Spencer Phillips

998887568

11/16/15

STA 141 A02

STA 141 Assignment 4

1. For this problem, I made a function to search up the prices of vehicles to see how accurate they were using the code str\_match(bodyval, "\\$[0-9]+\\,?[0-9][0-9]+") as the regular expression. Using my method, I got 13064 prices from the bodies. Of these 13064 options, 77.32% of them were the same as the actual price from the prices column by taking a table of the two FALSE being not the same as the source, TRUE being the same.

FALSE TRUE

0.2268065 0.7731935

2. For finding the VIN number, I used str\_match again (str\_match(bodyval,"[A-Z0-9a-z]{11}[0-9]{6}")), and after looking up how a VIN number is formatted, allowed any type of alphanumeric beginning to the first 11 digits of the value, and then had the last 6 digits of the 17 digit VIN number to be all numbers. This kept from there just being random 17 digit things which may have no spaces in it or be an email address or something. This gave me 7734 VIN numbers of the 34,677 or 22.3% of all the cars

3. For phone numbers being extracted from the body, I was able to extract 16205 number from the 34677 possible values, a success rate of approximately 46.7%. I did this through checking for standard American phone numbers with area codes included seeing that not having the area codes of these values would make it impossible to actually call any of the numbers. I did this by first checking for parenthesis around the area code which was optional, then three numbers with the first one not being 0 or 1, then there was either a dash or space (both optional) and three numbers with the first not being 0 or one, then a dash or space optional then 4 numbers. Code: str\_match(bodyval, "\\(?[2-9][0-9]{2}\\)?-?[:blank:]?[2-9][0-9]{2}-?[:blank:]?[0-9]{4}")

4. For emails, I was able to get 107 of the 34677, a rate of 0.3% of emails found. This gives me the idea that not many people are putting their emails into the body of their descriptions, opting to instead just go with the obvious contact button that is on Craigslist.

5. For the years, I used a function that would search for the years with an or statement that would look for relevant years, starting with 1930 being the earliest possible one and the maximum being 2019. This was done using the following statement: (([1][9][3456789][0-9])|([2][0][01][0-9])) From this, I searched in the headers tab and I was able to get a total of 34649 values, a percentage of 99.2% percent. Of those that were found, comparing them to the true year got me a success rate of 91.9% among those that were found. A fantastic rate!

6. For number 6, I used the header part of the data since most of the data was in a format which was “YEAR MAKER MODEL” Which made it easy to gsub it apart in order to get a collection of all of the different car models. This did lead to a lot of generalizing since not all of the postings were in that same format, meaning that there was possibly loss due to the discrepancy in how cars were advertised.

Part 2:  
1. In this part I decided for my model to use the Honda Civic and the Toyota Camry since they were some of the most prevalent cars and also are very similar in that they are inexpensive 4 door sedans.

For the Civic, I used the lm function to test the price against age, odometer, condition, and city. I got the following results. The y intercept for the price was approximately $8800 with age leading to a $231 depreciation per year, while driving the car led to three-tenths of a cent increase in price per mile driven! Having the car in fair or good condition led to $5000 drops in price while a new car would get an approximately $3478 boost to the cost. The best place to sell a Civic would be in the SF Bay Area, followed by Denver due to the respective $4412 and $4185 boosts just for selling it there. Sacramento and New York City also had over $1000 boosts to price but selling a Civic in Las Vegas is a bad idea due to the loss of $1444 to the car’s value.

For the Camry however, the data was all over the place, the y intercept for the data was actually -$801,900. The data got even weirder though, with there being said that there is an approximately $4 a mile increase in price for the odometer and having the car in good condition could net you $245,000 boost in the price while in new condition you get only a measly $27,830 boost. While age still gives a logical -$417 drop in car value per year, selling the car in Sacramento or New York City can give a boost in about $565,000 while Chicago only gives a paltry bonus of $202,600.

In conclusion, I feel like I would not use this method to find the price of a car, especially the model for the Camry, however I feel like the Civic model would come in handy for a few things. I especially found interesting the vast difference in prices among the different places in how there could be a difference in almost $5000 depending on where the car is sold, which seems extreme in my mind. I feel this model is appropriate because you can see where the price differences are which is a good starting point on finding where your individual car fits into the market. However, these differences can be easily manipulated if there are some extreme outliers that can damage the outcome easily or data that was put in incorrectly, such as an odometer reading that got one too many zeros or missing a number with the year or something.