Alexandria University
Faculty of Engineering
Computer & Systems Engineering
CSE111: Probability Theory

LAB #6 – Report

Team Members:

Student Name	Student ID
Abdelrahman Khayri Saad	19015906
Mahmoud Tarek Mahmoud Embaby	20011800

Lab Requirements:

Applying R concepts that we have learnt in the previous labs. (Detailed Explanation Link).

1. Loading Data:

We were required to load a built-in data set called `mtcars` into the R distribution, then discover the data set. Accordingly, we discovered and analyzed the dataset to know the following:

Number of Observations	32		
Number of Variables	11		
Variables Names	{"mpg", "cyl", "disp", "hp", "drat", "wt", "qsec", "vs", "am", "gear", "carb"}		
Sample observations	Variable	Data Type	Values
	mpg	num	21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
	cyl	num	6 6 4 6 8 6 8 4 4 6 ...
	disp	num	160 160 108 258 360 ...
	hp	num	110 110 93 110 175 105 245 62 95 123 …
	drat	num	3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 ...
	wt	num	2.62 2.88 2.32 3.21 3.44 ...
	qsec	num	16.5 17 18.6 19.4 17 ...
	vs	num	0 0 1 1 0 1 0 1 1 1 ...
	am	num	1 1 1 0 0 0 0 0 0 0 ...
	gear	num	4 4 4 3 3 3 3 4 4 4 ...
	carb	num	4 4 1 1 2 1 4 2 2 4 ...

Thanks to the very useful built-in functions in R, we were able to get an analytical summary of each variable in the dataset which really helped us answer many questions in the lab.

Variable Name	Min.	1st Quantity	Median	Mean	3rd Quantity	Max.
mpg	10.40	15.43	19.20	20.09	22.80	33.90
cyl	4.000	4.000	6.000	6.188	8.000	8.000
disp	71.1	120.8	196.3	230.7	326.0	472.0
hp	52.0	96.5	123.0	146.7	180.0	335.0
drat	2.760	3.080	3.695	3.597	3.920	4.930
wt	1.513	2.581	3.325	3.217	3.610	5.424
qsec	14.50	16.89	17.71	17.85	18.90	22.90
vs	0.000	0.000	0.000	0.4375	1.000	1.000
am	0.000	0.000	0.000	0.4062	1.000	1.000
gear	3.000	3.000	4.000	3.688	4.000	5.000
carb	1.000	2.000	2.000	2.812	4.000	8.000

We were also able to fetch the head and tail of the dataset.

2. Extracting Information:

We used several built-in functions as well as dplyer library to extract information from mtcars dataset.

Requirement #1: Display the head of each type of transmission separately.

Knowing that `am` variable refers to the transmission of vehicle (0 = automatic, 1 = manual), we used dplyer's filter() function to filter cars according to transmission type then passed the output to head() function that returns the first `num` rows of a dataset.

Head of **Automatic Transmission** Cars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2

Head of **Manual Transmission** Cars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	

Requirement #2: Display the top 10 cars according to: Displacement, HP & DRAT.

We had two options to solve this problem, either sort the whole dataset one time according to the values of the three columns at the same time or three times by sorting one column each time. Accordingly, we got four outputs.

Method #1:
We used the built-in order() function that returns a permutation which rearranges its first argument into ascending or descending order, then passed the output to head() function that returns the first `num` rows of a dataset.

Method #2:
We used dplyer's arrange() function that orders the rows of a data frame by the values of selected columns, then passed the output to head() function that returns the first `num` rows of a dataset.

Output:
The two methods returned same output; a data-frame sorted by certain column as shown below.

SORTED BY {DISP, HP & DRAT}

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Cadillac Fleetwood	10.4	8	472	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Duster 360	14.3	8	360	245	3.21	3.570	15.84	0	0	3	4
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Ford Pantera L	15.8	8	351	264	4.22	3.170	14.50	0	1	5	4
Camaro Z28	13.3	8	350	245	3.73	3.840	15.41	0	0	3	4
Dodge Challenger	15.5	8	318	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.30	0	0	3	2

SORTED BY DISPLACEMENT

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Cadillac Fleetwood	10.4	8	472	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Duster 360	14.3	8	360	245	3.21	3.570	15.84	0	0	3	4
Ford Pantera L	15.8	8	351	264	4.22	3.170	14.50	0	1	5	4
Camaro Z28	13.3	8	350	245	3.73	3.840	15.41	0	0	3	4
Dodge Challenger	15.5	8	318	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.30	0	0	3	2

SORTED BY HORSEPOWER (HP)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3

SORTED BY DRAT

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4

We can observe in the data sorted by values of three columns that Displacement is dominating the sort. That's because displacement has a relatively large number compared to other variables values.

Requirement #3: Display cars whose mpg is above average only.

Average value could be calculated by two methods:

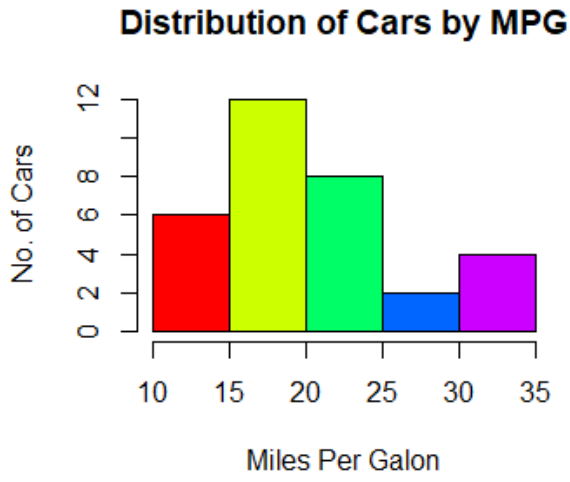
1. Using the built-in function: `mean()`.
2. Getting the sum of values and dividing by number of rows.

We used the second method. Value was confirmed by data analysis in part #1.

Having the average value of mpg, we were able to filter the dataset using dplyer's filter() function to find cars whose mpg is above average only.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

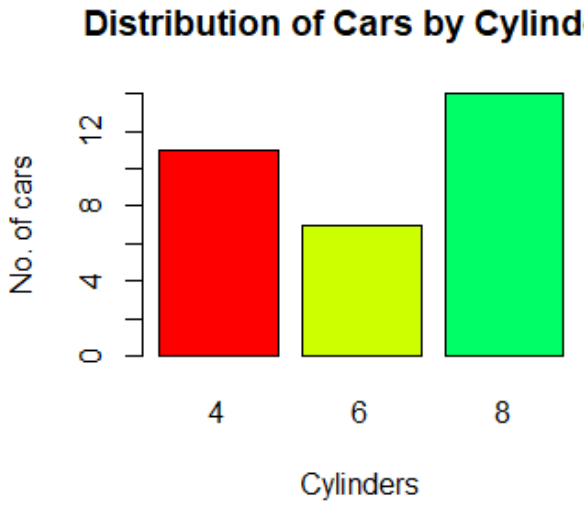
Requirement #4: What is the best type of chart to describe each feature of the mtcars dataset? Plot each chart type using R and state the reason behind each choice.



We decided to visualize **Milage (MPG)** using **HISTOGRAM** because:

A histogram is useful when visualizing discrete and continuous data where it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

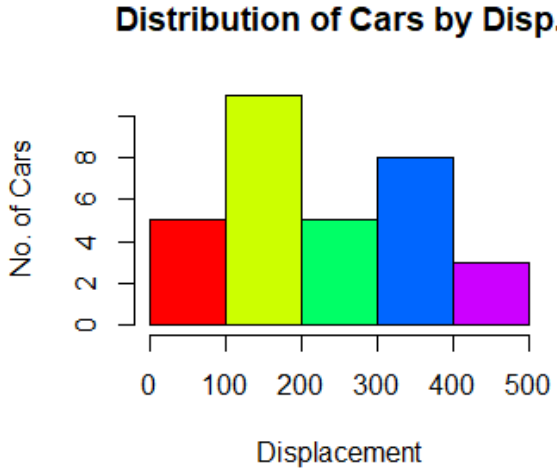
We can determine the median and distribution of the data.



We decided to visualize **Cylinders** using **BARPLOT** because:

The data is divided into three data points and bar plot helps us perform a comparison of metric values across the sub-groups of data. We can then know which group is the highest, most common, or lowest among the other groups in our dataset.

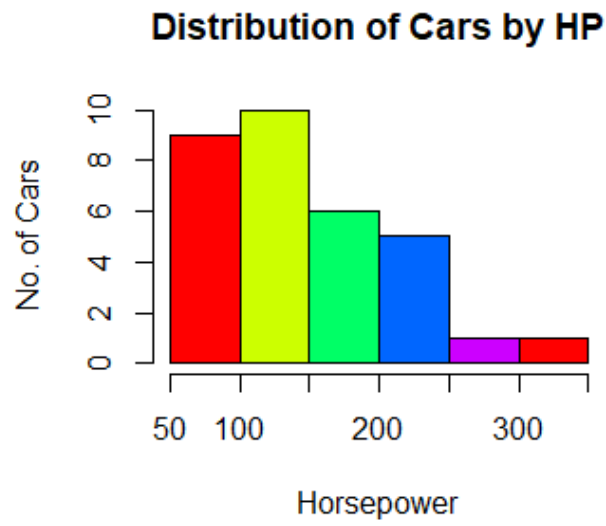
We can determine that most of the cars have 8 cylinders.



We decided to visualize **Displacement** using **HISTOGRAM** because:

A histogram is useful when visualizing discrete and continuous data where it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

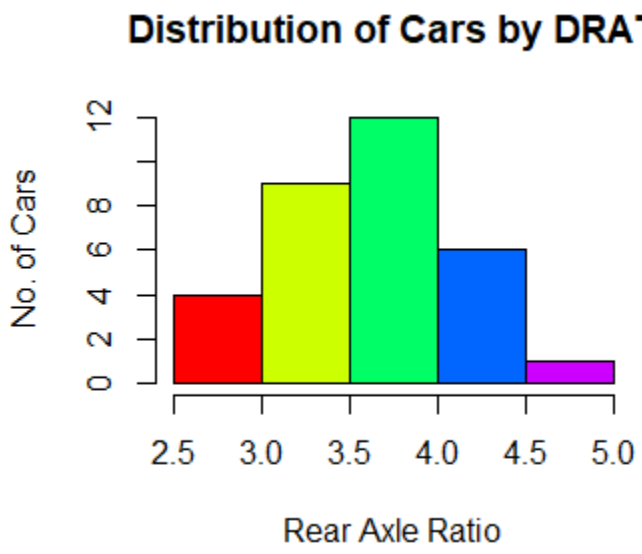
We can determine the median and distribution of the data.



We decided to visualize **Horsepower (HP)** using **HISTOGRAM** because:

A histogram is useful when visualizing discrete and continuous data where it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

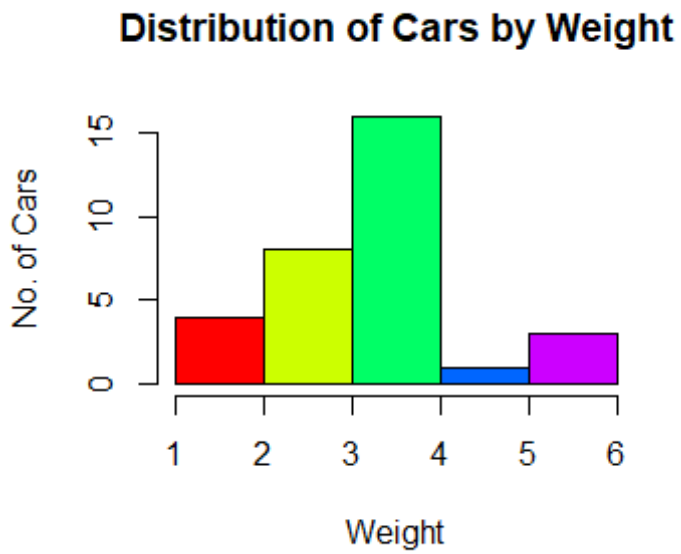
We can determine the median and distribution of the data.



We decided to visualize **DRAT** using **HISTOGRAM** because:

A histogram is useful when visualizing discrete and continuous data where it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

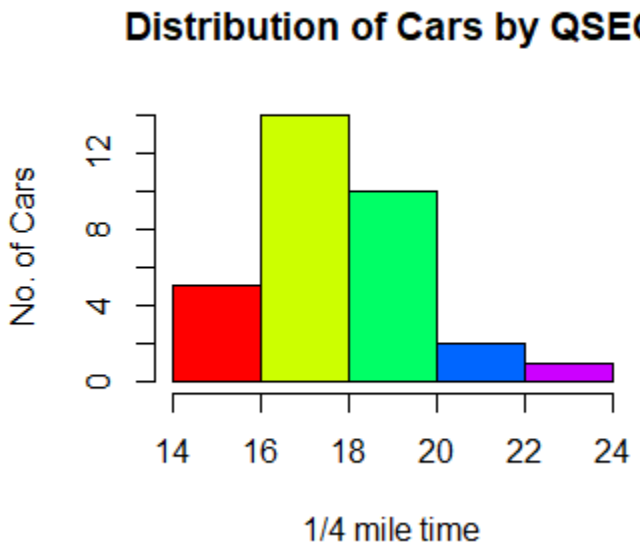
We can determine the median and distribution of the data.



We decided to visualize **WEIGHT** using **HISTOGRAM** because:

A histogram is useful when visualizing discrete and continuous data where it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

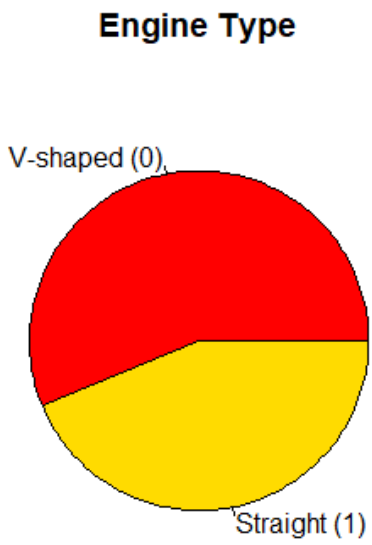
We can determine the median and distribution of the data.



We decided to visualize **QSEC** using **HISTOGRAM** because:

A histogram is useful when visualizing discrete and continuous data where it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

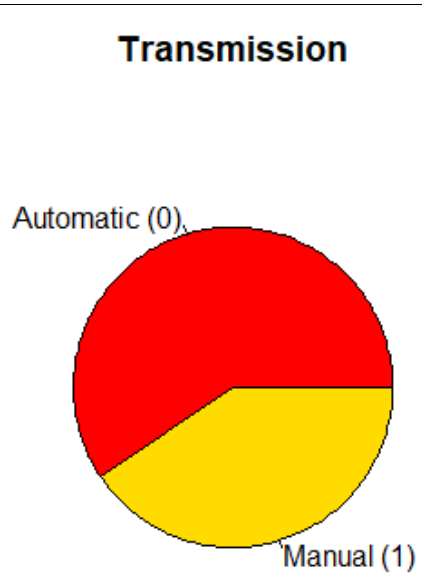
We can determine the median and distribution of the data.



We decided to visualize **ENGINE TYPE (VS)** using **PIE CHART** because:

We have a small number of possible values (0 and 1). Pie chart can efficiently show the dominating values, the percentage of each value and it is best used when we have small set of possible values. Finally, they look nice!

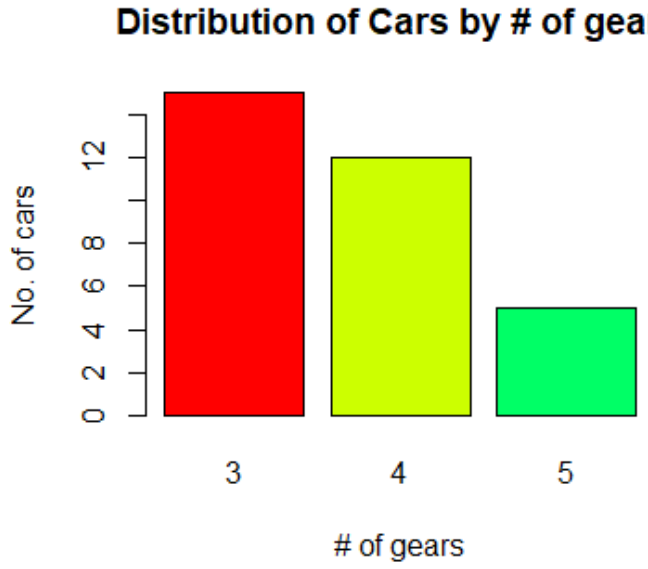
We can determine that the V-shaped engine types are more common.



We decided to visualize **TRANSMISSION (AM)** using **PIE CHART** because:

We have a small number of possible values (0 and 1). Pie chart can efficiently show the dominating values, the percentage of each value and it is best used when we have small set of possible values. Finally, they look nice!

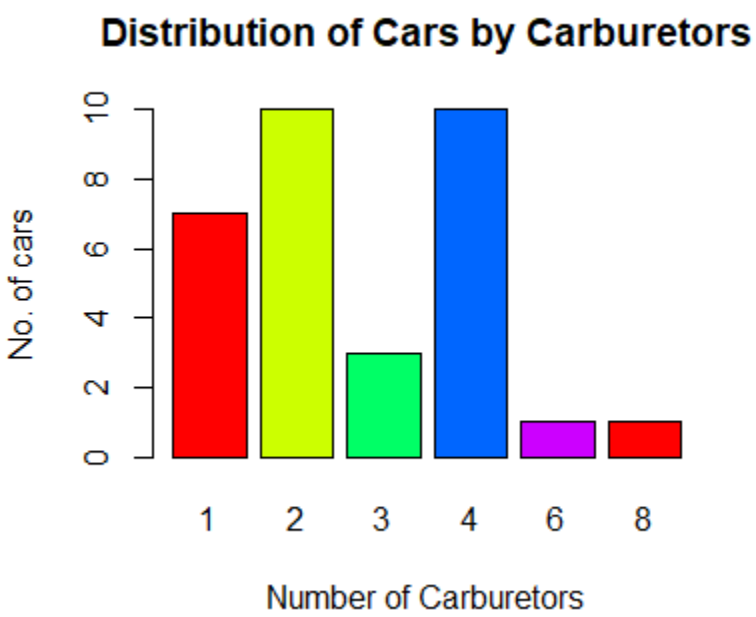
We can determine that the automatic cars are more common.



We decided to visualize **Number of Gears (gear)** using **BARPLOT** because:

The data is divided into three data points and bar plot helps us perform a comparison of metric values across the sub-groups of data. We can then know which group is the highest, most common, or lowest among the other groups in our dataset.

We can determine that most of the cars have 3 gears.



We decided to visualize **Number of Gears (gear)** using **BARPLOT** because:

The data is divided into 6 data points and bar plot helps us perform a comparison of metric values across the sub-groups of data. We can then know which group is the highest, most common, or lowest among the other groups in our dataset.

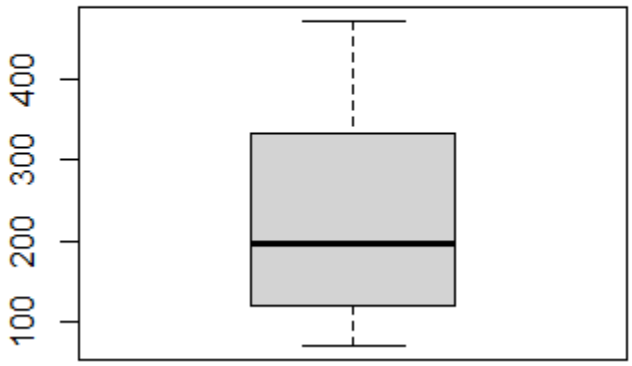
We can determine that most of the cars have 2 or 4 carburetors.

We used the built-in functions to create all the graphs and add some colors to make it look good. We also validated all the results by visualizing a distribution of each feature and it looked very similar to the created chart.

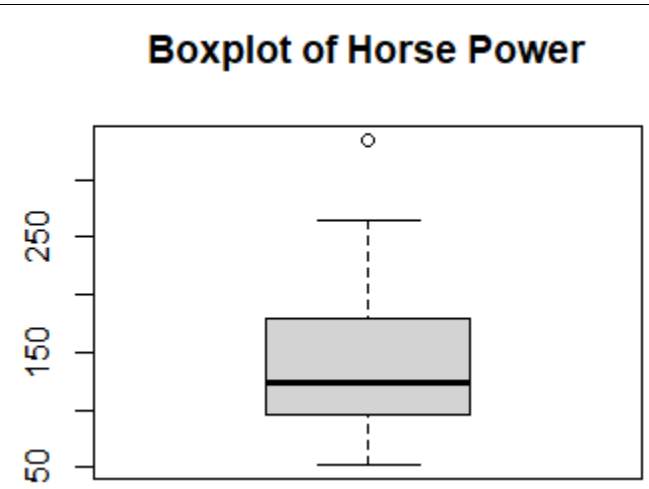
Requirement #5: Plot the boxplots for the following features: disp, hp and qsec. Extract the 3 main percentiles. What can you deduce?

We used the built-in function to create boxplots for the three features: Displacement, Horsepower and QSEC.

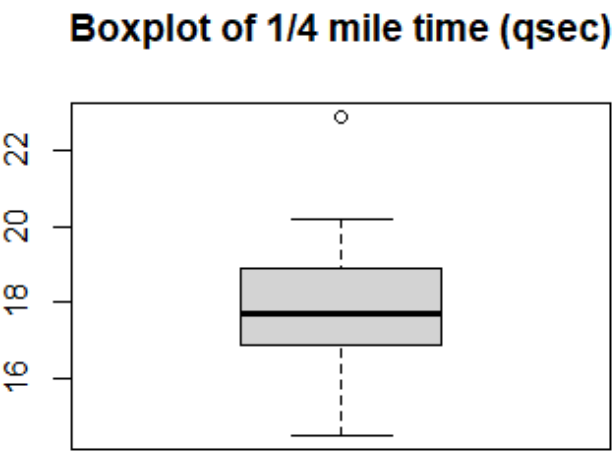
Then to get the percentiles, we used the built-in function quantile()



PERCENTILES	DEDUCTION
Q_1 = 120.825 (25% Percentile)	We can conclude that the variance of displacement is **very large**.
Q_2 = 196.300 (50% Percentile)	
Q_3 = 326.000 (75% Percentile)	



PERCENTILES	DEDUCTION
Q_1 = 96.5 (25% Percentile)	We can conclude that the variance of hp is relatively **large**.
Q_2 = 123.0 (50% Percentile)	
Q_3 = 180.0 (75% Percentile)	



PERCENTILES	DEDUCTION
Q_1 = 16.8925 (25% Percentile)	We can conclude that the variance of qsec is **low**.
Q_2 = 17.7100 (50% Percentile)	
Q_3 = 18.9000 (75% Percentile)	

3. Distributions:

Requirement #A: Assume that the weight fits a normal distribution. Find the percentage of cars having 3.4 lbs or more.

Algorithm

The weight data is retrieved, and the mean and standard deviation are calculated from it. The cumulative probability distribution at 3.4 lb is calculated using pnorm() then is subtracted from 1.

Output

```
"Percentage of cars having weight of 3.4 lb or more is 42.5919081041855%"
```

Answer

42.5919%

Requirement #B: What is the probability of getting 18 or less manual cars using these 32 observations? Assume that the probability of getting a manual car in an infinite series of cars is equal to the probability of getting a manual car from this dataset.

Algorithm

The number of manual cars in the dataset is counted and divided by the total number of data points to get the probability of a car having manual transmission. The probability is then used in pbinom() to calculate the probability of having 18 or less cars have manual transmission.

Output

```
"Probability of 18 or less cars out of 32 being manuals is 0.945029440815751"
```

Answer

0.945029

Requirement #C: Suppose there are twelve spots in a car parking area. Each spot is suitable for five possible car types, and only one of them fits perfectly. Find the probability of having four or less spots filled with the corresponding car type if the garagist attempts to park in each spot at random.

Algorithm

The number of outcomes in the sample space is calculated with 5^{12}. The number of favored outcomes (4 or less parking spots having the correct car type parked in it) is calculated with a for loop in which each iteration calculates the number of ways of having a specific number of cars being parked in the correct spot (For example, the first iteration calculates the number of ways of no cars being parked in the correct spot, the second is the ways of 1 car in the correct spot, etc). Each value is then added to a favored outcomes count and is then divided by the sample space to get the desired probability.

Output

```
"Probability of 4 or less parking spots being having the correct car type is 0.92744450048"
```

Answer

0.9274445

4. Permutations and Combinations:

Requirement #A: Given that we have a number in the ternary numeral system, this number has 3 digits. Use R to find all the permutations for such number. Solve using 2 different methods.

Algorithm

The first method to find the possible permutations for a ternary based number that is 3 digits long is to use permutations() to get all possible permutations of the digits 0,1,2 with replacement, then remove all the permutations with the digit 0 as the first digit to assure that the number is 3 digits long.

The second method is to use a loop to iterate through the digits 0,1 that can be placed in the first digit (The leftmost digit). A nested loop is added to it to iterate through the digits 0,1,2 that can be placed in the second digit. A nested loop is added to the second level to iterate through the digits 0,1,2 that can be placed in the third digit.

Output

```
      [,1] [,2] [,3]
 [1,]    1    0    0
 [2,]    1    0    1
 [3,]    1    0    2
 [4,]    1    1    0
 [5,]    1    1    1
 [6,]    1    1    2
 [7,]    1    2    0
 [8,]    1    2    1
 [9,]    1    2    2
[10,]    2    0    0
[11,]    2    0    1
[12,]    2    0    2
[13,]    2    1    0
[14,]    2    1    1
[15,]    2    1    2
[16,]    2    2    0
[17,]    2    2    1
[18,]    2    2    2
```

Both methods yield the same permutations.

Requirement #B: Given that we have a number in the ternary numeral system, this number has 3 digits. Use R to find all the permutations for such number. Solve using 2 different methods.

Algorithm

The sample space count is calculated with this formula

$$S = 9!/((9-3)!\,3!)$$

To calculate the favored outcomes, we first take the numbers 2 and 5 for the maximum and minimum numbers which leaves one number to be chosen. The number must be between 2 and 5 to maintain the required max and min. This means there are 2 ways to choose 3 numbers, the max of which is 5 and the min is 2. Dividing this number by S gets the desired probability.

Output

```
"Probability of getting 3 numbers where the max is 5 and the min is 2 is 0.0238095238095238"
```

This is the resulting probability from applying the above algorithm.

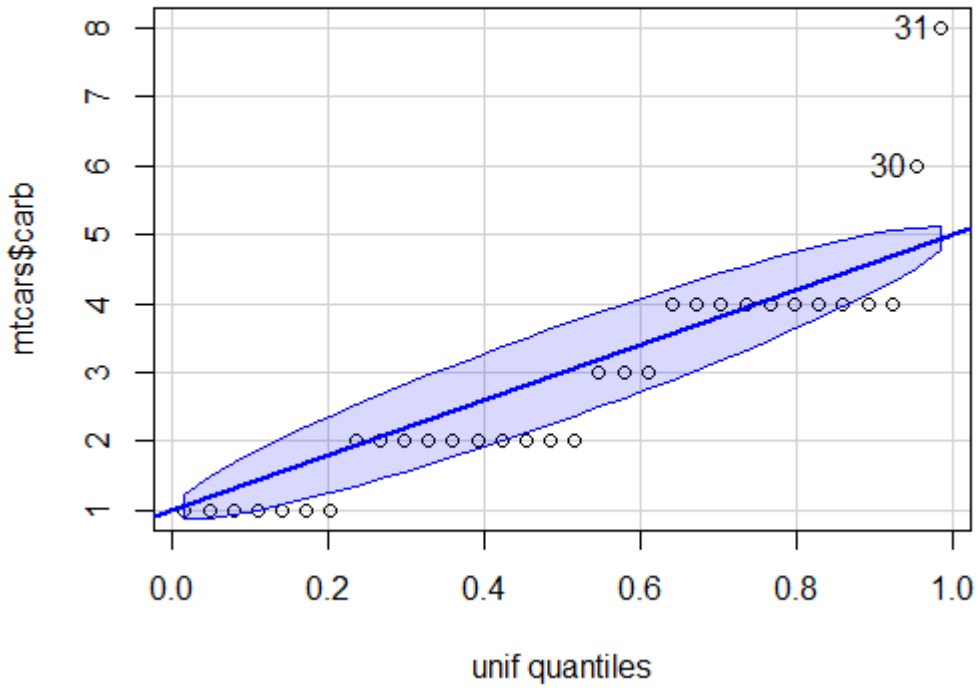
Answer

0.023809

5. Bonus:

Requirement #A: Plot Q-Q plot for the mtcars dataset.

With the help of library “car”, we could create a Q-Q plot for the mtcars using the function `qqPlot`.



Requirement #B: Extract all the information you can deduce from the plotted graph.

MPG follows the uniform distribution

CYL follows the uniform distribution

DISP follows the uniform distribution

HP follows the uniform distribution

DRAT follows the uniform distribution

WT follows the normal distributed

QSEC follows the uniform distribution

AM follows the poisson distribution

GEAR follows the uniform distribution

CARB follows the uniform distribution

References:

https://www.rdocumentation.org/

https://www.coursera.org/learn/r-programming

https://r-dir.com/community/forums.html

https://stackoverflow.com

https://www.programmingr.com/animation-graphics-r/qq-plot/