**SIT 741**

**Assignment 2**

**Task 1 : Source weather data**

1. I planned to use the **NOAA data source** since the Api calls could easily be made through R code and the data could be used in form of a data frame after using the names of the parameters in the code. The weather data could then be customized according to our stations, date and limit.

Whereas, the bureau of metrology required different runs of data and specifying different type of weather conditions. After that when we used to get the output for different conditions, they have to be combined in order to form a data frame which is much more tiring and requires a lot of steps if compared to download from the NOAA.

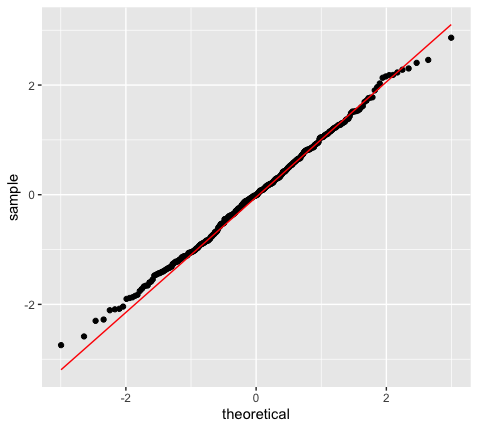
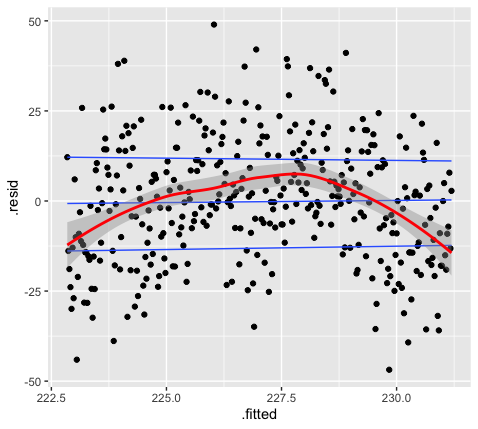
1. I downloaded the weather data for the station of ‘Perth Airport’ for the relevant time as the ED demand data in Assignment 1. The time period Is written below.
2. I divided the download of data into data frames, perth1 and perth2 since the data was more than of 1000 rows. Perth 1 covered the time period from 1st July 2013 to 31st December 2013 and had 736 rows. Perth 2 covered the time period from 1st January 2014 to 30th June 2014 and had 724 rows. I then combined both of the data frames into one.

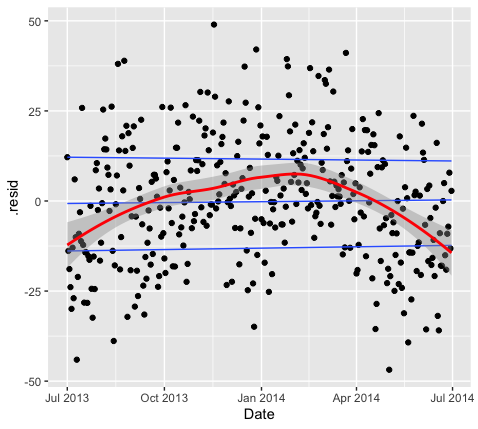
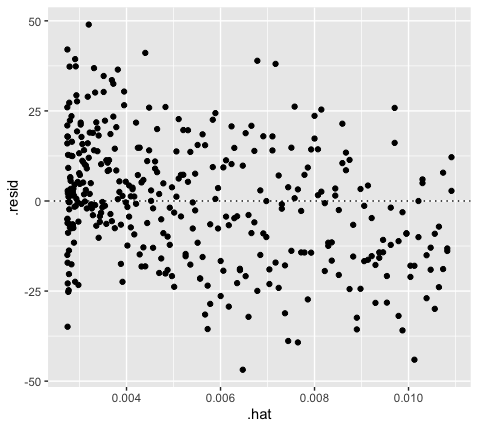
**Task 2: Model Planning**

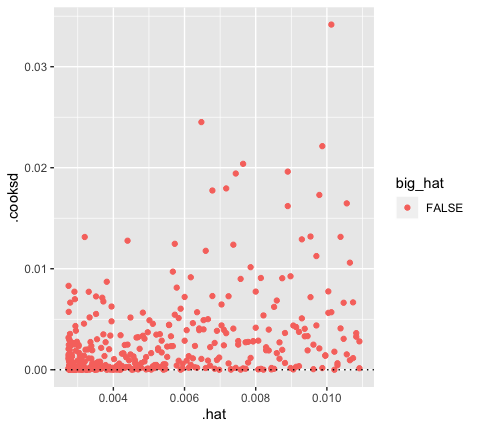
1. The final model will be able to justify the ED demands during different days and different EHF values. Through this model it would be easier to interpret which days the ED demand would be higher, which could be useful metric for hospitals and patients as well as the patients as they can then see which hospitals ED are more crowded and compare them. Moreover, if hospitals could understand which time and days were to have more demand for ED, then they could make pre arrangements for that in order to control the overcrowding. As explained earlier, the potential users for the model would be the ‘Hospitals’. Due to this knowledge from data, specific resources including nursing staff, housekeeping, doctors could be alerted beforehand.
2. We want to predict the relationship between the attendance of the Emergency department of hospitals along with the EHF(Excess Heat Factor) values for the area. The response variable for the model will be ***“Attendance”*** which will be taken from the Hospital dataset made in Assignment 1. The Attendance would be showing the number of people who came in the hospital, irrelevant of the fact if there they were admitted or not. The predictors would be the ***“date”*** of the Attendances, and the “***EHF”.*** Extreme Heat Value is calculated on the basis of weather temperature. The EHF values would be collected by the government and stored on various mediums, in a routinely manner. This data collected could then be used for predictions.
3. Yes, we are likely to build our model on historical data. The historical data is made up of temperature and precipitation and the number of people coming to hospitals. This defines a pattern, and it could be said that the data could have similar features in the future as it could be learned for instance which months and which weeks the weather could be like, along with the demand of ED.
4. There will be a couple of statistical methods that would be used in order to build the model, such as checking if a linear relationship exists between the response and predictors. Then GAM(Generalised Additive Model) is used to fit the degree of smoothness in the model. This smoothness is necessary before the model could be passed on to for regression analysis. By smoothing the predictor values, this function is expected to show a better relation if compared to the previous models relationsip.

**Task 3: Model the ED demands**

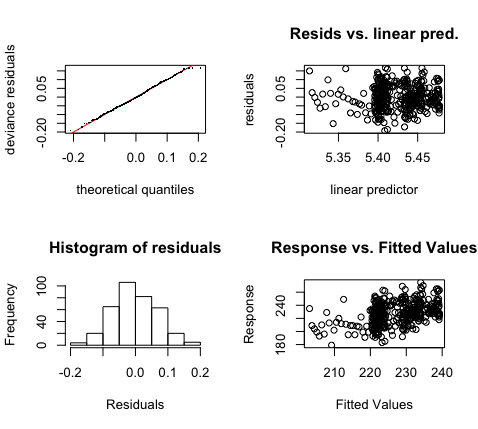
1. I selected the Royal Perth Hospital.
2. The linear model will be fitted on the Hospitals data with the Date as the predictor value. By fitting the linear model there were various things that could be interpreted. Starting from the R2 which is 1.9% which means that 1.9% of variations is explained by the variation of independent variables and the rest of them is due to the factors that are not included in the model, which takes us to the conclusion there more variables have to added in order for the model to have some predictive power. Moreover, at 5 % significance level the p- value of the model less than 0.5 which means the model was significant and Date does have some sort of predictive power in the linear model. Lastly the AIC is 3114. There were a couple of plots that were made in order to conclude that the linear model was not sufficient as per the plots.

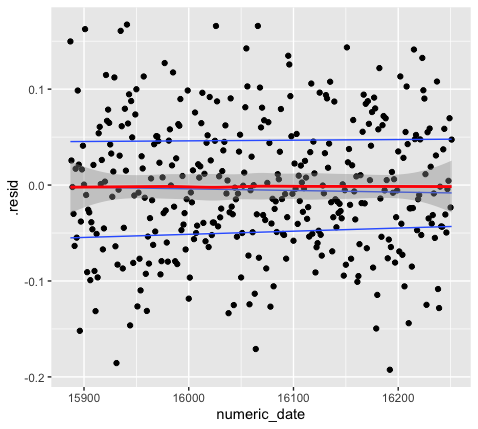




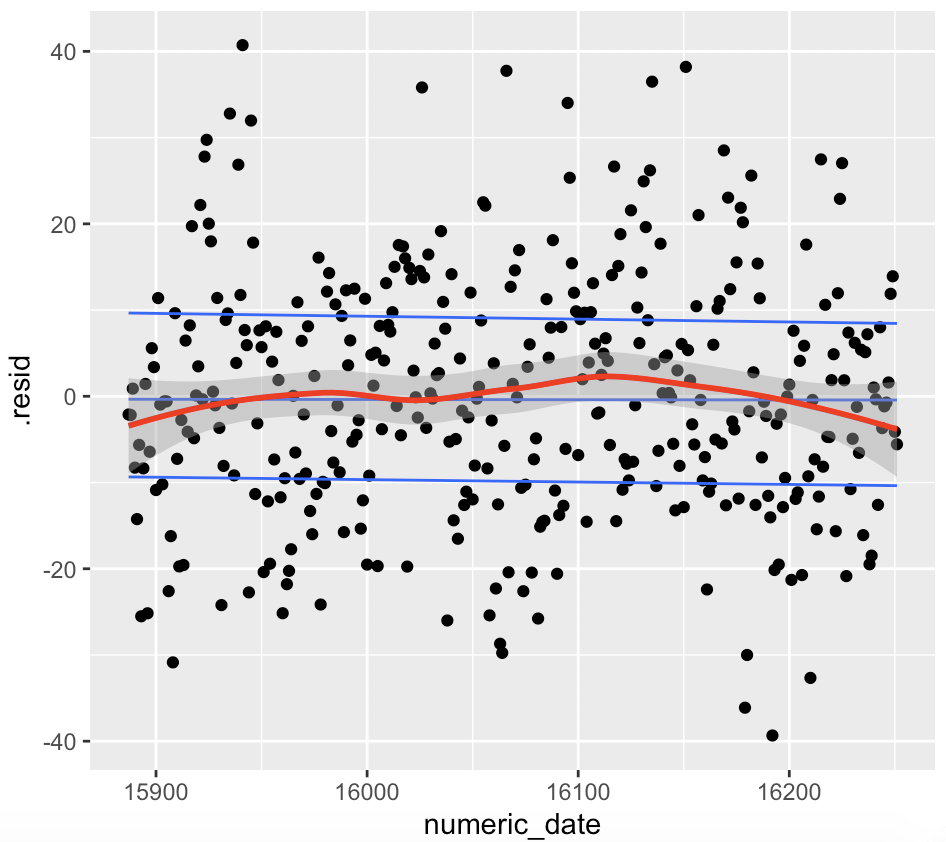


1. I fitted the GAM on the Hospital data with respect to Attendance in the hospital. After looking at this model it could be deduced that the R2 was 17.6% which was a lot better than the previous model. This means that 17.6% of the variations is explained by the variation of independent variables, whereas, the rest of the variation is due to the factors not included in this model. Lastly the value for AIC is 3057 for this model which is lower than of what was for the linear model, which indicates that this one was a better fit. The histogram below shows that the distribution is normal and it could be concluded that the residuals follow a normal distribution. The residual plot against the date shows that that the residuals seem to be independent along with the Date as the predictor.





1. The weekly seasonality was fitted and then augmented and picture is below for the augment weekly gam fit. The R2 for this model is 29.5% which means that 29.5% of the variations is explained by the variations in the independent variables, and the rest of the variation is explained by factors not included in the model. The AIC for this model is 3013. Based on the models R2 this model has more predictive power due to the variable of weekly seasonality and since AIC is lower than the previous model, which means this model was better fit.

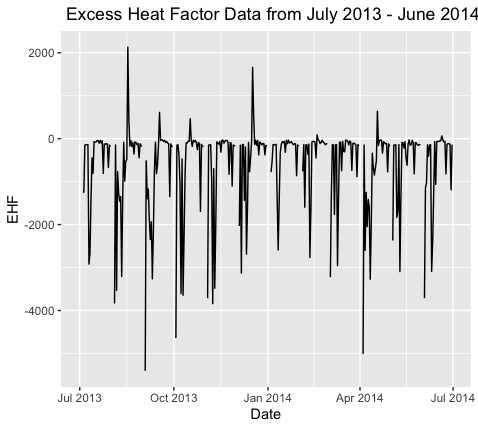


1. Seeing the residuals plot in the plots of the models, I would say there was no correlation between the residuals since the models weren’t that powerful.
2. The day of the week variable is a categorical variable as the class type of this variable is factor. The results would have been different if it wasn’t categorical.

**Task 4 – Heatwaves and ED demands**

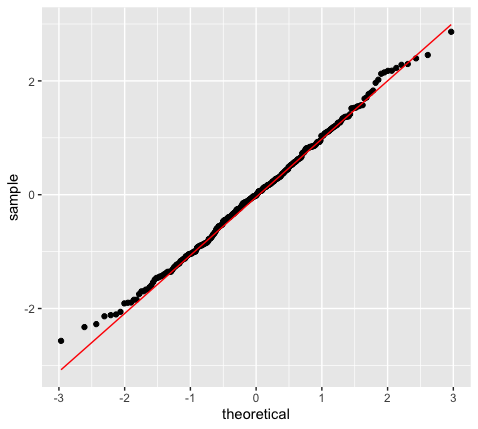
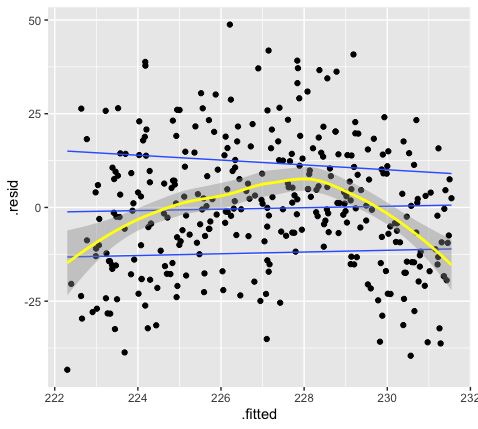
4.1 : Measuring Heatwaves

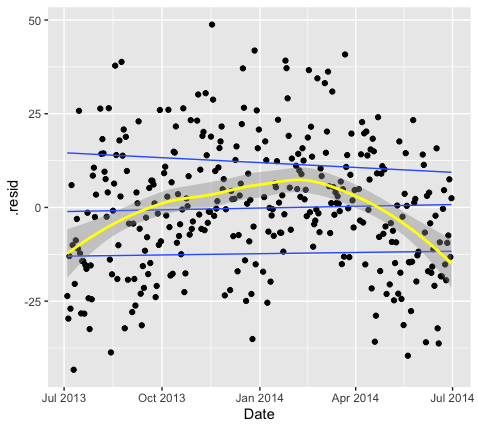
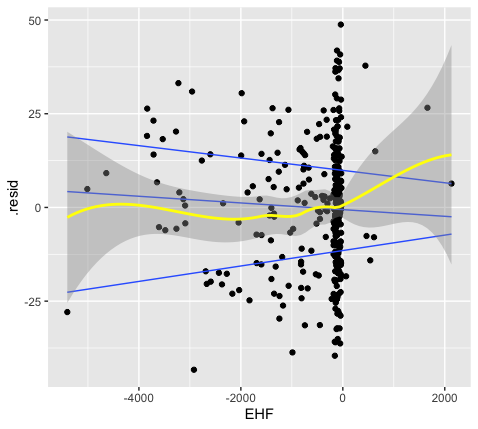
I measured the heatwaves buy the metric called Excess heat factor. As seen below the months that had an EHF more than 0 were September 2013, December 2013 and May 2014.

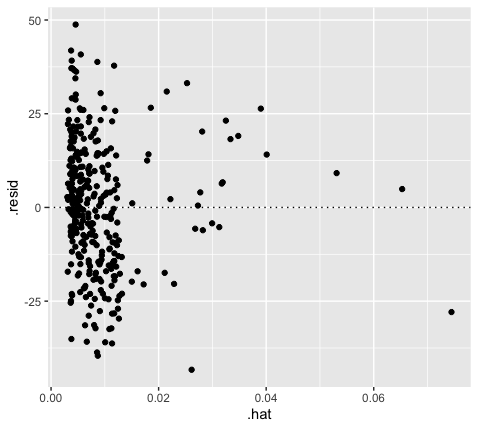
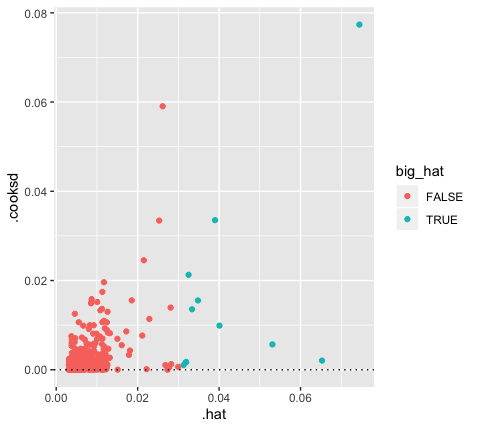


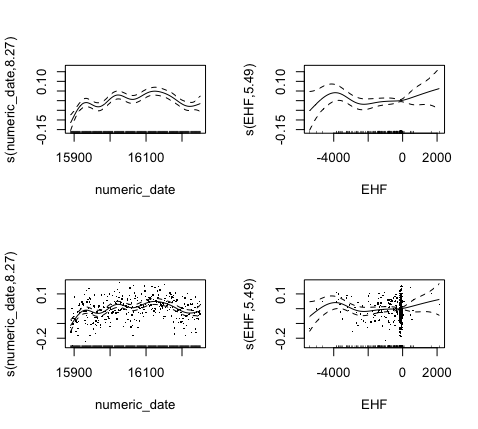
4.2 : Models with EHF

I added the EHF Values and then fitted the Lm and GAM models with the additional predictor. Firstly, after running the linear model, with EHF as an additional parameter the multiple R2 was 0.02 which means that 2.3% of the variation is explained by the variation in the independent variables and the rest of the variation is explained by factors not included in the model. The value of R2  Is higher than of the value was for R2 with Date only as a predictor Value. Also, the AIC value is 2814 which is lower than the linear model before which means that with respect to the linear model previous model was a better fit with only one predictor. Now running the GAM model on it. The R2 Value is 19.9% which means that 19.9% of the variations is explained by the variations in the independent variable, and the rest is explained by factors not included in the model. The AIC value for the GAM model with EHF as an additional parameter is 2783. After comparing the R2 and AIC values with the model which only had Date as a predictor, it turned out that the predictor power of the that model was better but the GAM model that was fitted with EHF parameter was fitted better.









4.3: Extra Weather Features

I incorporated the weekly seasonality into the model and fitted the GAM model. After adding the weekly seasonally into the EHF, the model had the R2 value of 0.337 which means that 33.7% of the variations is explained by the variations in the independent variables and the rest of the variations is explained by the factors not included in the model. The deviance explained was 40.7% which is a good value. So far, this was the best R2 Value and deviance explained for all the models compared which means that EHF does has some predictive value. Also, the AIC was 2739 which is lower than the model with only seasonality, which means this was a good fit. Lastly looking at the coefficients it was learned that the Days and EHF have a negative relation with the Attendance of the Emergency department.

**Task 5 : Reflection**

1. Historical data for the weather helped us to build models and predict certain things like seasonality and the relations. The models are generated through the historical data showing how fit the model is and whether it has some sort of predictive power or not. However, with such historical data there tends to arise some problems that the older the data becomes it tends to get outdated and could make inaccurate predictions then, for instance due to global warming there might be a huge rise in temperatures in some region but if this rise wasn’t there years ago than the historical data might be giving inaccurate predictions.
2. Regression models helps us with understanding the process and making predictions. In this assignment we were more focused towards understanding the process as we had to see the outcome of the models and understand the coefficients and independent variables, and the relation of independent variables with dependent variables. We studied the trend of different analysis and processes included the features related to date and EHF. Now considering that we still have to use the regression to understand the process then the it won’t be affecting the models, because it would only be affecting the model if our decision was to make predictions.
3. The analysis, I would confidently say that my analysis helped my back up my interpretations regarding the models and scenarios that were asked. It started with building the linear model and analysing its residuals. Then fitting a GAM model and analysing the residuals and coefficients. Then Augmenting the weekly seasonality and fitting the model. After that I added Excess heat factor as an additional parameter in the data and ran the linear and GAM models to see if the additional parameter helped me improve my analysis or not. Then added the feature of weekly seasonality to see its effect in the model.