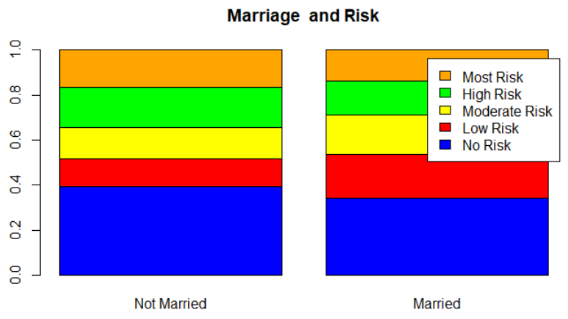
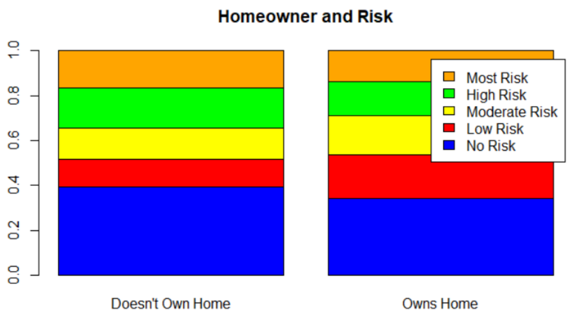
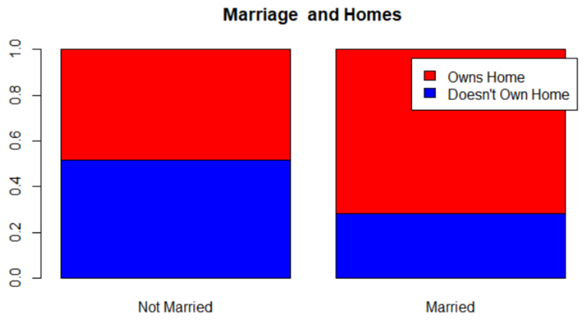
Undercutting Based on Analytics
Section B Team 9: Alper Sayiner, Annabelle Nguyen, Vatsal Sanghavi, and Sophia Rowland

1. Pick two (or more) variables and attempt to show a relation between them via visualization. As discussed before, this requires one to formulate a question, and to communicate clearly a conclusion based on data visualization (specify the why, what, how).

Are groups that include married couples or homeowners more responsible by being less risky? As the first stacked proportion bar graph below demonstrates, there is a smaller proportion of married couples than unmarried couples in the no-risk group, but overall there is a larger proportion of unmarried couples than married couples in the high-risk and most-risk groups. As the second stack proportion bar graph below demonstrates, there again is a smaller proportion of homeowners than non-homeowners in the no-risk group, but overall there is a larger proportion of non-homeowners than homeowners in the high-risk and most risk groups. Perhaps the parallels between married couples and homeowners indicate that there is a relationship between the two, and as the third stacked proportion bar graph shows, married couples tend to be homeowners more often than not.







2. Provide a model based on linear regression to forecast the quoting procedure from ALLSTATE based on the observed variables. Pick two variables of your model, describe their marginal impact on the quote, and comment the interpretation from the business perspective.

We compared the below models and picked the one with the highest adjusted R square value. The sample models were chosen based on examples provided, intuition, and also by checking which variables have the maximum correlation with the cost variable:

- model1 <-lm(cost~.^2, data=DATA)
- model2 <-lm(cost~., data=DATA)
- model3 <- lm(cost~ day+state+group_size+homeowner+car_age, data=DATA)
- model4 <- lm(cost~risk_factor+car_age+car_value+group_size+age_youngest, data=DATA)
- model5 <- lm(cost ~ .+(A+B+C+D+E+F+G)^2, data=DATA)
- model6 <- lm(cost ~ A+B+C+D+E+F+G, data=DATA)
- model7 <- lm(cost ~ homeowner+car_age+age_oldest+age_youngest+A+B+C+D+E+F+G, data=DATA)

Among these, the first model has the maximum value of adjusted R square so we pick that. Now, it only makes sense to interpret significant variables. Two examples are given below.

- Those who purchased insurance on day 1 (Tuesday) were quoted $212.4 on average more compared to the excluded day (day 0 that is Monday). From a business perspective, our quotes on average should only be $211 (or less) higher for customers asking for them on a Tuesday compared to those asking for a quote on Monday. This will lure them away from ALLSTATE to our company.
- Customers shopping from state CO (Colorado) were on average quoted a price of $105.4 lower than customers from state AL (Alabama). From a business perspective, our quote for people from Colorado should be even lower, for example lower by $106 compared to those from Alabama on average to lure them away from ALLSTATE to our company.

3. Suppose that a customer will pick the lowest between the quote you provide and that ALLSTATE provides. Build a model framework (follow/adapt steps in Model Framework in Class 3 for the Churn Problem) to maximize expected revenue from a customer given the observed characteristics. This includes the mathematical model, description of a decomposition strategy, the associated core tasks, and specific data mining methods you would choose. For each core task comment if it can and if it cannot be implemented with the available data.

In this problem, we are given information about policy offered, customer demographics, and a quote given by ALLSTATE. Using this data, we are trying to predict ALLSTATE's model for finding quotes so that we can predict what price ALLSTATE will offer a customer. With this information, we can offer the customer a quote that will maximize our revenue, but be lower than ALLSTATE's quote. First, we need to try and find ALLSTATE's model, so we create various predictive models using the data provided and pick the model with the best cross-validation performance. Unfortunately, our data did not include gender, which often plays an important role in car insurance quoting, which may make our model less like ALLSTATE's model. The model we found in questions two is model1 <-lm(cost~.^2, data=DATA) with an adjusted R^2 of 0.6571. Next, when we use our model to predict ALLSTATE's price in a real-time quote comparison, we should create a 95% confidence interval around our prediction, take the lowest quote ALLSTATE could offer in that interval, and offer $1 less, as long as that number is more than our costs of providing the insurance to ensure that we are maximizing revenue by getting more customers.

4. Suppose that a customer will pick the lowest between the quote you provide and that ALLSTATE provides. Aiming to maximize expected revenue, provide quotes for each of the three customers specified in "new.customers". Clearly state which core task and which data mining method you used to provide the quote.

In order to predict the quotes for each of the 3 customers, we use a supervised predictive method, specifically linear regression. Using the ALLSTATE data and model1 as discussed in Question 2, we predict that ALLSTATE will present these quotes to the 3 new customers.

	Fit (average)	lwr (lower range)	upr (upper range)
1	595.1728	539.7348	650.6108
2	666.0726	607.7809	724.3643
3	641.5417	585.7131	697.3703

We predict that ALLSTATE will also want to maximize expected revenue from customers. Therefore, assuming that our predictions are accurate, ALLSTATE will likely present the quote for Customer 1 as $595.17 - $650.61, because customers will likely pick the lowest between the quote ALLSTATE provides.

In order to undercut ALLSTATE, our company should charge customers on the lower half of the 95% prediction interval of ALLSTATE. For example, for Customer 1, we'll provide a quote of $539.73 - $595.17 and the customer will likely choose the lowest cost. However, in order to be completely certain that we will undercut ALLSTATE, we should charge at least $1 less than the lower range of ALLSTATE'S estimate for each customer.

5. Suppose next that the customer might not accept either of the two quotes (but he will consider only the smallest of the quotes). Build a model framework (follow/adapt steps in Model Framework in Class 3 for the Churn Problem) to maximize expected profit from

a customer given the observed characteristics. This includes the mathematical model, description of a decomposition strategy, the associated core tasks, and specific data mining methods you would choose. For each core task comment if it can and if it cannot be implemented with the available data.

In this problem, we are given information about policy offered, customer demographics, and a quote given by ALLSTATE. Using this data, we are trying to predict ALLSTATE's model for finding quotes so that we can predict what price ALLSTATE will offer a customer. Using this information, we can offer the customer a quote lower than ALLSTATE's quote, but the customer still may not pick that quote. This indicates that there may be a lower quote from another insurance company, for which we have no data. To combat this problem we should seek data for quote generation in other car insurance companies and try to find their models. If that is not possible, perhaps we could gather information about the quote comparison process, including our quote, customer demographics, whether or not the customer chooses our insurance, and the lowest quote offered to the customer (if possible). Using this data, we can try to build a model that predicts the lowest price that is offered by another insurance company. We can do this by creating various predictive models using the comparison data and picking the model with the best cross-validation performance. Next, when we use our model to predict the lowest comparison price in a real-time quote comparison, we should create a 95% confidence interval around our prediction, take the lowest quote offered in that interval, and offer $1 less, as long as that number is more than our costs of providing the insurance to ensure that we maximize profit by getting the most customers.