Project Stage - III (Distributions and Hypothesis Testing)

Goals

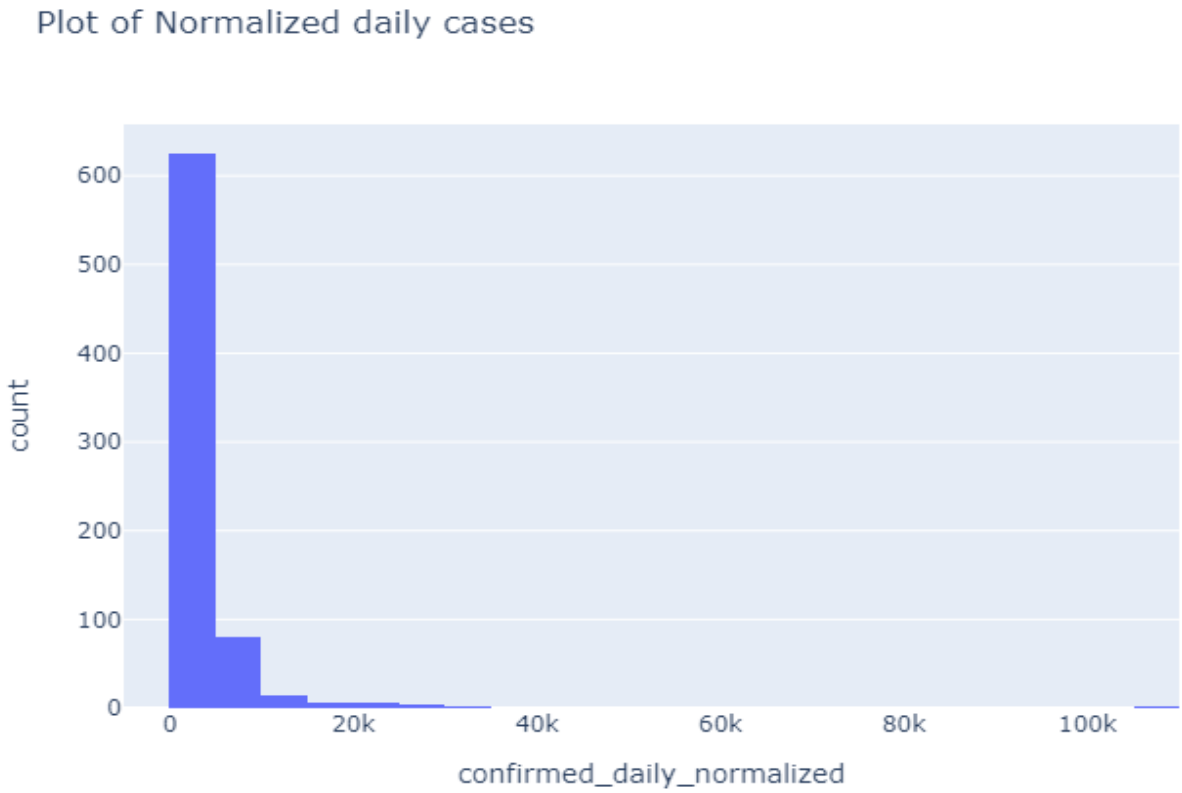
The goal of Stage II is to develop distributions and formal hypothesis tests for the intuitions you had in Stage I and II and utilize statistical modeling to prove/disprove them.

Task1:

- 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.

Solution: For this task, I have used superdataset generated in stage_1 of the project. I have added columns of cases_daily, deaths_daily, cases_normalized_daily, deaths_daily_normalized.

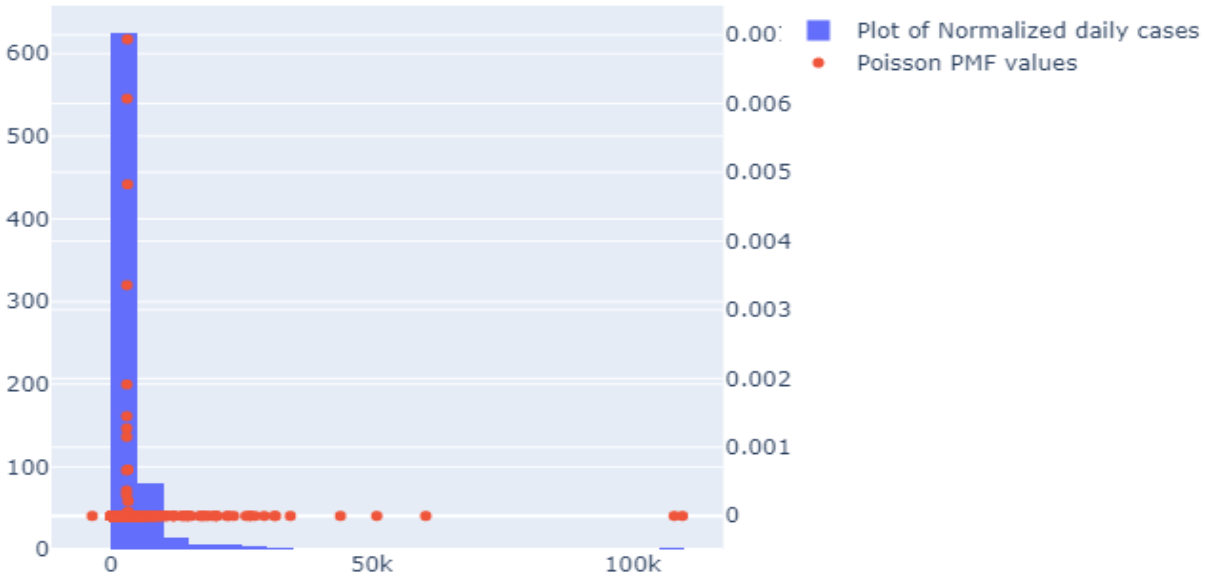
Then plotted the distribution of the data using the histogram.



From the plot, it can be seen that:

- The data is positively skewed.
- It has a discrete distribution.
- It gives information about number of covid cases occurred in a day.
- Poisson distribution provides the probability of seeing certain number of successes in a given time interval. Let us try to plot Poisson distribution on the data. And it is also works for discrete values.

So, plotting the Poisson model for the given data using the pmf values generated for each value of normalized_cases_daily. After plotting we get the output like:



From this plot, it can be seen that the Poisson model fit the model approximately. So, Poisson model will be a good fit for the data.

Task2:

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

Solution: For the measures of center, I have calculated mean, variance, skewness, kurtosis for each state of selection in stage_2. The states are NC, AL, FL, CA, IL, LA. The details about it are as below:

North Carolina:

- **Mean:** 21773
- **Variance:** 982267196.1566926
- **Skewness:** 3.319234715061097
- **Kurtosis:** 12.68782960871529

Alabama:

- **Mean:** 15476
- **Variance:** 475530219.7905452
- **Skewness:** 3.9983361783762095
- **Kurtosis:** 20.948549511750667

California:

- **Mean:** 9209
- **Variance:** 179599711.20938087
- **Skewness:** 3.120498009716973
- **Kurtosis:** 11.161163900685484

Florida:

- **Mean:** 15153
- **Variance:** 415175947.68177265
- **Skewness:** 2.421748461234236
- **Kurtosis:** 6.7486887316609465

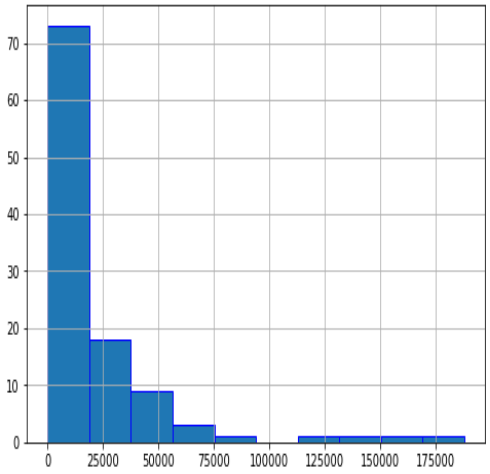
Illinois:

- **Mean:** 26015
- **Variance:** 1380148472.299737
- **Skewness:** 3.0350599046728486
- **Kurtosis:** 11.058040490419375

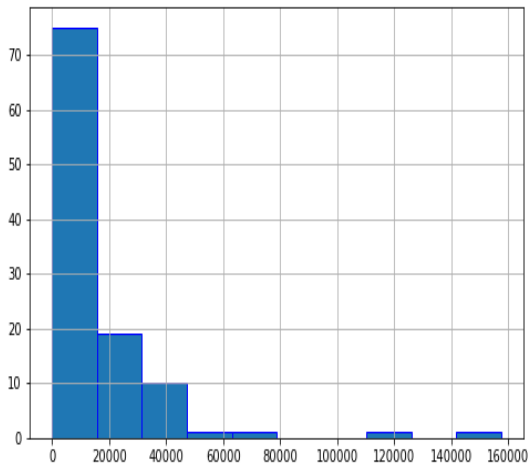
Louisiana:

- **Mean:** 15303
- **Variance:** 505261145.90029436
- **Skewness:** 3.428304090029175
- **Kurtosis:** 14.364807769618462

For distribution, I have selected the same states and plot the distributions and given information about the distribution in the notebook.



NC



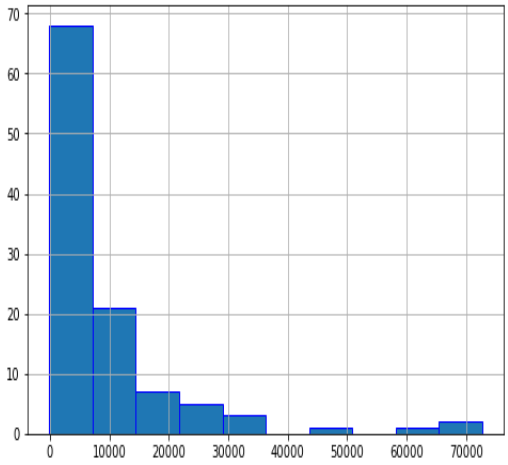
AL

Description:

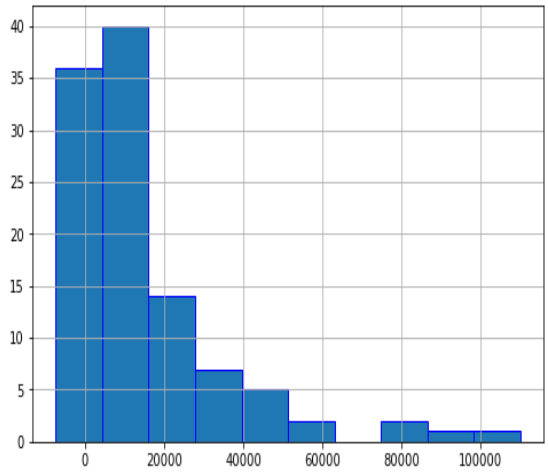
- From the histogram, it can be seen that **North Carolina** have very high number of cases initially which starts decreasing eventually and becomes very low in the end.
- It is a positive skew type distribution because it has skew to the right side.
- It is unimodal distribution
- It is exponential distribution as the random variable is taking values after performing normalization.

Description:

- From the histogram, it can be seen that **Alabama** have very high number of cases initially which starts decreasing eventually, becomes zero and then start increasing again.
- It is a positive skew type distribution because it has skew to the right side.
- It is unimodal distribution
- It is exponential distribution as the random variable is taking values after performing normalization.



CA



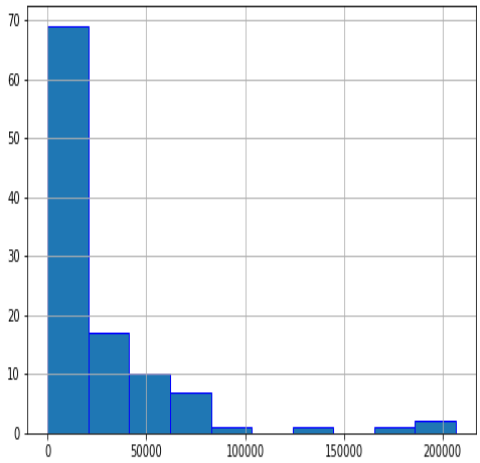
FL

Description:

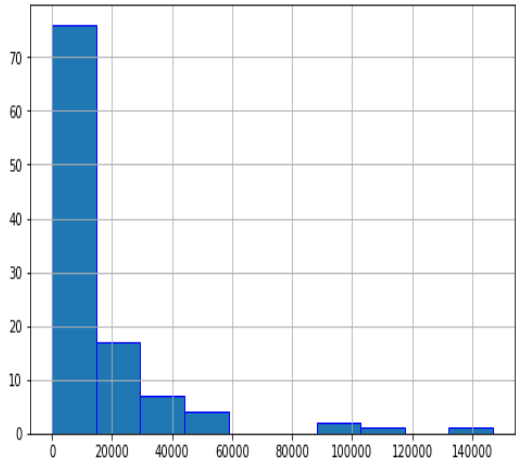
- From the histogram, it can be seen that **California** have very high number of cases initially which starts decreasing eventually, becomes zero and then start increasing again.
- It is a positive skew type distribution because it has skew to the right side.
- It is unimodal distribution
- It is exponential distribution as the random variable is taking values after performing normalization.

Description:

- From the histogram, it can be seen that **Florida** have very high number of cases initially which starts decreasing eventually, becomes zero and then start increasing again.
- It is a positive skew type distribution because it has skew to the right side.
- It is unimodal distribution
- It is exponential distribution as the random variable is taking values after performing normalization.



IL



LA

Description:

- From the histogram, it can be seen that **Illinois** have very high number of cases initially which starts decreasing eventually, becomes zero and then start increasing again.
- It is a positive skew type distribution because it has skew to the right side.
- It is unimodal distribution
- It is exponential distribution as the random variable is taking values after performing normalization.

Description:

- From the histogram, it can be seen that **Louisiana** have very high number of cases initially which starts decreasing eventually, becomes zero and then start increasing again.
- It is a positive skew type distribution because it has skew to the right side.
- It is unimodal distribution
- It is exponential distribution as the random variable is taking values after performing normalization.

Task 3:

- Model a Poisson distribution of COVID-19 cases and deaths of a state and compare to other 5 states. For example, number of new cases and deaths per 100,000 population. Hint - the parameter for a poison's distribution will be its mean value. Then for the minimum and maximum range of covid cases you are calculating probability mass function to observe the probability at different points.

Solution: For this I have plotted the poisson distribution for cases and deaths by stats.poisson.rvs(). The major parameter of this function is the mean values. The mean values calculated in the above task are used for cases and deaths.
The plots are shown in the jupyter notebook.

Task 4:

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

Solution: For this part, I am using the **ACS demographic dataset.** It contains census information about population, population divided by age, sex, age range, race in a countywide fashion.

I have added columns for cases_daily and deaths_daily and also normalize the column values. After that, I have shown correlation among the **non-normalized values** (cases_daily) and other parameter of enrichment dataset. The result of the correlation is shown in the notebook.

Observation:

- From above correlations it can be observed that there exists a positive correlation between non normalized cases with the demographic data.
- Positive correlations implies that there is a linear relationship among the variables. As one value increase, the other value also increases.

For **normalized values**, it is negative. The correlations are mentioned in the notebook.

Observation:

- After normalizing the data for a population of 100000, it is observed there exists a negative correlation between confirmed cases normalized with the total population, Male_total_population, Female_total_population, Population 35 to 44 years old, Population 45 to 54 years old, Population 55 to 59 years old.
- This imply that when normalized cases increases, the other factors decrease. There exists an inverse relationship among them.

Task 5:

- Formulate hypothesis between Enrichment data and number of cases to be compared against states. Choose 3 different variables to compare against. For example: Does higher employment data lead to higher covid case numbers or more rapid increase in covid cases.

Hypothesis:

- States with higher **Total Population** will lead to higher number of covid cases and increase in number of covid cases.
- This implies that states with higher **Male Population** will lead to higher number of covid cases and increase in number of covid cases.
- It also implies that states with higher **Female Population** will lead to higher number of covid cases and increase in number of covid cases.
- States with higher **Middle aged Population** will lead to higher number of covid cases and increase in number of covid cases.

Thank you:

Manish Shah