# STATISTICAL DATA ANALYSIS

MOHIT RANGHOLIYA
mranghol@deakin.edu.au

# Task-1

## Task-1.1

Scanning through the http://archive.ics.uci.edu/ml/datasets.html datasets , it can be observed that there are total **362 datasets** and each dataset is described using **7 attributes.** These attributes are as follows :

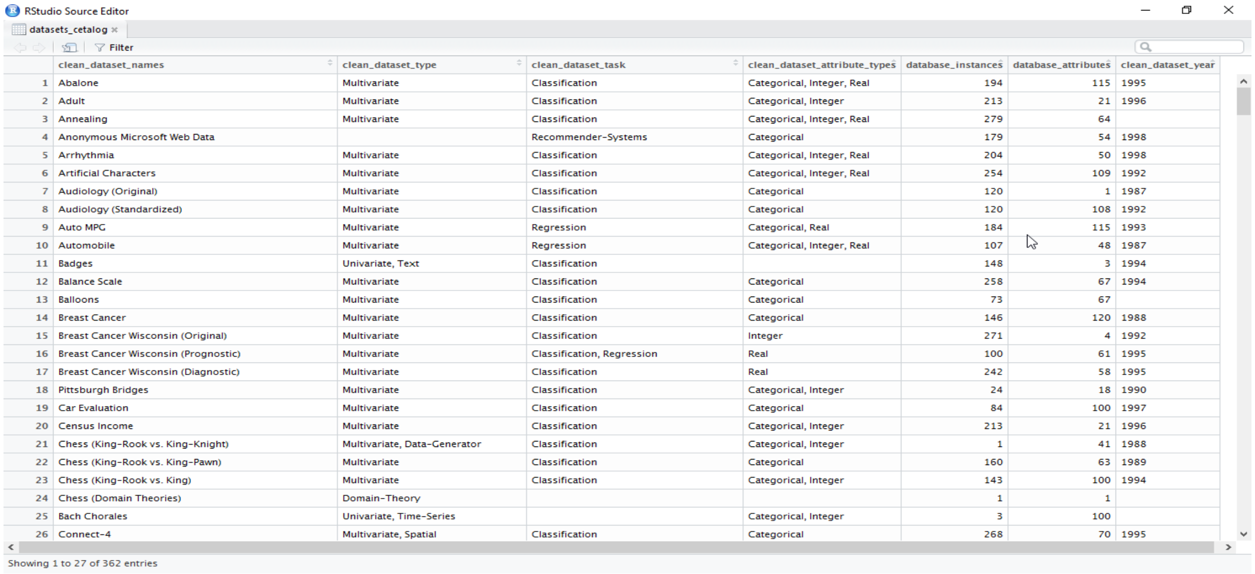
1. Name – Dataset name
2. Data Types – Type of the data
3. Default Task – Task performed on the dataset
4. Attribute Types – Datatype Type of attributes
5. #Instances – Number of Instances in the dataset
6. #Attributes – Number of Attributes in the dataset
7. Year – Year in which it is added in the catalog

## Task-1.2

Scraping the table from http://archive.ics.uci.edu/ml/datasets.html for our analysis in R. Here the following code snippet is used to scrap each attributes and store into the separate dataframe.

```r
library(rvest)
uci_html <- read_html("http://archive.ics.uci.edu/ml/datasets.html")
#Database Names
clean_dataset_names <- uci_html %>%
  xml_find_all("/html/body/table[2]/tr/td[2]/table[2]/tr/td/table/tr/td[2]/p/b/a") %>%
  html_text()
#clean_dataset_names
database_names <- data.frame(clean_dataset_names)
```

Below is the snapshot of whole table after merging them together.



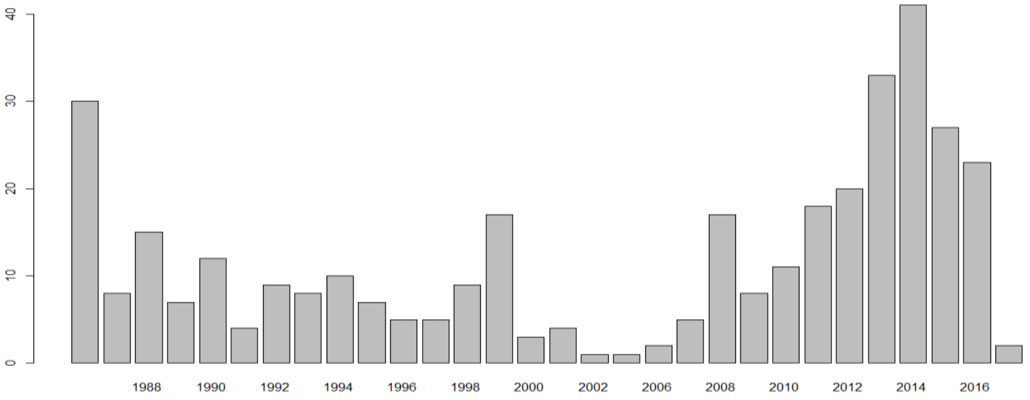
| | clean_dataset_names | clean_dataset_type | clean_dataset_task | clean_dataset_attribute_types | database_instances | database_attributes | clean_dataset_year |
|---|---|---|---|---|---|---|---|
| 1 | Abalone | Multivariate | Classification | Categorical, Integer, Real | 194 | 115 | 1995 |
| 2 | Adult | Multivariate | Classification | Categorical, Integer | 213 | 21 | 1996 |
| 3 | Annealing | Multivariate | Classification | Categorical, Integer, Real | 279 | 64 | |
| 4 | Anonymous Microsoft Web Data | | Recommender-Systems | Categorical | 179 | 54 | 1998 |
| 5 | Arrhythmia | Multivariate | Classification | Categorical, Integer, Real | 204 | 50 | 1998 |
| 6 | Artificial Characters | Multivariate | Classification | Categorical, Integer, Real | 254 | 109 | 1992 |
| 7 | Audiology (Original) | Multivariate | Classification | Categorical | 120 | 1 | 1987 |
| 8 | Audiology (Standardized) | Multivariate | Classification | Categorical | 120 | 108 | 1992 |
| 9 | Auto MPG | Multivariate | Regression | Categorical, Real | 184 | 115 | 1993 |
| 10 | Automobile | Multivariate | Regression | Categorical, Integer, Real | 107 | 48 | 1987 |
| 11 | Badges | Univariate, Text | Classification | | 148 | 3 | 1994 |
| 12 | Balance Scale | Multivariate | Classification | Categorical | 258 | 67 | 1994 |
| 13 | Balloons | Multivariate | Classification | Categorical | 73 | 67 | |
| 14 | Breast Cancer | Multivariate | Classification | Categorical | 146 | 120 | 1988 |
| 15 | Breast Cancer Wisconsin (Original) | Multivariate | Classification | Integer | 271 | 4 | 1992 |
| 16 | Breast Cancer Wisconsin (Prognostic) | Multivariate | Classification, Regression | Real | 100 | 61 | 1995 |
| 17 | Breast Cancer Wisconsin (Diagnostic) | Multivariate | Classification | Real | 242 | 58 | 1995 |
| 18 | Pittsburgh Bridges | Multivariate | Classification | Categorical, Integer | 24 | 18 | 1990 |
| 19 | Car Evaluation | Multivariate | Classification | Categorical | 84 | 100 | 1997 |
| 20 | Census Income | Multivariate | Classification | Categorical, Integer | 213 | 21 | 1996 |
| 21 | Chess (King-Rook vs. King-Knight) | Multivariate, Data-Generator | Classification | Categorical, Integer | 1 | 41 | 1988 |
| 22 | Chess (King-Rook vs. King-Pawn) | Multivariate | Classification | Categorical | 160 | 63 | 1989 |
| 23 | Chess (King-Rook vs. King) | Multivariate | Classification | Categorical, Integer | 143 | 100 | 1994 |
| 24 | Chess (Domain Theories) | Domain-Theory | | | 1 | 1 | |
| 25 | Bach Chorales | Univariate, Time-Series | | Categorical, Integer | 3 | 100 | |
| 26 | Connect-4 | Multivariate, Spatial | Classification | Categorical | 268 | 70 | 1995 |

Showing 1 to 27 of 362 entries

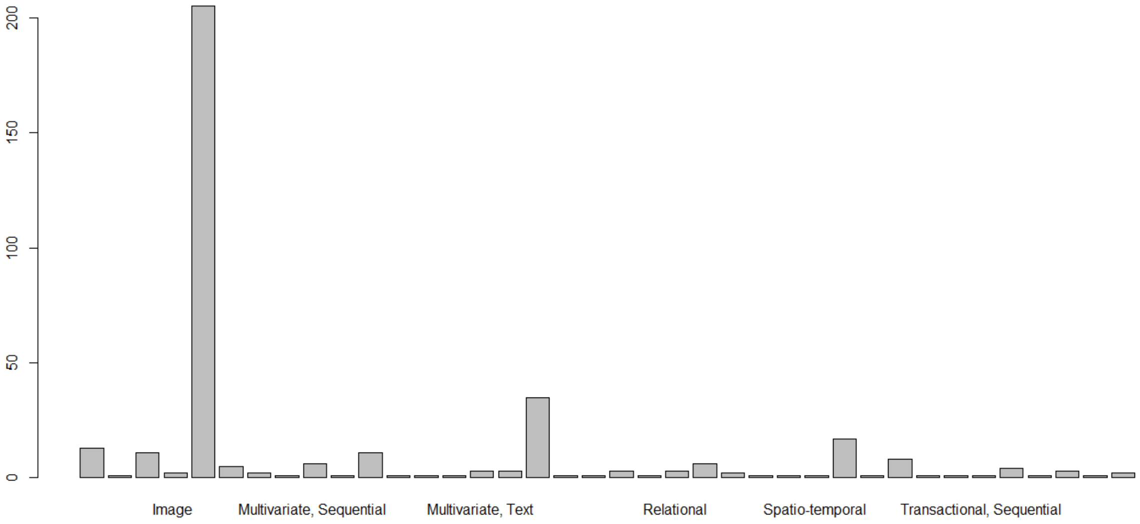
## Task-1.3

Exploratory data analysis on scrapped table has given some more insights of the given datasets. Following are some the useful results.

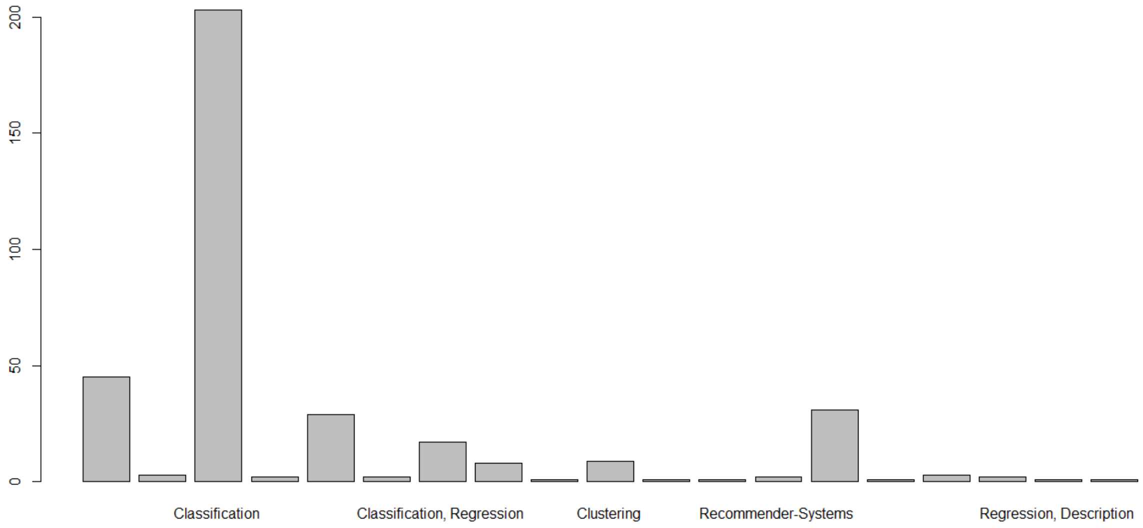
- It is found that , in the year of 2014 , highest number of datasets were added in to this catalog.



- In the given catalog of datasets it seems that dataset datatype is mostly Multivariate. Out of 362 datasets ,282 are Multivariate.



- It is also seen that most of the datasets are suitable for Classification task. From the graph below, total 263 datasets are suitable for Classification.



## Task-1.4(a) Dataset having default task as Regression

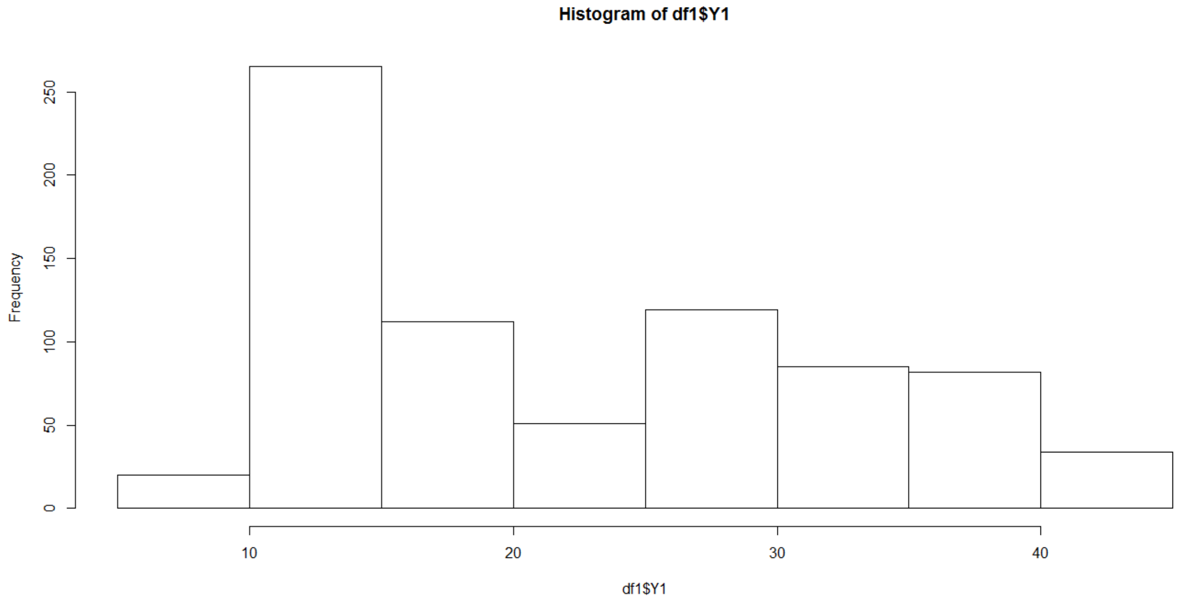
Selected Dataset : **Energy Efficiency**

The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses.

X1 Relative Compactness
X2 Surface Area
X3 Wall Area
X4 Roof Area
X5 Overall Height
X6 Orientation
X7 Glazing Area
X8 Glazing Area Distribution
**y1 Heating Load**
**y2 Cooling Load**

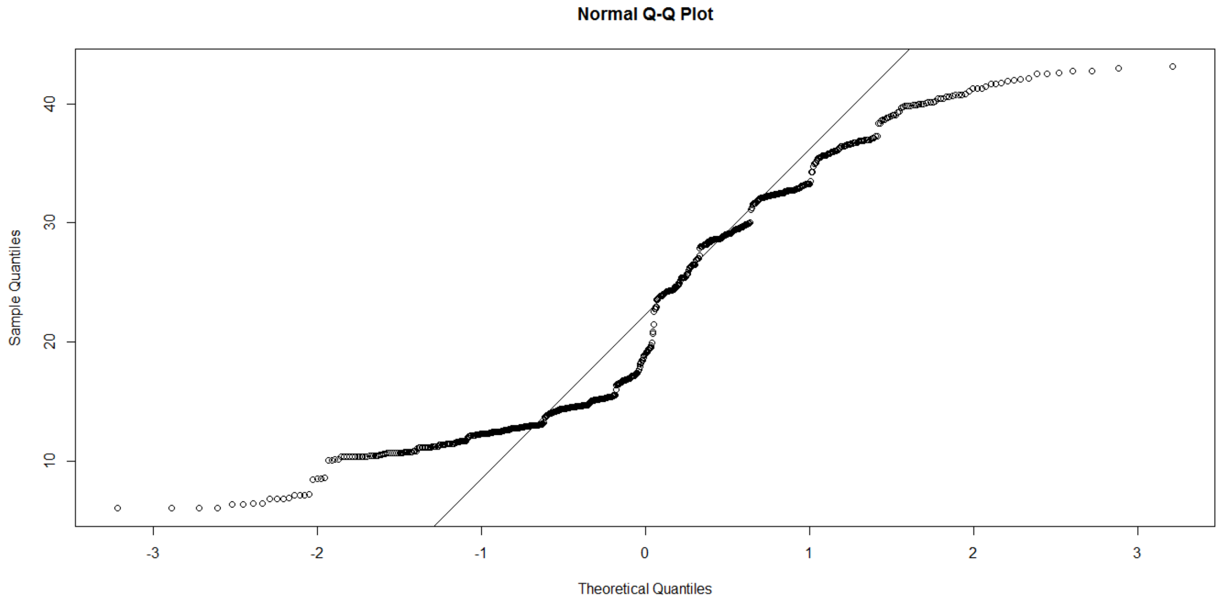
## Task-1.4(b) Visualization of response variables

Histogram of Heating Load



Q-Q Normal with Q-Q Line for Heating Load

Visualization shows that Heating load is not normally distributed.

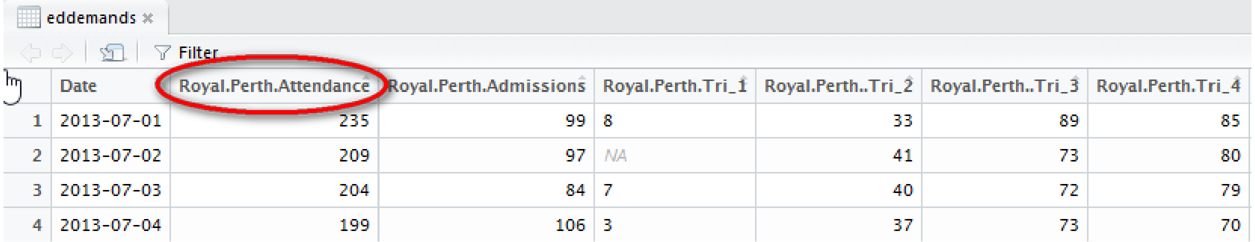


# Task-2

## Task-2.1

### Task-2.1.1

Emergency Department (ED) Demand Data has been downloaded from the provided source. The data needed to be transformed. There is subtitle and title which has been transformed for one header only by combining the name of Hospital and attribute name of that Hospital. So after this this data can be loaded into one dataframe in R for analysis. The snapshot below shows this transformation.



| | Date | Royal.Perth.Attendance | Royal.Perth.Admissions | Royal.Perth.Tri_1 | Royal.Perth..Tri_2 | Royal.Perth..Tri_3 | Royal.Perth.Tri_4 |
|---|---|---|---|---|---|---|---|
| 1 | 2013-07-01 | 235 | 99 | 8 | 33 | 89 | 85 |
| 2 | 2013-07-02 | 209 | 97 | NA | 41 | 73 | 80 |
| 3 | 2013-07-03 | 204 | 84 | 7 | 40 | 72 | 79 |
| 4 | 2013-07-04 | 199 | 106 | 3 | 37 | 73 | 70 |

### Task-2.1.2

Time period covered by the dataset : **2013-07-01 to 2014-06-30 (1 Year)**

Hospitals Involved :

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

**Attendances** - the number of patients recorded as arriving at a public emergency department.

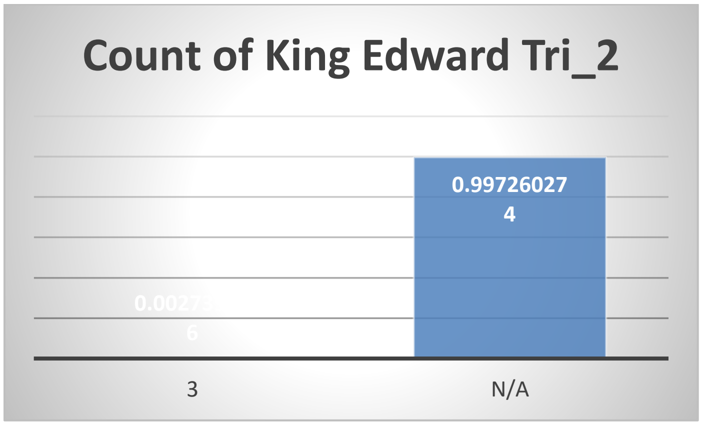
Attendance = Triage1 + Triage2 + Triage3 + Triage4 + Triage5

**Admissions** - the number of patients who are subsequently admitted to the hospital for care and/or treatment.

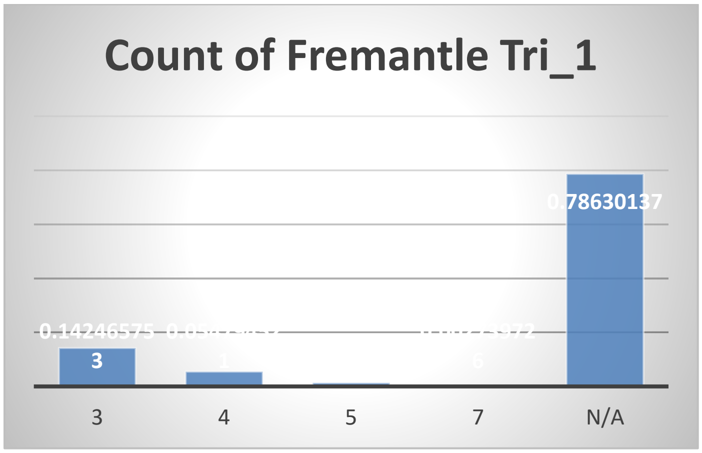
Triage categories are allocated to each patient based on an assessment of their presenting conditions, generally by the triage nurse, with triage 1 being the most urgent and triage 5 being the least urgent.
**Triage 1:** Resuscitation- immediate, within seconds;
**Triage 2:** Emergency- within 10 minutes;
**Triage 3:** Urgent- within 30 minutes;
**Triage 4:** Semi-urgent- within 60 minutes;
**Triage 5:** Non-urgent - within 120 minutes

### Task-2.1.3 Data Cleaning on ED Demands data

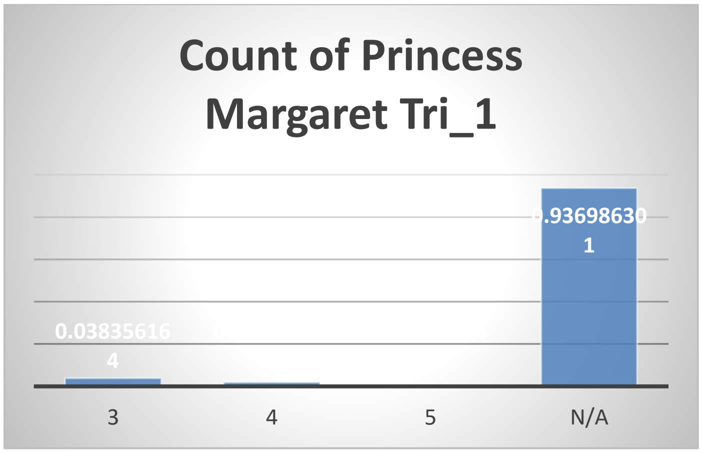
- King Edward Tri_1 has 100% N/A values **-> Remove King Edward Tri_1**
- King Edward Tri_2 has 99.3% N/A values **-> Remove King Edward Tri_2**



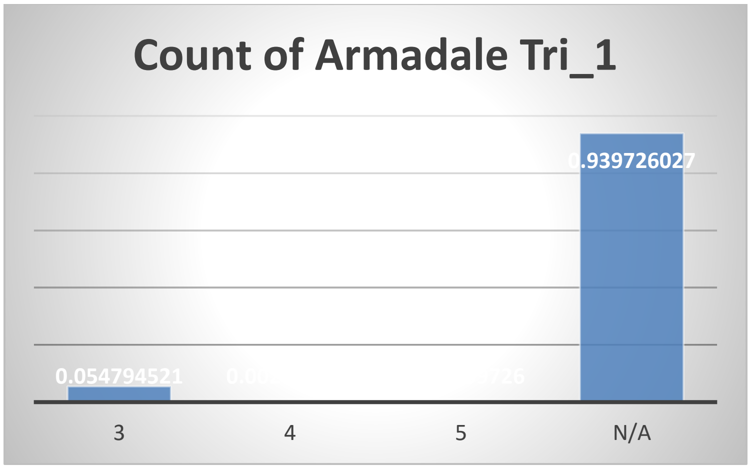
- 78.63% records have N/A values for Fremantle Hospital Triage-1 -> **Remove Fremantle Tri_1**



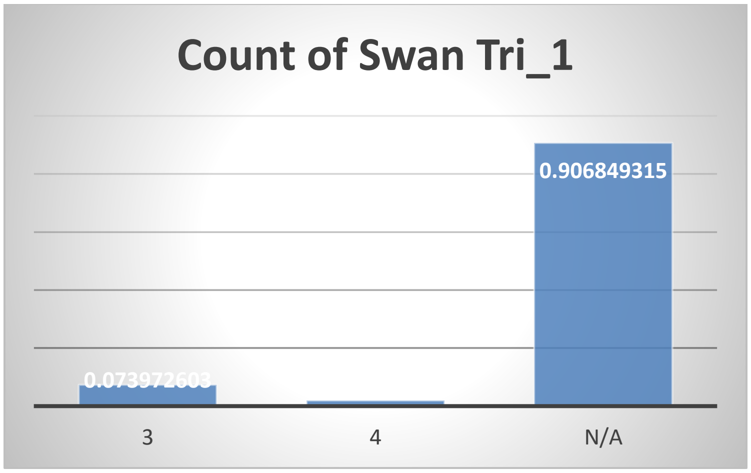
- 78.63% records have N/A values for Princess Margaret Hospital Triage-1 -> **Remove Princess Margaret Tri_1**



- Armadale Tri_1 has 93.97% N/A values **-> Remove Armadale Tri_1**



- Swan Hospital Triage-1 has 90.68% N/A values -> **Remove Swan Tri_1**

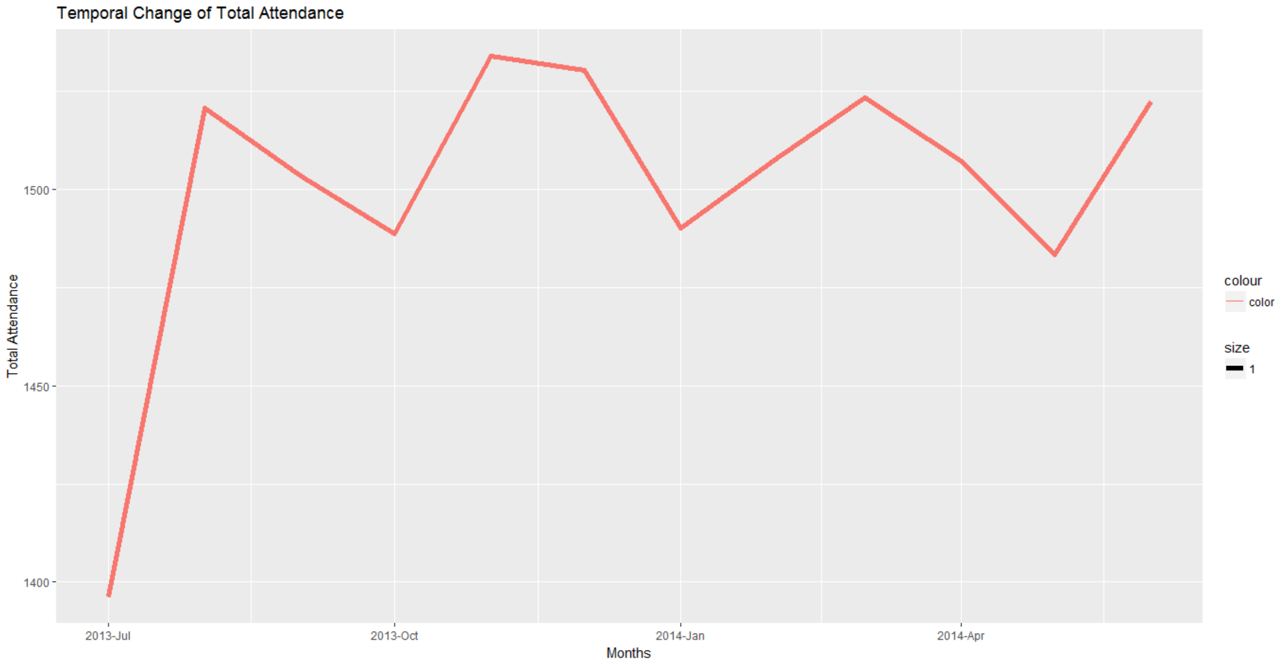


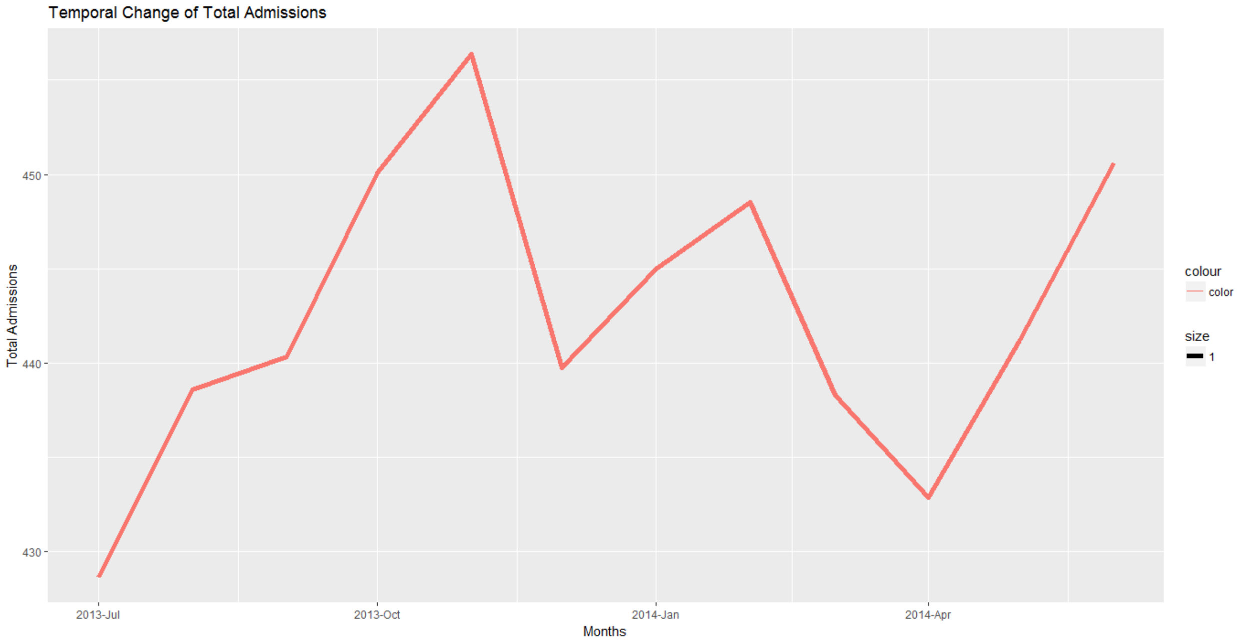
## Task-2.2 Exploratory Data Analysis

### Task-2.2.1

**Target Variable** :

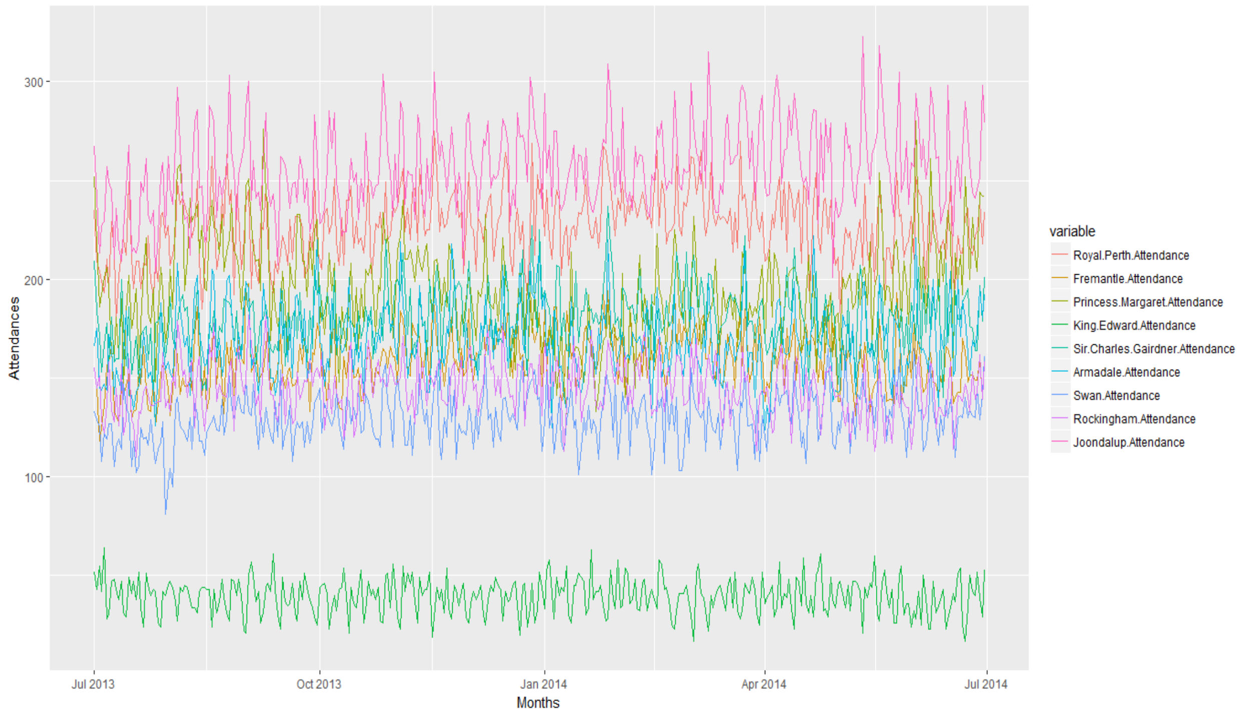
1. **Attendance** of patients arriving at Emergency Departments.
2. **Admissions** - the number of patients who are subsequently admitted to the hospital for care and/or treatment





## Task-2.2.2

Comparison of the volume of ED demands of different hospitals at different times.



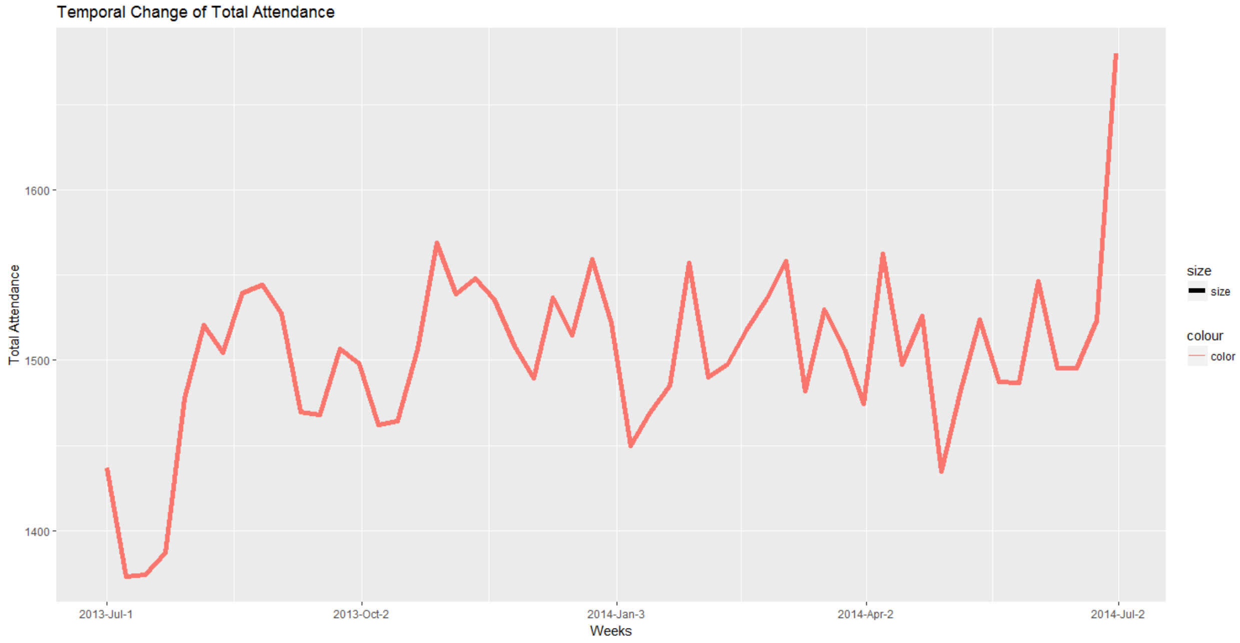
The above line plot represents month wise attendance of particular hospitals with different color.

The useful insight is Royal Perth hospital had the highest attendance throughout the year and particularly it was highest in the month of May,2014 and June,2014.

The lowest attendance is marked at King Edward Hospital throughout the year.

## Task-2.2.3

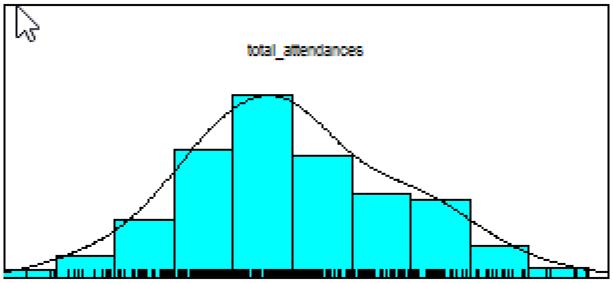
Change of ED demands during a week

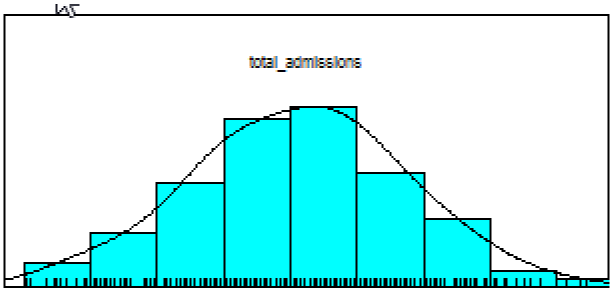


The above graph represents the useful insight that the second week of July,2014 has the highest total attendance of all hospitals.

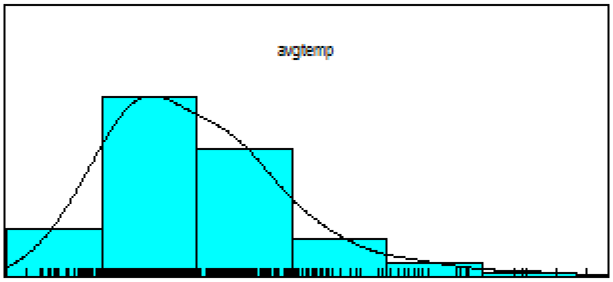
## Task-2.2.4

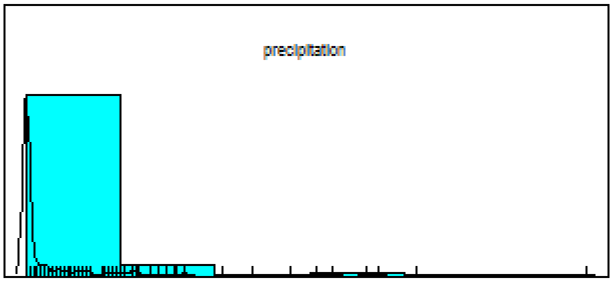
**Normal Distribution** is more suitable for attendances and admissions. Below snaps are the distributions for total attendance and admissions.





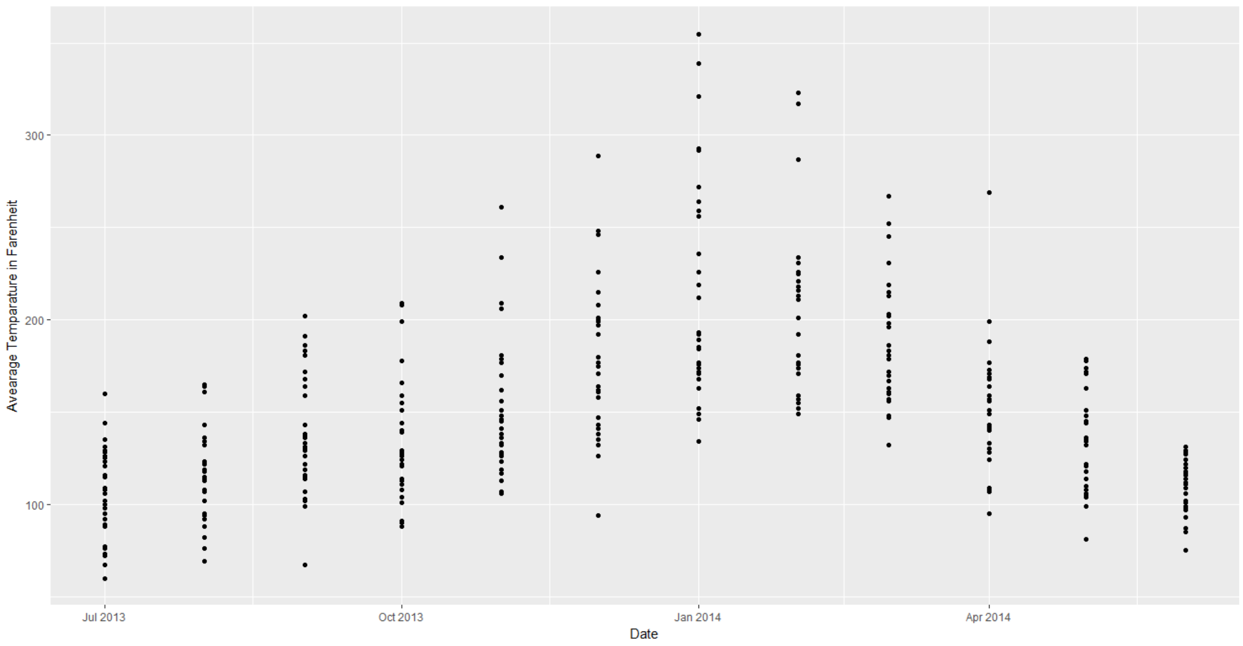
Average Temperature and Precipitation satisfies the poison distribution. Below are the distribution for them.





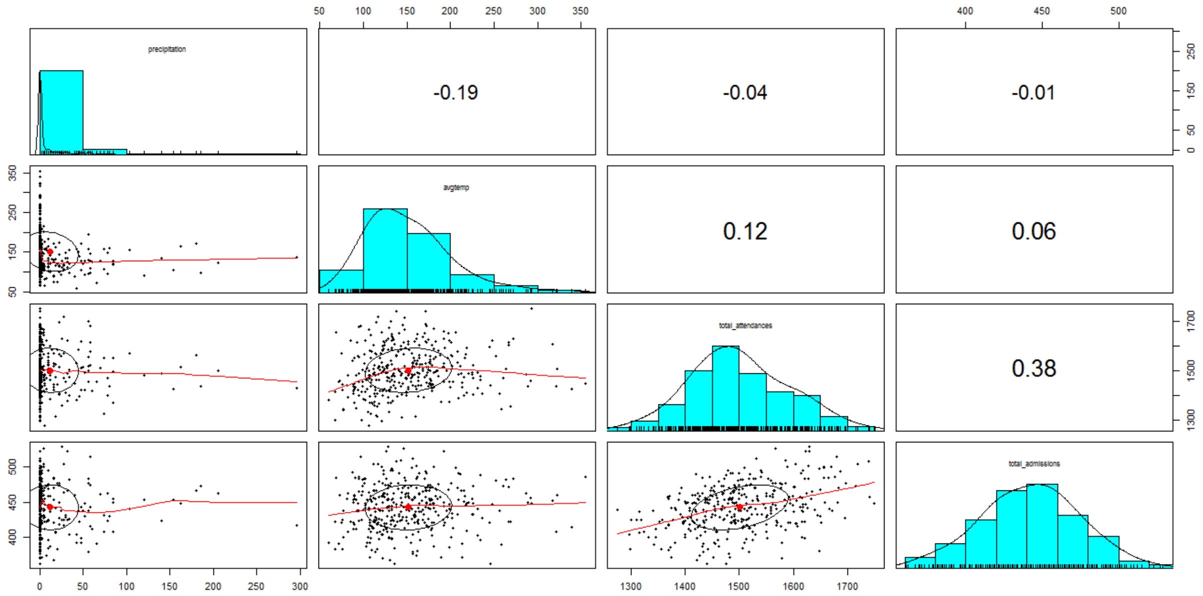
## Task-2.2.5 Visualization of change of the weather data in the year

The change of average temperature during the whole year. The useful insight got from here is in the month of Jan,2014 the average temperature was the highest. From July,2013 to Jan,2014 it increased slowly around every month an average hike of 20 Fahrenheit - 30 Fahrenheit.



## Task-2.2.6

From the given dataset , it seems that there is not much correlation between average temperature and total attendance/total admissions at ED Demands.


precipitation
avgtemp
total_attendances
total_admissions