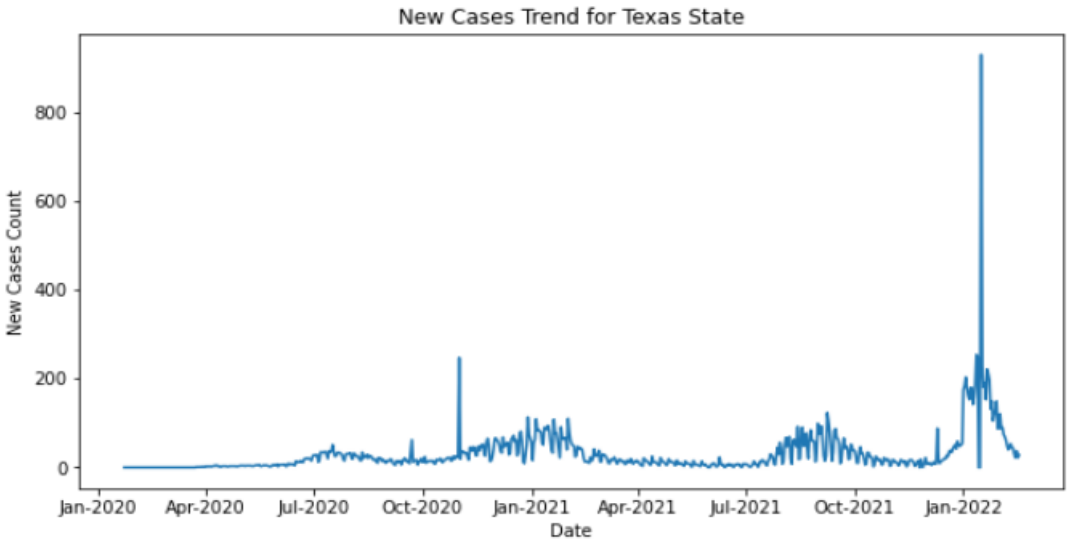
TASK 1

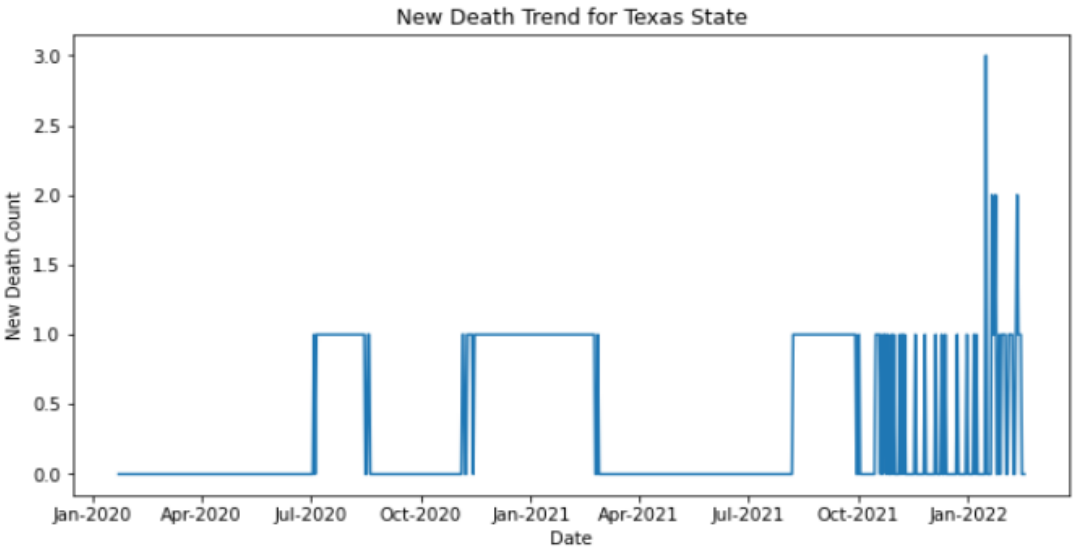
As part of **Member Task 1**, we are required to

- Generate weekly statistics (mean, median, mode) for number of new cases and deaths across a specific state. I have generated the statistics for Texas state.
- Compare the data against other states. (Normalize by population)
- Identify counties within the previous state with high case and death rates. I have compared Texas data with California, Florida, New York, North Carolina, and Illinois states. In this process, I have normalized the population count at 100000.
- Plot daily trends (new cases and deaths) of state and top 5 infected counties. I have utilized normalized, and log normalized values to plot the trends.

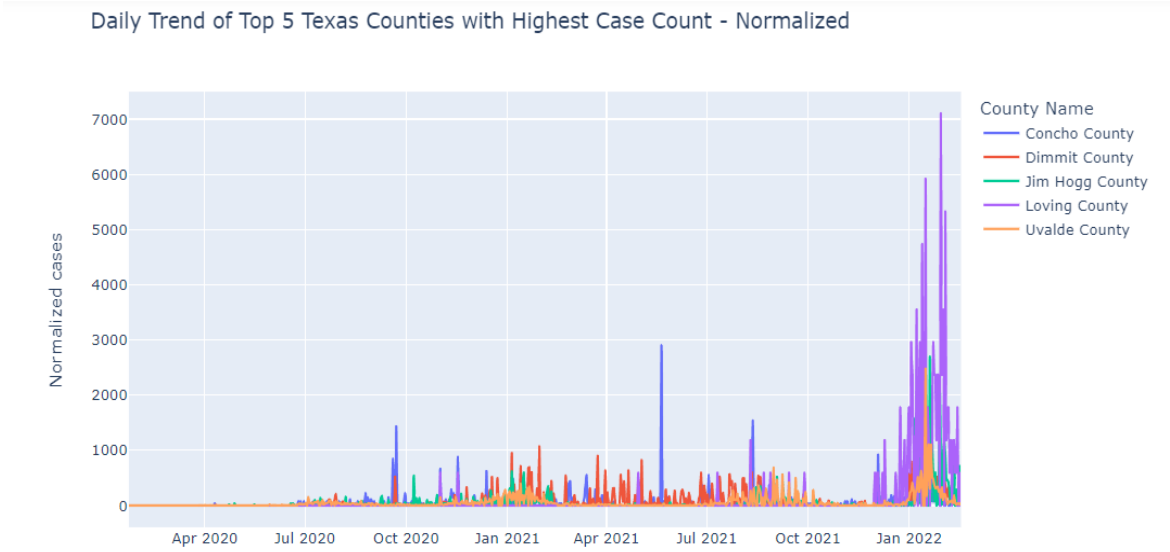
Daily Confirmed Cases [Normalized] – Texas State



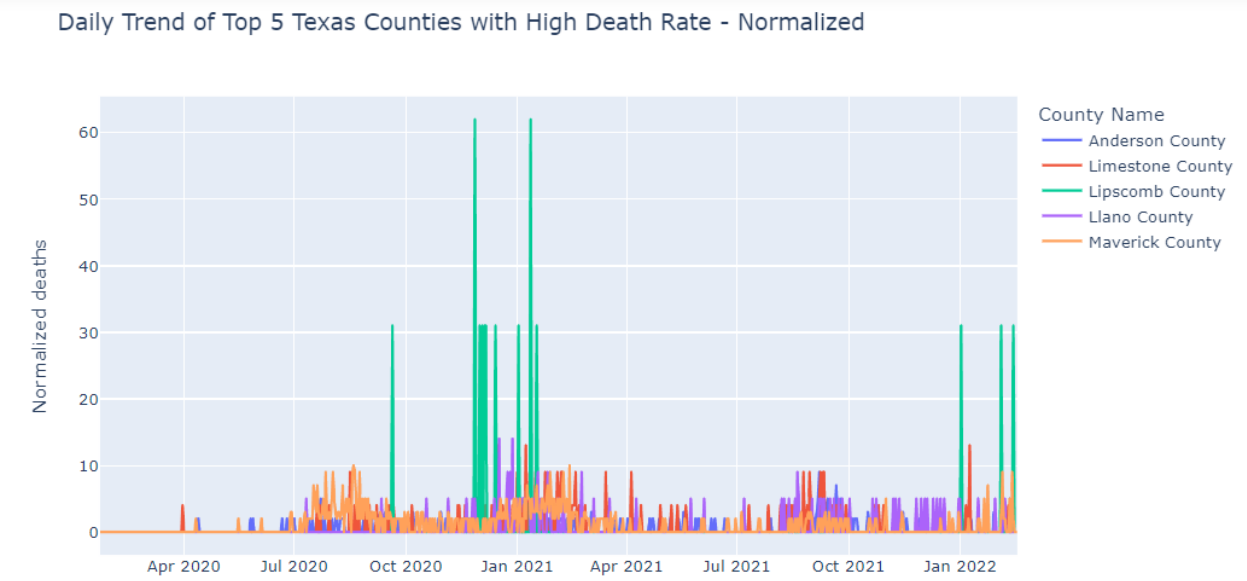
Daily Confirmed Deaths [Normalized] – Texas State



Top 5 Counties of Texas state with high number of cases [Normalized]



Top 5 Counties of Texas state with high number of deaths [Normalized]



TASK 2

Statistics of Cases in Texas State – Day Level

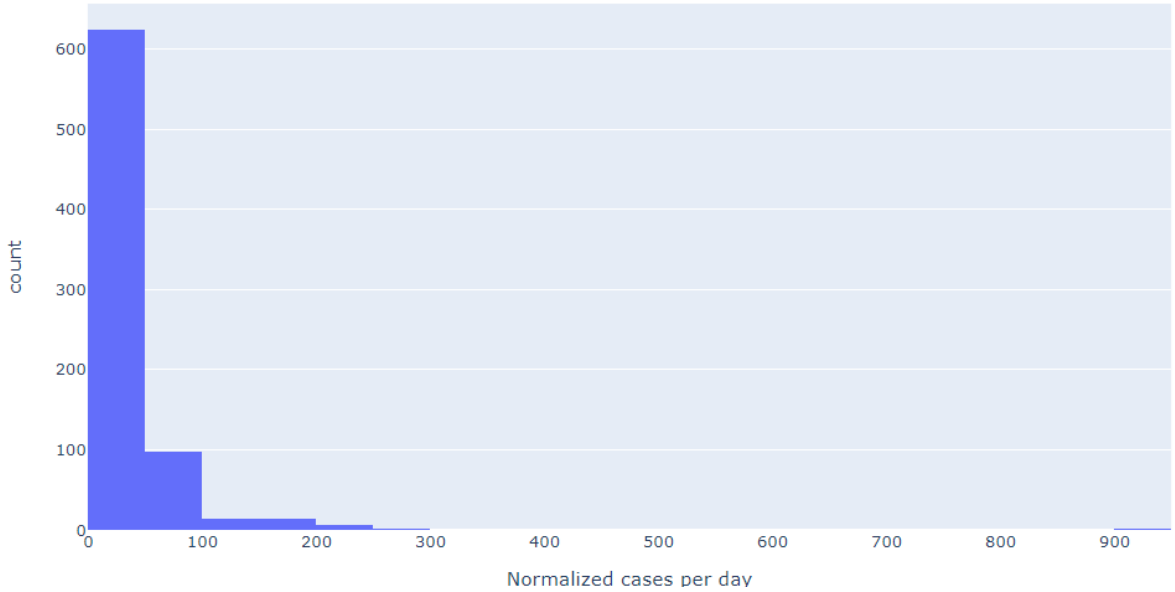
```
print("mean =",TX_data_day["Normalized cases per day"].mean())
print("mean =",TX_data_day["Normalized cases per day"].median())
print("variance =",TX_data_day["Normalized cases per day"].var())
print("skewness =",TX_data_day["Normalized cases per day"].skew())
print("kurtosis =",TX_data_day["Normalized cases per day"].kurt())
```

```
mean = 29.688243064729193
mean = 16.0
variance = 2503.3233139725876
skewness = 8.90120011041223
kurtosis = 140.35915510086807
```

Observations

- As Mean is greater than Median, the distribution is not uniform and is positively skewed.
- Since Kurtosis value is a positive value, we can say that the data has sharp peak.
- The high variance values indicate the number of cases is very spread out from the mean and from each other.

Histogram for the Normalized Number of Cases of Texas state



Observations

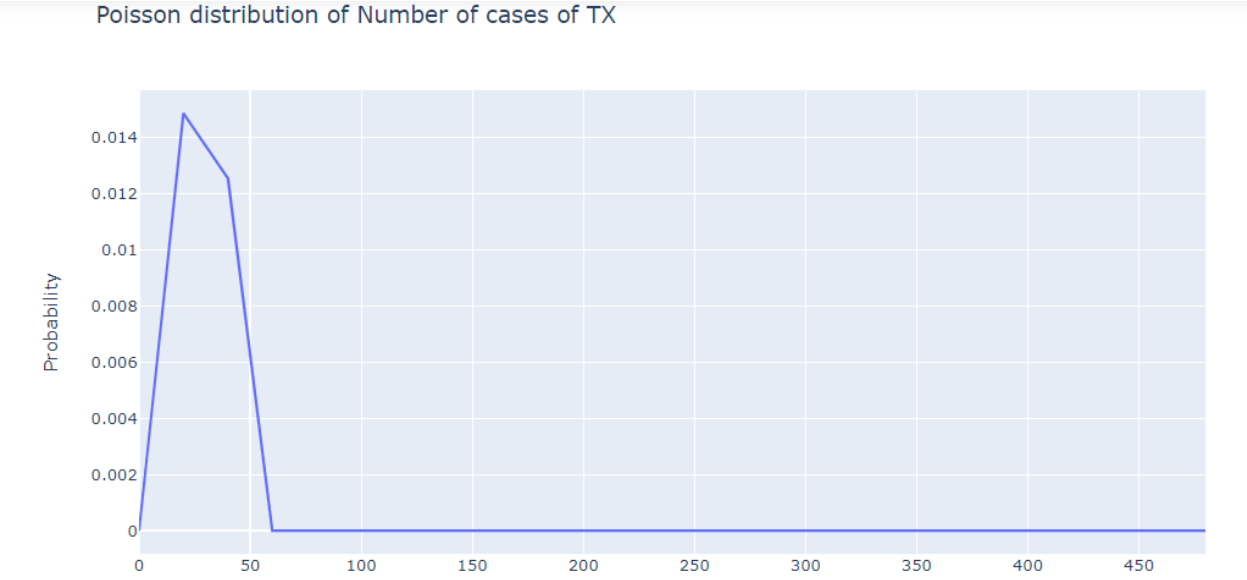
- The number of cases is a discrete value.
- A discrete Poisson probability distribution gives the probability of given number of events occurring in a time interval. Here, we have number of times specific number of cases that occurred in a day.
- The data is left skewed.

Considering all these points, I have fit the data into Poisson distribution.

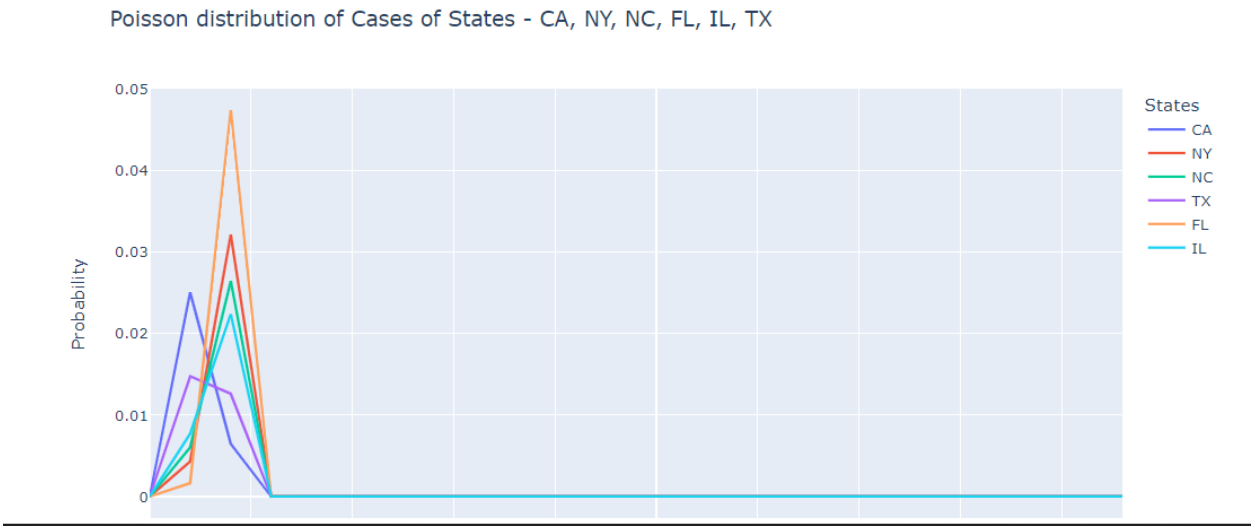
Plotting Poisson Distribution

- Mean has been calculated for number of cases per day which can be considered as lambda value.
- By looking at the histogram, we have taken min and max values that are used in calculating k value.
- For each k value, we calculate pmf using lambda.
- Finally, plot the Poisson distribution with ‘k’ on X-axis and probability on Y-axis.

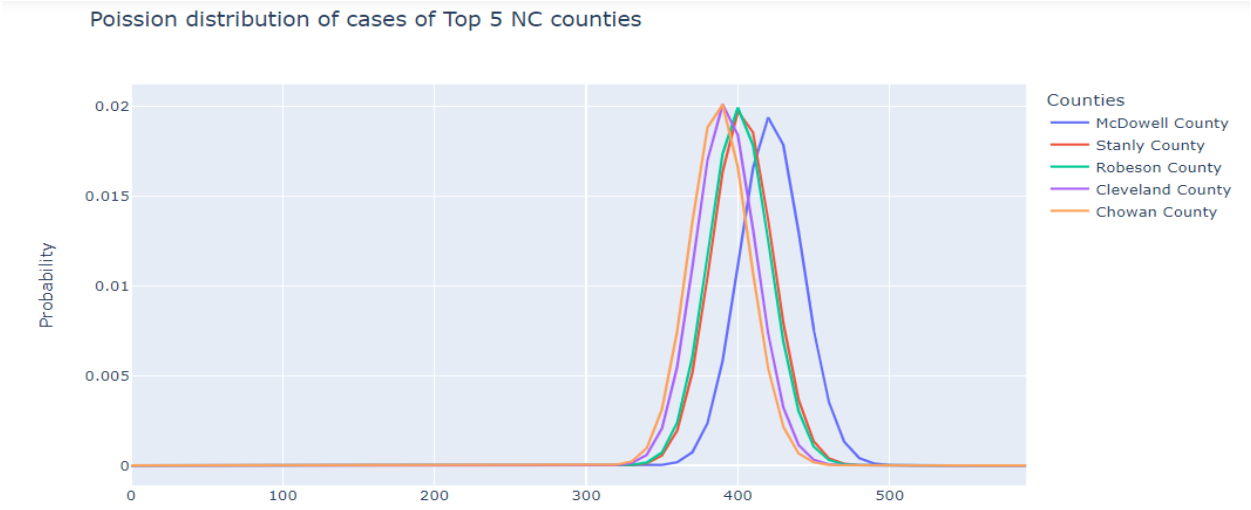
Poisson Distribution of Case Count of Texas



Poisson Distribution of Case Count of CA, NY, NC, FL, IL, TX



Poisson Distribution of Case Count of Top 5 NC Counties



Performing correlation between Enrichment data variables and COVID-19 cases to observe any patterns

```
#calculating correlation
correlation = Merged_Demographic_Data_LatestDate_byState.corr().round(2)
```

```
correlation.style.background_gradient(cmap='coolwarm')
```

	Unnamed: 0	countyFIPS	StateFIPS	Cases	Death	population	Total_Population	Total_Male	Total_Female	Gender_ratio	Age_unc
Unnamed: 0	1.000000	1.000000	1.000000	0.350000	0.410000	0.360000	0.350000	0.350000	0.350000	0.790000	0.400
countyFIPS	1.000000	1.000000	1.000000	0.370000	0.430000	0.380000	0.370000	0.370000	0.370000	0.830000	0.420
StateFIPS	1.000000	1.000000	1.000000	0.370000	0.430000	0.380000	0.370000	0.370000	0.370000	0.830000	0.420
Cases	0.350000	0.370000	0.370000	1.000000	0.950000	0.990000	0.990000	0.990000	0.990000	0.460000	0.980
Death	0.410000	0.430000	0.430000	0.950000	1.000000	0.950000	0.950000	0.950000	0.950000	0.510000	0.950
population	0.360000	0.380000	0.380000	0.990000	0.950000	1.000000	1.000000	1.000000	1.000000	0.470000	0.990
Total_Population	0.350000	0.370000	0.370000	0.990000	0.950000	1.000000	1.000000	1.000000	1.000000	0.460000	0.990
Total_Male	0.350000	0.370000	0.370000	0.990000	0.950000	1.000000	1.000000	1.000000	1.000000	0.460000	1.000
Total_Female	0.350000	0.370000	0.370000	0.990000	0.950000	1.000000	1.000000	1.000000	1.000000	0.460000	0.990
Gender_ratio	0.790000	0.830000	0.830000	0.460000	0.510000	0.470000	0.460000	0.460000	0.460000	1.000000	0.500
Age_under5	0.400000	0.420000	0.420000	0.980000	0.950000	0.990000	0.990000	1.000000	0.990000	0.500000	1.000

From the correlation matrix, we can say that,

- There is a high correlation between Number of cases with Population, Sex, different age groups(particularly Male and Female over 65 yrs of age)

  > Both Male and Female population are equally prone to Covid

- There is a weak correlation between covid case or deaths with different races

Hypothesis

- Do seniors have a greater death rate than those in lower age groups.
- Is there a particular race that was hit the hardest by the virus when the various races were compared.
- Is any specific gender highly prone to virus.