Ashutosh R Shukla Student ID:-217669865

SIT741 Assignment 1

Task 1: Obtaining ED demand data

Task 1.1 Download the data set using the link below.

→Done in R Markdown file

Task 1.2 Answer the following questions:

- How many rows are in the data?
  →There are 365 rows in the data set.
- What data types are in the data?
  →There are all numeric and character data types in the data and the csv data is arranged as a Data frame.
- What time period does the data cover?
  →The time period this data cover is from 30-06-2014 to 01-07-2013.That is Time difference of 364 days.
- What's the difference between "Attendance" and "Admissions"?
  →the major difference between Attendances and admissions could be understood through their definitions i.e. attendance are the number of patients recorded as arriving at a public emergency department in our case Hospitals whereas Admissions are the number of patients who are admitted to the hospital for care and/or treatment subsequently.
- What do the variables `Tri_1`, `Tri_2`, … represent?
  →So, there are a series of steps that happen Upon arrival in the ED, where people undergo a brief triage, or interview aka Triage's, that helps to determine the nature and severity of their illness.
  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). N/A - Values is less than 3 and has been suppressed. (Australia, 2015)
  Individuals with serious illnesses are then seen by a physician more rapidly than those with less severe symptoms or injuries

Task 2: Tidy data (5 points)

Task 2.1 Cleaning up columns

→Done in R Markdown File.

Task 2.2 Tidying data

- Does each variable have its own column?

  →Yes, each unique variable (data from different hospitals) has its own column.

- Does each observation have its own row?

→Yes, here each observation (dates in our case) has its own row.

- Does each value have its own cell?

  →Yes, there lies each unique value corresponding to each variable and its observation(date) in a separate cell.

- How many spreading operations do you need?

  →We need (9[for no of hospitals]x5[for number of triads])45 operations in the data set.

- How many gathering operations do you need?

  → We need (9[for no of hospitals]x5[for number of triads])45 operations in the data set.

- Explain the steps.

  →So, for this we need to download the Tidyverse database: -

  # The first step involves

  install.packages("tidyverse")

  #Or Alternatively we can also install just tidyr:

  install.packages("tidyr")

  There are two fundamental verbs of data tidying:

  gather() function that takes multiple columns and gathers them into key-value pairs. It is actually responsible for making the data wider and longer.

  spread() function that takes two columns (key & value) and spreads them into multiple columns. This makes the data longer and in that terms wider.

  So, we apply these two functions on our variables Tri1_1, Tri1_2, Tri1_3, Tri1_4 and Tri1_5(for all hospitals) and generate an immersive and tidy data.

3. Are the variables having the expected variable types in R? Clean up the data types.

→No, not all the variables are having expected variable types in R as there are several variables such as Triage_1(Tri_1) and Traige_2(Tri_2) that should have "numeric" data types but due to several null/Not available(N/A) values it treats the variable type as "character".

Cleaning up done in R code.

4. Are there any missing values? Fix the missing data. Justify your actions.

→Yes, there are several values in various columns that are “N/A” that are missing or not available.

We can fix this by removing the NAs as a missing value likely represents no data in that group and make them equivalent to zero count signifying that they don’t count to any value in our data.

Task 3: Exploratory Data Analysis

Task 3.1 Select a hospital

→Done in R markdown file.

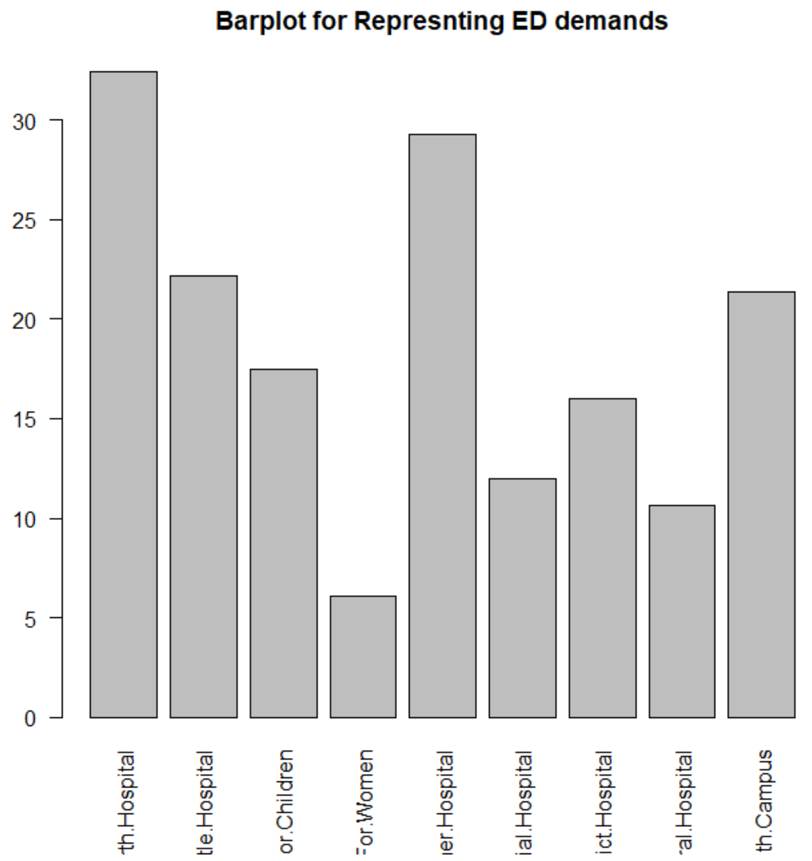
Task 3.2 For the hospital selected, if we want to compare the volume of ED demands across the year, which plot can we use? Show your plot and explain what the plot shows. (Hint: Which variables measure the ED demands?)

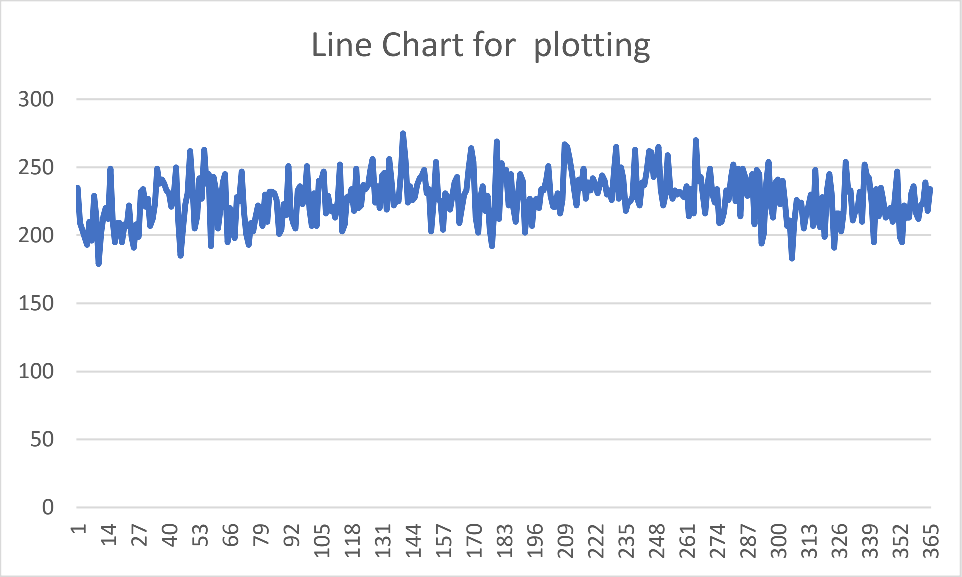
→If we want to compare the volume of ED Demands over the year we can measure it using Attendance and Admissions variable.

The plot we could use to efficiently differentiate between the ED demand over a year is a line chart as it shows day by day estimate of variable progression for a variable of a particular hospital.

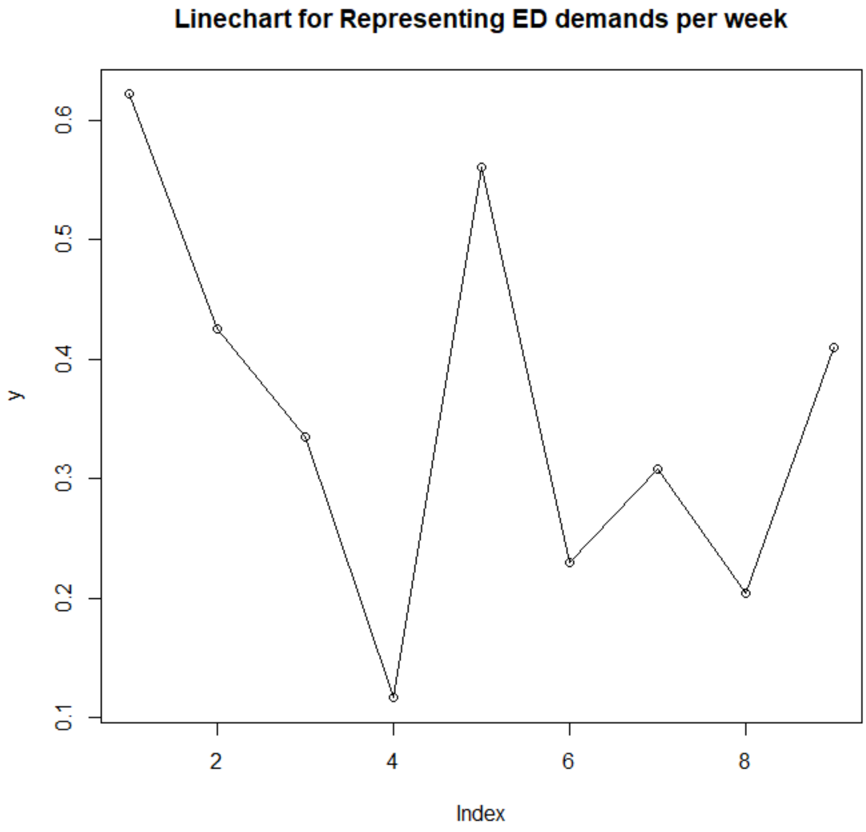
The plot we could use to efficiently differentiate between the ED demand over a year between different hospitals is a Barplot because as from the diagram we could see that each different value of means of Admissions of various hospitals given can be observed and analysed distinctively and uniquely.

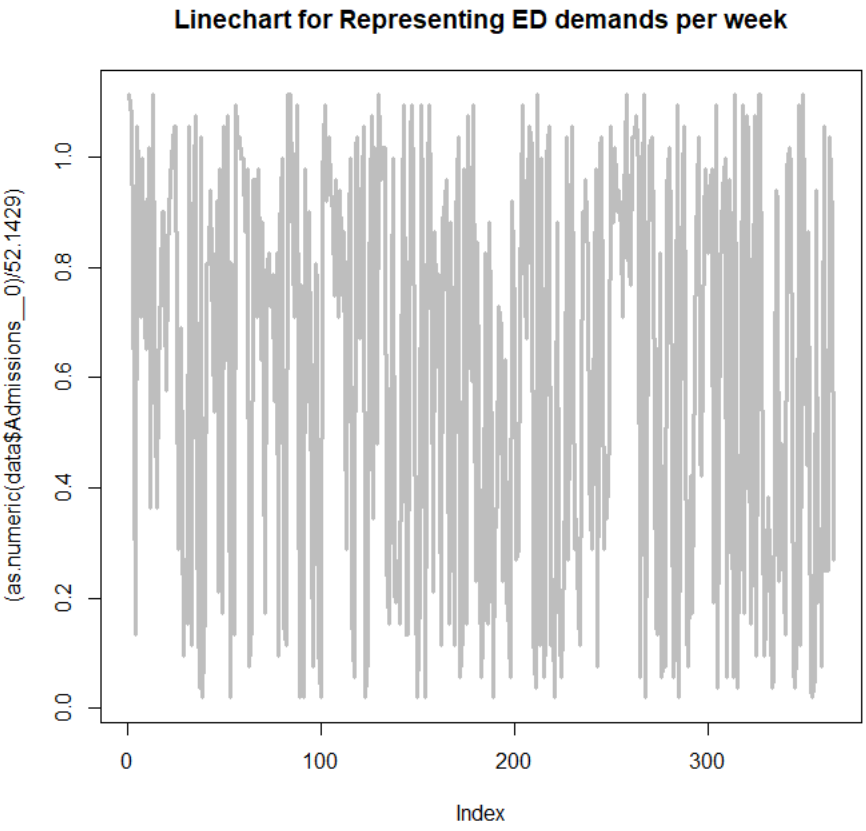
Here, we took mean for the consideration of the values for each hospital as it would be the perfect statistical measure that would give us an average number of people Admitting in the hospitals for the year; in result giving us a comparative distribution for comparing demand of ED across the year.


Barplot for Represnting ED demands
30
25
20
15
10
5
0
th.Hospital
tle.Hospital
or.Children
For.Women
er.Hospital
ial.Hospital
ict.Hospital
al.Hospital
th.Campus


Line Chart for plotting
300
250
200
150
100
50
0
1
14
27
40
53
66
79
92
105
118
131
144
157
170
183
196
209
222
235
248
261
274
287
300
313
326
339
352
365

Task 3.3 How do the ED demands change during a week? Show it visually.





Here is the change among all the hospitals observed over the ED's demand over the week as wee can see it fluctuates and goes down nearly at the 4th day(which is the lowest) and is at peak during the start of the week and after 4th day it goes on peaking high and low simultaneously.

Task 3.4 Which distributions are appropriate for modelling the ED demand? Which variables meet the assumptions for the Poisson distribution? (For simplicity, here we will make a "naive" assumption that counts on consecutive days are independent. We will relax this assumption later in the unit.)

→From the implementation On R Markdown file we can conclude that weather Poisson Distribution or Lognormal distribution is appropriate for modelling ED Demand as here, an event(Admittance in hospitals as we've taken this variable for the evaluation) can occur any number of times during a time period plus events occur independently(assumption). In other words, if an event occurs, it does not affect the probability of another event occurring in the same time period. So, satisfying all the conditions for Poisson Distribution; I believe it is appropriate for ED Demand.

The variable that meet the assumptions of Poisson Distribution are Attendance and Admissions as they are independent and ones occurrence does not affect the probability of another event.[which is not true in terms of Triads variable as if one triad(or tri_5) does show 0 value then there is no scope for further testing as patient is highly unlikely to have any value for Tri_1 as he/she's not critically emergent.

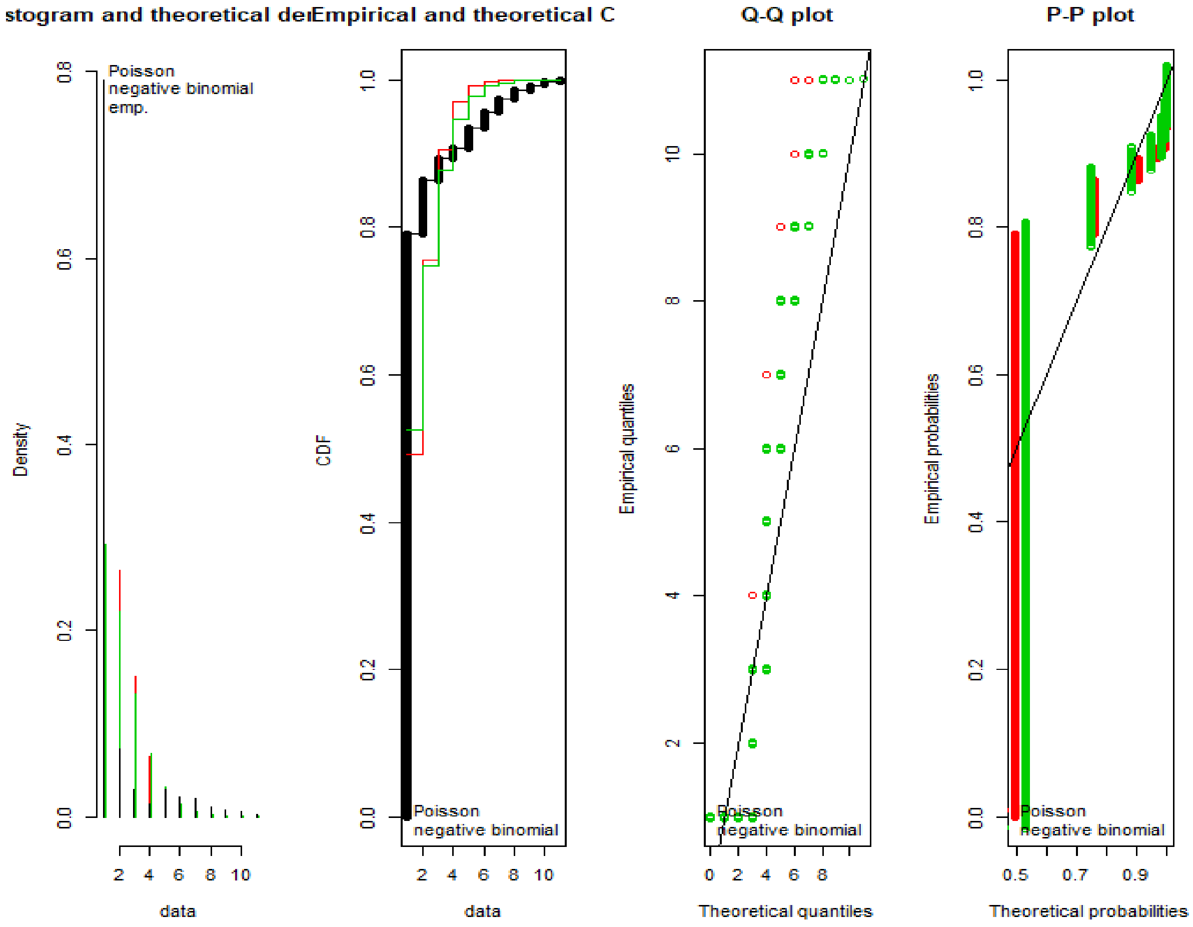
- The rate of occurrence is constant; that is, the rate does not change based on time.

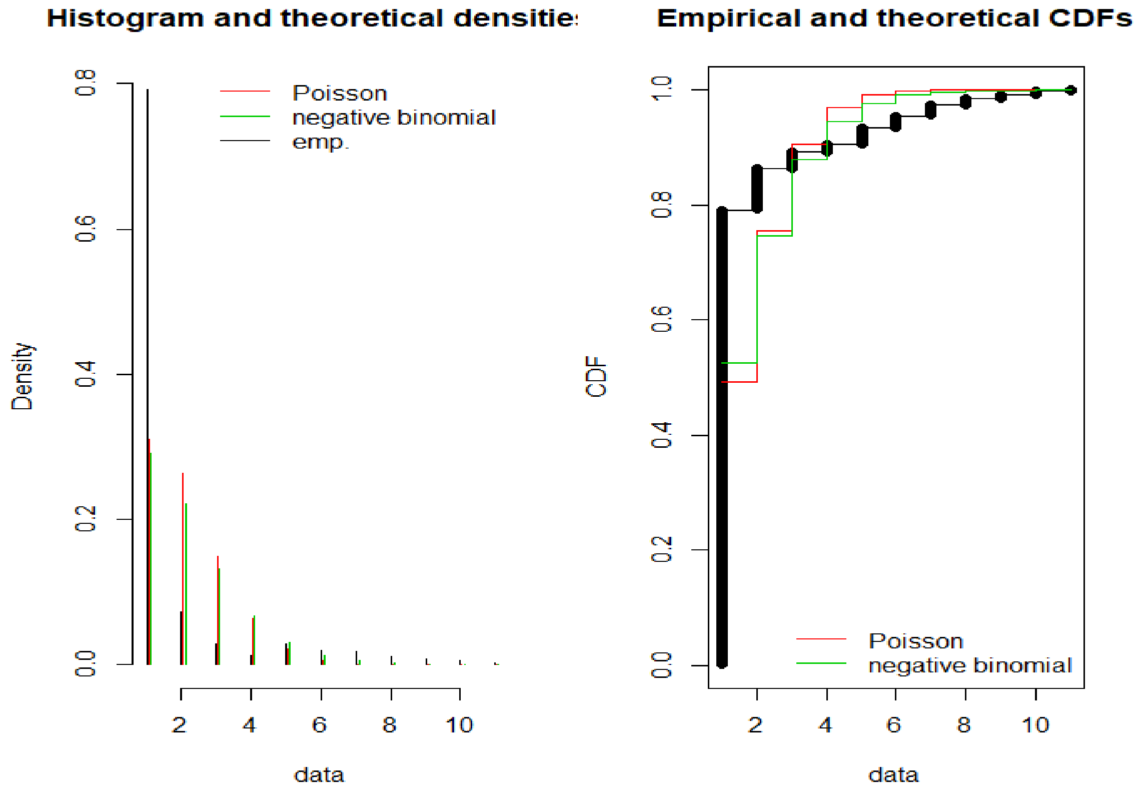
Task 4: Fitting distributions

Task 4.1: Fitting distributions

→Done in R Markdown file.

Task 4.2: Compare distributions→





As we can see Negative Binomial is a better fit for the given data as it fits passably than Poisson distribution as seen from the graph plots generated above.

Poisson does somewhat fir the data but more accurately we could see Negative Binomial dist. Fits it more closely(all the green points in the data graph).

Task 5: Research question

The important distributions are:-

- The Bernoulli distribution, which takes value 1 with probability p and value 0 with probability q = 1 − p.

  The Bernoulli distribution is a special case of the binomial distribution with n = 1. The kurtosis goes to infinity for high and low values of p, but for p = 1 / 2 the two-point distributions including the Bernoulli distribution have a lower excess kurtosis than any other probability distribution, namely −2. (WikiPedia, 2019)

  The Bernoulli distributions for 0 ≤ p ≤ 1 forms an exponential family.

  The maximum likelihood estimator of p based on a random sample is the sample mean.
- The binomial distribution, which describes the number of successes in a series of independent Yes/No experiments all with the same probability of success.

  The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used. (WikiPedia, 2019)
- The Poisson binomial distribution, which describes the number of successes in a series of independent Yes/No experiments with different success probabilities.

  The ordinary binomial distribution is a special case of the Poisson binomial distribution, when all success probabilities are the same, that is p 1 = p 2 = ⋯ = p n. (WikiPedia, 2019)
- The negative binomial distribution or Pascal distribution, a generalization of the geometric distribution to the nth success.

  The Pascal distribution (after Blaise Pascal) and Polya distribution (for George Pólya) are special cases of the negative binomial distribution. A convention among engineers, climatologists, and others is to use "negative binomial" or "Pascal" for the case of an integer-valued stopping-time parameter r, and use "Polya" for the real-valued case.

  For occurrences of "contagious" discrete events, like tornado outbreaks, the Polya distributions can be used to give more accurate models than the Poisson distribution by allowing the mean and variance to be different, unlike the Poisson. "Contagious" events have positively correlated occurrences causing a larger variance than if the occurrences were independent, due to a positive covariance term. (WikiPedia, 2019)
- The Poisson distribution, which describes a very large number of individually unlikely events that happen in a certain time interval. Related to this distribution are a number of other distributions: the displaced Poisson, the hyper-Poisson, the general Poisson binomial and the Poisson type distributions.

  The Poisson distribution is an appropriate model if the following assumptions are true.

  k is the number of times an event occurs in an interval and k can take values 0, 1, 2, ….

  The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently. The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals. Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur. Or The actual probability distribution is given by a binomial distribution and the number of trials is sufficiently bigger than the number of successes one is asking about (see Related distributions).

If these conditions are true, then k is a Poisson random variable, and the distribution of k is a Poisson distribution. (WikiPedia, 2019)

- The Beta distribution on [0,1], a family of two-parameter distributions with one mode, of which the uniform distribution is a special case, and which is useful in estimating success probabilities. The beta distribution has been applied to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines.

  In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions. For example, the beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success such as the probability that a space vehicle will successfully complete a specified mission. The beta distribution is a suitable model for the random behaviour of percentages and proportions.

  The usual formulation of the beta distribution is also known as the beta distribution of the first kind, whereas beta distribution of the second kind is an alternative name for the beta prime distribution. (WikiPedia, 2019)

Task 6: Ethics question

During your work, have you identified any issues that have ethical implications? Does it concern security or privacy? How do you mitigate the risk?

- ➔ Yes, there are some issues that actually have a major implication ethically; one of them being the "Global Warming" leading to extremely heating conditions around places. With the inevitable rise of extreme weather events, it is crucial that we better understand its potential impact on our everyday life.
- ➔ Yes, it concerns majorly on our security as more and more people face to be the victims of such crisis caused as a result of degradation and harmful activities caused by humans to the nature.
- ➔ In order to mitigate the risk we could take precautionary measure in order to safeguard the environment we life in and sort of protect it sustainably to maintain harmony and balance in the nature.

Task 7: Reflection

1. What help did you receive from other students? What did you learn from them?

→I received motivation from other students to work quickly and perfectly towards the assignment task given and explore more content on what needs to be done and moreover, learn additional content as well from extra materials widening my scope in the field.

2. Please estimate the mark that you will receive for assignment 1. Please provide both a point estimate and an interval estimate (a confidence interval). You don't need to provide a mathematical model, but please explain how you use conditional information to reach the estimates. Based on the conditional information, explain what you would have done differently to improve that mark?

→I should receive 90-95% of the marks in this assignment as I believe I have tried my level best to deliver the best quality work through research, learning a completely new language, excelling at it by working in real term data and models.

My confidence level is 95% and my point estimate would be 89.I have used the information provided and gone and delivered results specific to the questions asked and answered them correctly to reach solutions so that forms my basis of my estimates above.

I would have started learning about the subject in inter trimester breaks and started developing small app and codes just for practise and getting good grip on the content.

References

Australia, D. o. H. (. A., 2015. *Emergency Department Attendances, Admissions, and Admissions by Triage for 1 June 2013 - 30 June 2014,* s.l.: https://data.gov.au/dataset/ds-dga-6bfec5ea-207e-4d67-8965-c7e72290844b/details.

WikiPedia, 2019. List of probability distributions. *List of probability distributions*, 28 March, p. 7.