Group members: Amy Fan, Zalma Gallardo, Mona Kim, Charisse Liu

Task 1:

a)BEAR Auto Insurance has failed to provide any policyholder data by your deadline. You only have claims data, so you must do your best as a consultant. Using the given claims data, calculate the amount of premiums required per individual to obtain an 80% loss ratio

Sources:

Data: Policyholder Claim Experience Data

Loss ratio concept from Lecture 2 slide 4: Claims/Premiums = (Losses/Gains) = 0.8

Excel skills: https://www.youtube.com/watch?v=7nEyAXiJ304

Example: For Policyholder 1, 4461/Premiums = 0.8. -> 4461/0.8 = Premiums. Must apply this formula to every data point using Excel skills from lecture 3

https://www.youtube.com/watch?v=7nEyAXiJ304

Final code: = B5:B8165/0.8. Click Control/Shift/Enter to enter values

b)After calculating the monthly premium per policyholder, explain why charging a flat rate across all policyholder's is not a smart decision. Explain why rate segmentation is necessary to keep a balanced risk pool (In other words, high-risk customers must pay more than low-risk customers on the aggregate).

Source: Lecture 2: Slides 14-17 Adverse Selection and Risk Pooling

The claims data shows 2 types of policyholders: one with non-zero claims/premiums and another one with zero claims/premiums. Policyholders with non-zero claims/premiums are high-risk(damage car more; require insurance to pay more to cover damage) whereas policyholders with zero claims/premiums are low-risk(little to no damage, so no insurance payment). Assuming we have these 2 types of policyholders, if we charge a flat rate across all policyholders, then low-risk policyholders will most likely opt out of insurance because they don't want to pay more for insurance than they need. This is due to the fact that many low-risk policyholders rarely use insurance. Many low-risk policyholders have a premium of 0, which means that they barely need to pay insurance. Thus, the remaining policyholders in the insurance pool are high-risk, which leads to higher premiums(higher insurance prices), which then drives more people to opt out. This increases the loss ratio from the loss of profit coming from low-risk policyholders. Therefore, rate segmentation is needed to reduce this loss ratio by preventing low-risk policyholders from opting out of insurance if they think they are paying what they deserve. Since low-risk policy makers now pay less, insurance companies need to charge high-risk policy makers more in order to still make profit.

Task 2 - Univariate Factor Research

BEAR Auto Insurance has collected data about its potential customers and has sent all of that data to you. It is your task to derive some meaning from the data while considering the business implications.

(a) Clean the given data using the Excel functions covered in lecture 2.

Control H, replace _Z with empty string, replace < with empty string

(b) Conduct exploratory data analysis and derive what data is relevant, and what data is irrelevant for pricing insurance premiums for individuals. Demonstrate visually for each of your factors why they are relevant.

Sources: Policyholder information data

https://medium.com/swlh/modeling-insurance-claim-frequency-a776f3bf41dc

https://www.youtube.com/watch?v=3hxk6xk3t2A

Claim count: We made a bar graph showing the number of zero and non-zero claims.

1. Assign each value in claim count 0 if claim amount is 0 and 1 if claim amount is not 0. Create a new group column that contains these values. The code for doing this is =IFS(I2:I8162 == 0, 0, I2:I8162< >0, 1).
2. Create a count column that represents 1 for each policyholder count
3. Create a pivot table that aggregates counts column by each group

Analysis: The higher the number of non-zero claims, the higher the risk. We could compare Bear Auto's percentage of ones with the average/standard percentage of ones for any car insurance. If Bear Auto's percentage of ones is greater, then it has a higher risk pool than average, meaning that Bear Auto attracts more high-risk people. If the standard percentage of ones is greater, then it has a lower risk pool on average since it attracts more low-risk people.

Claim Amount excluding 0s: We made a histogram distribution of non-zero claim amounts (see "Claim amount excluding 0s"). The distribution is mostly right skewed with few outliers greater than $11302. Most non-zero claims are between 3202 and 5902, so we can compare values of these claims in the histogram to the expected value of premiums for non-zero claims. Setting premium price at expected value induces adverse selection because low-risk policyholders with claims less than expected value will opt out of insurance. Thus, if a policyholder's claims are greater than the expected value of premium, then they should pay more than the expected value of premium. Likewise, if a policyholder's claims are less than the expected value of premium, they should pay less than expected value of the premium. The expected value is just an estimate since the real value depends on the company's profit, risk. Alternatively, we can set different premium rates for different categories of people based on the claim

amount of a particular category relative to the base category(category that contributes highest to claim amount). This technique is rate segmentation, which will be described more in detail in Task 3.
Claim Amount with 0s: The amount of zero claims is greater than the amount of non-zero claims, so these people make up a majority of the policyholders. However, they do not make up a majority of the total claim amount, which is offset by the policyholders with non-zero claims, who make claims that are a larger part of the total claim amount. Therefore, the premium price for policyholders with zero claims should be set separately from policyholders with non-zero claims. Their premium price should be set lower than the 25th percentile claim amount with 0s because policyholders in this group should carry less of the risk burden than non-zero claim policyholders since they are not contributing to the loss of profit that the insurance company is making. Note 25 percentile is just an estimate. The real percentile depends on the company's financial risk and profit. Alternatively, we can set different premium rates for different categories of people based on the claim amount of a particular category relative to the base category(category that contributes highest to claim amount). Rate segmentation will be described more in detail in Task 3. This is the concept of risk-pooling where low-risk policyholders pay less to counter the negative effects of adverse selection.

Claim severity: The claim severity is calculated by dividing the total claim amount per policyholder over the number of claim counts labeled as 1.
Code: I2:I8162/3152 because there are 3152 claims made to Bear Auto Insurance
The histogram for claim severity has a similar distribution as claim amount. The majority of policyholders have a claim severity value between 0 and .48, which means that the average cost of a claim for the majority of people insured at Bear Auto is less than the actual average. Those who have higher severity make claims that are much more expensive than the average claim cost, so this is why the average is a lot higher even though the majority of people do not claim much. The higher the claim severity is, the more expensive a particular policyholder is to the insurance company. Therefore, if we only have claim severity data, we can set the premium price just like how we set premium price for claim amount except that the scaling of values will change.

Task 3 - The Univariate Approach

(a) After you have an idea of how much information the rating variables give you about the risk levels of individuals, select 3-4 of the 7 rating variables and explain how you might set the rate relativities for the bins(levels) using a loss ratio approach. To do this, ask yourself how you can find out how costly each bin is relative to the total losses incurred by the insurer. Try to derive one set of rate relativities for any one of three chosen rating factors.

Rating variables:

1. Since location is a categorical variable, the levels are rural and urban. The category that contributes most to the total claim amount is the base rate since we want policyholders in the base rate category to pay higher rates than other categories. Thus, the base rate category would be urban because urban policyholders have a higher claim amount since more people live in urban environments than rural. The rate formula for the base category is (total claim amount of urban policyholders/ total claim amount of all policyholders). The rating factor of rural policyholders relative to urban policyholders is (total claim amount of rural / total claim amount of urban)*base rate because we think rural policyholders should pay lower rates since they have less claim amounts. Also, since urban policyholders have a larger population, raising their rates will not affect individual policyholders as much as rural policyholders but help reduce adverse selection at the same time. I derived the base rate of urban policyholders on Urban spreadsheet and rating factor of rural policyholders on the Rural spreadsheet on my Excel file using the formulas above.
2. Income: Since income is a numerical variable, each level of income corresponds to a set or bin of income values instead of a single income value. We can define certain bins of income corresponding to low-income, middle class, and high-income. We then calculate the claim amounts of each group. Unlike the location variable, if the low-income category is making the greatest claim amounts, then the base category should be the income class with the greatest number of policyholders instead, which is probably middle class, because it would be harder for low-income people to pay higher premium rates. Ideally, high-income people should pay more but since they are less likely to have a high claim amount. Additionally, raising the rates of the group with largest policyholders will not affect each individual policyholder that much since risk is spread among a large group of policyholders. The base rate formula is the (claim amount of policyholders in income bin with most policyholders/claim amount of all policyholders). The rate of any other income bin relative to the base category is just the (claim amount of that income

bin/claim amount of base bin) * base rate. We repeat this process for other bins by calculating loss relative to base level.

3. MVR points are a numerical variable. An increase in traffic tickets or accidents leads to an increase of MVR points to a person's driving record. Each point has its own bin of MVR point values. The base level is the bin that corresponds to the bin of MVR point values that contributes most to the claim amount, which is likely to be the bin of highest MVR points since MVR and claim amount are positively correlated. If that is the case, then the base rate formula is composed of the highest MVR point claim/ claim amount of all policyholders. The rating factor of the any MVR bin is just the (claim amount of that bin/claim amount of base bin) * base rate. We repeat this process for other bins by calculating the loss of each bin relative to base level. A problem is that MVR points and income level might be correlated so charging low income policyholders more, so we might need to determine separate rates for low-income and not low-income policyholders in the high MVR bins.
4. Car age is also a numerical variable and a car's age corresponds to a specific bin or a set of bins. We then determine the claim amount of each bin and the bin with the greatest claim amount is the base category. Older cars are probably more likely to have a greater claim amount because of its tendency to malfunction and get into accidents. Thus, the base rate formula is the claim amount of bin with the greatest claim amount/ claim amount of all policyholders. The rate of the any MVR bin is just the (claim amount of that bin/claim amount of base bin) * base rate.

(b) Your manager congratulates your work for having selected an intuitive set of factors. However, he points out that he can't deliver these to the client because of the shortcomings in univariate analysis. Research and explain some of the shortcomings of univariate analysis.

> Here is an example of the distortion created with univariate methods when selecting factors for a pricing model. A one-way pure premium analysis may show for a personal auto insurance book of business that older cars have high claims experience relative to newer cars. In reality, however, this analysis is distorted by the fact that older cars tend to be driven by younger drivers—who tend to have high claims experience. The experience for both young drivers and old cars looks unfavorable despite the fact that this may be driven primarily by the youthful driver effect. As a result, we can see that the problem is caused by correlations between rating factors.

While univariate analysis can analyze attributes of a single variable, it fails to describe the relationship between multiple variables simultaneously. Univariate analysis is less comprehensive which means that it only takes one variable into account but in the real world there are numerous variables that contribute to

how data is shaped. Then since univariate analysis takes one variable at a time it makes it challenging to show correlations between multiple variables at the same time. For instance, MVR points track a driver's performance and that score is then used to determine car insurance rates. The more tickets and infractions accumulated the more MVR points a driver has. This analysis is distorted because it could be that some drivers are having a higher MVR score due to struggling with mental health issues, not so much because they are under the influence or due to cell phone usage. There are other factors that contribute to driving performance such as stress or fatigue which are not often taken into account by the univariate analysis. Overall, univariate analyses are a better option when there is a simple, straightforward relationship between a variable.

Task 4 - Rate Segmentation & Business Impacts, Multivariate Approach

Consider the example given in task 3. As a univariate method ignores the correlations between older vehicles and younger drivers, it would count the effect of drivers' age twice. This is called a "double counting" effect caused by the distortion of the relativities among rating variables. This example focuses on only two rating variables; however, in reality, the numbers of rating variables affect the risk potential. As a result, a distribution bias in some—if not all—of the other variables could yield unfair premium rates or a lack of accurate rate segmentations. An alternative method that minimizes such shortcomings of univariate methods is to take a multivariate approach.

Multivariate methods, such as generalized linear models, consider all rating variables simultaneously and accommodate for correlations between rating variables:

1. The multivariate methods withdraw any irrelevant and unsystematic effects of rating variables and capture only the useful ones. This process increases efficiency in analyzing tens or hundreds of variables at once.
2. They provide model diagnostics about the accuracy of results and the suitability of the model.
3. They consider the interdependency between two or more rating variables. Interdependencies occur when one rating variable's effect ranges depending on the levels of other variables. By analyzing interdependencies between variables, the actuary can improve the predictive value of rates.

(Source: Basic Ratemaking(4th Edition) by Geoff Werner)