STAGE-3

1) Use the state data (the state of your choice) generated in Stage II to fit a distribution to the number of COVID-19 **new** cases

   i) Graphically plot the distribution and describe the distribution statistics. If using discrete values, calculate the Probability Mass Function for the individual values or range (if using histogram) and plot that.

   The State which I chose in the stage 2 is NC. I imported the weekly data of NC generated in the stage2 and the plotted the histogram on Normalized new cases. The histogram is as follows:

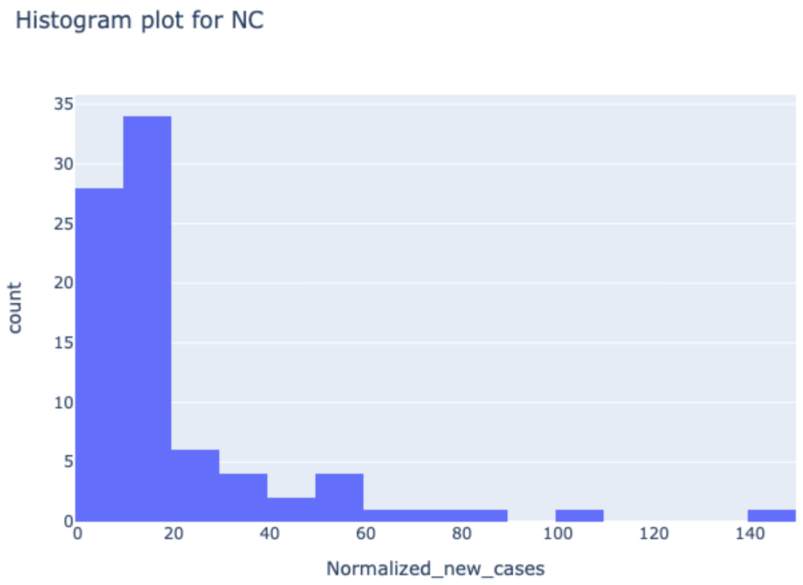


Fig.1

In general, I applied gamma distribution to the above histogram plot by calculating the pdf and the gamma distribution is as follows:

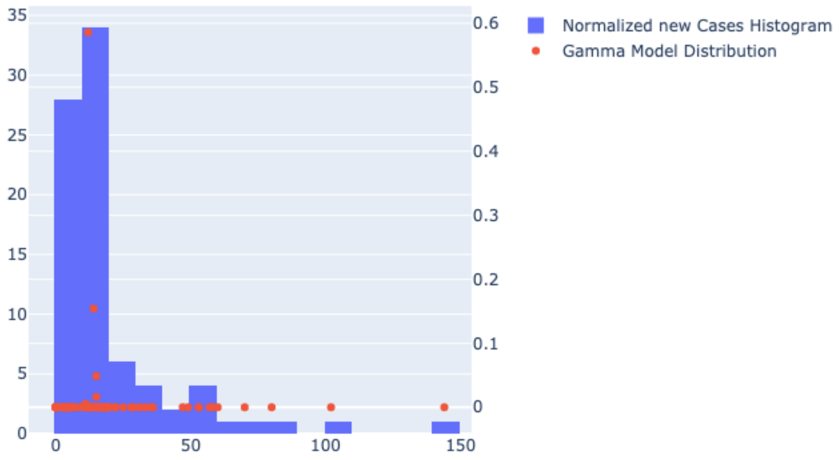


Fig. 2

From the Fig.2 we can observe that the gamma distribution is not a good fit for our data. Some of the observations of our data is:

1) The data we have are discrete values.
2) The data has positive values as we calculated the covid cases per week.
3) It starts of high and has long tail.
4) We can also observe that the data is right skewed.

We know that the Poisson distribution model gives us the probability of a given number of events happening in a fixed interval of time and our data is also of same kind the poisson distribution is the best fit for our data.

Applying poisson distribution to the North Carolina weekly cases data by calculating the pmf using the range of the histogram, then the poisson distribution is as follows:

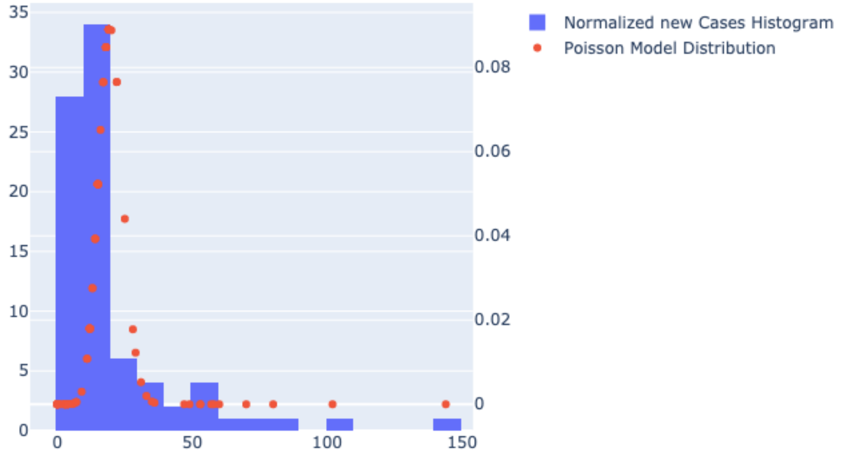


Fig.3

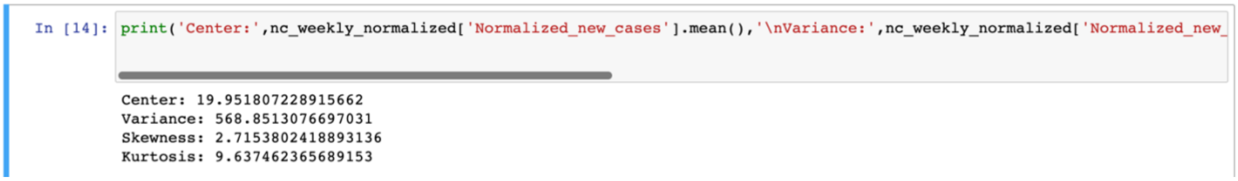
From the Fig.3 we can see that poisson distribution almost fits for our data. It missed some peaks, but it is fitting most of the data.

ii) Describe the type of distribution (modality) and its statistics (moments of a distribution - center, variance, skewness, kurtosis) in the report and the notebook.

From the fig.2 we can say that the gamma distribution didn't have any peaks. So, for the gamma distribution we cannot mention the modality.

But from the fig.3 we can observe that the poisson distribution has one peak so we can say that it is unimodal.

Moments of distribution for North Carolina weekly cases is as follows:



```
In [14]: print('Center:',nc_weekly_normalized['Normalized_new_cases'].mean(),'\nVariance:',nc_weekly_normalized['Normalized_new_
```

```
Center: 19.951807228915662
Variance: 568.8513076697031
Skewness: 2.7153802418893136
Kurtosis: 9.637462365689153
```

iii) Compare the distribution and its statistics to 5 other states of your choosing. Describe if the distributions look different and what does that imply.

The histogram of five other states are as follows:

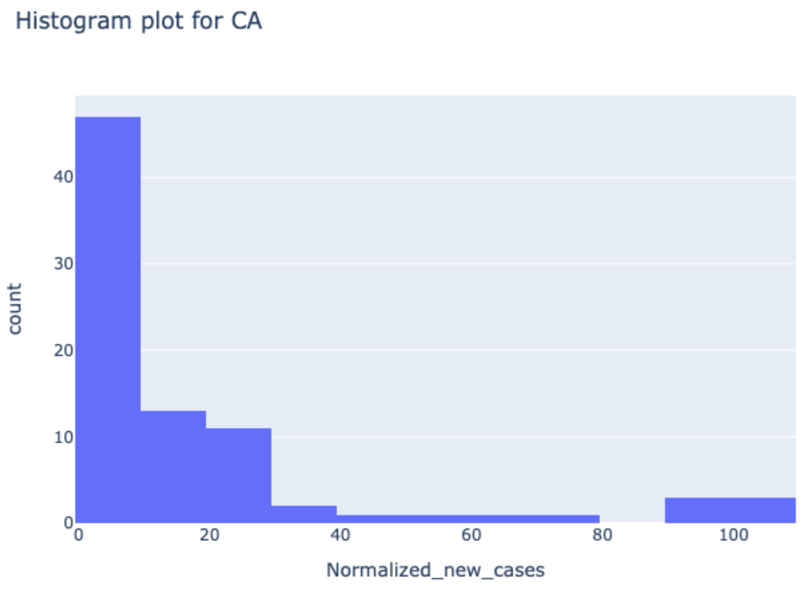


Fig.4

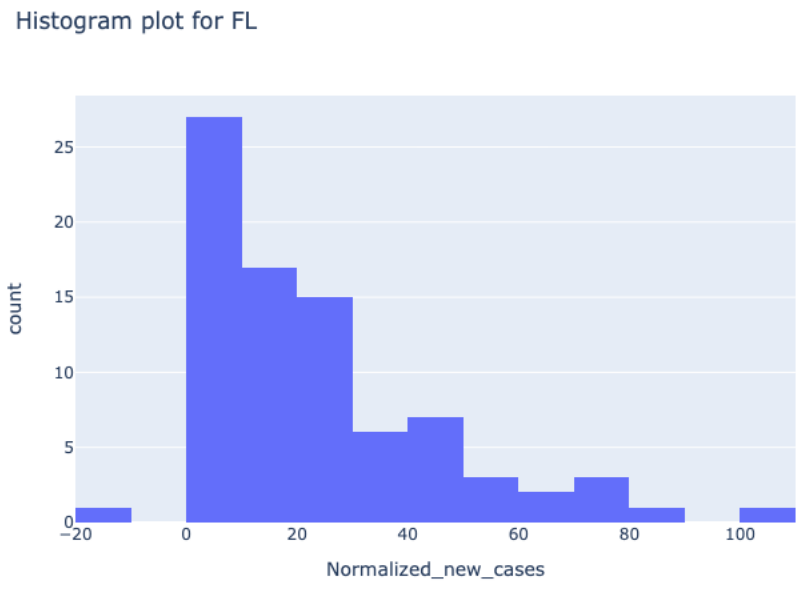


Fig.5

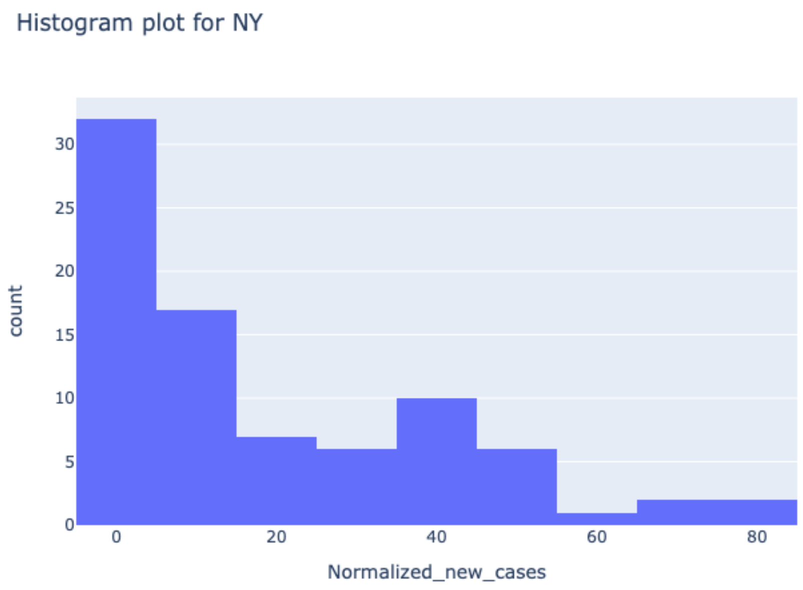

Histogram plot for NY
count
0
5
10
15
20
25
30
0
20
40
60
80
Normalized_new_cases

Fig.6

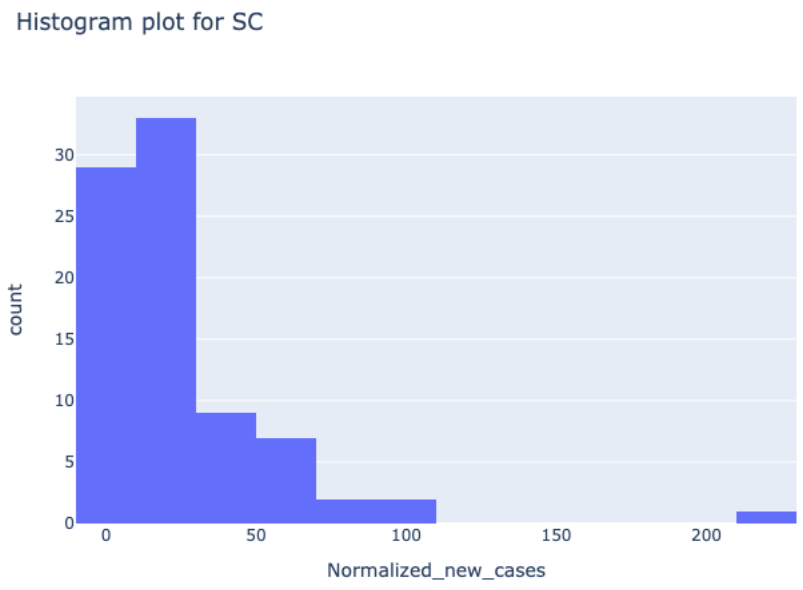

Histogram plot for SC
count
0
5
10
15
20
25
30
0
50
100
150
200
Normalized_new_cases

Fig.7

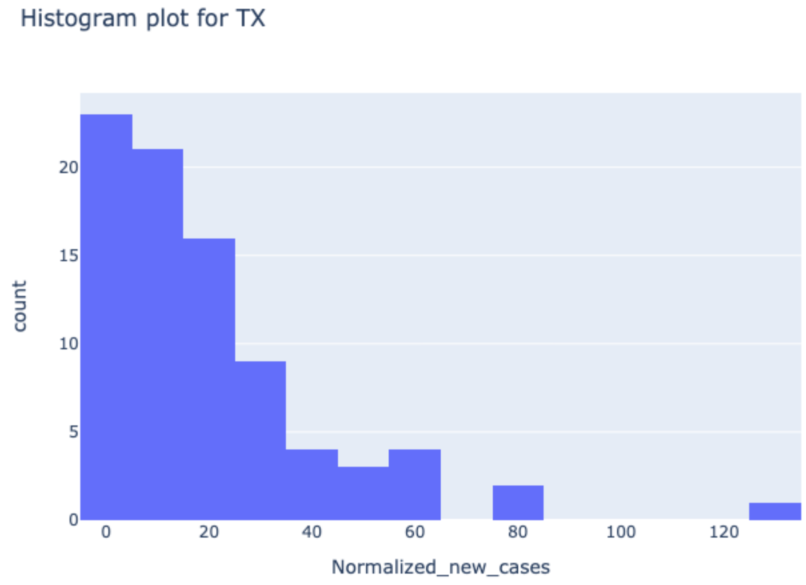
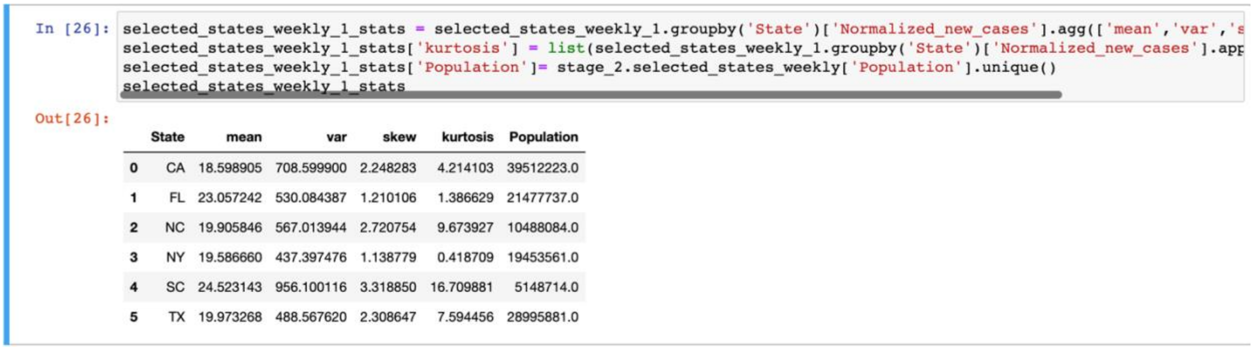


Fig.8

From the Fig.4, Fig.5, Fig.6, Fig.7 & Fig.8 we can say that for all the selected states weekly cases are right skewed. The values start off high and has long tail at the end. And all the values are discrete. But the count of weekly cases is different. This is because those selected states have different population. Some states like South Carolina have less population but states like California has more population.

But from all the above hist plots we can say that poisson distribution would be a good fit for all states.

Comparison of Statistics among selected states is as follows:



```
selected_states_weekly_1_stats = selected_states_weekly_1.groupby('State')['Normalized_new_cases'].agg(['mean','var','s
selected_states_weekly_1_stats['kurtosis'] = list(selected_states_weekly_1.groupby('State')['Normalized_new_cases'].app
selected_states_weekly_1_stats['Population']= stage_2.selected_states_weekly['Population'].unique()
selected_states_weekly_1_stats
```

	State	mean	var	skew	kurtosis	Population
0	CA	18.598905	708.599900	2.248283	4.214103	39512223.0
1	FL	23.057242	530.084387	1.210106	1.386629	21477737.0
2	NC	19.905846	567.013944	2.720754	9.673927	10488084.0
3	NY	19.586660	437.397476	1.138779	0.418709	19453561.0
4	SC	24.523143	956.100116	3.318850	16.709881	5148714.0
5	TX	19.973268	488.567620	2.308647	7.594456	28995881.0

From the above statistics we can observe that the South Carolina state has high kurtosis. It means that’s its data is highly tailed. New York has very low kurtosis value.

Coming to the skewness all states skew values are positive. So, we can say that all states data are right skewed. It means right tails are long compared to left.

2) Model a poission distribution of **new** COVID-19 cases and deaths of a state and compare to other 5 states. Describe how the poission modeling is different from the first modeling you did

The poisson distribution models of new cases among all the 6 states including North Carolina is as follows:

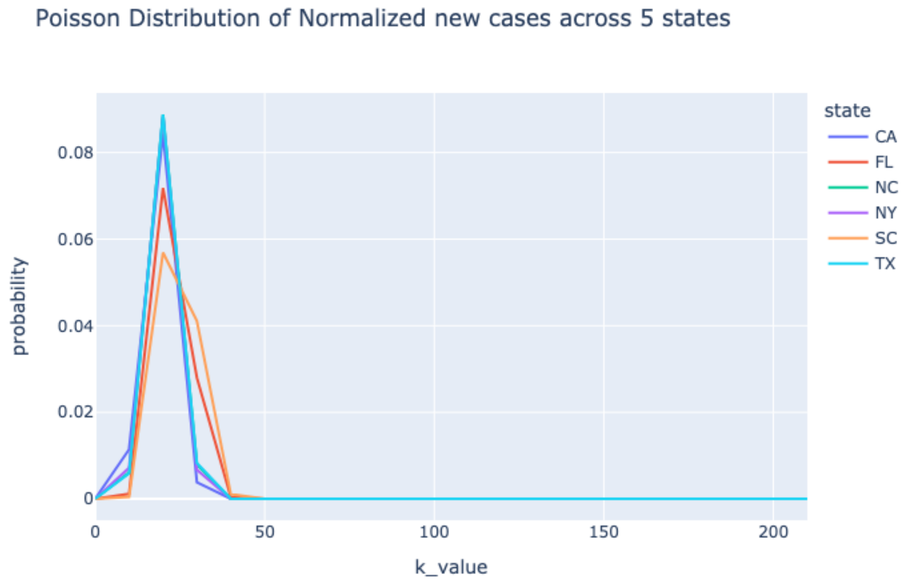
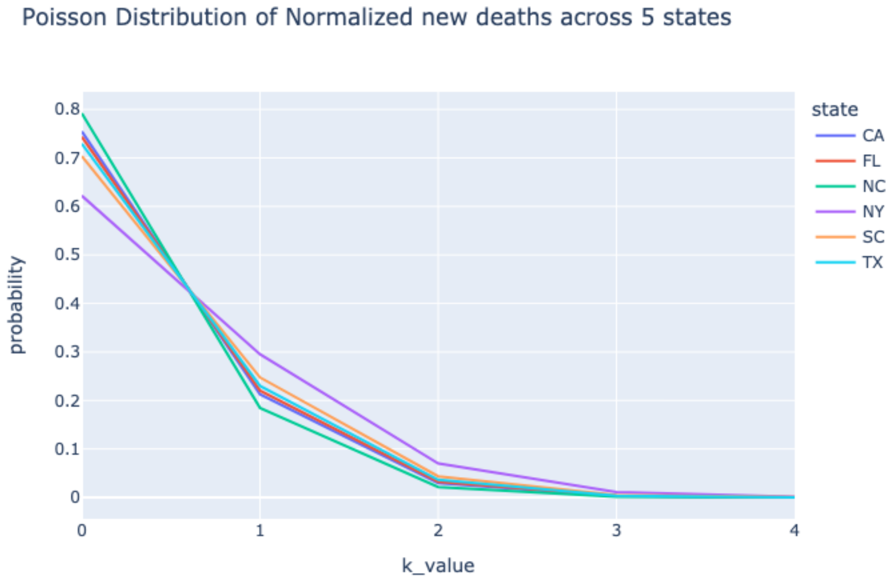


Fig.9

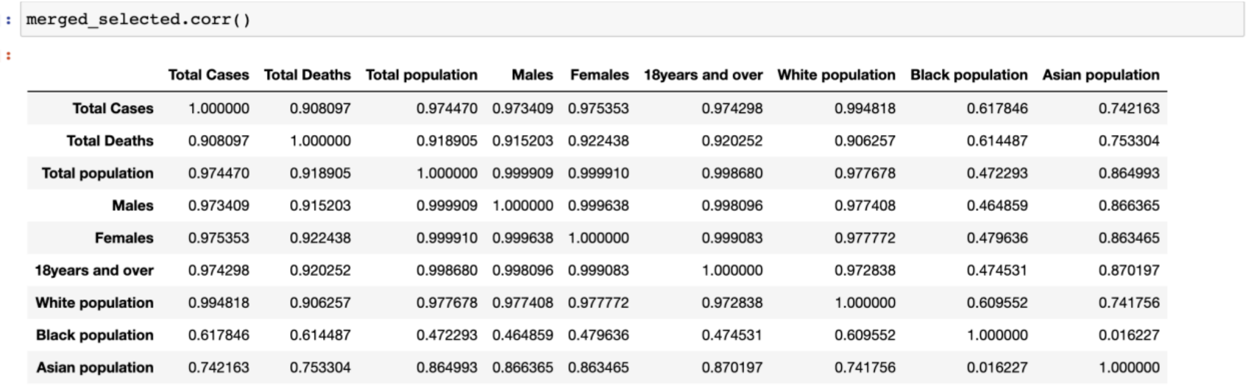
The poisson distribution models of new deaths among all the 6 states is as follows:



The poisson distribution model is different from the gamma distribution model as the poisson distribution model deals with discrete values and even the data we have is discrete. So, poisson distribution is best fit for the comparison.

3) Perform corelation between Enrichment data valiables and COVID-19 cases to observe any patterns

   Correlation between Enrichment data variables and covid-19 cases among the selected states is as follows:



```
: merged_selected.corr()
```

:

	Total Cases	Total Deaths	Total population	Males	Females	18years and over	White population	Black population	Asian population
Total Cases	1.000000	0.908097	0.974470	0.973409	0.975353	0.974298	0.994818	0.617846	0.742163
Total Deaths	0.908097	1.000000	0.918905	0.915203	0.922438	0.920252	0.906257	0.614487	0.753304
Total population	0.974470	0.918905	1.000000	0.999909	0.999910	0.998680	0.977678	0.472293	0.864993
Males	0.973409	0.915203	0.999909	1.000000	0.999638	0.998096	0.977408	0.464859	0.866365
Females	0.975353	0.922438	0.999910	0.999638	1.000000	0.999083	0.977772	0.479636	0.863465
18years and over	0.974298	0.920252	0.998680	0.998096	0.999083	1.000000	0.972838	0.474531	0.870197
White population	0.994818	0.906257	0.977678	0.977408	0.977772	0.972838	1.000000	0.609552	0.741756
Black population	0.617846	0.614487	0.472293	0.464859	0.479636	0.474531	0.609552	1.000000	0.016227
Asian population	0.742163	0.753304	0.864993	0.866365	0.863465	0.870197	0.741756	0.016227	1.000000

Task 4: **Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose 3 different variables to compare against**

1) Are the covid cases higher among the males?
2) Does the covid cases increases if there is increase in population?
3) Are the covid cases higher among the white population?
4) Are covid cases higher for 18years and over age group?