Contributors: Jacob Spano, Joseph Cicero, Leo Morifuji, Christain Cuevas, Shane Fitzpatrick

Car Specs: How are they Important to the Consumer?

Introduction

Recently, cars have risen in price and demand significantly. The market is ever-changing, and many are misled by overzealous dealerships looking to sell the most expensive, top-spec vehicle without taking into account the personal needs of the consumer. The average consumer cannot afford to purchase a car outright and often must take out a loan to be paid off over the course of several years. Beyond just cost, there has been a higher demand for cars due to their undeniable convenience and the fact that many cities are nearly impossible to traverse by foot (Benfield 2014). Many rely on their vehicles for everyday life, which is why consumers need to be well-informed before their purchase. After all, the average person will own their car for 8-10 years (Chase 2024).

Background Information

The search for finding improved fuel efficiency is a key factor that drives the research and development of cars/car parts in respect to consumer needs. Most rational consumers seek to purchase as fuel efficient a vehicle as possible, while still fulfilling their other needs/wants of the vehicle. This is, in large part, due to the volatility of fuel prices over time; particularly, as of events in recent years, such as the COVID-19 pandemic in 2020. Fuel prices can change very rapidly and unpredictably due to factors such as domestic and international political changes, economic reasons, and issues in the production or delivery of fuel. This reasoning highlights why the research we are conducting is important and functional in today's world. However, we are not the first group to research these topics. Many others have conducted research under the same basis as our group. Some believe that alternative fuel options may be the future of our

transportation and can thus solve our fuel efficiency problem. Yet, a group of researchers found that with all the up-and-coming alternative options, “99.8% of all the world's transport is powered by Internal Combustion Engines or ICE’s” with more than 94% of transport energy coming from liquid fuels made from petroleum. They found that Alternative fuel options such as Electric, Hydrogen based engines, and Biofuel engines lack the significant manufacturing and production to overtake the prominence and overall convenience of ICE’s and petroleum based fuel.

As well as alternative fuel options, researchers have also explored other trains of thought such as engine specifications and engine performance. One group of researchers conducted a study and found that on average for cars of unmodified stock, that the number of Horsepower that an engine can output has a negative relationship with the miles per gallon that a car can travel. This means that they found that for every increase in horsepower for an engine, the miles per gallon tends to decrease and becomes worse. Increased horsepower is common in larger vehicles such as Semi’s or trucks, or any car that tends to often pull loads and muscle cars.

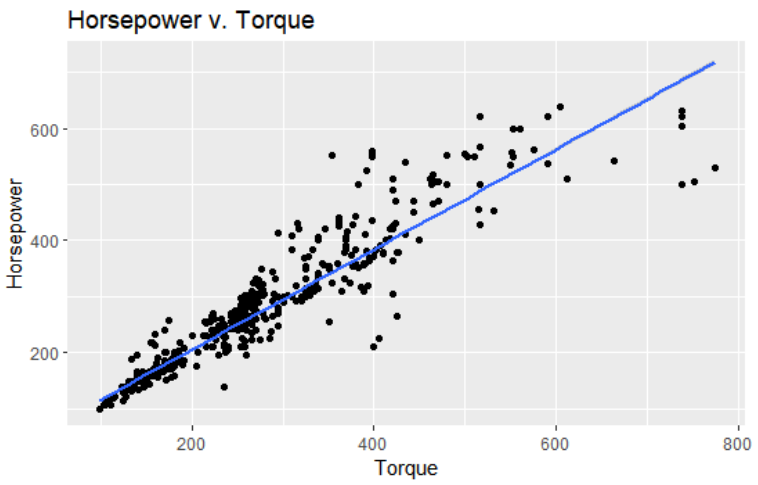
Lastly, one group of researchers directly measured the miles per gallon of cars related to the year of model and car. In their study that contained over 50 years of data, the trend they found was that there is a positive correlation between year of the car and the miles per gallon. This means that as the year of the car became newer, the more miles it could travel per gallon. This is largely due to innovation and the advancement of technology over the years. It also shows how fuel efficiency improvement is a never ending variable when it comes to consumer needs.

Description of Data and Methods

This data manipulation project seeks to find the best outcome for the consumer that fits their specific needs, whether longevity, power, or efficiency, by using car specifications. The cars in this data set vary significantly, with a range of fuel types, car brands, model years, dimensions, torque, horsepower, number of gears, miles per gallon, and transmission types (manual/automatic). The data was collected and analyzed in RStudio to perform a variety of tests and graphs to promote transparency within the automotive industry.

An effective way to find the best car for the consumer is to compare its horsepower to the torque it produces. Horsepower is a measurement of energy expenditure over time in this instance, with each horsepower being equivalent to 746 watts, and torque is a force acting at a certain distance. The level of horsepower in a vehicle essentially tells the consumer how strong it is, which is important to know if they are looking for a powerful car.

When comparing torque and horsepower, the two seem to either increase or decrease with a direct relationship. In this model, horsepower seems to be capped at around 650. This could mean a variety of different things, but it is most likely that the lack of technology from the time these cars were produced explains the upper bound of horsepower. Another possible explanation is that these companies had to sacrifice higher horsepower for higher quality in other areas.



Comparing different variables and their correlation to one another can reveal a great amount of information about the variables themselves. This was accomplished in this experiment with an ANOVA. An ANOVA is

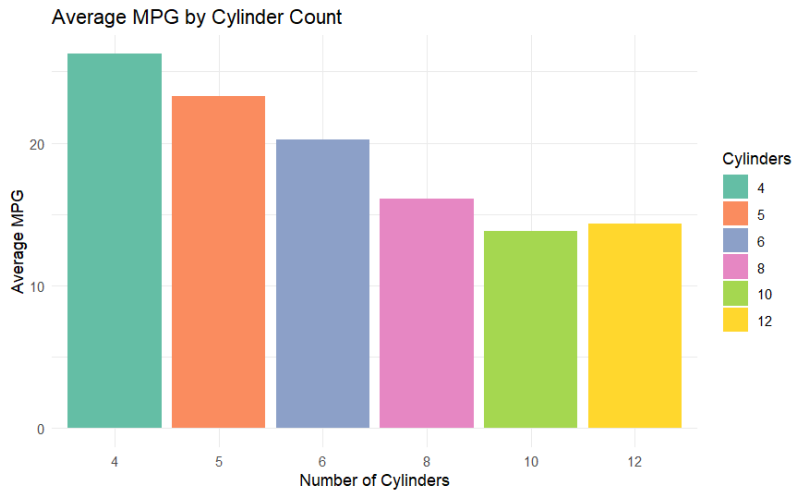
```
Analysis of Variance Table

Response: Car_Data$Identification.Year
                                                  Df  Sum Sq Mean Sq F value          Pr(>F)
Car_Data$Volume                                    1    0.75  0.7451  1.2406          0.2654
Car_Data$Fuel.Information.City.mpg                 1   26.89 26.8880 44.7685 0.0000000002457 ***
Car_Data$Fuel.Information.Highway.mpg              1   12.03 12.0325 20.0340 0.0000077426761 ***
Car_Data$Engine.Information.Engine.Statistics.Horsepower   1   25.41 25.4089 42.3059 0.0000000008557 ***
Car_Data$Engine.Information.Engine.Statistics.Torque       1    0.92  0.9158  1.5249          0.2169
Residuals                                       5070 3045.05  0.6006
---
```

essentially an analysis of different variables and their correlation to another variable. In this experiment, the variables were the volume of the vehicle, miles per gallon in the city and on the highway, horsepower, and torque. These variables were compared to the year of creation of the vehicle.

When examining the correlation between the variables, it was found that miles per gallon was a great way to predict the year of creation of the vehicle, along with horsepower. Surprisingly, torque and volume were not statistically significant. The reason why volume was not statistically significant is most likely due to the fact that cars have always been made in different shapes and sizes and have not tended to change over time. Why torque, however, is not statistically significant is not as simple to reason, but car companies may want to create cars with different abilities to do work, which is a product of force and distance, to satisfy the different needs of a variety of consumers.

The next variables we were interested in using to create a visualization to perform an analysis was the number of cylinders vs the average MPG. First, we needed to isolate the number of cylinders and the MPG into their own columns, this would make plotting much more simple. The first obstacle we ran into was isolating the number of cylinders since it was part of a long string within the "Engine.Information.Engine.Type" column, but luckily each entry was formatted the same. So, using the R library called "stringr" which comes with the "string_extract" function, we were able to look for the number that came before the word "cylinders" in each entry of the column, then create a new column that was numeric and acted as a categorical variable to later be used in our plot. For the dependent variable, we needed to create a new column that was composed of "average MPG" which was just simply created by adding the "Fuel.Information.City.mpg" and "Fuel.Information.Highway.mpg" columns together, then

dividing by two. We added these values into their own column we named “AvgMPG”. Now that we had the two columns we needed to perform the analysis we desired, we created a plot of the number of cylinders in a vehicle vs the average miles per gallon that vehicle would get. This



visualization immediately made it clear that there was a strong trend between the variables. According to this visualization, we can confidently say that as the number of cylinders increases, the fuel efficiency decreases. This conclusion is actually something we expected since according to other articles such as the one written by Lingampally Shalini, it has been seen before that cars that have more cylinders tend to consume more fuel. Looking at our graph you may note this trend cannot be seen for the 10 cylinder verse 12 cylinder values, the reasoning for this is likely because cars that have that many cylinders tend to be very high performance whether it be sports related or work related, such as a supercar or a semi truck, so it is more likely we didn’t have enough data points

```
car_data <- read.csv("cars.csv")

# Clean and extract the number of cylinders
car_data <- car_data %>%
  mutate(
    Cylinders = str_extract(
      `Engine.Information.Engine.Type`,
      "\\d+\\s*([cC]ylinder|[cC]ylinders)"
    ) %>%
      str_extract("\\d+") %>%  # Extract the number only
      as.numeric()
  )

# Create a new column for the average MPG (city and highway)
car_data <- car_data %>%
  mutate(
    AvgMPG = (`Fuel.Information.City.mpg` + `Fuel.Information.Highway.mpg`) / 2
  )

# Filter out rows with missing data for cylinders or AvgMPG
car_data_clean <- car_data %>%
  filter(!is.na(Cylinders) & !is.na(AvgMPG))

# Create the bar plot
ggplot(car_data_clean, aes(x = factor(Cylinders), y = AvgMPG, fill = factor(Cylinders))) +
  geom_bar(stat = "summary", fun = "mean", position = position_dodge()) +
  labs(
    title = "Average MPG by Cylinder Count",
    x = "Number of Cylinders",
    y = "Average MPG"
  ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2", name = "Cylinders")
```

on these two engine types to develop accurate averages. This visualization was simple yet very compelling in the story it told, it allowed us to make the connection between cylinder and fuel efficiency which will be extremely useful and relevant in proving the rest of our argument. This is because we also used cylinder count in our analysis of “cylinders vs horsepower” then used horsepower to determine its connection with torque. Ultimately, this provided the connection between performance and efficiency we needed to effectively convey to the reader which choice would be the best balance in fuel efficiency without sacrificing performance.

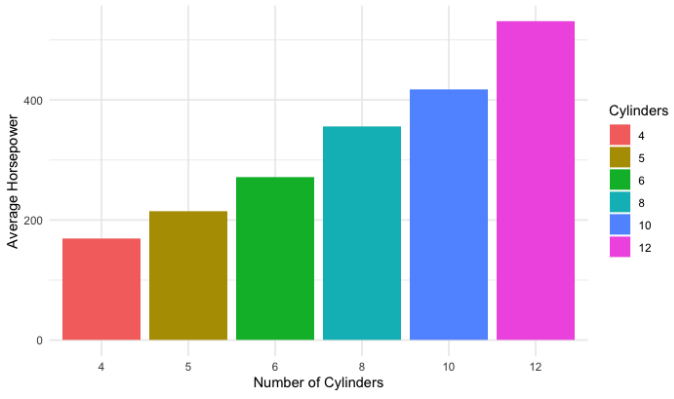
Our final graph takes a look at the relationship between the average horsepower of a vehicle based on its cylinder count. Similarly to Christian’s findings, the largest hurdle in creating this histogram was separating the cylinder count from the "Engine Information.Engine Type" variable. This was accomplished by using the “stringr" package to isolate strings that contained a numerical value followed by “cylinders.” The inputs vary slightly in the dataset; for instance, some values read as “cylinder,” while others are “cylinders.” It was crucial to ensure that the function accounted for this, and also for the function to be case-insensitive since there was variance there as well. A user by the name of “Jim” on the R-Bloggers forum was particularly helpful in demonstrating how to use the package properly, ensuring our visuals/conclusion were accurate and fully representative of the datasets’ information (Jim 2022).

The histogram, created using the ggplot2 package, presents a linearly positive relationship between the number of cylinders a car has and its horsepower. This means that as the cylinder count increases, the average horsepower also increases. Such is to be expected– power output in vehicles with combustion engines is strongly influenced by cylinder count. Within each cylinder, little “explosions” (precise mixtures of air/fuel) occur, which then drive the piston of the vehicle, creating energy, and ultimately powering it (BGAuto 2024). Therefore, generally

speaking, it is logical that more cylinders = more power. However, it is important to note that this is not always the case. Outliers in this data exist mostly due to differences in engine design; for instance, vehicles with forced induction (e.g. turbocharged models, as represented in the dataset) are capable of generating more power on smaller engines. This means that a 6-cylinder isn't always going to have more horsepower than a 5-cylinder, for example. Something that this dataset didn't account for but that would have been interesting to analyze is the average maintenance cost per year of these vehicles. You would likely see vehicles with higher cylinder/horsepower counts spending far more than those with lower values. This is due to the fact that the more complex a vehicle's engine is designed, the more "picky" it generally becomes. High performance vehicles, particularly when being driven as intended, are under extremely high stress as compared to vehicles with economy builds. Thus, by and large they must undergo more frequent–and expensive–servicing intervals.

Appendix 4: Visualization + Code

Figure 4: Average Horsepower by Cylinder Count



```r


#Extracting only the number of cylinders from the engine type column
cars2 <- cars2 %>%
  mutate(Cylinders = str_extract(`Engine.Information.Engine.Type`, "\\d+\\s*[Cc]ylinders?")) %>%
  #Get rid of spaces + convert cylinders to numeric
  mutate(Cylinders = as.numeric(str_extract(Cylinders, "\\d+")))

#Filtering + calculating average horsepower by cylinder count
cars2_clean <- cars2 %>%
  filter(!is.na(Cylinders) & !is.na(`Engine.Information.Engine.Statistics.Horsepower`)) %>%
  group_by(Cylinders) %>%
  summarise(Avg_Horsepower = mean(`Engine.Information.Engine.Statistics.Horsepower`, na.rm = TRUE))

#Plotting the average horsepower by cylinder count
ggplot(cars2_clean, aes(x = as.factor(Cylinders), y = Avg_Horsepower, fill = as.factor(Cylinders))) +
  geom_bar(stat = "identity") +
  labs(title = "Figure 4: Average Horsepower by Cylinder Count",
       x = "Number of Cylinders",
       y = "Average Horsepower") +
  scale_fill_discrete(name = "Cylinders") +
  theme_minimal()

```

Conclusion

From our research, we can conclude that the “sweet spot” for most consumers in regard to cylinder counts is likely going to be in the 4-6 range. Vehicles in this range generally have between 170-270 horsepower, a reasonable amount for your average consumer. In this context we define “average consumer” as someone needing a grocery-getter and a way to get to work/university. In a mostly flat area, a 4 cylinder engine will be more than sufficient while allowing the owner to save on fuel and maintenance costs. Individuals in more hilly areas, who perhaps do more highway driving, hauling, etc, likely will want to invest in higher horsepower vehicles. According to the ANOVA, horsepower and MPG were essentially directly correlated with year created, which implies that consumers should be looking for a car from around 2009 to get a car with lower horsepower and therefore fuel consumption. The direct relationship between torque and horsepower implies that torque will be lower as well as horsepower. This is primarily for fuel efficiency and those looking for a more powerful car should look towards the newer ones.

Works Cited

Bart, Austin Cory. "Cars CSV File." *CORGIS Datasets Project*, November 3 2015, corgis-edu.github.io/corgis/csv/cars/.

Ben Boland. *The Relationship Between Horsepower & Fuel Efficiency*. https://scholars.fhsu.edu/cgi/viewcontent.cgi?article=1427&context=sacad.

Benfield, Kaid. "The Daunting Challenge of Unwalkable America." *Smart Cities Dive*, 2014, www.smartcitiesdive.com/ex/sustainablecitiescollective/daunting-challenge-unwalkable-america-excerpted-people-habitat/333351/.

BG. "Do More Cylinders Always Mean More Engine Power?: BG Automotive." *Fort Collins, Longmont, and Loveland Auto Repair - BG Automotive*, 19 Nov. 2024, www.bgautomotiveinc.com/blog/do-more-cylinders-always-mean-more-engine-power#:~:text=Engines%20with%20more%20cylinders%20tend,larger%20and%20high%2Dperformance%20vehicles.

Chase. "What Is the Average Length of Car Ownership?: Chase." *Credit Card, Mortgage, Banking, Auto*, www.chase.com/personal/auto/education/maintenance/average-length-of-car-ownership.

Jim. "Extract Patterns in R?: R-Bloggers." *R-Bloggers*, 15 Oct. 2022, www.r-bloggers.com/2022/10/extract-patterns-in-r/#google_vignette.

Leach, Felix. *The scope for improving the efficiency and environmental impact of internal combustion engines*, 2020. *Science Direct*, vol. 1, https://www.sciencedirect.com/science/article/pii/S2666691X20300063#sec0005.

Shalini, Lingampally. *Prediction of Automobile MPG using Optimization Techniques,*28/8/2021, *IEEE*, *https://ieeexplore.ieee.org/abstract/document/9563597/authors#authors*

Sivak, Michael. *Actual fuel economy of cars and light trucks: 1966-2017*. 2019. *Green Car Congress*, https://www.greencarcongress.com/2019/09/20190930-sivak.html.

str_extract: Extract the complete, 2020, *RDocumentation,* matchhttps://www.rdocumentation.org/packages/stringr/versions/1.5.1/topics/str_extract