Car Analysis

Manan Bhatia

2022-05-23

Question 1.1:

Test if the mean number of cylinder is different between Mazda and Isuzu vehicles

Since the question asks specifically if there is any different in the mean number of cylinders between Mazda and Isuzu.

To find the evidence of differences in the mean number of cylinders between Mazda and Isuzu, I will specifically use the t.test techniques with the method: alternative = "two.sided" and compute a hypothesis test, as I only want to find if there is any difference in means between two pairs of data. If the result gives out a p-value of less than 0.05, it means there is evidence of differences between Mazda and Isuzu's mean number of cylinders.

```
vehicles= read.csv("light_vehicles.csv")
head(vehicles)
```

```
##   Year   Make Colour    Fuel.type Number.of.cylinders Number.of.seats
## 1 2003  Mazda    Red       Petrol                   7              11
## 2 2012 Suziki  Beige       Diesel                   7               1
## 3 2020    Kia   Blue     Electric                   4               9
## 4 2012 Toyota   Grey       Diesel                   3               9
## 5 2006  Isuzu Silver Petrol - Gas                  12              13
## 6 2014   Ford  Black       Diesel                   2               2
##   GVM.weight Tare.weight
## 1       6340        4380
## 2       2040        3996
## 3       2471        1627
## 4       3568        5080
## 5       6989        4755
## 6       6405        2516
```

H0: No difference between number of cylinders of both Mazda and Isuzu vehicles HA: A difference between number of cylinders of both Mazda and Isuzu vehicles

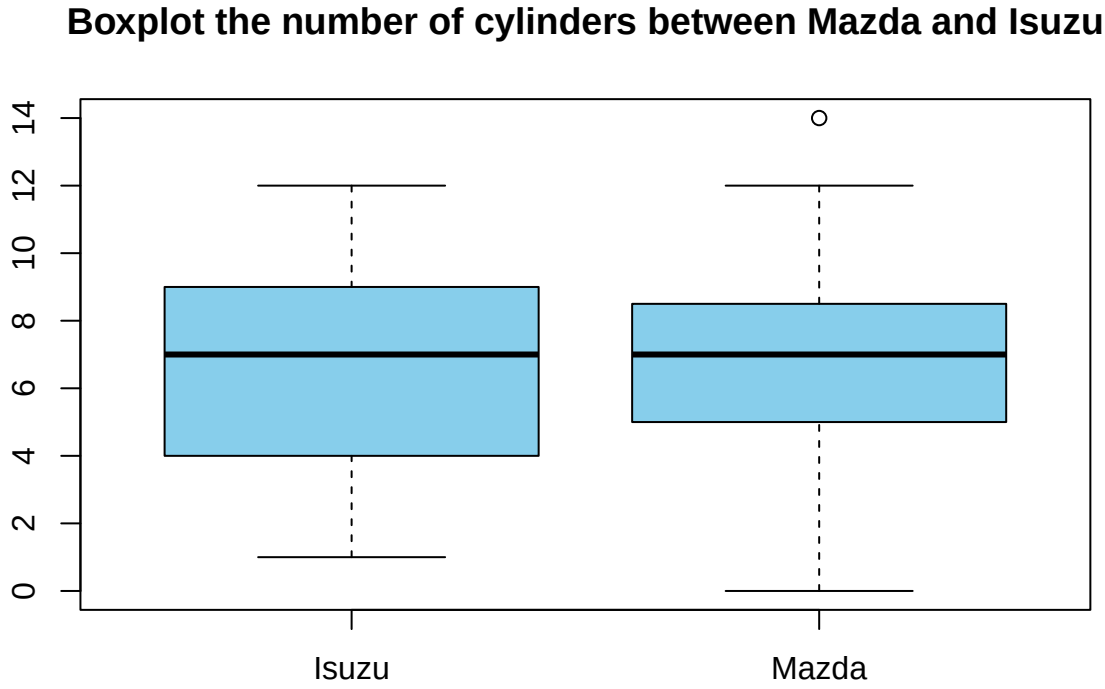
```
vehicles.Isuzu <- subset(vehicles, Make=='Isuzu', Number.of.cylinders, drop=TRUE)
mean(vehicles.Isuzu)
```

```
## [1] 6.52381
```

```r
vehicles.Mazda <- subset(vehicles, Make=='Mazda', Number.of.cylinders, drop=TRUE)
mean(vehicles.Mazda)
```

```
## [1] 6.5625
```

```r
label<-c("Isuzu", "Mazda")
boxplot(vehicles.Isuzu, vehicles.Mazda, names=label, col="skyblue", main="Boxplot the number of cylinde
```



- The box plot displays an overview of all number of cylinders of Mazda and Isuzu from the max value to the min value, the mean, the range, the quartile and quartile data.
- This graphing techniques on two groups of value allows viewers to have a better overview of the overall data within the two groups.

Then the t.test is proceeded on the two extracted data of Number of cylinders from Mazda and Isuzu, to find the probability of the occurrence of any differences in the number of cylinders. if p is smaller than 0.05, then it will reject the null hypothesis and accept that there is an evidence of difference.

```r
vehicles.t.test <- t.test(vehicles.Isuzu,vehicles.Mazda, var.equal=TRUE, alternative="two.sided")
vehicles.t.test$p.value
```

```
## [1] 0.9512909
```

P value is more than 5%, The null hypothesis test is not rejected as there is a very slight difference between Mazda and Isuzu ####

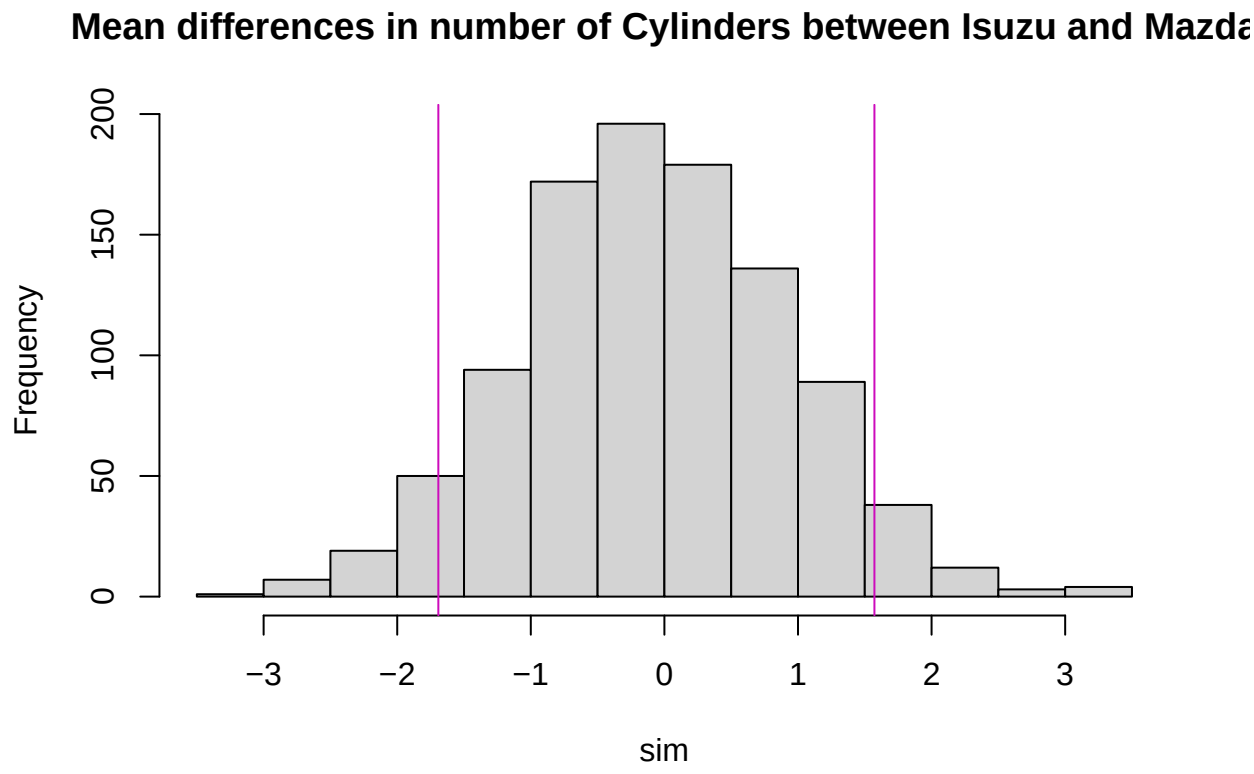
```r
sim= replicate(1000, {
  Isuzu.resamp= sample(vehicles.Isuzu, replace=TRUE)
  Mazda.resamp= sample(vehicles.Mazda, replace=TRUE)
  t.test(Isuzu.resamp, Mazda.resamp, var.equal=TRUE)$statistic
})
hist(sim, breaks=20, main= "Mean differences in number of Cylinders between Isuzu and Mazda")
quantile(sim, c(0.05, 0.95))
```

```
##        5%       95%
## -1.692271  1.572108
```

```r
quantile (sim, 0.50)
```

```
##         50%
## -0.08108335
```

```r
abline(v=quantile(sim, c(0.05, 0.95)), col=6)
```



- In graphing techniques, I used the distribution histogram to represent the data because, the data is replicated 1000 times. Therefore, the distribution histogram can attribute the occurrence frequency of data in a most observable way when replicating

- In fact, from the graph alone, we can see the mean differences in number of cylinders between Isuzu and Mazda is already centralised at approximate -0.0810834.This means Mazda has almost 4 more numbers of cylinders than Isuzu. And from the function quantile(sim, c(0.05, 0.95)), We are 95% confident that the mean difference in number of Cylinders between Isuzu and Mazda is approximately between -1.6922714 to 1.5721081.

Question 1.2:

Test if the mean number of seats is different for each colour. If so, determine which colour has a statistically different mean

The question is asking for evidences of different of the mean number of seats for each colour and then shows which colour is statistically different. Therefore, in this example, I will use oneway test and then TukeyHSD test to find the different for each colour. As one way test is specifically used for finding any significant differences in means between two or more groups, we set the threshold of 0.05, if p-value is less than 0.05, meaning there is a significant differences between means of two groups or more. Meanwhile, TukeyHSD test enables users to see the differences in means, the p adjacent between every possible paired groups. If p-value is less than 0.05, there is a different between means of that paired groups.

```
F.test <- oneway.test(vehicles$Number.of.seats~vehicles$Colour, data=vehicles)
F.test #p-value is small, therefore F.test is big
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  vehicles$Number.of.seats and vehicles$Colour
## F = 3.4104, num df = 8.00, denom df = 191.75, p-value = 0.001092
```

```
tu <- aov(vehicles$Number.of.seats~vehicles$Colour, data=vehicles)
tu
```

```
## Call:
##    aov(formula = vehicles$Number.of.seats ~ vehicles$Colour, data = vehicles)
##
## Terms:
##                 vehicles$Colour Residuals
## Sum of Squares          474.241  7628.007
## Deg. of Freedom               8       491
##
## Residual standard error: 3.94153
## Estimated effects may be unbalanced
```

- Since we want an anova table to proceed to the next step of using TukeyHSD test, we will call the function aov().

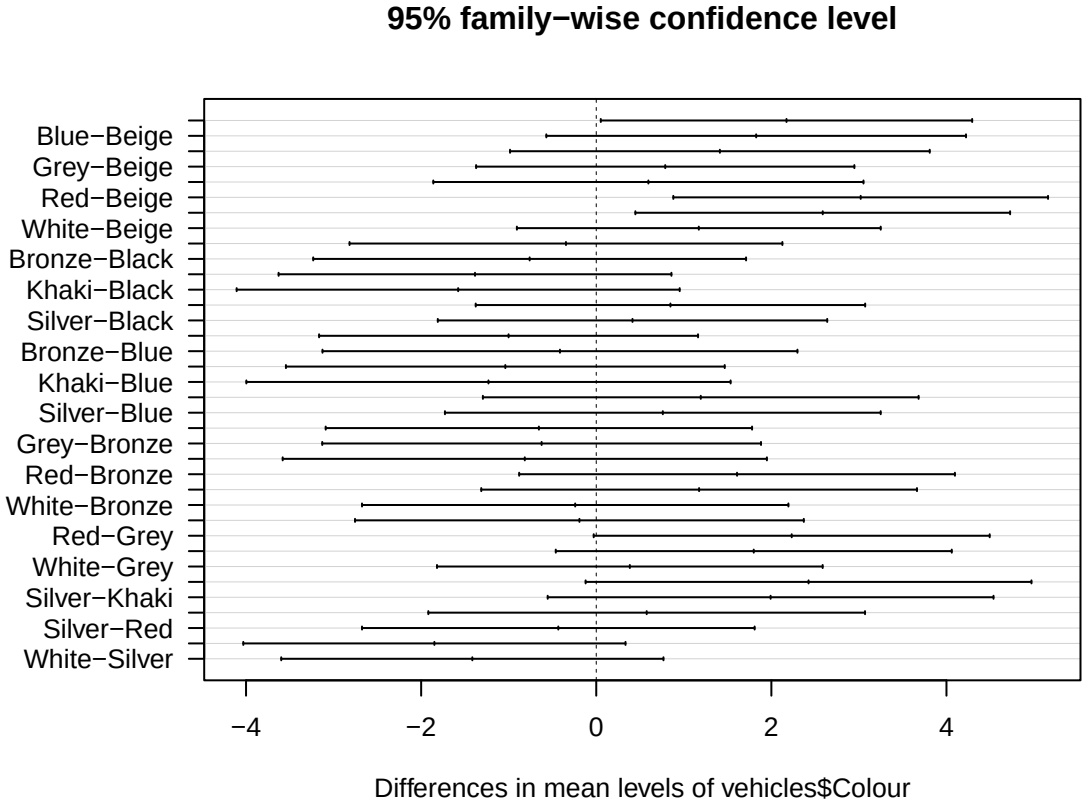
```
summary(tu)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## vehicles$Colour   8    474   59.28   3.816 0.00023 ***
## Residuals       491   7628   15.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tuke.test<-TukeyHSD(tu)
tuke.test
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = vehicles$Number.of.seats ~ vehicles$Colour, data = vehicles)
##
## $`vehicles$Colour`
##                      diff         lwr       upr     p adj
## Black-Beige     2.1727795  0.05184105 4.2937179 0.0398643
## Blue-Beige      1.8265954 -0.57012584 4.2233166 0.3003830
## Bronze-Beige    1.4119612 -0.98475999 3.8086825 0.6583956
## Grey-Beige      0.7879074 -1.37221656 2.9480314 0.9684098
## Khaki-Beige     0.5955299 -1.86102635 3.0520862 0.9978962
## Red-Beige       3.0200913  0.88012361 5.1600590 0.0004554
## Silver-Beige    2.5867580  0.44679028 4.7267257 0.0057664
## White-Beige     1.1713351 -0.90636873 3.2490389 0.7107926
## Blue-Black     -0.3461841 -2.81818883 2.1258206 0.9999646
## Bronze-Black   -0.7608183 -3.23282297 1.7111865 0.9892073
## Grey-Black     -1.3848721 -3.62823345 0.8584893 0.5978045
## Khaki-Black    -1.5772496 -4.10730953 0.9528104 0.5845492
## Red-Black       0.8473118 -1.37664781 3.0712715 0.9590267
## Silver-Black    0.4139785 -1.80998115 2.6379381 0.9996933
## White-Black    -1.0014444 -3.16555804 1.1626693 0.8810472
## Bronze-Blue    -0.4146341 -3.12695789 2.2976896 0.9999306
## Grey-Blue      -1.0386880 -3.54439406 1.4670181 0.9336022
## Khaki-Blue     -1.2310655 -3.99640381 1.5342729 0.9024787
## Red-Blue        1.1934959 -1.29485477 3.6818466 0.8579877
## Silver-Blue     0.7601626 -1.72818810 3.2485133 0.9897240
## White-Blue     -0.6552603 -3.09027171 1.7797511 0.9956334
## Grey-Bronze    -0.6240538 -3.12975992 1.8816523 0.9974543
## Khaki-Bronze   -0.8164313 -3.58176966 1.9489070 0.9918069
## Red-Bronze      1.6081301 -0.88022062 4.0964808 0.5345902
## Silver-Bronze   1.1747967 -1.31355396 3.6631475 0.8685480
## White-Bronze   -0.2406261 -2.67563757 2.1943853 0.9999976
## Khaki-Grey     -0.1923775 -2.75537556 2.3706206 0.9999997
## Red-Grey        2.2321839 -0.02917678 4.4935446 0.0561834
## Silver-Grey     1.7988506 -0.46251011 4.0602113 0.2450598
## White-Grey      0.3834277 -1.81910349 2.5859589 0.9998143
## Red-Khaki       2.4245614 -0.12147184 4.9705946 0.0761485
## Silver-Khaki    1.9912281 -0.55480518 4.5372613 0.2664423
## White-Khaki     0.5758052 -1.91812279 3.0697332 0.9985145
## Silver-Red     -0.4333333 -2.67544805 1.8087814 0.9995944
## White-Red      -1.8487562 -4.03152277 0.3340103 0.1733468
## White-Silver   -1.4154229 -3.59818944 0.7673437 0.5298095
```

```
par(mar=c(5.1,10,4.1,2.1), cex=0.8)
plot(TukeyHSD(tu), las=1)
```



- I used plot(TukeyHSD()) graphing since it can visually show the comparison between paired data.However, the data is too large, so before that, I need to set a default parameter with margin size that will enable to compact all the data of fit to the plot(TukeyHSD()).
- there are not much differences in means of number of sets in each colour, using the Tukey test the most difference I can see with is Blue-Beige, Bronze-Beige, White-Beige, Red-Blue, Red-Bronze, Silver-Bronze, Silver-Grey, and Silver-Khaki.

Question 1.3:

1/ Use Bootstrapping to compute a 88% confidence interval for the difference between GVM and TareWeights for Volkswagen vehicles?

2/ Compute a 88% cI for the difference between GVM and Tare weights for Volkswagen vehicle by using approximation.

3/ Can we conclude that GVM weights are different than Tare weights for Volkswagen vehicles (Dont do a hypothesis test)?

4/ Test the hypothesis that GVM weights are greater than Tare weights for Volkswagen vehicles

```
#1/
#There is no record on Volkswagen, a simulation boot to approximate data is needed:
count.Volkswagen.GVM.weight= length(subset(vehicles, Make=="Volkswagen", GVM.weight, drop=TRUE))
count.Volkswagen.GVM.weight
```

```
## [1] 29
```

```
count.Volkswagen.Tare.weight= length(subset(vehicles, Make=="Volkswagen", Tare.weight, drop=TRUE))
count.Volkswagen.Tare.weight
```

```
## [1] 29
```

- There are only 29 recorded data of the GVM.weight of Volkswagen.
- There are only 29 recorded data of the Tare.weight of Volkswagen

I do the simulation by bootstrapping the shuffled data from GVM and Tare weight of Volkswagen vehicles.

```
Volkswagen.GVM = subset(vehicles, Make=="Volkswagen", GVM.weight, drop=TRUE)
Volkswagen.GVM
```

```
##  [1] 5273  885 1127 3291 3889 1931 3151 5686 3176 2022 4734 3931 5166 2544 4610
## [16] 6001 6880 1553 5453 2417 2861 3956 1051 3056 6786 4484 6261 1040 5099
```

```
Volkswagen.Tare= subset(vehicles, Make=="Volkswagen", Tare.weight, drop=TRUE)
Volkswagen.Tare
```

```
##  [1] 1446 2305 3666 4444 3537 4604 2364 2943 4648 3285 3110 3331 1770 1155 3420
## [16] 2721 4521 2508 2091 1839 3551 1358 1505 3758 4938 5002 4986 3605 5048
```

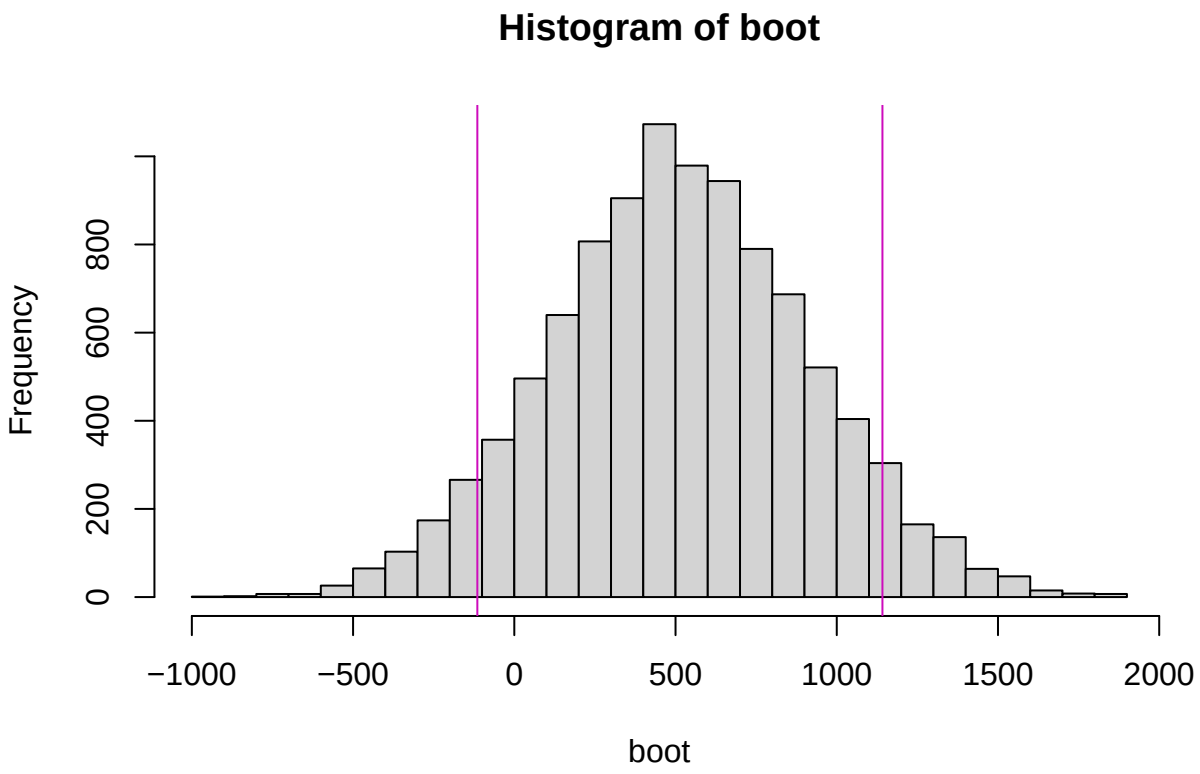
```
d0= mean(Volkswagen.GVM-Volkswagen.Tare)
d0
```

```
## [1] 512.2414
```

```r
boot= replicate(10000, {
  Volkswagen.GVM.sampled= sample(Volkswagen.GVM, replace=TRUE)
  Volkswagen.Tare.sampled= sample(Volkswagen.Tare, replace=TRUE)
  mean(Volkswagen.GVM.sampled-Volkswagen.Tare.sampled)
})
hist(boot, breaks=20)
quantile(boot, c(0.06, 0.94))
```

```
##        6%       94%
## -114.5683 1141.7207
```

```r
abline(v=quantile(boot, c(0.06, 0.94)), col=6)
```



I used distribution histogram to represent the bootstrapping data since it can show the frequency of occurrence of the bootstrapping data. And then the function abline shows the horizontal vector to highlight the range of 88% confident interval. In this boot distribution, the abline lines up between -114.5682759, 1141.7206897. And the highest frequency occurrence of the difference between sampled GVM and Tare weight of Volkswagen is at 506.7931034.

The Wilcox.test is used show the difference between two pairs of data. In this case, we use Wilcoxon-Mann-Whitney test to find the range of difference between GVM and Tare Weight of Volkswagen vehicles that has 88% confident level.

```r
#2/
CI_88_approx=wilcox.test(Volkswagen.GVM,Volkswagen.Tare, conf.int= TRUE, conf.level=0.88)
CI_88_approx
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Volkswagen.GVM and Volkswagen.Tare
## W = 493, p-value = 0.2651
## alternative hypothesis: true location shift is not equal to 0
## 88 percent confidence interval:
##  -268 1242
## sample estimates:
## difference in location
##                    511
```

```r
CI_88_approx$conf.int
```

```
## [1] -268 1242
## attr(,"conf.level")
## [1] 0.88
```

```r
#The approximate conf level of 88%CI

#3/
#To conclude that the GVM weight are different than Tare weights of Volkswagen vehicles since 88% of da
```

So 88% confident level locates between -268, 1242.

I will use the Wilcoxon-Mann-Whitney test to find if GVM weights are greater than Tare weights for the vehicle Volkswagen. Before that, I need to conduct the hypothesis.

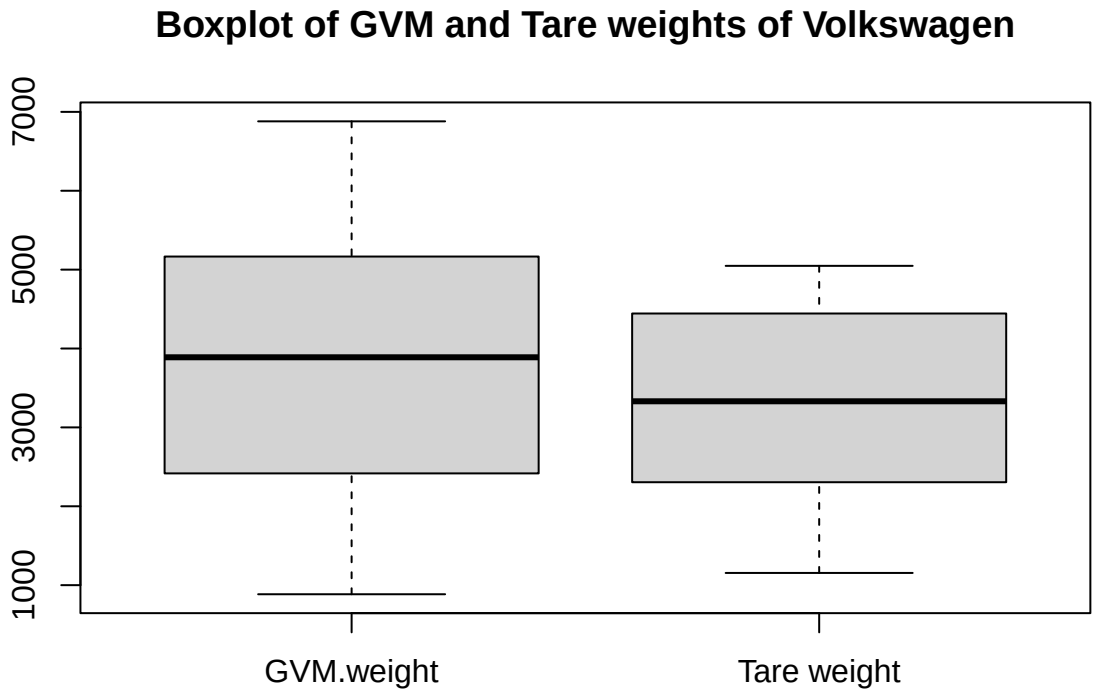
```r
#4/
#Using Wilcoxon-Mann Whitney test to find if  GVMweight are greater than TareWeights for Volkswagen

#H0: GVM.weight is not greater to Tare.Weights
#HA: GVM.weight is greater than Tare.weights

wilcox.test(Volkswagen.GVM, Volkswagen.Tare, alternative = "greater")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Volkswagen.GVM and Volkswagen.Tare
## W = 493, p-value = 0.1325
## alternative hypothesis: true location shift is greater than 0
```

```r
boxplot(Volkswagen.GVM, Volkswagen.Tare, names= c("GVM.weight", "Tare weight"), main="Boxplot of GVM an
```




```
#Since the p-value is more than 5%, the null hypothesis is favoured
```

- The box plot displays an overview of all GVM.weight and Tare.weight of Volkswagen vehicles from the max value to the min value, the mean, the range, the quartile data.
- I used box plot technique to summarise both data of GVm and Tare weight of Volkswagen in to two box plot techniques, so that we observe the general differences between GVM and Tare weight of the vehicles Volkswagen. And we can see the GVM weight is generally higher than the Tare weight but at the same time, it has a higher range from max value to min value than the Tare weight of the vehicle Volkswagen
- Since p-value is small ($p = 0.1325254$), therefore reject the null hypothesis, and we can conclude that GVM.weight is greater than Tare.weights

Question 1.4:

Test if there is a difference in proportions of the Silver vehicles between Landrover and Mercedes

```
l=length(subset(vehicles, Make=="Landrover", Colour, drop=TRUE))
l
```

```
## [1] 44
```

```
m=length(subset(vehicles,Make=="Mercedes", Colour, drop=TRUE))
m
```

```
## [1] 21
```

```
Colour.vehicles.table=table(vehicles$Make[vehicles$Colour=="Silver"])
Colour.vehicles.table
```

```
##
##        BMW     Holden      Honda      Isuzu        Kia  Landrover      Mazda
##          2          9          4          6          3          2          6
##   Mercedes Mitsubishi     Nissan      Skoda     Suziki     Toyota Volkswagen
##          3          4          3          3          3          6          6
```

```
Silver.Landrover=Colour.vehicles.table[7]
Silver.Landrover
```

```
## Mazda
##     6
```

```
Silver.Mercedes=Colour.vehicles.table[9]
Silver.Mercedes
```

```
## Mitsubishi
##          4
```

```
diff.proportion=(Silver.Landrover/l) - (Silver.Mercedes/m)
diff.proportion
```

```
##       Mazda
## -0.05411255
```

The difference in proportion between Silver Landrover and Mercedes would be -0.0541126. Meaning the proportion of Silver Landrover is -0.0541126 higher than Silver Mercedes.

Question 1.5:

The recent trend shows that people tend to buy more powerful vehicles. We would like to investigate whether there is a linear relationship between the registration year and the mean of the number of cylinders

a/ Decide if the mean numbers of cylinders and the registration year are linearly related?

b/ If so, compute the equation to predict mean number of cylinders by using the registration year and discuss the significance of this equation? What is your estimate of the population mean number of cylinders when the year is 1984?

```
#a/
min(vehicles$Year)
```

```
## [1] 1985
```

```
max(vehicles$Year)
```

```
## [1] 2021
```

```
mean.calculate <- tapply(vehicles$Number.of.cylinders, vehicles$Year, mean)
mean.calculate
```

```
##     1985     1986     1987     1988     1989     1990     1991     1992
## 7.090909 6.333333 7.800000 6.363636 4.888889 6.100000 6.454545 6.384615
##     1993     1994     1995     1996     1997     1998     1999     2000
## 7.368421 6.888889 6.923077 6.538462 5.142857 6.909091 6.600000 6.812500
##     2001     2002     2003     2004     2005     2006     2007     2008
## 7.583333 6.777778 7.000000 7.250000 7.875000 7.625000 8.473684 9.000000
##     2009     2010     2011     2012     2013     2014     2015     2016
## 7.190476 6.375000 7.250000 7.000000 7.722222 6.100000 5.625000 7.625000
##     2017     2018     2019     2020     2021
## 6.545455 6.428571 6.500000 5.578947 4.689655
```

```
year.register<- c(1985:2021)
```

```
#plotting the mean in N.cylinders depended on registration year
mean.calculated=tapply(vehicles$Number.of.cylinders,vehicles$Year, mean)
mean.calculated#this calculate the mean of cylinders of vehicles which has the same registration year
```

```
##     1985     1986     1987     1988     1989     1990     1991     1992
## 7.090909 6.333333 7.800000 6.363636 4.888889 6.100000 6.454545 6.384615
##     1993     1994     1995     1996     1997     1998     1999     2000
## 7.368421 6.888889 6.923077 6.538462 5.142857 6.909091 6.600000 6.812500
##     2001     2002     2003     2004     2005     2006     2007     2008
## 7.583333 6.777778 7.000000 7.250000 7.875000 7.625000 8.473684 9.000000
##     2009     2010     2011     2012     2013     2014     2015     2016
## 7.190476 6.375000 7.250000 7.000000 7.722222 6.100000 5.625000 7.625000
##     2017     2018     2019     2020     2021
## 6.545455 6.428571 6.500000 5.578947 4.689655
```

```r
year.registrated= c(1985:2021)

#h0: r  = 0
#hA: r != 0
r.Correlation=cor(mean.calculated,year.registrated, method="pearson")
r.Correlation # Correlation
```

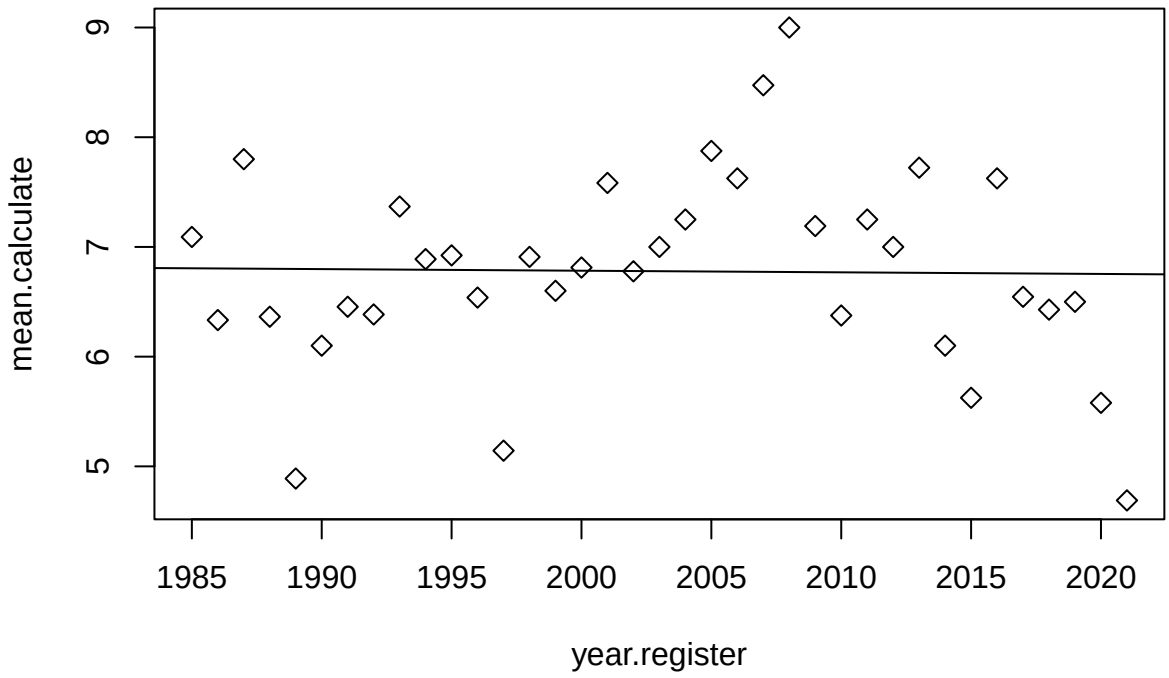
```
## [1] -0.01753745
```

r is a "really slight decreasing linear relationship", as r is -0.0175375, which has intermediate downward slope.

It is a slight decreasing linear relationship, we can still plot the linear equation to predict the next value of y or x. And we can also find the slope coefficient and intercept value through the function lm() as below:

```r
plot(y=mean.calculate, x=year.register, pch=5, main="Linear relationship between the number of cylinders
abline(lm(mean.calculate~year.register))
```

near relationship between the number of cylinders and the registration




```r
#There is a little to no downward trend but correlation coefficient is very weak.
```

```r
#b/
fit= lm(mean.calculate~year.register)
fit
```

```
##
## Call:
## lm(formula = mean.calculate ~ year.register)
##
## Coefficients:
##   (Intercept)  year.register
##      9.724992      -0.001471
```

```r
slope= fit[[1]][2]
slope
```

```
## year.register
##  -0.001470907
```

```r
intercept = fit[[1]][1]
intercept
```

```
## (Intercept)
##    9.724992
```

```r
x = 1990
axis= slope*x + intercept
axis
```

```
## year.register
##      6.797888
```

- Therefore, y is the predict number of cylinder when choosing the registration year.
- Therefore, axis = -0.0014709x + 9.7249919 is the new equation to find the predict number of cylinder depending on registration year. -when the year is 1984 then the number of cylinder will equal to 6.7978879

Question 2.1: Can a genuine causal relationship be established from this study? Justify your answer

This test genuinely cannot be concluded yet, multiple reasons are drawn: -Small Sample Sizes: The dataset includes a relatively small number of marijuana users—approximately 30 individuals—compared to an even smaller group of non-users, around 20 individuals. In a quantitative experiment, a small sample size limits the generalisability of the findings, as fewer occurrences are measured. This increases variability in estimating the drug's effects and makes it more challenging to reach definitive conclusions. However, the study mitigates some errors by ensuring that both groups come from the same city and age range, reducing potential biases related to population and age differences. Nonetheless, applying these findings to different populations or age groups could lead to varying results, potentially reducing the accuracy and reliability of the analysis.

Question 2.2: Can the results be generalised to other 14- to 16-year-olds? Justify your answer

This experiment cannot be generalised to other 14 to 16 years old because: A larger sample size and data collected from diverse populations, including different subgroups of 14- to 16-year-olds, are necessary to establish a consistent conclusion regarding the effects of marijuana use. Once a common pattern emerges, the findings can be reliably generalised to adolescents in this age range, allowing for a more professional and statistically confident assessment of marijuana's impact on short-term memory. However, several factors contribute to variability in the results:

Variabilities:

1. Temporal Variation in Brain Development: Although the experiment was conducted on 14- to 16-year-olds, individual differences in brain development exist. Some adolescents may develop more slowly than their peers, making them more susceptible to poor test results. Given the small sample size, chance could lead to a disproportionate number of slower-developing individuals in the drug-user group, while the non-user group might consist of individuals with faster cognitive development, skewing the results.
2. Population Differences: The findings cannot be applied to all 14- to 16-year-olds without testing a broader and more diverse sample, including different ethnicities and genders. Since the study was conducted on two sub-populations from the same larger population without specifying demographic details, the design is flawed. Drug effects may vary across genders and ethnic backgrounds, meaning the results may not be universally applicable.
3. Variation in Dosage and Age: The effects of marijuana depend on both the user's age and the dosage consumed. Younger individuals are generally more sensitive to the drug's effects, and higher doses can have a greater impact on memory function. If these factors are not carefully considered in the quantitative analysis, measurement errors may arise, further limiting the accuracy of the findings.

Question 2.3: What are some potential confounding factors

There are some potential confounding factors occurred in the experiments, such as:

The designation of the sample sizes of the drug effects' experiment: -A small sample size increases the likelihood that the selected sub-populations coincidentally share certain characteristics, meaning the observed effects of marijuana in the two groups may have occurred by chance. As a result, differences in outcomes could stem from the random selection of participants from the same population rather than the actual effects of the drug, introducing confounding factors in the study.

The average dosage per users depending on their current ages: -The impact of marijuana varies based on both dosage and the user's age, but it is unclear which factor plays a more significant role. Since both variables can contribute to the observed effects, they act as confounding factors in the analysis.

Population variations and gender types: -As previously mentioned, different populations and genders may respond differently to drug effects—some individuals may exhibit strong reactions, while others may show no response at all. Therefore, the neuropsychological test comparing drug users and non-users aged 14 to 16 from the same population may have produced results by chance. These findings may not necessarily apply to other populations or genders, making population variation and gender differences additional confounding factors in the study.