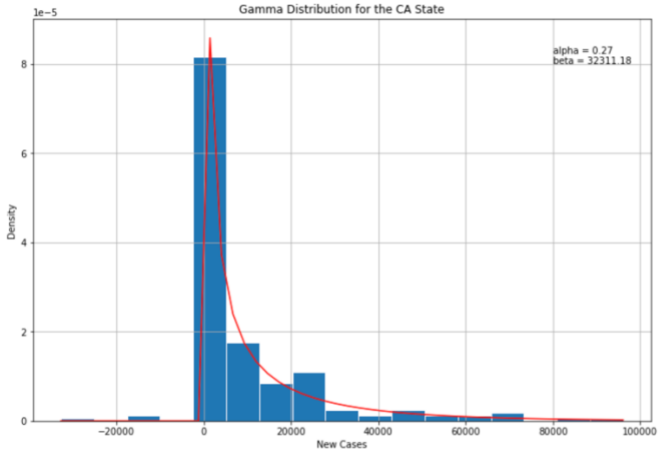
SAIPAVAN TADIKONDA STAGE-3 REPORT

1) In Stage-2 I took California State. So, plotting various distributions and calculating moment of distributions for CA state for each used distribution:

Using Gamma Distribution:

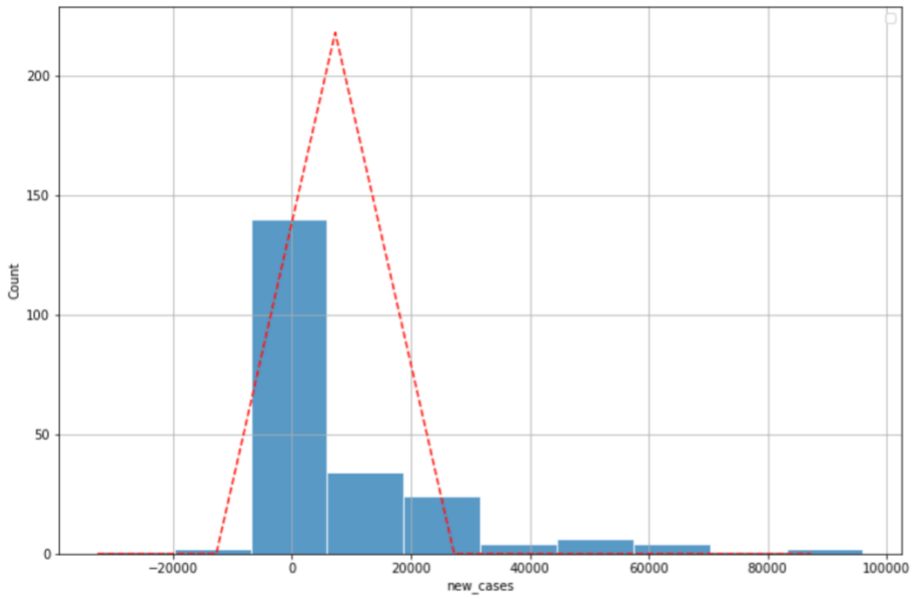
```
Moment of Distribution = 0.27426724620239956 0.27426724620239956 3.818941648931559 21.87647297691614
```



From the above plot we can observe that gamma distribution using mom didn't fit correctly for some part of the histogram on the new cases.

Using Poisson Distribution:

```
Moments of Distribution =  8861.899082568807 8861.899082568807 0.010622741175524742 0.00011284263008218878
```

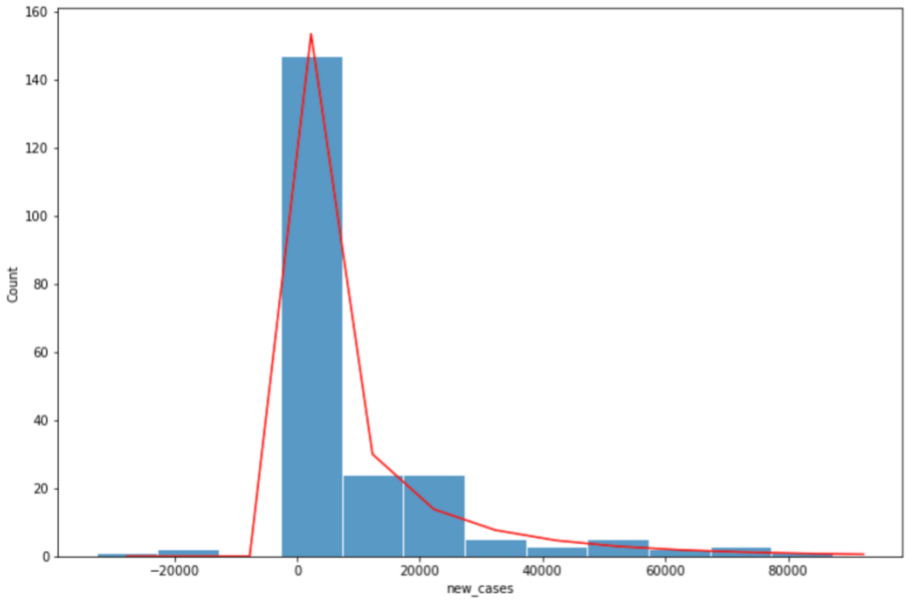


From the above plot we can observe that poisson distribution is not so good fit for the California State data as the data is widely spread i.e., data is over-dispersed. If the data is widely spread then that poisson distribution is not good. Also if you observe the mean and variance differ very much. But in poisson mean and variance are equal. So poisson distribution is not a good fit for this data

As we came to know that our data is widely spread let us know try negative binomial distribution which is somewhat good in this type of cases. Now let us try to fit the negative binomial distribution using the MoM method. For that let us first find the variance and mean. Using this mean and variance let us find the n, p and r values. After that using these values plot the negative binomial distribution on the new cases histogram.

Using Negative Binomial Distribution:

```
Moment of Distribution=  8861.899082568807 286338439.74092007 3.818882552647585 21.87579592462802
[<matplotlib.lines.Line2D at 0x7f800f8bcc10>]
```



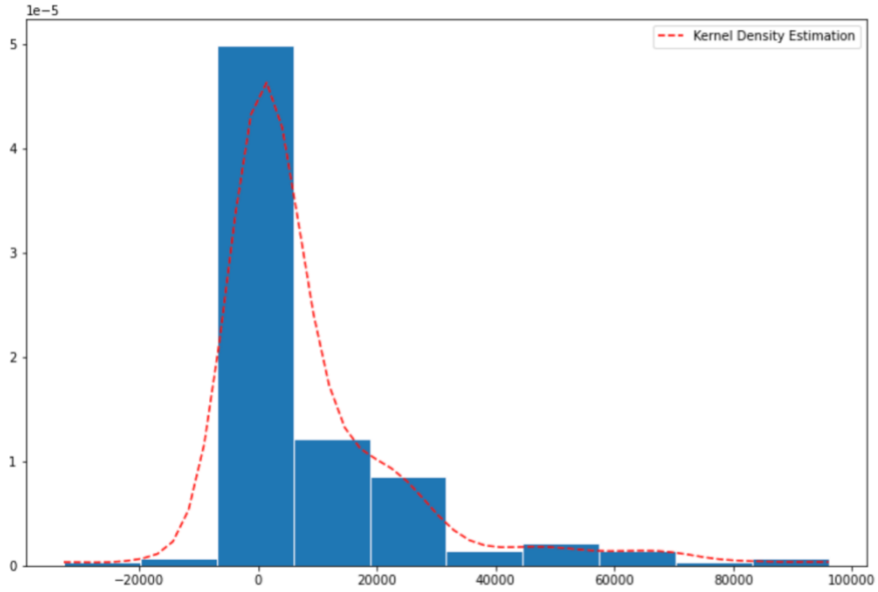
From the above plot we can see that negative binomial distribution using MoM fitted good for the California state new cases data compared to the poisson and gamma using MoM.

Till now we tried to implement the distributions using MoM method. Now let us try to implement using the Kernel Density Estimation Method. We know that the Kernel Density Estimation is a non-parametric method of estimating the PDF function. It is non-parametric method because it does not assume any underlying distribution. In Kernel Density estimation we use the gaussian kernel. The advantage of Gaussian kernel is that it automatically determines the bandwidth which is the important parameter of the kernel density estimator. Now plotting the kernel density estimation using the scipy stats kde on the new cases histogram.

Using Kernel Density Estimation:

```
Moments of Distribution =  0.0 285024960.6595404 2.3749707776792013 6.803399438550649
```

```
: <matplotlib.legend.Legend at 0x7f801106ba90>
```



From the above plot we can observe that the Kernel Density Estimation method is a very good fit compared to the MoM method. If we try to adjust the bandwidth between 0.1-0.5 then we could get optimally smoothed kernel density estimate for the new cases data of CA State.

After this repeated the same distributions and calculated the moments of distributions for the selected states NC, NJ, and NY.

Compare the distribution and its statistics to 3 other states of your choosing.

Let us see the moment of distributions obtained for each selected state in a data frame:

	State	Mean	Variance	Skewness	Kurtosis
0	CA	8861.899083	285024960.7	2.374971	6.803399
1	NC	2623.298165	53952872.0	2.799712	6.647819
2	NJ	2380.316514	19482974.7	2.367865	6.013621
3	NY	4831.096330	45247698.0	2.720358	10.184112

From the above moments of distribution for each state we can see each state follows the same trends. Coming, to the Skewness the values are almost same for each state and are positive that means each state data are right skewed. Right tails are more compared to

the left tails. Coming to the kurtosis the NY state has the highest kurtosis remaining states are almost similar. So, the NY state data is highly tailed.

Describe if the distributions look different and what does that imply.

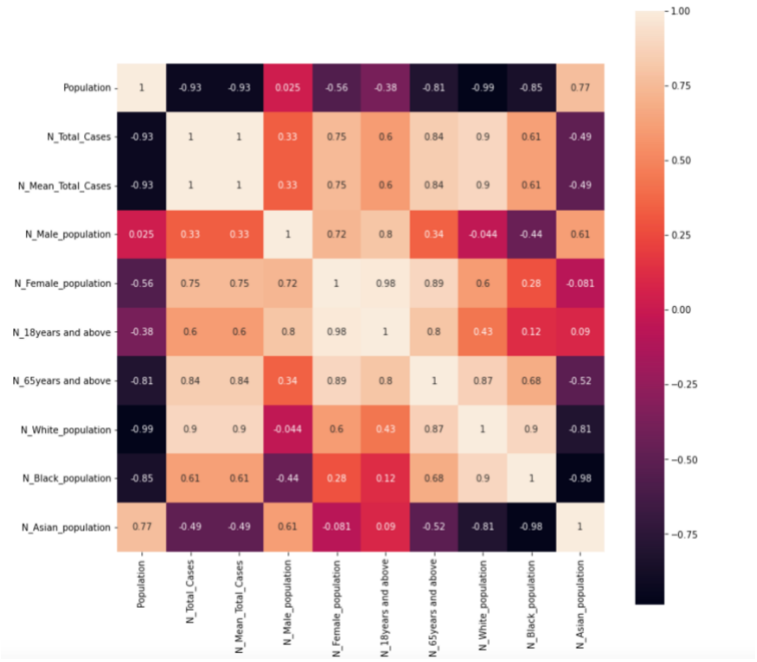
Distributions for each state looks different because each state registers different number of cases. Histogram spread depends on the number of new cases so the spread would be different from one state to another.

Coming to the distributions for each state, negative binomial distribution using MoM is more suitable for state whose data is more widely spread followed by the gamma distribution. If state data is not widely spread, then the Poisson distribution will be a very good fit. Also, if we observe the state data which we have with us it is a discrete data not continuous. So discrete distribution fits good for our data.

Apart from all these distribution Kernel density estimation method works very good compared to the method of moments method. We can get best kernel density estimate by giving the bandwidth parameter. If we select between 0.1 - 0.5 then we can see optimally smoothed kernel density estimate.

2) Correlation:

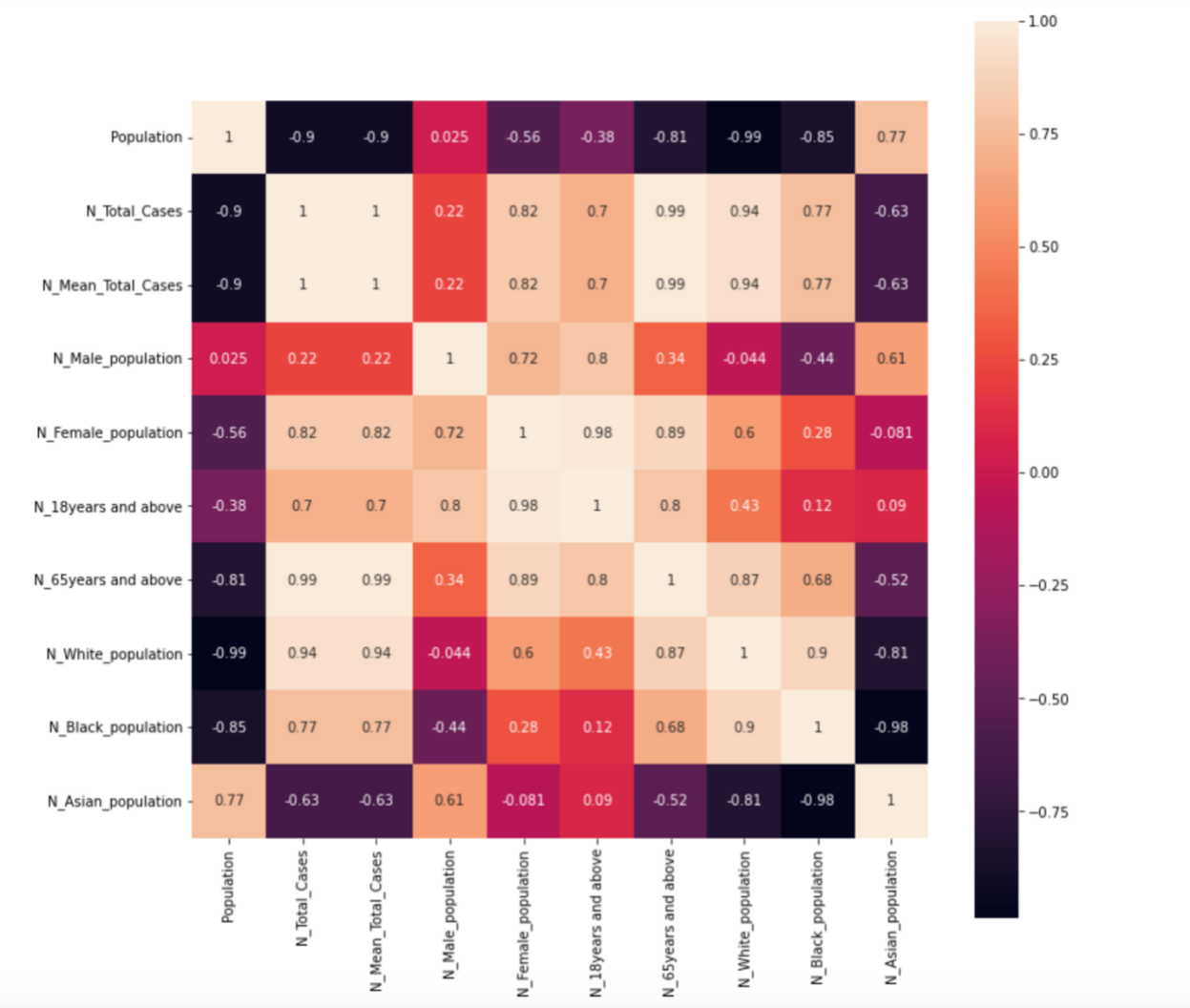
We know that the correlation coefficient always lies in between 1 and -1. Like covariance, positive correlation coefficient indicates that the columns are directly related, and negative coefficient indicates that the columns are inversely related to each other. If the correlation coefficient is near to zero, then that columns are weakly related to each other.



From the first heatmap which is on the Normalized data from date range May 30th, 2022, to Jan 2nd, 2023, if we observe the correlation coefficient is high between Mean_total_Cases and Female population, White population, 65years and above, Black population,18years and above. And correlation coefficient is very less (Nearly zero) between Mean_Total_Cases and Male population and correlation coefficient is negative between Mean_Total_Cases and Asian population.

Correlation coefficients are also similar between Total_Cases and other columns as mentioned in the above description.

Now let us see the second heatmap which is performed on date range from Jan 22nd, 2020, till Jan 16th, 2023.



Coming to the second heatmap the results obtained is similar to the results obtained in the first heatmap. The relationship between columns mentioned in the first heatmap valid's for the second heatmap too.

3) Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose 3 different variables to compare against.

- Does the increase in white population increases in the spread of covid-19 cases?
- Covid cases are more in the age group of 65 years and above.
- Covid cases spread is low in the Male population.
- Covid cases are less among the Asian population.