The exam contains FIVE questions. ALL questions must be answered. The exam is worth 100 marks in total.

SECTION A

Write about a quarter of a page each on any **four** of the following topics. Clearly state if you agree or disagree with each statement.

1. Events like COVID-19 show that forecasting is never a good idea because the future is too unpredictable.
2. Taking logarithms of the data is useful for stablizing the variance of a time series.
3. Regression models are better than ARIMA models because the coefficients are more interpretable.
4. The best forecasting model has white noise residuals.
5. Choosing a model using the AICc is better than choosing a model on the basis of a test set because it involves all the data.
6. The MAPE is the best accuracy measure because it is easy to understand and is independent of the scale of the data.

Total: 20 marks

— END OF SECTION A —

SECTION B

Figures 1, 2 and 3 relate to the daily page views on the OTexts website from 1 January 2018 to 10 April 2022.

1. Using Figures 1, 2 and 3, describe the daily page views on the OTexts website. Carefully comment on the interesting features of all three plots.

6 marks

```
otexts %>%
  autoplot(Pageviews) +
  labs(subtitle = "Daily Pageviews on OTexts.com", y = "Thousands")
```

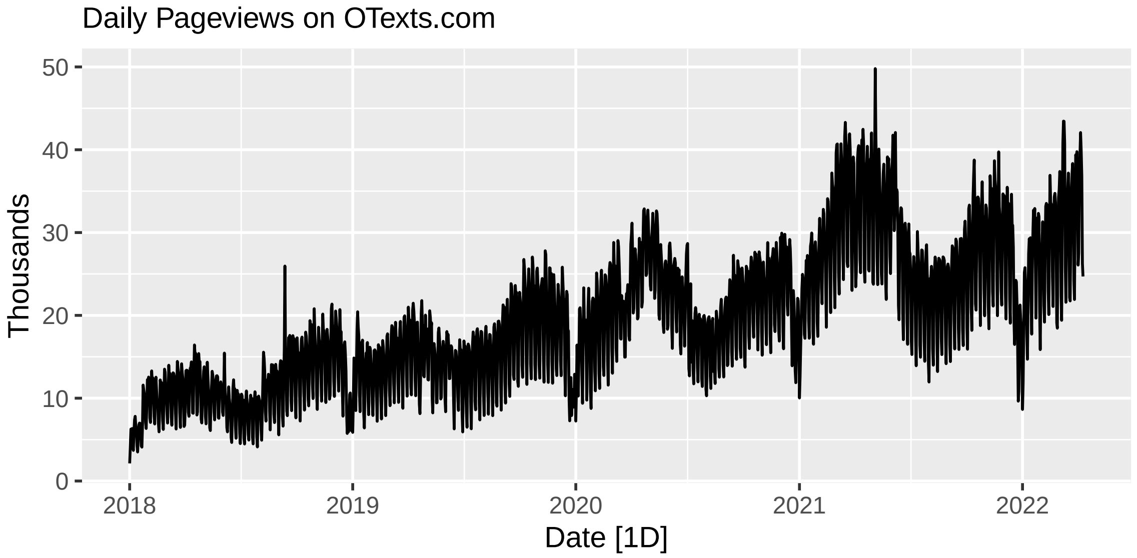

Figure 1:

```
otexts %>%
  gg_season(Pageviews, period = "year") +
  labs(subtitle = "Daily Pageviews on OTexts.com", y = "Thousands")
```

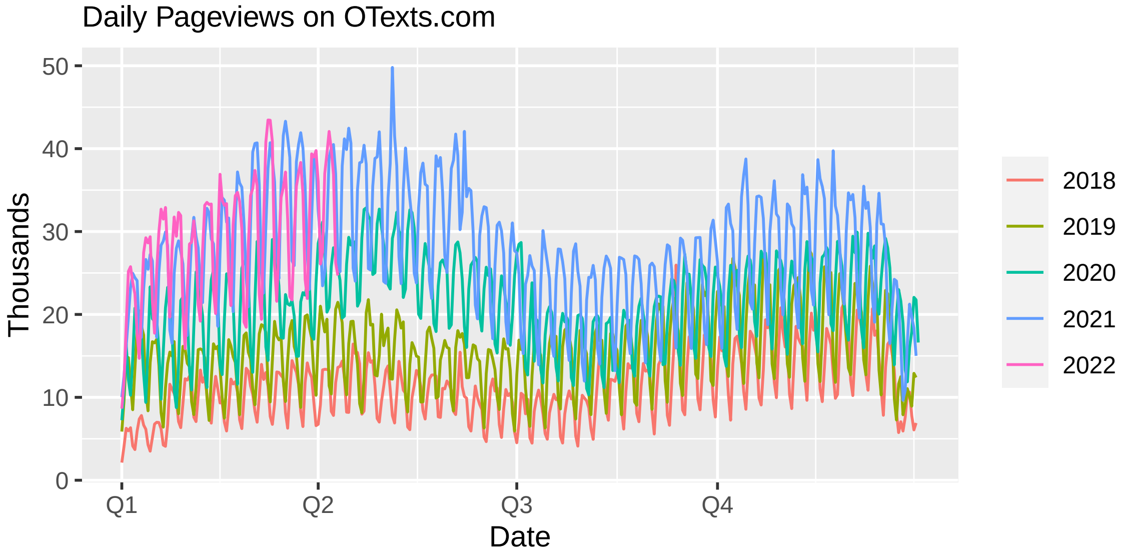

Figure 2:

```
otexts %>%
  gg_season(Pageviews, period = "week") +
  labs(subtitle = "Daily Pageviews on OTexts.com", y = "Thousands")
```

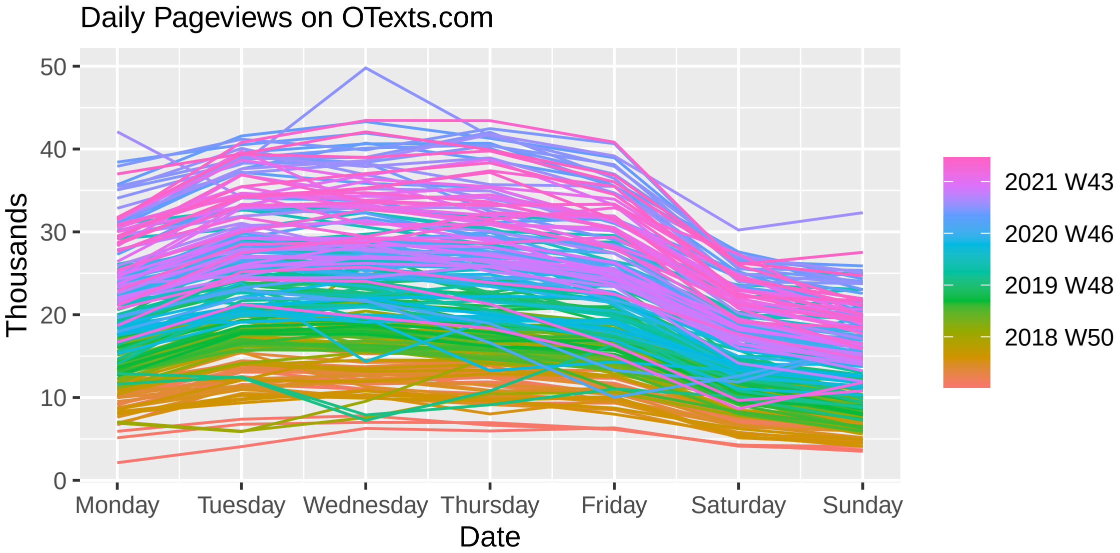

Figure 3:

2. Figure 4 was created using the code below. Why has a logarithm been used? Comment on the choice of the `window` argument in each term. What would you expect if the `window` values were substantially changed?

4 marks

```
otexts %>%
  model(STL(log(Pageviews) ~ trend(window = 99) +
    season(period = "week", window = 99) +
    season(period = "year", window = 9),
    robust = TRUE
  )) %>%
  components() %>%
  autoplot()
```

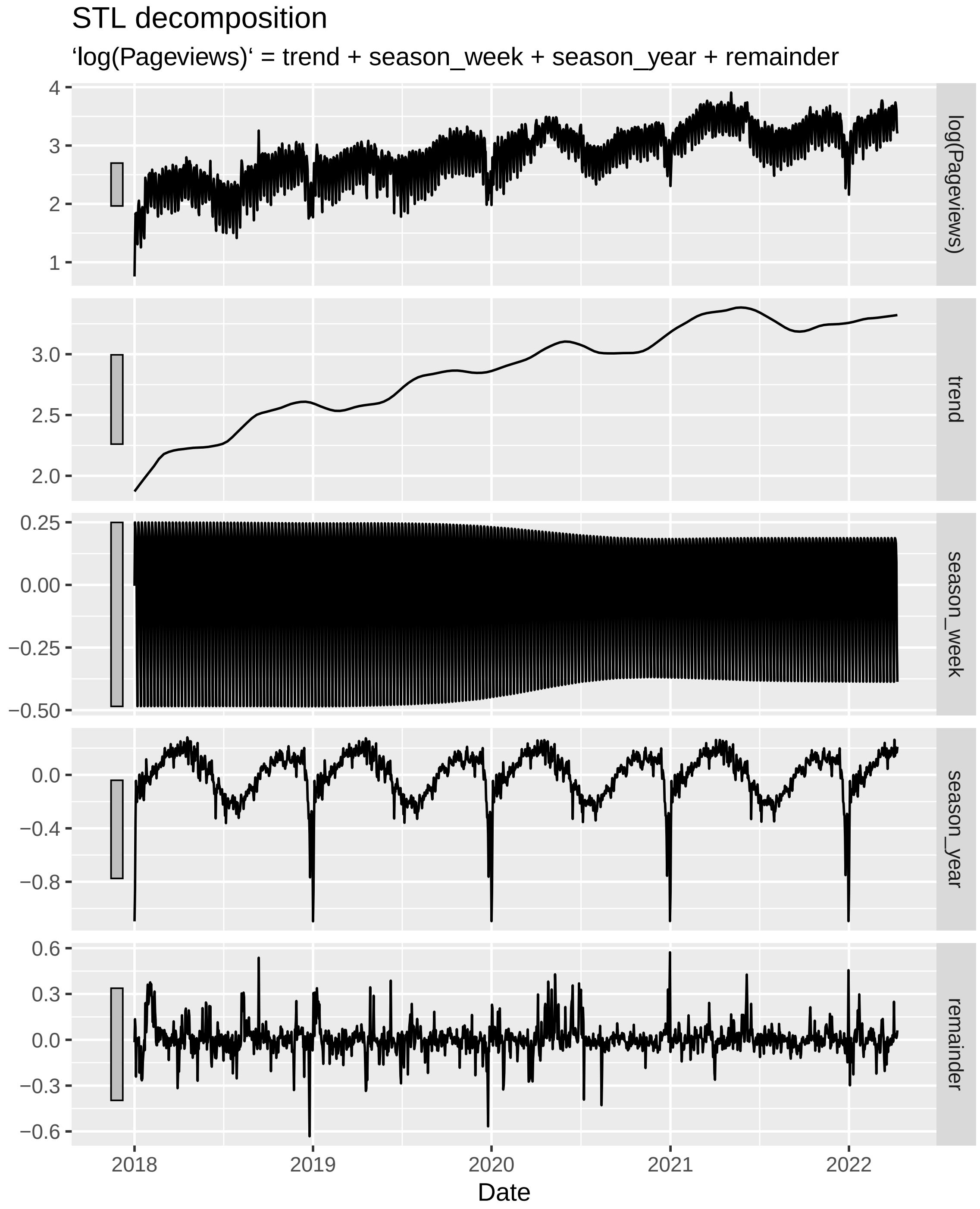
STL decomposition
'log(Pageviews)' = trend + season_week + season_year + remainder
4
3
2
1
log(Pageviews)
3.0
2.5
2.0
trend
0.25
0.00
−0.25
−0.50
season_week
0.0
−0.4
−0.8
season_year
0.6
0.3
0.0
−0.3
−0.6
remainder
2018
2019
2020
2021
2022
Date

Figure 4:

3. You have been asked to provide forecasts for the next four weeks for OTexts pageviews.

   Consider applying each of the methods and models below. Comment, in a few words each, on whether each one is appropriate for forecasting the data. No marks will be given for simply guessing whether a method or a model is appropriate without justifying your choice.

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

   (a) Seasonal naïve method using annual seasonality.

   (b) Seasonal naïve method using weekly seasonality.

   (c) An STL decomposition on the log transformed data combined with an ETS 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(2,2,2)(0,0,0)$_7$ applied to the log transformed data.

   (h) ARIMA(0,1,1)(0,1,1)$_7$ 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.

10 marks

Total: 20 marks

— END OF SECTION B —

SECTION C

The output below shows the results of fitting an ETS model to the `Pageviews` variable.

```
fit <- otexts %>%
  model(ETS(Pageviews))
report(fit)
```

```
## Series: Pageviews
## Model: ETS(M,A,M)
##   Smoothing parameters:
##     alpha = 0.5737414
##     beta  = 0.0001000033
##     gamma = 0.1266783
##
##   Initial states:
##      l[0]      b[0]      s[0]     s[-1]    s[-2]    s[-3]    s[-4]   s[-5]
##  3.605899 0.1449521 0.6318171 0.7210449 1.035443 1.163588 1.300494 1.20237
##      s[-6]
##  0.9452431
##
##   sigma^2:  0.0101
##
##      AIC     AICc      BIC
## 13337.34 13337.54 13401.58
```

1. Write down the observation and state equations for the model, specifying which is the observation equation, what parameters have been optimized, and explaining why this particular model was chosen.

   6 marks

2. Figure 5 shows the components of the model. Explain what has been plotted, and how these plots relate to the equations shown earlier. Explain how the values of `beta` and `gamma` shown earlier correspond to features of these plots.

   5 marks

   ```
   fit %>%
     components() %>%
     autoplot()
   ```

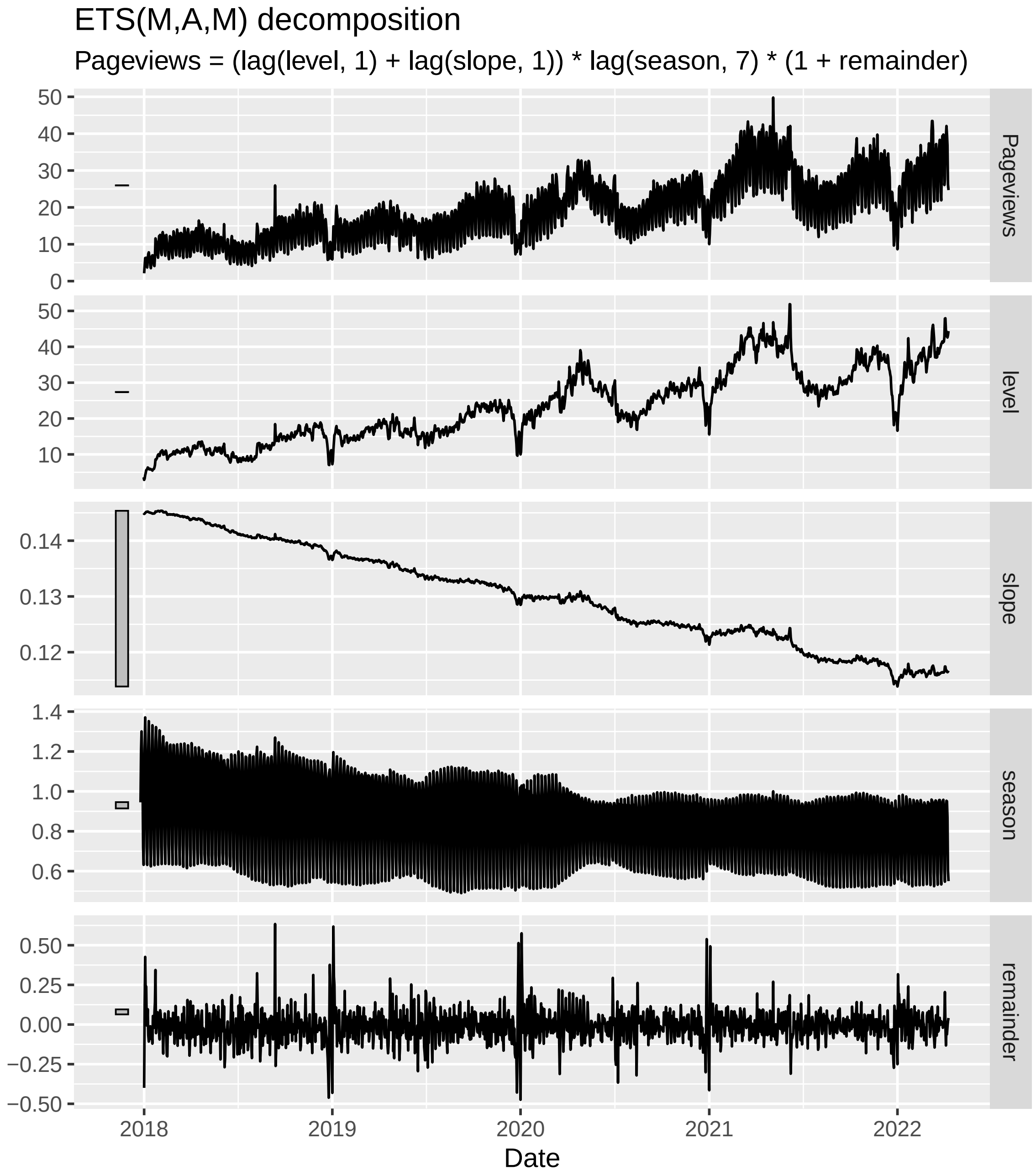

Figure 5:

3. How have the annual and weekly seasonalities been handled by this ETS model?

3 marks

4. Comment on the large residuals seen at the end of each year, and the dip in the level at the end of each year. What is causing these?

2 marks

5. Do you expect this model to produce good forecasts for the next 3 weeks? What about for the next 12 months? Explain.

4 marks

Total: 20 marks

— END OF SECTION C —

SECTION D

It is decided to fit a dynamic regression model to the data, with Fourier terms to handle the annual seasonality, and a seasonal ARIMA error to handle the weekly seasonality.

```r
fit_arima <- otexts %>%
  model(
    dr = ARIMA(log(Pageviews) ~ fourier(period = "year", K = 4) + PDQ(period = "week"))
  )
report(fit_arima)
```

```
## Series: Pageviews
## Model: LM w/ ARIMA(1,0,2)(2,1,1)[7] errors
## Transformation: log(Pageviews)
##
## Coefficients:
##          ar1      ma1      ma2    sar1    sar2     sma1
##       0.8858  -0.1213  -0.1750  0.1600  0.0247  -0.9046
## s.e.  0.0236   0.0361   0.0341  0.0307  0.0298   0.0182
##       fourier(period = "year", K = 4)C1_365
##                                      0.0071
## s.e.                                 0.0318
##       fourier(period = "year", K = 4)S1_365
##                                      0.0928
## s.e.                                 0.0319
##       fourier(period = "year", K = 4)C2_365
##                                     -0.1580
## s.e.                                 0.0251
##       fourier(period = "year", K = 4)S2_365
##                                     -0.0818
## s.e.                                 0.0255
##       fourier(period = "year", K = 4)C3_365
##                                     -0.0440
## s.e.                                 0.0228
##       fourier(period = "year", K = 4)S3_365
##                                      0.0183
## s.e.                                 0.0229
##       fourier(period = "year", K = 4)C4_365
##                                     -0.0575
## s.e.                                 0.0208
##       fourier(period = "year", K = 4)S4_365
##                                      0.0018
## s.e.                                 0.0209
##
## sigma^2 estimated as 0.009012:  log likelihood=1456.1
## AIC=-2882.21   AICc=-2881.9   BIC=-2801.98
```

1. Write down the model using backshift notation.

6 marks

2. Comment on the model diagnostics shown in Figure 6 and the output below. How might the model be improved? Do you think the resulting forecasts will be reliable?

5 marks

```
augment(fit_arima) %>%
  gg_tsdisplay(.innov, "histogram")
```

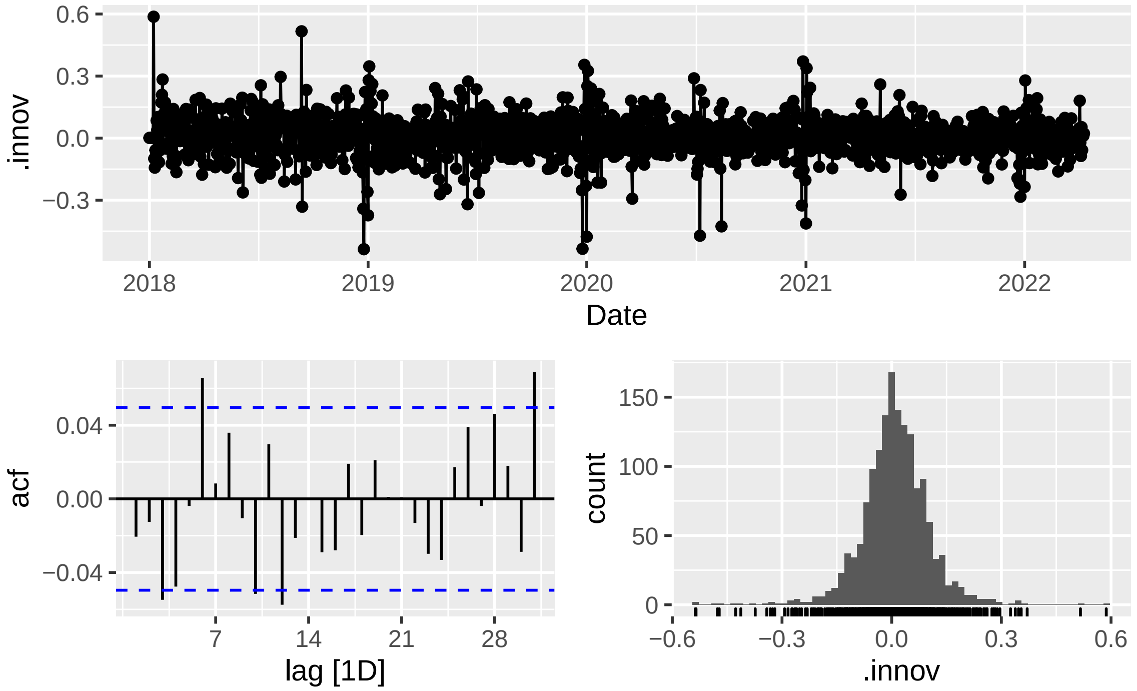

Figure 6:

```
augment(fit_arima) %>%
  features(.innov, ljung_box, lag = 28, dof = 14)
```

```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 dr        43.9 0.0000617
```

3. Figure 7 shows forecasts with prediction intervals for the next four weeks, along with the data from 2022. The forecasts appear to have no trend. Why is that? If you wanted to include a *local* trend, how would you modify the model?

3 marks

```
fit_arima %>%
  forecast(h = "4 weeks") %>%
  autoplot(otexts %>% filter(year(Date) == 2022)) +
  labs(title = "Data and forecasts for 2022")
```

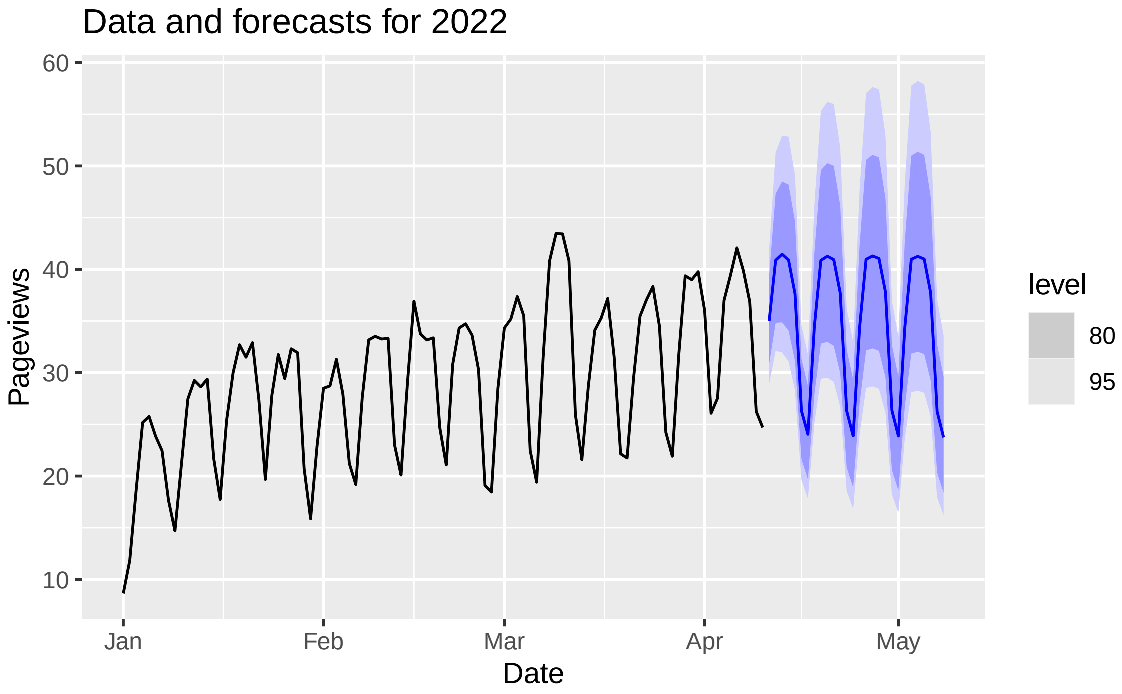

Figure 7:

4. The one-step-ahead forecast is 35.0 pageviews. Give a 95% prediction interval assuming Gaussian innovation residuals.

4 marks

5. It is thought that pageviews will be higher during teaching semesters. How would you modify this model to allow for a semester effect?

2 marks

Total: 20 marks

— END OF SECTION D —

SECTION E

1. You decide to compare three different models on this data set: (a) the ETS model from Section C; (b) the dynamic regression model from Section D; and (c) the STL decomposition shown in Figure 4 with ETS applied to the seasonally adjusted data, and seasonal naive methods applied to both seasonal components.

   The following code uses a test set of the last 4 weeks to compare the three models, along with two benchmark methods.

```
fit <- otexts %>%
  head(-28) %>%
  model(
    ets = ETS(Pageviews),
    dr = ARIMA(log(Pageviews) ~ fourier(period = "year", K = 4) +
                                PDQ(period = "week")),
    stl = decomposition_model(
            STL(log(Pageviews) ~ trend(window = 99) +
                  season(period = "week", window = 99) +
                  season(period = "year", window = 9),
                  robust = TRUE),
            ETS(season_adjust)
    ),
    naive = NAIVE(Pageviews),
    snaive = SNAIVE(Pageviews ~ lag("week"))
  )
fc <- fit %>%
  forecast(h = 28)
fc %>%
  accuracy(otexts, measures=list(RMSE=RMSE, MAPE=MAPE)) %>%
  arrange(RMSE)
```

```
## # A tibble: 5 x 4
##   .model .type  RMSE  MAPE
##   <chr>  <chr> <dbl> <dbl>
## 1 dr     Test   3.19  7.75
## 2 ets    Test   3.67  9.87
## 3 stl    Test   3.96  9.72
## 4 snaive Test   4.50 11.1
## 5 naive  Test  12.8  31.4
```

   What do you conclude from the above output about the five models? Explain the two accuracy measures used. Why is the naive method so bad?

   4 marks

2. An alternative approach to comparing the forecast accuracy of models would be to use time series cross-validation. Explain the concept of time series cross-validation. You may use an annotated diagram.

   6 marks

3. **ETC3550 students only**: You decide the `snaive` model is good enough even though it is not as accurate as the first three models. Let the observations be $y_1, \dots, y_T$ and the residual variance be denoted by σ^2. Write down the forecast distribution for an h-step forecast. What assumptions have you made?

   5 marks

4. **ETC5550 students only**: Someone proposes a fancy new forecasting method that uses some predictor variables such as the number of students in each university that is using an OTexts book as a recommended text, and the relative wealth of the countries they live in. This new method reduces the cross-validated one-step RMSE by 2%, but takes about a day to estimate the forecasts. Would you recommend using the new method? Explain your reasons.

   5 marks

5. OTexts also needs forecasts of the maximum monthly traffic that could arise, in order to make sure their internet server will cope with the extreme demand. They want to choose an internet plan that allows for a maximum of P pageviews per month, and they are happy to allow a 1% chance of being above this level. Explain how you would choose P based on forecasts over the next 12 months.

   5 marks

Total: 20 marks

— END OF SECTION E —