STAT 331 Lab 7

Instructions

Submit your HTML markdown document by the beginning of class.

Introduction

This data set (mariokart.sas7bdat) includes data about some auctions on Ebay. Auctions were included in the data set if they satisfied a number of conditions. (1) They were included in a search for "wii mario kart" on ebay.com, (2) items were in the Video Games > Games > Nintendo Wii section of Ebay, (3) the listing was an auction and not exclusively a "Buy it Now" listing (sellers sometimes offer an optional higher price for a buyer to end bidding and win the auction immediately, which is an optional Buy it Now auction), (4) the item listed was the actual game, (5) the item was being sold from the US, (6) the item had at least one bidder, (7) there were no other items included in the auction with the exception of racing wheels, either generic or brand- name being acceptable, and (8) the auction did not end with a Buy It Now option. All prices are in US dollars. Our goal for this lab to create models for the total selling price of the Ebay package (totalPr).

Variables

- ID: Auction ID assigned by Ebay.
- duration: Auction length, in days.
- nBids: Number of bids.
- cond: Game condition, either new or used.
- startPr: Starting price of the auction.
- shipPr: Shipping price.
- totalPr: Total price, which equals the auction price plus the shipping price.
- shipSp: Shipping speed or method.
- sellerRate: The seller's rating on Ebay (number of positive ratings minus the number of negative ratings).
- stockPhoto: Whether or not the auction feature photo was a "stock" photo.
- wheels: Number of Wii wheels included in the auction.
- title: The title of the auctions.

Exercises

1) Begin by exploring the dataset. Print out the variable names and the first few rows of the dataset.
2) There are two observations that really don't fit the pattern of the rest of the data with respect to total selling price. Identify these observations through numerical or graphical summaries, and explain why they are outliers.

3) Remove the outliers identified above.

4) Create two new indicator variables:

- condI: an indicator for condition of the game (cond), where 1 = new and 0 = used
- photoI: an indicator for stockPhoto, where 1 = yes and 0 = no

5) Note that Mario Kart packages can come with 0 through 4 wheels, but only a couple of packages come with 3 or 4 wheels. Create a new variable called wheelsnew that is coded as 0 wheels, 1 wheel, or 2+ wheels.

6) Examine the relationship between total selling price and number of wheels. Use numerical and graphical summaries to assess whether there's a difference in total selling price across different numbers of wheels. Be sure to use your new variable, wheelsnew.

7) Let's look at the relationship between total selling price and some of the other variables.

(a) Use the cor() function to look at the correlation between the numeric variables in the dataset (totalPr, duration, nBids, startPr, shipPr, sellerRate).

(b) Since we're only interested in the relationship with totalPr, comment on the correlations between each of the numeric variables and totalPr. Which variable has the strongest linear association with totalPr?

(c) Validate your answer to (b) with graphs of each pair of variables: totalPr vs. duration; totalPr vs. nBids; and so on. Try adding linear fits to each of these graphs as well. Do they confirm your answer to (b)? Do any of the variables have a non-linear relationship with totalPr?

(d) If you had to pick one variable to use to predict the total selling price of a new Mario Kart package, which one would you pick and why?