MySQL with R ‘auto’ database

Amin Fesharaki

8/1/2021

```
library(ggplot2) #Data Visualization with ggplot2
library(dplyr) #Data Transformation with dplyr
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(RMySQL)
```

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

```
con <- dbConnect(MySQL(), user='root', password='AIctex92#1aB', dbname='auto', host='localhost')


# Send query to pull requests in batches
res <- dbSendQuery(con, "SELECT * FROM mpg")
data <- dbFetch(res)
dbDisconnect(con)
```

[1] TRUE

```
#Pull mpg, horsepower, and weight data
SubD <- subset(data,select=c(mpg, horsepower,weight))

#correlation matrix
round(cor(SubD), digits = 4)
```

```
           mpg horsepower  weight
```

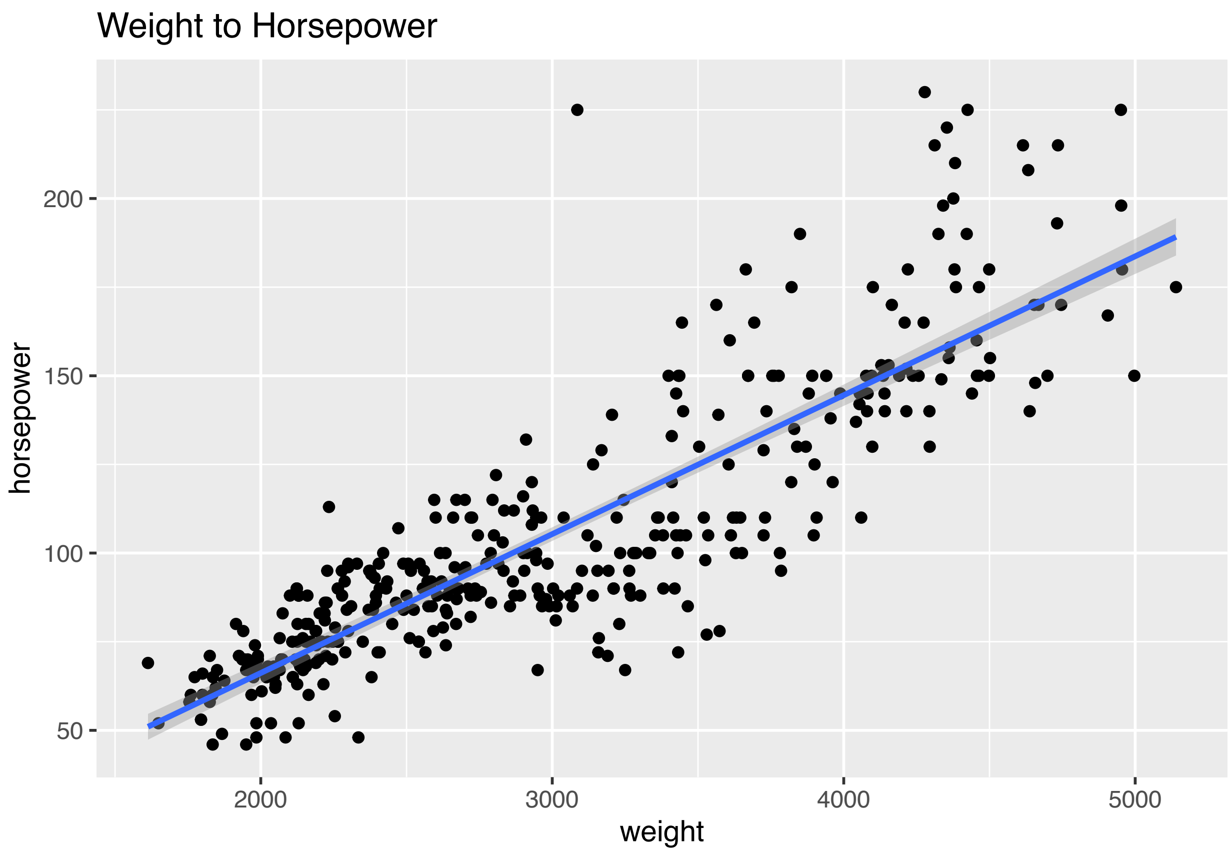
mpg 1.0000 -0.7777 -0.8315 horsepower -0.7777 1.0000 0.8645 weight -0.8315 0.8645 1.0000

```
#full range: fit spans the full range of plot // se: confidence interval is displayed around smooth //
geom_smooth(method="auto", se=TRUE, fullrange=FALSE, level=0.95)
```

geom_smooth: na.rm = FALSE, orientation = NA, se = TRUE stat_smooth: na.rm = FALSE, orientation = NA, se = TRUE, fullrange = FALSE, level = 0.95, method = auto position_identity

```
#Plotting the data points and adding a regression line (method lm = linear model)
Graph <- ggplot(SubD, aes(x=weight,y=horsepower)) + geom_point() + geom_smooth(method=lm) + labs(title=
Graph
```

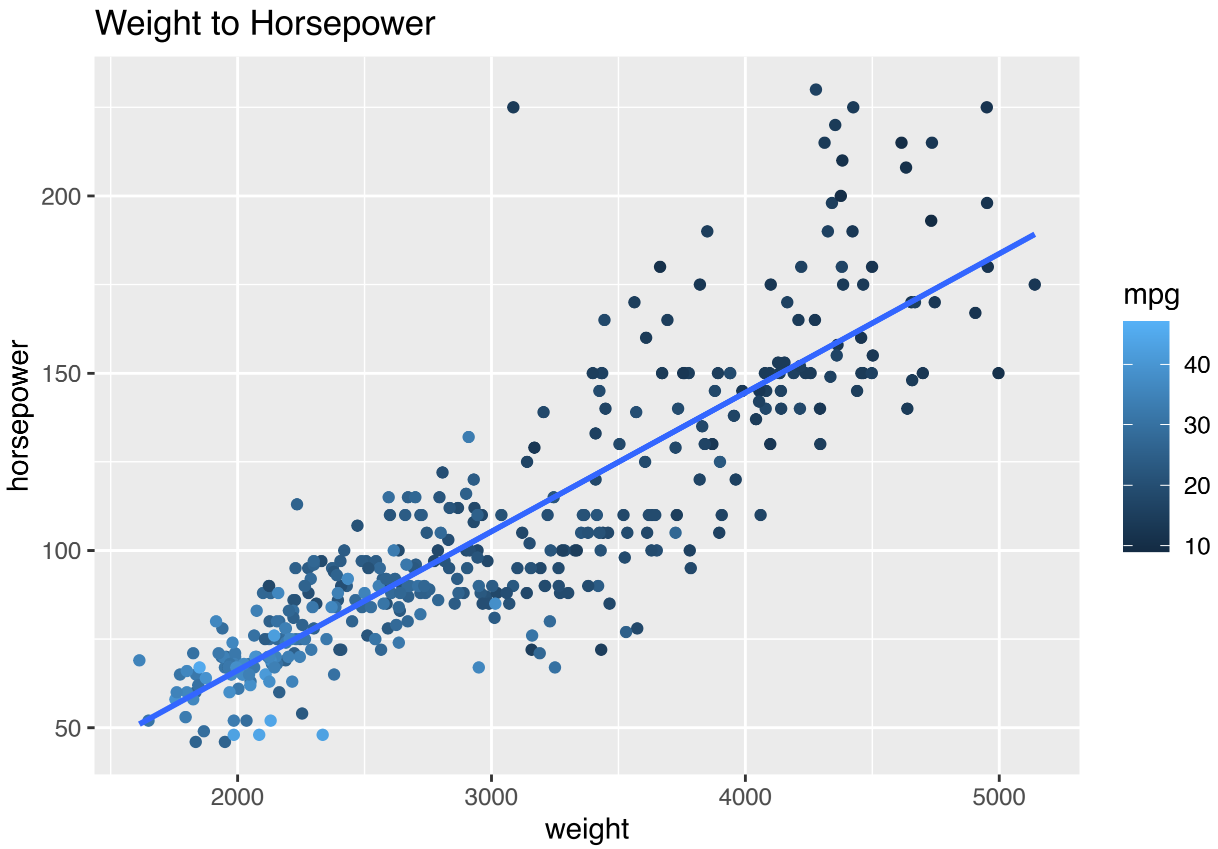
```
## `geom_smooth()` using formula 'y ~ x'
```



This scatter plot graphs weight vs horsepower. From this chart, we can see that these two variables are moderately positive correlated with each other. It is more correlated towards the less weight/horsepower, and starts spreading more towards higher values

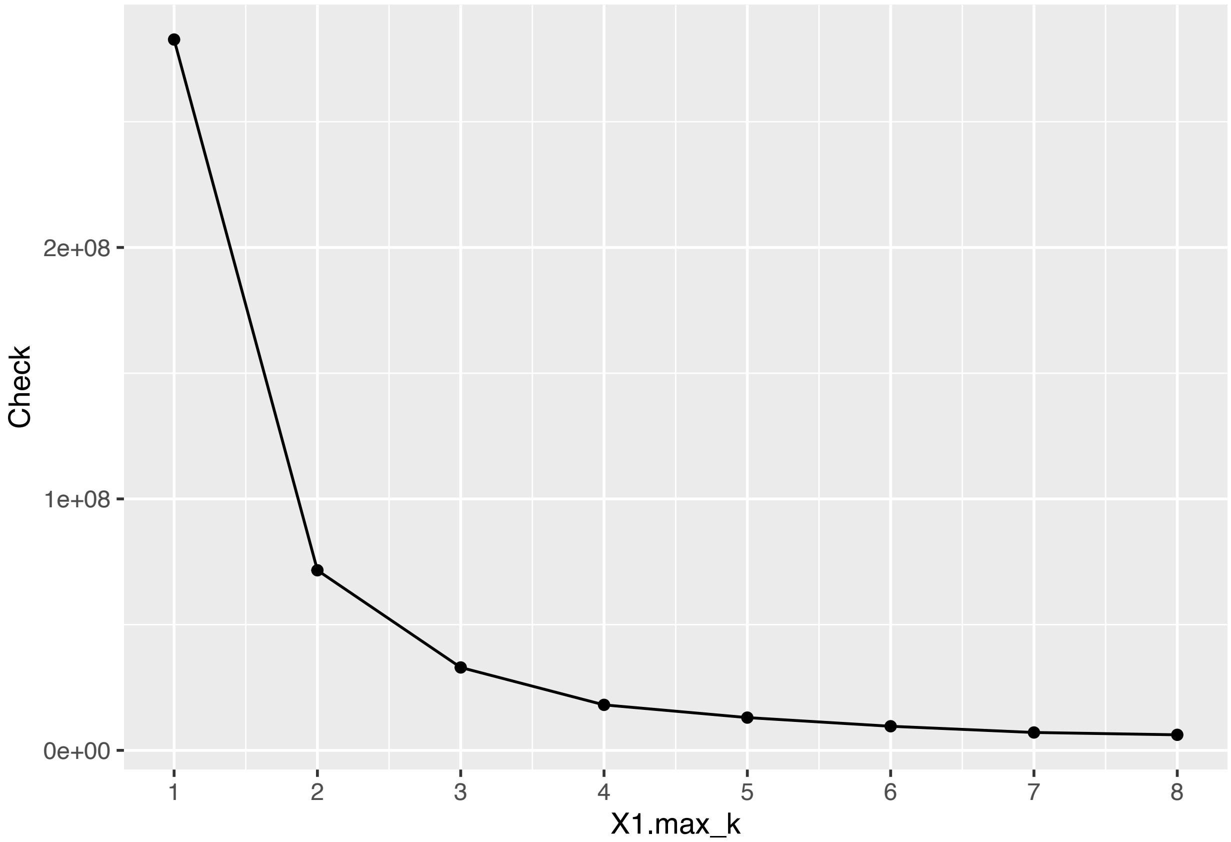
```
#Create Graph with labels, extend the regression line, and color coordinate the 'Likes'
Group <- ggplot(SubD, aes(x=weight,y=horsepower, color=mpg)) + geom_point() + geom_smooth(method=lm, se
Group
```

```
## `geom_smooth()` using formula 'y ~ x'
```



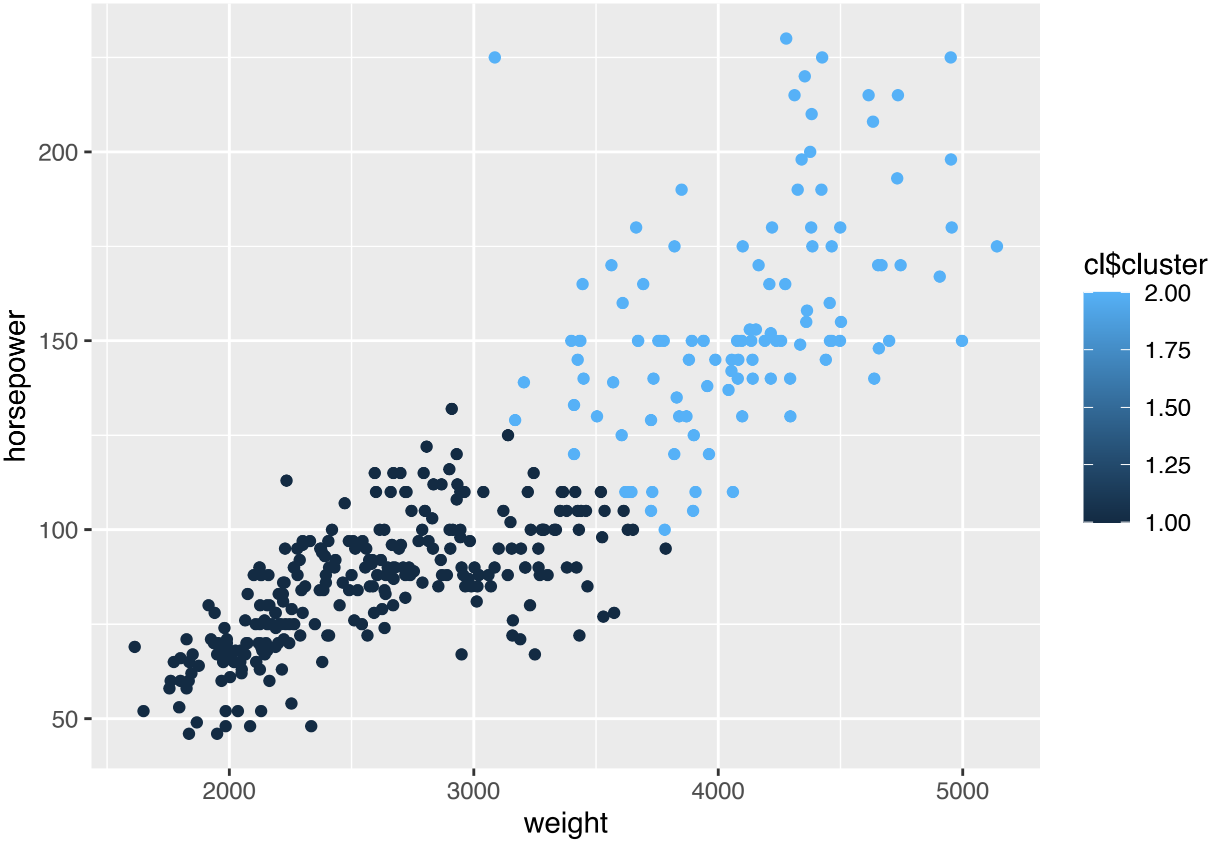
This scatter plot graphs weight to horsepower with color points using mpg. We can see that higher mpg values tend to be stronger positively correlated with horsepower and weight than smaller mpg values

```
##Determining the optimal number of clusters for the data (using the elbow method)##
#Create a function to compute and store the total within clusters sum of squares
kmean_withinss <- function(k) {
    cluster <- kmeans(SubD, k)
    return (cluster$tot.withinss)
}

# Set maximum cluster to 8
max_k <-8
# Run function over a range 1 to 8
Check <- sapply(1:max_k, kmean_withinss)

# Create a data frame to store the algorithm in order to plot the graph
elbow <-data.frame(1:max_k, Check)

# Plot the graph with gglop
ggplot(elbow, aes(x = X1.max_k, y = Check)) +
    geom_point() +
    geom_line() +
    scale_x_continuous(breaks = seq(1, 20, by = 1))
```



From the elbow method, we can conclude that 2 is the optimal number of clusters for this data set (where the curve starts to have a diminishing return)

```
#Scale the data, only include weight and horsepower Column, and use optimal number of clusters (2). Cal
cl <- kmeans(scale(SubD[,c(2,3)]),2,nstart=25)

clGraph= ggplot(SubD,aes(weight,horsepower,color=cl$cluster))+geom_point()
clGraph
```



From the cluster graph, we can see that one cluster is more tightly packed than the other cluster. This can be interpreted as the smaller weight and horsepower values (with higher mpg values) is more correlated with each other than the higher weight and horsepower values (with lower mpg values). In other words, the data points tend to spread out from eachother as you increase weight and horsepower