Module 5 Exercise 2

Filipp Krasovsky

11/30/2020

2.2 Use R to understand how horsepower and weights are related to each other. Plot them using a scatter plot and color the data points using mpg. Do you see anything interesting/useful here? Report your observations with this plot. Now let us cluster the data on this plane in a "reasonable" number of groups. Show your plot where the data points are now colored with the cluster information and provide your interpretations. (10 points)

```
#install packages
library(RMySQL)
```

```
## Loading required package: DBI
```

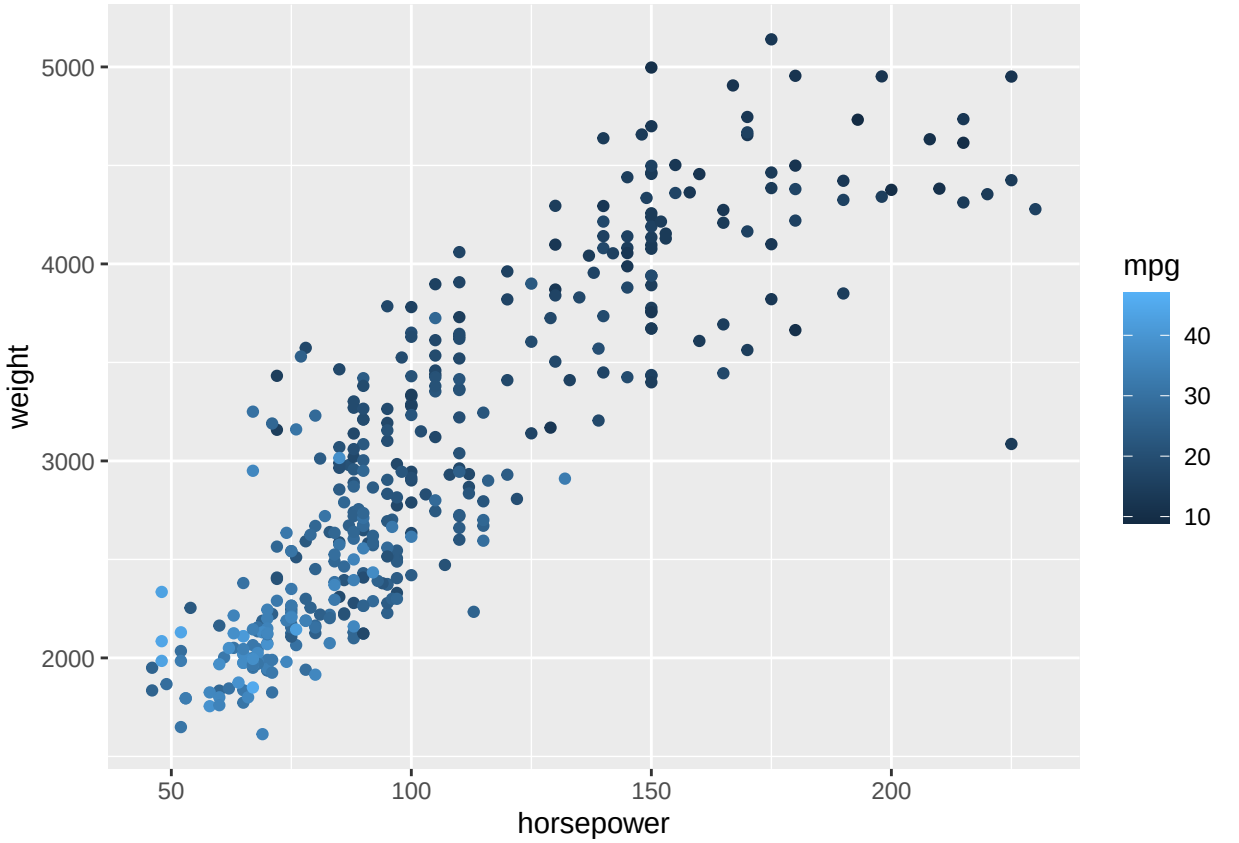
```
#connect to db
default_authentication_plugin='admin'
db = dbConnect(MySQL(),user='root',password='admin', dbname='500b', host='127.0.0.1')

#extract mpg data.
mpg_df = fetch(dbSendQuery(db,"SELECT * FROM mpg"),n=-1)
mpg_df = as.data.frame(mpg_df)

#print head for sanity check
head(mpg_df)
```

```
##   mpg cylinders displacement horsepower weight acceleration model year origin
## 1  18         8          307        130   3504           12         70      1
## 2  15         8          350        165   3693         11.5         70      1
## 3  18         8          318        150   3436           11         70      1
## 4  16         8          304        150   3433           12         70      1
## 5  17         8          302        140   3449         10.5         70      1
## 6  15         8          429        198   4341           10         70      1
##                   car name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
#plot horsepower and weight, color with mpg.
library(ggplot2)
ggplot(mpg_df,aes(horsepower,weight,color=mpg)) + geom_point()
```



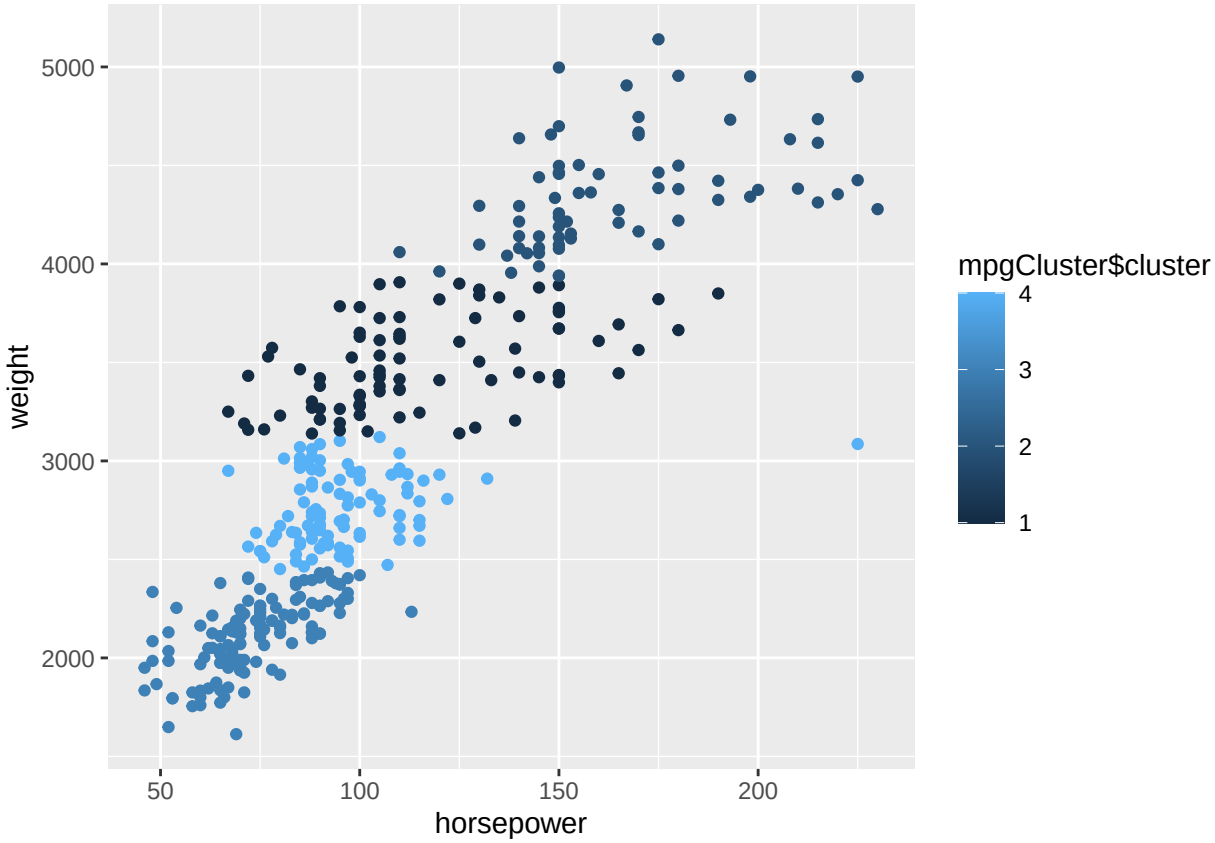
Initial observations demonstrate two major points of interest: 1. there is a very strong positive correlation between weight and horsepower. This is rather intuitive as the heavier a vehicle is, the more horsepower is required to propel it.

2. Combinations of horsepower and weight with increased values for both are associated, on face value, with a lower mpg.

Given that the gradient of mpg falls between ~10 and ~50, we can create 4 different clusters for our kmeans.

```
set.seed(20)
mpgCluster <- kmeans(mpg_df[,4:5],4,nstart=20)

ggplot(mpg_df,aes(horsepower,weight,color=mpgCluster$cluster)) + geom_point()
```



In this instance, the output of the kmeans clustering process can be interpreted as follows: Cluster 1 = 20-30 mpg Cluster 2 = 10-20 mpg Cluster 3 = 40-50 mpg Cluster 4 = 30-40 mpg

Overall, the learning model captures the noise in the data well but makes some mistakes at smaller values of horsepower/weight.