Lab 3: Simple forecasting tools and linear regression

Introduction to Time Series Analysis

Name:
This lab is to be done in class (completed outside of class if need be). You can collaborate with your classmates, but you must identify their names above, and you must submit **your own** lab as an knitted pdf file. To answer the questions, display the results and write your answers if asked.

1. For each of the following time series, calculate the residuals from a naive or seasonal naïve forecast. Test if the residuals are white noise and normally distributed. What do you conclude?

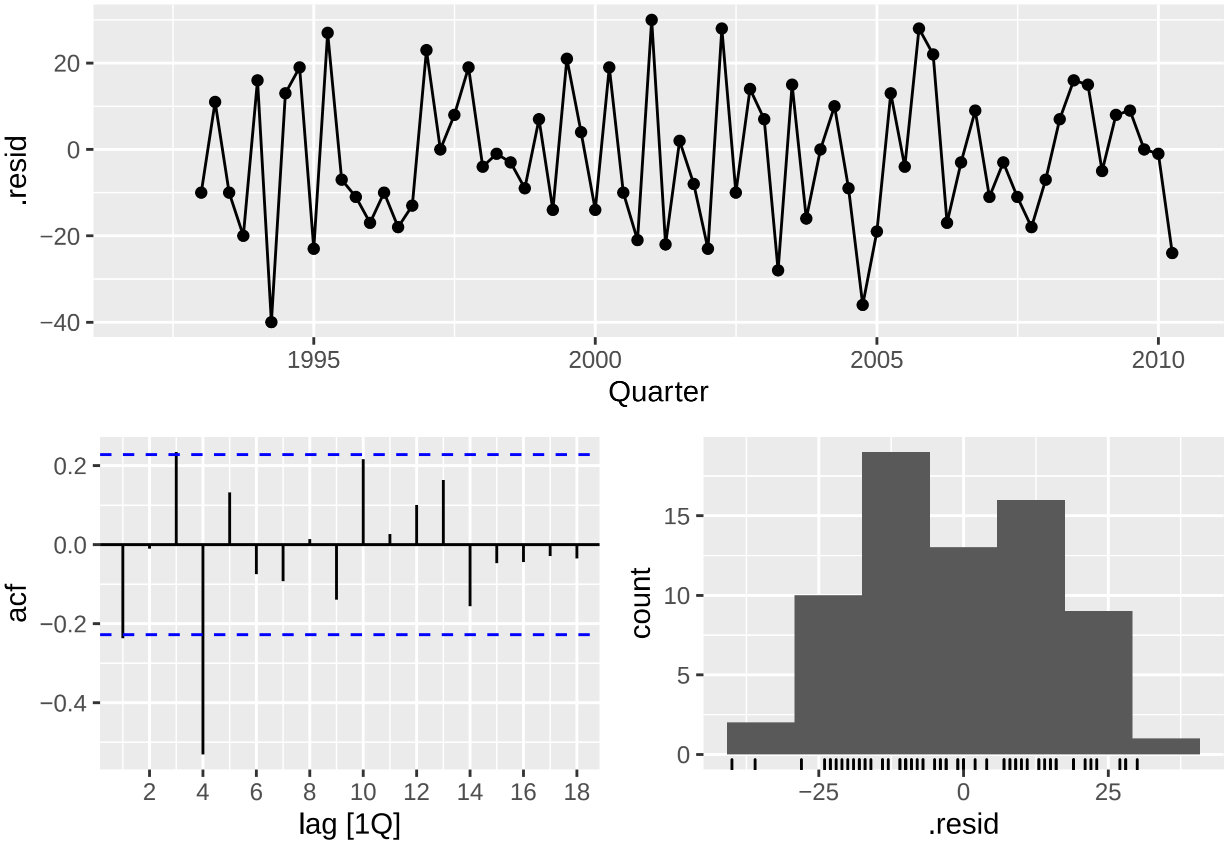
- Quarterly Australian beer production data from 1992
- Australian Exports series from `global_economy`
- Bricks series from `aus_production`

```
# Quarterly Australian beer production data from 1992
recent_production <- aus_production %>% filter(year(Quarter) >= 1992)
fit <- recent_production %>% model(SNAIVE(Beer))
fit %>% gg_tsresiduals()
```

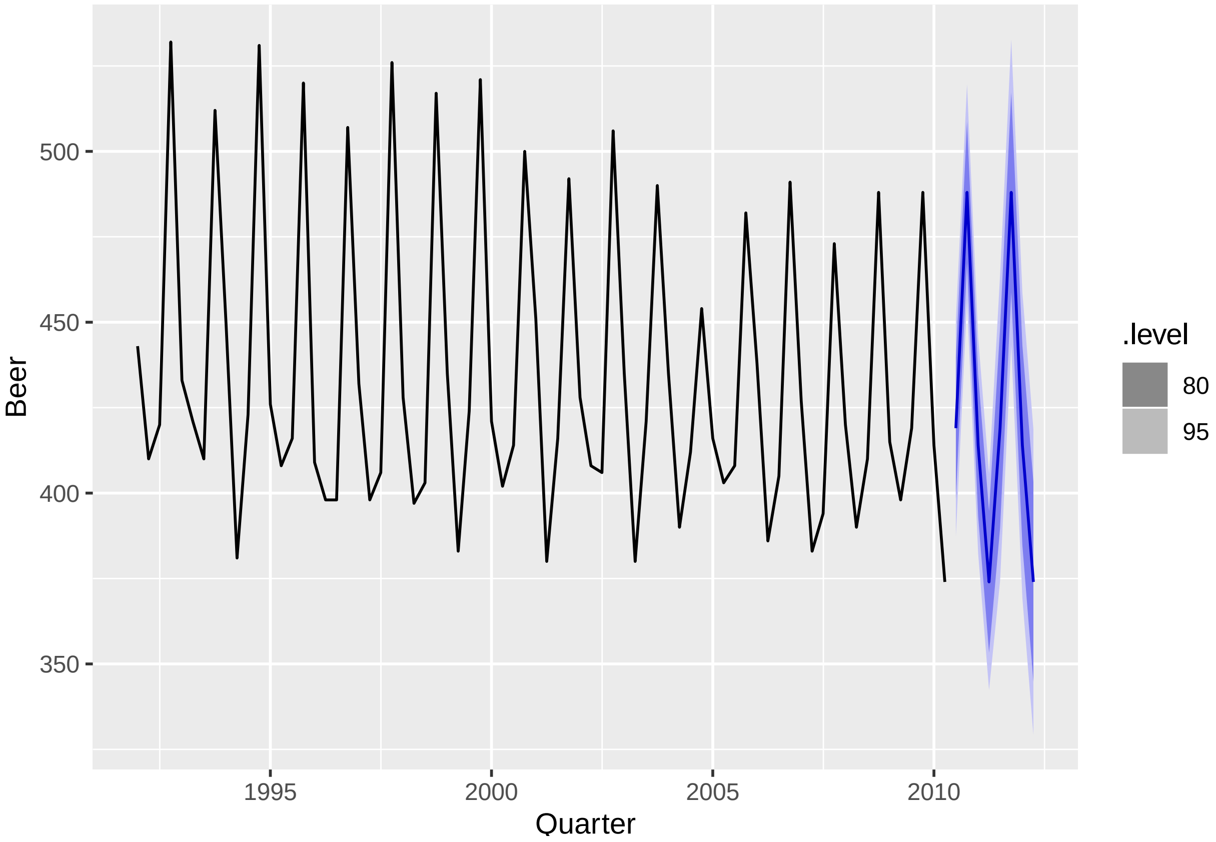
```
## Warning: Removed 4 rows containing missing values (geom_path).

## Warning: Removed 4 rows containing missing values (geom_point).

## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



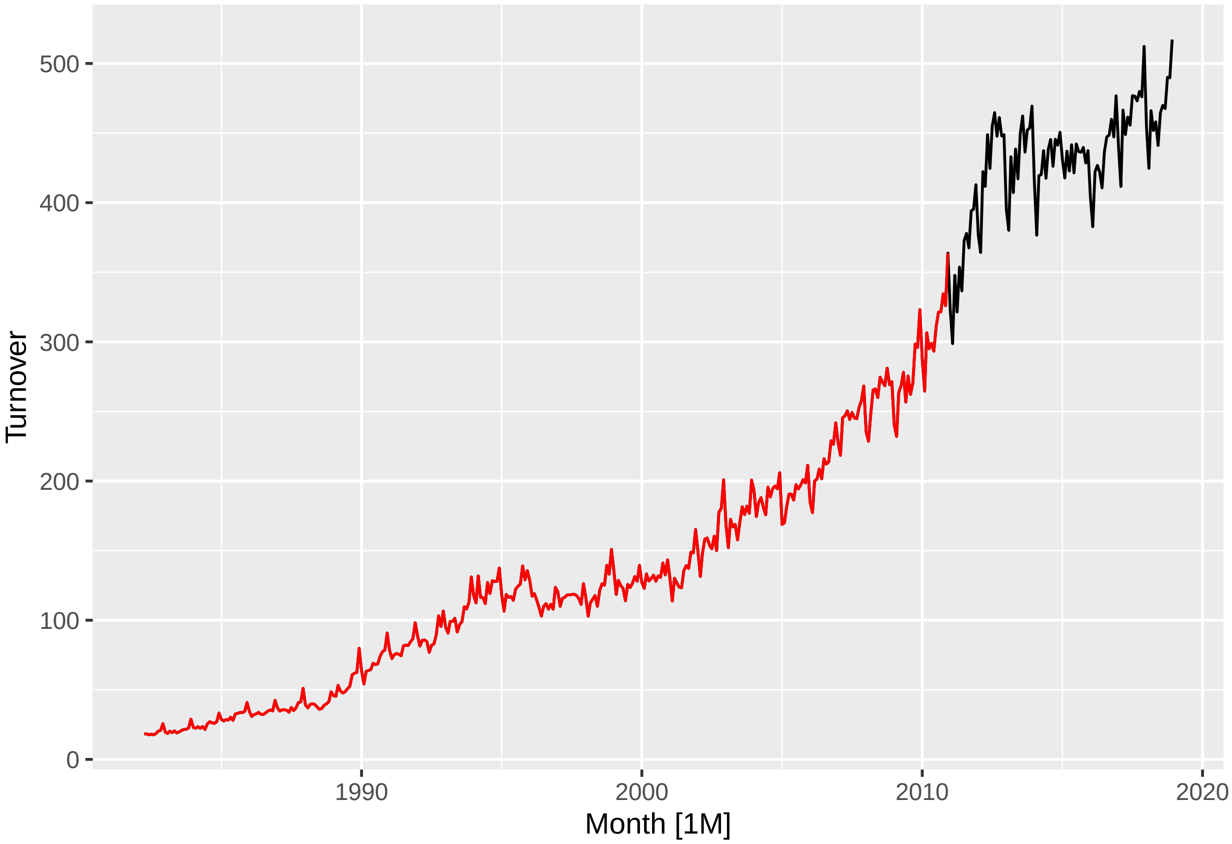
```
fit %>% forecast() %>% autoplot(recent_production)
```



2. For the turnover time series in the following dataset, conduct the following tasks:

```
set.seed(12345678)
myts <- aus_retail %>%filter(`Series ID` == sample(aus_retail$`Series ID`,1))

myts_train <- myts %>%filter(Month <= yearmonth("2010 Dec"))
myts%>%autoplot(Turnover) +autolayer(myts_train,Turnover, colour = "red")
```



```
fit <- myts_train %>% model(SNAIVE(Turnover))
fc <- fit %>% forecast()
fit %>% accuracy()
```

```
## # A tibble: 1 x 11
##   State   Industry        .model  .type    ME  RMSE   MAE   MPE  MAPE  MASE  ACF1
##   <chr>   <chr>           <chr>   <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Wester~ Cafes, restau~ SNAIVE~ Trai~  10.5  16.5  12.4  9.03  10.6     1 0.830
```

```
fc %>% accuracy(myts)
```

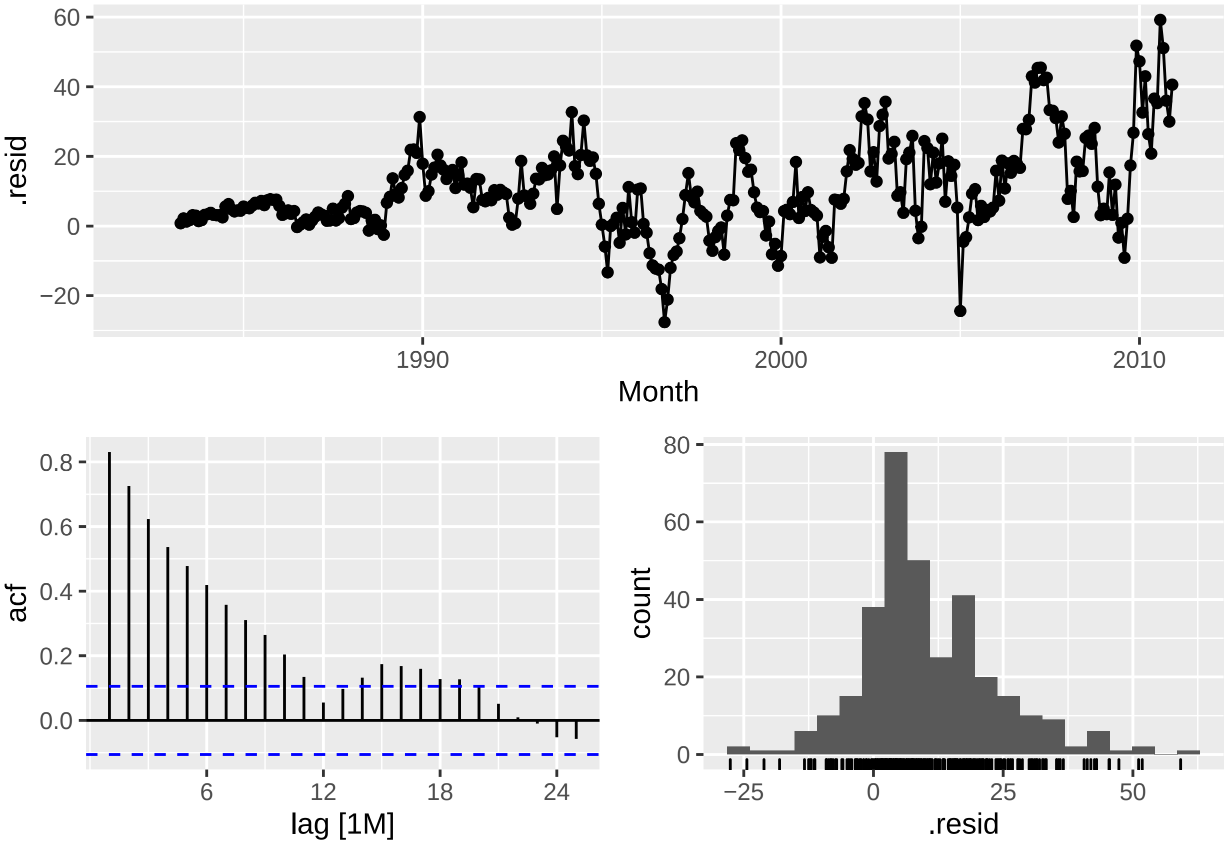
```
## # A tibble: 1 x 11
##   .model    State   Industry      .type    ME  RMSE   MAE   MPE  MAPE  MASE  ACF1
##   <chr>     <chr>   <chr>         <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 SNAIVE(~ Wester~ Cafes, resta~ Test   84.5  93.4  84.5  20.6  20.6  6.83 0.868
```

```
fit %>% gg_tsresiduals()
```

```
## Warning: Removed 12 rows containing missing values (geom_path).

## Warning: Removed 12 rows containing missing values (geom_point).

## Warning: Removed 12 rows containing non-finite values (stat_bin).
```



- Create a training dataset consisting of observations before 2011
- Calculate seasonal naïve forecasts using `SNAIVE()` applied to your training data
- Compare the accuracy (e.g. using RMSE, MAE, MASE) of your forecasts against the actual values.
- Check the residuals. Do the residuals appear to be uncorrelated and normally distributed?
- How sensitive are the accuracy measures to the amount of training data used?

3. Consider the number of pigs slaughted in New South Wales (data set `aus_livestock`).

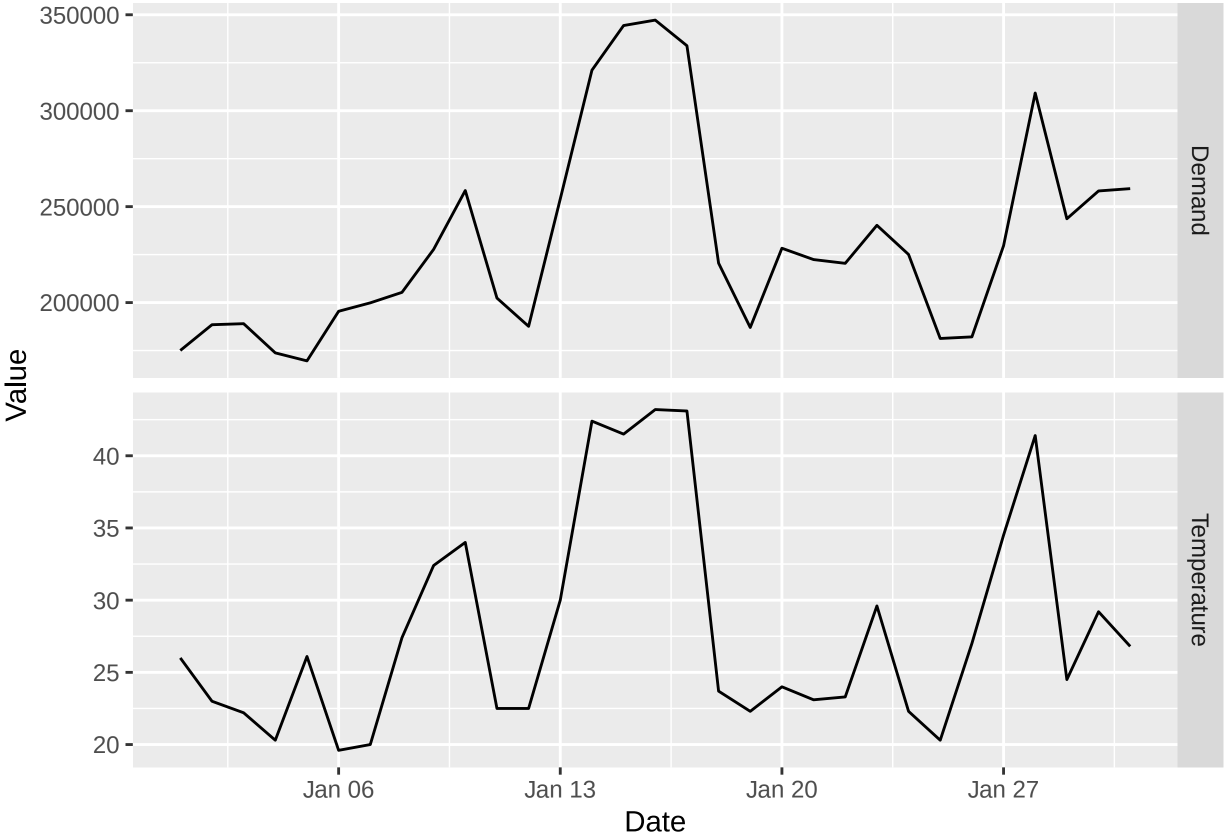
    - Produce some plots of the data in order to become familiar with it.
    - Create a training set of 486 observations, witholding a test set of 72 observations (6 years).
    - Try using various simple methods to forecast the training set and compare the results on the test set. Which method did best?
    - Check the residuals of your preferred method. Do they resemble white noise?

4. Half-hourly electricity demand for Victoria, Australia is contained in vic_elec. Extract the January 2014 electricity demand, and aggregate this data to daily with daily total demands and maximum temperatures.

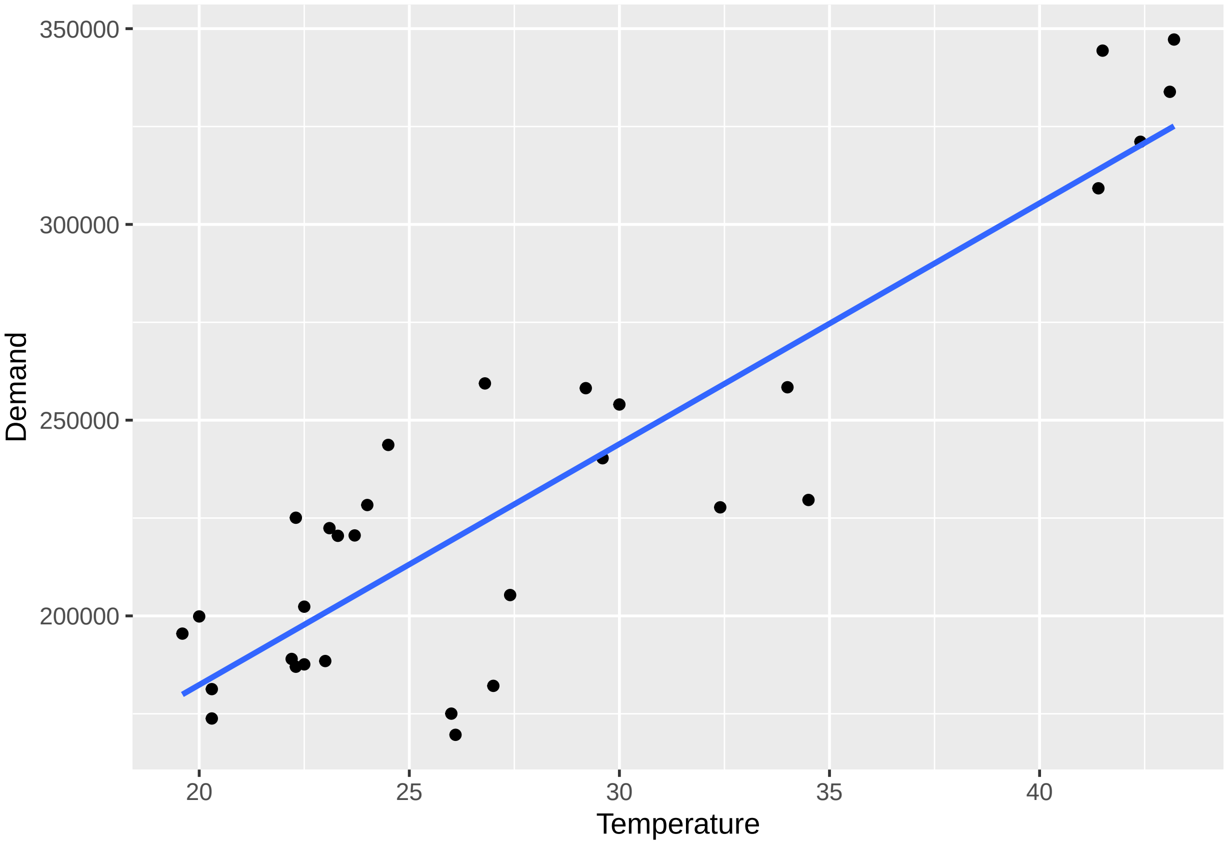
```
jan14_vic_elec <- vic_elec %>%
 filter(yearmonth(Time) == yearmonth("2014 Jan")) %>%
 index_by(Date = as_date(Time)) %>%
 summarise(Demand = sum(Demand), Temperature = max(Temperature))
```

- Plot the data and find the regression model for Demand with temperature as an explanatory variable. Why is there a positive relationship?
- Produce a residual plot. Is the model adequate? Are there any outliers or influential observations?
- Use the model to forecast the electricity demand that you would expect for the next day if the maximum temperature was 15^o and compare it with the forecast if the with maximum temperature was 35^o.
- Comment on these forecasts? Give prediction intervals for your forecasts.

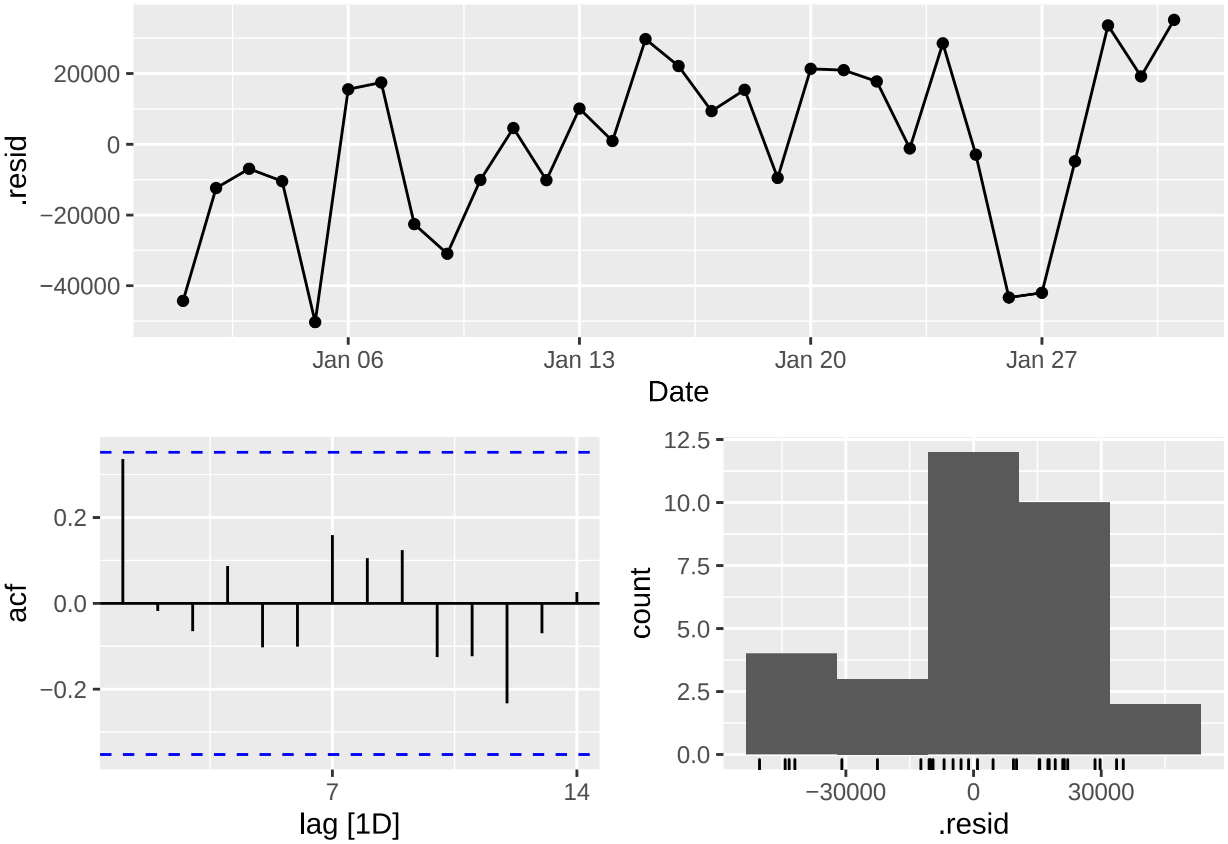
```
jan14_vic_elec %>%
  gather("Variable", "Value", Demand, Temperature) %>%
  ggplot(aes(x = Date, y = Value)) +
  geom_line() +
  facet_grid(vars(Variable), scales = "free_y")
```



```
jan14_vic_elec %>%
  ggplot(aes(x=Temperature, y=Demand)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
```



```
#> `geom_smooth()` using formula 'y ~ x'
fit <- jan14_vic_elec %>% model(TSLM(Demand ~ Temperature))
fit %>% gg_tsresiduals()
```



```
fit %>% forecast(new_data(jan14_vic_elec, 1) %>% mutate(Temperature = 15)) %>%
  mutate(interval = hilo(.distribution, 80))
```

```
## # A fable: 1 x 6 [1D]
## # Key:     .model [1]
##   .model          Date       Demand .distribution   Temperature         interval
##   <chr>           <date>      <dbl> <dist>                <dbl>           <hilo>
## 1 TSLM(Demand ~ ~ 2014-02-01 151634. N(151634, 6.8e~          15 [118255.5, 185~
```

5. The `us_gasoline` series from package `fpp3` consists of weekly data for supplies of US finished motor gasoline product, from 2 February 1991 to 20 January 2017. The units are in “million barrels per day”. Consider only the data to the end of 2004.

   - Fit a harmonic regression with trend to the data. Experiment with changing the number Fourier terms. Plot the observed gasoline and fitted values and comment on what you see.
   - Select the appropriate number of Fourier terms to include by minimising the AICc or CV value.
   - Check the residuals of the final model using the gg_tsresiduals() function. Use a Ljung-Box test to check for correlation in the residuals. Even though the residuals fail the correlation tests, the results are probably not severe enough to make much difference to the forecasts and prediction intervals. (Note that the correlations are relatively small, even though they are significant.)
   - Forecast the next year of data.
   - Plot the forecasts along with the actual data for 2005. What do you find?

6. Graduate students only:

Using matrix notation it was shown that if $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{e}$ has mean $\boldsymbol{0}$ and variance matrix $\sigma^2\boldsymbol{I}$, the estimated coefficients are given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and a forecast is given by $\hat{y} = \boldsymbol{x}^*\hat{\boldsymbol{\beta}} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ where $\boldsymbol{x}^*$ is a row vector containing the values of the regressors for the forecast (in the same format as $\boldsymbol{X}$), and the forecast variance is given by $var(\hat{y}) = \sigma^2\left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right]$.

Consider the simple time trend model where $y_t = \beta_0 + \beta_1 t$. Using the following results,

$$\sum_{t=1}^{T} t = \frac{1}{2}T(T+1), \quad \sum_{t=1}^{T} t^2 = \frac{1}{6}T(T+1)(2T+1)$$

derive the following expressions:

a. $\boldsymbol{X}'\boldsymbol{X} = \frac{1}{6}\begin{bmatrix} 6T & 3T(T+1) \\ 3T(T+1) & T(T+1)(2T+1) \end{bmatrix}$

b. $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{2}{T(T^2-1)}\begin{bmatrix} (T+1)(2T+1) & -3(T+1) \\ -3(T+1) & 6 \end{bmatrix}$

c. $\hat{\beta}_0 = \frac{2}{T(T-1)}\left[(2T+1)\sum_{t=1}^{T} y_t - 3\sum_{t=1}^{T} ty_t\right]$

$\hat{\beta}_1 = \frac{6}{T(T^2-1)}\left[2\sum_{t=1}^{T} ty_t - (T+1)\sum_{t=1}^{T} y_t\right]$

d. $\text{Var}(\hat{y}_t) = \hat{\sigma}^2\left[1 + \frac{2}{T(T-1)}\left(1 - 4T - 6h + 6\frac{(T+h)^2}{T+1}\right)\right]$