The exam contains FIVE sections. ALL sections must be completed. The exam is worth 100 marks in total.

Below are the State Space equations for each of the models in the ETS framework.

ADDITIVE ERROR MODELS

Trend	Seasonal N	Seasonal A	Seasonal M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathbf{A_d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$

MULTIPLICATIVE ERROR MODELS

Trend	Seasonal N	Seasonal A	Seasonal M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
$\mathbf{A_d}$	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

SECTION A

Write about a quarter of a page each on any **four** of the following topics.

1. The AIC is better than the MSE for selecting a forecasting model.

   5 marks

2. The Ljung-Box test is useful for selecting a good forecasting model.

   5 marks

3. The MAPE is better than the RMSE for measuring forecast accuracy because it is easier to explain.

   5 marks

4. Dynamic regression models are only useful for forecasting if the predictors are deterministic or easy to forecast, and there is no serial correlation in the residuals.

   5 marks

5. ETS and ARIMA models are largely interchangeable, so it does not matter which one we use.

   5 marks

6. STL is not useful for forecasting because it only provides a decomposition of the data.

   5 marks

Total: 20 marks

SECTION B

Figures 1–3 relate to the daily use of public transport in Canberra, from July 2019 – March 2024. The variable plotted is the total number of passenger boardings each day on all forms of public transport except for school buses.

```
act_pt |>
  autoplot(Boardings) +
  labs(
    title = "Daily passengers using public transport in Canberra",
    y = "Boardings (thousands)"
  )
```

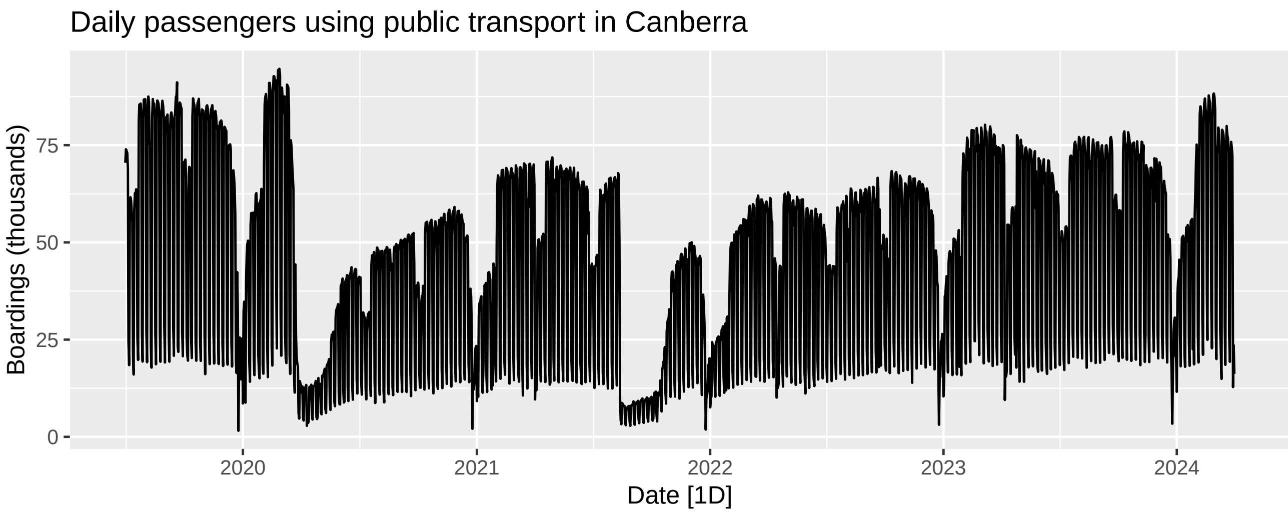

Figure 1

```
act_pt |>
  gg_season(Boardings, period = "week") +
  labs(
    title = "Daily passengers using public transport in Canberra",
    y = "Boardings (thousands)"
  )
```

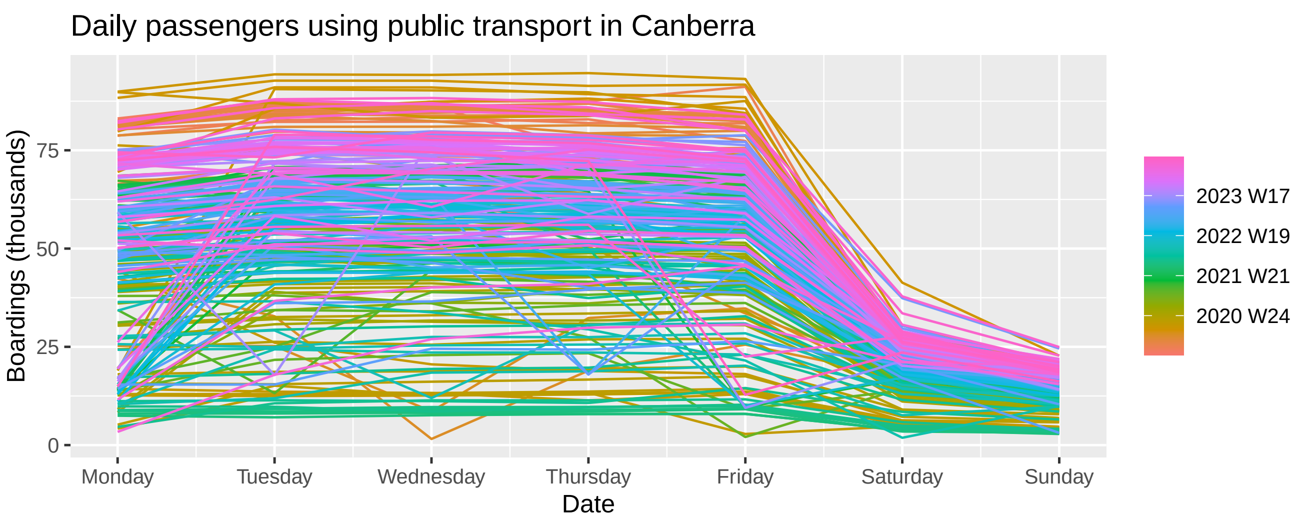

Figure 2

1. Using Figures 1–2, describe the daily passenger boardings for public transport in Canberra. Canberra had two major periods of "COVID-19 lockdowns" where there was substantially reduced travel, and has four school terms per year. Carefully comment on the interesting features of both plots, and how the lockdowns, school terms and other holidays are evident.

5 marks

2. For the STL decomposition in Figure 3, discuss what is shown in each panel. Why has a log transformation been used? Describe how COVID-19 lockdowns and holidays have affected the trend, seasonal and remainder components.

5 marks

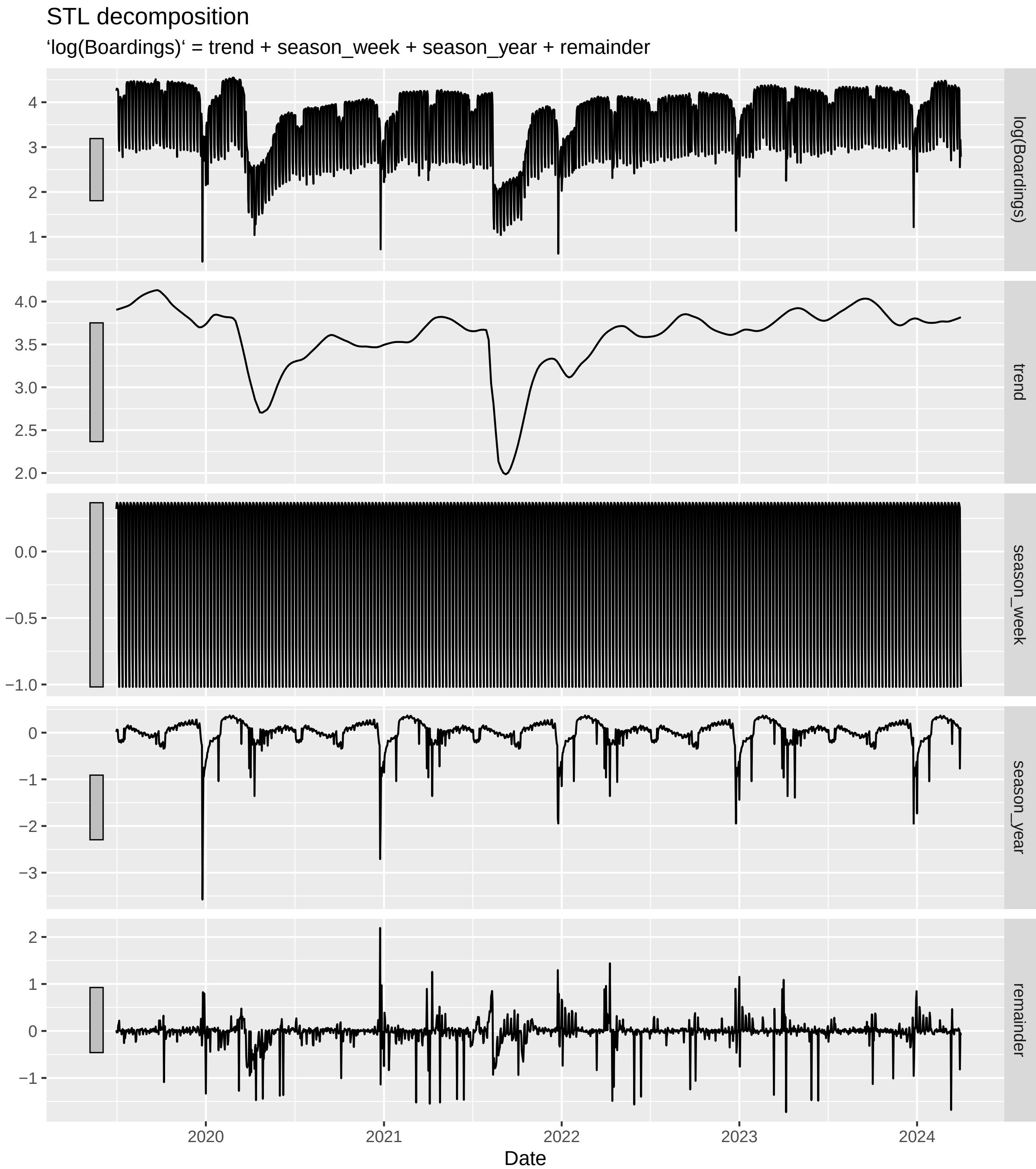

Figure 3

3. You have been asked to provide forecasts for the next four weeks for daily passenger boardings. Consider applying each of the methods and models below to the data from March 2022 onwards. Comment, in a few words each, on whether each one is appropriate for forecasting the next two weeks of data. No marks will be given for simply guessing whether a method or a model is appropriate without justifying your choice.

10 marks

Start your response by stating: **suitable** or **not suitable**.

(a) Seasonal naïve method using weekly seasonality.
(b) Naïve method.
(c) An STL decomposition on the log transformed data combined with an ARIMA to forecast the seasonally adjusted component, and seasonal naïve methods for both seasonal components.
(d) Holt-Winters method with damped trend and multiplicative weekly seasonality.
(e) ETS(A,N,A).
(f) ETS(M,A,M) with annual seasonality.
(g) ARIMA(2,4,2)$(1,1,0)_7$ applied to the log transformed data.
(h) ARIMA(1,0,1)$(1,1,0)_7$ applied to the log transformed data.
(i) Regression with time and Fourier terms for both weekly and annual seasonality.
(j) Dynamic regression on the log transformed data with Fourier terms for the annual seasonality and a seasonal ARIMA model to handle the weekly seasonality and other dynamics.

Total: 20 marks

SECTION C

An ETS model is fitted to the time series shown in Figure 1, but only using data from March 2022 onwards.

```r
act_pt_recent <- act_pt |>
  filter(Date >= ymd("2022-03-01"))
fit <- act_pt_recent |>
  model(ets = ETS(Boardings))
report(fit)
```

```
Series: Boardings
Model: ETS(M,N,M)
  Smoothing parameters:
    alpha = 0.208
    gamma = 1e-04

  Initial states:
 l[0] s[0] s[-1] s[-2] s[-3] s[-4] s[-5] s[-6]
 43.5 1.17  0.35  0.49   1.2  1.24  1.28  1.27

  sigma^2:  0.0338

 AIC AICc  BIC
8256 8256 8302
```

1. Write down the equations for the model, including specifying the values of all model parameters.

   5 marks

2. What features of the data has the model ignored?

   2 marks

3. What does the value of γ tell you?

   1 marks

4. The data, remainder (i.e., residuals), and estimated states, are shown below for the last week of observations, along with the forecasts for the next day. Show how the forecast mean has been obtained, and give a 95% prediction interval for this day.

   5 marks

```r
components(fit) |> tail(7)
```

```
# A dable: 7 x 6 [1D]
# Key:     .model [1]
# :        Boardings = lag(level, 1) * lag(season, 7) * (1 + remainder)
  .model Date       Boardings level season remainder
  <chr>  <date>         <dbl> <dbl>  <dbl>     <dbl>
1 ets    2024-03-25      72.5  58.9  1.17     0.0712
2 ets    2024-03-26      75.8  59.0  1.28     0.0107
3 ets    2024-03-27      74.5  58.9  1.28    -0.0103
4 ets    2024-03-28      72.2  58.7  1.24    -0.0118
5 ets    2024-03-29      12.8  48.7  1.20    -0.819
6 ets    2024-03-30      23.6  48.6  0.490   -0.0125
7 ets    2024-03-31      16.2  48.1  0.350   -0.0485
```

```r
forecast(fit, h = 1)
```

```
# A fable: 1 x 4 [1D]
# Key:     .model [1]
  .model Date         Boardings .mean
  <chr>  <date>          <dist> <dbl>
1 ets    2024-04-01 N(56, 106)   56.1
```

5. Some plots of the residuals are shown in Figure 4. Discuss what these tell you about the model?

5 marks

```r
gg_tsresiduals(fit)
```

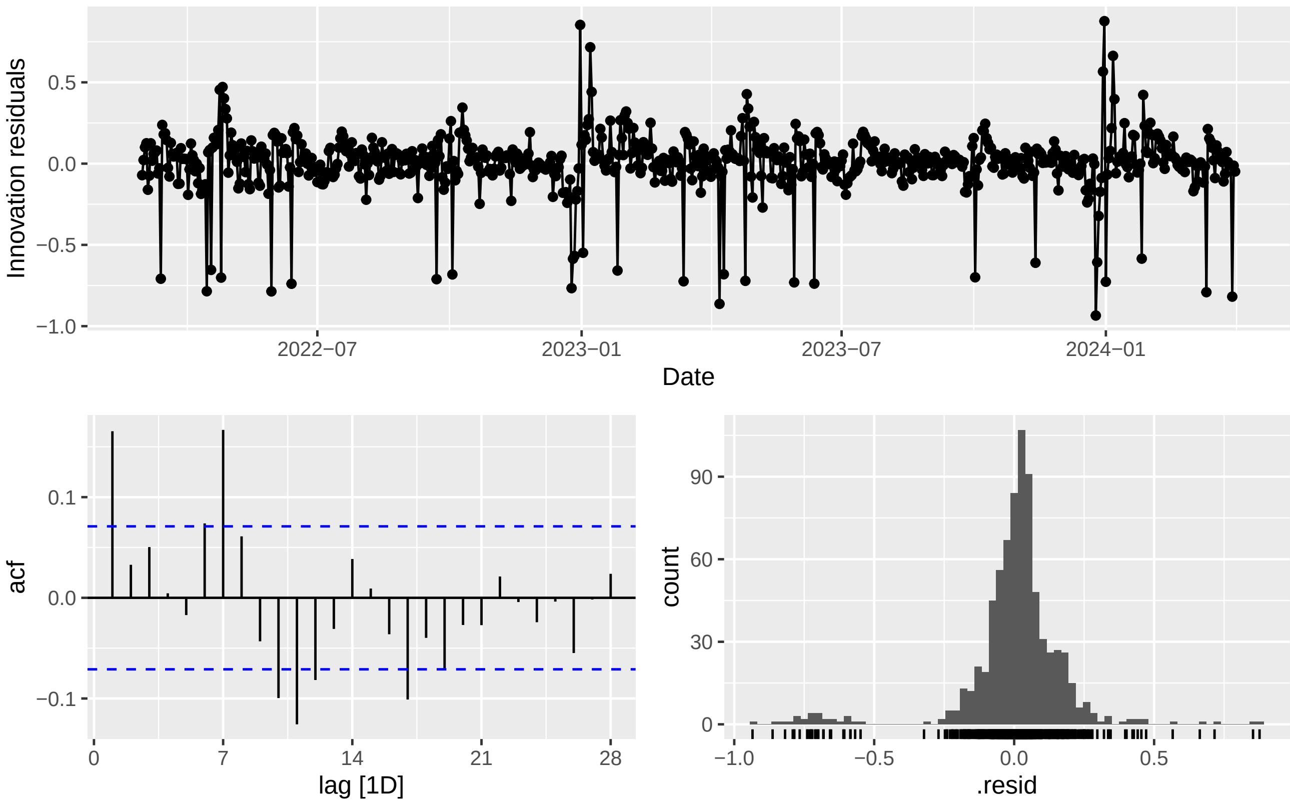

Figure 4:

6. If you conducted a Ljung-Box test of the residuals using 14 lags, what do you think the p-value would be? Why?

2 marks

Total: 20 marks

SECTION D

An ARIMA model is fitted to the time series shown in Figure 1, but only using data from March 2022 onwards.

```
fit <- act_pt_recent |>
  model(arima = ARIMA(log(Boardings)))
report(fit)
```

```
Series: Boardings
Model: ARIMA(1,0,1)(1,1,0)[7]
Transformation: log(Boardings)

Coefficients:
         ar1      ma1     sar1
      0.8165  -0.5920  -0.4574
s.e.  0.0399   0.0518   0.0336

sigma^2 estimated as 0.1045:  log likelihood=-218
AIC=444   AICc=444   BIC=463
```

1. Write down the equations for the model using backshift notation, including specifying the values for all model parameters.

5 marks

2. This model suggests that the weekly differences of the logged data are stationary. What aspects of the model lead to that conclusion?

3 marks

3. Let the observed data be given by y_t, and define $x_t = \log(y_t) - \log(y_{t-7})$. The x_t series is shown in Figure 5. What features of these plots suggest that x_t is stationary?

2 marks

```
act_pt_recent |>
  mutate(x = difference(log(Boardings), lag = 7)) |>
  gg_tsdisplay(x, plot_type = "partial")
```

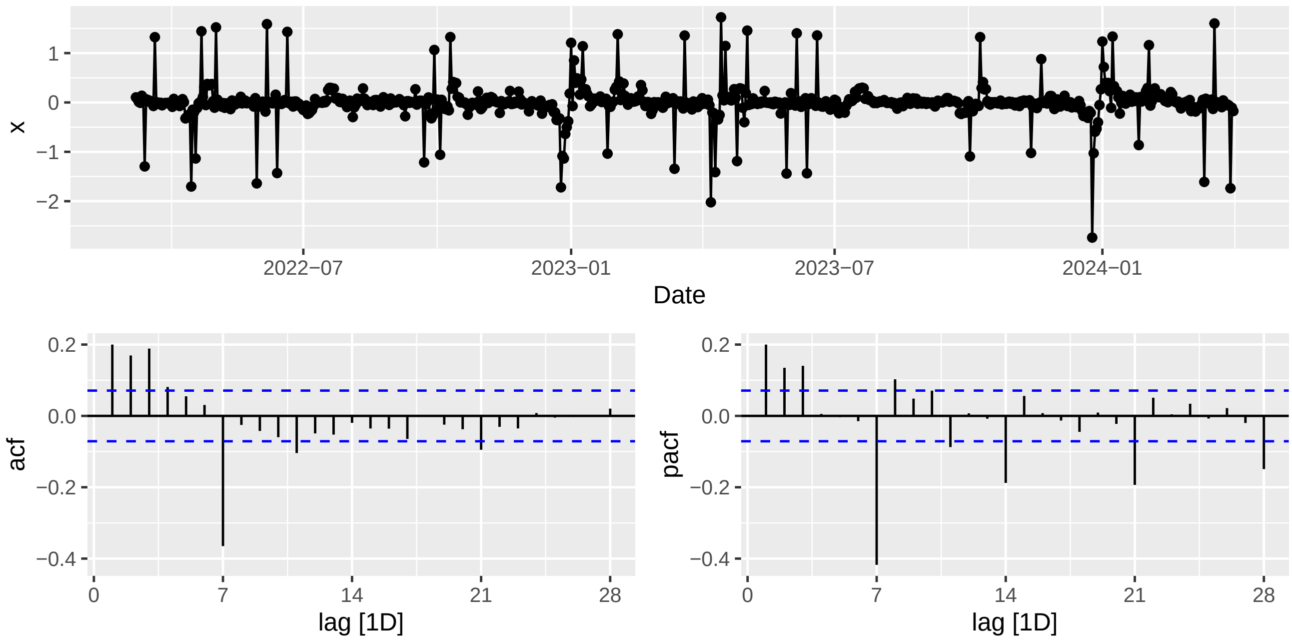

Figure 5:

4. The residuals shown in Figure 6 show a lot of large outliers. What might be causing these outliers?

2 marks

```
gg_tsresiduals(fit)
```

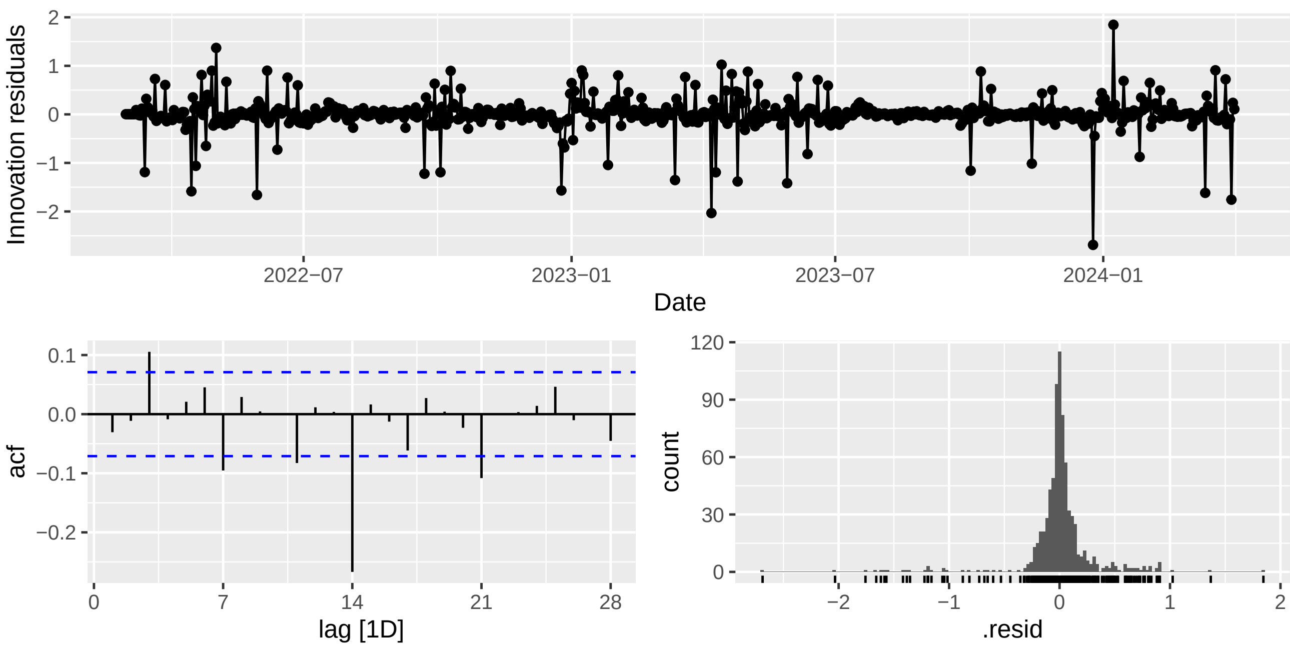

Figure 6:

5. The residuals shown in Figure 6 are clearly not white noise, and are not normally distributed. What features of the plots suggest this?

2 marks

6. If you produced forecasts from this model, how reliable do you think the point forecasts and prediction intervals would be? Why?

3 marks

7. Another model is found with 10 parameters, a slightly lower AICc value, but with no improvement in the residuals. Would you prefer this model to the one you have already fitted? Why or why not?

3 marks

Total: 20 marks

SECTION E

A dynamic regression model is fitted to the time series shown in Figure 1, but only using data from March 2022 onwards.

```
fit <- act_pt_recent |>
  model(dynreg = ARIMA(log(Boardings) ~ fourier("year", K = 10)))

report(fit)
```

```
Series: Boardings
Model: LM w/ ARIMA(0,0,1)(1,1,2)[7] errors
Transformation: log(Boardings)

Coefficients:
         ma1     sar1     sma1     sma2  fourier("year", K = 10)C1_365
      0.1040   -0.698   -0.235   -0.648                        -0.0385
s.e.  0.0357    0.957    0.949    0.883                         0.0179
      fourier("year", K = 10)S1_365  fourier("year", K = 10)C2_365
                            -0.0261                        -0.0925
s.e.                         0.0167                         0.0150
      fourier("year", K = 10)S2_365  fourier("year", K = 10)C3_365
                             0.0503                        -0.1035
s.e.                         0.0152                         0.0146
      fourier("year", K = 10)S3_365  fourier("year", K = 10)C4_365
                             0.0105                        -0.1579
s.e.                         0.0147                         0.0146
      fourier("year", K = 10)S4_365  fourier("year", K = 10)C5_365
                            -0.0152                        -0.0328
s.e.                         0.0145                         0.0144
      fourier("year", K = 10)S5_365  fourier("year", K = 10)C6_365
                            -0.0215                        -0.0297
s.e.                         0.0144                         0.0144
      fourier("year", K = 10)S6_365  fourier("year", K = 10)C7_365
                             0.0119                        -0.0606
s.e.                         0.0144                         0.0144
      fourier("year", K = 10)S7_365  fourier("year", K = 10)C8_365
                            -0.0016                        -0.0367
s.e.                         0.0143                         0.0144
      fourier("year", K = 10)S8_365  fourier("year", K = 10)C9_365
                            -0.0076                         0.0147
s.e.                         0.0144                         0.0143
      fourier("year", K = 10)S9_365  fourier("year", K = 10)C10_365
                             0.0895                         -0.0306
s.e.                         0.0144                          0.0143
      fourier("year", K = 10)S10_365
                              0.0448
s.e.                          0.0143

sigma^2 estimated as 0.07058:  log likelihood=-65.4
AIC=181   AICc=183   BIC=296
```

1. Write down the equations for the regression part of the model. There is no need to give numerical values for the coefficients.

   5 marks

2. Which coefficients in the model relate to annual seasonality, and which coefficients relate to weekly seasonality? What does the remaining coefficient handle?

   3 marks

3. The model could be improved by changing the number of Fourier terms used. Explain how you could determine the optimal number of Fourier terms to use.

   3 marks

4. It is thought that days on which rain is forecast may have fewer passengers using public transport. How could you incorporate this information into your forecasts?

   3 marks

5. **ETC3550 only** To compare all the models used so far, you decide to use a test set comprising the observations from January 2024 to March 2024. Explain how you would use this test set to compare the models, and what you would look for in the comparison. Discuss whether this is the most reliable way to compare the models.

   6 marks

6. **ETC5550 only** You decide to compare all the models used so far, in a time series cross-validation comparison. Explain what this means, and why it is a useful way to compare the models in this way.

   6 marks

Total: 20 marks