EXPLORATORY DATA ANALYSIS

1)We took out the following analysis as part of exploratory data analysis-

a)Histogram-A **histogram** is a bar graph of raw data that creates a picture of the data distribution.There is a peak at a point where intensity is e^6.When intensity is e^6 it reaches to a 0.8 density point.There is a maxima at a point where intensity is e^6.Density increases when intensity increases from 0 to e^6 and density decreases when intensity increases from e^6 to infinity.Histogram was plotted of the microarray data set (info).

b)QQ Plot-A **Q-Q plot** is a scatterplot created by **plotting** two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.This QQ Plot is roughly straight as both quantiles come from the same distribution.We have plotted sample quantiles with theoretical quantiles of the distribution of micro array data set.QQ Plot was plotted for gdata i.e.data frames of microarray data set (info).

QQ Line Plot-It has a red line drawn on QQ Plot showing the slope of QQ Plot.

c)Clustering Analysis-Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.Here there is a decreasing trend i.e. as the number of clusters increases within group sum of squares decreases.Clustering analysis was done for gdata i.e. data frames of microarray data set.

d)Correlation Plot-Here the matrix shows how similar are two arrays.There are 16 arrays and correlation of each array to each array is displayed in the matrix.So (1,1),(2,2), …. Coordinates of the matrix have red boxes showing full correlation as it has array being compared to itself.Here we plotted the data of 16 columns of the data set which were treated as 16 arrays.There were only 16 columns in the data set so a 16*16 matrix was drawn in correlation plot.

e)RNA Degradation Plot-Here as probe number increases, Mean Intensity also increases.It was plotted for the microarray data set.

f)Scatter Plot-A **scatter plot** is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.I have plotted scatter plot of FPKM and also plotted a scatter plot of FPKM vs count.In scatter plot of FPKM majority of the dots have 0 FPKM and some have FPKM slightly above 0.I have also plotted a scatter plot of all combinations of gene symbol,FPKM and count along with a red line showing their slope.

g)Bar Plot-A **bar plot** is a **plot** that presents categorical data with rectangular **bars** with lengths proportional to the values that they represent.I have plotted two types of bar plot-1)Horizontal Bar Plot and 2)Vertical Bar Plot.In vertical bar plot there are few bars bigger than the average height of the bars.Same observation is being observed in horizontal bar plot.I have plotted the bar plot for data having column name as “count”.

h)Density Plot-A **Density Plot** visualises the distribution of data over a continuous interval or time period.A **density plot** is a representation of the distribution of a numeric variable. It uses a kernel **density** estimate to show the probability **density** function of the variable .It shows that for all “FPKM’s” we have N=0 and for all values of N we have 0 “FPKM’s” so we get a L shaped graph.
In rugged density plot we have small vertical lines showing the concentration of points.

i)Line Plot-A **line plot** is a **graph** that shows frequency of data along a number **line**. Continuous line plot and dotted line plot are two line plots plotted for the data.Both line plots were plotted for “count” variable.

j)BoxPlot-The **box plot** is a standardised way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.I have plotted two Box Plots.One before the data was normalised and one after the data was normalised.Normalisation of data removes its outliers and so the data gets transformed into a normalised data. In the vertical column of the plot we have log2Intensity.We can see that horizontal line of each box plot gets aligned after normalisation.Box plot was plotted for the data of microarray data set which was normalised for further analysis.