Question 1

Part A

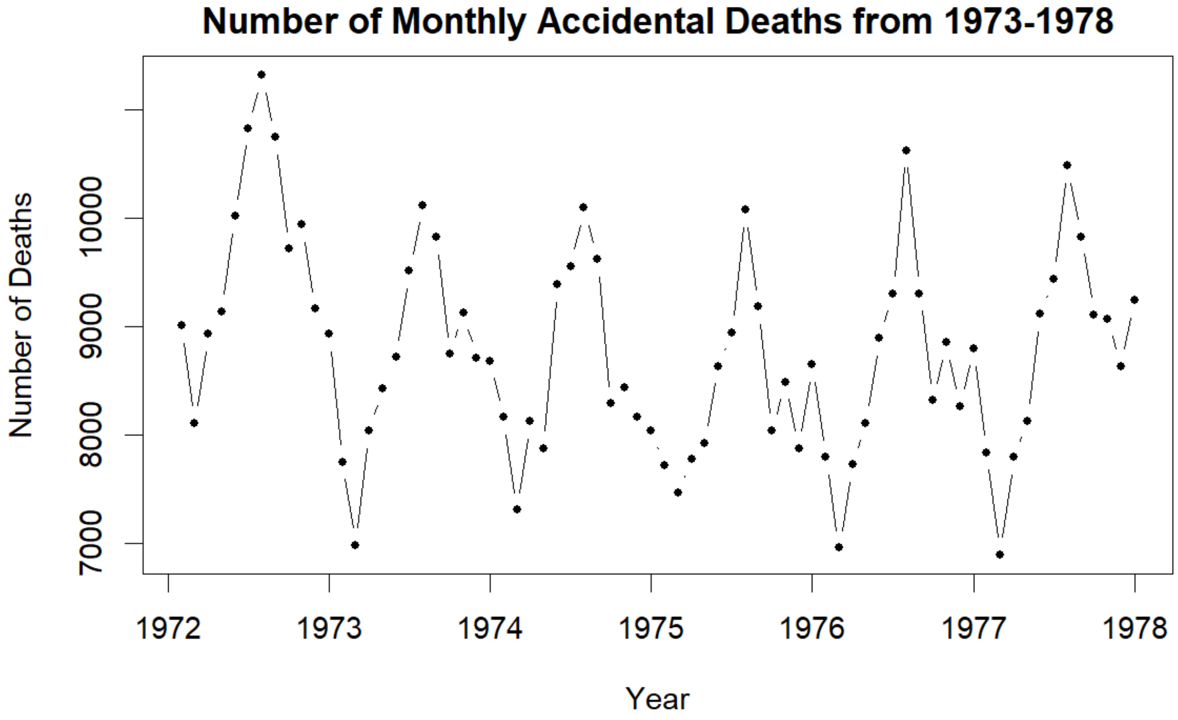
The data comes from the monthly number of accidental deaths between 1973 and 1978. Plot the time series of the data. Explain what each part of your code is doing.

plot( ) creates a plot in R
deaths$num takes the num variable from the deaths dataset and marks that number for each observation provided in the dataset
type=”b” tells R to create a plot with both lines and points
pch=20 specifies a bullet for each observation that the default empty circle point
xlab=“Year” labels the x-axis
ylab=“Number of Deaths” labels the y-axis
xaxt=“n” suppresses plotting the x=axis, as the next line of code will give more details for how the x-axis should look like
axis( ) adds an axis to the current plot
side=1 specifies that this axis should be placed below the plot
at=c(0, 12, 24, 36, 48, 60, 72) tells R at which observations the tick marks should be drawn, which in this case is every 12 months/1 year
labels=c(“1972”, “1973”, “1974”, “1975”, “1976”, “1977”, “1978”) gives R the exact labels to be added below each tick mark, which are the years the observations were taken

Part B

What do you notice about the data in terms of periodicity?

There is an obvious periodicity pattern occurring in the data because the line connecting the observations resembles a sin curve. The peak number of deaths occurs in July, with a steady decrease until the lowest number of deaths in February. The cycle repeats and follows this pattern throughout the whole timeline.



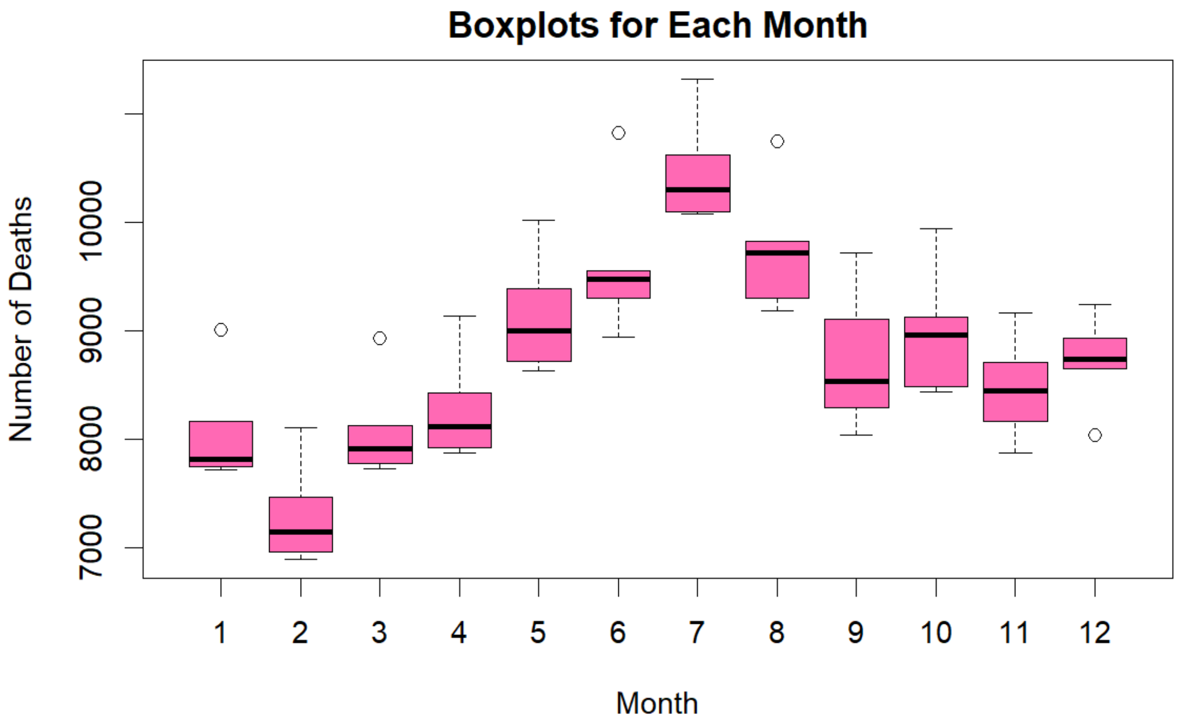
Question 2 – Investigating Seasonality

Part A

Create a boxplot showing the number of deaths by month of the year. What do your observations about the plot tell you about this time series?

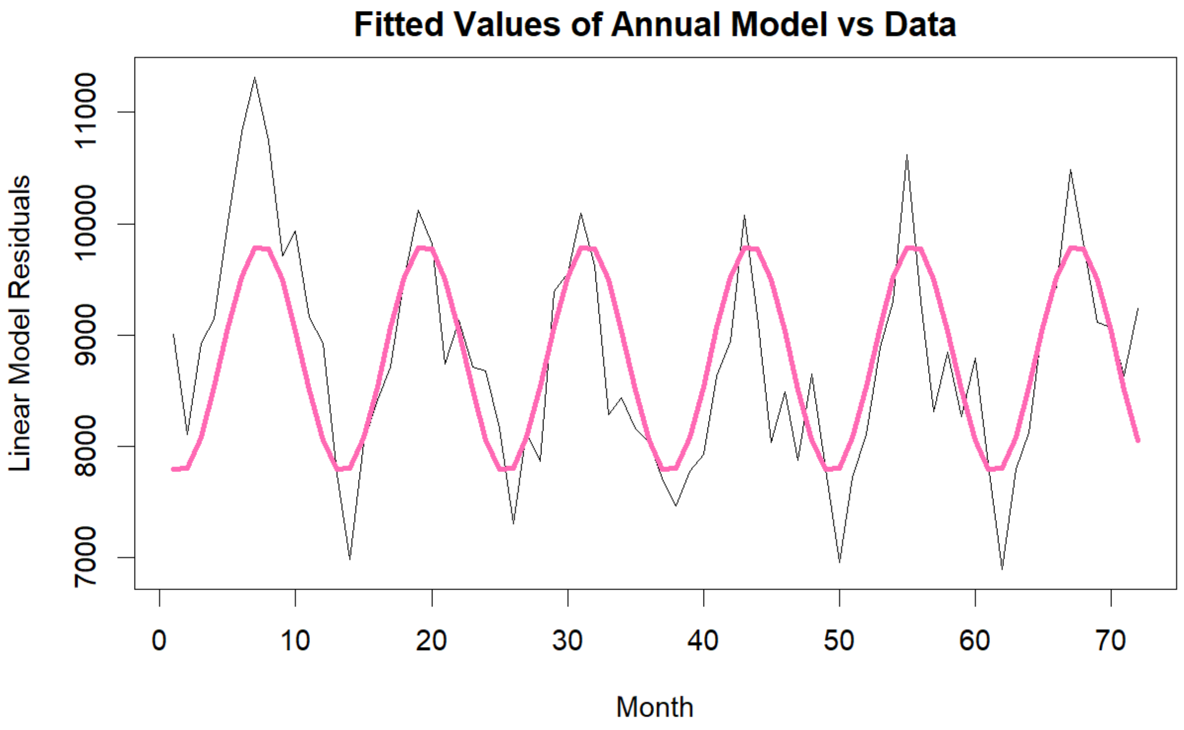
These boxplots make it very obvious that there is a seasonal pattern to the number of deaths per month. The lowest number of deaths occur in February, gradually increase to July, and then decrease again. The number of deaths in September, October, November, and December oscillate a little below 9000 deaths, so there may be another period going on there. The confidence interval from July do not overlap with any other month, so a statistical test will most likely show a significant difference between the means of these months.

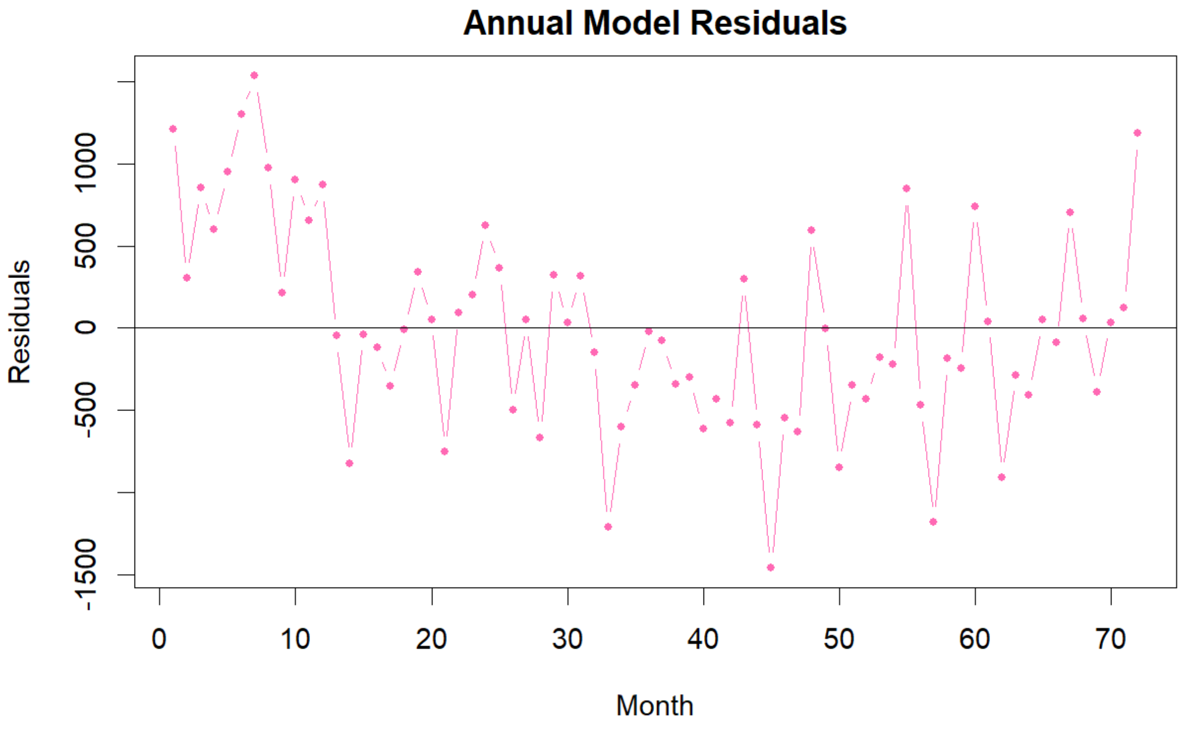
The seasonality observed in the plot above implies that this time series will not have stationarity and the residuals most likely will not have homoscedasticity and/or independence.

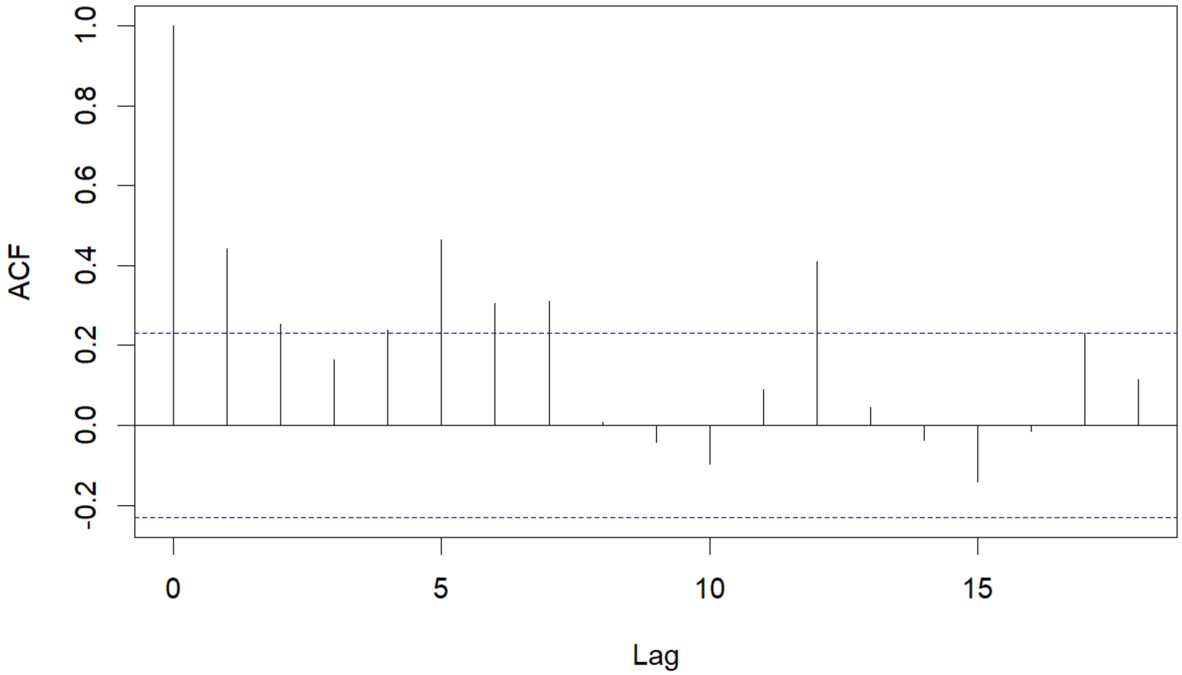


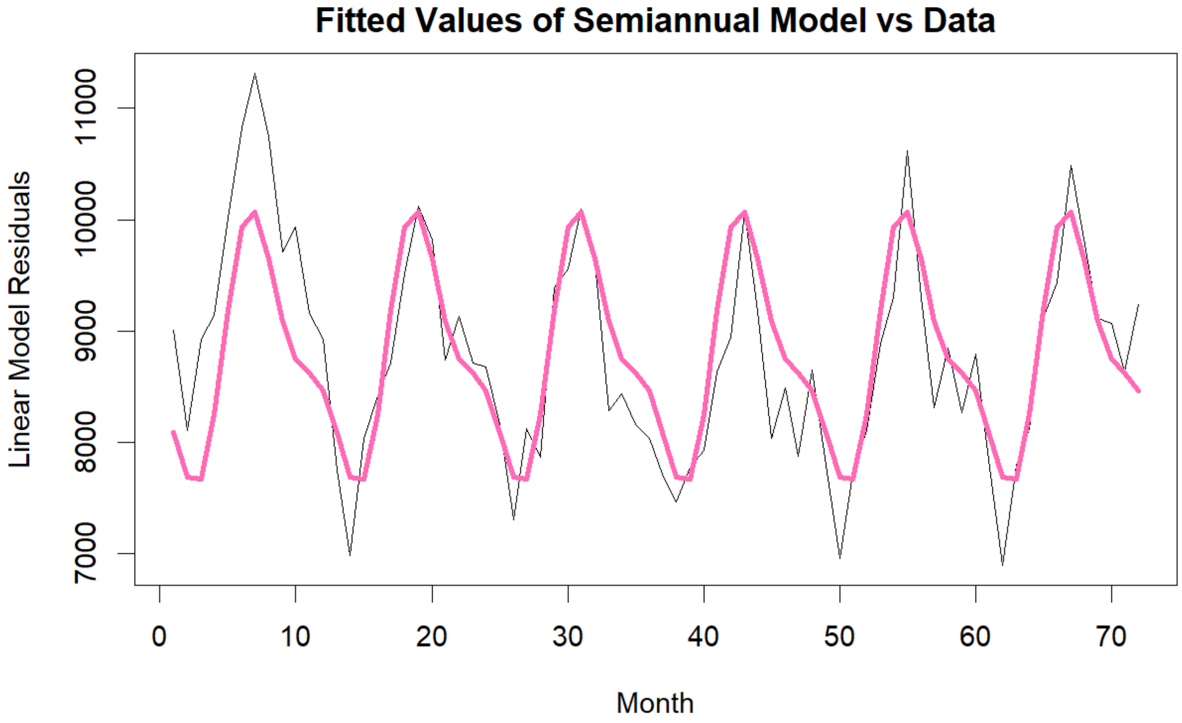
Part B

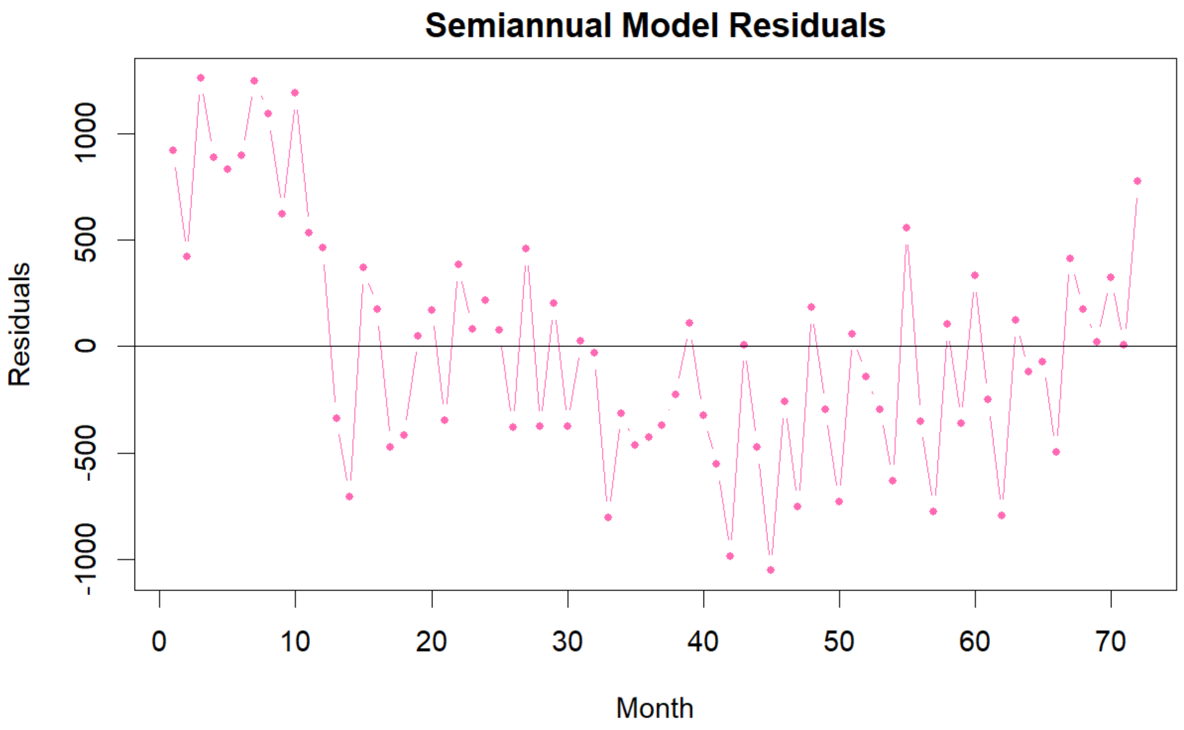
Consider a model with only an annual periodicity and a model with both an annual and semi-annual term. Plot the fitted values against the observed data. Include the residual plots and diagnostics for each model. Based on what you see, which model would you prefer and why?

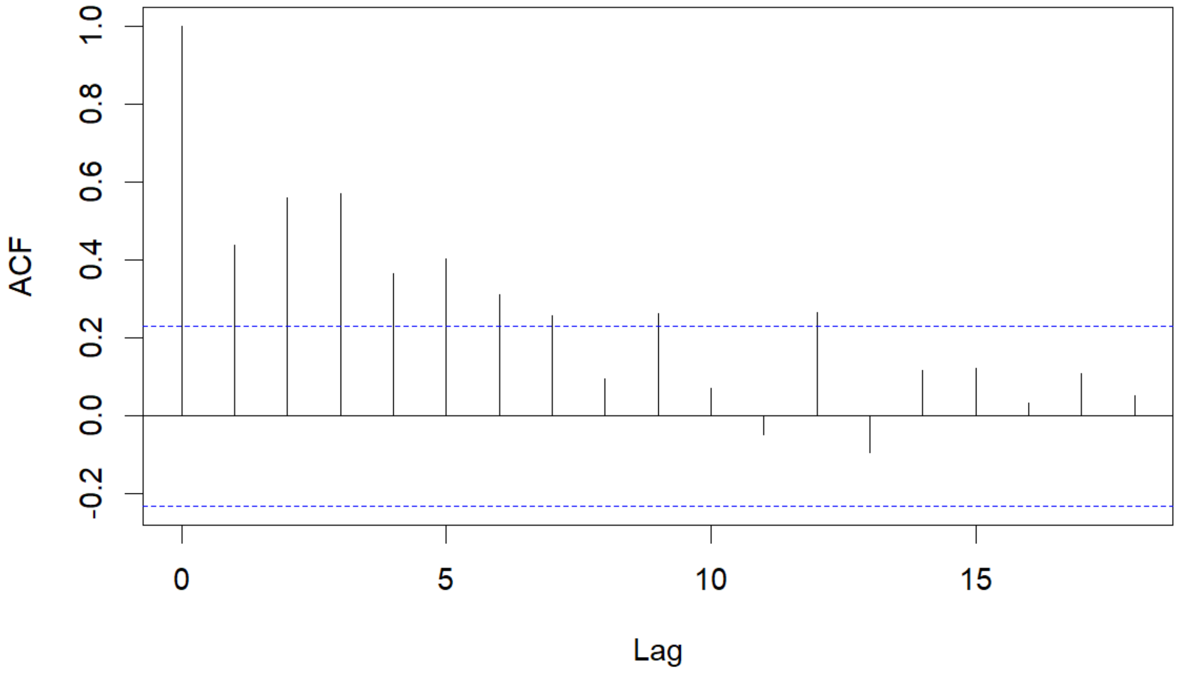





ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
Lag


Fitted Values of Semiannual Model vs Data
Linear Model Residuals
11000
10000
9000
8000
7000
0
10
20
30
40
50
60
70
Month



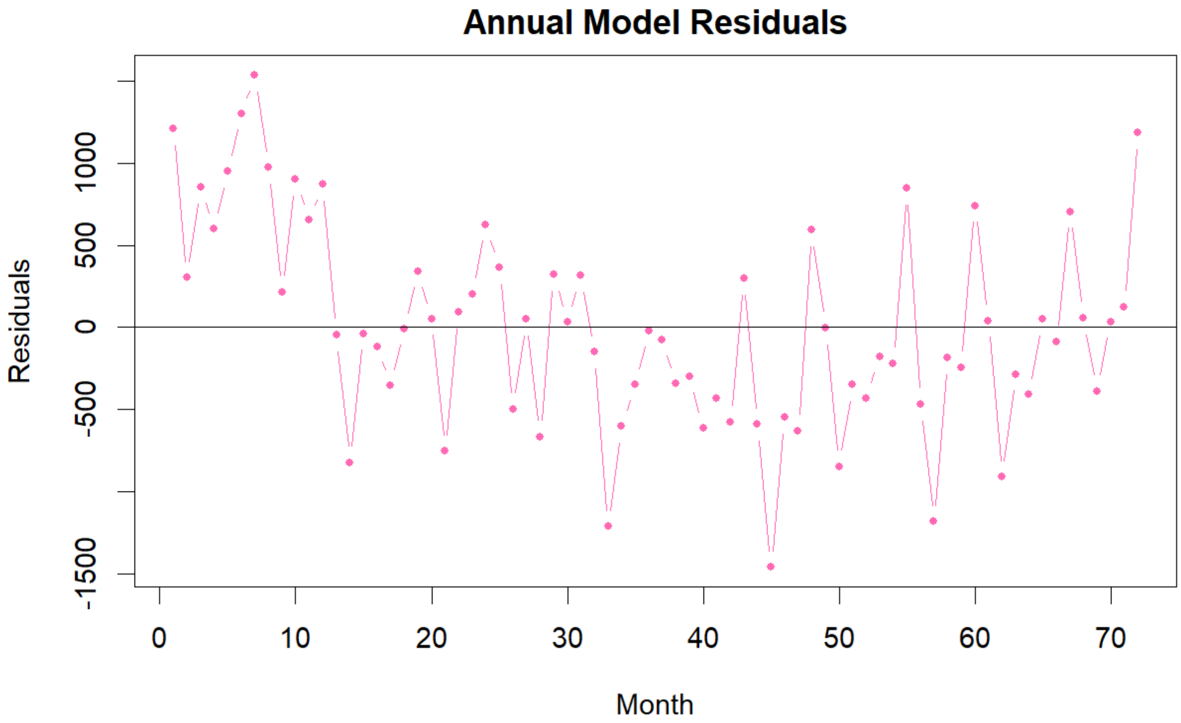


Just from looking at the fitted values of each model over the observations, the model with both the annual and the semiannual terms appear to fit the data better. The residual plots from both models show the same pattern of not being randomly distributed around 0. The ACF plot for both annual model shows a high correlation in 7 autoregressive models. An argument can be made for choosing either model, but I would go with putting both terms term in the final model.

Question 4 – Investigating Trend

Part A

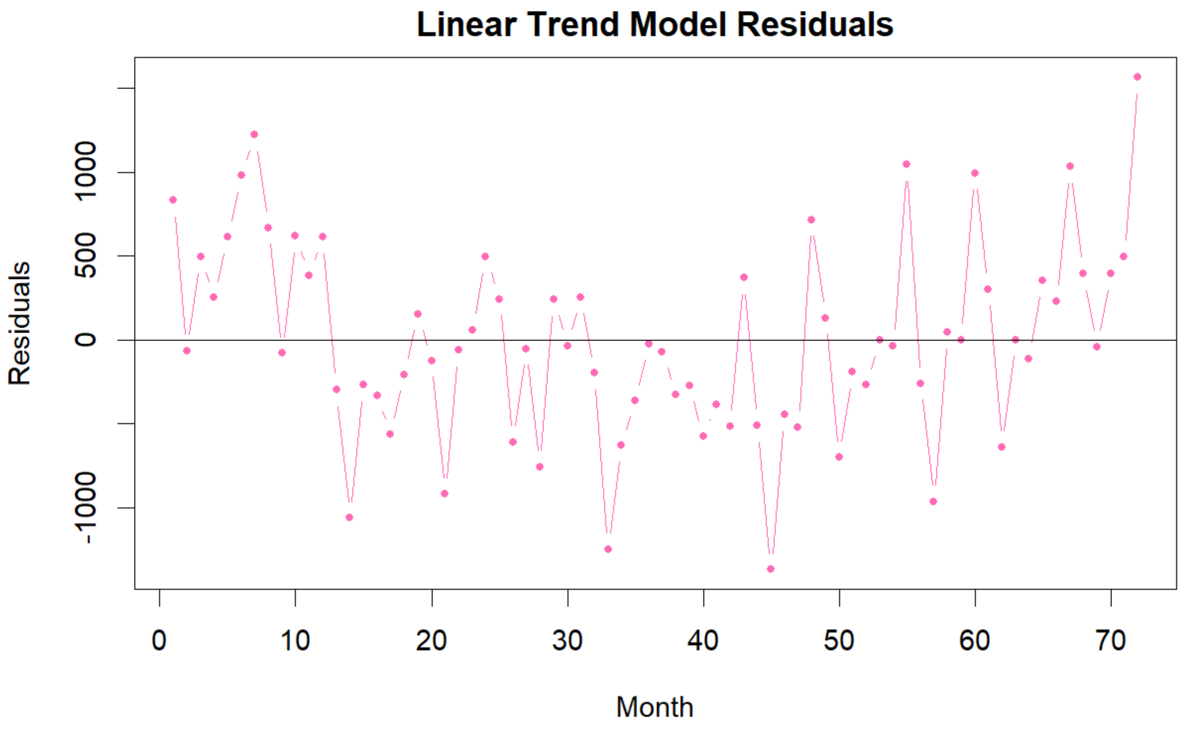
Regardless of what you decided in the previous question, proceed with the annual periodicity model. What kind of trend do you observe from the residuals?

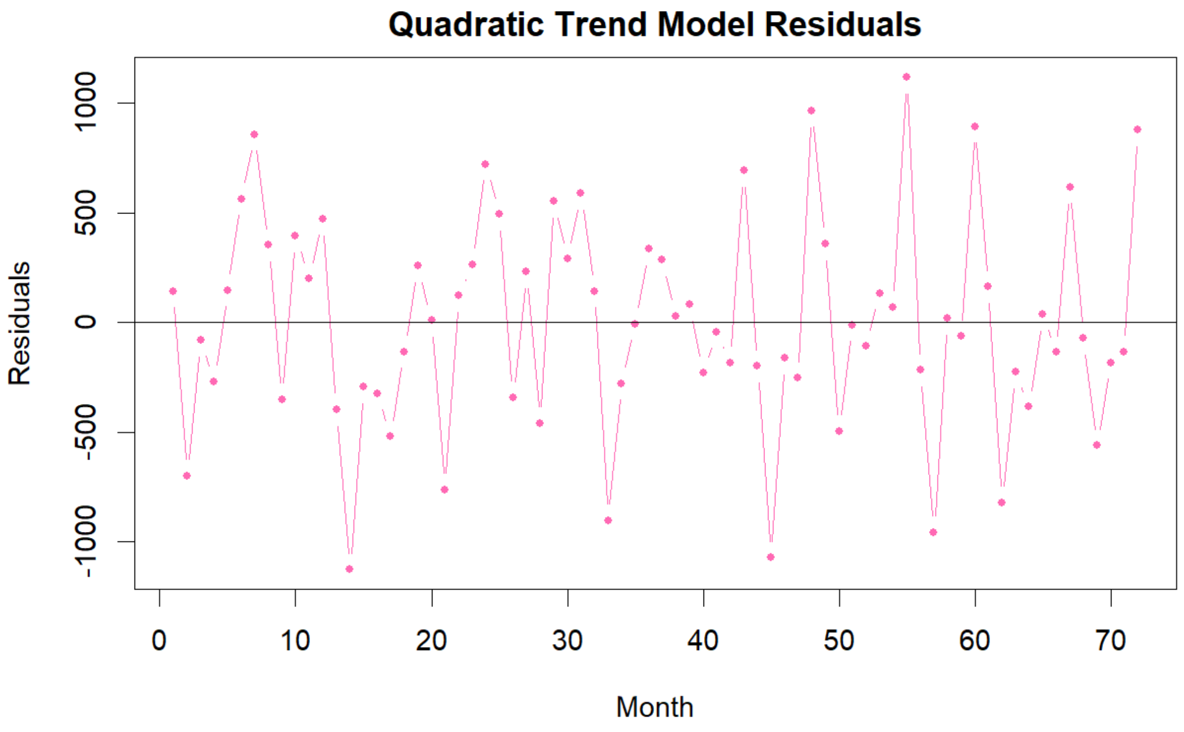


The residuals from the annual model are not randomly distributed around 0. The residuals for the first 13 observations are well above 0, which indicates the model is overestimating those fitted values. The residuals for the rest of the observations are closer to 0 but usually are in the negative range, which means that the model is underestimating those fitted values. There seems to be a systematic pattern to the individual, so the current annual model isn't doing the best job.

Part B

Model the trend in the residuals using a linear and quadratic model. From the residuals of each model, which model would you prefer?

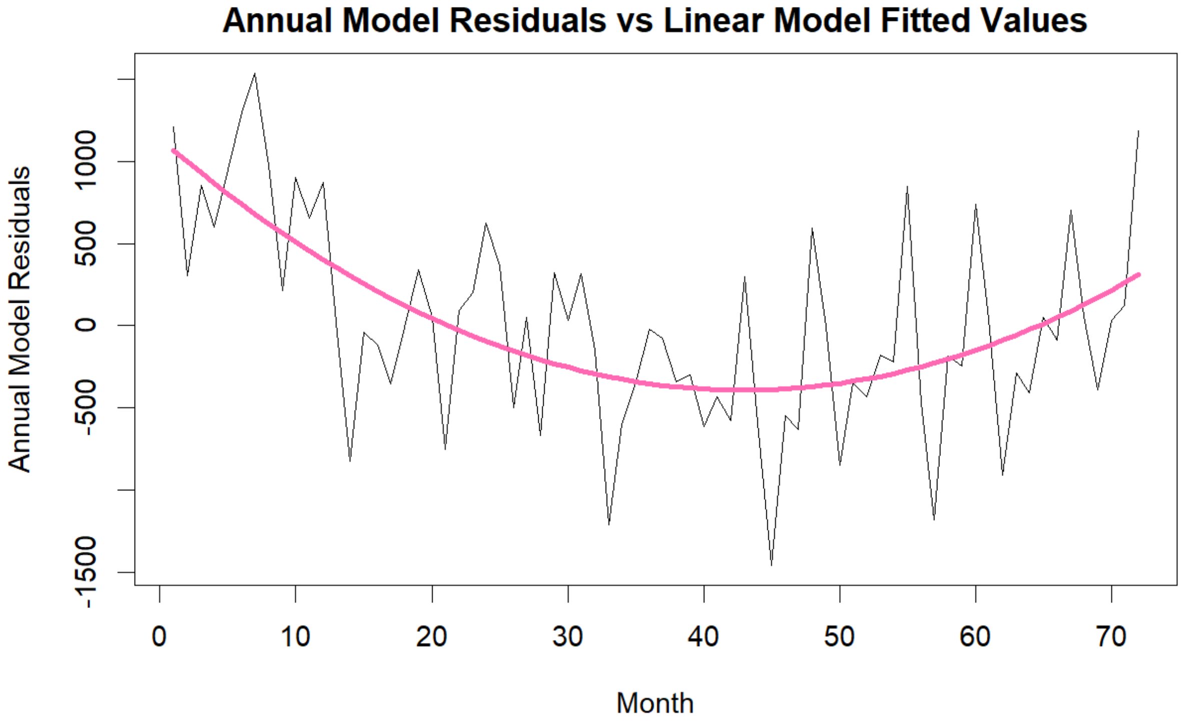




The residuals are more randomly scattered around 0 in the quadratic trend model than the linear trend model. The linear trend model's residuals are above 0 on the extremes. In the linear trend model, the p-value for the time term was 0.00193 and $R^2 = 0.1292$. In the quadratic trend model, the p-values for the time and time^2 covariates were 2.13×10^{-8} and 4.89×10^{-7}, respectively. The covariates in the quadratic trend model were much more associated with the number of deaths outcome, had a larger R^2 value of 0.3981, and a better residual plot, which indicates that the quadratic model fits the data better.

Part C

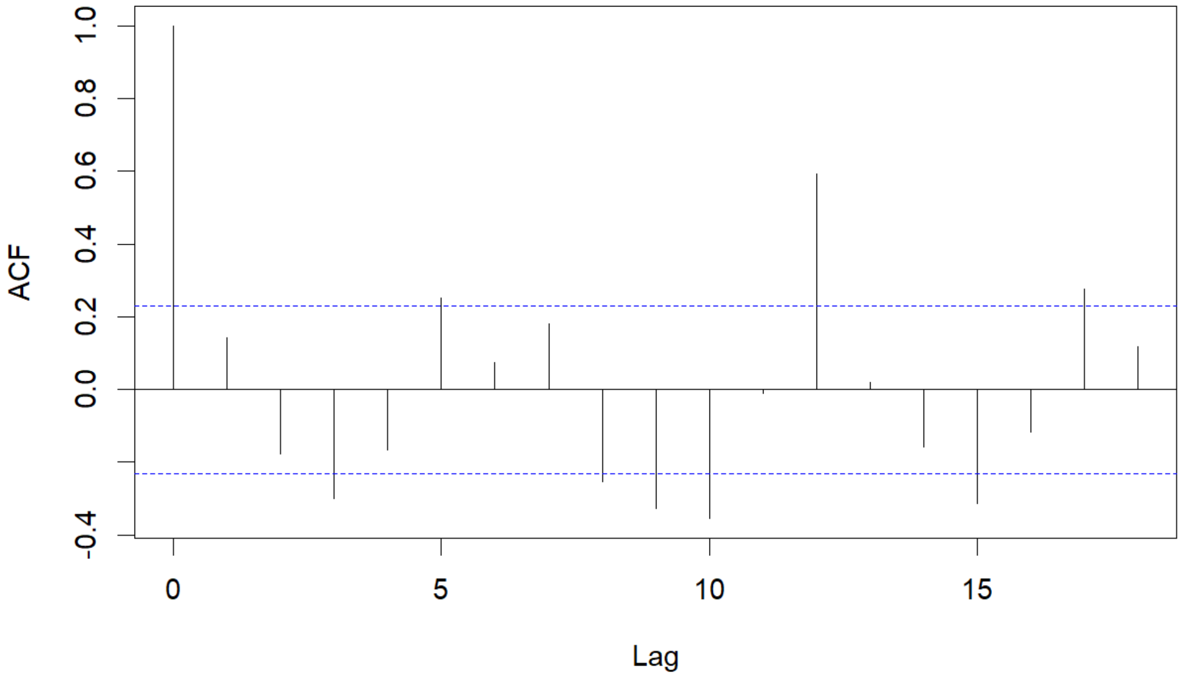
Plot the residuals from the seasonally adjusted model and overlay the predictions from your preferred trend model.



Question 5 – Investigating Autocorrelation

Part A

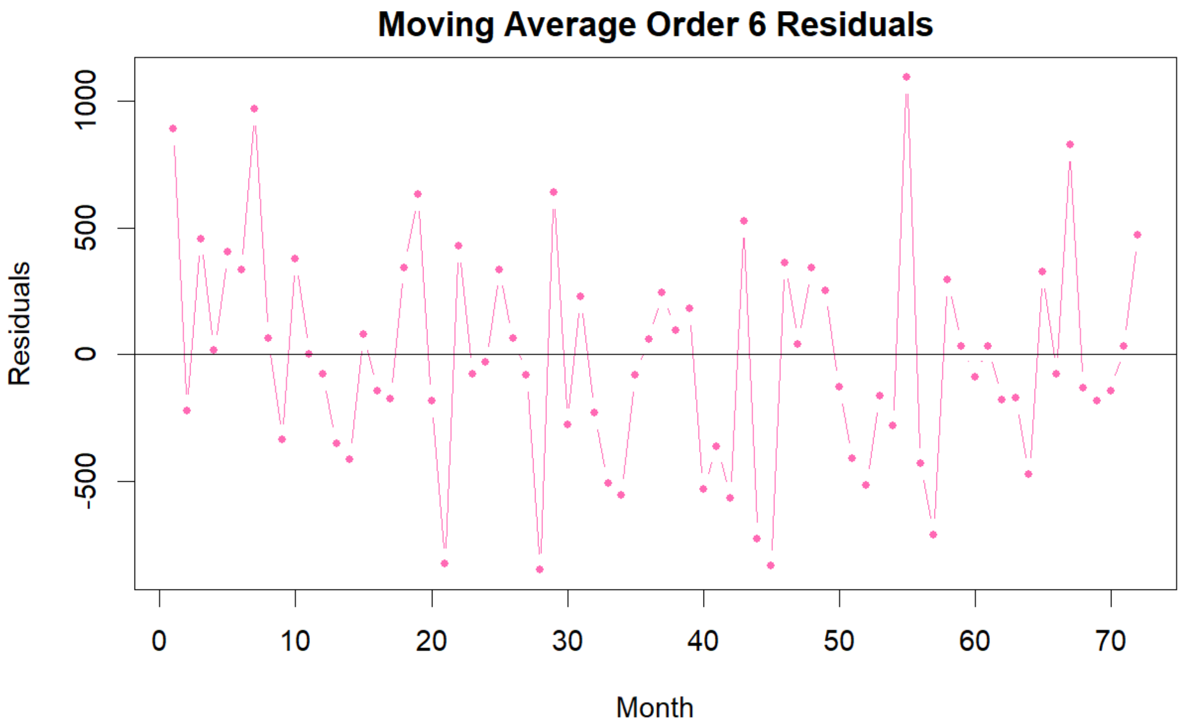
Based on what you see in the ACF, explain why a moving average model will likely be the most appropriate approach.

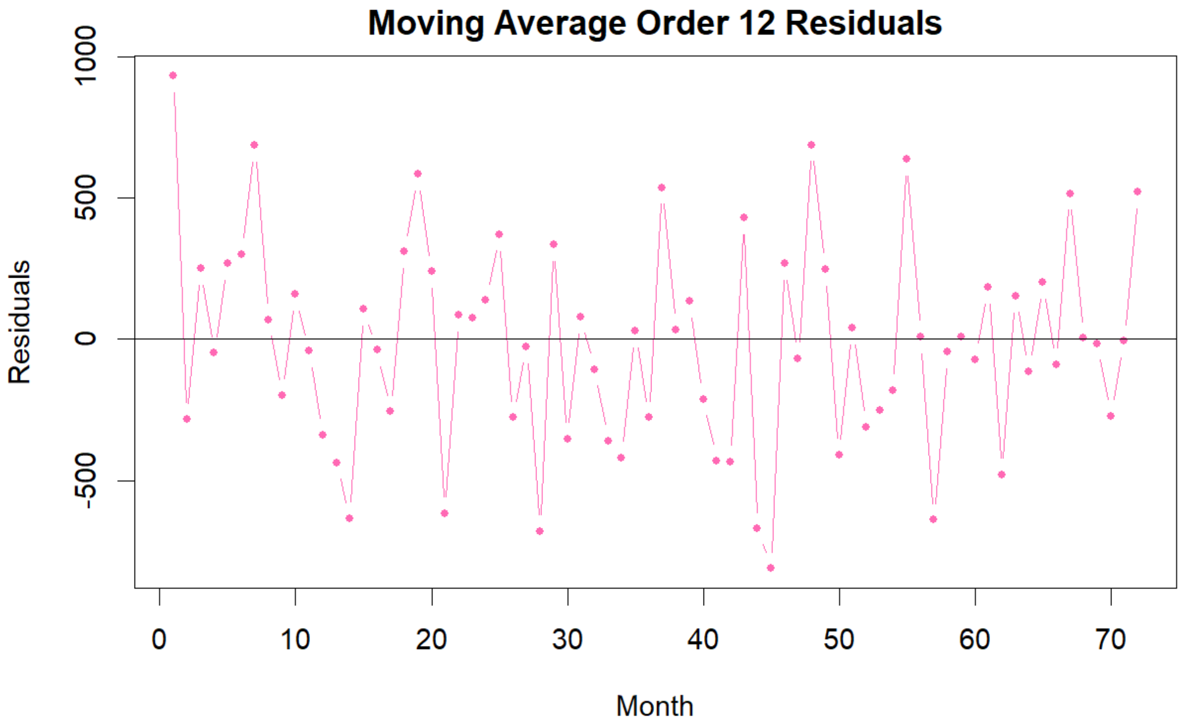


The ACF plot shows that there is one large correlation of approximately 0.6 when the lag is 12 months, which corresponds to the observed cyclical pattern in the data every year. There are a few other spikes that violate the no association threshold, but not by much. Because the ACF plot shows only 1 large spike, the rest being essentially 0, a moving average model would be the most appropriate.

Part B

Consider moving average models of order 6 and 12. Determine which model you think best fits the data and justify your choice.





The residuals for the order 12 model are slightly more randomly distributed around 0 than the order 6 model. The AIC for the order 6 model is 1104.89, while the AIC for the order 12 model is 1097.04. Although the order 12 model has more terms, some of which were not significantly associated with the number of deaths, the residual plot and smaller AIC value indicates that it is a better fit for the data.

Part C

Based on the final model you have come to, what do you learn about this data? What statements could you make about accidental deaths during the time period modeled?

Using the order 12 model, the number of deaths in 1 month is highly correlated with the number of deaths that occurred exactly 1 year ago, which makes sense because the ACF plot showed a huge spike at a lag of 12 months. That means the data is not independent and the number of deaths is highly dependent on what time of the year it is.