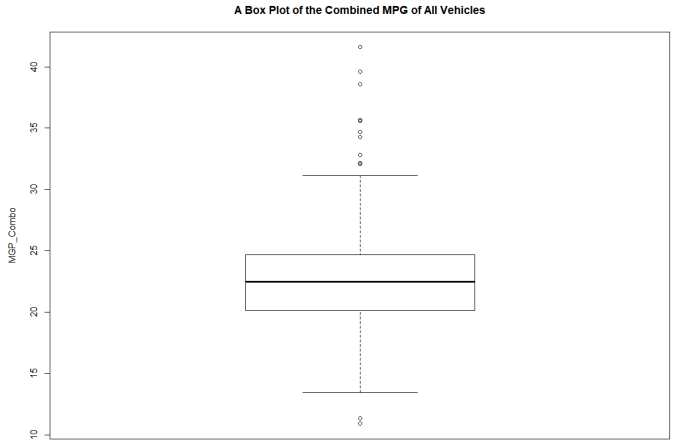
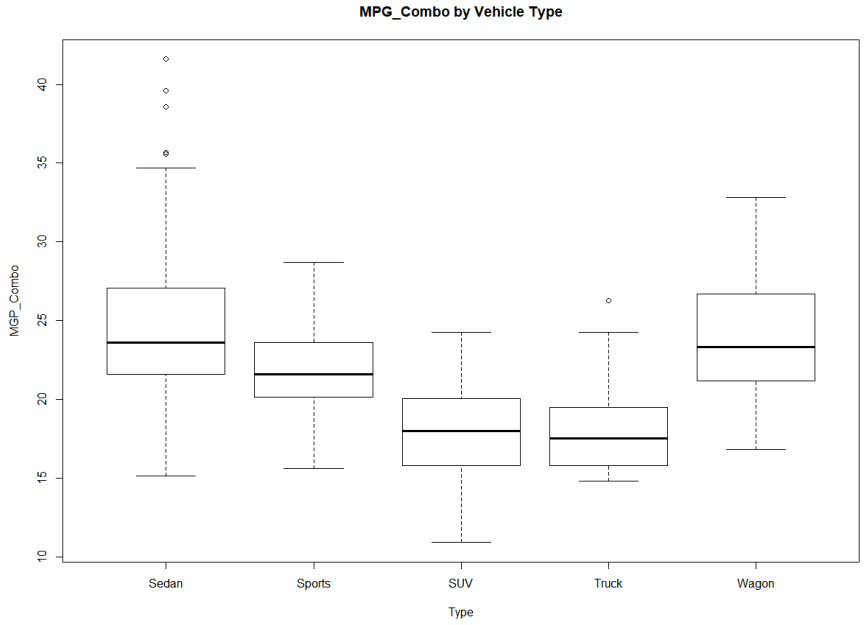
STA 6443 – Algorithms 1 HW1 solution

Exercise 1

(a) The box plot of the combined MPG has several outliers above and below the quartiles. Ignoring those outliers, the mean and median are quite close together (roughly 22 combined miles per gallon but slightly larger values of mean) and the spread of the distribution is not wide. It is slightly right skewed and most vehicles get between 20 and 25 miles per gallon (combined)

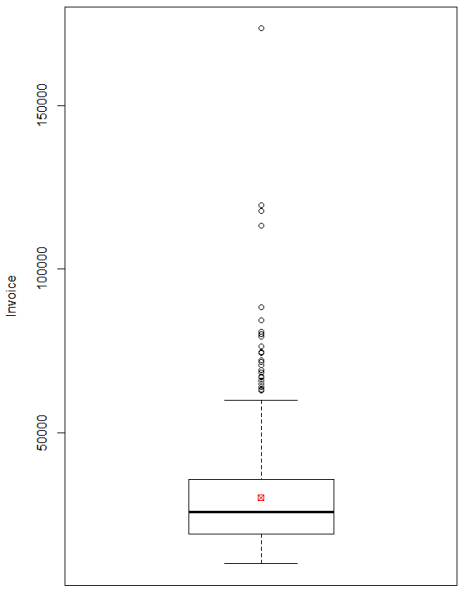


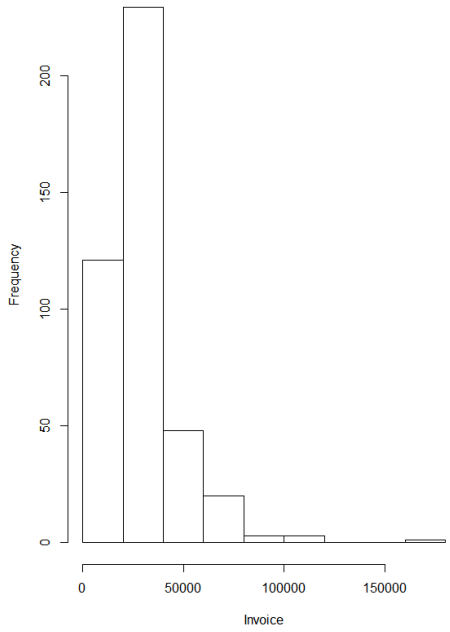
(b) When viewing the distribution of combined MPG separated by vehicle type, we see some interesting things. The combined MPG of SUVs and sports cars appear to be most like normal distributions - seemingly symmetric, with not wide spreads. Sedans' combined MPG have lots of variability and the most outliers. Combined MPG for wagons also has a widespread but may not be symmetric. Trucks appear to be least fuel efficient among the vehicle types. The distribution of combined MPG of trucks is right skewed and has at least one outlier.



(c) The distribution of Invoice is not normal as visible from the qqplot and strongly skewed as visible from the histogram. Some very expensive cars are all European vehicles including the Porsche 911 GT2 and four Mercedes-Benz vehicles - CL500, CL600, SL55 AMG, and SL 600. The mean and median are quite different from each other and it clarifies the asymmetry of the invoice variable's distribution. The qqplot does not follow the straight line and shapiro-Wilk test shows very small p-value, thus Invoice does not follow the Normal distribution.

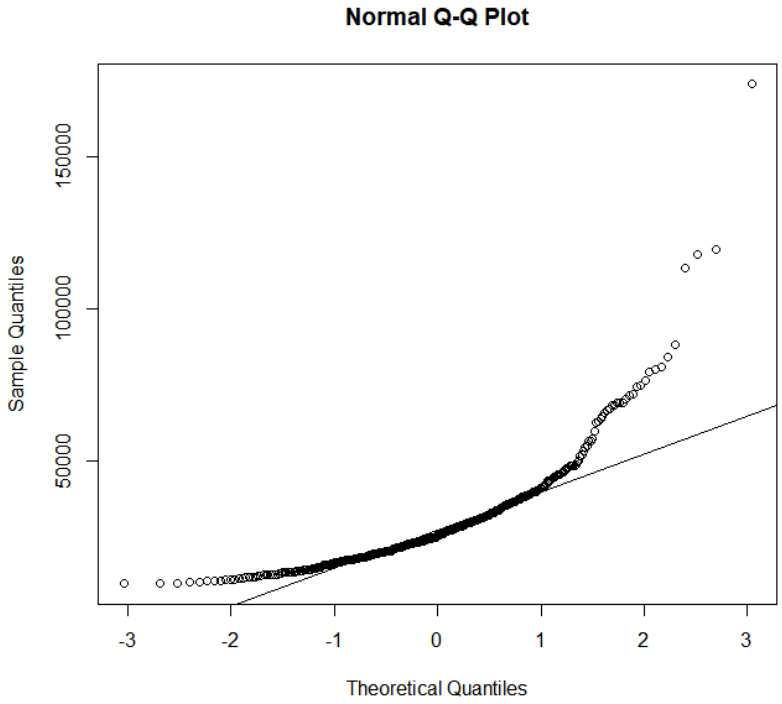
```
> summary(cars$Invoice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9875   18973   25672   30096   35777  173560
> mean(cars$Invoice); var(cars$Invoice)
[1] 30096.48
[1] 312488924
> range(cars$Invoice) # [min,max]
[1]   9875 173560
> skewness(cars$Invoice) # skewness
[1] 2.806001
```





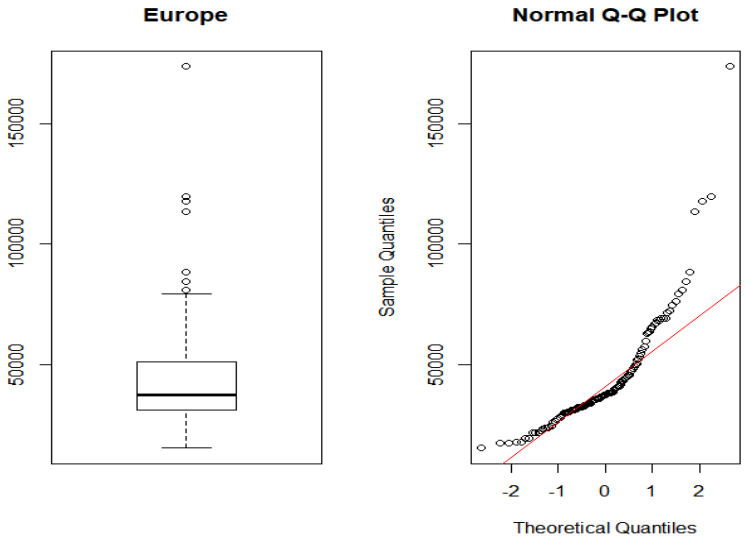

```
        Shapiro-Wilk normality test

data:  cars$Invoice
W = 0.77353, p-value < 2.2e-16
```



(d) There are 155 cars originating from Asia, 123 European cars, and 147 originating from USA in the data set. None of the distributions of Invoice by Origin are normal according to the tests of normality (all p-values are very small, less than 0.05), histograms, and qqplots. The distribution of Invoice (in dollars) is skewed and asymmetric for each of the 3 continents. The respective mean, median, and mode for each of the 3 distributions are quite different from each other.

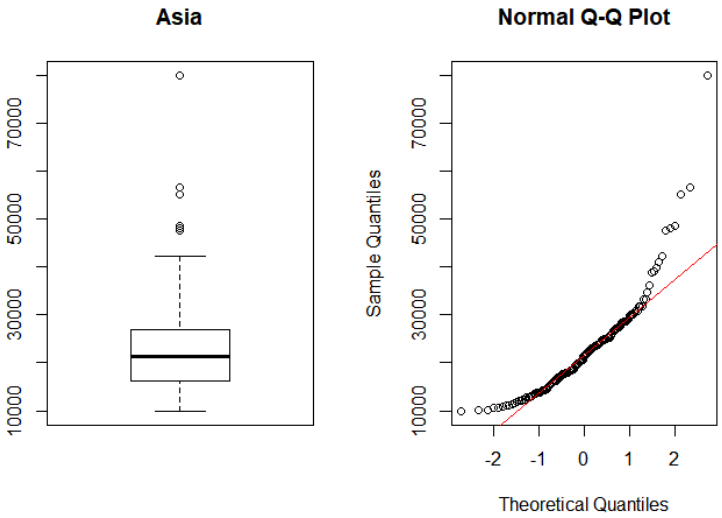
Europe




```
Shapiro-Wilk normality test
```

```
data:  Europe$Invoice
W = 0.79809, p-value = 1.024e-11
```

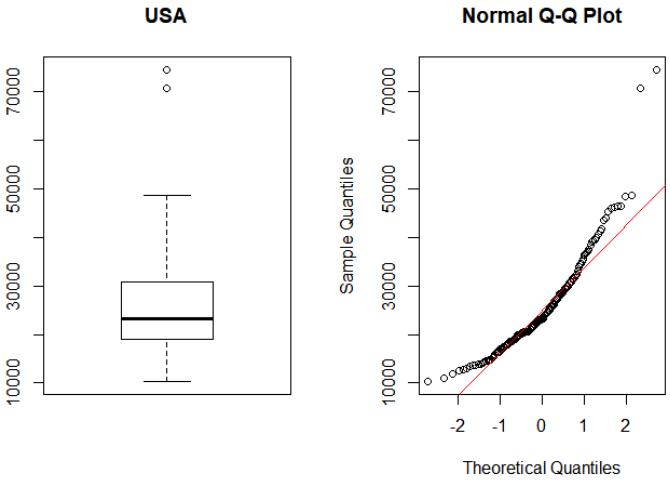
Asian




```
        Shapiro-Wilk normality test

data:  Asia$Invoice
W = 0.84696, p-value = 2.012e-11
```

USA




```
        Shapiro-Wilk normality test

data:  USA$Invoice
W = 0.89222, p-value = 6.42e-09
```

Exercise 2

(a) The distributions of Invoice for Europe and Asia are not Normal from the normality tests above. Thus we should perform nonparametric **Wilcoxon rank-sum test**.

(b) H0: Distributions of Invoice for Europe and Asian cars are from the same distribution
H1: One of the group tends to be more expensive (either Europe or Asia)

(c) We see the Wilcoxon rank-sum test rejects the null hypothesis with very small p-value (less than 0.05). Thus we conclude that one of the cars (either from Europe or Asia) tends to be more expensive. The box plots (in Exercise 1d) yield similar conclusions and it seems that the European cars have higher invoices

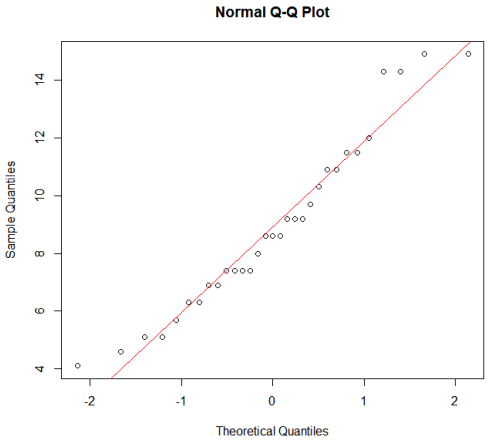
```
Wilcoxon rank sum test with continuity correction

data:  Europe$Invoice and Asia$Invoice
W = 16721, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Exercise 3

(a) The distributions of wind from July and August both follow Normal as we see the almost straight line in qqplots and large p-values (greater than significance level 0.05) on Shapiro-Wilk test. Thus we perform two-sample t-test. Then we check equal variance of two groups through equal variance test and find that two groups have the same variance with large p-values. So we can perform **pooled two-sample t-test**.

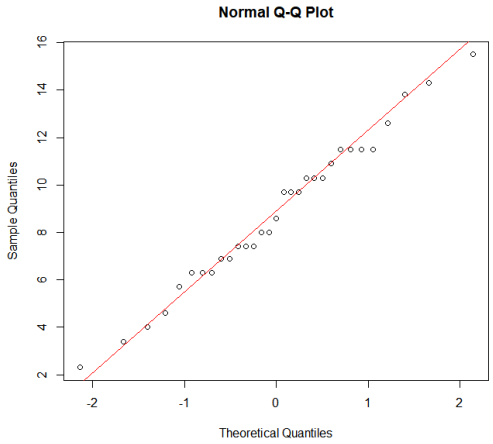
July




```
        Shapiro-Wilk normality test

data:  July$Wind
W = 0.95003, p-value = 0.1564
```

August




```
        Shapiro-Wilk normality test

data:  Aug$Wind
W = 0.98533, p-value = 0.937
```

```
        F test to compare two variances

data:  July$Wind and Aug$Wind
F = 0.8857, num df = 30, denom df = 30, p-value = 0.7418
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4270624 1.8368992
sample estimates:
ratio of variances
         0.8857035
```

(b) H0: mean(Wind of July) = mean(Wind of Aug)
H1: mean(Wind of July) != mean(Wind of Aug)

(c) We see large p-values on pooled two sample t-test and do not have enough evidence to reject the null hypothesis. The mean of Wind from July is equal to the mean of wind from August.

```
> t.test(July$Wind, Aug$Wind, var.equal = TRUE)

        Two Sample t-test

data:  July$Wind and Aug$Wind
t = 0.1865, df = 60, p-value = 0.8527
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.443108  1.739883
sample estimates:
mean of x mean of y
 8.941935  8.793548
```