  0       3rd Qu.:  31
##  Max.   :1848      Max.   :987       Max.   :2454
##  NA's   :5860736   NA's   :5860736   NA's   :5860736
```

```
summary(d19)
```

```
##        FL_DATE        OP_UNIQUE_CARRIER OP_CARRIER_FL_NUM     ORIGIN
##  2019-12-01:  22784   WN     :1363946   Min.   :   1       ATL    : 395009
##  2019-08-02:  22440   DL     : 991986   1st Qu.:1025       ORD    : 339606
##  2019-08-05:  22423   AA     : 946776   Median :2158       DFW    : 304344
##  2019-08-01:  22403   OO     : 836445   Mean   :2557       DEN    : 252026
##  2019-07-26:  22392   UA     : 625910   3rd Qu.:3917       CLT    : 235496
##  2019-07-19:  22387   YX     : 329149   Max.   :7933       LAX    : 219952
##  (Other)   :7287208   (Other):2327825                      (Other):5675604
##       DEST            DEP_TIME       DEP_DELAY          TAXI_OUT
##  ATL    : 395026   Min.   :   1     Min.   : -82.00   Min.   :  1.00
##  ORD    : 339569   1st Qu.: 914     1st Qu.:  -5.00   1st Qu.: 11.00
##  DFW    : 304346   Median :1327     Median :  -2.00   Median : 15.00
##  DEN    : 252064   Mean   :1335     Mean   :  10.92   Mean   : 17.39
##  CLT    : 235490   3rd Qu.:1746     3rd Qu.:   7.00   3rd Qu.: 20.00
##  LAX    : 219996   Max.   :2400     Max.   :2710.00   Max.   :227.00
##  (Other):5675546   NA's   :130086   NA's   :130110    NA's   :133977
##    WHEELS_OFF       WHEELS_ON         TAXI_IN          ARR_TIME
##  Min.   :   1     Min.   :   1     Min.   :  1.00   Min.   :   1
##  1st Qu.: 930     1st Qu.:1042     1st Qu.:  4.00   1st Qu.:1046
##  Median :1340     Median :1500     Median :  6.00   Median :1504
##  Mean   :1358     Mean   :1459     Mean   :  7.74   Mean   :1463
##  3rd Qu.:1801     3rd Qu.:1912     3rd Qu.:  9.00   3rd Qu.:1917
##  Max.   :2400     Max.   :2400     Max.   :316.00   Max.   :2400
##  NA's   :133977   NA's   :137647   NA's   :137647   NA's   :137646
##    ARR_DELAY          AIR_TIME         DISTANCE      CARRIER_DELAY
##  Min.   : -99.00   Min.   :   4.0   Min.   :  31.0   Min.   :   0
##  1st Qu.: -15.00   1st Qu.:  60.0   1st Qu.: 369.0   1st Qu.:   0
##  Median :  -6.00   Median :  93.0   Median : 640.0   Median :   0
##  Mean   :   5.41   Mean   : 111.6   Mean   : 800.5   Mean   :  21
##  3rd Qu.:   7.00   3rd Qu.: 141.0   3rd Qu.:1034.0   3rd Qu.:  18
##  Max.   :2695.00   Max.   :1557.0   Max.   :5095.0   Max.   :2695
##  NA's   :153805    NA's   :153805                    NA's   :6032784
##  WEATHER_DELAY       NAS_DELAY       SECURITY_DELAY    LATE_AIRCRAFT_DELAY
##  Min.   :   0      Min.   :   0      Min.   :   0      Min.   :   0
##  1st Qu.:   0      1st Qu.:   0      1st Qu.:   0      1st Qu.:   0
##  Median :   0      Median :   2      Median :   0      Median :   3
##  Mean   :   4      Mean   :  17      Mean   :   0      Mean   :  27
##  3rd Qu.:   0      3rd Qu.:  20      3rd Qu.:   0      3rd Qu.:  33
##  Max.   :1847      Max.   :1741      Max.   :1078      Max.   :2206
##  NA's   :6032784   NA's   :6032784   NA's   :6032784   NA's   :6032784
##     X
##  Mode:logical
##  NA's:7422037
##
##
##
##
##
```

```r
summary(d20)
```

```
##        FL_DATE          OP_CARRIER     OP_CARRIER_FL_NUM     ORIGIN
##  2020-03-20:  22172   WN     :531577   Min.   :   1      ATL    : 128140
##  2020-03-19:  22157   AA     :320328   1st Qu.:1033      DFW    : 112089
##  2020-03-13:  22100   DL     :306623   Median :2172      ORD    : 106249
##  2020-03-16:  22071   OO     :304032   Mean   :2599      DEN    :  94666
##  2020-03-12:  22023   UA     :180724   3rd Qu.:4087      CLT    :  89780
##  2020-03-23:  21985   YX     :113653   Max.   :9888      LAX    :  69184
##  (Other)   :2415066   (Other):790637                     (Other):1947466
##       DEST          CRS_DEP_TIME       DEP_TIME        DEP_DELAY
##  ATL    : 128238   Min.   :   1     Min.   : -80.0   Min.   :  1.00
##  DFW    : 112038   1st Qu.: 920     1st Qu.:  -7.0   1st Qu.: 11.00
##  ORD    : 106282   Median :1322     Median :  -4.0   Median : 14.00
##  DEN    :  94703   Mean   :1325     Mean   :   3.4   Mean   : 15.85
##  CLT    :  89746   3rd Qu.:1731     3rd Qu.:   0.0   3rd Qu.: 18.00
##  LAX    :  69227   Max.   :2400     Max.   :2814.0   Max.   :189.00
##  (Other):1947340   NA's   :263935   NA's   :263990   NA's   :264407
##     TAXI_OUT        WHEELS_OFF       WHEELS_ON         TAXI_IN
##  Min.   :   1     Min.   :   1     Min.   :  1.00   Min.   :   1
##  1st Qu.: 934     1st Qu.:1058     1st Qu.:  4.00   1st Qu.:1102
##  Median :1334     Median :1506     Median :  6.00   Median :1510
##  Mean   :1349     Mean   :1473     Mean   :  6.96   Mean   :1478
##  3rd Qu.:1744     3rd Qu.:1905     3rd Qu.:  8.00   3rd Qu.:1910
##  Max.   :2400     Max.   :2400     Max.   :194.00   Max.   :2400
##  NA's   :264407   NA's   :265285   NA's   :265285   NA's   :265285
##   CRS_ARR_TIME        ARR_TIME        ARR_DELAY      CANCELLED
##  Min.   :-117.00   Min.   :  7      Min.   :  29   Min.   :   0.0
##  1st Qu.: -20.00   1st Qu.: 60      1st Qu.: 368   1st Qu.:   0.0
##  Median : -11.00   Median : 93      Median : 632   Median :   0.0
##  Mean   :  -4.63   Mean   :110      Mean   : 782   Mean   :  25.2
##  3rd Qu.:  -1.00   3rd Qu.:140      3rd Qu.:1020   3rd Qu.:  23.0
##  Max.   :2794.00   Max.   :698      Max.   :5095   Max.   :2560.0
##  NA's   :268764    NA's   :268764                  NA's   :2295172
##  CANCELLATION_CODE    DIVERTED       CRS_ELAPSED_TIME  ACTUAL_ELAPSED_TIME
##  Min.   :   0.0    Min.   :   0.0    Min.   :   0.0    Min.   :   0.0
##  1st Qu.:   0.0    1st Qu.:   0.0    1st Qu.:   0.0    1st Qu.:   0.0
##  Median :   0.0    Median :   2.0    Median :   0.0    Median :   0.0
##  Mean   :   3.8    Mean   :  14.5    Mean   :   0.1    Mean   :  20.7
##  3rd Qu.:   0.0    3rd Qu.:  19.0    3rd Qu.:   0.0    3rd Qu.:  22.0
##  Max.   :1525.0    Max.   :1462.0    Max.   :1185.0    Max.   :2228.0
##  NA's   :2295172   NA's   :2295172   NA's   :2295172   NA's   :2295172
##  AIR_TIME       DISTANCE       CARRIER_DELAY  WEATHER_DELAY  NAS_DELAY
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:2547574   NA's:2547574   NA's:2547574   NA's:2547574   NA's:2547574
##
##
##
##
##
##  SECURITY_DELAY LATE_AIRCRAFT_DELAY Unnamed..27
##  Mode:logical   Mode:logical        Mode:logical
##  NA's:2547574   NA's:2547574        NA's:2547574
```

```
##
##
##
##
##
```

- Row binding to create data frame for time series using 'ARR_DELAY' to forecast

```
d15 <- d15[,c("FL_DATE","ARR_DELAY")]
d16 <- d16[,c("FL_DATE","ARR_DELAY")]
d17 <- d17[,c("FL_DATE","ARR_DELAY")]
d18 <- d18[,c("FL_DATE","ARR_DELAY")]
d19 <- d19[,c("FL_DATE","ARR_DELAY")]
d20 <- d20[,c("FL_DATE","ARR_DELAY")]

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(backports)
```

```
## Warning: package 'backports' was built under R version 3.6.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.3.1
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

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

```
library(fpp2)
```

```
## Warning: package 'fpp2' was built under R version 3.6.3
```

```
## -- Attaching packages --------------------------------------- fpp2 2.4 --
```

```
## v fma       2.4     v expsmooth 2.3
```

```
## Warning: package 'fma' was built under R version 3.6.3
```

```
## Warning: package 'expsmooth' was built under R version 3.6.3
```

```
##
```

```
library(seasonal)
```

```
## Warning: package 'seasonal' was built under R version 3.6.3
```

```
##
## Attaching package: 'seasonal'
```

```
## The following object is masked from 'package:tibble':
##
##     view
```

```
d <- rbind(d15,d16,d17,d18,d19,d20)
t <- d %>% filter(ARR_DELAY != "NA")
```

- Transform date format

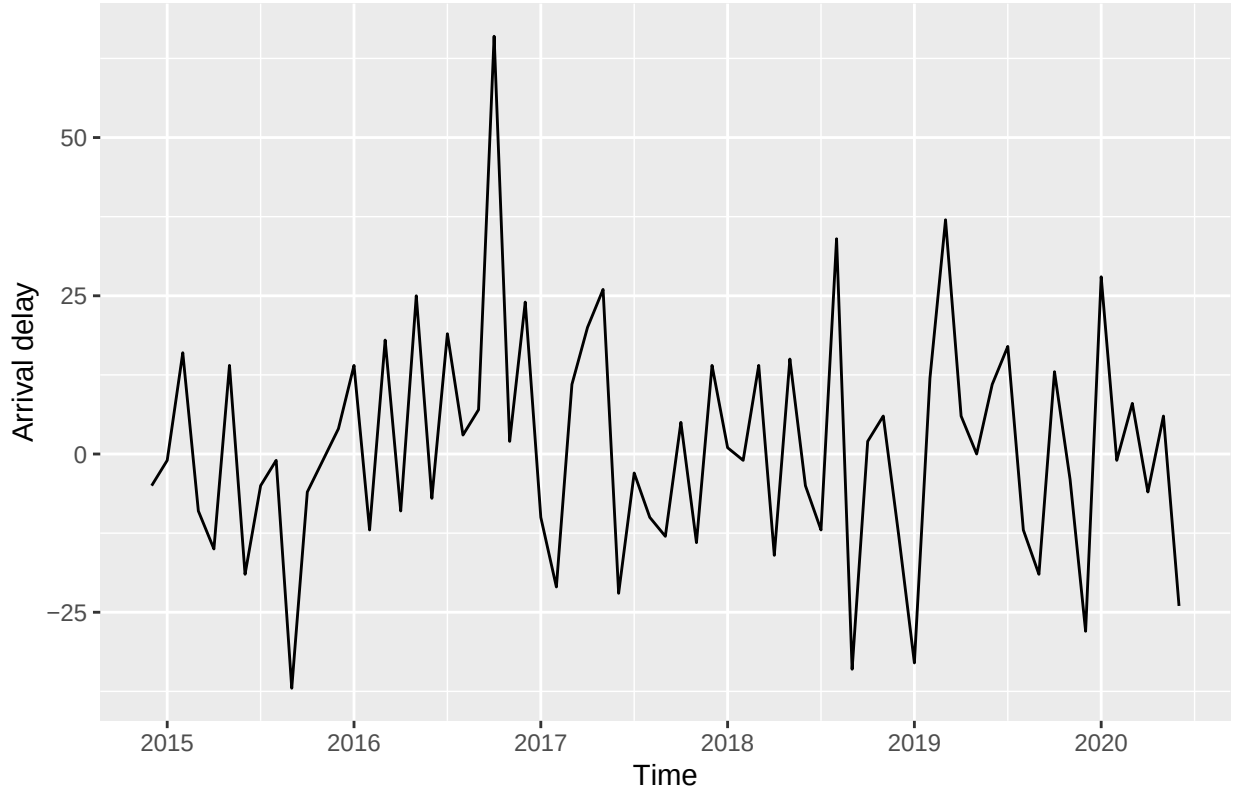
```
date_change <- as.Date(t$FL_DATE)
y<- as.POSIXct(date_change, format = "%m/%d/%Y")

t$year <- format(y, "%Y")
t$month <- format(y,"%b")
```

```
t2 <- t[,c(3,4,2)]
```

```
s <- ts(data = t2$ARR_DELAY, start = c(2014,12), end = c(2020,6), frequency = 12)
```

Plot total arrival delay times over the years

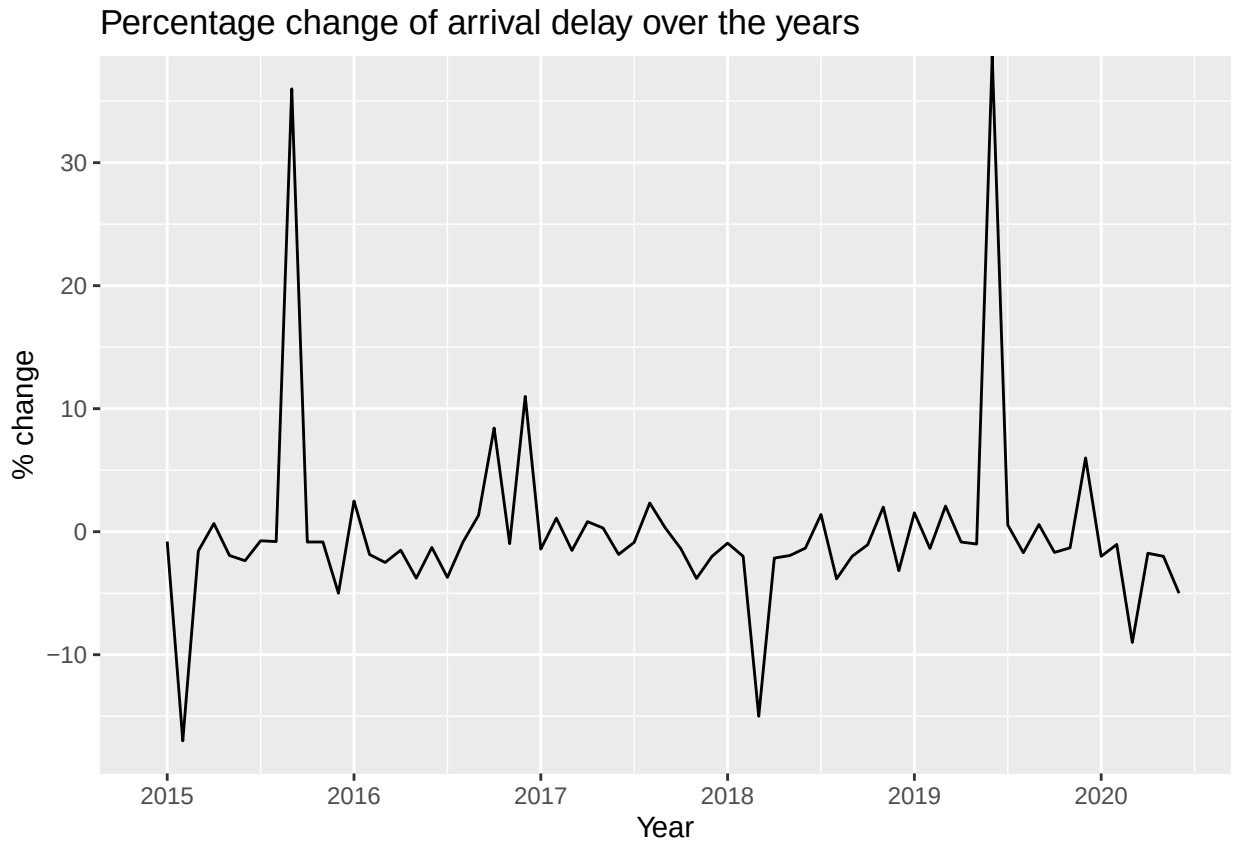
```
autoplot(s) + ylab("Arrival delay")
```



Percentage change of Arrival delay over the month over years

```
library(tidyverse)
c <- t2 %>% group_by(year) %>% group_by(month) %>%
  mutate(pct_change = (ARR_DELAY/lag(ARR_DELAY) - 1))
```

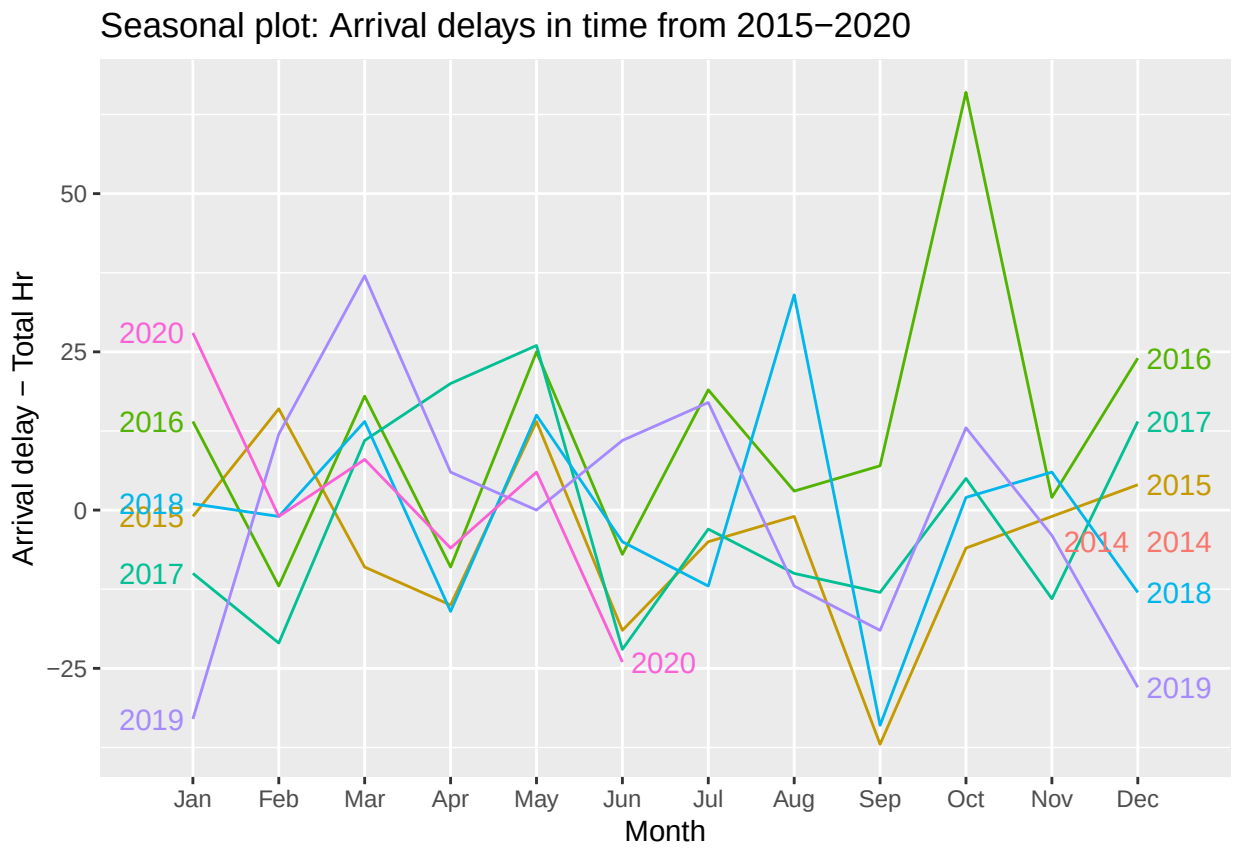
```r
sc <- ts(data = c$pct_change, start = c(2014,12), end = c(2020,6), frequency = 12)

autoplot(sc) + ylab("% change") + xlab("Year") + ggtitle("Percentage change of arrival delay over the y
```



3. Forecast models

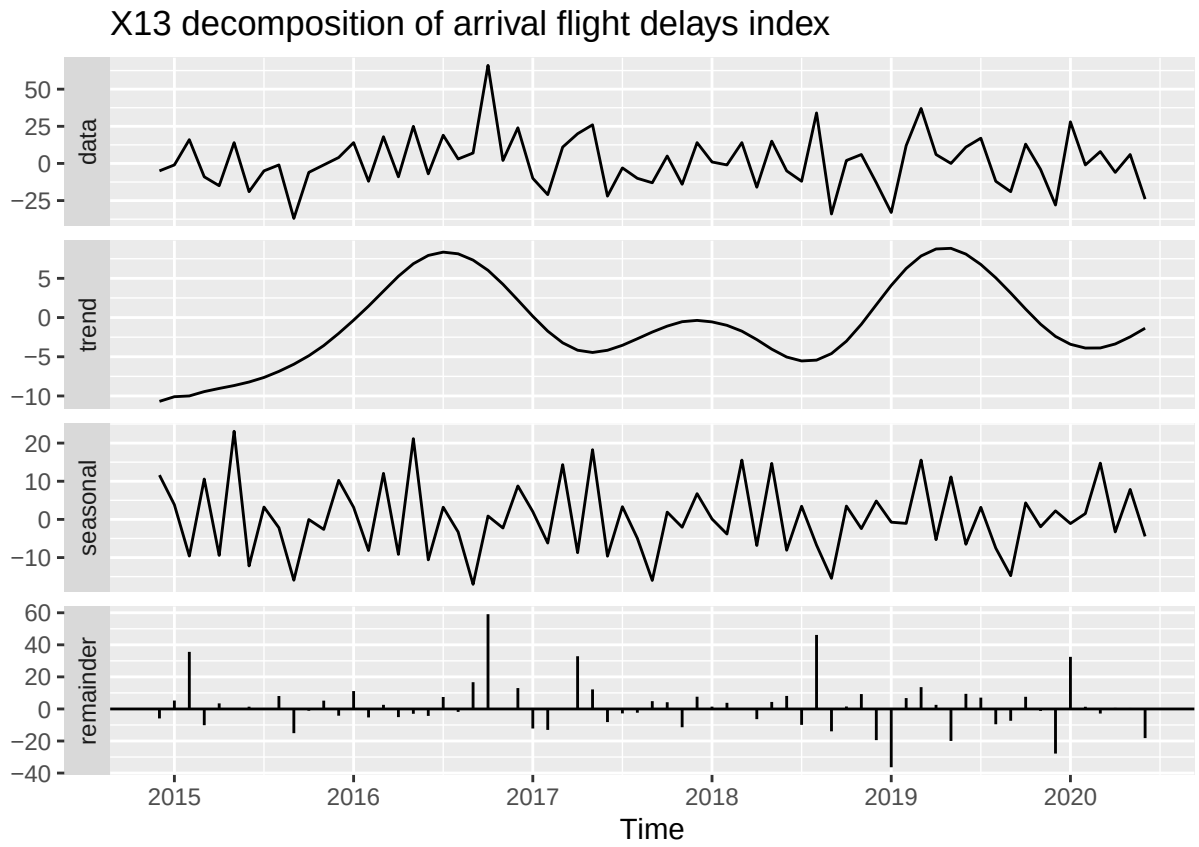
- Seasonal plot

```r
ggseasonplot(s, year.labels=TRUE, year.labels.left=TRUE) +
  ylab("Arrival delay - Total Hr") +
  ggtitle("Seasonal plot: Arrival delays in time from 2015-2020")
```



- Time seriese decomposition method - X13

```
fit <- s %>% seas(x11="")
autoplot(fit) +
  ggtitle("X13 decomposition of arrival flight delays index")
```



- Moving Average Partition data

```
train <- window(s, end = c(2017, 12))
test <- window(s, start = c(2018, 1))

train
```

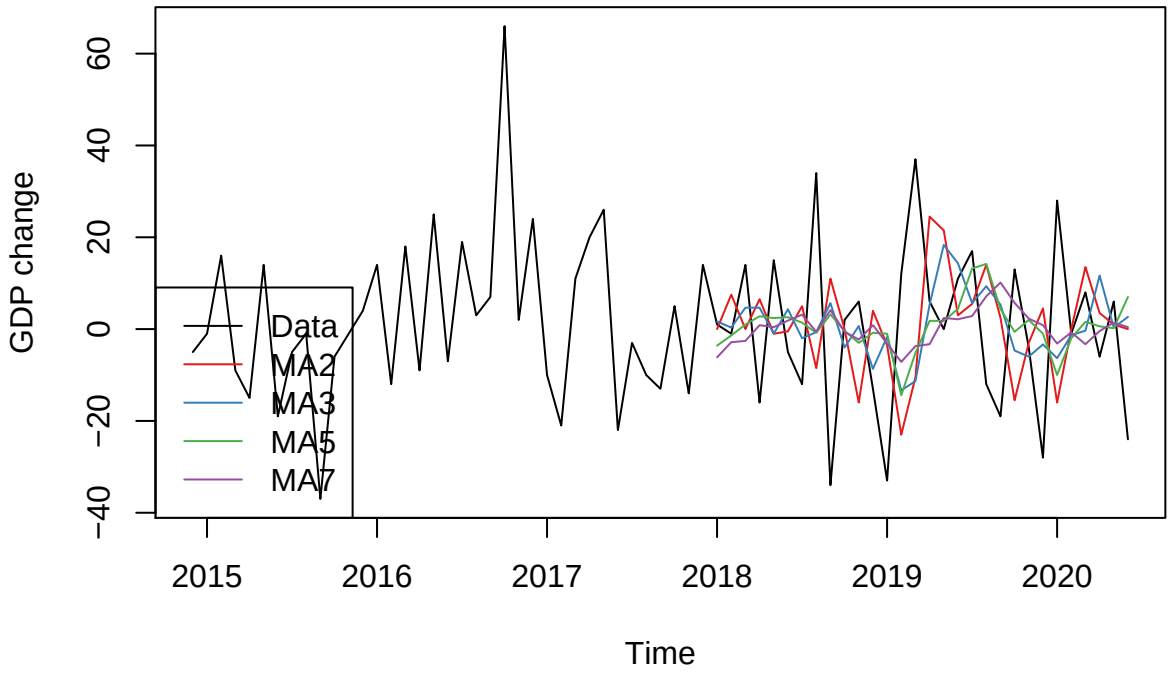
```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 2014                                              -5
## 2015  -1  16  -9 -15  14 -19  -5  -1 -37  -6  -1   4
## 2016  14 -12  18  -9  25  -7  19   3   7  66   2  24
## 2017 -10 -21  11  20  26 -22  -3 -10 -13   5 -14  14
```

```
test
```

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 2018   1  -1  14 -16  15  -5 -12  34 -34   2   6 -13
## 2019 -33  12  37   6   0  11  17 -12 -19  13  -4 -28
## 2020  28  -1   8  -6   6 -24
```

```
library(RColorBrewer)
n <- length(s)
m <- length(test)
```

```
maorder <- c(2,3,5,7)
frc <- array(NA,c(m,4),dimnames=list(time(test),paste0("MA",maorder)))

for (i in 1:m){
  fitsample <- s[1:(n-m+i-1)]
  fitsample <- ts(fitsample,frequency=frequency(s),start=start(s))
   for (j in 1:length(maorder)){
     frc[i,j] <- mean(tail(fitsample,maorder[j]))}
}

frc <- ts(frc,frequency=frequency(test),start=start(test))
cmp <- brewer.pal(4,"Set1")

plot(s,ylab="GDP change")
for (j in 1:length(maorder)){
  lines(frc[,j],col=cmp[j])
}
legend("bottomleft",c("Data",colnames(frc)),col=c("black",cmp),lty=1)
```




```
e <- matrix(rep(test,4),ncol=4) - frc
RMSE <- sqrt(apply(e^2,2,mean))
MAE <- apply(abs(e),2,mean)
E <- rbind(RMSE,MAE)
print(round(E,3))
```

```
##         MA2    MA3    MA5    MA7
## RMSE 23.782 20.436 20.049 19.315
## MAE  19.400 15.967 15.600 15.629
```

- Holt-Winter's seasonal method

  Additive & multiplicative model

```r
hw.add <- HoltWinters(train, seasonal = "additive")
hw.add
```

```
## Holt-Winters exponential smoothing with trend and additive seasonal component.
##
## Call:
## HoltWinters(x = train, seasonal = "additive")
##
## Smoothing parameters:
##  alpha: 0.34144
##  beta : 0
##  gamma: 0.5512733
##
## Coefficients:
##            [,1]
## a     2.7801620
## b     1.3189103
## s1   -0.9049750
## s2  -16.7209214
## s3   14.4794829
## s4   -1.2144506
## s5   16.1694492
## s6  -20.5502280
## s7   -0.2461459
## s8   -4.4599910
## s9  -17.1798307
## s10  11.0870101
## s11 -11.5143243
## s12   8.6541585
```

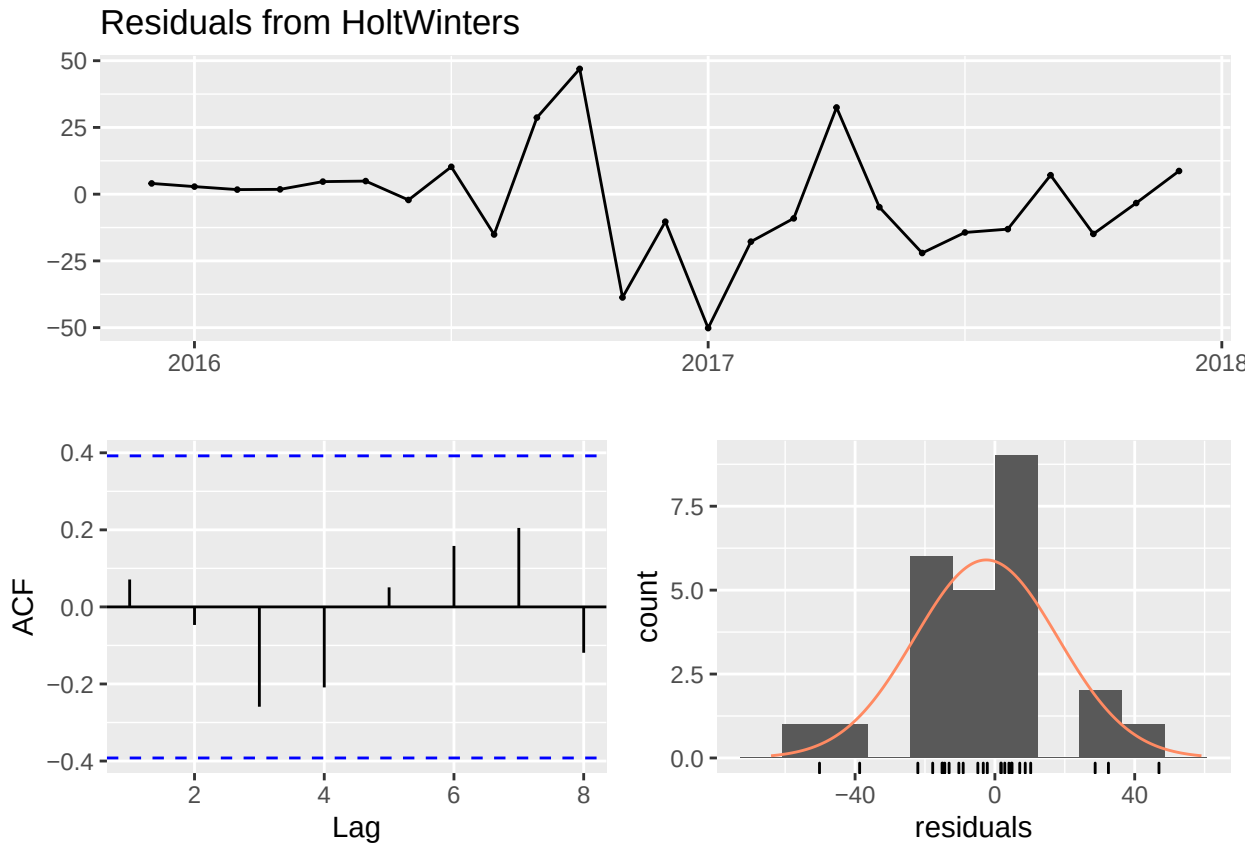
```r
# multiplicative
hw.mult <- HoltWinters(train, seasonal = "multiplicative")
hw.mult
```

```
## Holt-Winters exponential smoothing with trend and multiplicative seasonal component.
##
## Call:
## HoltWinters(x = train, seasonal = "multiplicative")
##
## Smoothing parameters:
##  alpha: 0.0414929
##  beta : 0.1184434
##  gamma: 0.01376287
##
## Coefficients:
```

```
##           [,1]
## a   14.2131773
## b    1.4113961
## s1  13.2772972
## s2   6.5406485
## s3  -2.7980289
## s4   0.8300969
## s5  -0.3778417
## s6  -1.3686303
## s7  -0.7290816
## s8  -0.1282666
## s9  -3.0967870
## s10 -0.8792444
## s11 -0.1591375
## s12  0.6787983
```

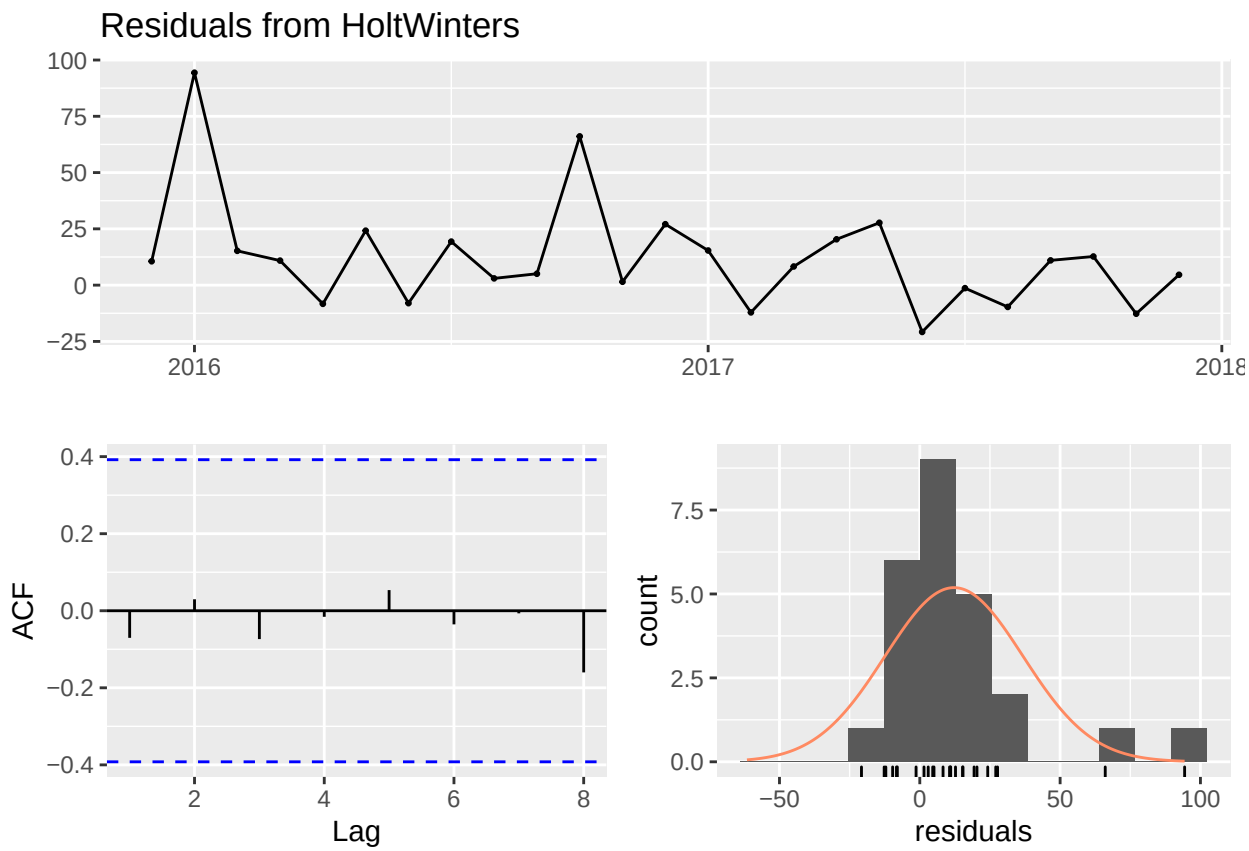
```
checkresiduals(hw.add)
```

```
## Warning in modeldf.default(object): Could not find appropriate degrees of
## freedom for this model.
```




```
checkresiduals(hw.mult)
```

```
## Warning in modeldf.default(object): Could not find appropriate degrees of
## freedom for this model.
```



Forecast with hold out sample is 2.5 years (h = 30) and compare prediction accuracy

```
for.add <- forecast(hw.add, h = 30)
for.mult <- forecast(hw.mult, h = 30)

accuracy(for.add, test)
```

```
##                      ME     RMSE      MAE       MPE     MAPE      MASE
## Training set  -2.468087 20.26032 14.80112 -29.6407 210.4702 0.7371077
## Test set     -21.371832 30.74107 24.90649     -Inf      Inf 1.2403633
##                    ACF1 Theil's U
## Training set 0.07116871        NA
## Test set     0.12716449         0
```

```
accuracy(for.mult, test)
```

```
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set  12.18028  26.99928  18.01328 98.17956 127.4167 0.8970757
## Test set     -41.72664 174.33110 102.71093      Inf      Inf 5.1150861
##                    ACF1 Theil's U
## Training set -0.07031602        NA
## Test set      0.28957267       NaN
```

ets as universal exponential smoothing model function

```
ets(train)
```

```
## ETS(A,N,N)
##
## Call:
##  ets(y = train)
##
##   Smoothing parameters:
##     alpha = 1e-04
##
##   Initial states:
##     l = 1.8366
##
##   sigma:  18.6396
##
##      AIC     AICc      BIC
## 354.0193 354.7466 358.8521
```

```
hw.mult2 <- ets(train, model = "ANN")
for.mult2 <- forecast(hw.mult2, h = 30)

accuracy(for.mult, test)
```

```
##                      ME      RMSE       MAE      MPE     MAPE      MASE
## Training set  12.18028  26.99928  18.01328 98.17956 127.4167 0.8970757
## Test set     -41.72664 174.33110 102.71093      Inf      Inf 5.1150861
##                      ACF1 Theil's U
## Training set -0.07031602        NA
## Test set      0.28957267       NaN
```

```
accuracy(for.mult2, test)
```

```
##                        ME     RMSE      MAE      MPE     MAPE      MASE
## Training set  0.003501979 18.12885 13.87926 114.1149 114.1149 0.6911984
## Test set     -1.769987491 17.77813 13.98911     -Inf      Inf 0.6966688
##                      ACF1 Theil's U
## Training set -0.06150155        NA
## Test set     -0.26829004         0
```

Run gamma and RMSE

```
gamma <- seq(0.0001, 0.95, 0.01)

RMSE <- NA

for(i in seq_along(gamma)) {
hw.opt <- ets(train, model = "ANN", gamma = gamma[i])
future <- forecast(hw.opt, h = 30)
RMSE[i] = accuracy(future, test)[2,2]
}
```

Graph gamma on forecast errors and with gamma minimum

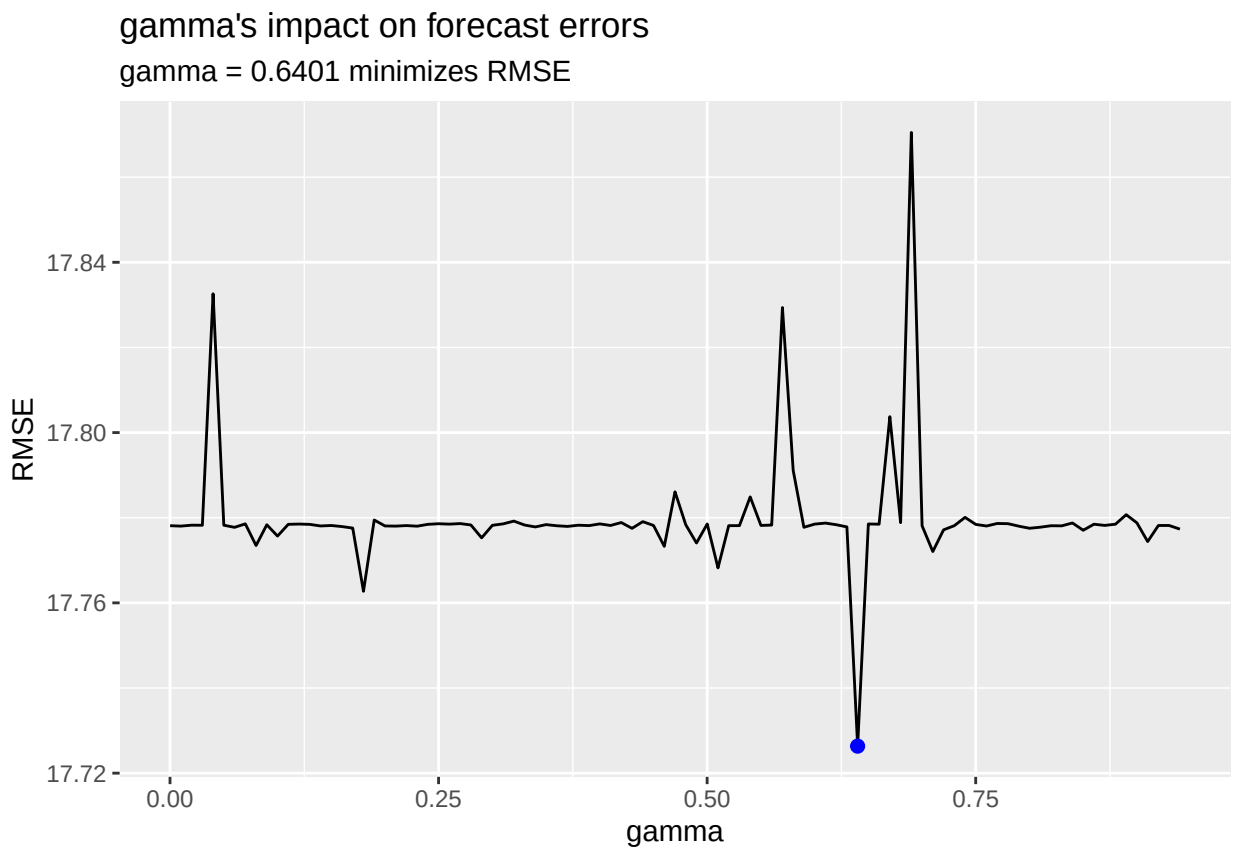
```
library(ggplot2)
error <- data_frame(gamma, RMSE)
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
minimum <- filter(error, RMSE == min(RMSE))
minimum
```

```
## # A tibble: 1 x 2
##   gamma  RMSE
##   <dbl> <dbl>
## 1 0.640  17.7
```

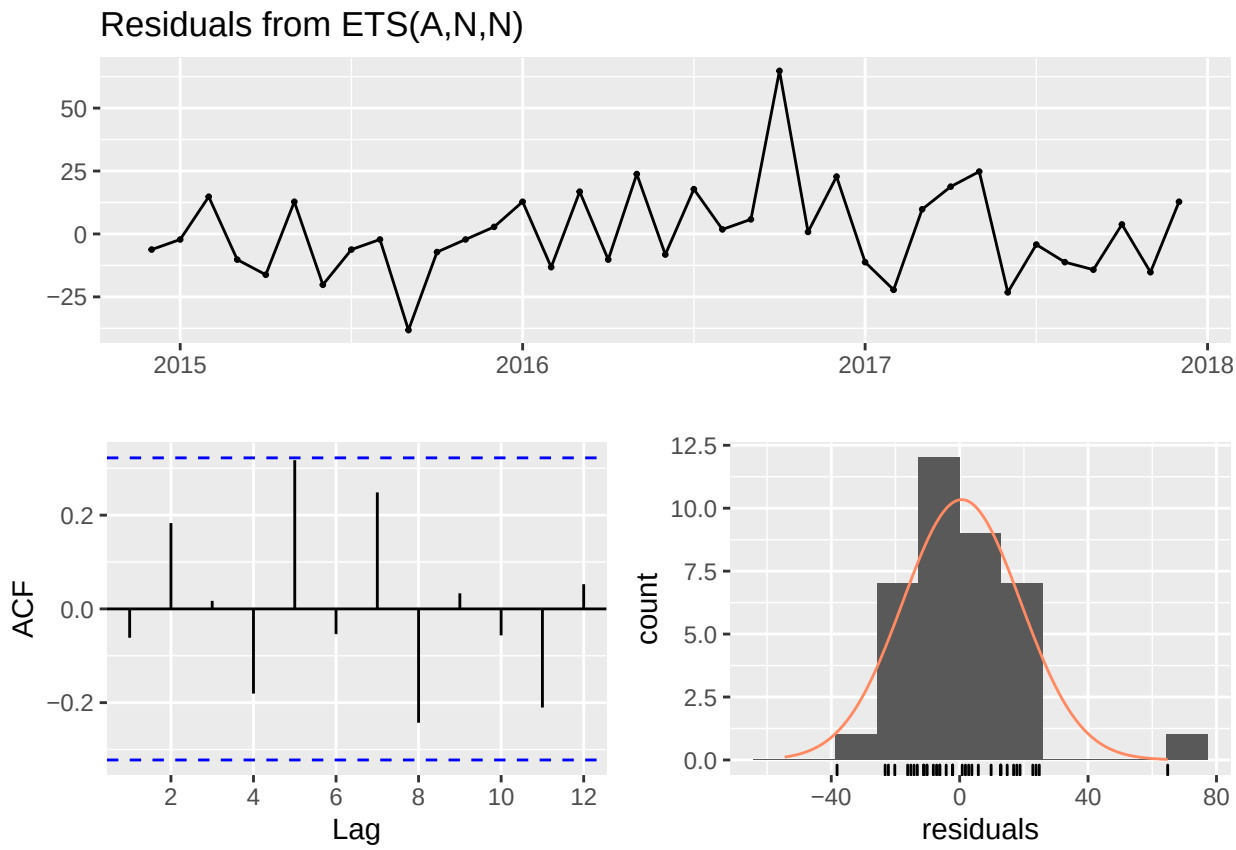
```
ggplot(error, aes(gamma, RMSE)) +
geom_line() +
geom_point(data = minimum, color = "blue", size = 2) +
ggtitle("gamma's impact on forecast errors",subtitle = "gamma = 0.6401 minimizes RMSE")
```



Forecast model is "A,N,N"

New exponential smothing model function with minimum gamma

```
hw.opt <- ets(train, model = "ANN", gamma = 0.6401)
checkresiduals(hw.opt)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(A,N,N)
## Q* = 10.618, df = 4, p-value = 0.03121
##
## Model df: 3.   Total lags used: 7
```

```
for.mult3 <- forecast(hw.opt, h = 30)
accuracy(for.mult2, test)
```

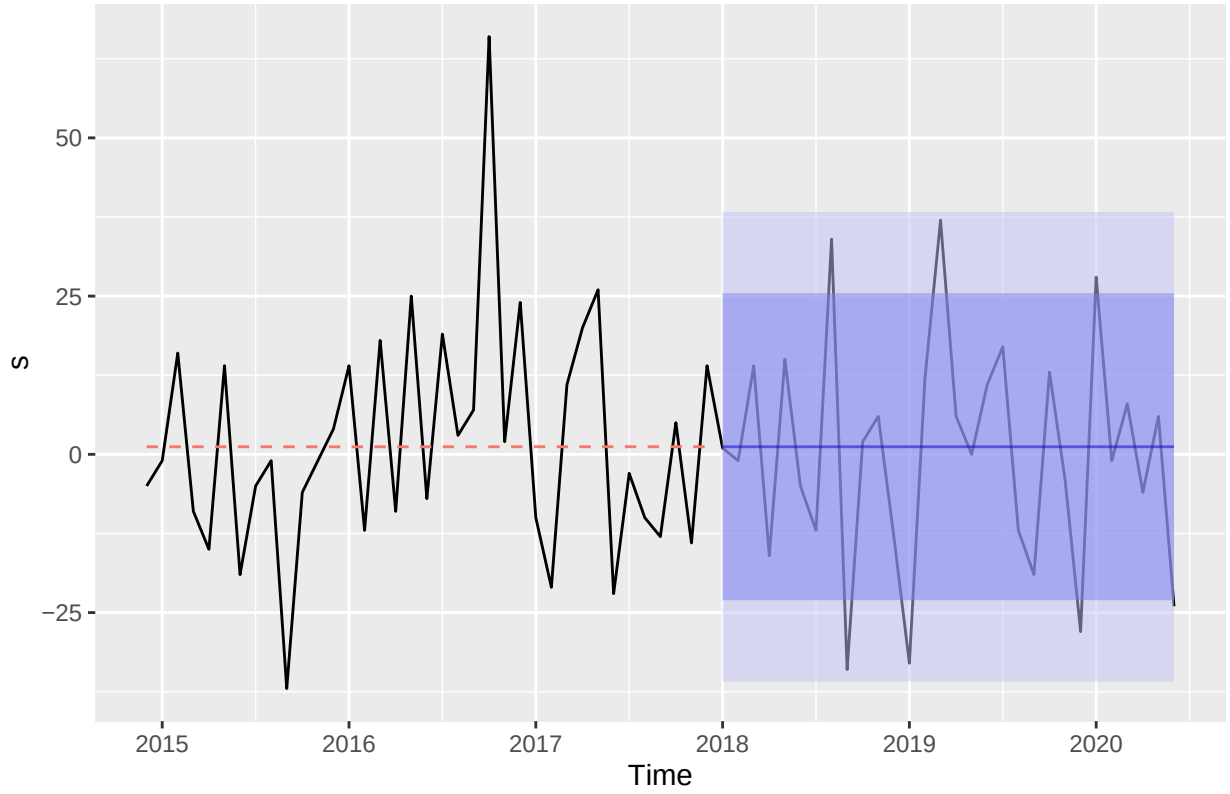
```
##                       ME     RMSE      MAE      MPE     MAPE      MASE
## Training set  0.003501979 18.12885 13.87926 114.1149 114.1149 0.6911984
## Test set     -1.769987491 17.77813 13.98911     -Inf      Inf 0.6966688
##                    ACF1 Theil's U
## Training set -0.06150155        NA
## Test set     -0.26829004         0
```

```
accuracy(for.mult3, test)
```

```
##                     ME     RMSE      MAE      MPE     MAPE      MASE
```

```
## Training set  0.6370097 18.13996 13.82776 109.2301 109.2301 0.6886333
## Test set     -1.1376837 17.72634 13.94696     -Inf      Inf 0.6945696
##                     ACF1 Theil's U
## Training set -0.06150913        NA
## Test set     -0.26829004         0
```

```
autoplot(s) +
  autolayer(fitted.values(for.mult3), linetype = "dashed", show.legend = FALSE) +
  autolayer(for.mult3, alpha = .50)
```



4. Conclusion

a. Compare all the models

- Use RMSE/MAE from Moving Average

```
e <- matrix(rep(test,4),ncol=4) - frc
RMSE <- sqrt(apply(e^2,2,mean))
MAE <- apply(abs(e),2,mean)
E <- rbind(RMSE,MAE)
print(round(E,3))
```

```
##         MA2    MA3    MA5    MA7
## RMSE 23.782 20.436 20.049 19.315
## MAE  19.400 15.967 15.600 15.629
```

- Use RMSE/MAE from Holt-Winter Multiplicative method

```
accuracy(for.mult3, test)
```

```
##                       ME     RMSE      MAE      MPE     MAPE      MASE
## Training set  0.6370097 18.13996 13.82776 109.2301 109.2301 0.6886333
## Test set     -1.1376837 17.72634 13.94696     -Inf      Inf 0.6945696
##                     ACF1 Theil's U
## Training set -0.06150913        NA
## Test set     -0.26829004         0
```

- Use AIC,BIC on tested data

```
ets(test)
```

```
## ETS(A,N,N)
##
## Call:
##  ets(y = test)
##
##   Smoothing parameters:
##     alpha = 1e-04
##
##   Initial states:
##     l = 0.0765
##
##   sigma:  18.3116
##
##      AIC     AICc      BIC
## 280.4182 281.3413 284.6218
```

b. Provide a suggestion of the best model.

- Holt-Winter model yields lowers RMSE & MAE than Moving Average. Holt-Winter RMSE and MAE for tested data is sequentially 17.72634 & 13.94696. Meanwhile, Moving Average RMSE(s) and MAE(s) for tested data are above 19.315 and 15600. RMSE is to measure how far off an actual value is from the mean. A good model should have better predictions than the naïve estimate of the mean for all predictions. Therefore, measure of variation (RMSE) should reduce the randomness better than the Standard Deviation. Thus, RMSE should be as smaller number as best. To conclude, in my opinion, Holt-Winter method is the best model

c. Conclude findings:

- X13 time series decomposition may not be the best model but it is helpful to have a view of how data appears yearly, seasonally, its autocorrelation function (ACF) and its trend
- MA is to smooth out the "noise" arrival delays over a specific of time by creating a constant updated of average arrival delays. Based on the graph above, I have created MA(2), MA(3), MA(5), MA(7) to reflect short term trend. The MAs forecast pretty well as they follow the pattern of original data set. But MA doubts me to consider the forecast result is high chance of accuracy
- Holt-Winter seasonal method to capture seasonality in the data set and it also provides RMSE and MAE. To me, it is like the combination of X13 time series decompostion and MA. Therefore, I choose Holt-Winter method as my forecast model