EAS 509: Statistical Data Mining II - Project

Sri Guna Kaushik Undru - srigunak   Shri Harsha Thirumala Adapala - sadapala

2023-11-26

Loading the dataset

```
complaints <- readr::read_csv("baggagecomplaints.csv",show_col_types = FALSE)
head(complaints)
```

```
## # A tibble: 6 x 8
##   Airline        Date    Month  Year Baggage Scheduled Cancelled Enplaned
##   <chr>          <chr>   <dbl> <dbl>   <dbl>     <dbl>     <dbl>    <dbl>
## 1 American Eagle 01/2004     1  2004   12502     38276      2481   992360
## 2 American Eagle 02/2004     2  2004    8977     35762       886  1060618
## 3 American Eagle 03/2004     3  2004   10289     39445      1346  1227469
## 4 American Eagle 04/2004     4  2004    8095     38982       755  1234451
## 5 American Eagle 05/2004     5  2004   10618     40422      2206  1267581
## 6 American Eagle 06/2004     6  2004   13684     39879      1580  1347303
```

```
complaints <- complaints %>%
  mutate(
    Date_new = paste(Year, Month, "01", sep = " "),
    Date_new = as.yearmon(Date_new, "%Y %m")
  ) %>%
  select(-c(Date, Month, Year)) %>%
  rename(Date = Date_new)
head(complaints)
```

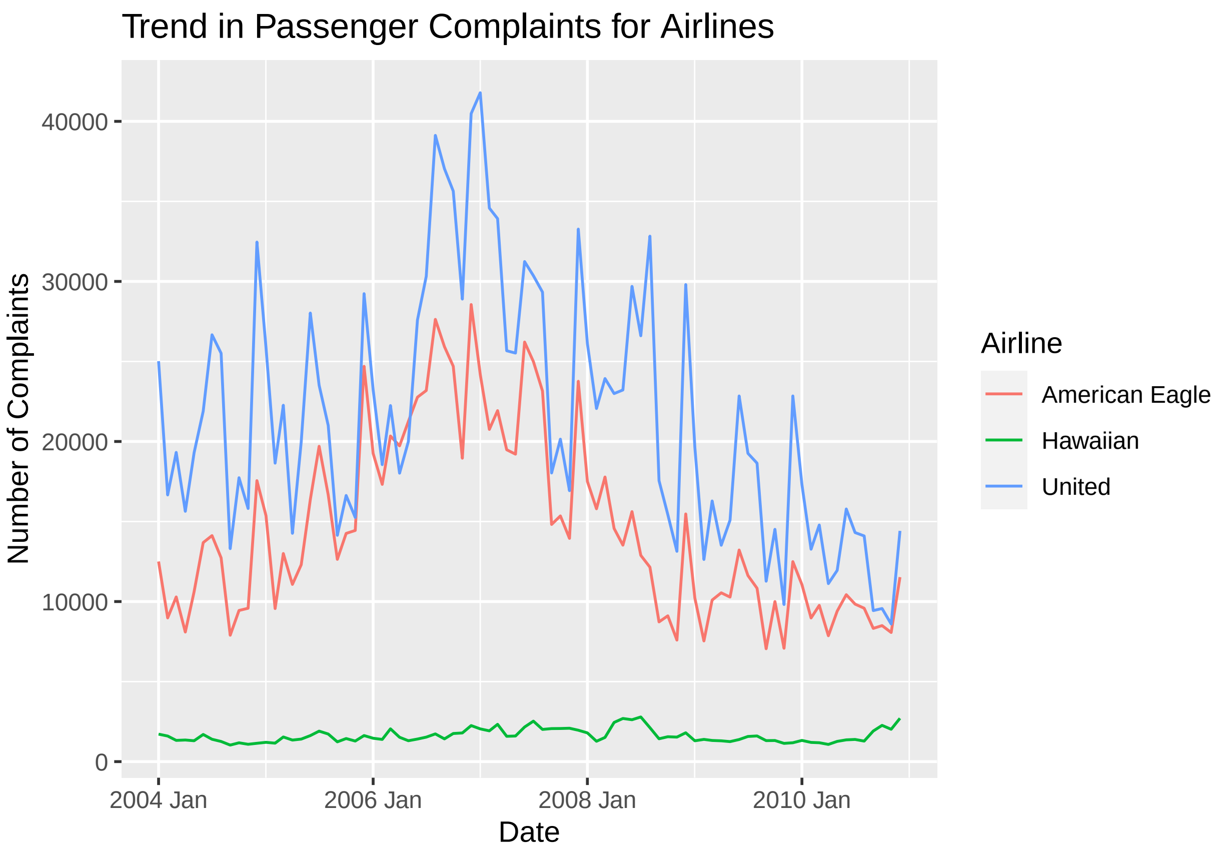
```
## # A tibble: 6 x 6
##   Airline        Baggage Scheduled Cancelled Enplaned Date
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl> <yearmon>
## 1 American Eagle   12502     38276      2481   992360 Jan 2004
## 2 American Eagle    8977     35762       886  1060618 Feb 2004
## 3 American Eagle   10289     39445      1346  1227469 Mar 2004
## 4 American Eagle    8095     38982       755  1234451 Apr 2004
## 5 American Eagle   10618     40422      2206  1267581 May 2004
## 6 American Eagle   13684     39879      1580  1347303 Jun 2004
```

```
complaints %>%
  mutate(Date=yearmonth(Date)) %>%
  tsibble(
    index = Date,
    key = Airline
    ) -> complaints
head(complaints)
```

```
## # A tsibble: 6 x 6 [1M]
## # Key:       Airline [1]
##   Airline        Baggage Scheduled Cancelled Enplaned     Date
```

```
##   <chr>             <dbl>      <dbl>      <dbl>     <dbl>    <mth>
## 1 American Eagle   12502      38276       2481    992360 2004 Jan
## 2 American Eagle    8977      35762        886   1060618 2004 Feb
## 3 American Eagle   10289      39445       1346   1227469 2004 Mar
## 4 American Eagle    8095      38982        755   1234451 2004 Apr
## 5 American Eagle   10618      40422       2206   1267581 2004 May
## 6 American Eagle   13684      39879       1580   1347303 2004 Jun
```

```r
complaints %>%
  autoplot(Baggage) +
  labs(x = "Date", y = "Number of Complaints", title = "Trend in Passenger Complaints for Airlines")
```



The data suggests that American and United Airlines experience more ups and downs in their baggage-related complaints. Additionally, since Hawaiian Airlines operates fewer flights, it's more meaningful to compare the number of complaints relative to the number of flights, rather than just looking at the total complaints. To get a clearer picture, we should identify the months with the most complaints over time, using a measure that takes into account the number of flights.

```r
complaints_summary <- complaints %>%
  group_by(Airline) %>%
  summarise(
    Scheduled = mean(Scheduled, na.rm = TRUE),
    Enplaned = mean(Enplaned, na.rm = TRUE),
    Count = n()
  )
complaints_summary
```

```
## # A tsibble: 252 x 5 [1M]
## # Key:       Airline [3]
```

```
##    Airline            Date Scheduled Enplaned Count
##    <chr>             <mth>     <dbl>    <dbl> <int>
##  1 American Eagle 2004 Jan     38276   992360     1
##  2 American Eagle 2004 Feb     35762  1060618     1
##  3 American Eagle 2004 Mar     39445  1227469     1
##  4 American Eagle 2004 Apr     38982  1234451     1
##  5 American Eagle 2004 May     40422  1267581     1
##  6 American Eagle 2004 Jun     39879  1347303     1
##  7 American Eagle 2004 Jul     41586  1396642     1
##  8 American Eagle 2004 Aug     42016  1339264     1
##  9 American Eagle 2004 Sep     40871  1292147     1
## 10 American Eagle 2004 Oct     42381  1393881     1
## # i 242 more rows
```

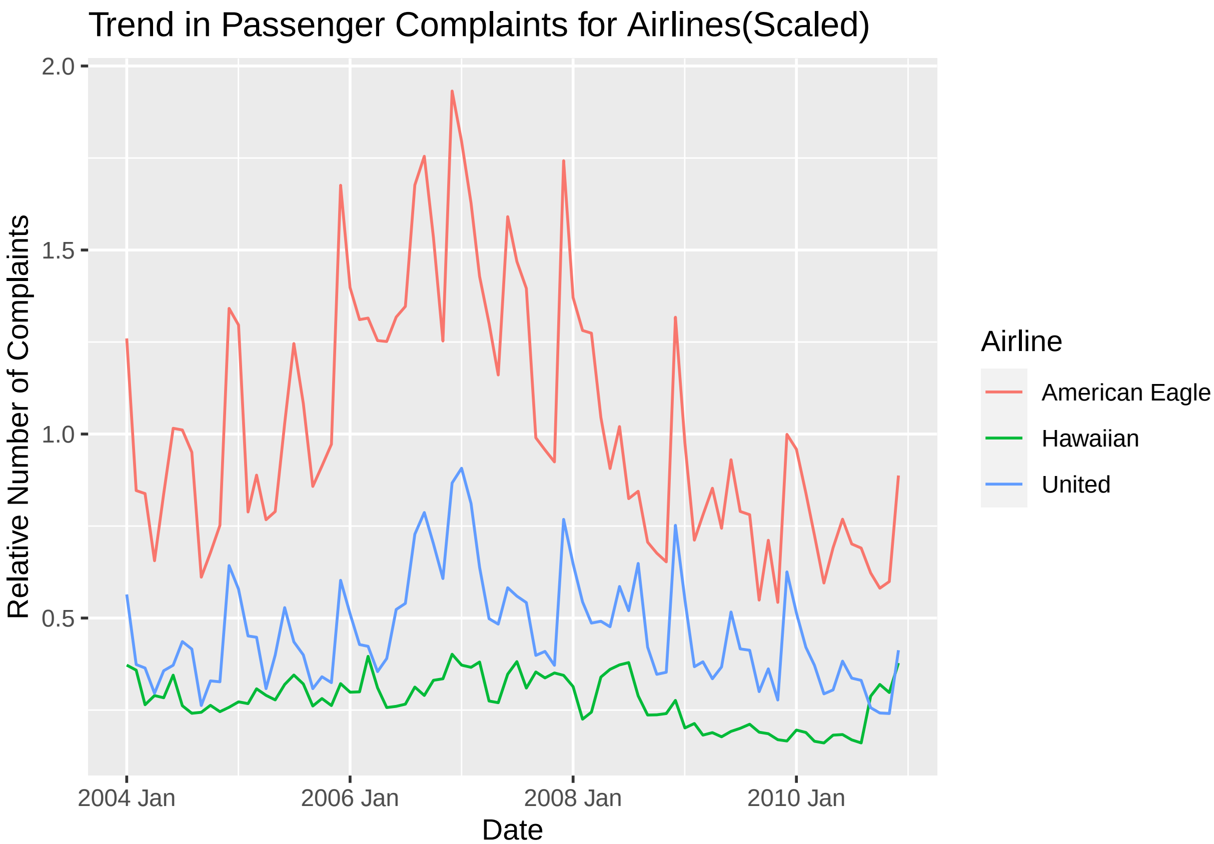
United Airlines is a lot larger than many other airlines. In the above summarizing data frame, one can see it has about three times as many flights and passengers as American Eagle, and about eight times more than Hawaiian Airlines. So, United Airlines ends up dealing with more bags simply because it serves a lot more passengers.

To cater with the company size/disparity among comparison. We will scale the complaints count with respect the Enplaned trips.

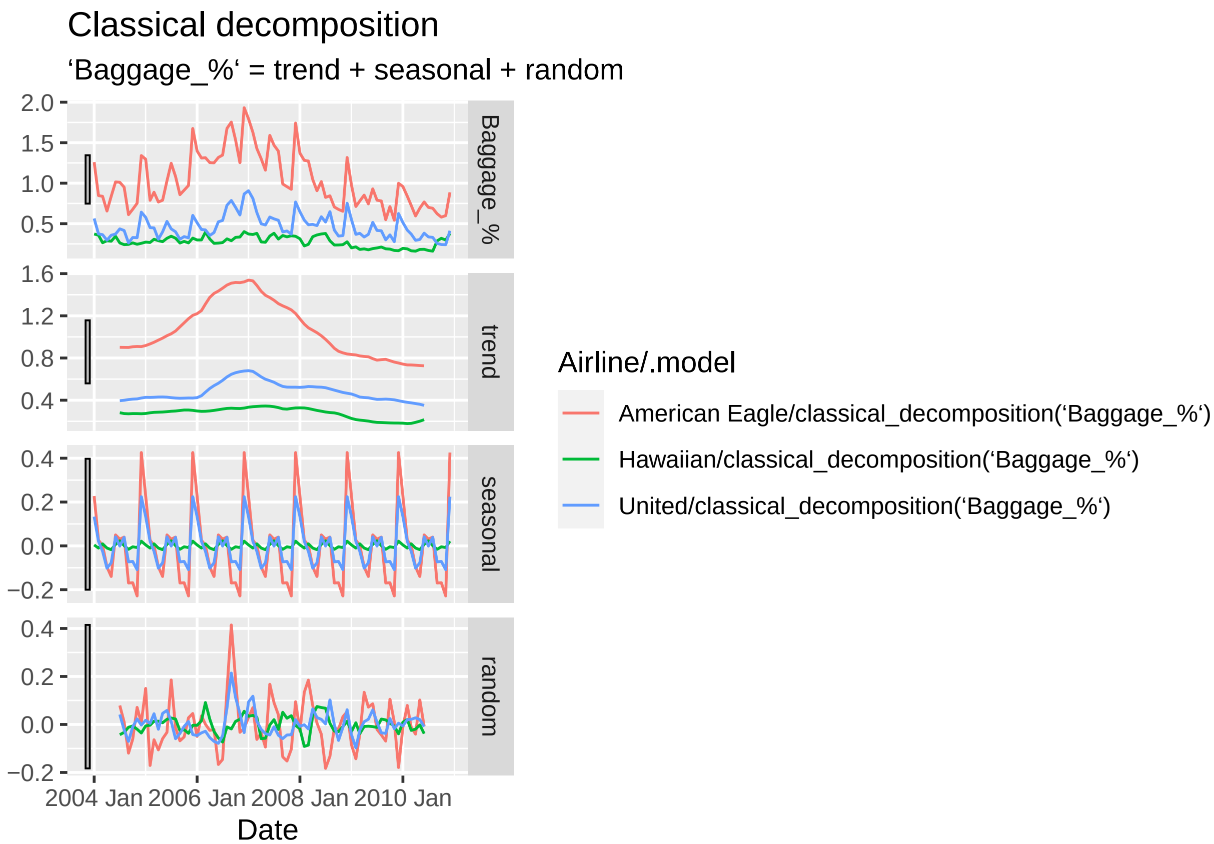
```r
complaints <- complaints %>%
  mutate(
    "Baggage_%" = (Baggage/Enplaned) * 100
  )
head(complaints)
```

```
## # A tsibble: 6 x 7 [1M]
## # Key:       Airline [1]
##   Airline        Baggage Scheduled Cancelled Enplaned     Date `Baggage_%`
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>    <mth>       <dbl>
## 1 American Eagle   12502     38276      2481   992360 2004 Jan       1.26
## 2 American Eagle    8977     35762       886  1060618 2004 Feb       0.846
## 3 American Eagle   10289     39445      1346  1227469 2004 Mar       0.838
## 4 American Eagle    8095     38982       755  1234451 2004 Apr       0.656
## 5 American Eagle   10618     40422      2206  1267581 2004 May       0.838
## 6 American Eagle   13684     39879      1580  1347303 2004 Jun       1.02
```

```r
complaints %>%
  autoplot(`Baggage_%`) +
  labs(x = "Date", y = "Relative Number of Complaints", title = "Trend in Passenger Complaints for Airli
```




```
complaints %>% model(
  classical_decomposition(`Baggage_%`)
  ) %>% components() %>%
  autoplot()
```



Upon doing classical additive decomposition, we can say that 1) All airlines show a generally stable or slightly increasing trend with United having the highest level and Hawaiian the lowest. 2) There is seasonality present and seasonal swings for American Eagle and United are quite similar and more pronounced than for Hawaiian.

Splitting the data frame into training and testing sets, where the testing set includes data from January 2010 to December 2010.

```
complaints_train <- complaints %>% filter(Date < yearmonth("2010 01"))
complaints_test <- complaints %>% filter(Date >= yearmonth("2010 01"))
```

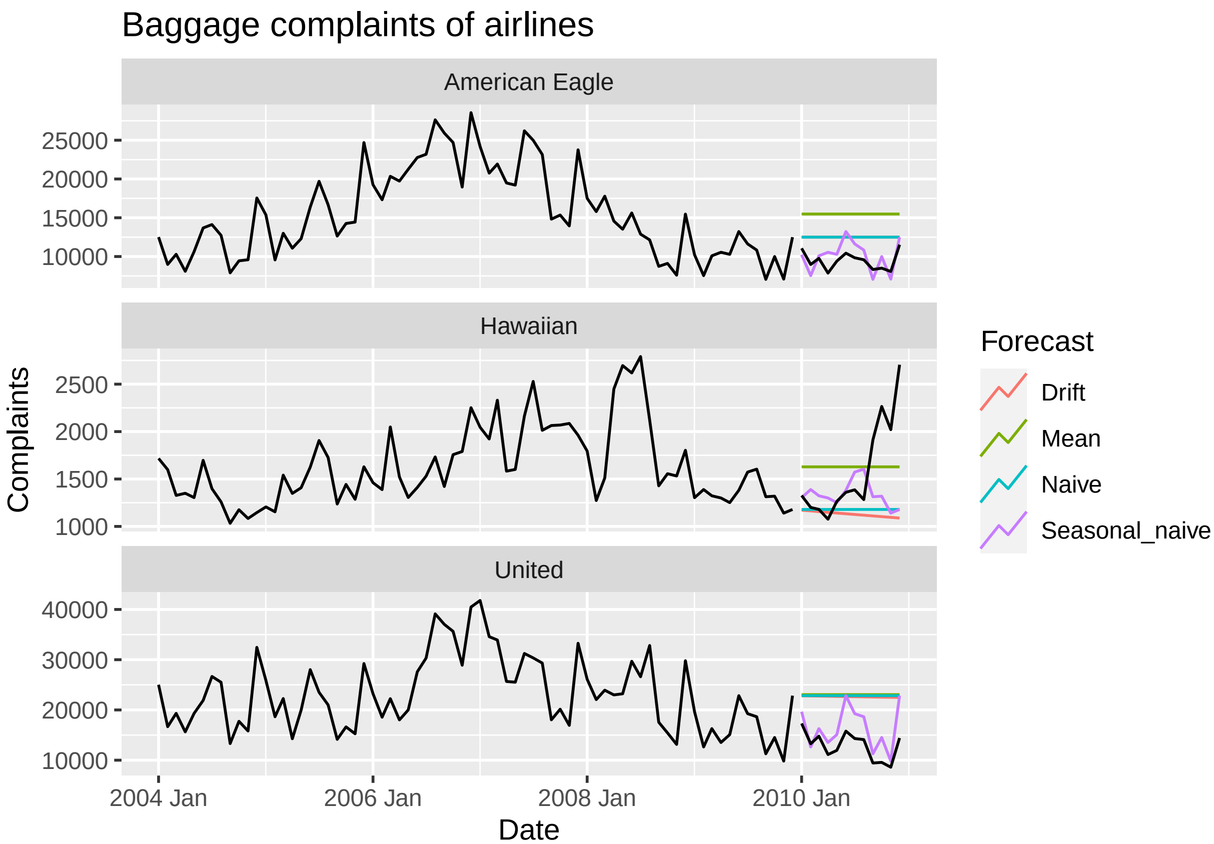
```
fit <- complaints_train %>%
  model(
    Seasonal_naive = SNAIVE(Baggage),
    Naive = NAIVE(Baggage),
    Drift = RW(Baggage ~ drift()),
    Mean = MEAN(Baggage)
  )

fc <- fit %>%
  forecast(h = "1 year")

z <- fc %>%
  hilo(level = 95) %>%
  pull(`95%`)
z$lower
```

```
##  [1]  -1780.53696  -4473.53696  -1918.53696  -1464.53696  -1732.53696
##  [6]   1208.46304   -386.53696  -1179.53696  -4959.53696  -2011.53696
```

```
##  [11]  -4930.53696   484.46304   4920.94186   1783.25005   -624.38556
##  [16]  -2654.11627  -4442.34493  -6059.02721  -7545.72000  -8929.49991
##  [21] -10229.17441 -11458.43712 -12627.62561 -13744.77113   4813.40436
##  [26]   1555.93033   -994.29973  -3186.19822  -5153.77387  -6965.18375
##  [31]  -8660.57758 -10265.89359 -11798.99390 -13272.77078 -14696.87119
##  [36] -16078.72321   4242.41377   4242.41377   4242.41377   4242.41377
##  [41]   4242.41377   4242.41377   4242.41377   4242.41377   4242.41377
##  [46]   4242.41377   4242.41377   4242.41377    233.29090    319.29090
##  [51]    253.29090    230.29090    181.29090    311.29090    503.29090
##  [56]    534.29090    243.29090    249.29090     70.29090    109.29090
##  [61]    571.76149    320.23507    127.23206    -35.47701   -178.82658
##  [66]   -308.42449   -427.60207   -538.52986   -642.71552   -741.25676
##  [71]   -834.98228   -924.53589    555.75451    287.13341     75.19008
##  [76]   -108.03696   -273.28701   -426.02210   -569.45980   -705.67876
##  [81]   -836.11043   -961.78799  -1083.48453  -1201.79533    817.92614
##  [86]    817.92614    817.92614    817.92614    817.92614    817.92614
##  [91]    817.92614    817.92614    817.92614    817.92614    817.92614
##  [96]    817.92614   2191.58668  -4834.41332  -1176.41332  -3944.41332
## [101]  -2378.41332   5382.58668   1797.58668   1184.58668  -6191.41332
## [106]  -2948.41332  -7647.41332   5383.58668   9464.86548   3922.63230
## [111]   -330.07280  -3915.26904  -7073.89033  -9929.50226 -12555.50844
## [116] -14999.73540 -17295.40355 -19466.70048 -21531.88584 -23505.14559
## [121]   9244.63948   3460.47364  -1074.47084  -4976.48595  -8482.27446
## [126] -11712.22375 -14737.25120 -17603.17211 -20341.53678 -22975.11708
## [131] -25520.95269 -27992.16401   8345.76374   8345.76374   8345.76374
## [136]   8345.76374   8345.76374   8345.76374   8345.76374   8345.76374
## [141]   8345.76374   8345.76374   8345.76374   8345.76374
```

```r
fc %>% autoplot(complaints_train, level = NULL) +
  labs(
    title = "Baggage complaints of airlines",
    y = "Complaints"
  ) +
  autolayer(complaints_test, color = "black") +
  guides(colour = guide_legend(title = "Forecast"))
```

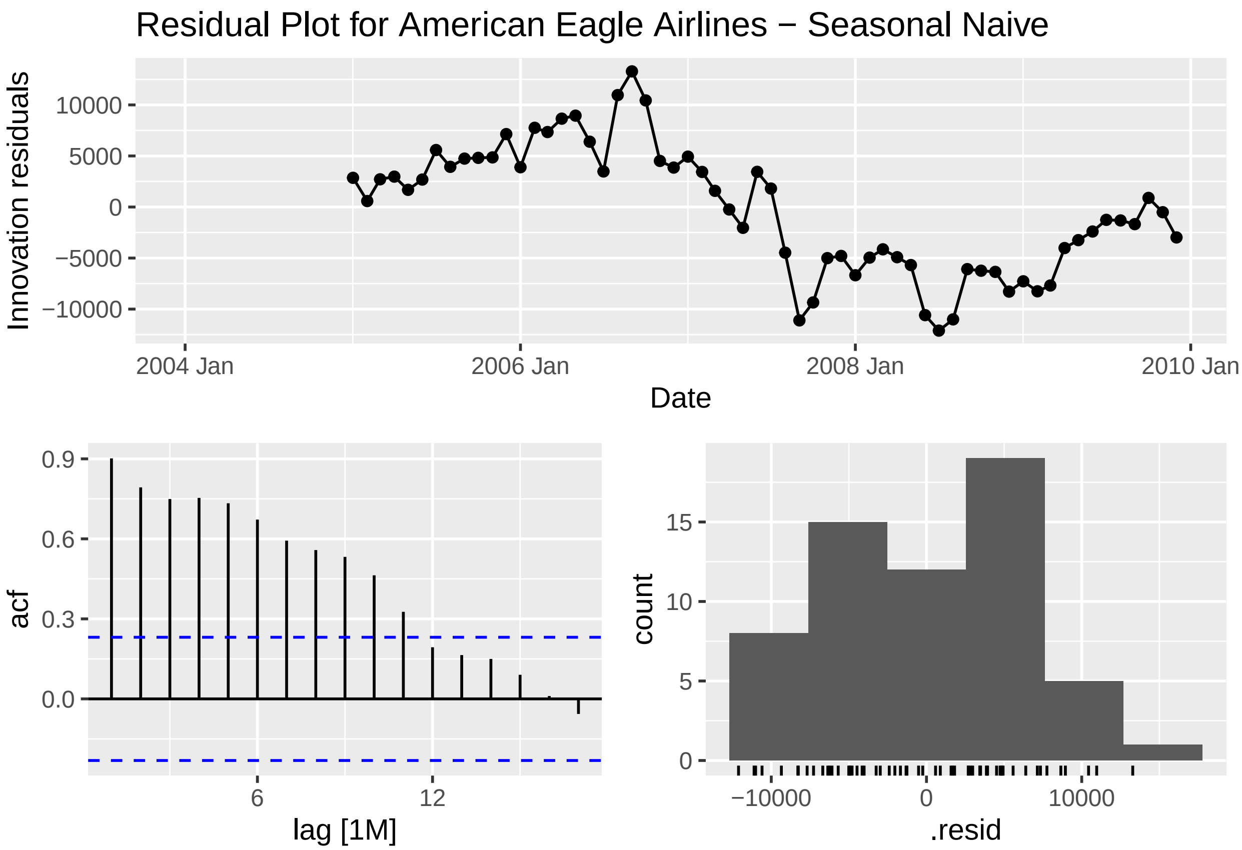
```
## Plot variable not specified, automatically selected `.vars = Baggage`
```



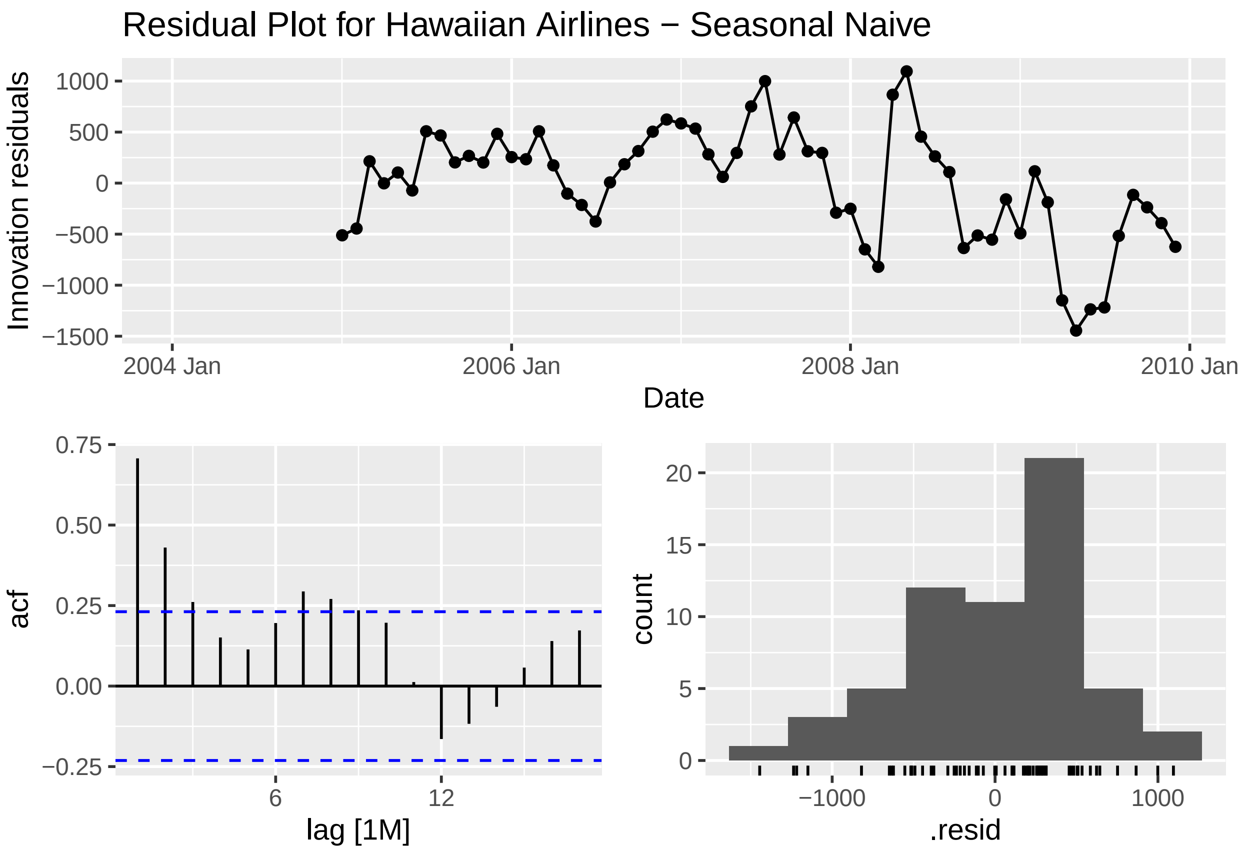
Based on visual inspection, it appears that the Seasonal Naive model most closely follows the pattern observed in the actual test data among the four basic models considered. However, this assessment is solely based on visual analysis.

```
sn_fit <- fit %>%
  select(Seasonal_naive)
num_models <- nrow(sn_fit)
```

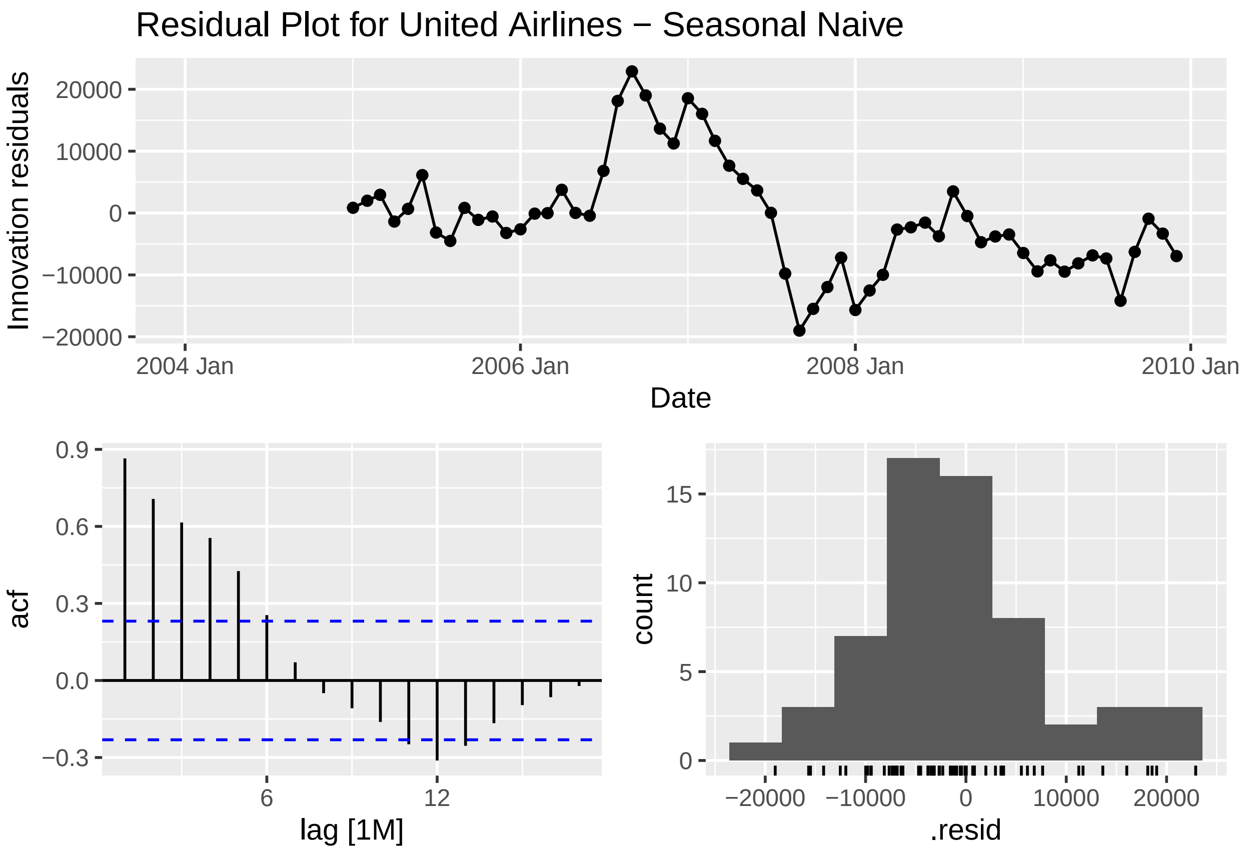
```
suppressWarnings({
  model <- sn_fit[1, ]
plot <- gg_tsresiduals(model) +
        labs(title = paste("Residual Plot for", model$Airline, "Airlines - Seasonal Naive"))

print(plot)
})
```




```
suppressWarnings({
  model <- sn_fit[2, ]
  plot <- gg_tsresiduals(model) +
          labs(title = paste("Residual Plot for", model$Airline, "Airlines - Seasonal Naive"))

  print(plot)
})
```




```r
suppressWarnings({
  model <- sn_fit[3, ]
  plot <- gg_tsresiduals(model) +
          labs(title = paste("Residual Plot for", model$Airline, "Airlines - Seasonal Naive"))

  print(plot)
})
```

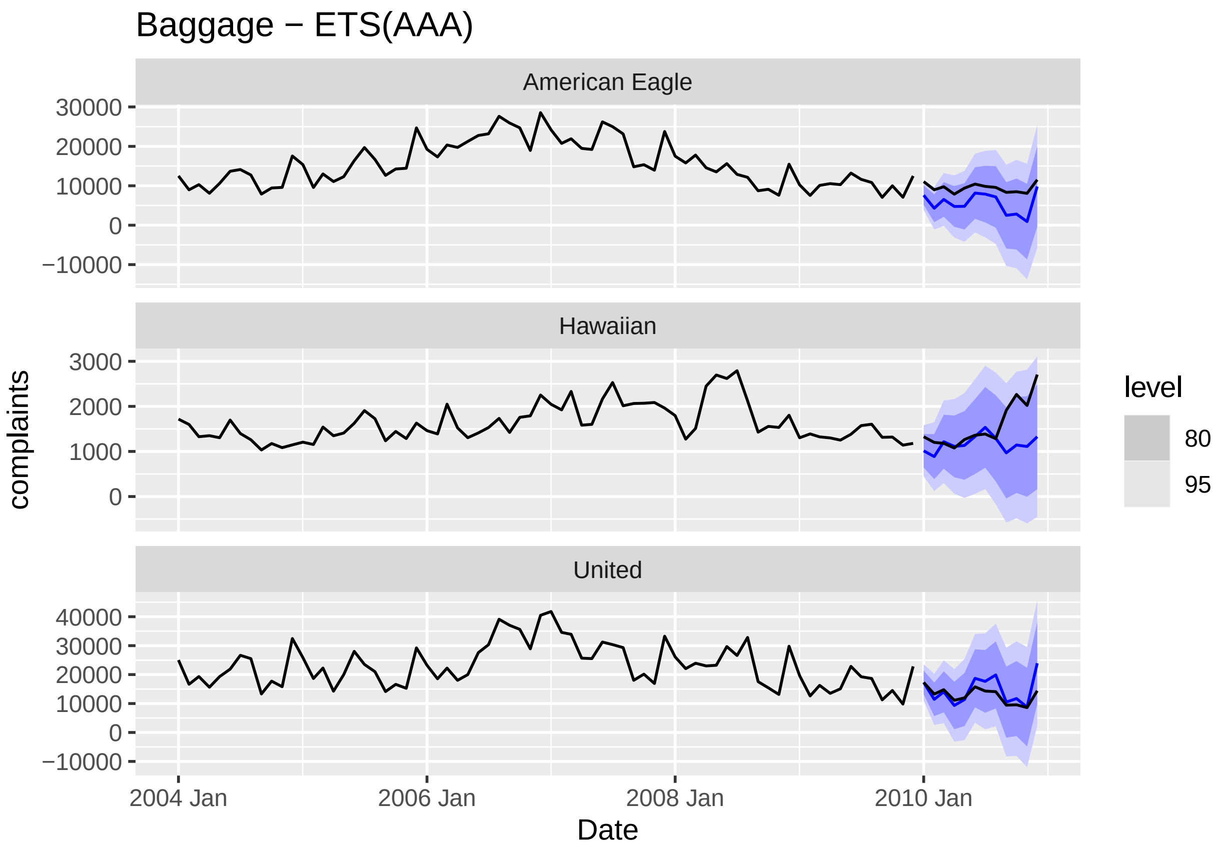


The presence of any systematic structure in the residuals plots or significant autocorrelation at various lags would suggest that the model can be further improved.

Now, lets try modelling with ETS(AAA) model

```
ets_fit <- complaints_train %>%
  model(additive = ETS(Baggage ~ error("A") + trend("A") + season("A")))

ets_fc <- ets_fit %>%
  forecast(h = "1 year")

ets_fc %>% autoplot(complaints_train) +
  labs(
    title = "Baggage - ETS(AAA)",
    y = "complaints"
  ) +
  autolayer(complaints_test, color = "black") +
  guides(colour = guide_legend(title = "Forecast"))
```

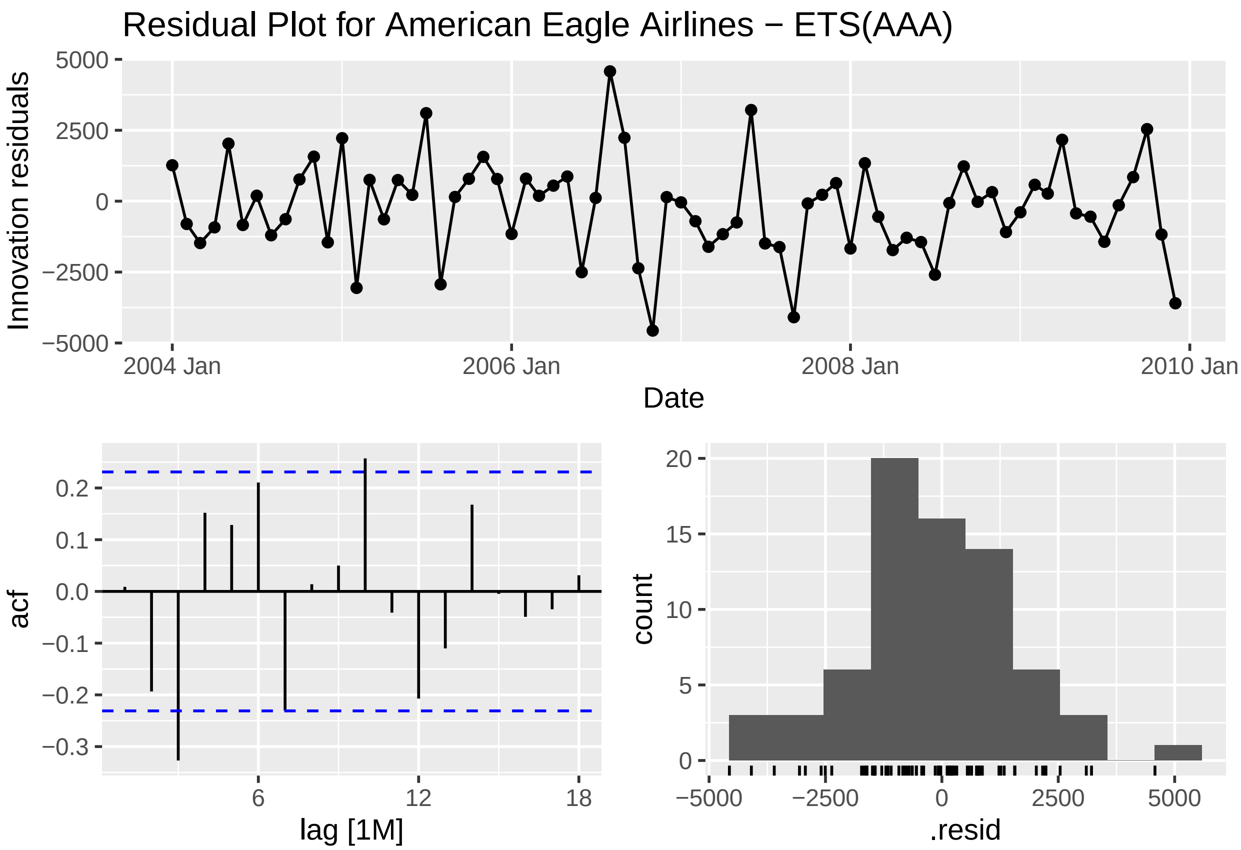
```
## Plot variable not specified, automatically selected `.vars = Baggage`
```



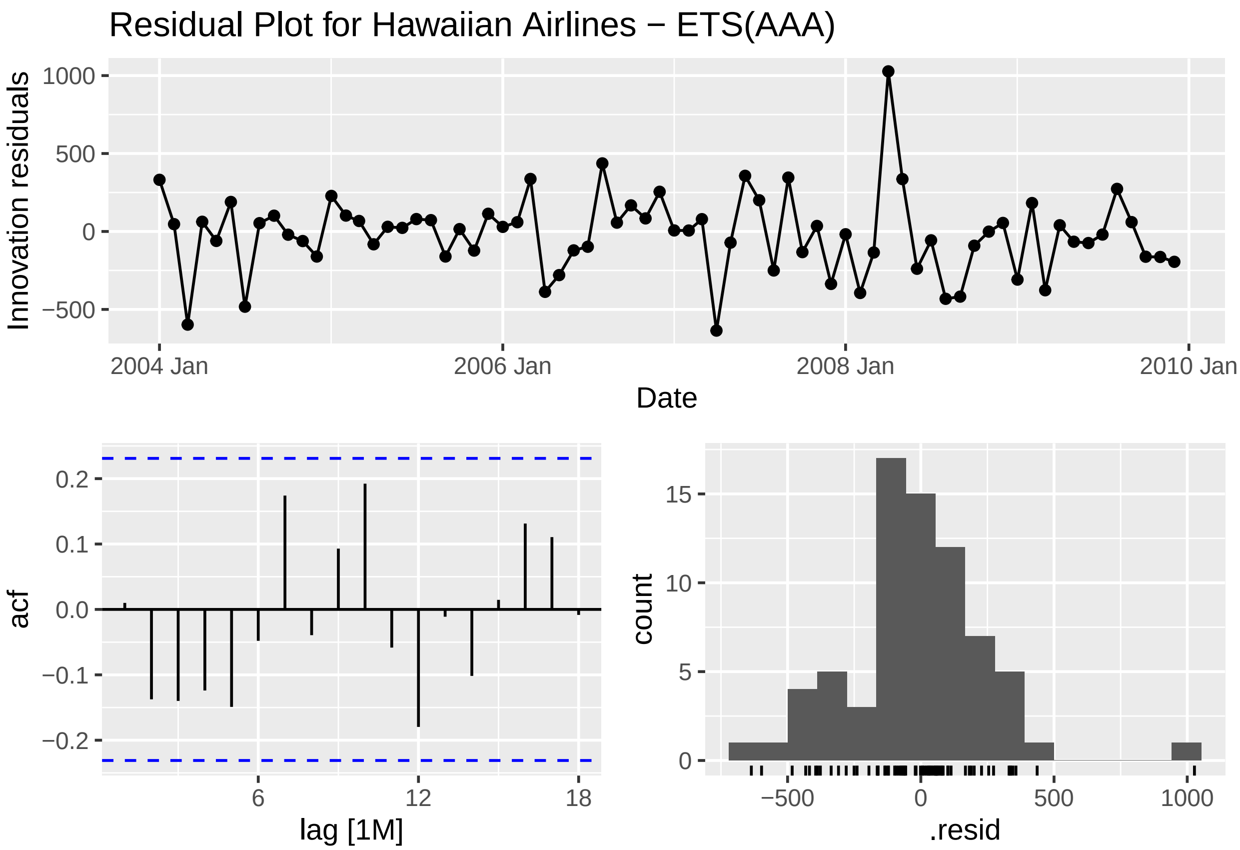
Based on visual inspection, it appears that the ETS(AAA) model closely follows the pattern observed in the actual test data compared to the four basic models considered before. However, this assessment is solely based on visual analysis.

```
# Extract the residuals
num_models <- nrow(ets_fit)
```

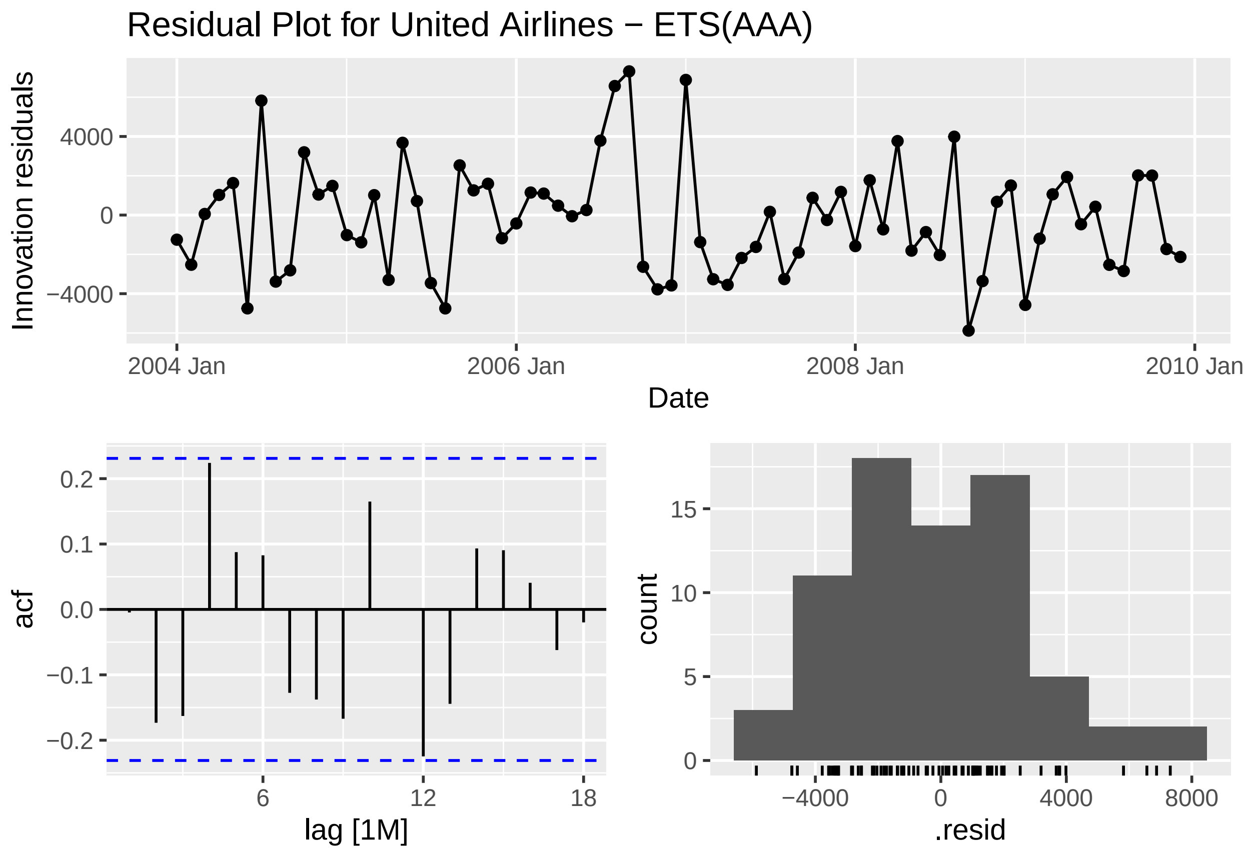
```
model <- ets_fit[1, ]

    # Generate the residual plot
    plot <- gg_tsresiduals(model) +
            labs(title = paste("Residual Plot for", model$Airline, "Airlines - ETS(AAA)"))
    print(plot)
```




```r
model <- ets_fit[2, ]

    # Generate the residual plot
    plot <- gg_tsresiduals(model) +
            labs(title = paste("Residual Plot for", model$Airline, "Airlines - ETS(AAA)"))
    print(plot)
```




```
model <- ets_fit[3, ]

    # Generate the residual plot
    plot <- gg_tsresiduals(model) +
            labs(title = paste("Residual Plot for", model$Airline, "Airlines - ETS(AAA)"))
    print(plot)
```



The above residual plots suggests that while the ETS(AAA) model has captured much of the data's behavior, but there are instances where it fails to predict accurately, as indicated by the spikes in the time series plot and the few significant autocorrelations in the ACF plot.The histogram's shape also suggests that the residuals may not be normally distributed.

Now lets try with the ARIMA model.

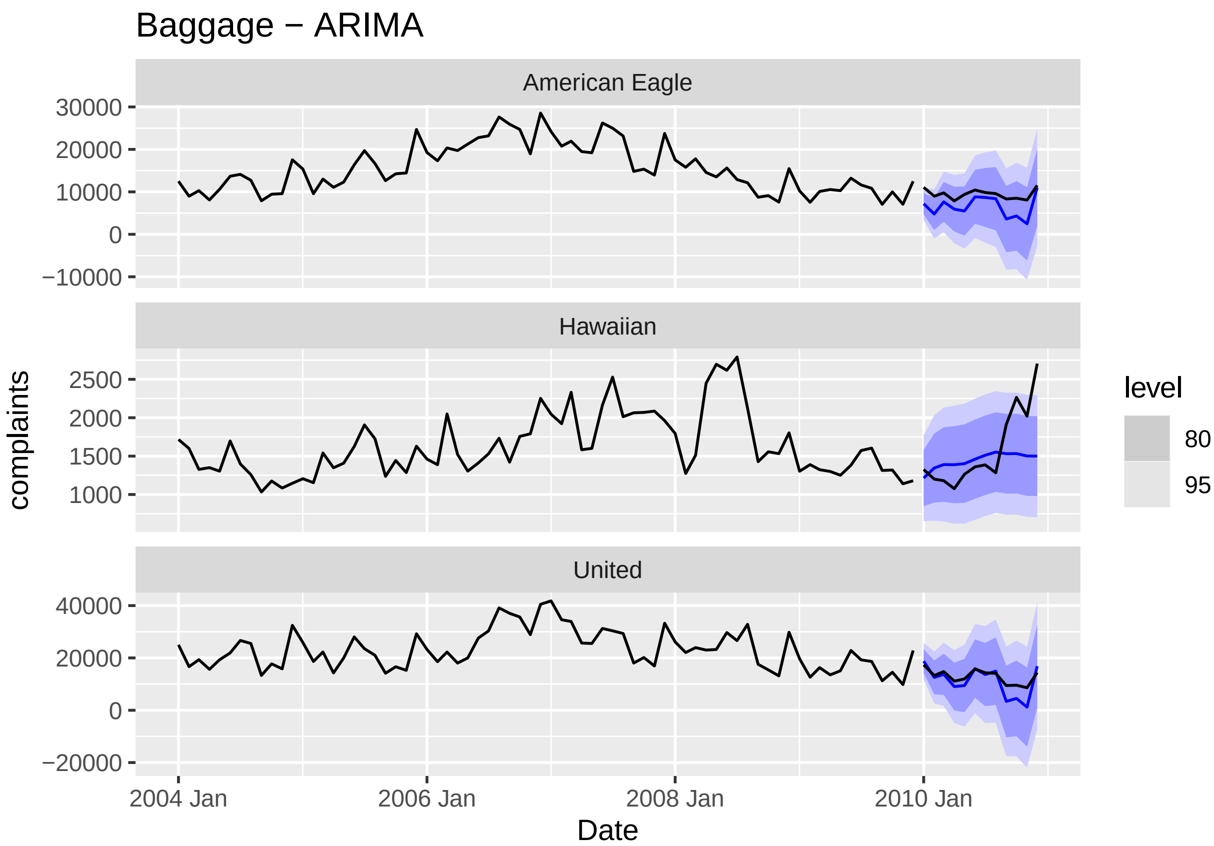
```
arima_fit <- complaints_train %>%
  model(arima = ARIMA(Baggage))

# Forecast for a 1-year horizon
arima_fc <- arima_fit %>%
  forecast(h = "1 year")

arima_fit %>%
  select(arima) %>%
  report()
```

```
## # A tibble: 3 x 9
##   Airline         .model    sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##   <chr>           <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
## 1 American Eagle arima   3780072.    -533. 1078. 1080. 1091. <cpl [2]>  <cpl>
## 2 Hawaiian       arima     81430.    -508. 1025. 1025. 1034. <cpl [1]>  <cpl>
## 3 United         arima  12677092.    -570. 1146. 1146. 1152. <cpl [24]> <cpl [0]>
```

```
# Plot the forecast and the original training data
suppressWarnings({
  arima_fc %>% autoplot(complaints_train) +
    labs(
```

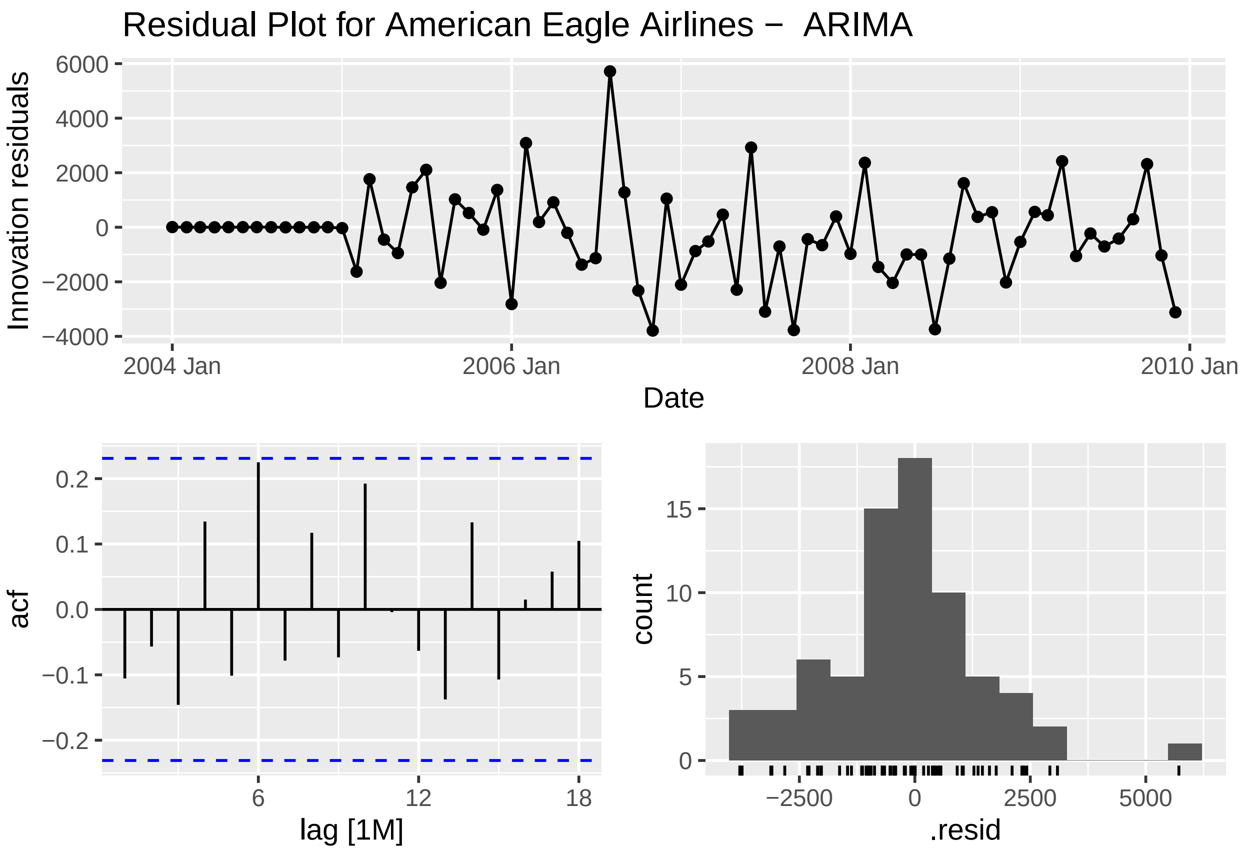
```r
      title = "Baggage - ARIMA",
      y = "complaints"
    ) +
    autolayer(complaints_test, color = "black") +
    guides(colour = guide_legend(title = "Forecast"))
})
```

```
## Plot variable not specified, automatically selected `.vars = Baggage`
```

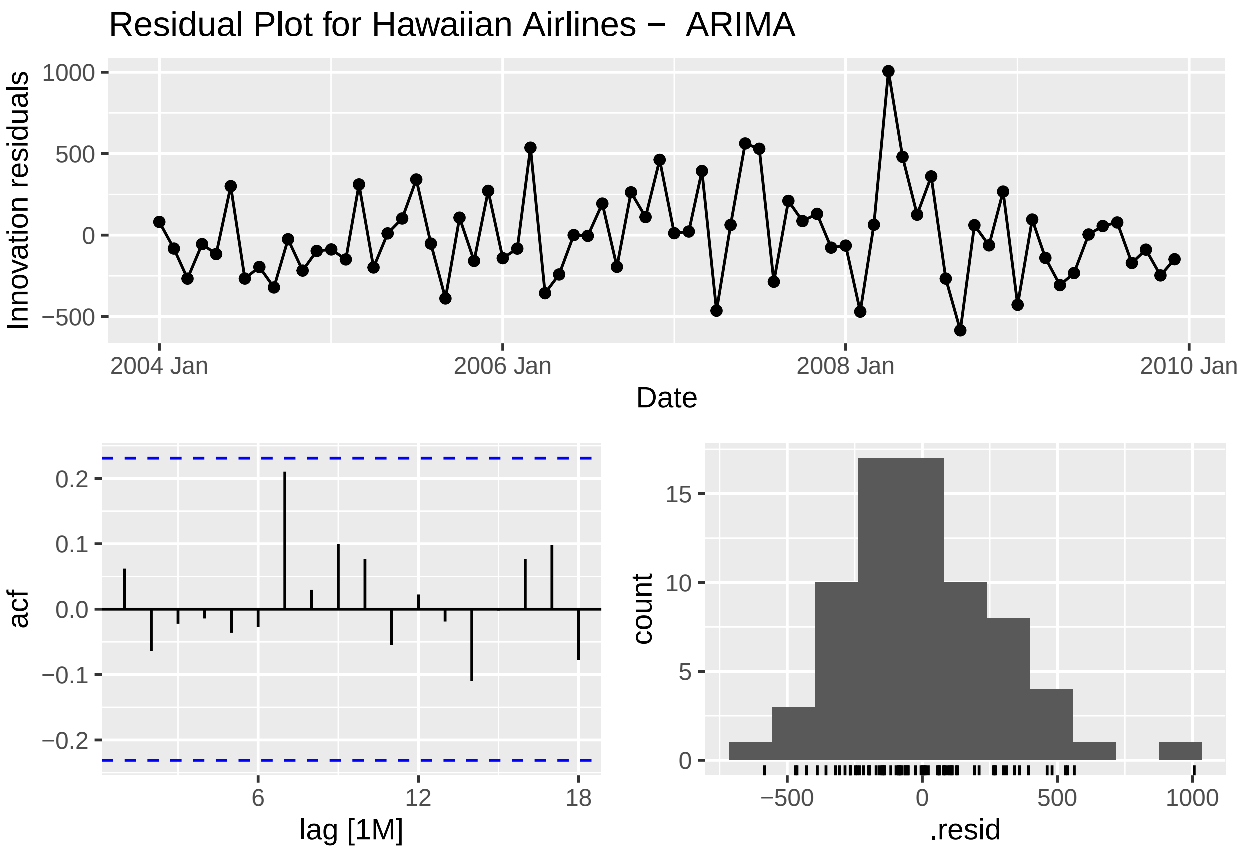



```r
# Extract the residuals
num_models <- nrow(arima_fit)
```

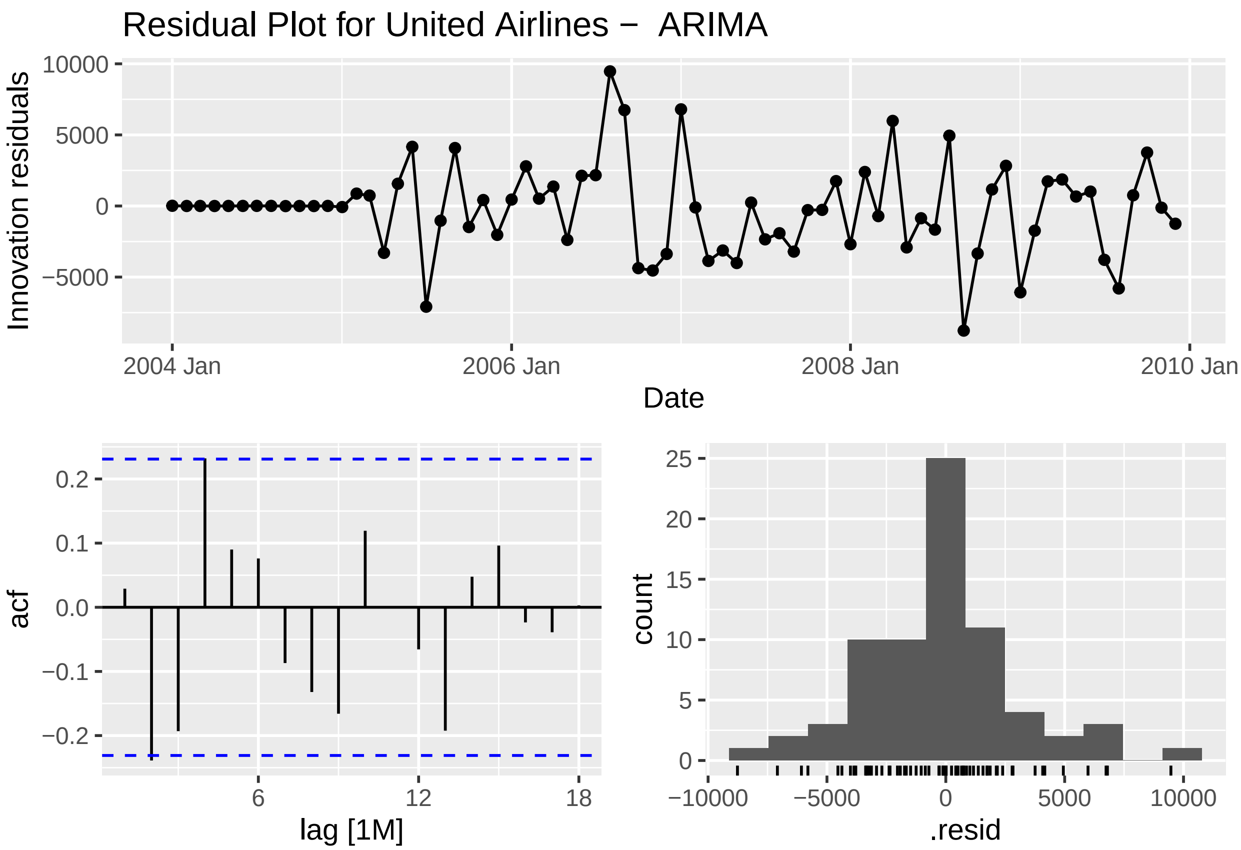
```r
model <- arima_fit[1, ]

    # Generate the residual plot
plot <- gg_tsresiduals(model) +
        labs(title = paste("Residual Plot for", model$Airline, "Airlines -  ARIMA"))
print(plot)
```




```r
model <- arima_fit[2, ]

    # Generate the residual plot
plot <- gg_tsresiduals(model) +
        labs(title = paste("Residual Plot for", model$Airline, "Airlines -  ARIMA"))
print(plot)
```




```
model <- arima_fit[3, ]

    # Generate the residual plot
plot <- gg_tsresiduals(model) +
        labs(title = paste("Residual Plot for", model$Airline, "Airlines -  ARIMA"))
print(plot)
```



The above plots depicts a residual plots for all the Airlines using an ARIMA model. The top plot shows the residuals (errors) over time, which do not display any obvious patterns or trends, suggesting the model's errors are random, which is a good sign in time series forecasting. The bottom left plot is the autocorrelation function (ACF) of residuals, showing that most lags are within the confidence interval, indicating little to no autocorrelation. The bottom right is a histogram of the residuals, which seems fairly normally distributed around zero. Overall, these diagnostics suggest the ARIMA model fits the data reasonably well, with no apparent autocorrelation issues and residuals that are approximately normally distributed. ## lets compare the BIC and AIC errors for both ETS(AAA) & SARIMA models

```r
report(ets_fit)
```

```
## # A tibble: 3 x 10
##   Airline        .model     sigma2 log_lik   AIC  AICc   BIC    MSE   AMSE   MAE
##   <chr>          <chr>       <dbl>   <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl>
## 1 American Eagle additive   3.65e6    -689. 1412. 1423. 1450. 2.84e6 5.56e6 1295.
## 2 Hawaiian       additive   8.43e4    -553. 1140. 1152. 1179. 6.55e4 1.14e5  183.
## 3 United         additive   1.02e7    -726. 1486. 1497. 1524. 7.94e6 1.51e7 2269.
```

```r
report(arima_fit)
```

```
## # A tibble: 3 x 9
##   Airline        .model    sigma2 log_lik   AIC  AICc   BIC ar_roots   ma_roots
##   <chr>          <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl> <list>     <list>
## 1 American Eagle arima   3780072.   -533. 1078. 1080. 1091. <cpl [2]>  <cpl>
## 2 Hawaiian       arima     81430.   -508. 1025. 1025. 1034. <cpl [1]>  <cpl>
## 3 United         arima  12677092.   -570. 1146. 1146. 1152. <cpl [24]> <cpl [0]>
```

You can see that, there is huge difference in the AIC and BIC values between the two models indicating better performance of SARIMA model

Lets see the p-values for ETS(AAA) and SARIMA model

```
augment(arima_fit %>% select(arima)) %>%
  features(.resid, ljung_box, lag = 24, dof = 16)
```

```
## # A tibble: 3 x 4
##   Airline        .model lb_stat lb_pvalue
##   <chr>          <chr>    <dbl>     <dbl>
## 1 American Eagle arima     28.6  0.000368
## 2 Hawaiian       arima     13.9  0.0835
## 3 United         arima     28.9  0.000330
```

```
augment(ets_fit %>% select(additive)) %>%
  features(.resid, ljung_box, lag = 24, dof = 16)
```

```
## # A tibble: 3 x 4
##   Airline        .model   lb_stat  lb_pvalue
##   <chr>          <chr>      <dbl>      <dbl>
## 1 American Eagle additive    38.0 0.00000749
## 2 Hawaiian       additive    25.6 0.00121
## 3 United         additive    32.1 0.0000887
```

The data indicates that p_values for SARIMA model are significant compared to the ETS(AAA) model.