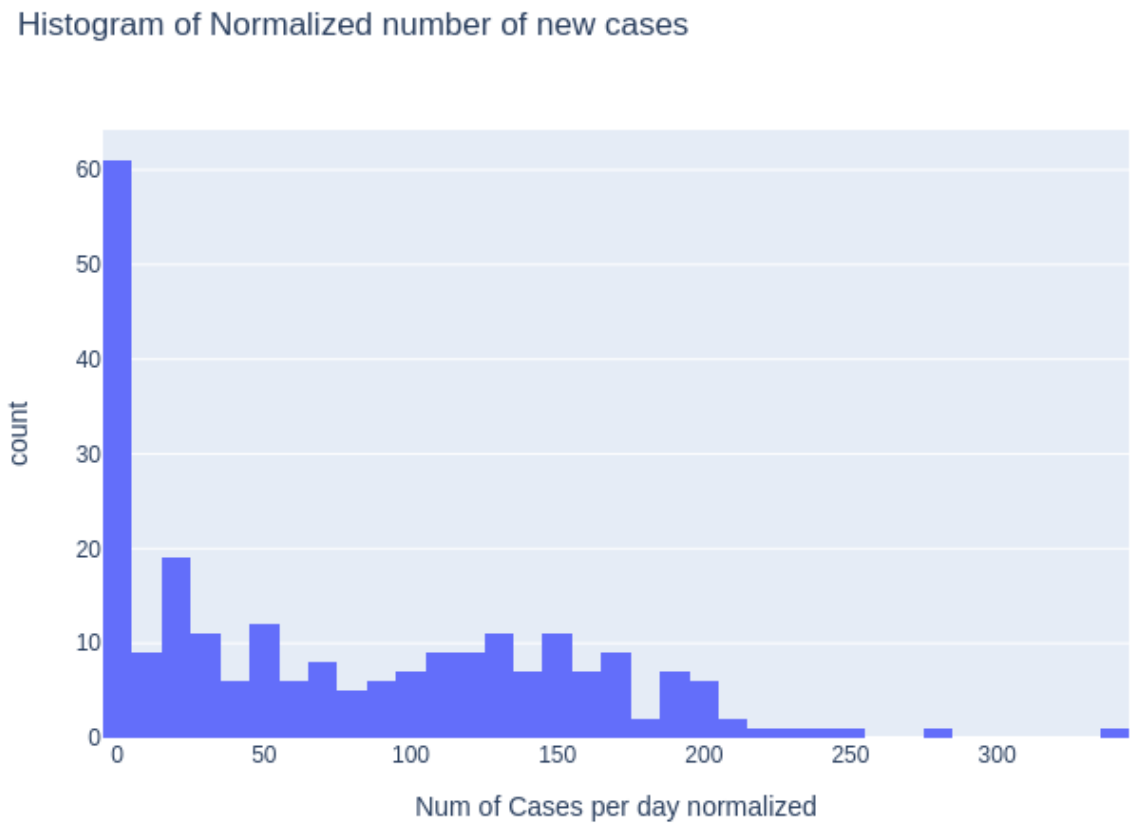
Data Science Project Stage-2
Submitted by: Rashmi Hassan Udaya Kumar

The first step in recognizing what sort of distribution to fit our data to, we observe that:

- The data is positive-valued, since we are measuring number of covid-19 cases per day.
- The data is discrete
- The data gives the number of covid-19 cases observed in period/time interval of a day
- The probability of the occurrence of the cases each day is equiprobable.

Next we plot the histogram of the Normalized Daily number of new cases.

The histogram is as shown below:



From the histogram we observe that

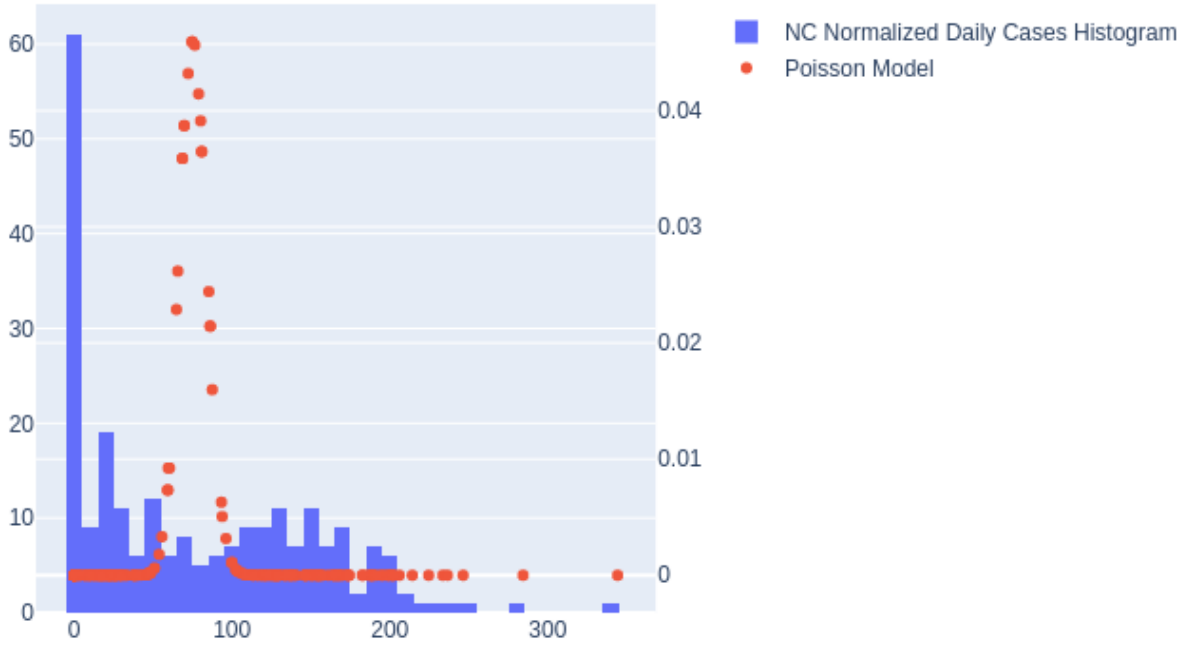
- The data is left skewed
- It starts off high and has a long tail

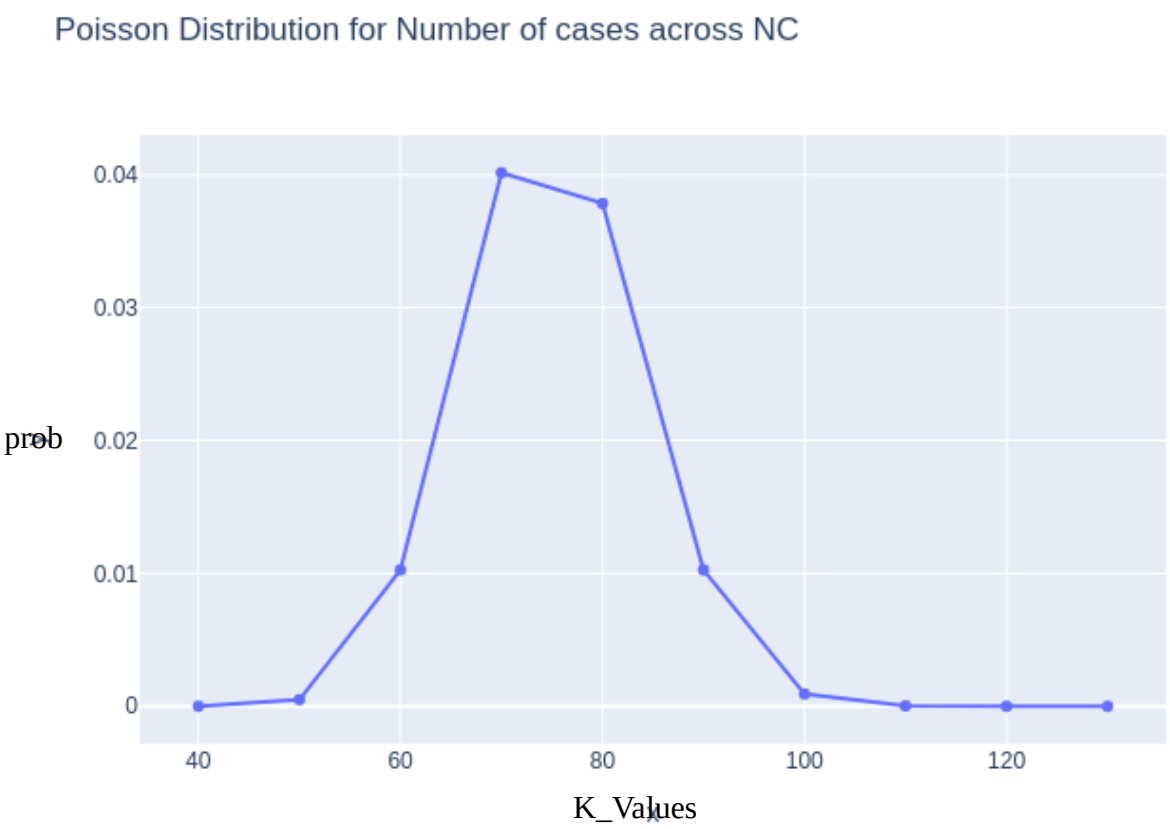
We know that the Poisson distribution models the probability of seeing a certain number of successes within a time interval, here we are measuring the number of covid-19 cases within a time interval of a day. Thus The Poisson distribution is a good fit.

Process of modeling the Poisson distribution

- Let us take the time period for the finding the number of cases to be a day
- We notice from the NC_weekly_data that the mean number of cases normalized for 1000000 people is 75.
- We use the mean value of the number of cases as the value of lambda
- We use poisson.pmf to obtain the probability of the number of cases and deaths for different values of k
- For finding the probability of number of cases we are using k = 40,50,60,70,80,90,100,110,120

By following the above procedure we obtain the following poisson distribution for the NC daily cases data.




Poisson Distribution for Number of cases across NC
prob
0.04
0.03
0.02
0.01
0
40
60
80
100
120
K_Values

Distribution Statistics of NC daily new cases is as shown below:

Distribution Statistics

Measure of center

```
In [50]: NC_data_groupedBydate['Num of Cases per day normalized'].mean()
Out[50]: 75.52396557038008
```

Measure of spread

```
In [51]: #variance
         NC_data_groupedBydate['Num of Cases per day normalized'].var()
Out[51]: 5252.865980896412
```

Skewness

```
In [52]: NC_data_groupedBydate['Num of Cases per day normalized'].skew()
Out[52]: 0.6997251548048288
```

Kurtosis

```
In [53]: NC_data_groupedBydate['Num of Cases per day normalized'].kurt()
Out[53]: -0.2861513970272158
```

Hypothesis :

- Is being Male or Female a factor for higher covid cases
- Does belonging to a particular age group influence increase in covid cases
- Does belonging to a particular ethnic group have a influence on the number of covid cases