CSS 490 Kevin Ulrich
Team Dota Robert Stewart
May 12, 2017 Danny Ly

# Iteration 2

## Features and Categories

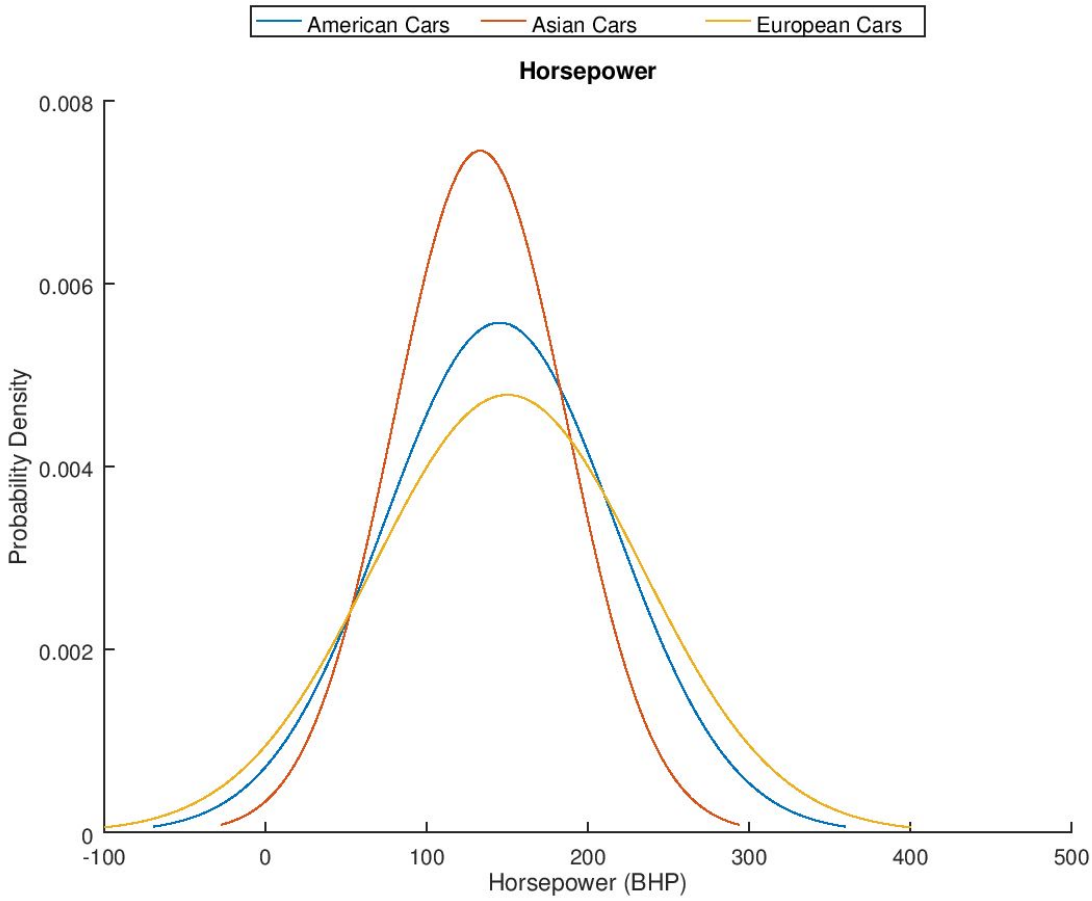
From the car observation data we have collected, we decided to limit the features used to maximum horsepower, maximum torque, fuel tank capacity, top speed, length, width, height, and curb weight. Every car observation has these features and they will help classify car observations into one of the three categories: American car, Asian cars, and European cars.

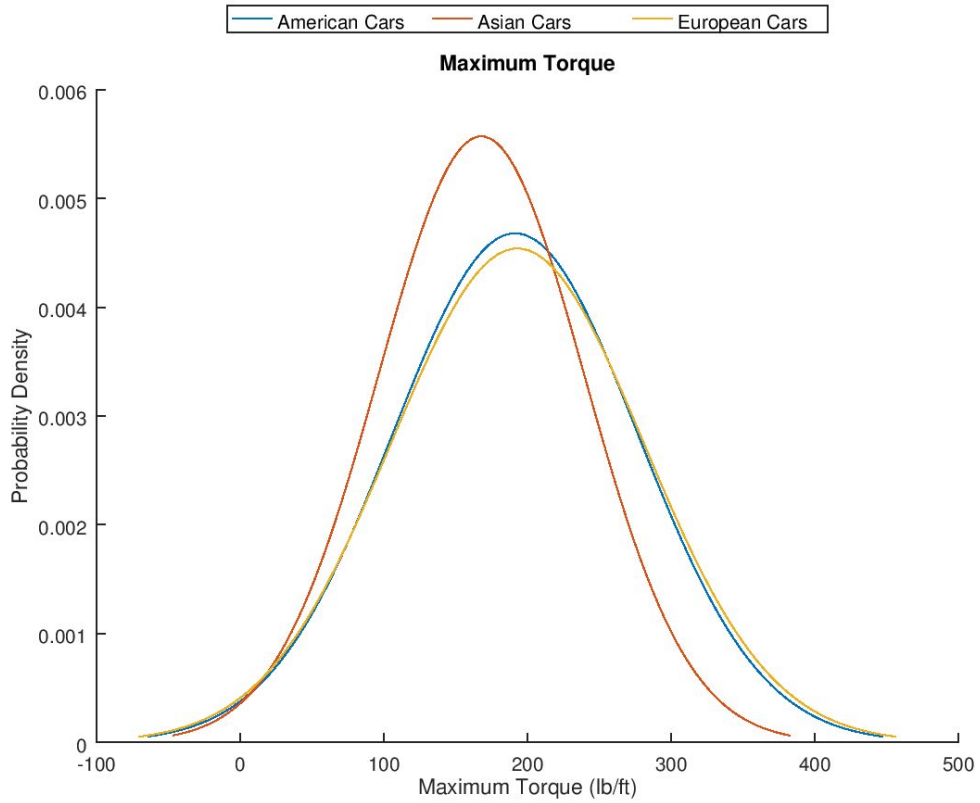
## Determination of Plots

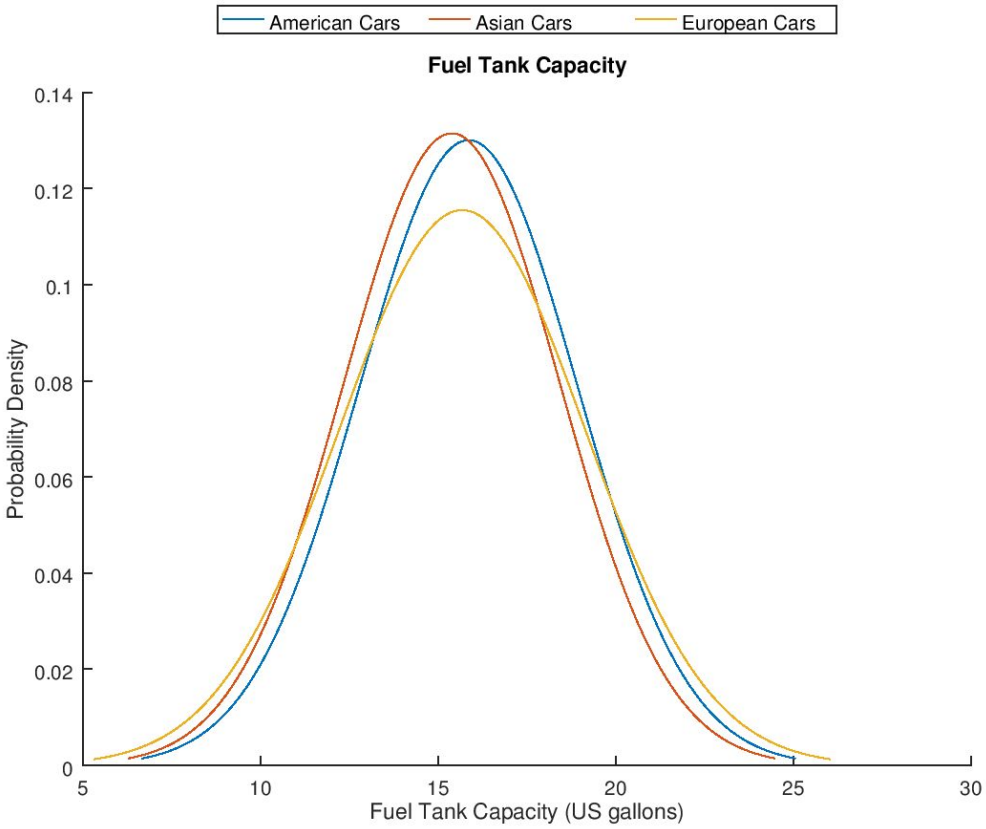
We decided that a 2D scatter plot summary, a probability density plot summary, and a histogram summary would be all be useful in analyzing our collected data.

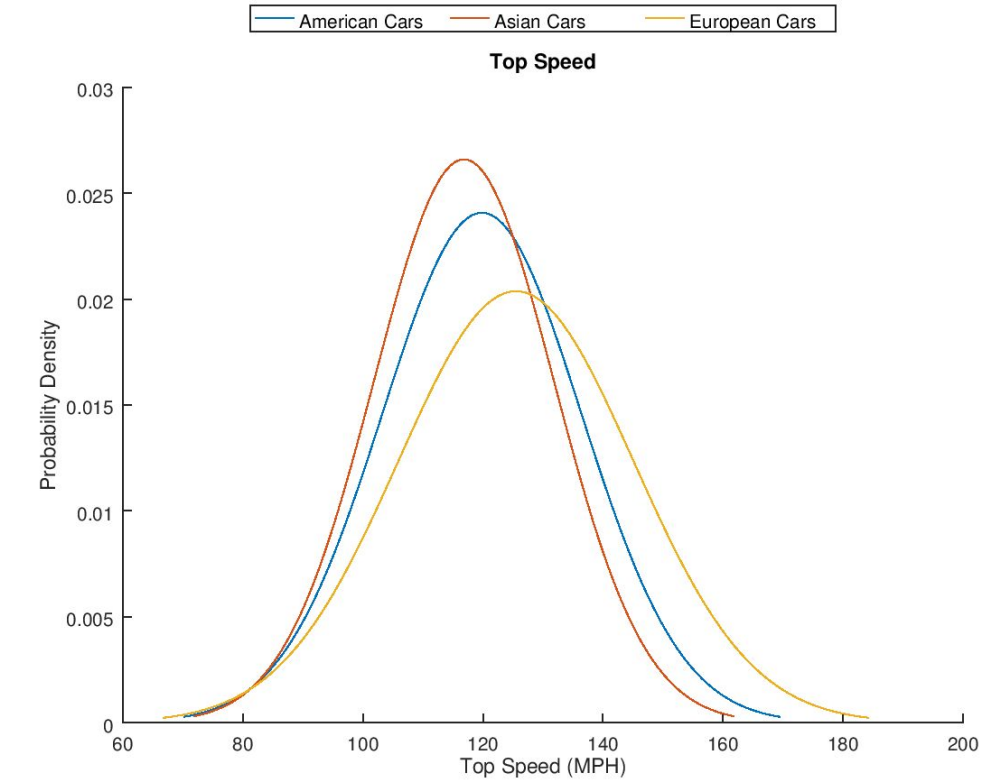
- We decided a histogram summary would useful for analyzing the spread of features across intervals as histograms capture frequency of observations. With the frequency of observations captured, features frequency can be closely analyzed in the bins that they are grouped in. For example, using a histogram would help in determining the normality of a feature as a normalized histogram should be in a bell shaped curve. Comparing the histograms between features will also help visualize patterns. For example, from the histograms that capture the frequency of car height, width, and length for american cars, it is clear that there are two groupings of sizes. This makes sense as car sizes can generally be grouped into larger SUV's/trucks and smaller compact cars.
- A 2D scatter plot summary is useful to be able to see correlation among the different classes. This will be useful for PCA, where we are interested in looking at inconsequential or highly correlated features, the latter being evident in scatter plots.
- We decided a probability density plot summary would be useful for informing the decision boundaries between the three different classes (regions). It's also a good way to tell, when combined with a dot diagram, whether or not normality can be assumed for a dataset.

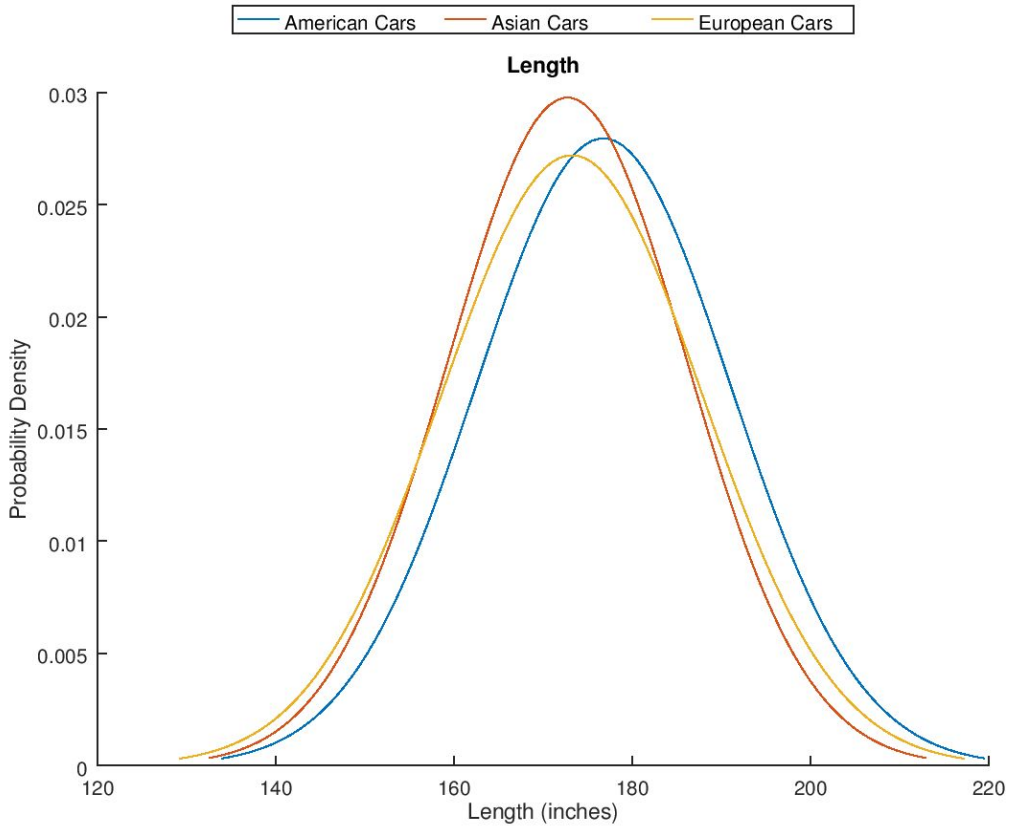
# Probability Density Plots

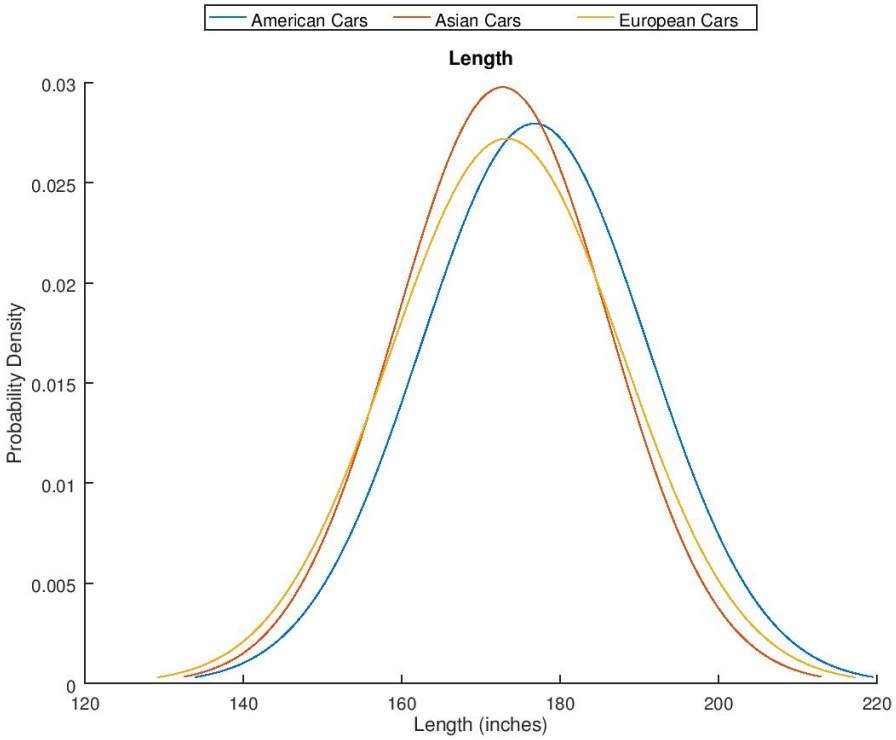


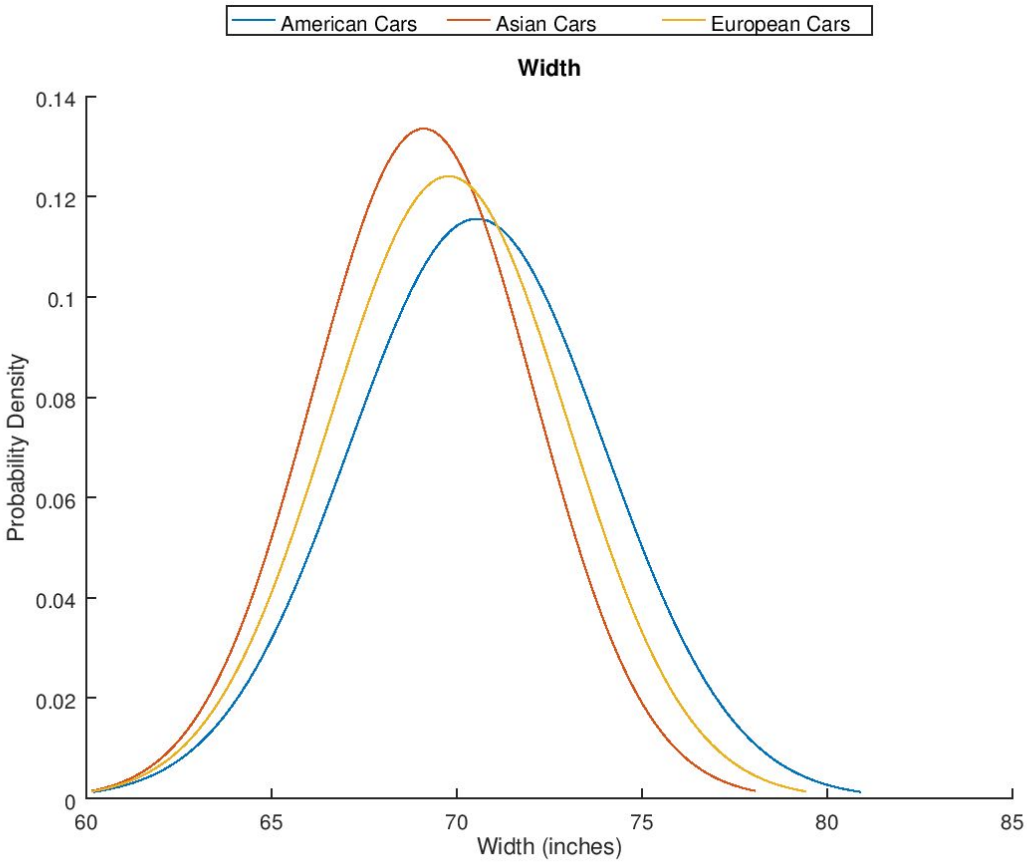

American Cars
Asian Cars
European Cars
Maximum Torque
Probability Density
0.006
0.005
0.004
0.003
0.002
0.001
0
-100
0
100
200
300
400
500
Maximum Torque (lb/ft)

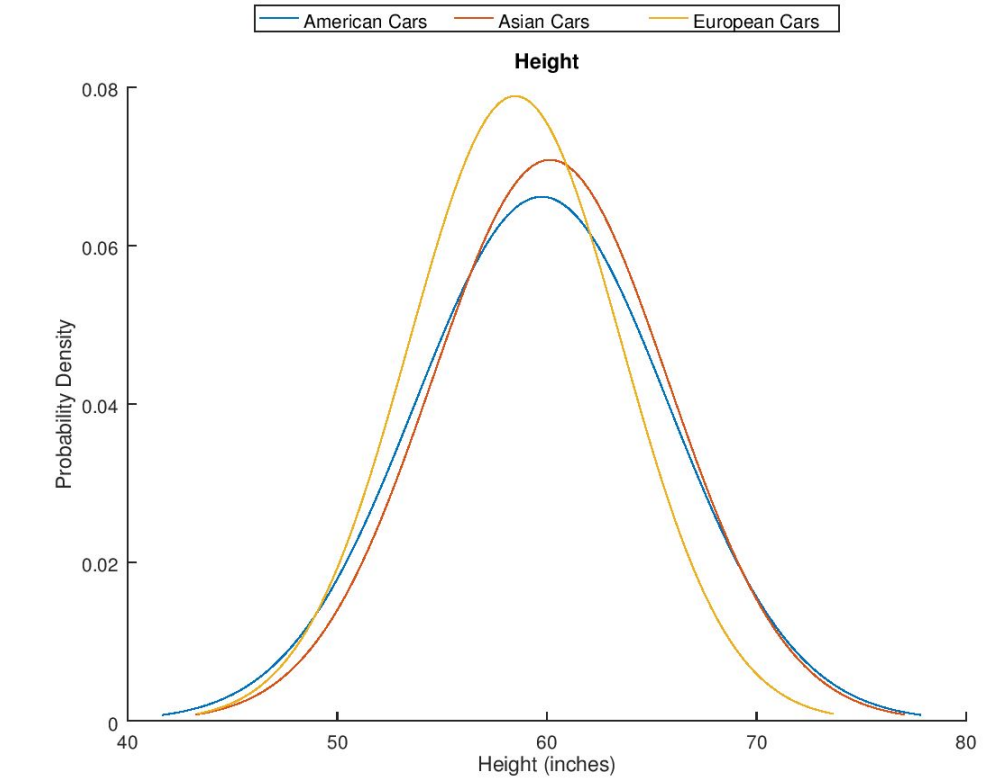

American Cars
Asian Cars
European Cars
Fuel Tank Capacity
Probability Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
15
20
25
30
Fuel Tank Capacity (US gallons)

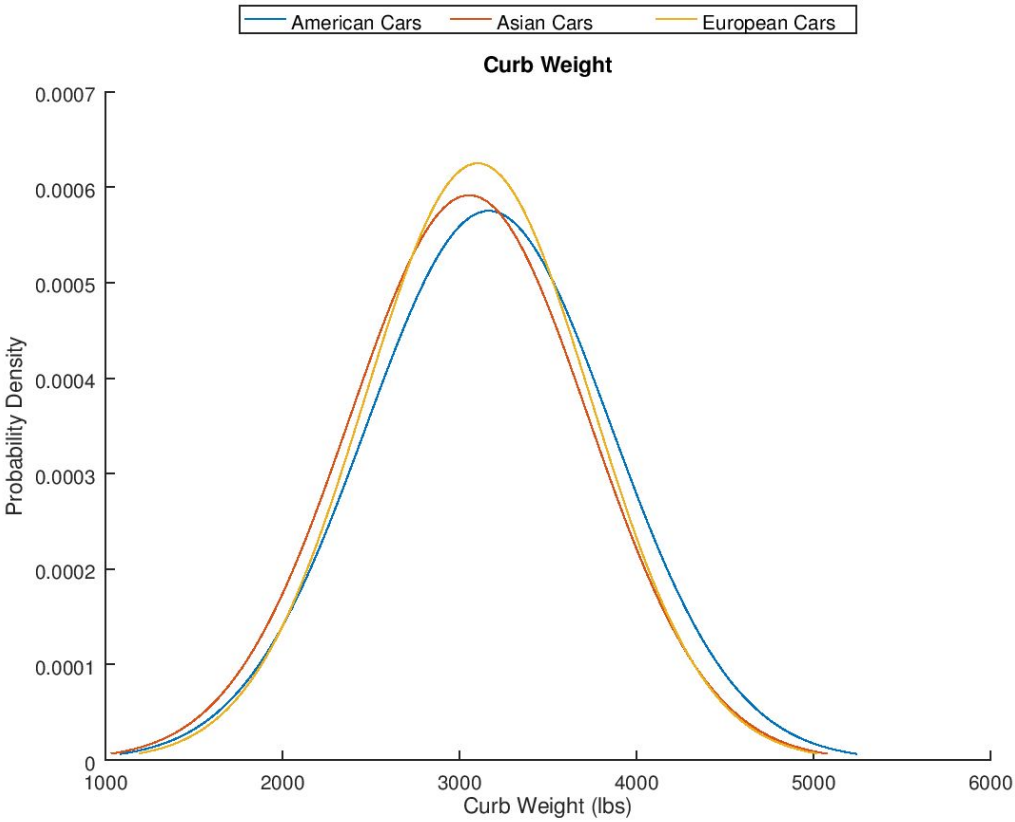

American Cars
Asian Cars
European Cars
Top Speed
Probability Density
0.03
0.025
0.02
0.015
0.01
0.005
0
60
80
100
120
140
160
180
200
Top Speed (MPH)


American Cars
Asian Cars
European Cars
Length
Probability Density
0.03
0.025
0.02
0.015
0.01
0.005
0
120
140
160
180
200
220
Length (inches)


American Cars
Asian Cars
European Cars
Length
Probability Density
0.03
0.025
0.02
0.015
0.01
0.005
0
120
140
160
180
200
220
Length (inches)

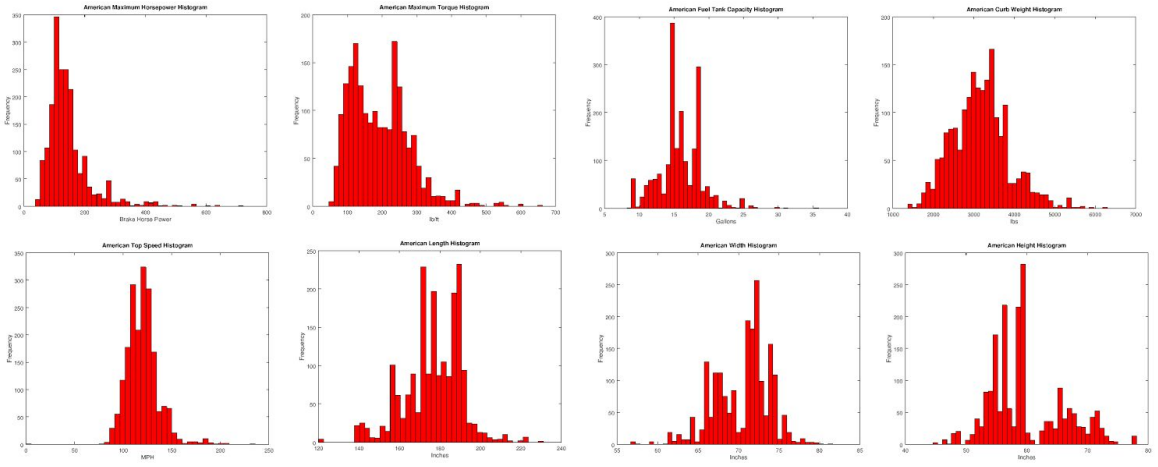

American Cars
Asian Cars
European Cars
Width
Probability Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
60
65
70
75
80
85
Width (inches)


American Cars
Asian Cars
European Cars
Height
Probability Density
0.08
0.06
0.04
0.02
0
40
50
60
70
80
Height (inches)

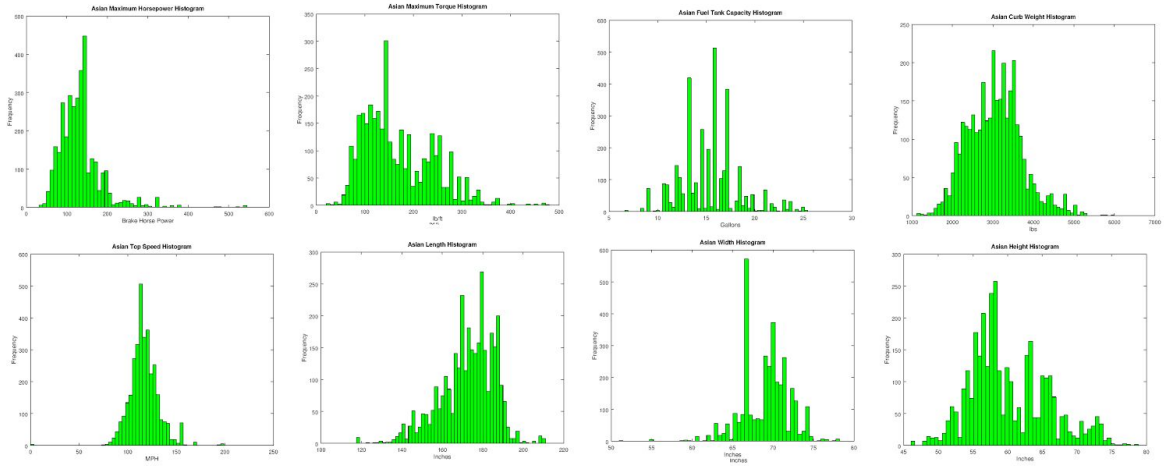

American Cars
Asian Cars
European Cars
Curb Weight
Probability Density
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
1000
2000
3000
4000
5000
6000
Curb Weight (lbs)

# Histogram

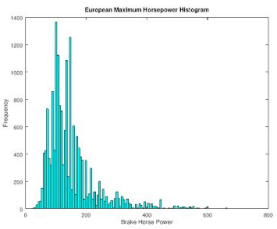
## American Feature Histograms

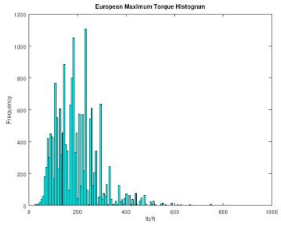


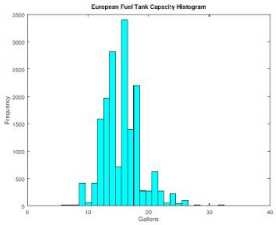

## Asian Feature Histograms

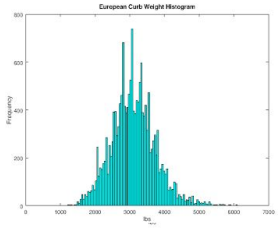


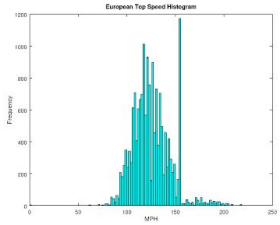
# Asian Feature Histograms

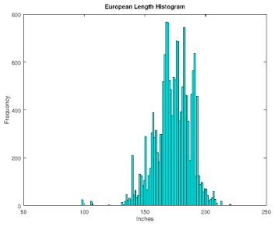


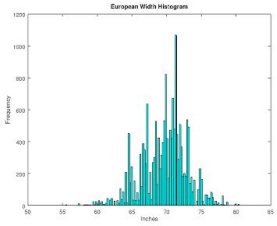



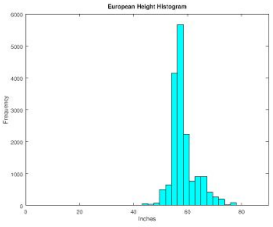



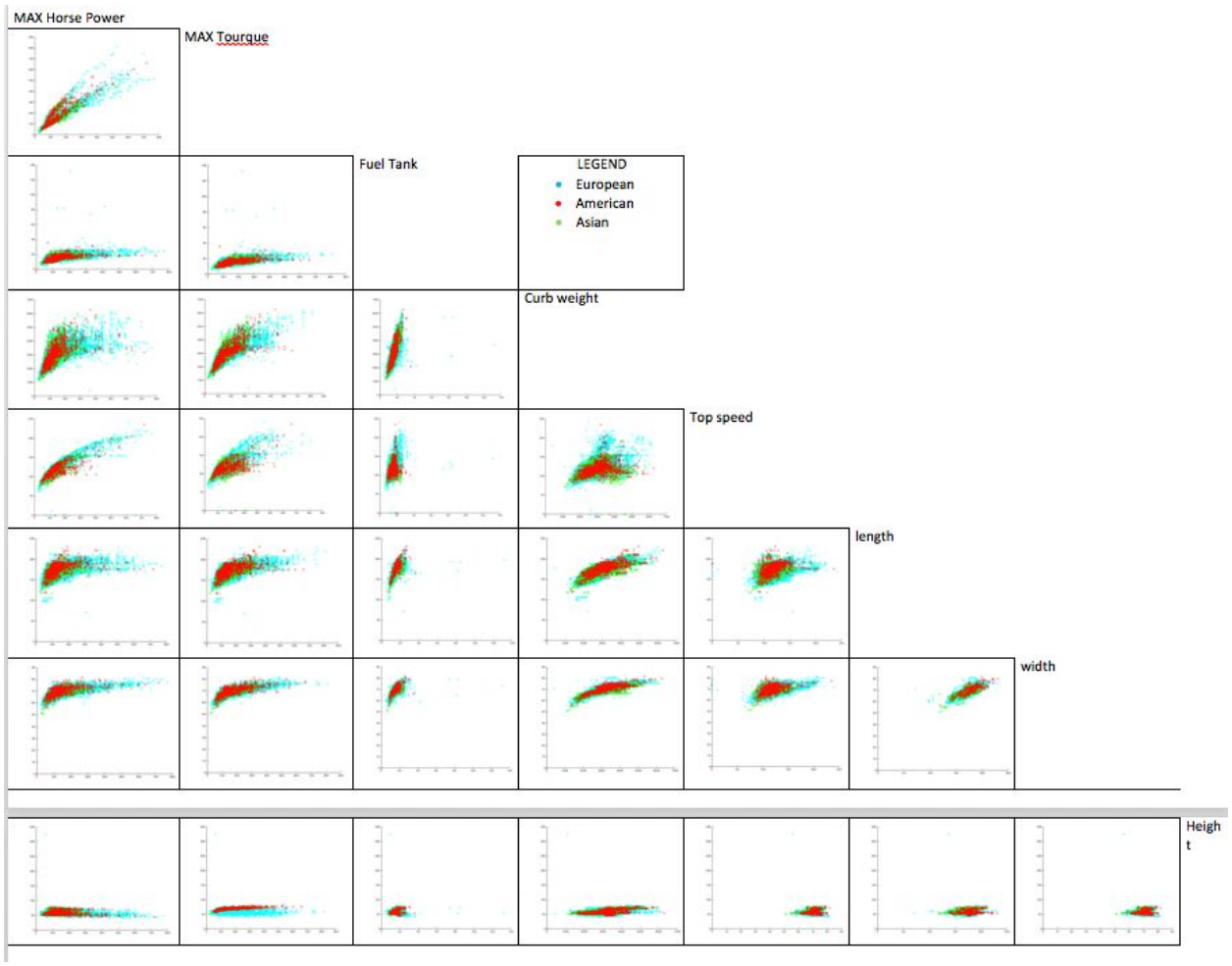



# Scatter Plots




# Scatter Plot Matrix Analysis

- Horsepower to Torque
    - Looking at this data, it is clear  that for any car in any region that horsepower has a strong correlation to torque of a vehicle.
- Max Horsepower to Fuel Tank Capacity
    - Looking at the data, at first glance we would think fuel tank has some correlation to max horsepower, however taking a deeper interpretation, if a fuel tank's weight, meaning if a fuel tank capacity is smaller, that should positively impact the abilities of a vehicle's max power or max horsepower allowing it to go faster because of the reduce in fuel weight.
- Max Horsepower to Curb Weight

    - To interpret this data, for our data we can see that there has been a medium correlation to max horsepower, if we think about fuel tank capacity, we can see that the data on the scatter plot for the max horsepower has stronger than the correlation between the fuel tank capacity, that is because curb weight is the total weight of the vehicle, whereas the fuel tank capacity is only accounting for the weight of the fuel tank.
- Horsepower to Top Speed
    - For this interpretation, our initial hypothesis was that horsepower does not affect a vehicle's ability of its top speed, but interpreting more of this data we can see that the data does tell us if our vehicle were to be heavier, without its correct max horsepower the vehicle would not reach its top speed. Therefore when looking at our scatter plot we can see a high correlation between horsepower and top speed, because in the event that a vehicle does have a high top speed, that does mean it has a higher top speed because more horsepower means a better ability to attain higher speeds.
- Horsepower and Length
    - Initially, our hypothesis was that there was no way that a vehicle's length could affect its horsepower or have correlation, looking at our data we can see our hypothesis is seldomly correct, and does prove our initial hypothesis, in order to understand the data, we need to understand that a longer vehicle, will in fact have more material to its mass, which is where a lot of a vehicle's weight is contributed too. Therefore we we can see the minor correlation in our scatter plot.
- Horsepower to Width
    - Rationally understanding a vehicle and how width our affect its horsepower, we do not see much correlation, the only rational reason we see this deviation from our initial idea of these two continuous variables being minimally correlated is, the aerodynamics of a vehicle, when a vehicle is traveling the width of its body can negatively affect its ability to perform(horsepower).
- Horsepower to Height
    - For our initial hypothesis we claimed the height of a vehicle will have a strong correlation to a vehicle's max horsepower, because the taller a vehicle is the less it is to its center of gravity, thus not allowing that vehicle to fully perform. However for this scatter plot data, we can see for all regions this to be untrue. There is no correlation between horsepower and height, to reflect on our initial though, we realize horsepower is a function of its ability to push the vehicle in a direction, the height shouldn't affect that ability, therefore we understand this data in the scatter plot.
- Torque to Fuel Tank Capacity
    - Just as we see the scatter plot between horsepower to fuel tank, these two variables show very similar behaviors, we believe mainly because of the size or capacity of a fuel tank, because torque is the measurement of how much force a vehicle will produce at initial throttle, as we can see for all regions in the event that a vehicle does possess a larger fuel capacity tank, there will be a correlation

between the two variables. Something to note, looking closely we can see that the correlation between fuel tank and torque are stronger with american cars than Asian or European. Mainly because american cars possess iron fuel tanks and typically have higher torque than asian and european cars.

- Torque and Curb Weight
    - For torque and curb weight, we can see a strong correlation between these two continuous variables, that is because the heavier a vehicle is the less torque it will be able to produce. It's important to note that the european vehicles have the most correlation between torque and its weight. This may be because european vehicles are not as heavy as asian and american vehicles, it's also important fact to note, most european vehicles do have an forced induction component , such as turbos, which give them better torque.
- Torque to Top Speed
    - Before analyzing this data it's important to note that torque is engaged during the initial burst of driving a vehicle, after initial throttle torque does not have much influence on the experience of driving. Therefore with this knowledge we can see that there isn't such a strong correlation between these two variables, mainly because torque only plays a role in top speed during the beginning of driving.
- Torque to Length
    - Looking at our data, just as max horsepower to length the correlation between the two are very similar, we can agree with this data, because just as the explanation of how max horsepower has a minor impact on length, torque also doesn't hold much correlation to length.
- Torque to width
    - Torque to width is just similar to max horsepower to width, the correlation between the two are not strong, a vehicle's width does not impact the amount of torque it will produce.
- Torque to height
    - Just as with width , and the data from max horsepower there is no correlation between a vehicle's ability to produce torque to its height, the minimal correlation we can see through the scatter plot data could be from initial aerodynamic force reducing the ability to perform fully is what we think.
- Fuel tank to curb weight
    - This correlation is very strong, just as the rationalization between horsepower to torque is, this is because fuel tank capacity is apart of the total curb weight of a vehicle. Therefore we can see the strong correlation, because curb weight is apart of the fuel tanks weight. For Example a lighter fuel tank will contribute to an overall lighter curb weight, we can especially see this rational for all three regions of car manufacturers.
- Fuel tank to top speed
    - We see a strong correlation here, why? Because looking at the fuel tank we can rationalize that with a larger fuel tank the vehicle will not have a higher top speed. This is because of the added extra weight to the vehicle which if we see the

correction from top speed to curb weight there is a strong correlation between them.

- Fuel Tank to length
    - There
- Fuel Tank Width
- Fuel Tank Height
    - Its important to note, that there is zero correlation between fuel tank and height that is mainly because, if we think aboutit the fuel tank height doesn't play any role in the height of a vheicales height
- Curb Weight to Top Speed
- Curb Weight Length
- Curb Weight Width
- Curb Weight Height
-

# Analysis

## Normality

Looking at the histograms, the data is in almost all cases normal. In a few cases, the maximum does not belong with the majority of the data, or the data is bivariate. In general though, we can assume normality. We were unable to plot a PDF curve with a dot diagram underneath due to the sheer number of data points for all the regions. Since we work out of Octave (Danny works with Matlab, and he did the scatter plots), the plotting is very slow with more than 1000 data points, and the European cars data set has over 16,000.

## Central Limit Theorem

To infer whether or not the central limit theorem applies to the data we have, we calculate a sample mean for each region, and then the population mean for each region.

For American cars, we have a population of 1921, so we took 500 as a statistically large sample. The sample mean is:

**147.62 191.89 15.735 3157.4 120.58 176.19**

And the population mean is:

**145.18 191.48 15.856 3165.2 119.82 176.76**

For Asian cars, we have a population of 3342, so we took 1000 as a statistically large sample. The sample mean is:

**133.21 167.66 15.319 3046.3 116.89 172.79**

And the population mean is:

**133.18 167.92 15.382 3054.7 116.83 172.73**

For European cars, we have a population of 16935, so we took 6000 as a statistically large sample. The sample mean is:

**149.4 192.08 15.655 3103.4 125.47 173.24**

And the population mean is:

**150.21 193.03 15.675 3103.4 125.5 173.23**

These all are normal according to the central limit theorem, since the variance between the sample mean for each is within a small finite range of the population mean.

## Independence

Independence cannot be assumed across all features analyzed. Looking at correlations between features using the 2D scatter plot summary, there appears to be strong correlation between some features. For example fuel tank capacity and curb weight seem to be strongly correlated as they are linearly fit. This is exemplified in Table 1 as the correlation between the fuel tank capacity and curb weight for American cars is over 0.8. This makes intuitive sense because the heavier a car weighs, the more energy it will need to move and therefore heavier cars will be fitted with larger fuel capacities. There are strong correlations between other features such as top speed and maximum horsepower. Principle component analysis will be used to limit the feature set by removing dependent features.

**Table 1: Correlation Matrix for American cars**

| | **maximum horsepower** | **maximum torque** | **fuel tank capacity** | **Curb weight** | **Top Speed** | **Length** | **Width** | **height** |
|---|---|---|---|---|---|---|---|---|
| **maximum** | 1.00000 | 0.81486 | 0.53188 | 0.58201 | 0.78550 | 0.44507 | 0.5094 | -0.0239 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **horsepower** | | | | | | | | |
| **maximum torque** | 0.81486 | 1.00000 | 0.63956 | 0.79090 | 0.60836 | 0.59011 | 0.70464 | 0.22724 |
| **fuel tank capacity** | 0.53188 | 0.63956 | 1.00000 | 0.80309 | 0.30196 | 0.77341 | 0.69491 | 0.35227 |
| **Curb weight** | 0.58201 | 0.79090 | 0.80309 | 1.00000 | 0.31155 | 0.75492 | 0.83292 | 0.60084 |
| **Top Speed** | 0.78550 | 0.60836 | 0.30196 | 0.31155 | 1.00000 | 0.39136 | 0.41443 | -0.2828 |
| **Length** | 0.44507 | 0.59011 | 0.77341 | 0.75492 | 0.39136 | 1.00000 | 0.76532 | 0.21576 |
| **Width** | 0.50941 | 0.70464 | 0.69491 | 0.83292 | 0.41443 | 0.76532 | 1.00000 | 0.44285 |
| **height** | -0.0239 | 0.22724 | 0.35227 | 0.60084 | -0.2828 | 0.21576 | 0.44285 | 1.00000 |

# Team Assessment

## Danny

I created the script to display all relations between regions and put the correlation scatter plot matrix together.

## Robert

I created the script to print the histogram summary plots. I also contributed to the analysis section, such as the independence and determination of plots sections.

## Kevin

I created the script to print the probability density plots. I contributed to the Analysis section.

# References

## Kevin's References


Dr. Erika Parsons
http://www.investopedia.com/terms/c/central_limit_theorem.asp
https://www.gnu.org/software/octave/doc/interpreter/Terminal-Output.html
https://www.mathworks.com/help/matlab/ref/csvread.html
https://www.ultimatespecs.com/
https://lists.gnu.org/archive/html/help-octave/2011-09/msg00179.html
http://octave.1599824.n4.nabble.com/remove-row-from-matrix-td2969149.html
https://www.gnu.org/software/octave/doc/interpreter/Manipulation-of-Plot-Windows.html
https://www.gnu.org/software/octave/doc/interpreter/Two_002dDimensional-Plots.html
https://en.wikipedia.org/wiki/Bhp
https://www.gnu.org/software/octave/doc/interpreter/Printing-and-Saving-Plots.html
https://www.mathworks.com/help/matlab/ref/genvarname.html
https://www.mathworks.com/examples/matlab/mw/graphics-ex99259345-plot-multiple-lines
https://www.mathworks.com/matlabcentral/answers/23591-examples-of-multiple-plots-on-one-graph
https://octave.sourceforge.io/octave/function/hold.html
http://octave.1599824.n4.nabble.com/Same-plot-for-two-functions-td1631271.html
https://www.gnu.org/software/octave/doc/v4.0.0/Variables.html
http://web.cecs.pdx.edu/~gerry/MATLAB/plotting/multipleCurves.html
https://www.gnu.org/software/octave/doc/interpreter/Distributions.html