Q. What type of data are we using ?

First we check that we are using a data frame.

```
> is.data.frame(mtcars)
[1] TRUE
```

We can type the entire dataset by using 'mtcars' but this is impractical for larger datasets. Instead we can use **head** to see the first 6 items, and **tail** to see the last 6 items.

```
> head (mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
> tail(mtcars)
                mpg cyl  disp  hp drat    wt qsec vs am gear carb
Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    2
Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1    5    4
Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
```

The purpose of the above is to get a sense of the data and from this we can see what type of data the individual columns hold:

- **mpg** is continuous and quantitive.
- **cyl** appears to be categorical even though it looks like numerical data. All the values appear to be either 4, 6 or 8. This makes sense for a cloumn that is displaying the number of cylinders.
- **disp** is continuous and quantitative.
- **hp** is discrete and quantitative.
- **drat** is discrete and quantitative.
- **gsec** is continuous and quantitative.
- **vs** is categorical
- **am** is categorical
- **gear** is categorical
- **carb** is categorical

We can also provide a seq using subsetting notation to view particular rows or columns. We can do this to check the hypothesis about the types of datas in the columns that we have made above.

```
> mtcars[7:12,]
             mpg cyl  disp  hp drat   wt  qsec vs am gear carb
Duster 360  14.3   8 360.0 245 3.21 3.57 15.84  0  0    3    4
Merc 240D   24.4   4 146.7  62 3.69 3.19 20.00  1  0    4    2
Merc 230    22.8   4 140.8  95 3.92 3.15 22.90  1  0    4    2
Merc 280    19.2   6 167.6 123 3.92 3.44 18.30  1  0    4    4
Merc 280C   17.8   6 167.6 123 3.92 3.44 18.90  1  0    4    4
Merc 450SE  16.4   8 275.8 180 3.07 4.07 17.40  0  0    3    3
```

Nothing here seems to go against our hypothesis but we can do more by choosing a column. For example 10 to check if all the values in gear are in categories.

```
mtcars[,10]
 [1] 4 4 4 3 3 3 3 4 4 4 4 3 3 3 3 3 3 4 4 4 3 3 3 3 3 4 5 5 5 5 5 4
```

```
So, yes, they are all in {3,4,5}. We could also have used column names in the
format:
      datasetName$columnName

> mtcars$carb
 [1] 4 4 1 1 2 1 4 2 2 4 4 3 3 3 4 4 4 1 2 1 1 2 2 4 2 1 2 2 4 6 8 2
```

We can omit the datsetName and $ if we use attach.

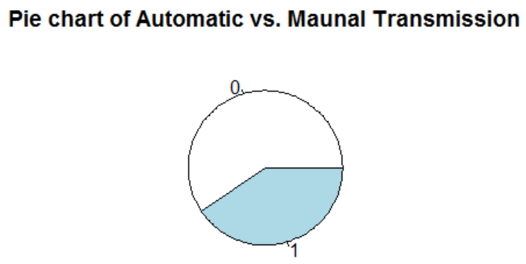
```
> attach(mtcars)
> carb
 [1] 4 4 1 1 2 1 4 2 2 4 4 3 3 3 4 4 4 1 2 1 1 2 2 4 2 1 2 2 4 6 8 2
```

This gives us the same output as above.

Do more cars have automatic or manual transmission ?

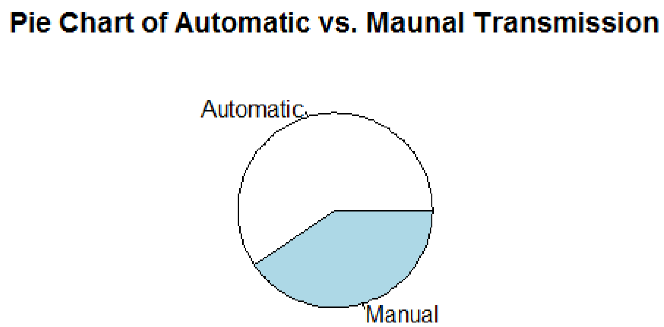
We can use **table** to aggregate the categorical data in the column **am**. When their are few categories and we just want a general idea of the relative size between them a pie chart is effective.

```
> amcounts <- table(mtcars$am)
> pie(amcounts, main = "Pie chart of Automatic vs. Maunal Transmission")
```



We can see a clear winner but I have already forgotten which is 0 and which is 1. Let's go back and add labels.

```
labels = c("Automatic", "Manual")
pie(amcounts, main = "Pie Chart of Automatic vs. Maunal Transmission", labels = labels)
```



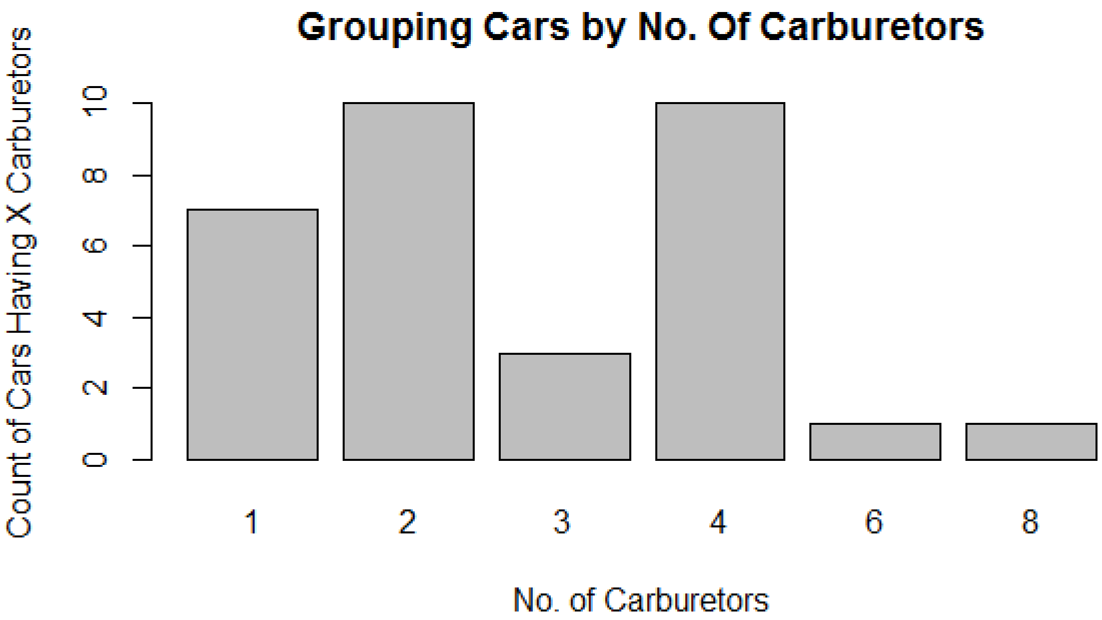
Ok, so more cars have automatic transmission but it is hardly uncommon to find one with manual.

Q. What is the distribution of cars by number of carburetors?

```
> carb_amounts <- table(carb)
> carb_amounts
carb
 1  2  3  4  6  8
 7 10  3 10  1  1
```

Now we have 6 groups which is too much for a pie chart. We have categorical data so we will use a bar chart rather than a histogram.

```
> barplot(carb_amounts, xlab="No. of Carburetors", ylab="Count of Cars Having X Carburetors", main="Grouping Cars by No. Of Carburetors")
```



1,2,4 are very frequent and take up most of the set. It is interesting that there are 3 cars with 3 carburetors. Let's have a look at them :

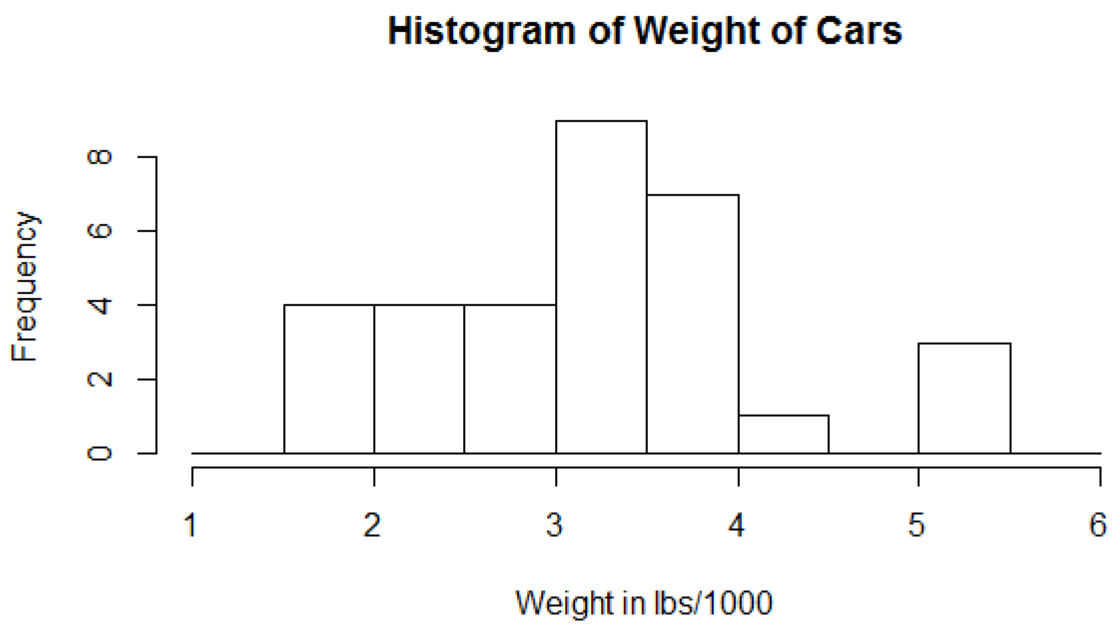
```
> mtcars[carb==3,]
             mpg cyl  disp  hp drat   wt qsec vs am gear carb
Merc 450SE  16.4   8 275.8 180 3.07 4.07 17.4  0  0    3    3
Merc 450SL  17.3   8 275.8 180 3.07 3.73 17.6  0  0    3    3
Merc 450SLC 15.2   8 275.8 180 3.07 3.78 18.0  0  0    3    3
```

That makes sense that they are from a similar car, Merc 450, – I expected just the same company – as their are too few for it to be common but too many for it to be an outlier.

Q. What is the distribution of the weight of cars ?

Now we are going to use continuous data so we must use a histogram. After some experimentation I settled on buckets of 500 lbs between 1,000 and 6,000 lbs.

```
hist(wt, breaks=seq(1,6, 0.5), main="Histogram of Weight of Cars", xlab="Weight in lbs/1000")
```



The most popular weight cars are between 3,000 to 4,000 lbs but when not in that range there are many more on the lighter side. I would theorize that since lighter cars are more efficient on fuel (checked in a later question) there are quite a number of people who have this as a strong factor in their decision to purchase a car. On the other side when mpg is not considered important then the other factors, perhaps safety or size, lead to the use of more and heavier materials which causes the weight to balloon.

```
> mtcars[wt > 4,]
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
```

The three heaviest are among the five worst for mpg. With the Cadillac Fleetwood and Lincoln Continental being extreme outliers.

```
> mtcars[mpg < 15,]
                     mpg cyl disp  hp drat    wt  qsec vs am gear carb
Duster 360          14.3   8  360 245 3.21 3.570 15.84  0  0    3    4
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
Camaro Z28          13.3   8  350 245 3.73 3.840 15.41  0  0    3    4
```

Q. Which of the quantile types is the default used in summary ?

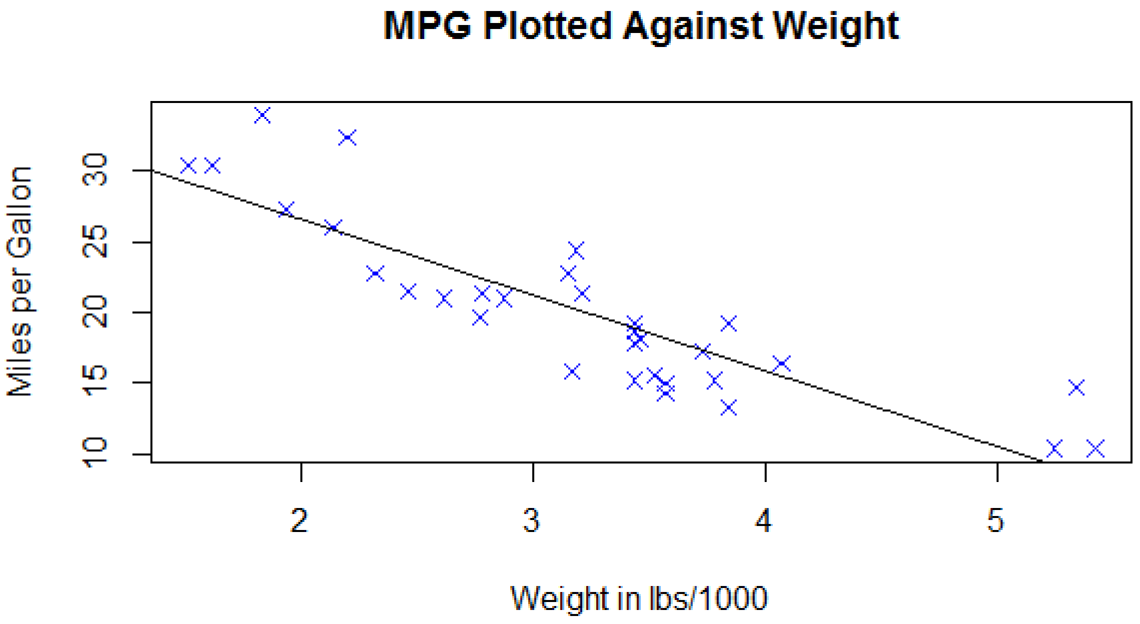
```
> summary(wt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.513   2.581   3.325   3.217   3.610   5.424 

> quantile(wt, type=1)
   0%   25%   50%   75%  100% 
1.513 2.465 3.215 3.570 5.424 
> quantile(wt, type=2)
    0%    25%    50%    75%   100% 
1.5130 2.5425 3.3250 3.6500 5.4240 
> quantile(wt, type=3)
   0%   25%   50%   75%  100% 
1.513 2.465 3.215 3.570 5.424 
> quantile(wt, type=4)
   0%   25%   50%   75%  100% 
1.513 2.465 3.215 3.570 5.424 
> quantile(wt, type=5)
    0%    25%    50%    75%   100% 
1.5130 2.5425 3.3250 3.6500 5.4240 
> quantile(wt, type=6)
     0%     25%     50%     75%    100% 
1.51300 2.50375 3.32500 3.69000 5.42400 
> quantile(wt, type=7)
     0%     25%     50%     75%    100% 
1.51300 2.58125 3.32500 3.61000 5.42400 
> quantile(wt, type=8)
      0%      25%      50%      75%     100% 
1.513000 2.529583 3.325000 3.663333 5.424000 
> quantile(wt, type=9)
      0%      25%      50%      75%     100% 
1.513000 2.532812 3.325000 3.660000 5.424000 
```

Type 7 is the one that matches summary.

Q Are lighter cars more fuel efficient?

```
plot(wt, mpg, main="MPG Plotted Against Weight", xlab="Weight in lbs/1000",
ylab="Miles per Gallon", pch=4, col="blue")
abline(lm(mpg~wt))
```



I would say clearly yes as their is a clear downward line of best fit.

Q. Are more cars below the mean weight or above it?

For this we will need both the median, to find a point that 50 % of cars above and below, and the mean weight of the cars.

```
> median(wt)
[1] 3.325

> mean(wt)
[1] 3.21725
```

Since median is higher than mean most cars' weight is above the mean. We could also have gotten these values from summary.

```
> summary(wt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.513   2.581   3.325   3.217   3.610   5.424 
```

Q. Which has the larger standard deviation and variance of drat and wt?

```
> sd(drat)
[1] 0.5346787
> sd(wt)
[1] 0.9784574
> var(drat)
[1] 0.2858814
> var(wt)
[1] 0.957379
```

The standard deviation and variance are larger for wt.