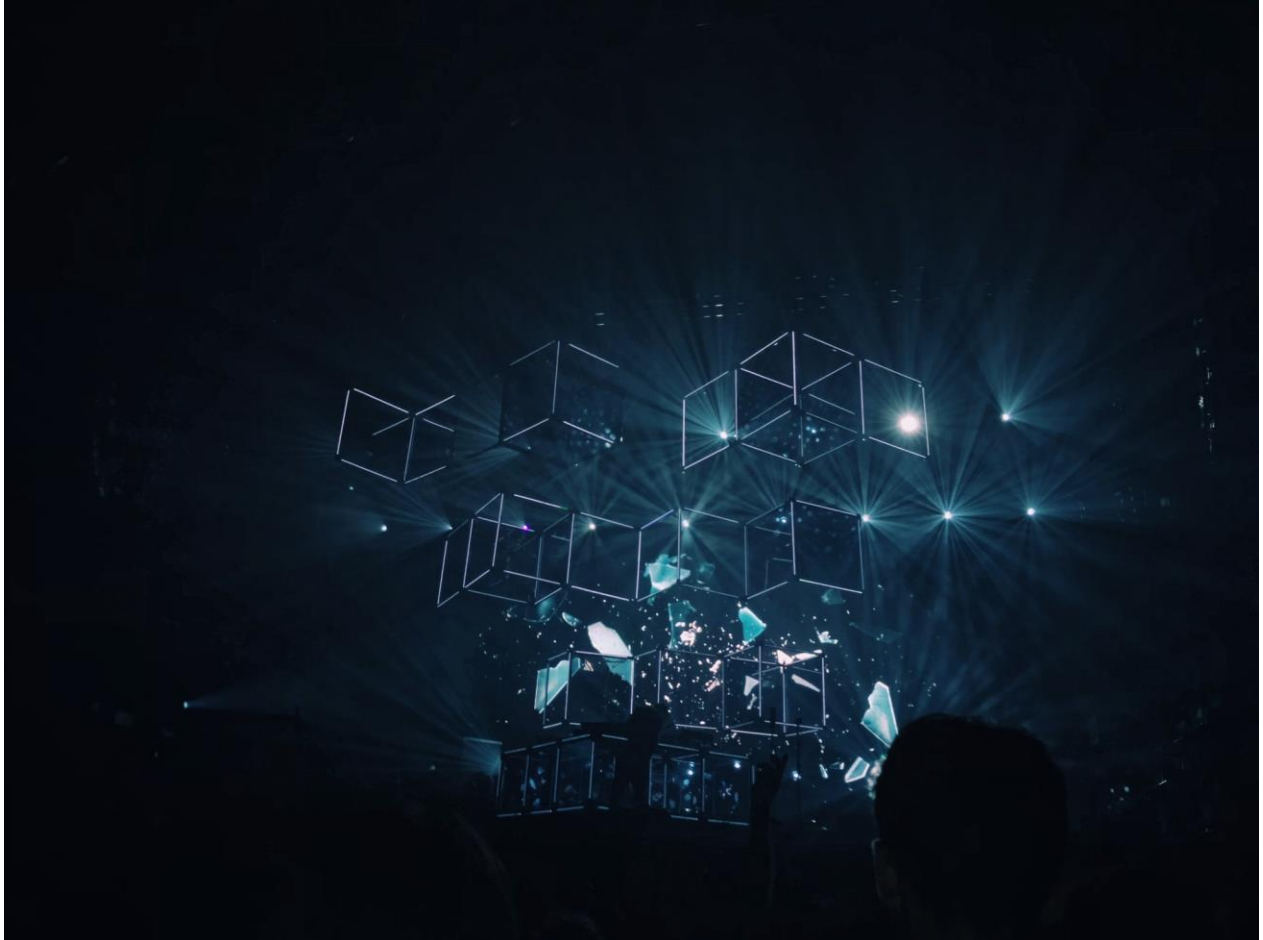


Assessment 3 Problem Solving Task 2

Statistical Data Analysis

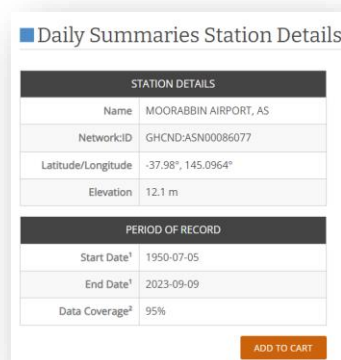


Question 1:**Road traffic accident dataset (16 points)****Q1.1: Which data source do you plan to use? Justify your decision. (4 points)**

Accident Data – This data is same as the one used in the assignment 1. The chosen region is Metropolitan Southeast region. Albeit I had used East region in the assignment 1, I have gone with Southeast region as the weather data I found has a coverage of 95% in the area. And analysing that would mean that the results will be somewhat accurate.

Weather Data – I have chosen a southeast region, “Moorabin Airport AS” using the NOAA token key that was requested on email and used “rnoaa” library to fetch the data. This data has a 95% coverage and covers the time range the accident data was recorded as well.

The implementation on accessing the data is shown in the R notebook



The screenshot shows the NOAA station details for Moorabin Airport AS. It includes station details like Name, NetworkID, Latitude/Longitude, and Elevation, as well as the period of record with start and end dates and data coverage percentage. An 'ADD TO CART' button is visible at the bottom right.

STATION DETAILS	
Name	MOORABBIN AIRPORT, AS
NetworkID	GHCND:ASN00086077
Latitude/Longitude	-37.98°, 145.0964°
Elevation	12.1 m
PERIOD OF RECORD	
Start Date ¹	1950-07-05
End Date ¹	2023-09-09
Data Coverage ²	95%

Link to data source - <https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:ASN00086077/detail>

This sets us to begin the analysis to find the relationship between weather and accidents happened in the region in the particular time range

Q1.2: From the data source identified, download daily temperature and precipitation data for the region during the relevant time period. (Hint: If you download data from NOAA <https://www.ncdc.noaa.gov/cdo-web/>, you need to request an NOAA web service token for accessing the data.) (2 points)

From the NOAA website, on the search filter, Air Temperature and precipitation was selected for the date range to identify the data. However, through R code, using `datatyped` the temperature and precipitation were filtered.

Q1.2: Answer the following questions (Provide the output from your R Studio):

• **How many rows are in your local weather data? (2 points)**

The data was not downloaded separately. Instead, data features such as TMIN, TMAX and PRCP were fetched using multiple API requests. As I have filtered the data to the time range I Require, it has 4925 rows and 8 columns.

```
Rows 4925 Cols 8
```

And the following is after merging the data with the southeast region accident data and weather data (Post removing unnecessary columns)

```
Tot_rows: 1644
Tot Cols 6
```

• **What time period does the data cover?**

As mentioned above as well, the original data covers 1950 July to 2023 September. However, I have joined the data to fit the time range we require, that is 2016 to 2020 as per the accidents data.

Note: The data was prepared in 5 steps as it had a limit of 1000 rows at a time. I have prepared the data in 5 chunks and have row bonded all to a single data.

Question 2:

Q2.1. Model planning:

a) What is the main goal of your model, how it will be used? (1 point)

- Main goal of the model is to predict the accidents as per the weather conditions to be extra vigilant and allocate resources and take other necessary precautionary steps. This in turn will help control/reduce the number of accidents in future.
- This is possible by using the right historic data and using/training the best models/algorithms and by evaluating the model using several performance metrics

b) How it will be relevant to the emergency services demand?

- If the model has a good performance, emergency services can utilize the available resources at the right time, this in turn will be helpful to cut cost, time, energy and can reap more benefits by alleviating several accidents by offering the best services.
- [1] As per this paper, Results of an analysis indicate two environmental factors were significantly associated with an increased risk of crash-related mortality and injury among taxi drivers.
- The same can be monitored for a time period to see if the services are really able to control accidents.
- For continuous improvements, the historical data has to be continuously used using different strategies to be able generalize better and predict.

c) Who are the potential users of your model? (1 point)

- General public, Local traffic officers, emergency services, corporates (when conducting events) can all benefit from this model as they can make a judgement to drive or take necessary precautions in particular weather conditions that predict a abnormally high accidents

Q2.2. Relationship and data:**a) What relationship do you plan to model or what do you want to predict? (1 point)**

Models are experimental until we compare and pick the best one out of all. Initially, linear model will be used to observe the performance and relationship. Reason being, linear model is easy to implement and visualize the relationship using graphs to understand and the fundamental connections before we build more complex models.

A linear model of Total_accidents vs Date will be used as a starting point of the model, and then with multi variate linear model using predictors, such as TMIN, TMAX and PRCP for the weather data and the response variable TOTAL_ACCIDENTS from the accidents data of south east region. Evaluate the model based on Residual standard error, multiple R-squared and adjusted r-squared.

Post observing the performance, we can go for GAM model to understand the non-linear relationships between the predictors and the response variables. Finally, compare the performance between linear model and GAM model based on evaluation metrics.

b) What is the response variable? (1 point)

- The total number of accidents is the response variable.

c) What are the predictor variables? (1 point)

- The predictor variables are Date, Week day, TMIN, TMAX, Temperature range and PRCP from the weather data of southeast region.

d) Will the variables in your model be routinely collected and made available soon enough for prediction? (1 point)

Yes, that was used in this assignment, is being collected everyday. The last record in it is of last month, (September 2023)

e) As you are likely to build your model on historical data, will the data in the future have similar characteristics? (1 point)

Yes, the data being used will have similar characteristics as the one that's being used is of historic data from the same region, from the National centers for environmental information.

Q2.3. What statistical method(s) will be applied to generate the model? Why? (2 points)

For the linear model used, following are some of the statistical methods being appended,

Mean - Mean is a fundamental statistical measure that provides the average value of a variable (e.g., TOTAL_ACCIDENTS, Air temperature). It helps in understanding the central tendency and overall trend in the data.

Standard Errors - The standard errors associated with the coefficient estimates. These are used in hypothesis testing and confidence interval construction.

t-values - t-values are the ratio of the coefficient estimate to its standard error. They are used for hypothesis testing regarding the significance of each coefficient.

p-values - The p-values associated with each coefficient test the null hypothesis that the true coefficient is zero (no effect). Lower p-values indicate greater evidence against the null hypothesis.

Residuals - Descriptive statistics for the residuals (differences between observed and predicted values). This includes minimum, 1st quartile, median, 3rd quartile, and maximum values.

However, for **GAM** the statistical methods used were,

Spline functions are frequently used to describe smooth functions in GAMs because they enable estimation of smooth, non-linear connections without requiring a predetermined functional form.

The level of smoothness in a smooth function is controlled by the smoothing parameter. It establishes how closely the calculated curve complies with the observed data.

Question 3:

Q3.1 Which region do you pick?

I picked metropolitan south-east region from car_accident data

Q3.2 Fit a linear model for Y using date as the predictor variable. Plot the fitted values and the residuals. Assess the model fit. Is a linear function sufficient for modelling the trend of Y? Support your conclusion with plots.

After fitting the linear model for Y, let's analyse the plots of fitted values and the residuals.

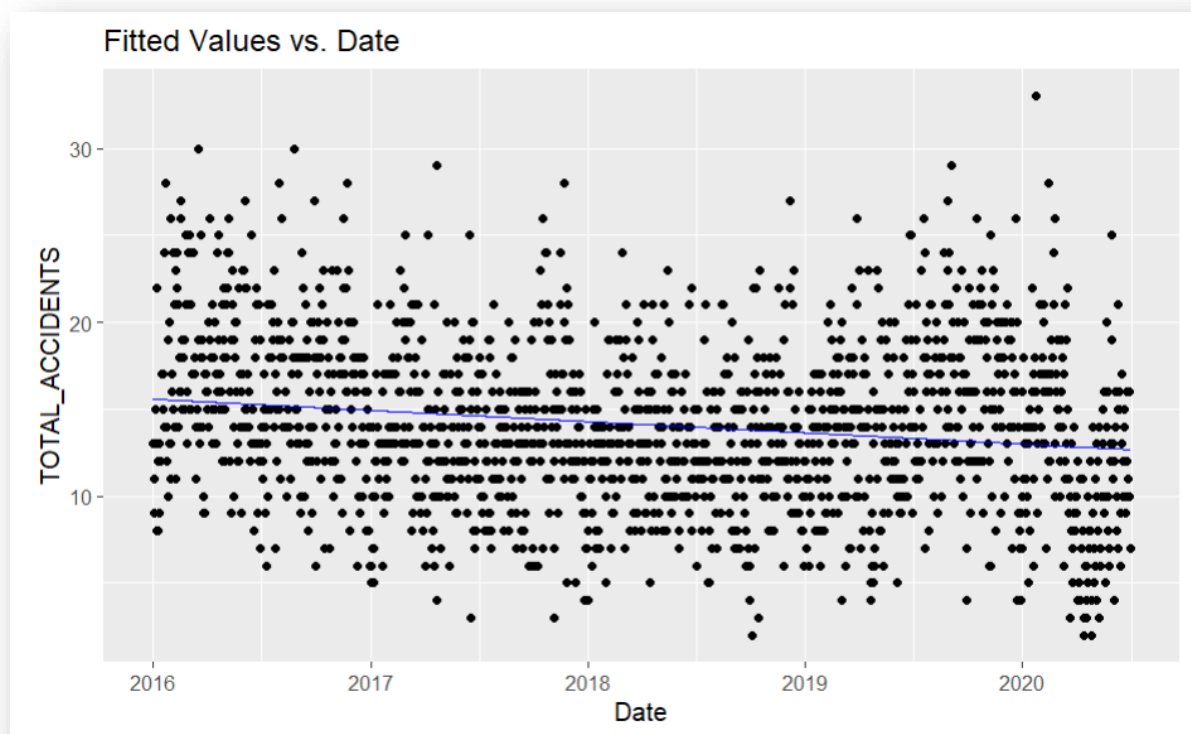


Fig 3.1

Assessing model fit: As per the above Residuals vs Date linear model plot, the model hardly utilizes between 12 to 16 accidental data points for prediction. The model does not fit the data well as several data points are being missed.

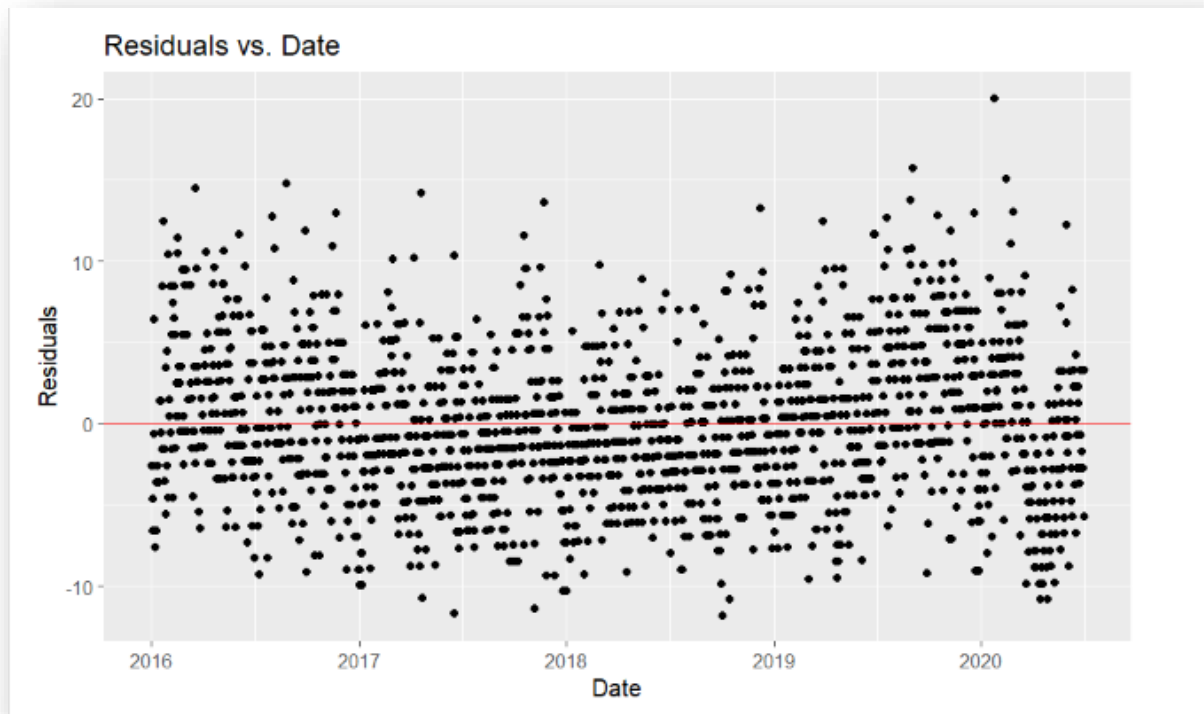


Fig 3.2 Residuals vs Date

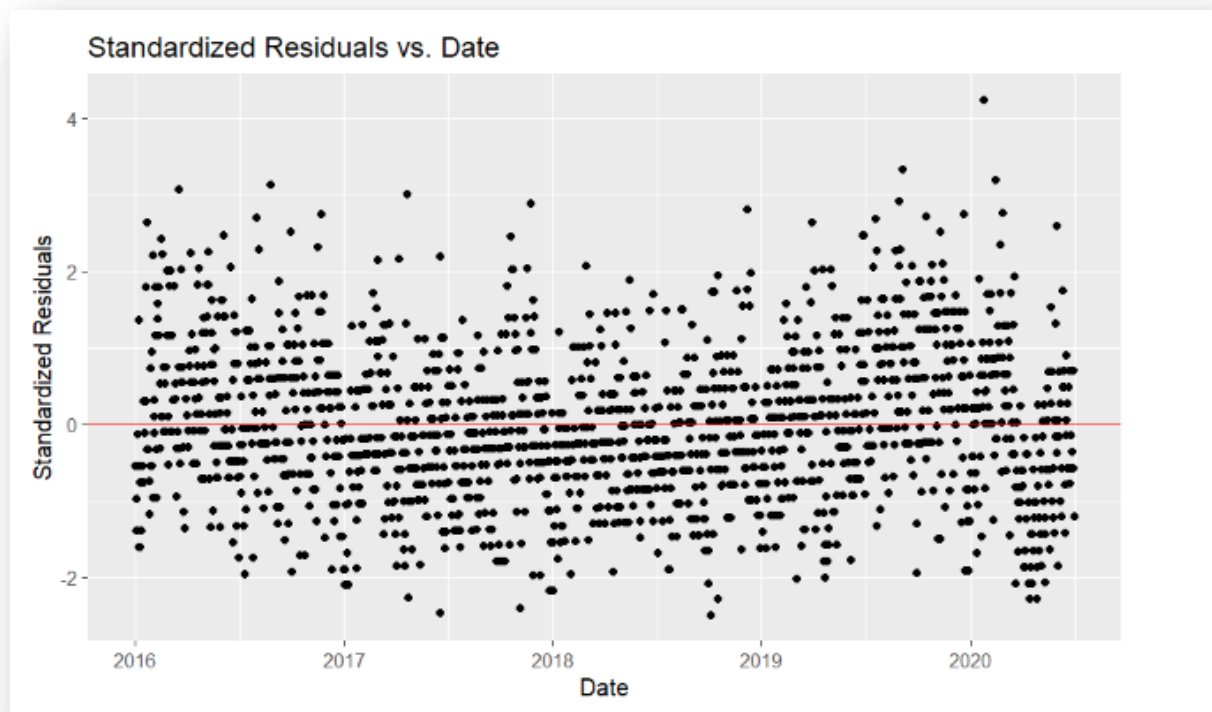


Fig 3.3 Standardized residuals vs Date

By observing the plot above, linear model doesn't seem to fit the datapoints well. Hence, some other models must be explored and experimented. Residuals are the difference from the observed value to fitted value.

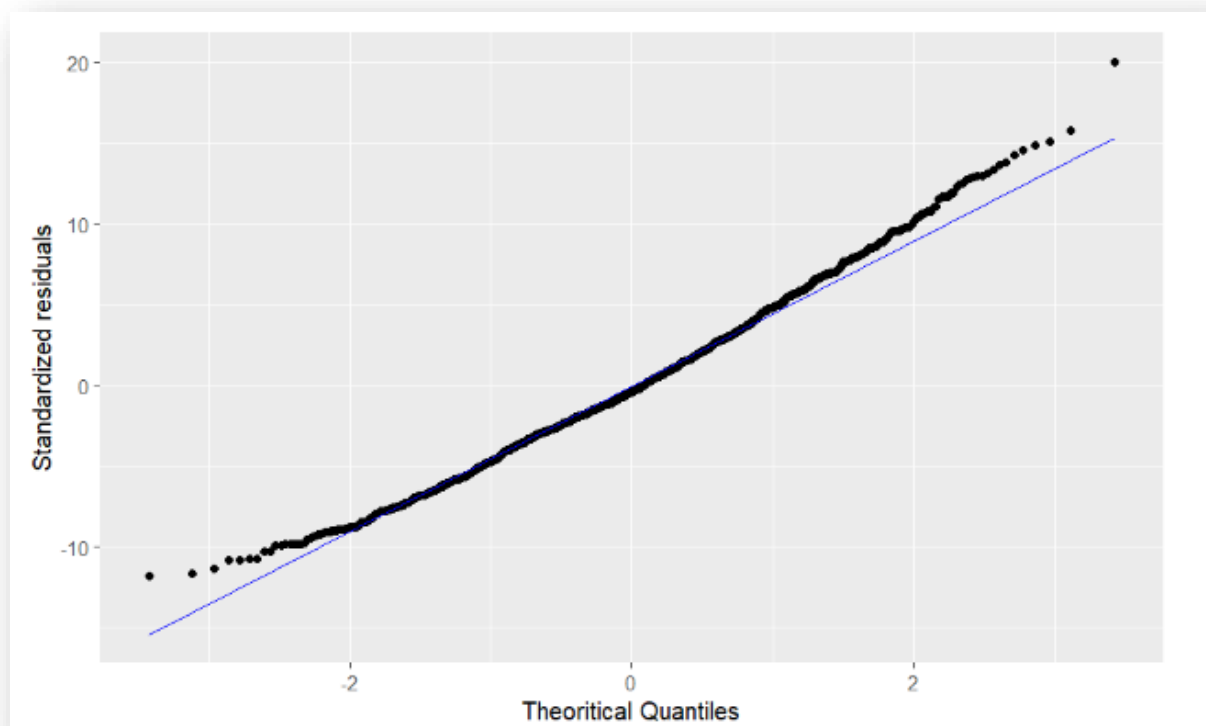


Fig 3.4 QQ plot of linear model

The QQ plot (Fig 3.4) above implies that one or more assumptions of linear regression may have been broken, there are outliers in the data, there is a non-linear connection between the independent and dependent variables, or other causes could be to blame.

Other potential justifications are –

Outliers in the data could exist. Outliers have the potential to skew model results and produce non-normally distributed residuals.

The relationship between the independent and dependent variables could not be linear. There may be a non-linear relationship between the independent and dependent variables. If the relationship between the independent and dependent variables is non-linear, then a linear regression model will not be able to capture the relationship accurately. Non-normally distributed residuals may result from this.

Q3.3 As we are not interested in the trend itself, relax the linearity assumption by fitting a generalised additive model (GAM). Assess the model fit. Do you see patterns in the residuals indicating insufficient model fit?

A generalized additive model is fitted to the model as linear model did not give out good results. Let us analyse the model fit using the model summary below (Fig 3.5)


```

Family: gaussian
Link function: identity

Formula:
TOTAL_ACCIDENTS ~ s(date_numeric)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.1308    0.1118   126.4   <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(date_numeric) 7.689  8.563 23.24  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.108   Deviance explained = 11.2%
GCV = 20.647  Scale est. = 20.537      n = 1644

```

Fig 3.5 model summary

- The adjusted R^2 value is 0.108, implying that the model explains about 10.8% of the variance in the response variable. (This is a low proportion of variance is explained, but that is the data and the GAM model we have in hand)
- The model explains 11.2% of the deviance, showing a moderate explanatory power.
- Generalized Cross Validation value is 20.647, providing an estimate of the model's prediction error. (Lower GCV values indicate better fitting model)
- The scale estimate is 20.537, providing an estimate of the error variance. This can be used to assess the goodness of fit.

The statistical significance of the terms and the moderate explanatory power suggest that the model is capturing meaningful relationships between the date and the total number of accidents.

As per the plot below (fig 3.6) we can the goodness of the fit. The model is able to capture non-linear relationships in the data

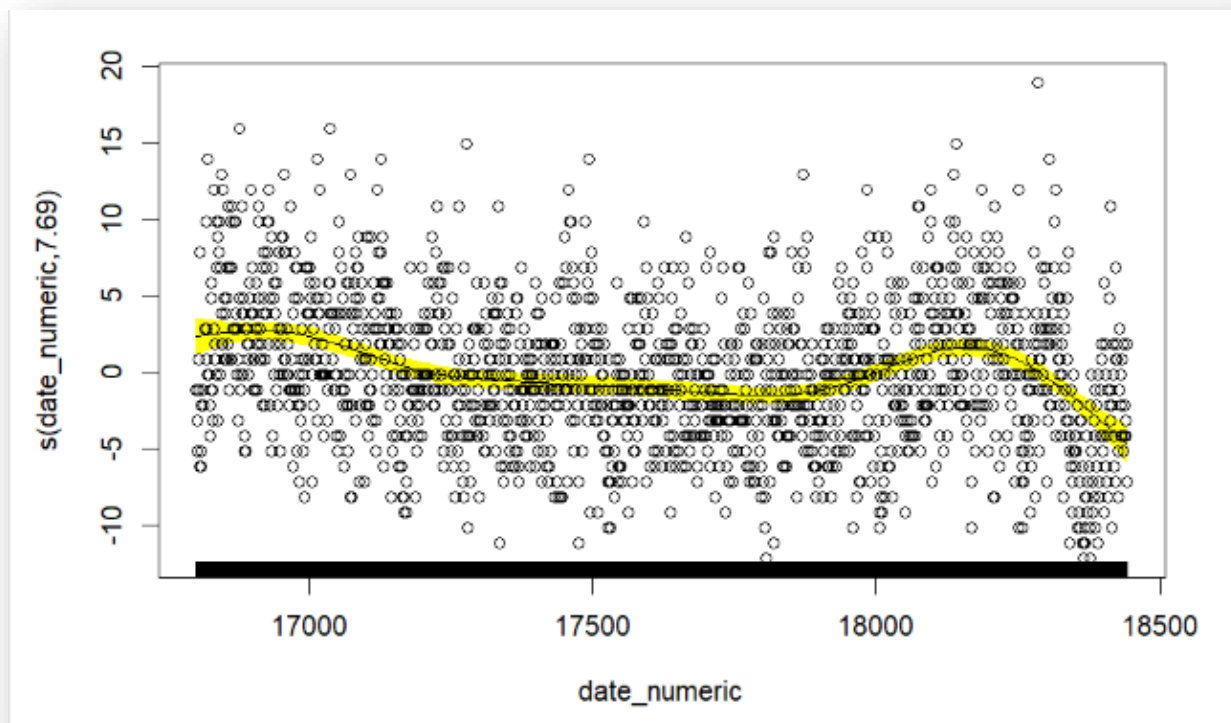


Fig 3.6 Gam fit plot

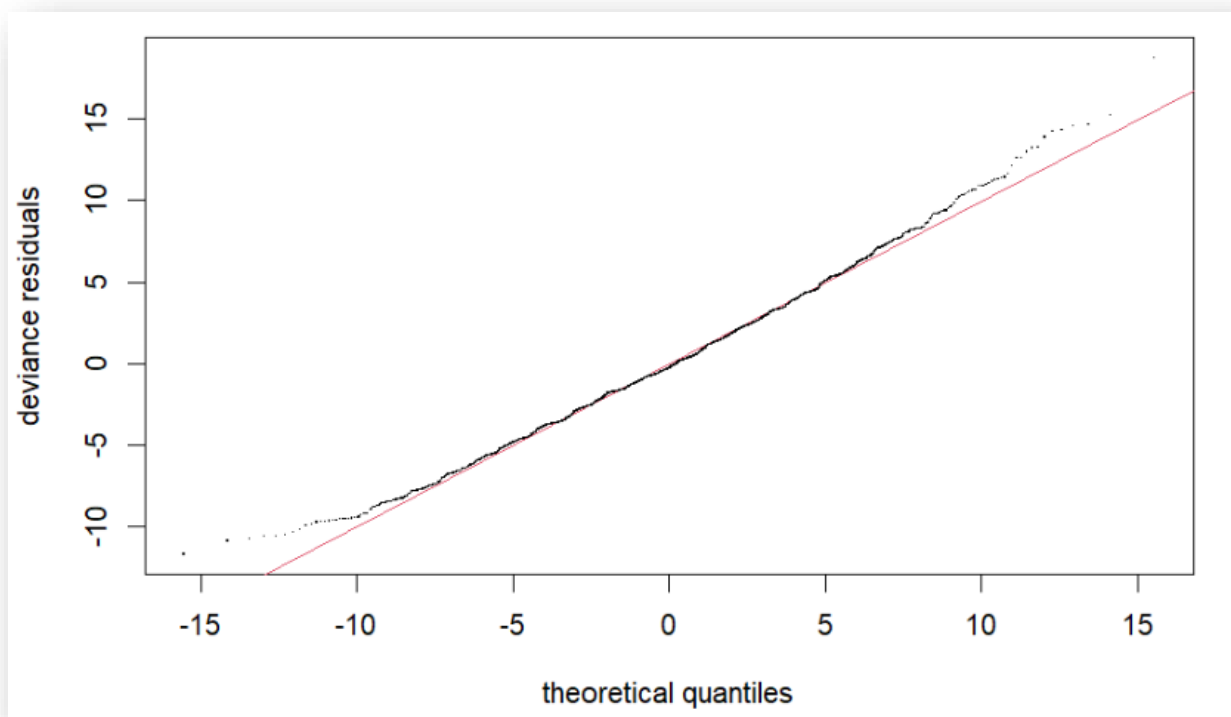


Fig 3.7 Deviance residuals vs theoretical quantiles

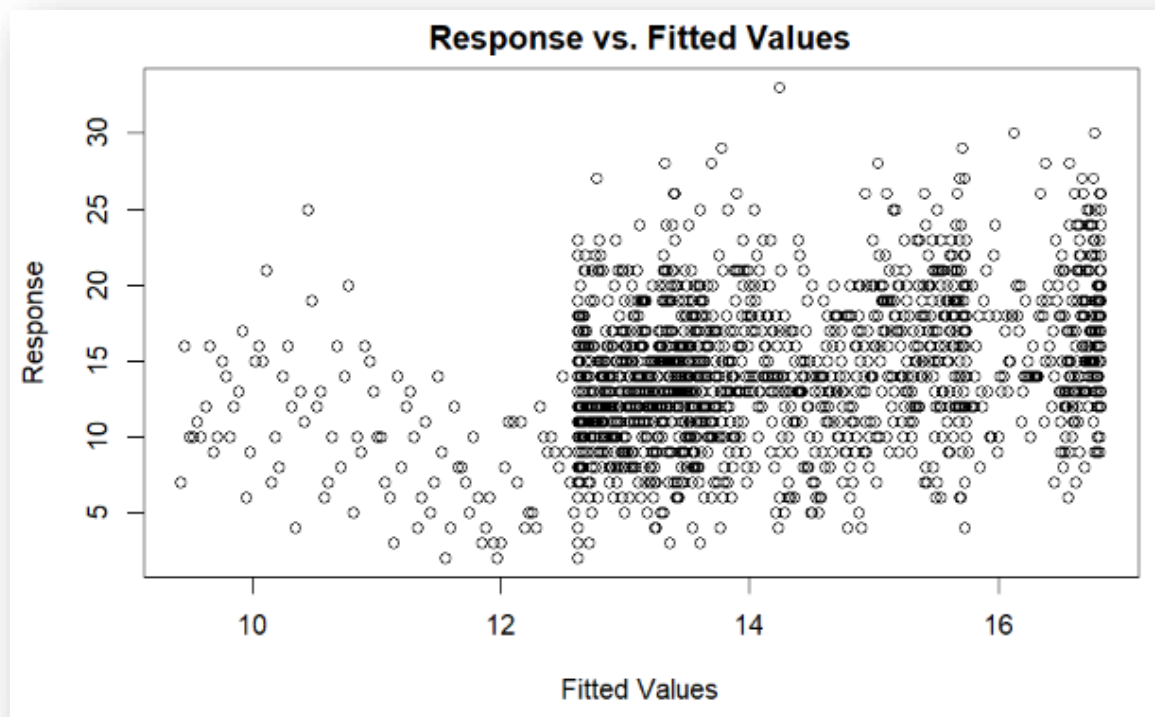


Fig 3.8 Response vs Fitted values

The above residual plot (fig 3.8) reveals that, When the residuals are randomly scattered (indicating no systematic overestimation or underestimation by the model), and there's an even spread of residuals across the range of fitted values, it implies that the model is doing a good job of capturing the underlying patterns in the data.

Lets break down the graph above. (fig 3.8) – Loosely scattered plot before 12.5 in x axis, the plot of response (actual values) against fitted values shows a loose scattering. This means that it does not closely match the actual number of car accidents in this period. The model might not capture the underlying pattern well in this region.

Tightly scattered plot after 12.5 in x-axis, the tight scattering of points after the change point indicates that the model predictions closely align with the actual number of car accidents during this period.

```

Method: GCV  Optimizer: magic
Smoothing parameter selection converged after 6 iterations.
The RMS GCV score gradient at convergence was 5.386008e-06 .
The Hessian was positive definite.
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(date_numeric) 9.00 7.69    0.82 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig 3.9 Optimizer: magic, Method GCV

Smoothing parameter selection process was used, and the model achieved convergence. The basis dimension chosen for the smooth term is appropriate based on the basis dimension checking results. The small p-value associated with the k-index confirms the suitability of the chosen basis dimension.

Q3.4 Augment the model to incorporate the weekly variations.

To have the weekly wise variations, we need to transform the date to week days. Using the library lubridate, mutate and get the data week wise

```

Family: gaussian
Link function: identity

Formula:
TOTAL_ACCIDENTS ~ s(date_numeric) + s(week_day_numeric, k = 7)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.1308     0.1059   133.5   <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(date_numeric)  7.788  8.622 25.73  <2e-16 ***
s(week_day_numeric) 5.567  5.926 31.95  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.2   Deviance explained = 20.6%
GCV = 18.587  Scale est. = 18.425    n = 1644

```

Fig 3.10 Gam summary for weekwise data

We can see the increase in the deviance from ~11% from the gam model without augmentation to 20.6% in the gam model. 20.6% of the deviance in the response variable is explained by the model.

GCV is an estimate of the prediction error and is used for smoothing parameter selection in GAMs. Here GCV value is 18.587. [6] In particular, it has been shown that the addition of imperceptible deviations to the input, called adversarial perturbations, can cause neural networks to make incorrect predictions with high confidence

The smoothing terms (day numeric) have a strong relationship with the response, and the model's goodness of fit is reflected in the adjusted R-squared value. The GCV helps in smoothing parameter selection, and the scale estimate provides an estimate of the model's scale or variability

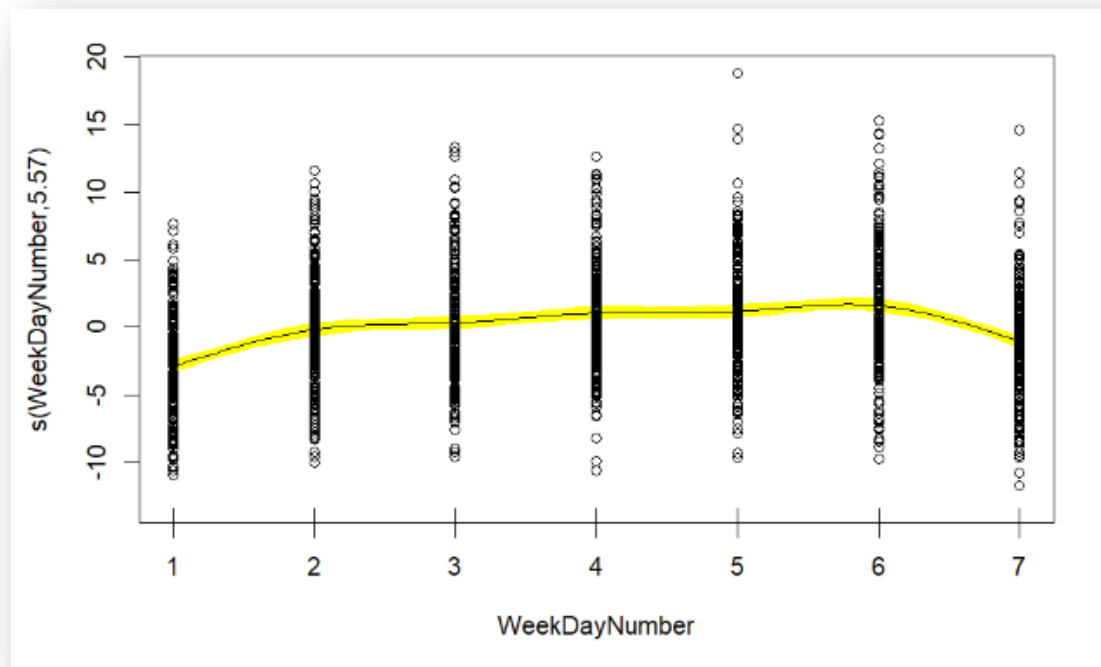


Fig 3.11 Gam summary for weekwise data

Q3.5 Compare the models using the Akaike information criterion (AIC). Report the best-fitted model through coefficient estimates and/or plots

```
[1] "AIC score of Linear Model:"
[1] 9775.729
[1] "AIC score of GAM Model:"
[1] 9644.718
[1] "AIC score of GAM Model with weekly variance:"
[1] 9471.866
```

Fig 3.12 AIC results output of each model

The Akaike Information Criterion (AIC) score is a measure of the relative quality of a statistical model. Lower the AIC values better is the fitted model. On comparing, the AIC scores of linear model, the GAM model, and the GAM model with weekly variance, "GAM Model with Weekly Variance" has the lowest AIC value (9471.866), indicating it is the best-fitted model among the three.

Q3.6 Analyse the residuals. Do you see any correlation patterns among the residuals?

Let us compare the QQ plots of each model

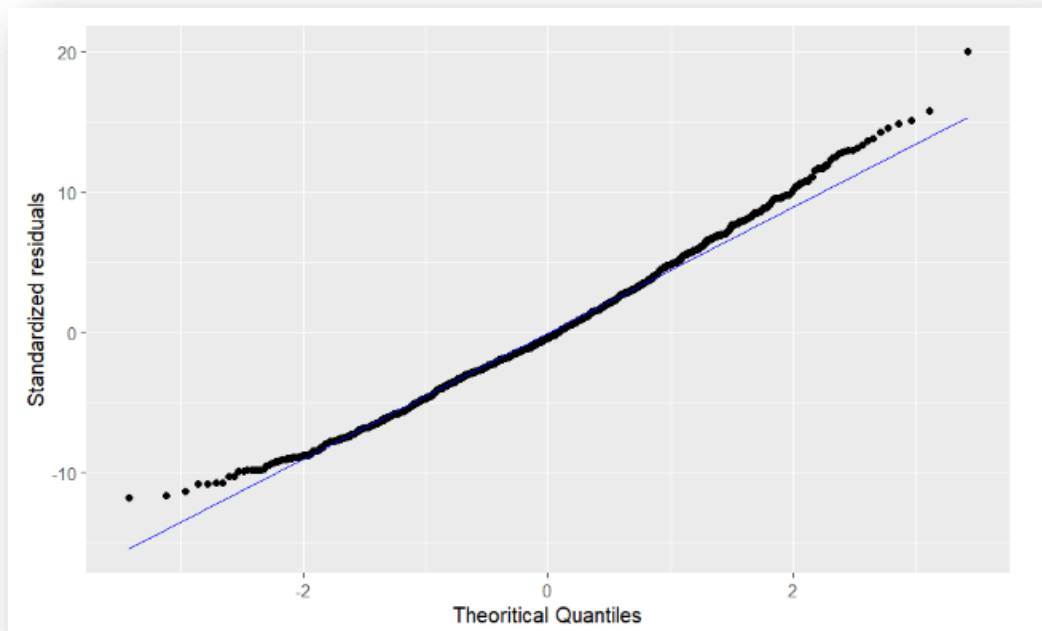


Fig 3.13 QQ plot of Linear model

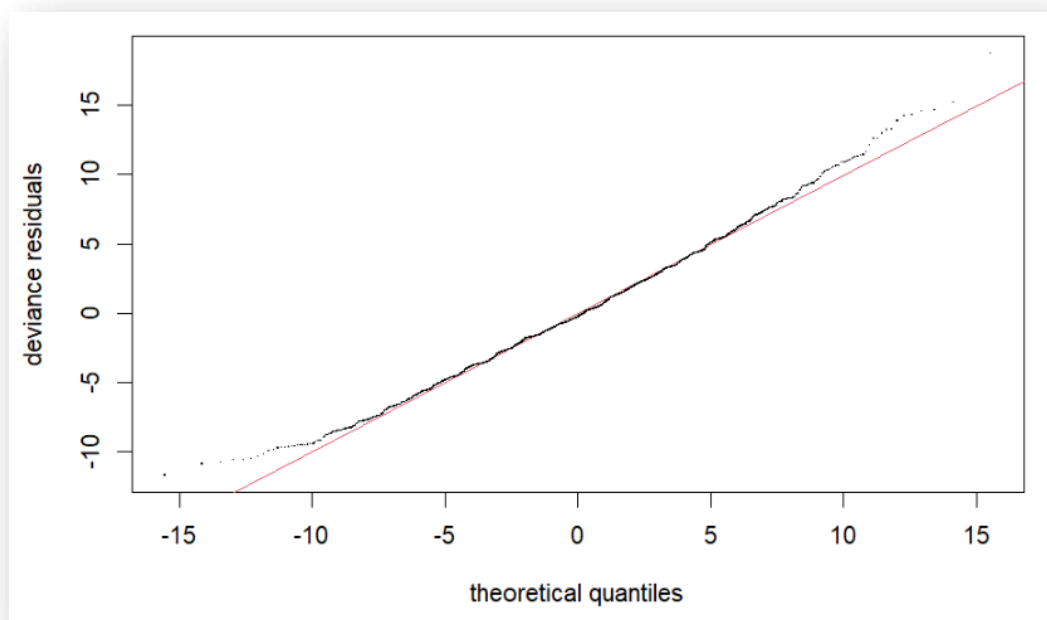


Fig 3.14 QQ plot of GAM model

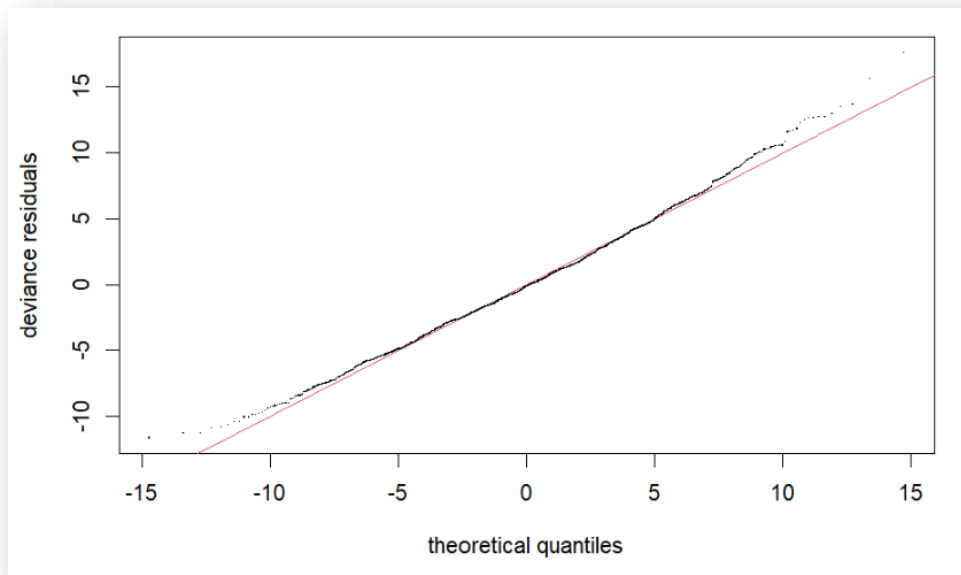


Fig 3.15 QQ plot of GAM model with weekly variance

From the plots above on fig 3.13, fig 3.14 fig 3.15 majority of the datapoints in the graphs tend to stick near the straight line, except a few deviations at the start and end.

Residuals Plots

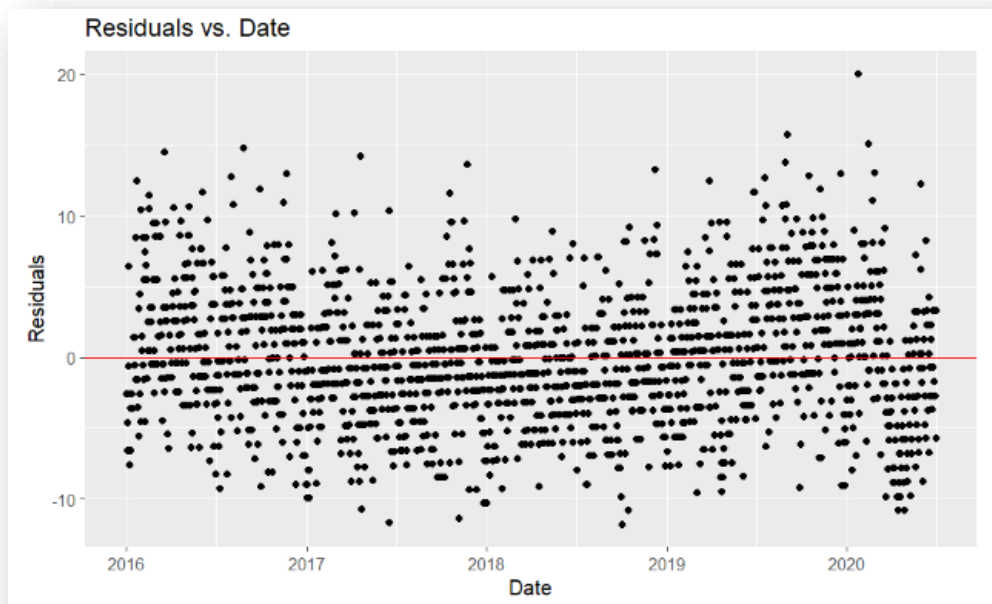


Fig 3.16 Response vs fitted values curve of linear model

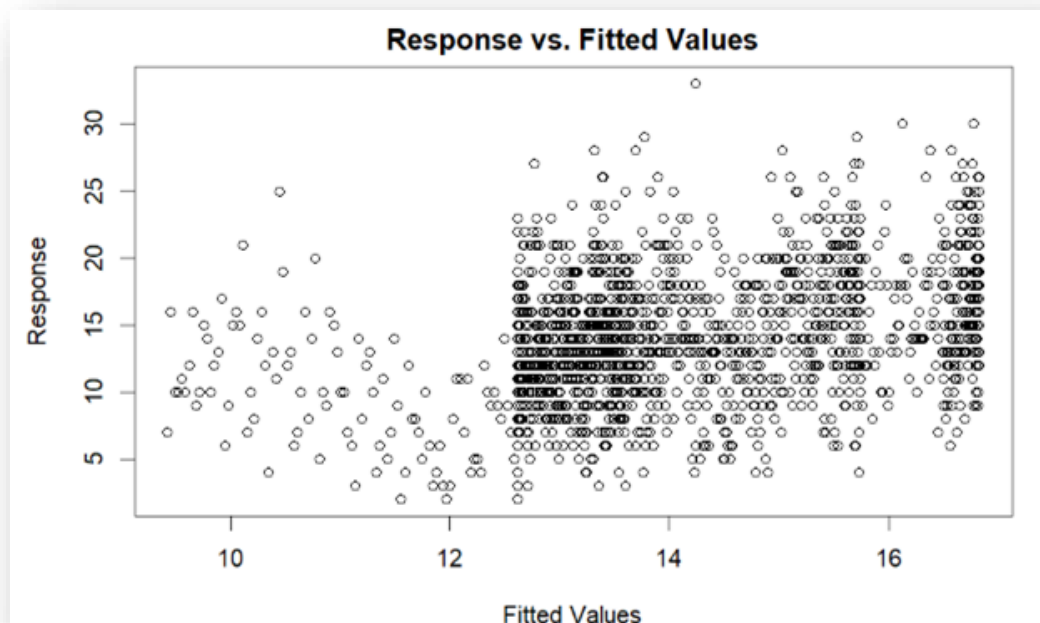


Fig 3.17 Response vs fitted values curve of normal gam_model

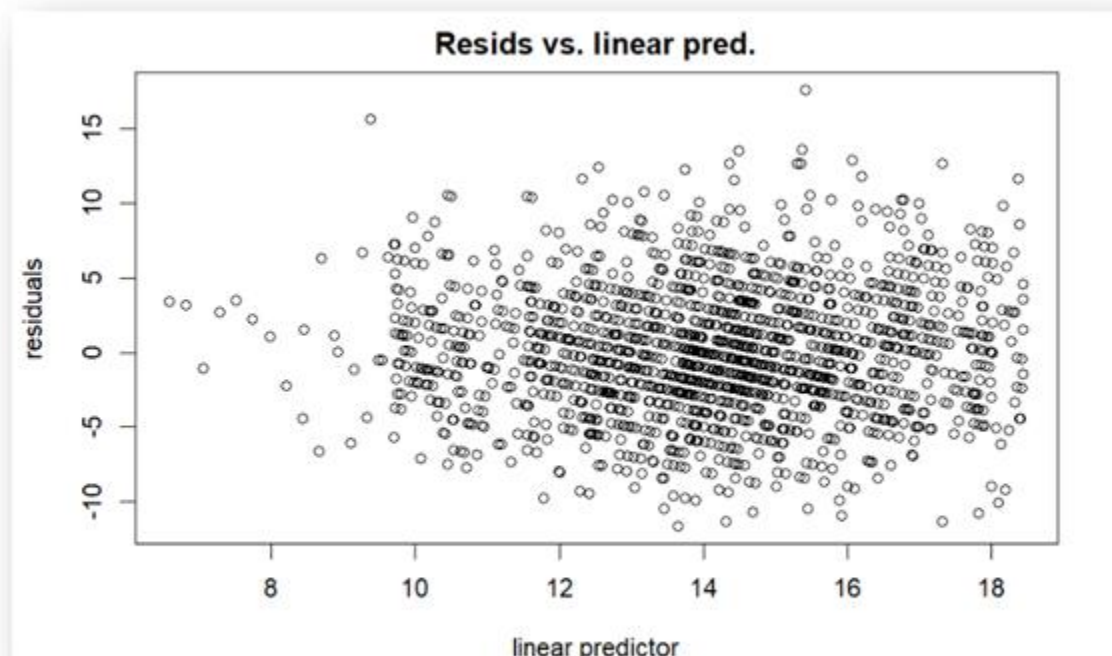


Fig 3.18 Response vs fitted values curve of normal gam_model weekday variance

Fig 3.16, 3.16 and 3.18 are the scatter plots of the residuals, there's an even spread of residuals across the range of fitted values, it implies that the model is doing a good job of capturing the underlying patterns in the data.

Histograms of Residuals:

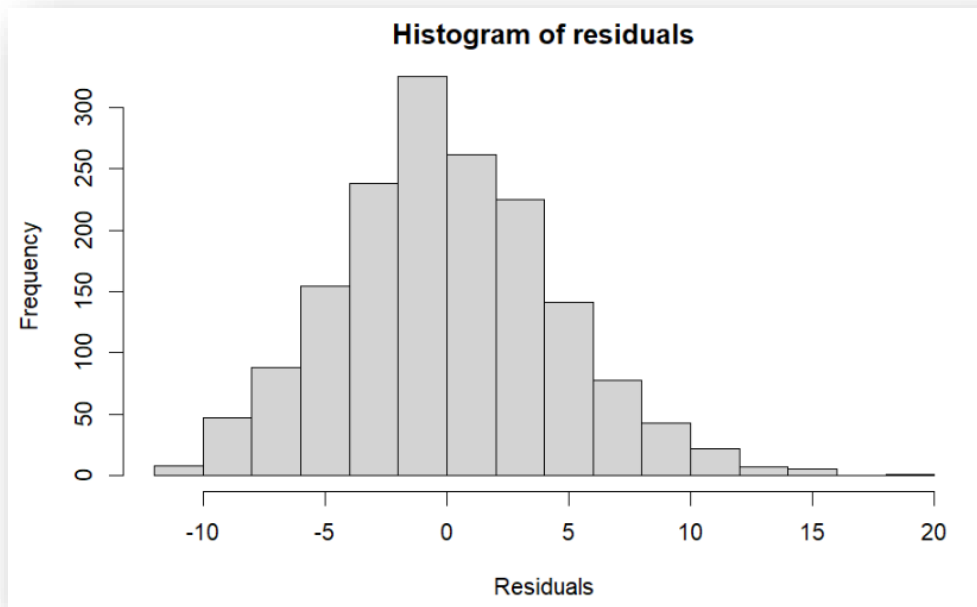


Fig 3.19 Histogram of the residuals of normal gam_model

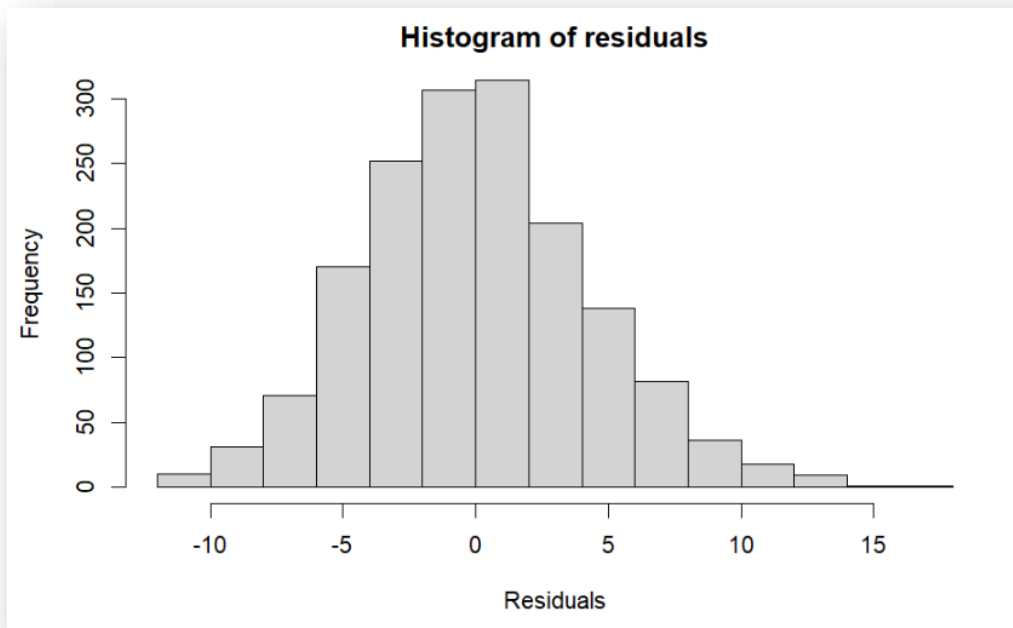


Fig 3.20 Histogram of the residuals of gam_model with augmentation

The histogram figures of the residuals obtained using gam.check above shows normal distribution (Bell curve).

Q3.7 What data type is your day-of-the-week variable? (3 points) Does the data type of this variable affect the model fit?

From the “Character” data type that was there in the original dataset, I had to convert them to numeric as only numeric columns are the ones suitable for such analysis. Because character datatype cannot be used to perform any kind of analysis or fit in the model. Hence, the data type of the variable does have an impact on the model fit.

Question 4:

Q4.1: Measuring heat wave. (3 points)

---This has been implemented in R notebook---

Key steps involved are,

1. Fetch the data of 30 years and am getting the temperature range using the TMAX and TMIN
2. Calculate the T95 value as mentioned in the paper

95%
182.3

3. Calculate the three-day average for the data I already had
4. Calculate the EHI (sig) excess heat index significance, by subtracting t95 from three days average.
5. Then I calculated Ehi(acclimatisation) using three-day average and 30 day average
6. Finally using the formula below, I calculated the EHI, as this formula is obtained directly from the paper itself.

$$EHF = EHI_{sig} \times \max(1, EHI_{accl})$$

EHF values for my dataset for the years 2016 to 2020

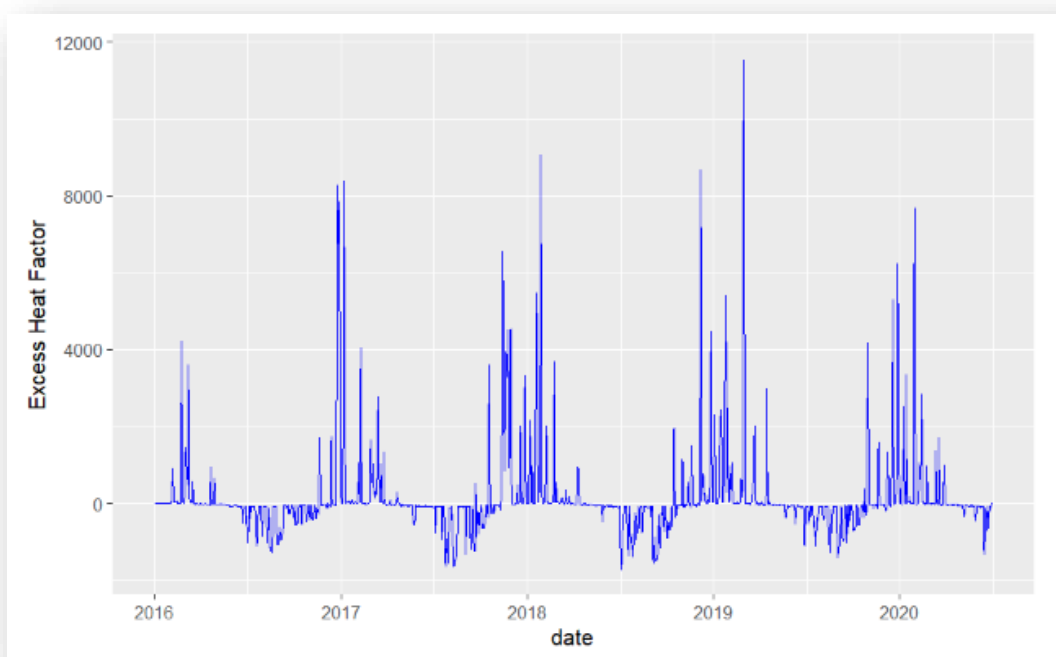


Fig 4.1 EHF values (2016 to 2020)

From Fig 4.1, The highest heat factor spike is observed in 2019. 2017 and 2018 share a similar peak whereas 2016 has the lowest heat factor. One point of view without an evidence would be that this could be due to the climate change, and 2020 was impacted by 2020 where certain media reported lower human activities helped the environment better

Q4.2: Models with EHF (3 points)

---The code has been implemented in R notebook---

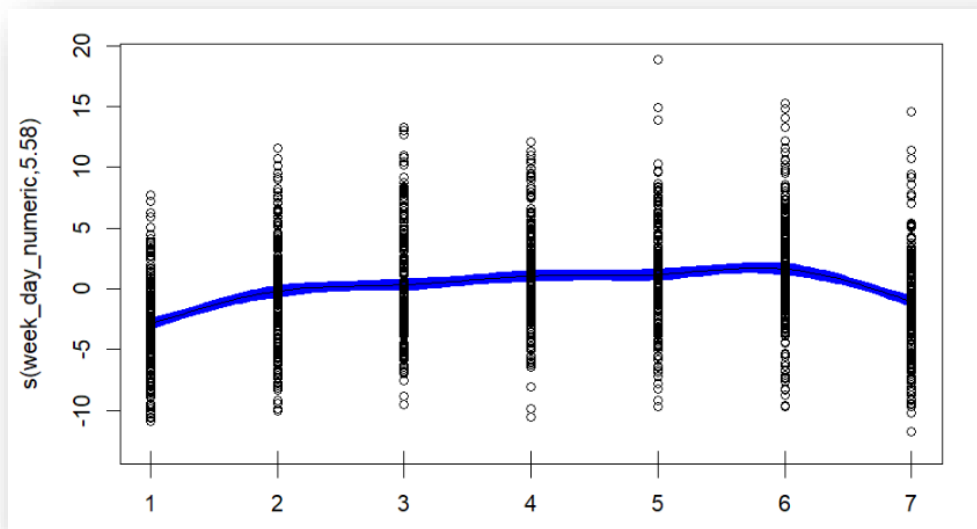


Fig 4.2

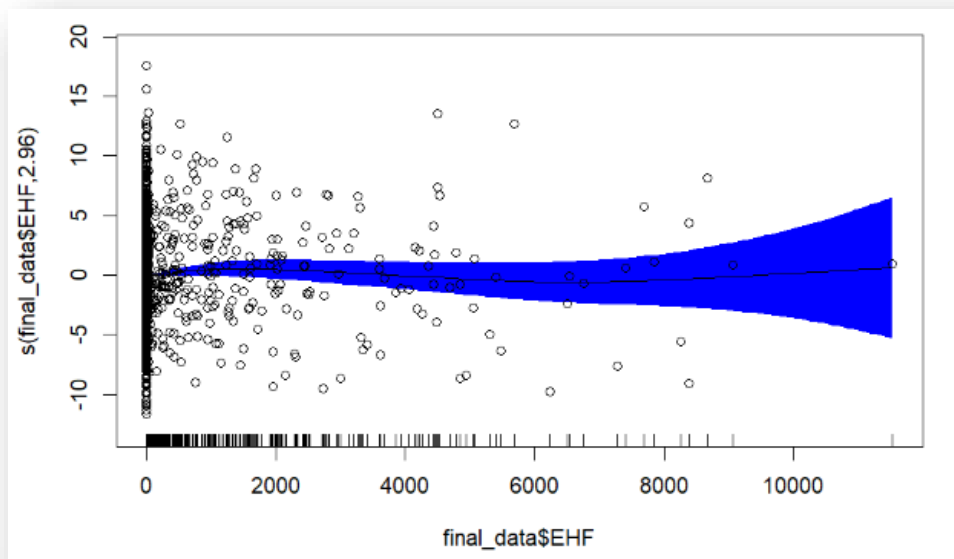


Fig 4.3 Inclusion of EHF in data

Fig 4.3 shows a non-linear relationship between EHF and the number of car accidents. This means that the relationship between the two variables cannot be described by a simple equation. However, the plot does show a trend, with the number of car accidents increasing as EHF increases.

Does the extra predictor improve the model fit?

Yes, not significantly, but the deviance has now increased a bit to 20.9%. The R-squared value is 20.1% (Read: Fig 4.4). And not much of a big difference in the AIC values as it is expected to not increase significantly.

```
Family: gaussian
Link function: identity

Formula:
TOTAL_ACCIDENTS ~ s(date_numeric) + s(week_day_numeric, k = 7) +
  s(final_data$EHF)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.1308    0.1058   133.6   <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(date_numeric)    7.920  8.682 25.865  <2e-16 ***
s(week_day_numeric) 5.582  5.931 32.005  <2e-16 ***
s(final_data$EHF)  2.955  3.681  0.763   0.496
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.201   Deviance explained = 20.9%
GCV = 18.598   Scale est. = 18.401    n = 1644
```

Fig 4.4 EHF model summary

```
[1] "AIC score of GAM Model with EHF is:"
[1] 9472.788
```

AIC Score – 9472.788

Other plots

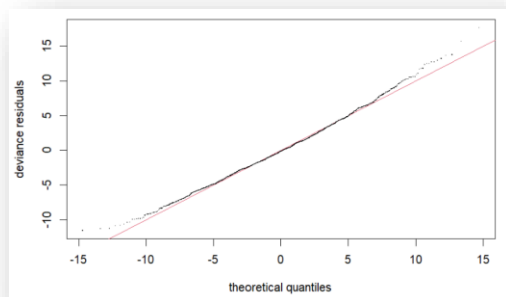


Fig 4.5 QQ plot

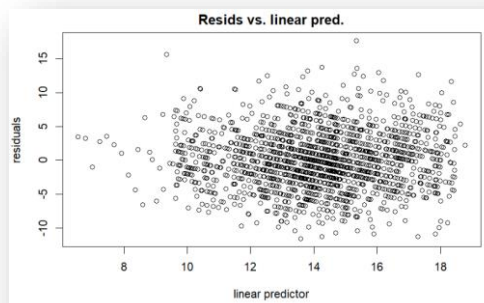


Fig 4.6 Residual Plot

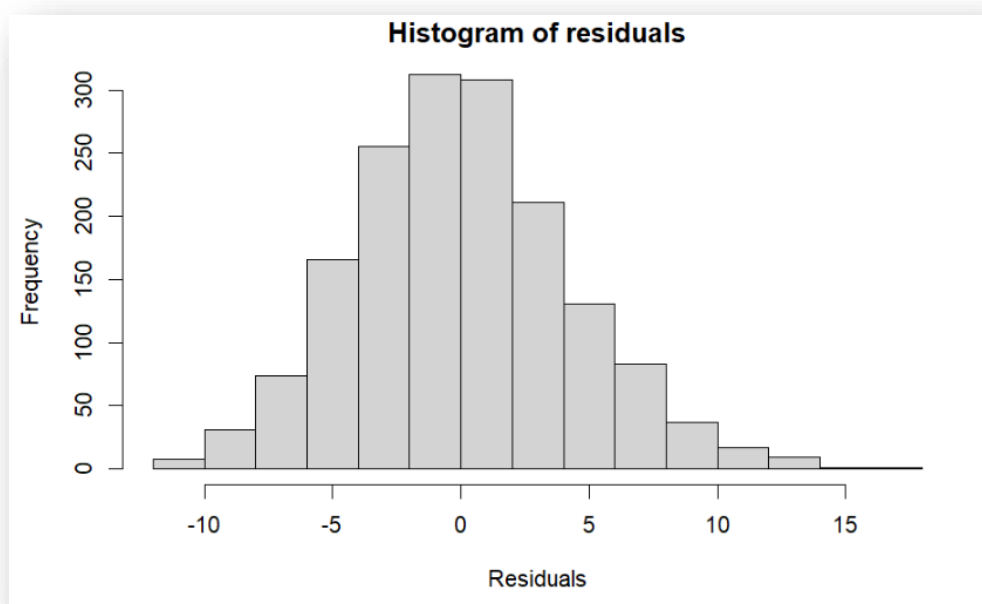


Fig 4.7 Histogram of the residuals

The histogram plot shows a good bell curve, proving that this follows a normal distribution.

What conclusions can you draw? (3 points)

Since we had a strong R-square result, improved deviance and the AIC score was good as well compared to other models combined, adding EHF to this analysis definitely improved the model performance. Additionally, this model contains no outliers.

The smooth term of the EHF is not as important as the date and weekday. Since model has improved to only a bit this cannot be concluded as the best model yet as well as considering the smooth term of EHF is not as vital compared to dates we used earlier.

[4] GAMs and other data-driven statistical models provide a useful, unbiased technique to forecast biomass and abundance across large geographic areas. However, until recently, the potential of GAMs for generating distribution maps for spatially detailed ecological models was underutilised.

Q4.3: Research question - extra weather features (15 points)

What extra weather features that may be more predictive of road traffic accident numbers?

- Traffic flow and the chance of accidents can both be severely impacted by poor visibility brought on by fog, mist, or other meteorological conditions.
- Road surfaces can be impacted by high humidity, especially in warm weather, which could have an impact on accident rates.
- Weather alerts or warnings, such as severe weather notifications, might offer useful information about potentially dangerous circumstances.
- Radiation - Particularly at dawn, dusk, or on bright sunny days, the intensity and length of the sun can have an impact on driver visibility and behaviour on the road.
- Wind direction and speed should be considered because they might have an impact on how well a vehicle handles and how visible it is on the road.

Plots from the extra weather predictor feature, that is temperature range.

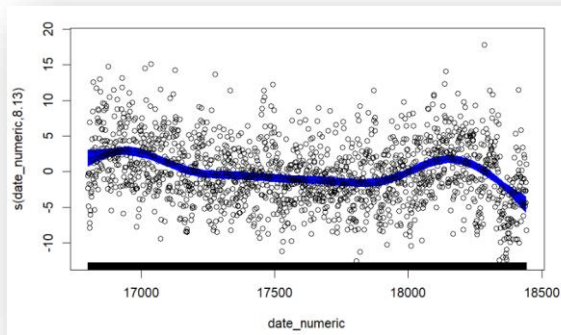


Fig 4.8

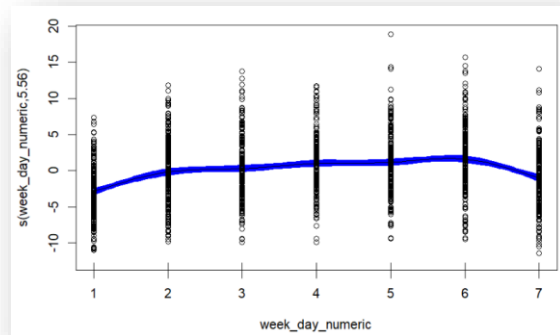


Fig 4.9

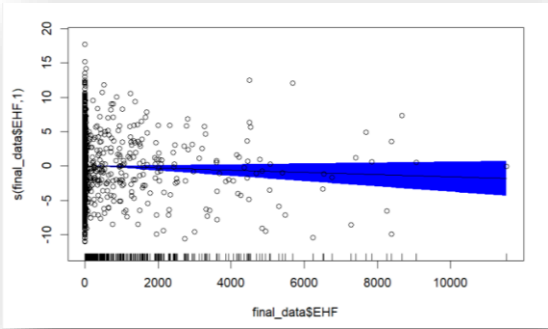


Fig 4.10 for EHF feature

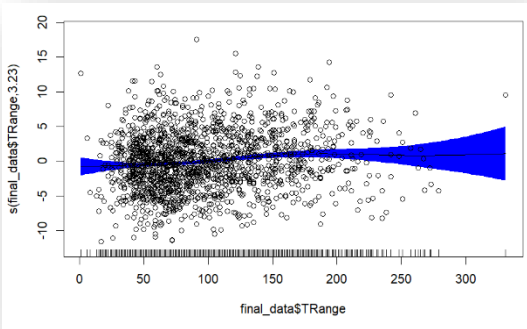


Fig 4.11 for Temperature range feature

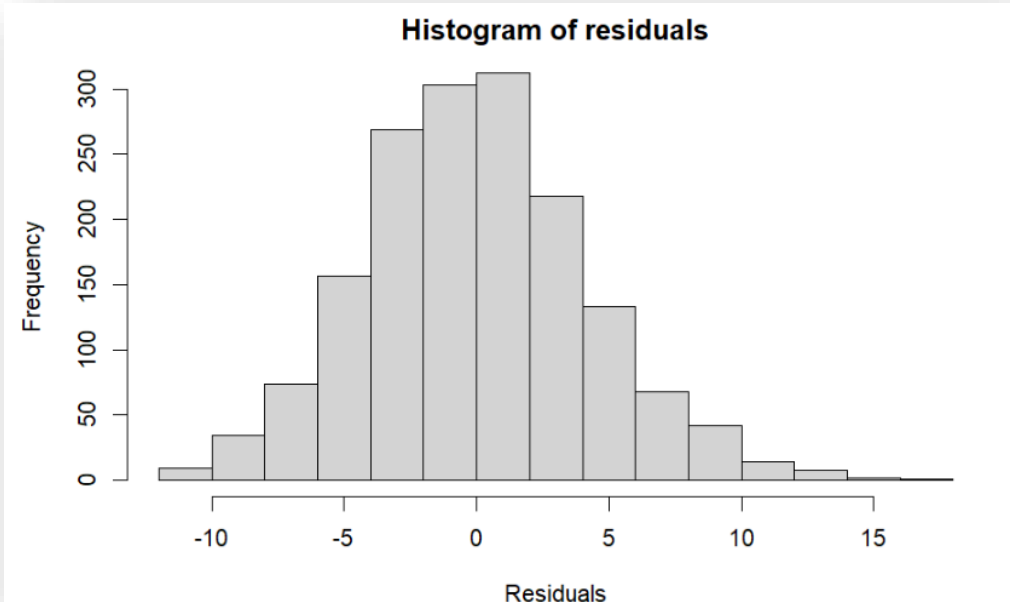


Fig 4.13 Histogram of residuals

Fig 4.13, Residuals follow a normal distribution as the histogram has a good bell shape

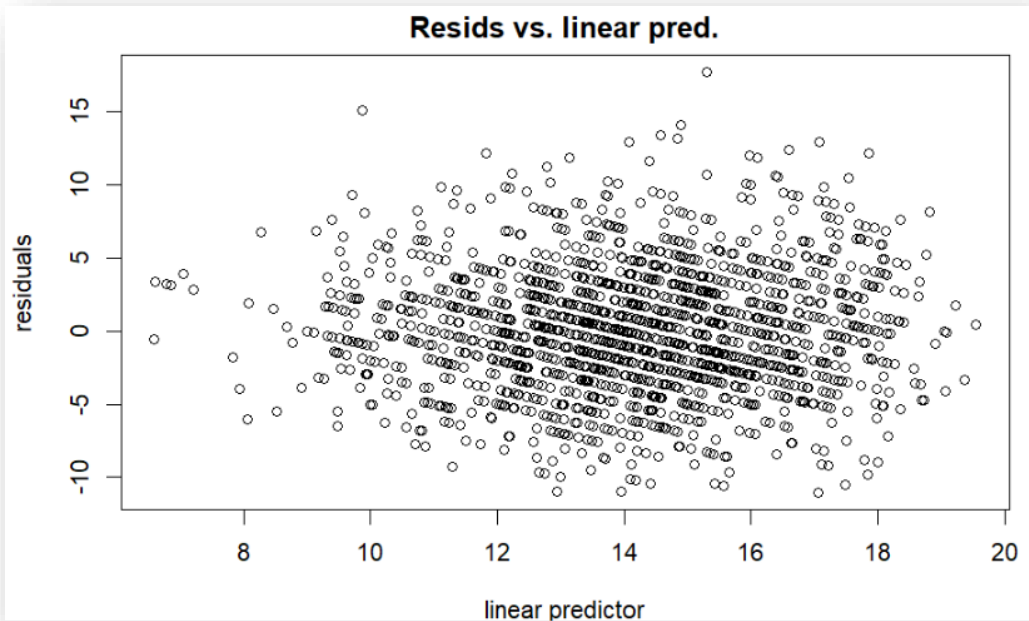


Fig 4.14 Residuals plot

```
Family: gaussian
Link function: identity

Formula:
TOTAL_ACCIDENTS ~ s(date_numeric) + s(week_day_numeric, k = 7) +
s(Final_data$EHF) + s(Final_data$TRange)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.131      0.105   134.5   <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(date_numeric)      8.128   8.788 26.638   < 2e-16 ***
s(week_day_numeric)  5.565   5.925 31.715   < 2e-16 ***
s(Final_data$EHF)     1.000   1.000  2.049    0.152
s(Final_data$TRange) 3.232   4.090  6.572  2.82e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.212  Deviance explained = 22.1%
GCV = 18.348  Scale est. = 18.137    n = 1644
```

Fig 4.15 Summary of the model

```
[1] "AIC score of GAM Model with EHF and Temperature range is:"
[1] 9450.497
```

Fig 4.16 AIC score

From the summary output, Deviance seems to be much better out of all, as it stands at 22.1%. AIC scores is a good measure of the goodness of the model and as per the paper, [5] The goal is to figure out how accurately models will predict new data when fitted to old. The criterion came to be called

the Akaike Information Criterion (AIC). And as per the AIC score shown, this has been the best model so far out of all the models being in this entire assignment with the lowest AIC score at 9450.497

Concluding that, as I observed so far, adding more and more weather features (predictor variables) seems to improve model step by step.

Question 5: Reflection**Q5.1: We used some historical data to fit regression models. What additional data could be used to improve your model? (5 points)**

From a data set point of view, there could be additional predictor variables such as

- a. Data on Traffic conditions - To better comprehend the effect of traffic on accident occurrences, data on traffic flow, congestion levels, road conditions (such as wet, icy, dry), construction activities, and accident-prone locations might be included.
- b. When analysing accident patterns, information on the types of vehicles on the road, driver behaviour (such as distracted driving and speeding offences), vehicle safety features, and vehicle age can be helpful to increase accuracy.
- c. [3] Road traffic accidents (RTAs) are among the life-threatening problems. Several factors contribute to RTAs ranging from human to technical and natural/environmental impacts. Anthropogenic air pollution and corresponding environmental factors also increase the probability of RTAs.
- d. The model's capacity to predict accidents based on expected weather conditions can be enhanced by using real-time or forecasted meteorological data (such as temperature, humidity, and precipitation projections) in addition to historical weather data.
- e. Insights into how population dynamics affect accident rates and severity can be gained by combining data on population density, demographics.

Q5.2: Regression models can be used for two main objectives: 1) understanding a process, and 2) making predictions.**a) In this assignment, do we have reasons to choose one objective over the other? (5 points)**

It is necessary to have both in balance to build a model with a capability predict accurately. Understanding a process is the main goal if the primary objective is to learn more about the dynamics and correlations between weather variables (such as EHF, temperature, and precipitation) and automobile accidents.

Relevant authorities can gain important insights by studying the influence of weather on accidents, identifying key variables, and understanding how these variables interact. This knowledge can direct the creation of safety precautions and targeted actions.

Making predictions is the objective if the main goal is to create a model that accurately forecasts the occurrence or severity of car accidents based on the weather and other relevant variables. Under various meteorological conditions, predictive models can be used to forecast accident frequency or severity. This can help with preventative actions like modifying traffic management, emergency services, or public awareness campaigns during bad weather.

b) How would the selection of one of these objectives affect our model? (5 points)

If creating the best model is taken as the only objective, the model would fail to capture the real underlying details. If understanding is the only objective, the strategies applied on the model would not make sense.

The development of a strong predictive model (objective 2) can be influenced by objective 1's understanding of the process. On the other hand, prediction models can also provide insight into the fundamental mechanisms underlying the link between weather and accidents.

In conclusion, a comprehensive strategy that provides both comprehension and predictive capacities for improved traffic safety and accident prevention may be possible by balancing the two goals.

Q5.3: Overall, have your analyses answered the objective/question that you set out to answer? (5 points)

Yes, the aim of this assignment was to first build a linear model to see the relationship between accidents and the dates. I analysed the relationship between them and learned that the model did not perform well (As the data had non-linear relationships).

This is when I moved to the next step of building and experimenting with another model that is GAM, which helps in capturing non-linear relationships as well. For each model, there has been thorough explanations of the graphs and various plots were drawn to understand the pattern and relationships.

The predictor variables used were minimum temperature, maximum temperature, precipitation and the temperature range along with the date and accident numbers. Additionally, EHIsig was computed by subtracting T95 of the historic data from the three-day average of my data. And the EHI (accl) was calculated by subtracting the three-day average with 30 day average and finally using the formula given in the paper, EHF value was found and used as an extra predictor in the model

And performance metrics and coefficients were analysed between the models including the AIC scores. As I have used the southeast region area and have analysed the accident data between them with deviance explained by each model, the model can be used by the authorities if the model fits their criteria of a valid model based on the metrics.

References:

1. Lam, L. T. (2004). Environmental factors associated with crash-related mortality and injury among taxi drivers in New South Wales, Australia. *Accident Analysis & Prevention*, 36(5), 905-908.
2. Nairn, J. R., & Fawcett, R. J. (2015). The excess heat factor: a metric for heatwave intensity and its use in classifying heatwave severity. *International journal of environmental research and public health*, 12(1), 227-253.
3. Hammad, H. M., Ashraf, M., Abbas, F., Bakhat, H. F., Qaisrani, S. A., Mubeen, M., ... & Awais, M. (2019). Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of Pakistan. *Environmental Science and Pollution Research*, 26, 11674-11685.
4. Grüss, A., Drexler, M., & Ainsworth, C. H. (2014). Using delta generalized additive models to produce distribution maps for spatially explicit ecosystem models. *Fisheries Research*, 159, 11-24.
5. Forster, M., & Sober, E. (2011). AIC scores as evidence: A Bayesian interpretation. In *Philosophy of statistics* (pp. 535-549). North-Holland.
6. Rebuffi, S. A., Goyal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34, 29935-29948.

The following libraries were used during the analysis in R notebook:

```
library(dplyr)
library(stringr)
library(ggplot2)
library(skimr)
library(rnoaa)
library(mgcp)
```