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)

# added libraries
library(caret)
library(stats)
library(ROCR)
```

Problem Set

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

```
#   Read
strokedata <- read_csv("strokedata.csv")
```

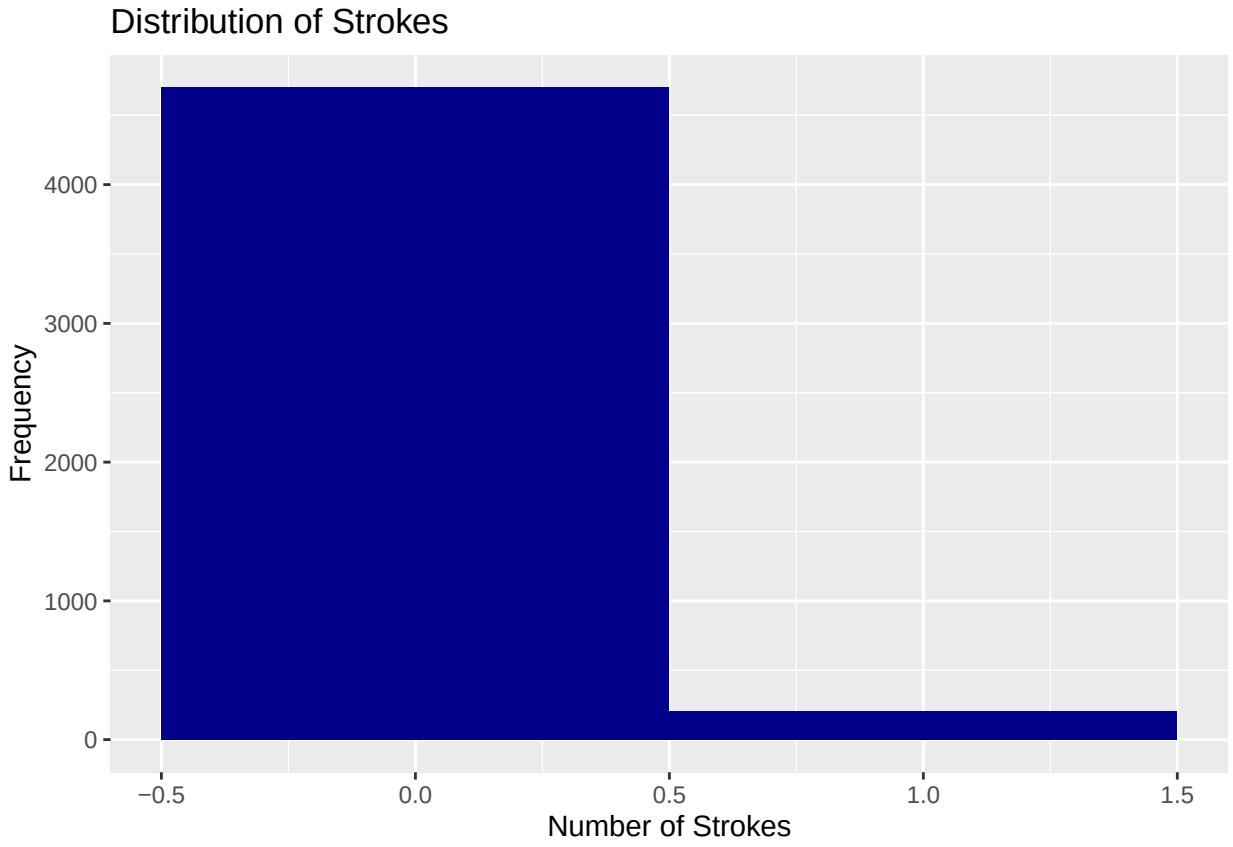
```
## Rows: 5109 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): gender, ever_married, work_type, Residence_type, smoking_status
## dbl (7): id, age, hypertension, heart_disease, avg_glucose_level, bmi, stroke
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# remove any rows with missing values.
strokedata_clean <- na.omit(strokedata)
```

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

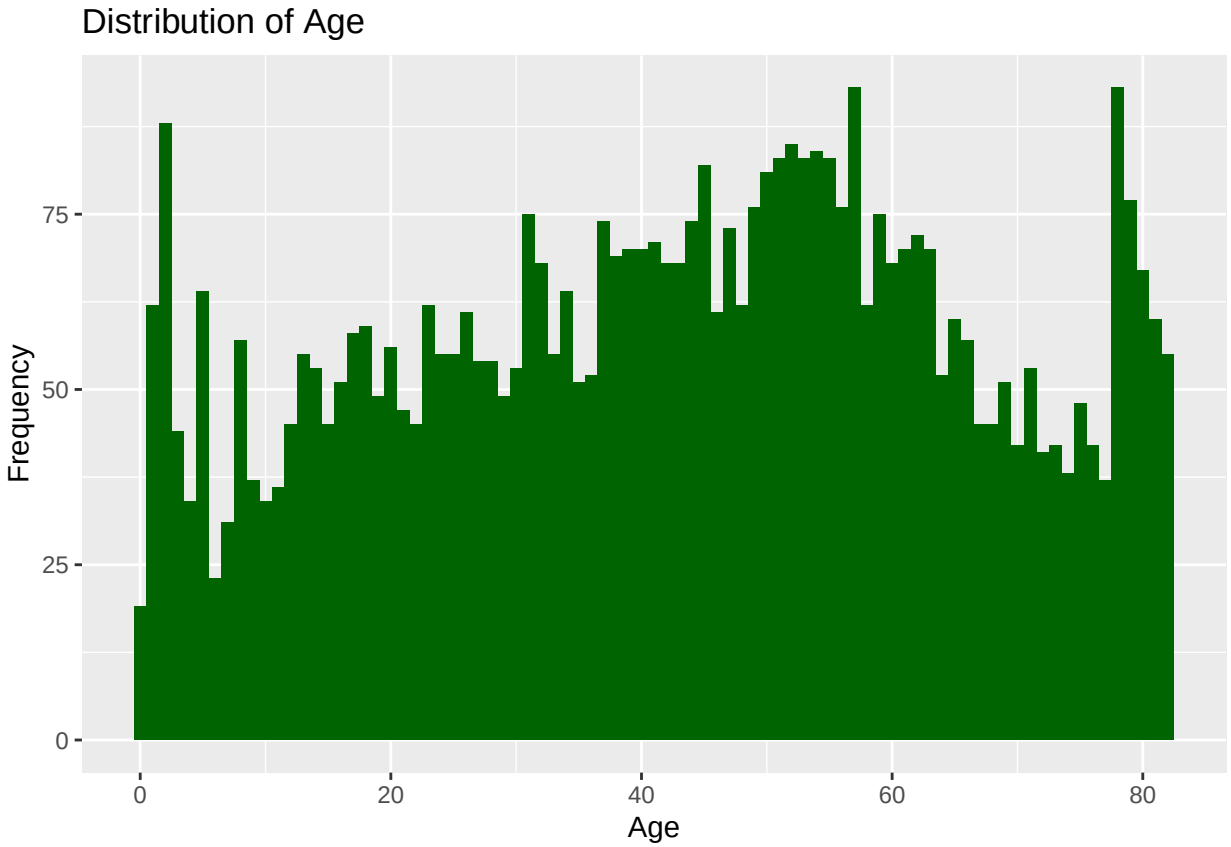
a. Showing the distribution of strokes

```
# Histogram showing the distribution of strokes
ggplot(strokedata_clean, aes(x=stroke)) +
  geom_histogram(fill="darkblue", binwidth=1) +
  labs(title="Distribution of Strokes", x="Number of Strokes", y="Frequency")
```



b. Showing the distribution of age

```
# Histogram showing the distribution of age
ggplot(strokedata_clean, aes(x=age)) +
  geom_histogram(fill="darkgreen", binwidth=1) +
  labs(title="Distribution of Age", x="Age", y="Frequency")
```



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

```
set.seed(123)
split <- createDataPartition(strokedata_clean$stroke, p = 0.85, list = FALSE)
strokedata_train <- strokedata_clean[split, ]
strokedata_test <- strokedata_clean[-split, ]
```

4. Complete the following:

1. Create a logistic regression model on the response variable `stroke` using all columns as features

```
# Create a logistic regression model
model <- glm(stroke ~ ., data=strokedata_train, family="binomial")
```

2. Print out a summary of your model

```
# Print out a summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = strokedata_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1896  -0.2895  -0.1483  -0.0728   3.4624
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -7.496e+00  1.088e+00  -6.889 5.63e-12 ***
```

```
## id                           -1.215e-06  3.886e-06  -0.313 0.754654
## genderMale                    8.469e-02  1.694e-01   0.500 0.617143
## age                           7.557e-02  7.114e-03  10.624  < 2e-16 ***
## hypertension                  4.765e-01  1.925e-01   2.475 0.013315 *
## heart_disease                 3.208e-01  2.283e-01   1.405 0.159922
## ever_marriedYes              -2.018e-01  2.676e-01  -0.754 0.450827
## work_typeGovt_job            -1.051e+00  1.142e+00  -0.920 0.357823
## work_typeNever_worked        -1.008e+01  3.410e+02  -0.030 0.976415
## work_typePrivate             -8.532e-01  1.125e+00  -0.758 0.448308
## work_typeSelf-employed       -1.248e+00  1.147e+00  -1.088 0.276565
## Residence_typeUrban           9.399e-03  1.651e-01   0.057 0.954604
## avg_glucose_level             4.777e-03  1.424e-03   3.354 0.000796 ***
## bmi                           1.345e-02  1.278e-02   1.052 0.292625
## smoking_statusnever smoked -2.571e-03  2.112e-01  -0.012 0.990289
## smoking_statussmokes          4.595e-01  2.532e-01   1.815 0.069504 .
## smoking_statusUnknown        -1.854e-01  2.717e-01  -0.683 0.494919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1433.7  on 4171  degrees of freedom
## Residual deviance: 1128.4  on 4155  degrees of freedom
## AIC: 1162.4
##
## Number of Fisher Scoring iterations: 14
```

3. Which features are significant?

```
# The following features are significant:
# - Age
# - Hypertension
# - Ever Married
# - Work Type (Govt Job)
# - Avg Glucose Level
# - Smoking Status (Smokes)

# These features are significant because they have a p-value of less than 0.05, indicating that they ar
```

5. Using the logistic regression model complete the following:

1. Predict on your testing data frame

```
# Predict on testing data frame
predictions <- predict(model, strokedata_test, type="response")
# Compute accuracy using AUC
pred <- prediction(predictions, strokedata_test$stroke)
pred
```

```
## A prediction instance
##   with 736 data points
```

2. Compute your testing accuracy

```
# Compute accuracy using AUC
performance <- performance(pred, "auc")
accuracy <- as.numeric(performance@y.values[[1]])
print(accuracy)
```

```
## [1] 0.8460735
```