 **Problem Statement**

**#1. Use the given link Data Set.**

yeast <- read.table("C:/Users/Mymaster/Desktop/Data Analytics/Assignments\_Questions/yeast.txt", quote="\"", comment.char="")

View(yeast)

yeastdata <- yeast

dim(yeastdata)

install.packages("reshape")

View(yeastdata)

library(reshape)

yeastdata <- rename (yeastdata, c(V1 = "Sequence Name",

V2 = "mcg",

V3 = "gvh",

V4 = "alm",

V5 = "mit",

V6 = "erl",

V7 = "pox",

V8 = "vac",

V9 = "nuc",

V10 = "Class Distribution"))

View(yeastdata)

#Answer the below questions:

**# a. What are the assumptions of ANOVA, test it out?**

To use the ANOVA test we made the following assumptions:

1) Each group sample is drawn from a normally distributed population

2) All populations have a common variance

3) All samples are drawn independently of each other

4) Within each sample, the observations are sampled randomly and independently of each other

5) Factor effects are additive

#for checking normality assumption

x <- yeastdata$nuc

h <- hist(x, breaks = 10, col = "red", xlab = "Score of discriminant analysis of nuclear localization signals

of nuclear and non-nuclear proteins", main = "Histogram of Score with normal curve")

xfit <- seq(min(x), max(x), length= 40)

yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))

yfit <- yfit\*diff(h$mids[1:2]\* length(x))

lines(xfit, yfit , col = "blue", lwd = 2)

#for checking skewness or kurtosis and variances

library(psych)

describe(yeastdata)

#for checking outliers

boxplot(yeastdata)

**# b. Why ANOVA test? Is there any other way to answer the above question?**

ANOVA allows researcher to evaluate all the mean differences in a single hypothesis test using a single ??-level and thereby keep the risk of a Type I error under control, no matter how many different means are being compared.

A regression analysis will accomplish the same goal as an ANOVA.