THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

STAT3622 DATA VISUALIZATION

Assignment 1, due on February 16

Use R for Questions 1-3 and Python for Question 4. Attach all the codes.

1. Load the file "seeds.txt" in R. This is the dataset for creating graphics. The detailed information of this dataset can be found at

   https://archive.ics.uci.edu/ml/datasets/seeds.

   For each question, first show the plots and then interpret them.

   (a) Visualize the distribution for the categorical attribute "variety".

   (b) Visualize the distributional difference for each continuous attribute among three varieties of seeds.

   (c) Visualize the Euclidean distance matrix of samples involving all continuous attributes by the heatmap.

   (d) Create a new variable "flag", which takes the value "True" if the "area" is larger than 15, and "False", otherwise. Show a stacked bar graph of the sample size for each variety of seeds and how they are further divided out by "flag".

   (e) Show the scattered graph for "length.of.kernel" (x-axis) and "width.of.kernel" (y-axis) of all samples, where the colors of points indicate the varieties of seeds.

   (f) Show the multipanel scattered plots for "length.of.kernel" (x-axis) and "width.of.kernel" (y-axis) conditional on "variety".

2. Load the file "Boston.txt" in R. The detailed information of this dataset can be found at

   https://www.kaggle.com/c/boston-housing.

   For each question, first show the plots and then interpret them.

   (a) Visualize the distribution for all attributes.

   (b) Visualize the relationship between the variable "medv" and any other attribute. Which attributes may have impacts on the median values of owner-occupied homes ("medv")?

(c) Fit a linear regression model, where the response variable is "medv" (transformation is allowed). You can choose a subset of attributes or create new attributes as covariates. Interpret the resulting model.

(d) Using plots to check whether the model in (c) satisfies the assumptions for linear regression. Interpret the results.

3. Load the file "mpg.txt" in R. The detailed information of this dataset can be found at

   https://www.rdocumentation.org/packages/ggplot2/versions/3.3.3/topics/mpg.

   (a) Display a table of the average displacement ("displ") for each year.

   (b) Display a table of the median highway mileage ("hwy") per year for each type of car.

   (c) Display the first five observations in descending order by two attributes, displacement ("displ") and number of cylinders ("cyl").

   (d) Visualize the number of samples for each type of car per year with number of cylinders larger than 4 and fuel type equal to "r".

   (e) Given the type of car, the attributes could also be affected by the manufacturer. Explore graphically if this is the case.

4. Use Python packages to visualize the data (LC_Accept.csv and LC_Decline.csv) and interpret the results. Remember to add useful information into the figure, such as titles, legends, etc. If necessary, choose the appropriate form of the chart, or use multiple forms, for better illustration.

   (a) Plot the acceptance rates of loan applications over the months.

   (b) Draw a side-by-side bar chart of loan purposes grouped by the status of acceptance and declination.

   (c) Visualize the distributional difference for each attribute (Amount_Requested, Risk_Score, Debt_Income_Ratio, Employment_Length) between acceptance and declination.