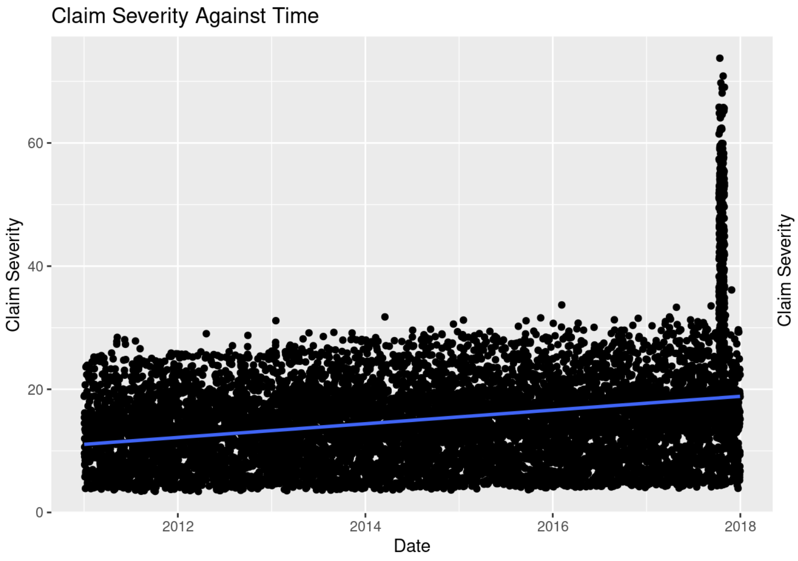
Group members: Amy Fan, Mona Kim, Zalma Gallardo

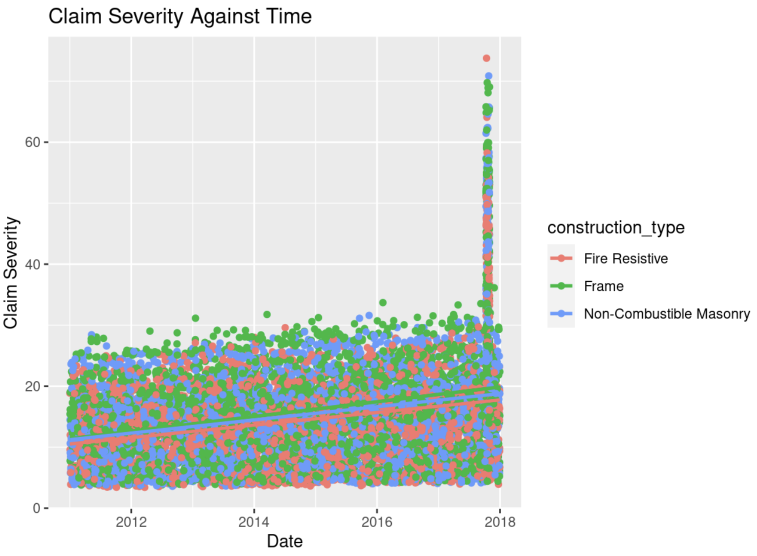
Task 1 - Define Exposure (Background)

Exposures are rating variables used to measure risk or loss (claim amount, frequency, severity) of insuring a certain policyholder. For instance, in Project 1, MVR points, income, location, and car age are used to determine expected risk or loss of a policyholder when insuring car insurance. Likewise, there are many exposures that determine a homeowner's risk, such as agency, construction/roof type, house age, square footage, public protection class, house price, credit score, location, etc. Exposure base is the base level of an exposure used to determine premium value. The rate for all non-base levels is expressed relative to the base level. The premium for the policyholder is the base rate multiplied by all the factor relativities (BaseRate x RatingFactor1 x RatingFactor2 x RatingFactor3..). In Project 1 Task 3, we defined different exposure basis for different exposures of car insurance premium, but exposure basis in insurance is usually the level of an exposure that contributes most to claim amount or insurance company loss because we want the base bin to pay the most premium for contributing to most of the company's loss and other levels to pay premium based on their loss relative to the exposure base loss. Thus, the most common exposure base in homeowner insurance would be the cluster of housing that poses the highest risk of loss(highest claim amount), such as homeowners with older house age, lower public protection class/housing price, and bad credit score. However, there are exceptions if the group that has the most claim amount cannot afford to pay the most insurance premium, such as low-income homeowners. In this case, the exposure base is the level that is the largest and most typical, such as middle-class homeowners.

Task 2 - Identify Catastrophe (EDA)

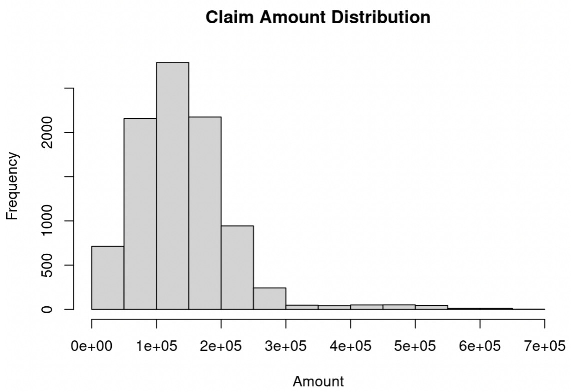
(a)The major catastrophes in the last 7 years have been fire destruction, and ground movements especially those caused by earthquakes, industrial accidents, and other environmental conditions. There is a slight positive association between claim severity and time. However, extremely high claim severity cases appear occasionally and they are not very common. There is a small claim severity increase with time and the catastrophes with major claim severity occurred in late 2017. The catastrophe with the total highest claim severity in 2017 was fire-related.

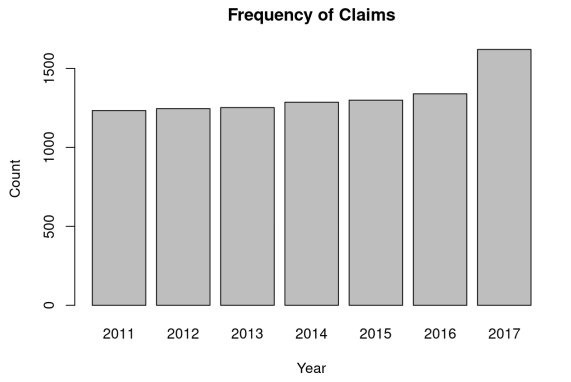




(b) Catastrophe risk management is different from traditional risk management because the former is for low-probability, high-risk events, whereas the latter is for lower-risk events with a higher probability of occurrence (ex. A tornado vs. a broken leg, one is less common and more severe). High deductibles and low premiums characterize catastrophe risk management.

(c) First, we thought of cleaning the policy and claims data to get the claim severity, frequency, and count for each class of all pricing factors. However, we already did that in Task 3 part d, so instead, we visualized and analyzed the claim amount and frequency of claims distributions. The claim severity is calculated by dividing the claim amount per policyholder by the total claim count. Both the claim severity and the total claim amount histogram have similar distributions. The majority of the policyholders have a claim severity between 0 and 20 and the average claim severity is 15. Thus, the average claim severity is in the lower spectrum and the cost of the claims is generally not that costly. The histograms show that low claim severities are more common than high claim severities which means that catastrophic events are less common while lower claim severity events are more common. The premium price for the claim amount can then be set lower because catastrophic events are less common. However, it is important to keep in mind that the frequency of claims has increased from 2011-2017 which means that high claim severity cases could become more common in the future, causing claims to become more costly.





Task 3 - Generalized Linear Model (Calculations And Analysis)

(a) Poisson is used for modeling discrete data that count the occurrence of an event(called *count data*), such as the number of claims filed during a specific timeframe, or express a rate of an event(called *rate data*). Poisson allows us to predict the frequency/rate of an event based on a given response variable, hence would be one of the options for the family distribution for frequency.

Gamma, on the other hand, is helpful when modeling continuous, right-skewed data. Gamma calculates the relativity between the waiting times and the occurrence of an event by using the Poisson process and gives us the prediction of the average severity of an event. Therefore, Gamma is an appropriate option for the family distribution severity.

(b) The Tweedie distribution is a family of continuous probability distributions, and it can have a mixture of zeros and positive continuous data values. Let Y be a response variable, and *p* be a parameter that takes control of the variance of the Tweedie distribution. If $1<p<2$, then the distribution becomes a Poisson which is continuous for $Y>0$, with a positive mass at $Y=0$; if $p>2$, then the distribution becomes a gamma which is continuous for $Y>0$. Note that Tweedie parameters can be converted into Poisson and gamma parameters and vice versa. Since the Tweedie distribution can handle both count data and continuous data, it is the ideal option to describe the pure premium, which is a product of frequency and severity.

(c) Chi-Square Test is one option to measure the correlations between categorical variables. First, we generate a dummy variable for each pricing factor and the claim occurrence per policy number. Then, the chisq.test() function in R returns a report that contains a chi-squared value and a p-value. From the observation of the reports generated for each pricing factor, we have found the following factors have less significant associations with claim occurrence:

- Agency: all
- County: Sonoma
- Construction year: 1986 to 1999
- Roof type: Shingle, Tile
- House value: all
- Public protection class: all

As the associations between claims and these factors are relatively less strong, these factors can work as “offsets” in the later analysis to correct for exposures to any unexpected risks in different periods of observation, for example, a widespread wildfire in 2017 in this case.

(d) (see .xlsx files)

Task 4 - Calculate Premium and Give Business Solutions (Final Proposal)

a) Pure premium base class is 5000 = frequency x severity = baserate x ratingfactor1 x ratingfactor2 …(Project 1 formula). Task 3 part d calculates the pure premium relatives for each pricing factor class relative to the base class by doing frequency relatives of each pricing factor x severity relatives of each pricing factor. Thus, for each policy, 5000 x (agency pure premium relativity) x (county pure premium relativity) x (construction type pure premium relativity) x (roof type pure premium relativity) x (construction year pure premium relativity) x (square footage pure premium relativity) x (public protection cause pure premium relativity) x (house value pure premium relativity). In R, I would first need to drop the severity and frequency relative columns in Pure Premium Relativities.xlsx by only selecting the pure premium column for each table(table$Pure Premium Relativity) and change the name of each pricing factor to pure premium column using the rename function(i.e. rename(table, Pure Premium Relativity = Pure Premium Relativity of Agency)). I would then need to join each table in Pure Premium Relatives to Exhibit 1 policy data of General Analytics data by using the merge function multiple times for each table, so I would have the pure premium relatives for each policy(i.e merge(premiums, table, by = 'Agency')). Then, I can just do 5000 * (data$Pure Premium Relativity of Agency) * (data$Pure Premium Relativity of County)*...*(data$Pure Premium Relativity of House Value)

b) Housing damage worsens the longer it is left unattended, so improving the speed and efficiency of insurance companies responding to the damage is a key strategy. One strategy is to create a faster, more efficient claims process that reduces the cost of claims by addressing inefficiencies and using people who have a reputation for responsiveness. The second strategy is to update technology since many insurance companies lose money due to outdated technology. For instance, instead of relying solely on agents to manually collect and input data, insurers can offer multichannel customer experience that allows customers to serve themselves, such as a robust claims processing system. This improves the speed and accuracy of liability decisions.