1 Question 2

The code used for these analyses is available at `https://github.com/AndrewTijua/IDA`.

Complete Case Analysis

Performing the complete case analysis we obtain

$\bar{y}$	19.27273
se(y)	2.603013
logdose $\sim$ recovtime	0.23912558
bloodp $\sim$ recovtime	-0.01952862

Mean Imputation

Performing mean imputation we obtain

$\bar{y}$	19.27273
se(y)	2.28413
logdose $\sim$ recovtime	0.21506117
bloodp $\sim$ recovtime	-0.01934126

Regression Imputation

Performing regression imputation we obtain

$\bar{y}$	19.44428
se(y)	2.312845
logdose $\sim$ recovtime	0.2801835
bloodp $\sim$ recovtime	-0.0111364

Stochastic Regression Imputation

We note that the recovery time cannot be less than 0, so we must resample from our distribution if this occurs.

We use the command `set.seed(1)` to get reproducible results. Using this we obtain

$\bar{y}$	19.75397
se(y)	2.60018
logdose $\sim$ recovtime	0.24563681
bloodp $\sim$ recovtime	-0.04772682

Predictive Mean Matching

Performing predictive mean matching we obtain

$\bar{y}$	19.44
se(y)	2.464467
logdose $\sim$ recovtime	0.30379446
bloodp $\sim$ recovtime	-0.03208685

An advantage of predictive mean matching is that the imputed value is a definite possible value of the data. There are methods that work with categorical data, of which there are not any in the same vein as stochastic regression.

Some problems with predictive mean matching is that there is no mathematical theory justifying it. There is no mathematical proof showing that it is valid in a given situation. It relies on MC simulation, which is inherently flawed in that it cannot explore all possibilities. Another is that the method implemented here chooses the nearest neighbour. The more accepted version randomly selects from k nearest neighbours. This is not enough to produce proper imputation, leading to artificially low standard errors and inflated test statistics (in our case the Pearson correlation).