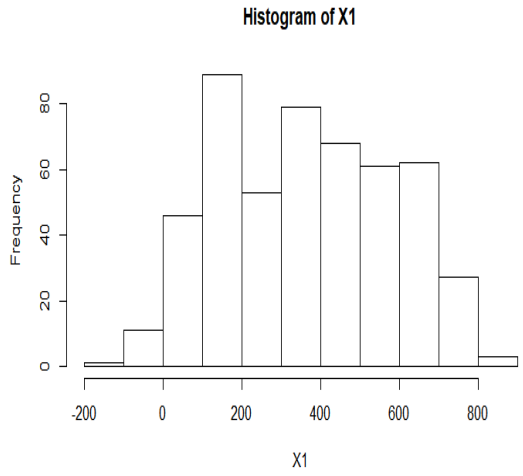
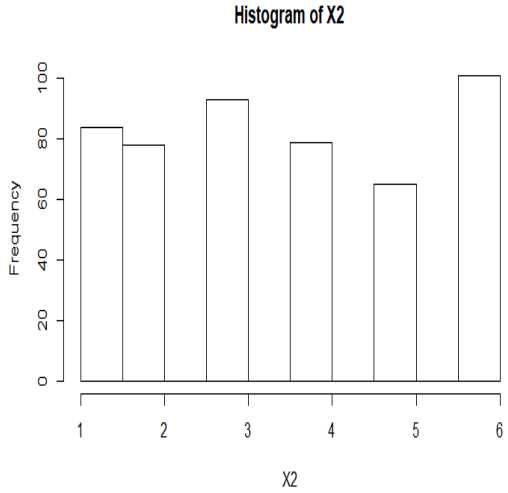
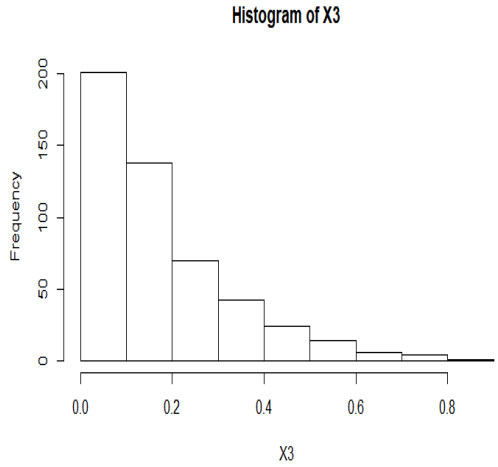
Task 1. Basic statistics analysis

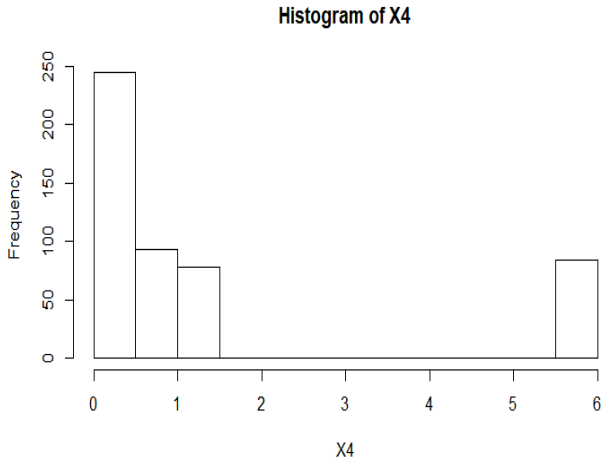
1.1. <u>For each variable Xi, i.e. column in the data set corresponding to Xi, calculate the following: Histogram, mean, variance.</u>

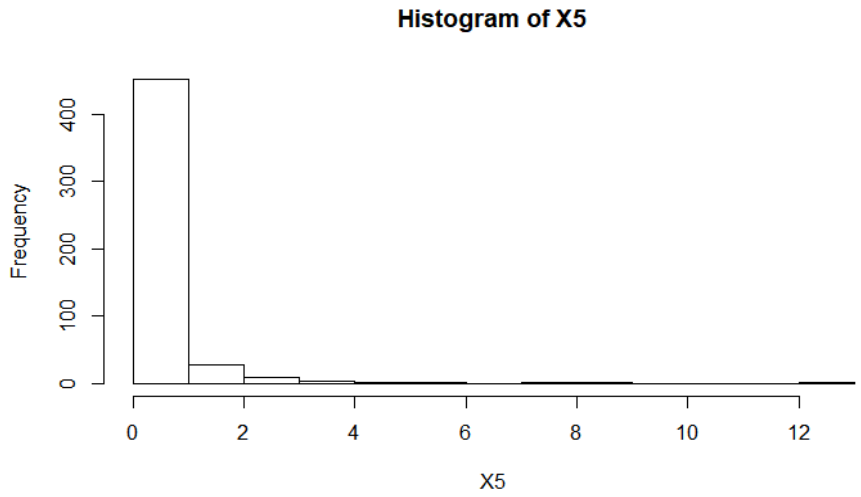
X_i	Mean	Variance
X_1	365.5421	46943.41
X_2	3.532	3.04707
X_3	0.1732864	0.02495294
X_4	1.490117	4.305147
X_5	0.3777813	1.094548



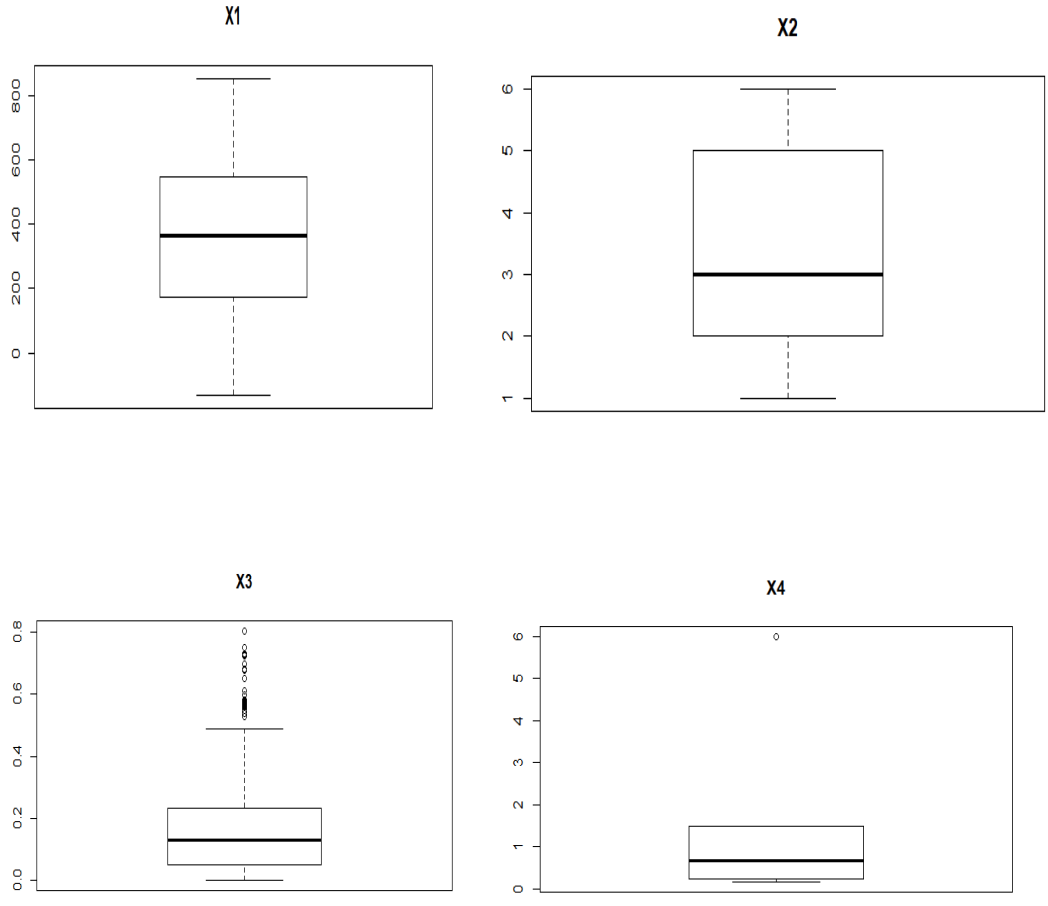


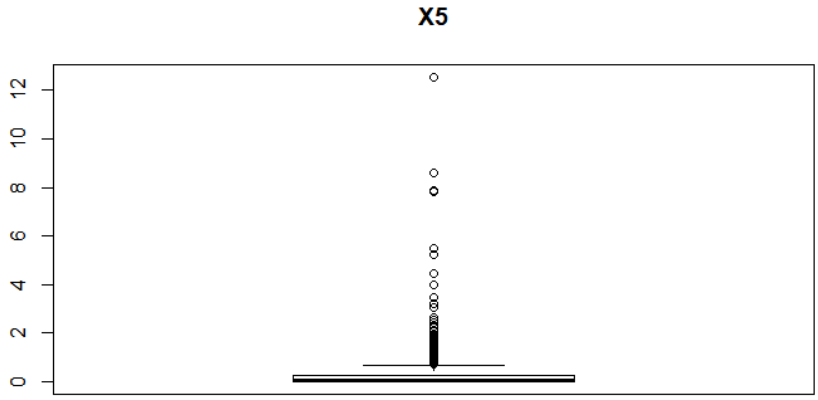






1.2 Use box plot or any other function to remove outliers (do not over do it !), or you can do that during the model building phase (tasks 2 and 3).





Outliers are not removed here as we don't know if these independent variables are statistically significant or not for our final regression model.
Later, we find that the variables having outliers are discarded from the final model.

1.3 Calculate the correlation matrix Σ among all variables, i.e., Y, X1, X2, X3, X4 and X5. Draw conclusions related to possible dependencies among these variables.

Our correlation matrix is as follows :

	X1	X2	X3	X4	X5	Y
X1	1.00	-0.04	0.02	0.04	0.33	1.00
X2	-0.04	1.00	-0.01	-0.78	0.06	-0.04
X3	0.02	-0.01	1.00	-0.01	0.08	0.02
X4	0.04	-0.78	-0.01	1.00	-0.04	0.04
X5	0.33	0.06	0.08	-0.04	1.00	0.33
Y	1.00	-0.04	0.02	0.04	0.33	1.00

The correlation co-efficient shows how strong the linear relationship between two variable s is. Positive correlation implies both the variables are moving in same direction. Negative c orrelation implies, when one variable increases the other variable decreases. If correlation i s close to +1 or -1, high degree of correlation or the association between the variables is str ong. Otherwise, it means weak correlation.

In our matrix, we can see that, X2 and X4 are strongly correlated as their correlation coeffic ient is 0.78. And Y(the dependent variable) is strongly correlated t independent variable X1 as their coefficient of correlation is 1.

1.4 Comment on the results

In the basic statistics analysis, we calculated the mean and variance for each predictor. Then we plotted them on histogram to analyze their trends and looked for outliers by using box plot. X3, X4 and X5 have outliers but later, we find that our final model do not have these predictors , so we don't need to remove their outliers.