Math 189 Homework 2: Exploratory Data Analysis of mtcars

Yunchun Pan, Siqing Lyu, Nathan Ng, Pudan Xu

1/18/2021

Introduction

In this homework, we examine the Motor Trend Car Road Tests dataset **mtcars.csv**[1]. The file contains data extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). This dataset has the following 11 variables:

- **mpg**: Miles per gallon
- **cyl**: Number of cylinders
- **disp**: Displacement(cubic inch)
- **hp**: Gross horsepower
- **drat**: Rear axle ratio
- **wt**: Weight
- **qsec**: 1/4 mile time
- **vs**: Engine (0 = V-shaped, 1 = straight)
- **am**: Transmission (0 = automatic, 1 = manual)
- **gear**: Number of forward gears
- **carb**: Number of carburetors

Our Work

We first load the Motor Trend Car Road Tests dataset into R to get the data we use throughout this assignment.

```
mtcars <- read.csv("~/Desktop/ma189/Data/mtcars.csv")
```

After we load in our dataset, we calculate sample mean and sample variance of each variable and store them into matrices to be M_m and M_v respectively.

```
ncols = dim(mtcars)[2]
varMean <- colMeans(mtcars[,2:ncols])
Mm <-t(as.matrix(varMean))
Mm <- Mm[1,]
Mm
```

```
##        mpg        cyl       disp         hp       drat         wt
##  20.090625   6.187500 230.721875 146.687500   3.596563   3.217250
##       qsec         vs         am       gear       carb
##  17.848750   0.437500   0.406250   3.687500   2.812500
```

[1]Source: The Motor Trend automobiles data are presented in *Building Multiple Regression Models Interactively* from *Biometrics* by Harold V. Henderson and Paul F. Velleman (1981).

```
Mv <- Mm
for (i in 2:dim(mtcars)[2]) {
  Mv[i-1] = var(mtcars[,i])
}
Mv
```

```
##          mpg          cyl         disp           hp         drat
## 3.632410e+01 3.189516e+00 1.536080e+04 4.700867e+03 2.858814e-01
##           wt         qsec           vs           am         gear
## 9.573790e-01 3.193166e+00 2.540323e-01 2.489919e-01 5.443548e-01
##         carb
## 2.608871e+00
```

Then, we calculate the sample variance-covariance matrix and the sample correlation matrix and store the result into matrices M_{vc} and M_{cor} respectively.

```
Mvc <- var(mtcars[,2:ncols])
Mvc
```

```
##              mpg         cyl        disp          hp         drat
## mpg    36.324103  -9.1723790  -633.09721 -320.732056   2.19506351
## cyl    -9.172379   3.1895161   199.66028  101.931452  -0.66836694
## disp -633.097208 199.6602823 15360.79983 6721.158669 -47.06401915
## hp   -320.732056 101.9314516  6721.15867 4700.866935 -16.45110887
## drat    2.195064  -0.6683669   -47.06402  -16.451109   0.28588135
## wt     -5.116685   1.3673710   107.68420   44.192661  -0.37272073
## qsec    4.509149  -1.8868548   -96.05168  -86.770081   0.08714073
## vs      2.017137  -0.7298387   -44.37762  -24.987903   0.11864919
## am      1.803931  -0.4657258   -36.56401   -8.320565   0.19015121
## gear    2.135685  -0.6491935   -50.80262   -6.358871   0.27598790
## carb   -5.363105   1.5201613    79.06875   83.036290  -0.07840726
##               wt         qsec          vs          am        gear
## mpg   -5.1166847   4.50914919   2.01713710   1.80393145   2.1356855
## cyl    1.3673710  -1.88685484  -0.72983871  -0.46572581  -0.6491935
## disp 107.6842040 -96.05168145 -44.37762097 -36.56401210 -50.8026210
## hp    44.1926613 -86.77008065 -24.98790323  -8.32056452  -6.3588710
## drat  -0.3727207   0.08714073   0.11864919   0.19015121   0.2759879
## wt     0.9573790  -0.30548161  -0.27366129  -0.33810484  -0.4210806
## qsec  -0.3054816   3.19316613   0.67056452  -0.20495968  -0.2804032
## vs    -0.2736613   0.67056452   0.25403226   0.04233871   0.0766129
## am    -0.3381048  -0.20495968   0.04233871   0.24899194   0.2923387
## gear  -0.4210806  -0.28040323   0.07661290   0.29233871   0.5443548
## carb   0.6757903  -1.89411290  -0.46370968   0.04637097   0.3266129
##             carb
## mpg  -5.36310484
## cyl   1.52016129
## disp 79.06875000
## hp   83.03629032
## drat -0.07840726
## wt    0.67579032
## qsec -1.89411290
## vs   -0.46370968
## am    0.04637097
```

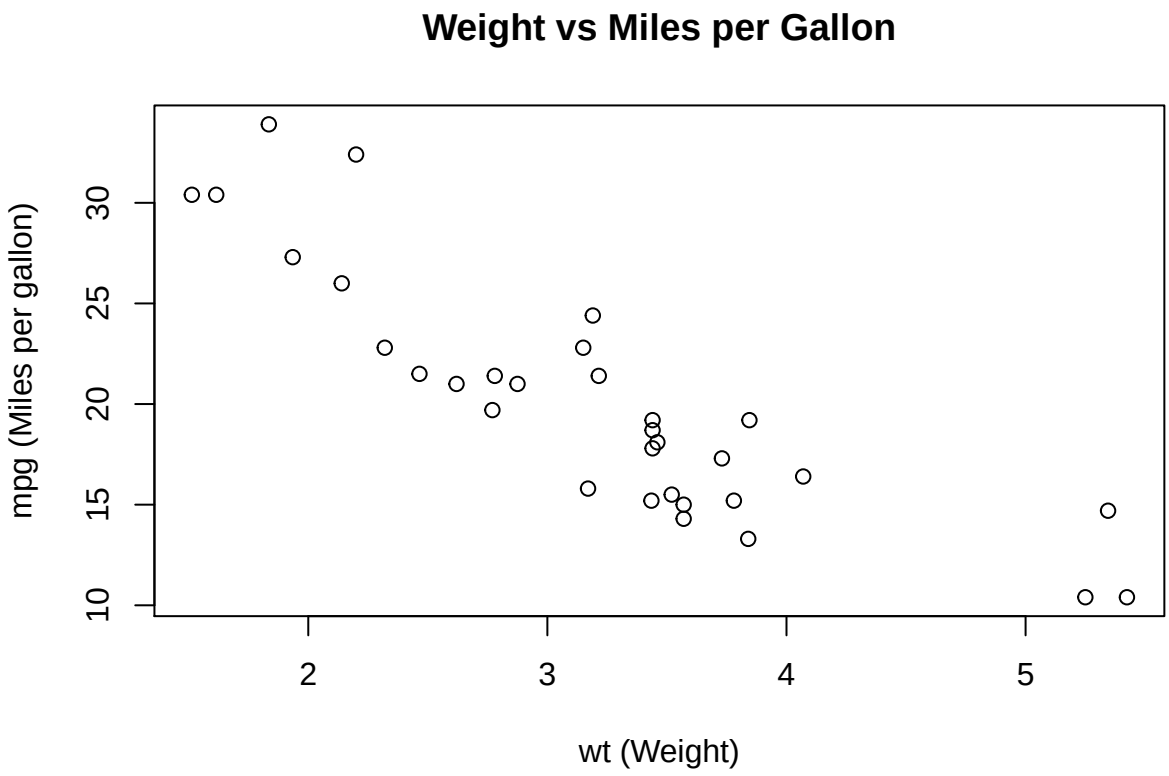
```
## gear  0.32661290
## carb  2.60887097
```

```
Mcor <- cor(mtcars[,2:ncols])
Mcor
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

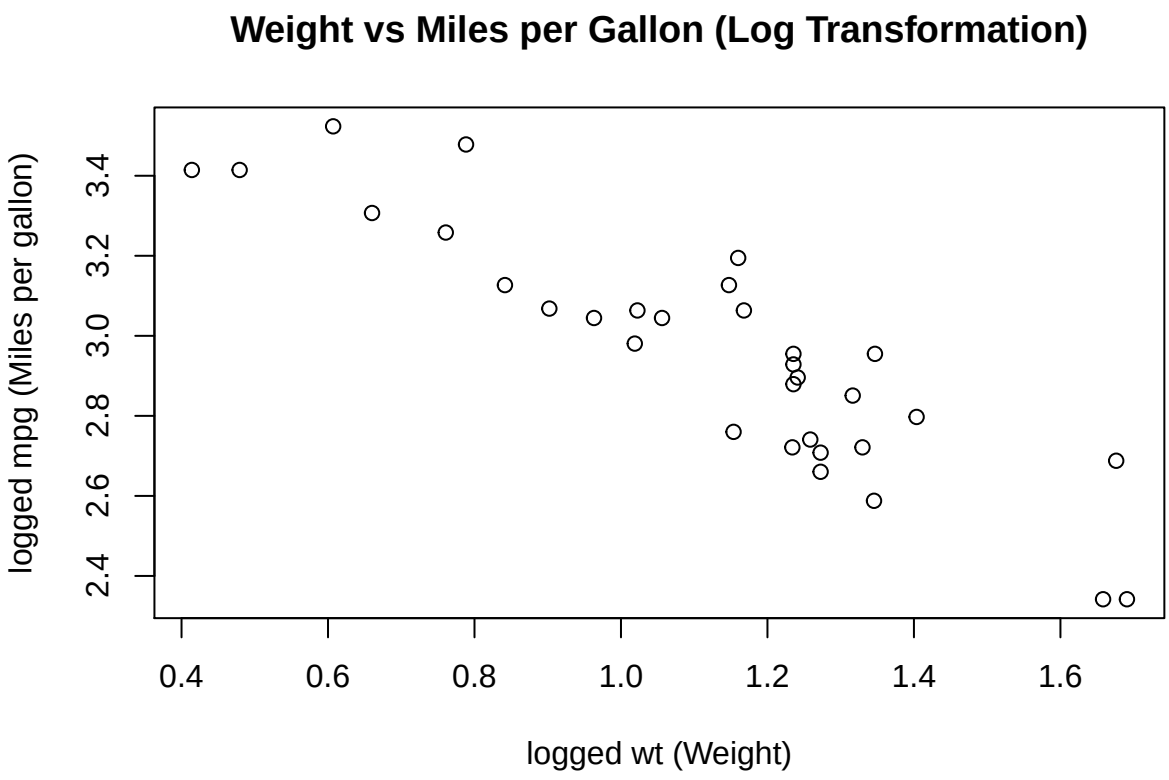
Next, we draw a scatter plot of weights of automobiles and their mileage per gallon. This plot shows a negative association between these two variables. As automobile weight increases, its mileage per gallon decreases. The range of automobile weights is approximately 4, while the range of mileage per gallon is about 24. There are three heavier automobiles in the dataset, and their data points are shown in the lower right corner of the plot.

```
plot(x = mtcars$wt, y = mtcars$mpg,
     xlab = "wt (Weight)", ylab = "mpg (Miles per gallon)",
     main = "Weight vs Miles per Gallon")
```



In addition, we apply log transformation of wt and mpg and create a scatter plot of log-transformed values. This plot shows that the log transformed weights and mileage per gallon have a more linear negative relationship than original weights and mpg. The variance of these transformed data points is smaller than that of original ones.

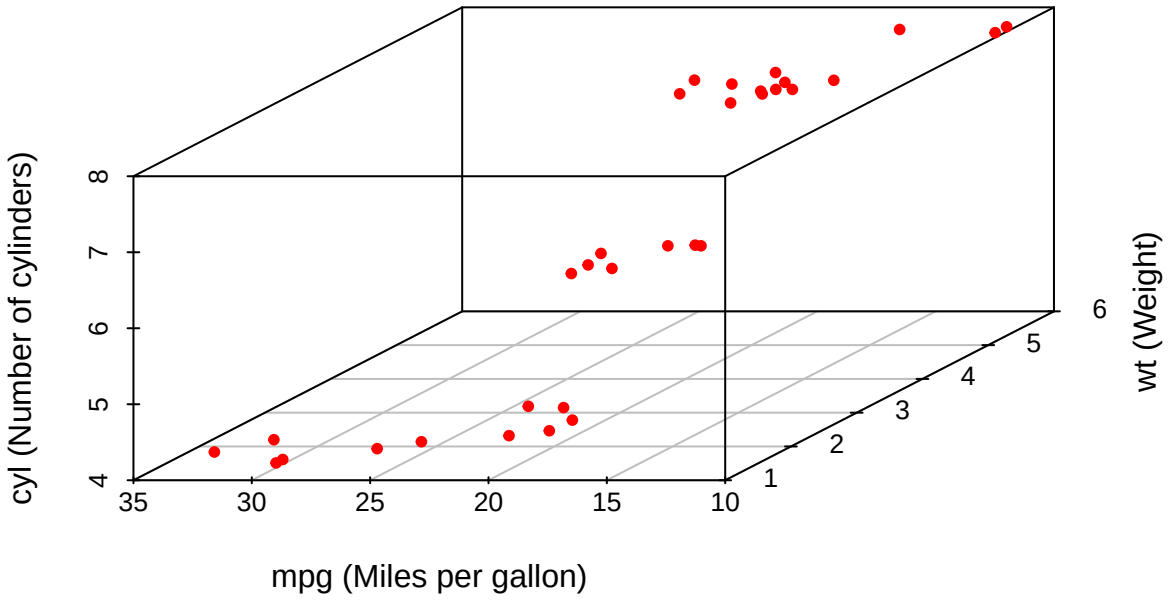
```
x <- log(mtcars[,c("wt","mpg")])
plot(x[,1],x[,2],xlab = "logged wt (Weight)", ylab = "logged mpg (Miles per gallon)",
     main = "Weight vs Miles per Gallon (Log Transformation)")
```



Then, we use a 3D scatter plot to show the relationship between wt (Weight), mpg (Miles per gallon), and cyl (Number of cylinders) of automobiles. When we plot the three variables against each other, data points are split into three distinct clusters based on the number of cylinders. This includes automobiles with 4, 6, and 8 cylinders, which is not surprising since these are the most common number of cylinders found in automobiles. There is also a negative relationship between weight and mpg that is shared by the three clusters, where the greater the weight of an automobile, the lower the mpg. This is the same relationship we find in our previous 2D scatter plots of weight and mpg for all observations.

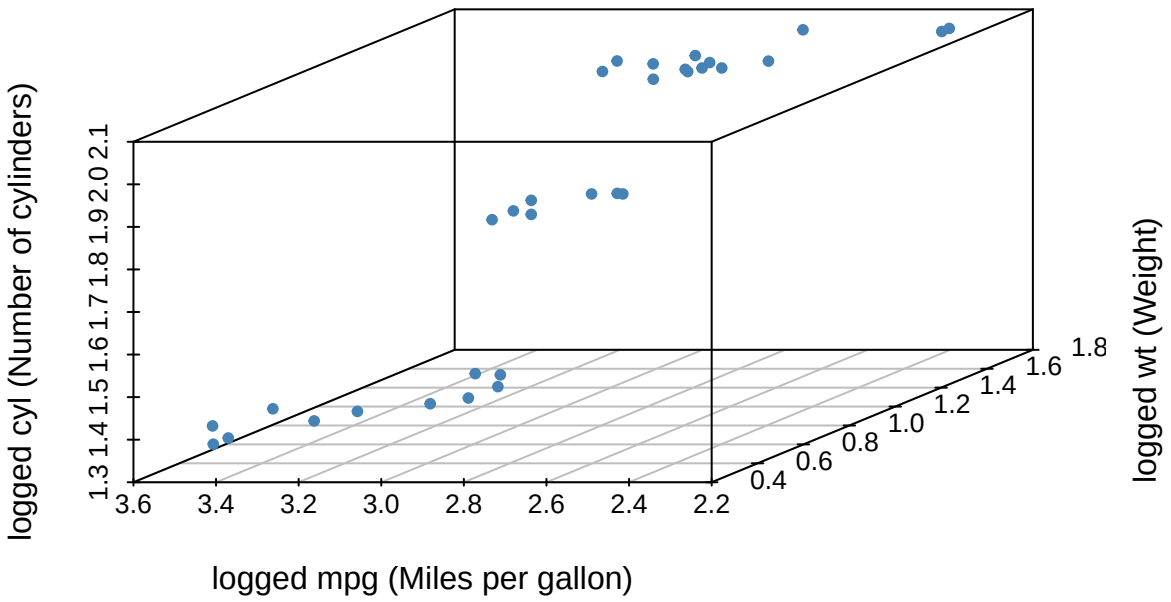
```
library("scatterplot3d")
scatterplot3d(x = mtcars$wt, y = mtcars$mpg, z = mtcars$cyl,
              xlab = "wt (Weight)", ylab = "mpg (Miles per gallon)",
              zlab = "cyl (Number of cylinders)",
              main = "Plot of Automobile Weight, Miles per Gallon, and Number of Cylinders",
              color="red", pch=20, angle=220)
```

Plot of Automobile Weight, Miles per Gallon, and Number of Cylinders



We also create a 3D scatter plot of the log transformations of the three variables to better analyze the data. Similar to the 3D scatter plot of the raw data of the three variables, we see three clusters of data points based on the three different counts of cylinders. There is also a similar negative relationship between the weight and mpg of the automobile.

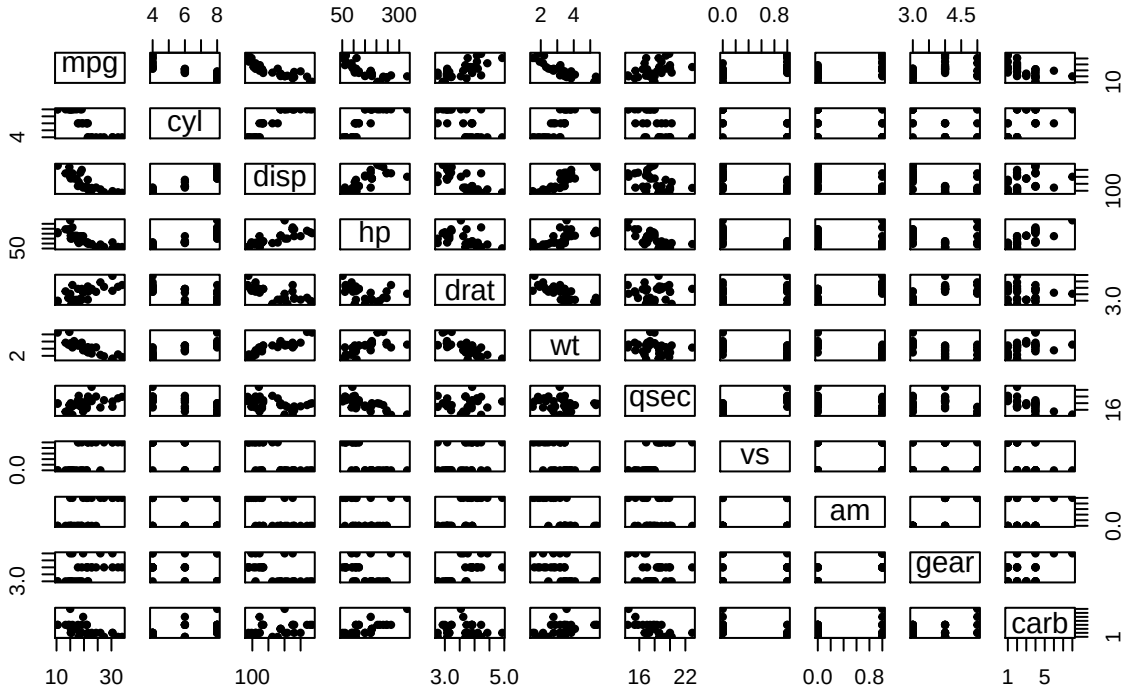
```
scatterplot3d(x = log(mtcars$wt), y = log(mtcars$mpg), z = log(mtcars$cyl),
              xlab = "logged wt (Weight)", ylab = "logged mpg (Miles per gallon)",
              zlab = "logged cyl (Number of cylinders)",
              main = " Plot of Logged Weight, Miles per Gallon, and Number of Cylinders",
              color="steelblue", pch=20, angle=220)
```



Although an engineer suggests that the relationship between the weight and mpg of an automobile is subject to its number of cylinders, our plots suggest that the number of cylinder has little to no effect on the weight and mpg relationship. As we see in both 3D scatterplots, data points are split into three clusters based on the discrete number of cylinders an automobile can have. Despite this, all three clusters in both plots share a similar negative relationship between the weight and mpg of the automobile. Since all three clusters have a similar relationship and do not differ based on the number of cylinders an automobile had, we believe that the relationship between the weight and mpg of an automobile is not subject to its number of cylinders.

Here, we draw pairwise scatter plots for all variables. Since the first column of mtcars dataset contains automobiles' names and they are not quantitative data, we ignore the first column and make pairwise graphs of the other 11 columns. From the matrix graph below, we find that paired scatterplots for disp, hp, drat, gear, and carb are skewed, while plots of mpg vs wt and mpg vs qsec are approximately normal.

```
ncols = dim(mtcars)[2]
pairs(mtcars[,2:ncols], pch=20, main="Scatterplot Matrix of mtcars Dataset")
```



Conclusion:

By performing exploratory analysis of mtcars dataset, we find a negative relationship between automobile weights and mileage per gallon. Applying log transformation of these two variables, we find a more linear negative relationship between these two features, and the transformed data points have less variance and are more normal. When we include the number of cylinders as a feature to create a 3D scatter plot, we observe three clusters in our plots based on the number of cylinders. This is not surprising since automobiles often have either 4, 6, or 8 cylinders. Within those clusters, we also observe a negative relationship between the weight and mpg of an automobile, similar with what we see in the 2D scatter plot. We also include a 3D scatter plot of the log transformation of the three variables. From the plot, we also see very similar clustering and negative relationships between weights and mpg among all three clusters. To explore relationship between all 11 variables, we plot their paired scatter plots. The resulting paired scatterplots for disp, hp, drat, gear, and carb are skewed, while plots of mpg vs wt and mpg vs qsec are approximately normal. In response to the engineer's suggestion that the relation between an automobile's weight and mpg is dependent on the number of cylinders, our 3D scatter plots show that the relationship between weight and mpg is still negative and similar among automobiles with different number of cylinders. This leads us to believe that the relationship between an automobile's weight and mpg is not affected by the number of cylinders the automobile has, and thus we do not accept the engineer's claim.