R Programming: Homework 2

Hieu Pham

Instructions

You can submit your homework in one of two ways.

1. You can fill in the missing code blocks directly in this .Rmd file (be sure to change the file name)
2. You can create a new .R file and clearly label you answers.

Getting Started

In this assignment, we will be working with a stroke dataset which provides details about people who had strokes. Most columns are self-explanatory, but for more data details visit this website: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

The packages listed below are simply suggestions, but please edit this list as you see fit.

```
## you can add more, or change...these are suggestions
library(tidyverse)
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
```

Problem Set

1. Read in the `strokedata.csv` dataset, and remove any rows with missing values.

2. Create two histograms using `ggplot2()`.

a. Showing the distribution of strokes
b. Showing the distribution of age

3. Split your stroke.df dataframe into an 85/15 train/test split with a seed of 123.

4. Complete the following:

1. Create a logistic regression model on the response variable `stroke` using all columns as features
2. Print out a summary of your model
3. Which features are significant?

5. Using the logistic regression model complete the following:

1. Predict on your testing data frame
2. Compute your testing accuracy