

# Nilkamal School of Mathematics, Applied Statistics & Analytics, NMIMS

**MSc. Statistics and Data Science (2023-25)**

TITLE –

FROM BLEND TO BENEFIT-

Fitting of mixture distributions on motor insurance claims

PREPARED BY –

GROUP-10

|  |  |  |  |
| --- | --- | --- | --- |
| NAME | SAP ID | ROLL NO | YEAR |
| Rohan Shah | 86062300001 | A058 | 2023-25 |
| Astha Sharma | 86062300064 | A059 | 2023-25 |
| Verni Sharma | 86062300020 | A060 | 2023-25 |
| Anshika Sharma | 86062300034 | A061 | 2023-25 |
| Anniruddha Shelke | 86062300008 | A062 | 2023-25 |

UNDER THE SUPERVISION OF

Dr. Pradnya Khandeparkar

TABLE OF CONTENTS –

1. ACKNOWLEDGMENT
2. ABSTRACT
3. INTRODUCTION
4. RATIONALE
5. AIM AND OBJECTIVE
6. DATA PREPARATION
7. METHODOLOGY
8. RESULTS AND DISCUSSION
9. SUMMARY AND CONCLUSION
10. LIMITATIONS
11. REFERENCES
12. APPENDIX

ACKNOWLEDGEMENT

The realization of any achievement necessitates the concerted labor and dedication of numerous individuals. The endeavor at hand, our project work, is no exception. Our profound appreciation extends to Dr. Pradnya Khandeparkar, whose inspiration and guidance have been instrumental in steering our efforts towards the successful execution of this research project.

We express our sincere gratitude to NMIMS for affording us the invaluable opportunity to undertake and conduct this research project. The support and resources provided by NMIMS have played a pivotal role in facilitating the progression and completion of our scholarly endeavors.

ABSTRACT

Problem statement: The modeling of claims is an important task of actuaries. Our problem is in modeling the actual motor insurance claim data set. In this study, we show that the actual motor insurance claim can be fitted by a finite mixture model**.** Approach: Firstly, we analyze the actual data set and then we choose the finite mixture of Lognormal distributions as our model. The estimated parameters of the model are obtained from the EM algorithm. Then, we use the K-S test to show how well the finite mixture Lognormal distributions fit the actual data set. Results: From the tests, we found that the finite mixture lognormal distributions fit the actual data set with a significant level of 0.10. Conclusion**:** The finite mixture Lognormal distributions can be fitted to motor insurance claims and this fitting is better when the number of components (k) is increasing.

INTRODUCTION

Finite mixtures of distributions have provided a mathematical approach to the statistical modeling of a wide variety of random phenomena. It is an extremely flexible method of modeling and has continued to receive increasing attention over the years from both practical and theoretical point of view. Areas in which mixture models have been successfully applied include astronomy, biology, genetics, medicine, psychiatry and economics. Very little literature is on the applications in general insurance setting. According to the motor insurance is an important branch of non-life insurance in many countries, with contributions amongst the total premium income category. It is a fact that, most insurance claims exhibit some level of clustering, and the usefulness of mixture distribution in modeling heterogeneity in a cluster analysis context is obvious. In practice, most motor insurance claims which occur with losses are modeled by unimodal loss models and . Motor insurance claims with multimodal loss distributions are more advance to apply common unimodal loss models. We therefore extend our knowledge on mixture distributions using finite mixtures of regression models to model such case. Finite mixtures of regression models are a popular method to model unobserved heterogeneity or to account for over dispersion in the claims data. They are flexible models and in theory it is easy to modify and extend them by using more complex models for the component distribution functions and estimate the corresponding parameters. Finite mixture models with a fixed number of components are usually estimated with the expectation-maximization (EM) algorithm within a maximum likelihood framework. Since there are many different modes for claim possibilities, a finite mixture model should work well, and compared (numerically) two approaches to the estimation of the parameters of the component densities in a univariate mixture of normal distributions; one approach is based on a constrained maximum likelihood (ML) algorithm; the other, is on the fuzzy c-means (FCM) clustering algorithm, [8]. Finite mixture models so far include components of the data structure. The purpose of this study is to determine an appropriate finite mixture model for the claims data. The results which can help us determine the expected reserves.

Rationale

Why did we go for Motor Insurance?

The insurance industry plays a crucial role in the economy by providing financial protection against risks. Research in this field can contribute to a better understanding of risk management, financial stability, and economic resilience. This focuses on improving the customer experience which includes studying consumer behavior, developing customer-centric products, and enhancing the claims processing experience.

The insurance industry generates vast amounts of data. Research in data science and analytics can lead to the development of predictive models, fraud detection techniques, and other data-driven innovations that benefit the insurance sector. It provides valuable insights and contributes to the development of expertise in a dynamic and growing industry. The motor insurance industry plays a crucial role in the overall functioning of the economy and society. Reasons being: Legal Requirements, Financial Protection, Protection Against Unforeseen Events Asset Protection, Medical Coverage, etc.

What is a Mixture Distribution?

**What is a Mixture Distribution?**

A mixture distribution is a statistical distribution that is composed of a mixture of two or more component distributions. Each component distribution is associated with a certain probability, and the overall probability distribution is a weighted sum (or mixture) of these components.

Picture 2

* πi ’s are the weights
* fi(x)’s are the distributions of the individual components
* it goes from 1 to k.

**Why Mixture Distribution on Claims?**

1. Claims data are often skewed or non-normal distributions. Mixture models, especially GMMs, are capable of capturing a wide range of distribution shapes, making them suitable for handling diverse claims data

Handling Skewed or Non-Normal Distributions

2. Heterogeneity in Claims Data:

Claims data often exhibits heterogeneity, which means that it may come from different underlying distributions. Mixture models can capture this heterogeneity by representing the data as a combination of multiple distributions.

3. Modeling Complex  Distributions:

Claims data may not be easily characterized by a single probability distribution. Mixture models provide a flexible framework for approximating complex data distributions by combining simpler component distributions

4. Fraud Detection

Mixture models can be used for fraud detection in insurance claims. By modeling the normal and abnormal behavior separately, anomalies can be detected more effectively.

5. Predictive Modeling:

Mixture models are useful for predictive modeling in insurance, helping companies anticipate future claim patterns and identify potential risks. This is particularly valuable for strategic planning and resource allocation.

AIM & OBJECTIVES

To provide financial coverage to policyholders in the event of accidental damage, theft, or loss of their vehicles and also to mitigate the financial risks associated with owning and operating a motor vehicle.

Objective:

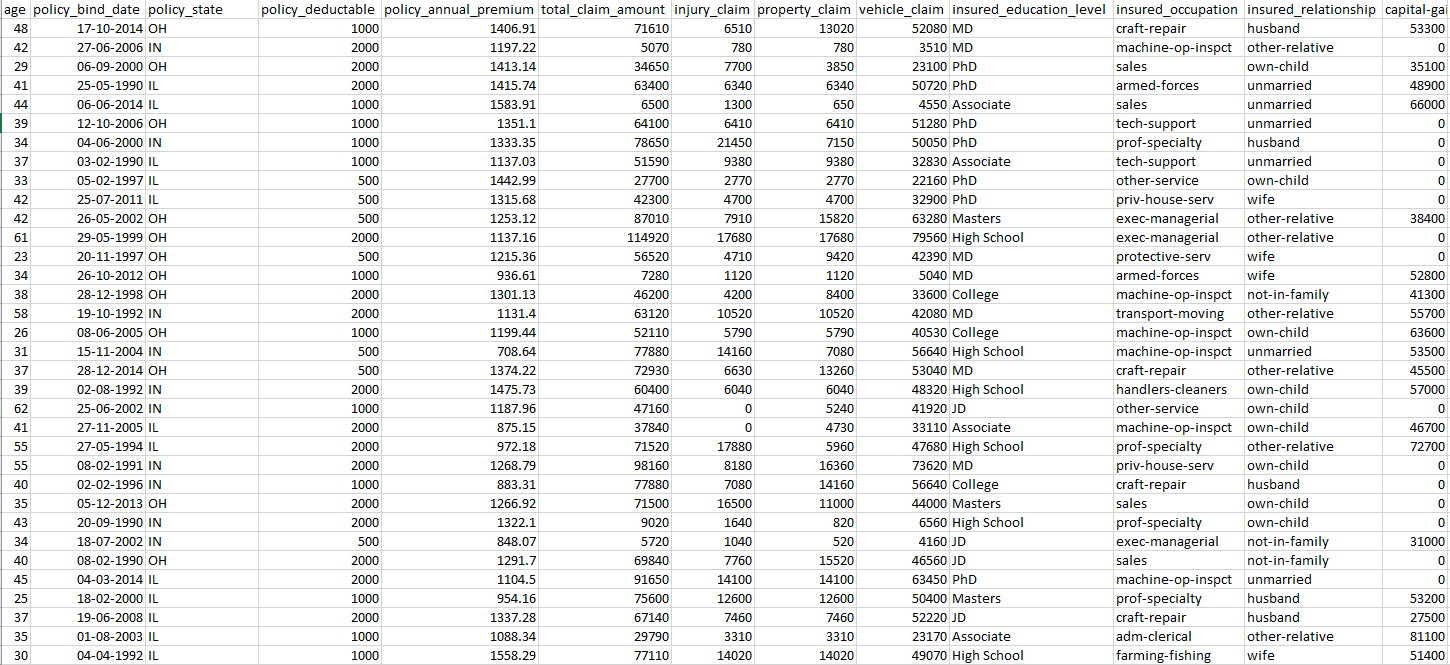
To Study Motor Insurance Claims and investigate its underlying distribution

Data Preparation

Data preparation is a crucial step in the data analysis process, and documenting it thoroughly is essential for the transparency and reproducibility of your work. When including a section on data preparation in your report, consider the following components:

DATA COLLECTION :

Data was obtained from Kaggle



IDENTIFYING THE VARIABLE :

Total claim Amount&Vehicle Claim Amount (here in dollars)

PREPROCESSING THE DATA:

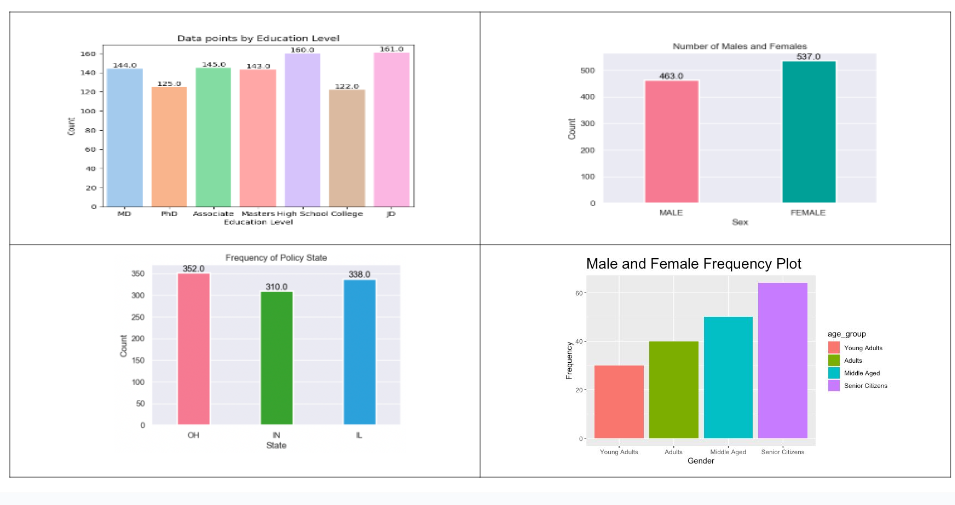
1. No missing values were found in our data
2. Some columns not used in our study were removed (“hobbies of insured” etc.)
3. Investigated bias: Data was not found to be biased (in terms of sex, education level or state variables, etc.)

**TOOLS AND SOFTWARE USED**

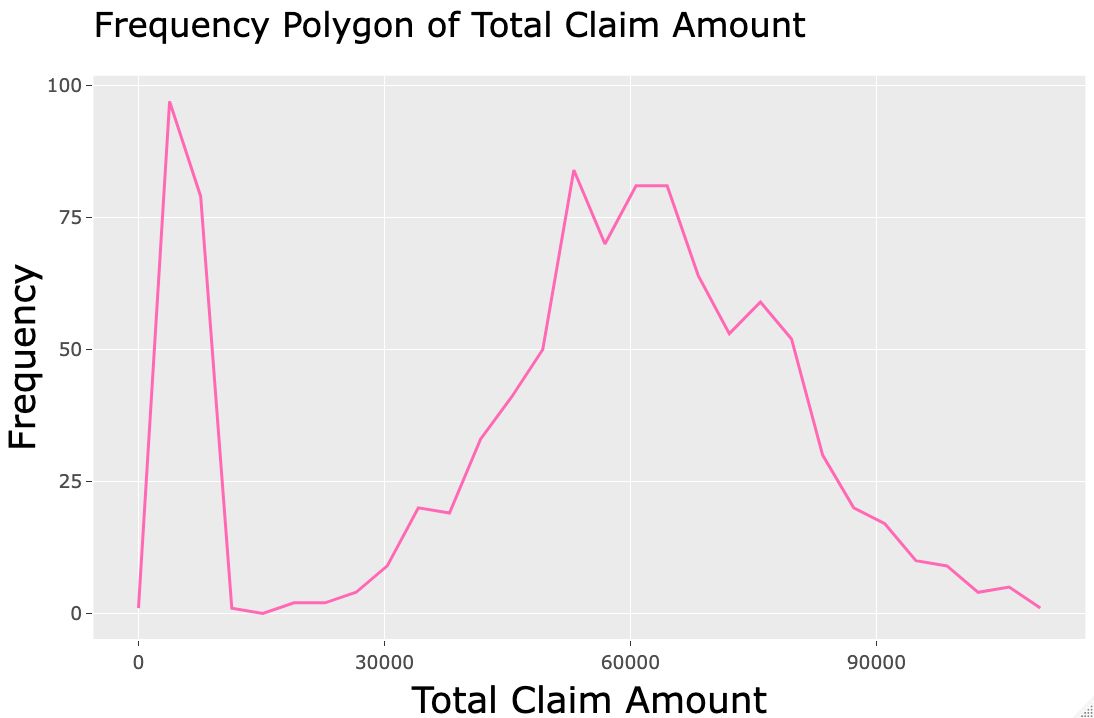
R and Python programming language

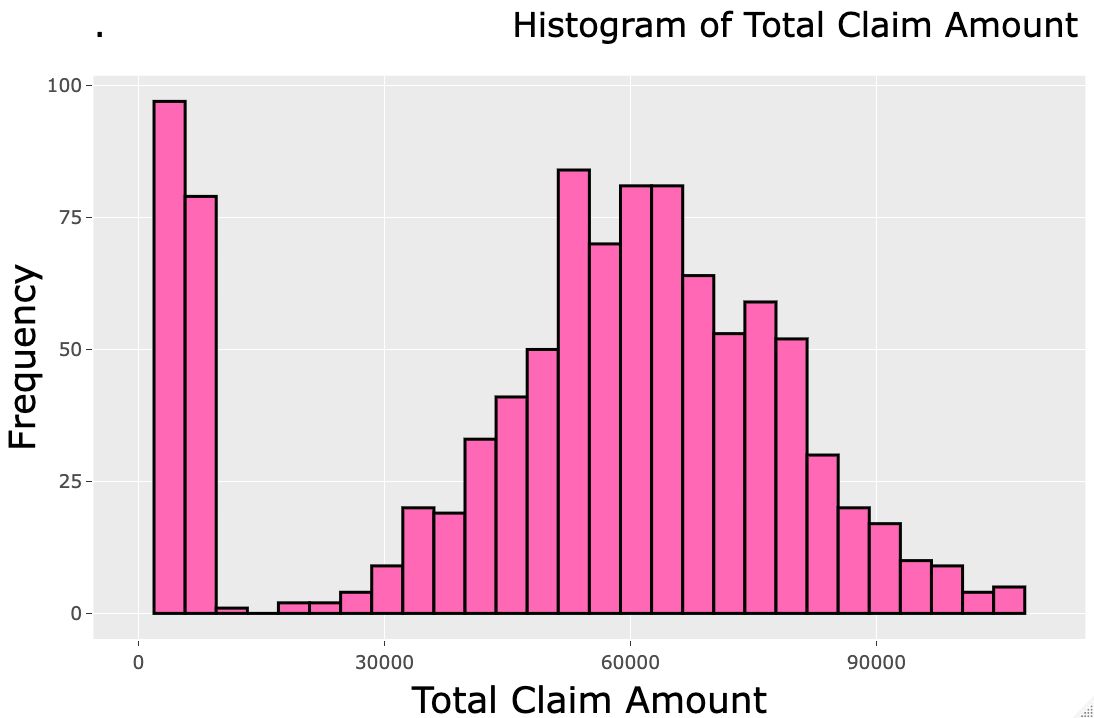
EXPLORATORY DATA ANALYSIS OF THE DATA

EDA on the given data and the findings were as follows ( we can interpret easily from the following graphs)



FREQUENCY POLYGON





METHODOLOGY

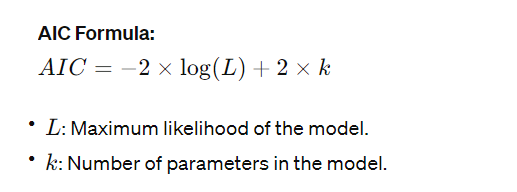
The principles, processes, and rules that guide our approach for designing and conducting our study are -

1)There are many algorithms available however literature suggests that the EM algorithm has been the most widely used for mixture distributions for its accuracy

2) Goodness of Fit: The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.

3)We took out the AIC Scores and the one with the least AIC values was preferred.

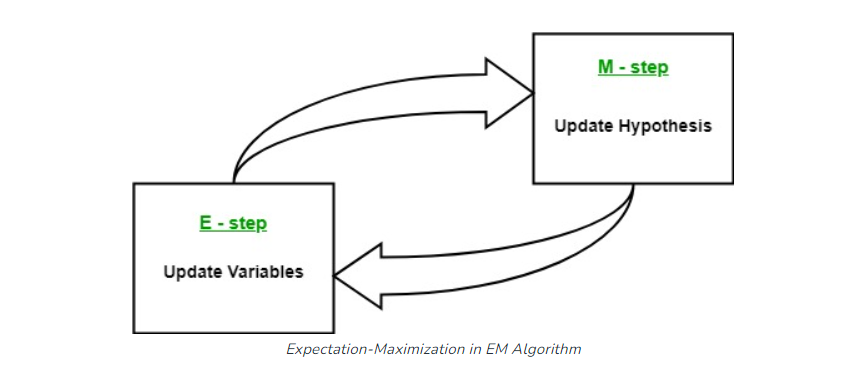
Akaike Information Criterion ( AIC) is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given data set. It estimates models relatively.



About EM algorithm-

The Expectation-Maximization (EM) algorithm is an iterative optimization method that combines different [unsupervised](https://www.geeksforgeeks.org/unsupervised-machine-learning-the-future-of-cybersecurity/) [machine learning](https://www.geeksforgeeks.org/machine-learning/) algorithms to find maximum likelihood or maximum posterior estimates of parameters in statistical models that involve unobserved latent variables. The EM algorithm is commonly used for latent variable models and can handle missing data. It consists of an estimation step (E-step) and a maximization step (M-step), forming an iterative process to improve model fit.

* In the E step, the algorithm computes the latent variables i.e., the expectation of the log-likelihood using the current parameter estimates.
* In the M step, the algorithm determines the parameters that maximize the expected log-likelihood obtained in the E step, and corresponding model parameters are updated based on the estimated latent variables.



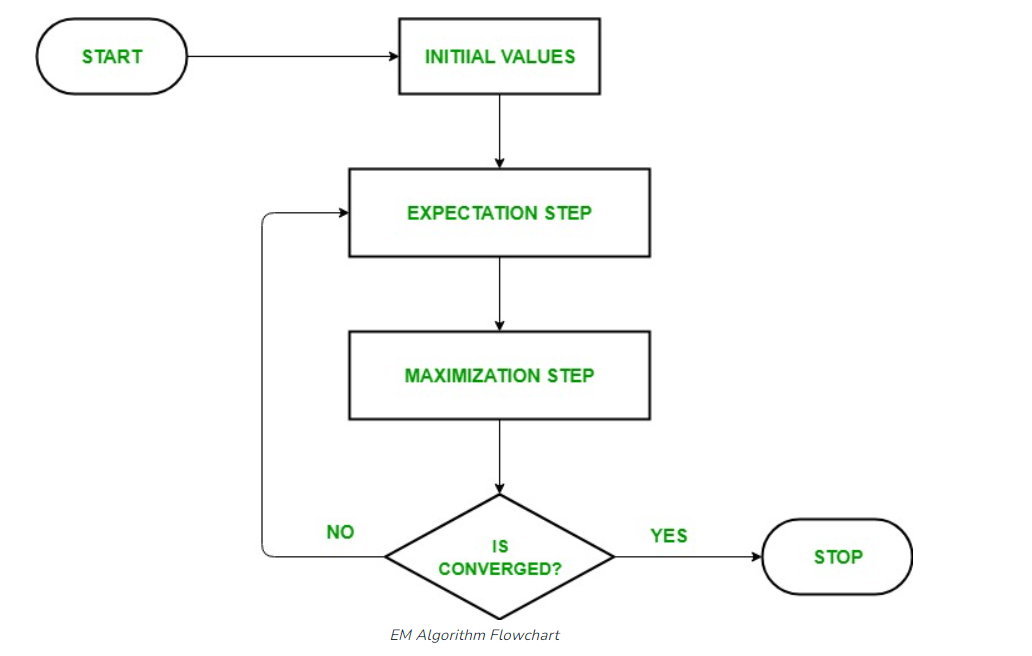
By iteratively repeating these steps, the EM algorithm seeks to maximize the likelihood of the observed data. It is commonly used for unsupervised learning tasks, such as clustering, where latent variables are inferred and has applications in various fields, including machine learning, computer vision, and natural language processing.

**Key Terms in Expectation-Maximization (EM) Algorithm**

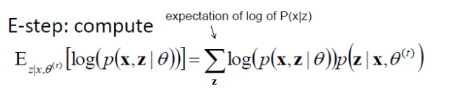
Some of the most commonly used key terms in the Expectation-Maximization (EM) Algorithm are as follows:

* **Latent Variables:** Latent variables are unobserved variables in statistical models that can only be inferred indirectly through their effects on observable variables. They cannot be directly measured but can be detected by their impact on the observable variables.
* **Likelihood:** It is the probability of observing the given data given the parameters of the model. In the EM algorithm, the goal is to find the parameters that maximize the likelihood.
* **Log-Likelihood:** It is the logarithm of the likelihood function, which measures the goodness of fit between the observed data and the model. EM algorithm seeks to maximize the log-likelihood.
* **Maximum Likelihood Estimation (MLE)**: MLE is a method to estimate the parameters of a statistical model by finding the parameter values that maximize the likelihood function, which measures how well the model explains the observed data.
* **Posterior Probability**: In the context of Bayesian inference, the EM algorithm can be extended to estimate the maximum a posteriori (MAP) estimates, where the posterior probability of the parameters is calculated based on the prior distribution and the likelihood function.
* **Expectation (E) Step**: The E-step of the EM algorithm computes the expected value or posterior probability of the latent variables given the observed data and current parameter estimates. It involves calculating the probabilities of each latent variable for each data point.
* **Maximization (M) Step**: The M-step of the EM algorithm updates the parameter estimates by maximizing the expected log-likelihood obtained from the E-step. It involves finding the parameter values that optimize the likelihood function, typically through numerical optimization methods.
* **Convergence:**Convergence refers to the condition when the EM algorithm has reached a stable solution. It is typically determined by checking if the change in the log-likelihood or the parameter estimates falls below a predefined threshold.
* **How the Expectation-Maximization (EM)  Algorithm Works:**

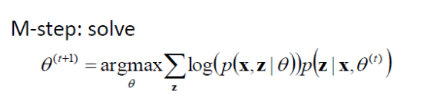
The essence of the Expectation-Maximization algorithm is to use the available observed data of the dataset to estimate the missing data and then use that data to update the values of the parameters. Let us understand the EM algorithm in detail.



1. **Initialization:**
   * Initially, a set of initial values of the parameters are considered. A set of incomplete observed data is given to the system with the assumption that the observed data comes from a specific model.
2. **E-Step (Expectation Step):** In this step, we use the observed data in order to estimate or guess the values of the missing or incomplete data. It is basically used to update the variables.
   * Compute the posterior probability or responsibility of each latent variable given the observed data and current parameter estimates.
   * Estimate the missing or incomplete data values using the current parameter estimates.
   * Compute the log-likelihood of the observed data based on the current parameter estimates and estimated missing data.



1. **M-step (Maximization Step):** In this step, we use the complete data generated in the preceding “Expectation” step in order to update the values of the parameters. It is basically used to update the hypothesis.
   * Update the parameters of the model by maximizing the expected complete data log-likelihood obtained from the E-step.
   * This typically involves solving optimization problems to find the parameter values that maximize the log-likelihood.
   * The specific optimization technique used depends on the nature of the problem and the model being used.



1. **Convergence**: In this step, it is checked whether the values are converging or not, if yes, then stop otherwise repeat *step-2* and *step-3* i.e., “Expectation” – step and “Maximization” – step until the convergence occurs.
   * Check for convergence by comparing the change in log-likelihood or the parameter values between iterations.
   * If the change is below a predefined threshold, stop and consider the algorithm converged.
   * Otherwise, go back to the E-step and repeat the process until convergence is achieved.

## Kolmogorov-Smirnov -

Kolmogorov–Smirnov Test is a completely efficient manner to determine if two samples are significantly one of a kind from each other. It is normally used to check the uniformity of random numbers. Uniformity is one of the maximum important properties of any random number generator and the Kolmogorov–Smirnov check can be used to check it. The Kolmogorov–Smirnov take a look at can also be used to check whether or not two underlying one-dimensional opportunity distributions differ. It is a totally green manner to determine if two samples are substantially distinct from each other. The Kolmogorov–Smirnov statistic quantifies the gap between the empirical distribution function of the pattern and the cumulative distribution feature of the reference distribution, or among the empirical distribution functions of samples.

## How Kolmogorov-Smirnov test works?

To answer this first we need to discuss the purpose of using this test. The main idea behind using this test is to check whether the two samples that we are dealing with follow the same type of distribution or if the shape of the distribution is the same or not.

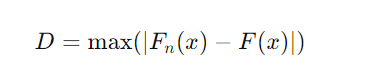
First of all, if we assume that the shape or the [probability distribution](https://www.geeksforgeeks.org/probability-distribution-function/) of the two samples is the same then the maximum value of the absolute difference between the cumulative probability distribution difference between the two functions will be the same. And higher the value the difference between the shape of the distribution is high.

The hypothesis taken –

H0: The sample is drawn from the reference distribution.

H1: The sample is not drawn from the reference distribution.

The formula for Test statistic for K-S test

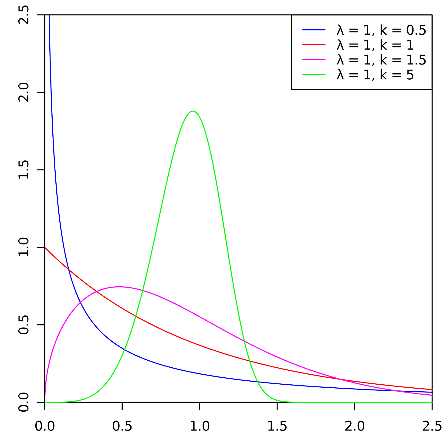


Distributions Investigated-

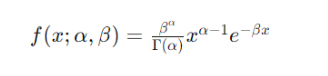
* 1. Weibull distribution



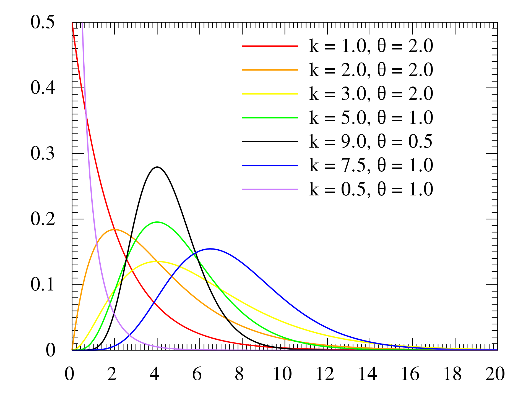
* γ is the **shape parameter**, also called as the Weibull slope or the threshold parameter.
* α is the **scale parameter**, also called the characteristic life parameter.
* μ is the **location parameter**, also called the waiting time parameter or sometimes the shift parameter.



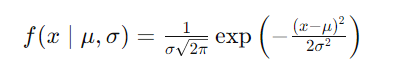
* 1. Gamma Distribution



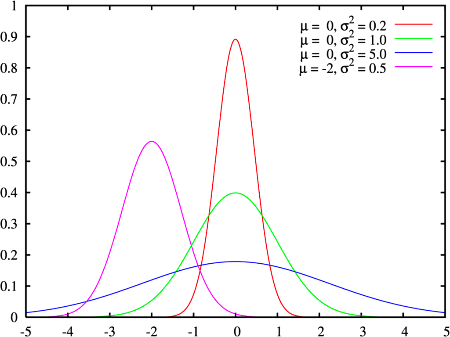
* *x* is the random variable.
* *α* is the shape parameter (also known as the "k shape parameter").
* *β* is the rate parameter (sometimes called the "theta scale parameter").
* Γ(*α*) is the gamma function evaluated at *α*.



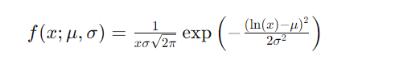
* 1. Normal Distribution



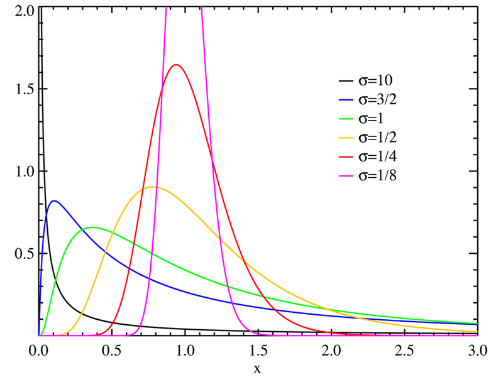
* *x* is the random variable.
* *μ* is the mean of the distribution.
* *σ* is the standard deviation of the distribution.
* *π* is the mathematical constant pi (approximately 3.14159).
* exp(⋅) represents the exponential function



* 1. Log normal distribution



* *x* is the random variable.
* *μ* is the mean of the natural logarithm of the distribution.
* *σ* is the standard deviation of the natural logarithm of the distribution.
* ln(*x*) denotes the natural logarithm of *x*.



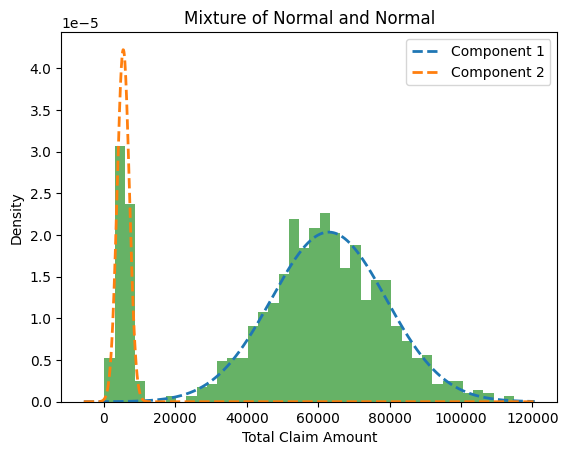
Results and Discussion

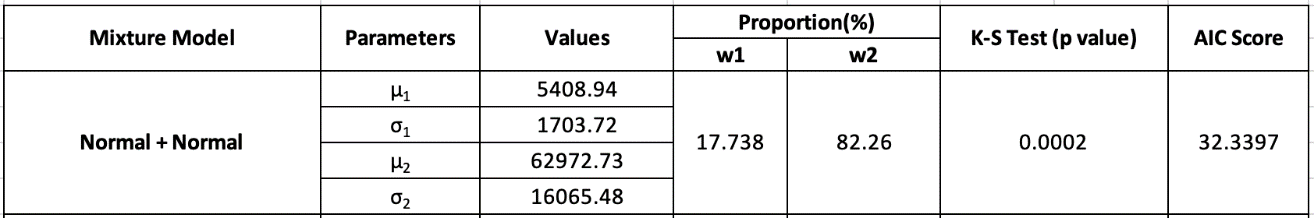
Theory suggests that the claim amount of data will tend to follow distributions that are positively skewed since the claim amount cannot take negative values

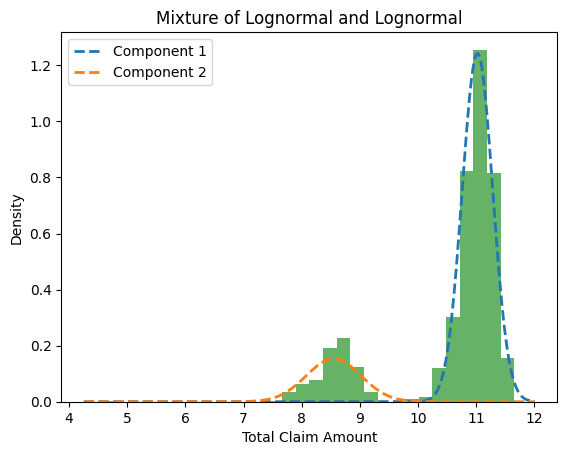
Hence, we have considered the distributions:

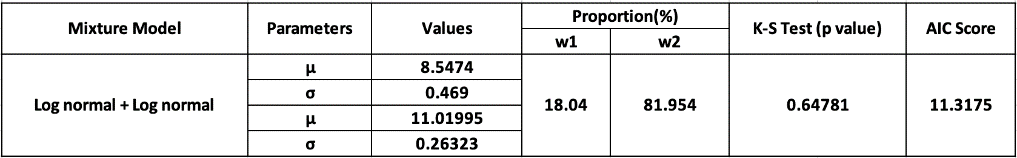
* Weibull
* Gamma
* Lognormal
* Normal (based on initial investigation)

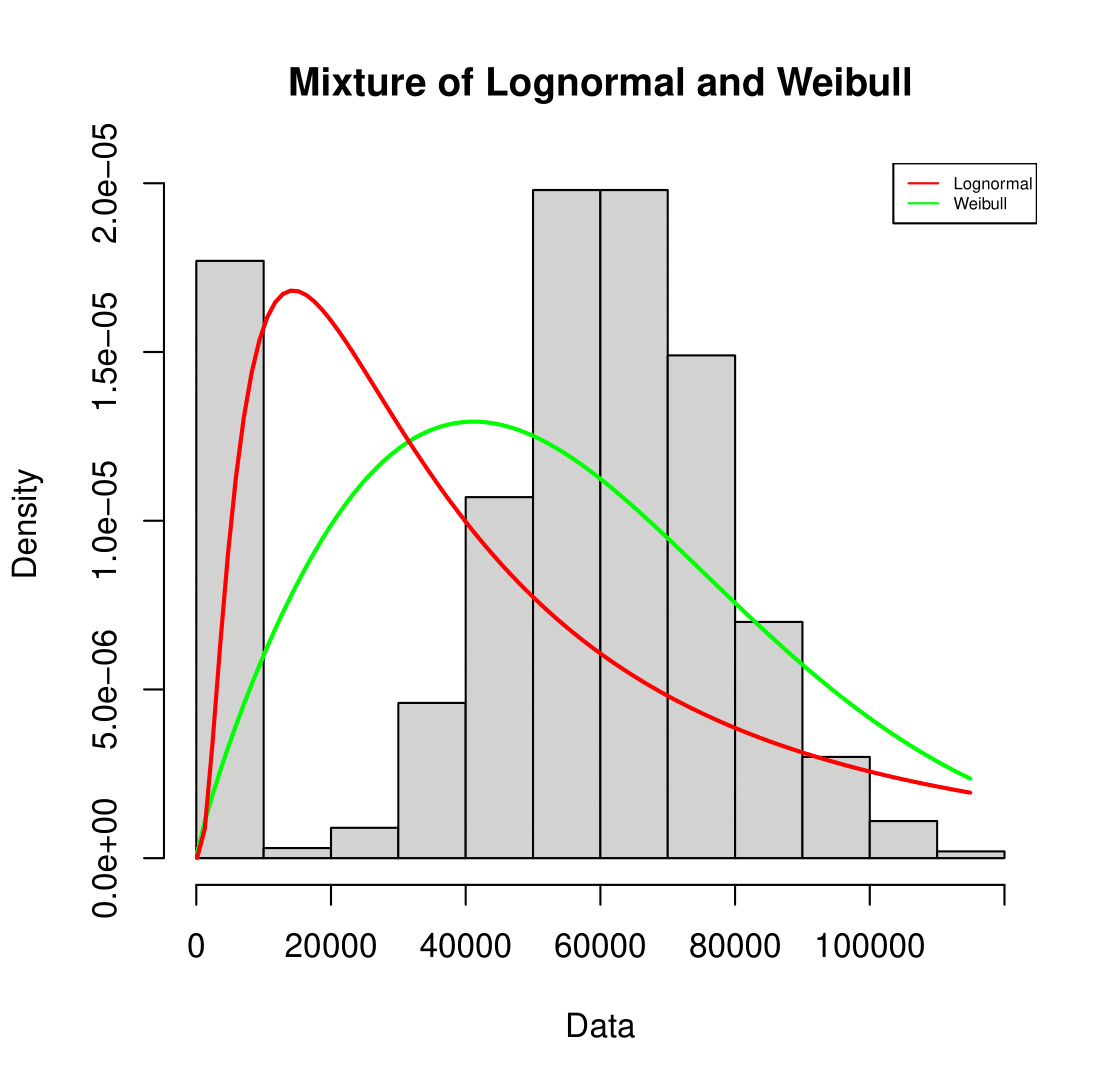
Fitted Distributions

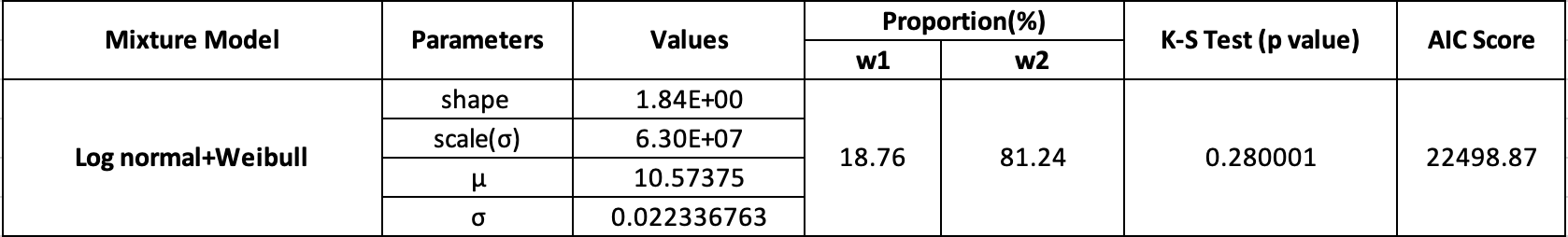


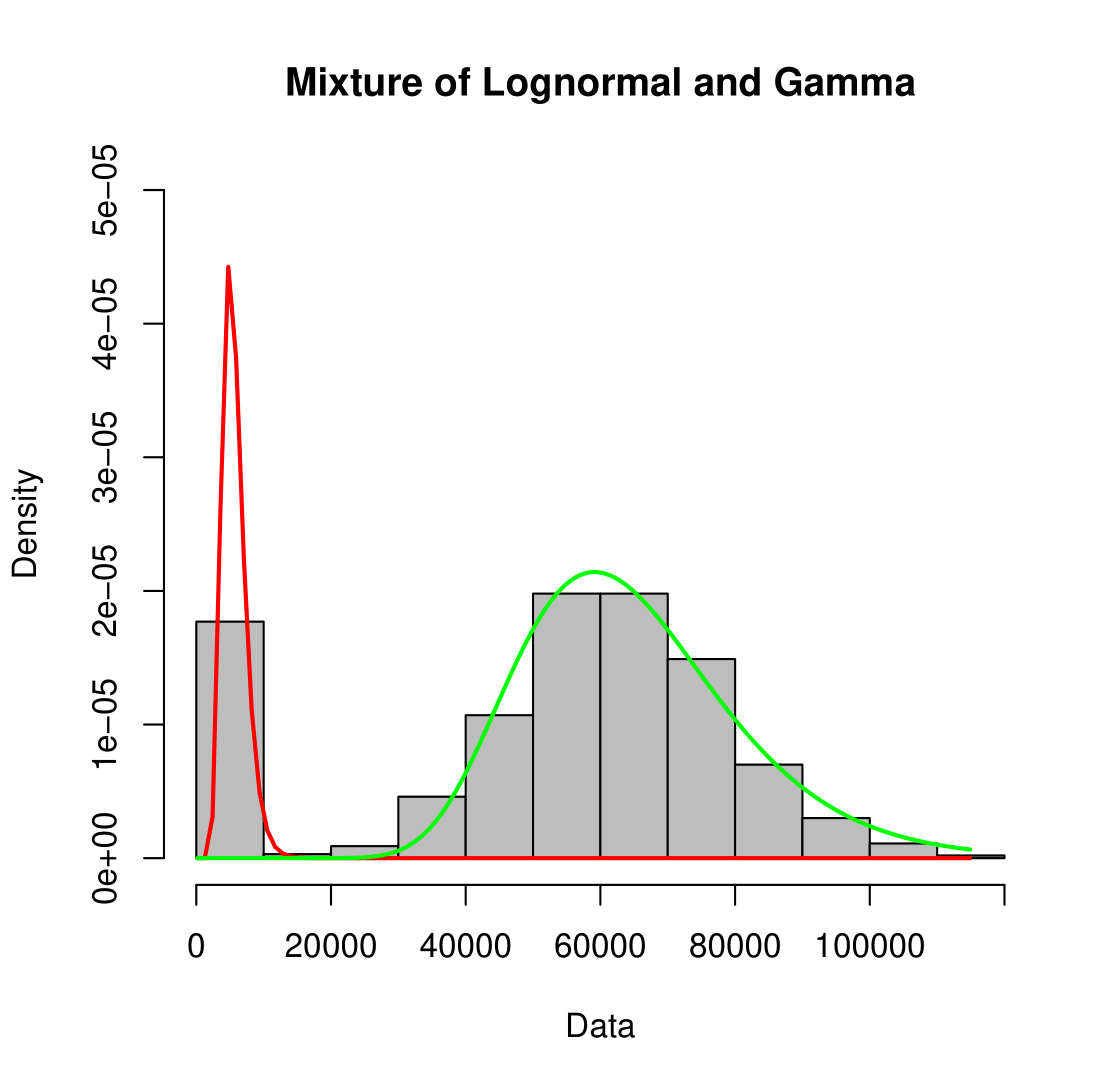


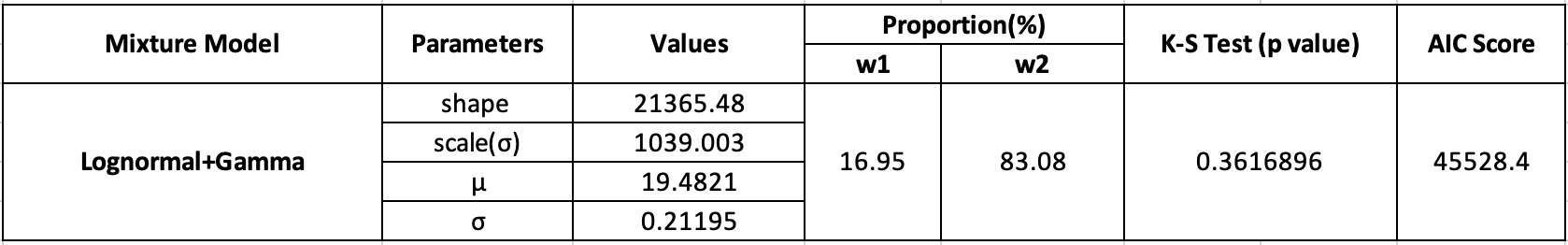






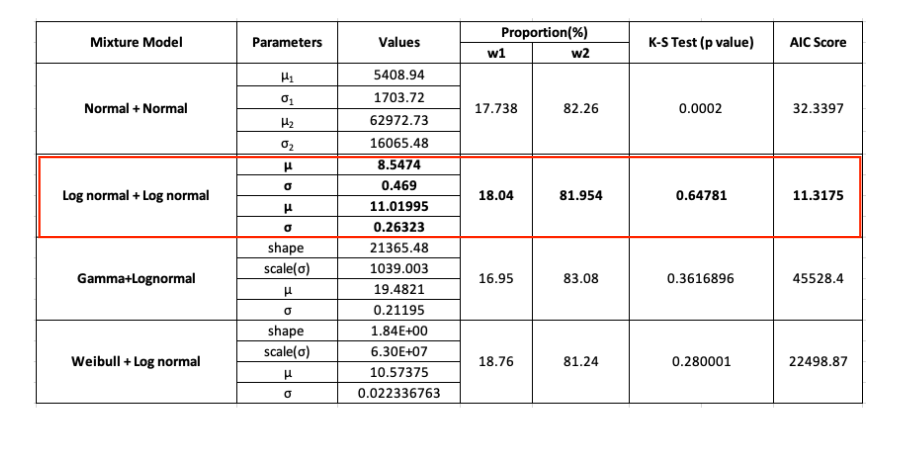






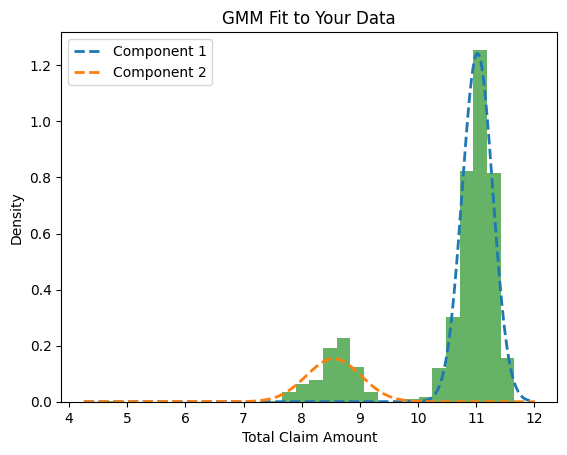
ESTIMATION RESULTS

Here we have the following table for the estimation results



In the following analysis, we observed that the Log normal+Log normal mixture model had the lowest AIC score among the different mixture model so we preferred it.

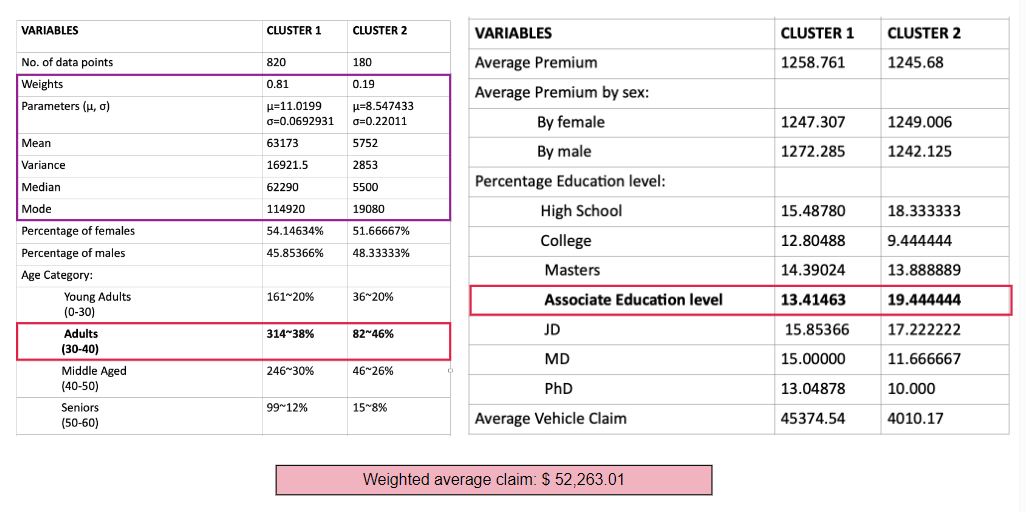
Mixture of a Lognormal distribution-





clusters were formed using the responsibilities provided by the GMM model. Responsibilities are nothing but the probability with which each data point is likely to have come from the mixture distributions

The results that were obtained

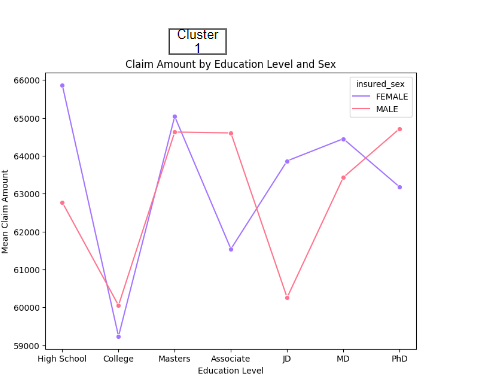


From our analysis, we did observe that the weighted average came out to be 52,263.01 dollars.

Summary and Conclusion

Thus, we can say that the finite mixture of Lognormal distributions can be fitted to motor insurance claims and we can take out the following inferences.

Line Plot between education level and mean claim amount

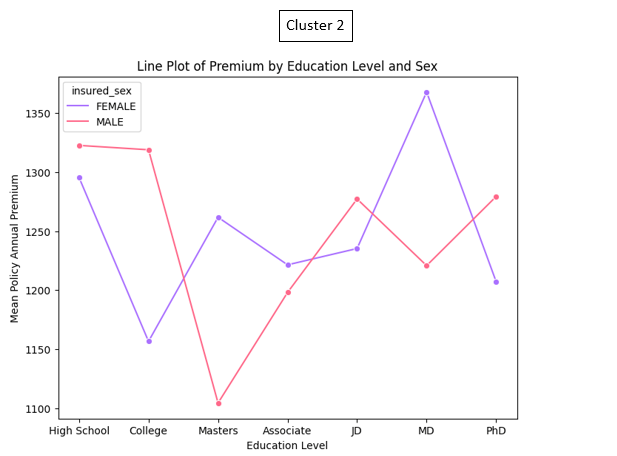


Conclusion obtained

High school males and males with associate and working professional, Ph.D. degrees are likely to go for higher premiums.

Females with MD degrees go for a higher premium

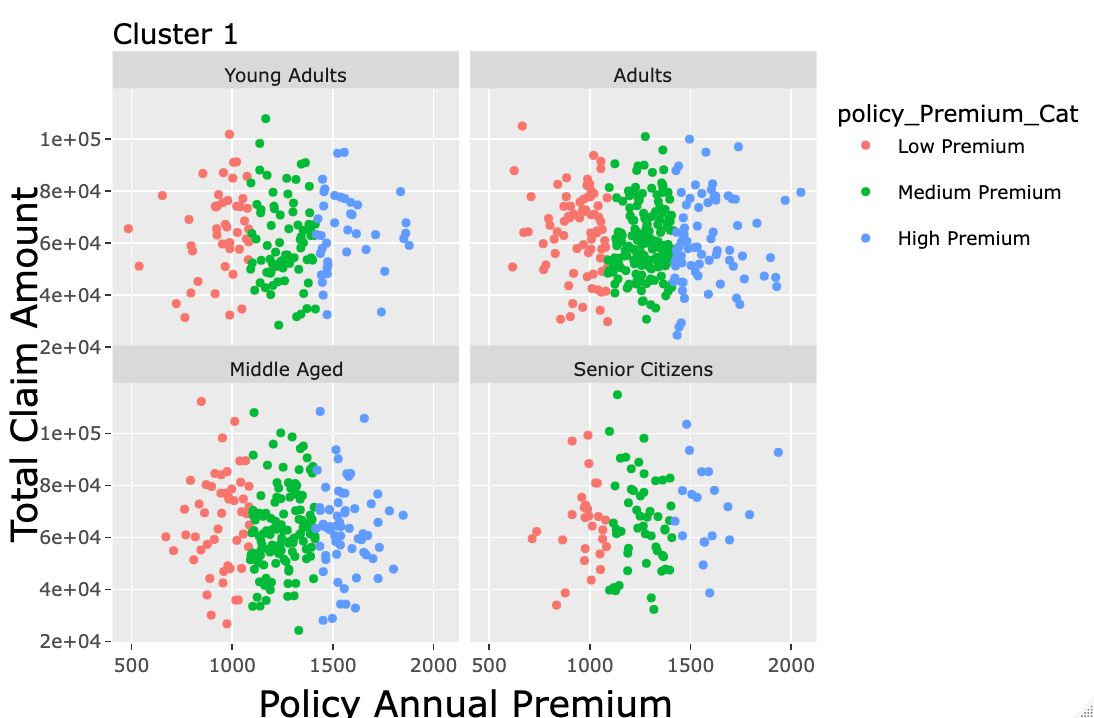
Line Plot of Premium by Education Level and sex



Males with PhD degrees are likely to go for higher premiums and those with Master’s degrees are likely to go for lesser premiums.

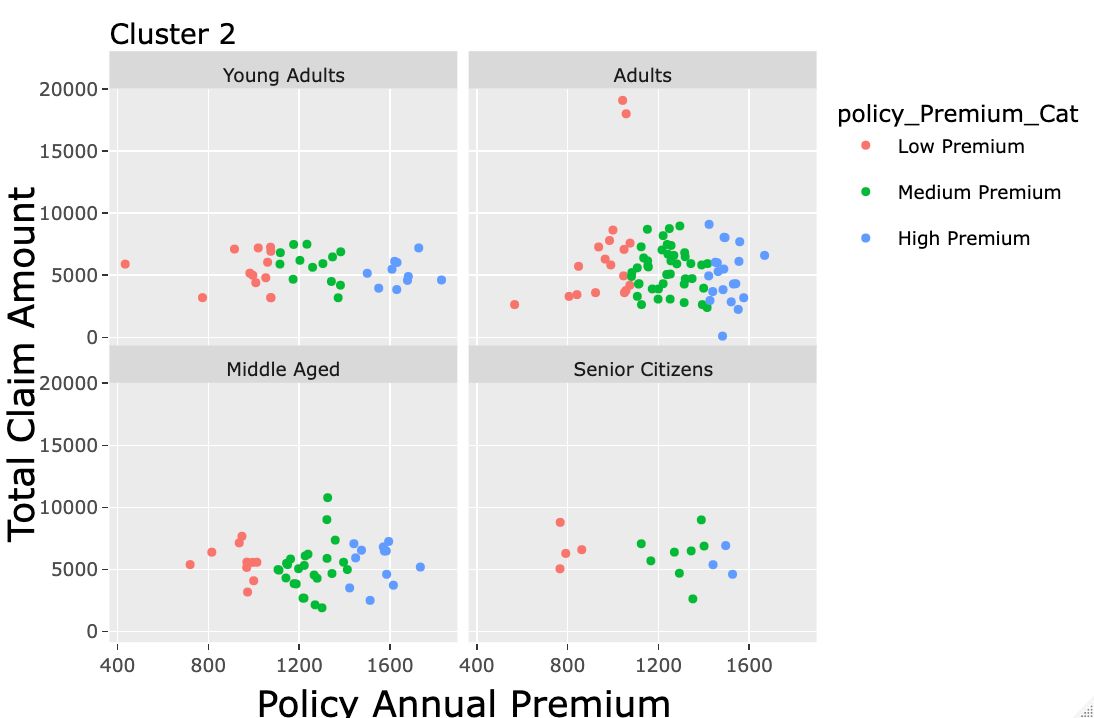
Females with MD and Master’s degrees go for a higher premium

Cluster Plot 1



For cluster 1 we observe that young adults with lack of knowledge and less spending power tends to choose revolve around any kind of policy plan but in case of middle aged , senior citizens having more spending power and more knowledge about the premium plans they stick with medium and high premium only.

Cluster Plot 2



For cluster 2 we observe that most of the policy premiums tend to lie around a specific limit of total claim amount which basically means there’s a miniscule change in total claim amount as your premium is increasing

LIMITATIONS

1. Data Volume and Computational Complexity:
   * Estimating the parameters of a mixture distribution, especially with a large number of claims, can be computationally expensive and may pose challenges in terms of processing time and resource requirements.
2. Number of Components
   * Choosing too few components may oversimplify the model, failing to capture the underlying data structure.
   * Choosing too many components can lead to overfitting and increased computational complexity.
3. Identifiability:
   * Different combinations of component distributions can produce the same overall distribution.
   * Challenges in interpreting parameters and estimating the true underlying distribution.
4. Initialization Sensitivity:
   * In EM algorithms, Poor initializations may lead to convergence to local optima instead of the global optimum.

References

RESEARCH PAPERS

# 1)Fitting Finite Mixtures of Generalized Linear Regressions on Motor Insurance Claims January 2017[International Journal of Statistical Distributions and Applications](https://www.researchgate.net/journal/International-Journal-of-Statistical-Distributions-and-Applications-2472-3487?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) 3(4):124 DOI:[10.11648/j.ijsd.20170304.19](http://dx.doi.org/10.11648/j.ijsd.20170304.19)

Authors:

[Nana Kena Frempong](https://www.researchgate.net/profile/Nana-Kena-Frempong) **(**[Kwame Nkrumah University Of Science and Technology](https://www.researchgate.net/institution/Kwame-Nkrumah-University-Of-Science-and-Technology?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19))

# 2) Fitting of Finite Mixture Distributions to Motor Insurance Claims December 2011[Journal of Mathematics and Statistics](https://www.researchgate.net/journal/Journal-of-Mathematics-and-Statistics-1549-3644?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19) 8(1):49 DOI:[10.3844/jmssp.2012.49.56](http://dx.doi.org/10.3844/jmssp.2012.49.56)

Authors: [P. Sattayatham](https://www.researchgate.net/profile/P-Sattayatham) **(**[Suranaree University of Technology](https://www.researchgate.net/institution/Suranaree_University_of_Technology?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19))

3)Modeling of Motor Insurance Extreme Claims through  Appropriate Statistical Distributions

V SELVAKUMAR, DIPAK KUMAR SATPATHI, P T V PRAVEEN KUMAR, V V HARAGOPAL,Department of Mathematics, Birla Institute of Technology and Science – Pilani, Hyderabad, India. BHAVAN’S VIVEKANANDA COLLEGE OF SCIENCE, HUMANITIES, AND COMMERCE, HYDERABAD, INDIA.

Email: vskselva79@gmail.com

LINKS

For EM Algorithm:

https://onlinelibrary.wiley.com/doi/book/10.1002/9780470191613

For Mixture Distributions:

https://medium.com/@smallfishbigsea/an-explanation-of-discretized-logistic-mixture-likelihood-bdfe531751f0#:~:text=Mixture%20of%20Distributions&text=A%20mixture%20of%20distriction%20is,power%20by%20introducing%20more%20parameters.

For Lognormal Distributions:

https://www.sciencedirect.com/science/article/abs/pii/S2211692318300663

APPENDIX

Some other conclusions -

BEHAVIOUR OF TOTAL CLAIM AMOUNT AND CAR MANUFACTURER FOR THE PREMIUM CATEGORY

CLUSTER 1

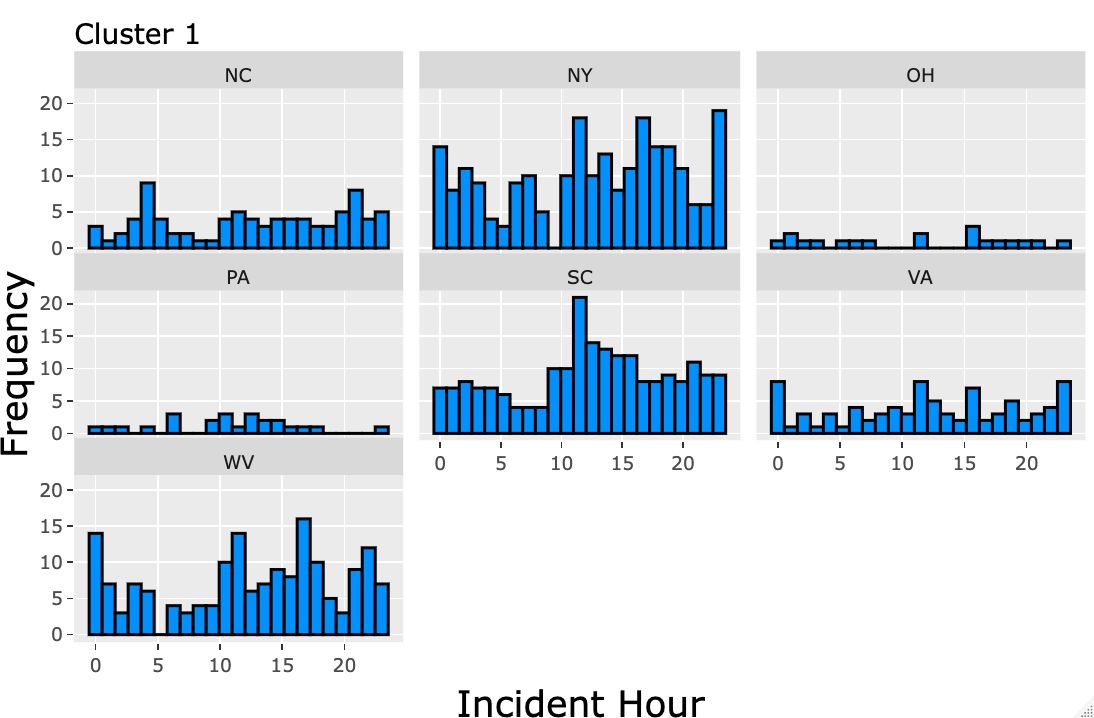


CLUSTER 2

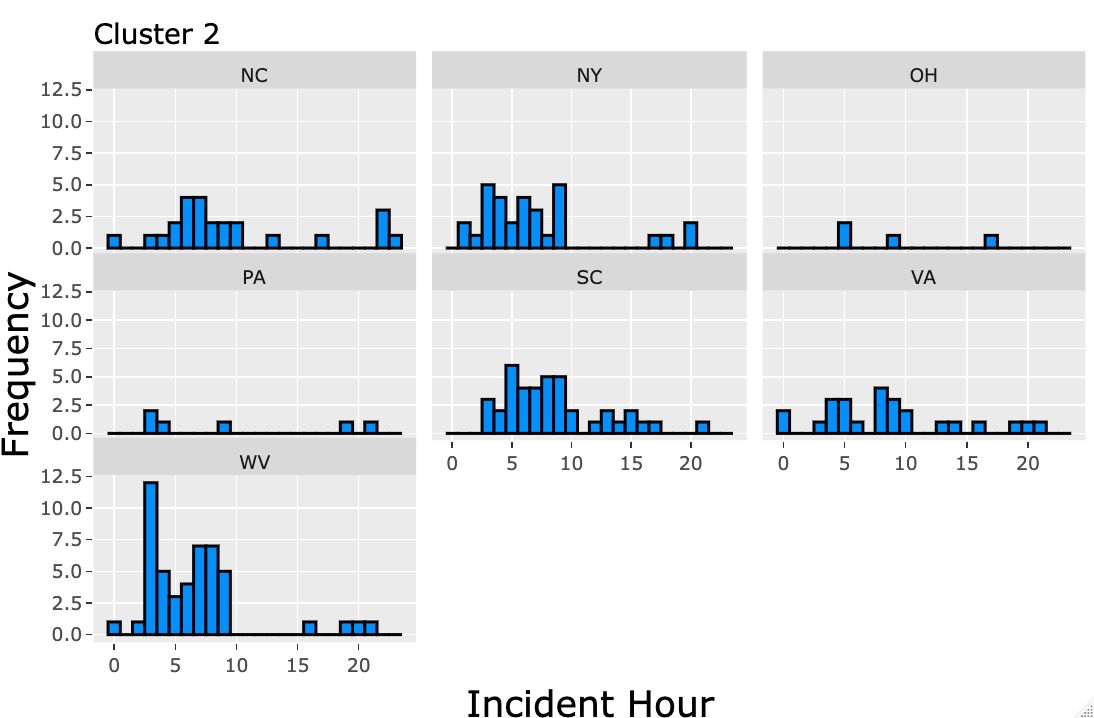


Some other conclusions -

For Cluster 1 we observe that in busy states like New York, South Carolina and West Virginia accidents can occur at any given interval of the day with higher frequency-



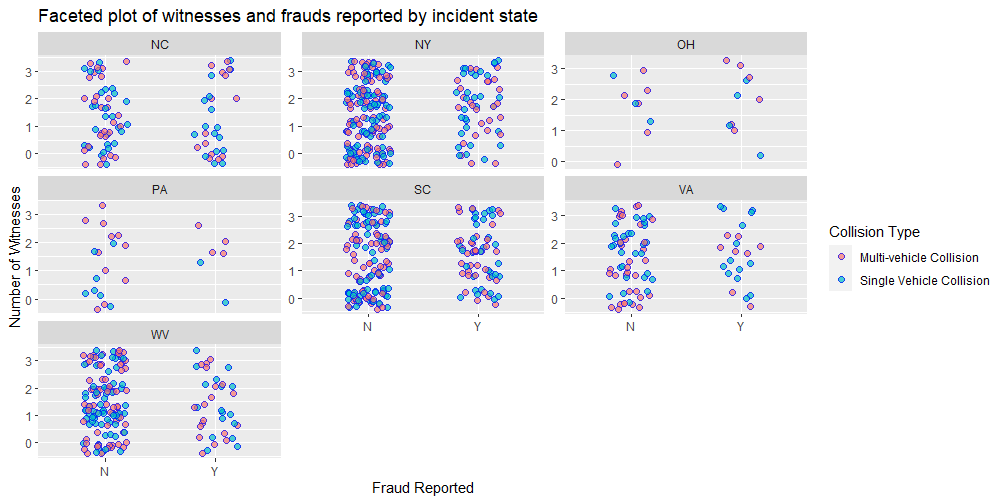
For Cluster 2 we observe that the majority of the accidents occur till 10th hour of the day but there’s a minuscule change in frequency as the day progresses



FACETED PLOTS CONCLUSION



Faced plot of witness and frauds by incident state



We can comment that in states such as New York, South Carolina and Virginia the ratio of fraud not getting committed and fraud getting committed are relatively similar.

Alternate Method

Alternate Method

Alternate Metho

Mixture of Lognormal

Mixture of Lognormal

Mixture of Lognormal

Mixture of Summary and Conclusion