Travelers Analytics:
U of M Stats 8053 Insurance Modeling Problem

October 30th, 2013

Nathan Hubbell, FCAS
Shengde Liang, Ph.D.

TRAVELERS

Agenda

- Travelers: Who Are We & How Do We Use Data?
- Insurance 101
  - Basic business terminology
- Insurance Modeling Problem
  - Introduction
  - Exploratory Data Analysis
  - Assignment Walk-through

How is data used at Travelers?

- Loss, Premium, and Financial Data
- Research & Development
- Unstructured

- Traditional Actuarial Usage
  - Univariate analysis
- Includes external data
  - Multivariate analysis
  - Example: GLMs allow for a non-linear approach in predictive modeling.
- Future development
  - Continued use of sophisticated statistical methods

Insurance 101

Basics of Insurance

Insurance companies sell insurance policies, which are the promise to pay in the event that a customer experiences a loss.

The unique challenge in insurance is that we don't know what the cost of insuring a customer is when we sell the policy.

Example: The cost to insure an auto customer

It's impossible to predict if someone is going to

- Get into an accident
- The type of accident (hit a telephone pole, hit another vehicle, bodily injury)
- How bad (cost) the accident will be

Business Impact of Loss Experience

To estimate the cost of insuring policyholders, we must predict losses

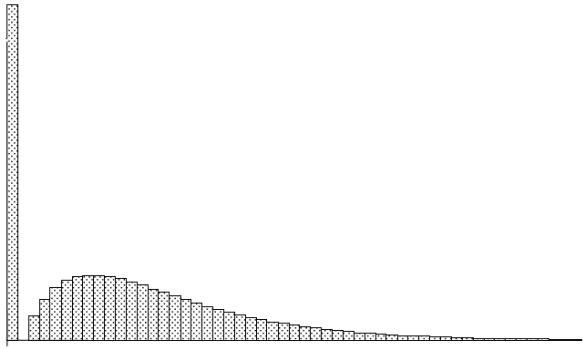
Two fundamental questions we must answer are:

1. **Ratemaking**: looking to the future
   - Setting rates for policies
   - How much do we need to charge customers for a policy in order to reach our target profit? *Basic idea: price = cost + profit*
2. **Reserving**: looking at the impact of past experience
   - Setting aside reserve money
   - How much money do we need to set aside to pay for claims?

Note: We cannot precisely predict losses for each individual or business. However, if we group our customers together, we can build statistical models to predict average loss over a group.

Model Building

- Generalized Linear Models (GLMs)
  - Potential response variables:
    - **Claims** – **Frequency** (# claims / **exposure**) (e.g. Poisson, Negative Binomial)
    - **Loss** – **Severity** (loss $ / claim) (e.g. Gamma, Inverse Gaussian)
    - **Pure Premium** = Frequency * Severity = loss $ / exposure
  - A common **link function** is g(x) = ln(x).
  - Probability distribution: **Tweedie**
    - Compound distribution of a **Poisson** claim #
    - And a **Gamma** claim size distribution
    - Large spike at 0 for policies with no claims
    - Wide range of amount in the claims



- Challenges include:
  - Variable selection
  - Bias-variance trade-off

So what is an example of an actual modeling problem in insurance?

* Source: “A Practitioner's Guide to Generalized Linear Models”

What questions do you have about:

- Travelers?
- Insurance?
- Statistics at Travelers?

Business Problem

- **Refer to the one page hand out "Kangaroo Auto Insurance Company Modeling Problem" for more details**
- You, as a statistician, work for Kangaroo Insurance, an Australian insurance company
- The underwriter in your company would like you to build a pricing model (pure premium) for the auto insurance product.
- The pricing needs to be **competitive**.
  - accurately reflect the risk your company is taking.
  - enough segmentation among customers.
- The data from policies written in 2004 and 2005 is provided.

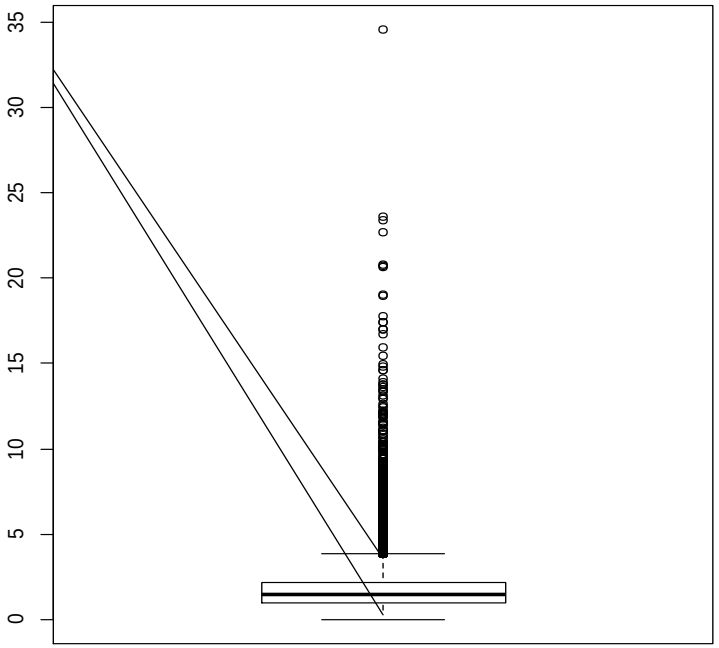
Data Information

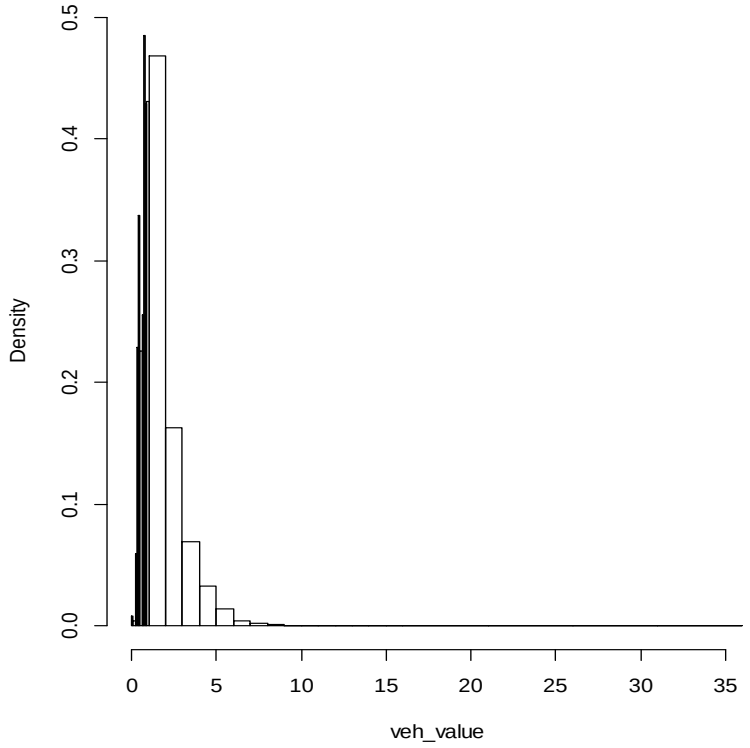
- Losses for each vehicle from policies written in 2004 and 2005.
- Each policy was written as one-year originally.
- There are 67,856 policies (vehicles) in the data.
- Ten (10) variables in the data.

veh_value	exposure	clm	numclaims	claimcst0	veh_body	veh_age	gender	area	agecat	_OBSTAT_			
1.06	0.303901	0	0	0	HBACK	3	F	C	2	01101	0	0	0
1.03	0.648871	0	0	0	HBACK	2	F	A	4	01101	0	0	0
3.26	0.569473	0	0	0	UTE	2	F	E	2	01101	0	0	0
4.14	0.317591	0	0	0	STNWG	2	F	D	2	01101	0	0	0
1.38	0.854209	0	0	0	HBACK	2	M	A	2	01101	0	0	0
1.22	0.854209	0	0	0	HBACK	3	M	C	4	01101	0	0	0
1	0.492813	0	0	0	HBACK	2	F	C	4	01101	0	0	0
7.04	0.314853	0	0	0	STNWG	1	M	A	5	01101	0	0	0
1.66	0.4846	1	1	669.51	SEDAN	3	M	B	6	01101	0	0	0
2.35	0.391513	0	0	0	SEDAN	2	M	C	4	01101	0	0	0
1.51	0.99384	1	1	806.61	SEDAN	3	F	F	4	01101	0	0	0
0.76	0.539357	1	1	401.8055	HBACK	3	M	C	4	01101	0	0	0
1.89	0.654346	1	2	1811.71	STNWG	3	M	F	2	01101	0	0	0

Variable Information – veh_value

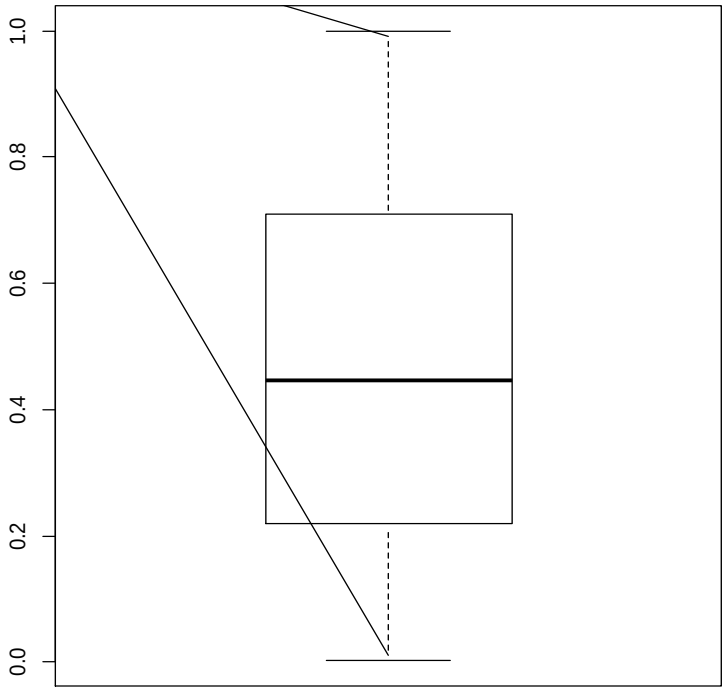
- vehicle value, in $10,000s, a numerical variable.

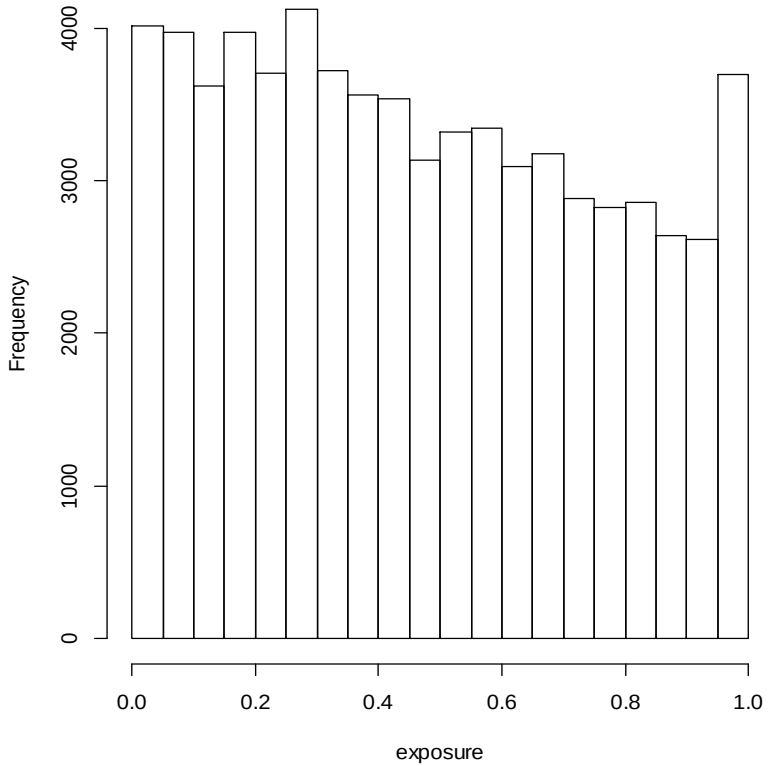




Variable Information – exposure

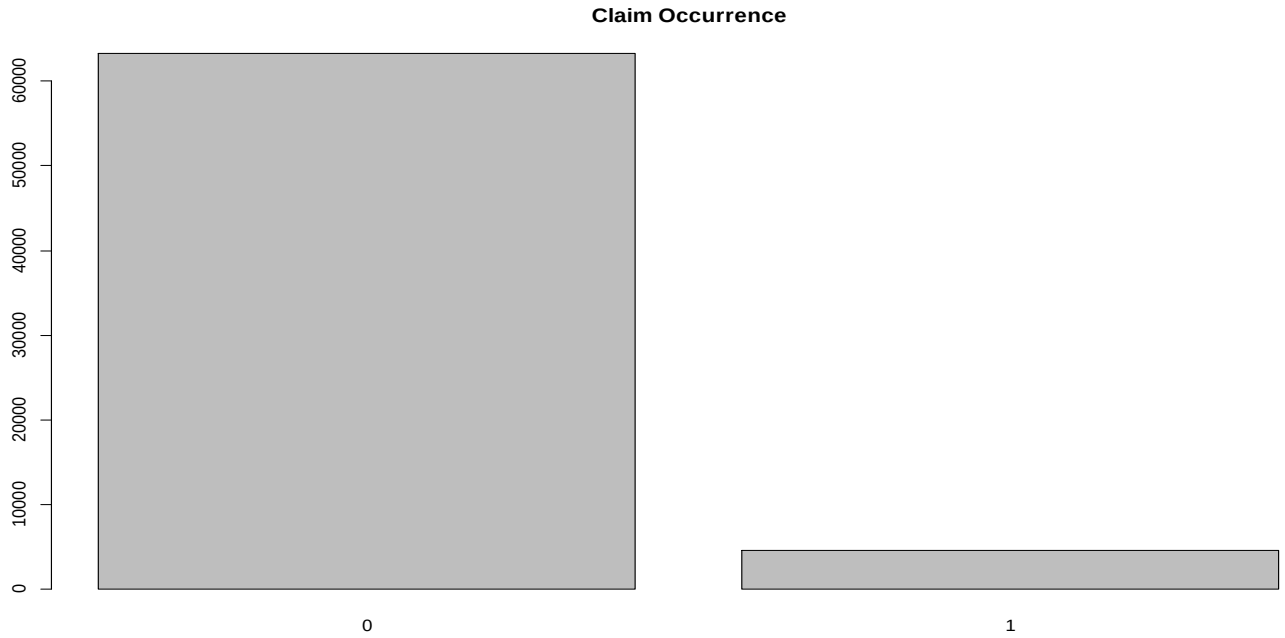
- The covered period, in years, a numerical variable (always between 0 and 1)
  - The amount of time a vehicle was “exposed” to potential accidents.





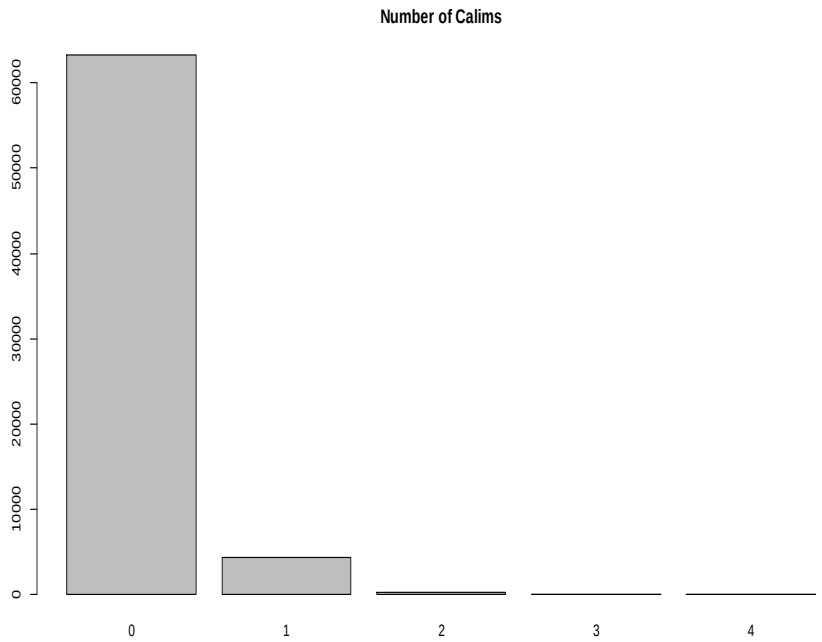
Variable Information – clm

- An indicator whether the vehicle/driver had at least one claim during the covered period, 0=No, 1=Yes.
- 4,624/67,856 had at least one claim.



Variable Information – numclaims

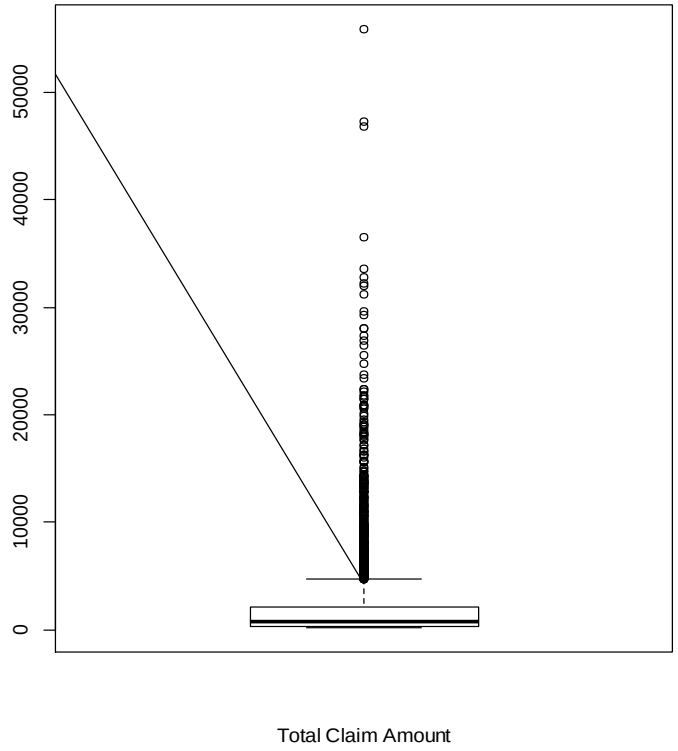
- Number of claims during covered period, integer values.
- 4,624/67,856 had at least one claim.

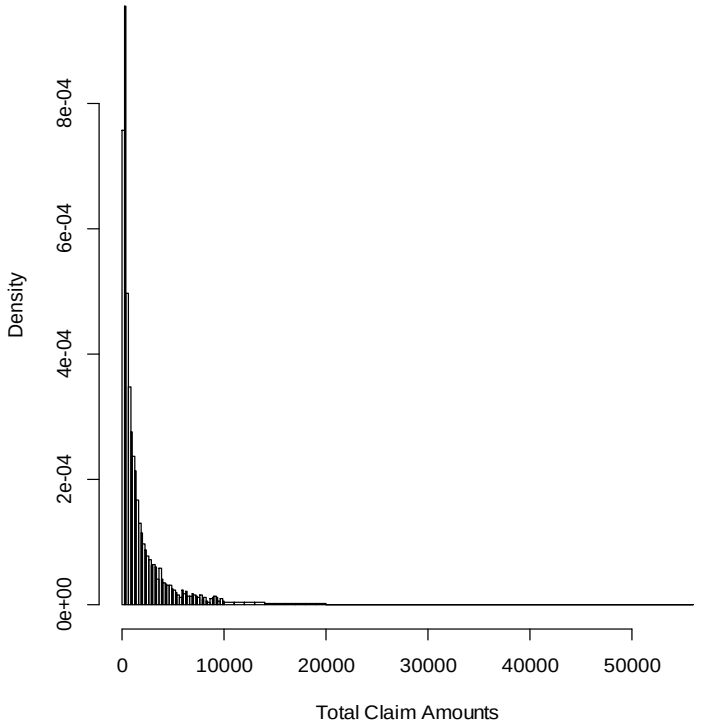


Number of Claims	Frequency
0	63,232
1	4,333
2	271
3	18
4	2

Variable Information – claimcst0 (**target variable**)

- The total amount of the claims, in dollars, numeric values.

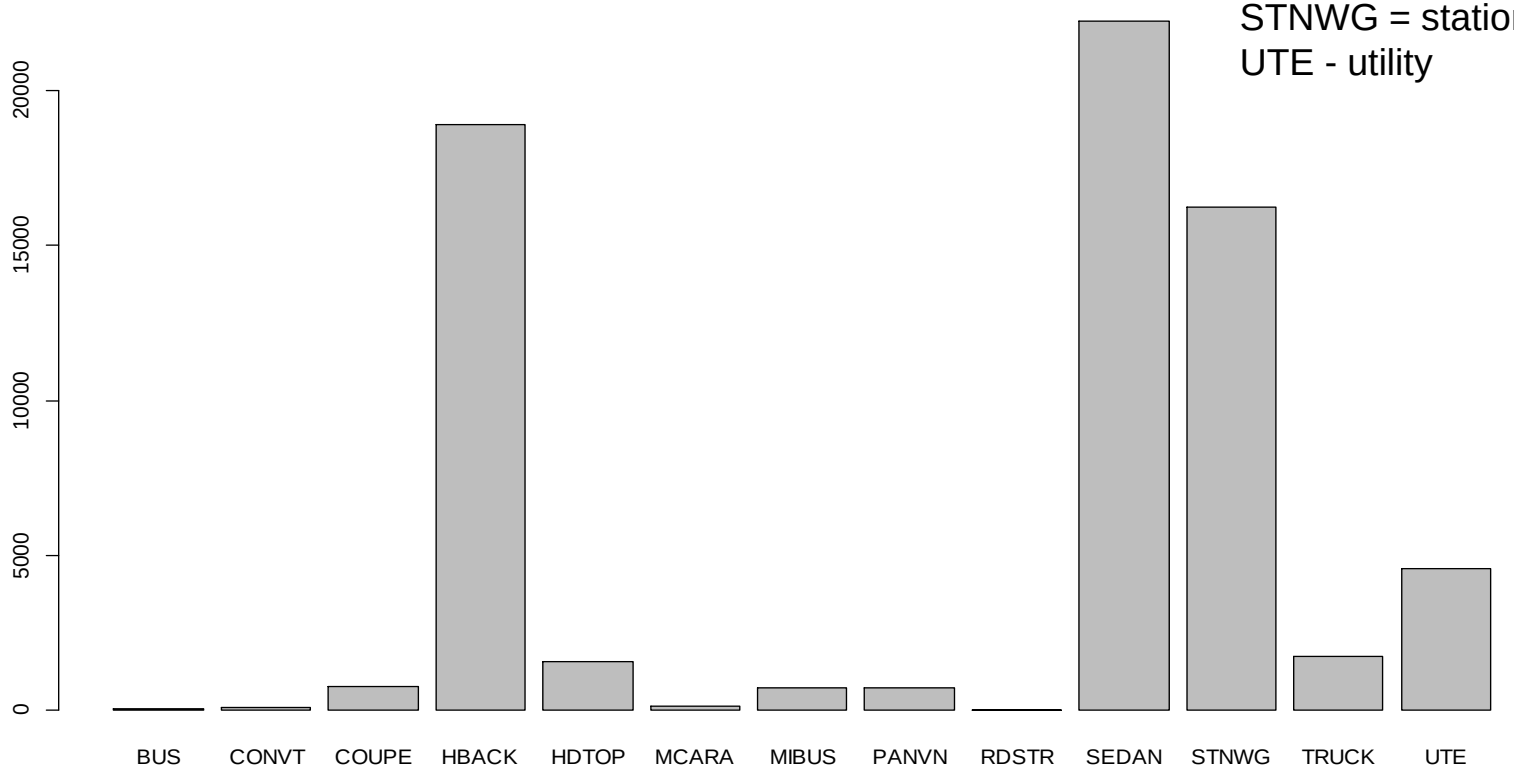




Variable Information – veh_body

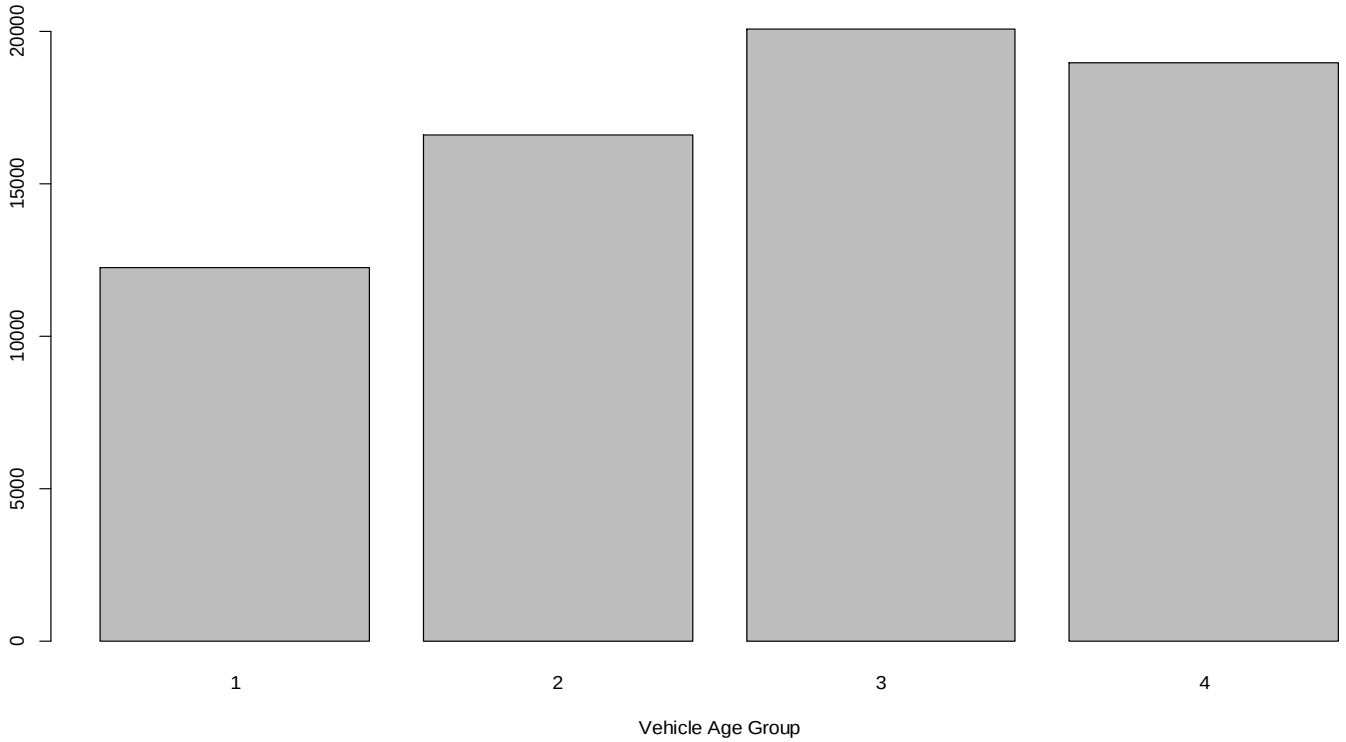
CONVT = convertible
HBACK = hatchback
HDTOP = hardtop
MCARA = motorized caravan
MIBUS = minibus
PANVN = panel van
RDSTR = roadster
STNWG = station wagon
UTE - utility

- The vehicle body code, a character string.



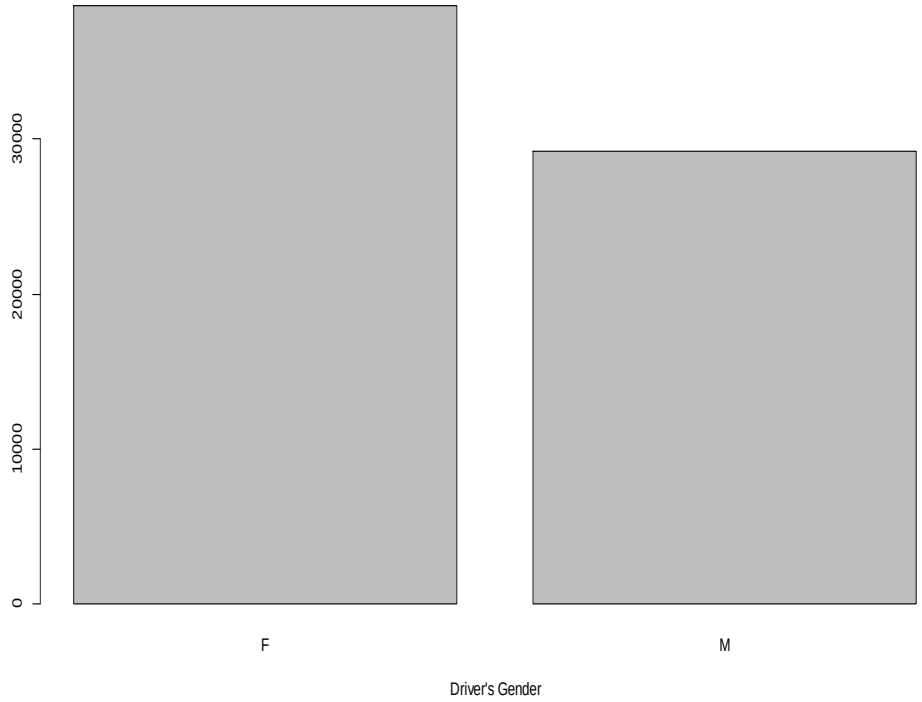
Variable Information – veh_age

- The age group of insured vehicle, coded as 1, 2, 3, and 4, with 1 being the youngest.



Variable Information – gender

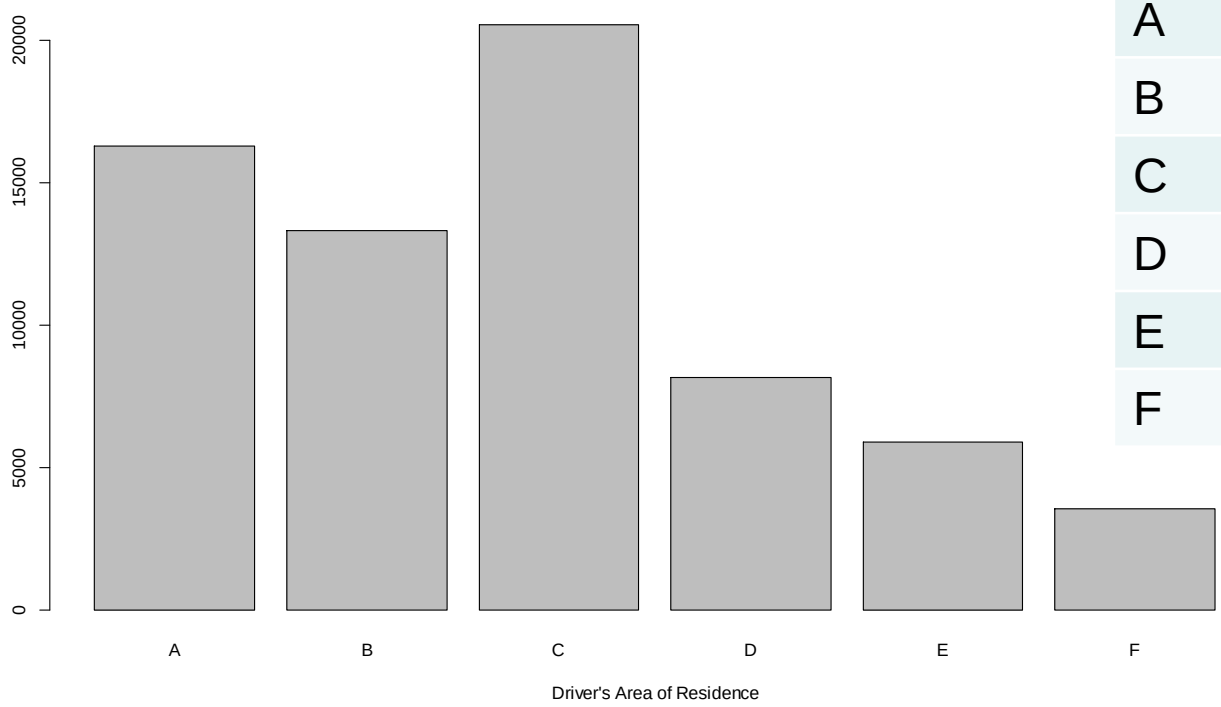
- The gender of driver, F (female) or M (male)



Gender	Frequency
F	38,603
M	29,253

Variable Information – area

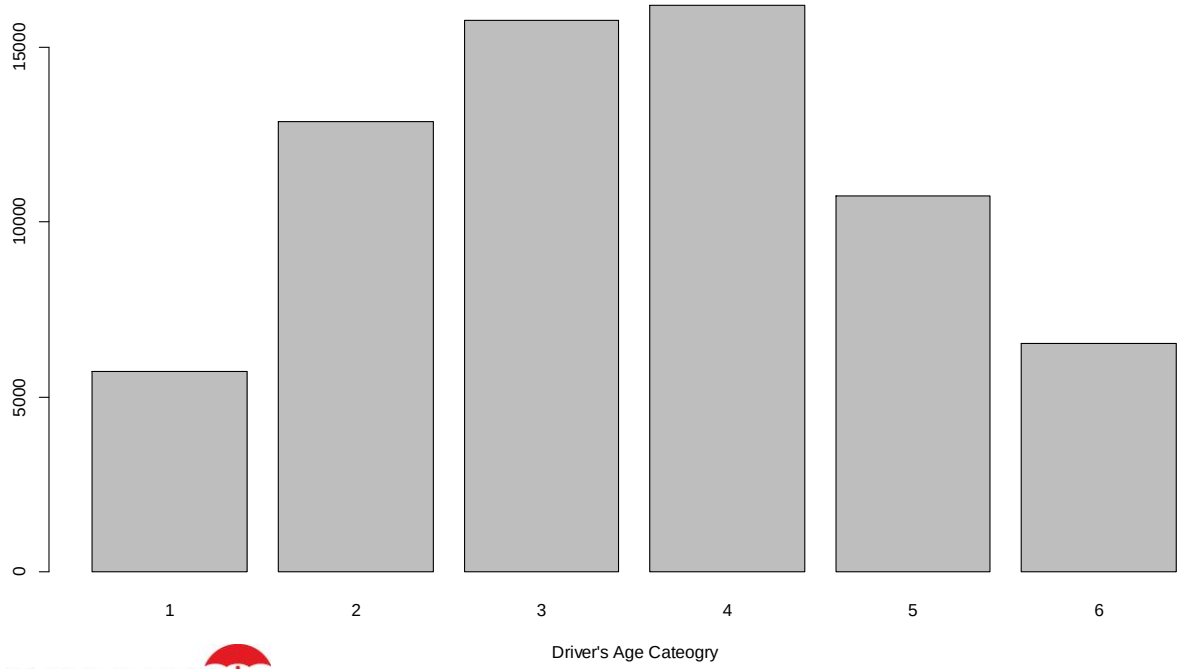
- Driver's area of residence, a character code.



Area Code	Frequency
A	16,312
B	13,341
C	20,540
D	8,173
E	5,912
F	3,578

Variable Information – agecat

- Driver's age category, coded as 1, 2, 3, 4, 5 and 6, with 1 being the youngest.



Driver Age Category	Frequency
1	5,742
2	12,875
3	15,767
4	16,189
5	10,736
6	6,547

Questions May Be Asked

- What models did you fit?
  - what is your assumption(s)?
  - is your assumption reasonable?
  - how do you check your assumption(s)?
- What is the impact of each variable?
  - are all variables equally important?
  - if not, which ones are more important? How do you measure it?
- How do you check your model actually works (genaralizability)?

What questions do you have about the
"Kangaroo Insurance Company Modeling Problem"?

References and Resources

- **Contacts**
  - **Nathan Hubbell – NHUBBELL@travelers.com**
  - **Shengde Liang – SLIANG@travelers.com**

- **Travelers Careers**
  - http://www.travelers.com/careers
  - Actuarial and Analytics Research Internship and Full Time

- **A Practitioner's Guide to Generalized Linear Models**
  - http://www.towerswatson.com/assets/pdf/2380/Anderson_et_al_Edition_3.pdf