MODELING LIABILITY INSURANCE CLAIM SEVERITY USING THE GAMMA FAMILY OF DISTRIBUTIONS

Stanley Sayianka Saitet

Lameck Kenyaga

Winfred Kinya Bundi

Amos Njagi Kaberia

Khalif Ali Hussein

Emmanuel Watia Wambua

A Research Project Submitted to the Department of Mathematics in Partial Fulfillment of the Requirements for the Award of Degree in Bachelors of Science in Actuarial Science of Egerton University

EGERTON UNIVERSITY

OCTOBER, 2022

DECLARATION AND RECOMMENDATION

DECLARATION

We declare that this is our original work and that it has not been presented for examination by anyone in this university or any other university for the award of any degree.

Name: Stanley Sayianka Saitet

Registration number: S19/03431/18

Sign: _____

Name: Lameck Kenyaga

Registration number: S19/03424/18

Sign: _____

Name: Winfred Kinya Bundi

Registration number: S19/03400/18

Sign: _____

Name: Amos Njagi Kaberia

Registration number: S19/03432/18

Sign: _____

Name: Khalif Ali Hussein

Registration number: S19/03415/18

Sign: _____

Name: Emmanuel Watia Wambua

Registration number: S19/09586/17

Sign: _____

RECOMMENDATION

This research project has been submitted with our approval as university supervisors.

Mr. Kenneth Langat

Department of Mathematics

Egerton University

Sign…… Date……

Dr. Cox L.Tamba (Ph.D.)

Department of Mathematics

Egerton University

Sign……. Date……

ACKNOWLEDGEMENT

We are grateful to God for giving us the chance to come this far, in our academic pursuits.

We are grateful to out project supervisors, Mr. Langat and Dr. Cox for their valuable support, suggestions, ideas and direction in finishing this work.

We would like to acknowledge the input of our fellow students, for the support and suggestions in numerous topics.

Finally, we are grateful to our families and friends for the encouragement, which went a long way in supporting our work.

ABSTRACT

Here comes the abstract

Table of Contents

Table of Figures

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Liability insurance is a risk-transfer mechanism. It exists to protect the insured from events such as injuries to people, and damage of third-party property. It differs from other forms of insurance in that in the event of the occurrence of the insured risks, the insurance company compensates the affected third parties rather than the insured. Forms of liability insurance include: customer injury lawsuit, property damage lawsuit, indemnity insurance, employer's liability, director's liability, professional indemnity insurance, product liability as well as operations and commercial liability.

Originally, individual companies that faced a common peril formed a group and created a self-help fund out of which to pay compensation should any member incur loss (in other words, a mutual insurance agreement). The modern system relies on dedicated carriers, usually for-profit, to offer protection against specified perils in consideration of a premium.

Liability insurance is one of the fastest growing insurance sectors with a global market size value of more than 25 billion dollars and a projected global market size of 433 billion dollars by 2031. This type of insurance is important in an economy in providing protection from actions, which arise out of unintended negligence, which often give rise to infrequent but large losses. The nature of claim severity coupled with the wide range of cover arising out of liability insurance sets it apart from other forms of general insurance, such as auto-insurance.

Liability insurance actuaries are often interested in accurately modeling claims severity, as well as the extreme events such as the possibility of larger than normal losses in an attempt to depict the uncertain behavior of future claims payments. This uncertainty necessitates the use of probability distributions to model the occurrence of claims, the timing of the settlement and the severity of the claims. In this study, we focus on modeling the severity of liability claims using gamma family distributions.

1.2 Statement of the Problem

For major liability insurers in Kenya, the rate of claim frequency is lower as compared to other policies such as motor-insurance, life and health insurance, however, when the insured risks occur, they often result in large catastrophic losses making liability insurance a sensitive sector to manage.

The nature of claim severity in this insurance sector necessitates the use of such distributions with the so-called heavy-tails, which are better equipped to handle extreme events such as large losses.

Accurate modeling using statistical distributions, which are capable of capturing the tails of the claim severity is necessary when dealing with liability claims due to the heavy-tailed nature of such claims. Choosing the best distribution to model the claim severity ensures accurate reserve allocation, premium revisions and profitability in the liability insurance sector.

1.3 Objectives

1.3.1 General Objective

To model claim severity data from liability insurance using five models from the gamma family, and compare them to gauge their accuracy using goodness of fit tests, and graphical tests.

1.3.2 Specific Objectives

The following specific objectives are of interest in our project:

1. To fit the models on the data and estimate the parameters of each of the models fitted using the training set of the claim severity data.
2. To compare the models fitted using the Goodness-of-fit test, and graphical methods.
3. To fit the models on the testing set of the claims severity data, obtain measures of fit and update model parameters.

1.4 Justification of the Study

Modeling claim severity of an insurance company is an essential part of insurance concerning: setting up reserves, pricing policies, forecasting future claims experience and reinsurance planning. Pricing of insurance products is important for both insurers and the insured. Well-priced products will ensure that insurance companies have adequate cash to settle claims and create sufficient reserves for future claims settling.

This research is crucial in that it enables insurance companies make educated choices on premium pricing, hence sufficient provisions would be made for reserves, which would decrease their risk of running into ruin.

This study will also act as a road map for regulators within the insurance space such as: The Insurance Regulatory Authority of Kenya when making financial judgments and formulating monetary policies regarding liability insurance. This study will lay a foundation for future academics who plan to research more about Claims modeling and provide more insight into current research on the gamma family of distributions in claims severity modeling.

1.5 Definition of Terms

Claim severity: This refers to the monetary loss of an insurance claim.

Likelihood: This is a function which indicates how likely a particular population is to produce an observed sample.

Tail Weight: In statistics, the specific information about the shape of a distribution contained in its extreme values is measures using the concept of tail weight.

Franchise: Provisions in an insurance policy, where the insurer does not pay unless damage or loss exceeds a given amount.

1.6 Limitations of the study

One of the limitations of the study is that it is not exhaustive in its application of the Gamma family of distributions to model claim severity. The Gamma family of distributions is a vast family of more than twenty distributions, however we restrict our attention to only five of them, which will act as a representative for the rest.

Another limitation of the study is the enforcement of franchise amount of 1000 shillings.

1.7 Scope of the study

This study utilizes data from a liability insurance company in Kenya with records from the year 1998 to 2020. We filter the data to include Liability insurance claims, by dropping the property claims. The variable of interest in this study is the claim severity of the liability insurance portfolio.

In this study, we cover five statistical distributions from the gamma family, which are: The exponential distribution, the gamma distribution, the Weibull distribution, the Pareto distribution and the burr distribution.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, we are interested in discussing the empirical and theoretical framework. It begins with the various studies that have been carried out on claim severity modeling in the first section. In the second section, the gamma family distribution is then presented with its various characteristics.

MeelisKaarik (2012), modeled third party liability claims from an insurance company using heavy tailed distributions. The distributions studied under the research included: a mixture of the lognormal distributions and the generalized Pareto distribution. The Generalized Pareto Distribution (GPD) was fitted to the extreme tails of the data. The research aimed at estimating parameters for risk measures and constructing threshold levels for various risk measures. The study recommends a quantile-based method for insurance loss modeling.

Robert (2015) studied modeling extreme claims using the Exponential, Uniform and Pareto Distribution, with a focus on extreme value theory. The insurance claims being modeled were from a fire insurance portfolio for a 10-year period in Kenya. The study applies Maximum Likelihood Estimation (MLE), method of moments (MoM) and Linear combination of moments method(L-Moment) in fitting the distributions to the data, and the best fit was later evaluated using the QQ plots. The pareto distribution emerged as the best fitting distribution, followed by the exponential distribution. The study concludes that the Pareto, and Generalized Pareto Distribution are useful in modeling heavy-tailed severity claims when they exceed higher threshold levels.

Packova and Brebera (2015) used Gamma, Weibull, Lognormal and Pareto distributions to model data obtained from a Czech insurance company for compulsory motor third-party liability insurance. The Maximum Likelihood Estimator method was used to estimate the parameters of the selected parametric distributions. They further used the Anderson.Darling, Chi-Square and Kolmogorov-Smirnov tests to determine whether the chosen distribution provides a good fit to the data. The finding of the study was that the Pareto distribution can be assumed to be a good model for the losses.

Das et al. (2016), attempted to use the Burr XII distribution as an actuarial risk model. The study fits the distribution to data, and focuses on computing the probability of ultimate ruin using a recursive algorithm, the fit of the distribution is evaluated using the empirical distribution function (EDF) and related statistics. The claim severity data used in the study was from property insurance of natural catastrophic events, and had the characteristics of large but infrequent claim sizes. The study concludes that, due to the characteristics of the data, the burr XII distribution provides a good fit for such kind of claim severity data.

Nath et al. (2016) conducted research on fitting two heavy tailed distributions on insurance claims data namely: the Weibull and Burr XII distribution. The study focuses on modeling and computing the probability of ultimate ruin for the insurance portfolio under the Classical Risk Model. The data used were from property insurance and fire insurance claim severity. The distributions were fitted to data using the Maximum Likelihood Estimation, and the fit was compared using the Anderson-Darling test and the Cramer Von Mises test. The distribution with the best fit found was the Weibull distribution.

Anyanumeh (2016), did a comparison of risk classification methods for claims severity. They compared several risk classification methods for claim severity data by using an equation which was written as the weighted difference between the observed and fitted values. The weighted equation was applied to estimate claim severities which was equivalent to the total claim cost divided by the number of claims. From their data, they also observed that, the classical and regression fitting procedures gave equal values for parameter estimates however, the regression procedure provided a faster convergence. The multiplicative and additive models gave similar parameter estimates. The smallest chi-squares were given by the minimum chi-squares model except for the exponential model. All models provided similar values for absolute difference. This study therefore was meant to come up with best model among the chosen.

Omari et al. (2018) modeled a sample of the automobile portfolio data-sets obtained from the insurance Data package in R. They used the Maximum Likelihood Estimation method to obtain parameter estimates for the fitted models. TI1e Anderson-Darling and Kolmogorov-Smirnov tests were then used as goodness-of-fit tests for the claim severity models. The Akaike Information Criterion and Bayesian Information Criterion were further applied to choose between competing distributions. The finding of their study was that the log normal distribution

provides a good model for claims severity on a short-term basis. The study recommended that for a long-term basis, insurers should adjust the distributions accordingly based on insurer-specific claims experience.

Ng'elechei et al. (2020), used secondary data from APA insurance company to model the severity of insurance claims. The study aimed at finding the best statistical distribution for fitting past claims motor data. The parameters were estimated using the MLE method. The chi-square and Anderson-darling tests were used to check the goodness of fit of the frequency distributions. The Pareto model was found to be the best fit model for the severity claim data among the models tested.

Okindo (2021), used the Exponential, Gamma, Log-normal, Pareto and Weibull distributions to model claims severity data of motor insurance in Kenya given that they are positively skewed distributions. The parameters for these distributions were estimated using the MLE method then the Anderson – Darling test was applied as a goodness-of- fit test on the fitted distributions. The study concludes that the Weibull and Gamma distributions are suitable for modeling motor commercial and motor private data respectively.

CHAPTER THREE

MATERIALS AND METHODS

This chapter aims to discuss the methodology, which includes: the discussion of gamma family distribution in an attempt to model liability claim severity, assumptions of parameter estimation and testing the goodness of fit.

The following are the assumptions used when modelling the liability claim severity.

i. The claims are independently distributed, so that the occurrence of a claim from a particular policy is not affected with other policy claim history.
ii. The minimum amounts of claim severity is one thousand shillings due to the enforcement of franchise.

3.1. Distributions

The gamma family of distributions is a large family of distributions with several distributions ranging from the simple one parameter exponential distribution to more complex distributions such as the burr distribution.

Gamma family of distribution is characterized by the gamma function as shown below

$$\Gamma(\alpha)=\int_0^\infty y^{\propto -1}e^{-y}\,dy$$

This family of distributions has high positive skewness, hence suitable for modeling strictly positive random variables. The distributions are also be categorized by tail weight. The concept of tail weight is used to compare distributions based on the probability assigned to large values, so that: distributions which assign higher probabilities to larger values are said to be heavier-tailed. In the gamma family of distributions, light-tailed distributions include: the exponential distribution, the chi-squared distribution and the gamma distribution, while heavy-tailed distributions include: the Weibull distribution, the Pareto distribution and the Burr distribution.

3.1.1 The Exponential distribution

The exponential distribution is one of the elementary models for claim severity since it is a simple distribution with one parameter. The distribution has the following properties:

The probability density function is:

$$f(x)=\theta e^{-\theta x}$$

The cumulative distribution function is:

$$F(x)=1-e^{-\theta x}$$

The maximum likelihood estimation for the distribution is derived as follows:

$$L(\theta)=\prod_{i=1}^{n}\theta e^{-\theta x_i}$$

$$\log L(\theta)=n\log\theta-\theta\sum_{i=1}^{n}x_i$$

$$\hat{\theta}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\acute{x}}$$

3.1.2 The Gamma distribution

The gamma distribution is a central distribution in loss modeling. The gamma distribution is a two-parameter distribution with the shape and scale parameters. The gamma distribution is an extension of the exponential distribution since it is the sum of independently distributed exponential random variables. This makes it more suitable for loss modeling since it has two tunable parameters. The distribution has the following properties:

The probability density function:

$$f(x)=\frac{\beta^{\alpha}}{\Gamma\alpha}x^{\alpha-1}e^{-\beta x}$$

The cumulative density function of the gamma distribution is approximated using the chi-squared tables, using the transformation that:

$$if\ X\ \ Gamma(\alpha,\beta)$$

$2\beta X \quad \chi^2_{2\alpha}$

The maximum likelihood estimation for the parameters of the gamma distribution are derived as follows:

$$L(\alpha,\beta)=\prod_{i=1}^{n}\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$$

$$¿(\frac{\beta^{\alpha}}{\Gamma(\alpha)})^{n}x_i^{n\alpha-n}e^{-\beta\sum_i^n x_i}$$

$$\log L(\alpha,\beta):n\alpha\log\beta-n\log\Gamma(\alpha)+n(\alpha-1)\log x_i-\beta\sum_{i=1}^{n}x_i$$

$F(\alpha)=\frac{\delta}{\delta\alpha}(\Gamma(\alpha))$, This is commonly referred to as the di-gamma function.

$$\hat{\beta}=\frac{\hat{\alpha}}{x}$$

3.1.3 The Weibull distribution

The Weibull distribution is a heavy tailed distribution with a wide variety of applications ranging from survival analysis to loss modelling. The Weibull distribution is a transformed distribution, obtained by raising the exponential distribution to a power. The distribution has the following properties:

The probability density function:

$$f(x)=\frac{\gamma\left(\frac{x}{\beta}\right)^{\gamma}e^{-\left(\frac{x}{\beta}\right)^{\gamma}}}{x}$$

The cumulative density function:

$$F(x)=1-e^{-\left(\frac{x}{\beta}\right)^{\gamma}}$$

The maximum likelihood estimation for the parameters of the Weibull distribution are derived as follows:

$$L(\gamma,\beta)=\prod_{i=1}^{n}\frac{\gamma\left(\frac{x}{\beta}\right)^{\gamma}e^{-\left(\frac{x}{\beta}\right)^{\gamma}}}{x}$$

$$\log L(\gamma,\beta)=\log[\left(\frac{\gamma}{\beta^{\gamma}}\right)^{n}\prod_{i=1}^{n}x_i^{(\gamma-1)}*\exp\left\{-\left(\frac{\sum_{i=1}^{n}x_i}{\beta}\right)^{\gamma}\right\}]$$

$$n\log(\gamma)-\gamma n\log(\beta)+(\gamma-1)\sum_{i=1}^{n}\log(x_i)-\sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\gamma}$$

The estimates for the parameters are then estimated from the above non-linear equation using Newton-Raphson's iterative algorithm, given possible starting points.

3.1.4 The Pareto distribution

The two parameter Pareto distribution also known as the Lomax or the Pareto Type II is a mixture distribution obtained when an exponential distribution is mixed with a gamma distribution. The Pareto distribution first emerged as a distribution for modeling the wealth distribution, due to its heavy tailed nature. The Pareto distribution is also used in loss modeling due to its heavy-tailed nature and its adaptability. The distribution has the following properties:

The probability density function:

$$f(x)=\frac{\alpha\,\lambda^{\alpha}}{(x+\lambda)^{\alpha+1}}$$

The cumulative density function:

$$F(x)=1-\left(\frac{\lambda}{x+\lambda}\right)^{\alpha}$$

The Maximum likelihood estimates for the parameters of this distribution are derived as shown below:

$$\log L(\alpha,\lambda)=nlog(\alpha)+\alpha n\log(\lambda)-(\alpha+1)\sum_{i=1}^{n}\log(\lambda+x_i)$$

$$We\,rewrite\,it\,as:\log(L(\alpha,\lambda))=nlog(\alpha)+\alpha n\log(\lambda)-(\alpha+1)S(\lambda)$$

We take the derivatives as shown below:

$$\hat{\alpha}=\frac{n}{S(\lambda)-nlog(\lambda)}$$

$$n\frac{\alpha(\lambda)}{\lambda}=(\alpha(\lambda)+1)S'(\lambda),\,where\,we\,consider\,\alpha=\alpha(\lambda)$$

We solve for λ using the Newton-Raphson's iterative algorithm.

3.1.5 Burr distribution

The Burr distribution also known as the Burr Type XII or the Singh–Maddala distribution is a common distribution used in loss modeling both for insurance and re-insurance events. This distribution I heavy-tailed and can be obtained by raising the two parameter Pareto distribution to a positive power, hence it can be regarded as a transformed Pareto distribution.

The probability density function:

$$f(x)=\frac{\alpha\gamma\left(\frac{x}{\theta}\right)^{\gamma}}{x\left[1+\left(\frac{x}{\theta}\right)^{\gamma}\right]^{\alpha+1}}$$

The cumulative density function:

$$F(x)=1-\pi^{\alpha}$$

$$\pi=\frac{1}{1+\left(\frac{x}{\theta}\right)^{\gamma}}$$

The maximum likelihood estimates for the parameters of the Burr distribution are shown below:

$$L(\alpha,\beta\vee x)=(\alpha\beta)^{n}\prod_{i=1}^{n}x_i^{\alpha-1}(1+x_i^{\alpha})^{-\beta-1}$$

The log-likelihood function reduces to:

$$l(\alpha,\beta\vee x)=n\log\alpha+n\log\beta+\sum_{i=1}^{n}\log\int_{0}^{\infty}x_i^{\alpha-1}(1+x_i^{\alpha})^{-\beta-1}\mu(x)\,dx$$

The MLE estimates for α,β are then obtained by taking partial derivatives and equating to 0, but since this is a highly non-linear likelihood function, then a closed form solution for the MLSE cannot be obtained and hence, the Newton-Raphson's iteration technique is used to find the solution.

3.2. Data

3.2.1 Scope of the Data

The data used for this study was fetched from a portfolio of liability insurance policies, from a major liability insurance company in Kenya. The data on claim severity ranges from 1998 to 2020, and the following variables are supplied: Claim ID, Claim Date, and Claim Amount.

3.2.2 Data analysis

The claim severity is first transformed using log-transformation, which takes care of the extreme skewness of the data. This is also useful in fitting the models as it ensures numerical stability.

For the model fitting process, the data is split into a training set and a testing set, whereby the training set comprises 85% of the total dataset, while the testing set comprises 15% of the data. The splitting was done after randomization of the claim severity data in order to remove any bias.

The training set is used to fit the model and tune the parameters of the five models, after which the three best models are then evaluated on the testing set to determine the best fitting model in terms of chosen metrics.

The data is also subjected to a graphical outlier-detection test, where any claim above 100,000,000 is regarded an outlier and is omitted from the training set. This ensures stability in the model fitting process, and also ensures that the measures of central tendency are not swayed wildly.

3.3. Model fitting process

The following process used when fitting the models chosen on the claim severity data is summarized below:

- Split the dataset into a training and testing dataset: The splitting is done based randomization to ensure 85% of the data is in the training set, while the testing set comprises 15% of the original data.
- Select a model from the model family chosen.
- Fit the model to the dataset, using maximum likelihood estimation: The five selected models will be fitted to the training dataset by means of Maximum Likelihood Estimation.

- Specify criteria for comparing the models fitted: The criteria to choose the model, will be based on graphical methods such as using QQ-plots, PP-plots, as well as using goodness of fit tests, and using information criteria such as the Bayesian Information Criteria, and the Akaike's Information criteria.

  In testing the model using the goodness of fit tests, a two-sided Kolmogorov-Smirnov test is employed of the form:

  Given the sample claims severity data from the training set of the form: $X_1, X_2, X_3, \ldots, X_n$ assumed to follow a particular population distribution F, for any particular chosen distribution F_0, the hypotheses statements are:

$H_0 : F = F_0$

$H_1 : F \neq F_0$

The test statistic used under the null hypothesis is denoted D and is given by:

$D := \sqrt{n} \vee F_n(x) - F_0(x) \vee ¿$

For this given test, if the null hypothesis is true, then the test statistic D tends to be small, while if the null hypothesis is not true, then the test statistic takes on large values.

The information criteria are given by:

$AIC = -2 * ¿(likehood) + 2k$

$BIC = -2 * ¿(likehood) + 2k + ¿(N) * K$

The Information criterion tend to take smaller values for better models, hence we prefer models with lower AIC and BIC scores.

- After fitting all models to the dataset, check the model fit, and decide on the best model for the claim's severity data.

CHAPTER FOUR

RESULTS AND DISCUSSIONS

4.1 Claim frequency over time

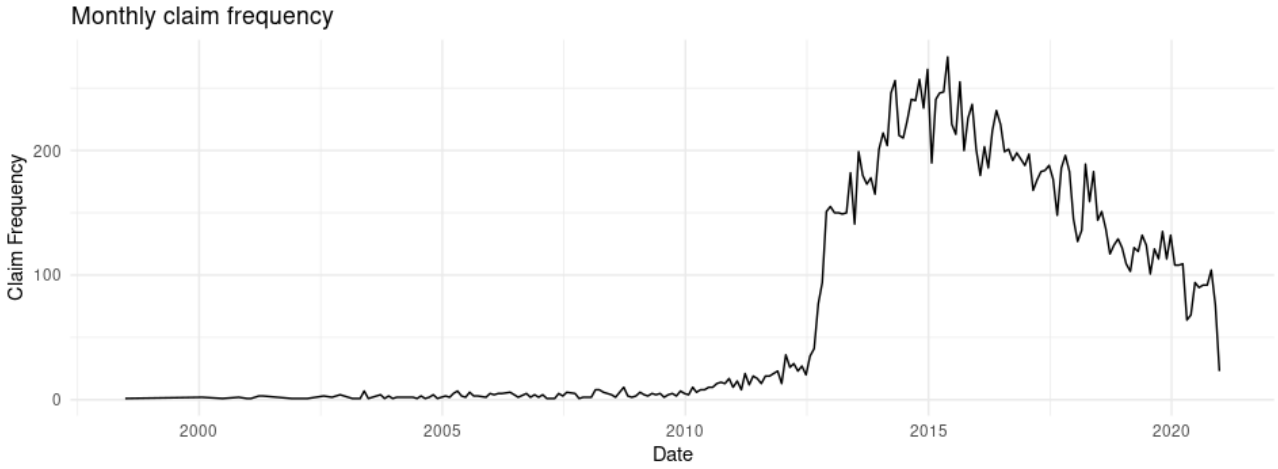

Figure 1: Claim frequency

We observe that the monthly claim frequency over the period 1990 to 2010 has been low, and below 50 claims per month, however the monthly claim frequency spiked in the year 2012 to 2015, and then begun gradually decreasing to 2020.

4.2 Claim severity data and the log-transformed claims

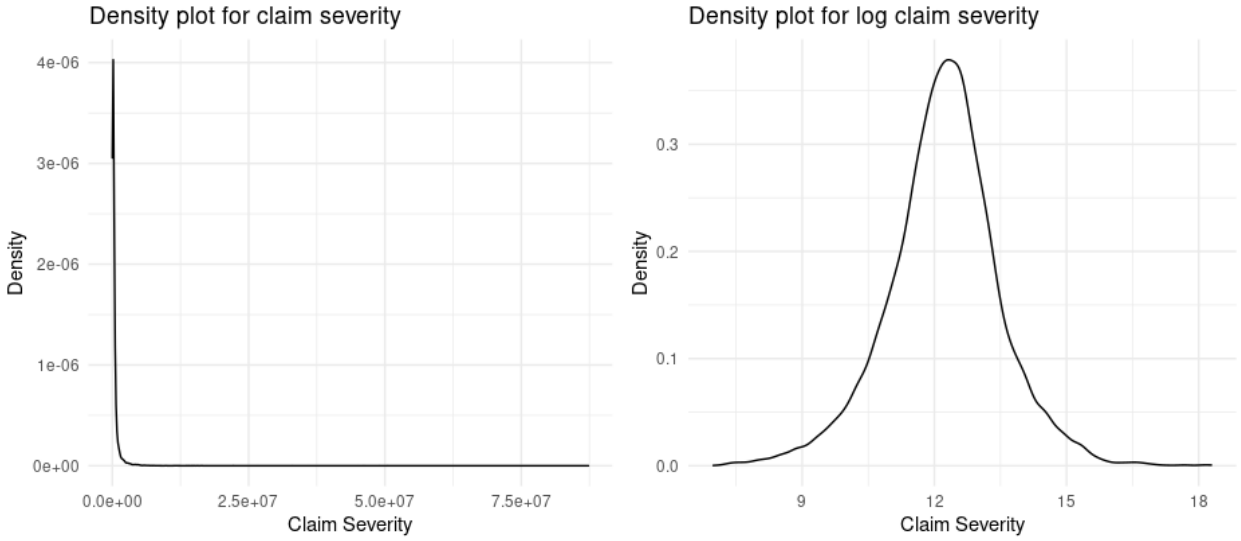

Figure 2: Overall Claim severity distribution

The claim severity density plot is highly skewed, and with fat tails with the distribution ranging from Kshs. 1000 to Kshs. 75,000,000, however after applying a log-transformation, the claim severity distribution becomes approximately symmetric, and within a good range of 6 to 18,

which is important for ensuring numerical stability in Maximum Likelihood Fitting of the models.

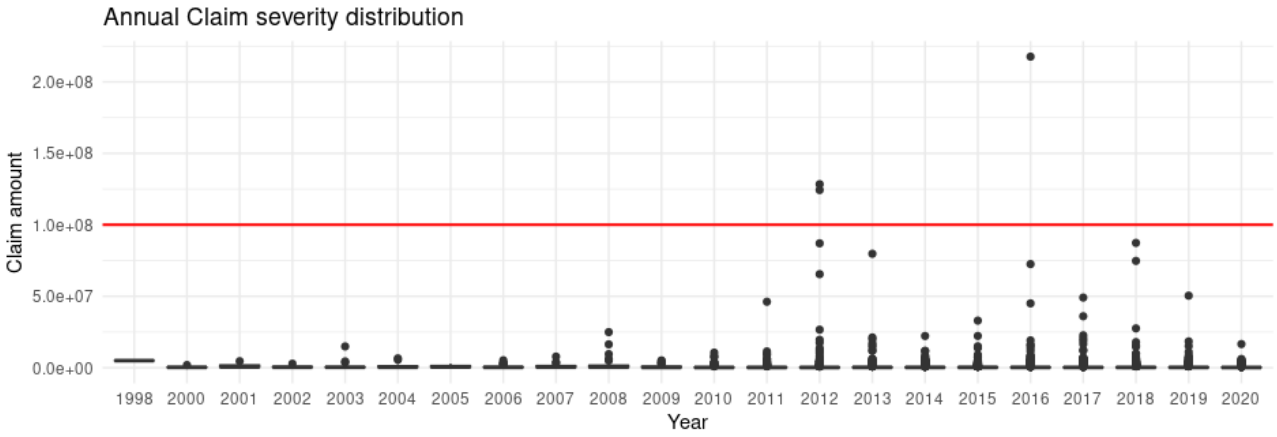

Figure 3: Annual claim severity distribution

We observe from the annual boxplot on claim severity data, that there are three outlier points (as shown by the points above the red line on 100,000,000). We omit the three outliers.

The summary statistics indicated below capture the distribution of the claim severity as well as the log-transformed version:

The original claim severity data:

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Skewness
1077	100906	211411	504696	422743	87314327	26.34648

Figure 4: Summary for claims severity

The log transformed claim severity data has the following summary statistics:

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Skewness
6.982	11.522	12.262	12.216	12.955	18.285	-0.1554675

Figure 5: Summary for log-claim severity

From the summary statistics above, it is evident that the claim severity data before transformation exhibits high positive skewness as indicated by the skewness statistic (26.34648), with a mean claim of 514,018, while for the log transformed claim severity data, the transformed claims have an approximately symmetric patters as shown by the skewness (almost 0), as well as the closeness of the mean, mode and median claim.

4.3 The fitted models statistics

This section covers the parameter estimation summaries for the five models by means of maximum likelihood estimation.

4.3.1 The exponential distribution

Fitting of the distribution ' exp ' by maximum likelihood

Parameters :

	estimate	Std. Error
rate	**0.08192395**	**0.000669633**

Loglikelihood: -52399.89 AIC: 104801.8 BIC: 104809.4

4.3.2 The Gamma distribution

Fitting of the distribution ' gamma ' by maximum likelihood

Parameters :

	estimate	Std. Error
shape	**91.003220**	**1.05018997**
rate	**7.455352**	**0.08627271**

Loglikelihood: -24866.33 AIC: 49736.65 BIC: 49751.88

Correlation matrix:

	shape	rate
shape	**1.0000000**	**0.9972541**
rate	**0.9972541**	**1.0000000**

4.3.3 The Weibull distribution

Fitting of the distribution ' weibull ' by maximum likelihood

Parameters :

	estimate	Std. Error
shape	**9.99203**	**0.05707005**
scale	**12.76716**	**0.01105460**

Loglikelihood: -25416.25 AIC: 50836.5 BIC: 50851.72

Correlation matrix:

	shape	scale
shape	**1.0000000**	**0.3275878**
scale	**0.3275878**	**1.0000000**

4.3.4 The Pareto distribution

The pareto distribution was fitted with the following starting values shape: 102115, and the scale: 468

Fitting of the distribution ' pareto ' by maximum likelihood

Parameters :

estimate Std. Error

shape 7050.712 90.58317

scale 85611.363 822.76250

Loglikelihood: -52401.14 AIC: 104806.3 BIC: 104821.5

Correlation matrix:

shape scale

shape 1.0000000 0.7706746

scale 0.7706746 1.0000000

4.3.5 The Burr distribution

The Burr distribution was fitted with the following starting values, shape1 2.5, shape2 2.5, and scale: 0.5.

Fitting of the distribution ' burr ' by maximum likelihood

Parameters :

estimate Std. Error

shape1 1.717064 0.06128418

shape2 15.043711 0.16732884

scale 12.809169 0.04534124

Loglikelihood: -24458.46 AIC: 48922.91 BIC: 48945.75

Correlation matrix:

shape1 shape2 scale

shape1 1.0000000 -0.7967202 0.9731561

shape2 -0.7967202 1.0000000 -0.8050158

scale 0.9731561 -0.8050158 1.0000000

4.4 Graphical Comparisons of Fit

4.4.1 The PP-plot technique

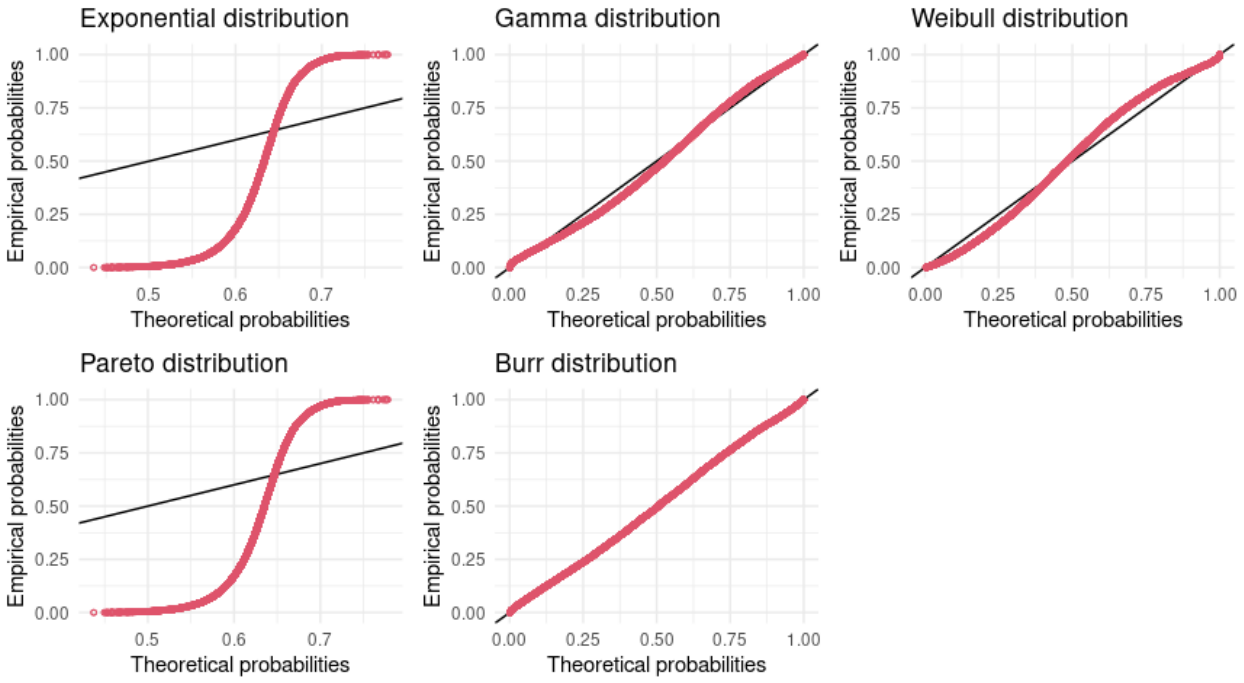

Figure 6: Probability-Probability plot

The chart on PP-plot was useful in comparing the cumulative density function for the log-transformed claims, and the fitted distributions, especially in the regions of high probability density in data. It was evident that the Burr, Gamma and Weibull distribution provided a good fit to the center of the data respectively, as compared to the rest which gave a bad fit.

4.4.2 The QQ-plot technique

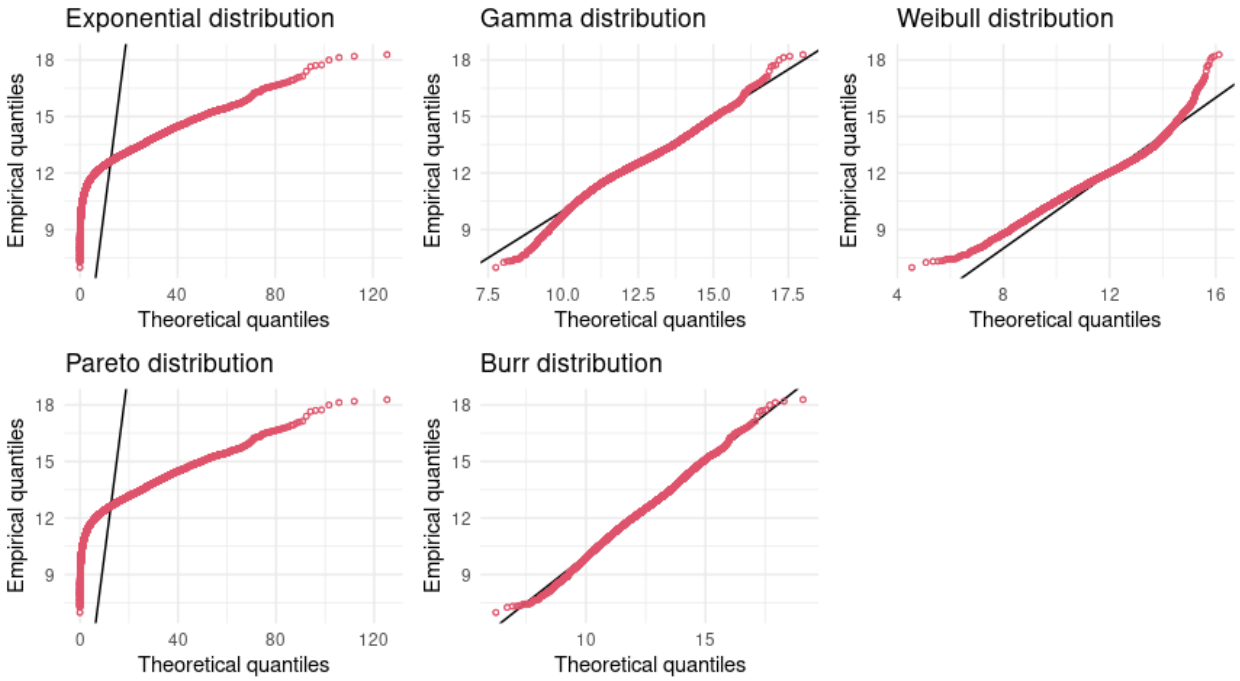

Figure 7: Quantile-Quantile plot

The QQ-plot chart above was useful in capturing the fit of the models with regards to the skewness of the log-transformed claims, as well as the location and scale of the data. It was evident that the Burr distribution provided the best fit, since it captured the tail of the data at hand quite well, as compared to the rest of the distributions. The Gamma and Weibull distribution come close in modelling accuracy, although they only capture the center of the distribution, and do not fit the tails of the distribution adequately.

4.5 Statistical Goodness-of-Fit tests

Goodness-of-fit statistics

	Exponential	Gamma	Weibull	Pareto	Burr
Kolmogorov-Smirnov statistic	0.516454	0.04984437	0.06733494	0.5183278	0.01683924
Cramer-von Mises statistic	1207.489743	12.10519413	27.07315268	1214.9639777	1.35732067
Anderson-Darling statistic	5583.245256	70.13492497	168.72334687	5613.1512300	8.47601424

Goodness-of-fit criteria

	Exponential	Gamma	Weibull	Pareto	Burr
Akaike's Information Criterion	104801.8	49736.65	50836.50	104806.3	48922.91
Bayesian Information Criterion	104809.4	49751.88	50851.72	104821.5	48945.75

Figure 8: Goodness of fit statistics

The goodness of fit statistics for the fit of the five distributions is shown below in tabular format:

From the above tables, the best fitting distribution based on the three tests: Kolmogorov-Smirnov, Cramer-von Mises, and Anderson Darling tests is the Burr distribution, as evidenced by the small statistics. The next best fitting distribution is the Gamma distribution, and finally the Weibull distribution. The Exponential and Pareto distribution did not provide a good fit to the data as indicated by their very large test statistics, and hence, we proceeded with only the top three distributions: Burr, Gamma and Weibull distribution.

The Akaike’s and Bayesian Information criterion prefer the Burr distribution as the best fitting distribution, and agree with the analysis generated by the tests above.

4.6 Comparison on test data

We proceeded with the three best fitting distributions the Burr, Gamma and Weibull distribution and attempted to compare their fit on previously unseen data (claim severities ranging from 2019 to 2020) in order to assess fit. The quantiles of the test claim severity distribution is compared to the quantiles of the three distributions by graphical methods, and then a formal two-sided Kolmogorov Smirnov test is applied.

4.6.1 Graphical fit

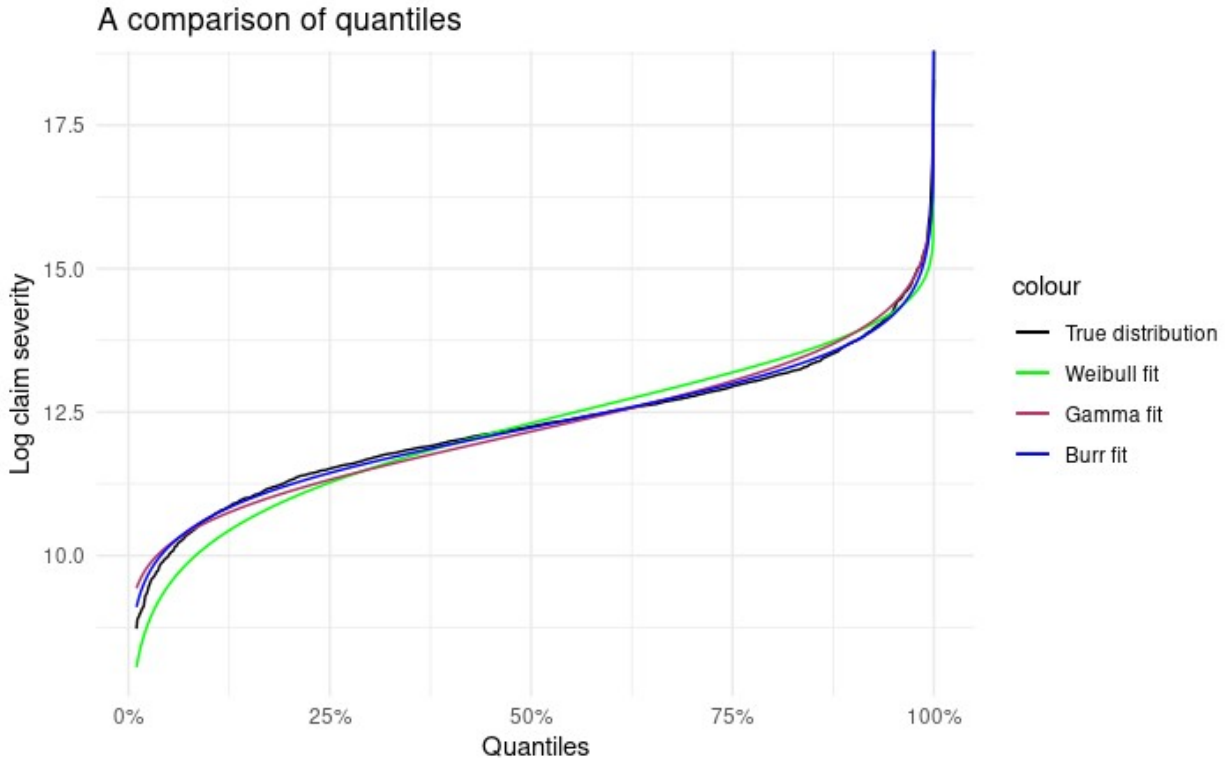

Figure 9: Comparison of quantiles

From the above comparison of quantiles plot, it is evident that the Gamma and Burr provide a better fit to the quantiles of the test claim severity data, both in terms of the center of the distribution and the tails of the distribution. The Weibull distribution is only able to give a good fit to the center of the distribution.

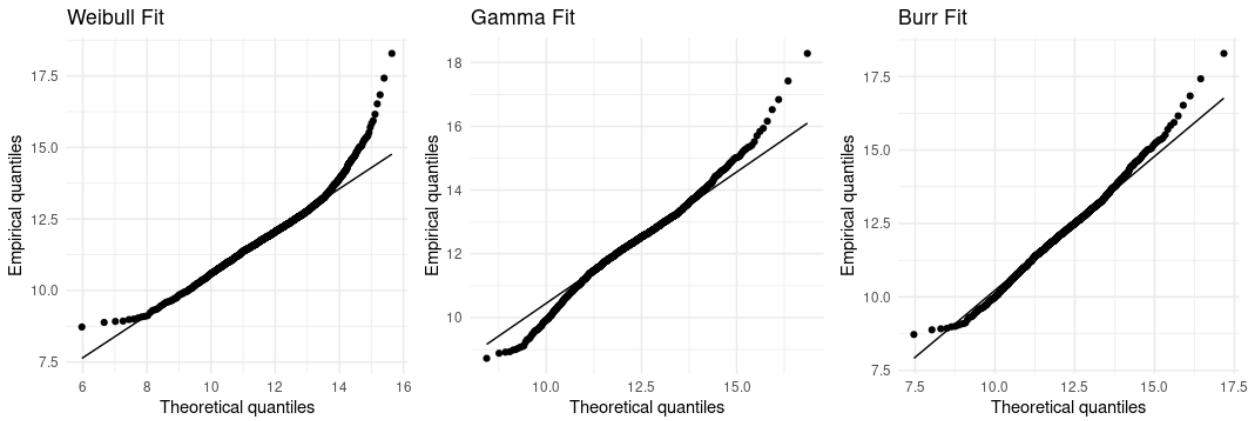

Figure 10: Comparison of fit on test data

The chart above gives a good visual representation and indicates that the Gamma and Burr distribution accurately capture the dynamics of the test claim severity data, by capturing the quantiles at the center of the distribution as well as the quantiles at the tails of the distributions.

4.6.2 Comparison by the Kolmogorov Smirnov test

The Kolmogorov Smirnov test is applied to test whether the quantiles of the test claim severity data came from the same distribution as the quantiles generated by either the Burr, Gamma or Weibull distribution. The test results are shown in the table below;

	Distribution	Statistic	P-Value
1	Gamma	0.05852674	0.06710870
2	Weibull	0.07164480	0.01235575
3	Burr	0.02522704	0.91075034

Figure 11: Kolmogorov-Smirnov statistics

From the test statistics and the P-values generated, we rejected the null-hypothesis and conclude that the quantile distribution of the test claim severity data is not significantly different from the quantile distribution of the fitted Burr and Gamma distribution at the 5% level.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

In this study, we examined the distribution of the log-claims severity from a portfolio of liability insurance policies using data ranging from the year 1998 to 2020. The distributions used to model the data were chosen from the Gamma family of distributions are Exponential, Gamma, Weibull, Pareto, and Burr distribution. The results indicated that the Burr and Gamma distribution provide the best fit to the data with the Gamma distribution providing a good fit mostly within the center of the data, and the Burr distribution providing a good fir to the tails of the data. The results are interesting since, the Burr is regarded a heavy-tailed distribution, while the Gamma distribution is regarded a light-tailed distribution, yet both give a good fit to liability claim severity data. These results are useful to academic researches and analysts working within liability insurance sector.

5.2 Recommendations

Based on the findings from this study, this paper presents the following recommendations to analysts and academic researches:

Analysts working with liability claim severity should analyze data in a two-step method, which involves separating the data into the center of the distribution of data, and the extreme tails of the data. The analyst should then proceed to fit a light-tailed distribution to the center of the data distribution, and fit a heavy tailed distribution to the tails of the data.

Future research should concentrate on using other families of distributions such as the Beta family, and extreme value theory distributions in modelling claim severity data.

REFERENCES

Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?. Insurance: Mathematics and Economics, 51(2), 239-248.

Karobia, R. J. (2015). Modelling extreme claims using generalised pareto distributions family in an insurance company (Doctoral dissertation, University of Nairobi).

Packová, V., & Brebera, D. (2015). Loss distributions in insurance risk management. Recent Advances on Economics and Business Administration, 17-22.

Das, J., & Nath, D. C. (2016). Burr distribution as an actuarial risk model and the computation of some of its actuarial quantities related to the probability of ruin. Journal of mathematical finance, 6(1), 213-231.

Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). Modeling the frequency and severity of auto insurance claims using statistical distributions.

Ng'elechei, J. K., Chelule, J. C., Orango, H. I., & Anapapa, A. O. (2020). Modeling Frequency and Severity of Insurance Claims in an Insurance Portfolio. American Journal of Applied Mathematics and Statistics, 8(3), 103-111.

Okindo, N. F. (2021). Oq-o2-2o21 (Doctoral dissertation, Strathmore University).

APPENDIX

```
# setting directory and loading necessary global data, files and
functions
setwd("/home/nc-workforce/Documents/Desktop/BACS 409")
source("global data.R")

require(pacman)

## Loading required package: pacman

p_load(MASS,
       dplyr,
       ggplot2,
       PerformanceAnalytics,
       plotly,
       lubridate,
       readxl,
       stringr,
       stringi,
       gridExtra,
       actuar,
       fitdistrplus)

# loading data
df <- read_xlsx("data/claims2.xlsx") %>%
  dplyr::select(REV_IRA_CLASS, REV_ACC_DATE, amount) %>%
  mutate(
    CLAIM.AMOUNT = ifelse(amount <= 1000, NA, amount),
    CLAIM.DATE = ymd("1900-01-01") + REV_ACC_DATE,
    Year = year(CLAIM.DATE),
    Month = month(CLAIM.DATE)
  ) %>%
  dplyr::filter(REV_IRA_CLASS == "Liability") %>%
  select(CLAIM.DATE, CLAIM.AMOUNT, Year, Month) %>%
  na.omit() %>%
  arrange(CLAIM.DATE)

# boxplots for detecting outliers
b2 <- df |>
  ggplot()+
  geom_boxplot(aes(x = as.factor(Year), y = `CLAIM.AMOUNT`)) +
  geom_hline(aes(yintercept = 1e8), col = "red", lwd = .7)+
  theme_minimal()+
  labs(title = "Annual Claim severity distribution", x = "Year", y =
"Claim amount")
b2

# omitting outliers i.e. points above 100,000,000
outliers <- which(df$CLAIM.AMOUNT > 1e8)
df <- df[-outliers, ]
```

```r
# constructing a training and testing dataset, using randomization

# setting seed for reproducibility
set.seed(82)
ttr <- sample(nrow(df), (.85*nrow(df)))
df_train <- df[ttr,]
df_test <- df[-ttr, ]

train_claim <- df_train$CLAIM.AMOUNT
test_claim <- df_test$CLAIM.AMOUNT

# Claim Frequency modelling
-----------------------------------------------

cfreq <- df %>%
  count(Year, Month) %>%
  arrange(Year, Month)
cfreq$Period <- str_c(cfreq$Year, "-", cfreq$Month)


# the line plot of monthly-claims frequency
ldf <- cfreq |>
  mutate(Date = ymd(str_c(Period,'-25'))) |>
  select(Date, n)
pl <- ggplot(data = ldf) +
  geom_line(aes(x = Date, y = n))+
  labs(title = "Monthly claim frequency", x = "Date", y = "Claim
Frequency") +
  theme_minimal()

# Claim severity modelling
-------------------------------------------------

# plotting the densitites for the claim severity
p1 <- ggplot(data = df)+
  geom_density(aes(x = (`CLAIM.AMOUNT`)),
               outline.type = "upper")+
  labs(title = "Density plot for claim severity",
       x= "Claim Severity", y = "Density")+
  theme_minimal()

# plotting the densitites for the log-claim severity
p2 <- ggplot(data = df)+
  geom_density(aes(x = log(`CLAIM.AMOUNT`)),
               outline.type = "upper")+
  labs(title = "Density plot for log claim severity",
       x= "Claim Severity", y = "Density")+
  theme_minimal()
```

```r
grid.arrange(p1, p2, nrow = 1, ncol = 2)


# Model fitting
-------------------------------------------------------------

# fitting the exponential
exp_fit <- fitdist(log(train_claim),
                   distr = "exp",
                   method = "mle")
summary(exp_fit)
```

```
## Fitting of the distribution ' exp ' by maximum likelihood
## Parameters :
##        estimate  Std. Error
## rate 0.08192395 0.000669633
## Loglikelihood:  -52399.89   AIC:  104801.8   BIC:  104809.4
```

```r
# fitting the gamma
gamma_fit <- fitdist(log(train_claim),
                     distr = "gamma",
                     method = "mle")
summary(gamma_fit)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape 91.003220 1.05018997
## rate   7.455352 0.08627271
## Loglikelihood:  -24866.33   AIC:  49736.65   BIC:  49751.88
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.9972541
## rate  0.9972541 1.0000000
```

```r
# fitting the weibull
weibull_fit <- fitdist(log(train_claim),
                       distr = "weibull",
                       method = "mle")
summary(weibull_fit)
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## shape  9.99203 0.05707005
## scale 12.76716 0.01105460
## Loglikelihood:  -25416.25   AIC:  50836.5   BIC:  50851.72
## Correlation matrix:
##           shape     scale
```

```
## shape 1.0000000 0.3275878
## scale 0.3275878 1.0000000
```

```
# fitting the pareto
pareto_fit <- fitdist(log(train_claim),
                      distr = "pareto",
                      method = "mle",
                      start = list(shape = 102115,
                                   scale = 468))
summary(pareto_fit)
```

```
## Fitting of the distribution ' pareto ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## shape  7050.712   90.58317
## scale 85611.363  822.76250
## Loglikelihood:  -52401.14   AIC:  104806.3   BIC:  104821.5
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.7706746
## scale 0.7706746 1.0000000
```

```
# fitting the pareto
burr_fit <- fitdist(log(train_claim),
                    distr = "burr",
                    method = "mle",
                    start = list(shape1 = 2.5,
                                 shape2 = 2.5,
                                 scale = .5))
summary(burr_fit)
```

```
## Fitting of the distribution ' burr ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape1  1.717064 0.06128418
## shape2 15.043711 0.16732884
## scale  12.809169 0.04534124
## Loglikelihood:  -24458.46   AIC:  48922.91   BIC:  48945.75
## Correlation matrix:
##            shape1     shape2      scale
## shape1  1.0000000 -0.7967202  0.9731561
## shape2 -0.7967202  1.0000000 -0.8050158
## scale   0.9731561 -0.8050158  1.0000000
```

```
# Comparison
-------------------------------------------------------------


# the graphical way: PP plot
grid.arrange(
  ppcomp_clone(exp_fit, main = "Exponential distribution", xlogscale =
```

```
F, ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  ppcomp_clone(gamma_fit, main = "Gamma distribution", xlogscale = F,
ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  ppcomp_clone(weibull_fit, main = "Weibull distribution", xlogscale =
F, ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  ppcomp_clone(pareto_fit, main = "Pareto distribution", xlogscale =
F, ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  ppcomp_clone(burr_fit, main = "Burr distribution", xlogscale = F,
ylogscale = F, plotstyle = "ggplot", addlegend = FALSE, ylegend =
NULL)+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(show.legend = F),
  nrow = 2
)



# QQ plots
grid.arrange(
  qqcomp_clone(exp_fit, main = "Exponential distribution", xlogscale =
F, ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  qqcomp_clone(gamma_fit, main = "Gamma distribution", xlogscale = F,
ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  qqcomp_clone(weibull_fit, main = "Weibull distribution", xlogscale =
F, ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(),
  qqcomp_clone(pareto_fit, main = "Pareto distribution", xlogscale =
F, ylogscale = F, plotstyle = "ggplot")+
    theme_minimal()+
    theme(legend.position="none")+
```

```
    geom_jitter(),
  qqcomp_clone(burr_fit, main = "Burr distribution", xlogscale = F,
ylogscale = F, plotstyle = "ggplot", addlegend = FALSE, ylegend =
NULL)+
    theme_minimal()+
    theme(legend.position="none")+
    geom_jitter(show.legend = F),
  nrow = 2
)



# goodness-of-fit statistics
gf <- gofstat(list(exp_fit, gamma_fit, weibull_fit, pareto_fit,
burr_fit),
              fitnames=c("Exponential", "Gamma", "Weibull", "Pareto",
"Burr"))

# Using the models fitted on the test data
--------------------------------

# we will only deal, with GAMMA, WEIBULL, BURR
test_claims <- (df_test$CLAIM.AMOUNT)


# Quantile estimation
-----------------------------------------------------

# generating the quantiles
prob_quantile <- seq(from = .01, to = 1, by = .001)

# extracting quantiles from fitted distributions
eq <- quantile(exp_fit,
               probs = prob_quantile)
gq <- quantile(gamma_fit,
               probs = prob_quantile)
wq <- quantile(weibull_fit,
               probs = prob_quantile)
pq <- quantile(pareto_fit,
               probs = prob_quantile)
bq <- quantile(burr_fit,
               probs = prob_quantile)

# merging the extracted quantiles
q_df <- data.frame(
  prob = prob_quantile,
  exponential_quantile = as.numeric(t(eq$quantiles)),
  gamma_quantile = as.numeric(t(gq$quantiles)),
  weibull_quantile = as.numeric(t(wq$quantiles)),
  pareto_quantile = as.numeric(t(pq$quantiles)),
  burr_quantile = as.numeric(t(bq$quantiles)),
```

```
  true_quantile = stats::quantile(log(test_claims), probs =
prob_quantile)
) |>
  na.omit()

# plotting the quantiles
cols <- c("True distribution" = "black",
          "Weibull fit" = "green",
          "Gamma fit" = "maroon",
          "Burr fit" = "blue")
ggplot(data = q_df)+
  geom_line(aes(x = prob, y = true_quantile, col = "True
distribution"))+
  geom_line(aes(x = prob, y = weibull_quantile, col = "Weibull fit"))+
  geom_line(aes(x = prob, y = gamma_quantile, col = "Gamma fit"))+
  geom_line(aes(x = prob, y = burr_quantile, col = "Burr fit"))+
  scale_x_continuous(labels = scales::percent_format())+
  scale_color_manual(values = cols)+
  theme_minimal()+
  labs(title = "A comparison of quantiles", x = "Quantiles", y = "Log
claim severity")



# using QQ plots for the extracted quantiles vs true distribution
quantile
t1 <- ggplot(data = q_df, aes(sample = true_quantile))+
  stat_qq(distribution = qweibull, dparams = weibull_fit$estimate)+
  stat_qq_line(distribution = qweibull, dparams =
weibull_fit$estimate)+
  labs(title = "Weibull Fit", x = "Theoretical quantiles", y =
"Empirical quantiles")+
  theme_minimal()

t2 <- ggplot(data = q_df, aes(sample = true_quantile))+
  stat_qq(distribution = qgamma, dparams = gamma_fit$estimate)+
  stat_qq_line(distribution = qgamma, dparams = gamma_fit$estimate)+
  labs(title = "Gamma Fit", x = "Theoretical quantiles", y =
"Empirical quantiles")+
  theme_minimal()


t3 <- ggplot(data = q_df, aes(sample = true_quantile))+
  stat_qq(distribution = qburr, dparams = burr_fit$estimate)+
  stat_qq_line(distribution = qburr, dparams = burr_fit$estimate)+
  labs(title = "Burr Fit", x = "Theoretical quantiles", y = "Empirical
quantiles")+
  theme_minimal()
```

```r
grid.arrange(t1, t2, t3,
             nrow = 1)


# Formal statistical test for comparison of test data quantiles and
fitted quantiles
kg <- ks.test(x = q_df$true_quantile,
              y = q_df$gamma_quantile)

kw <- ks.test(x = q_df$true_quantile,
              y = q_df$weibull_quantile)

kb <- ks.test(x = q_df$true_quantile,
              y = q_df$burr_quantile)

ks_tbl <- rbind(
  broom::tidy(kg),
  broom::tidy(kw),
  broom::tidy(kb)
) |>
  mutate(Distribution = c("Gamma", "Weibull", "Burr")) |>
  select(Distribution, statistic, p.value)
colnames(ks_tbl) <- c("Distribution", "Statistic", "P-Value")


# End of analysis
------------------------------------------------------------
```