Create data frame with these two column vectors

x = 1:30

y = x^3

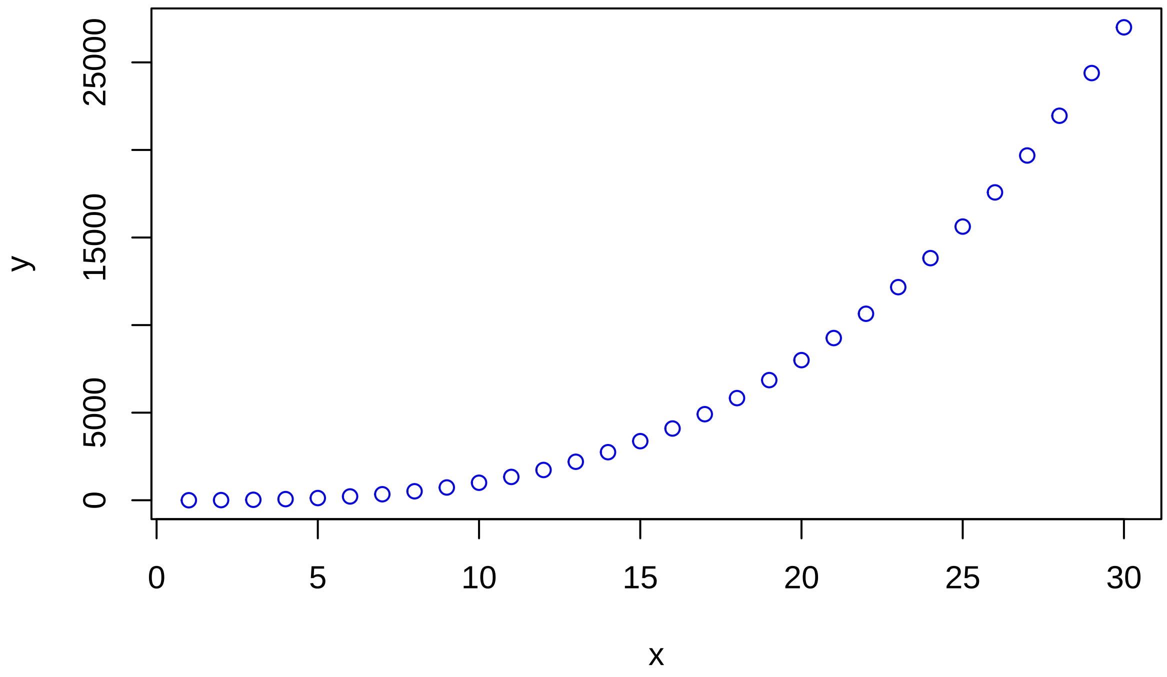
A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

```
x <- 1:30
y <- x^3
df <- data.frame(x,y)
print(head(df))
```

```
##   x   y
## 1 1   1
## 2 2   8
## 3 3  27
## 4 4  64
## 5 5 125
## 6 6 216
```

Create plot of x and y variables and interpret it carefully.

```
plot(df$x, df$y, xlab="x", ylab="y", col="blue")
```



The plot shows nonlinear relationship between dependent variable x and independent variable y having positive correlation.

### Get appropriate correlation coefficient of this data in and interpret it carefully.

Since the relation is nonlinear so use spearman correlation coefficient. Spearman correlation evaluates the monotonic relationship.

```
corr <- cor.test(x=df$x, y=df$y, method='spearman')
print(corr)
```

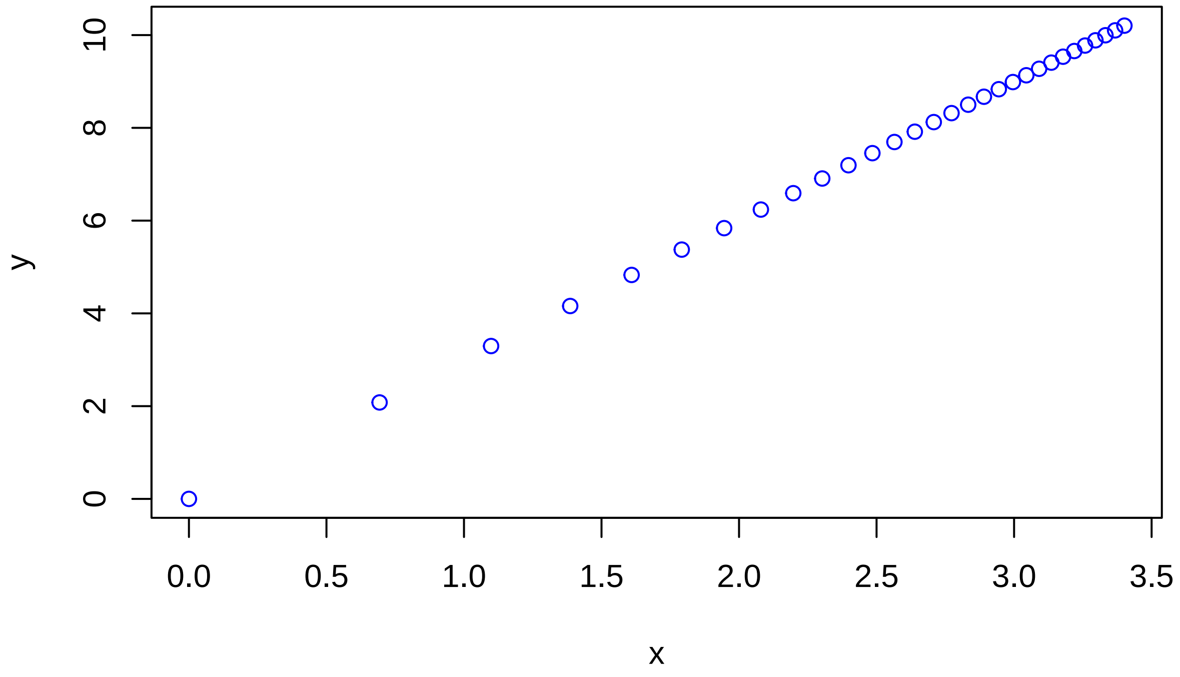
```
##
##  Spearman's rank correlation rho
##
## data:  df$x and df$y
## S = 0, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
##   1
```

A Spearman correlation 1 results when the two variables being compared are monotonically related, even if the relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well.

Transform the plot to linear using appropriate mathematical function.

Transforming non linear relation to linear relation using log().

```
plot(log(df$x), log(df$y), xlab = 'x', ylab = 'y', col="blue")
```



## Get appropriate correlation coefficient now in R Studio and interpret it carefully too.

```
pear_method = cor(log(df$x), log(df$y), method='pearson')
print(pear_method)
```

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

Pearson's Correlation Coefficient is a linear correlation coefficient that returns a value of between -1 and +1. +1 means that there is a strong positive correlation.

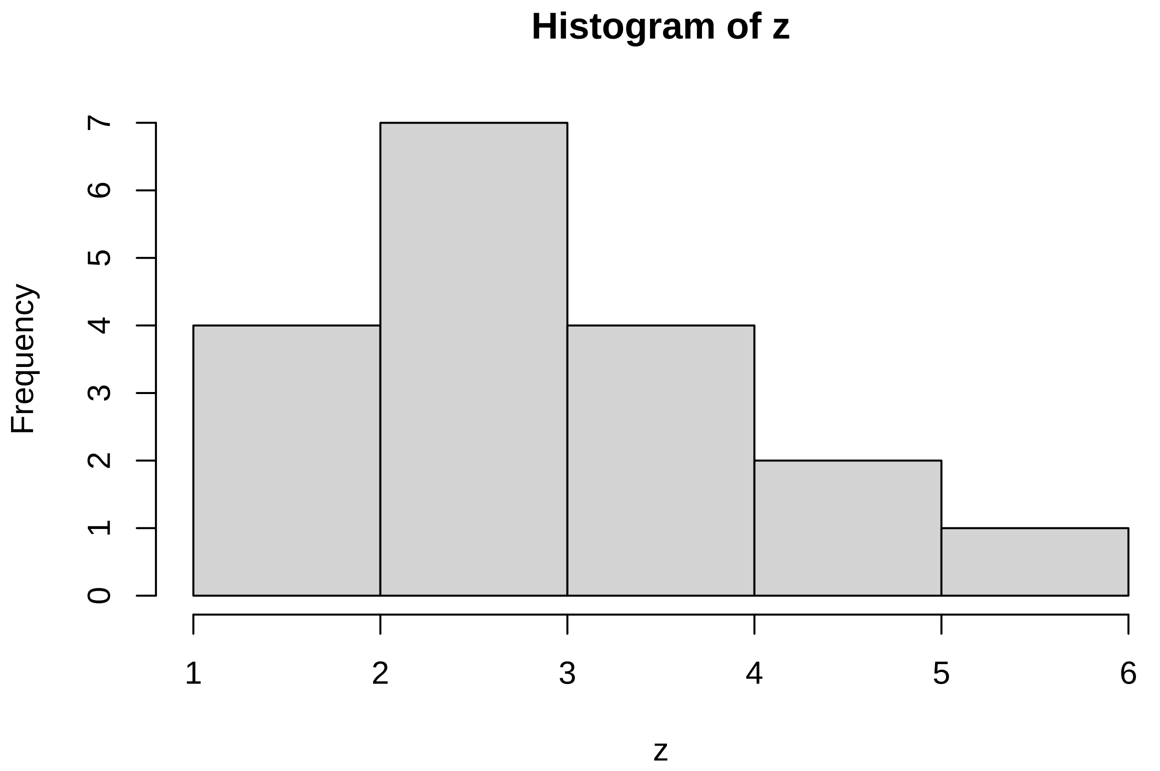
Create a new column vector z defined in the slide 18 of session two slide deck.

```
z <- c(1,2,2,2,3,3,3,3,3,3,3,4,4,4,4,5,5,6)
print(z)
```

```
##  [1] 1 2 2 2 3 3 3 3 3 3 3 4 4 4 4 5 5 6
```

Create a histogram of z variable and interpret it carefully.

```
hist(z)
```



Z is positive or right skewed. Here the distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer.

Get summary statistics of z variable and interpret it carefully.

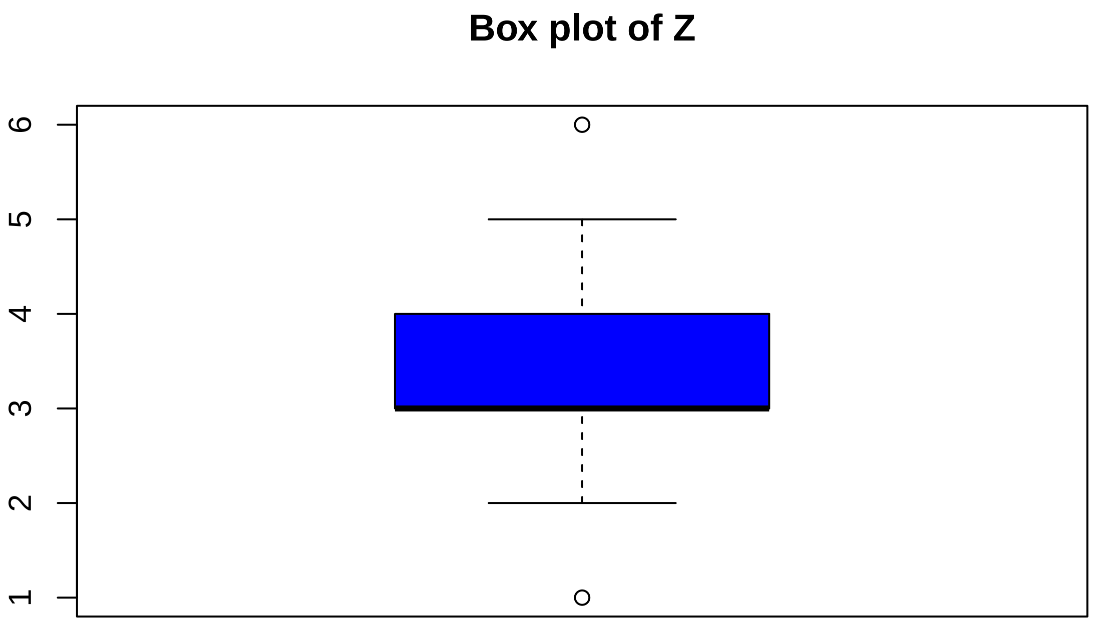
```
summary(z)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.000    3.000    3.000    3.333    4.000    6.000
```

Summary() gives the better idea of distribution of the variable. The summary of data gives Q1, Q2, Q3, Q4 (quartiles) and min max of the data. Interquartile range (IQR) of data = Q3-Q1 = 1 i.e the middle 50% of data betweem 3 and 4 is 1. The minimum and maximum of data will be 1 and 6 i.e the data lies between 1 and 6.

Get box-plot of z variable and interpret the result carefully.

```
boxplot(z, col = "blue", main="Box plot of Z")
```



A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It tells about outliers and what their values are. We can also find inter quartile range (IQR) = Q3-Q1.