Data Science for Business Team Case 2

Team 18

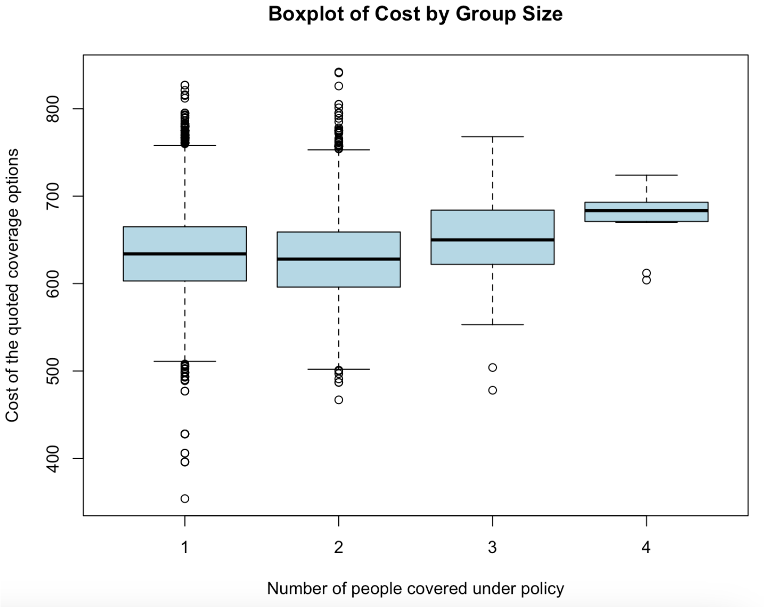
Yinan Chen, Caryl Alexis Cohen, Tirth Pravin Gala,
Yuhe(Tiffany) Jin, Chenjie (Angelina) Sun

Question 1

Why: Assessing the number of people covered is essential for insurance companies to decide the policy. How does the cost of the quoted coverage options vary with the number of people covered under the policy?

What: From this box plot, we can see the distribution of the cost of the quoted coverage options for each group size. 1indicates the smallest group size while 4 indicates the largest under the policy. Specifically, the average costs and variance of the group size 4 are significantly higher than those of other groups. That is to say, the cost of insurance coverage tends to increase as the number of people covered under the policy increases.

How: More people covered under the policy implies a more diverse risk pool that may result in higher risks for the insurance companies.



Question 2

After running the regression model, our team chose *homeowner* and *married_couple* as our two variables for this case.

	Estimate	**Std. Error**	**t-value**	**Pr(>|t|)**
married_couple1	-9.16056	1.23442	-7.421	1.22e-13
homeowner1	-13.23331	0.64396	-20.550	< 2e-16

'married_couple' has a value of -9.16056, it indicates that being part of or considered as a married couple results in an average quote that is $9.16 less than for those who are not married. All other factors will stay the same. From a **business perspective**, this may suggest that married couples may be perceived as less risky or more stable customers, because they are more likely to have multiple policies within the family member or they are statistically less likely to make claims. For example, married couples sometimes represent an opportunity to cross-sell other insurance products, such as

homeowners, renters, or life insurance. If one of them has a positive experience with ALLSTATE, it might lead to the other one getting their policy with the company as well, doubling the customer value. The t-value suggests that this estimate is very statistically significant. Thus, it might make sense for insurance marketers to target married couples with certain incentives or discounts, given their potentially lower risk profile.

'homeowner' has a coefficient of -13.23331, suggesting that homeowners receive a quote that is $13.23 less than non-homeowners, provided all other factors remain the same.
For **business perspective**, homeowners might be seen as more responsible or financially stable customers. Being a homeowner may correlate with other factors like having a more steady income, which can translate to timely premium payments and potentially fewer claims. The extremely significant t-value indicates that being a homeowner is a strong predictor in this model for quoting purposes.

The insurance companies like ALLSTATE provide better rates for *homeowner* and *married_couple* groups due to the stability and lower risk associated with them.

Question 3

1. Mathematical model

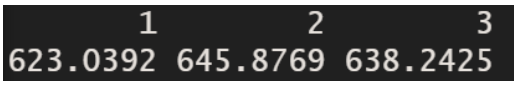
For each customer with features X:
Customer: pick the quote you provide (P) or pick that ALLSTATE provides (notP)
Gains = E[Revenue | X,P] - E[Revenue | X,notP]

2. Associated core tasks & Decomposition Strategy
   Predictive Goal: Predict whether the customers would pick the quote you provide or pick that ALLSTATE provides
   - Regression: Predict the quote price *(cost)* given the observed features of customers by running multiple linear regression *(other variables)*
     - This core task could be implemented within the available data.
   - Classification: Predict the class an individual belongs to by estimating the features as independent variables with a focus on the marginal impact of features
     - This core task could be implemented within the available data.
3. Specific data mining methods
   The methods we use are regression analysis and classification. More specifically, they are multiple linear regression, linear regression with interaction terms, CART, and Random Forest.

Question 4

1. Core Task: Estimate value of quote considering customer characteristics

```
result_model =
glm(cost ~
        day + state + group_size + homeowner + car_value + car_age + risk_factor + age_oldest +
        age_youngest + married_couple + C_previous + duration_previous + A + B + C + D + E + F
        + G + A:D + A:E + A:F + B:C + B:D + B:E + B:G + C:E + C:G + D:E + D:F + E:F + F:G,
        data = DATA)
predict (result_model, newdata= new.customers)
```

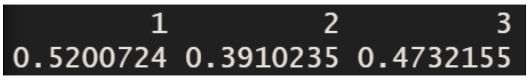


```
       1        2        3 
623.0392 645.8769 638.2425 
```

2. Core Tasks: Find the probability of winning the customer. (Estimate Cost < Actual Cost)

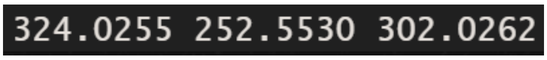
```
DATA$cost_prediction = predict(result_model,newdata=DATA)
DATA$residual = DATA$cost-DATA$cost_prediction
DATA$win = ifelse(DATA$residual>0,1,0)
DATA$customer_win = ifelse(DATA$residual>0,1,0)

win_prediction=glm(win~day + state + group_size + homeowner + car_age + car_value + risk_factor + age_oldest + age_youngest + married_couple + C_previous + duration_previous + A + B + C + D + E + F + G + A:D + A:E + A:F + B:C + B:D + B:E + B:G + C:E + C:G + D:E + D:F + E:F + F:G, data = DATA, family = "binomial")
predict(win_prediction,newdata= new.customers, type="response")
```



```
        1         2         3 
0.5200724 0.3910235 0.4732155 
```

```
value1=623.0392*0.5200724
value2=645.8769*0.3910235
value3=638.2425*0.4732155
result_values = c(value1,value2,value3)
result_values
```



```
324.0255 252.5530 302.0262
```

Question 5

1. Mathematical model

For each customer with features X:

Customer: accept (Ac) either one of the two quotes with a lower price or not accept(notAc) both quotes

Gains = E[Profit | X,Ac] - E[Profit | X,notAc]

2. Associated core tasks & Decomposition Strategy

   Predictive Goal: Predict whether the customers would accept(Ac) one of the two quotes (ours and ALLSTATE’s) or not accept(notAc) both quotes

- Regression: Predict the quote price *(cost)* given the observed features of customers by running multiple linear regression *(other variables)*
  - This core task could be implemented within the available data.
- Classification: Predict the class an individual belongs to by estimating the features as independent variables
  - This core task could be implemented within the available data.

3. Specific data mining methods

   The methods we use are multiple linear regression, linear regression with interaction terms, logistic regression, logistic regression with interaction terms, Classification Tree, and Random Forest. Then, we will choose the most fitting classification model based on their performance.