**Team Number:**

**Team Captain:**

**Team Members:**

The data set used in this activity is “HeartDisease.txt”. The data set used in this study is from South African Heart Disease Data (http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html), a subset `of the Coronary Risk-Factor Study (CORIS) of the Western Cape, South African. The aim of this study was to establish the intensity of ischemic heart disease risk factors in the high-incidence region. (Note: the target variable for this study is “chd”.

**Problem 1 Programming and Reporting (20 Points)**

**PART I Programming (10 Points)**

1. Read the data into your software system
2. Examine univariate statistics for the following variables: sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, and age. (not including the target variable)
3. Produce histogram of each of the following variables with imposing normal curve: sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, and age.
4. Produce quantile plot of each of the following variables: sbp, tobacco, ldl, adiposity, typea, obesity, alcohol, and age.
5. Build a logistic regression model with all predictors.
6. Perform power transformation on the following variables: sbp (power = -2), tobacco (power = 0.4), ldl (power = 0.1), obesity (power = -0.4), and alcohol (power = 0.4).
7. Produce histogram of each of the following transformed variables with imposing normal curve: sbp, tobacco, ldl, obesity, and alcohol.
8. Produce quantile plot of each of the following transformed variables: sbp, tobacco, ldl, obesity, and alcohol.
9. Build a logistic regression model with all predictors (transformed and **three** remaining original variables do not perform any transformation).
10. Build another logistic regression model with all predictors as in Part 9 except using significant predictors only.

**PART II Reporting**

1. After completion of this activity, complete the following table. (3 Points)

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Mean | Median | Skewness |
| sbp |  |  |  |
| tobacco |  |  |  |
| ldl |  |  |  |
| adiposity |  |  |  |
| typea |  |  |  |
| obesity |  |  |  |
| alcohol |  |  |  |
| age |  |  |  |

1. Display the histogram and quantile plot of “tobacco”.
2. Display the histogram and quantile plot of “alcohol”.
3. Complete the following table. (3 Points)

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Mean | Median | Skewness |
| Sbp (Power=-2) |  |  |  |
| Tobacco (Power=0.4) |  |  |  |
| Ldl (Power=0.1) |  |  |  |
| obesity (power = -0.4) |  |  |  |
| alcohol (Power=0.4) |  |  |  |

1. Display the histogram and quantile plot of “tobacco” after power transformation.
2. Display the histogram and quantile plot of “alcohol” after power transformation.
3. Find the 95% confidence interval on the likelihood of heart disease if one more kilogram of tobacco consumed using the first model.
4. Find the 95% confidence interval on the likelihood of heart disease if one more kilogram of alcohol consumed using the first model.
5. Which model perform better based on the c-statistics?

**Problem #2 (5 Points)** Suppose that we collect data from a group of students in statistics with variables “Hours of Study” () and “Undergraduate GPA” (), and “Receiving A as Final Grade” (). We fit a logistic regression model with the following estimates:

1. Estimate the probability of a student who spent 10 hours per week with 3.5 undergraduate GPA.
2. How many hours would the student in part (a) need to study to have 80% or higher chance of getting A in this class?

**Problem #3 (5 Points)**

1. What is the fraction of people with an odd of 0.3 of defaulting on their credit card payment?
2. Suppose that John has 10% chance default in his credit card payment. What is the odd that John will default?