STATISTICAL DATA ANALYSIS

Assignment 2

NOTE: All questions below are italicized and underlined for an easy understanding and locating the answers.

Task 1: Source weather data

1. *Which data source do you plan to use? Justify your decision.*

   I am going to use NOAA Climate dataset. I chose it because it is convenient to integrate the download into R coding. It means we can get the data dynamically during runtime. Secondly it comes with cleaner format and have the features we need. It just needs a few simple steps to clean up the dataset.

2. *Download the dataset (see R codes)*
3. *How many rows are in the data? What time period does the data cover?*

   The data contains 365 rows, equivalent to a one-year-time period from 2013-Jul-01 to 2014-Jun-30.

   Summary results of temperature data taken from Perth Airport station:
   [see the output (highlighted in yellow) below for references]

```
-- Data Summary ------------------------
                           Values
Name                       perth_temp
Number of rows             365
Number of columns          5
```

```
-- Variable type: Date -------------------------------------------------
-
# A tibble: 1 x 7
  skim_variable n_missing complete_rate min        max        median
* <chr>             <int>         <dbl> <date>     <date>     <date>
1 date                  0             1 2013-07-01 2014-06-30 2013-12-30
```

Task 2: Model planning

1. *How will the final model be used? How will it be relevant to the overcrowding problems at our EDs? Who are the potential users of your model?*

   The final model should be beneficial to as much users as possible. In this sense, it could be integrated into a universal app for mobile devices or webservice which easily access by many people.

   The model will be able to provide the foreseeable information of the facility regarding their current service capacity and help users to prepare themselves for queuing time they are going to face. In addition, users can target a different facility for a faster waiting time. This will help balance the overcrowding situation throughout the state.

   This model will be suitable for three types of users. The first type is the patients and their families. The model helps them to decides which facility is the most convenient for them. The second type is technical staffs who work at the facility. They can use the model to plan their resources based on the forecast demand. Lastly, the third type of user is government officials. Based on the model outcomes, the government can identify whether more facilities are needed for a certain region to reduce the pain of queuing time and ultimately improve healthcare system.

2. *What relationship do you plan to model or what do you want to predict? What is the response variable? What are the predictor variables? Will the variables in your model be routinely collected and made available soon enough for prediction?*

   The expected model is supposed to reveal the relationship between environmental factors and the amount of service demands needed for health facilities. Therefore, the response variable is considered to be the ED demands and the predictor variables are the climate elements.

   Regarding the data collection, there will not be any problem at all since the climate data is updated daily by local weather stations and can be retrieved instantly via web service. ED demands can be updated by technical staffs of the involved health facilities in regular manner or be retrieved from health department website. This way, the prediction of model is current and reliable enough to serve the daily need of users.

3. *As you are likely to build your model on historical data, will the data in the future have similar characteristics?*

   There is always variation between the existing data and the future data. We cannot expect 100% similarity between them, but the similarity becomes closer once we have collected more and more data of the past. In other words, large historical data can cover most of the outliers that may appear in the future.

4. *What statistical method(s) will be applied to generate the model? Why?*

   To predict the future outcome, we are using linear regression model as they can estimate the future trend of data distribution based on a set of training data (historical). The trend is normally located through the indication of relationship between the predictor and the target variable. If the predictor and the target variable are in relation, hypothetically regression model can reveal their bonding effect and be able to predict based on the change of their properties.

Task 3: Model the ED demands

1. *Which hospital do you pick?*

   Royal Perth Hospital is selected for the modeling.

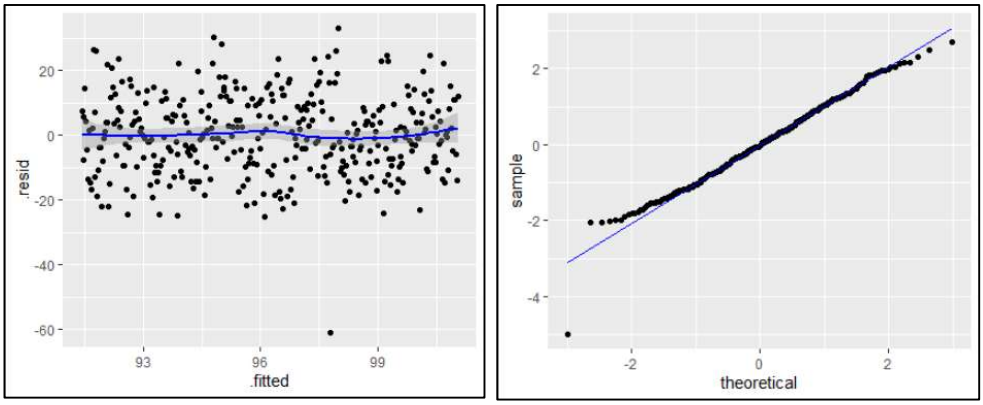
2. *Fit a linear model for* Y *using date as the predictor variable. Plot the fitted values and the residuals. Assess the model fit. Is a linear function sufficient for modelling the trend of* Y*? Support your conclusion with plots.*

   The outcomes of the model fit:

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>
0.04980653	0.04718892	12.21531	19.02746	1.681194e-05

estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
-329.10475393	97.511427819	-3.375038	8.177914e-04
0.02646962	0.006068164	4.362048	1.681194e-05

Residual plottings:



Depending on the outcomes and plots of the model fit, this linear model can be said to have sufficient evidence that can predict the trend of Y variable. Specific p-value less than 0.05 makes Date variable an important factor that contributes to the trend of ED demands. It can be further seen by the regularity of the residual plot against fitted values, although there are some outliers. Theoretical plot adds more evidence that the linear model hits the majority of the standard linear line.

But there is a concern regarding the low value of R-Squared at only about 0.05. It interprets the large gaps between the distance of errors of the observations. Therefore, we can conclude that the model is significant for inpredicting the trend of ED demands but has limitation due to the large error variation.

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?*

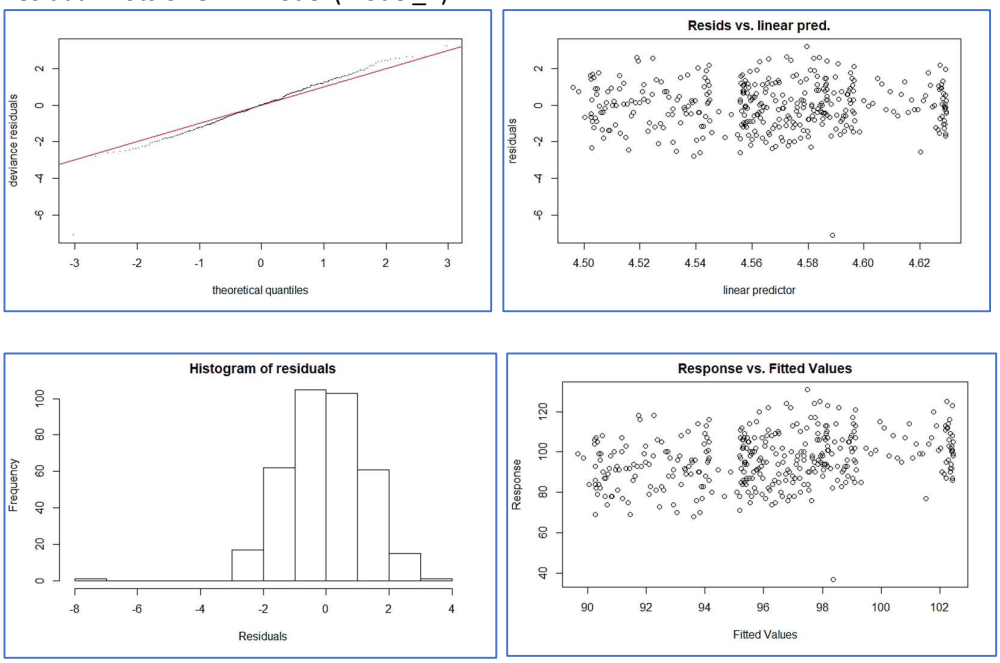
Model fitness (model_1):

```
Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.566192   0.005339   855.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
               edf Ref.df Chi.sq p-value
s(date_index) 8.134  8.798  43.53   1e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0605   Deviance explained = 7.96%
UBRE = 0.57408  Scale est. = 1         n = 365
```

Residual Plots of GAM model (model_1):



This fitted model uses generalize adaptive model. Due to the fact that our response variable is numeric and represents the count amount of admission at a 24-hours basis, Poisson function family is the best candidate together with its link function "log". As the results, we can see the significant evidence of p-value (almost zero) of our predictor variable which explains its potentiality in the model.

All the plotting above shows a good level of regularity of residuals, except some outliers. For instance; Theoretical Quantiles plot, the line of residuals follows quite closely to the standard line. However, the increase of variation at each end can be noticed. Residual vs. Fitted plot projects no identifiable pattern, except the a few outliers. It gives the idea that the residuals are normally distributed. Response vs. Fitted Values plot indicates some level of linear relationship depicted by the model.

Beyond the regularity of the model, the R-squared and Deviance explained are noticeably low, which tells us that there is much more space of improvement.

4. *Augment the model to incorporate the weekly seasonality. Compare the models using the Akaike information criterion (AIC). Report the best-fitted model through coefficient estimates and/or plots.*

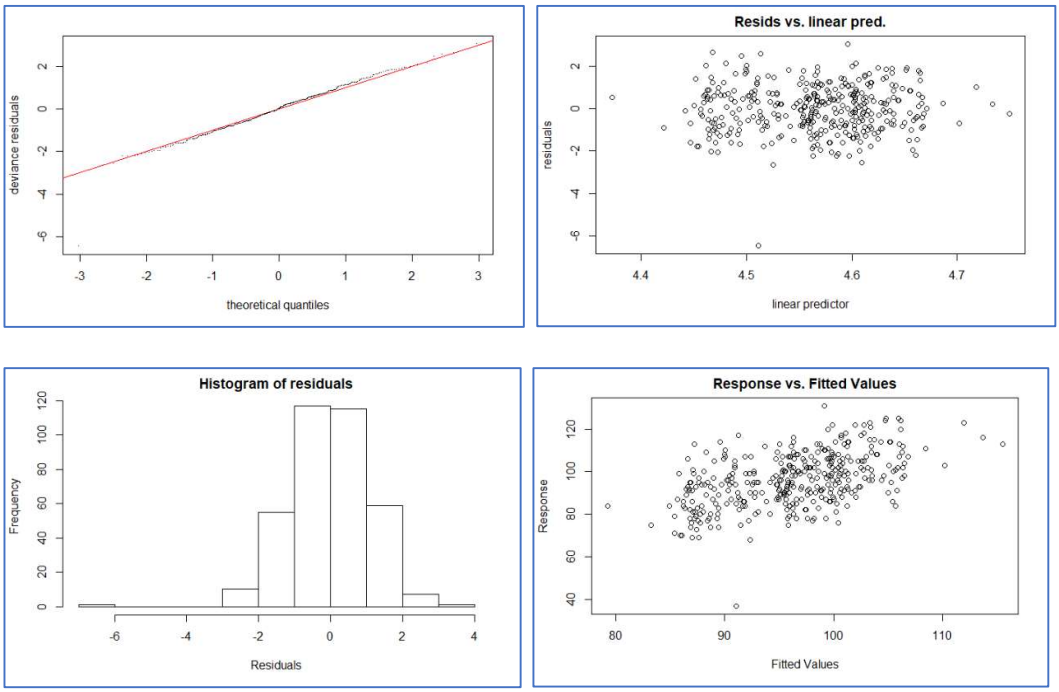
Model fitness (model_2 with weekly seasonal effect):

```
R-sq.(adj) =  0.197   Deviance explained = 25.7%
UBRE = 0.39717  Scale est. = 1         n = 365
```

AIC comparison between model_1 and model_2:

	df <dbl>	AIC <dbl>
gam_model_1	9.133683	2909.633
gam_model_2	30.408101	2845.061

Residual Plots of GAM model (model_2):



As seen in the fitness results, model_2 incorporates the effect of seasonality which helps explain better and increase the prediction reliability. R_Squared and Deviance explained have improved and at the same time, we can see the drop of UBRE value. More evidentially, AIC comparison indicates the improved performance of model_2 over the previous model_1.

In addition, the Theoretical Quantiles plot reveals a closer residuals distribution to the standard line. Residuals are in a more condensed and randomized state. We can also notice that in the Histogram plot, more outliers have been relocated to the center (zero) compared to the previous plot of model_1.

5. *Analyse the residuals. Do you see any remaining correlation patterns among the residuals?*

There is no recognizable pattern of residuals. They are completely random. So, we can say there is no remaining correlation among them.

6. *Is your day-of-the-week variable numeric, ordinal, or categorical? Does the decision affect the model fit?*

AIC results of different values of wday(day-of-the-week):

	df <dbl>	AIC <dbl>
gam_model_2	30.40810	2845.061
gam_model_3	30.40810	2845.061
gam_model_4	28.44063	2847.287

To answer the question, three model have been tested. The model_2, as we used it in our previous section, uses [wday] as categorical data type. The model_3 uses [wday] as numerical data type. The last model_3 uses [wday] as ordinal data type. Based on the tested results, all kind of data type will not disrupt the computing capability of the model. However, it can be seen from the benchmarking result,

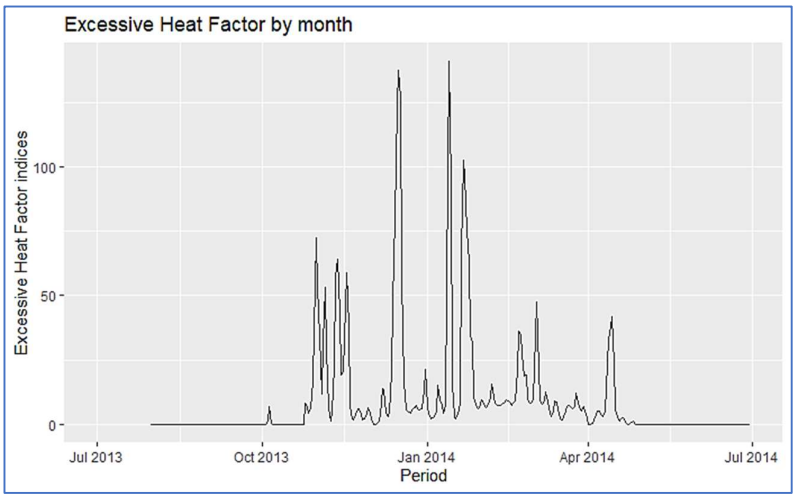
the ordinal type of data will cause the fitness of the model to go down at some level while the other two data type demonstrate the same performance.

Task 4 Heatwaves and ED demands

Task 4.1: Measuring heatwave:

1. *John Nairn and Robert Fawcett from the Australian Bureau of Meteorology have proposed a measure for the heatwave, called the excess heat factor (EHF). Read the following article to understand the definition of the EHF.*
2. *Use the NOAA data to calculate the daily EHF values for the Perth area during the relevant time period. Plot the daily EHF values.*

EHF indices vs. Time period plot using TMAX variable of NOAA data points:



This figure explains the existence of heatwaves occurring during an estimated of 6 months from November 2013 to May 2014. More remarkably, the result shows high penetration of heatwaves during summer (December-February). The peak index was recorded at 140 in 2014 January which is equivalent to 40 degree Celsius.

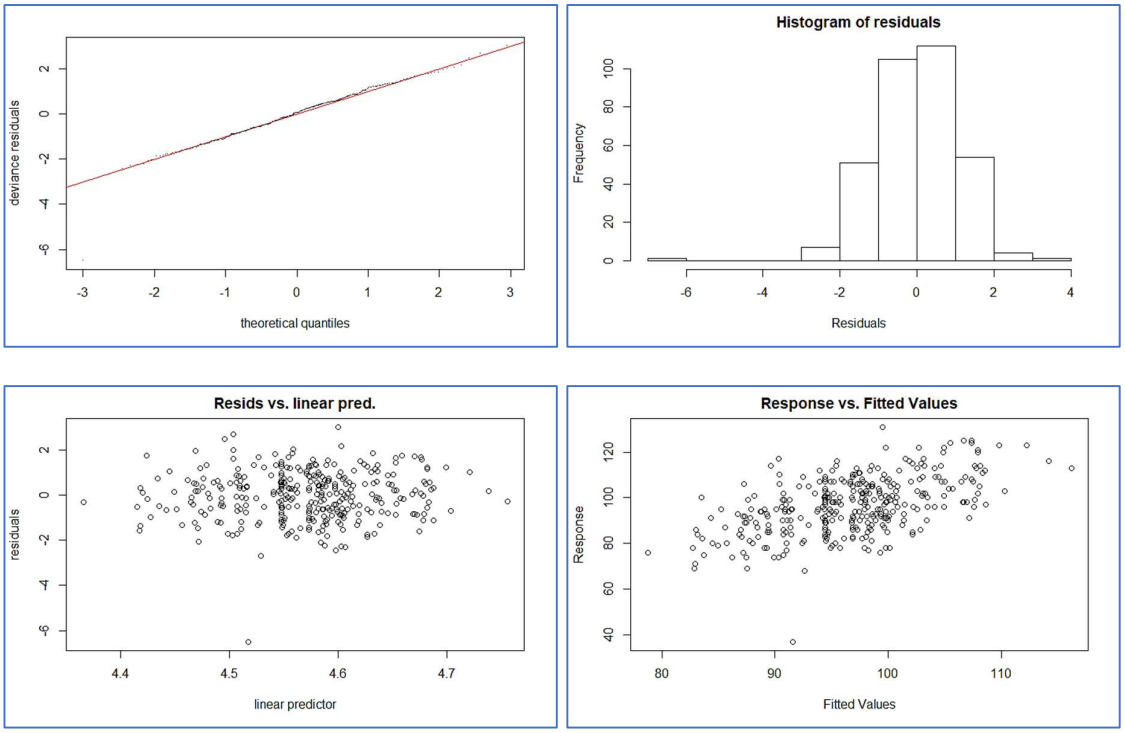
Task 4.2: Models with EHF:

Use the EHF as an additional predictor to augment the model(s) that you fitted before. Report the estimated effect of the EHF on the ED demand. Does the extra predictor improve the model fit? What conclusions can you draw?

AIC comparison between existing model (model_2) and the new model (model_5) with addition variable of EHF:

	df <dbl>	AIC <dbl>
gam_model_2	30.40810	2845.061
gam_model_5	33.53655	2593.366

Residual Plots of the new model (model_5) with additional variable of EHF:



The additional feature of EHF dramatically brings new level of improvement for the new model_5. As seen, the residuals stick very closely to the standard line in Theoretical Quantiles plot which shows a well fitness of the model and we can also see the obvious reduction in residual variation in Histogram plot. It does interpret the bonding relationship between Admissions and EHF variables. In other words, Heatwaves is likely to be the responsible factor which drives the proportion of hospital visits during the Summer time.

Task 4.3: Extra weather features

Can you think of extra weather features that may be more predictive of ED demands? Try incorporating your feature into the model and see if it improves the model fit.

AIC comparison between existing model (model_5) and the new model (model_6) with addition variable of precipitation (PRCP):

	df <dbl>	AIC <dbl>
gam_model_5	33.53655	2593.366
gam_model_6	34.86700	2593.263

I am wondering whether the precipitation play a role in the contribution the ED demands. So, I added the PRCP into the model to test its effect on ED demands. After updating the model with additional PRCP variable (model_6), we find out not much improvement in AIC metrical comparison. Therefore, we can imply that PRCP does not involve with the cause of patient visits at the medical facility and should be removed from the model. The subtle improvement of AIC of the model_6 is not worth the additional complexity of the mode by incorporating PRCP variable.

Task 5: Reflection

1. *We used some historical data to fit regression models. What are the limitations of such data, if any?*

   Historical data may or may not be able to represent the real scenario of the current problem. For instance, people's health issues may be caused by nutrition deficiency in the past, but this factor is no longer the main issue for today society. In the same way, the model that built using historical data might end up with no use at all and required a complete reconstruction. Therefore, in the absence of the current data, model should be built based on the latest historical data possible to minimize the variations that the future brings.

2. *Regression models can be used for 1) understanding a process, or 2) making predictions. In this assignment, do we have reasons to choose one objective over the other? How would the decision affect our models?*

   In this assignment, we really need to choose a clear object over another. If the model is meant for exploring the insight of a process, it would just involve in finding the right variables/factors which contribute to a certain event. However, our model here is built to predict the ED demands and it relates strongly the trend over a period of time. Therefore, in addition to finding the right variables, we need also to explore the seasonal and/or cyclical effects which distort the true nature of the trend.

3. *Overall, have your analyses answered the questions that you set out to answer?*

   At this point, the built model has included some major factors which are responsible for the increase in ED demands. I believe that there are unaccounted factors, besides the weather and weekday effects, which contributes to a better prediction of ED demands. Due to the limitation of the data features, I can say that the analyses in this assignment have significantly produced the optimal outcomes which I meant to explore.

The end