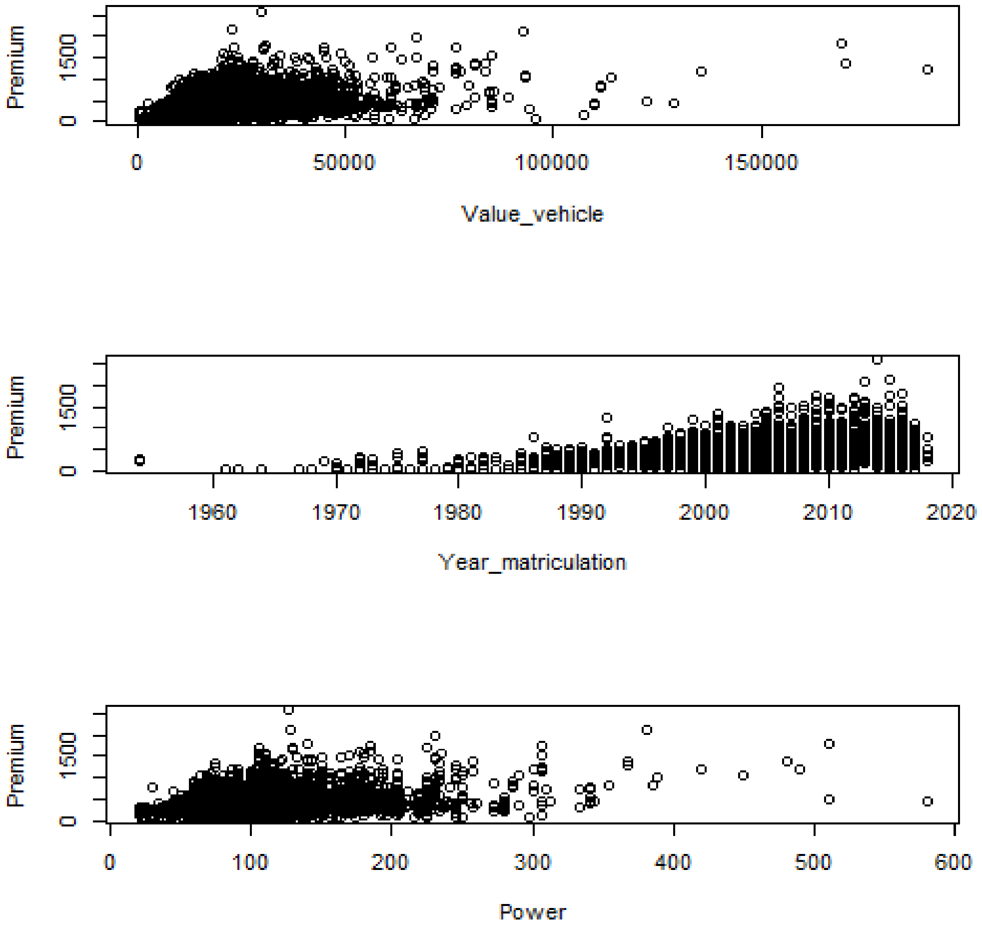
# Driving Data:

## Building Predictive Models for Auto Insurance Premiums

By: Kylie Wilkin

# Problem Statement: Pricing in Non-Life Insurance

- Pricing in non-life insurance: determining the premium the insured pays for risk transfer
  - Significance: insurers can price policies competitively while covering risk effectively
- Calculation is complex because of highly correlated variables, unknown values, non-numerical data, asymmetric distributions, and other external factors
  - Outliers distorting predictions in traditional models
- As we shift to unprecedented shifts, risk categories such as cyber emerge in the era of autonomous vehicles
- This study applies predictive linear and logistic regression models to calculate premiums for auto insurance claims




# Data Collection and Preparation

- Data source: Spanish non-life motor insurance company, published in 2024, records made from 11/2015-12/2019
- Coding platform: RStudio
- Over 31,000 records
- Training set and test split
- Exploratory Data Analysis (EDA): identify patterns, correlations, and variable distributions of predictors to the response variable as premium

# Distinctions of Europe

## Europe

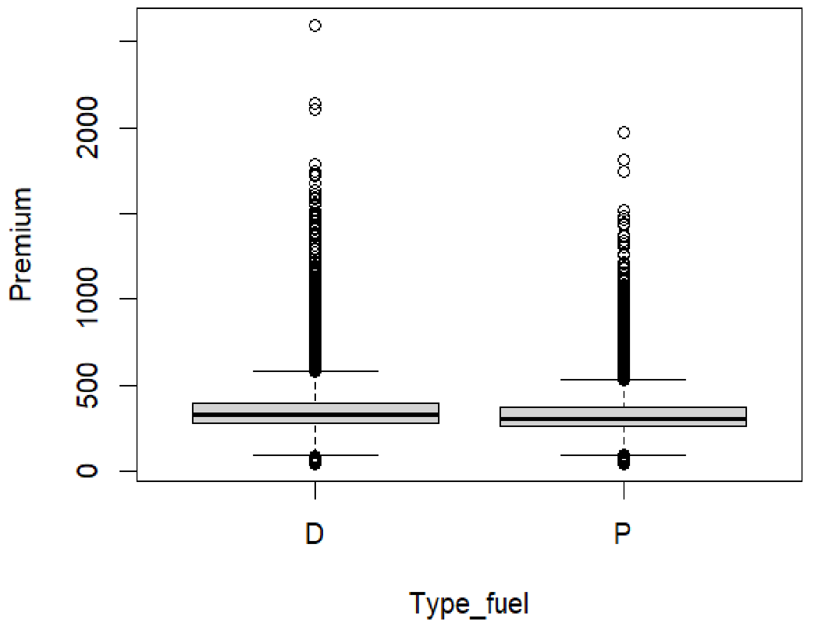

- Driving habits: Urban design favoring public transportation, stricter speed limits
- Regulatory environments: stricter data protection, may rely more on anonymous and vehicle-type data
- Risk types: high prevalence of theft, small city-type cars

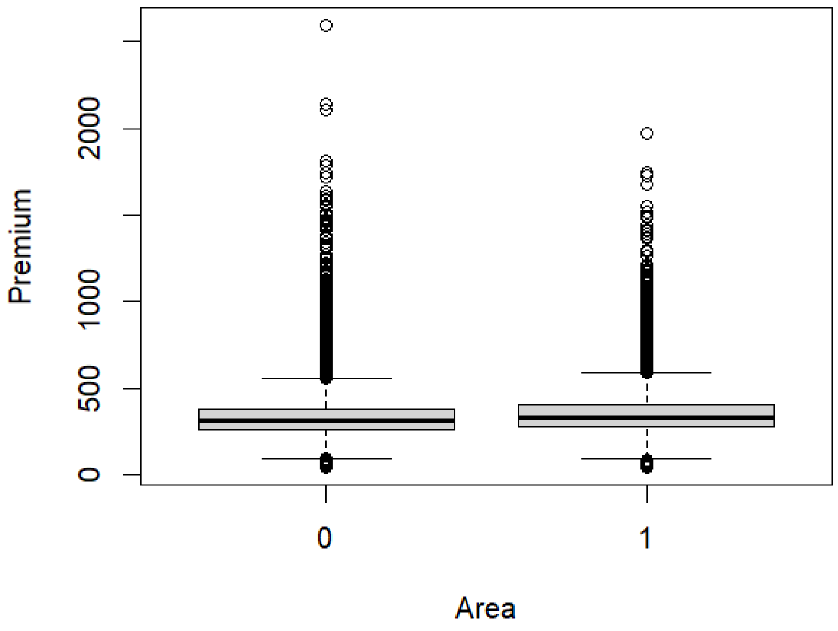
## America


- Driving habits: longer, more frequent highway commutes, higher speed limits
- Regulatory environments: less strict regulations, supports more data-intensive models
- Risk types: severe weather events (like hurricanes), high-power vehicles

# Data: New Variables

- Removing N/As
- Date variables into quantitative predictors (i.e. date of birth into age)
- Removing all the blanks (unknown values) in the date of lapse variable
- Factors: Type of fuel, vehicle type, area of use, whether more drivers are declared, type of payment
- Clean data file: [autopremclean.csv]()



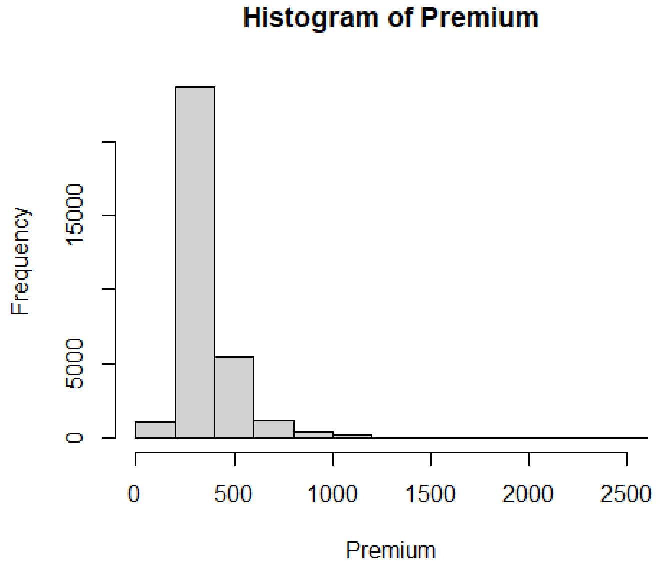



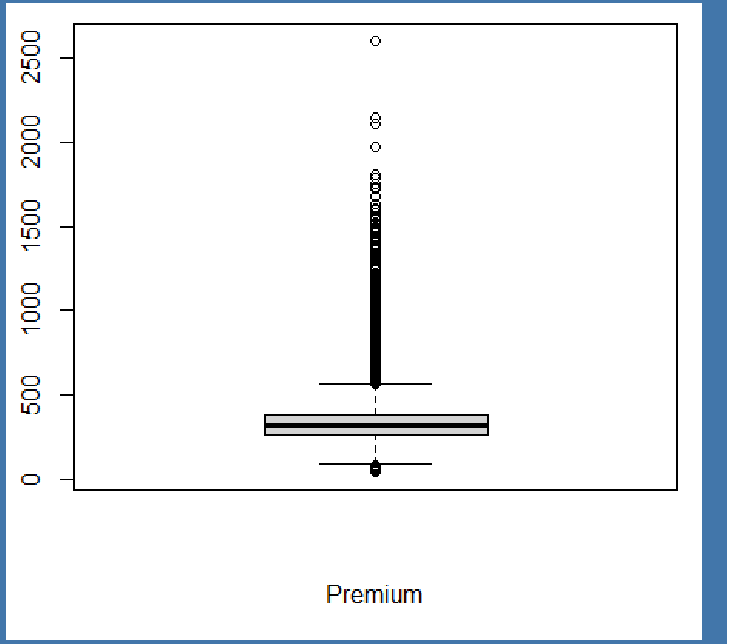

| |
|---|
| Years_since_start_contract |
| Years_since_last_renewal |
| Years_till_next_renewal |
| Years_since_license |
| Years_till_lapse |

# Response Variable: Premium

Distribution:

- Net amount associated with policy during the current year
- Mostly near €500 or less
- Outliers, skewered to right
- Average of €350
- Range of €42- €2597

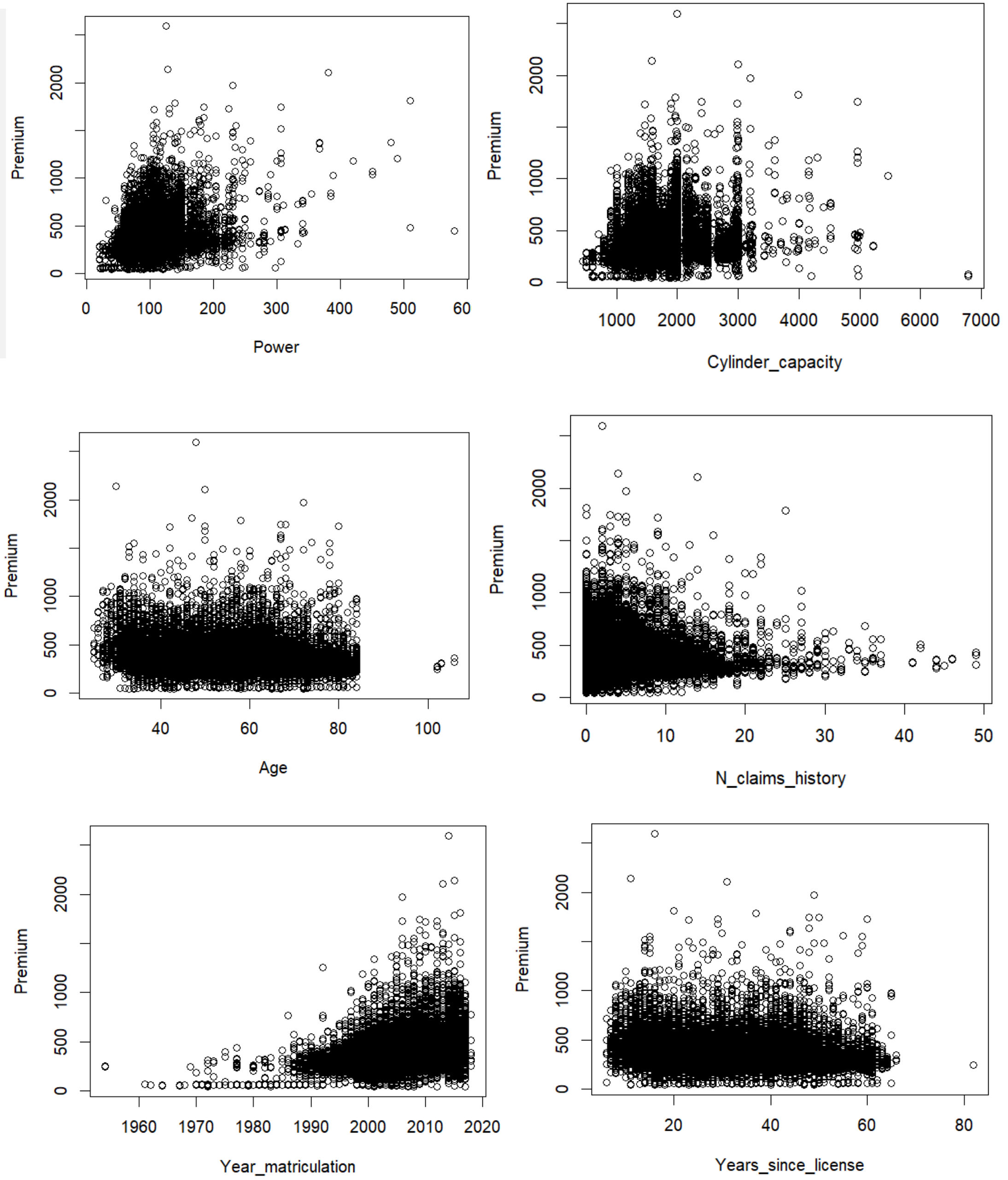
# Predictor Insights

## Vehicle Attributes

- Value of vehicle
- Weight- kilograms
- Length- meters
- Cylinder capacity
- Horsepower
- Type of vehicle- motorbikes, vans, passenger cars, agricultural vehicles
- Type of fuel- gas vs diesel
- Year of registration

## Driver Characteristics

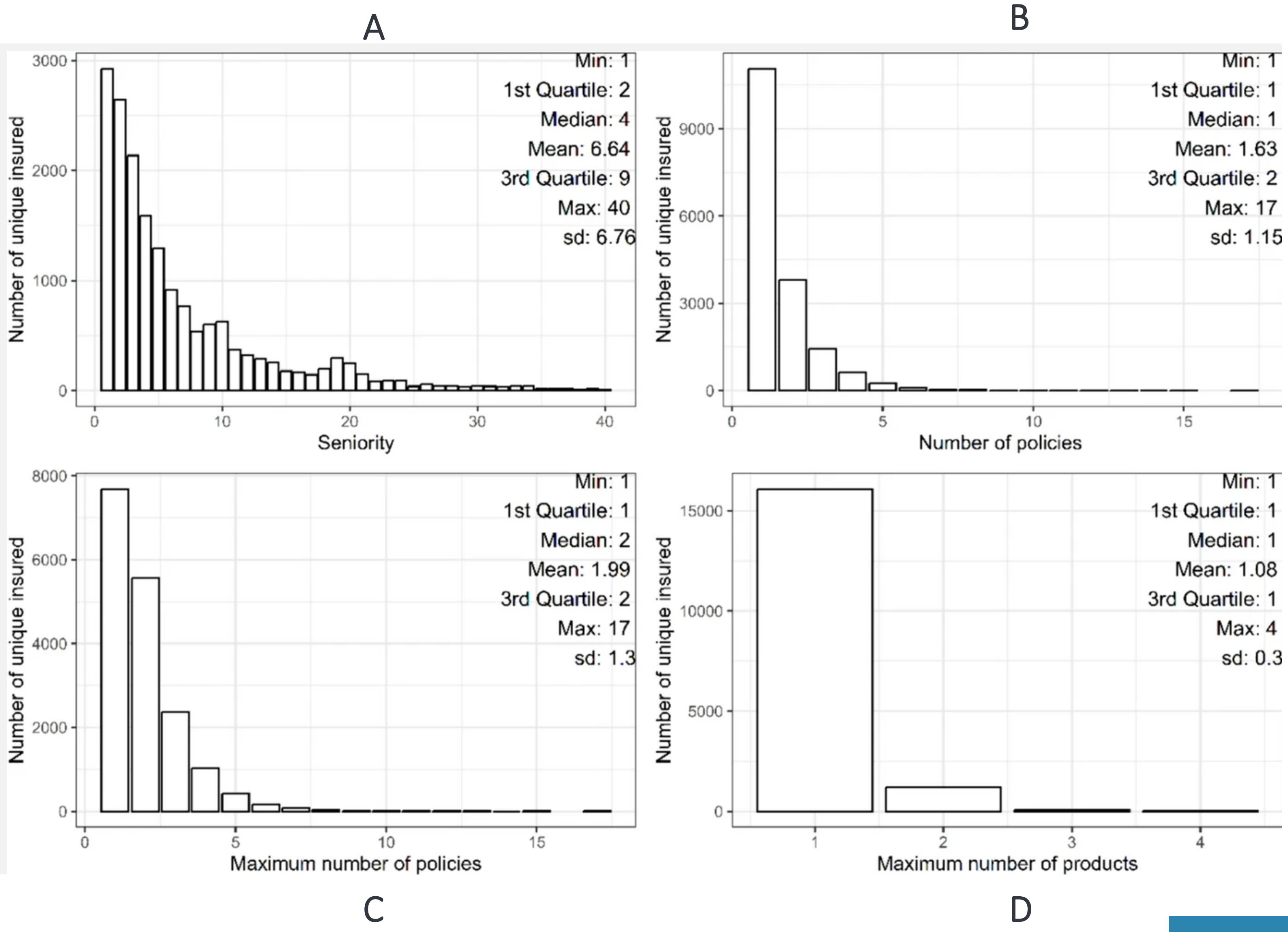
- Age
- Years since license
- Payment method- semiannual vs annual
- Channel the policy was contracted- agent vs brokers
- Policy’s claims frequency history



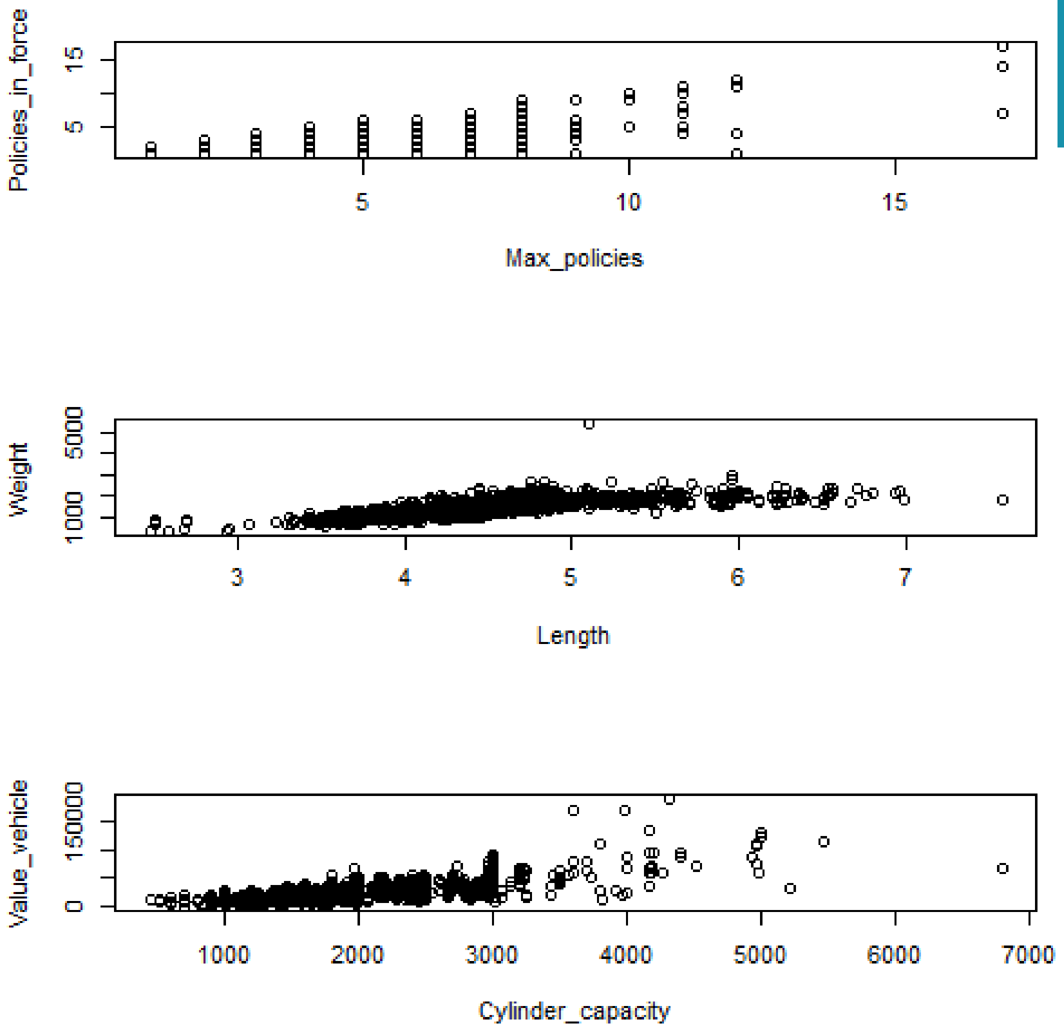
# Predictor Insights

## Policyholder's Affiliation

- The difference between graphs B and C is how much their insurance coverage has decreased over time
- The company offers 4 different products (car, household, commerce, and personal accident) for its policyholders (refer to graph D)
- Lapse and number of years since lapse
    - Number of policies that the customer has canceled including ones that were replaced



# Data: Interaction terms



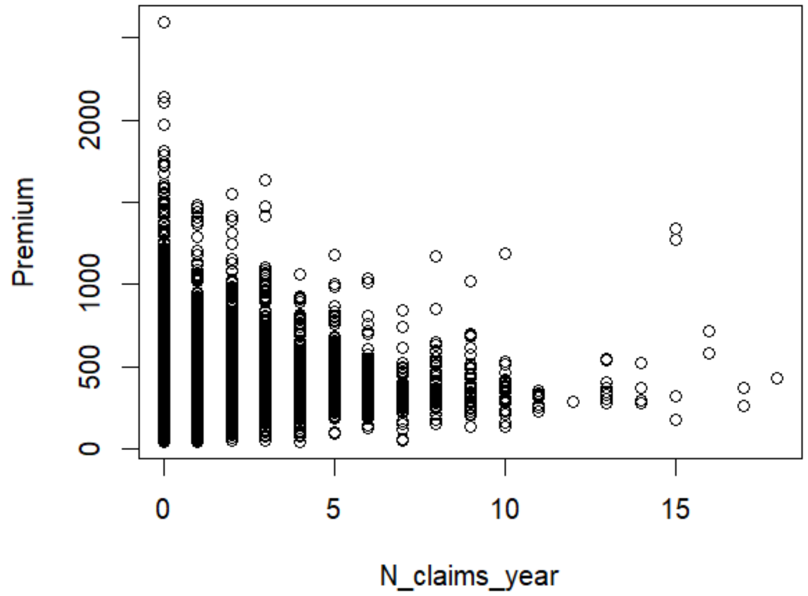

## Addressing Multicollinearity

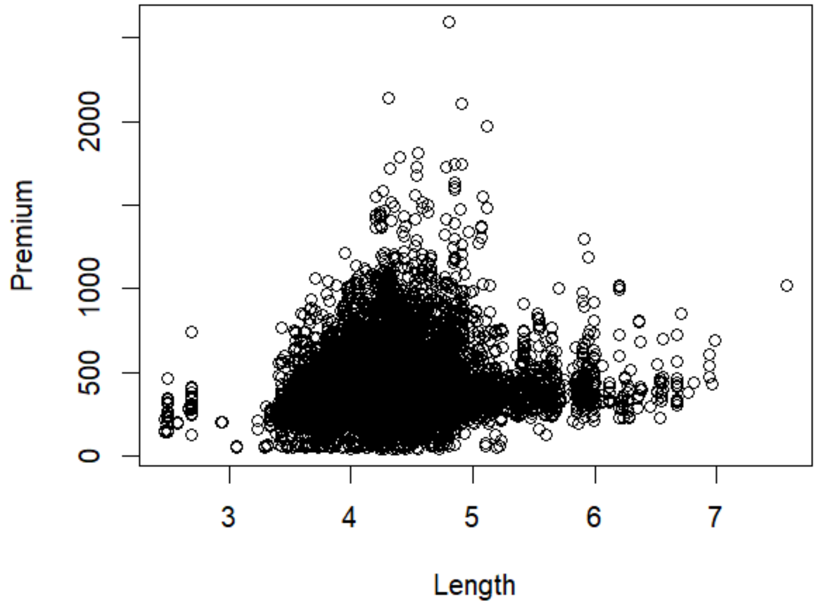
Notable correlations: Weight and length, cylinders and horsepower, policies in force and maximum policies, value of vehicle and weight, value of vehicle and cylinder capacity, etc.

# Data: Notable transformations

**Logarithmic Transformations:**

- Examples: the value of the vehicle, horsepower, claims history, years since the contract started
- Binary variables
- Threshold effects



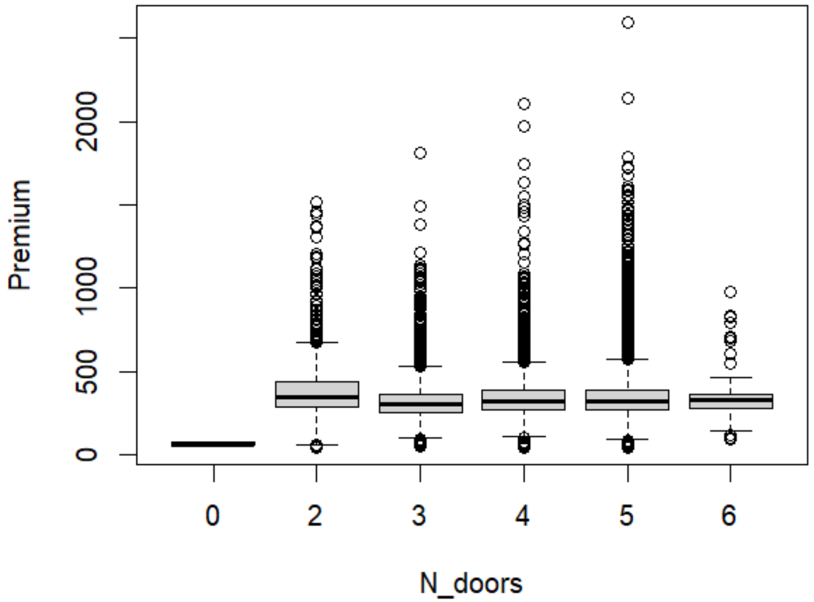



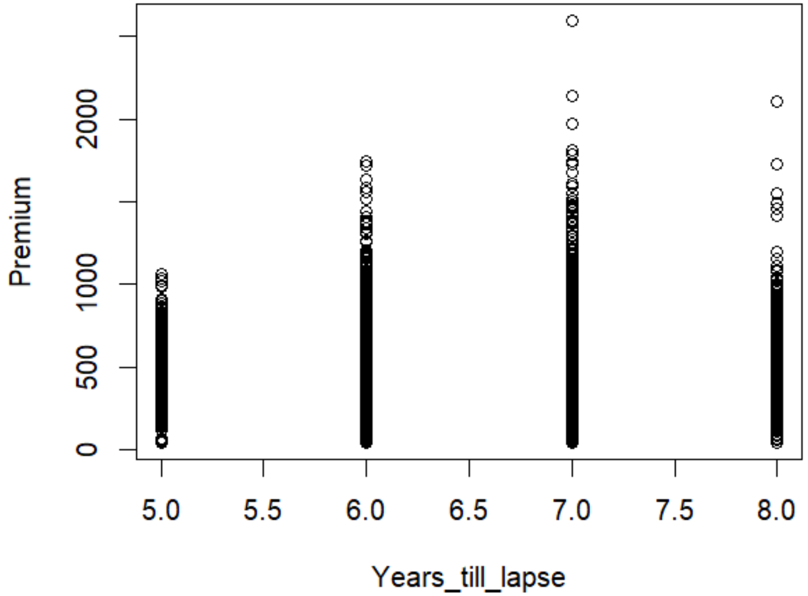

**Quadratic Transformations:**

- Examples: weight, age, cylinder capacity, length
- Turning points
- Physical limits

# Data: Insignificant Predictors

After adding all variables to the first model, these were the least significant predictors




Low Predictive Power

- Unhelpful, shows little risk

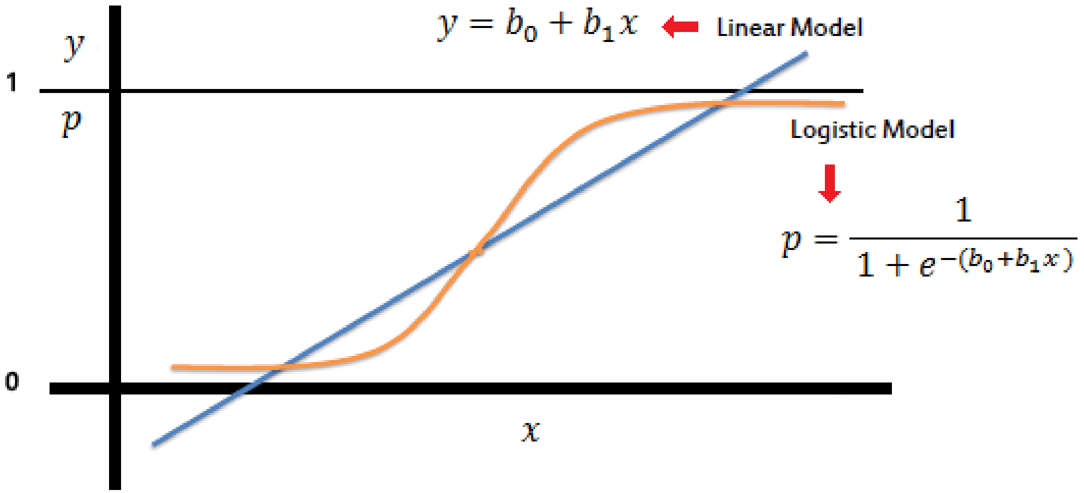
Redundancy

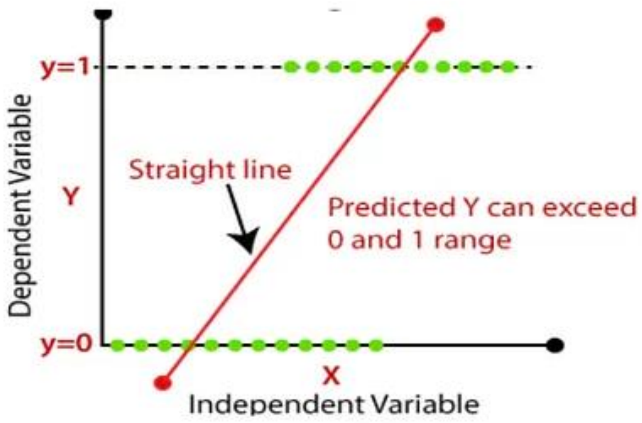
- If highly correlated with other terms (i.e. policies)

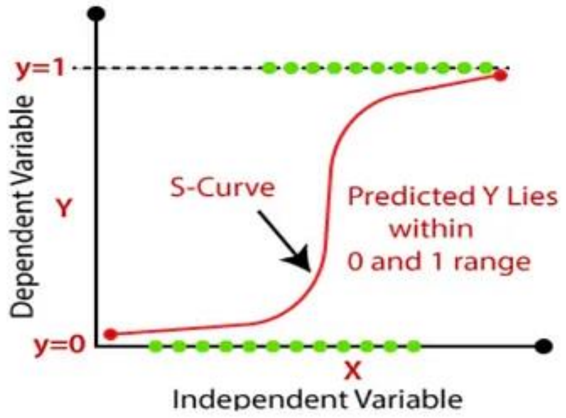
Focus on Significant Drivers of Premiums:

- Prioritize variables that directly affect premium

# Model types: Linear vs Logistic




Linear:

- Predict continuous outcomes
- Assumed linear relationship between predictors and response
- Sensitive to outliers
- Weak because more likely to be non-linear relationships

Logistic:

- Handles binary outcomes effectively
- Valued for segmenting different risk levels
- Probability that the premium is higher or lower
- Less interpretable for complex relationships

# Model Building: Refining for Optimal Fit

- Begin with models including all relevant predictors.
- Drop insignificant variables
- Add interaction terms, transform variables

**Starting Simple & Refinement**

- VIF to detect multicollinearity
- Drop or combine highly correlated variables
- To not get unreliable coefficient estimates

**Handling Multicollinearity**

- R^2
- AIC (Akaike Information Criterion)
- K-Fold Cross Validation: Splits data into folds

**Model Comparison**

Residuals analysis to confirm:

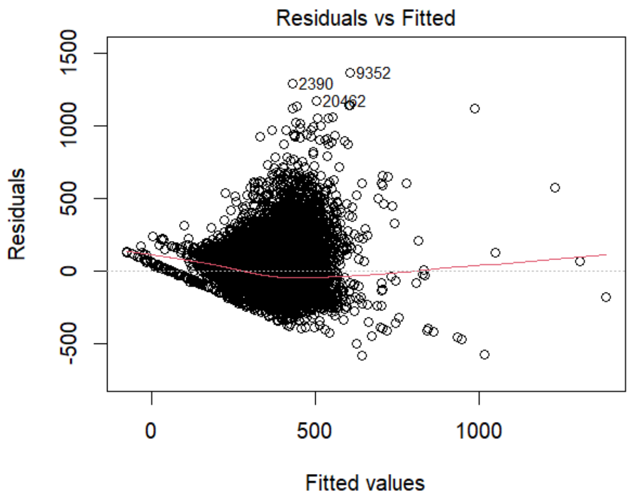
- No patterns (linear models)
- Homoscedasticity
- No influential outliers skewing the results

**Assumptions Validation**

# Building Linear Models

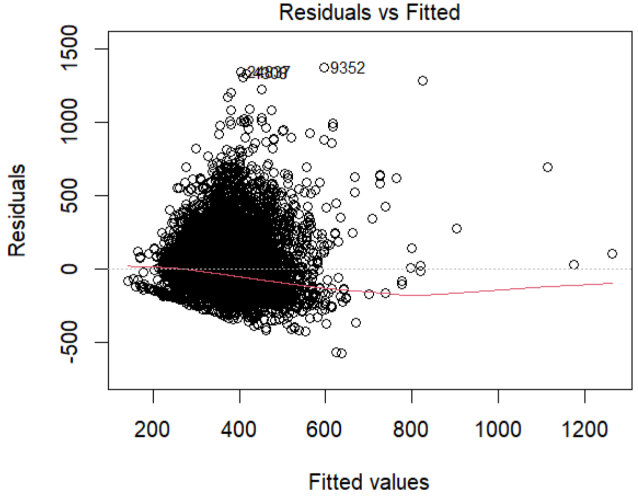
## Model 1

- Added variables highly correlated with premium all the way to the last one
- Included all 26 variables
- $R^2$ = 0.2753
- AIC = 316893.2
- VIF: weight > 5




Model 1

Model 2




## Model 2

- Added interaction terms
- Took out insignificant terms
- $R^2$ = 0.1636
- AIC = 320509.1
- VIF: 3 terms above 50

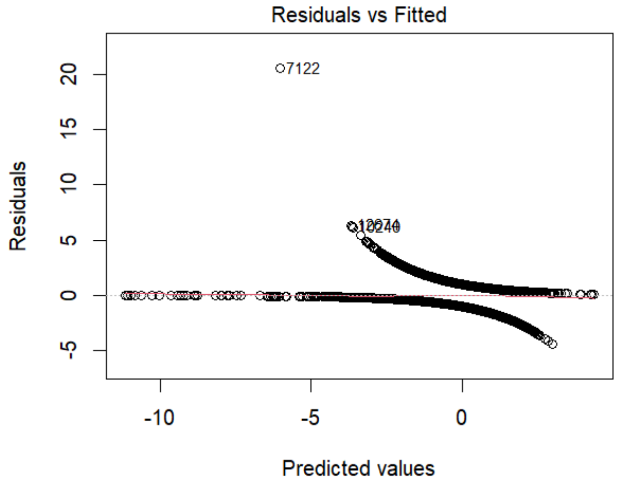
## Model 3

- Took out insignificant variables until all had the highest significance
- $R^2$ = 0.2751
- AIC = 316890.6
- VIF: weight > 5

# Building Logistic Models

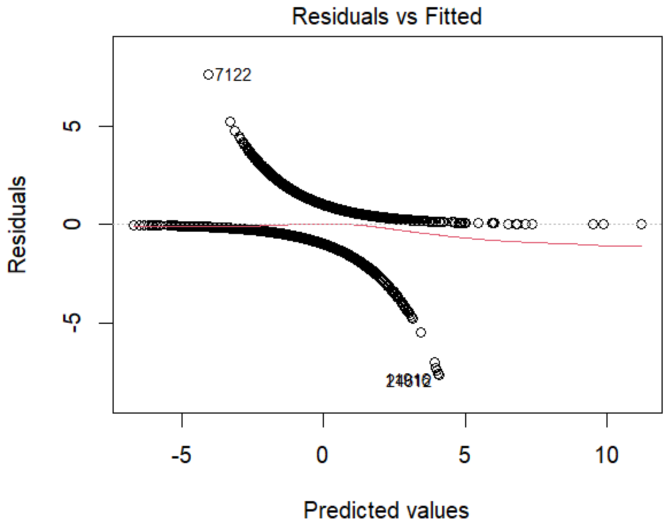
## Model 4

- Logged non-binary variables
- 12 predictors total (3 interaction terms)
- Only cylinders are insignificant
- AIC = 29193
- VIF: no evidence of multicollinearity
- 69% accurate




Models 4

Model 5




## Model 5

- 9 predictors total (1 interaction term)
- Length and weight are insignificant
- AIC = 29548.82
- VIF: length and weight > 16
- 69.1% accurate

## Model 6

- Logged non-binary variables
- 9 predictors total (1 interaction term)
- Only age is insignificant
- AIC = 29295
- VIF: no evidence of multicollinearity
- 68.5 % accurate

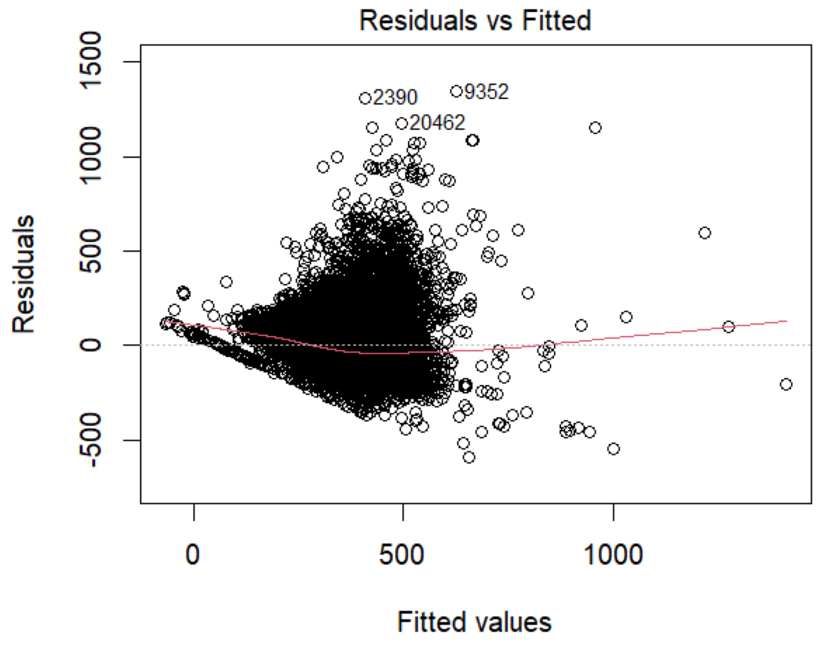
# Final Models

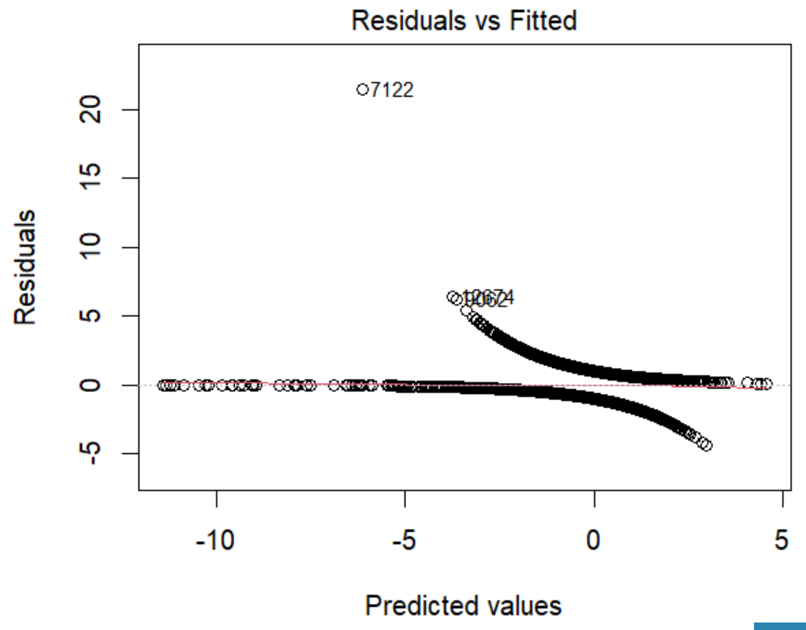
## Model A

- Top performing linear model
- 24 variables included- 5 are interaction terms
- Seniority and length are not significant
- R^2 of 0.2888
- AIC = 316090.8
- VIF: no evidence of multicollinearity
- Performed best out of all models for K=1-10 folds

## Model B

- Top performing logistic model
- 10 variables included- 2 are interaction terms
- All variables p-value < 0.05
- Logged non-binary terms
- AIC = 29194
- VIF: no evidence of multicollinearity
- 69.5% accuracy
- K-folds better than other logistic models



# Conclusion and Implications for Future Work

Next Steps:

- Enhanced data collection and model refinement
- Working with different data
- Build similar models for motorbikes

Self-Driving Cars:

- Potential for predictive analytics on their premiums
- Results show vehicle characteristics had an impact on this model compared to descriptions of the driver
- Fully autonomous vehicle insurance may focus more on software and hardware reliability.

# References

- https://www.openicpsr.org/openicpsr/project/193182/version/V1/view;jsessionid=2178B9C43D61DBCCC82A693D74106160
- https://link.springer.com/article/10.1007/s13385-024-00398-0
- https://online.stat.psu.edu/stat504/lesson/6/6.1
- https://www.casstudentcentral.org/about-our-profession/what-is-a-property-and-casualty-actuary/
- https://www.actuary.org/sites/default/files/files/publications/BigDataAndTheRoleOfTheActuary.pdf
- https://www.soa.org/globalassets/assets/files/static-pages/research/opportunities/2019-student-case-sub-bu.pdf
- https://www.casact.org/sites/default/files/2021-02/pubs_forum_18spforum_01_avtf_2018_report.pdf

# Questions?

Thank you for your attention.

Questions?

LinkedIn: www.linkedin.com/in/kylie-wilkin
Email: kyliejwilkin@gmail.com