Muhammad Ahsan Asif

218606833

SIT 741 Assignment

**Task 1**

Task 1.2

* There are 366 rows in the dataset and 64 columns in the dataset.
* There are 9 Hospitals in the dataset as followed:

1. Royal.Perth.Hospital
2. Fremantle.Hospital
3. Princess.Margaret.Hospital.For.Children
4. King.Edward.Memorial.Hospital.For.Women
5. Sir.Charles.Gairdner.Hospital
6. Armadale.Kelmscott.District.Memorial.Hospital
7. Swan.District.Hospital
8. Rockingham.General.Hospital
9. Joondalup.Health.Campus

* Data Types in the Dataset:

1. Hospital Names : String, Categorical
2. Attendance/Admission/Triage : Integer , Numerical
3. Date: Date

* The dataset covers the time period starting from July 2013 to June 2014.
* The difference between Attendance and Admission is that **Attendance** is recorded whenever a patient is registered for any manner in the system. Attendance count also includes patients who are dead on arrival or who does not wait to be seen and then leave. It gives a total number of them. Whereas, **Admission** refers to all those patients who were admitted to emergency department for some sort of treatment or further check up by the doctor.
* **Tri\_1 :** The condition is definitely life threatening and requires immediate medical care and should be seen in less than or equal to 2 mins.

**Tri\_2 :** It is a probable life or limb threat and should be seen in less than or equal to 10 mins.

**Tri\_3:** It is a possible life or limb threat and should be seen in less than or equal to 30 mins.

**Tri\_4:** There is no threat to life or limb but some incapacity or injury and should be seen in less than or equal to one hour.

**Tri\_5:** There is no threat to life or limb and should be seen in less than or equal to two hours.

**Task 2:**

Task 2.2

1. No, each variable does not have its own column.

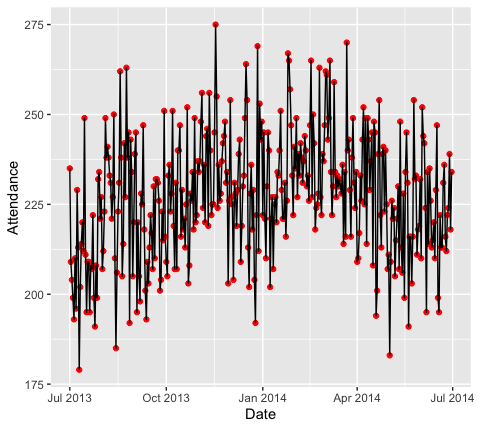
Yes, each observation does have its own row.

Yes, each value does have its own row.

1. We need to do one spread function and one gather function. I started out by gathering all the columns in one column except Date and then I used the separate function to separate the variable names from their hospital numbers that is \_\_0,\_\_1 etc. After separating I used spread to spread out the rest of the columns that were Admissions, Attendance, Tri\_1 etc
2. Variables were not clean. For instance, all of the variables had the datatype **chr**. So, I changed the date variable to **Date** and the rest of them to **numeric**.
3. Yes, they were missing values. I dealt them by replacing the missing values by 0, since I didn’t want to remove the entire rows and exclude them from the dataset so I thought replacing them will a null value would be better.

**Task 3**

Task 3.2



By using the variables of Date and Attendance since in Attendance the total number of patients were present. I used a line chart with red points as indicators of start and end. This plot gave a clear visual of the ED demands across the year. It showed that in the months of October, November, February march and April there was a quite high number of patient attendance in the emergency departments. In the summer that are June and July the number of attendances decreased as compared to the rest of the year.

Task 3.4

After plotting different variables of the Royal Perth hospital, in the histogram I could see a bell shaped curve in the graph which made me to conclude that in order to model the ED demand Normal distribution would be appropriate. The variables of Attendance and Admissions meet the assumptions of poisson distribution.

**Task 4**

Based upon the values in summary and the plot and hist function in the summary I would conclude that negative binomial distribution model fitted well into the data.

**Task 5:**

Fitting distribution are usually used through mathematical functions to find that statistical analysis of the models. There are various distribution methods when it comes to the analysis of data using R. I would like to talk about three of them which are:

* Normal Distribution
* Poisson Distribution
* Binomial Distribution

Talking about normal distribution it is a continuous probability distribution, and the pros would include the theorem of central limit theorem which applies on it that is after a certain level of observations in the sample, the sample is known to be random. The cons of this distribution would include that there is always there is a probability of a negative value. Poisson Distribution deals basically with rates whereas, the binomial Distribution deals with the sum of events with the respect of occurring or not. Pros of poisson distribution would include we can take the limits of the events going to infinity and when you adjust the events the interpretation becomes clearer and efficient to explain. In binomial distribution you can have a look at a number of events that are not relative to time. Cons could include Both of these distributions requires their events to be independent otherwise these models don’t work.

**Task 6:**

During my work with the assignment I noticed that the dataset could be used in accordance with other dataset and could be threat to the security of hospitals whose data sets we have been using. These Admission and Attendances could be linked with other delicate information in order to make sense and take meaningful data out of it. In order to mitigate the risk, the emergency department data could be scaled down or be limited to people who have access to it.

**Task 7:**

I received the help of how to plot the data into the line charts, histograms etc from other students. I did know how to plot these functions but to modify them according to my own need was taught by some of the other students. I learned on how to plot different types charts,plots,diagram and then to add my own aesthetics into it.

I estimate that I should be getting 20/25 and interval estimate of between 80% to 90% based upon the work im submitting today. I reached to this estimate based upon the theoretical answers I gave which could have been made better and I was unable to explain the log likelihood through the code which then would have been easier to explain the log likelihood of the two models, otherwise except that I am happy with my work today.