HW4 - Nachi Lieder 314399114

Introduction

In this report we will analyze the case distribution of the covid cases across the world. We will assess France's time series of covid cases.

Part 1

Lets look at the time series of France covid cases. Looking at the straight up number of weekly cases we can see the time series spiking around October-November of 2020 , and another small spike during April 2020.

Generally speaking , it is a difficult task to fit a model with an abnormality such as the one we see in this projection. We will attempt in any case to fit a forcasting model.

To create an autoregressive model we need to verify whether the given series is stationary. Foor this I ran a Dickey Fuller test to define how strongly the series is defined by a trend. We recieve a series with a pvalue of 0.3 meaning that we reject the null hypothesis and that the data has a unit root and is non-stationary.

from the initial view of the plot we dont exactly see any seasonality.

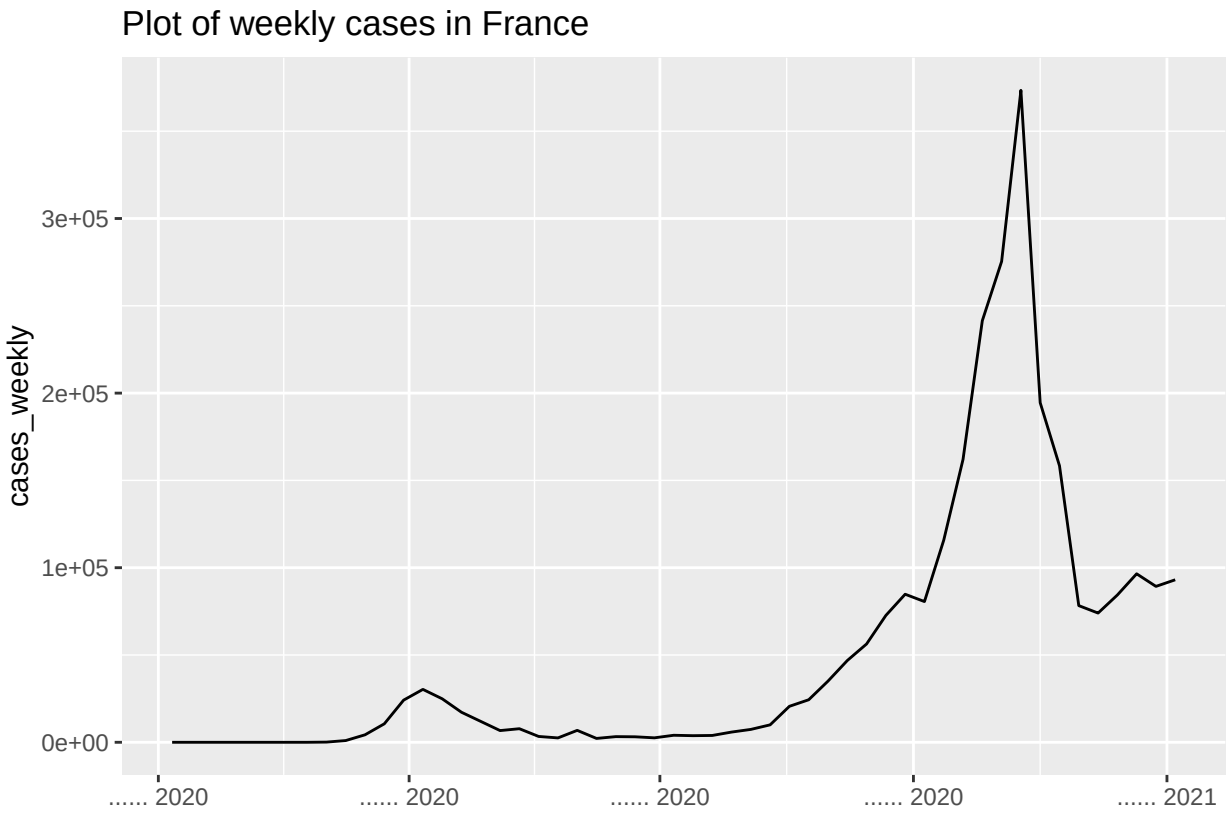
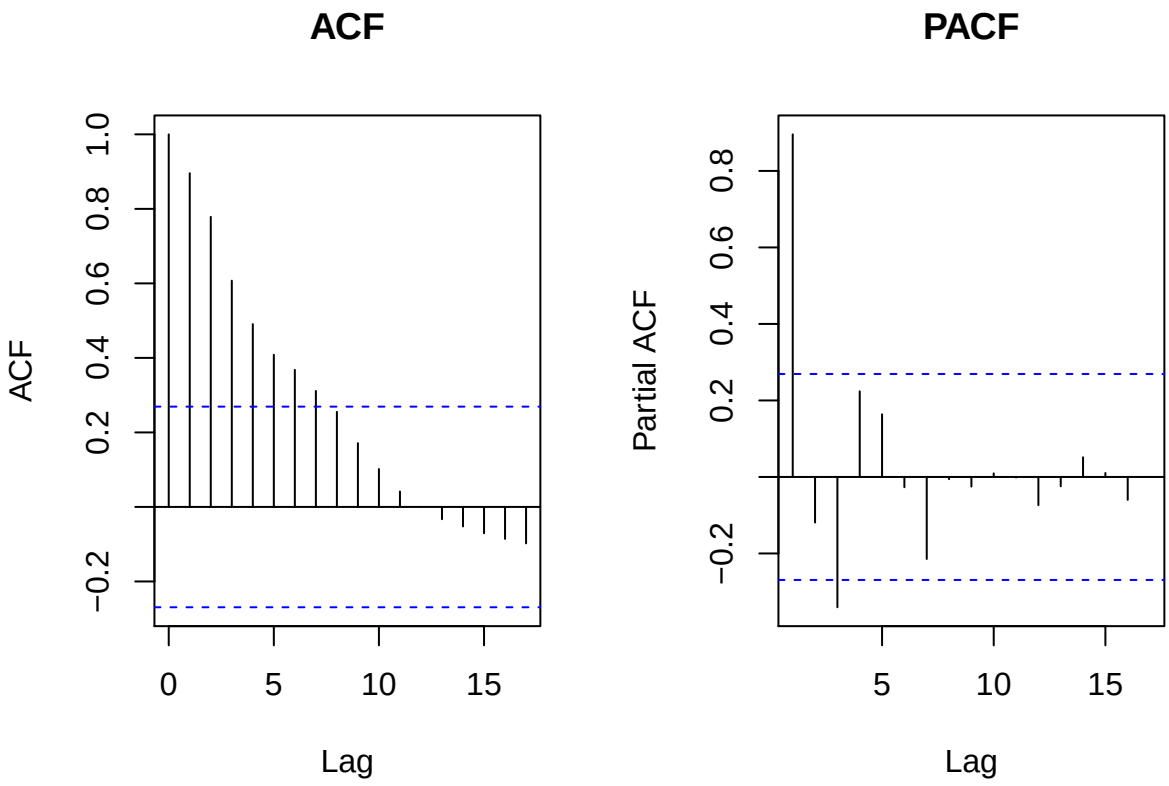


Table 1: Dickey Fuller Test

Test statistic	Lag order	P value	Alternative hypothesis
-2.674	3	0.303	stationary

Lets look at the Auto Correlation and Partial Auto Correlation plots to understand what potential lags we would like to include in our ARIMA model. From the following plots , we can see that the ACF would consider including lags of up to 7 weeks (potentially because each observation is still influencing the previous one). Though when running the auto correlation plot, we can see that after the 3rd week , we can see insignificance in terms of contribution and corellation. We can also derive that indeed there is a correlation between following observations which causes the PACF to have weaker signals and correaltions. From the PACF we will decide to reduce our lags from 7 to 3. We will attempt to fit the model of ARIMA(3,1,0).

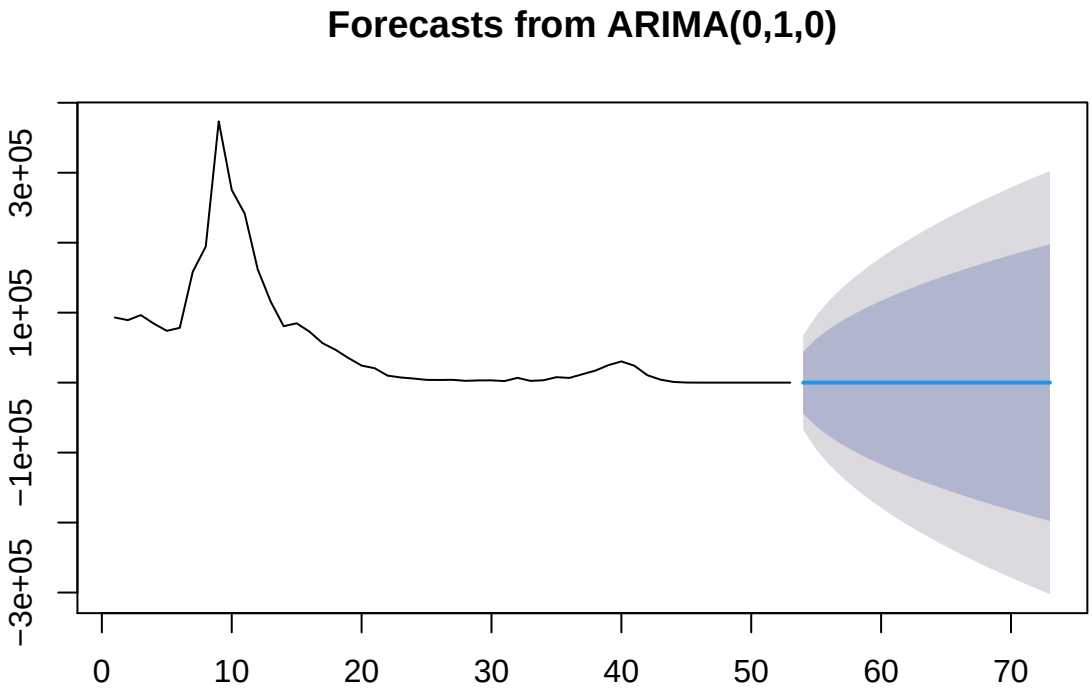


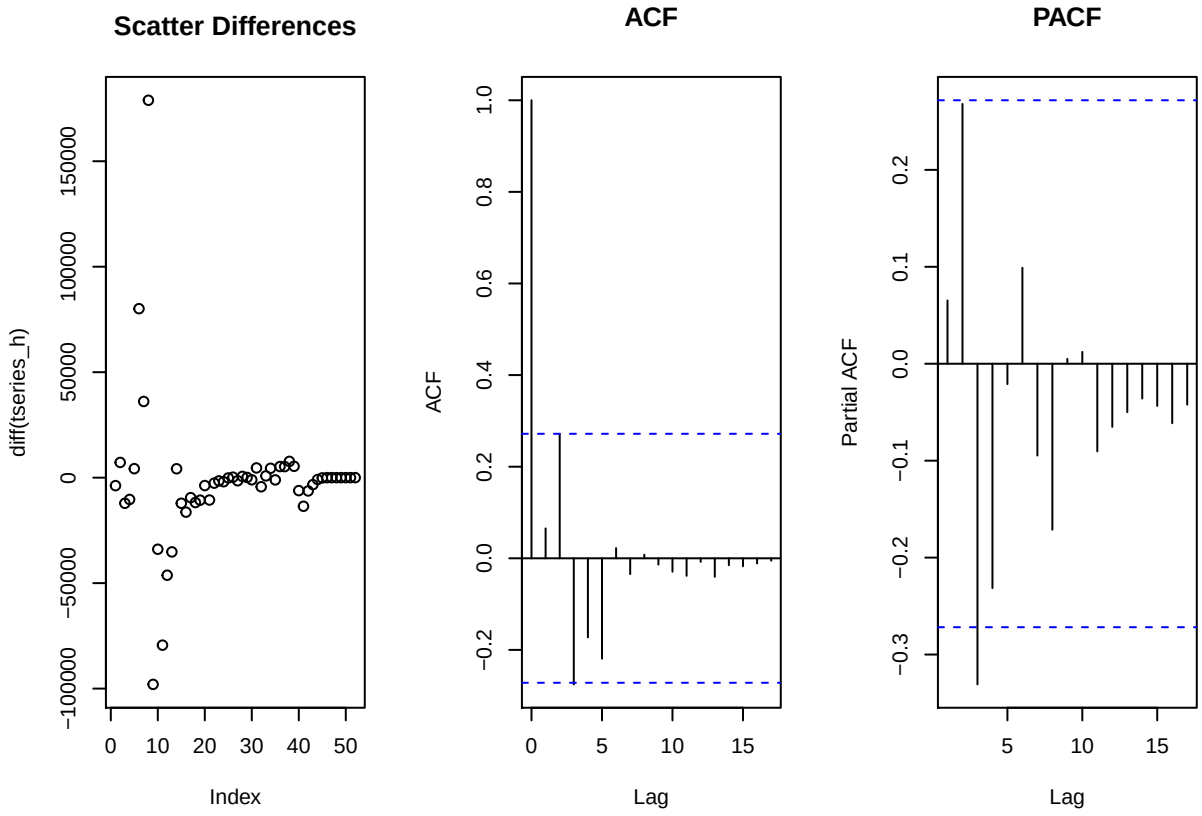
Lets try and explore the Auto ARIMA to understand the behaviour and optimized parameters according to the corrected AIC. We will explore two methods here , one using stepwise methodology and the other not. Interestingly enough the results are quite different. Lets compare the results:

Approach 1: Using Stepwise

In this approach , we can see that the optimized paramters for our ARIMA give us a model of (0,1,0) which is nothing but the differences. In addition we can see an AIC of 1236.25. Lets observe the plot of the differences of our time series, In the plot below we can see the scatter of the differences, having most of them aaround 0 , excluding the abnormal spike we observed in the former part of the exploration. While looking at the ACF and PACF , with the PACF we can see that indeed there is no auto correlation that has a strong enough coefficient. (having the 1 day lag with with a strong signal for ACF , though we will ignore this due to the weak PACF signal.)

This could be an acceptable model, though we will explore the second option as well for reference and decide from the two.


Forecasts from ARIMA(0,1,0)
−3e+05
−1e+05
1e+05
3e+05
0
10
20
30
40
50
60
70



Lets observe the second approach:

Approach 2: Not Stepwise

Here we can see that the optinal model is quite different. The model we receive is and ARIMA of (0,1,2). Looking at the AIC we see 1231 which is already in improvement. In addition , lets look at the ma_1 and ma_2 coefficients, and we can see that the first coeffieicnet is 0.14 and the second 0.51.

While observing the residuals of our ARIMA model , we can a normal behavior , and whileobserving our ACF we see no auto correlation effect.

ma1	ma2
0.1425	0.5168

Conclusion

To conclude this part of the analysis, we will agree to advance and utilize the ARIMA model in the form of (0,1,2). This is based on our improved model defined by the AIC reduction.

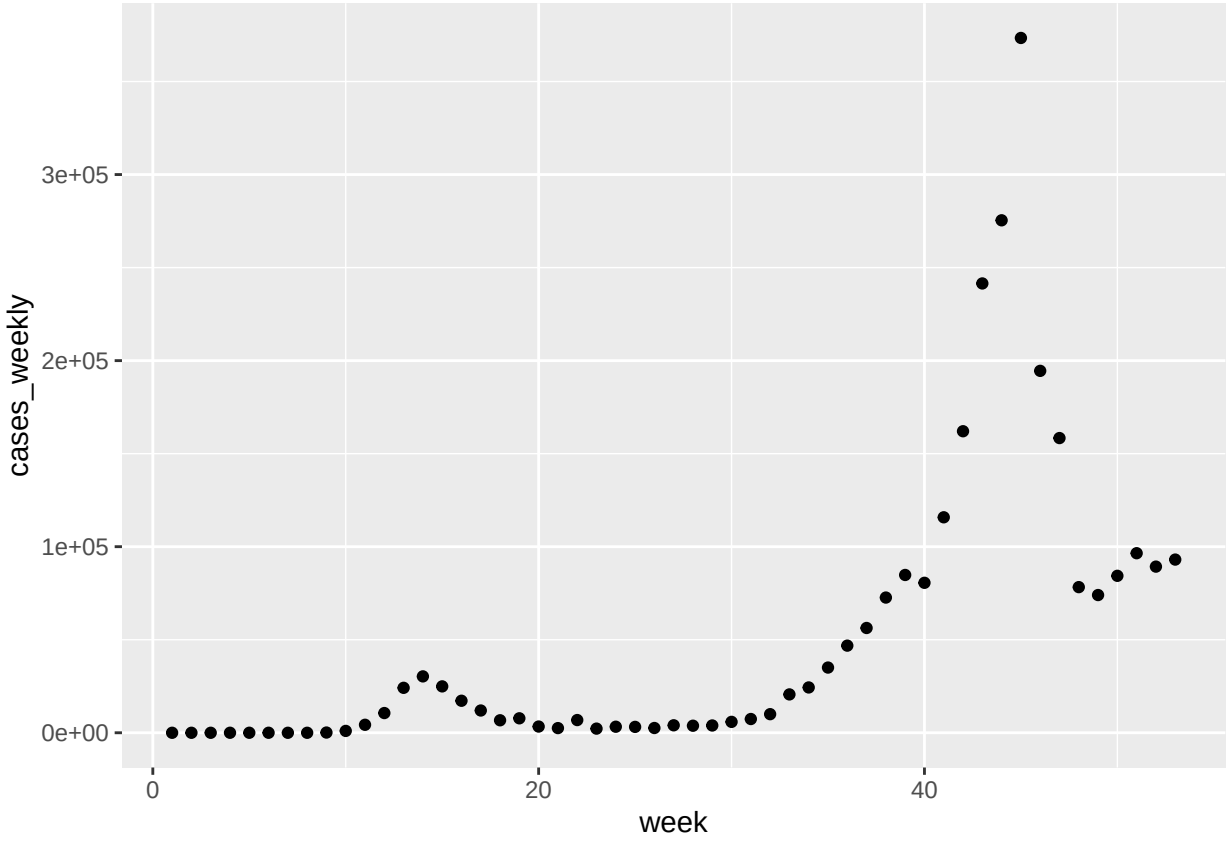
Part 2

Here we will attempt to fit a spline smoother to the data of the french time series.

Lets look at first at the basis functions . We will test several types of basis functions. We will asses 3 different functions and base our fit by viewing the plots fit.

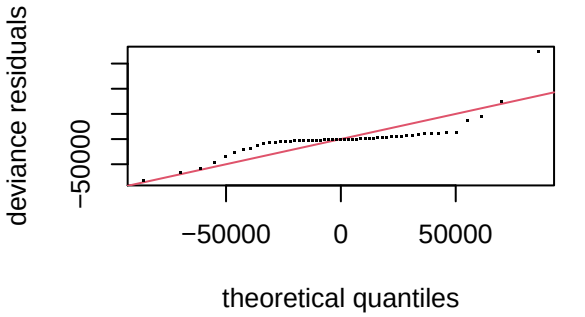
Below we can see the three plots, we can see that the model of cases_weekly ~ ti(week) has a fairly weak fit and the smooth is too strong. The model of cases_weekly ~ s(week) gives us a much beter fit.

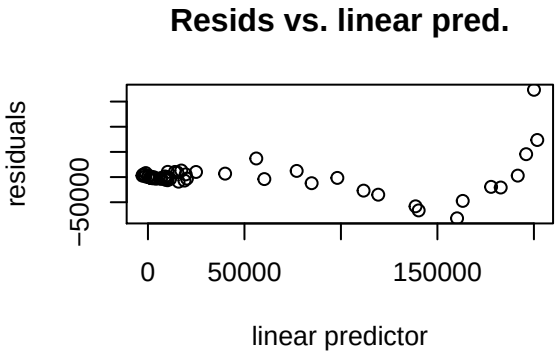
Also , when combining the basis functions together we can see a strong smooth terms for the thin plate spline.

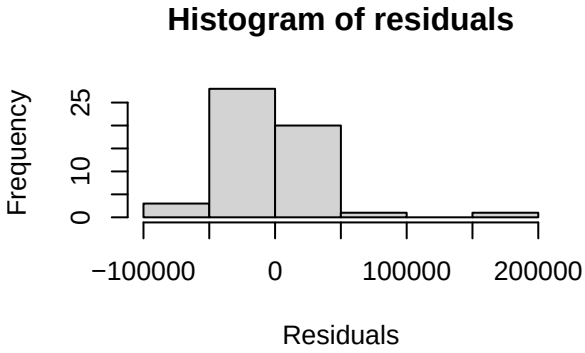


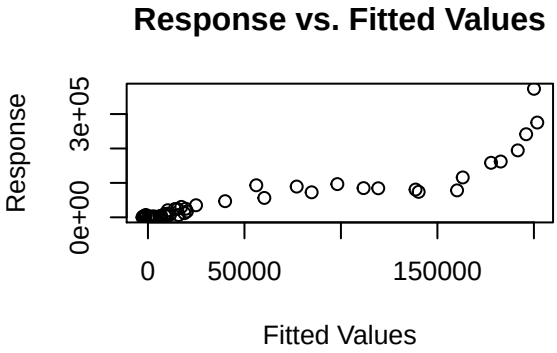
Lets first look at the model of $cases - weekly \sim s(week)$ First off the histogram of residuals presents a gausian behavior , though the residuals vs linear predictions is not exactly as we hoped. as the predictor increases we see a diviance in the residuals. This will be considered a benchmark towards the other models that will be evaluated below.

In addition , we can see a fairly nice fit with a slight generalization towards the outlier peak during weeks 40-50.



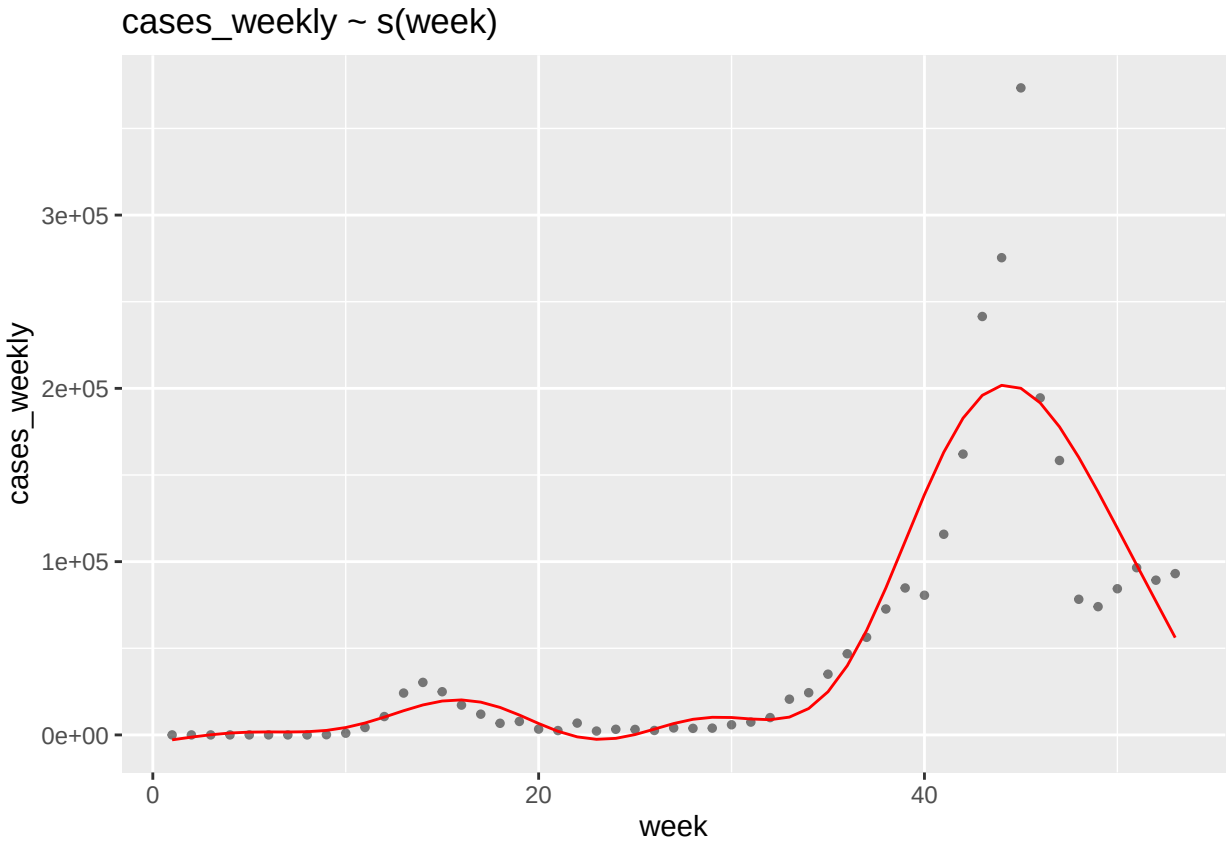







```
##
## Method: REML   Optimizer: outer newton
## full convergence after 6 iterations.
## Gradient range [-8.890937e-05,-3.239133e-06]
## (score 622.9536 & scale 1339850678).
## Hessian positive definite, eigenvalue range [1.760196,25.92942].
## Model rank =  10 / 10
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##           k'  edf k-index p-value
## s(week) 9.00 7.39    0.46   0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_weekly ~ s(week)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    50108       5028   9.966 6.36e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(week) 7.385  8.364 22.19  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.781   Deviance explained = 81.2%
## -REML = 622.95  Scale est. = 1.3399e+09  n = 53
```



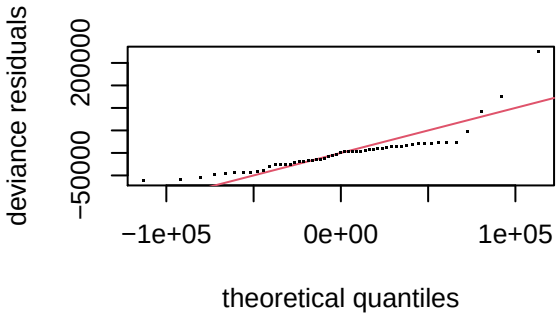
Lets evaluate the following model formats and comapre the results.

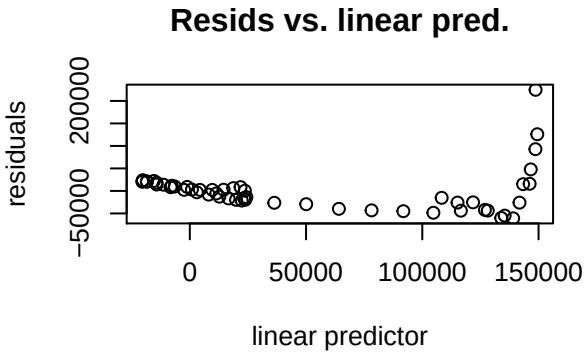
- Case b: cases_weekly ~ ti(week)
- Case c: cases_weekly ~ ti(week) +s(week)
- Case d: cases_weekly ~ te(week)

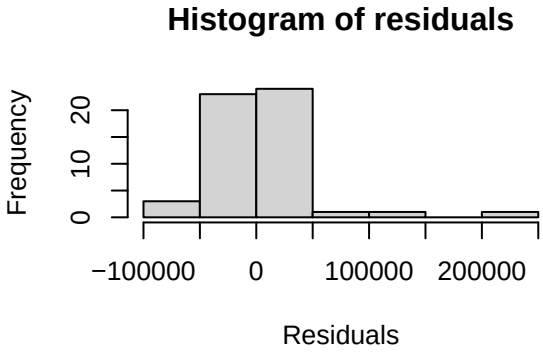
Under case B we see strong results coming frm a very low p value, significantly high F stat and a high R squared adj. of 0.62. When looking at the plot of predictions we can see that the fit doesnt seem as stong and well fit as the Case A , and is much more smooth.

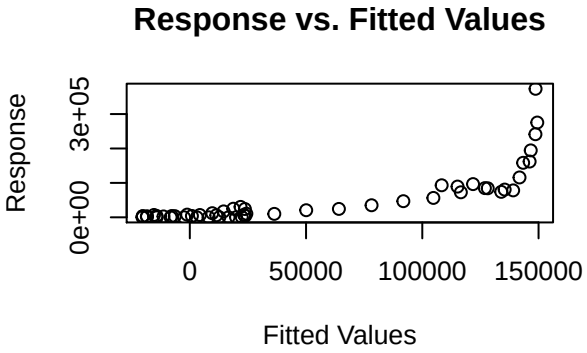
With Case C we can see even stronger results , in particular with the R squared. improving from 0.62 to 0.78. The fit is a lot less smooth and well fit. This is a strong candidate for the selected model.

Model D is similar to model B giving similar results.

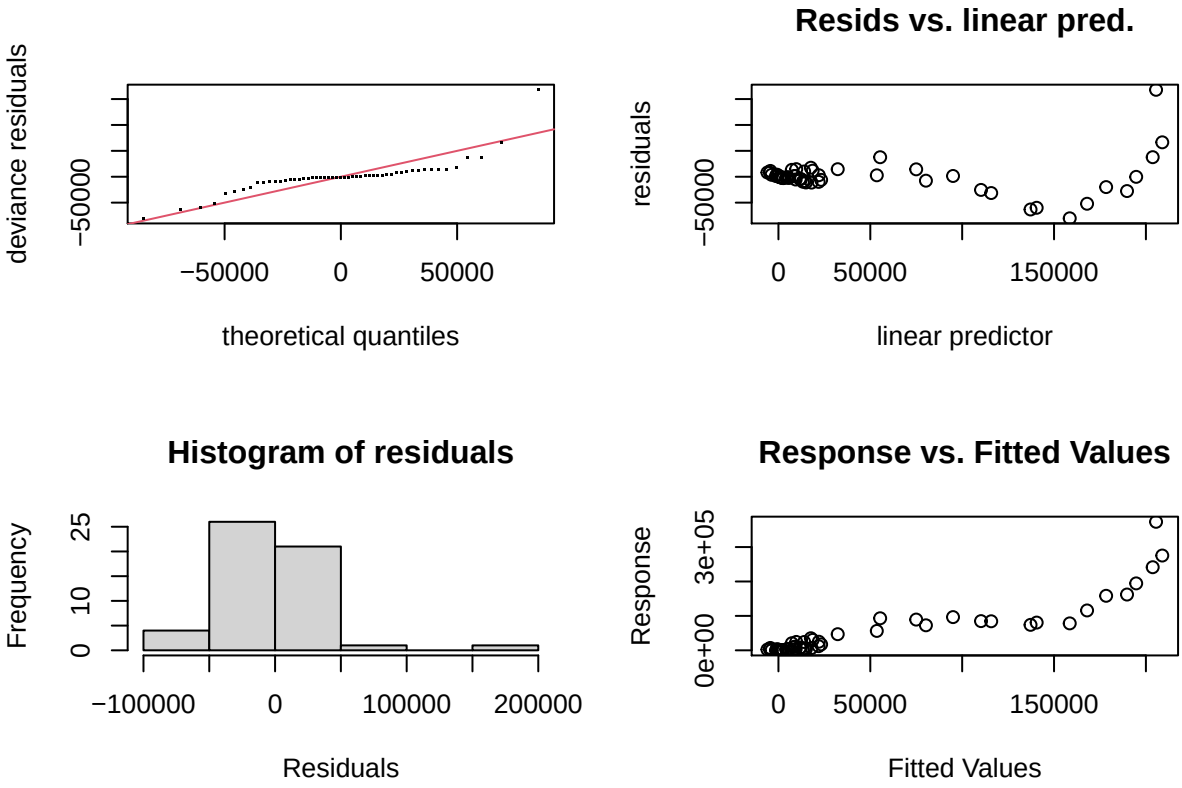




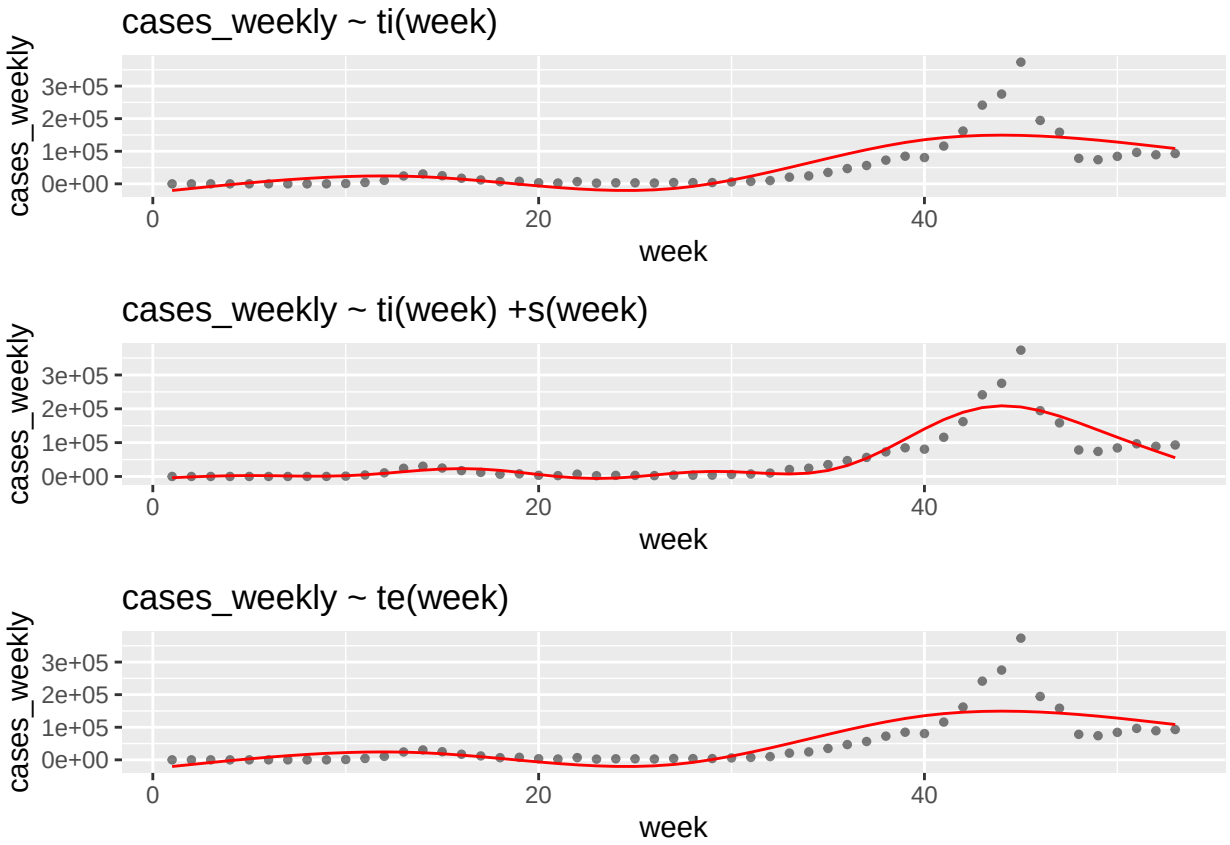





```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 7 iterations.
## The RMS GCV score gradient at convergence was 788.1816 .
## The Hessian was positive definite.
## Model rank =  5 / 5
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##            k'  edf k-index p-value
## ti(week) 4.00 3.91    0.27  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 14 iterations.
## The RMS GCV score gradient at convergence was 136.0888 .
## The Hessian was positive definite.
## Model rank =  13 / 13
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##           k' edf k-index p-value
## ti(week) 4.0 1.0    0.48  <2e-16 ***
## s(week)  8.0 7.1    0.48  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



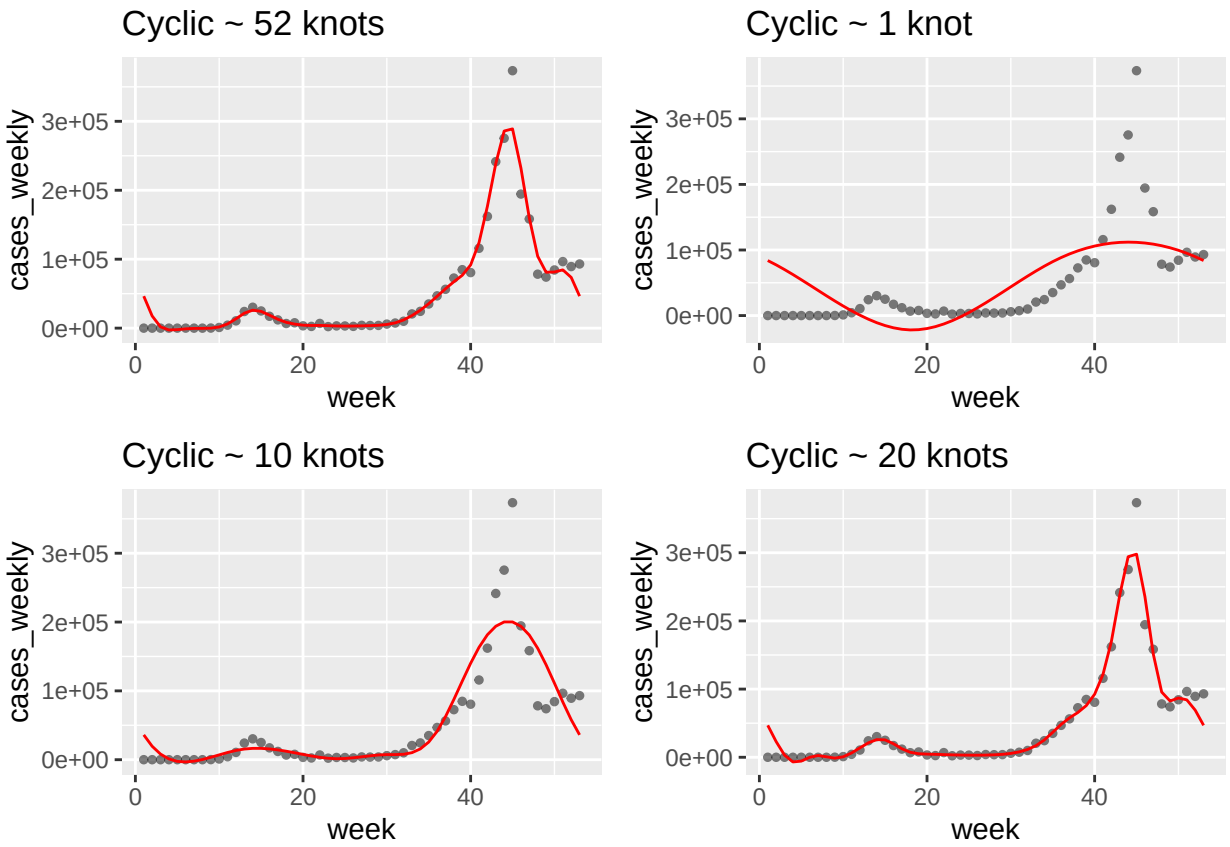
Lets review models with the change of the basis function. We will change the amount of knots and assess the plots.

Lets asses the change in the parameter of the amount of knots ranging (52,1,10,20).

- Model E: Model with 52 knots
- Model F: Model with 1 knots
- Model G: Model with 10 knots
- Model H: Model with 20 knots

Looking at the fit for 52 knots we can see a very good fit, though this may be understandable since the entire cycle of our data is 52 weeks. We will keep this in mind when fully evaluating all the models together.

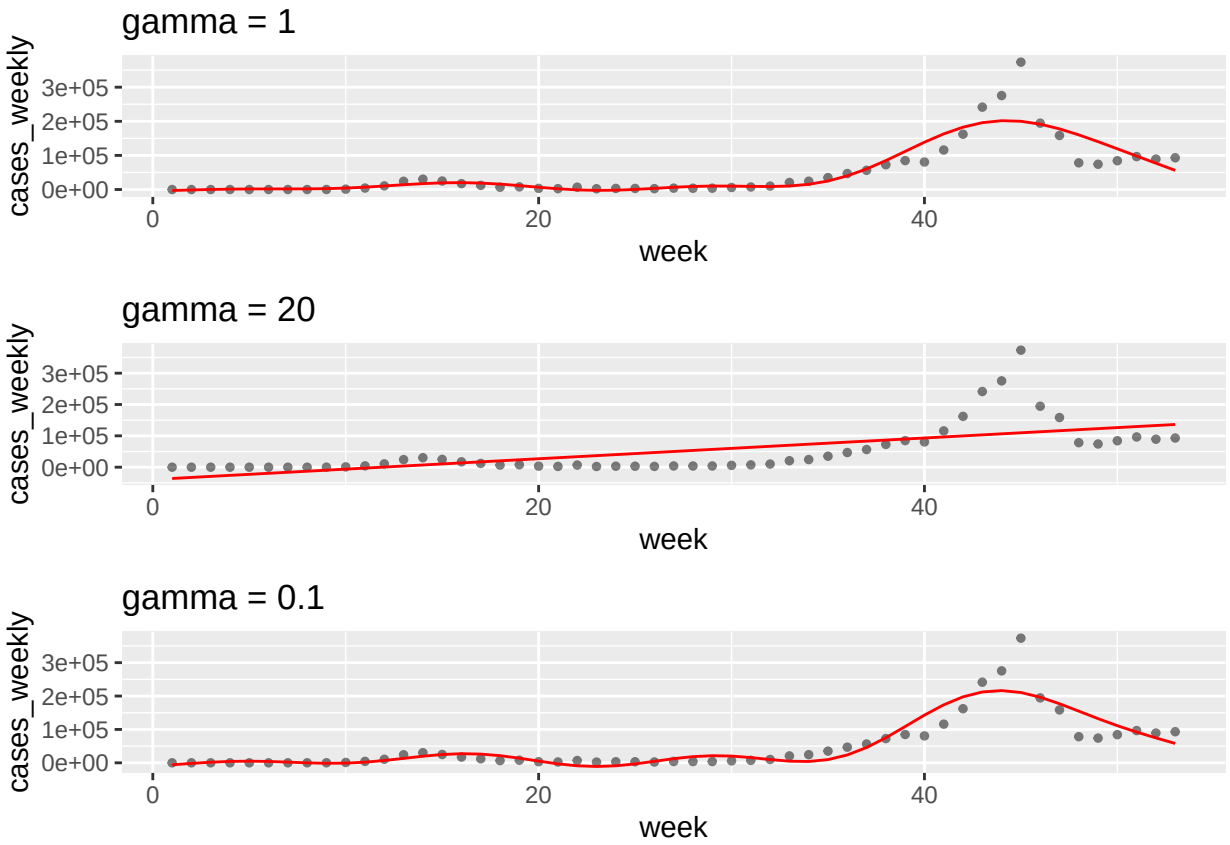
As expected the single knot gives the smoothest plot which is somewhat irrelevant to our analysis. What is interesting is that 20 and 52 knots give similar plots. Eventually we will dive into this to understand why this may occur.



Now lets observe the changes in the smotthing parameter. We will be adjusting our gamma parameter to the following models:

- Model I: Gamma = 1
- Model J: Gamma = 20
- Model K: Gamma = 0.1

Here when viewing the plot , we dont see any model which may be a solid contender to our model selection.



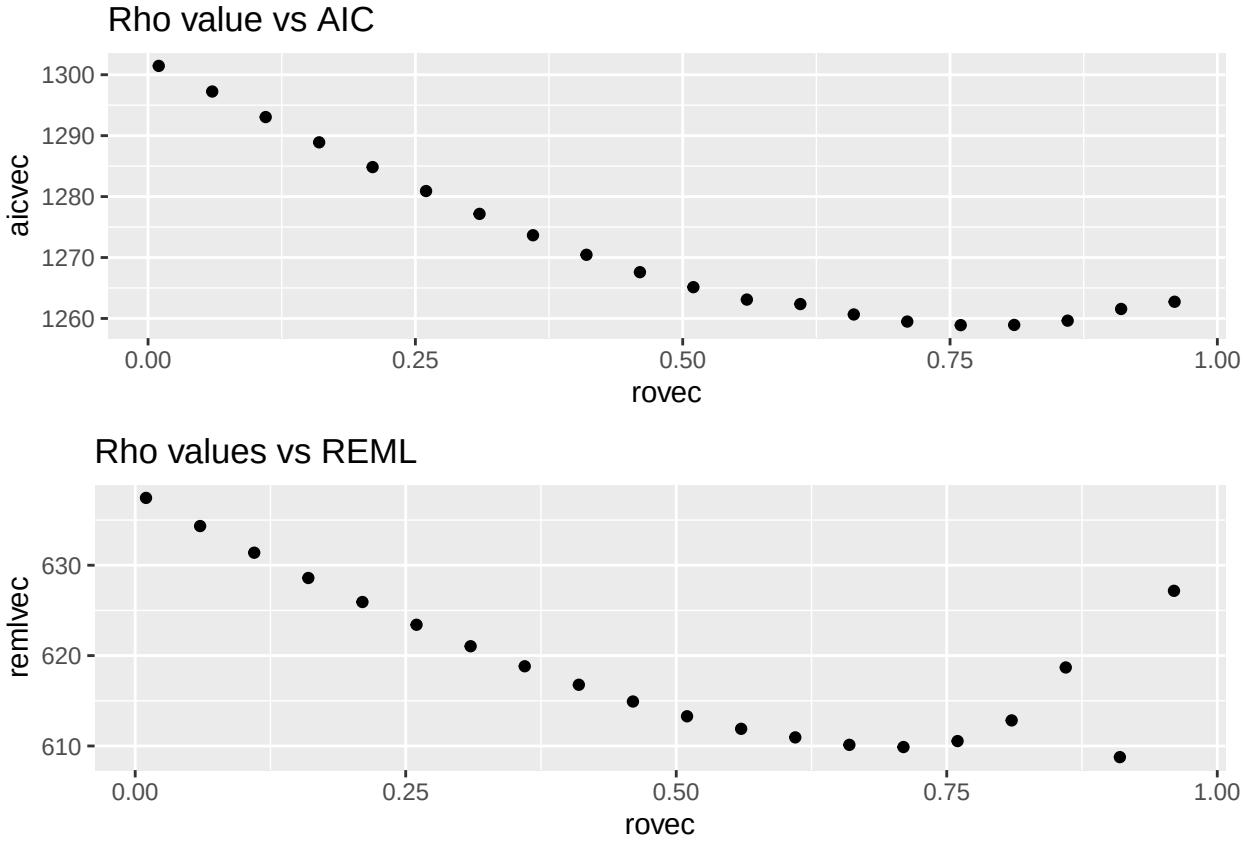
Lets present the model comparison in terms of AIC. We can see that model which is the model of 20 knots with a cyclic smoother presents the best fit with the lowest AIC of 1253.

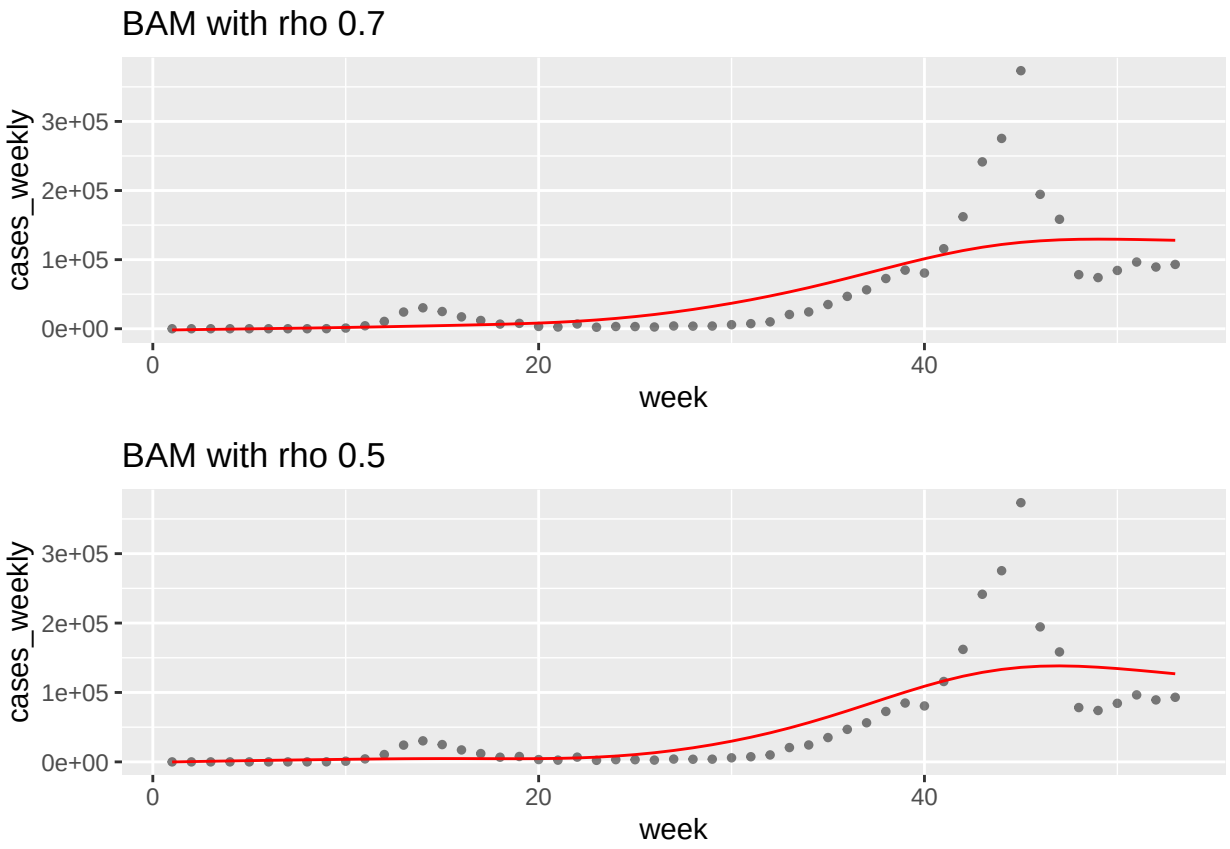
```
##            Model df      AIC      BIC    logLik   Test      L.Ratio p-value
## model1$lme     1  4 1309.679 1317.561 -650.8397
## model2$lme     2  4 1297.288 1305.169 -644.6441
## model3$lme     3  5 1299.288 1309.140 -644.6441 2 vs 3 1.552239e-07  0.9997

##            Model df      AIC      BIC    logLik
## model4$lme     1  3 1256.254 1262.108 -625.1270
## model5$lme     2  3 1309.069 1314.922 -651.5343
## model6$lme     3  3 1274.716 1280.570 -634.3582
## model7$lme     4  3 1253.097 1258.950 -623.5484

##            Model df      AIC      BIC    logLik   Test L.Ratio p-value
## model7$lme     1  3 1253.097 1258.950 -623.5484
## model8$lme     2  4 1253.907 1261.634 -622.9536 1 vs 2 1.18954  0.2754
## model9$lme     3  4 1253.907 1261.634 -622.9536
```

Now let us compare the prior results vs the models based on BAM models. The following analysis reviews the models with different values oh ρ evaulating the AIC. What may interest us from the plot below is looking at the ρ with the values 0.5 and 0.7 since there we can see a significant change in the relation between ρ and REML.

When reviewing the plot, it seems like the optimal fit from the GAM model is a better fit.Though on the other hand , we know that for these values of ρ we get a much more improved model in terms of AIC.

Generally speaking I would reccomend to further investigate both the optimal from the GAM and model with $\rho = 0.7$ as primary candidates for our model selection.





Part 3 - Varying coefficient model

Lets look at the information regarding continents , and observe models that review weekly cases per continent in a scaled form of per capita 1:100,000. The tests we will look at will involve Asia and Europe.

We will approach this in two ways:

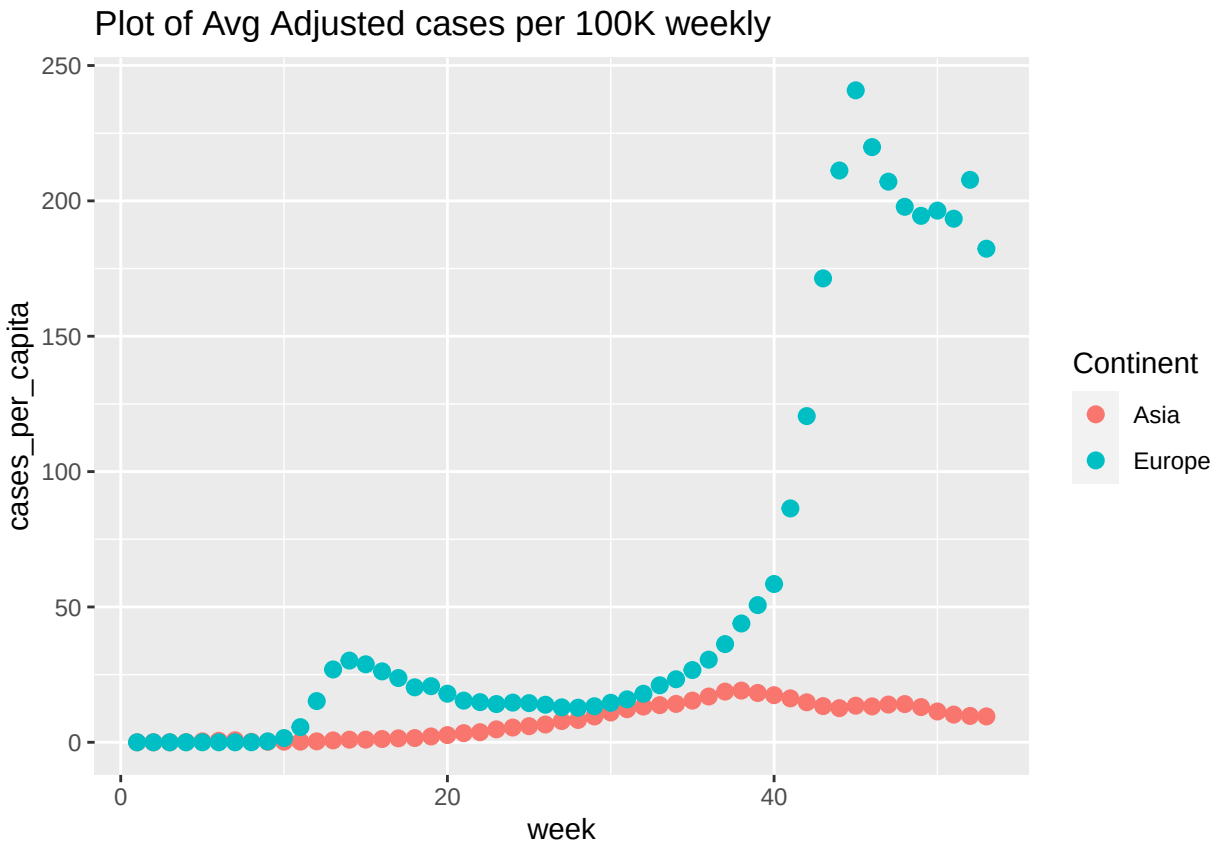
Approach:

We will treat each continent as a union of all cases. This means that we will sum per week all the weekly cases and adjust by the sum of all the population. Or in other words:

$$\frac{\Sigma_{countries-in-continent} weekly-cases}{\Sigma_{countries-in-continent} Population}$$

From here we will have two sets of groups, one for Asia and one for Europe. Once we have there two sets, we will have 2 constructed time series of weekly observation and their adjusted cases per 100K observants.

First lets addres the general behavior of the time series. We can see below that within year 2020 , until week 30 both continents held an average of below 25 cases per capita ,and from there the drastic growth within Europe took an increase , while in Asia it stayed steady. A thought on this would be to examine the methods of containment within the two continents and break down by country to understand which contries are responsible for maintaining this ratio.



Per this approach , lets try to create a coefficient varying model differed by the two groups. We will examine several models and create a model selection based on several criterias.

- Looking at the model 4 we get a very high R squared of 97%. In addition we can observe a high t value fr our intercept regarding the continent indicator with T =28.The REML is 394, lowest among all the model that were checked (see appendix).

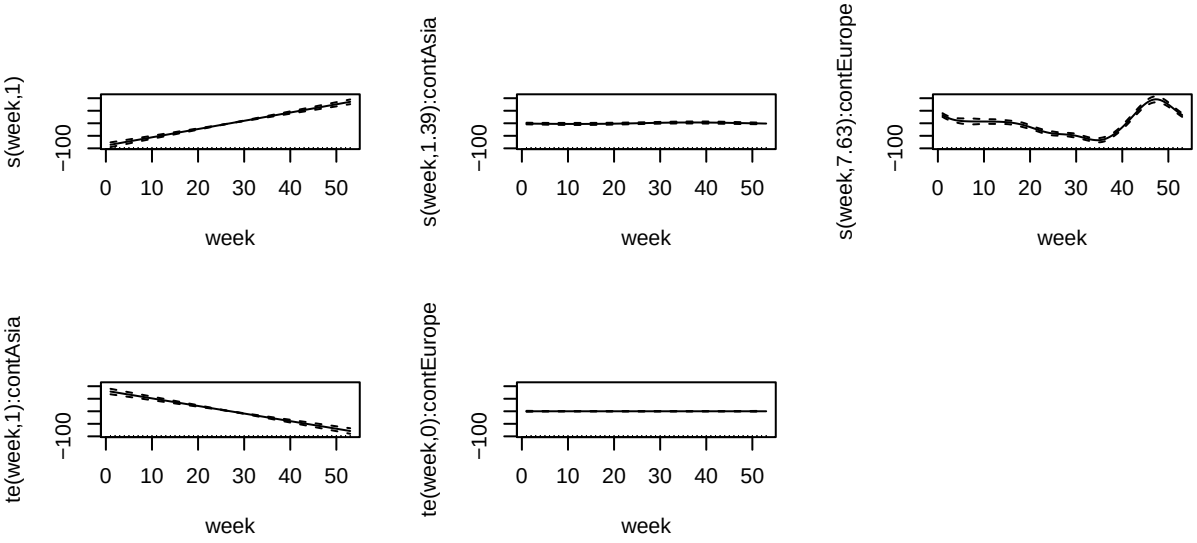
Running a quick anova test we can see that our selected model produces the lowest residuals deviance with the value 8150.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_per_capita ~ cont + s(week) + s(week, by = cont, bs = "cc") +
##     te(week, by = cont)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.661      1.286   5.957 4.54e-08 ***
## contEurope    51.068      1.819  28.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
```

```
##                            edf    Ref.df       F p-value
## s(week)              1.0000155 1.0000228 335.384  <2e-16 ***
## s(week):contAsia     1.3856393 8.0000000   0.345    0.12
## s(week):contEurope   7.6348046 8.0000000 135.019  <2e-16 ***
## te(week):contAsia    1.0001614 1.0002334 207.871  <2e-16 ***
## te(week):contEurope 0.0003691 0.0005511   0.308    0.99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 34/35
## R-sq.(adj) =  0.976   Deviance explained = 97.9%
## -REML =  394.9  Scale est. = 87.658    n = 106
```

Table 3: Anova 4 models

Resid. Df	Resid. Dev	Df	Deviance
96.12	97073	NA	NA
99.81	104228	-3.694	-7155
95.71	79840	4.096	24388
90.94	8150	4.776	71690



Extra Characterisitcs:

Now lets preform the T Test to compare means of both groups. Since the data is stationary (below - dickey fuller test indicating stationary), we will convert this by dealing ewith the weekly lag differences , and compare these two paired sets of values. our null hypothesis would be the the groups have the same means and we will execute a two sided T test to validate the significance of the test. From the time series plot show above, the first hint ould be that there is a different mean , but we will conclude this below.

From the results below we see a p value that is of value 0.07 , this is pretty high and in our case we can conclude a characteristic that the mean of the lag differences is different between the two groups. This is not suprising since we can see that the general behavior of each continent is different and there is a peak in Europe which is not seen in Asia.

Table 4: Dickey Fuller Test - Europe Lag Differences

Test statistic	Lag order	P value	Alternative hypothesis
-1.685	3	0.7014	stationary

Table 5: Dickey Fuller Test- Asia Lag Differences

Test statistic	Lag order	P value	Alternative hypothesis
-1.024	3	0.9261	stationary

Table 6: Welch Two Sample t-test: `diff(asia, 1)` and `diff(europe, 1)`

Test statistic	df	P value	Alternative hypothesis	mean of x	mean of y
-1.856	51.39	0.06926	two.sided	0.1839	3.506

Part 4

In this part we will try to characterize and cluster the different time series by country. The gaol is to try to identify if there are meaningful clusters and their underlying meaning. For this I decided to use the approach from class where I used a K Mediods model to cluster the countries. Since my time series were only 52 weeks long, and somecountries even have less, I decided to hold a small amount of clusters . First off we can view the plot of the dimension reduction using the first two principal components. Here there seems to be one large cluster and the rest of the observations are somewhat spread out. We will see how the model treats these other observations , with the expectation that the cluster we are able to see will probably stay the same.

Next we will perform the K meadiods on 3,4,5 clusters , and I found that the model doesnt seem to find enough observations to assign to the last clusters (5), and therefor I decided to stay with 4.

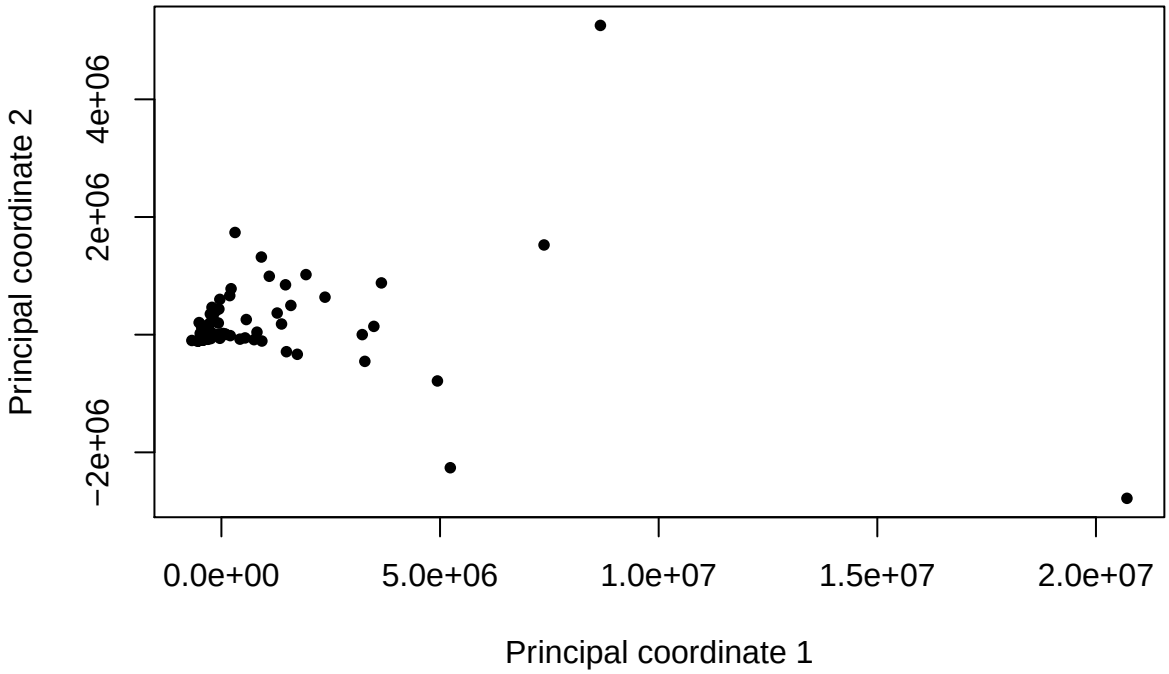
What we can see is quite interesting. First off , the vast majority of the countries are 79%-15%-4$-0.5% . I would interpret this as most of the countries dealing with the Covid 19 the same way , having the same behavior of the weekly cases. Looking at some of the other clusters, Cluster 2 seems mostly dominated by european countries, and cluster 3 looks like the set of countries which seemed to have drastic issues with the covid.(Example: Italy, UK , USA, Russia, France. . .) We can split this into 3 characters, having the low key, mid key and severe cases of covid aligned with the timing. These countries might have dealt with covid the

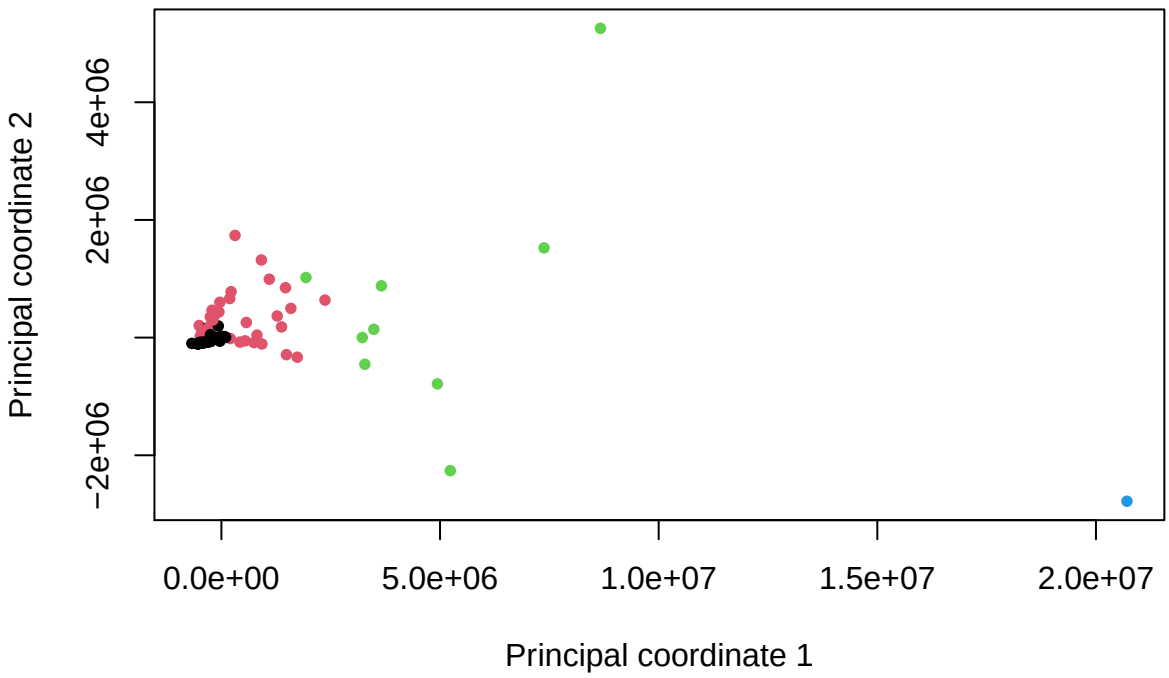
same way during the same time. Cluster 2 might be characterized as the european culture. We can imagine that europe would more or less deal with the covid the same way having a collaborative approach.

You can see in the appendix the cluster spread and assigning per country as well. In addition the distribution of the continents accross the clusters.

Another note from the appendix is the distribution per continent across the clusters. What we can see here is that African contries are 96% located in cluster 1 ,while only 60% of european countries. 12% of European countries are in cluster 3.

Observation and Assumption: Since we are analyzing time series clusters, i would assume that countries that are bordered one next to the other, might have similar issues - especially if the borders are open. So countries in Europe would mostly be tied together under the same clusters, and so in asia and Africa.






```
##  countriesAndTerritories       1                2                3
##  Length:214              Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  Class :character        1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Mode  :character        Median :1.0000   Median :0.0000   Median :0.00000
##                          Mean   :0.7991   Mean   :0.1542   Mean   :0.04206
##                          3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##                          Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##        4            continentExp
##  Min.   :0.000000   Length:214
##  1st Qu.:0.000000   Class :character
##  Median :0.000000   Mode  :character
##  Mean   :0.004673
##  3rd Qu.:0.000000
##  Max.   :1.000000
```

Appendix:

The following prints are of the summary results from the models discussed above.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

```
## cases_weekly ~ ti(week)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    50108       6616   7.573 9.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df     F p-value
## ti(week) 3.905  3.994 21.68  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.62   Deviance explained = 64.9%
## GCV = 2.5568e+09  Scale est. = 2.3202e+09  n = 53

##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_weekly ~ ti(week) + s(week)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    50108       4966   10.09 5.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df      F p-value
## ti(week) 1.000      1  1.927   0.172
## s(week)  7.097      8 11.364  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.786   Deviance explained = 81.9%
## GCV = 1.5778e+09  Scale est. = 1.307e+09  n = 53

##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_weekly ~ te(week)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    50108       6616   7.573 9.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
```

```
##            edf Ref.df     F p-value
## te(week) 3.905  3.994 21.68  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.62   Deviance explained = 64.9%
## GCV = 2.5568e+09  Scale est. = 2.3202e+09  n = 53
```

The following summaries are of the models from part 3 with the time varying coefficient models

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_per_capita ~ s(week) + s(week, by = cont, bs = "cc")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.195      3.067   10.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                          edf Ref.df     F  p-value
## s(week)             1.000088      1 13.95 0.000317 ***
## s(week):contAsia    0.004576      8  0.00 0.416699
## s(week):contEurope 6.635251      8 19.98  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.726   Deviance explained = 74.6%
## -REML =  522.7  Scale est. = 997.05    n = 106
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_per_capita ~ s(week) + te(week, by = cont)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.195      3.129   10.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                         edf Ref.df      F p-value
## s(week)             1.000  1.000 176.00  <2e-16 ***
## te(week):contAsia   1.000  1.001  85.50  <2e-16 ***
## te(week):contEurope 2.548  2.882  22.11  <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 17/18
## R-sq.(adj) =  0.715   Deviance explained = 72.8%
## -REML = 508.47  Scale est. = 1037.6    n = 106
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## cases_per_capita ~ s(week) + s(week, by = cont, bs = "cc") +
##     te(week, by = cont)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.20       2.78   11.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf    Ref.df      F p-value
## s(week)              1.000e+00 1.000e+00  1.946   0.166
## s(week):contAsia     2.807e-04 8.000e+00  0.000   0.879
## s(week):contEurope   5.560e+00 8.000e+00 14.497  <2e-16 ***
## te(week):contAsia    1.321e-05 2.641e-05  0.037   0.999
## te(week):contEurope  1.000e+00 1.000e+00 34.314  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 33/34
## R-sq.(adj) =  0.775   Deviance explained = 79.1%
## -REML = 500.72  Scale est. = 819.38    n = 106
```

Below we can see the distribution and assignment of the clusters per country and contient.

```
##
##                                          1 2 3 4
##   Afghanistan                            1 0 0 0
##   Albania                                1 0 0 0
##   Algeria                                1 0 0 0
##   Andorra                                1 0 0 0
##   Angola                                 1 0 0 0
##   Anguilla                               1 0 0 0
##   Antigua_and_Barbuda                    1 0 0 0
##   Armenia                                1 0 0 0
##   Aruba                                  1 0 0 0
##   Australia                              1 0 0 0
##   Bahamas                                1 0 0 0
##   Bahrain                                1 0 0 0
##   Barbados                               1 0 0 0
##   Belarus                                1 0 0 0
##   Belize                                 1 0 0 0
```

```
##   Benin                                       1 0 0 0
##   Bermuda                                     1 0 0 0
##   Bhutan                                      1 0 0 0
##   Bolivia                                     1 0 0 0
##   Bonaire, Saint Eustatius and Saba           1 0 0 0
##   Bosnia_and_Herzegovina                      1 0 0 0
##   Botswana                                    1 0 0 0
##   British_Virgin_Islands                      1 0 0 0
##   Brunei_Darussalam                           1 0 0 0
##   Bulgaria                                    1 0 0 0
##   Burkina_Faso                                1 0 0 0
##   Burundi                                     1 0 0 0
##   Cambodia                                    1 0 0 0
##   Cameroon                                    1 0 0 0
##   Cape_Verde                                  1 0 0 0
##   Cases_on_an_international_conveyance_Japan 1 0 0 0
##   Cayman_Islands                              1 0 0 0
##   Central_African_Republic                    1 0 0 0
##   Chad                                        1 0 0 0
##   China                                       1 0 0 0
##   Comoros                                     1 0 0 0
##   Congo                                       1 0 0 0
##   Costa_Rica                                  1 0 0 0
##   Cote_dIvoire                                1 0 0 0
##   Cuba                                        1 0 0 0
##   Cura\303§ao                                   1 0 0 0
##   Cyprus                                      1 0 0 0
##   Democratic_Republic_of_the_Congo            1 0 0 0
##   Denmark                                     1 0 0 0
##   Djibouti                                    1 0 0 0
##   Dominica                                    1 0 0 0
##   Dominican_Republic                          1 0 0 0
##   Ecuador                                     1 0 0 0
##   Egypt                                       1 0 0 0
##   El_Salvador                                 1 0 0 0
##   Equatorial_Guinea                           1 0 0 0
##   Eritrea                                     1 0 0 0
##   Estonia                                     1 0 0 0
##   Eswatini                                    1 0 0 0
##   Ethiopia                                    1 0 0 0
##   Falkland_Islands_(Malvinas)                 1 0 0 0
##   Faroe_Islands                               1 0 0 0
##   Fiji                                        1 0 0 0
##   Finland                                     1 0 0 0
##   French_Polynesia                            1 0 0 0
##   Gabon                                       1 0 0 0
##   Gambia                                      1 0 0 0
##   Ghana                                       1 0 0 0
##   Gibraltar                                   1 0 0 0
##   Greece                                      1 0 0 0
##   Greenland                                   1 0 0 0
##   Grenada                                     1 0 0 0
##   Guam                                        1 0 0 0
##   Guatemala                                   1 0 0 0
```

```
##   Guernsey                          1 0 0 0
##   Guinea                            1 0 0 0
##   Guinea_Bissau                     1 0 0 0
##   Guyana                            1 0 0 0
##   Haiti                             1 0 0 0
##   Holy_See                          1 0 0 0
##   Honduras                          1 0 0 0
##   Iceland                           1 0 0 0
##   Ireland                           1 0 0 0
##   Isle_of_Man                       1 0 0 0
##   Jamaica                           1 0 0 0
##   Jersey                            1 0 0 0
##   Kazakhstan                        1 0 0 0
##   Kenya                             1 0 0 0
##   Kosovo                            1 0 0 0
##   Kuwait                            1 0 0 0
##   Kyrgyzstan                        1 0 0 0
##   Laos                              1 0 0 0
##   Latvia                            1 0 0 0
##   Lebanon                           1 0 0 0
##   Lesotho                           1 0 0 0
##   Liberia                           1 0 0 0
##   Libya                             1 0 0 0
##   Liechtenstein                     1 0 0 0
##   Lithuania                         1 0 0 0
##   Luxembourg                        1 0 0 0
##   Madagascar                        1 0 0 0
##   Malawi                            1 0 0 0
##   Malaysia                          1 0 0 0
##   Maldives                          1 0 0 0
##   Mali                              1 0 0 0
##   Malta                             1 0 0 0
##   Marshall_Islands                  1 0 0 0
##   Mauritania                        1 0 0 0
##   Mauritius                         1 0 0 0
##   Moldova                           1 0 0 0
##   Monaco                            1 0 0 0
##   Mongolia                          1 0 0 0
##   Montenegro                        1 0 0 0
##   Montserrat                        1 0 0 0
##   Mozambique                        1 0 0 0
##   Myanmar                           1 0 0 0
##   Namibia                           1 0 0 0
##   New_Caledonia                     1 0 0 0
##   New_Zealand                       1 0 0 0
##   Nicaragua                         1 0 0 0
##   Niger                             1 0 0 0
##   Nigeria                           1 0 0 0
##   North_Macedonia                   1 0 0 0
##   Northern_Mariana_Islands          1 0 0 0
##   Norway                            1 0 0 0
##   Oman                              1 0 0 0
##   Palestine                         1 0 0 0
##   Panama                            1 0 0 0
```

```
##   Papua_New_Guinea                          1 0 0 0
##   Paraguay                                  1 0 0 0
##   Puerto_Rico                               1 0 0 0
##   Qatar                                     1 0 0 0
##   Rwanda                                    1 0 0 0
##   Saint_Kitts_and_Nevis                     1 0 0 0
##   Saint_Lucia                               1 0 0 0
##   Saint_Vincent_and_the_Grenadines          1 0 0 0
##   San_Marino                                1 0 0 0
##   Sao_Tome_and_Principe                     1 0 0 0
##   Saudi_Arabia                              1 0 0 0
##   Senegal                                   1 0 0 0
##   Seychelles                                1 0 0 0
##   Sierra_Leone                              1 0 0 0
##   Singapore                                 1 0 0 0
##   Sint_Maarten                              1 0 0 0
##   Slovakia                                  1 0 0 0
##   Slovenia                                  1 0 0 0
##   Solomon_Islands                           1 0 0 0
##   Somalia                                   1 0 0 0
##   South_Korea                               1 0 0 0
##   South_Sudan                               1 0 0 0
##   Sri_Lanka                                 1 0 0 0
##   Sudan                                     1 0 0 0
##   Suriname                                  1 0 0 0
##   Syria                                     1 0 0 0
##   Taiwan                                    1 0 0 0
##   Tajikistan                                1 0 0 0
##   Thailand                                  1 0 0 0
##   Timor_Leste                               1 0 0 0
##   Togo                                      1 0 0 0
##   Trinidad_and_Tobago                       1 0 0 0
##   Tunisia                                   1 0 0 0
##   Turks_and_Caicos_islands                  1 0 0 0
##   Uganda                                    1 0 0 0
##   United_Arab_Emirates                      1 0 0 0
##   United_Republic_of_Tanzania               1 0 0 0
##   United_States_Virgin_Islands              1 0 0 0
##   Uruguay                                   1 0 0 0
##   Uzbekistan                                1 0 0 0
##   Vanuatu                                   1 0 0 0
##   Venezuela                                 1 0 0 0
##   Vietnam                                   1 0 0 0
##   Wallis_and_Futuna                         1 0 0 0
##   Western_Sahara                            1 0 0 0
##   Yemen                                     1 0 0 0
##   Zambia                                    1 0 0 0
##   Zimbabwe                                  1 0 0 0

##
##               1 2 3 4
##   Argentina   0 1 0 0
##   Austria     0 1 0 0
##   Azerbaijan  0 1 0 0
```

```
##   Bangladesh   0 1 0 0
##   Belgium      0 1 0 0
##   Canada       0 1 0 0
##   Chile        0 1 0 0
##   Colombia     0 1 0 0
##   Croatia      0 1 0 0
##   Czechia      0 1 0 0
##   Georgia      0 1 0 0
##   Hungary      0 1 0 0
##   Indonesia    0 1 0 0
##   Iran         0 1 0 0
##   Iraq         0 1 0 0
##   Israel       0 1 0 0
##   Japan        0 1 0 0
##   Jordan       0 1 0 0
##   Mexico       0 1 0 0
##   Morocco      0 1 0 0
##   Nepal        0 1 0 0
##   Netherlands  0 1 0 0
##   Pakistan     0 1 0 0
##   Peru         0 1 0 0
##   Philippines  0 1 0 0
##   Poland       0 1 0 0
##   Portugal     0 1 0 0
##   Romania      0 1 0 0
##   Serbia       0 1 0 0
##   South_Africa 0 1 0 0
##   Sweden       0 1 0 0
##   Switzerland  0 1 0 0
##   Ukraine      0 1 0 0
```

```
##
##                  1 2 3 4
##   Brazil         0 0 1 0
##   France         0 0 1 0
##   Germany        0 0 1 0
##   India          0 0 1 0
##   Italy          0 0 1 0
##   Russia         0 0 1 0
##   Spain          0 0 1 0
##   Turkey         0 0 1 0
##   United_Kingdom 0 0 1 0
```

```
## 1 2 3 4
## 0 0 0 1
```

```
## # A tibble: 6 x 4
##   continentExp `1_1`  `2_1`  `3_1`
##   <chr>        <dbl>  <dbl>  <dbl>
## 1 Africa       0.964 0.0364 0
## 2 America      0.837 0.122  0.0204
## 3 Asia         0.738 0.238  0.0238
## 4 Europe       0.6   0.273  0.127
```

```
## 5 Oceania      1     0      0
## 6 Other        1     0      0
```

```
## ----setup, include=FALSE----------------------------------------------------------------------------------------


## ---- echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-----------------------------------------------------
# import libraries
library(rmarkdown)
library(plyr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(pivottabler)
library(gtsummary)
library(ggpubr)
library(ggfortify)
library(cluster)
library(MASS)
library(lmtest)
library(fBasics)
library(rcompanion)
library(gridExtra)
library(cowplot)
library(kableExtra)
library(haven)
library(tidyverse)
library(rstatix)
library(ggpubr)
library(lme4)
library(reshape2)
library(kableExtra)
library(pander)
library(performance)
library(pROC)
library(sqldf)
library(nlme)
library(ggeffects)
library(doBy)
library(tseries)
library(forecast)




## ---- echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-----------------------------------------------------
# data <- read.csv("https://opendata.ecdc.europa.eu/covid19/casedistribution/csv", na.strings = # "", f

#setwd("School/courses/applied_stats/p4")
#write.csv( data, 'data.csv')
data <- read.csv('download')


## ----initial_plot , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------------
```

```r
#convert dates columns
# plot the weekly cases in france
data$dateRep<-as.Date(data$dateRep, '%d/%m/%Y')
france <-  data[ data$countriesAndTerritories=='France' ,]
p <- ggplot(france, aes(x=dateRep, y=cases_weekly)) +
  geom_line() +
  xlab("")+ ggtitle("Plot of weekly cases in France")
grid.arrange(p)

par(mfrow=c(1,1))
# dickey fuller test
options(warn=-1)
pander(adf.test(france$cases_weekly), caption = "Dickey Fuller Test")



## ----exlplore_data , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-----------------------------

par(mfrow=c(1,2))
# dickey fuller test
options(warn=-1)
#pander(adf.test(france$cases_weekly), caption = "Dickey Fuller Test")

#par(mfrow=c(1,2))
p1<- acf(france$cases_weekly, plot=FALSE)
plot(p1,main = "ACF")

p2<- pacf(france$cases_weekly, plot=FALSE)
plot(p2,main = "PACF")

# box cox transofrmation
tseries_h<- france$cases_weekly
#bx<- BoxCox(tseries_h, lambda = 0.5)
#plot.ts(bx)
#lambda <- BoxCox.lambda(tseries_h)
#adf.test(bx)



## ----auto_arima1 , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-------------------------------
#Aproach 1
par(mfrow=c(1,1))
options(warn = -1)
f<- auto.arima(tseries_h)
plot(forecast(f,h=20))
#pander(summary(f) , caption ='Summary Stepwise ARIMA')
#pander(f$coef , caption ='Coefficients non Stepwise ARIMA')
#pander(f$aic , caption ='AIC non Stepwise ARIMA')
#pandoc.table(f$aic, keep.line.breaks = FALSE,caption ='AIC Stepwise ARIMA',style = 'rmarkdown')


par(mfrow=c(1,3))
# lewts check auto coreelation since we are looking at the diff (0,1,0)
```

```r
plot(diff(tseries_h),main = "Scatter Differences")
p1<- acf(diff(tseries_h),plot = FALSE)
plot(p1,main = "ACF")
p2<- pacf(diff(tseries_h),plot = FALSE)
plot(p2,main = "PACF")


## ----auto_arima2 , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-------------------------------

#Aproach 2
options(warn = -1)
f<- auto.arima(tseries_h, stepwise = FALSE,seasonal=FALSE)
pander(f$coef , caption ='Coefficients non Stepwise ARIMA')
#pandoc.table(f$aic, keep.line.breaks = FALSE,caption ='AIC non Stepwise ARIMA',style = #'rmarkdown')



#pander(f$aic , caption ='AIC non Stepwise ARIMA')
# lewts check auto coreelation since we are looking at the diff (0,1,0)
#resid<- checkresiduals(f, plot=FALSE,test=FALSE)
#par(mfrow=c(1,1))
#plot(forecast(f,h=20))

#checkresiduals(f,test = FALSE)


## ----function , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------

# define plotting function
plot_predictions <- function(france, m , title){
  pred <- data.frame(week = france$week,
                          cases_weekly = france$cases_weekly,
                          predicted_values = predict(m, newdata = france))

 ggplot(pred, aes(x = week)) +
  geom_point(aes(y = cases_weekly), size = 1, alpha = 0.5) + geom_line(aes(y = predicted_values), colou
}



## ----gam , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------------



library(mgcv)
library(stringr)
# split the column to use the week as an input
france$week <- str_split_fixed(france$year_week,'-',2)[,2]
france$week <- as.numeric(france$week)


p1 <- ggplot(france, aes(week, cases_weekly)) + geom_point()
p1
```

```r
## ----s_week , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------
############################################################
# s(week)
m1 <- gam(cases_weekly ~ s(week), data = france , method = "REML")
# plot diagnostics
par(mfrow = c(2,2))
gam.check(m1)

#the larger the number, the more wiggly the fitted model.
summary(m1)

#model of s(week)
#p1<- ggplot(france, aes(week, cases_weekly)) + geom_point() + geom_smooth(method = "gam", formula = y

p1 <- plot_predictions(france,m1,"cases_weekly ~ s(week)")
grid.arrange(p1)


## ----ti_week , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------
############################################################
# ti(week)

m2 <- gam(cases_weekly ~ ti(week), data = france)
# plot diagnostics
par(mfrow = c(2,2))
gam.check(m2)

#the larger the number, the more wiggly the fitted model.
#summary(m2)

#model of s(week)
#ggplot(france, aes(week, cases_weekly)) + geom_point() + geom_smooth(method = "gam", formula = #y ~ti(
#grid.arrange(p1)
p2<- plot_predictions(france,m2,"cases_weekly ~ ti(week)")



############################################################
# ti(week) +s(week)

m3 <- gam(cases_weekly ~ ti(week) +s(week), data = france)
# plot diagnostics
par(mfrow = c(2,2))
gam.check(m3)

#the larger the number, the more wiggly the fitted model.
#summary(m3)

#model of s(week)
#p2<- ggplot(france, aes(week, cases_weekly)) + geom_point() + geom_smooth(method = "gam", #formula = y
#grid.arrange(p2)
```

```r
p3<- plot_predictions(france,m3,"cases_weekly ~ ti(week) +s(week)")


############################################################
m11 <- gam(cases_weekly ~ te(week), data = france)

#the larger the number, the more wiggly the fitted model.
#summary(m11)

#model of te(week)
#ggplot(france, aes(week, cases_weekly)) + geom_point() + geom_smooth(method = "gam", formula = y ~te(x

p11<- plot_predictions(france,m11,"cases_weekly ~ te(week) ")
############################################################


grid.arrange(p2,p3,p11)



## ----k , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------------
# change the number of basis functions

m4 <- gam(cases_weekly ~ s(week, bs = 'cc', k = 52), data = france , method = "REML")
m5 <- gam(cases_weekly ~ s(week, bs = 'cc', k = 1), data = france , method = "REML")
m6 <- gam(cases_weekly ~ s(week, bs = 'cc', k = 10), data = france , method = "REML")
m7 <- gam(cases_weekly ~ s(week, bs = 'cc', k = 20), data = france , method = "REML")



#model of s(week)
#ggplot(france, aes(week, cases_weekly)) + geom_point() + geom_smooth(method = "gam", formula = y ~s(x,


p4<- plot_predictions(france,m4,"Cyclic ~ 52 knots")
p5<-plot_predictions(france,m5,"Cyclic ~ 1 knot")
p6<-plot_predictions(france,m6,"Cyclic ~ 10 knots")
p7<-plot_predictions(france,m7,"Cyclic ~ 20 knots")

grid.arrange(p4,p5,p6,p7)



## ----gamma , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE------------------------------------

############################################################
# change the smoothing parameter - gamma


m8 <- gam(cases_weekly ~ s(week), gamma=1 ,data = france , method = "REML")
p8 <-plot_predictions(france,m8, 'gamma = 1')

m9 <- gam(cases_weekly ~ s(week), gamma=20 ,data = france , method = "REML")
```

```r
p9 <-plot_predictions(france,m9, 'gamma = 20')

m10 <- gam(cases_weekly ~ s(week), gamma=0.1 ,data = france , method = "REML")
p10 <- plot_predictions(france,m10, 'gamma = 0.1')


grid.arrange(p8,p9,p10)


## ----anova , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-----------------------------------

model1 <- gamm(cases_weekly ~ ti(week), data = france)
model2<- gamm(cases_weekly ~ s(week), data = france)
model3<- gamm(cases_weekly ~ s(week) + ti(week), data = france)
model4<- gamm(cases_weekly ~ s(week, bs = 'cc', k = 52), data = france , method = "REML")
model5<- gamm(cases_weekly ~ s(week, bs = 'cc', k = 1), data = france , method = "REML")
model6<- gamm(cases_weekly ~ s(week, bs = 'cc', k = 10), data = france , method = "REML")
model7<- gamm(cases_weekly ~ s(week, bs = 'cc', k = 20), data = france , method = "REML")
model8<- gamm(cases_weekly ~ s(week), gamma=1 ,data = france , method = "REML")
model9<- gamm(cases_weekly ~ s(week), gamma=20 ,data = france , method = "REML")
model10 <- gamm(cases_weekly ~ s(week), gamma=0.1 ,data = france , method = "REML")
model11 <- gamm(cases_weekly ~ te(week) ,data = france , method = "REML")

anova(model1$lme,
      model2$lme,
      model3$lme)

anova(model4$lme,
      model5$lme,
      model6$lme,
      model7$lme)

anova(model7$lme,
      model8$lme,
      model9$lme)


## ----bam , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE--------------------------------------

# check where the drop of the rho is
#once we hit a high rho , we see that the trend is random  - and they are due to auto correlated errors

aicvec<-remlvec<-c()
rovec<- seq(.01,1,0.05)
for (k in 1:length(rovec)){
  tmp<- bam( cases_weekly ~ s(week) , rho = rovec[k],data = france)
  aicvec[k] <- AIC(tmp)
  remlvec[k] <- tmp$gcv.ubre
  #cat(rovec[k], aicvec[k], remlvec[k],"\n")


}
```

```r
#matplot(rovec, cbind(aicvec,remlvec))
q1<- qplot(rovec, aicvec)+ ggtitle('Rho value vs AIC ')
q2<- qplot(rovec, remlvec)+ ggtitle('Rho values vs REML' )

# check certain rhos

m13 <- bam(cases_weekly ~ s(week) , rho =0.7,data = france)
p13 <- plot_predictions(france,m13, 'BAM with rho 0.7')

m14 <- bam(cases_weekly ~ s(week) , rho =0.5,data = france)
p14 <- plot_predictions(france,m14, 'BAM with rho 0.5')
grid.arrange(q1,q2)
grid.arrange(p13,p14)



## ----vc , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------------
data$cases_per_capita <- data$cases_weekly / data$popData2019 * 100000

data$week <- str_split_fixed(data$year_week,'-',2)[,2]
data$week <- as.numeric(data$week)


# crete set of data aggregated by continent and sum weekly cases and population (per week)
agg_data <- data %>%
  group_by(continentExp,week) %>%
  summarise(cases_weekly = sum(cases_weekly),
            popData2019 = sum(popData2019))

agg_data$cases_per_capita <- agg_data$cases_weekly / agg_data$popData2019 * 100000

agg_data <- agg_data[(agg_data$continentExp == 'Asia')|((agg_data$continentExp == 'Europe')),]


agg_data$cont <- factor(agg_data$continentExp)
agg_data<- na.omit(agg_data)

cases_per_capita<- as.vector(agg_data$cases_per_capita)
dateRep <- agg_data$dateRep
cont <- agg_data$cont

cont_plot <- ggplot(agg_data, aes(x = week, y = cases_per_capita, colour = factor(cont)))+
  geom_point(size=2.5)+ ggtitle('Plot of Avg Adjusted cases per 100K weekly')
cont_plot$labels$colour <- "Continent"
grid.arrange(cont_plot)


## ----vc_gam , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE------------------------------------
m3_1 <- gam(cases_per_capita ~ s(week)+ s(week,by=cont,bs="cc"), data = agg_data , method = "REML")
m3_2 <- gam(cases_per_capita ~ s(week)+ te(week,by=cont), data = agg_data , method = "REML")
m3_3 <- gam(cases_per_capita ~ s(week)+ s(week,by=cont,bs="cc")+ te(week,by=cont), data = agg_data , me
summary(m3_4 <- gam(cases_per_capita ~ cont + s(week)+ s(week,by=cont,bs="cc")+ te(week,by=cont), data =
method = "REML"))
```

```r
par(mfrow=c(3,3))
plot(m3_4)

pander(anova(m3_1,m3_2,m3_3,m3_4),caption = 'Anova 4 models')



## ----tstat , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-----------------------------------

##### -------- aggregating all countries in same continent a- sum all weekly cases and adjust
asia <-  agg_data[(agg_data$continentExp == 'Asia'),]$cases_per_capita
europe <-  agg_data[(agg_data$continentExp == 'Europe'),]$cases_per_capita

pander(adf.test(europe), caption = "Dickey Fuller Test - Europe Lag Differences")
pander(adf.test(asia), caption = "Dickey Fuller Test- Asia Lag Differences")

# low p val -> not equal
pander(t.test(diff(asia,1), y = diff(europe,1), alternative = c("two.sided"), paired = FALSE, var.equal



## ----ts_clustering , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------
cols <- c('countriesAndTerritories','week','cases_weekly')
pv <- data[cols]
library(reshape2)

#pv <- pv[pv$week > 10 , ]
rr <- recast(pv, countriesAndTerritories  ~ week, id.var = c( "week",'countriesAndTerritories'))
rr[is.na(rr)] <- 0

require(dtw)
jj <- dist(rr[,-1], method="dtw")
mds <- cmdscale(jj, eig=TRUE)
plot(mds$points[,1:2], pch=16, cex=.8, xlab="Principal coordinate 1", ylab="Principal coordinate 2")

require(cluster)
p3 <- pam(jj,3)      # k-medoids clustering
t_table <- table(p3$clust,rr[,1])

#merge with continent and compare
tr_table <- t(t_table)

plot(mds$points[,1:2], pch=16, cex=.8, xlab="Principal coordinate 1", ylab="Principal coordinate 2",col=



continents_and_countries <- as.data.frame(unique(data[c('countriesAndTerritories','continentExp')]))
rownames(continents_and_countries) <- continents_and_countries$countriesAndTerritories

df_tr_table <- as.data.frame.matrix(tr_table)
df_tr_table$countriesAndTerritories <- rownames(df_tr_table)

m<- merge(x = df_tr_table,y = continents_and_countries, by.x='countriesAndTerritories', by.y='countries
```

```r
## ----summaries , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE----------------------------------
summary(m2)
summary(m3)
summary(m11)


## ----p3_summaries , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-------------------------------
summary(m3_1)
summary(m3_2)
summary(m3_3)



## ----cluster , echo=FALSE , echo=FALSE, warning=FALSE,message=FALSE-----------------------------------
# all cluster 1
tr_table[tr_table[,1] ==1,]
# all cluster 2
tr_table[tr_table[,2] ==1,]
# all cluster 3
tr_table[tr_table[,3] ==1,]

m %>%
  group_by(continentExp) %>%
  summarise(across(c(2,3,4), list(mean)))
```