Data Science Project phase – 2 report

Nikitha Narsing

Statistics of daily Cases in Texas state :

Mean, Median

```
state_stats = state_data_tx["Num of Cases Per Day Normalized"].agg(["mean","median"]).round()
state_stats
```

```
29]: mean      205.0
     median     76.0
     Name: Num of Cases Per Day Normalized, dtype: float64
```

Skewness

```
state_data_tx["Num of Cases Per Day Normalized"].skew()
```

```
30]: 1.4404417960889893
```

Kurtosis

```
state_data_tx["Num of Cases Per Day Normalized"].kurt()
```

```
31]: 0.9841217963066313
```

Variance

```
2]: state_data_tx["Num of Cases Per Day Normalized"].var()
```

```
[1232]: 72970.48862243061
```

```
3]: state_data_tx["Num of Cases Per Day Normalized"].describe()
```

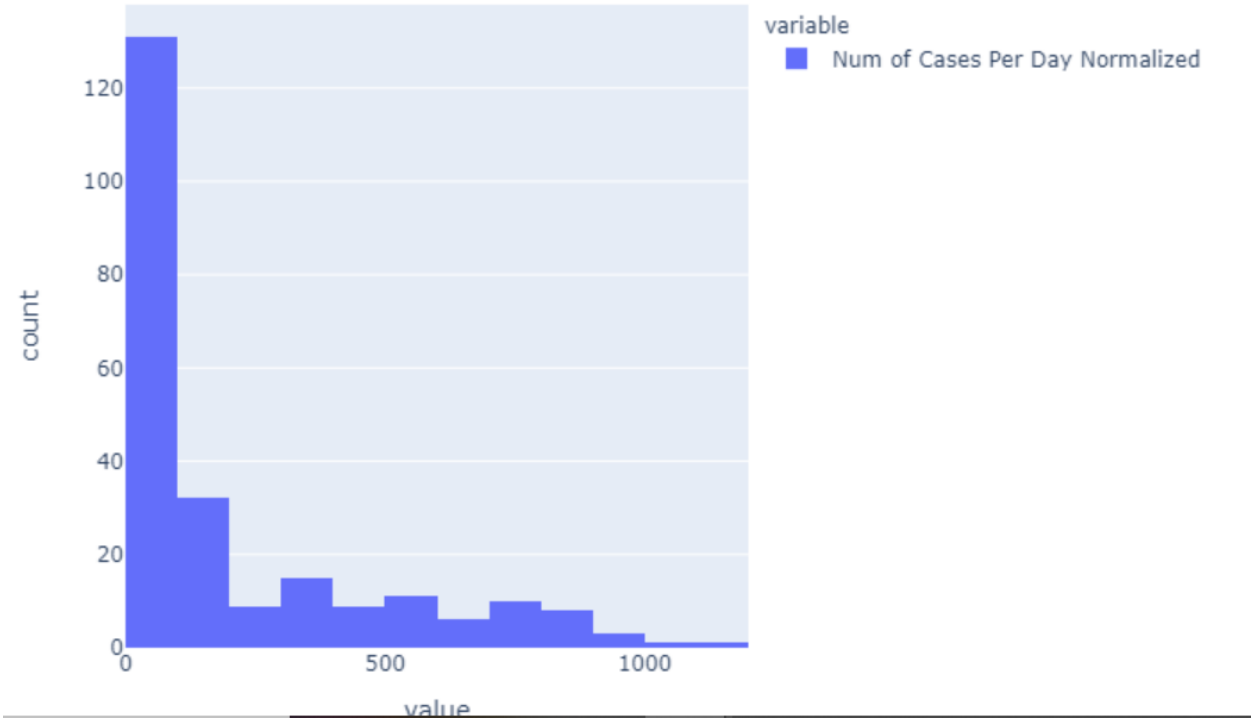
```
[1233]: count     236.000000
        mean      205.207627
        std       270.130503
        min         0.000000
        25%         3.500000
        50%        76.500000
        75%       320.500000
        max      1112.000000
        Name: Num of Cases Per Day Normalized, dtype: float64
```

Points from the statistics:

- Since **Mean is greater than Median** the distribution is not uniform, and it is positively skewed.

- Since the **skewness** is >1 we can say data is highly skewed and the skew is on the left and tail on the right.
- **Kurtosis** measures the peakedness of the distribution. Since we have kurtosis value as a positive value, we can say that the data has sharper peak.
- We can see that we have high **variance** which indicates that the number of cases are very spread out from the mean, and from one another.

Histogram for the Normalised Number of Cases of Texas state :

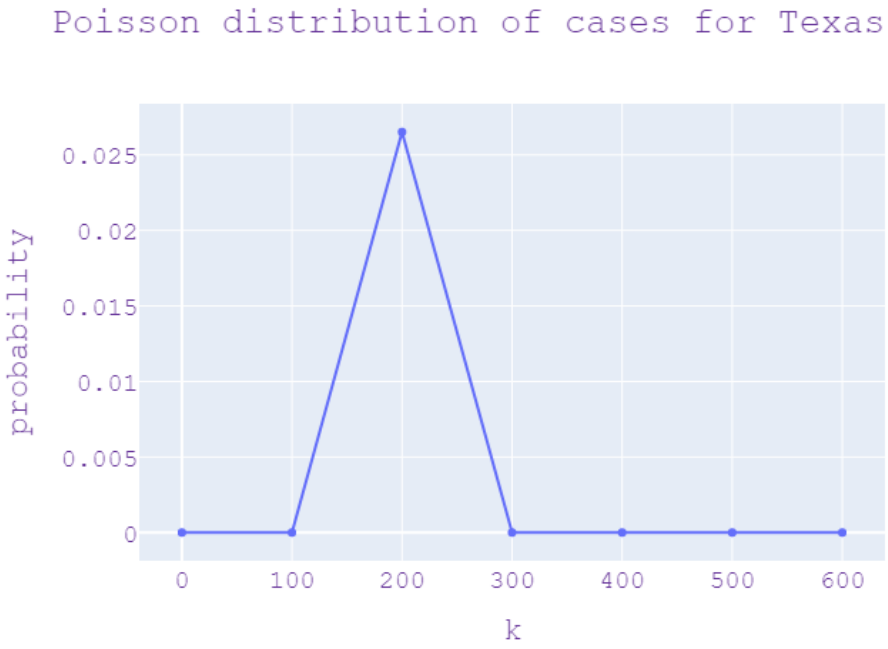


Distribution

- We can see the values(Number of Cases Per Day) are discrete
- A discrete Poisson probability distribution gives the probability of a given number of events occurring in a fixed interval of time, so here we have the number of times specific number of cases that occurred in a day.
- And we can see that the data is left-skewed with the tail to the right. Taking all these points into consideration I feel the Texas state data follows **Poisson distribution**.

Plotting the poisson distribution using pmf:

- We have calculated the mean of Number of new cases per day and taken that value as lambda.
- Then seeing the histogram we have taken the min and max values and using the number of bins take the range for ‘k’ value.
- For each k value calculate pmf using lambda and with the probabilities plot the Poisson distribution with ‘k’ on X-axis and the probability on Y-axis.


Poisson distribution of cases for Texas
probability
0.025
0.02
0.015
0.01
0.005
0
0
100
200
300
400
500
600
k