MT Cars Analysis

Jayesh Gokhale

5/1/2021

Analysis on Motor Trend Cars Dataset

We have to answer two questions -

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

I looked up the help file of mtcars data set and based on data, I gathered that following are the numeric variables

mpg (Miles/(US) gallon), disp (Displacement), hp (Gross Horsepower), drat (Gear Axle Ratio), wt (Weight in 1000 lbs) and qsec (1/4 mile time). Folliwng are

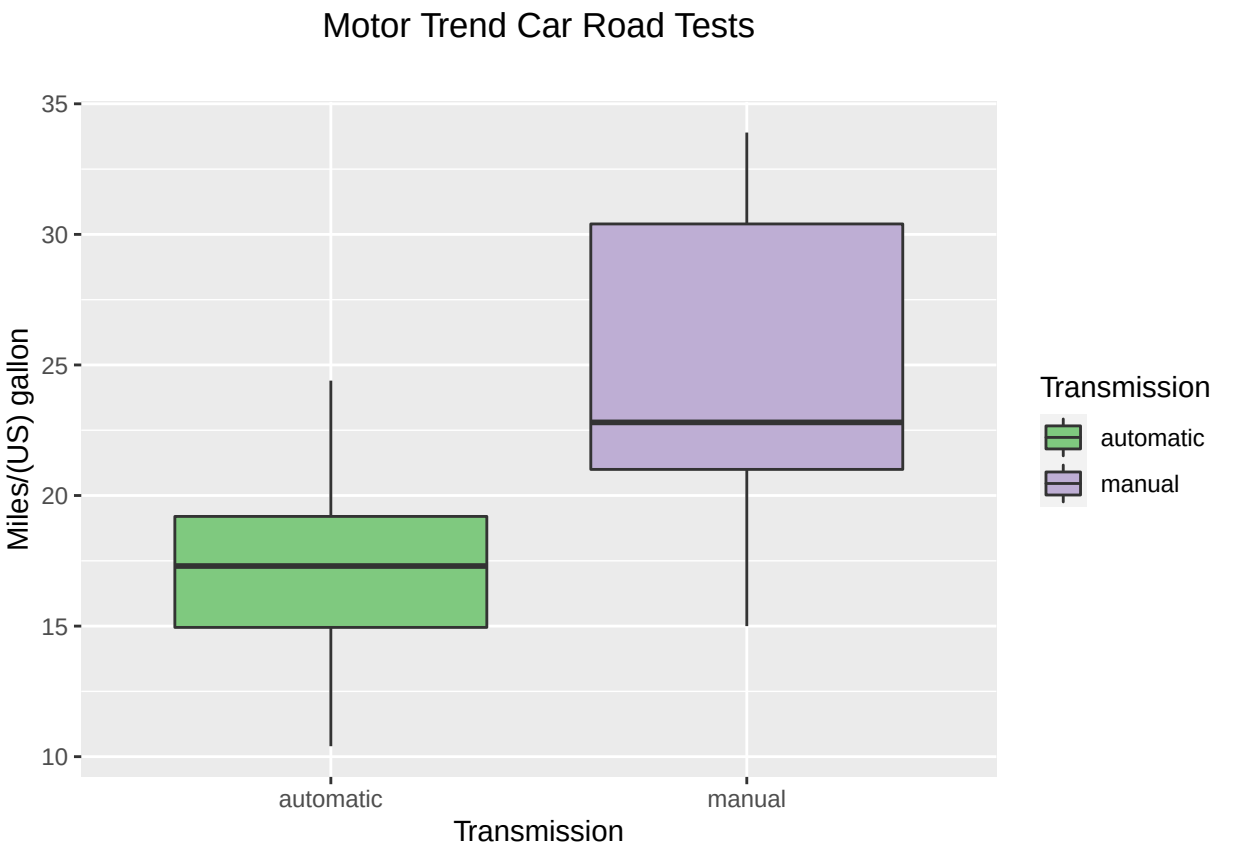
Following are the categorical variables -

cyl (Number of cylinders), vs (Engine: 0 = v-shaped, 1 = straight), **am (Transmission: 0 = automatic, 1 = manual)**, gear (Number of forward gears), carb (Number of Carburetors)

Let us convert all categorical variables to factors.

Now since the question is whether automatic/manual transmission is better for MPG, let us begin with a box plot of mpg vs am

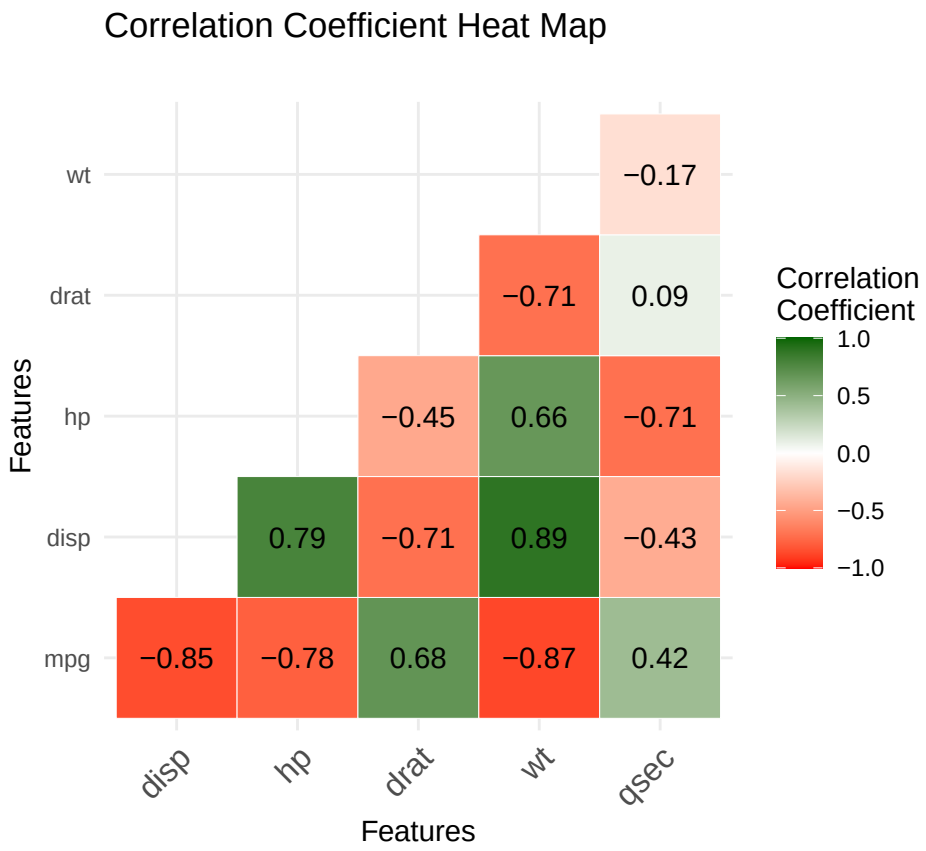
```
ggplot(data=mt, aes(x=am,y=mpg,fill=am)) +
  geom_boxplot(outlier.colour="black",
               outlier.size=2,position=position_dodge(1)) +
  labs(title = "Motor Trend Car Road Tests",
       subtitle = "",
       y = "Miles/(US) gallon", x = "Transmission") +
    scale_fill_brewer(name = "Transmission",palette="Accent")  +
  theme(plot.title = element_text(hjust = 0.5))
```



On the face of it, manual transmission does appear to have significantly higher MPG. But the results may be misleading unless we take into account the effect of other variables.

Let us first draw a heatmap of the continuous variables.

```
correlation.matrix <- cor(mt[,c("mpg","disp","hp","drat","wt","qsec")])
correlation.matrix <- round(correlation.matrix,2)
correlation.matrix[upper.tri(correlation.matrix)] <- NA
diag(correlation.matrix) <- NA
row.corr.matrix <- melt(correlation.matrix,na.rm=TRUE)
colnames(row.corr.matrix) <- c("F1","F2","CORR")
ggplot(data = row.corr.matrix, aes(x=F1, y=F2, fill=CORR)) +
  geom_tile(color="white")+
  labs(title = "Correlation Coefficient Heat Map",
       subtitle = "",
       y = "Features", x = "Features") +
  scale_fill_gradient2(low = "red", high = "darkgreen", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Correlation\nCoefficient") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()+
  geom_text(aes(x=F1, y=F2, label = CORR), color = "black", size = 4)
```



From the heatmap we can see that wt, disp and hp are negatively highly correlated to mpg and drat is positively correlated to mpg. Now we can also observe that wt, disp and hp are correlated amongst themselves. drat is also negatively correlated to wt.

Let us now look at the relationship between am (Transmission) and mpg while adjusting for each of the four numeric variables viz. wt, disp, hp and drat.

```
plotWeight <- ggplot(data=mt, aes(x=wt,y=mpg,fill=am)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title = "",
       subtitle = "",
       y = "MPG", x = "Weight") +
  scale_fill_brewer(name = "Transmission",palette="OrRd")  +
  theme(plot.title = element_text(hjust = 0.5))

plotDisp <- ggplot(data=mt, aes(x=disp,y=mpg,fill=am)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title = "",
       subtitle = "",
       y = "MPG", x = "Displacement") +
  scale_fill_brewer(name = "Transmission",palette="OrRd")  +
  theme(plot.title = element_text(hjust = 0.5))

plotHP <- ggplot(data=mt, aes(x=hp,y=mpg,fill=am)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title = "",
       subtitle = "",
```

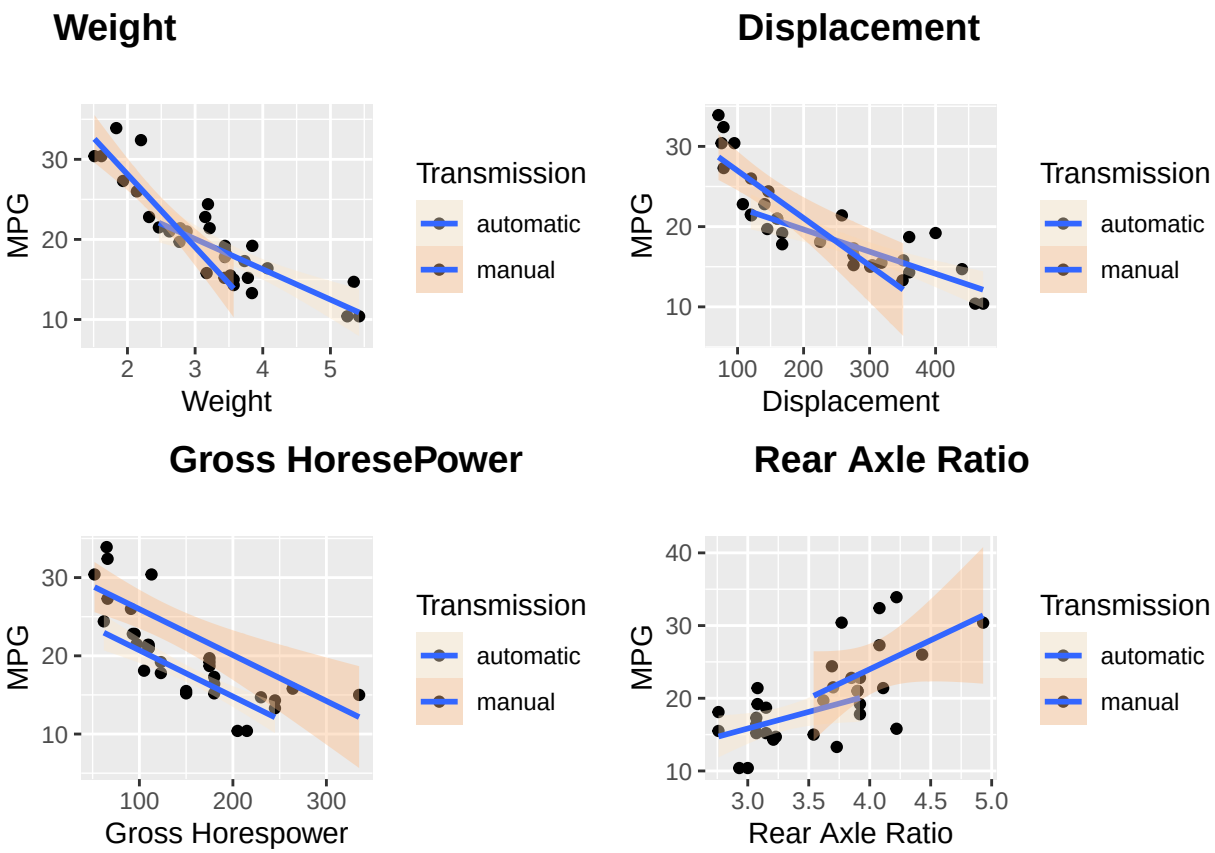
```
       y = "MPG", x = "Gross Horespower") +
  scale_fill_brewer(name = "Transmission",palette="OrRd")  +
  theme(plot.title = element_text(hjust = 0.5))

plotDRAT <- ggplot(data=mt, aes(x=drat,y=mpg,fill=am)) +
  geom_point() + geom_smooth(method=lm) +
  labs(title = "",
       subtitle = "",
       y = "MPG", x = "Rear Axle Ratio") +
  scale_fill_brewer(name = "Transmission",palette="OrRd")  +
  theme(plot.title = element_text(hjust = 0.5))

figure <- ggarrange(plotWeight, plotDisp, plotHP, plotDRAT,
                    labels = c("Weight", "Displacement","Gross HoresePower","Rear Axle Ratio"),
                    ncol = 2, nrow = 2)
```

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

```
figure
```



Let us build a model from numeric to numeric first

```
lr.numeric <- glm(mpg~disp+hp+drat+wt+qsec,data=mt)
summary(lr.numeric)
```

```
##
## Call:
## glm(formula = mpg ~ disp + hp + drat + wt + qsec, data = mt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5404  -1.6701  -0.4264   1.1320   5.4996
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.53357   10.96423   1.508  0.14362
## disp         0.00872    0.01119   0.779  0.44281
## hp          -0.02060    0.01528  -1.348  0.18936
## drat         2.01578    1.30946   1.539  0.13579
## wt          -4.38546    1.24343  -3.527  0.00158 **
## qsec         0.64015    0.45934   1.394  0.17523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6.543428)
##
##     Null deviance: 1126.05  on 31  degrees of freedom
## Residual deviance:  170.13  on 26  degrees of freedom
## AIC: 158.28
##
## Number of Fisher Scoring iterations: 2
```

From factors we will choose cylinder since as per plots, gear, carb, vs are not adding much value it seems

This is clearly a trick question. The thing is that we have to find some other variable which are impacting MPG and Transmission is just a co-incidence. That is the crux of the problem

Similar to Breath mint