# Pricing and Severity Analysis

Ryan Gomberg

# Overview

1. Abstract
2. Introduction to the Dataset
3. General Statistics
4. Exposure Effects
5. GLMs Methodology
6. Feature Interaction and Visualization
7. Results and Interpretation
8. Business Implications
9. Conclusion

# Abstract


**Motivation**

- Auto insurance pricing and risk management is contingent on not only understanding claim frequency and severity separately, but also how they interact when used synchronously across policyholders.
- It is important to ascertain how different demographics and differences in driver, vehicle, and claim attributes impact expected severity and potential loss/profit.
- These serve as a basis for ensuring long-term stability in the event of future claims, disasters, or demographic shifts.

**Overview**

- This project aims to analyze auto insurance claims through a synthesized modeling and visualization approach.
  - **Generalized Linear Models (GLMs)** are used to quantify and predict risk between key features and claim information.
  - An **interactive Excel dashboard** is employed to dynamically explore patterns, validate assumptions posed in the GLM model, and identify or verify potential nonlinear behavior.
- Together, they demonstrate how actuarial techniques motivate data-driven pricing and informed decision-making.

# Introduction to the Dataset

**The dataset was generated in R and is my own, not taken from any online sources.**
Our sample contains 5,000 policyholders, each with multiple features:

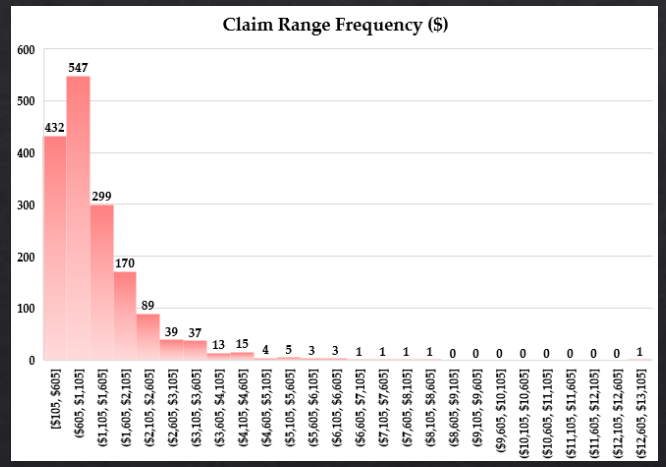
- Age: Between 18 and 90 years.
- Vehicle Age: Between 0 and 15 years.
- Territory: Urban, Suburban, and Rural with assigned probabilities of 0.5, 0.3, and 0.2 in parity with real-world population densities.
- Prior Claims: The amount of claims the policyholder has had previously. Generated with a Poisson(0.4) distribution to reflect insurance claims.
- Exposure: The percentage of a year in which the policy is active. Uniformly distributed between 0.5 to 1, making these exposure periods equally likely.
- Claim Count: The amount of claims a policyholder has currently has. Follows a Poisson distribution whose mean is dependent on the previous features.
- Severity and Average Severity: Lognormally distributed (most claims are concentrated toward lower values)
- Claim Amount: Total dollar amount in claims for each policyholder.

# General Statistics

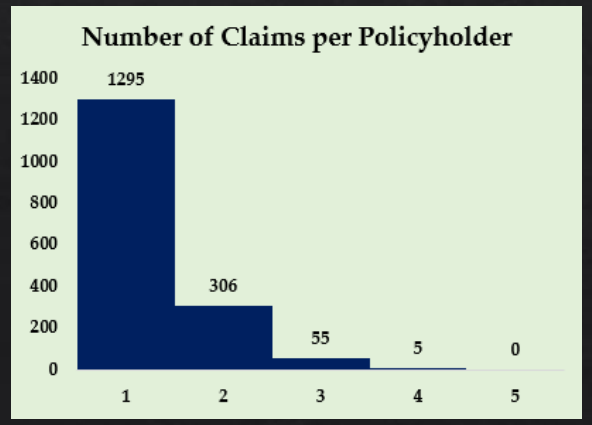
1. 1,661 of the policyholders have at least one claim. We will mainly be focused with this subgroup.
2. The median, mean, and maximum claim amount are $936.60, $1,225.30, and $13,054.00, respectively. This agrees with the right-skewedness of a lognormal distribution.
3. The maximum number of claims for any given policyholder is 4. This is exceptionally rare by construction, with only 5 policyholders having 4 claims.

```
NA's   :3339
```

3,339 policyholders have zero claims.



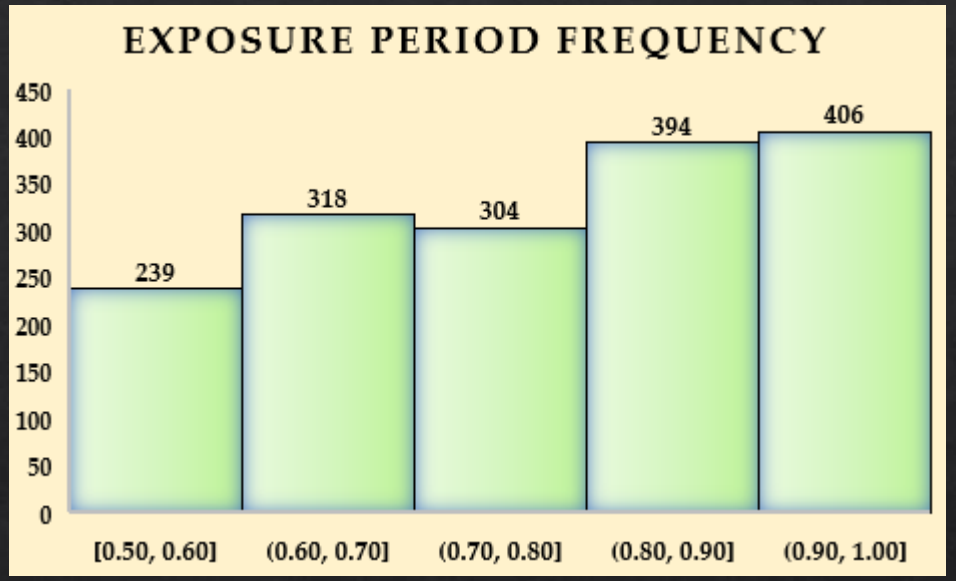

```
 claim_amount
Min.    :  104.6
1st Qu.:  594.4
Median :  936.6
Mean    : 1225.3
3rd Qu.: 1531.0
Max.    :13054.0
```




```
 claim_count
Min.    :0.0000
1st Qu.:0.0000
Median :0.0000
Mean    :0.4184
3rd Qu.:1.0000
Max.    :4.0000
```

# Exposure Effects

- The GLMs will reveal that exposure length is handled as an offset term to explain claim frequency as a rate that is adjusted for varying periods of active policies.
- The exposure period frequency chart shows that the majority of policies tend to higher exposure values. This reflects typical policy lifespans, where most policies are carried out for most of the year, with fewer policies terminated mid-term.
- Looking at claim amount across exposure bands, the observed claim amounts generally increase as exposure bands increase. This pattern reflects greater period of risk: policies exposed for longer periods are open to more opportunity for losses to develop and be recognized.




**% Change in Claim Amount Between Adjacent Exposure Bands**

| | Territory | | |
|---|---|---|---|
| Exposure (in years) | Urban | Suburban | Rural |
| 0.5-0.6 | 100.00% | 100.00% | 100.00% |
| 0.6-0.7 | ↑ 125.65% | ↑ 136.28% | ↑ 143.55% |
| 0.7-0.8 | ↓ 98.02% | ↓ 74.42% | ↑ 110.77% |
| 0.8-0.9 | ↑ 109.57% | ↑ 143.88% | ↑ 130.29% |
| 0.9-1 | ↑ 116.30% | ↑ 109.66% | ↓ 95.75% |

# Methodology

We propose GLMs to individualize model frequency and severity.

## Claim Frequency GLM – Poisson

- Claim frequency is a discrete outcome which measures the number of claims per policy over a fixed period of time.
- This makes a Poisson GLM a great starting point for modeling this type of data because events occur independently and policies have a constant average claim rate over the exposure period.
- Poisson GLMs include an offset term. We purposefully set the offset term to be policy exposure.
- Not all policies are exposed to risk for the same amount of time. To adjust for differences in which policies are active for the full year, or canceled mid-term, including exposure as an offset ensures we are measuring the **claim rate per unit of exposure** as opposed to raw claim counts.

## Claim Severity GLM – TBD

- Severity data is known to be strictly positive, right-skewed, and infrequent large losses (long tail).
- We want to consider multiple distributions, namely the Gamma, Exponential, Lognormal, and Weibull distributions because they offer varying levels of flexibility in modeling the positive and right-skewed nature of our loss behavior.
- In this comparison, we are looking for differences in goodness of fit, capturing tail behavior, and overdispersion. This will be achieved through AIC scores, ANOVA tests, and QQ-plots.
- **Note that because claim severities were generated from a lognormal distribution, there will be a bias in favor of the lognormal regression model and is expected to outperform the other candidates.**

# Methodology – Frequency GLMs

- Although the *p*-value for the Suburban territory is greater than 0.05, given that the other features are in the model, the ANOVA test verifies that each feature is statistically significant.
- The deviance reduction of 331.7 indicates that the features in the model result in a better fit.
- The QQ-plot fits the Poisson model well and has low variance, but appears to fall off as the claim count approaches 0 and then follows a line different from the desired 45-degree line.
- If we choose this model, it follows that living Urban or Suburban region decreases the claim frequency by 12.78% and 7.26%, respectively, compared to living in a Rural region. As a result, this dataset assumes rural drivers file more claims, which can be attributed to longer travel, wildlife, or harsher road and weather conditions.

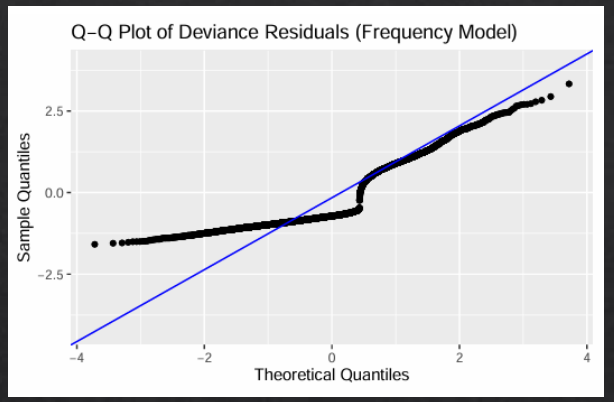
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.827727   0.093299 -19.590  < 2e-16 ***
age                0.015423   0.001086  14.198  < 2e-16 ***
vehicle_age        0.044241   0.005144   8.600  < 2e-16 ***
prior_claims       0.236442   0.030868   7.660 1.86e-14 ***
territorySuburban -0.075415   0.061377  -1.229    0.219
territoryUrban    -0.136705   0.055607  -2.458    0.014 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4828.2  on 4999  degrees of freedom
Residual deviance: 4496.5  on 4994  degrees of freedom
AIC: 8079.2
```

```
           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        4999     4828.2
age         1  199.211       4998     4629.0 < 2.2e-16 ***
vehicle_age 1   71.741       4997     4557.3 < 2.2e-16 ***
prior_claims 1  54.629       4996     4502.6 1.456e-13 ***
territory   2    6.107       4994     4496.5   0.04719 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

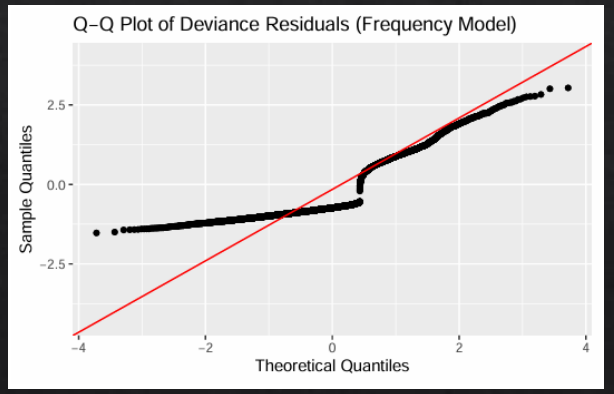
ANOVA results

# Methodology – Frequency GLMs

- Despite an overdispersion coefficient of 0.9, we must raise skepticism and watch out for signs of **zero-inflation**: the model greatly over or under-predicts the probability of having no claims, but accurately models the claim frequency among policyholders with claims.
- A Negative Binomial model was also constructed by similarity to a Poisson model. However, this model performed slightly worse than the Poisson model and exhibited the same symptoms of zero-inflation.
- Since 66.78% of policyholders have no claims, this well justifies the zero-inflation observed here.




Poisson QQ-Plot




Negative Binomial QQ-Plot

# Methodology – Severity GLMs

- Unsurprisingly, the lognormal regression model showed the highest promise in fitting the dataset. Not only does it have the lowest AIC by a significant margin, but it also offers the best convergence to the 45-degree line on the QQ-plot.
- The Gamma GLM comes up in second, albeit from pretty far, but it could still be an acceptable model due to its similarities with a lognormal model. Its AIC steers closer to Weibull and Exponential models, but it tracks the 45-degree line pretty well.
- The Weibull and Exponential distributions are poor choices due to their high AIC values (compared to the others) and divergence on the QQ-plot.

Gamma

AIC: 25405

Lognormal

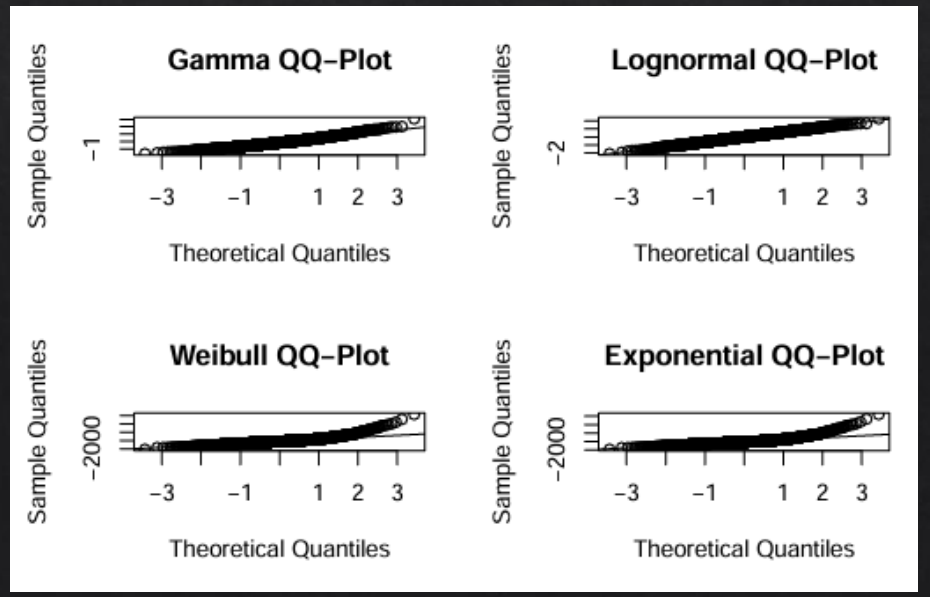
3070.845

Weibull

25565.64

Exponential

26184.03



# Methodology – Severity GLMs

- Using both models, it follows that:
  - Both higher age and vehicle age lead to higher severity.
  - Compared to Rural areas, living in a Suburban area reduces severity by 8.5-12% and living in an Urban area further reduces the severity 13-14% (or 21.5-26% compared to Rural). The effects on territory are moderate, but still worth nothing.
  - The difference between lognormal regression and Gamma GLM gets larger as policyholder age and vehicle age increase. More precisely, when increasing policyholder age by 30 years and vehicle age by 5 years, this difference increases by 65.38% for Rural areas, 78.72% for Suburban areas, and 77.23% for Urban areas.
  - The lognormal regression is more conserved in its approximation, whereas the Gamma GLM will start diverging to larger values as these features increases (exhibited by the QQ-plot).

**(1) Controlling for: Age 30 years, Vehicle Age of 5 years. Comparing over all 3 territories:**

- **Lognormal**: Rural $= e^{6.2938} \approx \$541.21$, Suburban $= e^{6.1679} \approx \$477.18$, Urban $= e^{6.0234} \approx \$412.98$.
- **Gamma GLM**: Rural $= e^{6.3832} \approx \$591.82$, Suburban $= e^{6.2938} \approx \$541.21$, Urban $= e^{6.1437} \approx \$465.77$.

**(2) Controlling for: Age 60 years, Vehicle Age of 10 years. Comparing over all 3 territories:**

- **Lognormal**: Rural $= e^{7.0528} \approx \$1156.09$, Suburban $= e^{6.9269} \approx \$1019.33$, Urban $= e^{6.7824} \approx \$882.18$.
- **Gamma GLM**: Rural $= e^{7.1227} \approx \$1239.79$, Suburban $= e^{7.0333} \approx \$1133.77$, Urban $= e^{6.8832} \approx \$975.74$.

# Methodology – Conclusion

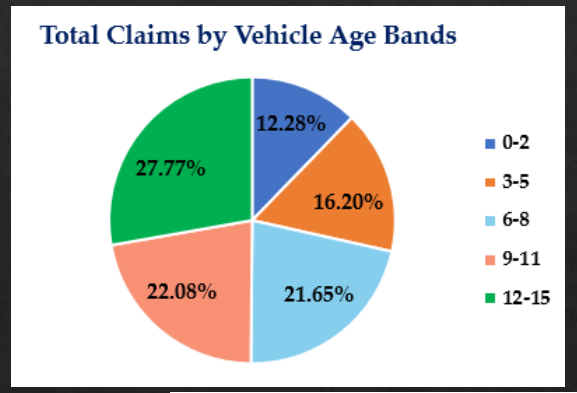
- The GLMs exhibited notable patterns in how the policyholder features influence claim frequency and severity. Individually, the direction of associations is consistent:
  - Claim frequency and severity increase as the policyholder's age increase, holding all other factors constant, thereby making age a strong predictive feature.
  - This also holds for vehicle age. Older vehicles are prone to more mechanical and technical failures, which can lead to more claims.
  - Compared to Rural areas, claim frequency decreases for urban and suburban regions

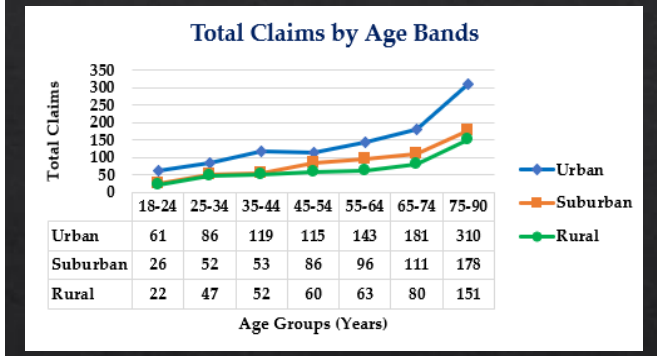
    (Urban < Suburban < Rural).
- Their combined effect on total premiums are mostly consistent:
  - Age and vehicle age increase leads to an increase in total premiums (severity and frequency increase).
  - Since claim frequency and severity move in the same direction, these effects are compounded, and Urban premiums < Suburban premiums < Rural premiums generally.
- The models can only reveal so much – interactive dashboards are crucial in understanding the dataset's behavior between different features and also reinforcing the GLM results.

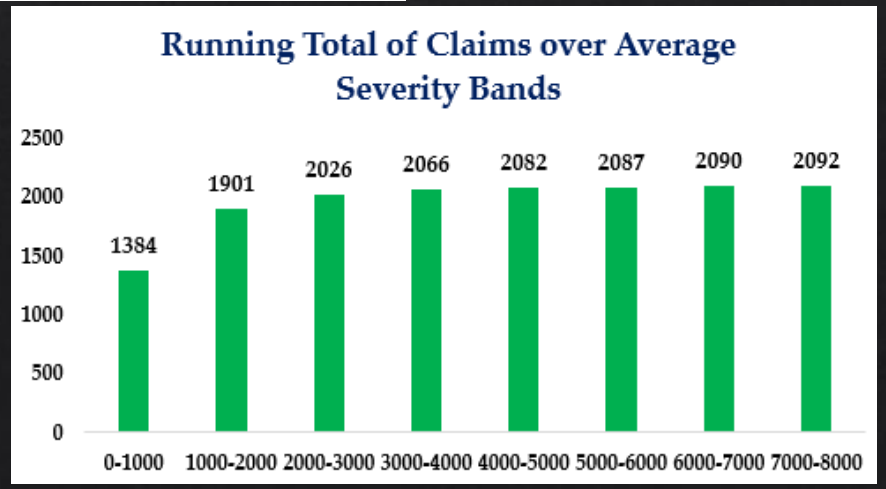
# Feature Interaction and Visualization

While the GLMs offer predictive power, its mileage goes as far as the data that supports it. That is to say, the existing data needs to support what both models have been predicting. This is where Excel dashboards shine.

- These charts are taken from the group of policyholders with claims (for consistency with severity charts).
- The frequency GLM predicts that claim frequency will increase as age and vehicle age increase. This pattern also holds in the data; claim counts almost always increase over each subsequent band.
- Moreover, the severity GLMs predicted that most policyholders would have losses frontloaded between $0 and $3,000, and that most claims would have losses in said range. The running total of claims chart reveals how claim frequency is frontloaded in this severity range.

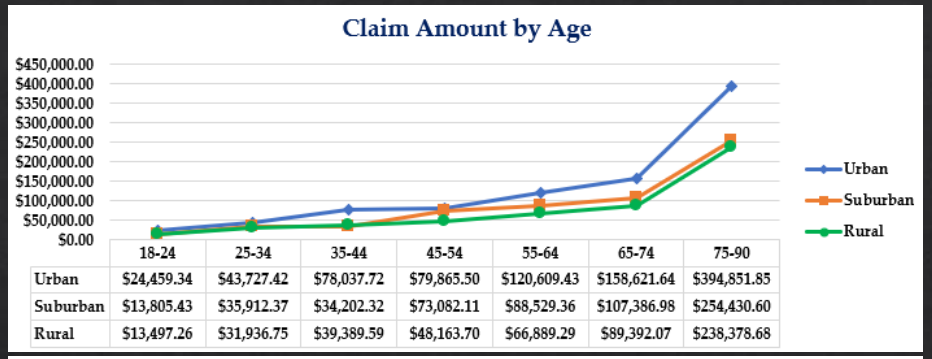


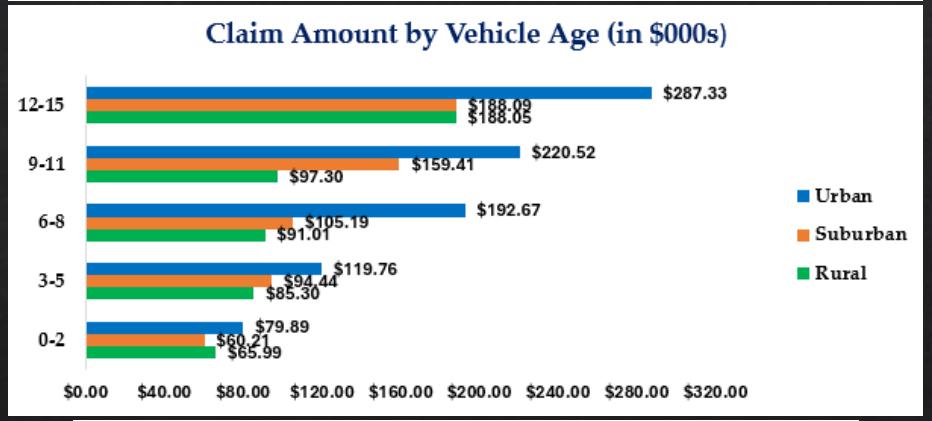



# Feature Interaction and Visualization

- Similarly, the dashboard confirms that the direction of severity association aligns with what is in the dataset.
- The Severity GLM predicted an increase in claim amount as policyholders and their vehicles get older, holding other factors constant. The dashboard confirms this pattern within the dataset; the claim amount mostly increases in subsequent age and vehicle age bands.
- Of the 1,661 policyholders with claims ,the observed average claim amounts mostly increase across subsequent exposure bands, indicating greater time at risk. In the GLM, exposure is treated as an offset, therefore making the model estimate severity on a per-unit exposure basis as opposed to attributing an effect to exposure itself (claim amount in this case).




| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65-74 | 75-90 |
|---|---|---|---|---|---|---|---|
| Urban | $24,459.34 | $43,727.42 | $78,037.72 | $79,865.50 | $120,609.43 | $158,621.64 | $394,851.85 |
| Suburban | $13,805.43 | $35,912.37 | $34,202.32 | $73,082.11 | $88,529.36 | $107,386.98 | $254,430.60 |
| Rural | $13,497.26 | $31,936.75 | $39,389.59 | $48,163.70 | $66,889.29 | $89,392.07 | $238,378.68 |

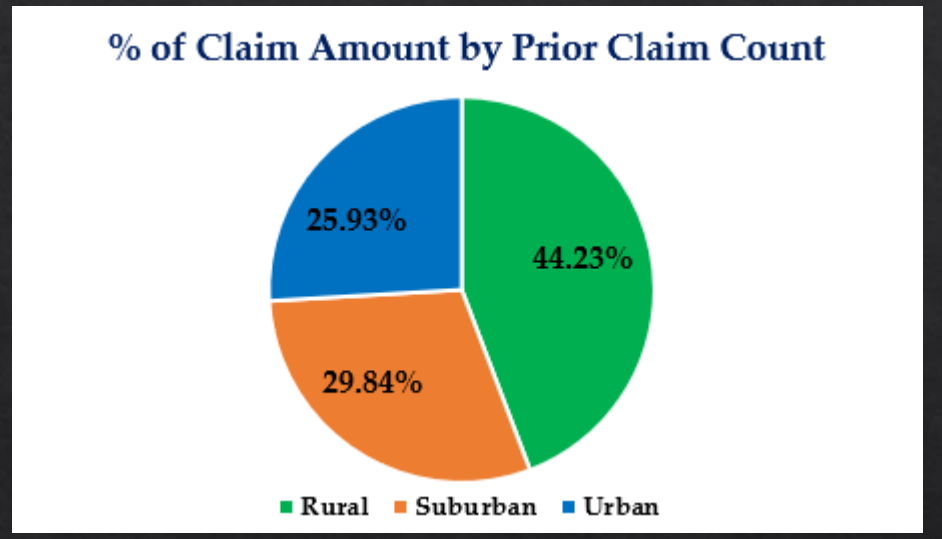


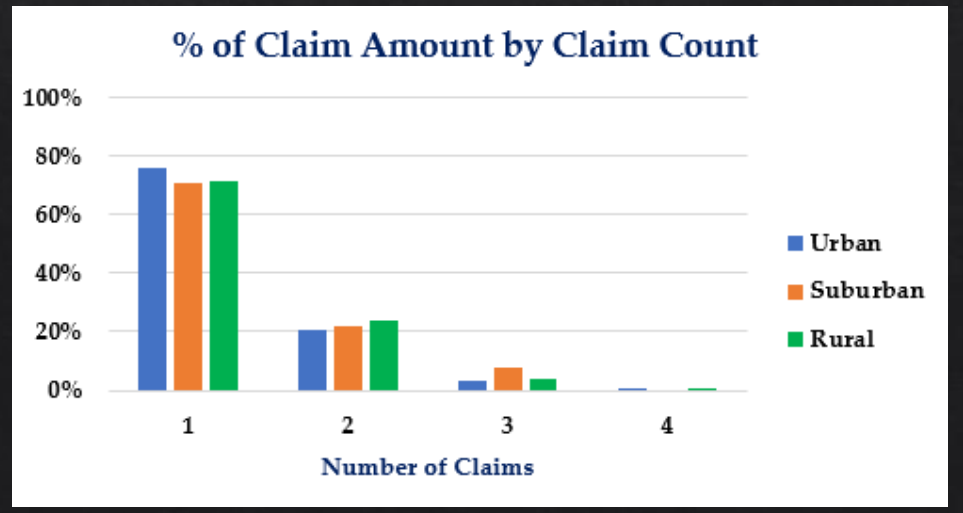

**% Change in Claim Amount Between Adjacent Exposure Bands**

| | Territory | | |
|---|---|---|---|
| Exposure (in years) | Urban | Suburban | Rural |
| 0.5-0.6 | 100.00% | 100.00% | 100.00% |
| 0.6-0.7 | ↑ 125.65% | ↑ 136.28% | ↑ 143.55% |
| 0.7-0.8 | ↓ 98.02% | ↓ 74.42% | ↑ 110.77% |
| 0.8-0.9 | ↑ 109.57% | ↑ 143.88% | ↑ 130.29% |
| 0.9-1 | ↑ 116.30% | ↑ 109.66% | ↓ 95.75% |

# Feature Interaction and Visualization

- So far, the dataset has justified what the GLMs have predicted, thereby establishing a connection between retrospective and prospective modeling.
- It is equally important to identify the features that will have the largest impact on pricing policies.
- For instance, policy prices should be higher for Rural areas as they are more prone to losses of higher magnitude.
- Looking at claim counts, pricing should gravitate towards policyholders who are likely to have at most 2 claims. Policyholders with 3 or 4 claims do not offer substantial contribution towards the total claim amount, due to their general infrequency.



# Results and Interpretations

- The frequency and severity models together describe a consistent and intuitive understanding of how risk varies among policyholders.
- Earlier we proved through the GLMs and visualizations that most premiums have strong dependence on age and vehicle age:
  - As policyholder gets older, their chances for accidents increase due to possible issues in hearing and/or vision.
  - If a policyholder maintains their vehicle for longer, they are more likely to have maintenance and repairs.
- Territory also plays an imperative role. Earlier we concluded that Urban premiums will be the cheapest, then Suburban, followed by Rural as the most expensive:
  - Urban: Fewer accidents and lower speeds = lowest expected total premium.
  - Suburban: Moderate accident frequency and severity = moderate total premium.
  - Rural: Highest accident frequency (higher speeds, wildlife, harsher weather/road conditions) = highest total premium.

# Business Implications

- By laying the groundwork, it is now possible to recommend a pricing structure.
- First, premiums are generally a product of the expected claim frequency and severity.

$$\text{Premium} \propto E[\text{Claims per year}] \times E[\text{Loss per claim}]$$

- Use the Poisson GLM to compute the expected number of claims each year $\hat{\lambda}_i$, and the Lognormal Regression model to compute the expected loss per claim $\hat{\mu}_i$. Then, scaling by exposure, obtain the expected loss

$$\text{Expected Loss}_i = \text{Exposure}_i \times \hat{\lambda}_i \times \hat{\mu}_i$$

- It is also important to understand that lending insurance itself has a cost, and that some policyholders have extreme losses that the insurer must compensate for. For instance, a head-on collision due to poor visibility could result in a major loss for both sides. An additional **policy cost** is typically added to mitigate or cover such situations. Operation and policy costs are added and then put into the expression as a multiplier.

$$\text{Expected Loss}_i = \text{Exposure}_i \times \hat{\lambda}_i \times \hat{\mu}_i \times (1 + \text{Operation Cost} + \text{Policy Cost})$$

# Business Implications

- Suppose a 40-year-old from a rural area wanted to buy an auto insurance policy 25% into the policy year. The to-be insured car was purchased 8 years ago and has had no previous claims on file. If we let the operation cost = 18% and the policy cost = 12%, then
  - The expected number of claims per year is 0.3936.
  - The expected loss per claim is \$610.
  - The insured is exposed to the policy for 75% of the year.
- This policyholder's policy should be priced at a premium of

$$0.3936 \times 610 \times .75 \times (1 + 0.18 + 0.12) \approx \$234.$$

# Conclusion

In this presentation, we have

1. Motivated the importance of modeling and visualization in pricing.
2. Explained the dataset, in terms of size, features, and their construction.
3. Offered basic statistics to reveal the general trend of claims and losses.
4. Emphasized the importance of exposure in a policy and as an offset in a Poisson frequency GLM.
5. Implemented claim frequency and severity GLMs, highlight the strengths and weaknesses in their predictive power.
6. Validated the predictive trend from the GLMs with Excel visualizations.
7. Reasoned the findings using real-world scenarios in each significant feature.
8. Uncovered a pricing structure for premiums to be used on future policyholders.

# Resources

- Pricing Severity Modeling in R (Used for GLMs and exploratory analysis):

  https://ryangomberg.github.io/ryangomberg/PricingSeverityModeling.pdf

- Excel visualization and interactive dashboard used for feature interactions and visualization:

  https://ryangomberg.github.io/ryangomberg/PricingSevDashboards.pdf

  Download the Excel file for personal use here.