This is a quick guide of interpretation for the results of ***lm*** R function

Let's get started by running one example:

```
mlr <- lm(BSAAM~., data = filter.water)
summary(mlr)

# Output

Call:
lm(formula = BSAAM ~ ., data = filter.water)

Residuals:
   Min     1Q Median     3Q    Max
-12690  -4936  -1424   4173  18542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15944.67    4099.80   3.889 0.000416 ***
APMAM         -12.77     708.89  -0.018 0.985725
APSAB        -664.41    1522.89  -0.436 0.665237
APSLAKE      2270.68    1341.29   1.693 0.099112 .
OPBPC          69.70     461.69   0.151 0.880839
OPRC         1916.45     641.36   2.988 0.005031 **
OPSLAKE      2211.58     752.69   2.938 0.005729 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7557 on 36 degrees of freedom
Multiple R-squared:  0.9248,    Adjusted R-squared:  0.9123
F-statistic: 73.82 on 6 and 36 DF,  p-value: < 2.2e-16
```

The model above is achieved by using the lm() function in R and the output is called using the summary() function on the model.

Below we define and briefly explain each component of the model output:

**Formula Call**

As you can see, the first item shown in the output is the formula R used to fit the data. Note the simplicity in the syntax: the formula just needs the predictors and the target/response variable, together with the data being used.

**Residuals**

The next item in the model output talks about the residuals. Residuals are essentially the difference between the actual observed response values (distance to stop dist in our case) and the response values that the model predicted. The Residuals section of the model output breaks it down into 5 summary points. When assessing how well the model fit the data, you should look for a symmetrical distribution across these points on the mean value zero (0). We could take this further consider plotting the residuals to see whether this normally distributed, etc. but will skip this for this example.

**Coefficients**

The next section in the model output talks about the coefficients of the model.

*Coefficient - Estimate*

We one line for each explicative variable or predictor plus one more row 'Intercept', Intercept is giving data when all the variables are 0 so all the measure done without considering any variable, this is again not much used in normal cases, it's average value of y when x = 0

```
#                Estimate    Std. Error t value Pr(>|t|)

# (Intercept) 15944.67    4099.80   3.889 0.000416 ***
```

*Coefficient - Standard Error*

The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We'd ideally want a lower number relative to its coefficients. The Standard Error can be used to compute an estimate of the expected difference in case we ran the model again and again. The Standard Errors can also be used to compute confidence intervals and to statistically test the hypothesis of the existence of a relationship between predictors and dependent variables.

*Coefficient - t value*

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis. In our example, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists. In general, t-values are also used to compute p-values.

*Coefficient - Pr(>t)*

The *Pr(>t)* acronym found in the model output relates to the probability of observing any value equal or larger than *t*. A small p-value indicates that it is unlikely we will observe a relationship between the predictor and response variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. Note the 'signif. Codes' associated to each estimate. Three stars (or asterisks) represent a highly significant p-value.

**Residual Standard Error**

Residual Standard Error is measure of the *quality* of a linear regression fit. Theoretically, every linear model is assumed to contain an error term *E*. Due to the presence of this error term, we are not capable of perfectly predicting our response variable from the predictor ones. The Residual Standard Error is the average amount that the response will deviate from the true regression line. It's also worth noting that the Residual Standard Error was calculated with 36 degrees of freedom. Simplistically, degrees of freedom are the number

of data points that went into the estimation of the parameters used after taking into account these parameters (restriction).

**Multiple R-squared, Adjusted R-squared**

The R-squared statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. It is a measure of the linear relationship between our predictor variables and our response / target variable. It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In our example, the R2 we get is 0.9248. Or roughly 92% of the variance found in the response variable can be explained by the predictor variables. Nevertheless, it's hard to define what level of R2 is appropriate to claim the model fits well. Essentially, it will vary with the application and the domain studied.

A side note: In multiple regression settings, the R2 will always increase as more variables are included in the model. That's why the adjusted R2 is the preferred measure as it adjusts for the number of variables considered.

**F-Statistic**

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship between speed and distance). The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables. In our example the F-statistic is **73.82** which is relatively larger than 1 given the size of our data.

**p-value**

p-value: < 2.2e-16

Overall p value on the basis of F-statistic, normally p value less than 0.05 indicate that overall model is significant