Part 1

1. What is the sample size of your data set?

Sample Size: 1023

1. Explain why the variable X should not be treated as a predictor variable. (What is X most likely?)

From the data set we can see that variable X contains serial number from 1 to 1023 numeric values. As X contains numeric values like indexing, they might not provide a meaningful update while predicting outputs.

1. If we remove the variable X, how many numerical variables are there?

13

1. If we remove the variable X, how many categorical variables are there?

4

1. For each numerical variable in the data set (excluding the variable X), determine whether the distribution is unimodal, bimodal, or multimodal. Also state whether each distribution is most likely uniform symmetric, not uniform but symmetric, left skewed, or right skewed.

X1:

unimodal

not uniform but symmetric

X2:

unimodal

not uniform but symmetric

X3:

Unimodal

Right skewed

X4:

Unimodal

Right skewed

X5:

Unimodal

not uniform but symmetric

X6:

Multimodal

not uniform but symmetric

X7:

Unimodal

not uniform but symmetric

Y1:

Unimodal

not uniform but symmetric

Y2:

Unimodal

Not uniform but symmetric

Y3:

Unimodal

Not uniform but symmetric

Y4:

Multimodal

Right skewed

Y5:

Bimodal

Not Uniform but symmetric

Y6:

Unimodal

Not Uniform but symmetric

1. For which categorical/numerical variable pairs) do the boxplots indicate a significant difference between groups (look at both center and spread)?

Boxplots that are indicating a significant difference between groups:

(X8,X1), (X8,X5),(X8,Y3),(X8,Y5),(X8,Y6),(X9,Y5),(X10,Y5),(Y7,Y6)

1. For each of the categorical variables, convert the counts in the summary display to percentages.

X8 Yes: 94.48%

X8 No: 5.52%

X9 Small: 48.77%

X9 Medium: 47.31%

X9 Large: 3.9%

X10 High: 14.66%

X10 Low:55.03

X10 Moderate:30.30%

Y7 Group1: 45.94%

Y7 Group2: 54.05%

1. For each of the numerical variables, determine whether the mean or the median is the appropriate summary measure to use for the center of the distribution. (Consult your histograms.)

Measure to use center of distribution

X1: Median & Mean

X2: Median

X3: Median

X4: Mean

X5: Mean

X6: Mean

X7: Median & Mean

Y1: Mean

Y2: Mean

Y3: Mean

Y4: Median

Y5: Median

Y6: Mean

Part 2

1. For each of the ten linear models that individual one created, determine whether the model utility test would reject or fail to reject the null hypothesis that the slope

parameter is zero. You may use an a = 0.05.

lmY2.1- Reject null hypothesis

lmY2.2- Reject null hypothesis

lmY2.3- Reject null hypothesis

lmY2.4- Reject null hypothesis

lmY2.5- Fail to Reject null hypothesis

lmY2.6- Fail to Reject null hypothesis

lmY2.7- Fail to Reject null hypothesis

lmY2.8- Fail to Reject null hypothesis

lmY2.9- Fail to Reject null hypothesis

lmY2.10- Fail to Reject null hypothesis

1. For each of the ten linear models, what proportion of variability is explained by the model?

Proportion of Variability explained by each model by R square

lmY2.1- 36.09%

lmY2.2- 32.12%

lmY2.3- 22.64%

lmY2.4- 24.37%

lmY2.5- 0.0897%

lmY2.6- 0.09312%

lmY2.7- 0.02198%

lmY2.8- 1.076e-03

lmY2.9- 0.08967%

lmY2.10- 0.05706%

1. For the multiple linear regression model that individual one created, determine whether there is a relationship between the collection of predictors and the response Y2. What single numerical value do you look at to determine if there is a relationship? What was that numerical value for this model?

lmY2.Full- For this multiple linear regression we have an F statistic 8.88 e+0.4 which is >1. As the F statistic is very large we reject null hypothesis and we conclude that there is a relationship between predictors and response variable Y2.

1. What proportion of variability is explained by the full model?

99.91% of variability is explained by full model

1. Which of the predictors (if any) should we consider eliminating from the model? Which of these predictors (if any) should be eliminated first?

For our model we should eliminate X3, X5, X7, X9small, X9 Medium, X8 Yes.

X3 has the highest p-value so we need eliminate X3 first.

1. Of the eleven models that individual one created for predicting Y2, which one performs best? Explain.

Out of 11 models lmY2.Full will perform better because it has high R square value.

1. For each of the six two-predictor linear models that individual two created, determine whether there is a relationship between the collection of predictors and the response Y6. What single numerical value do you look at to determine if there is a relationship?

What was that numerical value for this model?

lmY6.12- yes there is significant relationship between predictors and Y6.(F-statistics: 5727)

lmY6.13- yes there is significant relationship between predictors and Y6.(F-statistics: 13.36)

lmY6.14- yes there is significant relationship between predictors and Y6.(F-statistics: 44.2)

lmY6.15- yes there is significant relationship between predictors and Y6.(F-statistics: 12.5)

lmY6.16- yes there is significant relationship between predictors and Y6.(F-statistics: 12.51)

lmY6.17- yes there is significant relationship between predictors and Y6.(F-statistics:11.53)

To determine relationship, we are looking at F-Statistic.

1. For each of the six two-variable linear models, what proportion of variability is explained by the model?

Proportion of variability explained by the model:

lmY6.12- 91.82%

lmY6.13- 2.55%

lmY6.14- 7.97%

lmY6.15- 2.43%

lmY6.16- 2.39%

lmY6.17- 2.17%

1. For the models with interaction terms that individual two created, determine whether there is a relationship between the collection of predictors and the response Y6. What single numerical value do you look at to determine if there is a relationship? What was that numerical value for this model?

lmY6int12- F-Statistic(4.299e+0.4).Yes there is a relationship between predictors and response Y6

lmY6int13- F-Statistic(9.137). Yes there is a relationship between predictors and response Y6

lmY6int14- F-Statistic(32.53). Yes there is a relationship between predictors and response Y6

lmY6int15- F-Statistic(9.516). Yes there is a relationship between predictors and response Y6

lmY6int16- F-Statistic(9.359). Yes there is a relationship between predictors and response Y6

lmY6int17- F-Statistic(7.975). Yes there is a relationship between predictors and response Y6

1. For each of the six two-variable linear models, what proportion of variability is explained by the model?

lmY6int12- 99.22%

lmY6int13- 2.61%

lmY6int14- 8.74%

lmY6int15- 2.72%

lmY6int16- 2.68%

lmY6int17- 2.29%

1. For each of the six models with interaction terms, determine if the interaction term

should remain in the model or if it should be eliminated. You may use an a = 0.05.

lmY6int12- Pvalue(<2.2e-16). Significant

lmY6int13- Pvalue(<5.175e-06). Significant

lmY6int14- Pvalue(<2.2e-16). Significant

lmY6int15- Pvalue(<3.346e-06). Significant

lmY6int16- Pvalue(<4.178e-06). Significant

lmY6int17- Pvalue(<2.943e-05). Significant

1. Of the twelve models that individual two created for predicting Y6, which one performs best? Explain.

Out of 12 models lmY6int12 is the best model because it has highest R square.

1. For each of the seven linear models that individual three created, determine whether the model utility test would reject or fail to reject the null hypothesis that the slope

parameter is zero. You may use an a = 0.05.

For X1 - fail to reject

For X2 - fail to reject

For X3 - reject

For X4- reject

For x5 - fail to reject

For X6 - fail to reject

For X7- fail to reject

1. For each of the seven linear models, what proportion of variability is explained by the model?

X1- 0.2136% of the variability in the dependent variable

X2 - 0.12% of the variability in the dependent variable

X3 - 24.58% of the variability in the dependent variable

X4 - 22.89% of the variability in the dependent variable

X5 - 0.0467% of the variability in the dependent variable

X6 - 0.03506% of the variability in the dependent variable

X7 - 0.3078% of the variability in the dependent variable.

1. For each of the seven quadratic models that individual three created, determine whether there is a relationship between the collection of predictors and the response Y4. What single numerical value do you look at to determine if there is a relationship? What was that numerical value for this model?

X1- p value (0.2956) > 0.05. So we fail to reject null hypothesis. Hence no relation between predictor and response.

X2 - p value (0.304) > 0.05. So we fail to reject null hypothesis. Hence no relation between predictor and response.

X3 - p value (~0) < 0.05. So we reject null hypothesis. Hence there is a relation between predictor and response.

X4 - p value (~0) < 0.05. So we reject null hypothesis. Hence there is a relation between predictor and response.

X5 - p value (0.6233) > 0.05. So we fail to reject null hypothesis. Hence no relation between predictor and response.

X6 - p value (0.7361) > 0.05. So we fail to reject null hypothesis. Hence no relation between predictor and response.

X7 - p value (0.03296) < 0.05. So we reject null hypothesis. Hence there is a relation between predictor and response.

1. For each of the seven quadratic models, what proportion of variability is explained by the model?

X1 - 0.2387% of the variability in the dependent variable

X2 - 0.2332% of the variability in the dependent variable

X3 - 74.25% of the variability in the dependent variable

X4 - 53.7% of the variability in the dependent variable

X5 - 0.09266% of the variability in the dependent variable

X6 - 0.06005% of the variability in the dependent variable

X7 - 0.6669% of the variability in the dependent variable

1. Of the fourteen models that individual three created for predicting Y4, which one performs best? Explain.

Quadratic model with X3 performs better compared to other model because the R^2 squared value is 0.7425.

Part 3

1. What are the prior probabilities of the two groups.

Prior probabilities of two groups:

Group1: 0.459433

Group2: 0.540567

1. Determine the proportion of time that the logistic regression model was accurate and the proportion of time that the logistic regression model was inaccurate on the full data set.

Proportion of time that logistic regression model was accurate:

93+475/1023=93.46

Proportion of time that logistic regression model was accurate:

377+78/1023=377.07

1. Determine the proportion of time that the Ida model was accurate and the proportion of time that the Ida model was inaccurate on the full data set.

Proportion of time that lda was accurate:

93+475/1023=93.46

Proportion of time that lda was accurate:

377+78/1023=377.07

1. Determine the proportion of time that the qda model was accurate and the proportion of time that the qda model was inaccurate on the full data set.

Proportion of time that qda was accurate:

198+387/1023=198.37

Proportion of time that qda was accurate:

272+166/1023=272.16

1. Which of the three techniques produced the best fit on the full data set?

QDA performs best when compared to LDA and Logistic because QDA's accuracy is 57.13%