STATS414_HW1

Ningyuan Wang

9/11/2020

Note: for the limitation of space, some results were hidden in the code but report in the text.

```
gas = read.csv("vwjetta_gas.csv") # read in data

# recode missing values
gas$mileage[gas$mileage==999999] <- NA
gas$vehage[gas$vehage==999] <- NA
gas$mpg[gas$mpg==99] <- NA

# declare categorical variable
gas$congrade = factor(gas$congrade)
levels(gas$congrade) <- c("rough", "average", "clean", "excellent")
```

a.

Missing values on mileage is 5, on vehage is 8, and on mpg is 6.

```
summary(gas)
```

b.

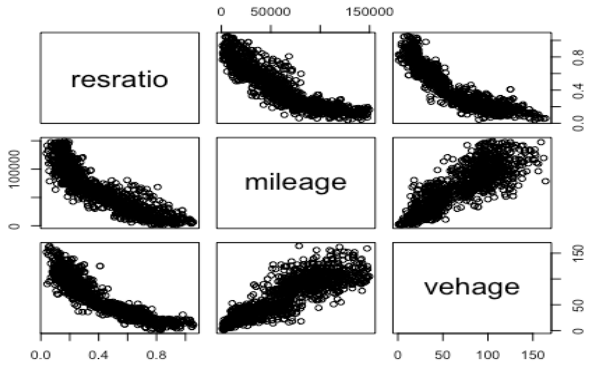
There are 578 rough vehicles, 378 average vehicles, 209 clean vehicles, and 31 excellent vehicles.

```
summary(gas$congrade)
```

```
##     rough   average     clean excellent
##       578       378       209        31
```

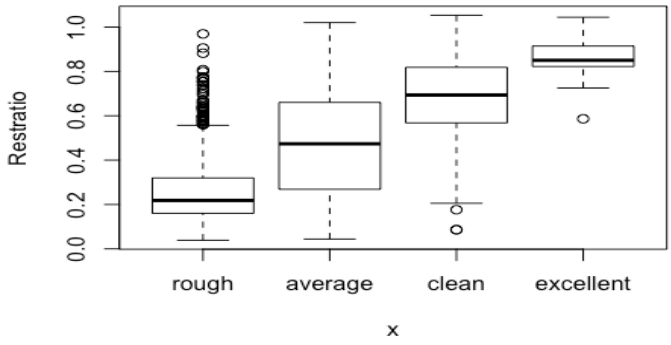
c.

```
pairs(gas[, c("resratio", "mileage", "vehage")])
```



d.

```
plot(gas$congrade, gas$resratio, ylab = "Restratio", main = "")
```

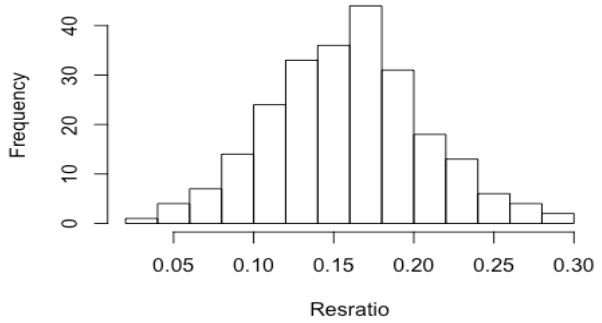


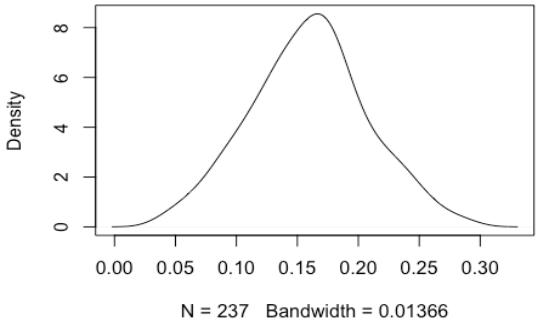

```
tapply(gas$resratio, gas$congrade, mean, na.rm = T)

##     rough   average     clean excellent
## 0.2722137 0.4774244 0.6732804 0.8596422
```

e.

The variable resratio for vehicles with mileage greater than 100,000 has a bell(normal) shape. The mean of resratio for vehicles with mileage greater than 100,000 is 0.16, the corresponding standard deviation is 0.05 (rounded to two digits).

```
hist(gas$resratio[gas$mileage>100000], xlab = "Resratio", main = "")

plot(density(gas$resratio[gas$mileage>100000], na.rm = T), main = "")
```






```r
mean(gas$resratio[gas$mileage>100000], na.rm = T)
sd(gas$resratio[gas$mileage>100000], na.rm = T)
```

f.

For vehicles with average condition grade only, the mean and standard deviation for resratio are 0.48 and 0.23; for mile15 are 3.65 and 2.22; for vehage are 58.18 and 34.79 (rounded to two digits).

```r
gas$mile15 = gas$mileage/15000
gas_avg = gas[gas$congrade=="average", c("resratio", "mile15", "vehage")]

# mean values for three variables
colMeans(gas_avg, na.rm = T)

# standard deviations for three variables
apply(gas_avg, 2, sd, na.rm = T)
```