Assignment 12.2

Name : Y Vasudev

Batch : DA with R , Excel and Tableau

1. Use the given link Data Set.

Answer the below questions:

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

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

yeastdata <- read.table("D:/BIG DATA/DATA ANALYTICS WITH R, EXCEL & TABLEAU/12 NON-LINEAR MODELS/yeast\_data.txt", quote="\"", comment.char="")

View(yeastdata)

dim(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)

#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.

#The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant

#differences between the means of three or more independent (unrelated) groups

#The one-way ANOVA compares the means between the groups you are interested in and determines whether

#any of those means are statistically significantly different from each other

#In the situation where there multiple response variables you can test them simultaneously using a

#multivariate analysis of variance (MANOVA)

#As in salescity dataset we see that our categorical variables has more than 2 levels hence we are using

#anova t test ,otherwise if our independent variables that is city here in salescity dataset

#has got only 2 levels than we used "independent sample t test

#Thus we can use independent sample t test also if our independent variables has got only 2 levels

#the other ways for testing which we can used are by plotting

#histogram

#scatterplot

#box plot

#qq plot

#by this too

#Levene's test

#Fligner Killeen test

#Bartlett's tes