STAA577: HW3

- 36 points
- Due: see Canvas for due dates
- Submit your HW by uploading a PDF or DOC file to Canvas. I recommend using the `HW3_yourname.Rmd` file as a template to submit your answers.
- Include all of your R code in an appendix, unless explicitly asked to show it. Your final document may include R output and graphics, but all code should be in the appendix.

---

1. (2 points) Textbook (An Introduction to Statistical Learning with Applications in R, by James, Witten, Hastie, and Tibshirani) problem: Exercises 4.7, Conceptual Question #2.

---

2. (3 points) Textbook (An Introduction to Statistical Learning with Applications in R, by James, Witten, Hastie, and Tibshirani) problem: Exercises 4.7, Conceptual Question #3.

---

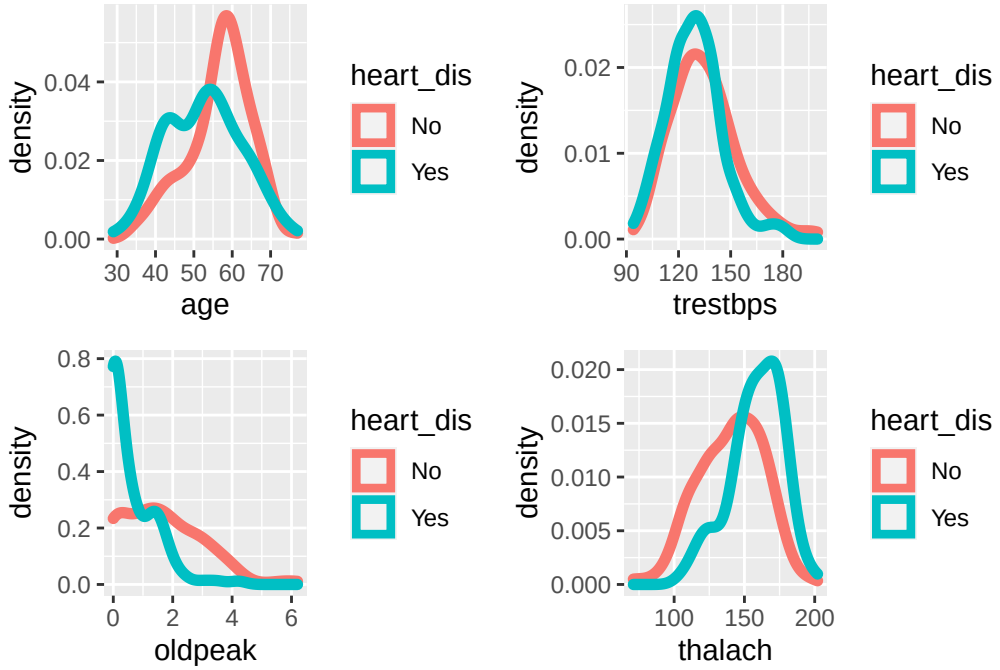
3. Recall the `heart` data from HW2. It contains information about 303 patients and whether or not they have heart disease. The variable `target` is 1 if the patient has heart disease and 0 if they do not. The data are split into a training and test set on Canvas. You can read more about the variables in the data here: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data. Be sure to treat the variables `sex, cp, fps, exang` and `restecg` as factors in both the training and test data sets.

   (a) (4 points) Use the training data to find a multiple logistic regression model using `age, sex, cp, trestbps, thalach, exang, oldpeak, ca`, and `thal` as predictors. Use a Bayes classificastion rule and report both the prediction **training accuracy** the prediction **test accuracy**.

   (b) (1 point) For plotting purposes, create a variable called `heart_dis` which is set to "yes" for patients with heart disease and "no" for patients without heart disease. Add this variable to your training data set.

```
heart_dis <- rep("Yes", dim(trainingdata)[1])
heart_dis[trainingdata$target == 0] <- "No"
trainingdata$heart_dis <- heart_dis
```

   (c) (5 points) Using `ggplot` and `geom_density()`, create four plots, one for each of the continuous predictors: `age, trestbps, oldpeak, and thalach` showing the density of the predictor for each of the two classes of heart disease status. Within each plot, use the `color` option in `geom_density()` to create two density plots, one for each of the heart disease groups. Save each figure as an individual graph, then use the `+` function in the `patchwork` package to arrange the four graphs into a single image (or you can use facetting). Here is an example of what your final plot should look like:



(d) (4 points) Use the training data to fit an LDA model using `age, sex, cp, trestbps, thalach, exang, oldpeak, ca`, and `thal` as predictors. Use a Bayes classification rule and report both the prediction **training accuracy** the prediction. **test accuracy**.

(e) (3 points) Repeat the previous problem using QDA.

(f) (2 points) Examine documentation for the `knn()` function in the `class} package. In order to use this function, you will need to create two new data sets (one for training, one for testing) that contain only the predictors`age, sex, cp, trestbps, thalach, exang, oldpeak, ca, and`thal. Create these two data sets. I recommend using a pipe and the`select()‘ function.

(g) (2 points) KNN predicts an observation's class by identifying the observations that are nearest to it. As a result, the scale of the variables matters. To control for differences in units and scale, we need to standardize our numerical variables. Use the `scale()` function to center and standardize the four numerical variables (`age, trestbps, oldpeak, thalach`) in both the training and test data sets.

(h) (4 points) Use the `knn()` function to use K-nearest-neighbors with $k = 14$ to classify the test data. Report only your **test accuracy**.

---

4. (6 points) Using the `Boston` data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median (use the code below). The testing and training data sets are provided in the code. Make sure to not include a crime rate as a predictor. Explore logistic regression, LDA, and KNN models using various subsets of the predictors. Describe your findings (at least 3 different takes is enough). Feel free to be creative.

```
library(MASS) ## data is in here
set.seed(10) ## reproducible

## take a look
head(Boston)

## create training and testing data (More on this in Week 4)
trn_samples <- sample(1:dim(Boston)[1], 440, replace=FALSE)
```

```r
training_Boston <- Boston[trn_samples,]
testing_Boston <- Boston[-trn_samples,]

## create a response variable
training_Boston$crimMedian <- training_Boston$crim > median(training_Boston$crim)
testing_Boston$crimMedian <- testing_Boston$crim > median(training_Boston$crim)
```
