Moody's Mega Math Challenge 2016: Share and (Car) Share Alike

Team #7356:
Rees Chang[1], Claudia Chen[2],
Dylan Gleicher[3], Emily Schussheim[3],
Jim Zhang[4]

[1]Cornell University

[2]Georgetown University

[3]Yale University

[4]California Institute of Technology

27 February 2016

Contents

Executive Summary

Dear Automotive Companies,

The invention of the commercial car in the early 20th century sparked a transportation revolution. Today, car sharing has become an attractive alternative to private car ownership. For automotive companies, tapping into this growing market is imperative to remaining relevant in the modern age of automobile transportation.

For car-sharing companies, the two main factors that contribute to consumer decisions in car sharing are (1) the amount of time using the car per day and (2) the miles driven per day, since car-sharing companies charge consumers per minute and per mile [1]. Thus, we built a mathematical model to determine the percentage of all current U.S. drivers in nine categories: high, medium, or low for each of the two main factors.

From this model, using a three-category classification system based on the driving behaviors of different age groups, 57.5% of Americans drive a "low" daily distance, 12.1% drive a "medium" daily distance, and 30.3% drive a "high" daily distance. In regards to daily driving duration, it was found that 67.2% of Americans have a high daily driving time, 23.91% have a low daily driving time, and 8.89% have a medium daily driving time. We also felt it would be helpful to automotive companies to know if a specific individual, in consideration of the types of people who live around the company's geographic location, to be in a low, medium, or high category for each factor. Consequently an algorithm was created that allows user-input values describing an individual to predict categorization of that individual for each of the two factors.

Multiple car-sharing options are available: (1) round trip car sharing, (2) one-way car sharing floating model, (3) one-way car sharing station model, (4) and fractional ownership. To determine the participation in different cities given the different car-sharing options, we created a model that ranked the participation for each option in four cities: Poughkeepsie, NY; Richmond, VA; Riverside, CA; and Knoxville, TN. Our model concluded that the one-way car station model would garner the most participation in any given city, and that the order car companies should proceed to investigate implementation is (1) Richmond, VA (2) Riverside, CA (3) Poughkeepsie, NY (4) Knoxville, TN. Recent studies suggest that self-driving and clean-energy vehicles are close to entering the mainstream [2][3]. We created a model to account for the inclusion of emerging automobile technologies. This model found that using current self-driving car accident rates, the ranking of the given cities remains the same. However, in a future where self-driving car accident rates reach nearly zero, the ranking changes to (1) Poughkeepsie, NY (2) Riverside, CA (3) Richmond, VA (4) Knoxville, TN.

Overall, our models produce the proportions of people who drive certain distances and durations per day in America; predict based on custom, personal variables via machine classification; and produce the best car sharing business models for car sharing participation rates given different cities. If these models are used in conjunction with cost data, the profitability and thus the viability of the car sharing market given the conditions for a specific car sharing company can be easily predicted and subsequently optimized.

- Team #7356

1 Introduction

1.1 Background

Owning a car is becoming decreasingly common among millenials due to the costs and responsibilities involved. According to the AAA Foundation for Traffic Safety, the number of cars purchased by people from ages 18 to 34 fell by nearly 30 percent from 2007 to 2011 [4]. Between 2001 and 2009, the average number of miles driven by people aged 16 to 34 dropped by 23 percent [5].

Modern day car traffic congestion creates some of the most agonizing situations, and are only worsening as the number of cars on the road increases. Finally, environmental degradation is increasingly becoming part of the international agenda, and the expansive and wasteful use of vehicular transport is no help to the quagmire. As environmental issues garner augmented national attention and the emphasis on alternative modes of transport becomes more pronounced, consumers seek solutions to the inefficiencies of car ownership.

Commercial car sharing began in Europe in the 1970's and is now considered the future of transport [6]. According to research from the University of California at Berkeley, there were over 1.3 million car-sharing customers in the United States in 2014, a market pool that eager transport businesses like General Motors are poised to tap [7].

Four methods of regulating and organizing car-sharing systems have developed. Round trip car sharing encompasses a process in which vehicles are rented by the mile, hour, or day and are taken from and returned to the same central point after usage. One-way floating models for car sharing vary in that cars can be rented from and then returned to varying locations within a designated area. Other one-way models function with stations, and a yet final strategy proposes a peer sharing model, in which people jointly own a private vehicle.

All four models are currently in various degrees of application, and a "greener," healthier, and more productive future demands that car sharing take hold in the national and international population to alleviate traffic and contamination woes.

1.2 Restatement of the Problem

With car-sharing companies on the rise, it is important to know how different variables in the operation of a car-sharing company affect the profitability. In order to determine this, multiple problems were considered.

1. The two main factors that drivers consider when making decisions about car-sharing are the amount of time using the car and the miles driven per day. Build a mathematical model to determine the percentage of current U.S. drivers in each category - low, medium, high - for all combinations of the two specified factors.

2. Multiple car-sharing business options other than Zipcar's original pay-by-the-hour option are emerging. Create a model to rank Poughkeepsie, NY; Richmond, VA; Riverside, CA; and Knoxville, TN by projected car-sharing participation in consideration of the following four business options: -Round trip car sharing: vehicles rented by the day, hour, or mile, or some combination of the three, and are picked up from and returned to the same point. -One-way car sharing floating model: cars are rented on demand and are returned to defined areas, usually requiring a "jockey" to manually reposition

vehicles. -One-way car sharing station model: customers pick up and drop off cars at existing stations. -Fractional ownership: multiple owners jointly purchase a private car.

3. Adjust the previous model to account for emerging technologies, such as self-driving cars and renewable/alternative fuel, impacting car sharing participation. Re-rank the four cities with these new variables in mind.

1.3 Global Assumptions and Justifications

Below are the assumptions that we have made that apply to our entire solution.

- **Assumption:** Future and present behavior can be accurately modeled using current data.
  Justification: Our collected data applies to general and average popualtion, implying that it would be an accurate indicator for future and interpolated situations.

- **Assumption:** All data collected from the private sector is valid under ideal business conditions; all professional organizations have made decisions for their best interest correctly and thus car usage participation can be implied by the number of car sharing stations in a given area.
  Justification: Since there has been no indication that car sharing businesses haven't been operating correctly and efficiently, we have reason to believe that they are all under ideal business conditions.

2 Who's driving?

When determining whether to utilize car sharing, people generally consider two factors: the amount of time one uses their car for, and the miles one drives per day. Therefore, to determine the potential of the car sharing business, it is necessary to determine the percentages of current U.S. drivers that have high, medium, and low levels of these two factors. Thus, we analyzed data to objectify what a low, medium, and high level for each of the two factors were.

2.1 Assumptions, Simplifications, and Justifications

In order to maintain versatility and robustness of our model, we made the following assumptions and simplifications:

- **Simplification:** The categorization system can be based off of driving behavior for different age groups.
  Justification: We believe that age is a good variable to provide the basis of our categorization system since teenagers are generally considered to commute short-distances to school and the elderly considered to stay at home more, while the middle-aged are generally known to work more and thus have to commute longer distances.

- **Assumption:** Census data can be extrapolated to the entire country.
  Justification: Sample sizes for census data are very large, satisfying the law of large numbers.

2.2 The Model and Results

Classifying Low, Medium, and High Levels

The first step in creating a solution to the first problem is to create ranges of values for the two factors — daily drive time and daily drive distance — that represent the three categories: low, medium, and high. In order to objectify such a subjective topic, we analyzed data on the possible ranges of averages of the respective factors.

Miles Driven per Day

In order to classify the categories for miles driven per day, we found data for the average annual miles per driver by age group M_{avgA} [8], and also the distribution of licensed drivers by age [9]. Average daily miles M_{avgD} was calculated as follows:

$$M_{avgD} = M_{avgA} \div 365.25 \ \frac{\text{days}}{\text{year}} \tag{1}$$

We reasoned that since the number of miles driven per age by age is bell-shaped, as can be seen in Figure 1, then we can classify people who have values of miles driven per day close to younger and older people as having "low" miles driven per day values while people whose number of miles driven per day closer to middle-aged people can be classified as "high." Thus we took the range of average daily miles per driver by age group (20.991) and divided it by three, and added that value (6.997) to the minimum value (20.873, the value for people aged 16-19) once to get the upper bound of the "low" category, and twice to get the lower bound of the "high" category. Consequently, we classify anyone who drives less than 27.870 miles per day as driving a "low" daily distance, anyone who drives between 27.870 and 34.867 miles per day as"medium", and anyone who drives greater than 34.867 miles per day as"high".

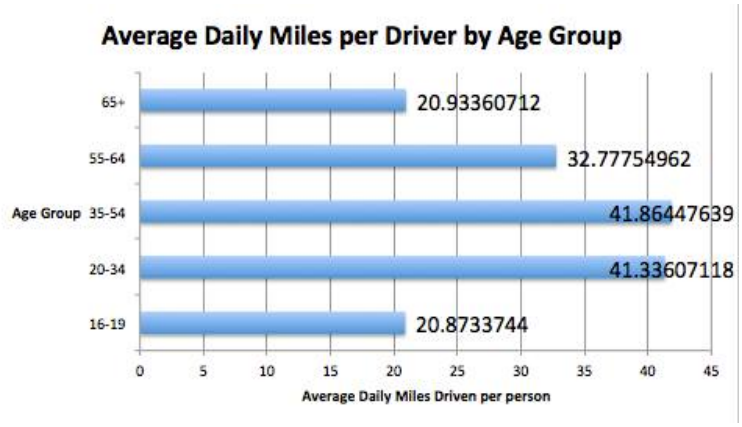

Figure 1: The average daily distances driven by age group reveals that driving distance for young and old people is low relative to the middle-aged.

Next, we determined the distribution of people among the low, medium, and high classifications of miles driven per day by analyzing trip data provided by the National Highway Traffic Safety Administration (NHTS) [10]. The NHTS provided data estimates for the annual miles driven M_A for over 300,000 people, which was converted to miles driven per

day M_D as follows in equation 2:

$$M_D = M_A \div 365.25 \ \frac{\text{days}}{\text{year}} \tag{2}$$

The distribution of M_D can be seen in the following graph:

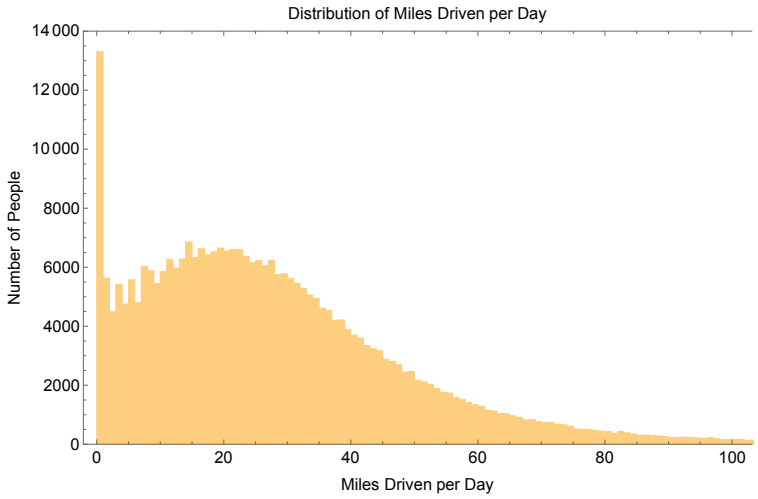

Figure 2: The distribution of miles driven per day for 300,000 people [10]

There is a notable spike at 0 miles driven per day, which is indicative of the greater number of people who choose not to drive as opposed to the number of people who drive small amounts. The distribution, discounting those who choose not to drive, suggests a skewed distribution of miles driven per day centered at around 30 miles/day.

Next, we had to take the data in this distribution and classify each person into one of three buckets: low, medium, or high. Using our calculated cut-offs for low, medium, and high miles driven per day, the data appears in proportions reflected in figure 3.

Figure 3 suggests that a majority of people (57.5%) drive a low amount of miles per day, which seems to be only minimally affected by the fact that 13300 (4.3%) of the people in the sample group drove less than 0.5 miles per day. Additionally, 30.3% of people in the sample group drove a high amount of miles per day and 12.1% of the people drove a medium amount.

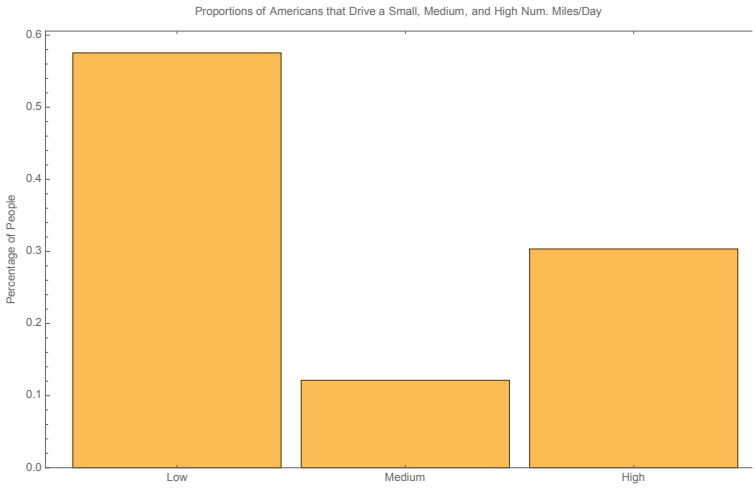

Figure 3: A bar graph showing the proportions of Americans with low, medium, and high daily driving distance.

Time Spent Driving per Day

In order to classify the categories for amount of time using the car per day, the exact same methods were applied as was used on the previous factor, miles driven per day, since the histogram for average daily driving duration by age group followed a similar bell shape as average daily distance driven by age. The histogram for average daily driving duration by age can be seen below. This method of classification by age group resulted in "low" daily driving duration being classified as less than 36.67 minutes, "medium" being between 36.67 and 45.33 minutes, and "high" being greater than 45.33 minutes.

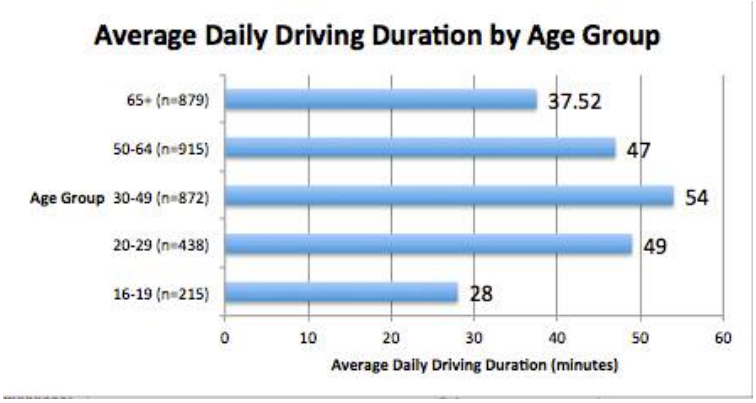

Figure 4: A bar graph showing the average daily driving duration by age group reveals that daily driving duration for young and old people is low relative to the middle-aged.

Next, we determined the distribution of people among the low, medium, and high classifications of amount of time spent driving per day by analyzing the same trip data provided by the NHTS [10]. The NHTS provided data on the durations of approximately one million car trips for specific individuals. The data was then manipulated in Ruby and Mathematica to find the sample distribution of daily driving duration, as shown below.

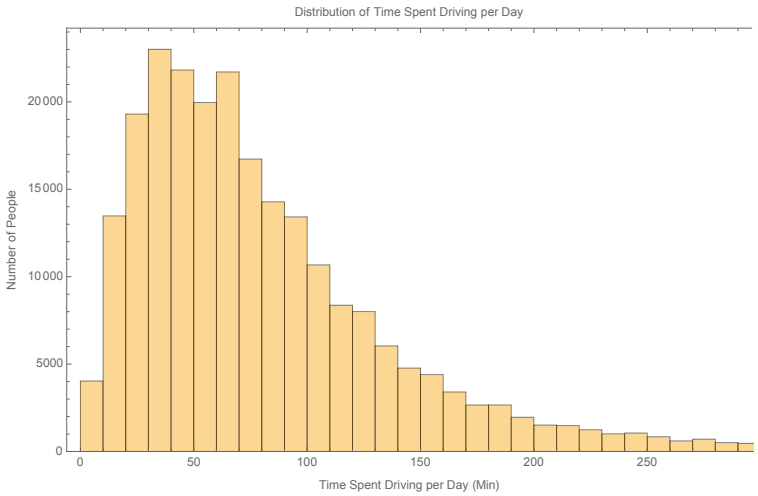

Figure 5: A bar graph showing a right-skewed distribution of time spent driving per day, as created in Mathematica.

Next, we had to take the data in this distribution and classify each person into one of three buckets: low, medium, or high. Using our aforementioned calculated cut-offs for low, medium, and high daily driving durations, the data appears in the following proportions:

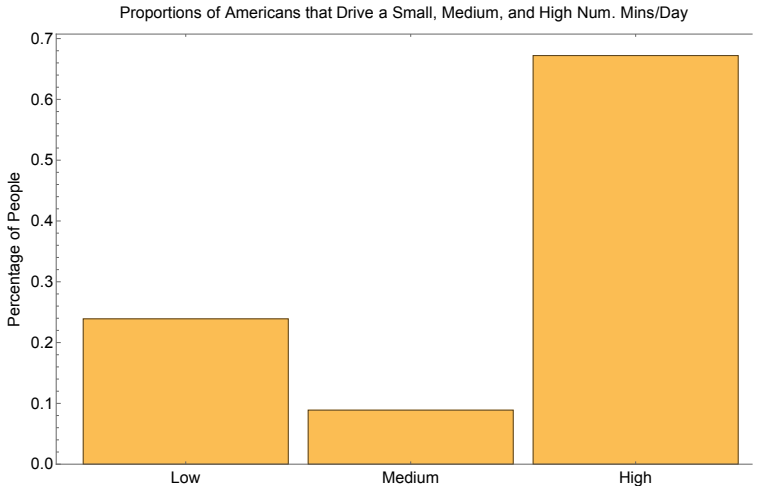

Figure 6: A bar graph showing the proportions of Americans with low, medium, and high daily driving durations.

This suggests that a majority of Americans have a high daily driving time (67.2%), while 23.91% have a low daily driving time and 8.89% have a medium daily driving time. This differs from daily distance driven, wherein a majority of people have a low daily driving distance (57.5%). This can be somewhat attributed to the possibility that driving shorter distances can have high time:distance ratios, since local roads would be taken rather than high-speed routes such as highways.

By finding the conditional proportion for any combination of the three, daily mileage

and daily driving duration categories, the probabilities of all combinations can be found via simple arithmetic. The results are shown below.

		Daily Duration Category		
		Low	Medium	High
Daily Mileage Category	Low	13.75%	5.11%	38.64%
	Medium	2.89%	1.08%	8.13%
	High	7.24%	2.69%	20.36%

Figure 7: A table showing the proportions of combinations of categories for daily driving time and daily driving distance.

While it makes sense that high daily mileage corresponds to high daily duration with a high probability, values with medium categorization result in very low probabilities, causing results to not follow a positive linear correlation when rising from low to medium to high categorizations, as one would expect the model to roughly follow since it generally takes larger amounts of time to drive long distances.

2.3 Case Study

In order to allow the companies to determine what a good city to use might be, it is wise to consider the average person in a city, and whether they would be prone to driving a low, medium, or high amount of time and travel.

In order to simplify this process for the company, we utilized Mathematica's machine learning capabilities in order to classify several variables for 1000 people from the NHTS survey along with their corresponding actual classifications [10]. We utilized Mathematica's ability to learn the trends in this data, and generated a nearest-neighbor classification approach which output a predicted classification given the following variables: Gender, whether the person lives in an urban or rural area, whether the person uses the Interstate, whether the person uses public transportation, the person's vehicle type, and the person's age, the person's state, and whether they're a worker.

We tested the following case study Billy:

```
In[43]:= billy = {sex["male"], urbrural["urban"], useintstate["no"], usepubtr["yes"],
           vehtype["car"], 40, "CT", worker["yes"]}
Out[43]= {1, 1, 2, 1, 1, 40, CT, 1}

In[47]:= milesPredictor[billy, "Probabilities"]
Out[47]= <|High → 0.0905797, Low → 0.090942, Medium → 0.818478|>

In[48]:= travelPredictor[billy, "Probabilities"]
Out[48]= <|High → 0.958031, Low → 0.0286836, Medium → 0.013285|>
```

Figure 8: Our classification of Billy where milesPredictor outputs the predicted classification for miles driven and travelPredictor outputs the predicted classification for time spent driving

The case study verifies with 82% certainty a medium distance driven per day and with 96% certainty a high amount of time spent per day. This classification makes sense as since Billy is a near-middle aged urban worker his profile seems like one likely to be driving a decent amount of mileage per day, and Billy also lives in an urban area, doesn't use the

interstate, and must commute by car to work so it's likely he's going to end up in some potentially long rush-hour traffic in the average case.

2.4 Assessment of Model

Sensitivity Analysis

To test the sensitivity of our model, we wanted to adjust the cutoff values of our three categories. Since our proportions for the "medium" categorization were low for both factors, we decided to adjust the size of the "medium" range by 1.0, 2.0, and 5.0 miles for daily driving distance. The effects on the proportions of predicted Americans in the "medium" category for daily driving distance is graphed below.

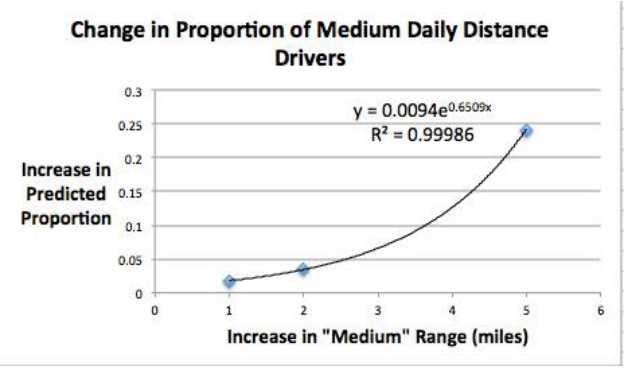

Figure 9: Sensitivity analysis of median daily driver distance data.

From figure 9 it can be seen that increasing the "medium" daily driving distance cutoff range by 1 or 2 only increases the predicted proportion of "medium" daily distance drivers by only 1.77% and 3.50% respectively, but greater increases in the cutoff range lead to an exponential increase in the predicted proportion, as can be seen by the best fit curve. A similar sensitivity would be found for daily driving duration since the variable being tested was the cutoff values for categorization, which was rooted in the same assumption (that categories can be created with data by age group) for both factors. However, the sensitivity analysis does not invalidate the age group-rooted basis of our categorization system, as categorizing quantitative values is a subjective process. That said, our results from figure 7 indicate that our cutoff values should be adjusted for the reasons discussed in section 2.2.

3 Zippity do or don't?

3.1 Assumptions, Simplifications, and Justifications

- **Assumption:** The availability of parking space in a given area has a negligible effect on the efficiency of a floating model.
  Justification: Floating and other such car sharing models minimize vehicle congestion and traffic. One Zipcar statistically takes fifteen to twenty personally owned vehicles off the road [11]. Any such geographic imbalance that could cause parking shortage is made up for in the gained available space.

- **Assumption:** To the consumer, the only differentiating factors between the one-way floating and station models are the increased convenience and costs associated with the floating model. As an extension of this assumption, the ratio of number of participants in the floating model to the number of participants in the station model holds for all locations, because costs and convenience increases should be consistent.
  Justification: The cost of "jockeys" that is added in the floating model have no effect on consumer preferences aside from making car-finding more convenient and increasing the cost of car rental, and aside from the addition of these "jockeys," there is no differentiating factor between the operation of the floating model and the station model. In addition, market forces of supply and demand should determine the proper ratios of consumer numbers in the station model to the consumer numbers in the floating model.

- **Assumption:** The number of vehicles per one-way car-sharinig station is approximately equal to the average in studies.
  Justification: This is due to the limited availibility of data on vehicles per station.

3.2 The Model and Results

Model Design

- **Round trip car sharing**
  Since our goal was to find the most likely participation rates, we wanted to find a model that most accurately described real-life data. Thus, we needed to first obtain real data to use for testing our model. We found that the round trip car sharing business model is both traditional for the car rental system, and also has prevalence in the car-sharing industry, since large car-sharing companies such as Zipcar (the largest) use a round trip, station-based system. [12][13] We were able to find literature that related one-way car sharing stations to round trip car sharing: one-way car sharing attracts approximately 3 to 4 times as many participants as round trip car sharing does. Therefore, we will first find an accurate model for the existing car sharing station model (round trip), and then multiply the participation values by 3 and 4 to see the participation rates for the one-way car sharing station model.

  Since this model for car sharing is the most prevalent in the real world, and thus was the most crucial part of our solution. We were most able to compare our equations in this section to the real world, and all regression techniques had a strong grounding in real life. We originally believed that many factors would affect the outcome of participation rates in any given city, listing variables such as the cost of gas, the number of vehicle crashes, the distance traveled by the cars, the population of the city, the population density of the city, and various other factors. The cost of gas increases overall costs for the participant due to increased travel expenses; an increased number of vehicle crashes should lead to increased traffic jams and therefore a smaller incentive to use the ride sharing service; the average distance traveled by cars, if small, should encourage walking rather than driving, and, if large, should encourage one-way travel; the population and population density of the city should affect the participation rates positively due to proximity of stations.

For our dependent variable, we used the number of stations in a given city as the measure of how popular a car-share service was in a city, since participation rates were not found during research. It is reasonable to assume that participation numbers are directly proportional to the number of sharing stations in a given city, because all plans to create stations must be reasonably supported by a volume of participants, or the business (sharing service) would be making a loss.

In order to account for all of our variables, we decided to employ multiple linear regression to find a fit among cities we have station data for, then apply the model to the four cities of Poughkeepsie, Richmond, Riverside, and Knoxville. To perform a multiple linear regression, we first eliminated independent variables that are correlated to others, like population and population density, as having a correlation within the independent variables create duplicates of the same factor and will reduce the accuracy of the model. Then, we fit the data using statistical analysis tools, and removed any independent variables that related to the dependent variable with little or no statistical significance and performed multiple linear regression on the remaining independent variables.

The multiple linear regression model is given by:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_n\beta_n, \tag{3}$$

where Y is the dependent variable, which is the number of stations there are in a given city, X_i are the independent variables, β_0 is the constant (the Y-intercept), and all other β_i's are the scaling factor for their corresponding independent variables. We must solve for the set of β values that create the smallest absolute error between the model and the data.

- **One-way car sharing station model**
  Since this model's participation is said by experts to attract 3 to 4 times as many participants as the round trip model [12], we can multiply the round trip model by a factor, either 3 or 4, to get the participation for this one-way model. We will relate the two business models in terms of their participation rates by

$$S = kR, \tag{4}$$

  where k is either 3 or 4, R is the participation for the round trip car sharing model, and S is the participation for the one-way car sharing station model.

- **One-way car sharing floating model**
  The one-way car sharing floating model is relatively new to the industry [12], and its only theoretical benefit to the customer over the one-way car sharing station model is the increased convenience and costs, so participation in this model should be some constant multiple of the participation in the station model. We will relate the two business models in terms of their participation rates by

$$F = jS, \tag{5}$$

where j is some factor to be determined by actual calculations, F is the participation for the one-way car sharing floating model, and S is the participation for the one-way car sharing station model.

- **Fractional ownership**
  In the United States, the majority of people are employed (4.9% of people are unemployed) [14], which implies sharing a car between more than two people is extremely inconvenient due to scheduling problems, maintainence, and responsibilities for keeping the vehicle in good condition. We reason that such inconveniences associated with owning a shared car would greatly decrease the participation in any city to sub-competitive percentages and numbers. For the reasons that people would want to participate in a car-sharing service, this model would not be logical.

Model Results

- **Round trip car sharing**
  To eliminate the multicolinearity of independent variables from our multiple linear regression model, we first take a qualitative approach. The factors we believed would affect participation were the cost of gas, the number of vehicle crashes, the distance traveled by the cars, the population of the city, and the population density of the city. The cost of gas should not be correlated with any of the other independent variables, the number of vehicle crashes should be related to the number of total drivers (and therefore the population), the distance traveled by cars per day should be relatively constant over the country, the population of the city would be correlated to the population density and the total number of drivers, so we must first eliminate the population variable from our analysis. Then, we can divide the number of vehicle crashes by the population to obtain a crash rate, which should not be related to the other variables. In the end, we are able to use the vehicle crash rate (we chose fatal crashes due to their significance in traffic interruptions), and the population density of the city. A correlation analysis of the data showed that gas price has a -0.05 correlation with crash rate and a 0.29 correlation with population density, and crash rate had a 0.15 correlation with population density.

  We took crash rate, gas costs, and population data from various sources [15][16][17] relating to several of the most popular car sharing cities. However, gas price turned out to have a very high p-value for multiple linear regression and a correlation of only 0.12 with the number of stations in a city (fatal crash rate and population density had a -0.70 and 0.64 correlation with station numbers, respectively), and was therefore deemed statistically insignificant.

  Our table of values for remaining used variables is shown below:

City	Fatal Crash Rate	Population Density	Number of Stations
New York	3.89E-05	27012.40	716
Chicago	6.39E-05	11957.08	630
Miami	1.74E-04	11996.99	297
Portland	5.81E-05	4641.83	537
Washington, D.C.	5.62E-05	10792.68	521
Seattle	5.09E-05	7962.14	489
Austin	7.01E-05	3064.09	392
San Diego	5.58E-05	4246.96	411

Table 1: Table of values gathered from sources on the fatal crash rate in each city (per capita), the population density (people per square mile), and the total number of car sharing stations in the city [16][15][18].

We use the Analysis ToolPak in Microsoft Excel to carry out the multiple regression analysis. For equation 3, we have $\beta_0 = 513.46, \beta_1 = -241200, \beta_2 = 0.01754$, where β_0 is the constant term in the equation (the Y-intercept, β_1 is the scaling factor for the fatal crash rate, and β_2 is the scaling factor for the population density. Therefore, our equation 3 becomes

$$Y = 513.46 - 241200X_1 + 0.01754X_2, \tag{6}$$

where Y is the predicted number of stations in a given city, the X_1 is the fatal crash rate in a given city per capita, and X_2 is the population density in a given city. We apply this model to the four cities of Poughkeepsie, NY; Richmond, VA; Riverside, CA; and Knoxville, TN using the independent variables corresponding to them and calculate the expected number of stations in each city. Our resulting data table is found below.

City	Fatal Crash Rate	Population Density	Number of Stations
Poughkeepsie, NY	1.97E-04	6363.90	201
Richmond, VA	9.64E-05	3642.42	378
Riverside, CA	1.10E-04	3937.69	355
Knoxville, TN	2.06E-04	1870.49	134

Table 2: The table of values that corresponds to the four cities found in the problem. Note that in this table, the number of stations column is calculated using equation 6 rather than observed. The fatal crash rate is in fatal crashes per capita and the population density is in population per square mile.

Since we assumed that the number of car-sharing stations in a given city is directly proportional to the number of car sharers in a given city, we can calculate this from data. A study of a Seattle one-way car sharing experiment showed that there were 5 people using one car every day [19]. We assume that the turnout rate for round-trip car

sharing will be approximately 3 to 4 times smaller, since the number of participants in a one-way service is 3 to 4 times as many as those in a round-trip car service [12]. Also, a study found that there were approximately 10 cars per station in a rental system [20]. Therefore, we can perform the following dimensional analysis:

$$\frac{\text{Participants}}{\text{City}} = \frac{\text{Participants}}{\text{Vehicle}} * \frac{\text{Vehicles}}{\text{Stations}} * \frac{\text{Stations}}{\text{City}} \tag{7}$$

We can then multiply our values of 10, 5/3 or 5/4, and the values in our earlier tables for number of stations per city together to obtain an approximate number of total participants in each city.

City	Number of Stations	Participation (Max)	Participation (Min)
Poughkeepsie, NY	201	3348	2511
Richmond, VA	378	6308	4731
Riverside, CA	355	5909	4431
Knoxville, TN	134	2239	1679

Table 3: The table of participant values corresponding to the four cities found in the problem. The participation (max) column assumes a factor of 3 times more participants in the one-way trip than the round trip, while the participation (min) column assumes a factor of 4 times more participants in the one-way trip than the round trip. From this table, we can conclude that Richmond, VA is most fit for this business model.

- **One-way car sharing station model**
  Since we have determined that the one-way car sharing station model can attract three to four times more participants than the round trip model, we simply multiply the values found in the round-trip table model by 3 and 4 (there are no more max and min values). Therefore, the table of values for the one-way car sharing station model is:

City	Number of Stations	Participation
Poughkeepsie, NY	201	10043
Richmond, VA	378	18923
Riverside, CA	355	17726
Knoxville, TN	134	6716

Table 4: The table of participant values corresponding to the four cities found in the problem. The participation column is derived from Table 3. From this table, we can conclude that Richmond, VA is most fit for this business model.

- **One-way car sharing floating model**
  Using the case study of one-way car sharing floating models in Seattle, we determine the total participation in Seattle for the one-way floating model, then compare this

number to the total participation in Seattle for the one-way station model, and find the ratio of the two. We assume that this ratio holds for all cities, justified by the earlier assumption that the only factors differentiating the floating from the station model are the increased convenience and costs of using the floating model (the consumer is the one responsible for determining this relationship). We found from the Seattle case study that there were 500 vehicles in the floating model and 5 rides per vehicle per day, which amounts to 2500 rides per day [19]. In Seattle, we also found that there were 489 stations for the one-way car sharing station model, which amounts to 24450 rides per day if we assume 10 vehicles per station and 5 rides per vehicle per day (rides per vehicle per day in Seattle and vehicles per station data come from earlier studies) [20][19]. This results in a ratio of participants in the floating model to participants in the station model of:

$$j = R_{FS} = 2500\frac{\text{rides}}{\text{day}} : 24450\frac{\text{rides}}{\text{day}} = 0.1022, \tag{8}$$

where j (from equation 5) and R_{FS} is the ratio of participants per day in the floating model to participants per day in the station model. If we extend this ratio to the four cities in the problem, we arrive at alarmingly low participation rates that are even lower than those of the round-trip station model. They are:

City	Participation (Station)	Participation (Floating)
Poughkeepsie, NY	10043	1026
Richmond, VA	18923	1934
Riverside, CA	17726	1812
Knoxville, TN	6716	686

Table 5: The table of participant values corresponding to the four cities found in the problem. The second column in the table is taken from table 4, and the third column is calculated from the second column based on the ratio R_{FS}, equal to 0.1022. From this table, we can conclude that Richmond, VA is most fit for this business model, but is not ideal, compared to either the round trip station model or the one-way car sharing station model.

Most likely, these decreased participation values and lack of popularity are most likely caused by the increased time spent searching for a parking spot in the floating model (perhaps this decrease in convenience outweighted the increase in convenience caused by versatile access to the vehicle), or the potential increased rental costs of using the car.

- **Fractional ownership** We have reasoned the fractional ownership model will be infeasible due to its associated inconveniences, so we will not be performing computations on data.

  Based on the outputs of our mathematical models of the four car-sharing business models, we found that the one-way car sharing station model would garner the most participation per day in any given city. Tables 3 and 5 can be used to compare the results.

3.3 Ranking

Utilizing the one-way station-organized car sharing option, which we determined to be the option that yields the most participation out of the four, the chosen cities rank (1) Richmond, VA (2) Riverside, CA (3) Poughkeepsie, NY (4) Knoxville, TN.

3.4 Case Study

In order to extend our model to another urban popualtion in the United States, we applied it to the population density [15] and occurence of fatal crashes [16] in Los Angeles, CA. In support of our model's validity, the resulting participation was the quantity 23637, or approximately twice that in Poughkeepsie, NY, which we interpreted as a rational and correct application due to increased population density and decreased fatality rates in LA compared to Poughkeepsie.

3.5 Assessment of Model

In the one-way car sharing station model, the multiple linear regression yielded p-values of 0.001 for the constants β_0, 0.022 for β_1, and 0.081 for β_2. According to the Handbook of Biological Statistics, a valid p-value for multiple linear regression satisfies the condition $p < 0.15$. Our values satisfy this and the regression as a whole has a Multiple R value of 0.881. This means that our model is an acceptable fit for the situation, especially given the large number of variables involved in the real world that we were not able to account for.

Normal Probability Plot of Residuals

For our multiple linear regression analysis, we found the normal probability plot in addition to the Multiple R value and the p-values. The plot is shown below.

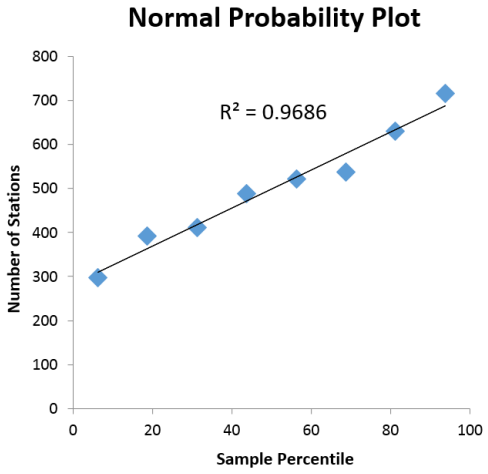

Figure 10: A normal probability plot generated for our multiple linear regression. The high R^2 value of 0.9686 shows that the regression was largely successful and not subject to data skewing, unlike some alternative outcomes were the data to be skewed [21].

Sensitivity Analysis

One key part of our model is that we assume the number of vehicles per station is equal to the average found in the studies, but we can vary this assumption and test for how much the model is affected by this assumption. We vary the average number of vehicles by ±10% and ±20% and see their effects on the projected participation numbers in a random city (our random number generator picked the third city in the problem, Riverside, CA). At 0% change, the number of participants remained at 100%. At a 10% change, the number of participants increased to 110% or decreased to 90%, and at a 20% change, the number of participants increased to 120% or decreased to 80% as well. This is due to the multiplicative nature of our model and its dependence on the number of vehicles per station. To solve this, we simply have the user (the company with an interest in participation rates) multiply by the number of vehicles in their stations rather than use the default average value. The multiple linear regression itself is done before this multiplication, and is generally free of flaws, as seen from the normal probability plot and the analysis of the model in earlier sections.

For the one-way floating model, we depend on another assumption that the ratio, R_{FS} was constant across all locations, and this was also part of a multiplicative model. Changing this ratio by a percentage resulted in the same percentage change in the participation number. However, the only way that the floating model would outcompete the station model is that the ratio changes by over 1000%, which is extremely unlikely.

4 Road map to the future.

Most car shares reduce the environmental impact of driving and transportation. They typically offer newer, low emission vehicles. Members report driving less, using public transit more, and opting out of private car ownership. Each car share vehicle replaces from 4 to 15 personal vehicles on the road. Car sharing also leads to reduced congestion and less time looking for a parking space, further reducing emissions.

4.1 Assumptions, Simplifications, and Justifications

- **Assumption:** While the fuel economy of renewable-energy cars will help car sharing companies financially, there is no reason to postulate that the sheer "green" appeal of renewable-energy cars will increase ridership.
  Justification: Firstly, according to a recent poll, only 32 percent of Americans personally worry about global warming and climate change, meaning that an even smaller percentage would change their habits to prevent global climates from rising. [22]. Additionally, car sharing with non-renewable energy cars already causes members to drive less, use public transit more, and opt out of private car ownership, resulting in shared vehicles replacing 4 to 15 personal vehicles on the road [1]. This results in less traffic and less time looking for a parking space, further reducing emissions [1][23]. Thus, people who care enough about the environment to switch from personal car driving to shared car driving will likely have already switched to shared car driving even without renewable-energy cars being offered by car share businesses.

- **Assumption:** While the current self-driving cars have technological difficulties engaging effectively in the human-consumer world, future models will be superior.

Justification: Self-driving car technology has only recently debuted, and with companies like Google investing in the innovation's progress, it is valid to picture a world in which they function appropriately.

4.2 The Model and Results

Model Design

To model the effect of a future environment in which cars are both self-driving and run on renewable or electric energy, we utilized the model we created for the method of car sharing organization that garnered the most participation. Within this model, the variables we analyzed were the popualtion density of a given area and the rate of fatal car accident occurences, which served as an implicit factor for analyzing vehicle congestion and the general automotive environemnt.
In the future described, in which self-driving and energy-efficient cars dominate these car sharing systems, our model would flux in that self-driving cars have different crash rates than traditional vehicles. Currently, self-driving cars' accident rates are double those of human-operated vehicles [24]. However, we also reasoned that in a farther hypothetical future, these cars will have accident rates of nearly zero as technology develops further and such engineering efforts see their goals realized. We thusly developed two models, one for which the fatal accident rate is doubled to reflect the present situation, and one in which the fatal accident rate is zero to account for technological advances in the farther future.

Model Results

City	Fatal Crash Rate	Number of Stations	Participation
Poughkeepsie, NY	3.93E-04	-203.20	-10160
Richmond, VA	1.93E-04	180.38	9019
Riverside, CA	2.19E-04	129.42	6471
Knoxville, TN	4.12E-04	-289.40	-14470

Table 6: The table of participant values corresponding to the four cities found in the problem, with the doubling of the fatal crash rate due to an implementation of self-driving cars. Note that this doubling is only temporary, and will likely level off over time. The participation values did become negative, which shows that the increased fatality rates will lead to an infeasible business model.

Doubling the vehicle accident rates did not change the rankings for user participation in the four given cities. The ranking remained the same and is (1) Richmond, VA (2) Riverside, CA (3) Poughkeepsie, NY (4) Knoxville, TN

City	Fatal Crash Rate	Number of Stations	Participation
Poughkeepsie, NY	0.00E+00	604.93	30246
Richmond, VA	0.00E+00	576.53	28827
Riverside, CA	0.00E+00	579.61	28981
Knoxville, TN	0.00E+00	558.05	27902

Table 7: The table of participant values corresponding to the four cities found in the problem, with the fatal crash rate going to zero due to an increase in the safety of self-driving cars over a long period of time. Note that this has changed the rankings of participation among the cities examined.

Farther future projections in which the fatal crash rate will be near zero resulted in a new ranking for participation in the four given cities. This new ranking is as follows: (1) Poughkeepsie, NY (2) Riverside, CA (3) Richmond, VA (4) Knoxville, TN.

Sensitivity Analysis

The model here is derived from the same multiple linear regression as the previous section, which was analyzed in-depth already. As a result, sensitivity analysis was omitted from this section.

5 Strengths and Weaknesses

We have assessed our solution's strengths and weaknesses as a whole.

5.1 Strengths

For part one of the problem, our model uses data of very large sample sizes from government agencies [10][8][9], making our predicted percentages of Americans in each of our categories very accurate. Additionally, our machine-learning algorithm is able to individualize the process of categorizing people's driving habits based on certain variables, making our model very adaptable to any business' geographical location, since the population demographics of any given city in America will vary.

For part two of the problem, our models utilized population density statistics from eight of the most car-sharing-active cities in the country, which provided an accurate array of data for national cities in which car sharing was not only extant but popular. This strategy enhanced the accuracy of our models and predictions.

For part three of the problem, we developed two models that projected both a near-future result of utilizing fuel-efficient and self-driving vehicles and a hypothetical, ideal future. Our choice to use two models enhanced the accuracy of our predictions and their dual viability.

5.2 Weaknesses

For part one of the problem, our categories were created based on age ranges. Categorizing quantitative data is a very subjective process, so if a company wanted to create categories based on another variable, our categories would not apply. Additionally, our bar graphs displaying the proportions of Americans in each category (figures 3 and 6), both showed relatively low values for the "medium"category, indicating that businesses may want to adjust

our category cut-off values to provide more evenly distributed proportions (for example, some of the higher values in the low category and some of the lower values in the high category could get moved to the medium category).

For part two of the problem, our models for the four car sharing options were modeled from the total number of car sharing stations in eight cities, which we interpreted to imply the participation in these cities. The number of these stations, however, excludes floating car sharing plans.

For part three of the problem, one of our models projects a hypothetical future in which the number of accidents involving self-driving cars is reduced to nearly zero. However, there is a possibilty that self-driving cars will be deemed uneconomical and be overtaken by other transportation innovations, transforming the national transport landscape that will rely on variables we have not taken into account.

6 Conclusion

In our rapidly evolving world, connections are defining the new status quo, and members of the rising generation will live in a world in which cooperation contribute to the wellbeing of international society. This trend is no less apparent in the popularity of car sharing, for which a near future indicates the increased value and popularity of more economic, green, and practical modes of transport.

Using age groups as the basis of our categorization system, it was found that 57.5% of Americans drive a "low" daily distance, 12.1% drive a "medium" daily distance, and 30.3% drive a "high" daily distance. In regards to daily driving duration, it was found that 67.2% of Americans have a high daily driving time, 23.91% have a low daily driving time, and 8.89% have a medium daily driving time. Since the probabilities for all nine combinations of do not have linearly increasing probabilities as the category of both factors are "increased" from low to medium to high, it is evident that the cutoff values for each category should be adjusted.

We utilized Mathematica's machine learning capabilities to create a program that can predict, with relatively good certainty, the classification of a person's driving habits based upon some easily-acquirable variables. This can be used by some company to find an ideal city based upon the average person/target audience in a city.

Of the four car sharing options, our models analyzed population density and the rate of accidents to determine that the one-way station model is the most efficient. Utilizing this model to maximize participation, we evaulated that the order in which car sharing companies should move forward investigating legalities and implementaton is (1) Richmond, VA (2) Riverside, CA (3) Poughkeepsie, NY (4) Knoxville, TN.

Finally, the transportation industry is ever growing and developing, and with this clairvoyant attitude, we analyzed the future of the car sharing industry with the entrance of self-driving and environmentally friendly vehicles. We developed two models that altered crash-rate statistics under the assumption that current self-driving cars engage in more accidents, whereas in a hypothetical future, self-driving cars will approch negligible amounts of accidents. This led us to conclude that with current crash-rate statistics on self-driving cars, the ranking of the four given cities remains constant; however, in the future, ideal scenario, the ranking would change to first Poughkeepsie, NY, followed by Riverside, CA; Richmond, VA; and Knoxville, TN.

7 References

[1] Institute for Transportation Development Policy. Car sharing. 2012. [Online; accessed 27-February-2016].

[2] Kara Swisher. Gm invests $500 million in lyft and strikes strategic autonomous car alliance. *re/code*, January 2016. [Online; accessed 27-February-2016].

[3] Reuters Matthew DeBord. Ford is betting $4.5 billion on radically transforming the company and its mission. *Business Insider*, December 2015. [Online; accessed 27-February-2016].

[4] Dominique Mosbergen. Most millennials won't own a car in 5 years, says lyft co-founder john zimmer. *The Huffington Post*, July 2015. [Online; accessed 27-February-2016].

[5] United States Public Interest Research Group. 21st century transportation. October 2014. [Online; accessed 27-February-2016].

[6] Susan Shaheen. One-way carsharing's evolution in the americas. *Move Forward*, July 2015. [Online; accessed 27-February-2016].

[7] Lauren Hepler. Zipcar, google and why the carsharing wars are just beginning. *GreenBiz*, July 2014. [Online; accessed 27-February-2016].

[8] U.S. Department of Transportation. Average annual miles per driver by age group. *OHPI*, February 2015. [Online; accessed 27-February-2016].

[9] U.S. Department of Transportation. Distribution of licensed drivers - 2010 by sex and percentage in each age group and relation to populatio n. *OHPI*, Sept 2011. [Online; accessed 27-February-2016].

[10] U.S. Department of Transportation. Our nation's travel. *National Household Travel Survey*, February 2011. [Online; accessed 27-February-2016].

[11] Zipcar. This idea is bigger than all of us. February 2015. [Online; accessed 27-February-2016].

[12] Chris Brown. Carsharing: State of the market and growth potential. *Auto Rental News*, March/April 2015. [Online; accessed 27-February-2016].

[13] Sally McGrane. Car sharing grows with fewer strings attached. *New York Times*, June 2013. [Online; accessed 27-February-2016].

[14] Bureau of Labor Statistics. Labor force statistics from the current population survey. *Databases, Tables Calculators by Subject*, February 2016. [Online; accessed 27-February-2016].

[15] US Census Bureau. Population estimates table. *Quickfacts*, July 2014. [Online; accessed 27-February-2016].

[16] Centers for Disease Control and Prevention. Motor vehicle crash deaths in metropolitan areas. *Morbidity and Mortality Weekly Report (MMWR)*, July 2012. [Online; accessed 27-February-2016].

[17] GasBuddy. Average regular gas price by us city. February 2016. [Online; accessed 27-February-2016].

[18] Jocelyn Milici Ceder. Car share infographic: Top 10 u.s. car sharing cities. *Walk Score*, February 2013. [Online; accessed 27-February-2016].

[19] Seattle Department of Transportation. 2013 seattle free-floating car share pilot program report. March 2014. [Online; accessed 27-February-2016].

[20] Tuenjai Fukuda Matthew Barth, Susan A. Shaheen and Atsushi Fukuda. Carsharing and station cars in asia. *Transportation Research Record: Journal of the Transportation Research Board*, March 2006. [Online; accessed 27-February-2016].

[21] Penn State University. Normal probability plot of residuals. 2016. [Online; accessed 27-February-2016].

[22] Jeffrey M. Jones. In u.s., concern about environmental threats eases. *Gallup*, March 2015. [Online; accessed 27-February-2016].

[23] Access Management Parking Strategies. On-street car share policy. September 2015. [Online; accessed 27-February-2016].

[24] Keith Naughton. Humans are slamming into driverless cars and exposing a key flaw. *Bloomberg Business*, December 2015. [Online; accessed 27-February-2016].

8 Appendix A

```
(* Import the data from the NHTS *)
trainData = Import[SystemDialogInput["FileOpen"]]
(* Remove the header from the datafeed *)
trainData = trainData[[2 ;; -1]]

(* Grab the mile and travel data from the training data *)
mileData = trainData[[All, {2, 3, 4, 5, 6, 7, 8, 9, 10}]]
travelData = trainData[[All, {1, 3, 4, 5, 6, 7, 8, 9, 10}]]

(* Set the limits for the classifications of miles driven per day and time travelled per day *)
milesLower = 27.87

milesMedium = 34.867

travelLower = 36.67

travelMedium = 45.33
(* Function classifies corresponding time or mile data *)

classifyMiles[x_] := If[x ≤ milesLower, "Low",
  If[x ≥ milesMedium, "High",
   "Medium"]]

classifyTravel[x_] := If[x ≤ travelLower, "Low",
  If[x ≥ travelMedium, "High",
   "Medium"]]

(* Initialize the miles training set *)
milesTrainingSet = {}

(* Populate the miles training set (correlating the variables to their proper classification *)
For[i = 1, i ≤ Length[travelData], i++,
 milesTrainingSet = Append[milesTrainingSet,
   {mileData[[i, {2, 3, 4, 5, 6, 7, 8, 9}]] → classifyMiles[mileData[[i, 1]]]}]]
(* Fix formatting of set *)

milesTrainingSet = milesTrainingSet[[All, 1]]
(* Initialize time training set *)

travelTrainingSet = {}

(* Populate and fix the formatting for the training set *)
For[i = 1, i ≤ Length[travelData], i++,
 travelTrainingSet = Append[travelTrainingSet,
   {travelData[[i, {2, 3, 4, 5, 6, 7, 8, 9}]] → classifyMiles[travelData[[i, 1]]]}]]

travelTrainingSet = travelTrainingSet[[All, 1]]

(* Use machine learning to create the predictors *)
milesPredictor = Classify[milesTrainingSet]

travelPredictor = Classify[travelTrainingSet]|
```

Figure 11: The Mathematica code used to create the machine learner that classified people to low, medium, or high miles driven and time spent driving based upon variables discussed in the Case Study section of the Who's driving section