Homework 1

Rudy Martinez, Jose Fernandez, Brenda Parnin

9/11/2020

Set Working Directory

```
setwd("/Users/rudymartinez/Desktop/MSDA/Fall 2020/STA 6443_Algorithms I/STAT-Algorithms-1/Week 2/HW1") # set your own path
```

Libraries

```
library(tidyverse)
```

Exercise 1

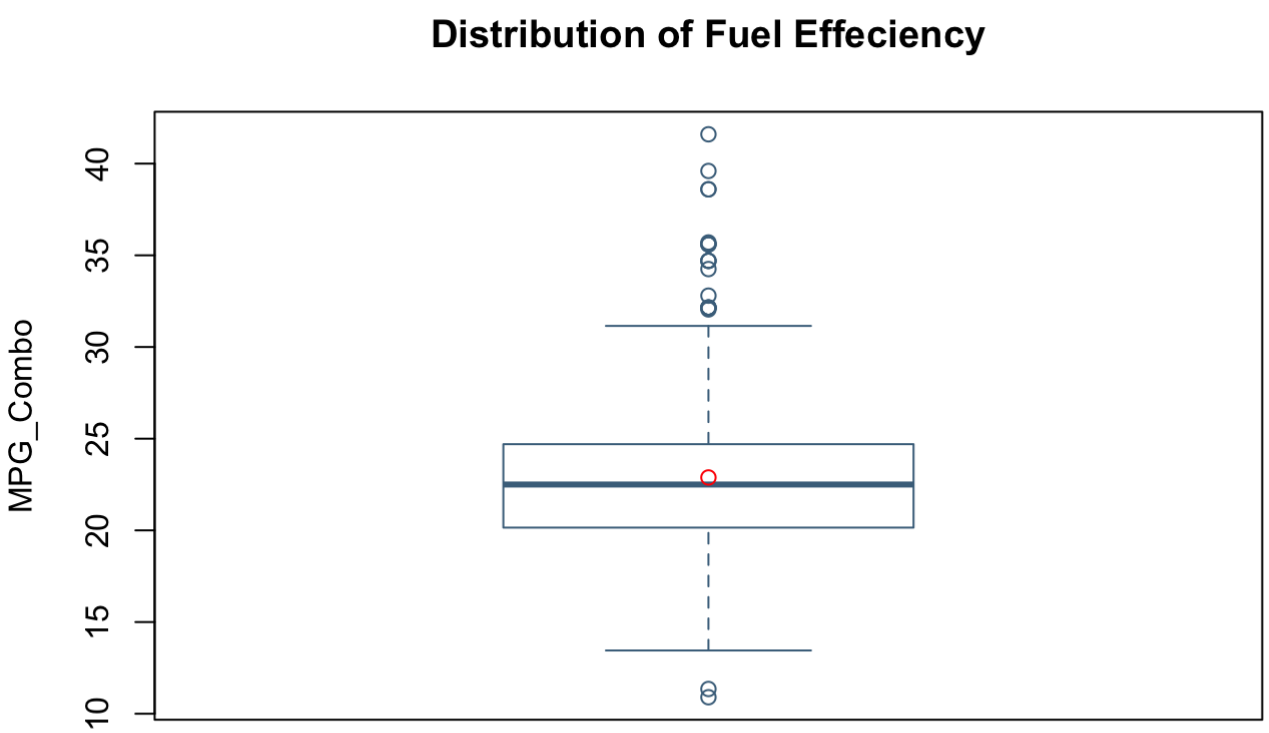
```
cars=read.csv("Cars.csv", header = TRUE)  # read data set
```

(1.a)

Create a combined mpg variable called MPG_Combo which combines 55% of the MPG_City and 45% of the MPG_Highway. Obtain a box plot for MPG_Combo and comment on what the plot tells us about fuel efficiencies.

```
MPG_Combo = 0.55*cars$MPG_City+0.45*cars$MPG_Highway  # combined mpg variable
cars = data.frame(cars, MPG_Combo)   # data frame with MPG_Combo
attach(cars)

boxplot(cars$MPG_Combo,
        main = "Distribution of Fuel Effeciency",
        ylab = "MPG_Combo",
        col = "White",
        border = "skyblue4",
        horizontal = FALSE
)
points(mean(cars$MPG_Combo, na.rm=TRUE), col="red")
```



Observations: The Boxplot above represents the distribution of fuel efficiency for all vehicles in the dataset:

• The Boxplot appears to be fairly symmetrical

• The fact that the mean is greater than the median indicates that the data is slightly skewed to the right

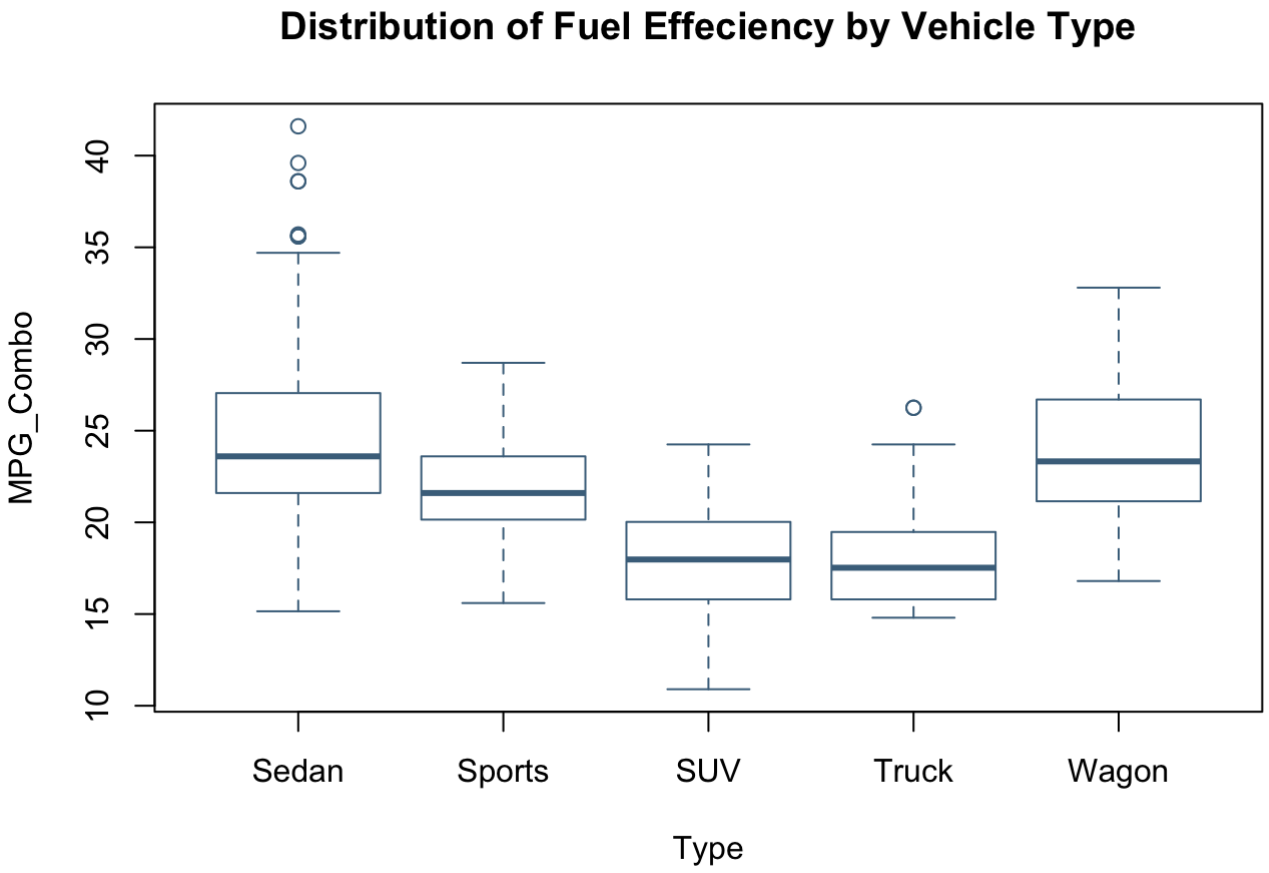
• There is a grouping of outliers positioned above the maximum

• In sum, the mean and median fuel efficiency of all vehicles are positioned in-between the range of 20 – 25 MPG_Combo

(1.b)

Obtain box plots for MPG_Combo by Type and comment on any differences you notice between the different vehicle types combined fuel efficiency.

```
boxplot(MPG_Combo ~ Type, data=cars,
        main = "Distribution of Fuel Effeciency by Vehicle Type",
        xlab = "Type",
        ylab = "MPG_Combo",
        col = "White",
        border = "skyblue4",
        horizontal = FALSE
)
```



Observations: A Boxplot visualization of the 5 vehicle types indicates the following:

• Sports and SUV vehicle types are fairly symmetrical.

• The Sedan, Wagon, and Truck vehicle types appear to be right skewed.

• The Sedan vehicle type data reflects more outliers in comparison to its peers

• In sum, through pure observation, we see that the Sedan offers the best fuel efficiency while the SUV has the worst fuel efficiency.

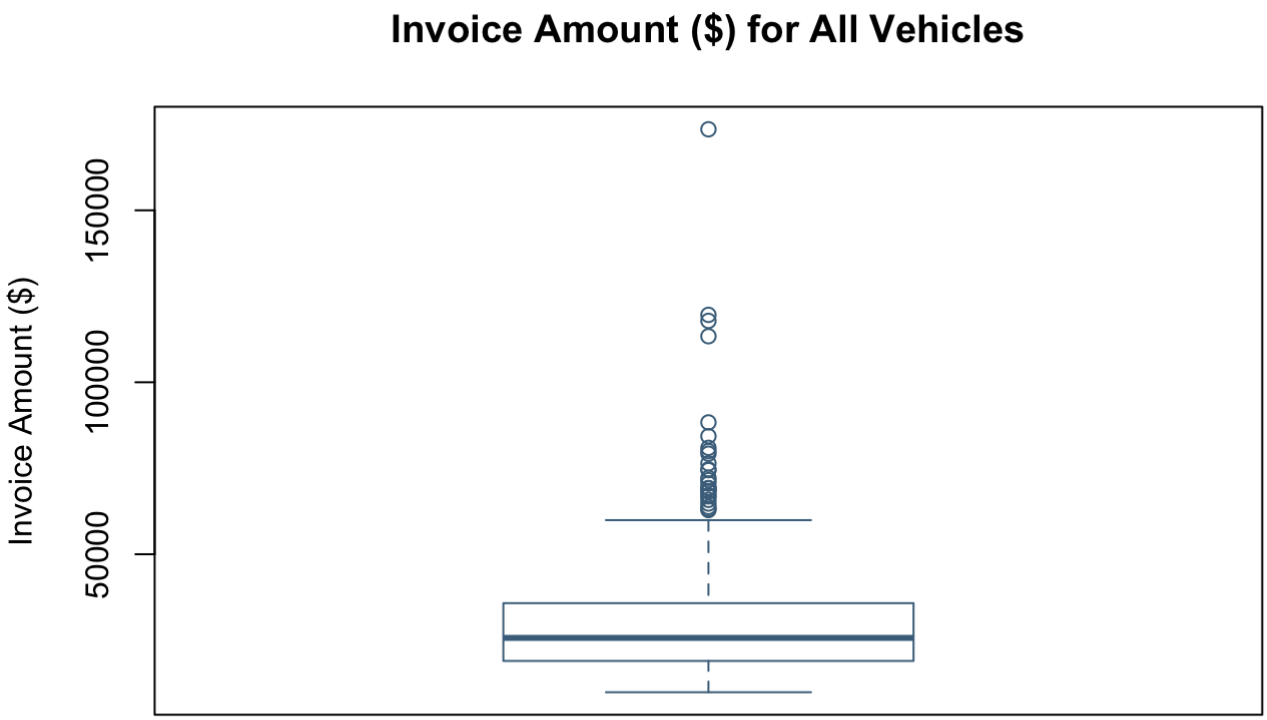
(1.c)

Obtain basic descriptive statistics for Invoice for all vehicles. Comment on any general features and statistics of the data. Use visual and quantitative methods to comment on whether an assumption of Normality would be reasonable for Invoice variable.

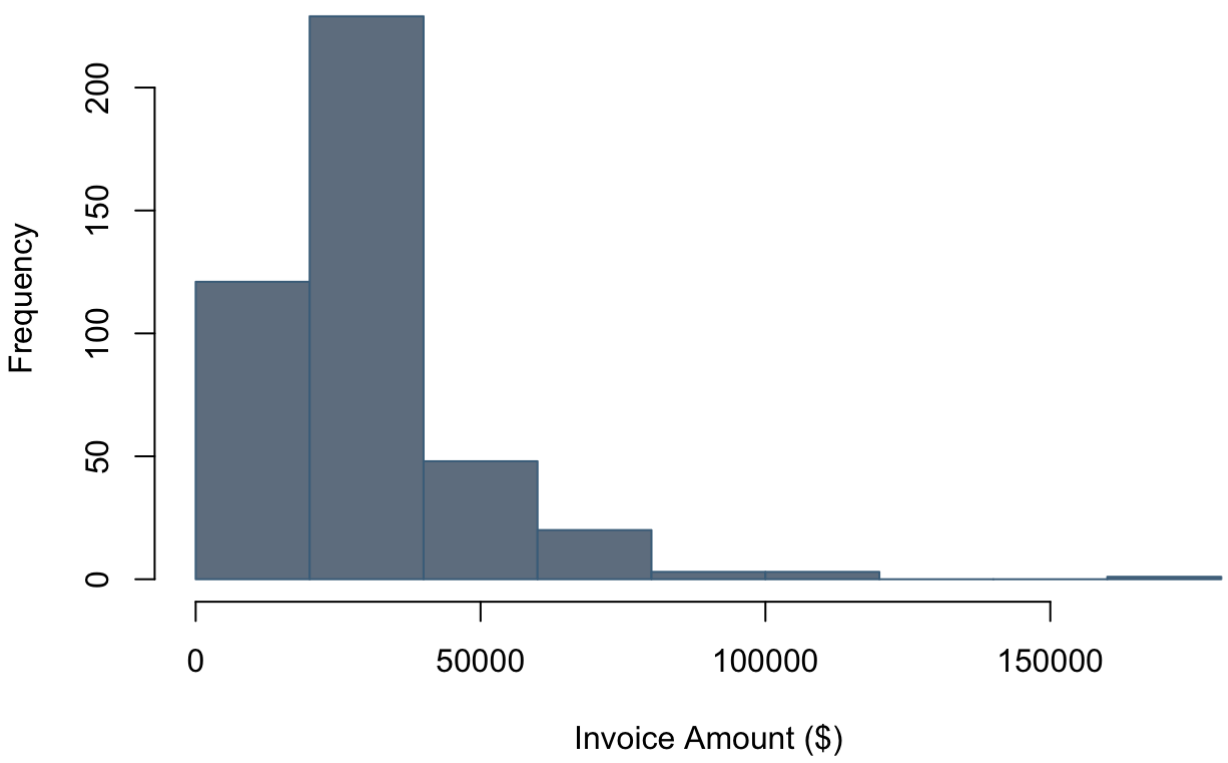
```
summary(cars$Invoice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9875   18973   25672   30096   35777  173560
```

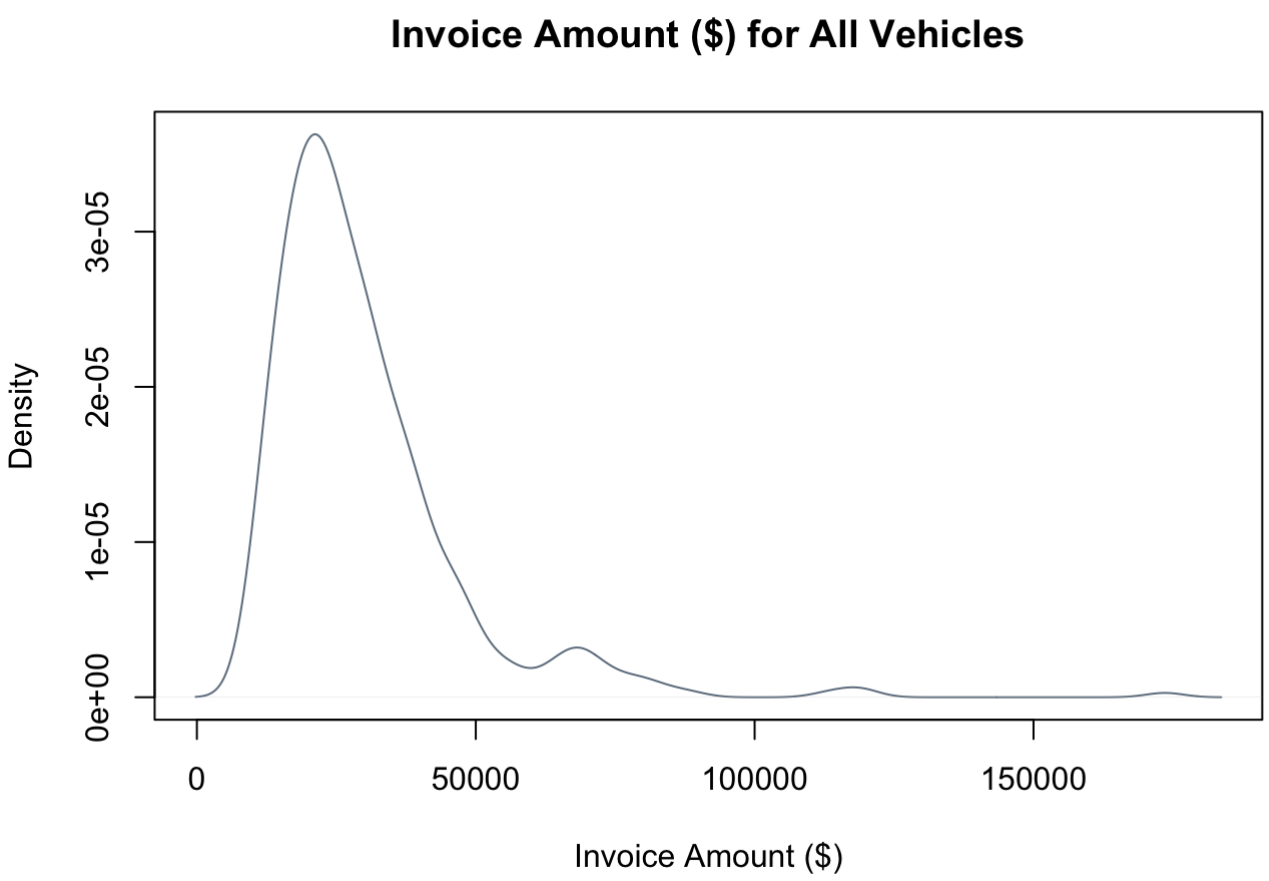
```
boxplot(cars$Invoice,
        main = "Invoice Amount ($) for All Vehicles",
        ylab = "Invoice Amount ($)",
        col = "White",
        border = "skyblue4",
        horizontal = FALSE
)
```



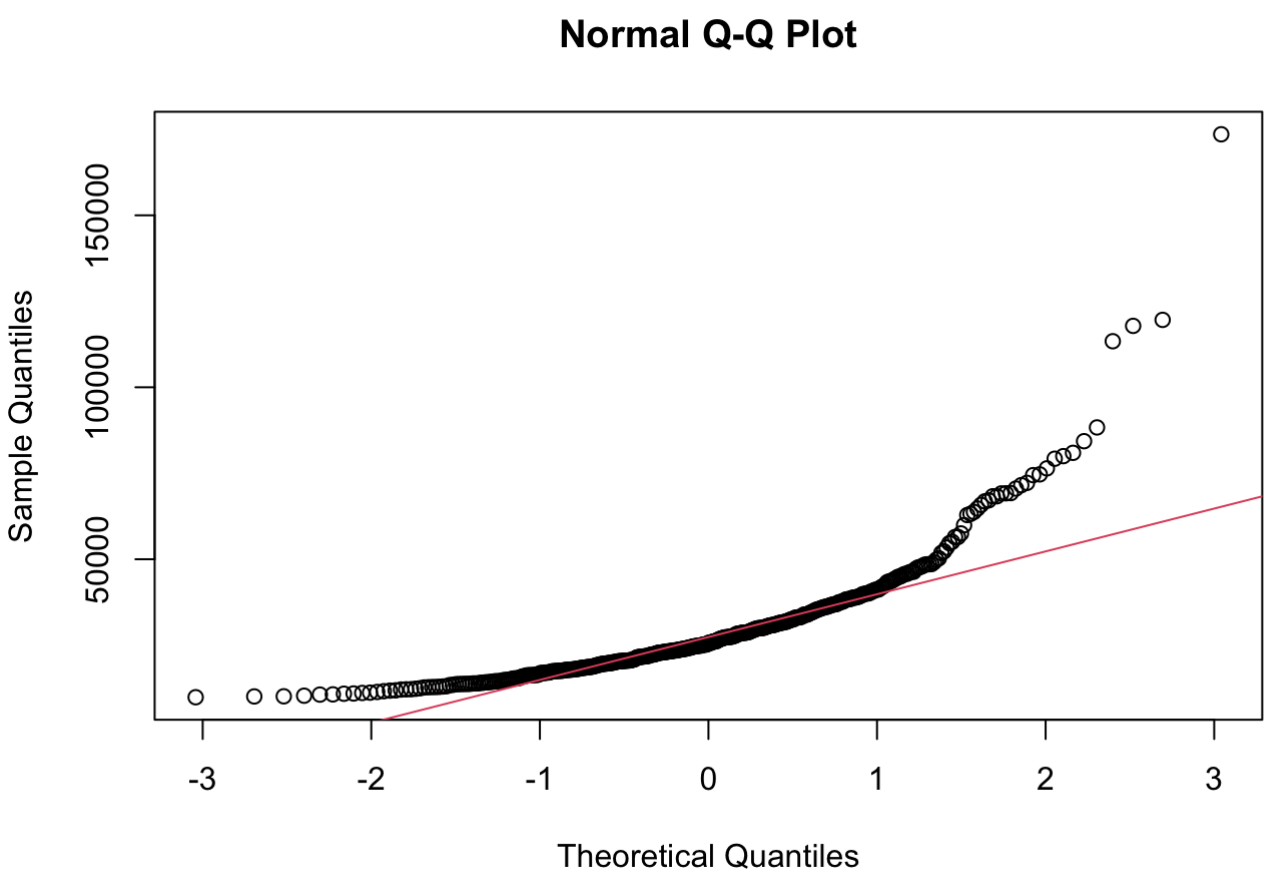

```
hist(cars$Invoice,
     main = "Invoice Amount ($) for All Vehicles ",
     xlab = "Invoice Amount ($)",
     col = "slategrey",
     border = "skyblue4")
```




```
plot(density(cars$Invoice),
     main="Invoice Amount ($) for All Vehicles",
     xlab = "Invoice Amount ($)",
     col = "slategrey")
```




```
qqnorm(cars$Invoice); qqline(cars$Invoice, col = 2)
```




```
shapiro.test(cars$Invoice)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cars$Invoice
## W = 0.77353, p-value < 2.2e-16
```

Observations: The visualizations detailed above indicate the following observations:

• The Invoice Amount ($) for all vehicles Boxplot, Histogram, Density Plot, and Normal QQ Plot visualizations reflect a right skewed distribution.

• This is reinforced by the descriptive statistics generated – a minimum Invoice Amount of $9,875, a median Invoice Amount of $25,672, a mean Invoice Amount of $30,096, and a maximum Invoice Amount of $173,560 – that indicate a long right tail. Additionally, the mean Invoice Amount is greater than the median Invoice Amount.

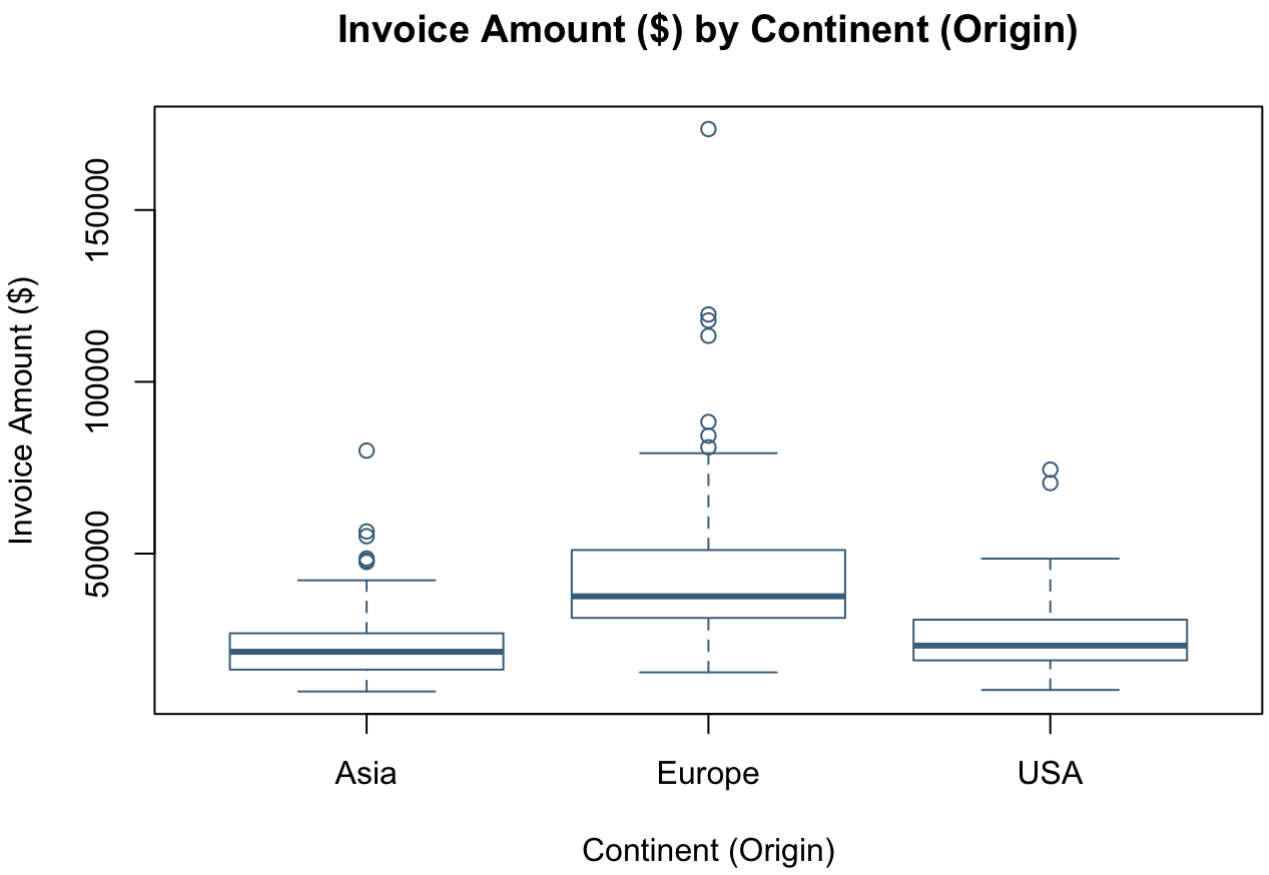
• The Shapiro-Wilk normality test results displayed in the table above show a p-value that is smaller than the significance level of 0.05. Due to this, it indicates that it does not follow normal distribution.

• Thus, our conclusion is that the combination of visual and quantitative methods detailed above indicate that an assumption of Normality is NOT reasonable.

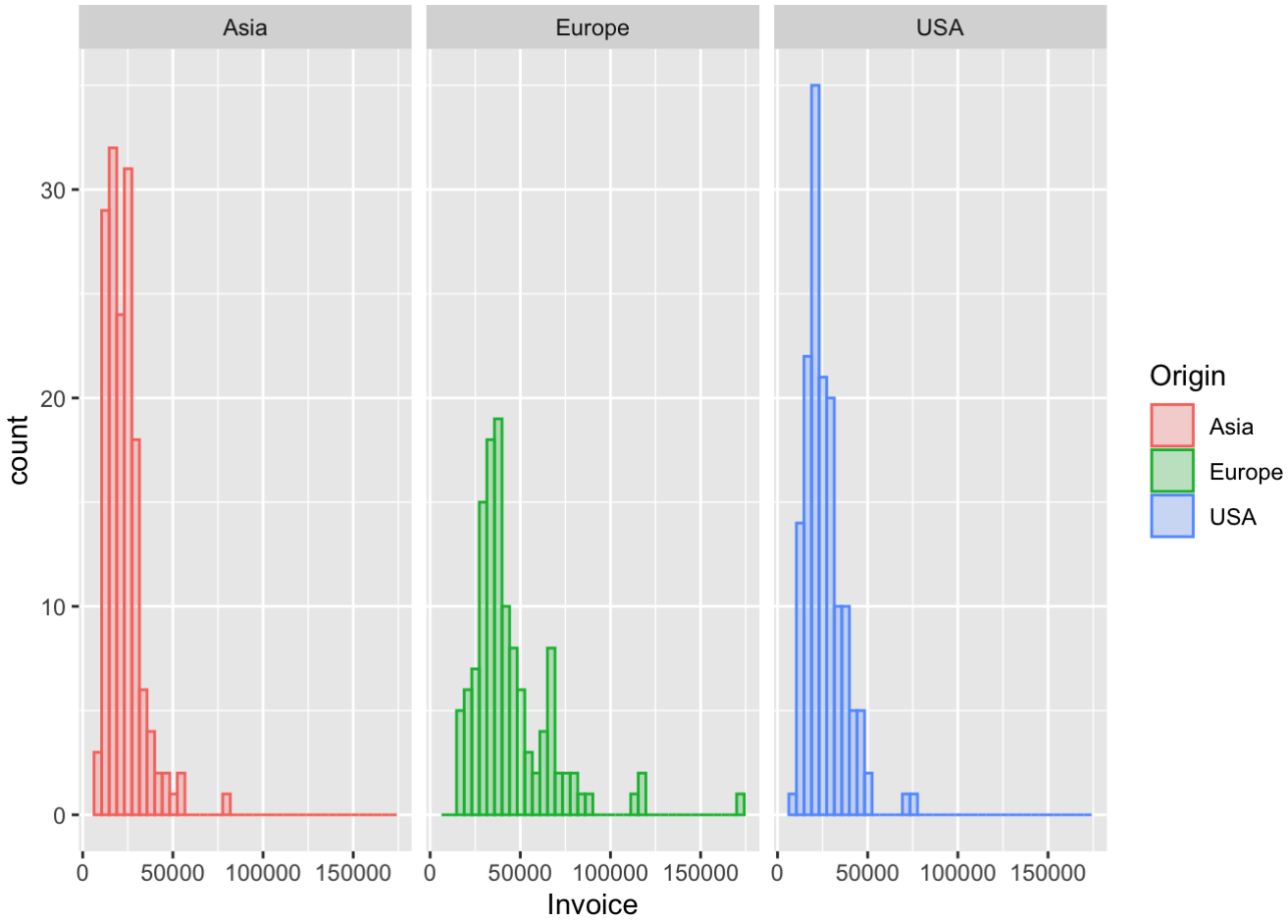
(1.d)

Use visual and quantitative methods to comment on whether an assumption of normality would be reasonable for Invoice variable by Origin. (i.e., check normality of Invoice from i) Europe, ii) Asian, and iii) USA cars.

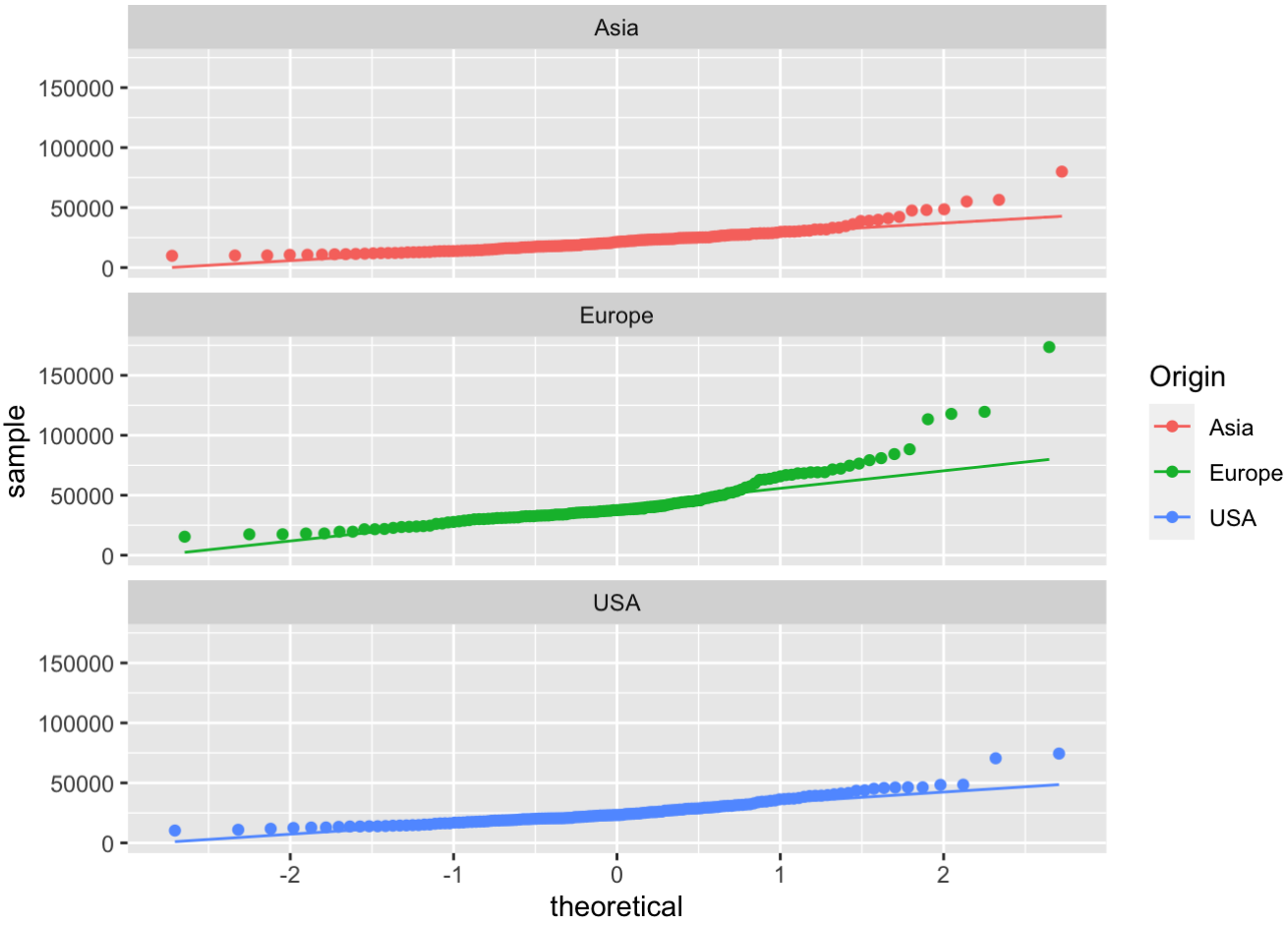
```
boxplot(Invoice ~ Origin, data=cars,
        main = "Invoice Amount ($) by Continent (Origin)",
        xlab = "Continent (Origin)",
        ylab = "Invoice Amount ($)",
        col = "White",
        border = "skyblue4",
        horizontal = FALSE
)
```




```
histogram_plot = ggplot(data=cars, mapping=aes(x=Invoice))+geom_histogram(aes(fill=Origi
n, color=Origin), alpha = 0.25, bins=40) + facet_wrap(Origin~.)
histogram_plot
```




```
qq_plot1 = ggplot(cars, aes(sample = Invoice, col=Origin)) + facet_wrap(~Origin, ncol=1)
qq_plot1 + stat_qq() + stat_qq_line() + theme_gray()
```




```
shapiro.test(cars[cars$Origin=="Asia", "Invoice"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cars[cars$Origin == "Asia", "Invoice"]
## W = 0.84696, p-value = 2.012e-11
```

```
shapiro.test(cars[cars$Origin=="Europe", "Invoice"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cars[cars$Origin == "Europe", "Invoice"]
## W = 0.79809, p-value = 1.024e-11
```

```
shapiro.test(cars[cars$Origin=="USA", "Invoice"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cars[cars$Origin == "USA", "Invoice"]
## W = 0.89222, p-value = 6.42e-09
```

Observations: The visualizations detailed above indicate the following observations:

• The Invoice Amount ($) for vehicles of different Origin (Country) Histogram, Boxplot, and Normal QQ Plot visualizations reflect a right skewed distribution.

- Each Origin's Boxplot is right skewed. This can be determined by the short distance between the lower quartile and the minimum value and the long distance between the upper quartile and the maximum value.
- The Invoice Histogram graph does not look bell shaped, meaning that it is not normally distributed. Each of the histogram graphs have a long right tail.

• The Shapiro-Wilk normality test results displayed in the table above show a p-value that is smaller than the significance level of 0.05 for each Origin. Due to this, it indicates that it does not follow normal distribution.

• Thus, our conclusions is that the combination of visual and quantitative methods detailed above indicate that an assumption of Normality is NOT reasonable for the Invoice variable by Origin.

Exercise 2

(2.a)

Which test should we perform, and why? Justify your answer based on findings on Exercise 1 (d).

Decision: Because the Invoice Amount ($) by Origin (Country) does not follow a normal distribution, the test that should be performed is the two-sample test, specifically the Wilcoxon Rank Sum Test.

(2.b)

Specify null and alternative hypotheses.

Ho: Cars originated in Asia have the same Median Invoice Amount compared to cars originated in Europe (both groups are from the same distribution).

Ha: One group (either Asia or Europe) tends to have larger Invoice Amounts (One group has larger median Invoice Amount than the other group)

(2.c)

State the conclusion based on the test result.

```
asia_europe = filter(cars, Origin == 'Asia' | Origin == 'Europe')

wilcox.test(Invoice ~Origin, data=asia_europe, exact=FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Invoice by Origin
## W = 2344, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Conclusions: Based on the Shapiro-Wilk Test from Exercise 1.d, the results show that both Asia and Europe do not follow normal distribution. Since normal distribution is not followed, we can conduct the Wilcoxon Rank-Sum Test. The Wilcoxon Rank-Sum Test yielded a p-value below the significance level of 0.05, meaning that we can reject the null hypothesis and support the alternative hypothesis of one group having a greater median Invoice Amount than the other group.

Exercise 3

(3.a)

Which test should we perform, and why? See QQ-plot and perform Shapiro-Wilk test for normality check.

Check for Normality

```
shapiro.test(airquality[airquality$Month==7, "Wind"])
```

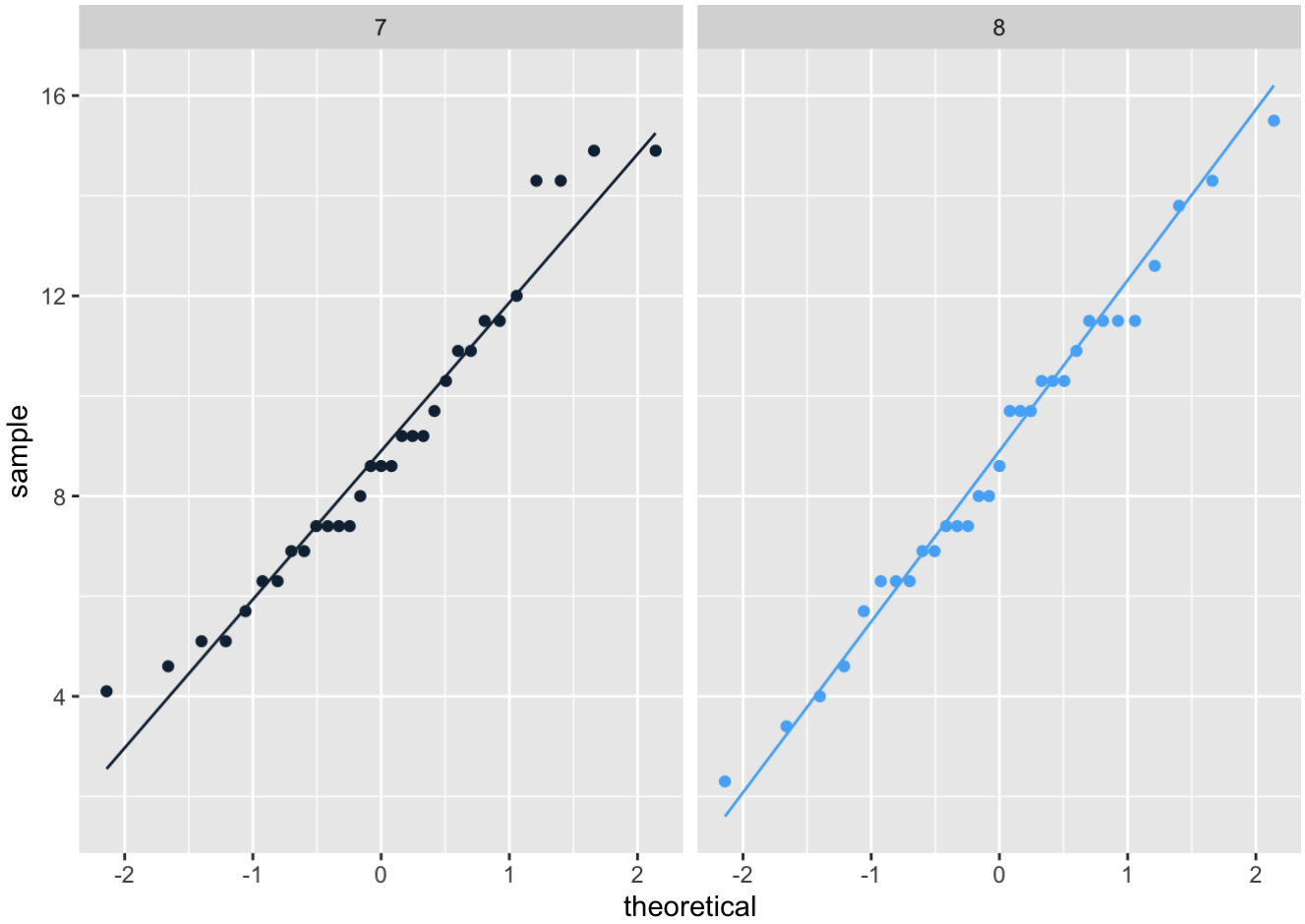
```
##
##  Shapiro-Wilk normality test
##
## data:  airquality[airquality$Month == 7, "Wind"]
## W = 0.95003, p-value = 0.1564
```

```
shapiro.test(airquality[airquality$Month==8, "Wind"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  airquality[airquality$Month == 8, "Wind"]
## W = 0.98533, p-value = 0.937
```

```
july_august = filter(airquality, Month == 7 | Month == 8)

qq_plot2 = ggplot(july_august, aes(sample = Wind, col=Month)) + facet_wrap(~Month, ncol=
2)
qq_plot2 + stat_qq() + stat_qq_line() + theme_gray() + theme(legend.position = "none")
```



• Through the Shapiro-Wilk Test, we determined that both the July and August Wind speed data follows a normal distribution per their respective p-values that exceed the significance level of 0.05.

• This is reinforced through the Q-Q Plot visualization above that shows how well the July and August Wind speed data distribution fits the theoretical distribution.

• Based on these findings (normal distribution), the next step is to perform an equal variance test.

Check for Equal Variance

```
var.test(Wind ~Month, july_august, alternative ="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  Wind by Month
## 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
```

• After conducting an F test (var.test), the results indicated a p-value of 0.7418. This high p-value means we can't reject the null hypothesis, indicating that the two groups have an equal variance.

• Based on these findings (equal variance), the next step is to conduct a Pooled T-test

Conduct Pooled T-test

```
t.test(Wind ~Month, july_august,alternative ="two.sided", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Wind by Month
## 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 in group 7 mean in group 8
##        8.941935        8.793548
```

(3.b)

Specify null and alternative hypotheses

Ho: Mean of Wind speed in July is equivalent to the Mean wind speed in August

Ha: Mean of Wind speed in July is not equivalent to the Mean of Wind speed in August

(3.c)

State the conclusion based on the test result.

Conclusions/Rationale:

• Through the Shapiro-Wilk Test, we determined that both the July and August Wind speed data follows a normal distribution per their respective p-values that exceed the significance level of 0.05.

• This is reinforced through the Q-Q Plot visualization above that shows how well the July and August Wind speed data distribution fits the theoretical distribution.

• After conducting an F test (var.test), the results indicated a p-value of 0.7418. This high p-value means we can't reject the null hypothesis, indicating that the two groups have an equal variance.

• Because our July and August Wind speed data follows a normal distribution and has an equal variance, we must perform a two-sample test, specifically a Pooled t-test.

• **After performing a Pooled t-test, we determined that the Mean of Wind speed for July and August are equivalent. Thus, we can't reject the null hypothesis.**