Exploring Car Parts and its Manufacturers

Stat 184 Final Project

Logan Camacho, Joshua Santiago

12/18/2024

Introduction

With the invention of the first automobile in 1886, the quality of cars has continued to improve. Many different companies have since developed new technologies to stand out from the rest of the crowd. A common way to show this is through the engine of the car. Engines have many different parts, and the quality of those parts can change the performance of the engine. Engines of higher quality tend to be better optimized in certain areas, such as power/weight ratio, fuel economy, and reliability. However, the average consumer looking into the car market will be led astray when overemphasizing the importance of one component of the car, or even worse being completely lost at what to begin researching before they even step foot into a dealership. Constant letters delivered to vehicle owners describing immediate recalls that must be amended for, malicious salespersons trying to wring you out for every dollar on a car bound to fall apart once it leaves the lot, and even manufacturers keeping quiet on reliability issues that only the most seasoned car enthusiast would pick up on often scare the average person into buying the car that truly fits their needs within the current economy. Stigmas/stereotypes around the top-rated brands often do as much damage to their reputation as much as the same misconceptions to lesser-known brands, and so within our data analysis we seek to clear up any grey matter concerning information that you need to know before considering whether or not your next vehicle is the best fit for your lifestyle.

Data Collection and FAIR and CARE Principles

We are working with a dataset that consists of cars across the world in production during the 2023 model-year, contained within a file named “Car_Models.csv”. This dataset came from Kaggle, an open access site dedicated to providing easily accessible data, was created by Peshimaam Mohammed Muzammil and last updated 2 years ago. This is due to the tendency of automotive manufacturers releasing cars under their listed model-year on a wide range of dates based on EPA guidelines; for example, a 2023 car could be on the market from January 2nd, 2022, up until December 31st, 2023. Considering the different countries that each manufacturer came from, there were some cells of data with values that went against the universal units of measure in favor of domestic units corresponding to local guidelines, such as kilometers per liter instead of miles per gallon. While cars are only machines meant to be sold and driven by humans, the data we used still adhere to CARE principles and DEI when considering the sample space spanning countries around the world with technical specifications accessible to the public.

Exploratory Data Analysis

After completing data wrangling and making the data tidy, we can observe that our data set contains 404 unique entries, with 13 columns. Each entry is a different car, and the 13 columns listed are company, model, horsepower, torque, transmission type, drivetrain, fuel economy, number of doors, price, model year, body type, engine type, and number of cylinders. Many of these columns revolve around the quality and performance of the engine, which is something that led us to select this dataset in the first place. Many top-rated brands pride themselves on how good their engine is, so selecting this dataset allows us to properly conduct our research questions. Figure __ shows the results from tidying the dataset. Each numerical value is in the same unit of measurement, allowing us to properly conduct analysis.

	Company	Model	Horsepower_(hp)	Torque_(lb-ft)	Transmission.Type	Drivetrain	Fuel.Economy
1	Ferrari	812 Superfast	789	530	7-speed automatic	RWD	13/20 mpg
2	Ferrari	F8 Tributo	710	568	7-speed automatic	RWD	15/19 mpg
3	Ferrari	Roma	611	560	8-speed automatic	RWD	16/24 mpg
4	Ferrari	Portofino M	612	561	8-speed automatic	RWD	16/22 mpg
5	Ferrari	SF90 Stradale	986	590	8-speed automatic	AWD	18 mpg (combine
6	Ferrari	488 Pista	711	568	7-speed automatic	RWD	15/20 mpg
7	Ferrari	GTC4Lusso	680	514	7-speed automatic	AWD	12/17 mpg
8	Ferrari	488 GTB	661	561	7-speed automatic	RWD	15/22 mpg
9	Ferrari	California T	553	557	7-speed automatic	RWD	16/23 mpg
10	Ferrari	LaFerrari	950	664	7-speed automatic	RWD	12 mpg (combine
11	Toyota	Corolla	169	151	Continuously variable-speed automatic	FWD	30-33 mpg

Showing 1 to 11 of 404 entries, 13 total columns

Figure 1

We first wanted to explore how many cars were a part of the same company. This can be visualized through a bar graph which can be seen in *Figure 2*. The x-axis is the company name, and the y-axis is the total count of cars per company in this dataset.

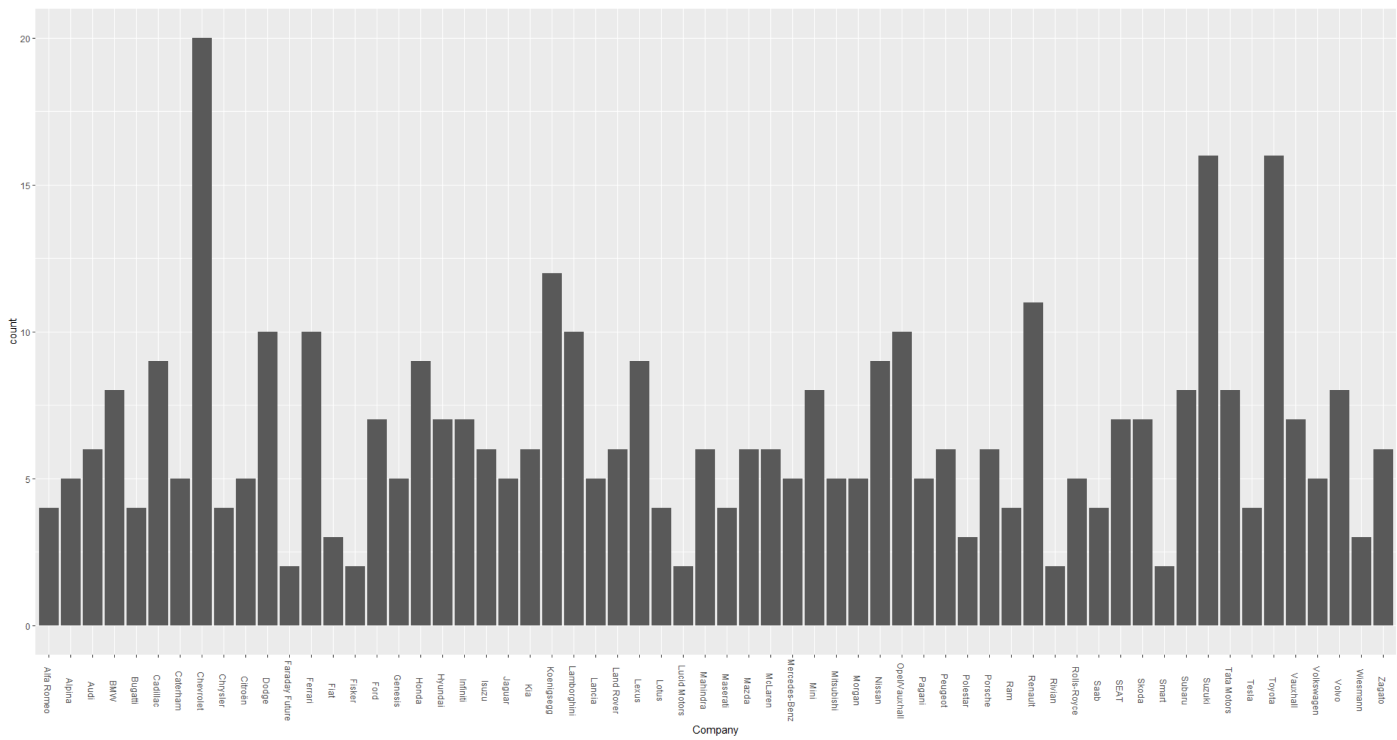

Figure 2

From this visualization, we can see that Chevrolet, Toyota, and Suzuki are the three companies with the most cars in this dataset. If we ever choose to conduct data analysis on car models of the same company, selecting one of these companies would give us the best range of models.

Another question we had upon looking at this dataset was whether or not there is a correlation between the torque and horsepower of a car. The horsepower of a car is the engine's maximum output and the speed at which it produces force. Torque measures how much power an engine can produce. If a car is able to produce more power, then the speeds it can reach are much higher. Torque and horsepower and measured differently and we aim to see if there is a correlation between them.

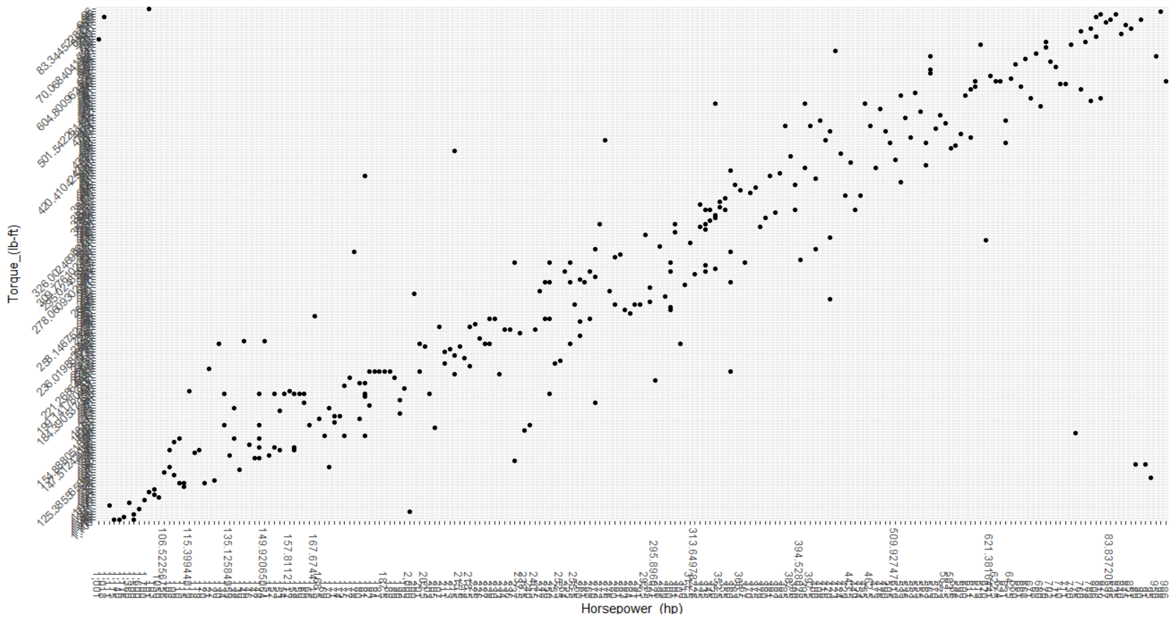

Figure 3

Figure 3 shows a scatterplot with the torque on the y-axis and the horsepower on the x-axis. We are able to notice that as horsepower increases, so does the torque. This confirms that there is a trend between a car's horsepower and torque.

Our exploratory data analysis gave us some insight on trends between engine performance, as well as showing the number of car models of each company in this dataset. Continuing with engines, we now aim to study potential trends with car parts, including engines and the car companies themselves.

Research Questions

Are manufacturers known for having transmission issues in their cars selling more CVT-default models than any other transmission type?

As one of the big Japanese auto makers in the United States, Nissan/Infiniti has slowly been losing their foothold in competition with their neighboring competitors Honda/Acura and Toyota/Lexus (slash denotes car brand/luxury brand from the same company). The sole reason why the latter two brands have sold well over a million vehicles within the 2023 Fiscal year (Reynolds) was due to their recognition in the market as the most reliable and affordable auto makers over the course of decades in the industry. Nissan was not too far behind on this chart, ranking 5th overall sitting right behind Honda, but not able to breach the one million mark, so we should not count them out entirely. All three of these companies started setting foot in the United States since the mid to late 1950s, and thus provided ample competition for American names once the assembly line became a requirement for every manufacturer to compete; despite their time in the mainland, Nissan in recent years has faced troublesome pushback on their reputation as a car maker.

Starting in 1992, Nissan introduced a new type of transmission to the world, and therefore their vehicle lineup named the “Continuous Variable Transmission” and has dealt with a mixed-response ever since then from their customers about the overall performance, reliability, and feel of the transmission. For the 2023 model-year, all of Nissan’s models only offered a CVT transmission for each of their cars

What type of drivetrain does the different companies use, and is there a correlation between drivetrain and horsepower?

Drivetrain is responsible for transferring power from the engine to the wheels. There are different kinds of drivetrain with the most common four being front-wheel drive (FWD), rear-wheel drive (RWD), all-wheel drive (AWD), four-wheel drive (4WD). Drivetrain can affect how quickly and efficiently a car accelerates, which is important when trying to reach top speed. RWD is considered the best when it comes to reaching speed quickly as it delivers power to the back wheels which can aid in proper acceleration. To test this, we looked at the different vehicles and the type of drivetrain they had. We also wanted to compare this between the different companies, as our ultimate goal is to guide viewers to which car best fits their lifestyle.

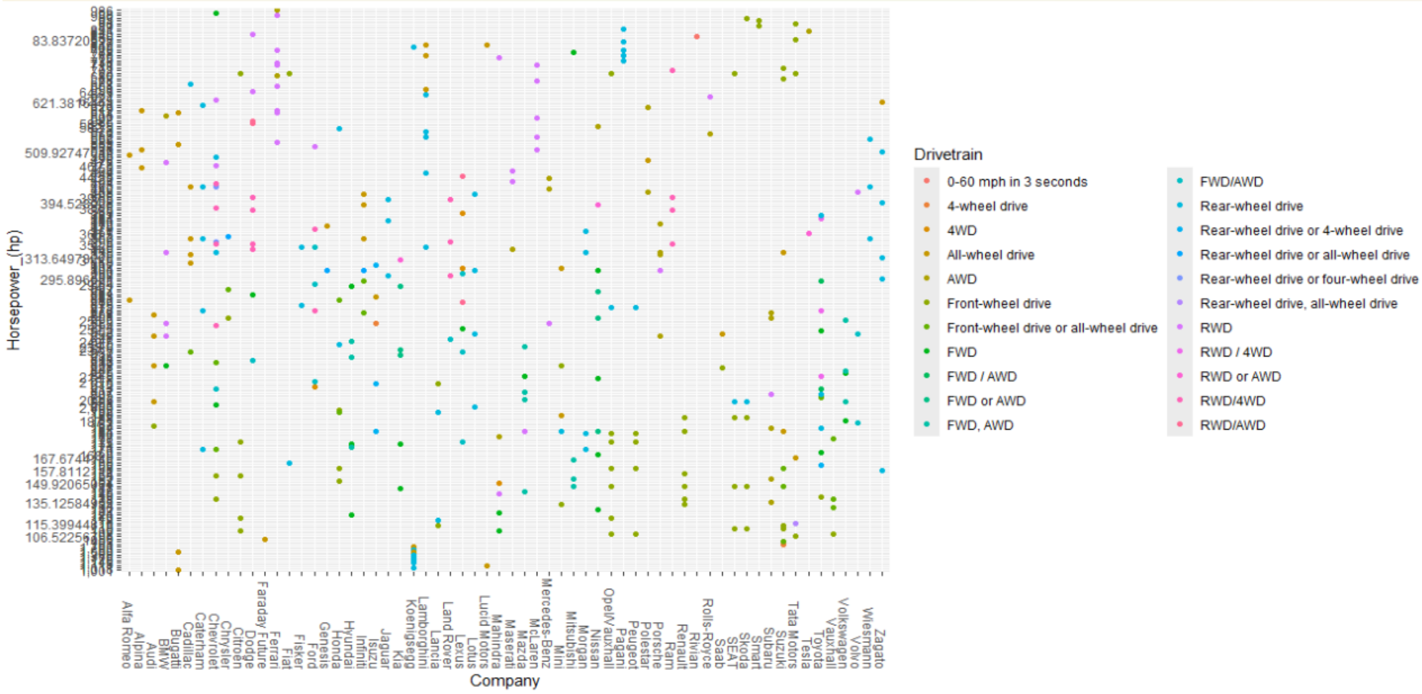

Figure 4

Figure 4 shows a scatterplot with the x-axis being the companies and the y-axis being the horsepower. Additionally, each point is color coded to represent the kind of drivetrain that vehicle has. We can see that vehicles with lower horsepower tend to have a FWD or AWD drivetrain. Another thing to note is that many luxury car brands like Ferarri, Maserati and Mcleran use RWD for their vehicles. These brands pride themselves on the quality and speed of their vehicles, and that can be seen through the high horsepower.

Conclusion

In the world of vehicles and having to endure the stress-inducing events of car buying, we discovered that there's more car companies out in the world than you'd expect, along with the intricacies of each and every one of their individual specifications. Although MPG is one of the most important features of a car, some enthusiasts might only want the best power output, and @fig-HpTorque showed us the direct correlation between the highest torque-horsepower pairs for each vehicle. Not to mention, links were able to be made between all sorts of drivetrain options and horsepower; most people only care for the practical application of the drivetrain, but the next buyer of that sports car could switch sides solely based on possible options of the drivetrain. Despite that, we learned that the lesser understood components of a car, specifically the transmission, could be the difference between thousands of dollars of repairs a year mitigating failures instead of only performing baseline maintenance that are universal aspects of car owning. When buying a car, we implore you to do your part in researching just as we did on each and every understandable technical aspect of the cars on the market to ensure that you truly learn where your money is being spent. One missed article, or one missed recall issues could spell out a year's worth of trouble concerning your car, and that was the most important lesson that we wanted to spread to all vehicle owners, racing enthusiasts, as well as future technicians looking for a lasting career.

Code Appendix

```r
raw_data <- read.table('Car_Models.csv', sep = ",", header = 1)
data_clean <- raw_data %>% #Separating unit of measurement from number
  separate_wider_delim(
    cols = Horsepower,
    delim = " ",
    names = c("Horsepower", "Unit"),
    too_few = "align_start",
    too_many = "merge"
  ) %>%
  separate_wider_delim(
    cols = Torque,
    delim = " ",
    names = c("Torque", "Units"),
    too_few = "align_start",
    too_many = "merge"
  )
#Taking the average value of rows with ranges given
averages_clean <- data_clean[c(17,19,20,22,24:28,30:32,35,37:41,43,47,49,69:73,91:97,114:115,117:118,132:145,227),] %>%
  separate_wider_delim(
    cols = Horsepower,
    delim = "-",
    names = c("lower", "upper"),
    too_few = "align_start"
  ) %>%
  separate_wider_delim(
    cols = Torque,
    delim = "-",
    names = c("lower1", "upper1"),
    too_few = "align_start"
  ) %>%
  #rowwise() %>%
  mutate( #setting na values to 0 and finding the averages
    lower = coalesce(as.numeric(lower), 0),
    upper = coalesce(as.numeric(upper), 0),
    Horsepower = (as.numeric(lower) + as.numeric(upper)) / 2,
    lower1 = coalesce(as.numeric(lower1), 0),
    upper1 = coalesce(as.numeric(upper1), 0),
    Torque = (as.numeric(lower1) + as.numeric(upper1)) / 2
  )
#Cleaning rows whose ranges were seperated from the initial sep
ranges_clean <- data_clean[c(100:104, 108:111),] %>%
  unite(
    col = "Temp",
    Horsepower,
    Unit,
    sep = " ",
    na.rm = TRUE
  ) %>%
  unite(
    col = "Temp1",
    Torque,
    Units,
    sep = " ",
    na.rm = TRUE
  ) %>%
  separate_wider_delim(
```

```r
    cols = Temp,
    delim = " - ",
    names = c("temp_lower", "temp_upper"),
    too_few = "align_start"
  ) %>%
  separate_wider_delim(
    cols = Temp1,
    delim = " - ",
    names = c("temp_lower1", "temp_upper1"),
    too_few = "align_start"
  ) %>%
  separate_wider_delim(
    cols = temp_lower,
    delim = " ",
    names = c("lower", "lower_unit"),
    too_few = "align_start"
  ) %>%
  separate_wider_delim(
    cols = temp_upper,
    delim = " ",
    names = c("upper", "Unit"),
    too_few = "align_start"
  ) %>%
  separate_wider_delim(
    cols = temp_lower1,
    delim = " ",
    names = c("lower1", "lower1_unit"),
    too_few = "align_start"

  ) %>%
  separate_wider_delim(
    cols = temp_upper1,
    delim = " ",
    names = c("upper1", "Units"),
    too_few = "align_start"
  ) %>%
  mutate(
    Horsepower = (as.numeric(lower) + as.numeric(upper)) / 2,
    Torque = (as.numeric(lower1) + as.numeric(upper1)) / 2
  )
#appending the ranges_clean and averages_clean back to the data_clean and dropping the uncleaned values
averages_clean_append <- averages_clean[-c(3:4, 6:7)]
ranges_clean_append <- ranges_clean[-c(3:5, 7:9)]
data_clean <- rbind(data_clean, averages_clean_append)
data_clean <- rbind(data_clean, ranges_clean_append)
data_clean <- data_clean[-c(17,19,20,22,24:28,30:32,35,37:41,43,47,49,69:73,91:97,114:115,117:118,132:145,227,100:104,108:111),]

#Converting Unit na values to hp and Units na values to lb-ft
data_clean <- data_clean %>%
  mutate_at(
    c("Unit"), ~replace_na(.,"hp")
  ) %>%
  mutate_at(
    c("Units"), ~replace_na(.,"lb-ft")
  )
#Converting all Horsepower values to hp measurements and Torque values to lb-ft measurements
hp_convert <- data_clean %>%

  filter(Unit != "hp") %>%
  mutate(
    Horsepower = as.numeric(Horsepower) * 0.98632007, #conversion value found online
    Unit = "hp"
  )
data_clean <- rbind(data_clean, hp_convert)

torque_convert <- data_clean %>%
  filter(Units != "lb-ft")%>%
  mutate(
    Torque = as.numeric(Torque) * 0.7375621493, #conversion value found online
    Units = "lb-ft"
  )
data_clean <- rbind(data_clean, torque_convert)
data_clean <- data_clean[-c(11:16, 80, 276:284, 336:343, 405, 407:411, 413:417, 420, 424:429, 433:436, 440),]
data_clean <- data_clean[-c(382),] #missed one value when removing duplicates

#Renaming Horsepower and Torque columns to include unit of measurement and dropping Unit and Units columns
names(data_clean)[3] <- "Horsepower_(hp)"
names(data_clean)[5] <- "Torque_(lb-ft)"
data_clean <- data_clean[-c(4,6)]

#Exploratory Data Analysis
#Getting counts of how many cars in each company
ggplot(data_clean) + geom_bar(aes(x = Company)) + theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust = 0.5))

#Exploring trend between horsepower and torque
ggplot(data_clean) + geom_point(aes(x = `Horsepower_(hp)`, y = `Torque_(lb-ft)`))
```

```
                                ) + theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust = 1)
                                          ) + theme(axis.text.y = element_text(angle = 45, vjust = 0.5, hjust = 1)
                                                     )
#Exploring trend between horsepower and company sorted by drivetrain
ggplot(data_clean, aes(x = Company, y = `Horsepower_(hp)`)) + geom_point(aes(color = Drivetrain)
                                                                          ) + theme(axis.text.x = element_text(angle = -90, vjust = 0.5, hjust

# Group by count of multiple columns
cvtTable <- data_clean %>%
  filter(str_detect(Transmission.Type, "CVT|Continuously variable|eCVT")) %>%     # Filter for "CVT" transmission
  select(Company, Model, Transmission.Type, Model.Year.Range) %>%
  group_by(Company)               # Group by company
print(cvtTable)
```

References

https://www.geeksforgeeks.org/how-to-split-column-into-multiple-columns-in-r-dataframe/

https://stackoverflow.com/questions/73674827/is-there-a-function-in-r-to-convert-the-units-of-multiple-columns-with-a-built-i

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://media.datacamp.com/legacy/image/upload/v1666806657/Marketing/Blog/ggplot2_cheat_sheet.pdf

2023 U.S. Vehicle Dependability Study(VDS) | J.D. Power

The Top Ten Models With Major Transmission Problems - Mister Transmission

National Auto Sales Numbers For All Automakers In 2023

XTRONIC CVT Continuously Variable Transmission | Nissan USA