Team 03

Mackenzie O’Brien, Madison Russell, & Madeline Signorelli

**Project #1: Targeted Marketing**

Companies routinely use surveys to estimate potential revenues in different markets. Survey responses are collected and aggregated in large databases where data analysis applications are used to generate reports to help in business decisions.

Suppose you are marketing manager attempting to optimize an advertising campaign. For years managers often send out bulk mailings (now bulk emails) to different prospect lists. Imagine you can measure cost for every mail sent. If every contact costs the company 5 dollars, sending a mail to 1000 people costs 5,000 dollars. This becomes very expensive. None of the prospects even guarantee a response, or even a response which leads to additional revenue. Companies employ data analysis techniques to select only those prospects most likely to respond to a campaign. By analyzing survey responses from prospects, we can segregate prospects into different groups of revenue probability.

Why are beer commercials shown during the Superbowl? Beer companies are showing targeted advertisements to a segment of the population they think are the most likely to go out and buy another 6 pack. Instead of showing commercials all year round and paying millions of dollars, they select the best potential sample of the population at the right time. How then, do they go about selecting the right audience? It seems it is becoming less common for people to make multi-million dollar decisions based on gut feeling and intuition.

In this case, you are a data analyst who works for the Land Rover automobile company. You desire to gain a better understanding of the lifestyle of potential SUV buyers. You commission a study of consumer attributes, interests, and opinions, and send out a questionnaire. The survey involves 30 questions, covering a variety of different consumer dimensions. The 31st question asked the consumer to rate themselves as to how likely they would be to purchase a SUV.

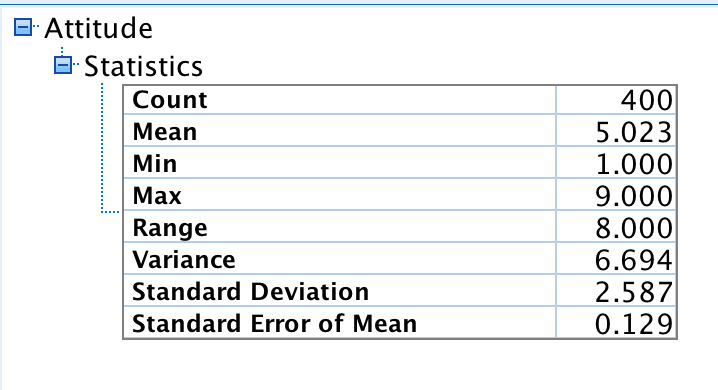
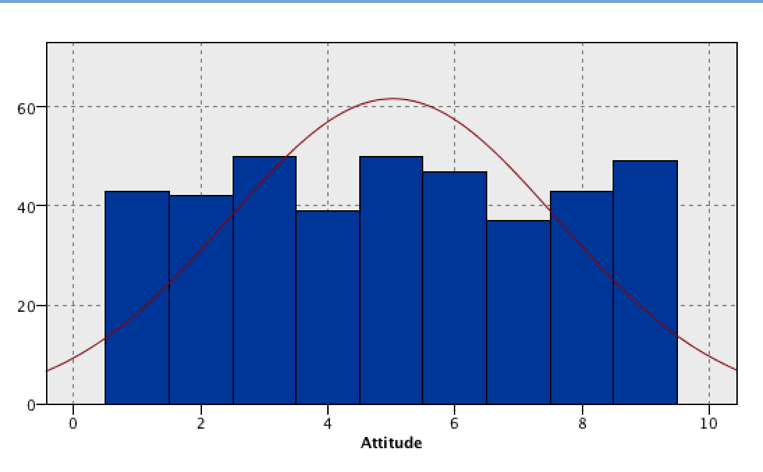
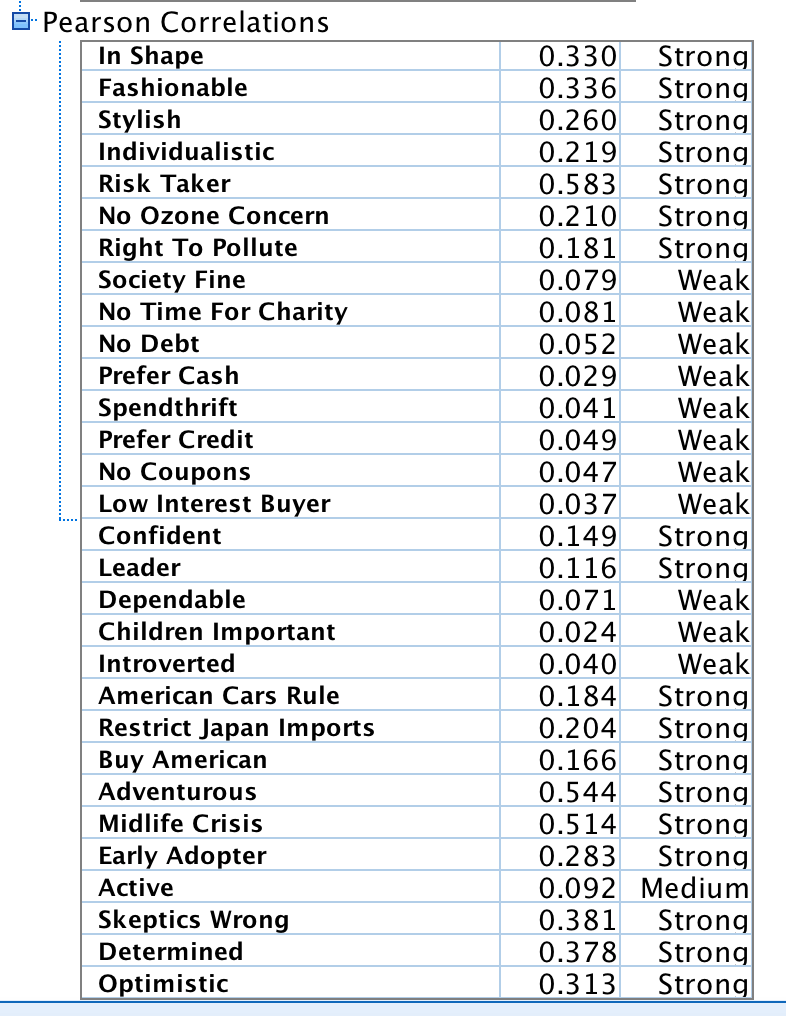
Respondents were asked to use a 9 point Likert scale to give their answer to each question. The value of "1" means the responder disagreed with the statement, and the value of "9" means the responder totally agrees. 400 consumers were surveyed. The profiles were obtained from the mailing lists of Car and Driver, Business Week, and Inc. magazines.

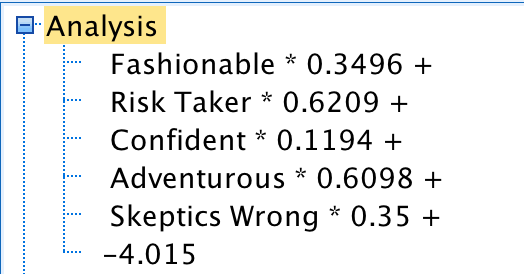
The survey questions (prefixed by their column name in the data file):

1. In Shape - I am in very good physical condition
2. Fashionable - When I must choose between the two, I dress for fashion, not comfort
3. Stylish - I have more stylish clothes than most of my friends
4. Individualistic - I want to look a little different from others
5. Risk Taker - Life is too short not to take some gambles
6. No Ozone Concern - I am not concerned about the ozone layer
7. Right To Pollute- I think the government is doing too much to control pollution
8. Society Fine - Basically, society today is fine
9. No Time For Charity - I don't have time to volunteer for charities
10. No Debt - Our family is not too heavily in debt today
11. Prefer Cash - I like to pay cash for everything I buy
12. Spendthrift - I pretty much spend for today and let tomorrow bring what it will
13. Prefer Credit - I use credit cards because I can pay the bill off slowly
14. No Coupons - I seldom use coupons when I shop
15. Low Interest Buyer - Interest rates are low enough to allow me to buy what I want
16. Confident - I have more self-confidence than most of my friends
17. Leader - I like to be considered a leader
18. Dependable - Others often ask me to help them out of a jam
19. Children Important - Children are the most important thing in a marriage
20. Introverted - I would rather spend a quiet evening at home than go out to a party
21. American Cars Rule - Foreign-made cars can't compare with American-made cars
22. Restrict Japan Imports - The government should restrict imports of products from Japan
23. Buy American - Americans should always try to buy American products
24. Adventurous - I would like to take a trip around the world
25. Midlife Crisis - I wish I could leave my present life and do something entirely different
26. Early Adopter - I am usually among the first to try new products
27. Active - I like to work hard and play hard
28. Skeptics Wrong - Skeptical predictions are usually wrong
29. Determined - I can do anything I set my mind to
30. Optimistic - Five years from now, my income will be a lot higher than it is now
31. Attitude - I would consider buying the Discovery made by Land Rover

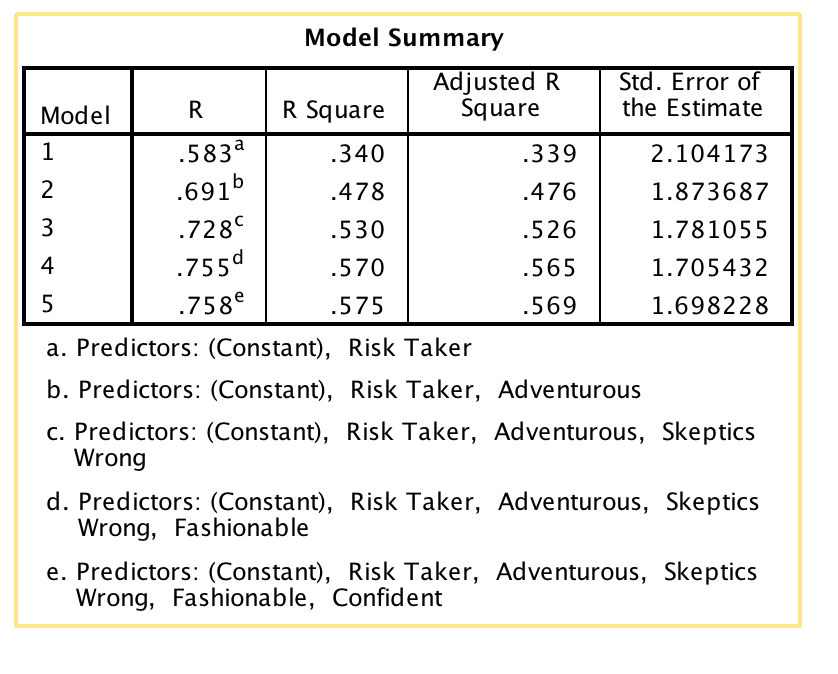
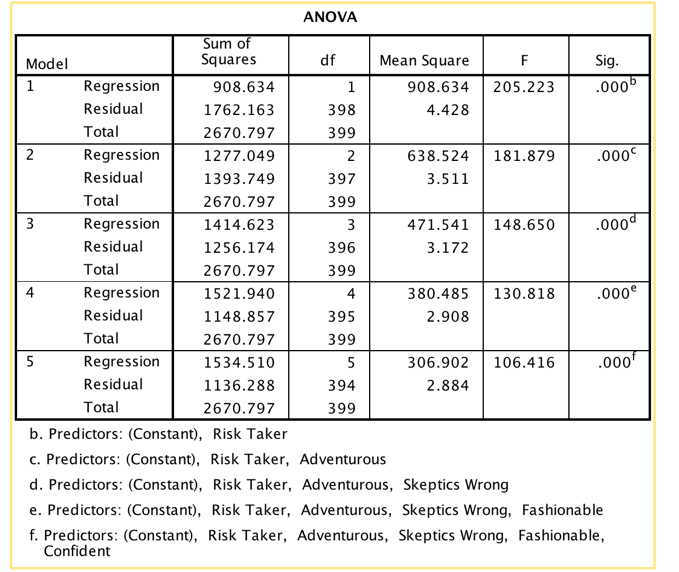
The data is provided in an Excel file named SurveyData.xls. By this time you should be pretty familiar with loading data from an Excel file or other sources.

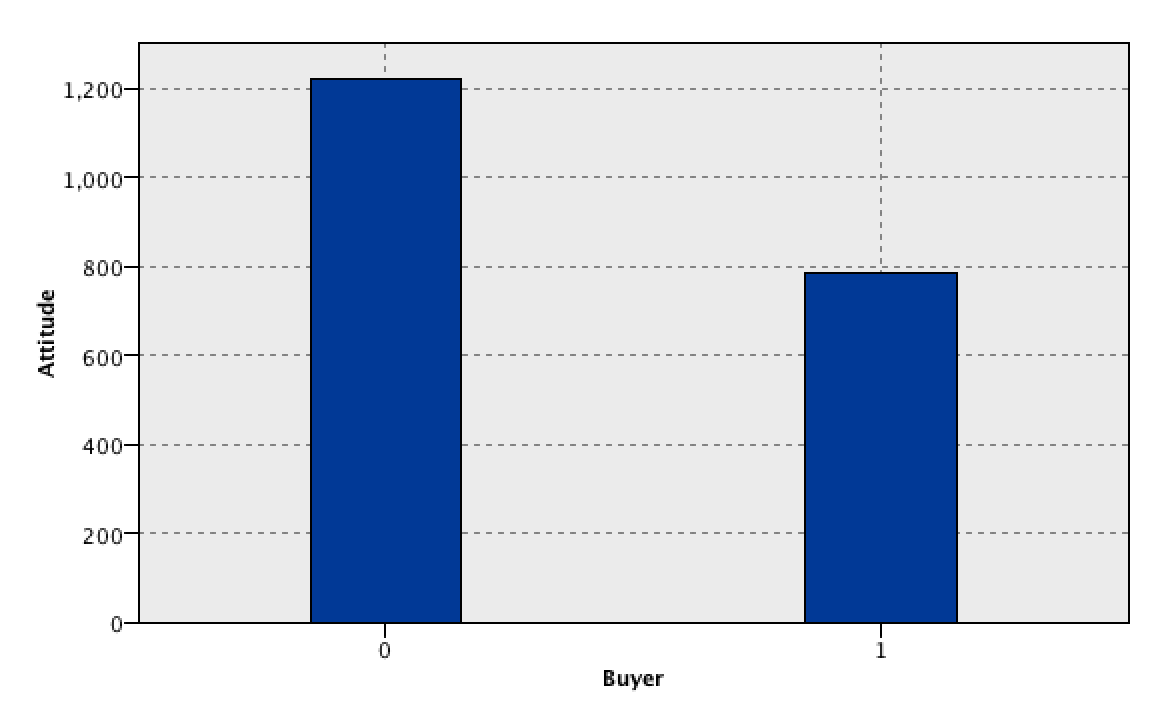
Tasks:

1. Explore the data to get some initial insights. Add a few chart nodes and a statistics node to visualize the data. Report on your findings of the shape of the data. (Note: we assume that a Likert scale of 1-9 points can approximate a continuous variable)
   1. The mean of Attitude is 5.023, which can be expected on a Likert scale of 1-9. The standard deviation is 2.587 and the variance is 6.694. The histogram of attitude sees spike at values near 3, 5, and 9 with the lowest value of 7 having a count of just above 35. This data set is not distributed normally.
2. Missing data can be a problem. Check for missing values (there may be some, there may none) and remediate the issue accordingly, if you think it is necessary.
   1. There are only a few outliers and no extreme outliers. There is no missing data, so we did not remediate anything.
3. Provide a table describing the relationship of each survey question (questions 1-30) with Attitude (question 31). Also, investigate the correlation amongst the predictor variables (questions 1-30) [[1]](#footnote-1). Report your findings.
   1. There seems to be weak correlations between attitude and society fine, no time for charity, no debt, prefer cash, spendthrift, prefer credit, no coupons, low interest buyer, dependable, children important, and introverted. There is a medium correlation with attitude and active and the rest are strong correlations according to the Pearson’s correlations model.
      1. In Shape is strongly correlated to fashionable, stylish, individualistic, risk taker, adventurous, midlife crisis, skeptics wrong, optimistic, and attitude
      2. Fashionable is strongly correlated to in shape, stylish, individualistic, risk taker, adventurous, midlife crisis, skeptics wrong, optimistic, and attitude
      3. Stylish is strongly correlated to in shape, fashionable, individualistic, risk taker, skeptics wrong, determined, optimistic, and attitude
      4. Individualistic is strongly correlated to in shape, fashionable, stylish, risk taker, adventurous, skeptics wrong, determined, optimistic, and attitude
      5. Risk Taker is strongly correlated to in shape, fashionable, stylish, risk taker, individualistic, no ozone concern, right to pollute, American cars rule, restrict japan imports, buy American, adventurous, mid life crisis, early adopter, skeptics wrong, determined, optimistic, and attitude
      6. No Ozone concern is strongly correlated to risk taker, right to pollute, society fine, no time for charity, confident, leader, adventurous, and attitude
      7. Right to Pollute is strongly correlated with risk taker, no ozone concern, society fine, no time for charity, and attitude
      8. Society fine strongly correlates to no ozone concern, right to pollute, no time for charity
      9. No time for charity strongly correlates to no ozone concern, right to pollute, society fine
      10. No debt, prefer cash, spendthrift, prefer credit, no coupons, low interest buyer all strongly correlate to each other.
      11. Confident strongly correlates to no ozone concern, leader, dependable, American cars rule, restrict japan imports, and attitude
      12. Leader strongly correlates to no ozone concern, confident, dependable, restrict japan imports, skeptics wrong, determined, optimistic, and attitude
      13. Dependable strongly correlates to confident, leader
      14. Children important and introverted are strongly correlated
      15. American cars rule is strongly correlated to risk taker, confident, restrict japan imports, buy American, adventurous, midlife crisis, and attitude
      16. Restrict japan imports strongly correlates to risk taker, confident, leader, American cars rule, buy American, adventurous, midlife crisis, skeptics wrong, optimistic, and attitude
      17. Buy American strongly correlates to risk taker, American cars rule, restrict japan imports, adventurous, midlife crisis, optimistic, and attitude
      18. Adventurous strongly correlates to in shape, fashionable, individualistic, risk taker, no ozone concern, American cars rule, restrict japan imports, buy American, midlife crisis, early adopter, active, skeptics wrong, determined, and attitude
      19. Midlife crisis strongly correlates to in shape, fashionable, risk taker, American cars rule, restrict japan imports, buy American, adventurous, early adopter, active, skeptics wrong, determined, optimistic, and attitude
      20. Early Adopter strongly correlates to risk taker, adventurous, midlife crisis, active, skeptics wrong, determined, optimistic, and attitude
      21. Active strongly correlates to adventurous, midlife crisis, early adopter, skeptics wrong, determined, optimistic.
      22. Skeptics wrong strongly correlates to in shape, fashionable, stylish, individualistic, risk taker, leader, restrict japan imports, adventurous, midlife crisis, early adopter, skeptics wrong, determined, optimistic.
      23. Determined strongly correlates to stylish, individualistic, risk taker, leader, adventurous, midlife crisis, early adopter, skeptics wrong, determined, optimistic.
      24. Optimistic correlates to in shape, fashionable, stylish, individualistic, risk taker, leader, restrict japan imports, buy American, midlife crisis, early adopter, skeptics wrong, determined, optimistic.
4. Partition the dataset: Use a 70% - 30% ratio for training and testing.
5. Use SPSS Modeler linear regression tool to investigate whether a linear relationship exists between attitude and the other variables. Use the method of your choice (all variables, a subset, stepwise / forwards / backwards) to build the multiple linear regression model, and justify your choice. For your model:
   1. Write out the estimated regression equation and explain the meaning of the coefficients



* + 1. We used stepwise regression. We chose it because it was easier to read in the advanced tab of SPSS and the fact that it tests for the possibility of variables becoming non-significant.
    2. The equation is Attitude=Fashionable \* 0.3496 + Risk Taker \* 0.6209 + Confident\*0.1194 + Adventurous \* 0.6098 + Skeptics Wrong \*0.35 + -4.015. This means that if all the variables are 0, the attitude is -4.015 as a base value, however this is impossible as attitude must be at least 1. As Fashionable increases on the Likert scale, it increases the likelihood of increasing attitude by 0.3496. This is true for the other variables and their slopes.
  1. Provide a full report of the chosen regression model and report its metrics (goodness of fit, predictive performance) and statistics on training and test data
     1. The goodness of fit (R2) for the final model in the step-wise regression (e/f) is .575, meaning that the model accounts for 57.5% variability of the data. We get this number from SSR/SST. The standard error of the estimate is 1.698228, meaning that the model prediction may be off by an average of 1.698228



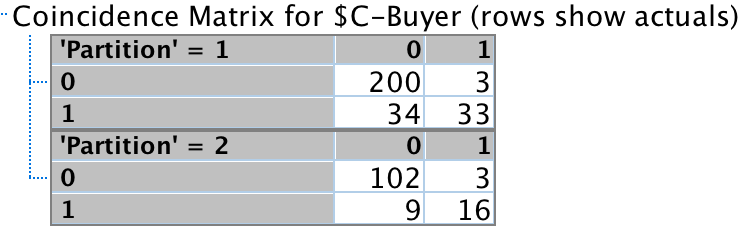
1. Derive “Buyer” attribute: pick a threshold Attitude value (Attitude = 7), above which you can consider that that the responder is a highly likely buyer. Create a rule that tests the value of Attitude, and use this to derive a new attribute called Buyer (the rule is something like “Attitude" > 7)
2. Check the distribution of Buyers and non-Buyers in the dataset to see if the data set is unbalanced (you can plot a graph or create a table).
   1. There is a distinct unbalance is the dataset in which there are proportionately less buyers that non- buyers based off of overall attitude.
3. Decision Tree Classification: Model the dependence of how likely someone will purchase a Land Rover using a C5.0 decision tree. With the resulting tree, create a profile of the potential Buyer (Note: when creating the tree remember to drop Attitude from the list of predictors). This means explaining the rules created by the tree (or walking down the branches of the tree, same thing)



The variable at the root of the tree is Adventurous. If the respondent is not very adventurous (<=5.5) and not much of a risk taker (<=5.5), then there is a 93.5% chance that they are a non-buyer. If the respondent is adventurous (>5.5), is determined (>2.5), believes that American cars are the best (>3.5), and makes skeptical predictions that are usually wrong (<4.5), then there is a 91.2% chance that they are a buyer. If the respondent is adventurous (>5.5), is determined (>2.5), believes that American cars are the best (>3.5), believes skeptical predictions are usually wrong (<=4.5), and is not much of a risk taker (<=5.5), then there is a 100% chance that they are a non-buyer.

(ENOUGH RULES? ^^^^)

1. Performance Evaluation (1) – Confusion Matrix and Derived Metrics. Derive proper performance metrics considering the (balanced / unbalanced) nature of the data set.



Since the data set is unbalanced, with proportionately less buyers that non-buyers based off of overall attitude, we will derive the performance metrics of recall, specificity, and precision for the training set.

Recall=33/(33+34)=0.493

Specificity=200/(200+3)=0.985

Precision=33/(33+3)=0.917

1. Performance Evaluation (2) – Gain Chart. Plot a Gain Chart and explain it



The gain chart shows the percentage improvement in the decision tree classification strategy compared to a random chance strategy. At about 15% of the sample we attain the largest gain with respect to a random sample strategy. The slope is steepest from 0-15%.

1. You can use the statistics node in SPSS, but you may profit from using the Data Analysis add-in in Excel. In there, choose the correlation option, and it will build for you a nice two-way table with all the correlations (aka correlation matrix) [↑](#footnote-ref-1)