Assignment 1 - Distribution Analysis

# Table of Contents

[Introduction 1](#_Toc62695998)

[Question 1: Construct a scatterplot with smoothing for ‘mpg’ vs. ‘disp’ 2](#_Toc62695999)

[Question 2: Construct a scatterplot with smoothing for ‘mpg’ vs. ‘hp’ 3](#_Toc62696000)

[Question 3: Construct a scatterplot with smoothing for ‘mpg’ vs. ‘wt’ 4](#_Toc62696001)

[Code from R Notebook 5](#_Toc62696002)

# Introduction

In this assignment, the correlations between **Fuel Economy** and a few other parameters in the *CarsDB* dataset are better understood. The assignment is divided into three parts, each discussing one of the following relationships:

1. Fuel efficiency vs engine displacement
2. Fuel efficiency vs gross horsepower
3. Fuel efficiency vs weight of the car

**Question 4** of the assignment will be answered next to each of these graphs.

Question 4:

a. Identifiable Trends for each of the three scatterplots  
b. Any anomalies that you see for each of the three scatterplots

# 

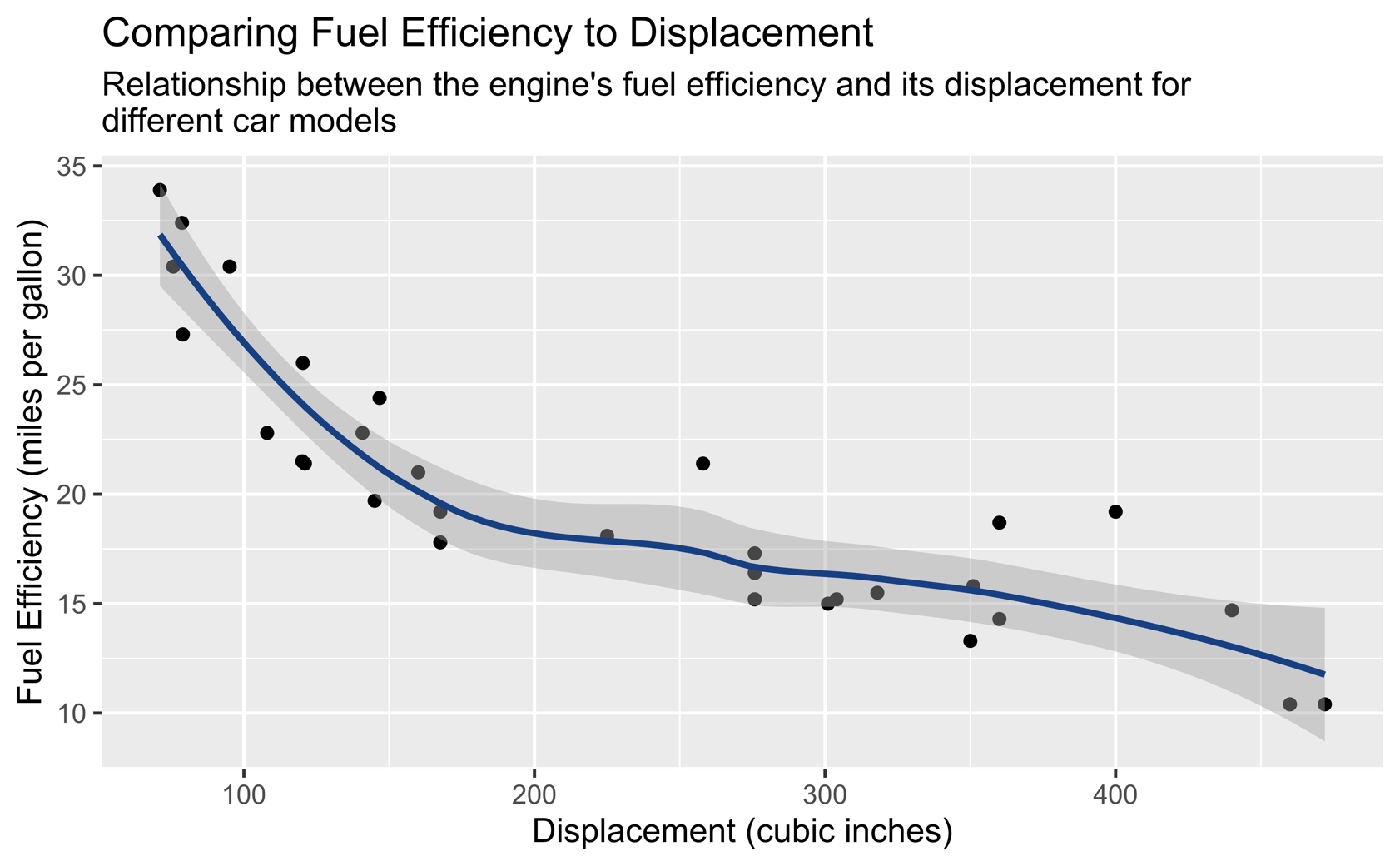
# Question 1: Construct a scatterplot with smoothing for ‘mpg’ vs. ‘disp’

Answer 4a:

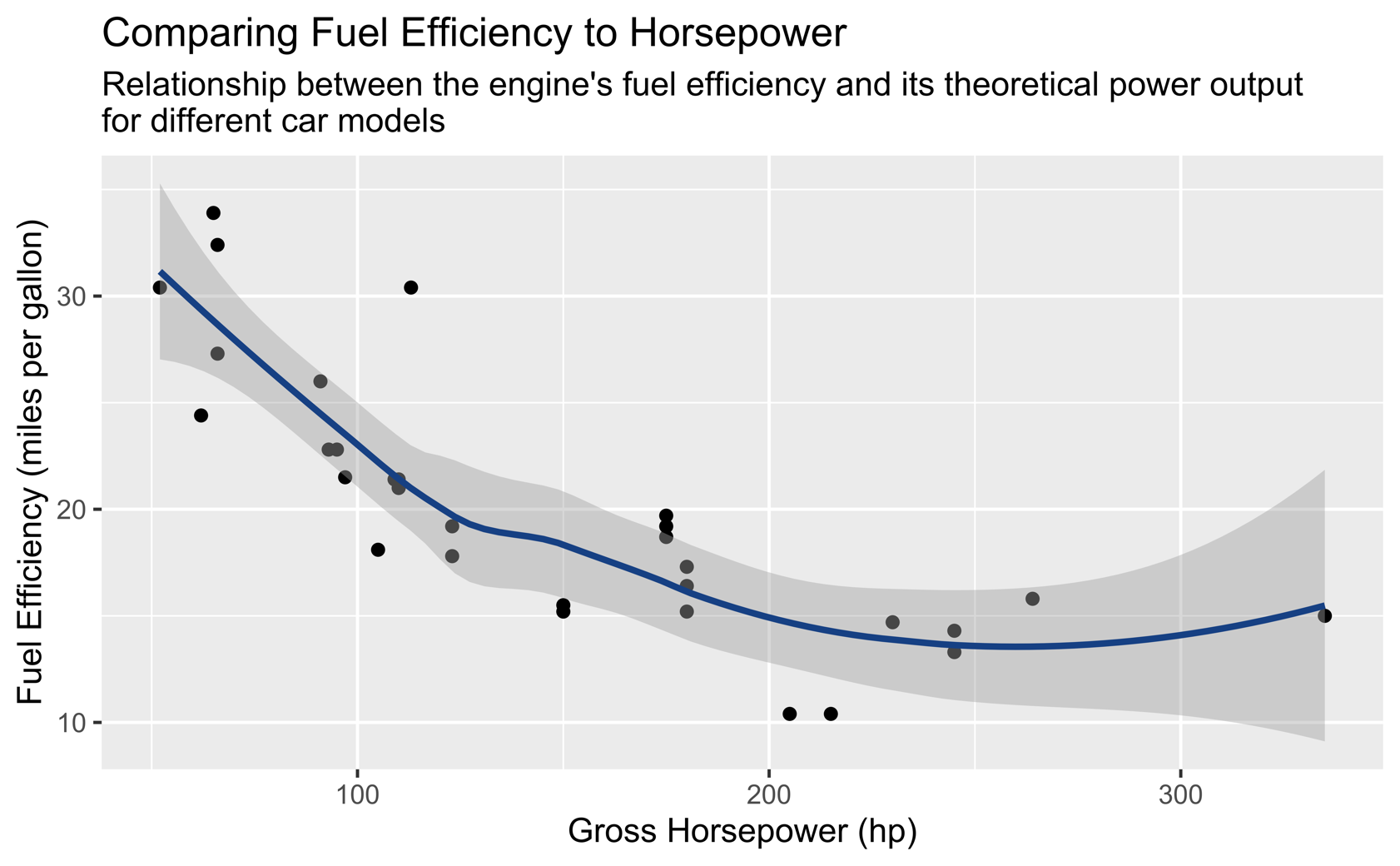
The fuel efficiency and engine displacement seem to have a **negative correlation**. As the displacement increases, the fuel efficiency decreases. The relationship is a **more logarithmic than linear in nature**. This means that the rate of change (fall) in fuel efficiency is larger for displacement values of around 100 cu.in., compared to around 300 cu.in.

Answer 4b:

The confidence intervals help identify potential outliers in the dataset. Three major anomalies or outliers fall within the displacement (*x-variable*) range of 250 cu.in to 400 cu.in. These **outliers** have been marked with a **green circle**. There are more points outside the confidence interval near 100 cu.in. displacement value but compared the three **marked in green** seemed to fall the furthest from the edge of their corresponding confidence intervals. Therefore, the probability of these points to be outliers is lower than the ones marked.



# Question 2: Construct a scatterplot with smoothing for ‘mpg’ vs. ‘hp’



Answer 4a:

Similar to the graph above, the fuel efficiency and gross horsepower share an almost **negative correlation**. As the gross horsepower increases, the fuel efficiency decreases. The relationship is **almost** **parabolic in nature**. However, as the change in trendline occurs over 300 hp, with only one point to determine the trend, a precise conclusion cannot be made. Though, it can be said with certainty that the **rate of change (fall) in fuel efficiency** decreases as the **gross horsepower** increases for hp values lesser than 270. To determine if the fuel economy increases after a certain threshold of gross horsepower is crossed, we will need more data points with hp greater than 270.

Answer 4b:

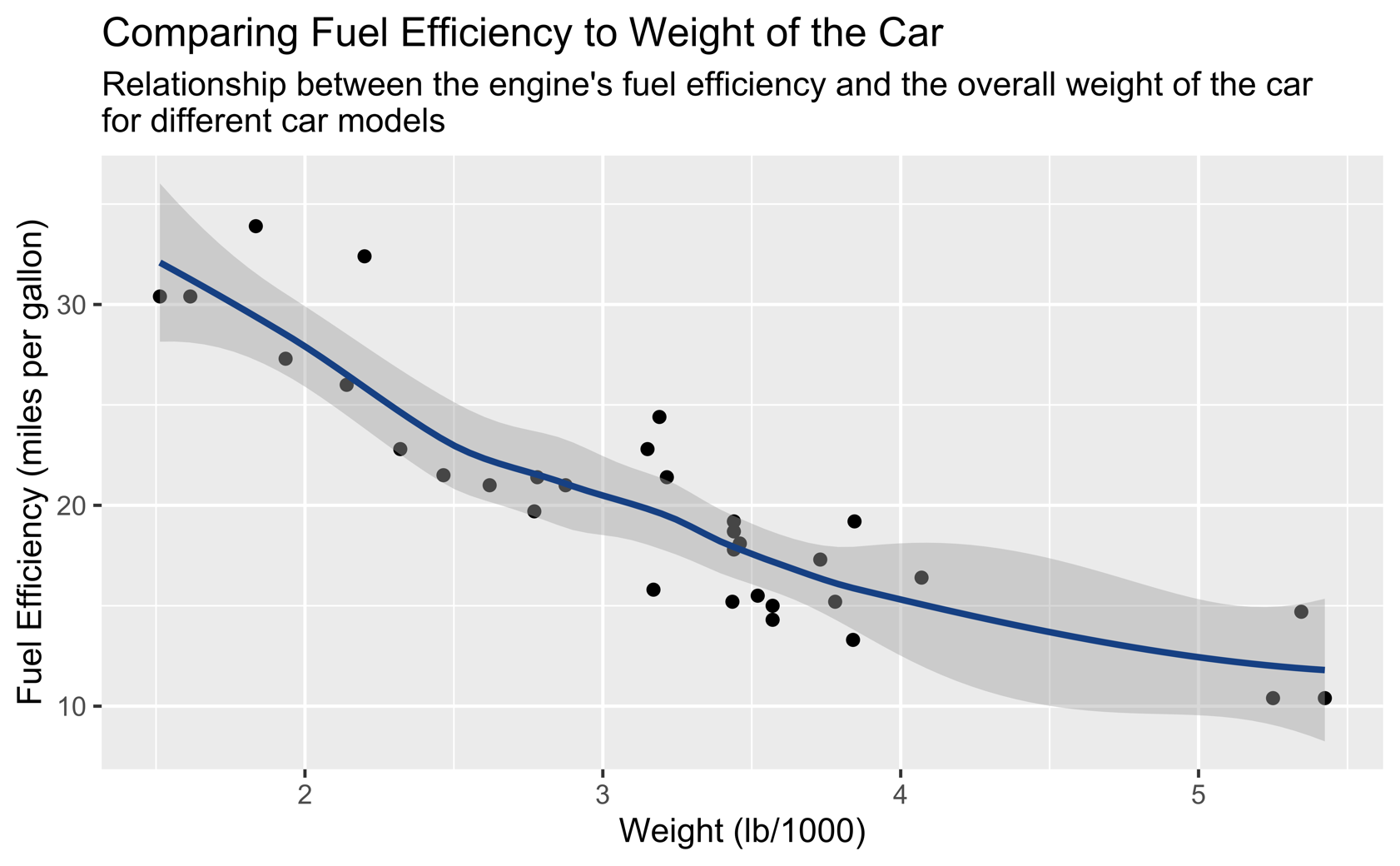
There are multiple potential outliers and a few anomalies in the dataset. **\*\***

**\*\*** The ones marked with a **green circle** are the ones furthest from the confidence interval and have the highest probability of being **outliers**.

On the other hand, there are two groups of points marked with a **purple circle** seem to be clustered in an odd way. They seem to be stacked over each other, signifying that for a particular value of gross horsepower, a few different values of fuel efficiencies are observed. Such clusters could represent improvement in fuel efficiency through the research and development, leading to better results for a specific horsepower. These points increase the bias in the dataset by having a major impact the trendline, and can be considered as anomalies. Some statistical manipulation, like taking the mean fuel efficiency of the points with same hp, can reduce the effect of the anomaly.

Finally, it is highly likely that the point **circled in red** is an outlier. Logically, higher gross horsepower means lower fuel efficiency, but the presence of this point creates an almost parabolic trendline. Another reason to suspect this point of being an anomaly is the large confidence intervals around the point. Lack of data around this point prevents us from making any concrete conclusions on the nature of the trendline.

# Question 3: Construct a scatterplot with smoothing for ‘mpg’ vs. ‘wt’



A group of the points around the weight of 3500lbs (**marked in red**) are outside the confidence interval as well. These could also be considered potential outliers as the confidence intervals is smaller there. A sign of weak correlation in the data, is the presence of many points outside the confidence interval.

A group of points marked with the **purple circle** highlights a potential anomaly. All three of these points have the potential to be an outlier due to their distance from the rest of the data points. There is more than 1000lbs difference between the points (marked with a purple line). The size of the gap and number of points (three) accentuates the possible presence of an anomaly. More data points with weight greater 4200lbs would give a better indication of the true nature of the relationship between the variables.

Answer 4a:

Similar to the above graphs, the fuel efficiency and weight of the car share a **negative correlation**. As the weight of the car increases, the fuel efficiency decreases. The relationship is **almost** **linear in nature**. Similar to the case discussed above, some anomalies prevent us from concluding with certainty the nature of the relationship. Usually in the real-world, not many variables have a truly linear relationship, but goal of modelling a relationship between two variables is to make predictions in a specified scope. In those terms, as long as the weight of the car is less than 4000lbs, one can use a linear model to represent the above relationship. To determine the same for weight greater than 4000lbs, more data points are needed.

Answer 4b:

The points **marked in green** are potential outliers.

# Code used to produce graphs

Install required packages install.packages(c(“lattice”, “psych”, “gmodels”, “readr”, “ggplot2”))

library(lattice)  
library(psych)  
library(gmodels)  
library(readr)  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

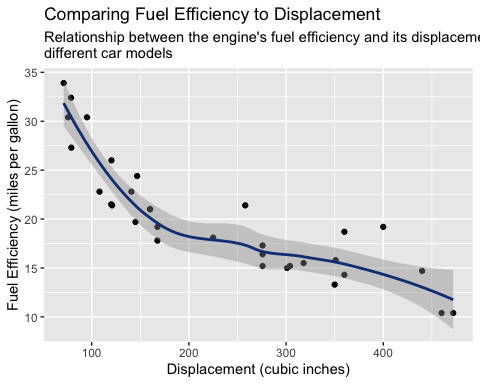
## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

Loading cars db data

cars\_db <- read.csv("carsDB.csv")  
cars\_db

## model mpg cyl disp hp drat wt qsec vs am gear carb  
## 1 Mazda RX4 21.0 6 160.0 110 3.90 2.620 16.46 0 1 4 4  
## 2 Mazda RX4 Wag 21.0 6 160.0 110 3.90 2.875 17.02 0 1 4 4  
## 3 Datsun 710 22.8 4 108.0 93 3.85 2.320 18.61 1 1 4 1  
## 4 Hornet 4 Drive 21.4 6 258.0 110 3.08 3.215 19.44 1 0 3 1  
## 5 Hornet Sportabout 18.7 8 360.0 175 3.15 3.440 17.02 0 0 3 2  
## 6 Valiant 18.1 6 225.0 105 2.76 3.460 20.22 1 0 3 1  
## 7 Duster 360 14.3 8 360.0 245 3.21 3.570 15.84 0 0 3 4  
## 8 Merc 240D 24.4 4 146.7 62 3.69 3.190 20.00 1 0 4 2  
## 9 Merc 230 22.8 4 140.8 95 3.92 3.150 22.90 1 0 4 2  
## 10 Merc 280 19.2 6 167.6 123 3.92 3.440 18.30 1 0 4 4  
## 11 Merc 280C 17.8 6 167.6 123 3.92 3.440 18.90 1 0 4 4  
## 12 Merc 450SE 16.4 8 275.8 180 3.07 4.070 17.40 0 0 3 3  
## 13 Merc 450SL 17.3 8 275.8 180 3.07 3.730 17.60 0 0 3 3  
## 14 Merc 450SLC 15.2 8 275.8 180 3.07 3.780 18.00 0 0 3 3  
## 15 Cadillac Fleetwood 10.4 8 472.0 205 2.93 5.250 17.98 0 0 3 4  
## 16 Lincoln Continental 10.4 8 460.0 215 3.00 5.424 17.82 0 0 3 4  
## 17 Chrysler Imperial 14.7 8 440.0 230 3.23 5.345 17.42 0 0 3 4  
## 18 Fiat 128 32.4 4 78.7 66 4.08 2.200 19.47 1 1 4 1  
## 19 Honda Civic 30.4 4 75.7 52 4.93 1.615 18.52 1 1 4 2  
## 20 Toyota Corolla 33.9 4 71.1 65 4.22 1.835 19.90 1 1 4 1  
## 21 Toyota Corona 21.5 4 120.1 97 3.70 2.465 20.01 1 0 3 1  
## 22 Dodge Challenger 15.5 8 318.0 150 2.76 3.520 16.87 0 0 3 2  
## 23 AMC Javelin 15.2 8 304.0 150 3.15 3.435 17.30 0 0 3 2  
## 24 Camaro Z28 13.3 8 350.0 245 3.73 3.840 15.41 0 0 3 4  
## 25 Pontiac Firebird 19.2 8 400.0 175 3.08 3.845 17.05 0 0 3 2  
## 26 Fiat X1-9 27.3 4 79.0 66 4.08 1.935 18.90 1 1 4 1  
## 27 Porsche 914-2 26.0 4 120.3 91 4.43 2.140 16.70 0 1 5 2  
## 28 Lotus Europa 30.4 4 95.1 113 3.77 1.513 16.90 1 1 5 2  
## 29 Ford Pantera L 15.8 8 351.0 264 4.22 3.170 14.50 0 1 5 4  
## 30 Ferrari Dino 19.7 6 145.0 175 3.62 2.770 15.50 0 1 5 6  
## 31 Maserati Bora 15.0 8 301.0 335 3.54 3.570 14.60 0 1 5 8  
## 32 Volvo 142E 21.4 4 121.0 109 4.11 2.780 18.60 1 1 4 2

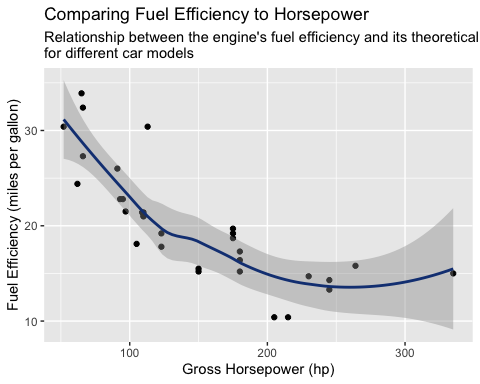
ggplot(data = cars\_db, aes(x=disp, y=mpg)) + geom\_point(color = "#000000") +   
 geom\_smooth(formula = y~x, method = 'loess', color = "#154083") +  
 labs(title = "Comparing Fuel Efficiency to Displacement",  
 subtitle = "Relationship between the engine's fuel efficiency and its displacement for \ndifferent car models",  
 x = "Displacement (cubic inches)",  
 y = "Fuel Efficiency (miles per gallon)")



ggsave("mpg\_vs\_disp.png", dpi = 1200)

## Saving 5 x 4 in image

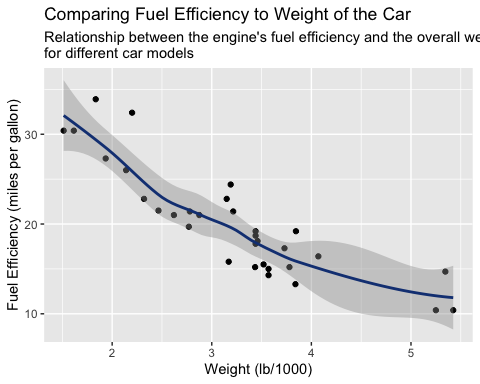
ggplot(data = cars\_db, aes(x=hp, y=mpg)) + geom\_point(color = "#000000") +   
 geom\_smooth(formula = y~x, method = 'loess', color = "#154083") +  
 labs(title = "Comparing Fuel Efficiency to Horsepower",  
 subtitle = "Relationship between the engine's fuel efficiency and its theoretical power output \nfor different car models",  
 x = "Gross Horsepower (hp)",  
 y = "Fuel Efficiency (miles per gallon)")



ggsave("mpg\_vs\_hp.png", dpi = 1200)

## Saving 5 x 4 in image

ggplot(data = cars\_db, aes(x=wt, y=mpg)) + geom\_point(color = "#000000") +  
 geom\_smooth(formula = y~x, method = 'loess', color = "#154083") +  
 labs(title = "Comparing Fuel Efficiency to Weight of the Car",  
 subtitle = "Relationship between the engine's fuel efficiency and the overall weight of the car \nfor different car models",  
 x = "Weight (lb/1000)",  
 y = "Fuel Efficiency (miles per gallon)")



ggsave("mpg\_vs\_wt.png", dpi = 1200)

## Saving 5 x 4 in image