1. Two models are applied to a dataset that has been partitioned. Model A is considerably more accurate than model B on the training data, but slightly less accurate than model B on the validation data. Which model are you more likely to consider for final deployment?

2. **Use your individual dataset.   
Comment on all results and include readable screenshots from RStudio.**

1. Set your work directory where you have your individual data.

A screen shot of a computer

Description automatically generated

1. Read your individual dataset into RStudio. For example:

A screenshot of a computer

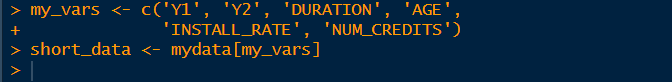
Description automatically generated

1. Install and load any libraries you need. For example:   
   **install.packages("ggplot2")**  
   **library(ggplot2)**

A screenshot of a computer

Description automatically generated

1. Select 7 variables ( Y1, Y2 and 5 other continuous variables) and create a new file.   
   Call the new file “**small**”. From now on use the “**small**” data file.





3.1. Produce a scatterplot of Y1 and one of the another continuous variables (not binary 0/1). Comment on the result.

A graph of numbers and dots

Description automatically generated with medium confidence

3.2. Produce a histogram of Y1. Comment on the result.

A graph of a graph

Description automatically generated

3.3. Produce a boxplot of Y1. Comment on the result.

A diagram of a graph

Description automatically generated

3.4. Produce a histogram of Y2. Comment on the result.

A graph with a bar and a number of numbers

Description automatically generated

3.5. Produce a heatmap with correlations for Y1 and another continuous variables. Comment on the result.

A yellow and orange squares with black text

Description automatically generated

|  |  |  |
| --- | --- | --- |
| VARS | CORRELATIONS | DESCRIPTIONS |
| Y1 and Y2 | -0.14 | Weak negative linear relationship |
| Y1 and CHK\_ACCT | -0.04 | Very weak negative linear relationship |
| Y1 and DURATION | 0.64 | Moderate positive linear relationship |
| Y1 and AGE | 0.02 | Very weak positive linear relationship |
| Y1 and INSTALL\_RATE | -0.27 | Moderate negative linear relationship |
| Y1 and NUM\_CREDITS | 0.02 | Very weak positive linear relationship |

3.6. Partition the dataset into training set (60%) and validation set (40%).   
Call the new datasets: "mytrain" and "myvalid" respectively.

A screenshot of a computer program

Description automatically generated

3.7. Produce histograms for Y1 separate for the training set and for validation set.

histograms for Y1 in the **training** set

A graph of a graph

Description automatically generated

histograms for Y1 separate in **validation** set.

**A graph of a graph

Description automatically generated with medium confidence**