

Name: Ishita More

UID: 2019130039

TE Comps

Subject: Data Analytics

ISE-Part(II)

Aim:

Probability distributions and hypothesis testing

Dataset:

Daily Age Group Outcomes-

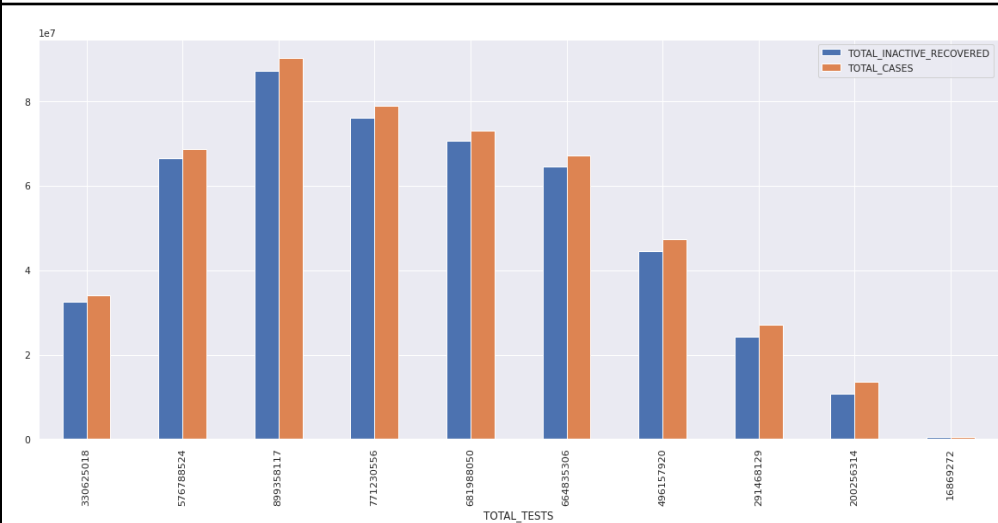
<https://www.tn.gov/content/dam/tn/health/documents/cedep/novel-coronavirus/datasets/Public-Dataset-Daily-Age-Group-Outcomes.XLSX>

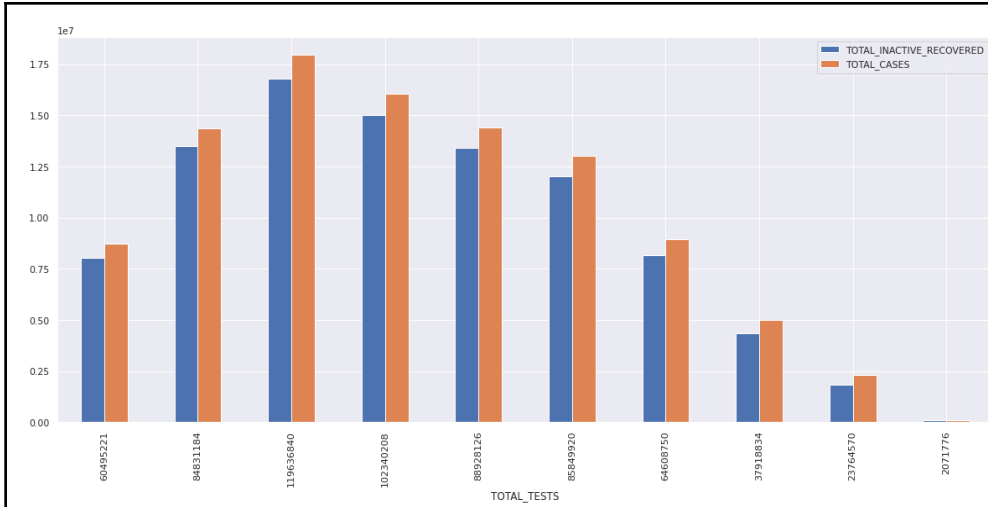
Daily case outcomes by age group (2/25/2022 - 4/9/2020)

In this Dataset, there is data of total tests done, total cases found, totally recovered, total hospitals, and total deaths on a particular day.

It contains the data from 2/26/2022 to 4/9/2020.

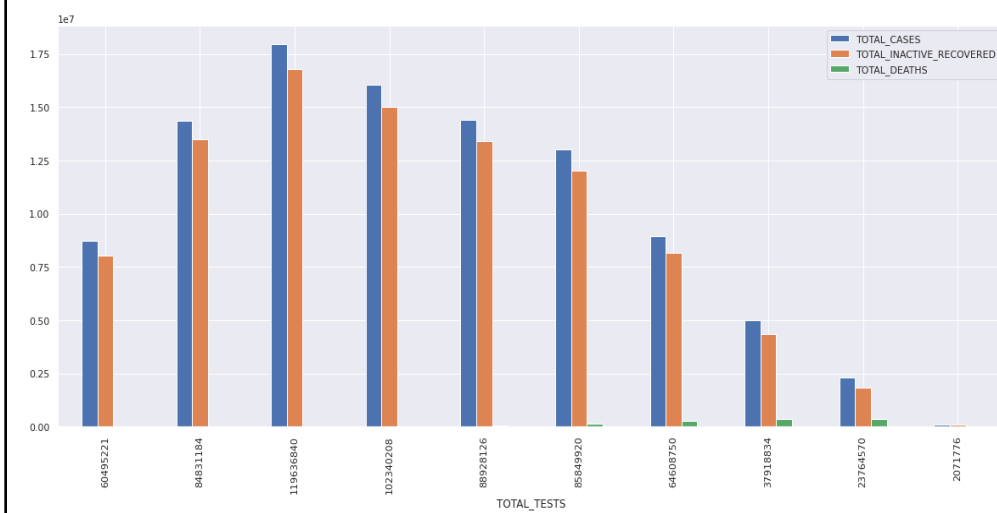
Inference:

Graphs	Inference																																	
 <table><tr><th>TOTAL_TESTS</th><th>TOTAL_INACTIVE_RECOVERED (1e7)</th><th>TOTAL_CASES (1e7)</th></tr><tr><td>330625018</td><td>3.2</td><td>3.4</td></tr><tr><td>576788524</td><td>6.6</td><td>6.8</td></tr><tr><td>899358117</td><td>8.6</td><td>8.8</td></tr><tr><td>771230556</td><td>7.6</td><td>7.8</td></tr><tr><td>681988050</td><td>7.0</td><td>7.2</td></tr><tr><td>664835306</td><td>6.4</td><td>6.6</td></tr><tr><td>496157920</td><td>4.4</td><td>4.6</td></tr><tr><td>291468129</td><td>2.4</td><td>2.6</td></tr><tr><td>200256314</td><td>1.0</td><td>1.2</td></tr><tr><td>16869272</td><td>0.1</td><td>0.2</td></tr></table>	TOTAL_TESTS	TOTAL_INACTIVE_RECOVERED (1e7)	TOTAL_CASES (1e7)	330625018	3.2	3.4	576788524	6.6	6.8	899358117	8.6	8.8	771230556	7.6	7.8	681988050	7.0	7.2	664835306	6.4	6.6	496157920	4.4	4.6	291468129	2.4	2.6	200256314	1.0	1.2	16869272	0.1	0.2	<p>Bar graphs are ideal for comparing two or more values, or values over time.</p> <p>Fig 1: Original Data</p> <p>Fig2: Recent Data</p> <p>Fig 1 and 2 is the comparison between recovered and total cases found. Here we can observe that the difference is very low. This implies cases were found daily, but the recovery rate is also good.</p>
TOTAL_TESTS	TOTAL_INACTIVE_RECOVERED (1e7)	TOTAL_CASES (1e7)																																
330625018	3.2	3.4																																
576788524	6.6	6.8																																
899358117	8.6	8.8																																
771230556	7.6	7.8																																
681988050	7.0	7.2																																
664835306	6.4	6.6																																
496157920	4.4	4.6																																
291468129	2.4	2.6																																
200256314	1.0	1.2																																
16869272	0.1	0.2																																



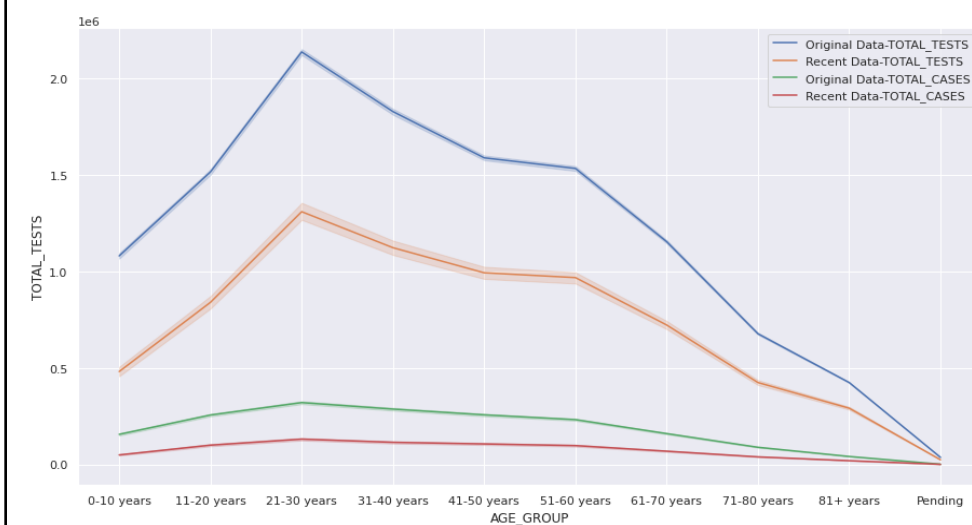
Comparison between Total Cases, Recovery, and Deaths on the basis of a particular age group.

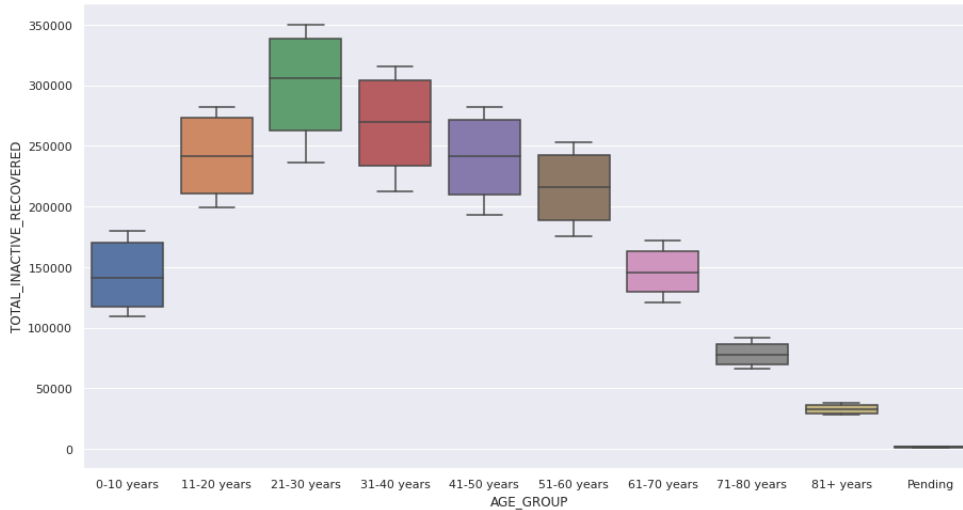
Total Cases found and recovery rate is high for the age group of 20-30 but the death rate is negligible. Whereas it is the opposite in the case of 70+ age.



The number of Total tests done has been reduced than before for all age groups.

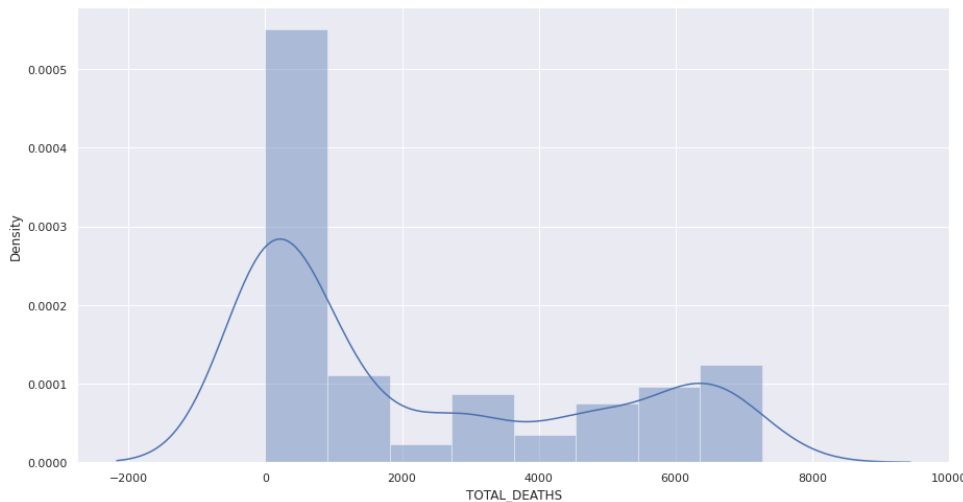
Similarly, the number of Total cases found has been reduced than before for all age groups except the 80+ age group.



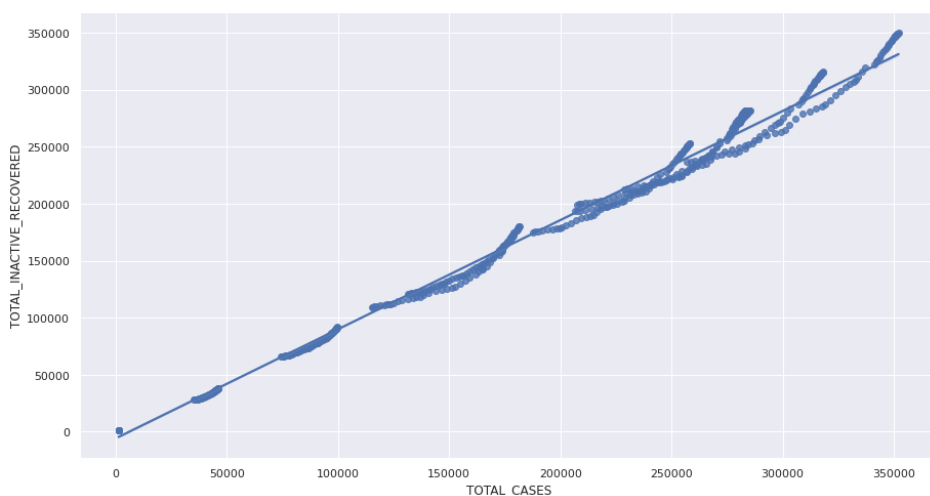


A boxplot is a standardized way of displaying the distribution of data based on a five-number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

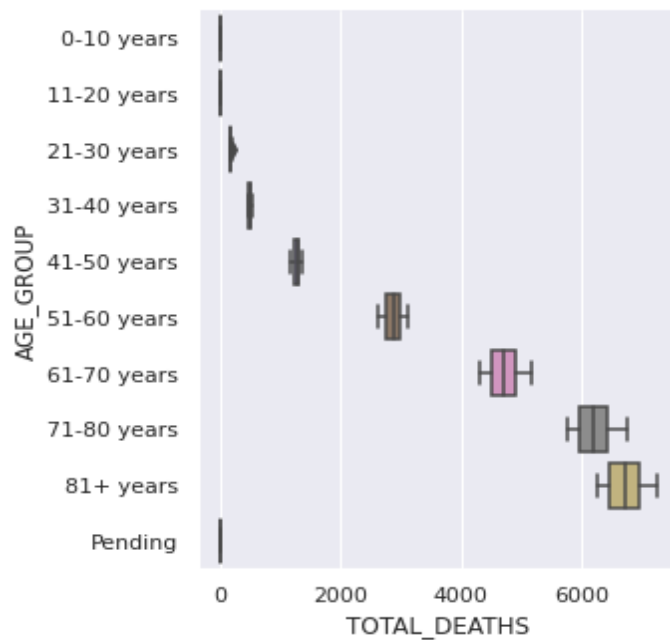
The recovery rate is high in the case of the 20-30 age group, whereas, it is low in the case of the 60+ age group.



As in this Dist plot, the bar is high at 0-1000, which implies that the number of deaths occurring recently is less i.e. between 0 to 1000.

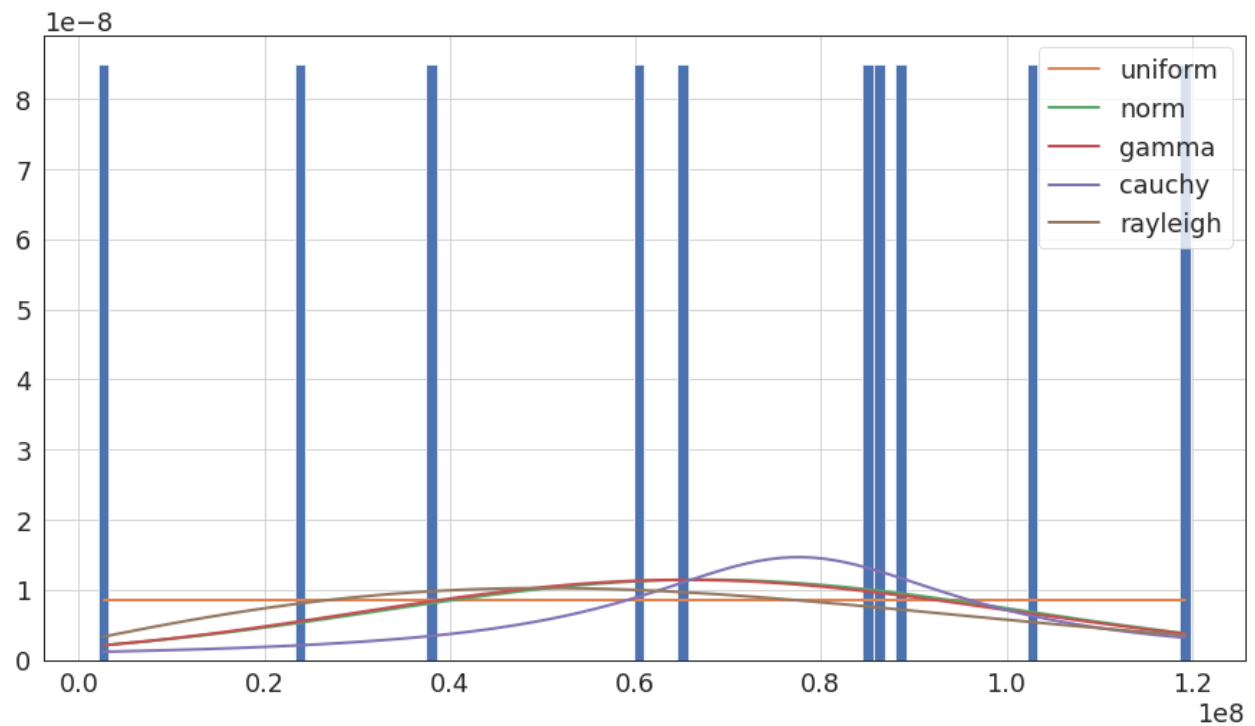


As in the Scatter plot, the total cases found and recovery rate is almost a straight line. So, we can say that the total cases found and recovery rate goes hand in hand.



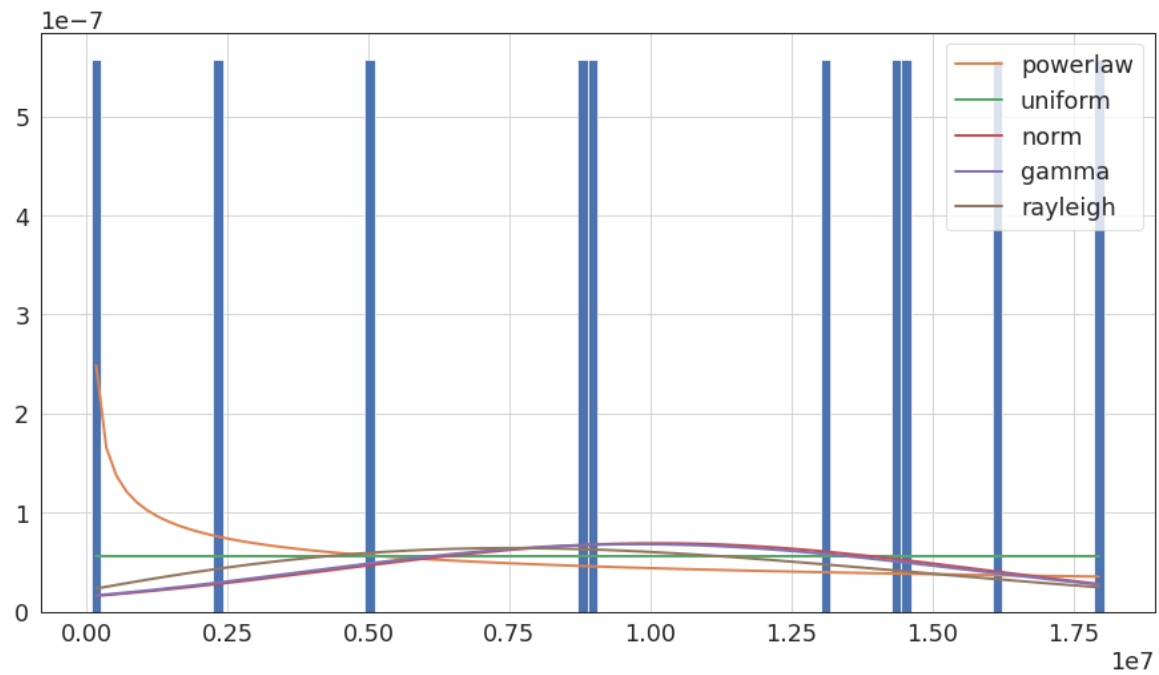
As in this Cat plot, shows frequencies is high for the 71+ age group, which implies that the number of deaths is high in this age group.

2. For the variable TOTAL_TESTS,



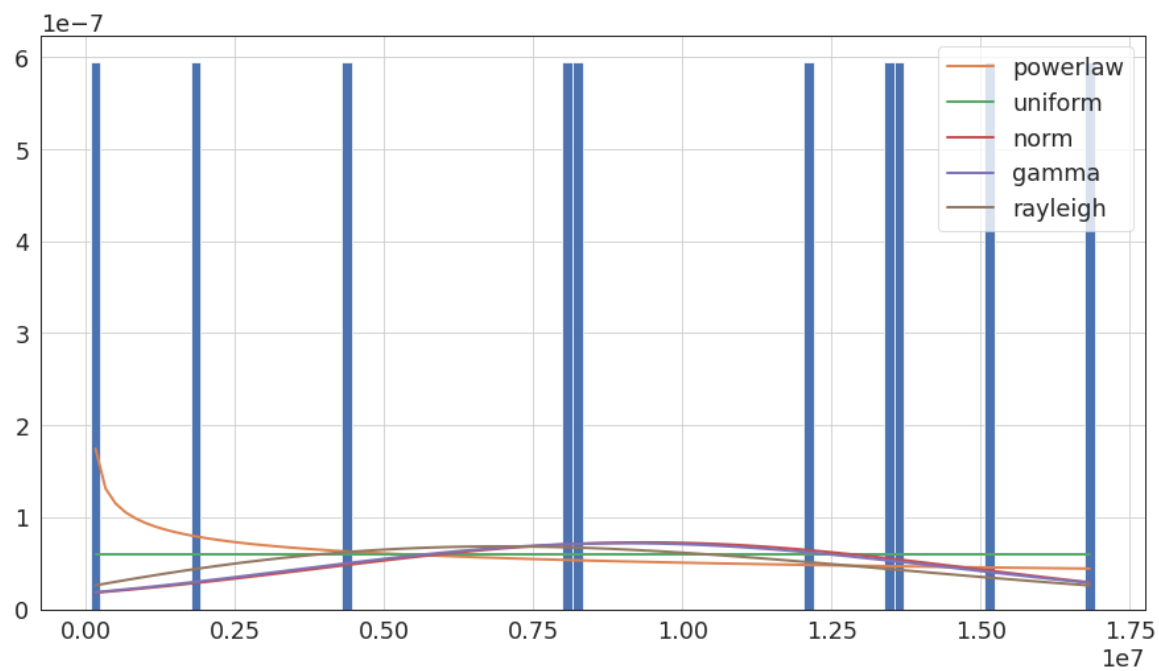
Using the fitter module, we can compute that the best fit for it is Uniform Distribution

For the variable TOTAL_CASES,



Using the fitter module, we can compute that the best fit for it is powerlaw Distribution.

For the variable TOTAL_INACTIVE_RECOVERED,



Using the fitter module, we can compute that the best fit for it is powerlaw Distribution

3.

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

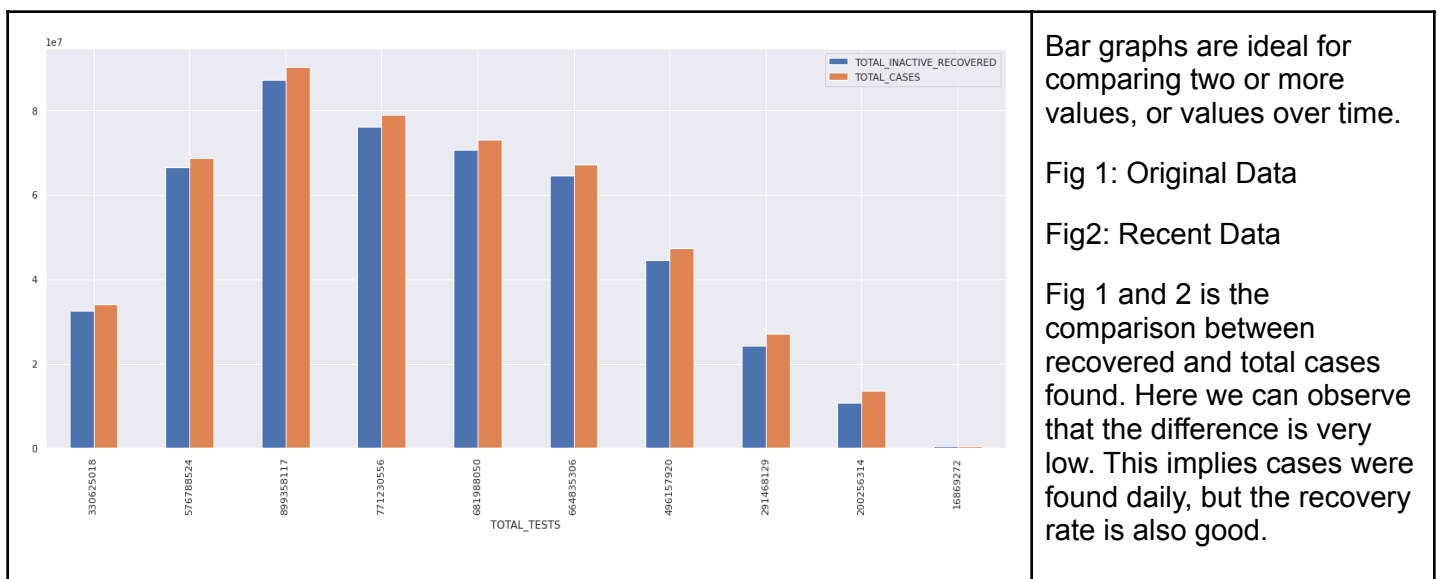
- Since our chi-squared statistic exceeds the critical value, we'd reject the null hypothesis that the two distributions are the same.
- Also, the p -values are less than $\alpha(0.05)$, we'd reject the null hypothesis that the two distributions are the same.

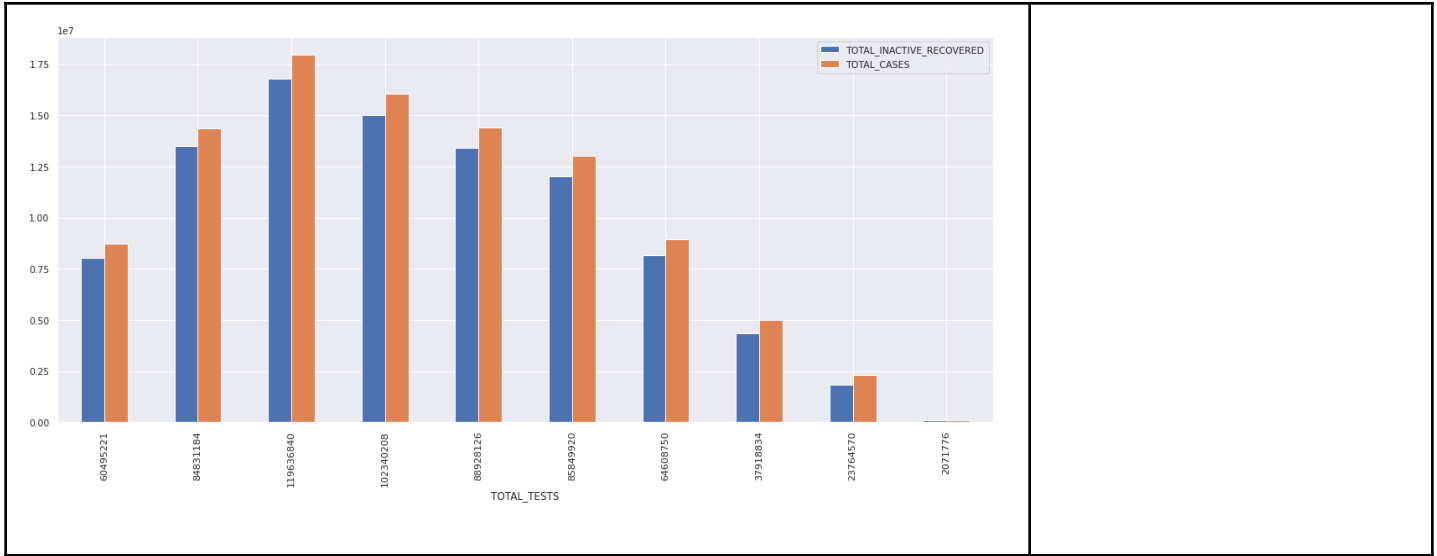
Using `data1.skew()`,

It is concluded that

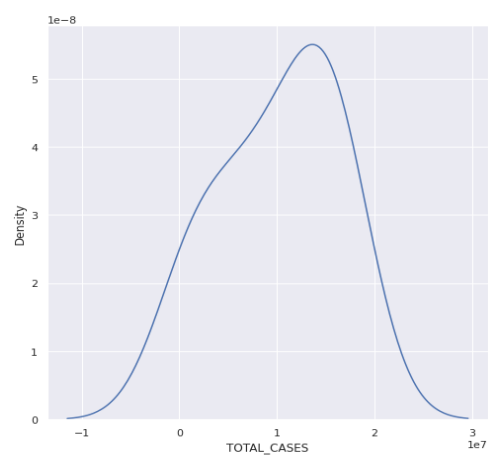
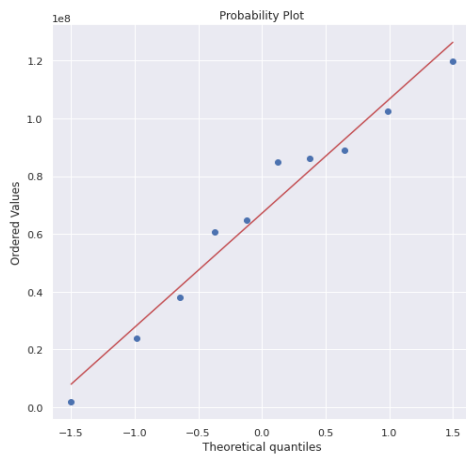
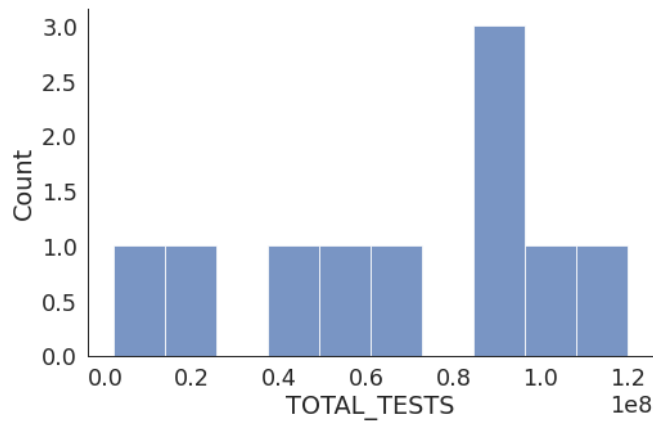
- The variables with skewness > 1 are highly positively skewed.
- The variables with skewness < -1 such as `TOTAL_TESTS`, `TOTAL_CASES`, and `TOTAL_INACTIVE_RECOVERED` are highly negatively skewed.
- The variables with $0.5 < \text{skewness} < 1$ such as `TOTAL_DEATHS` are moderately positively skewed.
- The variables with $-0.5 < \text{skewness} < -1$ are moderately negatively skewed.
- And, the variables with $-0.5 < \text{skewness} < 0.5$ are symmetric i.e normally distributed such as `TOTAL_HOSP`.

4 & 5





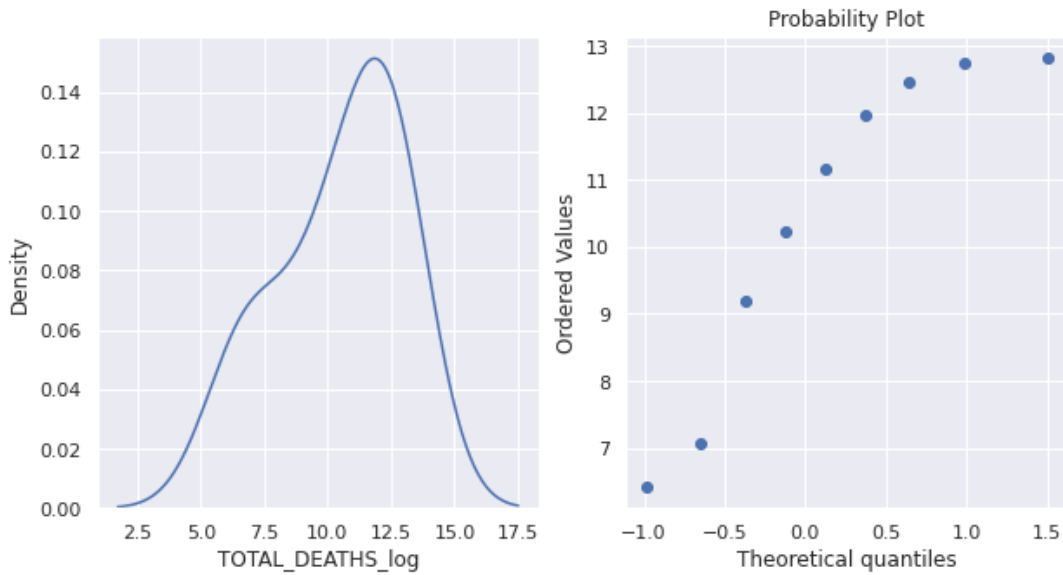
6. As from the above, it is clear that the data use is not a normal distribution. Therefore, to transform them so that they fit a normal distribution. Also, using a histogram, prob plot, and kde plot we can observe that the histogram is not in the shape of a bell.



To transform it for TOTAL_DEATHS,

Using logarithmic transformation,

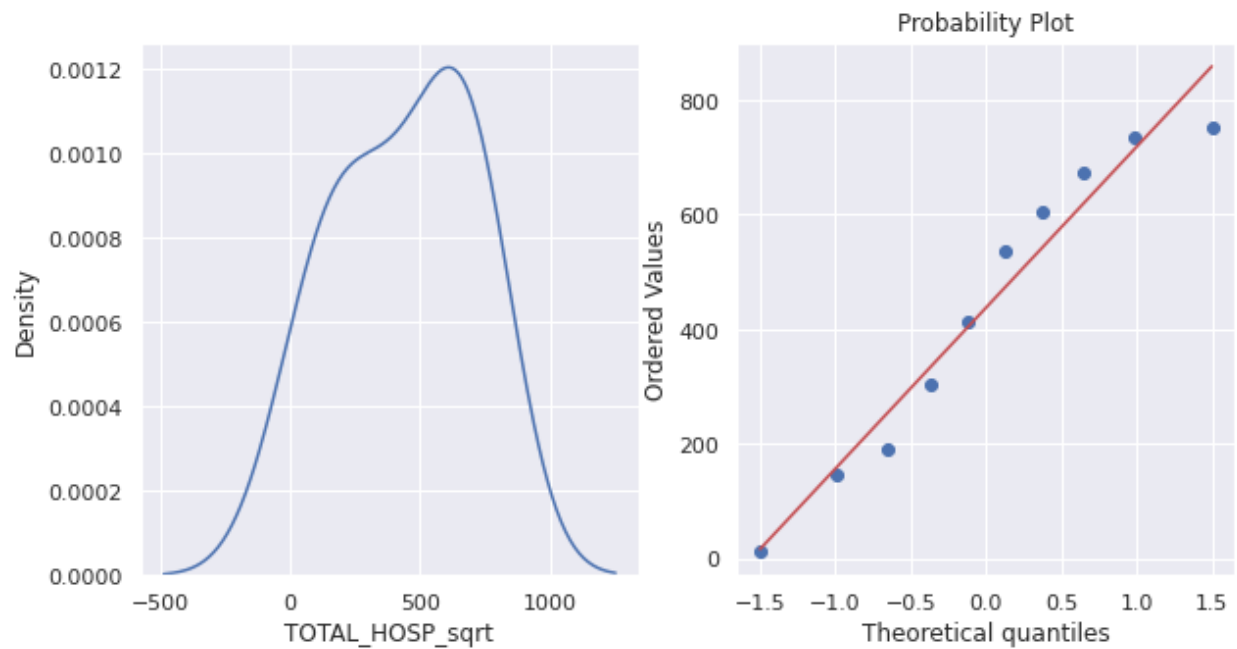
```
data1["TOTAL_DEATHS_log"] = np.log(data1["TOTAL_DEATHS"])
```



To transform it for TOTAL_HOSP,

Using Square transformation,

```
data1["TOTAL_HOSP_sqrt"] = np.sqrt(data1["TOTAL_HOSP"])
```



7.

```
if chi_squared_stat >=critical then, Reject Hypothesis 0 (null Hypothesis)'
elif chi_squared_stat < critical then, Fail to Reject Hypothesis 0 (null Hypothesis)')
if p_value <= alpha then, ("Dependent : Reject Hypothesis 0 (null Hypothesis)")
Else 'Independent : Fail to Reject Hypothesis 0 (null Hypothesis)'
```

Result:

1. TOTAL_TESTS
chi_squared_stat: 46.179802602055005
Reject Hypothesis 0 (null Hypothesis)
Significance 0.05, 5.576495836655756e-07
Reject Hypothesis 0 (null Hypothesis)
2. TOTAL_CASES
chi_squared_stat: 77.21893404040038
Reject Hypothesis 0 (null Hypothesis)
Significance 0.05, 5.754585591598041e-13
Reject Hypothesis 0 (null Hypothesis)
3. TOTAL_INACTIVE_RECOVERED
chi_squared_stat: 92.41296699514167
Reject Hypothesis 0 (null Hypothesis)
Significance 0.05, 5.333480866301491e-16
Reject Hypothesis 0 (null Hypothesis)
4. TOTAL_HOSP_sqrt
chi_squared_stat: 8.60317621165327
Fail to Reject Hypothesis 0 (null Hypothesis)
Significance 0.05, 0.4746803040277736
Fail to Reject Hypothesis 0 (null Hypothesis)

Statistical Power

Statistical significance: Statistical significance refers to the claim that a result from data generated by testing or experimentation is not likely to occur randomly or by chance but is instead likely to be attributable to a specific cause.

Statistical power: It is only relevant when the null hypothesis is false. The statistical power of a hypothesis test is the probability of correctly rejecting a null hypothesis or the likeliness of accepting the alternative hypothesis if it is true. So, the higher the statistical power for a given test, the lower the probability of making a Type II (false negative) error.

Significance (p-value) is the probability that we reject the null hypothesis while it is true. Power is the probability of rejecting the null hypothesis while it is false. Significance is thus the probability of Type I error, whereas 1–power is the probability of Type II error. Mathematically these are not complementary probabilities: p-value is calculated using the probability distribution for the null hypothesis, while the power with the probability distribution for the alternative hypothesis.