Predicting Purchased Insurance Options

Chris Li

Problem Overview

- Goal: Predict the purchased coverage options using a limited subset of the total interaction history
- Each product has 7 options, each with 2, 3, or 4 ordinal possible values
- Cost of product is related to product options and customer characteristics

Data

- 52011 customers, 25 variables
- Each customer has 3-13 entries
- Customer characteristics: location, group size, homeowner, risk factor, age of oldest/youngest person in the group, married, previous option C value, duration of previous coverage
- Car data: car age, car value
- Product data: which option was viewed, cost

Clean Data

- Day: Weekday, Weekend
- Time: Morning (6am-11am), Afternoon (12pm-5pm), Evening (6pm-11pm), Night(12am-5am)

Predict Missing Risk Factor

- 19963 customers had missing Risk Factor, 1570 of them filled risk factor before purchasing
- Predict risk factor with all non-missing entries at purchase point
- Use predicted value for all missing entries resembles the actual situation best, as insurance company used predicted information to calculate cost

Predict Missing Previous Coverage Information

- 3325 customers missed PrevC and PrevDuration at some point in their shopping history

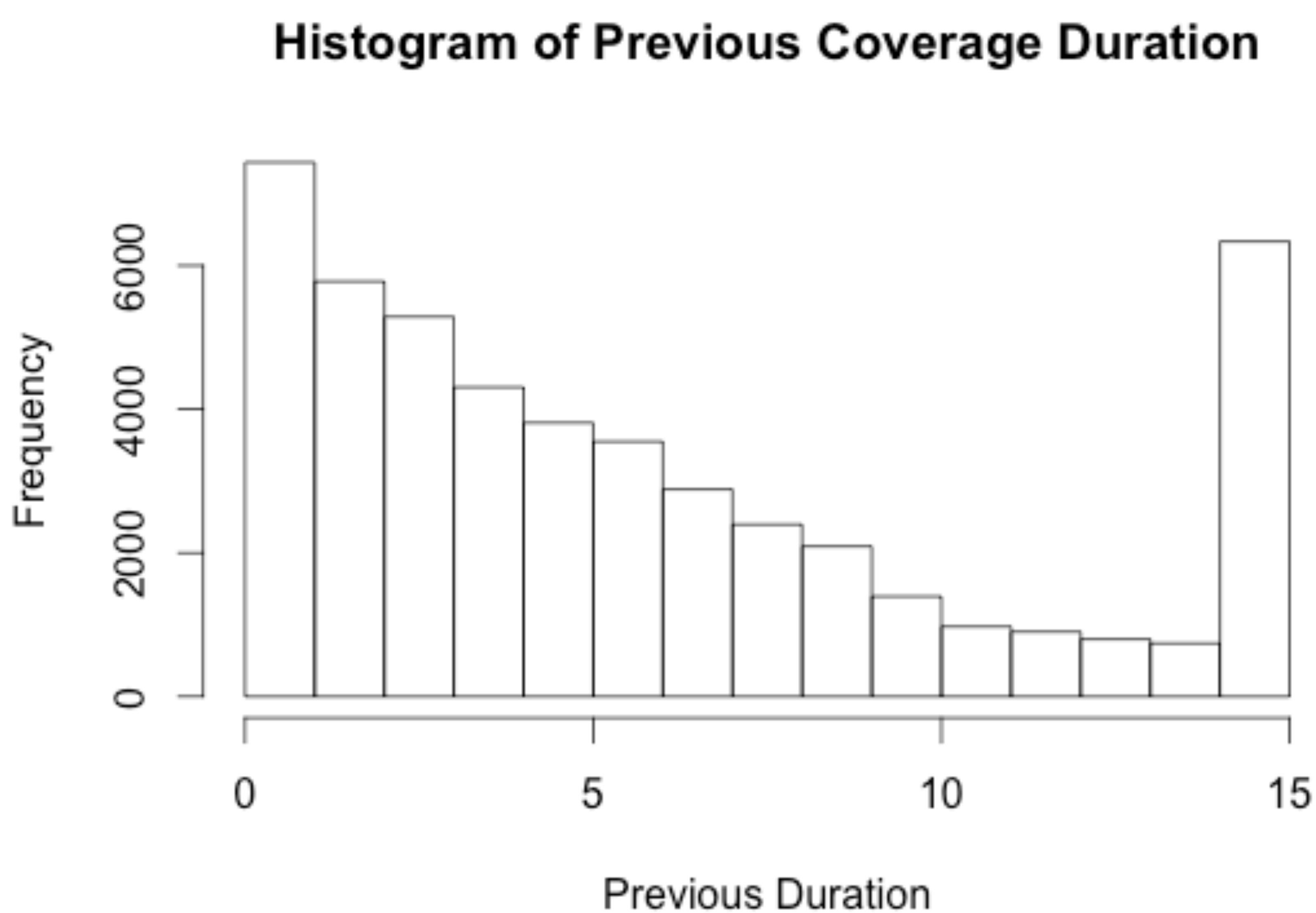
- 444 customers didn’t have this information at the time of purchase

- Assume: there was reason that they didn’t want to reveal their previous coverage information—the distributions of PrevC and PrevDuration for these 3325 customers were different from those for other customers

- T test for PrevDuration indicates the two groups are significantly different ($p<2.2e-16$). People who revealed their previous coverage information at the beginning on average had 1.47 years longer coverage duration than people who didn’t

- Geometric distribution with right censored data (censored at PrevDuration=15)

- Fit a geometric distribution with people who didn’t reveal previous coverage information at first, then predict for missing entries



- Missing Previous C information: multinomial regression with people who didn’t reveal their previous coverage of option C. Then predict.

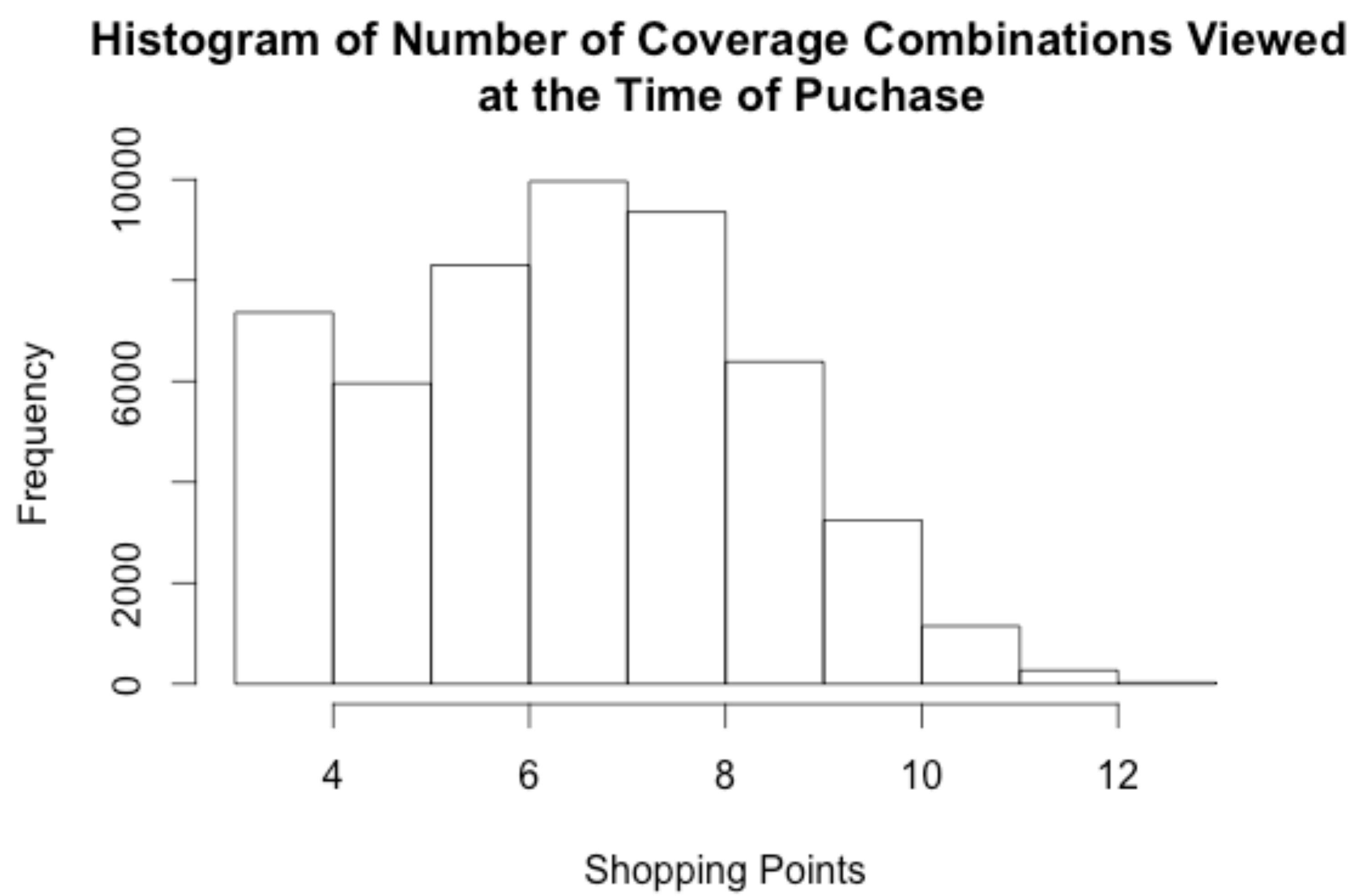
- Missing Car Value: multinomial regression with all non-missing entries (only 98 missing)

Add New Variables

- Age difference: oldest age in group - youngest age
    - To deal with problem of one person group with different ages
- Family: At least two group members, age difference greater than 15, married member
- Couple: Two group members, age difference less than 15, married
- Individual: One person group

- People viewed at least 3 combinations before making purchase

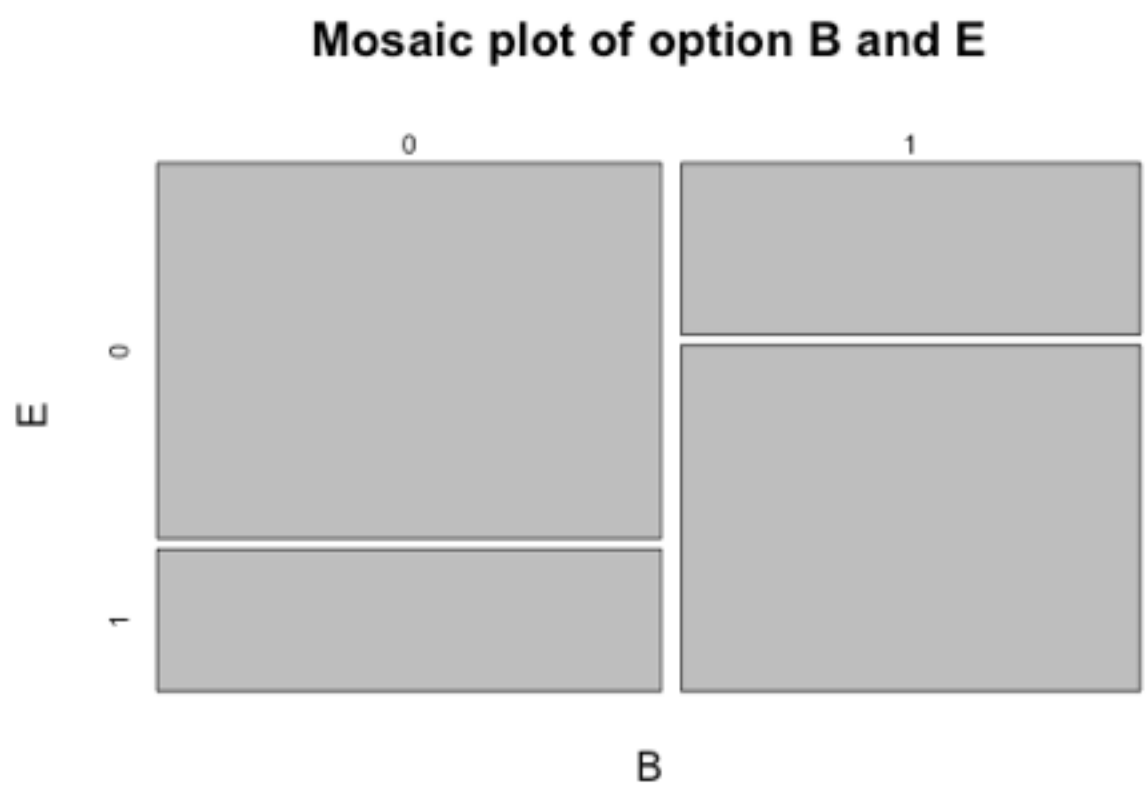
- Can create several subsets based on shopping points: first view, second last view, last view, and purchase

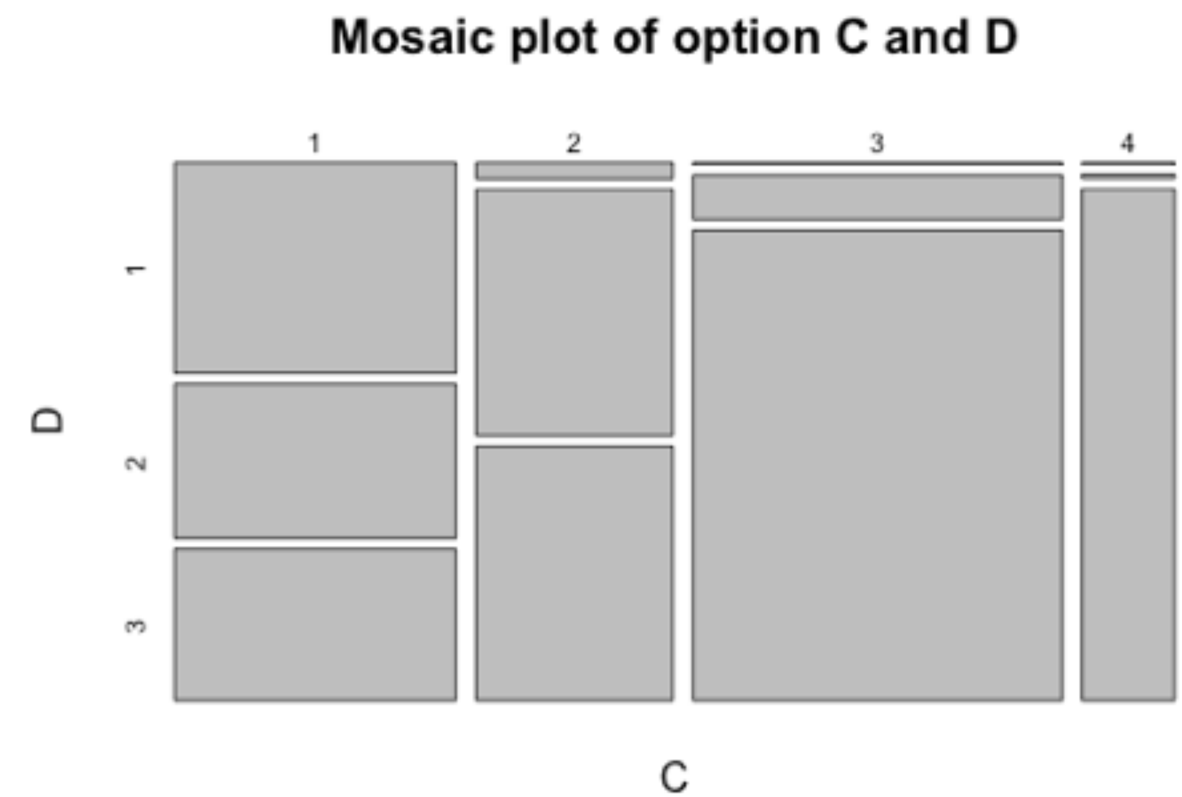


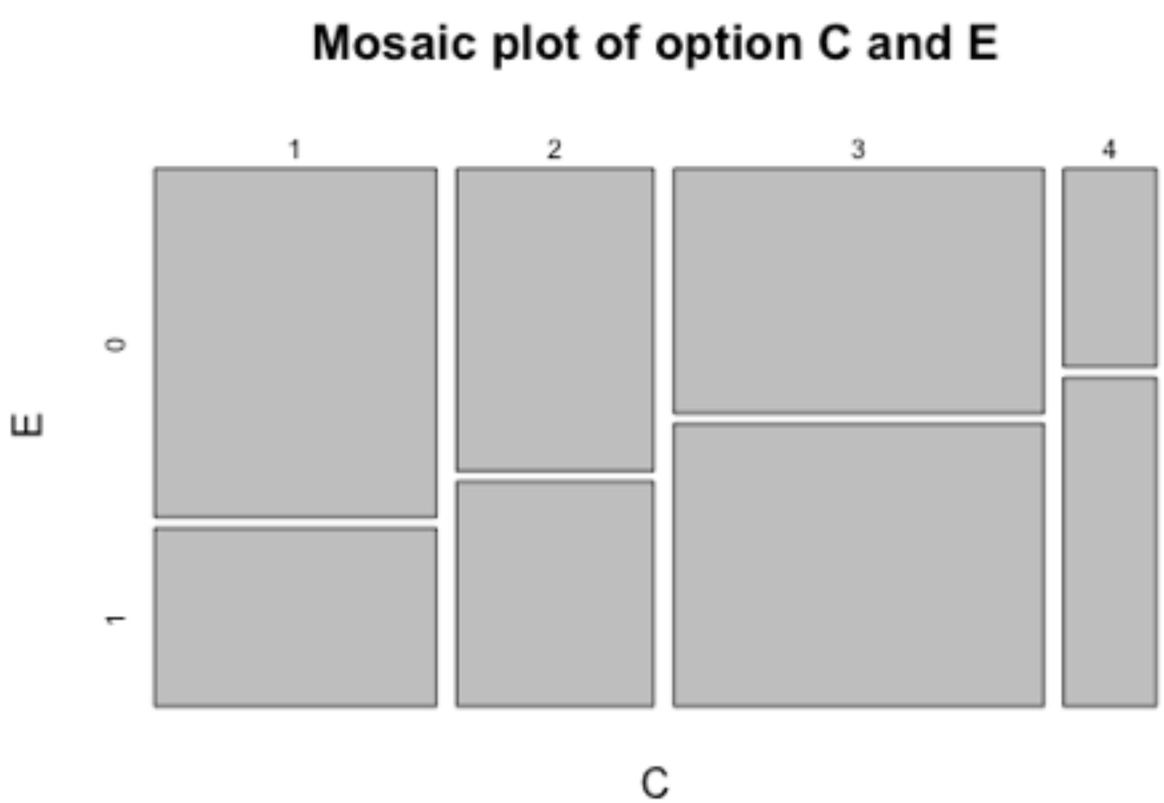
Options by State

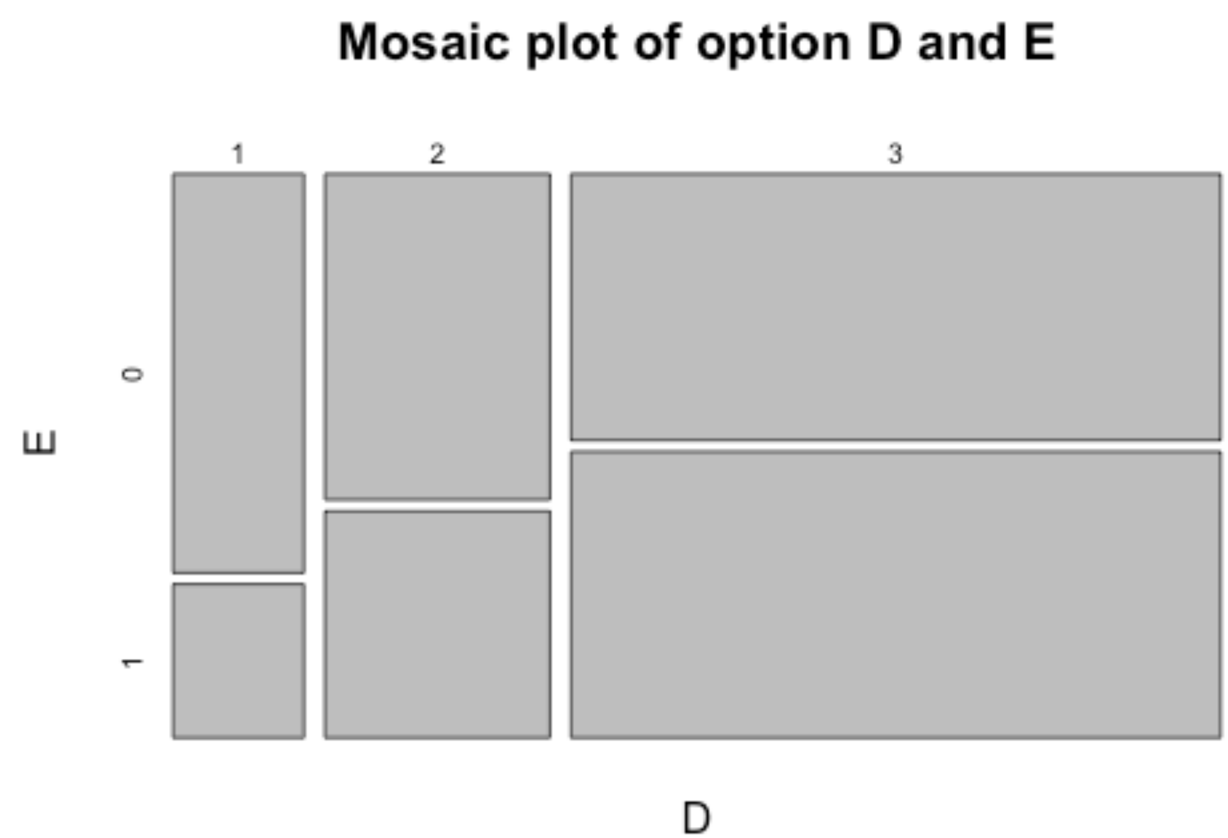
- Option C = 1 was not available in Georgia or Maine
- Option D =1 was not available in Georgia
- Option G =2 was the only option in North Dakota and South Dakota
- Option G = 1 or 2 was not available in Florida, G = 3 is more likely than G = 4
- Option G = 1 was not available in Ohio

Relationship between Options

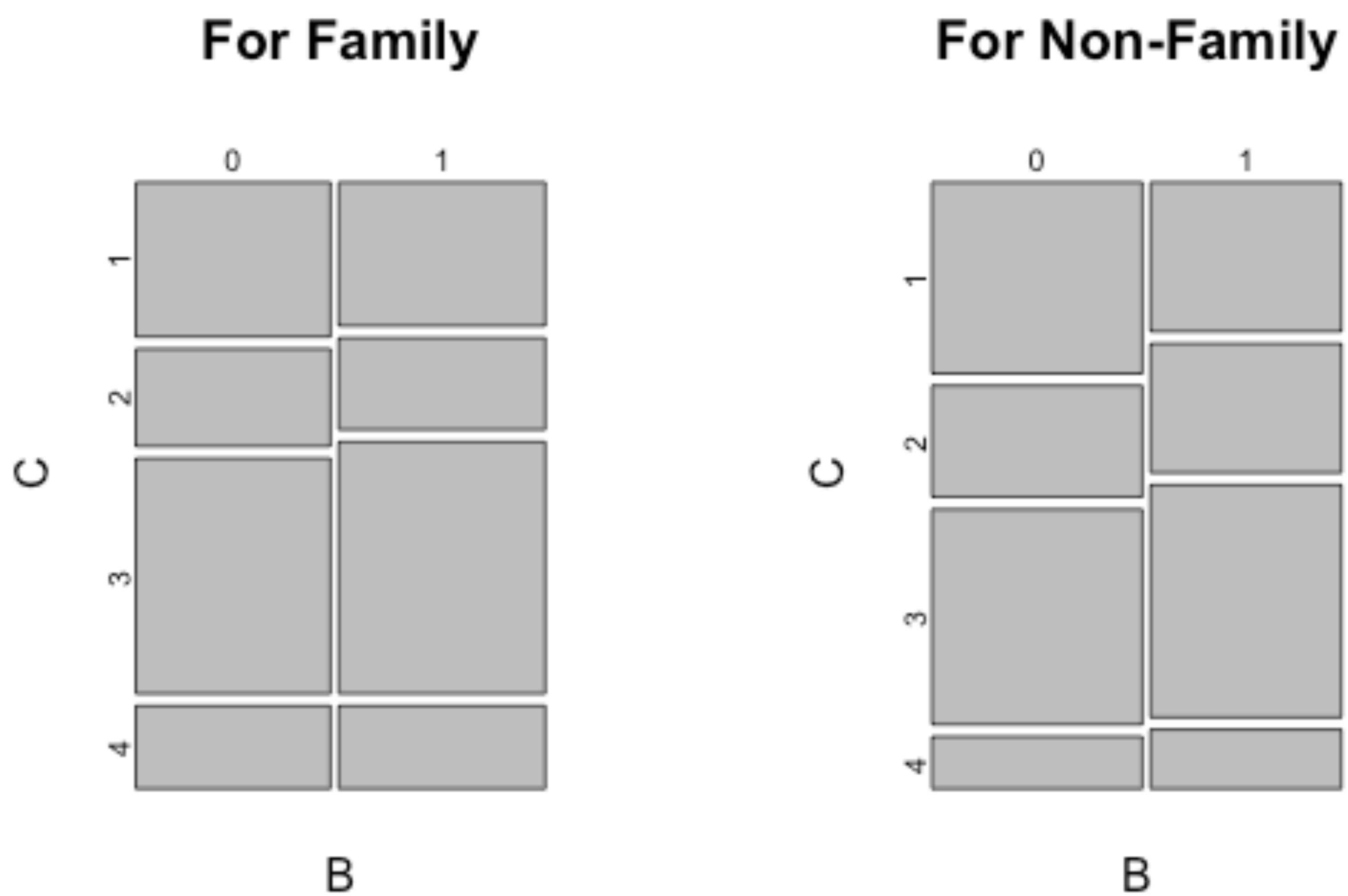




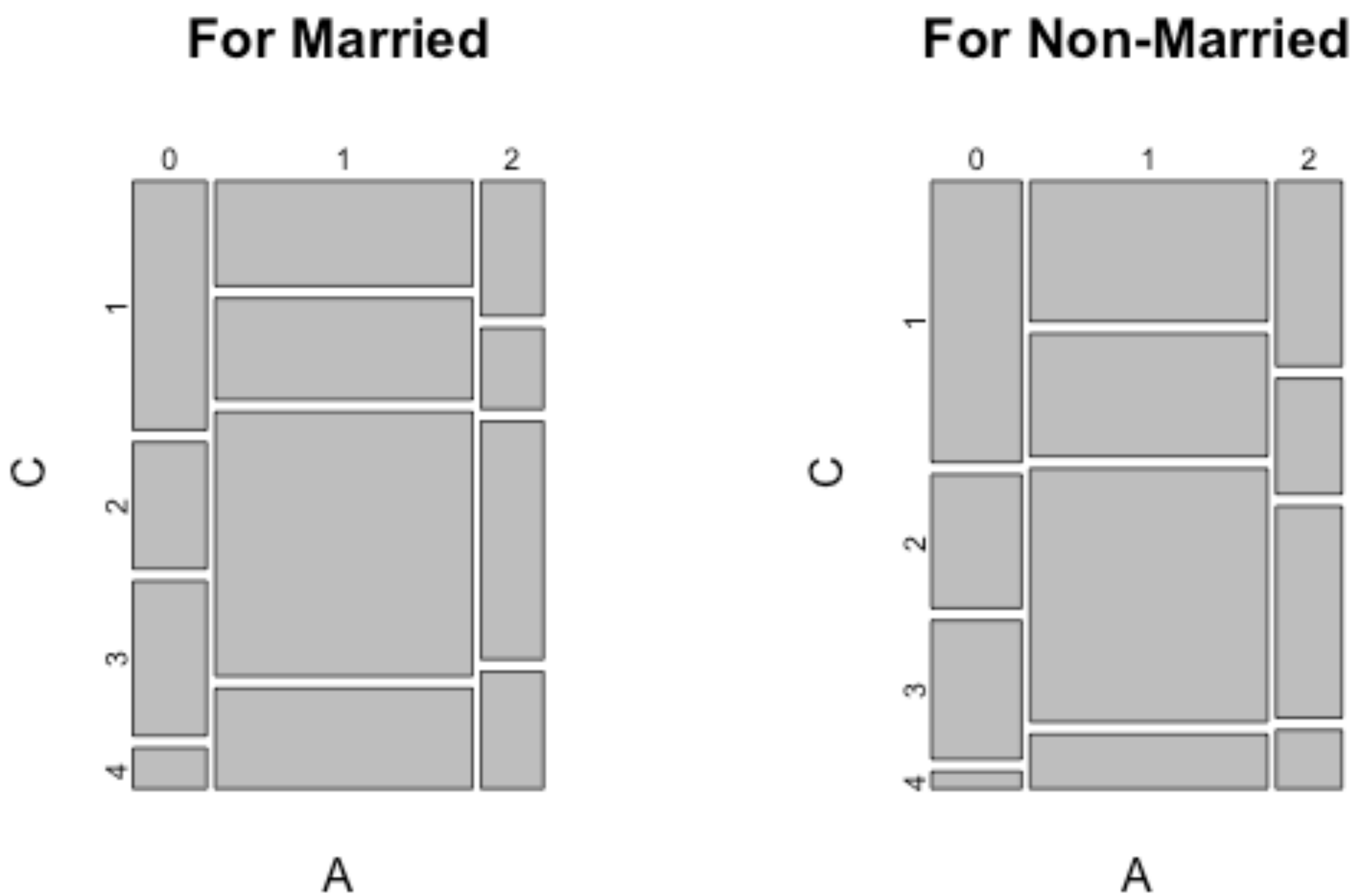




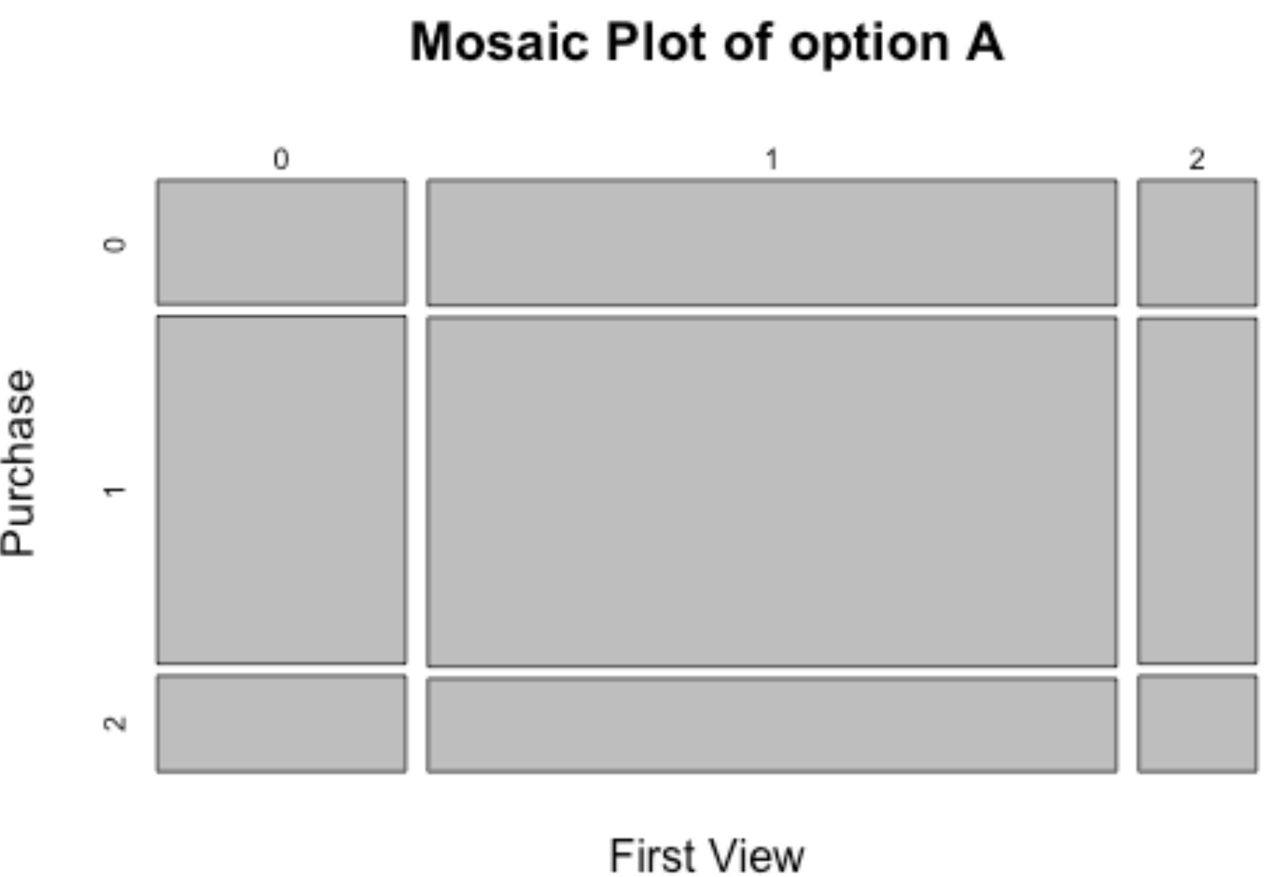
Relationship given Family

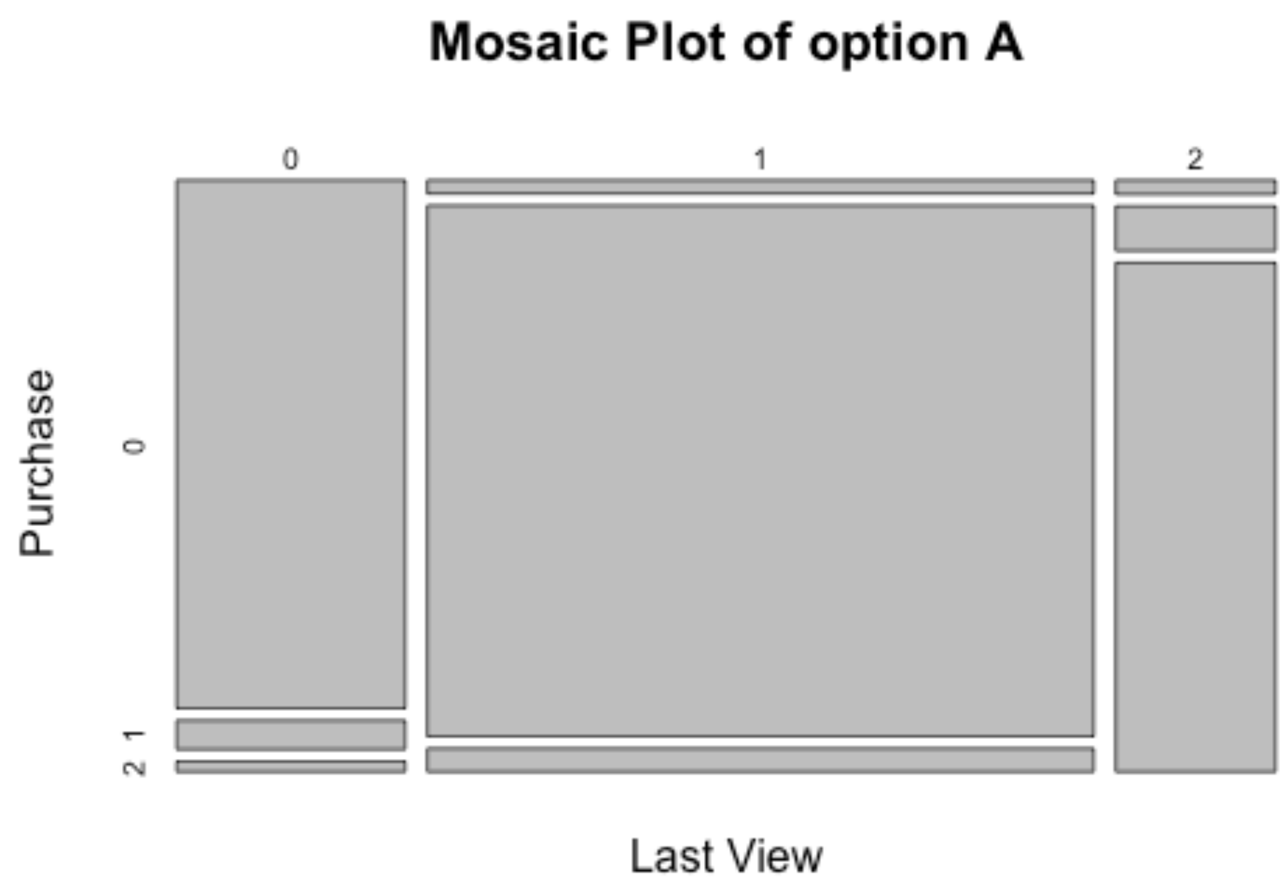


Relationship given If Married

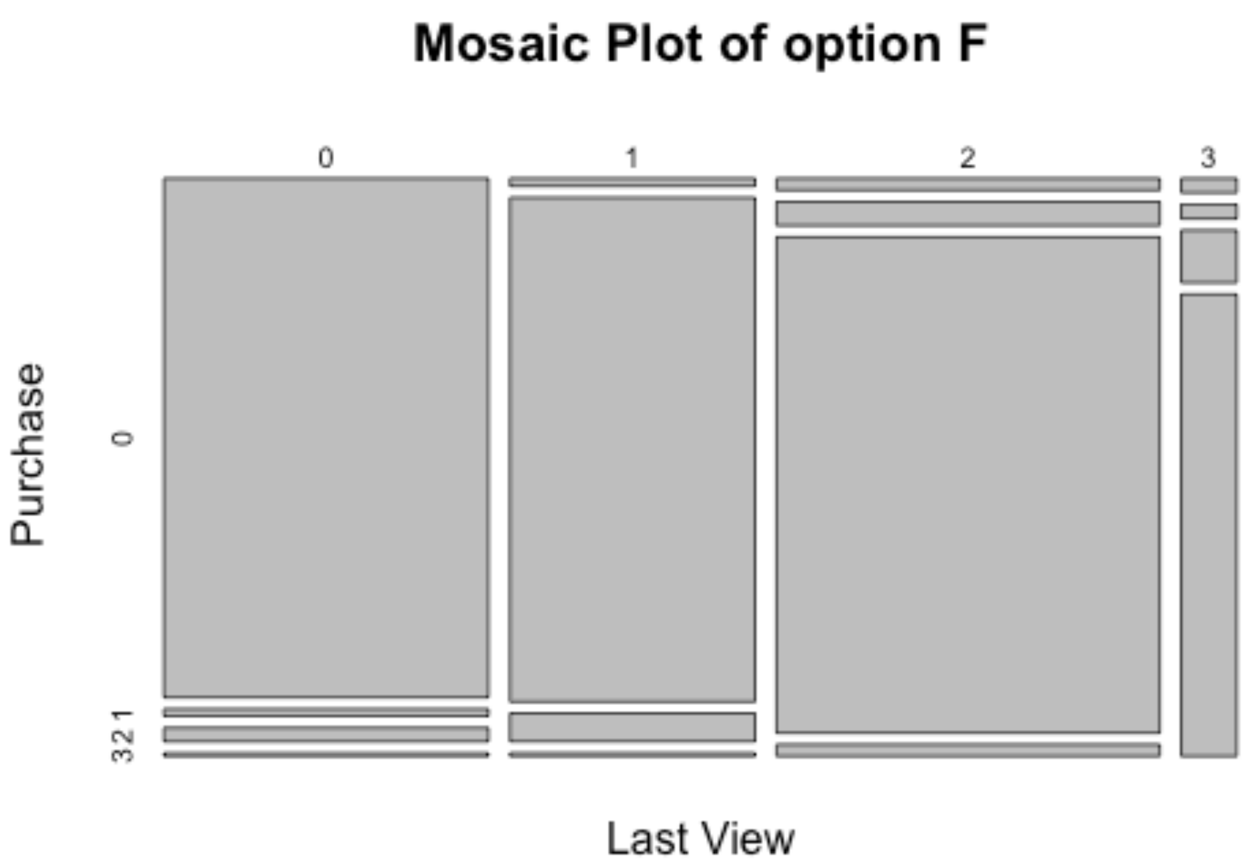


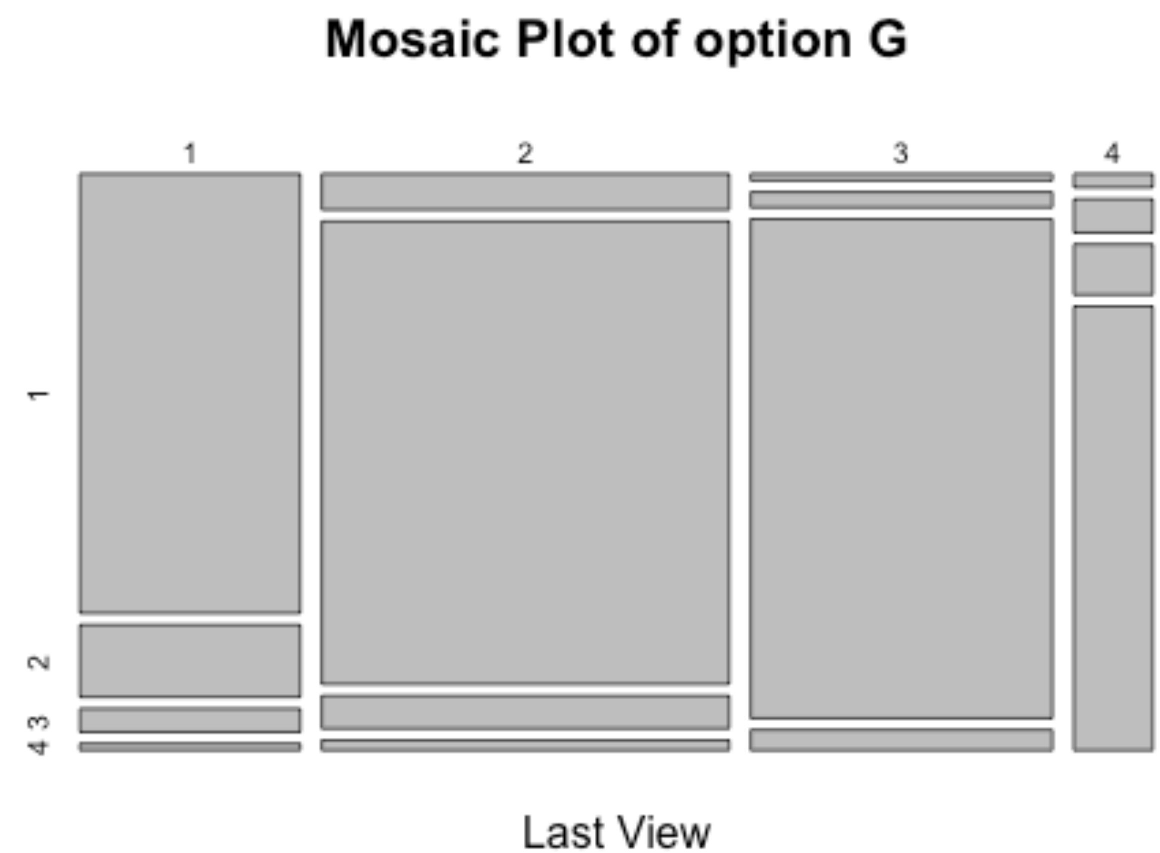
Customers were highly likely to buy the option they viewed last



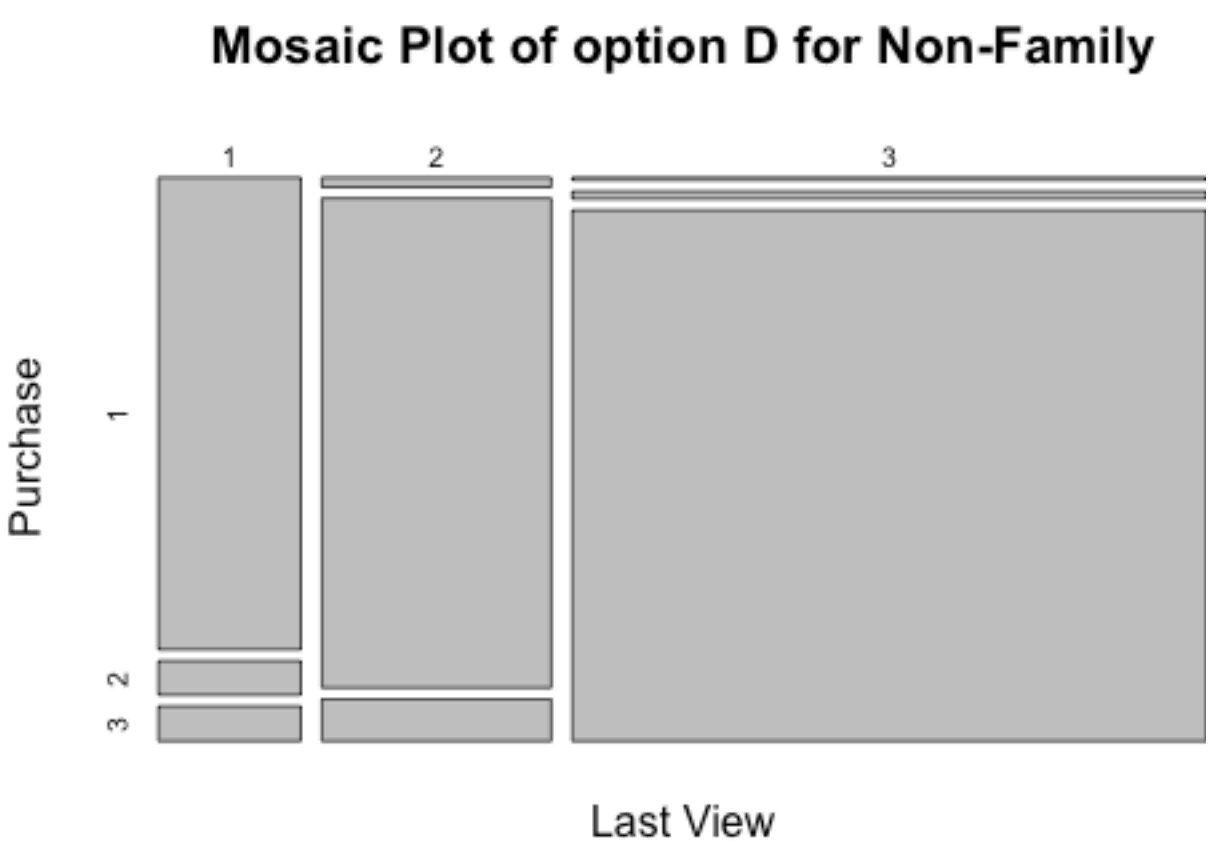


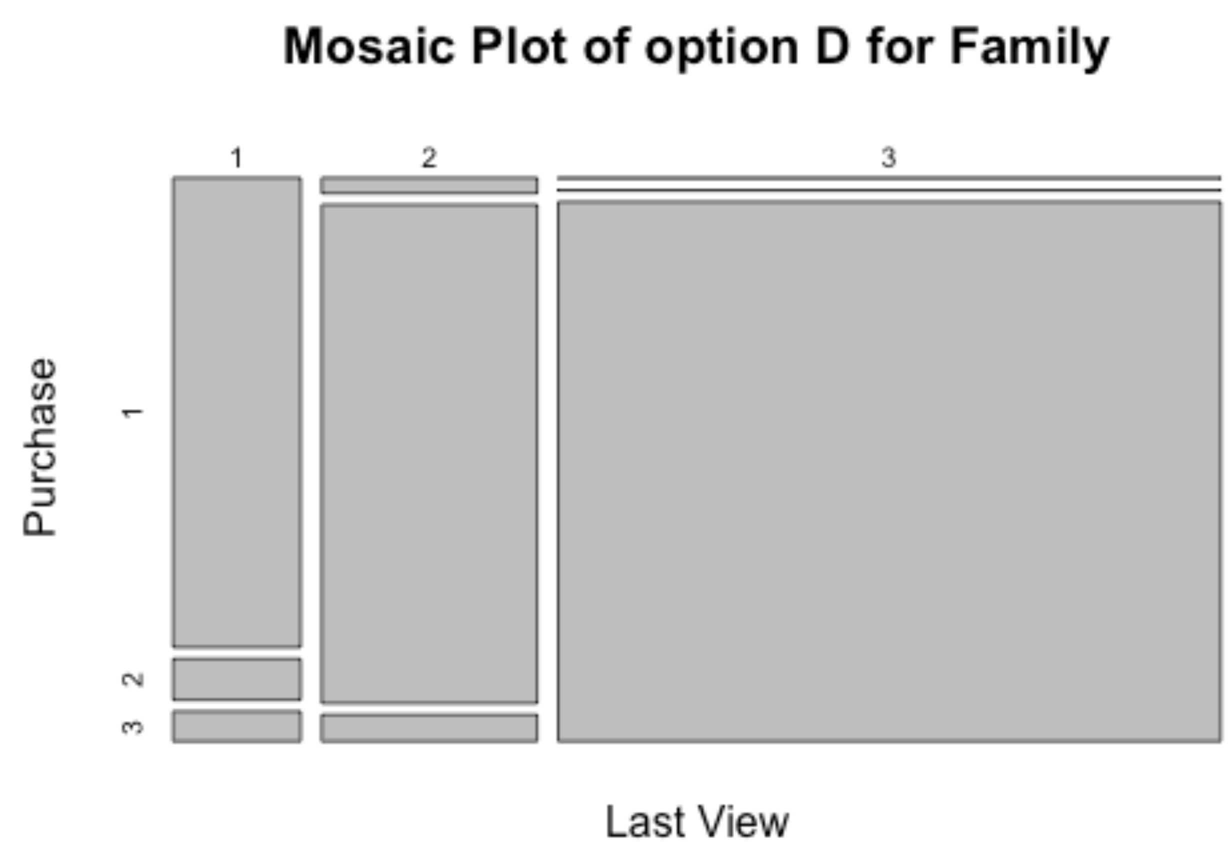
…except for option G





Other Variables Also have Effect on What People Chose





Simplest Model

- Use last viewed information to predict final purchase

- Completely correct rate is 70.6%

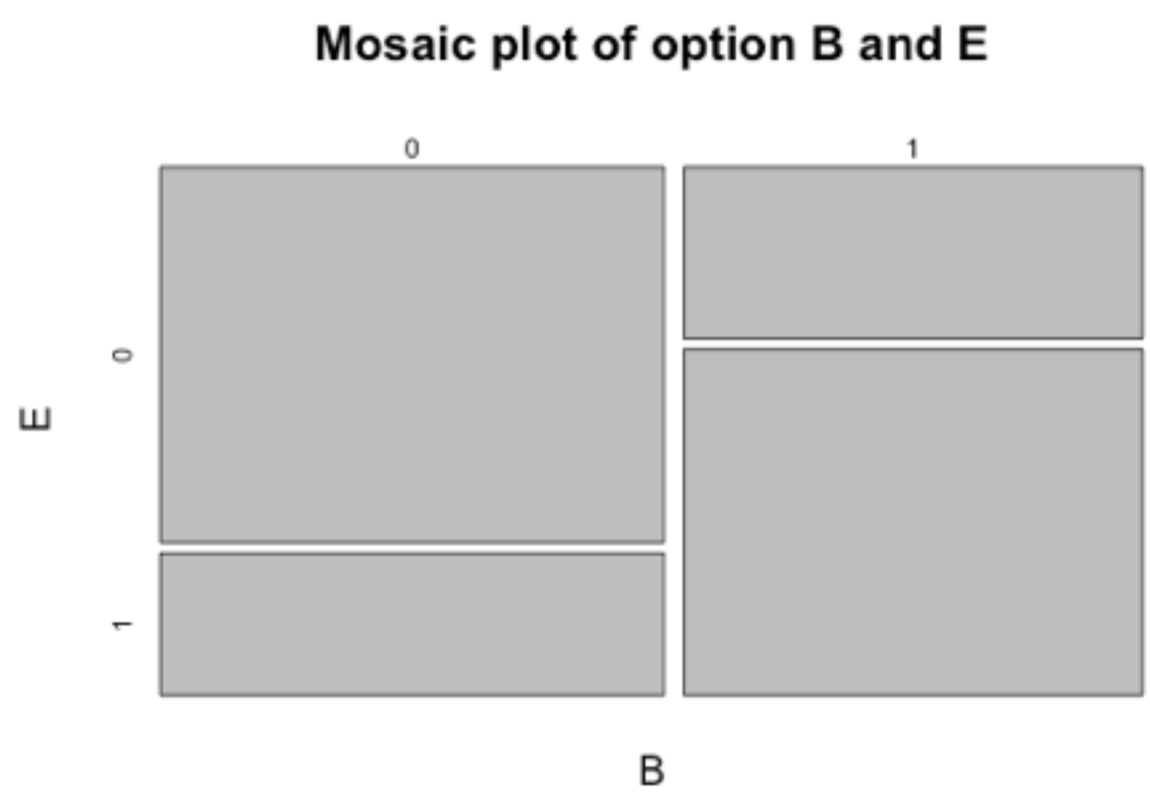
- If don't require to correctly predict G, correct rate increases to 78.4% (compare to correct rate increases to 72% if don't require to correctly predict F)

Predict G

- Use the last viewed values in A-F, but predicted value for G
- Correct rate increase to 70.66%
- Alternatively, random forest method. Correct rate is still 70.66%.
- Most important variables for predicting G are last viewed G, second last viewed G, state and cost.

Predict Everything

- Fit a random forest model for each option…
- Some pattern we found in mosaic plot did proved to be useful. Ex. (second) last viewed E was important predictor for option B

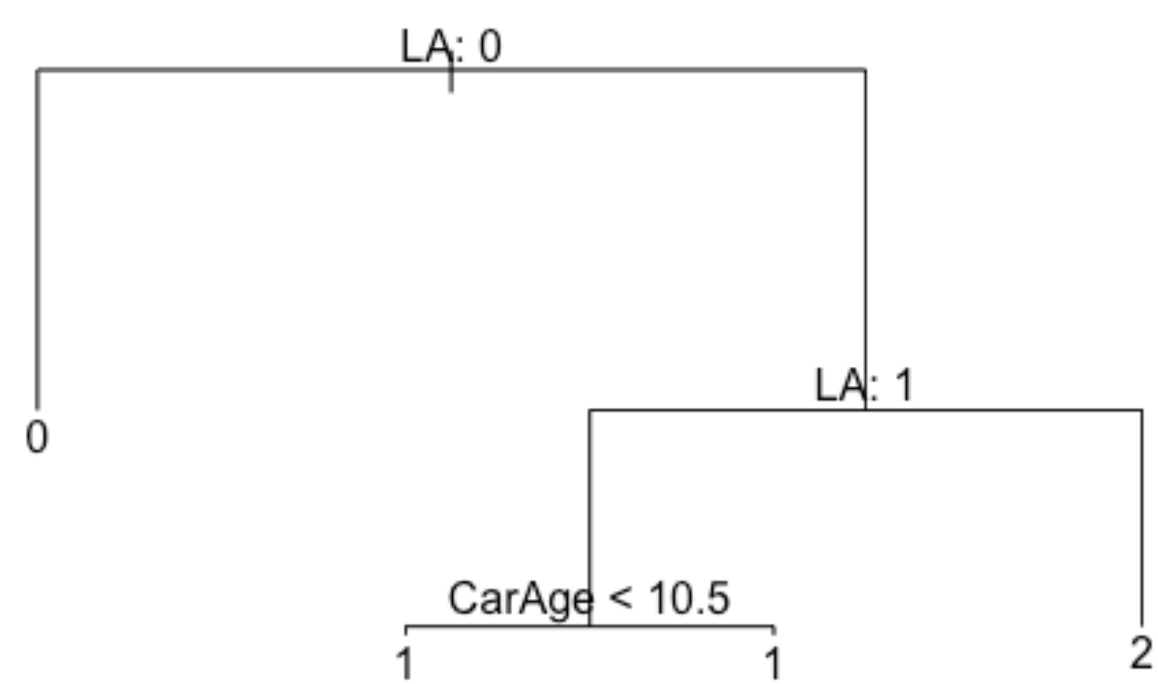


Doesn’t work well…

- Correct rate drops to 70.42%, lower than naive model (70.6%)

Tree

- Use binary tree to determine classification for each option
- Tree function in R only allows 32 levels—cannot use state data
- Except for option A, final decision for each option only determined by last viewed value of corresponding option
- Correct rate is 70.6%



Maybe…

- …The predicting power is mitigated by people who changed their purchased option from last viewed option

A	B	C	D	E	F	G
0	0	1	1	0	0	4
0	0	1	1	0	0	1
0	0	1	1	0	0	1
1	1	1	1	1	2	1
1	1	1	1	1	2	1
2	1	1	1	1	3	1

Predict Who will Change

- New variable: people who changed their final decision from last viewed option
- Predict who will change.
- For people who will change their decision, try to find a pattern for how they will change

Who will Change

- LASSO regression on change indicator with available previous information
- Use Bernoulli distribution with predicted change probability to get predicted change indicator
- Correct rate for change indicator = 57.5%
- If we assume no one would change decision… correct rate = 70.6%

Predict for People who will Change

- A random forest model for each option

- This time, the final decision of each option is less dependent on values of other options (option C and D depend on last viewed value of the other, option E and F depend on last viewed value of option A)

- However…correct rate is 66.12% (worse than naive model)

Conclusion

- Naive model works good enough…
- If want something a little better, try using predicted G value

Other Possibilities

- We have complete data here, we know true “last viewed” options. In test data, probably we only know first several views. Last possible entry always has higher correlation with final purchase.

- Tree model for each state

- Try predicting combination (2304-level factor) instead of predicting each option

Questions?

Thank you!