CAPSTONE PRESENTATION

Rafah Asadi, Brett Bejcek, Peter Jacobs, Kyle Voytovich

APRIL 18, 2017

THE OHIO STATE UNIVERSITY

UNDERGRADUATE
DATA ANALYTICS MAJOR

OVERVIEW

SCOPE --

- RECAP OF COURSE OBJECTIVE
- PROBLEM DEFINITION

APPROACH, TECHNIQUES, AND FINDINGS ------------

- INITIAL STRATEGY
- LOGISTIC REGRESSION, DECISION TREES, LINEAR REGRESSION
- VISUALIZATIONS

RECOMMENDATIONS ------------------------------------

- KEEPING THE GOOD
- IMPROVING THE BAD

CONCLUSION ---

- REFLECTION ON EXPERIENCE

COMPUTER SCIENCE

DEMONSTRATE AN ABILITY TO APPLY COMPUTER SCIENCE PRINCIPLES RELATING TO DATA REPRESENTATION, RETRIEVAL, PROGRAMMING, AND ANALYSIS.

STATISTICS

DEMONSTRATE AN ABILITY TO APPLY STATISTICAL MODELS AND CONCEPTS TO ANALYZE DATA AND DRAW CONCLUSIONS BASED ON DATA.

LEARNING OBJECTIVES

CRITICAL THINKING

DEMONSTRATE CRITICAL THINKING SKILLS ASSOCIATED WITH PROBLEM IDENTIFICATION, DECISION MAKING, AND SYNTHESIS OF INFORMATION.

COMMUNICATION

DEMONSTRATE AN ABILITY TO COMMUNICATE FINDINGS TO INDIVIDUALS WITH VARYING LEVELS OF TECHNICAL KNOWLEDGE.

CSE 4193: DATA ANALYTICS CAPSTONE

UNDERSTAND WHAT SEGMENTS OF THE CUSTOMERS INSURED HAVE GOOD LOSS RATIOS AND WHAT SEGMENTS OF CUSTOMERS INSURED HAVE POOR LOSS RATIOS.

SCOPE OF PROBLEM

SCHEMA DEFINITION

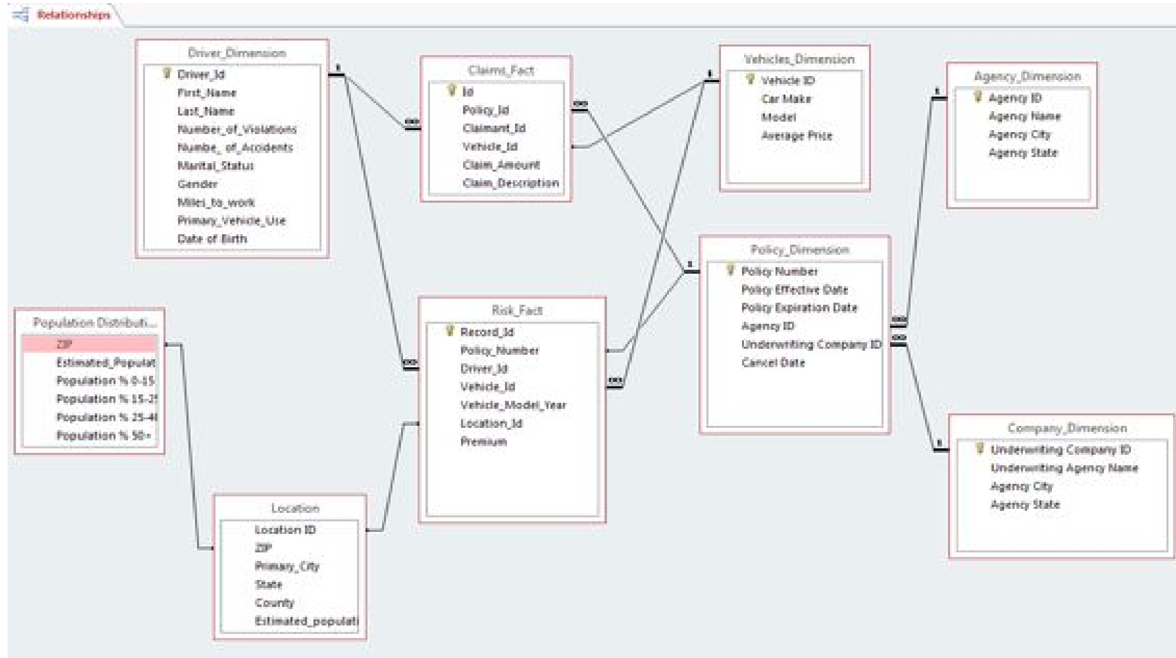
DATABASE CREATION

TABLE AGGREGATION

FIRST LOOKS

EDA TOOL CREATION

INTERACTION DISCOVERY



id	PolicyID	DriverID	VehicleID	LocationID	Premium	ClaimsAmount	id:1	Fname	Lname	Violations	Accidents	MaritalStatus
35566	1014425	1041	9	316	134	0	1041	Shalon	Manalo	0	0	S
35567	1015462	1146	82	110	351	7470	1146	Odette	Truby	5	0	S
35563	1002901	1274	73	82	366	0	1274	Latesha	Besong	4	1	S
35564	1009972	1328	63	369	178	6400	1328	Lakesha	Schuh	0	0	M
35565	1014384	1446	28	207	152	0	1446	Bennie	Altomari	1	0	S
11	1000001	3001	34	229	374	0	3001	Josefa	Turnbill	6	2	S
22644	1000002	3002	61	401	403	0	3002	Deidre	Whilden	6	2	S
20253	1000003	3003	61	576	228	0	3003	Fernande	Lashbrook	2	0	S
7801	1000004	3004	42	405	224	0	3004	Arica	Relihan	3	1	M
21658	1000005	3005	84	259	341	0	3005	Kimberli	Dumont	1	0	S
27517	1000006	3006	68	211	233	0	3006	Isaac	Eget	0	0	D
23458	1000007	3007	48	144	212	0	3007	Nada	Schueler	3	0	S
29257	1000008	3008	82	188	271	0	3008	Cortez	Sammer	3	0	S
11754	1000009	3009	49	522	284	0	3009	Adan	Meyer	1	0	M

GENERAL APPROACH

SCHEMA DEFINITION

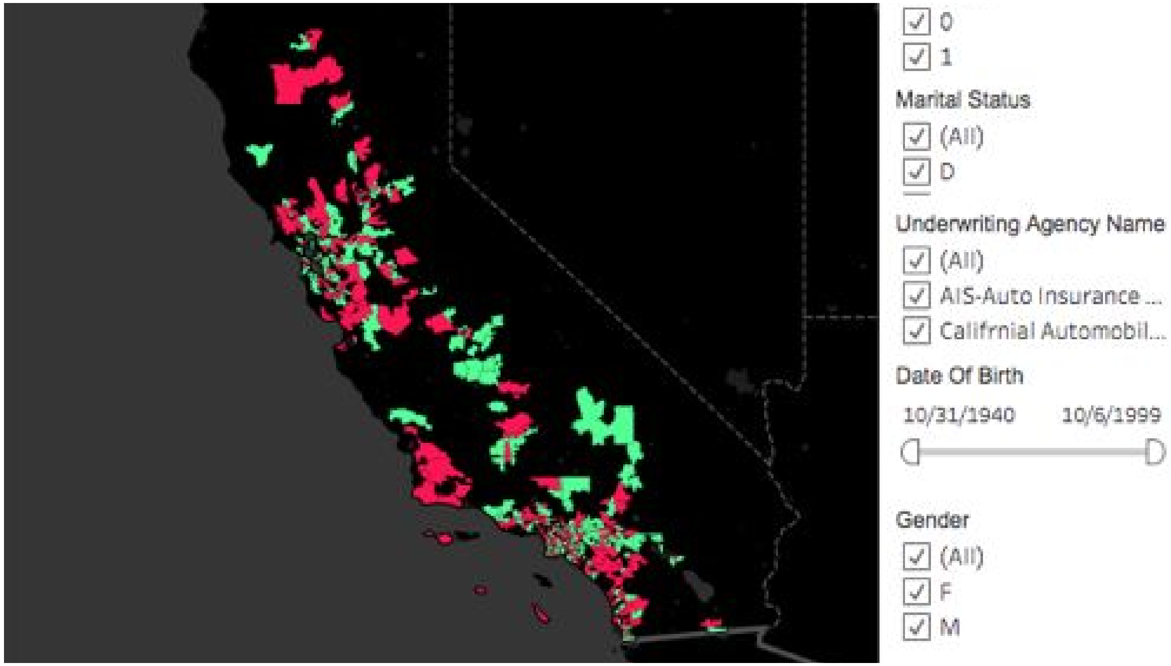
DATABASE CREATION

TABLE AGGREGATION

FIRST LOOKS

EDA TOOL CREATION

INTERACTION DISCOVERY



Violations

Violations	Losses	Premiums	Loss.Ratio
0	1995650	1297718	1.537815
1	3144328	1298587	2.421346
2	2254577	1392465	1.619126
3	2969347	1817335	1.633902
4	1264474	1047767	1.206827
5	1423964	890721	1.598664
6	1269859	903930	1.404820

GENERAL APPROACH

PART 1: Use modeling techniques to help identify profitable and non-profitable population segments

PART 2: Visualize these population segments to ensure model output is sensible

PART 3: Derive insights from model output and visualizations.

PART 4: Use these insights to make recommendations about where to change/(not change) premium pricing

ANALYTICAL STRATEGY

LOGISTIC REGRESSION

Model Claim/No Claim

- 35,712 total covered drivers
- 1,236 drivers with a claim

The Model

- Let Y_i be whether driver i had a claim
- $Y_i \sim \text{Bernoulli}(p_i)$
- $\text{Logit}(p_i) = X_i^T\beta$

The Predictors (X)

- Approximately 15 predictors appropriate for the model (premium, violations, etc.)
- Search the model space (including single variables and two-way interactions)

LOGISTIC REGRESSION

Purpose of Model

- Significant predictors can be indicators for risk groups that are more/less likely to make a claim (feed into the visualization team)
- Does **not** take into account claim severity

Results

- Little significant when premium is already in the model
- Predictors have a weak relationship with claim frequency

Next steps

- Different approaches to learn the connection between the descriptors and claim frequency/severity
- Look into Decision Trees

DECISION TREES

Step 1:

Use decision trees to discover interactions between variables that influence

- Whether or not someone has a claim (which segments of the population are more likely to have claims than others?)

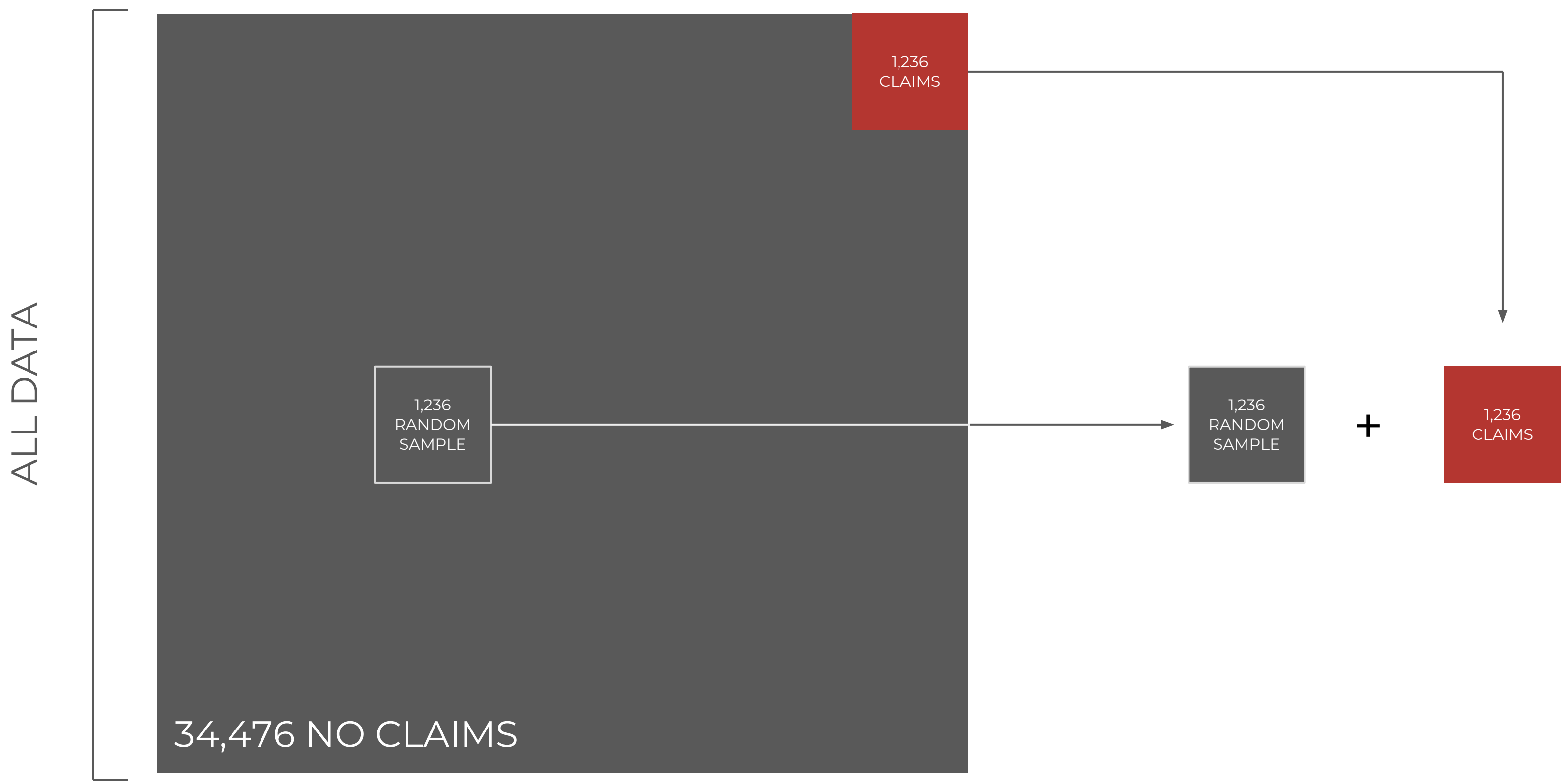
- Severity of the claim (conditional on having a claim, which segments of the population have more severe claims?)

Step 2:

Feed insights about interactions to the visualization team to further explore these interactions

DECISION TREE #1

TREE ON WHETHER OR NOT SOMEONE HAS A CLAIM

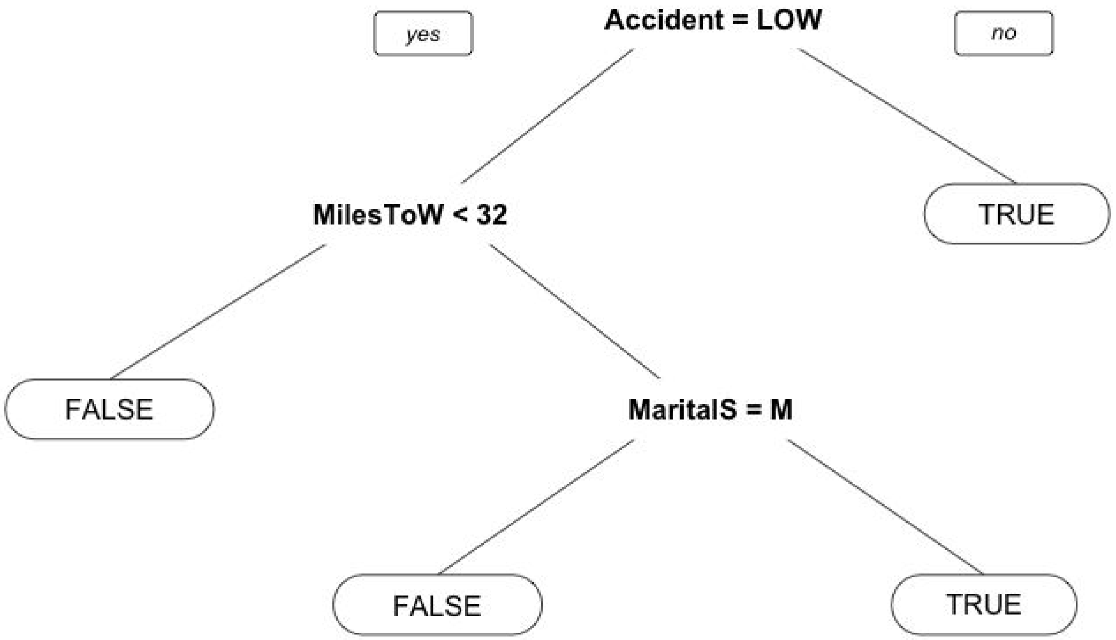


DECISION TREE #1A

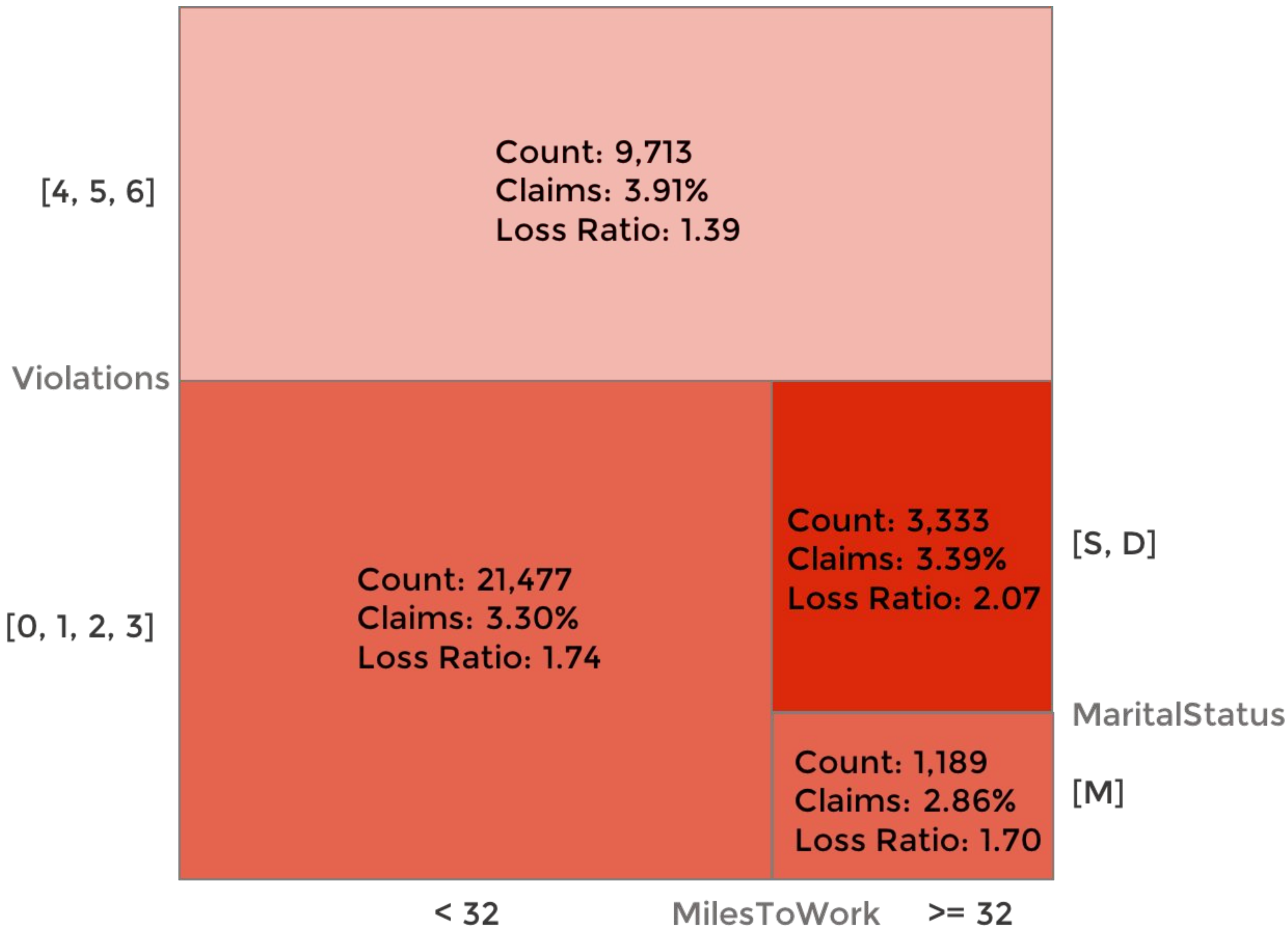
INVESTIGATING DRIVER FACTORS

- Dependent Variable: Whether or not someone had a claim.

Independent Variables Used
Age
of Violations
of Accidents
Gender
Miles to Work
Primary Vehicle Usage
Marital Status



DECISION TREE #1B



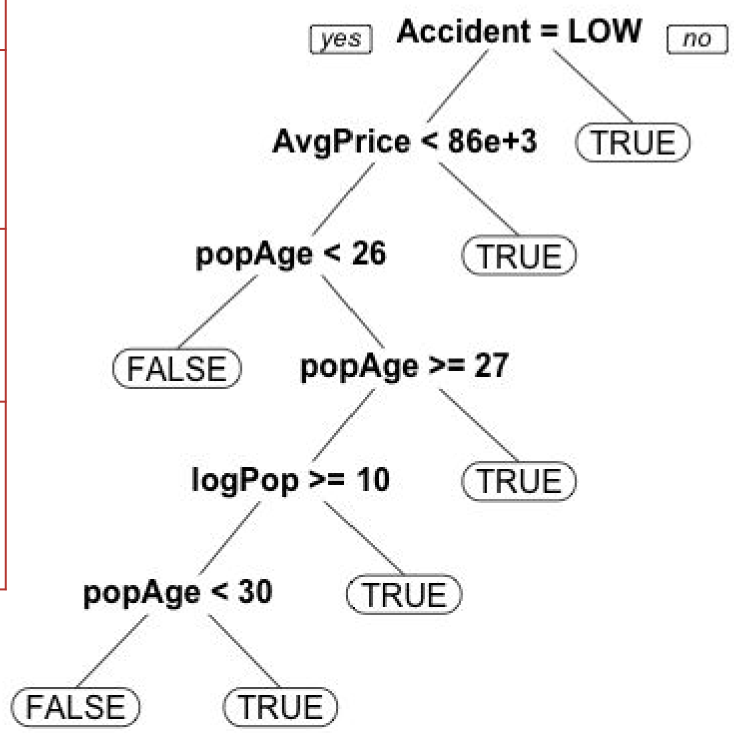
DECISION TREE #2

INVESTIGATING DRIVER FACTORS AND ADDITIONAL ATTRIBUTES

- Dependent Variable: Whether or not someone has claim

Driver Factors
Age
Number of Violations
Gender
Miles to Work
Primary Vehicle Usage
Marital Status

Additional Attributes
Expected age of population risk lives in
Size of population risk lives in
Average price of car risk drives



DECISION TREE #3

CLAIM SEVERITY

- By filtering the out outliers, we have a better chance of capturing general trends in the data. There are about 25 points (defined as those points with Claim Amount - Premium above the 98th percentile).

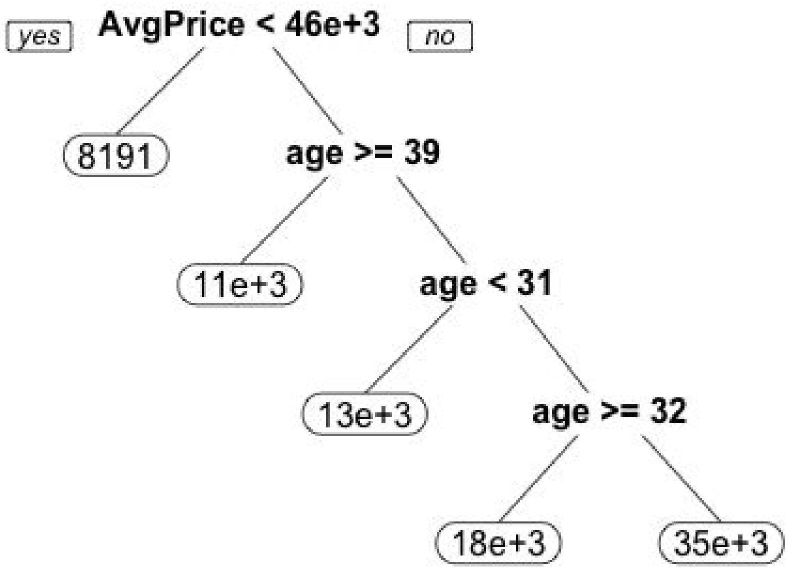
DECISION TREE #3

INVESTIGATING DRIVER FACTORS AND ADDITIONAL ATTRIBUTES (BY SEVERITY)

- Dependent Variable: Claim Amount - Premium

Driver Factors
Age
Number of Violations
Gender
Miles to Work
Primary Vehicle Usage
Marital Status

Additional Attributes
Expected age of population risk lives in
Size of population risk lives in
Average price of car risk drives

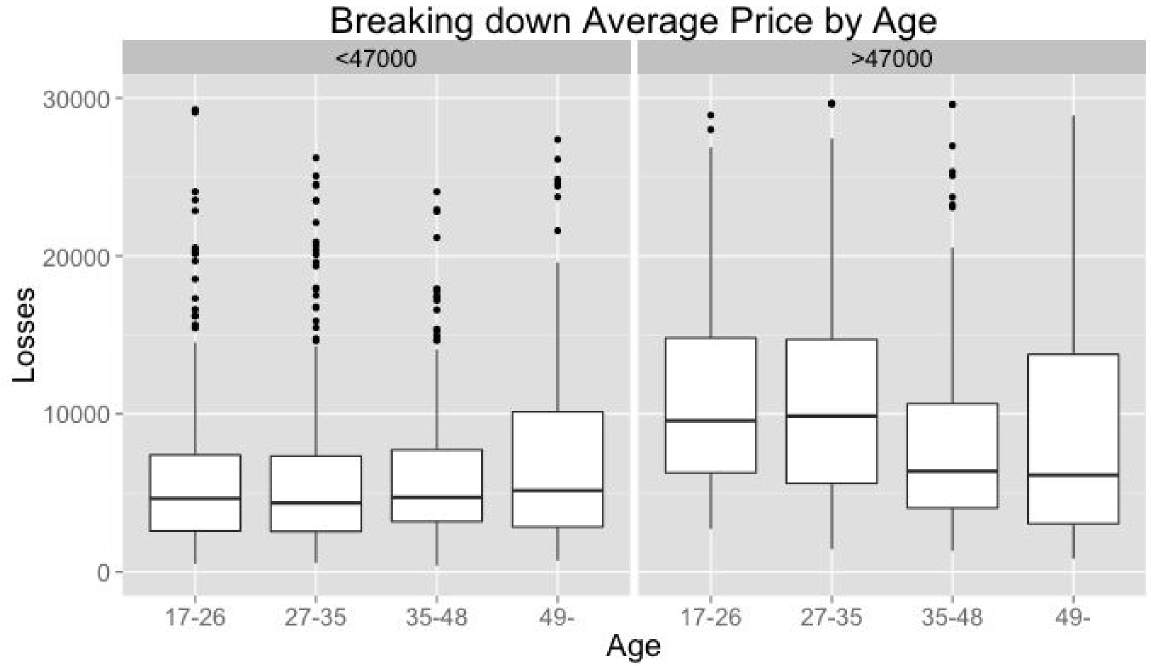


GENERATING HYPOTHESES

EXAMPLE HYPOTHESIS (FROM TREE #3)

- People with higher priced cars tend to have higher losses than people with lower priced cars
- Conditional on people having higher priced cars, those who are younger tend to have higher losses

Visualization falls in line with hypothesis



DECISION TREE INSIGHT OVERVIEW

CLAIM FREQUENCY INSIGHTS

- Individuals with more accidents tend to be more likely to have a claim
  - Conditional on an individual having many accidents, being divorced or single seems particularly strongly associated with an individual having a claim
  - Conditional on an individual having fewer accidents, the higher the price of the vehicle, the more likely an individual will have a claim [Does not hold for higher accidents]
- Divorced and Single individuals in general tend to be more likely to have a claim than married individuals
- Individuals with more expensive vehicles tend to have more claims

CLAIM SEVERITY INSIGHTS

- Individuals with higher priced cars tend to have more severe claims
  - Conditional on the car being higher priced, young people tend to have more claims than old people

LINEAR REGRESSION

WHY LINEAR REGRESSION?

- The goal is to identify predictor variables and interactions between variables that influence the the response (LossRatio , Profit)

WHY ANALYSIS BY ZIPCODE?

- To explore how the features of an area impact the response variable in that area,
- Because population and age_group percentages data are by Zipcode
- There are 600 Zipcodes vs. (39 counties, 20 agency, 7 companies)

THE METHOD

- Given all variables, let Lasso (a variable selection and shrinkage method) find the linear regression model (of main effects and two-way interaction) with the smallest MSE.

LINEAR REGRESSION

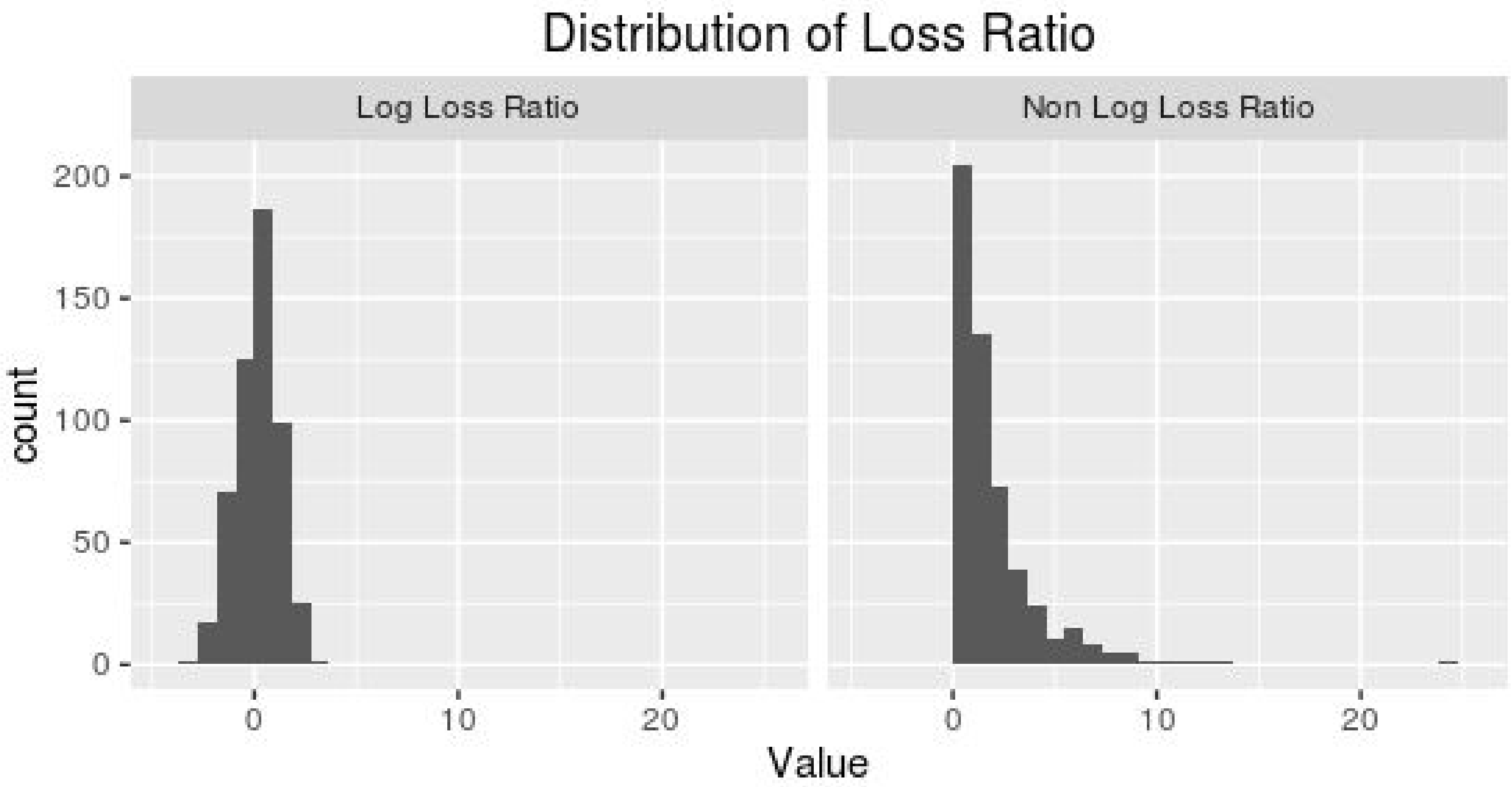
The Covariates (X)

- 23 covariates
- Grouped by zipcode as sums or percentages.
- “ClaimsFreq", "Violations","Accidents", "MilesToWork", "Population", "Percent0to15", "Percent15to25", "Percent25to40", "Percent50", "PercentFemale", "PercentMale", "PercentSingle", "PercentMarried", "PercentDivorced","PercentPriVehUsage_Work", "PercentPriVehUsage_Leisure", "AvgPrice_grp1", "AvgPrice_grp2", "AvgPrice_grp3","AvgPrice_grp4", "NumUW_Agency", "NumAgents", "numberDriversOnPolicy"
- CarAvgPrice - based on quantile of the variable

	grp1		grp2		grp3		grp4	
0%		25%		50%		75%		100%
11000		28000		35000		47000		166000

MODEL.1 - LOGLOSSRATIO

Why LogLossRatio?



MODEL.1 - LOGLOSSRATIO

```
lm(formula = LogLossRatio ~ ClaimsFreq + Population + PercentFemale +
    PercentMarried + PercentPriVehUsage_Work + AvgPrice_grp4 +
    ClaimsFreq:Population + ClaimsFreq:PercentFemale + ClaimsFreq:PercentMarried +
    ClaimsFreq:PercentPriVehUsage_Work + ClaimsFreq:AvgPrice_grp4,
    data = ZipCodeSet)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.27656 -0.53896 -0.06002  0.51598  2.35473 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)  
(Intercept)                       -8.737e-01  1.124e+00  -0.777   0.4374  
ClaimsFreq                         5.789e+00  2.200e+01   0.263   0.7926  
Population                         9.098e-07  5.926e-06   0.154   0.8780  
PercentFemale                     -9.156e-01  1.172e+00  -0.781   0.4351  
PercentMarried                     5.549e-01  1.233e+00   0.450   0.6530  
PercentPriVehUsage_Work            7.982e-01  1.268e+00   0.629   0.5294  
AvgPrice_grp4                     -2.941e-02  1.449e-02  -2.030   0.0429 *
ClaimsFreq:Population              3.939e-05  1.315e-04   0.299   0.7647  
ClaimsFreq:PercentFemale           3.310e+01  2.423e+01   1.366   0.1725  
ClaimsFreq:PercentMarried          4.122e+00  2.492e+01   0.165   0.8687  
ClaimsFreq:PercentPriVehUsage_Work -8.137e+00  2.470e+01  -0.329   0.7419  
ClaimsFreq:AvgPrice_grp4           6.405e-01  3.309e-01   1.936   0.0534 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8229 on 514 degrees of freedom
Multiple R-squared:  0.4179,	Adjusted R-squared:  0.4055 
F-statistic: 33.55 on 11 and 514 DF,  p-value: < 2.2e-16
```

MODEL.1 - LOGLOSSRATIO

Summary of Model.1

- Adg.R_squared = 0.46, relatively speaking a good model.
- Significant predictors
    - AvgPrice grp4: there is a positive linear relationship between the group of customers owning very expensive cars ($47K-$166K) and LossRatio
    - Interpretation: For every additional individual in the zipcode who owns a car >47K, there is a 0.97 increase in the loss ratio.
    - Interaction between ClaimsFrequncy and AvgPrice_grp4
    - The effect of ClaimsFrequncy on the logLossRatio is different for different values of AvgPrice_grp4
- Non-significant predictors
    - Population,PercentFemale,PercentMarried, PercentPriVehUsage_Work
    - Importance in terms of interaction with ClaimsFrequncy.

MODEL.11 - LOGLOSSRATIO

Model.11: One standard error from MSE of Model.1

```
Call:
lm(formula = LogLossRatio ~ ClaimsFreq + PercentFemale + ClaimsFreq:PercentFemale,
    data = ZipCodeSet)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23531 -0.54384 -0.06329  0.50073  2.39174

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.5817     0.4244  -1.371    0.171
ClaimsFreq                12.9488     8.5971   1.506    0.133
PercentFemale             -1.2124     1.1556  -1.049    0.295
ClaimsFreq:PercentFemale  38.4231    23.7815   1.616    0.107

Residual standard error: 0.8232 on 522 degrees of freedom
Multiple R-squared:  0.4084,    Adjusted R-squared:  0.405
F-statistic: 120.1 on 3 and 522 DF,  p-value: < 2.2e-16
```

MODEL.2 - LOGLOSSRATIO

```
Call:
lm(formula = LogLossRatio ~ ClaimsFreq + PercentViolations +
    Population + PercentFemale + PercentMarried + PercentPriVehUsage_Work +
    AvgPrice_grp4 + ClaimsFreq:PercentViolations + ClaimsFreq:Population +
    ClaimsFreq:PercentFemale + ClaimsFreq:PercentMarried + ClaimsFreq:PercentPriVehUsage_Work +
    ClaimsFreq:AvgPrice_grp4, data = ZipCodeSet)

Residuals:
     Min       1Q   Median       3Q      Max
-2.22912 -0.54209 -0.06271  0.51552  2.32383

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -1.453e+00  1.352e+00  -1.075   0.2829
ClaimsFreq                         1.396e+01  2.704e+01   0.516   0.6060
PercentViolations                  2.329e-01  2.938e-01   0.793   0.4284
Population                         9.065e-07  5.947e-06   0.152   0.8789
PercentFemale                     -8.782e-01  1.175e+00  -0.748   0.4550
PercentMarried                     5.453e-01  1.235e+00   0.442   0.6590
PercentPriVehUsage_Work            7.689e-01  1.274e+00   0.604   0.5464
AvgPrice_grp4                     -2.887e-02  1.453e-02  -1.988   0.0474 *
ClaimsFreq:PercentViolations      -3.216e+00  5.732e+00  -0.561   0.5750
ClaimsFreq:Population              4.340e-05  1.321e-04   0.329   0.7427
ClaimsFreq:PercentFemale           3.271e+01  2.429e+01   1.346   0.1788
ClaimsFreq:PercentMarried          4.299e+00  2.495e+01   0.172   0.8633
ClaimsFreq:PercentPriVehUsage_Work -8.181e+00  2.487e+01  -0.329   0.7423
ClaimsFreq:AvgPrice_grp4           6.323e-01  3.318e-01   1.906   0.0572 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.824 on 512 degrees of freedom
Multiple R-squared:  0.4187,    Adjusted R-squared:  0.4039
F-statistic: 28.37 on 13 and 512 DF,  p-value: < 2.2e-16
```

MODEL.3 - LOGLOSSRATIO

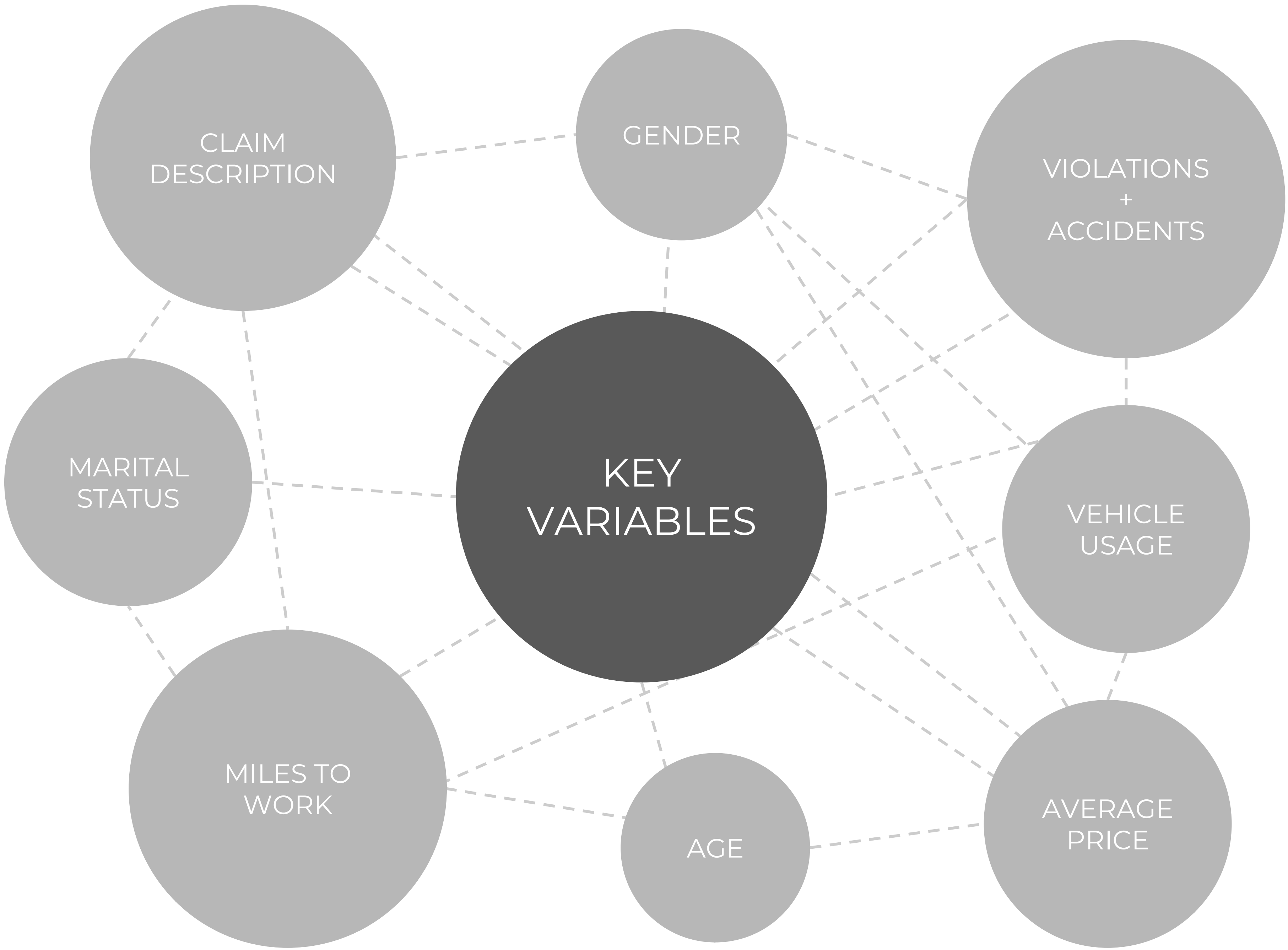
```
Call:
lm(formula = LogLossRatio ~ ClaimsFreq + PercentViolations +
    Population + PercentFemale + PercentMarried + PercentPriVehUsage_Work +
    ClaimsFreq:PercentViolations + ClaimsFreq:Population + ClaimsFreq:PercentFemale +
    ClaimsFreq:PercentMarried + ClaimsFreq:PercentPriVehUsage_Work,
    data = ZipCodeSet)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12544 -0.53675 -0.06628  0.52147  2.31240

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.860e+00  1.336e+00  -1.392    0.164
ClaimsFreq                        2.145e+01  2.684e+01   0.799    0.425
PercentViolations                 2.514e-01  2.941e-01   0.855    0.393
Population                       -3.335e-06  5.581e-06  -0.598    0.550
PercentFemale                    -1.141e+00  1.169e+00  -0.976    0.330
PercentMarried                    7.000e-01  1.235e+00   0.567    0.571
PercentPriVehUsage_Work           7.492e-01  1.276e+00   0.587    0.557
ClaimsFreq:PercentViolations     -3.293e+00  5.735e+00  -0.574    0.566
ClaimsFreq:Population             1.458e-04  1.200e-04   1.215    0.225
ClaimsFreq:PercentFemale          3.806e+01  2.420e+01   1.573    0.116
ClaimsFreq:PercentMarried         1.476e+00  2.497e+01   0.059    0.953
ClaimsFreq:PercentPriVehUsage_Work -7.793e+00  2.492e+01  -0.313    0.755

Residual standard error: 0.8257 on 514 degrees of freedom
Multiple R-squared:  0.414,     Adjusted R-squared:  0.4015
F-statistic: 33.01 on 11 and 514 DF,  p-value: < 2.2e-16
```

ANALYSIS OF PROFIT BY ZIP - MODEL.1

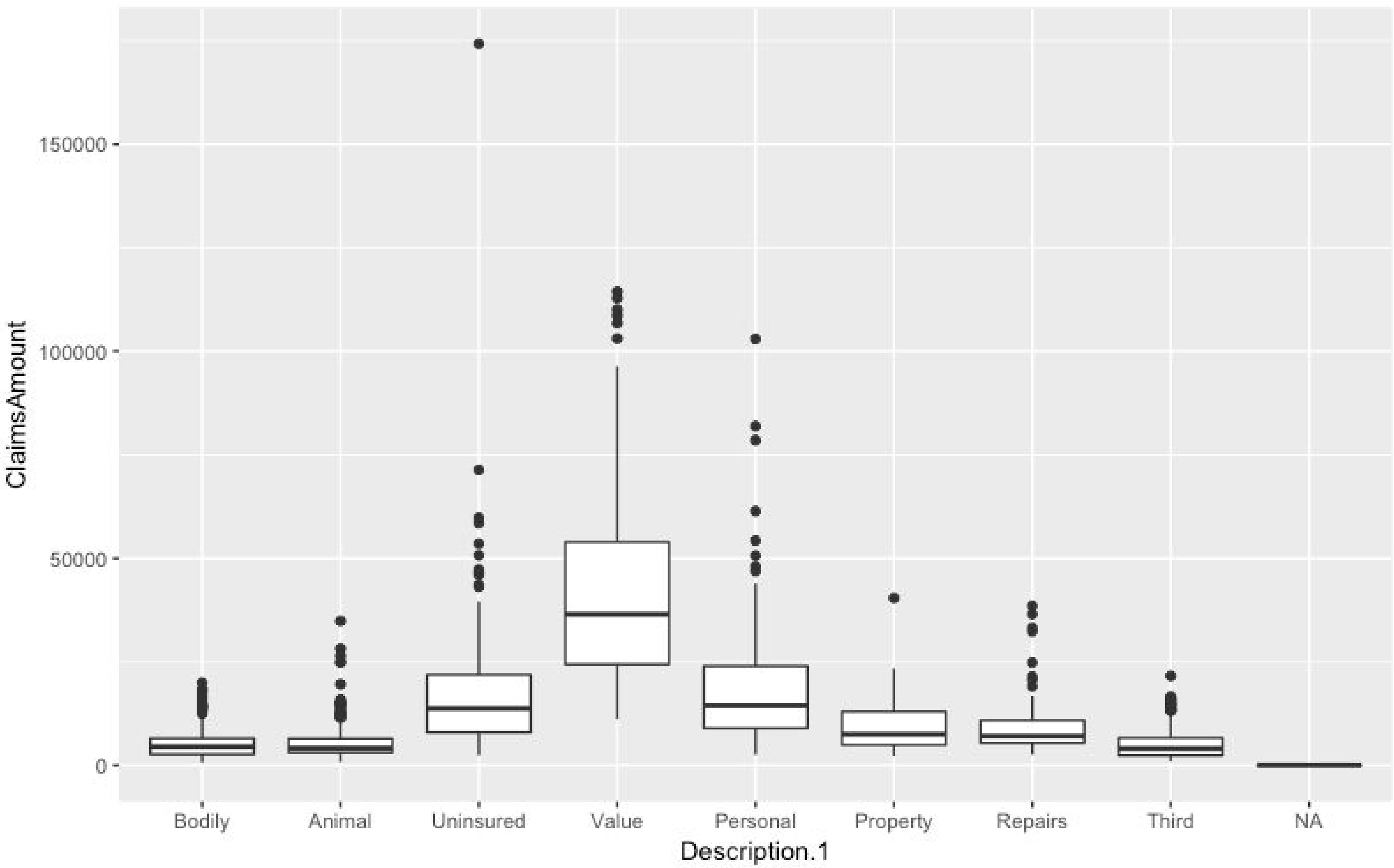
```
Call:
lm(formula = Profit ~ NumAgents + ClaimsFreq:Percent15to25 +
    PercentFemale + PercentDivorced + numberDriversOnPolicy +
    ClaimsFreq:PercentFemale + ClaimsFreq:PercentDivorced + NumAgents:numberDriversOnPolicy,
    data = ZipCodeSet)

Residuals:
    Min      1Q  Median      3Q     Max
-133250   -3564    6677   12019   47108

Coefficients: (2 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -9751.3    18401.9  -0.530   0.5964
NumAgents                           397.4      859.4   0.462   0.6440
PercentFemale                     37290.4    22026.5   1.693   0.0911 .
PercentDivorced                    4883.9    64638.4   0.076   0.9398
numberDriversOnPolicy                  NA         NA      NA       NA
ClaimsFreq:Percent15to25         -99543.7   284220.2  -0.350   0.7263
ClaimsFreq:PercentFemale       -1425476.7   361837.1  -3.940 9.28e-05 ***
ClaimsFreq:PercentDivorced       131945.2  1292135.6   0.102   0.9187
NumAgents:numberDriversOnPolicy        NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23760 on 519 degrees of freedom
Multiple R-squared:  0.2586,    Adjusted R-squared:   0.25
F-statistic: 30.17 on 6 and 519 DF,  p-value: < 2.2e-16
```

ANALYSIS OF INV-PROFIT BY ZIP - MODEL.2

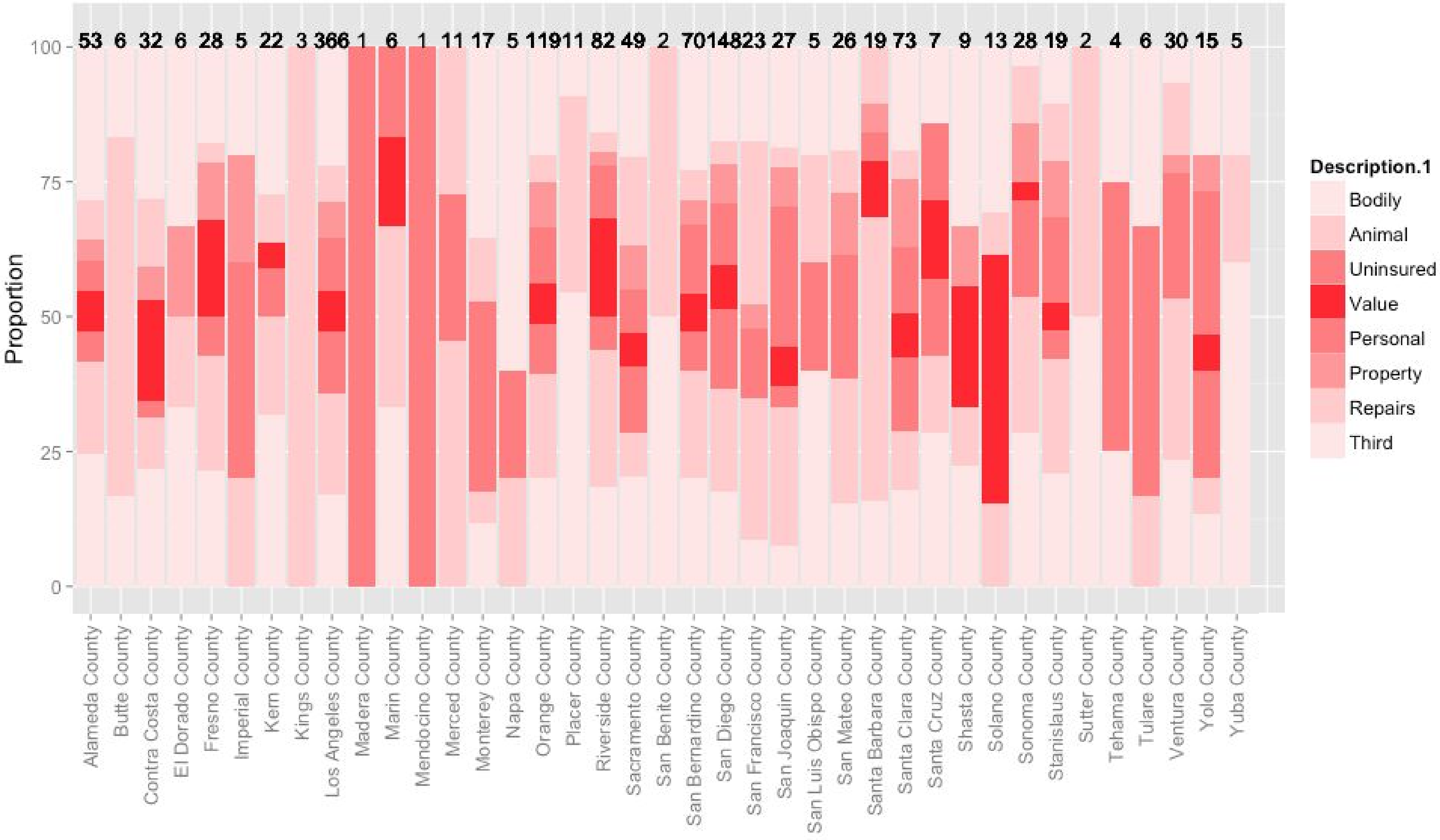
```
Call:
lm(formula = InvProfit ~ NumAgents + ClaimsFreq + PercentFemale +
    PercentDivorced + +ClaimsFreq:PercentFemale + ClaimsFreq:PercentDivorced +
    NumAgents:numberDriversOnPolicy + ClaimsFreq:Percent15to25,
    data = ZipCodeSet)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0089501 -0.0000590  0.0000354  0.0001600  0.0043869

Coefficients: (1 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.129e-05  6.214e-04  -0.018   0.9855
NumAgents                       -5.484e-06  2.610e-05  -0.210   0.8337
ClaimsFreq                      -3.823e-03  8.365e-03  -0.457   0.6478
PercentFemale                    1.071e-03  1.015e-03   1.055   0.2917
PercentDivorced                 -4.515e-03  1.995e-03  -2.263   0.0241 *
ClaimsFreq:PercentFemale        -7.507e-03  2.094e-02  -0.359   0.7201
ClaimsFreq:PercentDivorced       6.310e-02  4.029e-02   1.566   0.1179
NumAgents:numberDriversOnPolicy         NA         NA      NA       NA
ClaimsFreq:Percent15to25         4.414e-03  9.281e-03   0.476   0.6345
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0007203 on 518 degrees of freedom
Multiple R-squared:  0.01761,   Adjusted R-squared:  0.004334
F-statistic: 1.326 on 7 and 518 DF,  p-value: 0.2354
```


CLAIM DESCRIPTION
GENDER
VIOLATIONS + ACCIDENTS
MARITAL STATUS
KEY VARIABLES
VEHICLE USAGE
MILES TO WORK
AGE
AVERAGE PRICE

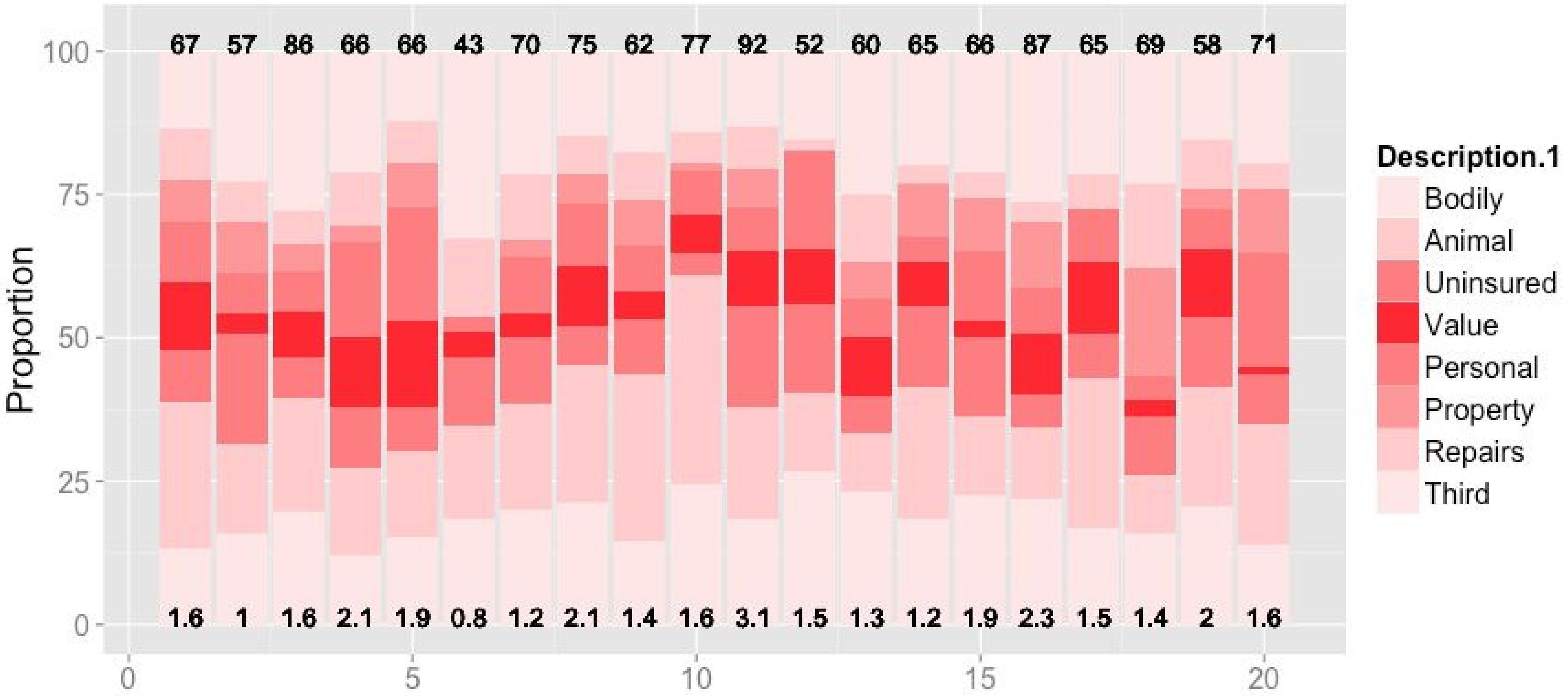
CLAIM DESCRIPTION VS. AMOUNT



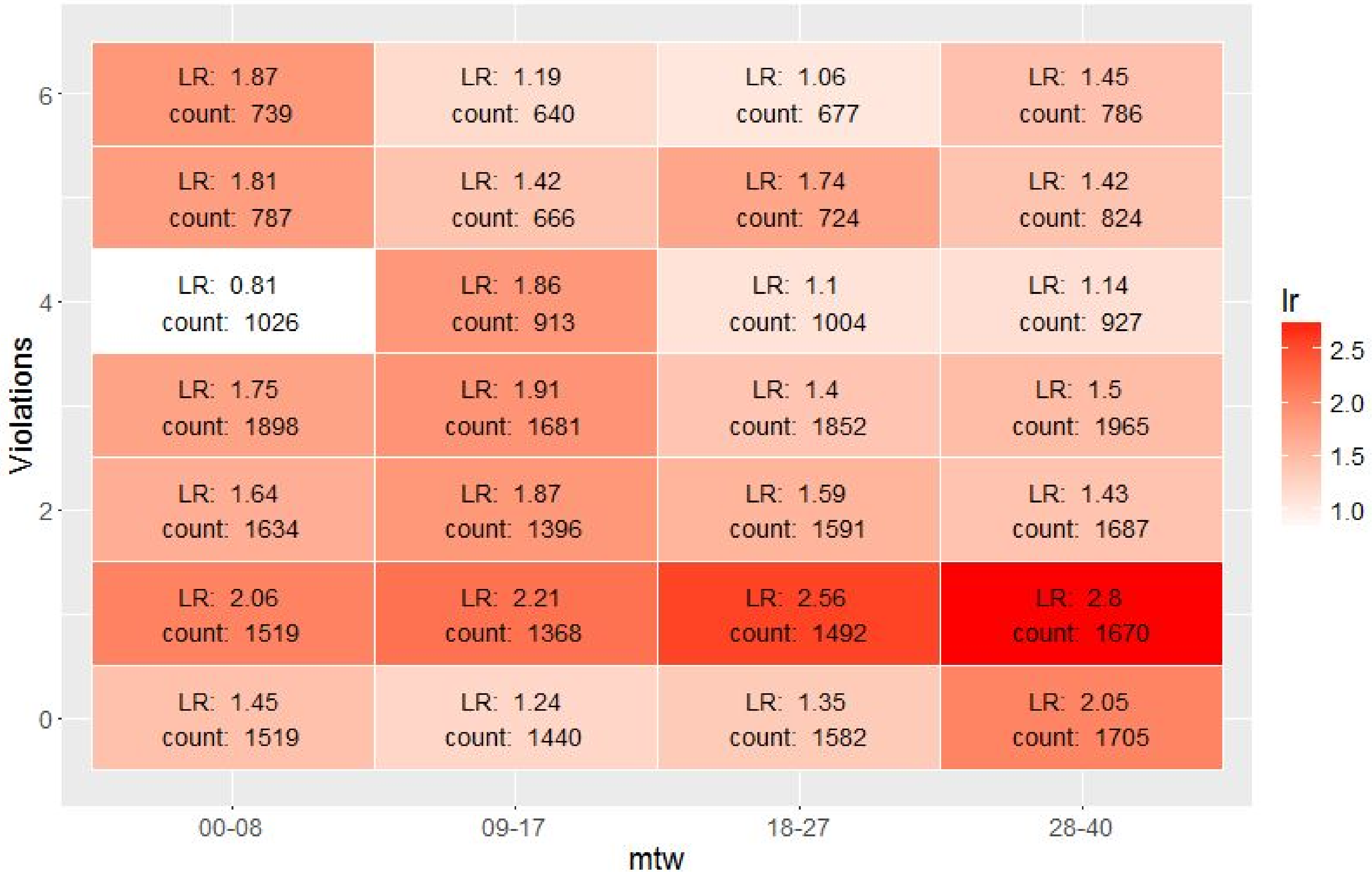
COUNTY VS. CLAIM DESCRIPTION



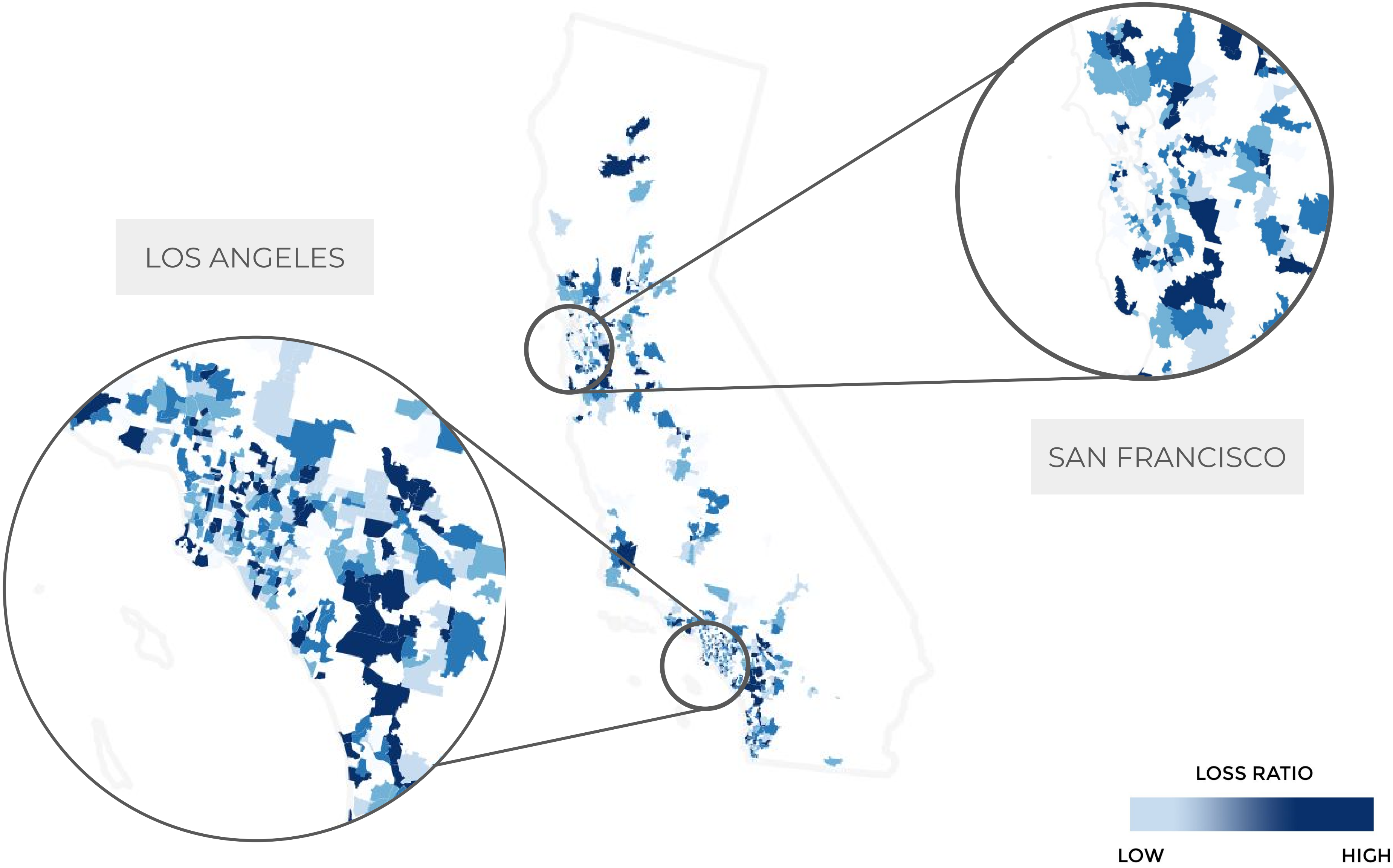
AGENCY VS. CLAIM DESCRIPTION



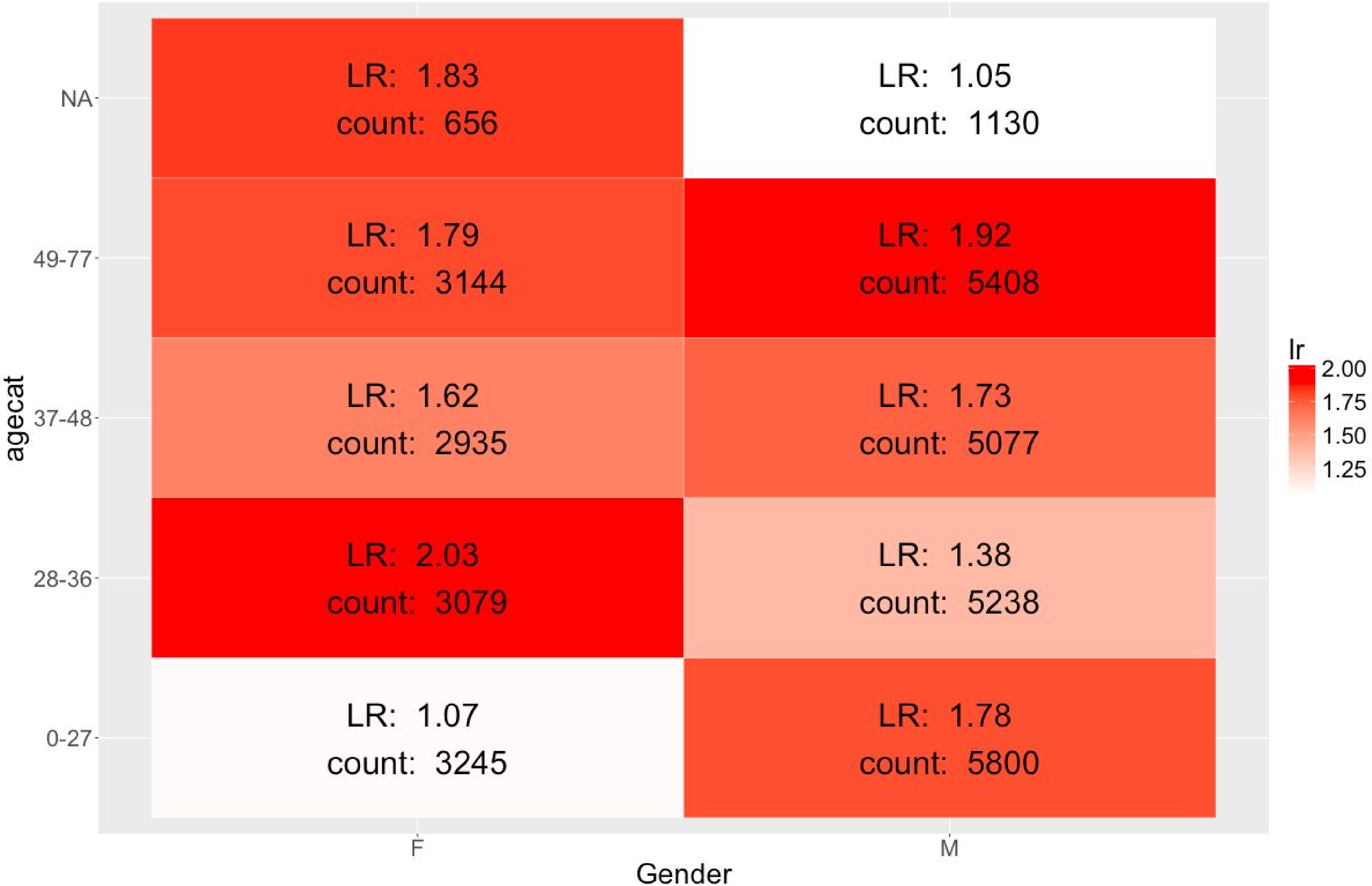
MILES TO WORK VS. VIOLATIONS



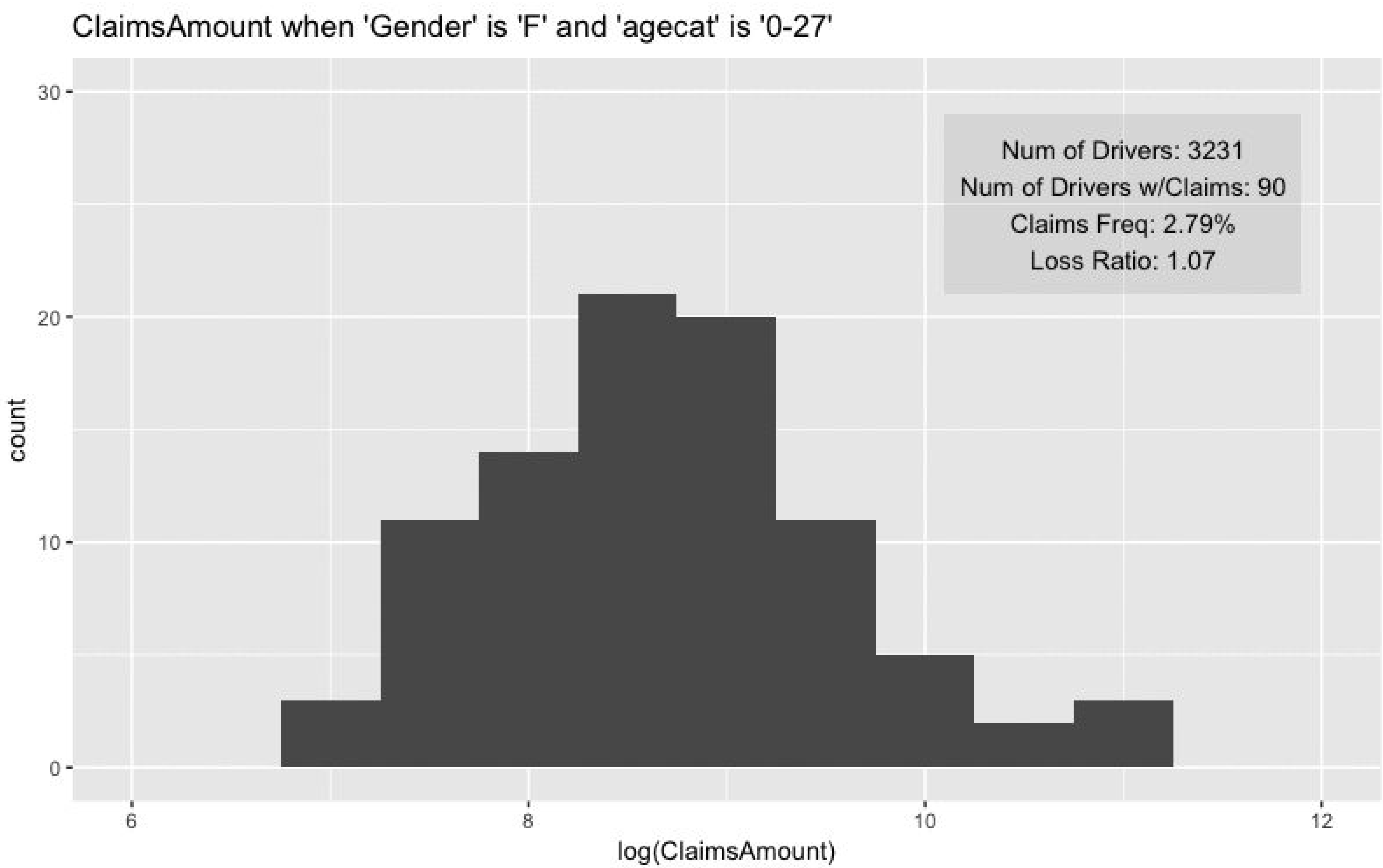
SPATIAL VISUALIZATION



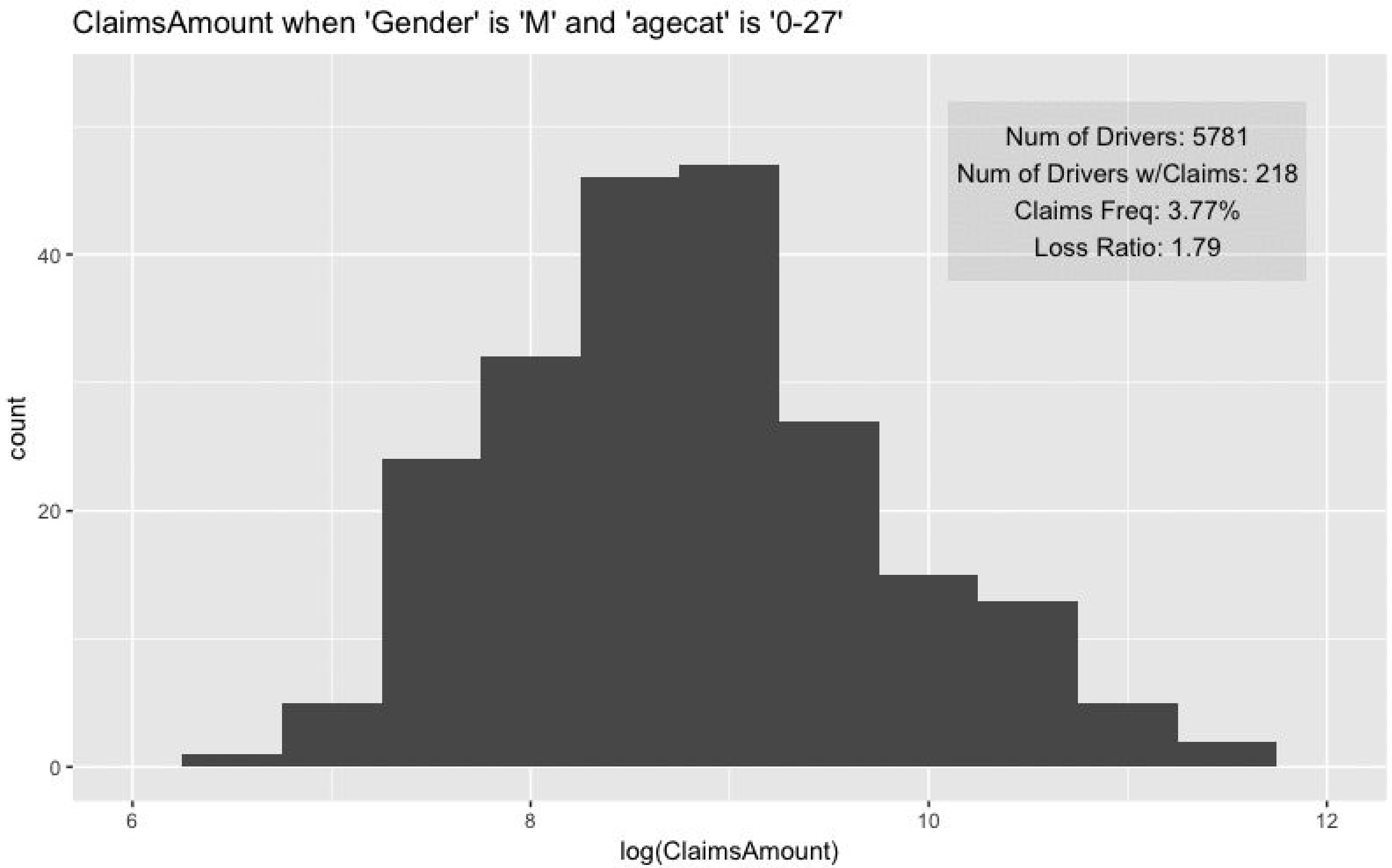
GENDER VS. AGE



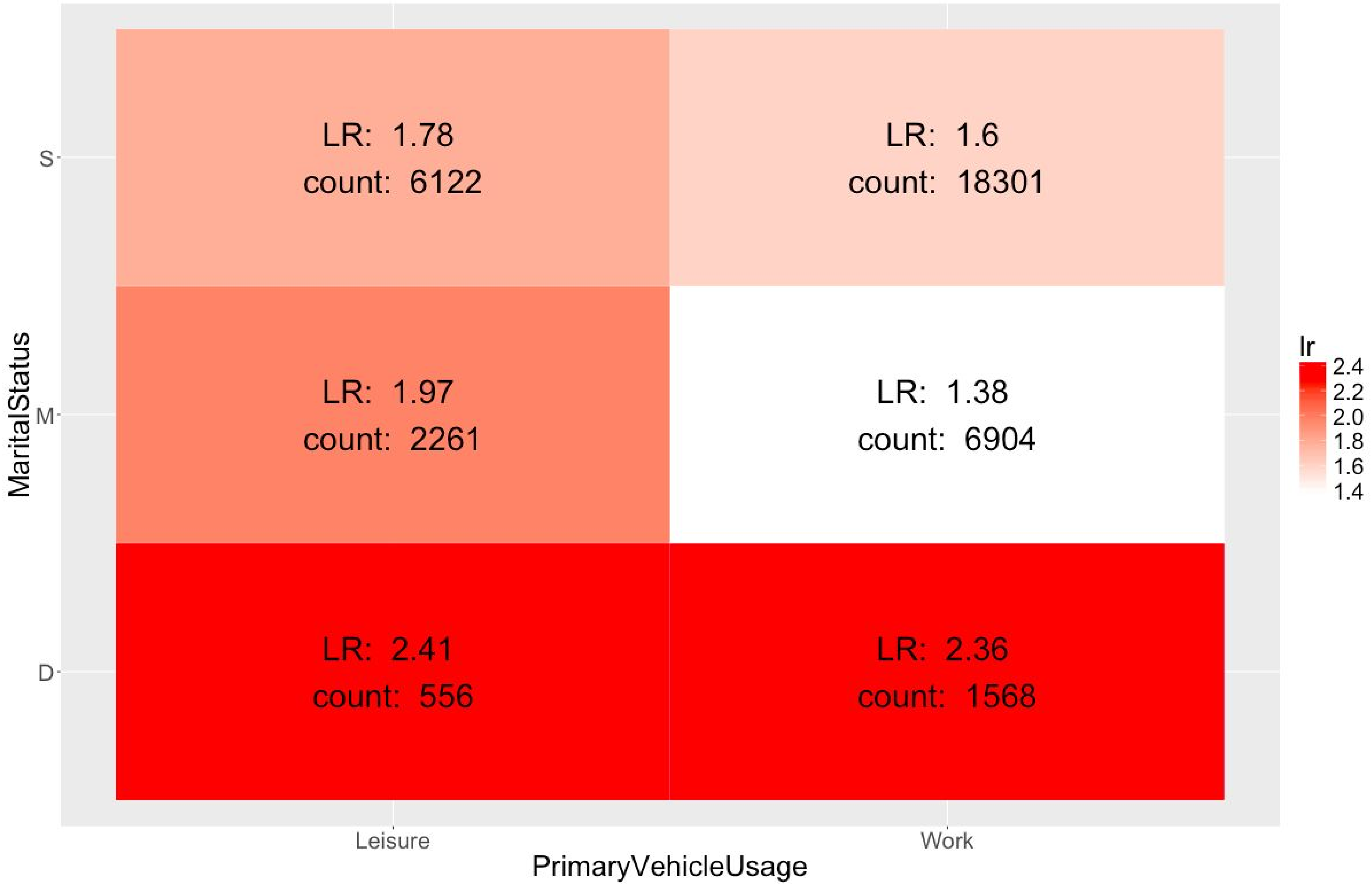
CLAIMS AMOUNT VS. GENDER



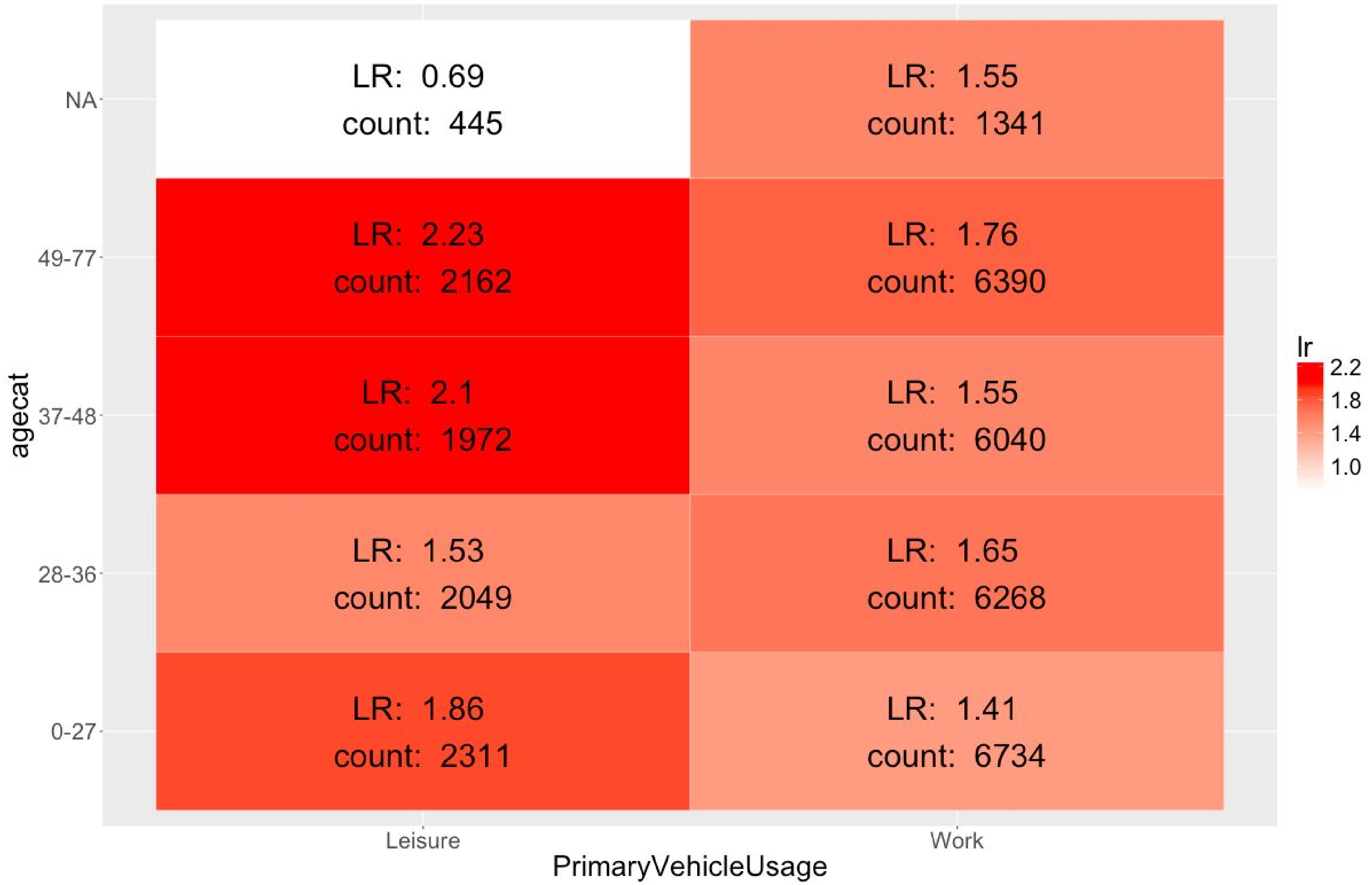
CLAIMS AMOUNT VS. GENDER



MARRIED VS. VEHICLE USAGE



VEHICLE USAGE VS. AGE



CONTINUE TO SEEK OUT:

- Females of 0-27 years
- Married people who work
- Zero violations of 0-27 years
- Males with no birthday
- Agencies “Auto AA Insurance” and “Clovis Insurance Agency”

RECOMMENDATIONS

BE WEARY OF:

- All (especially young) customers with expensive cars
- Divorced people
- Customers with one violation (especially those who drive further to work)
- Females aged 28-36
- Agency “Yugine C Sport Insurance”

RECOMMENDATIONS

HOW WE WOULD CHANGE OUR APPROACH

- We expected more obvious results

- Complex data allowed for a challenging problem but also a challenging analysis

- Distribution for the claim amount with low number of claims proved challenging for the techniques we know

- We would focus on more simple modelling techniques and more extensive exploratory analysis

- Focusing on interactions earlier on would be useful

SUGGESTIONS FOR IMPROVING THIS EXPERIENCE

- Have a predictive element of the capstone
- More data about the claims
- Examples of previous or sample analysis
- Helping students avoid common mistakes (ex. summing loss ratios across individuals vs. looking at the loss ratio of a given group)
- Be able to give out the formula for how the insurance company is currently pricing without giving away any hints
- It is unlikely that an insurance company will not have birthday information

REFLECTION ON EXPERIENCE

??

QUESTIONS

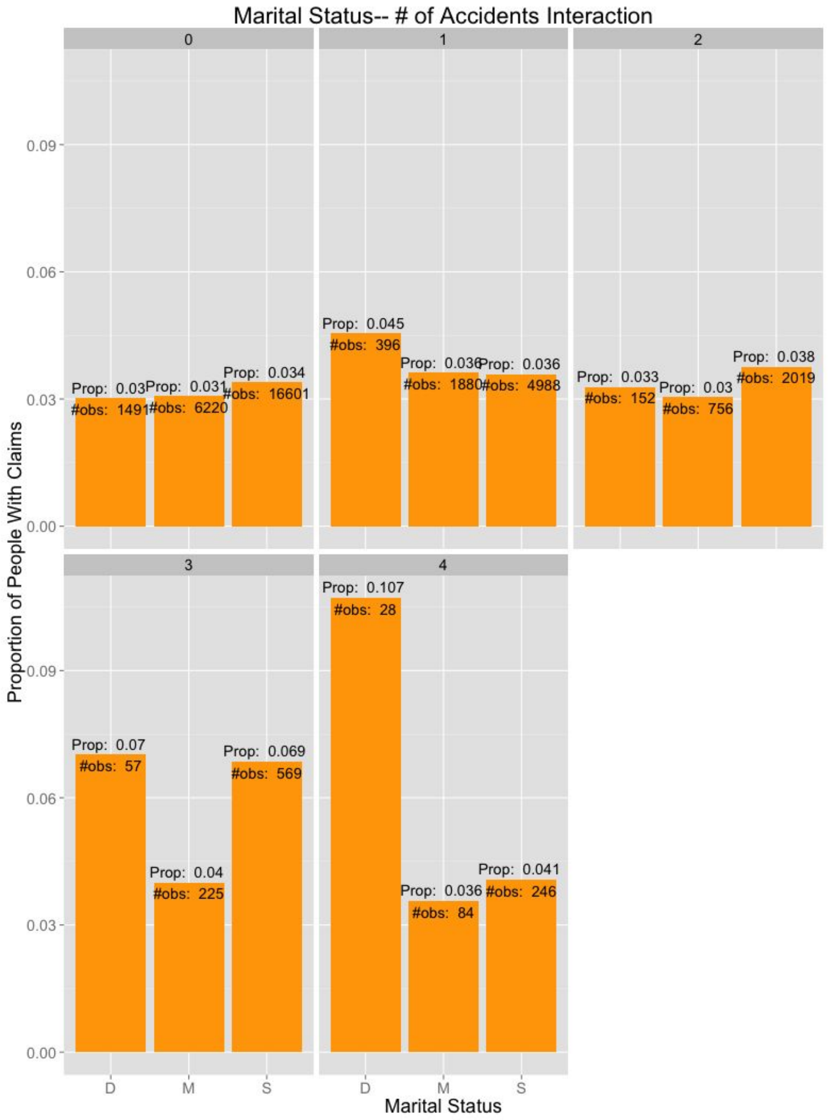
pwc

THE OHIO STATE UNIVERSITY

UNDERGRADUATE
DATA ANALYTICS MAJOR

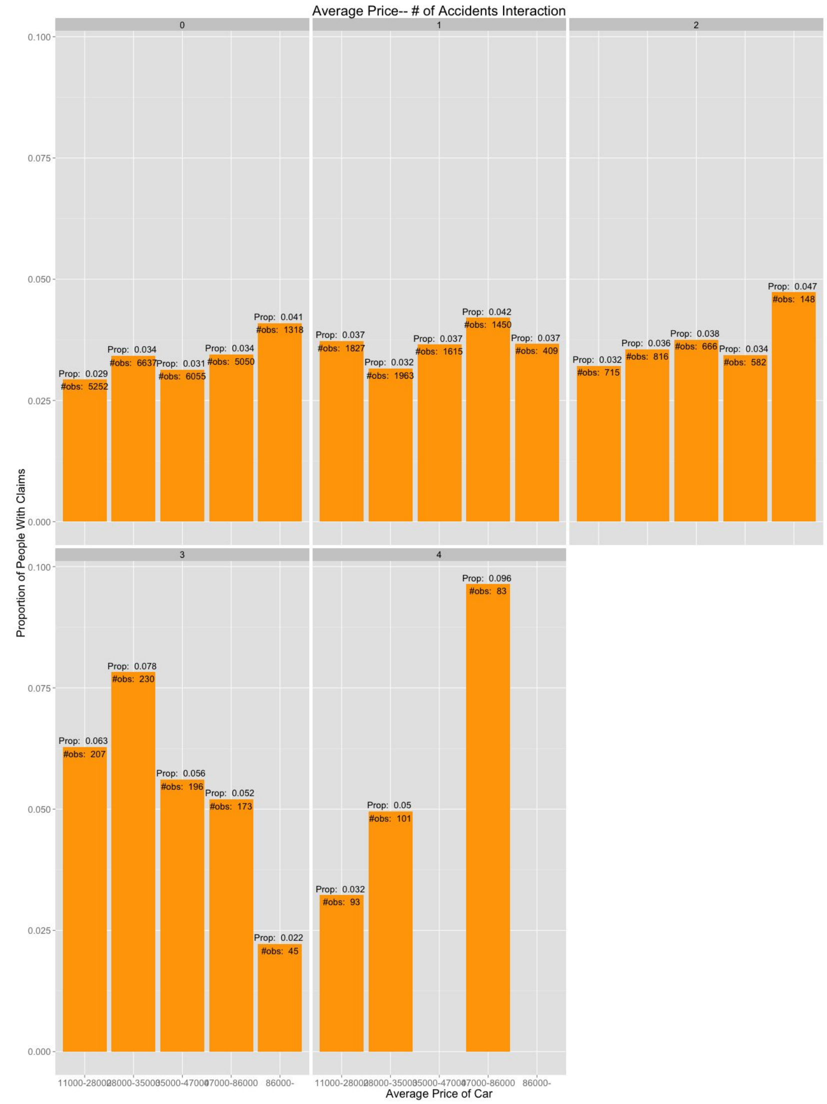
THANK YOU FOR A GREAT EXPERIENCE

Hypothesis: Conditional on an individual having many accidents, being divorced or single seems particularly strongly associated with an individual having a claim



JUSTIFICATION FOR CLAIM 1 (SLIDE 18)

Hypothesis: <u>Conditional on an individual having fewer accidents</u>, the higher the price of the vehicle, the more likely an individual will have a claim [Does not hold for higher accidents]



JUSTIFICATION FOR CLAIM 2 (SLIDE 18)