DESAUTELS Faculty of Management
Faculté de gestion

# MGSC 662
# Multivariate Statistics

# Final Project
# Predicting Potentially Collectible Cars

## Table of Contents



Name: Avi Malhotra
Presented to: Prof. Juan Camilo Serpa

# Introduction

The automotive industry, renowned for its technological innovation and significant market influence, presents an intriguing domain for in-depth data analysis. This study aims to move beyond conventional aspects such as manufacturing and consumer preferences, focusing on a niche yet captivating objective: identifying potential collectible cars. Collectability in automobiles is a nuanced concept, often driven by unique attributes and historical significance that appeal to collectors and enthusiasts. By analyzing a comprehensive dataset encompassing a variety of car attributes - ranging from make and body style to engine specifications and pricing - this report endeavors to unveil patterns and insights that delineate a car's potential as a collectible item. Employing a blend of statistical methods and advanced machine learning techniques, the analysis is meticulously designed to predict which cars might hold special value for collectors, thereby serving as a valuable resource for stakeholders like collectors, automotive enthusiasts, and market analysts.

# Exploratory Data Analysis

The dataset under examination comprises a diverse array of characteristics pertaining to automobiles. These attributes cover various aspects, including but not limited to, physical dimensions, engine specifications, performance metrics, and economic factors. *Note: Please refer to the Appendix for supplementary graphs.*

## Structure and Content

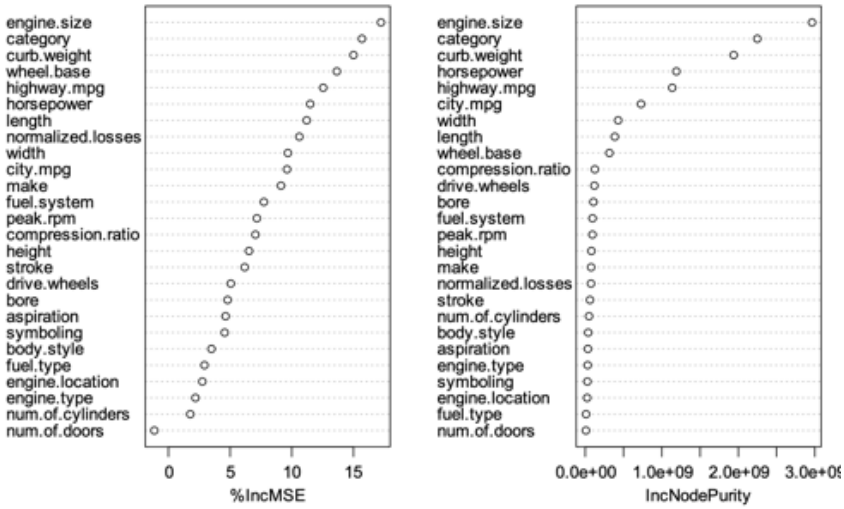
- **Rows and Columns**: The dataset includes a collection of records, each representing an individual car, and various columns depicting the attributes of these cars.
- **Attributes**: Key attributes in the dataset include 'make', 'body.style', 'num.of.doors', 'engine.size', 'horsepower', 'peak.rpm', 'price', and others. These variables are a mix of categorical (e.g., 'body.style', 'make') and numerical (e.g., 'engine.size', 'price') types.
- **Non standard formatting:** It was noted that the null values were demarcated with a '?', so they were converted to NA first.
- **Missing Values**:
  1) "normalized-losses": 41 missing
  2) "num-of-doors": 2 missing
  3) "bore": 4 missing
  4) "stroke" : 4 missing
  5) "horsepower": 2 missing
  6) "peak-rpm": 2 missing
  7) "price": 4 missing
- **Variable Types**: The variables are a mix of categorical and continuous data types. This diversity necessitates different strategies for data processing and analysis.

## Data Preprocessing

- ***'num.of.doors'***: Imputed missing values with *'four'*, assuming it as the most common category amongst the car manufacturer.
- ***'horsepower'***: Missing values imputed using the range of 110 to 115, based on the correlation with engine size (127 to 137) and other attributes like fuel type and aspiration.
- ***'peak.rpm'***: Imputed using a range between 4800 to 5500, determined by mode, considering the consistency of seven categorical attributes.
- ***'bore'***: For engine type *'rotor'* with NA values, imputed using a range of 3.3 to 3.65 based on mode, due to close similarity between mean and mode and the attribute's categorization.
- ***'stroke'***: Missing values imputed within the range of 3.1 to 3.4, using mode, considering the similarity between *'bore'* and *'stroke'* values.
- ***'normalized.losses'***: Given the lack of correlation with other numerical columns and varied categories in missing values, imputed using the median due to the distribution's resemblance to normal distribution.
- ***'price'***: Missing values replaced with the average price of cars with the same make, ensuring consistency in pricing based on make.
- ***'category'***: A derived attribute from price (*'luxury'* for cars more expensive than $30,000 and *'affordable'* otherwise). The threshold was derived from the outlier entries in the boxplot.

# Methodology

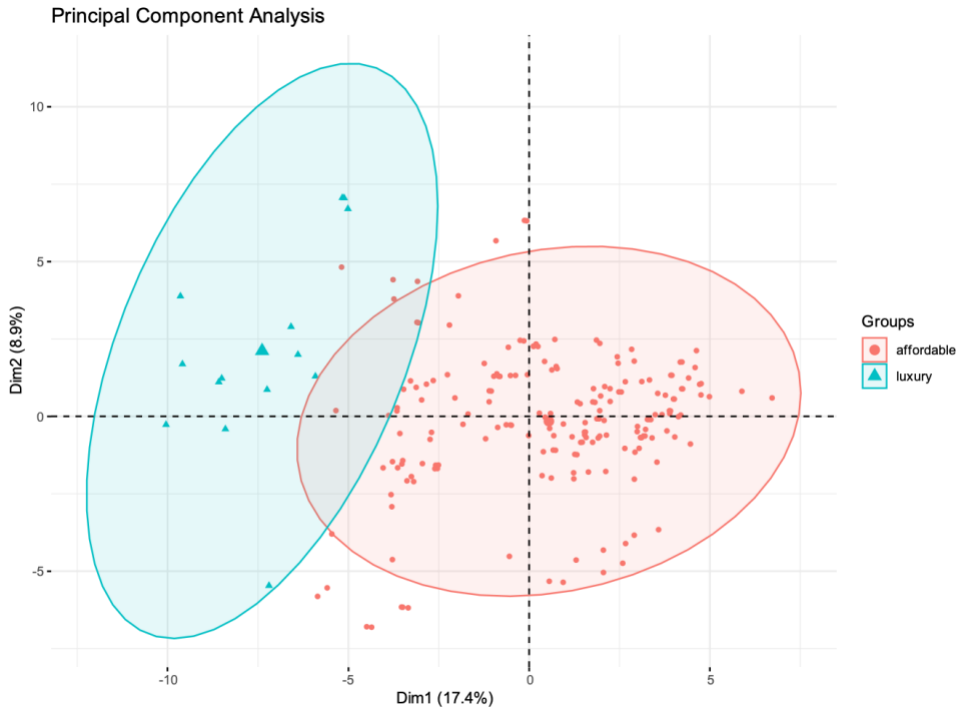
## Feature Extraction using Random Forest




*Feature importance list for random forest model*

The Random Forest model, with a commendable accuracy of 90%, was instrumental in identifying key predictors that influence a car's market value and potential collectability. It is worth noting that the underlying assumption was that prices correlate to collectability, hence the feature importance matrix above pertains to predicting price. Nonetheless, the model underscored *"engine.size"*, *"curb.weight"*, *"highway.mpg"*, *"horsepower"*, and *"categoryluxury"* as significant features. This informed selection of attributes provides a foundational understanding of the elements that contribute to a car's appeal and market positioning.

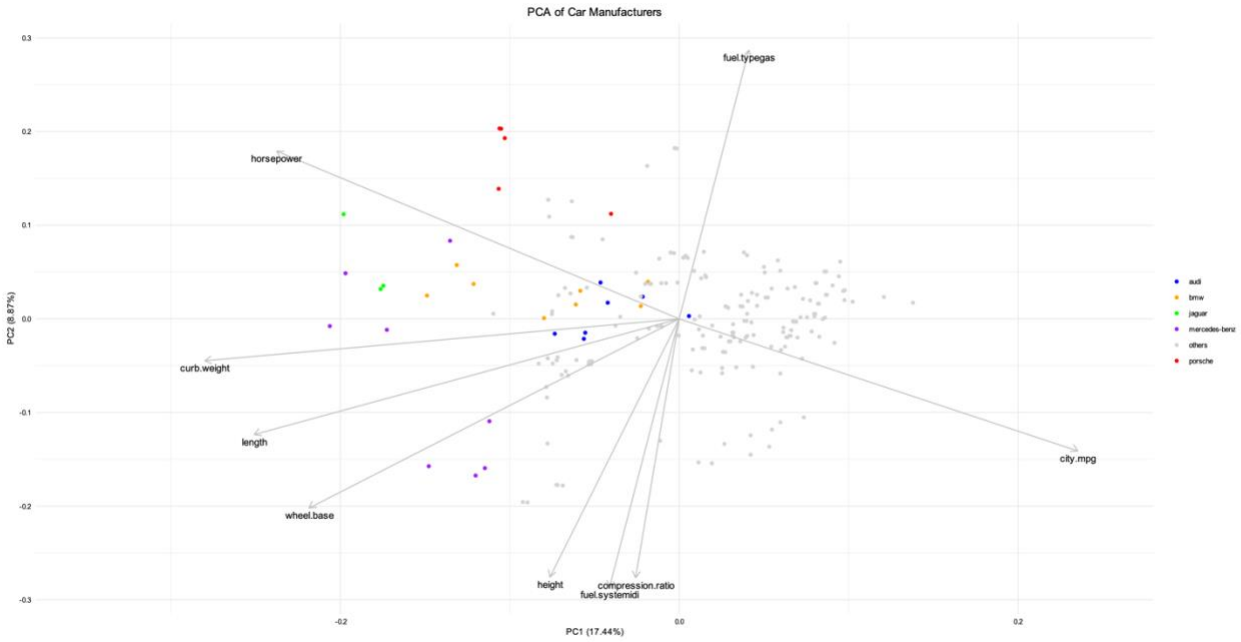
## Feature Extraction using PCA




*Generic PCA plot for 2 components*

The PCA plot above offers a compelling visualization of the variance within the dataset, showcasing the distribution of cars along the first two principal components. This graphical representation is particularly insightful for understanding the spread and concentration of data points. The first principal component accounts for 17.4% of the variance in the dataset, while the second component captures an additional 8.9%. The color-coded points, enhanced by the ellipses, distinctly demarcate the *'affordable'* from the *'luxury'* categories, underscoring their unique feature sets. The clear patterns observed along the principal components highlight the defining attributes that distinguish affordable cars from their luxury counterparts.

The secondary PCA plot, illustrated below, further dissected the luxury segment, revealing key attributes of high-end brands appealing to collectors. Porsche was characterized by high horsepower, while Mercedes-Benz, Jaguar, and BMW were identified by a combination of high curb weight and horsepower, aligning with their luxury status. Audi, with lower scores on these attributes, stood apart, potentially affecting its collectability. Notably, luxury brands were also associated with lower city mpg, indicating a possible preference for performance over fuel efficiency. As visible, only a subset of the features is plotted. A threshold of 0.27 was established in order to extract the top-10 features contributing to the 2 principal components. Such a filtration automatically removed high collinearities (such as between *'curb.weight'* and *'engine.size'*). Nonetheless, there is still the exception of 'compression ratio' due to its numerical nature and high collinearity with *'fuel.systemsidi'*. The former was thus removed from the list of features used for k-means. This step was crucial in refining the set of predictors that truly signify a car's desirability as a collectible.



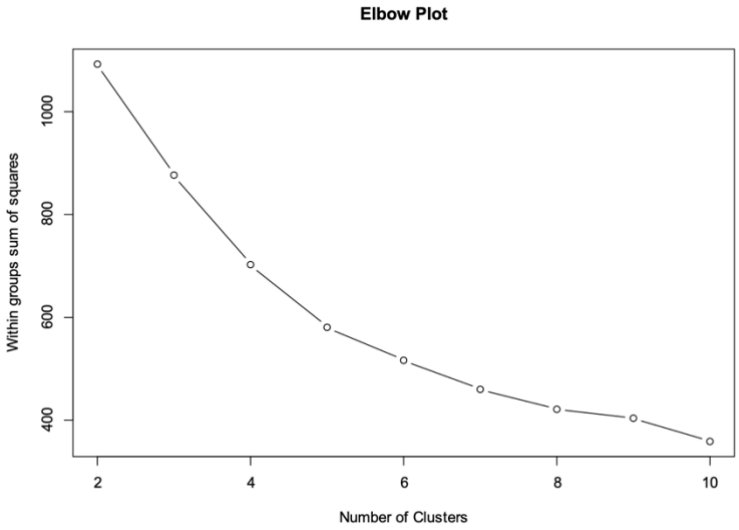

*Useful features from PCA*

Thus ,from PCA, additional predictors such as *"fuel.typegas"*, *"fuel.systemidi"*, *"wheel.base"*, *"length"*, *"height"*, *"curb.weight"*, *"horsepower"*, and *"city.mpg"* were identified.

## Final Model: K-Means Clustering

K-Means clustering was employed using a curated list of predictors derived from both Random Forest and PCA, allowing for a targeted clustering analysis. The predictors were:

- *"engine.size"*,
- *"curb.weight"*,
- *"highway.mpg"*,
- *"horsepower"*,
- *"categoryluxury"*,
- *"fuel.typegas"*,
- *"fuel.systemidi"*,
- *"wheel.base"*,
- *"length"*,
- *"height"*,
- *"curb.weight"*,
- *"horsepower"*,
- *"city.mpg"*



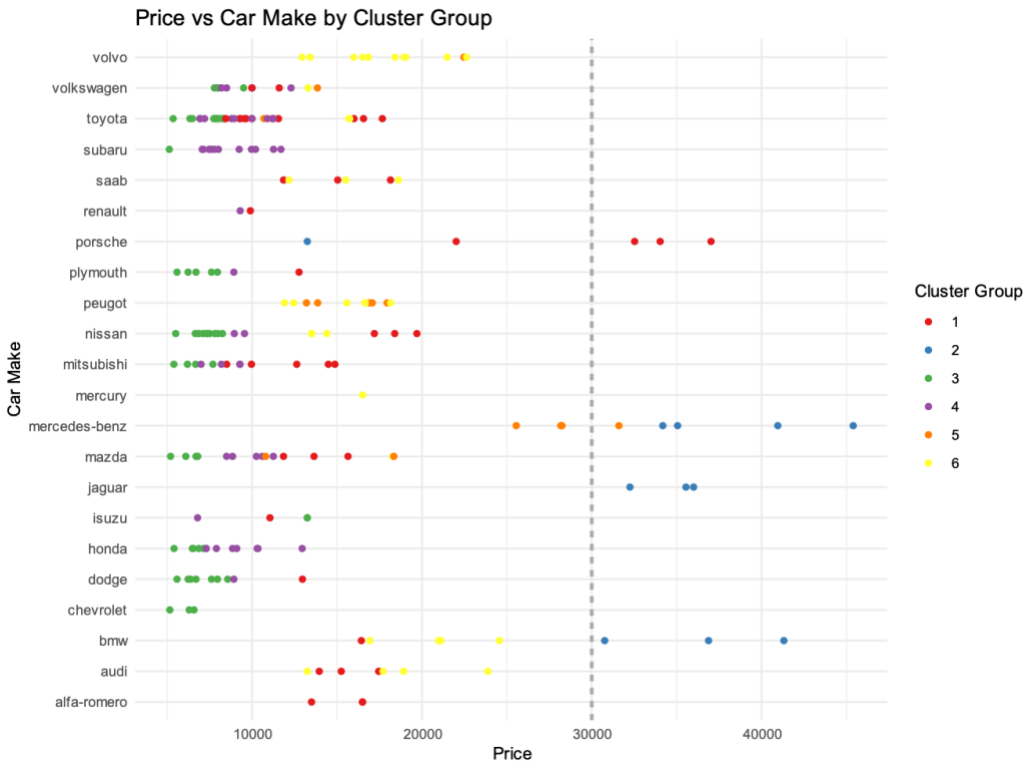

The optimal number of clusters was ascertained using the elbow method, pinpointing the most coherent grouping structure without overcomplicating the model. While it is not clearly visible via the elbow plot, the number of clusters was capped at 6 due to diminishing returns in the sum of squares (which indicates within-cluster cohesion) .This balance between detail and simplicity was key in revealing the natural groupings within the data, providing actionable insights for stakeholders.

The amalgamation of these methods highlights a strategic and insightful approach to understanding the complex landscape of automotive features in relation to collectability.

## Results & Insights

### Identifying Collectible Cars

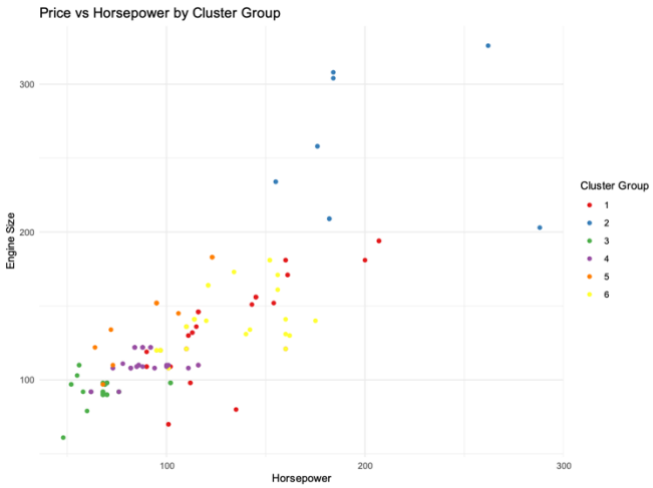
Cluster 2 has emerged as the segment with vehicles that could be deemed worthy of collection. This cluster predominantly features the most expensive cars on average, with a significant representation of luxury brands (10/14 in total). Interestingly, the classification within this unsupervised model does not solely align with price tags; while luxury is often synonymous with a higher price point, the model has not clustered all high-priced cars together. Notably, Porsche, a brand with substantial cachet, has a surprisingly limited presence in this collectible category, with just one model deemed collectible. Meanwhile, all Jaguar models, most from Mercedes-Benz, and a select few from BMW, were predictably categorized within Cluster 2, aligning with conventional wisdom regarding luxury collectibles.



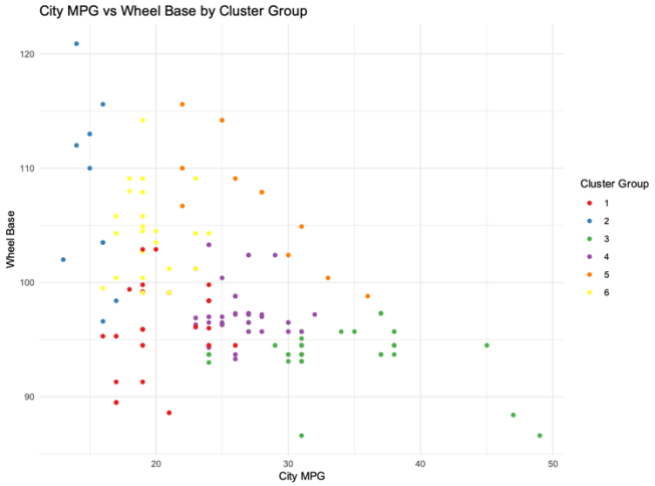

*K-means model result*

## Cluster Insights

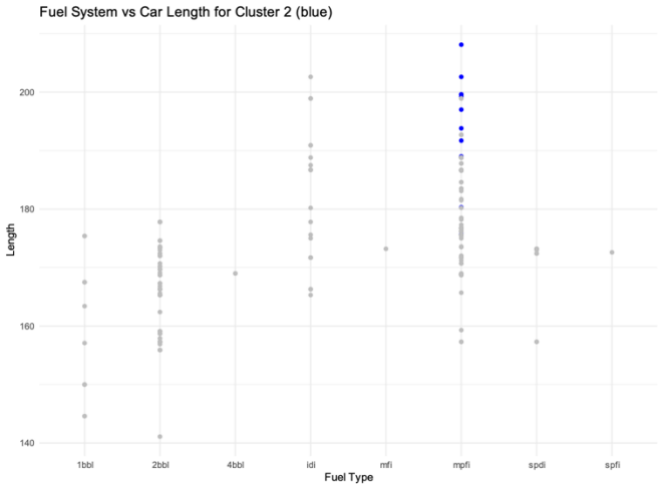
The following 3 info-graphs highlight the unique characteristics of collectible cars (i.e. cars in cluster-2).




**Power Attributes:** Vehicles in Cluster 2 are distinguished by their superior horsepower and engine size, traits that are quintessential for performance-oriented collectible cars. This suggests a pattern where power and performance are key determinants for potential collectability.




**Size and Efficiency Trade-Off:** The cars within Cluster 2 also display a trend towards larger wheelbases and lower city MPG. This observation could be indicative of a design philosophy that favors size and presence over fuel efficiency, a characteristic often found in luxury vehicles that are designed to make a statement rather than serve as utilitarian transports.




**Design and Fuel System:** The elongated length and the preference for 'mpfi' fuel systems within Cluster 2 further underscore the luxurious nature of these vehicles. This combination of design elements and engineering choices paints a picture of cars that are not just meant to be driven but to be experienced and admired.

## Business Insights

From a business perspective, Cluster 2's characteristics provide a blueprint for what attributes contribute to a car's status as a potential collector's item. Automakers can leverage this information to guide the development of future models that align with these collectible traits. For collectors and investors, Cluster 2 offers a predictive lens through which to identify cars that are likely to appreciate in value or become sought-after classics.

Additionally, the insights from this cluster analysis could inform marketing strategies that target demographics likely to value the specific attributes associated with these collectible models. For automotive historians and archivists, the data presents an opportunity to track the evolution of luxury cars and predict which current models might be revered by future generations.

In sum, the clustering not only sheds light on current market dynamics but also suggests a trajectory for the future of car collecting, offering a valuable resource for various stakeholders in the automotive ecosystem.

## Conclusion

This report marks a personal milestone in my journey as a data scientist, where my professional pursuit intersects with my lifelong enthusiasm for cars. It's been a venture into the heart of the automotive industry, not just as a market of commodities but as a canvas of collectible artistry. Through the comprehensive analysis of car attributes, I've sought to uncover the subtle patterns and distinct characteristics that signal a car's potential to transcend from a mere vehicle to a collector's gem.

Navigating through the intricacies of data, from initial exploration to in-depth modeling, has been both challenging and rewarding. It required a balance of technical expertise and a nuanced understanding of what makes a car stand out, especially since the concept of what makes a car admirable is subjective. The process has been a testament to the potential of data science to unravel complex questions and provide insights that resonate with both the heart and the mind.

Cluster 2 emerged from our models as a narrative of power, luxury, and desirability, capturing the essence of what might catch a collector's eye. Yet, in the pursuit of identifying collectibles, the data told a story that was broader than the sum of its parts. It wasn't just about the cars that fell into this cluster, but about the ones that didn't. This reflects a reality that the value of collectability is as diverse and multifaceted as the people who cherish these vehicles.

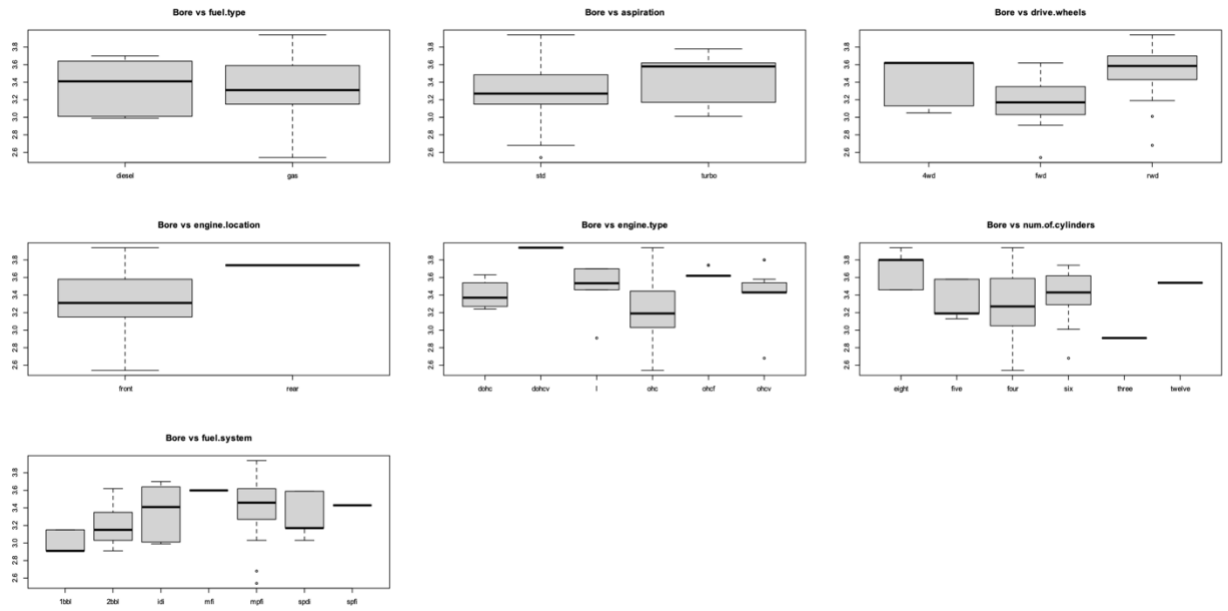
## Citations

Information on the visualization packages used. Click *here*.

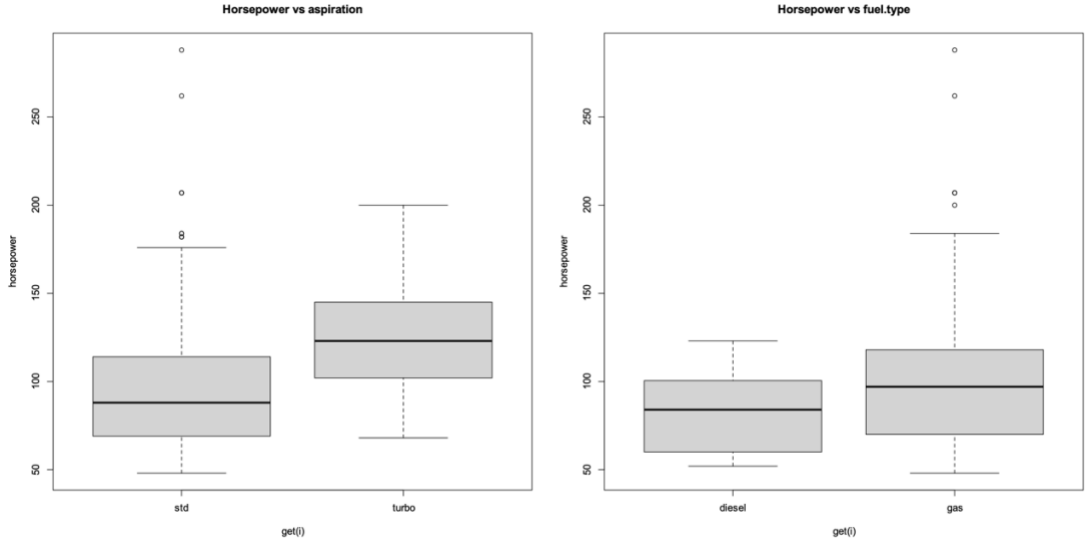
# Appendix

Plots for Exploratory Data Analysis

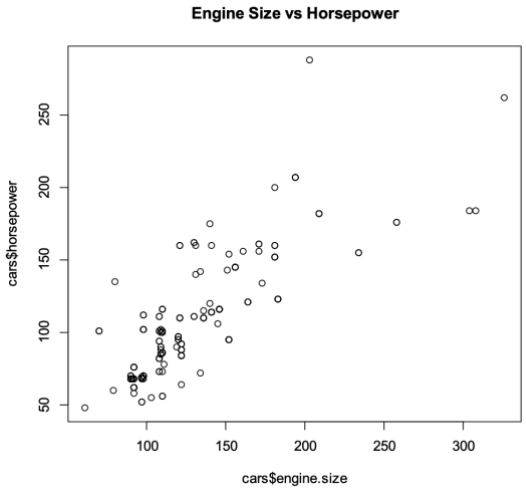
---

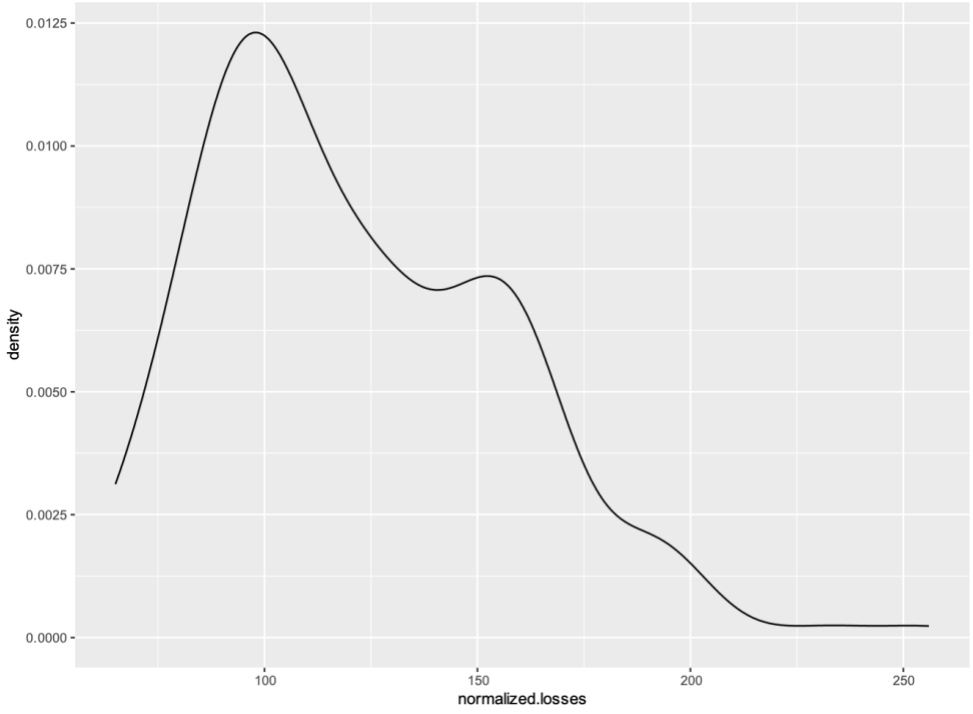


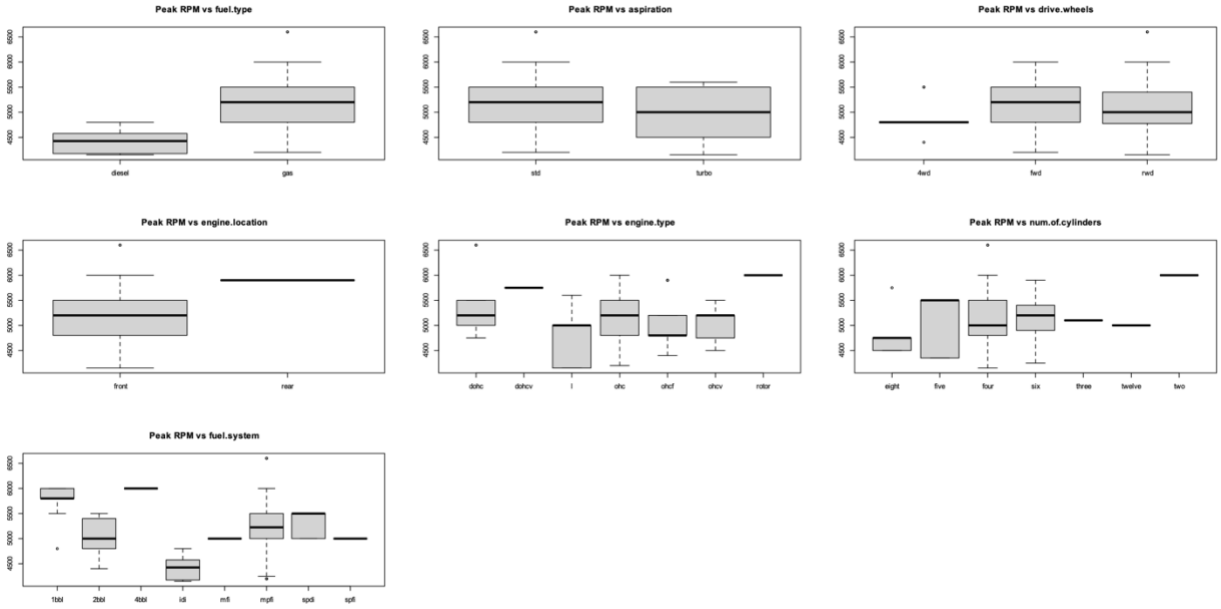

---

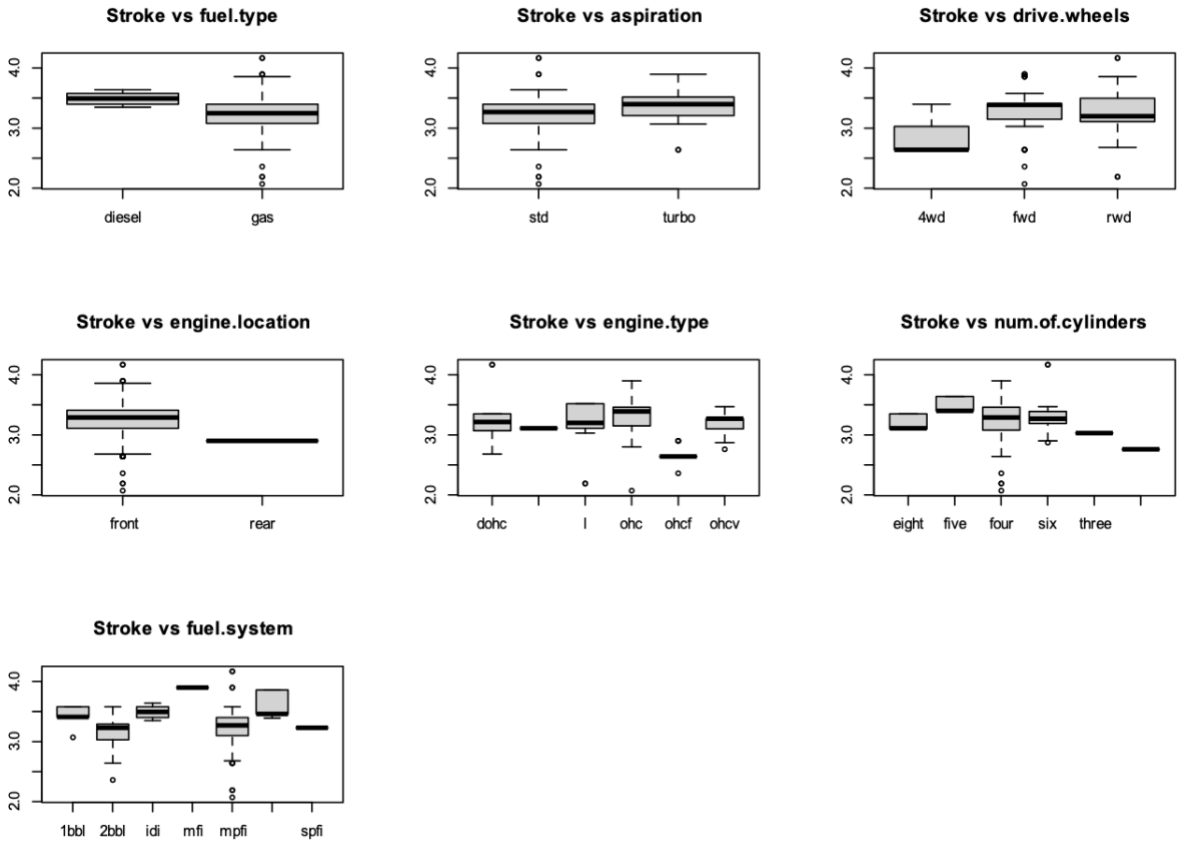


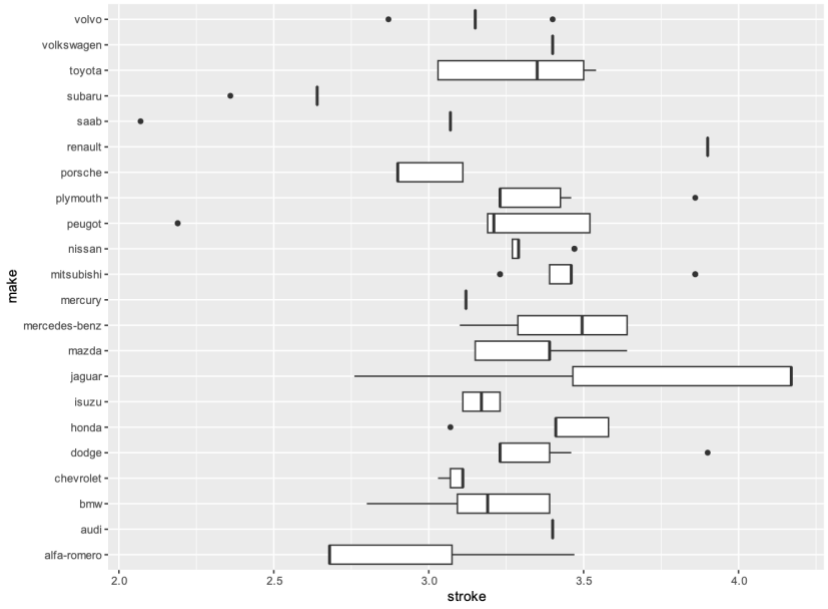

Price Boxplot
10000
20000
30000
40000

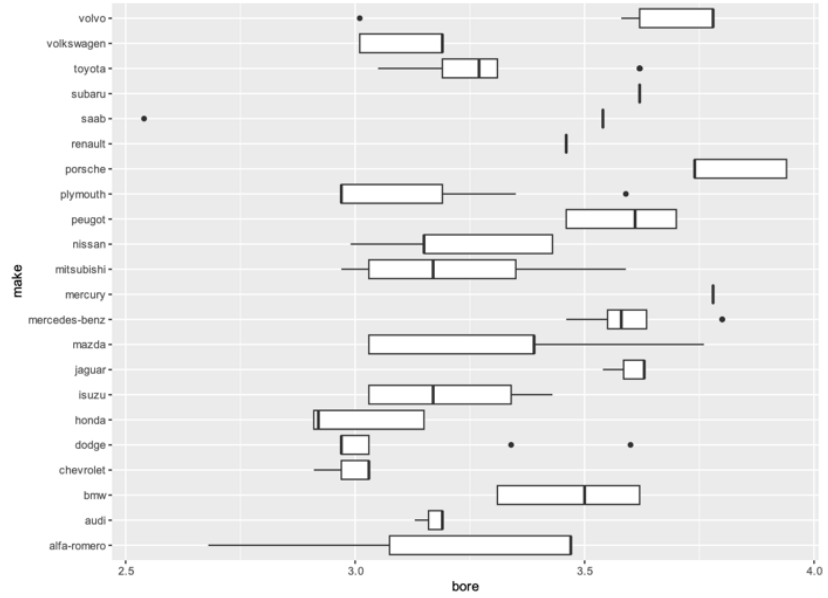

Engine Size vs Horsepower
cars$horsepower
50
100
150
200
250
100
150
200
250
300
cars$engine.size


density
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
100
150
200
250
normalized.losses


Peak RPM vs fuel.type
diesel
gas
Peak RPM vs aspiration
std
turbo
Peak RPM vs drive.wheels
4wd
fwd
rwd
Peak RPM vs engine.location
front
rear
Peak RPM vs engine.type
dohc
dohcv
l
ohc
ohcf
ohcv
rotor
Peak RPM vs num.of.cylinders
eight
five
four
six
three
twelve
two
Peak RPM vs fuel.system
1bbl
2bbl
4bbl
idi
mfi
mpfi
spdi
spfi


Stroke vs fuel.type
diesel
gas
Stroke vs aspiration
std
turbo
Stroke vs drive.wheels
4wd
fwd
rwd
Stroke vs engine.location
front
rear
Stroke vs engine.type
dohc
l
ohc
ohcf
ohcv
Stroke vs num.of.cylinders
eight
five
four
six
three
Stroke vs fuel.system
1bbl
2bbl
idi
mfi
mpfi
spfi


volvo
volkswagen
toyota
subaru
saab
renault
porsche
plymouth
peugot
nissan
mitsubishi
mercury
mercedes-benz
mazda
jaguar
isuzu
honda
dodge
chevrolet
bmw
audi
alfa-romero
make
2.0
2.5
3.0
3.5
4.0
stroke


volvo
volkswagen
toyota
subaru
saab
renault
porsche
plymouth
peugot
nissan
mitsubishi
mercury
mercedes-benz
mazda
jaguar
isuzu
honda
dodge
chevrolet
bmw
audi
alfa-romero
make
2.5
3.0
3.5
4.0
bore